Lecture Notes in Econometrics

Tobias J. Klein
Tilburg University

this version: May 6, 2026

Optimized for reading on a tablet.

Contents

Chapter 1

Introduction

1.1 These Lecture Notes

Put into one sentence, the idea of these lecture notes is to collect material that is useful for graduate students when they perform empirical work for their dissertation. Nowadays, most research in economics is empirical.[1] Here, we focus on analyzing behavior of firms or individuals at the individual level—as opposed to the aggregate level. We will do so based on models, either implicitly or explicitly. This makes us economists and this is why there is an "Econo" in "Econometrics". Models are simplifications of reality that help us to focus on the most important aspects of the problem we are analyzing, or the key determinants of behavior, with the ultimate goal to design rules or institutions that foster the well-being of individuals when they make decisions or interact in market or other environments. One way to think about this is that individuals want to travel from A to B. Most of them take the route *via* C. But emprical analysis shows that they are much faster when they take the route *via* D. A model here is simply a map together with the assumption that individuals value shorter travel times. It does not have to contain every single detail of the landscape, every elevation, or every curve.

[1] This may not be the impression students get when they attend courses. Here, of course, the focus is on methods. But this claim can easily be verified by browsing through recent tables of contents of the five leading journals in economics, which are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Review of Economic Studies*, and the *Quaterly Journal of Economics*.

But the map has to be a good depiction of the relationship between the points A, B, C and D. Empirical analysis here amounts to measuring the travel time between the varios points on that map. This is enough to fully understand the situation when the goal is to minimize travel time. And the welfare-improving institution would be a very simple one that informs individuals about the fact that the way *via* D is faster. Or a new set of street signs.

I've originally put these lecture notes together for an advanced course in microeconometrics that I taught in the academic year 2007/2008, just after I joined Tilburg University. I realized at that time that there are many good textbooks around. But I was always missing something in those, such as a more in-depth discussion of identification, links to the literature, a few words on the history, something on structural estimation, and well-documented computer code that one could use as a starting point for, say, a paper or thesis.

That is why, in these notes, rather than putting together another discussion of the standard material that is well covered in textbooks, I focus on those other topics. The general idea is to stay a bit closer to the literature than it is usually done, to sometimes discuss the history of the ideas a bit more, to provide a little bit of advice on how to do things in practice—or, in general research—, and overall, to look at a few things in more detail.

This deeper knowledge is important. Having it distinguishes you from regular master students and helps you even more to accomplish the following mission: always know what exactly you are doing and why you are doing it. Ask yourselves: what exactly do I learn from the estimates I just obtained? Why am I using this particular specification for the regression I am running? Why can I estimate this parameter in that way? And so on. It will become clear that I am convinced that one often needs a model—theory-–to accomplish this mission—either in the back of one's mind or explicitly spelled-out in the research paper.

The structure of the lecture notes is as follows. In the first chapter, I provide a glimpse into the history of econometrics and discuss why econometrics is not simply statistics applied to economic data. I will also give some advice in the form of references to readings and tools that are useful for learning how to do research. Then, in Part I, I turn to the basics of any empirical analysis, which are identification and estimation. Identification is the question whether some quantity of interest can be estimated in prin-

ciple. Then, we turn to estimation, in particular maximum likelihood estimation and the generalized method of moments; and how one can combine them with simulation techniques. Also nonparametric estimation and indirect inference are discussed. Many textbooks focus on estimation and keep the discussion of identification implicit. These lecture notes are a complement in that they first treat the two completely separately and they bring them together using, for instance, the *analogy principle*.

After the basics, identification and estimation, we turn to econometric models. The most well-known econometric model is the linear one that we already use when discussing identification and estimation. But more generally, an econometric model in its simplest form describes how an observed outcome or choice y_i made by individual or firm i comes about given observed variables x_i and unobserved variables ε_i. In the above example, the choice would be to travel from A to B either *via* C or D and the question would be how this observed choice depends on circumstances. x_i could be the weather conditions when i choses, or the time of the day, and ε_i could be factors related to individual i that influence his choice for a particular route. In this example, choice is between two alternatives, to go *via* C or D, which is also called binary choice. Part II focusses on econometric models of binary, ordered and multinomial choice; then censoring, truncation, sample selection and duration analysis; and finally policy evaluation. Along the way, it will become apparent that these models are the basis for more involved structural models, and we will see that understanding the basic models is of great help when it comes to understanding those more complex models and to constructing and estimating one oneself.

Before moving on, I would like to grasp this opportunity to thank my students in Tilburg for pointing out typos and errors in very early versions thereof. I am especially grateful to Péter Cziráki, Jan Kabátek, Geng Niu, Kebin Ma, and Georgios Petropoulos for their detailed comments. I should also say that despite the effort to eliminate errors, these lecture notes will still contain a number of them. Therefore, I would be grateful for further comments, also if they are of a more general nature, as well as suggestions of references and for topics that would fit and are not covered yet.

1.2 Textbooks and Further Readings

There are many comprehensive textbooks in econometrics. Wooldridge (2002) is often used in Ph.D. level courses in the U.S. and is a very classical textbook in the sense that it features prominently the linear model, asymptotics, testing, and so on. Davidson and MacKinnon (2003) is similar, but focuses more on derivations and is very stringent. Another excellent classic book is Goldberger (1991). Hayashi (2000) shows that many estimators can be expressed within the framework of the generalized method of moments. A very nice, more advanced, but comprehensive read in that respect is the Handbook of Econometrics chapter by Newey and McFadden (1994). Another very nice, advanced but still very clearly written book discussing the general approach to estimation by means of sample analogs is Manski (1988a).

Cameron and Trivedi (2005) has a stronger focus on microeconometrics. The price is that it covers the work horse, linear model and the basics less prominently. But it is still comprehensive in nature.

Verbeek (2004) and Stock and Watson (2003) cover a similar range of topics and are also accessible to students at the bachelor level. Still, Stock and Watson's book is a good pick for those who want to conduct less technical empirical research that involves natural experiments and differences-in-differences analysis. A book that has recently become popular among applied researchers is Angrist and Pischke (2009). It specializes in those topics.

Maddala (1983) is a classic book on microeconometrics and also McFadden (1984) provides a classic discussion. Train (2003) is a more recent book on discrete choice with a focus on modeling heterogeneity and estimation using simulation techniques. Gourieroux and Monfort (1996) is more comprehensive and in-depth.

The books by Pagan and Ullah (1999) and Li and Racine (2007) are good starting points for learning more about nonparametricand semiparametric estimation. Pagan and Ullah (1999) cover methods for density estimation, conditional moment estimation, nonparametric estimation of derivatives, as well as semiparametric estimation of single and simultaneous equation models, discrete choice models, selectivity models, and of censored regression models. Li and Racine (2007) is more comprehensive and maybe even more clearly written than Pagan and Ullah (1999).

1.3 Notation

The notation in these lecture notes is not fully unified. The reason is that in various places I wanted to stick to the conventions in the respective literature that I am discussing.

The vector of explanatory variables, x_i, is typically a row vector. The reason for this choice is that then, we can write $x_i\beta$ for the linear combination of elements of x_i, with β being a column vector. For the definition of matrix derivatives, we follow Magnus (2010). In brief, the $M \times N$ matrix of derivatives of a function $f : \mathbb{R}^N \mapsto \mathbb{R}^M$ with respect to its $N \times 1$ argument x is the Jacobian matrix $\partial f(x)/\partial x'$.[2] If f is scalar-valued, that is if $M = 1$, then this matrix will be a $1 \times N$ row vector. Its transpose will be denoted by $\partial f(x)'/\partial x$. Note that the Jacobian matrix is the matrix of first partial derivatives, as opposed to the Jacobian, which is its determinant. If f is scalar valued, then $\partial f(x)/\partial x$ is its gradient, a column vector of length M, and $\partial^2 f(x)/\partial x \partial x'$ is its Hessian. Appendix A.3 provides more details on matrix differentiation.

[2] Since x_i is already a row vector, x_i will take the place of x'.

Chapter 2

The Big Picture and Useful Resources

2.1 Econometrics

Wikipedia says that econometrics is the discipline that "aim(s) to give empirical content to economic relations", citing the *New Palgrave Dictionary in Economics*. A subdiscipline of it is theoretical econometrics, which is concerned with statistical properties of econometric procedures, among other things. But econometrics is much broader and also contains applied econometrics.

The most prestigious journal of our field is *Econometrica*, and when you visit the website looking for the "aims and scope" of it, you will find that it "promotes studies that aim at the unification of the theoretical-quantitative and the empirical-quantitative approach to *economic problems* and that are penetrated by constructive and rigorous thinking" (emphasis added; you find a similar statement for the *Journal of Econometrics*; see also Amemiya (2009) for an analysis of trends within the field). Interestingly, the ultimate goal is given here, namely to study economic problems. Studying statistical properties of estimators is part of it, but ultimately we want to answer questions that are interesting in the eyes of economists.

The beginnings go back more than 100 years. In the late 19th century, economists were already concerned with the empirical analysis of business cycles. They were also

interested in estimating demand, as can be seen in Philip G. Wright's (1928) book and instrumental variables analysis can be traced back at least to an appendix in this book.

Another important development was the foundation of the Cowles Commission in 1932 in Colorado Springs, Colorado. Alfred Cowles, a businessman and economist, initiated this. One reason was that he was frustrated by stock market forecasts, and indeed he finds in Cowles (1933), entitled "Can stock market forecasters forecast?" that from 1928 to 1932 there is no evidence that forecasts were any better than random guesses. An early member of the Cowles Commission was Harold Davis, who is often cited with the statement "Science is measurement," highlighting the importance of empirical work.

Around the same time, in 1936, the Dutch economist Jan Tinbergen developed the first national comprehensive macroeconomic model. Years later, in 1969, he obtained the first Nobel prize together with the Norwegian economist Ragnar Frisch "for having developed and applied dynamic models for the analysis of economic processes."

Another important development was the foundation of the Econometric Society in 1930, amongst others by Yale economist Irving Fischer and the Norwegian economist Ragnar Frisch. Fisher worked on time series (1927) and regression analysis (1934), and shaped monetary economics. As an aside, he infamously claimed just prior to the stock market crash in 1929 that the stock market had reached "a permanently high plateau." Yet another early contributor is Trygve Haavelmo who got the Nobel prize in 1989 "for his clarification of probability theory foundations of econometrics and his analyzes of simultaneous economic structures."

Tjalling C. Koopmans, a Dutch economist born in 's-Graveland, North Holland, and a student of Tinbergen, joined the Cowles Commission in 1940 in Chicago, became its director in 1948 and initiated a move to Yale in 1955. In the 1950s, the Cowles Commission started to talk about "theory and measurement." He obtained the Nobel prize jointly with Leonid Kantorovich for his contributions to the field of resource allocation, specifically the theory of optimal use of resources.

In case you are interested in diving deeper into this, there are many articles and books on the history, including Stigler (1949), Stigler (1965), Goldberger (1972), Cargill (1974), Hildreth (1986), Epstein (1987), Morgan (1990), Christ (1994), and Ambirajan (1995). We will occasionally take a more historical perspective in the following chapters, beginning with the next subsection.

2.2 Structural Models and Reduced Forms

The general approach in economics is to study problems by means of economic models. An economic model is an abstraction of the world, which is made to focus on just a few aspects of human behavior and interaction. It is phrased in terms of assumptions, and one can think of it as a tale. We tell a story, and thereby hopefully learn about the big picture. Econometrics is of course concerned with the empirical side to this.

When you read an applied article in the *Quarterly Journal of Economics*, you will often see that people exploit, in one way or another, exogenous variation to learn something about the average causal effect this variation has on some outcome. They then call this a reduced form approach, because they do not estimate what the mechanism is through which something has an effect on something else. But they then move on and interpret their findings, having particular mechanisms or a class of models in mind. For example, when interpreting findings on unemployment duration, they will interpret their findings against the background of job search theory. This means that one attributes the exogenous variation to some primitive parameter in the model that has changed, such as for instance search costs.

Sometimes, there is a heated debate between followers of this approach and followers of the so-called structural approach to econometrics. The structural approach is to spell out an economic model, and to estimate parameters of that model. This quantifies how exogenous changes influence the primitives of the model. This allows one to go further: having the estimated parameters in hand, one can simulate how people would react to a policy change such as a tax reform that has *not* yet taken place. Michael Keane (2010) propagates this approach, in a somewhat provocative article. This appeared in a special issue of the *Journal of Econometrics*, and actually the whole issue is worth reading in this context. Another good read, if one wants to learn more about structural models, is volume 156, issue 1 of the *Journal of Econometrics*, on "Structural Models of Optimization Behavior in Labor, Aging, and Health."

What may seem confusing in this context is that every structural model has so-called "reduced forms", which one gets by solving the model and expressing some of the variables as functions of the other variables. Under the right assumptions, this yields an equation that is linear in parameters, and that can also be estimated. Then, one estimates structural parameters from a reduced form equation. But these are often not

the same reduced forms people have in mind when they read the *Quarterly Journal of Economics*. So, one way to confuse them is to ask what the equation they are looking at is a reduced form of (but many of the authors—probably not all of the readers, though,—will actually have a good answer to this question).

We will go through a list of examples that will make clear what this means. For now, I shall end this section by pointing you to Marschak (1953) who starts his article with the words "[k]nowledge is useful if it helps to make the best decisions." He then explains, by means of examples, which knowledge is useful in which situations, and depending on who uses this knowledge and thereby provides a reconciliatory perspective. In a nutshell, he argues that whether estimates of structural parameters are needed to answer a specific question, or whether "reduced-form evidence" is enough really depends on the context—which seems to make a lot of sense.

2.3 Experiments

A very clean approach to learning about reduced-form economic relationships is to conduct a lab or field experiment. The reason is that then, one can be sure that variation in circumstances under which outcomes came about are truly random. A good place to learn more about field experiments is Vol. 25, Issue 3 of the *Journal of Economic Perspectives*.

In contrast to that, some of the topics in these lecture notes are concerned with learning from observational data, so they are concerned with the challenges one faces if one *cannot* conduct such experiments. Then, variation may not be exogenous and therefore one has to pursue another identification strategy, such as exploiting exogenous variation in instrumental variables analysis or conducting a differences-in-differences analysis. These could in turn again be related to natural experiments, which are experiments nature has conducted for us. For instance, it may be random whether individuals are born just before or just after a threshold date. Good starting points for further reading are the survey paper by Angrist and Krueger (2001) and the textbook by Angrist and Pischke (2009).

In any case, even if we conduct an experiment, the methods discussed here are helpful. For instance, binary choice models are useful to describe behavior of individuals when they chooswe between two options and how they react to changes in circum-

stances. On top of that, structural models can be used in the contexts of experiments to relate differences in choices to deeper, structual parameters. That way, one can simulate how individuals would react to other changes in the environment that they have not been exposed to, and—more importantly even—welfare effects can be quantified. Without a model in hand, this is fundamentally impossible.

2.4 Useful (and Fun) Readings for the Young Economist

I have especially benefited from reading the following books. The first one I would like to mention is Silvia (2007). The author starts with describing a challenge many people face, namely to organize themselves, and get to working effectively on their projects. Then, he talks about many ways in which one could achieve that. I highly recommend this book to everybody—the time to read it is extremely well invested, and the sooner one reads it the better. I believe that even students studying for exams can benefit a lot from it, even though they are not the main target group.

Next, I would like to mention two guides for young economists who want to pursue an academic career. Thomson (2001) and Cawley (2011). Cawley focuses on the academic job market for junior positions after a Ph.D. I think one should actually read it already early on, maybe as early as in the first year as a Ph.D. student.

Then, once it comes to writing papers or a thesis and improving the English Strunk and White (1999) is definitely a classic—a very elegant and beautiful one I would say. There is, by the way, a corresponding one that is concerned with conventions when it comes to Matlab coding, Johnson (2010).

Well, and at some point one gets to publishing papers. This involves peer-review, meaning that you submit the paper, you get it back after half a year or so, along with some comments by reviewers. Then, you revise it, you send it back, wait for another few months, and so on. A recent article containing some critique on current practices is Spiegel (2012). It gives you a little bit of an idea how bumpy this process can be.

Finally, I keep adding Blog posts on this topic at https://kleintob.wordpress.com/category/advice-to-ph-d-students/.

2.5 Useful Tools

There are far too many tools that deserve to be mentioned. Still, I'd like to mention those I find most useful for my daily work. First of all, for writing papers and slides (and also these lecture notes) I very much like an editor called LyX. It is open source and fully compatible with LaTeX, and I am using it despite knowing how to write plain TeX files—it just saves a lot of time and is at the same time just as powerful as TeX, and easier to get into for those of you who have never used TeX. For managing my bibliography, I use JabRef on my Windows PC and BibDesk on the Mac together with Google Scholar, where you can export BibTeX entries.

Once it comes to working on many computers with the same files, I very much like Dropbox. It works in the background and your files are even accessible *via* a web browser. Dropbox is also great for collaborating, as it allows you to share a folder. Another tool I find very useful is Subversion (SVN). It is part of every Unix or Linux distribution (and therefore also comes with MacOS), and there are also clients for Windows PCs (Turtoise SVN). What it does is to keep track in changes of your files, so instead of saving manually several versions of the same file, you always save it in the same file, and store the different versions in the so-called repository. This one can be either on a server (which is useful when several people work together), or on your hard disk, for example in your Dropbox folder. You can think of it as a professional version of Time Machine for the Mac.

Besides, for empirical work and simulations, I use Stata, MATLAB, and the solver KNITRO.

For reading texts I use the iPad, with a combination of Goodreader and Dropbox to manage my files. The format of these lecture notes is optimized for reading on a tablet such as the iPad.

Part I

Identification and Estimation

Chapter 3

Identification

3.1 Introduction

The question of identification is a fundamental one: what can be recovered from observations, and under which conditions? Here, we will approach this topic from a historic perspective, which shows that studying identification often involves an understanding of economic theory.

It all started with economic theory. Microeconomic theory as we know it today can be traced back to Cournot who was the first to draw supply and demand curves in a graph. His book *Researches on the Mathematical Principles of the Theory of Wealth* was published in 1838, about thirty years before Alfred Marshall and Leon Walras refined his contributions to demand theory. Thereafter, at the beginning of the 20th century, scholars started to conduct empirical work in the way we know it today. Christ (1985) provides an interesting historical account on this. He ends with a discussion on identification in the context of supply and demand, which here is (p. 1985) "the question of whether a line (or equation) fitted to data for prices and quantity is a demand curve, or a supply curve, or some unknown mixture of both." Let us start by looking into this in a bit more detail.

3.2 A historic perspective

Also Henry Moore's work on business cycles, in his 1914 book *Economic Cycles: Their Law and Cause*, involved a regression of the quantity of pig iron sold in a market on its price. He called the resulting upward-sloping curve a "new type" of dynamic demand curve. The point in time at which economists started to work on identification was when Philip Wright (1915) subsequently reviewed that book in the *Quarterly Journal of Economics* and argued against Moore's interpretation (p. 638):

> But how about the "new type," the ascending demand curve for pig iron, is it hopelessly irreconcilable with theory? Not at all. The conditions of demand are changed (very probably by improved business conditions) in the direction of a rapid and continuous increase.

He then shows a figure in which the supply curve is being held constant and traced out by a shifting demand curve and argues that this shows that "the need of checking statistical inductions by abstract reasoning is quite as great as that of verifying abstract reasoning by statistics (p. 638)." Put differently, he advocates to *interpret* the data in light of abstract reasoning, or theory, or to look at the data through the lens of theory. In his case, he looks at a regression of quantities on prices, and he asks what the interpretation of the upward sloping line is. It is actually interesting to read Schultz (1925), who describes how demand estimation was done in practice at that time (p. 578f):

> The common method of fitting a straight line to data involves the arbitrary selection of one of the variables as the independent variable X and the assumption that an observed point fails to fall on the line because of an "error" or deviation in the dependent variable Y alone, the X variable being allowed no deviation.

Interestingly, here he acknowledges that measurement error may cause estimates to be inconsistent. And as irritating it is at first glance, it is still true that if both X and Y are measured with error, it is indeed the case that there is no reason to prefer one of the two regressions described above to the other. As a little aside, Wald (1940) cites this paper in the paper that later become known as the paper in which the "Wald IV estimator," an instrumental variables estimator with binary instruments, was introduced.

Coming back to Schultz (1925), he describes an important novel idea in the concluding remarks of his paper (p. 630), namely that

> [t]he "errors" with which we have to deal are not only, or even mainly, the accidental errors which are due to no known cause of systematic or constant error and which play such an important role in the theory of least squares. But they are treated as though they were true accidental errors. That is to say, we first eliminate such constant or systematic "errors" as may be eliminated through the use of index numbers, trend ratios, or link relatives. (Such constant or systematic "errors" are due, as a rule, to population growth and to changes in the general price level.) We are quite certain that there are others still, but we can not measure them. We therefore assume that they are eliminated by the graduation process involved in fitting the demand curve.

This now involved the idea that what is referred to as "errors" comes from somewhere and actually has a meaning over and above error that is due to inaccurate measurement of quantities. His proposal to eliminate or reduce shifts in the demand curve to trace it out using price movements that come from elsewhere is closely related to a proposal by Working (1927) to trace out either the supply or demand curve by holding the other one constant. In this 1927 paper, we actually find similar diagrams as the one in Wright (1915) in which demand and supply curves are drawn. It is there that Working explains that "fluctuations" in the supply curve trace out the demand curve and *vice versa*:

> [I]t is not evident that Professor Moore's "law of demand" for pig iron is in reality a "law of supply" instead? The original observation of prices and corresponding quantities are the resultant of both supply and demand. Consequently, they do not necessarily reflect the influence of demand any more than that of supply. The methods used in constructing demand curves (particularly if the quantity data are of quantity sold) may, under some conditions, yield a demand curve, under others, a supply curve, and, under still different conditions, no satisfactory result may be obtained.

He then continuous by discussing in which markets demand is more likely to fluctuate, and in which markets supply is. An example for the latter is the market for agricultural

commodities where weather conditions shift the supply curve, but demand is largely independent of that. He refers to Wright's work, in footnote 9, saying that (p. 223)

> [H]is analysis bears some resemblance to the above. However, his specific argument is unfortunate in that sense he says "the conditions of demand are changed (very probably by improved business conditions) in the direction of a rapid and continuous increase."...Mr. Wright, to whom the present paper has been submitted, now concurs that the result is due to a shifting back and forth rather than to a continuous shift of the demand curve to the right.

Here, Working realizes that it is independent variation in either of the curves, rather than a continuous increase that traces out one or the other. Actually, either of the two may or may not work in general. Then, he points out more clearly than Schultz (1925) that (p. 224f)

> [w]hether a demand or a supply curve is obtained may also be affected by the nature of the corrections applied to the original data. The corrections may be such as to reduce the effect of the shifting of the demand schedules without reducing the effect of the shifting of the supply schedules...By intelligently applying proper refinements, and making corrections to eliminate separately those factors which cause demand functions to shift and those factors which cause supply functions to shift, it may be possible even to obtain both a demand curve and a supply curve for the same product and from the same original data.

This is again the idea of controlling for factors shifting one of the two curves. Then, variation in the other curve can be used to trace out the one curve that is held constant. Factors shifting one curve, say the supply curve, while the other is held constant are—in modern terminology—instruments for price in a demand equation where quantity is regressed on price.

What is remarkable here is that footnote 9 in Working (1927) documents that Wright knew about this by the time worked on his book *The Tariff on Animal and Vegetable Oils*, which was published in 1928. The first 285 pages of that book contain a detailed discussion on vegetable oils and how they are treated. Then, in Appendix B,

Wright provides two derivations of the instrumental variables estimator that we know today, formalizing these ideas.

There is actually a related debate on the relative size of the contributions of the father, Philip Wright and his son Sewall. Goldberger (1972), in his survey on structural equations methods, gives a lot of credit to the son, stating already in the abstract that he "attempts to regress economists' neglect of the [related] work [on path analysis] of Sewall Wright." Goldberger then goes though Sewall Wrights contributions, noting that it remained largely unnoticed even though they are related to important subsequent developments such as the modern literature on identification, factor analysis and simultaneous equations models. Stock and Trebbi (2003) use stylometric techniques to investigate who wrote Appendix B and conclude that "[t]he stylometric evidence clearly points to Philip G. Wright. Who first thought of using the instrumental variable estimator to solve the identification problem in econometrics? Of this we cannot be sure: perhaps it was collaborative work, or even Sewall's idea which Philip simply wrote up." Maybe Sewall Wright played a role in formalizing ideas that were around for already quite some time and can be found in his father's discussion of Wright (1915). This Appendix B, in any case, concludes what is probably the first episode of research on identification. Remarkably, this first episode involved a deep understanding of economic theory that was in the beginning helpful for interpreting empirical relationships, and later provided guidance for developing estimators.

This already shows that studying identification is central to conducting empirical work. Nowadays, identification is probably more important than ever, as a paper that does not clearly explain why we can learn about the quantity of interest from data will most likely be considered incomplete.

3.3 Well-Known Identification Problems

Before moving on, let me briefly point you to some well known identification problems. The first one I would like to mention is the so-called "endogeneity problem." Here, in a nutshell, one is interested in the coefficient on a right hand side variable x_i, typically in a linear model, and that variable is correlated with the error term. This could be because one variable has been omitted that is related to both that variable and the error term, and therefore one faces what is called an omitted variables problem. One solution

to this is to use panel data methods (see Section 3.6.3) or exploit properties of so-called instrumental variables (see Section 4.4.2).

In a dynamic context, an important problem is to distinguish state dependence from heterogeneity (Bates and Neyman, 1952; Heckman, 1981). For example, we observe that unemployed individuals are the less likely to find a job the longer they have been unemployed (Heckman and Borjas, 1980). The question now is whether staying unemployed makes one less likely to find a job in the next period, for example because human capital depreciates or because employers take this as a bad signal—which would be state dependence—, or whether there is individual heterogeneity in the likelihood of finding a new job so that knowing that somebody has been unemployed for some time makes one conclude that this person actually is of a type that has a lower likelihood of finding a new job in any period. In the second case, we would still find the empirical relationship that individuals who are unemployed for a longer time will be less likely to find a job because of dynamic selection: over time, more and more individuals are left who are of the type with a lower likelihood to find a job. This example is in the context of duration models (see Section 6.4) but could also arise in the context of linear panel data models (see Section 3.6.3) when a past value of the left hand side variable enters as a regressor on the right hand side. If the coefficient on that regressor is non-zero, then there is state dependence.

A somewhat similar problem in a static context is the reflection problem described by Manski (1995, p.1ff). Suppose we observe that individuals belonging to the same group behave similarly. This could be for two reasons, either because the propensity of an individual to behave in some way varies with the prevalence of that behavior in the group (this is the so-called endogenous effects), or because individuals in the same group tend to behave similarly because they face similar institutional environments or have similar individual characteristics (this is called correlated effects). Without prior knowledge we cannot distinguish between the two explanations. This becomes a relevant problem if different explanations have different implications for public policy: understanding how students interact in classrooms is critical to the evaluation of many aspects of educational policy. Finding out which one of the two mechanisms is at play is the question of identification.

3.4 General Idea

We first need to start by saying what we want to estimate. This could be a *structural parameter* such as the parameter vector β in a standard linear regression equation

$$y_i = x_i\beta + \varepsilon_i \tag{3.4.1}$$

or the mean of a vector of random coefficients, β_i, in an otherwise similar model

$$y_i = x_i\beta_i + \varepsilon_i.$$

If such an equation represents a causal link rather than a mere empirical association we could call it a *structural equation* (Goldberger, 1972). This structural equation may, or may not, be related to a structural model (whether it is or not is not important here). Throughout we call y_i the *dependent variable* and x_i the vector of *independent variables*. Sometimes we will refer to the former as the outcome variable or the left hand side variable and the latter as covariates or right hand side variables.[1]

In general, different economic data generating processes, or *structures*, can generate the same distribution of y_i given x_i. In that case we call them *observationally equivalent*. No amount of data can distinguish between observationally equivalent structures. For example, the structures $y_i = x_i\beta + \varepsilon_i$ and $y_i = x_i\beta - \varepsilon_i$ are observationally equivalent if the distribution of ε_i conditional on x_i is symmetric about zero, because they both generate the same distribution of y_i given x_i.

An econometric model such as (3.4.1) with accompanying assumptions is a set of restrictions that define the set of structures which are admitted by the model. At this point it is important to note that distinguishing between a model and a structure is sometimes a bit artificial. Below it will become clear, however, what is meant. The model that defines the set of *admissible structures* involves, for example, functional form restrictions such as equations being linear in parameters, assumptions on the smoothness

[1] Occasionally, the left hand side variable is called the endogenous variable. However, in other contexts it is said that there is an endogenous variable on the right hand side as well. Therefore, we will try to avoid this appellation. Still, we will sometimes refer to some of the right hand side variables as being exogenous. It will be made clear what this means in the specific context. Often, it means that they are uncorrelated with, or independent of, the error term. See for example Engle et al. (1983) for details.

of functions, exclusion restrictions, other stochastic restrictions, and monotonicity assumptions. Here, we usually aim at making as few assumptions as possible, in order to get down to the question *what* drives identification. However, when it comes to estimation it might be preferable to make additional assumptions to improve on the precision of the estimates. These could be distributional assumptions or additional functional form restrictions.

If an econometric model rules out all but one structure that could generate the observed distribution of the outcome variable y_i conditional on the covariates x_i the model identifies the structure. If only one particular feature, a structural parameter, of the data generating process is of interest, this requirement can be weakened: if the structural parameter is the same for all admissible structures which could generate the distribution of the outcome given covariates, then we say that the structural parameter is identified. In Section 3.6.2 below I show that structural parameter β in (3.4.1) is identified once we assume that $\mathbb{E}[\varepsilon_i|x_i] = 0$.

In the remainder of this chapter, we formally discuss these ideas. They have already been laid out in the 1940s when Koopmans (1949, p. 132) called for a "separation between problems of statistical inference...and problems of identification." He describes identification (p. 125) as "inference from that distribution [the joint distribution of observations] to the parameters of the structural equations describing economic behavior". So, put differently, the question of identification is *why* we can implement a particular estimator in order to recover a relationship that is of interest.

3.5 Definitions

We are now ready to define what a structure is. For this, we follow the classic papers by Koopmans (1949), Hurwicz (1950), and Koopmans and Reiersøl (1950).[2]

Definition 1. A *structure* S consists of (i) a system of equations delivering a value of a vector outcome, Y, given a value of a vector covariate, X, and a value of a vector

[2] I use the usual notation in this literature. Uppercase letters denote random variables. You can think of Y as corresponding to y_i, that is used in most textbooks covering classical econometric theory, and X as corresponding to x_i. Throughout, we can think of the observations as being random draws from the same distribution.

of unobservable random variables, ε, (ii) a conditional distribution function, $F_{\varepsilon|X}$ for the unobservables given the covariates such that the conditional distribution function of outcomes, $F_{Y|X}$, is well defined.

This definition allows the outcome to be a vector rather than a scalar. Importantly, this definition is satisfied for any model that is studied in these lecture notes. Identification problems arise because different structures may be observationally equivalent.

Definition 2. Two structures S and S' are said to be *observationally equivalent (indistinguishable)* if, conditional on exogenous variables, the two distributions of outcome variables generated by the structures S and S' are identical for all possible values of the exogenous variables.

Conversely, if there are no observationally equivalent admissible structures to the structure that has generated the data then it is identified.

Definition 3. The structure S is *identified* by a model if there is no observationally equivalent structure S' within the set of admissible structures.

Trivially, since S is the structure that we seek to identify the observed distribution of Y given X is the distribution that is generated by S. All observationally equivalent structures generate this distribution. The set of observationally equivalent admissible structures is a subset. If it contains just one structure, S, then S is identified.

Definition 4. A *structural parameter* $\theta(S)$ *of a structure* S *is identified* by a model if it is the same for all structures S' that are admitted by the model and are observationally equivalent to S. It is *uniformly identified* by a model if it is identified for every structure S admitted by the model.

Again, S is the structure that has generated the data. If the structural parameter is the same for all observationally equivalent structures, then it is identified. This is a statement for a particular S. If it holds for all S that are admitted by the model then we say that the structural parameter is uniformly identified over S. This is the typical case in the models we will look at.

Chesher (2006) proves the following useful lemma which says that once we can find a functional (a function of a function) which links a structural characteristic to the conditional distribution of Y given X, then the structural characteristic is identified.

Lemma 1. *Consider a model, let S^a be the set of structures admitted by the model such that $\theta(S) = a$ and let A be the set of all values of $\theta(S)$ generated by admissible structures. Let $F^S_{Y|X}$ denote the conditional distribution function generated by a structure S. Suppose there exists a functional of the conditional distribution function of Y given X, $\mathscr{G}(F_{Y|X})$, such that for each $a \in A$, $\mathscr{G}(F^S_{Y|X}) = a$ for all $S \in S^a$. Then $\theta(S)$ is uniformly identified by the model.*

Proof. Consider any value $a_0 \in A$ and any structure S_0 with $\theta(S_0) = a_0$ and let S_0^* be the set of structures observationally equivalent to S_0. Consider any $S' \in S_0^*$ and let $\theta(S') = a'$. If a functional $\mathscr{G}$ with the stated property exists then $\mathscr{G}(F^{S'}_{Y|X}) = a'$ and $\mathscr{G}(F^{S_0}_{Y|X}) = a_0$. Since S' and S_0 are observationally equivalent $F^{S'}_{Y|X} = F^{S_0}_{Y|X}$ and therefore $a' = a_0$. Therefore, if a functional $\mathscr{G}$ with the stated property exists then, for any $a_0 \in A$, all structures observationally equivalent to any structure S_0 with $\theta(S_0) = a_0$ have the same value, a_0, of the structural characteristic, and so $\theta(S)$ is uniformly identified by the model. □

In the following examples we will see that the lemma provides a constructive way to prove identification.

3.6 Examples

3.6.1 Bivariate Linear Model

We observe y_i and a scalar x_i. In addition, there is an unobservable error term. Suppose that we are interested in the average effect x_i has on y_i. The job of an econometric model is to impose enough restrictions so that this effect is identified. In applied work it is key to justify why these restrictions are plausible. In this example, the model consists of the functional form restriction

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and the assumption that

$$cov(\varepsilon_i, x_i) = 0.$$

This defines the set of admissible structures. We get one admissible structure for every pair of values b_1 and b_2 of β_1 and β_2, respectively.

Our parameter of interest is β_2. It is identified if there is only one value of it that produces the observed distribution of y_i given x_i. Following the logic of Lemma 1, once we can find a functional that gives us this unknown parameter as a function of the distribution of y_i given x_i then it is identified. Observe[3]

$$cov(y_i, x_i) = cov(\beta_1 + \beta_2 x_i + \varepsilon_i, x_i) = \beta_2 var(x_i) + cov(\varepsilon_i, x_i).$$

By assumption, $cov(\varepsilon_i, x_i) = 0$ so that

$$\beta_2 = \frac{cov(y_i, x_i)}{var(x_i)}.$$

This establishes identification.

More formally, along the lines of Lemma 1, consider all structures satisfying the above mentioned restrictions for which $\beta_2 = b_2$, a particular numerical value. We have just shown that b_2 is a known functional of the distribution of y_i given x_i, that is

$$b_2 = \theta(S) = \frac{cov(y_i, x_i)}{var(x_i)}.$$

As this is the case for all b_2 and all structures with $\beta_2 = b_2$ we can conclude that β_2 is uniformly identified.

Notice that there could be another observationally equivalent structure S' with $b_2 = \theta(S) = \theta(S')$ but with a different value of β_1. This shows that the structure S itself is not identified because for this we would require β_1 to be identified. This is not the case because we have not made any assumptions on $\mathbb{E}[\varepsilon_i]$. Finally, notice that the restriction $cov(\varepsilon_i, x_i) = 0$ is very natural in the context of experiments in which x_i is randomly assigned to individuals, for example the dose of some medication, since it assumes that all randomness in the outcome is uncorrelated with x_i. Conversely, it is actually very unnatural if x_i is a choice made by i that is probably made in light of at least parial knowledge of ε_i. In the latter case, we face an endogeneity problem.

[3]The second equality follows from standard rules for covariances, see Section C.5. You can show this by writing down the definition of the covariance in this case.

Finally, observe that in this example, we have required that the covariance between x_i and ε_i is zero or, more generally, takes on a particular known value. For example, we could just as well have assumed that $cov(\varepsilon_i, x_i) = 3$, for example Then, we would have had

$$\beta_2 = \frac{cov(y_i, x_i) - 3}{var(x_i)}.$$

Alternatively, we could have assumed that the error term is mean independent of x_i, that is $\mathbb{E}[\varepsilon_i | x_i] = e$, where e is any fixed value. This implies that the covariance between x_i and ε_i is zero. To show this, we use the law of iterated expectations, in particular $\mathbb{E}[\varepsilon_i] = \mathbb{E}[\mathbb{E}[\varepsilon_i | x_i]]$. Using this,

$$\begin{aligned}
cov(x_i, \varepsilon_i) &= \mathbb{E}[(x_i - \mathbb{E}[x_i])(\varepsilon_i - \mathbb{E}[\varepsilon_i])] \\
&= \mathbb{E}[x_i \varepsilon_i - x_i \mathbb{E}[\varepsilon_i]] \\
&= \mathbb{E}\left[\mathbb{E}\left[x_i \varepsilon_i - x_i \mathbb{E}[\varepsilon_i] \middle| x_i\right]\right] \\
&= \mathbb{E}\left[x_i \mathbb{E}\left[\varepsilon_i \middle| x_i\right] - x_i \mathbb{E}\left[\mathbb{E}[\varepsilon_i] \middle| x_i\right]\right] \\
&= \mathbb{E}[x_i](e - e) \\
&= 0,
\end{aligned}$$

where the first equality follows from the definition of the covariance, the second equality follows from $\mathbb{E}[x_i]$ being a scalar which can be taken out of the expectation (alternatively we could have done this with $\mathbb{E}[\varepsilon_i]$), the third equality applies the law of iterated expectations, the fourth uses that the (conditional) expectation of a sum is equal to the sum of the (conditional) expectations, and the fifth uses the mean independence assumption and its implication $\mathbb{E}[\varepsilon_i] = e$.

3.6.2 Multivariate Linear Model

Our second example is taken from Chesher (2006, p.8f). Let

$$y_i = x_i \beta + \varepsilon_i,$$

where now x_i is a vector

$$x_i = (1, x_{i2}, \ldots, x_{iK}),$$

β is a vector, and y_i and ε_i are scalars. Assume that $\mathbb{E}[\varepsilon_i|x_i] = 0$. This implies that $\mathbb{E}[y_i|x_i] = x_i\beta$. Assume that there are K values of x_i, $x_1,\ldots,x_K$, in the data such that the $K \times K$ matrix

$$\tilde{x} \equiv \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \ldots & x_{1K} \\ 1 & x_{22} & \ldots & x_{2K} \\ & \vdots & & \\ 1 & x_{K2} & \ldots & x_{KK} \end{pmatrix}$$

has full rank. This implies that $\tilde{x}$ is invertible. Define

$$\mu_y(\tilde{x}) \equiv \begin{pmatrix} \mathbb{E}[y_i|x_i = x_1] \\ \vdots \\ \mathbb{E}[y_i|x_i = x_K] \end{pmatrix} = \begin{pmatrix} x_1\beta \\ \vdots \\ x_K\beta \end{pmatrix} = \tilde{x}\beta.$$

This is a system of K equations with K unknowns as the vector $\mu_y(\tilde{x})$ is observed. Then, since $\tilde{x}$ has been assumed to have full rank we have that

$$\beta = \tilde{x}^{-1}\mu_Y(\tilde{x}).$$

Again, as in Section 3.6.1, consider all structures satisfying the conditions of the econometric model in which $\beta = b$, a particular numerical value. Observe that b is a functional of the conditional distribution of y_i given x_i since

$$b = \tilde{x}^{-1}\mu_y(\tilde{x}).$$

As this is true for all b it follows from Lemma 1 that β is uniformly identified. Notably, just as in Example 3.6.1 we can use exogenous variation in x_i to identify β. As an aside notice that for the bivariate case we have $x_i = (1, x_{i2})$ so that for any two values x_1 and x_2 for which we have variation in x_{i2}, that is $x_{12} \neq x_{22}$, we have

$$\tilde{x}^{-1} = \begin{pmatrix} 1 & x_{12} \\ 1 & x_{22} \end{pmatrix}^{-1} = \frac{1}{x_{22} - x_{12}} \begin{pmatrix} x_{22} & -x_{12} \\ -1 & 1 \end{pmatrix}.$$

Hence,

$$\begin{aligned}\beta &= \tilde{x}^{-1}\mu_Y(\tilde{x}) \\ &= \frac{1}{x_{22}-x_{12}}\begin{pmatrix} x_{22} & -x_{12} \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \mathbb{E}[y_i|x_i=x_1] \\ \mathbb{E}[y_i|x_i=x_2] \end{pmatrix} \\ &= \frac{1}{x_{22}-x_{12}}\begin{pmatrix} x_{22}\mathbb{E}[y_i|x_i=x_1]-x_{12}\mathbb{E}[y_i|x_i=x_2] \\ -\mathbb{E}[y_i|x_i=x_1]+\mathbb{E}[y_i|x_i=x_2] \end{pmatrix}.\end{aligned}$$

Using $\mathbb{E}[y_i|x_i=x_1]=\beta_1+\beta_2 x_{12}$ and $\mathbb{E}[y_i|x_i=x_2]=\beta_1+\beta_2 x_{22}$ we see that the first element of β is equal to

$$\begin{aligned}\frac{x_{22}\mathbb{E}[y_i|x_i=x_1]-x_{12}\mathbb{E}[y_i|x_i=x_2]}{x_{22}-x_{12}} &= \frac{x_{22}(\beta_1+\beta_2 x_{12})-x_{12}(\beta_1+\beta_2 x_{22})}{x_{22}-x_{12}} \\ &= \frac{\beta_1(x_{22}-x_{12})}{x_{22}-x_{12}}+\frac{\beta_2(x_{12}x_{22}-x_{12}x_{22})}{x_{22}-x_{12}} \\ &= \beta_1+0 \\ &= \beta_1,\end{aligned}$$

the intercept. The second element is equal to

$$\frac{\mathbb{E}[y_i|x_i=x_2]-\mathbb{E}[y_i|x_i=x_1]}{x_{22}-x_{12}},$$

the difference in the expected values of y_i for different values of x_i over the difference in the second component of x_i. Using $\mathbb{E}[y_i|x_i=x_1]=\beta_1+\beta_2 x_{12}$ and $\mathbb{E}[y_i|x_i=x_2]=\beta_1+\beta_2 x_{22}$ we get that this is equal to

$$\frac{\beta_1+\beta_2 x_{22}-\beta_1-\beta_2 x_{12}}{x_{22}-x_{12}}=\beta_2.$$

Identification in this example is shown using variation of the conditional expectation of y_i given different values of x_i. Before, in Section 3.6.1, we have instead used a summary statistic for the co-variation between y_i and x_i, the covariance. This covariance measures the same underlying dependence and it is therefore worth noting that both approaches are very similar in nature. Clearly, once it comes to estimation, the

approach discussed above is inferior because an estimation procedure that is based on conditional averages could only make use of as many values of x_i as there are variables.

To complete the picture we finish with two observations. First, a multivariate version of the result in Section 3.6.1 can be obtained from

$$y_i = x_i\beta + \varepsilon_i,$$

by pre-multiplying both sides of it with the $K \times 1$-vector x_i' and taking expectations. If the covariance between all components of x_i and ε_i is zero we have $\mathbb{E}[x_i'\varepsilon_i] = 0$ and thus obtain

$$\beta = \mathbb{E}[x_i'x_i]^{-1}\mathbb{E}[x_i'y_i] \tag{3.6.1}$$

if $\mathbb{E}[x_i'x_i]$ is invertible, that is if there is enough variation in x_i. There is again a measure of the co-movement between x_i and y_i, namely $\mathbb{E}[x_i'y_i]$. Moreover, the variation in x_i is summarized by $\mathbb{E}[x_i'x_i]$.

The second and final observation is that we can show that all results still hold if β is in fact a random coefficient β_i with $\mathbb{E}[\beta_i|x_i] = \beta$. This is because in this case

$$y_i = x_i\beta_i + \varepsilon_i \tag{3.6.2}$$

which we can rewrite as

$$y_i = x_i\beta + \nu_i$$

where $\nu_i \equiv \varepsilon_i + x_i(\beta_i - \beta)$. If $\mathbb{E}[\varepsilon_i|x_i] = 0$ it follows that

$$\begin{aligned}
\mathbb{E}[\nu_i|x_i] &= \mathbb{E}[\varepsilon_i + x_i(\beta_i - \beta)|x_i] \\
&= \mathbb{E}[\varepsilon_i|x_i] + \mathbb{E}[x_i(\beta_i - \beta)|x_i] \\
&= 0 + x_i\mathbb{E}[(\beta_i - \beta)|x_i] \\
&= 0 + x_i(\mathbb{E}[\beta_i|x_i] - \beta) \\
&= 0 + x_i \cdot 0,
\end{aligned}$$

where the first equality follows from the definition of ν_i, for the second equality we use that the conditional expectation of a sum is the sum of conditional expectations (linearity of conditional expectations), the third equality follows from the assumption

that $\mathbb{E}[\varepsilon_i|x_i] = 0$ and that x_i can be taken out of an expectation conditional on x_i, the fourth equality follows again from the linearity of conditional expectations, and the last equality follows from the assumption that $\mathbb{E}[\beta_i|x_i] = \beta$. Notice that this shows us that a linear regression of y_i on x_i estimates the population average random effect, $\mathbb{E}[\beta_i]$, if this random effect is independent of x_i. This is because the randomness of β_i, $\beta_i - \beta$, enters a new error term v_i and we just have shown that v_i is mean independent of x_i under the assumption that $\beta_i - \beta$ is mean independent of x_i.[4]

3.6.3 Multivariate Linear Model for Panel Data

Now suppose we repeatedly observe individuals so that we have so-called panel, or longitudinal, data. Our model is now

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it},$$

where $i = 1,\ldots,N$ indexes individuals and $t = 1,\ldots,T$ indexes the time period in which these individuals are observed.

One possibility is to think of $u_{it} \equiv \alpha_i + \varepsilon_{it}$ as an error term that consists of a time-invariant component α_i and a time-varying component ε_{it}. Assume that the unobservables are uncorrelated across individuals, that α_i and ε_{it} are uncorrelated, that ε_{it} is uncorrelated across time, and that α_i and ε_{it} are both mean zero (which is innocuous if x_{it} includes a constant).

In principle, we can pool all observations across individuals and time and assume that the composite error term u_{it} is not correlated with the explanatory variables x_{it}. Then, we can show identification in exactly the same way as in Section 3.6.2. But this estimator is inefficient because it ignores the correlation structure, within individuals, between the error terms. An efficient version of this estimator that takes the correlation structure of the error terms into account is the random effects estimator that we present in Section 4.2.2. It requires the stronger assumption that x_{is} is uncorrelated with u_{it} for all combinations of s and t. This is called strict exogeneity (Engle et al., 1983) and is stronger than the uncorrelatedness assumption made in Section (3.6.3). But it is worth noting that we make it only out of efficiency considerations, not for identification.

[4]For identification it is actually sufficient to show that the covariance between v and x_i is zero. For this the mean independence assumptions can be replaced by covariance restrictions.

One could also make the assumption that $\mathbb{E}[\varepsilon_{it}|x_i] = 0$ and exploit the fact that α_i does not vary across time periods (include the contribution of all components of x_{it} that do not vary over time into α_i). For this, one does not need to assume that α_i and x_{it} are unrelated, which is a major advantage. But we need to assume, like for the random effects estimator, that ε_{it} is mean indpeendent of all components of x_i, which means that it is also mean independent of x_{is} in some other time period.

To show identification denote variables from which individual-specific averages have been subtracted with a tilde, and observe that $\tilde{\alpha}_i = 0$. Then,

$$\tilde{y}_{it} = \tilde{x}_{it}\beta + \tilde{\varepsilon}_{it}.$$

If $\mathbb{E}[\tilde{\varepsilon}_{it}|\tilde{x}_{it}] = 0$, then we are back in the very same case as before in Section 3.6.2 so that β is identified. A sufficient condition for this requirement to be fulfilled is that strict exogeneity holds, which is why we have made this assumption. We need strict exogeneity here because $\tilde{x}_{it}$ and $\tilde{\varepsilon}_{it}$ are functions of values of x_{it} and ε_{it}, respectively, in all time periods, as we subtract the individual averages in both cases.

Subtracting individual-specific means is called the within-transformation and the corresponding estimator is the within-estimator, which is sometimes also referred to as the fixed effects estimator. Instead, one can also use first differencing, where the value of a variable in $t-1$ is subtracted from the value in t so that we get for example $\Delta y_{it} = y_{it} - y_{it-1}$ for all but the first t. Then, one can proceed as before. The so-called differences-in-differences estimator in Section 7 is a special case of this, where we have just two time periods and compare the change in the outcome for one group that received some treatment to the change in the outcome for another group that did not receive that treatment. The underlying assumption is that in the absence of the treatment the time trend for those who have been treated would have been the same as the one for those who have not been treated.

3.7 Predictions and Conditional Expectation Functions

We have just seen examples in which structural parameters are identified. I would like to point out, at this point, that certain predictions can even be made if key parameters of the model are *not* identified.

Suppose once more that $y_i = x_i\beta + \varepsilon_i$ and $\mathbb{E}[\varepsilon_i] = 0$, but not necessarily $\mathbb{E}[\varepsilon_i|x_i] = 0$ or $cov(\varepsilon_i, x_i) = 0$. Suppose that β is not identified. Still, we know the conditional expectation function $\mathbb{E}[y_i|x_i]$ because $F_{y_i|x_i}$ is known (this is the usual thought experiment in identification analysis). Therefore, we can predict y_i given the actual x_i in our data. However, a *ceteris paribus* prediction in which we predict y_i for values $\hat{x}_i$ other than x_i is in general not feasible because for this we would have to know β and $\mathbb{E}[\varepsilon_i|\hat{x}_i]$.

There is a difference between the two because individuals that are characterized by a particular value of x_i are fundamentally different from individuals which are characterized by another value of x_i, for example $\hat{x}_i$, in the sense that $\mathbb{E}[\varepsilon_i|x_i] \neq \mathbb{E}[\varepsilon_i|\hat{x}_i]$.

The familiar condition $\mathbb{E}[\varepsilon_i|x_i] = \mathbb{E}[\varepsilon_i|\hat{x}_i]$ means that this is not the case. To see this, suppose that $y_i = x_i\beta + \varepsilon_i$ and that $\mathbb{E}[\varepsilon_i|x_i] = 0$. The latter condition means that individuals are actually the same in terms of $\mathbb{E}[\varepsilon_i|x_i]$. Then, β is identified, as we have shown in Section 3.6.2 above and we can now predict

$$\mathbb{E}[y_i|\hat{x}_i] = \hat{x}_i\beta$$

for *any* $\hat{x}_i$. If $\hat{x}_i \neq x_i$ we predict the so-called *counterfactual outcome* that we would have observed had we exogenously set x_i to $\hat{x}_i$. Consequently, *ceteris paribus* effects of changes in x_i are identified.

Juxtaposing these two cases shows that sufficient knowledge of structural parameters, β in this case, may enable us to predict counterfactual outcomes. This is key for making policy recommendations where we often aim at predicting changes in outcomes (y_i) for individuals or firms that are induced to change their behavior (x_i) by some change in policy.

3.8 Normalizations

In many models that are discussed in the remainder of these lecture notes normalizations are made. It is important to understand how they differ from assumptions. I would like to discuss this by means of an example.

Consider again the bivariate model $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. Suppose that $cov(\varepsilon_i, x_i) = 0$ and $\mathbb{E}[\varepsilon_i] = e$ and that $var(x_i) > 0$. We have shown in Section 3.6.1 that under these conditions β_2 is identified. As $\mathbb{E}[y_i|x_i] = \beta_1 + \beta_2 x_i + e$ we have that

$$\beta_1 + e = \mathbb{E}[y_i|x_i] - \beta_2 x_i.$$

The right hand side of this equation is known. This shows that we can at most identify $\beta_1 + e$, not β_1. Therefore, we need to fix e in order to identify β_1. This is called a normalization and usually we normalize e to be zero. The difference between an assumption and a normalization is that the latter does not impose any restrictions on the distribution of y_i conditional on *any* value of x_i. This is because for any value of e (say 1 instead of 0) that we pick β_1 changes accordingly (here it would decrease by 1). The emphasized "any" is important here, because counterfactual predictions still need to be correct. This is not the case, for example, if we wrongly "normalize" the covariance between ε_i and x_i to be 3, say. Then, counterfactual predictions will be wrong.

3.9 Further Readings

The literature on identification is growing at a fast pace since the early 2000's. For additional references and an overview over the study of identification in the first half of the 2000's see also Chesher (2007). Among others he cites early contributions such as Koopmans and Reiersøl (1950), Hurwicz (1950), Koopmans et al. (1950), Wald (1950), and Fisher (1966). See also Hsiao (1983) for a more classic treatment of identification.

Identification analysis is related to the concept of causality. A philosophical perspective on causality is provided in Holland (1986). The first two chapters in Wooldridge (2002) contain a classic discussion of causality in econometrics. Heckman (2008) reviews the approach to causal modeling in econometrics. Also Angrist and Pischke (2009) put a lot of emphasis on recovering causality in their guide for the applied researcher.

Manski (1995) discusses many relevant identification problems including the selection problem we discuss in Section 7.1.1, the mixing problem, response based sampling, simultaneity, and the reflection problem. In this chapter, we have focussed on point identification, which means that we look for a model such that the set of permissible structures has only one element (likewise for the identification of structural parameters). Manski (2003) is on partial identification where bounds on structural parameters are derived for situations in which they are not identified in the sense of the definitions given above, or more generally the set of permissible structures is described for a given set of assumptions. Manski argues that this approach is often preferable,

as the assumptions that are made to obtain point estimates are often incredible. In the same vein, Manski (2010) criticizes that empirical studies often leave the unjustified impression of certainty, as they are based on strong assumptions and the validity of those assumptions are taken for granted. Major contributions to this literature have been made in the meantime, but the applied branch of it is still in its infancy.

In the meantime, from an applied perspective, a possible response to the challenge put up by Manski is to conduct sensitivity analyses, as already propagated by Leamer (1983) in a well-known article. But generally, this is a fundamental problem without a simple solution.

3.10 Exercises

1. Let x_i and y_i be random variables and a, b, c, and d be scalars. Show that $cov(a+bx_i, c+dy_i) = bdcov(x_i, y_i)$.

2. Assume the multivariate linear model, $y_i = x_i\beta + \varepsilon_i$, and that there is enough variation in x_i so that $\mathbb{E}[x_i'x_i]$ is invertible. Moreover, assume that $\mathbb{E}[\varepsilon_i|x_i] = 0$. Derive the expression given in equation (3.6.1) by minimizing

$$\mathbb{E}[(y_i - x_i\beta)^2]$$

over β.

3. An estimator for the multivariate linear model that is based on the identification result in equation (3.6.1) is the well-known ordinary least squares estimator

$$\hat{\beta} = (X'X)^{-1}X'y,$$

where X is a $N \times K$ matrix of explanatory variables (each row contains one x_i) and y is a vector of N outcome variables so that $y = X\beta$. We have a data set with observations of the wage rate y_i for individual i, a variable m_i which is an indicator variable that takes on the value 1 if i is male (and 0 otherwise), and a variable f_i is an indicator variable that takes on the value 1 if i is female (0 otherwise). There are N_M male and N_F female individuals. For this exercise it is

useful to think of the data as being sorted by gender so that individual 1 through N_M is male and individual $N_M + 1$ through N is female. Then, X is given by

$$\underset{N\times 2}{X} = \begin{pmatrix} \mathbf{1}_M & \mathbf{0}_M \\ \mathbf{0}_F & \mathbf{1}_F \end{pmatrix},$$

where $\mathbf{1}_M$ and $\mathbf{1}_F$ are ones vectors of length N_M and N_F, respectively, and $\mathbf{0}_M$ and $\mathbf{0}_F$ are the corresponding null vectors.

(a) Write down what the estimator of β_1 and β_2 for the linear model

$$y_i = \beta_1 m_i + \beta_2 f_i + \varepsilon_i,$$

are.

(b) Let $\hat{\sigma}^2$ be an estimate of the variance of ε_i. Calculate $var(\hat{\beta}) = (X'X)^{-1}\hat{\sigma}^2$.

(c) Is it an assumption or a normalization that an intercept term is not part of the model? Why?

4. Consider the bivariate model $y_i = \beta_{1i} + \beta_{2i}x_i + \varepsilon_i$ and assume there is variation in x_i. β_{1i} and β_{2i} are random coefficients with $\mathbb{E}[\beta_{1i}|x_i] = \beta_1$ and $\mathbb{E}[\beta_{2i}|x_i] = \beta_2$. Assume $\mathbb{E}[\varepsilon_i|x_i] = 0$.

(a) Show that $\beta_2 = cov(x_i, y_i)/var(x_i)$.

(b) Find an expression for β_1.

5. Consider the multivariate model $y_i = x_i\beta_i$, where β_i is a vector of random coefficients with $\mathbb{E}[\beta_i|x_i] = \beta$. $\mathbb{E}[x_i'x_i]$ is of full rank. Show that $\beta = \mathbb{E}[x_i'x_i]^{-1}\mathbb{E}[x_i'y_i]$.

Chapter 4

Estimation

4.1 Ordinary Least Squares

Write

$$y \equiv (y_1, \ldots, y_N)',$$

$$\varepsilon \equiv (\varepsilon_1, \ldots, \varepsilon_N)',$$

and

$$X \equiv (x_1', \ldots, x_N')'.$$

Then, the ordinary least squares (OLS) estimator in matrix notation can be written as

$$\hat{\beta} = (X'X)^{-1}X'y$$

provided that $X'X$ is invertible. Substituting in the model in (4.4.5) yields

$$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon = \beta + (X'X)^{-1}X'\varepsilon.$$

We say that the OLS estimator is consistent if $\hat{\beta} \xrightarrow{p} \beta$. This is the case if $(X'X)^{-1}X'\varepsilon \xrightarrow{p} 0$ which is usually ensured by the assumption that the error term and the regressors are uncorrelated. In Chapter 3 we have shown that uncorrelatedness is implied by

$$\mathbb{E}[\varepsilon_i | x_i] = 0.$$

For this we have considered

$$cov(x_i, \varepsilon_i) = \mathbb{E}[x_i \varepsilon_i] - \mathbb{E}[x_i]\mathbb{E}[\varepsilon_i]$$

and used iterated expectations to write

$$\mathbb{E}[x_i \varepsilon_i] = \mathbb{E}[x_i \mathbb{E}[\varepsilon_i | x_i]]$$

and

$$\mathbb{E}[\varepsilon_i] = \mathbb{E}[\mathbb{E}[\varepsilon_i | x_i]]$$

which are both equal to zero if $\mathbb{E}[\varepsilon_i | x_i] = 0$. Moreover, we have shown that the converse does not hold because mean independence does not impose any restrictions on $\mathbb{E}[\varepsilon_i]$.

4.2 Generalized Least Squares

4.2.1 General Idea

The generalized least squares (GLS) estimator is an estimator for linear models and is motivated by the idea that estimators such as the OLS estimator are often not efficient because the Gauß-Markov condition that the error terms are i.i.d. across observations does not hold. This could be because our data are panel data and for the same individual the error term is correlated across time periods, or because the variance of the error term varies across observations. If one nevertheless manages to transform the data such that they are again i.i.d., then the resulting estimator will again be efficient, by the Gauß-Markov theorem.

Suppose for now that observations are indexed by $i = 1, \ldots, N$ (in the panel data context this will be *it*) and let the model be

$$y_i = x_i \beta + \varepsilon_i. \tag{4.2.1}$$

Assume that we know the variance-covariance matrix of the error terms, Ω. For the following argument, it is essential that the inverse of this matrix exists, which we henceforth assume.

For the $N \times N$ matrix Ω we can always find corresponding Cholesky factors P that are defined by $\Omega^{-1} = P'P$. Then, $\Omega = (P'P)^{-1} = P^{-1}(P')^{-1}$ and[1]

$$P\Omega P' = PP^{-1}(P')^{-1}P' = I_N.$$

Using this, we have that (4.2.1) is equivalent to

$$Py_i = Px_i\beta + P\varepsilon_i,$$

which means that we can estimate β by performing a linear regression of Py_i on Px_i. These estimates will be efficient because $P\varepsilon_i$ is i.i.d., as

$$var(P\varepsilon_i) = P\Omega P' = I_N.$$

In practice, Ω is typically unknown and the estimator that uses an estimate of Ω is called feasible GLS (FGLS) estimator. Newey and McFadden (1994, p. 2180) show that "under certain regularity conditions, the first-step estimator affects second-step standard errors…if and only if inconsistency in the first step leads to inconsistency in the second step."

4.2.2 Random Effects Estimator for Panel Data as a Special Case

A prominent example of a GLS estimator is the random effects estimator for panel data.

Consider again the model of Section (3.6.3),

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it},$$

where $i = 1,\ldots,N$ indexes individuals and $t = 1,\ldots,T$ indexes the time period in which these individuals are observed. Define $u_{it} \equiv \alpha_i + \varepsilon_{it}$ and assume that the unobservables are uncorrelated across individuals, that α_i and ε_{it} are uncorrelated, that ε_{it} is uncorrelated across time, that α_i and ε_{it} are both mean zero (which is innocuous if x_{it} includes a constant), and that the variances of α_i and ε_{it} are constant and given by σ_α^2 and σ_ε^2, respectively.

[1] Recall $(AB)^{-1} = B^{-1}A^{-1}$.

Write y_i and u_i for the T-vector of dependent variables and error terms, respectively, x_i for the $T \times K$ matrix of explanatory variables for individual i, ι_T for the ones vector of length T and I_T is the identity matrix of size $T \times T$.

Using this, we have

$$y_i = x_i\beta + u_i. \tag{4.2.2}$$

We will derive the GLS estimator by finding a transformation that undoes the correlation between the T elements of u_i, for a given i, because it will be zero across individuals, by assumption. A first step is to look at the covariance between u_{is} and u_{it}, which—as u_{is} and u_{it} are both mean zero—is given by

$$\begin{aligned}\mathbb{E}[u_{is}u_{it}] &= \mathbb{E}[(\alpha_i + \varepsilon_{it})(\alpha_i + \varepsilon_{it})] \\ &= \mathbb{E}\left[\alpha_i^2 + \alpha_i\varepsilon_{it} + \varepsilon_{is}\alpha_i + \varepsilon_{is}\varepsilon_{it}\right].\end{aligned}$$

The assumptions imply that the last three elements are zero for $s \neq t$ such that $\mathbb{E}[u_{is}u_{it}] = \sigma_\alpha^2$. Conversely, for $s = t$, we have that he variance is $\mathbb{E}[u_{it}u_{it}] = \sigma_\alpha^2 + \sigma_\varepsilon^2$. Hence,

$$var(u_i) = \begin{pmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

or, in more compact notation,

$$var(u_i) = \sigma_\alpha^2 \iota_T \iota_T' + \sigma_\varepsilon^2 I_T.$$

Denote this variance-covariance matrix by Ω. The next step, that we omit here, is to show that the Cholesky factor associated with Ω^{-1} is

$$P = \frac{1}{\sqrt{\sigma_\varepsilon^2}}\left(\sqrt{\Psi} \cdot \frac{\iota_T \iota_T'}{T} + \left(I_T - \frac{\iota_T \iota_T'}{T}\right)\right), \tag{4.2.3}$$

where

$$\Psi \equiv \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_\alpha^2 + \sigma_\varepsilon^2}}$$

characterizes the relative importance of the variance of ε_i. As we have seen before, the random effects estimator will estimate β after transforming the data using (4.2.3).

The transformation of y_i involves calculating

$$\left(I_T - \frac{\iota_T \iota_T'}{T}\right) y_i,$$

which is the derivation of y_i from its individual mean, as $I_T y_i = y_i$ and $(\iota_T \iota_T'/T) y_i$ is the average of y_i. Just using this yields the within group estimator in Section (3.6.3). On top of this, we add the average y_i,

$$\frac{\iota_T \iota_T'}{T} y_i,$$

the more the higher the relative importance of σ_ε^2, as measured by Ψ.

This yields transformed error terms that are i.i.d. To be precise, in order to efficiently estimate β from the transformed data by running OLS on $Py_i = Px_i\beta + P\varepsilon_i$, we need to assume that the unobservables are uncorrelated across individuals, that α_i and ε_{it} are uncorrelated, and that the composite error term $u_{it} \equiv \alpha_i + \varepsilon_{it}$ is not correlated with x_{is} for all s. The last requirement is referred to as strict exogeneity because it holds across time periods (Engle et al., 1983). It is stronger than the uncorrelatedness assumption made in Section (3.6.3) for the fixed effects estimator. The reason is that here, we make the assumption for u_{it}, as opposed to requiring strict exogeneity only for x_i with respect to ε_{it}.

This estimator can also be seen as a GMM estimator—that we discuss below in Section 4.4—that is based on the moment conditions

$$\mathbb{E}\left[\left((y_i - x_i\beta)\iota_K'\right) \otimes x_i\right],$$

where $\otimes$ is the Kronecker product multiplying all elements of A in $A \otimes B$ with all elements of B. However, So Im et al. (1999) show that these are unnecessarily many moment conditions in the sense that some of them are redundant. Dropping the redundant conditions yields the random effects estimator.

4.3 Maximum Likelihood Estimation

This review is a modified version of Chapter 13 in Wooldridge (2002). Some of the derivations have been made more explicit.[2]

Maximum likelihood estimators are built on the idea that we can specify a probability model for the variation in observables. For parametric maximum likelihood estimation the model is formulated up to a finite set of parameters.[3] Identification means in this context that only the true parameters maximize the likelihood (according to the model) for observing the data (in the limit, when we have an infinite amount of data). Maximum likelihood estimation is efficient if the model is correctly specified. Otherwise, the estimator is generally inconsistent. Hence, there is a tradeoff between precision of the estimation results and biases that one might incur.

It is worth mentioning that the estimator may still be consistent even if the model is misspecified, as explained for example in White (1982). We will see in Section 4.3.7 that the OLS estimator for the linear model $y_i = x_i\beta + \varepsilon_i$ is at the same time the maximum likelihood estimator if we specify ε_i to be normally distributed. But then, even if we misspecify the model because the true distribution of ε_i is not the normal one, then the estimator will still be consistent for β as long as x_i and ε_i are uncorrelated.

The traditional version of maximum likelihood theory in statistics formulates a probability model for the joint distribution of $y_i \in \mathbb{R}^G$ and $x_i \in \mathbb{R}^K$, where the observations are identically and independently distributed for $i = 1,\ldots,N$. In economic applications y_i is an outcome, or a vector of outcomes (a vector of dependent variables), and x_i contains variables we condition on. As we are typically not interested in the distribution of those conditioning variables we leave it unspecified and formulate a model for y_i conditional on x_i.

We assume that the probability density function (p.d.f.) of y_i conditional on x_i is

[2] See also Chapters 14 through 17 in Ruud (2000) for a more details on the asymptotic theory and computational issues, Cameron and Trivedi (2005) for a more condensed presentation, and Engle (1984) for hypothesis tests. The interested reader is also referred to Newey and McFadden (1994) for details on the asymptotic theory and the relationship between the generalized method of moments and maximum likelihood estimation.

[3] There are generalizations such as semiparametric maximum likelihood estimation, see for example Gallant and Nychka (1987). They allow us to additionally estimate the shape of the distribution of unobservables.

known up to a finite set of parameters, θ, and denote it by $f(y_i|x_i;\theta)$. The simplest possible example I can think of is the following. It does not incorporate covariates.

Example 1 (Red and Blue Balls)**.** Suppose we draw balls with replacement out of a pool. There are red balls ($y_i = 1$) and blue balls ($y_i = 0$). We want to estimate the fraction of red balls that we denote by p. p is the only element of θ. According to the probability model the likelihood to observe $y_i = 1$ is p and the likelihood to observe $y_i = 0$ is $1 - p$. Hence,

$$f(y_i|\theta) = p^{y_i} \cdot (1-p)^{1-y_i}$$

with $\theta = p$. □

In this example the data is directly informative about the unknown parameter p since this parameter is actually identical to $\Pr(y_i = 1)$. But I think it is nevertheless worthwhile to follow up on this example because it is so simple. Notice that here we do not make parametric assumptions, which means that we have a nonparametric model.

4.3.1 Identification

In Chapter 3 we have said that a structure is identified if there is no other observationally equivalent structure within the set of admissible structures. Here, this means that only one value of the parameter vector, θ, is compatible with the observed distribution of y_i given x_i.

For Example 1 this means that the structure is identified if the parameter p is identified. This is the case because we know the distribution of y_i and $p = \Pr(y_i = 1)$ by definition. Here it is obvious that there is a one to one relationship between the unknown parameter, p, and the distribution of y_i, very much in the spirit of Lemma 1. This might be less obvious in other cases, in particular if we incorporate covariates.

We now start our discussion with introducing the notation and some definitions. Denote the support of x_i by $\mathscr{X}$ and the support of y_i by $\mathscr{Y}$.[4] Moreover, let $f(y_i|x_i;\theta_0)$ denote the true conditional p.d.f. of y_i given x_i. That is, the true value of θ is denoted by θ_0. $\Theta \subset \mathbb{R}^P$ is the parameter space. We assume that $f(y_i|x_i;\theta_0)$ is a density with respect

[4] Loosely speaking, the support of a random variable is the set of values that it might take on.

to the σ-finite measure $\nu(dy_i)$. Most measures we usually deal with in econometrics are σ-finite. We will work with integrals in this section which are written using $\nu(dy_i)$, for example $\int_{\mathscr{Y}} y_i \, f(y_i|x_i;\theta) \, \nu(dy_i)$ for the mean of y_i for a particular value of θ and conditional on x_i.[5] Finally, denote the conditional log likelihood for observation i by

$$\ell_i(\theta) \equiv \log(f(y_i|x_i;\theta)).$$

An important property of such a model is the *conditional Kullback-Leibler information inequality* which establishes that for any non-negative function $f(\cdot|x_i;\theta)$ that integrates to 1, that is

$$\int_{\mathscr{Y}} f(y_i|x_i;\theta) \, \nu(dy_i) = 1, \tag{4.3.1}$$

we have that

$$\int_{\mathscr{Y}} \log\left(\frac{f(y_i|x_i;\theta_0)}{f(y_i|x_i;\theta)}\right) \, f(y_i|x_i;\theta_0) \, \nu(dy_i) \geq 0.$$

Since $f(y_i|x_i;\theta_0)$ is the true density of y_i given x_i the integral here is an expectation. The log of a fraction is equal to the log of the numerator minus the log of the denominator so that we get that the left hand side is the difference between two expectations, or

$$\mathbb{E}[\log(f(y_i|x_i;\theta_0))|x_i] \geq \mathbb{E}[\log(f(y_i|x_i;\theta))|x_i].$$

Using the notation that we have introduced before we have

$$\mathbb{E}[\ell_i(\theta_0)|x_i] \geq \mathbb{E}[\ell_i(\theta)|x_i].$$

That is, the expected conditional log likelihood is maximized at the true parameter value. Hence, θ_0 solves

$$\max_{\theta\in\Theta} \mathbb{E}[\ell_i(\theta)|x_i]$$

for all $x_i \in \mathscr{X}$. By iterated expectations this implies that it also solves

$$\max_{\theta\in\Theta} \mathbb{E}[\ell_i(\theta)]. \tag{4.3.2}$$

[5] Readers who are not familiar with this can think of $\nu(dy_i)$ as corresponding to the usual dy_i in integrals.

This shows that it makes sense to estimate θ_0 by maximizing the sample counterpart of $\mathbb{E}[\ell_i(\theta)]$ provided that θ_0 is identified.

In the terminology of the previous chapter θ_0 fully characterizes the structure S that has generated the data. It is identified by the model if there is no other structure S' admitted by the model that solve the maximization problem (4.3.2). Here, this is the case if θ_0 is the unique maximizer. If we are only interested in a subvector θ_0^1 this subvector is identified by the model if it is the same for all structures admitted by the model that solve the same maximization problem.

Example 2 (Red and Blue Balls II)**.** The log likelihood is, using $\theta = p$,

$$\ell_i(p) = \log\left(p^{y_i} \cdot (1-p)^{1-y_i}\right) = y_i \cdot \log(p) + (1-y_i) \cdot \log(1-p).$$

The expected log likelihood as a function of p is

$$\mathbb{E}[\ell_i(p)] = \mathbb{E}[y_i] \cdot \log(p) + \mathbb{E}[1-y_i] \cdot \log(1-p).$$

$\mathbb{E}[y_i] = \Pr(y_i = 1)$ and $\mathbb{E}[1-y_i] = 1 - \Pr(y_i = 1)$ so that

$$\mathbb{E}[\ell_i(p)] = \Pr(y_i = 1) \cdot \log(p) + (1 - \Pr(y_i = 1)) \cdot \log(1-p).$$

Now (4.3.2) says that the true value of p solves, using $\theta = p$ and $\Theta = [0,1]$,

$$\max_{p \in [0,1]} \mathbb{E}[\ell_i(p)] = \Pr(y_i = 1) \cdot \log(p) + (1 - \Pr(y_i = 1)) \cdot \log(1-p).$$

This is a differentiable, concave function. The first order condition for a maximum is[6]

$$\Pr(y_i = 1) \cdot \frac{1}{p} + (1 - \Pr(y_i = 1)) \cdot \frac{1}{1-p} \cdot (-1) = 0. \tag{4.3.3}$$

Solving for p yields $p = \Pr(y_i = 1)$. □

In this simple example we have seen that indeed the log likelihood function is maximized at the true parameter value. Notice that taking the log of a function is a positive monotone transformation. Therefore, the true parameter value not only maximized the log of the likelihood function, but also the likelihood function itself. Hence the name maximum likelihood estimation.

[6] Strictly speaking this is the first order condition of the unconstrained problem, ignoring the constraint $p \in [0,1]$. However, we will see that the constraint is not binding here.

4.3.2 Estimation

The analogy principle in econometrics is to estimate expectations using sample analogs (Manski, 1988a). It also applies here, as the estimator is defined by the sample analog of the maximization problem (4.3.2),

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta).$$

It is typically obtained by numerical optimization but there are cases in which an analytic solution exists. Next, we study important properties of the expected likelihood and then discuss asymptotic properties of the estimator. Throughout, we assume that θ is a P-vector that lies in the interior of Θ and that ℓ_i is twice continuously differentiable on the interior of Θ. Moreover, we assume that integration and differentiation can be interchanged when necessary.

Example 3 (Red and Blue Balls III)**.** The maximum likelihood estimator of p, $\hat{p}$, solves

$$\max_{p \in [0,1]} \frac{1}{N} \sum_{i=1}^{N} \ell_i(p) = \left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) \cdot \log(p) + \left(1 - \frac{1}{N} \sum_{i=1}^{N} y_i \right) \cdot \log(1-p).$$

The first order condition for a maximum is

$$\left(\frac{1}{N} \sum_{i=1}^{N} y_i \right) \cdot \frac{1}{\hat{p}} + \left(1 - \frac{1}{N} \sum_{i=1}^{N} y_i \right) \cdot \frac{1}{1-\hat{p}} \cdot (-1) = 0.$$

Solving for $\hat{p}$ yields

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{4.3.4}$$

That is, quite intuitively, the maximum likelihood estimator of the fraction of red balls is the sample fraction of red balls. □

4.3.3 Properties of the Expected Log Likelihood

Define the score of the log likelihood as

$$s_i(\theta) \equiv \frac{\partial \ell_i(\theta)}{\partial \theta},$$

a P-vector.

We assume that integration and differentiation can be interchanged so that

$$\frac{\partial}{\partial \theta}\left(\int_{\mathscr{Y}} f(y_i|x_i;\theta)\ \nu(dy_i)\right) = \int_{\mathscr{Y}} \frac{\partial}{\partial \theta} f(y_i|x_i;\theta)\ \nu(dy_i).$$

Then, the derivative of the identity

$$\int_{\mathscr{Y}} f(y_i|x_i;\theta)\ \nu(dy_i) = 1$$

is

$$\int_{\mathscr{Y}} \frac{\partial}{\partial \theta} f(y_i|x_i;\theta)\ \nu(dy_i) = 0. \tag{4.3.5}$$

$\ell_i(\theta) = \log(f(y_i|x_i;\theta))$ implies that

$$s_i(\theta) = \frac{\partial \ell_i(\theta)}{\partial \theta} = \frac{\partial f(y_i|x_i;\theta)/\partial \theta}{f(y_i|x_i;\theta)}. \tag{4.3.6}$$

so that we can rewrite (4.3.5) as

$$\int_{\mathscr{Y}} s_i(\theta) f(y_i|x_i;\theta)\ \nu(dy_i) = 0, \tag{4.3.7}$$

the *conditional score identity*. The expectation thereof, over x_i, gives the *unconditional score identity*. We speak of an identity here because (4.3.7) holds for all values of θ.

Evaluated at $\theta = \theta_0$ we have

$$\mathbb{E}[s_i(\theta_0)|x_i] = 0,$$

which implies, by iterated expectations,

$$\mathbb{E}[s_i(\theta_0)] = 0.$$

Both are systems of K equations with K unknowns. Notice that these are only expectations once we evaluate the density function $f(y_i|x_i;\theta)$ at $\theta = \theta_0$. It is intuitive that the derivative of the expected log likelihood is zero at the true parameter value because it is maximized at this value.

Next, define the Hessian to be the symmetric $P \times P$ matrix

$$H_i(\theta) \equiv \frac{\partial s_i(\theta)}{\partial \theta'}.$$

Here, we take the derivative of a P-column vector with respect to a P-row vector. This yields a $P \times P$ matrix.

We assume again that we can interchange integration and differentiation so that the derivative of (4.3.7) with respect to θ' is given by

$$\int_{\mathscr{Y}} \left(\frac{\partial}{\partial \theta'} s_i(\theta) f(y_i|x_i;\theta) + s_i(\theta) \frac{\partial}{\partial \theta'} f(y_i|x_i;\theta) \right) \nu(dy_i) = 0.$$

Using (4.3.6) we can write this as

$$\int_{\mathscr{Y}} \left(\frac{\partial}{\partial \theta'} s_i(\theta) f(y_i|x_i;\theta) + s_i(\theta) s_i(\theta)' f(y_i|x_i;\theta) \right) \nu(dy_i) = 0 \tag{4.3.8}$$

so that for $\theta = \theta_0$

$$-\mathbb{E}[H_i(\theta_0)|x_i] = \mathbb{E}[s_i(\theta_0) s_i(\theta_0)'|x_i],$$

the *conditional information matrix equality*. By the score identity the expectation of the score is zero so that the right hand side is the variance-covariance matrix of the score, $var(s_i(\theta_0)|x_i)$.

By iterated expectations we get

$$-\mathbb{E}[H_i(\theta_0)] = \mathbb{E}[s_i(\theta_0) s_i(\theta_0)'],$$

the *unconditional information matrix equality*. Observe that whereas (4.3.7) holds for any value of θ the (conditional) information matrix equality only holds for $\theta = \theta_0$.

This is because only at the true parameter vector the right hand side of (4.3.8) is an expectation.

Finally, define

$$A_0 \equiv -\mathbb{E}[H_i(\theta_0)].$$

By the unconditional information matrix equality and the score identity this is the variance-covariance matrix of the score of the unconditional log likelihood.

Example 4 (Red and Blue Balls IV)**.** The first order condition (4.3.3) with $\Pr(y_i = 1) = p$ shows directly that the score identity holds here. As for the information matrix equality we have that the score is given by

$$s_i(p) = y_i \cdot \frac{1}{p} + (1 - y_i) \cdot \frac{1}{1-p} \cdot (-1)$$

so that the outer product of the score is

$$\begin{aligned} s_i(p)s_i(p)' &= \left(y_i \cdot \frac{1}{p}\right)^2 - 2 \cdot y_i \cdot \frac{1}{p} \cdot (1 - y_i) \cdot \frac{1}{1-p} + \left((1 - y_i) \cdot \frac{1}{1-p}\right)^2 \\ &= y_i \cdot \frac{1}{p^2} + (1 - y_i) \cdot \frac{1}{(1-p)^2}. \end{aligned}$$

Here, the second equality holds because $y_i \cdot (1 - y_i) = 0$, $y_i^2 = y_i$, and $(1 - y_i)^2 = (1 - y_i)$.

The Hessian is given by

$$H_i(p) = \frac{\partial s_i(p)}{\partial p} = -y_i \cdot \frac{1}{p^2} - (1 - y_i) \cdot \frac{1}{(1-p)^2}.$$

It follows directly that the information matrix equality, $-\mathbb{E}[H_i(p)] = \mathbb{E}[s_i(p)s_i(p)']$, holds because the negative of the Hessian is equal to the outer product of the score.

This, however, is not true in general—only if $p = \Pr(y_i = 1)$. Then,

$$
\begin{aligned}
(4.3.9) \quad \mathbb{E}[s_i(p)s_i(p)'] &= -\mathbb{E}[H_i(p)] \\
&= \Pr(y_i = 1) \cdot \frac{1}{p^2} + (1 - \Pr(y_i = 1)) \cdot \frac{1}{(1-p)^2} \\
&= \frac{1}{\Pr(y_i = 1)} + \frac{1}{1 - \Pr(y_i = 1)} \\
&= \frac{1 - \Pr(y_i = 1)}{\Pr(y_i = 1) \cdot (1 - \Pr(y_i = 1))} + \frac{\Pr(y_i = 1)}{\Pr(y_i = 1) \cdot (1 - \Pr(y_i = 1))} \\
&= \frac{1}{\Pr(y_i = 1) \cdot (1 - \Pr(y_i = 1))}. \quad \square
\end{aligned}
$$

4.3.4 Asymptotic Properties of the Estimator

The estimator is consistent and normally distributed with

$$\sqrt{N}(\hat{\theta} - \theta_0) \to^d \mathcal{N}(0, A_0^{-1}).$$

Newey and McFadden (1994) provide elegant proofs for the unconditional likelihood case. See Ruud (2000) and Wooldridge (2002) for the conditional case.

The variance of the estimator is A_0^{-1}/N and is equal to the Cramér-Rao lower bound for the variance of a parametric estimator. This means that the maximum likelihood estimator is efficient.

Next, we discuss the intuition behind consistency and asymptotic normality. For the former we have that the average sample likelihood function converges in probability, for every θ, to the expected likelihood function, $\mathbb{E}[l_i(\theta)]$. This function is maximized at the true parameter value, θ_0. Therefore, under conditions for interchanging the maximization and limiting operations the limit of the maximizing $\hat{\theta}$ is the maximum of the limit.

For asymptotic normality observe that the maximum likelihood estimator sets the average sample score to zero, that is

$$\frac{1}{N} \sum_{i=1}^{N} s_i(\theta) = 0.$$

The right hand side can be approximated by a first order Taylor series expansion in θ about $\theta = \theta_0$,

$$\frac{1}{N}\sum_{i=1}^{N} s_i(\theta_0) + \frac{1}{N}\sum_{i=1}^{N} H_i(\theta_0)(\hat{\theta} - \theta_0) \approx 0.$$

Hence,

$$\hat{\theta} \approx \theta_0 - \left(\frac{1}{N}\sum_{i=1}^{N} H_i(\theta_0)\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} s_i(\theta_0)\right). \tag{4.3.10}$$

The negative of the average Hessian in the sample converges to its expectation, A_0, that is

$$-\frac{1}{N}\sum_{i=1}^{N} H_i(\theta_0) \to^p A_0,$$

and $\sqrt{N}$ times the average sample score converges in distribution to a normal distribution with mean zero and, by the information matrix equality, variance A_0, that is

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N} s_i(\theta_0) \to^d \mathcal{N}(0, A_0).$$

Therefore, the maximum likelihood estimator $\hat{\theta}$ is normally distributed with mean θ_0 and variance A_0^{-1}/N.

We illustrate this using the example from above.

Example 5 (Red and Blue Balls V)**.** We have established in (4.3.4) that the maximum likelihood estimator of p is given by

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

By the law of large numbers the right hand side converges to $\Pr(y_i = 1)$. Hence, the estimator is consistent. It is asymptotically normally distributed by a central limit theorem. Since y_i is binomial the variance of the estimator is

$$var(\hat{p}) = \Pr(y_i = 1) \cdot (1 - \Pr(y_i = 1))/N.$$

The variance of the estimator is given by the inverse of the negative of the expected Hessian, divided by N. Using (4.3.9) we have

$$(-\mathbb{E}[H_i(p)])^{-1}/N = \Pr(y_i = 1) \cdot (1 - \Pr(y_i = 1))/N. \quad \square$$

4.3.5 Variance Estimation

The information matrix equality establishes that the negative of the expected Hessian is equal to the expectation of the outer product of the score. Therefore, two equally valid estimators for the variance of the estimator are given by $1/N$ times the inverse of

$$\frac{1}{N}\sum_{i=1}^{N} -H_i(\hat{\theta})$$

or

$$\frac{1}{N}\sum_{i=1}^{N} s_i(\hat{\theta})s_i(\hat{\theta})'.$$

These are the average negative of the Hessian evaluated at the parameter estimate and the average outer product of the score evaluated at the parameter estimate, respectively.

Example 6 (Red and Blue Balls VI)**.** The first estimator for the variance of $\hat{p}$ is

$$\begin{aligned}
\left(\frac{1}{N}\sum_{i=1}^{N} -H_i(\hat{p})\right)^{-1} \Big/ N &= \left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i}{\hat{p}^2} + \frac{1-y_i}{(1-\hat{p})^2}\right)\right)^{-1} \Big/ N \\
&= \left(\frac{\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)}{\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2} + \left(\frac{\frac{1}{N}\sum_{i=1}^{N}(1-y_i)}{\left(1-\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2}\right)\right)^{-1} \Big/ N \\
&= \left(\frac{\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)}{\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2} + \left(\frac{1-\frac{1}{N}\sum_{i=1}^{N} y_i}{\left(1-\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2}\right)\right)^{-1} \Big/ N \\
&= \left(\frac{1}{\frac{1}{N}\sum_{i=1}^{N} y_i} + \frac{1}{1-\frac{1}{N}\sum_{i=1}^{N} y_i}\right)^{-1} \Big/ N \\
&= \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) \cdot \left(\frac{1}{N}\left(1 - \frac{1}{N}\sum_{i=1}^{N} y_i\right)\right) \Big/ N
\end{aligned}$$

where the last equality follows from arguments similar to the ones in (4.3.9). So the estimator for the variance of $\hat{p}$ is the sample variance of y_i divided by N. This is the sample counterpart of the theoretical variance.
The second estimator is the same as the first estimator because in this specific example the outer product of the score is equal to the negative Hessian. Usually, this only holds in expectation. □

4.3.6 Goodness of Fit

Goodness of fit measures aim at summarizing how well a model is able to predict an outcome once we feed in covariates. The best known such measure is the R^2 measure for the linear model. For the OLS estimator it is the analog of the variance of the predicted value $x_i\beta$ in the sample divided by the variance of y_i. It is zero if the covariates are not able to explain any of the variation in y_i and one if they are perfectly able to explain this variation. This latter case is more of a theoretical possibility since it requires that ε_i is degenerate.

The best known measure for models that are estimated using maximum likelihood is McFadden's Pseudo-R^2 measure. It compares the sample likelihood for the fitted model,

$$L_{fit} = \mathscr{L}(\hat{\theta})$$

to the likelihood for a model without covariates, denoted by L_0. The measure is

$$R^2_{\text{McFadden}} = 1 - \frac{L_{\text{fit}}}{L_0}.$$

Table 4.1 illustrates this. The first column contains different values for the estimated likelihood to observe the sample $\{y_i\}_{i=1}^N$, using $\{x_i\}_{i=1}^N$. This is denoted by $\prod f(y_i|x_i;\hat{\theta})$, where $\prod$ is shorthand notation for the product operator $\prod_{i=1}^N$. $\hat{\theta}$ is the maximum likelihood estimate of θ. The estimated likelihood to observe the sample $\{y_i\}_{i=1}^N$ is denoted by $\prod \hat{f}(y_i)$ and is assumed to be equal to 0.1.[7] $\prod f(y_i|x_i;\hat{\theta})$ can never be below 0.1 and if it is equal to 0.1 then the covariates contain no information. The third and fourth column contain L_{fit} and L_0, respectively. L_{fit} is the log of the value

[7] The maximum likelihood estimate of $f(y_i)$, $\hat{f}(y_i)$, could be the sample frequency if y_i is discrete or a parametric estimate of $f(y_i)$ if it is continuous.

Table 4.1: McFadden's goodness of fit measure.

$\prod f(y_i\|x_i;\hat{\theta})$	$\prod \hat{f}(y_i)$	L_{fit}	L_0	L_{fit}/L_0	$1-L_{\text{fit}}/L_0$
0.1	0.1	-2.30	-2.30	1.00	0.00
0.2	0.1	-1.61	-2.30	0.70	0.30
0.3	0.1	-1.20	-2.30	0.52	0.48
0.4	0.1	-0.92	-2.30	0.40	0.60
0.5	0.1	-0.69	-2.30	0.30	0.70
0.6	0.1	-0.51	-2.30	0.22	0.78
0.7	0.1	-0.36	-2.30	0.15	0.85
0.8	0.1	-0.22	-2.30	0.10	0.90
0.9	0.1	-0.11	-2.30	0.05	0.95
1.0	0.1	0.00	-2.30	0.00	1.00

in the first column and L_0 is the log of the value in the second column. The fifth column contains L_{fit}/L_0 and the last column contains $R^2_{\text{McFadden}} = 1 - L_{\text{fit}}/L_0$. This table shows that R^2_{McFadden} must lie between zero and one. This is because $\prod f(y_i|x_i;\hat{\theta}) \geq \prod \hat{f}(y_i)$ and $\prod f(y_i|x_i;\hat{\theta}) \leq 1$. It also shows why the measure is not simply L_{fit}/L_0. The reason for this is that the log of a number below one is always negative, as can be seen in the third and fourth column. If we divide the value in the third column by the one in the fourth column the result is always a number between zero and one that is *de*creasing in $\prod f(y_i|x_i;\hat{\theta})$. To offset this the measure is one minus this ratio.

4.3.7 Example: Linear Regression

For further illustration consider the linear model $y_i = x_i\beta_0 + \varepsilon_i$ where y_i is a scalar and x_i' as well as β_0 are K-vectors. Let ε_i be normally distributed with mean zero conditional on x_i and variance σ_0^2 so that the density of y_i given x_i as a function of θ is given by

$$f(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right).$$

Here, $\theta = (\beta', \sigma^2)'$. The log likelihood is given by

$$\ell_i(\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}.$$

The maximum likelihood estimator maximizes the average sample likelihood over the choice of θ, that is

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta \in \mathbb{R}^{K+1}} \frac{1}{N} \sum_{i=1}^{N} \ell_i(\theta) \\
&= \arg\max_{\theta \in \mathbb{R}^{K+1}} \frac{1}{N} \sum_{i=1}^{N} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2} \right) \\
&= \arg\max_{\theta \in \mathbb{R}^{K+1}} -\frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - x_i\beta)^2 \\
&= \arg\max_{\theta \in \mathbb{R}^{K+1}} -\frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)
\end{aligned}$$

where the third equality follows from the fact that the maximizing value of θ is invariant to positive monotone transformations of the objective function. Here, we have dropped the first term in parenthesis and have multiplied the remaining expression by N. For the forth equality we write $y = (y_1, \ldots, y_N)'$ and $X = (x_1', \ldots, x_K')'$.

We obtain the sample sum of score functions by differentiating with respect to β and σ^2, which yields[8]

$$\sum_{i=1}^{N} s_i(\theta) = \begin{bmatrix} \frac{1}{\sigma^2}(X'y - X'X\beta) \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{bmatrix}.$$

The first order conditions at $\theta = \hat{\theta}$ are that the score is equal to zero, that is[9]

$$\begin{aligned}
\frac{1}{\hat{\sigma}^2}(X'y - X'X\hat{\beta}) &= 0 \\
-\frac{N}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}(y - X\hat{\beta})'(y - X\hat{\beta}) &= 0.
\end{aligned}$$

[8] For the first equation the steps are similar to the ones for obtaining the ordinary least squares estimator in matrix notation. At this point one can derive the score function and the Hessian for observation i and show that the score identity holds.

[9] Notice that there are $K+1$ equations and $K+1$ unknowns.

From these we can get analytic expressions for the maximum likelihood estimators of $\hat{\beta}$ and $\hat{\sigma}^2$ provided that $X'X$ has full rank so that it is invertible. They are

$$\hat{\beta} = (X'X)^{-1}X'y$$

and

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\hat{\beta})^2.$$

Interestingly, in this particular case the OLS estimator for β is maximizing the sample likelihood. Notice that the estimator for the variance of the residual is not unbiased but consistent.[10]

The Hessian is given by the derivative of both elements of the score with respect to both β and σ^2. As we have derived the sample sum of scores before we get

$$\sum_{i=1}^{N} H_i(\theta) = \begin{bmatrix} -\frac{1}{\sigma^2}X'X & -\frac{1}{\sigma^4}(X'y - X'X\beta) \\ -\frac{1}{\sigma^4}(X'y - X'X\beta) & \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta) \end{bmatrix}.$$

From this we get

$$A_0 = -\mathbb{E}[H_i(\theta_0)|X] = \begin{bmatrix} \frac{1}{\sigma_0^2}\mathbb{E}[x_i'x_i] & 0 \\ 0 & \frac{N}{2\sigma_0^4} \end{bmatrix}$$

where we use that

$$\mathbb{E}[X'X] = N\mathbb{E}[x_i'x_i],$$

$$\mathbb{E}[X'y - X'X\beta_0|X] = X'X\beta_0 + \mathbb{E}[X'u|X] - X'X\beta_0 = 0,$$

and

$$\mathbb{E}\left[\frac{1}{\sigma_0^6}(y - X'\beta_0)'(y - X'\beta_0)\right] = \frac{1}{\sigma_0^6}N\sigma_0^2.$$

[10] The unbiased estimator is

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i - x_i\hat{\beta})^2.$$

Finally, we get that the variance covariance matrix of the maximum likelihood estimator is given by

$$var(\hat{\theta}) = A_0^{-1}/N = \begin{bmatrix} \sigma_0^2 \mathbb{E}[x_i' x_i]^{-1}/N & 0 \\ 0 & 2\sigma_0^4/N \end{bmatrix}.$$

This matrix is block-diagonal, that is the covariance between the estimator for the slope coefficients and the estimator for the variance of the error term is zero. □

4.3.8 Hypothesis Testing

For maximum likelihood estimators we have already seen that under some regularity conditions $\sqrt{N}(\hat{\theta} - \theta_0) \to^a \mathcal{N}(0, A_0^{-1})$ provided that θ_0 is identified.[11] We shall now discuss two-sided tests of Q nonlinear restrictions where the null is given by $c(\theta_0) = 0$ with $c(\theta_0)$ being a Q-vector. Throughout we assume that θ_0 is in the interior of Θ under the null and that $C(\theta_0) \equiv \partial c(\theta)/\partial \theta|_{\theta=\theta_0}$ has rank Q.

We denote the unconstrained estimator by $\hat{\theta}$ and let $\tilde{\theta}$ be the estimator under the the constraints that are imposed by the null. Moreover, we let $\hat{V}$ be an estimator of the variance of $\hat{\theta}$, for example $N^{-1}\sum_{i=1}^N -H_i(\hat{\theta})$, and $\hat{C} \equiv C(\hat{\theta})$. We proceed similarly for $\tilde{\theta}$ and define $\tilde{s}_i \equiv s_i(\tilde{\theta})$ in addition. Finally, define

$$\mathscr{L}(\theta) \equiv \log\left(\prod_{i=1}^N f(y_i|x_i;\theta)\right) = \sum_{i=1}^N \ell_i(\theta)$$

to be the log likelihood for the entire sample.

Denote the kth element of θ_0 by θ_{0k}. An example of a null hypothesis is

$$c(\theta_0) = \begin{pmatrix} 3\theta_{01} - \sqrt{\theta_{02}} \\ \log\theta_{03} + \theta_{04} - 5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Here, we have 2 nonlinear constraints that involve four parameters.

There are three tests that are asymptotically equivalent. Under the null they follow a χ^2 distribution with Q degrees of freedom.

[11] This review is based on Engle (1984).

The first test is the *likelihood ratio test*. The idea behind this test is that if the null hypothesis is correct, then the log likelihood evaluated at the constrained estimator should be as high as the log likelihood evaluated at the unconstrained estimator. The latter can never be lower because every parameter value that is feasible under the constraints is also feasible without the constraints. So the test is whether $\mathscr{L}(\hat{\theta})$ is equal to $\mathscr{L}(\tilde{\theta})$. If not, the null is rejected. The statistic is

$$LR \equiv 2\left(\mathscr{L}(\hat{\theta}) - \mathscr{L}(\tilde{\theta})\right).$$

For this test one needs to obtain both the constrained and the unconstrained estimator.

The second test is the *Wald test*. For this test we estimate the unconstrained model and then test directly whether the constraints hold. So, we test whether $c(\hat{\theta})$ is equal to zero. For this we use the quadratic form

$$W \equiv c(\hat{\theta})'(\hat{C}\hat{V}\hat{C}')^{-1}c(\hat{\theta}).$$

$c(\hat{\theta})$ is the deviation of the restrictions from their value under the null, which is zero, evaluated at $\hat{\theta}$. $(\hat{C}\hat{V}\hat{C}')$ is the variance of $c(\hat{\theta})$ that is due to estimation error.

The third test is built on the idea that under the null the average (unconstrained) score should be equal to zero when we evaluate it at the constrained estimator (by construction, it is always zero at the unconstrained estimator provided that θ_0 is in the interior of Θ). This is why this test is sometimes referred to as the *score test*. It is also referred to as the *lagrange multiplier test* because we can test whether the constraints are binding by testing whether the lagrange multiplier is equal to zero. The test statistic is

$$LM \equiv \left(\sum_{i=1}^{N} \tilde{s}_i\right)' \tilde{V} \left(\sum_{i=1}^{N} \tilde{s}_i\right).$$

This is a quadratic form where $\tilde{V}$ is the inverse of the variance of the score for the constrained estimator.

Engle (1984) shows that the three test statistics are identical if the likelihood function is quadratic, that is if

$$\frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta) = a - \frac{1}{2}(\hat{\theta} - \theta)'V^{-1}(\hat{\theta} - \theta)$$

for some scalar a and a positive definite matrix V^{-1}. Consider a linear (or linearlized, which is always possible) null hypothesis that is of the form $\theta_0 = \tilde{\theta}$. Observe

$$\frac{\partial}{\partial\theta}\frac{1}{N}\sum_{i=1}^{N}\ell_i(\theta)' = \frac{1}{N}\sum_{i=1}^{N}s_i(\theta) = (\hat{\theta}-\theta)'V^{-1}$$
$$\frac{\partial}{\partial\theta}\frac{1}{N}\sum_{i=1}^{N}s_i(\theta) = \frac{1}{N}\sum_{i=1}^{N}H_i(\theta) = -V^{-1},$$

that is the sample variance of the maximum likelihood estimator,

$$\left(\frac{1}{N}\sum_{i=1}^{N}H_i(\theta)\right)^{-1} = V,$$

does not depend on θ. Therefore,

$$\begin{aligned}
W &= (\hat{\theta}-\tilde{\theta})'\hat{V}^{-1}(\hat{\theta}-\tilde{\theta}) = (\hat{\theta}-\tilde{\theta})'V^{-1}(\hat{\theta}-\tilde{\theta})\\
LM &= \left(\sum_{i=1}^{N}s_i(\tilde{\theta})\right)'\tilde{V}\left(\sum_{i=1}^{N}s_i(\tilde{\theta})\right) = (\hat{\theta}-\tilde{\theta})'V^{-1}(\hat{\theta}-\tilde{\theta})\\
LR &= -2\left(a-\frac{1}{2}(\hat{\theta}-\hat{\theta})'\hat{V}^{-1}(\hat{\theta}-\hat{\theta})\right.\\
&\qquad\left.-\left(a-\frac{1}{2}(\hat{\theta}-\tilde{\theta})'\tilde{V}^{-1}(\hat{\theta}-\tilde{\theta})\right)\right)\\
&= (\hat{\theta}-\tilde{\theta})'V^{-1}(\hat{\theta}-\tilde{\theta}),
\end{aligned}$$

the desired result. Observe that this example nicely illustrates how different expressions for the variance of the estimator enter the test statistics. Essentially, these test statistics coincide here because $\hat{V} = \tilde{V} = V$.

Even if the likelihood function is not quadratic these tests are asymptotically equivalent. To see this suppose θ_0 is close to the value under the null. Then, the likelihood function will be approximately quadratic for large samples in a neighborhood around the value of θ under the null. Moreover, under appropriate conditions the score and the Hessian of the average likelihood evaluated at the estimate for θ converge to the score

and the Hessian evaluated at θ_0. This is due to the consistency of the estimator. Then, the quadratic approximation becomes more and more accurate and, in essence, by the argument we made for quadratic likelihood functions the test statistics converge to one another.

4.4 Generalized Method of Moments

The generalized method of moments (GMM) is based upon properties of moment conditions which are known *a priori*.[12] It provides a unified asymptotic theory for a general class of estimators including OLS, instrumental variables and maximum likelihood. GMM itself is a member of the class of *extremum estimators* as some objective function is minimized.

Examples of moment conditions are given by

$$\mathbb{E}[x_i'(y_i - x_i\beta_0)] = 0 \tag{4.4.1}$$

for OLS—the error term and the vector of regressors are uncorrelated at the true parameter vector β_0. For instrumental variables estimation we have

$$\mathbb{E}[z_i'(y_i - x_i\beta_0)] = 0, \tag{4.4.2}$$

namely that the error term and the vector of instrumental variables are uncorrelated, and

$$\mathbb{E}[s_i(\beta_0)] = 0, \tag{4.4.3}$$

for maximum likelihood estimation where the score of the likelihood function is zero in expectation.

The moment conditions could also come from a structural model. The example in Hansen and Singleton (1982) is a stochastic Euler equation.

Let (x_i, z_i, y_i) be independently and identically distributed (i.i.d.) and let Θ denote the parameter space with typical element θ and $\theta_0 \in \Theta$ being the population P-vector

[12] I draw on some lecture notes by Ariel Pakes. Useful further readings are the original article by Hansen (1982) as well as Newey and McFadden (1994) and the whole book by Hayashi (2000), which is centered around GMM. From a historical perspective, an early application is Hansen and Singleton (1982).

of parameters to be estimated. $g_i(\theta) \equiv g(x_i, z_i, y_i; \theta)$ is a known Q-vector of functions for which

$$\mathbb{E}[g_i(\theta)] = 0 \tag{4.4.4}$$

if, and only if, $\theta = \theta_0$. This is a necessary condition for identification similar to the one being assumed for maximum likelihood estimation in Section 4.3. As for regularity conditions we require, among others and without discussing them any further,

$$\Phi \equiv \mathbb{E}[g_i(\theta_0) g_i(\theta_0)'] < \infty$$

and that

$$D \equiv \mathbb{E}\left[\frac{\partial g_i(\theta_0)}{\partial \theta'}\right]$$

is of rank P.

The sample analog of the vector of moment conditions, (4.4.4), is

$$\frac{1}{N}\sum_i g_i(\theta) = 0.$$

These are Q equations with P unknown parameters. If $Q < P$ identification fails. If $Q = P$ the system is just-identified and we can solve for θ. Finally, if $Q > P$ the system is over-identified.

The GMM estimator is

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \left[\frac{1}{N}\sum_i g_i(\theta)\right]' A_n^{-1} \left[\frac{1}{N}\sum_i g_i(\theta)\right],$$

where A_n is a $Q \times Q$ weighting matrix which is symmetric and positive definite with the property that $A_n \xrightarrow{p} \Psi$. It minimizes a weighted sum of squared deviations of the empirical moment conditions from zero and can be shown to be consistent and asymptotically normal,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Lambda),$$

where

$$\Lambda = (D'\Psi^{-1}D)^{-1} (D'\Psi^{-1}\Phi\Psi^{-1}D) (D'\Psi^{-1}D)^{-1}$$

and

$$\Phi \equiv \mathbb{E}[g_i(\theta_0)g_i(\theta_0)'].$$

The minimum of the variance, in a positive definite sense, is $(D'\Phi^{-1}D)^{-1}$ and is attained for the optimal weighting matrix A_n that satisfies $\Psi = \Phi$ in the over-identified case. If $Q = P$ the choice of Ψ does not matter.

To implement this estimator, we estimate the parameter vector in a first step using for instance the identity weighting matrix, that is $A_n = I$. Note that those estimates are consistent. Calculate $\hat{\Phi} = \sum_i g_i(\hat{\theta})g_i(\hat{\theta})'$, reset $A_n = \hat{\Phi}$, and use A_n for the efficient second step estimates of the parameters.

4.4.1 Ordinary Least Squares as a Special Case

Suppose that

$$y_i = x_i\beta + \varepsilon_i,$$

where x_i is $1 \times K$ and $\mathbb{E}[x_i\varepsilon_i] = 0$. Then,

$$g_i(\beta) = \mathbb{E}[x_i'(y_i - x_i\beta)] = 0$$

is a set of K valid moment conditions. The sample analog is $N^{-1}\sum_i x_i'(y_i - x_i\beta) = 0$ and we are in the just-identified case since this is a system of K equations with K unknowns. Hence, we can directly solve for β and get

$$\hat{\beta} = \left(\sum_i x_i'x_i\right)^{-1}\left(\sum_i x_i'y_i\right).$$

Furthermore, we have

$$\Phi = \mathbb{E}[g_i(\beta)g_i(\beta)'] = \mathbb{E}[x_i'\varepsilon_i\varepsilon_i x_i]$$

and

$$D = \mathbb{E}\left[\frac{\partial g_i(\beta)}{\partial \beta}\right] = \mathbb{E}[-x_i'x_i].$$

So, we get the White-Eicker variance-covariance matrix

$$\Lambda = D^{-1}\Phi D^{-1} = (\mathbb{E}[x_i'x_i])^{-1}\mathbb{E}[x_i'\varepsilon_i\varepsilon_i x_i](\mathbb{E}[x_i'x_i])^{-1}$$

that can be estimated by its sample counterpart.

4.4.2 Instrumental Variables Estimation as a Special Case

Consider a regression model

$$y_i = x_i\beta + \varepsilon_i, \tag{4.4.5}$$

where $i = 1,\dots,N$, x_i' and β are K-vectors, and ε_i represents unobserved heterogeneity. We start in the case in which the number of instruments is equal to the number of endogenous regressors. Then, we discuss estimators which can be used if the number of instruments exceeds the number of endogenous variables.

Throughout, we denote the L-vector of instrumental variables by z_i' and include the exogenous variables that we use in the regression in this vector. The moment conditions are in both cases

$$\mathbb{E}\left[z_i'(y_i - x_i\beta)\right] = 0.$$

4.4.2.1 Just-Identified Case

Assume that a vector of instruments exists, that is that

$$\mathbb{E}[\varepsilon_i|z_i] = 0 \tag{4.4.6}$$

and

$$\operatorname{rank}(\mathbb{E}[z_i'x_i]) = K, \tag{4.4.7}$$

and that the number of endogenous variables is equal to the number of instruments, that is $K = L$. This is called the just-identified case.

Then, we can pre-multiply (4.4.5) by z_i' and get

$$z_i'y_i = z_i'x_i\beta + z_i'\varepsilon_i.$$

Taking expectations yields

$$\mathbb{E}[z_i'y_i] = \mathbb{E}[z_i'x_i]\beta + \mathbb{E}[z_i'\varepsilon_i]$$

and by (4.4.6) and (4.4.7) we get

$$\beta = \mathbb{E}[z_i'x_i]^{-1}\mathbb{E}[z_i'y_i].$$

As an aside notice that this is the OLS estimator if there are no endogenous regressors so that $z_i = x_i$.

Finally, by writing $Z = (z_1', \ldots, z_N')'$ we can express the instrumental variables estimator as[13]

$$\hat{\beta} = (Z'X)^{-1}Z'y.$$

However, this is not possible in the so called over-identified case where $L > K$, simply because the number of rows of Z' is not equal to the number of columns of X so that $Z'X$ is not a square matrix. Of course, it is still possible to use only K of the L instruments. Another more efficient strategy is used in the following two stage approach.

4.4.2.2 Two Stage Least Squares

Suppose $L > K$. Then, following the idea underlying (**??**) we could project x_i on z_i by means of a linear regression (Kelejian, 1971) and then use the projection (the fitted value) to replace x_i in a second step. As the instrument is uncorrelated with the error term this projection will be uncorrelated as well.

Formally, write

$$\hat{x}_i \equiv \mathbb{E}[x_i|z_i]. \tag{4.4.8}$$

Then, we could pre-multiply the equation

$$y_i = \hat{x}_i\beta + \varepsilon_i$$

by $\hat{x}_i'$ and take expectations to get

$$\mathbb{E}[\hat{x}_i'y_i] = \mathbb{E}[\hat{x}_i'\hat{x}_i]\beta + \mathbb{E}[\hat{x}_i'\varepsilon_i].$$

Using iterated expectations we have

$$\mathbb{E}[\hat{x}_i'\varepsilon_i] = \mathbb{E}[\mathbb{E}[\hat{x}_i'\varepsilon_i|z_i]] = \mathbb{E}[\hat{x}_i'\mathbb{E}[\varepsilon_i|z_i]] = 0,$$

[13] Notice that z_i is $1 \times L$, x_i is $1 \times K$, Z is $N \times L$, and X is $N \times K$. y is $N \times 1$.

where the second equality follows from (4.4.8) and the last equality follows from (4.4.6). Hence

$$\mathbb{E}[\hat{x}_i' y_i] = \mathbb{E}[\hat{x}_i' \hat{x}_i]\beta$$

so that

$$\beta = \mathbb{E}[\hat{x}_i' \hat{x}_i]^{-1} \mathbb{E}[\hat{x}_i' y_i]$$

provided that $\mathbb{E}[\hat{x}_i' \hat{x}_i]$ has full rank.

The sample analog yields the estimator

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}y$$

where we can use an OLS regression of X on Z in the first stage to fit

$$\hat{X} = Z(Z'Z)^{-1}Z'X.$$

Here, $(Z'Z)^{-1}Z'X$ is the vector of regression coefficients that are obtained in such a regression. This estimator is referred to as the two stage least squares estimator (2SLS) and is attributed to Theil (1953).

Equivalently, we could have used $\hat{x}_i$ as an instrument and proceed as if we were in the just-identified case (Basmann, 1957). This is because

$$\mathbb{E}[\hat{x}_i' x_i] = \mathbb{E}[\hat{x}_i' \mathbb{E}[x_i|z_i]] = \mathbb{E}[\hat{x}_i' \hat{x}_i].$$

Finally, let $\hat{\eta}$ be the vector of residuals that we obtain in the first stage regression of X on Z,

$$\hat{\eta} \equiv (I - Z(Z'Z)^{-1}Z')X.$$

We will show that the 2SLS estimator, where we replace X by $\hat{X}$ is identical to an estimation procedure in which we include the first stage residual for the endogenous variable, $\hat{\eta}$, as an additional regressor into a regression of y on X.[14]

For this define

$$r_y \equiv (I - \hat{\eta}(\hat{\eta}'\hat{\eta})^{-1}\hat{\eta}')y,$$

the vector of residuals from a regression of y on $\hat{\eta}$, and

$$R_X \equiv (I - \hat{\eta}(\hat{\eta}'\hat{\eta})^{-1}\hat{\eta}')X,$$

[14] See also Ruud (2000, p.504) for an alternative derivation.

the matrix of residuals from a regression of each column of X on $\hat{\eta}$. Notice that

$$R_X = X - \hat{\eta} = \hat{X}$$

because the regression fit of a regression of X on $\hat{\eta}$ yields $\hat{\eta}$ itself and X minus $\hat{\eta}$ is the fitted value $\hat{X}$ by definition.

To prove that 2SLS is the same as regressing y on X and $\hat{\eta}$ we draw on the Frisch and Waugh (1933) theorem that states that a regression of y on X and $\hat{\eta}$ is numerically identical to a regression of r_y on R_X. In this new regression the dependent variable is replaced by the residuals from a regression of the original dependent variable on η and the covariates X are replaced by the residuals from a regression of the original X on $\hat{\eta}$. This yields

$$\hat{\beta} = \left(R_X' R_X\right)^{-1} R_X' r_y$$

which is

$$\hat{\beta} = \left(R_X' R_X\right)^{-1} X'(I - \hat{\eta}(\hat{\eta}'\hat{\eta})^{-1}\hat{\eta}')'(I - \hat{\eta}(\hat{\eta}'\hat{\eta})^{-1}\hat{\eta}')y.$$

This simplifies to

$$\hat{\beta} = \left(R_X' R_X\right)^{-1} R_X' y.$$

Finally, as $R_X = \hat{X}$ we get the 2SLS estimator.

Holly and Sargan (1982) developed a test for endogeneity that is related to the Hausman (1978) test and the idea that 2SLS is the same as the above regression. The is a test for the significance of the included residual.

Again the model is

$$y_i = x_i\beta + \varepsilon_i \tag{4.4.9}$$

with K-vector x_i and a vector of instrumental variables with $\mathbb{E}[z_i'(y_i - x_i\beta_0)] = 0$ and $\mathbb{E}[z_i' x_i] \neq 0$. In the just-identified case we can solve directly for

$$\hat{\beta} = \left(\sum_i z_i' x_i\right)^{-1} \left(\sum_i z_i' y_i\right).$$

In the over-identified case, we can pre-multiply both sides of (4.4.9) with $(z_i' x_i)' = x_i' z_i$ to get

$$\hat{\beta} = \left(\sum_i x_i' z_i z_i' x_i\right)^{-1} \left(\sum_i x_i' z_i z_i' y_i\right).$$

This is the two stage least squares estimator. To see this, write the empirical moment conditions as

$$\sum_i z_i'(y_i - x_i\beta_0) = 0$$

and choose

$$A_n = \sum_i z_i' z_i.$$

Then,

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\min} \left(n^{-1} \sum_i g_i(\theta) \right)' \left(\sum_i z_i' z_i \right)^{-1} \left(n^{-1} \sum_i g_i(\theta) \right)$$

and the first order condition is

$$\left(\sum_i z_i' x_i \right)' \left(\sum_i z_i' z_i \right)^{-1} \left(\sum_i z_i'(y_i - x_i\beta_0) \right) = 0,$$

which can be written as

$$\hat{\Pi}' \left(\sum_i z_i'(y_i - x_i\beta_0) \right),$$

where $\hat{\Pi}$ are the coefficients from a regression of x_i on z_i. Writing $\hat{x}_i' = \hat{\Pi}' z_i'$ for the fitted values, we get the two stage least squares estimator. In this case, we can use the identity weighting matrix and the GMM variance-covariance matrix is in fact the well-known White-Eicker variance-covariance matrix. The instrumental variables estimator for the just-identified case is

$$\hat{\beta} = \left(\sum_i \hat{x}_i' \hat{x}_i \right)^{-1} \left(\sum_i \hat{x}_i' y_i \right)$$

and we get the usual expression for the variance-covariance matrix,

$$\Lambda = D^{-1}\Phi D^{-1} = (\mathbb{E}[x_i z_i'])^{-1} \mathbb{E}[z_i \varepsilon_i \varepsilon_i z_i'] (\mathbb{E}[z_i x_i'])^{-1}.$$

4.4.3 Nonlinear Least Squares as a Special Case

The nonlinear least squares estimator is for the parametric model

$$y_i = m(x_i; \theta) + \varepsilon_i,$$

where the K-vector x_i satisfies $\mathbb{E}[x_i \varepsilon_i] = 0$. The corresponding GMM moment conditions are

$$g_i(\theta) = \mathbb{E}[x_i(y_i - m(x_i; \theta))] = 0$$

and we are again in the just-identified case since this is a system of K equations with K unknowns.

4.4.4 Over-Identifying Restrictions Test

We can carry out a specification test in the over-identified case. Intuitively, the reason is that if the number of moment conditions, L, exceeds the number of parameters, K, then we can estimate the parameters many times—always with a different subset of moment conditions—and then test whether the estimates are the same in a statistical sense.

Formally, it can be shown that under the null that the moment conditions are true the GMM objective function evaluated at the estimate $\hat{\theta}$ of θ_0,

$$\left[n^{-1} \sum_i g_i(\hat{\theta}) \right]' A_n^{-1} \left[n^{-1} \sum_i g_i(\hat{\theta}) \right],$$

follows a χ^2 distribution with $L - K$ degrees of freedom. This test is sometimes also referred to as a Sargan, or "J" test.

The idea behind the test goes back to a paper by Sargan (1958) on instrumental variables estimation based on the moment conditions (4.4.2), and papers by Durbin (1954), Wu (1973), and Hausman (1978). Here, if the number of instruments exceeds the number of endogenous regressors, that is if $L > K$, we can test whether the model is correctly specified. An incorrect specification could stem from, for example, a failure of the linearity assumption or from the fact that some instruments are not valid.

Here, we can alternatively calculate the test statistic by regressing the residual from the 2SLS regression of y_i on x_i using z_i again on z_i. The test statistic is equal to N times the R^2 of this regression.

4.5 Nonparametric Regression

A nonparametric regression aims at estimating the function m in

$$y_i = m(x_i) + \varepsilon_i,$$

where ε_i is mean independent of x_i and normalized to have a mean of zero. Here, x_i may be multi-dimensional. If all of the elements in x_i are discrete, then one can simply estimate m by calculating the average y_i among all observations with a particular value of x_i, say x. This puts zero weight on observations with values of x_i that are different from x, and equal weight to all the remaining observations.

If at least one element of x_i is continuously distributed, then equality occurs with probability zero, and therefore it is more appealing to estimate m as a local, or weighted average. For constructing the weight, one can use the concept of a kernel K, which (typically) is a weighting function that integrates to one and is symmetric about zero. Its argument is the distance between x_i and x divided by a bandwidth h so that one gets

$$K\left(\frac{x_i - x}{h}\right).$$

If one wishes to put the most weight on an observation if x_i is equal to x, then K needs to have its maximum at zero and decrease in the distance to zero. There are many kernel functions that have those properties, but in practice there is often little difference in the the results if one is chosen over another.

Keeping the function K as it is, the lower the bandwidth h, the more deviations between x_i and x are inflated, so that the more we move away from zero and the lower the weight will be. To obtain weights that sum to one it is necessary to divide $K\left((x_i - x)/h\right)$ by their sum across observations. Then, one can write the estimator for $m(x)$ as

$$\hat{m}(x) = \sum_{i=1}^{N} \left(\frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x_j - x}{h}\right)} \right) \cdot y_i,$$

where N is the number of observations.

This idea is attributed to both, Nadaraya (1964) and Watson (1964). Pagan and Ullah (1999, p. 84ff) summarizes the asymptotic results. Behind them stands the thought

experiment of letting N go to infinity at a faster rate at which h go to zero. They also discuss how one can select the bandwidth in practice.

4.6 Simulated Maximum Likelihood and Simulated Method of Moments

Sometimes, when obtaining an estimator, one needs to evaluate integrals that do not have an analytic solution. Then, one possibility is to simulate these integrals. For example, the likelihood for observing a particular outcome y_i given x_i may depend on the realization β_i of a random coefficient. That is, it is $f(y_i|x_i,\beta_i;\theta)$ instead of $f(y_i|x_i,\beta;\theta)$. However, β_i is not observed and therefore, we need to integrate over the distribution of β_i to obtain the likelihood to then later estimate moments of this distribution. Assume that the random coefficient is normally distributed. Then, the likelihood contribution we will use in the estimation step is

$$f(y_i|x_i;\theta,\mu_{\beta_i},\sigma^2_{\beta_i}) = \int f(y_i|x_i,\beta_i;\theta)\, F_{\beta_i;\mu_{\beta_i},\sigma^2_{\beta_i}}(d\beta_i). \tag{4.6.1}$$

This integral can be simulated using draws from a standard normal distribution, or Halton sequences, or other methods, and is then used to obtain simulated maximum likelihood estimates of the unknown parameters, including the mean and the variance of the random coefficient. The logic behind the simulated GMM is the same. Sometimes, empirical moment conditions need to be simulated, but then one can proceed as if the underlying integral was known. Train (2003) discusses this in great detail and the interested reader is referred to that book, or the even more comprehensive treatment in Gourieroux and Monfort (1996). Early contributions to the literature include McFadden (1989), Pakes and Pollard (1989), and Duffie and Singleton (1993). These also contain asymptotic results. The general idea is that when we let the number of simulation draws go to infinity, then the simulation error will vanish and we can treat the integral as known.

Returning to the example with a normally distributed random coefficient, if one uses draws from a standard normal distribution for this, one would add μ_β to them and multiply them by σ_β to obtain draws from $F_{\beta_i}(\beta_i;\mu_\beta,\sigma_\beta)$. If instead one uses draws from the uniform distribution or a Halton sequence, or the like, one applies the inverse

standard normal distribution $F_{\beta_i}(\beta_i;0,1)^{-1}$ and then adds μ_β and multiplies by σ_β to obtain draws from $F_{\beta_i}(\beta_i;\mu_\beta,\sigma_\beta)$. Or one applies directly $F_{\beta_i}(\beta_i;\mu_\beta,\sigma_\beta)^{-1}$.

This idea is also useful in the multivariate case, although there, one has to transform random draws using the Cholesky factors. Define the Cholesky factors L to satisfy $var(\beta_i) = P'P$. Then, we can draw a K-vector u_i that is of the same length as β_i from a multivariate uniform distribution (with independent elements), post-multiply them by P and add μ_β to them.

An important practical matter is that it is important, when iterating to numerically maximize the likelihood function, to use the same draws in each iteration. Otherwise, the solver will either not convergence at all, or when the convergence criterion is not strict, it will converge to the wrong values. We will look more into this in the ordered probit example in Section 5.2.5.

Another important practical point is that one has to be careful when calculating the average simulated log likelihood when there are multiple observations per individual. In (4.6.1) we have integrated over the distribution of the random coefficient. The log thereof is

$$\ell_i\left(\theta,\mu_{\beta_i},\sigma^2_{\beta_i}\right) = \log\left(\int f(y_i|x_i,\beta_i;\theta)\, F_{\beta_i;\mu_{\beta_i},\sigma^2_{\beta_i}}(d\beta_i)\right).$$

This means that in practice we first have to calculate the likelihood, then average it over simulation draws and then take the log. If y_i is a vector with a sequence of choices, as in (4.2.2), then we first have calculate the likelihood to observe the particular sequence of choices for one draw of the random coefficient—$f(y_i|x_i,\beta_i;\theta)$ becomes $\prod_{t=1}^{T} f(y_{it}|x_i,\beta_i;\theta)$ and x_{it} contains values of the covariates in all periods—and then average over simulation draws to obtain an approximation to

$$\ell_i\left(\theta,\mu_{\beta_i},\sigma^2_{\beta_i}\right) = \log\left(\int\left(\prod_{t=1}^{T} f(y_{it}|x_i,\beta_i;\theta)\right) F_{\beta_i;\mu_{\beta_i},\sigma^2_{\beta_i}}(d\beta_i)\right).$$

The danger is to calculate instead first the average likelihood within each period, then take the log and then sum across periods. So, here, one has to be extremely careful when programming the estimator.

4.7 Indirect Inference

Sometimes, especially when it comes to complex structural models, it is difficult to estimate the unknown parameters directly. An alternative to doing so is to use the method of indirect inference. Here, I follow Gourieroux et al. (1993).

The idea is that we can simulate, for given parameters θ, data from the model

$$y_{it} = m(x_{it}, \varepsilon_{it}; \theta).$$

This can be any model, for example a structural life cycle model in which y_{it} is consumption, x_{it} is a vector of state variables including i's age at t, but also exogenous state variables, and ε_{it} is a vector of taste shocks for which we know the distribution, up to a finite set of parameters that we include in θ. The remaining parameters in θ are, for example, preference parameters. For a given θ, we can simulate choices y_{it}^s, $s = 1, \ldots, S$. Our data are the actual choices individuals i made in periods $t = 1, \ldots, T$, given the observed x_{it}.

Denote the unknown, true distribution of y_{it} given x_{it} by $F_{y_{it}|x_{it};\theta_0}$. This distribution depends on the unknown parameter θ_0 and even if we knew θ_0, it may not be possible to derive a closed form of this distribution. But we can simulate data for any given θ, also for θ_0. Denote the distribution of the simulated data in simulation round s with NT observations, by $\hat{F}^s_{y_{it}|x_{it};\theta}$. We denote this function with a hat because it is the "empirical" distribution that is based on a finite number of observations.

Then, there is an auxiliary model, with an auxiliary parameter β. If y_{it} is a binary choice, then the auxiliary model could for example be a parametric binary choice model. Or any other model for which we can find a criterion function Q such that the so-called binding function that was introduced by Gourieroux and Montfort (1995),

$$b(\theta) = \arg\max_{\beta} Q\left(\beta, F_{y_{it}|x_{it};\theta}\right),$$

can be defined. Here, $F_{y_{it}|x_{it};\theta}$ is the true distribution of y_{it} given the exogenous state variables in x_{it} (such as the interest rate), but not the endogenously determined ones (such as wealth), for the parameter value θ. Q could be a likelihood function or the negative of the GMM objective function.[15] Assume that Q has a unique maximizer for

[15] It would have to be the negative of the GMM objective function because usually we minimize the GMM objective function.

any given θ.

Define β_0 to be the value of the auxiliary parameter once we evaluate the binding function at the true value of the structural parameters, θ_0,

$$\beta_0 = b(\theta_0).$$

Assume, for identification, that $b(\cdot)$ is invertible and denote its inverse by $b^{-1}(\cdot)$. Moreover, assume that the Jacobian matrix $\partial b(\theta)/\partial \theta'$ is of full column rank when evaluated at the true parameter value θ_0.

At this point, it is useful to make the following thought experiment. Suppose $b(\theta)$ was known. Then, we could use data to estimate β_0, and use the parameter estimates

$$\hat{\beta} = \arg\max_{\beta} Q\left(\beta, \hat{F}_{y_{it}|x_{it}}\right)$$

to obtain

$$\hat{\theta} = b^{-1}(\hat{\beta}).$$

However, this is generally not feasible because the binding function depends on the unknown $F_{y_{it}|x_{it};\theta_0}$. The idea of indirect inference is to instead use the distribution of the simulated data, $\hat{F}^s_{y_{it}|x_{it};\theta}$, over multiple simulation rounds s, to estimate θ.

In round s, and for each value of the structural parameters we can simulate data and then estimate the auxiliary parameters as

$$\hat{\beta}^s(\theta) = \arg\max_{\beta} Q\left(\beta, \hat{F}^s_{y_{it}|x_{it};\theta}\right).$$

Here, we implicitly use the empirical distribution function of y_{it} given x_{it} for the simulated data set of size NT. Asymptotically, that is when the number of observations in the original data, or in one simulated data set, goes to infinity, we have that $\hat{F}^s_{y_{it}|x_{it};\theta}$ goes to $F_{y_{it}|x_{it};\theta_0}$ and hence $\hat{\beta}^s(\theta) = b(\theta)$. The idea of the indirect estimator of θ is to calibrate the value of θ such that

$$\frac{1}{S}\sum_{s=1}^{S} \hat{\beta}^s(\theta)$$

is close to $\hat{\beta}$, which is the estimated parameter of the auxiliary model for the actual data.

Formally, it is the solution of a minimum distance problem and akin to the GMM estimator,

$$\hat{\theta} = \arg\min_{\theta} \left[\hat{\beta} - \frac{1}{S}\sum_{s=1}^{S} \hat{\beta}^s(\theta)\right]' \hat{\Psi}^{-1} \left[\hat{\beta} - \frac{1}{S}\sum_{s=1}^{S} \hat{\beta}^s(\theta)\right], \tag{4.7.1}$$

where $\hat{\Psi}$ is a positive definite matrix converging to a positive definite matrix Ψ. Under the aforementioned assumptions it is a consistent estimator of θ.

This shows that the estimator is obtained by evaluating $\hat{\beta}^s(\theta)$ only at specific values appearing in the optimization. It is based on simulation and therefore, as already explained in Section 4.6, the same draws of ε_{it} should be used throughout. As for the elements of x_{it} that are exogenous to the model, one always uses the ones from the original data.

For fixed S, the estimator is asymptotically normally distributed with

$$\sqrt{NT}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Lambda),$$

where

$$\Lambda = \left(1 + \frac{1}{S}\right) (D'\Psi^{-1}D)^{-1} (D'\Psi^{-1}\Phi\Psi^{-1}D) (D'\Psi^{-1}D)^{-1} \tag{4.7.2}$$

with

$$D = \left.\frac{\partial b(\theta)}{\partial \theta'}\right|_{\theta=\theta_0}$$

and

$$\Phi = J_0^{-1}(I_0 - K_0)J_0^{-1}$$

for some I_0, J_0 and K_0. $I_0 - K_0$ can consistently be estimated by

$$N \cdot \frac{1}{S}\sum_{s=1}^{S} (W_s - \bar{W})(W_s - \bar{W})'$$

with

$$W_s = \left. \frac{\partial Q\left(\hat{\beta}, \hat{F}^s_{y_{it}|x_{it};\tilde{\theta}}\right)}{\partial \beta} \right|_{\beta=\hat{\beta}}$$

$$\bar{W} = \frac{1}{S}\sum_{s=1}^{S} W_s.$$

where $\tilde{\theta}$ is a consistent estimator for θ, for instance $\hat{\theta}$. J_0 can be consistently estimated by

$$\left. \frac{\partial^2 Q\left(\beta, \hat{F}_{y_{it}|x_{it}}\right)}{\partial \beta \partial \beta'} \right|_{\beta=\hat{\beta}},$$

the Hessian of the objective function of the auxiliary model evaluated at $\hat{\beta}$.

From (4.7.2), we can see that the optimal choice of the weighting function is

$$\left(\Psi^{-1}\right)^* = J_0(I_0 - K_0)^{-1}J_0,$$

assuming that $I_0 - K_0$ is invertible. Then,

$$\begin{aligned}\Lambda^* &= \left(1+\frac{1}{S}\right)\left(D'\left(\Psi^{-1}\right)^* D\right)^{-1}\left(D'\left(\Psi^{-1}\right)^* \Phi \left(\Psi^{-1}\right)^* D\right)\left(D'\left(\Psi^{-1}\right)^* D\right)^{-1} \\ &= \left(1+\frac{1}{S}\right)\left(D'\left(\Psi^{-1}\right)^* D\right)^{-1}.\end{aligned}$$

In addition to this, Gourieroux et al. (1993) show that the estimator in (4.7.1) is equivalent to an estimator that uses $S = 1$ but simulated data sets that are proportionally bigger so that the same number of simulated observations is used. Moreover, they show that instead of (4.7.1) an asymptotically equivalent estimator is given by

$$\hat{\theta} = \arg\min_{\theta} \left[\left. \frac{\partial Q\left(\beta, \hat{F}_{y_{it}|x_{it};\theta}\right)}{\partial \beta} \right|_{\beta=\hat{\beta}}\right]' \Sigma \left[\left. \frac{\partial Q\left(\beta, \hat{F}_{y_{it}|x_{it};\theta}\right)}{\partial \beta} \right|_{\beta=\hat{\beta}}\right],$$

where the optimal value of Σ is given by $J_0^{-1}\left(\Psi^*\right)^{-1}J_0^{-1} = (I_0 - K_0)^{-1}$. This goes back to the estimator proposed by Gallant and Tauchen (1996) who point out that it

necessitates only one optimization in θ if the gradient of the objective function of the auxiliary model has a closed form. Moreover, they provide results relating the indirect inference estimator to the simulated GMM estimator of Duffie and Singleton (1993), as well as specification tests and a set of examples. One of these is the approximation of a multidimensional normal distribution, which is useful for multinomial choice models.

4.8 Hypothesis Testing and Multiple Comparisons

Suppose we are interested in the question whether the data favor or disfavor a particular description of nature. Specifically, we are concerned with one particular *null hypothesis*. If the data fall into a particular region of the sample space called the *critical region* then the test is said to *reject* the null hypothesis, otherwise it *accepts*.

We make a *type 1 error* if the null hypothesis is falsely rejected. Conversely, a *type 2 error* is made if the null hypothesis is incorrectly accepted. For any test we call α the *size* of the test which is the probability of a Type 1 error. If we denote the probability of a Type 2 error by β the probability of rejecting the null when it is false is given by $1-\beta$. This probability is called the *power* of a test. A test is said to be *best* if it has the maximum power among all tests with size less than or equal to some particular value. Finally, a test is said to be *consistent* if it always rejects the null when it is false provided that the number of observations goes to infinity.

To choose among tests we can examine the rate at which the power function approaches its limiting value. The most common limiting argument is to consider the power of the test to distinguish *local* alternatives for tests of fixed size.

The last topic I would like to briefly mention is the one of multiple comparisons. Suppose one is interested in estimating β in

$$y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i.$$

Then, if the significance level is 5 percent and the true value of β is equal to zero, one will falsely reject the null hypothesis of a zero coefficient in 5 percent of the cases. This is a type 1 error.[16]

[16] Conversely, if the true value of β is different from zero, a type 2 error occurs.

This logic usually underlies statistical testing. However, it ignores the fact that researchers often experiment with different specifications, which means that they change the set of variables that enter z_i. If one does this often enough, then the chance to falsely reject the null hypothesis will increase. This is the problem of multiple comparisons. Strictly speaking, standard errors need to be adjusted for this, but one hardly ever sees this in practice. Tukey (1991) discusses this in more detail. Newson et al. (2003) show how one could adjust for multiple comparisons in Stata and provide additional references.

Exercises

1. Firms i produce quantities y_i according to the production function

$$y_i = k_i^{\alpha} l_i^{\beta} \varepsilon_i, \tag{4.8.1}$$

where $\alpha + \beta < 1$ and $\alpha, \beta > 0$. Our data set contains information on two inputs, capital k_i and labor l_i. Input prices are equal for all firms and given by p_K for capital and p_L for labor. The equilibrium price on the product market is p and there are infinitely many firms so that they take this price as given. ε_i is independently identically distributed across firms and $\mathbb{E}[\ln \varepsilon_i] = 0$. ε_i is known to firm i. Firms maximize profits.
One can show that i's optimal inputs are given by

$$k_i = \left[\frac{p_K}{p \cdot \alpha \cdot \varepsilon_i} \left(\frac{p_L \cdot \alpha}{p_K \cdot \beta} \right)^{\beta} \right]^{\frac{1}{\alpha+\beta-1}}$$
$$l_i = \left[\frac{p_L}{p \cdot \beta \cdot \varepsilon_i} \left(\frac{p_K \cdot \beta}{p_L \cdot \alpha} \right)^{\alpha} \right]^{\frac{1}{\alpha+\beta-1}}.$$

Transform the model in (4.8.1) so that it is linear in the parameters.

   (a) Is it possible to estimate the parameters α and β by ordinary least squares? Explain!

(b) Suppose that ε_i is independent of p_L and p_K. Are these input price appropriate instruments for input quantities? Explain!

(c) Now suppose we observe firm specific input prices p_{Ki} and p_{Li}, which are independent of ε_i. Are p_{Ki} and p_{Li} appropriate instruments for capital and labor? Explain!

2. Show that $Z'X = \sum_{i=1}^{N} z_i' x_i$ and $Z'y = \sum_{i=1}^{N} z_i' y_i$.

Part II

Econometric Models

Chapter 5

Discrete Choice

5.1 Binary Choice

5.1.1 General Model

The simplest discrete choice model is the binary choice model. It is for an outcome y_i that takes on the value 0 or 1. Labeling the two possible outcomes in this way is just a convention that turns out to be convenient. In general, any two numbers would do the job at some notational cost. The model relates the outcome to a $1 \times K$ vector of covariates x_i and an error term ε_i. Throughout, and also in later sections, we will assume that x_i is exogenous in the sense that x_i and ε_i are fully independent from one another. The model is then estimated using a linear regression or maximum likelihood. In either case the distribution of ε_i is assumed to be known.

The model is

$$y_i = 1\{x_i\beta \geq \varepsilon_i\}, \tag{5.1.1}$$

where $1\{\cdot\}$ is the indicator function that takes on the value one if the argument is true and zero otherwise. β is the unknown parameter vector.

The model is restrictive in that it assumes that there is only a scalar error term that does not interact with x_i. This can be relaxed. For example, one can allow β to be a random coefficient β_i. We will discuss this in Section 5.1.16. Later, we will see that

the binary choice model is a special case of both, the ordered choice model and the multinomial choice model. In Section 5.2.5, we will look at the more general (in that respect) ordered probit model with a random coefficient and in Section Section 5.3.9, we will look at the mixed multinomial logit model.

A second restriction is that the so-called *latent index*, $x_i\beta$, is linear.[1] In many places we can think of this index as being replaced by a more general nonparametric function $p(x_i)$. Commonly, however, the model is not formulated in that way because most applied researcher would typically use the linear specification. This specification is still very flexible as we can think of x_i as containing basis functions, for example for a polynomial, in which case we could interpret $x_i\beta$ as an approximation to $p(x_i)$.

There are three popular distributional assumptions for ε_i that each define a particular binary choice model. For the *probit model* it is assumed that ε_i is standard normally distributed. The *logit model* assumes that ε_i follows the logistic distribution. Finally, the *linear probability model* assumes that it is uniformly distributed. Below we discuss these assumptions and their implications from a practical perspective. Before that we look at properties of the general model and its random utility foundation.

5.1.2 Properties of the General Model

For now we assume that the distribution of ε_i is known. Denote the cumulative distribution function (c.d.f.) of ε_i by F_{ε_i}.[2]

For binary choice models we are interested in the dependence of the probability to observe a particular y_i on x_i. For the choice probability we have that

$$\Pr(y_i = 1|x_i) = \Pr(1\{x_i\beta \geq \varepsilon_i\}|x_i) = \Pr(\varepsilon_i \leq x_i\beta|x_i) = \Pr(\varepsilon_i \leq x_i\beta) = F_{\varepsilon_i}(x_i\beta). \tag{5.1.2}$$

[1] Here, we call $x_i\beta$ the latent index. Another possibility is to call $y_i^* \equiv x_i\beta - \varepsilon_i$ the latent index and formulate the model as $y_i = 1\{y_i^* \geq 0\}$. Clearly, this is the same model as the one in (5.1.1). Yet another possibility is to define the index as $y_i^* \equiv x_i\beta + \varepsilon_i$ and again formulate the model as $y_i = 1\{y_i^* \geq 0\}$. If the distribution of ε_i is symmetric about 0, which one usually assumes in practice, then the last two ways of defining the index yield the same likelihood contributions. Intuitively, this is because drawing the value e of ε_i is as likely as drawing the value $-e$. This can be shown using (5.1.3) below.

[2] For any real valued random variable the c.d.f. is given by

$$F_{\varepsilon_i}(e) \equiv \Pr(\varepsilon_i \leq e).$$

Here the first equality is by the model in (5.1.1), the second equality is by the definition of a probability, the third is by the independence between ε_i and x_i, and the fourth is by the distributional assumption for ε_i. For $y_i = 0$ we have that

$$\Pr(y_i = 0|x_i) = 1 - \Pr(y_i = 1|x_i) = 1 - F_{\varepsilon_i}(x_i\beta).$$

If the distribution of ε_i is symmetric about zero then we have in addition that

$$1 - F_{\varepsilon_i}(x_i\beta) = F_{\varepsilon_i}(-x_i\beta). \tag{5.1.3}$$

This will be the case for the probit and the logit model.

The model in (5.1.1) is only identified up to location and scale since for any constant c_l and any positive constant c_s we have that

$$x_i\beta \gtreqless \varepsilon_i,$$

which determines whether $y_i = 0$ or $y_i = 1$, is equivalent to

$$c_l + c_s x_i\beta \gtreqless c_l + c_s\varepsilon_i.$$

That means that the model prediction for y_i will be the same for any value of c_l and any positive c_s. So, we need to impose two normalizations. These are not assumptions because they do not restrict the distribution of y_i for a given x_i. The first normalization is on the location. Typically, we impose $c_l = 0$ implicitly by saying that x_i contains a constant so that β contains the intercept while ε_i has a mean of zero. The second normalization is on the scale. It is typically imposed *via* the distributional assumption for ε_i, which imposes that the standard deviation s takes on a particular value, for example one if we assume that ε_i is standard normally distributed.

5.1.3 Random Utility Foundation

The binary choice model can be derived from a richer model that involves utility comparisons. i receives utility

$$u_{i0} = z_{i0}\alpha_0 + w_i\gamma_0 + \varepsilon_{i0}$$

if she chooses alternative $y_i = 0$ and

$$u_{i1} = z_{i1}\alpha_1 + w_i\gamma_1 + \varepsilon_{i1}$$

if she chooses $y_i = 1$.

There are two taste shocks, ε_{i0} and ε_{i1}, one for each alternative, and two types of explanatory variables. z_{i0} and z_{i1} are of the first type. These are vectors of *alternative varying* characteristics such as the respective price or characteristics that are associated with the two alternatives. So here, following Lancaster (1966), utility is defined on characteristics, not on goods. w_i is of the second type. This is a vector of *alternative invariant* characteristics. These could be characteristics of the decision maker, such as years of education or income, or characteristics of the decision situation.

In addition there are alternative specific coefficients, α_0, α_1, γ_0, and γ_1. Usually, we impose that the utility i derives from alternative varying characteristics does not depend on the actual alternative, that is $\alpha_0 = \alpha_1 = \alpha$. This is akin to assuming that all differences between two alternatives with the same characteristics are captured by the two error terms.

The model is a model of utility maximization: i chooses $y_i = 1$ if $u_{1i} \geq u_{0i}$, or equivalently

$$z_{i1}\alpha_1 + w_i\gamma_1 + \varepsilon_{i1} \geq z_{i0}\alpha_0 + w_i\gamma_0 + \varepsilon_{i0}.$$

This can be rewritten as

$$y_i = 1\{z_{i1}\alpha_1 - z_{i0}\alpha_0 + w_i(\gamma_1 - \gamma_0) \geq -(\varepsilon_{i1} - \varepsilon_{i0})\}. \tag{5.1.4}$$

or

$$y_i = 1\{x_i\beta \geq \varepsilon_i\},$$

where

$$\begin{aligned} x_i &= \begin{pmatrix} z_{i1} & z_{i0} & w_i \end{pmatrix} \\ \varepsilon_i &= -(\varepsilon_{i1} - \varepsilon_{i0}) \\ \beta &= \begin{pmatrix} \alpha_1 \\ -\alpha_0 \\ (\gamma_1 - \gamma_0) \end{pmatrix}. \end{aligned}$$

This is the binary choice model in (5.1.1) and once we specify a distribution for the error terms ε_{i0} and ε_{i1}, respectively, this implies a distribution for the negative of their difference.

Observe at this point that choices can only be informative about the sign of utility differences, which is closely related to the idea that utility is an ordinal and not a cardinal concept. So only β can be identified which in turn means that at most $(\gamma_1 - \gamma_0)$, α_1 and α_0 can be identified. Normally we impose the normalization that $\gamma_0 = 0$. Then, γ_1 can be interpreted as the difference in the coefficients on w_i, which are part of β. If $\alpha_0 = \alpha_1 = \alpha$ then this is the coefficient on the difference in the characteristics, $z_{i1} - z_{i0}$. It also enters β. We have argued before that we also need to impose an assumption on the scaling. This is normally done implicitly by making a distributional assumption for the error terms which implies a distribution for the negative of their difference.

5.1.4 Foundation by a Structural Economic Model

In the previous subsection, we have shown that a binary choice model can be motivated by utility comparisons: an individual chooses alternative 1 whenever the random utility of doing so exceeds the random utility of choosing alternative 0. If utilities are linear, then we have a binary choice models of the form $y_i = 1\{x_i\beta \geq \varepsilon_i\}$.

It is conceptually not much different to instead estimate models of the form

$$y_i = 1\{g(x_i;\beta) \geq \varepsilon_i\},$$

where $g(x_i;\beta)$ is a function of x_i that depends on a finite set of parameters collected in β. This is particularly appealing if these are implied by economic models.

An early example of such a model is the one by Wolpin (1984). The binary choice in his model is a Malaysian couple's decision to have another child. This is a dynamic decision problem, as it involves the decision between having a child today or tomorrow. In his model, starting at age 15 or marriage and ending after 30 years, in any 18 month period t, a woman may give birth with certainty. The child survives the first period with probability P_t and all subsequent periods with probability 1. The household maximizes expected life time utility. Its period utility function is

$$\begin{aligned} U_t(M_t, X_t) &= (\alpha_1 + \xi_t) \cdot M_t - \alpha_2 M_t^2 + \beta_1 X_t - \beta_2 X_t^2 \\ &\quad + \gamma_1 M_t^i X_t^i + \gamma_2 M_t^i S \end{aligned}$$

where M_t is the number of children, X_t is the level of goods consumption, S is the number of years of schooling of the mother, and ξ_t is a taste shock specific to period t, which is uncorrelated over time.

The main trade-off for a household is that having a child in a given period yields utility from having that child, but there are also fixed cost of giving birth, which are specific to the age of the mother, and "maintenance cost." Wolpin shows that the decision to have a child is determined by the sign of

$$J_t = \mathbb{E}\left[\Delta LU_t\right] + P_t \xi_t,$$

where ΔLU_t is the change in the life time utility, for $\xi_t = 0$, that is due to having another child. This depends on state variables such as the number of children in t. Hence, this is a model of the form $y_i = 1\{g(x_i;\beta) \geq \varepsilon_i\}$ with $\varepsilon_i = -P_t \xi_t^i$ and $g(x_i;\beta) = \mathbb{E}\left[\Delta LU_t\right]$. He then assumes that ξ_t follows a normal distribution with mean zero and variance σ^2. Then, the probability to have another child is

$$\Pr(n_t = 1|M_{t-1}) = \Pr\left(\mathbb{E}\left[\Delta LU_t\right] + P_t \xi_t \geq 0\right) = 1 - \Phi\left(\mathbb{E}\left[\Delta LU_t\right]/\sigma P_t\right).$$

This is remarkably similar to the choice probabilities we have derived for the standard case. This example shows that if one manages to set up a structural model that has a structure similar to the one here and to solve that model for given parameters, then estimating that structural model is not as big a step as one may first think. The advantage of doing so is that the estimated parameters may then be used to predict the responsiveness of fertility to external factors, or in general to changes in welfare policies that have not been observed before. Here, Wolpin uses maximum likelihood

5.1.5 Identification

We now show that β in the general binary choice model of Section 5.1.2 is identified. Our starting point is

$$\Pr(y_i = 1|x_i) = F_{\varepsilon_i}(x_i\beta)$$

along with the assumption that F_{ε_i} is known and continuous. Data are informative about the left hand side and we are interested in estimating β.

Continuity of the distribution of ε_i implies that F_{ε_i} is strictly increasing. Consequently, its inverse, $F_{\varepsilon_i}^{-1}(\cdot)$, exists. Hence,

$$F_{\varepsilon_i}^{-1}(\Pr(y_i = 1|x_i)) = x_i\beta.$$

Assume, as in Section 3.6.2 that there are K values of x_i such that

$$\tilde{x} \equiv \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_K \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1K} \\ 1 & x_{22} & \dots & x_{2K} \\ & \vdots & & \\ 1 & x_{K2} & \dots & x_{KK} \end{pmatrix}$$

has full rank. This implies that its inverse exists. Define

$$\mu_y(\tilde{x}) \equiv \begin{pmatrix} F_{\varepsilon_i}^{-1}(\Pr(y_i = 1|x_1)) \\ \vdots \\ F_{\varepsilon_i}^{-1}(\Pr(y_i = 1|x_K)) \end{pmatrix} = \begin{pmatrix} x_1\beta \\ \vdots \\ x_K\beta \end{pmatrix} = \tilde{x}\beta.$$

Since $\tilde{x}$ has full full rank

$$\beta = \tilde{x}^{-1}\mu_Y(\tilde{x}).$$

The similarity the proof of identification in the multivariate case that has been presented in Section (3.6.2) is striking. Again, exogenous variation in x_i is exploited to identify θ.

In fact, also here we can extend the above result and $\tilde{x}$ contains more than K values of x_i that are linearly independent. Then, it follows from $\mu_y(\tilde{x}) = \tilde{x}\beta$ that

$$\beta = \left(\tilde{x}'\tilde{x}\right)^{-1}\tilde{x}'\mu_Y(\tilde{x}).$$

5.1.6 Estimation

We have seen above that the binary choice model can be estimated by maximum likelihood since we have specified that

$$\Pr(y_i = 1|x_i) = F_{\varepsilon_i}(x_i\beta)$$

and assume that F_{ε_i} is known.

The likelihood to observe y_i given x_i is

$$f(y_i|x_i;\theta) = F_{\varepsilon_i}(x_i\beta)^{y_i} \cdot (1 - F_{\varepsilon_i}(x_i\beta))^{1-y_i},$$

where $\theta = \beta$, and the log likelihood is

$$\ell_i(\beta) = y_i \cdot \log(F_{\varepsilon_i}(x_i\beta)) + (1 - y_i) \cdot \log(1 - F_{\varepsilon_i}(x_i\beta)).$$

The sample log likelihood function is

$$\mathscr{L}(\beta) = \sum_{i=1}^{N} y_i \cdot \log(F_{\varepsilon_i}(x_i\beta)) + (1 - y_i) \cdot \log(1 - F_{\varepsilon_i}(x_i\beta)). \tag{5.1.5}$$

Alternatively, since we have that

$$\mathbb{E}[y_i|x_i] = F_{\varepsilon_i}(x_i\beta)$$

we could estimate the model

$$y_i = F_{\varepsilon_i}(x_i\beta) + u_i$$

using a nonlinear least squares estimator. Here, u_i is a residual. This approach, however, seems unattractive as u_i is heteroskedastic for the same reason as it is in the linear probability model that we discuss in Section (5.1.11) below, while maximum likelihood estimation is efficient and an assumption on the distribution on ε_i has already been made.

5.1.7 Goodness of Fit

The goodness of fit can be summarized by R^2_{McFadden}. For this the maximum likelihood estimate of $\hat{f}(y_i)$ is the sample probability to observe $y_i = 1$ or $y_i = 0$, respectively. Another simple way to assess how well the model explains the data is to predict y_i to be one if $F_{\varepsilon_i}(x_i\beta) \geq 0.5$, and zero otherwise, and then report the percentage of correctly predicted choices. This can be done separately for individuals who chose $y_i = 0$ and $y_i = 1$.

5.1.8 Parameters of Interest and Reporting of Results

In linear regression models such as

$$y_i = x_i\beta + \varepsilon_i$$

we are typically interested in β. Under the assumption that $\mathbb{E}[y_i|x_i] = 0$ we have

$$\mathbb{E}[y_i|x_i] = x_i\beta,$$

so β is the effect a unit change in x_i has on the average y_i, or, following up on the discussion of (3.6.2), it is the average effect a change in x_i has on y_i. All these effects coincide because the model is linear in β.

The binary choice model, (5.1.1), however, is nonlinear. y_i changes only at the point at which $x_i\beta$ crosses ε, so most of the time the effect of x_i on y_i is zero. Therefore, it is useful to look at the effect x_i has on the choice probability. This is the so-called vector of *marginal effects*

(5.1.6)
$$\frac{\partial \Pr(y_i = 1|x_i)}{\partial x_i} = \frac{\partial F_{\varepsilon_i}(x_i\beta)}{\partial x_i} = \frac{\partial F_{\varepsilon_i}(x_i\beta)}{\partial x_i\beta} \cdot \frac{\partial x_i\beta}{\partial x_i} = \frac{\partial F_{\varepsilon_i}(x_i\beta)}{\partial x_i\beta} \cdot \beta = f_{\varepsilon_i}(x_i\beta) \cdot \beta.$$

The derivative of $F_{\varepsilon_i}(x_i\beta)$ with respect to $x_i\beta$ is the density $f_{\varepsilon_i}(x_i\beta)$ of ε_i. The marginal effects vary across individuals because $f_{\varepsilon_i}(x_i\beta)$ varies across individuals. In practice, either marginal effects are calculated for every individual in the sample and then the average is reported or they are reported for an individual with average characteristics.

An important property of the binary choice model is that the sign of the marginal effects is always equal to the sign of the coefficients. This follows from (5.1.6) because $f_{\varepsilon_i}(x_i\beta)$ is always positive. So if one is interested in the signs of the effects, then it suffices to look at the coefficients.

Alternatively, ratios of coefficients can be reported as they are equal to ratios of marginal effects. For this a baseline covariate x_{i1} has to be chosen and then we have for any other covariate x_{ik}

$$\frac{\partial \Pr(y_i = 1|x_i)/\partial x_{ik}}{\partial \Pr(y_i = 1|x_i)/\partial x_{i1}} = \frac{\partial F_{\varepsilon_i}(x_i\beta)/\partial(x_i\beta) \cdot \partial(x_i\beta)/\partial x_{ik}}{\partial F_{\varepsilon_i}(x_i\beta)/\partial(x_i\beta) \cdot \partial(x_i\beta)/\partial x_{i1}} = \frac{\beta_k}{\beta_1}.$$

Finally, it is worth noting that Ai and Norton (2003) point out that when interaction effects between two variables are included into x_i the marginal effect of them interacting is not the marginal effect of the interaction term.[3] To see this, let there be additional

[3] In the same context, Puhani (2012) showed that the interaction term should still be at the center of attention in the context of differences-in-differences estimation that we will discuss in Chapter (7).

variables x_{1i} and x_{2i} and first consider the linear case with

$$y_i = x_i\beta + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12} + \varepsilon_i.$$

Here, the average interaction effect is

$$\frac{\partial^2 \mathbb{E}[y_i|x_{1i},x_{2i},x_i]}{\partial x_{1i}\partial x_{2i}} = \beta_{12}.$$

In the case of the binary choice model, we have

$$\Pr(y_i = 1|x_{1i},x_{2i},x_i) = 1\{x_i\beta + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12} \geq \varepsilon_i\}$$

and

$$\begin{aligned}\frac{\partial^2 \mathbb{E}[y_i|x_{1i},x_{2i},x_i]}{\partial x_{1i}\partial x_{2i}} =& f_{\varepsilon_i}(x_i\beta + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12}) \cdot \beta_{12} \\ &+ f'_{\varepsilon_i}(x_i\beta + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{1i}x_{2i}\beta_{12}) \cdot (\beta_1 + x_{2i}\beta_{12}) \cdot (\beta_2 + x_{1i}\beta_{12}).\end{aligned}$$

5.1.9 Probit Model

For the probit model we assume that ε_i is distributed according to the standard normal distribution

$$\Phi(\varepsilon_i) \equiv \int_{-\infty}^{\varepsilon_i} \phi(z)\, dz,$$

where

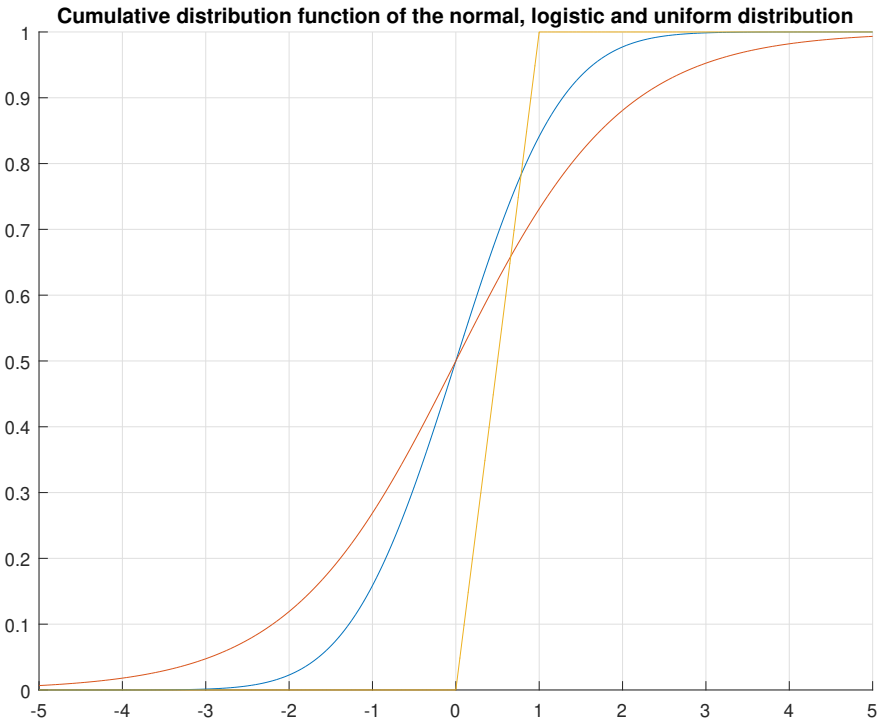
$$\phi(z) = (1/\sqrt{2\pi}) \cdot \exp(-z^2/2)$$

is the standard normal density function.[4]Figure (5.1.9) shows the c.d.f. Then, the choice probability (5.1.2) becomes

$$\Pr(y_i = 1|x_i) = \Phi(x_i\beta). \tag{5.1.7}$$

To derive this model from a random utility model assume that ε_{i0} and ε_{i1} are both normally distributed with mean zero and variance 0.5. This implies that the difference, $\varepsilon_i = -(\varepsilon_{i1} - \varepsilon_{i0})$, is also normally distributed and has mean zero and variance 1.

[4]There is no analytic solution to this integral and therefore it needs to be calculated using simulation. This, however, is automatically performed by all statistical packages.



The figure shows the cumulative distribution function of the normal distribution (green), standard normal distribution (blue) and uniform distribution (red).

Figure 5.1.1: Standard normal, logistic and uniform distribution

By using the standard normal distribution we impose the normalizations that ε_i has a mean of 0 and a standard deviation of 1. In Section 5.1.2, we have shown that such normalizations have to be imposed. To see this, suppose that we would instead not fix the mean of ε_i at 0 (or any other value), but would leave it as a free parameter μ, which we then try to estimate. Furthermore, suppose that we tried to estimate the standard deviation σ of ε_i as well, instead of normalizing it to 1. Then, we could—in the special case of the normal distribution—still express the choice probability using the standard

normal c.d.f., but now

$$\Pr(y_i = 1|x_i) = \Phi\left(\frac{x_i\tilde{\beta} - \mu}{\sigma}\right). \tag{5.1.8}$$

Here, we write $\tilde{\beta}$ instead of β because it is related to μ and σ. This has to be the case because the parameters are always defined relative to the normalizations one imposes. In particular, it holds that

$$x_i\beta = \frac{x_i\tilde{\beta} - \mu}{\sigma} \tag{5.1.9}$$

so that for all parameters β_k and $\tilde{\beta}_k$ except for the intercept it holds that $\beta_k = \tilde{\beta}_k/\sigma$. As before, x_i contains a constant term so that $\tilde{\beta}$ contains the intercept. But of course only the difference between the intercept and μ are identified because in (5.1.8) we can always change both the intercept and μ by any number. Hence, we can—without imposing any further restrictions on the distribution of y_i given x_i—set μ to 0. Having done that, we see that we can always multiply $\tilde{\beta}$ by any positive number, if we multiply σ at the same time—because this will cancel out in (5.1.8).

But unlike those normalizations, the distributional assumption of course imposes restrictions on the data generating process and has implications for the interpretation of coefficient estimates. We have established in (5.1.6) that the vector of marginal effects is

$$\frac{\partial \Pr(y_i = 1|x_i)}{\partial x_i} = f_{\varepsilon_i}(x_i\beta) \cdot \beta.$$

For the standard normal distribution we have that the mode is at 0 so that

$$f_{\varepsilon_i}(x_i\beta) \leq \phi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4$$

and consequently marginal effects are no bigger than 0.4β. The real structure normality imposes, however, is how the marginal effect of a change in one component of x_i varies with $x_i\beta$, or the choice probability. This will always differ across models, but typically only slightly so between the probit and the logit model.

Turning to estimating the model by maximum likelihood, we next derive the score and the Hessian of the log likelihood contribution. This provides some intuition how the

maximum likelihood estimator actually estimates β. The log likelihood for observation i is

$$\ell_i(\beta) = y_i \log \Phi(x_i\beta) + (1-y_i)\log(1-\Phi(x_i\beta)).$$

The score is given by the derivative with respect to β,

$$\begin{aligned}
s_i(\beta) &\equiv \frac{\partial}{\partial \beta}\ell_i(\beta) \\
&= \left(y_i \frac{\phi(x_i\beta)x_i}{\Phi(x_i\beta)} + (1-y_i)\frac{-\phi(x_i\beta)x_i}{1-\Phi(x_i\beta}\right)' \\
&= \frac{y_i\phi(x_i\beta)x_i'(1-\Phi(x_i\beta)) - (1-y_i)\phi(x_i\beta)x_i'\Phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))} \\
&= \frac{y_i\phi(x_i\beta)x_i' - \phi(x_i\beta)x_i'\Phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))} \\
&= \frac{\phi(x_i\beta)x_i'(y_i-\Phi(x_i\beta))}{\Phi(x_i\beta)(1-\Phi(x_i\beta))}.
\end{aligned}$$

Define the residual as $u_i \equiv y_i - \Phi(x_i\beta)$. Then, the maximum likelihood condition that the average score is equal to zero,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))}x_i'u_i\right) = 0,$$

can be interpreted as the empirical analog of a weighted orthogonality condition between the residuals and the explanatory variables. x_i and β are both K-dimensional, so this is a system of K equations with K unknowns and we can solve for β, very much in the spirit of Section 5.1.5. In finite samples the weighting ensures efficiency of the estimator. It exploits our knowledge on the distribution of ε_i.

By the distributional assumption we have that $\mathbb{E}[u_i|x_i] = 0$ because $\mathbb{E}[y_i|x_i] = \Pr(y_i = 1|x_i] = \Phi(x_i\beta)$. Therefore, we can verify that the score identity holds in this particular case:

$$\mathbb{E}[s_i|x_i] = \mathbb{E}\left[\left.\frac{\phi(x_i\beta)x_i'u_i}{\Phi(x_i\beta)(1-\Phi(x_i\beta))}\right| x_i\right] = 0.$$

Next, we calculate the Hessian by taking the derivative of the score with respect to β. This yields

$$H_i(\beta) = \frac{\left(\phi'(x_i\beta)x_i'x_i(y_i - \Phi(x_i\beta)) - \phi(x_i\beta)^2 x_i'x_i\right)\Phi(x_i\beta)(1-\Phi(x_i\beta))}{(\Phi(x_i\beta)(1-\Phi(x_i\beta)))^2} - \frac{\phi(x_i\beta)x_i'(y_i - \Phi(x_i\beta))\left(\phi(x_i\beta)x_i(1-\Phi(x_i\beta)) - \Phi(x_i\beta)\phi(x_i\beta)x_i\right)}{(\Phi(x_i\beta)(1-\Phi(x_i\beta)))^2}$$

where $\phi'(\cdot)$ is the derivative of the standard normal density. This can be written as

$$-\frac{\phi(x_i\beta)^2 x_i'x_i}{\Phi(x_i\beta)(1-\Phi(x_i\beta))} + u_i L(x_i\beta)$$

where $L(\cdot)$ is a function of $x_i\beta$ only. From this we get

$$\begin{aligned} A_0 &\equiv -\mathbb{E}[H_i(\beta_0)] \\ &= -\mathbb{E}\left[\mathbb{E}[H_i(\beta_0)|x_i]\right] \\ &= \mathbb{E}\left[\frac{\phi(x_i\beta_0)^2 x_i'x_i}{\Phi(x_i\beta_0)(1-\Phi(x_i\beta_0))}\right] - \mathbb{E}\left[\mathbb{E}[u_i L(x_i\beta_0)|x_i]\right] \\ &= \mathbb{E}\left[\frac{\phi(x_i\beta_0)^2 x_i'x_i}{\Phi(x_i\beta_0)(1-\Phi(x_i\beta_0))}\right] \end{aligned}$$

where the last equality uses $\mathbb{E}\left[\mathbb{E}[u_i L(x_i\beta_0)|x_i]\right] = \mathbb{E}\left[\mathbb{E}[u_i|x_i]L(x_i\beta_0)\right] = 0$ which follows from $\mathbb{E}[u_i|x_i] = 0$. Finally, notice that

$$\frac{\phi(x_i\beta_0)^2}{\Phi(x_i\beta_0)(1-\Phi(x_i\beta_0))} > 0$$

so that the sample average of the Hessian evaluated at β_0 is almost surely negative definite in large samples so that the maximization problem is well defined.

5.1.10 Logit Model

The distributional assumption for the logit model is that ε_i follows the logistic distribution

$$\Lambda(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{1+\exp(\varepsilon_i)},$$

which has a mean of zero, is symmetric (about zero), and has a variance of $\pi^2/3 \approx 3.2899$. This is also apparent in Figure (5.1.9). Using this, we have

$$\Pr(y_i = 1|x_i) = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$$

and the log likelihood contribution is[5]

$$\log\left(y_i \cdot \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} + (1-y_i)\cdot\left(1-\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)\right).$$

The score contribution is the derivative thereof with respect to β,

$$y_i \cdot \left(1-\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)\cdot x_i - (1-y_i)\cdot\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\cdot x_i.$$

The logit model is implied by a random utility model if we assume that both error terms are distributed according to the the type 1 extreme value distribution. This distribution is, for example for ε_{i0},

$$F_{\varepsilon_{i0}}(\varepsilon_{i0}) = \exp(-\exp(-\varepsilon_{i0}))$$

and has the density function

$$f_{\varepsilon_{i0}}(\varepsilon_{i0}) = \exp(-\varepsilon_{i0})\exp(-\exp(-\varepsilon_{i0})).$$

It is also referred to as the log Weibull distribution and its mean is given by Euler's constant, $\gamma \approx 0.5772\ldots$. Figure 5.1.10 shows that this distribution is not symmetric. However, the difference between two type 1 extreme value random variables follows the logistic distribution, which, as we have pointed out above, is symmetric about zero.

[5] This is equal to

$$y_i \cdot \log\left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right) + (1-y_i)\cdot\log\left(\left(1-\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)\right)$$

because here we are in the special case in which y_i is either zero or one.

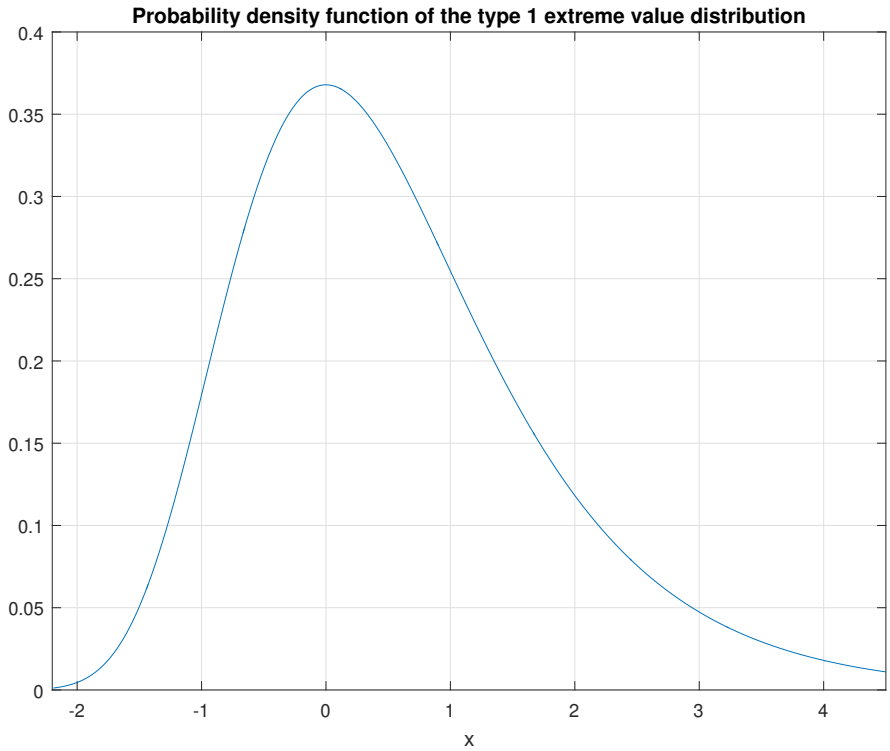


Figure 5.1.2: Type 1 extreme value distribution

The vector of marginal effects of changes in x_i is

$$\begin{aligned}\frac{\partial \Pr(y_i = 1|x_i)}{\partial x_i} &= \frac{\exp(x_i\beta)\cdot\beta\cdot(1+\exp(x_i\beta)) - \exp(x_i\beta)\cdot\exp(x_i\beta)\cdot\beta}{(1+\exp(x_i\beta))^2} \\ &= \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\cdot\left(1-\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}\right)\cdot\beta \\ &= \Pr(y_i = 1|x_i)\cdot(1-\Pr(y_i = 1|x_i))\cdot\beta.\end{aligned}$$

This is a well known formula that we derive for the more general multinomial logit

model in Section 5.3.6.

Like for the probit model the distributional assumption implies that the vector of marginal effects can be bounded. From the formula for the marginal effect of changes in x_i it follows that it is never bigger than $0.5 \cdot (1-0.5) = 0.25$ times β. Moreover, if we know, for example, that the probability lies between 0.3 and 0.7 then the marginal effects are in addition no smaller than 0.21 times β.

One can also interpret the slope coefficients in β as the log odds ratios. This is based on the observation that

$$\log\left(\frac{\Pr(y_i=1|x_i)}{\Pr(y_i=0|x_i)}\right) = \log\left(\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} \Big/ \frac{1}{1+\exp(x_i\beta)}\right) = x_i\beta,$$

so that

$$\partial \log\left(\frac{\Pr(y_i=1|x_i)}{\Pr(y_i=0|x_i)}\right) \Big/ \partial x_i = \beta.$$

Here,

$$\frac{\Pr(y_i=1|x_i)}{\Pr(y_i=0|x_i)}$$

is the odds ratio, indicating "how many times more likely it is that we observe $y_i = 1$ instead of $y_i = 0$". The coefficients in β are therefore approximately the percentage changes in the odds ratio that is associated with a unit change in the components of x_i, respectively. Norton (2012) shows in a Monte Carlo study that odds ratios are sensitive to the choice of specification and the way unobserved heterogeneity is incorporated into a model, while marginal effects seem to be affected less by this.

5.1.11 Linear Probability Model

The third popular binary choice model is the linear probability model where we assume that $F(\varepsilon_i) = \varepsilon_i$, as depicted in Figure (5.1.9), so that

$$\Pr(y_i=1|x_i) = F(x_i\beta) = x_i\beta.$$

This directly reveals a property of this model that some believe is unattractive, namely that predicted probabilities may lie outside the unit interval $[0,1]$. This can however

be avoided by including enough higher order terms in x_i. In Section 5.1.14 I therefore argue that the model is still useful to estimate marginal effects.

In this model $\mathbb{E}[y_i|x_i] = x_i\beta$, so the vector of marginal effects in this model is given by β which can be consistently estimated by regressing y_i on x_i using the OLS estimator. Define the regression residual, which is not to be confused with ε_i, as $e_i \equiv y_i - x_i\beta$. The OLS estimator is not efficient because this residual is heteroskedastic. In particular,

$$\begin{aligned} var(e_i|x_i) &= var(y_i - x_i\beta|x_i) \\ &= var(y_i|x_i) \\ &= \Pr(y_i = 1|x_i) \cdot (1 - \Pr(y_i = 1|x_i)) \\ &= x_i\beta \cdot (1 - x_i\beta) \end{aligned}$$

and this depends on x_i. Here, the third equality holds because y_i is a Bernoulli random variable.[6] This shows that in practice, one should at least use heteroskedasticity consistent standard errors. To obtain an efficient estimator we can either use maximum likelihood estimation instead, or proceed in three steps to obtain the feasible generalized least squares estimator that we have introduced in Section (4.2.1). In the first step, we estimate β inefficiently but consistently. Then, we calculate a weight

$$\hat{w}_i = 1 \Big/ \sqrt{x_i\hat{\beta} \cdot (1 - x_i\hat{\beta})}\,,$$

which is an estimate of the inverse of the standard deviation of e_i. And third, we regress $\hat{w}_i y_i$ on $\hat{w}_i x_i$. This yields efficient estimates of β because asymptotically the variance of the residual in the last regression, $\hat{w}_i e_i = \hat{w}_i y_i - \hat{w}_i x_i\beta$, is equal to 1 since $\hat{w}_i$ converges to

$$w_i = \frac{1}{\sqrt{x_i\beta \cdot (1 - x_i\beta)}}$$

so that $var(\hat{w}_i e_i)$ converges to

$$var(w_i e_i) = w_i^2 var(e_i) = \frac{1}{x_i\beta \cdot (1 - x_i\beta)} \cdot x_i\beta \cdot (1 - x_i\beta) = 1.$$

[6] This is a random variable that takes on only two values, one with probability p. It always has variance $p \cdot (1 - p)$.

5.1.12 Monte Carlo

The idea of a Monte Carlo study is that one specifies a data generating process and then estimates the unknown parameters many times so that one can assess how an estimator performs by looking, for example, at the mean and the variance of the estimates across simulated data sets.

We now go through the Matlab code of a Monte Carlo study for the logit model. Exercise 4 will follow up on this. We discuss the code in a bit more detail, as it is the first time we look at computer code. We start with the function `nll_logit.m`, which calculates the negative of the average log likelihood, `nll`, from a logit model for a given value of the parameter value, `beta`, an $N \times 1$ vector of choices, `y`, one for each observation, and an $N \times K$ matrix of explanatory variables, X. The function also calculates the negative of the average score, `ns`, which we will use to speed up the numerical optimization procedure. A function in Matlab can be called by the main program or by other functions. We now present the code.

```
1 function [nll ns] = nll_logit(beta,y,X)
  % negative average log likelihood and score for logit model
3
  prob1=exp(X*beta)./(1+exp(X*beta)); %probability to choose y=1
5 l=log(y.*prob1+(1-y).*(1-prob1)); %likelihood
  s= (y.*(1-prob1).*X-(1-y).*prob1.*X); %score
7
  nll=-mean(l); %negative of the average log likelihood
9 ns=-mean(s); %negative of the average score
```

The first line defines the output elements `nll` and `ns`, the name of the function, `nll_logit.m`, and the function arguments, `beta`, `y`, and `X`. In the fourth line, the $N \times 1$ vector of choice probabilities is calculated. `exp(X*beta)` is the exponential function applied to the N-vector `X*beta`, that we get by multiplying the $N \times K$ matrix `X` with the K-vector `beta`. Then, we divide the result element-wise by 1 plus the same expression. Element-wise division is achieved by using `./` and Matlab recognizes that the `1` in `1+exp(X*beta)` is the ones-vector of the same dimension as `exp(X*beta)`. In a similar fashion we then calculate the log likelihood contribution in line 5 and the score in line 6. Finally, we calculate the negative of the average log likelihood and score, respectively. This function is used in the Monte Carlo. Next, we take a look at the main program

that we will run.

```
clear all

% parameters for data generating process
N=100;
beta=-0.1;

% parameters for optimization
startvalues = 0;
options = optimset('Display','off','GradObj','on');

% parameters and initialization for Monte Carlo
repetitions = 1000;
betahat = NaN(repetitions,1);
nll = NaN(repetitions,1);
ns = NaN(repetitions,1);
nH = NaN(repetitions,1);

% Monte Carlo
for i = 1:repetitions
    % generate data
    x=chi2rnd(10,N,1); %years of education
    epsilon0=-evrnd(0,1,N,1);
    epsilon1=-evrnd(0,1,N,1);
    epsilon=epsilon0-epsilon1; %difference between 2 type 1
        extreme value variables follows logistic distribution
    y=beta*x>epsilon;
    objfun = @(b)nll_logit(b(1),y,x); %define objective
        function with scalar b as argument
    [betahat(i),nll(i),~,~,ns(i),nH(i)] = fminunc(objfun,
        startvalues,options); %minimization of minus the average
         log likelihood
end
```

In our Monte Carlo, the number of years of education has a negative impact on the probability that an individual has $y_i = 1$. First, in line 4, we say that there are 100 individuals and that the coefficient on x_i, `beta`, is equal to -0.1. We will use numerical optimization techniques built into Matlab and parameters for this are specified in line 8 and 9. These consist of the starting value in the variable `startvalues`. Here, this is

just a scalar, but in general, Matlab can also optimize over more than 1 variable. Then, in line 9, we specify that no output is shown as we will run the procedure many times (`'Display','off'`; you can change this to `'Display','iter'`) and that we supply the gradient of the objective function. This will be the negative of the average score, as the objective function is the negative of the average log likelihood. This has to be the second output to the objective function, which will be `nll_logit.m`.

In line 12 we put how many times we want to create data anew. Then, in the following four lines we initialize the vectors in which we will save the results. We will always save the coefficient estimate in a new line of the vector `betahat`, as well as the negative average log likelihood, score and numerical Hessian in the respective rows of the vectors `nll`, `ns`, and `nH`.

The actual Monte Carlo is carried out by looping over lines 20 till 28. In line 21, we draw x_i from a χ^2 distribution with 10 degrees of freedom, then we draw vectors consisting of ε_{0i} and ε_{1i} from a type 1 extreme value distribution, and then code that an individual smokes if β times x_i is greater or equal to $\varepsilon_i = \varepsilon_{0i} - \varepsilon_{1i}$. For the last step notice that we do not need to put an indicator function, as Matlab does this automatically when we use the logical operator `>`.

In line 26, we define the objective function, `objfun`. Here, `b` is the parameter we will optimize over and `smoking` and `yedu` will be taken as given. The actual optimization is carried out in line 27, and the results are saved in the vectors between square brackets, always in row `i`. When we run this program in Matlab, then we find that across replications, the mean is about -0.1, which means that the estimator is consistent. The variance across replications is about 0.0006.

5.1.13 Choice-Based Sampling

We now briefly discuss choice-based sampling in the context of the logit model because it turns out that the logit model can still be used in this situation, unlike other models. We face choice-based sampling if there is oversampling of observations with $y_i = 0$ or $y_i = 1$. This arises, for example, when college graduates are oversampled and we are interested in estimating a logit model to explore the determinants of graduating from college.

For $y_i = 1$ let q_1 be the fraction of observations in the population and h_1 the fraction

in the sample. Similarly for $y_i = 0$ let $q_0 = 1 - q_1$ the fraction in the population and $h_0 = 1 - h_1$ the fraction in the sample. Then, in general (that is, also for models other than the logit model) a weighted maximum likelihood estimator can be used where the weights are given by q_1/h_1 if $y_i = 1$ and q_0/h_0 if $y_i = 0$.[7] Re-weighting the observations makes intuitive sense because in the sample

$$\Pr(y_i = y|x_i)^{\text{sample}} = \left(\frac{h_1}{q_1} \cdot \Pr(y_i = 1|x_i)\right)^y \cdot \left(\frac{h_0}{q_0} \cdot (1 - \Pr(y_i = 1|x_i))\right)^{1-y}$$

and re-weighting offsets this by eliminating the two fractions so that the likelihood contribution is the correct one.

In the case of the logit model[8]

$$\Pr(y_i = 1|x_i)^{\text{sample}} = \frac{(h_1/q_1) \cdot \exp(x_i\theta)}{(h_0/q_0) + (h_1/q_1) \cdot \exp(x_i\theta)}$$

which we can rewrite as

$$\Pr(y_i = 1|x_i)^{\text{sample}} = \frac{\exp\left(x_i\theta + \log\left((h_1/q_1)/(h_0/q_0)\right)\right)}{1 + \exp\left(x_i\theta + \log\left((h_1/q_1)/(h_0/q_0)\right)\right)}.$$

This shows that we can estimate the slope coefficients using a standard logit model. For the intercept we have to subtract $\log\left((h_1/q_1)/(h_0/q_0)\right)$ from the estimate of the intercept to get the population intercept.

5.1.14 Distributional Assumptions in Applied Work

We have just discussed three binary choice models that are commonly used in applied work. At this point one might wonder which one we should use in order to obtain estimates of marginal effects or do predictions.

A first thing to notice is that those three models impose different normalizations on the location and the scale. While the probit and the logit model impose that the mean of ε_i is zero, the linear probability model imposes that it is 0.5. Moreover, the

[7] See Cameron and Trivedi (2005), p. 479, for details and references.

[8] For this I draw on some lecture notes by Daniel McFadden. See also Xie and Manski (1989).

probit model imposes that the variance is one, while the logit model imposes that it is equal to $\pi^2/3 \approx 3.29$. For the uniform distribution it is $1/12$. These differences in the normalizations have no implications on marginal effects or predicted values because they do not restrict the distribution of y_i given x_i.

However, they have implications on the relationship between parameters. In particular, logit slope coefficients should be about $\sqrt{\pi^2/3} \approx 1.81$ times as big as probit slope coefficients. The ones of a linear probability model should be $\sqrt{1/12} \approx 0.29$ times those of probit model.[9] Moreover, the intercepts differ across models because of the differences in the scale and location normalizations. We can follow from this that we should not compare coefficients across models. What we could do, however, is to compare ratios of slope coefficients because they are equal to ratios of marginal effects, so the normalizations are offset. Moreover, these differences in location and scale will cancel out when we look at fitted probabilities or marginal effects. Still, the shape may matter. However, as Figure 5.1.9 shows, for probabilities between 0.2 and 0.8 this is not likely to matter because in that range, all three distribution functions are very close to being linear—hence, the implied choice probabilities and their dependence on x_i will be very similar.

5.1.15 Relaxing Distributional Assumptions

In principle, distributional assumptions can easily be relaxed. Manski (1988b) and Matzkin (1992) study the question of identification and distribution free estimation of discrete choice models.

In fact, we have that

$$\Pr(y_i = 1|x_i) = \mathbb{E}[y_i|x_i],$$

so one can do so by performing a nonparametric regression of y_i on x_i that is very briefly discussed in Section 4.5. The problem is easy when all elements of x_i are discrete. Then, one can use a simple frequency estimator,

$$\widehat{\Pr(y_i = 1|x_i = x)} = \sum_{i=1}^{N} \left(\frac{1\{x_i = x\}}{\sum_{j=1}^{N} 1\{x_j = x\}} \right) \cdot y_i,$$

[9] Amemiya (1981) suggests to multiply them by 1.6 instead. He also suggests to multiply the coefficients of the linear probability model by 4 to compare them to the logit ones.

where N is the number of observations and the expression in parentheses is the (equal) weight that is implicitly used when calculating an average among the observations with $x_i = x$.

If x_i is continuously distributed, then one would instead use a weighting scheme in which the weight depends on the distance between x and x_i,

$$\widehat{\Pr(y_i = 1|x_i = x)} = \sum_{i=1}^{N} \left(\frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x_j - x}{h}\right)} \right) \cdot y_i.$$

Here, $K(\cdot)$ is a so-called Kernel function that generally takes on the highest value at 0 and h is the so-called bandwidth determining how fast the weight decreases in the distance between x_i and x. For further details see, for example, Pagan and Ullah (1999, p. 84ff) and Li and Racine (2007, p. 60ff).

The problem in either case is the curse of dimensionality. To see this, think of the case in which each element of x_i is discrete and can take on 10 values, with equal probability. Then, when there are 1000 observations, we will estimate $\widehat{\Pr(y_i = 1|x_i = x)}$ from $1000/10$ observations when the covariate is a scalar, from $1000/10^2 = 10$ observations when it has two elements, and $1000/10^3 = 1$ observations when it has three elements. This motivates so-called semiparametric models in which some parametric structure is imposed, in the following that x_i affects y_i only through the linear index $x_i\beta$. This is not as restrictive as it may sound, as x_i may contain basis functions so that $x_i\beta$ is an arbitrarily good (in principle) approximation to any function $p(x_i)$.

One of the first estimators that has been proposed was Manski's (1975) maximum score estimator.[10] The key assumptions he builds on is that the median of ε_i given x_i is equal to zero. Thereby, we allow for heteroskedasticity, avoid requiring full independence and specifying a functional form for the distribution. The idea is that we start with a candidate parameter vector β and predict y_i to be 1 if $x_i\beta \geq 0$ and 0 otherwise. Then, we add 1 to the objective function if the prediction is correct, and -1 if not. This can be done for any candidate parameter value. Then, we maximize the objective function subject to a scale normalization such as $\beta'\beta = 1$. Manski (1975) shows that this estimator is consistent, but he does not show asymptotic normality. Manski

[10] This estimator, as well as others that we describe below, are discussed in more detail in Chapter 7 of Pagan and Ullah (1999)

and Thompson (1986) show that the estimator does not converge at the parametric $\sqrt{N}$ rate. However, Horowitz (1992) suggests a smoothed version that is asymptotically normally distributed and converges almost at the parametric rate.

To obtain predicted probabilities or marginal effects after having estimated the parameter vector β, one can then perform a nonparametric regression of y_i on $x_i\hat{\beta}$.

Klein and Spady (1993) build on this important insight and suggested an efficient estimator that is based on maximizing a likelihood function that is built on exactly this idea. They start with the stronger assumption that x_i and ε_i are independent. Then, for any candidate parameter vector β they perform a nonparametric regression of y_i on $x_i\beta$. This provides an estimate of F_{ε_i} because $\mathbb{E}[y_i|x_i] = F_{\varepsilon_i}(x_i\beta)$. Then, they use this estimate to construct the likelihood function, which is then maximized. This means that within each iteration F_{ε_i} is estimated anew.

An alternative is to perform semiparametric least squares, as suggested by Ichimura. For this, we perform a nonparametric regression of y_i on $x_i\beta$ and minimize the variance of the residual by choosing β. This is equivalent to Klein and Spady's estimator if optimal weighting is used.

Gerfin (1996) implements the Klein and Spady (1993) estimator, Horowitz' (1992) smoothed maximum score estimator, the probit model, and another semiparametric estimator. His results suggest that the exact choice of the estimator is not as important as one may first think. However, this is clearly specific to the particular application.

5.1.16 Random coefficients

So far, we have assumed that the coefficient vector is the same for every individual i. If instead it is individual-specific—say β_i drawn from $F_{\beta_i}(\cdot;\mu_\beta,\sigma_\beta^2)$, a normal distribution with mean μ_β and variance-covariance matrix σ_β^2 independent of x_i—our mode is of the form

$$y_i = 1\{x_i\beta_i \geq \varepsilon_i\}.$$

Then, the choice probability is, analogously to (5.1.2),

$$\Pr(y_i = 1|x_i,\beta_i) = F_{\varepsilon_i}(x_i\beta_i).$$

It is of the same form because we additionally condition on the unobserved realization of the random coefficient. The unconditional likelihood contribution for observations

with $y_i = 1$ is given by the expectation over the distribution of the random coefficient,

$$f(1|x_i;\mu_\beta,\sigma_\beta) = \int F_{\varepsilon_i}(x_i\beta_i)\, dF_{\beta_i}(\beta_i;\mu_\beta,\sigma_\beta^2).$$

This can be simulated as explained in Section (4.6). Then, one can proceed as usual when performing maximum likelihood estimation.

5.2 Ordered Choice

5.2.1 General Model

The ordered choice model is a generalization of the binary choice model to $J > 2$ alternatives, $j = 1,\dots,J$. The ordering of these alternatives is the same for all individuals and known. Examples for this are outcomes such as the number of children, or educational choice between no high school, high school, and college.

There is a *latent*, that is unobserved, variable

$$y_i^* = x_i\beta + \varepsilon_i \tag{5.2.1}$$

which corresponds to $x_i\beta - \varepsilon_i$ in (5.1.1). Again, x_i is a vector of exogenous covariates and ε_i is assumed to have a known distribution.[11]

Besides, there are $J-1$ unobserved thresholds $\alpha_1,\dots,\alpha_{J-1}$.[12] These thresholds translate the latent index into reports and are, in the basic variant of the model, the same for all individuals. We observe

$$y_i = \begin{cases} 1 & \text{if } y_i^* \le \alpha_1 \\ 2 & \text{if } \alpha_1 < y_i^* \le \alpha_2 \\ \quad\vdots & \\ J & \text{if } \alpha_{J-1} < y_i^*. \end{cases} \tag{5.2.2}$$

[11] See footnote 1 on the similarity between those two formulations if the distribution ε_i is symmetric.

[12] There are also so-called interval regression models for situations in which the thresholds are known. This is the case if the latent variable is, for example, income and the observed variable is an income bracket. See Beresteanu and Molinari (2008) for recent work on regressions with so-called set valued random variables on the left hand side.

For example, if y_i^* is perceived life satisfaction, then y_i is the report on life satisfaction on, say, a 10 point scale. The binary choice model is a special case with $J = 2$.[13]

5.2.1 illustrates this for the case of four outcomes. For four outcomes we have three unobserved thresholds, α_1, α_2, and α_3. The curve is the density density of ε_i that is centered at zero. The three vertical lines are the thresholds ε_i has to cross according to (5.2.2) such that we observe certain outcomes: $y_i = 1$ in the leftmost interval to the left of $\alpha_1 - x_i\beta$, $y_i = 2$ between $\alpha_1 - x_i\beta$ and $\alpha_2 - x_i\beta$, and so on. The corresponding probabilities for this to happen are the respective areas underneath the density. For $y_i = 1$ this is, by definition, the c.d.f. of ε_i evaluated at $\alpha_1 - x_1\beta$. For $y_i = 2$ it is equal to the c.d.f. evaluated at $\alpha_2 - x_i\beta$ minus the c.d.f. evaluated at $\alpha_2 - x_i\beta$, which is the probability that the outcome is either $y_i = 1$ or $y_i = 2$ minus the probability that it is $y_i = 1$, which of course gives the probability that it is $y_i = 2$.

As before we need to impose normalizations on the location and scale. To see why this is necessary notice that we can always add a constant c_l to all the thresholds and to y_i^* and the observed outcome would still be the same. Moreover, we can always multiply y_i^* and all thresholds by a positive constant c_s without changing the observed outcome. for example, for $j = 2$ we have that

$$\alpha_1 < y_i^* \leq \alpha_2$$

is equivalent to

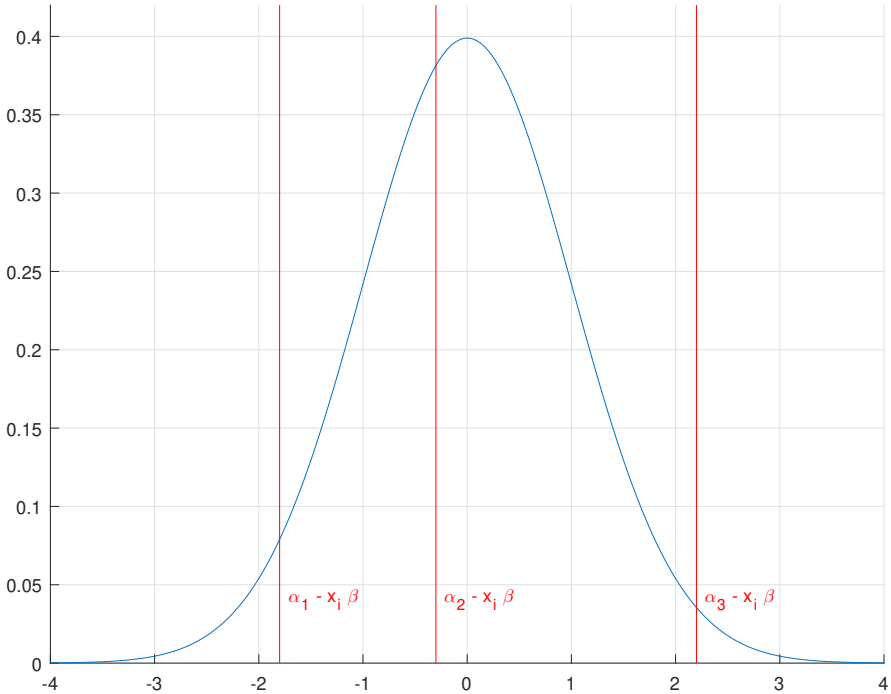
$$c_s \cdot (c_l + \alpha_1) < c_s \cdot (c_l + y_i^*) \leq c_s \cdot (c_l + \alpha_2).$$

The normalizations that are typically imposed differ slightly from the normalizations in the binary choice model. Here we impose that x_i does not include a constant term whereas in the binary choice model we have instead imposed that the threshold (here $\alpha_1 = 0$) is equal to zero. The scale normalization is imposed, as in the binary

[13] To see this, start with the ordered choice model with $J = 2$, no intercept in β, and therefore no constant term in x_i. Then, $y_i^* = x_i\beta + \varepsilon_i$. Relabel the outcomes into 0 and 1. Define $\tilde{x}_i = (1, x_i)$, $\tilde{\beta} = (-\alpha_1, \beta)'$, and $\tilde{\varepsilon}_i = -\varepsilon_i$. Then we can write

$$y_i = \begin{cases} 0 & \text{if } \tilde{x}_i\tilde{\beta} \leq \tilde{\varepsilon}_i \\ 1 & \text{if } \tilde{x}_i\tilde{\beta} > \tilde{\varepsilon}_i. \end{cases}$$

This model generates the same probabilities as the one in (5.1.1) as long as $\tilde{\varepsilon}_i$ is continuously distributed such that $\tilde{x}_i\tilde{\beta} = \tilde{\varepsilon}_i$ occurs with probability zero.



The figure shows the density of ε_i of an ordered probit model, along with three thresholds for ε_i so that we observe $y_i = 1$ in the leftmost area, followed by $y_i = 2$ to the right of this, and so on.

Figure 5.2.1: Ordered probit model with four possible outcomes

choice model, by specifying a distribution for ε_i. For the ordered probit model we assume that ε_i is standard normally distributed and for the ordered logit model we assume that it follows a logistic distribution.

5.2.2 Identification

To see that the model is identified we write

$$
\begin{aligned}
(5.2.3) \qquad & \Pr(y_i = 1|x_i) = \Pr(\varepsilon_i \leq \alpha_1 - x_i\beta) = F_{\varepsilon_i}(\alpha_1 - x_i\beta) \\
& \Pr(y_i = 2|x_i) = \Pr(\alpha_1 < x_i\beta + \varepsilon_i \leq \alpha_2) = F_{\varepsilon_i}(\alpha_2 - x_i\beta) - F_{\varepsilon_i}(\alpha_1 - x_i\beta) \\
& \qquad\qquad \vdots \\
& \Pr(y_i = J|x_i) = \Pr(\alpha_{J-1} < x_i\beta + \varepsilon_i) = 1 - F_{\varepsilon_i}(\alpha_{J-1} - x_i\beta).
\end{aligned}
$$

For a given x_i these are J observed choice probabilities that sum to one, so we essentially have $J-1$ equations with $K+J-1$ unknown parameters, K in β and $J-1$ unobserved thresholds. Once we have two values of x_i we have $2(J-1)$ equations and still $K+J-1$ unknown parameters. So eventually there will be more equations than parameters and in this case the model is identified.

Like in Section 5.1.5 we assume that the distribution of ε_i is continuous. Then, $F_{\varepsilon_i}(\cdot)$ is strictly increasing and hence invertible. Denote the inverse function by $F_{\varepsilon_i}^{-1}(\cdot)$. Then, it follows from (5.2.3) that for a given value of x_i we have

$$
\begin{aligned}
F_{\varepsilon_i}^{-1}(\Pr(y_i = 1|x_i)) &= \alpha_1 - x_i\beta \\
F_{\varepsilon_i}^{-1}(\Pr(y_i \leq 2|x_i)) &= \alpha_2 - x_i\beta \\
&\vdots \\
F_{\varepsilon_i}^{-1}(\Pr(y_i \leq J-1|x_i)) &= \alpha_{J-1} - x_i\beta.
\end{aligned}
$$

These are $J-1$ equations. Now use the variation in x_i to select as many linearly independent equations as there are unknowns.

This shows that we can identify the parameters β and $\alpha_1, \ldots, \alpha_{J-1}$. One way to think about this is that the independence between ε_i and x_i implies that the shape of the distribution of ε_i stays the same, which allows us to apply the mapping $F_{\varepsilon_i}^{-1}$ to invert the observed choice probabilities.

5.2.3 Estimation

Estimation is straightforward. From the choice probabilities we can construct the log likelihood function

$$\mathscr{L}(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{J} 1\{y_i = j\} \cdot \ln\left(\Pr(y_i = j|x_i)\right)$$

where $1\{y_i = j\}$ is an equal to 1 if $y_i = j$ and 0 otherwise.

5.2.4 Reporting of Results

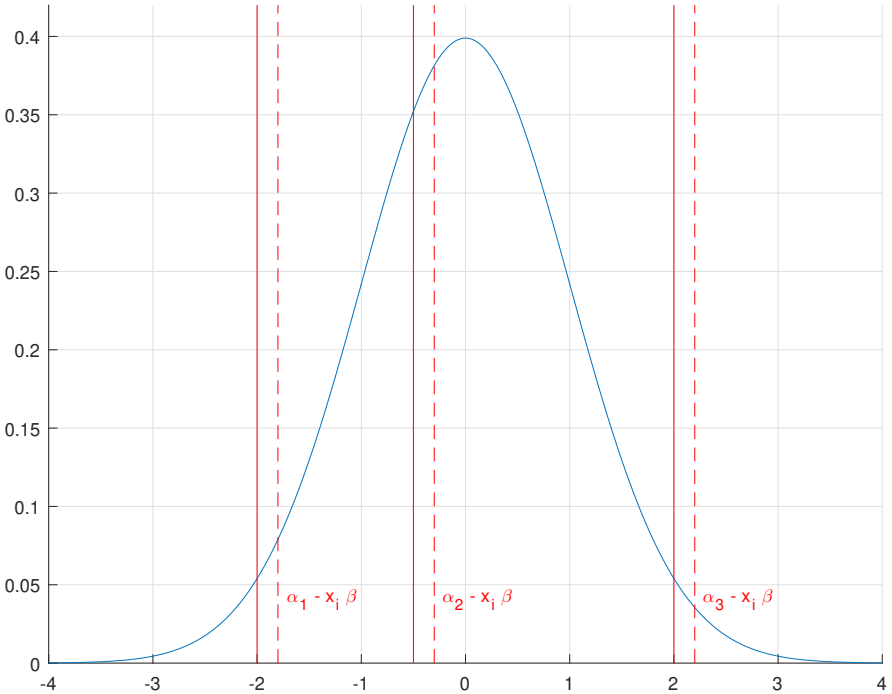
In the ordered choice model changes in x_i shift the mean of the distribution of y_i^* but leave the thresholds α_j unchanged. Figure 5.2.2 shows what the effect of this is. It is similar to Figure 5.2.1 except that the cutoff points for ε_i are shifted to the left, from the dashed vertical lines to the solid ones, respectively. They shift to the left because $x_i\beta$ has increased.

Marginal effects are changes in the probability that are due to marginal changes in those thresholds. Suppose we change $x_i\beta$ by a small amount. The marginal effect on the probability to observe $y_i = 1$ will then be minus (because the vertical line moves to the left so that probability to choose $y_i = 1$ becomes smaller) the area underneath the density, between the leftmost solid line and the leftmost dashed line, divided by the distance between the two vertical lines. We divide by the size of the change because marginal effects are always for a unit change in $x_i\beta$. This gives, in the limit in which the two vertical lines are infinitely close to one another, the density evaluated at $\alpha_1 - x_i\beta$. If we now want to look at the vector of marginal effects of changes in components of x_i, then we need to multiply this by the coefficient vector β. This can best be seen by taking the derivative of

$$\Pr(y_i = 1|x_i) = F_{\varepsilon_i}(\alpha_1 - x_i\beta)$$

in (5.2.3) with respect to the kth component of x_i denoted by x_{ki}. Applying the chain rule gives

$$\frac{\partial \Pr(y_i = 1|x_i)}{\partial x_{ki}} = f_{\varepsilon_i}(\alpha_1 - x_i\beta) \cdot \beta_k,$$



The figure shows the change of the thresholds that is caused by an increase in the mean of the latent index.

Figure 5.2.2: Effect of change in index in ordered choice model

where β_k is the kth component of β. Similar arguments hold for marginal effects on the other choice probabilities,

$$\frac{\partial \Pr(y_i = j|x_i)}{\partial x_i} = \begin{cases} -f_{\varepsilon_i}(\alpha_1 - x_i\beta)\cdot\beta & \text{if } j = 1 \\ -\left(f_{\varepsilon_i}(\alpha_j - x_i\beta) - f_{\varepsilon_i}(\alpha_{j-1} - x_i\beta)\right)\beta & \text{if } j = 2,\ldots,J-1 \\ f_{\varepsilon_i}(\alpha_{J-1} - x_i\beta)\cdot\beta & \text{if } j = J. \end{cases}$$

Figure 5.2.2 also shows that the marginal effect of a positive change in the mean of the index $x_i\beta$, which could for example originate in a positive change of one of the components of x_i with a positive coefficient, will always be negative on the probability

to choose $y_i = 1$ and positive on the probability to choose $y_i = 4$. For the other choices, $y_i = 2$ and $y_i = 3$, the sign of the effect depends on the value of $x_i\beta$. Generally, the sign of the marginal effect of a change in a component x_{ki} of x_i on the probability to observe $y_i = J$ will be equal to the sign of the coefficient β_k and the sign of the marginal effect on the probability to observe $y_i = 1$ will be the opposite of the sign of the coefficient β_k. In-between, the sign of the marginal effect depends on the value of $x_i\beta$. The reason for this is that choice probabilities sum to one so that the sum of the marginal effects has to sum to zero.

In the binary choice model, there are just two possible choices and we typically only consider the marginal effect on the probability to observe $y_i = 1$ (instead of $y_i = 0$), so also there the sign of the marginal effect is equal to the sign of the y_i with the highest "j."

5.2.5 An Ordered Probit Model with a Random Coefficient

Here, by means of a Monte Carlo study similar to the one performed in Section 5.1.12, we look at an ordered probit model with a random coefficient. In particular, we look at a model an index

$$y_i^* = x_{1i}\beta_1 + x_{2i}\beta_{2i} + \varepsilon_i,$$

where β_{2i} is a log normally distributed random coefficient with mean $\mu_{\beta_{2i}}$ and standard deviation $\sigma_{\beta_{2i}}$.

The likelihood function for this model is programmed up in `nll_oprobit.m`.

```
function [nll] = nll_oprobit(pars,y,X,draws)
% negative average log likelihood and score for multinomial
    logit model
% with random coefficient

beta1=pars(1);
beta2_mu=pars(2);
beta2_sig=pars(3);
alpha1=pars(4);
alpha2=pars(5);

f=0; %initialize likelihood contribution

```

```
for draw=1:size(draws,2)
    beta2=exp(beta2_mu+beta2_sig*draws(:,draw));
    for j=1:size(X,3)
        beta2(:,1,j)=beta2(:,1,1);
    end
    Xb=X(:,1,:).*beta1+X(:,2,:).*beta2;
    prob=(y==1).*normcdf(alpha1-Xb)...
        +(y==2).*(normcdf(alpha2-Xb)-normcdf(alpha1-Xb))...
        +(y==3).*(1-normcdf(alpha2-Xb)); %likelihood
    prob=max(0.0000001,prob); %to avoid numerical problems due
        to zero probabilities
    probseq=prod(prob,3); %likelihood contribution of the
        sequence of choices
    f=f+probseq/size(draws,2); %calculate average likelihood
        contribution, across simulation draws
end

nll=-mean(log(f)); %negative of the average simulated log
    likelihoods
```

The only output element is `nll`, which is the average negative log likelihood. In principle, we could also program a gradient in a similar was as for the binary probit model, but it is not necessary for the numerical optimization procedure (but it speeds it up). The first input element is a vector containing the parameters, `pars`. In line 5 through 9 we can see that the first element is β_1, the second one is the mean of β_{2i}, the third one is the standard deviation of β_{2i}, the fourth parameter is the first threshold α_1, and the last parameter is the second threshold α_2. The second input is the $N \times 2$ matrix, N being the number of observations, `X` of explanatory variables.

There is an additional input to the function `nll_oprobit.m`., namely `draws`. This is a matrix that contains a number of draws for each individual that we will generate in the main program and then pass over to the likelihood function. It is important to use the same draws in different iterations of the likelihood function, as explained in Section (4.6).

The average log likelihood is now calculated across individuals and draws of the random coefficient, as described in Section (4.6). The loop between line 13 and 21 is over draws of the random coefficient. In line 14, we first calculate the particular draw from the log normal distribution for each individual that we then store in the

N-vector `beta2`. then, we calculate `Xb` as $x_{1i}\beta_1 + x_{2i}\beta_{2i}$. Thereafter, we calculate the choice probabilities according to (5.2.3). In line 20 we then add one over the number of draws, `size(draws,2)`, times the likelihood contribution from this particular draw to the vector `f` that we initialize in line 11. This loop runs `size(draws,2)` times, which is the number of draws, hence we get the average log likelihood for each individual in the end.

The following program runs the Monte Carlo analysis.

```
% parameters for data generating process
N=1000;
T=5;

% parameters we seek to estimate
beta1=0.4;
beta2_mu=-0.3; %mean of beta2
beta2_sig=0.2; %standard deviation of beta2
alpha1=-1; %first threshold
alpha2=0.5; %second threshold

% parameters for optimization
startvalues = [0.4;-0.3;0.2;-1;0.5]; %use the true values as
    starting values
options = optimset('Algorithm','interior-point','Display','iter
    ','GradObj','off');
lb=-inf(5,1); %define lower bounds on parameters
lb(3)=0; %third parameter is variance, so it should be
    nonnegative

% parameters and initialization for Monte Carlo
repetitions = 2; %number of repetitions
numberdraws=100; %number of random draws for the simulation
    step
bhat = NaN(5,repetitions); %bhat are the estimates
vhat = NaN(5,5,repetitions); %estimate of variance-covariance
    matrix
nH = NaN(5,5,repetitions); %negative Hessian

% Monte Carlo
for i = 1:repetitions
    % generate data
```

```
    X=poissrnd(3,N,2,T); %covariates come from poisson
        distribution with parameter 3
29  beta2=exp(mvnrnd(beta2_mu,beta2_sig^2,N)); %beta2 is log
        normally distributed
    % stardard normal random draws for optimization, important
        to draw them
31  % before maximizing the likelihood and not anew within each
         iteration;
    % otherwise the procedure stops to early because of a
        different set of
33  % random draws that leads to a lower value of the
        likelihood even
    % though the candidate parameters in that iteration should
        actually
35  % lead to a higher value; such things are due to simulation
         error
    % draws=normrnd(0,1,N,numberdraws); %these are the old
        draws for (a)-(c)
37  % this is for part (d)
    p = haltonset(N);
39  p = scramble(p,'RR2');
    uniformdraws=p(1:N,1:numberdraws);
41  draws=norminv(uniformdraws,0,1);
    for j=1:T
43      beta2(:,1,j)=beta2(:,1,1);
        draws(:,1,j)=draws(:,1,1);
45  end
    epsilon=normrnd(0,1,N,1,T);
47  ystar=X(:,1,:).*beta1+X(:,2,:).*beta2+epsilon;
```

In line 1 through 9, we specify that there are $N = 1000$ observations and pick parameters which we then seek to estimate from data that we generate, as in Section (5.1.12), within each iteration between line 27 and 47. Parameters for the optimization are given in line 11 through 15 and parameters for the Monte Carlo are given in line 17 through 20. In line 20 we set a so-called seed so that the same random numbers are generated whenever we run this file. This is done so that the reader can replicate the results reported in Table 5.1 below. The random draws, `draws`, are generated in line 40. In line 45, we obtain an estimate of the variance covariance matrix of the estimates that uses the numerical Hessian `nH`, that is obtained as an output in line 44. We will only use the

Table 5.1: Ordered probit Monte Carlo results

	θ	estimates $\hat{\theta}$				estimates of ste. of $\hat{\theta}$		
		5th perc.	avg.	95th perc.	std.	5th perc.	avg.	95th perc.
β_1	0.40	0.30	0.42	0.56	0.08	0.06	0.08	0.10
$\mu_{\beta_{2i}}$	-0.30	-0.54	-0.23	0.16	0.22	0.13	0.22	0.38
$\sigma_{\beta_{2i}}$	0.20	0.00	0.19	0.55	0.20	0.17	0.47	0.69
α_1	-1.00	-4.89	-1.27	-0.53	1.25	0.25	0.43	0.44
α_2	0.50	0.15	0.57	1.06	0.28	0.21	0.28	0.39

This table contains the results from the Monte Carlo study for the ordered probit model. The first column contains the parameters of the underlying data generating process. The next four columns describe the moments of the distribution of estimates across Monte Carlo replications. The last three columns describe the variation of the estimate of the standard errors from the numerical Hessian, also across replications.

estimates of the variance-covariance matrix if the numerical Hessian is invertible. This can be checked by obtaining the reciprocal of the condition number in line 46. The condition number is the ratio of the largest singular value of a matrix to the smallest. Our criterion is that the reciprocal of the condition number is above 10^{-6}. This will be the case in about 95 percent of the replications.

Running this program takes considerably longer because of the simulation step. It is roughly proportional to the number of random draws we use. Table 5.1 summarizes the results when we use 1000 draws to approximate the likelihood function. It shows that the true parameters in the first column lie well between the 5th and 95th percentile of the estimated ones, across Monte Carlo replication. On average, the parameter estimates are close to the true value and the standard deviation is reasonably low. The accuracy can generally be improved by increasing the number of observations and the number of simulation draws. The mean and the standard deviation of β_{2i} are estimated less precisely, which is typical for a random coefficient model. Comparing the standard deviation of the estimates across replications to properties of the distribution of the estimates thereof across replications shows that the estimates are roughly correct.

5.2.6 Differences in Reporting Behavior

The model can be generalized to allow for differences in reporting behavior that give rise to differences in reports y_i if the underlying index y_i^* is the same. This could be for cultural reasons which yield to higher reports by individuals from the U.S., for example. See Kapteyn et al. (2007) for an application to the reporting differences between the U.S. and The Netherlands for work disability. They use so-called vignette questions in surveys to identify the differences in reporting behavior across individuals. The idea here is that a person is described to all individuals `and` then they are asked to evaluate this individual's work disability on the same scale as the one they will use to report on their own work disability. Then, two assumptions are made, namely that individuals evaluate the vignette in the same way as they evaluate themselves, and that the underlying distribution of y_i^* is the same for everybody. This is a very powerful approach once the assumptions hold, and even allows one to relate differences in reporting behavior to observed characteristics of individuals. The key underlying idea is that systematic differences in reports on the work disability of the vignette person must be due to differences in reporting behavior. Having recovered these differences in reporting behavior one can then control for those when estimating the determinants of the underlying index, y_i^*.

5.2.7 Cardinality and Fixed Effects

The model specifies y_i to be a nonlinear step function of y_i^* and imposes a normalization on the scale of the coefficients. Therefore, clearly, estimates of β in (1) will generally differ from those obtained when regressing y_i on x_i using OLS. However, ratios of coefficients can still be similar in specific contexts.

Ferrer-i Carbonell and Frijters (2004) look into this question in the context of happiness, where answers are typically given on an 11 point scale between 0 and 10 and panel data are often available, for example in the German Socio-Economic Panel. They find that results are very similar between both models in the sense of ratios of coefficients being very similar. However, they point out that it assuming cardinality makes it easier to control fixed effects by means of the fixed effects . They also explain how one can allow for fixed effects in an ordered logit model, but again, comparing results allowing for fixed effects shows that there is not much of a difference between a linear

and a non-linear model.

5.2.8 An Example of a Structural Model

A rich example of a structural model with an ordered choice type of structure is the entry and exit model of Bresnahan and Reiss (1991). In words, there are markets that differ in their market size. For a given number of firms, the bigger the market the bigger the variable profits will be. At the same time, holding the market size fixed, the more firms are in that market the lower the variable profits per firm and also the lower industry profits will be, because competition becomes more fierce the more firms are in the market.

The model is a static model. Firms earn variable profits in each periods and need to cover their fixed costs. Otherwise, they exit or do not enter in the first place. This means that in each market, variable profits need to be high enough to cover fixed costs, and therefore there will be a finite number of firms.

The data consists of market characteristics, foremost the number of inhabitants, demographics, as well as the number of firms of a particular type in that market. The aim is to determine how average profits vary with the number of firms in the market. This is interesting from a competition policy point of view because informs policy makers about the number of firms that is necessary to ensure that competition is strong enough. Bresnahan and Reiss specify normalized profits of the Nth entrant to be (p. 990)

$$\Pi_N = \bar{\Pi}_N + \varepsilon,$$

where average profits

$$\bar{\Pi}_N = population \cdot \left(\alpha_1 + X\beta - \sum_{n=2}^{N} \alpha_N \right) - F_N.$$

that are weakly decreasing in N, and assume that ε is standard normally distributed. Denoting the c.d.f. of the standard normal distribution by Φ and using the symmetry,

$1-\Phi(a)=\Phi(-a)$, we have

$$\begin{aligned}\Pr(N=0) &= \Pr(\Pi_1<0)\\ &= \Pr(\varepsilon<-\bar{\Pi}_1)\\ &= \Phi(-\bar{\Pi}_1)\\ &= 1-\Phi(\bar{\Pi}_1)\\ &= 1-\Phi\left(population\cdot(\alpha_1+X\beta)-F_1\right).\end{aligned}$$

In words, the probability that there are no firms, $N=0$, in a market is the probability that profits of a monopolist, Π_1, would be negative because fixed costs are too high. The last line shows that this is the less likely to be the case the higher the population, as Φ is a c.d.f. and hence increasing in its argument and variable profits of a monopolist, $\alpha_1+X\beta$, are positive. Still, total profits can of course be negative because of fixed costs.

The probability to observe a monopolist is equal to the probability that a monopolist earns positive profits, but profits for the second entrant would be negative. This gives

$$\begin{aligned}\Pr(N=1) &= \Pr(\Pi_1\geq 0\wedge\Pi_2<0)\\ &= \Pr(\bar{\Pi}_1+\varepsilon\geq 0\wedge\bar{\Pi}_2+\varepsilon<0)\\ &= \Pr(\varepsilon<-\bar{\Pi}_2)-\Pr(\varepsilon<-\bar{\Pi}_1)\\ &= \Phi(-\bar{\Pi}_2)-\Phi(-\bar{\Pi}_1)\\ &= \left(1-\Phi(\bar{\Pi}_2)\right)-\left(1-\Phi(\bar{\Pi}_1)\right)\\ &= \Phi(\bar{\Pi}_1)-\Phi(\bar{\Pi}_2)\\ &= \Phi\left(population\cdot(\alpha_1+X\beta)-F_1\right)-\Phi\left(population\cdot(\alpha_1+X\beta-\alpha_2)-F_2\right).\end{aligned}$$

The other choice probabilities have a similar structure.

This model shares some of the features of the standard ordered probit model with an index structure such as in (5.2.1), where the main explanatory variable is the population, but is more general in that the coefficient on that variable, which is related to variable profits, actually depends on the number of firms in the market, which is the "report" y_i on the left hand side, as in (5.2.2).

5.3 Multinomial Choice

5.3.1 The Measurement of Urban Travel Demand

So far, we have considered models for choice among alternatives that can be ordered. Now we turn to situations in which the J alternatives are not ordered. There are many examples for this in the area of consumer choice, such as brand choice. Other examples are the choice of an undergraduate major, or, famously the choice of travel mode for commuting to work.

In a ground-breaking and thoughtful study, Nobel laureate Daniel McFadden showed in 1974b how the multinomial logit model, which is an extension to the binary logit model that we have seen in Section 5.1.10, could be adapted for this. Nowadays, this is still an important area of public policy, as transportation systems are critical components of every urban economy. Public transportation projects are often massive and mutually exclusive, with irreversible cumulative effects over long periods. Therefore, it is important to obtain accurate forecasts under alternative transportation policies and to precisely calculate respective benefits to calculate welfare effects. McFadden models travel demand as the result of aggregation over the urban population, modeling each member's decision making based on his personal needs and environment.

The multinomial logit model he used possesses the independence of irrelevant alternatives (IIA) property. We will discuss this model and see why this property allowed McFadden (1974b) to forecast demand for an alternative that had not been introduced before, the Bay Area Rapid Transport (BART) commuter train. At the moment he undertook the analysis, the only options available to commuters were traveling by car or bus. McFadden's approach was then to use survey data to estimate utility parameters, such as the dependence of utility on travel time and travel costs. Based on these he predicted at the individual level how likely it was that the individual chose to commute by BART once it became available. The IIA property in this context imposes that the ratio between the probability to travel by bus and the probability to travel by car remains unchanged once the BART alternative becomes available.

In textbooks, the IIA property is typically seen as a "shortcoming of the model." I find this slightly misleading, as the model was brought to economics as the model that precisely possesses this property, for reasons of tractability and because it allowed researchers to predict demand for alternatives that have not been introduced thus far.

The latter two motivations are still important nowadays. We will discuss below how one can test whether this property of the model is rejected by the data and how it can be relaxed.

5.3.2 Luce's Axiom and Independence of Irrelevant Alternatives

An important early contribution that precedes McFadden's is the book by Luce (1959), in which he develops a probabilistic theory of choice relating the probability that a particular choice is made to the choice sets that are available to a decision maker. One of his first observations is that (p. 3) "[c]omplete data concerning the choices that a person makes from each possible pair of alternatives taken from a set of three or more alternatives do not appear to determine what choices he will make when the whole set is provided." This is the general problem that also McFadden (1974b) faced: it is *a priori*, that is without prior knowledge or without making assumptions, not clear how many individuals will choose to travel by BART. Luce then explains how he proposes to overcome this problem. "The method of attack is to introduce a single axiom relating the various probabilities of choices from finite sets of alternatives." The axiom is that the probability to choose an alternative from a subset R, when the full choice set is T, is equal to the probability that an alternative is chosen out of the subset R when the choice set is S, times the probability that an alternative is chosen out of subset S when the choice set is T, no matter what the set S is. This may seem tautological as it sounds like a "true conditional probability statement" we are very much used to, but—to the contrary—this axiom turns out to be useful because it restricts choice behavior in important ways. He shows this on p. 9. The first step is to show that the axiom implies that for two alternatives x and y,

$$\Pr(x \succeq x' \forall x' \in S) = \Pr(x \succeq y) \cdot \left(\Pr(x \succeq x' \forall x' \in S) + \Pr(y \succeq x' \forall x' \in S)\right),$$

where $\Pr(x \succeq x' \forall x' \in S)$ is the probability that x is the preferred choice out of all choices in S and $\Pr(x \succeq y)$ is the probability that x is preferred over y. The last probability is defined accordingly. This means that the probability to choose x out of S is equal to the probability to prefer x over y times the probability that x or y is the preferred choice in S. From this it follows that

$$\Pr(x \succ x' \forall x' \in S)\,(1 - \Pr(x \succ y)) = \Pr(x \succ y) \cdot \Pr(y \succ x' \forall x' \in S).$$

This implies

$$\Pr(x \succ x' \forall x' \in S) \cdot \Pr(y \succ x) = \Pr(x \succ y) \cdot \Pr(y \succ x' \forall x' \in S),$$

which gives

$$\frac{\Pr(x \succ y)}{\Pr(y \succ x)} = \frac{\Pr(x \succ x' \forall x' \in S)}{\Pr(y \succ x' \forall x' \in S)} \tag{5.3.1}$$

and means that the ratio on the left hand side is independent of S. *This is the IIA property* that is always talked about in the context of multinomial choice models.[14] In Luce's words (p. 9) "the idea states that if one is comparing two alternatives according to some...preference, this comparison should be unaffected by the addition of new alternatives or the subtraction of old ones." But he also writes that "[i]t should be noted that it is only the ratio of the two probabilities, not the probabilities themselves, that is invariant with changes of the irrelevant alternatives."

He then derives that under his axiom that there exists a positive real-valued function v, which is unique up to multiplication by a positive constant, such that for every subset S of the full choice set

$$\Pr(x \succ x' \forall x' \in S) = \frac{v(x)}{\sum_{x' \in S} v(x')}. \tag{5.3.2}$$

5.3.3 McFadden's Utility Foundation

Luce's equation (5.3.2) is not directly useful, but an important point of departure for McFadden (1974a), who proposes to think of an individual i as having a utility function that can be written as

$$u_{ij} = \bar{u}_{ij} + \varepsilon_{ij}.$$

It gives the utility for every alternative j and consists of a non-stochastic part $\bar{u}_{ij}$ that "reflects the 'representative' tastes of the population," while ε_{ij} "is stochastic and reflects the idiosyncrasies of this individual in tastes for the alternative". Later, in Section

[14] Luce is not the inventor of this approach, but writes that the idea "was brought to the fore" by Arrow in 1951.

(5.3.5), we will see how structure can be imposed on the utility function. The important point to bear in mind here is that systematic utility $\bar{u}_{ij}$ can be defined to depend on characteristics of the product, which are travel cost and time in the travel demand problem of McFadden (1974b). This then ultimately allows extrapolation and prediction of demand for new alternatives.

The individual chooses the alternative that maximizes utility. Then, the probability that this individual will choose alternative $y_i = j$ is

$$\begin{aligned} \Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) &= \Pr\left(u_{ij} > u_{ij'} \text{ for all } j' \neq j\right) \\ &= \Pr\left(\bar{u}_{ij} + \varepsilon_{ij} > \bar{u}_{ij'} + \varepsilon_{ij'} \text{ for all } j' \neq j\right) \end{aligned} \tag{5.3.3}$$

This can be written as

$$\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) = \int_{-\infty}^{\infty} F^{(j)}_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}\left(\bar{u}_{ij} + \varepsilon_{ij} - \bar{u}_{i1},\ldots,\bar{u}_{ij} + \varepsilon_{ij} - \bar{u}_{iJ}\right) d\varepsilon_{ij}, \tag{5.3.4}$$

where the choice set is $j = 1,\ldots,J$ and $F^{(j)}_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}$ is the partial derivative of the joint c.d.f. of ε_{i1} through ε_{iJ} with respect to its jth argument. This is because in (5.3.3) we require all ε'_{ij} to be less than $\bar{u}_{ij} + \varepsilon_{ij} - \bar{u}_{ij'}$ and equality occurs with probability zero, and because we integrate out ε_{ij}—hence, we have to take the partial derivative with respect to the jth argument. In Section (5.3.4) we provide an alternative derivation that clarifies this a bit more.

Starting from there, he notes, one can specify that joint distribution of taste shocks, impose structure on the $\bar{u}_{ij}$'s and estimate the model by maximum likelihood.

Instead of picking a distribution, McFadden proposes to impose Luce's (1959) axiom, which imposes that choice probabilities are of the form in (5.3.2). He then shows that if the ε_{ij}'s are independent from one another and each of them follows the type 1 extreme value distribution, then we have

$$\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) = \frac{\exp\left(\bar{u}_{ij}\right)}{\sum_{j'}\exp\left(\bar{u}_{ij'}\right)}.$$

He refers to Luce and Suppes (1965) for a derivation, who attribute the following argument to E. Holman and A. Marley. The resulting choice probabilities were first derived by Marschak (1959). One way to derive them is to substitute

(5.3.5)
$$\begin{aligned} & F^{(j)}_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}\left(\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{i1},\ldots,\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{iJ}\right) \\ &= \exp\left(-\varepsilon_{ij}-\bar{u}_{ij}+\bar{u}_{ij}\right)\prod_{j'=1}^{J}\exp\left(-\exp\left(-\varepsilon_{ij}-\bar{u}_{ij}+\bar{u}_{ij'}\right)\right) \\ &= \exp\left(-\varepsilon_{ij}\right)\prod_{j'=1}^{J}\exp\left(-\exp\left(-\varepsilon_{ij}-\bar{u}_{ij}+\bar{u}_{ij'}\right)\right) \\ &= \exp\left(-\varepsilon_{ij}\right)\exp\left(-\exp\left(-\varepsilon_{ij}\right)\left(\sum_{j=1}^{J}\exp\left(\bar{u}_{ij'}-\bar{u}_{ij}\right)\right)\right) \end{aligned}$$

into (5.3.4), which gives

(5.3.6)
$$\begin{aligned} \Pr(y_i=j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) &= \int_{-\infty}^{\infty}\exp\left(-\varepsilon_{ij}\right)\exp\left(-\exp\left(-\varepsilon_{ij}\right)\left(\sum_{j=1}^{J}\exp\left(\bar{u}_{ij'}-\bar{u}_{ij}\right)\right)\right)d\varepsilon_{ij} \\ &= \int_{-\infty}^{\infty}\exp\left(-\varepsilon_{ij}\right)\exp\left(-\exp(-\varepsilon_{ij}-c)\right)d\varepsilon_{ij}, \end{aligned}$$

where
$$-c=\log\left(\sum_{j=1}^{J}\left(\exp\left(\bar{u}_{ij'}-\bar{u}_{ij}\right)\right)\right).$$

This is equal to
$$\int_{-\infty}^{\infty}\exp\left(-\exp\left(-(\varepsilon_{ij}+c)\right)\right)\exp(-(\varepsilon_{ij}+c)+c)\,d(\varepsilon_{ij}+c).$$

Notice that here we integrate over $\varepsilon_{ij}+c$ instead of ε_{ij} before.[15] This is called "change in variables" or "substitution" because we can define another variable which is equal to $\varepsilon_{ij}+c$ and integrate over this new variable instead of the original variable ε_{ij}. Using this we have

(5.3.7)
$$\Pr(y_i=j|\bar{u}_{i1},\ldots,\bar{u}_{iJ})=\exp(c)\int_{-\infty}^{\infty}\exp\left(-\exp\left(-(\varepsilon_{ij}+c)\right)\right)\exp(-\varepsilon_{ij}+c)\,d(\varepsilon_{ij}+c).$$

[15] For this I am following the lecture notes by Guido Imbens.

As already discussed in Section 5.1.10 the type 1 extreme value distribution is

$$F_{\varepsilon_{ij}}(\varepsilon_{ij}) = \exp(-\exp(-\varepsilon_{ij})) \tag{5.3.8}$$

and has the density function

$$f_{\varepsilon_{ij}}(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \cdot \exp(-\exp(-\varepsilon_{ij})). \tag{5.3.9}$$

(5.3.9) implies that the integral in (5.3.7) is equal to 1 because $\exp(-\exp(-(\varepsilon_{ij}+c))) \cdot \exp(-\varepsilon_{ij}+c)$ is the p.d.f. of $\varepsilon_{ij}+c$. Therefore,

$$\begin{aligned}\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) &= \exp(c) \\ &= \frac{1}{\sum_{j'=1}^{J}\exp\left(\bar{u}_{ij'}-\bar{u}_{ij}\right)} \\ &= \frac{1}{\exp(-\bar{u}_{ij})\sum_{j'=1}^{J}\exp(\bar{u}_{ij'})},\end{aligned}$$

from which we get the well-known choice probability

$$\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) = \frac{\exp(\bar{u}_{ij})}{\sum_{j'=1}^{J}\exp(\bar{u}_{ij'})}. \tag{5.3.10}$$

The revolutionary contribution was to go beyond Luce (1959), who derived a similar expression in (5.3.2), and relate this choice probability to "'representative' tastes of the population". At the same time, McFadden is very clear about the limitations related to the IIA property and writes (p. 113)

> The primary limitation of the model is that the independence of irrelevant alternatives axiom is implausible for alternative sets containing choices that are close substitutes. An example illustrates this point. Suppose a population faces the alternatives of travel by auto and by bus, and two-thirds choose to use auto. Suppose now a second "brand" of bus travel is introduced that is in all essential respects the same as the first. Intuitively, two-thirds of the population will still choose auto, and the remainder will split between the bus alternatives. However, if the selection probabilities

satisfy Axiom 1 [IIA; Luce's axiom], only half the population will use auto when the second bus is introduced. The reason this is counter-intuitive is that we expect individuals to lump the two bus alternatives together in making the auto-bus choice. This example suggests that application of the model should be limited to situations where the alternatives can plausibly be assumed to be distinct and weighted independently in the eyes of each decision-maker.

This is what most people have in mind when they hear about IIA. But what they often don't realize is that this was made very explicit when the model was made popular by McFadden.

Finally, we can indeed verify that the ratio of two choice probabilities does indeed not depend on the choice set:

$$\frac{\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ})}{\Pr(y_i = k|\bar{u}_{i1},\ldots,\bar{u}_{iJ})} = \frac{\exp(\bar{u}_{ij})}{\sum_{j'=1}^{J}\exp(\bar{u}_{ij'})} \Bigg/ \frac{\exp(\bar{u}_{ik})}{\sum_{j'=1}^{J}\exp(\bar{u}_{ij'})} = \frac{\exp(\bar{u}_{ij})}{\exp(\bar{u}_{ik})}.$$

5.3.4 An alternative Starting Point for the Derivation

An alternative is to start with

$$\begin{aligned}
(5.3.11)\qquad & \Pr(y_i = 1|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) \\
&= \Pr(u_{i1} \geq u_{i2}, u_{i1} \geq u_{i3},\ldots,u_{i1} \geq u_{iJ}) \\
&= \Pr\left(\varepsilon_{i2} \leq \bar{u}_{i1} + \varepsilon_{i1} - \bar{u}_{i2}, \varepsilon_{i3} \leq \bar{u}_{i1} + \varepsilon_{i1} - \bar{u}_{i3},\ldots,\varepsilon_{i3} \leq \bar{u}_{i1} + \varepsilon_{i1} - \bar{u}_{iJ}\right) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{i2}}\int_{-\infty}^{\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{i3}}\cdots\int_{-\infty}^{\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{iJ}} \\
&\qquad f_{\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3},\ldots,\varepsilon_{iJ}}(\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3},\ldots,\varepsilon_{iJ})\, d\varepsilon_{iJ}\cdots d\varepsilon_{i3}\, d\varepsilon_{i2}\, d\varepsilon_{i1}.
\end{aligned}$$

Independence of the ε_{ij} from one another implies that we can replace $f_{\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3},\ldots,\varepsilon_{iJ}}$ in (5.3.11) by a product of marginal distributions,

$$f_{\varepsilon_{i1}}(\varepsilon_{i1})\cdot f_{\varepsilon_{i2}}(\varepsilon_{i2})\cdot f_{\varepsilon_{i3}}(\varepsilon_{i3})\cdots f_{\varepsilon_{iJ}}(\varepsilon_{iJ}).$$

Using

$$\int_{-\infty}^{\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{ij'}} f_{\varepsilon_{ij'}}(\varepsilon_{ij'})\,d\varepsilon_{ij'} = F_{\varepsilon_{ij'}}(\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{ij'})$$

and rearranging gives

$$\begin{aligned}\Pr(y_i = j|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) &= \int_{-\infty}^{\infty} \prod_{j'\neq j} F_{\varepsilon_{ij'}}(\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{ij'})\, f_{\varepsilon_{ij}}(\varepsilon_{ij})\,d\varepsilon_{ij} \\ &= \int_{-\infty}^{\infty} \left(\prod_{j'\neq j}^{J} \exp(-\exp(-(\bar{u}_{ij}+\varepsilon_{ij}-\bar{u}_{ij'})) \right) \exp(-\varepsilon_{ij})\cdot\exp(-\exp(-\varepsilon_{ij}))\,d\varepsilon_{ij}\end{aligned}$$

This is equal to (5.3.6) and from there one can proceed as before.

5.3.5 Imposing Structure on the Utility Function

In practice, we will impose structure on the utility functions. As in the binary case in Section 5.1.3, we can specify that the utility i receives when choosing alternative j is a function of characteristics of j when chosen by i, a $1\times K_1$ vector z_{ij}, characteristics of the choice situation, a $1\times K_2$ vector w_i, which includes characteristics of the decision maker or the environment, and an alternative and individual specific unobserved error term, ε_{ij}. This error term captures unobserved characteristics of the alternatives and preference heterogeneity across decision makers. This shows that the multinomial choice model is a generalization of the additive random utility model that underlies the binary choice model,

$$y_i = 1\{z_{i1}\alpha_1 - z_{i0}\alpha_0 + w_i(\gamma_1-\gamma_0) + (\varepsilon_{i1}-\varepsilon_{i0}) > 0\}.$$

To simplify the notation define a $1\times JK_2$ vector $\tilde{w}_{ij}$ that consists of J blocks of size $1\times K_2$. All elements are zero except for the jth block which is given by w_i, that is

$$\tilde{w}_{ij} \equiv \begin{pmatrix} 0 & \cdots & 0 & w_i & 0 & \cdots & 0 \end{pmatrix}.$$

Define

$$x_{ij} \equiv \begin{pmatrix} z_{ij} & \tilde{w}_{ij} \end{pmatrix}$$

and denote the parameters on z_{ij} by α and the ones on w_i for alternative j by γ_j. Here, we assume that α is alternative invariant, see the discussion in Section 5.1.3.

Last, define a $1 \times JK_2$ vector containing the coefficients on w_i for all alternatives,

$$\gamma' \equiv \begin{pmatrix} \gamma_1 & \cdots & \gamma_J \end{pmatrix}$$

and write $\beta \equiv (\alpha', \gamma')'$. β is a $K_1 + JK_2$-vector. Then,

$$\tilde{w}_{ij}\gamma = \begin{pmatrix} 0 & \cdots & 0 & w_i & 0 & \cdots & 0 \end{pmatrix} \cdot \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_{j-1} \\ \gamma_j \\ \gamma_{j+1} \\ \vdots \\ \gamma_J \end{pmatrix} = w_i\gamma_j.$$

Using this notation we can specify the functional form of the utility function in terms of x_{ij} and β. Here, we will focus on the additive utility function

$$u_{ij} = x_{ij}\beta + \varepsilon_{ij} = z_{ij}\alpha + w_i\gamma_j + \varepsilon_{ij}. \tag{5.3.12}$$

The model is that alternative j is chosen by i if the utility from choosing j is at least as high as the utility from choosing all j' in the choice set C, that is

$$y_i = j \text{ if } u_{ij} \geq u_{ij'} \text{ for all } j' \in C.$$

This shows that, as in the binary choice model, data are only informative about the sign of utility differences. To see this take $y_i = 1$ as the base alternative and observe that it follows from (5.3.12) that for each j

$$u_{ij} \geq u_{i1}$$

is equivalent to

$$(z_{ij} - z_{i1})\alpha + w_i(\gamma_j - \gamma_1) + \varepsilon_{ij} - \varepsilon_{i1} \geq 0.$$

Hence, we need to impose that z_{ij} does not include a constant because only the coefficient on the difference $z_{ij} - z_{ij'}$ is identified.[16] Moreover, we need to impose that one γ_j, say γ_1, is equal to zero because the coefficient on w_i is given by the difference between γ_j and γ_1.

[16] If we instead allow α to vary across alternatives then we need to impose that the constant term is zero for one of the alternatives.

5.3.6 Marginal Effects in the Multinomial Logit Model

Individual marginal effects can be obtained from coefficient estimates. Like for the binary choice model we can then average across individuals to get estimates of population average marginal effects or obtain marginal effects for an average individual right away. In industrial organization, elasticities are sometimes calculated instead. This is straightforward once the marginal effects have been obtained.

In the following, we derive marginal effects for a utility function of the type described in Section 5.3.5. It is possible to relax this, at some notational cost. It is convenient to define $p_{ij} \equiv \Pr(y_i = j|x_i)$. Then, we have

$$p_{ij} = \frac{\exp(z_{ij}\alpha + w_i\gamma)}{\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)}.$$

Hence,

$$\frac{\partial p_{ij}}{\partial z_{ij}} = \frac{\exp(z_{ij}\alpha + w_i\gamma)\alpha \left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)}{\left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2} - \frac{\exp(z_{ij}\alpha + w_i\gamma)\exp(z_{ij}\alpha + w_i\gamma)\alpha}{\left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2}.$$

But this is just equal to

$$p_{ij}\alpha - p_{ij}^2\alpha = p_{ij}(1 - p_{ij})\alpha.$$

For the derivative with respect to z_{ik} we get

$$\frac{\partial p_{ij}}{\partial z_{ik}} = -\frac{\exp(z_{ij}\alpha + w_i\gamma)\exp(z_{ik}\alpha + w_i\gamma)\alpha}{\left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2} = -p_{ij}p_{ik}\alpha.$$

Finally,

$$\frac{\partial p_{ij}}{\partial w_i} = \frac{\exp(z_{ij}\alpha + w_i\gamma)\gamma_j \left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)}{\left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2} - \frac{\exp(z_{ij}\alpha + w_i\gamma)\left(\sum_{j'' \in C} \exp(z_{ij''}\alpha + w_i\gamma)\gamma_j\right)}{\left(\sum_{j' \in C} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2}.$$

This is equal to

$$p_{ij}\gamma_j - p_{ij}\sum_{j''\in C} p_{ij}\gamma_j = p_{ij}\left(\gamma_j - \sum_{j''\in C} p_{ij}\gamma_j\right).$$

To summarize, the marginal effects are given by

$$\frac{\partial p_{ij}}{\partial z_{ik}} = \begin{cases} p_{ij}(1-p_{ik})\alpha & \text{if } j = k \\ -p_{ij}p_{ik}\alpha & \text{otherwise} \end{cases}$$

and

$$\frac{\partial p_{ij}}{\partial w_i} = p_{ij}\left(\gamma_j - \sum_{l\in C} p_{il}\gamma_l\right).$$

Notably, the sign of the effect of alternative specific regressors on the respective choice probability of that alternative is equal to the sign of the coefficient, whereas for alternative invariant regressors this is not necessarily true. For the latter it depends on the relative size of γ_j and the average γ_l, $\sum_{l\in C} p_{il}\gamma_l$, where we average across all alternatives. This corresponds nicely to the finding that in the ordered choice model the sign of the marginal effect depends on the value of the explanatory variables.

Finally, notice that from these results we get the binary logit marginal effects in the special case of $\alpha_0 = \alpha_1$ using the normalization $\gamma_0 = 0$. They are

$$\frac{\partial p_{i1}}{\partial (z_{i1} - z_{i0})} = p_{i1}(1-p_{i1})\alpha$$

and

$$\frac{\partial p_{i1}}{\partial w_i} = p_{i1}(1-p_{i1})\gamma_1.$$

5.3.7 Welfare Analysis

In the multinomial logit model it is particularly easy to conduct a welfare analysis. Suppose that we are interested in the effect of a change in the characteristics of all goods. We stack into a big vector x_i and consider a change from $\dot{x}_i = (\dot{x}'_{i1}, \ldots, \dot{x}'_{iJ})$ to $\ddot{x}_i = (\ddot{x}'_{i1}, \ldots, \ddot{x}'_{iJ})$. In economics, one of the most important characteristic of an alternative is often its price. But the change in characteristics could be anything, for example the

reduction of the time it takes to travel between two cities if you take the train. A welfare measure of this is the monetary amount by which i needs to be compensated, or that i would be willing to give up, such that she is as well off with characteristics $\ddot{x}_i$ as she is with characteristics $\dot{x}_i$. This is the so-called compensating variation (CV).

Denote utility that is received when characteristics are x_i and i receives the amount CV by $V(\ddot{x}_i, CV)$. CV is then implicitly defined by

$$\max_j V(\dot{x}_i, 0) = \max_j V(\ddot{x}_i, CV). \tag{5.3.13}$$

If CV is negative then i would be willing to pay for the change in characteristics, which seems reasonable in the context of a favorable change in alternative varying characteristics.

The error terms are unobserved, and therefore we usually consider $\mathbb{E}[CV]$ to be a reasonable measure for the change in welfare. Once we have estimated a multinomial choice model with utilities $\bar{u}_{ij} + \varepsilon_{ij}$ and choice set S this can be estimated. The basis for this is the insight that, if the ε_{ij}'s are distributed type 1 extreme value independently from one another, then maximal utility is itself distributed type 1 extreme value, as

$$\begin{aligned}
\Pr\left(\max_{j\in S} u_{ij} \leq u\right) &= \Pr(u_{i1} \leq u, \ldots, u_{iJ} \leq u) \\
&= \prod_{j\in S} \Pr(\varepsilon_{ij} \leq u - \bar{u}_{ij}) \\
&= \prod_{j\in S} \exp(-\exp(-u + \bar{u}_{ij})) \\
&= \exp\left(-\sum_{j\in S} \exp(-u + \bar{u}_{ij})\right) \\
&= \exp\left(-\exp\left\{-u + \log\left(\sum_{j\in S} \exp(\bar{u}_{ij})\right)\right\}\right)
\end{aligned}$$

is a type 1 extreme value random variable with mean equal to Euler's constant, $\gamma \approx 0.5772\ldots$, plus

$$I = \log\left(\sum_j \exp(\bar{u}_{ij})\right). \tag{5.3.14}$$

I is also referred to as the "inclusive value" of the choice set S that will also play a role in the nested logit model in Section 5.3.10 below. It can be calculated once we specify $\bar{u}_{ij}$ and estimate its parameters. Combining (5.3.13) with (5.3.14) gives that for the multinomial logit model, the change in the expected maximal utility that is due to a change in the characteristics from $\ddot{x}_{ij}$ to $\dot{x}_{ij}$ is

$$\log\left(\sum_{j=1}^{J}\exp(\ddot{x}_{ij}\beta)\right)-\log\left(\sum_{j=1}^{J}\exp(\dot{x}_{ij}\beta)\right).$$

Let α be the coefficient on "money", which in economic models of multinomial choice would usually be alternative specific because it is (discounted expected) lifetime income minus the price of alternative j. Then, the expected compensating variation would be $1/\alpha$ times this quantity, or

$$\frac{1}{\alpha}\left\{\log\left(\sum_{j=1}^{J}\exp(\ddot{x}_{ij}\beta)\right)-\log\left(\sum_{j=1}^{J}\exp(\dot{x}_{ij}\beta)\right)\right\}.$$

Let I_i be (discounted lifetime) income and p_j be the price of good j. Then, $I_i - p_j$ is a characteristic of alternative j if chosen by i and hence enters z_{ij}. The underlying idea here is that a price change should affect utility in the same way as an income change. For example, i should be exactly as well off as before if all prices increase by 100, but at the same time income increases by 100.

Let the coefficient on $I_i - p_j$ be α_1. Then, we have for the multinomial logit model model

$$\mathbb{E}[CV]=\frac{1}{\alpha_1}\left\{\ln\left(\sum_{j=1}^{J}\exp(\ddot{x}_{ij}\beta)\right)-\ln\left(\sum_{j=1}^{J}\exp(\dot{x}_{ij}\beta)\right)\right\}.$$

For other models, this can be simulated, by drawing ε_{ij}'s, calculating the corresponding compensating variation, and repeating this many times. An estimate of the expected compensating variation is then given by the average of the simulated ones.

5.3.8 Testing for Violations of IIA

McFadden (1974a), when introducing the multinomial logit model, pointed out that the IIA property of the model is strong such "that application of the model should be

limited to situations where the alternatives can plausibly be assumed to be distinct and weighted independently in the eyes of each decision-maker". Sometimes, it is not clear whether this is the case. In those situations, one can test for the model's validity using the Hausman and McFadden (1984) test.

The idea is that when we have individuals whom we observe to chose one of the alternatives in the subset S of alternatives, then the ratio of two choice probabilities should not depend on the subset—very much like in Luce (1959) as we have seen in Section 5.3.2.

Hausman and McFadden (1984) suggest to implement this by specifying a utility function as in Section 5.3.5 and then estimating the parameter vector β in subsamples in which choices are made out of the subset S of alternatives. One of these subsamples could of course be the whole sample. If the multinomial logit model is correctly specified, then the estimates of the slope coefficients in β will always be the same (up to estimation error), which is the null hypothesis of the test.[17]

5.3.9 Mixed Logit Model

It is possible to allow the coefficient vector to be a vector of random variables β_i with a joint distribution, also across alternatives, which is specified up to a finite set of parameters. For example, this parameter vector includes the mean of β_i and the corresponding variance-covariance matrix. Then, instead of the multinomial logit probability (5.3.10), we have a similar expression conditional on β_i,

$$\Pr(y_i = j|x_{ij}, \beta_i) = \frac{\exp(x_{ij}\beta_i)}{\sum_{j' \in C} \exp(x_{ij'}\beta_i)}.$$

The expectation thereof is a function of unknown parameters and replaces (5.3.10) in the likelihood function. In practice, this integral can be simulated, as explained in Section 4.6.

As an alternative, Chesher and Santos Silva (2002)—who also provide a nice overview over the literature—derive a probability model which approximates this random coefficient logit model. This is appealing because the model can then be estimated using

[17] This is related to our discussion of the binary logit model when there is choice-based sampling. Also here, the slope coefficients will not be affected when there is oversampling of observations where individuals chose $y_i = 0$ or $y_i = 1$.

standard software, such as Stata, for the multinomial logit model. The difference is that "mean utilities" are then specified as

$$\bar{u}_{ij} = x_{ij}\beta + \sum_{s=1}^{J}\sum_{t=s}^{J} \omega_{st} z_i^{st}(x_i, \beta).$$

Here, it is important that in the double sum we have that $s,t \neq i^*$, where i^* is the base alternative that is picked for normalization. The coefficients ω_{st} on the additional variables $z_i^{st}(x_i,\beta)$ correspond to variances and covariances of the random coefficients in the random coefficient logit model. The additional variables are

$$z_i^{st}(x_i,\beta) = \begin{cases} \frac{1}{2} - p_{it} & \text{if } i \text{ has chosen } s \text{ and } s = t \\ -p_{is} & \text{if } i \text{ has chosen } s \text{ and } s \neq t \\ -p_{it} & \text{if } i \text{ has chosen } t \text{ and } s \neq t \\ 0 & \text{if } i \text{ has neither chosen } s \text{ nor } t. \end{cases}$$

They depend on x_i and the parameter vector β because the choice probabilities p_{it} depend on those variables. These can be estimated in a first step, ignoring the random coefficients. The resulting error is then part of the overall approximation error.

5.3.10 Nested Logit Model

In the nested logit model—another model attributed to McFadden (1978)—, alternatives are grouped into mutually exclusive nests B_s, $s = 1,\ldots,S$. There could also be multiple layers of nests, which gives a tree structure as in the application in Goldberg (1995). But still, an alternative would ultimately only be in one nest. The nesting structure is assumed to be known *a priori*. Here, for simplicity, we will only discuss the case of one layer of nests.

Choice within a nest is—as before—described by a variant of the multinomial logit model with choice probability

$$\Pr(y_i = j | x_{i1},\ldots,x_{iJ}, y_i \in B_s) = \frac{\exp(x_{ij}\beta/\rho_s)}{\sum_{j' \in B_s} \exp(x_{ij'}\beta/\rho_s)}.$$

The difference is that we now have coefficients β/ρ_s, which means that the scaling of the coefficients is different across nests. These scaled coefficients β/ρ_s (but not β and

ρ_s separately) can be estimated using a conventional multinomial logit model within the nest.

ρ_s is called dissimilarity parameter for reasons that will become clear below. It is usually assumed to lie between zero and one.[18]

Choice between nests B_s will be driven by differences in the expected maximal utility across nests. The latter is equal to Euler's constant $\gamma = 0.5772\ldots$, plus the same inclusive value as in Section 5.3.7,

$$I_{is} = \log\left(\sum_{j\in B_s} \exp(x_{ij}\beta/\rho_s)\right). \tag{5.3.15}$$

Only now, it depends on the re-scaled coefficients β/ρ_s that we can estimate from choice behavior within nests.

On top of that there may be characteristics q_{is} with alternative-invariant coefficients δ that are common to all alternatives in nest B_s and will therefore not affect choice within the nest. But they may influence choice between nests next to differences in expected maximal utilities.

There is a nice similarity between the nested logit model and a Russian Matryoshka doll. This is a set of set of wooden dolls of decreasing size placed one inside the other. In the nested logit model, the probability that nest s is chosen is again a multinomial logit model with

$$\Pr(y_i \in B_s | I_{i1},\ldots,I_{iS},q_{i1},\ldots,q_{iS}) = \frac{\exp(q_{is}\delta + \rho_s I_{is})}{\sum_{s'=1}^{S}\exp(q_{is'}\delta + \rho_{s'} I_{is'})}, \tag{5.3.16}$$

where q_{is} are the nest specific variables described before.

5.3.10.1 An example

For an example consider a situation in which an individual is choosing between buying Coca Cola (C), Pepsi (P), and No-Name (N) cola. There is one nest of branded colas

[18] McFadden (1978) shows that a necessary and sufficient condition for the model to be consistent with utility maximization is that dissimilarity parameters lie in the unit interval for each nest. Börsch-Supan (1990) argues that this requirement is too strong and provides a re-interpretation. More generally, McFadden and Richter (1971) and follow-ups on this paper provide testable implications of models with random utility maximization.

(B) that contains C and P, and another nest containing just N, a trivial nest.

Suppose first that mean utility (what has been denoted by $x_{ij}\beta$ above) is 10 for each of those three colas. Then, if $\rho_B = 1$ the nesting structure becomes irrelevant and we get—just as we would in the multinomial logit model—that the choice probability is $1/3$ for either of the three colas. This can be seen in the last row of the upper panel of Table 5.2.

Next, turning to the trivial nest that only contains the no-name cola, observe that we can set the corresponding ρ_N to any value we like (except for zero); it'll always cancel out as can be seen from (5.3.15) and (5.3.16). The reason is that in (5.3.15) we will only sum over one element and therefore the log will cancel with the exponential function, leaving $x_{ij}\beta/\rho_N$. Once we multiply this by ρ_N in (5.3.16) we will always get $x_{ij}\beta$, no matter what the value of ρ_N is. Put differently, the value of the nest is given by $x_{ij}\beta + \varepsilon_N$ and the expected maximal utiltiy is given by $x_{ij}\beta$ plus the expectation of ε_N, which is Euler's constant.

Now consider choice between the non-branded and the two branded colas. Mean utility is 10 for each of the three alternatives, but because of ρ_B, in the eyes of the decision maker the branded colas, C and P, become the more similar to one another the smaller ρ_B. The word "similar" here means that choice is determined less and less by the taste shocks the smaller ρ_B. This is because when choosing between the two branded colas within nest B the average utility that is relevant for that choice, $10/\rho_B$, will dominate choice the more the smaller ρ_B, because the contribution of the taste shocks stays the same while the contribution of the average utility increases. To see this consider first the case in which ρ_B is very close to zero. Then, the random utility from choosing C within the nest, $10/\rho_B + \varepsilon_C$, is always very close to the random utility from choosing P, $10/\rho_B + \varepsilon_P$, no matter what the taste shocks are, so that the random utility from choosing the nest of branded colas, B, does not depend on whether it actually contains only one of them or both. Therefore, is also equal to 10, or formally ρ_B times utility from choice within the nest,$10/\rho_B$, when we ignore the expectation of the taste shock for the moment. We show in Section 5.3.10.2 below for a more thorough look.

Looking at the table, we have that the probability to choose one of the two branded colas is always $1/2$ conditional on choosing among the two. However, as we have just seen, the probability to choose B goes to, so that then the overall choice probability is $1/4$ for each of the branded colas and $1/2$ for N. This shows why ρ_B is called

dissimilarity parameters: the higher the value of ρ_B, the less similar are alternatives within a nest relative to alternatives in other nests. Put differently, the lower ρ_B, the closer the expected maximimal utility is to the maximal expected utility for alternatives that are in nest B.

Table 5.2: Cola example

mean utility is always 10				
ρ_B	$\Pr(B)$	$\Pr(N)$	$\Pr(C)$	$\Pr(P)$
0.1	0.52	0.48	0.26	0.26
0.2	0.53	0.47	0.27	0.27
0.3	0.55	0.45	0.28	0.28
0.4	0.57	0.43	0.28	0.28
0.5	0.59	0.41	0.29	0.29
0.6	0.60	0.40	0.30	0.30
0.7	0.62	0.38	0.31	0.31
0.8	0.64	0.36	0.32	0.32
0.9	0.65	0.35	0.33	0.33
1	0.67	0.33	0.33	0.33

mean utility is 11 for C, 10.5 for P, and 10 for N				
ρ_B	$\Pr(B)$	$\Pr(N)$	$\Pr(C)$	$\Pr(P)$
0.1	0.73	0.27	0.73	0.00
0.2	0.73	0.27	0.68	0.06
0.3	0.74	0.26	0.62	0.12
0.4	0.75	0.25	0.58	0.17
0.5	0.76	0.24	0.56	0.20
0.6	0.77	0.23	0.54	0.23
0.7	0.78	0.22	0.53	0.26
0.8	0.79	0.21	0.52	0.28
0.9	0.80	0.20	0.51	0.29
1	0.81	0.19	0.51	0.31

The lower panel of the same table shows what the effect is once mean utilities are different. Then, the lower ρ_B, the higher the probability that the alternative with the highest mean utility is chosen within the nest of branded colas.

Finally, notice that IIA holds by construction when $\rho = 1$, because then the model

is the standard multinomial logit model. However, the model is able to produce choice patterns that do not obey IIA for $\rho < 1$. Intuitively, when ρ gets smaller, then the ratio between the probability to choose C and N depends on whether an alternative P is available.

5.3.10.2 A more detailed look

The model is called "nested" logit because choice between the nests is again given by a multinomial logit model. This suggest that choice is made in a sequential way: first, a decision maker chooses which nest he will pick an alternative from, and only then he will choose among the alternatives in that nest. This is slightly misleading, as the choice probability will be exactly the same if he chooses from all alternatives from the start. To show this, define

$$\varepsilon_{is} \equiv \max_{j \in B_s} \left\{ x_{ij}\beta/\rho_s + \varepsilon_{ij} \right\} - I_{is} = \max_{j \in B_s} \left\{ x_{ij}\beta/\rho_s + \varepsilon_{ij} - I_{is} \right\}.$$

This is the difference between the actual maximal utility and the expected maximal utility (the latter net of Euler's constant) when i chooses among alternatives in nest s. To show (5.3.16), we need to show that ε_{is} is type 1 extreme value, that is that

$\Pr(\varepsilon_{is} \leq e) = \exp(-\exp(-e))$. Denoting the logical "for all" by $\forall$, we have

$$\begin{aligned}
\Pr(\varepsilon_{is} \leq e) &= \Pr(x_{ij}\beta/\rho_s + \varepsilon_{ij} - I_{is} \leq e \,\forall\, j \in B_s) \\
&= \Pr(\varepsilon_{ij} \leq e - x_{ij}\beta/\rho_s + I_{is} \,\forall\, j \in B_s) \\
&= \prod_{j \in B_s} \Pr(\varepsilon_{ij} \leq e - x_{ij}\beta/\rho_s + I_{is}) \\
&= \prod_{j \in B_s} \exp(-\exp(-e + x_{ij}\beta/\rho_s - I_{is})) \\
&= \prod_{j \in B_s} \exp(-\exp(-e)\exp(x_{ij}\beta/\rho_s)\exp(-I_{is})) \\
&= \exp\left(-\exp(-e)\left(\sum_{j \in B_s} \exp(x_{ij}\beta/\rho_s)\right)\exp(-I_{is})\right) \\
&= \exp\left(-\exp(-e)\left(\sum_{j \in B_s} \exp(x_{ij}\beta/\rho_s)\right)\exp(I_{is})^{-1}\right) \\
&= \exp\left(-\exp(-e)\left(\sum_{j \in B_s} \exp(x_{ij}\beta/\rho_s)\right)\left(\sum_{j \in B_s} \exp(x_{ij}\beta/\rho_s)\right)^{-1}\right) \\
&= \exp(-\exp(-e)))
\end{aligned}$$

where the first equality follows from the definition of ε_{is}, the second equality from rearranging terms, the third equality is due to the independence of ε_{ij} across alternatives, the fourth equality from ε_{ij} being type 1 extreme value, that is from (5.3.8), the fifth and sixth, respectively, from using the fact that the exponential function applied to a sum is equal to the product of exponential functions of the elements of the sum, and the eighth from (5.3.15). Finally, (5.3.16) follows from the observation that the utility associated with choosing nest s is $\rho_s I_{is} + \varepsilon_{is}$, where ε_{is} is type 1 extreme value, as has just been shown.

Cardell (1997) shows that an alternative interpretation of the nested logit model is that it is a "variance components structured multinomial logit model". Following Berry (1994), Cardell (1997) shows that nested logit probabilities can be derived from a model in which

$$u_{ij} = x_{ij}\beta + \zeta_{is} + \rho_s \varepsilon_{ij},$$

where ε_{ij} is distributed type 1 extreme value and the distribution of ζ_{is} depends on ρ_s and is such that $\zeta_{is} + \rho_s \varepsilon_{ij}$ is distributed type 1 extreme value as well. $1 - \rho_s$ has the interpretation of a correlation of $\zeta_{is} + \rho_s \varepsilon_{ij}$ across j within the same nest.[19]

Moving on, it follows from the above derivation that the probability that i chooses alternative $j \in B_s$ is

$$p_{ij} \equiv \frac{\exp\left(\frac{z_{ij}\alpha + w_i\gamma_j}{\rho_s}\right)}{\sum_{j' \in B_s} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma}{\rho_s}\right)} \cdot \frac{\left(\sum_{j' \in B_s} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma_{j'}}{\rho_{s'}}\right)\right)^{\rho_{s'}}}{\sum_{s'} \left(\sum_{j' \in B_{s'}} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma_{j'}}{\rho_{s'}}\right)\right)^{\rho_{s'}}}.$$

Observe that if $\rho_{s'} = 1$ for all s', then the denominator in the first fraction cancels with the numerator in the second fraction. In that case, the double sum in the denominator of the second fraction, over nests and then alternatives within nests, will be the same as the usual summation over all alternatives in the entire choice set.

Alternatively, we can write this probability as

$$p_{ij} = p_{ij|s} \cdot p_{is},$$

where

$$p_{ij|s} \equiv \frac{\exp\left(\frac{z_{ij}\alpha + w_i\gamma_j}{\rho_s}\right)}{\sum_{j' \in B_s} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma_{j'}}{\rho_s}\right)}$$

is the probability that i chooses j out of the alternatives in choice set B_s and

$$p_{is} \equiv \frac{\left(\sum_{j' \in B_s} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma_{j'}}{\rho_{s'}}\right)\right)^{\rho_{s'}}}{\sum_{s'} \left(\sum_{j' \in B_{s'}} \exp\left(\frac{z_{ij'}\alpha + w_i\gamma_{j'}}{\rho_{s'}}\right)\right)^{\rho_{s'}}}$$

is the probability that i chooses one of the alternatives in s.

[19] Formally, in the paper, Cardell (1997) shows the error terms in the nested logit model can be written as

$$\nu_{1,ij} + \lambda_{1,ij}\left(\nu_{2,ij} + \lambda_{2,ij}\left(\nu_{3,ij} + \cdots\right)\right),$$

where the $\nu_{\ell,ij}$ are distributed as what he calls $C(\lambda)$, with parameter λ and density $f(\nu;\lambda) = (1/\lambda)\sum_{n=0}^{\infty}[((-1)^n \exp(-n\nu)/(n!\Gamma(-\lambda n))]$.

5.3.10.3 Marginal Effects

One can show that the marginal effects of changes in alternative varying regressors are

$$\frac{\partial p_{ij}}{\partial z_{ij}} = p_{ij}\left(\frac{1}{\rho_s} - p_{ij} - \frac{1-\rho_s}{\rho_s}\, p_{ij|s}\right)\alpha$$

and

$$\frac{\partial p_{ij}}{\partial z_{ik}} = \begin{cases} -p_{ij}\left(p_{ik} + \frac{1-\rho_s}{\rho_s}\, s_{k|g}\right)\alpha & \text{for } j \neq k \text{ and } k \in B_s \\ -p_{ij}p_{ik}\alpha & \text{for } j \neq k \text{ and } k \notin B_s \end{cases},$$

where p_{ij} and $p_{ij|s}$ are as defined in Section 5.3.10 above. Notice that for $\rho_s = 1$ they are equal to the ones in the multinomial logit model in Section 5.3.6.

5.3.11 Generalized Extreme Value Taste Shocks

The nested logit model, it turns out, is a particular member of class of models that allow for violations of IIA while preserving some of the computational properties of the multinomial logit model, which is because they are still analytically tractable. In those models, the distribution of the unobservables ε_{ij} is a member of a more general class of distributions, termed generalized extreme value distributions. Here, I follow McFadden (1977, 1978, 1984, 1981) and Small (1987).

In those models, there is a function $G(\exp(\varepsilon_{i1}), \ldots, \exp(\varepsilon_{iJ}))$ defining the c.d.f.

$$F_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}(\varepsilon_{i1}, \ldots, \varepsilon_{iJ}) = \exp\left(-G(\exp(\varepsilon_{i1}), \ldots, \exp(\varepsilon_{iJ}))\right)$$

whose marginal distribution with respect to each ε_{ij} is the extreme value distribution

$$\begin{aligned} \frac{\partial F_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}(\varepsilon_{i1}, \ldots, \varepsilon_{iJ})}{\partial \varepsilon_{ij}} &= \lim_{\varepsilon_{ik}\to\infty, k\neq j} F_{\varepsilon_{i1},\ldots,\varepsilon_{iJ}}(\varepsilon_{i1}, \ldots, \varepsilon_{iJ}), \\ &= \exp\left(-G(0, \ldots, 0, 1, 0, \ldots, 0)\exp(-\varepsilon_{ij})\right) \end{aligned}$$

where $G(0, \ldots, 0, 1, 0, \ldots, 0)$ is the function G with all but the jth argument evaluated at zero and the jth argument evaluated at 1. McFadden (1978) provides conditions on this function G such that

$$\Pr(y_i = j) = \exp(\bar{u}_{ij}) \cdot \frac{\partial G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3})) \Big/ \partial(\exp(\bar{u}_{ij}))}{G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3}))}$$

defines a probabilistic choice model which is consistent with utility maximization. He terms it the generalized extreme value model. The conditions are that (i) $G(a_1,\ldots,a_J)$ is a nonnegative, (ii) homogeneous-of-degree-one function of $(a_1,\ldots,a_J) > 0$, that (iii)

$$\lim_{a_j\to\infty} G(a_1,\ldots,a_J) = +\infty$$

for $j = 1,\ldots,J$, that for any distinct $(j_1,\ldots,j_k)$ from $\{1,\ldots,J\}$ we have that (iv) the cross partial derivatives

$$\frac{\partial^k G}{\partial a_{j_1}\cdots\partial a_{j_k}}$$

have alternating signs in the sense that they are nonnegative if the order k is odd and non-positive if k is even.

He also shows that welfare in utility terms is given by

$$\mathbb{E}\left[\max_j \bar{u}_{ij} + \varepsilon_{ij}\right] = \gamma + \log\left(G\left(\exp(\bar{u}_{i1}),\exp(\bar{u}_{i2}),\exp(\bar{u}_{i3})\right)\right),$$

where, again, $\gamma = 0.5772\ldots$ is Euler's constant, and

$$\Pr(y_{ij}|\bar{u}_{i1},\ldots,\bar{u}_{iJ}) = \frac{\partial\mathbb{E}\left[\max_j \bar{u}_{ij} + \varepsilon_{ij}\right]}{\partial\bar{u}_{ij}}.$$

These results are useful because they allow us to tailor multinomial choice models to particular applications.

5.3.11.1 Multinomial and Nested Logit Model

The multinomial logit model has one of the simplest possible function

$$G^{MNL}(a_1,\ldots,a_J) = \sum_{j'} a_{j'},$$

and we use $a_j = \exp(\varepsilon_{j1})$.

The nested logit model uses

$$G^{NL}(a_1,\ldots,a_J) = \sum_{s=1}^{S}\left[\sum_{j'\in B_s} y_{j'}^{1/\rho_s}\right]^{\rho_s}$$

and we can immediately see that the multinomial logit model is obtained as the special case in which $\rho_s = 1$. Finally, the nested logit model with two layers of nests, a top layer with nests B_r and dissimilarity parameter ρ_t and a bottom layer with nests B_s, within those top layer nests, and dissimilarity parameters ρ_r has

$$G(a_1,\ldots,a_J) = \left[\sum_r \left(\sum_{j\in B_r} a_j^{1/\rho_r}\right)^{\rho_r/\rho_t}\right]^{\rho_t}.$$

5.3.11.2 An Example of a Nested Logit Model

Let us now look at an example with three alternatives. Omitting the conditioning on $\bar{u}_{i1},\bar{u}_{i2},\bar{u}_{i3}$ we have that in the nested logit model in which alternative 2 and 3 are in one nest, $B = \{2,3\}$ with dissimilarity parameter ρ,

$$\Pr(y_i = 1|, C = \{1,2,3\}) = \frac{\exp(\bar{u}_{i1})}{\exp(\bar{u}_{i1}) + [\exp(\bar{u}_{i2}/\rho) + \exp(\bar{u}_{i3}/\rho)]^{\rho}}. \tag{5.3.17}$$

Again, for $\rho = 1$, we get the standard multinomial logit probabilities. Here,

$$G(a_1,a_2,a_3) = a_1 + \left[a_2^{1/\rho} + a_3^{1/\rho}\right]^{\rho}, \tag{5.3.18}$$

with $a_j = \exp(\bar{u}_{ij})$, $j = 1,2,3$.

Let us first verify that G satisfies the conditions given above. For $a_j = \exp(\bar{u}_{ij})$, $j = 1,2,3$, the function is nonnegative, so (i) holds. It is (ii), homogeneous of degree one, as

$$\begin{aligned} G(\lambda a_1,\lambda a_2,\lambda a_3) &= \lambda a_1 + \left[(\lambda a_2)^{1/\rho} + (\lambda a_3)^{1/\rho}\right]^{\rho} \\ &= \lambda a_1 + \left[\lambda^{1/\rho}\cdot\left(a_2^{1/\rho} + a_3^{1/\rho}\right)\right]^{\rho} \\ &= \lambda\left(a_1 + \left[a_2^{1/\rho} + a_3^{1/\rho}\right]^{\rho}\right) \\ &= \lambda G(a_1,a_2,a_3). \end{aligned}$$

One can also see directly that condition (iii) holds, and for (iv) we have that the function is always positive if the first partial derivative is with respect to a_1 and zero thereafter,

so then the condition holds. If we first take the derivative with respect to, say, a_2, then we have

$$\begin{aligned}\frac{\partial G(a_1,a_2,a_3)}{\partial a_2} &= \rho\left[a_2^{1/\rho}+a_3^{1/\rho}\right]^{\rho-1}\cdot\left(\frac{a_2^{1/\rho-1}}{\rho}\right)\\ &= \left[a_2^{1/\rho}+a_3^{1/\rho}\right]^{\rho-1}\cdot a_2^{1/\rho-1}>0.\end{aligned}$$

The derivative thereof with respect to a_1 is zero, so then the condition holds. If we take the derivative with respect to a_3, then we have that this is

$$\frac{\partial^2 G(a_1,a_2,a_3)}{\partial a_2\partial a_3} = (\rho-1)\cdot\left[a_2^{1/\rho}+a_3^{1/\rho}\right]^{\rho-2}\cdot a_2^{(1/\rho)-1}\cdot\frac{a_2^{(1/\rho)-1}}{\rho}\leq 0,$$

provided that $0<\rho\leq 1$. The derivative thereof with respect to a_1 is zero. Hence, condition (iii) holds.

5.3.11.3 Ordered Generalized Extreme Value Model

The general results we have discussed above are useful because they allow us to tailor multinomial choice models to particular applications. An example is the ordered generalized extreme value model of Small (1987), where

$$G(a_1,\ldots,a_J)=\sum_{r=1}^{J+M}\left(\sum_{j\in B_r} w_{r-j}a_j^{1/\rho_r}\right)^{\rho_r}.$$

Here, M is a positive integer and ρ_r and w_m are constants satisfying $0<\rho_r\leq 1$ for $r=1,\ldots,J+M$, $w_m\geq 0$ for $m=0,\ldots,M$, and $\sum_{m=0}^{M} w_m=1$, and where

$$B_r=\{j\in\{1,\ldots,J\}|r-M\leq j\leq r\}.$$

There are $J+M$ subsets B_r of the overall choice set and they overlap. For example, with $M=2$ the subsets are $B_1=\{1\}$, $B_2=\{1,2\}$, $B_3=\{1,2,3\}$, $B_4=\{2,3,4\}$, $B_J=\{J-2,J-1,J\}$, $B_{J+1}=\{J-1,J\}$, $B_{J+2}=\{J\}$. This ensures that each alternative belongs to exactly $M+1$ different subsets.

Subsets achieve the same as the nests in a nested logit model. If two alternatives belong to the same subset B_r, then the correlation between their ε_{ij}'s will be positive provided that $\rho_r < 1$. Conversely, any two stochastic elements ε_{ij} and ε_{ik} are independent if $|j-k| > M$. ε_{ij} and ε_{ik} are the more correlated the smaller $j-k$. This is because the closer they are, the more subsets they have in common, provided that the relevant parameters ρ_s are strictly less than one.

The model is a multinomial logit model within each subset B_r, except that there are weights w_{r-j} such that the probability that i chooses j when choosing from the choice set B_r is

$$p_{ij|r} = \frac{w_{r-j} \exp\left(\bar{u}_{ij}/\rho_r\right)}{\sum_{j' \in B_r} w_{r-j'} \exp\left(\bar{u}_{ij'}/\rho_r\right)}.$$

However, since alternatives can belong to more than one subset, instead of aggregating those probabilities in a hierarchical way as it is done in the nested logit model, the model mixes over nests B_r in the sense that an alternative is a member of more than one nest. This gives for the probability that j is chosen

$$p_{ij} = \sum_{r=j}^{j+M} p_{ij|r} \cdot p_{ir},$$

where the probability that i chooses from subset r is

$$p_{ir} = \frac{\exp\left(\rho_r I_r\right)}{\sum_{r'=1}^{J+M} \exp\left(\rho_{r'} I_{r'}\right)},$$

with

$$I_r \equiv \sum_{j' \in B_r} w_{r-j'} \exp\left(\bar{u}_{ij'}/\rho_r\right).$$

5.3.12 Multinomial Probit Model and Other Generalizations

Our discussion so far has been on variants and generalizations of the multinomial logit model. These are motivated by the desire to allow for a correlation between the taste shocks ε_{ij} between the alternatives. Of course, in principle, it is also possible to perform maximum likelihood estimation using a multivariate normal distribution. For this one

has to resort to simulation-based methods because higher dimensional integrals are hard to compute. In fact, this led McFadden (1989) to develop the simulated method of moments. See Section 4.6.

Alternatively, indirect inference may be used, as described in Section (4.7.2). Section 9 in Gourieroux et al. (1993) is specifically on this topic.

5.3.13 Estimating Multinomial Choice Models from Market Level Data

Up to now, our focus has been on maximum likelihood estimation of multinomial choice models using micro level data. However, the common practice in industrial organization is to estimate those models from market level data. This is clearly described in the classic article by Berry (1994).

We will discuss this in the context of the demand for goods, say cars. Suppose, for simplicity, that we have data on market shares s_j for products $j = 1,\ldots,J$ with attributes x_j for just one market at one point in time. The idea is that the attributes of the cars, for example the weight of the car, are the same for all consumers in that market, hence the omission of the subscript i in x_j. In addition, there are unobserved attributes ξ_j that, for example, describe the attractiveness or desirability of a car model, and that mean utilities are

$$\bar{u}_{ij} = x_j\beta + \xi_j.$$

Suppose we also know the market size, M, and that each consumer buys at most once in the time period we have the data for. Then, market shares s_j are given by quantities divided by the market size.

5.3.13.1 Basic Idea

Assume that consumers choose according to a multinomial logit model and that they can also choose an additional alternative, not to buy any product at all. This is commonly referred to as the outside good. Impose the normalization that the mean utility derived from the outside good is zero. Then, market shares are equal to multinomial logit choice probabilities of the form (5.3.10),

$$s_j = \frac{\exp(x_j\beta + \xi_j)}{1 + \sum_{j'=1}^{J} \exp\left(x_{j'}\beta + \xi_{j'}\right)},$$

where the additional "1" in the denominator is the exponential function applied to the mean utility of the outside good. The market share of the outside good is

$$s_0 = \frac{1}{1 + \sum_{j'=1}^{J} \exp\left(x_{j'}\beta + \xi_{j'}\right)}.$$

Both s_j and s_0 are observed. The insight is now that

$$\log(s_j) - \log(s_0) = x_j\beta + \xi_j.$$

This means that when there are enough products in the market in the sense that J is at least equal to the number of elements in β, that then we can estimate β by regressing $\log(s_j) - \log(s_0)$ on x_j, treating ξ_j like an error term and normalizing its mean to zero.

Of course this example suggests that when car producers know that their car is in high demand because it is particularly attractive, that is if ξ_j is high, then they will also set a higher price, which enters x_j. This means that at least some components of x_j are endogenous in the sense that they are correlated with the error term ξ_j. For this, classic instrumental variables estimators can be used because the estimation equation is linear in the parameters. These instruments need to be related to price, but should not be related to unobserved characteristics. Instruments that have been used are cost shifters (Working, 1927)—see also Section 3.2—, pre-determined product characteristics and the number of products in the market (Berry et al., 1995), and geographical variation in prices (Hausman, 1996; Nevo, 2001).

5.3.13.2 Refinements

The assumption in the model outlined above is that individual utilities depend on a linear index $x_j\beta + \xi_j$. The approach to estimation, because of the endogeneity of price, is to invert the market shares and then estimate β using conventional methods. A more refined model with random coefficients is proposed by Berry et al. (1995), also termed "BLP". Their main contribution is to show how one can invert the market shares to back out the residual ξ_j when β is replaced by a vector of random coefficients. The next step is then to construct the GMM objective function for instrumental variables estimation.

Exercises

1. Show that the logistic distribution,

$$F_{\varepsilon_i}(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{1+\exp(\varepsilon_i)},$$

is symmetric about zero. Show also that the two logit models

$$y_i = 1\{x_i\theta \geq \varepsilon_i\}$$

and

$$y_i = 1\{x_i\theta + \varepsilon_i > 0\}$$

are observationally equivalent.

2. We know that the maximum likelihood estimator behaves well if the log likelihood is globally concave. A sufficient condition for this is that the Hessian is negative definite. In Section 5.1.9 this was shown for the probit model. Show it for the logit model.

3. Derive that in the logit model

$$\Pr(y_i = 1|x_i) = \frac{\exp(x_i\theta)}{1+\exp(x_i\theta)}$$

from a random utility model. Explicitly state the normalizations you use.

4. In Section 5.1.12, we have performed a Monte Carlo study for the binary logit model.

   (a) Here, the model has only one explanatory variable, years of education, and $x_i\beta$ is just the years of education times -0.1. Instead, x_i could also include a constant term so that β is a vector with a constant term and the same slope coefficient. Extend the likelihood function and the Monte Carlo so that it includes such a constant term, and set the constant term equal to -0.2 when generating the data. Adapt all the code and then run the Monte Carlo. Use this adapted code for the remaining exercises.

(b) Now we want to see what happens if we instead generate the data using a rescaled version of a probit model. For this, draw ε_i from a normal distribution with mean 0 and variance $\pi^2/3$ so that the scale is comparable. Then estimate the same logit model as before. Do you get very different coefficient estimates? Repeat this for a constant term equal to -1.5.

(c) Go back to generating the data using the logit model. Program the McFadden R^2 measure. What is its mean over the Monte Carlo replications?

(d) Within each Monte Carlo replication, calculate individual marginal effects and compare the average marginal effect to the marginal effect at average characteristics, respectively. What is the mean of those two across Monte Carlo replications respectively?

5. We obtain logit estimates in a choice-based sample consisting of all 444 individuals that chose option 1 and a random sample of 606 out of 7587 individuals who chose option 0. In the choice-based sample the model is

$$\Pr(\text{exit}_{it}|x_{it})^{\text{choice based}} = \frac{\frac{444/1050}{444/8031}\exp(x_{it}\theta)}{\frac{606/1050}{7587/8031}1+\frac{444/1050}{444/8031}\exp(x_{it}\theta)}.$$

(a) Show that we can use a standard logit model to estimate the slope coefficients in θ.

(b) Why can't we use the results to predict the probability to chose 1 for a randomly drawn individual in the full sample?

(c) How can we correct our estimates from the choice-based sample?

6. In Section 5.2.5, we have looked at an ordered probit model with a random coefficient.

(a) Extend the code so that it can handle so-called panel data, say 5 observations of the dependent and the independent variables per individual, assuming that the random coefficient will be the same in all periods. Then run a Monte Carlo analysis. Pay attention when putting together the likelihood function, because the log likelihood contribution of this individual will be the log of the average product of the per-period likelihood contributions.

(b) Here, random numbers are draws from a normal distribution using Matlab's built-in random number generator. Alternatively, one can use "draws" from a Halton quasi random point set with reverse-radix scrambling (Matlab does that for you, see `http://www.mathworks.nl/help/toolbox/stats/haltonset.html`). These "draws" have a uniform distribution, so to convert them into a "normally distributed draws" you have to apply the inverse of the normal distribution to them. Do this and show what is normally claimed, namely that one needs much fewer of those draws when simulating an expectation.

7. McFadden (1978) provides conditions on a function G such that

$$\Pr(y_i|) = \exp(\bar{u}_{j1}) \cdot \frac{\partial G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3})) \Big/ \partial(\exp(\bar{u}_{ij}))}{G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3}))}$$

defines a probabilistic choice model which is consistent with utility maximization. He also shows that

$$\mathbb{E}\left[\max_j u_{ij} + \varepsilon_{ij}\right] = \gamma + \log\left(G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3}))\right),$$

where $\gamma = 0.5772\ldots$ is Euler's constant, and

$$\Pr(y_{ij}|\bar{u}_{i1}, \ldots, \bar{u}_{iJ}) = \frac{\partial \mathbb{E}[\max_j u_{ij} + \varepsilon_{ij}]}{\partial u_{ij}}.$$

(a) What is G for the multinomial logit model with J alternatives?

(b) Verify this last equation, step by step, for the multinomial logit model.

(c) Now consider the nested logit model with three alternatives, in which alternative 2 and 3 are in a separate nest. Verify the last equation also for this case.

(d) For the nested logit model, what happens to the choice probabilities for all three choices when $\rho \to 0$? What when $\rho = 1$? Provide an intuition and explain.

8. Show that the multinomial logit model is the special case of the nested logit model when we set $\rho = 1$.

9. Consider the following setup. Individuals choose between three brands. Log prices are drawn from a normal distribution with means 2.9, 3 and 3, respectively, and variance-covariance matrix

$$\begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.2 & 0.1 \\ 0 & 0.1 & 0.2 \end{pmatrix}.$$

This means that we obtain prices from these log prices by applying the exponential function to these draws. The price coefficient in the utility function is $\beta_2 = -0.1$ and the intercept for the first alternative is $\beta_1 = -2$; for the other ones it is $\beta_1 = 0$. This is a setup in which alternative 1 is sold at a lower price, but people value it also less.

   (a) Program up a Monte Carlo study in Matlab, with 100 observations and $1,000$ Monte Carlo replications. In particular, generate the data from a multinomial logit model with type 1 extreme value errors. Then estimate the parameters $1,000$ times. What are their means across repetitions?

   (b) Program also up the estimate of the variance-covariance matrix of these estimates. This is the inverse of the average negative Hessian (which is an output argument to the maximization, see the code for the binary logit model). Compare the average estimate of this variance-covariance matrix to the variances of the actual estimates across replications.

Chapter 6

Censoring, Truncation, Sample Selection and Duration Analysis

6.1 Introduction

Censoring means that the outcome is only observed if it takes on certain values. However, covariates are still observed. An example for this are expenditures which are constrained to be nonnegative. for example for luxury goods they are zero, that is censored, for a sizable part of the population.[1]

We say that there is *truncation* if not only the outcome variable is unobserved if it takes on certain values, but also the covariates. In the expenditures example this means that we only observe individuals with positive expenditures.

Tobin (1958) was the first to propose what he called "a hybrid of probit analysis and multiple regression." The model has subsequently been generalized by Gronau (1973, 1974) and Heckman (1974). The generalization is that different factors are allowed to influence the decision to buy at all and how much to spend, or in their case the decision of a women whether to work and, if she decides to work, how many hours. Heckman (1976, 1979) noted that both in the case of censoring and truncation we face selected

[1] Duration analysis, which is not discussed in these lecture notes, is particularly prone to censoring because a duration has typically not ended at the moment the individual is observed.

samples. This selection is nonrandom if, for a given set of observable characteristics, individuals for whom we observe the outcome variable are fundamentally different from individuals for whom we do not observe the outcome variable.

For instance, market wages are observed only for women who decide to work. If characteristics of non-working women are still observed we face censoring. If we only observe women who work then we face truncation. In this example, it is plausible that the sample of women who work, and thereby the sample of women who don't, are nonrandomly selected because the decision to work might be related to potential wages even if we condition on observable characteristics.

In the remainder I discuss models for censoring and truncation from below. The case of censoring and truncation from above and combinations of the two are analogous.[2] In a well-known survey Amemiya (1984) classifies different variants of the model into 5 types. I discuss the type 1 and type 2 Tobit model. The type 3 and type 4 model are not discussed, essentially because both are straightforward extensions of the type 2 model. A generalization of the type 5 model is used for policy evaluation. There, we face two selected samples, the sample of treated and the sample of untreated individuals. An outcome is observed in either case and the treatment decision is made in light of both possible outcomes. This model is discussed in Section 6.3.

As for further readings, Kiefer (1988) is a well-known survey on duration models. Powell (1994) and van den Berg (2001) are recent overviews for the literature on censoring and truncation. Finally, Kalbfleisch and Prentice (2002) is a comprehensive reference for duration analysis.

6.2 Standard Tobit Model

6.2.1 Model

In the original model by Tobin (1958) there is a latent dependent variable

$$y_i^* = x_i\beta + \varepsilon_i,$$

and ε_i is normally distributed with mean 0 and variance σ^2.

[2] The notions "from below" and "from above" refer to censoring low and high values of the outcome variable, respectively.

In the case of *censoring* from below at 0 we have that y_i^* is observed if it exceeds 0, otherwise we observe 0, that is

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{otherwise.} \end{cases}$$

As we have already pointed out the model can be adapted to allow for censoring from above, for censoring from below and above, and for censoring thresholds other than 0.

In the case of *truncation* y_i^* is observed if it exceeds 0, that is

$$y_i = y_i^* \text{ if } y_i^* > 0.$$

6.2.2 Properties of the Model

$x_i\beta$ is the mean of y_i^* for a given x_i and we are interested in the vector of marginal effects on this mean, which is given by β. We can estimate it by OLS if

$$\mathbb{E}[y_i|x_i] = x_i\beta.$$

However, by the law of total probability, we have in the case of censoring

$$\begin{aligned} \mathbb{E}[y_i|x_i] &= \Pr(y_i = 0|x_i) \cdot 0 + \Pr(y_i > 0|x_i) \cdot \mathbb{E}[y_i|x_i, y_i > 0] \\ &= \Pr(y_i > 0|x_i) \cdot (x_i\beta + \mathbb{E}[\varepsilon_i|x_i, \varepsilon_i > -x_i\beta]). \end{aligned}$$

It always holds that $\mathbb{E}[\varepsilon_i|x_i, \varepsilon_i > -x_i\beta] > \mathbb{E}[\varepsilon_i|x_i] = 0$ so that $\mathbb{E}[y_i|x_i] \neq x_i\beta$. Hence, a regression of y_i on x_i yields biased estimates.

Similarly, we have in the case of truncation:

$$\mathbb{E}[y_i|x_i, y_i > 0] = x_i\beta + \mathbb{E}[\varepsilon_i|x_i, \varepsilon_i > -x_i\beta] > x_i\beta$$

so that also here a regression of y_i on x_i yields biased estimates.

The probability to observe y_i^* is

$$\Pr(y_i^* > 0|x_i) = \Pr(x_i\beta + \varepsilon_i > 0) = \Pr(\varepsilon_i > -x_i\beta).$$

This probability increases in the mean of y_i^*, $x_i\beta$. So, if we hold the variance of ε_i constant, then the higher y_i^* the lower the bias in OLS estimates of β because the less

likely it is that censoring or truncation occurs. Intuitively, one might consider it unproblematic to use OLS to regress total expenditures on characteristics of the individual, because the probability that they are zero is sufficiently low, while considering it problematic to regress expenditures for luxury goods on the same characteristics because the probability of having zero expenditures for luxury good is non-negligible.

6.2.3 Identification

It is easy to see that β is identified under a median restriction, rather than mean independence. If we denote the conditional median of ε_i by $med(\varepsilon_i|x_i)$ and assume that it is zero then we have that

$$med(y_i|x_i) = med(y_i^*|x_i) = x_i\beta$$

provided that less than half of the observations are censored. This argument is due to Powell (1984).

Alternatively, one can assume that the distribution is symmetric and artificially censor the upper tail of the distribution of y_i. Then, the mean of the artificially censored data is equal to the mean of y_i*, see Powell (1986) for details. This idea is also useful for the case of truncation where artificial truncation, rather than censoring, is performed.

Both the median restriction and the symmetry of the distribution are implied by the normality assumption that is made in the Tobit model.

6.2.4 Maximum Likelihood Estimation

The likelihood function for the type 1 Tobit model is a combination of the likelihood function of the probit model and the linear regression model with normally distributed error terms.

We start with censoring. The censoring probability is

$$\Pr(y_i = 0|x_i) = \Pr(x_i\beta + \varepsilon_i \leq 0) = \Pr(\varepsilon_i \leq -x_i\beta).$$

Notice that here we do not need to normalize the variance of ε_i to one because we observe y_i^* whenever it is positive. This fixes the scale.

So, since ε_i is assumed to be normally distributed with variance σ^2 we have that this is equal to

$$\Pr(\varepsilon_i/\sigma \leq -x_i\beta/\sigma) = \Phi(-x_i\beta/\sigma).$$

By the symmetry of the standard normal c.d.f. we have

$$\Pr(\varepsilon_i/\sigma > -x_i\beta/\sigma) = 1 - \Phi(-x_i\beta/\sigma) = \Phi(x_i\beta/\sigma).$$

This is the part of the model that is similar to the probit model we discussed in Section 5.1.9.

Now let d_i be an indicator for y_i^* being observed. Then, the density of y_i^* is the derivative of the c.d.f. $\Phi((y_i^* - x_i\beta)/\sigma)$ with respect to y_i^*,

$$f(y_i^*|x_i;\beta,\sigma) = \frac{1}{\sigma}\phi\left(\frac{y_i^* - x_i\beta}{\sigma}\right).$$

This is the part of the model that corresponds to the multivariate linear model in Section 4.3.7.

In the case of *censoring* the joint density of y_i and d_i given x_i is

$$f(y_i|x_i,d_i;\beta,\sigma) = \begin{cases} \Phi(-x_i\beta/\sigma) & \text{if } d_i = 0 \\ \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right) & \text{if } d_i = 1, \end{cases}$$

so we can write

$$f(y_i|x_i,d_i;\beta,\sigma) = \Phi(-x_i\beta/\sigma)^{1-d_i} \cdot \left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right)^{d_i}.$$

The log likelihood function for the sample is

$$\mathscr{L}(\beta,\sigma) = \sum_{i=1}^{N}(1-d_i)\log(\Phi(-x_i\beta/\sigma)) + d_i\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right).$$

For *truncation* the joint density of y_i and $d_i = 1$ is still

$$f(y_i, 1|x_i;\beta,\sigma) = \frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right).$$

The probability that y_i is observed is

$$\Pr(y_i > 0|x_i) = \Phi(x_i\beta/\sigma)$$

so that by Bayes' rule we get the density of y_i is in the truncated sample is

$$f(y_i|x_i, d_i = 1; \beta, \sigma) = \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)}{\Phi(x_i\beta/\sigma)}.$$

From this we get the sample log likelihood function

$$\mathscr{L}(\beta, \sigma) = \sum_{i=1}^{N} \log\left(\frac{1}{\sigma}\phi\left(\frac{y_i - x_i\beta}{\sigma}\right)\right) - \log(\Phi(x_i\beta/\sigma)).$$

6.2.5 Relaxing Distributional Assumptions

It is possible to relax the distributional assumptions we made here and estimate censored regression models semiparametrically. See Chapter 9 of Pagan and Ullah (1999) or Chapter 11 of Li and Racine (2007) for an overview.

6.3 Tobit Model with Selection Equation

6.3.1 Model

In the type 1 Tobit model we have assumed that the probability to observe y_i^* depends in the same way on explanatory variables as y_i^*. This is because we have assumed that y_i^* is observed whenever it takes on certain values.[3] The type 2 Tobit model generalizes this. For female labor supply, for example, it allows the relative importance of characteristics such as the number of children to be different for the decision to work at all and how many hours to work.

In this model y_{1i} indicates whether the outcome of interest is observed, for example indicates whether a woman is working. This was indicated by d_i in the type 1 Tobit model. In particular, there is a latent index

$$y_{1i}^* = x_{1i}\beta_1 + \varepsilon_{1i} \tag{6.3.1}$$

[3] In the type 1 Tobit model the probability to observe $y_i > 0$ given explanatory variables x_i is $\Phi(x_i\beta/\sigma)$. Moreover, we have assumed that $y_i^* = x_i\beta + \varepsilon_i$. For both the explanatory variables and the coefficients are the same.

and the selection model for y_{1i} is

$$y_{1i} = 1\{y_{1i}^* > 0\}. \tag{6.3.2}$$

There is a second latent index, for example for the desired amount of hours a woman likes to work,

$$y_{2i}^* = x_{2i}\beta_2 + \varepsilon_{2i}, \tag{6.3.3}$$

and we observe

$$y_{2i} = y_{2i}^* \text{ if } y_{1i} = 1. \tag{6.3.4}$$

The error terms are jointly normally distributed, that is

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right),$$

where, like in the probit model, we normalize the variance of ε_{1i} to 1.

The model is a generalization of the type 1 Tobit model because it is obtained for $x_{1i} = x_{2i}$, $\beta_1 = \beta_2$ and $\varepsilon_{1i} = \varepsilon_{2i}$, without the scale normalization.

6.3.2 Identification

We can see that β_2 is identified when there is censoring by writing

$$\begin{aligned} \mathbb{E}[y_{2i}|x_{1i}, x_{2i}, y_{1i} = 1] &= \mathbb{E}[y_{2i}|x_{1i}, x_{2i}, x_{1i}\beta_1 + \varepsilon_{1i} > 0] \\ &= x_{2i}\beta_2 + \mathbb{E}[\varepsilon_{2i}|x_{1i}\beta_1 + \varepsilon_{1i} > 0]. \end{aligned}$$

Here, the first equality follows from (6.3.1) and (6.3.2). The second equality follows from (6.3.3) and (6.3.4) and the independence between $(\varepsilon_{1i}, \varepsilon_{2i})$ and (x_{1i}, x_{2i}).

$\mathbb{E}[\varepsilon_{2i}|x_{1i}\beta_1 + \varepsilon_{1i} > 0]$ depends on x_{1i} once ε_{1i} and ε_{2i} are not independent of one another. Denote this expectation by $\kappa(x_{1i}\beta_1)$.

β_1 is identified once we choose a distribution of ε_{1i}. Then, we can condition on $x_{1i}\beta_1$, thereby holding $\kappa(x_{1i}\beta_1)$ constant, and use variation in x_{2i} to identify the slope coefficients in β_2 using the same logic as in Section 3.6.2.

To identify the intercept in β_2 there needs to be enough variation in x_{1i} so that $\Pr(y_{1i} = 1|\tilde{x}_{1i}) = 1$ for some $\tilde{x}_{1i}$. Then, $\kappa(\tilde{x}_{1i}\beta_1) = 0$ because we have imposed the normalization $\mathbb{E}[\varepsilon_{1i}] = 0$ and for $x_i = \tilde{x}_i$ censoring occurs with probability zero. So

$$\mathbb{E}[\varepsilon_{1i}|\tilde{x}_{1i}\beta_1 + \varepsilon_{1i} > 0] = \mathbb{E}[\varepsilon_{1i}],$$

where the equality holds because $\tilde{x}_{1i}\beta_1 + \varepsilon_{1i} > 0$ is true with probability one. This argument has been used by Heckman (1990). Chamberlain (1986) calls it "identification at infinity."[4]

6.3.3 Heckman Correction

Heckman (1976, 1979) maintains the normality assumption and points out that it implies that

$$\mathbb{E}[\varepsilon_{2i}|x_{1i}\beta_1 + \varepsilon_{1i} > 0] = \sigma_{12}\lambda(x_{1i}\beta_1),$$

where $\lambda(x_{1i}\beta_1) \equiv \phi(x_{1i}\beta_1)/\Phi(x_{1i}\beta_1)$ is called inverse Mills ratio.

This suggests that we can estimate β_2 in two steps. First, estimate β_1 using a probit model with all observations. This estimate, $\hat{\beta}_1$, can then be used to calculate the inverse Mills ratio

$$\hat{\lambda} = \lambda(x_{1i}\hat{\beta}_1).$$

Finally, estimate β_2 in

$$y_{2i} = x_{2i}\beta_2 + \sigma_{12}\hat{\lambda} + v_i$$

using the OLS estimator on uncensored observations.[5]

In practice it is important to correct the second stage standard errors because the Mills ratio term has been estimated in the first stage.[6] However, for a simple test for selection, which consists of testing whether the coefficient on the Mills ratio term is zero, we do not need to do this because under the null of no selection the standard errors are valid.

[4] See also Ahn and Powell (1993) and Powell (1994) on semiparametric estimation of such models and Andrews and Schafgans (1988) on semiparametric estimation of the intercept.

[5] This estimator is less efficient than the maximum likelihood estimator which is obtained under the same assumptions.

[6] See Heckman (1979) and Newey and McFadden (1994) for derivations.

A second point that is important in practice is that the inverse Mills ratio is typically close to being a linear function in $x_{1i}\beta_1$ so that we run the risk of close multicollinearity between the Mills ratio term and x_{2i} if $x_{1i} = x_{2i}$. Therefore, it is favorable to include some variables in x_{1i} which are not included in x_{2i}.

6.3.4 Maximum Likelihood Estimation

The likelihood to observe $y_{1i} = 0$ is $\Phi(-x_{1i}\beta_1)$. Moreover, the likelihood to observe y_{2i} together with $y_{1i} = 1$ is

$$\Phi(x_{1i}\beta_1) \cdot f_{\varepsilon_{2i}|\varepsilon_{1i}}(y_{2i} - x_{2i}\beta | \varepsilon_{1i} > -x_{1i}\beta_1)$$

where $f_{\varepsilon_{2i}|\varepsilon_{1i}}(\cdot|\cdot)$ is the density of ε_{2i} conditional on ε_{1i}.[7] We can combine this to get

$$\begin{aligned} &f(y_{2i}|x_i, y_{1i}; \beta, \sigma) \\ &\quad = \Phi(-x_{1i}\beta_1)^{1-y_{1i}} \cdot \left(\Phi(x_{1i}\beta_1) \cdot f_{\varepsilon_{2i}|\varepsilon_{1i}}(y_{2i} - x_{2i}\beta | \varepsilon_{1i} > -x_{1i}\beta_1)\right)^{y_{1i}}. \end{aligned}$$

and

$$\begin{aligned} \mathscr{L}(\beta, \sigma) = \sum_{i=1}^{N} &(1 - y_{1i}) \log(\Phi(-x_{1i}\beta_1)) \\ &+ y_{1i} \left(\log(\Phi(x_{1i}\beta_1)) + \log\left(f_{\varepsilon_{2i}|\varepsilon_{1i}}(y_{1i} - x_{2i}\beta | \varepsilon_{1i} > -x_{1i}\beta_1)\right)\right). \end{aligned}$$

6.3.5 Relaxing Distributional Assumptions

It is possible to relax the distributional assumptions the Heckman correction estimator is based on. Chapter 8 of Pagan and Ullah (1999) and Chapter 10 of Li and Racine (2007) discusses how one can do that in order to semiparametrically estimate selection models.

[7] There is a closed form for the bivariate normal distribution that depends on the unknown σ_{12}.

6.4 Duration Analysis

Like some of the models we have already discussed above, duration analysis deals with data that are typically right censored.[8] This happens if events have not ended at the time an individual is observed, or if individuals drop out of the sample before the duration ends.

Our leading example will be the unemployment duration of individuals and its dependence on individual characteristics. We say that their *initial state* is that they are unemployed. It is important to distinguish between two different sampling schemes, *stock sampling* and *flow sampling*. If our sample consists of all individuals that are unemployed at a given point in time, then we have a stock sample. If, to the contrary, our sample consists of all individuals that became unemployed in a given time interval, then we face flow sampling.

Naturally, stock samples are selected samples because those individuals who became unemployed at a given time but found a job before the sampling date are not in the sample. In contrast, those individuals who did not find a job are in the sample. This is a case of left truncation and can be dealt with in a straightforward way. However, as before in our discussion of the Tobit model, we will mainly focus on flow samples in the remainder.

Throughout, we will assume that the duration of the event does not depend on the exact starting time conditional on covariates, which, in addition, we assume to be exogenous. Moreover, we assume censoring occur at random. This excludes, for example, the case in which individuals with unobserved characteristics that yield particularly long unemployment durations drop out of the sample after some time. However, we will allow for the case in which the likelihood to find a new job at a given time depends on the time.

We proceed as follows. We first introduce the standard notation and present some standard relationships between defined measures. Then, we study the proportional hazard model for accurately measured durations as well as for grouped data.

[8] The notions of "right" and "left" stem from the imagination of a horizontal time line. Before, we have been expressing this in terms of the words "below" and "above" which are sensible in the context of a scale for, say, earnings.

6.4.1 Notation and Key Relationships

Denote the time spent in the initial state by T and assume that T is continuously distributed. T has a c.d.f.

$$F(t) = \Pr(T \leq t)$$

and p.d.f.

$$f(t) = \frac{\partial F(t)}{\partial t}. \tag{6.4.1}$$

The survivor function gives the probability that an individual is still in the initial state at t, that is

$$S(t) \equiv 1 - F(t).$$

To express the mean duration in terms of the survivor function observe that

$$\frac{\partial (sF(s))}{\partial s} = F(s) + sf(s).$$

Subtracting $F(s)$ from both sides and taking the integral yields that $\mathbb{E}[T]$ is equal to

$$\int_0^\infty sf(s)\,ds = \int_0^\infty \frac{\partial (sF(s))}{\partial s}\,ds - \int_0^\infty F(s)\,ds = [sF(s)]_0^\infty - \int_0^\infty F(s)\,ds.$$

Observe that $0F(0) = 0$ and $F(\infty) = 1$. Then,

$$\mathbb{E}[T] = \infty - \int_0^\infty F(s)\,ds = \int_0^\infty 1 - F(s)\,ds = \int_0^\infty S(s)\,ds,$$

that is the mean duration is given by the area under the survivor function.

The hazard rate is rate of leaving the initial state in t conditional on surviving at least until t, that is

$$\lambda(t) \equiv \lim_{\Delta \downarrow 0} \frac{\Pr(t \leq T < t + \Delta)}{\Delta}.$$

The numerator is the probability that the duration ends between t and $t + \Delta$ conditional on surviving until t. It can also be written as $(F(t+\Delta) - F(t))/S(t)$. The denominator is the length of the time interval. Here, we consider a limit for $\Delta \to 0$. (6.4.1) implies

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{6.4.2}$$

If this function is increasing in t then we say that there is positive duration dependence, and otherwise we say that there is negative duration dependence. for example if the probability to find a new job for the ones who are still unemployed in a given period is increasing in the unemployment duration then we say that there is positive duration dependence.

At t the probability to survive decays with rate $\lambda(t)$. Therefore, we can write

$$S(t) = \exp\left(-\int_0^t \lambda(s)\, ds\right). \tag{6.4.3}$$

Conversely, (6.4.3) implies that

$$-\frac{\partial \log S(t)}{\partial t} = \frac{\partial \int_0^t \lambda(s)\, ds}{\partial t} = \lambda(t).$$

where the last equality follows from Leibnitz' rule for the differentiation of integrals. In both formulas

$$\Lambda(s) \equiv \int_0^t \lambda(s)\, ds$$

is the so-called integrated hazard. At $t = 0$ no duration has ended so that $S(0) = 1$.

In an important special case we have a constant hazard rate $\lambda(t) = \lambda$. Then,

$$S(t) = \exp\left(-\int_0^t \lambda\, ds\right) = \exp(-\lambda t) \tag{6.4.4}$$

which is the exponential decay formula for a rate of λ and a starting value of 1. A direct consequence is that T follows the exponential distribution since

$$F(t) = 1 - \exp(-\lambda t).$$

In practice, the Weibull distribution is often preferred over the exponential distribution because on the one hand it is more flexible in allowing the hazard rate to depend on t. On the other hand, it nests the exponential distribution as a special case. The survivor function for the Weibull distribution is

$$S(t) = \exp(-\lambda t^{\alpha})$$

with hazard rate

$$\lambda(t) = \lambda \alpha t^{\alpha-1}$$

and integrated hazard

$$\Lambda(t) = \int_0^t \lambda \alpha s^{\alpha-1}\, ds = [\lambda s^{\alpha}]_0^t = \lambda t^{\alpha}.$$

The corresponding quantities for the exponential distribution are obtained for $\alpha = 1$.

Like in the type 1 Tobit model the likelihood contribution of an uncensored observation for individual i is $f(t_i)$. (6.4.2) implies that this is equal to $\lambda(t_i) \cdot S(t_i)$. For a censored observation, also at t_i, it is $S(t_i)$. Hence, we can define an indicator d_i that takes on the value 1 if the observation is uncensored and 0 if it is censored. This yields for the density of the observed t_i given the censoring indicator

$$f(t_i|d_i) = (\lambda(t_i) \cdot S(t_i))^{d_i} \cdot (S(t_i))^{1-d_i}$$

and hence the log likelihood is given by

$$l_i(\theta) = d_i \log(f(t_i)) + (1-d_i)\log(S_i(t_i)) = d_i \log(\lambda(t_i)) + \log(S(t_i)). \tag{6.4.5}$$

The second equality follows because $\log(\lambda(t_i) \cdot S(t_i)) = \log(\lambda(t_i)) + \log(S(t_i))$. The log likelihood is very similar to the one for the type 1 Tobit model. In case of left truncation, which occurs in stock samples, an adjustment similar to the one in the truncated regression model can be made.

An assumption on $f(t)$ or $F(t)$ will typically imply a particular functional form for the baseline hazard function, and vice versa. For example, if the Weibull distribution is chosen the set of parameters consists of α, λ and β.

6.4.2 Proportional Hazard Model

One key characteristic of duration analysis is that the hazard rate, rather than the duration time itself, is modeled. Typically, a likelihood function is derived from this. We have just studied two particular special cases in which the duration is exponentially distributed and follows the Weibull distribution, respectively. However, such parametric restrictions are not necessary for identification because the hazard function is, by

definition, a property of the observed distribution of T. It is identified and so is $S(t)$. Hence, by (6.4.2) $f(t)$ is identified.

In economic applications, we are interested in the relationship between these quantities and exogenous covariates. Notice that the identification argument holds conditional on covariates as well so that these effects are identified as well.

The early literature has mainly focused on modeling the hazard rate and estimating it using maximum likelihood techniques. For this, so called proportional hazard models are the most prominent choice. They specify the hazard as a function of covariates x and a baseline hazard $\lambda_0(t)$ and impose

$$\lambda(t,x) = \lambda_0(t)\exp(x\beta) \tag{6.4.6}$$

Here, x does not include a constant term so that β does not include an intercept—a normalization. β is the derivative of the log of the hazard rate with respect to x, hence a semi-elasticity since this is equal to the derivative of the hazard rate over the hazard rate itself.

Here, we do not consider the case of time varying covariates. In practice, we can model one individual that is uncensored but has two values of x as two observations, one for each value. In principle, the hazard rate at t can also depend on past values of x. For these generalizations see, for example, Cameron and Trivedi (2005, p. 598). We will also not discuss the mixed proportional hazard model where

$$\lambda(t,x,v) = v\lambda_0(t)\exp(x\beta)$$

and v is unobserved. Here, identification is not automatically ensured. This model is estimated, like the mixed multinomial logit model, using simulation techniques. See van den Berg (2001) for a discussion of this model.

Next, we discuss the Cox (1972) proportional hazard model where a specification of the baseline hazard can be circumvented. Thereafter, we discuss models for grouped data where the baseline hazard is sometimes approximated using a step function. Alternatively, an ordered logit model can be estimated.

6.4.3 Cox Proportional Hazard Model

The Cox (1972) proportional hazard model exploits the fact that the hazard function in (6.4.6) is a multiplicative function of the baseline hazard and the part that depends on

covariates. Cox (1972) based his estimation procedure for the latter on a quantity that does not depend on the baseline hazard.

In particular, consider all individuals whose duration has not ended at t. We say that these individuals are at risk and form the risk set $R(t)$. We now focus on the case in which there is only one individual whose duration ends at t. This is a clear limitation. Extensions to the case in which multiple durations end at t are possible but not trivial.[9] Now, at t, since the observations are independently sampled the probability that i's duration ends at t provided that $i \in R(t)$ is given by

$$\frac{\lambda(t, x_i)}{\sum_{i' \in R(t)} \lambda(t_{i'}, x_{i'})}.$$

(6.4.6) implies that this is equal to

$$\frac{\exp(x_i \beta)}{\sum_{i' \in R(t)} \exp(x_{i'} \beta)}.$$

This quantity is the likelihood contribution for i at t. It resembles the likelihood function for the multinomial logit model except that now the set of alternatives is given by the risk set which, usually, contains more observations.

6.4.4 Proportional Hazard Model and Grouped Data

A clear limitation of the Cox (1972) model is that it is not straightforward to use it once there are only discrete measurements of the individual durations. This is often the case in applications where, for example, the unemployment duration is measured in months. Moreover, it is typically the case that many durations end after, say, 10 months. Hence, the assumption that only one duration ends at a time is clearly violated. We call such coarsely recorded data grouped data.

Kalbfleisch and Prentice (1973) develop results for the Cox (1972) model that are applicable for grouped data. Subsequently, Prentice and Gloeckler (1978) developed easily interpretable expressions that can be used for maximum likelihood estimation. Let durations be grouped in intervals $A_j \equiv [a_{j-1}, a_j)$ with $a_0 = 0$.

[9] See Kiefer (1988) for a discussion.

Define

$$\alpha_j \equiv \exp\left(-\int_{a_{j-1}}^{a_j} \lambda_0(s)\, ds\right). \tag{6.4.7}$$

(6.4.6) implies that for an individual with all covariates artificially set to zero this is the probability that a duration ends between $t = a_{j-1}$ and $t = a_j$ conditional on having survived up to $t = a_{j-1}$. The well-known Kaplan and Meier (1958) "product limit" estimator estimates this probability for all j and thereby estimates the survivor function and the discrete time hazard rates. Prentice and Gloeckler (1978) adjust this estimation procedure for covariates. For this observe that (6.4.3) and (6.4.6) imply that

$$\log(S(t)) = -\int_0^t \lambda(s,x)\, ds = -\exp(x\beta)\int_0^t \lambda_0(s)\, ds.$$

Hence, by (6.4.7),

$$\log(S(t)) = -\exp(x\beta)\sum_{j=1}^{t} \alpha_j.$$

Moreover, assume that the baseline hazard is constant in a time interval, that is

$$\lambda(t,x) = \frac{\alpha_t}{\sum_{s=1}^{t} \alpha_s}\exp(x\beta).$$

Then, (6.4.5) takes on the form

$$l_i(\theta) = d_i(x_i\beta + \log(\alpha_{t_i})) - \exp(x_i\beta)\sum_{j=1}^{t_i} \alpha_j$$

and the unknown parameters that are to be estimated are the set of α_j's and β.

Han and Hausman (1990) take a similar approach. In the proportional hazard model the integrated hazard is

$$\Lambda(t,x) = \exp(x\beta)\Lambda_0(t)$$

where

$$\Lambda_0(t) \equiv \int_0^t \lambda_0(s)\, ds$$

is the integrated baseline hazard. Then, we have for the negative of the log of the integrated hazard that

$$-\log(\Lambda(t,x)) = -x\beta - \log(\Lambda_0(t)).$$

Define $\varepsilon \equiv -\log(\Lambda(t,x))$. Then, the survivor function of the proportional hazard model can be written as

$$S(t) = \exp(-\Lambda_0(t)\exp(x\beta)) = \exp(-\exp(\varepsilon)).$$

Hence, $S(t)$ follows a type 1 extreme value distribution. The probability to observe that a duration ends at t is given by

$$\lambda(t,x) = \frac{\alpha_t}{\sum_{s=1}^{t}\alpha_s}\exp(x\beta).\int_{\Lambda_0(t-1)+x\beta}^{\Lambda_0(t)+x\beta} f(\varepsilon)\,d\varepsilon.$$

This is an ordered logit model. Let y_{it} be a indicator for the duration ending at t. Then, the log likelihood for the uncensored case is given by

$$l_i(\theta) = \sum_t y_{it}\log\left(\int_{\Lambda_0(t-1)+x_i\beta}^{\Lambda_0(t)+x_i\beta} f(\varepsilon_i)\,d\varepsilon_i\right).$$

Censoring can be accounted for by including a term which specifies the cumulative probability of survival if the observation is censored.

Part III

Policy Evaluation

Chapter 7

Policy Evaluation

In this chapter we discuss the literature on policy evaluation. This literature mainly studies the case of a binary endogenous treatment variable which is sometimes allowed to depend on the effect of the treatment on an outcome, thus allowing for very general dependence structures. In the end, we separately discuss the case of continuous treatments, that is continuous endogenous variables. This chapter can be regarded as an extension of the discussion of well-known identification problems in Chapter 3.3.

7.1 Formal Framework and Parameters of Interest

The notation in this literature is usually different to the one in classical econometrics. This is sometimes confusing, but there are also some advantages to it. In particular, we can easily represent individual causal effects of choosing one alternative relative to another in terms of the difference between the two hypothetical outcomes. See Holland (1986) for a discussion and a philosophical perspective.

In this so-called potential outcomes notation uppercase letters denote random variables and lowercase letters realizations. We observe an outcome Y, a vector of covariates X, a binary variable D and possibly a vector of instruments Z. If $D = 1$ the observed outcome Y is a realization of Y_1 and if $D = 0$ it is a realization of Y_0. Hence, we observe

$$Y = D \cdot Y_1 + (1 - D) \cdot Y_0.$$

Throughout, the vector of covariates is assumed to be exogenous.[1]

In a cross section there are no repeated observations and hence, we never observe realizations of Y_0 and Y_1 for the same unit. Holland (1986) calls this the "fundamental problem of causal inference." A consequence of this is that the distribution of $Y_1 - Y_0$ is not observed. For this reason researchers became interested in recovering features of $Y_1 - Y_0$, possibly as a function of X, from observations. Important features which are sought to be estimated are the population average treatment effect, $\mathbb{E}[Y_1 - Y_0]$, the treatment effect on the treated, $\mathbb{E}[Y_1 - Y_0|D = 1]$, and the treatment effect on the untreated, $\mathbb{E}[Y_1 - Y_0|D = 0]$. Many of them are of direct relevance to policy makers as they answer the question, for example, what would have been the gain in the outcome for the ones who did not take the treatment had they been forced to do so. See, for example, Heckman and Vytlacil (2000), Heckman and Vytlacil (2005), and Heckman et al. (2006) for a discussion.

In the remainder we go through several sets of assumptions under which these parameters of interest are identified. The main focus in this literature is on the heterogeneity of treatment effects and therefore, we focus on this point as well. Before moving on, we shall emphasize that in program evaluation we face a missing data problem.

7.1.1 Missing Data Problem

The missing data problem is best discussed using the type 5 Tobit model which is presented in classic notation to establish the link to our earlier discussion of censoring and truncation in Chapter 6. In fact, this model is a straightforward generalization of the type 2 Tobit model that was discussed in Section 6.3. For now we switch back to the notation that is more common in econometrics.

Depending on a variable d_i we observe an outcome

$$y_i = \begin{cases} y_{0i} = g_0(x_i) + \varepsilon_{0i} \text{ if } d_i = 0 \\ y_{1i} = g_1(x_i) + \varepsilon_{1i} \text{ if } d_i = 1. \end{cases}$$

[1] Then, everything can be discussed conditional on X. A weaker version of exogeneity of X is a no-feedback condition, see Heckman and Vytlacil (2005) for details. This condition is that D has no effect on X. At the same time, X is allowed to affect D.

Since $g_0(x_i)$ and $g_1(x_i)$ are general nonparametric functions we can normalize ε_{0i} and ε_{1i} to be mean zero. d_i is determined by the selection model

$$d_i = 1\{p(z_i) \geq \nu_i\}.$$

We can actually normalize ν_i to be uniformly distributed when we use the general function $p(z_i)$ instead of a linear index. Then, $p(z_i)$ is the probability to observe $d_i = 1$, which we interpret as the probability of being treated. To see this, write

$$\Pr(d_i = 1|z_i) = \Pr(\nu_i \leq p(z_i)|z_i) = p(z_i),$$

where the last equality holds when ν_i is uniformly distributed *independent* of z_i. The probability $p(z_i)$ is sometimes also referred to as the propensity score (Rosenbaum and Rubin, 1983). So far, the only differences to the type 2 Tobit model that has been studied before is that an outcome is observed in either case and that we use a more general functional form $p(z_i)$ instead of a linear index.

In policy evaluation we are interested in the effect of d_i on y_i,

$$\mathbb{E}[y_{1i} - y_{0i}] = g_1(x_i) - g_0(x_i).$$

The missing data problem is that we don't observe y_{0i} for those individuals who chose $d_i = 1$, whereas we don't observe y_{1i} for those who chose $d_i = 0$. If ε_{0i} and ε_{1i} are not independent of ν_i, which is the most plausible case in applied contexts, the missing data problem cannot be ignored because

$$\mathbb{E}[\varepsilon_{0i}] \neq \mathbb{E}[\varepsilon_{0i}|d_i = 0]$$

and

$$\mathbb{E}[\varepsilon_{1i}] \neq \mathbb{E}[\varepsilon_{1i}|d_i = 1].$$

Assume that the unobservables $(\varepsilon_{0i}, \varepsilon_{1i}, \nu_i)$ are jointly independent of the observables (x_i, z_i) and notice that this still allows for a correlation between x_i and z_i and between ε_{0i}, ε_{1i} and ν_i. Under this assumption we can write, like for the type 2 Tobit model,

$$\mathbb{E}[y_{1i}|x_i, z_i, d_i = 1] = \mathbb{E}[y_{1i}|x_i, p(z_i) \geq \nu_i] = g_1(x_i) + \kappa_1(p(z_i)).$$

In addition we have now

$$\mathbb{E}[y_{0i}|x_i, z_i, d_i = 0] = \mathbb{E}[y_{0i}|x_i, p(z_i) < \nu_i] = g_0(x_i) + \kappa_0(p(z_i)).$$

This shows that conditional on $p(z_i)$, $g_0(x_i)$ and $g_1(x_i)$ are identified up to respective constants. Unlike in many contexts in which we are interested in the dependence of $g_0(x_i)$ and $g_1(x_i)$ on x_i, in policy evaluation we are mostly interested in those constants (Heckman, 1990). For this, like for the type 2 Tobit model, an "identification at infinity" argument can be made.

In particular, $\mathbb{E}[\varepsilon_{0i}] = 0$ implies that for $p(z_i) = 0$

$$\kappa_0(0) = \mathbb{E}[\varepsilon_{0i}|d_i = 0, p(z_i) = 0] = \mathbb{E}[\varepsilon_{0i}|0 < \nu_i] = \mathbb{E}[\varepsilon_{0i}] = 0.$$

Similarly, $\mathbb{E}[\varepsilon_{1i}] = 0$ implies that for $p(z_i) = 1$

$$\kappa_1(1) = \mathbb{E}[\varepsilon_{1i}|d_i = 1, p(z_i) = 1] = \mathbb{E}[\varepsilon_{1i}|1 \geq \nu_i] = \mathbb{E}[\varepsilon_{1i}] = 0.$$

Hence, $g_0(x_i)$ is identified if $p(z_i) = 0$ for some z_i and $g_1(x_i)$ is identified if $p(z_i) = 1$ for some other z_i. This result is very intuitive because in cases in which $p(z_i) = 0$ there is no missing data problem for y_{0i}. Conversely, $p(z_i) = 1$ implies that there is none for y_{1i}.

Finally, a Heckman (1976, 1979) correction procedure procedure is feasible under the assumption that $(\varepsilon_{0i}, \varepsilon_{1i}, \nu_i)$ is trivariate normally distributed with

$$var \begin{pmatrix} \varepsilon_{0i} \\ \varepsilon_{1i} \\ \nu_i \end{pmatrix} = \begin{pmatrix} \sigma^2_{\varepsilon_{0i}} & \sigma_{\varepsilon_{0i},\varepsilon_{1i}} & \sigma_{\varepsilon_{0i},\nu_i} \\ \sigma_{\varepsilon_{0i},\varepsilon_{1i}} & \sigma^2_{\varepsilon_{1i}} & \sigma_{\varepsilon_{1i},\nu_i} \\ \sigma_{\varepsilon_{0i},\nu_i} & \sigma_{\varepsilon_{1i},\nu_i} & 1 \end{pmatrix}.$$

It is well known that this implies

$$\kappa_0(p(z_i)) = -\sigma_{\varepsilon_{0i},\nu_i} \cdot \frac{\phi(p(z_i))}{1 - \Phi(p(z_i))}$$
$$\kappa_1(p(z_i)) = \sigma_{\varepsilon_{1i},\nu_i} \cdot \frac{\phi(p(z_i))}{\Phi(p(z_i))}.$$

Such a correction procedure can be implemented in very much the same way as in the type 2 Tobit case. A nice application for this is Willis and Rosen (1979) where

corrected estimates of earnings as a function of college education are obtained and then included into a structural probit model to study the dependence of college choice on earnings.

This establishes the link to results that have been known before the 1990's. Next, we turn to models that have genuinely been designed for policy evaluation. For this we use the notation in that literature.

7.2 Random Assignment Conditional on Covariates

7.2.1 Key Assumption

The easiest case arises if D is independent of the pair (Y_0, Y_1) conditional on X. This assumption is sometimes written as

$$D \perp\!\!\!\perp (Y_0, Y_1) | X, \tag{7.2.1}$$

where "$\perp\!\!\!\perp$" denotes joint independence.[2] This means that all individuals that are characterized by a given value of X are equally likely to choose $D = 1$, no matter what their draws of Y_0 and Y_1 are. Otherwise put, the ones who chose $D = 1$ are a random sample of the population conditional on X. Under this assumption D is exogenous given X and we have for the c.d.f.'s of the potential outcomes that

$$F_{Y_1|X} = F_{Y|X,D=1}$$

and

$$F_{Y_0|X} = F_{Y|X,D=0}.$$

To establish the link to the models that have been discussed in Chapter 3 consider the following example in standard notation.

Example 7. Let

$$y_{0i} = x_i\beta_0 + \varepsilon_{0i}$$
$$y_{1i} = x_i\beta_1 + \varepsilon_{1i}$$

[2] Joint independence means that D is independent of both Y_0 and Y_1 given X. At the same time it is *not* assumed that Y_0 is independent of Y_1.

so that the effect of d_i on y_i is random and given by

$$x_i(\beta_1 - \beta_0) + \varepsilon_{1i} - \varepsilon_{0i}.$$

Hence,

$$y_i = x_i\beta_0 + \varepsilon_{0i} + d_i\Big(x_i(\beta_1 - \beta_0) + \varepsilon_{1i} - \varepsilon_{0i}\Big).$$

Conditional on x_i both error terms, ε_{0i} and ε_{1i}, are independent of d_i by assumption so that the conditional distribution of y_i given x_i and d_i is equal to the conditional distribution of y_{0i} given x_i for $d_i = 0$ and y_{1i} given x_i for $d_i = 1$. Here, the assumption that x_i is exogenous means that the error terms are uncorrelated with x_i. Therefore, we can simply compare differences in outcomes for different combinations of d_i and x_i. In practice, we can use a simple regression of y_i on x_i and $d_i x_i$ to estimate the average effect of d_i on y_i as a function of x_i. □

7.2.2 Propensity Score Matching

Define the propensity score as

$$P(X) \equiv \Pr(D = 1|X)$$

and notice that it is (quasi) observable (because we can estimate it). Rosenbaum and Rubin (1983) have shown that exogeneity of D given X, (7.2.1), implies

$$(Y_0, Y_1) \perp\!\!\!\perp D|P(X)$$

and

$$X \perp\!\!\!\perp D|P(X).$$

Therefore, we can first estimate

$$\mathbb{E}[Y_1 - Y_0|P(X)] = \mathbb{E}[Y_1|P(X)] - \mathbb{E}[Y_0|P(X)]$$

and then average those estimates over the population distribution of $P(X)$ to identify the mean effect of D on Y. Formally, this follows from the law of iterated expectations because

$$\begin{aligned}\mathbb{E}[Y_1 - Y_0] &= \mathbb{E}[\mathbb{E}[Y_1 - Y_0|P(X)]] \\ &= \mathbb{E}[\mathbb{E}[Y|D = 1, P(X)] - \mathbb{E}[Y|D = 0, P(X)]].\end{aligned}$$

If we were instead interested in the average treatment effect on the treated we would instead average over the distribution of $P(X)$ given $D = 1$. In a similar manner we would use the distribution of $P(X)$ given $D = 0$ for the average treatment effect on the untreated.

This proceeding crucially depends, however, on the assumption that the support of $P(X)$ does not depend on D. This means that for each value of $P(X)$ there are observations with $D = 0$ and $D = 1$ in the sample.[3]

Estimation procedures consist of several steps. Here, I present a stylized version for the proceeding of many of them.[4] In a first step $P(X)$ is estimated. Typically, the fitted probability from a probit or logit model is used here. In the second step the expected value of Y is estimated as a function of $P(X)$, separately for those observations with $D = 0$ and $D = 1$. Then, the difference is taken for a given $P(X)$ and finally, those differences are averages using the population distribution of $P(X)$ as weights.

Clearly, the upside of this approach is that it allows for heterogeneous effects of D, and that it is intuitive and robust while imposing minimal structure in the estimation step. In particular, no linearity in X assumption is needed. However, the downside is that it relies on the strong assumption of the choice of D being random conditional on X. Moreover, the support condition can be demanding in applications. Finally, no general asymptotic theory is available thus far.[5]

7.3 Differences-in-Differences Estimation

Suppose that there is a treatment and a control group and the treated individuals only received the treatment in the second of two time periods. The outcome equation is

$$y_{it} = \alpha_i + \beta d_{it} + \varepsilon_{it}.$$

[3] If this does not hold, then one can impose functional form assumptions for $\mathbb{E}[Y_0|P(X)]$ or $\mathbb{E}[Y_1|P(X)]$ and extrapolate to values p of $P(X)]$ that lie outside of the support of $P(X)]$ given D. This then allows one to estimate $\mathbb{E}[Y_0|P(X) = p]$ even though p is not in the support of $P(X)$ given $D = 0$.

[4] Sometimes, several steps are taken at once. Well known procedures are Kernel smoothing, nearest neighbor matching, one to one matching, one to many matching, and many others.

[5] There are some results by Abadie and Imbens (2006). Another important point is that in the estimation step the propensity score could be misspecified. Battistin and Chesher (2004) discuss the consequences of a such a misspecification.

In principle, we can also allow for covariates here. Suppose that strict exogeneity holds for the treatment in the sense that d_{it} and ε_{is} are uncorrelated for $s,t = 1,2$. Then, the average treatment effect is the coefficient β in

$$y_{it} = \alpha + \beta 1\{i \text{ is treated and } t = 2\} + \gamma 1\{t = 2\} + \delta 1\{i \text{ is treated}\} + u_{it},$$

which can be estimated by OLS. This is called differences-in-differences estimation because the treatment effect is given by the difference in the expected outcome between the change in the expected outcome in the treatment and the control group from period $t = 1$ to period $t = 2$. An alternative interpretation is that we actually perform a form of fixed effects estimation.

7.4 Nonrandom Assignment and Instrumental Variables

7.4.1 Instrumental Variables

Let us again switch back to classical notation. For a linear model,

$$y_i = x_i\beta + \varepsilon_i,$$

mean independence of ε_i from x_i or uncorrelatedness between x_i and ε_i is often violated in economic applications. Then, we say that "we face an endogeneity problem", and OLS estimates the sum of the parameter vector and an additional term which depends on the correlation between the error term and the right hand side variables. The terminology reflects the fact that the regressor is not exogenous to the model, but rather determined within the model.

A correlation between ε_i and elements of x_i can occur for at least two reasons. First, it could be that there is a so-called confounding variable which is not included into the set of regressors but has an impact on both x_i and ε_i. This is very close to the second reason, namely omitting relevant variables when specifying the regression equation.[6]

If we denote these variables by w_i and their coefficient by γ then the true data generating process is

$$y_i = x_i\beta + w_i\gamma + e_i$$

[6] See, for example, Fisher (1935, Ch. 7) and Yates (1937) for early discussions.

with $\mathbb{E}[e_i|x_i] = 0$. In this case, the error term in (4.4.5) is in fact

$$\varepsilon_i \equiv w_i\gamma + e_i$$

and is correlated with x_i if $w_i\gamma$ is correlated with x_i. In both cases we face an endogeneity problem.

The general idea of instrumental variables estimation is to exploit variation in those "instrumental" variables that translates into variation in the endogenous variable. This is a fruitful approach if the instrumental variable itself is unrelated to the error term in the equation which is to be estimated. Assume for a moment that x_i is a scalar. Then, the idea is that y_i is a function of x_i and x_i is a function of an instrument z_i. Therefore,

$$\frac{dy_i}{dz_i} = \frac{dy_i}{dx_i}\frac{dx_i}{dz_i}$$

and hence

$$\frac{dy_i}{dx_i} = \frac{dy_i}{dz_i} \Big/ \frac{dx_i}{dz_i}.$$

There is a long tradition of using this idea in econometrics.[7] One of the earliest applications has been the estimation of supply and demand elasticities from market data, as described in Section (3.2).

7.4.2 Homogeneous Effects

A natural starting point once we want to allow for more complex assignment mechanisms is to consider homogeneous effects in a first step while allowing the treatment indicator to be endogenous. We say that effects are homogeneous if they are constant across individuals given x_i, that is if

$$y_i = x_i\beta_0 + d_i x_i(\beta_1 - \beta_0) + \varepsilon_i$$

for some constants β_0 and β_1. $x_i(\beta_1 - \beta_0)$ is the effect of d_i on y_i. For an instrumental variables approach assume that there is an instrument (or a vector of instruments) z_i which is correlated with d_i conditional on x_i while being uncorrelated with ε_i. Then, we

[7] See Angrist and Krueger (2001) and Goldberger (1972).

can proceed by estimating β_0 and β_1 using standard instrumental variables techniques. Here, we would regress y_i on a full set of interactions between x_i and d_i, instrumenting d_i with z_i. Importantly, the choice of instrument does not matter for identification if several ones are available.

7.4.3 Heterogeneous Effects and no Selection on Unobservables

Suppose now that

$$y_i = x_i\beta_{0i} + d_i x_i(\beta_{1i} - \beta_{0i}),$$

where x_i is uncorrelated with (β_{0i}, β_{1i}) and includes 1 as the first element. In this general formulation effects are idiosyncratic and given by β_{0i} and β_{1i}, random vectors, with means $\bar{\beta}_0$ and $\bar{\beta}_1$.[8]

We assume for now that there is no selection on unobservables, that is that d_i and its effect, $x_i(\beta_{1i} - \beta_{0i})$, are uncorrelated conditional on x_i. This assumption has been criticized by many authors, for example Heckman (1997) who points out that

> if responses to treatment vary, and if we are interested in estimating the mean effect of treatment on the treated, or the effect of treatment on randomly selected persons, instrumental variables identify these parameters only when agents do not select into the program on the basis of the idiosyncratic component of their response to the program. This is a strong assumption that forces the analyst to assume either irrationality or ignorance on the part of persons whose behavior is being studied.

In addition we assume that z_i is uncorrelated with $x_i\beta_{0i}$ and $d_i(x_i\beta_{1i} - x_i\beta_{0i})$. Then, the instrument is uncorrelated with all random components in the regression equation conditional on x_i and hence we can estimate the average effect of d_i on y_i as a function of x_i using a conventional instrumental variables estimator.[9]

To see this let z_i be a scalar and consider

$$\begin{aligned} &cov(z_i, x_i\beta_{0i} + d_i x_i(\beta_{1i} - \beta_{0i})) \\ &\quad = cov(z_i, x_i\beta_{0i}) + cov(z_i, d_i x_i(\beta_{1i} - \beta_{0i})). \end{aligned}$$

[8] The first element of β_{0i} and β_{1i} can be regarded as the sum of an intercept term, $\bar{\beta}_{0i}$ and $\bar{\beta}_{1i}$, respectively, and an error term with mean zero, ε_{0i} and ε_{1i}, respectively.

[9] See Heckman and Vytlacil (1998) and Heckman et al. (2006) for detailed discussions of this finding.

Both elements of this sum are equal to zero by assumption and hence the instrument is uncorrelated with all randomness in the outcome equation conditional on x_i.

7.4.4 Heterogeneous Effects and Selection on Unobservables

Starting from the above discussion it is interesting to ask what an instrumental variables estimator estimates if, conditional on x_i, d_i and its effect, $y_{1i} - y_{0i} = x_i(\beta_{1i} - \beta_{0i})$, are in fact correlated.[10]

Think of a 2 stage procedure conditional on x_i where we first replace d_i by $p_i \equiv \Pr(d_i = 1|x_i, z_i)$ and then regress y_i on p_i while holding x_i constant. Yitzhaki (1989) shows that we have for the slope coefficient

$$\frac{cov(y_i, p_i|x_i)}{var(p_i|x_i)} = \int_{-\infty}^{\infty} \frac{\partial \mathbb{E}[y_i|x_i, p_i = t]}{\partial t} w(t)\, dt \tag{7.4.1}$$

where the weights depend on the distribution of p_i and are given by

$$w(t) = \frac{\mathbb{E}[p_i - \mathbb{E}[p_i|x_i]|x_i, p_i > t] \cdot \Pr(p_i > t)}{var(p_i|x_i)}.$$

This is a weighted average of covariances of partial derivatives which is in general not related to any quantity that is of interest. Hence, more structure is needed to relate structural parameters of interest to the distribution of observables.

It is noteworthy that if there is no selection on unobservables then it follows that a standard instrumental variables estimator estimates the average treatment effect. This is because by construction we have

$$\mathbb{E}[y_i|x_i, p_i] = \mathbb{E}[y_{0i} + d_i \cdot (y_{1i} - y_{0i})|x_i, p_i].$$

By the independence between d_i and $(y_{1i} - y_{0i})$ conditional on x_i we have that this is equal to

$$\mathbb{E}[y_{0i}|x_i, p_i] + \mathbb{E}[d_i|x_i, p_i] \cdot \mathbb{E}[(y_{1i} - y_{0i})|x_i, p_i].$$

By the condition that the instrument is independent of the potential outcomes and noting that $\mathbb{E}[d_i|x_i, p_i] = p_i$ we get

$$\mathbb{E}[y_i|x_i, p_i] = \mathbb{E}[y_{0i}|x_i] + p_i \cdot \mathbb{E}[(y_{1i} - y_{0i})|x_i]$$

[10] See again Heckman et al. (2006) for a discussion.

so that

$$\frac{\partial \mathbb{E}[y_i|x_i, p_i = t]}{\partial t} = \mathbb{E}[y_{1i} - y_{0i})|x_i].$$

The result then follows from (7.4.1).

7.4.5 Local Average Treatment Effects

Starting from the insight that instrumental variables might fail to identify treatment effect parameters Imbens and Angrist (1994) impose an additional condition on the way individuals select into the treatment. We discuss this approach using the non-classical notation. Everything can be thought of as being conditional on exogenous covariates X.[11]

Assume there is a binary instrument, Z, that takes on the values 0 and 1, define $P(Z) \equiv \Pr(D = 1|Z)$ and write P shorthand for $P(Z)$. Moreover, assume that $D(P) = 1\{P \geq V\}$ where we normalize V to be uniformly distributed.

This model implies what has been called monotonicity or uniformity: a change in P from p to $p' > p$ that is induced by a change in Z from 0 to 1 can never change D from 1 to 0.[12] Finally, assume that $Z \perp\!\!\!\perp (Y_0, Y_1, V)$. This implies

$$P \perp\!\!\!\perp (Y_0, Y_1, V).$$

Consider the following example.[13] Two officials screen applicants for a social program. For every set of characteristics of the applicants (the covariates X in the outcome equation) the admission rate differs between the two officials. If it is unlikely that the identity of the official affects the outcome of participation or nonparticipation in the program, then conditional on the characteristics of the applicant this identity qualifies as an instrument. Suppose the admission rate for official A was higher than for official B. Then, in this setup monotonicity holds whenever *any* applicant who would have been accepted by official B *is* accepted by official A. Imbens and Angrist (1994) note

[11] See Frölich (2007) for incorporating covariates when estimating the model and Angrist et al. (1996) for a discussion of the model.

[12] Vytlacil (2002) shows that there are two equivalent ways to formalize this assumption. See Klein (2010) for an analysis of the case in which monotonicity is assumed but does not hold.

[13] This is Example 2 taken from Imbens and Angrist (1994).

	$D=0$ if $Z=0$	$D=1$ if $Z=0$
$D=0$ if $Z=1$	never taker	defiar
$D=1$ if $Z=1$	complier	always taker

Table 7.1: Subpopulations

that "this is unlikely to hold if admission is based on a number of criteria." In this case, monotonicity is violated. However, there are other cases in which monotonicity holds naturally. for example, eligibility rules typically imply that becoming eligible to take a treatment can never change participation from 1 to 0 since participation was not possible before. See Battistin and Rettore (2008a) for details.

Table 7.1 is taken from Angrist et al. (1996) and illustrates that the response of the treatment variable to changes in the instrument can be used to partition the population into 4 subpopulations. Monotonicity implies that there are no defiars in the population.

After having introduced this monotonicity assumption Imbens and Angrist (1994) notice that, for $p' > p$,

$$\begin{aligned}
&\mathbb{E}[Y|P=p'] - \mathbb{E}[Y|P=p] \\
&\qquad = \mathbb{E}[D(P)Y_1 + (1-D(P))Y_0|P=p'] \\
&\qquad\qquad - \mathbb{E}[D(P)Y_1 + (1-D(P))Y_0|P=p] \\
&\qquad = \mathbb{E}[Y_0 + D(p')(Y_1 - Y_0)] - \mathbb{E}[Y_0 + D(p)(Y_1 - Y_0)] \\
&\qquad = \mathbb{E}[(D(p') - D(p))(Y_1 - Y_0)]
\end{aligned}$$

where the second equality follows from the independence assumption. This can be rewritten in terms of the effects for compliers and defiars

$$\begin{aligned}
&\Pr(D(p') - D(p) = 1) \cdot \mathbb{E}[Y_1 - Y_0 | D(p') - D(p) = 1] \\
&\qquad - \Pr(D(p') - D(p) = -1) \cdot \mathbb{E}[Y_1 - Y_0 | D(p') - D(p) = -1].
\end{aligned}$$

The first line is the probability of being a complier times the mean effect for compliers. The second line is the probability of being a defiar times the mean effect for defiars. The minus in front appears because D changes from 1 to 0. This representation, so far, is not meaningful and corresponds to what we have stated for the instrumental variables

estimator in (7.4.1). However, under monotonicity $\Pr(D(p') - D(p) = -1) = 0$ so that

$$\mathbb{E}[Y_1 - Y_0 | p < V < p'] = \frac{\mathbb{E}[Y|P = p'] - \mathbb{E}[Y|P = p]}{p' - p}. \tag{7.4.2}$$

That is, the average effect of D on Y for the subpopulation of compliers is given by the conventional instrumental variables estimate if the instrument takes on only two values.[14] The numerator is the difference that is due to a change in the values of the instrument, which in turn changes P from p to p'. The denominator is the change in P.

7.4.6 Other Treatment Effect Parameters

Recently, Heckman and Vytlacil (1999, 2001, 2000, 2005) have related other treatment effects parameters of interest to local versions of the local average treatment effect.

In particular, starting with

$$\mathbb{E}[Y_1 - Y_0 | p < V < p'] = \frac{\mathbb{E}[Y|P = p'] - \mathbb{E}[Y|P = p]}{p' - p}$$

and taking limits for $p' \to p$ yields for the so-called marginal treatment effect, $m(v) \equiv \mathbb{E}[Y_1 - Y_0 | V = v]$, that

$$m(p) = \frac{\partial \mathbb{E}[Y|P = p]}{\partial p}.$$

This parameter is of economic interest because it relates the mean effect of D on Y to a particular value of V, which is unobserved.

Starting from this identification result notice that many treatment effect parameters

[14] This in fact is close to the Wald (1940) instrumental variables estimator for binary instruments. See Angrist et al. (2000) for a generalization of this result to simultaneous equations.

can be expressed in terms of the marginal treatment effect:

$$\begin{aligned}
\mathbb{E}[Y_1 - Y_0] &= \int_0^1 m(v)\, dv \\
\mathbb{E}[Y_1 - Y_0 | p < V < p'] &= \frac{1}{p' - p} \int_p^{p'} m(v)\, dv \\
\mathbb{E}[Y_1 - Y_0 | D = 1] &= \int_0^1 h_{TT}(v) m(v)\, dv \\
\mathbb{E}[Y_1 - Y_0 | D = 0] &= \int_0^1 h_{TUT}(v) m(v)\, dv,
\end{aligned}$$

where $h_{TT}(v)$ and $h_{TUT}(v)$ are weights for the treatment effect on the treated and the treatment effect on the untreated, respectively.

Overall, exploiting monotonicity is an elegant approach that does not rely on a parametric probability model. However, the monotonicity assumption is fundamentally untestable and not at all innocuous as it imposes a substantial amount of structure on the choice process.[15] Furthermore, strong support conditions need to hold for identification of the average treatment effect as we need the derivative of $\mathbb{E}[Y|P = p]$ with respect to p to be identified at any value p in the (open) unit interval.

On the other hand, in some cases monotonicity holds by construction. An example for this is an institutional change that changes the eligibility to choose $D = 1$. In this case, individuals can never change their decision from $D = 0$ to $D = 1$. Then, exploiting monotonicity is a natural approach.

7.4.7 Natural Experiments

More recent applied work has focussed on using credible exogenous variation in order to estimate policy-relevant treatment effects parameters, such as sometimes local average treatment effects.

These studies often exploit *natural experiments*, which, technically speaking, generate instrumental variables. An extremely well-received textbook in this community of researchers is the one by Angrist and Pischke (2009).

[15] Klein (2010) derives an approximation to the bias one incurs if monotonicity is assumed but fails to hold.

Natural experiments are characterized by a situation in which nature (or politicians) randomly affect choices (the endogenous variable) of some individuals. This could be due to institutional details which affect individuals in a distinct way because of their birth date. For example, eligibility rules could state that only individuals born before a certain date are eligible to choose a certain alternative.[16]The important feature is that within a known subgroup, here it is all individuals born shortly before and after the threshold date, it is random whose choice is affected in the sense that this (being affected) is not related to the error term of the equation which is to be estimated.

In practice, it is of crucial importance to base the choice of instruments on good knowledge of the institutional details. This ultimately determines the quality of the subsequent analysis. One famous example for an instrument is the quarter of birth for the years of schooling which is considered to be endogenous in an earnings equation (Angrist and Krueger, 1991). The rationale behind this is that most states require students to enter school in the calendar year in which they turn six. Consequently, those born late in the year are younger when they enter than the ones who are born in the beginning of the year. Furthermore, compulsory schooling laws typically require students to remain in school until their 16th birthday. Because of the regulations for entry those born late in the year are in 10th grade when this happens whereas those born early in the year are in 9th grade. The quarter of birth is a valid instrument if it is unrelated to a person's ability, motivation and other factors determining wages conditional on observed covariates. Bound and Jaeger (2000), however, cast serious doubt on the validity of the quarter of birth instrument.

7.4.8 Regression Discontinuity Design

Natural experiments, in particular when they have to do with threshold rules, are closely related to the notion of the so-called regression discontinuity design (RDD). It has first been promoted by Thistlethwaite and Campbell (1960) who study the effect of student scholarships on career aspirations, exploiting the fact that awards are only made if a test score exceeds a threshold. In this situation, individuals who are close to the threshold but on opposite sides differ only with respect to the probability of having received the

[16]This can be interpreted as a binary instrument (Wald, 1940). See also Battistin and Rettore (2008b) on making use of eligibility rules.

treatment or not, and therefore the difference in the expected outcome between them must be due to the effect of the treatment.

Hahn et al. (2001) discuss the relationship of this early work to the set of assumptions made by Imbens and Angrist (1994). Essentially, the estimator that is used is an instrumental variables estimator in (7.4.2). To see this, denote the so-called "running variable", which is the scalar measure that determines whether or not somebody is eligible for a treatment, by z, and already subtract the treshold so that individuals are actually eligible when $z \geq 0$. Then, one way to express the RDD estimator is

$$\mathbb{E}[Y_1 - Y_0 | p < V < p'] = \frac{\lim_{z \to 0^+} \mathbb{E}[Y|Z=z] - \lim_{z \to 0^-} \mathbb{E}[Y|P=z]}{\lim_{z \to 0^+} \mathbb{E}[D|Z=z] - \lim_{z \to 0^-} \mathbb{E}[D|P=z]},$$

where $\lim_{z \to 0^+}$ denotes taking the limit from the right, $\lim_{z \to 0^-}$ denotes taking the limit from the left.

Notice that the denominator is equal to the difference in the probability to receive the treatment, $p' - p$ in (7.4.2). The case in which this difference is 1 is referred to as a sharp regression discontinuity design, and the case in which the difference is less than 1 is referred to as a fuzzy regression discontinuity design. From our discussion in Section 7.4.5 it follows that in general, we need to assume monotonicity for the estimator to be valid. However, as argued by Battistin and Rettore (2008a), it automatically holds when the probability to reveive the treatment is zero below the threshold, that is if $p' = 0$, because then becoming eligible can never change the treatment decision from 1 to 0.

In practice it is important to ask the question whether the estimated parameter is actually of interest. Thinking again of the RDD estimator as the Imbens and Angrist (1994) instrumental variables estimator discussed in Section (7.4.5), we see that a natural drawback is of exploiting a RDD is that the exogenous variation that arises through the existence of the threshold is local and has only an effect on individuals who are on the edge of participating because these are the compliers. Sometimes, however, this effect is of particular interest because it is related to the question what the effect of marginally changing the threshold would be—a question that could be highly policy-relevant. On top of this practical issue, Imbens and Lemieux (2008) discuss how one can address many technical issues related to estimating treatment effects exploiting an RDD in practice.

7.5 Continuous Endogenous Variables

The case of continuous endogenous is conceptually closely related, but somewhat different in terms of the approach.

Next, we discuss one approach that can be taken if a particular structure can be imposed. This approach has recently been promoted by Chesher (2002) and Imbens and Newey (2003). We follow the exposition in the latter paper. Consider the triangular structure

$$Y = g(X, \varepsilon) \tag{7.5.1}$$

$$X = h(Z, \eta) \tag{7.5.2}$$

with scalar η and possibly vector valued ε and assume that Z is independent of (ε, η).

As the function h is thus far not restricted we can normalize η to be uniformly distributed. Moreover, assume that h is strictly increasing in its second argument. Then, h is invertible and η is identified since it is given by the percentile of X given Z,

$$\eta = h^{-1}(Z, X).$$

To see this consider

$$\begin{aligned} F_{X|Z=z}(x) &= \Pr(X \leq x | Z = z) \\ &= \Pr(h(Z, \eta) \leq x | Z = z) \\ &= \Pr(\eta \leq h^{-1}(Z, x) | Z = z) \\ &= \Pr(\eta \leq h^{-1}(z, x)) \\ &= F_{\eta}(h^{-1}(z, x)) \\ &= h^{-1}(z, x) \\ &= \eta \end{aligned}$$

and note that the left hand side is observable. η is sometimes referred to as a control function. The key observation for identification of

$$\int g(X, \varepsilon)\, dF_{\varepsilon}(\varepsilon)$$

is then that the independence assumption implies

$$\varepsilon \perp\!\!\!\perp Z|\eta.$$

Hence,

$$\varepsilon \perp\!\!\!\perp h(Z,\eta)|\eta$$

so that

$$\varepsilon \perp\!\!\!\perp X|\eta.$$

Therefore, if the support of η does not depend on X we can identify

$$\int g(X,\varepsilon)\, dF_\varepsilon(\varepsilon) = \mathbb{E}[\mathbb{E}[Y|X,\eta]|X=x].$$

Notice that X is allowed to be correlated with ε so that a nonparametric regression of Y on X cannot be used here.

Overall, this approach allows the researcher to estimate the dependence between the endogenous variable and its effect in an easy way. The assumptions that are needed, however, are not weak. In particular, it has to be assumed that η is a scalar error term which is very restrictive in many contexts.

Mathematical Appendix

This mathematical appendix is not meant to be comprehensive. Instead, I would like to remind the reader of the results that are useful to have in mind when going through these lecture notes. There are many excellent books on the material, for example Abadir and Magnus (2005) for matrix algebra. Pagan and Ullah (1999) and Li and Racine (2007) both contain an Appendix A with a review of the most important important statistical results and definitions. So do many other econometrics textbooks.

Appendix A

Linear Algebra

A.1 Matrices

The $K \times L$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1L} \\ a_{21} & a_{22} & \cdots & a_{2L} \\ \vdots & & \ddots & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KL} \end{pmatrix}$$

is symmetric if $K = L$ and $a_{ij} = a_{ji}$ for all i, j. It is square if $K = L$.

The transpose of A is

$$A = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{K1} \\ a_{12} & a_{22} & \cdots & a_{K2} \\ \vdots & & \ddots & \vdots \\ a_{1L} & a_{2L} & \cdots & a_{KL} \end{pmatrix}.$$

A diagonal matrix is a square matrix that has no off-diagonal elements, that is $a_{ij} = 0$ if $i \neq j$. The diagonal of a matrix is the vector collecting the diagonal elements, $a_{11}, a_{22}, \ldots$.

A.2 Products between Vectors and Matrices

The product between a $K \times L$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1L} \\ a_{21} & a_{22} & \cdots & a_{2L} \\ \vdots & & \ddots & \vdots \\ a_{K1} & a_{K2} & \cdots & a_{KL} \end{pmatrix}$$

and an $L \times M$ matrix

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1M} \\ b_{21} & b_{22} & \cdots & b_{2M} \\ \vdots & & \ddots & \vdots \\ b_{L1} & b_{L2} & \cdots & b_{LM} \end{pmatrix}$$

is of dimension $K \times M$. The ijth element of this matrix AB is given by

$$\sum_{\ell=1}^{L} a_{i\ell} b_{\ell j}.$$

A can also be a row vector with $K = 1$ and B can be a column vector with $M = 1$.

The transpose of the product between two matrices is

$$(AB)' = B'A'.$$

Its inverse is

$$(AB)^{-1} = B^{-1}A^{-1}$$

if both, A and B are square matrices.

Elementwise multiplication of elements of two matrices is usually denoted by $\circ$. These matrices must be of the same dimension.

The Kronecker product is denoted by $\otimes$. If is used, for example for the $K \times L$ matrix A and the $M \times N$ matrix C, then it results in the $KM \times LN$ matrix forming all combinations of the elements of A and C.

A.3 Vector and Matrix Differentiation

A good paper on this is Magnus (2010). Suppose there is a column vector of length K,

$$a(b) = \begin{pmatrix} a_1(b) \\ a_2(b) \\ \vdots \\ a_K(b) \end{pmatrix}$$

whose elements depend on another column vector that is of length L. Then, we can define the $K \times L$ Jacobian matrix as

$$\frac{\partial a}{\partial b'} = \begin{pmatrix} \frac{\partial a_1}{\partial b_1} & \frac{\partial a_1}{\partial b_2} & \cdots & \frac{\partial a_1}{\partial b_L} \\ \frac{\partial a_2}{\partial b_1} & \frac{\partial a_2}{\partial b_2} & \cdots & \frac{\partial a_2}{\partial b_L} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial a_K}{\partial b_1} & \frac{\partial a_K}{\partial b_2} & \cdots & \frac{\partial a_K}{\partial b_L} \end{pmatrix}.$$

Observe that this notation reflects that the resulting matrix will be of dimension $K \times L$. The transpose of this matrix can then be denoted by $\partial a'/\partial b$.

If $a(b)$ is not a vector, but a scalar, that is if $K = 1$, then

$$\frac{\partial a}{\partial b} = \begin{pmatrix} \frac{\partial a}{\partial b_1} \\ \frac{\partial a}{\partial b_2} \\ \vdots \\ \frac{\partial a}{\partial b_L} \end{pmatrix}$$

is the gradient of that function and

$$\frac{\partial a}{\partial b \partial b'} = \begin{pmatrix} \frac{\partial^2 a}{\partial b_1 \partial b_1} & \frac{\partial^2 a}{\partial b_1 \partial b_2} & \cdots & \frac{\partial^2 a}{\partial b_1 \partial b_L} \\ \frac{\partial^2 a}{\partial b_2 \partial b_1} & \frac{\partial^2 a}{\partial b_2 \partial b_2} & \cdots & \frac{\partial^2 a}{\partial b_2 \partial b_L} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 a}{\partial b_L \partial b_1} & \frac{\partial^2 a}{\partial b_L \partial b_2} & \cdots & \frac{\partial^2 a}{\partial b_L \partial b_L} \end{pmatrix}$$

is the square, $L \times L$ Hessian.

To illustrate this, let A be of dimension $K \times L$ and x be of dimension $L \times 1$. Then,

$$Ax = \begin{pmatrix} \sum_{\ell=1}^{L} A_{1\ell} x_\ell \\ \sum_{\ell=1}^{L} A_{2\ell} x_\ell \\ \vdots \\ \sum_{\ell=1}^{L} A_{K\ell} x_\ell \end{pmatrix}$$

and therefore,

$$\frac{\partial Ax}{\partial x'} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1L} \\ A_{21} & A_{22} & \cdots & A_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & \cdots & A_{KL} \end{pmatrix} = A. \tag{A.3.1}$$

Furthermore, if B is $L \times K$,

$$x'B = \begin{pmatrix} \sum_{\ell=1}^{L} x_\ell B_{\ell 1} & \sum_{\ell=1}^{L} x_\ell B_{\ell 2} & \cdots & \sum_{\ell=1}^{L} x_\ell B_{\ell K} \end{pmatrix}.$$

Taking the derivative with respect to x gives

$$\frac{\partial x'B}{\partial x} = \begin{pmatrix} B_{11} & B_{12} & \cdots & B_{1K} \\ B_{21} & B_{22} & \cdots & B_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ B_{K1} & B_{K2} & \cdots & B_{LK} \end{pmatrix} = B.$$

Taking the derivative of $x'B$ with respect to x is the same as taking the transpose of the derivative of $B'x$ with respect to x', which according to (A.3.1) is the transpose of B', B.

Finally, if $K = L$ so that A is square and $L \times L$, then the derivative of the quadratic

form

$$x'Ax = \begin{pmatrix} \sum_{\ell=1}^{L} x_\ell A_{\ell 1} & \sum_{\ell=1}^{L} x_\ell A_{\ell 2} & \cdots & \sum_{\ell=1}^{L} x_\ell A_{\ell L} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_L \end{pmatrix}$$

$$= \sum_{\ell=1}^{L} x_\ell A_{\ell 1} x_1 + \sum_{\ell=1}^{L} x_\ell A_{\ell 2} x_2 + \cdots + \sum_{\ell=1}^{L} x_\ell A_{\ell L} x_L$$

$$= \sum_{\ell=1}^{L} \sum_{m=1}^{L} x_\ell A_{\ell m} x_m$$

is the gradient

$$\frac{\partial x'Ax}{\partial x} = \begin{pmatrix} \sum_{m=1}^{L} A_{1m} x_m + \sum_{\ell=1}^{L} x_\ell A_{\ell 1} \\ \sum_{m=1}^{L} A_{2m} x_m + \sum_{\ell=1}^{L} x_\ell A_{\ell 2} \\ \vdots \\ \sum_{m=1}^{L} A_{Lm} x_m + \sum_{\ell=1}^{L} x_\ell A_{\ell L} \end{pmatrix} = (A + A')x.$$

Appendix B

Analysis

B.1 Partial Derivative

Let there be two functions $u(x)$ and $v(x)$. Then, the product rule says that

$$\frac{\partial\left(u(x)\cdot v(x)\right)}{\partial x}=\frac{\partial u(x)}{\partial x}v(x)+u(x)\frac{\partial v(x)}{\partial x}.$$

and the quotient rule says

$$\frac{\partial\left(\frac{u(x)}{v(x)}\right)}{\partial x}=\frac{\frac{\partial u(x)}{\partial x}v(x)-u(x)\frac{\partial v(x)}{\partial x}}{\left(v(x)\right)^2}.$$

Leibnitz' rule says that for integrals of the form

$$\int_{a(x)}^{b(x)}f(x,y)\,dy$$

we have

$$\frac{\partial}{\partial x}\int_{a(x)}^{b(x)}f(x,y)\,dy=f\left(x,b(x)\right)\frac{\partial b(x)}{\partial x}-f\left(x,a(x)\right)\frac{\partial a(x)}{\partial x}+\int_{a(x)}^{b(x)}\frac{\partial f(x,y)}{\partial x}\,dy.$$

B.2 Total Derivative

For a function f of several variables, say x and y, the total derivative evaluated at x and y is

$$df(x,y) = \frac{\partial f(x,y)}{\partial x}\,dx + \frac{\partial f(x,y)}{\partial y}\,dy.$$

B.3 Implicit Function Theorem

For a vector-valued function $f(x,y)$ with $y = g(x)$ we have that

$$\frac{\partial g(x)}{\partial x_k} = -\left(\left.\frac{\partial f(x,y)}{\partial y}\right|_{y=g(x)}\right)^{-1} \left.\frac{\partial f(x,y)}{\partial x_k}\right|_{y=g(x)},$$

provided that

$$\left.\frac{\partial f(x,y)}{\partial y}\right|_{y=g(x)}$$

is invertible.

Appendix C

Statistics

C.1 Random Variables and Distribution Functions

Let X be a random variable, possibly of dimension K. Then, X is real valued and we can define the joint distribution function

$$F_X(x) \equiv \Pr(X \leq x).$$

The distribution function is non-decreasing and its limits are given by zero and one.

The density function is the derivative of the distribution function with respect to its argument, which may be vector valued with components x_k,

$$f_X(x) = \prod_k \frac{\partial F_X(x)}{\partial x_k}.$$

It may not exist.

C.2 Conditional Distributions

The conditional distribution of X given Y is defined as

$$F_{X|Y=y}(x) \equiv \frac{\Pr(X \leq x, Y \leq y)}{\Pr(Y \leq y)}$$

and the corresponding density is denoted by $f_{X|Y=y}$.

C.3 Independence

Two random variables are independent if

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

C.4 First Moments

The mean of a scalar random variable X is

$$\mathbb{E}[X] = \int x dF_X(x) = \int x f_X(x),$$

where the second equality holds if the density function exists. Higher order moments such as the variance can be defined accordingly. Likewise, we have

$$\mathbb{E}[X|Y=y] = \int x dF_{X|Y=y}(x) = \int x f_{X|Y=y}(x)$$

and

$$\mathbb{E}[g(X)] = \int g(x) dF_{X|Y=y}(x) = \int g(x) f_{X|Y=y}(x).$$

For two random variables X and Y, the law of iterated expectations is

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

C.5 Second Moments

The variance of a random variable X is defined as

$$var(X) \equiv \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

One can show that

$$var(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

and that, for two constants a and b,

$$var(a+bX) = b^2 var(X).$$

The covariance between two random variables X and Y is defined as

$$cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

and we can show that

$$cov(X,Y) = \mathbb{E}\left[X(Y - \mathbb{E}[Y])\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Moreover,

$$cov(a + bX, c + dY) = bc \cdot cov(X,Y)$$

provided that a, b, c and d are constants. The results for variances follow for the special case in which $X = Y$, $a = c$ and $b = d$.

If X is a column vector with elements $X_1, \ldots, X_K$, then

$$var(X) \equiv \begin{bmatrix} var(X_1) & cov(X_1,X_2) & \cdots & cov(X_1,X_K) \\ cov(X_2,X_1) & var(X_2) & \cdots & cov(X_2,X_K) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_K,X_1) & cov(X_K,X_2) & \cdots & var(X_K) \end{bmatrix}$$

is the corresponding variance-covariance matrix. It is symmetric, as $cov(X_k,X_l) = cov(X_l,X_k)$.

Bibliography

Abadie, A. and G. Imbens (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica 74*(1), 235–267.

Abadir, K. M. and J. R. Magnus (2005). *Matrix Algebra*. Econometric Exercises. New York, USA: Cambridge University Press.

Ahn, H. and J. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 2–29.

Ai, C. and E. C. Norton (2003). Interaction terms in logit and probit models. *Economics Letters 80*(1), 123–129.

Ambirajan, S. (1995). The delayed emergence of econometrics as a separate discipline. *Measurement, quantification, and economic analysis: numeracy in economics*, 198.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic L 19*, 1483–1536.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics 24*(1-2), 3–61.

Amemiya, T. (2009). Thirty-five years of journal of econometrics. *Journal of Econometrics 148*(2), 179–185.

Andrews, D. W. K. and M. M. A. Schafgans (1988). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies 65*(3), 497–517.

Angrist, J. and J. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies 67*(3), 499–527.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996, June). Identification of causal effects using instrumental variables. *Journal of the American Statistical Society 91*(434), 444–455.

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings. *Quarterly Journal of Economics 106*(4), 979–1014.

Angrist, J. D. and A. B. Krueger (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives 15*(4), 69–85.

Arrow, K. J. (1951). *Social cChoice and Individual Values*. New York, NY, USA: Wiley and Sons.

Basmann, R. L. (1957, jan). A generalized classical method of linear estimation of coefficients in a

structural equation. *Econometrica 25*(1), 77–83.

Bates, G. and J. Neyman (1952). *Contributions to the theory of accident proneness II: True or false contagion.*, pp. 255–275. Number 1. University of California Publications in Statistics 1.

Battistin, E. and A. Chesher (2004). The impact of measurement error on evaluation methods based on strong ignorability. Working Paper, University College London, London, UK.

Battistin, E. and E. Rettore (2008a). Ineligible and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics 142*(2), 715–730.

Battistin, E. and E. Rettore (2008b). Ineligible and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics 142*(2), 715–730.

Beresteanu, A. and F. Molinari (2008). Asymptotic properties for a class of partially identified models. *Econometrica 76*(4), 763–814.

Berry, S., J. Levinsohn, and A. Pakes (1995, July). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–90.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics 25*(2), 242–262.

Börsch-Supan, A. (1990). On the compatibility of nested logit models with utility maximization. *Journal of Econometrics 43*(3), 373–388.

Bound, J. and D. A. Jaeger (2000). Do compulsory school attendance laws alone explain the association between quarter of births and earnings? *Research in Labor Economics 19*, 83–108.

Bresnahan, T. F. and P. C. Reiss (1991). Entry and competition in concentrated markets. *Journal of Political Economy 99*(5), 977–1009.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cardell, N. S. (1997, April). Variance components structures for the extreme-value and logistic distributions with application to models of heterogeneity. *Econometric Theory 13*(2), 185–213.

Cargill, T. (1974). Early applications of spectral methods to economic time series. *History of Political Economy 6*(1), 1.

Cawley, J. (2011). A guide (and advice) for economists on the us junior academic job market. Working PaperJa.

Chamberlain, G. (1986). Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics 32*, 189–218.

Chesher, A. (2002, September). Instrumental values. cemmap Working Paper CWP17/02.

Chesher, A. (2006). Notes on identification. Mimeograph.

Chesher, A. (2007). Instrumental values. *Journal of Econometrics 139*(1), 15–34.

Chesher, A. and J. M. C. Santos Silva (2002). Taste variation in discrete choice models. *Review of Economic Studies 69*(1), 147–168.

Christ, C. (1985). Early progress in estimating quantitative economic relationships in america. *American Economic Review 75*(6), 39–52.

Christ, C. (1994). The cowles commission's contributions to econometrics at chicago, 1939-1955. *Journal of Economic Literature 32*(1), 30–59.

Cournot, A. A. (1838). *Researches into the Mathematical Principles of the Theory of Wealth* (1897 ed.).

New York: Macmillan.

Cowles, A. (1933). Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, 309–324.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Davidson, R. and J. G. MacKinnon (2003). *Econometric Theory and Methods*. New York: Oxford University Press.

Duffie, D. and K. Singleton (1993). Simulated moments estimation of markov models of asset prices. *Econometrica 61*(4), 929–952.

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute 22*, 975–813.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume II, Chapter 13, pp. 775–826. Elsevier.

Engle, R. F., D. F. Hendry, and J.-F. Richard (1983, March). Exogeneity. *Econometrica 51*(2), 277–304.

Epstein, R. (1987). *A history of econometrics*. Amsterdam.

Ferrer-i Carbonell, A. and P. Frijters (2004). How important is methodology for the estimates of the determinants of happiness?*. *Economic Journal 114*(497), 641–659.

Fisher, F. (1966). *The Identification Problem in Econometrics*. McGraw Hill.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Frisch, R. and F. V. Waugh (1933, oct). Partial time regressions as compared with individual trends. *Econometrica 1*(4), 387–401.

Frölich, M. (2007). Nonparametric iv estimation of local average treatment effects with covariates. *Journal of Econometrics 139*(1), 35–75.

Gallant, A. and G. Tauchen (1996). Which moments to match? *Econometric Theory 12*, 657–681.

Gallant, A. R. and D. W. Nychka (1987, mar). Semi-nonparametric maximum likelihood estimation. *Econometrica 55*(2), 363–390.

Gerfin, M. (1996). Parametric and semi-parametric estimation of the binary response model of labour market participation. *Journal of Applied Econometrics 11*(3), 321–339.

Goldberg, P. K. (1995, jul). Product differentiation and oligopoly in international markets: The case of the u.s. automobile industry. *Econometrica 63*(4), 891–951.

Goldberger, A. (1991). *A course in econometrics*. Harvard Univ Pr.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica 40*(6), 979–1001.

Gourieroux, C. and A. Monfort (1996). *Simulation-Based Econometric Methods*. Oxford University Press, USA.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 8*(S1), S85–S118.

Gourieroux, C. and A. Montfort (1995). Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory 11*, 195–228.

Gronau, R. (1973). The effect of children on the housewife's value of time. *Journal of Political Economy 81*(2, Part 2: New Economic Approaches to Fertility), S168–S199.

Gronau, R. (1974, nov). Wage comparisons–a selectivity bias. *The Journal of Political Economy 82*(6),

1119–1143.

Hahn, J., P. Todd, and W. van der Klaauw (2001, January). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Han, A. and J. A. Hausman (1990, jan). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics 5*(1), 1–28.

Hansen, L. and K. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 1269–1286.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Hausman, J. and D. McFadden (1984, sep). Specification tests for the multinomial logit model. *Econometrica 52*(5), 1219–1240.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica 46*(6), 1251–1271.

Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. In *The economics of new goods*, pp. 207–248. University of Chicago Press.

Hayashi, F. (2000). Econometrics. *Princeton University Press*.

Heckman, J. (1974, jul). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679–694.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement 5*, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–162.

Heckman, J. J. (1981). Heterogeneity and state dependence. In S. Rosen (Ed.), *Studies in Labor Markets*. University of Chicago Press.

Heckman, J. J. (1990, May). Varieties of selection bias. *American Economic Review: Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association 80*(2), 313–318.

Heckman, J. J. (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources 32*(3), 441–462.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review 76*(1), 1–27.

Heckman, J. J. and G. Borjas (1980). Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence. *Economica 47*, 247–283.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006, December). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. J. Vytlacil (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources 33*(4), 974–987.

Heckman, J. J. and E. J. Vytlacil (1999, April). Local instrumental variables and latent variables models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences 96*, 4730–4734.

Heckman, J. J. and E. J. Vytlacil (2000). The relationship between treatment parameters within a latent variable framework. *Economics Letters 66*(1), 33–39.

Heckman, J. J. and E. J. Vytlacil (2001). Local instrumental variables. In C. Hsiao, K. Morimune, and J. Powell (Eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, pp. 1–46. Cambridge: Cambridge University Press.

Heckman, J. J. and E. J. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica 73*(3), 669–738.

Hildreth, C. (1986). *The Cowles Commission in Chicago, 1939-1955*, Volume 271. Springer.

Holland, P. W. (1986, December). Statistics and causal inference. *Journal of the American Statistical Association 81*(396), 945–960.

Holly, A. and J. Sargan (1982). Testing for exogeneity in a limited information framework. Cahiers de Recherches Economiques, No. 8204. Université de Lauseanne.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 505–531.

Hsiao, C. (1983). Identification. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume I, Chapter 4, pp. 223–283. North-Holland.

Hurwicz, L. (1950). Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10. New York: John Wiley and Sons.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics 58*(1), 71–120b.

Imbens, G. and T. Lemieux (2008). Regression discontinuity designs: A guide to pratice. *Journal of Econometrics 142*, 615 – 635.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and W. K. Newey (2003, December). Identification and estimation of triangular simultaneous equations models without additivity. MIT Working Paper. Presented at the 2003 EC2 conference held in London.

Johnson, R. K. (2010). *The elements of MATLAB style*. Cambridge University Press.

Kalbfleisch, J. D. D. and R. L. Prentice (1973, aug). Marginal likelihoods based on cox's regression and life model. *Biometrika 60*(2), 267–278.

Kalbfleisch, J. D. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kaplan, E. L. and P. Meier (1958, jun). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Kapteyn, A., J. Smith, and A. van Soest (2007). Vignettes and self-reports of work disability in the united states and the netherlands. *The American Economic Review 97*(1), 461–473.

Keane, M. P. (2010). A structural perspective on the experimentalist school. *Journal of Economic Perspectives 24*(2), 47–58.

Kelejian, H. H. (1971). Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association 66*(334), 373–374.

Kiefer, N. M. (1988, June). Economic duration data and hazard functions. *Journal of Economic Literature 26*(2), 646–679.

Klein, R. and R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 387–421.

Klein, T. J. (2010). Heterogeneous treatment effects: Instrumental variables without monotonicity? *Journal of Econometrics 155*(2), 99–116.

Koopmans, T. C. (1949). Identification problems in economic model construction. *Econometrica 17*(2), 125–144.

Koopmans, T. C. and O. Reiersøl (1950). The identification of structural characteristics. *Annals of Mathematical Statistics 21*(2), 165–181.

Koopmans, T. C., H. Rubin, and R. B. Leipnik (1950). Measuring the equation system of dynamic economies. In *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10. New York: John Wiley and Sons.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy 74*(2), 132–157.

Leamer, E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*(1), 31–43.

Li, Q. and J. S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Luce, R. (1959). *Individual Choice Behavior.* John Wiley.

Luce, R. D. and P. Suppes (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. H. Galanter (Eds.), *Handbook of Mathematical Psychology*, Volume 3, New York, NY, pp. 249–410. Wiley.

Maddala, G. S. (1983). *Qualitative and Limited Dependent Variable Models in Econometrics*. Cambridge University Press.

Magnus, J. R. (2010). On the concept of matrix derivative. *Journal of Multivariate Analysis 101*, 2200–2206.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3*(3), 205–228.

Manski, C. (2010). Policy analysis with incredible certitude. Technical report, National Bureau of Economic Research.

Manski, C. and T. Thompson (1986). Operational characteristics of maximum score estimation. *Journal of Econometrics 32*(1), 85–108.

Manski, C. F. (1988a). *Analog Estimation Methods in Econometrics*. Chapman and Hall.

Manski, C. F. (1988b, sep). Identification of binary response models. *Journal of the American Statistical Association 83*(403), 729–738.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. New York: Springer.

Marschak, J. (1953). Economic measurements for policy and prediction. In W. C. Hood and T. C. Koopmans (Eds.), *Studies in Econometric Methods*, Number 14 in Cowles Commission Monograph, Chapter 1, pp. 1–48. Wiley.

Marschak, J. (1959). Binary choice constraints on random utility indicators. In K. Arrow (Ed.), *Mathematical Methods in the Social Sciences*, Stanford, CA, pp. 312–329. Stanford University Press.

Matzkin, R. L. (1992, mar). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

McFadden, D. (1974a). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of Econometrics*, New York, NY. Academic Press.

McFadden, D. (1974b). The measurement of urban travel demand. *Journal of Public Economics 4*(4), 303–328.

McFadden, D. (1977). A closed-form multinomial choice model without the independence from irrelevant alternatives restrictions. WP 7703, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley, CA.

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull (Eds.), *Spatial Interaction Theory and Planning Models*. London, U.K.: Edward Elgar.

McFadden, D. (1981). Econometric models of probabilistic choice. In C. F. Manski and McFa (Eds.), *Structural Analysis of Discrete Data*. Cambridge, MA, USA: MIT.

McFadden, D. (1984). Econometric analysis of qualitative response models. In Z. Griliches and M. D. Intriligator (Eds.), *Handbook of Econometrics*, Volume II, Chapter 24, pp. 1395–1457. Elsevier.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*(5), 995–1026.

McFadden, D. and M. K. Richter (1971). On the extension of a probability to the boolean algebra generated by a family of events, with a application to choice theory. Working Paper.

Moore, H. L. (1914). *Economic Cycles: Their Law and Cause*. New York: Macmillan.

Morgan, M. S. (1990). *The History of Econometric Ideas*. Cambridge, MA: Cambridge University Press.

Nadaraya, E. (1964). Some new estimates for distribution functions. *Theory of Probability and Its Applications 9*(3), 497–500.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. *Econometrica 69*(2), 307–342.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2112–2245. Amsterdam: North Holland.

Newson, R. et al. (2003). Multiple test procedures and smile plots. *The Stata Journal 3*(2), 100–32.

Norton, E. (2012). Log odds and ends. NBER Working Paper No. 18525.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge, United Kingdom: Cambridge University Press.

Pakes, A. and D. Pollard (1989). Simulation and the asymptotics of optimization estimators. *Econometrica 57*(5), 1027–1057.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics 25*(3), 303–325.

Powell, J. L. (1986, nov). Symmetrically trimmed least squares estimation for tobit models. *Econometrica 54*(6), 1435–1460.

Powell, J. L. (1994). Estimation of semiparametric models. In R. F. Engle and D. McFadden (Eds.), *Handbook of Econometrics*, Volume 4, pp. 2443–2521.

Prentice, R. and L. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics 34*(1), 57–67.

Puhani, P. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters 115*, 85–87.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Ruud, P. A. (2000). *An Introduction to Classical Econometric Theory*. Oxford, New York: Oxford University Press.

Sargan, J. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 393–415.

Schultz, H. (1925). The statistical law of demand as illustrated by the demand for sugar. *Journal of Political Economy 33*(6), 577–631.

Silvia, P. (2007). *How to write a lot: A practical guide to productive academic writing.* American Psychological Association.

Small, K. (1987). A discrete choice model for ordered alternatives. *Econometrica 55*(2), 409–424.

So Im, K., S. Ahn, P. Schmidt, and J. Wooldridge (1999). Efficient estimation of panel data models with strictly exogenous explanatory variables. *Journal of Econometrics 93*(1), 177–201.

Spiegel, M. (2012). Reviewing less—progressing more. *Review of Financial Studies*.

Stigler, G. (1949). A survey of contemporary economics. *The Journal of Political Economy 57*(2), 93–105.

Stigler, G. (1965). Statistical studies in the history of economic thought. *Essays in the History of Economics*, 31–50.

Stock, J. and M. Watson (2003). *Introduction to econometrics*. Addison Wesley New York.

Stock, J. H. and F. Trebbi (2003). Who invented instrumental variable regression? *Journal of Economic Perspectives 17*(3), 177–194.

Strunk, Jr., W. and E. B. White (1999). *The Elements of Style* (4th ed.). Longman.

Theil, H. (1953). Repeated least squares applied to complete equation systems. Den Hague: Central Planning Bureau.

Thistlethwaite, D. and D. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology 51*, 309–317.

Thomson, W. (2001). *A Guide for the Young Economist*. MIT Press.

Tobin, J. (1958, jan). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge, UK: Cambridge University Press.

Tukey, J. (1991). The philosophy of multiple comparisons. *Statistical Science 6*(1), 100–116.

van den Berg, G. (2001). Duration models: Specification, identification and multiple durations. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 5, pp. 3381–3460. Elsevier.

Verbeek, M. (2004). *A Guide to Modern Econometrics*. John Wiley & Sons.

Vytlacil, E. (2002, January). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Wald, A. (1940). The fitting of straight lines of both variables are subject to error. *Annals of Mathematical*

Statistics 11, 284–300.

Wald, A. (1950). Note on the identification of economic relations. In *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10. New York: John Wiley and Sons.

Watson, G. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A 26*(4), 359–372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

Willis, R. J. and S. Rosen (1979). Education and self-selection. *Journal of Political Economy 87*(5, Part 2: Education and Income Distribution), S7–S36.

Wolpin, K. (1984). An estimable dynamic stochastic model of fertility and child mortality. *The Journal of Political Economy 92*(5), 852–874.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Working, E. (1927). What do statistical "demand curves" show? *Quarterly Journal of Economics 41*(2), 212.

Wright, P. (1915). Moore's economic cycles. *Quarterly Journal of Economics 29*(3), 631–641.

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oil*. New York: MacMillan.

Wu, D. (1973). Alternative tests of independence between stochastic regressors and disturbances. *Econometrica 41*(4), 733–750.

Xie, Y. and C. F. Manski (1989). The logit model and response-based samples. *Sociological Methods and Research 17*(3), 283–302.

Yates, F. (1937). The design and analysis of factorial experiments. Technical report, Technical Communication no. 35 of the Commonwealth Bureau of Soils.

Yitzhaki, S. (1989). On using linear regression in welfare economics. Working paper No. 217, Department of Economics, Hebrew University, Jerusalem, Israel.

Nomenclature

ι_T	ones vector of length T
$\otimes$	Kronecker product
$\partial f(x)/\partial x'$	Jacobian matrix
$\partial^2 f(x)/\partial x \partial x'$	Hessian, where f is scalar valued and x is a column vector
$\theta(S)$	structural parameter
I_T	identity matrix of size $T \times T$
S	structure
$s_i(\theta)$	score of the likelihood function
x_i	row vector of explanatory variables
c.d.f.	cumulative distribution function $F_{\varepsilon_i}(e) \equiv \Pr(\varepsilon_i \leq e)$
p.d.f.	probability density function

Index