Microeconometrics (CentER)

Part 2: Binary and Multinomial Choice

Tobias J. Klein, Tilburg University

May 9, 2026

Introduction

Economic research

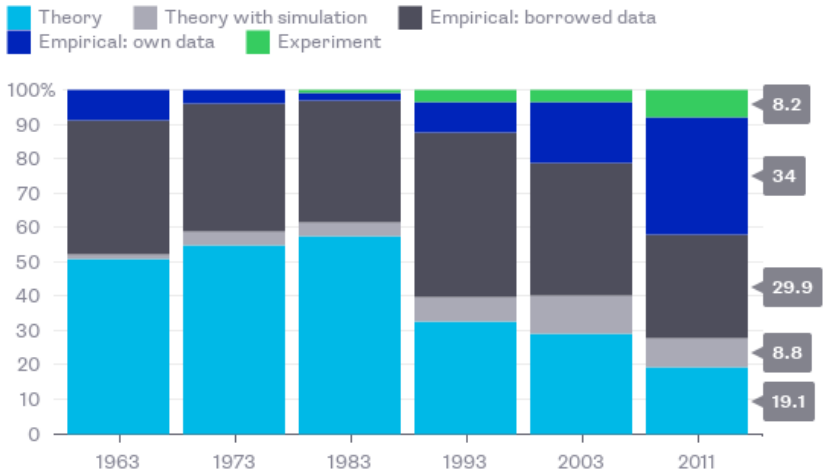
- A big part of economic research is analyzing data.
- For this, we need tools and guidance:
  - tools: econometrics
  - guidance: economic theory.
- Today it is more exciting than ever to do empirical work:
  - data availability (admin data, public data)
  - lots of methods (econometrics)
  - AI tools for coding, data work.

Ph.D. in Economics

- Overall goal of a Ph.D. trajectory: help you become a productive researcher.
- Intermediate goal: great job market paper and great first job.
- Most important choices:
  - field of interest
  - research questions
  - supervisor (look at placement record, talk to current students, look at his or her current research productivity, traveling activity, academic background).
- Ph.D. "core": Micro, Macro, Econometrics; goal: give you general background, teach you methods used in economic research. But not representative of research interests in economics. Think of it as the toolbox.

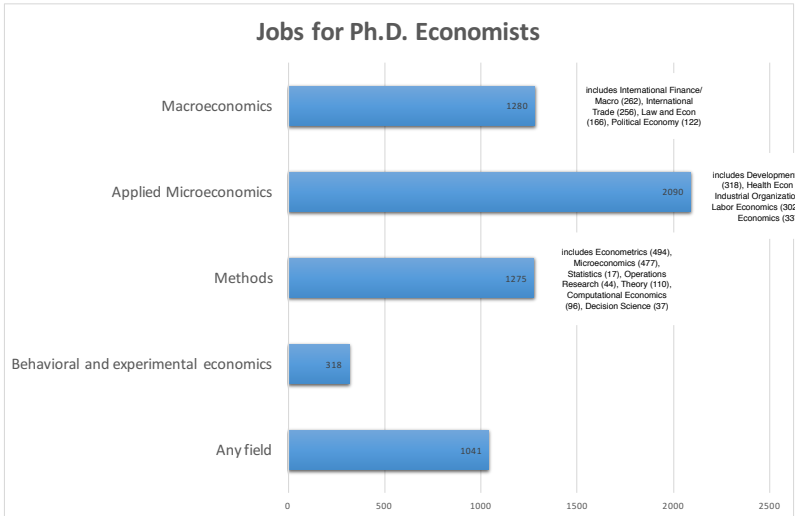
The Changing Nature of Economic Research

Methodology of articles in top economics journals, as percent of total

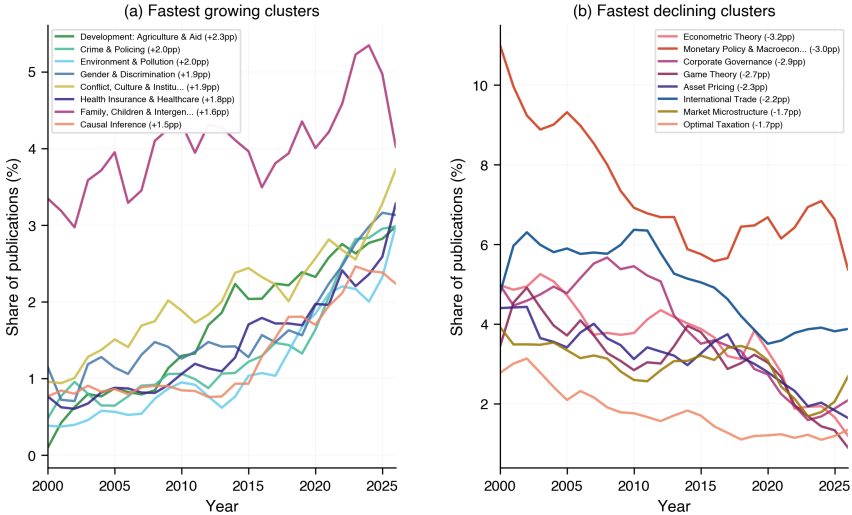


Source: Daniel S. Hamermesh, Journal of Economic Literature

BloombergView



Source: https://econjobmarket.org/stats.php, accessed January 10, 2016



Notes: Share of publications for the fastest-growing (left) and fastest-declining (right) individual clusters. Percentage-point changes in legend compare 2000–04 to 2021–26 shares (3-year rolling average).

See Harris (2026)

Most Ph.D. economists do not work at universities

Table 1

Number of *AER*-Equivalent Publications of Graduating Cohorts from 1986 to 2000

	Percentiles of graduates' AER-equivalent publications 6 years after PhD									*Average cohort size*	*Publishing grads (%)*
	99th	*95th*	*90th*	*85th*	*80th*	*75th*	*70th*	*60th*	*50th*		
Harvard	4.31	2.36	1.47	1.04	0.71	0.41	0.30	0.12	0.04	30.5	66.3
Chicago	2.88	1.71	1.04	0.72	0.51	0.33	0.19	0.06	0.01	27.3	59.4
U Penn	3.17	1.52	1.01	0.60	0.40	0.27	0.22	0.06	0.02	19.3	59.5
Stanford	3.43	1.58	1.02	0.67	0.50	0.33	0.23	0.08	0.03	24.7	67.9
MIT	4.73	2.87	1.66	1.24	0.83	0.64	0.48	0.20	0.07	25.5	70.0
UC Berkeley	2.37	1.08	0.55	0.35	0.20	0.13	0.08	0.04	0.02	28.0	62.4
Northwestern	2.96	1.92	1.15	0.93	0.61	0.47	0.30	0.14	0.06	10.1	65.8
Yale	3.78	2.15	1.22	0.83	0.57	0.39	0.19	0.08	0.03	15.7	64.8
U MI, Ann Arbor	1.85	0.77	0.48	0.29	0.17	0.09	0.05	0.02	0.01	19.1	54.0
Columbia	2.90	1.15	0.62	0.34	0.17	0.10	0.06	0.01	0.01	17.4	54.8
Princeton	4.10	2.17	1.79	1.23	1.01	0.82	0.60	0.36	0.19	16.2	76.1
UCLA	2.59	0.89	0.49	0.26	0.14	0.06	0.04	0.02	0	17.9	48.5
NYU	2.05	0.89	0.54	0.20	0.07	0.03	0.02	0.01	0	11.7	46.0
Cornell	1.74	0.65	0.40	0.23	0.12	0.07	0.05	0.02	0.01	17.3	57.9
U WI, Madison	2.39	0.89	0.51	0.31	0.20	0.11	0.06	0.03	0.01	25.0	60.3
Duke	1.37	1.03	0.59	0.49	0.23	0.19	0.11	0.05	0.02	7.8	59.8
Ohio State U	0.69	0.41	0.13	0.07	0.04	0.02	0.02	0.01	0	15.9	47.9
U Maryland	1.12	0.37	0.23	0.10	0.07	0.05	0.03	0.01	0.01	13.5	56.2
Rochester	2.93	1.94	1.56	1.21	1.14	0.98	0.70	0.34	0.17	8.7	78.5
U TX, Austin	0.92	0.53	0.21	0.06	0.05	0.02	0.01	0	0	10.3	38.3
Minnesota	2.76	1.20	0.68	0.46	0.29	0.21	0.12	0.04	0.01	22.2	59.5
U IL, Urbana-Ch	1.00	0.38	0.21	0.10	0.06	0.04	0.03	0.01	0.01	26.4	54.8
UC Davis	1.90	0.66	0.42	0.27	0.12	0.08	0.05	0.02	0.01	6.2	53.8
Toronto	3.13	1.85	0.80	0.61	0.29	0.19	0.15	0.07	0.03	6.4	64.6
British Columbia	1.51	1.05	0.71	0.60	0.52	0.45	0.26	0.22	0.11	4.5	73.1
UC San Diego	2.29	1.69	1.17	0.88	0.74	0.60	0.46	0.30	0.18	6.1	78.3
U Southern CA	3.44	0.34	0.14	0.09	0.03	0.02	0.02	0.01	0	4.9	43.8
Boston U	1.59	0.49	0.21	0.08	0.05	0.02	0.02	0	0	12.5	41.0
Penn State U	0.93	0.59	0.25	0.12	0.08	0.06	0.02	0.01	0.01	7.1	51.4
Carnegie Mellon	2.50	1.27	1.00	0.86	0.71	0.57	0.52	0.21	0.09	2.9	66.7
Non-Top-30	1.05	0.31	0.12	0.06	0.04	0.02	0.01	0	0	16.8	40.1

Source: Based on the authors own calculations using the data described in the paper.
Note: We order the table using the Coupé (2003) ranking of economics departments.

See Conley and Önder (2014, JEP)

This course

- Methods course, from an applied perspective.
- Focus:
  - first half: estimating causal effects (when exogeneity does not hold)
  - second half: binary and multinomial choice, bring together economic theory and econometric methods, "structural approach".

Prerequisites

- Formal prerequisites: Quantitative Methods (CentER) and Foundation of Econometrics (CentER).
- Also, note that "All non-CentER students should ask formal permission from the Director of Graduate Studies in Economics BEFORE the start of the course. Please send your request for permission including grade list, CV and motivation letter to CentER Graduate School at tisem-msc-rm@tilburguniversity.edu. Note that asking permission is not just a formality."
- Reason: you need to be familiar with
  - maximum likelihood estimation
  - regressions, testing, IV estimation.

Overview

- Binary choice
  - Illustration: Logit model
  - Binary choice models, more generally
  - Generalization: ordered choice
- Multinomial choice
  - Multinomial logit model
  - Generalizations

Binary choice

Overview

- Binary choice
  - Illustration: Logit model
  - Binary choice models, more generally
  - Generalization: ordered choice
- Multinomial choice
  - Multinomial logit model
  - Generalizations

Binary choice

- Binary choice: y_i takes on two values.
- Examples: college attendance, smoking, working, being healthy, having children.
- Typical models: probit, logit, linear probability model.
- Will first look at the standard linear index class of models and in particular the logit model, then discuss generalizations.

Standard (linear index) binary choice model

- The econometric model is
$$y_i = 1\{x_i\beta \geq \varepsilon_i\}.$$
- Unobserved ε_i, is assumed to be distributed independently of x_i with known cumulative distribution function (c.d.f.) F_{ε_i} (the c.d.f. is defined as $F_{\varepsilon_i}(e) \equiv \Pr(\varepsilon_i \leq e)$).

Sample log likelihood function

- Recall that the model is $y_i = 1\{x_i\beta \geq \varepsilon_i\}$ with ε_i independent of x_i, and that $F_{\varepsilon_i}(e) \equiv \Pr(\varepsilon_i \leq e)$, so that

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

  (first equality: substitute in the model, second: swap order and use independence, third: definition c.d.f.)
- There are only two possible values y_i can take on, so $\Pr(y_i = 0|x_i) = 1 - \Pr(y_i = 1|x_i)$.
- From this we can calculate the log likelihood contributions by taking logs.
- The first derivative of those log likelihood contributions are the score contributions.

Binary choice

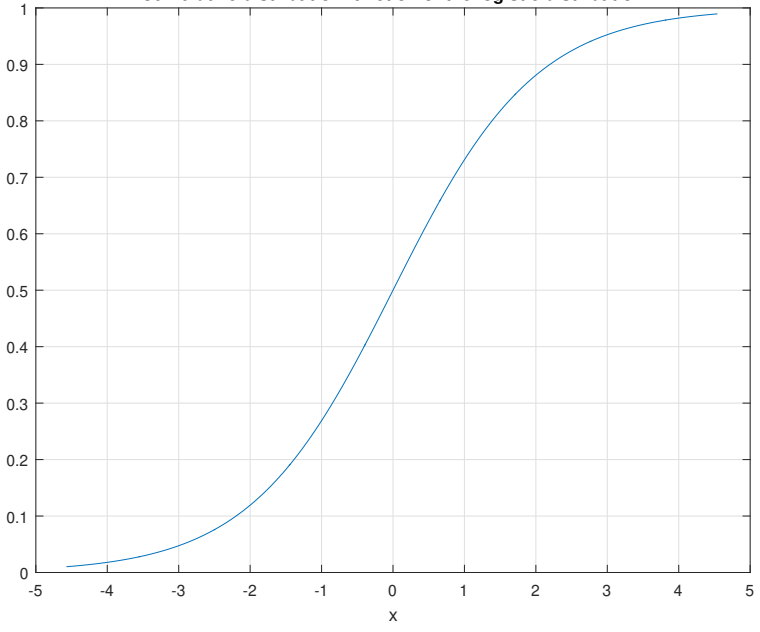
Illustration: Logit model

Logit model

- Assuming that the distribution of ε_i is logistic (with location parameter $\mu = 0$ and scale parameter $s = 1$) gives the logit model.
- The logistic distribution has c.d.f.

$$F_{\varepsilon_i}(e) = \frac{\exp(e)}{1 + \exp(e)}.$$

- ε_i has a mean of zero and a variance of $\pi^2/3 \approx 3.2899$.


Cumulative distribution function of the logistic distribution
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
x

Log likelihood and score contributions

- We have

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

  and the log likelihood contribution is

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

- The score contribution is the derivative thereof with respect to β (show this yourself),

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

Likelihood function in Matlab

```
function [nll ns] = nll_logit(beta,y,X)
% negative average log likelihood and score for logit model

prob1=exp(X*beta)./(1+exp(X*beta)); %probability to choose y=1
l=log(y.*prob1+(1-y).*(1-prob1)); %likelihood
s= (y.*(1-prob1).*X-(1-y).*prob1.*X); %score

nll=-mean(l); %negative of the average log likelihood
ns=-mean(s); %negative of the average score
```

Monte Carlo

- Generate many data sets and estimate the parameters each time by ML.
- Setup here: x_i has a negative effect on $\Pr(y_i = 1|x_i)$
  - x_i drawn from χ^2 distribution with 10 degrees of freedom
  - $y_i = 1$ if -0.1 times x_i is greater or equal to $\varepsilon_i = \varepsilon_{0i} - \varepsilon_{1i}$, where ε_{0i} and ε_{1i} are type 1 extreme value (you can think of ε_i as being drawn from a logistic distribution)
  - Monte Carlo with 1000 replications and $N = 100$ observations.

Intermezzo: Drawing from distributions

- A computer can easily generate (pseudo-)random draws from a uniform distribution.
- We can translate uniform draws u into draws from a distribution F_Z by calculating $F_Z^{-1}(u)$.
- Reason: $\Pr(F_Z^{-1}(U) \leq z) = \Pr(U \leq F_Z(z)) = F_Z(z)$.
- For example, for the logistic distribution, we can show that

$$F_{\varepsilon_i}^{-1}(u) = \log\left(\frac{u}{1-u}\right)$$

for $u \in (0, 1)$. So, we can take uniform draws and convert them "manually".

Monte Carlo: Setting the stage

```
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: Loop

```
% Monte Carlo
for i = 1:repetitions
    % generate data
    x=chi2rnd(10,N,1);
    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
```

Monte Carlo: Results

- Across replications, the mean is about -0.1, so the estimator appears to be consistent.
- The variance is about 0.0006.

Binary choice

Binary choice models, more generally

Overview

Binary choice

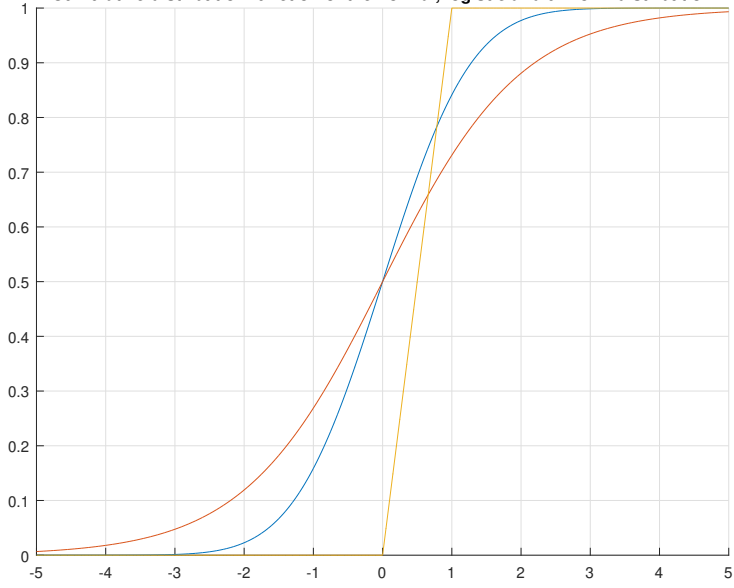
- Illustration: Logit model
- Binary choice models, more generally
- Generalization: ordered choice

Multinomial choice

- Multinomial logit model
- Generalizations

Alternative distributional assumptions

- Have assumed that $y_i = 1\{x_i\beta \geq \varepsilon_i\}$—which is a linear-in-$x_i$ index/threshold crossing model—and that ε_i follows logistic distribution, gave logit model.
- Can maintain functional form assumption but instead assume that ε_i follows the standard normal distribution, gives probit model.
- Or that it follows the uniform distribution; then, we obtain the so-called Linear Probability Model (for reasons that will become clear below).


Cumulative distribution function of the normal, logistic and uniform distribution
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Normalizations

- In the model $y_i = 1\{x_i\beta \geq \varepsilon_i\}$, the parameter β is only identified up to so-called normalizations: the inequality

$$x_i\beta \geq \varepsilon_i$$

still holds when we perform positive monotone transformations to each side, such as adding a constant or multiplying by a positive constant.

- By picking a distribution for ε_i that has a given mean and variance, we impose such normalizations.
- For example, if we pick the same distribution with a higher variance, say a normal distribution with a variance of 4 instead of 1, then this will re-scale β by 2 (the square root of $4/1$).

Let the row vector x_i contain 1 as the first element so that the first element of a column vector post-multiplying x_i is an intercept. Think of two versions of the probit model:

1. $y_i = 1\{x_i\beta \geq \varepsilon_i\}$ with $\varepsilon_i \sim \mathcal{N}(0,1)$ independent of x_i; gives

$$\Pr(y_i = 1|x_i) = \Pr(\varepsilon_i \leq x_i\beta|x_i) \overset{\text{indep}}{=} \Pr(\varepsilon_i \leq x_i\beta) = \Phi(x_i\beta),$$

   where Φ is the standard normal c.d.f.

2. $y_i = 1\{x_i\tilde{\beta} \geq \tilde{\varepsilon}_i\}$ with $\tilde{\varepsilon}_i \sim \mathcal{N}(\mu, \sigma^2)$ independent of x_i; gives

$$\begin{aligned}\Pr(y_i = 1|x_i) &= \Pr(\tilde{\varepsilon}_i \leq x_i\tilde{\beta}|x_i) \overset{\text{indep}}{=} \Pr(\tilde{\varepsilon}_i \leq x_i\tilde{\beta}) \\ &= \Pr\left(\underbrace{\frac{\tilde{\varepsilon}_i - \mu}{\sigma}}_{\sim\mathcal{N}(0,1)} \leq \frac{x_i\tilde{\beta} - \mu}{\sigma}\right) = \Phi\left(\frac{x_i\tilde{\beta} - \mu}{\sigma}\right).\end{aligned}$$

Why we need normalizations

- Model 1 is a special case of model 2:
  - location normalization: $\mu = 0$
  - scale normalization: $\sigma = 1$.
- Model 2 is too general:
  - if we increase μ and at the same time the first element of $\tilde{\beta}$ (the intercept) by the same amount, then we get the same implied choice probability
  - if we multiply μ, σ and $\tilde{\beta}$ by the same factor, we again get the same implied choice probability.

Normalizations are necessary for identification

- Identification: we can learn about the parameters from choice probabilities.
- If the choice probabilities can be the same for different values of μ and σ, then this is not possible. The model is not identified.
- If the model is not identified, then we cannot find a unique maximum of the likelihood function.
- Bottom line: we need to fix the location and the scaling.

Assumptions and normalizations

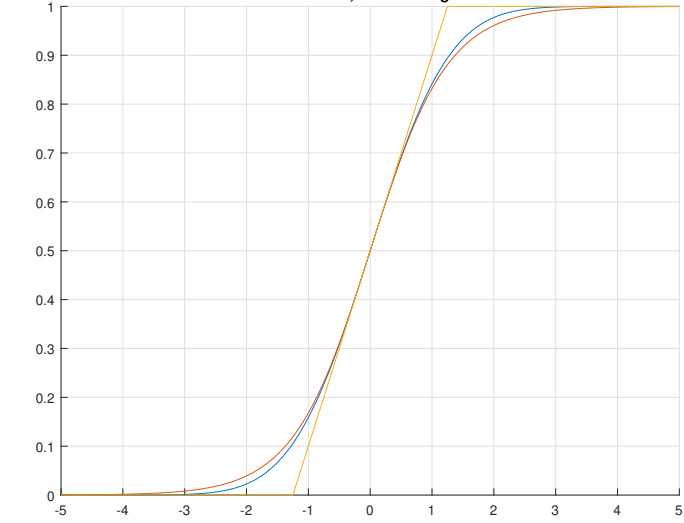
- When is an assumption restrictive? When it restricts the empirical content $\Pr(y_i = 1|x_i)$ of the model. Examples of restrictive assumptions here are linearity in x_i and that $\tilde{\varepsilon}_i$ is normally distributed.
- Unlike assumptions, normalizations do not restrict the empirical content of the model. Given the other assumptions, fixing μ and σ just changes the location and scaling of the parameter vector $\tilde{\beta}$.

Are distributional assumptions likely to matter?

- When there are only differences in scale, then they cancel out when we look at implied probabilities, as we could see in the probit example above.
- Ignoring differences in shapes for the moment, recall that variance of logistic distribution is $\pi^2/3 \approx 3.2899$, hence one could argue that probit coefficients are comparable to logit coefficients when multiplied by $\sqrt{3.2899} \approx 1.8138$.
- But this ignores differences in the shape of the distribution. Amemiya (1981) suggests to multiply probit coefficients by 1.6 to compare them with logit ones. The underlying idea is to match choice probabilities near the center of the distribution. Following the same idea, he suggests to multiply coefficients from the linear probability model by 4 to compare them to logit ones.

Are distributional assumptions likely to matter?

- The figure on the next slide shows rescaled and re-centered versions of the three c.d.f.'s.
- We can see that for probabilities between 0.2 and 0.8 differences in shape are 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.


Cumulative distribution function of the normal, rescaled logistic and rescaled uniform distribution
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

Reporting of results: Marginal effects

- Model is $y_i = 1\{x_i\beta \geq \varepsilon_i\}$ so that a change in x_i either has no effect on y_i or changes it from 0 to 1 (provided that β is positive, otherwise it's the opposite). But we don't observe this, so we don't know. Hence, it is not meaningful to think about this.
- A more meaningful measure is how the probability to observe $y_i = 1$ depends on x_i. For discrete x_i, this is the difference in the probabilities. For continuously distributed x_i, this is the marginal effect.

Reporting of results: Marginal effects

- We have $\Pr(y_i = 1|x_i) = F_{\varepsilon_i}(x_i\beta)$ so that marginal effects for small changes in continously distiributed x_i are given by

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

  where f_{ε_i} is the probability density function (p.d.f.) of ε_i.
- Observe that ratios of marginal effects are equal to ratios of coefficients.
- Commonly, either average marginal effects (differences in probabilities) are reported, or marginal effects (differences in probabilities) at sample average x_i. These are not the same because $\Pr(y_i = 1|x_i)$ is not linear in x_i.

Peculiarities of the logit model

- Have

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

- Marginal effects are given by

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

Peculiarities of the Linear Probability Model

- Recall $y_i = 1\{x_i\beta \geq \varepsilon_i\}$. In the linear probability model, we have that ε_i is uniformly distributed: $F_{\varepsilon_i}(e) = e$.
- Therefore, $\Pr(y_i = 1|x_i) = F_{\varepsilon_i}(x_i\beta) = x_i\beta$.
- Observe that marginal effects are equal to coefficients.
- Also, β can be estimated by ordinary least squares (OLS) with estimation equation

$$y_i = x_i\beta + e_i.$$

- Reason: $\mathbb{E}[y_i|x_i] = x_i\beta$. This is all we need for OLS to be consistent/unbiased.
- Note that e_i is not the same as ε_i.

Peculiarities of the Linear Probability Model

- The error term e_i is heteroskedastic.
- Reason: y_i is a Bernoulli random variable (a random variable that takes on 2 values).
- $\Pr(y_i = 1|x_i) = x_i\beta$ and a Bernoulli random variable has variance $\Pr(y_i = 1|x_i) \cdot (1 - \Pr(y_i = 1|x_i)))$. Therefore,

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

  and this depends on x_i.
- Therefore, need to use heteroskedasticity consistent standard errors; can do generalized least squares instead (theoretically efficient).

Additive random utility foundation

- A model of the form $y_i = 1\{x_i\beta \geq \varepsilon_i\}$ is implied by a richer model that involves utility comparisons.
- i receives utility

$$u_{0i} = z_{i0}\alpha_0 + w_i\gamma_0 + \varepsilon_{i0}$$

when choosing $y_i = 0$ and utility

$$u_{1i} = z_{i1}\alpha_1 + w_i\gamma_1 + \varepsilon_{i1}$$

when choosing $y_i = 1$.

Additive random utility foundation

- Here, there are several types of explanatory variables:
  - z_{i0} are characteristics of i's situation when $y_i = 0$
  - z_{i1} are characteristics of the situation when $y_i = 1$
  - So, e.g., when z_{i0} and z_{i0} indicate how much money i has in her pocket, respectively, then $z_{i0} - z_{i1}$ is the cost of choosing $y_i = 1$.
  - w_i are characteristics that are alternative invariant, e.g. i's years of education.

Additive random utility foundation

- Moreover, there are alternative specific coefficients:
  - α_0 is the value i puts on characteristics of her situation if $y_i = 0$ and α_1 is the value i puts on these characteristics if $y_i = 1$
  - Usually, we impose that decision makers only value characteristics of the alternatives irrespective of which alternative they choose, i.e. $\alpha_0 = \alpha_1 = \alpha$
  - γ_0 and γ_1 indicate how much utility i derives from alternative invariant characteristics when he chooses $y_i = 0$ or $y_i = 1$, respectively. E.g., a more educated individual may have a stronger preference for going to the opera.

Additive random utility foundation

- i chooses $y_i = 1$ if this maximizes his utility. Then,

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

- Can be rewritten as

$$y_i = 1\{\underbrace{z_{i1}\alpha_1 - z_{i0}\alpha_0 + w_i(\gamma_1 - \gamma_0)}_{x_i\beta} \geq \underbrace{-(\varepsilon_{i1} - \varepsilon_{i0})}_{\varepsilon_i}\}.$$

where

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

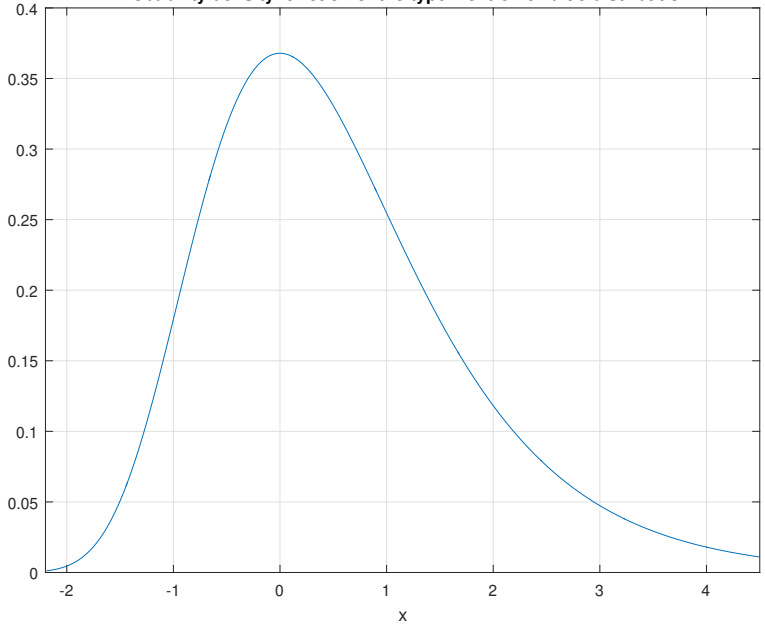
Additive random utility foundation

- Here we have two error terms and an assumption on the distribution of ε_{1i} and ε_{0i} implies how ε_i is distributed:
  - assuming that they are both normally distributed implies that their difference is normally distributed; if they are independently distributed and their variance is 0.5, respectively, then the difference will have variance 1
  - assuming that they are both type 1 extreme value distributed (with mean equal to Euler's constant $0.5772\ldots$, respectively) implies that their difference follows the logistic distribution (with mean zero). Type 1 extreme value distribution has c.d.f.

$$F_{\varepsilon_{ij}}(e) = \exp(-\exp(-e))$$

and density

$$f_{\varepsilon_{ij}}(e) = \exp(-e) \cdot \exp(-\exp(-e)).$$


Probability density function of the type 1 extreme value distribution
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-2
-1
0
1
2
3
4
x

Structural Economic Model

- So far, we have considered 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. For instance, $g(x_i; \beta)$ could be a difference in discounted sums of utilities.

Nonparametric Estimation

- In order to relax distributional assumptions estimate $\Pr(y_i = 1|x_i) = \mathbb{E}[y_i|x_i]$ by performing a nonparametric regression of y_i on x_i.
- Use e.g. frequency estimator when x_i is discrete.
- Or Nadaraya-Watson Kernel regression when x_i is continuously distributed (see, e.g. Pagan and Ullah, 1999, p. 84ff):

$$\tilde{m}(x) = \sum_{i=1}^{N} \underbrace{\left(\frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{x_j - x}{h}\right)} \right)}_{\text{weight for observation } i} \cdot y_i,$$

where $K(\cdot)$ is a so-called Kernel (a weighting function putting most weight on values close to zero) and h is the so-called bandwidth.

Nonparametric Estimation

- Curse of dimensionality (think of the discrete case with 10 values for each covariate and 1000 observations; then $1000/10 = 100$ observations for each value of the covariate when there is one covariate, and $1000/10^2 = 10$ when there are two and they are independent, $1000/10^3 = 1$ for 3, and so on).
- 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 $x_i\beta$.

Manski's (1975) semiparametric Maximum Score estimator

- Key assumptions: $y_i = 1\{x_i\beta \geq \varepsilon_i\}$ and median of ε_i given x_i is equal to zero. Allows for heteroskedasticity and does not require full independence.
- For a candidate parameter vector β
  - predict y_i to be 1 if $x_i\beta \geq 0$ and 0 otherwise
  - add 1 to the objective function if the prediction is correct, and -1 if not.
- Maximize the objective function subject to a scale normalization such as $\beta'\beta = 1$.
- To obtain predicted probabilities or marginal effects perform a nonparametric regression of y_i on $x_i\hat{\beta}$.
- Manski (1975) shows consistency, but not asymptotic normality. Manski and Thompson (1986) show that the estimator does not converge at the parametric $\sqrt{N}$ rate. Horowitz (1992) suggests a smoothed version that is asymptotically normally distributed and converges almost at the parametric rate.

Klein and Spady's (1993) semiparametric maximum likelihood estimator

- Assumes that x_i and ε_i are independent.
- For any candidate parameter vector perform a nonparametric regression of y_i on $x_i\beta$.
- This provides an estimate of F_{ε_i} because $\mathbb{E}[y_i|x_i] = F_{\varepsilon_i}(x_i\beta)$. Use this to construct the likelihood function, which is then maximized. This means that within each iteration F_{ε_i} is estimated anew.
- Alternative: Perform semiparametric least squares (Ichimura, 1993). For this, perform nonparametric regression of y_i on $x_i\beta$, i.e. minimize the variance of the residual. 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 "semi-nonparametric" estimator. Scale normalization: one coefficient equal to -1.

Table III. Estimation results, Switzerland (N = 873)

	Klein–Spady[a]		Smoothed maximum score[a]		Probit[a]		SNP[b]	
Variable	Coeff.	Std.err	Coeff.	Std.err	Coeff.	Std.err	Coeff.	Std.err
Intercept	—	—	5·83	(1·78)	5·62	(1·35)	5·99	(2·20)
AGE	2·98	(0·90)	2·84	(0·98)	3·11	(0·77)	3·23	(0·87)
AGESQ	−0·44	(0·12)	−0·40	(0·13)	0·44	(0·10)	−0·47	(0·12)
EDUC	0·02	(0·03)	0·03	(0·05)	0·03	(0·03)	0·02	(0·03)
NYC	−1·32	(0·33)	−0·80	(0·43)	−1·07	(0·26)	−1·26	(0·24)
NOC	−0·25	(0·11)	−0·16	(0·20)	−0·22	(0·09)	−0·26	(0·10)
NLINC	−1·0	—	−1·0	—	−1·0	—	−1·0	
FOREIGN	1·06	(0·32)	0·91	(0·57)	1·07	(0·29)	0·98	(0·26)
Bandwidth	0·40[c]		0·70					

[a] Results based on scale normalization $b_{NLINC} = -1$.

[b] Obtained by dividing the respective coefficients by the absolute value of the coefficient of NLINC. Standard errors computed by Delta method.

[c] Multiplied by the standard deviation of the index xb_{ks}.

Random coefficients

- Suppose coefficient vector is individual-specific (say drawn from a normal distribution with mean μ_β and variance-covariance matrix Σ_β) and $y_i = 1\{x_i\beta_i \geq \varepsilon_i\}$.
- Then, likelihood contribution for observations with $y_i = 1$ is (for $y_i = 0$ analogously)

$$f(1|x_i; \mu_\beta, \sigma_\beta) = \int F_{\varepsilon_i}(x_i\beta_i)\, dF_{\beta_i}(\beta_i; \mu_\beta, \Sigma_\beta).$$

- This can be simulated using draws from the distribution of β_i for given parameters μ_β, Σ_β.

Binary choice

Generalization: ordered choice

Overview

Ordered choice

- There are J possible outcomes $y_i = 1, \ldots, J$ that have no cardinal meaning, but a natural ordering (e.g., number of children, level of schooling, health).
- There is a latent (i.e. unobserved) variable $y_i^* = x_i\beta + \varepsilon_i$.
- We observe the outcome

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & \\ J & \text{if } \alpha_{J-1} < y_i^*. \end{cases}$$

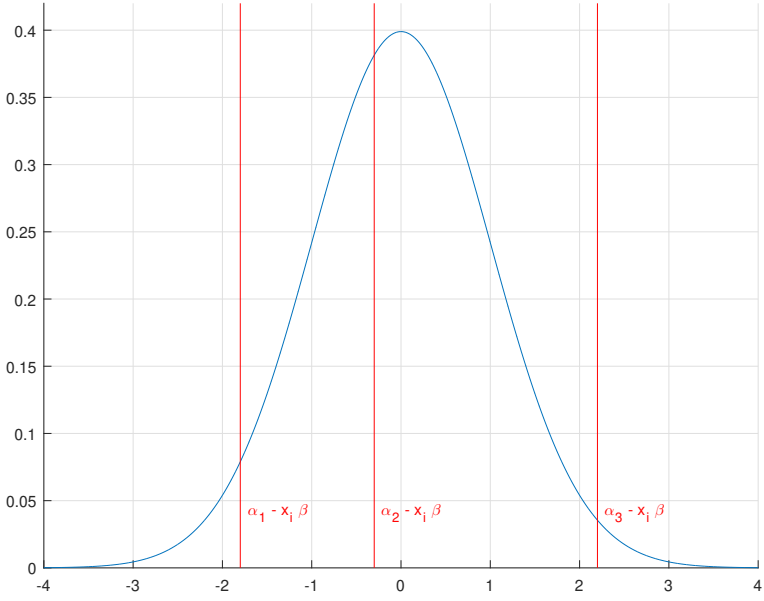
- We will treat the thresholds as parameters. (If they are observed, we have instead an interval regression model.)

Likelihood contributions

- Model implies

$$\begin{aligned}
\Pr(y_i = 1|x_i) &= \Pr(\varepsilon_i \leq \alpha_1 - x_i\beta) = F_{\varepsilon_i}(\alpha_1 - x_i\beta) \\
\Pr(y_i = 2|x_i) &= \Pr(\alpha_1 < x_i\beta + \varepsilon_i \leq \alpha_2) = F_{\varepsilon_i}(\alpha_2 - x_i\beta) - F_{\varepsilon_i}(\alpha_1 - x_i\beta) \\
&\vdots \\
\Pr(y_i = J|x_i) &= \Pr(\alpha_{J-1} < x_i\beta + \varepsilon_i) = 1 - F_{\varepsilon_i}(\alpha_{J-1} - x_i\beta).
\end{aligned}$$

- Usual distributional assumptions for ε_i:
  - normally distributed: ordered probit model
  - logistic: ordered logit model (next slide).


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
α_1 - x_i β
α_2 - x_i β
α_3 - x_i β

Normalizations

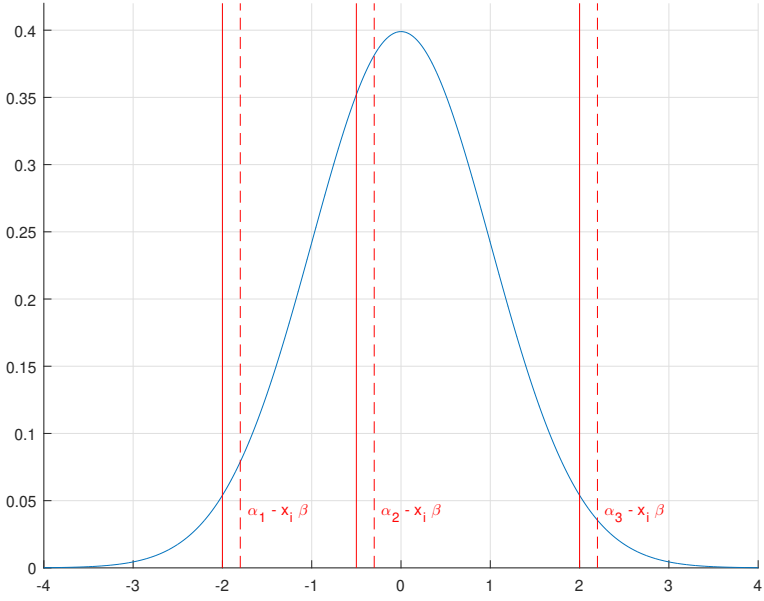
- Model is only identified up to location and scale.
- Usually impose the following normalizations:
  - x_i does not include a constant term (in the binary choice model we have instead imposed $\alpha_1 = 0$)
  - fix the variance of ε_i at a particular value; like in the binary choice model this is done by making a distributional assumption.

Reporting of results

- Marginal effects of change in x_{ij} are

$$\frac{\partial \Pr(y_i = j|x_i)}{\partial x_{ik}} = \begin{cases} -f_{\varepsilon_i}(\alpha_1 - x_i\beta) \cdot \beta_k & \text{if } j = 1 \\ -\left(f_{\varepsilon_i}(\alpha_j - x_i\beta) - f_{\varepsilon_i}(\alpha_{j-1} - x_i\beta)\right)\beta_k & \text{if } j = 2, \ldots, J-1 \\ f_{\varepsilon_i}(\alpha_{J-1} - x_i\beta) \cdot \beta_k & \text{if } j = J. \end{cases}$$

- The sign of the marginal effect for a given j is in general not equal to the sign of β_k.
- For $j = 1$ the marginal effect and β_k are always of opposite sign whereas for $j = J$ they are always of the same sign.
- Reason: Choice probabilities sum to 1.
- Therefore focus on coefficients in some cases more meaningful.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4
$\alpha_1 - x_i \beta$
$\alpha_2 - x_i \beta$
$\alpha_3 - x_i \beta$

Multinomial choice

Overview

Multinomial choice

Multinomial logit model

Multinomial choice

- McFadden (1974).
- J alternatives, cannot be ordered. Additional examples:
  - type of health insurance (differ in many dimensions such as coverage and deductibles)
  - brand of a soft drink (e.g. between Coca Cola, Pepsi, and no-name cola).
- Want to relate choice probabilities to
  - alternative varying regressors z_{ij}; these are characteristics of j if chosen by i such as income minus price
  - alternative invariant regressors w_i; these are characteristics of the decision situation, or decision maker i.

Random utility maximization

- i's utility when choosing j is

$$u_{ij} = \underbrace{z_{ij}\alpha + w_i\gamma_j}_{x_{ij}\beta} + \varepsilon_{ij}.$$

  - consists of average utility $\bar{u}_{ij} = x_{ij}\beta$ (can be generalized to be non-linear in x_{ij}) and taste shock ε_{ij}
  - usual assumption: coefficient on z_{ij} is the same across alternatives; example: the utility of traveling by bus depends in the same way on travel time as the utility of traveling by car.
- Utility maximization (ignoring ties):

$$y_i = j \text{ if } u_{ij} \geq u_{ij'} \text{ for all } j'.$$

- Objects of interest: β and marginal effects (related to e.g. price elasticities).

Random utility maximization

- Probability to choose $j = 1$ depends on all x_{ij} and all ε_{ij} and is

$$\begin{aligned}
&\Pr(y_i = 1 | x_{i1}, \ldots, x_{iJ}) \\
&= \Pr(u_{i1} \geq u_{i2}, u_{i1} \geq u_{i3}, \ldots, u_{i1} \geq u_{iJ}) \\
&= \Pr\left(\varepsilon_{i2} \leq (x_{i1} - x_{i2})\beta + \varepsilon_{i1}, \varepsilon_{i3} \leq (x_{i1} - x_{i3})\beta + \varepsilon_{i1}, \ldots, \right. \\
&\qquad \left. \varepsilon_{iJ} \leq (x_{i1} - x_{iJ})\beta + \varepsilon_{i1}\right) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{(x_{i\mathbf{1}} - x_{i\mathbf{2}})\beta + \varepsilon_{i\mathbf{1}}} \int_{-\infty}^{(x_{i\mathbf{1}} - x_{i\mathbf{3}})\beta + \varepsilon_{i\mathbf{1}}} \cdots \int_{-\infty}^{(x_{i\mathbf{1}} - x_{iJ})\beta + \varepsilon_{i\mathbf{1}}} \\
&\qquad f_{\varepsilon_{i\mathbf{1}}, \varepsilon_{i\mathbf{2}}, \varepsilon_{i\mathbf{3}}, \ldots, \varepsilon_{iJ}}(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{iJ})\, d\varepsilon_{iJ} \cdots d\varepsilon_{i3}\, d\varepsilon_{i2}\, d\varepsilon_{i1},
\end{aligned}$$

where $f_{\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \ldots, \varepsilon_{iJ}}$ is the joint density of the taste shocks.

Choice probability in the logit model

- The multinomial logit choice probabilities are

$$\Pr(y_i = j | x_{i1}, \ldots, x_{iJ}) = \frac{\exp(x_{ij}\beta)}{\sum_{j'} \exp(x_{ij'}\beta)},$$

  where the summation in the denominator is over alternatives j' in the relevant choice set.

- For our specification of the utility function, $\bar{u}_{ij} = x_{ij}\beta = z_{ij}\alpha + w_i\gamma_j$ we thus have

$$\Pr(y_i = j | z_{i1}, \ldots, z_{iJ}, w_i) = \frac{\exp(z_{ij}\alpha + w_i\gamma_j)}{\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})}.$$

Normalizations in the logit model

- Choices are informative about utility differences. Observe that

$$\frac{\exp(z_{ij}\alpha + w_i\gamma_j)}{\sum_{j'}\exp(z_{ij'}\alpha + w_i\gamma_{j'})} = \frac{\exp\left((z_{ij} - z_{i1})\alpha + w_i(\gamma_j - \gamma_1)\right)}{\sum_{j'}\exp\left((z_{ij'} - z_{i1})\alpha + w_i(\gamma_{j'} - \gamma_1)\right)}$$

  (we get this by dividing both the numerator and the denominator of the left hand side by $\exp(z_{i1}\alpha + w_i\gamma_1)$).

- Interpretation: difference to utility of first alternative, $j = 1$ acts as base alternative. Usual normalization is to set utility of one alternative to 0, e.g. $\gamma_1 = 0$.

- z_{ij} cannot have a constant term it it (first element equal to 1). But we can include alternative-specific constants using alternative dummies (for $j \neq 1$).

Marginal effects

- Define

$$p_{ij} \equiv \Pr(y_i = j | x_{i1}, \ldots, x_{iJ}) = \frac{\exp(z_{ij}\alpha + w_i\gamma_j)}{\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})}.$$

- We have

$$\frac{\partial p_{ij}}{\partial z_{ij}} = \frac{\exp(z_{ij}\alpha + w_i\gamma_j) \cdot \alpha \cdot \left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)}{\left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)^2} - \frac{\exp(z_{ij}\alpha + w_i\gamma_j) \cdot \exp(z_{ij}\alpha + w_i\gamma_j) \cdot \alpha}{\left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)^2},$$

  which is equal to

$$p_{ij}\alpha - p_{ij}^2\alpha = p_{ij}(1 - p_{ij})\alpha.$$

Marginal effects

- Recall

$$p_{ij} = \frac{\exp(z_{ij}\alpha + w_i\gamma_j)}{\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})}.$$

- For the derivative with respect to z_{ik}, $k \neq j$, we get

$$\frac{\partial p_{ij}}{\partial z_{ik}} = -\frac{\exp(z_{ij}\alpha + w_i\gamma_j) \cdot \exp(z_{ik}\alpha + w_i\gamma_k) \cdot \alpha}{\left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma)\right)^2} = -p_{ij}p_{ik}\alpha.$$

Marginal effects

- Starting again from

$$p_{ij} = \frac{\exp(z_{ij}\alpha + w_i\gamma_j)}{\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})}$$

we get

$$\frac{\partial p_{ij}}{\partial w_i} = \frac{\exp(z_{ij}\alpha + w_i\gamma_j) \cdot \gamma_j \cdot \left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)}{\left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)^2}$$

$$- \frac{\exp(z_{ij}\alpha + w_i\gamma_j) \cdot \left(\sum_{j''} \exp(z_{ij''}\alpha + w_i\gamma_{j''})\gamma_{j''}\right)}{\left(\sum_{j'} \exp(z_{ij'}\alpha + w_i\gamma_{j'})\right)^2}$$

$$= p_{ij}\gamma_j - p_{ij} \sum_{j''} p_{ij''}\gamma_{j''} = p_{ij} \left(\gamma_j - \sum_{j''} p_{ij''}\gamma_{j''}\right).$$

Marginal effects

- So, to summarize, 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

$$\frac{\partial p_{ij}}{\partial w_i} = p_{ij}\left(\gamma_j - \sum_{j''}^{J} p_{ij''}\gamma_{j''}\right).$$

- Sign of the effect of z_{ij} on p_{ij} is same as sign of the coefficient, sign of the effect of z_{ik} on p_{ij} is the opposite of sign of the coefficient.
- Sign of effect of w_i on p_{ij} is the sign of the coeficient relative to the average coefficient.

Welfare analysis in general

- Welfare effects are closely related to (expected) maximal utility.
- They arise from introducing new alternatives or changing characteristics of existing alternatives.
- Question: What is the transfer that is needed to make i as well off with the new alternative or the changed characteristics, as compared to before? This amount is the so-called "compensating variation" (CV) and is measured in terms of money. One is usually interested in its expectation, $\mathbb{E}[CV]$.

Expected maximal utility in the logit model

Consider J independent type 1 extreme value variables $u_{ij} \equiv \bar{u}_{ij} + \varepsilon_{ij}$. Then,

$$
\begin{aligned}
\Pr\left(\max_j u_{ij} \leq u\right) &= \Pr\left(u_{i1} \leq u, \ldots, u_{iJ} \leq u\right) = \prod_j \Pr\left(u_{ij} \leq u\right) = \prod_j \Pr\left(\varepsilon_{ij} \leq u - \bar{u}_{ij}\right) \\
&= \prod_j \exp\left(-\exp(-u + \bar{u}_{ij})\right) = \exp\left(-\sum_j \exp(-u + \bar{u}_{ij})\right) \\
&= \exp\left(-\sum_j \exp(-u) \cdot \exp(\bar{u}_{ij})\right) = \exp\left(-\exp(-u) \cdot \sum_j \exp(\bar{u}_{ij})\right) \\
&= \exp\left(-\exp(-u) \cdot \exp\left(\log\left(\sum_j \exp(\bar{u}_{ij})\right)\right)\right) \\
&= \exp\left(-\exp\left\{-u + \log\left(\sum_j \exp(\bar{u}_{ij})\right)\right\}\right).
\end{aligned}
$$

Expected maximal utility in the logit model

- We have just shown that $\max_j u_{ij}$ has c.d.f.

$$\exp\left(-\exp\left\{-u+\log\left(\sum_j \exp(\bar{u}_{ij})\right)\right\}\right).$$

- This means that $\max_j u_{ij}$ is a type 1 extreme value random variable with mean equal to $0.5772\ldots$ plus

$$I \equiv \log\left(\sum_j \exp\left(\bar{u}_{ij}\right)\right).$$

  I is the expected maximal utility, or "inclusive value".
- If α is the price coefficient (a negative number), I is the initial inclusive value and I' is the changed one, then $-1/\alpha$ is the value of one util in monetary terms and we have

$$\mathbb{E}[CV] = -\left(-\frac{1}{\alpha}\right)\cdot\left(I'-I\right).$$

Limitations

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 [Independence of irrelevant alternatives], 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.

McFadden (1973, p. 113)

Hausman-McFadden (1984) test

- Multinomial logit model has necessary and sufficient characterization given by the independence of irrelevant alternatives (IIA) property.
- IIA property means that the ratio of the probabilities of choosing any two alternatives is independent of the attributes or the availability of a third alternative.
- Means that when mean utilities are of the form $\bar{u}_{ij} = x_{ij}\beta$, then we can estimate all but the constant terms from a subset of choices.
- Test: Compare parameter estimates obtained from choice data from the full choice set with estimates obtained from conditional choice data from a restricted choice set.

Multinomial choice

Generalizations

Overview

Nested multinomial logit model

- In this model alternatives are grouped into mutually exclusive nests B_s with $s = 1, \ldots, S$ (here we discuss only the case with one layer of nests). The nesting structure is assumed to be known *a priori*.
- Utility is specified as $u_{ij} = x_{ij}\beta + \rho_s \varepsilon_{ij}$, where ρ_s is a scaling parameter and ε_{ij} is type 1 extreme value.
- Can also write this as a multinomial logit model with mean utilities $x_{ij}\beta/\rho_s$. Therefore, choice probability for choice within the nest is of the multinomial logit form:

$$\Pr(y_i = j | x_{i1}, \ldots, x_{iJ}, y_i \in B_s) = \frac{\exp(x_{ij}\beta/\rho_s)}{\sum_{j' \in B_s} \exp(x_{ij'}\beta/\rho_s)}.$$

  Can estimate $\tilde{\beta} \equiv \beta/\rho_s$ using a conventional multinomial logit model within the nest.

Nested multinomial logit model

- The expected maximal utility within nest B_s is $0.5772\ldots$ plus the so-called inclusive value

$$I_{is} = \log\left(\sum_{j \in B_s} \exp(x_{ij}\beta/\rho_s)\right)$$

—we have seen a similar expression in the context of welfare analysis.

- Probability that nest s is chosen is

$$\Pr(y_i \in B_s | I_{i1}, \ldots, I_{iS}) = \frac{\exp(\rho_s I_{is})}{\sum_{s'=1}^{S} \exp(\rho_{s'} I_{is'})}.$$

- Can add nest specific variables q_{is} that take on the same value for all alternatives in the nest and do not affect the choice between the alternatives in the nest.

Nested multinomial logit model

- ρ_s is the so-called dissimilarity parameter:
  - for $\rho_s = 1$ we obtain the multinomial logit model
  - for $0 < \rho_s < 1$ alternatives in the same nest become closer substitutes.
- ρ_s is identified from choice between nests.
- 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 (McFadden, 1977, 1978; Boersch-Supan, 1990).

Independence of irrelevant alternatives

- Consider following example: 3 alternatives, first forms a trivial nest, second and third form a true nest.
- Have (omitting here and in the following the conditioning on $\bar{u}_{i1}, \bar{u}_{i2}, \bar{u}_{i3}$)

$$\begin{aligned}
\Pr(y_i = 1 | C = \{1,2,3\}) &= \frac{\exp(\bar{u}_{i1})}{\exp(\bar{u}_{i1}) + [\exp(\bar{u}_{i2}/\rho) + \exp(\bar{u}_{i3}/\rho)]^{\rho}} \\
\Pr(y_i = 2 |, C = \{1,2,3\}) &= \frac{\exp(\bar{u}_{i2}/\rho) \cdot [\exp(\bar{u}_{i2}/\rho) + \exp(\bar{u}_{i3}/\rho)]^{\rho-1}}{\exp(\bar{u}_{i1}) + [\exp(\bar{u}_{i2}/\rho) + \exp(\bar{u}_{i3}/\rho)]^{\rho}}.
\end{aligned}$$

- Observe that we get multinomial logit probabilities for $\rho = 1$. In that case, IIA holds because

$$\frac{\Pr(y_i = 1 |, C = \{1,2,3\})}{\Pr(y_i = 2 |, C = \{1,2,3\})} = \frac{\Pr(y_i = 1 |, C = \{1,2\})}{\Pr(y_i = 2 |, C = \{1,2\})} = \frac{\exp(\bar{u}_{i1})}{\exp(\bar{u}_{i2})}.$$

- This is not true anymore for $\rho \neq 1$, so IIA does not hold in that case and the nested logit model is more general in that respect.

Goldberg (1995)

- Estimates nested logit model for demand for cars.
- Then specifies a model for the supply side of this industry and infers marginal cost from the first order conditions of each of the firms. Example with a monopolist with fixed cost f selling quantity q at price p and produces at marginal cost mc:
  - profits are $\pi = (p - mc)q - f$
  - first order condition is $\partial\pi/\partial p = q + (p - mc) \cdot \partial q/\partial p = 0$
  - hence
$$mc = \frac{q}{\partial q/\partial p} + p.$$
- This allows her to simulate the effect of quota restrictions on market shares, prices and quality upgrading; and the effect of changes in the exchange rate on car prices ("pass through").

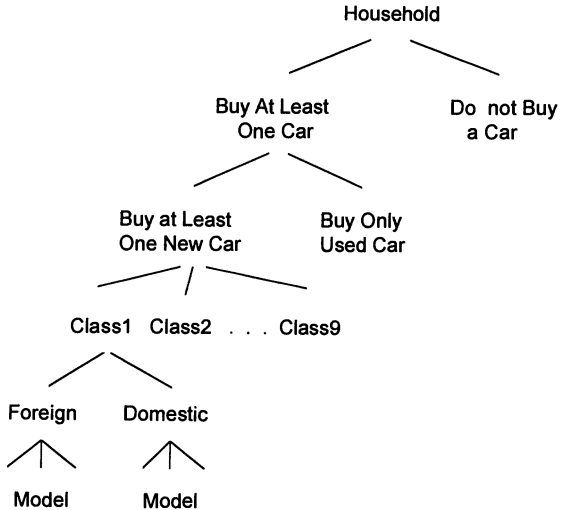


FIGURE 1.—Automobile choice model.

TABLE II

PRICE ELASTICITIES OF DEMAND (AVERAGE BY CLASS)

Class	Origin	Elasticity	Elasticity (first time buyer)	Elasticity (repeat buyer)
Subcompacts	DOM	−3.2857	−3.6245	−2.9816
	FOR	−3.6797	−5.2531	−2.9488
Compacts	DOM	−3.419	−4.8722	−3.1546
	FOR	−4.0319	−5.7229	−3.3733
Intermediate	DOM	−4.1799	−5.3153	−2.8420
	FOR	−5.1524	−6.2232	−4.9274
Standard	DOM	−4.7121	−5.932	−4.3730
Luxury	DOM	−1.9121	−2.5981	−1.1137
	FOR	−2.7448	−3.1272	−1.9959
Sports	DOM	−1.0654	−2.3468	−1.3959
	FOR	−1.5254	3.0211	−1.1429
Pick-ups	DOM	−3.5259	−5.1391	−3.1647
	FOR	−2.6883	−3.9822	−2.1483
Vans	DOM	−4.3633	−5.4977	−3.9790
	FOR	−4.6548	−4.8837	−2.4376
Other	DOM	−4.0884	−4.3185	−3.5694
	FOR	−3.0271	−3.3185	−2.3345

TABLE IV

MARGINAL COSTS AND MARKUPS

Class	Origin	Cost	Price	Markup	(Price − Cost)
1	DOM	3906	6628	0.36	2722
1	FOR	5688	7840	0.27	2152
2	DOM	3213	6391	0.43	3178
2	FOR	5430	6610	0.19	1180
3	DOM	4773	7134	0.33	2361
3	FOR	9300	12781	0.30	3421
4	DOM	4866	8632	0.40	3766
5	DOM	7247	13458	0.46	6301
5	FOR	10379	18499	0.43	8129
6	DOM	3715	10105	0.69	6390
6	FOR	5822	12823	0.56	7001
7	DOM	5101	8229	0.37	3128
7	FOR	2758	5611	0.41	2583
8	DOM	6937	9634	0.30	2697
8	FOR	12691	15291	0.17	2600
9	DOM	8333	10121	0.15	1788
9	FOR	2750	5174	0.44	2424
Model		**Cost**	**Price**	**Markup**	**(Price − Cost)**
Civic		4884	5680	0.14	796
Escort		3068	4565	0.33	1497
Lynx		3069	4325	0.29	1256
Accord		5286	5854	0.10	567
Audi 5000		7353	14165	0.48	6812
Oldsmobile 98		5372	11295	0.52	5923
Jaguar		10768	19091	0.44	8323
Mercedes 300		13188	22662	0.42	9474
Porsche 944		5714	13136	0.56	7422
Ferrari		7679	19698	0.61	12018

Generalized extreme value model

- The denominator in the previously shown choice probabilities, e.g.

$$\Pr(y_i = 1|, C = \{1, 2, 3\}) = \frac{\exp(\bar{u}_{i1})}{\exp(\bar{u}_{i1}) + [\exp(\bar{u}_{i2}/\rho) + \exp(\bar{u}_{i3}/\rho)]^{\rho}},$$

  is a special case of a function

$$G(a_1, a_2, a_3) = a_1 + \left[a_2^{1/\rho} + a_3^{1/\rho}\right]^{\rho},$$

  with $a_j = \exp(\bar{u}_{ij})$, $j = 1, 2, 3$.

- McFadden (1978) provides conditions on G such that

$$\Pr(y_i = j|\bar{u}_{i1}, \ldots, \bar{u}_{iJ}) = \exp(\bar{u}_{ij}) \cdot \frac{\partial G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3})) \Big/ \partial(\exp(\bar{u}_{ij}))}{G(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3}))}$$

  and that the resulting probabilistic choice model is consistent with utility maximization.

Generalized extreme value model

- He shows that

$$\mathbb{E}\left[\max_j \bar{u}_{ij} + \varepsilon_{ij}\right] = \gamma + \log\left(G\left(\exp(\bar{u}_{i1}), \exp(\bar{u}_{i2}), \exp(\bar{u}_{i3})\right)\right),$$

  where, again, $\gamma = 0.5772\ldots$ is Euler's constant, and

$$\Pr(y_i = j | \bar{u}_{i1}, \ldots, \bar{u}_{iJ}) = \frac{\partial \mathbb{E}\left[\max_j \bar{u}_{ij} + \varepsilon_{ij}\right]}{\partial \bar{u}_{ij}}.$$

Multinomial probit model

- Assumes that

$$u_{ij} = x_{ij}\beta + \varepsilon_{ij}.$$

  and that the ε_{ij}'s are jointly normally distributed.
- We can define one base alternative and re-express the problem so that choice is based on $J-1$ utility differences. As a scale normalization we can set one variance parameter to 1.
- Hence we need to estimate $J-2$ variance parameters and $(J-1)\cdot(J-2)/2$ covariance parameters on top of the parameters of the utility function.

Mixed logit model

- Maintain assumption that ε_{ij}'s are type 1 extreme value.
- Use a random coefficient specification

$$\Pr(y_i = j | x_{ij}, \beta_i) = \frac{\exp(x_{ij}\beta_i)}{\sum_{j'} \exp(x_{ij'}\beta_i)}.$$

- The coefficients are allowed to be correlated with one another, but (typically) not with x_i.
- Expectation thereof can be simulated and enters likelihood function.
- Called "mixed" because it averages over different multinomial logit models, one for each draw of the random coefficients.
- Any choice model that is derived from random utility maximization has choice probabilities that can in principle be approximated as closely as one pleases by this model (McFadden and Train, 2000).

Estimation of the mixed logit model

- Suppose β_i has only one normally distributed element. We want to estimate the mean μ_β and the variance σ_β^2 of this distribution.
- Take a set of, say, 100 standard normal draws d_{is} for each individual. Do this before you evaluate the likelihood function.
- Program the likelihood function. Inputs are the data (choices y_i and characteristics z_{ij}), the parameters you want to estimate (μ_β, σ_β^2), and the standard normal draws.
- Inside the likelihood function
  - for each individual i and each s turn standard normal draw d_{is} into normal draws with desired mean and variance using $\beta_{is} = \mu_\beta + \sigma_\beta \cdot d_{is}$
  - calculate probabiliy p_{is} that i chooses y_i given the z_{ij}
  - do this for all 100 values of s and then average over resulting p_{is} to get the likelihood contribution for i
  - only then calculate log likelihood contribution for i by taking log.

Epilogue

Exam preparation

1. Start with the slides, make sure that you fully understand every single one of them.
2. Then go to the lecture notes and look for the derivations that you were pointed to in class. Also read the first two sections.
3. After that, move on to the problem sets. Everything on them is fully relevant for the exam.
4. Digest and reflect.

Epilogue

- Most of economics is empirical, not econometric theory and not economic theory. There is also not much applied theory without any empirical application.
- Empirical work should, in my view, be closely related to theory. Theory gives guidance for carrying out the empirical analysis, and it helps us to interpret the obtained results. Think of McFadden (1974).
- Whether structural models or reduced-form analysis should be used depends on the purpose of the analysis and data availability. See Marshak, J. (1953): "Economic measurements for policy and prediction," in Studies in Econometric Method, ed. by W. Hood, and T. C. Koopmans, pp. 1–26. Wiley, New York, available at http://cowles.econ.yale.edu/P/cm/m14/m14-01.pdf.