Summer Program in Quantitative Methods of Social Research
Advanced Topics in Maximum Likelihood
Categorical Choice Models and Models for Qualitative Dependent Variables with Repeated Observations

Gregory Wawro

Associate Professor

Department of Political Science

Columbia University

420 W. 118th St.

New York, NY 10027

phone: (212) 854-8540

fax: (212) 222-0598

email: gjw10@columbia.edu

ACKNOWLEDGEMENTS

This course draws on material from the following works.

Alvarez, R. Michael and Jonathan Nagler. 1995. "Economics, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election." *American Journal of Political Science* 39:714-744

Amemiya, Takeshi. 1985 *Advanced econometrics.* Cambridge: Harvard University Press.

Baltagi, Badi H. 1995. *Econometric Analysis of Panel Data.* New York: John Wiley & Sons.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. "Taking Time Seriously: Time-SeriesCross-Section Analysis with a Binary Dependent Variable." *American Journal of Political Science* 42:12601288.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics.* New York: Oxford University Press.

Greene, William H. 1997. *Econometric Analysis*, 3rd ed. Upper Saddle River, N.J.: Prentice Hall

Hsiao, Cheng. 1986. *Analysis of Panel Data.* Cambridge: Cambridge University Press.

King, Gary. 1989. *Unifying Political Methodology.* New York: Cambridge University Press

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks: Sage Publications.

Maddala, G.S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics.* New York: Cambridge University Press.

Zorn, Christopher J. W. 2001. "Generalized Estimating Equations Models for Correlated Data: A Review With Applications." *American Journal of Political Science* 45: 470–90.

Zorn, Christopher J. W. Notes for Advanced Maximum Likelihood, ICPSR Summer Program in Quantitative Methods (http://www.polisci.emory.edu/zorn/Classes/ICPSR2003/index.html).

TABLE OF CONTENTS

Part I

Probabilistic Choice Models

Section 1

Introduction

- The first class of categorical choice models we'll focus on are based on the logistic distribution.
- To fix ideas, consider the binary logit model.

$$y_i^* = \boldsymbol{\beta}'\mathbf{x}_i + u_i$$

where y_i^* is a scalar, $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$, $\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix}$, and $u_i \sim$ logistic.

$$\boldsymbol{\beta}'\mathbf{x}_i = [\beta_1 \cdots \beta_k] \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

$$y_i^* > 0 \Rightarrow \boldsymbol{\beta}'\mathbf{x}_i + u_i > 0 \Rightarrow u_i > -\boldsymbol{\beta}'\mathbf{x}_i$$

$$\begin{aligned} \Pr(y_i^* > 0|\mathbf{x}_i) = \Pr(y_i = 1|\mathbf{x}_i) &= \Pr(u_i > -\boldsymbol{\beta}'\mathbf{x}_i) \\ &= \Pr\left(u_i \leq \boldsymbol{\beta}'\mathbf{x}_i\right) \end{aligned}$$

- Assume u_i follows a standard logistic distribution w/ mean 0 and variance $\pi^2/3$.
- Then
$$\Pr(y_i = 1) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}$$
- The likelihood function for the logit model then is
$$L = \Pi_{i=1}^{n} \left[\frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}\right]^{y_i} \left[\frac{1}{1 + \exp(\boldsymbol{\beta}'\mathbf{x})}\right]^{1-y_i}$$
- We also use models based on the logistic distribution when we have dependent variables that are categorical and the categories are unordered.
- For example:
  - Transportation choices: bus, car, train.
  - Voting choices in elections with more than two candidates.
  - Occupational categories: craft, blue collar, white collar, professional.

Section 2

The Multinomial Logit Model

- The set-up:
  - n decision makers indexed by i
  - J alternatives: $1, 2, \ldots, J$.
  - We observe $y_i = j$ if i chooses alternative j.
  - We observe a vector of characteristics $\mathbf{z}_i$ for each i.
  - The object of interest is $P_{ij} = \Pr(y_i = j|\mathbf{z}_i)$.
- We know $P_{ij}\epsilon[0, 1]$, so we want P_{ij} to be proportional to some function that is nonnegative:

$$P_{ij} \propto \exp(\tilde{\boldsymbol{\gamma}}'_j \mathbf{z}_i)$$

- The probabilities across choices must sum to one so

$$\sum_{j=1}^{J} P_{ij} = 1 \Rightarrow P_{ij} = \frac{\exp(\tilde{\boldsymbol{\gamma}}'_j \mathbf{z}_i)}{\sum_{k=1}^{J} \exp(\tilde{\boldsymbol{\gamma}}'_k \mathbf{z}_i)}$$

2.1 Identification Issue

- Unknown parameters in model are $\tilde{\boldsymbol{\gamma}}_1, \tilde{\boldsymbol{\gamma}}_2, \ldots, \tilde{\boldsymbol{\gamma}}_J$, but can't estimate all of these parameters.
- With J alternatives, if know the choice probabilities for $J-1$ of these alternatives, then know the probability of choosing the remaining alternative.
- Don't have J independent pieces of information regarding the choice probabilities.
- $\Rightarrow$ there can't be J separate parameter vectors to estimate—only have $J-1$ identifiable parameter vectors.

2.2 Normalization

- We need some parameter restrictions to knock out one of the probabilities as an unknown piece of information.
- One way to do this: set $\tilde{\boldsymbol{\gamma}}_1 = \mathbf{0}$.
- What we're really doing is something like this:

$$\begin{aligned} P_{ij} &= \frac{\exp(\tilde{\boldsymbol{\gamma}}'_j \mathbf{z}_i) / \exp(\tilde{\boldsymbol{\gamma}}'_1 \mathbf{z}_i)}{\sum_{k=1}^{J} \exp(\tilde{\boldsymbol{\gamma}}'_k \mathbf{z}_i) / \exp(\tilde{\boldsymbol{\gamma}}'_1 \mathbf{z}_i)} \\ &= \frac{\exp\left([\tilde{\boldsymbol{\gamma}}'_j - \tilde{\boldsymbol{\gamma}}'_1]\mathbf{z}_i\right)}{[\exp(\tilde{\boldsymbol{\gamma}}'_1 \mathbf{z}_i) + \exp(\tilde{\boldsymbol{\gamma}}'_2 \mathbf{z}_i) + \ldots + \exp(\tilde{\boldsymbol{\gamma}}'_J \mathbf{z}_i)] / \exp(\tilde{\boldsymbol{\gamma}}'_1 \mathbf{z}_i)} \\ &= \frac{\exp\left([\tilde{\boldsymbol{\gamma}}'_j - \tilde{\boldsymbol{\gamma}}'_1]\mathbf{z}_i\right)}{1 + \sum_{k=2}^{J} \exp\left([\tilde{\boldsymbol{\gamma}}'_k - \tilde{\boldsymbol{\gamma}}'_1]\mathbf{z}_i\right)} \\ &= \frac{\exp\left(\boldsymbol{\gamma}'_j \mathbf{z}_i\right)}{1 + \sum_{k=2}^{J} \exp\left(\boldsymbol{\gamma}'_k \mathbf{z}_i\right)} \end{aligned}$$

  where $\boldsymbol{\gamma}'_j = \tilde{\boldsymbol{\gamma}}'_j - \tilde{\boldsymbol{\gamma}}'_1$ for $j = 2, 3, \ldots, J$.

- To write down the likelihood function, first define a set of dummy variables, d_{ij}, where

$$d_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases}$$

- The likelihood function for an individual then is

$$\begin{aligned} L_i &= P_{i1}^{d_{i1}} P_{i2}^{d_{i2}} \cdots P_{iJ}^{d_{iJ}} \\ &= \prod_{j=1}^{J} P_{ij}^{d_{ij}} \end{aligned}$$

- The log likelihood for the sample is

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \ln P_{ij}$$

- Interpretation of the coefficients in this model is tricky:

$$\frac{\partial P_j}{\partial \mathbf{z}_i} = P_j \left[\boldsymbol{\gamma}_j - \sum_{k=1}^{J} P_k \boldsymbol{\gamma}_k \right]$$

- Another way to interpret effects is to compute log-odds ratios:

$$\begin{aligned} \frac{P_j}{P_k} &= \frac{\exp(\boldsymbol{\gamma}_j' \mathbf{z}_i) / \left(1 + \sum_{k=2}^{J} \exp(\boldsymbol{\gamma}_k' \mathbf{z}_i)\right)}{\exp(\boldsymbol{\gamma}_k' \mathbf{z}_i) / \left(1 + \sum_{k=2}^{J} \exp(\boldsymbol{\gamma}_k' \mathbf{z}_i)\right)} \\ &= \frac{\exp(\boldsymbol{\gamma}_j' \mathbf{z}_i)}{\exp(\boldsymbol{\gamma}_k' \mathbf{z}_i)} = \exp\left([\boldsymbol{\gamma}_j' - \boldsymbol{\gamma}_k'] \mathbf{z}_i\right) \end{aligned}$$

- Taking logs gives
$$\ln\left(\frac{P_j}{P_k}\right) = [\boldsymbol{\gamma}'_j - \boldsymbol{\gamma}'_k]\mathbf{z}_i$$
- Interpretation: we are taking a base case and comparing the probability of making other choices relative to the base case.
- Consider category 1, where $\boldsymbol{\gamma}_1 = \mathbf{0}$:
$$\ln\left(\frac{P_j}{P_1}\right) = \boldsymbol{\gamma}'_j\mathbf{z}_i$$
- So $\boldsymbol{\gamma}_j$ gives you the change in the log odds of being in category j instead of category 1.

2.3 Application of MNL

- Canonical data set for MNL is occupational attainment data from the General Social Survey (see Table 6.1 on p. 152 in Long):
  - The dependent variable is occupation (OCC=1 for menial; =2 for blue collar; =3 for craft; =4 for white collar; =5 for professional).
  - The explanatory variables are possible years of work experience (EXP=age-years of education - 5), number of years of formal education (ED) and race ($WHITE$=1 for white; =0 otherwise).
- We can estimate the model using `Stata`'s `mlogit` command:

```
mlogit occ exper ed white, basecategory(1);
```

```
Multinomial regression                            Number of obs   =        337
                                                  LR chi2(12)     =     166.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -426.80048                       Pseudo R2       =     0.1629

------------------------------------------------------------------------------
         occ |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
       exper |   .0047212   .0173984     0.27   0.786    -.0293789    .0388214
          ed |  -.0994247   .1022812    -0.97   0.331    -.2998922    .1010428
       white |   1.236504   .7244352     1.71   0.088    -.1833631    2.656371
       _cons |   .7412336    1.51954     0.49   0.626     -2.23701    3.719477
-------------+----------------------------------------------------------------
3            |
       exper |   .0276838   .0166737     1.66   0.097     -.004996    .0603636
          ed |   .0938154    .097555     0.96   0.336    -.0973888    .2850197
       white |   .4723436   .6043097     0.78   0.434    -.7120817    1.656769
       _cons |  -1.091353   1.450218    -0.75   0.452    -3.933728    1.751022
-------------+----------------------------------------------------------------
4            |
       exper |   .0345959   .0188294     1.84   0.066     -.002309    .0715007
          ed |   .3531577   .1172786     3.01   0.003     .1232959    .5830194
       white |   1.571385   .9027216     1.74   0.082    -.1979166    3.340687
       _cons |  -6.238608   1.899094    -3.29   0.001    -9.960764   -2.516453
-------------+----------------------------------------------------------------
5            |
       exper |   .0356509    .018037     1.98   0.048      .000299    .0710028
          ed |   .7788519   .1146293     6.79   0.000     .5541826    1.003521
       white |   1.774306   .7550543     2.35   0.019     .2944273    3.254186
       _cons |  -11.51833   1.849356    -6.23   0.000      -15.143   -7.893659
------------------------------------------------------------------------------
(Outcome occ==1 is the comparison group)
```

2.4 Test for combination of categories

- It may be the case that some of the categories for the dependent variable can be combined. We can test for this:

  1. Subset the data by the observations with the two outcomes under consideration.
  2. Estimate a binary logit on the subsample.
  3. Compute a likelihood ratio (LR) test that the slope coefficients are simultaneously 0.

  If a statistically significant LR statistic is obtained, then we reject the hypothesis that the categories are indistinguishable.

Section 3

The Conditional Logit Model

- The conditional logit model is derived from random utility models.
- Key difference between the MNL model and the conditional logit model: the latter considers the effects of the choice characteristics on the determinants of choice probabilities.
- Suppose an individual i can choose from J alternatives, we can write down the utility of choice j as

$$U_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \varepsilon_{ij}$$

- We assume that if i chooses j then the utility associated with alternative j is the largest of the utilities obtained from the various choices. So we are interested in

$$\Pr(U_{ij} > U_{ik}) \qquad \text{for } k \neq j$$

- For example,

$$\begin{aligned}
\Pr(y_i = 1) &= \Pr(U_{i1} > U_{i2}) \\
&= \Pr(\boldsymbol{\beta}'\mathbf{x}_{i1} + \varepsilon_{i1} > \boldsymbol{\beta}'\mathbf{x}_{i2} + \varepsilon_{i2}) \\
&= \Pr(\varepsilon_{i1} - \varepsilon_{i2} > \boldsymbol{\beta}'\mathbf{x}_{i2} - \boldsymbol{\beta}'\mathbf{x}_{i1})
\end{aligned}$$

- To derive the statistical model, we make some distributional assumptions about ε_{ij}. Specifically, we assume it has an extreme value distribution:

$$F(\varepsilon_{ij}) = \exp(\exp(-\varepsilon_{ij}))$$
$$f(\varepsilon_{ij}) = \exp(-\varepsilon_{ij})\exp(\exp(-\varepsilon_{ij}))$$

- Key things to note:
  1. We assume that the ε_{ij} are iid. This is a very strong assumption: not only are they iid across decision makers, but iid across alternatives for each decision-maker.
  2. We have common parameters across choices.

- Let Y_i indicate the choice of an individual. Then given our assumptions, we obtain the **conditional logit** model as

$$\Pr(Y_i = j) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}'\mathbf{x}_{ik})}$$

- The log likelihood for this model looks the same as for the MNL model.

3.1 Equivalence of the MNL model and conditional logit model

- It is important to note that the MNL model and conditional logit model are algebraically equivalent. To see this let

$$\boldsymbol{\gamma}_2' = \begin{pmatrix}\mathbf{0} & \boldsymbol{\beta} & \mathbf{0}\end{pmatrix} \qquad \boldsymbol{\gamma}_1' = \begin{pmatrix}\boldsymbol{\beta} & \mathbf{0} & \mathbf{0}\end{pmatrix} \qquad \mathbf{z}_i = \begin{pmatrix}\mathbf{x}_{i1}\\ \mathbf{x}_{i2}\\ \mathbf{x}_{i3}\end{pmatrix}$$

- Then

$$\begin{aligned}\frac{P_2}{P_1} &= \exp\left[\left(\begin{pmatrix}\mathbf{0} & \boldsymbol{\beta} & \mathbf{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\beta} & \mathbf{0} & \mathbf{0}\end{pmatrix}\right)\begin{pmatrix}\mathbf{x}_{i1}\\ \mathbf{x}_{i2}\\ \mathbf{x}_{i3}\end{pmatrix}\right]\\ &= \exp\left[\begin{pmatrix}-\boldsymbol{\beta} & \boldsymbol{\beta} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{i1}\\ \mathbf{x}_{i2}\\ \mathbf{x}_{i3}\end{pmatrix}\right]\\ &= \exp\left[\boldsymbol{\beta}(\mathbf{x}_{i2} - \mathbf{x}_{i1})\right]\end{aligned}$$

- We could combine the MNL and conditional logit models to get something like

$$P_{ij} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\gamma}_j'\mathbf{z}_i)}{\sum_{k=1}^{J}\exp(\boldsymbol{\beta}'\mathbf{x}_{ik} + \boldsymbol{\gamma}_k'\mathbf{z}_i)}$$

3.2 Application of a model combining both choice-level and individual-level characteristics (from `clogit` section in Stata manual)

- We have data on consumer's choices of automobiles:
  - **car**: American, European, Japanese.
  - **dealer**: number of dealers for each type of car in consumer's city.
  - **sex**: gender of consumer.
  - **income**: income of consumer.

  In order to examine the effects of consumer attributes on the choices, we must interact them with dummy variables for the choices:

```
gen japan = (car==2);

gen europe = (car==3);

gen sexJap = sex*japan;

gen sexEur = sex*europe;

gen incJap = income*jap;

gen incEur = income*europe;

clogit choice japan europe sexJap sexEur incJap incEur dealer, group(id);
```

This produces

```
Iteration 0:   log likelihood = -284.51561
Iteration 1:   log likelihood = -251.47313
Iteration 2:   log likelihood = -250.78678
Iteration 3:   log likelihood =  -250.7794
Iteration 4:   log likelihood =  -250.7794

Conditional (fixed-effects) logistic regression   Number of obs   =        885
                                                  LR chi2(7)      =     146.62
                                                  Prob > chi2     =     0.0000
Log likelihood =  -250.7794                       Pseudo R2       =     0.2262

------------------------------------------------------------------------------
      choice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       japan |  -1.352189   .6911829    -1.96   0.050    -2.706882    .0025049
      europe |  -2.355249   .8526681    -2.76   0.006    -4.026448   -.6840502
      sexJap |  -.5346039   .3141564    -1.70   0.089    -1.150339    .0811314
      sexEur |   .5704111   .4540247     1.26   0.209     -.319461    1.460283
      incJap |   .0325318    .012824     2.54   0.011     .0073973    .0576663
      incEur |    .032042   .0138676     2.31   0.021      .004862    .0592219
      dealer |   .0680938   .0344465     1.98   0.048       .00058    .1356076
------------------------------------------------------------------------------
```

To get the results in terms of odds ratios do:

```
clogit, or;

------------------------------------------------------------------------------
      choice | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       japan |   .2586735   .1787907    -1.96   0.050     .0667446    1.002508
      europe |   .0948699   .0808925    -2.76   0.006     .0178376    .5045692
      sexJap |   .5859013   .1840647    -1.70   0.089     .3165294    1.084513
      sexEur |   1.768994    .803167     1.26   0.209     .7265405    4.307179
      incJap |   1.033067    .013248     2.54   0.011     1.007425    1.059361
      incEur |   1.032561   .0143191     2.31   0.021     1.004874    1.061011
      dealer |   1.070466   .0368737     1.98   0.048      1.00058    1.145232
------------------------------------------------------------------------------
```

3.3 Independence of Irrelevant Alternatives

- The multinomial logit model has the property known as **independence of irrelevant alternatives** which follows from the iid assumption for the disturbances.
- From above
$$\ln\left(\frac{P_j}{P_k}\right) = [\boldsymbol{\gamma}_j' - \boldsymbol{\gamma}_k']\mathbf{z}_i$$
- This means that the odds ratio between any two choices does not depend on the other choices. This can be problematic. To see why consider the classic "red bus, blue bus" problem.
- Suppose individuals have 2 modes of transportation they can choose from:
    1. Car.
    2. Red bus.
- Assume individuals are indifferent between traveling by car and by bus. This implies
$$\Pr(\text{car}) = \Pr(\text{red bus}) = 1/2.$$
- The odds of taking the car v. the red bus:
$$\Pr(\text{car})/\Pr(\text{red bus}) = 1.$$

- Suppose we add the choice of a blue bus. IIA requires

$$\Pr(\text{car}) = \Pr(\text{red bus}) = \Pr(\text{blue bus}) = 1/3$$

since $\Pr(\text{car})/\Pr(\text{red bus})$ must still equal 1.

- But if individuals simply split evenly in their choice of the red v. blue bus (which makes sense), then

$$\Pr(\text{red bus}) = \Pr(\text{blue bus}) = 1/4$$

- But this means

$$\frac{\Pr(\text{car})}{\Pr(\text{red bus})} = 2$$

  - Intuition: we have to be careful what we define as an "irrelevant alternative."
  - The event that an individual chooses the car over the red bus should make the event that an individual chooses the car over the blue bus more likely.
    - But we've essentially ruled this out by assuming iid errors.
  - If individual gets higher utility out of driving a car than taking red bus, should also get higher utility for driving over taking blue bus

- The IIA assumption renders the MNL model highly problematic for many settings in the social sciences.
- For example:
  - Suppose a country has a liberal and conservative party, and a new conservative party enters.
  - IIA $\Rightarrow$ entrance of the 2nd conservative party should not affect the relative probability of an individual choosing b/t the liberal party and the first conservative party.
  - Not likely if voters view the two conservative parties as similar.

3.4 IIA test

- Intuition: if IIA holds, then you can estimate the model leaving out one of the alternatives and the coefficient estimates will not change significantly.
- If you estimate the model excluding an irrelevant alternative then the parameter estimates will be consistent, but not efficient.
- But if the relative odds ratios between the choices that remain are not truly independent, then we will get inconsistent estimates after we eliminate the choice.
- To perform this test we compute the statistic

$$\chi_k^2 = \left(\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f\right)' \left[\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f\right]^{-1} \left(\hat{\boldsymbol{\beta}}_s - \hat{\boldsymbol{\beta}}_f\right)$$

  where s indicates the estimator using the subset of alternatives, f indicates the estimator using the full set of alternatives, and $\hat{\mathbf{V}}_s$ and $\hat{\mathbf{V}}_f$ are the respective asymptotic covariance matrices (k = number of rows in $\hat{\boldsymbol{\beta}}_s$).
- Note that if $\hat{\mathbf{V}}_s - \hat{\mathbf{V}}_f$ is not positive semidefinite, the statistic we compute can be negative. This is usually taken as evidence that IIA holds.
- Test is generally of low power.

Section 4

The Nested Logit Model

- IIA is a problem for the MNL model when some of the alternatives are closely related. The nested MNL (NMNL) alleviates this problem to a certain extent.
- One way to think about the NMNL model is that we are treating decisions as sequential.
- For example, suppose we want to study votes for candidates in multimember districts with multiple parties. We can think of voters first deciding which party they will vote for and then which of the party's candidates to vote for ($\Rightarrow$ tree-like structure to the decision making process).
- Suppose individuals can choose from parties indexed by $i = 1, 2, \ldots, C$ and candidates indexed by $j = 1, 2, \ldots, N_i$ from party i.
- Let

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

where

  - U_{ij} is the voters utility for alternative (i, j).
  - V_{ij} is a function of the measured attributes of this alternative.
  - ε_{ij} includes other factors not measured that affect utility.

- If we assume ε_{ij} has an extreme value distribution then the probability P_{ij} that the (i,j)th alternative will be chosen is given by

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{m=1}^{C} \sum_{n=1}^{N_m} \exp(V_{mn})}$$

- Let

$$V_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\gamma}'\mathbf{z}_i$$

where $\mathbf{x}_{ij}$ is a vector of variables that vary with **both** party and candidates and $\mathbf{z}_i$ is a vector of variables that vary only with the party.

- We can write P_{ij} as $P_{j|i} \cdot P_i$ where

$$\begin{aligned} P_{j|i} &= \frac{\exp(V_{ij})}{\sum_{k=1}^{N_i} \exp(V_{ik})} \\ &= \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i)\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\exp(\boldsymbol{\gamma}'\mathbf{z}_i)\sum_{k=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ik})} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{k=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ik})} \end{aligned}$$

and

$$\begin{aligned} P_i &= \frac{\sum_{j=1}^{N_i} \exp(V_{ij})}{\sum_{m=1}^{C} \sum_{n=1}^{N_m} \exp(V_{mn})} \\ &= \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i)\sum_{j=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{m=1}^{C} \exp(\boldsymbol{\gamma}'\mathbf{z}_m)\sum_{n=1}^{N_m} \exp(\boldsymbol{\beta}'\mathbf{x}_{mn})} \end{aligned}$$

- Next, define an inclusive value for the ith choice:

$$I_i = \ln \sum_{j=1}^{N_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ij})$$

- Then the above equations become

$$P_{j|i} = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\exp(I_i)}$$
$$P_i = \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i + I_i)}{\sum_{m=1}^{C} \exp(\boldsymbol{\gamma}'\mathbf{z}_m + I_m)}$$

- We can rewrite P_i as

$$P_i = \frac{\exp(\boldsymbol{\gamma}'\mathbf{z}_i + \tau_i I_i)}{\sum_{m=1}^{C} \exp(\boldsymbol{\gamma}'\mathbf{z}_m + \tau_i I_m)}$$

  where τ_i is a parameter that we will estimate. Note that above we restricted $\tau_i = 1$. If we let τ vary then we have the nested logit model.

- To estimate this model we can use a 2-stage procedure:
  1. Estimate $\boldsymbol{\beta}$ in $P_{j|i}$ by doing conditional logit.
  2. Compute the inclusive values using the estimates of $\boldsymbol{\beta}$ for all i categories. Estimate $\boldsymbol{\gamma}$ and τ by conditional logit.

- As with Heckman's two step method for selection models, we need to correct the standard errors from the second stage.

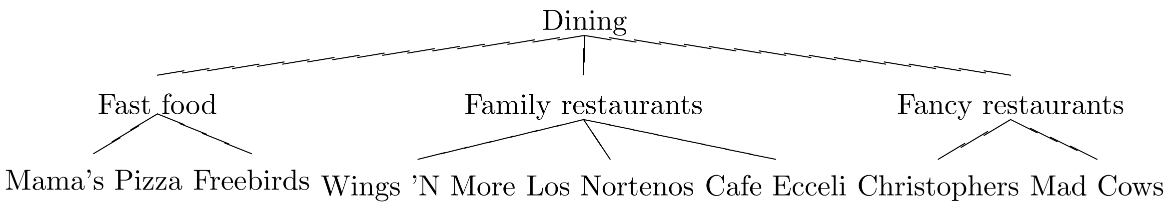
- Another alternative would be to just do full information maximum likelihood, which involves maximizing the log-likelihood

$$\ln L = \sum_{i=1}^{n} \ln\left(P_{j|i} \cdot P_i\right)$$

- Note that this model can be generalized to more choice-levels.

4.1 Application of NMNL (from **Stata** Manual)

- Suppose we are interested in a family's choice of restaurants. The tree structure is as follows:



- The model we will estimate is

$$\begin{aligned}\Pr(\text{restaurant}|\text{type}) &= \Pr(\beta_{\text{cost}}\text{cost} + \beta_{\text{rating}}\text{rating} + \beta_{\text{distance}}\text{distance})\\ \Pr(\text{type}) &= \Pr(\alpha_{\text{iFast}}\text{incFast} + \alpha_{\text{iFancy}}\text{incFancy}\\ &\quad + \alpha_{\text{kFast}}\text{kidFast} + \alpha_{\text{kFancy}}\text{kidFancy}\\ &\quad + \tau_{\text{fast}}\text{IV}_{\text{fast}} + \tau_{\text{family}}\text{IV}_{\text{family}} + \tau_{\text{fancy}}\text{IV}_{\text{fancy}})\end{aligned}$$

- To estimate the effects of certain explanatory variables on dining choice using `nlogit` in Stata, do the following:

1. Set up the tree structure:

```
nlogitgen type = restaurant(fast: Freebirds | MamasPizza, family:
  CafeEccell | LosNortenos | WingsNmore, fancy: Christophers | MadCows);
```

2. Define some new variables for the "top" level choice (i.e., $\mathbf{z_i}$):

```
. gen incFast = (type==1) * income;

. gen incFancy = (type==3) * income;

. gen kidFast = (type==1) * kids;

. gen kidFancy = (type==3) * kids;
```

3. Run the model:

```
nlogit chosen (restaurant = cost rating distance )
(type = incFast incFancy kidFast kidFancy), group(family_id) nolog;
```

This produces:

```
Nested logit
Levels             =          2                 Number of obs      =      2100
Dependent variable =     chosen                 LR chi2(10)        =  199.6293
Log likelihood     =  -483.9584                 Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
restaurant   |
        cost |  -.0944352     .03402    -2.78   0.006    -.1611131   -.0277572
      rating |   .1793759    .126895     1.41   0.157    -.0693338    .4280855
    distance |  -.1745797   .0433352    -4.03   0.000    -.2595152   -.0896443
-------------+----------------------------------------------------------------
type         |
     incFast |  -.0287502   .0116242    -2.47   0.013    -.0515332   -.0059672
    incFancy |   .0458373   .0089109     5.14   0.000     .0283722    .0633024
     kidFast |  -.0704164   .1394359    -0.51   0.614    -.3437058    .2028729
    kidFancy |  -.3626381   .1171277    -3.10   0.002    -.5922041   -.1330721
-------------+----------------------------------------------------------------
(IV params)  |
             |
type         |
       /fast |   5.715758   2.332871     2.45   0.014     1.143415     10.2881
     /family |   1.721222   1.152002     1.49   0.135    -.5366608    3.979105
      /fancy |   1.466588   .4169075     3.52   0.000     .6494642    2.283711
------------------------------------------------------------------------------
LR test of homoskedasticity (iv = 1): chi2(3)=    9.90    Prob > chi2 = 0.0194
------------------------------------------------------------------------------
```

Section 5

The Multinomial Probit Model

- The multinomial logit model and the conditional logit model are unattractive because they don't allow for correlation between the error terms (i.e., assume $\varepsilon_{ij} \perp \varepsilon_{ik}$ for $j \neq k$).
- The **multinomial probit (MNP) model** relaxes this assumption, though as we will see, not without substantial cost.
- The derivation of the MNP model is similar to the derivation of the conditional logit model in that we start from a random utility model. To keep the derivation simple, we will look at the special case of three alternatives:

$$U_1 = \boldsymbol{\beta}'\mathbf{x}_1 + \boldsymbol{\gamma}_1'\mathbf{z} + \varepsilon_1$$
$$U_2 = \boldsymbol{\beta}'\mathbf{x}_2 + \boldsymbol{\gamma}_2'\mathbf{z} + \varepsilon_2$$
$$U_3 = \boldsymbol{\beta}'\mathbf{x}_3 + \boldsymbol{\gamma}_3'\mathbf{z} + \varepsilon_3$$

  (Note we are suppressing the i subscript.)

- Let $\bar{U}_j = \boldsymbol{\beta}'\mathbf{x}_j + \boldsymbol{\gamma}_j'\mathbf{z}$.
- Assume $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$
- The probability of choosing the first alternative is

$$\Pr(U_1 > U_2, U_1 > U_3) = \Pr(\varepsilon_2 - \varepsilon_1 < \bar{U}_1 - \bar{U}_2, \varepsilon_3 - \varepsilon_1 < \bar{U}_1 - \bar{U}_3)$$

- Let $\eta_{21} = \varepsilon_2 - \varepsilon_1$ and $\eta_{31} = \varepsilon_3 - \varepsilon_1$. Then η_{21} and η_{31} have a bivariate normal distribution with covariance matrix

$$\Omega_1 = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} & \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} \\ \sigma_1^2 - \sigma_{13} - \sigma_{12} + \sigma_{23} & \sigma_1^2 + \sigma_3^2 - 2\sigma_{13} \end{bmatrix}$$

- Then the probability of choosing category 1 is

$$P_1 = \int_{-\infty}^{\bar{U}_1 - \bar{U}_2} \int_{-\infty}^{\bar{U}_1 - \bar{U}_3} f(\eta_{21}, \eta_{31}) d\eta_{31}, d\eta_{21}$$

- The probabilities for the other categories are defined similarly.
- The likelihood function for the MNP model looks the same as the likelihood function for the MNL and conditional logit models.
- Traditionally, the main barrier to using the MNP model has been its computational difficulty.
- Some advances have been made recently using simulation methods, but there is still not much progress beyond 3 categories (at least in political science).
- Fundamental identification issues remain.

5.1 Identification of variance-covariance matrix

- Not all of the parameters in Σ can be estimated.
- For the model written in utility differences, it was traditionally thought that $J(J-1)/2$ free parameters could be estimated. More recent treatments claim that it is possible to estimate only $J(J-1)/2-1$ elements. So need to make assumptions about restrictions.
- Easy way to test if you've got an identification problem: nonsensical parameters or parameters on the boundary.
- A standard assumption is $\sigma_j^2 = 1 \ \forall \ j$ (i.e., homoskedasticity across choices).

5.2 Application: Alvarez and Nagler "Economics, Issues, and the Perot Candidacy: Voter Choice in the 1992 Presidential Election" '95 *AJPS*

- Key question of concern: how did the candidacy of Ross Perot—a rare serious 3rd party candidate—affect the outcome of the 1992 presidential election
- Did Perot take away votes from Clinton, Bush, or both?
- Obviously, a model that assumes IIA would have difficulty answering these questions.
- They estimate an 3-choice MNP using Nat'l Election Survey data. Their model includes both choice-specific and individual-specific characteristics.
- Normalize on the Perot choice, which makes sense given the stated goals of the study.
- See table for results.
- In a similar study, Lacy and Burden ("The Vote-Stealing and Turnout Effects of Ross Perot in the 1992 U.S. Presidential Election" '99 *AJPS*) include the choice to vote in the election.
- However, some questions are raised about the model's identification.

Table 3. Multinomial Probit Estimates for a Three-Candidate Model (Perot Coefficients Normalized to Zero)

Independent Variables	Coefficients for Bush		Coefficients for Clinton
Ideological Distance		−.09* .02	
Constant	.50 .44		−.44 .58
Felt personal finances were worse	−.04 .05		.02 .05
Felt national economy was worse	−.14** .08		.21* .10
Oppose government jobs	.07 .05		−.01 .05
Oppose government health care	.10* .05		.06 .04
Oppose government minority assistance	.01 .05		−.17* .05
Abortion	−.35* .14		.01 .11
Region (East)	−.15 .17		.32 .21
Region (South)	.25 .18		.50* .19
Region (West)	−.11 .18		−.03 .21
New or returning voter	.28** .15		−.23* .17
Term limits	.06 .13		.08 .11
Felt deficit was a major problem	−.58* .22		−.003 .18
Democrat	−.19 .17		1.34* .28
Republican	1.00* .43		−.74 .46
Gender (Female)	.38* .19		.21 .14
Respondent's education	.14* .07		.004 .06
Age: 18–29	−.86* .41		−.57* .26
Age: 30–44	−.64* .30		−.54* .19
Age: 45–59	−.51* .24		−.10 .21
σ_{BC}		−.08 .28	
σ_{BP}		.27 .54	
σ_{CP}		−.07 .26	

Note: Maximum-likelihood estimates with their estimated standard errors below.
LL = −568.18; % correct = 70.6; number of observations = 909.
*indicates an estimate significant at the $p = .05$ level.
**indicates an estimate significant at the $p = .10$ level.

Table 3. Multinomial Probit Estimates for Four–Choice Model

	Clinton	Perot	Abstain
Constant	.313	−1.510	1.820**
	(.519)	(.981)	(.602)
Personal Finances (worse)	.102*	.141*	.007
	(.046)	(.061)	(.051)
National Economy (worse)	.207**	.094	.133*
	(.067)	(.082)	(.064)
National Health Care (oppose)	−.167**	−.193**	−.137**
	(.042)	(.054)	(.046)
Abortion (pro–choice)	.320**	.309**	.119
	(.069)	(.101)	(.076)
Democrat	.917**	0.00	0.00
	(.179)	—	—
Independent	0.00	.399**	0.00
	—	(.202)	—
Liberal	.781**	.405	.356
	(.218)	(.275)	(.226)
Conservative	−.716**	−.702**	−.528**
	(.180)	(.234)	(.201)
No Ideology	−.078	−.437	.289
	(.203)	(.306)	(.203)
Voted in 1988	0.00	.137	−1.86**
	—	(.309)	(.437)
Clinton Moral (disagree)	−.468**	0.00	0.00
	(.103)	—	—
South	−.122	−.486*	.072
	(.154)	(.225)	(.155)
Gender (male)	−.290*	−.561**	−.348*
	(.147)	(.203)	(.156)
Race (African-American)	1.040**	−.287	.478*
	(.286)	(.597)	(.289)
Income	−.030*	−.007	−.057*
	(.014)	(.019)	(.017)
Age 18–29	.134	1.02**	−.022
	(.234)	(.368)	(.246)
Age 30–44	.059	.133	.231
	(.220)	(.283)	(.263)
Age 45–59	−.022	.410	.279
	(.202)	(.270)	(.209)
College Educated	0.00	0.00	−.429*
	—	—	(.191)
External Efficacy	0.00	0.00	−.076*
	—	—	(.046)
$\hat{\sigma}_{Bush,Perot}$	−.99		
	(.91)		
$\hat{\sigma}_{Clinton,Perot}$	.11		
	(.47)		
$\hat{\sigma}_{Bush,Abstain}$	.23		
	(.30)		
$\hat{\sigma}_{Clinton,Abstain}$	−.09		
	(.42)		
$\hat{\sigma}_{Perot,Abstain}$	−.59		
	(.96)		

(continued)

Table 3. Multinomial Probit Estimates for Four–Choice Model (*continued*)

Final Log Likelihood	–1314.12
$\chi^2(57)$	1450.24**
Number of Iterations	35
Percent Correctly Predicted	50.6
Number of Cases	1471

Note: Bush coefficients normalized to zero. Coefficients are maximum likelihood estimates with standard errors in parentheses. Parameters fixed at zero are indicated by 0.00 coefficients and (—) standard errors. * indicates $p < .05$, ** indicates $p < .01$, one–tailed.
Source: 1992 National Election Study

Part II

Models for Repeated Observations

Section 6

Introduction

- Having multiple observations of cross-sectional units over time presents both new opportunities and new challenges.
- One of the key advantages of repeated observation models: allow you to exploit the structure of the data to estimate individual specific effects which you do not get to observe.
- Account for heterogeneity in the data to get cleaner estimates of explanatory variables of interest.
- Models have been developed for a variety of the data situations, but we'll focus on models for dichotomous dependent variables.
- Representative regression model for panel data:

$$y_{it} = \beta x_{it} + \alpha_i + u_{it} \tag{6.1}$$

  where

  - i denotes the cross-sectional units ($i = 1, \ldots, N$),
  - t denotes the time period ($t = 1, \ldots, T$),
  - x_{it} is an exogenous explanatory variable,
  - β are parameters to be estimated,
  - α_i is an individual specific effect which we do not get to observe,
  - u_{it} is a random disturbance term.

Section 7

General Issues with Repeated Observations Data

- Standard distinction between types of repeated observations data:
  1. T is large relative to N: Time-Series Cross-Section (TSCS) data.
  2. N is large relative to T: Panel data.
- How to model individual specific effects:
  1. Fixed effects: like having separate intercepts for each individual.
  2. Random effects: effects have a distribution with a mean and a variance.
- Non-standard error structures (TSCS):
  1. Panel heteroskedasticity:
  $$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$
  2. Contemporaneous correlation of the errors
  $$E(u_{it}, u_{js}) = \begin{cases} \sigma_i^2 & \text{if } i = j \text{ and } s = t \\ \sigma_{ij} & \text{if } i \neq j \text{ and } s = t \\ 0 & \text{otherwise} \end{cases}$$
  3. Serially correlated errors: $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$.

Section 8

Dichotomous Dependent Variables

- The basic set up of the panel model for dichotomous dependent variables is similar to the standard models.

$$y_{it}^* = \beta \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.1}$$

where

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0 \\ 0 & \text{if } y_{it}^* \leq 0 \end{cases}$$

and $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and u_{it} is assumed to be iid with mean zero and variance σ_u^2.

- The choice we make about the distribution for the disturbance term matters a lot with panel data and is based on our beliefs about the correlation between the explanatory variables and the individual specific effect.

- If we think there is no correlation between α_i and $\mathbf{x}_{it}$ then we can estimate a **random effects model**. With this model we assume that $\alpha \sim IID(0, \sigma_\alpha^2)$.

- If we think there is correlation between α_i and $\mathbf{x}_{it}$ then we can estimate a **fixed effects model**. Here we assume that the α_i are fixed parameters to be estimated.

- If $T \to \infty$, then it is possible to get consistent estimates of $\boldsymbol{\beta}$ and α_i.
- However, if T is fixed and $N \to \infty$, then we have the incidental parameters problem—i.e., since the number of parameters increases with N, we cannot consistently estimate α_i for fixed T.
- Unfortunately, the inconsistency in α_i is transmitted to $\boldsymbol{\beta}$.
- In the linear regression case we take out the individual effect by subtracting off the means of the variables. That is, if we take the average over time of (8.1), we get

$$\bar{y}_i^* = \beta \bar{x}_i + \alpha_i + \bar{u}_i, \tag{8.2}$$

where $\bar{y}_i^* = \sum_{t=1}^{T} y_{it}^*/T$. Subtracting (8.2) from (8.1) gives

$$y_{it}^* - \bar{y}_i^* = \beta(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i), \tag{8.3}$$

- But this transformation is not valid with a qualitative limited dependent variable model.
- Such models are nonlinear and so mean-differencing will not work to remove the individual specific effects.

8.1 Fixed Effect Logit

- Chamberlain ('80 *Rev. of Econ. Studies*) has derived a conditional maximum likelihood estimator (CMLE) that works in a similar way. The conditional likelihood function can be written as

$$L = \prod_{i=1}^{N} \Pr\left(y_{i1}, \ldots, y_{iT} | \sum_{t=1}^{T} y_{it}\right)$$

- Consider the case where $T = 2$. The unconditional likelihood is

$$L = \prod_{i=1}^{N} \Pr(y_{i1}) \Pr(y_{i2})$$

- Note that:

$$\Pr[y_{i1} = 0, y_{i2} = 0 | y_{i1} + y_{i2} = 0] = 1$$
$$\Pr[y_{i1} = 1, y_{i2} = 1 | y_{i1} + y_{i2} = 2] = 1$$

  which means these probabilities add no information to the conditional log likelihood so we can ignore them.

- But

$$\Pr[y_{i1} = 0, y_{i2} = 1 | y_{i1} + y_{i2} = 1] = \frac{\Pr[y_{i1} = 0, y_{i2} = 1 \text{ and } y_{i1} + y_{i2} = 1]}{\Pr[y_{i1} + y_{i2} = 1]}$$

$$= \frac{\Pr[y_{i1} = 0, y_{i2} = 1 \text{ and } y_{i1} + y_{i2} = 1]}{\Pr[y_{i1} = 0, y_{i2} = 1] + \Pr[y_{i1} = 1, y_{i2} = 0]}$$

- If we assume that the data follow a logistic distribution then we can rewrite this as

$$\frac{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}{\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}+\frac{\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i1})}\frac{1}{1+\exp(\alpha_i+\boldsymbol{\beta}'\mathbf{x}_{i2})}}$$

  which simplifies to

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}{\exp(\boldsymbol{\beta}'\mathbf{x}_{i1})+\exp(\boldsymbol{\beta}'\mathbf{x}_{i2})}$$

- The expression for the remaining probability is similarly derived. These probabilities then constitute the conditional log-likelihood function, which can be maximized using standard techniques.
- This can be extended to T of arbitrary size but the computations are excessive for $T > 10$.

- Can't use standard specification test like LR for checking unit heterogeneity b/c likelihoods are not comparable (CML uses a restricted data set).
- But can use this estimator to do a Hausman test for the presence of individual effects.
- Intuition: in the absence of individual specific effects, both the Chamberlain estimator and the standard logit maximum likelihood estimator are consistent, but the former is inefficient. If individual specific effects exist, then the Chamberlain estimator is consistent while the standard logit MLE is inconsistent.
  - Inefficiency is due to loss of information/throwing away observations.
- We compute the following statistic for the test:

$$\chi_k^2 = \left(\hat{\boldsymbol{\beta}}_{\mathrm{CML}} - \hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)' \left[\mathbf{V}_{\mathrm{CML}} - \mathbf{V}_{\mathrm{ML}}\right]^{-1} \left(\hat{\boldsymbol{\beta}}_{\mathrm{CML}} - \hat{\boldsymbol{\beta}}_{\mathrm{ML}}\right)$$

  If we get a significant χ^2 value we reject the null of no individual specific effects.
- If $\mathbf{V}_{\mathrm{ML}} > \mathbf{V}_{\mathrm{CML}}$, assume zero χ^2 statistic.

8.2 Application: Unionization of Women in the U.S. (from **Stata** manual)

- Chamberlain's fixed effects logit model can be estimated using `xtlogit` with the `fe` option.
- Sample is a panel (unbalanced) of 4,434 women from 1970–1988.
- Dep. var. is a dummy indicating union membership.
- Ind. vars include: age, years of schooling, amount of time spent living outside an SMSA, years spent living in the South, and an interaction between this variable and year.
- First, let's estimate a pooled logit model.

```
. logit union age grade not_smsa south southXt;

Iteration 0:   log likelihood =  -13864.23
Iteration 1:   log likelihood = -13550.511
Iteration 2:   log likelihood =  -13545.74
Iteration 3:   log likelihood = -13545.736

Logit estimates                                   Number of obs   =      26200
                                                  LR chi2(5)      =     636.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -13545.736                       Pseudo R2       =     0.0230

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0099931   .0026737     3.74   0.000      .0047527    .0152335
       grade |   .0483487   .0064259     7.52   0.000      .0357541    .0609432
    not_smsa |  -.2214908   .0355831    -6.22   0.000     -.2912324   -.1517493
       south |  -.7144461   .0612145   -11.67   0.000     -.8344244   -.5944678
     southXt |   .0068356   .0052258     1.31   0.191     -.0034067    .0170779
       _cons |  -1.888256    .113141   -16.69   0.000     -2.110009   -1.666504
------------------------------------------------------------------------------
```

Now let's estimate the fixed effects model, accounting for unit heterogeneity.

```
. xtlogit union age grade not_smsa south southXt, i(id) fe;

note: multiple positive outcomes within groups encountered.
note: 2744 groups (14165 obs) dropped due to all positive or
      all negative outcomes.
Iteration 0:   log likelihood = -4541.9044
Iteration 1:   log likelihood = -4511.1353
Iteration 2:   log likelihood = -4511.1042

Conditional fixed-effects logit                 Number of obs      =     12035
Group variable (i) : idcode                     Number of groups   =      1690

                                                Obs per group: min =         2
                                                               avg =       7.1
                                                               max =        12

                                                LR chi2(5)         =     78.16
Log likelihood  = -4511.1042                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       union |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0079706   .0050283     1.59   0.113    -.0018848    .0178259
       grade |   .0811808   .0419137     1.94   0.053    -.0009686    .1633302
    not_smsa |   .0210368    .113154     0.19   0.853    -.2007411    .2428146
       south |  -1.007318   .1500491    -6.71   0.000    -1.301409   -.7132271
     southXt |   .0263495   .0083244     3.17   0.002      .010034    .0426649
------------------------------------------------------------------------------
```

To run the Hausman test, simply type

```
.hausman
```

In this case, the test gives:

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     Prior       Current       Difference         S.E.
-------------+----------------------------------------------------------------
         age |   .0079706     .0099931        -.0020225      .0042586
       grade |   .0811808     .0483487         .0328321      .0414182
    not_smsa |   .0210368    -.2214908         .2425276      .1074136
       south |  -1.007318    -.7144461        -.2928718      .1369945
     southXt |   .0263495     .0068356         .0195139      .0064797
------------------------------------------------------------------------------
                b = less efficient estimates obtained previously from clogit
                B = fully efficient estimates obtained from logit

    Test:  Ho:  difference in coefficients not systematic

                chi2(  5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =    20.50
                          Prob>chi2 =     0.0010
```

Note that the **hausman** command requires the less efficient model to be estimated first.

8.3 Random Effects Probit

- The probit model does not lend itself to the fixed effects treatment because there is no way to sweep out the individual specific effects. But we can estimate a probit model if we assume random effects.
- Let

$$\varepsilon_{it} = \alpha_i + u_{it}$$

and assume $\alpha_i \sim N(0, \sigma_\alpha^2)$; $u_{it} \sim N(0, \sigma_u^2)$, and α_i and u_{it} are independent of each other. Then

$$\text{var}[\varepsilon_{it}] = \sigma_u^2 + \sigma_\alpha^2 = 1 + \sigma_\alpha^2$$

and

$$\text{corr}[\varepsilon_{it}, \varepsilon_{is}] = \rho = \frac{\sigma_\alpha^2}{1 + \sigma_\alpha^2}$$

for $t \neq s$. This implies $\sigma_\alpha^2 = \rho/(1+\rho)$.

- We can write the probability associated with an observation as

$$\Pr[y_{it}] = \int_{-\infty}^{q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}} f(\varepsilon_{it}) d\varepsilon_{it} = \Phi[q_{it}\boldsymbol{\beta}'\mathbf{x}_{it}]$$

- Because of the α_i, the T observations for i are jointly normally distributed. The individual's contribution to the likelihood is

$$\begin{aligned} L_i &= \Pr[y_{i1}, y_{i2}, \ldots, y_{iT}] \\ &= \int_{-\infty}^{q_{i1}\boldsymbol{\beta}'\mathbf{x}_{i1}} \int_{-\infty}^{q_{i2}\boldsymbol{\beta}'\mathbf{x}_{i2}} \cdots \int_{-\infty}^{q_{iT}\boldsymbol{\beta}'\mathbf{x}_{iT}} f(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{it}) d\varepsilon_{iT} \cdots \varepsilon_{i2}\varepsilon_{i1} \end{aligned}$$

- Rather than evaluating multiple integrals, a simplification is possible. Consider the joint density:

$$f(\varepsilon_{i1},\ldots,\varepsilon_{it},\alpha_i) = f(\varepsilon_{i1},\ldots,\varepsilon_{it}|\alpha_i)f(\alpha_i)$$

- We can then integrate over α_i:

$$f(\varepsilon_{i1},\ldots,\varepsilon_{it}) = \int_{-\infty}^{\infty} f(\varepsilon_{i1},\ldots,\varepsilon_{it}|\alpha_i)f(\alpha_i)d\alpha_i$$

- Conditioned on α_i, the ε_is are independent:

$$f(\varepsilon_{i1},\ldots,\varepsilon_{it}) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} f(\varepsilon_{it}|\alpha_i)f(\alpha_i)d\alpha_i$$

- This gives i's contribution to the likelihood as

$$L_i = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-r_i^2} \left\{ \prod_{t=1}^{T} \Phi[q_{it}(\boldsymbol{\beta}'\mathbf{x}_{it} + \theta r_i)] \right\} dr_i \tag{8.4}$$

$$\theta = \sqrt{\frac{2\rho}{1-\rho}} \tag{8.5}$$

- Things to note:
  - The assumption that the α_i and $\mathbf{x}_i$ are uncorrelated is very restrictive. We are also assuming that the within-cross section correlation is the same across all time periods.
  - ρ can be interpreted as the proportion of the variance contributed by the unit effects.
  - We can test for unit heterogeneity by checking the statistical significance of ρ. One way to do this is with a likelihood ratio ratio test of the random effects probit and pooled probit models.
  - The standard way to evaluate the integral in the likelihood is by Gauss-Hermite quadrature. This raises some concerns about how the size of T and N affect the accuracy of the quadrature approximation, and some checks of the performance of the approximation are in order.
  - Stata's `xtprobit` command can be used to estimate this model.
  - We could derive this model for the logistic distribution rather than the normal distribution.

8.4 Application: RE Probit for PAC contributions and roll call votes

- A major concern is the effects of campaign contributions on the behavior of members of the U.S. Congress.
- Assessing these effects is complicated because it is methodologically difficult to account for members' predispositions to vote in favor of PACs' interests.
- Estimating a RE probit model can help overcome this problem because it enables us to account for individual specific effects, such as the predisposition to vote for or against a particular piece of legislation, which are too costly or impossible to measure.
- The data set includes PAC contributions and roll-call votes that are of particular interest to certain types of PACs (see Wawro '01 *AJPS*).
- Let's estimate a pooled model and RE probit for USCC votes in the 102nd Congress, 2nd Session.

The pooled model:

```
. probit vote labor corp unemp;

Iteration 0:   log likelihood =  -1385.746
Iteration 1:   log likelihood = -1215.6587
Iteration 2:   log likelihood = -1214.1783
Iteration 3:   log likelihood =  -1214.178

Probit estimates                                  Number of obs   =       2002
                                                  LR chi2(3)      =     343.14
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1214.178                       Pseudo R2       =     0.1238

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       labor |  -.1197816   .0075367   -15.89   0.000    -.1345532    -.10501
        corp |   .0970168    .009864     9.84   0.000     .0776837    .1163499
       unemp |  -6.325817   1.252524    -5.05   0.000    -8.780719   -3.870916
       _cons |   .3775498   .1226202     3.08   0.002     .1372186     .617881
------------------------------------------------------------------------------
```

From these results, we would infer a strong relationship between contributions and votes.

And now the RE probit model:

```
. xtprobit vote labor corp unemp, i(icpsr) nolog;

Random-effects probit                           Number of obs      =      2002
Group variable (i) : icpsr                      Number of groups   =       286

Random effects u_i ~ Gaussian                   Obs per group: min =         7
                                                               avg =       7.0
                                                               max =         7

                                                Wald chi2(3)       =     44.06
Log likelihood  = -1158.7298                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       labor |  -.0606706   .0124172    -4.89   0.000    -.0850078   -.0363333
        corp |   .0494033   .0140651     3.51   0.000     .0218362    .0769705
       unemp |  -8.208123   2.193177    -3.74   0.000    -12.50667   -3.909576
       _cons |   .6124466   .2044377     3.00   0.003     .2117559    1.013137
-------------+----------------------------------------------------------------
    /lnsig2u |  -.5200013   .1903423                     -.8930653   -.1469373
-------------+----------------------------------------------------------------
     sigma_u |   .7710511   .0733818                      .6398429    .9291653
         rho |   .3728519   .0445084                      .2904777    .4633316
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =   110.90 Prob >= chibar2 = 0.000
```

These results indicate a less strong—but still statistically significant—relationship between contributions and votes.

8.5 Correlated Random Effects Probit

- The big drawback of the random effects model is the unattractive assumption that the α_i and $\mathbf{x}_i$ are uncorrelated.
- Chamberlain has proposed a correlated random effects (CRE) model that gets around this problem by assuming a specific functional relationship between α_i and $\mathbf{x}_i$. That is,

$$\alpha_i = \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \eta_i = \mathbf{a}' \mathbf{x}_i + \eta_i$$

  where $\mathbf{a}' = (\mathbf{a}_1', \ldots, \mathbf{a}_T')$, $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, and η_i is normally distributed (with mean zero and variance σ_η^2) and is independent of the $\mathbf{x}_{it}$.
- The $\mathbf{a}_t$ are vectors of parameters to be estimated and capture the nature of the relationship between α_i and $\mathbf{x}_{it}$.
- Equation (8.1) then becomes

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it} + \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \varepsilon_{it}$$

  where $\varepsilon_{it} = \eta_i + u_{it}$.

- Estimation of the CRE model proceeds in a sequential manner:

  1. Estimate separate probit equations by maximum likelihood for each time period, regressing the dependent variable in each period on all of the leads and lags of the explanatory variables.
  2. Stack the estimates from each of these probits into a vector $\hat{\pi}$ and construct the joint covariance matrix of all of the estimates. The vector $\hat{\pi}$ is a vector of reduced form estimates.
  3. Use a minimum distance estimator to impose restrictions on $\hat{\pi}$ to back out estimates of the structural parameters $\boldsymbol{\beta}$ and $\mathbf{a}_t$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{a}')$ and choose $\boldsymbol{\theta}$ to minimize

     $$[\hat{\pi} - \mathbf{f}(\theta)]' \hat{\boldsymbol{\Omega}}^{-1} [\hat{\pi} - \mathbf{f}(\theta)],$$

     where $\hat{\boldsymbol{\Omega}}$ is an estimate of the asymptotic variance-covariance matrix for the reduced-form estimates.
  4. We then conduct our standard hypothesis tests to make inferences about the effects of the variables of interest.
  5. Marginal effects can be determined by simulating probabilities:

     $$\Pr(y_{it} = 1) = \Phi \left[(1 + \sigma_\eta^2)^{-1/2} (\boldsymbol{\beta}' \mathbf{x}_{it} + \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it}) \right] \tag{8.6}$$

- One key advantage of the CRE estimator: allows for an explicit test for correlation between the individual specific effect and the explanatory variables.
- One drawback of the CRE estimator: requires us to impose a good deal of structure on the relationship between α_i and $\mathbf{x}_{it}$ and restricts us to including only time-varying variables in $\mathbf{x}_{it}$.
- Including time invariant variables in $\mathbf{x}_{it}$ in effect induces perfect collinearity since the values of the leads and the lags of these variables will be the same within each cross-sectional unit.
- Its drawbacks aside, this estimator is part of a class of generalized method of moments estimators that give substantial efficiency gains over pooled probit estimators, which ignore individual specific effects.
- For details of the estimation procedure for the CRE model, see Hsiao's treatment in *Analysis of Panel Data*, which is more accessible than Chamberlain's original derivation.
- Can be estimated w/ standard software with some more restrictive assumptions; GAUSS code is available for the general model.

8.6 Application: CRE Probit for PAC contributions and roll call votes

- Before we assumed that unit effects were uncorrelated with explanatory variables—a pretty heroic assumption.
- The CRE probit model is attractive for this analysis, especially since it explicitly models the CW that contributions and voting predispositions are correlated.
- I conducted a CRE probit analysis of roll-call votes and PAC contributions using sets of roll-call votes that are of particular interest to certain types of PACs (Wawro '01 *AJPS*).
  - Tables 3 and 4 report results of the CRE estimation.
  - Figures 2 & 4 plot simulated probabilities based on the CRE estimates.

TABLE 3 Panel Probit Model Results for Voting Behavior on AFL-CIO Roll-Call Votes, 102nd–104th Congresses

	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Variable	*Estimated coefficients**					
ln Labor PAC contributions	−0.006	0.002	−0.012	0.010	0.008	0.012
	(0.004)	(0.011)	(0.009)	(0.006)	(0.001)	(0.009)
ln Corporate PAC contributions	0.031	0.036	0.006	0.010	0.018	0.030
	(0.003)	(0.011)	(0.007)	(0.006)	(0.002)	(0.011)
Unemployment rate	−0.003	−0.008	0.001	−0.006	−0.011	−0.025
	(0.004)	(0.011)	(0.008)	(0.007)	(0.002)	(1.617)
	Tests of no correlation between individual effect and regressors†					
ln Labor PAC contributions	923.348	371.868	342.397	481.322	4,384.273	159.388
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
ln Corporate PAC contributions	170.824	33.342	106.950	32.075	3,986.840	102.342
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
Unemployment rate	112.456	3.222	117.859	31.739	892.259	12.804
	(< 0.001)	(.666)	(< 0.001)	(< 0.001)	(< 0.001)	(.025)
N	315	319	356	332	324	343
T	8	5	7	7	9	5

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

TABLE 4 Panel Probit Model Results for Voting Behavior on USCC Roll-Call Votes, 102nd–104th Congresses

	102nd Congress		103rd Congress		104th Congress	
	1st Session	2nd Session	1st Session	2nd Session	1st Session	2nd Session
Variable	*Estimated coefficients**					
ln Labor PAC contributions	0.025	−0.005	−0.012	0.043	0.008	0.022
	(0.004)	(0.007)	(0.003)	(0.011)	(0.001)	(0.004)
ln Corporate PAC contributions	−0.030	0.004	0.003	0.010	−0.011	0.015
	(0.005)	(0.006)	(0.004)	(0.009)	(0.001)	(0.004)
Unemployment rate	0.002	−0.010	0.023	0.018	0.004	0.002
	(0.005)	(0.007)	(0.004)	(0.011)	(0.001)	(0.005)
	Tests of no correlation between individual effect and regressors†					
ln Labor PAC contributions	676.705	267.680	119.077	55.355	6,441.995	606.078
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
ln Corporate PAC contributions	459.730	189.141	87.816	37.028	16,767.967	387.557
	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)	(< 0.001)
Unemployment rate	106.871	223.512	82.972	15.364	7,427.728	237.892
	(< 0.001)	(< 0.001)	(< 0.001)	(0.031)	(< 0.001)	(< 0.001)
N	302	288	357	344	340	338
T	8	7	7	7	10	8

Note: *Standard errors in parentheses.

† Entries are χ^2_T statistics. *p* values in parentheses.

FIGURE 2 **Effect of Labor PAC Contributions on the Probability of Voting in Favor of the AFL-CIO Positions, 104th Congress, 1st Session**

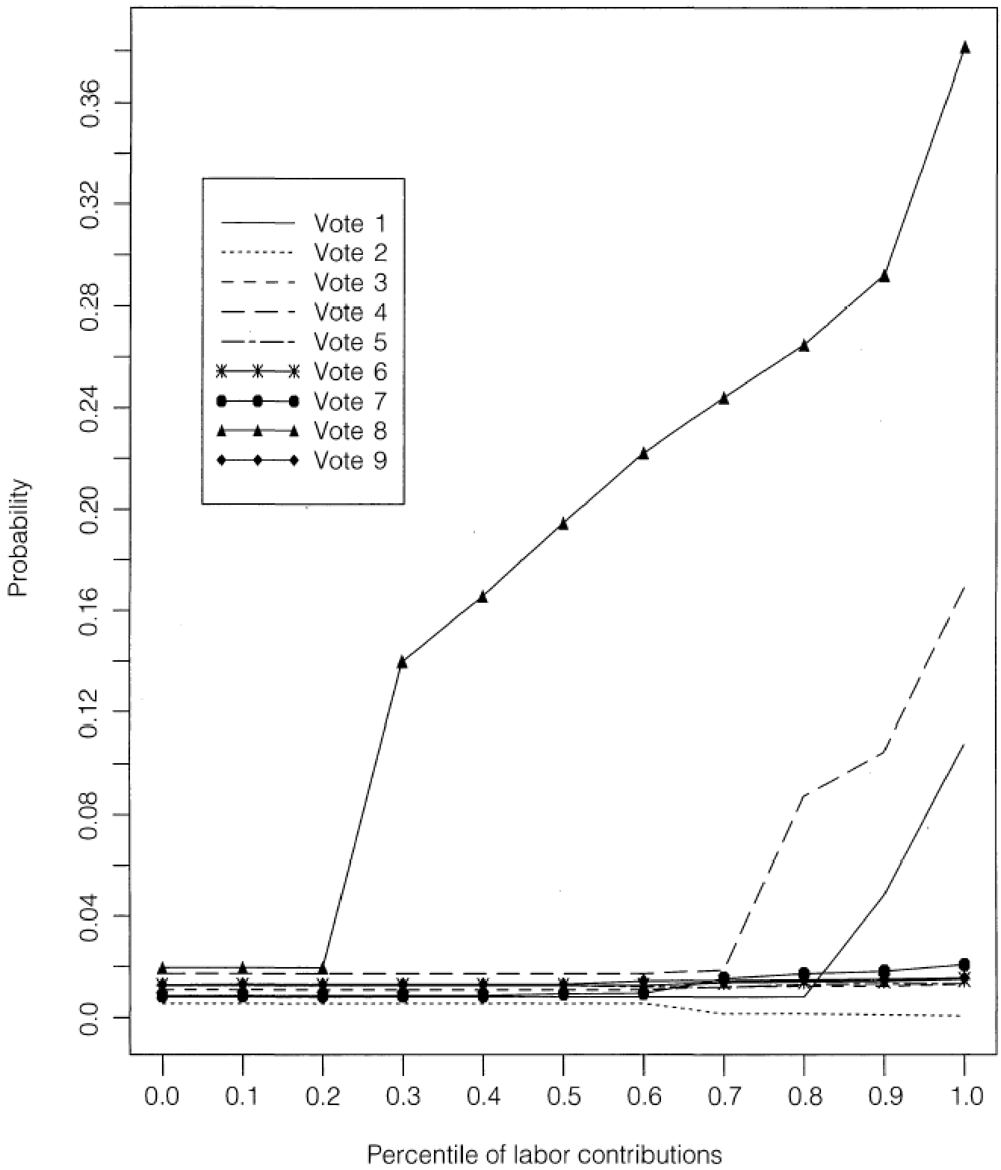
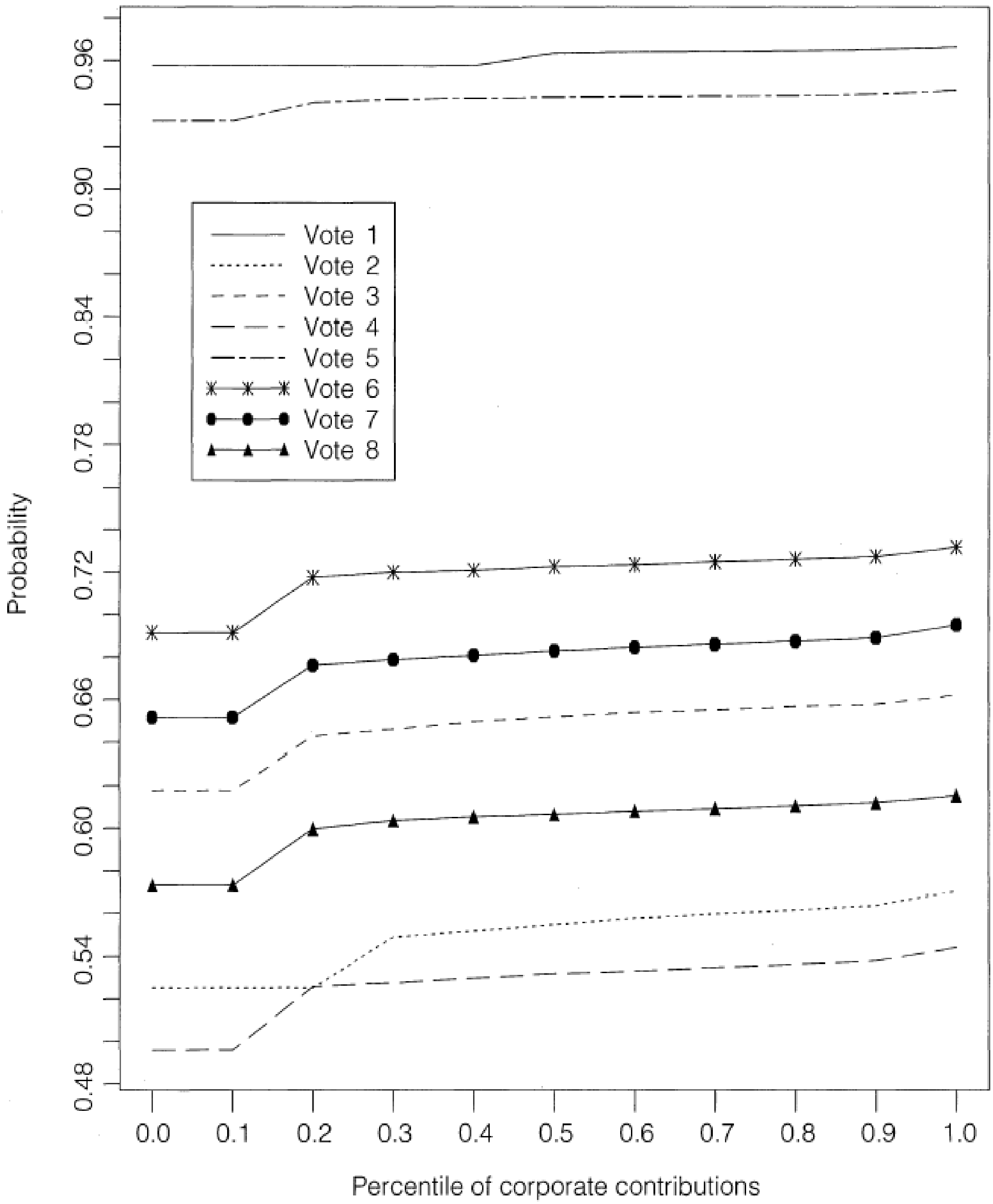


FIGURE 4 **Effect of Corporate PAC Contributions on the Probability of Voting in Favor of the USCC Positions, 104th Congress, 2nd Session**



Section 9

Binary Time-Series Cross-Section (BTSCS) Data

- The methods above are appropriate when N is large and T is small. Beck, Katz, and Tucker ('98 *AJPS*) derive a method for when T is large.

- The method is based on the observation that BTSCS data is identical to grouped duration data. That is, we get to observe whether an event occurred or not only after the end of some discrete period (e.g., a year).
- Thus, we can use duration methods to correct for the problem of temporal dependence.
- Start from the hazard rate for the continuous time Cox proportional hazard model:

$$\lambda(t) = \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)$$

- The survival function is given by

$$S(t) = \exp\left(-\int_0^t \lambda(\tau)d\tau\right)$$

- Assuming we get to observe only whether or not an event occurred between time $t_k - 1$ and t_k, we can write

$$
\begin{aligned}
\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\int_{t_k-1}^{t_k} \lambda_i(\tau)d\tau\right) \\
&= 1 - \exp\left(-\int_{t_k-1}^{t_k} \exp(\boldsymbol{\beta}'\mathbf{x}_{it})\lambda_0(t)d\tau\right) \\
&= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\int_{t_k-1}^{t_k} \lambda_0(t)d\tau\right)
\end{aligned}
$$

- Let

$$
\begin{aligned}
\alpha_{t_k} &= \int_{t_k-1}^{t_k} \lambda_0(t)d\tau \\
\kappa_{t_k} &= \ln(\alpha_{t_k})
\end{aligned}
$$

- Then

$$
\begin{aligned}
\Pr\left(y_{it_k} = 1\right) &= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it})\alpha_{t_k}\right) \\
&= 1 - \exp\left(-\exp(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t_k})\right)
\end{aligned}
$$

- This is a binary model with a complimentary log-log (cloglog) link. The cloglog link is identical to a logit link function when the probability of an event is small ($< 25\%$) and extremely similar when the probability of an event is moderate ($< 50\%$).

- For ease of application then, Beck, Katz, and Tucker recommend using the logistic analogue

$$\Pr\left(y_{it}=1|\mathbf{x}_{it}\right)=\frac{1}{1+\exp(-(\boldsymbol{\beta}'\mathbf{x}_{it}+\kappa_{t-t0}))}$$

where κ_{t-t0} is a dummy variable marking the length of the sequence of zeros that precede the current observation. For example,

t	1	2	3	4	5	6	7	8	9
y	0	0	0	1	0	1	1	0	0
κ	κ_1	κ_2	κ_3	κ_4	κ_1	κ_2	κ_1	κ_1	κ_2

- The intuition behind why ordinary logit is inadequate for BTSCS data is that it doesn't allow for a nonconstant baseline hazard.
- Including the κ dummies allows duration dependence by allowing for a time-varying baseline hazard.
- To see how the κ dummies are interpretable as baseline probabilities or hazards, note

$$\Pr\left(y_{it}=1|\mathbf{x}_{it}=0,t0\right)=\frac{1}{1+\exp(-\kappa_{t-t0})}$$

- The κ dummies are essentially time fixed effects that account for duration dependence. Thus when we estimate the model we need to create a matrix of dummies and concatenate it with the matrix of explanatory variables. For the example given above, this matrix would look like

$$\mathbf{K}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note there are 4 columns because the longest spell is 4 periods long.

$$\Pr\left(y_{it} = 1 | \mathbf{x}_{it}\right) = \frac{1}{1 + \exp(-(\boldsymbol{\beta}'\mathbf{x}_{it} + \kappa_{t-t0}))}$$

Section 10

Generalized Estimating Equations (GEEs)

- This class of models allows us to account for unobserved correlation among observations, without including unit-specific effects.
- GEEs relax assumptions about the independence of observations: observations are grouped into clusters and then parameters are estimated to model the correlations among observations in the cluster.
- This class of models is derived from the General Linear Model (GLM) approach.
  - Models in which one specifies the "link" function $E(Y_i) = \mu_i = h(\boldsymbol{\beta}'\mathbf{x}_i)$ and the relationship between the mean and the variance (e.g. $\mathbf{V}_i = g(\mu_i)$).
  - Standard GLMs obtain estimates by solving the "score equations":

$$\mathbf{U}_k(\beta) = \sum_{i=1}^{N} \mathbf{D}_i \mathbf{V}_i^{-1}(Y_i - \mu_i) = 0 \tag{10.1}$$

where $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$ and $\mathbf{V}_i$ is the variance matrix.

10.1 GLMs for Correlated Data

- Consider $E(Y_{it}) = \mu_{it} = g(\boldsymbol{\beta}'\mathbf{x}_{it})$, $T > 1$.
- Must make some provision for dependence within i, across t.
- Specify the conditional within-unit or within-cluster correlation:
  - Define the "working" $T \times T$ correlation matrix $\mathbf{R}_i(\alpha)$ as a function of α.
  - Structure (but not the elements) of $\mathbf{R}_i(\alpha)$ determined by the investigator.
  - Then redefine the variance matrix in Equation 10.1 as:

$$\mathbf{V}_i = \frac{(\mathbf{A}_i)^{\frac{1}{2}}\mathbf{R}_i(\alpha)(\mathbf{A}_i)^{\frac{1}{2}}}{\phi} \tag{10.2}$$

    where the $\mathbf{A}_i$ are $T \times T$ diagonal matrices w/ $g(\mu_{it})$ along the diagonal.

- Intuition:
  - Choose β so that μ_{it} is "close" to Y_{it} on average,
  - Optimally weight each residual $(Y_{it} - \mu_{it})$ by the inverse of $\text{cov}(Y_i)$.

10.2 Options for specifying within-cluster correlation

- "Working Independence": $\mathbf{R}_i(\alpha) = \mathrm{I}$
  - No within-unit correlation (completely pooling observations).
  - Just like standard logit/probit.
- "Exchangeable": $\mathbf{R}_i(\alpha) = \rho$
  - Within-unit correlation is the same across units.
  - Similar to "random effects," in the sense that we're assuming constant within-unit marginal covariance; but do not need orthogonality assumption of RE.
- Autoregressive : $\mathbf{R}_i(\alpha) = \rho^{|t-s|}$
  - Here, AR(1).
  - Correlation decays over "time."
  - Can also do "banded" / "stationary" correlations.
- "Unstructured" : $\mathbf{R}_i(\alpha) = \alpha_{st}, t \neq s$
  - $\boldsymbol{\alpha}$ is a $T \times T$ matrix.
  - Allows $\frac{T(T-1)}{2}$ unique pairwise correlations.
  - Very flexible, but can be hard to estimate, esp. w/ large T.

10.3 "Robust" Standard Errors

- $\hat{\beta}_{GEE}$ is robust to misspecification of $\mathbf{R}_i(\alpha)$.
- $\widehat{\text{var}(\hat{\beta}_{GEE})}$ is not.
- Can compute "sandwich" estimator: $\widehat{\text{var}(\hat{\beta}_{GEE})} =$

$$N\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{S}}_i\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)\left(\sum_{i=1}^{N}\hat{\mathbf{D}}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} \tag{10.3}$$

  where $\hat{\mathbf{S}}_i = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$.
- Similar to the Huber/White estimator.
- One drawback of GEE is that it does not produce standard errors for elements of $\mathbf{R}_i(\alpha) \Rightarrow$ we should be cautious about drawing inferences about these parameters (treated as "nuisance").
- A number of software packages can estimate GEEs: `xtgee` in `Stata`, `proc genmod` w/ repeated option in SAS, `gee, geepack` in S-Plus/R, `GEE` in Riemann Library for Gauss.

10.4 GEE2

- Other options exist for estimating the $m = \frac{T(T-1)}{2}$ elements of $\mathbf{R}_i(\alpha)$.
- Can do this with a separate estimating equation:

$$\mathbf{U}_{\mathrm{m}}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \mathbf{E}_i'\mathbf{W}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\eta}_i) \tag{10.4}$$

  where

  - $\mathbf{Z}_i' = (Z_{i12}, Z_{i13}, \ldots, Z_{i1T}, Z_{i23}, \ldots, Z_{iT,(T-1)})$; the $\frac{T(T-1)}{2}$ "observed" sample pairwise correlation.
  - $\boldsymbol{\eta}_i$ is the column vector of the expected values of the pairwise intracluster correlation for observation i.
  - $\mathbf{E}_i = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}_i}$
  - $\mathbf{W}_i$ is a square diagonal matrix of rank $\frac{T(T-1)}{2}$ containing the variances and covariances of the $\mathbf{Z}_i$s.

- Can be estimated either separately from $\mathbf{U}_k(\boldsymbol{\beta})$, assuming $\mathrm{cov}[U_k(\boldsymbol{\beta}), U_m(\alpha)] = 0$, or allowing the two to covary.
- Drawbacks: Requires correct specification of $\mathbf{R}_i(\alpha)$ for consistent estimates of $\hat{\boldsymbol{\beta}}$.

10.5 Application: Panel Decision-making on the Court of Appeals

- Question of interest: do gender and race affect the decisions of judges on the U.S. Court of Appeals?
- Almost all cases are decided by (more or less) randomly selected panels of three judges.
- “Norm of unanimity” suggests there will be significant correlation among judges on the same panel.
- GEE approach is useful for accounting for correlation, treating each three judge panel as a cluster.
- Farhang and Wawro (*JLEO* ’04) use this approach to examine decisions in employment discrimination cases.

Table 10.1: GEE analysis of judges' votes in Appeals Court decisions

	Logit		GEE	
Variable	Coefficient	Std. Err.	Coefficient	Std. Err.
Judge and Panel Level Variables				
Intercept	1.126	0.362	1.226	0.613
Gender	0.877	0.269	0.791	0.276
One female colleague (female judge)	−0.373	0.454	−0.398	0.594
One female colleague (male judge)	0.800	0.180	0.789	0.274
Two female colleagues (male judge)	−0.337	0.575	−0.362	0.595
Race	−0.329	0.363	−0.212	0.359
One nonwhite colleague (nonwhite judge)	−0.033	0.558	−0.458	0.764
One nonwhite colleague (white judge)	−0.136	0.236	−0.151	0.357
Two nonwhite colleagues (white judge)	−0.443	0.720	−0.859	1.081
NOMINATE score	−0.570	0.192	−0.643	0.176
Panel colleagues' NOMINATE scores	−1.098	0.243	−1.100	0.342
Author	0.122	0.188	−0.023	0.040
Gender × author	−0.174	0.421	0.054	0.120
Race × author	0.345	0.549	0.069	0.117
NOMINATE score × author	−0.173	0.338	0.069	0.082
Case Specific Variables				
Race discrimination	0.285	0.197	0.294	0.310
Gender discrimination	−0.190	0.185	−0.171	0.308
Harassment	0.761	0.206	0.717	0.354
Age discrimination	−0.062	0.190	−0.047	0.330
Religious discrimination	−0.752	0.624	−0.715	1.138
Nationality discrimination	0.069	0.468	0.045	0.784
Reverse gender discrimination	0.576	0.391	0.487	0.716
Reverse race discrimination	−1.282	0.609	−1.388	1.016
Government defendant	0.139	0.159	0.091	0.268
EEOC plaintiff	0.150	0.620	−0.083	1.423
Plaintiff appeal	−1.597	0.235	−1.570	0.390
Both appeal	0.578	0.282	0.586	0.486
Posture	−0.079	0.207	−0.093	0.316
Circuit Level Variables				
1st Circuit dummy	−1.048	0.409	−1.102	0.641
2nd Circuit dummy	−0.624	0.373	−0.636	0.650
3rd Circuit dummy	−1.358	0.549	−1.386	0.773
4th Circuit dummy	−1.885	0.578	−1.973	0.839
5th Circuit dummy	−2.053	0.400	−2.038	0.688
6th Circuit dummy	0.093	0.397	−0.023	0.703
7th Circuit dummy	−1.112	0.320	−1.156	0.541
8th Circuit dummy	−0.870	0.326	−0.923	0.564
10th Circuit dummy	−1.315	0.377	−1.344	0.657
11th Circuit dummy	−0.207	0.334	−0.252	0.583
D.C. Circuit dummy	−1.410	0.489	−1.350	0.794
$\hat{\rho}$	—	—	0.893	—

Note: Table entries are GEE estimates. $N = 1200$.