Discrete Choice Modeling

Panel Data Binary Choice Models

0 Introduction
1 Summary
2 Binary Choice
3 Panel Data
4 Bivariate Probit
5 Ordered Choice
6 Count Data
7 Multinomial Choice
8 Nested Logit
9 Heterogeneity
10 Latent Class
11 Mixed Logit
12 Stated Preference
13 Hybrid Choice

William Greene
Stern School of Business
New York University
The British Household Panel Survey began in 1991 and is a multi-purpose study whose unique value resides in the fact that:

- it follows the same representative sample of individuals – the panel – over a period of years;
- it is household-based, interviewing every adult member of sampled households;
- it contains sufficient cases for meaningful analysis of certain groups such as the elderly or lone parent families.

The wave 1 panel consists of some 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. Additional samples of 1,500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2,000 households was added in Northern Ireland, making the panel suitable for UK-wide research.

- BHPS wave 18 data and documentation are available from the UK Data Archive.
HILDA Survey

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is a household-based panel study which began in 2001. It has the following key features:

- It collects information about economic and subjective well-being, labour market dynamics and family dynamics.
- Special questionnaire modules are included each wave.
- The wave 1 panel consisted of 7,682 households and 19,914 individuals. In wave 11 this was topped up with an additional 2,153 households and 5,477 individuals.
- Interviews are conducted annually with all adult members of each household.
- The panel members are followed over time.
- The funding has been guaranteed for sixteen waves, though the survey is designed to continue for longer than this.
- Academic and other researchers can apply to use the General Release datasets for their research.
About SOEP

The SOEP Service Group

SOEP Quicklinks:
- SOEPinfo
- SOEPlit
- SOEPmonitor
- SOEPdata Documents
- SOEPdata FAQ

About SOEP:

Team
Contact
SOEP-Overview
Mission
SOEP Survey Committee

---

Short Description

The German Socio-Economic Panel Study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin. Every year, there were nearly 11,000 households, and more than 20,000 persons sampled by the fieldwork organization TNS Infratest Sozialforschung.

The data provide information on all household members, consisting of Germans living in the Old and New German States, Foreigners, and recent Immigrants to Germany. The Panel was started in 1984.

Some of the many topics include household composition, occupational biographies, employment, earnings, health, and satisfaction indicators.
The Panel Study of Income Dynamics - PSID - is the longest running longitudinal household survey in the world.

The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics. The PSID is directed by faculty at the University of Michigan, and the data are available on this website without cost to researchers and analysts.

The data are used by researchers, policy analysts, and teachers around the globe. Over 3,000 peer-reviewed publications have been based on the PSID. Recognizing the importance of the data, numerous countries have created their own PSID-like studies that now facilitate cross-national comparative research. The National Science Foundation recognized the PSID as one of the 60 most significant advances funded by NSF in its 60 year history.
Discrete Choice Modeling
Panel Data Binary Choice Models

Survey of Income and Program Participation

SIPP

(Formerly, DEWS)

URL: http://www.census.gov/sipp/

Source: U.S. Census Bureau, Demographic Survey Division, Survey of Income and Program Participation Branch
Created: February 14, 2002
Last revised: January 2, 2009

Measuring America—People, Places, and Our Economy
Discrete Choice Modeling
Panel Data Binary Choice Models

European Community Household Panel (ECHP)

Description of dataset

The European Community Household Panel (ECHP) is a panel survey in which a sample of households and persons have been interviewed year after year.

These interviews cover a wide range of topics concerning living conditions. They include detailed income information, financial situation in a wider sense, working life, housing situation, social relations, health and biographical information of the interviewed.

The total duration of the ECHP was 8 years, running from 1994-2001 (8 waves).

ECHP based data in the database

99% of the "income and living conditions" domain under theme "Population and social conditions" is derived from ECHP. This includes many indicators of relative monetary poverty and of income inequality, analysed in different ways (eg. different cut-off thresholds, by age, gender, activity status, tenure status...).

It also includes a selection of indicators of social exclusion and non-monetary deprivation derived from ECHP, notably on housing.

Of these, 4 have been chosen as structural indicators, namely the at-risk-of-poverty rate before cash social transfers, the persistent at-risk-of-poverty rate and the s80/s20 income quintile share ratio. The at-risk-of-poverty rate after social transfers is a headline indicator.

A selection of indicators in the "health status" and "health care" collections of the "public health" domain also under the above-mentioned same theme are derived from ECHP as well.
National Longitudinal Surveys

The National Longitudinal Surveys (NLS) are a set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women. For more than 4 decades, NLS data have served as an important tool for economists, sociologists, and other researchers.

On This Page
- NLS General Overviews
- NLS News Releases
- NLS Tables
- NLS Data
- NLS Publications
- NLS FAQs
- NLS Related Links
- Contact NLS

NLS General Overviews

- National Longitudinal Survey of Youth 1997 (NLSY97) -- Survey of young men and women born in the years 1980-84; respondents were ages 12-17 when first interviewed in 1997.
- National Longitudinal Survey of Youth 1979 (NLSY79) -- Survey of men and women born in the years 1957-64; respondents were ages 14-22 when first interviewed in 1979.
- NLSY79 Children and Young Adults -- Survey of the biological children of women in the NLSY79.
- National Longitudinal Surveys of Young Women and Mature Women (NSW) -- The Young Women's survey includes women who were ages 14-24 when first interviewed in 1968. The Mature Women's survey includes women who were ages 30-44 when first interviewed in 1967. These surveys were discontinued in 2003.
- National Longitudinal Surveys of Young Men and Older Men -- The Young Men's survey, which was discontinued in 1981, includes men who were ages 14-24 when first interviewed in 1966. The Older Men's survey, which was discontinued in 1990, includes men who were ages 45-59 when first interviewed in 1966.
Overview

The annual Agricultural Resource Management Survey (ARMS) is USDA's primary source of information on the financial condition, production practices, and resource use of America's farm businesses and the economic well-being of America's farm households. ARMS data are essential to USDA, congressional, administration, and industry decision makers when weighing alternative policies and programs that touch the farm sector or affect farm families.

Sponsored jointly by ERS and the National Agricultural Statistics Service (NASS), ARMS is the only national survey that provides observations of field-level farm practices, the economics of the farm businesses operating the field (or dairy herd, green house, nursery, poultry house, etc.), and the characteristics of farm operators and their households (age, education, occupation, farm and off-farm work, types of employment, family living expenses, etc.)—all collected in a representative sample. Information about crop production, farm production, business, and households includes data for selected surveyed States where available. See more background on ARMS...
Application: Health Care Panel Data

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods
Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. ( Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987).

Variables in the file are

- **DOCTOR** = 1(Number of doctor visits > 0)
- **HOSPITAL** = 1(Number of hospital visits > 0)
- **HSAT** = health satisfaction, coded 0 (low) - 10 (high)
- **DOCVIS** = number of doctor visits in last three months
- **HOSPVIS** = number of hospital visits in last calendar year
- **PUBLIC** = insured in public health insurance = 1; otherwise = 0
- **ADDON** = insured by add-on insurance = 1; otherwise = 0
- **HHNINC** = household nominal monthly net income in German marks / 10000.
  (4 observations with income=0 were dropped)
- **HHKIDS** = children under age 16 in the household = 1; otherwise = 0
- **EDUC** = years of schooling
- **AGE** = age in years
- **MARRIED** = marital status
Unbalanced Panels

Most theoretical results are for balanced panels.

Most real world panels are unbalanced.

Often the gaps are caused by attrition.

The major question is whether the gaps are ‘missing completely at random.’ If not, the observation mechanism is endogenous, and at least some methods will produce questionable results.

Researchers rarely have any reason to treat the data as nonrandomly sampled. (This is good news.)
Unbalanced Panels and Attrition ‘Bias’

  - Variable addition test using covariates of presence in the panel
  - Nonconstructive – what to do next?

  - Stringent assumptions about the process
  - Model based on probability of being present in each wave of the panel
Panel Data Binary Choice Models

Random Utility Model for Binary Choice

\[ U_{it} = \alpha + \beta'x_{it} + \varepsilon_{it} + \text{Person i specific effect} \]

Fixed effects using “dummy” variables

\[ U_{it} = \alpha_i + \beta'x_{it} + \varepsilon_{it} \]

Random effects using omitted heterogeneity

\[ U_{it} = \alpha + \beta'x_{it} + \varepsilon_{it} + u_i \]

Same outcome mechanism: \[ Y_{it} = 1[U_{it} > 0] \]
Pooled Model
Ignoring Unobserved Heterogeneity

Assuming strict exogeneity; \( \text{Cov}(x_{it}, u_i + \varepsilon_{it}) = 0 \)

\[ y_{it}^* = x_{it}'\beta + u_i + \varepsilon_{it} \]

\[ \text{Prob}[y_{it} = 1 | x_{it}] = \text{Prob}[u_i + \varepsilon_{it} > -x_{it}'\beta] \]

Using the same model format:

\[ \text{Prob}[y_{it} = 1 | x_{it}] = F\left(x_{it}'\beta / \sqrt{1+\sigma_u^2}\right) = F(x_{it}'\delta) \]

This is the 'population averaged model.'
Ignoring Heterogeneity in the RE Model

Ignoring heterogeneity, we estimate $\delta$ not $\beta$. Partial effects are $\delta f(x_{it}\delta)$ not $\beta f(x_{it}\beta)$.

$\beta$ is underestimated, but $f(x_{it}\beta)$ is overestimated. Which way does it go? Maybe ignoring $u$ is ok?
Not if we want to compute probabilities or do statistical inference about $\beta$. Estimated standard errors will be too small.
Ignoring Heterogeneity (Broadly)

- Presence will generally make parameter estimates look smaller than they would otherwise.
- Ignoring heterogeneity will definitely distort standard errors.
- Partial effects based on the parametric model may not be affected very much.
- Is the pooled estimator ‘robust?’ Less so than in the linear model case.
## Pooled vs. RE Panel Estimator

### Binomial Probit Model

**Dependent variable**

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|----------|---------|-----------|
| Constant | .02159      | .05307         | .407     | .6842   |
| AGE      | .01532***   | .00071         | 21.695   | .0000   | 43.5257   |
| EDUC     | -.02793***  | .00348         | -8.023   | .0000   | 11.3206   |
| HHNINC   | -.10204**   | .04544         | -2.246   | .0247   | .35208    |

Unbalanced panel has 7293 individuals

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|----------|---------|-----------|
| Constant | -.11819     | .09280         | -1.273   | .2028   |
| AGE      | .02232***   | .00123         | 18.145   | .0000   | 43.5257   |
| EDUC     | -.03307***  | .00627         | -5.276   | .0000   | 11.3206   |
| HHNINC   | .00660      | .06587         | .100     | .9202   | .35208    |
| Rho      | .44990***   | .01020         | 44.101   | .0000   |           |
Partial Effects

Partial derivatives of $E[y] = F[*]$ with respect to the vector of characteristics. They are computed at the means of the Xs. Observations used for means are All Obs.

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Elasticity |
|----------|-------------|----------------|---------|----------|------------|
| AGE      | .00578***   | .00027         | 21.720  | .0000    | .39801     |
| EDUC     | -.01053***  | .00131         | -8.024  | .0000    | -.18870    |
| HHNINC   | -.03847**   | .01713         | -2.246  | .0247    | -.02144    |

<table>
<thead>
<tr>
<th>Based on the panel data estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>EDUC</td>
</tr>
<tr>
<td>HHNINC</td>
</tr>
</tbody>
</table>
Effect of Clustering

- $Y_{it}$ must be correlated with $Y_{is}$ across periods
- Pooled estimator ignores correlation
- Broadly, $y_{it} = E[y_{it}|x_{it}] + w_{it}$,
  - $E[y_{it}|x_{it}] = \text{Prob}(y_{it} = 1|x_{it})$
  - $w_{it}$ is correlated across periods
- Assuming the marginal probability is the same, the pooled estimator is consistent. (We just saw that it might not be.)
- Ignoring the correlation across periods generally leads to underestimating standard errors.
‘Cluster’ Corrected Covariance Matrix

\[ C = \text{the number of clusters} \]
\[ n_c = \text{number of observations in cluster } c \]
\[ H^{-1} = \text{negative inverse of second derivatives matrix} \]
\[ g_{ic} = \text{derivative of log density for observation} \]

\[ V = H^{-1} \left( \frac{C}{C - 1} \right) \left( \sum_{c=1}^{C} \left( \sum_{i=1}^{n_c} g_{ic} \right) \left( \sum_{i=1}^{n_c} g'_{ic} \right) \right) H^{-1} \]
## Cluster Correction: Doctor

**Binomial Probit Model**

**Dependent variable**: DOCTOR

**Log likelihood function**: -17457.21899

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| Constant | -.25597***  | .05481         | -4.670  | .0000   | 43.5257   |
| AGE      | .01469***   | .00071         | 20.686  | .0000   | 43.5257   |
| EDUC     | -.01523***  | .00355         | -4.289  | .0000   | 11.3206   |
| HHNINC   | -.10914**   | .04569         | -2.389  | .0169   | .35208    |
| FEMALE   | .35209***   | .01598         | 22.027  | .0000   | .47877    |

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| Constant | -.25597***  | .07744         | -3.305  | .0009   | 43.5257   |
| AGE      | .01469***   | .00098         | 15.065  | .0000   | 43.5257   |
| EDUC     | -.01523***  | .00504         | -3.023  | .0025   | 11.3206   |
| HHNINC   | -.10914*    | .05645         | -1.933  | .0532   | .35208    |
| FEMALE   | .35209***   | .02290         | 15.372  | .0000   | .47877    |
Random Effects
Quadrature – Butler and Moffitt (1982)

This method is used in most commercial software since 1982

$$\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} F(y_{it}, \alpha + \beta' x_{it} + \sigma_u v_i) \right] \phi(v_i) dv_i$$

$$= \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} g(v) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{v^2}{2} \right) dv_i$$

(make a change of variable to $w = v/\sqrt{2}$)

$$= \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} g(\sqrt{2}w) \exp(-w^2) dw_i$$

The integral can be computed using Hermite quadrature.

$$\approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} \log \sum_{h=1}^{H} w_h g(\sqrt{2}z_h)$$

The values of $w_h$ (weights) and $z_h$ (nodes) are found in published tables such as Abramovitz and Stegun (or on the web). $H$ is by choice. Higher $H$ produces greater accuracy (but takes longer).

\[ u_i \sim N[0, \sigma_u^2] \]

\[ = \sigma_u v_i \]

where $v_i \sim N[0,1]$
Quadrature Log Likelihood

After all the substitutions, the function to be maximized:
Not simple, but feasible.

\[
\log L = \sum_{i=1}^{N} \log \left( \frac{1}{\sqrt{\pi}} \right) \sum_{h=1}^{H} \frac{1}{\sqrt{\pi}} \prod_{t=1}^{T_i} F(y_{it}, \alpha + \beta' x_{it} + (\sigma_u \sqrt{2}) z_h) \\
= \sum_{i=1}^{N} \log \left( \frac{1}{\sqrt{\pi}} \right) \sum_{h=1}^{H} \prod_{t=1}^{T_i} F(y_{it}, \alpha + \beta' x_{it} + \theta z_h)
\]
Simulation Based Estimator

\[
\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \left[ \prod_{t=1}^{T_i} F(y_{it}, \alpha + \beta' x_{it} + \sigma_u v_{ir}) \right] \phi (v_i) \, dv_i \\
= \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} g(v_i) \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{v_i^2}{2} \right) \, dv_i
\]

This equals \( \sum_{i=1}^{N} \log E[g(v_i)] \)

The expected value of the function of \( v_i \) can be approximated by drawing \( R \) random draws \( v_{ir} \) from the population \( N[0,1] \) and averaging the \( R \) functions of \( v_{ir} \). We maximize

\[
\log L_s = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \left[ \prod_{t=1}^{T_i} F(y_{it}, \alpha + \beta' x_{it} + \sigma_u v_{ir}) \right]
\]
## Random Effects Model: Quadrature

Random Effects Binary Probit Model

Dependent variable: DOCTOR

Log likelihood function: -16290.72192  \( \Rightarrow \) Random Effects

Restricted log likelihood: -17701.08500  \( \Rightarrow \) Pooled

Chi squared [1 d.f.]: 2820.72616

Estimation based on N = 27326, K = 5

Unbalanced panel has 7293 individuals

| Variable | Coefficient | Standard Error | b/St.Er. | P[|Z|>z] | Mean of X |
|----------|-------------|----------------|---------|---------|-----------|
| Constant | -0.11819    | 0.09280        | -1.273  | .2028   |           |
| AGE      | 0.02232***  | 0.00123        | 18.145  | .0000   | 43.5257   |
| EDUC     | -0.03307*** | 0.00627        | -5.276  | .0000   | 11.3206   |
| HHNINC   | 0.00660     | 0.06587        | .100    | .9202   | .35208    |
| Rho      | 0.44990***  | 0.01020        | 44.101  | .0000   |           |

<table>
<thead>
<tr>
<th>Pooled Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>EDUC</td>
</tr>
<tr>
<td>HHNINC</td>
</tr>
</tbody>
</table>
Random Parameter Model

Random Coefficients Probit Model

Dependent variable DOCTOR (Quadrature Based)

Log likelihood function -16296.68110 (-16290.72192)
Restricted log likelihood -17701.08500
Chi squared [1 d.f.] 2808.80780
Simulation based on 50 Halton draws

|Variable| Coefficient| Standard Error| b/St.Er.| P[|Z|>z]|
|---------|------------|---------------|--------|--------|
|Nonrandom parameters| | | | |
|AGE| .02226***| .00081| 27.365| .0000 ( .02232) |
|EDUC| -.03285***| .00391| -8.407| .0000 (-.03307) |
|HHNINC| .00673| .05105| .132| .8952 (.00660) |
|Means for random parameters| | | | |
|Constant| -.11873**| .05950| -1.995| .0460 (-.11819) |
|Scale parameters for dists. of random parameters| | | | |
|Constant| .90453***| .01128| 80.180| .0000 |

Using quadrature, a = -.11819. Implied ρ from these estimates is .90454^2/(1+.90453^2) = .449998 compared to .44990 using quadrature.
A Dynamic Model

\[ y_{it} = 1[\mathbf{x}_{it}' \beta + \gamma y_{i,t-1} + \varepsilon_{it} + u_i > 0] \]

Two similar 'effects'

Unobserved heterogeneity

State dependence = state 'persistence'

\[ \Pr(y_{it} = 1 \mid y_{i,t-1}, \ldots, y_{i0}, x_{it}, u] = F[\mathbf{x}_{it}' \beta + \gamma y_{i,t-1} + u_i] \]

How to estimate \( \beta, \gamma, \) marginal effects, \( F(.) \), etc?

(1) Deal with the latent common effect
(2) Handle the lagged effects:

This encounters the initial conditions problem.
Dynamic Probit Model: A Standard Approach

(1) Conditioned on all effects, joint probability

\[ P(y_{i1}, y_{i2}, ..., y_{iT} | y_{i0}, x_i, u_i) = \prod_{t=1}^{T} F(x'_{it}\beta + \gamma y_{i,t-1} + u_i, y_{it}) \]

(2) Unconditional density; integrate out the common effect

\[ P(y_{i1}, y_{i2}, ..., y_{iT} | y_{i0}, x_i) = \int_{-\infty}^{\infty} P(y_{i1}, y_{i2}, ..., y_{iT} | y_{i0}, x_i, u_i) h(u_i | y_{i0}, x_i) du_i \]

(3) Density for heterogeneity

\[ h(u_i | y_{i0}, x_i) = N[\alpha + \theta y_{i0} + x'_i\delta, \sigma_u^2], x_i = [x_{i1}, x_{i2}, ..., x_{iT}], \text{ so} \]

\[ u_i = \alpha + \theta y_{i0} + x'_i\delta + \sigma_u w_i \]

(4) Reduced form

\[ P(y_{i1}, y_{i2}, ..., y_{iT} | y_{i0}, x_i) = \int_{-\infty}^{\infty} \prod_{t=1}^{T} F(x'_{it}\beta + \gamma y_{i,t-1} + \alpha + \theta y_{i0} + x'_i\delta + \sigma_u w_i, y_{it}) h(w_i) dw_i \]

This is a random effects model.
Simplified Dynamic Model

Projecting $u_i$ on all observations expands the model enormously.

(3) Projection of heterogeneity only on group means

$$h(u_i \mid y_{i0}, x_i) = N[\alpha + \theta y_{i0} + \bar{x}_i \delta, \sigma_u^2] \quad \text{so} \quad u_i = \alpha + \theta y_{i0} + \bar{x}_i \delta + w_i$$

(4) Reduced form

$$P(y_{i1}, y_{i2}, ..., y_{iT} \mid y_{i0}, x_i) =$$

$$\int_{-\infty}^{\infty} \prod_{t=1}^{T} F(\alpha + x_{it}' \beta + \gamma y_{i,t-1} + \theta y_{i0} + \bar{x}_i' \delta + \sigma_u w_i, y_{it})h(w_i)dw_i$$

Mundlak style correction with the initial value in the equation. This is (again) a random effects model.
A Dynamic Model for Public Insurance

setpanel ; group =id:pds=ti$
create ; obs =ndx(id,1)$
namelist ; xit =age,income,hhkids,hsat$
create ; agebar=0;incbar=0;kidsbar=0;hsatbar=0$
namelist ; means =agebar,incbar,kidsbar,hsatbar$
create ; means =groupmean(xit,obs=pds=ti)$
create ; yi0 =groupobs1(public,obs=pds=ti)$
probit ; if[obs > 1];Panel ; lhs=public
;rhs=xit.means,one,yi0.public[-1]$

Add initial value, lagged value, group means
Dynamic Common Effects Model

Random Effects Binary Probit Model
Dependent variable PUBLIC
Log likelihood function -2588.02882
Restricted log likelihood -2696.91167
Chi squared [ 1](P= .000) 217.76570
Significance level .00000

(Cannot compute pseudo R2. Use RHS=one to obtain the required restricted logL)
Estimation based on N = 20033, K = 12
Inf.Cr.AIC = 5200.1 AIC/N = .260
Unbalanced panel has 5768 individuals
- ChiSqd[1] tests for random effects -
LM ChiSqd 111.854 P value .00000
LR ChiSqd 217.766 P value .00000
Wald ChiSqd 474.563 P value .00000

<table>
<thead>
<tr>
<th>PUBLIC</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>z</th>
<th>Prob.</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGE</td>
<td>-.01568</td>
<td>.01212</td>
<td>-1.29</td>
<td>.1957</td>
<td>-.03944 - .00808</td>
</tr>
<tr>
<td>INCOME</td>
<td>-.67330***</td>
<td>.25494</td>
<td>-2.64</td>
<td>.003</td>
<td>-.17305 - -.17371</td>
</tr>
<tr>
<td>MHKIDS</td>
<td>.01281</td>
<td>.11641</td>
<td>-1.11</td>
<td>.9124</td>
<td>-.24095 - .21534</td>
</tr>
<tr>
<td>HSAT</td>
<td>.00336</td>
<td>.02000</td>
<td>-1.7</td>
<td>.8666</td>
<td>-.04625 - .03854</td>
</tr>
<tr>
<td>AGEBAR</td>
<td>.04392***</td>
<td>.01258</td>
<td>3.49</td>
<td>.0005</td>
<td>.01926 - .06858</td>
</tr>
<tr>
<td>INCBAR</td>
<td>-1.71950***</td>
<td>.38878</td>
<td>-4.42</td>
<td>.0000</td>
<td>-2.48149 - -.95751</td>
</tr>
<tr>
<td>KIDSBAR</td>
<td>.26462*</td>
<td>.15011</td>
<td>1.76</td>
<td>.0779</td>
<td>-.02959 - .55883</td>
</tr>
<tr>
<td>HSATBAR</td>
<td>-.05228</td>
<td>.03223</td>
<td>-1.62</td>
<td>.1048</td>
<td>-.11545 - .01089</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.57400***</td>
<td>.31443</td>
<td>-5.01</td>
<td>.0000</td>
<td>-2.19038 - -.95762</td>
</tr>
</tbody>
</table>

Y10 4.02429*** .28588 14.08 .0000 3.46398 4.58460
YLAG .95309*** .09358 10.18 .0000 .76967 1.13650
Rho  .66459*** .03143 21.78 .0000 .62300 .74618

***, **, * => Significance at 1%, 5%, 10% level.
Fixed Effects
Fixed Effects Models

- Estimate with dummy variable coefficients
  \[ U_{it} = \alpha_i + \beta'x_{it} + \varepsilon_{it} \]

- Can be done by “brute force” for 10,000s of individuals

\[
\log L = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log F(y_{it}, \alpha_i + \beta'x_{it})
\]

- \( F(.) = \) appropriate probability for the observed outcome

- Compute \( \beta \) and \( \alpha_i \) for \( i=1,\ldots,N \) (may be large)

- See FixedEffects.pdf in course materials.
Unconditional Estimation

- Maximize the whole log likelihood

- Difficult! Many (thousands) of parameters.

- Feasible – NLOGIT (2001) (“Brute force”)
### Fixed Effects Health Model

Groups in which \( y_{it} \) is always = 0 or always = 1. Cannot compute \( \alpha_i \).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-.0649</td>
<td>.0045</td>
<td>-14.418</td>
<td>.0000</td>
<td>-.0232</td>
<td>.0008</td>
<td>-28.991</td>
<td>.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>AGE</td>
<td>.0027</td>
<td>.0506</td>
<td>.054</td>
<td>.9570</td>
<td>.0573</td>
<td>.0037</td>
<td>15.467</td>
<td>.0000</td>
<td>43.5257</td>
</tr>
<tr>
<td>EDUC</td>
<td>.3530</td>
<td>.1161</td>
<td>3.040</td>
<td>.0024</td>
<td>.3425</td>
<td>.0481</td>
<td>7.118</td>
<td>.0000</td>
<td>11.3206</td>
</tr>
<tr>
<td>INCOME</td>
<td>-.0609</td>
<td>.0666</td>
<td>-.915</td>
<td>.3600</td>
<td>.0129</td>
<td>.0206</td>
<td>.627</td>
<td>.5307</td>
<td>.75862</td>
</tr>
<tr>
<td>MARRIED</td>
<td>-.0118</td>
<td>.0475</td>
<td>-.249</td>
<td>.8032</td>
<td>.0666</td>
<td>.0186</td>
<td>3.581</td>
<td>.0003</td>
<td>.40273</td>
</tr>
<tr>
<td>KIDS</td>
<td>-.0248</td>
<td>.0049</td>
<td>-5.087</td>
<td>.0000</td>
<td>-.0089</td>
<td>.0003</td>
<td>-29.012</td>
<td>.0000</td>
<td>43.5257</td>
</tr>
<tr>
<td>Partial Effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>.0010</td>
<td>.0192</td>
<td>.054</td>
<td>.9567</td>
<td>.0219</td>
<td>.0014</td>
<td>15.478</td>
<td>.0000</td>
<td>11.3206</td>
</tr>
<tr>
<td>EDUC</td>
<td>.1349</td>
<td>.0515</td>
<td>2.617</td>
<td>.0089</td>
<td>.1309</td>
<td>.0184</td>
<td>7.118</td>
<td>.0000</td>
<td>.35208</td>
</tr>
<tr>
<td>INCOME</td>
<td>-.0233</td>
<td>.0010</td>
<td>-22.562</td>
<td>.0000</td>
<td>.0049</td>
<td>.0079</td>
<td>.626</td>
<td>.5311</td>
<td>.75862</td>
</tr>
<tr>
<td>MARRIED</td>
<td>-.0045</td>
<td>.0004</td>
<td>-10.792</td>
<td>.0000</td>
<td>.0254</td>
<td>.0071</td>
<td>3.589</td>
<td>.0003</td>
<td>.40273</td>
</tr>
</tbody>
</table>

### LogLikelihood Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects</td>
<td>-8500.704</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogLR</td>
<td>-17365.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pooled</td>
<td>-18279.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7293 Individuals
3289 Individuals Bypassed
Conditional Estimation

- Principle: $f(y_{i1}, y_{i2}, \ldots | \text{some statistic})$ is free of the fixed effects for some models.
- Maximize the conditional log likelihood, given the statistic.
- Can estimate $\beta$ without having to estimate $\alpha_i$.
- Only feasible for the logit model. (Poisson and a few other continuous variable models. No other discrete choice models.)
Binary Logit Conditional Probabilities

\[
\text{Prob}(y_{it} = 1 | x_{it}) = \frac{e^{\alpha_i + x'_i \beta}}{1 + e^{\alpha_i + x'_i \beta}}.
\]

\[
\text{Prob}\left( Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \Bigg| \sum_{t=1}^{T_i} y_{it} \right) = \frac{\exp\left( \sum_{t=1}^{T_i} y_{it} x'_{it} \beta \right)}{\sum \exp\left( \sum_{t=1}^{T_i} d_{it} x'_{it} \beta \right)}.
\]

Denominator is summed over all the different combinations of \( T_i \) values of \( y_{it} \) that sum to the same sum as the observed \( \sum_{t=1}^{T_i} y_{it} \). If \( S_i \) is this sum, there are \( \binom{T}{S_i} \) terms. May be a huge number. An algorithm by Krailo and Pike makes it simple.
Example: Two Period Binary Logit

\[
\text{Prob}(y_{it} = 1 \mid x_{it}) = \frac{e^{\alpha_i + x_{it}'\beta}}{1 + e^{\alpha_i + x_{it}'\beta}}.
\]

\[
\text{Prob}\left\{ Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iT_i} = y_{iT_i} \right\} = \frac{\exp\left(\sum_{t=1}^{T_i} y_{it} x_{it}'\beta\right)}{\sum_{i=1}^{\sum_{i=S_i}^{T_i}} \exp\left(\sum_{t=1}^{T_i} d_{it} x_{it}'\beta\right)}.
\]

<table>
<thead>
<tr>
<th>Prob</th>
<th>( Y_{i1} = 0, Y_{i2} = 0 )</th>
<th>( \sum_{t=1}^{2} y_{it} = 0, \text{data} )</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob</td>
<td>( Y_{i1} = 1, Y_{i2} = 0 )</td>
<td>( \sum_{t=1}^{2} y_{it} = 1, \text{data} )</td>
<td>( \frac{\exp(x_{i1}'\beta)}{\exp(x_{i1}'\beta) + \exp(x_{i2}'\beta)} )</td>
</tr>
<tr>
<td>Prob</td>
<td>( Y_{i1} = 0, Y_{i2} = 1 )</td>
<td>( \sum_{t=1}^{2} y_{it} = 1, \text{data} )</td>
<td>( \frac{\exp(x_{i2}'\beta)}{\exp(x_{i1}'\beta) + \exp(x_{i2}'\beta)} )</td>
</tr>
<tr>
<td>Prob</td>
<td>( Y_{i1} = 1, Y_{i2} = 1 )</td>
<td>( \sum_{t=1}^{2} y_{it} = 2, \text{data} )</td>
<td>1</td>
</tr>
</tbody>
</table>
Estimating Partial Effects

“The fixed effects logit estimator of $\beta$ immediately gives us the effect of each element of $x_i$ on the log-odds ratio… Unfortunately, we cannot estimate the partial effects… unless we plug in a value for $\alpha_i$. Because the distribution of $\alpha_i$ is unrestricted – in particular, $E[\alpha_i]$ is not necessarily zero – it is hard to know what to plug in for $\alpha_i$. In addition, we cannot estimate average partial effects, as doing so would require finding $E[\Lambda(x_{it} \beta + \alpha_i)]$, a task that apparently requires specifying a distribution for $\alpha_i$.”

(Wooldridge, 2010)
Logit Constant Terms

Step 1. Estimate $\boldsymbol{\beta}$ with Chamberlain's conditional estimator

Step 2. Treating $\boldsymbol{\beta}$ as if it were known, estimate $\alpha_i$ from the first order condition

$$
\bar{y}_i = \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{e^{\alpha_i e^{x_{it}' \hat{\beta}}}}{1 + e^{\alpha_i e^{x_{it}' \hat{\beta}}}} = \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{\delta_i c_{it}}{1 + \delta_i c_{it}} = \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{c_{it}}{\mu_i + c_{it}}
$$

Estimate $\mu_i = 1 / \exp(\alpha_i) \Rightarrow \alpha_i = -\log \mu_i$

c_{it} = \exp(x_{it}' \hat{\beta})$ is treated as known data.

Solve one equation in one unknown for each $\alpha_i$.

Note there is no solution if $\bar{y}_i = 0$ or $1$.

Iterating back and forth does not maximize logL.
### Fixed Effects Logit Health Model: Conditional vs. Unconditional

#### Table 2.14 Estimated Fixed Effects Logit Models

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unconditional Estimator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogL</td>
<td>-8506.164</td>
<td>LogLR</td>
<td>-17365.15</td>
<td>7293 Individuals</td>
<td>3289 Individuals Bypassed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conditional Estimator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LogL</td>
<td>-5669.541</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mean of X**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Partial Effects</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>-.0259</td>
<td>.0063</td>
<td>-4.102</td>
<td>.0000</td>
<td>-.0012</td>
<td>.0009</td>
<td>-13.961</td>
<td>.0000</td>
<td>43.5257</td>
</tr>
<tr>
<td>EDUC</td>
<td>.0021</td>
<td>.0193</td>
<td>.110</td>
<td>.9122</td>
<td>.0002</td>
<td>.0010</td>
<td>.176</td>
<td>.8605</td>
<td>11.3206</td>
</tr>
<tr>
<td>INCOME</td>
<td>.1429</td>
<td>.0582</td>
<td>2.455</td>
<td>.0141</td>
<td>.0066</td>
<td>.0023</td>
<td>2.920</td>
<td>.0035</td>
<td>3.5208</td>
</tr>
<tr>
<td>MARRIED</td>
<td>-.0258</td>
<td>.0015</td>
<td>-17.531</td>
<td>.0000</td>
<td>-.0011</td>
<td>.0014</td>
<td>-.789</td>
<td>.4303</td>
<td>7.5862</td>
</tr>
<tr>
<td>KIDS</td>
<td>-.0039</td>
<td>.0008</td>
<td>-5.225</td>
<td>.0000</td>
<td>-.00088</td>
<td>.0010</td>
<td>-.084</td>
<td>.9331</td>
<td>4.0273</td>
</tr>
</tbody>
</table>
Advantages and Disadvantages of the FE Model

- Advantages
  - Allows correlation of effect and regressors
  - Fairly straightforward to estimate
  - Simple to interpret

- Disadvantages
  - Model may not contain time invariant variables
  - Not necessarily simple to estimate if very large samples (Stata just creates the thousands of dummy variables)
  - The incidental parameters problem: Small T bias
Incidental Parameters Problems: 
Conventional Wisdom

- **General**: The unconditional MLE is biased in samples with fixed $T$ except in special cases such as linear or Poisson regression (even when the FEM is the right model).

  The conditional estimator (that bypasses estimation of $\alpha_i$) is consistent.

- **Specific**: Upward bias (experience with probit and logit) in estimators of $\beta$
A Monte Carlo Study of the FE Estimator: Probit vs. Logit

Estimates of Coefficients and Marginal Effects at the Implied Data Means

Means of Empirical Sampling Distributions, $N = 1000$ Individuals Based on 200 Replications.

<table>
<thead>
<tr>
<th></th>
<th>$T=2$</th>
<th>$T=3$</th>
<th>$T=5$</th>
<th>$T=8$</th>
<th>$T=10$</th>
<th>$T=20$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>$\delta$</td>
<td>$\beta$</td>
<td>$\delta$</td>
<td>$\beta$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>Logit Coeff</td>
<td>2.020, 2.027</td>
<td>1.698, 1.668</td>
<td>1.379, 1.323</td>
<td>1.217, 1.156</td>
<td>1.161, 1.135</td>
<td>1.069, 1.062</td>
</tr>
<tr>
<td>Logit M.E.</td>
<td>1.676 1.660</td>
<td>1.523 1.477</td>
<td>1.319 1.254</td>
<td>1.191 1.128</td>
<td>1.140 1.111</td>
<td>1.034 1.052</td>
</tr>
<tr>
<td>Probit Coeff</td>
<td>2.083, 1.938</td>
<td>1.821, 1.777</td>
<td>1.589, 1.407</td>
<td>1.328, 1.243</td>
<td>1.247, 1.169</td>
<td>1.108, 1.068</td>
</tr>
<tr>
<td>Ord. Probit</td>
<td>2.328, 2.605</td>
<td>1.592, 1.806</td>
<td>1.305, 1.415</td>
<td>1.166, 1.220</td>
<td>1.131, 1.158</td>
<td>1.058, 1.068</td>
</tr>
</tbody>
</table>

Results are scaled so the desired quantity being estimated ($\beta$, $\delta$, marginal effects) all equal 1.0 in the population.
Bias Correction Estimators

- **Motivation:** Undo the incidental parameters bias in the fixed effects probit model:
  - (1) Maximize a penalized log likelihood function, or
  - (2) Directly correct the estimator of $\beta$

- **Advantages**
  - For (1) estimates $\alpha_i$ so enables partial effects
  - Estimator is consistent under some circumstances
  - (Possibly) corrects in dynamic models

- **Disadvantage**
  - No time invariant variables in the model
  - Practical implementation
  - Extension to other models? (Ordered probit model (maybe) – see JBES 2009)
A Mundlak Correction for the FE Model

Fixed Effects Model:

\[ y_{it}^* = \alpha_i + \beta' x_{it} + \varepsilon_{it}, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T_i \]

\( y_{it} = 1 \) if \( y_{it} > 0 \), 0 otherwise.

Mundlak (Wooldridge, Heckman, Chamberlain),...

\[ \alpha_i = \gamma + \theta' \overline{x}_i + u_i \] (Projection, not necessarily conditional mean)

where \( u \) is normally distributed with mean zero and standard deviation \( \sigma_u \) and is uncorrelated with \( \overline{x}_i \) or \((x_{i1}, x_{i2}, \ldots, x_{iT})\)

Reduced form random effects model

\[ y_{it}^* = \gamma + \theta' \overline{x}_i + \beta' x_{it} + \varepsilon_{it} + u_i, \quad i = 1, \ldots, N; \quad t = 1, \ldots, T_i \]

\( y_{it} = 1 \) if \( y_{it} > 0 \), 0 otherwise.
### Mundlak Correction

#### Table 2.17 Random Effects Model with Mundlak Correction

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Coef.</th>
<th>S.E.</th>
<th>t</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>.9459</td>
<td>.1116</td>
<td>8.473</td>
<td>.0000</td>
<td>.6551</td>
<td>.1232</td>
<td>5.320</td>
<td>.0000</td>
<td>43.5257</td>
</tr>
<tr>
<td>AGE</td>
<td>-.0365</td>
<td>.0015</td>
<td>-24.279</td>
<td>.0000</td>
<td>-.0521</td>
<td>.0036</td>
<td>-14.582</td>
<td>.0000</td>
<td>43.5257</td>
</tr>
<tr>
<td>EDUC</td>
<td>.0817</td>
<td>.0073</td>
<td>11.230</td>
<td>.0000</td>
<td>.0031</td>
<td>.0421</td>
<td>0.73</td>
<td>.9415</td>
<td>11.3206</td>
</tr>
<tr>
<td>INCOME</td>
<td>.3207</td>
<td>.0717</td>
<td>4.474</td>
<td>.0000</td>
<td>.2937</td>
<td>.0959</td>
<td>3.064</td>
<td>.0022</td>
<td>.35208</td>
</tr>
<tr>
<td>MARRIED</td>
<td>.0188</td>
<td>.0346</td>
<td>5.544</td>
<td>.5863</td>
<td>-.0429</td>
<td>.0534</td>
<td>-.803</td>
<td>.4220</td>
<td>.75862</td>
</tr>
<tr>
<td>KIDS</td>
<td>.0430</td>
<td>.0298</td>
<td>1.443</td>
<td>.1490</td>
<td>-.0019</td>
<td>.0397</td>
<td>-.048</td>
<td>.9614</td>
<td>.40273</td>
</tr>
<tr>
<td>AGEBAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0193</td>
<td>.0039</td>
<td>4.895</td>
<td>.0000</td>
<td></td>
</tr>
<tr>
<td>EDUCBAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0790</td>
<td>.0427</td>
<td>1.848</td>
<td>.0646</td>
<td></td>
</tr>
<tr>
<td>INCMBAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.3451</td>
<td>.1496</td>
<td>2.307</td>
<td>.0211</td>
<td></td>
</tr>
<tr>
<td>MARRBAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0499</td>
<td>.0717</td>
<td>.695</td>
<td>.4871</td>
<td></td>
</tr>
<tr>
<td>KIDSBAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.0936</td>
<td>.0616</td>
<td>1.520</td>
<td>.1285</td>
<td></td>
</tr>
<tr>
<td>Rho</td>
<td>.5404</td>
<td>.0100</td>
<td>53.842</td>
<td>.0000</td>
<td>.5389</td>
<td>.0100</td>
<td>53.822</td>
<td>.0000</td>
<td></td>
</tr>
</tbody>
</table>
A Variable Addition Test for FE vs. RE

The Wald statistic of 45.27922 and the likelihood ratio statistic of 40.280 are both far larger than the critical chi squared with 5 degrees of freedom, 11.07. This suggests that for these data, the fixed effects model is the preferred framework.
Fixed Effects Models Summary

- Incidental parameters problem if $T < 10$ (roughly)
- Inconvenience of computation
- Appealing specification
- Alternative semiparametric estimators?
  - Theory not well developed for $T > 2$
  - Not informative for anything but slopes (e.g., predictions and marginal effects)
- Ignoring the heterogeneity definitely produces an inconsistent estimator (*even with cluster correction!*)
- A Hobson’s choice
- Mundlak correction is a useful common approach.
A Study of Health Status in the Presence of Attrition

Research Article

The dynamics of health in the British Household Panel Survey

Paul Contoyannis¹, Andrew M. Jones².*, Nigel Rice³

Article first published online: 9 AUG 2004
DOI: 10.1002/jae.755
Copyright © 2004 John Wiley & Sons, Ltd.
Model for Self Assessed Health

- **British Household Panel Survey (BHPS)**
  - Self assessed health on 0,1,2,3,4 scale
  - Sociological and demographic covariates
  - Dynamics – inertia in reporting of top scale

- **Dynamic ordered probit model**
  - Balanced panel – analyze dynamics
  - Unbalanced panel – examine attrition
Dynamic Ordered Probit Model

Latent Regression - Random Utility
\[ h_{it}^* = \beta'x_{it} + \gamma'H_{i,t-1} + \alpha_i + \varepsilon_{it} \]
\[ x_{it} = \text{relevant covariates and control variables} \]

\[ H_{i,t-1} = 0/1 \text{ indicators of reported health status in previous period} \]
\[ H_{i,t-1}(j) = 1[\text{Individual i reported } h_{it} = j \text{ in previous period}], j=0,...,4 \]

Ordered Choice Observation Mechanism
\[ h_{it} = j \text{ if } \mu_{j-1} < h_{it}^* \leq \mu_j, j = 0,1,2,3,4 \]

Ordered Probit Model - \( \varepsilon_{it} \sim N[0,1] \)

Random Effects with Mundlak Correction and Initial Conditions
\[ \alpha_i = \alpha_0 + \alpha_1'H_{i,1} + \alpha_2'x_i + u_i, \ u_i \sim N[0,\sigma^2] \]
Random Effects Dynamic Ordered Probit Model

Random Effects Dynamic Ordered Probit Model

\[ h_{it}^* = x_{it}' \beta + \sum_{j=1}^{J} \gamma_j h_{i,t-1}(j) + \alpha_i + \epsilon_{i,t} \]

\[ h_{i,t} = j \text{ if } \mu_{j-1} < h_{it}^* < \mu_j \]

\[ h_{i,t}(j) = 1 \text{ if } h_{i,t} = j \]

\[ P_{it,j} = P[h_{it} = j] = \Phi(\mu_j - x_{it}' \beta - \sum_{j=1}^{J} \gamma_j h_{i,t-1}(j) - \alpha_i) \]

\[ - \Phi(\mu_{j-1} - x_{it}' \beta - \sum_{j=1}^{J} \gamma_j h_{i,t-1}(j) - \alpha_i) \]

Parameterize Random Effects

\[ \alpha_i = \alpha_0 + \sum_{j=1}^{J} \alpha_{1,j} h_{i,1}(j) + \mathbf{x}' \bar{\alpha}_i + u_i \]

Simulation or Quadrature Based Estimation

\[ \ln L = \sum_{i=1}^{N} \ln \int_{\alpha_i} \prod_{t=1}^{T_i} P_{it,j}(\alpha_j) d\alpha_j \]
## Data

Table I. Variable definitions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAH</td>
<td>Self-Assessed Health: 5 if excellent, 4 if good, 3 if fair, 2 if poor, 1 if very poor</td>
</tr>
<tr>
<td>WIDOW</td>
<td>1 if widowed, 0 otherwise</td>
</tr>
<tr>
<td>SINGLE</td>
<td>1 if never married, 0 otherwise</td>
</tr>
<tr>
<td>DIV/SEP</td>
<td>1 if divorced or separated, 0 otherwise</td>
</tr>
<tr>
<td>NON-WHITE</td>
<td>1 if a member of ethnic group other than white, 0 otherwise</td>
</tr>
<tr>
<td>DEGREE</td>
<td>1 if highest academic qualification is a degree or higher degree, 0 otherwise</td>
</tr>
<tr>
<td>HND/A</td>
<td>1 if highest academic qualification is HND or A level, 0 otherwise</td>
</tr>
<tr>
<td>O/CSE</td>
<td>1 if highest academic qualification is O level or CSE, 0 otherwise</td>
</tr>
<tr>
<td>HSIZE</td>
<td>Number of people in household including respondent</td>
</tr>
<tr>
<td>NCHO4</td>
<td>Number of children in household aged 0–4</td>
</tr>
<tr>
<td>NCH511</td>
<td>Number of children in household aged 5–11</td>
</tr>
<tr>
<td>NCH1218</td>
<td>Number of children in household aged 12–18</td>
</tr>
<tr>
<td>INCOME</td>
<td>Equivalized annual real household income in pounds</td>
</tr>
<tr>
<td>AGE</td>
<td>Age in years at 1st December of current wave</td>
</tr>
</tbody>
</table>
Variable of Interest

- SAH = excellent
- SAH = good
- SAH = fair
- SAH = poor
- SAH = very poor

Frequency

Wave 1 | Wave 2 | Wave 3 | Wave 4 | Wave 5 | Wave 6 | Wave 7 | Wave 8
---|---|---|---|---|---|---|---
0.5 | 0.45 | 0.4 | 0.35 | 0.3 | 0.25 | 0.2 | 0.15

Figure 1. Self-assessed health status by wave
Dynamics

Table II. Transition matrices, balanced panel

(a) Men

<table>
<thead>
<tr>
<th>SAH</th>
<th>EX</th>
<th>GOOD</th>
<th>FAIR</th>
<th>POOR</th>
<th>VERY POOR</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>EX</td>
<td>0.600</td>
<td>0.342</td>
<td>0.046</td>
<td>0.010</td>
<td>0.002</td>
<td>5485</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.184</td>
<td>0.651</td>
<td>0.142</td>
<td>0.019</td>
<td>0.004</td>
<td>9263</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.055</td>
<td>0.361</td>
<td>0.471</td>
<td>0.100</td>
<td>0.012</td>
<td>3433</td>
</tr>
<tr>
<td>POOR</td>
<td>0.029</td>
<td>0.120</td>
<td>0.340</td>
<td>0.418</td>
<td>0.093</td>
<td>1031</td>
</tr>
<tr>
<td>VERY POOR</td>
<td>0.032</td>
<td>0.073</td>
<td>0.133</td>
<td>0.423</td>
<td>0.339</td>
<td>248</td>
</tr>
<tr>
<td>N</td>
<td>5231</td>
<td>9287</td>
<td>3565</td>
<td>1111</td>
<td>266</td>
<td>19460</td>
</tr>
</tbody>
</table>

(b) Women

<table>
<thead>
<tr>
<th>SAH</th>
<th>EX</th>
<th>GOOD</th>
<th>FAIR</th>
<th>POOR</th>
<th>VERY POOR</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>EX</td>
<td>0.572</td>
<td>0.353</td>
<td>0.059</td>
<td>0.013</td>
<td>0.004</td>
<td>5164</td>
</tr>
<tr>
<td>GOOD</td>
<td>0.150</td>
<td>0.657</td>
<td>0.162</td>
<td>0.026</td>
<td>0.005</td>
<td>11306</td>
</tr>
<tr>
<td>FAIR</td>
<td>0.040</td>
<td>0.362</td>
<td>0.465</td>
<td>0.116</td>
<td>0.017</td>
<td>4928</td>
</tr>
<tr>
<td>POOR</td>
<td>0.021</td>
<td>0.156</td>
<td>0.360</td>
<td>0.365</td>
<td>0.098</td>
<td>1587</td>
</tr>
<tr>
<td>VERY POOR</td>
<td>0.014</td>
<td>0.106</td>
<td>0.192</td>
<td>0.326</td>
<td>0.362</td>
<td>423</td>
</tr>
<tr>
<td>N</td>
<td>4884</td>
<td>11329</td>
<td>5082</td>
<td>1649</td>
<td>464</td>
<td>23408</td>
</tr>
</tbody>
</table>
### Attrition

#### Table V. Sample size, drop-outs and attrition rates by wave

<table>
<thead>
<tr>
<th>Wave</th>
<th>FULL SAMPLE</th>
<th></th>
<th>( \text{EX at } t - 1 )</th>
<th>( \text{GOOD at } t - 1 )</th>
<th>( \text{FAIR at } t - 1 )</th>
<th>( \text{POOR at } t - 1 )</th>
<th>( \text{VPOOR at } t - 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. individuals</td>
<td>Survival rate</td>
<td>Drop-outs</td>
<td>Attrition rate</td>
<td>Attrition rate</td>
<td>Attrition rate</td>
<td>Attrition rate</td>
</tr>
<tr>
<td>1</td>
<td>10256</td>
<td>87.33%</td>
<td>1299</td>
<td>12.67%</td>
<td>11.54%</td>
<td>12.57%</td>
<td>13.01%</td>
</tr>
<tr>
<td>2</td>
<td>8957</td>
<td>79.58%</td>
<td>795</td>
<td>8.88%</td>
<td>8.08%</td>
<td>8.13%</td>
<td>9.65%</td>
</tr>
<tr>
<td>3</td>
<td>8162</td>
<td>76.30%</td>
<td>337</td>
<td>4.13%</td>
<td>6.67%</td>
<td>6.54%</td>
<td>6.73%</td>
</tr>
<tr>
<td>4</td>
<td>7825</td>
<td>76.30%</td>
<td>337</td>
<td>4.13%</td>
<td>6.67%</td>
<td>6.54%</td>
<td>6.73%</td>
</tr>
<tr>
<td>5</td>
<td>7430</td>
<td>72.45%</td>
<td>395</td>
<td>5.05%</td>
<td>6.21%</td>
<td>6.18%</td>
<td>7.87%</td>
</tr>
<tr>
<td>6</td>
<td>7238</td>
<td>70.57%</td>
<td>192</td>
<td>2.58%</td>
<td>3.11%</td>
<td>3.24%</td>
<td>5.06%</td>
</tr>
<tr>
<td>7</td>
<td>7102</td>
<td>69.25%</td>
<td>136</td>
<td>1.88%</td>
<td>3.15%</td>
<td>3.85%</td>
<td>4.79%</td>
</tr>
<tr>
<td>8</td>
<td>6839</td>
<td>66.68%</td>
<td>263</td>
<td>3.70%</td>
<td>3.43%</td>
<td>3.82%</td>
<td>5.30%</td>
</tr>
</tbody>
</table>


Testing for Attrition Bias

Table 9: Verbeek and Nijman tests for attrition: based on dynamic ordered probit models with Wooldridge specification of correlated effects and initial conditions

<table>
<thead>
<tr>
<th></th>
<th>MEN</th>
<th></th>
<th></th>
<th></th>
<th>WOMEN</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>Std.err.</td>
<td>t-test</td>
<td>p-value</td>
<td>β</td>
<td>Std.err.</td>
<td>t-test</td>
<td>p-value</td>
</tr>
<tr>
<td>NEXT WAVE</td>
<td>.199</td>
<td>.035</td>
<td>5.67</td>
<td>.000</td>
<td>.060</td>
<td>.034</td>
<td>1.77</td>
<td>.077</td>
</tr>
<tr>
<td>ALL WAVES</td>
<td>.139</td>
<td>.031</td>
<td>4.46</td>
<td>.000</td>
<td>.071</td>
<td>.029</td>
<td>2.45</td>
<td>.014</td>
</tr>
<tr>
<td>NUMBER OF WAVES</td>
<td>.031</td>
<td>.009</td>
<td>3.54</td>
<td>.000</td>
<td>.016</td>
<td>.008</td>
<td>1.88</td>
<td>.060</td>
</tr>
</tbody>
</table>

Three dummy variables added to full model with unbalanced panel suggest presence of attrition effects.
Probability Weighting Estimators

- A Patch for Attrition
- (1) Fit a participation probit equation for each wave.
- (2) Compute $p(i,t) = \text{predictions of participation for each individual in each period.}$
  - Special assumptions needed to make this work
- Ignore common effects and fit a weighted pooled log likelihood: $\Sigma_i \Sigma_t [d_{it}/p(i,t)] \log LP_{it}$. 
Attrition Model with IP Weights

Assumes (1) \( \text{Prob(attrition|all data)} = \text{Prob(attrition|selected variables)} \) (ignorability)

(2) Attrition is an ‘absorbing state.’ No reentry.

Obviously not true for the GSOEP data above.

Can deal with point (2) by isolating a subsample of those present at wave 1 and the monotonically shrinking subsample as the waves progress.
Inverse Probability Weighting

Panel is based on those present at WAVE 1, N1 individuals. Attrition is an absorbing state. No reentry, so N1 ≥ N2 ≥ ... ≥ N8. Sample is restricted at each wave to individuals who were present at the previous wave.

\[ d_{it} = 1 \text{[Individual is present at wave t].} \]

\[ d_{i1} = 1 \quad \forall \ i, d_{it} = 0 \implies d_{i,t+1} = 0. \]

\( \tilde{x}_{i1} \) = covariates observed for all i at entry that relate to likelihood of being present at subsequent waves.

(health problems, disability, psychological well being, self employment, unemployment, maternity leave, student, caring for family member, ...)

Probit model for \( d_{it} = 1[\delta'\tilde{x}_{i1} + w_{it}], \ t = 2,\ldots,8. \) \( \hat{\pi}_{it} = \text{fitted probability}. \)

Assuming attrition decisions are independent, \( \hat{P}_{it} = \prod_{s=1}^{t} \hat{\pi}_{is} \)

Inverse probability weight \( \hat{W}_{it} = \frac{d_{it}}{\hat{P}_{it}} \)

Weighted log likelihood \( \log L_w = \sum_{i=1}^{N} \sum_{t=1}^{8} \log L_{it} \) (No common effects.)
Estimated Partial Effects by Model

Table 12: Average partial effects on probability of reporting excellent health for selected variables

<table>
<thead>
<tr>
<th>a) Men</th>
<th>(1) Pooled model, balanced sample</th>
<th>(2) Pooled model, unbalanced sample</th>
<th>(3) Pooled model, IPW-1</th>
<th>(4) Pooled model, IPW-2</th>
<th>(5) Random effects, balanced sample</th>
<th>(6) Random effects, unbalanced sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ln(INCOME)</td>
<td>.009 (.004)</td>
<td>.009 (.004)</td>
<td>.009 (.004)</td>
<td>.011 (.005)</td>
<td>.013 (.006)</td>
<td>.012 (.005)</td>
</tr>
<tr>
<td>Mean Ln(INCOME)</td>
<td>.049 (.024)</td>
<td>.043 (.022)</td>
<td>.042 (.021)</td>
<td>.045 (.022)</td>
<td>.066 (.028)</td>
<td>.056 (.025)</td>
</tr>
<tr>
<td>DEGREE</td>
<td>.010 (.005)</td>
<td>.017 (.009)</td>
<td>.018 (.009)</td>
<td>.018 (.009)</td>
<td>.015 (.006)</td>
<td>.027 (.012)</td>
</tr>
<tr>
<td>HND/A</td>
<td>.019 (.009)</td>
<td>.021 (.011)</td>
<td>.021 (.010)</td>
<td>.022 (.011)</td>
<td>.028 (.011)</td>
<td>.030 (.013)</td>
</tr>
<tr>
<td>O/CSE</td>
<td>.016 (.008)</td>
<td>.020 (.010)</td>
<td>.020 (.010)</td>
<td>.020 (.010)</td>
<td>.024 (.010)</td>
<td>.028 (.012)</td>
</tr>
<tr>
<td>SAHEX(t-1)</td>
<td>.234 (.087)</td>
<td>.231 (.090)</td>
<td>.231 (.090)</td>
<td>.230 (.089)</td>
<td>.082 (.031)</td>
<td>.085 (.035)</td>
</tr>
<tr>
<td>SAHFAIR(t-1)</td>
<td>-.170 (.085)</td>
<td>-.163 (.084)</td>
<td>-.162 (.084)</td>
<td>-.162 (.083)</td>
<td>-.080 (.034)</td>
<td>-.077 (.036)</td>
</tr>
<tr>
<td>SAHPOOR(t-1)</td>
<td>-.242 (.167)</td>
<td>-.233 (.163)</td>
<td>-.232 (.162)</td>
<td>-.232 (.162)</td>
<td>-.151 (.077)</td>
<td>-.145 (.078)</td>
</tr>
<tr>
<td>SAHVPOOR(t-1)</td>
<td>-.260 (.198)</td>
<td>-.253 (.197)</td>
<td>-.255 (.199)</td>
<td>-.255 (.200)</td>
<td>-.184 (.104)</td>
<td>-.179 (.106)</td>
</tr>
</tbody>
</table>
These are 4 dummy variables for state in the previous period. Using first differences, the 0.234 estimated for SAHEX means transition from EXCELLENT in the previous period to GOOD in the previous period, where GOOD is the omitted category. Likewise for the other 3 previous state variables. The margin from ‘POOR’ to ‘GOOD’ was not interesting in the paper. The better margin would have been from EXCELLENT to POOR, which would have had (EX,POOR) change from (1,0) to (0,1).