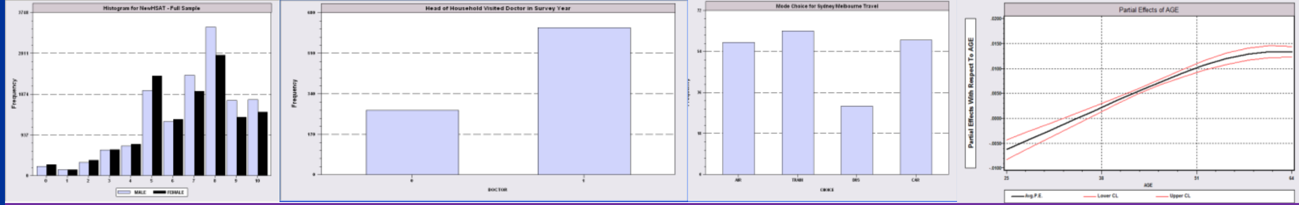
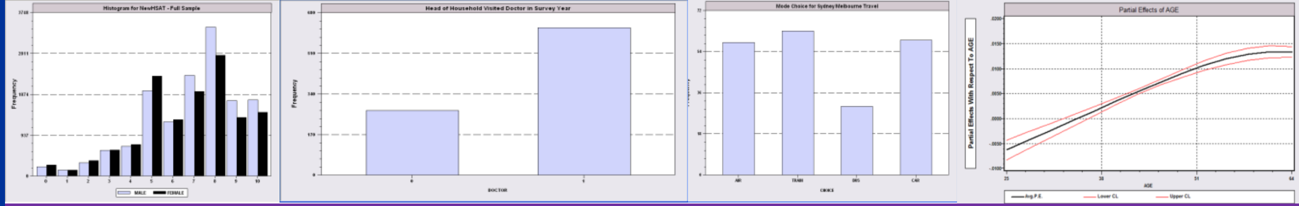


Discrete Choice Modeling

William Greene
Stern School of Business
New York University



University of Essex

iSER

About centres & surveys

Research projects & publications

Study Masters & PhDs

News updates & events

Home → BHPS

British Household Panel Survey

BHPS

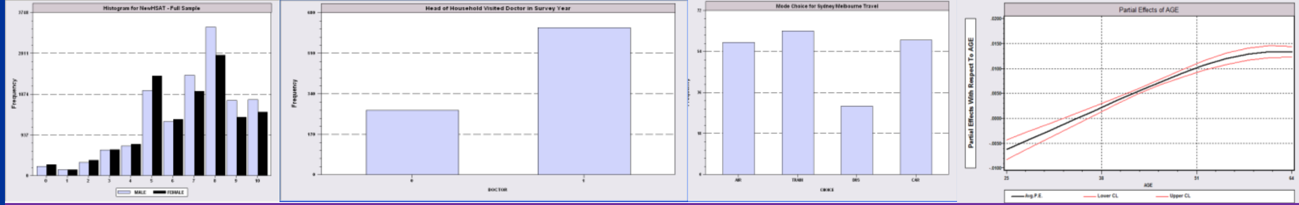
The British Household Panel Survey began in 1991 and is a multi-purpose study whose unique value resides in the fact that:

BHPS British Household Panel Survey

- 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.



FACULTY OF BUSINESS & ECONOMICS

THE UNIVERSITY OF MELBOURNE

Melbourne Institute

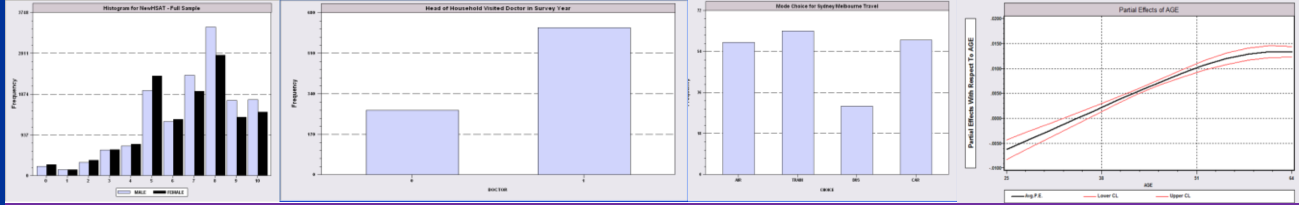
The Household, Income and Labour Dynamics in Australia (HILDA) Survey

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.

HILDA Home
News
Ordering the Data
Documentation and Support
HILDA Publications
Research Conference



A A A Deutsch Sitemap Newsletter Contact Imprint Data Protection DIW Berlin Suche

SOEP

About SOEP | Research Data Center SOEP

About SOEP

The SOEP Service Group

SOEP Quicklinks:

- SOEPinfo
- SOEPlit
- SOEPnewsletter
- SOEPmonitor
- SOEPdata Documents
- SOEPdata FAQ

About SOEP >

- Team
- Contact
- SOEP-Overview
- Mission
- SOEP Survey Committee

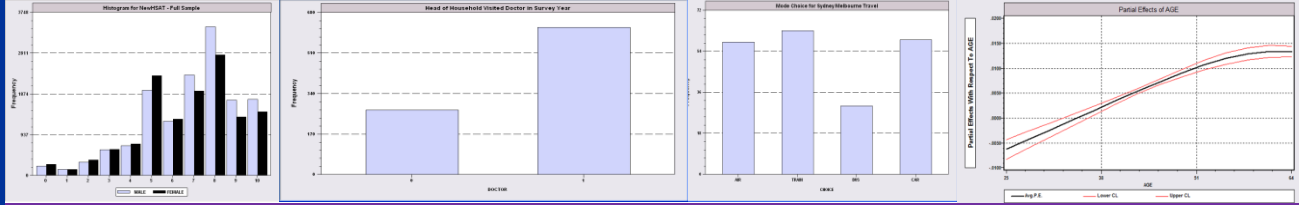
- Short Description
- Services of the Research Data Center SOEP
- Organization & Financing

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.



PSID

A national study of socioeconomics and health over lifetimes and across generations

STUDIES | DOCUMENTATION | DATA | PUBS, MEETINGS & MEDIA | PEOPLE | NEWS

Home

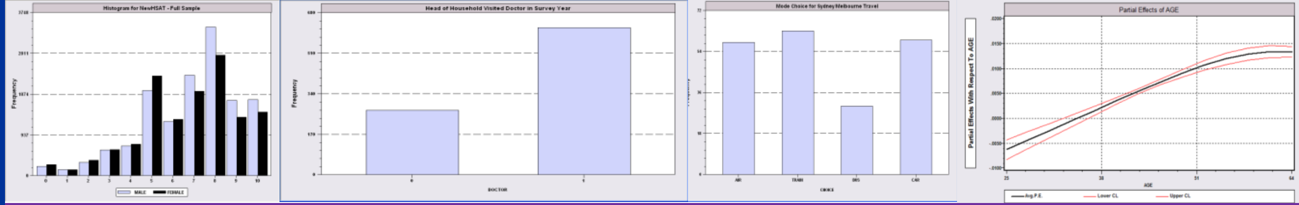
The Panel Study of Income Dynamics - PSID - is the longest running longitudinal household survey in the world.

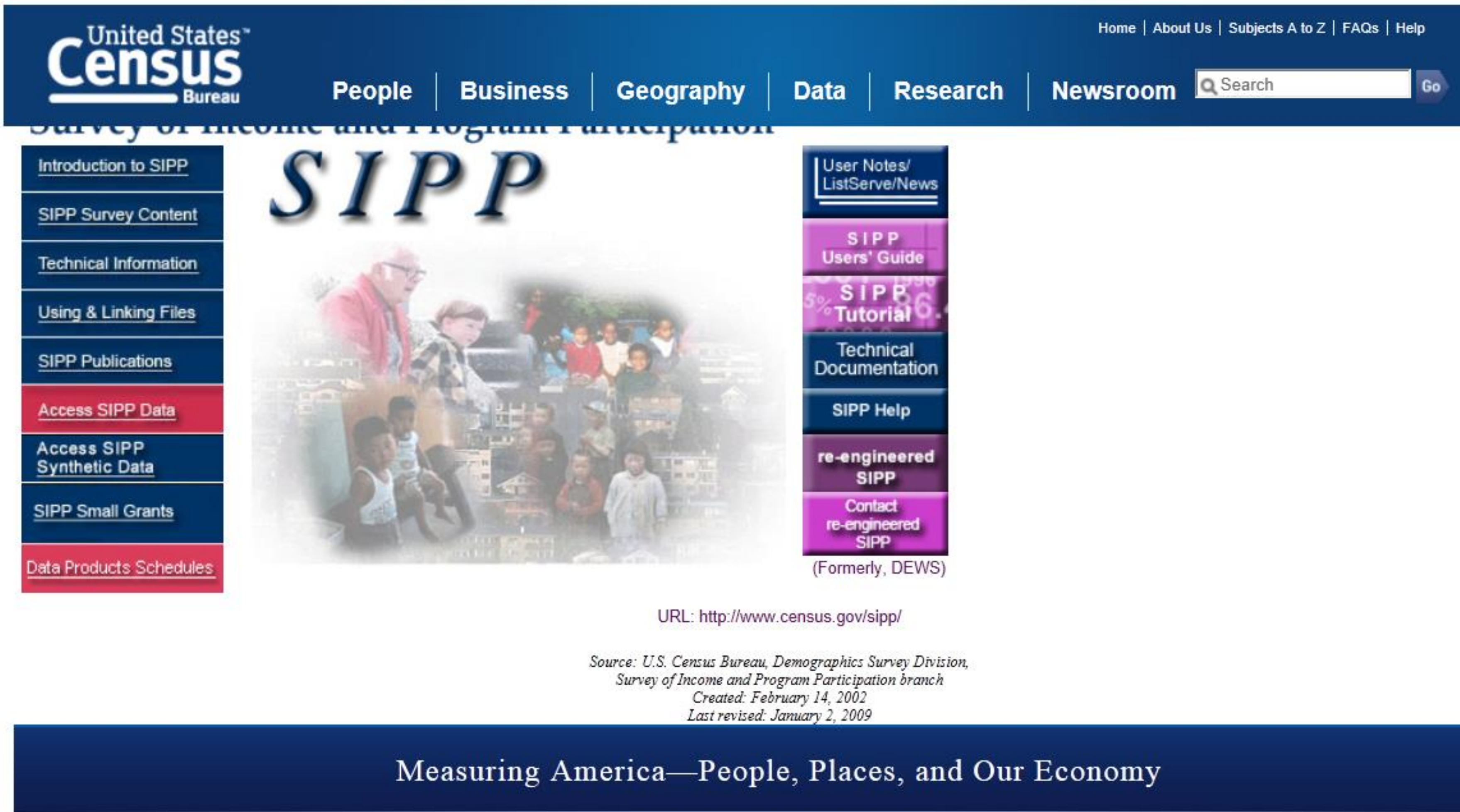
The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, 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.

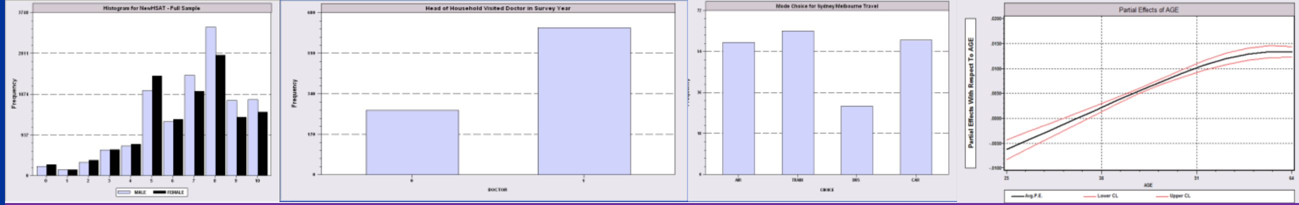
RECENT PUBLICATIONS

- Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentr...
  PODCAST
- Multigenerational Households and the School Readiness of Children Born to Unmarried Mother...
- Cumulative Effects of Job Characteristics on Health
- Essays on the Empirical Implications of Performance Pay Contracts





United States Census Bureau
Home | About Us | Subjects A to Z | FAQs | Help
People
Business
Geography
Data
Research
Newsroom
Search
Go
Introduction to SIPP
SIPP Survey Content
Technical Information
Using & Linking Files
SIPP Publications
Access SIPP Data
Access SIPP Synthetic Data
SIPP Small Grants
Data Products Schedules
SIPP
User Notes/ ListServe/News
SIPP Users' Guide
SIPP Tutorial
Technical Documentation
SIPP Help
re-engineered SIPP
Contact re-engineered SIPP
(Formerly, DEWS)
URL: http://www.census.gov/sipp/
Source: U.S. Census Bureau, Demographics 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



Register | Links | Contact | Important legal notice English (en)

European Commission
eurostat Your key to European statistics

European Commission > Eurostat > Access to microdata > European Community Household Panel

Home | Statistics | Publications | About Eurostat | User support

Access to microdata

- Introduction
- **European Community Household Panel**
  - Publications
- European Union Labour Force Survey
- Community Innovation Statistics
  - Publications
- European Union Statistics on Income and Living Conditions
  - Publications
- Structure of Earnings Survey
  - Publications
- Adult Education Survey
  - Publications

European Community Household Panel (ECHP)

ECHP microdata for scientific purposes: how to obtain them?

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...).

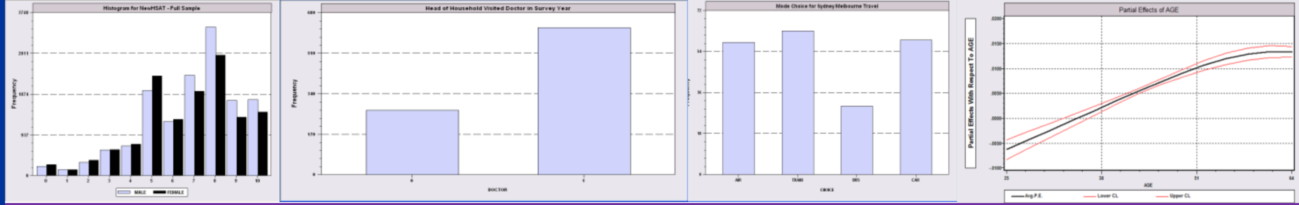
It also includes a selection of indicators of social exclusion and non-monetary deprivation derived from ECHP, notably on housing.

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.

See Also

- Additional information on ECHP
- Income, Social Inclusion and Living Conditions



UNITED STATES DEPARTMENT OF LABOR | A to Z Index | FAQs | About BLS | Contact Us | Subscribe to E-mail Updates GO

BUREAU OF LABOR STATISTICS | Follow Us | What's New | Release Calendar | Site Map | Search BLS.gov

Home | Subject Areas | Databases & Tools | Publications | Economic Releases | Beta

National Longitudinal Surveys

SHARE ON: | NLS | FONT SIZE: | PRINT:

BROWSE NLS

- NLS HOME
- NLS GENERAL OVERVIEWS
- NLS NEWS RELEASES
- NLS TABLES
- NLS PUBLICATIONS
- NLS FAQS
- CONTACT NLS

SEARCH NLS Go

NLS TOPICS

- NLSY97
- NLSY79
- NLSY79 CHILD & YOUNG ADULT
- NLS ORIGINAL COHORTS
- OBTAIN DATA
- DOCUMENTATION

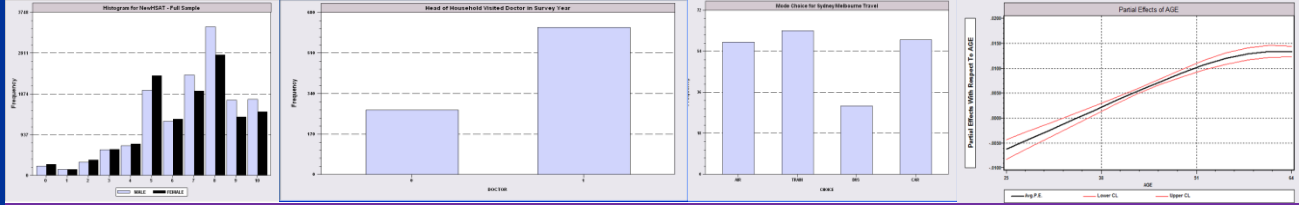
The **National Longitudinal Surveys (NLS)** are a set of surveys designed to gather information at multiple points in time on the labor market activities and other significant life events of several groups of men and women. For more than 4 decades, NLS data have served as an important tool for economists,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

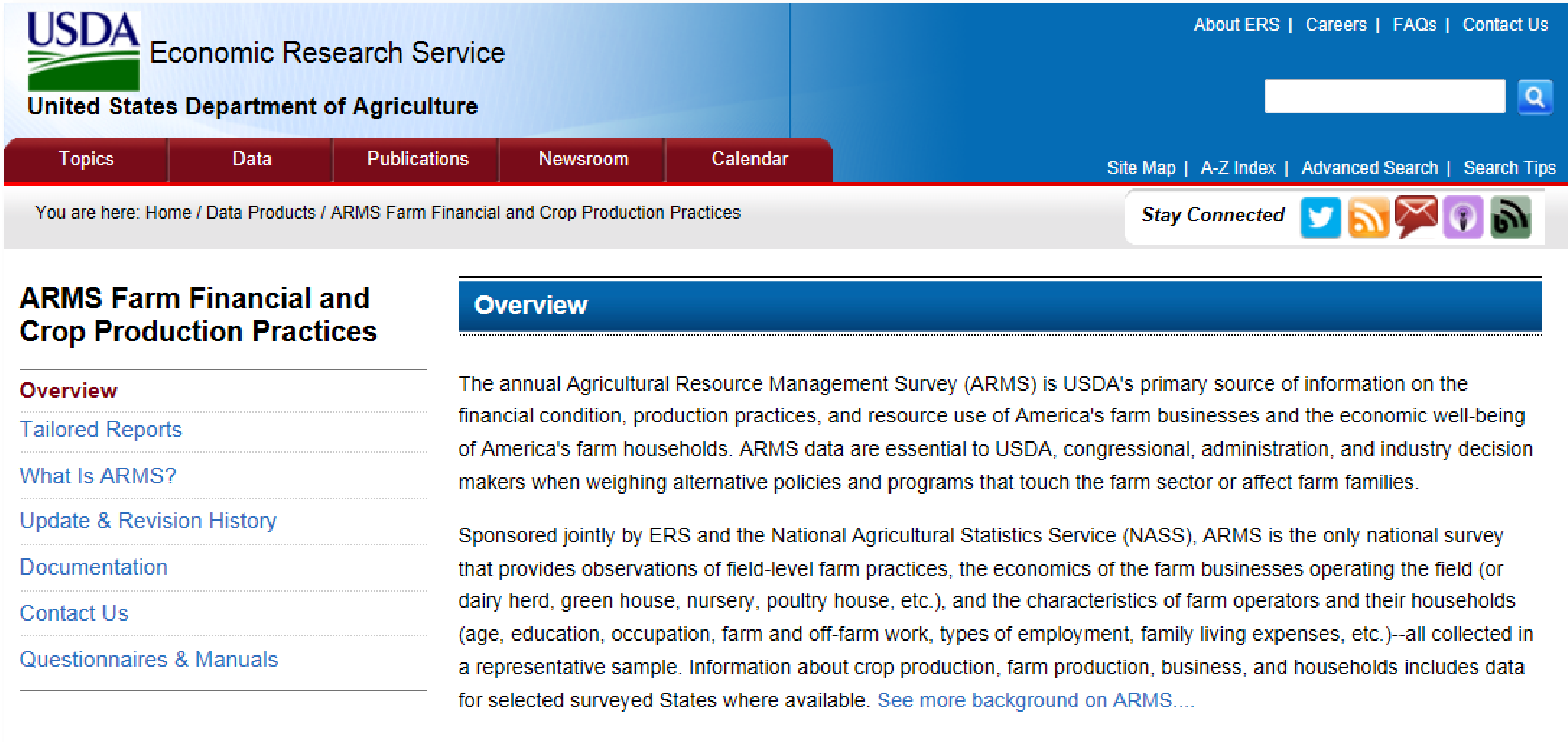
On This Page

- » 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 (NLSW)-- 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.





USDA Economic Research Service

United States Department of Agriculture

About ERS | Careers | FAQs | Contact Us

Topics | Data | Publications | Newsroom | Calendar

Site Map | A-Z Index | Advanced Search | Search Tips

You are here: Home / Data Products / ARMS Farm Financial and Crop Production Practices

Stay Connected

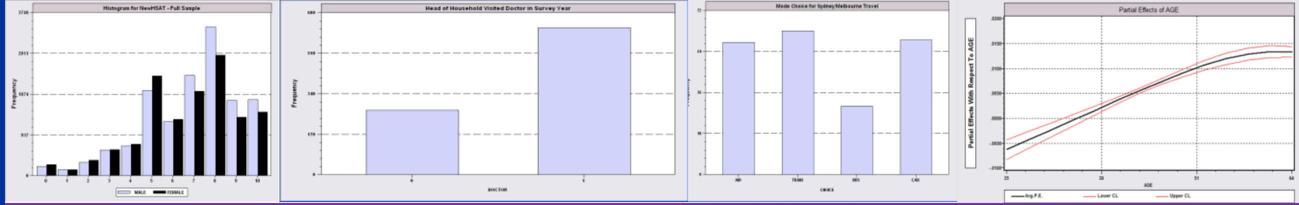
ARMS Farm Financial and Crop Production Practices

- **Overview**
- Tailored Reports
- What Is ARMS?
- Update & Revision History
- Documentation
- Contact Us
- Questionnaires & Manuals

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....



U.S. Department of Health & Human Services

AHRQ Agency for Healthcare Research and Quality

Advancing Excellence in Health Care

AHRQ Home | Questions? | Contact Us | Site Map | What's New | Browse | Información en español | E-mail

MEPS

Medical Expenditure Panel Survey

Contact MEPS | MEPS FAQ | Español | MEPS Site Map

Search MEPS Go

Font Size: S M L X

MEPS Home

About MEPS

- :: Survey Background
- :: Workshops & Events
- :: Data Release Schedule

Survey Components

- :: Household
- :: Insurance/Employer
- :: Medical Provider
- :: Survey Questionnaires

Data and Statistics

- :: Data Overview
- :: MEPS Topics
- :: Publications Search
- :: Summary Data Tables
- :: MEPSnet Query Tools

The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States. MEPS is the most complete source of data on the cost and use of health care and health insurance coverage. Learn more about MEPS.

Contact MEPS

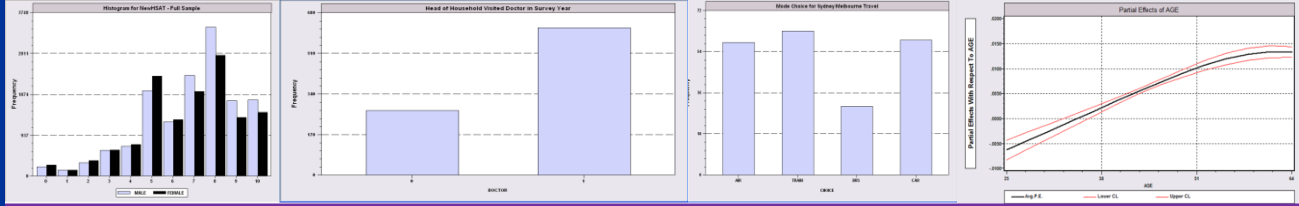
New to MEPS?

Select a profile:

- General user
- Researcher
- Policymaker
- Media
- Survey participant

MEPS Topics

- Access to Health Care
- Children's Health
- Children's Insurance Coverage
- Elderly Health Care
- Health Care Costs/Expenditures
- Health Care Disparities
- Health Insurance
- Medical Conditions
- Medicare/Medicaid/SCHIP
- Men's Health
- Mental Health
- Obesity
- Prescription Drugs
- Projected Data/Expenditures
- Quality of Health Care
- State and Metro Area Estimates
- The Uninsured
- Women's Health



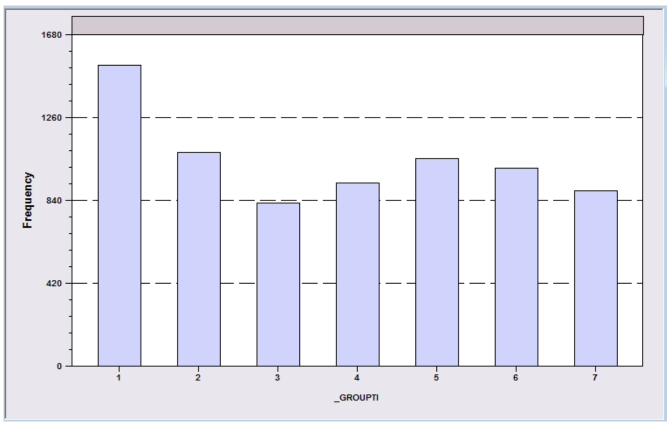
Application: Health Care Panel Data

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; otherswise = 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



Group Sizes

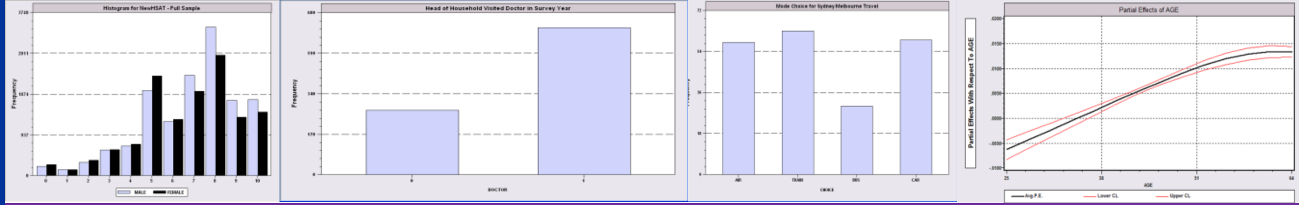
Most theoretical results are for balanced panels.

Most real world panels are unbalanced.

Often the gaps are caused by attrition.

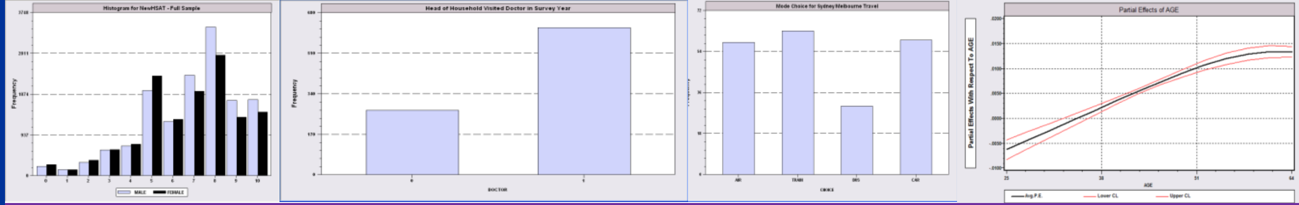
The major question is whether the gaps are 'missing completely at random.' If not, the observation mechanism is endogenous, and at least some methods will produce questionable results.

Researchers rarely have any reason to treat the data as nonrandomly sampled. (This is good news.)



Unbalanced Panels and Attrition ‘Bias’

- Test for ‘attrition bias.’ (Verbeek and Nijman, Testing for Selectivity Bias in Panel Data Models, International Economic Review, 1992, 33, 681-703.
  - Variable addition test using covariates of presence in the panel
  - Nonconstructive – what to do next?
- Do something about attrition bias. (Wooldridge, Inverse Probability Weighted M-Estimators for Sample Stratification and Attrition, Portuguese Economic Journal, 2002, 1: 117-139)
  - 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

$$U_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + \text{Person i specific effect}$$

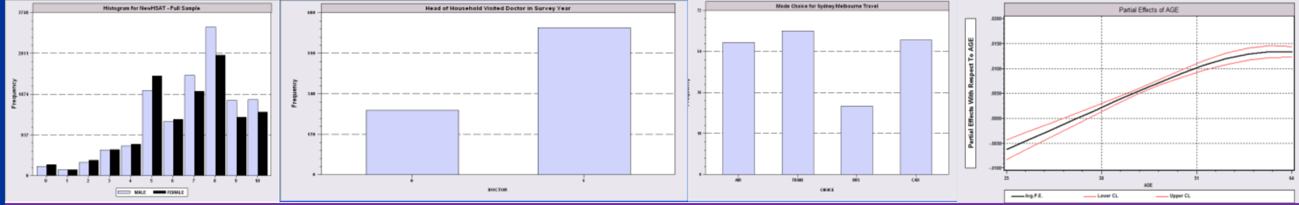
Fixed effects using "dummy" variables

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

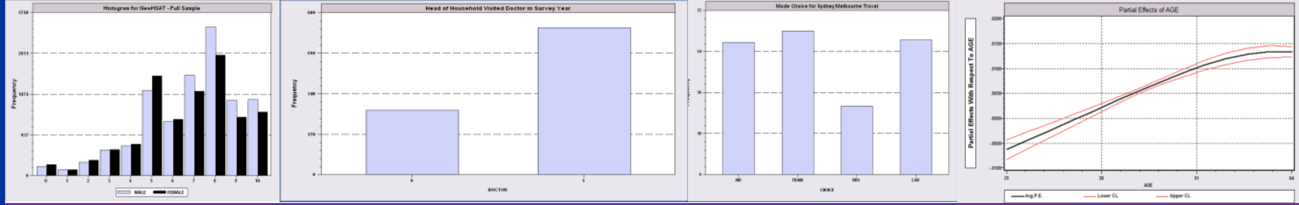
Random effects using omitted heterogeneity

$$U_{it} = \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i$$

Same outcome mechanism: $Y_{it} = 1[U_{it} > 0]$



Pooled Model



Ignoring Unobserved Heterogeneity

Assuming strict exogeneity; $\text{Cov}(\mathbf{x}_{it}, u_i + \varepsilon_{it}) = 0$

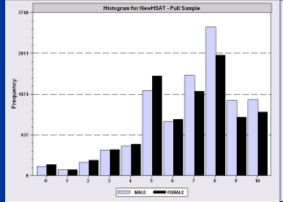
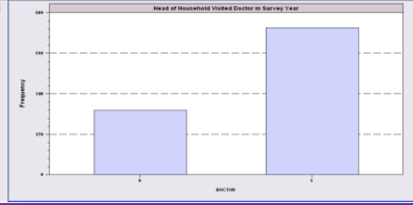
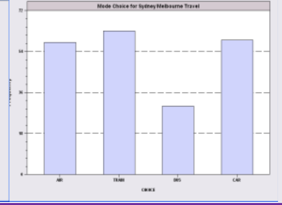
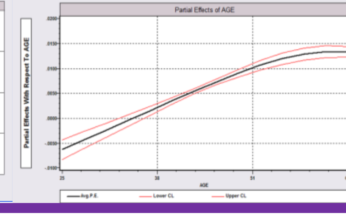
$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{it}$$

$$\text{Prob}[y_{it} = 1 \mid x_{it}] = \text{Prob}[u_i + \varepsilon_{it} > -\mathbf{x}_{it}'\boldsymbol{\beta}]$$

Using the same model format:

$$\text{Prob}[y_{it} = 1 \mid x_{it}] = F\left(\mathbf{x}_{it}'\boldsymbol{\beta} \,/\, \sqrt{1+\sigma_u^2}\right) = F(\mathbf{x}_{it}'\boldsymbol{\delta})$$

This is the 'population averaged model.'

Ignoring Heterogeneity in the RE Model

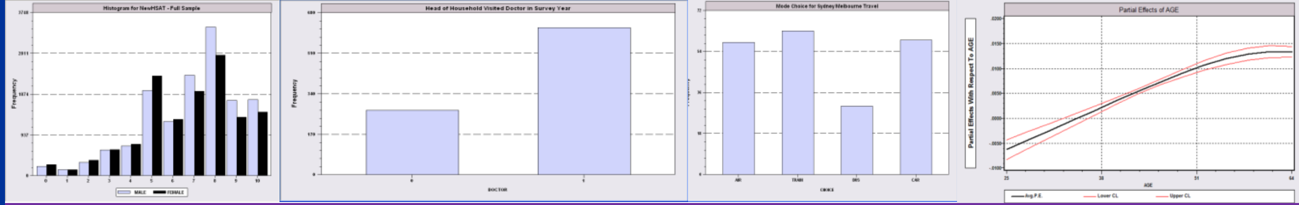
Ignoring heterogeneity, we estimate $\boldsymbol{\delta}$ not $\boldsymbol{\beta}$.

Partial effects are $\boldsymbol{\delta}\, f(\mathbf{x}_{it}'\boldsymbol{\delta})$ not $\boldsymbol{\beta} f(\mathbf{x}_{it}'\boldsymbol{\beta})$

$\boldsymbol{\beta}$ is underestimated, but $f(\mathbf{x}_{it}'\boldsymbol{\beta})$ is overestimated.

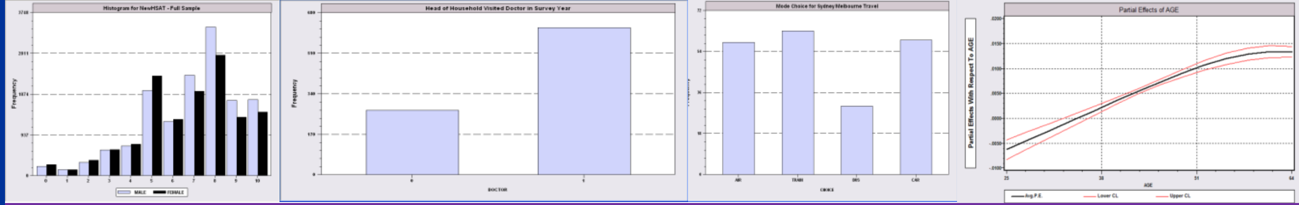
Which way does it go? Maybe ignoring u is ok?

Not if we want to compute probabilities or do statistical inference about $\boldsymbol{\beta}$. 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

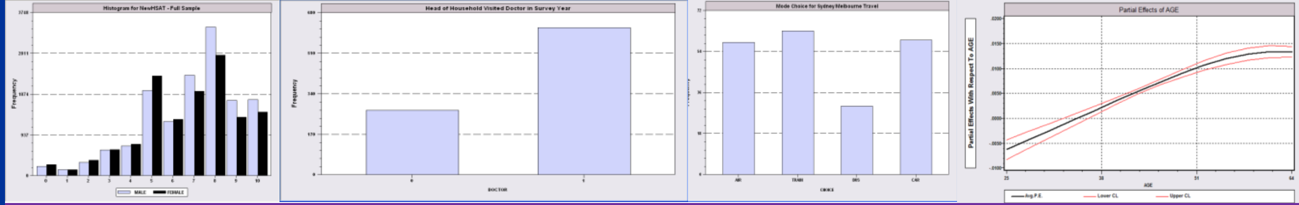
Binomial Probit Model

Dependent variable DOCTOR

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

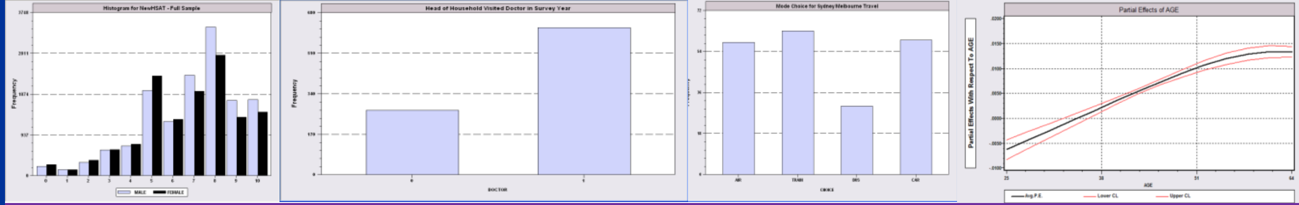
Unbalanced panel has 7293 individuals

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	



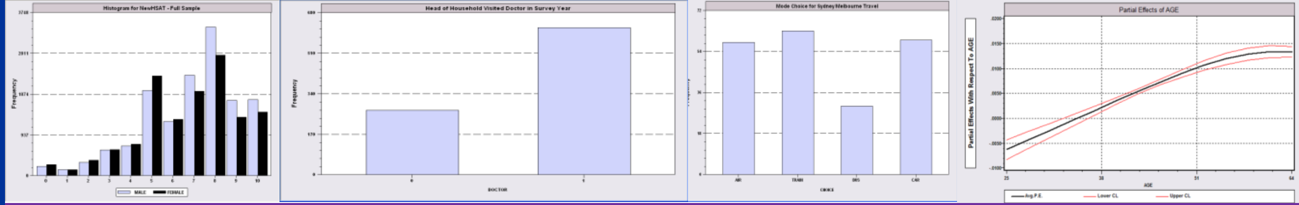
Partial Effects

```
----------------------------------------------------------------------
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
--------+-------------------------------------------------------------
        |Pooled
     AGE|    .00578***        .00027       21.720    .0000      .39801
    EDUC|   -.01053***        .00131       -8.024    .0000     -.18870
  HHNINC|   -.03847**         .01713       -2.246    .0247     -.02144
--------+-------------------------------------------------------------
        |Based on the panel data estimator
     AGE|    .00620***        .00034       18.375    .0000      .42181
    EDUC|   -.00918***        .00174       -5.282    .0000     -.16256
  HHNINC|    .00183           .01829         .100    .9202      .00101
--------+-------------------------------------------------------------
```



Effect of Clustering

- Y_{it} must be correlated with Y_{is} across periods
- Pooled estimator ignores correlation
- Broadly, $y_{it} = E[y_{it}|\mathbf{x}_{it}] + w_{it}$,
  - $E[y_{it}|\mathbf{x}_{it}] = \text{Prob}(y_{it} = 1|\mathbf{x}_{it})$
  - 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

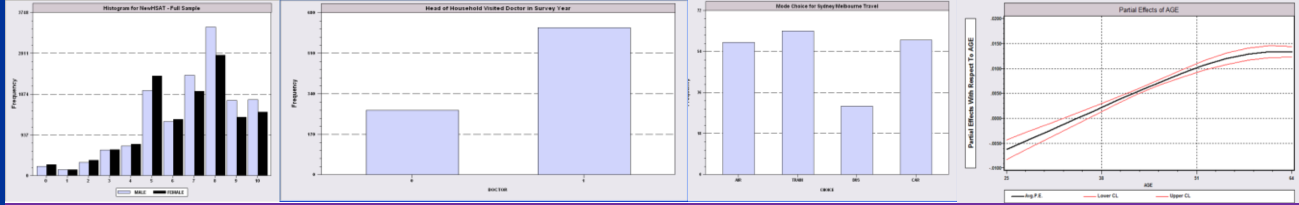
$C =$ the number if clusters

$n_c =$ number of observations in cluster c

$\mathbf{H}^{-1} =$ negative inverse of second derivatives matrix

$\mathbf{g}_{ic} =$ derivative of log density for observation

$$\mathbf{V} = \mathbf{H}^{-1}\left(\frac{C}{C-1}\right)\left(\sum_{c=1}^{C}\left(\sum_{i=1}^{n_c}\mathbf{g}_{ic}\right)\left(\sum_{i=1}^{n_c}\mathbf{g}_{ic}'\right)\right)\mathbf{H}^{-1}$$

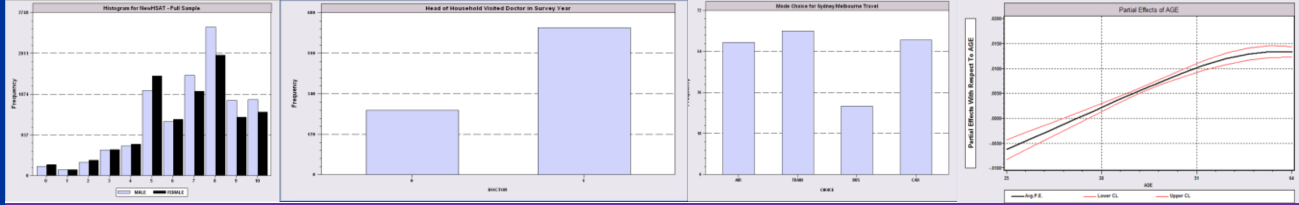


Cluster Correction: Doctor

Binomial Probit Model
Dependent variable DOCTOR
Log likelihood function -17457.21899

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Conventional Standard Errors				
Constant	-.25597***	.05481	-4.670	.0000	
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
	Corrected Standard Errors				
Constant	-.25597***	.07744	-3.305	.0009	
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 commerical software since 1982

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

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

$$u_i \sim N[0, \sigma_u^2]$$
$$= \sigma_u v_i$$
$$\text{where } v_i \sim N[0,1]$$

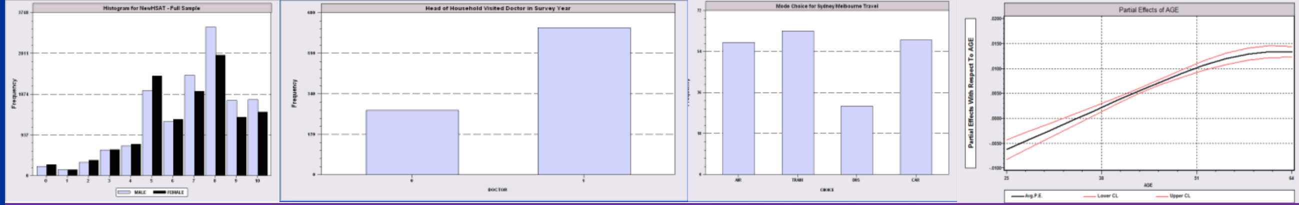
(make a change of variable to $w = v/\sqrt{2}$

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

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).



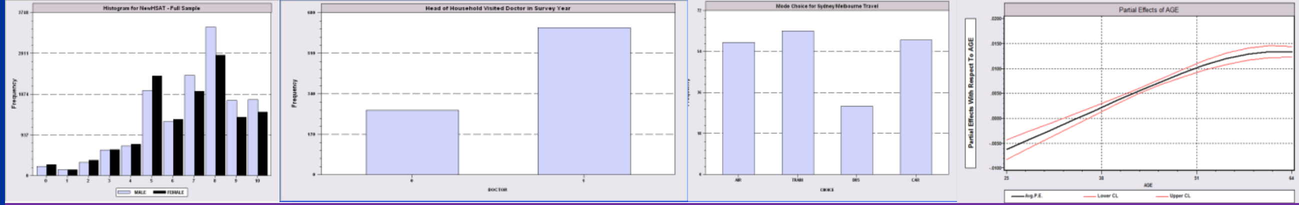
Quadrature Log Likelihood

After all the substitutions, the function to be maximized:

Not simple, but feasible.

$$\log L = \sum_{i=1}^{N} \log \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} w_h \left[\prod_{t=1}^{T_i} F\left(y_{it}, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \left(\sigma_u \sqrt{2}\right) z_h \right) \right]$$

$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} w_h \left[\prod_{t=1}^{T_i} F\left(y_{it}, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \theta z_h \right) \right]$$



Simulation Based Estimator

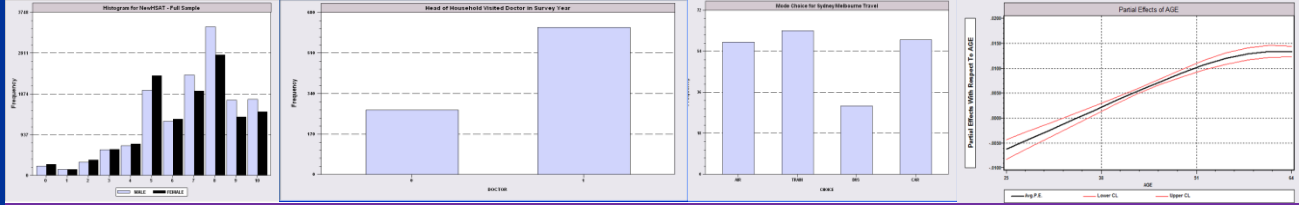
$$\log L = \sum_{i=1}^{N} \log \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T_i} F(y_{it}, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_u v_i) \right] \phi(v_i) dv_i$$

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

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

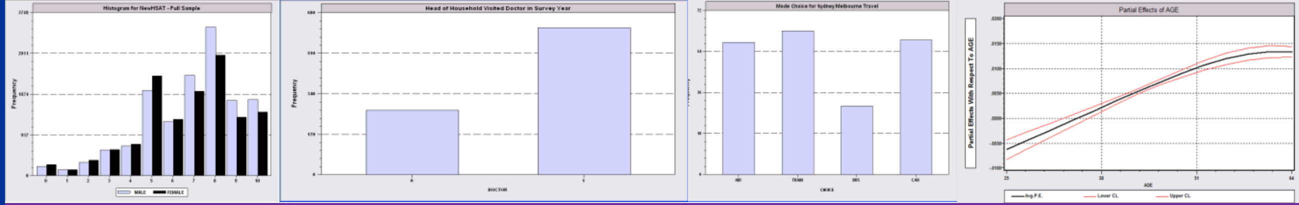
The expected value of the function of v_i can be approximated by drawing R random draws v_{ir} from the population N[0,1] and averaging the R functions of v_{ir}. We maximize

$$\log L_S = \sum_{i=1}^{N} \log \frac{1}{R} \sum_{r=1}^{R} \left[\prod_{t=1}^{T_i} F(y_{it}, \alpha + \boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_u v_{ir}) \right]$$



Random Effects Model: Quadrature

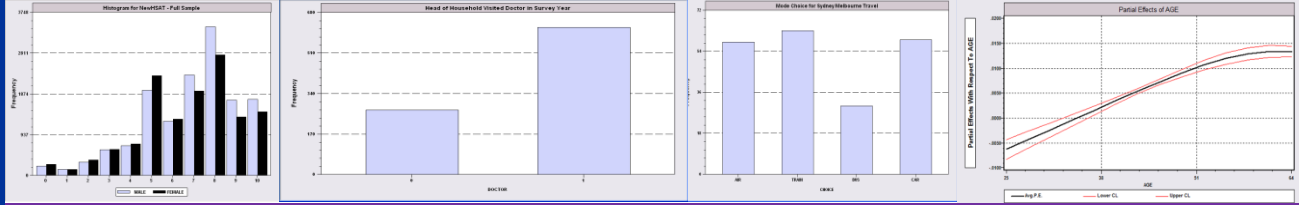
```
----------------------------------------------------------------------
Random Effects Binary Probit Model
Dependent variable                 DOCTOR
Log likelihood function      -16290.72192  ← Random Effects
Restricted log likelihood    -17701.08500  ← Pooled
Chi squared [    1 d.f.]       2820.72616
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|    -.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
--------+-------------------------------------------------------------
        |Pooled Estimates
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
--------+-------------------------------------------------------------
```



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'

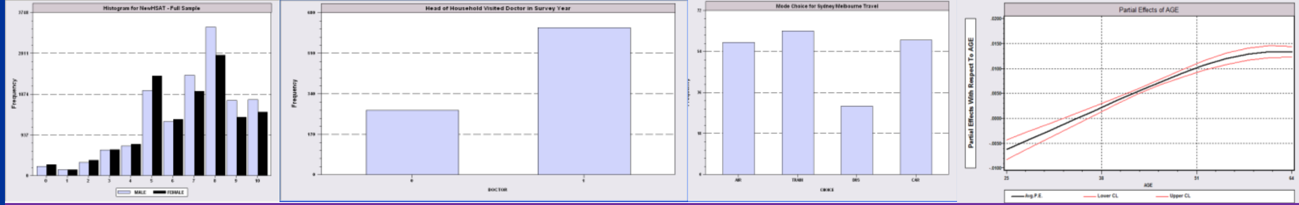
$$\Pr(y_{it} = 1 \mid y_{i,t-1}, \ldots, y_{i0}, x_{it}, u] = F[\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + u_i]$$

How to estimate β, γ, 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

$$\boxed{P(y_{i1}, y_{i2}, \ldots, y_{iT} \mid y_{i0}, \mathbf{x_i}, u_i) = \prod_{t=1}^{T} F(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + u_i, y_{it})}$$

(2) Unconditional density; integrate out the common effect

$$P(y_{i1}, y_{i2}, \ldots, y_{iT} \mid y_{i0}, \mathbf{x_i}) = \int_{-\infty}^{\infty} P(y_{i1}, y_{i2}, \ldots, y_{iT} \mid y_{i0}, \mathbf{x_i}, u_i) h(u_i \mid y_{i0}, \mathbf{x_i}) du_i$$

(3) Density for heterogeneity

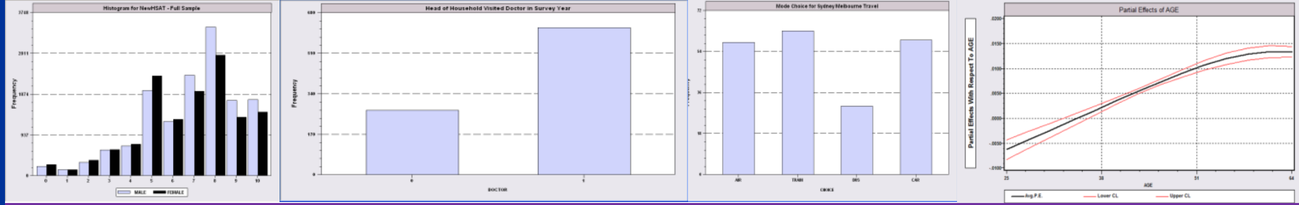
$$h(u_i \mid y_{i0}, \mathbf{x_i}) = N[\alpha + \theta y_{i0} + \mathbf{x_i}'\boldsymbol{\delta}, \sigma_u^2], \mathbf{x}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}], \text{ so}$$

$$\boxed{u_i = \alpha + \theta y_{i0} + \mathbf{x_i}'\boldsymbol{\delta} + \sigma_u w_i}$$ (contains every period of $\mathbf{x}_{it}$)

(4) Reduced form

$$\boxed{\begin{array}{l} P(y_{i1}, y_{i2}, \ldots, y_{iT} \mid y_{i0}, \mathbf{x_i}) = \\ \int_{-\infty}^{\infty} \prod_{t=1}^{T} F(\mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \alpha + \theta y_{i0} + \mathbf{x_i}'\boldsymbol{\delta} + \sigma_u w_i, y_{it}) h(w_i) dw_i \end{array}}$$

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

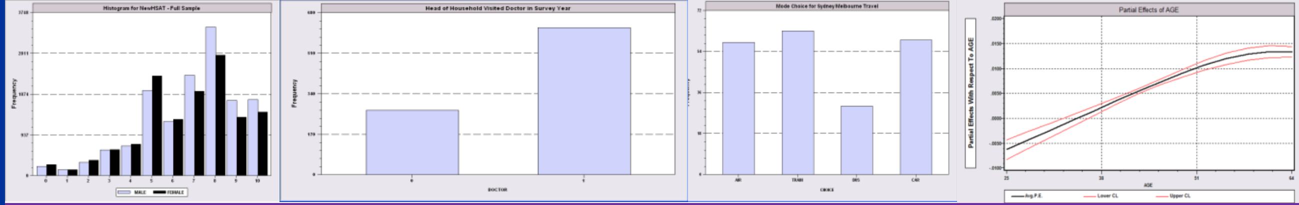
$$h(u_i \mid y_{i0}, \mathbf{x}_i) = N[\alpha + \theta y_{i0} + \bar{\mathbf{x}}_i'\boldsymbol{\delta}, \sigma_u^2] \text{ so}$$
$$u_i = \alpha + \theta y_{i0} + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + w_i$$

(4) Reduced form

$$P(y_{i1}, y_{i2}, \ldots, y_{iT} \mid y_{i0}, \mathbf{x}_i) =$$
$$\int_{-\infty}^{\infty} \prod_{t=1}^{T} F(\alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \gamma y_{i,t-1} + \theta y_{i0} + \bar{\mathbf{x}}_i'\boldsymbol{\delta} + \sigma_u w_i, y_{it}) h(w_i) dw_i$$

Mundlak style correction with the initial value in the equation.

This is (again) a random effects model



A Dynamic Model for Public Insurance

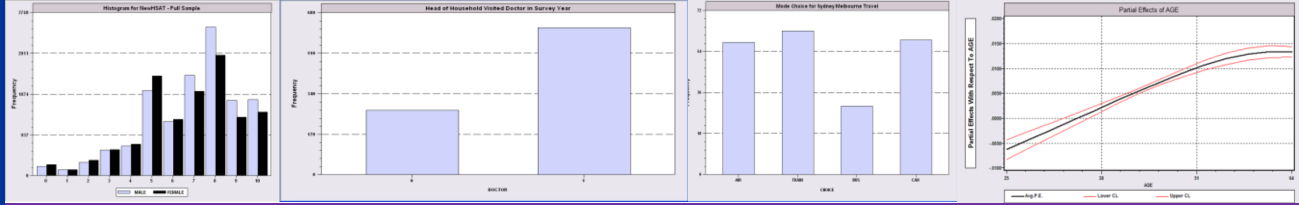
Untitled 2 *

Insert Name:

```
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,pds=ti)$
create    ; yi0   =groupobs1(public,pds=ti)$
probit    ; if[obs > 1];Panel ; lhs=public
            ;rhs=xit,means,one,yi0,public[-1] $
```

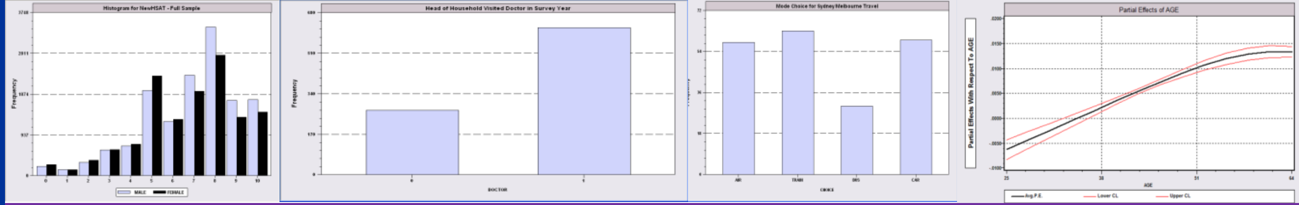
Age
Household Income
Kids in the household
Health Status

Add initial value, lagged value, group means

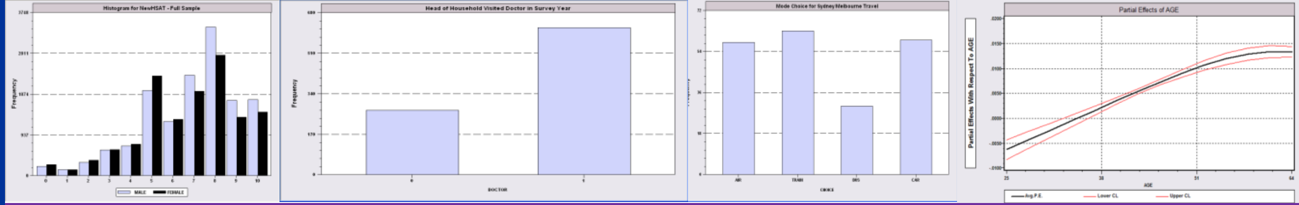


Dynamic Common Effects Model

```
-----------------------------------------------------------------------------
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
--------+--------------------------------------------------------------------
        |                 Standard            Prob.       95% Confidence
  PUBLIC|  Coefficient      Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
     AGE|   -.01568        .01212    -1.29  .1957     -.03944     .00808
  INCOME|   -.67338***     .25494    -2.64  .0083    -1.17305    -.17371
  HHKIDS|   -.01281        .11641     -.11  .9124     -.24095     .21534
    HSAT|   -.00336        .02000     -.17  .8666     -.04256     .03584
  AGEBAR|    .04392***     .01258     3.49  .0005      .01926     .06858
  INCBAR|  -1.71950***     .38878    -4.42  .0000    -2.48149    -.95751
 KIDSBAR|    .26462*       .15011     1.76  .0779     -.02959     .55883
 HSATBAR|   -.05228        .03223    -1.62  .1048     -.11545     .01089
Constant|  -1.57400***     .31448    -5.01  .0000    -2.19038    -.95762
     YI0|   4.02429***     .28588    14.08  .0000     3.46398    4.58460
    YLAG|    .95309***     .09358    10.18  .0000      .76967    1.13650
     Rho|    .68459***     .03143    21.78  .0000      .62300     .74618
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```



Fixed Effects



Fixed Effects Models

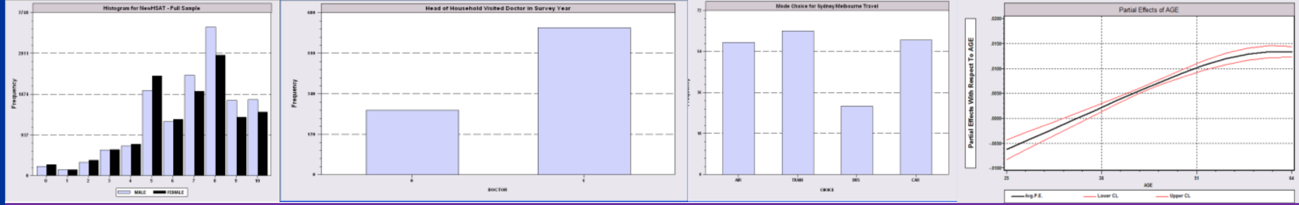
- Estimate with dummy variable coefficients

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

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

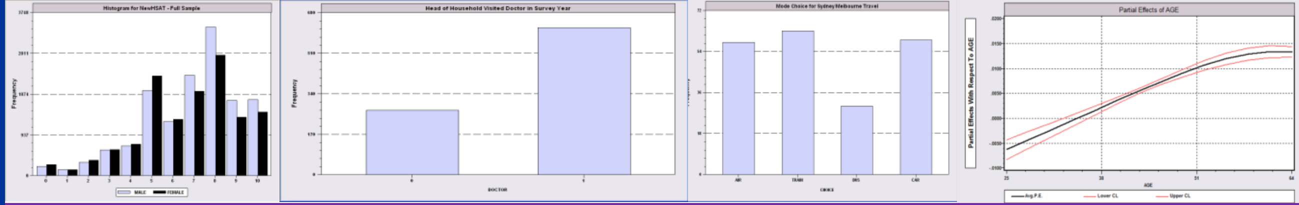
$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log F(y_{it}, \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it})$$

- F(.) = appropriate probability for the observed outcome
- Compute β and α_i for i=1,…,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

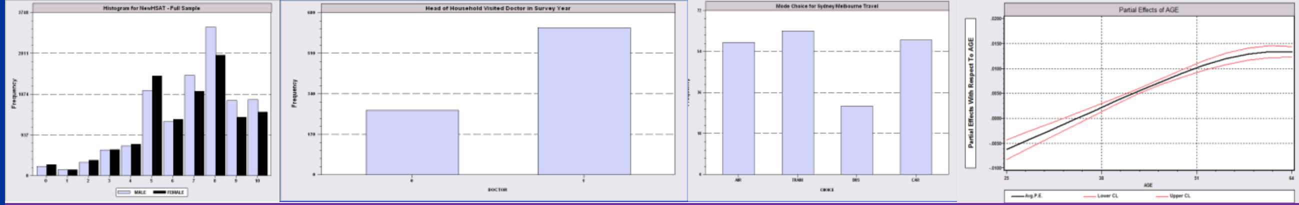
Groups in which y_{it} is always = 0 or always = 1. Cannot compute α_i.

```
Fixed Effects                     Pooled
LogL   = -8500.704                LogL   = -17365.76
LogLR  = -17365.76                LogL0  = -18279.95
7293 Individuals
3289 Individuals Bypassed
```

Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant					.4963	.0589	8.425	.0000	1.0000
AGE	-.0649	.0045	-14.418	.0000	-.0232	.0008	-28.991	.0000	43.5257
EDUC	.0027	.0506	.054	.9570	.0573	.0037	15.467	.0000	11.3206
INCOME	.3530	.1161	3.040	.0024	.3425	.0481	7.118	.0000	.35208
MARRIED	-.0609	.0666	-.915	.3600	.0129	.0206	.627	.5307	.75862
KIDS	-.0118	.0475	-.249	.8032	.0666	.0186	3.581	.0003	.40273
	Partial Effects				Partial Effects				
AGE	-.0248	.0049	-5.087	.0000	-.0089	.0003	-29.012	.0000	43.5257
EDUC	.0010	.0192	.054	.9567	.0219	.0014	15.478	.0000	11.3206
INCOME	.1349	.0515	2.617	.0089	.1309	.0184	7.118	.0000	.35208
MARRIED	-.0233	.0010	-22.562	.0000	.0049	.0079	.626	.5311	.75862
KIDS	-.0045	.0004	-10.792	.0000	.0254	.0071	3.589	.0003	.40273

Conditional Estimation

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



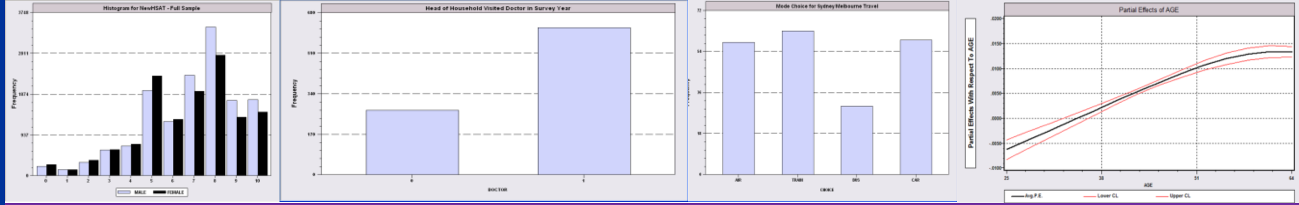
Binary Logit Conditional Probabiities

$$\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it}) = \frac{e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

$$\text{Prob}\left(Y_{i1}=y_{i1}, Y_{i2}=y_{i2}, \ldots, Y_{iT_i}=y_{iT_i} \,\middle|\, \sum_{t=1}^{T_i} y_{it}\right)$$

$$= \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it}=S_i} \exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)} = \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\substack{\text{All } \binom{T_i}{S_i} \text{ different ways that} \\ \Sigma_t d_{it} \text{ can equal } S_i}} \exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

Denominator is summed over all the different combinations of T_i values of y_{it} that sum to the same sum as the observed $\Sigma_{t=1}^{T_i} y_{it}$. 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

$$\text{Prob}(y_{it}=1\,|\,\mathbf{x}_{it}) = \frac{e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}{1+e^{\alpha_i+\mathbf{x}_{it}'\boldsymbol{\beta}}}.$$

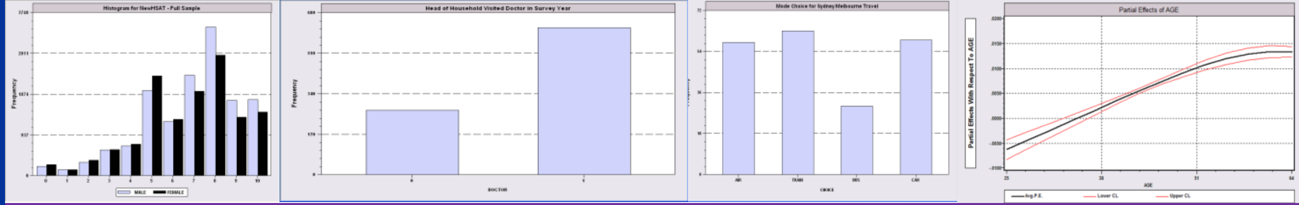
$$\text{Prob}\left(Y_{i1}=y_{i1}, Y_{i2}=y_{i2}, \ldots, Y_{iT_i}=y_{iT_i}\,\middle|\,\sum_{t=1}^{T_i} y_{it}, \text{data}\right) = \frac{\exp\left(\sum_{t=1}^{T_i} y_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}{\sum_{\Sigma_t d_{it}=S_i}\exp\left(\sum_{t=1}^{T_i} d_{it}\mathbf{x}_{it}'\boldsymbol{\beta}\right)}.$$

$$\boxed{\text{Prob}\left(Y_{i1}=0, Y_{i2}=0\,\middle|\,\sum_{t=1}^{2} y_{it}=0, \text{data}\right) = 1.}$$

$$\text{Prob}\left(Y_{i1}=1, Y_{i2}=0\,\middle|\,\sum_{t=1}^{2} y_{it}=1, \text{data}\right) = \frac{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})+\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}$$

$$\text{Prob}\left(Y_{i1}=0, Y_{i2}=1\,\middle|\,\sum_{t=1}^{2} y_{it}=1, \text{data}\right) = \frac{\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}{\exp(\mathbf{x}_{i1}'\boldsymbol{\beta})+\exp(\mathbf{x}_{i2}'\boldsymbol{\beta})}$$

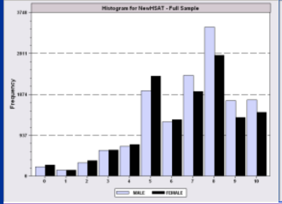
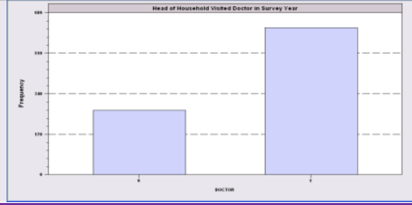
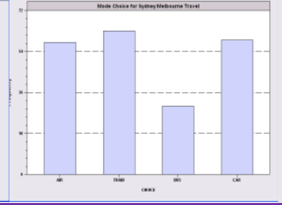
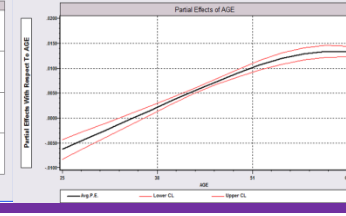
$$\boxed{\text{Prob}\left(Y_{i1}=1, Y_{i2}=1\,\middle|\,\sum_{t=1}^{2} y_{it}=2, \text{data}\right) = 1.}$$



Estimating Partial Effects

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

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

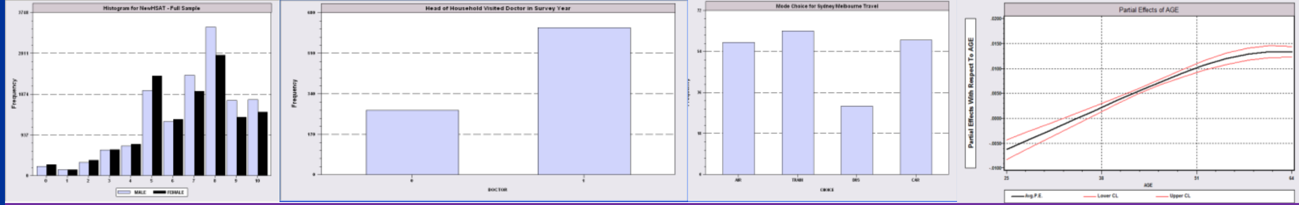
Estimate $\mu_i = 1/\exp(\alpha_i) \Rightarrow \alpha_i = -\log \mu_i$

$c_{it} = \exp\left(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}\right)$ is treated as known data.

Solve one equation in one unknown for each α_i.

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

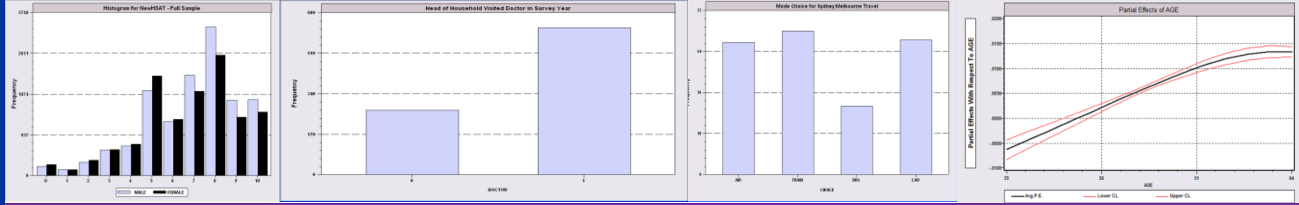
Iterating back and forth does not maximize logL.



Fixed Effects Logit Health Model: Conditional vs. Unconditional

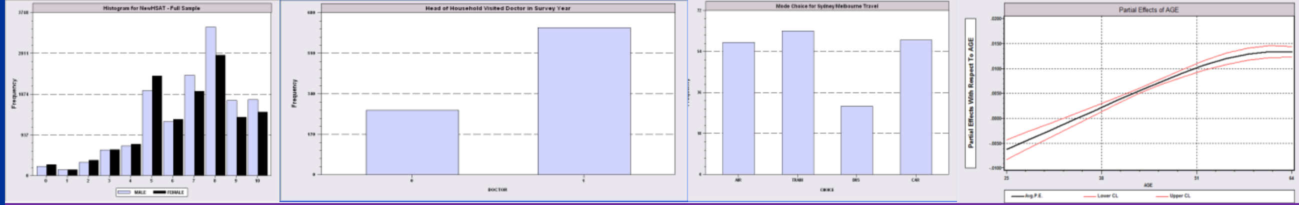
Table 2.14 Estimated Fixed Effects Logit Models

	Unconditional Estimator LogL = -8506.164 LogLR = -17365.15 7293 Individuals 3289 Individuals Bypassed				Conditional Estimator LogL = -5669.541				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
AGE	-.1095	.0076	-14.405	.0000	-.0881	.0068	-12.984	.0000	43.5257
EDUC	.0090	.0835	.108	.9141	.0126	.0718	.176	.8604	11.3206
INCOME	.6038	.1968	3.068	.0022	.4767	.1750	2.724	.0064	.35208
MARRIED	-.1091	.1114	-.979	.3276	-.0772	.0983	-.785	.4322	.75862
KIDS	-.0167	.0793	-.210	.8337	-.0059	.0706	-.084	.9331	.40273
	Partial Effects				Partial Effects				
AGE	-.0259	.0063	-4.102	.0000	-.0012	.00009	-13.961	.0000	43.5257
EDUC	.0021	.0193	.110	.9122	.0002	.0010	.176	.8605	11.3206
INCOME	.1429	.0582	2.455	.0141	.0066	.0023	2.920	.0035	.35208
MARRIED	-.0258	.0015	-17.531	.0000	-.0011	.0014	-.789	.4303	.75862
KIDS	-.0039	.0008	-5.225	.0000	-.00008	.0010	-.084	.9331	.40273



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**

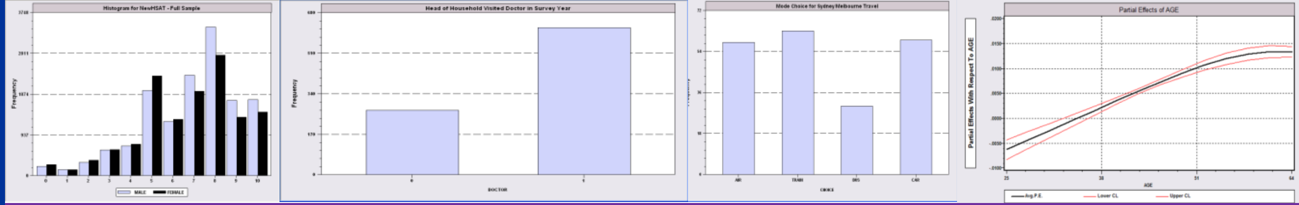


Incidental Parameters Problems: Conventional Wisdom

- 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 α_i) is consistent.

- Specific: Upward bias (experience with probit and logit) in estimators of β



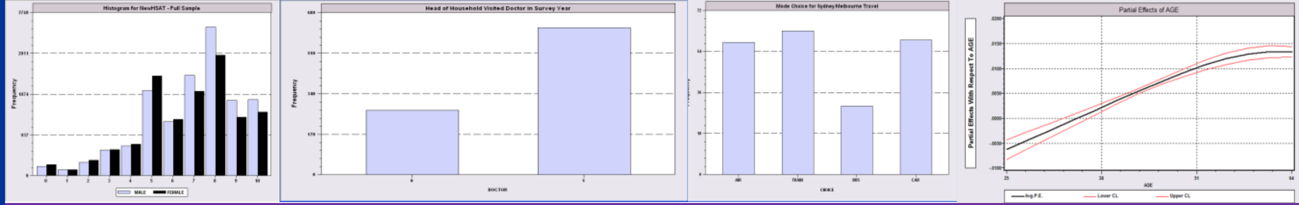
A Monte Carlo Study of the FE Estimator: Probit vs. Logit

Estimates of Coefficients and Marginal Effects at the Implied Data Means

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

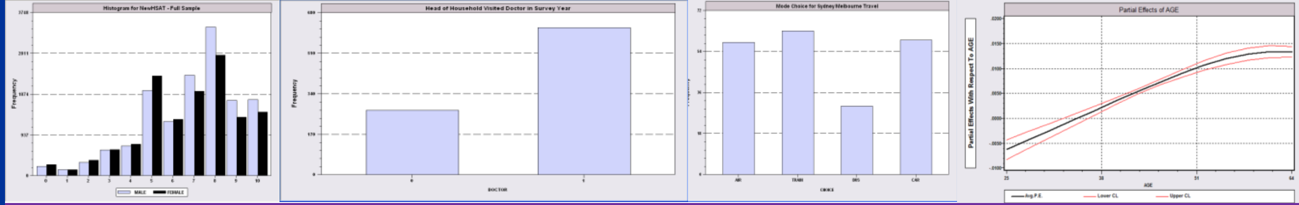
	T=2	*T*=3	*T*=5	*T*=8	*T*=10	*T*=20
	β δ	β δ	β δ	β δ	β δ	β δ
Logit Coeff	2.020, 2.027	1.698, 1.668	1.379, 1.323	1.217, 1.156	1.161, 1.135	1.069, 1.062
Logit M.E.	1.676 1.660	1.523 1.477	1.319 1.254	1.191 1.128	1.140 1.111	1.034 1.052
Probit Coeff	2.083, 1.938	1.821, 1.777	1.589, 1.407	1.328, 1.243	1.247, 1.169	1.108, 1.068
Probit M.E.	1.474 1.388	1.392 1.354	1.406 1.231	1.241 1.152	1.190 1.110	1.088 1.047
Ord. Probit	2.328, 2.605	1.592, 1.806	1.305, 1.415	1.166, 1.220	1.131, 1.158	1.058, 1.068

Results are scaled so the desired quantity being estimated (β, δ, 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 β
- Advantages
  - For (1) estimates α_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 :

$\boxed{y_{it}^* = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it}}$, i = 1,...,N; t = 1,...,T_i

$y_{it} = 1$ if $y_{it} > 0$, 0 otherwise.

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

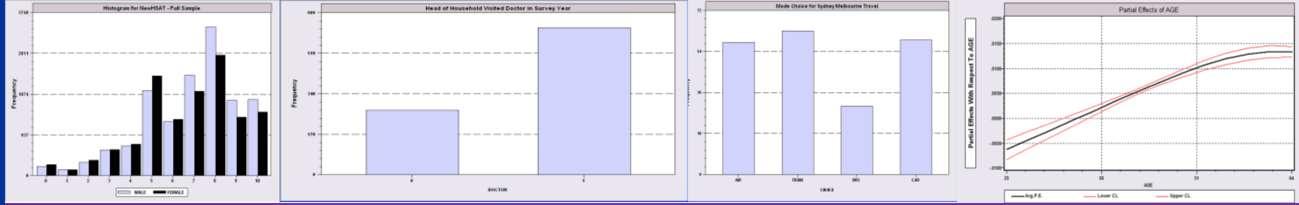
$\boxed{\alpha_i = \gamma + \boldsymbol{\theta}'\overline{\mathbf{x}}_i + u_i}$ (Projection, not necessarily conditional mean)

where u is normally distributed with mean zero and standard deviation σ_u and is uncorrelated with $\overline{\mathbf{x}}_i$ or $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$

Reduced form random effects model

$\boxed{y_{it}^* = \gamma + \boldsymbol{\theta}'\overline{\mathbf{x}}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \varepsilon_{it} + u_i}$, i = 1,...,N; t = 1,...,T_i

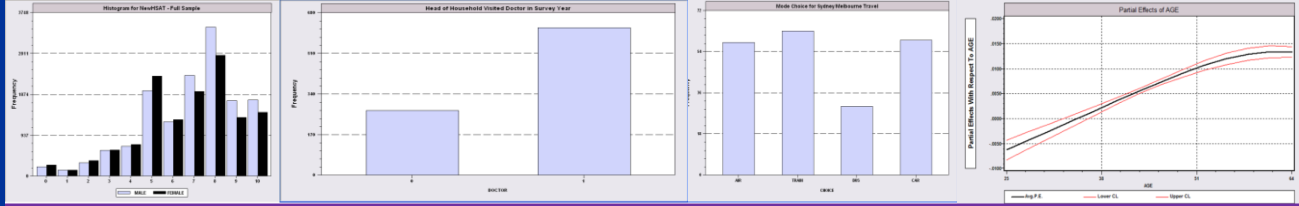
$y_{it} = 1$ if $y_{it} > 0$, 0 otherwise.



Mundlak Correction

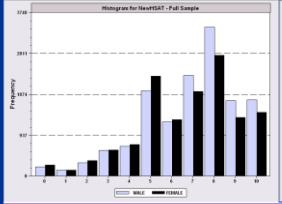
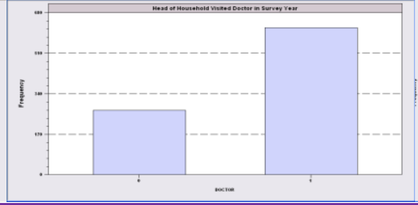
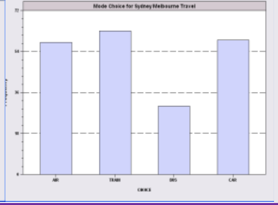
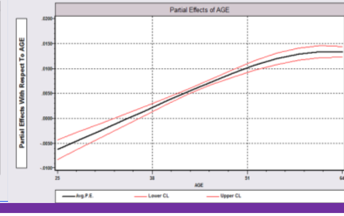
Table 2.17 Random Effects Model with Mundlak Correction

	Random Effects Probit LogL = -15424.40 LogL0 = -17365.76 7293 Individuals				Group Means Addition LogL = -15404.26 LogL0 = -17365.76				
Variable	Coef.	S.E.	t	P	Coef.	S.E.	t	P	Mean of X
Constant	.9459	.1116	8.473	.0000	.6551	.1232	5.320	.0000	43.5257
AGE	-.0365	.0015	-24.279	.0000	-.0521	.0036	-14.582	.0000	43.5257
EDUC	.0817	.0073	11.230	.0000	.0031	.0421	.073	.9415	11.3206
INCOME	.3207	.0717	4.474	.0000	.2937	.0959	3.064	.0022	.35208
MARRIED	.0188	.0346	.544	.5863	-.0429	.0534	-.803	.4220	.75862
KIDS	.0430	.0298	1.443	.1490	-.0019	.0397	-.048	.9614	.40273
AGEBAR					.0193	.0039	4.895	.0000	
EDUCBAR					.0790	.0427	1.848	.0646	
INCMBAR					.3451	.1496	2.307	.0211	
MARRBAR					.0499	.0717	.695	.4871	
KIDSBAR					.0936	.0616	1.520	.1285	
Rho	.5404	.0100	53.842	.0000	.5389	.0100	53.822	.0000	



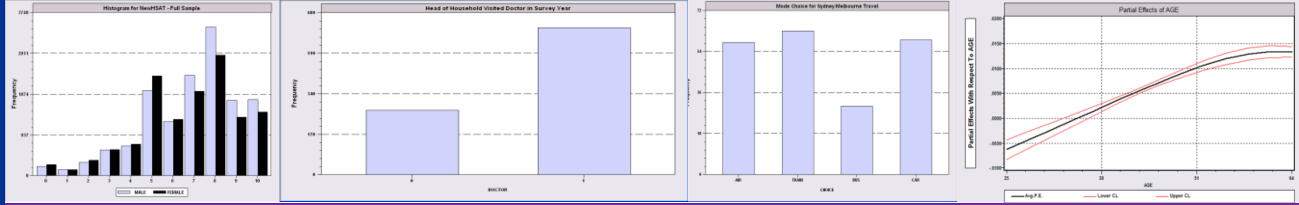
A Variable Addition Test for FE vs. RE

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

Journal of
APPLIED ECONOMETRICS

Research Article

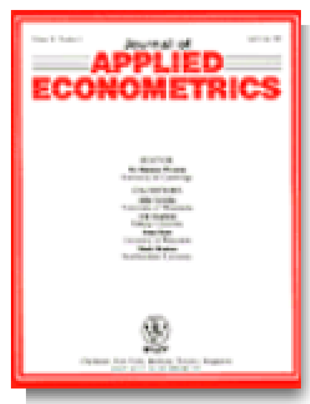
The dynamics of health in the British Household Panel Survey

Paul Contoyannis[1], Andrew M. Jones[2,*], Nigel Rice[3]

Article first published online: 9 AUG 2004

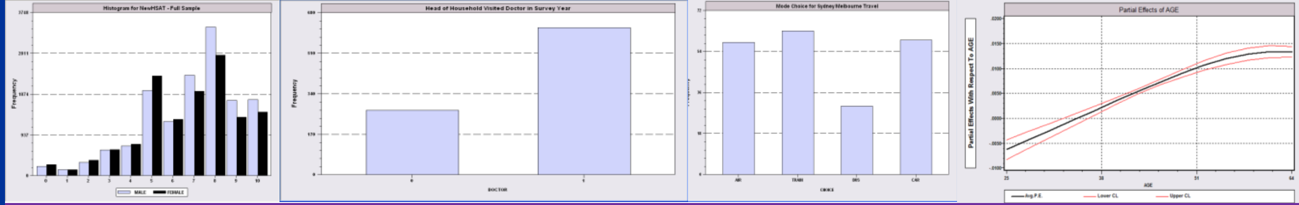
DOI: 10.1002/jae.755

Issue



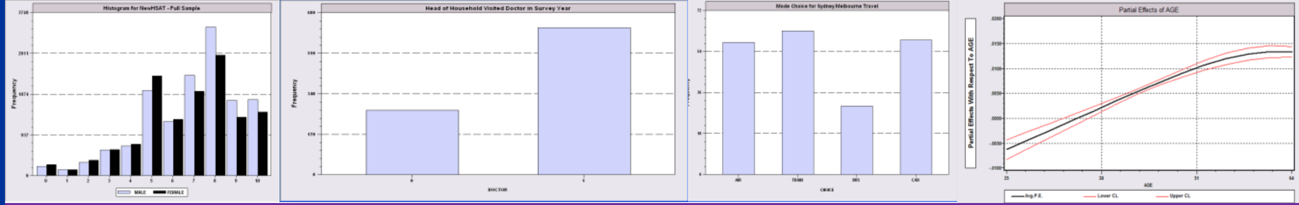
Journal of Applied Econometrics

Volume 19, Issue 4, pages 473–503, July/August 2004



Model for Self Assessed Health

- British Household Panel Survey (BHPS)
  - Waves 1-8, 1991-1998
  - 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

$$h^*_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \boxed{\boldsymbol{\gamma}'\mathbf{H}_{i,t-1}} + \alpha_i + \varepsilon_{it}$$

$\mathbf{x}_{it}$ = relevant covariates and control variables

$\mathbf{H}_{i,t-1}$ = 0/1 indicators of reported health status in previous period

$\mathrm{H}_{i,t-1}(j)$ = 1[Individual i reported $h_{it} = j$ in previous period], j=0,...,4

It would not be appropriate to include $h_{i,t-1}$ itself in the model as this is a label, not a measure

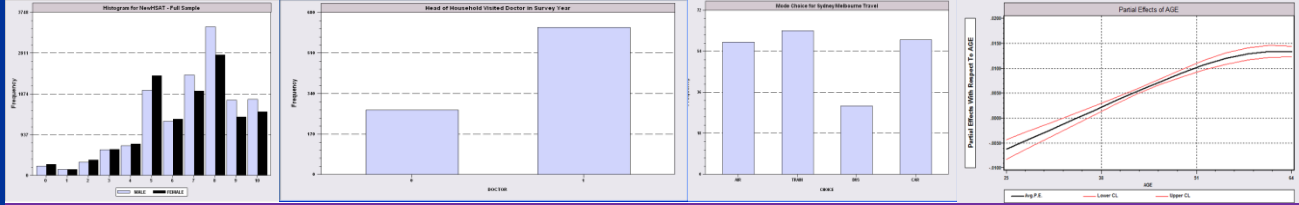
Ordered Choice Observation Mechanism

$$h_{it} = j \text{ if } \mu_{j-1} < h^*_{it} \leq \mu_j,\ j = 0,1,2,3,4$$

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

Random Effects with Mundlak Correction and Initial Conditions

$$\alpha_i = \alpha_0 + \boxed{\boldsymbol{\alpha}_1'\mathbf{H}_{i,1} + \boldsymbol{\alpha}_2'\overline{\mathbf{x}}_i} + u_i,\quad u_i \sim N[0,\sigma^2]$$



Random Effects Dynamic Ordered Probit Model

Random Effects Dynamic Ordered Probit Model

$$h_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) + \alpha_i + \varepsilon_{i,t}$$

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

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

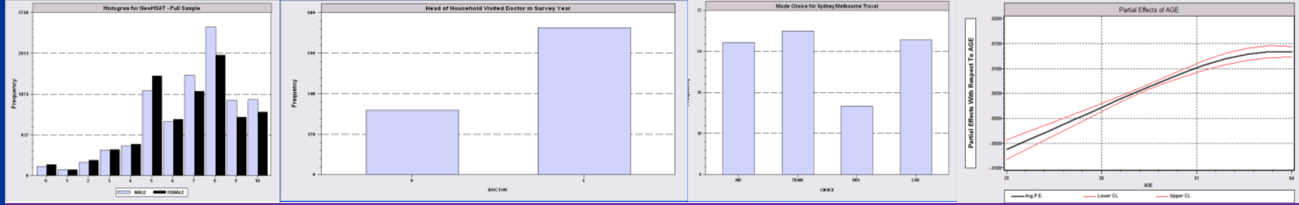
$$P_{it,j} = P[h_{it} = j] = \Phi(\mu_j - \mathbf{x}_{it}'\boldsymbol{\beta} - \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) - \alpha_i) - \Phi(\mu_{j-1} - \mathbf{x}_{it}'\boldsymbol{\beta} - \Sigma_{j=1}^{J}\gamma_j h_{i,t-1}(j) - \alpha_i)$$

Parameterize Random Effects

$$\alpha_i = \alpha_0 + \Sigma_{j=1}^{J}\alpha_{1,j} h_{i,1}(j) + \boldsymbol{\alpha}'\overline{\mathbf{x}}_i + u_i$$

Simulation or Quadrature Based Estimation

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



Data

Table I. Variable definitions

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

Variable of Interest

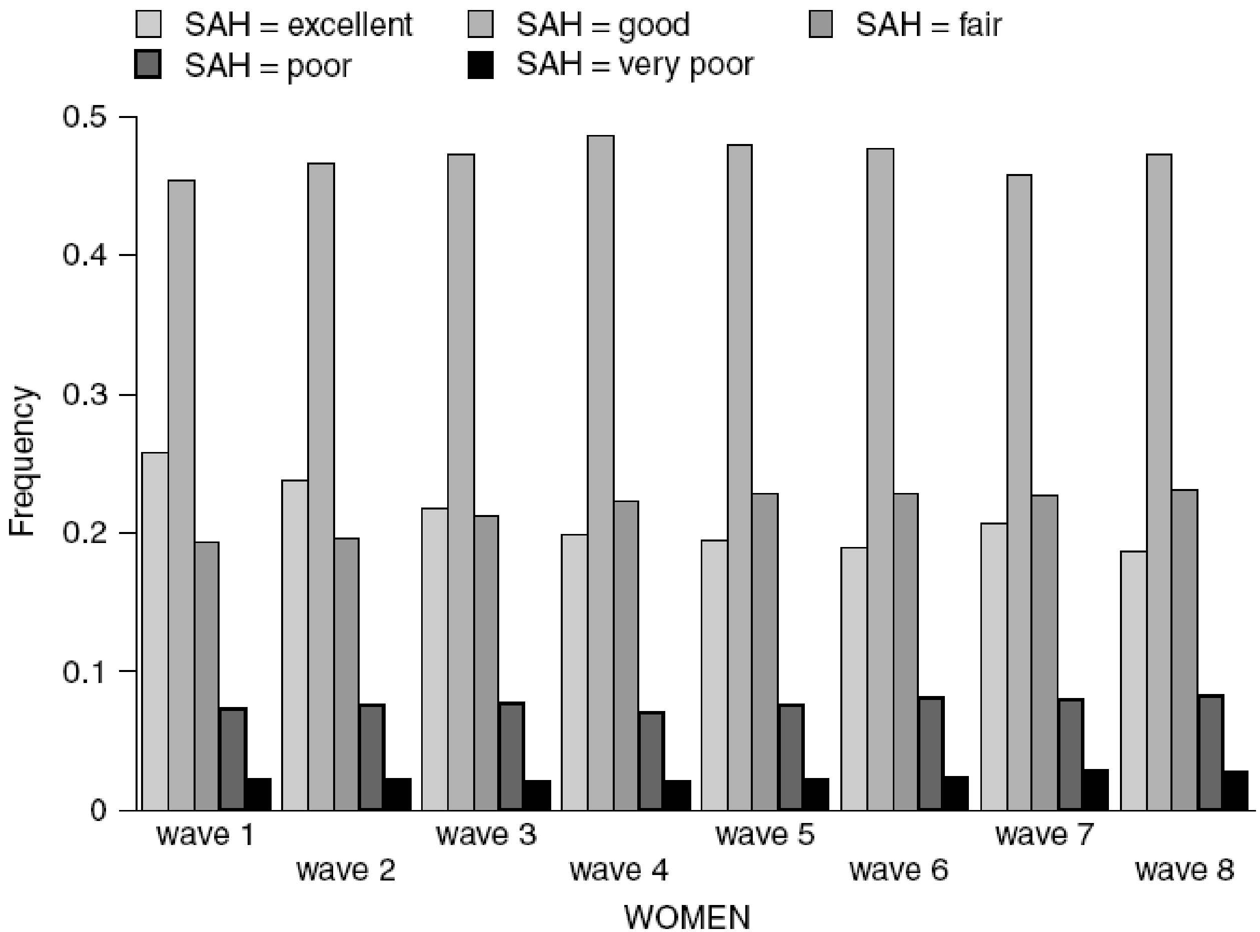
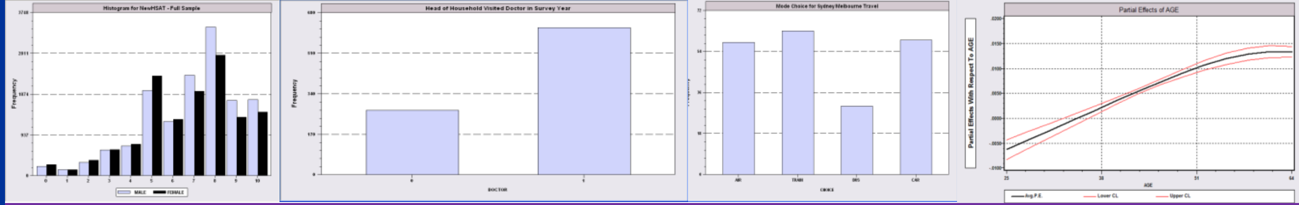


Figure 1. Self-assessed health status by wave



Dynamics

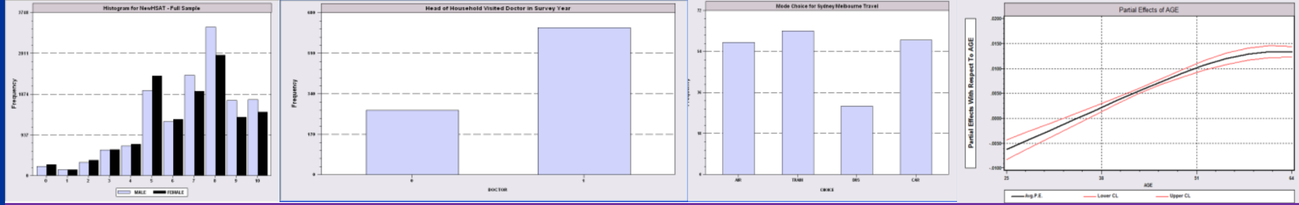
Table II. Transition matrices, balanced panel

(a) Men

SAH	EX	GOOD	FAIR	POOR	VERY POOR	*N*
EX	0.600	0.342	0.046	0.010	0.002	5485
GOOD	0.184	0.651	0.142	0.019	0.004	9263
FAIR	0.055	0.361	0.471	0.100	0.012	3433
POOR	0.029	0.120	0.340	0.418	0.093	1031
VERY POOR	0.032	0.073	0.133	0.423	0.339	248
N	5231	9287	3565	1111	266	19 460

(b) Women

SAH	EX	GOOD	FAIR	POOR	VERY POOR	*N*
EX	0.572	0.353	0.059	0.013	0.004	5164
GOOD	0.150	0.657	0.162	0.026	0.005	11 306
FAIR	0.040	0.362	0.465	0.116	0.017	4928
POOR	0.021	0.156	0.360	0.365	0.098	1587
VERY POOR	0.014	0.106	0.192	0.326	0.362	423
N	4884	11 329	5082	1649	464	23 408

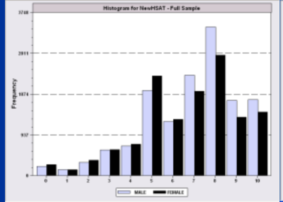
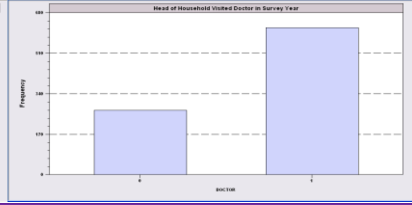
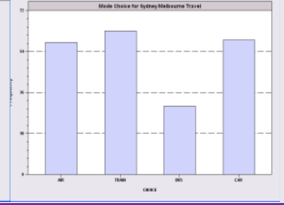
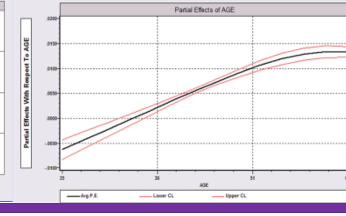


Attrition

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

(a) All data

	FULL SAMPLE				EX at $t-1$	GOOD at $t-1$	FAIR at $t-1$	POOR at $t-1$	VPOOR at $t-1$
Wave	No. individuals	Survival rate	Drop-outs	Attrition rate	Attrition rate	Attrition rate	Attrition rate	Attrition rate	Attrition rate
1	10 256								
2	8957	87.33%	1299	12.67%	11.54%	12.57%	13.01%	13.73%	23.74%
3	8162	79.58%	795	8.88%	8.08%	8.13%	9.65%	12.62%	19.46%
4	7825	76.30%	337	4.13%	6.67%	6.54%	6.73%	10.35%	14.74%
5	7430	72.45%	395	5.05%	6.21%	6.18%	7.87%	9.11%	16.34%
6	7238	70.57%	192	2.58%	3.11%	3.24%	5.06%	10.47%	18.83%
7	7102	69.25%	136	1.88%	3.15%	3.85%	4.79%	8.83%	8.75%
8	6839	66.68%	263	3.70%	3.43%	3.82%	5.30%	5.88%	17.01%

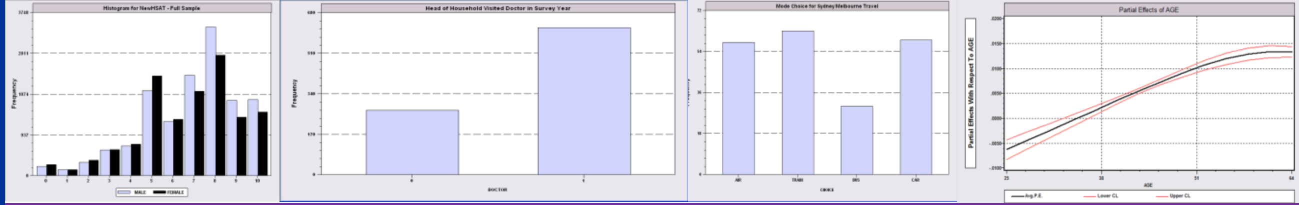





Testing for Attrition Bias

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

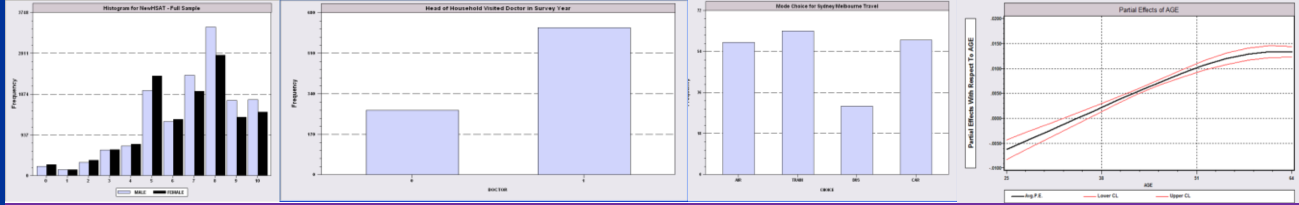
	MEN				WOMEN			
	β	Std.err.	t-test	p-value	β	Std.err.	t-test	p-value
NEXT WAVE	.199	.035	5.67	.000	.060	.034	1.77	.077
ALL WAVES	.139	.031	4.46	.000	.071	.029	2.45	.014
NUMBER OF WAVES	.031	.009	3.54	.000	.016	.008	1.88	.060

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

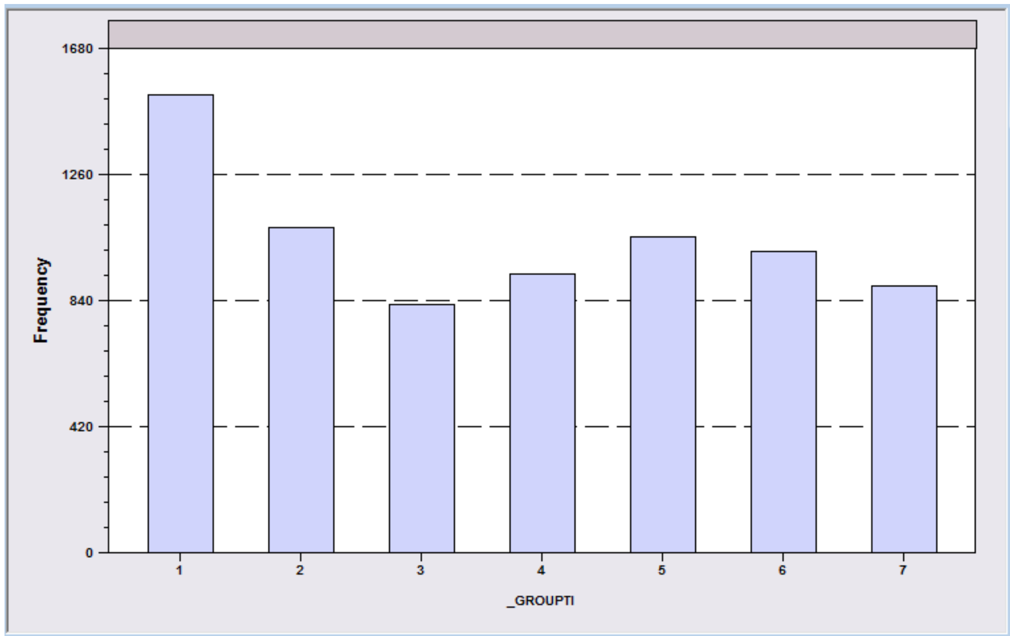


Probability Weighting Estimators

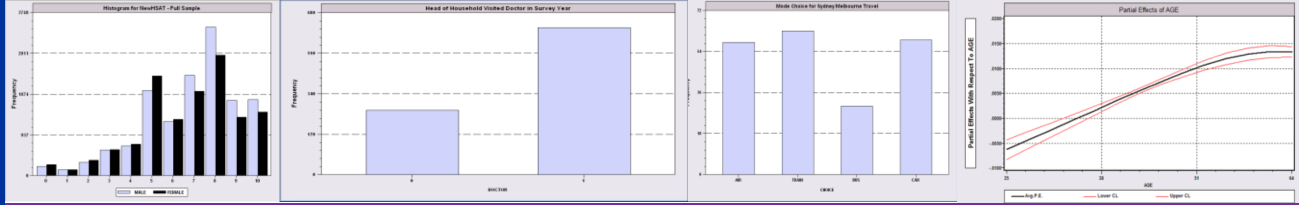
- A Patch for Attrition
- (1) Fit a participation probit equation for each wave.
- (2) Compute p(i,t) = predictions of participation for each individual in each period.
  - Special assumptions needed to make this work
- Ignore common effects and fit a weighted pooled log likelihood: $\Sigma_i\ \Sigma_t\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



Attrition Model with IP Weights



Assumes (1) Prob(attrition|all data) = 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 $\geq$ N2 $\geq$ ... $\geq$ N8.

Sample is restricted at each wave to individuals who were present at the previous wave.

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

$\text{d}_{i1} = 1 \;\; \forall \;\; \text{i, d}_{it} = 0 \;\Rightarrow\; \text{d}_{i,t+1} = 0.$

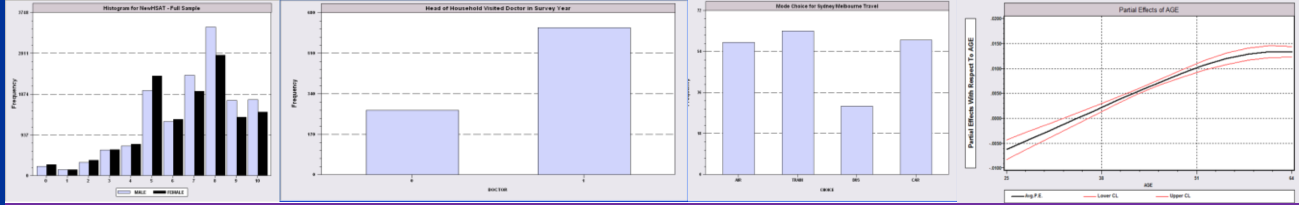
$\tilde{\mathbf{x}}_{i1}$ = covariates observed for all i at entry that relate to likelihood of being present at subsequent waves.
(health problems, disability, psychological well being, self employment, unemployment, maternity leave, student, caring for family member, ...)

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

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

Inverse probability weight $\hat{\text{W}}_{it} = \dfrac{\text{d}_{it}}{\hat{\text{P}}_{it}}$

Weighted log likelihood $\log\text{L}_W = \sum_{i=1}^{N} \sum_{t=1}^{8} \log L_{it}$ (No common effects.)

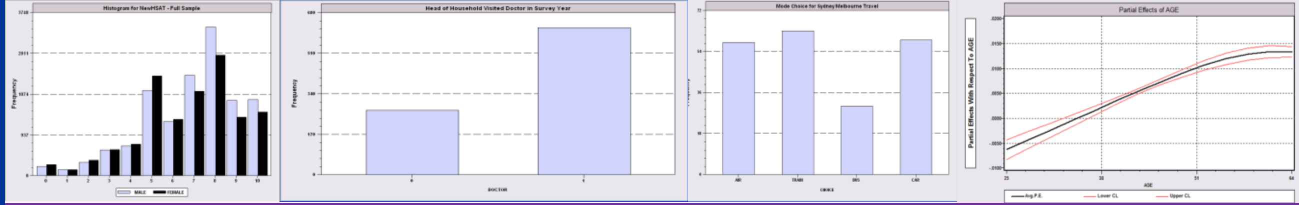


Estimated Partial Effects by Model

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

a) Men

	(1) Pooled model, balanced sample	(2) Pooled model, unbalanced sample	(3) Pooled model, IPW-1	(4) Pooled model, IPW-2	(5) Random effects, balanced sample	(6) Random effects, unbalanced sample
Ln(INCOME)	.009 (.004)	.009 (.004)	.009 (.004)	.011 (.005)	.013 (.006)	.012 (.005)
Mean Ln(INCOME)	.049 (.024)	.043 (.022)	.042 (.021)	.045 (.022)	.066 (.028)	.056 (.025)
DEGREE	.010 (.005)	.017 (.009)	.018 (.009)	.018 (.009)	.015 (.006)	.027 (.012)
HND/A	.019 (.009)	.021 (.011)	.021 (.010)	.022 (.011)	.028 (.011)	.030 (.013)
O/CSE	.016 (.008)	.020 (.010)	.020 (.010)	.020 (.010)	.024 (.010)	.028 (.012)
SAHEX(t-1)	.234 (.087)	.231 (.090)	.231 (.090)	.230 (.089)	.082 (.031)	.085 (.035)
SAHFAIR(t-1)	-.170 (.085)	-.163 (.084)	-.162 (.084)	-.162 (.083)	-.080 (.034)	-.077 (.036)
SAHPOOR(t-1)	-.242 (.167)	-.233 (.163)	-.232 (.162)	-.232 (.162)	-.151 (.077)	-.145 (.078)
SAHVPOOR(t-1)	-.260 (.198)	-.253 (.197)	-.255 (.199)	-.255 (.200)	-.184 (.104)	-.179 (.106)



Partial Effect for a Category

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

a) Men

	(1) Pooled model, balanced sample	(2) Pooled model, unbalanced sample	(3) Pooled model, IPW-1	(4) Pooled model, IPW-2	(5) Random effects, balanced sample	(6) Random effects, unbalanced sample
Ln(INCOME)	.009 (.004)	.009 (.004)	.009 (.004)	.011 (.005)	.013 (.006)	.012 (.005)
Mean Ln(INCOME)	.049 (.024)	.043 (.022)	.042 (.021)	.045 (.022)	.066 (.028)	.056 (.023)
DEGREE	.010 (.005)	.017 (.009)	.018 (.009)	.018 (.009)	.015 (.006)	.027 (.012)
HND/A	.019 (.009)	.021 (.011)	.021 (.010)	.022 (.011)	.028 (.011)	.030 (.013)
O/CSE	.016 (.008)	.020 (.010)	.020 (.010)	.020 (.010)	.024 (.010)	.028 (.012)
SAHEX(t-1)	.234 (.087)	.231 (.090)	.231 (.090)	.230 (.089)	.082 (.031)	.085 (.035)
SAHFAIR(t-1)	-.170 (.085)	-.163 (.084)	-.162 (.084)	-.162 (.083)	-.080 (.034)	-.077 (.036)
SAHPOOR(t-1)	-.242 (.167)	-.233 (.163)	-.232 (.162)	-.232 (.162)	-.151 (.077)	-.145 (.078)
SAHVPOOR(t-1)	-.260 (.198)	-.253 (.197)	-.255 (.199)	-.255 (.200)	-.184 (.104)	-.179 (.106)

SAHEX(t-1)	.234 (.087)
SAHFAIR(t-1)	-.170 (.085)
SAHPOOR(t-1)	-.242 (.167)
SAHVPOOR(t-1)	-.260 (.198)

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 (EX,POOR) change from (1,0) to (0,1).