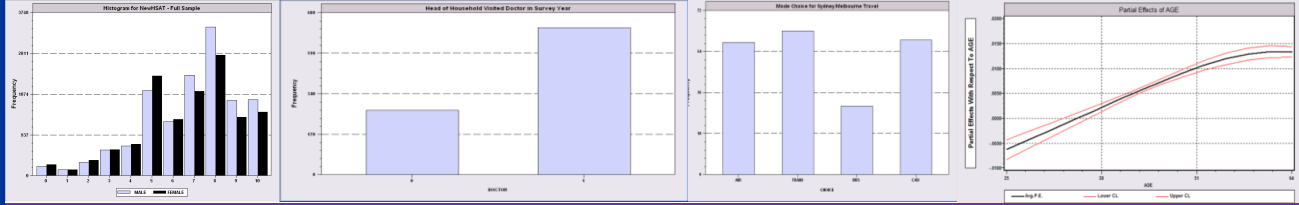
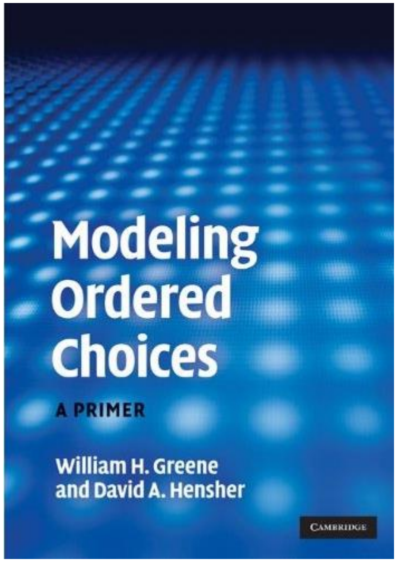
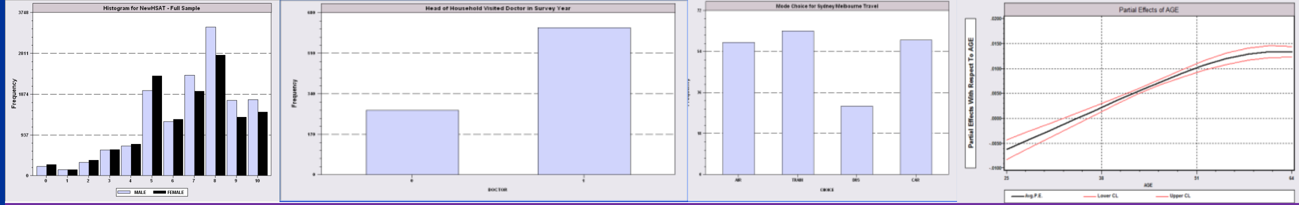


Discrete Choice Modeling

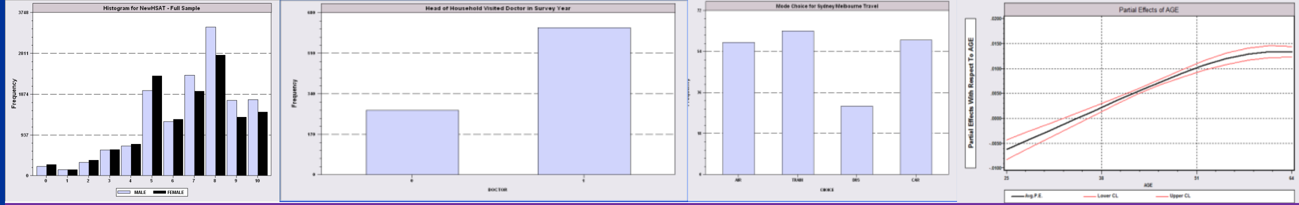
William Greene
Stern School of Business
New York University





Ordered Discrete Outcomes

- E.g.: Taste test, credit rating, course grade, preference scale
- Underlying random preferences:
  - Existence of an underlying continuous preference scale
  - Mapping to observed choices
- Strength of preferences is reflected in the discrete outcome
- Censoring and discrete measurement
- The nature of ordered data



Ordered Preferences at IMDB.com

User ratings for

National Treasure: Book of Secrets

Your Vote | vote

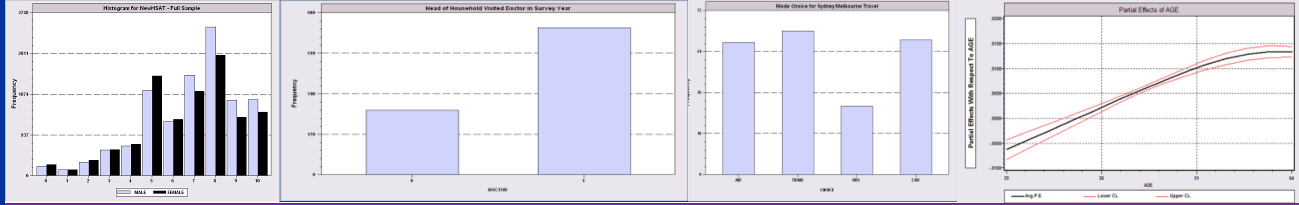
41,771 IMDb users have given a weighted average vote of 6.6 /

Demographic breakdowns are shown below.

Votes	Percentage	Rating
4,795	11.5%	10
3,286	7.9%	9
7,179	17.2%	8
10,636	25.5%	7
7,729	18.5%	6
3,646	8.7%	5
1,788	4.3%	4
940	2.3%	3
538	1.3%	2
1,234	3.0%	1

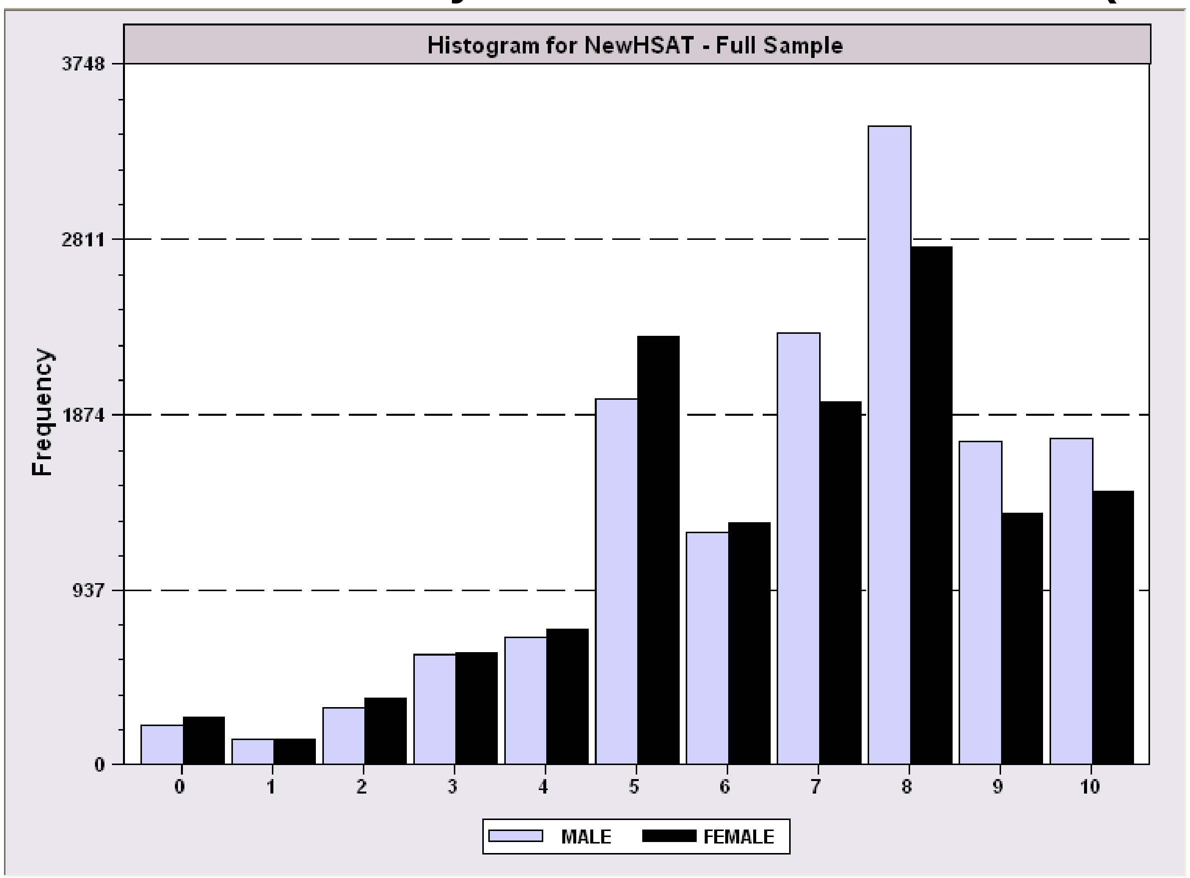
Arithmetic mean = 6.9. Median = 7

	Votes	Average
Males	33,644	6.5
Females	5,464	7.2
Aged under 18	2,492	7.6
Males under 18	1,795	7.5
Females under 18	695	8.1
Aged 18-29	26,045	6.7
Males Aged 18-29	22,603	6.6
Females Aged 18-29	3,372	7.3
Aged 30-44	8,210	6.3
Males Aged 30-44	7,216	6.3
Females Aged 30-44	936	6.7
Aged 45+	2,258	6.6
Males Aged 45+	1,814	6.5
Females Aged 45+	420	7.0
IMDb staff	8	6.4
Top 1000 voters	309	6.0
US users	14,792	6.8
Non-US users	24,283	6.5
IMDb users	41,771	6.6

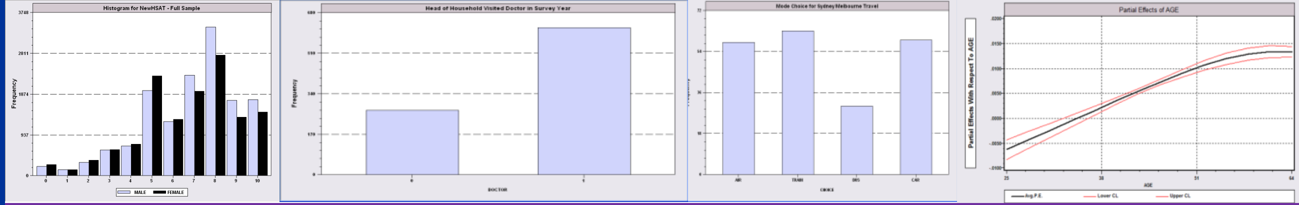


Health Satisfaction (HSAT)

Self administered survey: Health Care Satisfaction? (0 – 10)



Continuous Preference Scale



Modeling Ordered Choices

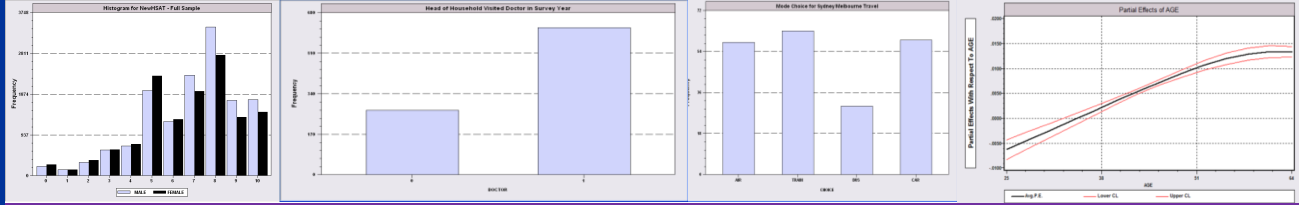
- Random Utility (allowing a panel data setting)

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

$$= a_{it} + \varepsilon_{it}$$

- Observe outcome j if utility is in region j
- Probability of outcome = probability of cell

$$\Pr[Y_{it}=j] = F(\mu_j - a_{it}) - F(\mu_{j-1} - a_{it})$$



Ordered Probability Model

$y^* = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$, we assume **x** contains a constant term

$y = 0$ if $y^* \leq 0$

$y = 1$ if $0 < y^* \leq \mu_1$

$y = 2$ if $\mu_1 < y^* \leq \mu_2$

$y = 3$ if $\mu_2 < y^* \leq \mu_3$

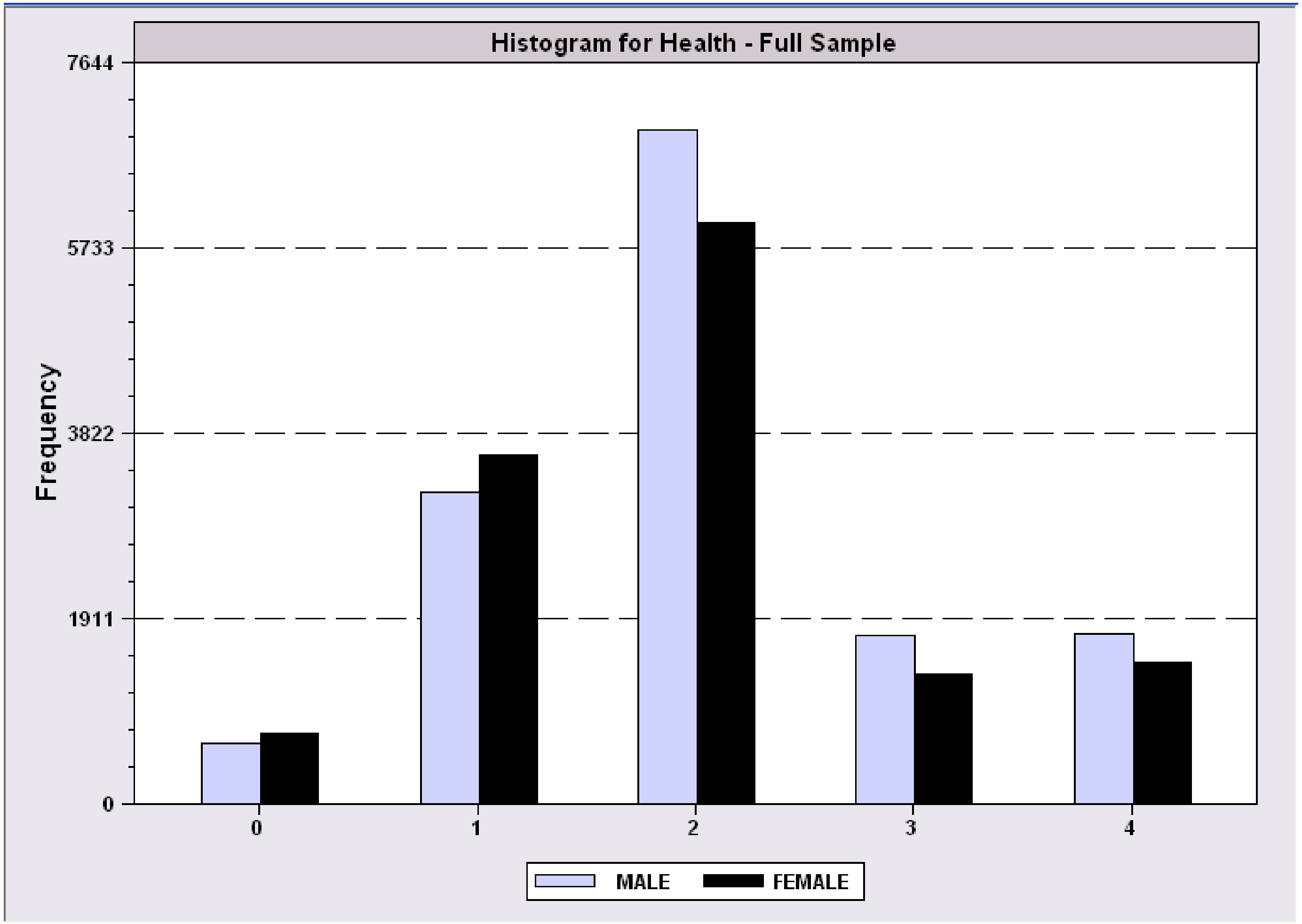
...

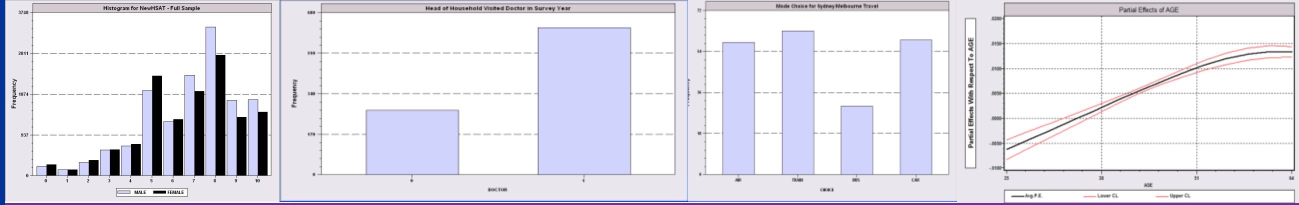
$y = J$ if $\mu_{J-1} < y^* \leq \mu_J$

In general: $y = j$ if $\mu_{j-1} < y^* \leq \mu_j$, $j = 0,1,\ldots,J$

$\mu_{-1} = -\infty,\ \mu_o = 0,\ \mu_J = +\infty,\ \mu_{j-1} < \mu_j,\ j = 1,\ldots,J$

Combined Outcomes for Health Satisfaction

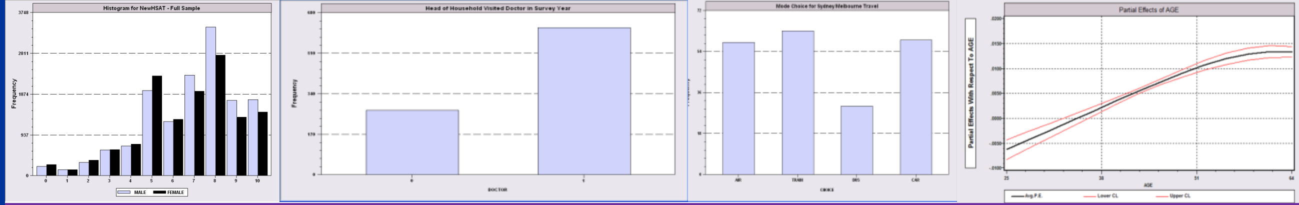


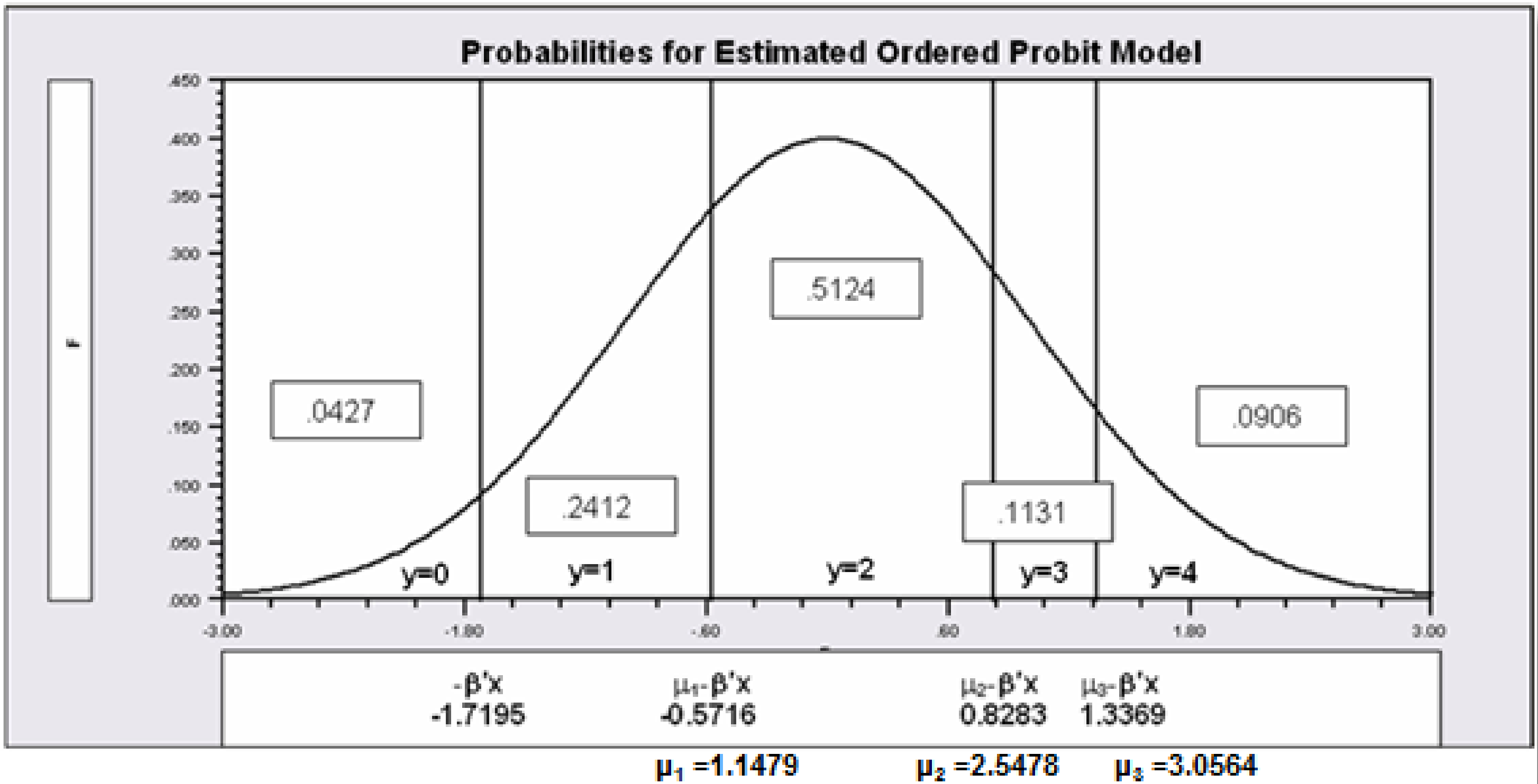


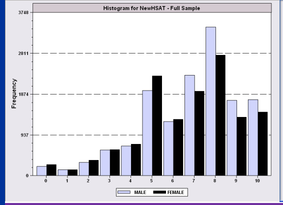
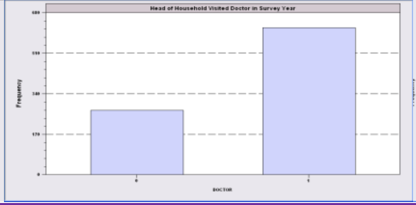
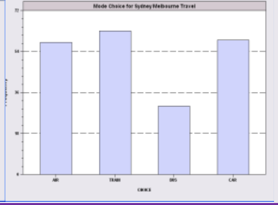
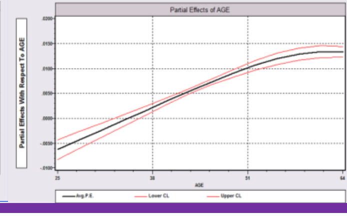
Ordered Probabilities

$$\begin{aligned}
\text{Prob}[y=j] = \text{Prob}[\mu_{j-1} < y^* \le \mu_j] \\
&= \text{Prob}[\mu_{j-1} < \boldsymbol{\beta}'\mathbf{x} + \varepsilon \le \mu_j] \\
&= \text{Prob}[\boldsymbol{\beta}'\mathbf{x} + \varepsilon \le \mu_j] - \text{Prob}[\boldsymbol{\beta}'\mathbf{x} + \varepsilon \le \mu_{j-1}] \\
&= \text{Prob}[\varepsilon \le \mu_j - \boldsymbol{\beta}'\mathbf{x}] - \text{Prob}[\varepsilon \le \mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}] \\
&= F[\mu_j - \boldsymbol{\beta}'\mathbf{x}] - F[\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}]
\end{aligned}$$

where $F[\varepsilon]$ is the CDF of ε.




Probabilities for Estimated Ordered Probit Model
.0427
.2412
.5124
.1131
.0906
y=0
y=1
y=2
y=3
y=4
-β'x
-1.7195
μ1-β'x
-0.5716
μ2-β'x
0.8283
μ3-β'x
1.3369
μ1 =1.1479
μ2 =2.5478
μ3 =3.0564

Coefficients

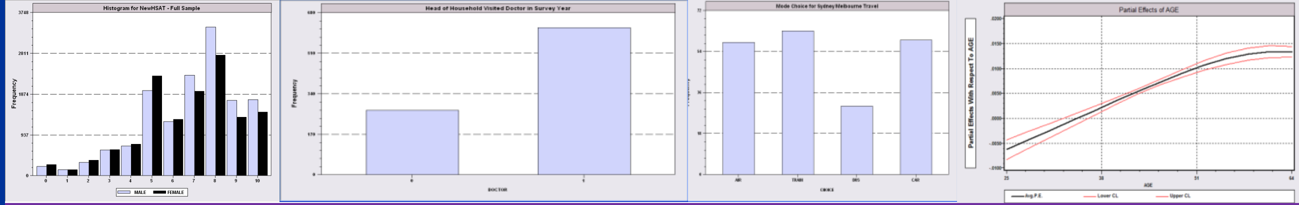
- What are the coefficients in the ordered probit model?

  There is no conditional mean function.

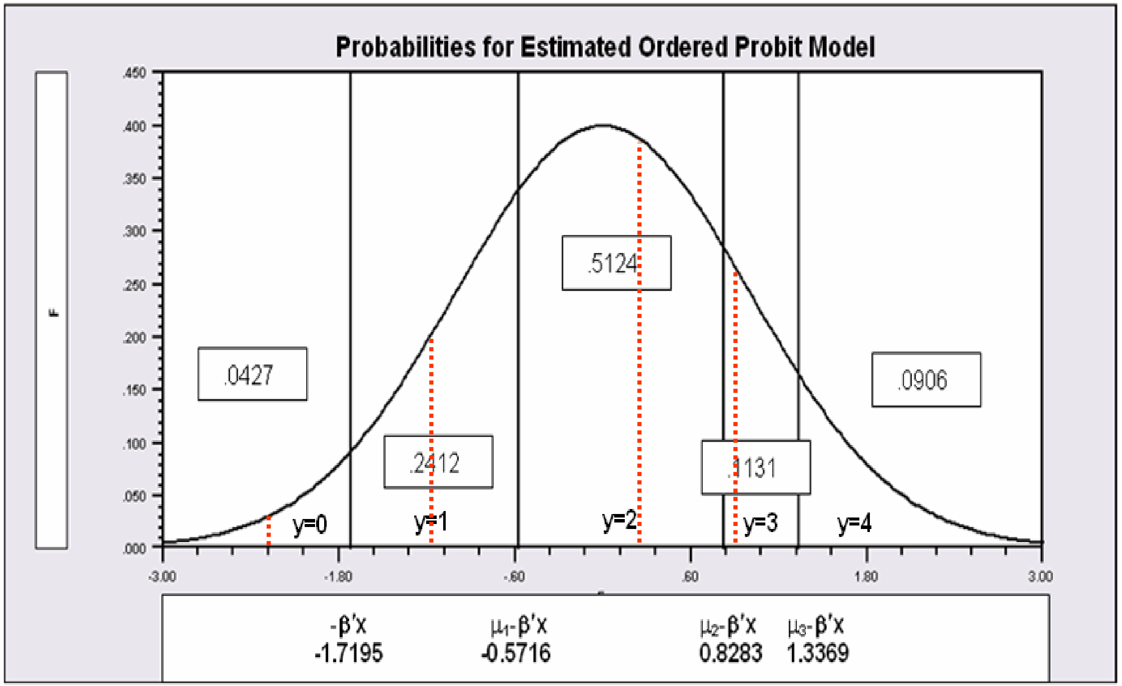
  $$\frac{\partial \text{Prob}[y=j|\mathbf{x}]}{\partial x_k} = [f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x})]\ \beta_k$$

  Magnitude depends on the scale factor and the coefficient.

  Sign depends on the densities at the two points!
- What does it mean that a coefficient is "significant?"



Partial Effects in the Ordered Choice Model



When $\beta_k > 0$, increase in x_k decreases Prob[y=0] and increases Prob[y=J]. Intermediate cells are ambiguous, but there is only one sign change in the marginal effects from 0 to 1 to … to J

Assume the β_k is positive.

Assume that x_k increases.

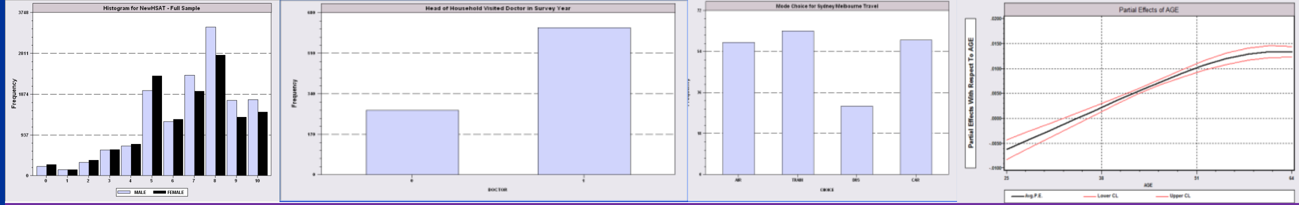
β'x increases. μ_j- **β'x** shifts to the left for all 5 cells.

Prob[y=0] decreases

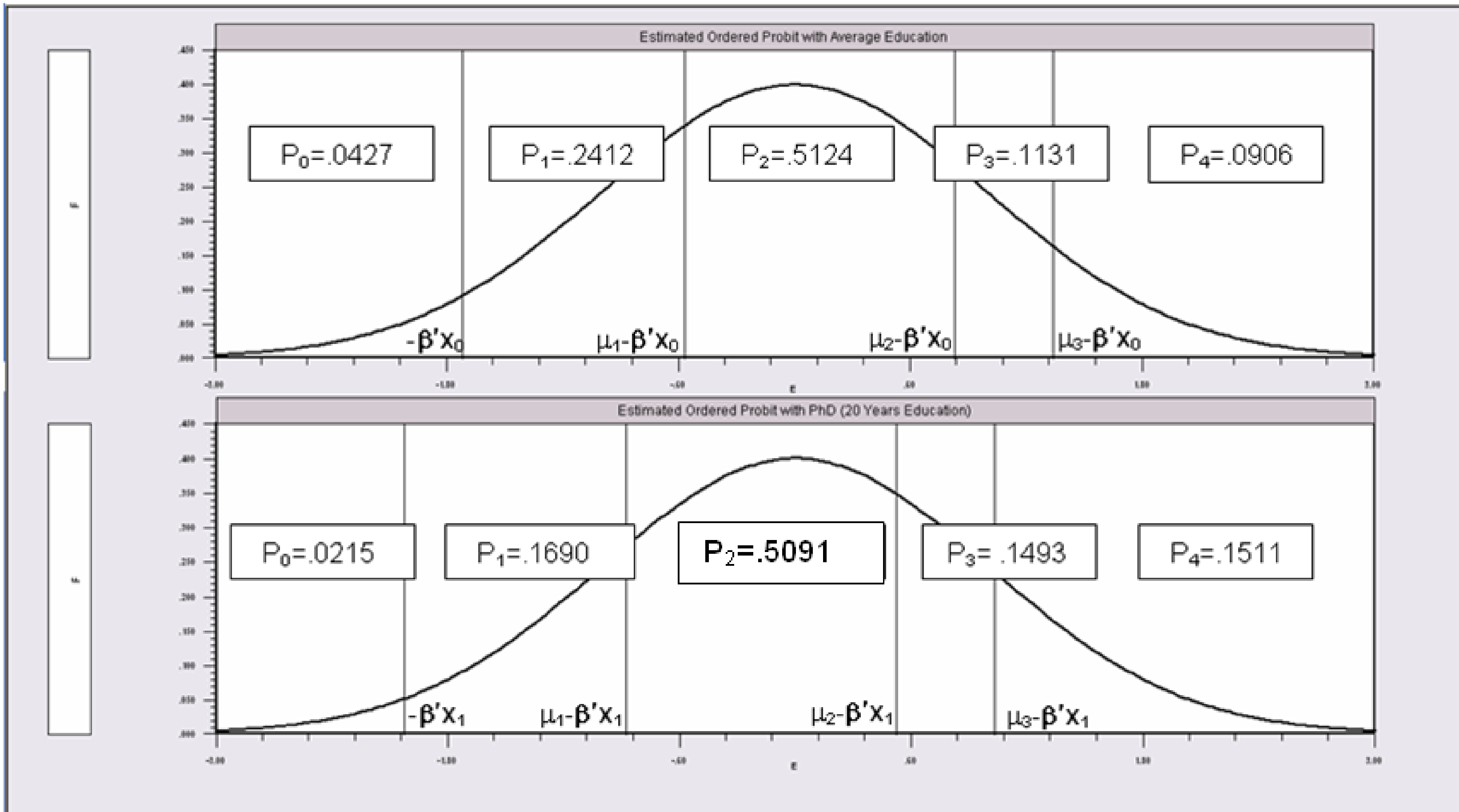
Prob[y=1] decreases – the mass shifted out is larger than the mass shifted in.

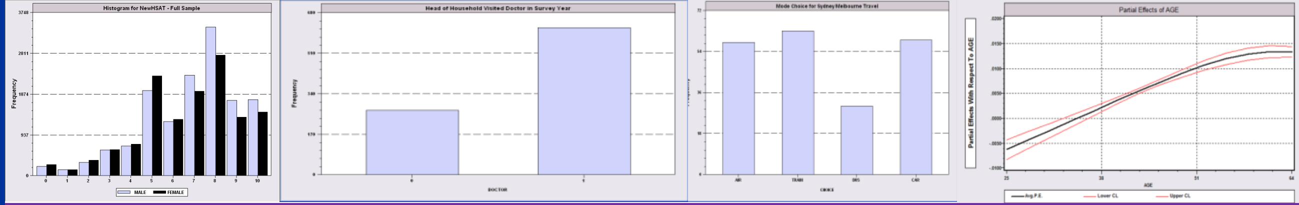
Prob[y=3] increases – same reason in reverse.

Prob[y=4] must increase.



Partial Effects of 8 Years of Education

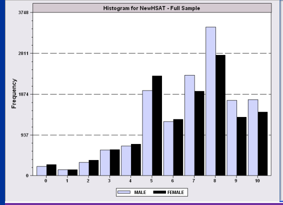
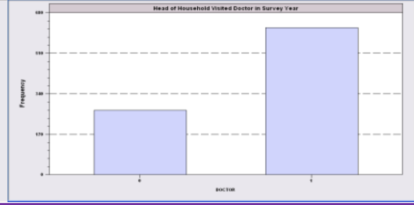
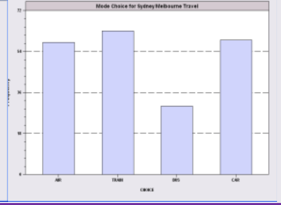
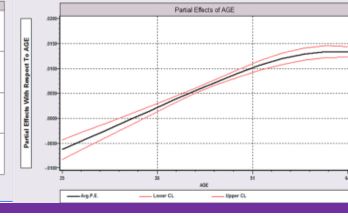




An Ordered Probability Model for Health Satisfaction

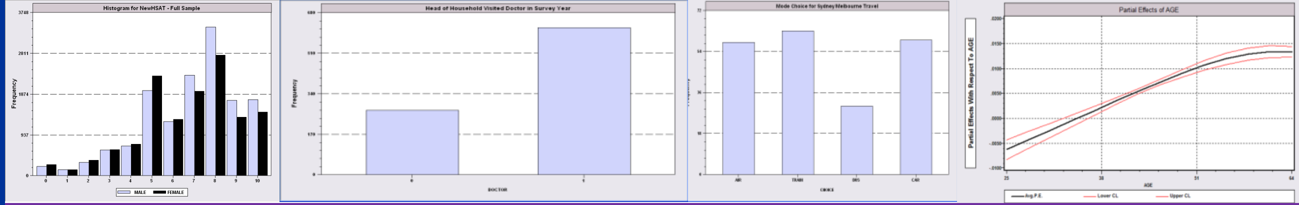
```
+---------------------------------------------+
| Ordered Probability Model                   |
| Dependent variable                HSAT      |
| Number of observations           27326      |
| Underlying probabilities based on Normal    |
|    Cell frequencies for outcomes            |
|  Y Count Freq  Y Count Freq  Y Count Freq   |
|  0   447 .016  1   255 .009  2   642 .023   |
|  3  1173 .042  4  1390 .050  5  4233 .154   |
|  6  2530 .092  7  4231 .154  8  6172 .225   |
|  9  3061 .112 10  3192 .116                 |
+---------------------------------------------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Index function for probability					
Constant	2.61335825	.04658496	56.099	.0000	
FEMALE	-.05840486	.01259442	-4.637	.0000	.47877479
EDUC	.03390552	.00284332	11.925	.0000	11.3206310
AGE	-.01997327	.00059487	-33.576	.0000	43.5256898
HHNINC	.25914964	.03631951	7.135	.0000	.35208362
HHKIDS	.06314906	.01350176	4.677	.0000	.40273000
Threshold parameters for index					
Mu(1)	.19352076	.01002714	19.300	.0000	
Mu(2)	.49955053	.01087525	45.935	.0000	
Mu(3)	.83593441	.00990420	84.402	.0000	
Mu(4)	1.10524187	.00908506	121.655	.0000	
Mu(5)	1.66256620	.00801113	207.532	.0000	
Mu(6)	1.92729096	.00774122	248.965	.0000	
Mu(7)	2.33879408	.00777041	300.987	.0000	
Mu(8)	2.99432165	.00851090	351.822	.0000	
Mu(9)	3.45366015	.01017554	339.408	.0000	



Ordered Probability Partial Effects

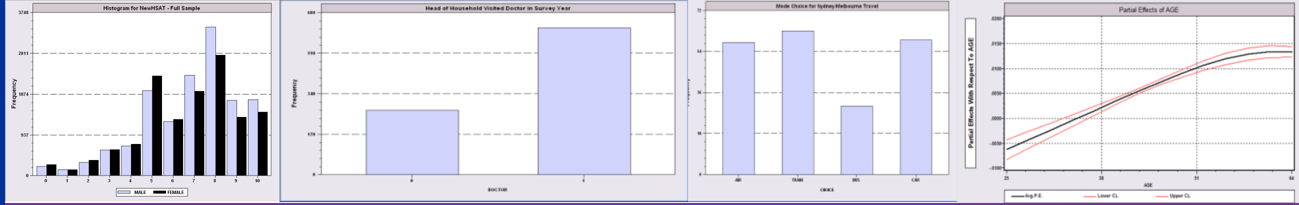
```
+-----------------------------------------------------+
| Marginal effects for ordered probability model      |
| M.E.s for dummy variables are Pr[y|x=1]-Pr[y|x=0]   |
| Names for dummy variables are marked by *.          |
+-----------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
         These are the effects on Prob[Y=00] at means.
 *FEMALE        .00200414       .00043473      4.610   .0000     .47877479
 EDUC          -.00115962     .986135D-04    -11.759   .0000    11.3206310
 AGE            .00068311     .224205D-04     30.468   .0000    43.5256898
 HHNINC        -.00886328       .00124869     -7.098   .0000     .35208362
 *HHKIDS       -.00213193       .00045119     -4.725   .0000     .40273000
         These are the effects on Prob[Y=01] at means.
 *FEMALE        .00101533       .00021973      4.621   .0000     .47877479
 EDUC          -.00058810     .496973D-04    -11.834   .0000    11.3206310
 AGE            .00034644     .108937D-04     31.802   .0000    43.5256898
 HHNINC        -.00449505       .00063180     -7.115   .0000     .35208362
 *HHKIDS       -.00108460       .00022994     -4.717   .0000     .40273000
 ... repeated for all 11 outcomes
         These are the effects on Prob[Y=10] at means.
 *FEMALE       -.01082419       .00233746     -4.631   .0000     .47877479
 EDUC           .00629289       .00053706     11.717   .0000    11.3206310
 AGE           -.00370705       .00012547    -29.545   .0000    43.5256898
 HHNINC         .04809836       .00678434      7.090   .0000     .35208362
 *HHKIDS        .01181070       .00255177      4.628   .0000     .40273000
```



Ordered Probit Marginal Effects

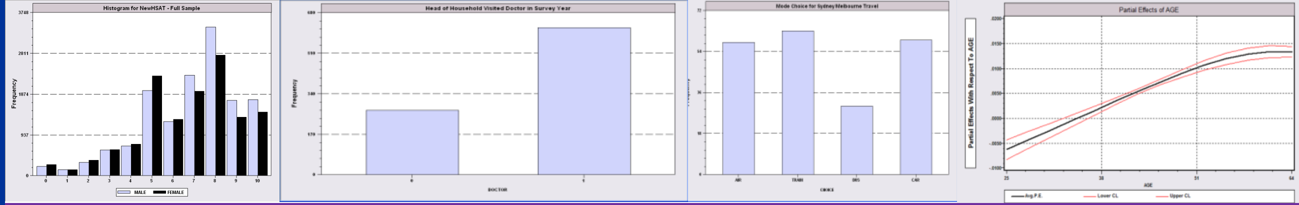
Summary of Marginal Effects for Ordered Probability Model (probit)

Variable	Y=00	Y=01	Y=02	Y=03	Y=04	Y=05	Y=06	Y=07
	Y=08	Y=09	Y=10	Y=11	Y=12	Y=13	Y=14	Y=15
*FEMALE	.0020	.0010	.0023	.0037	.0036	.0074	.0024	.0008
	-.0059	-.0064	-.0108					
EDUC	-.0012	-.0006	-.0013	-.0021	-.0021	-.0043	-.0014	-.0005
	.0034	.0037	.0063					
AGE	.0007	.0003	.0008	.0012	.0012	.0025	.0008	.0003
	-.0020	-.0022	-.0037					
HHNINC	-.0089	-.0045	-.0103	-.0162	-.0159	-.0329	-.0105	-.0035
	.0262	.0283	.0481					
*HHKIDS	-.0021	-.0011	-.0025	-.0039	-.0039	-.0080	-.0026	-.0009
	.0063	.0069	.0118					

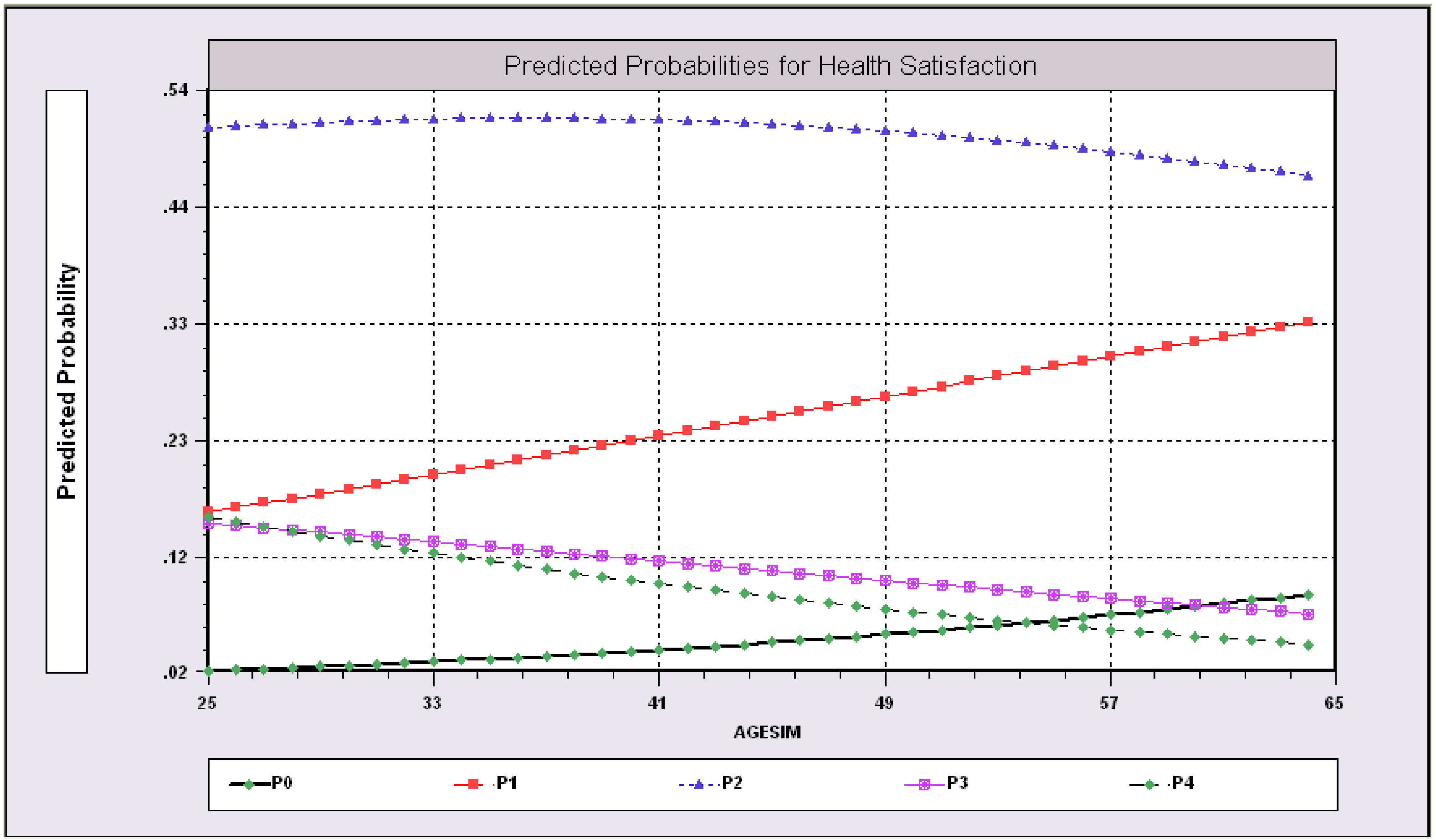


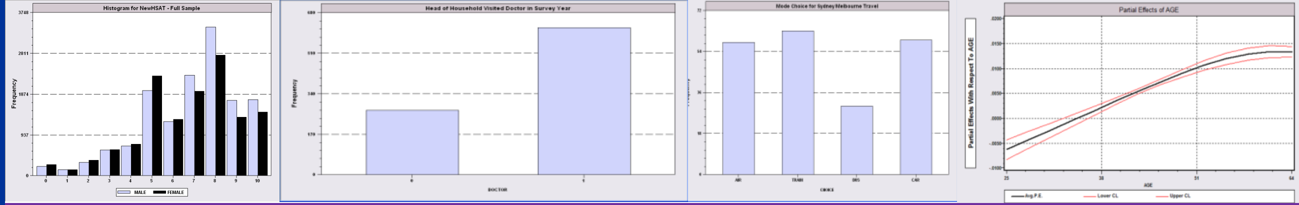
Analysis of Model Implications

- Partial Effects
- Fit Measures
- Predicted Probabilities
  - Averaged: They match sample proportions.
  - By observation
  - Segments of the sample
  - Related to particular variables



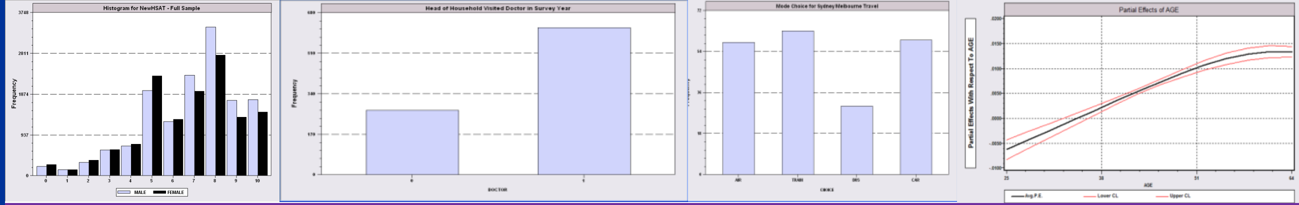
Predictions from the Model Related to Age





Fit Measures

- There is no single “dependent variable” to explain.
- There is no sum of squares or other measure of “variation” to explain.
- Predictions of the model relate to a set of J+1 probabilities, not a single variable.
- How to explain fit?
  - Based on the underlying regression
  - Based on the likelihood function
  - Based on prediction of the outcome variable



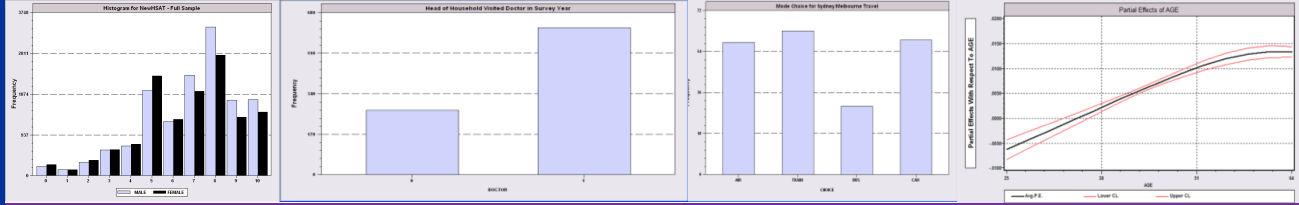
Log Likelihood Based Fit Measures

$$R_{Pseudo}^{2} = 1 - \log L_{Model} / \log L_{No\ Model}.$$

A degrees of freedom adjusted version is sometimes reported,

$$Adjusted\ R_{Pseudo}^{2} = 1 - [\log L_{No\ Model} - M] / \log L_{Model},$$

$$\begin{aligned}
&\textit{Log Akaike Information Criterion} &&= AIC &&= (-2\log L + 2M)/n,\\
&\textit{Finite Sample AIC} &&= AIC_{FS} &&= AIC + 2M(M+1)/(n - M - 1),\\
&\textit{Bayes Information Criterion} &&= BIC &&= (-2\log L + M/\log n)/n\\
&\textit{Hannan-Quinn IC} &&= HQIC &&= (-2\log L + 2\,M\,\log\log n)/n.
\end{aligned}$$



$$Count\ R^2 = \frac{Number\ of\ Correct\ Predictions}{n}$$

and

$$Adjusted\ Count\ R^2 = \frac{Number\ of\ Correct\ Predictions - n_j*}{n - n_j*},$$

where n_j* is the count of the most frequent outcome.

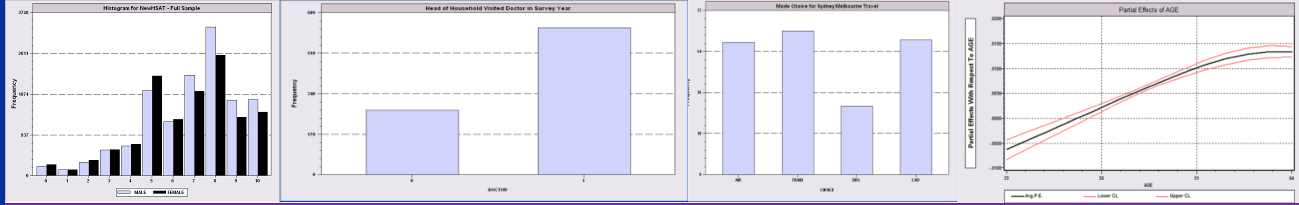
$$\hat{y}_i = j*\ \ such\ that\ estimated$$
$$\text{Prob}(y_i = j*|\mathbf{x}_i) > \ estimated\ \text{Pr}(y_i = j|\mathbf{x}_i)\ \forall\ j \neq j*$$

That is, put the predicted y in the cell with the highest probability.

Predicted vs. Actual Outcomes for Ordered Probit Model

Cross tabulation of predictions.
Row is actual, column is predicted.
Model=Probit. Prediction=most likely cell.

Actual	0	1	2	3	4	Row Sum
0	0	0	230	0	0	220
1	0	0	1113	0	0	1113
2	0	0	2226	0	0	2226
3	0	0	500	0	0	500
4	0	0	414	0	0	414
Col Sum	0	0	4483	0	0	4483

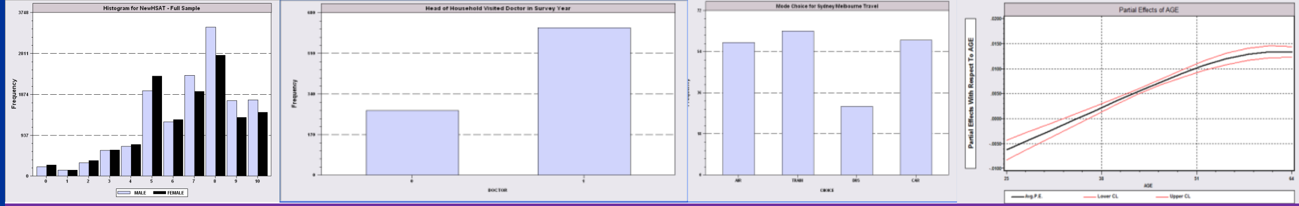


A Somewhat Better Fit

www.stata-press.com/data/r8/fullauto.dta
1977 repair records of 66 foreign and domestic cars.
The variable *rep77* takes values *poor, fair, average, good* and *excellent*.
Explanatory variables in the model are *foreign* (origin of manufacture), *length* (a proxy for size) and *mpg*.
The McFadden *Pseudo* R^2 is 0.1321.
The *Count* R^2 is (1+0+21+7+1)/66 = 0.454.
The adjusted value is (30 – 27)/(66-27) = 0.077.

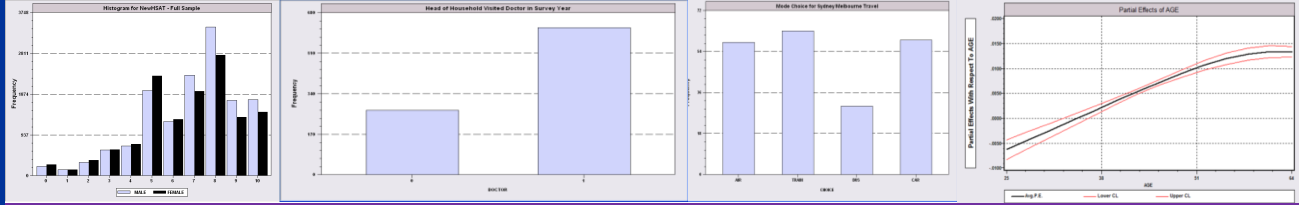
Cross tabulation of predictions.
Row is actual, column is predicted.
Model=Probit. Prediction=most likely cell.

Actual	0	1	2	3	4	Row sum
0	1	0	2	0	0	3
1	0	0	9	2	0	11
2	0	1	21	5	0	27
3	0	0	11	7	2	20
4	0	0	2	2	1	5
Col Sum	1	1	45	16	3	66



Different Normalizations

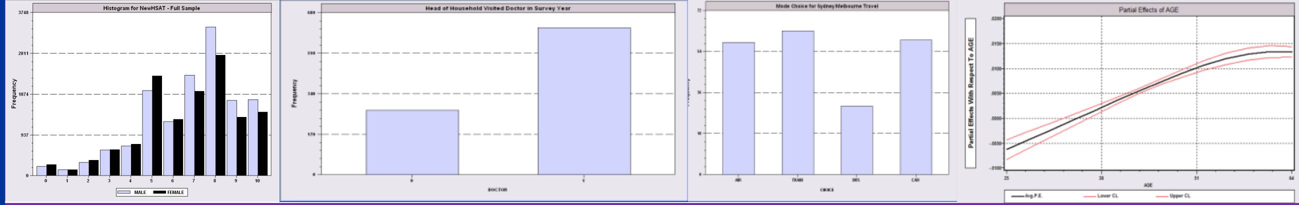
- NLOGIT
  - $Y = 0,1,\ldots,J$, $U^* = \alpha + \beta'x + \varepsilon$
  - One overall constant term, α
  - J-1 "cutpoints;" $\mu_{-1} = -\infty$, $\mu_0 = 0$, $\mu_1,\ldots \mu_{J-1}$, $\mu_J = +\infty$
- Stata
  - $Y = 1,\ldots,J+1$, $U^* = \beta'x + \varepsilon$
  - No overall constant, $\alpha=0$
  - J "cutpoints;" $\mu_0 = -\infty$, $\mu_1,\ldots \mu_J$, $\mu_{J+1} = +\infty$




```
(The data on rep77 contain 8 missing observations)
Ordered Probit ; Lhs = rep77 ; Rhs=one,foreign,length,mpg $
+---------------------------------------------+
| Ordered Probability Model                   |
| Dependent variable               REP77      |
| Number of observations              66      |
| Log likelihood function      -78.02002      |
| Number of parameters                 7      |
| Info. Criterion: AIC =         2.57636      |
| Restricted log likelihood    -89.89510      |
| McFadden Pseudo R-squared     .1320992      |
| Chi squared                   23.75015      |
| Degrees of freedom                   3      |
| Prob[ChiSqd > value] =        .2816655E-04  |
| Underlying probabilities based on Normal    |
+---------------------------------------------+
+--------+--------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |b/St.Er.|P[|Z|>z]| Mean of X|
+--------+--------------+----------------+--------+--------+----------+
---------+Index function for probability
 Constant|   -10.1589039      3.03379286    -3.349   .0008
 FOREIGN |    1.70486053      .41520516      4.106   .0000    .31818182
 LENGTH  |     .04686753      .01228262      3.816   .0001   189.121212
 MPG     |     .13045591      .03696460      3.529   .0004   21.3333333
---------+Threshold parameters for index
 Mu(1)   |    1.05112609      .18720281      5.615   .0000
 Mu(2)   |    2.38670648      .18420739     12.957   .0000
 Mu(3)   |    3.82169002      .28935433     13.208   .0000
```

$\hat{\alpha}$

$\hat{\mu}_j$

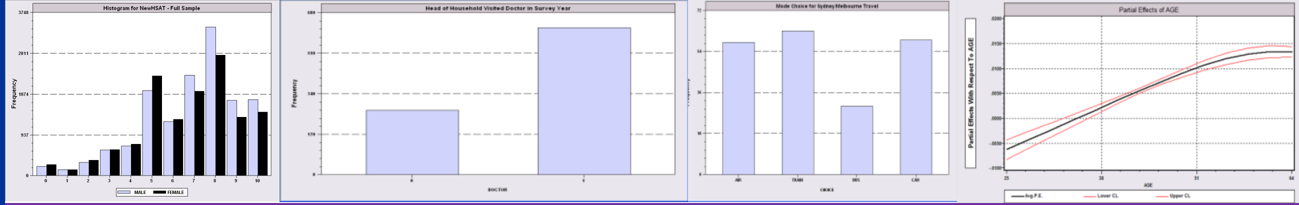


Stata and *NLOGIT* Estimates of an Ordered Probit Model

```
. oprobit rep77 foreign length mpg
Iteration 0:   log likelihood = -89.895098
Iteration 1:   log likelihood = -78.141221
Iteration 2:   log likelihood = -78.020314
Iteration 3:   log likelihood = -78.020025
Ordered probit regression                          Number of obs   =        66
                                                   LR chi2(3)      =     23.75
                                                   Prob > chi2     =    0.0000
Log likelihood = -78.020025                        Pseudo R2       =    0.1321
------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   1.704861   .4246786     4.01    0.000      .8725057    2.537215
      length |   .0468675    .012648     3.71    0.000       .022078    .0716571
         mpg |   .1304559   .0378627     3.45    0.001      .0562464    .2046654
-------------+----------------------------------------------------------------
       /cut1 |    10.1589   3.076749                        4.128586    16.18922
       /cut2 |   11.21003   3.107522                        5.119399    17.30066
       /cut3 |   12.54561   3.155228                        6.361476    18.72974
       /cut4 |   13.98059   3.218786                        7.671888     20.2893
------------------------------------------------------------------------------
```

$-\hat{\alpha}$

$\hat{\mu}_j - \hat{\alpha}$



Generalizing the Ordered Probit with Heterogeneous Thresholds

Index $= \boldsymbol{\beta}'\mathbf{x}_i$

Threshold parameters

Standard model : $\mu_{-1} = -\infty, \mu_0 = 0, \mu_j > \mu_{j-1} > 0, \mu_J = +\infty$

Preference scale and thresholds are homogeneous

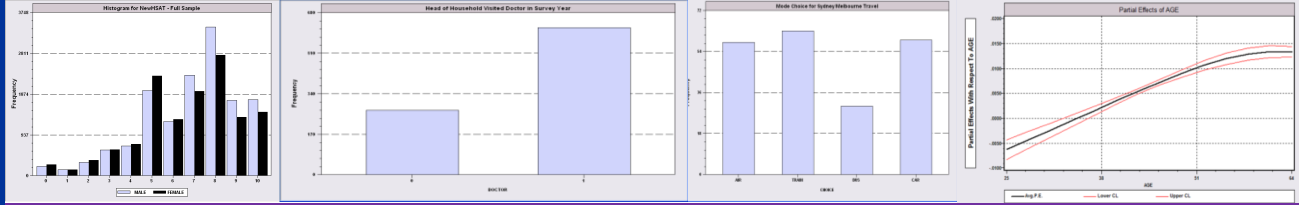
A generalized model (Pudney and Shields, JAE, 2000)

$\mu_{ij} = \alpha_j + \boldsymbol{\gamma}_j'\mathbf{z}_i$

Note the identification problem. If z_{ik} is also in x_i (same variable)

then $\mu_{ij} - \boldsymbol{\beta}'\mathbf{x}_i = \alpha_j + \gamma z_{ik} - \beta z_{ik} + \ldots$ No longer clear if the variable

is in **x** or **z** (or both)



Hierarchical Ordered Probit

Index $= \boldsymbol{\beta}'\mathbf{x}_i$

Threshold parameters

Standard model : $\mu_{-1} = -\infty, \mu_0 = 0, \mu_j > \mu_{j-1} > 0, \mu_J = +\infty$

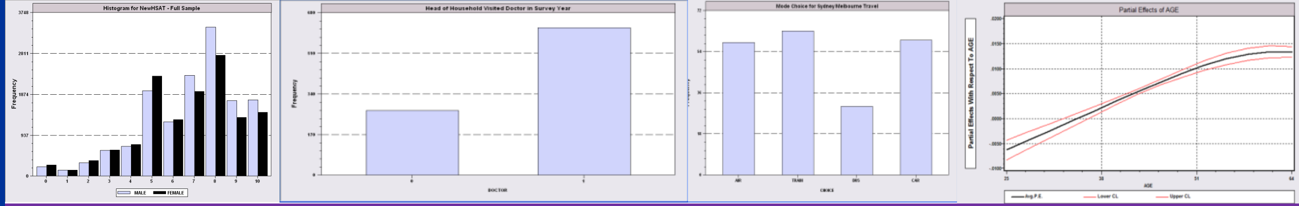
Preference scale and thresholds are homogeneous

A generalized model (Harris and Zhao (2000), NLOGIT (2007))

$$\mu_{ij} = \exp[\alpha_j + \boldsymbol{\gamma}_j'\mathbf{z}_i]$$

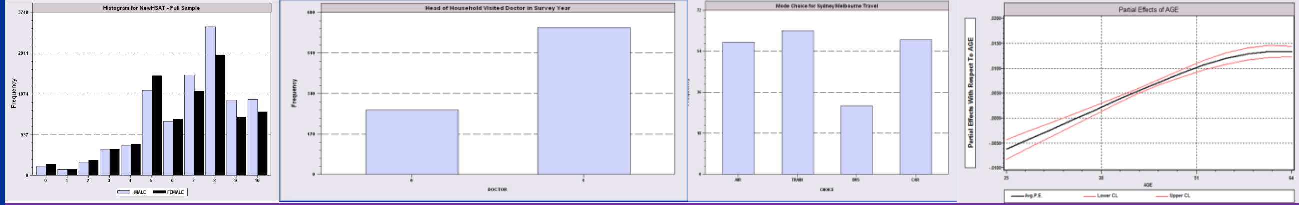
An internally consistent restricted modification

$$\mu_{ij} = \exp[\alpha_j + \boldsymbol{\gamma}'\mathbf{z}_i], \ \alpha_j = \alpha_{j-1} + \exp(\theta_j)$$



Ordered Choice Model

```
+---------------------------------------------+
| Ordered Probability Model                   |
| Dependent variable                 HSAT     |
| Weighting variable                 None     |
| Number of observations             1939     |
| Iterations completed                 14     |
| Log likelihood function       -2622.995     |
| Number of parameters                  9     |
| Info. Criterion: AIC =          2.71480     |
|   Finite Sample: AIC =          2.71484     |
| Info. Criterion: BIC =          2.74065     |
| Info. Criterion:HQIC =          2.72430     |
| Restricted log likelihood     -2634.772     |
| Chi squared                    23.55427     |
| Degrees of freedom                    4     |
| Prob[ChiSqd > value] =      .9810369E-04 |
| Underlying probabilities based on Normal    |
|    Cell frequencies for outcomes            |
|  Y Count Freq  Y Count Freq  Y Count Freq   |
|  0    89 .045  1    55 .028  2   158 .081   |
|  3   267 .137  4   336 .173  5  1034 .533   |
+---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Index function for probability
 Constant      1.98785052       .23413738     8.490   .0000
 AGE           -.01237740       .00293176    -4.222   .0000    46.3140794
 EDUC           .01798743       .01553417     1.158   .2469    10.5102719
 FEMALE         .09599181       .05307917     1.808   .0705     .52037133
 HHNINC         .13604164       .18479073      .736   .4616     .33073435
          Threshold parameters for index
 Mu(1)          .24287214       .02704191     8.981   .0000
 Mu(2)          .67874854       .03076131    22.065   .0000
 Mu(3)         1.15094430       .02944830    39.084   .0000
 Mu(4)         1.61429661       .03187311    50.648   .0000
```

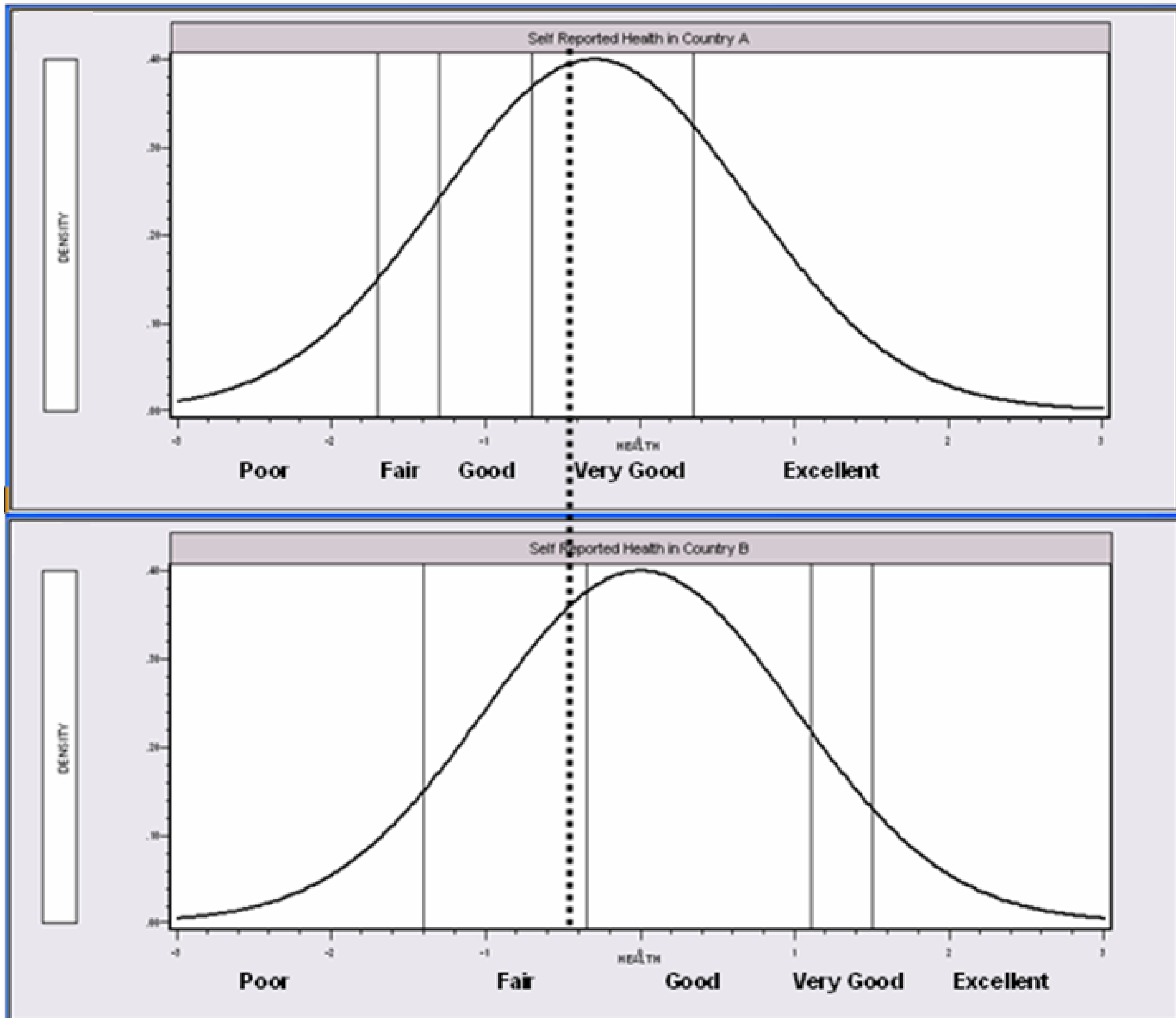


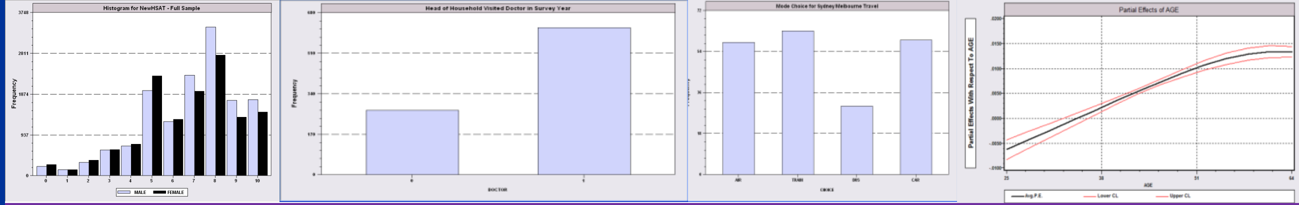
HOPit Model

```
+---------------------------------------------+
| Ordered Probability Model                   |
| Dependent variable                HSAT      |
| Weighting variable                None      |
| Number of observations            1939      |
| Iterations completed                16      |
| Log likelihood function      -2622.269      |
| Number of parameters                11      |
| Info. Criterion: AIC =         2.71611      |
|   Finite Sample: AIC =         2.71618      |
| Info. Criterion: BIC =         2.74771      |
| Info. Criterion:HQIC =         2.72773      |
| Restricted log likelihood    -2634.772      |
| Chi squared                   25.00667      |
| Degrees of freedom                   4      |
| Prob[ChiSqd > value] =        .5015413E-04 |
| Underlying probabilities based on Normal    |
|    Cell frequencies for outcomes            |
|  Y Count Freq  Y Count Freq  Y Count Freq   |
|  0    89 .045  1    55 .028  2   158 .081   |
|  3   267 .137  4   336 .173  5  1034 .533   |
| HOPIT (covariates in thresholds) model      |
+---------------------------------------------+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Index function for probability				
Constant	1.92291596	.25039254	7.680	.0000	
AGE	-.01097834	.00313151	-3.506	.0005	46.3140794
EDUC	.01790608	.01703842	1.051	.2933	10.5102719
FEMALE	.09800495	.05307626	1.846	.0648	.52037133
HHNINC	.13128664	.17744955	.740	.4594	.33073435
	Estimates of t(j) in mu(j)=exp[t(j)+d*z]				
Theta(1)	-1.44465511	.14963340	-9.655	.0000	
Theta(2)	-.41703432	.09847125	-4.235	.0000	
Theta(3)	.11112389	.08371803	1.327	.1844	
Theta(4)	.44971160	.07809864	5.758	.0000	
	Threshold covariates mu(j)=exp[t(j)+d*z]				
HHKIDS	-.03932779	.03943145	-.997	.3186	
INSURANC	.04440841	.07060503	.629	.5294	

Differential Item Functioning



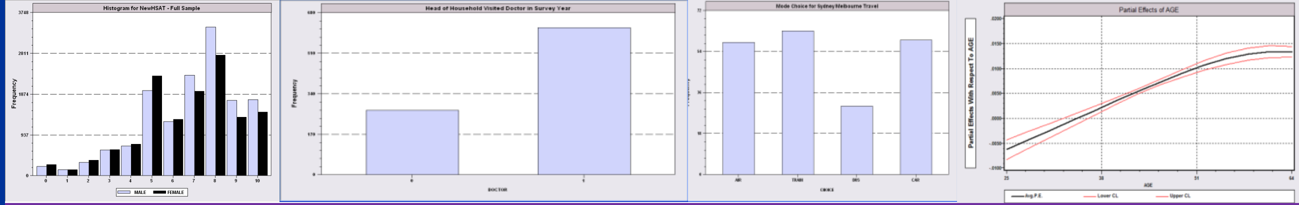


A Vignette Random Effects Model

To use all the information in the sample, the log likelihood function is the sum of the two parts, with the restriction on the common threshold parameters,

$$\log L = \sum_{i=1}^{N} \log \sum_{j=0}^{J} B_{i,j} \begin{bmatrix} \Phi\left(\tau_{i,j} - (\beta_0 - \lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) - \\ \Phi\left(\tau_{i,j-1} - (\beta_0 - \lambda_0) - (\boldsymbol{\beta}'\mathbf{x}_i - \boldsymbol{\gamma}_0'\mathbf{z}_i)\right) \end{bmatrix}$$

$$+ \sum_{q=1}^{Q} \sum_{m=1}^{M} \log \sum_{j=0}^{J} B_{q,j,m} \begin{bmatrix} \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} - \\ \Phi\left\{\frac{1}{\sigma}\left[\tau_{q,j-1} - (\theta_m - \lambda_0) - \boldsymbol{\gamma}_0'\mathbf{z}_q\right]\right\} \end{bmatrix}$$

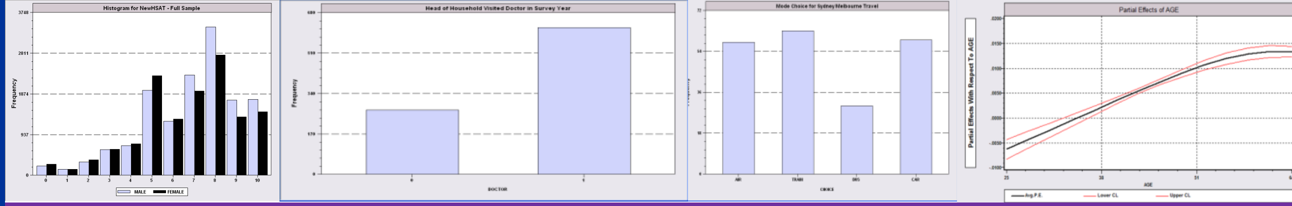


Vignettes

TABLE 2. Comparing Political Efficacy in Mexico and China

Eq.	Variable	Ordered Probit Coeff.	Ordered Probit (SE)	Our Method Coeff.	Our Method (SE)
μ	**China**	**0.670**	**(0.082)**	**−0.364**	**(0.090)**
	Age	0.004	(0.003)	0.006	(0.003)
	Male	0.087	(0.076)	0.114	(0.081)
	Education	0.020	(0.008)	0.020	(0.008)
τ^1	**China**			**−1.059**	**(0.059)**
	Age			0.002	(0.001)
	Male			0.044	(0.036)
	Education			−0.001	(0.004)
	Constant	0.425	(0.147)	0.431	(0.151)
τ^2	**China**			**−0.162**	**(0.071)**
	Age			−0.002	(0.002)
	Male			−0.059	(0.051)
	Education			0.001	(0.006)
	Constant	−0.320	(0.059)	−0.245	(0.114)
τ^3	**China**			**0.345**	**(0.053)**
	Age			−0.001	(0.002)
	Male			0.044	(0.047)
	Education			−0.003	(0.005)
	Constant	−0.449	(0.074)	−0.476	(0.105)
τ^4	**China**			**0.631**	**(0.083)**
	Age			0.004	(0.002)
	Male			−0.097	(0.072)
	Education			0.027	(0.007)
	Constant	−0.898	(0.119)	−1.621	(0.149)
Vignettes	θ_1			1.284	(0.161)
	θ_2			1.196	(0.160)
	θ_3			0.845	(0.159)
	θ_4			0.795	(0.159)
	θ_5			0.621	(0.159)
$\ln \sigma$				−0.239	(0.042)

Note: Ordered probit indicates counterintuitively and probably incorrectly that the Chinese have higher political efficacy than the Mexicans, whereas our approach reveals that this is because the Chinese have comparatively lower standards (τ's) for moving from one categorical response into the next highest category. The result is that although the Chinese give higher reported levels of political efficacy than the Mexicans, it is the Mexicans who are in fact more politically efficacious.

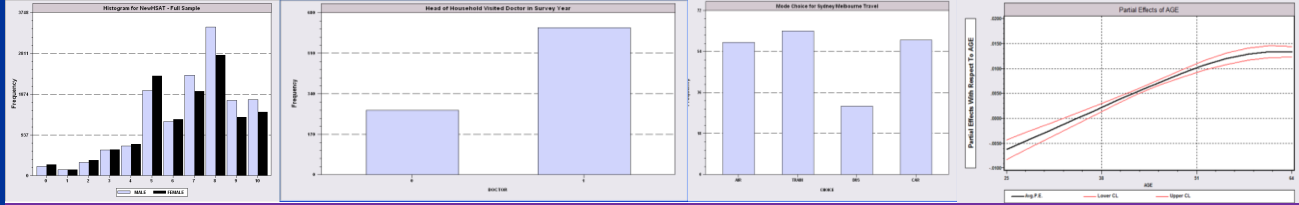


Qual Life Res
DOI 10.1007/s11136-013-0615-2

Anchoring vignettes for health comparisons: an analysis of response consistency

Nicole Au · Paula K. Lorgelly

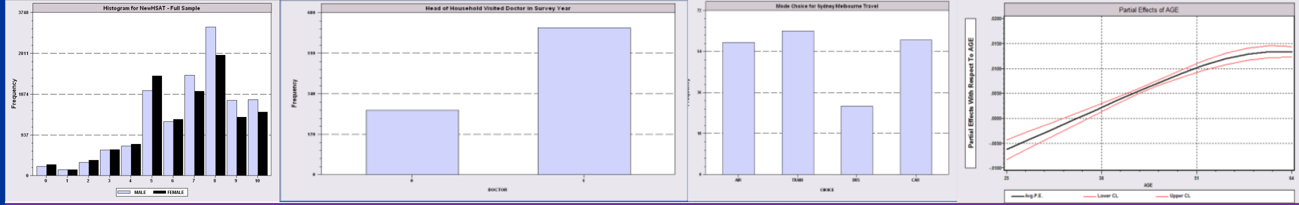
Accepted: 20 December 2013



Abstract

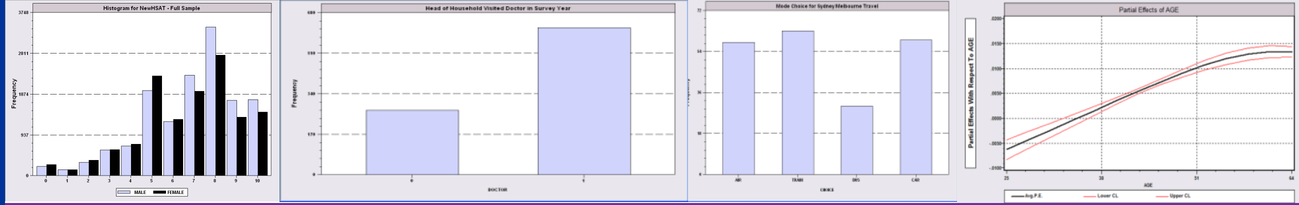
Purpose Self-rated health (SRH) is widely used to measure and compare the health status of different groups of individuals. However, SRH can suffer from heterogeneity in reporting styles, making health comparisons problematic. Anchoring vignettes is a promising technique for improving inter-group comparisons of SRH. A key identifying assumption of the approach is response consistency—that respondents rate themselves using the same underlying response scale that they rate the vignettes. Despite growing research into response consistency, it remains unclear *how* respondents rate vignettes and *why* respondents may not assess vignettes and themselves consistently.

Method Vignettes for the EQ-5D-5L were developed and included in an online survey. In-depth interviews were conducted with participants following survey completion. Response consistency was examined through qualitative analysis of the interview responses and quantitative coding of participants' thought processes.



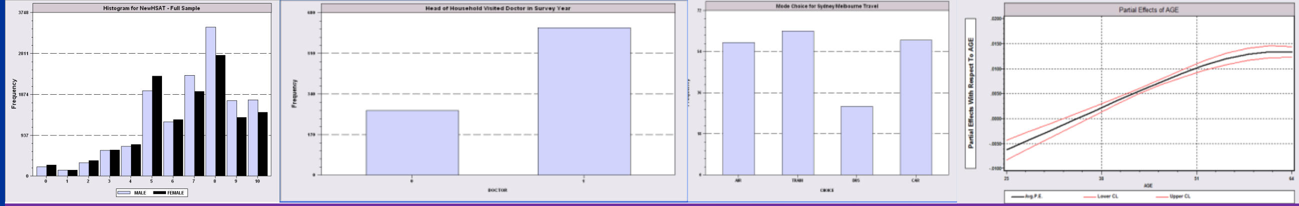
Introduction

Self-rated measures of general health have been shown to be good predictors of mortality and morbidity [1–3], and are widely used to measure and compare the health status of individuals. Self-rated health (SRH) measures have been used in a range of applications, including evaluations of health programs [4, 5], patient-reported outcomes [5–8] and monitoring of population health [9, 10]. In its simplest form, SRH asks individuals to evaluate their overall health on a five-point scale, while more comprehensive measures of SRH, such as the EQ-5D [11], aim to capture overall health-related quality of life (HRQL) using a number of questions about specific health dimensions (such as mobility and pain).



Panel Data

- Fixed Effects
  - The usual incidental parameters problem
  - Practically feasible but methodologically ambiguous
  - Partitioning Prob(y_{it} > j|$\mathbf{x}_{it}$) produces estimable binomial logit models. (Find a way to combine multiple estimates of the same $\boldsymbol{\beta}$.
- Random Effects
  - Standard application
  - Extension to random parameters – see above



Incidental Parameters Problem

Table 9.1 Monte Carlo Analysis of the Bias of the MLE in Fixed Effects Discrete Choice Models (Means of empirical sampling distributions, *N* = 1,000 individuals, *R* = 200 replications)

	T = 2		*T* = 3		*T* = 5	
	β	δ	β	δ	β	δ
Logit	2.020	2.027	1.698	1.668	1.379	1.323
Probit	2.083	1.938	1.821	1.777	1.589	1.407
Ordered Probit	2.328	2.605	1.592	1.806	1.305	1.415

	T = 8		*T* = 10		*T* = 20	
	β	δ	β	δ	β	δ
Logit	1.217	1.156	1.161	1.135	1.069	1.062
Probit	1.328	1.243	1.247	1.169	1.108	1.068
Ordered Probit	1.166	1.220	1.131	1.158	1.058	1.068

A Dynamic Ordered Probit Model

Journal of APPLIED ECONOMETRICS

Research Article

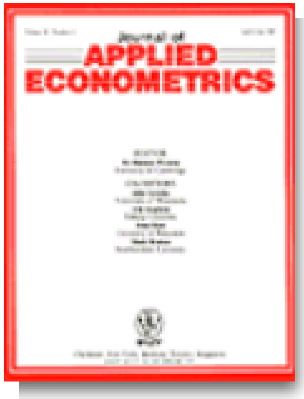
The dynamics of health in the British Household Panel Survey

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

Article first published online: 9 AUG 2004

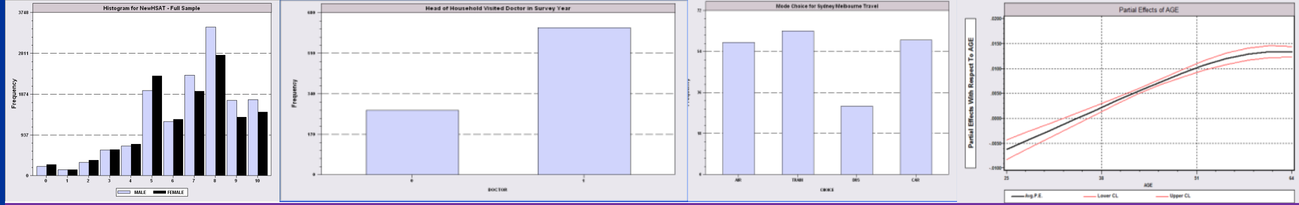
DOI: 10.1002/jae.755

Issue



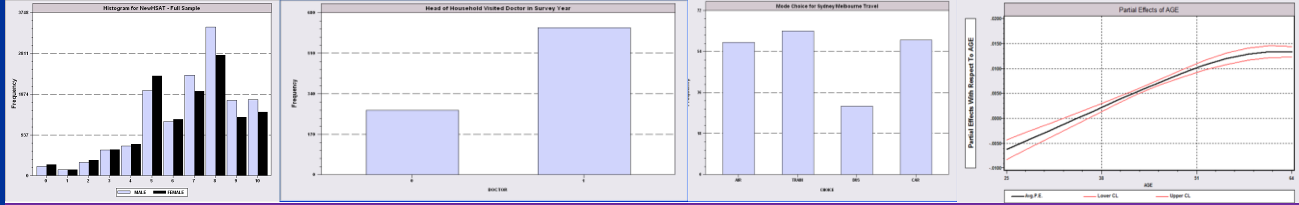
Journal of Applied Econometrics

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

$\text{H}_{i,t-1}(j) = 1[\text{Individual i reported } \text{h}_{it} = j \text{ in previous period}]$, 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

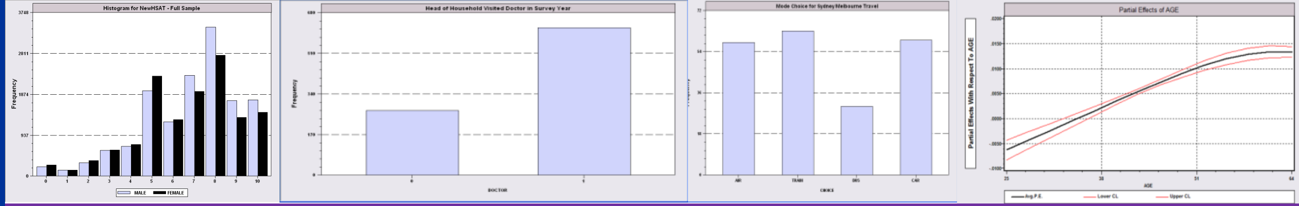
Ordered Choice Observation Mechanism

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

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

Random Effects with Mundlak Correction and Initial Conditions

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

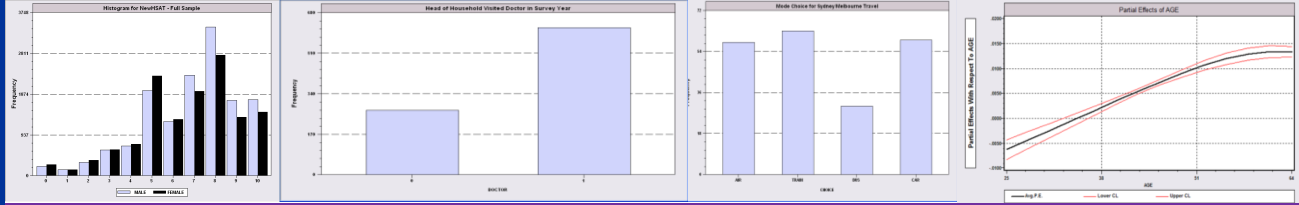


Testing for Attrition Bias

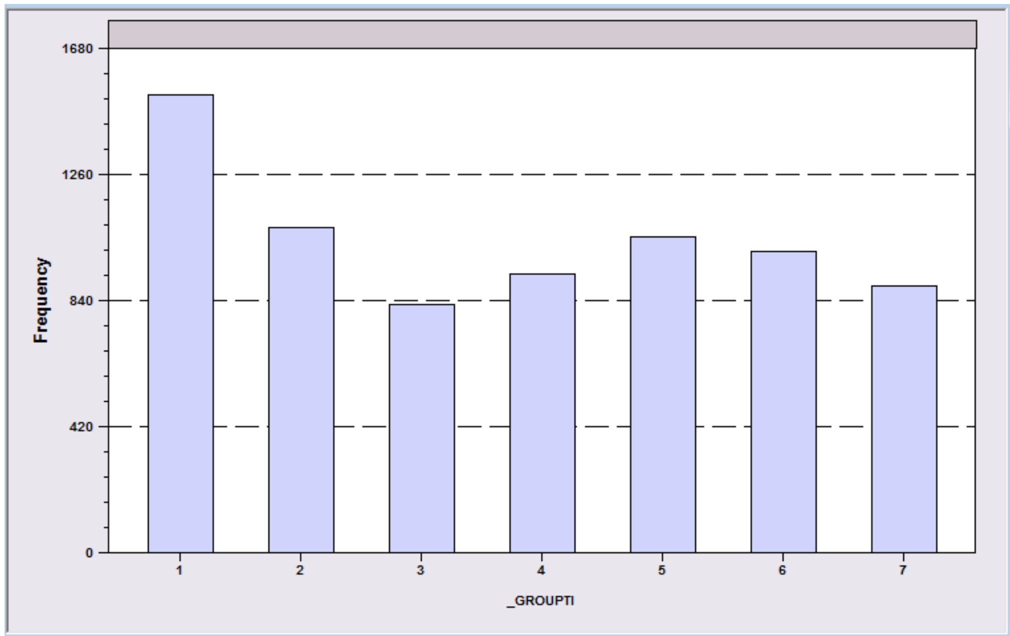
Table 9: Verbeek and Nijman tests for attrition: based on dynamic ordered probit models with Wooldridge specification of correlated effects and initial conditions

	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

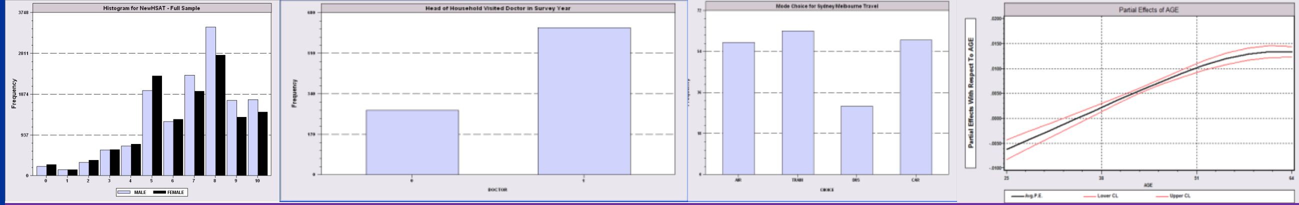
Three dummy variables added to full model with unbalanced panel suggest presence of attrition effects.



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 ≥ N2 ≥ ... ≥ 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].

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

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

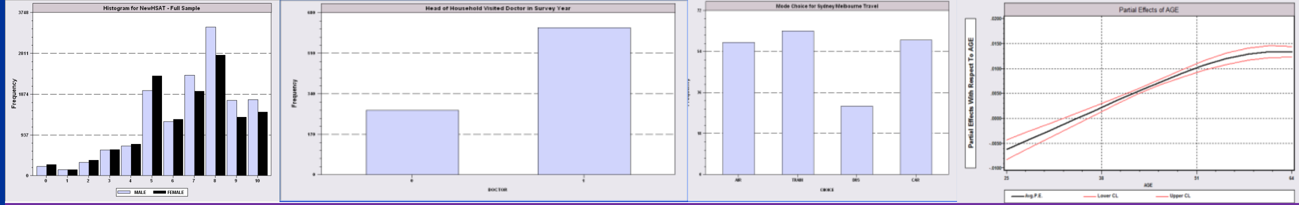
(health problems, disability, psychological well being, self employment, unemployment, maternity leave, student, caring for family member, ...)

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

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

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

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

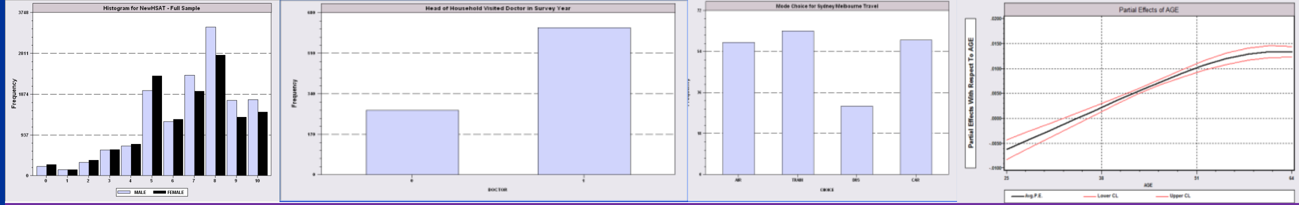


Estimated Partial Effects by Model

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

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

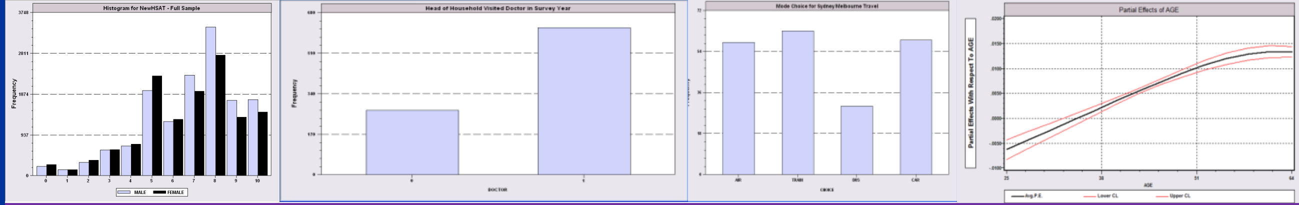
Table 12: Average partial effects on probability of reporting excellent health for selected variables

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 (.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)

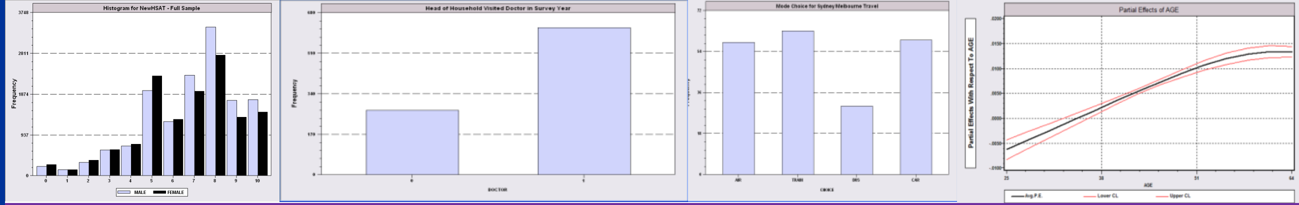
SAHEX(t-1)	.234 (.087)
SAHFAIR(t-1)	-.170 (.085)
SAHPOOR(t-1)	-.242 (.167)
SAHVPOOR(t-1)	-.260 (.198)

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).



Model Extensions

- Multivariate
  - Bivariate
  - Multivariate
- Inflation and Two Part
  - Zero inflation
  - Sample Selection
  - Endogenous Latent Class



A Sample Selection Model

Estimated Ordered Probit Sample Selection Model

$$PUBLIC_i^* = \alpha_1 + \alpha_2 AGE_i + \alpha_3 EDUC_i + \alpha_4 HANDDUM_i + u_i,$$
$$PUBLIC_i = 1[PUBLIC_i^* > 0],$$
$$HEALTH_i^* = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i,$$
$$HEALTH_i = j \text{ if } \mu_{j-1} < HEALTH_i^* \leq \mu_j,$$
$$(HEALTH_i, \mathbf{x}_i) \text{ observed when } PUBLIC_i = 1,$$
$$(u_i, \varepsilon_i) \sim N_2[(0,1), (1,1,\rho)],$$

```
+---------------------------------------------+
| Binomial Probit Model                       |
| Dependent variable               PUBLIC     |
| Number of observations             4483     |
| Log likelihood function       -1471.427     |
| Restricted log likelihood     -1711.545     |
+---------------------------------------------+
+---------+---------+--------+--------+------+---------+--------+--------+------+
|Variable| Coeff.   |Standard|b/St.Er.|Prob. | Coeff.  |Standard|b/St.Er.|Prob. |
|        |          |Error   |        |      |         |Error   |        |      |
+---------+---------+--------+--------+------+---------+--------+--------+------+
+--------+Index function for probability     |  Single Equation probit          |
|Constant|  3.4512   .1622    21.267   .0000| 3.5925   .1651   21.758   .0000|
|AGE     |  -.0054   .0025    -2.181   .0292| -.0027   .0024   -1.110   .2670|
|EDUC    |  -.1804   .0093   -19.394   .0000| -.1967   .0094  -21.016   .0000|
|HANDDUM |   .6710   .0803     8.353   .0000|  .2881   .0980    2.939   .0033|
+--------+Index function for ordered probit  |                                  |
|Constant|  2.2347   .1270    17.590   .0000|Binary Choice Model Predictions   |
|AGE     |  -.0160   .0016    -9.780   .0000|        Predicted                 |
|EDUC    |  -.0314   .0092    -3.398   .0007|Actual    0        1     Total     |
|INCOME  |   .2384   .0994     2.399   .0164|   0     164     408       572     |
|MARRIED |  -.0093   .0386     -.242   .8089|   1     141    3770      3911     |
|KIDS    |   .0545   .0371     1.466   .1427| Total   305    4178      4483     |
+--------+Threshold parameters for index     |                                  |
|Mu(1)   |   .9695   .0394    24.581   .0000|                                  |
|Mu(2)   |  2.2399   .0524    42.718   .0000|                                  |
|Mu(3)   |  2.7091   .0547    49.519   .0000|                                  |
|Rho(u,e)|   .8080   .0452    17.880   .0000|                                  |
+--------+-----------------------------------+----------------------------------+
```