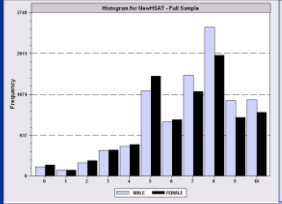
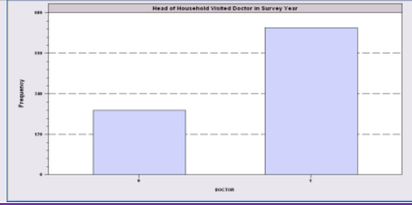
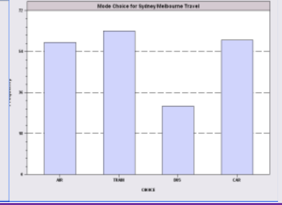
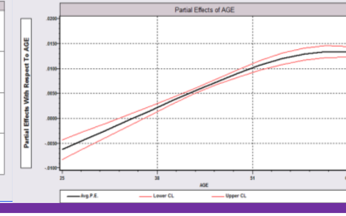
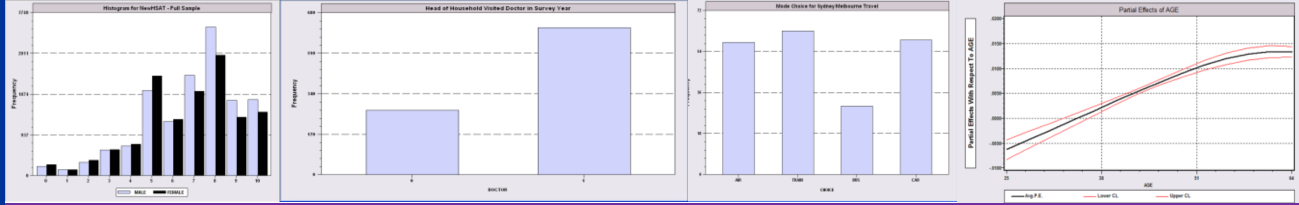





Discrete Choice Modeling

William Greene
Stern School of Business
New York University



Application: Major Derogatory Reports

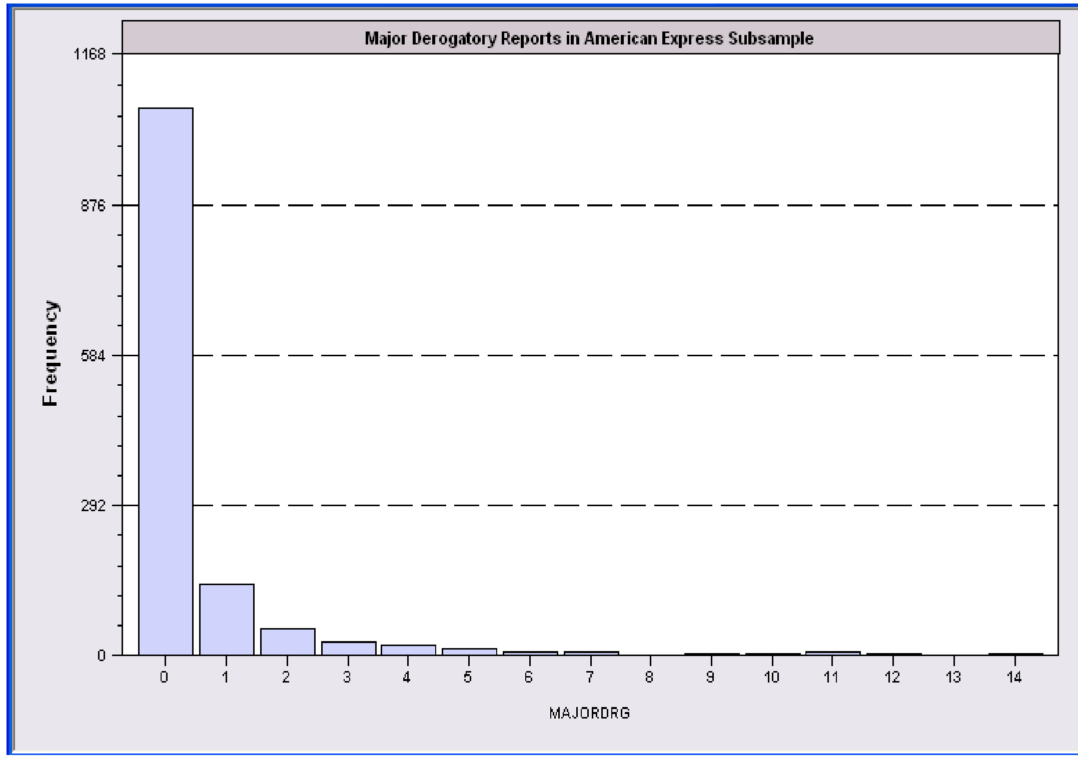
AmEx Credit Card Holders

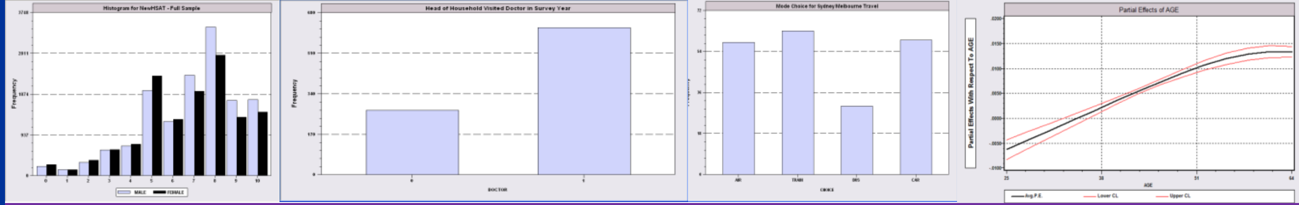
N = 1310 (of 13,777)

Number of major derogatory reports in 1 year

Issues:

- Nonrandom selection
- Excess zeros

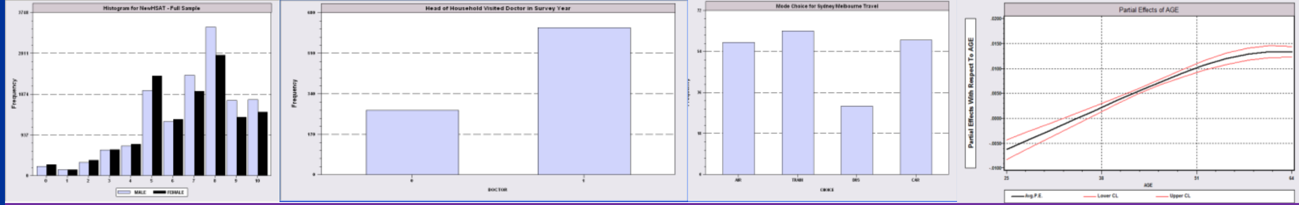




Histogram for Credit Data

Histogram for MAJORDRG NOBS= 1310, Too low: 0, Too high: 0

Bin	Lower limit	Upper limit	Frequency	Cumulative Frequency
0	.000	1.000	1053 (.8038)	1053(.8038)
1	1.000	2.000	136 (.1038)	1189(.9076)
2	2.000	3.000	50 (.0382)	1239(.9458)
3	3.000	4.000	24 (.0183)	1263(.9641)
4	4.000	5.000	17 (.0130)	1280(.9771)
5	5.000	6.000	10 (.0076)	1290(.9847)
6	6.000	7.000	5 (.0038)	1295(.9885)
7	7.000	8.000	6 (.0046)	1301(.9931)
8	8.000	9.000	0 (.0000)	1301(.9931)
9	9.000	10.000	2 (.0015)	1303(.9947)
10	10.000	11.000	1 (.0008)	1304(.9954)
11	11.000	12.000	4 (.0031)	1308(.9985)
12	12.000	13.000	1 (.0008)	1309(.9992)
13	13.000	14.000	0 (.0000)	1309(.9992)
14	14.000	15.000	1 (.0008)	1310(1.0000)

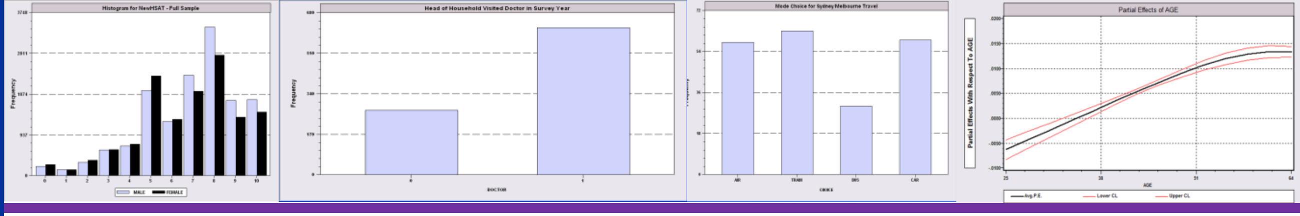


JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **18**: 387–405 (2003)
Published online 8 October 2002 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.680

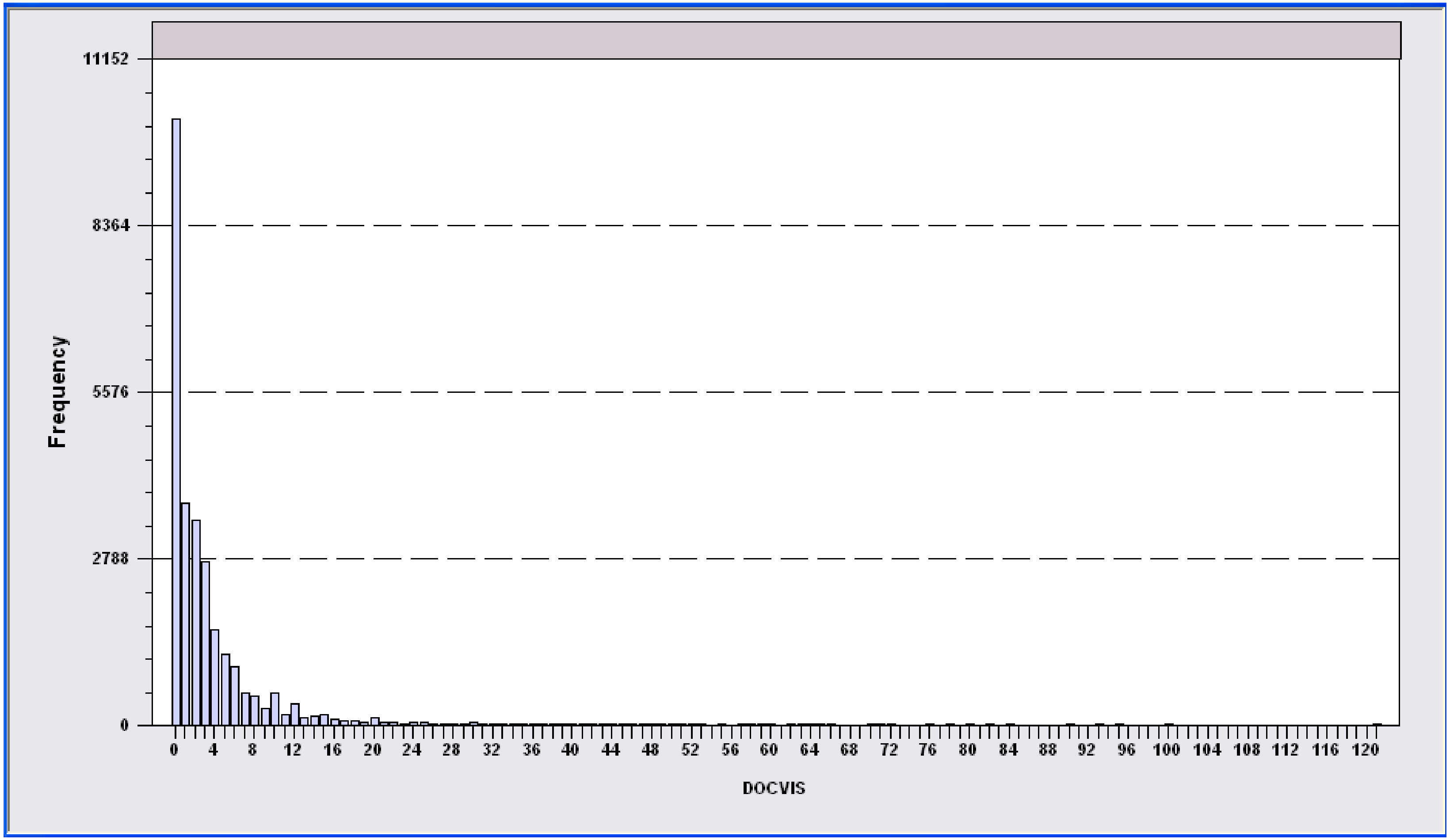
INCENTIVE EFFECTS IN THE DEMAND FOR HEALTH CARE: A BIVARIATE PANEL COUNT DATA ESTIMATION

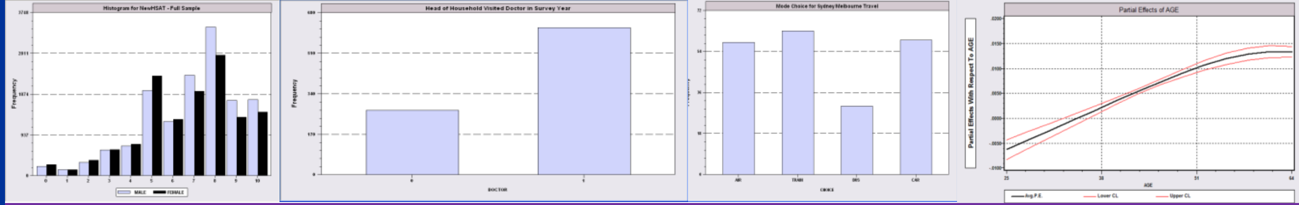
REGINA T. RIPHAHN,[a–c]* ACHIM WAMBACH[b,d] AND ANDREAS MILLION[a,c]

[a] *University of Munich, Germany*
[b] *CEPR, London, UK*
[c] *IZA, Bonn, Germany*
[d] *University of Erlangen-Nuremberg, Germany*



Doctor Visits



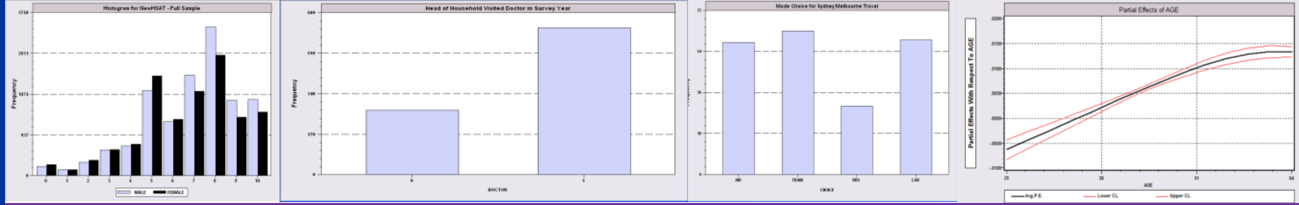


Basic Modeling for Counts of Events

- E.g., Visits to site, number of purchases, number of doctor visits
- Regression approach
  - Quantitative outcome measured
  - Discrete variable, model probabilities
- Poisson probabilities – "loglinear model"

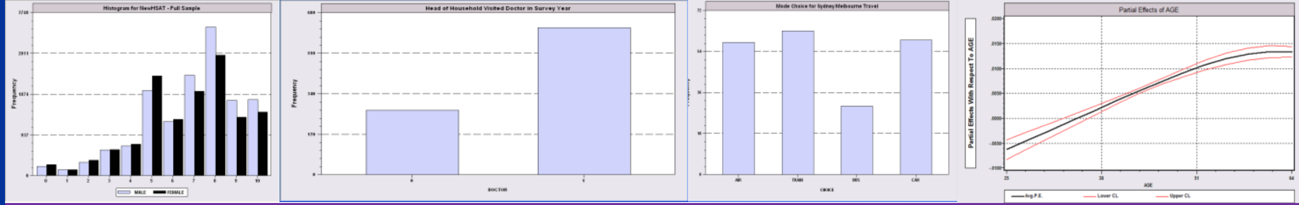
$$\text{Prob}[Y_i = j \mid \mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}$$

$$\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i) = E[y_i \mid \mathbf{x}_i]$$



Poisson Model for Doctor Visits

```
----------------------------------------------------------------------
Poisson Regression
Dependent variable                    DOCVIS
Log likelihood function    -103727.29625
Restricted log likelihood -108662.13583
Chi squared [    6 d.f.]       9869.67916
Significance level                 .00000
McFadden Pseudo R-squared        .0454145
Estimation based on N =   27326, K =    7
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              7.59235   207468.59251
Chi- squared =255127.59573  RsqP= .0818
G  - squared =154416.01169  RsqD= .0601
Overdispersion tests: g=mu(i)   : 20.974
Overdispersion tests: g=mu(i)^2: 20.943
--------+-------------------------------------------------------------
Variable| Coefficient   Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
Constant|     .77267***       .02814       27.463   .0000
     AGE|     .01763***       .00035       50.894   .0000     43.5257
    EDUC|    -.02981***       .00175      -17.075   .0000     11.3206
  FEMALE|     .29287***       .00702       41.731   .0000      .47877
 MARRIED|     .00964          .00874        1.103   .2702      .75862
  HHNINC|    -.52229***       .02259      -23.121   .0000      .35208
  HHKIDS|    -.16032***       .00840      -19.081   .0000      .40273
--------+-------------------------------------------------------------
```



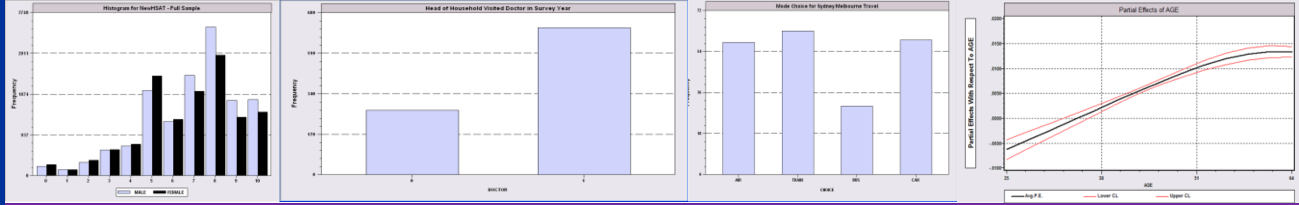
Partial Effects

Partial derivatives of expected val. with respect to the vector of characteristics.
Effects are averaged over individuals.
Observations used for means are All Obs.

Conditional Mean at Sample Point	3.1835
Scale Factor for Marginal Effects	3.1835

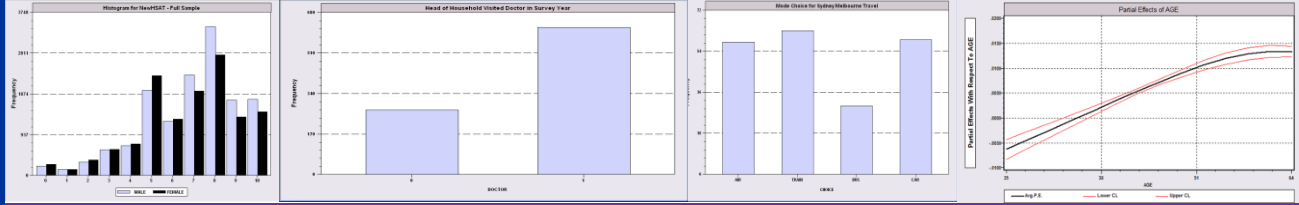
$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial \mathbf{x}_i} = \lambda_i \boldsymbol{\beta}$$

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
AGE	.05613***	.00131	42.991	.0000	43.5257
EDUC	-.09490***	.00596	-15.923	.0000	11.3206
FEMALE	.93237***	.02555	36.491	.0000	.47877
MARRIED	.03069	.02945	1.042	.2973	.75862
HHNINC	-1.66271***	.07803	-21.308	.0000	.35208
HHKIDS	-.51037***	.02879	-17.730	.0000	.40273



Poisson Model Specification Issues

- Equi-Dispersion: $\text{Var}[y_i|\mathbf{x}_i] = E[y_i|\mathbf{x}_i]$.
- Overdispersion: If $\lambda_i = \exp[\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i]$,
  - $E[y_i|\mathbf{x}_i] = \gamma\exp[\boldsymbol{\beta}'\mathbf{x}_i]$
  - $\text{Var}[y_i] > E[y_i]$ (overdispersed)
  - ε_i ~ log-Gamma ➔ Negative binomial model
  - ε_i ~ Normal$[0,\sigma^2]$ ➔ Normal-mixture model
  - ε_i is viewed as unobserved heterogeneity (“frailty”).
    - Normal model may be more natural.
    - Estimation is a bit more complicated.



$$\text{Prob}[Y_i = j \mid \mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}$$

$$\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i) = E[y_i \mid \mathbf{x}_i]$$

Estimation:

Nonlinear Least Squares: $\textbf{Min}_\beta \sum_{i=1}^{N} (y_i - \lambda_i)^2$

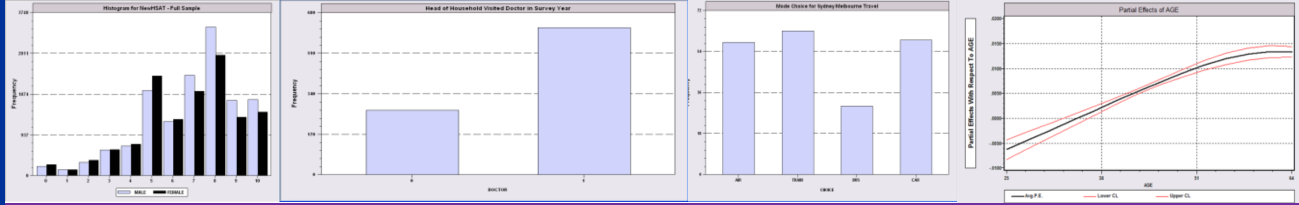
Moment Equations : $\sum_{i=1}^{N} \lambda_i (y_i - \lambda_i)\mathbf{x}_i$

Inefficient but robust if nonPoisson

Maximum Likelihood: $\textbf{Max}_\beta \sum_{i=1}^{N} [-\lambda_i + y_i \log \lambda_i - \log(y_i!)]$

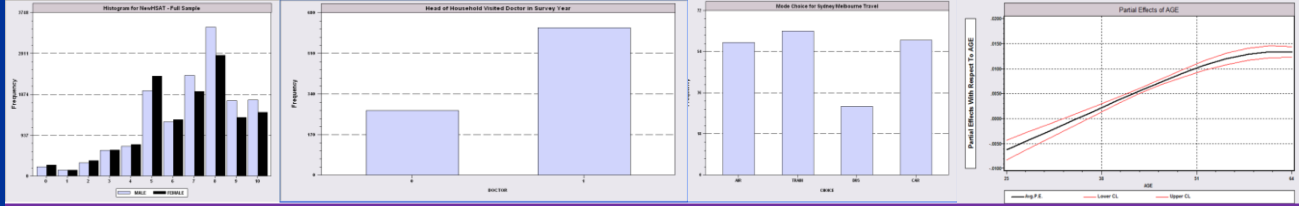
Moment Equations : $\sum_{i=1}^{N} (y_i - \lambda_i)\mathbf{x}_i$

Efficient, also robust to some kinds of NonPoissonness

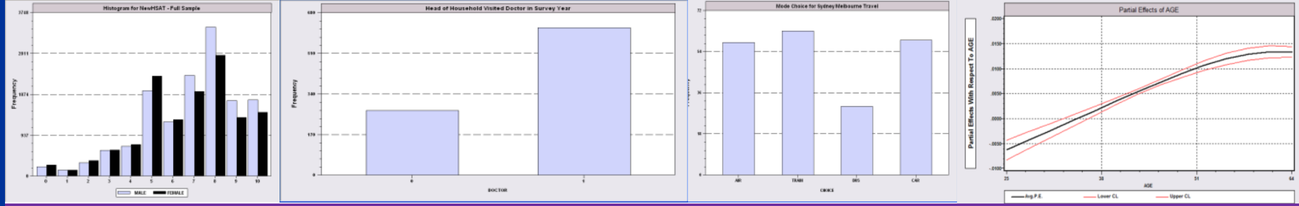


Poisson Model for Doctor Visits

```
----------------------------------------------------------------------
Poisson Regression
Dependent variable                    DOCVIS
Log likelihood function     -103727.29625
Restricted log likelihood -108662.13583
Chi squared [    6 d.f.]       9869.67916
Significance level                 .00000
McFadden Pseudo R-squared        .0454145
Estimation based on N =   27326, K =    7
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              7.59235   207468.59251
Chi- squared =255127.59573  RsqP= .0818
G   - squared =154416.01169  RsqD= .0601
Overdispersion tests: g=mu(i)   : 20.974
Overdispersion tests: g=mu(i)^2: 20.943
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
Constant|      .77267***        .02814      27.463   .0000
     AGE|      .01763***        .00035      50.894   .0000     43.5257
    EDUC|     -.02981***        .00175     -17.075   .0000     11.3206
  FEMALE|      .29287***        .00702      41.731   .0000      .47877
 MARRIED|      .00964           .00874       1.103   .2702      .75862
  HHNINC|     -.52229***        .02259     -23.121   .0000      .35208
  HHKIDS|     -.16032***        .00840     -19.081   .0000      .40273
--------+-------------------------------------------------------------
```

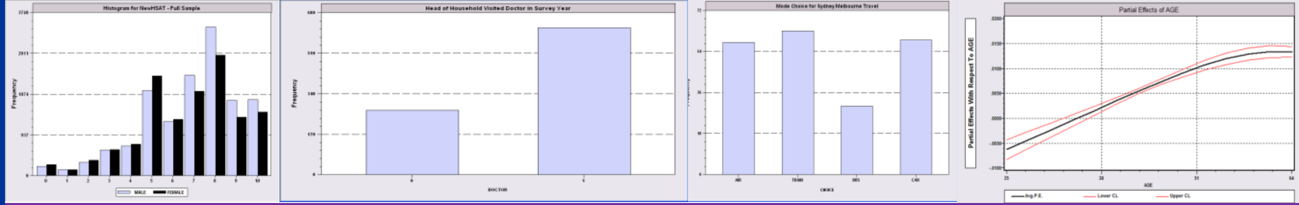


Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
	Standard – Negative Inverse of Second Derivatives				
Constant	.77267***	.02814	27.463	.0000	
AGE	.01763***	.00035	50.894	.0000	43.5257
EDUC	-.02981***	.00175	-17.075	.0000	11.3206
FEMALE	.29287***	.00702	41.731	.0000	.47877
MARRIED	.00964	.00874	1.103	.2702	.75862
HHNINC	-.52229***	.02259	-23.121	.0000	.35208
HHKIDS	-.16032***	.00840	-19.081	.0000	.40273
	Robust – Sandwich				
Constant	.77267***	.08529	9.059	.0000	
AGE	.01763***	.00105	16.773	.0000	43.5257
EDUC	-.02981***	.00487	-6.123	.0000	11.3206
FEMALE	.29287***	.02250	13.015	.0000	.47877
MARRIED	.00964	.02906	.332	.7401	.75862
HHNINC	-.52229***	.06674	-7.825	.0000	.35208
HHKIDS	-.16032***	.02657	-6.034	.0000	.40273
	Cluster Correction				
Constant	.77267***	.11628	6.645	.0000	
AGE	.01763***	.00142	12.440	.0000	43.5257
EDUC	-.02981***	.00685	-4.355	.0000	11.3206
FEMALE	.29287***	.03213	9.116	.0000	.47877
MARRIED	.00964	.03851	.250	.8023	.75862
HHNINC	-.52229***	.08295	-6.297	.0000	.35208
HHKIDS	-.16032***	.03455	-4.640	.0000	.40273



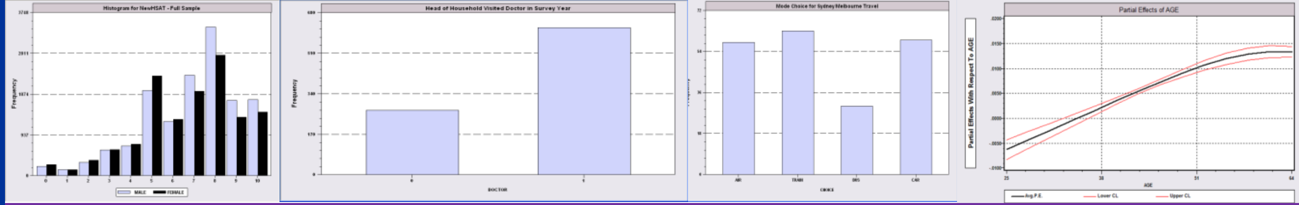
Negative Binomial Specification

- Prob(Y_i=j|$\mathbf{x}_i$) has greater mass to the right and left of the mean
- Conditional mean function is the same as the Poisson: $E[y_i|\mathbf{x}_i] = \lambda_i = \text{Exp}(\boldsymbol{\beta}'\mathbf{x}_i)$, so marginal effects have the same form.
- Variance is $\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + \alpha\,\lambda_i)$, α is the overdispersion parameter; $\alpha = 0$ reverts to the Poisson.
- Poisson is consistent when NegBin is appropriate. Therefore, this is a case for the ROBUST covariance matrix estimator. (Neglected heterogeneity that is uncorrelated with $\mathbf{x}_i$.)



NegBin Model for Doctor Visits

```
----------------------------------------------------------------------
Negative Binomial Regression
Dependent variable                      DOCVIS
Log likelihood function     -60134.50735  NegBin  LogL
Restricted log likelihood -103727.29625  Poisson LogL
Chi squared [    1 d.f.]     87185.57782  Reject Poisson model
Significance level                .00000
McFadden Pseudo R-squared       .4202634
Estimation based on N =   27326, K =    8
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              4.40185   120285.01469
NegBin form 2; Psi(i) = theta
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
Constant|     .80825***       .05955        13.572   .0000
     AGE|     .01806***       .00079        22.780   .0000     43.5257
    EDUC|    -.03717***       .00386        -9.622   .0000     11.3206
  FEMALE|     .32596***       .01586        20.556   .0000      .47877
 MARRIED|    -.00605          .01880        -.322    .7477      .75862
  HHNINC|    -.46768***       .04663       -10.029   .0000      .35208
  HHKIDS|    -.15274***       .01729        -8.832   .0000      .40273
        |Dispersion parameter for count data model
   Alpha|    1.89679***       .01981        95.747   .0000
--------+-------------------------------------------------------------
```



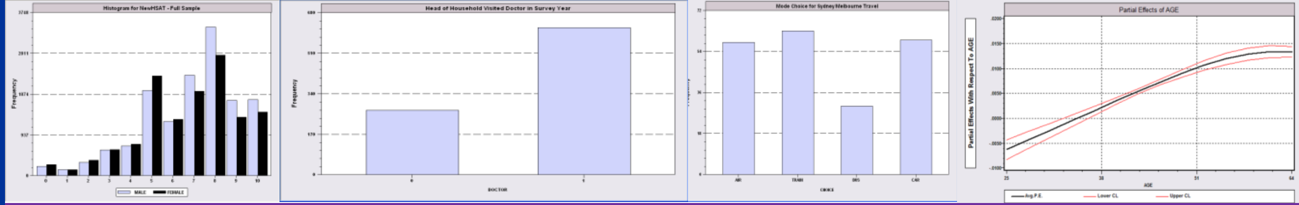
Marginal Effects

Scale Factor for Marginal Effects 3.1835 POISSON

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
AGE	.05613***	.00131	42.991	.0000	43.5257
EDUC	-.09490***	.00596	-15.923	.0000	11.3206
FEMALE	.93237***	.02555	36.491	.0000	.47877
MARRIED	.03069	.02945	1.042	.2973	.75862
HHNINC	-1.66271***	.07803	-21.308	.0000	.35208
HHKIDS	-.51037***	.02879	-17.730	.0000	.40273

Scale Factor for Marginal Effects 3.1924 NEGATIVE BINOMIAL

Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
AGE	.05767***	.00317	18.202	.0000	43.5257
EDUC	-.11867***	.01348	-8.804	.0000	11.3206
FEMALE	1.04058***	.06212	16.751	.0000	.47877
MARRIED	-.01931	.06382	-.302	.7623	.75862
HHNINC	-1.49301***	.16272	-9.176	.0000	.35208
HHKIDS	-.48759***	.06022	-8.097	.0000	.40273



Journal of Health Economics 28 (2009) 265–279

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Health care utilisation in Europe: New evidence from the ECHP

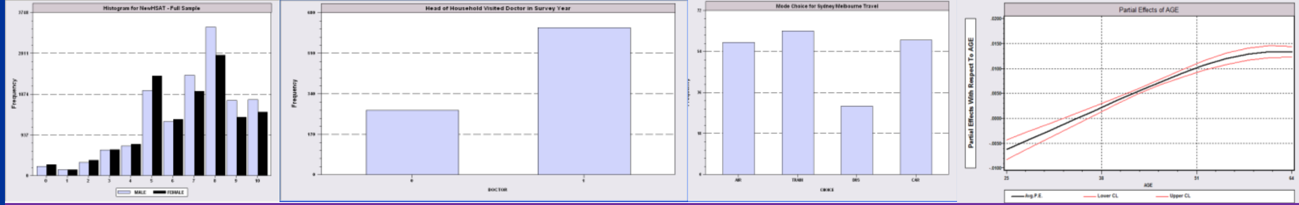
Teresa Bago d'Uva[a,b,*], Andrew M. Jones[c]

[a] Department of Applied Economics (Room H13-09), Erasmus School of Economics, PB 1738, 3000 DR Rotterdam, The Netherlands
[b] Netspar, The Netherlands
[c] Department of Economics and Related Studies, University of York, York, YO10 5DD, United Kingdom

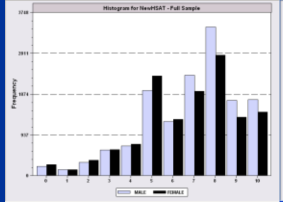
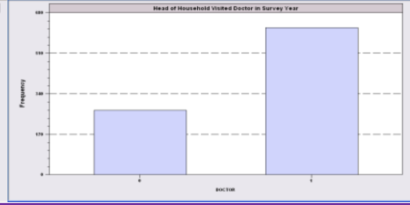
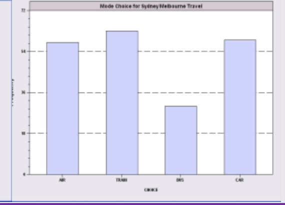
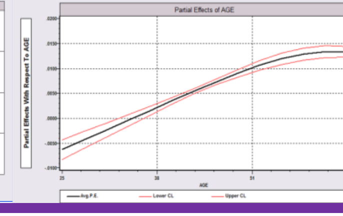
Following Bago d'Uva (2006), the class-specific density of the number of visits in a given year, $f_j(y_{it}|x_{it},\theta_j)$, is defined as in the standard hurdle model, using a negative binomial as the parent distribution in both stages. Formally, for each component j, $j=1,\ldots,C$, the probability of zero visits and the probability of observing y_{it} visits, given $y_{it}>0$, are given by

$$
\boxed{
\begin{aligned}
f_j(0|x_{it};\theta_{j1}) &= P[Y_{it}=0|x_{it},\theta_{j1}] = (\lambda_{j1,it}^{1-k}+1)^{-\lambda_{j1,it}^{k}}\\
f_j(y_{it}|y_{it}>0,x_{it};\theta_{j2}) &= \frac{\Gamma\left(y+(\lambda_{j2,it}^{k}/\alpha_j)\right)\left(\alpha_j\lambda_{j2,it}^{1-k}+1\right)^{-(\lambda_{j2,it}^{k}/\alpha_j)}[1+(\lambda_{j2,it}^{k-1}/\alpha_j)]^{-y_{i,t}}}{\Gamma(\lambda_{j2,it}^{k}/\alpha_j)\Gamma(y_{it}+1)\left[1-(\alpha_j\lambda_{j2,it}^{1-k}+1)^{-(\lambda_{j2,it}^{k}/\alpha_j)}\right]},
\end{aligned}
}
\qquad (2)
$$

where $\lambda_{j1,it}=\exp(x'_{it}\beta_{j1})$, $\lambda_{j2,it}=\exp(x'_{it}\beta_{j2})$, α_j are overdispersion parameters and k is an an arbitrary constant (most commonly set equal to 1 or 0, corresponding to the NegBin1 and NegBin2 models, respectively; we use the NegBin2 model). The vectors of parameters driving the probability of seeking care, β_{j1}, and the number of visits, given that this is positive, β_{j2}, are allowed to be different which means that the determinants of care may have different effects on the two stages of the decision process. This is also the case in the standard hurdle model, which corresponds to a (degenerate) LC model with only




```
-----------------------------------------------------------------------------
Negative Binomial Regression
Dependent variable                DOCVIS
Log likelihood function     -60134.40266
Restricted log likelihood -103726.69548
Chi squared [   1 d.f.]      87184.58563
Significance level                .00000
McFadden Pseudo R-squared       .4202611
Estimation based on N =  27326, K =    8
Inf.Cr.AIC  = 120284.8 AIC/N =     4.402
NegBin form 2; Psi(i) = theta
Tests of Model Restrictions on Neg.Bin.
Model               Logl ChiSquared[df]
Poisson(b=0)  -108662.14  ********* [**]
Poisson       -103726.70     9870.9 [ 6]
NB v. Poisson  -60134.40    87184.6 [ 1]
NB (b=0)       -60835.72     1402.6 [ 6]
--------+--------------------------------------------------------------------
        |                Standard            Prob.       95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
Constant|     .80836***    .05956    13.57  .0000       .69163      .92508
     AGE|     .01806***    .00079    22.78  .0000       .01651      .01962
    EDUC|    -.03716***    .00386    -9.62  .0000     -.04473     -.02958
 MARRIED|    -.00601       .01880     -.32  .7493     -.04285      .03083
  FEMALE|     .32594***    .01586    20.55  .0000       .29486      .35702
  INCOME|    -.46841***    .04662   -10.05  .0000     -.55978     -.37703
  HHKIDS|    -.15273***    .01729    -8.83  .0000     -.18662     -.11883
        |Dispersion parameter for count data model
   Alpha|    1.89677***    .01981    95.75  .0000      1.85794     1.93560
--------+--------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```

Model Formulations

Poisson

$$\text{Prob}[Y = y_i \mid \mathbf{x}_i] = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{\Gamma(1+y_i)},$$

$$\lambda_i = \exp(\alpha + \mathbf{x}_i'\boldsymbol{\beta}),\ y_i = 0,1,\ldots,\ i = 1,\ldots,N$$

$$E[y \mid \mathbf{x}_i] = Var[y \mid \mathbf{x}_i] = \lambda_i$$

$$E[y_i \mid \mathbf{x}_i] = \lambda_i$$

Negative Binomial – 1

$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i + \kappa\, \lambda_i^{1} = \lambda_i\,[1 + \kappa],$

Replace θ with $\theta\lambda_i$ in NB-2.

$$\text{Prob}[Y = y_i \mid \mathbf{x}_i] = \frac{\Gamma(\theta\lambda_i + y_i)q^{\theta\lambda_i}(1-q)^{y_i}}{\Gamma(y_i+1)\Gamma(\theta\lambda_i)}$$

$$y_i = 0,1,\ldots;\ q = 1/1(1+\theta).$$

$E[y_i|\mathbf{x}_i] = \lambda_i$

Negative Binomial – 2

$E[y_i|\mathbf{x}_i,\varepsilon_i] = \exp(\alpha + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i) = h_i\lambda_i,$

$$\text{Prob}[Y = y_i|\mathbf{x}_i] = \frac{\Gamma(\theta + y_i)r_i^{\theta}(1-r_i)^{y_i}}{\Gamma(1+y_i)\Gamma(\theta)},$$

$$y_i = 0,1,\ldots,\theta > 0,$$

$r_i = \theta/(\theta+\lambda_i)$

$E[y_i|\mathbf{x}_i] = \lambda_i,\quad \text{Var}[y_i|\mathbf{x}_i] = \lambda_i\,[1 + (1/\theta)\lambda_i]$

$= \lambda_i\,[1 + \kappa\,\lambda_i]$

$\kappa = \text{Var}[h_i].$

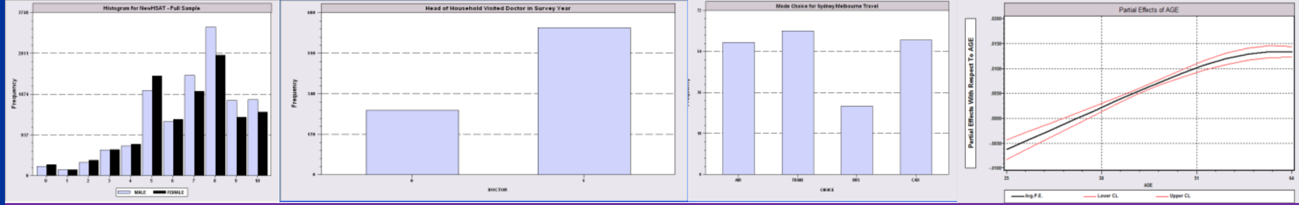
Replace θ with $\theta\lambda_i^{2-P}$ in NB-1

$$\text{Prob}[Y = y_i \mid \mathbf{x}_i] = \frac{\Gamma(\theta\lambda_i^{2-P} + y_i)s_i^{\theta\lambda_i^{2-P}}(1-s_i)^{y_i}}{\Gamma(y_i+1)\Gamma(\theta\lambda_i^{2-P})}$$

$$s_i = \frac{\lambda_i}{\lambda_i + \theta\lambda_i^{2-P}}.$$

$E[y_i|\mathbf{x}_i] = \lambda_i$

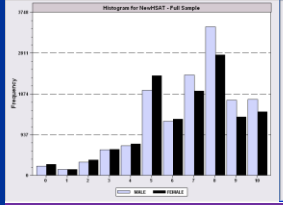
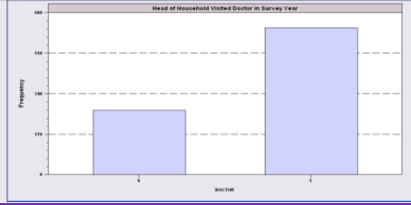
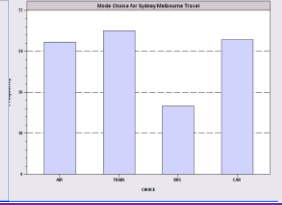
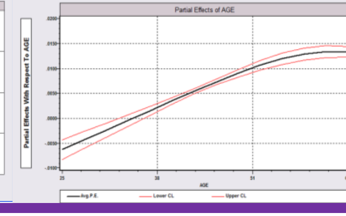
$\text{Var}[y_i|\mathbf{x}_i] = \lambda_i\,[1 + (1/\theta)\lambda_i^{P-1}].$



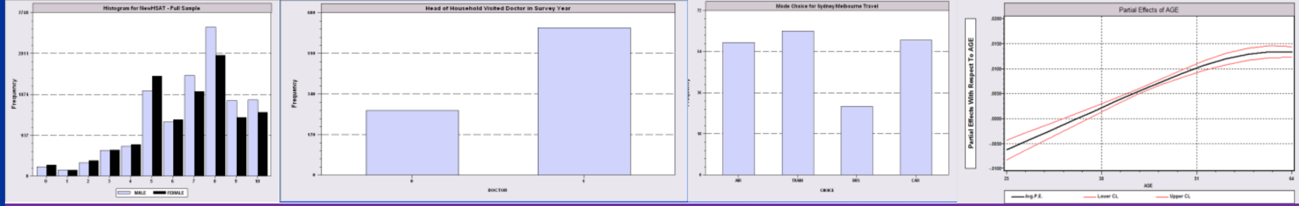
NegBin-P Model

```
----------------------------------------------------------------------
Negative Binomial (P) Model
Dependent variable                  DOCVIS
Log likelihood function     -59992.32903
Restricted log likelihood -103727.29625
Chi squared [    1 d.f.]      87469.93445
--------+-------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er.
--------+-------------------------------------------
Constant|     .60840***       .06452        9.429
     AGE|     .01710***       .00082       20.782
    EDUC|    -.02313***       .00414       -5.581
  FEMALE|     .36386***       .01640       22.187
 MARRIED|     .03670*         .02030        1.808
  HHNINC|    -.35093***       .05146       -6.819
  HHKIDS|    -.16902***       .01911       -8.843
        |Dispersion parameter for count data model
   Alpha|    3.85713***       .14581       26.453
        |Negative Binomial. General form, NegBin P
       P|    1.38693***       .03142       44.140
--------+-------------------------------------------------------------
```

NB-2	NB-1	Poisson
.80825***	.62584***	.77267***
.01806***	.01428***	.01763***
-.03717***	-.01549***	-.02981***
.32596***	.33028***	.29287***
-.00605	.04324**	.00964
-.46768***	-.24543***	-.52229***
-.15274***	-.14877***	-.16032***
Dispersion	Dispersion	
1.89679***	6.09246***	

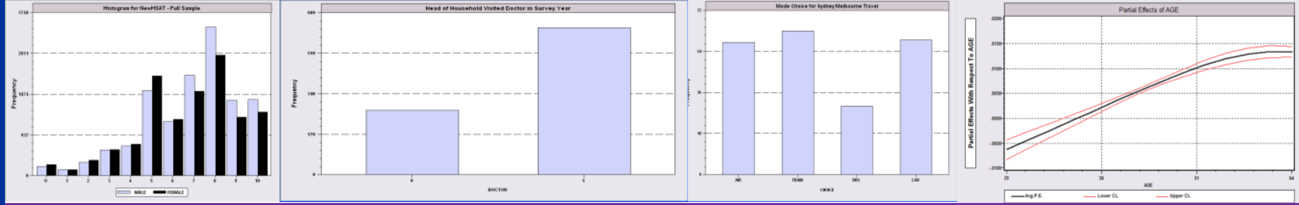





Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
Scale Factor for Marginal Effects 3.1835 POISSON					
AGE	.05613***	.00131	42.991	.0000	43.5257
EDUC	-.09490***	.00596	-15.923	.0000	11.3206
FEMALE	.93237***	.02555	36.491	.0000	.47877
MARRIED	.03069	.02945	1.042	.2973	.75862
HHNINC	-1.66271***	.07803	-21.308	.0000	.35208
HHKIDS	-.51037***	.02879	-17.730	.0000	.40273
Scale Factor for Marginal Effects 3.1924 NEGATIVE BINOMIAL - 2					
AGE	.05767***	.00317	18.202	.0000	43.5257
EDUC	-.11867***	.01348	-8.804	.0000	11.3206
FEMALE	1.04058***	.06212	16.751	.0000	.47877
MARRIED	-.01931	.06382	-.302	.7623	.75862
HHNINC	-1.49301***	.16272	-9.176	.0000	.35208
HHKIDS	-.48759***	.06022	-8.097	.0000	.40273
Scale Factor for Marginal Effects 3.1835 NEGATIVE BINOMIAL - 1					
AGE	.04547***	.00263	17.285	.0000	43.5257
EDUC	-.04933***	.01196	-4.125	.0000	11.3206
FEMALE	1.05145***	.05456	19.272	.0000	.47877
MARRIED	.13766**	.06154	2.237	.0253	.75862
HHNINC	-.78134***	.15139	-5.161	.0000	.35208
HHKIDS	-.47361***	.05885	-8.048	.0000	.40273
Scale Factor for Marginal Effects 3.0077 NEGATIVE BINOMIAL - P					
AGE	.05143***	.00246	20.934	.0000	43.5257
EDUC	-.06957***	.01241	-5.605	.0000	11.3206
FEMALE	1.09436***	.04968	22.027	.0000	.47877
MARRIED	.11038*	.06109	1.807	.0708	.75862
HHNINC	-1.05547***	.15411	-6.849	.0000	.35208
HHKIDS	-.50835***	.05753	-8.836	.0000	.40273



Zero Inflation – ZIP Models

- Two regimes: (Recreation site visits)
  - Zero (with probability 1). (Never visit site)
  - Poisson with Pr(0) = exp[- β'x_i]. (Number of visits, including zero visits this season.)
- Unconditional:
  - Pr[0] = P(regime 0) + P(regime 1)*Pr[0|regime 1]
  - Pr[j | j >0] = P(regime 1)*Pr[j|regime 1]
- "Two inflation" – Number of children
- These are "latent class models"



Zero Inflation Models

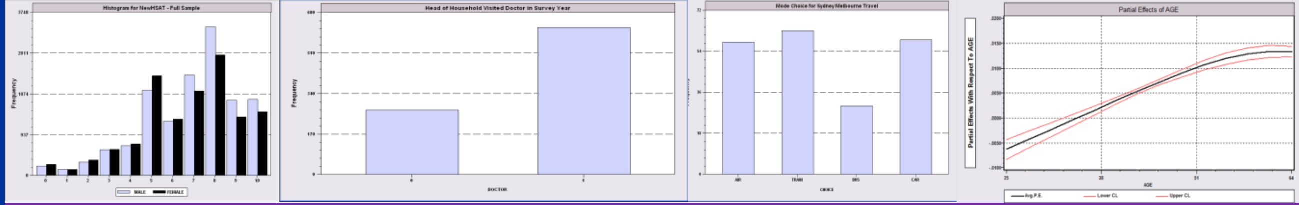
$$\boxed{\text{Prob}(y_i = j \mid x_i) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}, \lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)}$$

$$\text{Zero Inflation} = \text{ZIP}$$

$$\text{Prob}(0 \text{ regime}) = \boxed{F(\boldsymbol{\gamma}'\mathbf{z}_i)}$$

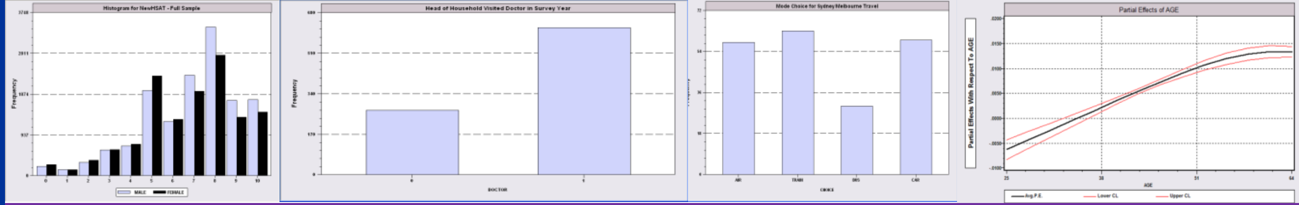
$$\text{ZIP - tau} = \text{ZIP}(\tau) \text{ [Not generally used]}$$

$$\text{Prob}(0 \text{ regime}) = \boxed{F(\tau\boldsymbol{\beta}'\mathbf{x}_i)}$$



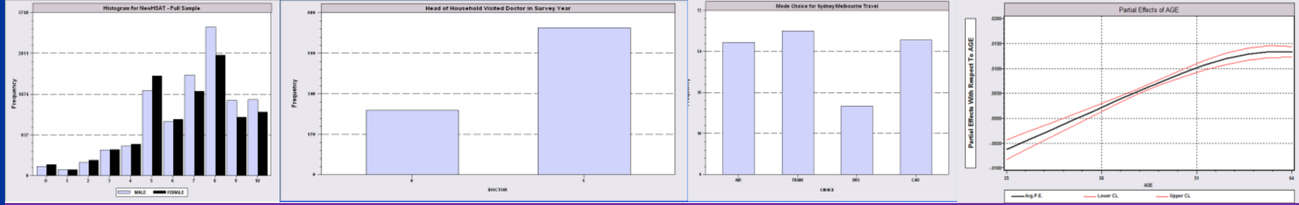
Notes on Zero Inflation Models

- Poisson is not nested in ZIP. γ = 0 in ZIP does not produce Poisson; it produces ZIP with P(regime 0) = ½.
  - Standard tests are not appropriate
  - Use Vuong statistic. ZIP model almost always wins.
- Zero Inflation models extend to NB models – ZINB is a standard model
  - Creates two sources of overdispersion
  - Generally difficult to estimate



ZIP Model

```
----------------------------------------------------------------------
Zero Altered Poisson      Regression Model
Logistic distribution used for splitting model.
ZAP term in probability is F[tau x Z(i)      ]
Comparison of estimated models
            Pr[0|means]       Number of zeros        Log-likelihood
Poisson           .04933   Act.= 10135 Prd.=  1347.9   -103727.29625
Z.I.Poisson       .36565   Act.= 10135 Prd.=  9991.8    -83843.36088
Vuong statistic for testing ZIP vs. unaltered model is      44.6739
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Poisson model.
A value less than -1.96 rejects the ZIP model.
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Poisson/NB/Gamma regression model
Constant|    1.47301***       .01123       131.119   .0000
     AGE|     .01100***       .00013        83.038   .0000      43.5257
    EDUC|    -.02164***       .00075       -28.864   .0000      11.3206
  FEMALE|     .10943***       .00256        42.728   .0000       .47877
 MARRIED|    -.02774***       .00318        -8.723   .0000       .75862
  HHNINC|    -.42240***       .00902       -46.838   .0000       .35208
  HHKIDS|    -.08182***       .00323       -25.370   .0000       .40273
        |Zero inflation model
Constant|    -.75828***       .06803       -11.146   .0000
  FEMALE|    -.59011***       .02652       -22.250   .0000       .47877
    EDUC|     .04114***       .00561         7.336   .0000      11.3206
--------+-------------------------------------------------------------
```



Marginal Effects for Different Models

Scale Factor for Marginal Effects 3.1835 POISSON

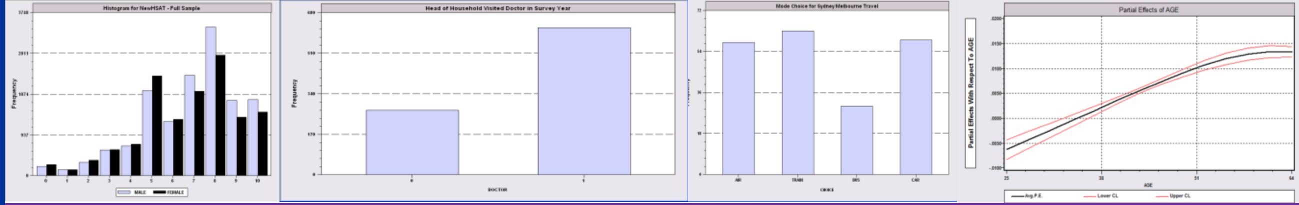
Variable	Coefficient	Standard Error	b/St.Er.	P[|Z|>z]	Mean of X
AGE	.05613***	.00131	42.991	.0000	43.5257
EDUC	-.09490***	.00596	-15.923	.0000	11.3206
FEMALE	.93237***	.02555	36.491	.0000	.47877
MARRIED	.03069	.02945	1.042	.2973	.75862
HHNINC	-1.66271***	.07803	-21.308	.0000	.35208
HHKIDS	-.51037***	.02879	-17.730	.0000	.40273

Scale Factor for Marginal Effects 3.1924 NEGATIVE BINOMIAL - 2

AGE	.05767***	.00317	18.202	.0000	43.5257
EDUC	-.11867***	.01348	-8.804	.0000	11.3206
FEMALE	1.04058***	.06212	16.751	.0000	.47877
MARRIED	-.01931	.06382	-.302	.7623	.75862
HHNINC	-1.49301***	.16272	-9.176	.0000	.35208
HHKIDS	-.48759***	.06022	-8.097	.0000	.40273

Scale Factor for Marginal Effects 3.1149 ZERO INFLATED POISSON

AGE	.03427***	.00052	66.157	.0000	43.5257
EDUC	-.11192***	.00662	-16.901	.0000	11.3206
FEMALE	.97958***	.02917	33.577	.0000	.47877
MARRIED	-.08639***	.01031	-8.379	.0000	.75862
HHNINC	-1.31573***	.03112	-42.278	.0000	.35208
HHKIDS	-.25486***	.01064	-23.958	.0000	.40273

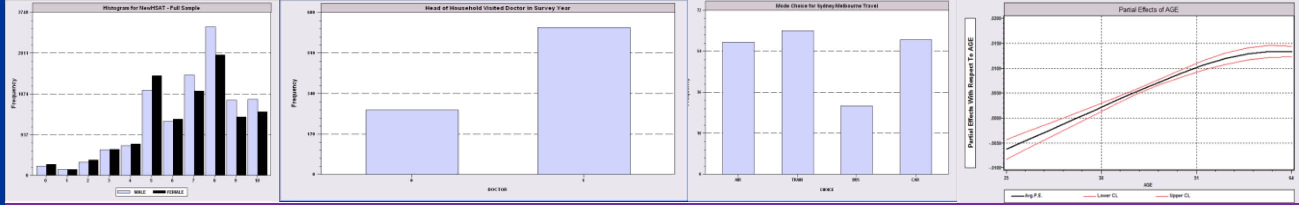


Zero Inflation Models

Zero Inflation = ZIP

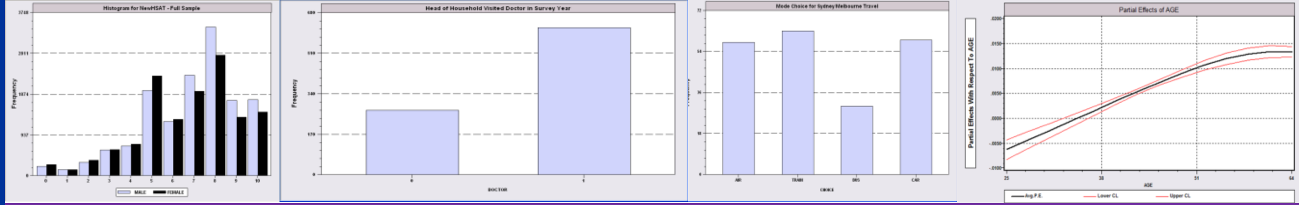
$$\text{Prob}(y_i = j \mid x_i) = \frac{\exp(-\lambda_i)\lambda_i^j}{j!}, \boxed{\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)}$$

$$\text{Prob}(0 \text{ regime}) = \boxed{F(\boldsymbol{\gamma}'\mathbf{z}_i)}$$



An Unidentified ZINB Model

```
---------------------------------------------------------------------------
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x Z(i)      ]
Comparison of estimated models
            Pr[0|means]      Number of zeros        Log-likelihood
Poisson          .04933    Act.= 10135 Prd.=  1348.0   -103726.69548
Neg. Bin.        .16540    Act.= 10135 Prd.=  4519.8    -60134.40266
Z.I.Neg_Bin      .32272    Act.= 10135 Prd.=  8818.5    -59964.41695
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is      8.7490
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.
--------+------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+------------------------------------------------------------------
        |Poisson/NB/Gamma regression model
Constant|    1.03203***     .06064    17.02  .0000      .91318   1.15088
     AGE|     .01556***     .00079    19.72  .0000      .01401    .01710
    EDUC|    -.03015***     .00397    -7.60  .0000     -.03793   -.02238
 MARRIED|    -.04791**      .01880    -2.55  .0108     -.08475   -.01107
  FEMALE|     .16415***     .01556    10.55  .0000      .13364    .19465
  INCOME|    -.48992***     .04562   -10.74  .0000     -.57934   -.40050
  HHKIDS|    -.12394***     .01741    -7.12  .0000     -.15807   -.08981
        |Dispersion parameter
   Alpha|    1.63685***     .01811    90.38  .0000     1.60136   1.67235
        |Zero inflation model
Constant|   -1.04171***     .19037    -5.47  .0000    -1.41483   -.66860
     AGE|    -.03193***     .00356    -8.97  .0000     -.03891   -.02495
    EDUC|     .07742***     .01161     6.67  .0000      .05467    .10016
 MARRIED|    -.64642***     .08468    -7.63  .0000     -.81239   -.48045
  FEMALE|   -25.0806       .1700D+10     .00 1.0000 *********** ***********
  INCOME|    -.33349*       .20013    -1.67  .0956     -.72574    .05876
  HHKIDS|     .66192***     .07974     8.30  .0000      .50564    .81819
--------+------------------------------------------------------------------
Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
---------------------------------------------------------------------------
```



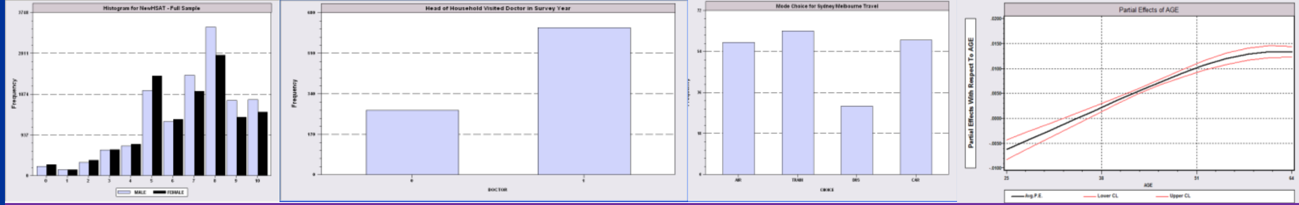
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x Z(i)]
Comparison of estimated models

	Pr[0|means]	Number of zeros		Log-likelihood
Poisson	.04933	Act.= 10135	Prd.= 1348.0	-103726.69548
Neg. Bin.	.16540	Act.= 10135	Prd.= 4519.8	-60134.40266
Z.I.Neg_Bin	.35100	Act.= 10135	Prd.= 9591.5	-60119.65754

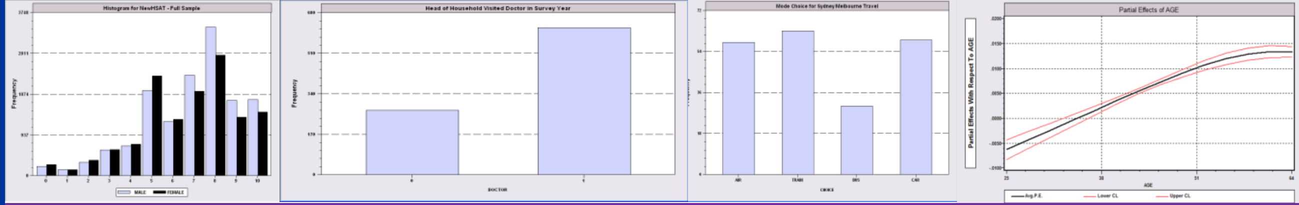
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is 2.5546
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Poisson/NB/Gamma regression model					
Constant	.91754***	.06026	15.23	.0000	.79943	1.03564
AGE	.01616***	.00080	20.17	.0000	.01459	.01773
EDUC	-.03320***	.00397	-8.37	.0000	-.04097	-.02542
MARRIED	-.01081	.01859	-.58	.5608	-.04724	.02562
FEMALE	.31349***	.01558	20.13	.0000	.28296	.34402
INCOME	-.53185***	.04506	-11.80	.0000	-.62016	-.44354
HHKIDS	-.15486***	.01715	-9.03	.0000	-.18848	-.12124
	Dispersion parameter					
Alpha	1.76738***	.02003	88.24	.0000	1.72812	1.80663
	Zero inflation model					
Constant	-1.41814***	.46248	-3.07	.0022	-2.32458	-.51170
AGE	-.05673***	.00988	-5.74	.0000	-.07609	-.03737
EDUC	.10051***	.02595	3.87	.0001	.04964	.15138
INCOME	-2.20648***	.58974	-3.74	.0002	-3.36234	-1.05061

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

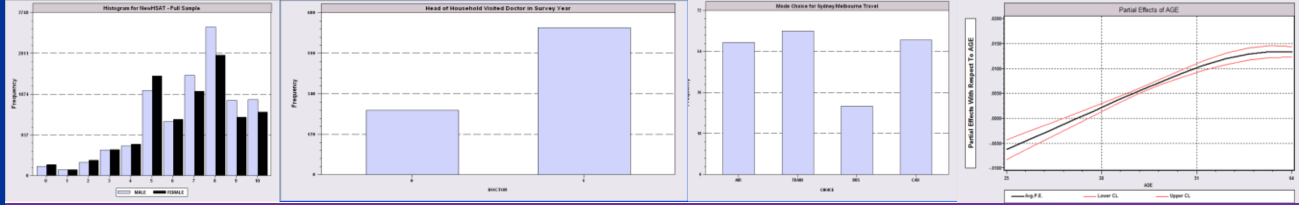



```
----------------------------------------------------------------------
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x Z(i)     ]
Comparison of estimated models
            Pr[0|means]       Number of zeros        Log-likelihood
Poisson          .04933    Act.= 10135 Prd.=  1348.0   -103726.69548
Neg. Bin.        .16540    Act.= 10135 Prd.=  4519.8    -60134.40266
Z.I.Neg_Bin      .35100    Act.= 10135 Prd.=  9591.5    -60119.65754
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is      2.5546
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.
--------+-------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------
        |Poisson/NB/Gamma regression model
Constant|     .91754***    .06026    15.23  .0000      .79943   1.03564
     AGE|     .01616***    .00080    20.17  .0000      .01459    .01773
    EDUC|    -.03320***    .00397    -8.37  .0000     -.04097   -.02542
 MARRIED|    -.01081       .01859     -.58  .5608     -.04724    .02562
  FEMALE|     .31349***    .01558    20.13  .0000      .28296    .34402
  INCOME|    -.53185***    .04506   -11.80  .0000     -.62016   -.44354
  HHKIDS|    -.15486***    .01715    -9.03  .0000     -.18848   -.12124
        |Dispersion parameter
   Alpha|    1.76738***    .02003    88.24  .0000     1.72812   1.80663
        |Zero inflation model
Constant|   -1.41814***    .46248    -3.07  .0022    -2.32458   -.51170
     AGE|    -.05673***    .00988    -5.74  .0000     -.07609   -.03737
    EDUC|     .10051***    .02595     3.87  .0001      .04964    .15138
  INCOME|   -2.20648***    .58974    -3.74  .0002    -3.36234  -1.05061
--------+-------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```



Partial Effects for Different Models

```
-----------------------------------------------------------------------------
Partial derivatives of expected val. with
respect to the vector of characteristics.
Effects are averaged over individuals.
Sample average conditional mean      3.1355
Effects of common variables in two part
models are added to obtain partial effect.
--------+--------------------------------------------------------------------
        |    Partial      Standard               Prob.       95% Confidence
  DOCVIS|     Effect        Error          z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
        |Index Function in Count Probability
     AGE|     .05962***     .00132     45.17   .0000       .05703     .06221
    EDUC|    -.11365***     .00662    -17.18   .0000      -.12661    -.10068
 MARRIED|    -.08626***     .00998     -8.64   .0000      -.10582    -.06670
  FEMALE|     .35856***     .00817     43.89   .0000       .34255     .37457
  INCOME|   -1.43400***     .08681    -16.52   .0000     -1.60414   -1.26385
  HHKIDS|    -.25824***     .01017    -25.40   .0000      -.27816    -.23831
        |Zero Inflation Probability
     AGE|     .05962***     .00132     45.17   .0000       .05703     .06221
    EDUC|    -.11365***     .00662    -17.18   .0000      -.12661    -.10068
  INCOME|   -1.43400***     .08681    -16.52   .0000     -1.60414   -1.26385
--------+--------------------------------------------------------------------
        |Poisson Scale;  Factor for Marginal Effects  3.1835
     AGE|     .05612***     .00112     50.15   .0000       .05393     .05832
    EDUC|    -.09482***     .00557    -17.03   .0000      -.10573    -.08391
 MARRIED|     .03070        .02770      1.11   .2677      -.02359     .08499  #
  FEMALE|     .92643***     .02215     41.82   .0000       .88302     .96985  #
  INCOME|   -1.66509***     .07217    -23.07   .0000     -1.80654   -1.52365
  HHKIDS|    -.49811***     .02552    -19.52   .0000      -.54812    -.44809  #
--------+--------------------------------------------------------------------
        |Negative Binomial;  Scale Factor for Marginal Effects  3.1924
     AGE|     .05766***     .00269     21.40   .0000       .05238     .06294
    EDUC|    -.11862***     .02068     -5.73   .0000      -.15916    -.07807
 MARRIED|    -.01920        .94978      -.02   .9839     -1.88073    1.84232  #
  FEMALE|    1.03174        .66197      1.56   .1191      -.26570    2.32918  #
  INCOME|   -1.49533***     .38141     -3.92   .0001     -2.24288    -.74777
  HHKIDS|    -.47626***     .07705     -6.18   .0000      -.62728    -.32524  #
```



The Vuong Statistic for Nonnested Models

Model 0: $\log L_{i,0} = \log f_0(y_i \mid x_i, \theta_0) = m_{i,0}$

Model 0 is the Zero Inflation Model

Model 1: $\log L_{i,1} = \log f_1(y_i \mid x_i, \theta_1) = m_{i,1}$

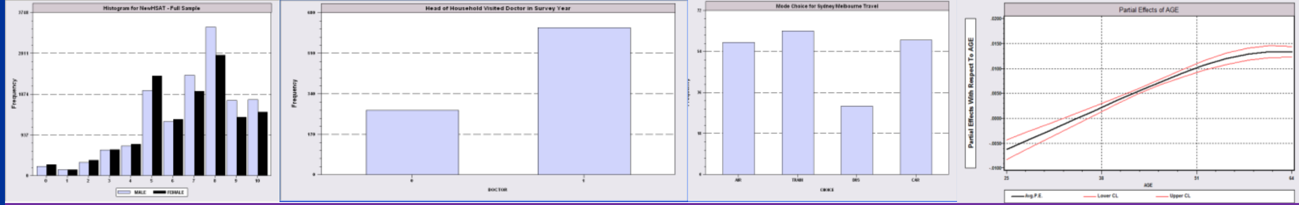
Model 1 is the Poisson model

(Not nested. α=0 implies the splitting probability is 1/2, not 1)

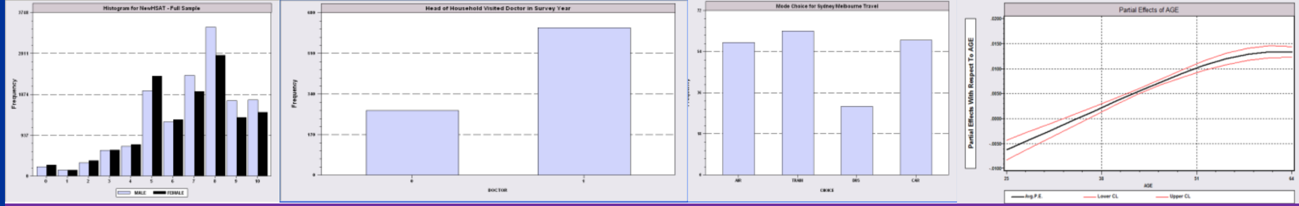
Define $a_i = m_{i,0} - m_{i,1} = \log \dfrac{f_0(y_i \mid x_i, \theta_0)}{f_1(y_i \mid x_i, \theta_1)}$

$$V = \frac{[\bar{a}]}{s_a / \sqrt{n}} = \frac{\sqrt{n}\left[\frac{1}{n}\Sigma_{i=1}^{n}\left(\log \frac{f_0(y_i \mid x_i, \theta_0)}{f_1(y_i \mid x_i, \theta_1)}\right)\right]}{\sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}\left[\log \frac{f_0(y_i \mid x_i, \theta_0)}{f_1(y_i \mid x_i, \theta_1)} - \overline{\log \frac{f_0(y_i \mid x_i, \theta_0)}{f_1(y_i \mid x_i, \theta_1)}}\right]^2}}$$

Limiting distribution is standard normal. Large + favors model 0, large - favors model 1, -1.96 < V < 1.96 is inconclusive.

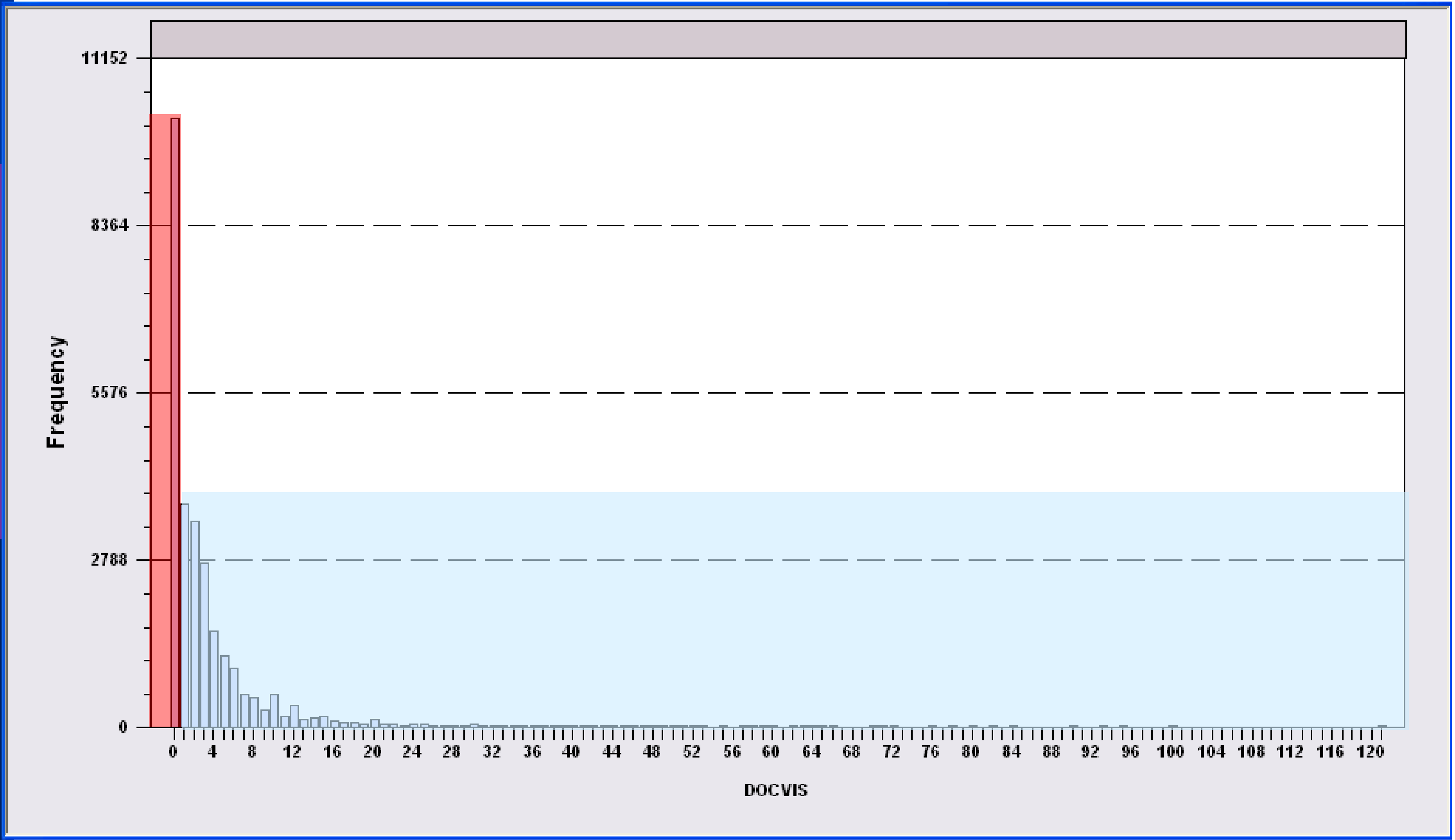


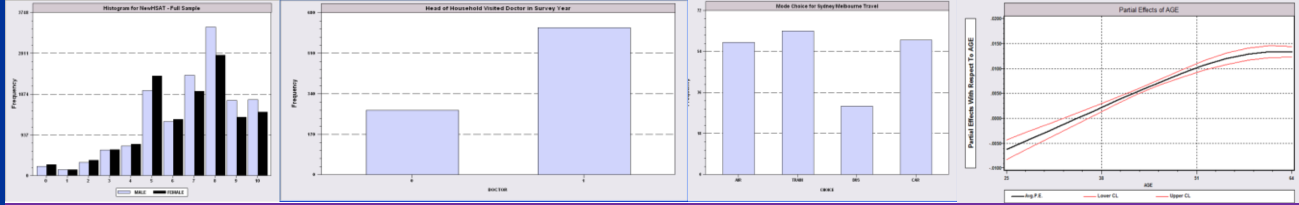
```
----------------------------------------------------------------------
Zero Inflated Neg.Binomial Regression Model
Logistic distribution used for splitting model.
ZIP term in probability is F[tau x Z(i)     ]
Comparison of estimated models
            Pr[0|means]       Number of zeros          Log-likelihood
Poisson          .04933    Act.= 10135 Prd.=  1348.0    -103726.69548
Neg. Bin.        .16540    Act.= 10135 Prd.=  4519.8     -60134.40266
Z.I.Neg_Bin      .35100    Act.= 10135 Prd.=  9591.5     -60119.65754
Note, the ZIP log-likelihood is not directly comparable.
ZIP model with nonzero Q does not encompass the others.
Vuong statistic for testing ZIP vs. unaltered model is      2.5546
Distributed as standard normal. A value greater than
+1.96 favors the zero altered Z.I.Neg_Bin model.
A value less than -1.96 rejects the ZIP model.
--------+-------------------------------------------------------------
```



A Hurdle Model

- Two part model:
  - Model 1: Probability model for more than zero occurrences
  - Model 2: Model for number of occurrences given that the number is greater than zero.
- Applications common in health economics
  - Usage of health care facilities
  - Use of drugs, alcohol, etc.


Frequency
11152
8364
5576
2788
0
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76
80
84
88
92
96
100
104
108
112
116
120
DOCVIS



Hurdle Model

Two Part Model

$$\text{Prob}[y > 0] = F(\boldsymbol{\gamma}'\mathbf{x})$$

$$\text{Prob}[y = j \mid y > 0] = \frac{\text{Prob}[y=j]}{\text{Prob}[y>0]} = \frac{\text{Prob}[y=j]}{1-\text{Prob}[y=0 \mid x]}$$

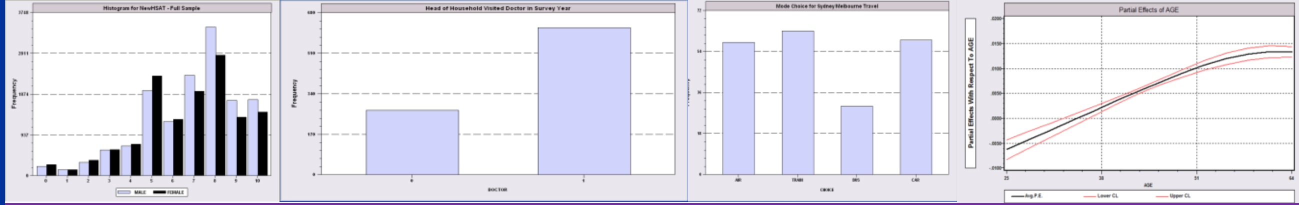
A Poisson Hurdle Model with Logit Hurdle

$$\text{Prob}[y>0]=\frac{\exp(\boldsymbol{\gamma}'\mathbf{x})}{1+\exp(\boldsymbol{\gamma}'\mathbf{x})}$$

$$\text{Prob}[y=j|y>0,x]=\frac{\exp(-\lambda)\lambda^{j}}{j![1-\exp(-\lambda)]},\ \ \lambda=\exp(\boldsymbol{\beta}'\mathbf{x})$$

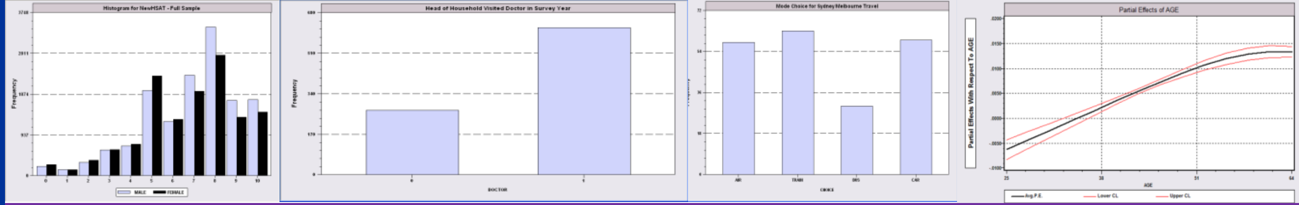
$$E[y|x] = 0 \times \text{Prob}[y=0]+\text{Prob}[y>0]\times E[y|y>0] = \frac{F(\boldsymbol{\gamma}'\mathbf{x})\exp(\boldsymbol{\beta}'\mathbf{x})}{1-\exp[-\exp(\boldsymbol{\beta}'\mathbf{x})]}$$

Marginal effects involve both parts of the model.



Hurdle Model for Doctor Visits

```
----------------------------------------------------------------------------
Poisson hurdle model for counts
Dependent variable               DOCVIS
Log likelihood function    -84241.29532
Restricted log likelihood -103726.69548
Chi squared [   1 d.f.]     38970.80031
Significance level               .00000
McFadden Pseudo R-squared      .1878533
Estimation based on N =  27326, K =  11
Inf.Cr.AIC  = 168504.6 AIC/N =    6.166
Tests of Model Restrictions on Neg.Bin.
Model               Logl ChiSquared[df]
Poisson(b=0)  -108662.14  ******** [**]
Poisson       -103726.70    9870.9 [ 6]
Hurdle Model   -84241.30   38970.8 [ 4]
LOGIT  hurdle equation
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
        |Parameters of count model equation
Constant|    1.53373***     .01053   145.61  .0000     1.51309    1.55437
     AGE|     .01088***     .00013    85.28  .0000      .01063     .01113
    EDUC|    -.02386***     .00072   -32.94  .0000     -.02528    -.02244
 MARRIED|    -.03458***     .00294   -11.77  .0000     -.04034    -.02883
  FEMALE|     .10240***     .00243    42.11  .0000      .09764     .10717
  INCOME|    -.46216***     .00874   -52.90  .0000     -.47929    -.44504
  HHKIDS|    -.07843***     .00301   -26.02  .0000     -.08434    -.07252
        |Parameters of binary hurdle equation
Constant|     .03578        .08574      .42  .6765     -.13228     .20383
     AGE|     .02471***     .00117    21.04  .0000      .02240     .02701
    EDUC|    -.04521***     .00561    -8.06  .0000     -.05620    -.03422
  INCOME|    -.16401**      .07373    -2.22  .0261     -.30851    -.01951
--------+-------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

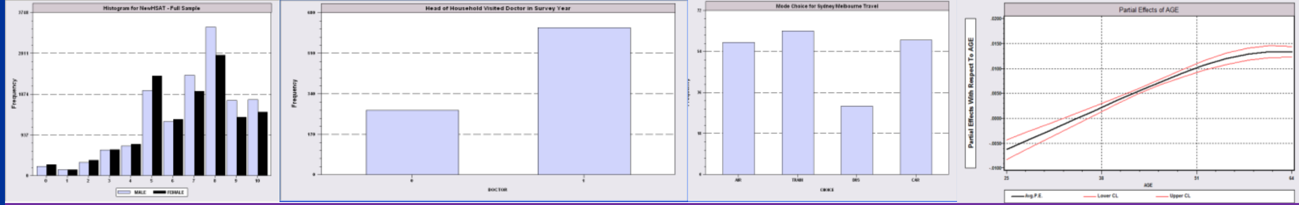


Partial Effects

Partial derivatives of expected val. with
respect to the vector of characteristics.
Effects are averaged over individuals.
Observations used for means are All Obs.
Sample average conditional mean 3.1632
Effects of common variables in two part
models are added to obtain partial effect.

DOCVIS	Partial Effect	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Effects in Count Model Equation					
AGE	.03306***	.00038	85.98	.0000	.03231	.03382
EDUC	-.07249***	.00218	-33.29	.0000	-.07676	-.06822
MARRIED	-.10508***	.00892	-11.78	.0000	-.12256	-.08759
FEMALE	.31113***	.00737	42.23	.0000	.29669	.32557
INCOME	-1.40418***	.02630	-53.39	.0000	-1.45573	-1.35263
HHKIDS	-.23829***	.00915	-26.04	.0000	-.25623	-.22036
	Effects in Binary Hurdle Equation					
AGE	.02737***	.00130	21.08	.0000	.02482	.02991
EDUC	-.05008***	.00621	-8.06	.0000	-.06225	-.03790
INCOME	-.18166**	.08166	-2.22	.0261	-.34171	-.02161
	Combined effect is the sum of the two parts					
AGE	.06043***	.00135	44.63	.0000	.05777	.06308
EDUC	-.12257***	.00659	-18.59	.0000	-.13549	-.10965
INCOME	-1.58584***	.08504	-18.65	.0000	-1.75252	-1.41916

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.



Journal of Health Economics 28 (2009) 265–279

Contents lists available at ScienceDirect

Journal of Health Economics

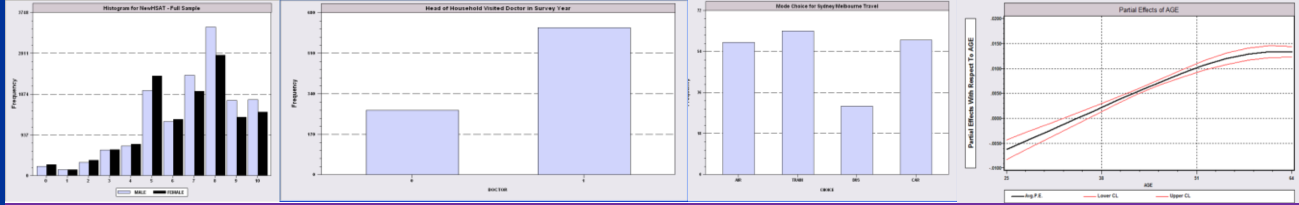
journal homepage: www.elsevier.com/locate/econbase

ELSEVIER

Health care utilisation in Europe: New evidence from the ECHP

Teresa Bago d'Uva [a,b,*], Andrew M. Jones [c]

[a] Department of Applied Economics (Room H13-09), Erasmus School of Economics, PB 1738, 3000 DR Rotterdam, The Netherlands
[b] Netspar, The Netherlands
[c] Department of Economics and Related Studies, University of York, York, YO10 5DD, United Kingdom



Following Bago d'Uva (2006), the class-specific density of the number of visits in a given year, $f_j(y_{it}|x_{it}, \theta_j)$, is defined as in the standard hurdle model, using a negative binomial as the parent distribution in both stages. Formally, for each component $j, j = 1, \ldots, C$, the probability of zero visits and the probability of observing y_{it} visits, given $y_{it} > 0$, are given by

$$f_j(0|x_{it};\theta_{j1}) = P[Y_{it} = 0|x_{it}, \theta_{j1}] = (\lambda_{j1,it}^{1-k} + 1)^{-\lambda_{j1,it}^{k}}$$

$$f_j(y_{it}|y_{it} > 0, x_{it};\theta_{j2}) = \frac{\Gamma\left(y + (\lambda_{j2,it}^{k}/\alpha_j)\right)\left(\alpha_j\lambda_{j2,it}^{1-k} + 1\right)^{-(\lambda_{j2,it}^{k}/\alpha_j)}[1 + (\lambda_{j2,it}^{k-1}/\alpha_j)]^{-y_{i,t}}}{\Gamma(\lambda_{j2,it}^{k}/\alpha_j)\Gamma(y_{it} + 1)\left[1 - (\alpha_j\lambda_{j2,it}^{1-k} + 1)^{-(\lambda_{j2,it}^{k}/\alpha_j)}\right]}, \tag{2}$$

where $\lambda_{j1,it} = \exp(x_{it}'\beta_{j1})$, $\lambda_{j2,it} = \exp(x_{it}'\beta_{j2})$, α_j are overdispersion parameters and k is an an arbitrary constant (most commonly set equal to 1 or 0, corresponding to the NegBin1 and NegBin2 models, respectively; we use the NegBin2 model). The vectors of parameters driving the probability of seeking care, β_{j1}, and the number of visits, given that this is positive, β_{j2}, are allowed to be different which means that the determinants of care may have different effects on the two stages of the decision process. This is also the case in the standard hurdle model, which corresponds to a (degenerate) LC model with only

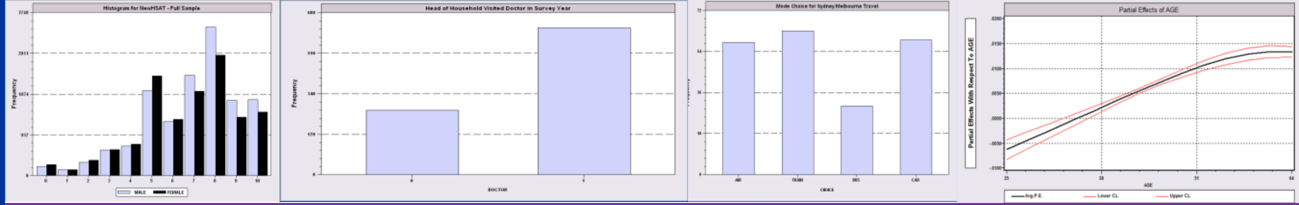
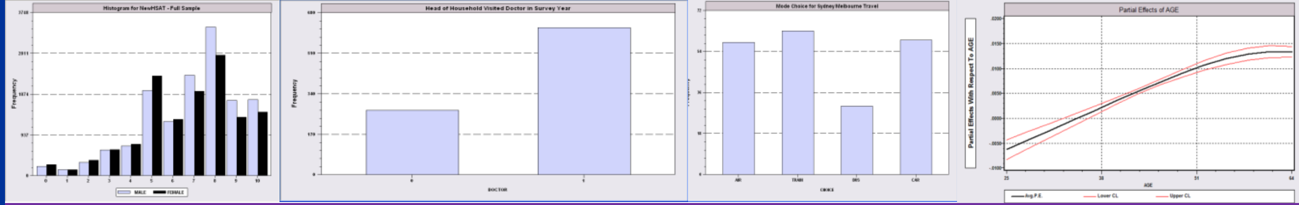


Table 2
Comparison of pooled models for GP and specialist visits.

Country	Hurdle		LC NegBin		LC hurdle	
	log L	BIC	log L	BIC	log L	BIC
GPs	−1189299.0	2379331.8	−1155544.2	2312125.4	−1147484.7	2296728.7
Specialists	−771788.0	1544309.9	−750839.7	1502716.2	−743979.6	1489718.2

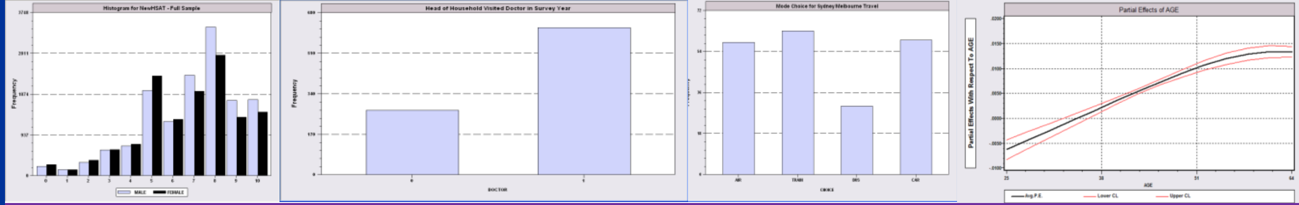


JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **19**: 455–472 (2004)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.764

HEALTH CARE REFORM AND THE NUMBER OF DOCTOR VISITS—AN ECONOMETRIC ANALYSIS

RAINER WINKELMANN*
Department of Economics, University of Zurich, Switzerland

Application of Several of the Models Discussed in this Section

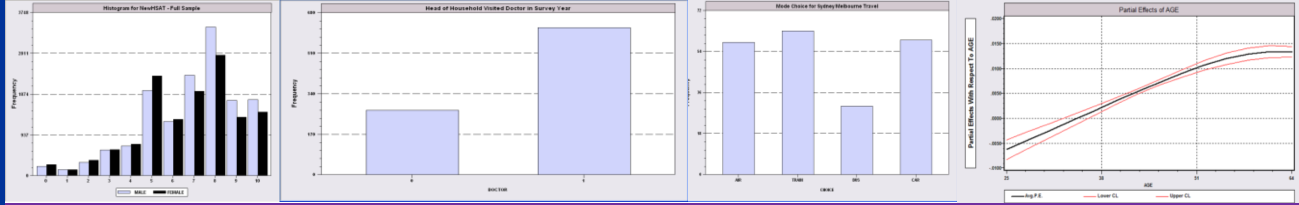


6.2. Adverse Selection

The above discussion of adverse selection in the German health insurance system yielded that those individuals who foresee higher health expenditures can be expected to purchase add-on insurance. For the empirical model this suggests two predictions. First, we should observe a positive correlation between holding add-on insurance and realized health care demand. Second, holding add-on insurance should be endogenous to subsequent health care demand.

6.3. Moral Hazard

Turning now to the incentives affecting demand for health care, moral hazard driven behaviour would imply that individuals with higher insurance coverage, and therefore lower opportunity cost, demand more health care. This has been tested numerous times in the literature.[15] However, the hypothesis could not be confirmed in panel data models for Germany so far.[16] Therefore we



JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **26**: 872–879 (2011)
Published online 2 June 2011 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.1252

REPLICATION SECTION

THE FREQUENCY OF VISITING A DOCTOR: IS THE DECISION TO GO INDEPENDENT OF THE FREQUENCY?

HANS VAN OPHEM*

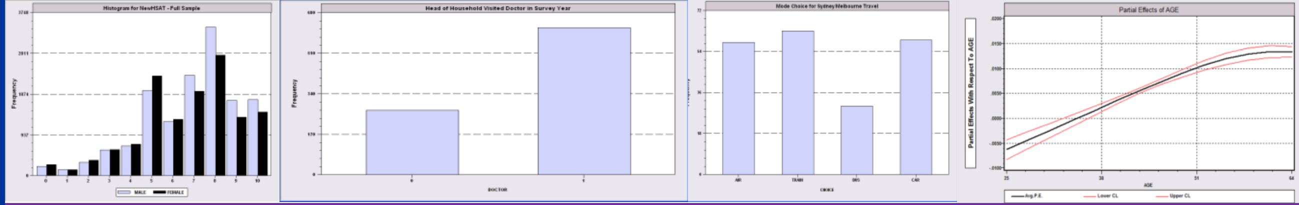
Department of Quantitative Economics, Faculty of Economics and Econometrics, University of Amsterdam, The Netherlands

SUMMARY

In his analysis of the effects of the reform of the German healthcare system, Winkelmann (*Journal of Applied Econometrics* 2004; **19**: 455–472) investigates the number of doctor visits. He makes a distinction between the decision to go to a physician and the number of times the physician is visited in the observed time period. Winkelmann finds that there is no correlation between both decisions. This result appears to be far from straightforward since the primary driving force in both decisions will be the health of the patient. From this perspective a significant correlation is expected. In this paper, I first replicate part of Winkelmann's research. I then set out to analyse whether the zero correlation is actually true or comes from the way the relation between both decisions is modelled. My empirical analysis confirms the latter, but nevertheless also corroborates Winkelmann's main conclusions on the relevance of the explanatory variables used in his investigation. Copyright © 2011 John Wiley & Sons, Ltd.

See also:
van Ophem H. 2000. Modeling selectivity in count data models. *Journal of Business and Economic Statistics* 18: 503–511.

Winkelmann finds that there is no correlation between the decisions… A significant correlation is expected … [T]he correlation comes from the way the relation between the decisions is modeled.



To model the potentially related decisions, Winkelmann (2004) decided to use a hurdle model. He recognizes that both decisions will usually not be made independently, and therefore Winkelmann allows for a correlation between both decision, although in a specific manner. First, let us discuss the reason for a correlation. Following Winkelmann (2004), I model the decision to go to a doctor or not by a simple probit model:

$$z_i = x_i'\gamma + \varepsilon_i \tag{1}$$

where z_i is an observation-specific latent indicator variable that exceeds zero if the individual decides not to visit a doctor ($d_i = 1$ in this case) and is smaller than or equal to zero if he decides to do so ($d_i = 0$). z_i could be seen as representing the health of the individual. Low values represent bad health and high values good health. This health indicator is dependent on exogenous factors

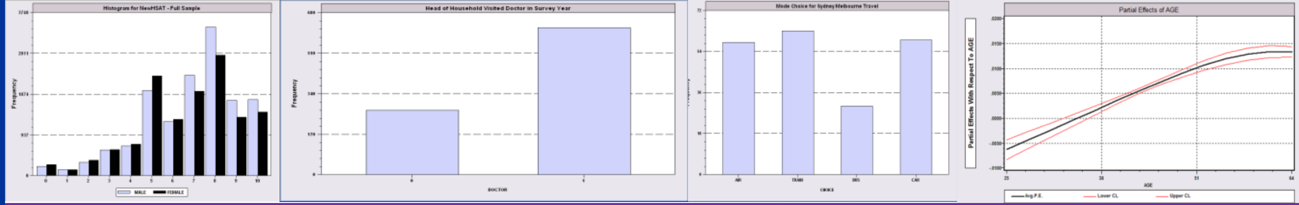
Probit Participation Equation

Poisson-Normal Intensity Equation

for example, factors that characterize the physician. We then arrive at the following specification of the expected frequency:

$$\lambda_i = \exp(\alpha z_i + w_i'\theta_i + \upsilon_i) = \exp(\alpha x_i'\gamma + w_i'\theta_i + (\upsilon_i + \alpha\varepsilon_i)) \tag{2}$$

where the substitution is made because the health indicator z_i is not actually observed. Winkelmann (2004) specifies $\lambda_i = \exp(x_i'\beta + u_i)$. Equation (2) shows two things. First, the error term introduced



Equation (2) also shows something else: if υ_i and ε_i are uncorrelated, then the correlation, in absolute value, between the error terms in (1) and (2) will be smaller than 1, except if $\text{var}(\upsilon_i) = \sigma^2 = 0$. To see this, calculate the correlation between ε_i and $\upsilon_i + \alpha\varepsilon_i$:

$$\text{corr}(\varepsilon_i, \upsilon_i + \alpha\varepsilon_i) = \frac{\rho\sigma + \alpha}{\sqrt{\sigma^2 + \alpha^2 + 2\rho\alpha\sigma}} \tag{3}$$

Only if the assumption of zero correlation between υ_i and ε_i is dropped, i.e. $\rho \neq 0$, is a correlation of $+1$ or -1 between ε_i and $\upsilon_i + \alpha\varepsilon$ possible. This overall correlation is not equal to the correlation between the hurdle and the count processes, however. For this last correlation, the additional randomness of the count should be taken into account. To illustrate this, consider the special case that $\upsilon_i = 0$. In this case, the correlation between the error terms in (1) and (2) is 1 if $\alpha > 0$. Despite this, conditioning on ε_i does not eliminate the randomness of the count process

Bivariate-Normal Heterogeneity in Participation and Intensity Equations

Gaussian Copula for Participation and Intensity Equations

To model the potential correlation between the hurdle and count fully, use will be made of a bivariate copula. The first application of the copula technique in economics on continuous variables is discussed in Lee (1983). Copulas offer a method to relate two (or more) marginal distributions, if necessary from different families, to each other and allow for the estimation of a correlation between the stochastic processes. I will use the Gaussian copula. The underlying stochastic variables, in the present case the decision to see a doctor or not (equation (1), where the error already is assumed to be normally distributed) and the number of doctor visits (a Poisson distributed count variable with unobserved heterogeneity) are transformed to normal distributed random variables and then evaluated using a bivariate normal distribution.[7] The transformation of an essentially discrete random variable like a count is not straightforward. The way to do this is

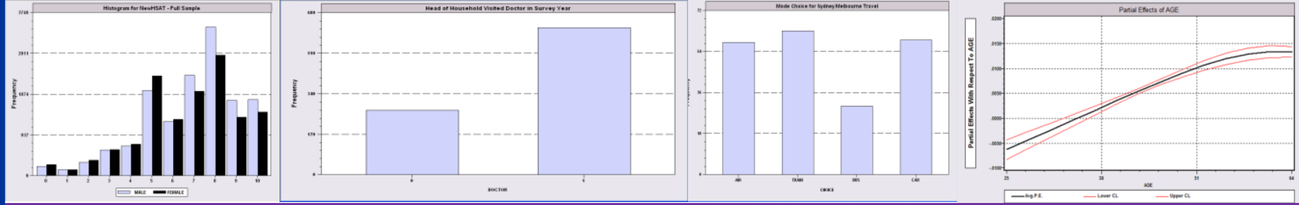


Table I. Estimation results of the Probit–Poisson–log-normal model with correlation

	Probit	Truncated Poisson–log-normal
Age	0.002 (0.002)	−0.001 (0.002)
Age^2	−0.046 (0.016)**	0.013 (0.016)
Male	0.442 (0.039)**	−0.100 (0.060)
Education	−0.014 (0.008)	−0.018 (0.008)*
Married	−0.115 (0.044)**	0.036 (0.045)
Household size	0.040 (0.015)**	−0.042 (0.016)**
Active sport	−0.131 (0.040)**	0.029 (0.043)
Good health	0.498 (0.040)**	−0.465 (0.069)**
Bad health	−0.567 (0.068)**	0.657 (0.065)**
Social assistance	−0.060 (0.101)	0.143 (0.094)
Log(income)	−0.109 (0.047)*	0.036 (0.048)
Variance unobserved heterogeneity		0.781 (0.016)**
Correlation		0.001 (0.280)

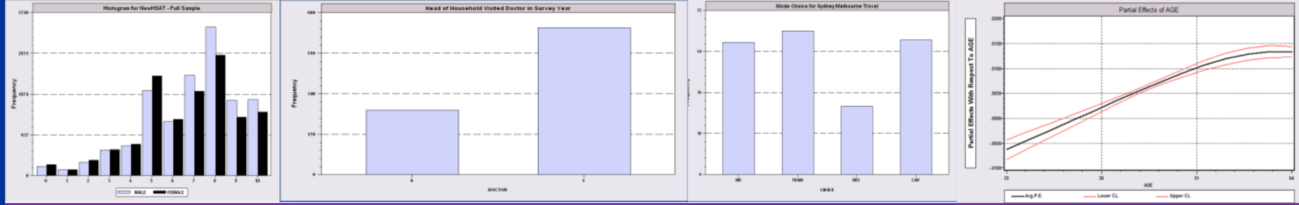
Log-likelihood = −11895.94. Absolute asymptotic standard errors between parentheses. Asterisks indicate significance at ** 1%; * 5% (two-sided test). The model also contains the variables: a constant, full-time, part-time, unemployed, spring, fall, winter in both the Probit and the truncated Poisson–log-normal.

Correlation between Heterogeneity Terms

Table II. Estimation results of the Probit–Poisson–log-normal model with correlation using a copula

	Probit	Truncated Poisson–log-normal
Age	0.001 (0.002)	−0.002 (0.002)
Age^2	−0.041 (0.015)**	0.024 (0.017)
Male	0.424 (0.040)**	−0.209 (0.046)**
Education	−0.016 (0.008)*	−0.015 (0.008)
Married	−0.108 (0.044)*	0.063 (0.046)
Household size	0.039 (0.014)**	−0.051 (0.015)**
Active sport	−0.127 (0.041)**	0.062 (0.043)
Good health	0.487 (0.040)**	−0.612 (0.049)**
Bad health	−0.564 (0.067)**	0.768 (0.055)**
Social assistance	−0.048 (0.105)	0.167 (0.100)
Log(income)	−0.108 (0.047)*	0.063 (0.050)
Variance unobs. heterogeneity (σ^2)		0.866 (0.028)**
Correlation		−0.467 (0.078)**

Correlation between Counts

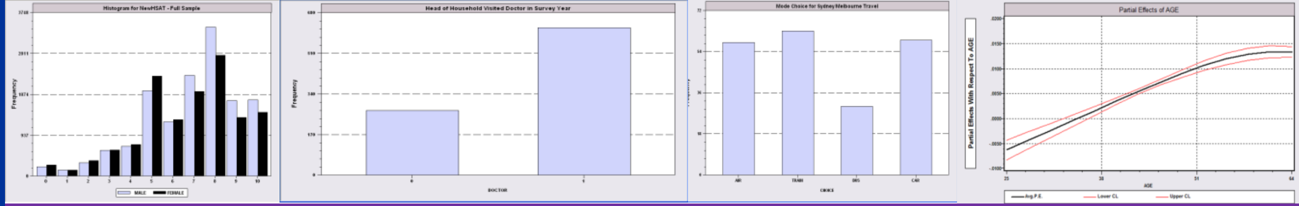


Panel Data Models

Heterogeneity; $\lambda_{it} = \exp(\boldsymbol{\beta}'\mathbf{x}_{it} + c_i)$

- Fixed Effects
  - Poisson: Standard, no incidental parameters issue
  - Negative Binomial
    - Hausman, Hall, Griliches (1984) put FE in variance, not the mean
    - Use “brute force” to get a conventional FE model
- Random Effects
  - Poisson
    - Log-gamma heterogeneity becomes an NB model
    - Contemporary treatments are using normal heterogeneity with simulation or quadrature based estimators
  - NB with random effects is equivalent to two “effects” one time varying one time invariant. The model is probably overspecified

Random Parameters: Mixed models, latent class models, hierercherical – all extended to Poisson and NB



JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. 18: 387–405 (2003)
Published online 8 October 2002 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.680

INCENTIVE EFFECTS IN THE DEMAND FOR HEALTH CARE: A BIVARIATE PANEL COUNT DATA ESTIMATION

REGINA T. RIPHAHN,[a–c*] ACHIM WAMBACH[b,d] AND ANDREAS MILLION[a,c]
[a] University of Munich, Germany
[b] CEPR, London, UK
[c] IZA, Bonn, Germany
[d] University of Erlangen-Nuremberg, Germany

Poisson (log)Normal Mixture

The restrictions of the basic Poisson model may more appropriately be overcome by an alternative set of distributional assumptions. Preston (1948) proposed a lognormal error distribution which leads to a Poisson lognormal distribution. This model has the following form:

$$p(y_i = a|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^a}{a!} \tag{5}$$

$$\ln(\lambda_i) \sim N(\beta' x_i, \sigma) \tag{6}$$

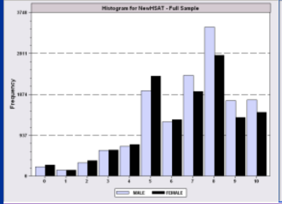
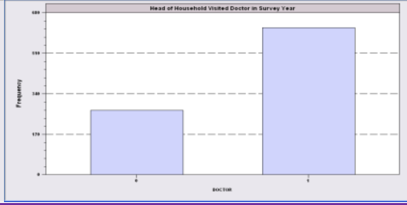
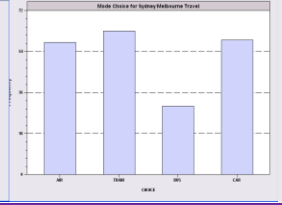
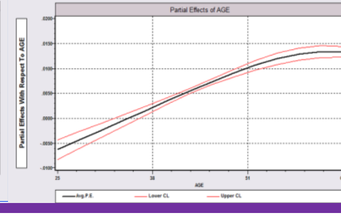
The likelihood function reveals immediately why this model has not been as popular as NEGBIN in the past:

$$L(\beta, \sigma) = \prod_{i=1}^{n} \int_0^{\infty} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \frac{1}{\sqrt{2\pi}\sigma\lambda_i} e^{-\frac{(\ln\lambda_i - \beta' x_i)^2}{2\sigma^2}} d\lambda_i \tag{7}$$

As in the NEGBIN case the error has to be integrated out. But here the resulting integrals cannot be solved analytically and thus the likelihood function contains integrals, which can only be solved using numerical integration procedures. Such procedures have been applied in econometrics before and can now be utilized at drastically reduced evaluation times due to the progress in computer processing power. Therefore assuming lognormally distributed errors is now a feasible alternative to the NEGBIN model. The advantage of using the assumption of lognormality, which we label the LogNormalP model is that it provides a flexible framework for fully parametric count data models. To see this we reformulate statement (6) such that:

$$\ln(\lambda_i) = \beta' x_i + \varepsilon_i \tag{8}$$

where ε_i is a normally distributed error term. The right-hand side of equation (8) is familiar

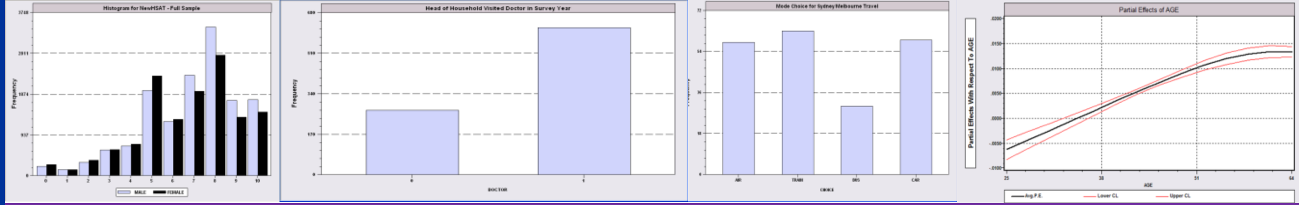






```
Poisson Model with Normal Heterogeneity
Dependent variable                 DOCVIS
Log likelihood function      -60118.36777
Restricted log likelihood -226132.77062
Chi squared [    1 d.f.]     332028.80572
Significance level                 .00000
McFadden Pseudo R-squared        .7341457
Estimation based on N =   27326, K =    8
Inf.Cr.AIC  = 120252.7 AIC/N =      4.401
Mean of LHS Variable =            3.18352
Vuong Stat. vs. Poisson =        35.87191
Vuong test favors extended model.
```

DOCVIS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Parameters of Poisson Probability					
Constant	-.33298***	.06698	-4.97	.0000	-.46426	-.20170
AGE	.02012***	.00086	23.41	.0000	.01843	.02180
EDUC	-.02565***	.00423	-6.06	.0000	-.03394	-.01736
MARRIED	.04342**	.02143	2.03	.0427	.00143	.08542
FEMALE	.46375***	.01718	26.99	.0000	.43006	.49743
INCOME	-.37823***	.05361	-7.06	.0000	-.48330	-.27316
HHKIDS	-.22331***	.02013	-11.09	.0000	-.26278	-.18385
	Standard Deviation of Heterogeneity					
Sigma	1.32247***	.00659	200.53	.0000	1.30954	1.33539

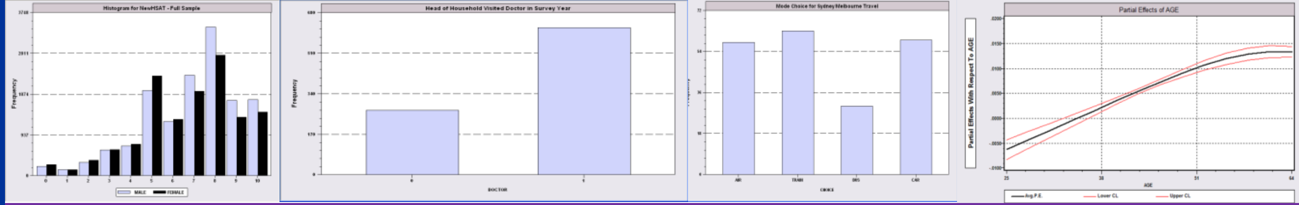
Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

Constant	.80836***	.05956	13.57	.0000	.69163	.92508
AGE	.01806***	.00079	22.78	.0000	.01651	.01962
EDUC	-.03716***	.00386	-9.62	.0000	-.04473	-.02958
MARRIED	-.00601	.01880	-.32	.7493	-.04285	.03083
FEMALE	.32594***	.01586	20.55	.0000	.29486	.35702
INCOME	-.46841***	.04662	-10.05	.0000	-.55978	-.37703
HHKIDS	-.15273***	.01729	-8.83	.0000	-.18662	-.11883
	Dispersion parameter for count data model					
Alpha	1.89677***	.01981	95.75	.0000	1.85794	1.93560



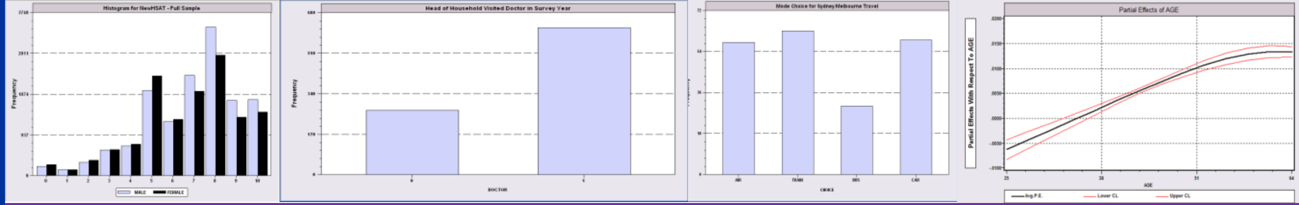
A Peculiarity of the FENB Model

- ‘True’ FE model has $\lambda_i=\exp(\alpha_i+x_{it}'\beta)$. Cannot be fit if there are time invariant variables.
- Hausman, Hall and Griliches (Econometrica, 1984) has α_i appearing in θ.
  - Produces different results
  - Implies that the FEM can contain time invariant variables.




```
----------------------------------------------------------------------
Panel Model with Group Effects
Dependent variable                DOCVIS
Log likelihood function     -33952.55746
Estimation based on N =  27326, K =    6
Inf.Cr.AIC  =  67917.1 AIC/N =     2.485
Model estimated: May 01, 2013, 18:34:56
Unbalanced panel has   7293 individuals
Neg.Binomial Regression - Fixed Effects
--------+-------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error        z   |z|>Z*         Interval
--------+-------------------------------------------------------------
Constant|  -1.49011***    .09565   -15.58  .0000   -1.67759  -1.30264
     AGE|    .02347***    .00119    19.66  .0000     .02113    .02580
    EDUC|    .02778***    .00628     4.42  .0000     .01546    .04009
  INCOME|    .01379       .05509      .25  .8024    -.09419    .12176
  HHKIDS|   -.03733*      .02123    -1.76  .0787    -.07894    .00428
  FEMALE|    .39355***    .02531    15.55  .0000     .34393    .44316
--------+-------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```

```
|-> negb;lhs=docvis;panel;fem;
    rhs=one,age,educ,income,hhkids,female$
+-----------------------------------------------------------------+
| Variable = ____________ Variable Groups     Max    Min   Average |
| TI         Group sizes  ID          7293       7      1       3.7 |
+-----------------------------------------------------------------+
+----------------------------------------------------+
| Frequency count for group sizes of TI              |
| Group size =    1    Pct =  20.91%   CumPct =  20.91% |
| Group size =    2    Pct =  14.80%   CumPct =  35.71% |
| Group size =    3    Pct =  11.31%   CumPct =  47.02% |
| Group size =    4    Pct =  12.70%   CumPct =  59.71% |
| Group size =    5    Pct =  14.41%   CumPct =  74.13% |
| Group size =    6    Pct =  13.71%   CumPct =  87.84% |
| Group size =    7    Pct =  12.16%   CumPct = 100.00% |
+----------------------------------------------------+
Variable FEMALE    has no within group variation.
It is not possible to fit the fixed effects model with
this sample of data with this variable(s) in the model.
Error    630: Cannot fit FEM with time invariant variables.
```



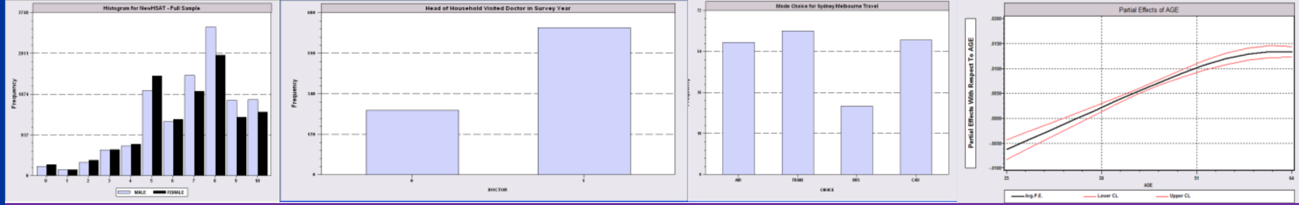
Available online at www.sciencedirect.com
ScienceDirect
Economics Letters 99 (2008) 63 – 66
economics letters
www.elsevier.com/locate/econbase

The fixed effects negative binomial model revisited

Paulo Guimarães *
Division of Research, Moore School of Business, University of South Carolina, 1705 College Street, Columbia, SC 29208, USA

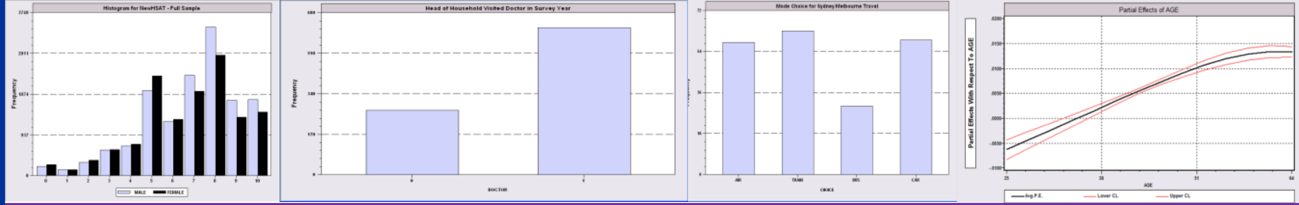
Received 6 February 2006; received in revised form 17 April 2007; accepted 24 May 2007
Available online 31 May 2007

See: Allison and Waterman (2002),
Guimaraes (2007)
Greene, Econometric Analysis (2012)




```
----------------------------------------------------------------------
Panel Model with Group Effects
Dependent variable                DOCVIS
Log likelihood function    -34017.10315
Estimation based on N =  27326, K =    5
Inf.Cr.AIC  =  68044.2 AIC/N =    2.490
Model estimated: May 01, 2013, 18:40:12
Unbalanced panel has   7293 individuals
Neg.Binomial Regression - Fixed Effects
--------+-------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------
Constant|  -1.15867***    .09377   -12.36  .0000   -1.34245   -.97488
     AGE|    .02393***    .00119    20.15  .0000     .02160    .02625
    EDUC|    .01435**     .00629     2.28  .0224     .00203    .02667
  INCOME|    .01922       .05527      .35  .7280    -.08911    .12755
  HHKIDS|   -.03385       .02116    -1.60  .1097    -.07533    .00763
--------+-------------------------------------------------------------
```

```
----------------------------------------------------------------------
FIXED EFFECTS NegBin Model
Dependent variable                DOCVIS
Log likelihood function    -49478.04763
Estimation based on N =  27326, K =6145
Inf.Cr.AIC  = 111246.1 AIC/N =    4.071
Model estimated: May 01, 2013, 18:40:37
Unbalanced panel has   7293 individuals
Skipped 1153 groups with inestimable ai
Negative binomial regression model
--------+-------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCVIS|  Coefficient     Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------
        |Index function for probability
     AGE|    .04471***    .00277    16.15  .0000     .03929    .05014
    EDUC|   -.04800       .02963    -1.62  .1053    -.10608    .01009
  INCOME|   -.20067***    .07321    -2.74  .0061    -.34416   -.05718
  HHKIDS|   -.00125       .02921     -.04  .9657    -.05850    .05599
        |Overdispersion parameter
   Alpha|   1.91945***    .02993    64.13  .0000    1.86079   1.97812
--------+-------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------
```



JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **18**: 387–405 (2003)
Published online 8 October 2002 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jae.680

INCENTIVE EFFECTS IN THE DEMAND FOR HEALTH CARE: A BIVARIATE PANEL COUNT DATA ESTIMATION

REGINA T. RIPHAHN,[a–c]* ACHIM WAMBACH[b,d] AND ANDREAS MILLION[a,c]
[a] *University of Munich, Germany*
[b] *CEPR, London, UK*
[c] *IZA, Bonn, Germany*
[d] *University of Erlangen-Nuremberg, Germany*

Bivariate Random Effects

Given the observed correlation between the two dimensions of health care demand, i.e. doctor and hospital visits, it appears plausible to formulate an estimation model which accounts for possible correlation in these two outcomes' unobservable determinants. In our bivariate panel model we assume that every person i has g different outcome observations y (i.e. doctor and hospital visits) in any period t.

$$y_{itg} \sim Po(\lambda_{itg}), \quad g = 1, 2 \tag{9}$$

$$\ln(\lambda_{it1}) = \beta_1' x_{it1} + u_{i1} + \varepsilon_{it1} \tag{10}$$

$$\ln(\lambda_{it2}) = \beta_2' x_{it2} + u_{i2} + \varepsilon_{it2} \tag{11}$$

$$\boxed{(\varepsilon_{it1}, \varepsilon_{it2}) \sim N_2(0, 0, \sigma_{\varepsilon 1}^2, \sigma_{\varepsilon 2}^2, \rho)} \tag{12}$$

$$u_{i1} \sim N(0, \sigma_{u1}^2) \tag{13}$$

$$u_{i2} \sim N(0, \sigma_{u2}^2) \tag{14}$$

$$E[\varepsilon_{itg} u_{jh}] = 0 \quad \forall\ i, t, g, j, h \tag{15}$$

$$E[\varepsilon_{itg} \varepsilon_{jsh}] = 0 \quad \text{if } t \neq s \vee i \neq j \vee g \neq h \tag{16}$$

$$E[u_{ig} u_{jh}] = 0 \quad \text{if } i \neq j \vee g \neq h \tag{17}$$

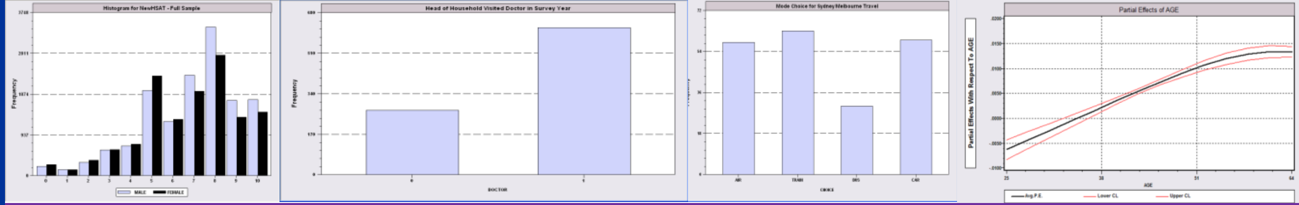


Table V. Bivariate PanelLogNormalP

	Males		Females	
	Doctor	Hospital	Doctor	Hospital
Constant	2.563**	−0.206	2.423**	−1.567**
	(8.52)	(−0.27)	(9.40)	(−2.29)
age	−0.060**	−0.077**	−0.040**	−0.032
	(−4.24)	(−2.35)	(−3.73)	(−1.11)
$age^2 \cdot 10^{-3}$	0.823**	0.942**	0.499**	0.234
	(5.10)	(2.49)	(4.15)	(0.73)
health	−0.237**	−0.243**	−0.191**	−0.196**
	(−33.07)	(−16.03)	(−35.81)	(−13.41)
handicap	−0.029	−0.086	0.063	0.039
	(−0.63)	(−0.73)	(1.44)	(0.34)
h_degree	0.007**	0.008**	0.004**	0.010**
	(6.90)	(3.27)	(4.75)	(4.67)
married	0.085	−0.054	0.009	−0.044
	(1.64)	(−0.49)	(0.26)	(−0.48)
schooling	−0.022**	−0.051**	0.014	−0.015
	(−2.39)	(−2.35)	(1.62)	(−0.71)
hhincome	−0.090	0.375	−0.107	0.407*
	(−0.91)	(1.55)	(−1.31)	(1.92)
child	−0.059	0.103	−0.117**	0.073
	(−1.44)	(1.08)	(−3.58)	(0.82)
σ_ε	0.996**	1.244**	0.822**	1.053**
	(64.85)	(25.18)	(80.89)	(22.82)
σ_u	0.795**	1.195**	0.701**	1.123**
	(34.75)	(19.90)	(42.23)	(21.88)
ρ	0.490**		0.386**	
	(14.27)		(9.94)	
lnL	−31,227.1		−34,391.8	
n	14,243		13,083	
time in h	17		14	