



NYU STERN

NEW YORK UNIVERSITY · LEONARD N. STERN SCHOOL OF BUSINESS

The 2010 Medici Summer School in Management Studies

William Greene

Department of Economics

Stern School of Business



Econometric Models When There Are Unusual Events



Part 5: Binary Outcomes



Agenda

- General modeling for binary choices
- Problem of unbalanced data – “rare events”
- A proposed statistical approach
- Application to credit card defaults

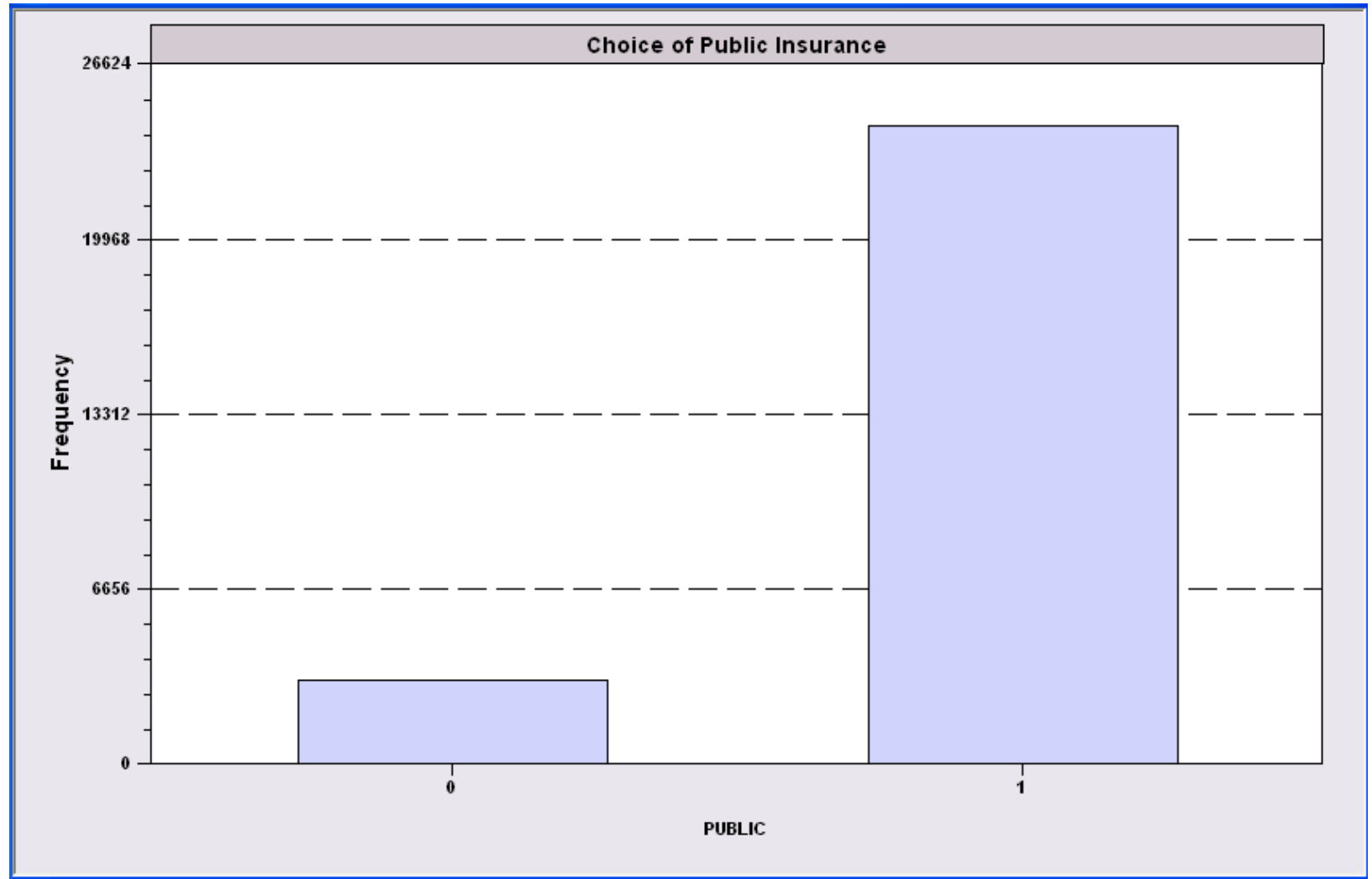


Model Framework

- Binary outcome: Default in time period $[t, t+\Delta]$ is 0/1, yes or no
- Covariates: Economic conditions, individual characteristics
- Linear regression is inappropriate

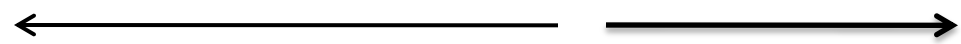
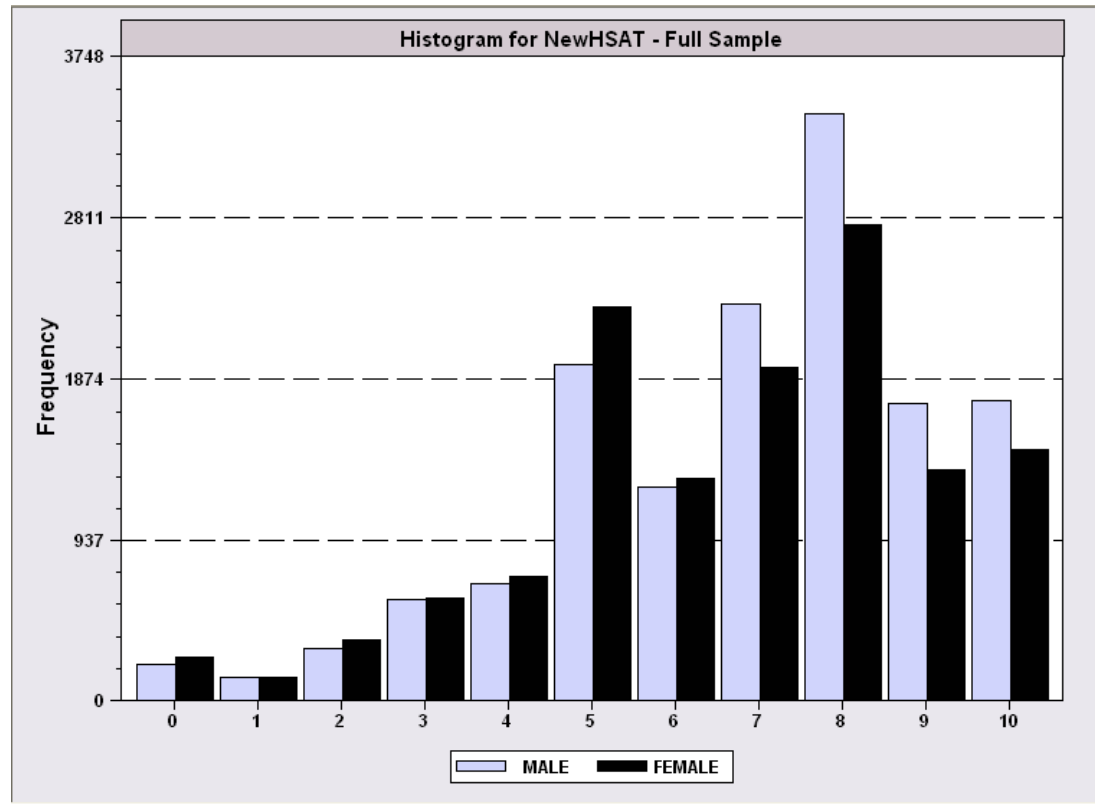


Simple Binary Choice: Public Insurance





Censored Health Satisfaction Scale

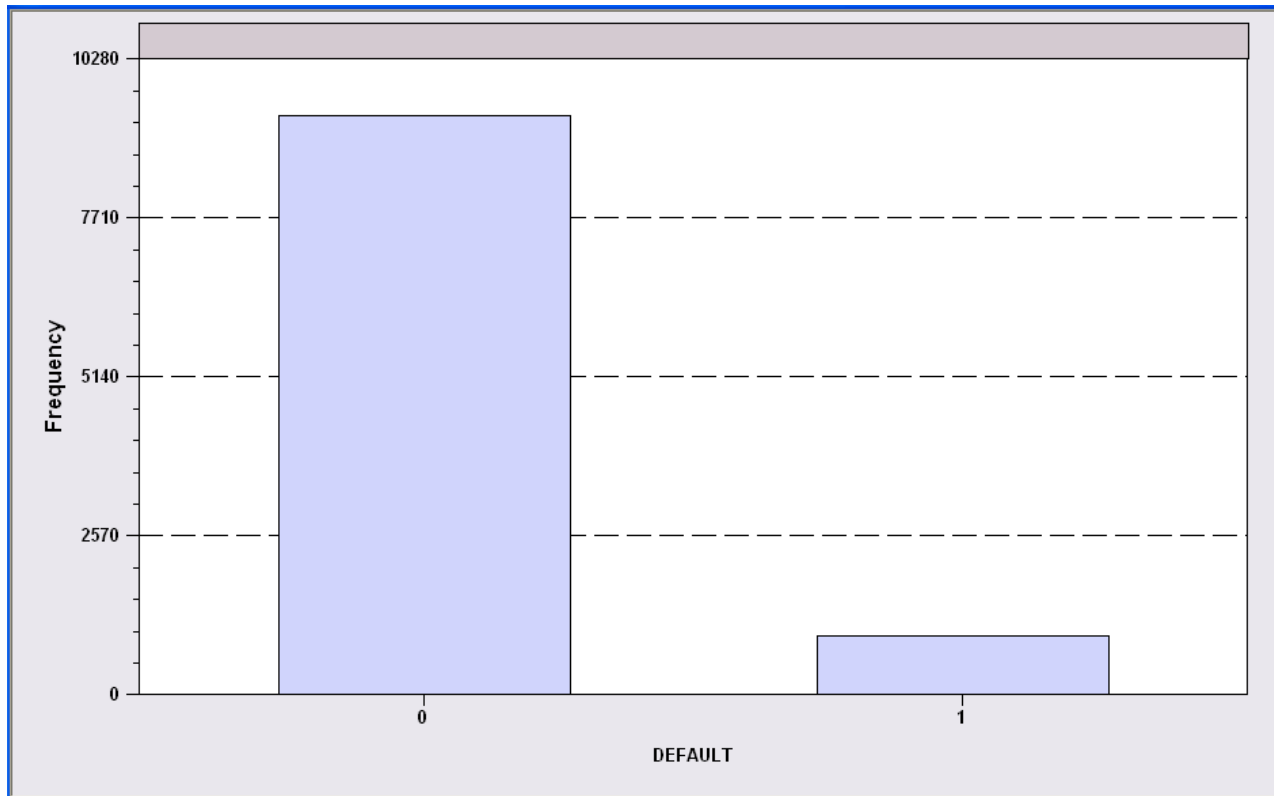


0 = Not Healthy

1 = Healthy



Default by Credit Cardholders





Modeling the Event

- Discriminant analysis and Z scores
 - Two populations
 - Membership is unknown a priori
 - “Discriminant function” $Z = a + bx$ is used to classify: If $Z > a$, classify as group 1 (default)
- Binary choice analysis
 - One population
 - Membership is only probabilistic
 - Random utility function, $U^* = a + bx + e$
 - Utility function implies a probability of group 1



A Random Utility Approach

- Underlying Preference Scale, $U^*(x_1 \dots)$
- Revelation of Preferences:
 - $U^*(x_1 \dots) \leq 0 \implies$ Choice “0”
 - $U^*(x_1 \dots) > 0 \implies$ Choice “1”



A Model for Binary Choice

- Yes or No decision (Buy/Not buy, Do/Not Do)
- Example, choose to visit physician or not
- Model: Net utility of visit at least once

$$U_{\text{visit}} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \gamma \text{Sex} + \varepsilon$$

Choose to visit if net utility is positive

$$\text{Net utility} = U_{\text{visit}} - U_{\text{not visit}}$$

- Data: $X = [1, \text{age}, \text{income}, \text{sex}]$
 $y = 1$ if choose visit, $\Leftrightarrow U_{\text{visit}} > 0$, 0 if not.



What Can Be Learned from the Data? (A Sample of Consumers, $i = 1, \dots, N$)

- Are the characteristics “relevant?”
- Predicting behavior
 - Individual – Will a person buy the add-on insurance?
Will a particular bondholder default?
 - Aggregate – What proportion of the population will buy the add-on insurance?
What proportion of bonds will default?
- Analyze changes in behavior when attributes change –
E.g., how will changes in education change the proportion who buy the insurance?



Choosing Between the Two Alternatives

Modeling the Binary Choice

$$U_{\text{visit}} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_1 \text{Sex} + \varepsilon$$

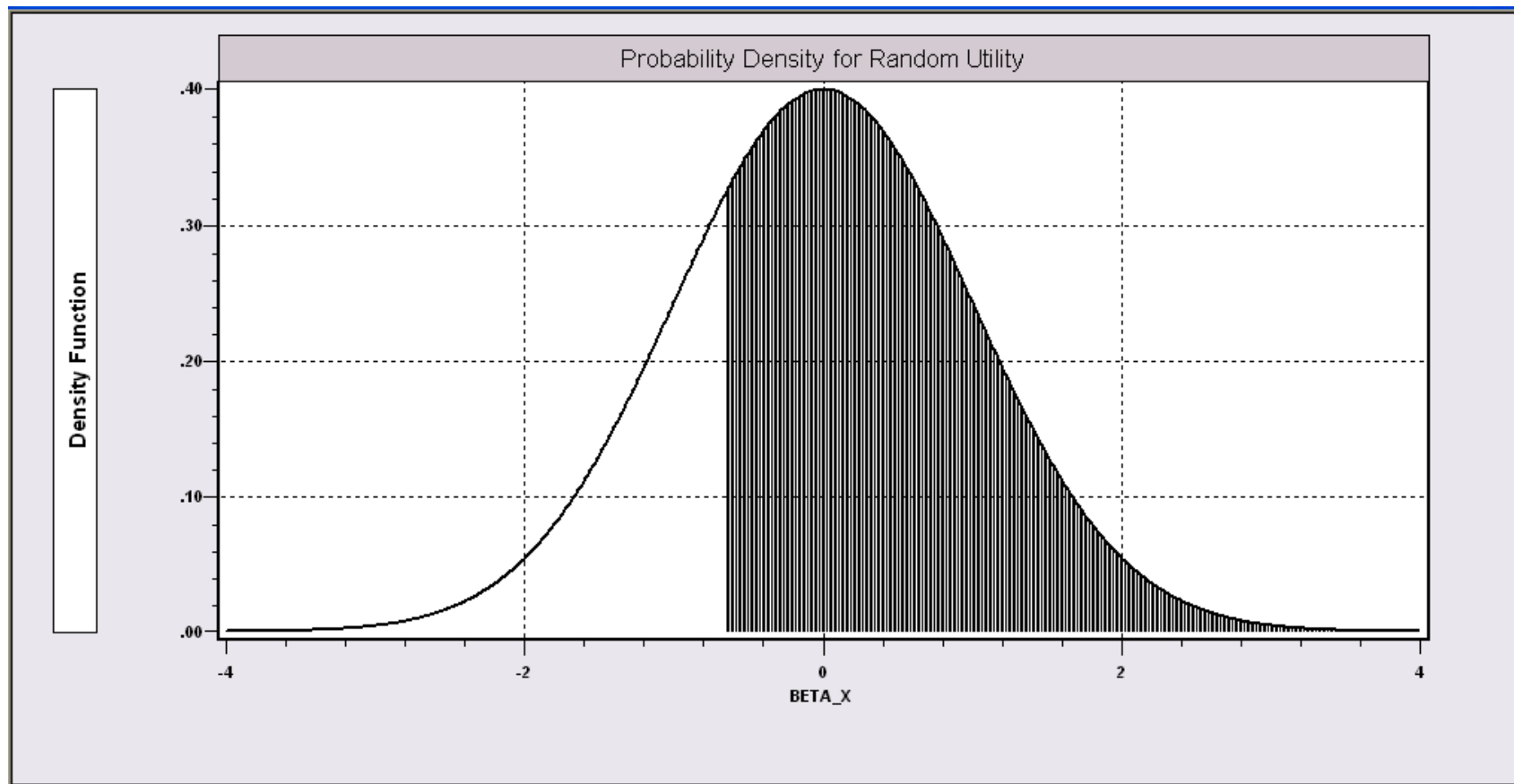
Chooses to visit: $U_{\text{visit}} > 0$

$$\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_1 \text{Sex} + \varepsilon > 0$$

$$\varepsilon \geq -[\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_1 \text{Sex}]$$



Probability Model for Choice Between Two Alternatives



$$\varepsilon > -[\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}]$$

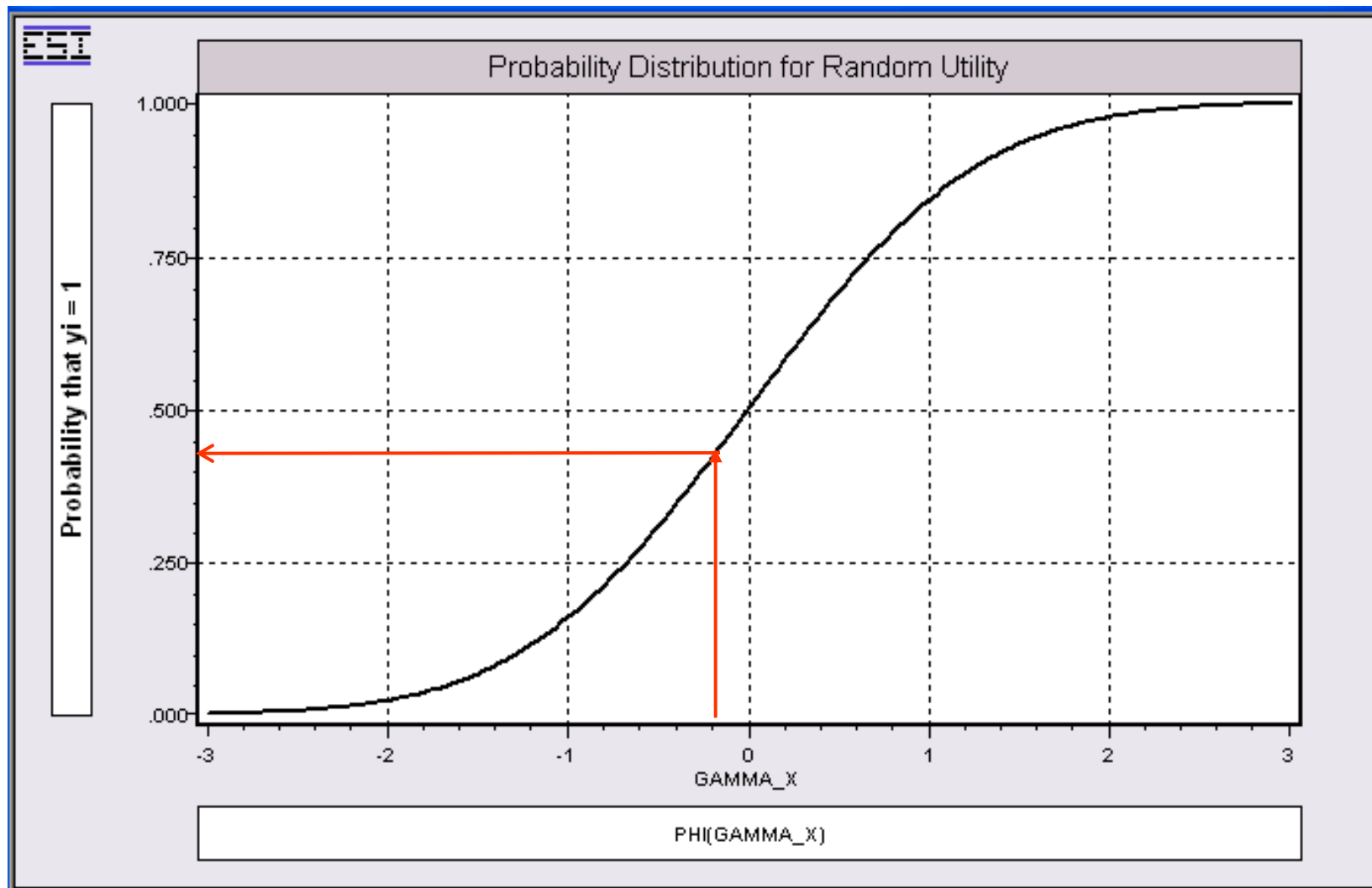


An Econometric Model

- Choose to visit iff $U_{\text{visit}} > 0$
 - $U_{\text{visit}} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon$
 - $U_{\text{visit}} > 0 \Leftrightarrow \varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})$
- Probability model: For any person observed by the analyst,
Prob(visit) =
 $\text{Prob}[\varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})]$
- Note the relationship between the unobserved ε and the outcome



$$\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex}$$



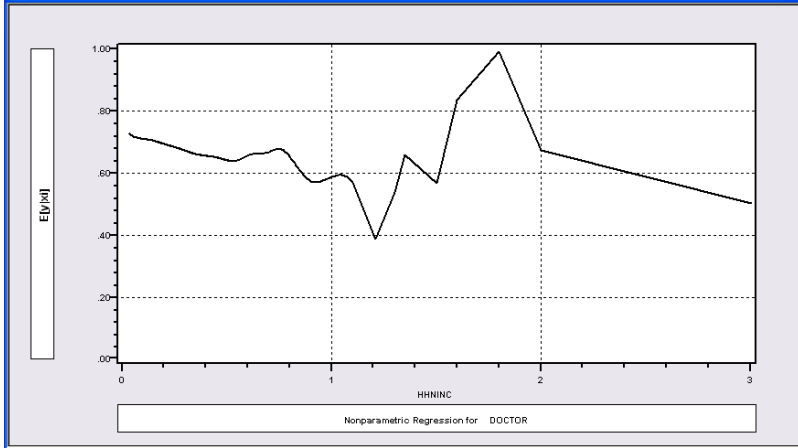


Modeling Approaches

- Nonparametric – “relationship”
 - Minimal Assumptions
 - Minimal Conclusions
- Semiparametric – “index function”
 - Stronger assumptions
 - Robust to model misspecification (heteroscedasticity)
 - Still weak conclusions
- Parametric – “Probability function and index”
 - Strongest assumptions – complete specification
 - Strongest conclusions
 - Possibly less robust. (Not necessarily)

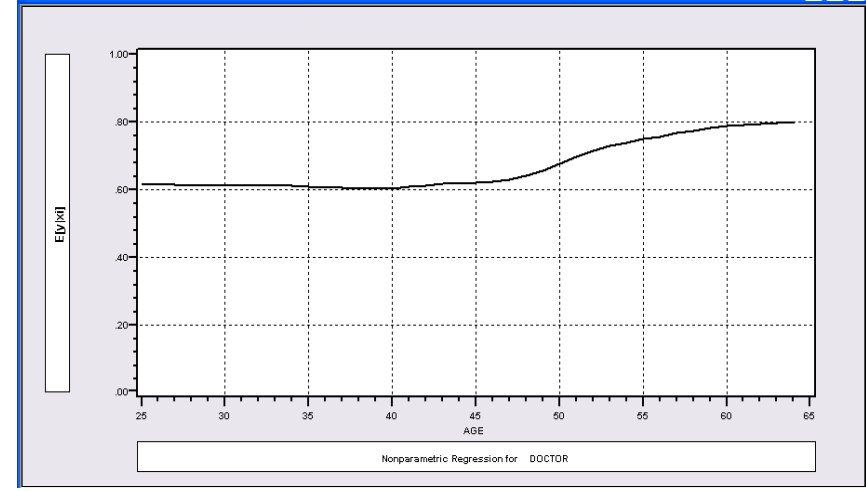


Nonparametric Regressions



$$P(\text{Visit})=f(\text{Income})$$

$$P(\text{Visit})=f(\text{Age})$$





Parametric Model Estimation

- How to estimate $\alpha, \beta_1, \beta_2, \beta_3$?

- It's not regression
- The technique of maximum likelihood

$$L = \prod_{y=0} \text{Prob}[y = 0] \prod_{y=1} \text{Prob}[y = 1]$$

- $\text{Prob}[y=1] =$
 $\text{Prob}[\varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})]$
 $\text{Prob}[y=0] = 1 - \text{Prob}[y=1]$

- Requires a model for the probability

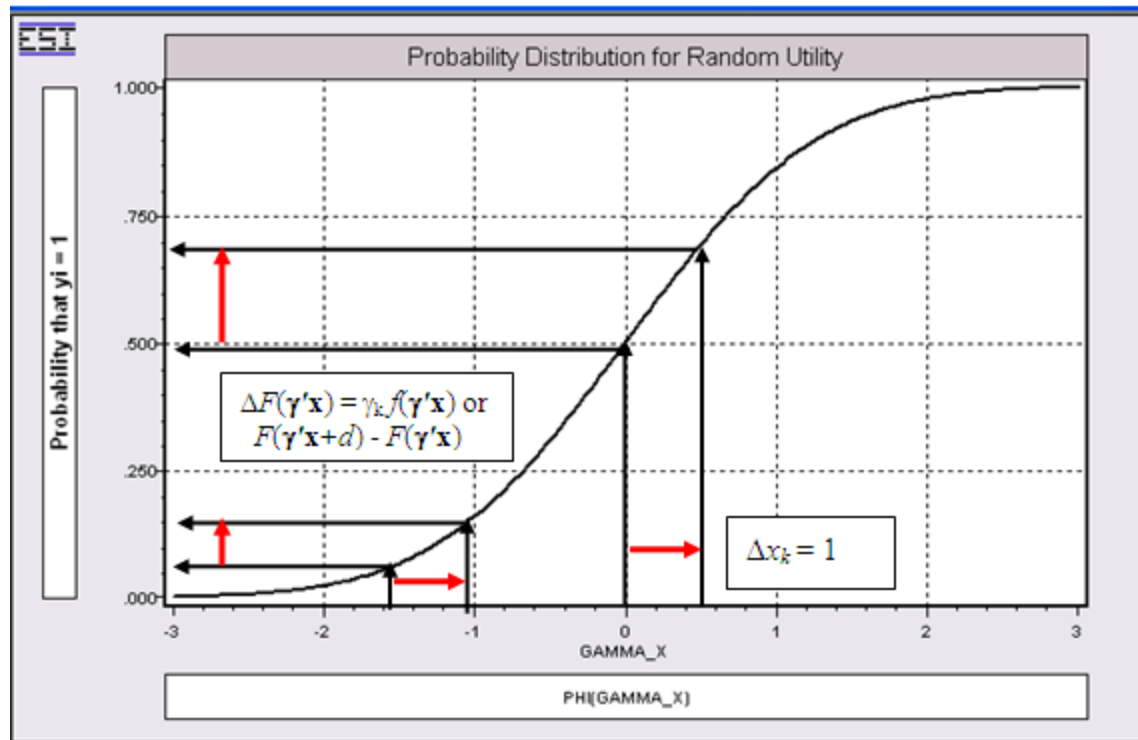


Estimated Binary Choice Models

	LOGIT		PROBIT		EXTREME VALUE	
Variable	Estimate	t-ratio	Estimate	t-ratio	Estimate	t-ratio
Constant	-0.42085	-2.662	-0.25179	-2.600	0.00960	0.078
Age	0.02365	7.205	0.01445	7.257	0.01878	7.129
Income	-0.44198	-2.610	-0.27128	-2.635	-0.32343	-2.536
Sex	0.63825	8.453	0.38685	8.472	0.52280	8.407
Log-L	-2097.48		-2097.35		-2098.17	
Log-L(0)	-2169.27		-2169.27		-2169.27	



Effect on Predicted Probability of an Increase in Age



$$\alpha + \beta_1 (\text{Age}+1) + \beta_2 (\text{Income}) + \beta_3 \text{Sex}$$

(β_1 is positive)



Marginal Effects in Probability Models

- $\text{Prob}[\text{Outcome}] = \text{some } F(\alpha + \beta_1 \text{Income} \dots)$
- "Partial effect" = $\partial F(\alpha + \beta_1 \text{Income} \dots) / \partial "x"$ (derivative)
 - Partial effects are derivatives
 - Result varies with model
 - Logit: $\partial F(\alpha + \beta_1 \text{Income} \dots) / \partial \mathbf{x} = \text{Prob} * (1 - \text{Prob}) * \beta$
 - Probit: $\partial F(\alpha + \beta_1 \text{Income} \dots) / \partial \mathbf{x} = \text{Normal density} * \beta$
 - Extreme Value: $\partial F(\alpha + \beta_1 \text{Income} \dots) / \partial \mathbf{x} = \text{Prob} * (-\log \text{Prob}) * \beta$
- Scaling usually erases model differences



Estimated Partial Effects

	LOGIT		PROBIT		EXTREME VALUE	
	Estimate	t ratio	Estimate	t ratio	Estimate	t ratio
Age	.00527	7.235	.00527	7.269	.00506	6.291
Income	-.09844	-2.611	-.09897	-2.636	-.09711	-2.527
Female	.14026	8.663	.13958	8.264	.13539	8.747



Marginal Effect for a Dummy Variable

- $\text{Prob}[y_i = 1 | \mathbf{x}_i, d_i] = F(\beta' \mathbf{x}_i + \gamma d_i)$
= conditional mean

- Marginal effect of d

$$\text{Prob}[y_i = 1 | \mathbf{x}_i, d_i=1] -$$

$$\text{Prob}[y_i = 1 | \mathbf{x}_i, d_i=0]$$

- Probit: $\delta(d_i) = \Phi(\hat{\beta}' \bar{\mathbf{x}} + \hat{\gamma}) - \Phi(\hat{\beta}' \bar{\mathbf{x}})$



Average Partial Effects

$$\text{Probability} = P_i = F(\beta' \mathbf{x}_i)$$

$$\text{Partial Effect} = \frac{\partial P_i}{\partial \mathbf{x}_i} = \frac{\partial F(\beta' \mathbf{x}_i)}{\partial \mathbf{x}_i} = f(\beta' \mathbf{x}_i) \times \beta = d_i$$

$$\text{Average Partial Effect} = \frac{1}{n} \sum_{i=1}^n d_i$$

are estimates of $\delta = E[d_i]$

under certain assumptions.



Nonlinear Effect

$$P = F(\text{age}, \text{age}^2, \text{income}, \text{female})$$

 Binomial Probit Model

Dependent variable DOCTOR
 Log likelihood function -2086.94545
 Restricted log likelihood -2169.26982
 Chi squared [4 d.f.] 164.64874
 Significance level .00000

-----+-----

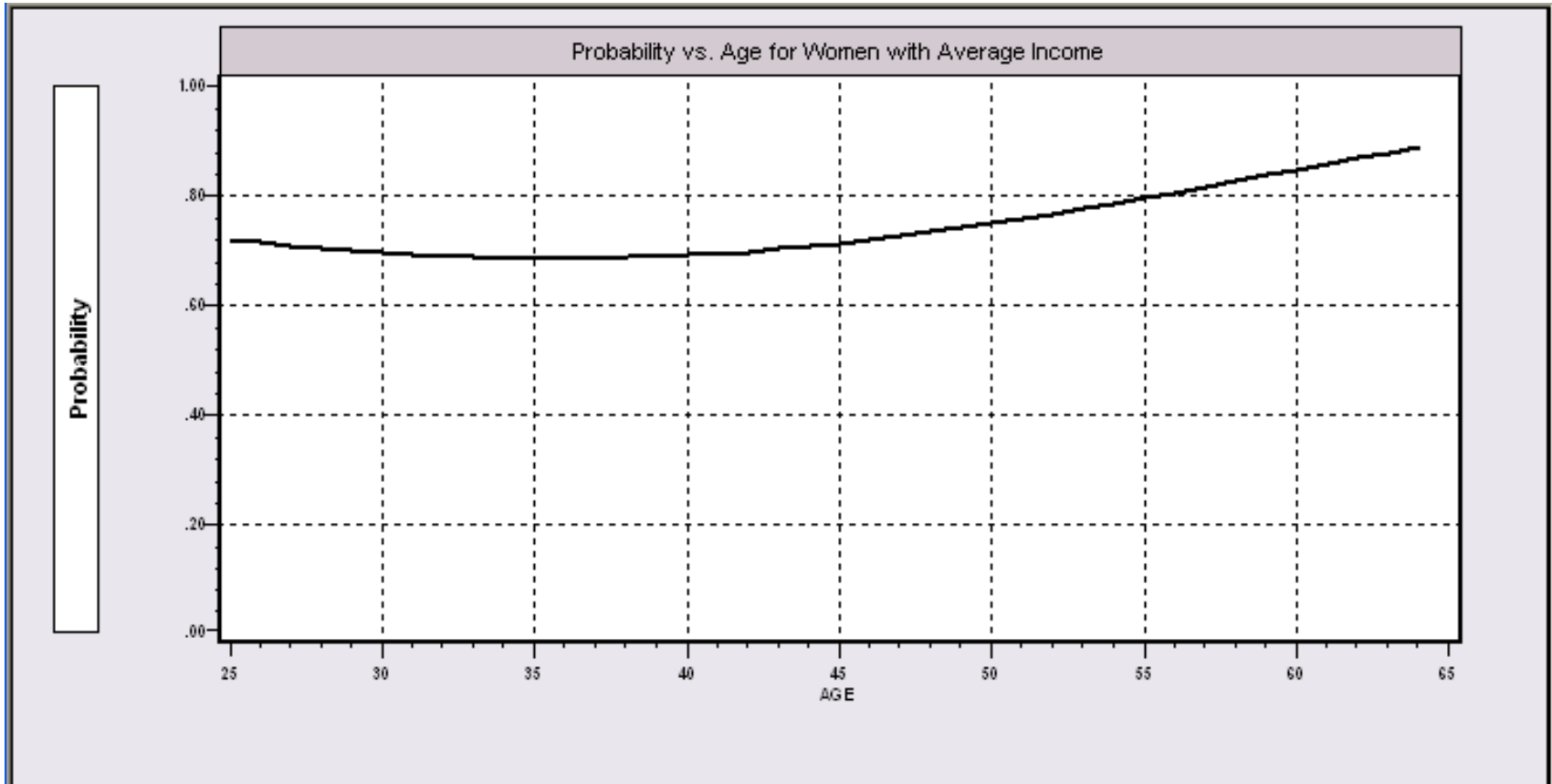
Variable	Coefficient	Standard Error	b/St.Er.	P[Z >z]	Mean of X
-----+-----					
Index function for probability					
Constant	1.30811***	.35673	3.667	.0002	
AGE	-.06487***	.01757	-3.693	.0002	42.6266
AGESQ	.00091***	.00020	4.540	.0000	1951.22
INCOME	-.17362*	.10537	-1.648	.0994	.44476
FEMALE	.39666***	.04583	8.655	.0000	.46343

-----+-----

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



Nonlinear Effects





Partial Effect for Nonlinear Terms

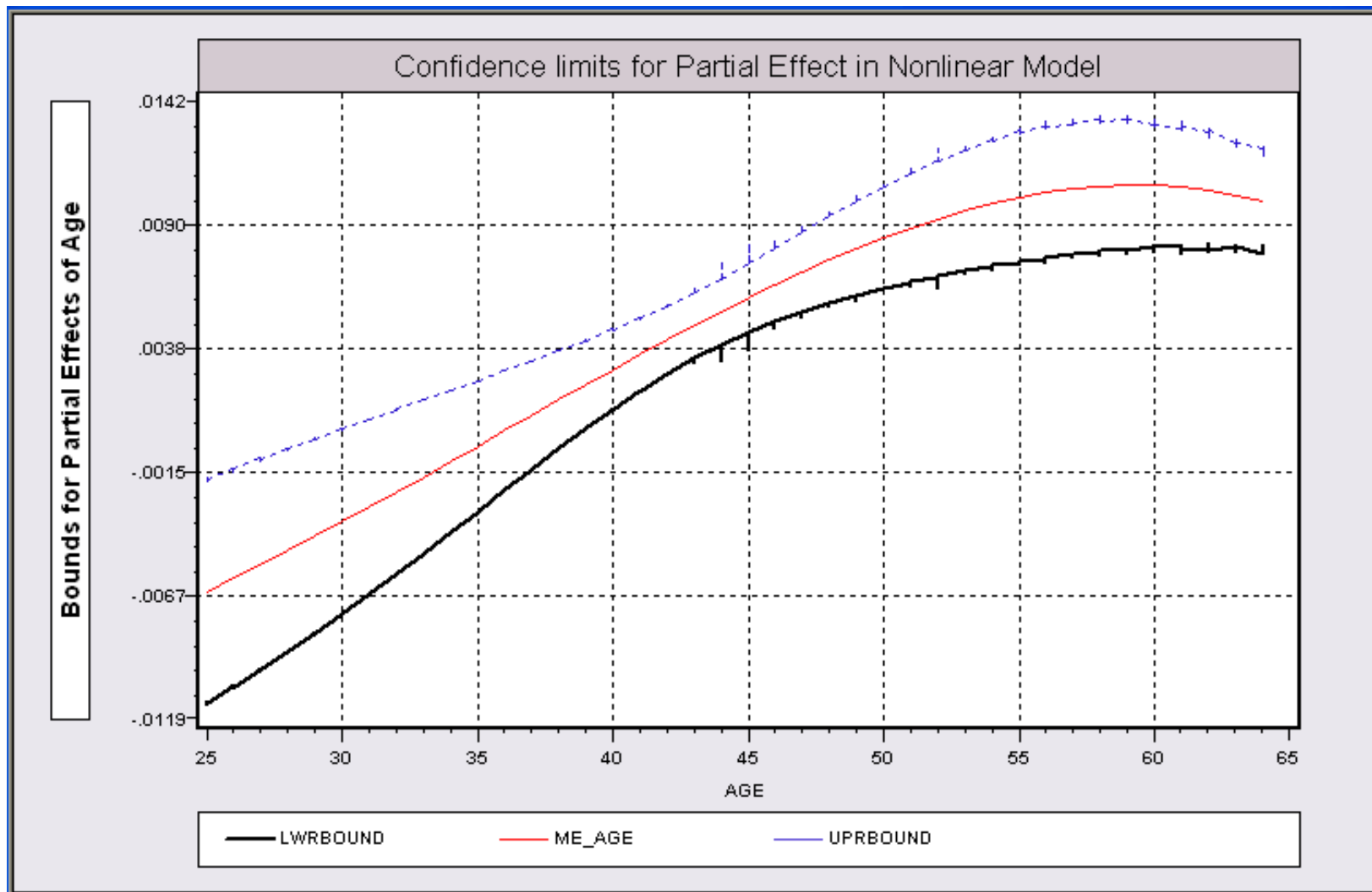
$$\text{Prob} = \Phi[\alpha + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \beta_3 \text{Income} + \beta_4 \text{Female}]$$

$$\frac{\partial \text{Prob}}{\partial \text{Age}} = \phi[\alpha + \beta_1 \text{Age} + \beta_2 \text{Age}^2 + \beta_3 \text{Income} + \beta_4 \text{Female}] \times (\beta_1 + 2\beta_2 \text{Age})$$

- (1) Must be computed for a specific value of Age
- (2) Compute standard errors using delta method or Krinsky and Robb.
- (3) Compute confidence intervals for different values of Age.
- (4) Test of hypothesis that this equals zero is identical to a test that $(\beta_1 + 2\beta_2 \text{Age}) = 0$. Is this an interesting hypothesis?



Confidence Limits for Partial Effects





Model for Visit Doctor

Probit Estimates for *Doctor*. (Absolute asymptotic *t* ratios in parentheses)

Variable	Mean	Std.Dev. (Range)	Model 0	Model 1	Model 2	Model 3	Model 4
Constant			-.1243 (2.138)	-.1423 (2.44)	-.2510 (4.03)	-.3058 (3.56)	-.4664 (5.18)
Female*	0.4788	.4996	.3559 (22.22)	.4552 (13.94)	.7082 (11.16)	.3453 (22.11)	.7647 (11.58)
Age	43.53	11.33 (25 - 64)	.01189 (14.95)	.01137 (14.05)	.01559 (15.20)	.01589 (9.89)	.01963 (10.96)
Income	.3521	.1769 (0.0 - 3.1)	-.1324 (2.85)	-.1197 (2.56)	-.1371 (2.94)	.4060 (2.09)	.4885 (2.51)
Married*	.7586	.4279	.07352 (3.56)	.1387 (4.99)	.06241 (3.01)	.07877 (3.80)	.1168 (4.11)
Young Kids*	.4027	.4905	-.1521 (8.30)	-.1613 (8.71)	-.1588 (8.64)	-.1525 (8.32)	-.1658 (8.94)
Education	11.32	2.325 (7 - 18)	-.01497 (4.19)	-.01587 (4.43)	-.01641 (4.578)	-.01452 (4.06)	-.0165 (4.59)
Female* Married				-.131 (3.49)			-.09607 (2.52)
Female* Age					-.00820 (5.74)		-.00787 (5.40)
Income* Age						-.01241 (2.86)	-.01418 (3.27)

* Binary Variable



Simple Partial Effects

Table 2 Estimated Partial Effects for Age, Female and Income Based on Model 2

Variable	Coefficient	t ratio	Average Partial Effect*	Minimum Effect**	Maximum Effect**	Minimum t ratio**	Maximum t ratio**
Age	0.01559	15.20	0.00433	0.00205	0.00622	6.07	18.87
Female	0.7072	11.16	0.313	0.225	0.387	17.71	65.42
Income	-0.1371	-2.94	-0.0499	-0.0547	-0.0381	-3.23	-2.56

* Averaged across all observations

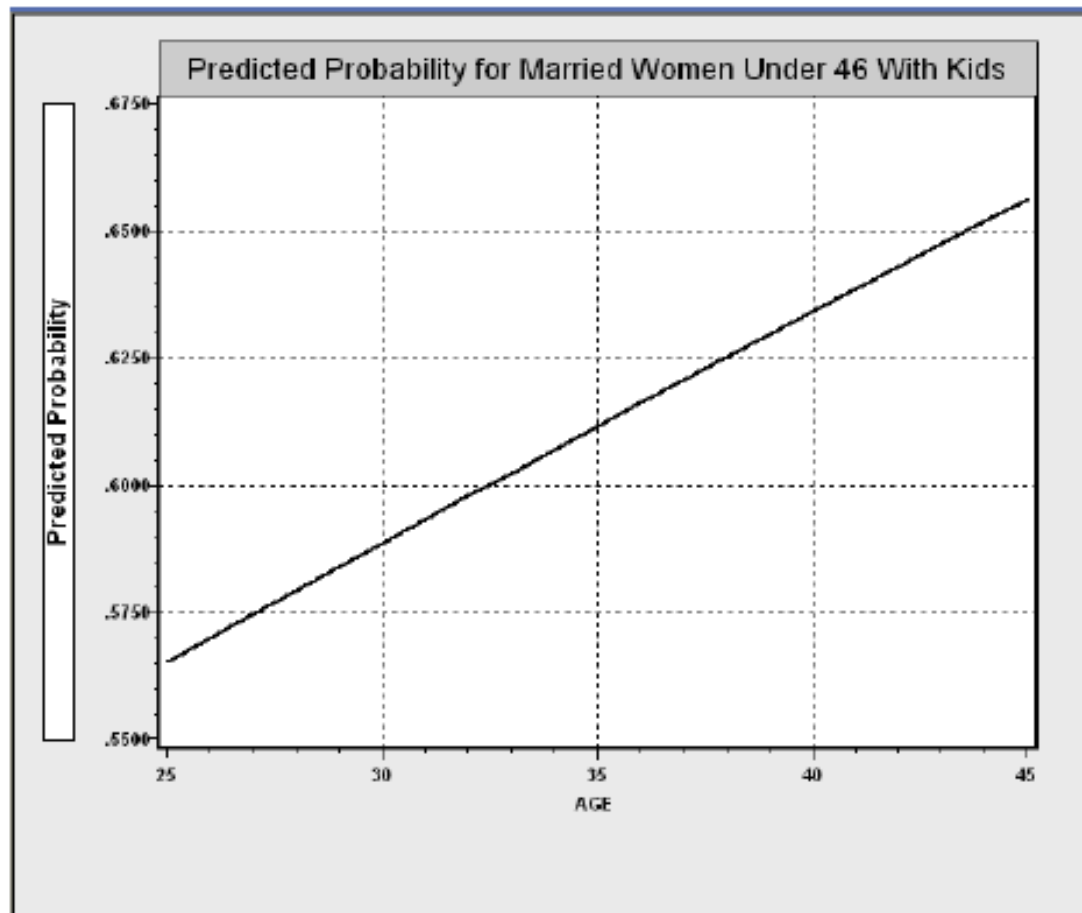
** Based on values computed for each observation

Table 3 Estimated Interaction Effects Between Age and Income in Model 3

	Mean	Standard Deviation	Minimum	Maximum
Probability	0.6291	0.1002	0.4015	0.8366
Interaction Effect	-0.004244	0.0007766	-0.005085	-0.001669
t Ratio	-2.73	0.12	-3.18	-2.17



Direct Effect of Age



Relationship between Age and Probability of Doctor Visit

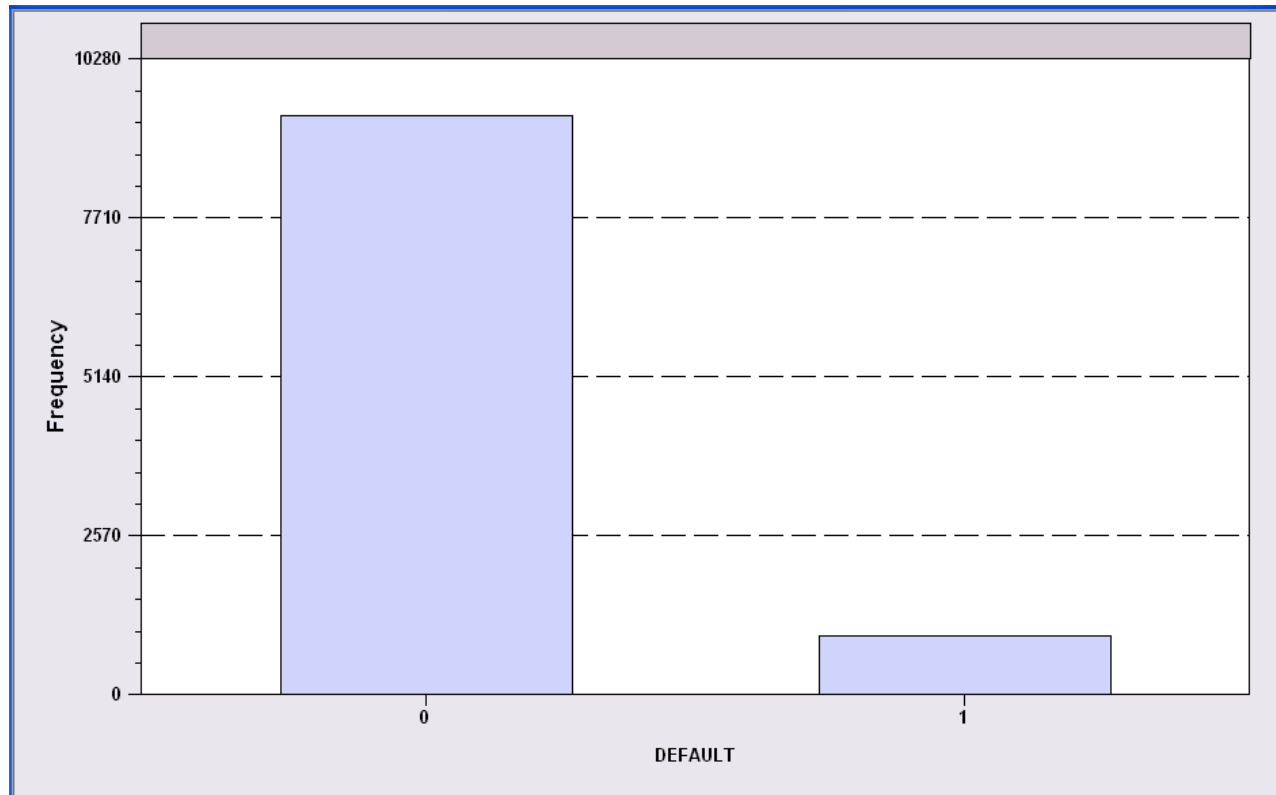


A Problem of Unbalanced Samples

- Either 0 or 1 heavily dominates the sample
- Regression methods work poorly or not at all
- Estimates are imprecise and highly variable
- Meanings of probabilities and model estimates are questionable

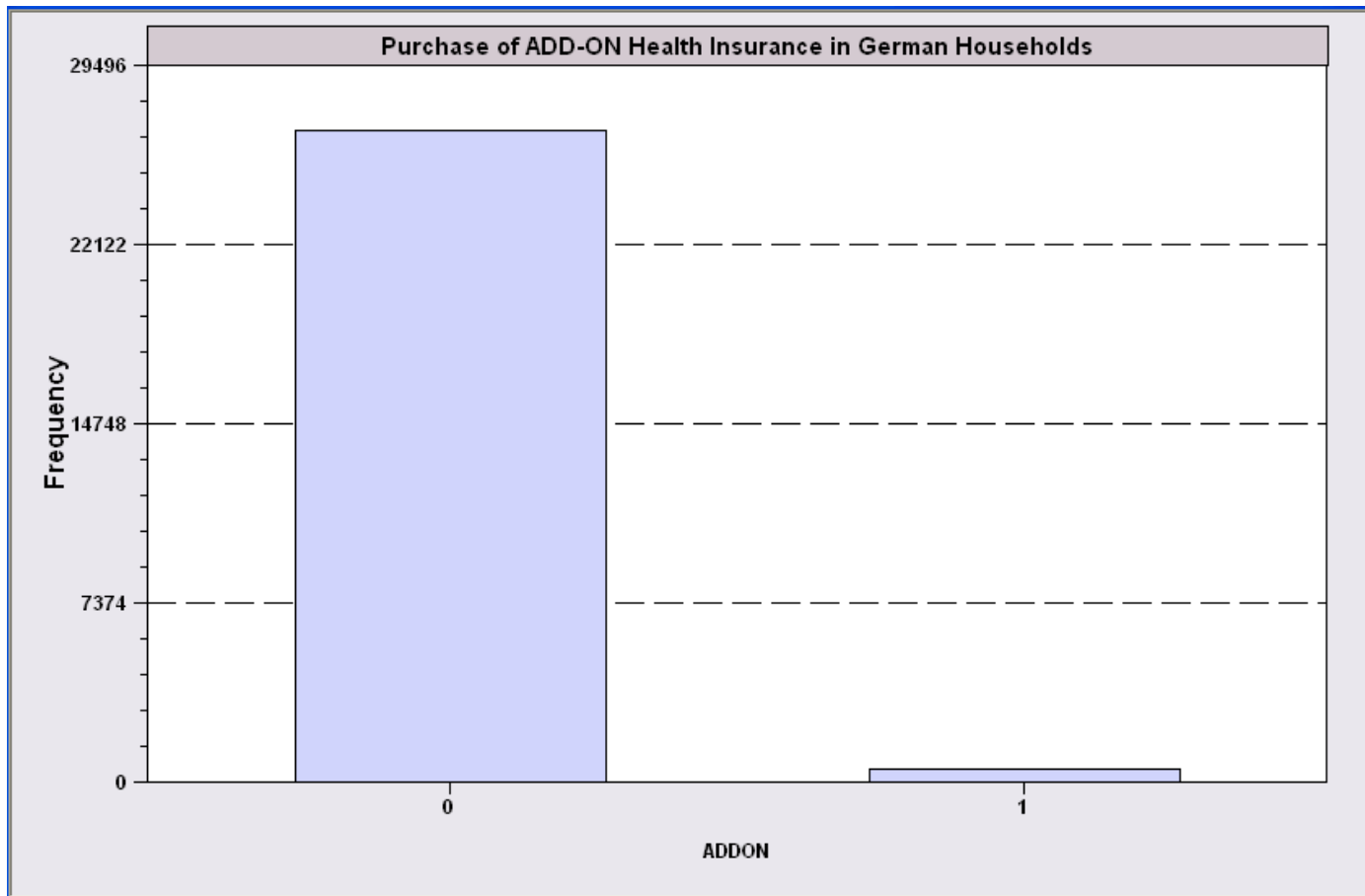


Default by Cardholders





Add On Insurance Purchase





King and Zeng on Rare Events

- King, G. and Zeng, L., "Logistic Regression in Rare Events Data. 2001 (Available online)
- King, G. and Zeng, L., "Explaining Rare Events in International Relations," *International Organization*, 55, 3, Summer 2001.

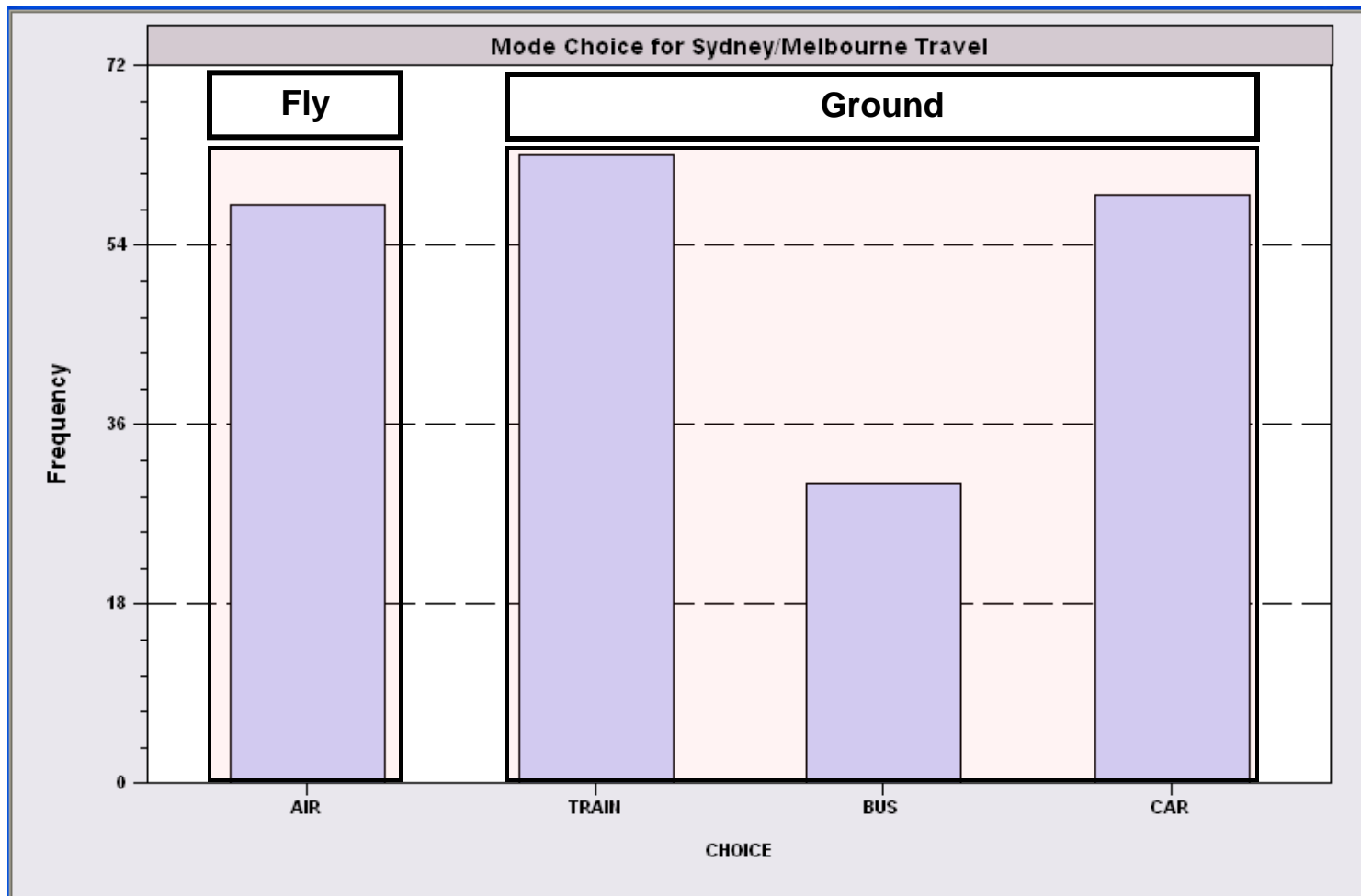


Proposed Approaches

- “Bias correction”
- Choice based sampling
 - Sample is “sweetened” to increase proportion of events that “occur”
 - Estimates and standard errors are corrected for the nonrandom sampling.



A Travel Application: Sydney/Melbourne





Choice Based Sample for a Travel Application

	Sample	Population	Weight
Fly	27.62%	14%	0.5068
Ground	72.38%	86%	1.1882



Choice Based Sampling Correction

- Maximize Weighted Log Likelihood
- Covariance Matrix Adjustment

$$V = \mathbf{H}^{-1} \mathbf{G} \mathbf{H}^{-1} \text{ (all three weighted)}$$

\mathbf{H} = Hessian

\mathbf{G} = Outer products of gradients



Effect of Choice Based Sampling

GC = a general measure of cost

TTME = terminal time

HINC = household income

Unweighted

Variable	Coefficient	Standard Error	b/St.Er.	P[Z >z]
Constant	1.784582594	1.2693459	1.406	.1598
GC	.02146879786	.006808094	3.153	.0016
TTME	-.09846704221	.016518003	-5.961	.0000
HINC	.02232338915	.010297671	2.168	.0302

| Weighting variable CBWT |
| Corrected for Choice Based Sampling |

Variable	Coefficient	Standard Error	b/St.Er.	P[Z >z]
Constant	1.014022236	1.1786164	.860	.3896
GC	.02177810754	.006374383	3.417	.0006
TTME	-.07434280587	.017721665	-4.195	.0000
HINC	.02471679844	.009548339	2.589	.0096

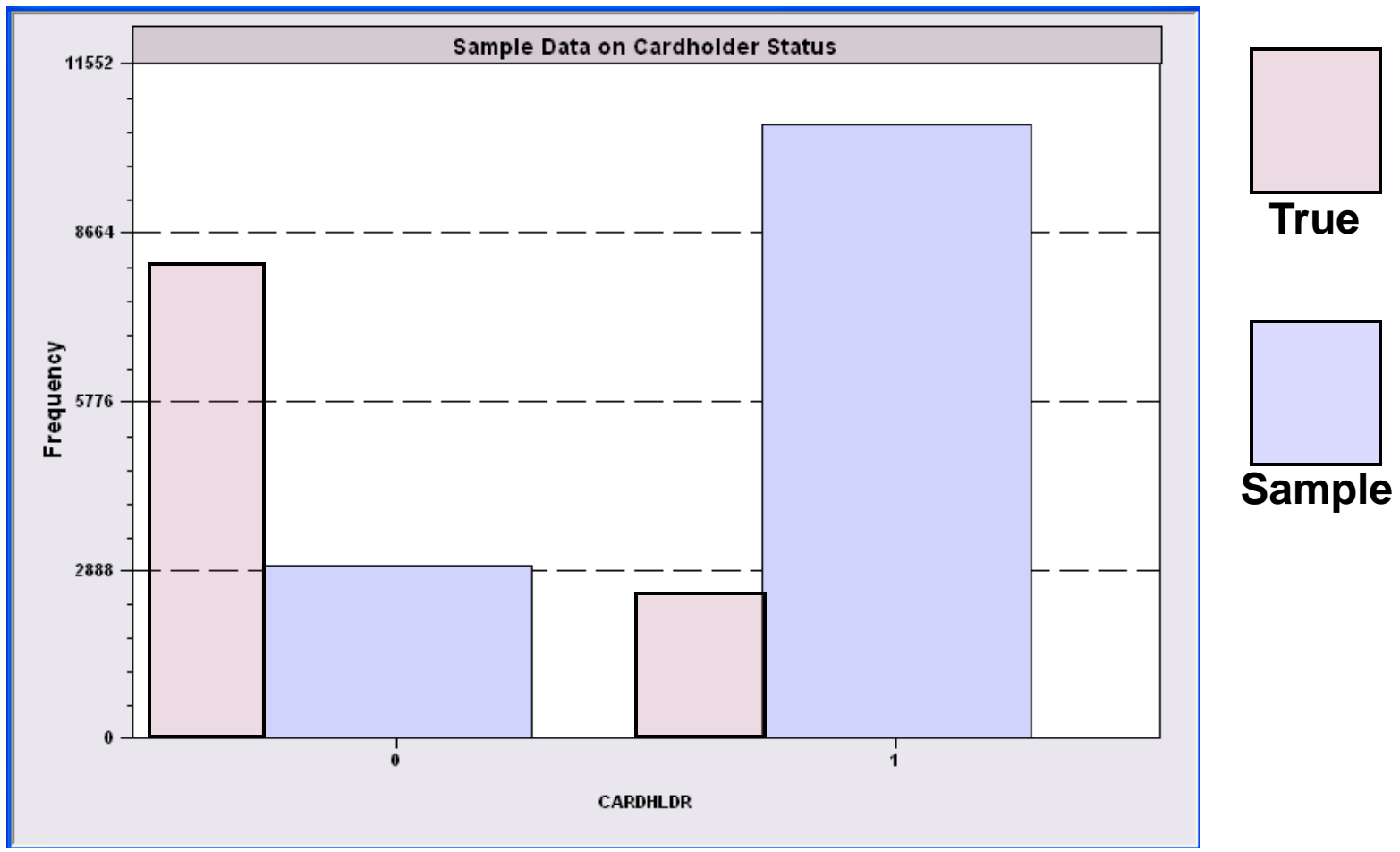


Modeling Default

- American Express Cardholders
 - Applications: 13,444
 - Acceptances: 10,499
 - Not representative of the population
- Default | Application Accepted
 - Acceptances: 10,499
 - Default: 996
 - Not representative of the population



Artificially Proportioned Sample





Application

Descriptive Statistics for Variables

Variable	Mean	Std. Dev.	Minimum	Maximum	Cases
CARDHLDR	.78094	.41362	0.0	1.000	13444
DEFAULT	.094866	.29304	0.0	1.000	10499

Sampling Weights for Choice Based Sampling

Event	w=sample	W=Population	$\Omega=W/w$
D=1, C=1	996/13444	.232 × .103	.32255
D=0, C=1	9503/13444	.232 × .897	.29441
C=0	2945/13444	.768	3.50594



Application to Default

Weighted and Unweighted Probit Cardholder Equations

	Choice based sampling		Unweighted	
Variable	Coefficient	t-ratio	Coefficient	t-ratio
ONE	-1.1175	-9.090	0.1070	1.390
AGE	-0.0021	-0.806	-0.0012	-0.672
MTHCURAD	0.0010	2.547	0.0011	3.943
DEPNENTS	-0.0947	-2.623	-0.0957	-4.079
MTHMPLOY	-0.0002	-0.410	-0.0002	-0.694
MAJORDRG	-0.7514	-13.922	-0.7796	-34.777
MINORDRG	-0.0609	-1.554	-0.0471	-2.005
OWNRENT	0.0514	0.947	-0.0042	-0.119
MTHPRVAD	0.0002	0.626	0.0001	0.767
PREVIOUS	0.1781	1.843	0.2089	2.967
INCOME	0.1153	4.353	0.1362	7.001
SELFEMPL	-0.3652	-3.711	-0.3634	-5.804

	Predicted				Predicted		
Actual	0	1	TOTAL	Actual	0	1	TOTAL
0	.208	.011	2945	0	.110	.109	2945
1	.420	.361	10499	1	.020	.761	10499
TOTAL	8448	4996	13444	TOTAL	1748	11696	13444



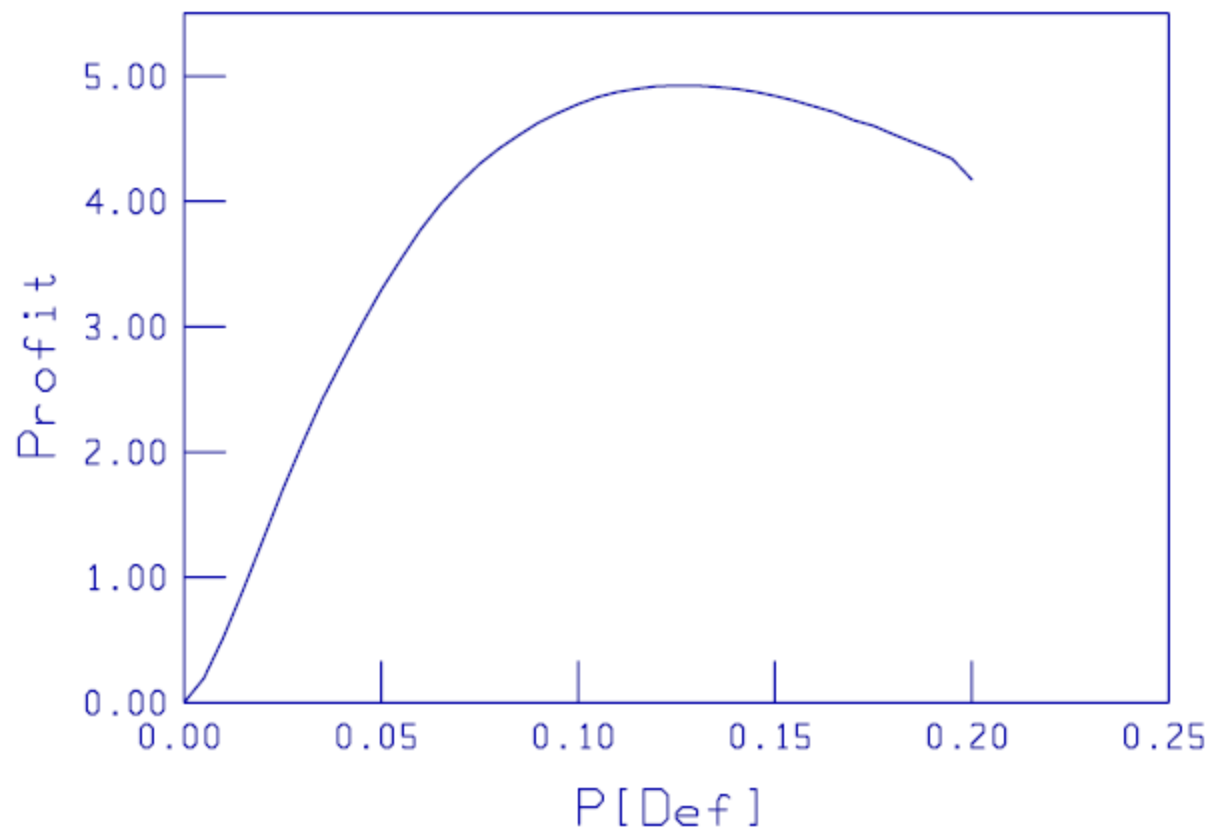
An Integrated Model With Default and Spending

Default Model							
Variable	Unconditional			Conditional			Partial
	Coeff.	Std.Err.	t-ratio	Coeff.	Std.Err.	t-ratio	
<u>Basic Default Specification</u>							
Constant	-1.1350	0.0984	-11.533	-1.3752	0.3945	-3.486	
AGE	-0.0031	0.0023	-1.342	-0.0054	0.0094	-0.582	-.0018
MTHCURAD	0.0003	0.0003	1.069	0.0002	0.0013	0.153	-.0001
DEPDNTS	0.0445	0.0294	1.512	-0.0217	0.1114	-0.195	.0073
MTHMPLOY	0.0007	0.0003	2.331	0.0007	0.0013	0.566	.0002
MAJORDRG	0.0592	0.0408	1.448	-0.2969	0.1985	-1.495	.0033
MINORDRG	0.0764	0.0296	2.586	0.1780	0.0993	1.793	.0488
OWNRENT	-0.0010	0.4312	-0.023	0.0908	0.1706	0.533	.0236
MTHPRVAD	0.0004	0.0002	1.817	0.0002	0.0009	0.274	.00002
PREVIOUS	-0.1507	0.0792	-1.902	-0.1112	0.3103	-0.358	-.0434
INCOME	-0.0168	0.0033	-5.608	-0.0072	0.0151	-0.476	.0062
SELFEMPL	0.0788	0.0850	0.927	-0.1969	0.3565	-0.552	-.0017
TRADACCT	0.0004	0.0044	0.109	0.0207	0.0205	1.009	-.0028
INCPER	-0.0228	0.0323	-0.706	-0.0545	0.1058	-0.515	-.0094
EXP_INC	-0.4761	0.1717	-2.774	-0.5790	0.5033	-1.150	-.1614
<u>Credit Bureau</u>							
CREDPEN	0.0138	0.0063	2.195	0.0199	0.0272	0.732	.0066
CREDACTV	-0.1218	0.0126	-9.657	-0.1500	0.0857	-2.695	-.0424
CRDDEL30	0.2841	0.0712	3.991	0.2829	0.2766	1.023	.1120
CR30DLNQ	0.0806	0.0177	4.559	0.0446	0.0757	0.589	.0225
AVGRVBAL	0.0011	0.0024	0.439	0.0156	0.0123	1.268	.0038
AVBALINC	0.0039	0.00042	9.192	0.0008	0.0021	0.398	.0004
<u>Expenditure</u>							
FITEXP	0.0014	0.0044	3.103	0.00064	0.0019	0.336	



Influence of the Crucial Variable

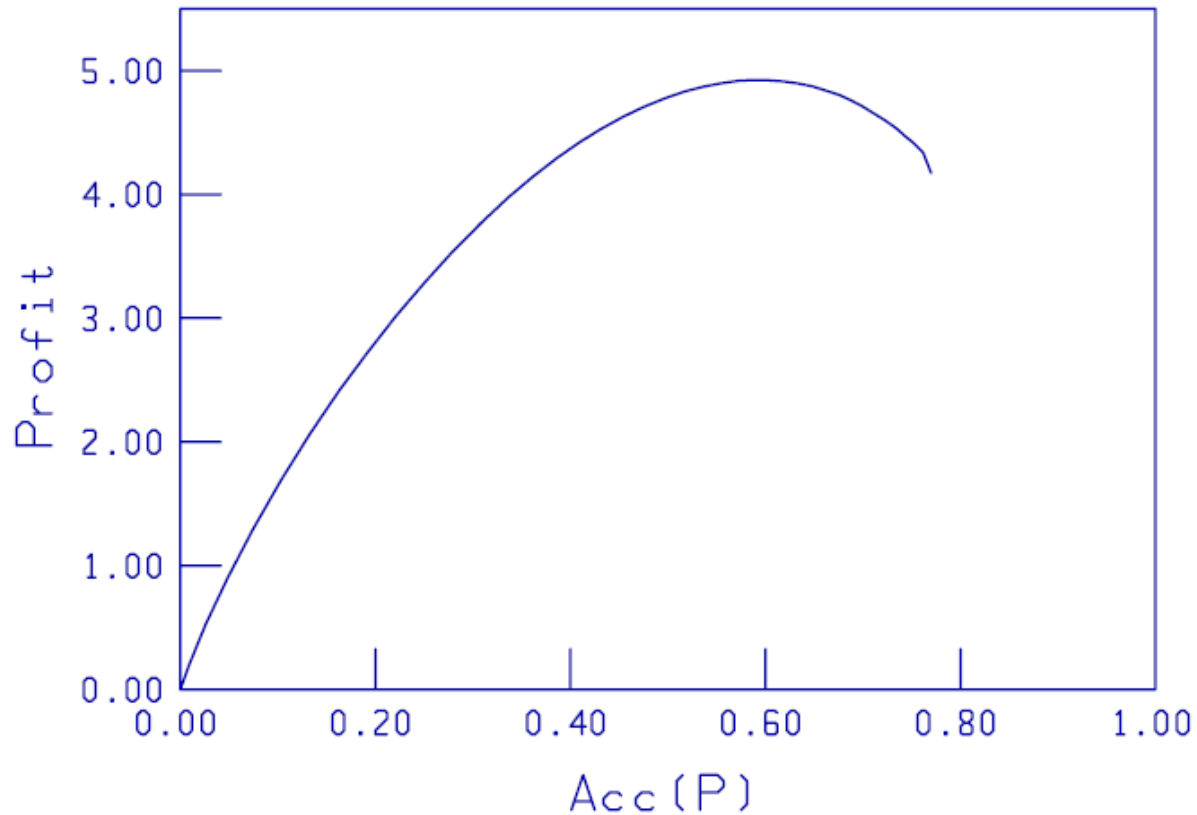
Figure 3. Profits vs. Default Probability





Implication for a Policy Rule

Figure 4. Profits vs. Acceptance Rate





Binary Choice Model Problems with Unusual Events

- Sparse ones
- Constant term correction
- WESML
- Implications for estimation and inference



What Did We Learn?

- Frailty of the model
- Role of crucial parameters
- Consequence of biased estimation
- A possible model/sample based improvement of the calculation



Part 6: Models for Counts



Application: Major Derogatory Reports

AmEx Credit Card
Holders

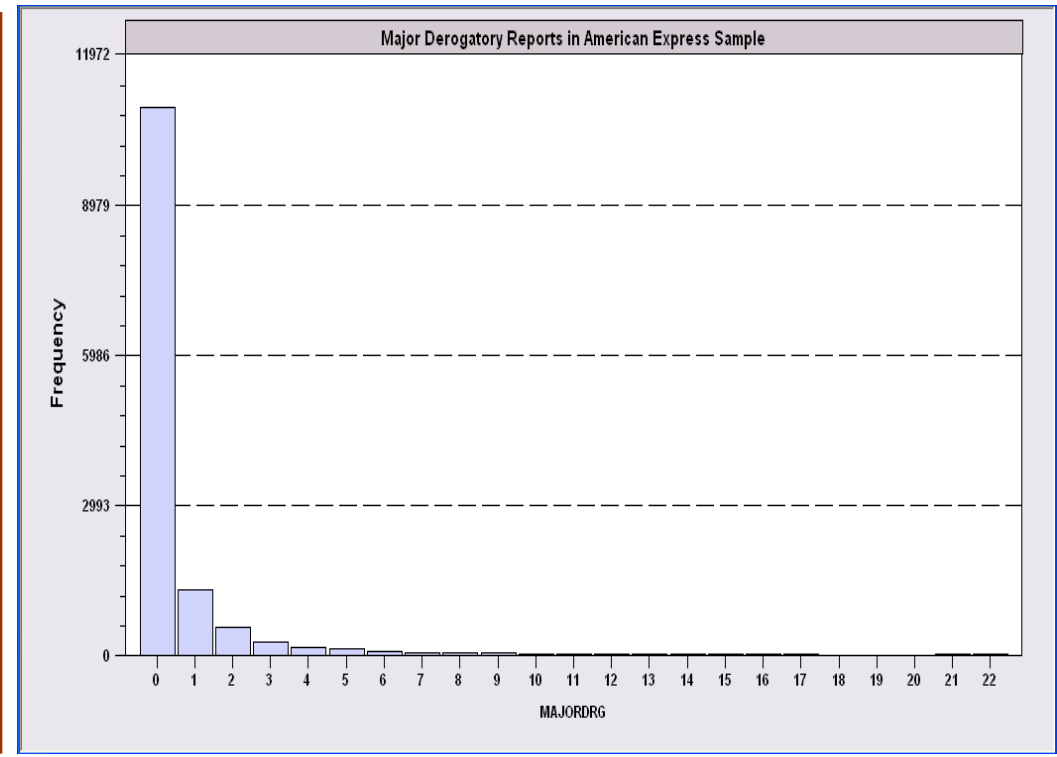
$N = 13,777$

Number of major
derogatory reports in 1
year

Issues:

Nonrandom selection

Excess zeros





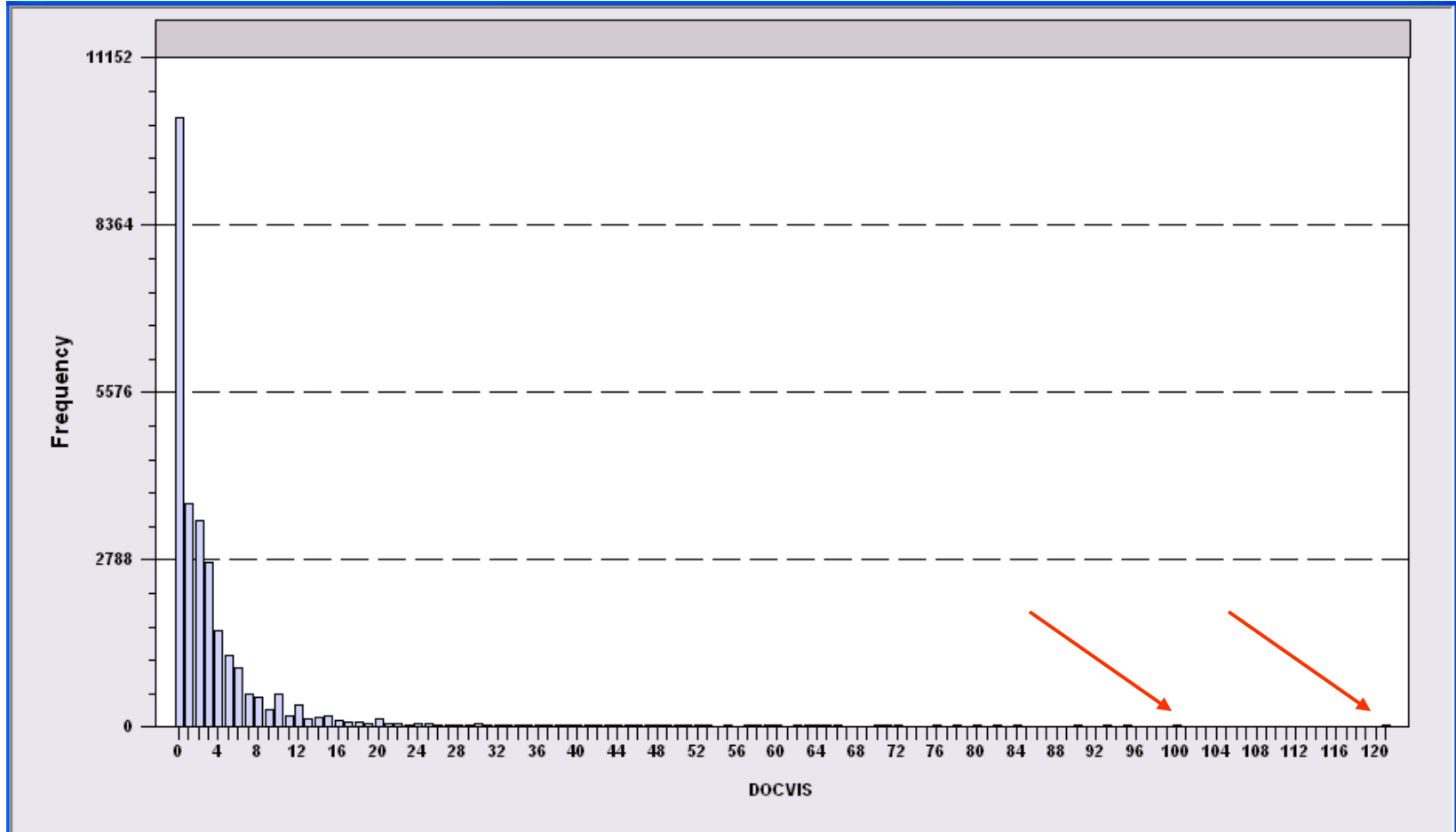
Histogram for Credit Data

Histogram for MAJORDRG NOBS= 13444, Too low: 0, Too high: 0

Bin	Lower limit	Upper limit	Frequency	Cumulative Frequency
0	0	10883 (.8095)	10883 (.8095)	
1	1	1306 (.0971)	12189 (.9066)	
2	2	534 (.0397)	12723 (.9464)	
3	3	244 (.0181)	12967 (.9645)	
4	4	140 (.0104)	13107 (.9749)	
5	5	110 (.0082)	13217 (.9831)	
6	6	58 (.0043)	13275 (.9874)	
7	7	38 (.0028)	13313 (.9903)	
8	8	32 (.0024)	13345 (.9926)	
9	9	28 (.0021)	13373 (.9947)	
10	10	17 (.0013)	13390 (.9960)	
11	11	22 (.0016)	13412 (.9976)	
12	12	5 (.0004)	13417 (.9980)	
13	13	10 (.0007)	13427 (.9987)	
14	14	5 (.0004)	13432 (.9991)	
15	15	3 (.0002)	13435 (.9993)	
16	16	3 (.0002)	13438 (.9996)	
17	17	2 (.0001)	13440 (.9997)	
18	18	0 (.0000)	13440 (.9997)	
19	19	0 (.0000)	13440 (.9997)	
20	20	0 (.0000)	13440 (.9997)	
21	21	3 (.0002)	13443 (.9999)	
22	22	1 (.0001)	13444 (1.0000)	



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”

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

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



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

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

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



Negative Binomial Specification

The Poisson estimator is consistent when there is unmeasured heterogeneity in the conditional mean. Therefore, this is a case for the ROBUST covariance matrix estimator. (Neglected heterogeneity that is uncorrelated with \mathbf{x}_i .)



Negative Binomial Specification

- $\text{Prob}(Y_i=j|\mathbf{x}_i)$ has greater mass to the right and left of the mean
- Conditional mean function is the same as the Poisson: $E[y_i|\mathbf{x}_i] = \lambda_i = \text{Exp}(\beta'\mathbf{x}_i)$, so marginal effects have the same form.
- Variance is $\text{Var}[y_i|\mathbf{x}_i] = \lambda_i(1 + a \lambda_i)$, a is the overdispersion parameter; $a = 0$ 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	



Model Formulations

Poisson

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

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

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

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

Negative Binomial – 1

$$\text{Var}[y_i | \mathbf{x}_i] = \lambda_i + \kappa \lambda_i^2 = \lambda_i [1 + \kappa],$$

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

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

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

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

Negative Binomial – 2

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

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

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

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

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

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

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

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

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

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

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

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



NegBin-1 Model

```

-----
Negative Binomial Regression
Dependent variable          DOCVIS
Log likelihood function     -60025.78734
Restricted log likelihood  -103727.29625
NegBin form 1; Psi(i) = theta*exp[bx(i)]
-----+-----
Variable| Coefficient      Standard Error  b/St.Er.  P[|Z|>z]  Mean of X
-----+-----
Constant|      .62584***      .05816        10.761    .0000
      AGE|      .01428***      .00073        19.462    .0000      43.5257
      EDUC|     -.01549***      .00359         -4.314    .0000      11.3206
      FEMALE|    .33028***      .01479        22.328    .0000       .47877
      MARRIED|   .04324**       .01852         2.335    .0196       .75862
      HHNINC|   -.24543***      .04540         -5.406    .0000       .35208
      HHKIDS|   -.14877***      .01745        -8.526    .0000       .40273
      |Dispersion parameter for count data model
      Alpha|    6.09246***      .06694         91.018    .0000
-----+-----

```



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***



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 3.1835 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 $\Pr(0) = \exp[-\beta'x_i]$. (Number of visits, including zero visits this season.)
- Unconditional:
 - $\Pr[0] = P(\text{regime 0}) + P(\text{regime 1}) * \Pr[0 | \text{regime 1}]$
 - $\Pr[j | j > 0] = P(\text{regime 1}) * \Pr[j | \text{regime 1}]$
- “Two inflation” – Number of children
- These are “latent class models”



Application: Major Derogatory Reports

AmEx Credit Card
Holders

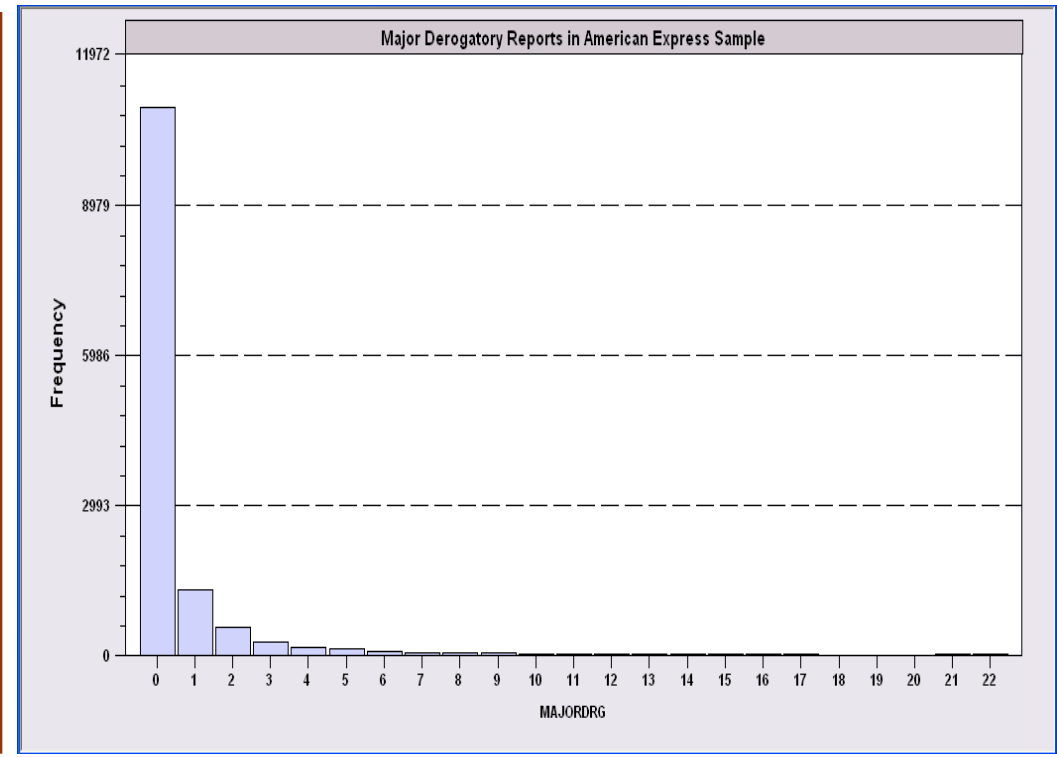
$N = 13,777$

Number of major
derogatory reports in 1
year

Issues:

Nonrandom selection

Excess zeros





Zero Inflation Models

ZIP - tau = ZIP(τ)

$$\text{Prob}(y_i = j | x_i) = \frac{\exp(-\lambda_i) \lambda_i^j}{j!}, \lambda_i = \exp(\beta' x_i)$$

$$\text{Prob}(0 \text{ regime}) = F(\tau \beta' x_i)$$

Zero Inflation = ZIP

$$\text{Prob}(y_i = j | x_i) = \frac{\exp(-\lambda_i) \lambda_i^j}{j!}, \lambda_i = \exp(\beta' x_i)$$

$$\text{Prob}(0 \text{ regime}) = F(\gamma' z_i)$$



Notes on Zero Inflation Models

- Poisson is not nested in ZIP. $\tau = 0$ in ZIP(τ) or $\gamma = 0$ in ZIP does not produce Poisson; it produces ZIP with $P(\text{regime } 0) = 1/2$.
 - Standard tests are not appropriate
 - Use Vuong statistic. ZIP model almost always wins.
- Zero Inflation models extend to NB models – ZINB(τ) and ZINB are standard models
 - Creates two sources of overdispersion
 - Generally difficult to estimate



ZIP(τ) Model

Zero Altered Poisson Regression Model
Logistic distribution used for splitting model.

ZAP term in probability is $F[\tau \times \ln \text{LAMBDA}]$

Comparison of estimated models				
	Pr[0 means]	Number of zeros		Log-likelihood
Poisson	.04933	Act.= 10135	Prd.= 1347.9	-103727.29625
Z.I.Poisson	.35944	Act.= 10135	Prd.= 9822.1	-84012.30960

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 44.5723

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	1.45145***	.01121	129.498	.0000	
AGE	.01140***	.00013	86.245	.0000	43.5257
EDUC	-.02306***	.00075	-30.829	.0000	11.3206
FEMALE	.13129***	.00256	51.357	.0000	.47877
MARRIED	-.02270***	.00317	-7.151	.0000	.75862
HHNINC	-.41799***	.00898	-46.527	.0000	.35208
HHKIDS	-.08750***	.00322	-27.189	.0000	.40273
Zero inflation model					
Tau	-.38910***	.00836	-46.550	.0000	



ZIP Model

 Zero Altered Poisson Regression Model
 Logistic distribution used for splitting model.

ZAP term in probability is $F[\tau \times Z(i)]$

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



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.



Hurdle Model

Two Part Model

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

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

A Poisson Hurdle Model with Logit Hurdle

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

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

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

Marginal effects involve both parts of the model.



Hurdle Model for Doctor Visits

```
-----
Poisson hurdle model for counts
Dependent variable          DOCVIS
Log likelihood function     -84211.96961
Restricted log likelihood   -103727.29625
Chi squared [ 1 d.f.]      39030.65329
Significance level          .00000
McFadden Pseudo R-squared  .1881407
Estimation based on N = 27326, K = 10
LOGIT hurdle equation
-----
```

Variable	Coefficient	Standard Error	b/St.Er.	P[Z >z]	Mean of X
-----+-----					
Parameters of count model equation					
Constant	1.53350***	.01053	145.596	.0000	
AGE	.01088***	.00013	85.292	.0000	43.5257
EDUC	-.02387***	.00072	-32.957	.0000	11.3206
FEMALE	.10244***	.00243	42.128	.0000	.47877
MARRIED	-.03463***	.00294	-11.787	.0000	.75862
HHNINC	-.46142***	.00873	-52.842	.0000	.35208
HHKIDS	-.07842***	.00301	-26.022	.0000	.40273
Parameters of binary hurdle equation					
Constant	.77475***	.06634	11.678	.0000	
FEMALE	.59389***	.02597	22.865	.0000	.47877
EDUC	-.04562***	.00546	-8.357	.0000	11.3206
-----+-----					



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 .0109 Scale Factor for Marginal Effects 3.0118

Variable	Coefficient	Standard Error	b/St.Er.	P[Z >z]	Mean of X
-----+-----					
Effects in Count Model Equation					
Constant	4.61864	2.84230	1.625	.1042	
AGE	.03278	.02018	1.625	.1042	43.5257
EDUC	-.07189	.04429	-1.623	.1045	11.3206
FEMALE	.30854	.19000	1.624	.1044	.47877
MARRIED	-.10431	.06479	-1.610	.1074	.75862
HHNINC	-1.38971	.85557	-1.624	.1043	.35208
HHKIDS	-.23620	.14563	-1.622	.1048	.40273
Effects in Binary Hurdle Equation					
Constant	.86178***	.07379	11.678	.0000	
FEMALE	.66060***	.02889	22.865	.0000	.47877
EDUC	-.05074***	.00607	-8.357	.0000	11.3206
Combined effect is the sum of the two parts					
Constant	5.48042*	2.85728	1.918	.0551	
EDUC	-.12264***	.04479	-2.738	.0062	11.3206
FEMALE	.96915***	.19441	4.985	.0000	.47877

-----+-----



Quantile Regression for Counts

Machado, A. and J. Santos Silva,
“Quantiles for Counts,”

Journal of the American Statistical
Association,
100, 472, 2005, pp. 1226-1237

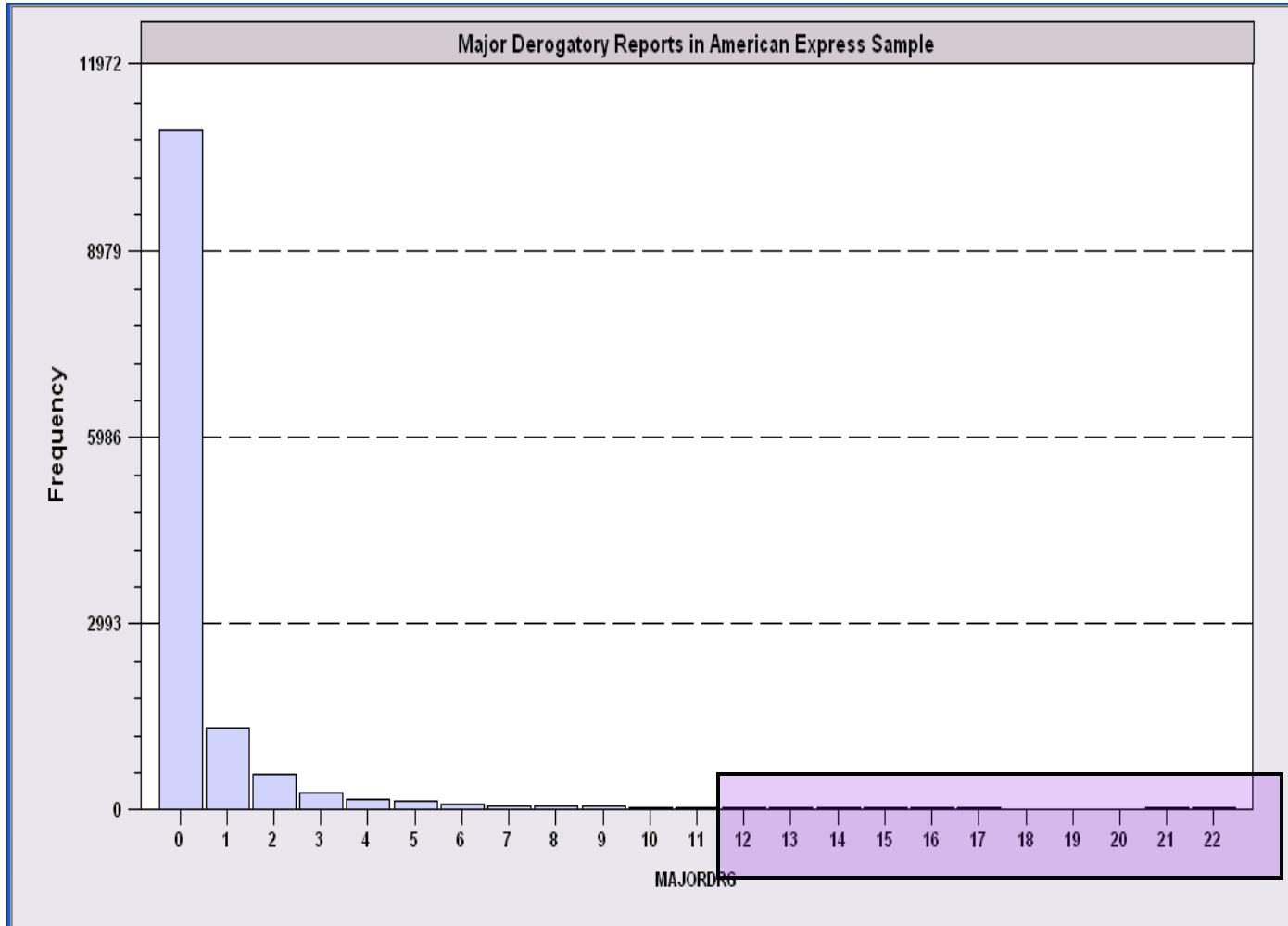


Quantile Regression for Counts

- Comparable to quantile regression for a continuous variable
- Sensitivity to outlying observations is less a problem for count data estimators than for regressions – ML, not least squares
- Quantiles for counts may be more interesting



Unusual Counts





Mean vs. Median Regression

-----+-----
Poisson Regression

LHS=MAJORDRG Mean = .46281
 Standard deviation = 1.43267
 Number of observs. = 13444

-----+-----

MAJORDRG	Coefficient	Standard Error	z	Prob. z> Z	Mean of X
Constant	-1.33233***	.04519	-29.48	.0000	
AGE	.01388***	.00133	10.42	.0000	33.4718
ACADMOS	.00180***	.00019	9.46	.0000	55.3189
OWNRENT	-.09684***	.02859	-3.39	.0007	.45597
HINC	.00569	.00871	.65	.5138	3.01143
SELFEMPL	.04334	.05210	.83	.4054	.05794

-----+-----

Quantile Regression Model. Quantile = .500000
 Minimum = .00000
 t= .50000 quantile = .00000
 Maximum = 22.00000

-----+-----

Constant	-2.78674***	.15140	-18.41	.0000	
AGE	.01249***	.00437	2.86	.0043	33.4743
ACADMOS	.00248***	.00058	4.27	.0000	55.3230
OWNRENT	-.15880	.10532	-1.51	.1316	.45600
HINC	.05112*	.02737	1.87	.0618	3.01166
SELFEMPL	-.15943	.21109	-.76	.4501	.05795

-----+-----



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 .4628 Scale Factor for Marginal Effects .4628

MAJORDRG	Coefficient	Standard Error	z	Prob. z> Z	Mean of X
AGE	.00642***	.00068	9.41	.0000	33.4718
ACADMOS	.00083***	.9631D-04	8.63	.0000	55.3189
OWNRENT	-.04482***	.01372	-3.27	.0011	.45597
HINC	.00263	.00414	.63	.5255	3.01143
SELFEMPL	.02006	.02480	.81	.4186	.05794

Partial Effects for Quantile Count Regression				
Variable	Value	Partial Effect	Semi-Elasticity	
AGE	33.474	.025	.010	
ACADMOS	55.323	.005	.002	
*OWNRENT	.000	-.298	-.118	
HINC	3.012	.104	.041	
*SELFEMPL	.000	-.299	-.118	

* = Dummy variable. Other variables fixed at means.

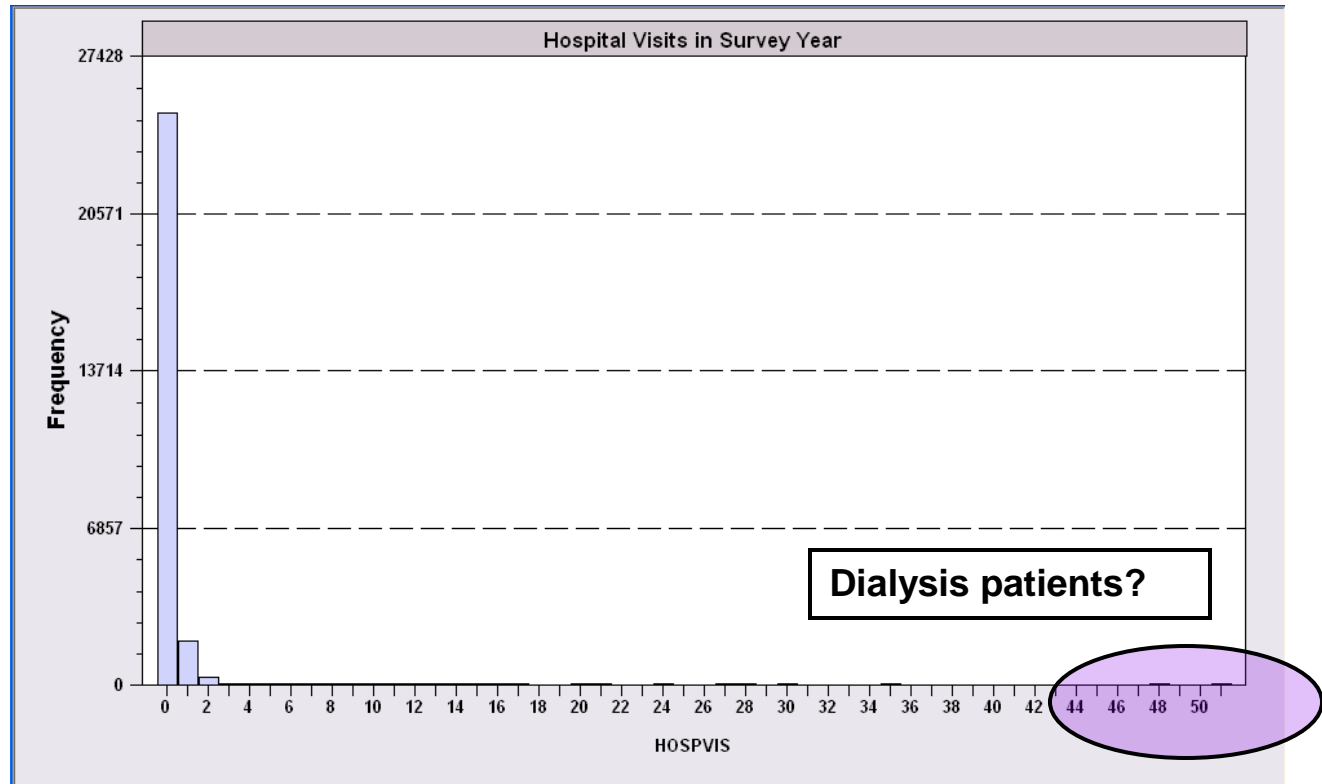


What Have We Learned?

- Models for Count Data
- Data sets contain “unusual” configurations
 - Preponderance of zeros
 - Unusually large observations
- For the preponderance of zeros case, build a richer specification
 - Zero inflation models
 - Two part or hurdle models
- For the unusually large observations, a quantile regression may be more interesting



Wild Observations



Any broad model will assign infinitesimally small probabilities. These observations will not be explained by the model.



Part 7: What Have We Learned?



Rare Events vs. Unusual Events

- Assigning probabilities:
 - What function do probabilities serve?
 - Using information from experts – a Bayesian approach
- Rare events are not merely events that have low probability in the context of the sampling frame.
- Rare events are essentially outside the realm of historical experience and therefore outside the reach of econometric models
- Events with low probability in that frame are “unusual”



Econometric Modeling

- Perhaps it is too ambitious to hope to build econometric “models” for rare events
- Models can be readily extended to accommodate unusual events within the context of the sampling frame.



Unusual Events and Outliers

- Outliers are “unusual” in the context of the model
- Outlier is a subjective term. Computers cannot appropriately determine that observations are outliers
- Models may be merely inadequate. Outliers may be a consequence of the specification.



Regression Modeling

- Outliers and unusual observations in the context of the linear regression model
- Quantile regression model
 - A way to immunize least squares from extreme observations
 - A tool to study different features of the population (other than the conditional mean function)



Binary Choice

- Standard methods of modeling for binary outcomes
- Nonstandard situations
 - Preponderance of ones or zeros
 - Adjusting binary choice analysis for “unusual events”
- An adjustment to standard inference procedures, not a new modeling framework



Models for Count Data

- Standard methods of analyzing counts
- Nonstandard data sets have preponderance of zeros
 - Two part models that accommodate two decisions
 - Zero inflation models that accommodate a richer data generating process
- Extreme values
 - Standard count methods are less affected
 - Quantile models for counts are useful in the same way that quantile regression for continuous data is.



Econometric Models for Rare Events

- Econometric models assume there is and has been order in the universe.
- Rare events are outside the realm of this modeling paradigm. An event, by its nature, is not a draw from a stable data generating process.
- Hence, we build econometric models that accommodate unusual events.



Thank You!

William Greene

Department of Economics

Stern School of Business

<http://pages.stern.nyu.edu/~wgreene>

wgreene@stern.nyu.edu