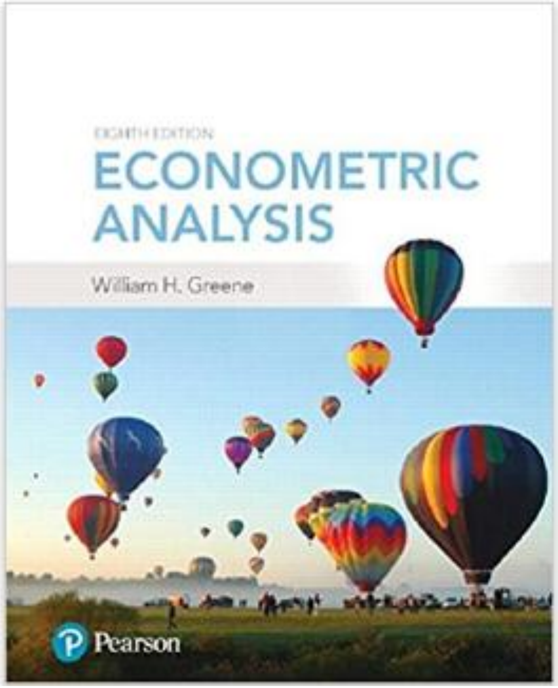
Econometrics I

Professor William Greene
Stern School of Business
Department of Economics



Econometrics I

Part 18 – Maximum Likelihood

Maximum Likelihood Estimation

This defines a class of estimators based on the particular distribution assumed to have generated the observed random variable.

Not estimating a mean – least squares is not available

Estimating a mean (possibly), but also using information about the distribution

Setting Up the MLE

The distribution of the observed random variable is written as a function of the parameters to be estimated

$P(y_i|\text{data},\boldsymbol{\beta})$ = Probability density | parameters.

The likelihood function is constructed from the density

Construction: Joint probability density function of the observed sample of data – generally the product when the data are a *random* sample.

(Log) Likelihood Function

- $f(y_i|\boldsymbol{\beta},\mathbf{x}_i)$ = probability density of observed y_i given parameter(s) and possibly data, $\mathbf{x}_i$.
- Observations are independent
- Joint density = $\Pi_i\, f(y_i|\boldsymbol{\beta},\mathbf{x}_i)$ = $L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$
- $f(y_i|\boldsymbol{\beta},\mathbf{x}_i)$ is the contribution of observation i to the likelihood.
- The MLE of $\boldsymbol{\beta}$ maximizes $L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X})$
- In practice it is usually easier to maximize

$$\log L(\boldsymbol{\beta}|\mathbf{y},\mathbf{X}) = \Sigma_i \log f(y_i|\boldsymbol{\beta},\mathbf{x}_i)$$

Average Time Until Failure

Estimating the average time until failure, θ, of light bulbs.

y_i = observed life until failure.

$f(y_i|\theta) = (1/\theta)\exp(-y_i/\theta)$

$L(\theta) = \Pi_i f(y_i|\theta) = \boldsymbol{\theta}^{\mathbf{-n}} \exp(-\Sigma y_i/\boldsymbol{\theta})$

$\log L(\theta) = -n\log(\theta) - \Sigma y_i/\theta$

Likelihood equation: $\partial \log L(\theta)/\partial\theta = -n/\theta + \Sigma y_i/\theta^2 = 0$

Solution: $\theta_{MLE} = \Sigma y_i /n$. Note: $E[y_i] = \theta$

Note, $\partial \log f(y_i|\theta)/\partial\theta = -1/\theta + y_i/\theta^2$

Since $E[y_i] = \theta$, $E[\partial \log f(\theta)/\partial\theta] = 0$.

Extension: Loglinear Model: $\theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}) = E[y_i|\mathbf{x}_i]$

The MLE

The log-likelihood function: $\log L(\boldsymbol{\beta}|\text{data})$

The likelihood equation(s):

First derivatives of logL equal zero at the MLE.

$(1/n)\Sigma_i \,\partial \log f(y_i|\boldsymbol{\beta},\mathbf{x}_i)/\partial\boldsymbol{\beta}_{\mathbf{MLE}} = \mathbf{0}.$

(Sample statistic.) (The 1/n is irrelevant.)

"First order conditions" for maximization

Usually a nonlinear estimator.

A moment condition - its counterpart is the fundamental theoretical result $E[\partial \log L/\partial\boldsymbol{\beta}] = \mathbf{0}$.

Properties of the MLE

- **Consistent**: Not necessarily unbiased, however
- **Asymptotically normally distributed**: Proof based on central limit theorems
- **Asymptotically efficient**: Among the possible estimators that are consistent and asymptotically normally distributed – counterpart to Gauss-Markov for linear regression
- **Invariant**: The MLE of g(θ) is g(the MLE of θ)

The Linear (Normal) Model

Definition of the likelihood function - joint density of the observed data, written as a function of the parameters we wish to estimate.

Definition of the maximum likelihood estimator as that function of the observed data that maximizes the likelihood function, or its logarithm.

For the model: $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$, where $\varepsilon_i \sim N[0,\sigma^2]$,

the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad \text{and} \quad s^2 = \mathbf{e}'\mathbf{e}/n.$$

That is, least squares is ML for the slopes, but the variance estimator makes no degrees of freedom correction, so the MLE is biased.

Normal Linear Model

The log-likelihood function

$= \Sigma_i \log f(y_i|\theta)$

= sum of logs of densities.

For the linear regression model with normally distributed disturbances

$$\log L = \Sigma_i \left[-\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log \sigma^2 - \tfrac{1}{2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/\sigma^2 \right].$$

$$= -n/2[\log 2\pi + \log\sigma^2 + v^2/\sigma^2]$$

$$v^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n$$

Likelihood Equations

The estimator is defined by the function of the data that equates
∂log-L/$\partial\boldsymbol{\theta}$ to **0**. (Likelihood equation)

The derivative vector of the log-likelihood function is the *score function*. For the regression model,

$$\mathbf{g} = [\partial \log L/\partial \boldsymbol{\beta} \,,\, \partial \log L/\partial \sigma^2]'$$

$$= \partial \log L/\partial \boldsymbol{\beta} = \Sigma_i \, [(1/\sigma^2)\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\,] \qquad = \mathbf{X}'\boldsymbol{\varepsilon}/\boldsymbol{\sigma}^2 \,.$$

$$\partial \log L/\partial \sigma^2 = \Sigma_i \, [-1/(2\sigma^2) + (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2/(2\sigma^4)] \;= -n/2\sigma^2 \, [1 - s^2/\sigma^2]$$

For the linear regression model, the first derivative vector of logL is

$$\underset{(K\times 1)}{(1/\sigma^2)\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})} \quad \text{and} \quad \underset{(1\times 1)}{(1/2\sigma^2)\, \Sigma_i \, [(y_i - \mathbf{x}_i'\boldsymbol{\beta})^{2/}\sigma^2 - 1]}$$

Note that we could compute these functions at *any* $\boldsymbol{\beta}$ and σ^2. If we compute them at **b** and **e′e**/n, the functions will be identically zero.

Maximizer of the log likelihood? Use the Information Matrix

The negative of the second derivatives matrix of the log-likelihood,

$$-\mathbf{H} = -\sum_i \frac{\partial^2 \log f_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

For a maximizer, **-H** is positive definite.

-H forms the basis for estimating the variance of the MLE.

It is usually a random matrix. –**H** is the information matrix.

Hessian for the Linear Model

$$-\frac{\partial^2 \log L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\begin{bmatrix} \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \frac{\partial^2 \log L}{\partial \sigma^2 \partial \boldsymbol{\beta}'} & \frac{\partial^2 \log L}{\partial \sigma^2 \partial \sigma^2} \end{bmatrix}$$

$$= \frac{1}{\sigma^2} \begin{bmatrix} \sum_i \mathbf{x}_i \mathbf{x}_i' & \frac{1}{\sigma^2} \sum_i \mathbf{x}_i (y_i - \mathbf{x}_i' \beta) \\ \frac{1}{\sigma^2} \sum_i (y_i - \mathbf{x}_i' \beta) \mathbf{x}_i' & \frac{1}{2\sigma^4} \sum_i (y_i - \mathbf{x}_i' \beta)^2 \end{bmatrix}$$

Note that the off diagonal elements have expectation zero.

Information Matrix

This can be computed at any vector $\boldsymbol{\beta}$ and scalar σ^2. You can take expected values of the parts of the matrix to get

$$-E[\mathbf{H}] = \begin{bmatrix} \frac{1}{\sigma^2}\sum_i \mathbf{x}_i\mathbf{x}_i' & \mathbf{0}' \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{bmatrix}$$

(which should look familiar). The off diagonal terms go to zero (one of the assumptions of the linear model).

Asymptotic Variance

- The asymptotic variance is $\{-E[\mathbf{H}]\}^{-1}$ i.e., the inverse of the information matrix.

$$\{-E[\mathbf{H}]\}^{-1} = \begin{bmatrix} \sigma^2\left[\sum_i \mathbf{x}_i\mathbf{x}_i'\right]^{-1} & \mathbf{0}' \\ \mathbf{0} & \dfrac{2\sigma^4}{n} \end{bmatrix} = \begin{bmatrix} \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} & \mathbf{0}' \\ \mathbf{0} & \dfrac{2\sigma^4}{n} \end{bmatrix}$$

- There are several ways to estimate this matrix
  - Inverse of negative of expected second derivatives
  - Inverse of negative of actual second derivatives
  - Inverse of sum of squares of first derivatives
  - Robust matrix for some special cases

Computing the Asymptotic Variance

We want to estimate $\{-E[\mathbf{H}]\}^{-1}$ Three ways:

(1) Just compute the negative of the actual second derivatives matrix and invert it.

(2) Insert the maximum likelihood estimates into the known expected values of the second derivatives matrix. Sometimes (1) and (2) give the same answer (for example, in the linear regression model).

(3) Since E[**H**] is the variance of the first derivatives, estimate this with the sample variance (i.e., mean square) of the first derivatives, then invert the result. This will almost always be different from (1) and (2).

Since they are estimating the same thing, in large samples, all three will give the same answer. Current practice in econometrics often favors (3).   Stata rarely uses (3). Others do.

Model for a Binary Dependent Variable

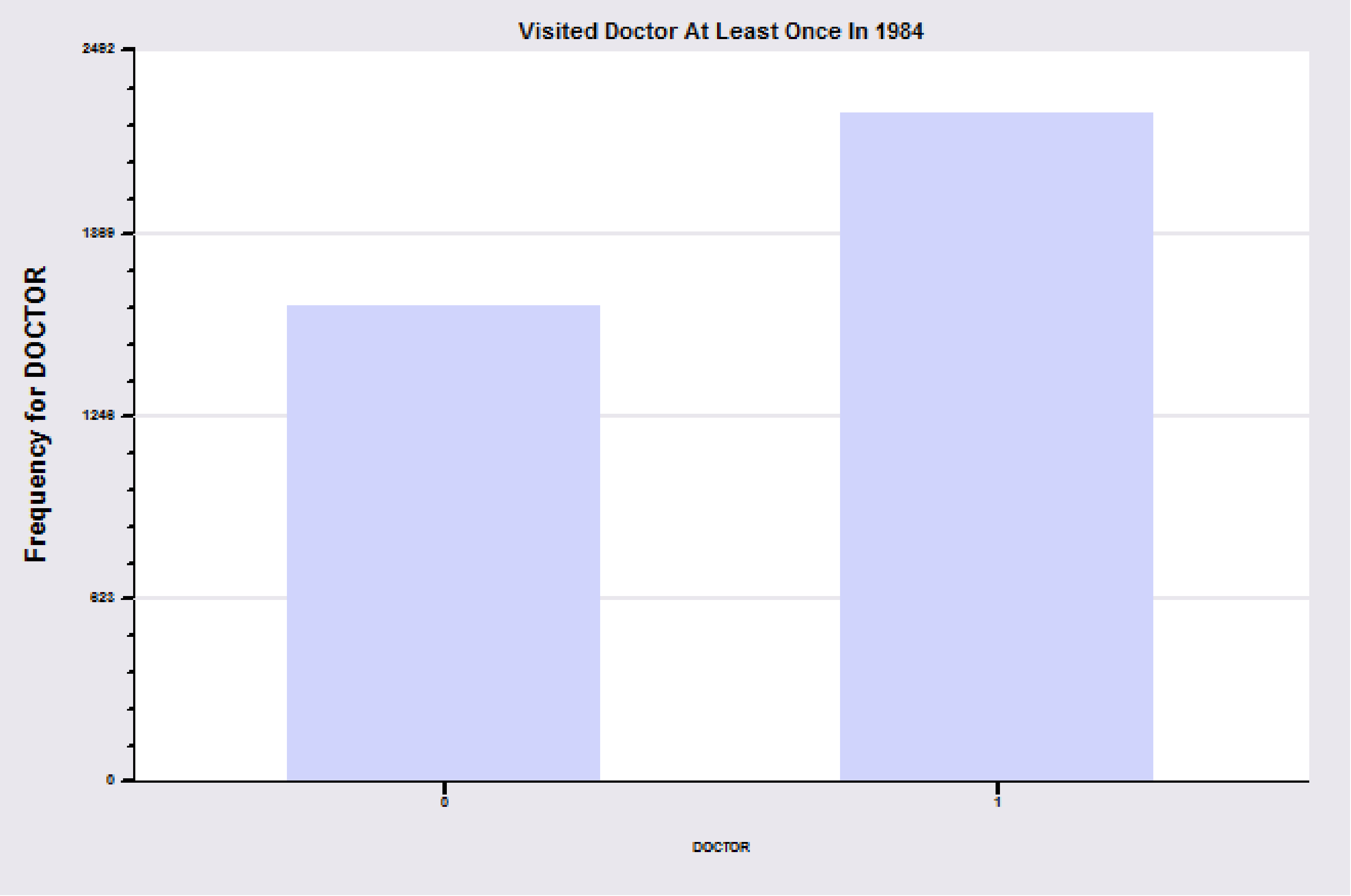
- Binary outcome.
  - Event occurs or doesn't (e.g., the person adopts green technology, the person enters the labor force, etc.)
  - Model the probability of the event. P(**x**)=Prob(y=1|**x**)
  - Probability responds to independent variables
- Requirements for a probability
  - 0 < Probability < 1
  - P(**x**) should be monotonic in **x** – it's a CDF

Behavioral Utility Based Approach

- Observed outcomes partially reveal underlying preferences
- There exists an underlying preference scale defined over alternatives, U*(choices)
- Revelation of preferences between two choices labeled 0 and 1 reveals the ranking of the underlying utility
  - U*(choice 1) > U*(choice 0) $\longrightarrow$ Choose 1
  - U*(choice 1) $\leq$ U*(choice 0) $\longrightarrow$ Choose 0
- Net utility = U = U*(choice 1) - U*(choice 0). U > 0 => choice 1

Binary Outcome: Visit Doctor

In the 1984 year of the GSOEP, 2265 of 3874 individuals visited the doctor at least once.



A Random Utility Model for the Binary Choice

- Yes or No decision | Visit or not visit the doctor
- Model: Net utility of visit at least once
- Net utility depends on observables and unobservables

Random Utility

U_{doctor} = Net utility = $U^*_{visit} - U^*_{not\ visit}$

$U_{doctor} = \alpha + \beta_1 Age + \beta_2 Income + \beta_3 Sex + \varepsilon$

Choose to visit at least once if net utility is positive

- Observed Data: **X** = Age, Income, Sex
  y = 1 if choose visit, $\Leftrightarrow U_{doctor} > 0$, 0 if not.

Modeling the Binary Choice Between the Two Alternatives

Net Utility $U_{doctor} = U^*_{visit} – U^*_{not\ visit}$

$U_{doctor} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon$

Chooses to visit: $U_{doctor} > 0$

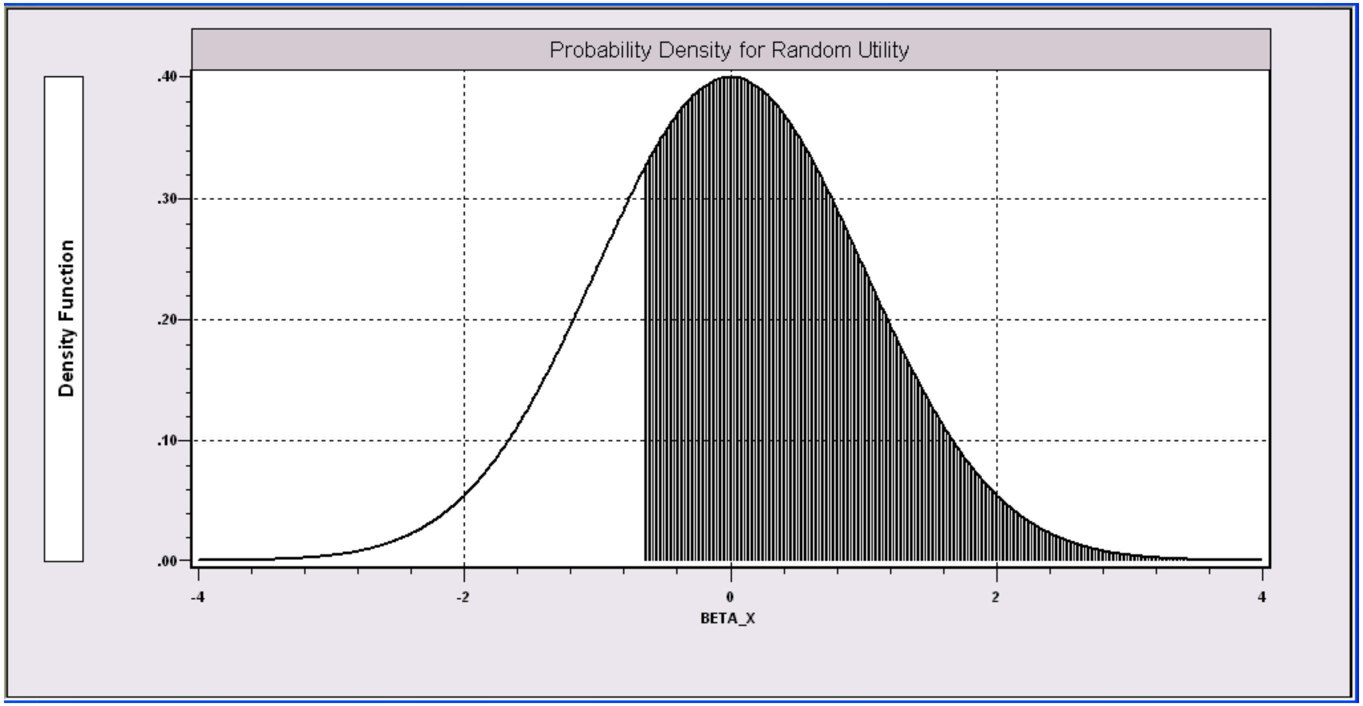
$$\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon > 0$$

Choosing to visit is a random outcome because of ε

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

Probability Model for Choice Between Two Alternatives

People with the same (Age,Income,Sex) will make different choices because ε is random. We can model the *probability* that the random event "visits the doctor"will occur.



Probability is governed by ε, the random part of the utility function.

Event DOCTOR=1 occurs if $\varepsilon > -(\alpha + \beta_1 Age + \beta_2 Income + \beta_3 Sex)$

We model the probability of this event.

An Application

27,326 Observations in GSOEP Sample

- 1 to 7 years, panel
- 7,293 households observed
- We use the 1994 year; 3,337 household observations

Descriptive Statistics for 4 variables

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases	Missing
DOCTOR	.657980	.474456	0.0	1.0	3377	0
AGE	42.62659	11.58599	25.0	64.0	3377	0
INCOME	.444764	.216586	.034000	3.0	3377	0
FEMALE	.463429	.498735	0.0	1.0	3377	0

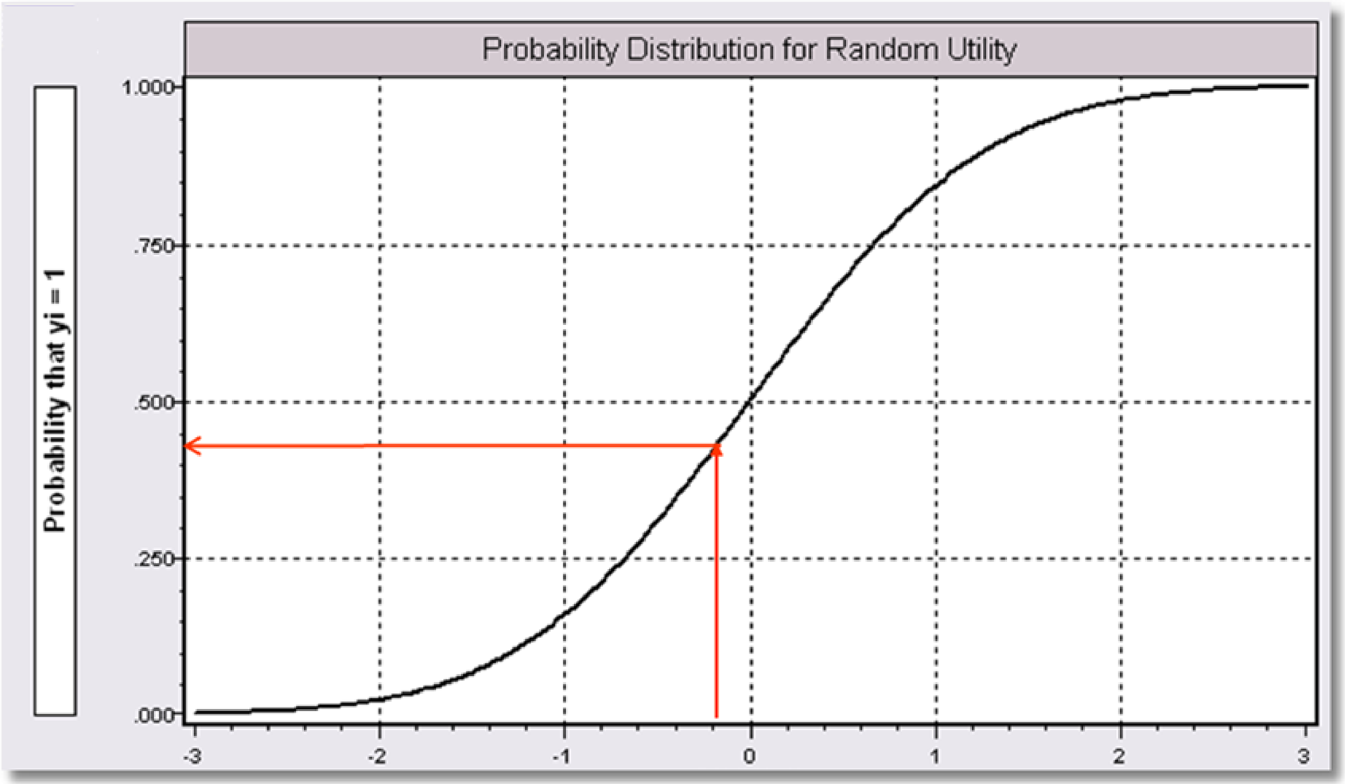
An Econometric Model

- Choose to visit iff $U_{doctor} > 0$
  - $U_{doctor} = \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex} + \varepsilon$
  - $U_{doctor} > 0 \Leftrightarrow \varepsilon > -(\alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})$
    $\varepsilon < \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})$
- Probability model: For any person observed by the analyst,

  $\text{Prob(doctor=1)} = \text{Prob}(\varepsilon \leq \alpha + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Sex})$
- Note the relationship between the unobserved ε and the observed outcome DOCTOR.

Index = $\alpha+\beta_1$Age + β_2 Income + β3 Sex
Probability = a function of the Index.
P(Doctor = 1) = f(Index)

Internally consistent probabilities:
(1) (Coherence) 0 < Probability < 1
(2) (Monotonicity) Probability increases with Index.



A Fully Parametric Model

- Index Function: $U = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$
- Observation Mechanism: $y = 1[U > 0]$
- Distribution: $\varepsilon \sim f(\varepsilon)$; Normal, Logistic, …
- Maximum Likelihood Estimation:

$$\text{Max}(\boldsymbol{\beta}) \log L = \Sigma_i \log \text{Prob}(Y_i = y_i | x_i)$$

A Parametric Logit Model

```
-----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                DOCTOR
Log likelihood function      -2097.48109
Restricted log likelihood    -2169.26982
Chi squared [   3](P= .000)    143.57744
Significance level               .00000
McFadden Pseudo R-squared      .0330935
Estimation based on N =   3377, K =   4
Inf.Cr.AIC  =    4203.0 AIC/N =    1.245
--------+--------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
  DOCTOR|  Coefficient       Error        z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
Constant|    -.42085***      .15810     -2.66  .0078     -.73072    -.11099
     AGE|     .02365***      .00328      7.21  .0000      .01722     .03008
  INCOME|    -.44198***      .16936     -2.61  .0091     -.77393    -.11003
  FEMALE|     .63825***      .07551      8.45  .0000      .49026     .78624
--------+--------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```

We examine the model components.

Parametric Model Estimation

- How to estimate α, β_1, β_2, β_3?
  - The technique of maximum likelihood

$$L = \prod_{y=0} \text{Prob}[y = 0 \mid \mathbf{x}] \times \prod_{y=1} \text{Prob}[y = 1 \mid \mathbf{x}]$$

  - Prob[doctor=1] = Prob[ε > -(α + β_1 Age + β_2 Income + β_3 Sex)]

    Prob[doctor=0] = 1 – Prob[doctor=1]

- Requires a model for the probability

Completing the Model: F(Ɛ)

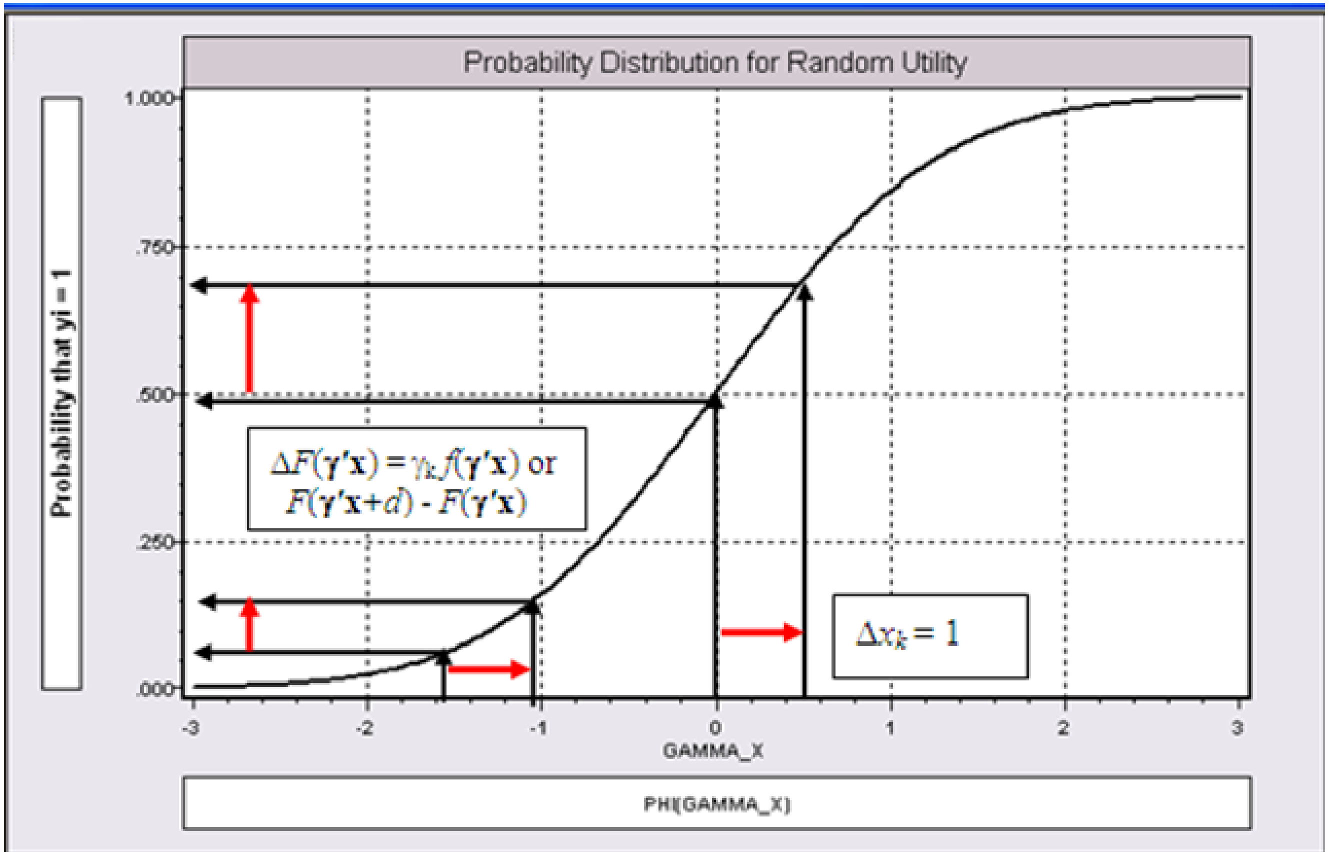
- The distribution
  - Normal: **PROBIT**, natural for behavior
  - Logistic: **LOGIT**, allows “thicker tails”
  - Gompertz: **EXTREME VALUE**, asymmetric
  - Others…
- Does it matter?
  - Yes, large difference in estimates
  - Not much, quantities of interest are more stable.

Estimated Binary Choice Models for Three Distributions

	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	

Log-L(0) = log likelihood for a model that has only a constant term.
Ignore the t ratios for now.

Effect on Predicted Probability of an Increase in Age



$\alpha + \boxed{\beta_1 \text{ (Age+1)}} + \beta_2 \text{ (Income)} + \beta_3 \text{ Sex}$ (β_1 is positive)

Partial Effects in Probability Models

- Prob[Outcome] = some $F(\alpha+\beta_1\text{Income}\ldots)$
- "Partial effect" = $\partial F(\alpha+\beta_1\text{Income}\ldots) / \partial\text{"x"}$ (derivative)
  - Partial effects are derivatives
  - Result varies with model

Logit: $\partial F(\alpha+\beta_1\text{Income}\ldots) / \partial\mathbf{x}$	=	Prob * (1-Prob) $\times\ \boldsymbol{\beta}$
Probit: $\partial F(\alpha+\beta_1\text{Income}\ldots) / \partial\mathbf{x}$	=	Normal density $\times\ \boldsymbol{\beta}$
Extreme Value: $\partial F(\alpha+\beta_1\text{Income}\ldots) / \partial\mathbf{x}$	=	Prob * (-log Prob) $\times\ \boldsymbol{\beta}$

  - Scaling usually erases model differences

Partial effect for the logit model

$$\text{Prob}(\text{doctor}=1) = \frac{\exp(\alpha+\beta_1\text{Age}+\beta_2\text{Income}+\beta_3\text{Sex})}{1+\exp(\alpha+\beta_1\text{Age}+\beta_2\text{Income}+\beta_3\text{Sex})}$$

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

$$= \Lambda(\boldsymbol{\beta}'\mathbf{x})$$

The derivative with respect to one of the variables is

$$\frac{\partial\Lambda(\boldsymbol{\beta}'\mathbf{x})}{\partial x_k} = \left[\Lambda(\boldsymbol{\beta}'\mathbf{x})\right]\left[1-\Lambda(\boldsymbol{\beta}'\mathbf{x})\right]\beta_k$$

(1) A multiple of the coefficient, not the coefficient itself

(2) A function of all of the coefficients and variables

(3) Evaluated using the data and model parts after the model is estimated.

Similar computations apply for other models such as probit.

Estimated Partial Effects for Three Models

(Standard errors to be considered later)

	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

Partial Effect for a Dummy Variable Computed Using Means of Other Variables

- Prob[y_i = 1|$\mathbf{x}_i$,d_i] = F($\boldsymbol{\beta}$'$\mathbf{x}_i$+γd_i) where d is a dummy variable such as Sex in our doctor model.
- For the probit model, Prob[y_i = 1|$\mathbf{x}_i$,d_i] = $\Phi(\boldsymbol{\beta}'\mathbf{x}+\gamma d)$, Φ = the normal CDF.
- Partial effect of d

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

$$= \delta(d_i) = \Phi\left(\hat{\boldsymbol{\beta}}'\overline{\mathbf{x}} + \hat{\gamma}\right) - \Phi\left(\hat{\boldsymbol{\beta}}'\overline{\mathbf{x}}\right)$$

Partial Effect – Dummy Variable

```
------------------------------------------------------------------------
Partial derivatives of E[y] = F[*]  with
respect to the vector of characteristics
They are computed at the means of the Xs
Observations used for means are All Obs.
--------+---------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]  Elasticity
--------+---------------------------------------------------------------
        |Index function for probability
Constant|    -.09186***        .03550      -2.588   .0097
     AGE|     .00527***        .00073       7.269   .0000        .33855
  INCOME|    -.09897***        .03755      -2.636   .0084       -.06632
        |Marginal effect for dummy variable is P|1 - P|0.
  FEMALE|     .13958***        .01618       8.624   .0000        .09745
--------+---------------------------------------------------------------
Note: ***, **, * = Significance at 1%, 5%, 10% level.
Elasticity for a binary variable = marginal effect/Mean.
------------------------------------------------------------------------
```

Computing Partial Effects

- **Compute at the data means (PEA)**
  - Simple
  - Inference is well defined.
  - Not realistic for some variables, such as Sex

- **Average the individual effects (APE)**
  - More appropriate
  - Asymptotic standard errors are slightly more complicated.

Partial Effects

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

Partial Effect $= \dfrac{\partial \mathrm{P_i}}{\partial \mathbf{x}_i} = \dfrac{\partial \mathrm{F}(\boldsymbol{\beta}'\mathbf{x}_i)}{\partial \mathbf{x}_i} = \mathrm{f}(\boldsymbol{\beta}'\mathbf{x}_i) \times \boldsymbol{\beta} = \mathbf{d}_i$

Partial Effect at the Means $= \mathrm{f}(\boldsymbol{\beta}'\overline{\mathbf{x}}) \times \boldsymbol{\beta} = \mathrm{f}\left(\boldsymbol{\beta}'\left(\frac{1}{n}\Sigma_{i=1}^{n}\mathbf{x}_i\right)\right) \times \boldsymbol{\beta}$

Average Partial Effect $= \frac{1}{n}\Sigma_{i=1}^{n}\mathbf{d}_i = \left(\frac{1}{n}\Sigma_{i=1}^{n}\mathrm{f}(\boldsymbol{\beta}'\mathbf{x}_i)\right) \times \boldsymbol{\beta}$

Both are estimates of $\boldsymbol{\delta} = \mathrm{E}[\mathbf{d}_i]$ under certain assumptions.

The two approaches usually give similar answers, though sometimes the results differ substantially.

	Average Partial Effects	Partial Effects at Data Means
Age	0.00512	0.00527
Income	-0.09609	-0.09871
Female	0.13792	0.13958

APE vs. Partial Effects at the Mean

Delta Method for Average Partial Effect

$$\text{Estimator of Var}\left[\frac{1}{N}\sum_{i=1}^{N}\text{PartialEffect}_i\right] = \bar{\mathbf{G}}\,\mathbf{Var}\left[\hat{\boldsymbol{\beta}}\right]\bar{\mathbf{G}}'$$

```
--> partials ; effects: hhninc/female ; summary $
--------------------------------------------------------------------
Partial Effects for Probit Probability Function
Partial Effects Averaged Over Observations
* ==> Partial Effect for a Binary Variable
--------------------------------------------------------------------
                      Partial     Standard
(Delta method)        Effect       Error       |t|  95% Confidence Interval
--------------------------------------------------------------------
      HHNINC         -.05496       .03762      1.46      -.12869       .01877
   *  FEMALE          .14021       .01599      8.77       .10886       .17155
--------------------------------------------------------------------
--> partials ; effects: hhninc/female ; summary ; means$
--------------------------------------------------------------------
Partial Effects for Probit Probability Function
Partial Effects Computed at data Means
* ==> Partial Effect for a Binary Variable
--------------------------------------------------------------------
                      Partial     Standard
(Delta method)        Effect       Error       |t|  95% Confidence Interval
--------------------------------------------------------------------
      HHNINC         -.06374       .04009      1.59      -.14232       .01484
      FEMALE          .15045       .01752      8.59       .11611       .18479
--------------------------------------------------------------------
```

Scottish Journal of Political Economy, Vol. 57, No. 1, February 2010

Published by Blackwell Publishing Ltd, 9600 Garsington Road, Oxford, OX4 2DQ, UK and 350 Main St, Malden, MA, 02148, USA

SHEDDING LIGHT ON THE LIGHT BULB PUZZLE: THE ROLE OF ATTITUDES AND PERCEPTIONS IN THE ADOPTION OF ENERGY EFFICIENT LIGHT BULBS

Corrado Di Maria, Susana Ferreira** and Emiliya Lazarova**

Abstract

Despite the potential energy savings and economic benefits associated with compact fluorescent light bulbs, their adoption by the residential sector has been limited to date. In this paper, we present a theoretical model that focuses on the agents' ability to perceive the correct cost of lighting and on the role of environmental attitudes as key determinants of the adoption decision. We use original data from Ireland to test our theoretical predictions. Our results emphasize the importance of education, information and environmental awareness in the adoption decision.

Table 1
Descriptive statistics

Variable	N	Mean	Standard deviation	Minimum	Maximum
Adoption of energy-efficient light bulbs	1392	0.30	0.46	0	1
Support Kyoto	1469	3.05	0.78	1	4
Importance of Environment	1496	2.51	0.59	1	3
Knowledge of Environment	1500	0.85	0.35	0	1
Education (reference = primary education)					
Lower secondary	1500	0.19	0.39	0	1
Upper secondary	1500	0.47	0.50	0	1
University degree	1500	0.17	0.38	0	1
Income (€)	1497	22,987	11,644	1852	57,138
Rural dwelling	1500	0.38	0.49	0	1
Own house	1480	0.78	0.41	0	1
Age	1492	43.61	17.10	18	90
Sex (1 = male)	1500	0.48	0.50	0	1
Marital status (1 = married)	1500	0.52	0.50	0	1
Number of dependent children	1500	0.88	1.29	0	8

[16] Due to missing observations the final sample in the probit regressions consists of 1339 observations. The effective response rate is 66.6%. The margin of error using the entire sample is $\pm$ 2.5 percent at a 95% confidence level (see UII, 2001).

Tables 3 and 4 only constitute a partial analysis of their actual behaviour. In order to investigate in more detail which factors determine the *individual* decision of adopting energy-efficient light bulbs, we estimate a probit model in which the probability of adopting CFLs is modelled as a function of (a vector of) environmental attitudes and awareness, education, logarithm of income, and other controls:

$$P(\text{adoption} = 1|\mathbf{x}) = G(\beta_0 + \beta_1\,\text{education} + \beta_2\,\log(\text{income}) + \boldsymbol{\chi}\,\textbf{attitudes} + \boldsymbol{\gamma}\,\textbf{controls}),$$

Table 5
Adoption of energy-efficient light bulbs, probit regressions

	(1)		(2)		(3)		(4)	
	Coefficient	Marginal effects	Coefficient	Marginal efficient	Coefficient	Marginal effects	Coefficient	Marginal effects
Age	0.003	0.001	0.003	0.001	0.003	0.001	0.003	0.001
	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)
Male	−0.071	−0.024	−0.079	−0.027	−0.084	−0.029	−0.077	−0.026
	(0.075)	(0.025)	(0.076)	(0.026)	(0.075)	(0.025)	(0.076)	(0.026)
Married	0.092	0.031	0.076	0.026	0.079	0.027	0.067	0.022
	(0.096)	(0.033)	(0.098)	(0.033)	(0.095)	(0.032)	(0.098)	(0.033)
Number of dependant children	−0.027	−0.009	−0.035	−0.012	−0.025	−0.008	−0.031	−0.011
	(0.034)	(0.012)	(0.035)	(0.012)	(0.034)	(0.011)	(0.035)	(0.012)
Lower secondary school	0.168	0.059	0.177	0.062	0.167	0.058	0.157	0.054
	(0.141)	(0.050)	(0.141)	(0.051)	(0.140)	(0.050)	(0.143)	(0.051)
Upper secondary school	0.428	0.146	0.401	0.137	0.368	0.126	0.336	0.114
	(0.129)***	(0.044)***	(0.129)***	(0.044)***	(0.128)***	(0.044)***	(0.131)**	(0.044)**
University degree	0.457	0.166	0.400	0.145	0.389	0.140	0.336	0.120
	(0.152)***	(0.058)***	(0.153)***	(0.058)**	(0.152)**	(0.057)**	(0.154)**	(0.057)**
log(Income)	0.294	0.100	0.306	0.104	0.305	0.104	0.285	0.096
	(0.087)***	(0.029)***	(0.087)***	(0.030)***	(0.087)***	(0.030)***	(0.088)***	(0.030)***
Rural	−0.193	−0.065	−0.206	−0.069	−0.210	−0.071	−0.204	−0.068
	(0.078)**	(0.026)**	(0.078)***	(0.026)***	(0.077)***	(0.026)***	(0.079)***	(0.026)***
Own house	0.232	0.076	0.251	0.082	0.255	0.083	0.243	0.079
	(0.109)**	(0.034)**	(0.109)**	(0.034)**	(0.108)**	(0.033)**	(0.110)**	(0.034)**
Importance of environment	0.337	0.115						
	(0.070)***	(0.023)***						
Support for Kyoto			0.205	0.070				
			(0.053)***	(0.018)***				

How Well Does the Model Fit the Data?

- **There is no R squared for a probability model.**
  - Least squares for linear models is computed to maximize R^2
  - There are no residuals or sums of squares in a binary choice model
  - The model is not computed to optimize the fit of the model to the data
- **How can we measure the “fit” of the model to the data?**
  - “Fit measures” computed from the log likelihood
    - **Pseudo R squared = 1 – logL/logL0**
    - **Also called the “likelihood ratio index”**
  - Direct assessment of the effectiveness of the model at predicting the outcome

Pseudo R^2 = Likelihood Ratio Index

$$\text{Pseudo } R^2 = 1 - \frac{\log L \text{ for the model}}{\log L \text{ for a model with only a constant term}}$$

The prediction of the model is $\hat{F} = F\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)$ = Estimated $\text{Prob}(y_i = 1 | x_i)$

Using only the constant term, $F(\alpha)$

$$\begin{aligned}\text{LogL}_0 &= \sum_{i=1}^{n}\{(1-y_i)\log[1-F(\alpha)] + y_i \log F(\alpha)\}\\ &= n_0 \log[1-F(\alpha)] + n_1 \log F(\alpha) < 0\end{aligned}$$

The log likelihood for the model is larger, but also < 0.

$$\text{LRI} = 1 - \frac{\log L}{\log L_0}. \text{ Since } \log L > \log L_0 \quad 0 \le \text{LRI} < 1.$$

The Likelihood Ratio Index

- Bounded by 0 and a number < 1
- Rises when the model is expanded
- **Specific values between 0 and 1 have no meaning**
- Can be strikingly low even in a great model
- Should not be used to compare models
  - Use logL
  - Use information criteria to compare nonnested models
- Can be negative if the model is not a discrete choice model. For linear regression,
  $\log L = -n/2(1+\log 2\pi+\log(\mathbf{e}'\mathbf{e}/n)]$; Positive if $\mathbf{e}'\mathbf{e}/n < 0.058497$

Fit Measures Based on LogL

```
----------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                 DOCTOR
Log likelihood function       -2085.92452      Full model          LogL
Restricted log likelihood     -2169.26982      Constant term only LogL0
Chi squared [    5 d.f.]       166.69058
Significance level                .00000
McFadden Pseudo R-squared       .0384209       1 - LogL/logL0
Estimation based on N =    3377, K =   6
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              1.23892     4183.84905       -2LogL + 2K
Fin.Smpl.AIC     1.23893     4183.87398       -2LogL + 2K + 2K(K+1)/(N-K-1)
Bayes IC         1.24981     4220.59751       -2LogL + KlnN
Hannan Quinn     1.24282     4196.98802       -2LogL + 2Kln(lnN)
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Characteristics in numerator of Prob[Y = 1]
Constant|    1.86428***       .67793        2.750   .0060
     AGE|    -.10209***       .03056       -3.341   .0008      42.6266
   AGESQ|     .00154***       .00034        4.556   .0000      1951.22
  INCOME|     .51206          .74600         .686   .4925       .44476
 AGE_INC|    -.01843          .01691       -1.090   .2756      19.0288
  FEMALE|     .65366***       .07588        8.615   .0000       .46343
--------+-------------------------------------------------------------
```

Fit Measures Based on Predictions

- Computation
  - Use the model to compute predicted probabilities
  - **P = F(a + b_1Age + b_2Income + b_3Female+…)**
  - Use a rule to compute predicted y = 0 or 1
  - Predict y=1 if P is “large” enough
  - Generally use 0.5 for “large” (more likely than not)

$$\hat{y} = 1 \text{ if } \hat{P} > P^*$$

- Fit measure compares predictions to actuals
- Count successes and failures

Computing test statistics requires the log likelihood and/or standard errors based on the Hessian of LogL

Logit: $g_i = y_i - \Lambda_i \quad H_i = \Lambda_i(1-\Lambda_i) \quad E[H_i] = \Psi_i = \Lambda_i(1-\Lambda_i)$

$(q_i = 2y_i - 1,\ z_i = q_i\boldsymbol{\beta}'\mathbf{x}_i.\ \Lambda_i = \exp(z_i)/[1+\exp(z_i)])$

Probit: $g_i = \dfrac{q_i\phi_i}{\Phi_i} \quad H_i = \dfrac{z_i\phi_i}{\Phi_i} + \left(\dfrac{\phi_i}{\Phi_i}\right)^2, \quad E[H_i] = \Psi_i = \dfrac{\phi_i^2}{\Phi_i(1-\Phi_i)}$

$\phi_i = \phi(z_i),\ \Phi_i = \Phi(z_i)$. Note, g_i is a "generalized residual."

Estimators: Based on H_i, $E[H_i]$ and g_i^2 all functions evaluated at z_i

Actual Hessian: $\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^N H_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

Expected Hessian: $\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^N \Psi_i\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

BHHH: $\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \left[\sum_{i=1}^N g_i^2\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

Robust Covariance Matrix

(Robust to the model specification? Latent heterogeneity? Correlation across observations? Not always clear)

"Robust" Covariance Matrix: $\mathbf{V} = \mathbf{A}\,\mathbf{B}\,\mathbf{A}$

$\mathbf{A}$ = negative inverse of second derivatives matrix

$$= \text{ estimated E}\left[-\frac{\partial^2 \log L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right]^{-1} = \left[-\sum_{i=1}^{N}\frac{\partial^2 \log \text{Prob}_i}{\partial\hat{\boldsymbol{\beta}}\partial\hat{\boldsymbol{\beta}}'}\right]^{-1}$$

$\mathbf{B}$ = matrix sum of outer products of first derivatives

$$= \text{ estimated E}\left[\frac{\partial \log L}{\partial\boldsymbol{\beta}}\frac{\partial \log L}{\partial\boldsymbol{\beta}'}\right] = \left[\sum_{i=1}^{N}\frac{\partial \log \text{Prob}_i}{\partial\hat{\boldsymbol{\beta}}}\frac{\partial \log \text{Prob}_i}{\partial\hat{\boldsymbol{\beta}}'}\right]^{-1}$$

For a logit model, $\mathbf{A} = \left[\sum_{i=1}^{N}\hat{P}_i(1-\hat{P}_i)\mathbf{x}_i\mathbf{x}_i'\right]^{-1}$

$$\mathbf{B} = \left[\sum_{i=1}^{N}(y_i-\hat{P}_i)^2\mathbf{x}_i\mathbf{x}_i'\right] = \left[\sum_{i=1}^{N}e_i^2\mathbf{x}_i\mathbf{x}_i'\right]$$

(Resembles the White estimator in the linear model case.)

Robust Covariance Matrix for Logit Model

Doesn't change much. The model is well specified.

DOCTOR	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Conventional Standard Errors						
Constant	1.86428***	.67793	2.75	.0060	.53557	3.19299
AGE	-.10209***	.03056	-3.34	.0008	-.16199	-.04219
AGE^2.0	.00154***	.00034	4.56	.0000	.00088	.00220
INCOME	.51206	.74600	.69	.4925	-.95008	1.97420
	Interaction AGE*INCOME					
_ntrct02	-.01843	.01691	-1.09	.2756	-.05157	.01470
FEMALE	.65366***	.07588	8.61	.0000	.50494	.80237
Robust Standard Errors						
Constant	1.86428***	.68518	2.72	.0065	.52135	3.20721
AGE	-.10209***	.03118	-3.27	.0011	-.16321	-.04098
AGE^2.0	.00154***	.00035	4.44	.0000	.00086	.00222
INCOME	.51206	.75171	.68	.4958	-.96127	1.98539
	Interaction AGE*INCOME					
_ntrct02	-.01843	.01705	-1.08	.2796	-.05185	.01498
FEMALE	.65366***	.07594	8.61	.0000	.50483	.80249

The Effect of Clustering

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

'Cluster' Corrected Covariance Matrix

$C =$ the number if clusters

$n_c =$ number of observations in cluster c

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

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

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

Cluster Correction: Doctor

```
----------------------------------------------------------------------
Binomial Probit Model
Dependent variable                    DOCTOR
Log likelihood function         -17457.21899
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        | Conventional Standard Errors
Constant|    -.25597***       .05481       -4.670   .0000
     AGE|     .01469***       .00071       20.686   .0000      43.5257
    EDUC|    -.01523***       .00355       -4.289   .0000      11.3206
  HHNINC|    -.10914**        .04569       -2.389   .0169       .35208
  FEMALE|     .35209***       .01598       22.027   .0000       .47877
--------+-------------------------------------------------------------
        | Corrected Standard Errors
Constant|    -.25597***       .07744       -3.305   .0009
     AGE|     .01469***       .00098       15.065   .0000      43.5257
    EDUC|    -.01523***       .00504       -3.023   .0025      11.3206
  HHNINC|    -.10914*         .05645       -1.933   .0532       .35208
  FEMALE|     .35209***       .02290       15.372   .0000       .47877
--------+-------------------------------------------------------------
```

Hypothesis Tests

- We consider "nested" models and parametric tests
- Test statistics based on the usual 3 strategies
  - Wald statistics: Use the unrestricted model
  - Likelihood ratio statistics: Based on comparing the two models
  - Lagrange multiplier: Based on the restricted model.
- Test statistics require the log likelihood and/or the first and second derivatives of logL

Base Model for Hypothesis Tests

```
----------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                DOCTOR
Log likelihood function      -2085.92452
Restricted log likelihood    -2169.26982
Chi squared [   5 d.f.]        166.69058
Significance level                .00000
McFadden Pseudo R-squared       .0384209
Estimation based on N =   3377, K =   6
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              1.23892     4183.84905
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Characteristics in numerator of Prob[Y = 1]
Constant|    1.86428***       .67793        2.750   .0060
     AGE|    -.10209***       .03056       -3.341   .0008      42.6266
   AGESQ|     .00154***       .00034        4.556   .0000      1951.22
  INCOME|     .51206          .74600         .686   .4925       .44476
 AGE_INC|    -.01843          .01691       -1.090   .2756      19.0288
  FEMALE|     .65366***       .07588        8.615   .0000       .46343
--------+-------------------------------------------------------------
```

H_0: Age is not a significant determinant of Prob(Doctor = 1)

H_0: $\beta_2 = \beta_3 = \beta_5 = 0$

Likelihood Ratio Test

Null hypothesis restricts the parameter vector

Alternative relaxes the restriction

Test statistic: Chi-squared =

2 (LogL|Unrestricted model – LogL|Restrictions) $\geq$ 0

Degrees of freedom = number of restrictions

LR Test of H_0: $\beta_2 = \beta_3 = \beta_5 = 0$

```
UNRESTRICTED MODEL
Binary Logit Model for Binary Choice
Dependent variable                    DOCTOR
Log likelihood function          -2085.92452
Restricted log likelihood        -2169.26982
Chi squared [    5 d.f.]           166.69058
Significance level                    .00000
McFadden Pseudo R-squared           .0384209
Estimation based on N =    3377, K =    6
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              1.23892     4183.84905
```

```
RESTRICTED MODEL
Binary Logit Model for Binary Choice
Dependent variable                    DOCTOR
Log likelihood function          -2124.06568
Restricted log likelihood        -2169.26982
Chi squared [    2 d.f.]            90.40827
Significance level                    .00000
McFadden Pseudo R-squared           .0208384
Estimation based on N =    3377, K =    3
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC              1.25974     4254.13136
```

Chi squared[3] = 2[-2085.92452 - (-2124.06568)] = 77.46456

Wald Test of H_0: $\beta_2 = \beta_3 = \beta_5 = 0$

Unrestricted parameter vector is estimated

Discrepancy: $\mathbf{q} = \mathbf{Rb} - \mathbf{m}$ is computed
(or $\mathbf{r}(\mathbf{b},\mathbf{m})$ if nonlinear)

Variance of discrepancy is estimated:

$$\text{Var}[\mathbf{q}] = \mathbf{R}\ \mathbf{V}\ \mathbf{R}'$$

Wald Statistic is $\mathbf{q}'[\text{Var}(\mathbf{q})]^{-1}\mathbf{q} = \mathbf{q}'[\mathbf{RVR}']^{-1}\mathbf{q}$

Wald Test

Rb – m = q

1.86428
-0.102093
0.00154004
0.512059
-0.018433
0.653659

0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	0	1	0

-

0
0
0

=

-0.102093
0.00154004
-0.018433

RVR'

0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	0	1	0

0.459584	-0.019886	0.000198178	-0.231031	0.00556395	0.00125442
-0.019886	0.000934022	-9.98629e-006	0.00516338	-0.00014392	-0.000160159
0.000198178	-9.98629e-006	1.14272e-007	-1.03054e-006	2.83849e-007	1.56007e-006
-0.231031	0.00516338	-1.03054e-006	0.556523	-0.0122686	-0.000805883
0.00556395	-0.00014392	2.83849e-007	-0.0122686	0.000285847	2.77535e-005
0.00125442	-0.000160159	1.56007e-006	-0.000805883	2.77535e-005	0.00575729

0	0	0
1	0	0
0	1	0
0	0	0
0	0	1

=

0.000934022	-9.98629e-006	-0.00014392
-9.98629e-006	1.14272e-007	2.83849e-007
-0.00014392	2.83849e-007	0.000285847

-0.102093	0.00154004	-0.018433

$\{$

0.000934022	-9.98629e-006	-0.00014392
-9.98629e-006	1.14272e-007	2.83849e-007
-0.00014392	2.83849e-007	0.000285847

$\}^{-1}$

-0.102093
0.00154004
-0.018433

= 69.0541

Chi squared[3] = 69.0541

Lagrange Multiplier Test of H_0: $\beta_2 = \beta_3 = \beta_5 = 0$

- Restricted model is estimated
- Derivatives of unrestricted model and variances of derivatives are computed at restricted estimates
- Wald test of whether derivatives are zero tests the restrictions
- Usually hard to compute – difficult to program the derivatives and their variances.

LM Test for a Logit Model

- Compute $\mathbf{b}_0$ (subject to restictions) (e.g., with zeros in appropriate positions.
- Compute $P_i(\mathbf{b}_0)$ for each observation.
- Compute $e_i(\mathbf{b}_0) = [y_i - P_i(\mathbf{b}_0)]$
- Compute $\mathbf{g}_i(\mathbf{b}_0) = \mathbf{x}_i e_i$ using full $\mathbf{x}_i$ vector
- $LM = [\Sigma_i \mathbf{g}_i(\mathbf{b}_0)]'[\Sigma_i \mathbf{g}_i(\mathbf{b}_0)\mathbf{g}_i(\mathbf{b}_0)']^{-1}[\Sigma_i \mathbf{g}_i(\mathbf{b}_0)]$

```
? Logit Model with quadratic and interaction
Namelist ; x=one,age,age*age,income,
              age*income,female $
Logit     ; if[year=1994]
          ; Lhs = doctor
          ; Rhs = x
? Constrained MLE. Force 3 coefficients to = 0
          ; cml:b(2)=0,b(3)=0,b(5)=0
          ; Prob = p$
? First derivative (scale part)
Create    ; gi= (doctor - p) ; gi2 = gi*gi $
? Second derivative (scale part)
Create    ; hi=p*(1-p)$
? LM statistic based on BHHH estimator
Matrix ;if[year=1994] ; list ; G = X'gi $
Matrix ;if[year=1994] ; List ; LM = g'*<X'[gi2]X>*g $
? LM statistic uses internal routine
Logit   ; if[year=1994] ; Lhs=doctor ; Rhs=x
        ; Start = b ; Maxit=0$
? LM statistic based on actual second derivatives
Matrix ;if[year=1994] ; List ; ML = g'*<X'[hi]X>*g $
```

(There is a built in function for this computation.)

Restricted Model

```
----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable               DOCTOR
Log likelihood function     -2124.06568
Restricted log likelihood   -2169.26982
Chi squared [  5](P= .000)     90.40827
Significance level               .00000
McFadden Pseudo R-squared      .0208384
Estimation based on N =   3377, K =   3
Inf.Cr.AIC  =   4254.1 AIC/N =    1.260
Linear constraints imposed            3
--------+-------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
  DOCTOR|  Coefficient      Error       z    |z|>Z*        Interval
--------+-------------------------------------------------------------------
Constant|     .52822***      .08978     5.88  .0000      .35227     .70418
     AGE|        0.0     .....(Fixed Parameter).....
 AGE*AGE|        0.0     .....(Fixed Parameter).....
  INCOME|    -.37810**       .16741    -2.26  .0239    -.70623    -.04998
        |Interaction AGE*INCOME
_ntrct02|        0.0     .....(Fixed Parameter).....
  FEMALE|     .67750***      .07483     9.05  .0000      .53084     .82416
--------+-------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

```
|-> Create   ; gi= (doctor - p) ; gi2 = gi*gi $
|-> Matrix ;if[year=1994] ; list ; G = X'gi $
       G|             1
--------+--------------
       1|    .239344E-05
       2|        2268.60
       3|        212205.
       4|    .968396E-06
       5|        849.705
       6|    .238041E-05
|-> Matrix ;if[year=1994] ; List ; ML = g'*<X'[gi2]X>*g $
      ML|             1
--------+--------------
       1|        81.4583
```

```
|-> Matrix ;if[year=1994] ; List ; ML = g'*<X'[hi]X>*g $

      ML|             1
--------+--------------
       1|        71.6745
```

```
B_AND_G|             1              2
--------+-----------------------------
       1|       .528225    .239344E-05
       2|       .000000        2268.60
       3|       .000000        212205.
       4|      -.378105    .968396E-06
       5|       .000000        849.705
       6|       .677500    .238041E-05
```

```
|-> Logit   ; if[year=1994] ; Lhs=doctor ; Rhs=x
    ; Start = b ; Maxit=0$
Maximum of     0 iterations. Exit iterations with status=1.
Maxit = 0. Computing LM statistic at starting values.
No iterations computed and no parameter update done.

-----------------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                    DOCTOR
LM Stat. at start values            71.67452
LM statistic kept as scalar         LMSTAT
Log likelihood function          -2124.06568
Restricted log likelihood        -2169.26982
Chi squared [   5](P= .000)        90.40827
Significance level                   .00000
McFadden Pseudo R-squared          .0208384
Estimation based on N =    3377, K =    6
Inf.Cr.AIC  =    4260.1 AIC/N =     1.262
--------+--------------------------------------------------------------------
        |                Standard            Prob.       95% Confidence
  DOCTOR|  Coefficient     Error       z    |z|>Z*          Interval
--------+--------------------------------------------------------------------
Constant|     .52822       .66783     .79  .4290     -.78069    1.83714
     AGE|      0.0         .02967     .00 1.0000 -.58161D-01  .58161D-01
 AGE*AGE|      0.0         .00032     .00 1.0000 -.63007D-03  .63007D-03
  INCOME|    -.37810       .72928    -.52  .6041    -1.80747    1.05126
        |Interaction AGE*INCOME
_ntrct02|      0.0         .01625     .00 1.0000 -.31844D-01  .31844D-01
  FEMALE|     .67750***    .07522    9.01  .0000      .53007     .82493
--------+--------------------------------------------------------------------
```

```
|-> Matrix ;if[year=1994] ; List ; ML = g'*<X'[hi]X>*g $

      ML|          1
--------+--------------
       1|     71.6745
```

I have a question. The question is as follows. We have a probit model. We used LM tests to test for the hetercodeaticiy in this model and found that there is heterocedasticity in this model...

How do we proceed now? What do we do to get rid of heterescedasticiy?

Testing for heteroscedasticity in a probit model and then getting rid of heteroscedasticit in this model is not a common procedure. In fact I do not remember seen an applied (or theoretical also) works which tests for heteroscedasticiy and then uses a method to get rid of it???

See Econometric Analysis, 7th ed. pages 714-714

Appendix

Properties of the Maximum Likelihood Estimator

We will sketch formal proofs of these results:

The log-likelihood function, again

The likelihood equation and the information matrix.

A linear Taylor series approximation to the first order conditions:

$$\mathbf{g}(\theta_{ML}) = \mathbf{0} \approx \mathbf{g}(\theta) + H(\theta)(\theta_{ML} - \theta)$$

(under regularity, higher order terms will vanish in large samples.)

Our usual approach. Large sample behavior of the left and right hand sides is the same.

A Proof of consistency. (Property 1)

The limiting variance of $\sqrt{n}(\theta_{ML} - \theta)$. We are using the central limit theorem here.

Leads to asymptotic normality (Property 2). We will derive the asymptotic variance of the MLE.

Estimating the variance of the maximum likelihood estimator.

Efficiency (we have not developed the tools to prove this.) The Cramer-Rao lower bound for efficient estimation (an asymptotic version of Gauss-Markov).

Invariance. (A ***VERY*** handy result.) Coupled with the Slutsky theorem and the delta method, the invariance property makes estimation of nonlinear functions of parameters very easy.

Regularity Conditions

- Deriving the theory for the MLE relies on certain “regularity” conditions for the density.
- What they are
  - 1. logf(.) has three continuous derivatives wrt parameters
  - 2. Conditions needed to obtain expectations of derivatives are met. (E.g., range of the variable is not a function of the parameters.)
  - 3. Third derivative has finite expectation.
- What they mean
  - Moment conditions and convergence. We need to obtain expectations of derivatives.
  - We need to be able to truncate Taylor series.
  - We will use central limit theorems

The MLE

The results center on the first order conditions for the MLE

$$\frac{\partial \log L}{\partial \hat{\boldsymbol{\theta}}_{MLE}} = \mathbf{g}\left(\hat{\boldsymbol{\theta}}_{MLE}\right) = \mathbf{0}$$

Begin with a Taylor series approximation to the first derivatives:

$$\mathbf{g}\left(\hat{\boldsymbol{\theta}}_{MLE}\right) = \mathbf{0} \approx \mathbf{g}(\boldsymbol{\theta}) + \mathbf{H}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) \text{ [+ terms o(1/n) that vanish]}$$

The derivative at the MLE, $\hat{\boldsymbol{\theta}}_{MLE}$, is exactly zero. It is close to zero at the true $\boldsymbol{\theta}$, to the extent that $\hat{\boldsymbol{\theta}}_{MLE}$ is a good estimator of $\boldsymbol{\theta}$.

Rearrange this equation and make use of the Slutsky theorem

$$\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) \approx \left[-\mathbf{H}(\boldsymbol{\theta})\right]^{-1} \mathbf{g}(\boldsymbol{\theta})$$

In terms of the original log likelihood

$$\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) \approx \left[-\sum_{i=1}^{n} \mathbf{H}_i(\boldsymbol{\theta})\right]^{-1} \left[\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta})\right]$$

where $\mathbf{g}_i(\boldsymbol{\theta}) = \dfrac{\partial \log f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and $\mathbf{H}_i(\boldsymbol{\theta}) = \dfrac{\partial^2 \log f_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$

Consistency of the MLE

$$\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) \approx \left[-\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}\right)\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}\right)\right]$$

Divide both sums by the sample size.

$$\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) = \left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}\right)\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}\right)\right] + o\left(\frac{1}{n}\right)$$

The approximation is now exact because of the higher order term.

As $n \to \infty$, the third term vanishes. The matrices in brackets are sample means that converge to their expectations.

$$\left[-\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i\left(\boldsymbol{\theta}\right)\right]^{-1} \to \left\{-E\left[\mathbf{H}_i\left(\boldsymbol{\theta}\right)\right]\right\}^{-1}, \text{ a positive definite matrix.}$$

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\left(\boldsymbol{\theta}\right)\right] \to E\left[\mathbf{g}_i\left(\boldsymbol{\theta}\right)\right] = \mathbf{0}, \text{ one of the regularity conditions.}$$

Therefore, collecting terms,

$$\left(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}\right) \to \mathbf{0} \quad \text{or} \quad \text{plim}\ \hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}$$

Asymptotic Variance

Multiply both sides by $\sqrt{n}$. Thus,

$$\sqrt{n}\left(\hat{\theta}_{ML} - \theta\right) \approx [-\mathbf{H}(\theta)/n]^{-1}\sqrt{n}\,[\mathbf{g}(\theta)/n].$$

The limiting variance of the thing on the LHS is the same as the limiting variance of the thing on the RHS. Remember that $[-\mathbf{H}(\theta)/n]^{-1}$ converges to a positive definite matrix. Suppose that **D** is the limiting variance of $\sqrt{n}\,[\mathbf{g}(\theta)/n]$. Then, the limiting variance of

$$\sqrt{n}\left(\hat{\theta}_{ML} - \theta\right) \text{ will be } [-\mathbf{H}(\theta)/n]^{-1} \times \mathbf{D} \times [-\mathbf{H}(\theta)/n]^{-1},$$

so to complete the derivation, we need to know what **D**, the limiting covariance matrix of the score vector. There is a proof in your text of the VIR,

Asymptotic Variance

the limiting variance of $\sqrt{n}\ [\mathbf{g}(\boldsymbol{\theta})/n]$ is $-(1/n)E[\mathbf{H}(\boldsymbol{\theta})]$

(It bears repeating. The variance of the first derivatives vector is the second derivatives matrix. Multiplying it out and using our usual transition from limiting variances to asymptotic variances,

$$\text{Asy.Var}\left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right) = (1/n)\,[-E[\mathbf{H}(\boldsymbol{\theta})/n]]^{-1}.$$

II. Estimating the Asymptotic Covariance Matrix of the Maximum Likelihood Estimator:

$\text{LogL} = \text{LogL}(\theta|\text{data}) \quad = \Sigma_i \log f(y_i|x_i,\theta) \quad = \Sigma_i \log f_i(\theta)$

$g(\theta) = \partial \log L/\partial\theta \quad = \Sigma_i \partial \log f(y_i|x_i,\theta)/\partial\theta \quad = \Sigma_i g_i(\theta)$

$H(\theta) = \partial^2 \log L/\theta\partial\theta' \quad = \Sigma_i \partial^2 \log f(y_i|x_i,\theta)/\partial\theta\partial\theta' \quad = \Sigma_i H_i(\theta)$

(1) Negative inverse of the expected Hessian - using estimates of the expectations (when known)

a. Requires knowledge of the expectation of $\partial^2 \log f(y_i|x_i,\theta)/\partial\theta\partial\theta'$, $-E[H_i(\theta)]$
b. Estimator is then computed by inserting MLE into these functions

$$\text{Est.Asy.Var}[.] = [\Sigma_i -E[H_i(\theta)]]^{-1} \text{ using the MLE}$$

(2) Negative inverse of the actual Hessian as an estimate of its population counterpart

a. Exact expected value of Hessian might be unknown, but we need to estimate the mean, so use the mean of the actual values.
b. Estimator is just the negative inverse of the actual Hessian

$$\text{Est.Asy.Var}[.] = [\Sigma_i -H_i(\theta)]^{-1} \text{ using the MLE}$$

(c. There are cases in which the Hessian does not involve the random variable, so that the actual Hessian equals the expected Hessian.)

(3) Inverse of sum of squares (outer products) of first derivatives, under the theory that the negative of the expected Hessian is the variance of the first derivatives.

a. Negative of expected Hessian is the variance of the first derivatives. Use an empirical estimator
b. Estimator is the sum of "squares" of the first derivatives

$$\text{Est.Asy.Var}[.] = [\Sigma_i g_i(\theta)g_i(\theta)']^{-1} \text{ using the MLE}$$

(This is called the BHHH - Berndt, Hall, Hall, Hausman - estimator

Asymptotic Distribution

You might guess (correctly) normal. Why?

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}\right) \approx [-\mathbf{H}(\boldsymbol{\theta})/n]^{-1}\sqrt{n}\,[\mathbf{g}(\boldsymbol{\theta})/n]\text{, on the}$$

right hand side, is a matrix which converges to something times root n times a sample mean. We can invoke the Lindberg-Feller version of the central limit theorem. The conclusion is

$$(\hat{\boldsymbol{\theta}}_{ML}) \xrightarrow{a} N\left[\ \boldsymbol{\theta},\ [-\mathbf{H}(\boldsymbol{\theta})]^{-1}\ \right]$$

Efficiency: Variance Bound

If the density of the observed random variable satisfies the regularity conditions, then there is a lower bound for the variance of a consistent, normally distributed estimators. This is the *Crame'r - Rao Lower* bound for a regular estimator:

If $f(y_i|\boldsymbol{\theta})$ satisfies the regularity conditions, then, if $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ which is consistent and asymptotically normally distributed and if $\mathbf{V}$ is the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$, then $\mathbf{V} - [-\mathbf{H}(\boldsymbol{\theta})]^{-1}$ is a nonnegative definite matrix. That is, there is no C.A.N. estimator which has a variance which is smaller than the inverse of the information matrix.

VVIR: This means that the MLE is efficient among C.A.N. estimators.

Invariance

The maximum likelihood estimator of a function of θ, say $h(\theta)$ is h(MLE). This is not always true of other kinds of estimators. To get the variance of this function, we would use the delta method. E.g., the MLE of $\boldsymbol{\theta}=(\boldsymbol{\beta}/\sigma)$ is $\mathbf{b}/(\mathbf{e}'\mathbf{e}/n)$

Food Policy 50 (2015) 11–19

Contents lists available at ScienceDirect

Food Policy

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

Does SNAP improve your health? ☆

Christian A. Gregory [a,*], Partha Deb [b,c]

[a] Diet, Safety and Health Economics Branch, Food Economics Division, Economic Research Service, USDA, Washington DC, United States
[b] Dept. of Economics, Hunter College, City University of New York, New York, United States

CrossMark

Table 2
Parameter estimates from ordered and count models.

	SAH
Female	0.034 (0.021)
Black	0.346*** (0.028)
Hispanic	−0.018 (0.029)
Other Race	0.021 * (0.051)
Married	−0.217*** (0.024)

One Vehicle Exempt per Adult	0.116** (0.049)
$tanh(\rho)$ / λ	0.305*** (.047)
$ln(\delta)$	
χ^2_{IV}	17.87*** (0.000)

The Linear Probability "Model"

$$\text{Prob}(y=1|\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}$$

$$E[y|\mathbf{x}] = 0 * \text{Prob}(y=1|\mathbf{x}) + 1\text{Prob}(y=1|\mathbf{x}) = \text{Prob}(y=1|\mathbf{x})$$

$$y = \boldsymbol{\beta}'\mathbf{x} + \varepsilon$$

ROTTEN APPLES: AN INVESTIGATION OF THE
PREVALENCE AND PREDICTORS
OF TEACHER CHEATING

Brian A. Jacob
Steven D. Levitt

Working Paper 9413
http://www.nber.org/papers/w9413

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2002

The Dependent Variable equals zero for 99.1% of the observations. In the sample of 163,474 observations, the LHS variable equals 1 about 1,500 times.

Table 9: OLS Estimates of the Relationship between Cheating and Classroom Characteristics

Independent variables	Dependent variable = Indicator of classroom cheating (1)	(2)	(3)	(4)
Social promotion policy	0.0011 (0.0013)	0.0011 (0.0013)	0.0015 (0.0013)	0.0023 (0.0009)
School probation policy	0.0020 (0.0014)	0.0019 (0.0014)	0.0021 (0.0014)	0.0029 (0.0013)
Prior classroom achievement	-0.0047 (0.0005)	-0.0028 (0.0005)	-0.0016 (0.0007)	-0.0028 (0.0007)
Social promotion*classroom achievement	--	-0.0049 (0.0014)	-0.0051 (0.0014)	-0.0046 (0.0012)
School probation*classroom achievement	--	-0.0070 (0.0013)	-0.0070 (0.0013)	-0.0064 (0.0013)
Mixed grade classroom	-0.0084 (0.0007)	-0.0085 (0.0007)	-0.0089 (0.0008)	-0.0089 (0.0012)
% of students included in official reporting	0.0252 (0.0031)	0.0249 (0.0031)	0.0141 (0.0037)	0.0131 (0.0037)
School*Year Fixed Effects	No	No	No	Yes
Number of observations	163,474	163,474	163,474	163,474

Notes: The unit of observation is classroom*grade*year*subject and the sample includes years eight years (1993 to 2000), four subjects (reading comprehension and three math sections) and five grades (three to seven). The dependent variable is the cheating indicator derived using the 95th percentile cutoff. Robust standard errors clustered by school*year are shown in parenthesis. Other variables included in the regressions in column 1 and 2 include a

2SLS for a binary dependent variable.

Table 10: In Cheating Classrooms, for Whom do Teachers Cheat?

Independent variables	**Dependent variable =** Teacher cheated for the student (1)	(2)	(3)	(4)
Prior achievement in the bottom quartile	0.011 (0.038)	--	-0.007 (0.075)	--
Prior achievement in the 2^{nd} quartile	0.057 (0.024)	--	0.069 (0.039)	--
Prior achievement in the 3^{rd} quartile	0.023 (0.067)	--	-0.012 (0.141)	--
Prior achievement (linear measure)	--	0.0004 (0.0003)	--	0.0005 (0.0004)
Prior achievement (linear) * High-stakes	--	-0.0007 (0.0004)	--	-0.0007 (0.0005)
Excluded from test reporting	-0.045 (0.014)	-0.048 (0.014)	-0.045 (0.021)	-0.052 (0.020)
Male	-0.009 (0.004)	-0.009 (0.004)	-0.014 (0.005)	-0.013 (0.005)
Black	0.005 (0.011)	0.006 (0.011)	0.004 (0.024)	0.001 (0.023)
Hispanic	-0.010 (0.010)	-0.008 (0.009)	0.006 (0.023)	0.004 (0.022)
Age	-0.010 (0.004)	-0.012 (0.004)	-0.015 (0.005)	-0.017 (0.005)
Sample	Full		Low-Achieving Schools	
Number of observations	39,216		23,010	

Notes: The sample includes only those classrooms that were categorized as cheating based on the 95th percentile cutoff in a particular subject and year. The dependent variable takes on the value of one if a *student's* answer string and test score pattern was suspicious at the 90^{th} percentile level, suggesting that the teacher had cheated for that student in the particular subject and year. All models include fixed effects for classroom*year. Low achieving schools are defined as those in which fewer than 25% of students met national norms in reading in 1995. The equations are estimated using 2SLS where a student's test scores at t-2 are used to instrument for the student's t-1 achievement level. Robust standard errors are shown in parenthesis.

Modeling a Binary Outcome

- Did firm *i* produce a product or process innovation in year *t* ? y_{it} : 1=Yes/0=No
- Observed N=1270 firms for T=5 years, 1984-1988
- Observed covariates: $\mathbf{x}_{it}$ = Industry, competitive pressures, size, productivity, etc.
- How to model?
  - Binary outcome
  - Correlation across time
  - Heterogeneity across firms

Application

$$y_{it}^{*} = \beta_1 + \sum_{k=2}^{8} x_{k,it}\beta_k + \varepsilon_{it}\,,\;\; y_{it} = \mathbf{1}\left(y_{it}^{*} > 0\right),$$

$i = 1,\ldots,1270,\ t = 1984,\ldots,1988.$

y_{it} = 1 if a product innovation was realized by German manufacturing firm i in year t, 0 otherwise,

$x_{2,it}$ = Log of industry sales in DM,

$x_{3,it}$ = Import share = ratio of industry imports to (industry sales plus imports),

$x_{4,it}$ = Relative firm size = ratio of employment in business unit to employment in the industry (times 30),

$x_{5,it}$ = FDI share = Ratio of industry foreign direct investment to (industry sales, plus imports),

$x_{6,it}$ = Productivity = Ratio of industry value added to industry employment,

$x_{7,it}$ = Raw materials sector = 1 if the firm is in this sector,

$x_{8,it}$ = Investment goods sector = 1 if the firm is in this sector

Probit and LPM

Variable	PROBIT		LINEARPM	
	Estimate	t ratio	Estimate	t ratio
Constant	-1.96031	-8.508	-.10424	-1.244
LOGSALES	.17711	7.966	.05198	6.524
IMUM	1.13384	7.506	.45284	8.065
SP	1.07274	7.549	.09492	4.093
FDIUM	2.85318	7.096	1.07787	7.567
PROD	-2.34116	-3.272	-.55012	-2.192
RAWMTL	-.27858	-3.452	-.09861	-3.317
INVGOOD	.18796	4.793	.07879	5.372
Log-L	-4114.05			
Log-L(0)	-4283.17			
Rsqrd			.04467	
s.d.e(i)			.47987	

|-> Maketable ; ProbitME,LinearME $

Variable	PROBITME		LINEARME	
	Estimate	t ratio	Estimate	t ratio
LOGSALES	.06573	8.083	.05198	6.524
IMUM	.42080	7.613	.45284	8.065
SP	.39812	7.632	.09492	4.093
FDIUM	1.05890	7.177	1.07787	7.567
PROD	-.86887	-3.278	-.55012	-2.192
RAWMTL	-.10569	-3.420	-.09861	-3.317
INVGOOD	.07045	4.774	.07879	5.372

OLS approximates the partial effects, “directly,” without bothering with coefficients.

Binomial Probit Model
Dependent variable DV

MLE

DV	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
	Index function for probability					
Constant	-2.23278***	.19860	-11.24	.0000	-2.62203	-1.84353
AGE	.01053***	.00186	5.65	.0000	.00688	.01418
EDUC	-.02047*	.01095	-1.87	.0616	-.04193	.00099
MARRIED	-.12096***	.04625	-2.61	.0089	-.21161	-.03030
PUBLIC	.29821***	.09436	3.16	.0016	.11327	.48314
HEALTHY	-.85776***	.04959	-17.30	.0000	-.95496	-.76057

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

Partial derivatives of E[y] = F[*] with
respect to the vector of characteristics
Average partial effects for sample obs.

Average Partial Effects

DV	Partial Effect	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval		
AGE	.00041***	.7425D-04	5.56	.0000	.00027	.00056	
EDUC	-.00080*	.00043	-1.87	.0621	-.00164	.00004	
MARRIED	-.00504**	.00205	-2.46	.0139	-.00906	-.00103	#
PUBLIC	.00919***	.00223	4.12	.0000	.00482	.01356	#
HEALTHY	-.03140***	.00186	-16.92	.0000	-.03503	-.02776	#

\# Partial effect for dummy variable is E[y|x,d=1] - E[y|x,d=0]

Ordinary least squares regression ....
LHS=DV Mean = .01749
Standard deviation = .13110
Fit R-squared = .01955 R-bar squared .01937

OLS Coefficients

DV	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	.02278***	.00682	3.34	.0008	.00942	.03614
AGE	.00044***	.7315D-04	5.98	.0000	.00029	.00058
EDUC	-.00059	.00037	-1.62	.1060	-.00131	.00013
MARRIED	-.00520***	.00187	-2.78	.0055	-.00887	-.00153
PUBLIC	.00700***	.00263	2.66	.0077	.00185	.01215
HEALTHY	-.03261***	.00166	-19.59	.0000	-.03588	-.02935

Odds Ratios

This calculation is not meaningful if the model is not a binary logit model

$$\text{Prob}(y=0\,|\,\mathbf{x},z)=\frac{1}{1+\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)},$$

$$\text{Prob}(y=1\,|\,\mathbf{x},z)=\frac{\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)}{1+\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)}$$

$$\begin{aligned}\text{OR}(\mathbf{x},z)&=\frac{\text{Prob}(y=1\,|\,\mathbf{x},z)}{\text{Prob}(y=0\,|\,\mathbf{x},z)}=\frac{\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)}{1}\\&=\exp(\boldsymbol{\beta}'\mathbf{x}+\gamma z)\\&=\exp(\boldsymbol{\beta}'\mathbf{x})\exp(\gamma z)\end{aligned}$$

$$\frac{\text{OR}(\mathbf{x},z+1)}{\text{OR}(\mathbf{x},z)}=\frac{\exp(\boldsymbol{\beta}'\mathbf{x})\exp(\gamma z+\gamma)}{\exp(\boldsymbol{\beta}'\mathbf{x})\exp(\gamma z)}=\exp(\gamma)$$

Odds Ratio

- Exp(γ) = **multiplicative** change in the odds ratio when z changes by 1 unit.
- dOR($\mathbf{x}$,z)/d$\mathbf{x}$ = OR($\mathbf{x}$,z)*$\boldsymbol{\beta}$, not exp($\boldsymbol{\beta}$)
- The “odds ratio” is not a partial effect – it is not a derivative.
- It is only meaningful when the odds ratio is itself of interest and the change of the variable by a whole unit is meaningful.
- “Odds ratios” might be interesting for dummy variables

Cautions About reported Odds Ratios

```
. logit  grade gpa tuce psi, nolog

Logit estimates                                   Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   2.826113   1.262941     2.24   0.025     .3507938    5.301432
        tuce |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
         psi |   2.378688   1.064564     2.23   0.025       .29218    4.465195
       _cons |  -13.02135   4.931325    -2.64   0.008    -22.68657    -3.35613
------------------------------------------------------------------------------
```

```
. logit  grade gpa tuce psi, or nolog

Logit estimates                                   Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   16.87972   21.31809     2.24   0.025     1.420194    200.6239
        tuce |   1.099832   .1556859     0.67   0.501     .8333651    1.451502
         psi |   10.79073   11.48743     2.23   0.025     1.339344    86.93802
------------------------------------------------------------------------------
```

Model for a Binary Dependent Variable

- Binary outcome.
  - Event occurs or doesn’t (e.g., the democrat wins, the person enters the labor force,…
  - Model the probability of the event. P(**x**)=Prob(y=1|**x**)
  - Probability responds to independent variables
- Requirements
  - 0 < Probability < 1
  - P(**x**) should be monotonic in **x** – it’s a CDF

Two Standard Models

- Based on the normal distribution:
  - Prob[y=1|**x**] = $\Phi(\boldsymbol{\beta}'\mathbf{x})$ = CDF of normal distribution
  - The "probit" model
- Based on the logistic distribution
  - Prob[y=1|x] = exp($\boldsymbol{\beta}'\mathbf{x}$)/[1+ exp($\boldsymbol{\beta}'\mathbf{x}$)]
  - The "logit" model
- Log likelihood
  - $P(y|x) = (1-F)^{(1-y)} F^y$ where F = the cdf
  - $\text{LogL} = \Sigma_i\, (1-y_i)\log(1-F_i) + y_i \log F_i$

    $= \Sigma_i\, F[(2y_i-1)\boldsymbol{\beta}'\mathbf{x}]$ since $F(-t)=1-F(t)$ for both.

Coefficients in the Binary Choice Models

$E[y|x] = 0*(1-F_i) + 1*F_i = P(y=1|x)$

$= F(\boldsymbol{\beta}'\mathbf{x})$

The coefficients are not the slopes, as usual in a nonlinear model

$$\partial E[y|x]/\partial x = f(\boldsymbol{\beta}'\mathbf{x})\boldsymbol{\beta}$$

These will look similar for probit and logit

Application: Female Labor Supply

1975 Survey Data: Mroz (Econometrica) 753 Observations

Descriptive Statistics

All observations in current sample

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases	Missing
LFP	.568393	.495630	.000000	1.00000	753	0
WHRS	740.576	871.314	.000000	4950.00	753	0
KL6	.237716	.523959	.000000	3.00000	753	0
K618	1.35325	1.31987	.000000	8.00000	753	0
WA	42.5378	8.07257	30.0000	60.0000	753	0
WE	12.2869	2.28025	5.00000	17.0000	753	0
WW	2.37457	3.24183	.000000	25.0000	753	0
RPWG	1.84973	2.41989	.000000	9.98000	753	0
HHRS	2267.27	595.567	175.000	5010.00	753	0
HA	45.1208	8.05879	30.0000	60.0000	753	0
HE	12.4914	3.02080	3.00000	17.0000	753	0
HW	7.48218	4.23056	.412100	40.5090	753	0
FAMINC	23080.6	12190.2	1500.00	96000.0	753	0
KIDS	.695883	.460338	.000000	1.00000	753	0

Estimated Choice Models for Labor Force Participation

```
----------------------------------------------------------------------
Binomial Probit Model
Dependent variable                    LFP
Log likelihood function      -488.26476  (Probit)
Log likelihood function      -488.17640  (Logit)
--------+-------------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]   Mean of X
--------+-------------------------------------------------------------
        |Index function for probability
Constant|     .77143          .52381        1.473   .1408
      WA|    -.02008          .01305       -1.538   .1241      42.5378
      WE|     .13881***       .02710        5.122   .0000      12.2869
    HHRS|    -.00019**      .801461D-04    -2.359   .0183      2267.27
      HA|    -.00526          .01285        -.410   .6821      45.1208
      HE|    -.06136***       .02058       -2.982   .0029      12.4914
  FAMINC|     .00997**        .00435        2.289   .0221      23.0806
    KIDS|    -.34017***       .12556       -2.709   .0067       .69588
--------+-------------------------------------------------------------
Binary Logit Model for Binary Choice
--------+-------------------------------------------------------------
        |Characteristics in numerator of Prob[Y = 1]
Constant|    1.24556          .84987        1.466   .1428
      WA|    -.03289          .02134       -1.542   .1232      42.5378
      WE|     .22584***       .04504        5.014   .0000      12.2869
    HHRS|    -.00030**        .00013       -2.326   .0200      2267.27
      HA|    -.00856          .02098        -.408   .6834      45.1208
      HE|    -.10096***       .03381       -2.986   .0028      12.4914
  FAMINC|     .01727**        .00752        2.298   .0215      23.0806
    KIDS|    -.54990***       .20416       -2.693   .0071       .69588
--------+-------------------------------------------------------------
```

Partial Effects

```
--------------------------------------------------------------------
Partial derivatives of probabilities with
respect to the vector of characteristics.
They are computed at the means of the Xs.
Observations used are All Obs.
--------+-----------------------------------------------------------
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]  Elasticity
--------+-----------------------------------------------------------
        |PROBIT:  Index function for probability
      WA|    -.00788          .00512       -1.538   .1240      -.58479
      WE|     .05445***       .01062        5.127   .0000      1.16790
    HHRS|-.74164D-04**      .314375D-04    -2.359   .0183      -.29353
      HA|    -.00206          .00504        -.410   .6821      -.16263
      HE|    -.02407***       .00807       -2.983   .0029      -.52488
  FAMINC|     .00391**        .00171        2.289   .0221       .15753
        |Marginal effect for dummy variable is P|1 - P|0.
    KIDS|    -.13093***       .04708       -2.781   .0054      -.15905
Variable| Coefficient    Standard Error  b/St.Er. P[|Z|>z]  Elasticity
--------+-----------------------------------------------------------
        |LOGIT:  Marginal effect for variable in probability
      WA|    -.00804          .00521       -1.542   .1231      -.59546
      WE|     .05521***       .01099        5.023   .0000      1.18097
    HHRS|-.74419D-04**      .319831D-04    -2.327   .0200      -.29375
      HA|    -.00209          .00513        -.408   .6834      -.16434
      HE|    -.02468***       .00826       -2.988   .0028      -.53673
  FAMINC|     .00422**        .00184        2.301   .0214       .16966
        |Marginal effect for dummy variable is P|1 - P|0.
    KIDS|    -.13120***       .04709       -2.786   .0053      -.15894
--------+-----------------------------------------------------------
```

Testing Hypotheses – A Trinity of Tests

The likelihood ratio test:

- Based on the proposition (Greene's) that restrictions always "make life worse"
- Is the reduction in the criterion (log-likelihood) large? Leads to the LR test.

The Wald test: The usual.

The Lagrange multiplier test:

- Underlying basis: Reexamine the first order conditions.
- Form a test of whether the gradient is significantly "nonzero" at the restricted estimator.

Testing Hypotheses

Wald tests, using the familiar distance measure

Likelihood ratio tests:

LogL_U = log likelihood without restrictions

LogL_R = log likelihood with restrictions

$\text{LogL}_U > \text{logL}_R$ for any nested restrictions

$2(\text{LogL}_U - \text{logL}_R) \rightarrow$ chi-squared [J]

Estimating the Tobit Model

Log likelihood for the tobit model for estimation of $\boldsymbol{\beta}$ and σ:

$$\log L=\sum_{i=1}^{n}\left[(1\text{-}d_i)\log\Phi\left(\frac{-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)+d_i\log\left(\frac{1}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right)\right]$$

$d_i = 1$ if $y_i > 0$, 0 if $y_i = 0$. Derivatives are very complicated, Hessian is nightmarish. Consider the Olsen transformation*:

$\theta=1/\sigma$, $\gamma=-\boldsymbol{\beta}/\sigma$. (One to one; $\sigma=1/\theta$, $\boldsymbol{\beta} = -\gamma/\theta$.)

$$\log L=\sum_{i=1}^{n}\log\left[(1\text{-}d_i)\log\Phi\left(\mathbf{x}_i'\gamma\right)+d_i\log\left(\theta\phi\left(\theta y_i+\mathbf{x}_i'\gamma\right)\right)\right]$$

$$\sum_{i=1}^{n}\log\left[(1\text{-}d_i)\log\Phi\left(\mathbf{x}_i'\gamma\right)+d_i(\log\theta+(1/2)\log 2\pi-(1/2)\left(\theta y_i+\mathbf{x}_i'\gamma\right)^2)\right]$$

$$\frac{\partial\log L}{\partial\gamma}=\sum_{i=1}^{n}\left[(1\text{-}d_i)\frac{\phi\left(\mathbf{x}_i'\gamma\right)}{\Phi\left(\mathbf{x}_i'\gamma\right)}-d_ie_i\right]\mathbf{x}_i$$

$$\frac{\partial\log L}{\partial\theta}=\sum_{i=1}^{n}d_i\left(\frac{1}{\theta}-e_iy_i\right)$$

*Note on the Uniqueness of the MLE in the Tobit Model," Econometrica, 1978.