CHAPTER 13
Limited Dependent Variables

13.1 Introduction

In labor economics, one is faced with explaining the decision to participate in the labor force, the decision to join a union, or the decision to migrate from one region to the other. In finance, a consumer defaults on a loan or a credit card debt, or purchases a stock or an asset like a house or a car. In these examples, the dependent variable is usually a dummy variable with values 1 if the worker participates (or consumer defaults on a loan) and 0 if he or she does not participate (or default). We dealt with dummy variables as explanatory variables on the right hand side of the regression, but what additional problems arise when this dummy variable appears on the left hand side of the equation? As we have done in previous chapters, we first study its effects on the usual least squares estimator, and then consider alternative estimators that are more appropriate for models of this nature.

13.2 The Linear Probability Model

What is wrong with running OLS on this model? After all, it is a feasible procedure. For the labor force participation example one regresses the dummy variable for participation on age, sex, race, marital status, number of children, experience and education, etc. The prediction from this OLS regression is interpreted as the likelihood of participating in the labor force. The problems with this interpretation are the following:

(i) We are predicting probabilities of participation for each individual, whereas the actual values observed are 0 or 1.

(ii) There is no guarantee that $\widehat{y}_i$, the predicted value of y_i is going to be between 0 and 1. In fact, one can always find values of the explanatory variables that would generate a corresponding prediction outside the $(0, 1)$ range.

(iii) Even if one is willing to assume that the true model is a linear regression given by

$$y_i = x_i'\beta + u_i \qquad i = 1, 2, \ldots, n. \tag{13.1}$$

what properties does this entail on the disturbances? It is obvious that $y_i = 1$ only when $u_i = 1 - x_i'\beta$, let us say with probability π_i, where π_i is to be determined. Then $y_i = 0$ only when $u_i = -x_i'\beta$ with probability $(1 - \pi_i)$. For the disturbances to have zero mean

$$E(u_i) = \pi_i(1 - x_i'\beta) + (1 - \pi_i)(-x_i'\beta) = 0 \tag{13.2}$$

Solving for π_i, one gets that $\pi_i = x_i'\beta$. This also means that

$$\text{var}(u_i) = \pi_i(1 - \pi_i) = x_i'\beta(1 - x_i'\beta) \tag{13.3}$$

which is heteroskedastic. Goldberger (1964) suggests correcting for this heteroskedasticity by first running OLS to estimate β, and estimating $\sigma_i^2 = \text{var}(u_i)$ by $\widehat{\sigma}_i^2 = x_i'\widehat{\beta}_{OLS}(1 - x_i'\widehat{\beta}_{OLS}) =$

B.H. Baltagi, *Econometrics*, Springer Texts in Business and Economics, DOI 10.1007/978-3-642-20059-5_13,

$\widehat{y}_i(1-\widehat{y}_i)$. In the next step a *Weighted Least Squares* (WLS) procedure is run on (13.1) with the original observations divided by $\widehat{\sigma}_i$. One cannot compute $\widehat{\sigma}_i$ if OLS predicts $\widehat{y}_i$ larger than 1 or smaller than 0. Suggestions in the literature include substituting 0.005 instead of $\widehat{y}_i < 0$, and 0.995 for $\widehat{y}_i > 1$. However, these procedures do not perform well, and the WLS predictions themselves are not guaranteed to fall in the $(0,1)$ range. Therefore, one should use the robust White heteroskedastic variance-covariance matrix option when estimating linear probability models, otherwise the standard errors are biased and inference is misleading.

This brings us to the fundamental problem with OLS, i.e., its functional form. We are trying to predict

$$y_i = F(x_i'\beta) + u_i \tag{13.4}$$

with a linear regression equation, see Figure 13.1, where the more reasonable functional form for this probability is an S-shaped cumulative distribution functional form. This was justified in the biometrics literature as follows: An insect has a tolerance to an insecticide I_i^*, which is an unobserved random variable with *cumulative distribution function* (c.d.f.) F. If the dosage of insecticide administered induces a stimulus I_i that exceeds I_i^*, the insect dies, i.e., $y_i = 1$. Therefore

$$\Pr(y_i = 1) = \Pr(I_i^* \leq I_i) = F(I_i) \tag{13.5}$$

To put it in an economic context, I_i^* could be the unobserved reservation wage of a worker, and if we increase the offered wage beyond that reservation wage, the worker participates in the labor force. In general, I_i could be represented as a function of the individuals characteristics, i.e., the x_i's. $F(x_i'\beta)$ is by definition between zero and 1 for all values of x_i. Also, the linear probability model yields the result that $\partial\pi_i/\partial x_k = \beta_k$, for every i. This means that the probability of participating (π_i) always changes at the same rate with respect to unit increases in the offer wage x_k. However, this probability model gives

$$\partial\pi_i/\partial x_k = [\partial F(z_i)/\partial z_i] \cdot [\partial z_i/\partial x_k] = f(x_i'\beta) \cdot \beta_k \tag{13.6}$$

where $z_i = x_i'\beta$, and f is the *probability density function* (p.d.f.). Equation (13.6) makes more sense because if x_k denotes the offered wage, changing the probability of participation π_i from 0.96 to 0.97 requires a larger change in x_k than changing π_i from 0.23 to 0.24.

If $F(x_i'\beta)$ is the true probability function, assuming it is linear introduces misspecification, and as Figure 13.1 indicates, for $x_i < x_\ell$, all the u_i's generated by a linear probability approximation are positive. Similarly for all $x_i > x_u$, all the u_i's generated by a linear probability approximation are negative.

13.3 Functional Form: Logit and Probit

Having pointed out the problems with considering the functional form F as linear, we turn to two popular functional forms of F, the *logit* and the *probit*. These two c.d.f.'s differ only in the tails, and the logit resembles the c.d.f. of a t-distribution with 7 degrees of freedom, whereas the probit is the normal c.d.f., or that of a t with ∞ degrees of freedom. Therefore, these two forms will give similar predictions unless there are an extreme number of observations in the tails.

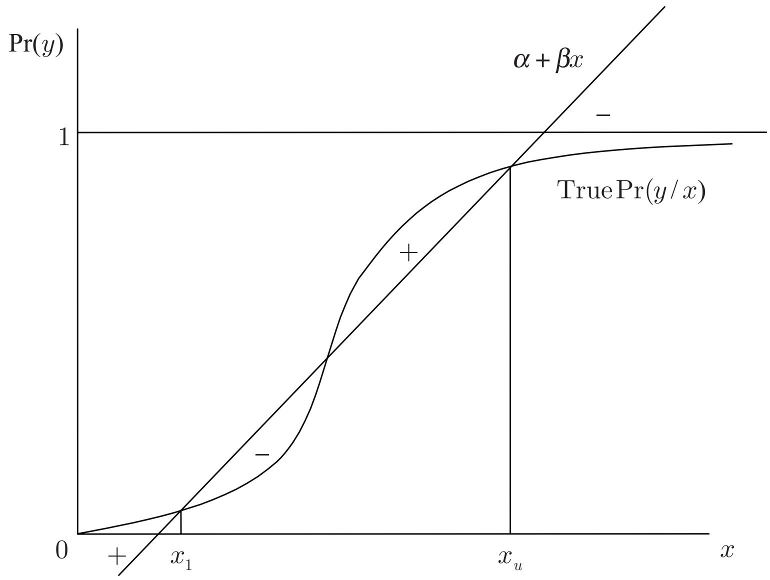

Figure 13.1 Linear Probability Model

We will use the conventional notation $\Phi(z) = \int_{-\infty}^{z} \phi(u)du$, where $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ for $-\infty < z < \infty$, for the *probit*. Also, $\Lambda(z) = e^z/(1+e^z) = 1/(1+e^{-z})$ for $-\infty < z < +\infty$, for the *logit*. Some results that we will use quite often in our derivations are the following: $d\Phi/dz = \phi$, and $d\Lambda/dz = \Lambda(1-\Lambda)$. The p.d.f. of the logistic distribution is the product of its c.d.f. and one minus this c.d.f. Therefore, the marginal effects considered above for a general F are respectively,

$$\partial\Phi(x_i'\beta)/\partial x_k = \phi_i\beta_k \tag{13.7}$$

and

$$\partial\Lambda(x_i'\beta)/\partial x_k = \Lambda_i(1-\Lambda_i)\beta_k \tag{13.8}$$

where $\phi_i = \phi(x_i'\beta)$ and $\Lambda_i = \Lambda(x_i'\beta)$.

One has to be careful with the computation of partial derivatives in case there is a dummy variable among the explanatory variables. For such models, one should compute the marginal effects of a change in one unit of a continuous variable x_k for both values of the dummy variable.

Illustrative Example: Using the *probit* model, suppose that the probability of joining a union is estimated as follows: $\widehat{\pi}_i = \Phi(2.5 - 0.06\ WKS_i + 0.95\ OCC_i)$ where WKS is the number of weeks worked and $OCC = 1$, if the individual is in a blue-collar occupation, and zero otherwise. Weeks worked in this sample range from 20 to 50. From (13.7), the marginal effect of one extra week of work on the probability of joining the union is given by:

For Blue-Collar Workers:

$$\begin{aligned}&-0.06\ \phi[2.5-0.06\ WKS+0.95]\\ &=-0.06\ \phi[2.25]=-0.002 \text{ at } WKS=20\\ &=-0.06\ \phi[1.35]=-0.010 \text{ at } WKS=35\\ &=-0.06\ \phi[0.45]=-0.022 \text{ at } WKS=50\end{aligned}$$

For Non Blue-Collar Workers:

$$\begin{aligned}&-0.06\ \phi[2.5-0.06\ WKS]\\ &=-0.06\ \phi[1.3]=-0.010 \text{ at } WKS=20\\ &=-0.06\ \phi[0.4]=-0.022 \text{ at } WKS=35\\ &=-0.06\ \phi[-0.5]=-0.021 \text{ at } WKS=50\end{aligned}$$

Note how different these marginal effects are for blue-collar versus non blue-collar workers even for the same weeks worked. Increasing weeks worked from 20 to 21 reduces the probability of joining the union by 0.002 for a Blue-Collar worker. This is compared to five times that amount for a Non Blue-Collar worker.

13.4 Grouped Data

In the biometrics literature, grouped data is very likely from laboratory experiments, see Cox (1970). In the insecticide example, every dosage level x_i is administered to a group of insects of size n_i, and the proportion of insects that die are recorded (p_i). This is done for $i = 1, 2, \ldots, M$ dosage levels.

$$P[y_i = 1] = \pi_i = P[I_i^* \leq I_i] = \Phi(\alpha + \beta x_i)$$

where I_i^* is the tolerance and $I_i = \alpha + \beta x_i$ is the stimulus. In economics, observations may be grouped by income levels or age and we observe the labor participation rate for each income or age group. For this type of grouped data, we estimate the probability of participating in the labor force π_i with p_i, the proportion from the sample. This requires a large number of observations in each group, i.e., a large n_i for $i = 1, 2, \ldots, M$. In this case, the approximation is

$$z_i = \Phi^{-1}(p_i) \cong \alpha + \beta x_i \tag{13.9}$$

for each p_i, we compute the standardized normal variates, the z_i's, and we have an estimate of $\alpha + \beta x_i$. Note that the *standard* normal distribution assumption is not restrictive in the sense that if I_i^* is $N(\mu, \sigma^2)$ rather than $N(0, 1)$, then one standardizes the $P[I_i^* \leq I_i]$ by subtracting μ and dividing by σ, in which case the new I_i^* is $N(0, 1)$ and the new α is $(\alpha - \mu)/\sigma$, whereas the new β is β/σ. This also implies that μ and σ are not separately estimable. A plot of the z_i's versus the x_i's would give estimates of α and β. For the biometrics example, one can compute $LD50$, which is the dosage level that will kill 50% of the insect population. This corresponds to $z_i = 0$, which solves for $x_i = -\widehat{\alpha}/\widehat{\beta}$. Similarly, $LD95$ corresponds to $z_i = 1.645$, which solves for $x_i = (1.645 - \widehat{\alpha})/\widehat{\beta}$. Alternatively, for the economic example, $LD50$ is the minimum reservation wage that is necessary for a 50% labor participation rate.

One could improve on this method by including more x's on the right hand side of (13.9). In this case, one can no longer plot the z_i values versus the x variables. However, one can run

OLS of z on these x's. One problem remains, OLS ignores the heteroskedasticity in the error term. To see this:

$$p_i = \pi_i + \epsilon_i = F(x_i'\beta) + \epsilon_i \tag{13.10}$$

where F is a general c.d.f. and $\pi_i = F(x_i'\beta)$. Using the properties of the binomial distribution, $E(p_i) = \pi_i$ and $\text{var}(p_i) = \pi_i(1-\pi_i)/n_i$. Defining $z_i = F^{-1}(p_i)$, we obtain from (13.10)

$$z_i = F^{-1}(p_i) = F^{-1}(\pi_i + \epsilon_i) \cong F^{-1}(\pi_i) + [dF^{-1}(\pi_i)/d\pi_i]\epsilon_i \tag{13.11}$$

where the approximation $\cong$ is a Taylor series expansion around π_i with $\epsilon_i \to 0$. Since F is monotonic $\pi_i = F(F^{-1}(\pi_i))$. Let $w_i = F^{-1}(\pi_i) = x_i'\beta$, differentiating with respect to π gives

$$1 = [dF(w_i)/dw_i]dw_i/d\pi_i \tag{13.12}$$

Alternatively, this can be rewritten as

$$dF^{-1}(\pi_i)/d\pi_i = dw_i/d\pi_i = 1/\{dF(w_i)/dw_i\} = 1/f(w_i) = 1/f(x_i'\beta) \tag{13.13}$$

where f is the probability density function corresponding to F. Using (13.13), equation (13.11) can be rewritten as

$$\begin{aligned} z_i &= F^{-1}(p_i) \cong F^{-1}(\pi_i) + \epsilon_i/f(x_i'\beta) \\ &= F^{-1}(F(x_i'\beta)) + \epsilon_i/f(x_i'\beta) = x_i'\beta + \epsilon_i/f(x_i'\beta) \end{aligned} \tag{13.14}$$

From (13.14), it is clear that the regression disturbances of z_i on x_i are given by $u_i \cong \epsilon_i/f(x_i'\beta)$, with $E(u_i) = 0$ and $\sigma_i^2 = \text{var}(u_i) = \text{var}(\epsilon_i)/f^2(x_i'\beta) = \pi_i(1-\pi_i)/(n_i f_i^2) = F_i(1-F_i)/(n_i f_i^2)$ since $\pi_i = F_i$ where the subscript i on f or F denotes that the argument of that function is $x_i'\beta$. This heteroskedasticity in the disturbances renders OLS on (13.14) consistent but inefficient. For the *probit*, $\sigma_i^2 = \Phi_i(1-\Phi_i)/(n_i\phi_i^2)$, and for the *logit*, $\sigma_i^2 = 1/[n_i\Lambda_i(1-\Lambda_i)]$, since $f_i = \Lambda_i(1-\Lambda_i)$. Using $1/\sigma_i$ as weights, a WLS procedure can be performed on (13.14). Note that $F^{-1}(p)$ for the *logit* is simply $\log[p/(1-p)]$. This is one more reason why the logistic functional form is so popular. In this case one regresses $\log[p/(1-p)]$ on x correcting for heteroskedasticity using WLS. This procedure is also known as the minimum logit chi-square method and is due to Berkson (1953).

In order to obtain feasible estimates of the σ_i's, one could use the OLS estimates of β from (13.14), to estimate the weights. Greene (1993) argues that one should *not* use the proportions p_i's as estimates for the π_i's because this is equivalent to using the y_i^2's instead of σ_i^2 in the heteroskedastic regression. These will lead to inefficient estimates. If OLS on (13.14) is reported one should use the robust White heteroskedastic variance-covariance option, otherwise the standard errors are biased and inference is misleading.

Example 1: Beer Taxes and Motor Vehicle Fatality. Ruhm (1996) used grouped logit analysis with fixed time and state effects to study the impact of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates. Ruhm collected panel data for 48 states (excluding Alaska, Hawaii and the District of Columbia) over the period 1982-1988. The dependent variable is a proportion p, denoting the total vehicle fatality rate per capita for state i at time t. One can perform the inverse logit transformation, $\log[p/(1-p)]$, provided p is not zero or one, and run the usual fixed effects regression described in Chapter 12. Denote this dependent variable by (LFVR). The explanatory variables included the real beer tax rate on 24 (12 oz.) containers of beer (BEERTAX), the minimum legal drinking age (MLDA) in years, the percentage of the population living in dry counties (DRY), the average number of vehicle miles per person aged 16 and over (VMILES), and the percentage of young drivers (15-24 years old) (YNGDRV). Also some dummy variables indicating the presence of alcohol regulations. These include BREATH test laws which is a dummy variable that takes the value 1 if the state authorized the police to administer pre-arrest breath test to establish probable cause for driving under the influence (DUI). JAILD which takes the value of 1 if the state passed legislation mandating jail or community service (COMSERD) for the first DUI conviction. Other variables included are the unemployment rate, real per capita income, and state and time dummy variables. Details on these variables are given in Table 1 of Ruhm (1996). Some of the variables in this data set can be downloaded from the Stock and Watson (2003) web site at www.aw.com/stock_watson. Table 13.1 replicate to the extent possible the grouped logit regression results in column (d) of Table 2, p. 444 of Ruhm (1996). This regression does not include some of the other alcohol regulations that were *not* provided in the data set. These were replicated using the robust White cross-section option in EViews.

Table 13.1 Grouped Logit, Beer Tax and Motor Vehicle Fatality

Dependent Variable: LVFR
Method: Panel Least Squares
Sample: 1982 1988
Cross-sections included: 48
Total panel (unbalanced) observations: 335
White cross-section standard errors & covariance (d.f. corrected)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	–9.361589	0.132383	–70.71570	0.0000
BEERTAX	–0.183533	0.077658	–2.363344	0.0188
MLDA	–0.004465	0.007814	–0.571427	0.5682
DRY	0.008677	0.002611	3.323523	0.0010
YNGDRV	0.493472	0.450802	1.094651	0.2746
VMILES	6.91E-06	4.23E-06	1.632849	0.1037
BREATH	–0.015930	0.027952	–0.569893	0.5692
JAILD	–0.012623	0.038973	–0.323888	0.7463
COMSERD	0.020238	0.024505	0.825867	0.4096
PERINC	0.060305	0.010275	5.868945	0.0000

Effects Specification

Cross-section fixed (dummy variables)
Period fixed (dummy variables)

R-squared	0.931852	Mean dependent var	–8.534768
Adjusted R-squared	0.916318	S.D. dependent var	0.276471
S.E. of regression	0.079977	Akaike info criterion	–2.046359
Sum squared resid	1.739808	Schwarz criterion	–1.329075
Log likelihood	405.7652	F-statistic	59.98854
Durbin-Watson stat	1.433777	Prob(F-statistic)	0.000000

Table 13.1 shows that the beer tax is negative and significant, while the minimum legal drinking age is not significant. Neither is the breath test law, JAILD or COMSERD variables, all of which represent state alcohol safety related legislation. Income per capita and the percentage of the population living in dry counties have a positive and significant effect on motor vehicle fatality rates. The state dummy variables are jointly significant with an observed F-value of 34.9 which is distributed as $F(47, 272)$. The year dummies are jointly significant with an observed F-value of 2.97 which is distributed as $F(6, 272)$. Problem 12 asks the reader to replicate Table 13.1. These results imply that increasing the minimum legal drinking age, or imposing stiffer punishments like mandating jail or community service are not effective policy tools for decreasing traffic related deaths. However, increasing the real tax on beer is an effective policy for reducing traffic related deaths.

For grouped data, the sample sizes n_i for each group have to be sufficiently large. Also, the p_i's cannot be zero or one. One modification suggested in the literature is to add $(1/2n_i)$ to p_i when computing the log of odds ratio, see Cox (1970).

Example 2: Fractional Response. Papke and Wooldridge (1996) argue that in many economic settings p_i may be 0 or 1 for a large number of observations. For example, when studying participation rates in pension plans or when studying high school graduation rates. They propose a fractional logit regression which handles fractional response variables based on quasi-likelihood methods. Fractional response variables are bounded variables. Without loss of generality, they could be restricted to lie between 0 and 1. Examples include the proportion of income spent on charitable contributions, the fraction of total weekly hours spent working. Papke and Wooldridge (1996) propose modeling the $E(y_i/x_i)$ as a logistic function $\Lambda(x_i'\beta)$. This insures that the predicted value of y_i lies in the interval $(0, 1)$. It is also well defined even if y_i takes the values 0 or 1 with positive probability. It is important to note that in case y_i is a proportion from a group of known size n_i, the quasi maximum likelihood method ignores the information on n_i. Using the Bernoulli log-likelihood function, one gets

$$L_i(\beta) = y_i \log[\Lambda(x_i'\beta)] + (1 - y_i) \log[1 - \Lambda(x_i'\beta)]$$

for $i = 1, 2, \ldots, n$, with $0 < \Lambda(x_i'\beta) < 1$.

Maximizing $\sum_{i=1}^{n} L_i(\beta)$ with respect to β yields the quasi-MLE which is consistent and $\sqrt{n}$ asymptotically normal *regardless* of the distribution of y_i conditional on x_i, see Gourieroux, Monfort and Trognon (1984) and McCullagh and Nelder (1989). The latter proposed the generalized linear models (GLM) approach to this problem in statistics. Logit QMLE can be done in Stata using the GLM command with the Binary family function indicating Bernoulli and the Link function indicating the logistic distribution.

Papke and Wooldridge (1996) derive robust asymptotic variance of the QMLE of β and suggest some specification tests based on Wooldridge (1991). They apply their methods to the participation in 401(K) pension plans. The data are from the 1987 IRS Form 5500 reports of pension plans with more than 100 participants. This data set containing 4734 observations can be downloaded from the *Journal of Applied Econometrics* Data Archive. We focus on a subset of their data which includes 3874 observations of plans with match rates less than or equal to one. Match rates above one may be indicating end-of-plan year employer contributions made to avoid IRS disqualification. Participation rates (PRATE) in this sample are high

Table 13.2 Logit Quasi-MLE of Participation Rates in 401(K) Plan

```
glm prate mrate log_emp log_emp2 age age2 sole if one==1, f(bin) l(logit) robust
note: prate has non-integer values

Iteration 0:          log pseudo-likelihood = −1200.8698
Iteration 1:          log pseudo-likelihood = −1179.3843
Iteration 2:          log pseudo-likelihood = −1179.2785
Iteration 3:          log pseudo-likelihood = −1179.2785

Generalized linear models                          Number of obs      =      3784
Optimization          : ML: Newton-Raphson         Residual df        =      3777
                                                   Scale parameter    =         1
Deviance              = 1273.60684                 (1/df) Deviance    =  .3372006
Pearson               = 724.4199889                (1/df) Pearson     =  .1917977

Variance function     : V(u) = u*(1-u)             [Bernoulli]
Link function         : g(u) = ln(u/(1-u))         [Logit]
Standard errors       : Sandwich

Log pseudo-likelihood = −1179.278516
BIC                   = −29843.34715               AIC                =  .6269971
```

prate	Coef.	Robust Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
mrate	1.39008	.1077064	12.91	0.000	1.17898	1.601181
log_emp	−1.001874	.1104365	−9.07	0.000	−1.218326	−.7854229
log_emp2	.0521864	.0071278	7.32	0.000	.0382161	.0661568
age	.0501126	.0088451	5.67	0.000	.0327766	.0674486
age2	−.0005154	.0002117	−2.43	0.015	−.0009303	−.0001004
sole	.0079469	.0502025	0.16	0.874	−.0904482	.1063421
_cons	5.057997	.4208646	12.02	0.000	4.233117	5.882876

averaging 84.8%. Over 40% of the plans have a participation proportion of one. This makes the log-odds ratio approach awkward since adjustments have to be made to more than 40% of the observations. The plan match rate (MRATE) averages about 41 cents on the dollar. Other explanatory variables include total firm employment (EMP), age of the plan (AGE), a dummy variable (SOLE) which takes the value of 1 if the 401(K) plan is the only pension plan offered by the employer. The 401(K) plans average 12 years in age, they are the SOLE plan in 37% of the sample. The average employment is 4622. Problem 14 asks the reader to replicate the descriptive statistic given in Table I of Papke and Wooldridge (1996, p. 627). Table 13.2 gives the Stata output for logit QMLE using the same specification given in Table II of Papke and Wooldridge (1996, p. 628). Note that it uses the GLM command, the Bernoulli variance function and the logit link function. The results show that there is a positive and significant relationship between match rate and participation rate. All the other variables included are significant except for SOLE. Problem 14 asks the reader to replicate this result and compare with OLS. The latter turns out to have a lower R^2 and fails a RESET test, see Chapter 8.

13.5 Individual Data: Probit and Logit

When the number of observations n_i in each group is small, one cannot obtain reliable estimates of the π_i's with the p_i's. In this case, one should not group the observations, instead these observations should be treated as individual observations and the model estimated by the maximum likelihood procedure. The likelihood is obtained as independent random draws from a Bernoulli distribution with probability of success $\pi_i = F(x_i'\beta) = P[y_i = 1]$. Hence

$$\ell = \prod_{i=1}^{n}[F(x_i'\beta)]^{y_i}[1 - F(x_i'\beta)]^{1-y_i} \tag{13.15}$$

and the log-likelihood

$$\log\ell = \sum_{i=1}^{n}\{y_i \log F(x_i'\beta) + (1 - y_i)\log[1 - F(x_i'\beta)]\} \tag{13.16}$$

The first-order conditions for maximization require the score $S(\beta) = \partial\log\ell/\partial\beta$ to be zero:

$$\begin{aligned} S(\beta) &= \partial\log\ell/\partial\beta = \sum_{i=1}^{n}\{[f_i y_i/F_i] - (1 - y_i)[f_i/(1 - F_i)]\}x_i \\ &= \sum_{i=1}^{n}(y_i - F_i)f_i x_i/[F_i(1 - F_i)] = 0 \end{aligned} \tag{13.17}$$

where the subscript i on f or F denotes that the argument of that function is $x_i'\beta$. For the logit model (13.17) reduces to

$$S(\beta) = \sum_{i=1}^{n}(y_i - \Lambda_i)x_i = 0 \qquad \text{since} \qquad f_i = \Lambda_i(1 - \Lambda_i) \tag{13.18}$$

If there is a constant in the model, the solution to (13.18) for $x_i = 1$ implies that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{\Lambda}_i$. This means that the number of participants in the sample, i.e., those with $y_i = 1$, will always be equal to the predicted number of participants from the logit model. Similarly, if x_i is a dummy variable which is 1 if the individual is male and zero if the individual is female, then (13.18) states that the predicted frequency is equal to the actual frequency for males and females. Note that (13.18) resembles the OLS normal equations if we interpret $(y_i - \widehat{\Lambda}_i)$ as residuals. For the probit model (13.17) reduces to

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n}(y_i - \Phi_i)\phi_i x_i/[\Phi_i(1 - \Phi_i)] \\ &= \sum_{y_i=0}\lambda_{oi}x_i + \sum_{y_i=1}\lambda_{1i}x_i = 0 \end{aligned} \tag{13.19}$$

where $\lambda_{oi} = -\phi_i/[1 - \Phi_i]$ for $y_i = 0$ and $\lambda_{1i} = \phi_i/\Phi_i$ for $y_i = 1$. Also, $\Sigma_{y_i=0}$ denotes the sum over all zero values of y_i. These λ_i's are thought of as *generalized residuals* which are orthogonal to x_i. Note that unlike the logit, the probit does not necessarily predict the number of participants to be exactly equal to the number of ones in the sample.

Equations (13.17) are highly nonlinear and may be solved using the scoring method, i.e., starting with some initial value β_o we revise this estimate as follows:

$$\beta_1 = \beta_o + [I^{-1}(\beta_o)]S(\beta_o) \tag{13.20}$$

where $S(\beta) = \partial\log\ell/\partial\beta$ and $I(\beta) = E[-\partial^2\log\ell/\partial\beta\partial\beta']$. This process is repeated until convergence. For the logit and probit models, $\log F(x_i'\beta)$ and $\log[1 - F(x_i'\beta)]$ are concave. Hence, the log-likelihood function given by (13.16) is *globally concave*, see Pratt (1981). Hence, for both the logit and probit, $[\partial^2\log\ell/\partial\beta\partial\beta']$ is negative definite for all values of β and the iterative procedure will converge to the unique maximum likelihood estimate $\widehat{\beta}_{MLE}$ no matter what starting values we use. In this case, the asymptotic covariance matrix of $\widehat{\beta}_{MLE}$ is estimated by $I^{-1}(\widehat{\beta}_{MLE})$ from the last iteration.

Amemiya (1981, p. 1495) derived $I(\beta)$ by differentiating (13.17), multiplying by a negative sign and taking the expected value, the result is given by:

$$I(\beta) = -E[\partial^2 \log\ell/\partial\beta\partial\beta'] = \sum_{i=1}^n f_i^2 x_i x_i'/F_i(1-F_i) \tag{13.21}$$

For the logit, (13.21) reduces to

$$I(\beta) = \sum_{i=1}^n \Lambda_i(1-\Lambda_i)x_i x_i' \tag{13.22}$$

For the probit, (13.21) reduces to

$$I(\beta) = \sum_{i=1}^n \phi_i^2 x_i x_i'/\Phi_i(1-\Phi_i) \tag{13.23}$$

Alternative maximization may use the Newton-Raphson iterative procedure which uses the Hessian itself rather than its expected value in (13.20), i.e., $I(\beta)$ is replaced by $H(\beta) = [-\partial^2\log\ell/\partial\beta\partial\beta']$. For the logit model, $H(\beta) = I(\beta)$ and is given in (13.22). For the probit model, $H(\beta) = \sum_{i=1}^n [\lambda_i^2 + \lambda_i x_i'\beta]x_i x_i'$ which is different from (13.23). Note that $\lambda_i = \lambda_{oi}$ if $y_i = 0$; and $\lambda_i = \lambda_{1i}$ if $y_i = 1$. These were defined below (13.19).

A third method, suggested by Berndt, Hall, Hall and Hausman (1974) uses the outer product of the first derivatives in place of $I(\beta)$, i.e., $G(\beta) = S(\beta)S'(\beta)$. For the logit model, this is $G(\beta) = \sum_{i=1}^n (y_i - \Lambda_i)^2 x_i x_i'$. For the probit model, $G(\beta) = \sum_{i=1}^n \lambda_i^2 x_i x_i'$. As in the method of scoring, one iterates starting from initial estimates β_o, and the asymptotic variance-covariance matrix is estimated from the inverse of $G(\widehat{\beta})$, $H(\widehat{\beta})$ or $I(\widehat{\beta})$ in the last iteration.

Test of hypotheses can be carried out from the asymptotic standard errors using t-statistics. For $R\beta = r$ type restrictions, the usual Wald test $W = (R\widehat{\beta} - r)'[RV(\widehat{\beta})R']^{-1}(R\widehat{\beta} - r)$ can be used with $V(\widehat{\beta})$ obtained from the last iteration as described above. Likelihood ratio and Lagrange Multiplier statistics can also be computed. $LR = -2[\log\ell_{restricted} - \log\ell_{unrestricted}]$, whereas, the Lagrange Multiplier statistic is $LM = S'(\beta)V(\beta)S(\beta)$, where $S(\beta)$ is the score evaluated at the restricted estimator. Davidson and MacKinnon (1984) suggest that $V(\widehat{\beta})$ based on $I(\widehat{\beta})$ is the best of the three estimators to use. In fact, Monte Carlo experiments show that the estimate of $V(\widehat{\beta})$ based on the outer product of the first derivatives usually performs the worst and is not recommended in practice. All three statistics are asymptotically equivalent and are asymptotically distributed as χ_q^2 where q is the number of restrictions. The next section discusses tests of hypotheses using an artificial regression.

13.6 The Binary Response Model Regression[1]

Davidson and MacKinnon (1984) suggest a modified version of the Gauss-Newton regression (GNR) considered in Chapter 8 which is useful in the context of a binary response model described in (13.5).[2] In fact, we have shown that this model can be written as a nonlinear regression

$$y_i = F(x_i'\beta) + u_i \tag{13.24}$$

with u_i having zero mean and $\text{var}(u_i) = F_i(1-F_i)$. The GNR ignoring heteroskedasticity yields

$$(y_i - F_i) = f_i x_i' b + \text{residual}$$

where b is the regression estimates when we regress $(y_i - F_i)$ on $f_i x_i'$.

Correcting for heteroskedasticity by dividing each observation by its standard deviation we get the *Binary Response Model Regression* (BRMR):

$$\frac{(y_i - F_i)}{\sqrt{F_i(1-F_i)}} = \frac{f_i}{\sqrt{F_i(1-F_i)}}\, x_i' b + \text{residual} \tag{13.25}$$

For the logit model with $f_i = \Lambda_i(1-\Lambda_i)$, this simplifies further to

$$\frac{y_i - \Lambda_i}{\sqrt{f_i}} = \sqrt{f_i}\, x_i' b + \text{residual} \tag{13.26}$$

For the probit model, the BRMR is given by

$$\frac{y_i - \Phi_i}{\sqrt{\Phi_i(1-\Phi_i)}} = \frac{\phi_i}{\sqrt{\Phi_i(1-\Phi_i)}}\, x_i' b + \text{residual} \tag{13.27}$$

Like the GNR considered in Chapter 8, the BRMR given in (13.25) can be used for obtaining parameter and covariance matrix estimates as well as test of hypotheses. In fact, Davidson and MacKinnon point out that the transpose of the dependent variable in (13.25) times the matrix of regressors in (13.25) yields a vector whose typical element is exactly that of $S(\beta)$ given in (13.17). Also, the transpose of the matrix of regressors in (13.25) multiplied by itself yields a matrix whose typical element is exactly that of $I(\beta)$ given in (13.21).

Let us consider how the BRMR is used to test hypotheses. Suppose that $\beta' = (\beta_1', \beta_2')$ where β_1 is of dimension $k-r$ and β_2 is of dimension r. We want to test H_o; $\beta_2 = 0$. Let $\widetilde{\beta}' = (\widetilde{\beta}_1, 0)$ be the restricted MLE of β subject to H_o. In order to test H_o, we run the BRMR:

$$\frac{y_i - \widetilde{F}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}} = \frac{\widetilde{f}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}}\, x_{i1}' b_1 + \frac{\widetilde{f}_i}{\sqrt{\widetilde{F}_i(1-\widetilde{F}_i)}}\, x_{i2}' b_2 + \text{residual} \tag{13.28}$$

where $x_i' = (x_{i1}', x_{i2}')$ has been partitioned into vectors conformable with the corresponding partition of β. Also, $\widetilde{F}_i = F(x_i'\widetilde{\beta})$ and $\widetilde{f}_i = f(x_i'\widetilde{\beta})$. The suggested test statistic for H_o is the explained sum of squares of the regression (13.28). This is asymptotically distributed as χ^2_r under H_o.[3] A special case of this BRMR is that of testing the null hypothesis that *all* the slope coefficients are zero. In this case, $x_{i1} = 1$ and β_1 is the constant α. Problem 2 shows that the restricted MLE in this case is $\widetilde{F}(\alpha) = \bar{y}$ or $\widetilde{\alpha} = F^{-1}(\bar{y})$, where $\bar{y}$ is the proportion of the sample with $y_i = 1$. Therefore, the BRMR in (13.25) reduces to

$$\frac{y_i - \bar{y}}{\sqrt{\bar{y}(1-\bar{y})}} = \frac{f_i(\widetilde{\alpha})}{\sqrt{\bar{y}(1-\bar{y})}}\, b_1 + \frac{f_i(\widetilde{\alpha})}{\sqrt{\bar{y}(1-\bar{y})}}\, x_{i2}' b_2 + \text{residual} \tag{13.29}$$

Note that $\bar{y}(1-\bar{y})$ is *constant* for all observations. The test for $b_2 = 0$ is not affected by dividing the dependent variable or the regressors by this constant, nor is it affected by subtracting a constant from the dependent variable. Hence, the test for $b_2 = 0$ can be carried out by regressing y_i on a constant and x_{i2} and testing that the slope coefficients of x_{i2} are zero using the usual least squares F-statistic. This is a simpler alternative to the likelihood ratio test proposed in the previous section and described in the empirical example in section 13.9. For other uses of the BRMR, see Davidson and MacKinnon (1993).

13.7 Asymptotic Variances for Predictions and Marginal Effects

Two results of interest after estimating the model are: the *predictions* $F(x'\widehat{\beta})$ and the *marginal effects* $\partial F/\partial x = f(x'\widehat{\beta})\widehat{\beta}$. For example, given the characteristics of an individual x, we can predict his or her probability of purchasing a car. Also, given a change in x, say income, one can estimate the marginal effect this will have on the probability of purchasing a car. The latter effect is constant for the linear probability model and is given by the regression coefficient of income, whereas for the probit and logit models this marginal effect will vary with the x_i's, see (13.7) and (13.8). These marginal effects can be computed with Stata using the *dprobit* command. The default is to compute them at the sample mean $\bar{x}$. There is also the additional problem of computing variances for these predictions and marginal effects. Both $F(x'\widehat{\beta})$ and $f(x'\widehat{\beta})\widehat{\beta}$ are nonlinear functions of the $\widehat{\beta}$'s. To compute standard errors, we can use the following linear approximation which states that whenever $\widehat{\theta} = F(\widehat{\beta})$ then the asy.var$(\widehat{\theta}) = (\partial F/\partial\widehat{\beta})'V(\widehat{\beta})(\partial F/\partial\widehat{\beta})$. For the predictions, let $z = x'\widehat{\beta}$ and denote by $\widehat{F} = F(x'\widehat{\beta})$ and $\widehat{f} = f(x'\widehat{\beta})$, then

$$\partial\widehat{F}/\partial\widehat{\beta} = (\partial\widehat{F}/\partial z)(\partial z/\partial\widehat{\beta}) = \widehat{f}x \qquad \text{and} \qquad \text{asy.var}(\widehat{F}) = \widehat{f}^2 x'V(\widehat{\beta})x.$$

For the marginal effects, let $\widehat{\gamma} = \widehat{f\beta}$, then

$$\text{asy.var}(\widehat{\gamma}) = (\partial\widehat{\gamma}/\partial\widehat{\beta}')V(\widehat{\beta})(\partial\widehat{\gamma}/\partial\widehat{\beta}')' \tag{13.30}$$

where $\partial\widehat{\gamma}/\partial\widehat{\beta}' = \widehat{f}I_k + \widehat{\beta}(\partial\widehat{f}/\partial z)(\partial z/\partial\widehat{\beta}') = \widehat{f}I_k + (\partial\widehat{f}/\partial z)(\widehat{\beta}x')$.

For the probit model, $\partial\widehat{f}/\partial z = \partial\widehat{\phi}/\partial z = -z\widehat{\phi}$. So, $\partial\widehat{\gamma}/\partial\widehat{\beta}' = \widehat{\phi}[I_k - z\widehat{\beta}x']$ and

$$\text{asy.var}(\widehat{\gamma}) = \widehat{\phi}^2[I_k - x'\widehat{\beta}\widehat{\beta}x']V(\widehat{\beta})[I_k - x'\widehat{\beta}\widehat{\beta}x']' \tag{13.31}$$

For the logit model, $\widehat{f} = \widehat{\Lambda}(1-\widehat{\Lambda})$, so

$$\begin{aligned}\partial\widehat{f}/\partial z &= (1-2\widehat{\Lambda})(\partial\widehat{\Lambda}/\partial z) = (1-2\widehat{\Lambda})(\widehat{f}) = (1-2\widehat{\Lambda})\widehat{\Lambda}(1-\widehat{\Lambda})\\ \partial\widehat{\gamma}/\partial\widehat{\beta}' &= \widehat{\Lambda}(1-\widehat{\Lambda})[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x']\end{aligned}$$

and (13.30) becomes

$$\text{asy.var}(\widehat{\gamma}) = [\widehat{\Lambda}(1-\widehat{\Lambda})]^2[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x']V(\widehat{\beta})[I_k + (1-2\widehat{\Lambda})\widehat{\beta}x']' \tag{13.32}$$

13.8 Goodness of Fit Measures

There are problems with the use of conventional R^2-type measures when the explained variable y takes on two values, see Maddala (1983, pp. 37–41). The predicted values $\widehat{y}$ are probabilities and the actual values of y are either 0 or 1 so the usual R^2 is likely to be very low. Also, if there is a constant in the model the linear probability and logit models satisfy $\sum_{i=1}^n y_i = \sum_{i=1}^n \widehat{y}_i$. However, the probit model does not necessarily satisfy this exact relationship although it is approximately valid.

Several R^2-type measures have been suggested in the literature, some of these are the following:

(i) The squared correlation between y and $\widehat{y}$: $R_1^2 = r_{y,\widehat{y}}^2$.

(ii) Measures based on the *residual sum of squares*: Effron (1978) suggested using

$$R_2^2 = 1 - [\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2] = 1 - [n \sum_{i=1}^n (y_i - \widehat{y}_i)^2 / n_1 n_2]$$

since $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = n_1 - n(n_1/n)^2 = n_1 n_2/n$, where $n_1 = \sum_{i=1}^n y_i$ and $n_2 = n - n_1$.

Amemiya (1981, p. 1504) suggests using $\left[\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \widehat{y}_i(1-\widehat{y}_i)\right]$ as the residual sum of squares. This weights each squared error by the inverse of its variance.

(iii) Measures based on likelihood ratios: $R_3^2 = 1 - (\ell_r/\ell_u)^{2/n}$ where ℓ_r is the restricted likelihood and ℓ_u is the unrestricted likelihood. This tests that all the slope coefficients are zero in the standard linear regression model. For the limited dependent variable model however, the likelihood function has a maximum of 1. This means that $\ell_r \leq \ell_u \leq 1$ or $\ell_r \leq (\ell_r/\ell_u) \leq 1$ or $\ell_r^{2/n} \leq 1 - R_3^2 \leq 1$ or $0 \leq R_3^2 \leq 1 - \ell_r^{2/n}$. Hence, Cragg and Uhler (1970) suggest a pseudo-R^2 that lies between 0 and 1, and is given by $R_4^2 = (\ell_u^{2/n} - \ell_r^{2/n})/[(1 - \ell_r^{2/n})/\ell_u^{2/n}]$. Another measure suggested by McFadden (1974) is $R_5^2 = 1 - (\log\ell_u/\log\ell_r)$.

(iv) Proportion of correct predictions: After computing $\widehat{y}$, one classifies the i-th observation as a success if $\widehat{y}_i > 0.5$, and a failure if $\widehat{y}_i < 0.5$. This measure is useful but may not have enough discriminatory power.

13.9 Empirical Examples

Example 1: Union Participation

To illustrate the logit and probit models, we consider the PSID data for 1982 used in Chapter 4. In this example, we are interested in modelling union participation. Out of the 595 individuals observed in 1982, 218 individuals had their wage set by a union and 377 did not. The explanatory variables used are: years of education (ED), weeks worked (WKS), years of full-time work experience (EXP), occupation ($OCC = 1$, if the individual is in a blue-collar occupation), residence ($SOUTH = 1$, $SMSA = 1$, if the individual resides in the South, or in a standard metropolitan statistical area), industry ($IND = 1$, if the individual works in a manufacturing industry), marital status ($MS = 1$, if the individual is married), sex and race ($FEM = 1$, $BLK = 1$, if the individual is female or black). A full description of the data is given in Cornwell and Rupert (1988). The results of the linear probability, logit and probit models are given in Table 13.3. These were computed using EViews. In fact Table 13.4 gives the probit output. We have already mentioned that the probit model normalizes σ to be 1. But, the logit model has variance $\pi^2/3$. Therefore, the logit estimates tend to be larger than the probit estimates although by a factor less than $\pi/\sqrt{3}$. In order to make the logit results comparable to those of the probit, Amemiya (1981) suggests multiplying the logit coefficient estimates by 0.625.

Similarly, to make the linear probability estimates comparable to those of the probit model one needs to multiply these coefficients by 2.5 and then subtract 1.25 from the constant term. For this example, both logit and probit procedures converged quickly in 4 iterations. The log-likelihood values and McFadden's (1974) R^2 obtained for the last iteration are recorded.

Table 13.3 Comparison of the Linear Probability, Logit and Probit Models: Union Participation*

Variable	**OLS**	**Logit**	**Probit**
EXP	−.005 (1.14)	−.007 (1.15)	−.007 (1.21)
WKS	−.045 (5.21)	−.068 (5.05)	−.061 (5.16)
OCC	.795 (6.85)	1.036 (6.27)	.955 (6.28)
IND	.075 (0.79)	.114 (0.89)	.093 (0.76)
SOUTH	−.425 (4.27)	−.653 (4.33)	−.593 (4.26)
SMSA	.211 (2.20)	.280 (2.05)	.261 (2.03)
MS	.247 (1.55)	.378 (1.66)	.351 (1.62)
FEM	−.272 (1.37)	−.483 (1.58)	−.407 (1.47)
ED	−.040 (1.88)	−.057 (1.85)	−.057 (1.99)
BLK	.125 (0.71)	.222 (0.90)	.226 (0.99)
Const	1.740 (5.27)	2.738 (3.27)	2.517 (3.30)
Log-likelihood		−312.337	−313.380
McFadden's R^2		0.201	0.198
χ^2_{10}		157.2	155.1

* Figures in parentheses are t-statistics

Note that the logit and probit estimates yield similar results in magnitude, sign and significance. One would expect different results from the logit and probit only if there are several observations in the tails. The following variables were insignificant at the 5% level: EXP, IND, MS, FEM and BLK. The results show that union participation is less likely if the individual resides in the South and more likely if he or she resides in a standard metropolitan statistical area. Union participation is also less likely the more the weeks worked and the higher the years of education. Union participation is more likely for blue-collar than non blue-collar occupations. The linear probability model yields different estimates from the logit and probit results. OLS predicts two observations with $\widehat{y}_i > 1$, and 29 observations with $\widehat{y}_i < 0$. Table 13.5 gives the actual versus predicted values of union participation for the linear probability, logit and probit models. The percentage of correct predictions is 75% for the linear probability and probit model and 76% for the logit model.

One can test the significance of all slope coefficients by computing the LR based on the unrestricted log-likelihood value ($\log\ell_u$) reported in Table 13.3, and the restricted log-likelihood value including only the constant. The latter is the same for both the logit and probit models and is given by

$$\log\ell_r = n[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})] \tag{13.33}$$

where $\bar{y}$ is the proportion of the sample with $y_i = 1$, see problem 2. In this example, $\bar{y} = 218/595 = 0.366$ and $n = 595$ with $\log\ell_r = -390.918$. Therefore, for the probit model,

$$LR = -2[\log\ell_r - \log\ell_u] = -2[-390.918 + 313.380] = 155.1$$

which is distributed as χ^2_{10} under the null of zero slope coefficients. This is highly significant and the null is rejected. Similarly, for the logit model this LR statistic is 157.2. For the linear probability model, the same null hypothesis of zero slope coefficients can be tested using a

Table 13.4 Probit Estimates: Union Participation

Dependent Variable:	UNION
Method:	ML – Binary Probit
Sample:	1 595
Included observations:	595
Convergence achieved after 5 iterations	
Covariance matrix computed using second derivatives	

Variable	Coefficient	Std. Error	z-Statistic	Prob.
EX	–0.006932	0.005745	–1.206491	0.2276
WKS	–0.060829	0.011785	–5.161666	0.0000
OCC	0.955490	0.152137	6.280476	0.0000
IND	0.092827	0.122774	0.756085	0.4496
SOUTH	–0.592739	0.139102	–4.261183	0.0000
SMSA	0.260700	0.128630	2.026741	0.0427
MS	0.350520	0.216284	1.620648	0.1051
FEM	–0.407026	0.277038	–1.469203	0.1418
ED	–0.057382	0.028842	–1.989515	0.0466
BLK	0.226482	0.228845	0.989675	0.3223
C	2.516784	0.762612	3.300217	0.0010

Mean dependent var	0.366387	S.D. dependent var	0.482222
S.E. of regression	0.420828	Akaike info criterion	1.090351
Sum squared resid	103.4242	Schwarz criterion	1.171484
Log likelihood	–313.3795	Hannan-Quinn criter.	1.121947
Restr. log likelihood	–390.9177	Avg. log likelihood	–0.526688
LR statistic (10 df)	155.0763	McFadden R-squared	0.198349
Probability(LR stat)	0.000000		
Obs with Dep=0	377	Total obs	595
Obs with Dep=1	218		

Table 13.5 Actual Versus Predicted: Union Participation

		Predicted		Total
		Union = 0	Union = 1	
Actual	Union = 0	OLS = 312 LOGIT = 316 Probit = 314	OLS = 65 LOGIT = 61 Probit = 63	377
	Union = 1	OLS = 83 LOGIT = 82 Probit = 86	OLS = 135 LOGIT = 136 Probit = 132	218
Total		OLS = 395 LOGIT = 398 Probit = 400	OLS = 200 LOGIT = 197 Probit = 195	595

Chow F-statistic. This yields an observed value of 17.80 which is distributed as $F(10, 584)$ under the null hypothesis. Again, the null is soundly rejected. This F-test is in fact the BRMR test considered in section 13.6. As described in section 13.8, McFadden's R^2 is given by $R_5^2 = 1 - [\log\ell_u/\log\ell_r]$ which for the probit model yields

$$R_5^2 = 1 - (313.380/390.918) = 0.198.$$

For the logit model, McFadden's R_5^2 is 0.201.

Example 2: Employment and Problem Drinking

Mullahy and Sindelar (1996) estimate a linear probability model relating employment and measures of problem drinking. The analysis is based on the 1988 Alcohol Supplement of the National Health Interview Survey. This regression was performed for Males and Females separately since the authors argue that women are less likely than men to be alcoholic, are more likely to abstain from consumption, and have lower mean alcohol consumption levels. They also report that women metabolize ethanol faster than do men and experience greater liver damage for the same level of consumption of ethanol. The dependent variable takes the value 1 if the individual was employed in the past two weeks and zero otherwise. The explanatory variables included the 90th percentile of ethanol consumption in the sample (18 oz. for males and 10.8 oz. for females) and zero otherwise. This variables is denoted by hvdrnk90. The state unemployment rate in 1988 (UE88), Age, Age^2, schooling, married, family size, and white. Health status dummies indicating whether the individual's health was excellent, very good, fair. Region of residence, whether the individual resided in the northeast, midwest or south. Also, whether he or she resided in center city (msa1) or other metropolitan statistical area (not center city, msa2). Three additional dummy variables were included for the quarters in which the survey was conducted. Details on the definitions of these variables are given in Table 1 of Mullahy and Sindelar (1996). Table 13.6 gives the probit results based on $n = 9822$ males using Stata. These results show a negative relationship between the 90th percentile alcohol variable and the probability of being employed, but this has a p-value of 0.075. Mullahy and Sindelar find that for both men and women, problem drinking results in reduced employment and increased unemployment. Table 13.7 gives the *marginal effects* computed in Stata using the *mfx* option after *probit* estimation. The marginal effects are computed at the *sample mean* of the variables, except in the case of dummy variables where it is done for a discrete change from 0 to 1. For example, the marginal effect of being a heavy drinker in the upper 90th percentile of ethanol consumption in the sample, (given that all the other variables are evaluated at their mean and dummy variables are changing from 0 to 1), is to decrease the probability of employment by 1.6%. These can also be computed at particular values of the explanatory variables with the option *at* in Stata. In fact Table 13.8 gives the *average marginal effect* for all males. This can be computed using the *margeff* command in Stata. In this case the average marginal effect for a heavy drinker (-.0165) did not change much from the marginal effect computed at the sample mean (-.0162) and neither did the standard error (.0096 compared with .0093). The goodness of fit as measured by how well this probit classifies the predicted probabilities is given in Table 13.9 using the *estat classification* option in Stata. The percentage of correct predictions is 90.79%. Problem 13 asks the reader to verify these results as well as those in the original article by Mullahy and Sindelar (1996).

Table 13.6 Probit Estimates: Employment and Problem Drinking

. probit emp hvdrnk90 ue88 age agesq educ married famsize white hlstat1 hlstat2 hlstat3 hlstat4 region1 region2 region3 msa1 msa2 q1 q2 q3, robust

Probit regression	Number of obs	=	9822
	Wald chi2(20)	=	928.33
	Prob > chi2	=	0.0000
Log pseudolikelihood = −2698.1797	Pseudo R2	=	0.1651

emp	Coef.	Robust Std. Err.	z	$P>\|z\|$	[95% Conf. Interval]	
hvdrnk90	−.1049465	.0589881	−1.78	0.075	−.2205612	.0106681
ue88	−.0532774	.0142025	−3.75	0.000	−.0811137	−.0254411
age	.0996338	.0171185	5.82	0.000	.0660821	.1331855
agesq	−.0013043	.0002051	−6.36	0.000	−.0017062	−.0009023
educ	.0471834	.0066739	7.07	0.000	.0341029	.0602639
married	.2952921	.0540858	5.46	0.000	.189286	.4012982
famsize	.0188906	.0140463	1.34	0.179	−.0086398	.0464209
white	.3945226	.0483381	8.16	0.000	.2997818	.4892634
hlstat1	1.816306	.0983447	18.47	0.000	1.623554	2.009058
hlstat2	1.778434	.0991531	17.94	0.000	1.584098	1.972771
hlstat3	1.547836	.0982637	15.75	0.000	1.355243	1.74043
hlstat4	1.043363	.1077279	9.69	0.000	.8322205	1.254506
region1	.0343123	.0620021	0.55	0.580	−.0872096	.1558341
region2	.0604907	.0537885	1.12	0.261	−.0449327	.1659142
region3	.1821206	.0542346	3.36	0.001	.0758227	.2884185
msa1	−.0730529	.0518719	−1.41	0.159	−.1747199	.0286141
msa2	.0759533	.0513092	1.48	0.139	−.0246109	.1765175
q1	−.1054844	.0527728	−2.00	0.046	−.2089171	−.0020516
q2	−.0513229	.0528185	−0.97	0.331	−.1548453	.0521995
q3	−.0293419	.0543751	−0.54	0.589	−.1359152	.0772313
_cons	−3.017454	.3592321	−8.40	0.000	−3.721536	−2.313372

Example 3: Fertility and Same Sex of Previous Children

Carrasco (2001) estimated a probit equation for fertility using PSID data over the period 1986–1989. The sample consists of 1,442 married or cohabiting women between the ages of 18 and 55 in 1986. The dependent variable fertility (f) is specified by a dummy variable that equals 1 if the age of the youngest child in the next year is 1. The explanatory variables are: (ags26l) which is a dummy variable that equals 1 if the woman has a child between 2 and 6 years old; education which has three levels (educ_1, educ_2 and educ_3), the female's age, race, and husband's income. An indicator of same sex of previous children (dsex), and its components: (dsexf) for girls, and (dsexm) for boys. This variable exploits the widely observed phenomenon of parental preferences for a mixed sibling-sex composition in developed countries. Therefore, a dummy for whether the sex of the next child matches the sex of the previous children provides a plausible predictor for additional childbearing. The data set can be obtained from the *Journal of Business & Economic Statistics* archive data web site. Problem 15 asks the reader to replicate some of the results obtained in the original article by Carrasco (2001). The estimates reveal that having children of the same sex has a significant and positive effect on the probability of having an additional child. The marginal effect of same sex children increases the probability of fertility by 3%, see Table 13.10. These are obtained using the *dprobit* command in Stata.

Table 13.7 Marginal Effects: Employment and Problem Drinking

. mfx compute

Marginal effects after probit
y = Pr(emp) (predict)
= .92244871

variable	dy/dx	Std. Err.	z	$P>\|z\|$	[95% Conf.	Interval]	X
hvdrnk90*	–.0161704	.00962	–1.68	0.093	–.035034	.002693	.099165
ue88	–.0077362	.00205	–3.78	0.000	–.011747	–.003725	5.56921
age	.0144674	.00248	5.83	0.000	.009607	.019327	39.1757
agesq	–.0001894	.00003	–6.37	0.000	–.000248	–.000131	1627.61
educ	.0068513	.00096	7.12	0.000	.004966	.008737	13.3096
married*	.0488911	.01009	4.85	0.000	.029119	.068663	.816432
famsize	.002743	.00204	1.35	0.179	–.001253	.006739	2.7415
white*	.069445	.01007	6.90	0.000	.049709	.089181	.853085
hlstat1*	.2460794	.01484	16.58	0.000	.216991	.275167	.415903
hlstat2*	.1842432	.00992	18.57	0.000	.164799	.203687	.301873
hlstat3*	.130786	.00661	19.80	0.000	.11784	.143732	.205254
hlstat4*	.0779836	.00415	18.77	0.000	.069841	.086126	.053451
region1*	.0049107	.00875	0.56	0.575	–.012233	.022054	.203014
region2*	.0086088	.0075	1.15	0.251	–.006092	.023309	.265628
region3*	.0252543	.00715	3.53	0.000	.011247	.039262	.318265
msa1*	–.0107946	.00779	–1.39	0.166	–.026061	.004471	.333232
msa2*	.0109542	.00735	1.49	0.136	–.003456	.025365	.434942
q1*	–.0158927	.00825	–1.93	0.054	–.032053	.000268	.254632
q2*	–.0075883	.00795	–0.95	0.340	–.023167	.007991	.252698
q3*	–.0043066	.00807	–0.53	0.594	–.020121	.011508	.242822

(*) dy/dx is for discrete change of dummy variable from 0 to 1

13.10 Multinomial Choice Models

In many economic situations, the choice may be among m alternatives where $m > 2$. These may be unordered alternatives like the selection of a mode of transportation, bus, car or train, or an occupational choice like lawyer, carpenter, teacher, etc., or they may be ordered alternatives like bond ratings, or the response to an opinion survey, which could vary from strongly agree to strongly disagree. Ordered response multinomial models utilize the extra information implicit in the ordinal nature of the dependent variable. Therefore, these models have a different likelihood than unordered response multinomial models and have to be treated separately.

13.10.1 Ordered Response Models

Suppose there are three bond ratings, A, AA and AAA. We sample n bonds and the i-th bond is rated A (which we record as $y_i = 0$) if its performance index $I_i^* < 0$, where 0 is again not restrictive. $I_i^* = x_i'\beta + u_i$, so the probability of an A rating or the $\Pr[y_i = 0]$ is

$$\pi_{1i} = \Pr[y_i = 0] = P[I_i^* < 0] = P[u_i < -x_i'\beta] = F(-x_i'\beta) \qquad (13.34)$$

Table 13.8 Average Marginal Effects: Employment and Problem Drinking

. margeff

Average partial effects after probit
y = Pr(emp)

variable	Coef.	Std. Err.	z	$P>\|z\|$	[95% Conf.	Interval]
hvdrnk90	–.0164971	.009264	–1.78	0.075	–.0346543	.00166
ue88	–.0078854	.0019748	–3.99	0.000	–.011756	–.0040149
age	.0147633	.0024012	6.15	0.000	.010057	.0194697
agesq	–.000193	.0000287	–6.73	0.000	–.0002493	–.0001368
educ	.0069852	.0009316	7.50	0.000	.0051593	.0088112
married	.048454	.0070149	6.91	0.000	.0347051	.0622028
famsize	.002796	.0019603	1.43	0.154	–.0010461	.0066382
white	.0685255	.0062822	10.91	0.000	.0562127	.0808383
hlstat1	.2849987	.0059359	48.01	0.000	.2733645	.2966328
hlstat2	.2318828	.0049776	46.59	0.000	.2221269	.2416386
hlstat3	.1725703	.0049899	34.58	0.000	.1627903	.1823502
hlstat4	.0914458	.0048387	18.90	0.000	.0819621	.1009295
region1	.0050178	.0083778	0.60	0.549	–.0114025	.021438
region2	.0088116	.0071262	1.24	0.216	–.0051556	.0227787
region3	.0259534	.0064999	3.99	0.000	.0132139	.0386929
msa1	–.0109515	.007632	–1.43	0.151	–.02591	.0040071
msa2	.0111628	.0067952	1.64	0.100	–.0021556	.0244811
q1	–.0160925	.0080458	–2.00	0.045	–.0318619	–.0003231
q2	–.0077086	.0076973	–1.00	0.317	–.0227951	.0073779
q3	–.0043814	.0077835	–0.56	0.573	–.0196368	.010874

Table 13.9 Actual vs Predicted: Employment and Problem Drinking

. estat class
Probit model for emp

	True		
Classified	D	~D	Total
+	8743	826	9569
–	79	174	253
Total	8822	1000	9822

Classified + if predicted Pr(D) >= .5
True D defined as emp != 0

Sensitivity	Pr(+| D)	99.10 %
Specificity	Pr(–|~D)	17.40 %
Positive predictive value	Pr(D| +)	91.37 %
Negative predictive value	Pr(~D| –)	68.77 %
False + rate for true ~D	Pr(+|~D)	82.60 %
False – rate for true D	Pr(–| D)	0.90 %
False + rate for classified	Pr(~D| +)	8.63 %
False – rate for classified	Pr(D| –)	31.23 %
Correctly classified		90.79 %

Table 13.10 Marginal Effects: Fertility and Same Sex of Previous Children

. dprobit f dsex ags26l educ_2 educ_3 age drace inc

Probit regression, reporting marginal effects	Number of obs = 5768
	LR chi2(7) = 964.31
	Prob > chi2 = 0.0000
Log likelihood = 1561.1312	Pseudo R2 = 0.2360

f	dF/dx	Std. Err.	z	$P > \|z\|$	x-bar	[95% Conf.	Interval]
dsex*	.0302835	.0069532	5.40	0.000	.256415	.016655	.043912
ags26l*	–.1618148	.0066629	–13.22	0.000	.377601	–.174874	–.148756
educ_2*	.0022157	.0090239	0.24	0.808	.717753	–.015471	.019902
educ_3*	.0288636	.0140083	2.45	0.014	.223994	.001408	.056319
age	–.0065031	.0007644	–16.65	0.000	32.8024	–.008001	–.005005
drace*	–.0077119	.0055649	–1.45	0.146	.773232	–.018619	.003195
inc	.0002542	.000241	1.06	0.289	12.8582	–.000218	.000727
obs. P	.1137309						
pred. P	.0367557	(at xbar)					

(*) dF/dx is for discrete change of dummy variable from 0 to 1
z and $P > |z|$ correspond to the test of the underlying coefficient being 0

The i-th bond is rated AA (which we record as $y_i = 1$) if its performance index I_i^* is between 0 and c where c is a positive number, with probability

$$\begin{aligned} \pi_{2i} &= \Pr[y_i = 1] = P[0 \leq I_i^* < c] \\ &= P[0 \leq x_i'\beta + u_i < c] = F(c - x_i'\beta) - F(-x_i'\beta) \end{aligned} \tag{13.35}$$

The i-th bond is rated AAA (which we record as $y_i = 2$) if $I_i^* \geq c$, with probability

$$\pi_{3i} = \Pr[y_i = 2] = P[I_i^* \geq c] = P[x_i'\beta + u_i \geq c] = 1 - F(c - x_i'\beta) \tag{13.36}$$

F can be the logit or probit function. The log-likelihood function for the ordered probit is given by

$$\begin{aligned} \log\ell(\beta, c) &= \textstyle\sum_{y_i=0} \log(\Phi(-x_i'\beta)) + \sum_{y_i=1} \log[\Phi(c - x_i'\beta) - \Phi(-x_i'\beta)] \\ &\quad + \textstyle\sum_{y_i=2} \log[1 - \Phi(c - x_i'\beta)]. \end{aligned} \tag{13.37}$$

For the probabilities given in (13.34), (13.35) and (13.36), the marginal effects of changes in the regressors are:

$$\partial\pi_{1i}/\partial x_i = -f(-x_i'\beta)\beta \tag{13.38}$$
$$\partial\pi_{2i}/\partial x_i = [f(-x_i'\beta) - f(c - x_i'\beta)]\beta \tag{13.39}$$
$$\partial\pi_{3i}/\partial x_i = f(c - x_i'\beta)\beta \tag{13.40}$$

Generalizing this model to m bond ratings is straight forward. The likelihood, the score and the Hessian for the m-ordered probit model are given in Maddala (1983, pp. 47–49).

Illustrative Example: Corporate Bond Rating. This data set is obtained from Baum(2006) by issuing the following command in Stata:

```
.use http://www.stata-press.com/data/imeus/panel84extract, clear
```

This data set contains ratings of 98 corporate bonds coded as 2 to 5 (*rating83c*). The rating 2 corresponds to the lowest rating BA_B_C and 5 to the highest rating AAA. These are given in Table 13.11.

Table 13.11 Corporate Bond Rating

. tab rating83c

Bond rating, 1982	Freq.	Percent	Cum.
BA_B_C	26	26.53	26.53
BAA	28	28.57	55.10
AA_A	15	15.31	70.41
AAA	29	29.59	100.00
Total	98	100.00	

This is modeled as an ordered logit with two explanatory variables: *ia83*, the income to asset ratio in 1983, and the change in that ratio from 1982 to 1983 (*dia*). The summary statistics are given in Table 13.12.

Table 13.12 Ordered Logit

Variable	Obs	Mean	Std. Dev.	Min	Max
rating83c	98	3.479592	1.17736	2	5
ia83	98	10.11473	7.441946	–13.08016	30.74564
dia	98	.7075242	4.711211	–10.79014	20.05367

. ologit rating83c ia83 dia

Ordered logistic regression			
	Number of obs	=	98
	LR chi2(2)	=	11.54
	Prob > chi2	=	0.0021
Log likelihood = –127.27146	Pseudo R2	=	0.0434

rating83c	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
ia83	.0939166	.0296196	3.17	0.002	.0358633	.1519699
dia	–.0866925	.0449789	–1.93	0.054	–.1748496	.0014646
/cut1	–.1853053	.3571432			–.8852932	.5146826
/cut2	1.185726	.3882099			.4248488	1.946604
/cut3	1.908412	.4164896			1.092108	2.724717

Income/assets has a positive effect on the rating, while the change in that ratio has a negative effect! Both are significant.

Table 13.13 Predicted Bond Rating

```
. predict pbabc pbaa paa paaa, pr
. sum pbabc pbaa paa paaa, separator(0)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
pbabc	98	.2729981	.1224448	.0388714	.7158453
pbaa	98	.2950074	.0456984	.0985567	.3299373
paa	98	.1496219	.0274841	.0449056	.1787291
paaa	98	.2823726	.1304381	.0466343	.7528986

The /cut1 to /cut3 give the estimated thresholds of the ratings categories using a logit specification. The first one is not significant, while the other two are. Note that the 95% confidence interval for these thresholds overlap.

We can predict the probability of achieving this rating using the *predict* command naming the values we want it to take for each category, see Table 13.13. The average of these predictions is pretty close to the actual frequencies observed in each category.

13.10.2 Unordered Response Models

There are m choices each with probability $\pi_{i1}, \pi_{i2}, \ldots, \pi_{im}$ for individual i. $y_{ij} = 1$ if individual i chooses alternative j, otherwise it is 0. This means that $\sum_{j=1}^{m} y_{ij} = 1$ and $\sum_{j=1}^{m} \pi_{ij} = 1$. The likelihood function for n individuals is a multinomial given by:

$$\ell = \prod_{i=1}^{n} (\pi_{i1})^{y_{i1}} (\pi_{i2})^{y_{i2}} .. (\pi_{im})^{y_{im}} \tag{13.41}$$

This model can be motivated by a utility maximization story where the utility that individual i derives from say the occupational choice j is denoted by U_{ij} and is a function of the job attributes for the i-th individual, i.e., some x_{ij}'s like the present value of potential earnings, and training cost/net worth for that job choice for individual i, see Boskin (1974).

$$U_{ij} = x'_{ij}\beta + \epsilon_{ij} \tag{13.42}$$

where β is a vector of implicit prices for these occupational characteristics. Therefore, the probability of choosing the first occupation is given by:

$$\begin{aligned} \pi_{i1} &= \Pr[U_{i1} > U_{i2}, U_{i1} > U_{i3}, \ldots, U_{i1} > U_{im}] \\ &= \Pr[\epsilon_{i2} - \epsilon_{i1} < (x'_{i1} - x'_{i2})\beta, \epsilon_{i3} - \epsilon_{i1} \\ &< (x'_{i1} - x'_{i3})\beta, \ldots, \epsilon_{im} - \epsilon_{i1} < (x'_{i1} - x'_{im})\beta] \end{aligned} \tag{13.43}$$

The normality assumption involves a number of integrals but has the advantage of not necessarily assuming the ϵ's to be independent. The more popular assumption computationally is the multinomial logit model. This arises if and only if the ϵ's are independent and identically distributed as a Weibull density function, see McFadden (1974). The latter is given by $F(z) = \exp(-\exp(-z))$. The difference between any two random variables with a Weibull distribution has a logistic distribution $\Lambda(z) = e^z/1 + e^z$, giving the *conditional logit model*:

$$\begin{aligned} \pi_{ij} &= \Pr[y_i = j] = \exp[(x_{ij} - x_{im})'\beta]/\{1 + \textstyle\sum_{j=1}^{m-1} \exp[(x_{ij} - x_{im})'\beta]\} \\ &= \exp[x'_{ij}\beta]/\textstyle\sum_{j=1}^{m} \exp[x'_{ij}\beta] \qquad \text{for} \qquad j = 1, 2, \ldots, m-1 \end{aligned} \tag{13.44}$$

and $\pi_{im} = \Pr[y_i = m] = 1/\{1 + \sum_{j=1}^{m-1} \exp[(x_{ij} - x_{im})'\beta]\} = \exp[x'_{im}\beta]/\sum_{j=1}^{m} \exp[x'_{ij}\beta]$. There are two consequences of this conditional logit specification. The first is that the odds of any two alternative occupations, say 1 and 2 is given by

$$\pi_{i1}/\pi_{i2} = \exp[(x_{i1} - x_{i2})'\beta]$$

and this does not change when the number of alternatives change from m to m^*, since the denominators divide out. Therefore, the odds are unaffected by an additional alternative. This is known as the *independence of irrelevant alternatives* property and can represent a serious weakness in the conditional logit model. For example, suppose the choices are between a pony and a bicycle, and children choose a pony two-thirds of the time. Suppose that an additional alternative is made available, an additional bicycle but of a different color, then one would still expect two-thirds of the children to choose the pony and the remaining one-third to split choices among the bicycles according to their color preference. In the conditional logit model, however, the proportion choosing the pony must fall to one half if the odds relative to either bicycle is to remain two to one in favor of the pony. This illustrates the point that when two or more of the m alternatives are close substitutes, the conditional logit model may not produce reasonable results. This feature is a consequence of assuming the errors ϵ_{ij}'s as independent. Hausman and McFadden (1984) proposed a Hausman type test to check for the independence of these errors. They suggest that if a subset of the choices is truly irrelevant then omitting it from the model altogether will not change the parameter estimates systematically. Including them if they are irrelevant preserves consistency but is inefficient. The test statistic is

$$q = (\widehat{\beta}_s - \widehat{\beta}_f)'[\widehat{V}_s - \widehat{V}_f]^{-1}(\widehat{\beta}_s - \widehat{\beta}_f) \tag{13.45}$$

where s indicates the estimators based on the restricted subset and f denotes the estimator based on the full set of choices. This is asymptotically distributed as χ^2_k, where k is the dimension of β.

Second, in this specification, none of the x_{ij}'s can be constant across different alternatives, because the corresponding β will not be identified. This means that we cannot include individual specific variables that do not vary across alternatives like race, sex, age, experience, income, etc. The latter type of data is more frequent in economics, see Schmidt and Strauss (1975). In this case the specification can be modified to allow for a differential impact of the explanatory variables upon the odds of choosing one alternative rather than the other:

$$\pi_{ij} = \Pr[y_i = j] = \exp(x'_{ij}\beta_j)/\sum_{j=1}^{m} \exp(x'_{ij}\beta_j) \quad \text{for} \quad j = 1, \ldots, m \tag{13.46}$$

where now the parameter vector is indexed by j. If the x_{ij}'s are the same for every j, then

$$\pi_{ij} = Pr[y_i = j] = \exp(x'_i\beta_j)/\sum_{j=1}^{m} \exp(x'_i\beta_j) \quad \text{for} \quad j = 1, \ldots, m \tag{13.47}$$

This is the model used by Schmidt and Strauss (1975). A normalization would be to take $\beta_m = 0$, in which case, we get the *multinomial logit* model

$$\pi_{im} = 1/\sum_{j=1}^{m} \exp(x'_i\beta_j) \tag{13.48}$$

and

$$\pi_{ij} = \exp(x'_i\beta_j)/[1 + \sum_{j=1}^{m-1} \exp(x'_i\beta_j)] \quad \text{for} \quad j = 1, 2, \ldots, m-1. \tag{13.49}$$

The likelihood function, score equations, Hessian and information matrices are given in Maddala (1983, pp. 36–37).

Multinomial Logit Model. Terza (2002) reconsidered the Mullahy and Sindelar (1996) data set for problem drinking described in section 13.9. However, Terza reclassified the dependent variable as follows: y=1 when the individual is out of the labor force, y=2 when this individual is unemployed, and y=3 when this individual is employed. Table 13.14 runs a multinomial logit model which replicates some of the results in Terza (2002, p. 399) for males using Stata. Although the health variables are still significant the problem drinking variable is not significant. Problem 16 asks the reader to replicate these results for females.

13.11 The Censored Regression Model

Suppose one is interested in the amount one is willing to spend on the purchase of a durable good. For example, a car. In this case, one would observe the expenditures only if the car is bought, so

$$y_i^* = x_i'\beta + u_i \quad \text{if} \quad y_i^* > 0 \tag{13.50}$$

where x_i denotes a vector of household characteristics, such as income, number of children or education. y_i^* is a latent variable, in this case the amount one is willing to spend on a car. We observe $y_i = y_i^*$ only if $y_i^* > 0$ and we set $y_i = 0$ if $y_i^* \leq 0$. The censoring at zero is of course arbitrary, and the u_i's are assumed to be $\text{IIN}(0, \sigma^2)$. This is known as the *Tobit* model after Tobin (1958). In this case, we have *censored* observations since we do not observe any y^* that is negative. All we observe is the fact that this household did not buy a car and a corresponding vector x_i of this household's characteristics. Without loss of generality, we assume that the first n_1 observations have positive y_i^*'s and the remaining $n_0 = n - n_1$ observations have non-positive y_i^*'s. In this case, OLS on the first n_1 observations, i.e., using only the positive observed y_i^*'s would be biased since u_i does not have zero mean. In fact, by omitting observations for which $y_i^* \leq 0$ from the sample, one is only considering disturbances from (13.50) such that $u_i > -x_i'\beta$. The distribution of these u_i's is a truncated normal density given in Figure 13.2. The mean of this density is not zero and is dependent on β, σ^2 and x_i. More formally, the regression function can be written as:

$$\begin{aligned} E(y_i^*/x_i, y_i^* > 0) &= x_i'\beta + E[u_i/y_i^* > 0] = x_i'\beta + E[u_i/u_i > -x_i'\beta] \\ &= x_i'\beta + \sigma\gamma_i \quad \text{for} \quad i = 1, 2, \ldots, n_1 \end{aligned} \tag{13.51}$$

where $\gamma_i = \phi(-z_i)/[1 - \Phi(-z_i)]$ and $z_i = x_i'\beta/\sigma$. See Greene (1993, p. 685) for the moments of a truncated normal density or the Appendix to this chapter. OLS on the positive y_i^*'s omits the second term in (13.51), and is therefore biased and inconsistent.

A simple two-step can be used to estimate (13.51). First, we define a dummy variable d_i which takes the value 1 if y_i^* is observed and 0 otherwise. This allows us to perform probit estimation on the whole sample, and provides us with a consistent estimator of (β/σ). Also, $P[d_i = 1] = P[y_i^* > 0] = P[u_i > -x_i'\beta]$ and $P[d_i = 0] = P[y_i^* \leq 0] = P[u_i \leq -x_i'\beta]$. Therefore, the likelihood function is given by

$$\begin{aligned} \ell &= \textstyle\prod_{i=1}^n [P(u_i \leq -x_i'\beta)]^{1-d_i}[P(u_i > -x_i'\beta)]^{d_i} \\ &= \textstyle\prod_{i=1}^n \Phi(z_i)^{d_i}[1 - \Phi(z_i)]^{1-d_i} \quad \text{where} \quad z_i = x_i'\beta/\sigma \end{aligned} \tag{13.52}$$

Table 13.14 Multinomial Logit Results: Problem Drinking

. mlogit y alc90th ue88 age agesq schooling married famsize white excellent verygood good fair northeast midwest south centercity othermsa q1 q2 q3, baseoutcome(1)

Multinomial logistic regression

	Number of obs	=	9822
	Wald chi2(20)	=	1276.47
	Prob > chi2	=	0.0000
Log likelihood = –3217.481	Pseudo R2	=	0.1655

y	Coef.	Std. Err.	z	$P > \|z\|$	[95% Conf.	Interval]
2						
alc90th	.1270931	.21395	0.59	0.552	–.2922412	.5464274
ue88	.0458099	.051355	0.89	0.372	–.0548441	.1464639
age	.1617634	.0663205	2.44	0.015	.0317776	.2917492
agesq	–.0024377	.0007991	–3.05	0.002	–.004004	–.0008714
schooling	–.0092135	.0245172	–0.38	0.707	–.0572664	.0388393
married	.4004928	.1927458	2.08	0.038	.022718	.7782677
famsize	.0622453	.0503686	1.24	0.217	–.0364753	.1609659
white	.0391309	.1705625	0.23	0.819	–.2951653	.3734272
excellent	2.91833	.4486757	6.50	0.000	2.038942	3.797719
verygood	2.978336	.4505932	6.61	0.000	2.09519	3.861483
good	2.493939	.4446815	5.61	0.000	1.622379	3.365499
fair	1.460263	.4817231	3.03	0.002	.5161027	2.404422
northeast	.0849125	.2374365	0.36	0.721	–.3804545	.5502796
midwest	.0158816	.2037486	0.08	0.938	–.3834583	.4152215
south	.1750244	.2027444	0.86	0.388	–.2223474	.5723962
centercity	–.2717445	.1911074	–1.42	0.155	–.6463081	.1028192
othermsa	–.0921566	.1929076	–0.48	0.633	–.4702486	.2859354
q1	.422405	.1978767	2.13	0.033	.0345738	.8102362
q2	–.0219499	.2056751	–0.11	0.915	–.4250657	.3811659
q3	–.0365295	.2109049	–0.17	0.862	–.4498954	.3768364
_cons	–6.113244	1.427325	–4.28	0.000	–8.910749	–3.315739
3						
alc90th	–.1534987	.1395003	–1.10	0.271	–.4269144	.1199169
ue88	–.0954848	.033631	–2.84	0.005	–.1614004	–.0295693
age	.227164	.0409884	5.54	0.000	.1468282	.3074999
agesq	–.0030796	.0004813	–6.40	0.000	–.0040228	–.0021363
schooling	.0890537	.0152314	5.85	0.000	.0592008	.1189067
married	.7085708	.1219565	5.81	0.000	.4695405	.9476012
famsize	.0622447	.0332365	1.87	0.061	–.0028975	.127387
white	.7380044	.1083131	6.81	0.000	.5257147	.9502941
excellent	3.702792	.1852415	19.99	0.000	3.339725	4.065858
verygood	3.653313	.1894137	19.29	0.000	3.282069	4.024557
good	2.99946	.1786747	16.79	0.000	2.649264	3.349656
fair	1.876172	.1885159	9.95	0.000	1.506688	2.245657
northeast	.088966	.1491191	0.60	0.551	–.203302	.3812341
midwest	.1230169	.1294376	0.95	0.342	–.130676	.3767099
south	.4393047	.1298054	3.38	0.001	.1848908	.6937185
centercity	–.2689532	.1231083	–2.18	0.029	–.510241	–.0276654
othermsa	.0978701	.1257623	0.78	0.436	–.1486195	.3443598
q1	–.0274086	.1286695	–0.21	0.831	–.2795961	.224779
q2	–.110751	.126176	–0.88	0.380	–.3580514	.1365494
q3	–.0530835	.1296053	–0.41	0.682	–.3071052	.2009382
_cons	–6.237275	.8886698	–7.02	0.000	–7.979036	–4.495515

(y==1 is the base outcome)

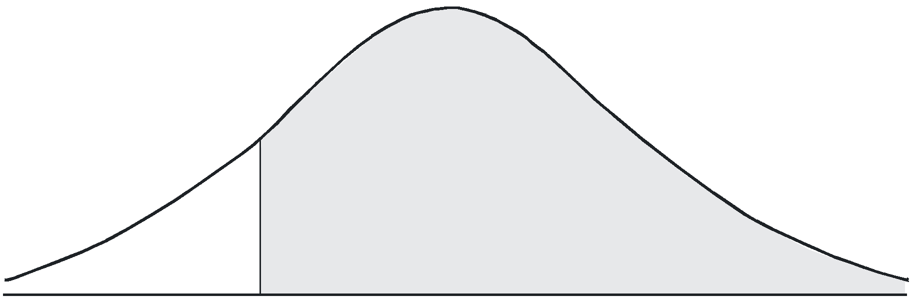

Figure 13.2 Truncated Normal Distribution

and once β/σ is estimated, we substitute these estimates in z_i and γ_i given below (13.51) to get $\widehat{\gamma}_i$. The second step is to estimate (13.51) using only the positive y_i^*'s with $\widehat{\gamma}_i$ substituted for γ_i. The resulting estimator of β is consistent and asymptotically normal, see Heckman (1976, 1979).

Alternatively, one can use maximum likelihood procedures to estimate the Tobit model. Note that we have two sets of observations: (i) the positive y_i^*'s with $y_i = y_i^*$, for which we can write the density function $N(x_i'\beta, \sigma^2)$, and (ii) the non-positive y_i^*'s for which we assign $y_i = 0$ with probability

$$\Pr[y_i = 0] = \Pr[y_i^* < 0] = \Pr[u_i < -x_i'\beta] = \Phi(-x_i'\beta/\sigma) = 1 - \Phi(x_i'\beta/\sigma) \tag{13.53}$$

The probability over the entire censored region gets assigned to the censoring point. This allows us to write the following log-likelihood:

$$\begin{aligned}\log\ell \;&=\; -(1/2)\textstyle\sum_{i=1}^{n_1}\log(2\pi\sigma^2) - (1/2\sigma^2)\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2 \\ &\quad + \textstyle\sum_{i=n_1+1}^{n}\log[1 - \Phi(x_i'\beta/\sigma)]\end{aligned} \tag{13.54}$$

Differentiating with respect to β and σ^2, see Maddala (1983, p. 153), one gets

$$\partial\log\ell/\partial\beta = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)x_i/\sigma^2 - \sum_{i=n_1+1}^{n}\phi_i x_i/\sigma[1 - \Phi_i] \tag{13.55}$$

$$\partial\log\ell/\partial\sigma^2 = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)^2/2\sigma^4 - (n_1/2\sigma^2) + \sum_{i=n_1+1}^{n}\phi_i x_i'\beta/[2\sigma^3(1 - \Phi_i)] \tag{13.56}$$

where Φ_i and ϕ_i are evaluated at $z_i = x_i'\beta/\sigma$.

Premultiplying (13.55) by $\beta'/2\sigma^2$ and adding the result to (13.56), one gets

$$\widehat{\sigma}^2_{MLE} = \textstyle\sum_{i=1}^{n_1}(y_i - x_i'\beta)y_i/n_1 = Y_1'(Y_1 - X_1\beta)/n_1 \tag{13.57}$$

where Y_1 denotes the $n_1 \times 1$ vector of non-zero observations on y_i, X_1 is the $n_1 \times k$ matrix of values of x_i for the non-zero y_i's. Also, after multiplying throughout by σ, (13.55) can be written as:

$$-X_0'\gamma_0 + X_1'(Y_1 - X_1\beta)/\sigma = 0 \tag{13.58}$$

where X_0 denotes the $n_0 \times k$ matrix of x_i's for which y_i is zero, γ_0 is an $n_0 \times 1$ vector of γ_i's $= \phi_i/[1 - \Phi_i]$ evaluated at $z_i = x_i'\beta/\sigma$ for the observations for which $y_i = 0$. Solving (13.58) one gets

$$\widehat{\beta}_{MLE} = (X_1'X_1)^{-1}X_1'Y_1 - \sigma(X_1'X_1)^{-1}X_0'\gamma_0 \tag{13.59}$$

Note that the first term in (13.59) is the OLS estimator for the first n_1 observations for which y_i^* is positive.

One can use the Newton-Raphson procedure or the method of scoring, for the second derivatives of the log-likelihood, see Maddala (1983, pp. 154–156). These can be computed with the *tobit* command in Stata. Note that for the Tobit specification, both β and σ^2 are identified. This is contrasted to the logit and probit specifications where only the ratio (β/σ^2) is identified. Wooldridge (2009, Chapter 17) recommends one obtain the estimates of (β/σ^2) from a probit and comparing those with the Tobit estimates generated by dividing $\widehat{\beta}$ by $\widehat{\sigma}^2$. If these estimates are different or have different signs, then the Tobit estimation may not be appropriate. Problem 13.17 illustrates the Tobit estimation for married women labor supply example using the Mroz (1987) data.

Maddala warns that the Tobit specification is not necessarily the right specification every time we have zero observations. It is applicable only in those cases where the latent variable can, in principle, take negative values and the observed zero values are a consequence of censoring and non-observability. In fact, one cannot have negative expenditures on a car, negative hours of work or negative wages. However, one can enter employment and earn wages when one's observed wage is larger than the reservation wage. Let y^* be the difference between observed wage and reservation wage. Only if y^* is positive will wages be observed. Final warning: The Tobit specification is heavily reliant on the *normality* and *homoskedasticity* assumptions. Failure of these assumptions leads to misleading inference.

13.12 The Truncated Regression Model

The truncated regression model excludes or truncates some observations from the sample. For example, in studying poverty we exclude the rich, say with earnings larger than some upper limit y^u from our sample. The sample is therefore not random and applying least squares to the truncated sample lead to biased and inconsistent results, see Figure 13.3. This differs from censoring. In the latter case, no data is excluded. In fact, we observe the characteristics of all households even those that do not actually purchase a car. The truncated regression model is given by

$$y_i^* = x_i'\beta + u_i \quad i = 1, 2, \ldots, n \quad \text{with} \quad u_i \sim \text{IIN}(0, \sigma^2) \tag{13.60}$$

where y_i^* is for example earnings of the i-th household and x_i contains determinants of earnings like education, experience, etc. The sample contains observations on individuals with $y_i^* \leq y^u$. The probability that y_i^* will be observed is

$$\Pr[y_i^* \leq y^u] = \Pr[x_i'\beta + u_i \leq y^u] = \Pr[u_i < y^u - x_i'\beta] = \Phi(\frac{1}{\sigma}(y^u - x_i'\beta)) \tag{13.61}$$

In addition, using the results of a truncated normal density, see Greene (1993, p. 685)

$$E(u_i/y_i^* \leq y^u) = \frac{-\sigma\phi((y^u - x_i'\beta)/\sigma)}{\Phi((y^u - x_i'\beta)/\sigma)} \tag{13.62}$$

which is not necessarily zero. From (13.60) one can see that $E(y_i^*/y_i^* \leq y^u) = x_i'\beta + E(u_i/y_i^* < y^u)$. Therefore, OLS on (13.60) using the observed y_i^* is biased and inconsistent because it ignores the term in (13.62).

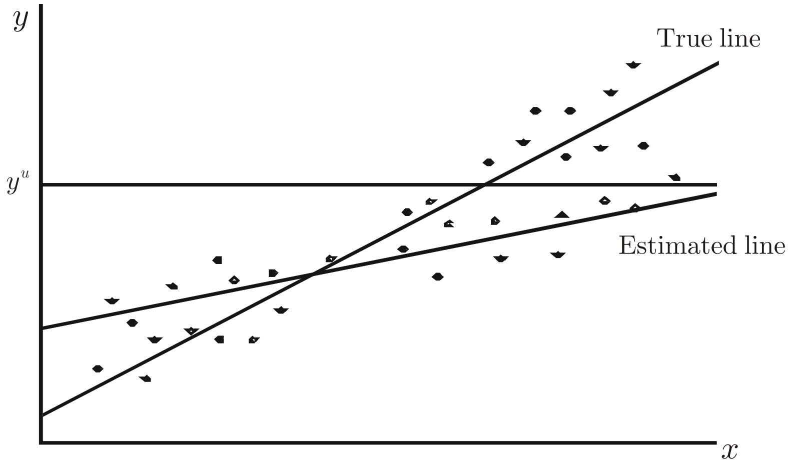

Figure 13.3 Truncated Regression Model

The density of y_i^* is normal but its total area is given by (13.61). A proper density function has to have an area of 1. Therefore, the density of y_i^* conditional on $y_i^* \leq y^u$ is simply the conditional density of y_i^* restricted to values of $y_i^* \leq y^u$ divided by the $\Pr[y_i^* \leq y^u]$, see the Appendix to this chapter:

$$\begin{aligned} f(y_i^*) &= \frac{\phi((y_i^* - x_i'\beta/\sigma)}{\sigma\Phi((y^u - x_i'\beta)/\sigma)} \quad \text{if} \quad y_i^* \leq y^u \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{13.63}$$

The log-likelihood function is therefore

$$\begin{aligned} \log\ell &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (y_i^* - x_i'\beta)^2 \\ &\quad - \textstyle\sum_{i=1}^n \log\Phi\left(\dfrac{y^u - x_i'\beta}{\sigma}\right) \end{aligned} \tag{13.64}$$

It is the last term which makes MLE differ from OLS on the observed sample. Hausman and Wise (1977) applied the truncated regression model to data from the New Jersey negative-income-tax experiment where families with incomes higher than 1.5 times the 1967 poverty line were eliminated from the sample.

13.13 Sample Selectivity

In labor economics, one observes the market wages of individuals only if the worker participates in the labor force. This happens when the worker's market wage exceeds his or her reservation wage. In a study of earnings, one does not observe the reservation wage and for non-participants in the labor force we record a zero market wage. This sample is censored because we observe the characteristics of these non-labor participants. If we restrict our attention to labor market participants only, then the sample is truncated. This example needs special attention because the censoring is not based directly on the dependent variable, as in section 13.11. Rather, it is based on the difference between market wage and reservation wage. This latent variable which determines the sample selection is correlated with the dependent variable. Hence, least squares on this model results in *selectivity bias*, see Heckman (1976, 1979). A sample generated by this type of self-selection may not represent the true population distribution no matter how big the sample size. However, one can correct for self-selection bias if the underlying sampling generating process can be understood and relevant identification conditions are available, see Lee (2001) for a excellent summary and the references cited there. In order to demonstrate this, let the earnings equation be given by

$$w_i^* = x_{1i}'\beta + u_i \qquad i = 1, 2, \ldots, n \tag{13.65}$$

and the labor participation (or selection) equation be given by

$$y_i^* = x_{2i}'\gamma + v_i \qquad i = 1, 2, \ldots, n \tag{13.66}$$

where u_i and v_i are bivariate normally distributed with mean zero and variance-covariance

$$\text{var}\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix} \tag{13.67}$$

Normalizing the variance of v_i to be 1 is not restrictive, since only the sign of y_i^* is observed. In fact, we only observe w_i and y_i where

$$\begin{aligned} w_i &= w_i^* \quad \text{if} \quad y_i^* > 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{13.68}$$

and

$$\begin{aligned} y_i &= 1 \quad \text{if} \quad y_i^* > 0 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

We observe $(y_i = 0, w_i = 0)$ and $(y_i = 1, w_i = w_i^*)$ only. The log-likelihood for this model is

$$\textstyle\sum_{y_i=0} \log \Pr[y_i = 0] + \sum_{y_i=1} \log \Pr[y_i = 1] f(w_i^*/y_i = 1) \tag{13.69}$$

where $f(w_i^*/y_i = 1)$ is the conditional density of w_i^* given that $y_i = 1$. The second term can also be written as $\Sigma_{y_i=1}\log \Pr[y_i = 1/w_i^*] f(w_i^*)$ which is another way of factoring the joint density function. $f(w_i^*)$ is in fact a normal density with conditional mean $x_{1i}'\beta$ and variance σ^2. Using properties of the bivariate normal density, one can write

$$y_i^* = x_{2i}'\gamma + \rho\left(\frac{1}{\sigma}(w_i^* - x_{1i}'\beta)\right) + \epsilon_i \tag{13.70}$$

where $\epsilon_i \sim \text{IIN}(0, \sigma^2(1 - \rho^2))$. Therefore,

$$\Pr[y_i = 1] = \Pr[y_i^* > 0] = \Phi\left(\frac{x_{2i}'\gamma + \rho((w_i - x_{1i}'\beta)/\sigma)}{\sqrt{1 - \rho^2}}\right) \tag{13.71}$$

where w_i has been substituted for w_i^* since $y_i = 1$. The likelihood function in (13.69) becomes

$$\begin{aligned} &\textstyle\sum_{y_i=0} \log(\Phi(-x_{2i}'\gamma)) + \sum_{y_i=1} \log\left(\frac{1}{\sigma}\phi(w_i - x_{1i}'\beta)\right) \\ &+ \textstyle\sum_{y_i=1} \log\Phi\left(\frac{x_{2i}'\gamma + \rho((w_i - x_{1i}'\beta)/\sigma}{\sqrt{1 - \rho^2}}\right) \end{aligned} \tag{13.72}$$

MLE may be computationally burdensome. Heckman (1976) suggested a two-step procedure which is based on rewriting (13.65) as

$$w_i^* = \beta x_{1i}' + \rho\sigma v_i + \eta_i \tag{13.73}$$

and replacing w_i^* by w_i and v_i by its conditional mean $E[v_i/y_i = 1]$. Using the results on truncated density, this conditional mean is given by $\phi(x_{2i}'\gamma)/\Phi(-x_{2i}'\gamma)$ known also as the *inverse Mills ratio*. Hence, (13.73) becomes

$$w_i = x_{1i}'\beta + \rho\sigma\frac{\phi(x_{2i}'\gamma)}{\Phi(-x_{2i}'\gamma)} + \text{residual} \tag{13.74}$$

Heckman's (1976) two-step estimator consists of (i) running a probit on (13.66) in the first step to get a consistent estimate of γ, (ii) substituting the estimated inverse Mills ratio in (13.74) and running OLS. Since σ is positive, this second stage regression provides a test for sample selectivity, i.e., for $\rho = 0$, by checking whether the t-statistic on the estimated inverse Mills ratio is significant. This statistic is asymptotically distributed as $N(0,1)$. Rejecting H_o implies there is a selectivity problem and one should not rely on OLS on (13.65) which ignores the selectivity bias term in (13.74). Davidson and MacKinnon (1993) suggest performing MLE using (13.72) rather than relying on the two-step results in (13.74) if the former is not computationally burdensome. Note that the Tobit model for car purchases given in (13.50) can be thought of as a special case of the sample selectivity model given by (13.65) and (13.66). In fact, the Tobit model assumes that the selection equation (the decision to buy a car) and the car expenditure equation (conditional on the decision to buy) are identical. Therefore, if one thinks that the specification of the selection equation is different from that of the expenditure equation, then one should *not* use the Tobit model. Instead, one should proceed with the two equation sample selectivity model discussed in this section. It is also important to emphasize that for the censored, truncated and sample selectivity models, normality and homoskedasticity are crucial assumptions. Suggested tests for these assumptions are given in Bera, Jarque and Lee (1984), Lee and Maddala (1985) and Pagan and Vella (1989). Alternative estimation methods that are more robust to violations of normality and heteroskedasticity include symmetrically trimmed least squares for Tobit models and least absolute deviations estimation for censored regression models. These were suggested by Powell (1984, 1986).

Notes

1. This is based on Davidson and MacKinnon (1993, pp. 523–526).

2. A binary response model attempts to explain a zero-one (or binary) dependent variable.

3. One should not use nR^2 as the test statistic because the total sum of squares in this case is not n.

Problems

1. *The Linear Probability Model.*

   (a) For the linear probability model described in (13.1), show that for $E(u_i)$ to equal zero, we must have $\Pr[y_i = 1] = x_i'\beta$.

   (b) Show that u_i is heteroskedastic with $\text{var}(u_i) = x_i'\beta(1 - x_i'\beta)$.

2. Consider the general log-likelihood function given in (13.16). Assume that *all the regression slopes are zero except for the constant* α.

   (a) Show that maximizing $\log\ell$ with respect to the constant α yields $\widehat{F}(\alpha) = \bar{y}$, where $\bar{y}$ is the proportion of the sample with $y_i = 1$.

   (b) Conclude that the value of the maximized likelihood is $\log\ell_r = n[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})]$.

   (c) Verify that for the union participation example in section 13.9 that $\log\ell_r = -390.918$.

3. For the *union participation* example considered in section 13.9:

   (a) Replicate Tables 13.3 and 13.5.

   (b) Using the measures of fit considered in section 13.8, compute $R_1^2, R_2^2, \ldots, R_5^2$ for the logit and probit models.

   (c) Compute the predicted value for the 10th observation using OLS, logit and probit models. Also, the corresponding standard errors.

   (d) The industry variable (IND) was not significant in all models. Drop this variable and run OLS, logit and probit. How do the results change? Compare with Tables 13.3 and 13.5.

   (e) Using the model results in part (d), test that the slope coefficients are all zero for the logit, probit, and linear probability models.

   (f) Test that the coefficients of IND, FEM and BLK in Table 13.3 are jointly insignificant using a LR test, a Wald test and a BRMR using OLS, logit and probit.

4. For the data used in the union participation example in section 13.9:

   (a) Run OLS, logit and probit using as the dependent variable OCC which is one if the individual is in a blue-collar occupation, and zero otherwise. For the independent variables use ED, WKS, EXP, SOUTH, SMSA, IND, MS, FEM and BLK. Compare the coefficient estimates across the three models. What variables are significant?

   (b) Using the measures of fit considered in section 13.8, compute $R_1^2, R_2^2, \ldots, R_5^2$ for the logit and probit models.

   (c) Tabulate the actual versus predicted values for OCC from all three model results, like Table 13.5 for Union. What is the proportion of correct decisions for OLS, logit and probit?

   (d) Test that the slope coefficients are all zero for the logit, probit, and linear probability models.

5. *Truncated Uniform Density.* Let x be a uniformly distributed random variable with density

$$f(x) = \frac{1}{2} \quad \text{for} \quad -1 < x < 1$$

   (a) What is the density function of $f(x/x > -1/2)$? **Hint:** Use the definition of a conditional density

$$f(x/x > -1/2) = f(x)/\Pr[x > -1/2] \quad \text{for} \quad -\frac{1}{2} < x < 1.$$

   (b) What is the conditional mean $\mathrm{E}(x/x > -1/2)$? How does it compare with the unconditional mean of x? Note that because we truncated the density from below, the new mean should shift to the right.

   (c) What is the conditional variance $\mathrm{var}(x/x > -1/2)$? How does it compare to the unconditional $\mathrm{var}(x)$? (Truncation reduces the variance).

6. *Truncated Normal Density.* Let x be $N(1,1)$. Using the results in the Appendix, show that:

(a) The conditional density $f(x/x > 1) = 2\phi(x-1)$ for $x > 1$ and $f(x/x < 1) = 2\phi(x-1)$ for $x < 1$.

(b) The conditional mean $E(x/x > 1) = 1 + 2\phi(0)$ and $E(x/x < 1) = 1 - 2\phi(0)$. Compare with the unconditional mean of x.

(c) The conditional variance $\text{var}(x/x > 1) = \text{var}(x/x < 1) = 1 - 4\phi^2(0)$. Compare with the unconditional variance of x.

7. *Censored Normal Distribution.* This is based on Greene (1993, pp. 692–693). Let y^* be $N(\mu, \sigma^2)$ and define $y = y^*$ if $y^* > c$ and $y = c$ if $y^* < c$ for some constant c.

(a) Verify the $E(y)$ expression given in (A.7).

(b) Derive the var(y) expression given in (A.8). **Hint:** Use the fact that

$$\text{var}(y) = E(\text{conditional variance}) + \text{var}(\text{conditional mean})$$

and the formulas given in the Appendix for conditional and unconditional means of a truncated normal random variable.

(c) For the special case of $c = 0$, show that (A.7) simplifies to $E(y) = \Phi(\mu/\sigma)\left[\mu + \dfrac{\sigma\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\right]$ and (A.8) simplifies to

$$\text{var}(y) = \sigma^2\Phi\left(\frac{\mu}{\sigma}\right)\left[1 - \delta\left(\frac{-\mu}{\sigma}\right) + \left(-\frac{\mu}{\sigma} - \frac{\phi(/\sigma)}{\Phi(\mu/\sigma)}\right)^2 \Phi\left(-\frac{\mu}{\sigma}\right)\right]$$

where $\delta\left(\dfrac{-\mu}{\sigma}\right) = \dfrac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\left[\dfrac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} + \dfrac{\mu}{\sigma}\right]$. Similar expressions can be derived for censoring of the upper part rather than the lower part of the distribution.

8. *Fixed and Adjustable Rate Mortgages.* Dhillon, Shilling and Sirmans (1987) considered the economic decision of choosing between fixed and adjustable rate mortgages. The data consisted of 78 households borrowing money from a Louisiana mortgage banker. Of these, 46 selected fixed rate mortgages and 32 selected uncapped adjustable rate mortgages. This data set can be downloaded from the Springer web site and is labelled DHILLON.ASC. It was obtained from Lott and Ray (1992). These variables include:

Y	= 0 if adjustable rate and 1 if fixed rate.
BA	= Age of the borrower.
BS	= Years of schooling for the borrower.
NW	= Net worth of the borrower.
FI	= Fixed interest rate.
PTS	= Ratio of points paid on adjustable to fixed rate mortgages.
MAT	= Ratio of maturities on adjustable to fixed rate mortgages.
MOB	= Years at the present address.
MC	= 1 if the borrower is married and 0 otherwise.
FTB	= 1 if the borrower is a first-time home buyer and 0 otherwise.
SE	= 1 if the borrower is self-employed and 0 otherwise.
YLD	= The difference between the 10-year treasury rate less the 1-year treasury rate.
MARG	= The margin on the adjustable rate mortgage.
CB	= 1 if there is a co-borrower and 0 otherwise.
STL	= Short-term liabilities.
LA	= Liquid assets.

The probability of choosing a variable rate mortgage is a function of personal borrower characteristics as well as mortgage characteristics. The efficient market hypothesis state that only cost variables and not personal borrower characteristics influence the borrower's decision between fixed and adjustable rate mortgages. Cost variables include FI, MARG, YLD, PTS and MAT. The rest of the variables are personal characteristics variables. The principal agent theory suggests that information is asymmetric between lender and borrower. Therefore, one implication of this theory is that the personal characteristics of the borrower will be significant in the choice of mortgage loan.

(a) Run OLS of Y on all variables in the data set. For this linear probability model what does the F-statistic for the significance of all slope coefficients yield? What is the R^2? How many predictions are less than zero or larger than one?

(b) Using only the cost variables in the restricted regression, test that personal characteristics are jointly insignificant. **Hint:** Use the Chow-F statistic. Do you find support for the efficient market hypothesis?

(c) Run the above model using the logit specification. Test the efficient market hypothesis. Does your conclusion change from that in part (b)? **Hint:** Use the likelihood ratio test or the BRMR.

(d) Do part (c) using the probit specification.

9. *Sampling Distribution of OLS Under a Logit Model.* This is based on Baltagi (2000).

Consider a simple logit regression model

$$y_t = \Lambda(\beta x_t) + u_t$$

for $t = 1, 2$, where $\Lambda(z) = e^z/(1 + e^z)$ for $-\infty < z < \infty$. Let $\beta = 1$, $x_1 = 1$, $x_2 = 2$ and assume that the u_t's are independent with mean zero.

(a) Derive the sampling distribution of the least squares estimator of β, i.e., assuming a linear probability model when the true model is a logit model.

(b) Derive the sampling distribution of the least squares residuals and verify the estimated variance of $\widehat{\beta}_{OLS}$ is biased.

10. *Sample Selection and Non-response.* This is based on Manski (1995), see the Appendix to this chapter. Suppose we are interested in estimating the probability than an individual who is homeless at a given date has a home six months later. Let $y = 1$ if the individual has a home six months later and $y = 0$ if the individual remains homeless. Let x be the sex of the individual and let $z = 1$ if the individual was located and interviewed and zero otherwise. 100 men and 31 women were initially interviewed. Six months later, only 64 men and 14 women were located and interviewed. Of the 64 men, 21 exited homelessness. Of the 14 women only 3 exited homelessness.

(a) Compute $\Pr[y = 1/\text{Male}, z = 1]$, $\Pr[z = 1/\text{Male}]$ and the bound on $\Pr[y = 1/\text{Male}]$.

(b) Compute $\Pr[y = 1/\text{Female}, z = 1]$, $\Pr[z = 1/\text{Female}]$ and the bound on $\Pr[y = 1/\text{Female}]$.

(c) Show that the width of the bound is equal to the probability of attrition. Which bound is tighter? Why?

11. *Does the Link Matter?* This is based on Santos Silva (1999). Consider a binary random variable Y_i such that

$$P(Y_i = 1|x) = F(\beta_0 + \beta_1 x_i), \qquad i = 1, \ldots, n,$$

where the link $F(\cdot)$ is a continuous distribution function.

(a) Write down the log-likelihood function and the first-order conditions of maximization with respect to β_0 and β_1.

(b) Consider the case where x_i only assumes two different values, without loss of generality, let it be 0 and 1. Show that $\widehat{F}(1) = \sum_{x_i=1} y_i/n_1$, where n_1 is the number of observations for which $x_i = 1$. Also, show that $\widehat{F}(0) = \sum_{x_i=0} y_i/(n-n_1)$.

(c) What are the maximum likelihood estimates of β_0 and β_1?

(d) Show that the value of the log-likelihood function evaluated at the maximum likelihood estimates of β_0 and β_1 is the same, independent of the form of the link function.

12. *Beer Taxes and Motor Vehicle Fatality.* Ruhm (1996) considered the effect of beer taxes and a variety of alcohol-control policies on motor vehicle fatality rates, see section 13.4. The data is for 48 states (excluding Alaska, Hawaii and the District of Columbia) over the period 1982–1988. This data set can be downloaded from the Stock and Watson (2003) web site at www.aw.com/stock_watson. Using this data set replicate the results in Table 13.1.

13. *Problem Drinking and Employment.* Mullahy and Sindelar (1996) considered the effect of problem drinking on employment and unemployment. The data set is based on the 1988 Alcohol Supplement of the National Health Interview Survey. This can be downloaded from the *Journal of Applied Econometrics* web site at http://qed.econ.queensu.ca/jae/2002-v17.4/terza/.

(a) Replicate the probit results in Table 13.6 and run also the logit and OLS regressions with robust White standard errors. The OLS results should match those given in Table 5 of Mullahy and Sindelar (1996).

(b) Compute marginal effects as reported in Table 13.7 and average marginal effects as reported in Table 13.8. Compute the classifications of actual vs predicted as reported in Table 13.9. Repeat these calculations for OLS and logit.

(c) Mullahy and Sindelar (1996) performed similar regressions for females and for the dependent variable taking the value of 1 if the individual is unemployed and zero otherwise. Replicate the OLS results in Tables 5 and 6 of Mullahy and Sindelar (1996) and perform the corresponding logit and probit regressions. Repeat part (b) for the female data set. What is your conclusion on the relationship between problem drinking and unemployment?

14. *Fractional Response.* Papke and Wooldridge (1996) studied the effect of match rates on participation rates in 401(K) pension plans. The data are from the 1987 IRS Form 5500 reports of pension plans with more than 100 participants. This data set can be downloaded from the *Journal of Applied Econometrics* web site at http://qed.econ.queensu.ca/jae/1996-V11.6/papke.

(a) Replicate Tables I and II of Papke and Wooldridge (1996).

(b) Run the specification tests (RESET) described in Papke and Wooldridge (1996).

(c) Compare OLS and logit QMLE using R^2, specification tests, and predictions for various values of MRATE as done in Figure 1 of Papke and Wooldridge (1996, p. 630).

15. *Fertility and Female Labor Supply.* Carrasco (2001) estimated a probit equation for fertility using PSID data over the period 1986-1989, see Table 13.10. The sample consists of 1,442 married or cohabiting women between the ages of 18 and 55 in 1986. The data set can be obtained from the *Journal of Business & Economic Statistics* archive data web site.

(a) Replicate Table 4, columns 1 and 2, of Carrasco (2001, p. 391). Show that having children of the same sex has a significant and positive effect on the probability of having an additional child.

(b) Compute the predicted probabilities from these regression and the percentage of correct decisions. Also compute the marginal effects.

(c) Replicate Table 5, columns 1 and 4, of Carrasco (2001, p. 392) which run a female labor force participation equation using OLS and probit. Compute the predicted probabilities from the probit regression and the percentage of correct decisions. Also compute the marginal effects.

(d) Replicate Table 5, column 5, of Carrasco (2001, p. 392) which runs 2sls on the female labor force participation equation using as instruments the same sex variables and their interactions with ags26l. Compute the over-identification test for this 2sls.

(e) Replicate Table 7, column 4, of Carrasco (2001, p. 393) which runs fixed effects on the female labor force participation equation with robust standard errors. Run also fixed effects 2sls using as instruments the same sex variables and their interactions with ags26l.

16. *Multinomial Logit.* Terza (2002) ran a multinomial logit model on the Mullahy and Sindelar (1996) data set for problem drinking described in problem 13, by explicitly accounting for the multinomial classification of the dependent variable. In particular, $y = 1$ when the individual is out of the labor force, $y = 2$ when this individual is unemployed, and $y = 3$ when this individual is employed.

   (a) Replicate the multinomial logit estimates for the male data reported in Table II of Terza (2002, p. 399) columns 3,4, 9 and 10.

   (b) Obtain the multinomial logit estimates for the female data. How does your conclusion on the relationship between problem drinking and employment/unemployment change from that in problem 13?

17. *Tobit Estimation of Married Women Labor Supply.* Wooldridge (2009, p. 593) estimated a Tobit equation for the Mroz (1987) data considered in problem 11.31. Using the PSID for 1975, Mroz's sample consists of 753 married white women between the ages of 30 and 60 in 1975, with 428 working at some time during the year. The wife's annual hours of work (*hours*) is regressed on nonwife income (*nwifeinc*); the wife's age (*age*), her years of schooling (*educ*), the number of children less than six years old in the household (*kidslt6*), and the number of children between the ages of five and nineteen (*kidsge6*). The data set was obtained from Wooldridge's (2009) data web site.

   (a) Give a detailed summary of hours of work and determine the extent of skewness and kurtosis.

   (b) Run OLS and Tobit estimation as described above and replicate Table 17.2 of Wooldridge (2009, p. 593).

   (c) Using the variable in the labor force (*inlf*), run OLS, logit and probit using the same explanatory variables given above and replicate Table 17.1 of Wooldridge (2009, p. 585).

   Give the predicted classification and the marginal effects at the mean as well as the average marginal effects for these three specifications.

   (d) Compare the estimates of (β/σ^2) from the probit in part (c) with the Tobit estimates given in part (b).

18. *Heckit Estimation of Married Women's Earnings.* Wooldridge (2009, p. 611) estimated a log wage equation for the Mroz (1987) data considered in problem 13.7. The wife's log wage (lwage) is regressed on her years of schooling (*educ*), her experience (*exper*) and its square (*expersq*). The probit equation to correct for sample selection includes the regressors plus the following additional variables: the number of children less than six years old in the household (*kidslt6*), the number of children between the ages of five and nineteen (*kidsge6*), nonwife income (*nwifeinc*) and the wife's age (*age*).

   (a) Run OLS and Heckit estimation as described above and replicate Table 17.5 of Wooldridge (2009, p. 611).

   (b) Test that the inverse Mills ratio is not significant. What do you conclude?

   (c) Run the MLE of this Heckman (1976) sample selection model.

References

This chapter is based on the material in Hanushek and Jackson (1977), Maddala (1983), Davidson and MacKinnon (1993) and Greene (1993). Additional references include:

Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19: 1481–1536.

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24: 3–61.

Baltagi, B.H. (2000), "Sampling Distribution of OLS Under a Logit Model," Problem 00.3.1, *Econometric Theory*, 16: 451.

Bera, A.K., C. Jarque and L.F. Lee (1984), "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic Review*, 25: 563–578.

Berkson, J. (1953), "A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function," *Journal of the American Statistical Association*, 48: 565–599.

Berndt, E., B. Hall, R. Hall and J. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 3/4: 653–665.

Boskin, M. (1974), "A Conditional Logit Model of Occupational Choice," *Journal of Political Economy*, 82: 389–398.

Carrasco, R. (2001), "Binary Choice with Binary Endogenous Regressors in Panel Data: Estimating the Effect Fertility on Female Labor Participation," *Journal of Business & Economic Statistics*, 19: 385–394.

Cornwell, C. and P. Rupert (1988), "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variables Estimators," *Journal of Applied Econometrics*, 3: 149–155.

Cox, D.R. (1970), *The Analysis of Binary Data* (Chapman and Hall: London).

Cragg, J. and R. Uhler (1970), "The Demand for Automobiles," *Canadian Journal of Economics,* 3: 386–406.

Davidson, R. and J. MacKinnon (1984), "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, 25: 241–262.

Dhillon, U.S., J.D. Shilling and C.F. Sirmans (1987), "Choosing Between Fixed and Adjustable Rate Mortgages," *Journal of Money, Credit and Banking*, 19: 260–267.

Effron, B. (1978), "Regression and ANOVA with Zero-One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73: 113–121.

Goldberger, A. (1964), *Econometric Theory* (Wiley: New York).

Gourieroux, C., A. Monfort and A. Trognon (1984), "Pseudo-Maximum Likelihood Methods: Theory," *Econometrica*, 52: 681–700.

Hanushek, E.A. and J.E. Jackson (1977), *Statistical Methods for Social Scientists* (Academic Press: New York).

Hausman, J. and D. McFadden (1984), "A Specification Test for Multinomial Logit Model," *Econometrica*, 52: 1219–1240.

Hausman, J.A. and D.A. Wise (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*, 45: 919–938.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, 5: 475–492.

Heckman, J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47: 153–161.

Lee, L.F. (2001), "Self-Selection," Chapter 18 in B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics* (Blackwell: Massachusetts).

Lee, L.F. and G.S. Maddala (1985), "The Common Structure of Tests for Selectivity Bias, Serial Correlation, Heteroskedasticity and Non-Normality in the Tobit Model," *International Economic Review*, 26: 1–20.

Lott, W.F. and S.C. Ray (1992), *Applied Econometrics: Problems with Data Sets* (The Dryden Press: New York).

Maddala, G. (1983), *Limited Dependent and Qualitative Variables in Econometrics* (Cambridge University Press: Cambridge).

Manski, C.F. (1995), *Identification Problems in the Social Sciences* (Harvard University Press: Cambridge).

McFadden, D. (1974), "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 3: 303–328.

McCullagh, P. and J.A. Nelder (1989), *Generalized Linear Models* (Chapman and Hall: New York).

Mroz, T.A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica,* 55: 765–799.

Mullahy, J. and J. Sindelar (1996), "Employment, Unemployment, and Problem Drinking," *Journal of Health Economics*, 15: 409–434.

Pagan, A.R. and F. Vella (1980), "Diagnostic Tests for Models Based on Individual Data: A Survey," *Journal of Applied Econometrics*, 4: S29-S59.

Papke, L.E. and J.M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with An Application to 401(K) Plan Participation Rates,"*Journal of Applied Econometrics*, 11: 619–632.

Powell, J. (1984), "Least Absolute Deviations Estimation of the Censored Regression Model," *Journal of Econometrics*, 25: 303–325.

Powell, J. (1986), "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica*, 54: 1435–1460.

Pratt, J.W. (1981), "Concavity of the Log-Likelihood," *Journal of the American Statistical Association*, 76: 103–109.

Ruhm, C.J. (1996), "Alcohol Policies and Highway Vehicle Fatalities," *Journal of Health Economics*, 15: 435–454.

Schmidt, P. and R. Strauss (1975), "Estimation of Models With Jointly Dependent Qualitative Variables: A Simultaneous Logit Approach," *Econometrica*, 43: 745–755.

Terza, J. (2002), "Alcohol Abuse and Employment: A Second Look," *Journal of Applied Econometrics*, 17: 393–404.

Wooldridge, J.M. (1991), "Specification Testing and Quasi-Maximum Likelihood Estimation," *Journal of Econometrics*, 48: 29–55.

Appendix

1. Truncated Normal Distribution

Let x be $N(\mu, \sigma^2)$, then for a constant c, the truncated density is given by

$$f(x/x > c) = \frac{f(x)}{\Pr[x > c]} = \frac{\frac{1}{\sigma}\phi([x-\mu]/\sigma)}{1 - \Phi\left(\frac{c-\mu}{\sigma}\right)} \quad c < x < \infty \tag{A.1}$$

where $\phi(z)$ denotes the p.d.f. and Φ denotes the c.d.f. of a $N(0,1)$ random variable. If the truncation is from above

$$f(x/x < c) = \frac{f(x)}{\Pr[x < c]} = \frac{\frac{1}{\sigma}\phi([x-\mu]/\sigma)}{\Phi\left(\frac{c-\mu}{\sigma}\right)} \quad -\infty < x < c \tag{A.2}$$

The conditional means are given by

$$E(x/x > c) = \mu + \sigma\frac{\phi(c^*)}{1 - \Phi(c^*)} \tag{A.3}$$

where $c^* = \frac{c-\mu}{\sigma}$, and

$$E(x/x < c) = \mu - \sigma\frac{\phi(c^*)}{\Phi(c^*)} \tag{A.4}$$

In other words, the truncated mean shifts to the right (left) if truncation is from below (above).

The conditional variances are given by $\sigma^2(1 - \delta(c^*))$ with $0 < \delta(c^*) < 1$ for all values of c^*.

$$\delta(c^*) = \frac{\phi(c^*)}{1-\Phi(c^*)}\left[\frac{\phi(c^*)}{1-\Phi(c^*)} - c^*\right] \quad \text{for} \quad x > c \tag{A.5}$$

$$= \frac{-\phi(c^*)}{\Phi(c^*)}\left[\frac{-\phi(c^*)}{\Phi(c^*)} - c^*\right] \quad \text{for} \quad x < c \tag{A.6}$$

In other words, the truncated variance is always less than the unconditional or untruncated variance. For more details, see Maddala (1983, p. 365) or Greene (1993, p. 685).

2. The Censored Normal Distribution

Let y^* be $N(\mu, \sigma^2)$, then for a constant c, define $y = y^*$ if $y^* > c$ and $y = c$ if $y^* < c$. Unlike the truncated normal density, the censored density assigns the entire probability of the censored region to the censoring point, i.e., $y = c$. So that $\Pr[y = c] = \Pr[y^* < c] = \Phi((c-\mu)/\sigma) = \Phi(c^*)$ where $c^* = (c-\mu)/\sigma$. For the uncensored region the probability of y^* remains the same and can be obtained from the normal density.

It is easy to show, see Greene (1993, p. 692) that

$$\begin{aligned} E(y) &= \Pr[y = c]E(y/y = c) + \Pr[y > c]E(y/y > c) \\ &= c\Phi(c^*) + (1 - \Phi(c^*))E(y^*/y^* > c) \\ &= c\Phi(c^*) + (1 - \Phi(c^*))\left[\mu + \sigma\frac{\phi(c^*)}{1-\Phi(c^*)}\right] \end{aligned} \tag{A.7}$$

where $E(y^*/y^* > c)$ is obtained from the mean of a truncated normal density, see (A.3).

Similarly, one can show, see problem 7 or Greene (1993, p. 693) that

$$\operatorname{var}(y) = \sigma^2 \left[1 - \Phi(c^*)\right] \left[1 - \delta(c^*) + \left(c^* - \frac{\phi(c^*)}{1 - \Phi(c^*)}\right)^2 \Phi(c^*)\right] \tag{A.8}$$

where $\delta(c^*)$ was defined in (A.5).

3. Sample Selection and Non-response

Non-response is a big problem plaguing survey data. Some individuals refuse to respond and some do not answer all the questions, especially on relevant economic variables like income. Suppose we interviewed randomly 150 individuals upon their graduation from high school. Among these, 50 were female and 100 were male. A year later, we try to re-interview these individuals to find out whether they are employed or not. Only 70 out of 100 males and 40 out of 50 females were located and interviewed a year later. Out of those re-interviewed, 60 males and 20 females were found to be employed. Let $y = 1$ if the individual is employed and zero if not. Let x be the sex of this individual and let $z = 1$ if this individual is located and interviewed a year later and zero otherwise.

Conditioning on sex of the respondent one can compute the probability of being employed a year after high school graduation as follows:

$$\Pr[y = 1/x] = \Pr[y = 1/x, z = 1]\ \Pr[z = 1/x] + \Pr[y = 1/x, z = 0]\ \Pr[z = 0/x]$$

In this case, $\Pr[y = 1/\text{Male},\ z = 1] = 60/70$, $\Pr[z = 1/\text{Male}] = 70/100$ and $\Pr[z = 0/\text{Male}] = 30/100$. But the sampling process is uninormative about the non-respondents or the censored observations, i.e., $\Pr[y = 1/\text{Male},\ z = 0]$. Therefore, in the absence of other information

$$\Pr[y = 1/\text{Male}] = (0.6) + (0.3)\ \Pr[y = 1/\text{Male},\ z = 0]$$

Manski (1995) argues that one can estimate bounds on this probability. In fact, replacing $0 \leq \Pr[y = 1/\text{Male},\ z = 0] \leq 1$ by its bounds, yields

$$0.6 \leq \Pr[y = 1/\text{Male}] \leq 0.9$$

with the width of the bound equal to the probability of non-response conditioning on Males, i.e., $\Pr[z = 0/\text{Male}] = 0.3$. Similarly, $0.4 \leq \Pr[y = 1/\text{Female}] \leq 0.6$ with the width of the bound equal to the probability of non-response conditioning on Females, i.e., $\Pr[z = 0/\text{Female}] = 10/50 = 0.2$. Manski (1995) argues that these bounds are informative and should be the starting point of empirical analysis. Researchers assuming that non-response is *ignorable* or *exogenous* are imposing the following restriction

$$\Pr[y = 1/\text{Male}, z = 1] = \Pr[y = 1/\text{Male}, z = 0] = \Pr[y = 1/\text{Male}] = 60/70$$
$$\Pr[y = 1/\text{Female}, z = 1] = \Pr[y = 1/\text{Female}, z = 0] = \Pr[y = 1/\text{Female}] = 20/40$$

To the extent that these probabilities are different casts doubt on the *ignorable* non-response assumption.