EC 327 PROBLEM SET 4

Prof. Baum, Mr. Dmitriev

17.1 Let m_0 denote the number (not the percent) correctly predicted when $y_i = 0$ (so the prediction is also zero) and let m_1 be the number correctly predicted when $y_i = 1$. Then the proportion correctly predicted is $(m_0 + m_1)/n$, where n is the sample size. By simple algebra, we can write this as $(n_0/n)(m_0/n_0) + (n_1/n)(m_1/n_1) = (1 - \overline{y})(m_0/n_0) + \overline{y}(m_1/n_1)$, where we have used the fact that $\overline{y} = n_1/n$ (the proportion of the sample with $y_i = 1$) and $1 - \overline{(y)} = n_0/n$ (the proportion of the sample with $y_i = 0$). But m_0/n_0 is the proportion correctly predicted when $y_i = 0$, and m_1/n_1 is the proportion correctly predicted when $y_i = 1$. Therefore, we have

$$m_0 + m_1)/n = (1 - \overline{y})(m_0/n_0) + \overline{y}(m_1/n_1).$$

If we multiply through by 100 we obtain

$$\hat{p} = (1 - \overline{y})\hat{q_0} + \overline{y}\hat{q_1},$$

where we use the fact that, by definition, $\hat{p} = 100[(m_0 + m_1)]$, $\hat{q_0} = 100(m_0/n_0)$, and $\hat{q_1} = 100(m_1/n_1)$.

(ii) We just use the formula part (i): $\hat{p} = 0.3(80) + .70(40) = 52$. Therefore, overall we correctly predict only 52 % of the outcomes. This is because, while 8% of the time we correctly predict y=0, $y_i = 0$ accounts for only 30% of the outcomes. More weight (.70) is given to the predictions when $y_i = 1$, and we do much less well predicting that outcome (getting it right only 40 % of the time).

17.2 We need to compute the estimated probability first at $hsGPA = 3.0$,$SAT = 1,200$, and $study = 10$ and subtract tis from the estimated probability at $hsGPA = 3.0$,$SAT = 1,200$, and $study = 5$. To obtain the first probability, we start by computing the linear function inside $\Lambda(.)$: $-1.17 + .24(3.0) + .00058(1,200) + .073(10) = .976$. Next, we plug this into the logit function: $\frac{\exp 0.976}{1+\exp 0.976} \approx .726$. This is the estimated probablity that a student-athlete with the given characteristics graduates in five years.

For student-athlete who attended study hall five hours a week, we compute $-1.17 + .24(3.0) + .00058(1,200) + .073(5) = .611$. Evaluating the logit function at this value gives $\exp(.611)/[1 + \exp(.611)] \approx .648$. Therefore, the difference in estimated probabilities is $.726 - .648 = .078$, or just under 0.10. [Note how far off the calculation would be if we simply use the coefficient on *study* to conclude that the difference in probabilities is $.073(10 - 5) = .365$]

17.6 (i) OLS will be unbiased, because we are coosing the sample on the basis of an exogenous explanatory variable. The population regression function for *sav* is the same as the regression function in the subpopulation with $age > 25$.

(ii) Assuming that marital status and number of children affect *sav* only through household size (*hhsize*), this is another example of exogenous sample selection. But, in the subpopulation of married people without children, $hhsize = 2$. Because the is no variation in *hhsize* in the subpopulation, we

would not be able to estimate]$beta_2$; effectively, the intercept in the subpopulation becomes $\beta_0 + 2\beta_2$ and that is all we can estimate. But, assuming there is variation in $inc, educ$, and age among married people without children (and that we have a sufficiently varied smple from this subpopulation), we can still estimate $\beta_1, \beta_3, \beta_4$.

(iii) This would be selecting the sample on the basis of the dependent variable, which caues OLS to be biased and inconsistent for estimating the β_j in the population model. We should instead use a truncated regression model.

C 17.2 (i) The probit estimates form *approve* on *white* are given in the following table:

```
. esttab, stats(r2 N, fmt(%9.3f %9.0g) )
──────────────────────────────
                      (1)
                  approve
──────────────────────────────
white              0.784***
                  (9.04)
_cons              0.547***
                  (7.25)
──────────────────────────────
r2
N                    1989
──────────────────────────────
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

As there is only one explanatory variable that takes on just two values, there are only two different predicted values: the estimated probabilities of loan approval for white and nonwhite applicants. Rounded to three decimal places these are .708 for nonwhites and .908 for whites. Without rounding errors, these are identical to the fitted values from the linear probability model. This must always be the case when the independent variables in a binary response model are mutually exclusive and exhaustive binary variables. Then, the predicted probabilities, whether we use the LPM, probit, or logit models, are simply the cell frequencies. (In other words, .708 is the proportion of loans approved for nonwhites and .908 is the proportion approved for whites.)

(ii) With the set of controls added, the probit estimate on *white* becomes about .520 (se$\approx$ .097). Therefore, there is still very strong evidence of dicrimination against nonwhites. We can divide this by 2.5 to make it roughly comparable to the LPM estimate in part (iii) of Computer Exercise C7.8: $.520/2/5 \approx .208$, compared with .129 in the LPM.

(iii) When we use logit instead of probit, the coefficient (standard error) on *white* becomes .938(.173).

(iv) Recall that, to make probit and logit estimates roughly comparable, we can multiply the logit estimates by 0.625. The scaled logit coefficient becomes .625(.938)$\approx$.586, which is reasonably close to the probit estimate. A better comparison would be to compare the predicted probabilities by setting the other controls at interesting values, such as their average values in the sample.

C17.3 (i) Out of 616 workers, 172, or about 18%, have zero pension benefits.

For the 444 workers reporting positive pension benefits, the range is from $7.28 to $2,880.27. Therefore, we have a nontrivial fraction of the sample with pensiont = 0, and the range of positive pension benefits is fairly wide. The Tobit model is well-suited to this kind of dependent variable.

(ii) The Tobit results are given in the following table:

Dependent variable: Pension

Independent variable	(1)	(2)
exper	5.20 (6.01)	4.39 (5.83)
age	-4.64(5.71)	-1.65(5.56)
tenure	36.02(4.56)	28.78 (4.50)
educ	93.21(10.89)	106.83(10.77)
depends	35.28(21.92)	41.47(21.21)
married	53.69(71.73)	19.75(69.50)
white	144.09 (102.08)	159.30(98.97)
male	308.15(69.89)	257.25(68.02)
union	———-	439.05(62.49)
constant	-1,252.43(219.07)	-1,571.51(218.54)
Number of Observations	616	616
Log likelihood value	-3,672.96	-3648.55
se	677.74	652.90

In column (1), which does not control for union, being white or male (or, of course, both) increases predicted pension benefits, although only male is statistically significant (t 4.41). =

(iii) We use equation (17.22) with exper = tenure = 10, age = 35, educ = 16, depends = 0, married = 0, white = 1, and male = 1 to estimate the expected benefit for a white male with the given characteristics. Using our shorthand, we have

xB = -1,252.5 + 5.20(10) – 4.64(35) + 36.02(10) + 93.21(16) + 144.09 + 308.15 = 940.90.Therefore, with se = 677.74 we estimate E(pension — x) as

Q(940.9/677.74).(940.9) + (677.74).(940.9/677.74) = 966.40.

For a nonwhite female with the same characteristics,

xB = -1,252.5 + 5.20(10) – 4.64(35) + 36.02(10) + 93.21(16) = 488.66.

Therefore, her predicted pension benefit is

Q(488.66/677.74).(488.66) + (677.74).(488.66/677.74) = 582.10.

The difference between the white male and nonwhite female is 966.40 – 582.10 = $384.30.

[Note: If we had just done a linear regression, we would add the coefficients on white and male to obtain the estimated difference. We get about 114.94 + 272.95 = 387.89, which is very close to the Tobit estimate. Provided that we focus on partial effects, Tobit and a linear model often give similar answers for explanatory variables near the mean values.]

(iv) Column (2) in the previous table gives the results with union added. The coefficient is large, but to see exactly how large, we should use equation (17.22) to estimate E(pension — x) with union = 1 and union = 0, setting the other explanatory variables at interesting values. The t statistic on union is over seven.

(v) When peratio is used as the dependent variable in the Tobit model, white and male are individually and jointly insignificant. The p-value for the test of joint significance is about .74. Therefore, neither whites nor males seem to have different tastes for pension benefits as a fraction of earnings. White males have higher pension benefits because they have, on average, higher earnings.

C17.6

```
. esttab, stats(r2 N, fmt(%9.3f %9.0g) )
──────────────────────────────
                      (1)
                    ldurat
──────────────────────────────
workprg           0.00876
                   (0.18)
priors            -0.0591***
                   (-6.44)
tserved          -0.00940***
                   (-7.23)
felon               0.179**
                    (3.06)
alcohol            -0.263***
                   (-4.39)
drugs             -0.0907
                   (-1.65)
black              -0.179***
                   (-3.78)
married             0.134*
                    (2.43)
educ              0.00539
                    (0.54)
age               0.00133***
                    (5.90)
_cons               3.569***
                   (25.87)
──────────────────────────────
r2                    0.109
N                      1445
──────────────────────────────
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

The results of an OLS regression using only the uncensored durations are given in the following table:

There are several important differences between the OLS estimates using the uncensored durations and the estimates from the censored regression in Table 17.4. For example, the binary indicator for drug usage, *drugs*, has become positive and insignificant, whereas it was negative (as we expect) and significant in Table 17.4. On the other hand, the work program dummy, *workprg*, becomes positive but is still insignificant. The remaining coefficients maintain the same sign, but they are all attenuated toward zero. The apparent attenuation bias of OLS for the coefficient on *black* is especially severe, where the estimate changes from -.543 in the (appropriate) censored regression estimation to -.00085 in the

(inappropriate) OLS regression using only the uncensored durations.

C17.9 (i) 248.

(ii) The distribution is not continuous: there are clear focal points, and rounding. For example, many more people report one pound than either two-thirds of a pound or 1 1/3 pounds. This violates the latent variable formulation underlying the Tobit model, where the latent error has a normal distribution. Nevertheless, we should view Tobit in this context as a way to possibly improve functional form. It may work better than the linear model for estimating the expected demand function.

(iii) The following table contains the Tobit estimates and, for later comparison, OLS estimates of a linear model:

```
. esttab, stats(r2 N, fmt(%9.3f %9.0g) ) mtitles (ecolbs-tobit ecolbs-ols)
```

	(1) ecolbs-tobit	(2) ecolbs-ols
main		
ecoprc	-5.821***	-2.903***
	(-6.57)	(-4.94)
regprc	5.655***	3.031***
	(5.31)	(4.26)
faminc	0.00664	0.00283
	(1.66)	(1.04)
hhsize	0.130	0.0537
	(1.37)	(0.84)
_cons	1.003	1.630***
	(1.50)	(3.62)
sigma		
_cons	3.441***	
	(27.18)	
r2		0.039
N	660	660

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Only the price variables, *ecoprc* and *regprc*, are statistically significant at the 1% level.

(iv) The signs of the price coefficients accord with basic demand theory: the own-price effect is negative, the cross price effect for the substitute good (regular apples) is positive.

(v) The null hypothesis can be stated as $H_0 : \beta_1 + \beta_2 = 0$. Define $\Theta_1 = \beta_1 + \beta_2$. Then $\hat{\Theta}_1 = -.16$. To obtain the t statistic, I write $\beta_2 = \Theta_1 - \beta_1$, plug in, and rearrange. This results in doing Tobit of *ecolbs* on $(ecoprc - regprc), regprc, faminc$, and $hhsize$. The coefficient on $regprc$ is $\hat{\Theta}_1$ and, of course we get its standard error: about .59. Therefore, the t statistic is about -.27 and p-value=.78. We do not reject the null.

(vi) The smallest fitted value is.798, while the largest is 3.327.

(vii) The squared correlation between $ecolbs_i$ and $\hat{ecolbs}_i$ is about .0369. This is one possible R-squared measure.

(viii) The linear model estimates are given in the table for part (ii). The OLS estimates are smaller than the Tobit estimates because of the OLS estimates are estimating partial effects on $E(ecolbs|x)$, whereas the Tobit coefficients must be scaled by the term in equation (17.27). The scaling factor is always between zero and one, and often substantially less than one. The Tobit model does not fit better, at least in terms of estmating $E(ecolbs|x)$:the linear model R-squared is a bit larger (.0393 versus .0369).

(ix) This is not a correct statement. We have another case where we have confidence in the ceteris paribus price effects (because the price variables are exogenously set), yet we cannot explain much of the variation in *ecolbs*. The fact that demand for a fictitious product is hard to explain is not very surprising. [Instructors Notes: This might be a good place to remind students about basic economics. You can ask them whether *reglbs* should be included as an additional explanatory variable in the demand equation for *ecolbs*, making the point that the resulting equation would no longer be a demand equation. In other words, *reglbs* and *ecolbs* are jointly determined, but it is not appropriate to write each as a function of the other. You could have the students compute heteroskedasticity-robust standard errors for the OLS estimates. Also, you could have them estimate a probit model for $ecolbs = 0$ versus $ecolbs > 0$, and have them compare the scaled Tobit slope estimates with the probit estimates.]

C17.13 (i) Using the entire sample, the estimated coefficient on *educ* is .1037 with standard error = .0097. (ii) 166 observations are lost when we restrict attention to the sample with $educ < 16$. This is about 13.5% of the original sample. The coefficient on *educ* becomes .1182 with standard error = .0126. This is a slight increase in the estimated return to education, and it is estimated less precisely (because we have reduced the sample variation in *educ*). (iii) If we restrict attention to those with wage < 20, we lose 164 observations [about the same number in part (ii)]. But now the coefficient on *educ* is much smaller, .0579, with standard error = .0093. (iv) If we use the sample in part (iii) but account for the known truncation point, log(20), the coefficient on *educ* is .1060 (standard error = .0168). This is very close to the estimate on the original sample. We obtain a less precise estimate because we have dropped 13.3% of the original sample.