Alex Miller

Problem Set 5

1.

(a)

An OLS regression performed to estimate the relationship between medical expenditures and medical insurance, controlling for several other factors. This relationship was modeled with the following equation:

*(Log medical expenses)= β0 + β1\*(insurance dummy) + β2\*(number of chronic illnesses) + β3\*(age \* 1/10) + β4\*(female dummy) + β5\*(education) + β6\*(black or Hispanic dummy)+ u*

The log of medical expenses was chosen as the dependent variable because it more closely approximates a normal distribution as evidenced in the histograms below. Since there are several observations with zero medical expenses, however, OLS will automatically drop them because the log transformation returns missing. This may be an issue if we are starting with a representative sample and we want generalizable results. In other words, the results of OLS will only be applicable to individuals with positive medical costs to begin with. This could be considered a censoring problem since medical expenses cannot be negative, but it could also be a sample selection problem, where those with zero expenses are not being observed because of some other process. The OLS regression, some data exploration, and the following histograms were produced with this code in STATA:

. sum \*

. tab dambexp ins, nofreq column

. reg lambexp ins totchr age female educ blhisp

The cross-tabulation showed that there may be some positive correlation between having insurance and having zero costs, so leaving out the zeroes would bias any coefficient on insurance. The output of the OLS and histograms are as follows:

**Source | SS df MS Number of obs = 2802**

**-------------+-------------------------- F( 6, 2795) = 110.58**

**Model | 1069.37332 6 178.23 Prob > F = 0.0000**

**Residual | 4505.06629 2795 1.6118 R-squared = 0.1918**

**-------------+-------------------------- Adj R-squared = 0.1901**

**Total | 5574.43961 2801 1.9902 Root MSE = 1.2696**

**----------------------------------------------------------------**

**lambexp | Coef. Std. Err. t P>|t|**

**-------------+--------------------------------------------------**

**ins | -.020827 .0500062 -0.42 0.677**

**totchr | .5618171 .0305078 18.42 0.000**

**age | .2172327 .0222225 9.78 0.000**

**female | .3793756 .0485772 7.81 0.000**

**educ | .0222388 .0097615 2.28 0.023**

**blhisp | -.2385321 .0551952 -4.32 0.000**

**\_cons | 4.907825 .1681512 29.19 0.000**

**----------------------------------------------------------------**

|  |  |
| --- | --- |
| **Histogram of medical expenses with superimposed normal distribution** | **Histogram of logged medical expenses with superimposed normal distribution** |
| hist_ambexp.jpg | hist_lambexp.jpg |

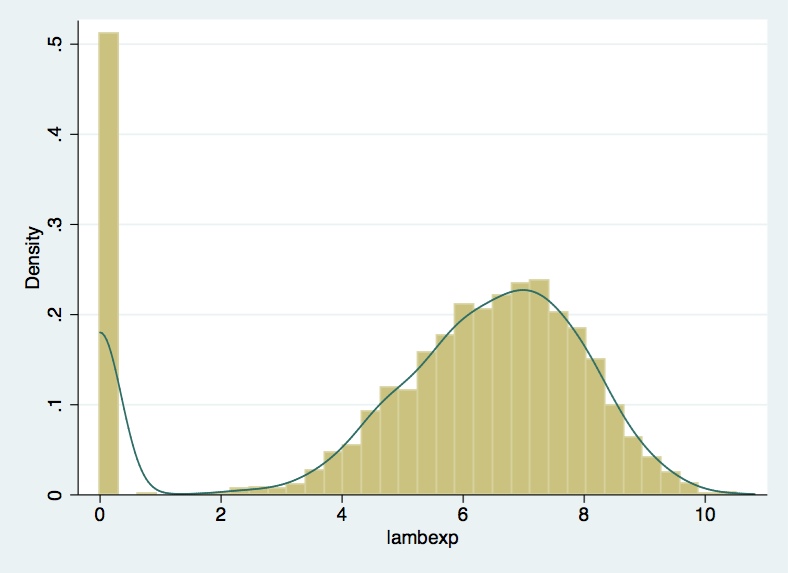
(b)

The Tobit model can only handle censored data if it is actually recoded as something and not just missing, so before we can perform another regression, we must first recode the missing values of the log of medical expenses to something just below the minimum observation (and something above zero). Since the minimum value for medical costs besides zero was one, the zeroes were recoded as 0.99999, and the missing values in the log transformation were recoded as log(0.99999) or about -0.00001. This was done with the following commands:

. quietly sum ambexp if ambexp>0

. scalar lbound=r(min)

. replace lambexp = log(lbound-0.00001) if lambexp==.

The following is a histogram of the new logged medical expense variable with a superimposed fit line:

The data still appears normally distributed from 0 up, but the probability stacked on -0.00001 makes it the clear point at which this new transformed data is censored. The probability of not being censored in a lower limit case is equal to (1- Φ((a – μ)/ σ)), where Φ is the cumulative density function, a is the censoring point, μ is the mean, and σ is the standard error. The cumulative density function is equal to the first derivative of the probability density function, which the superimposed line gives a rough idea of for this data. Therefore, all of the area under this curve, from any number greater than -0.00001 to positive infinity, is the probability of being selected and not censored. As it stands, 526 observations in this dataset are being censored.

(c)

The formula for the expected value of health care expenditures in a Tobit framework is as follows:

E(yi | xi) = (1 – Φ(a-xiβ/σ))(xiβ + σ \* ϕ(a-xiβ/σ)/(1- Φ(a-xiβ/σ))

Where ϕ is the probability density function, β is the set of coefficients that maximize the log-likelihood function, and the rest of the symbols are the same as above. In this case, a will be -0.00001. In this equation, the first half represents the probability of being selected, and the second half represents the expected value. The inverse Mills ratio, “σ \* ϕ(a-xiβ/σ)/(1- Φ(a-xiβ/σ)” comes into play in the second half as a correction factor for the observed mean.

(d)

The inverse Mills ratio for the sample was calculated with the following commands in STATA:

. tobit lambexp ins totchr age female educ blhisp, ll(-0.00001)

. gen sigma = [sigma]\_b[\_cons]

. gen logLike=e(ll)

. predict ystar, xb

. gen b = (-0.00001-ystar)/sigma

. gen probnotcensored\_m = 1-normal(b)

. gen inversemills = sigma\*(normalden(b)/(1-normal(b)))

. gen ycond\_m = ystar + inversemills

. list ambexp lambexp inversemills in 1/10

The output of these commands, including the results of the Tobit model, is as follows:

**Tobit regression Number of obs = 3328**

**LR chi2(6) = 831.03**

**Prob > chi2 = 0.0000**

**Log likelihood = -7494.2937 Pseudo R2 = 0.0525**

**----------------------------------------------------------------**

**lambexp | Coef. Std. Err. t P>|t|**

**-------------+--------------------------------------------------**

**ins | .2612206 .1026131 2.55 0.011**

**totchr | 1.161269 .0649656 17.88 0.000**

**age | .3630701 .0453222 8.01 0.000**

**female | 1.34181 .0986075 13.61 0.000**

**educ | .1384462 .0196568 7.04 0.000**

**blhisp | -.8731621 .1102505 -7.92 0.000**

**\_cons | .9237114 .3350347 2.76 0.006**

**-------------+--------------------------------------------------**

**/sigma | 2.781238 .0392269**

**----------------------------------------------------------------**

**Obs. summary: 526 left-censored at lambexp<=-.00001**

**2802 uncensored observations**

**0 right-censored observations**

**+-----------------------------------+**

**| ambexp lambexp inversemills |**

**|-----------------------------------|**

**1. | 760 6.633318 .1128537 |**

**2. | 497 6.20859 .8110353 |**

**3. | 1002 6.909753 .1421172 |**

**4. | 745 6.613384 .1632737 |**

**5. | 2728 7.911324 .0866 |**

**|-----------------------------------|**

**6. | 636 6.455199 .4442371 |**

**7. | 2 .6931472 .1773275 |**

**8. | 0 -.00001 .3131925 |**

**9. | 0 -.00001 .2669132 |**

**10. | 2826 7.946618 .1493369 |**

**+-----------------------------------+**

As the observed value of medical expenditures approaches the censoring point, the inverse Mills ratio rises. Similarly, as the censoring point for a lower-limit case increases, the inverse Mills ratio will increase in order to account for the shift in the observed mean away from the mean of the full (unobserved) sample.

(e)

The varying marginal effects for the Tobit model described above were produced by running the following commands immediately after the Tobit regression:

. mfx compute, predict(e(-0.00001,.)) eyex

. mfx compute, predict(pr(-0.00001,.)) eyex

. mfx compute, predict(ystar(-0.00001,.)) eyex

. mfx, eyex

The results of these commands are represented in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) |
|  | **Marginal effect on truncated mean** | **Marginal effect on probability of being observed** | **Conditional marginal effect** | **Unconditional marginal effect** |
|  |  |  |  |  |
| ins | 0.0151\* | 0.00226\* | 0.0173\* | 0.0179\* |
|  | (0.00592) | (0.000896) | (0.00681) | (0.00704) |
|  |  |  |  |  |
| totchr | 0.0887\*\*\* | 0.0133\*\*\* | 0.102\*\*\* | 0.105\*\*\* |
|  | (0.00500) | (0.00106) | (0.00576) | (0.00599) |
|  |  |  |  |  |
| age | 0.233\*\*\* | 0.0348\*\*\* | 0.268\*\*\* | 0.277\*\*\* |
|  | (0.0291) | (0.00478) | (0.0335) | (0.0347) |
|  |  |  |  |  |
| female | 0.108\*\*\* | 0.0161\*\*\* | 0.124\*\*\* | 0.128\*\*\* |
|  | (0.00795) | (0.00151) | (0.00916) | (0.00951) |
|  |  |  |  |  |
| educ | 0.293\*\*\* | 0.0439\*\*\* | 0.337\*\*\* | 0.349\*\*\* |
|  | (0.0417) | (0.00674) | (0.0480) | (0.0497) |
|  |  |  |  |  |
| blhisp | -0.0426\*\*\* | -0.00637\*\*\* | -0.0489\*\*\* | -0.0506\*\*\* |
|  | (0.00538) | (0.000886) | (0.00619) | (0.00642) |
|  |  |  |  |  |
| N | 3328 | 3328 | 3328 | 3328 |
|  |  |  |  |  |
| Note: Standard errors in parentheses. Marginal effect on dummy variable represents a discrete change of dummy variable from 0 to 1. Significance is denoted by: \* p<0.05, \*\* p<0.01, \*\*\* p<0.001. | | | | |

The first marginal effect represents the elasticity of E[Y|Y>a] (expected value of y given y is greater than our censoring point) with respect to changes in X. By far the largest effects are age and education. The second marginal effect represents the elasticity of the probability of Prob(Y>a) (probability of being selected) with respect to changes in X. Again, age and education are large and significant, but each variable’s marginal contribution to the probability of being selected is much smaller than its contribution to the expected value once selected. The third marginal effect represents the elasticity of E(Y) given X. This is the most intuitive marginal effect, since it uses the expected value conditional on the censoring and will conform to whatever the real-life censoring rule is. Finally, we have the elasticity of the unconditional expected value with respect to X. This would be the expected value if we could observe beyond the censoring point.

(f)

The log-likelihood estimation was manually calculated with the following commands:

. gen d\_censored = (lambexp<=-0.00001)

. gen contrib\_censored = (1-probnotcensored\_m)^(d\_censored)

. gen contrib\_not\_censored = ((1/sigma)\*normalden((lambexp-ystar)/sigma))^(1-d\_censored)

. gen Like\_i\_m = contrib\_censored\*contrib\_not\_censored

. gen logLike\_i\_m = log(Like\_i\_m)

. egen logLike\_m = sum(logLike\_i\_m)

. sum logLike\_m logLike

These commands returned a log-likelihood value of -7494.293, exactly equal to the one calculated automatically by STATA.

2.

As stated above, using OLS for this type of data will created biased results when making inference about the population as a whole. By excluding those who have not paid money for medical costs, you are excluding an interesting, large, and non-random portion of the sample. The Tobit model fairs alright as long as we can assume that the selection model is the same as the model that determines amounts. It is difficult to think of a scenario where there is another model driving whether someone incurs medical costs or not, independent of the model that determines how many medical costs they incur. In this scenario, the Tobit model did require a transformation of the data, but that was just due to the nature of the distribution and log transformations. Finally, the Heckman model allows for there to be different processes for the selection and the amount. Especially with the limited number of independent variables to choose from in this dataset, it would have been difficult to make an argument for why one variable would affect whether someone incurs medical costs at all, and then go on to make an argument for why it wouldn’t also belong in the amount setting equation. Without such a variable, the Heckman model would not be suitable.