MLE: Categorical and Limited Dependent Variables

Unit 2: Ordered, Multinomial, Count, and Limited Dependent Variables

4. Models for Censored Outcomes

PS2730-2021

Week 9

Professor Steven Finkel

Limited Dependent Variables: (Continued)

- So far concerned with one kind of limited dependent variable – non-negative integer *count outcomes* which represent the number of times something occurs in a given time interval
- Poisson, Negative Binomial, Zero-Inflated depending on the distribution, the degree of overdispersion, and the assumed data-generating process for the 0s
- There are other kinds of limited dependent variables, however
- Consider: situations where we have a linear relationship between X and a *continuous* Y* (latent variable) – unlike the count situation where the true dependent variable is non-continuous – but we do not observe all of Y* for some reason or another
- That is, Y (observed) is not the complete Y* but some non-continuous portion of it
- Most prominent examples: *censored* data and *sample-selected* data

- Censored Data
  - We observe the true Y* for some values, but for other values that don’t make it past some threshold (τ) you only observe one value, usually the threshold value itself
  - Example: Student Final Exams. Failure is 60 and you don’t record anything below that. So you have a bunch of 60s which really represents any value below 60. Anything above 60 receives the true value, below 60 receives a 60. This is a *censored* variable, with the censoring value being the threshold below which true values are not recorded.
  - This is called “censoring from below”; there can also be “censoring from above” (demand for concert tickets, e.g.), or both
  - Thermometer ratings -- people at 0 feel various degrees of extreme coldness toward the stimulus, people at 100 feel various degrees of extreme warmness, so have a double threshold model here, censored at 0 from below and 100 from above

- In all of these cases, there is a Y* which is continuous but an observed Y which is not identical to Y*, due to some aspect of the measurement scheme or some aspect of "reality" (the capacity of the stadium, e.g.) which prevents the full range of Y* from being realized

- Note: this is related but different from the latent variable Y* situation in probit. In probit we had a continuous Y* but we only observed 0 or 1 depending on whether the case made it past the threshold τ, which we arbitrarily set to 0. Here we know the full Y* if the case makes it past τ (if there is censoring from below), and the value of the remaining cases which do not make it past the threshold are usually set at τ, which can be 0 or may not be (if it is 0 the math is easier, though!)

- Second way that the observation of Y* might be limited: we only observe Y* if it is above a threshold on **some other variable**, not Y* itself. In censored example, if the exam grade is >60, we observe the exact score, and if not, we observe 60. In another kind of situation, we observe, for example, reported participation in politics in a survey **only** if the respondent decided to answer the survey. So there are a set of factors that lead individuals to be **selected into the observation sample**, and **only** if those factors put you over a threshold τ do we observe Y* in the data. Otherwise we don't observe a value for Y
- This is called **"sample-selected"** data.
- Example: You want to model how education affects voter turnout in the US, and examine a sample of individuals who are registered to vote. Since registered individuals are already much more likely to vote than average individuals, *and* since registration depends on education and a host of other factors that may be related to education, the estimation of turnout on education in the registered sample will not represent the true effect of education on turnout

- Can see the problem here: the observation sample Y is a potentially non-random sample of Y* in the population, so estimating the effect of any independent variable X on Y in the **sample** will lead to potentially biased estimates of the effect of X on Y* in the population (which is our true goal)
- The sample-selection problem relates more generally to problem of estimating treatment effects when the assignment to the treatment is not random. This is a common situation in social science.
- Example: Estimating the effect of accepting public funding on state legislative candidates' success – but candidates who accept public funding are weaker and would not succeed as much, independent of funding – So if we regress outcome against public funding, we obtain a negative relationship but **not** due to public funding.
- Will start with censoring where don't observe Y* in all cases, then relate them **in Unit 4** to treatment effect models where we have all Y* but with non-random selection into treatment

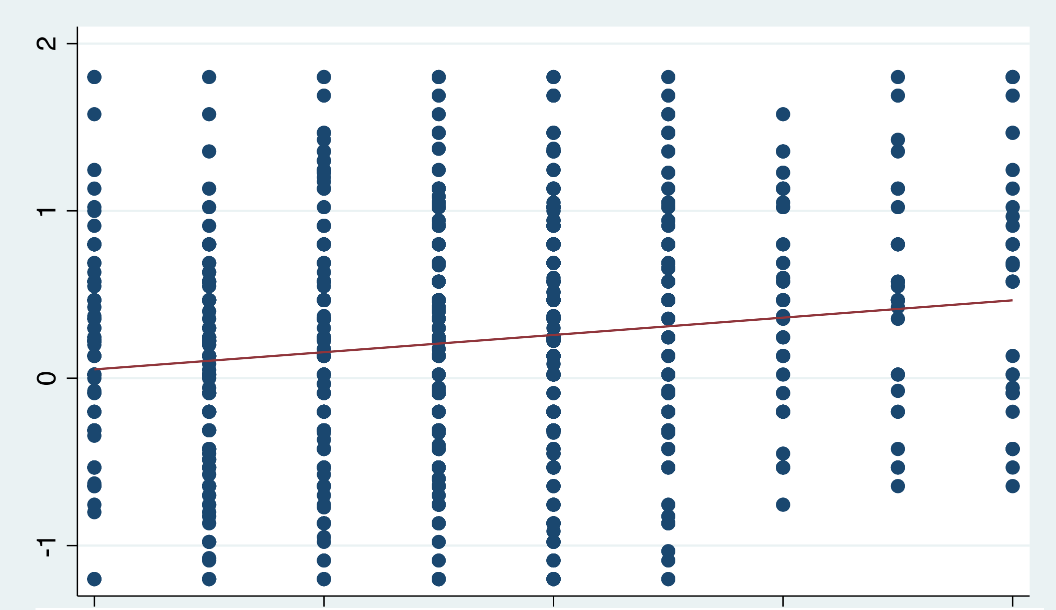
- With censoring (or sample-selection), using OLS to estimate effects of X on Y will lead to biased inferences whenever factors of interest are related to the probability of censoring or of being in the observation sample.

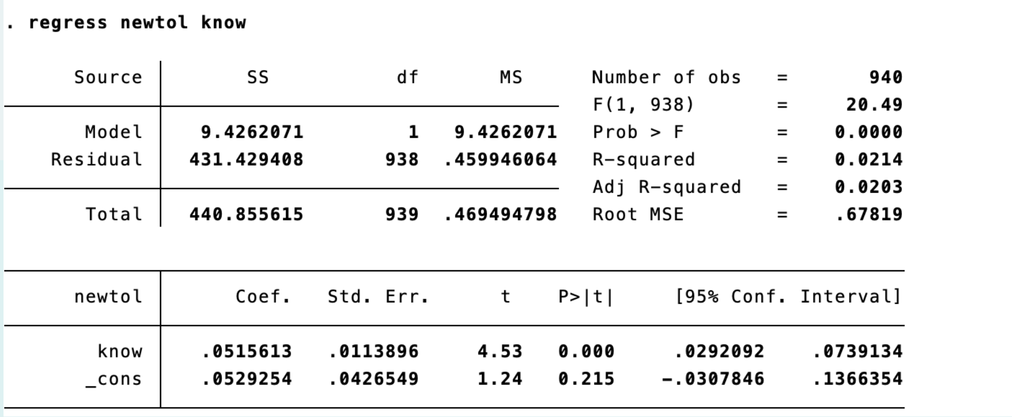
- Solution: model the two stages of the process. For censored data: probability of being censored is one stage, and the value of Y, conditional on not being censored is the second stage.

- (For sample-selected data, probability of being observed on Y is one stage, and modeling Y, conditional on having been observed, is the second stage.)

Modeling Censored Data

- Example: The true relationship between political knowledge (X) and political tolerance (Y*) in the South African data



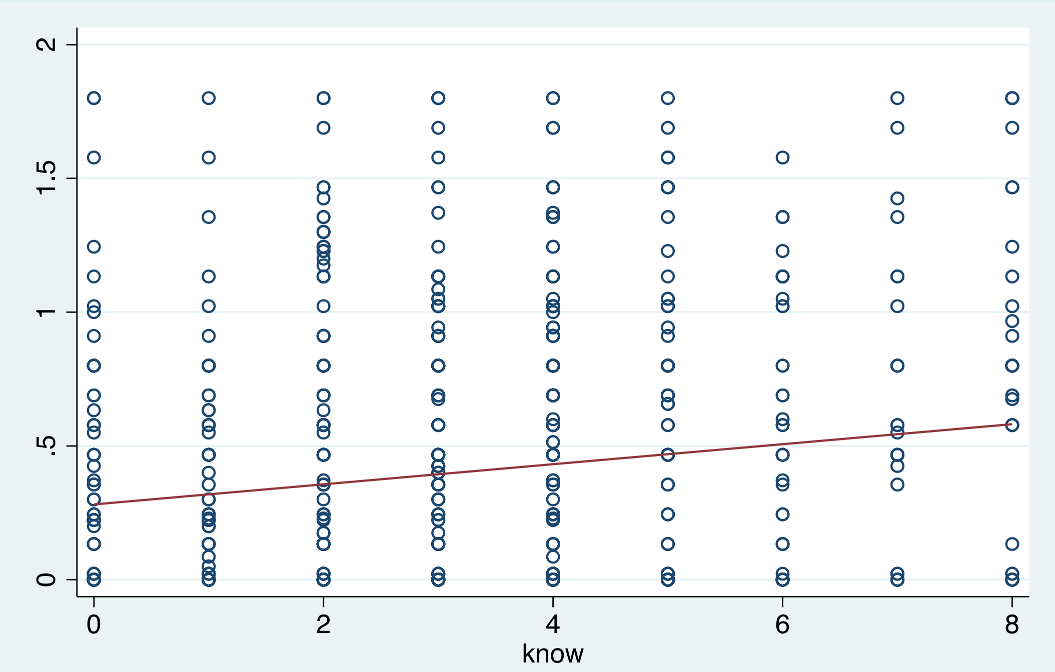


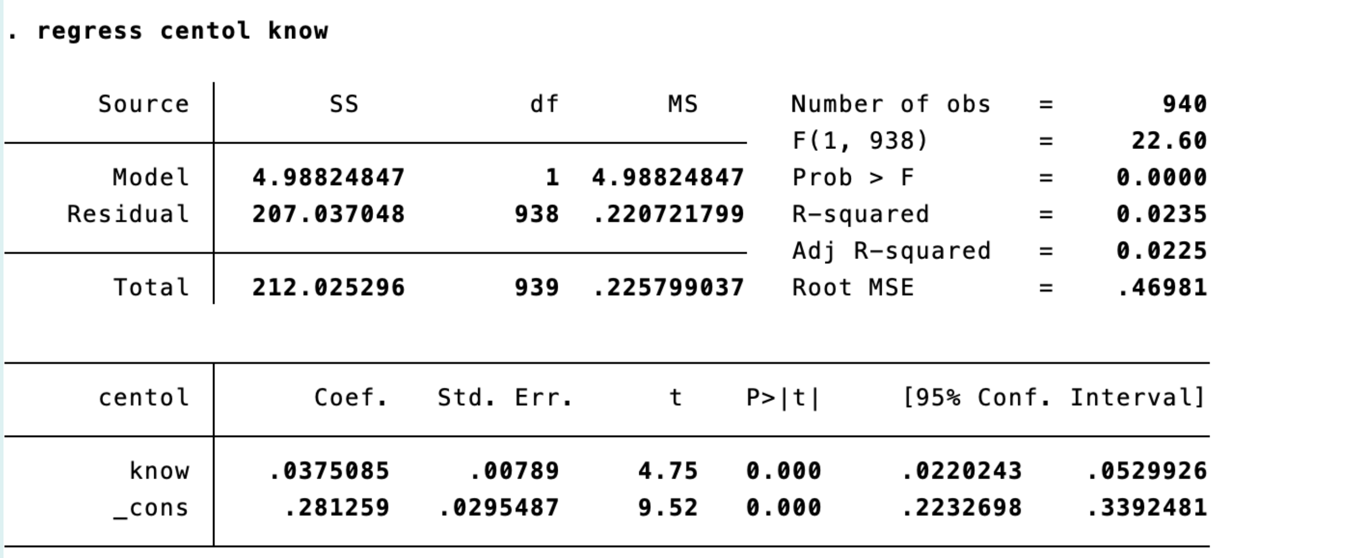
```
. regress newtol know

      Source |       SS           df       MS      Number of obs   =       940
-------------+----------------------------------   F(1, 938)       =     20.49
       Model |   9.4262071         1   9.4262071   Prob > F        =    0.0000
    Residual |  431.429408       938  .459946064   R-squared       =    0.0214
-------------+----------------------------------   Adj R-squared   =    0.0203
       Total |  440.855615       939  .469494798   Root MSE        =    .67819

------------------------------------------------------------------------------
      newtol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .0515613   .0113896     4.53   0.000     .0292092    .0739134
       _cons |   .0529254   .0426549     1.24   0.215    -.0307846    .1366354
------------------------------------------------------------------------------
```

- What if the lowest recorded value of tolerance was 0? Anybody who scored lower than 0 on the scale during the interview was coded as 0 in the data set (for whatever reason)

- Then we have 361, or 38.4% of all observations censored from below at 0

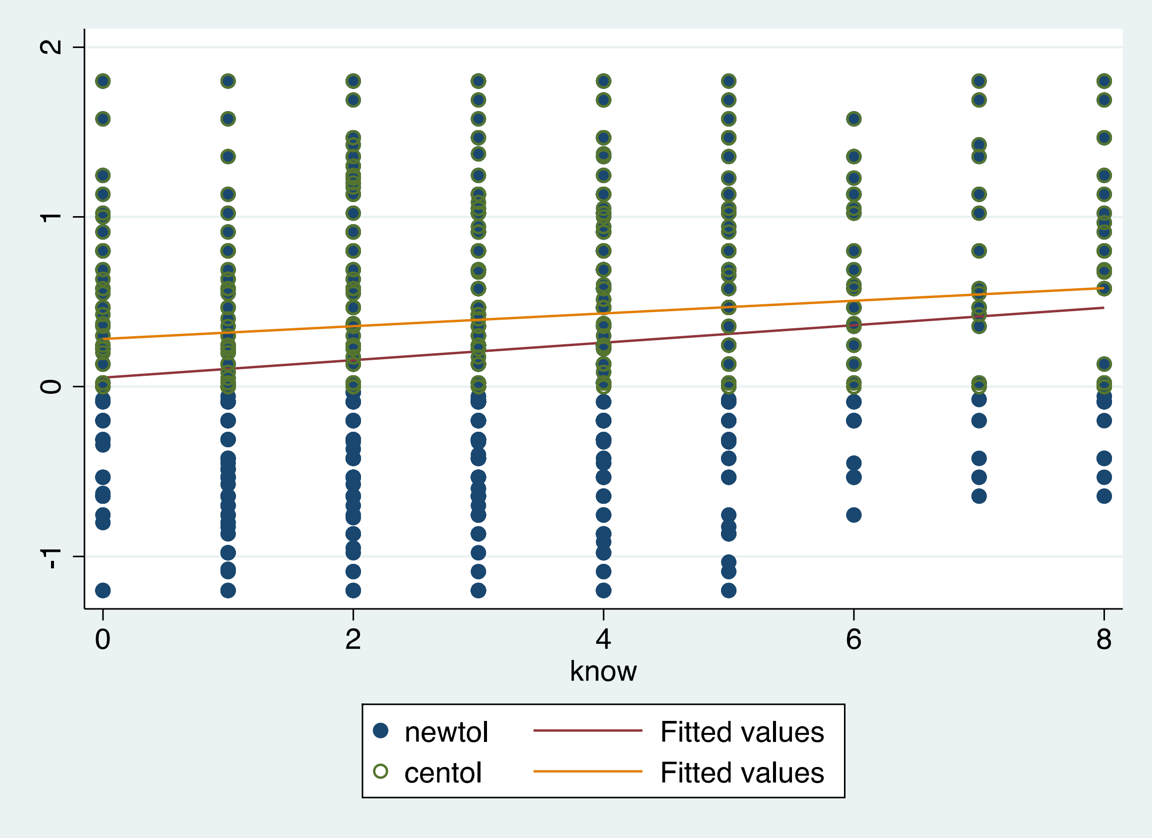
- For the remaining 579 cases, censored tolerance=tolerance, or Y*





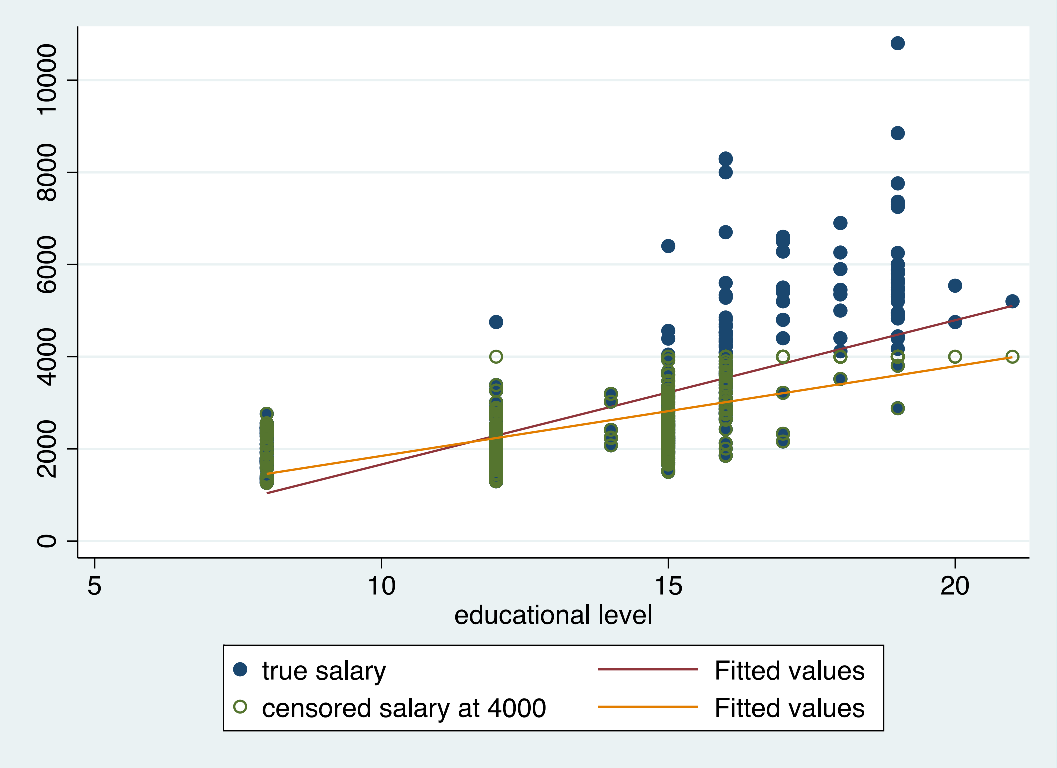
```
. regress centol know

      Source |       SS           df       MS      Number of obs   =       940
-------------+----------------------------------   F(1, 938)       =     22.60
       Model |  4.98824847         1  4.98824847   Prob > F        =    0.0000
    Residual |  207.037048       938  .220721799   R-squared       =    0.0235
-------------+----------------------------------   Adj R-squared   =    0.0225
       Total |  212.025296       939  .225799037   Root MSE        =    .46981

------------------------------------------------------------------------------
      centol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .0375085     .00789     4.75   0.000     .0220243    .0529926
       _cons |    .281259   .0295487     9.52   0.000     .2232698    .3392481
------------------------------------------------------------------------------
```

- What happened?
- Censored data put a "floor" on the dependent variable at 0, so OLS tilted to be flatter. Regression coefficient went from .052 to .038, a 27% decrease in magnitude!



Red=Y*
Orange=Censored Y*, Observed Y

- Can also see the effects of censoring from above. Consider a salary data set where the salaries are recorded up to $4000 and then only $4000 for all salary at that point or above. We regress education on Y* (true salary) and then education on Y (censored salary)



Same issue: regression coefficient for true salary is 312.7; for censored salary it is 194.5

- Problem in first graph: Low X is positively related to the probability of being censored; so low knowledge individuals are more likely to not have Y^* recorded and to see 0 instead; so the effect of X on censored Y is dampened, since low X has *even lower* Y^* than was observed as Y
- More formally:
  - X is negatively related to P(censored)
  - OLS on censored variable will be linking $E(Y|X)$ which is not $E(Y^*|X)$
  - Since X figures (negatively) in the censoring process, $E(Y|X)$ will be greater at low levels of X than the true $E(Y^*|X)$ and this will push the regression line upwards at low levels of X, thus dampening the β regression coefficient
  - Suggests we need to find a way to “add back” the difference between $E(Y^*|X)$ and $E(Y|X)$ to the estimation

- Problem in second graph: high X is positively related to the probability of being censored; high educated individuals are more like to not have Y* recorded and to see 4000 instead; so the effect of X on censored Y is dampened, since high X should see *even higher* Y* than was observed as Y
- So E(Y|X) at high levels of X will be lower than E(Y*|X); this also dampens the β effect of X on Y compared to the true effect of X on Y*
- Solution: think of the process as two-stages -- model the probability of being censored, and then model Y* among the non-censored cases.
- But we need to take the first stage into account to get correct estimates of the second stage!
- This was the contribution of Nobel economist James Tobin, who developed "Tobit" model ("Tobin's Probit") to handle censored data

The Tobit Model

- Tobit model linking Xs to Y* and then to observed Y

$$Y_i^* = \Sigma \beta X_i + \varepsilon_i$$

$$Y_i^* = X\mathrm{B} + \varepsilon_i \quad \text{in matrix type notation}$$

$$\mathrm{E}(Y_i^* \mid X) = X\mathrm{B}$$

- This is a linear model linking X to continuous Y*. Now we map to the observed Y via the "measurement model". If Y* is above a certain threshold, we observe Y*; if Y* is below a threshold, observed Y will be the threshold value

$$Y_i = Y^* \quad \text{if } \mathrm{Y}_i^* > \tau$$

$$Y_i = \tau \quad \text{if } \mathrm{Y}_i^* \leq \tau$$

- This is just like the probit model with a measurement twist (right?)

- Assume τ=0: it is the case in many applications, and it is much easier to deal with mathematically (but it is not intrinsically necessary)
- Then

$$Y_i = Y^* = XB + \varepsilon \ \text{ if } \mathrm{Y}_i^* > 0$$

$$Y_i = 0 \ \text{ if } \mathrm{Y}_i^* \leq 0$$

- We want to model the expected value of Y conditional on the X, as we do it all regression models. How does E(Y) change as a function of X?
- With censored Y, the conditional expectation of Y becomes

$$E(Y|x) = P(Y>0|x) * E(Y|Y>0,x) + P(Y=0|x) * 0$$

- The conditional mean of Y, given X, is equal to the probability of Y being beyond the threshold, multiplied by the average value of Y past the threshold, plus the probability of not being past the threshold, multiplied by the threshold value (which is 0, so this part will fall out)
- Can see the two stages of the model (right?)

Latent

$$Y^* = XB + \varepsilon$$

$$E(\varepsilon \mid X) = 0, E(\varepsilon^2) = \sigma^2$$

Observed

$$Y_i = XB + \varepsilon \text{ if } \mathrm{Y}_i^* > 0$$

$$Y_i = 0 \text{ if } \mathrm{Y}_i^* \leq 0$$

$$E(Y \mid X) = P(Y > 0 \mid x) * E(Y \mid Y > 0, X)$$

- We're going to model two things: P(Y>0), and then E(Y|Y>0) as a function of the Xs
- Will see formally all the intuitive problems we've pointed out about not correcting for the censoring process

- Begin with P(Y>0) portion of the model

$$P(Y_i > 0) = P(Y^* > 0) = P(XB + \varepsilon > 0)$$

$$P(Y_i > 0) = P(\varepsilon > -XB) = P(\varepsilon < XB) = \Phi(XB)$$

- Under standard probit assumptions that ε is normally distributed; since we have information on Y* we don't need to set the variance to be 1 but can estimate it (see previous slide)

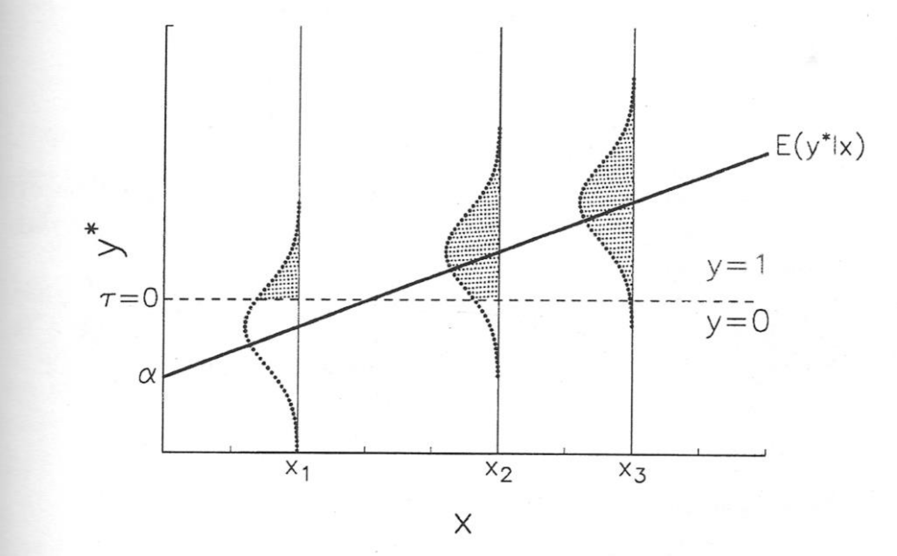


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

This gives the probability of being anywhere over the threshold, as in regular probit. **But we can do more since we know the Y* for all of the "over the threshold" cases**

- Now, what about the E(Y|Y>0) part of the model? That is, given that Y is above the threshold, what is its average value, conditioned on the Xs?
- Think about that portion of the distribution over the threshold: it is a "truncated normal distribution" – a normal distribution with some portion cut off and deleted (the portion at or below 0 or whatever τ is)
- In a truncated normal distribution:

$$E(Y \mid Y > 0, X) = X\mathrm{B} + E(\varepsilon \mid Y > 0, X)$$

- That is, the expected value of Y, conditioned on X is equal to the "true" mean of Y*, conditioned on X, *plus* an additional term which equals the average error term, conditioned on X, among the units greater than the threshold.
- While E(ε|x)=0 in the entire Y* distribution (see previous slide); it **cannot be zero in a truncated distribution** since we are focusing on that part of the distributed greater than some threshold value (0)

$$\mathrm{E}(Y \mid Y > 0, X) = X\mathrm{B} + \mathrm{E}(\varepsilon \mid Y > 0, X)$$

$$= X\mathrm{B} + \sigma \frac{\phi\left(\frac{XB}{\sigma}\right)}{\Phi\left(\frac{XB}{\sigma}\right)}$$

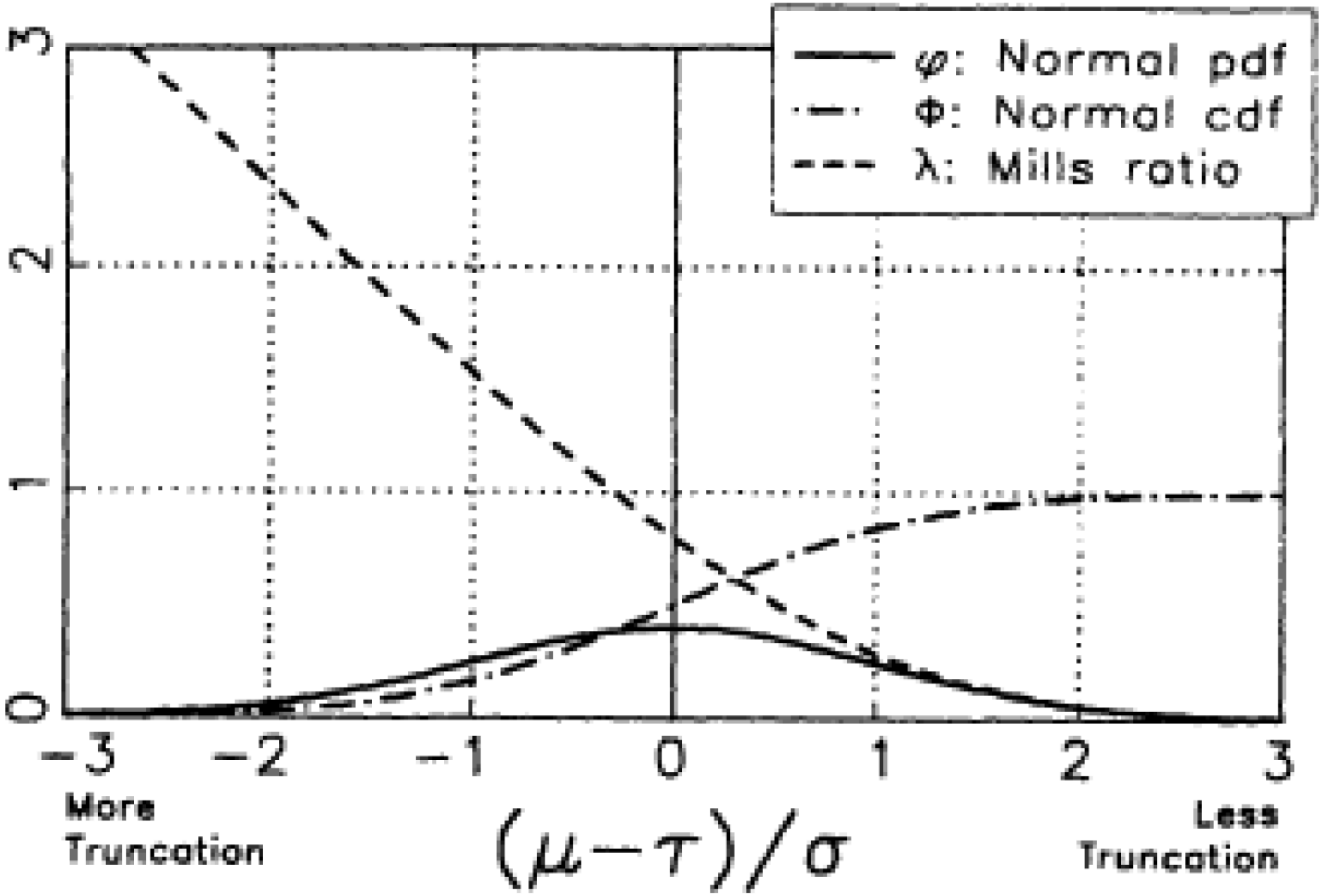
$$= X\mathrm{B} + \lambda\sigma$$

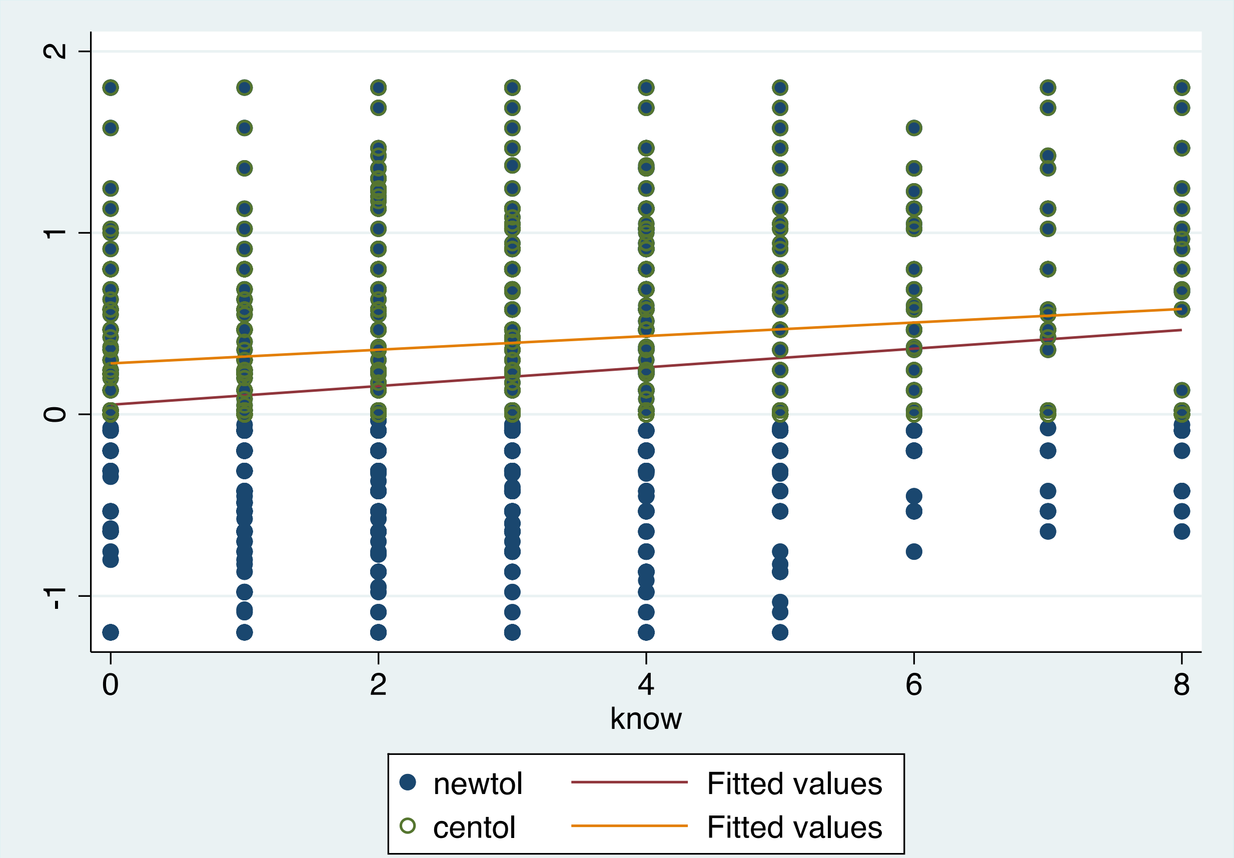
$$\text{where } \lambda = \frac{\phi\left(\frac{XB}{\sigma}\right)}{\Phi\left(\frac{XB}{\sigma}\right)} = \text{"Inverse Mills Ratio"}$$

So: E(ε|Y>0, x) in a truncated distribution with a normally distributed error term with variance σ^2 is: σ * ϕ/Φ (with the last term being the "INVERSE MILLS RATIO," (IMR), the most important new statistical quantity for these kind of models.

- IMR is the height of the normal curve evaluated at a given point XB, divided by the cumulative probability of being **uncensored**. (This works out nicely because the threshold was set to 0)
- When IMR is large, it means that: at a given point XB, there is more truncation, as the probability of being uncensored (the denominator) is smaller. This happens when XB puts the mean closer or under the threshold. Then the pdf in the numerator divided by the cumulative cdf in the denominator gets big.
- As XB puts the case farther from the threshold, there is less truncation, denominator is bigger, pdf in the numerator also smaller. So low IMR.
- So the size of the IMR, weighted by the standard deviation, is the difference between $E(\varepsilon|x)$ from the Y^* and $E(\varepsilon|Y>0,x)$

Long 1997


φ: Normal pdf
Φ: Normal cdf
λ: Mills ratio
0
1
2
3
−3
−2
−1
0
1
2
3
More
Truncation
$(\mu - \tau)/\sigma$
Less
Truncation



So when KNOW is high, fewer censored observations, and gap between E(Y|X) and E(Y*|X) is smaller

When KNOW is low, more censored observations, and gap between between E(Y|X) and E(Y*|X) is larger

That means the IMR at KNOW=8 is smaller than the IMR when KNOW=0

- Putting this together gives the censored regression model:

$$E(Y|X) = P(Y>0|x)*E(Y|Y>0,x)+P(Y=0|x)*0$$

$$E(Y|X) = \Phi\left(\frac{XB}{\sigma}\right)\left(XB+\lambda\sigma\right)+(1-\Phi)\left(\frac{XB}{\sigma}\right)*0$$

$$E(Y|X) = \Phi\left(\frac{XB}{\sigma}\right)\left(XB+\lambda\sigma\right)$$

- Which is the probability of being uncensored, multiplied by the expected value of Y, given that the unit is uncensored
- When IMR →∞, the probability of being uncensored is smaller and smaller, so the expectation of Y converges to the threshold 0
- When IMR→0, the probability of being uncensored is larger and larger, so the expectation of Y converges to XB, as it would with Y*
- We want to estimate this model to correct for the problems caused by censoring. Two methods: two-step tobit, and MLE

Two-Step Tobit

- See that the estimation is essentially an omitted variable problem. We need an estimate of the IMR (λ) which can correct the bias in OLS estimates on the censored data and recover the "true" β
- Parenthetically: can see that if omit the IMR from consideration and just estimate OLS on the censored data, X will be intrinsically related to the error term (which would then include λσ). If X is negatively related to the probability of being censored (as it is in our case), we have a negative relationship between X and the composite error term.
- Again, that's why OLS underestimates the "true" relationship

$$E(Y \mid X) = \Phi\left(\frac{XB}{\sigma}\right)\left(XB + \lambda\sigma\right)$$

- So correcting for "endogeneity" due to omitted variable bias is another way of looking at tobit! (We'll look at endogeneity further in Units 3 and 4)

- Procedure:

1. Run probit on all cases to generate Φ (probability of being uncensored) and ϕ, pdf associated with XB.

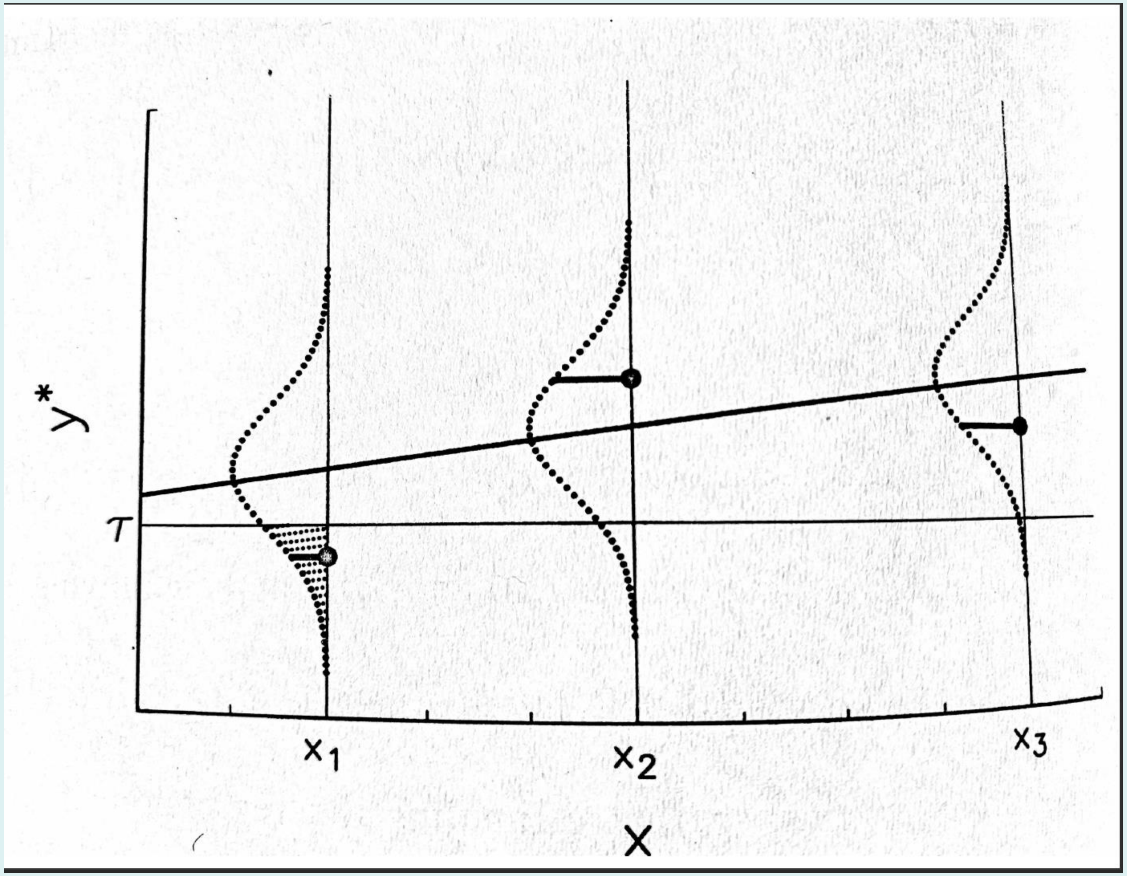
2 . Use estimated Φ and ϕ to calculate the IMR for each case

3. Run OLS on cases where Y>0 as Y=XB+σ*IMR

- This is running a probit to obtain estimates of IMR, and then inserting that estimate into the equation for Y>0 as an omitted variable correction in order to recover the true coefficients

- Makes intuitive sense, but it is less efficient than ML estimation
- Standard errors in the two-step procedure need to be corrected

- Maximum Likelihood Tobit
- Divide the observations into two sets: the censored cases, and the uncensored cases, and see what each contributes to the likelihood of having observed the sample of Ys that we did observe
- See Long, p.204



3 cases: x1 is censored on y (y*<0); and x2/x3 have observed y as Y*

What do each contribute to the overall likelihood function?

- For x1, we only know that it is censored on y, so we can't calculate a pdf (height of the normal curve). We can only calculate the probability that it was censored given x
- Since

$$P(y_i^* > 0) = P(\varepsilon > -\frac{XB}{\sigma}) = P(\varepsilon < \frac{XB}{\sigma}) = \Phi(\frac{XB}{\sigma})$$

Then $P(y_i<0)=1\text{-}\Phi(\frac{XB}{\sigma})$

- Summing over all the censored observations and taking logs:

$$L_c(\beta,\sigma^2) = \Pi\left(1\text{-}\Phi(\frac{XB}{\sigma})\right)$$

$$\ln L(\beta,\sigma^2) = \sum_{censored}\left(1\text{-}\Phi(\frac{XB}{\sigma})\right)$$

- For uncensored observations x1 and x2, their contribution to the likelihood function is the height of the normal curve evaluated at XB, just like in normal regression

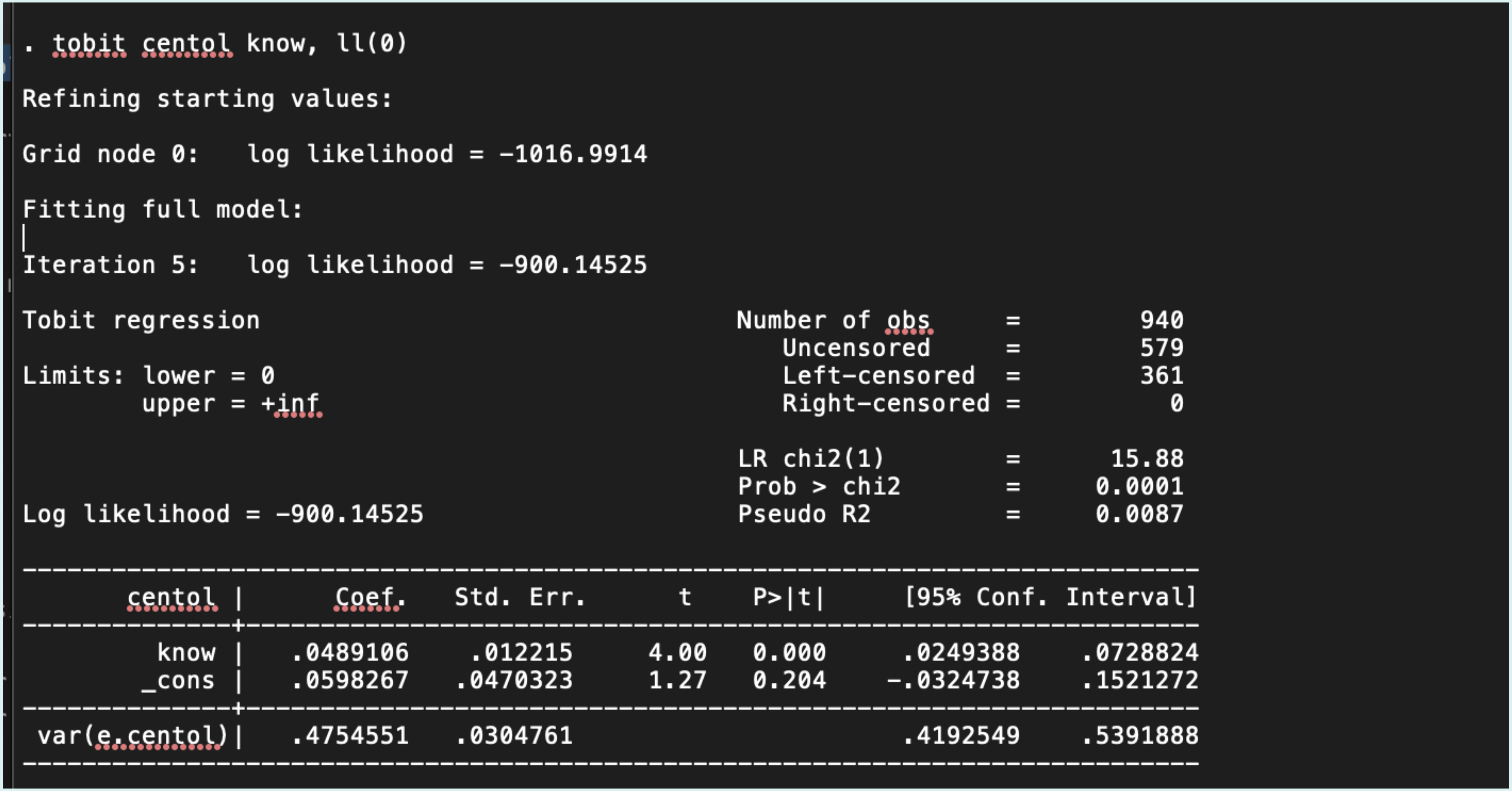
$$L_u(\beta,\sigma^2) = \Pi\left(\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{y_i - XB}{\sigma}\right)^2}\right)$$

$$\ln L_u(\beta,\sigma^2) = \sum_{uncensored} \ln\frac{1}{\sigma}\phi\left(\frac{y_i - XB}{\sigma}\right)$$

where ϕ=height of the normal curve for y_i

- Putting together for all cases, censored and censored gives the Tobit Likelihood function, which is maximized wrt β and σ as usual

$$\ln L_u(\beta,\sigma^2) = \sum_{uncensored} \ln\frac{1}{\sigma}\phi\left(\frac{y_i - XB}{\sigma}\right) + \sum_{censored} \ln\left(1 - \Phi\left(\frac{y_i - XB}{\sigma}\right)\right)$$

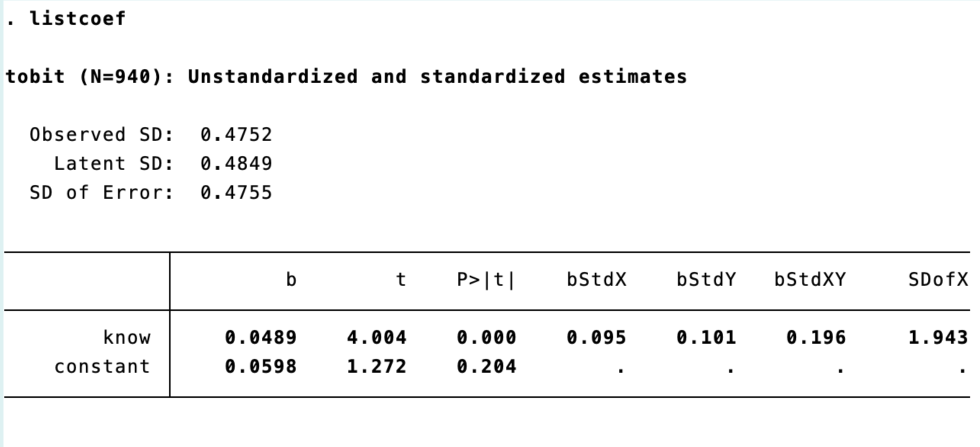


```
. tobit centol know, ll(0)

Refining starting values:

Grid node 0:   log likelihood = -1016.9914

Fitting full model:

Iteration 5:   log likelihood = -900.14525

Tobit regression                                Number of obs     =        940
                                                   Uncensored     =        579
Limits: lower = 0                                  Left-censored  =        361
        upper = +inf                               Right-censored =          0

                                                LR chi2(1)        =      15.88
                                                Prob > chi2       =     0.0001
Log likelihood = -900.14525                     Pseudo R2         =     0.0087

------------------------------------------------------------------------------
      centol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .0489106    .012215     4.00   0.000     .0249388    .0728824
       _cons |   .0598267   .0470323     1.27   0.204    -.0324738    .1521272
-------------+----------------------------------------------------------------
 var(e.centol)|   .4754551   .0304761                      .4192549    .5391888
------------------------------------------------------------------------------
```

- Coefficient for knowledge is .049, *much* closer to the “true” value of .051
- Corrects for the biasing effect of censoring at 0 for the 361 censored cases

Interpretation of Tobit coefficients

1. Effect of X on Y* = β. This is what we want to obtain, the recovery of the “true” β in the context of a partially censored sample. what we wanted. Can also use LISTCOEF to get standardized Y* estimate



```
. listcoef

tobit (N=940): Unstandardized and standardized estimates

  Observed SD:  0.4752
    Latent SD:  0.4849
  SD of Error:  0.4755

--------------+---------------------------------------------------------------------
              |        b         t     P>|t|     bStdX     bStdY    bStdXY     SDofX
--------------+---------------------------------------------------------------------
         know |   0.0489     4.004     0.000     0.095     0.101     0.196     1.943
     constant |   0.0598     1.272     0.204         .         .         .         .
--------------+---------------------------------------------------------------------
```

- So one additional unit of KNOW changes E(Y*) by .10 standard deviations (.049/.485), and a standard deviation change in X (1.94) units changes standardized Y* by (.049*1.94/.485)= .196
- Here you can add additional variables and don’t need the KHB method since Y* variance is fixed, estimated from the uncensored part of the distribution

2. Marginal Effect of X on Y for exceeding the threshold and being uncensored = P(uncensored)(1-P(uncensored)*β. = $\Phi(XB/\sigma)(1- \Phi(XB/\sigma))^* \beta$

This is the effect of very small changes in X on the probability of being uncensored, and is the slope of the tangent to the probit curve predicting being uncensored, evaluated at XB. Reaches its maximum value when probability of being uncensored (or censored)=.5

Can also use discrete change as appropriate

3. Effect of X on **actual Y** = $\beta^* \Phi(XB/\sigma)$= Probability of being uncensored multiplied by β. This is the effect of changes in X (marginal or discrete) on the expected value of Y (which is censored), not Y*.

Can see that as probability of being censored increases, the effect of a unit/marginal/standardized change in X on **actual (censored) Y** is very low, because the change won't be enough to put the case past the threshold so will likely keep it as 0. Similarly, when P(uncensored)=1, the effect is β, because the case is already over the threshold.

McDonald Moffit Decomposition

- One well known quantity in Tobit regression is the "McDonald-Moffit Decomposition", which takes the effect of X on actual (censored) Y -- see previous slide – and decomposes it into two parts:

  a) the average marginal change in Y for cases over the threshold, weighted by P(uncensored)

  b) the marginal change in P(uncensored), weighted by average (expected) value of Y for those cases over the threshold

$$\frac{\partial E(Y|x)}{\partial x_k} = \Phi\left(\frac{xB}{\sigma}\right) * \frac{\partial E(y|y>0,x)}{\partial x_k} + E(y|y>0,x)\frac{\partial \Phi(xB/\sigma)}{\partial x_k}$$

  – Stata (2) (3) (4) (1)

$$\frac{\partial E(Y|x)}{\partial x_k} = \Phi\left(\frac{xB}{\sigma}\right) * \frac{\partial E(y|y>0,x)}{\partial x_k} + E(y|y>0,x)\frac{\partial \Phi(xB/\sigma)}{\partial x_k}$$

- Stata (2) (3) (4) (1)

- So first part (Stata do file parts 2 and 3) gives the effect of the marginal change in X on the expected value of Y for the uncensored cases, weighted by the probability of being uncensored, and the second part (Stata parts 1 and 4) gives the effect of the marginal change in X on the probability of being uncensored, weighted by the expected value of y for the uncensored.

- Put Xs at their mean and calculate this quantity, then see which one is bigger effect. (See this week's do file for an example).