# Tobit Estimation in gretl

## Quantitative Microeconomics

R. Mora

Department of Economics
Universidad Carlos III de Madrid

# Outline

1. Introduction
2. Tobit in gretl
3. Maximum Likelihood of an Interval Regression Model

# The Tobit Model and ML Estimation

## The Tobit Model

- $h^* = \beta x + \varepsilon$
- $\varepsilon \sim N\left(0, \sigma^2\right)$
- $\begin{cases} \text{if } h^* > 0 \Rightarrow h = \beta x + \varepsilon \\ \text{if } h^* \leq 0 \Rightarrow h = 0 \end{cases}$

$$\hat{\beta}^{ML} = \arg\max \quad \sum_i \left\{ 1\left(h_i > 0\right) \log\left(\left(\tfrac{1}{\sigma}\right) \phi\left(\tfrac{h_i - \beta x_i}{\sigma}\right)\right) + 1\left(h_i = 0\right) \log\left(1 - \Phi\left(\tfrac{\beta x_i}{\sigma}\right)\right)\right\}$$

- in gretl, a quasi-Newton algorithm is used (the BFGS algorithm)

# Basic Commands in gretl for Tobit Estimation

- `tobit`: computes Maximum Likelihood tobit estimation
- `omit/add`: tests joint significance
- `$yhat`: estimates the depdendent variable
- `$lnl`: returns the log-likelihood for the last estimated model
- `intreg`: computes Maximum Likelihood of an interval regression model with normal disturbances

- in this Session, we are going to learn how to use `tobit` and `intreg`

# tobit *depvar indvars* ——verbose

- "censoring": the dependent variable only takes nonnegative values, with 0 reserved for the censured observations
- for other cases (censoring from above, at a point different from zero) the dependent variable must be re-defined appropriately
- for two-sided censoring the `intreg` command may be used
- it is possible to trace all iterations with the `--verbose` option
- output shows the $\chi^2_q$ statistic test for the null that all slopes are zero
- by default, standard errors are computed using the negative inverse of the Hessian

# Example: Simulated Data

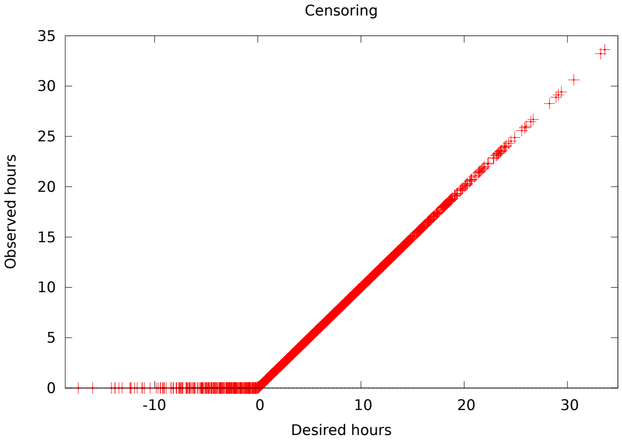
**The Tobit Model**

- $h^* = 10 + 0.5 * educ - 5 * kids + \varepsilon$
- $\varepsilon \sim N(0, 49)$

- education makes you willing to work more
- having a kid makes you willing to work less
- $\beta x = 5 + 0.5 * educ - 5 * kids$
- $\sigma = 7$

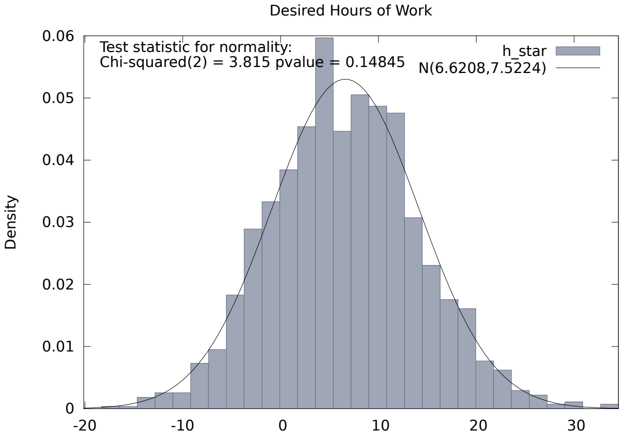
# Observed Hours vs. Desired Hours

$$h^* = 5 + 0.5 * educ - 5 * kids + \varepsilon, \varepsilon \sim N(0, 49)$$



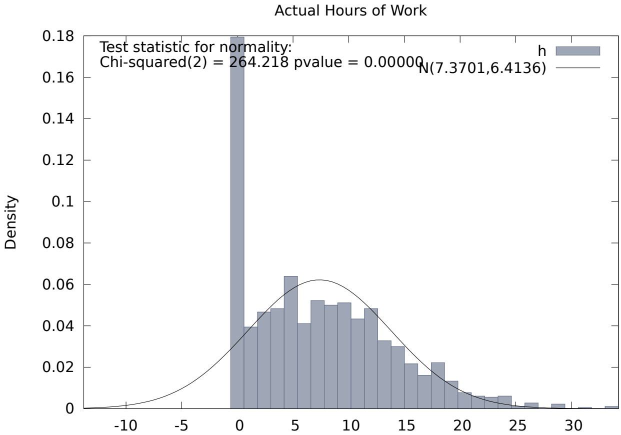
# Histogram of Desired Hours of Work

$$h^* = 5 + 0.5 * educ - 5 * kids + \varepsilon, \varepsilon \sim N(0, 49)$$



# Censoring in the Tobit Model

$$h^{*} = 5 + 0.5 * educ - 5 * kids + \varepsilon, \varepsilon \sim N(0, 49)$$



# ols with the Full Sample

$$h^{*}=5+0.5*educ-5*kids+\varepsilon,\varepsilon\sim N(0,49)$$

```
Model 1: OLS, using observations 1-1500
Dependent variable: h

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        6.60589     0.757742      8.718    7.39e-18 ***
  educ         0.387944    0.0751737     5.161    2.79e-07 ***
  kids        -4.48766     0.382296    -11.74     1.68e-30 ***

Mean dependent var   7.370101   S.D. dependent var   6.413578
Sum squared resid    54501.48   S.E. of regression   6.033833
R-squared            0.116094   Adjusted R-squared   0.114913
F(2, 1497)           98.30967   P-value(F)           7.67e-41
Log-likelihood      -4822.980   Akaike criterion     9651.960
Schwarz criterion    9667.899   Hannan-Quinn         9657.898
```

# ols with the Restricted Sample

$$h^* = 5 + 0.5 * educ - 5 * kids + \varepsilon, \varepsilon \sim N(0, 49)$$

```
Model 2: OLS, using observations 1-1223
Dependent variable: h

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        7.35385     0.773674      9.505    1.02e-20 ***
  educ         0.333273    0.0764336     4.360    1.41e-05 ***
  kids        -2.84425     0.389046     -7.311    4.79e-13 ***

Mean dependent var   9.134221   S.D. dependent var   5.678892
Sum squared resid    37539.67   S.E. of regression   5.547091
R-squared            0.047441   Adjusted R-squared   0.045879
F(2, 1220)           30.38012   P-value(F)           1.33e-13
Log-likelihood      -3829.194   Akaike criterion     7664.387
Schwarz criterion    7679.715   Hannan-Quinn         7670.156
```

# tobit Output

$$h^* = 5 + 0.5 * educ - 5 * kids + \varepsilon, \varepsilon \sim N(0, 49)$$

```
Function evaluations: 48
Evaluations of gradient: 9

Model 3: Tobit, using observations 1-1500
Dependent variable: h

             coefficient   std. error     z      p-value
  -------------------------------------------------------
  const        4.89145     0.893211      5.476   4.34e-08 ***
  educ         0.461985    0.0891748     5.181   2.21e-07 ***
  kids        -4.44067     0.470267     -9.443   3.63e-21 ***

Mean dependent var   7.447435   S.D. dependent var   6.233858
Censored obs              277   sigma                7.089319
Log-likelihood      -4412.070   Akaike criterion     8832.141
Schwarz criterion    8853.394   Hannan-Quinn         8840.058

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 3.57203
  with p-value = 0.167627
```

# Predicting Actual Hours of Work for Those who Work

## computing $\hat{h}_i^*$ and $\hat{h}_i$

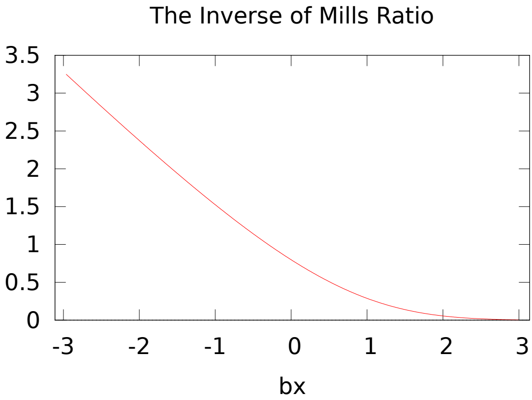
- $\hat{h}_i^*$: genr h_star_hat=$yhat
- for each observation, $\hat{h}_i = \max\left\{0, \hat{\beta} x_i\right\}$

## $E\left[h \,|h > 0, x\right]$

- $E\left[h \,|h > 0, x\right] = \beta x + E\left[\varepsilon \,|\beta x + \varepsilon > 0, x\right]$
- it can be shown that: $E\left[h \,|h > 0, x\right] = \beta x + \sigma \frac{\phi(\frac{\beta x}{\sigma})}{\Phi(\frac{\beta x}{\sigma})}$

- $\frac{\phi(\frac{\beta x}{\sigma})}{\Phi(\frac{\beta x}{\sigma})}$ is usually referred to as the inverse of Mills ratio
  - incidentally, John Mills actually used $\frac{1-\Phi(\frac{\beta x}{\sigma})}{\phi(\frac{\beta x}{\sigma})}$

# The Inverse of Mills Ratio




the higher $\frac{\beta x}{\sigma}$, the higher the probability of participation and the lower the correction

# Predicting Actual Hours of Work

## $E[h|x]$

- $E[h|x] = \Pr(h > 0)\, E[h|h > 0, x]$
- it can be shown that: $E[h|x] = \Phi\left(\frac{\beta x}{\sigma}\right)\left[\beta x + \sigma \frac{\phi(\frac{\beta x}{\sigma})}{\Phi(\frac{\beta x}{\sigma})}\right]$

# Understanding the Coefficients and the Slopes

- the Tobit estimates for the coefficients, $\hat{\beta}$, give the marginal effects on the desired number of hours
- frequently, we also want an estimate of the marginal effects on the probability of working and on the actual hours worked

# Algebraic Marginal Effects

## Probability to Participate

- $\frac{\partial \Pr(h_i > 0)}{\partial x_j} = \phi(\frac{\beta x}{\sigma}) \left(\frac{\beta_j}{\sigma}\right)$

## Actual Hours Worked

- $\frac{\partial E(h_i | x)}{\partial x_j} = \beta_j \Phi \left(\frac{\beta x}{\sigma}\right)$
- aprox. estimates of this effect can be obtained using OLS over the full sample

# Individual Marginal Effects: Discrete Change

- we may want to get individual marginal effects

## Discrete change

- store estimated coefficients in a vector
- generate a matrix with the controls under scenario 0, $x_0$, and another one with the controls under scenario 1, $x_1$
- predict index functions $\hat{\beta}^{ML} x_0$ and $\hat{\beta}^{ML} x_1$
- simulate censuring
- generate the individual marginal effects

# Example: The Effect of Having an Extra Kid

```
# numerical individual marginal effects: having an extra kid
genr beta=$coeff
genr kids1=kids+1
matrix x0={const,educ,kids}
matrix x1={const,educ,kids1}
series h0 = (x0*beta>0)*x0*beta
series h1 = (x1*beta>0)*x1*beta
series Mg_kid=h1-h0
summary Mg_kid --by=educ --simple
```

# Example: summary Mg_kid --by=educ --simple

## the effect is also smaller with higher education

| | Mean | Minimum | Maximum | Std. Dev. |
|---|---|---|---|---|
| educ = 8 | -9.4738 | -9.4738 | -9.4738 | 0.0000 |
| educ = 12 | -7.1477 | -11.164 | -5.9182 | 2.2237 |
| educ = 16 | -5.0951 | -7.6081 | -2.3627 | 2.6231 |
| educ = 21 | -4.4750 | -4.4750 | -4.4750 | 0.0000 |

# The Interval Regression Model

$m_i \leq x_i\beta + \varepsilon_i \leq M_i$

- the dependent variable is unobserved for some (possibly all) observations
- we observe instead an interval in which the dependent variable lies

- In practice, each observation belongs to one of four categories:
  - left-unbounded
  - right-unbounded
  - bounded
  - point observations

# The Interval Regression Model in gretl

**intreg minvar maxvar X**

- `minvar` contains $m_i$, with NAs for left-unbounded observations
- `maxvar` contains $M_i$, with NAs for right-unbounded observations
- for some observations $m_i$ may equal $M_i$
- standard errors are computed using the negative inverse of the Hessian
- if the −−robust flag is given, a "sandwich" estimator for standard errors are calculated
- if you wish to construct a likelihood ratio test, this is easily done by estimating both the full model and the null model

# An Example for intreg

```
nulldata 100
# generate artificial data
set seed 201449
x = normal()
epsilon = 0.2*normal()
ystar = 1 + x + epsilon
lo_bound = floor(ystar)
hi_bound = ceil(ystar)
# run the interval model
intreg lo_bound hi_bound const x
```

# intreg Output

```
Model 1: Interval estimates, using observations 1-100
Lower limit: lo_bound, Upper limit: hi_bound

             coefficient   std. error     z      p-value
  --------------------------------------------------------
  const       0.977619     0.0360159    27.14   2.97e-162 ***
  x           0.993708     0.0375009    26.50   1.02e-154 ***

Chi-square(1)        702.1575   p-value             1.0e-154
Log-likelihood      -40.45572   Akaike criterion    86.91144
Schwarz criterion    94.72695   Hannan-Quinn        90.07452

sigma = 0.225354
Left-unbounded observations: 0
Right-unbounded observations: 0
Bounded observations: 100
Point observations: 0

Test for normality of residual -
  Null hypothesis: error is normally distributed
  Test statistic: Chi-square(2) = 27.391
  with p-value = 1.1275e-06
```

# Summary

- gretl allows for ML estimation of the Tobit estimation
- the Tobit model identifies how each control affects both the probability of not censoring and the expectation of the dependent variable given that it is observed
- interval regression by ML can also be conducted in gretl