Intermediate Econometrics

10/12/2025 - Vincenzo Gioia

Censored and truncated models

Introduction

- **Cases where the value of the response is continuous and observed only in a certain range**
- These variables are truncated for a certain value, which can be on the left side of the distribution (l), on the right side (u) or on both sides
- The distribution of such a variable is a mix of a discrete and a continuous distribution:
  - the value of y is continuous on the $]l, u[$ interval, and its distribution can be described by a density function $f_Y(y)$
  - there is a mass of probability on l or/and on u, which is described by a probability $P(Y = l)$ or/and $P(Y = u)$
- Truncated responses in economics:
  - Corner solution
  - Missing data problem (censoring)
  - Selection problem (incidental truncation)

Corner solution

The consumer problem

- Let's suppose to have goods: y = vacations and z = food
- Utility function describing the consumer preference: (q_y and q_z the quantities of the two goods)

$$ U(q_y,q_z) = (q_y + \mu)^\beta q_z^{1-\beta}, \quad 0<\beta<1, \; \mu>0 $$

- The consumer seeks to maximize their utility subject to the budget constraint (x is the income and p_y and p_z are the prices): $x = p_y q_y + p_z q_z$
- Condition for an interior solution :

$$ \frac{\beta}{1-\beta}\frac{q_z}{q_y+\mu} = \frac{p_y}{p_z} $$

- Interior solution

$$ \begin{cases} q_y = \beta \dfrac{x}{p_y} - (1-\beta)\mu \\ q_z = (1-\beta)\dfrac{x}{p_z} + (1-\beta)\dfrac{p_y}{p_z}\mu \end{cases} $$

Corner Solution and Truncation

Income Threshold

- Demand for y can become **negative** → not economically admissible
- Minimum income level for positive consumption of y:

$$\bar x = \frac{1-\beta}{\beta} p_y \mu$$

- If $x < \bar x$:

$$q_y = 0, \quad q_z = \frac{x}{p_z} \quad \Rightarrow \textbf{corner solution}$$

- Expenditure on y (starts when the income is greater than this level, C in the next slide)

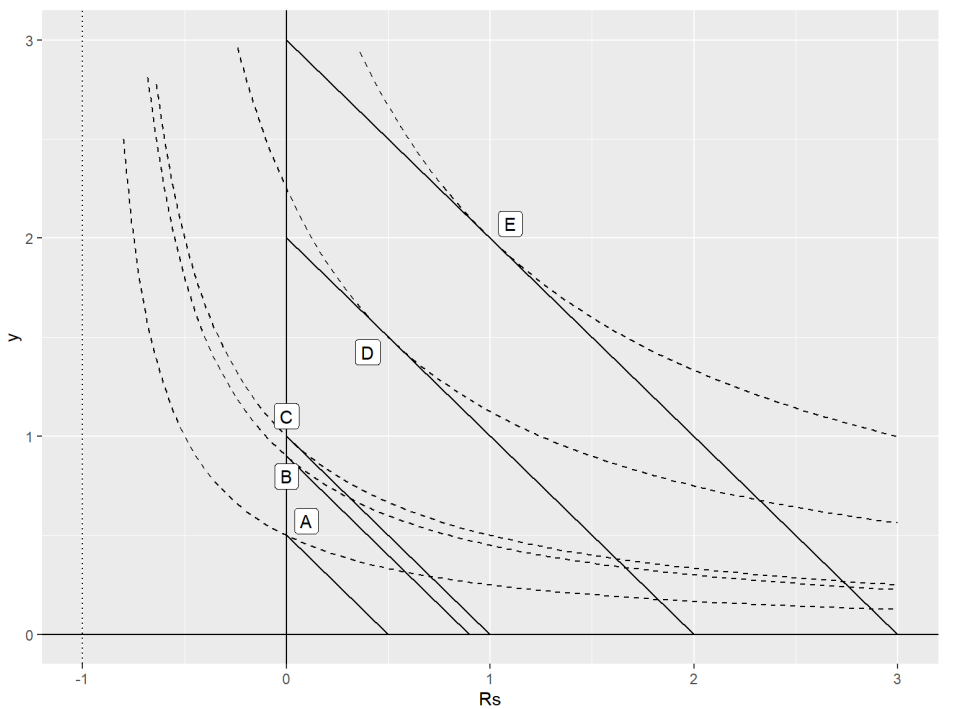
$$y = p_y q_y = \beta x - \beta \bar{x} = \beta_1 + \beta_2 x$$

- Interpretation:
  - y is a **left-truncated-at-zero variable**
  - β_2 = marginal propensity to consume good y
  - β_1 = minimum income required for positive consumption

Corner Solution and Truncation

Solution

- D and E: interior solutions
- A and B: corner soultions
- C: level of income that leads to a corner solution but for which the marginal rate of substitution equals the price ratio



Data censoring and truncation

Key Concepts

- **Data censoring** occurs when the dependent variable is only observed within a given range \(]l, u[\):
  - Values below \(l\) are reported as \(l\)
  - Values above \(u\) are reported as \(u\)
- Example: **Top-coding**
  - In Dutch household data, food expenditure above a threshold was replaced by an average value
  - The variable is **right-censored**
  - The censoring value is **not economically meaningful**
- **Data truncation** occurs when:
  - Observations outside a given range are **completely excluded** from the sample
  - Example: Only households below 1.5 times the poverty line are observed
- Key difference:
  - **Censoring** → values are modified
  - **Truncation** → observations are removed

Sample selection

Key concepts

- **Sample selection** occurs when the observation of the dependent variable is **not random** due to a **self-selection process**
- Example:
  - Outcome: **wage offers**
  - Wages are **only observed for women who participate** in the labor market
  - Participation decision determines whether wages enter the sample
- Key feature:
  - The outcome is **missing systematically**, not randomly
  - This leads to **selection bias**
- Relation to truncation and corner solutions:
  - It is a **different data-generating process**
  - But the **econometric treatment is similar**
  - Hence, these models are studied together

Censored vs Truncated Samples

Tobit Models

- Cases where the dependent variable is a **truncated variable**, but the **sample** can be either **censored** or **truncated**
- Differences on the vacation-food expenditure example
  - a **censored sample** consists of households for which the expenditure on vacations is positive and on household for which this expenditure is 0 (households that don't have any expenditure on vacations during the survey)
  - a **truncated sample** consists only of households for which the expenditure is strictly positive (individual surveyed in a travel agency or in an airport are truncated samples)
- **Censored regression model - Tobit model (Tobin, 1958)**:
  - Used with censored samples
  - Focus on Tobit-1 model: single equation; Jointly models of the probability of being observed and value of the response
  - We do not explore the Tobit-2 model, which is a bivariate model (2 equations)
- **Truncated regression model**:
  - Used with truncated samples

Tobit-1 model

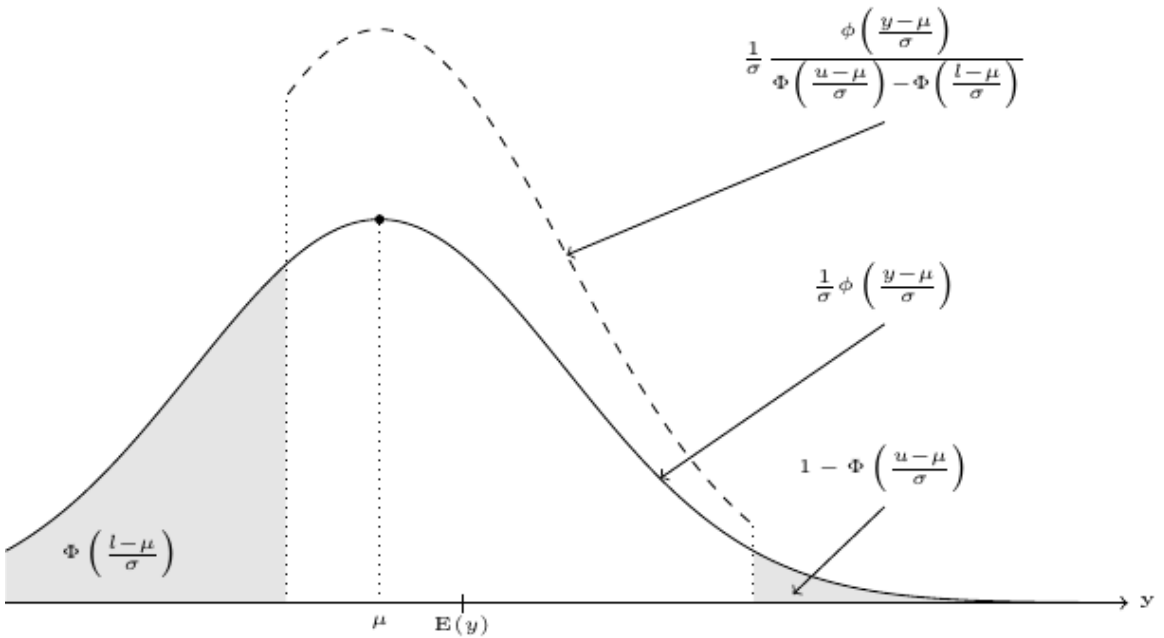
Definition

- Tobit-1 (or tobit, for short) is a linear model: $y_i = x^\top_i \beta + \epsilon_i$
  - y is only observed in a certain range, say $y \in] l, u[$
- In general, the tobit name is restricted to models estimated in a censored sample
  - In a semi-parametric setting, no hypotheses are made on the distribution of ϵ_i
  - under a fully parametric model we specify the distribution of ϵ; for example, it will suppose that $\epsilon_i \sim \mathcal{N} (0, \sigma_\epsilon^2)$, i.e., that the errors of the model are normal and homoskedastic
- In the context of the linear regression model, violation of these assumptions is not too severe, as the estimator is still consistent. This is not the case for the model studied here, as wrong assumptions of homoskedasticity and normality will lead to biased and inconsistent estimators

Truncated normal distribution

Truncated normal

- Under linear regression models we assume that the (conditional) distribution of the response is normal
- The fact that the response is truncated implies that the distribution of \(y\) is truncated normal
- The density of \(y\) is therefore: \[f(y) = \frac{1}{\sigma} \frac{\phi\left(\frac{y - \mu}{\sigma}\right)} {\Phi\left(\frac{u-\mu}{\sigma}\right) - \Phi\left(\frac{l-\mu}{\sigma}\right)} \]



Truncated normal distribution

Truncated normal

- As y is truncated, its expected value and its variance are not μ and $\sigma^ 2$
- Left(-right) truncation will lead to an expected value greater(-lower) than μ.
- For a general normal variable $y \sim \mathcal{N}(\mu, \sigma)$, denoting $\tilde{l} = (l - \mu) / \sigma$, the expectation is:

$$E(Y | Y > l) = \mu + \sigma \lambda_{\tilde{l}}$$

$$V(Y | Y > l)= \sigma ^ 2\left[1 - \lambda_{\tilde{l}}(\lambda_{\tilde{l}}-\tilde{l})\right] \] where $\lambda_l = \frac{\phi(l)}{1 - \Phi(l)}$ is the **inverse mills ratio**

Left-truncation at 0

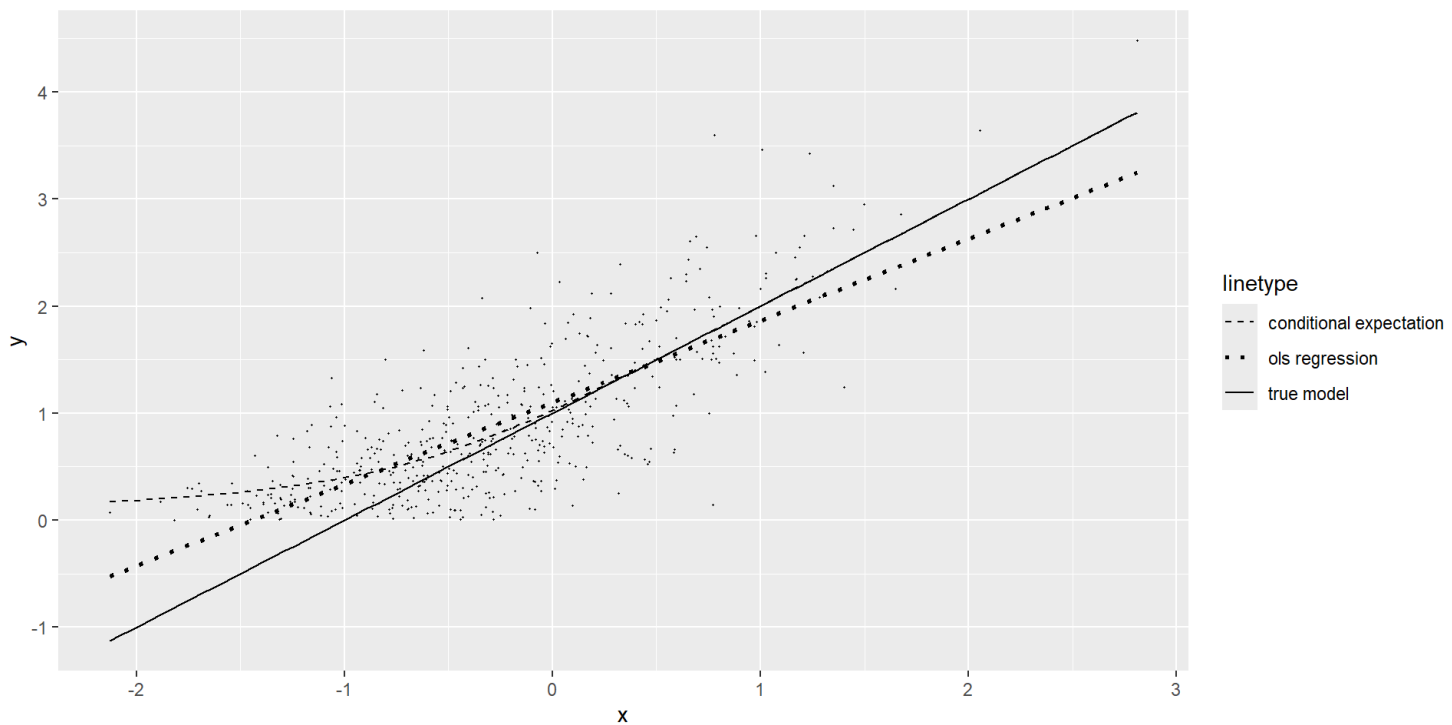
Consequences

- The expectation and the variance of y left-truncated at 0 can then be written as: $$\begin{array}{rcl} \mbox{E}(Y_i \mid x_i, Y_i > 0) &=& \mu_i + \sigma r(\mu_i / \sigma)\\ \mbox{V}(Y_i \mid x_i, Y_i > 0) &=& \sigma ^ 2 \left[1 + r'(\mu_i / \sigma)\right]\\ \end{array}$$ where $r(x)=\phi(x)/\Phi(x)$
- truncation has two consequences for the linear regression model:
  - the conditional variance depends on x so that **the errors of the model are heteroskedastic** $\implies$ **OLS inefficient**
  - the conditional expectation of Y is no longer equal to $\mu_i =x^\top_i \beta$, but to $\mu_i + \sigma r(\mu_i / \sigma)$ or, stated differently, the errors of the model are correlated with the covariate as $\mbox{E}(\epsilon \mid x) = \sigma r(\mu_n / \sigma)$ $\implies$ **OLS is biased and inconsistent**

Left-truncation at 0

Consequences

- For large values of \(x\), the expected value and the OLS regression line are similar
- Conversely, for low values of \(x\), the gap increases. This gap is positive and is particularly high for very low values of \(x\)



Truncated normal

Censored sample

- Up to now, we have considered a truncated sample, which is a sample containing only observed values of y
- Consider now that the underlying variable is: $y^*\mid x\sim \mathcal{N}(\mu, \sigma)$ with the following rule of observation:

$$\begin{cases}\begin{array}{rclccc} y&=&0 &\mbox{ if }& y^* < 0\\ y&=&y^* &\mbox{ if }& y^* \geq 0\\ \end{array}\end{cases}$$

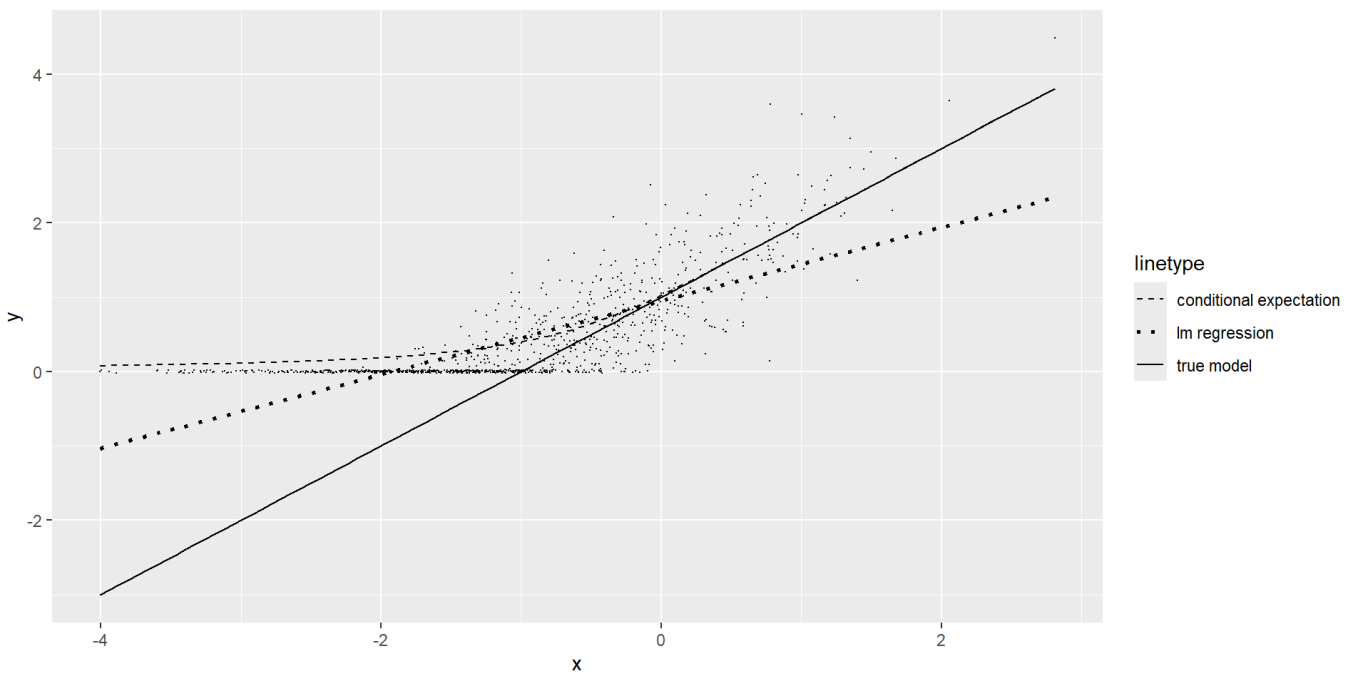
- In this case, the conditional expected value of Y can be computed as the weighted average of the expected value given that y is greater or lower than 0, the first one being the expected value of Y left-truncated at 0 and the second one being 0. With $\mu_i = \beta_1 + \beta_2 x_i$:

$$ \begin{array}{rcl} \mbox{E}(Y\mid x_n) &=& \left[1 - \Phi\left(\frac{\mu_n}{\sigma}\right)\right] \times 0 + \Phi\left(\frac{\mu_n}{\sigma}\right) \times \mbox{E}(Y\mid x, y > 0) \\ &=& \mu_n \Phi\left(\frac{\mu_n}{\sigma}\right) + \sigma \phi\left(\frac{\mu_n}{\sigma}\right) \end{array} $$

Truncated normal

Censored sample

- As for the previous case, the conditional expected value of \(Y\) is not \(\mu_i\), which implies that the **OLS estimator is biased and inconsistent**
- The downward bias of the slope seems more severe than for the truncated sample because there are much more observations for very low values of \(x\), i.e., in the range of the values of \(x\) where the correlation between \(x\) and \(\epsilon\) is severe



Interpretation of Coefficients

Corner Solution Models

- Let’s consider only the case of corner solution and not the case of data censoring (like top-coding)
- In both cases, the regression function: $\mu_i = x^\top_i \beta$ returns the mean of the distribution of the untruncated distribution of y
- In the data censoring case, which is just a problem of missing values of the response, this is the relevant distribution to consider and therefore β_k is the marginal effect of covariate x_k that we have to consider
- On the contrary, for corner solution models, the relevant distributions that we have to consider is on the one hand the probability of $y >0$ and on the other hand the zero left-truncated distribution of y

Interpretation of Coefficients

Corner Solution Models

- Therefore, μ_i is the mean of an untruncated latent variable, β_k is the marginal effect of x_k on this latent variable and none of these values are particularly meaningful.
- For a corner solution model, the effect of a change in x_k is actually twofold:
  - firstly, it changes the probability that the value of y is positive: $\mbox{P}(y > 0 \mid x)$,
  - secondly, it changes the expected value of y if it is positive: $\mbox{E}(y\mid x, y > 0)$.
- The probability that y is positive and the conditional expectation for positive values of y are, denoting as usual $\mu_i = x^\top_i \beta$:

$$ \begin{array}{rcl} \mbox{P}(Y_i > 0\mid x_i) &=& \Phi\left(\frac{\mu_i}{\sigma}\right)\\ \mbox{E}(Y_i\mid x_i, Y_i > 0) &=& \mu_i + \sigma r\left(\frac{\mu_i}{\sigma}\right) \end{array} $$

and the unconditional expectation of y is just the product of these two expressions:

$$ \mbox{E}(Y_i\mid x_i) = \mbox{P}(Y_i > 0\mid x_i) \times \mbox{E}(Y_i\mid x_i, Y_i > 0) $$

Interpretation of Coefficients

Corner Solution Models

- Its derivative with respect to x_k gives:

$$\begin{cases} \begin{array}{lcl} \frac{\displaystyle\partial\mbox{P}(Y_i > 0\mid x_i)}{\displaystyle\partial x_{ik}} &=& \frac{\beta_k}{\sigma} \phi\left(\frac{\mu_i}{\sigma}\right) \\ \frac{\displaystyle\partial \mbox{E}(Y_i\mid x_i, Y_i>0)}{\displaystyle\partial x_{ik}} &=& \beta_k \left[1 + r'\left(\frac{\mu_i}{\sigma}\right) \right] \end{array} \end{cases}$$

- Assuming $\beta_2>0$, we have
  - An increase of the probability that $Y>0$ (an increase of on the extensive margin)
  - An increase of the conditional expectation of Y, which is multiplied by the probability that y is observed (an increase of on the intensive margin)
- The sum of these two components gives the marginal effect of a variation of x on the unconditional expected value of Y $$\frac{\partial \mbox{E}(Y_i\mid x_i)}{\partial x_{ik}} = \beta_k\Phi\left(\frac{\mu_i}{\sigma}\right)$$

Estimation methods

Consistent estimators

- Several consistent estimators are available for the truncated and the censored model
- Inefficient estimators:
  - non-linear least squares
  - probit and two-step estimators
- The maximum likelihood estimator is asymptotically efficient if the conditional distribution of \(y\) is normal and homoskedastic
- The symmetrically trimmed least squares estimator, which is consistent even if the distribution of \(y\) is not normal and heteroskedastic

Estimation methods

Non-linear least squares

- The conditional expected value of \(y\): \(\mbox{E}(Y_i\mid x_i) = x^\top_i \beta + \sigma r\left(\frac{x^\top_i \beta }{\sigma}\right)\) is non-linear in \(x\)
- The parameters can be consistently estimated using non-linear least squares, by minimizing:

\[\sum_{i=1} ^ n \left[y_i - x^\top_i \beta - \sigma r\left(\frac{x^\top_i \beta}{\sigma}\right)\right] ^ 2 \]

Estimation methods

Probit and two-step estimators

- The probability that y is positive is $\Phi\left(\frac{x^\top_i \beta}{\sigma}\right)$, therefore, a probit model can be used to estimate the vector of coefficients $\frac{\beta}{\sigma}$
- σ is not identified, and each element of β is only estimated up to a $1/\sigma$ factor
- The probit estimation can only be performed for a censored sample, and not a truncated sample for which all the values of y are positive
- This idea leads to the **two-step estimator**:
  - first estimate the coefficient of the probit model $\hat{\delta}$ (with $\delta=\beta/\sigma$) and estimate r_i by $\hat{r}_i = r(x^\top_i\hat{\delta})$,
  - then regress y on x and $\hat{r}$ and estimate $\hat{\beta}$ and $\hat{\sigma}$.

Estimation methods

Maximum-likelihood estimation

- Estimating the model on the truncated sample leads to the log-likelihood of the normal gaussian model
- For the censored sample, the individual contribution to the likelihood sample will depend on whether $y=0$ or not and so the likelihood is simply the product of:
  - the likelihood of a probit model which explains that $y=0$ or $y > 0$
  - the likelihood of y for the truncated sample
- We can easily derive parameter estimates

Estimation methods

Semi-parametric estimators

- Only the **regression function is specified parametrically**: No distributional assumption on the error term (except symmetry)
- The **semi-parametric estimator is valid under much weaker assumptions**
- Main idea of the estimator:
  - Restore symmetry by **trimming or censoring the upper tail**
  - Then run **OLS on the symmetrically adjusted sample**
- Advantages:
  - Robust to **non-normality**
  - Valid for both **truncated and censored samples**
- Drawback:
  - Estimation is **computationally more complex**
  - The trimming depends on **unknown parameters**

Estimation of the Tobit-1 model in R

Model fitting

- The estimation of the tobit-1 model is available in functions of different packages:
  - `AER::tobit`
  - `censReg::censReg`
  - `micsr::tobit1`
- The three functions return identical results, except that they are parametrized differently: `micsr::tobit1` estimates σ as the two other functions estimate $\ln \sigma$.
- In addition to the formula and data arguments, they allow to specify
  - a `left` and a `right` argument to indicate the truncation points
  - by default these two arguments are 0 and $+\infty$, which correspond to the most usual zero left-truncated case
- The `micsr::tobit1` function allows to use either a censored or a truncated sample by setting the `sample` argument either to `"censored"` or `"truncated"`.
- The truncated regression model can also be estimated using the `truncreg::truncreg` function

Estimation of the Tobit-1 model in R

Model fitting

- We focus only on the `micsr::tobit1`, which also has the advantage of providing several different estimators selected using the `method` argument
  - `"ml"` for maximum likelihood (the only method available for the other towo functions)
  - `"lm"` for linear model
  - `"twostep"` for the two-step estimator
  - `"trimmed"` for the trimmed estimator
  - `"nls"` for the non-linear least squares estimator

Estimation of the Tobit-1 model in R

Example

- We consider the `charitable` data set concerning charitable giving (Wilhelm, 2008)
- Cross-section of 2384 households from 2001:
  - donation: the amount of (annual) charitable giving in US dollars (**outcome**)
  - donparents: the amount of charitable giving of the parents
  - education: the level of education of household's head, a factor with levels "less_high_school", "high_school", "some_college", "college", "post_college"
  - religion: a factor with levels "none", "catholic", "protestant", "jewish" and "other"
  - income: income
  - married: a dummy for married couples
  - south: a dummy for households living in the south

```
1 library(micsr)
2 data("charitable")
```

Estimation of the Tobit-1 model in R

Example

- Donation variable is left-censored for the value of 25, as this value corresponds to the item “less than $25 donation”
- For this value, we have households who didn’t make any charitable giving and some who made a small giving (from $1 to $25)

```
1  summary(charitable)
```

```
   donation       donparents                education         religion
Min.   :   25   Min.   :    25   less_high_school:242   none       : 322
1st Qu.:   25   1st Qu.:   125   high_school     :847   catholic   : 537
Median :  275   Median :   775   some_college    :670   protestant:1174
Mean   : 1234   Mean   :  2569   college         :396   jewish     :  66
3rd Qu.: 1125   3rd Qu.:  2368   post_college    :229   other      : 285
Max.   :76825   Max.   :491525
    income            married           south
Min.   :   855.5   Min.   :0.0000   Min.   :0.0000
1st Qu.: 31101.4   1st Qu.:0.0000   1st Qu.:0.0000
Median : 50712.5   Median :1.0000   Median :0.0000
Mean   : 63391.0   Mean   :0.6355   Mean   :0.3058
3rd Qu.: 78329.7   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :785385.2   Max.   :1.0000   Max.   :1.0000
```

Estimation of the Tobit-1 model in R

Example

- We consider the value of the donation in logs and subtracts from it \(\ln 25\), so that the response is 0 for households who gave no donation or a small donation
- The model can either be estimated using `logdon` as the response and the default values of `left` (0) or `right` (\(+\infty\)) or by using `log(donation)` as the response and setting `left` to `log(25)`
- We specify the model formula and we fit the model via ML

```
charitable$logdon <- log(charitable$donation) - log(25)

char_form <- logdon ~ log(donparents) + log(income) +
    education + religion + married + south
ch_ml <- tobit1(char_form, data = charitable)
```

Estimation of the Tobit-1 model in R

Example

- `log(income)` and `log(donparents)` positively influence charitable giving; Higher education levels is associated to higher donations; Religious groups donate more than reference group; Married households donate more; Region is not significant.

```
1 summary(ch_ml)
```

```
Maximum likelihood estimation
                        Estimate Std. Error  z-value  Pr(>|z|)
(Intercept)           -17.617804   0.898027 -19.6184 < 2.2e-16 ***
log(donparents)         0.200352   0.025235   7.9394 2.031e-15 ***
log(income)             1.453386   0.087030  16.6999 < 2.2e-16 ***
educationhigh_school    0.622148   0.188142   3.3068 0.0009437 ***
educationsome_college   1.100389   0.194320   5.6628 1.490e-08 ***
educationcollege        1.325042   0.214808   6.1685 6.895e-10 ***
educationpost_college   1.727244   0.235683   7.3287 2.324e-13 ***
religioncatholic        0.638635   0.171421   3.7255 0.0001949 ***
religionprotestant      1.257030   0.154226   8.1506 3.623e-16 ***
religionjewish          1.001090   0.307026   3.2606 0.0011118 **
religionother           0.836793   0.193670   4.3207 1.555e-05 ***
married                 0.766903   0.116755   6.5685 5.084e-11 ***
south                   0.112612   0.104586   1.0767 0.2815971
sigma                   2.113606   0.040971  51.5877 < 2.2e-16 ***
```

Estimation of the Tobit-1 model in R

Example

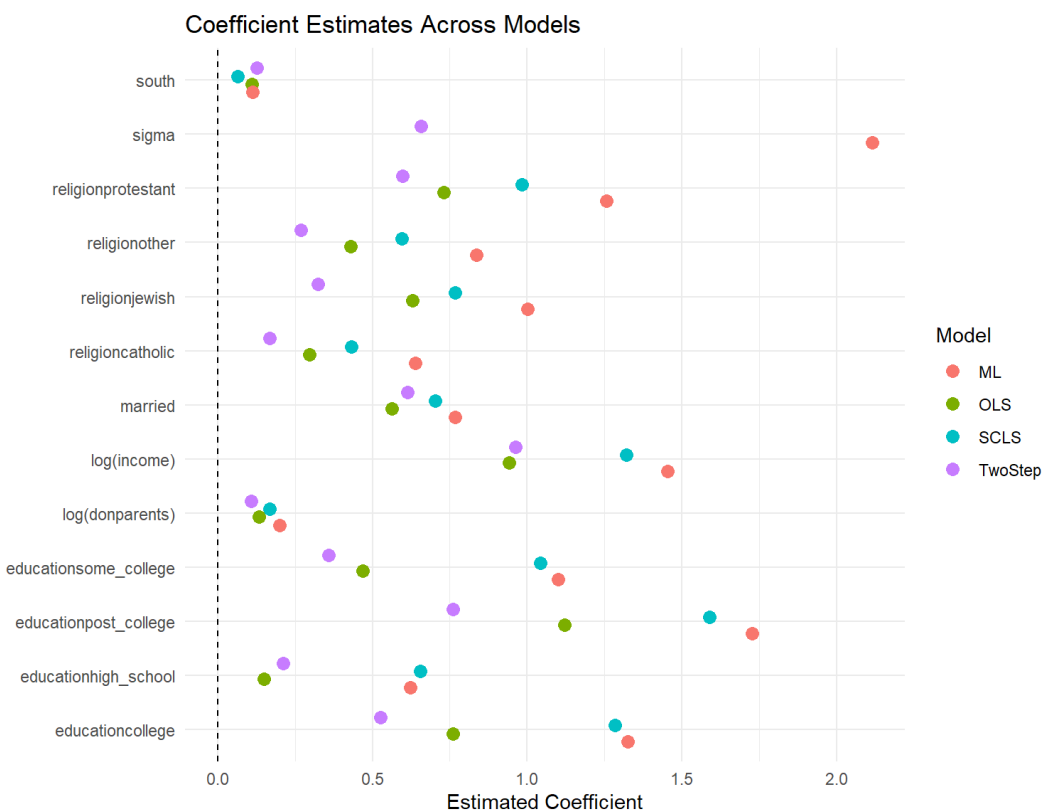
- Using `micsr::tobit1`, we also estimate the two-step, the **SCLS** (symmetrically censored least squares) and the OLS estimators.
- Let's just compare graphically the estimated coefficients under the four estimation approaches (we do not explore the standard errors because except forr the ML case, I guess they are affected by a bug)

```
ch_twostep <- tobit1(char_form, data = charitable, method = "twostep")
ch_scls <- tobit1(char_form, data = charitable, method = "trimmed")
ch_ols <- tobit1(char_form, data = charitable, method = "lm")
```

Estimation of the Tobit-1 model in R

Example

- Note that the OLS estimates are in general lower in absolute values than those of the three other estimators, which illustrates the fact that OLS estimators are biased toward zero when the response is censored



Conditional moment tests

Checks

- The most popular method of estimation for the tobit-1 model is the fully parametric maximum likelihood method
- Contrary to the OLS model, the estimator is only consistent if the generating process is perfectly described by the likelihood function, i.e., if \(\epsilon_n \sim \mathcal{N}(0,\sigma)\)
- In particular, the consistency of the estimator rests on the hypothesis of normality and homoskedasticity.
- The conditional moment tests are based on residuals (here, the residuals are partially observed) and can be computed using the `micsr::cmtest` function

Conditional moment tests

Normality and heteroschedasticity

- To test respectively the hypothesis of normality and of homoskedasticity
- Normality and heteroskedasticity are strongly rejected

```
1  cmtest(ch_ml, test = "normality")
```

```
    Conditional Expectation Test for Normality

data:  logdon ~ log(donparents) + log(income) + education + religion +  ...
chisq = 116.35, df = 2, p-value < 2.2e-16
```

```
1  cmtest(ch_ml, test = "heterosc")
```

```
    Heteroscedasticity Test

data:  logdon ~ log(donparents) + log(income) + education + religion +  ...
chisq = 103.59, df = 12, p-value < 2.2e-16
```

Conditional moment tests

Skewness and kurtosis

- Non-normality can be further investigated by testing separately the fact that the skewness and kurtosis indicators are respectively different from 0 and 3
- The hypothesis that the conditional distribution of the response is mesokurtic is not rejected at the 1% level and the main problem seems to be the asymmetry of the distribution, even after taking the logarithm of the response (This can be illustrated by plotting the (unconditional) distribution of the response (for positive values) and adding to the histogram the normal density curve)

```
1 cmtest(ch_ml, test = "skewness")
```

```
    Conditional Expectation Test for Skewness

data:  logdon ~ log(donparents) + log(income) + education + religion +  ...
z = 10.393, p-value < 2.2e-16
```

```
1 cmtest(ch_ml, test = "kurtosis")
```

```
    Conditional Expectation Test for Kurtosis

data:  logdon ~ log(donparents) + log(income) + education + religion +  ...
z = 2.3294, p-value = 0.01984
```

Right-truncated response

Example

- Let's consider the `food` data set
- Goal: estimate the demand for food in the Netherlands
- Two surveys are available: 1980 and 1988
- We will focus only on the first one

```
food <- as.data.frame(micsr.data::food)
food <- food[food$year == 1980,]
```

Right-truncated response

Example

- Cross-section of 4611 households
  - year: year of the survey, either 1980 or 1988
  - weights: weight in the survey
  - hsize: number of persons in the household
  - ageh: age of the head of the household in 5 classes
  - income: total income
  - food: food expenditure
  - midage: a dummy for household for which the head is aged between 35 and 64 years old

```
1 head(food)
```

```
  year weights hsize ageh income  food midage
1 1980  0.6152     2    1  55900  5280      0
2 1980  1.5214     1    1  10600  1700      0
3 1980  1.1618     2    5  34800  6730      0
4 1980  0.7749     5    2  34300 10640      1
5 1980  1.2303     1    2  35200  7410      1
6 1980  1.3042     1    1  15600  2040      0
```

Right-truncated response

Example

- Food expenses are top-coded for the top 5 percentiles, which corresponds to an expense of 13030 Dfl. The value reported for these observations is 17670 Dfl, which is the mean value of the expense for the top 5 percentile.
- The percentage of censored observations is not exactly 5%, because some observations have been excluded due to missing values for some covariates.

```
sum(food$food==17670)
```

```
[1] 127
```

```
mean(food$food==17670)
```

```
[1] 0.04534095
```

Right-truncated response

Example

- The response being expressed in logarithms, the right threshold is set to \(\log(13030)\)and the left one to \(-\infty\) as the response is not left-truncated
- The main coefficient of interest is the one associated with the log(income) covariate. Remember that in this data censoring case, the coefficient is the marginal effect, in the present context the income elasticity of food which is equal to 0.34.

```
food_tobit <- tobit1(log(food) ~ log(income) + log(hsize) + midage,
                         data = food, subset = year == 1980,
                         left = -Inf, right = log(13030))
food_tobit
```

```
Call:
tobit1(formula = log(food) ~ log(income) + log(hsize) + midage,
    data = food, subset = year == 1980, left = -Inf, right = log(13030))

Coefficients:
(Intercept)  log(income)   log(hsize)       midage        sigma
    4.70629      0.34005      0.47304      0.09501      0.36716
```

Two-sided tobit models

Example

- Source: Hochguertel (2003)
- Share of riskless assets, which can be either an internal solution, or a corner solution with the share equal to 0 or 1 ($l = 0$ and $u=1$) $\implies$ two-sided tobit model.
- Goal: explain the low share of risky assets in portfolios of Dutch households.
- The data set is called `portfolio`.

```
1 portfolio <- as.data.frame(micsr.data::portfolio)
2 names(portfolio)
```

```
 [1] "id"               "year"             "share"
 [4] "uncert"           "expinc"           "finass_10"
 [7] "finass_10_100"    "finass_more"      "networth"
[10] "noncapinc"        "mtrate"           "high_inc_oversmpl"
[13] "age"              "educ"             "diploma"
[16] "female"           "adults"           "child_0_12"
[19] "child_13_more"    "occup"            "riskav"
[22] "feeling"          "flex"             "smoke"
[25] "alcohol"          "body_mass"        "habits"
```

Two-sided tobit models

Example

- This data set is a panel of annual observations from 1993 to 1998 (yearly observations of 3348 households)
- The response is `share`, and the two covariates of main interest are `uncert` and `expinc`.
  - `uncert` indicates the degree of uncertainty felt by the household; it is a factor with levels `low`, `moderate` and `high`,
  - `expinc` indicates the prediction of the household concerning the evolution of their income in the next 5 years; it is a factor with levels `increase`, `constant` and `decrease`.
- Covariates: net worth, the age of household's head and its square, and a dummy for households whose head is a woman
- We use the `micsr::tobit1` function and we set the `left` and the `right` arguments respectively to 0 and 1.

```
1 portfolio$agesq  <- portfolio$age^2/100
2 prec_ml <- tobit1(share ~ uncert + expinc + networth +
3                     age + agesq + female,
4                   left = 0, right = 1, data = portfolio)
```

Two-sided tobit models

Example

- As expected, high uncertainty and pessimistic expectations about future income increase the share of riskless assets
- Net worth has a negative effect on the share of riskless assets
- Households headed by a woman have a higher share of riskless assets
- Finally the effect of age is U-shaped.

```
1 prec_ml
```

```
Call:
tobit1(formula = share ~ uncert + expinc + networth + age + agesq +
    female, data = portfolio, left = 0, right = 1)

Coefficients:
(Intercept)    uncertmod   uncerthigh    expinccst   expincdecr     networth
    1.48600      0.04410      0.08155      0.04059      0.04631     -0.02940
        age        agesq       female        sigma
   -0.02848      0.03055      0.14153      0.45578
```

Two-sided tobit models

Example

- Further investigations (Hochguertel, 2003) deal with the problem of eteroschedasticity
- The author modelled \(\log \sigma_i = w^\top_i \delta\) including all the covariates above, except `expinc` and `uncert`
- The function crch of the R package allows to do that

```
prec_ht <- crch::crch(share ~ uncert + expinc + networth +
         age + agesq + female | networth +
         age + agesq + female, left = 0, right = 1,
         data = portfolio)
```

Two-sided tobit models

```
1 summary(prec_ht)$coefficients
```

```
$location
               Estimate  Std. Error    z value      Pr(>|z|)
(Intercept)  2.66330267 0.081415222  32.712589 1.034277e-234
uncertmod    0.03473738 0.013854048   2.507381  1.216295e-02
uncerthigh   0.05364941 0.016387136   3.273874  1.060840e-03
expinccst    0.02644687 0.010686066   2.474893  1.332761e-02
expincdecr   0.04222966 0.014608545   2.890751  3.843229e-03
networth    -0.14346949 0.004353413 -32.955634 3.513004e-238
age         -0.02202405 0.003080029  -7.150600  8.639953e-13
agesq        0.02669050 0.003034251   8.796404  1.412698e-18
female       0.09196394 0.015422812   5.962852  2.478736e-09

$scale
               Estimate  Std. Error    z value      Pr(>|z|)
(Intercept)  0.41494455 0.126065307   3.291505  9.965294e-04
networth    -0.09291492 0.002554462 -36.373570 1.114519e-289
age         -0.01574593 0.005108319  -3.082410  2.053320e-03
```

Two-sided tobit models

Example

- It is clear that the heteroskedastic specification is supported by the data, the log-likelihood value of the second specification being much larger than the one of standard tobit model
- The value of some coefficients are strikingly different. For example, the coefficient of networth is -0.029 for the tobit model, but -0.1435 for the heteroskedastic model, as this covariate also has a huge effect on the conditional variance of the response

```
1 summary(prec_ht)$loglik
```

```
[1] -5960.59
```

```
1 summary(prec_ml)$logLik
```

```
    model
-6618.959
```

Tobit-1 model

Endogeneity

- Endogeneity can be treated in the **Tobit model** in a way very similar to the **Probit model**
- Key difference with Probit:
  - In **Probit**, we only observe whether the outcome is positive or not
  - In **Tobit**, we observe zero outcomes (corner solutions) and the **actual value** of the dependent variable when it is positive
- Because the positive values are observed:
  - The **variance of the error term is identified**
  - Unlike in the Probit model, where it must be normalized
- Available estimation methods (tobit1 function can be used for the purpose)
  - **Maximum Likelihood**
  - **Two-step estimator**
  - **Minimum Chi-square estimator**
- **Testing exogeneity** can be done using a **Wald test** based on the two-step estimator (endogtest can be used)
- Extra: see the example in the online book Microeconometrics with R