The Standard Tobit Model

censored regression model

$$
\begin{aligned}
y_i^* &= x_i'\beta + \varepsilon_i, \quad i = 1, 2, \dots, N, \\
y_i &= y_i^* \quad \text{if } y_i^* > 0 \\
&= 0 \quad \text{if } y_i^* \leq 0,
\end{aligned} \tag{7.60}
$$

$$
\begin{aligned}
P\{y_i = 0\} &= P\{y_i^* \leq 0\} = P\{\varepsilon_i \leq -x_i'\beta\} \\
&= P\left\{\frac{\varepsilon_i}{\sigma} \leq -\frac{x_i'\beta}{\sigma}\right\} = \Phi\left(-\frac{x_i'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right).
\end{aligned} \tag{7.61}
$$

$$
\frac{\partial P\{y_i = 0\}}{\partial x_{ik}} = -\phi(x_i'\beta/\sigma)\frac{\beta_k}{\sigma}. \tag{7.63}
$$

$$
E\{y_i|y_i > 0\} = x_i'\beta + E\{\varepsilon_i|\varepsilon_i > -x_i'\beta\} = x_i'\beta + \sigma\frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}. \tag{7.62}
$$

$$
E\{y_i\} = x_i'\beta\Phi(x_i'\beta/\sigma) + \sigma\phi(x_i'\beta/\sigma). \tag{7.64}
$$

$$
\frac{\partial E\{y_i\}}{\partial x_{ik}} = \beta_k\Phi(x_i'\beta/\sigma). \tag{7.65}
$$

The conditional expectaction of y (for y>0):

$$
\begin{aligned}
\mathrm{E}(y|y > 0, \mathbf{x}) &= \mathbf{x}\beta + \mathrm{E}(u|u > -\mathbf{x}\beta) \\
&= \mathbf{x}\beta + \sigma\mathrm{E}[(u/\sigma)|(u/\sigma) > -\mathbf{x}\boldsymbol{\beta}/\sigma] \\
&= \mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma)/\Phi(\mathbf{x}\beta/\sigma) \\
&= \mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)
\end{aligned}
$$

λ is called the inverse Mill's ratio.

The unconditional expectaction of y (for all y):

$$
\begin{aligned}
E(y|\mathbf{x}) &= P(y > 0|\mathbf{x}) \cdot E(y|y > 0, \mathbf{x}) + P(y = 0|\mathbf{x}) \cdot 0 \\
&= \mathrm{P}(y > 0|\mathbf{x}) \cdot \mathrm{E}(y|y > 0, \mathbf{x}) \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma) \cdot \mathrm{E}(y|y > 0, \mathbf{x}) \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)] \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)
\end{aligned}
$$

Estimation

$$
\begin{aligned}
\log L_1(\beta, \sigma^2) &= \sum_{i \in I_0} \log P\{y_i = 0\} + \sum_{i \in I_1} [\log f(y_i|y_i > 0) + \log P\{y_i > 0\}] \\
&= \sum_{i \in I_0} \log P\{y_i = 0\} + \sum_{i \in I_1} \log f(y_i),
\end{aligned} \tag{7.67}
$$

$$\log L_1(\beta, \sigma^2) = \sum_{i \in I_0} \log \left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right) \right] + \sum_{i \in I_1} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - x_i'\beta)^2}{\sigma^2} \right\} \right]. \tag{7.68}$$

truncated regression model

$$y_i^* = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, N, \tag{7.69}$$
$$y_i = y_i^* \quad \text{if } y_i^* > 0$$
$$(y_i, x_i) \text{ not observed if } y_i^* \leq 0,$$

Estimation:

$$\log L_2(\beta, \sigma^2) = \sum_{i \in I_1} \log f(y_i | y_i > 0) = \sum_{i \in I_1} [\log f(y_i) - \log P\{y_i > 0\}], \tag{7.70}$$

$$\log L_2(\beta, \sigma^2) = \sum_{i \in I_1} \left\{ \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - x_i'\beta)^2}{\sigma^2} \right\} \right] - \log \Phi\left(\frac{x_i'\beta}{\sigma}\right) \right\}. \tag{7.71}$$

Illustration: Expenditures on Alcohol and Tobacco (Part 1)

Table 7.7 Tobit models for budget shares alcohol and tobacco

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	−0.1592	(0.0438)	0.5900	(0.0934)
age class	0.0135	(0.0109)	−0.1259	(0.0242)
nadults	0.0292	(0.0169)	0.0154	(0.0380)
nkids ≥ 2	−0.0026	(0.0006)	0.0043	(0.0013)
nkids < 2	−0.0039	(0.0024)	−0.0100	(0.0055)
$\log(x)$	0.0127	(0.0032)	−0.0444	(0.0069)
age $\times \log(x)$	−0.0008	(0.0088)	0.0088	(0.0018)
nadults $\times \log(x)$	−0.0022	(0.0012)	−0.0006	(0.0028)
$\hat{\sigma}$	0.0244	(0.0004)	0.0480	(0.0012)
Loglikelihood	4755.371		758.701	
Wald test (χ_7^2)	117.86	($p = 0.000$)	170.18	($p = 0.000$)

For tabacco, age is an important factor in explaining the budget share. For alcoholic beverages, the number of children and total expenditures are individually significant.

Wald tests for the hypothesis that "all coefficients, except the intercept term, are equal to zero", produce highly significant values for both goods.

The Tobit II Model

The traditional model to describe sample selection problems is the **tobit II model**,[26] also referred to as the **sample selection model**. In this context, it consists of a linear wage equation

$$w_i^* = x_{1i}'\beta_1 + \varepsilon_{1i}, \tag{7.80}$$

where x_{1i} denotes a vector of exogenous characteristics (age, education, gender, ...) and w_i^* denotes person i's wage. The wage w_i^* is not observed for people that are not working (which explains the *). To describe whether a person is working or not a second equation is specified, which is of the binary choice type. That is,

$$h_i^* = x_{2i}'\beta_2 + \varepsilon_{2i}, \tag{7.81}$$

where we have the following observation rule:

$$w_i = w_i^*, h_i = 1 \quad \text{if } h_i^* > 0 \tag{7.82}$$

$$w_i \text{ not observed}, h_i = 0 \quad \text{if } h_i^* \leq 0, \tag{7.83}$$

The conditional expected wage, given that a person *is* working, is given by

$$\begin{aligned} E\{w_i|h_i = 1\} &= x_{1i}'\beta_1 + E\{\varepsilon_{1i}|h_i = 1\} \\ &= x_{1i}'\beta_1 + E\{\varepsilon_{1i}|\varepsilon_{2i} > -x_{2i}'\beta_2\} \\ &= x_{1i}'\beta_1 + \frac{\sigma_{12}}{\sigma_2^2}E\{\varepsilon_{2i}|\varepsilon_{2i} > -x_{2i}'\beta_2\} \\ &= x_{1i}'\beta_1 + \sigma_{12}\frac{\phi(x_{2i}'\beta_2)}{\Phi(x_{2i}'\beta_2)}, \end{aligned} \tag{7.84}$$

the inverse Mill's ratio: Heckman's lambda

Although the tobit II model can be motivated in different ways, we shall more or less follow Gronau (1974) in his reasoning. Assume that the utility maximization problem of the individual (in Gronau's case: housewives) can be characterized by a **reservation wage** w_i^r (the value of time). An individual will work if the actual wage she is offered exceeds this reservation wage. The reservation wage of course depends upon personal characteristics, via the utility function and the budget constraint, so that we write (assume)

$$w_i^r = z_i'\gamma + \eta_i,$$

where z_i is a vector of characteristics and η_i is unobserved. Usually the reservation wage is not observed.

Now assume that the wage a person is offered depends on her personal characteristics (and some job characteristics) as in (7.80), i.e.

$$w_i^* = x_{1i}'\beta_1 + \varepsilon_{1i}.$$

If this wage is below w_i^r individual i is assumed not to work. We can thus write her labour supply decision as

$$\begin{aligned} h_i &= 1 \quad \text{if } w_i^* - w_i^r > 0 \\ &= 0 \quad \text{if } w_i^* - w_i^r \leq 0 \end{aligned}$$

$$h_i^* \equiv w_i^* - w_i^r = x_{1i}'\beta_1 - z_i'\gamma + (\varepsilon_{1i} - \eta_i) = x_{2i}'\beta_2 + \varepsilon_{2i}, \tag{7.85}$$

$$y_i^* = x_{1i}'\beta_1 + \varepsilon_{1i} \tag{7.86}$$
$$h_i^* = x_{2i}'\beta_2 + \varepsilon_{2i} \tag{7.87}$$
$$y_i = y_i^*, h_i = 1 \quad \text{if } h_i^* > 0 \tag{7.88}$$
$$y_i \text{ not observed}, h_i = 0 \quad \text{if } h_i^* \leq 0, \tag{7.89}$$

where

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim NID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}\right). \tag{7.90}$$

Estimation:

$$\log L_3(\beta, \sigma_1^2, \sigma_{12}) = \sum_{i \in I_0} \log P\{h_i = 0\} + \sum_{i \in I_1} [\log f(y_i | h_i = 1) + \log P\{h_i = 1\}]. \tag{7.91}$$

Heckman's two-step estimation

$$y_i = x_{1i}'\beta_1 + \sigma_{12}\lambda_i + \eta_i, \tag{7.97}$$

where

$$\lambda_i = \frac{\phi(x_{2i}'\beta_2)}{\Phi(x_{2i}'\beta_2)}.$$

- Step 1: Estimate the selection equation(7.87) using Probit model. Obtain the estimate $\hat{\beta}_2$ and compute the Inverse Mills Ratio.
- Step 2: Run OLS on the outcome equation(7.97) using only the selected sample $(h_i = 1)$. Obtain estimates $\hat{\beta}_1$ and $\hat{\lambda}$.

In Heckman two-step estimation, since the error from the first step carries over to the second step, its efficiency is lower than the overall estimation of MLE. The advantage of the two-step method lies in its simplicity of operation and weaker assumptions about the distribution. Moreover we can get the probability $P(h_i = 1)$ in the first stage.

Illustration: Expenditures on Alcohol and Tobacco (Part 2)

Table 7.8 Models for budget shares alcohol and tobacco, estimated by OLS using positive observations only

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	0.0527	(0.0439)	0.4897	(0.0741)
age class	0.0078	(0.0110)	−0.0315	(0.0206)
nadults	−0.0131	(0.0163)	−0.0130	(0.0324)
nkids ≥ 2	−0.0020	(0.0006)	0.0013	(0.0011)
nkids < 2	−0.0024	(0.0023)	−0.0034	(0.0045)
$\log(x)$	−0.0023	(0.0032)	−0.0336	(0.0055)
age $\times \log(x)$	−0.0004	(0.0008)	0.0022	(0.0015)
nadults $\times \log(x)$	0.0008	(0.0012)	0.0011	(0.0023)
	$R^2 = 0.051$	$s = 0.0215$	$R^2 = 0.154$	$s = 0.0291$
	$N = 2258$		$N = 1036$	

Table 7.9 Probit models for abstention of alcohol and tobacco

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	−15.882	(2.574)	8.244	(2.211)
age	0.6679	(0.6520)	−2.4830	(0.5596)
nadults	2.2554	(1.0250)	0.4852	(0.8717)
$nkids \geq 2$	−0.0770	(0.0372)	0.0813	(0.0308)
$nkids < 2$	−0.1857	(0.1408)	−0.2117	(0.1230)
$\log(x)$	1.2355	(0.1913)	−0.6321	(0.1632)
$age \times \log(x)$	−0.0448	(0.0485)	0.1747	(0.0413)
$nadults \times \log(x)$	−0.1688	(0.0743)	−0.0253	(0.0629)
blue collar	−0.0612	(0.0978)	0.2064	(0.0834)
white collar	0.0506	(0.0847)	0.0215	(0.0694)
Loglikelihood	−1159.865		−1754.886	
Wald test (χ_9^2)	173.18	($p = 0.000$)	108.91	($p = 0.000$)

For alcoholic beverages, total expenditure, the number of adults in the household as well as the number of children older than 2 are statistically significant in explaining abstention.

For tobacco, total expenditure, number of children older than 2, age and being a blue-collar worker are statistically important explanators for abstention.

Table 7.10 Two-step estimation of Engel curves for alcohol and tobacco (tobit II model)

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	0.0543	(0.0487)	0.4516	(0.0735)
age class	0.0077	(0.0110)	−0.0173	(0.0206)
nadults	−0.0133	(0.0166)	−0.0174	(0.0318)
$nkids \geq 2$	−0.0020	(0.0006)	0.0008	(0.0010)
$nkids < 2$	−0.0024	(0.0023)	−0.0021	(0.0045)
$\log(x)$	−0.0024	(0.0035)	−0.0301	(0.0055)
$age \times \log(x)$	−0.0004	(0.0008)	0.0012	(0.0015)
$nadults \times \log(x)$	0.0008	(0.0012)	−0.0041	(0.0023)
λ	−0.0002	(0.0028)	−0.0090	(0.0026)
$\hat{\sigma}_1$	0.0215	n.c.	0.0291	n.c.
Implied ρ	−0.01	n.c.	−0.31	n.c.
	$N = 2258$		$N = 1036$	

For alcoholic beverages the inclusion of λ does not affect the results very much and we obtain estimates that are pretty close to those reported in Table 7.8. The t-statistic on the coefficient for λ does not allow us to reject the null hypothesis of no correlation, while the estimation results imply an estimated correlation coefficient of only −0.01.

$$\rho = \frac{\hat{\sigma}_1}{\lambda}$$

For tobacco, on the other hand, we do find a significant impact of the sample selection term λ, with an implied estimated correlation coefficient of −0.31.