Profesor: Héctor Bahamonde, PhD.

e:hector.bahamonde@uoh.cl

w:www.hectorbahamonde.com

Curso: MLE.

TA: Gonzalo Barría.

I. Outcomes Censurados/Truncados

En esta clase pensaremos en la variables dependientes truncadas o censuradas $y_\star$. *Censored* es diferente a *truncated*. Como veremos en esta clase,

1. **Censuradas**: Las variables dependientes censuradas $y_i^\star$ son aquellas donde a pesar de observar todo el dataset, sólo tenemos información parcial de y_i.
2. **Truncadas**: Las variables dependientes truncadas $y_i^\star$ son aquellas donde (por algún motivo) no vemos todo el dataset, y en consecuencia sólo tenemos información parcial de y_i.

También es posible ver las diferencias conceptuales entre variables truncadas y censuradas de manera gráfica:

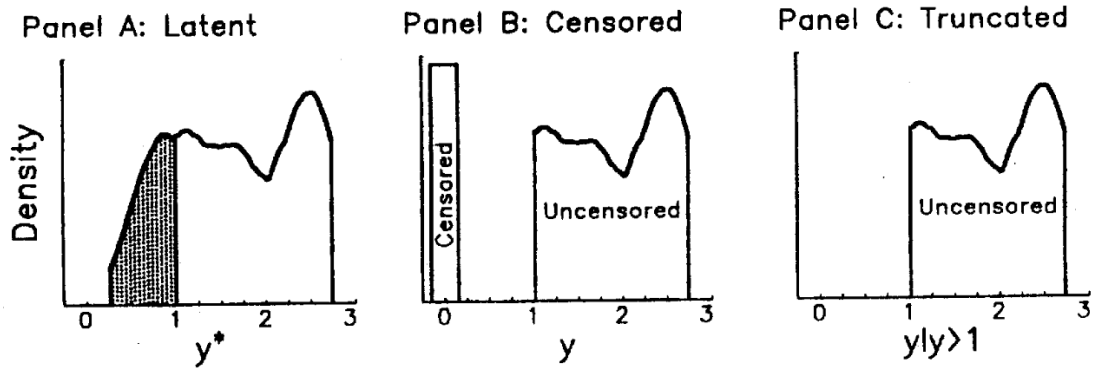


Figure 7.1. Latent, Censored, and Truncated Variables

Afortunadamente tenemos los modelos `tobit` para estimarlos. Como veremos, ambos procesos (*censored*/*truncated*) asumen supuestos distribucionales diferentes.

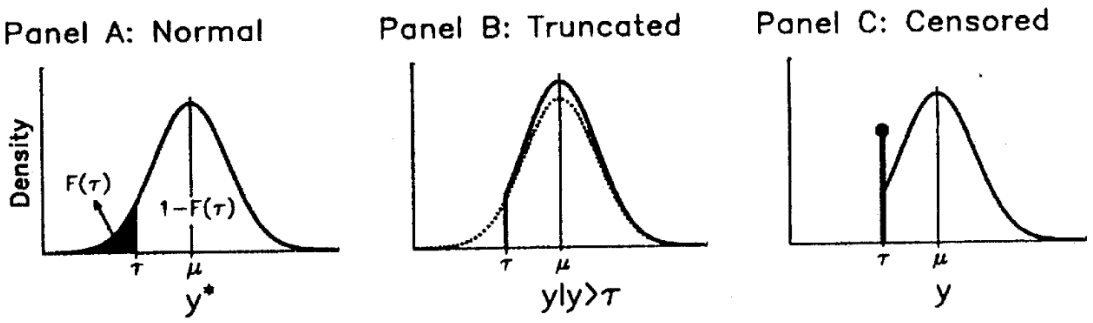


Figure 7.3. Normal Distribution With Truncation and Censoring

Como ves en la Figura 7.3, las distribuciones de $y_i^\star$ varían según el tipo de ignorancia con el que estamos lidiando. **Nota que uno de los supuestos de la especificación del modelo `tobit` es que $y_i^\star$ está normalmente distribuido**.

I. Censored Data

Las variables dependientes censuradas $y_i^\star$ son aquellas donde a pesar de observar todo el dataset, sólo tenemos información parcial de y_i. El típico ejemplo es cuando en las encuestas preguntan por los ingresos, y las escalas están truncadas: *Menos de $100* (1) o *Más de $1.000* (10). En el primer caso, respondentes que ganan $25 serán codificados como "1" al igual que alguien que gana $95. Esto se llama *left-censoring*. A la derecha (*right-censoring*) sería el caso de alguien que gane $1.500: ella será codificada como "10" al igual que alguien que gane $999. Nota que observas todo el dataset: es sólo que la estructura de la variable "censura" cierta información, convirtiendo y_i en $y_i^\star$.

Continuaremos motivando el modelo `tobit` vía modelos latentes. Existe una variable no censurada o truncada y_i que no podemos observar. En vez, observamos la variable censurada $y_i^\star$, donde,

$$y_i = \begin{cases} y_i^\star & \text{si } y_i^\star > 1 \\ 0 & \text{si } y_i^\star \leq 1 \end{cases}$$

En este ejemplo seguimos la Figura 7.2 para el caso específico de que veamos $y_i^\star$ sólo cuando $y_i^\star \geq 1$, pero *no* veamos nada cuando $y_i^\star \leq 1$. Obviamente, el umbral puede variar según la especificación.

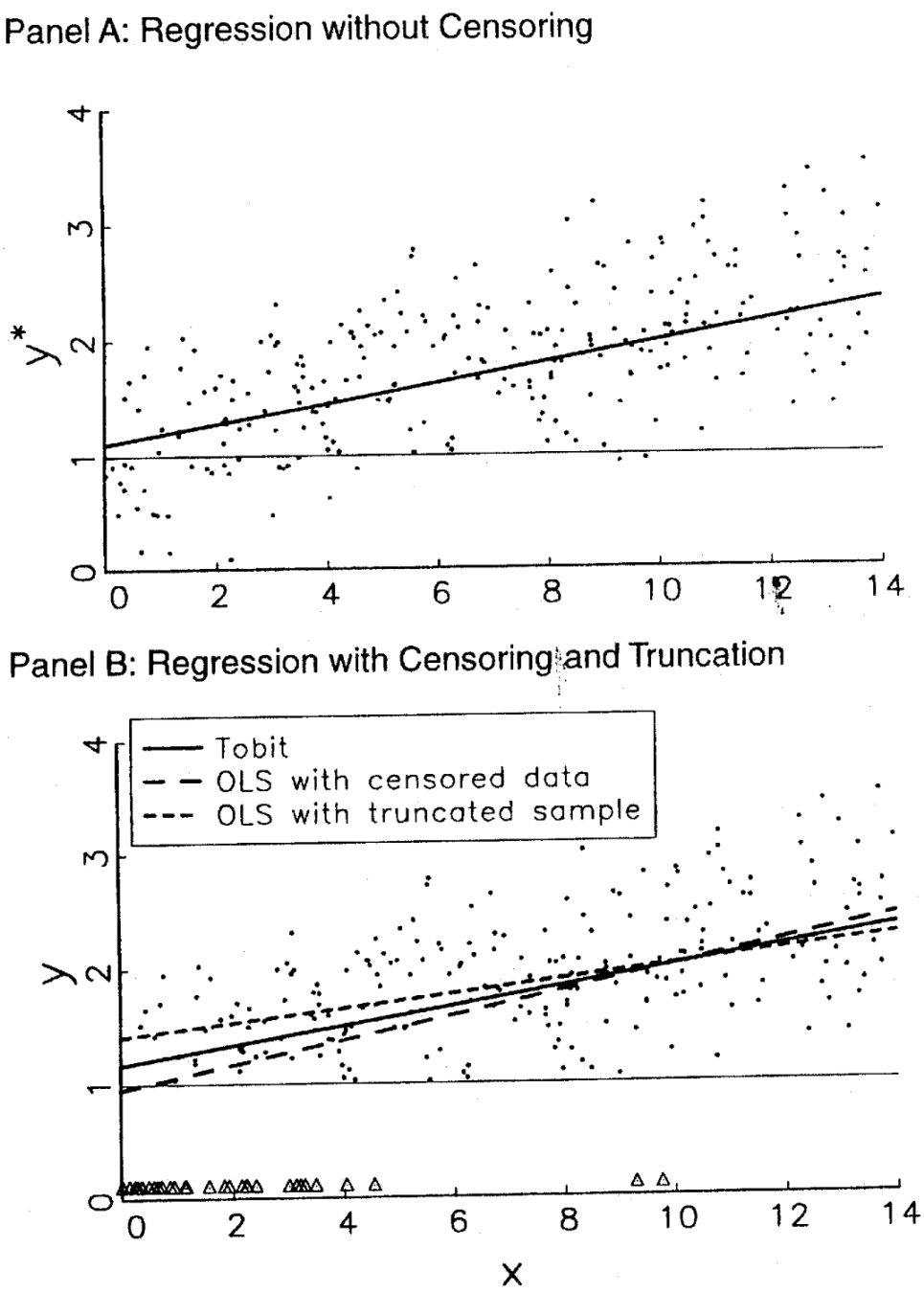


Figure 7.2. Linear Regression Model With and Without Censoring and Truncation

Como lo muestra Long (1997, p. 189), los coeficientes $\boldsymbol{\beta}$ estarán mal estimados si se usan métodos lineales. Para ello existe el modelo `tobit`.

Supuestos distribucionales La distribución normal truncada:

$$f(y|y > \tau, \mu, \sigma) = \frac{\frac{1}{\sigma}\Phi(\frac{y_i^\star - \mu}{\sigma})}{1 - \Phi(\frac{\tau - \mu}{\sigma})} \tag{1}$$

donde Φ es el *probability density function* de la distribución normal estándard, y τ es un "threshold" (o "umbral") bajo el cual se censuran valores de y_i.

Es importante también notar que $\Phi(\frac{\tau-\mu}{\sigma})$ en Equation 1 es la probabilidad de que y_i esté censurado (Long 1997, p. 195). Esto implica que que la probabilidad de *no* estar censurado es $1 - \Phi(\frac{\tau-\mu}{\sigma})$.

Como explica Long (1997, p. 199), existe un link entre el modelo `tobit` y el *probit*. Nota que no sólo ambos modelos estructurales son iguales, sino que también (y tal como lo mostramos en Equation 3), la probabilidad de que una observación sea censurada se calcula usando un modelo parecido al *probit*.

Estimación El modelo estructural `tobit` está dado por algo que ya debiera ser familiar,

$$y_i^\star = \boldsymbol{x}_i\boldsymbol{\beta} + e_i \tag{2}$$

donde e_i es homoesquedástico y $e_i \sim N(0, \sigma^2)$ (Long 1997, p. 206). Esto implica que un modelo `tobit` con *right-censoring* está dado por,

$$y_i = \begin{cases} y_i^\star = \boldsymbol{x}_i\boldsymbol{\beta} + e_i & \text{si } y_i^\star < 1 \\ \tau_y & \text{si } y_i^\star \geq \tau \end{cases}$$

Lo interesante de la especificación `tobit` es que modela (mediante supuestos distribucionales) la probabilidad de que un sujeto sea censurado, de manera tal que,

$$Pr(y_i^\star \leq \tau|\boldsymbol{x}_i) = Pr(\epsilon_i \leq \tau - \boldsymbol{x}_i\boldsymbol{\beta}|\boldsymbol{x}_i) = \Phi(\frac{\tau - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}) \tag{3}$$

Esta información es muy útil. Debido a que no sabemos los valores de $y_i^\star \leq \tau$, nosotros usamos

la proabilidad de ser censurado como si fuera el likelihood (Long 1997, pp. 204–205). En otras palabras, debido a que no conocemos los valores de $y_i^\star \leq \tau$, no podemos usar el alto de la curva normal para estimar el likelihood (ver Figura 7.8). Afortunadamente, debido a que conocemos la igualdad $y_i^\star \leq \tau$, y que observamos $\boldsymbol{x}$, podemos calcular Equation 3.

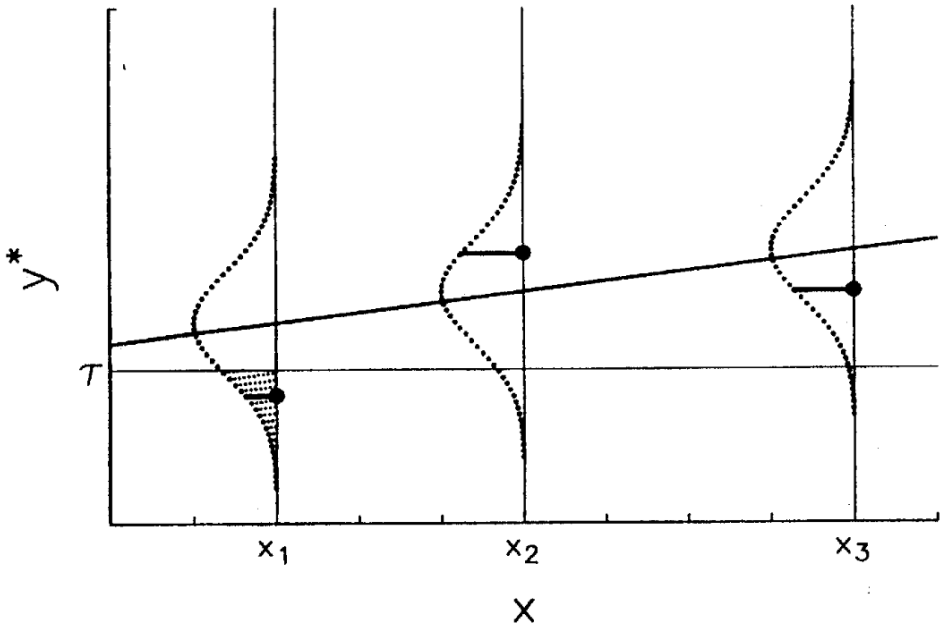


Figure 7.8. Maximum Likelihood Estimation for the Tobit Model

En consecuencia, la función likelihood para *uncensored outcomes* ("U") está dada por,

$$lnL_U(\boldsymbol{\beta}, \sigma^2) = \sum_U ln\frac{1}{\sigma}\Phi - (\frac{y_i - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}) \tag{4}$$

mientras que la función likelihood para *censored outcomes* ("C") está dada por,

$$lnL_C(\boldsymbol{\beta}, \sigma^2) = \sum_C ln\Phi - (\frac{\tau - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma}) \tag{5}$$

II. Truncated Data

El modelo estructural de los procesos truncados es igual. Lo mismo ocurre con su likelihood. La única diferencia es que el likelihood debe ser ajustado por el área de la distribución normal que ha

sido truncada. **Nota, en consecuencia, que el supuesto de la normalidad es fundamental en el modelo `tobit`.**

II. Programación

Greene (2011, p. 811) explica que la base de datos **`Affairs`** pregunta el número de relaciones extra maritales en un año.

Carguemos los datos:

```
library(AER)
data("Affairs")
```

Hagamos un resumen de la variable dependiente,

```
table(Affairs$affairs)

##
##   0   1   2   3   7  12
## 451  34  17  19  42  38
```

Como ves, los códigos "7" y "12" están censurados. El primero ("7") significa 4-10 *affairs*, y "12" una categoría superior.

El paquete de `R` que usaremos se llama **`AER`**, y la función se llama **`tobit`**.

```
tobit.m <- tobit(affairs ~ age + yearsmarried + religiousness + occupation + rating,
                 right = 4,
                 data = Affairs)
```

Fíjate que hemos declarado que tenemos *right-censoring* desde el código "4". Hagamos una tabla.

```
p_load(texreg)
texreg(tobit.m) # usa "screenreg" no "texreg".
```

	Model 1
(Intercept)	7.90^{**}
	(2.80)
age	-0.18^{*}
	(0.08)
yearsmarried	0.53^{***}
	(0.14)
religiousness	-1.62^{***}
	(0.42)
occupation	0.32
	(0.25)
rating	-2.21^{***}
	(0.45)
Log(scale)	2.07^{***}
	(0.11)
AIC	1014.09
BIC	1044.88
Log Likelihood	−500.04
Deviance	581.31
Total	601
Left-censored	451
Uncensored	70
Right-censored	80
Wald Test	42.56

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: *Statistical models*

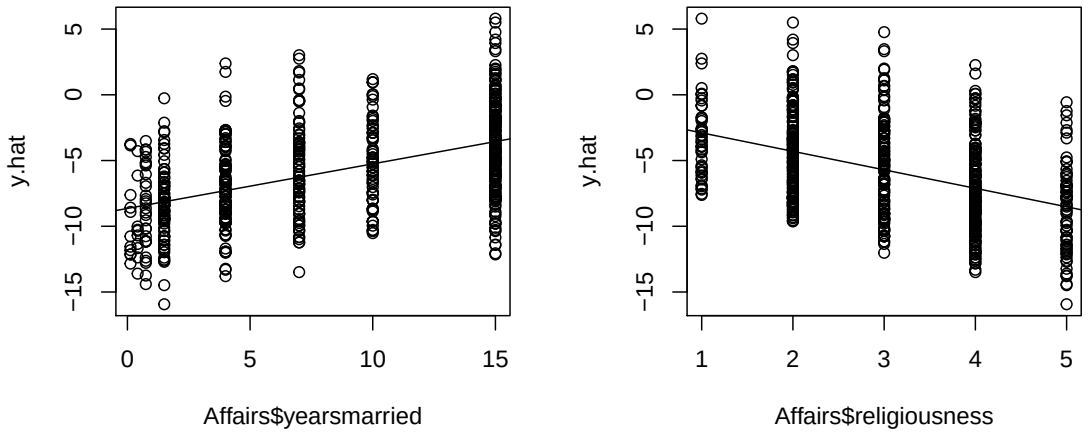
Interpretación Debido a que nuestra variable es continua, los modelos `tobit` se pueden interpretar igual que un modelo OLS. **Interpretemos la tabla**.

Estimemos las *predicted probabilities*,

```
y.hat = predict(tobit.m, interval = "prediction")
```

Ahora hagamos dos gráficos, que por propositos demostrativos, sólo seleccionaremos dos variables.

```
par(mfrow=c(1,2))
plot(Affairs$yearsmarried, y.hat);abline(lm(y.hat ~ Affairs$yearsmarried))
plot(Affairs$religiousness, y.hat);abline(lm(y.hat ~ Affairs$religiousness))
```



Finalmente, hagamos un diagnóstico rápido. Veamos los residuos. Primero, hagamos el cálculo.

```
e = Affairs$affairs - y.hat
```

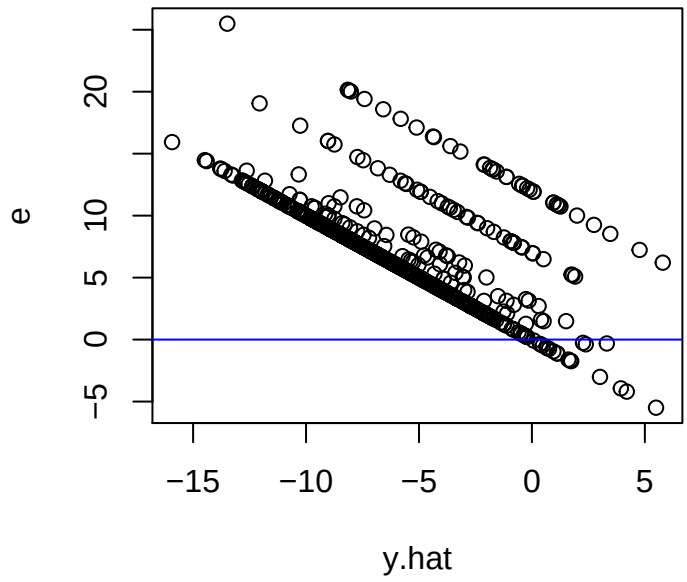
Veamos si tienen promedio cero.

```
mean(e)

## [1] 7.326525
```

Ahora veamos el gráfico. Siempre gráfica tus residuos.

```
plot(y.hat, e)
abline(h=0, col="blue")
```




```
knitr::purl('Tobit.Rnw')

## [1] "Tobit.R"

Stangle('Tobit.Rnw')

## Writing to file Tobit.R

## Error in match.arg(options$results, c("verbatim", "tex", "hide")):  'arg' should
be one of "verbatim", "tex", "hide"
```

References

Greene, William (2011). *Econometric Analysis*. 7th. Prentice Hall.

Long, Scott (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage, p. 297.