Tobit Regression: modelos lineales para datos censurados

Joaquín Amat Rodrigo j.amatrodrigo@gmail.com

Abril, 2018

Tabla de contenidos

Versión PDF: Github

Introducción

Tal y como se ha visto en documentos anteriores, el método de regresión por mínimos cuadrados ordinarios OLS (regresión lineal simple y regresión lineal múltiple) es muy útil para estudiar una variable respuesta continua en función de uno o más predictores. Sin embargo, bajo determinadas circunstancias y a pesar de existir una relación lineal entre las variables, el método de mínimos cuadrados ordinarios puede no ser adecuado.

Datos censurados:

Se considera que los datos están censurados (*censored*) cuando existe un determinado límite en la variable respuesta (superior, inferior o ambos) a partir del cual a todas las observaciones se les asigna un mismo valor. Algunos ejemplos de datos censurados son:

- Un instrumento de medida, por ejemplo una balanza, tiene un límite de detección por debajo del cual todo valor se considera de 0 (censura inferior).
- En una encuesta se pregunta por el nivel de ingresos de las personas. Se divide la escala en múltiples intervalos, el último de los cuales, contempla como iguales a todas aquellas personas que cobren 5000 euros o más (censura superior).

La característica fundamental de un escenario censurado es que hay una población subyacente en la que sí existen observaciones fuera de los límites de censura, sin embargo, debido a la incapacidad para detectarlas/seleccionarlas en el muestreo, la población observada parece no contenerlas.

Datos truncados:

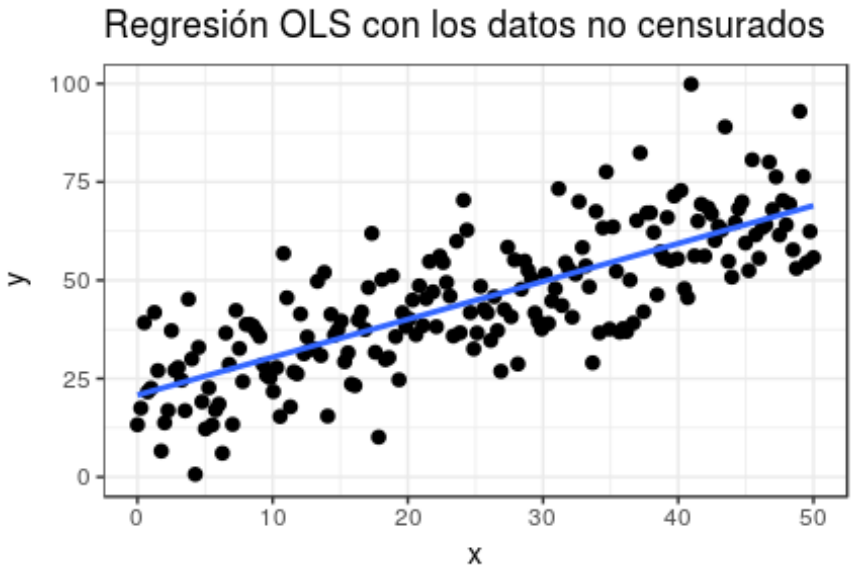
Las situaciones con datos truncados aparecen cuando en una población a partir de un determinado límite no existen observaciones. La diferencia fundamental respecto a los datos censurados es que, en estos últimos, las observaciones sí existen en la población latente pero no se pueden captar en el muestreo.

Aunque la diferencia puede parecer mínima, es muy importante tenerla en cuenta puesto que, si el objetivo último de la inferencia es obtener información sobre la población real, en el caso de escenarios censurados, hay que incluir de alguna forma esos eventos que existen pero no se observan. Este documento se centra en el tratamiento de casos censurados.

Las siguientes imágenes muestran una representación gráfica e intuitiva del impacto que tienen los datos censurados sobre la regresión por mínimos cuadrados. Supónganse dos variables X e Y que tienen una relación lineal. (Para poder simular los datos se considera que la relación sigue la ecuación $Y = X + 2 + error$).

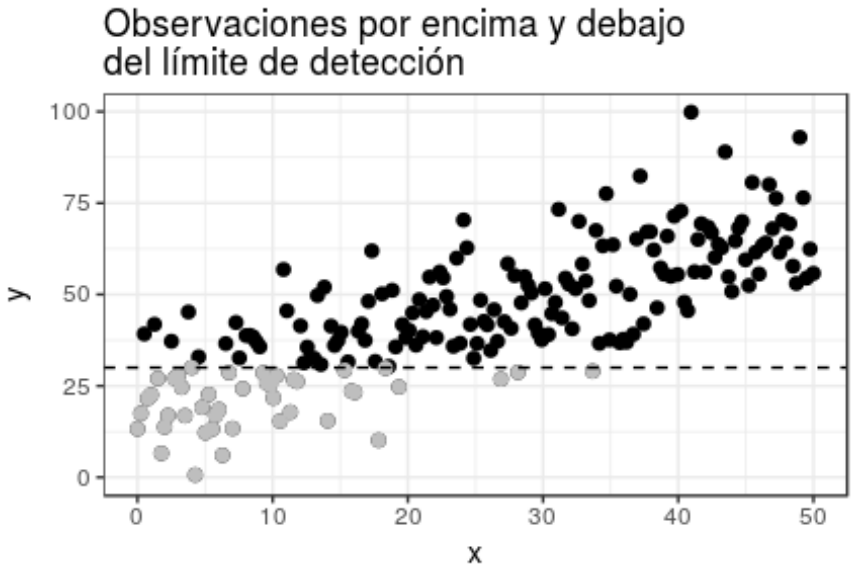
```r
library(ggplot2)
library(dplyr)
library(gridExtra)
simulador <- function(x){x + 20}
observaciones <- simulador(seq(0,50,length.out = 200))
# Para introducir la variancia propia de los datos observacionales se añade
# ruido aleatorio y normal a cada observación
set.seed(123)
observaciones <- observaciones + rnorm(n = 200, mean = 0, sd = 12)

datos <- data.frame(x = seq(0,50,length.out = 200), y = observaciones)
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point(size = 2) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw() +
  lims(y = c(0,100)) +
  labs(title = "Regresión OLS con los datos no censurados")
```



Supóngase ahora que los investigadores no son capaces de cuantificar valores de la variable Y inferiores a 30, a pesar de que sí que son capaces de detectarlos y por lo tanto saben que existen. La muestra observada pasaría a ser la siguiente:

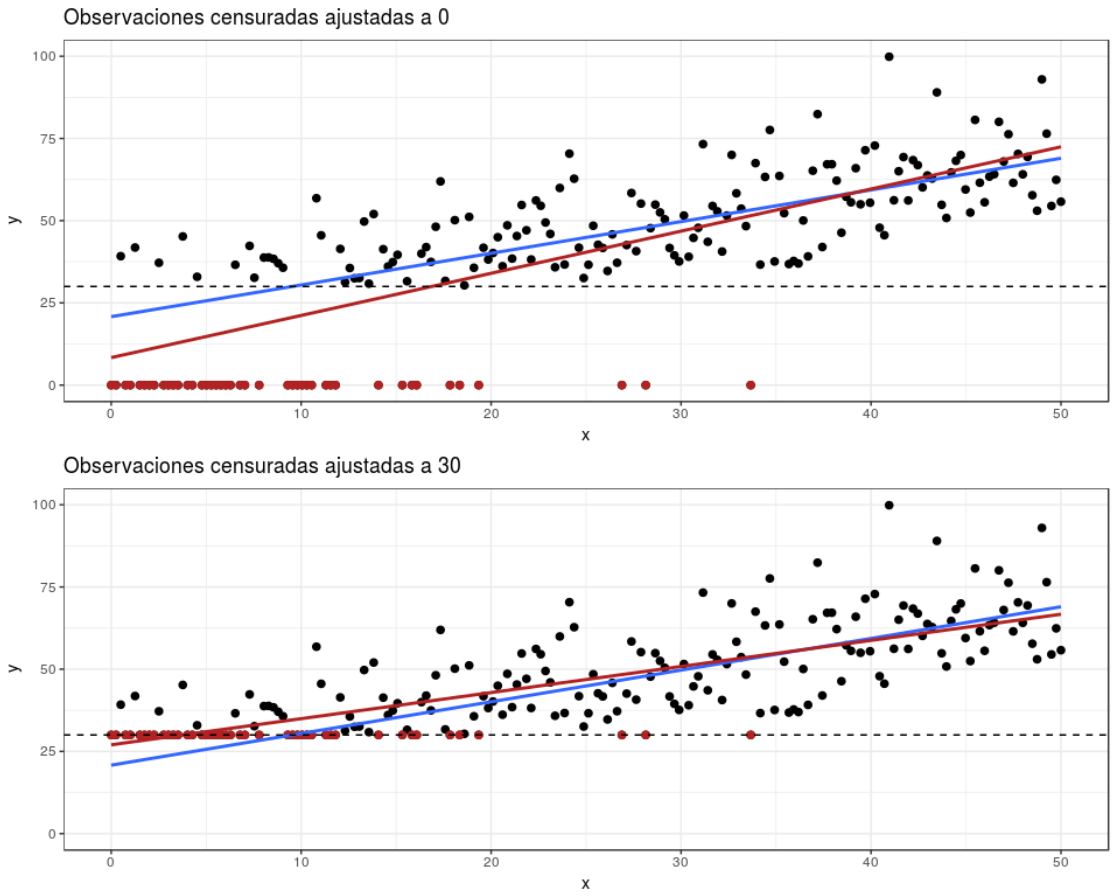
```
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point(size = 2) +
  geom_point(data = filter(datos, y < 30), size = 2, color = "grey") +
  geom_hline(yintercept = 30, linetype = "dashed") +
  theme_bw() +
  lims(y = c(0,100)) +
  labs(title = "Observaciones por encima y debajo\ndel límite de detección")
```



Dada esta situación, los investigadores tienen 2 opciones: considerar todos los valores por debajo del límite de detección como 0 o como 30. Cuál de las dos opciones es más correcta depende del tipo de estudio que se esté haciendo. En las dos imágenes siguientes se puede ver cómo impacta cada tipo de censura al ajuste por mínimos cuadrados.

```
datos_0 <- datos
datos_0$y[datos_0$y < 30] <- 0
p1 <- ggplot(data = datos_0, aes(x = x, y = y)) +
      geom_point(size = 2) +
      geom_point(data = filter(datos_0, y == 0), size = 2,
                 color = "firebrick") +
      geom_smooth(data = datos,method = "lm", se = FALSE) +
      geom_smooth(data = datos_0, method = "lm", se = FALSE,
                  color = "firebrick") +
      geom_hline(yintercept = 30, linetype = "dashed") +
      lims(y = c(0,100)) +
      theme_bw() +
      labs(title = "Observaciones censuradas ajustadas a 0")
```

```
datos_30 <- datos
datos_30$y[datos_30$y < 30] <- 30
p2 <- ggplot(data = datos_30, aes(x = x, y = y)) +
      geom_point(size = 2) +
      geom_point(data = filter(datos_30, y == 30), size = 2,
                 color = "firebrick") +
      geom_smooth(data = datos,method = "lm", se = FALSE) +
      geom_smooth(data = datos_30, method = "lm", se = FALSE,
                  color = "firebrick") +
      geom_hline(yintercept = 30, linetype = "dashed") +
      theme_bw() +
      lims(y = c(0,100)) +
      labs(title = "Observaciones censuradas ajustadas a 30")
grid.arrange(p1, p2, ncol = 1)
```



En el primer caso (imagen izquierda), la recta de regresión aumenta su inclinación debido a la influencia que ejercen todas aquellas observaciones que antes tenían valores entre 0 y 30 y que ahora son exactamente 0. En el segundo caso (imagen derecha) el efecto es el contrario, ya que muchas observaciones han pasado a tener el valor de 30 cuando antes eran inferiores.

Además del evidente impacto sobre la pendiente de la recta de regresión, se unen otros dos factores que invalidan la utilización de mínimos cuadrado OLS con datos censurados. A medida que se aproxima el límite de censura y más observaciones son fijadas al valor establecido, la varianza se reduce, perdiéndose así la condición de homocedasticidad (varianza constante de los residuos) y de independencia. Ambas necesarias para considerar válido el método de regresión por mínimos cuadrados.

Modelo de Tobit

En la regresión de Tobit (1958) se considera que existe una variable latente Y^* no observable y una variable Y observable, formada por la parte no censurada de Y^*. El objetivo es ser capaz de estimar parámetros de Y^* empleando solo una muestra de la parte observable. Para ello, asumiendo censura por el límite inferior (misma idea para la superior), considera que el valor esperado de la variable censurada Y puede definirse como:

$$E(y) = [P(No\ censurado)\ x\ E(y|y > \tau)] + [P(censurado)\ x\ E(y|y = \tau_y)]$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > \tau \\ \tau_y & \text{if } y_i^* \leq \tau \end{cases}$$

donde P es probabilidad, τ límite de censura, τ_y el valor que se le asigna a la variable latente $Y*$ cuando se le aplica la censura, $y|y > \tau$ y $y|y > \tau_y$ son el condicional de que Y sea mayor que el límite de censura y el valor asignado a la censura respectivamente.

Lo que hace útil al método de Tobit es que permite hacer estimaciones de Y^* a pesar de tener observaciones únicamente de Y.

Ejemplo

Considérese un estudio en el que se pretende crear un modelo que permita predecir el nivel académico (escala 200-800) de los estudiantes de una universidad. Para ello se emplea la nota obtenida en un examen de lectura, un examen de matemáticas y el tipo de programa al que están matriculados (académico, general o vocacional).

Se trata de un modelo censurado ya que todos aquellos estudiantes que contesten correctamente todas las preguntas obtendrán un 800, a pesar de que no todos ellos tengan exactamente el mismo nivel académico. Lo mismo ocurre para los estudiantes que contesten mal todas las preguntas, todos obtendrán un 200.

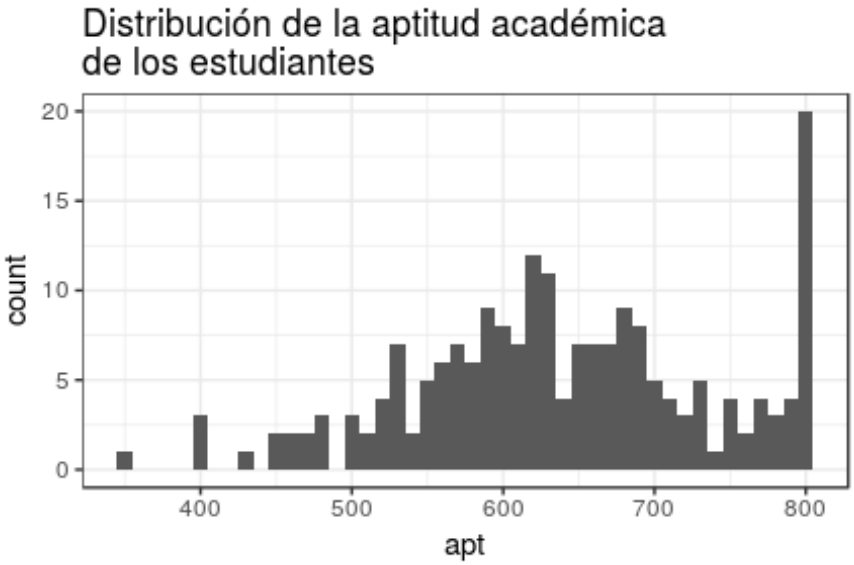
```
library(ggplot2)
datos <- read.csv("https://stats.idre.ucla.edu/stat/data/tobit.csv")
head(datos)
```

```
##   id read math       prog apt
## 1  1   34   40 vocational 352
## 2  2   39   33 vocational 449
## 3  3   63   48    general 648
## 4  4   44   41    general 501
## 5  5   47   43    general 762
## 6  6   47   46    general 658
```

```
summary(datos)
```

```
##        id              read            math              prog
##  Min.   :  1.00   Min.   :28.00   Min.   :33.00   academic  : 45
##  1st Qu.: 50.75   1st Qu.:44.00   1st Qu.:45.00   general   :105
##  Median :100.50   Median :50.00   Median :52.00   vocational: 50
##  Mean   :100.50   Mean   :52.23   Mean   :52.65
##  3rd Qu.:150.25   3rd Qu.:60.00   3rd Qu.:59.00
##  Max.   :200.00   Max.   :76.00   Max.   :75.00
##       apt
##  Min.   :352.0
##  1st Qu.:575.5
##  Median :633.0
##  Mean   :640.0
##  3rd Qu.:705.2
##  Max.   :800.0
```

```
ggplot(data = datos, aes(x = apt)) + geom_histogram(binwidth = 10) + theme_bw() +
    labs(title = "Distribución de la aptitud académica\nde los estudiantes")
```



La exploración de los datos muestra que la puntuación mínima obtenida es de 352, por lo que, aun siendo posible la censura por el límite inferior, no ocurre en este set de datos. En el extremo superior de la distribución sí se observa incremento de eventos con una puntuación de 800 muy por encima de lo esperado acorde a una distribución normal (la que suele seguir este tipo de puntuaciones), lo que pone de manifiesto que sí hay censura.

Al conocerse la naturaleza de los datos, se sabe que se trata de datos censurados y no de datos truncados. A pesar de ello, una pista que ayuda a diferenciar ambos escenarios es que, cuando se trata de datos truncados en los que no existen observaciones más allá de un límite, no hay acumulación en el límite, simplemente se corta la distribución.

La función vglm() del paquete VGAM permite ajustar modelos mediante el método de Tobit.

```
library(VGAM)
modelo_tobit <- vglm(apt ~ read + math + prog, tobit(Upper = 800), data = datos)
summary(modelo_tobit)
```

```
##
## Call:
## vglm(formula = apt ~ read + math + prog, family = tobit(Upper = 800),
##     data = datos)
##
```

```
##
## Pearson residuals:
##                Min      1Q   Median     3Q   Max
## mu       -2.5684 -0.7311 -0.03976 0.7531 2.802
## loge(sd) -0.9689 -0.6359 -0.33365 0.2364 4.845
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  209.55956   32.54590   6.439 1.20e-10 ***
## (Intercept):2    4.18476    0.05235  79.944  < 2e-16 ***
## read             2.69796    0.61928   4.357 1.32e-05 ***
## math             5.91460    0.70539   8.385  < 2e-16 ***
## proggeneral    -12.71458   12.40857  -1.025 0.305523
## progvocational -46.14327   13.70667  -3.366 0.000761 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  2
##
## Names of linear predictors: mu, loge(sd)
##
## Log-likelihood: -1041.063 on 394 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
```

Los coeficientes de regresión devueltos por un modelo Tobit se interpretan igual que los devueltos por un modelo de mínimos cuadrados OLS, pero referidos a la variable latente NO censurada.

La función censReg() del paquete censreg es otra de las múltiples implementaciones en R de modelos de regresión censurados.

```
library(censReg)
modelo_tobit <- censReg(apt ~ read + math + prog, right = 800, data = datos)
summary(modelo_tobit)
```

```
##
## Call:
## censReg(formula = apt ~ read + math + prog, right = 800, data = datos)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##            200              0            183              17
##
```

```
## Coefficients:
##                  Estimate Std. error t value  Pr(> t)
## (Intercept)    209.56597   32.77196   6.395 1.61e-10 ***
## read             2.69794    0.61881   4.360 1.30e-05 ***
## math             5.91448    0.70982   8.332  < 2e-16 ***
## proggeneral    -12.71476   12.40646  -1.025 0.305434
## progvocational -46.14390   13.72419  -3.362 0.000773 ***
## logSigma         4.18474    0.05301  78.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Newton-Raphson maximisation, 8 iterations
## Return code 1: gradient close to zero
## Log-likelihood: -1041.063 on 6 Df
```

Al comparar los resultados obtenidos por censReg() y vglm() parece haber pequeñas diferencias. Posiblemente se deban a que emplean diferentes métodos para maximizar la función de verosimilitud.

Bibliografía

An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)

Linear Models with R, Julian J.Faraway

Points of Significance: Simple linear regression by Naomi Altman & Martin Krzywinski

Points of Significance: Multiple linear regression Martin Krzywinski & Naomi Altman

*https://menghublog.wordpress.com/2014/12/28/the-use-of-tobit-and-truncated-regressions-for-limited-dependent-variables/

https://stats.idre.ucla.edu/r/dae/tobit-models/

Estimating Censored Regression Models in R using the censReg Package, Arne Henningsen University of Copenhagen

This work by Joaquín Amat Rodrigo is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This work by Joaquín Amat Rodrigo is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.