



**Universitat Oberta
de Catalunya**

Datos con límite de detección bajo. Varios enfoques metodológicos

Pedro Camacho Martínez

Plan de Estudios del Estudiante

Estadística y Bioinformática

Rebeca Sanz Pamplona

Nuria Pérez Alvarez

Fecha Entrega: 21/06/17



Esta obra está sujeta a una licencia de
Reconocimiento-NoComercial [3.0](#)
[España de Creative Commons](#)

FICHA DEL TRABAJO FINAL

<i>Título del trabajo:</i>	<i>Datos con límite de detección bajo. Varios enfoques metodológicos</i>
Nombre del autor:	<i>Pedro Camacho Martínez</i>
Nombre del consultor/a:	<i>Rebeca Sanz Pamplona</i>
Nombre del PRA:	<i>Nuria Perez Alvarez</i>
Fecha de entrega (mm/aaaa):	06/2017
Titulación::	<i>Plan de estudios del estudiante</i>
Área del Trabajo Final:	<i>Estadística y Bioinformática</i>
Idioma del trabajo:	<i>Español</i>
Palabras clave	Límites de detección, censura, Tobit
<p style="text-align: center;">Resumen del Trabajo (máximo 250 palabras): <i>Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.</i></p>	
<p>Muchas variables en estudios epidemiológicos corresponden a medidas continuas obtenidas mediante aparatos de medición con determinados límites de detección, produciendo distribuciones censuradas. La censura, a diferencia del truncamiento, se produce por un defecto de los datos de la muestra.</p> <p>La distribución de una variable censurada es una mezcla entre una distribución continua y otra discreta. En este caso, no es adecuado utilizar el modelo de regresión lineal estimado para mínimos cuadrados ordinarios, ya que proporciona estimaciones</p>	

sesgadas. Con un único punto de censura debe utilizarse el modelo de regresión censurado (modelo Tobit), mientras que cuando hay varios puntos de censura se utiliza la generalización de este modelo. La ilustración de estos modelos se presenta a través del análisis de las concentraciones de VHC, correspondientes al estudio sobre los efectos para la salud.

Abstract (in English, 250 words or less):

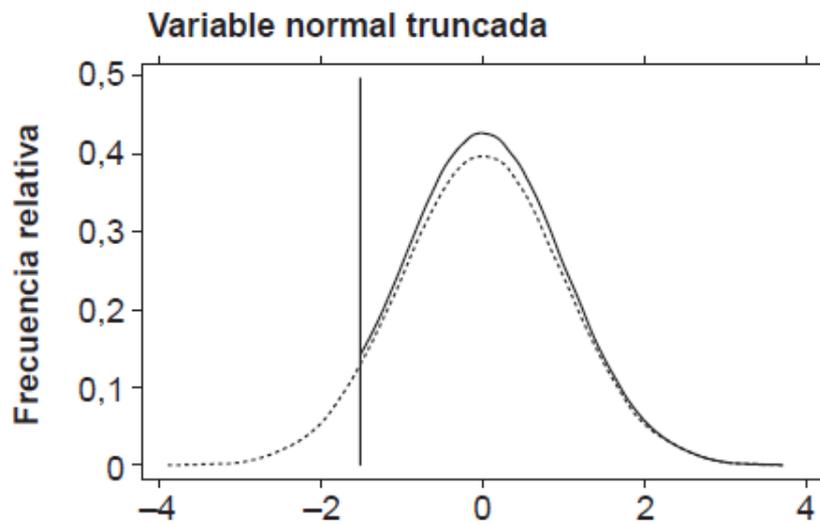
Many variables in epidemiological studies are continuous measures obtained by means of measurement equipments with detection limits, generating censored distributions. The censorship, opposite to the truncation, takes place for a defect of the data of the sample. The distribution of a censored variable is a mixture between a continuous and a categorical distributions. In this case, results from lineal regression models, by means of ordinary least squares, will provide biased estimates. With one only censorship point the Tobit model must be used, while with several censorship points this model's generalization should also be used. The illustration of these models is presented through the analysis of the levels of VHC in the study about health effects.

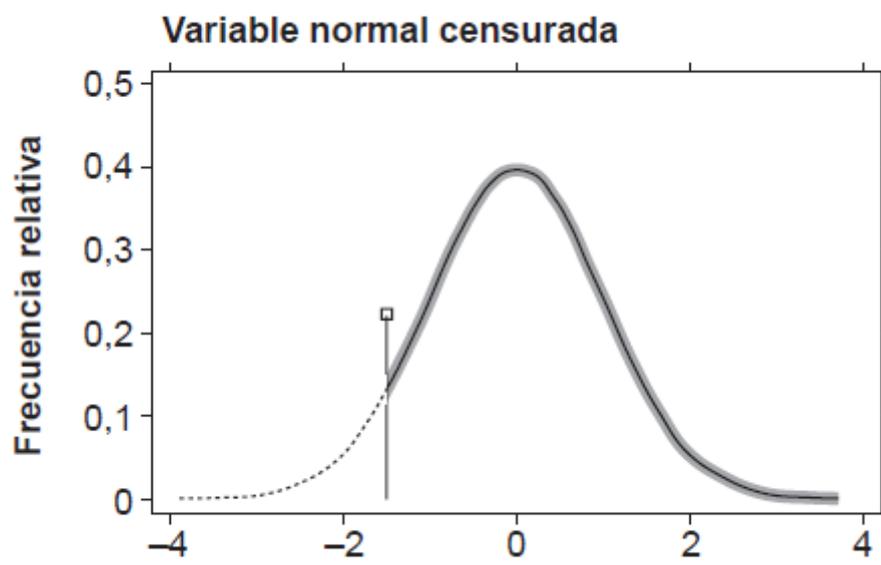
ÍNDICE

1. Introducción.....	1
1.1 Contexto y justificación del Trabajo.....	1
1.2 Objetivos del Trabajo.....	5
1.3 Enfoque y método seguido.....	5
1.4 Planificación del Trabajo.....	7
1.5 Breve resumen de productos obtenidos.....	8
Modelo de regresión censurado con un único punto de censura (modelo Tobit):.....	8
Interpretación de los coeficientes.....	11
Análisis estadístico.....	12
Software estadístico.....	13
2. Resultados.....	13
3. Discusión.....	17
4. Bibliografía.....	19
5. Anexos.....	21
Códigos usados en R:.....	21

Lista de figuras

Figura 1. Distribución normal estándar $N(0,1)$, con un punto de truncamiento inferior y con un único punto de censura inferior.





1.Introducción

1.1 Contexto y justificación del Trabajo

En muchas ocasiones, variables objeto de investigación en estudios epidemiológicos se corresponden a medidas continuas obtenidas mediante aparatos de medición debidamente ajustados y calibrados. Es habitual que dichos aparatos tengan determinados límites de detección, tanto inferiores como superiores. Estos límites pueden hacer que, a pesar de que la variable que nos interesa estudiar tenga una distribución determinada, los valores que realmente se observen en la muestra no sean representativos. Algunos ejemplos los podemos encontrar en la distribución de los niveles de inmunoglobulina E en sangre o los niveles de metales medidos en sangre (1) u orina (2).

El truncamiento es una característica intrínseca de la distribución de la variable objeto de estudio, de la cual se extraen los datos de la muestra. Se produce cuando sólo la parte de la distribución de la variable que se encuentra por encima (o por debajo) del denominado punto de truncamiento contiene la información relevante que se desea estudiar. Un ejemplo de variable truncada sería el valor de hemoglobina cuando el interés reside en estudiar a aquellos pacientes con valores inferiores a 8 g/dl en la población. El punto de truncamiento es 8 g/dl y la variable se dice que está truncada. A nivel teórico, para que la función de densidad de una variable aleatoria truncada integre la unidad, se divide su función de densidad entre la probabilidad de que una observación no pertenezca al área truncada. En la figura 1 se

representa gráficamente cómo afecta el truncamiento a la función de densidad de una distribución normal estándar, con punto de truncamiento inferior $a = -1,5$.

La censura, por el contrario, no es una característica intrínseca de la distribución de la variable objeto de estudio, sino un defecto de los datos de la muestra, que si no estuvieran censurados constituirían una muestra representativa de la población de interés no censurada. Un ejemplo habitual de censura es el que se produce cuando la variable objeto de estudio es el tiempo de supervivencia desde el diagnóstico de una enfermedad hasta la fecha de muerte (evento). En la práctica el estudio tendrá definida una fecha de finalización (punto de censura) en la que ocurrirá que no todos los sujetos de la muestra escogida habrán muerto (algunos seguirán vivos). A pesar de que el objetivo sería estudiar el tiempo de supervivencia en la población de enfermos diagnosticados de dicha enfermedad, no es posible disponer en la muestra de los tiempos de supervivencia de todos los enfermos. La variable tiempo de supervivencia se dice entonces que está censurada superiormente. Cuando la variable está censurada, la distribución que siguen los datos de la muestra es una mezcla entre una distribución continua y otra discreta, existiendo una acumulación de probabilidad en el punto de censura. También en la figura 1, se presenta la función de densidad de una distribución normal estándar, censurada, con un único punto de censura inferior $a = -1,5$.

Si la variable objeto de estudio es una medición continua que se distribuye según una ley normal, en la que existen uno o varios puntos de truncamiento y/o censura, no es posible utilizar los habituales modelos de regresión lineal estimados por mínimos

cuadrados ordinarios (MCO), porque proporcionan estimaciones incorrectas del efecto y de su variabilidad (3,4). Cuando la variable de interés tiene un punto de truncamiento se debe utilizar el denominado modelo de regresión truncado (3,4). Análogamente si tiene un único punto de censura tiene que utilizarse el llamado modelo de regresión censurado o modelo Tobit (5). Cuando existen varios puntos de truncamiento o censura, o cuando coexisten al mismo tiempo censura y truncamiento, se utilizan las respectivas generalizaciones de estos modelos que, desarrollados originalmente en el campo de la econometría, se han aplicado con frecuencia en el campo de la economía de la salud (6-9).

Las concentraciones de VHC en sangre, tienen la particularidad de tener varios puntos de censura inferior en la cola izquierda de la distribución, debidos al límite de detección inferior del aparato de medición. Por eso, se hace necesario utilizar modelos alternativos a los modelos de regresión lineal, estimados por MCO.

La transmisión del VHC es principalmente parenteral, y se hace por contacto directo de la sangre del sujeto infectado con la sangre de la persona expuesta (11), uso de material de inyección no estéril y transfusión de sangre o de productos sanguíneos no analizados (12). La evolución de una infección por el VHC es frecuentemente el paso a una hepatitis C crónica (aproximadamente 80%), exponiendo al individuo infectado a los riesgos de complicaciones hepáticas, como el desarrollo de una cirrosis o la aparición de un carcinoma hepatocelular (12). El tratamiento de referencia actualmente es la asociación de interferon pegilado y ribavirina pero dada la gran variabilidad genética del VHC (13), su

eficacia se mantiene solo parcialmente: eliminación viral para menos del 50% de los pacientes infectados por un virus de genotipo 1 frente a aproximadamente el 80% para los pacientes infectados por el virus de genotipo 2 o 3. Se encuentran en estudio nuevas opciones terapéuticas para proponer tratamientos personalizados más eficaces y más seguros (14,15). Para el diagnóstico de una infección por el VHC, pueden utilizarse dos categorías de pruebas virológicas, indirectas y directas (16). Las primeras, pruebas serológicas inmunoenzimáticas de 3ª generación, ponen en evidencia los anticuerpos (Acs). Los antígenos utilizados en las pruebas para detectar los Acs son regiones estructurales y no estructurales del VHC (17) (cápside, proteasa, cofactores, polimerasa,...). Los Acs anti-HCV detectados son testigo de una infección actual o pasada. Para verificar la presencia activa del virus, la serología positiva puede completarse por pruebas llamadas directas (ej: pruebas moleculares que detectan el ARN genómico). Los resultados se van a utilizar para orientar la gestión del paciente y la duración óptima del tratamiento. La prueba VIDAS Anti-HCV es una determinación de tercera generación que utiliza antígenos que representan las proteínas del core, NS3 y NS4 para la determinación cualitativa de los anticuerpos anti-VHC.

El principio de determinación asocia el método inmunoenzimático por método sandwich en 2 etapa(s) con una detección final por fluorescencia (ELFA). El cono (SPR®) de un solo uso sirve a la vez de fase sólida y de sistema de pipeteo. El resto de los reactivos de la reacción inmunológica están listos al empleo y previamente distribuidos en el cartucho. Todas las etapas de la prueba se realizan automáticamente por el sistema. Están

constituidas por una sucesión de ciclos de aspiración/expulsión del medio de reacción. En la primera etapa la muestra es diluida, después se aspira y expulsa del interior del cono. Los anticuerpos anti-HCV presentes en la muestra se fijan en los antígenos que representan las proteínas del core, NS3 y NS4 fijados en el interior del cono. Las etapas de lavado eliminan los compuestos no fijados. Durante la segunda etapa las IgG monoclonales (ratón) anti-inmunoglobulinas (anti-IgG) humanas en forma de Fab' conjugadas con la fosfatasa alcalina recombinante (levadura) son aspiradas y expulsadas del interior del cono y se ligan a las Ig humanas fijadas en las moléculas de la fase sólida. Unas nuevas etapas de lavado eliminan los compuestos no fijados. Durante la etapa final de revelado, el substrato (4-Metilumbeliferil fosfato) es aspirado y expulsado del cono; la enzima del conjugado cataliza la reacción de hidrólisis de este substrato en un producto (4-Metil-umbeliferona) cuya fluorescencia emitida se mide a 450 nm. El valor de la señal de fluorescencia es proporcional a la concentración de anticuerpos presentes en la muestra. Al finalizar la prueba, los resultados se calculan automáticamente por el sistema respecto al S1 memorizado, y después se imprimen.

1.2 Objetivos del Trabajo

El objetivo de este trabajo es describir las potenciales aplicaciones de la familia de modelos de regresión censurada en la modelización de variables censuradas estudiadas en el laboratorio clínico.

1.3 Enfoque y método seguido

Para contrastar si pudieran existir diferencias estadísticamente significativas entre los sujetos con censura y sin censura, se puede utilizar el test de la suma de rangos de Wilcoxon para las variables continuas, que pone a prueba si los datos de ambos grupos de sujetos proceden de poblaciones con la misma distribución. Para las variables categóricas se puede utilizar el estadístico de contraste de la χ^2 de Pearson, el cual pone a prueba si las filas y las columnas en una tabla de contingencia son independientes.

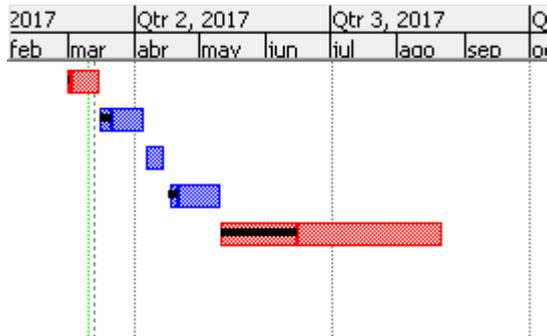
Seguidamente, para cada uno de los tres modelos de regresión analizados se hacen los respectivos modelos de regresión univariantes para cada una de las variables explicativas x_j consideradas. En el primer modelo de regresión analizado, se consideran tan sólo aquellos individuos con valores detectados (la muestra con valores observados) y se estima un modelo de regresión lineal por MCO. En el segundo modelo, se consideraron todos los individuos, asumiendo que todos los sujetos con valores censurados toman el mismo valor mínimo de censura.

Se estima un modelo de regresión lineal censurado con un único punto de censura o modelo Tobit. Por último, en el tercer modelo se consideraron de nuevo todos los individuos, aunque los individuos con valores censurados toman sus respectivos valores de censura. Se estima un modelo de regresión lineal censurado con varios puntos de censura, que es la generalización del modelo Tobit anterior.

Para cada uno de los tres análisis se construyen a continuación los modelos multivariados. Se incluyen todas aquellas

variables cuyo valor de la t de Student para el coeficiente estimado resulta en valor absoluto mayor que 1 en los correspondientes modelos univariados y, posteriormente, se eliminan una a una las variables no significativas (9) hasta configurar los modelos finales.

1.4 Planificación del Trabajo



PEC1
 Duración: 11 days?
 Inicio: 1/03/17 8:00
 Termin...: 15/03/17 17:00

PEC2
 Duración: 15 days?
 Inicio: 16/03/17 8:00
 Termin...: 5/04/17 17:00

VACACIONES(SEMANA SA...
 Duración: 7 days?
 Inicio: 6/04/17 7:00
 Termin...: 14/04/17 17:00

PEC3
 Duración: 18 days?
 Inicio: 17/04/17 8:00
 Termin...: 10/05/17 17:00

PEC4
 Duración: 73 days?
 Inicio: 11/05/17 8:00
 Termin...: 21/08/17 17:00

Se dedicarán 2 horas diarias salvo los domingos.

- 15/03/17: PEC1. Planificación y objetivos.
- 05/04/17: PEC2. Búsqueda de información, contexto y justificación del Trabajo, enfoque y método seguido
- 10/04/17-17/04/17: Vacaciones (Semana Santa).
- 10/05/17: PEC3: Breve resumen de productos obtenidos, breve descripción de los otros capítulos de la memoria.
- 24/05/17: PEC4: Conclusiones, resto de capítulos y correcciones finales.
- 21/06/17: Anexo bibliografía glosario. Entrega TFM.

La planificación, se ha hecho utilizando en todos los plazos 2 días más de trabajo para así, solventar los posibles imprevistos.

1.5 Breve resumen de productos obtenidos

Modelo de regresión censurado con un único punto de censura (modelo Tobit):

Para definir la distribución de la variable censurada, que se denominará y , con un único punto de censura inferior a , es necesaria la utilización de la variable aleatoria original subyacente (latente) y^* . Entonces, la variable censurada y tomará los valores:

$$y = a_y \text{ cuando la variable subyacente } y^* \leq a$$

$$y = y^* \text{ cuando la variable subyacente } y^* > a$$

Cabe notar la diferencia entre los valores a_y y a . El punto de censura a determina si y^* está censurada, mientras que a_y es el

valor asignado a la variable y si y^* está censurada. Usualmente el valor a_y es igual al valor del punto de censura a , pero podría no serlo. Por simplicidad se supondrán iguales.

Si además se realiza la asunción de que la distribución de la variable subyacente es $y^* \sim N(\mu, \sigma^2)$ la probabilidad de que una observación esté censurada o no lo esté será:

$$\begin{aligned} \Pr(\text{censurada}) &= \Pr(y^* \leq a) = \Pr(N(\mu, \sigma^2) \leq a) = \\ &= \Pr(N(0, 1) \leq \left(\frac{a - \mu}{\sigma}\right)) = \Phi\left(\frac{a - \mu}{\sigma}\right) \\ \Pr(\text{no censurada}) &= \Pr(y^* > a) = \\ &= 1 - \Pr(y^* \leq a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - a}{\sigma}\right) \end{aligned}$$

donde $\Phi(\cdot)$ representa la función de distribución de $a \sim N(0, 1)$ evaluada en el punto en cuestión.

La función de densidad de la variable censurada será entonces:

$$\Pr(y = a) = \Pr(y^* \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right) \text{ cuando } y^* \leq a$$

La misma densidad de y^* cuando $y^* > a$

La misma densidad de y^* cuando $y^* > a$

Esta distribución es una mixtura entre una distribución continua y otra discreta, donde se asigna toda la probabilidad contenida en el área censurada al punto de censura a . Por esta razón, se habla de un punto de acumulación de probabilidad en el punto de censura.

El interés en un modelo Tobit reside habitualmente en estudiar la variable latente y^* . La formulación general del modelo es que el valor medio de esta variable y^* es una función lineal de las variables explicativas $E [y_i^* | x_i] = X_i' \beta$. Dado que los valores de y^* son desconocidos, y tan sólo se conocen los valores de la variable censurada y , se modelizará la $E [y_i | x_i]$ expresándola en función de $E [y_i^* | x_i]$ como:

$$E [y_i | x_i] = E [y_i^* | x_i, y_i^* > a] \cdot \Pr [y_i^* > a | x_i] + a \cdot \Pr [y_i^* \leq a | x_i]$$

La estimación de este modelo utilizando el método de MCO proporciona estimaciones sesgadas de los coeficientes. Sin embargo, las estimaciones por el método de máxima verosimilitud facilitan estimaciones de los coeficientes eficientes y consistentes (3,4), ya que la función de verosimilitud que se maximiza integra información tanto de las observaciones censuradas como de las no censuradas:

$$l(\beta, \sigma^2) = \ln L(\beta, \sigma^2) = \sum_{y_i > a} -\frac{1}{2} \left[\ln(2\pi) + \ln(\sigma^2) + \frac{(y_i - x_i' \beta)^2}{\sigma^2} \right] + \sum_{y_i \leq a} \ln \left[\Phi\left(\frac{a - x_i' \beta}{\sigma}\right) \right]$$

En esta función se observa cómo se podrán identificar las estimaciones de los efectos sobre la variable latente y^* ($\hat{\beta}$) utilizando únicamente la variable censurada y .

En este modelo la no normalidad afecta en mayor medida que en los modelos de regresión lineal habituales y produce que los estimadores $\hat{\beta}$ sean inconsistentes. Los fundamentos teóricos

presentados en el modelo Tobit son generalizables a situaciones en las que la variable dependiente pueda tener varios puntos de censura, ya sean todos inferiores, todos superiores o inferiores y superiores (3,4).

Interpretación de los coeficientes

El interés en un modelo Tobit puede centrarse en la estimación de diferentes medidas de efecto:

- a) Cuando el interés reside en el estudio de las variables x asociadas con la variable latente y^* , las estimaciones $\hat{\beta}$ obtenidas en el modelo Tobit representan directamente el efecto marginal que cada una de las variables x tiene en el valor medio de y^* .
- b) Sin embargo, si el interés reside en el estudio de las variables x asociadas con la variable censurada y , las estimaciones $\hat{\beta}$ obtenidas en el modelo Tobit deberán ponderarse por la probabilidad de que una observación no esté censurada:

$$\hat{\beta} \cdot \Phi \left(\frac{x'_i \hat{\beta} - a}{\hat{\sigma}} \right)$$

Esta probabilidad de no censura depende de los valores que tome cada uno de los sujetos i en cada una de las variables x , por lo que habitualmente se evalúa en la media, mínimo y/o máximo de dichas variables.

Una posible limitación importante del modelo Tobit, al menos en ciertas aplicaciones, es que el valor de la esperanza condicionada a $y > 0$ se relaciona estrechamente con la probabilidad de que $y > 0$.

Análisis estadístico

Para contrastar si existen diferencias estadísticamente significativas entre los sujetos con censura y sin censura, se utiliza el test de la suma de rangos de Wilcoxon para las variables continuas, que pone a prueba si los datos de ambos grupos de sujetos proceden de poblaciones con la misma distribución. Para las variables categóricas se utiliza el estadístico de contraste de la χ^2 de Pearson, el cual pone a prueba si las filas y las columnas en una tabla de contingencia son independientes.

Seguidamente, para cada uno de los dos modelos de regresión analizados se realizan los respectivos modelos de regresión univariantes para cada una de las variables explicativas x_j consideradas. En el primer modelo de regresión analizado, se consideran tan sólo aquellos individuos con valores detectados (la muestra con valores observados) y se estima un modelo de regresión lineal por MCO. En un segundo modelo, se consideran todos los individuos, aunque se asume que todos los sujetos con valores censurados toman el mismo valor mínimo de censura ($a = 0,03 \mu\text{g/l}$).

Se estima un modelo de regresión lineal censurado con un único punto de censura o modelo Tobit.

Para cada uno de los dos análisis se construyen a continuación los modelos multivariados. Se incluyen todas aquellas variables cuyo valor de la t de Student para el coeficiente estimado resulta en valor absoluto mayor que 1 en los correspondientes modelos univariados y, posteriormente, se van eliminando una a una las variables no significativas hasta configurar los modelos finales.

Software estadístico

El análisis estadístico se realiza utilizando el paquete estadístico R Version 1.0.136

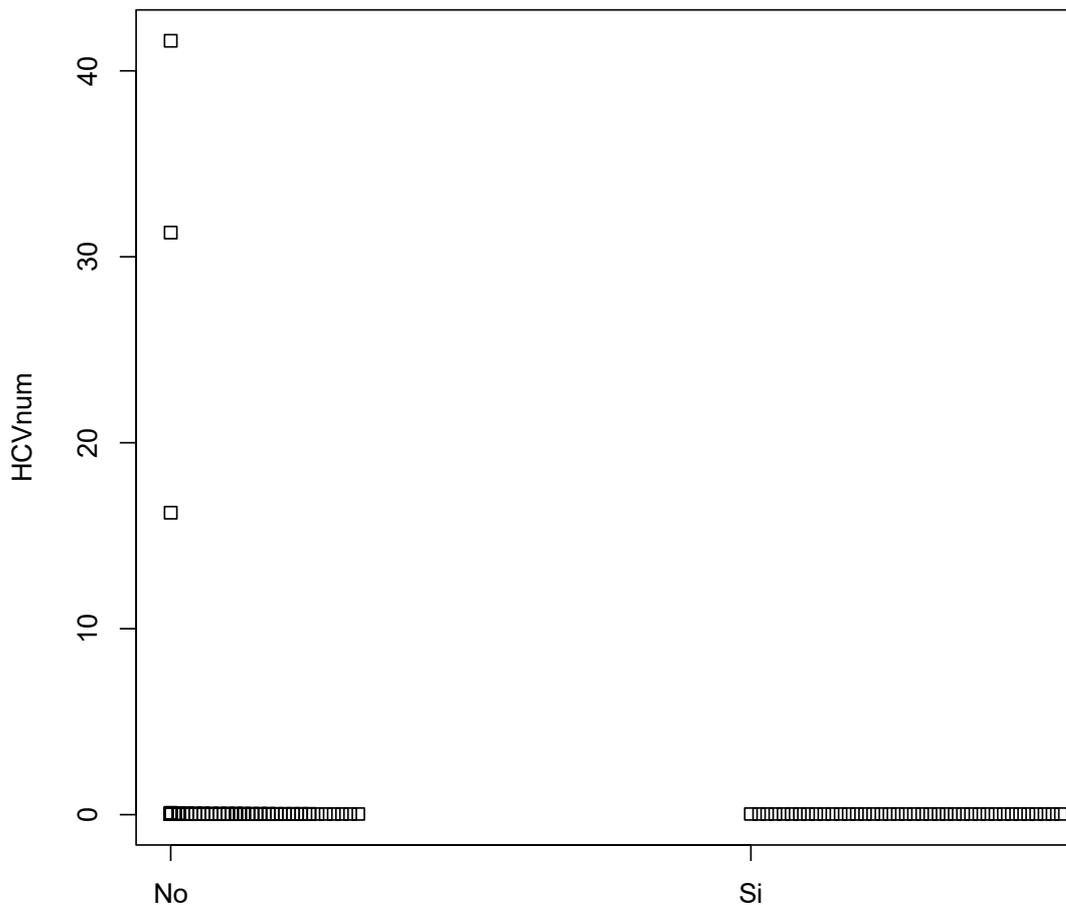
2. Resultados

En 39 de los 145 sujetos (26.90%) no se detectó la concentración de HCV debido al límite de detección inferior del aparato de medición (tabla 1). Para estos sujetos, el valor de censura se correspondió al límite inferior de detección. Además, para normalizar los valores de HCV, éstos fueron transformados logarítmicamente debido a la forma asimétrica de la distribución (fig. 2)

Tabla 1. Descripción de las concentraciones de HCV (en mcg/ml) para los individuos con valores censurados y no censurados

	mean	sd	IQR	0%	25%	50%	75%	100%	data:
No	0.8774528	5.249822	0.00475	0.033	0.034	0.035	0.03875	41.62	100
Si	0.0300000	0.000000	0.00000	0.030	0.030	0.030	0.03000	0.03	39

Fig.2. Distribución de las concentraciones de HCV sin punto de censura y para un único punto de censura.



En el análisis descriptivo para los sujetos censurados y no censurados (tabla 3), la comparación de los valores de las variables incluidas en el análisis no objetivó diferencias estadísticamente significativas a un nivel de significación $\alpha= 0,05$ para sífilis numérica, VIH, sexo y sífilis y sí hubo diferencias estadísticamente significativas para la edad ($p=0.02128$)

Tabla 3. Análisis descriptivo para la muestra censurada y no censurada		
	Censurado (39)	No censurado (106)
Edad, media (sd)	35.80070(6.741)	40.67715(11.071)
Sífilis, media (sd)	3.834564(21.54)	10.776425(48.89)
Sexo, n(%)		
Hombre	21 (53.85)	71 (66.99)
Mujer	18 (46.15)	35 (33.01)
VIH		
Positivo	0	2
Negativo	39	104
Sífilis		
Positivo	2	0
Negativo	37	106

En los modelos de regresión, tanto univariantes como multivariantes, muy pocas variables demostraron estar asociadas con los valores de HCV (tabla 4).

En el modelo de regresión lineal, estimado por MCO, la variable SIFNUM resultó estadísticamente significativas ($p = 0.00377$). En el modelo Tobit, considerando un mismo punto de censura en $0,03 \mu\text{g/dl}$ para los 39 sujetos con valores censurados, la variable SIFNUM continuó siendo estadísticamente significativa ($p = 0,00123$) (tabla 5). El resto de variables en ambos modelos no fueron estadísticamente significativas.

Tabla 4. Resultados del modelo de regresión lineal con la muestra no censurada

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.79115	2.01843	1.383	0.16980
EDAD	-0.04456	0.04580	-0.973	0.33294
Sexo[T.M]	-0.96025	1.06695	-0.900	0.37029
SIFSCR[T.POS]	-3.22473	2.78132	-1.159	0.24904
SIFNUM	0.04468	0.01506	2.967	0.00377
VIH1.2[T.POS]	2.05879	4.44960	0.463	0.64459

Tabla 5. Resultados del modelo Tobit considerando un único punto de censura.

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.448722	1.863027	-0.241	0.80967
EDAD	0.004442	0.043132	0.103	0.91797
Sexo[T.M]	-1.202519	0.947498	-1.269	0.20439
SIFSCR[T.POS]	-3.072530	2.551942	-1.204	0.22859
SIFNUM	0.046794	0.014483	3.231	0.00123
VIH1.2[T.POS]	2.423135	4.300938	0.563	0.57316
Log(scale)	1.610535	0.069263	23.252	< 2e-16

Como el interés residía en investigar los valores de HCV, de la cual se extrajo una muestra representativa, cada estimación $\hat{\beta}$ asociada a una variable x y obtenida en el modelo Tobit representa directamente el efecto marginal que cada una de las variables independientes tiene en el valor medio de la variable subyacente y^* cuando varían en una unidad, manteniendo constantes el resto de variables. Si el interés hubiese residido en la variable censurada y , la interpretación de los coeficientes en estos dos modelos no hubiese sido directa y se habría tenido que calcular el efecto

marginal de las variables incluidas en el modelo, corrigiendo por la probabilidad de no censura.

Comparando los resultados obtenidos en el primer modelo (regresión lineal) con las estimaciones $\hat{\beta}$ obtenidas a través del modelo Tobit, se observa cómo las estimaciones de los dos tipos de modelos, exceptuando la edad, van en la misma dirección (tienen el mismo signo), aunque difieren bastante en su magnitud. Las estimaciones obtenidas en el modelo Tobit son en general, sustancialmente mayores (en valor absoluto).

La estimación de los errores estándar asociados a estos coeficientes, fueron mayores en el modelo de regresión lineal que en el modelo Tobit.

3. Discusión.

Muchas variables epidemiológicas que no miden el tiempo transcurrido desde un momento dado hasta que se produce el evento de interés presentan también distribuciones con censura para las cuales los modelos de regresión lineal no deberían utilizarse, porque proporcionan estimaciones sesgadas e inconsistentes. En esta situación es aconsejable la utilización de modelos más adecuados a la naturaleza de la variable de estudio que tengan en cuenta la existencia de censura.

La familia de modelos de regresión censurada permite tratar este problema, ya sea con un único o con varios puntos de censura, y con censura inferior, superior o de intervalo. En comparación con los resultados que facilita el modelo de regresión lineal, los que se obtienen utilizando los modelos de regresión censurados no

cambian la dirección del efecto estimado. Las principales diferencias se encuentran al cuantificar la estimación de los efectos, tal como se ilustra en el análisis de los valores de HCV, donde los coeficientes estimados pueden variar en gran medida, así como en la estimación de los errores estándar de dichas estimaciones que intervienen en la significación estadística de estos estimadores. Esto debe ser tenido en cuenta, ya que en la mayoría de los estudios epidemiológicos la cuantificación del efecto es de tanto interés como su significación.

Otro punto a destacar, que pone de manifiesto la importancia de tener en cuenta la censura, es el hecho de que ignorar todas las observaciones censuradas y trabajar exclusivamente con observaciones detectadas hace que la variable que se desea estudiar a escala poblacional tenga una distribución diferente de la variable resultante al obtener la muestra. En particular, el valor medio calculado con la muestra resulta mayor que el valor medio poblacional, si los valores no detectados se sitúan en la cola inferior de la distribución, y resulta menor si los valores no detectados se sitúan en la cola superior de la distribución.

Así, si fuese posible realizar un modelo de regresión lineal conocida la variable latente y^* en la población, se obtendrían los valores reales en las estimaciones $\hat{\beta}$. Pero el efecto de eliminar las observaciones censuradas y de estimar un modelo de regresión lineal es que las estimaciones $\hat{\beta}$ MCO que se obtienen a través del modelo de regresión lineal, estimado por MCO, serán menores (en valor absoluto) que las anteriores y menos precisas. El efecto de introducir las observaciones censuradas y de estimar un modelo que tiene en cuenta la censura es que las estimaciones obtenidas

son generalmente mayores (en valor absoluto) que las estimaciones $\hat{\beta}^{\text{MCO}}$ y más precisas, por lo que serán más próximas a las verdaderas β^* . Por ser modelos muy sensibles a la falta de normalidad, es muy importante tener en cuenta este aspecto antes de realizar cualquier análisis. La ausencia de normalidad de los errores del modelo ocasiona que los estimadores obtenidos sean inconsistentes.

En un futuro, sería interesante ampliar la N de la base de datos y utilizar nuevas variables explicativas para dar más consistencia al trabajo, así como utilizar otros modelos de censura diferentes, como podrían ser el modelo con varios puntos de censura y el modelo en dos partes.

4. Bibliografía

1. Soriano JB, Antó JM, Sunyer J, Tobías A, Kogevinas M, Almar E, et al. Risk of asthma in the general Spanish population attributable to specific immunoresponse. *Int J Epidemiol* 1999;28:728-34.
2. Bleda MJ, González CA, Kogevinas M, Huici A, Gadea E, Ladona M, et al. Niveles séricos basales de dioxinas, furanos, PCB's y metales en una muestra de población general en una ciudad española en que se ha instalado una incineradora de residuos sólidos [abstract]. *Gac Sanit* 1996;10(Supl):56.
3. Greene WH. *Análisis econométrico*. 3.a ed. Prentice Hall Iberia, Madrid, 1999.
4. Long JS. *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage, 1997.
5. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958;26:24-36.
6. Van der Gaag J, Van der Ven W. The demand for primary health care. *Med Care* 1978;16:299-312.
7. Luoma K, Jarvio ML, Suoniemi I, Hjerppe RT. Financial incentives and productive efficiency in Finnish health centres. *Health Econ* 1996;5:435-45.
8. Grootendorst PV. Health care policy evaluation using longitudinal insurance claim data: a Tobit estimator. *Health Econ* 1997;6:365-82.
9. Rosko MD. Impact of internal and external environmental pressures on hospital. *Health Care Manag Sci* 1999;2:63-74.
10. Rosko MD. Impact of internal and external environmental pressures on hospital. *Health Care Manag Sci* 1999;2:63-74.
11. Sáez M, Barceló MA. Un criterio para omitir variables superfluas en modelos de regresión. *Gac Sanit* 1998;12:281-3.

12. ALTER MJ. Epidemiology of hepatitis C virus infection, *World J Gastroenterol* 2007; 13(17): 2436-2441.
13. GURTSEVITCH VE. Human Oncogenic Viruses: Hepatitis B and Hepatitis C Viruses and Their Role in Hepatocarcinogenesis, *Biochemistry (Moscow)*, 2008, 73(5): 504-513.
14. SIMMONDS P. et al. Consensus Proposals for a Unified System of Nomenclature of Hepatitis C Virus Genotypes, *HEPATOLOGY* 2005;42: 962-973.
15. VERMEHREN J, SARRAZIN C. New HCV therapies on the horizon, *Clin Microbiol Infect* 2011; 17: 122–134. 6.
16. WEBSTER D.P, KLENERMAN P., COLLIER J., JEFFERY K. JM. Development of novel treatments for hepatitis C, *Lancet Infect Dis* 2009; 9: 108-17.
17. CHEVALIEZ S. Virological tools to diagnose and monitor hepatitis C virus infection, *Clin Microbiol Infect* 2011; 17: 116–121.
18. PENIN F, DUBUISSON, REY F-A, MORADPOUR D, ET PAWLITSKY JM. Structural Biology of Hepatitis C Virus, *HEPATOLOGY* 2004; 39: 5-19.

5. Anexos

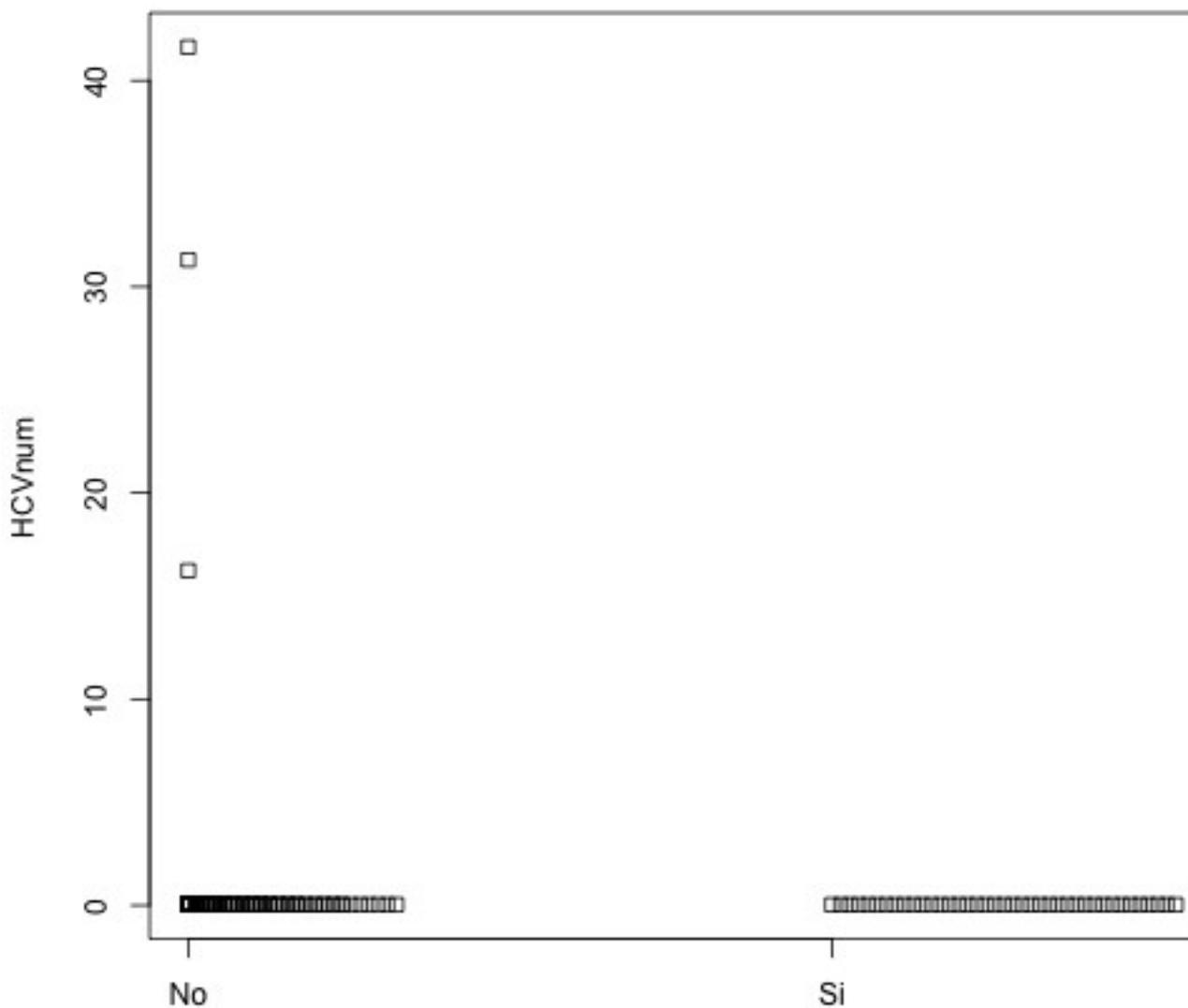
Códigos usados en R:

```
Dataset <-
  read.table("/Users/pcm/Google Drive/master bioinformat-
ica/TFM/definitivos/Dataset.txt",
  header=TRUE, sep=";", na.strings="NA", dec=".",
strip.white=TRUE)
CENSURADOS <-
  read.table("/Users/pcm/Google Drive/master bioinformat-
ica/TFM/definitivos/CENSURADOS.txt",
  header=TRUE, sep=";", na.strings="NA", dec=".",
strip.white=TRUE)
NOCENSURADOS <-
  read.table("/Users/pcm/Google Drive/master bioinformat-
ica/TFM/definitivos/NOCENSURADOS.txt",
  header=TRUE, sep=";", na.strings="NA", dec=".",
strip.white=TRUE)
summary(Dataset)
## Sexo          EDAD          VIH1.2          HCVnum          SIF-
SCR
## H:92   Min.      :19.68   NEG:143   Min.      : 0.0300
NEG:133
## M:53   1st Qu.:33.22   POS: 2    1st Qu.: 0.0300   POS:
12
##           Median :37.73           Median : 0.0340
##           Mean   :39.37           Mean   : 0.6495
##           3rd Qu.:43.72           3rd Qu.: 0.0370
##           Max.   :72.41           Max.   :41.6200
##           SIFNUM          censurados
## Min.      : 0.062   No:106
## 1st Qu.: 0.076   Si: 39
## Median : 0.083
## Mean   : 8.909
## 3rd Qu.: 0.094
## Max.   :300.400
summary(CENSURADOS)
## Sexo          EDAD          VIH1.2          HCVnum          SIFSCR
SIFNUM
## H:21   Min.      :19.68   NEG:39   Min.      :0.03   NEG:37
Min.      : 0.065
## M:18   1st Qu.:32.48           1st Qu.:0.03   POS: 2
1st Qu.: 0.076
##           Median :36.31           Median :0.03
Median : 0.082
##           Mean   :35.80           Mean   :0.03
Mean   : 3.835
```

```

##          3rd Qu.:38.71          3rd Qu.:0.03
3rd Qu.: 0.087
##          Max.    :51.40          Max.    :0.03
Max.    :134.400
## censurados
## Si:39
##
##
##
##
##
summary(NOCENSURADOS)
## Sexo          EDAD          VIH1.2          HCVnum
SIFSCR
## H:71  Min.    :20.67  NEG:104  Min.    : 0.03300
NEG:96
## M:35  1st Qu.:33.70  POS: 2   1st Qu.: 0.03400
POS:10
##          Median :38.70          Median : 0.03500
##          Mean   :40.68          Mean   : 0.87745
##          3rd Qu.:45.15          3rd Qu.: 0.03875
##          Max.   :72.41          Max.   :41.62000
##          SIFNUM          censurados
## Min.    : 0.0620  No:106
## 1st Qu.: 0.0760
## Median : 0.0835
## Mean   : 10.7764
## 3rd Qu.: 0.1020
## Max.   :300.4000
numSummary(Dataset[, "HCVnum"], groups=Dataset$censurados,
  statistics=c("mean", "sd", "IQR", "quantiles"), quan-
  tiles=c(0, .25, .5, .75, 1))
##          mean          sd          IQR          0%          25%          50%          75%
100% data:n
## No 0.8774528 5.249822 0.00475 0.033 0.034 0.035 0.03875
41.62 106
## Si 0.0300000 0.000000 0.00000 0.030 0.030 0.030 0.03000
0.03 39
stripchart(HCVnum ~ censurados, vertical=TRUE,
method="stack",
  ylab="HCVnum", data=Dataset)

```



```

with(Dataset, tapply(EDAD, censurados, median, na.rm=TRUE))
##      No      Si
## 38.69589 36.30685
wilcox.test(EDAD ~ censurados, alternative="two.sided",
data=Dataset)
##
## Wilcoxon rank sum test with continuity correction
##
## data:  EDAD by censurados
## W = 2584, p-value = 0.02128
## alternative hypothesis: true location shift is not equal
to 0
with(Dataset, tapply(SIFNUM, censurados, median,
na.rm=TRUE))

```

```

##      No      Si
## 0.0835 0.0820
wilcox.test(SIFNUM ~ censurados, alternative="two.sided",
data=Dataset)
##
## Wilcoxon rank sum test with continuity correction
##
## data:  SIFNUM by censurados
## W = 2270, p-value = 0.3664
## alternative hypothesis: true location shift is not equal
to 0
local({
  .Table <- xtabs(~censurados+Sexo, data=Dataset)
  cat("\nFrequency table:\n")
  print(.Table)
  .Test <- chisq.test(.Table, correct=FALSE)
  print(.Test)
})
##
## Frequency table:
##      Sexo
## censurados  H  M
##           No 71 35
##           Si 21 18
##
## Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 2.121, df = 1, p-value = 0.1453
local({
  .Table <- xtabs(~censurados+SIFSCR, data=Dataset)
  cat("\nFrequency table:\n")
  print(.Table)
  .Test <- chisq.test(.Table, correct=FALSE)
  print(.Test)
})
##
## Frequency table:
##      SIFSCR
## censurados NEG POS
##           No  96  10
##           Si  37   2
## Warning in chisq.test(.Table, correct = FALSE): Chi-
squared approximation
## may be incorrect
##
## Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 0.69631, df = 1, p-value = 0.404
local({

```

```

.Table <- xtabs(~censurados+VIH1.2, data=Dataset)
cat("\nFrequency table:\n")
print(.Table)
.Test <- chisq.test(.Table, correct=FALSE)
print(.Test)
})
##
## Frequency table:
##      VIH1.2
## censurados NEG POS
##      No 104  2
##      Si  39  0
## Warning in chisq.test(.Table, correct = FALSE): Chi-
squared approximation
## may be incorrect
##
## Pearson's Chi-squared test
##
## data:  .Table
## X-squared = 0.74614, df = 1, p-value = 0.3877
LinearModel.1 <- lm(HCVnum ~ EDAD + Sexo + SIFSCR + SIFNUM
+ VIH1.2,
  data=NOCENSURADOS)
summary(LinearModel.1)
##
## Call:
## lm(formula = HCVnum ~ EDAD + Sexo + SIFSCR + SIFNUM +
VIH1.2,
##      data = NOCENSURADOS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.276  -1.078  -0.707  -0.025  40.224
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.79115    2.01843   1.383  0.16980
## EDAD          -0.04456    0.04580  -0.973  0.33294
## Sexo[T.M]     -0.96025    1.06695  -0.900  0.37029
## SIFSCR[T.POS] -3.22473    2.78132  -1.159  0.24904
## SIFNUM         0.04468    0.01506   2.967  0.00377 **
## VIH1.2[T.POS]  2.05879    4.44960   0.463  0.64459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## Residual standard error: 5.04 on 100 degrees of freedom
## Multiple R-squared:  0.1222, Adjusted R-squared:
0.07826
## F-statistic: 2.783 on 5 and 100 DF,  p-value: 0.0214
t <- tobit(HCVnum ~ EDAD + Sexo + SIFSCR + SIFNUM + VIH1.2,

```

```
left = 0.03, data = NOCENSURADOS)
```

```
summary(t)
```