

*Estimadores del Total Poblacional en Muestras  
con Observaciones Censuradas*

CARLOS ALBERTO HERNÁNDEZ LOZANO  
LICENCIADO EN MATEMÁTICAS



UNIVERSIDAD NACIONAL DE COLOMBIA  
FACULTAD DE CIENCIAS  
DEPARTAMENTO DE ESTADÍSTICA  
BOGOTÁ, D.C.



*Estimadores del Total Poblacional en Muestras  
con Observaciones Censuradas*

CARLOS ALBERTO HERNÁNDEZ LOZANO  
LICENCIADO EN MATEMÁTICAS

TRABAJO PRESENTADO PARA OPTAR AL TÍTULO DE  
MAGISTER EN CIENCIAS-ESTADÍSTICA

DIRECTOR  
LEONARDO TRUJILLO OYOLA, PH.D.  
PROFESOR UNIVERSIDAD NACIONAL DE COLOMBIA.

CODIRECTOR  
GUILLERMO MARTÍNEZ FLÓREZ, PH.D.  
PROFESOR UNIVERSIDAD DE CÓRDOBA.



UNIVERSIDAD NACIONAL DE COLOMBIA  
FACULTAD DE CIENCIAS  
DEPARTAMENTO DE ESTADÍSTICA  
BOGOTÁ, D.C.



**Título en español**

Estimadores del Total Poblacional en Muestras con Observaciones Censuradas.

**Title in English**

Population Total Estimators in Censored Data Samples

**Resumen:** La presente tesis plantea el diseño de varios estimadores que permitan hacer inferencias sobre el total poblacional de una variable de interés en muestras que contienen observaciones censuradas. Los estimadores propuestos consideran la censura (entendida como una restricción en los valores de la variable de interés) como una característica propia de los datos que es incorporada en el proceso de estimación.

**Abstract:** This thesis proposes some estimators allowing to make inference about the population total of a variable in samples with censored observations. The proposed estimators take into account the censoring (understood like a restriction over the values of the variable of interest) like a data characteristic being incorporated in the process of estimation.

**Palabras clave:** Estimador asistido por modelos, estimador de regresión, modelo tobit, datos censurados

**Keywords:** Model-assisted estimator, regression estimator, Tobit's model, censored data



# Nota de aceptación

---

Jurado  
Luz Mery González García

---

Jurado  
Luis Hernando Vanegas Penagos

---

Director  
Leonardo Trujillo Oyola

---

Codirector  
Guillermo Martínez Flórez

Bogotá, D.C., 24 de Abril de 2017



Dedicado A

A Arturo, todo mi amor. Gracias por estar ahí junto a mí.  
A mi Madre. Siempre incondicional, creyente y fuerte.

DEDICADO A

---

## Agradecimientos

Agradezco a mi profesor Leonardo Trujillo, por su dedicado trabajo con este proyecto, ha sabido ser paciente durante estos años. Al profesor Guillermo Martínez por haber aceptado vincularse a nuestro trabajo y aportar valiosamente al mismo. A Arturo por sus especiales consejos y guía matemática, sin él hubiera perdido el rumbo y la esperanza en muchos momentos.



# Índice General

Índice General	III
Índice de Tablas	VII
Índice de Figuras	IX
INTRODUCCIÓN	XI
<b>1. PRELIMINARES TEÓRICOS</b>	<b>1</b>
1.1. Notación e Ideas Básicas . . . . .	1
1.2. Estadístico y Estimador . . . . .	3
1.3. Estimadores para el Total Poblacional . . . . .	5
1.3.1. Estimador de Horvitz-Thompson (H-T) o $\pi$ -estimador . . . . .	5
1.3.2. Estimadores de Regresión . . . . .	7
1.3.3. Estimadores Asistidos por Modelos Lineales Generalizados (MLG). . . . .	10
1.4. Censura . . . . .	12
1.4.1. Distribución Normal Censurada . . . . .	15
1.4.2. Modelo de Regresión Censurado o Modelo Tobit . . . . .	16
1.4.2.1. Estimación de los Parámetros del Modelo . . . . .	18
1.4.2.2. Estimación de los Parámetros en el Modelo Tobit con Pesos Muestrales . . . . .	20
1.4.2.3. Estimación de la Matriz de Varianzas y Covarianzas . . . . .	22

<b>2. DISEÑO DE ESTIMADORES PARA EL TOTAL POBLACIONAL</b>	<b>25</b>
2.1. Problema de Estimación . . . . .	25
2.2. Estimadores para el Total . . . . .	27
2.2.1. El Total de la Variable Latente y el Total de la Variable Censurada . . . . .	29
2.2.2. Estimadores para el Total de la Pseudovariable Latente . . . . .	30
2.2.2.1. Estimador Tobit Sintético . . . . .	30
2.2.2.2. $\pi$ -Estimador Tobit Sintético . . . . .	31
2.2.3. Estimadores para el Total de la Variable Censurada y la Pseudovariable Mixta . . . . .	32
2.2.3.1. $\pi$ -Estimador o Estimador de Horvitz-Thompson (H-T) . . . . .	32
2.2.3.2. Estimador de Regresión o Estimador GREG . . . . .	33
2.2.3.3. Estimador de Regresión Tobit . . . . .	34
2.2.3.4. Estimador Particionado . . . . .	34
2.3. Simulación . . . . .	36
2.3.1. Diseño de la Simulación General . . . . .	36
2.3.2. Resultados de la Simulación . . . . .	38
2.4. Estimadores para el Total Poblacional. . . . .	42
2.4.1. Esperanza y Varianza para el Estimador Tobit Sintético . . . . .	43
2.4.2. Esperanza y Varianza para el $\pi$ -Estimador Tobit Sintético . . . . .	44
2.4.3. Segunda Simulación . . . . .	46
2.4.4. Análisis de la Censura en el Comportamiento del Estimador . . . . .	49
2.5. Conclusiones . . . . .	52
<b>3. EJEMPLO DE APLICACIÓN</b>	<b>55</b>
3.1. Ejemplo de Aplicación. "Población CO124" . . . . .	55
3.1.1. Procedimiento . . . . .	55
3.2. Ejemplo de Aplicación 2. . . . .	60
3.2.1. Estimadores Bajo Muestreo ESTMAS . . . . .	60
3.2.2. Procedimiento . . . . .	61
<b>4. CONCLUSIONES Y TRABAJO FUTURO</b>	<b>65</b>
4.1. Conclusiones . . . . .	65
4.2. Trabajo Futuro . . . . .	66

---

<b>A. Distribución de los Estimadores Simulados</b>	<b>67</b>
<b>B. Conjuntos De Datos</b>	<b>73</b>
B.1. ‘Población CO124’ .....	73
<b>C. Código de Simulación</b>	<b>77</b>
C.1. Información Técnica .....	77
C.2. Código .....	80
C.2.1. Función que Permite Calcular el Modelo Tobit .....	80
C.2.2. Función que Permite Ejecutar la Simulación de Montecarlo (2)	85
<b>Bibliografía</b>	<b>95</b>



# Índice de Tablas

1.1. Estimación de $\beta$ y $\mu$ . . . . .	11
1.2. Formas de Clasificación para Muestras con Observaciones Censuradas	14
2.1. Ejemplo de Estimación con Datos Censurados y no Censurados. . . . .	26
2.2. Algunas Medidas Estadísticas de los Datos Poblacionales Simulados. .	36
2.3. Resultados de la simulación. Esperanza, Sesgo, Sesgo relativo y ECM para $t_{y^*}$ . . . . .	39
2.4. Resultados de la simulación. Esperanza, Sesgo, Sesgo relativo y ECM para $t_y$ . . . . .	41
2.5. Medidas resumen para los estimadores $\hat{t}_{sin}$ y $\hat{t}_{\pi TS}$ . $M = 50000$ , $n =$ (200, 400, 600, 800). . . . .	48
3.1. Estimaciones del valor de la variable P83 mediante los estimadores $\hat{t}_{sin}$ , $\hat{t}_{\pi TS}$ , $\hat{t}_{\pi y}$ y $\hat{t}_{yr}$ . Estimación de la varianza y del intervalo de con- fianza. . . . .	60
B.1. Datos para la Población CO124 . . . . .	76



## Índice de Figuras

2.1. Histograma con curva de densidad kernel (a) y (b). . . . .	37
2.2. Gráfico de dispersión de la variable latente con línea de regresión clásica (a) y la variable censurada (b) con línea de regresión censurada. . . . .	37
2.3. Agrupamiento de Muestras por Porcentaje de Censura y Gráfico de Boxplot. Estimador (1). . . . .	50
2.4. Agrupamiento de Muestras por Porcentaje de Censura y Gráfico de Boxplot. Estimador (3). . . . .	51
A.1. Histograma de frecuencias para los Estimadores (1) y (2) de la Tabla 2.3. . . . .	68
A.2. Histograma de frecuencias para los Estimadores (3) y (4) de la Tabla 2.3. . . . .	69
A.3. Histograma de frecuencias para los Estimadores (5) y (6) de la Tabla 2.3. . . . .	70
A.4. Histograma de frecuencias para los Estimadores (7) y (8) de la Tabla 2.3. . . . .	71
A.5. Histograma de frecuencias para los Estimadores (9) y (10) de la Tabla 2.3. . . . .	72



# INTRODUCCIÓN

El estudio de la censura y el truncamiento tiene una larga historia en la estadística. Cohen (1991, p. 2) comenta que el estudio de muestras truncadas y censuradas inicia con los trabajos de sir Francis Galton en el año de 1887 sobre registros de velocidad de caballos americanos de carreras. Sobre estos datos asume normalidad y proporciona un estimador de la media. Karl Pearson hacia el año de 1902 ajusta parábolas al logaritmo de las frecuencias muestrales sin que esto represente una mejora del trabajo de Galton. Hacia 1908, Karl Pearson y Alice Lee aplican el método de momentos para estimar la media y la desviación estándar de la distribución normal de una muestra simplemente truncada a izquierda. R. A. Fischer en 1931, emplea el método de máxima verosimilitud basado en muestras simplemente truncadas a izquierda para los parámetros de una distribución normal. En 1937 Stevens deriva ecuaciones de máxima verosimilitud para muestras simple y doblemente censuradas del Tipo I donde la cantidad de datos censurados es conocido. Cohen en 1940, presenta su tesis doctoral en la cual deriva estimadores de momentos para las distribuciones de Pearson. Tobin (1958) propone el primer modelo de regresión censurado, empleando el método de máxima verosimilitud para la estimación de los parámetros. Unos años después Amemiya (1973) y Olsen (1978), uno tras otro, demuestran propiedades asintóticas sobre los estimadores derivados por Tobin. A partir de este momento numerosos autores han continuado desarrollando teoría relacionada con las muestras truncadas y censuradas. Amemiya (1984) presenta un resumen de los trabajos realizados hasta ese momento y Cohen (1991) presenta su libro sobre muestras truncadas y censuradas.

A pesar del amplio estudio existente alrededor de la censura y el truncamiento, la mayor parte de la literatura en muestreo concierne a muestras irrestrictas o completamente observadas. En la realidad el panorama es muy distinto. Es probable que los investigadores, en la práctica, se encuentren con muestras que son censuradas o truncadas. En algunos casos, este fenómeno puede ser de menor impacto en los análisis y se justifica que pueda ser ignorado. Esto no siempre es posible y por tanto representa un aspecto a tener en cuenta al momento de trabajar con este tipo de

muestras. Ejemplos de encuestas donde se presenta censura se tienen en los datos de las *encuestas actuales de población*<sup>1</sup> del Bureau de Censos de los Estados Unidos.

Precisamente ese es el contexto donde se considera la elaboración del presente trabajo de tesis. Situaciones de muestreo en donde no es posible ignorar el fenómeno de la censura. Es por eso que esta tesis plantea la construcción de un estimador que permita hacer inferencia sobre el total poblacional de una variable de interés en muestras que contengan observaciones censuradas. Esto es, diseñar un estimador que considere la censura (entendida como una restricción en los valores de la variable de interés) como una característica propia de los datos y la incorpore en el proceso de estimación del total poblacional.

Para la consecución de este objetivo suponemos que la aproximación que puede servir para solucionar el problema de estimación se encuentra en el uso de estimadores asistidos por modelos de regresión o que involucran modelos de regresión. En particular se supone que el uso del modelo de regresión Tobit permitirá el diseño de un estimador del total apto para el trabajo en muestras con observaciones censuradas. Para probar la idea anterior, se dividió el trabajo de investigación en dos grandes partes. La primera consiste en el diseño, con base en los elementos teóricos, de posibles estimadores que puedan servir para abordar el problema propuesto. La segunda consiste en el diseño y aplicación de un experimento simulado por computador, donde se ponga a prueba los estimadores diseñados y se evalúe la solución óptima al problema así como su aplicación en un contexto real.

El presente documento se construyó para dar cuenta del trabajo mencionado anteriormente. Éste se encuentra organizado en tres capítulos. En el capítulo uno se abordan las ideas y conceptos teóricos que dan sustento a la propuesta. El capítulo se desarrolla en cuatro secciones iniciando por la presentación general de la notación que será usada a lo largo del trabajo así como de algunas ideas básicas sobre muestreo probabilístico. La segunda sección se dedica a precisar el concepto de estadístico y estimador. La sección tres se centra en presentar tres estimadores comunes en la teoría de muestreo: el  $\pi$ -estimador, el estimador de regresión y los estimadores asistidos por modelos lineales generalizados. La sección cuatro trata exclusivamente ideas, definiciones y resultados existentes en la literatura sobre datos censurados. Esta sección hace énfasis en el modelo Tobit como el modelo principal para trabajar con datos censurados.

El capítulo dos aborda el problema de estimación en muestras con datos censurados y presenta la solución al mismo. El capítulo se desarrolla en cinco secciones. La primera sección presenta en detalle el problema de estimación en muestras censuradas proponiendo ejemplos y determinando la pregunta de investigación. En la segunda sección se proponen algunos estimadores del total especialmente diseñados para dar solución al problema de estimación. La sección tres presenta el experimento que fue diseñado para estudiar y describir el comportamiento de los estimadores propuestos. El experimento hace uso de simulaciones de Monte Carlo de los estimadores propuestos sobre miles de muestras extraídas de un conjunto de datos especialmente

---

<sup>1</sup>En Inglés se conocen como *Current Population Survey of Income and Program Participation (SIPP)*, ver (Short et al., 1991) .

construido para el experimento. En la sección cuatro, se usan los resultados obtenidos en la simulación y se hace la selección de los estimadores que dan solución al problema de estimación analizando en detalle sus propiedades, especialmente el sesgo y la varianza. La sección final está dedicada a presentar las conclusiones de todos los análisis realizados a lo largo del capítulo.

El capítulo tres presenta un ejemplo en el cual es posible hacer la aplicación de los resultados obtenidos en el capítulo dos. En cada caso se hace un desarrollo detallado del proceso de aplicación de los estimadores, mostrando las rutinas implementadas para los cálculos y los resultados de la aplicación.

Las conclusiones generales del trabajo, así como posibles rutas de trabajo futuro se presentan en el último capítulo. En los apéndices siguientes se presentan algunos resultados anexos de la simulación así como la rutina de programación que fue diseñada para la realización de los cálculos presentados en el capítulo dos y los conjuntos de datos utilizados para la aplicación.



# Capítulo 1

## PRELIMINARES TEÓRICOS

En este capítulo se exponen algunos elementos conceptuales y teóricos básicos para el desarrollo del presente trabajo de tesis. Inicialmente se exponen elementos teóricos de los estimadores para totales poblacionales, seguido de algunos resultados de la teoría existente sobre las observaciones censuradas o por debajo del límite de detección.

En la sección sobre los estimadores del total poblacional, se abordan dos grandes conceptos, el de  $\pi$ -estimador y el de estimador asistido o estimador de regresión. En cada caso, se presentan resultados clásicos y posiciones recientes con respecto a estas temáticas. Se siguen principalmente las ideas de Särndal, Swensson & Wretman (1992) y las de Rondón, Ferraz & Vanegas (2012). Se toman también algunas referencias de los textos de Lohr (1999) y Groves, Fowler-Jr., Couper, Lepkowski, Singer & Tourangeau (2004).

Para la censura se hace una breve exposición de su significado, de los tipos de censura que se pueden presentar y de los modelos de regresión existentes para este tipo de datos. Para esto se sigue lo expuesto por Tobin (1958); Amemiya (1973); Amemiya (1984); Olsen (1978); Millard & Neerchal (2000); Greene (2007); Bolfarine, Santos, Correia, Martínez, Gómez, Bazan & ABE-Associação Brasileira de Estatística (2013).

### 1.1. Notación e Ideas Básicas

Sea  $\mathbb{U}$  la población objetivo o universo sobre la que se quiere realizar algún proceso de muestreo. Se puede definir la población objetivo como el conjunto

$$\mathbb{U} = \{1, \dots, k, \dots, N\}, \quad (1.1.1)$$

donde  $N$  es el tamaño de la población. Sobre  $\mathbb{U}$  se puede definir una variable  $y$  denominada *variable de interés o de estudio*. Con  $y_k$  se representa el valor de  $y$  para el  $k$ -ésimo elemento de  $\mathbb{U}$ .

El interés inicial es poder calcular funciones de los valores de  $y$  sobre toda la población. Tales valores se denominan *parámetros poblacionales* y se representan de forma general con  $\theta$ . Ejemplos de estos parámetros son el *total* o la *media* de  $y$  sobre  $\mathbb{U}$  definidos como

$$t = \sum_{\mathbb{U}} y_k = \sum_{k \in \mathbb{U}} y_k, \quad \text{parámetro para el total de } y \quad (1.1.2)$$

$$\bar{y}_{\mathbb{U}} = \frac{t}{N} = \frac{\sum_{\mathbb{U}} y_k}{N}. \quad \text{parámetro para la media de } y \quad (1.1.3)$$

De  $\mathbb{U}$  se extrae una muestra aleatoria  $s$  de tamaño  $n$  seleccionada de acuerdo con algún esquema probabilístico de selección. Bajo este esquema de selección se puede definir la función  $p(\cdot)$  tal que  $p(s)$  es la probabilidad de seleccionar  $s$ . La función  $p(\cdot)$  se denomina *diseño muestral*. Sea  $\mathcal{S}$  el conjunto de todas las posibles muestras  $s$  y sea  $S$  una variable aleatoria donde  $s$  es una realización de dicha variable. Se tiene entonces que  $p(\cdot)$  es la función de probabilidad de  $S$ .

El evento aleatorio de que un elemento  $k$  de la población sea incluido en la muestra  $s$  es expresado por medio de una *variable indicadora de pertenencia muestral*  $I_k$ , que es una variable aleatoria definida como

$$I_k = \begin{cases} 1, & \text{si } k \in S \\ 0, & \text{en otro caso.} \end{cases} \quad (1.1.4)$$

La probabilidad  $\pi_k$  de que un elemento  $k$  sea incluido en una muestra o *probabilidad de inclusión de primer orden* es determinada por el diseño muestral  $\mathbf{p}(\cdot)$  como sigue

$$\pi_k = \Pr(k \in S) = \Pr(I_k = 1) = \sum_{k=1}^N I_k \mathbf{p}(s) = \sum_{s \ni k} \mathbf{p}(s). \quad (1.1.5)$$

La probabilidad de inclusión de dos elementos  $k$  y  $l$  con  $k \neq l$  se llama *probabilidad de inclusión de segundo orden* o  $\pi_{kl}$  y se define como

$$\pi_{kl} = \Pr(k, l \in S) = \Pr(I_k I_l = 1) = \sum_{\substack{k=1 \\ l=1}}^N I_k I_l \mathbf{p}(s) = \sum_{s \ni k, l} \mathbf{p}(s). \quad (1.1.6)$$

Para la variable indicadora de pertenencia dada por (1.1.4) se tiene las siguientes propiedades

**1.1 Proposición** (Esperanza, Varianza y Covarianza). *Para la variable  $I_k$  son verdaderas las siguientes propiedades*

$$\mathbf{E}[I_k] = \pi_k \quad (1.1.7)$$

$$\mathbf{Cov}(I_k, I_k) = \mathbf{Var}(I_k) = \pi_k(1 - \pi_k) \quad (1.1.8)$$

$$\mathbf{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l \quad (1.1.9)$$

Estas propiedades son fácilmente demostrables si se tiene en cuenta el hecho de que la variable  $I_k$  es una variable aleatoria *Bernoulli*. Acorde con la notación dada por

Särndal et al. (1992, p. 36) se tiene que  $\mathbf{Cov}(I_k, I_k) = \mathbf{Var}(I_k) = \Delta_{kk}$  y  $\mathbf{Cov}(I_k, I_l) = \Delta_{kl}$ .

En lo que sigue se exponen los conceptos relacionados con los estimadores en muestreo<sup>1</sup>.

## 1.2. Estadístico y Estimador

**1.1 Definición** (Estadístico). Sea la muestra  $s$  una realización de la variable aleatoria  $S$  bajo un diseño muestral  $p(\cdot)$ , cualquier función  $Q = Q(S)$  de la forma

$$\begin{aligned} Q : \mathcal{S} &\mapsto \mathbb{R} \\ s &\mapsto Q(s) \end{aligned}$$

es un *estadístico*.

En la definición anterior se establece como condición, para que una función puede ser llamada *estadístico*, que se pueda calcular el valor de tal función para la muestra una vez sea seleccionada y los valores de los distintos individuos de la muestra hallan sido recolectados. En lo que sigue, la escritura  $Q(S)$  se usará para el estadístico y  $Q(s)$  para la realización del estadístico.

La naturaleza aleatoria de  $Q(S)$  se deriva del hecho de que  $S$  es una variable aleatoria. Por tanto  $Q(S)$  es también una variable aleatoria sobre la cual se puede definir las siguientes características.

**1.2 Definición** (Esperanza y Varianza). La *esperanza* y la *varianza* de un estadístico  $Q = Q(S)$  son definidas respectivamente por:

$$\mathbf{E}[Q] = \sum_{s \in \mathcal{S}} \mathbf{p}(s)Q(s) \quad (\text{Esperanza}), \quad (1.2.1)$$

$$\begin{aligned} \mathbf{Var}[Q] &= \mathbf{E}\{[Q - \mathbf{E}(Q)]^2\} \\ &= \sum_{s \in \mathcal{S}} \mathbf{p}(s)[Q(s) - \mathbf{E}(Q)]^2 \quad (\text{Varianza}). \end{aligned} \quad (1.2.2)$$

**1.3 Definición** (Covarianza). La *covarianza* entre dos estadísticos  $Q_1 = Q_1(S)$  y  $Q_2 = Q_2(S)$  es definida como

$$\begin{aligned} \mathbf{Cov}[Q_1, Q_2] &= \mathbf{E}\{[Q_1 - \mathbf{E}(Q_1)][Q_2 - \mathbf{E}(Q_2)]\} \\ &= \sum_{s \in \mathcal{S}} \mathbf{p}(s)[Q_1(s) - \mathbf{E}(Q_1)][Q_2(s) - \mathbf{E}(Q_2)] \quad (\text{Covarianza}). \end{aligned} \quad (1.2.3)$$

Para definir el concepto de *estimador* piénsese en un estadístico diseñado con el propósito de dar un valor aproximado a una cantidad desconocida de la población

<sup>1</sup>Para estudiar en mayor detalle lo expuesto en esta sección se pueden consultar los textos de Särndal, Swensson & Wretman (1992); Lohr (1999); Groves, Fowler-Jr., Couper, Lepkowski, Singer & Tourangeau (2004) entre otros.

$\mathbb{U}$ . Esto es, piénsese en diseñar una función de valor real que permita *estimar* el valor de la cantidad desconocida con base en la información contenida en la muestra. La función que permite esta estimación es llamada *estimador*.

Llámesese *parámetro* a la cantidad desconocida de la población  $\mathbb{U}$  y nótese de manera general con  $\theta$ . Supóngase que es posible observar y medir una variable de interés, por ejemplo  $y$ , sobre todo el universo. Esto es, se puede medir el valor de  $y$  para cada uno de los individuos del universo desde  $1, \dots, k, \dots, N$ . Se puede pensar a  $\theta$  como una función de  $y_1, \dots, y_k, \dots, y_N$ , así

$$\theta = \theta(y_1, \dots, y_k, \dots, y_N).$$

Algunos ejemplos de lo anterior se encuentran expresados en las ecuaciones (1.1.2) y (1.1.3) de la sección anterior. Tanto  $t$  como  $\bar{y}_{\mathbb{U}}$  son función de los valores de  $y$  para todos los elementos del universo. Por tanto, ambos son parámetros. Retomando el concepto de *estimador*, este se puede definir de la siguiente forma.

**1.4 Definición** (Estimador). Sea  $\theta$  un parámetro de la población y  $s$  una realización de la variable aleatoria  $S$  bajo un diseño muestral  $\mathbf{p}(\cdot)$ , cualquier función  $\hat{\theta} = \hat{\theta}(S)$  de la forma

$$\begin{aligned} \hat{\theta} : \mathcal{S} &\mapsto \mathbb{R} \\ s &\mapsto \hat{\theta}(s) \end{aligned}$$

será un *estimador* de  $\theta$ .

De acuerdo a lo anterior,  $\hat{\theta}$  es un estadístico diseñado para calcular un valor que aproxime el valor desconocido del parámetro poblacional  $\theta$  en función de  $s$ . Otra forma útil de ver a  $\hat{\theta}$  es como función de los valores  $y$  presentes en la muestra.

$$\hat{\theta} = \hat{\theta}(y_1, \dots, y_k, \dots, y_n) \quad \text{para } k \in s.$$

Dado que  $\hat{\theta}$  es una variable aleatoria, se puede definir, al igual que con  $Q$ , las mismas propiedades que se establecieron en la Definición 1.2.

**1.5 Definición** (Esperanza y Varianza de un Estimador). La *esperanza* y la *varianza* de  $\hat{\theta}$  están dadas por

$$\mathbf{E}[\hat{\theta}] = \sum_{s \in \mathcal{S}} \mathbf{p}(s) \hat{\theta}(s) \quad (\text{Esperanza}), \quad (1.2.4)$$

$$\mathbf{Var}(\hat{\theta}) = \sum_{s \in \mathcal{S}} \mathbf{p}(s) \{\hat{\theta}(s) - \mathbf{E}(\hat{\theta})\}^2 \quad (\text{Varianza}). \quad (1.2.5)$$

Dos medidas de la precisión del estimador se definen a continuación

**1.6 Definición** (Sesgo y Error Cuadrático Medio de un Estimador). El *sesgo* de  $\hat{\theta}$  se define como

$$\mathbf{B}(\hat{\theta}) = \mathbf{E}[\hat{\theta}] - \theta \quad (\text{Sesgo}). \quad (1.2.6)$$

$\hat{\theta}$  se dice insesgado para  $\theta$  si

$$\mathbf{B}(\hat{\theta}) = 0,$$

de donde se tiene que  $\mathbf{E}[\hat{\theta}] = \theta$ . El *error cuadrático medio* (**ECM**) de  $\hat{\theta}$  se define como

$$\mathbf{ECM}(\hat{\theta}) = \mathbf{E}[\hat{\theta} - \theta]^2 = \sum_{s \in \mathcal{S}} \mathbf{p}(s) [\hat{\theta}(s) - \theta]^2 \quad (\mathbf{ECM}) \quad (1.2.7)$$

De donde se puede verificar que

$$\mathbf{ECM}(\hat{\theta}) = \mathbf{Var}(\hat{\theta}) + [\mathbf{B}(\hat{\theta})]^2. \quad (1.2.8)$$

Si  $\hat{\theta}$  es insesgado, entonces  $\mathbf{ECM}(\hat{\theta}) = \mathbf{Var}[\hat{\theta}]$ .

Para finalizar es necesario precisar la diferencia entre *estimador* y *estimación*. Por *estimación* se entenderá que el número  $\hat{\theta}(s)$  es calculado después de que una muestra  $s$  del conjunto aleatorio  $S$  es seleccionada y han sido observados y recopilados los valores para la variable de estudio  $y$  para cada elemento  $k \in s$ . El estimador no es otra cosa que la función  $\hat{\theta}(S)$  que permite estimar el valor del parámetro  $\theta$ .

## 1.3. Estimadores para el Total Poblacional

Se presenta en esta sección tres estimadores para la estimación del total  $t$  de una variable  $y$ . Se expondrá brevemente el estimador de *Horvitz-Thompson* o  $\pi$ -estimador, el estimador general de regresión o *GREG* y una extensión de estos últimos, los estimadores *GEREG*. Del primero se resalta la idea de la  $\pi$ -expansión como medio de alcanzar la población y así estimar el total. De los otros dos se resalta la implementación de la información auxiliar en el proceso de estimación del parámetro.

### 1.3.1. Estimador de Horvitz-Thompson (H-T) o $\pi$ -estimador

**1.7 Definición** ( $\pi$ -expansión). Sea  $y_k$  el valor de  $y$  para un elemento de la muestra  $s$  de la población  $\mathbb{U}$ . Al resultado de dividir  $y_k$  entre la probabilidad de inclusión de primer orden  $\pi_k$  definida por la ecuación (1.1.5) de la forma

$$\check{y}_k = \frac{y_k}{\pi_k} \quad (1.3.1)$$

se denomina  $\pi$ -expansión, y al valor  $\check{y}_k$  se le denomina  $y_k$ -expandido.

La  $\pi$ -expansión tiene el efecto de expandir el valor de los elementos en la muestra. Dado que la muestra tiene pocos elementos, esta expansión es necesaria para alcanzar el nivel de la población. En general, un estimador que involucre la  $\pi$ -expansión como modo de alcanzar la población total, se considerará un  $\pi$ -estimador.

**1.8 Definición** (El  $\pi$ -Estimador del Total). Sea  $t = \sum_{\mathbb{U}} y_k$  el total poblacional de la variable de interés  $y$ , el estimador

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k} \quad (1.3.2)$$

es el  $\pi$ -estimador del total poblacional de  $y$ .

Equivalentes a la expresión anterior, se tienen las siguientes formas

$$\hat{t}_\pi = \sum_{\mathbb{U}} I_k \frac{y_k}{\pi_k}, \quad (1.3.3)$$

$$\hat{t}_\pi = \sum_s \check{y}_k = \sum_{\mathbb{U}} \check{y}_k I_k, \quad (1.3.4)$$

donde  $I_k$  es la variable indicadora de pertenencia definida en la ecuación (1.1.4) y  $\check{y}_k = y_k/\pi_k$ . Para el  $\pi$ -estimador del total se tienen las siguientes propiedades.

**1.2 Proposición** (Esperanza). *La esperanza del estimador de Horvitz-Thompson o  $\pi$ -estimador es  $t$ .*

$$E[\hat{t}_\pi] = t \quad (1.3.5)$$

*Demostración.*

$$\begin{aligned} E[\hat{t}_\pi] &= E \left[ \sum_{\mathbb{U}} I_k \frac{y_k}{\pi_k} \right] \\ &= \sum_{\mathbb{U}} E \left[ I_k \frac{y_k}{\pi_k} \right] \\ &= \sum_{\mathbb{U}} \frac{y_k}{\pi_k} E [I_k] \\ &= \sum_{\mathbb{U}} \frac{y_k}{\pi_k} \pi_k \\ &= \sum_{\mathbb{U}} y_k \\ &= t \quad \square \end{aligned}$$

**1.9 Definición** (Insesgamiento y Varianza). De la ecuación (1.2.6) y de la proposición 1.2 se deriva el hecho de que el  $\pi$ -estimador definido por (1.3.2) es insesgado para  $t = \sum_{\mathbb{U}} y_k$  y tiene como varianza

$$V(\hat{t}_\pi) = \sum \sum_{\mathbb{U}} \Delta_{kl} \check{y}_k \check{y}_l \quad (1.3.6)$$

donde  $\Delta_{kl}$  está dado por la ecuación (1.1.9)<sup>2</sup>. Un estimador insesgado de  $V(\hat{t}_\pi)$  está dado por

$$\hat{V}(\hat{t}_\pi) = \sum \sum_s \check{\Delta}_{kl} \check{y}_k \check{y}_l \quad (1.3.7)$$

donde  $\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$  y  $\pi_{kl}$  es la probabilidad de inclusión de segundo orden definida por la ecuación (1.1.6).

<sup>2</sup>Ver Särndal et al. (1992, Resultado 2.6.1, p. 36) donde se demuestra este resultado.

### 1.3.2. Estimadores de Regresión

La idea de involucrar *información auxiliar* de la población en el proceso de estimación del parámetro  $\theta$  tiene como fin diseñar un estimador del parámetro de interés que sea más preciso y eficiente en comparación con otro tipo de estimador. A continuación se presentan algunas ideas al respecto.

**1.10 Definición** (Información Auxiliar). Una variable o conjunto de variables se considerará *información auxiliar* si la información para estas variables se encuentra disponible, antes del proceso de muestreo, para toda la población objetivo en el marco muestral. En general se tiene que el vector de información auxiliar para el  $k$ -ésimo elemento de  $\mathbb{U}$  está dado por

$$\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kj}, \dots, x_{kJ})$$

donde  $k = 1, \dots, N$  y  $j = 1, \dots, J$  donde  $J$  es la cantidad de variables auxiliares existentes. El conjunto de los  $k$  vectores de información auxiliar conforman la matriz  $\mathbb{X}$  de diseño o de información auxiliar donde  $\mathbb{X} = \{\mathbf{x}_k\}$ , siendo ésta una matriz de tamaño  $N \times J$ .

La definición anterior no restringe de ninguna manera la naturaleza de la información auxiliar. Es decir, se puede tener dentro del marco de datos en uso variables cualitativas o cuantitativas que funcionarán como variables auxiliares, siempre y cuando éstas estén disponibles para todos los elementos dentro del marco muestral. Otra manera de abordar el uso de la información auxiliar, en especial para las variables de tipo cuantitativo, es cuando solamente se dispone de los valores muestrales del conjunto de variables auxiliares  $\mathbf{x}_k$  y se conoce de una fuente confiable el total poblacional para cada una de las variables auxiliares,  $t_{\mathbf{x}_k} = (t_{x_{k1}}, \dots, t_{x_{kJ}})$ . Para el caso de las variables cualitativas se debería contar con el total de cada una de las categorías que contenga dicha variable.

Teniendo en cuenta que se desea poder estimar el parámetro  $t_y = \sum_{\mathbb{U}} y_k$  cuando se han observado parejas  $(y_k, \mathbf{x}_k)$  para  $k \in s$  y que  $\mathbf{x}_k$  es conocido para toda la población, se puede definir un estimador acorde con lo expuesto anteriormente.

**1.11 Definición** (Estimador de Regresión<sup>3</sup>). El estimador de regresión, denotado por  $\hat{t}_{yr}$  se define formalmente como

$$\hat{t}_{yr} = \hat{t}_{y\pi} + \sum_{j=1}^J \hat{\beta}_j (t_{x_j} - \hat{t}_{x_j\pi}), \quad (1.3.8)$$

donde

$$\hat{t}_{y\pi} = \sum_s \frac{y_k}{\pi_k} = \sum_s \check{y}_k$$

<sup>3</sup>En la construcción del estimador de regresión Särndal et al. (1992) primero desarrollan un estimador denominado *estimador de diferencia* y a partir de éste definen el estimador de regresión (Särndal et al., 1992, p. 221-225)

es el  $\pi$ -estimador de  $t_y = \sum_{\mathbb{U}} y_k$ ,

$$\hat{t}_{x_j \pi} = \sum_s \frac{x_{kj}}{\pi_k} = \sum_s \check{x}_{kj}$$

es el  $\pi$ -estimador del total conocido de  $\mathbf{x}_{.j}$ ,

$$t_{x_j} = \sum_{\mathbb{U}} x_{kj}$$

es el total conocido de  $\mathbf{x}_{.j}$  y

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_J)' = \left( \sum_s \frac{\mathbf{x}_k \cdot \mathbf{x}'_k}{\sigma_k^2 \pi_k} \right)^{-1} \sum_s \frac{\mathbf{x}'_k \cdot y_k}{\sigma_k^2 \pi_k} \quad (1.3.9)$$

donde  $\sigma_k^2$  son varianzas asociadas a cada una de las observaciones  $y_k$  consideradas como variables aleatorias.

*1.1 Nota.* La forma particular del vector de parámetros dado por la ecuación (1.3.9) puede ser deducida de la siguiente manera.

$$\begin{aligned} \boldsymbol{\beta} &= (\beta_1, \dots, \beta_J)' \\ &= \left( \sum_{\mathbb{U}} \frac{\mathbf{x}_k \cdot \mathbf{x}'_k}{\sigma_k^2} \right)^{-1} \sum_{\mathbb{U}} \frac{\mathbf{x}'_k \cdot y_k}{\sigma_k^2}. \end{aligned}$$

o

$$\boldsymbol{\beta} = (\mathbb{X} \boldsymbol{\Sigma}^{-1} \mathbb{X}')^{-1} \mathbb{X}' \boldsymbol{\Sigma}^{-1} \mathbb{Y}$$

donde

$$\mathbb{Y} = (y_1, \dots, y_N)$$

y  $\boldsymbol{\Sigma}$  es una matriz diagonal de tamaño  $J \times J$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_J^2 \end{pmatrix}.$$

Haciendo

$$\boldsymbol{\beta} = \mathbf{T}^{-1} \mathbf{t}$$

donde

$$\mathbf{T} = \sum_{\mathbb{U}} \frac{\mathbf{x}_k \cdot \mathbf{x}'_k}{\sigma_k^2} \quad \text{y} \quad \mathbf{t} = \sum_{\mathbb{U}} \frac{\mathbf{x}'_k \cdot y_k}{\sigma_k^2}.$$

El  $\pi$ -estimador para  $\mathbf{T}$  y  $\mathbf{t}$  estará dado, respectivamente por

$$\hat{\mathbf{T}} = \sum_s \frac{\mathbf{x}_k \cdot \mathbf{x}'_k}{\sigma_k^2} \quad \text{y} \quad \hat{\mathbf{t}} = \sum_s \frac{\mathbf{x}'_k \cdot y_k}{\sigma_k^2}.$$

y por tanto

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \hat{\mathbf{T}}^{-1}\hat{\mathbf{t}} \\ &= \left( \sum_s \frac{\mathbf{x}_k \cdot \mathbf{x}'_k}{\sigma_k^2 \pi_k} \right)^{-1} \sum_s \frac{\mathbf{x}'_k \cdot y_k}{\sigma_k^2 \pi_k}.\end{aligned}\tag{1.3.10}$$

de donde se deduce la expresión dada por (1.3.9).

Con respecto a la varianza del estimador de regresión se tiene la siguiente definición

**1.12 Definición.** El estimador de regresión dado por la ecuación (1.3.8) es aproximado por linealización de Taylor como

$$\begin{aligned}\hat{t}_{yr0} &= \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})\boldsymbol{\beta} \\ &= \sum_{\mathbb{U}} y_k^0 + \sum_s \check{E}_k,\end{aligned}\tag{1.3.11}$$

donde  $\check{E}_k = E_k/\pi_k$ ,  $E_k = y_k - y_k^0$  y  $y_k^0 = \mathbf{x}_k \hat{\boldsymbol{\beta}}$ . El estimador  $\hat{t}_{yr}$  es aproximadamente insesgado para  $t = \sum_{\mathbb{U}} y_k$  con varianza aproximada<sup>4</sup> dada por

$$\mathbf{AV}(\hat{t}_{yr}) = \sum \sum_{\mathbb{U}} \Delta_{kl} \check{E}_k \check{E}_l.\tag{1.3.12}$$

El estimador de la varianza de (1.3.12) viene dado por

$$\hat{\mathbf{V}}(\hat{t}_{yr}) = \sum \sum_s \check{\Delta}_{kl} (g_{ks} \check{e}_{ks}) (g_{ls} \check{e}_{ls}),\tag{1.3.13}$$

donde

$$\begin{aligned}\check{e}_{ks} &= \frac{e_{ks}}{\pi_k}, \\ e_{ks} &= y_k - \hat{y}_k, \\ \hat{y}_k &= \mathbf{x}_k \hat{\boldsymbol{\beta}}, \\ &= \sum_{j=1}^J \hat{\beta}_j x_{kj}, \\ g_{ks} &= 1 + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})' \hat{\mathbf{T}}^{-1} \mathbf{x}_k / \sigma_k^2.\end{aligned}$$

En lo que respecta a la definición del estimador de regresión se pueden deducir seis expresiones alternativas a la mostrada en (1.3.8)<sup>5</sup>. Como ejemplo se tiene

$$\hat{t}_{yr} = \sum_{\mathbb{U}} \hat{y}_k + \sum_s \check{e}_{ks}\tag{1.3.14}$$

que es equivalente a la expresión dada en (1.3.8).

<sup>4</sup>La demostración de este resultado se encuentra en (Särndal et al., 1992, p. 236).

<sup>5</sup>Para estudiar otras formas del estimador de regresión ver (Särndal et al., 1992, p. 234).

Un aspecto fundamental en el planteamiento del estimador de regresión es la elección de los coeficientes  $\hat{\beta}_j$  en la ecuación (1.3.8). La escogencia está basada en la suposición sobre la forma de la dispersión de los puntos de la población finita  $\{(y_k, x_{k1}, \dots, x_{kJ}) : k = 1, \dots, N\}$ . Es decir, se supone que la dispersión de los  $N$  puntos de la población luce como si hubiera sido generados por un modelo de regresión lineal, denominado  $\xi$ . Aquí  $y_k$  y  $x_1, \dots, x_J$  son respectivamente las variables respuesta y regresoras del modelo. El modelo  $\xi$  puede ser formalmente caracterizado como sigue.

**1.13 Definición** (Modelo de regresión). El modelo  $\xi$  que asiste al estimador de regresión tiene las siguiente características

- i  $y_1, \dots, y_N$  se asumen como realizaciones de variables aleatorias independientes  $Y_1, \dots, Y_N$ .
- ii  $\mathbf{E}_\xi(Y_k) = \sum_{j=1}^J \beta_j x_{kj}$  con  $k = 1, \dots, N$ ; y
- iii  $\mathbf{Var}_\xi(Y_k) = \sigma_k^2$  con  $k = 1, \dots, N$

donde  $\mathbf{E}_\xi$  y  $\mathbf{Var}_\xi$  denotan el valor esperado y la varianza con respecto del modelo  $\xi$  y donde  $\beta_1, \dots, \beta_J$  y  $\sigma_1^2, \dots, \sigma_N^2$  son los parámetros del modelo.

El papel del modelo es describir puntualmente la dispersión de la población finita. En este sentido se espera que el modelo  $\xi$  se ajuste razonablemente bien a la población o que se pueda pensar que la población finita parezca haber sido generada en conformidad con el modelo  $\xi$ . Sin embargo, estos supuestos no significan de ninguna manera que se piense que la población fue realmente generada por el modelo  $\xi$ . Las conclusiones sobre los parámetros de la población finita son por lo tanto independientes de los supuestos que se tengan sobre el modelo.

Por otra parte, el modelo sirve como medio para encontrar apropiados  $\hat{\beta}$  que puedan ser puestos en la formula de la ecuación (1.3.8) del estimador de regresión. Por último, la eficiencia del estimador en comparación con el  $\pi$ -estimador depende de la bondad de ajuste del modelo a la dispersión de los datos. Sin embargo, las propiedades del estimador de regresión (su aproximada insesgadez o los estimadores de su varianza) no dependen de si el modelo  $\xi$  ajusta o no.

En lo que sigue se mostrará una extensión sobre el modelo  $\xi$  usado en el estimador de regresión hacia modelos probabilísticos de la familia exponencial.

### 1.3.3. Estimadores Asistidos por Modelos Lineales Generalizados (MLG).

En esta sección se aborda el caso general del estimador de regresión en el cual el modelo  $\xi$  que explica la relación existente entre la variable de interés y las variables auxiliares puede ser descrita de mejor forma por modelos lineales generalizados diferentes al normal. El resultado es un estimador de regresión asistido por modelos

lineales generalizados o *GEREG* (en inglés *generalized linear model regression estimator*). Para lo anterior, tomaremos como referencia el artículo de Rondón et al. (2012) el cual desarrolla estos aspectos.

En los modelos lineales generalizados, el vector  $(y_1, \dots, y_N)$  son realizaciones del vector  $(Y_1, \dots, Y_N)$  de variables aleatorias independientes e idénticamente distribuidas, donde  $Y$  es la variable de interés que tiene una función de distribución de probabilidad de la familia exponencial. Es decir la función de distribución de probabilidad para  $y$  es de la forma

$$f(y; \theta_k, \phi_k) = \exp \left\{ \frac{[y\theta_k - b(\theta_k)]}{\phi_k} + c(y, \phi_k) \right\} \quad (1.3.15)$$

en donde  $c(\cdot)$  es una función conocida,  $\theta_k$  es el parámetro canónico del modelo y  $\mathbf{E}(Y_k) = \mu_k = b'(\theta_k)$ . Esta función corresponde al componente aleatorio del modelo lineal generalizado. El componente sistemático esta dado por

$$g(\mu_k) = \eta_k = \sum_{j=1}^J \beta_j x_{kj} = \mathbf{x}_k \cdot \boldsymbol{\beta}_j \quad (1.3.16)$$

donde  $\mathbf{x}_k = (x_{k1}, \dots, x_{kJ})$  es un vector de  $J$  variables explicativas para los  $k$  elementos,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$  un vector de parámetros y  $g(\cdot)$  una función monótona diferenciable denominada función de enlace.

Para la estimación de los parámetros se emplea el método de máxima verosimilitud y se puede considerar en tres escenarios distintos: estimación sobre la población total, estimación sobre la muestra o estimación sobre la muestra incluyendo pesos muestrales. Lo anterior implica que para los parámetros poblacionales  $\mu$  y  $\beta$  se tendrán estimadores  $\hat{\mu}_k^U$ ,  $\hat{\mu}_k^s$ ,  $\hat{\mu}_k^{s\pi}$ ,  $\hat{\beta}_U$ ,  $\hat{\beta}_s$  y  $\hat{\beta}_s^\pi$  respectivamente. Teniendo en cuenta que estos usan la información contenida en la población, en la muestra o en la muestra con pesos muestrales. La Tabla 1.1 resume lo anterior (Rondón et al., 2012, p. 682).

	Con información sobre $s$	
	Pesos muestrales	Sin pesos muestrales
$\beta$	$\hat{\beta}_U = \arg \max L_U(\beta)$	$\hat{\beta}_s^\pi = \arg \max L_s^\pi(\beta)$ $\hat{\beta}_s = \arg \max L_s(\beta)$
$\mu_k$	$\hat{\mu}_k^U = g^{-1}(\mathbf{x}'_k \hat{\beta}_U)$	$\hat{\mu}_k^{s\pi} = g^{-1}(\mathbf{x}'_k \hat{\beta}_s^\pi)$ $\hat{\mu}_k^s = g^{-1}(\mathbf{x}'_k \hat{\beta}_s)$

TABLA 1.1. Estimación de  $\boldsymbol{\beta}$  y  $\mu$ .

Bajo este último escenario (estimación sobre la muestra incluyendo pesos muestrales) los autores hacen uso de la idea de pseudoverosimilitud, que no es más que la inclusión de los pesos muestrales  $1/\pi_k$  en la función de verosimilitud definida sobre

la muestra  $s$  seleccionada bajo el diseño muestral  $\mathbf{p}(\cdot)$ <sup>6</sup>. El logaritmo de la función de verosimilitud que incluye los pesos muestrales puede ser escrito como

$$\mathcal{L}_s^\pi(\beta) = \sum_{k \in s} \frac{\phi_k}{\pi_k} [y_k \theta(\beta; \mathbf{x}_k) - b(\theta(\beta; \mathbf{x}_k))], \quad (1.3.17)$$

de donde se tiene que  $\hat{\beta}_s^\pi = \arg \max \mathcal{L}_s^\pi(\beta)$  y  $\hat{\mu}_k^s = g^{-1}(\mathbf{x}_k \hat{\beta}_s^\pi)$  son los estimadores pseudomáximos verosímiles de  $\beta$  y  $\mu$ . Con lo anterior es posible definir el estimador de regresión asistido por modelos lineales generalizados o GEREG.

**1.14 Definición** (GEREG). El estimador de regresión asistido por modelos lineales generalizados para  $t_y = \sum_{\cup} y_k$  se define como

$$\hat{t}_G = \sum_{k \in U} k \in U \hat{\mu}_k^s + \sum_{k \in s} \frac{(y_k - \hat{\mu}_k^s)}{\pi_k} \quad (1.3.18)$$

donde  $\hat{\mu}_k^s = g^{-1}(\mathbf{x}_k \hat{\beta}_s^\pi)$  y con  $\hat{\beta}_s^\pi$  como el estimador basado en la muestra  $s$ .

Ahora bien, el modelo  $\xi$  que describe la relación existente entre la variable respuesta y las variables auxiliares y que asiste al estimador de regresión, se define del siguiente modo.

**1.15 Definición.** El modelo  $\xi$  que describe la dispersión de los datos de la población se expresa como

$$\begin{cases} \mathbf{E}_\xi[Y_k] = \mu_k \\ \mathbf{Var}_\xi(Y_k) = \phi_k^{-1} \mathbf{Var}(\mu_k) \\ g(\mu_k) = \sum_{j=1}^J \beta_j x_{kj} = \mathbf{x}_k \cdot \boldsymbol{\beta} \end{cases} \quad (1.3.19)$$

donde  $\boldsymbol{\beta}$  es un vector columna que contiene los parámetros desconocidos del modelo,  $g(\cdot)$  es la función de enlace,  $\mathbf{x}_k = (x_{k1}, \dots, x_{kJ})$  es el vector de información auxiliar para el  $k$ -ésimo elemento de la población.  $\mathbf{E}_\xi(\cdot)$  y  $\mathbf{Var}_\xi(\cdot)$  son respectivamente, el valor esperado y la varianza de la variable  $Y_k$  bajo modelo  $\xi$ .

## 1.4. Censura

La censura es un fenómeno que se presenta sobre una variable cuando, para los posibles valores que ésta puede tomar, sólo se sabe que estos están por encima o por debajo de un valor *límite de censura*  $T$ . En ocasiones a este tipo de datos, en especial en ciencias medioambientales, se les conoce como *observaciones por debajo del límite de detección* (Millard & Neerchal, 2000, p. 600).

<sup>6</sup>La función de verosimilitud o de log-verosimilitud en la muestra  $s$  que considera los pesos muestrales se denominará como verosimilitud  $\pi$ -ponderada o pseudoverosimilitud y logaritmo de la verosimilitud  $\pi$ -ponderada o logaritmo de la pseudoverosimilitud respectivamente.

Existen cuatro grandes formas de clasificar muestras que contienen observaciones censuradas, que pueden ser combinadas para la descripción detallada de las mismas. Cohen (1991) y Millard & Neerchal (2000) hacen esta distinción. La Tabla 1.2 presenta de forma sencilla estas cuatro formas. Las definiciones formales se dejarán para los casos correspondientes a *censura a izquierda* y *censura a derecha*. Estos casos son estudiados en el presente trabajo.

*1.2 Nota.* Para poder definir algunas ideas presentes en este apartado es necesario introducir notación extra. Se considerará de ahora en adelante que la muestra  $s$  será una muestra censurada de acuerdo a algunas de las categorías expuestas en la Tabla 1.2. En estas  $c$  denotará la cantidad de observaciones que se consideran censuradas y  $\tilde{c}$  la cantidad de observaciones no censuradas. Como consecuencia de lo anterior tendremos que la diferencia  $n - \tilde{c} = c$ . En el caso simple, el punto de censura o truncamiento se denotará como  $T$  independientemente de si se tiene censura por derecha o por izquierda.

**1.16 Definición** (Observaciones censuradas). Sea  $s$  una muestra aleatoria de tamaño  $n$  y  $y_k$  una variable observada y medida para cada individuo  $k \in s$ . Si se tiene que

$$y_k = \begin{cases} y_k & \text{si } y_k > T \\ T & \text{si } y_k \leq T \end{cases} \quad (1.4.1)$$

entonces  $y_k = T$  será una *observación censurada a izquierda* para ciertos  $k \in s$ . Si por el contrario se tiene que

$$y_k = \begin{cases} y_k & \text{si } y_k < T \\ T & \text{si } y_k \geq T \end{cases} \quad (1.4.2)$$

entonces  $y_k = T$  será una *observación censurada a derecha* para ciertos  $k \in s$ .

Relacionada con  $y$  existe una variable no observada denominada  $y^*$  la cual es conocida, en algunas ocasiones, como *variable latente*. Esto es debido a que la variable no puede ser observada directamente para algunos valores, pero estos se pueden inferir de los valores de otras variables que sí fueron medidas directamente a través de un modelo (ver Greene (2007, p. 871)). En este sentido, en su definición de censura, Bolfarine et al. (2013, pág. 1) mencionan lo siguiente:

*“Censura ocurre cuando los datos sobre una variable dependiente (respuesta) no están disponibles para algunas unidades de la muestra. Sin embargo para estas unidades, los datos para variables independientes (regresoras) están disponibles. Por ejemplo, personas de todos los niveles de renta son incluidas en una muestra pero, por alguna razón, las personas con un alto nivel de ingreso tienen una misma categoría \$1'000.000”*<sup>7</sup>.

La justificación de la existencia de  $y^*$  está en el hecho de que la variable  $y$  es el resultado de un proceso de observación bajo una condición de censura  $T$  de la variable  $y^*$ . En otras palabras  $y^*$  y  $y$  serían la misma variable si la condición de censura

<sup>7</sup>Traducción libre del portugués

Forma de Clasificación	Clasificación	Descripción
Forma 1	Observado	El caso general en el que no se presenta ningún tipo de censura. Los valores se usan tal y como son originalmente.
	Censurado	Los <i>valores censurados</i> son los que se presentan como menores que algún valor límite $T$ , mayores que algún valor límite $T$ o en un intervalo $(T_1, T_2)$ .
	Truncado	Los <i>valores truncados</i> son aquellos que no se presentan si el valor excede cierto límite.
Forma 2	Censura Izquierda	Un <i>valor censurado por izquierda</i> es aquel del que sólo se sabe que es menor que cierto valor límite $T$ .
	Censura derecha	Un <i>valor censurado por derecha</i> es aquel del que sólo se sabe que es mayor que cierto valor límite $T$ .
	Censura Intervalo	Un valor censurado en intervalo es aquel que se presenta como si estuviera dentro de un intervalo específico.
Forma 3	Tipo I	Muestras censuradas tipo I son aquellas en las que el valor límite de censura es conocido previamente, pero la cantidad de datos censurados y los no censurados en la muestra son aleatorios.
	Tipo II	Muestras censuradas tipo II son aquellas en las que el número de observaciones censuradas y no censuradas en la muestras son conocidas previamente, pero el valor del límite de censura es aleatorio.
	Censura Aleatoria	Muestras aleatoriamente censuradas son aquellas en las que tanto el número de observaciones censuradas, no censuradas y el valor del límite de censura son valores aleatorios no conocidos.
Forma 4	Censura Simple	Una muestra presenta <i>censura simple</i> si sólo existe un límite de censura $T$ .
	Censura Múltiple	Una muestra presenta <i>censura múltiple</i> si existe más de un límite de censura $T_1, \dots, T_n$ .

TABLA 1.2. Formas de clasificación para muestras con observaciones censuradas.

no existiera. En consecuencia, el papel de la variable  $y^*$  será determinante para la definición del modelo de regresión Tobit, además de ser ésta la variable sobre la que se está interesado en hacer inferencia del total. Este hecho se aclarará más adelante cuando se presente el problema de investigación.

Por último, se usarán indistintamente las expresiones *variable latente* o *variable no observada* para referirse a la variable  $y^*$  y las expresiones *variable censurada* o *variable observada* para referirse a la variable  $y$ .

### 1.4.1. Distribución Normal Censurada

Otro factor de clasificación para las muestras censuradas (así como para las truncadas) es el tipo de distribución que sigue la variable de interés. Por ejemplo, muestras obtenidas de una variable que se distribuye normalmente y que incluyen censura en para algunas observaciones, serán consideradas como muestras normales censuradas. Cohen (1991) presenta distintas opciones de distribuciones que permiten modelar el comportamiento de la variable de interés cuando ésta contiene observaciones censuradas. En lo que sigue, se expondrá el desarrollo para el caso donde se asume normalidad sobre la variable de interés  $y^*$  o variable latente.

Como se mencionó antes, se supone que  $y^*$  es una variable latente que se distribuye  $N(\mu, \sigma^2)$  con punto de censura a izquierda  $T$  igual a una constante  $a$  y que existe una variable censurada  $y$  tal que

$$y = \begin{cases} y^* & \text{si } y^* > a \\ a & \text{si } y^* \leq a. \end{cases} \quad (1.4.3)$$

Se tiene entonces que si  $y = a$

$$f_Y(y) = \Pr(y = a) = \Pr(y^* \leq a) = \Pr\left(\frac{y^* - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right). \quad (1.4.4)$$

Cuando  $y > a$

$$f_Y(y) = \frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right). \quad (1.4.5)$$

Aquí  $\phi$  y  $\Phi$  denotan las funciones de densidad y distribución acumulada de una normal estándar respectivamente. La función  $f_Y$  es una combinación de una función de distribución de probabilidad discreta en  $y = a$  y una función de densidad continua para  $y > a$ . Se define ahora una variable indicadora de censura de la siguiente forma

$$\delta = \begin{cases} 1 & \text{si } y^* > a \\ 0 & \text{si } y^* \leq a. \end{cases} \quad (1.4.6)$$

De acuerdo a (1.4.4) y (1.4.5) se puede definir lo siguiente.

**1.17 Definición** (Distribución Normal Censurada a Izquierda). Sea  $y^* \sim N(\mu, \sigma^2)$  una variable latente y  $y$  la variable censurada a izquierda con punto de censura

$T = a$  según (1.4.3). La función de distribución censurada para  $y$  está dada por

$$f(y) = \Phi\left(\frac{a - \mu}{\sigma}\right)^{1-\delta} \left(\frac{1}{\sigma}\phi\left(\frac{y - \mu}{\sigma}\right)\right)^\delta = \begin{cases} \Phi\left(\frac{a - \mu}{\sigma}\right) & \text{si } y = a \\ \frac{1}{\sigma}\phi\left(\frac{y - \mu}{\sigma}\right) & \text{si } y > a. \end{cases} \quad (1.4.7)$$

## 1.4.2. Modelo de Regresión Censurado o Modelo Tobit

El modelo deriva su nombre de Tobin (1958) por ser pionero en la investigación de modelos de datos censurados. A partir de ese año diversos autores continuaron extendiendo los resultados sobre datos censurados. En lo que sigue se exponen algunas ideas sobre el modelo Tobit.

Sea  $y_k$  una variable censurada a izquierda en el punto  $T$  y sea  $y_k^*$  su correspondiente variable latente. Suponiendo, sin pérdida de generalidad, que el punto de censura es  $T = 0$  se tiene que

$$y_k = \begin{cases} y_k^* & \text{si } y_k^* > 0 \\ 0 & \text{si } y_k^* \leq 0. \end{cases} \quad (1.4.8)$$

Sea

$$\delta_k = \begin{cases} 1 & \text{si } y_k^* > a \\ 0 & \text{si } y_k^* \leq a, \end{cases} \quad (1.4.9)$$

y  $\mathbb{X}$  la matriz de variables auxiliares con  $\mathbf{x}_k = (x_{k1}, \dots, x_{kj}, \dots, x_{kJ})$  sus vectores filas y  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_j, \dots, \beta_J)'$  un vector columna de parámetros. Consideremos el modelo lineal  $\psi$

$$\begin{cases} y_k = \begin{cases} y_k^* = \mathbf{x}_k \boldsymbol{\beta} + \varepsilon_k & \text{si } y_k^* > T \\ T & \text{si } y_k^* \leq T \end{cases} \\ \text{Donde } \varepsilon_1, \dots, \varepsilon_N \text{ son variables aleatorias iid.} \\ \text{Con } \varepsilon_k \sim N(0, \sigma^2). \end{cases} \quad (1.4.10)$$

Notemos que

$$y_k^* \sim N(\mathbf{x}_k \boldsymbol{\beta}, \sigma^2). \quad (1.4.11)$$

*1.3 Nota.* Se considerará para este trabajo que el modelo  $\psi$  es un modelo de regresión multivariado con intercepto. En general cuando se haga referencia al vector de parámetros  $\boldsymbol{\beta}$  o a su estimador  $\hat{\boldsymbol{\beta}}$  se asumirá que su primer elemento corresponderá al intercepto del modelo y que la matriz  $\mathbb{X}$  de información auxiliar contará con un vector columna de unos. Adicionalmente se considerará que el modelo es de varianza constante u homocedástico. El caso heterocedástico es un caso especial para este tipo de modelos y supone grandes dificultades en la estimación de los parámetros del modelo. Amemiya (1984, p. 23) menciona que para el caso heterocedástico la consistencia de los estimadores se pierde. La consideración de modelos con estructuras distintas a la especificada anteriormente, es decir, modelos sin intercepto o

modelos heterocedásticos exceden los alcances de este trabajo, razón por la cual no son tratados en el presente documento.

Por (1.4.7) y  $T = 0$  y dado que las  $y_k$  son independientes, se tiene que la función de verosimilitud viene dada por

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \prod_{k=1}^N \left[ \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right)^{1-\delta_k} \frac{1}{\sigma} \phi \left( \frac{y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right)^{\delta_k} \right] \quad (1.4.12)$$

y su logaritmo estará dado por

$$\begin{aligned} \ln \mathcal{L}(\boldsymbol{\beta}, \sigma^2) &= \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) + \delta_k \ln \left( \sigma^{-1} \phi \left( \frac{y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) \right) \right] \\ &= \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) \right. \\ &\quad \left. + \delta_k \left( -\ln \sigma^{2^{1/2}} + \ln \left( \frac{1}{\sqrt{2\pi}} \exp \left( \frac{-\left( \frac{y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right)^2}{2} \right) \right) \right) \right] \\ &= \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) \right. \\ &\quad \left. + \delta_k \left( -\frac{1}{2} \ln \sigma^2 - \frac{1}{2} \ln 2\pi - \frac{1}{2\sigma^2} (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta})^2 \right) \right] \\ &= \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) - \frac{\delta_k}{2} \ln \sigma^2 - \frac{\delta_k}{2} \ln 2\pi \right. \\ &\quad \left. - \frac{\delta_k}{2\sigma^2} (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta})^2 \right]. \end{aligned} \quad (1.4.13)$$

*1.4 Nota.* El modelo de regresión Tobit descrito en esta sección estudia el caso de variables simplemente censuradas a izquierda. Sin embargo este no es el único caso que se puede presentar. Thompson & Nelson (2003) estudian modelos de regresión cuando la variable observada presenta dos tipos de censura a la vez, censura a izquierda y en intervalo. Estos autores presentan y evalúan una aproximación mediante estimación máximo verosímil a los parámetros del modelo estudiando el impacto del sesgo y el porcentaje de censura en las estimaciones.

*1.5 Nota.* Desarrollos recientes sobre modelos de regresión en datos con observaciones censuradas pueden ser encontrados en los trabajos de Michalek, Gupta, Kulkarni, Tripathi & Selvavel (1998); Aboueissa & Stoline (2004) y Carson & Sun (2007). Tanto estos trabajos como el de (Thompson & Nelson, 2003), presentan extensiones o variaciones del modelo clásico propuesto por (Tobin, 1958).

### 1.4.2.1. Estimación de los Parámetros del Modelo

Varios métodos pueden ser usados para la estimación de los parámetros  $\beta$  y  $\sigma$ . Amemiya (1984) describe algunos métodos para realizar esta tarea, pero en general el método más usado corresponde al de Máxima Verosimilitud. A continuación se presenta en detalle este método.

Maximizar (1.4.13) es equivalente a maximizar

$$\ln \mathcal{L}(\beta, \sigma^2) = \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \beta}{\sigma} \right) - \frac{\delta_k}{2} \ln \sigma^2 - \frac{\delta_k}{2\sigma^2} (y_k - \mathbf{x}_k \cdot \beta)^2 \right] \quad (1.4.14)$$

dado que ambas funciones tendrán su máximo para los mismos valores de  $\beta$  y  $\sigma^2$ . Haciendo que

$$z = \frac{-\mathbf{x}_k \cdot \beta}{\sigma} \quad (1.4.15)$$

y derivando parcialmente e igualando a cero la ecuación (1.4.14) se tiene lo siguiente

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\beta, \sigma^2)}{\partial \beta_j} &= \sum_{k=1}^N \left[ (1 - \delta_k) \left( \frac{\phi(z)}{\Phi(z)} \right) \left( -\frac{1}{\sigma} \right) x_{kj} + \delta_k \left( \frac{1}{\sigma^2} (y_k - \mathbf{x}_k \cdot \beta) x_{kj} \right) \right] \\ &= -\frac{1}{\sigma} \sum_{k=1}^N (1 - \delta_k) \left( \frac{\phi(z)}{\Phi(z)} \right) x_{kj} + \frac{1}{\sigma^2} \sum_{k=1}^N \delta_k (y_k - \mathbf{x}_k \cdot \beta) x_{kj} = 0 \end{aligned} \quad (1.4.16)$$

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\beta, \sigma^2)}{\partial \sigma^2} &= \sum_{k=1}^N \left[ (1 - \delta_k) \left( \frac{\phi(z)}{\Phi(z)} \right) \left( -\frac{1}{2\sigma^3} \right) (-\mathbf{x}_k \cdot \beta) - \frac{\delta_k}{2\sigma^2} + \frac{\delta_k}{2\sigma^4} (y_k - \mathbf{x}_k \cdot \beta)^2 \right] \\ &= \frac{1}{2\sigma^3} \sum_{k=1}^N (1 - \delta_k) \left( \frac{\phi(z)}{\Phi(z)} \right) \mathbf{x}_k \cdot \beta - \sum_{k=1}^N \frac{\delta_k}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^N \delta_k (y_k - \mathbf{x}_k \cdot \beta)^2 \\ &= \frac{1}{2\sigma^3} \sum_{k=1}^N (1 - \delta_k) \left( \frac{\phi(z)}{\Phi(z)} \right) \mathbf{x}_k \cdot \beta - \frac{\tilde{c}}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^N \delta_k (y_k - \mathbf{x}_k \cdot \beta)^2 = 0. \end{aligned} \quad (1.4.17)$$

Donde  $\tilde{c} = \sum_{k=1}^N \delta_k$  o la cantidad total de datos no censurados. Amemiya (1973) demostró que el estimador máximo verosimil del modelo Tobit es fuertemente consistente y asintóticamente normal y aclaró, al igual que Tobin (1958), que las ecuaciones (1.4.16) y (1.4.17) no son lineales en sus parámetros y por tanto deben aproximarse sus soluciones por métodos iterativos. Olsen (1978)<sup>8</sup>, por su parte, demuestra la concavidad global del  $\ln \mathcal{L}$  en términos de los parámetros transformados  $\alpha = \beta/\sigma$  y  $h = \sigma^{-1}$ , lo que implica que los métodos iterativos como Newton-Raphson o el método de scoring siempre convergen al máximo global del  $\ln \mathcal{L}$ .

<sup>8</sup>Gourieroux (2000, p. 176) presenta de otra manera la demostración del resultado dado por Olsen (1978). Además anexa algunos detalles y comentarios útiles para la interpretación del resultado.

La ecuación (1.4.14) queda expresada en términos de los nuevos parámetros  $\alpha$  y  $h$  de la siguiente manera

$$\ln \mathcal{L}(\alpha, h) = \sum_{k=1}^N \left[ (1 - \delta_k) \ln \Phi(-\mathbf{x}_k \cdot \alpha) + \delta_k \ln h - \frac{\delta_k}{2} (hy_k - \mathbf{x}_k \cdot \alpha)^2 \right] \quad (1.4.18)$$

Esta transformación provee de expresiones mucho más simples para las derivadas del logaritmo de la función de verosimilitud expresada por (1.4.14).

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\alpha, h)}{\partial \alpha_j} &= \sum_{k=1}^N \left[ (1 - \delta_k) \frac{1}{\Phi(-\mathbf{x}_k \cdot \alpha)} \phi(-\mathbf{x}_k \cdot \alpha) (-x_{kj}) + \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) x_{kj} \right] \\ &= \sum_{k=1}^N \left[ -(1 - \delta_k) \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} x_{kj} + \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) x_{kj} \right] \\ &= - \sum_{k=1}^N \left[ (1 - \delta_k) \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} x_{kj} \right] + \sum_{k=1}^N \left[ \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) x_{kj} \right] = 0 \end{aligned} \quad (1.4.19)$$

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\alpha, h)}{\partial h} &= \sum_{k=1}^N \left[ 0(1 - \delta_k) + \delta_k \frac{1}{h} + \frac{\delta_k}{2} 2(hy_k - \mathbf{x}_k \cdot \alpha) y_k \right] \\ &= \sum_{k=1}^N \left[ \frac{\delta_k}{h} - \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) y_k \right] \\ &= \frac{\tilde{c}}{h} - \sum_{k=1}^N [\delta_k (hy_k - \mathbf{x}_k \cdot \alpha) y_k] = 0 \end{aligned} \quad (1.4.20)$$

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}(\alpha, h)}{\partial \alpha_i \partial \alpha_j} &= \frac{\partial}{\partial \alpha_i} \left[ - \sum_{k=1}^N \left[ (1 - \delta_k) \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} x_{kj} \right] + \sum_{k=1}^N [\delta_k (hy_k - \mathbf{x}_k \cdot \alpha) x_{kj}] \right] \\ &= - \sum_{k=1}^N \left[ (1 - \delta_k) x_{kj} \frac{\partial}{\partial \alpha_i} \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} \right] + \sum_{k=1}^N \left[ \delta_k x_{kj} \frac{\partial}{\partial \alpha_i} (hy_k - \mathbf{x}_k \cdot \alpha) \right] \\ &= \sum_{k=1}^N \left[ (1 - \delta_k) \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} \left( \mathbf{x}_k \cdot \alpha - \frac{\phi(-\mathbf{x}_k \cdot \alpha)}{\Phi(-\mathbf{x}_k \cdot \alpha)} \right) x_{kj} x_{ki} \right] \\ &\quad - \sum_{k=1}^N [\delta_k x_{kj} x_{ki}] \end{aligned} \quad (1.4.21)$$

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}(\alpha, h)}{\partial \alpha_i \partial h} &= \frac{\partial}{\partial \alpha_i} \left[ \frac{\tilde{c}}{h} - \sum_{k=1}^N \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) y_k \right] \\ &= 0 - \frac{\partial}{\partial \alpha_i} \left[ \sum_{k=1}^N \delta_k (hy_k - \mathbf{x}_k \cdot \alpha) y_k \right] \end{aligned}$$

$$= \sum_{k=1}^N [\delta_k y_k x_{ki}] \quad (1.4.22)$$

$$\begin{aligned} \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\alpha}, h)}{\partial h \partial h} &= \frac{\partial}{\partial h} \left[ \frac{\tilde{c}}{h} - \sum_{k=1}^N \delta_k (h y_k - \mathbf{x}_k \boldsymbol{\alpha}) y_k \right] \\ &= -\frac{\tilde{c}}{h^2} - \sum_{k=1}^N \left[ \delta_k y_k \frac{\partial}{\partial h} (h y_k - \mathbf{x}_k \boldsymbol{\alpha}) \right] \\ &= -\frac{\tilde{c}}{h^2} - \sum_{k=1}^N [\delta_k y_k^2] \end{aligned} \quad (1.4.23)$$

Para la estimación de los parámetros se procede usando las primeras derivadas con respecto a  $\boldsymbol{\beta}$  y  $\sigma^2$  (ecuaciones (1.4.16) y (1.4.17)) o más simple con respecto de  $\boldsymbol{\alpha}$  y  $h$  (ecuaciones (1.4.19) y (1.4.20)). Independientemente de la elección, las ecuaciones anteriores no pueden ser resueltas analíticamente y se deben emplear métodos numéricos. En párrafos anteriores se comentó que Newton-Raphson era una elección adecuada.

Para la maximización de la función expresada por la ecuación (1.4.14) o (1.4.18) existen rutinas de programación que se encuentran en algunos paquetes del software estadístico R. Especialmente se tienen los paquetes **AER** diseñado por Kleiber & Zeileis (2008), **survival** diseñado por Therneau & Grambsch (2000); Therneau (2015) y **censReg** diseñado por Henningsen (2016). En general estas rutinas hacen uso de la función **maxlik** del paquete **maxlik** diseñado por Henningsen & Toomet (2011) que está especialmente programado para la maximización de funciones de verosimilitud usando el método de *Newton-Raphson*.

#### 1.4.2.2. Estimación de los Parámetros en el Modelo Tobit con Pesos Muestrales

La inclusión de los pesos muestrales en el proceso de estimación de los parámetros del modelo tiene como propósito dar estimadores de los parámetros del modelo en la población. Chambers & Skinner (2003, p. 22-23) denominan a este tipo de estimación como estimación analítica, distinguiéndola de la estimación descriptiva encargada de hacer inferencia de parámetros descriptivos de la población como el total o el promedio de una variable. El proceso de estimación se conoce como *pseudoverosimilitud* y no es otra cosa que la inclusión de pesos muestrales en la función de verosimilitud que permite la estimación de los parámetros del modelo Rondón et al. (2012) utilizan este método para hacer la estimación de los parámetros del modelo que asiste al estimador GREG <sup>9</sup>.

<sup>9</sup>Chambers & Skinner (2003) y Särndal et al. (1992, p. 517) abordan en más detalle lo relacionado con la pseudoverosimilitud y los contextos en los que este método de inferencia puede ser usado.

Usando la expresión (1.4.14) e incluyendo los pesos muestrales se tiene

$$\begin{aligned} \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\beta}, \sigma^2) = & \sum_{k=1}^n \left[ \frac{(1 - \delta_k)}{\pi_k} \ln \Phi \left( \frac{-\mathbf{x}_k \cdot \boldsymbol{\beta}}{\sigma} \right) - \frac{\delta_k}{2\pi_k} \ln \sigma^2 \right. \\ & \left. - \frac{\delta_k}{2\pi_k \sigma^2} (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta})^2 \right] \end{aligned} \quad (1.4.24)$$

donde el superíndice  $\pi$  representa la inclusión de los pesos muestrales en la función de log-verosimilitud. La expresión (1.4.24) no es otra cosa que el  $\pi$ -estimador del logaritmo de la función de verosimilitud dado por la ecuación (1.4.14). Las expresiones para las primeras y segundas derivadas no presentan mayores modificaciones

$$\begin{aligned} \frac{\partial \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\beta}, \sigma^2)}{\partial \beta_j} = & -\frac{1}{\sigma} \sum_{k=1}^n \frac{(1 - \delta_k)}{\pi_k} \left( \frac{\phi(z)}{\Phi(z)} \right) x_{kj} \\ & + \frac{1}{\sigma^2} \sum_{k=1}^n \frac{\delta_k}{\pi_k} (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta}) x_{kj}, \end{aligned} \quad (1.4.25)$$

$$\begin{aligned} \frac{\partial \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} = & \frac{1}{2\sigma^3} \sum_{k=1}^n \frac{(1 - \delta_k)}{\pi_k} \left( \frac{\phi(z)}{\Phi(z)} \right) \mathbf{x}_k \cdot \boldsymbol{\beta} - \frac{\tilde{c}}{2\sigma^2} \sum_{k=1}^n \frac{1}{\pi_k} \\ & + \frac{1}{2\sigma^4} \sum_{k=1}^n \frac{\delta_k}{\pi_k} (y_k - \mathbf{x}_k \cdot \boldsymbol{\beta})^2. \end{aligned} \quad (1.4.26)$$

De manera similar se tiene el  $\pi$ -estimador del logaritmo de la función de versomilitud dada por (1.4.18)

$$\ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h) = \sum_{k=1}^n \left[ \frac{(1 - \delta_k)}{\pi_k} \ln \Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha}) + \frac{\delta_k}{\pi_k} \ln h - \frac{\delta_k}{2\pi_k} (hy_k - \mathbf{x}_k \cdot \boldsymbol{\alpha})^2 \right] \quad (1.4.27)$$

y las primeras y segundas derivadas para la expresión (1.4.27) estarán dadas por

$$\begin{aligned} \frac{\partial \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h)}{\partial \alpha_j} = & -\sum_{k=1}^n \left[ \frac{(1 - \delta_k)}{\pi_k} \frac{\phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})}{\Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})} x_{kj} \right] \\ & + \sum_{k=1}^n \left[ \frac{\delta_k}{\pi_k} (hy_k - \mathbf{x}_k \cdot \boldsymbol{\alpha}) x_{kj} \right], \end{aligned} \quad (1.4.28)$$

$$\frac{\partial \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h)}{\partial h} = \frac{\tilde{c}}{h} \sum_{k=1}^n \frac{1}{\pi_k} - \sum_{k=1}^n \left[ \frac{\delta_k}{\pi_k} (hy_k - \mathbf{x}_k \cdot \boldsymbol{\alpha}) y_k \right], \quad (1.4.29)$$

$$\begin{aligned} \frac{\partial^2 \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h)}{\partial \alpha_i \partial \alpha_j} = & \sum_{k=1}^n \left[ \frac{(1 - \delta_k)}{\pi_k} \frac{\phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})}{\Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})} \left( \mathbf{x}_k \cdot \boldsymbol{\alpha} - \frac{\phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})}{\Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})} \right) x_{kj} x_{ki} \right] \\ & - \sum_{k=1}^n \left[ \frac{\delta_k}{\pi_k} x_{kj} x_{ki} \right] \end{aligned} \quad (1.4.30)$$

$$\frac{\partial^2 \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h)}{\partial \alpha_i \partial h} = \sum_{k=1}^n \left[ \frac{\delta_k}{\pi_k} y_k x_{ki} \right], \quad (1.4.31)$$

y

$$\frac{\partial^2 \ln \hat{\mathcal{L}}_s^\pi(\boldsymbol{\alpha}, h)}{\partial h \partial h} = -\frac{\tilde{c}}{h^2} \sum_{k=1}^n \frac{1}{\pi_k} - \sum_{k=1}^n \left[ \frac{\delta_k}{\pi_k} y_k^2 \right]. \quad (1.4.32)$$

### 1.4.2.3. Estimación de la Matriz de Varianzas y Covarianzas

Amemiya (1973) prueba que los estimadores máximo verosímiles son fuertemente consistentes y asintóticamente normales con matriz asintótica de varianzas-covarianzas igual a

$$-\left( \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'} \right) \quad (1.4.33)$$

donde  $\theta = (\boldsymbol{\beta}', \sigma^2)$  y  $\ln \mathcal{L}$  está dado por las expresiones (1.4.14) o (1.4.18). Amemiya (1984, pg. 17) sugiere como estimador de (1.4.33) el uso de la inversa de la *Matriz de Información de Fisher* o MIF dada por

$$-\left( \mathbf{E} \left[ \frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'} \right] \right)^{-1}. \quad (1.4.34)$$

En adelante se denotará la MIF con  $\mathcal{F}$  y a su inversa como  $\mathcal{F}^{-1}$ . Para el cálculo de  $\mathcal{F}$  es necesario el uso de las siguientes expresiones (ver (Arellano-Valle et al., 2012, p. 455))

$$\mathbf{E}[\delta_k] = \Pr(y_k > 0) = \Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha}) \quad (1.4.35)$$

$$\mathbf{E}[\delta_k y_k] = \mathbf{E}[\delta_k] \mathbf{E}[y_k | y_k > 0] = \frac{1}{h} ((-\mathbf{x}_k \cdot \boldsymbol{\alpha}) \Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha}) + \phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})) \quad (1.4.36)$$

$$\mathbf{E}[\delta_k y_k^2] = \frac{1}{h^2} ((1 + -\mathbf{x}_k \cdot \boldsymbol{\alpha}) \Phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha}) + (-\mathbf{x}_k \cdot \boldsymbol{\alpha})^2 \phi(-\mathbf{x}_k \cdot \boldsymbol{\alpha})) \quad (1.4.37)$$

Partiendo de (1.4.18) y de las expresiones de su segunda derivada y haciendo  $q_k = \mathbf{x}'_k \boldsymbol{\alpha}$  se tendrá entonces que los elementos de  $\mathcal{F}$  vienen dados por

$$\begin{aligned} -\mathbf{E} \left[ \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\alpha}, h)}{\partial \alpha_i \partial \alpha_j} \right] &= -\mathbf{E} \left[ \sum_{k=1}^N (1 - \delta_k) \frac{\phi(-q_k)}{\Phi(-q_k)} \left( q_k - \frac{\phi(-q_k)}{\Phi(-q_k)} \right) x_{kj} x_{ki} \right. \\ &\quad \left. - \sum_{k=1}^N \delta_k x_{kj} x_{ki} \right] \\ &= -\left[ \mathbf{E} \left[ \sum_{k=1}^N (1 - \delta_k) \frac{\phi(-q_k)}{\Phi(-q_k)} \left( q_k - \frac{\phi(-q_k)}{\Phi(-q_k)} \right) x_{kj} x_{ki} \right] \right. \\ &\quad \left. - \mathbf{E} \left[ \sum_{k=1}^N \delta_k x_{kj} x_{ki} \right] \right] \\ &= -\left[ \sum_{k=1}^N \mathbf{E}[(1 - \delta_k)] \frac{\phi(-q_k)}{\Phi(-q_k)} \left( q_k - \frac{\phi(-q_k)}{\Phi(-q_k)} \right) x_{kj} x_{ki} \right] \end{aligned}$$

$$\begin{aligned}
& - \sum_{k=1}^N \mathbf{E}[\delta_k] x_{kj} x_{ki} \Big] \\
&= - \left[ \sum_{k=1}^N (1 - \Phi(-q_k)) \frac{\phi(-q_k)}{\Phi(-q_k)} \left( q_k - \frac{\phi(-q_k)}{\Phi(-q_k)} \right) x_{kj} x_{ki} \right. \\
&\quad \left. - \sum_{k=1}^N \Phi(-q_k) x_{kj} x_{ki} \right] \\
&= - \sum_{k=1}^N \left[ \left( (1 - \Phi(-q_k)) \frac{\phi(-q_k)}{\Phi(-q_k)} \left( q_k - \frac{\phi(-q_k)}{\Phi(-q_k)} \right) \right. \right. \\
&\quad \left. \left. - \Phi(-q_k) \right) x_{kj} x_{ki}, \right. \tag{1.4.38}
\end{aligned}$$

$$\begin{aligned}
- \mathbf{E} \left[ \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\alpha}, \mathbf{h})}{\partial \alpha_i \partial h} \right] &= - \mathbf{E} \left[ \sum_{k=1}^N \delta_k y_k x_{ki} \right] \\
&= - \sum_{k=1}^N \mathbf{E}[\delta_k y_k] x_{ki} \\
&= - \sum_{k=1}^N \frac{1}{h} \left( (-q_k) \Phi(-q_k) + \phi(-q_k) \right) x_{ki} \tag{1.4.39}
\end{aligned}$$

y

$$\begin{aligned}
- \mathbf{E} \left[ \frac{\partial^2 \ln \mathcal{L}(\boldsymbol{\alpha}, \mathbf{h})}{\partial h^2} \right] &= - \mathbf{E} \left[ -\frac{\tilde{c}}{h^2} - \sum_{k=1}^N \delta_k y_k^2 \right] \\
&= - \mathbf{E} \left[ \sum_{k=1}^N -\frac{\delta_k}{h^2} - \delta_k y_k^2 \right] \\
&= - \sum_{k=1}^N -\frac{1}{h^2} \mathbf{E}[\delta_k] - \mathbf{E}[\delta_k y_k^2] \\
&= - \sum_{k=1}^N -\frac{1}{h^2} \Phi(-q_k) - \frac{1}{h^2} \left( (1 + (-q_k)) \Phi(-q_k) + (-q_k)^2 \phi(-q_k) \right) \\
&= - \sum_{k=1}^N -\frac{1}{h^2} \left( \Phi(-q_k) + (1 + (-q_k)) \Phi(-q_k) + (-q_k)^2 \phi(-q_k) \right) \\
&= - \sum_{k=1}^N -\frac{1}{h^2} \left( 2\Phi(-q_k) + (-q_k) \Phi(-q_k) + (-q_k)^2 \phi(-q_k) \right). \tag{1.4.40}
\end{aligned}$$



# Capítulo 2

## DISEÑO DE ESTIMADORES PARA EL TOTAL POBLACIONAL

### 2.1. Problema de Estimación

Las muestras censuradas son de frecuente aparición en muchos contextos disciplinares como la econometría, la ecología, la biología, etc. En estos contextos la presencia de observaciones censuradas en las muestras puede ser ocasionada por problemas con los instrumentos de medición o por problemas en el diseño metodológico y de recolección de la información. Los siguientes ejemplos ilustrarán estas situaciones.

#### *2.1 Ejemplo.*

Medición de sustancias o compuestos químicos para los cuales las máquinas encargadas de su medición tienen un límite de detección por debajo del cual no es posible detectar la cantidad del compuesto y éste se registra con dicho valor límite. Ver (Helsel, 2012, p. 1).

#### *2.2 Ejemplo.*

En evaluaciones medioambientales, las medidas de las cantidades de contaminantes en ocasiones son reportadas por los laboratorios como “no detectado” o “rastros”, en cuyo caso los datos pueden ser censurados a izquierda o en intervalo respectivamente. Muestras debajo del límite de “no detección” tienen niveles de contaminante que se consideran por debajo del límite de detección y muestras con “rastros” tienen niveles que están por encima del límite de detección pero debajo del límite de cuantificación. Thompson & Nelson (2003, p. 224) presenta como ejemplo de esta situación un estudio que buscaba la exposición a pesticidas en niños mayores de 6 años.

#### *2.3 Ejemplo.*

Encuestas en las que se preguntan por los ingresos de una persona en un determinado periodo. Se puede tener situaciones en las que ingresos por debajo de un salario mínimo se registren en una categoría *un salario mínimo o menos* o por el contrario,

ingresos por encima de cierto tope se registren en categorías del estilo *10 salarios mínimos o más*.

#### 2.4 Ejemplo.

Un caso interesante de censura se debe a la capacidad de almacenamiento de datos. Hacia la década de los setenta la Oficina de Censos de los Estados Unidos trunca los datos a cinco dígitos particularmente para los datos de ingresos salariales, ingresos independientes e ingresos agrícolas. En la actualidad la capacidad de almacenamiento no es un problema, pero la Oficina de Censos sigue censurado la información de manera interna, en parte debido a las preocupaciones sobre la fiabilidad de los datos de las personas que reportan un valor de ingreso extremadamente alto y en parte debido a la confidencialidad. Burkhauser, Feng, Jenkins & Larrimore (2011, p. 8).

En los ejemplos anteriores la solución óptima consiste en evitar la presencia de observaciones censuradas. Para el caso del ejemplo 2.3 se podría evitar las observaciones censuradas desde la etapa de diseño y planeación de la encuesta. Corregir de antemano, eliminando preguntas o categorías de respuesta que generen observaciones censuradas o capacitando mejor a los encuestadores para que hagan un registro más preciso de la información. En el caso del ejemplo 2.1 la situación no es tan sencilla de solucionar. Se podría pensar en usar nuevos instrumentos de medición o técnicas diferentes para la recolección de la información como forma de evitar las observaciones censuradas.

Sin embargo no siempre es posible controlar la aparición de observaciones censuradas en las muestras como en los casos anteriores. Factores económicos, de organización o procedimiento pueden hacer imposible esta corrección previa. Por tanto es necesario considerar la censura como un fenómeno a tener en cuenta en el proceso de estimación de un parámetro poblacional. Para ilustrar la situación anterior veamos el siguiente ejemplo.

*2.5 Ejemplo.* Supóngase que se tiene una muestra  $s$  conformada por  $n = 10$  elementos extraídos de manera aleatoria acorde con un muestreo aleatorio simple sin reemplazamiento de una población de tamaño  $N = 100$ . En la Tabla 2.1 se registran los elementos de la muestra. En la primera fila se encuentran las observaciones no censuradas. En la segunda fila las mismas observaciones pero censuradas a izquierda en un punto límite de censura  $T = 0$ . Al final de las filas se encuentran las esti-

---

No censura	0.41	-0.11	0.23	-0.03	-1.01	-0.68	-0.72	1.01	1.93	0.37	$\hat{t}_y^* = 14$
Censura	0.41	0	0.23	0	0	0	0	1.01	1.93	0.37	$\hat{t}_y = 39.5$

---

TABLA 2.1. Ejemplo de estimación con datos censurados y no censurados.

maciones hechas mediante el estimador de *Horvitz-Thompson* del total de las dos variables. Dado que se conoce el verdadero valor del parámetro,  $t = 16.4$ , se sabe que la estimación  $\hat{t}_y^*$  es la más cercana al verdadero valor poblacional. Sin embargo, en la práctica solamente se cuenta con la información contenida en la segunda fila

esto conlleva a una mala estimación del parámetro dado que sólo se cuenta con la información de la variable para observaciones por encima del límite de censura.

La situación expresada en el anterior ejemplo muestra la dificultad que supone la realización de estimaciones en muestras con observaciones censuradas. Esta dificultad radica en el hecho de que el interés real está en estimar el total de la variable latente para lo cual sólo se dispone de la información de la variable censurada. Esto es, deseamos hacer inferencia sobre la variable latente o no observada contando únicamente con la información de la variable censurada u observada. Lo anterior lleva, naturalmente, a plantear la siguiente pregunta: *¿Cómo hacer estimaciones del total de una variable de interés cuando se sabe que la información con la que se cuenta en la muestra presenta observaciones censuradas?*

La respuesta a este interrogante estará mediada por la intervención del modelo de regresión Tobit y el diseño de estimadores especiales para muestras con observaciones censuradas.

## 2.2. Estimadores para el Total

En la sección anterior se describió el problema de estimación del total en muestras con observaciones censuradas. En lo que sigue se presentarán varias opciones de estimadores como posibles soluciones al problema de estimación. Se considerarán los siguientes grupos de estimadores:

- 1 Estimadores para el total de la variable censurada.
- 2 Estimadores para el total de la pseudovariante mixta.
- 3 Estimadores para el total de la pseudovariante latente.

En el primer grupo de estimadores se usa la variable censurada  $y_k$  definida por (1.4.8) como aproximación de la variable latente  $y_k^*$ . Dentro de este primer grupo se consideran los siguientes estimadores.

- Estimador de Horvitz-Thompson (H-T) o  $\pi$ -estimador.
- Estimador de regresión o estimador GREG.
- Estimador de regresión Tobit.

En el segundo grupo de estimadores se usará una pseudovariante  $y'_k$  como aproximación de  $y_k^*$ . La pseudovariante  $y'_k$  será una mezcla de valores por encima del límite de censura y de estimaciones de los valores censurados hechas a través del modelo Tobit. En este grupo se considera el siguiente estimador:

- Estimador mixto.

Para el tercer grupo de estimadores se usará una pseudovariante latente  $y_k^{*'}$  como aproximación de  $y_k^*$ . La pseudovariante  $y_k^{*'}$  estará completamente conformada por estimaciones hechas a través del modelo Tobit tanto de los valores censurados como de los no censurados. En este grupo se consideran los siguientes estimadores:

- Estimador Tobit sintético.
- $\pi$ -estimador Tobit sintético.

Para los estimadores mencionados anteriormente se consideraran algunos supuestos o condiciones generales bajo las cuales se asume el diseño y el comportamiento de los mismos. En general, se tienen las siguientes condiciones

C.1  $y^*$  (la variable latente) se distribuye  $N(\mu, \sigma^2)$ .

C.2 Existe una matriz  $\mathbb{X}$  de información auxiliar disponible para toda la población  $\mathbb{U}$ .

C.3  $y$  es una variable observada a través de la muestra  $s$  y es simplemente censurada a izquierda en el punto límite de censura  $T$ .

C.4  $T = 0$ .

La primera condición será requisito indispensable dado que, sobre esta condición es que se deriva todo el análisis de regresión Tobit. La condición C.2 es una condición deseable y será indispensable para algunos estimadores, pero su negación dará lugar al planteamiento de otro tipo de estimadores. La condición C.3 es el caso de estudio para el presente trabajo de tesis. Otros casos, como censura múltiple o progresiva exceden los alcances del presente trabajo. El caso de censura por derecha no representa mayores complicaciones o teoría adicional a la presentada en la sección 1.4. Basta con hacer uso de la definición de *censura a derecha* dada en (1.4.2) en las ecuaciones (1.4.4) y (1.4.7) y consecuentemente redefinir las expresiones derivadas de estas. La condición C.4 es el caso más común en la literatura relacionada. Cualquier otro límite de censura puede ser considerado y llevado al caso general de censura en cero. Basta con restar el valor del límite de censura a cada uno de los valores de la variable censurada y sobre esta variable trabajar. El proceso puede ser invertido sumando el valor del punto de censura, con lo que se restaura la dimensión original de los datos.

Por último, para algunos estimadores se considerará la estimación de los parámetros del modelo de regresión haciendo uso de los pesos muestrales y sin estos. Esto se verá en la sección de la simulación y tiene implicaciones serias en la estimación del parámetro.

### 2.2.1. El Total de la Variable Latente y el Total de la Variable Censurada

En la sección 2.1 se hace evidente la diferencia que existe entre el total asociado a la variable latente y el total asociado a la variable censurada. Las siguientes proposiciones precisaran esta diferencia.

**2.1 Proposición.** *Sea  $t_{y^*} = \sum_{\cup} y_k^*$  y  $t_y = \sum_{\cup} y_k$  los totales de las variables latente y censurada respectivamente. Se tiene que*

$$t_{y^*} < t_y \quad (2.2.1)$$

*siempre que  $T$  esté definido y  $y_k$  sea una variable simplemente censurada a izquierda en el punto  $T$ .*

*Demostración.* Dado el límite de censura  $T$  se puede descomponer

$$t_{y^*} = \sum_{y_k^* \leq T} y_k^* + \sum_{y_k^* > T} y_k^* \quad (2.2.2)$$

como la suma de los valores de  $y_k^* \leq T$  más los valores  $y_k^* > T$ . Por la ecuación (1.4.3) y consecuentemente con la descomposición anterior

$$t_y = \sum_{y_k \leq T} T + \sum_{y_k > T} y_k. \quad (2.2.3)$$

De la ecuación (1.4.3) se tiene que si  $y_k^* > T$  entonces  $y_k = y_k^*$  por tanto

$$\sum_{y_k^* > T} y_k^* = \sum_{y_k > T} y_k. \quad (2.2.4)$$

De forma semejante se tiene que si  $y_k^* \leq T$  entonces  $y_k = T$  de donde se deduce que

$$\sum_{y_k^* \leq T} y_k^* < \sum_{y_k \leq T} T \quad (2.2.5)$$

y por tanto

$$t_{y^*} < t_y. \quad (2.2.6)$$

□

Siguiendo el mismo argumento usado para deducir (2.2.1) se puede deducir

**2.2 Proposición.** *Sea  $t_{y^*} = \sum_{\cup} y_k^*$  y  $t_y = \sum_{\cup} y_k$  los totales de las variables latente y censurada respectivamente. Se tiene que*

$$t_{y^*} > t_y \quad (2.2.7)$$

*siempre que  $T$  esté definido y  $y_k$  sea una variable simplemente censurada a derecha en el punto  $T$ .*

## 2.2.2. Estimadores para el Total de la Pseudovariable Latente

Los estimadores considerados en este grupo tienen en común el uso de una *pseudovariable latente* construida a partir del modelo de regresión Tobit. Estos estimadores serán considerados como apropiados en la tarea de la estimación del total poblacional en muestras con observaciones censuradas. En secciones posteriores, primero bajo simulación y luego de forma analítica, se deducirán sus propiedades y se mostrará la razón por la cual estos estimadores se consideran adecuados. En lo que sigue se expondrá la deducción de estos estimadores.

### 2.2.2.1. Estimador Tobit Sintético

Sea  $t_{y^*} = \sum_{\mathbb{U}} y^*$  el parámetro que se desea estimar. Bajo las condiciones C.1 a C.4 y considerando el modelo  $\psi$  dado por (1.4.10) para la estimación de los parámetros  $\beta$  se puede definir la *pseudovariable latente*

$$y_k^* = \mathbf{x}_k \cdot \hat{\beta}. \quad (2.2.8)$$

Siguiendo lo expuesto en la sección 1.4.2.1 sobre la estimación de los parámetros de un modelo Tobit ( $\psi$ ), se tendrá que el vector de parámetros  $\hat{\beta}$  estará dado por la maximización del logaritmo de la función de verosimilitud o por la maximización del logaritmo de la función de pseudo-verosimilitud, es decir  $\hat{\beta}$  estará dado por

$$\hat{\beta}_s = \arg \max(\log \mathcal{L}_s(\beta)) \quad (2.2.9)$$

$$\hat{\beta}_s^\pi = \arg \max(\log \mathcal{L}_s^\pi(\beta)) \quad (2.2.10)$$

En (2.2.9),  $\beta$  es el argumento que maximiza el logaritmo de la función de verosimilitud dada por (1.4.14) o por (1.4.18). En (2.2.10),  $\beta$  es el argumento que maximiza (1.4.24) o (1.4.27).

*2.1 Nota.* De acuerdo a la notación dada por Rondón et al. (2012) en adelante se usará el superíndice  $\pi$  cuando se haga una estimación que involucre pesos muestrales. Se notará con  $\hat{\beta}_s^\pi$  a los estimadores de los parámetros que incluyan los pesos muestrales y con  $\hat{\beta}_s$  a los estimadores que no los incluyan.

A partir de esta pseudovariable se puede definir el estimador

$$\begin{aligned} \hat{t}_{sin} &= \sum_{\mathbb{U}} \mathbf{x}_k \cdot \hat{\beta} \\ &= \sum_{\mathbb{U}} y_k^* \end{aligned} \quad (2.2.11)$$

como estimador del total de la variable latente  $y^*$ . Si en (2.2.11) se usa (2.2.9) para calcular  $\hat{\beta}$ , el estimador toma la forma

$$\begin{aligned} \hat{t}_{sin} &= \sum_{\mathbb{U}} \mathbf{x}_k \cdot \hat{\beta}_s \\ &= \sum_{\mathbb{U}} y_k^*. \end{aligned} \quad (2.2.12)$$

Si por el contrario se usa (2.2.10), el estimador toma la forma

$$\begin{aligned}\hat{t}_{sin}^{\pi} &= \sum_{\mathbb{U}} \mathbf{x}_k \cdot \hat{\beta}_s^{\pi} \\ &= \sum_{\mathbb{U}} y_k^{\pi *'}.\end{aligned}\quad (2.2.13)$$

Tanto (2.2.12) como (2.2.13) se llamarán estimadores *Tobit Sintético* con o sin pesos muestrales. El nombre particular de estos estimadores se debe por un lado a la semejanza que tienen con los estimadores sintéticos que menciona Särndal et al. (1992) cuando habla de los estimadores de regresión aplicados a la estimación de dominios y por otro, al uso del modelo Tobit dentro del estimador.

### 2.2.2.2. $\pi$ -Estimador Tobit Sintético

Los estimadores (2.2.12) y (2.2.13) son la suma sobre toda la población de valores estimados a través del modelo Tobit. Para esto, es necesario hacer uso de la información auxiliar que se encuentra disponible para todos los elementos de la población. Sin embargo ésta no es una situación común, en la mayoría de los casos la información auxiliar estará disponible únicamente para los elementos contenidos en la muestra  $s$ , razón por la cual los estimadores *Tobit Sintético* dados por (2.2.12) y (2.2.13) no pueden ser usados. En tal circunstancia se puede proponer un estimador usando el principio de  $\pi$ -expansión como forma de alcanzar el nivel poblacional a partir de los datos contenidos en la muestra.

Sea  $t_{y^*} = \sum_{\mathbb{U}} y^*$  el parámetro a estimar. Bajo las condiciones C.1, C.3 y C.4 y considerando el modelo  $\psi$  dado por (1.4.10) para la estimación de los parámetros  $\beta$  se puede definir

$$\begin{aligned}\hat{t}_{\pi TS} &= \sum_s \frac{\mathbf{x}_k \cdot \hat{\beta}}{\pi_k} \\ &= \sum_s \frac{y_k^{*'}}{\pi_k}\end{aligned}\quad (2.2.14)$$

como el estimador de  $t_{y^*}$ . Usando (2.2.9) para calcular  $\hat{\beta}$  el estimador (2.2.14) se expresa como

$$\begin{aligned}\hat{t}_{\pi TS} &= \sum_s \frac{\mathbf{x}_k \cdot \hat{\beta}}{\pi_k} \\ &= \sum_s \frac{y_k^{*'}}{\pi_k}.\end{aligned}\quad (2.2.15)$$

Por el contrario, si se usa (2.2.10) para calcular  $\hat{\beta}$ , el estimador se puede expresar como

$$\hat{t}_{\pi TS}^{\pi} = \sum_s \frac{\mathbf{x}_k \cdot \hat{\beta}_s^{\pi}}{\pi_k}$$

$$= \sum_s \frac{y_k^{\pi_k^*}}{\pi_k}. \quad (2.2.16)$$

Los estimadores dados por (2.2.15) y (2.2.16) se llamarán  $\pi$ -Estimador Tobit Sintéticos con o sin pesos muestrales. Su nombre deriva de los estimadores *Tobit Sintéticos* definidos anteriormente y del hecho de que usan el principio de  $\pi$ -expansión como forma de alcanzar el nivel poblacional, lo que los convierte en  $\pi$  estimadores del total.

### 2.2.3. Estimadores para el Total de la Variable Censurada y la Pseudovariable Mixta

Los estimadores considerados en esta sección no son adecuados para el problema de estimación del total poblacional de una variable en muestras con observaciones censuradas. A pesar de esto, se considera que su exposición es importante ya que: primero, los estimadores que se expondrán son considerados como aproximaciones naturales a la solución del problema y segundo, estos estimadores permiten analizar el fenómeno de la censura y el comportamiento de ésta en la estimación del total poblacional. En secciones posteriores, especialmente bajo simulación, se darán argumentos del porque estos estimadores no son considerados como soluciones al problema de estimación.

#### 2.2.3.1. $\pi$ -Estimador o Estimador de Horvitz-Thompson (H-T)

Una aproximación natural a la solución del problema es hacer inferencia del total de la variable latente usando la variable censurada. Bajo las condiciones C.2 a C.4, se puede definir el estimador de Horvitz-Thompson o  $\pi$ -estimador dado por la ecuación (1.3.2) sobre la variable censurada  $y_k$ . Sobre esta variable el parámetro que se estima es

$$t_y = \sum_{\mathbb{U}} y_k \quad (2.2.17)$$

y su correspondiente  $\pi$ -estimador es

$$\hat{t}_{\pi y} = \sum_s \check{y}_k = \sum_s \frac{y_k}{\pi_k} \quad (2.2.18)$$

donde  $\pi_k$  son las probabilidades de inclusión de primer orden definidas en (1.1.5). Sin embargo, el total dado por (2.2.17) no es el parámetro buscado, de hecho (2.2.17) siempre será mayor que  $t_{y^*} = \sum_{\mathbb{U}} y^*$  como se demostró en (2.2.1).

Como se verá más adelante, el  $\pi$  - estimador definido por (2.2.18) será un buen estimador para el total de la variable censurada y no de la variable latente. Por otro lado, se puede interpretar esta situación de forma distinta y ver al  $\pi$ -estimador como un estimador de  $y_k^*$  que sobreestima el valor de su total. Esta interpretación permite dar cabida al estimador que se mostrará a continuación.

### 2.2.3.2. Estimador de Regresión o Estimador GREG

Bajo las condiciones C.1 a C.4, se puede aplicar el estimador de regresión dado por la definición 1.3.8 sobre la variable censurada  $y_k$ . El uso de este estimador como una solución natural para el problema de estimación se sustenta bajo el hecho de que se dispone de información auxiliar (condición C.2) y sobre la interpretación que se puede dar al estimador de regresión, por la cual el estimador de regresión es el  $\pi$ -estimador del total más un término de ajuste negativamente correlacionado (Särndal et al., 1992, p. 230). Es decir, se puede pensar que el estimador de regresión ajusta la estimación del total corrigiendo la sobreestimación que hace el  $\pi$ -estimador dado por (2.2.18) cuando se usa la variable censurada.

Usando (1.3.14) como la forma del estimador de regresión, se tiene que

$$\hat{t}_{yr} = \sum_{\mathbb{U}} \hat{y}_k + \sum_s \frac{y_k - \hat{y}_k}{\pi_k} \quad (2.2.19)$$

con  $\hat{y}_k = \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}$  como el estimador de regresión aplicado sobre la variable  $y_k$ . Como se mencionó anteriormente se consideran dos opciones en el cálculo del vector  $\hat{\boldsymbol{\beta}}$ . La primera opción está dada por

$$\begin{aligned} \hat{\boldsymbol{\beta}}_s &= (\mathbb{X}'_s \mathbb{X}_s)^{-1} \mathbb{X}_s y_s \\ &= \left( \sum_s \mathbf{x}_k \cdot \mathbf{x}_k \cdot \right)^{-1} \left( \sum_s \mathbf{x}_k \cdot y_k \right), \end{aligned} \quad (2.2.20)$$

que es la estimación de mínimos cuadrados ordinarios. La segunda opción será por mínimos cuadrados ponderados

$$\begin{aligned} \hat{\boldsymbol{\beta}}_s^\pi &= (\mathbb{X}'_s \Pi_s^{-1} \mathbb{X}_s)^{-1} \mathbb{X}_s \Pi_s^{-1} y_s \\ &= \left( \sum_s \frac{\mathbf{x}_k \cdot \mathbf{x}_k \cdot}{\sigma^2 \pi_k} \right)^{-1} \left( \sum_s \frac{\mathbf{x}_k \cdot y_k}{\sigma^2 \pi_k} \right). \end{aligned} \quad (2.2.21)$$

Usando (2.2.20) el estimador dado por (2.2.19) puede escribirse como

$$\begin{aligned} \hat{t}_{yr} &= \sum_{\mathbb{U}} \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s + \sum_s \frac{y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s}{\pi_k} \\ &= \sum_{\mathbb{U}} \hat{y}_k + \sum_s \frac{y_k - \hat{y}_k}{\pi_k}, \end{aligned} \quad (2.2.22)$$

que es la versión del estimador de regresión sin pesos muestrales y usando (2.2.21) puede escribirse como

$$\begin{aligned} \hat{t}_{yr}^\pi &= \sum_{\mathbb{U}} \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s^\pi + \sum_s \frac{y_k - \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s^\pi}{\pi_k} \\ &= \sum_{\mathbb{U}} \hat{y}_k^\pi + \sum_s \frac{y_k - \hat{y}_k^\pi}{\pi_k}, \end{aligned} \quad (2.2.23)$$

que es el estimador de regresión con pesos muestrales o usual. El modelo  $\xi$  que asiste a los estimadores dados por las ecuaciones (2.2.22) y (2.2.23) se define como

$$\begin{cases} y_k = \mathbf{x}_{k \cdot} \boldsymbol{\beta} + \varepsilon_k \\ \text{Donde } \varepsilon_1, \dots, \varepsilon_N \text{ son variables aleatorias iid.} \\ \text{Con } \varepsilon_k \sim N(0, \sigma^2). \end{cases} \quad (2.2.24)$$

### 2.2.3.3. Estimador de Regresión Tobit

Bajo las condiciones C.1 a C.4 y haciendo uso del modelo  $\psi$  dado por (1.4.10) se tendrá un estimador de regresión asistido por el modelo Tobit. El estimador de regresión asistido por el modelo Tobit será

$$\hat{t}_{tob} = \sum_{\mathbb{U}} \hat{y}_k + \sum_s \frac{y_k - \hat{y}_k}{\pi_k} \quad (2.2.25)$$

donde  $\hat{y}_k = \mathbf{x}_{k \cdot} \hat{\boldsymbol{\beta}}$ . Si  $\hat{\boldsymbol{\beta}}$  esta dado por (2.2.9) el estimador  $\hat{t}_{tob}$  tendrá la forma

$$\begin{aligned} \hat{t}_{tob} &= \sum_{\mathbb{U}} \mathbf{x}_{k \cdot} \hat{\boldsymbol{\beta}}_s + \sum_s \frac{y_k - \mathbf{x}_{k \cdot} \hat{\boldsymbol{\beta}}_s}{\pi_k} \\ &= \sum_{\mathbb{U}} \hat{y}_k + \sum_s \frac{y_k - \hat{y}_k}{\pi_k}. \end{aligned} \quad (2.2.26)$$

Por otro lado, si se usa (2.2.10) el estimador  $\hat{t}_{tob}$  será

$$\begin{aligned} \hat{t}_{tob} &= \sum_{\mathbb{U}} \mathbf{x}_{k \cdot} \hat{\boldsymbol{\beta}}_s^\pi + \sum_s \frac{y_k - \mathbf{x}_{k \cdot} \hat{\boldsymbol{\beta}}_s^\pi}{\pi_k} \\ &= \sum_{\mathbb{U}} \hat{y}_k^\pi + \sum_s \frac{y_k - \hat{y}_k^\pi}{\pi_k}. \end{aligned} \quad (2.2.27)$$

### 2.2.3.4. Estimador Particionado

Sea  $t_{y^*} = \sum_{\mathbb{U}} y^*$  el total de la variable latente. Bajo las condiciones C.1 a C.3 y con  $T \neq 0$  se puede descomponer este parámetro de la siguiente manera

$$\begin{aligned} t_{y_k^*} &= \sum_{\mathbb{U}} y_k^* \\ &= \sum_{\mathbb{U}} \delta_k y_k^* + \sum_{\mathbb{U}} (1 - \delta_k) y_k^* \end{aligned} \quad (2.2.28)$$

donde  $\delta_k$  esta definido por la ecuación (1.4.9). El primer término del lado derecho de la ecuación (2.2.28) corresponde a la suma de los valores  $y_k^*$  cuando  $y_k^* > a$  ( $\delta_k = 1$ ) y el segundo término corresponde a la suma de los valores  $y_k^*$  cuando  $y_k^* \leq a$  ( $\delta_k = 0$ ). Es decir que la ecuación (2.2.28) representa una partición entre los valores que se encuentran por encima del límite de censura  $T = a$  más los valores que son iguales o inferiores a este límite. Aplicando este mismo razonamiento al total de la variable

censurada  $y_k$  se tiene que

$$\begin{aligned} t_{y_k} &= \sum_{\mathbb{U}} y_k \\ &= \sum_{\mathbb{U}} \delta_k y_k + \sum_{\mathbb{U}} (1 - \delta_k) T. \end{aligned} \quad (2.2.29)$$

Si para este estimador se asume la condición C.4 donde  $T = 0$  se tiene que

$$t_{y_k} = \sum_{\mathbb{U}} \delta_k y_k. \quad (2.2.30)$$

Usando el modelo de regresión Tobit podemos calcular valores  $\hat{y}_k$ , que son estimaciones de  $y_k$  a través del modelo. En particular podemos tener estimaciones  $\hat{y}_k$  para aquellas observaciones en donde  $y_k = 0$ , es decir cuando  $y_k^* \leq 0$ . Entonces, es razonable suponer que bajo el modelo Tobit se tiene que

$$\hat{y}_k = \hat{y}_k^* = \mathbf{x}_k \cdot \hat{\boldsymbol{\beta}} \quad \text{para } y_k^* \leq 0 \quad (2.2.31)$$

donde  $\hat{\boldsymbol{\beta}}$  está dado por (2.2.9) o (2.2.10). Se puede definir una variable  $y'_k$  tal que

$$y'_k = \begin{cases} y_k = y_k^* & \text{si } y_k^* > 0 \\ \hat{y}_k = \hat{y}_k^* & \text{si } y_k^* \leq 0. \end{cases} \quad (2.2.32)$$

Usando (2.2.28) y (2.2.32) se puede escribir  $t_{y'_k}$  como

$$t_{y'_k} = \sum_{\mathbb{U}} \delta_k y_k^* + \sum_{\mathbb{U}} (1 - \delta_k) \hat{y}_k^*. \quad (2.2.33)$$

Usando el principio de  $\pi$ -expansión, se puede definir el estimador

$$\begin{aligned} \hat{t}_{par} = \hat{t}_{y'_k} &= \sum_s \delta_k \check{y}_k^* + \sum_s (1 - \delta_k) \check{\hat{y}}_k^* \\ &= \sum_s \delta_k \frac{y_k^*}{\pi_k} + \sum_s (1 - \delta_k) \frac{\hat{y}_k^*}{\pi_k} \end{aligned} \quad (2.2.34)$$

como el estimador para el total expresado por (2.2.33). Si  $\hat{\boldsymbol{\beta}}$  está dado por (2.2.9) el estimador será

$$\begin{aligned} \hat{t}_{par} = \hat{t}_{y'_k} &= \sum_s \delta_k \frac{y_k^*}{\pi_k} + \sum_s (1 - \delta_k) \frac{\mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s}{\pi_k} \\ &= \sum_s \delta_k \frac{y_k^*}{\pi_k} + \sum_s (1 - \delta_k) \frac{\hat{y}_k^*}{\pi_k} \\ &= \sum_s \delta_k \check{y}_k^* + \sum_s (1 - \delta_k) \check{\hat{y}}_k^*, \end{aligned} \quad (2.2.35)$$

pero si está dado por (2.2.10) es estimador será

$$\begin{aligned} \hat{t}_{par}^{\pi} = \hat{t}_{y'_k} &= \sum_s \delta_k \frac{y_k^*}{\pi_k} + \sum_s (1 - \delta_k) \frac{\mathbf{x}_k \cdot \hat{\boldsymbol{\beta}}_s^{\pi}}{\pi_k} \\ &= \sum_s \delta_k \frac{y_k^*}{\pi_k} + \sum_s (1 - \delta_k) \frac{\hat{y}_k^*}{\pi_k} \end{aligned}$$

$$= \sum_s \delta_k \check{y}_k^* + \sum_s (1 - \delta_k) \check{y}_k^*. \quad (2.2.36)$$

## 2.3. Simulación

En lo que sigue se presentará una simulación de Monte Carlo que permite estudiar el comportamiento de los estimadores propuestos y así seleccionar el estimador más adecuado para la estimación del parámetro  $t = \sum_{\mathbb{U}} y_k^*$ .

### 2.3.1. Diseño de la Simulación General

Para la simulación se creó una población de tamaño  $N = 1000$  conformada por una variable  $y^*$  (variable latente) y una variable  $x$  (variable auxiliar) relacionadas mediante la ecuación  $y^* = 5 + 2x + \varepsilon$  donde  $\varepsilon \sim N(0, 1)$  y  $x \sim N(1, 1)$ .

En la Tabla 2.2 se presentan algunas medias resumen para las variables  $y^*$  y  $x$ . Además se presentan algunas medidas resumen para la variable  $y$ , que es la variable que resulta de censurar artificialmente la variable  $y^*$  en el punto límite de censura  $T = 4$ . La Figura 2.1 (a) y (b) muestra un histograma de frecuencias con densidad estimada por kernel para la variable latente y para la variable censurada respectivamente. La figura 2.2 (a) y (b) muestra la dispersión de la variable latente y censurada cada una con respecto a la variable auxiliar. La Figura 2.2 (a) muestra el límite de censura mediante una línea punteada sobre la ordenada 4, se incluye la línea de regresión estimada mediante regresión clásica. La figura 2.2 (b), muestra la línea de regresión estimada mediante regresión censurada. Se realizan cuatro

$y^*$	$y$	$x$
Min. : 0.03556	Min. : 4.000	Min. :-2.5091
1st Qu.: 5.44955	1st Qu.: 5.450	1st Qu.: 0.3575
Median : 7.07091	Median : 7.071	Median : 1.0185
Mean : 6.99809	Mean : 7.107	Mean : 0.9949
3rd Qu.: 8.45880	3rd Qu.: 8.459	3rd Qu.: 1.6691
Max. :14.09196	Max. :14.092	Max. : 3.9340
$t_{y^*} = 6998.0940$	$t_y = 7106.6347$	$t_x = 994.8667$

TABLA 2.2. Algunas medidas estadísticas de los datos poblacionales simulados.

ejecuciones distintas de las simulaciones de Monte Carlo. En cada una se extraen  $M = 20000$  muestras bajo un muestreo aleatorio simple sin reemplazamiento sobre toda la población  $\mathbb{U}$  simulada. En cada ejecución de la simulación se varía el tamaño de las muestras  $n = (200, 400, 600, 800)$ . En el Apéndice A como resultado de la simulación, se presentan histogramas de frecuencia que aproximan las distribuciones de los 10 estimadores propuestos.

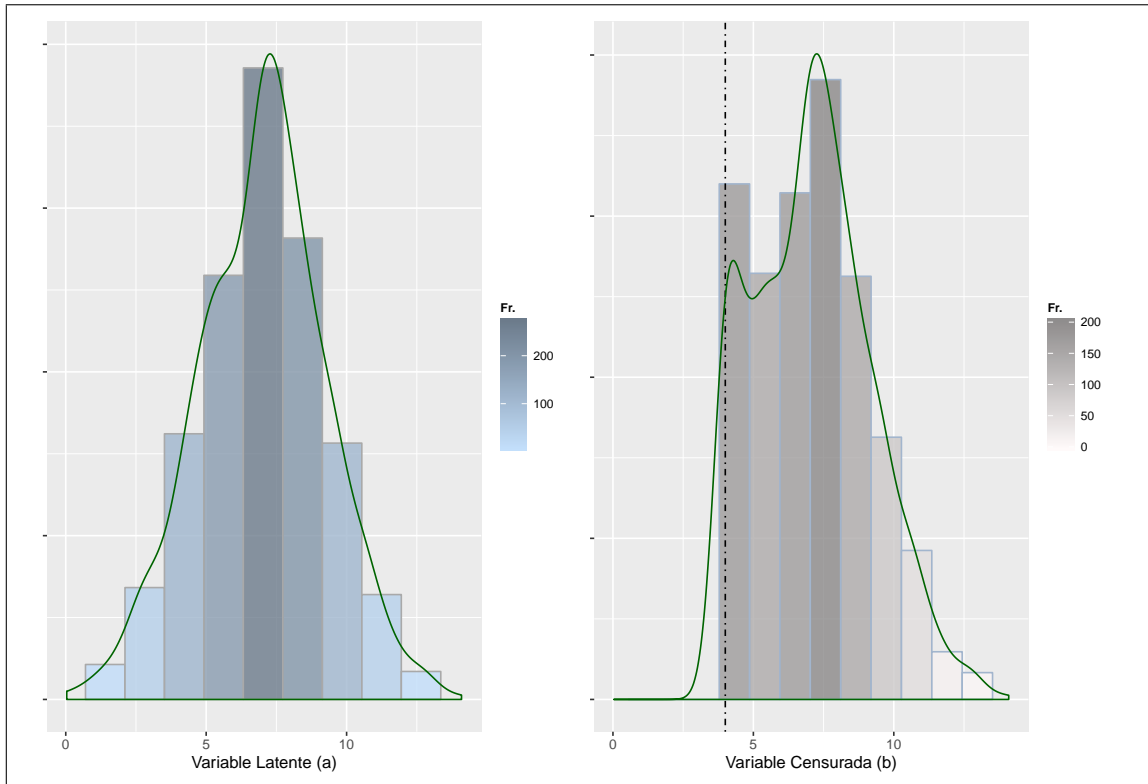


FIGURA 2.1. Histograma con curva de densidad kernel (a) y (b).

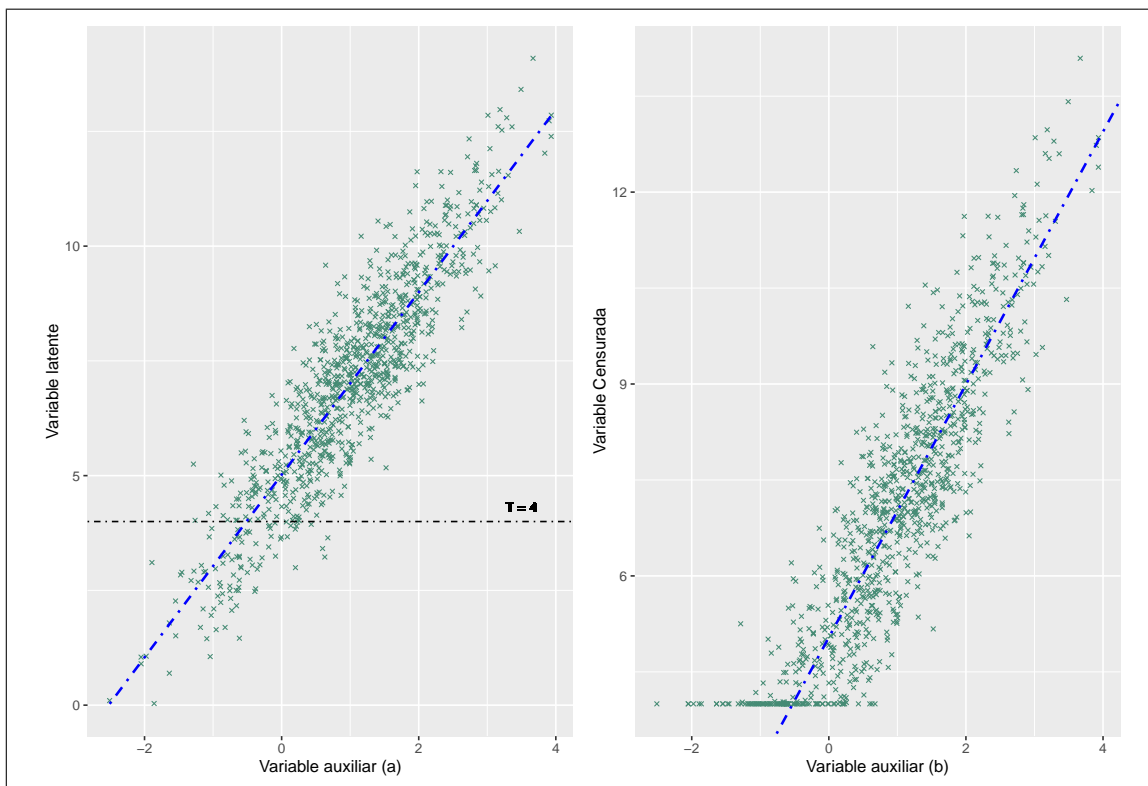


FIGURA 2.2. Gráfico de dispersión de la variable latente con línea de regresión clásica (a) y la variable censurada (b) con línea de regresión censurada.

Para evaluar los estimadores y determinar el más adecuado como solución al problema de estimación se calcularán las siguientes medidas:

$$\bar{\hat{t}} = \frac{1}{M} \sum_{m=1}^M \hat{t}_m \quad (2.3.1)$$

el cual es una estimador del valor esperado de los estimadores  $E[\hat{t}]$ ,

$$\hat{\mathbf{B}} = \frac{1}{M} \sum_{m=1}^M (\hat{t}_m - t) \quad (2.3.2)$$

como estimador del sesgo e indicará la exactitud de los estimadores, el sesgo relativo dado por

$$\mathbf{SR} = \frac{1}{M} \sum_{m=1}^M \frac{\hat{t}_m - t}{t} \quad (2.3.3)$$

como medida del tamaño del estimador del sesgo y por último se presentará un estimador del error cuadrático medio dado por

$$\widehat{\mathbf{ECM}} = \frac{1}{M} \sum_{m=1}^M (\hat{t}_m - t)^2 \quad (2.3.4)$$

el cual es una medida de precisión para los estimadores propuestos.

Para la selección de los estimadores más adecuados para la estimación del total de la variable latente  $t_{y^*}$  se tendrán en cuenta los siguientes criterios.

- I El estimador cuyo valor  $\bar{\hat{t}}$  se encuentre más cerca del valor  $t_{y^*}$ .
- II El estimador cuyo valor  $\hat{\mathbf{B}}$  en comparación con el resto de los estimadores sea menor.
- III El estimador cuyo valor  $\widehat{\mathbf{ECM}}$  en comparación con el resto de los estimadores sea menor.

### 2.3.2. Resultados de la Simulación

La Tabla 2.3 muestra las medidas resumen calculadas mediante las ecuaciones (2.3.1), (2.3.2), (2.3.3), (2.3.4) para cada uno de los estimadores propuestos en cada una de las cuatro ejecuciones de la simulación con respecto al parámetro de interés  $t_{y^*}$ . De los resultados de la simulación presentados en la Tabla 2.3 e implementando los criterios dados por I, II y III se puede mencionar lo siguiente

- 1 El verdadero valor del total poblacional para la variable latente es  $t_{y^*} = 6998.0904$ . De los diez estimadores propuestos, sólo los estimadores (1) y (3) tienen el valor de  $\bar{\hat{t}}$  más cercano al valor  $t_{y^*}$ . Los demás estimadores se encuentran considerablemente más lejos. Una característica interesante de los

		Estimadores									
$n$		(2.2.12)	(2.2.13)	(2.2.15)	(2.2.16)	(2.2.18)	(2.2.22)	(2.2.26)	(2.2.27)	(2.2.35)	(2.2.36)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\bar{t}$	200	7005,16	7105,72	7007,50	7109,15	7109,15	7105,72	7106,81	7105,72	7116,63	7074,81
	400	7005,14	7106,12	7005,38	7106,82	7106,82	7106,12	7106,58	7106,12	7114,09	7073,18
	600	7004,82	7106,07	7005,17	7106,60	7106,60	7106,07	7106,26	7106,07	7113,76	7073,10
	800	7005,31	7106,62	7005,46	7106,83	7106,83	7106,62	7106,69	7106,62	7113,95	7073,45
$\hat{B}$	200	7,06	107,63	9,4	111,06	111,06	107,63	108,72	107,63	118,53	76,72
	400	7,05	108,02	7,28	108,72	108,72	108,02	108,49	108,02	115,99	75,09
	600	6,73	107,98	7,07	108,5	108,5	107,98	108,16	107,98	115,67	75,01
	800	7,22	108,52	7,36	108,74	108,74	108,52	108,59	108,52	115,86	75,35
SR	200	0,001	0,0154	0,0013	0,0159	0,0159	0,0154	0,0155	0,0154	0,0169	0,011
	400	0,001	0,0154	0,001	0,0155	0,0155	0,0154	0,0155	0,0154	0,0166	0,0107
	600	0,001	0,0154	0,001	0,0155	0,0155	0,0154	0,0155	0,0154	0,0165	0,0107
	800	0,001	0,0155	0,0011	0,0155	0,0155	0,0155	0,0155	0,0155	0,0166	0,0108
$\widehat{ECM}$	200	4169,66	15305,04	20685,05	29379,94	29379,94	15305,04	15715,64	15305,04	31024,11	24974,17
	400	1621,08	13091,85	7712,15	18178,49	18178,49	13091,85	13259,62	13091,85	19798,61	12754,4
	600	724,04	12283,8	3474,26	14621,04	14621,04	12283,8	12347,69	12283,8	16213,77	8800,69
	800	312,54	12014,88	1340,42	12888,42	12888,42	12014,88	12041,48	12014,88	14486,86	6872,24

TABLA 2.3. Estimación de la esperanza, el sesgo, el sesgo relativo y el error cuadrático medio con respecto del parámetro  $t_{y^*}$ . Simulaciones con  $M = 20000$  y  $n = (200, 400, 600, 800)$ .

estimadores propuestos con respecto al valor  $\bar{\hat{t}}$ , es que este valor se mantiene estable mientras el tamaño de muestra aumenta. En el caso particular de los estimadores (1) y (3) esto significa que la variación de los tamaños de muestras no produce alteraciones significativas en el valor  $\bar{\hat{t}}$  manteniéndose cerca del valor  $t_{y^*}$

- 2 Para los estimadores (1) y (2) el valor  $\hat{\mathbf{B}}$  es el más pequeño de todos los estimadores propuestos. Sin embargo el valor de  $\hat{\mathbf{B}}$  para estos estimadores no es cero ni cercano a cero. Lo cual es un indicio de que los estimadores (1) y (3) no son insesgados. Esto mismo se tiene para los otros estimadores, sólo que el valor del sesgo es mucho más grande y en todos los casos positivo.
- 3 Dado que todos los estimadores son sesgados, el valor  $\mathbf{SR}$  permite determinar si la medida del sesgo es grande o pequeña. Para los estimadores (1) y (3) el valor del sesgo relativo es del orden de 0,1 %, muy cercano a cero, mientras que para los otros estimadores este valor es del orden de 1.5 %. Lo anterior indica que a pesar del sesgo de los estimadores (1) y (3), este sesgo es pequeño, al punto de que se podría hablar que son aproximadamente insesgados.
- 4 Comparando los valores  $\widehat{\mathbf{ECM}}$  para los diez estimadores se observa que para (1) y (3) sus valores  $\widehat{\mathbf{ECM}}$  son los más pequeños. Siendo menor para el caso del estimador (1). En general, para todos los estimadores, se tiene que con forme el tamaño de muestra aumenta el valor de  $\widehat{\mathbf{ECM}}$  disminuye. Esto indica que la precisión de los estimadores mejora conforme el  $n$  aumenta, especialmente en el caso de los estimadores (1) y (3).

De los análisis anteriores es claro que, de los diez estimadores propuestos, solamente los estimadores (1) y (3) (dados por (2.2.12) y (2.2.15) y simbolizados como  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$ ) pueden ser considerados como apropiados para la tarea de estimación del total  $t_{y^*}$  de acuerdo con los criterios de selección (I, II, III) dados en la página 38. Ahora bien, con el propósito de comparar entre estos dos estimadores, se usará la eficiencia relativa definida por

$$\mathbf{ER}(\hat{t}_{sin}, \hat{t}_{\pi TS}) = \frac{\widehat{\mathbf{ECM}}(\hat{t}_{sin})}{\widehat{\mathbf{ECM}}(\hat{t}_{\pi TS})} \quad (2.3.5)$$

como modo de determinar el más eficiente. Valores cercanos a uno indican que los dos estimadores son igualmente eficientes, menores que uno indican que se gana eficiencia usando el estimador  $\hat{t}_{sin}$  y mayores que uno indican que se gana eficiencia usando el estimador  $\hat{t}_{\pi TS}$ . Usando los datos de la simulación del valor  $\widehat{\mathbf{ECM}}$  cuando  $n = 800$  para estos dos estimadores se tiene que

$$\mathbf{ER}(\hat{t}_{sin}, \hat{t}_{\pi TS}) = \frac{312,54}{1340,42} = 0.233165.$$

Dado que  $\mathbf{ER}(\hat{t}_{sin}, \hat{t}_{\pi TS}) < 1$  se concluye que se gana mayor eficiencia al usar el estimador  $\hat{t}_{sin}$  y por tanto este estimador es el mejor estimador para el total  $t_{y^*}$ .

Las medias presentadas en la Tabla 2.3 muestran el desempeño de los estimadores en relación con el total  $t_{y^*}$ , que es el objetivo del presente trabajo de tesis. Por otro lado, es interesante observar el desempeño de los estimadores con respecto al total de la variable censurada  $t_y$ . La tabla 2.4 muestra las medidas (2.3.1), (2.3.2), (2.3.3), (2.3.4) calculadas con relación al total  $t_y$ . De la Tabla 2.4 se puede mencionar lo

		Estimadores									
$n$		(2.2.12)	(2.2.13)	(2.2.15)	(2.2.16)	(2.2.18)	(2.2.22)	(2.2.26)	(2.2.27)	(2.2.35)	(2.2.36)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
$\bar{t}$	200	7005,16	7105,72	7007,50	7109,15	7109,15	7105,72	7106,81	7105,72	7116,63	7074,81
	400	7005,14	7106,12	7005,38	7106,82	7106,82	7106,12	7106,58	7106,12	7114,09	7073,18
	600	7004,82	7106,07	7005,17	7106,60	7106,60	7106,07	7106,26	7106,07	7113,76	7073,10
	800	7005,31	7106,62	7005,46	7106,83	7106,83	7106,62	7106,69	7106,62	7113,95	7073,45
$\hat{B}$	200	-101,48	-0,91	-99,14	2,51	2,51	-0,91	0,18	-0,91	9,99	-31,82
	400	-101,49	-0,52	-101,26	0,18	0,18	-0,52	-0,05	-0,52	7,45	-33,45
	600	-101,81	-0,56	-101,47	-0,04	-0,04	-0,56	-0,38	-0,56	7,13	-33,53
	800	-101,32	-0,02	-101,18	0,2	0,2	-0,02	0,05	-0,02	7,32	-33,19
SR	200	-0,0143	-0,0001	-0,014	0,0004	0,0004	-0,0001	0	-0,0001	0,0014	-0,0045
	400	-0,0143	-0,0001	-0,0142	0	0	-0,0001	0	-0,0001	0,001	-0,0047
	600	-0,0143	-0,0001	-0,0143	0	0	-0,0001	-0,0001	-0,0001	0,001	-0,0047
	800	-0,0143	0	-0,0142	0	0	0	0	0	0,001	-0,0047
$\widehat{ECM}$	200	14417,22	3721,98	30425,22	17052,95	17052,95	3721,98	3896,03	3721,98	17073,5	20101,7
	400	11872,4	1423,1	17911,93	6357,64	6357,64	1423,1	1490,3	1423,1	6399,41	8235,47
	600	11044,42	624,45	13720,11	2847,85	2847,85	624,45	649,04	624,45	2885,38	4299
	800	10526,32	237,22	11523,18	1064,83	1064,83	237,22	248,93	237,22	1117,36	2295,94

TABLA 2.4. Estimación de la esperanza, el sesgo, el sesgo relativo y el error cuadrático medio con respecto del parámetro  $t_y$ . Simulaciones con  $M = 20000$  y  $n = (200, 400, 600, 800)$ .

siguiente:

1. El comportamiento general de los estimadores propuestos mostrado en la Tabla 2.4 es completamente opuesto al mostrado en la Tabla 2.3. Los estimadores (1) y (3) con respecto al total  $t_y$  no son estimadores óptimos.
2. El verdadero valor del total poblacional de la variable censurada es  $t_y = 7106.6347$ . El valor  $\bar{t}$  para los estimadores (2), (4), (5), (6), (7) y (8) se encuentra muy cerca del total  $t_y$ . Para estos estimadores el valor de  $\bar{t}$  se mantiene estable con forme el tamaño de muestra  $n$  aumenta. Para el estimador (10) el valor  $\bar{t}$  no se encuentra cerca de ninguno de los totales  $t_{y^*}$  ni  $t_y$ .
3. El valor del estimador del sesgo  $\hat{B}$  para los estimadores (2), (4), (5), (6), (7) y (8) es cercano a cero y se acerca a este valor conforme  $n$  aumenta. Esto indica

que estos estimadores son insesgados o aproximadamente insesgados para el total  $t_y$ . Para los estimadores (2), (6), (7) y (8) el sesgo es negativo a diferencia del estimador (5) para el cual el sesgo es positivo.

4. Para (2), (4), (5), (6), (7) y (8) el valor **SR** es prácticamente cero, indicando que los estimadores mencionados son insesgados para el valor de  $t_y$ . En el caso del estimador (9) el sesgo es del orden de 0.1%, siendo este muy pequeño, indicando que (9) podría ser un estimador aproximadamente insesgado para el valor  $t_y$ .
5. De los estimadores (2), (4), (5), (6), (7) y (8) que mostraron ser estimadores para el total  $t_y$ , sólo los estimadores (2), (6), (7) y (8) tienen un valor de  $\widehat{ECM}$  menor en comparación con los otros estimadores. Estos cuatro estimadores son más precisos para estimar el valor de la variable censurada y en general serán los mejores estimadores para el verdadero valor de esta variable.
6. Para los estimadores (2) y (4), que involucran pesos muestrales en la estimación del vector de parámetros  $\hat{\beta}$ , el valor de  $\bar{t}$  está más cerca de  $t_y$  que de  $t_{y^*}$ . Esto indica que la inclusión de los pesos muestrales en la estimación de los parámetros  $\hat{\beta}$  no mejora la exactitud de la estimación para  $t_{y^*}$ . La anterior afirmación tiene lugar si se comparan estos estimadores con sus homólogos (1) y (3). Los parejas de estimadores (1)-(2) y (3)-(4) tienen la misma estructura, se diferencian en la forma de calcular el vector de parámetros que hace parte de su definición. Para (1) y (3) (que no involucran pesos muestrales) el valor  $\bar{t}$  es más cercano al valor real  $t_y^*$  en cambio para (2) y (4) (que si involucran pesos muestrales) el valor  $\bar{t}$  está más lejos de  $t_y^*$  y más cerca de  $t_y$ .

## 2.4. Estimadores para el Total Poblacional en Muestras con Observaciones Censuradas.

De acuerdo con los resultados de la simulación presentados en la Sección 2.3.2, los estimadores apropiados para la tarea de estimación de la variable latente en muestras con observaciones censuradas corresponden a los estimadores *Tobit Sintético* y  $\pi$ -*estimador Tobit Sintético* dados por las ecuaciones (2.2.12) y (2.2.15) y presentados en la Tabla 2.3 como estimadores (1) y (3). En esta sección se presentará un análisis más detallado del comportamiento de estos dos estimadores. Por un lado, se darán expresiones analíticas para su esperanza y su varianza. Y por otro, se presentará una nueva simulación que permitirá estudiar con detalle el comportamiento de los estimadores.

### 2.4.1. Esperanza y Varianza para el Estimador Tobit Sintético

Para el estimador dado por (2.2.12) y que involucra el modelo de regresión Tobit  $\psi$  dado por la expresión (1.4.10) tiene como esperanza

$$\begin{aligned}
\mathbf{E}_\psi[\hat{t}_{sin}] &= \mathbf{E}_\psi\left[\sum_{k=1}^N y_k^{*'}\right] \\
&= \mathbf{E}_\psi\left[\sum_{k=1}^N \mathbf{x}_{k\cdot} \hat{\boldsymbol{\beta}}\right] \\
&= \sum_{k=1}^N \mathbf{E}_\psi\left[\mathbf{x}_{k\cdot} \hat{\boldsymbol{\beta}}\right] \\
&= \sum_{k=1}^N \mathbf{x}_{k\cdot} \mathbf{E}_\psi\left[\hat{\boldsymbol{\beta}}\right]
\end{aligned} \tag{2.4.1}$$

Dado que este estimador no involucra la información de los pesos muestrales, la esperanza de este estimador no depende del diseño muestral. Únicamente depende del modelo  $\psi$  lo cual queda indicado por la expresión  $\mathbf{E}_\psi[\hat{\boldsymbol{\beta}}_s]$  en la ecuación (2.4.1). Con respecto al comportamiento de la estimación de los parámetros  $\hat{\boldsymbol{\beta}}$ , Amemiya (1973, p. 1004) demuestra la fuerte consistencia de los estimadores máximo verosímil. En particular demuestra que para muestras finitas los estimadores de los parámetros convergen al verdadero valor del parámetro. Acorde con esto, se espera que conforme el tamaño de muestra  $n$  aumente, el estimador  $\hat{\boldsymbol{\beta}}$  converja a  $\boldsymbol{\beta}$  y por ende la esperanza del estimador  $\hat{t}_{sin}$  se acerque al verdadero valor  $t_{y^*}$ . Sin embargo, el hecho anterior no garantiza el insesgamiento de las estimaciones, lo cual se comprobó en las simulaciones.

Para este estimador la varianza estará dada por

$$\begin{aligned}
\mathbf{Var}[\hat{t}_{sin}] &= \mathbf{Var}\left[\sum_{k=1}^N y_k^{*'}\right] \\
&= \sum_{k=1}^N \sum_{m=1}^N \mathbf{Cov}(y_k^{*'}, y_m^{*'}) \\
&= \sum_{k=1}^N \sum_{m=1}^N \mathbf{Cov}(\mathbf{x}_{k\cdot} \hat{\boldsymbol{\beta}}, \mathbf{x}_{m\cdot} \hat{\boldsymbol{\beta}}) \\
&= \sum_{k=1}^N \sum_{m=1}^N \mathbf{Cov}\left(\sum_{i=1}^J x_{ki} \hat{\beta}_i, \sum_{j=1}^J x_{mj} \hat{\beta}_j\right) \\
&= \sum_{k=1}^N \sum_{m=1}^N \sum_{i=1}^J x_{ki} \mathbf{Cov}\left(\hat{\beta}_i, \sum_{j=1}^J x_{mj} \hat{\beta}_j\right) \\
&= \sum_{k=1}^N \sum_{m=1}^N \sum_{i=1}^J \sum_{j=1}^J x_{ki} x_{mj} \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^J \sum_{j=1}^J \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \sum_{k=1}^N x_{ki} \sum_{m=1}^N x_{mj} \\
&= \sum_{i=1}^J \sum_{j=1}^J \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) t_{x_{ki}} t_{x_{mj}} \tag{2.4.2}
\end{aligned}$$

De la ecuación (2.4.2) se observa que la varianza del estimador  $\hat{t}_{sin}$  depende de la matriz de varianzas-covarianzas de los estimadores  $\hat{\beta}$  de los parámetros del modelo. Un estimador de  $\mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j)$  está dado por la inversa de la matriz de información de Fisher  $\mathcal{F}^{-1}$  dada por la ecuación(1.4.34). De acuerdo con lo anterior se sigue que

$$\hat{\mathbf{V}}[\hat{t}_{sin}] = \sum_{i=1}^J \sum_{j=1}^J \mathcal{F}_{ij}^{-1} t_{x_{ki}} t_{x_{mj}} \tag{2.4.3}$$

será un estimador de la varianza dada por (2.4.2).

### 2.4.2. Esperanza y Varianza para el $\pi$ -Estimador Tobit Sintético

El estimador dado por (2.2.15) involucra en su cálculo el modelo de regresión Tobit dado por (1.4.10) y el principio de  $\pi$ -expansión expresado en la Definición 1.7. Por tal razón el calculo de su esperanza dependerá de dos elementos aleatorios. Uno debido al modelo Tobit ( $\psi$ ) y otro debido al diseño muestral ( $\mathbf{p}(\cdot)$ ). Razón por la cual la esperanza de este estimador estará dada por

$$\begin{aligned}
\mathbf{E}[\hat{t}_{\pi TS}] &= \mathbf{E}_p[\mathbf{E}_\psi[\sum_s \check{y}_k^{*'}]] \\
&= \mathbf{E}_p[\mathbf{E}_\psi[\sum_{k=1}^N I_k \check{y}_k^{*'}]] \\
&= \mathbf{E}_p[\sum_{k=1}^N \mathbf{E}_\psi[I_k \check{y}_k^{*'}]] \\
&= \mathbf{E}_p[\sum_{k=1}^N I_k \mathbf{E}_\psi[\check{y}_k^{*'}]] \\
&= \sum_{k=1}^N \mathbf{E}_p[I_k] \mathbf{E}_\psi[\check{y}_k^{*'}] \\
&= \sum_{k=1}^N \pi_k \frac{\mathbf{E}_\psi[y_k^{*'}]}{\pi_k} \\
&= \sum_{k=1}^N \mathbf{E}_\psi[y_k^{*'}] \\
&= \sum_{k=1}^N \mathbf{E}_\psi[\mathbf{x}_k \cdot \hat{\beta}]
\end{aligned}$$

$$= \sum_{k=1}^N \mathbf{x}_k \cdot \mathbf{E}_\psi [\hat{\beta}] \quad (2.4.4)$$

En la expresión (2.4.4) el subíndice  $p$  y  $\psi$  denotan el cálculo de la esperanza con respecto a los elementos aleatorios del diseño y del modelo. La esperanza de este estimador presenta la misma dificultad que la esperanza del estimador dado por (2.4.1). Por esta razón la esperanza del estimador  $\hat{t}_{\pi TS}$  convergerá al verdadero valor del total  $t_{y^*}$  conforme el tamaño de las muestras aumenten, pero no necesariamente serán insesgados. La varianza de este estimador estará dada por

$$\mathbf{Var}[\hat{t}_{\pi TS}] = \mathbf{Var}_p(\mathbf{E}_\psi[\hat{t}_{\pi TS} | s]) + \mathbf{E}_p[\mathbf{Var}_\psi(\hat{t}_{\pi TS} | s)] \quad (2.4.5)$$

tomando el primer término del lado derecho de la ecuación (2.4.5) se tiene que

$$\begin{aligned} \mathbf{Var}_p[\mathbf{E}_\psi[\hat{t}_{\pi TS} | s]] &= \mathbf{Var}_p(\mathbf{E}_\psi[\sum_s \check{y}_k^* | s]) \\ &= \mathbf{Var}_p(\mathbf{E}_\psi[\sum_{k=1}^N I_k \check{y}_k^* | s]) \\ &= \mathbf{Var}_p(\sum_{k=1}^N \mathbf{E}_\psi[I_k \check{y}_k^* | s]) \\ &= \mathbf{Var}_p(\sum_{k=1}^N I_k \mathbf{E}_\psi[\check{y}_k^* | s]) \\ &= \sum_{k=1}^N \mathbf{Var}_p(I_k) \mathbf{E}_\psi^2[\check{y}_k^* | s] \\ &\quad + 2 \sum_k \sum_m \mathbf{Cov}_p(I_k \mathbf{E}_\psi[\check{y}_k^* | s], I_m \mathbf{E}_\psi[\check{y}_m^* | s]) \\ &= \sum_{k=1}^N \mathbf{Var}_p(I_k) \mathbf{E}_\psi^2[\check{y}_k^* | s] \\ &\quad + 2 \sum_k \sum_m \mathbf{Cov}_p(I_k, I_m) \mathbf{E}_\psi[\check{y}_k^* | s] \mathbf{E}_\psi[\check{y}_m^* | s] \\ &= \sum_{k=1}^N \Delta_{kk} \mathbf{E}_\psi^2[\check{y}_k^* | s] + 2 \sum_k \sum_m \Delta_{km} \mathbf{E}_\psi[\check{y}_k^* | s] \mathbf{E}_\psi[\check{y}_m^* | s] \\ &= \sum_{k=1}^N \sum_{m=1}^N \Delta_{km} \mathbf{E}_\psi[\check{y}_k^*] \mathbf{E}_\psi[\check{y}_m^*] \\ &= \sum_{k=1}^N \sum_{m=1}^N \Delta_{km} \frac{\mathbf{x}_k \cdot \mathbf{E}_\psi[\hat{\beta}]}{\pi_k} \frac{\mathbf{x}_m \cdot \mathbf{E}_\psi[\hat{\beta}]}{\pi_m} \end{aligned} \quad (2.4.6)$$

con  $\Delta_{kk}$  y  $\Delta_{ik}$  dados por las expresiones (1.1.8) y (1.1.9) respectivamente y con  $I_k$  definido por (1.1.4). La expresión (2.4.6) no es otra cosa que la varianza del  $\pi$ -estimador dada en la ecuación (1.3.6). Por otra parte, tomando el segundo término del lado derecho (2.4.5) se tiene que

$$\mathbf{E}_p[\mathbf{Var}_\psi(\hat{t}_{\pi TS} | s)] = \mathbf{E}_p[\mathbf{Var}_\psi(\sum_s \check{y}_k^* | s)]$$

$$\begin{aligned}
&= \mathbf{E}_p \left[ \sum_{k=1}^n \sum_{m=1}^n \mathbf{Cov}(\check{y}_k^*, \check{y}_m^*) \right] \\
&= \mathbf{E}_p \left[ \sum_{k=1}^n \sum_{m=1}^n \sum_{i=1}^J \sum_{j=1}^J \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \right] \\
&= \mathbf{E}_p \left[ \sum_{k=1}^N \sum_{m=1}^N \sum_{i=1}^J \sum_{j=1}^J I_k I_m \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \right] \\
&= \sum_{k=1}^N \sum_{m=1}^N \sum_{i=1}^J \sum_{j=1}^J \mathbf{E}_p [I_k I_m] \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \\
&= \sum_{k=1}^N \sum_{m=1}^N \sum_{i=1}^J \sum_{j=1}^J \pi_{km} \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \\
&= \sum_{i=1}^J \sum_{j=1}^J \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \sum_{k=1}^N \sum_{m=1}^N \pi_{km} \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \quad (2.4.7)
\end{aligned}$$

juntando (2.4.6) y (2.4.7) se puede reescribir la varianza del estimador  $\hat{t}_{\pi TS}$  dada por (2.4.5) de la siguiente manera

$$\begin{aligned}
\mathbf{Var}[\hat{t}_{\pi TS}] &= \sum_{k=1}^N \sum_{m=1}^N \Delta_{km} \frac{\mathbf{x}_k \cdot \mathbf{E}_\psi[\hat{\beta}]}{\pi_k} \frac{\mathbf{x}_m \cdot \mathbf{E}_\psi[\hat{\beta}]}{\pi_m} \\
&\quad + \sum_{i=1}^J \sum_{j=1}^J \mathbf{Cov}(\hat{\beta}_i, \hat{\beta}_j) \sum_{k=1}^N \sum_{m=1}^N \pi_{km} \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \quad (2.4.8)
\end{aligned}$$

Un estimador de la varianza dada por la ecuación (2.4.8) estará dado por

$$\begin{aligned}
\hat{\mathbf{V}}[\hat{t}_{\pi TS}] &= \sum_{k=1}^n \sum_{m=1}^n \check{\Delta}_{km} \frac{\mathbf{x}_k \cdot \hat{\beta}}{\pi_k} \frac{\mathbf{x}_m \cdot \hat{\beta}}{\pi_m} + \sum_{i=1}^J \sum_{j=1}^J \mathcal{F}_{ij}^{-1} \sum_{k=1}^n \sum_{m=1}^n \frac{x_{ki}}{\pi_k} \frac{x_{mj}}{\pi_m} \\
&= \sum_{k=1}^n \sum_{m=1}^n \check{\Delta}_{km} \check{y}_k^* \check{y}_m^* + \sum_{i=1}^J \sum_{j=1}^J \frac{\mathcal{F}_{ij}^{-1}}{\pi_k} \check{t}_{x_{ki}} \check{t}_{x_{mj}} \quad (2.4.9)
\end{aligned}$$

De la ecuación (2.4.8) es claro que la varianza del estimador  $\hat{t}_{\pi TS}$  es la suma de la varianza debida a la aleatoriedad del diseño muestral  $\mathbf{p}(\cdot)$  más la varianza ocasionada por la aleatoriedad del modelo  $\psi$ .

### 2.4.3. Segunda Simulación

Se realizó una nueva simulación con el propósito de analizar únicamente los estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$ . En esta nueva simulación se extrajeron  $M = 50.000$  muestras de tamaño  $n = (200, 400, 600, 800)$  de la población. Las muestras fueron seleccionadas mediante un muestreo aleatorio simple sin remplazamiento.

En esta nueva simulación, además de las medidas dadas por (2.3.1), (2.3.2), (2.3.3), (2.3.4), se calcula

$$S_{\hat{t}}^2 = \frac{1}{M-1} \sum_{m=1}^M (\hat{t}_m - \bar{\hat{t}})^2 \quad (2.4.10)$$

como estimador de la varianza,

$$\bar{\hat{V}} = \frac{1}{M} \sum_{m=1}^M \hat{V}(\hat{t}) \quad (2.4.11)$$

que indicará el valor esperado del estimador de la varianza de los estimadores. Un intervalo de confianza para la estimación  $\hat{t}_m$  con nivel de significancia  $\alpha = 0.05$

$$\hat{t} \pm Z_{1-\alpha/2} (\hat{V}(\hat{t}))^{1/2}. \quad (2.4.12)$$

Y por último, se hará el conteo de intervalos de confianza  $R$  que contienen al verdadero valor del total de la variable latente  $t_{y^*}$  y así determinar la cantidad  $R/M$  o tasas de cobertura empírica (TCE) como una estimación del valor del nivel de confianza. La Tabla 2.5 muestra los resultados de esta simulación.

Con respecto a los resultados de la segunda simulación y en concordancia con los resultados de la primera, se puede comentar lo siguiente:

- 1 Se mantiene que el valor de  $\bar{\hat{t}}$  para los estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$  se aproxima al valor real de la variable latente  $t_{y^*}$ .
- 2 El valor de  $\hat{\mathbf{B}}$  que estima el sesgo para los dos estimadores es aproximadamente de 6.8 y 7.25 respectivamente. Acorde con el valor **SR** estos valores de sesgo son muy pequeños. Esto indica que los estimadores son aproximadamente insesgados.
- 3 Con respecto a la precisión de estos dos estimadores  $\widehat{\mathbf{ECM}}$ , se tiene que el estimador  $\hat{t}_{sin}$  es más preciso que  $\hat{t}_{\pi TS}$ .
- 4 La cantidad  $S_{\hat{t}}^2$  estima el verdadero valor de la varianza con un grado de precisión obtenido con  $M = 50000$  muestras. Acorde con este valor el estimador  $\hat{t}_{sin}$  es más eficiente que el estimador  $\hat{t}_{\pi TS}$ . La eficiencia de los dos estimadores aumenta conforme el tamaño de muestra aumenta.
- 5 El valor  $\bar{\hat{V}}(\hat{t})$  que es la esperanza del estimador de la varianza esta próximo al valor  $S_{\hat{t}}^2$ . Esto indica que el estimador  $\hat{V}(\hat{t}_{sin})$  y  $\hat{V}(\hat{t}_{\pi TS})$  son estimadores insesgados o aproximadamente insesgados de la varianza.
- 6 El valor TEC que indica la tasa de cobertura estimada para los estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$  es muy alta para los dos estimadores y se encuentra muy cerca de la tasa nominal de 95% alcanzada por la técnica de intervalos de confianza. Conforme  $n$  aumenta, la tasa de cobertura aumenta.

	$n$	Estimadores	
		$\hat{t}_{sin}$	$\hat{t}_{\pi TS}$
$\bar{t}$	$n = 200$	7004,36	7004,72
	$n = 400$	7004,99	7005,19
	$n = 600$	7005,1	7005,29
	$n = 800$	7005,01	7005,04
$\hat{B}$	$n = 200$	6,27	6,63
	$n = 400$	6,89	7,09
	$n = 600$	7	7,2
	$n = 800$	6,93	6,94
RB	$n = 200$	0,0009	0.0009
	$n = 400$	0.001	0.001
	$n = 600$	0.001	0.001
	$n = 800$	0.001	0.001
$\widehat{ECM}$	$n = 200$	4227.46	20761.60
	$n = 400$	1599.27	7791.01
	$n = 600$	737.06	3490.36
	$n = 800$	301.93	1334.94
$S_{\hat{t}}^2$	$n = 200$	237107,01	3552862,02
	$n = 400$	22426,95	192751,7
	$n = 600$	4342,42	18119,36
	$n = 800$	927,87	1478,96
$\bar{V}(\hat{t})$	$n = 200$	5112,64	16662,69
	$n = 400$	2561,07	6571,61
	$n = 600$	1708,05	3352,09
	$n = 800$	1281,58	1846,47
TCE	$n = 200$	0,97	0,92
	$n = 400$	0,99	0,93
	$n = 600$	1	0,95
	$n = 800$	1	0,98

TABLA 2.5. Medidas resumen para los estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$ .  $M = 50000$ ,  
 $n = (200, 400, 600, 800)$ .

De los resultados comentados anteriormente se debe resaltar que los dos estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$  junto con sus estimadores de varianza  $\hat{V}(\hat{t}_{sin})$  y  $\hat{V}(\hat{t}_{\pi TS})$  se encuentran diseñados bajo el cumplimiento de los supuestos expresados por C.1, C.2, C.3 y C.4 de la página 28.

#### 2.4.4. Análisis de la Censura en el Comportamiento del Estimador

Un aspecto importante de analizar corresponde con el porcentaje de censura dentro de la muestra y el comportamiento de los estimadores  $\hat{t}_{sin}$  y  $\hat{t}_{\pi TS}$ .

Para este análisis hay que tener presente el porcentaje máximo de observaciones que pueden ser censuradas. En la población, la cantidad de observaciones que son iguales o menores al límite de censura  $T = 4$  corresponde a 92 observaciones. Es decir que el porcentaje máximo de observaciones censuradas será de 9.2%. Es claro que el porcentaje de censura dentro de las muestras puede variar en un rango de (0 %,46 %), (0 %,23 %), (0 %,15.3 %) o (0 %, 11.13 %) para los tamaños de muestras dados.

Para analizar el comportamiento de los estimadores a diferentes niveles de censura se construyeron las gráficas 2.3 y 2.4 con base en los datos de la simulación de las muestras de tamaño  $n = 200$ . Cada gráfica agrupa las muestras de acuerdo al porcentaje de censura que contienen. En esta simulación las muestras se separaron en 30 grupos con porcentajes que varían entre el 2.5 % y el 17 % de censura en las muestras. La línea roja horizontal representa el valor verdadero del total de la variable latente  $\hat{t}_{y^*}$  y la línea negra horizontal representa el valor verdadero del total de la variable censurada  $\hat{t}_y$ . El rombo representa el valor esperado del estimador del total en cada grupo calculado mediante la ecuación (2.3.1). La línea negra vertical representa el valor de censura máximo en la población.

De las gráficas 2.3 y 2.4 se puede comentar lo siguiente

1. El valor  $\bar{\hat{t}}$ , que es la estimación de la esperanza del total y está representado por un rombo en las gráficas, varía conforme el porcentaje de censura cambia dentro de la muestra. En general se aprecia que para porcentajes de censura muy bajos o muy altos el valor  $\bar{\hat{t}}$  se encuentra muy alejado del valor real del total. Esto es, para porcentajes de censura muy bajos tanto el estimador  $\hat{t}_{sin}$  como el estimador  $\hat{t}_{\pi TS}$  tienden a sobreestimar el valor del total y para porcentajes muy altos de censura estos estimadores subestiman el valor del total. Esto indica que la precisión del estimador varía conforme el porcentaje de censura cambia.
2. Comparando las dos gráficas, se evidencia que está sobreestimación o subestimación del parámetro cambia más rápido para el estimador  $\hat{t}_{\pi TS}$  en comparación con estimador  $\hat{t}_{sin}$ . Se puede decir que el estimador  $\hat{t}_{sin}$  es más estable ante el cambio del porcentaje de censura.

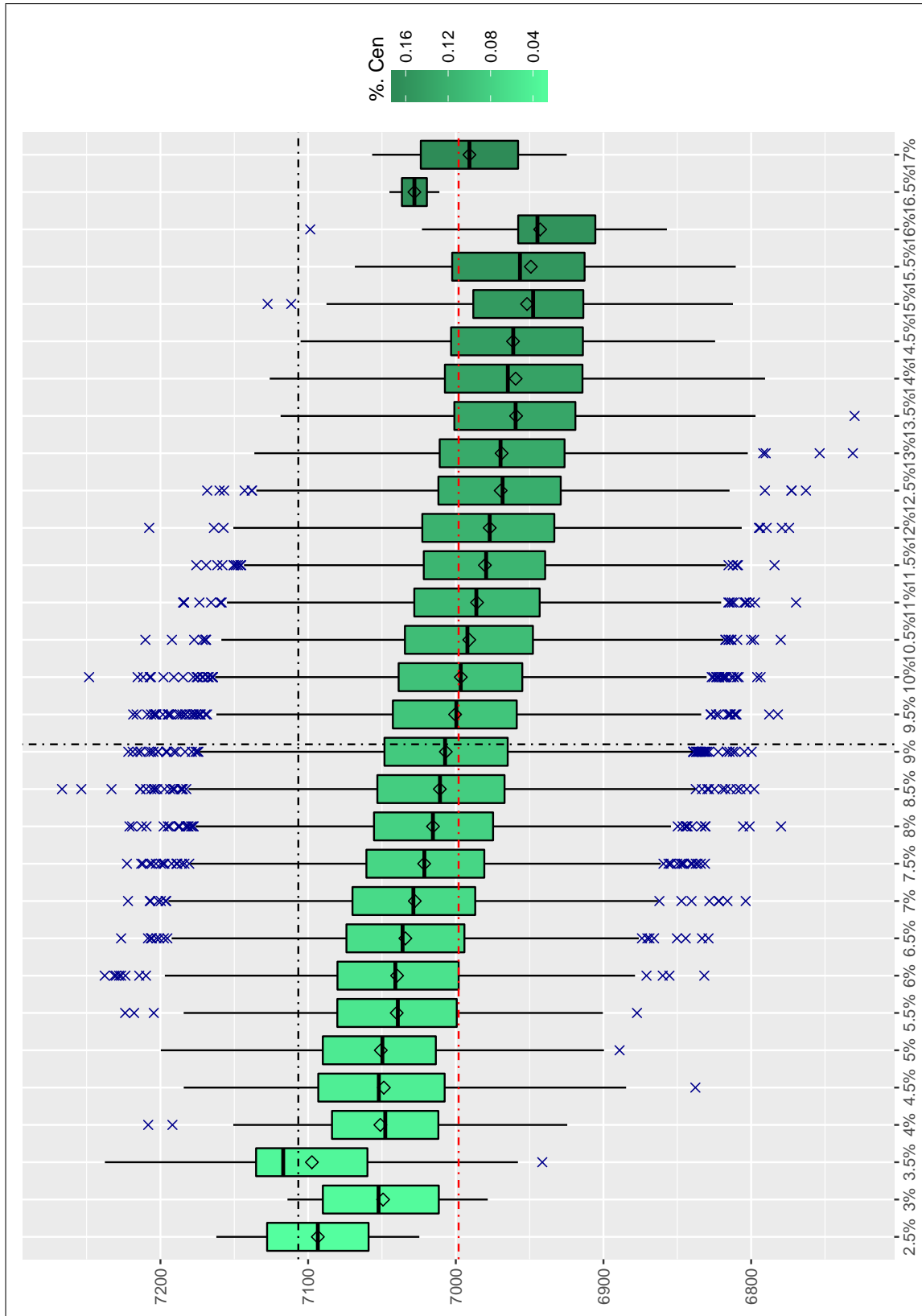


FIGURA 2.3. Agrupamiento de muestras por porcentaje de censura y gráfico de Boxplot. Estimador  $\hat{t}_{sin}$ .

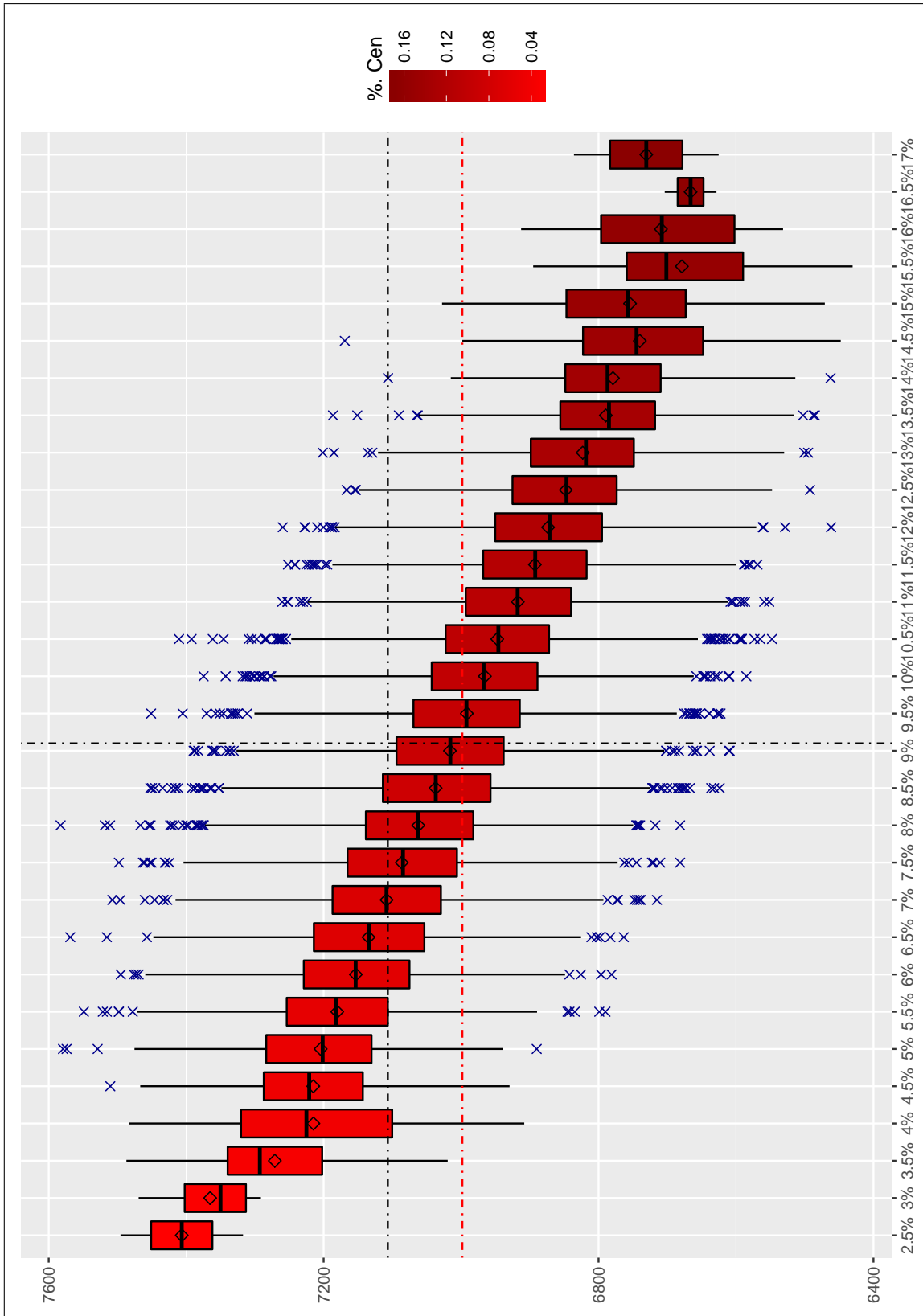


FIGURA 2.4. Agrupamiento de muestras por porcentaje de censura y gráfico de Boxplot. Estimador  $\hat{t}_{\pi TS}$ .

3. El comportamiento anterior indica la existencia de una relación inversa entre el porcentaje de censura y la exactitud del estimador.
4. Tanto para  $\hat{t}_{sin}$  como para  $\hat{t}_{\pi TS}$  las estimaciones son más precisas cuando el porcentaje de censura dentro de la muestra se acerca al porcentaje de censura máximo en la población.

## 2.5. Conclusiones

De la simulaciones presentadas anteriormente se pueden concluir los siguiente

1. Como se comento en la Sección 2.1 en el proceso de estimación del total en muestras con observaciones censuradas la variable observada es la variable censurada. Sin embargo esto representa un problema, ya que como se ha mencionado a largo del presente trabajo, el interés no está en hacer inferencia sobre la variable censurada sino sobre la variable latente o no observada.

La solución a este problema se aproxima, haciendo uso de la asociación natural que existe entre estas dos variables. La variable censurada adquiere valores de acuerdo a los valores que toma la variable latente en relación con un límite de censura. Esta asociación sugiere que un punto de partida para dar solución al problema de estimación es asumir la variable censurada como una aproximación de la variable latente y a partir de esta hacer la estimación. La simulación mostró que esta aproximación y la aproximación que crea una pseudovariable latente conformada por observaciones de la variable latente cuando estas son mayores que el límite de censura y estimaciones de las observaciones a través de un modelo de regresión Tobit cuando son menores o iguales que el límite de censura son aproximaciones erróneas ya que llevan a estimar, contrario a la intuición, el valor del total de la variable censurada y no el de la variable latente. Bajo estas aproximaciones fueron diseñados los estimadores (5) al (10) de la Tabla 2.3 y que como se comentó fueron estimadores del total censurado y en algunos casos (6), (7) y (8) fueron buenos estimadores de este total.

Se comprobó que la estrategia que permite hacer la estimación del total de la variable latente se basa en la creación de una pseudo variable latente  $y_k^*$ , conformada enteramente por observaciones estimadas a través del modelo de regresión Tobit. Creada esta variable, se asume que es una aproximación de la variable latente y a partir de ella se hace la estimación del total. Bajo esta aproximación se diseñaron los estimadores (1) al (4). De estos, sólo los estimadores (1) y (3) mostraron ser estimadores óptimos para el total de la variable latente.

2. De lo anterior se tiene que los estimadores más cercanos al total de la variable latente  $t = \sum_U y_k^*$  son los estimadores (1) y (3). De de estos dos estimadores, el mejor es el estimador (1), dado que presenta los sesgos, varianzas y errores cuadráticos medios más pequeños que el estimador (3). Además es más estable

al cambio del porcentaje de censura dentro de las muestras. Sin embargo el estimador (3) puede ser una buena opción si se cuenta con muestras de tamaños grandes, ya que el sesgo y la varianza estimada disminuyen y se incrementa su exactitud conforme aumenta el tamaño de la muestra.

3. Por otra parte, con respecto al uso de pesos muestrales en la estimación de los parámetros del modelo regresión Tobit. Se puede decir que esta inclusión se presenta como un problema en la estimación del total de la variable latente. Särndal et al. (1992, p. 518-519) discuten el uso de los pesos muestrales en la estimación de los parámetros de un modelo superpoblacional. Al respecto comentan que bajo los supuestos 1 y 2 del modelo expresado por (2.2.24) se cumple que

$$E_{\xi}[(\hat{\beta}_s - \beta)^2 | s, \mathbb{X}] \leq E_{\xi}[(\hat{\beta}_s^{\pi} - \beta)^2 | s, \mathbb{X}]. \quad (2.5.1)$$

Dado que esto se cumple para todo  $s$ , se sigue que

$$E_{\xi} E_p[(\hat{\beta}_s - \beta)^2 | s, \mathbb{X}] \leq E_{\xi} E_p[(\hat{\beta}_s^{\pi} - \beta)^2 | s, \mathbb{X}]. \quad (2.5.2)$$

Estas relaciones se mantienen inclusive si  $\hat{\beta}_s$  es el estimador máximo verosímil de  $\beta$ . La desigualdad dada en (2.5.2) conlleva a que

$$E_{\xi} E_p[(\mathbb{X}_k \hat{\beta}_s - \mathbb{X}_k \beta)^2 | s, \mathbb{X}] \leq E_{\xi} E_p[(\mathbb{X}_k \hat{\beta}_s^{\pi} - \mathbb{X}_k \beta)^2 | s, \mathbb{X}]. \quad (2.5.3)$$

Por tanto los totales que no involucran pesos muestrales en la estimación de los valores de  $y_k^*$  o  $y_k$  serán menores que los totales que si los involucran. Esto se evidencia en la Tabla 2.3 si se comparan los valores de las esperanzas para estimadores que usan pesos muestrales contra los que no los usan. Los primeros siempre son más grandes que los segundos. En el caso particular de la estimación del total de la variable latente, los estimadores que involucran pesos muestrales mostraron ser estimadores de la variable censurada y no de la latente como se podría suponer.



# Capítulo 3

## EJEMPLO DE APLICACIÓN

### 3.1. Ejemplo de Aplicación. “Población CO124”

El primer ejemplo que se expondrá se basa en el análisis del conjunto de datos *The CO124 population* presente en el libro de Särndal et al. (1992, p. 662)(ver sección B). Para este ejemplo se considera que la variable de interés es la población para el año de 1983 y como información auxiliar las variables importaciones, exportaciones, producto interno bruto y población del año 1980.

**P83** Población en 1983 (en millones).

**IMP** Importaciones en 1983 (en millones de dolares).

**EXP** Exportaciones en 1983 (en millones de dolares).

**GNP** Producto interno bruto en 1982 (en decenas de millones de dolares).

**P80** Población en 1980 (en millones)

El problema es estimar el total de la variable de la población para el año de 1983. Para esto se supondrá que la información con la que se cuenta es una muestra de  $n = 40$  en la que la variable de interés se encuentra simplemente censurada a derecha del límite de censura  $T = 100$ . Es decir, valores de la variable iguales o mayores que cien fueron registrados en la muestra con este valor.

#### 3.1.1. Procedimiento

Para la solución de este ejemplo se comentará paso a paso el procedimiento desarrollado mostrando el código que fue usado en el software R para la realización de los cálculos. En general se mostrará cómo aplicar la censura, extraer la muestra y realizar los cálculos sobre la muestra para encontrar la estimación.

a Se llaman las librerías necesarias y se cargan los datos al software.

```
> library(censReg)
> library(optimx)
> library(sampling)
> library(maxLik)
> source("codigo_funciones.R")
> set.seed(98)
> datos<-read.table("datos_theco124population.txt",header=T)
```

b Se crea una copia de la variable de interés para censurarla y se crea además un nuevo marco de datos solamente con la información necesaria.

```
> P83C<-datos$P83
> T<-100
> N=124
> n=40
> for(i in 1:N){
+   if(datos[i,3]>=T){
+     P83C[i]<-T
+   }
+ }
> datos2<-data.frame(cbind(datos$P83,P83C,datos$IMP,
datos$EXP,datos$GNP,datos$P80))
> colnames(datos2)<-c("P83", "P83C", "IMP", "EXP", "GNP", "P80")
> head(datos2)
  P83 P83C  IMP  EXP  GNP  P80
1 20.5 20.5 10395 11163 4535 18.7
2  3.7  5.0   131    63  122  3.4
3  6.6  6.6   288    57  135  6.1
4  4.4  5.0   194    76  122  4.1
5  9.1  9.1  1217   940  715  8.5
6  2.4  5.0   127   109   76  2.3
> (totallatente<-sum(datos2$P83))
[1] 4544.8
> (totalcensurado<-sum(datos2$P83C))
[1] 2573.4
```

Dado que se tiene el conjunto de datos completos, se sabe que el total de la variable P83 (variable latente o  $y^*$ ) es  $t_{y^*} = 4544.8$  millones y de la variable P83C (variable censurada o  $y$ ) es  $t_y = 2573.4$  millones.

c Del nuevo marco de datos se extrae una muestra  $n = 40$  bajo un muestreo aleatorio simple

```
> a<-sample(1:dim(datos2)[1],40,replace = F)
```

```

> mues<-datos2[a,-1]
> head(mues)
      P83C  IMP  EXP  GNP  P80
87    0.2 1456 3384  595  0.2
121 15.3 19393 20594 16904 14.7
72 100.0 13562 8304 18413 663.6
43   5.2  891  735  356  4.7
8    1.6  807  977  234  1.5
112  4.1 13890 18920 5884  4.1
> summary(mues)
P83C          IMP          EXP
Min.   : 0.20   Min.   : 127.0   Min.   :  57.0
1st Qu.: 3.40   1st Qu.: 647.8   1st Qu.: 477.5
Median : 7.90   Median : 1460.5   Median : 1553.5
Mean   : 20.35   Mean   : 12336.2   Mean   :12344.1
3rd Qu.: 24.98   3rd Qu.: 10233.0   3rd Qu.:11537.0
Max.   :100.00   Max.   :103734.0   Max.   :93310.0
GNP          P80
Min.   :  76.0   Min.   :  0.200
1st Qu.: 256.2   1st Qu.:  3.275
Median : 816.5   Median :  7.100
Mean   :10176.6   Mean   : 37.815
3rd Qu.: 7390.2   3rd Qu.: 23.800
Max.   :156300.0   Max.   :663.600

```

d Para la estimación de los parámetros del modelo de regresión Tobit se usa la función `censReg`. Para llevar los datos al caso genérico de censura en cero, a la variable P83C se le resta el punto de censura  $T = 100$ . La función `censReg` internamente hace las transformaciones necesarias y devuelve los resultados de las estimaciones en la magnitud original.

```

> mod2<-censReg(P83C~IMP+EXP+GNP+P80,data = mues,right = 100)
> mod2$estimate
(Intercept)          IMP          EXP          GNP
-1.970808e-02 -1.698455e-05  7.170272e-06 -3.125745e-05
P80          logSigma
1.086087e+00 -5.884320e-01

```

e Se calculan los estimadores propuestos como mejores estimadores para el total

$$\hat{t}_{sin} = \sum_{\mathbb{U}} y_k^{*'} \quad \text{Ec. (2.2.12)}$$

$$\hat{t}_{\pi TS} = \sum_s \frac{y_k^{*'}}{\pi_k} \quad \text{Ec. (2.2.15)}$$

También se calculan el estimador de *Horvitz-Thompson* y el GREG para comparar resultados

$$\hat{t}_{\pi y} = \sum_s \check{y}_k = \sum_s \frac{y_k}{\pi_k} \quad \text{Ec. (2.2.18)}$$

$$\hat{t}_{yr}^{\pi} = \sum_{\cup} \hat{y}_k^{\pi} + \sum_s \frac{y_k - \hat{y}_k^{\pi}}{\pi_k} \quad \text{Ec. (2.2.23)}$$

f En el orden mostrado anteriormente las estimaciones son

```
#Estimador Tobit Sintético
> m<-as.matrix(cbind(1,datos2[,c(3,4,5,6)]))
> (mest<-sum(m%*%mod2$estimate[-length(mod2$estimate)]))
[1] 4618.77
> covmat<-vcov(mod2)
> cs<-colSums(m)
> va<-0
> for(i in 1:dim(m)[2]){
+   for(j in 1:dim(m)[2]){
+     va<-va + covmat[i,j]*cs[i]*cs[j]
+   }
+ }
> va
219.716
> (incmest<-c(mest-1.96*sqrt(va),mest+1.96*sqrt(va)))
4588.748 4648.792
#pi-Estimador Tobit Sintético
> f<-as.matrix(cbind(1,mues[,2:5]))
> (mestimacion2<-
sum((f%*%mod2$estimate[-length(mod2$estimate)])/(n/N)))
[1] 2648.704
> css<-colSums(f)
> vai<-0
> for(i in 1:dim(f)[2]){
+   for(j in 1:dim(f)[2]){
+     vai<-vai + covmat[i,j]*css[i]*css[j]
+   }
+ }
> vai
1
14.90048
> pssd<-N^2*((1-(n/N))/n)*
var(f%*%mod2$estimate[-length(mod2$estimate)])
> ddd<-vai+pssd
> (incmest<-
c(mestimacion2-1.96*sqrt(ddd),mestimacion2+1.96*sqrt(ddd)))
[1] 1737.703 3559.705
```

```

> #pi estimador
> (ht<-124/40*sum(mues[,1]))
[1] 2580.44
> (varht<-N^2*((1-(n/N))/n)*var(mues[,1]))
[1] 189657.8
> (intht<-c(ht-1.96*sqrt(varht),ht+1.96*sqrt(varht)))
[1] 1726.865 3434.015
>
> #greg
> lmr<-lm(P83C~IMP+EXP+GNP+P80,data = mues)
> summary(lmr)

Call:
lm(formula = P83C ~ IMP + EXP + GNP + P80, data = mues)

Residuals:
Min       1Q   Median       3Q      Max
-2.6404 -0.3596 -0.1237  0.1459  5.4904

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.778e-02  3.038e-01  0.124  0.90174
IMP           1.852e-04  5.178e-05  3.577  0.00104 **
EXP          -4.053e-05  5.768e-05 -0.703  0.48698
GNP          -3.608e-04  4.664e-05 -7.736  4.4e-09 ***
P80           1.097e+00  1.176e-02  93.237 < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.416 on 35 degrees of freedom
Multiple R-squared:  0.9975, Adjusted R-squared:  0.9972
F-statistic: 3534 on 4 and 35 DF,  p-value: < 2.2e-16

>
> a<-sum(m%*%lmr$coefficients)
> b<-sum((mues[,1]-(lmr$fitted.values))/(40/124))
> (greg<-a+b)
[1] 4534.114
> (vargreg<-N^2*((1-(n/N))/n)*var(mues[,1])*(1-0.7694))
[1] 43735.1
> (intgreg<-c(greg-1.96*sqrt(vargreg),greg+1.96*sqrt(vargreg)))
[1] 4124.220 4944.007

```

g La Tabla 3.1 muestra las anteriores estimaciones, sus varianzas y el intervalo de confianza para la estimación.

Estimador	Estimación	Varianza	Int. Confianza
$\hat{t}_{sin}$	4618.77	219.716	(4588.748, 4648.792)
$\hat{t}_{\pi TS}$	2648.704	216035.7	(1737.703, 3559.705)
$\hat{t}_{\pi y}$	2580.44	189657.8	(1726.965, 3434.015)
$\hat{t}_{yr}$	2534.114	40560.46	(2141.980, 2931.454)

TABLA 3.1. Estimaciones del valor de la variable P83 mediante los estimadores  $\hat{t}_{sin}$ ,  $\hat{t}_{\pi TS}$ ,  $\hat{t}_{\pi y}$  y  $\hat{t}_{yr}$ . Estimación de la varianza y del intervalo de confianza.

En este ejemplo, a pesar de que no se cumplen los supuestos para el modelo de regresión Tobit, el estimador dado  $\hat{t}_{sin} = 4618.77$  produjo la estimación más cercana al valor real de la variable que es de 4544.8.

## 3.2. Ejemplo de Aplicación 2.

El presente ejemplo se desarrollará con el mismo conjunto de datos trabajados en la sección anterior pero variando el diseño muestral  $\mathbf{p}(\cdot)$ . Para este ejemplo se usará un diseño estratificado MAS o ESTMAS.

### 3.2.1. Estimadores Bajo Muestreo ESTMAS

Antes de aplicar los estimadores a los datos veamos primero la forma particular de los estimadores. Bajo un diseño muestral ESTMAS, el estimador Tobit Sintético (2.2.12) del total de la variable latente  $t_{y^*}$  es

$$\begin{aligned}
 \hat{t}_{sin} &= \sum_{\mathbb{U}} y_k^{*'} \\
 &= \sum_{h=1}^H \hat{t}_{sinh} \\
 &= \sum_{h=1}^H y_{kh}^{*'} \\
 &= \sum_{h=1}^H \mathbf{x}'_{k \cdot h} \hat{\beta}_h
 \end{aligned} \tag{3.2.1}$$

donde  $\hat{t}_{sinh} = \sum_{\mathbb{U}_h} y_k^{*'}$  es el estimador del total del estrato  $h$ . Se tiene que  $\hat{\beta}_h$  será la estimación de los parámetros del modelo de regresión Tobit para los datos de la muestra para cada estrato  $h$ . Es decir que para cada estrato  $h$  se tendrá una estimación  $\hat{\beta}_h$  de los parámetros  $\beta$ . De manera análoga, se tiene que el  $\pi$ -estimador Tobit Sintético dado por (2.2.15) y bajo un diseño muestral ESTMAS será de la forma

$$\hat{t}_{\pi TS} = \sum_s \check{y}_k^{*'}$$

$$\begin{aligned}
&= \sum_{h=1}^H \hat{t}_{\pi TSh} \\
&= \sum_{h=1}^H \frac{y_{kh}^*}{\pi_k} \\
&= \sum_{h=1}^H \frac{\mathbf{x}'_{k \cdot h} \hat{\beta}_h}{\pi_k}
\end{aligned} \tag{3.2.2}$$

donde  $\hat{t}_{\pi TSh} = \sum_{\mathbb{U}_h} \frac{y_k^*}{\pi_k}$  es el  $\pi$ -estimador Tobit Sintético para el estrato  $h$ .  $\pi_k = n_h/N_h$ , con  $n_h$  como el tamaño de cada estrato en la muestra  $s$  y  $N_h$  el tamaño del estrato en la población  $\mathbb{U}$ . Dado lo anterior se tiene que bajo un ESTMAS el  $\pi$ -estimador Tobit Sintético es

$$\hat{t}_{\pi TS} = \sum_{h=1}^H N_h \bar{y}_{sh}^* \tag{3.2.3}$$

con  $\bar{y}_{sh}^* = \sum_{s_h} y_k^*/n_h$ .

### 3.2.2. Procedimiento

Del ejemplo anterior, que consideraba un muestreo aleatorio simple, se mantienen sin modificación los literales (a) y (b) del procedimiento. El conjunto de datos *CO124* tiene una variable categórica que representa los continentes a los cuales pertenecen cada uno de los países. Se usará esta variable para estratificar la población y a partir de ahí realizar el proceso de estimación. Partiendo de los literales (a) y (b) tenemos

- c Las categorías 6 y 7 de la variable CONT tienen en total sólo 4 países. Junta-mos estas categorías en la categoría 4 correspondiente al continente asiático. De esta manera el total de países de la base de datos quedarán agrupados en cinco estratos.

```

> datos3<-data.frame(cbind(datos$P83,P83C,datos$IMP,
datos$EXP,datos$GNP,datos$P80,datos$CONT))
> colnames(datos3)<-c("P83","P83C","IMP","EXP","GNP","P80",
"CONT")
> datos3$CONT<-as.factor(datos3$CONT)
> # Level is the stratifying variable
> summary(datos3$CONT)
 1  2  3  4  5  6  7
38 14 11 33 24  3  1
> levels(datos3$CONT)[c(6,7)]<-c("4","4")
> levels(datos3$CONT)
[1] "1" "2" "3" "4" "5"

```

- d Calculamos el tamaño de cada estrato en la población y con base en esto determinamos la asignación óptima del tamaño de muestra. Para este ejemplo optaremos por determinar que el tamaño de la muestra  $n = 50$ .

```
> N1<-summary(datos3$CONT)[[1]]
> N2<-summary(datos3$CONT)[[2]]
> N3<-summary(datos3$CONT)[[3]]
> N4<-summary(datos3$CONT)[[4]]
> N5<-summary(datos3$CONT)[[5]]
> N1;N2;N3;N4;N5
[1] 38
[1] 14
[1] 11
[1] 37
[1] 24
> Nh <- c(N1,N2,N3,N4,N5)
> (nh<-ceiling(50*Nh/124))
[1] 16 6 5 15 10
```

- e Con base en estos datos podemos extraer la muestra acorde con los tamaños establecidos.

```
> library(TeachingSampling)
> sam <- S.STSI(datos3$CONT, Nh, nh)
> data <- datos3[sam,]
```

- f Con la muestra seleccionada, se puede proceder a calcular los estimadores. El siguiente bucle for esta diseñado para calcular el modelo a lo largo de los diferentes estratos hacer las estimaciones necesarias.

```
> st<-levels(datos3$CONT)
> ests1<-0
> espits<-0
> for(i in 1:length(st)){
+   a<-subset(data[data$CONT==st[i],])
+   if(sum(a$P83C==100)>0){
+     mrt<-censReg(P83C~IMP+EXP+ GNP+ P80,data = a,
right = 100)
+     coefi<-mrt$estimate[-length(mrt$estimate)]
+   }else{
+     mrt<-lm(P83C~IMP+EXP+ GNP+ P80,data = a)
+     coefi<-mrt$coefficients
+   }
+   #estimador tobit sintético
+   matdis<-model.matrix(P83C~IMP+EXP+ GNP+ P80,
```

```

data = datos3[datos3$CONT==i,]
+ ests1<-ests1+sum(matdis%*%coefi)
+ #pi-estimador tobit sintético
+ matdis2<-model.matrix(P83C~IMP+EXP+ GNP+ P80,data = a)
+ espits<-espits+Nh[i]*mean(matdis2%*%coefi)
+ }
> ests1
[1] 4509.816
> espits
[1] 3713.577

```

El bucle calcula para cada estrato en la muestra un vector de parámetros  $\hat{\beta}$ . Usa la información auxiliar disponible para calcular los valores  $y_k^*$ . Dependiendo de si se usa el estimador Tobit Sintético o el  $\pi$ -estimador Tobit Sintético, la información auxiliar corresponderá a toda la información disponible para el estrato  $h$  o corresponderá solamente con la información auxiliar disponible en la muestra.

- g El valor del estimador  $\hat{t}_{sin} = 4509.86$  y el valor del estimador  $\hat{t}_{\pi TS} = 3713.577$ . Nuevamente, el estimador más cercano con el verdadero valor de la variable latente es el estimador Tobit Sintético.



# CONCLUSIONES Y TRABAJO FUTURO

## 4.1. Conclusiones

De lo expuesto en los capítulos anteriores podemos concluir lo siguiente

1. Como solución al problema de estimación del parámetro total en muestras con observaciones censuradas se encontró que el estimador dado en la ecuación (2.2.12) y que tiene la forma

$$\hat{t}_{sin} = \hat{t}_{y_k^*} = \sum_{\cup} y_k^*$$

es el mejor estimador posible. Bajo simulación se encontró que en comparación con los demás estimadores propuestos, el estimador  $\hat{t}_{sin}$  tiene el sesgo, la varianza y el error cuadrático medio más bajo y además es bastante estable ante la variación del porcentaje de censura de dentro de la muestra. Sin embargo no es un estimador insesgado.

2. El estimador dado en la ecuación (2.2.15) y que tiene la forma

$$\hat{t}_{\pi TS} = \hat{t}_{y_k^*} = \sum_s \frac{y_k^*}{\pi_k}$$

es una opción aceptable como solución al problema de estimación. Bajo simulación mostró que conforme aumenta el tamaño de la muestra el sesgo, la varianza y el error cuadrático medio disminuían. Sin embargo mostró un comportamiento poco estable frente a la variación del porcentaje de censura dentro de la muestra. A pesar de que este estimador tampoco es insesgado la tasa de cobertura estimada muestra que los intervalos de confianza calculados a partir de este estimador tienen mejor cobertura del valor real del total de la variable latente.

3. La escogencia del estimador depende enteramente de la disponibilidad de información auxiliar para toda la población objetivo. Es decir, la aplicabilidad

de la solución óptima depende enteramente de la existencia de la información auxiliar para la población.

4. La aplicabilidad de estos estimadores depende del cumplimiento de supuestos muy fuertes. Se debe cumplir supuestos relacionados con el modelo, distribución de la variable latente, varianza constante de las variables, poseer información auxiliar para toda la población y conocer el porcentaje máximo de censura en la población para determinar si la estimación esta más cerca del valor real del parámetro.

## 4.2. Trabajo Futuro

De lo desarrollado en el presente trabajo quedan pendientes los siguientes aspectos

1. Siempre se trabajó bajo el supuesto de normalidad en los errores, independencia y homocedasticidad en el modelo que asiste al estimador propuesto. Queda pendiente entonces estudiar el comportamiento del estimador bajo la modificación de algunos de estos supuestos. Especialmente queda pendiente trabajar con el modelo Tobit cuando no se tiene el supuesto de varianza constante dado que en este caso se pierde la propiedad de consistencia en los estimadores  $\hat{\beta}$  y por tanto se afectará las propiedades de los estimadores del total. Algunos autores que han presentado desarrollos sobre el modelo Tobit en presencia de heterocedasticidad son Hurd (1979); Arabmazar & Schmidt (1981); Dooley (1983).
2. En la literatura existen otro tipo de modelos de regresión propuestos para trabajar muestras con datos censurados que presentan comportamiento distintos al normal. Cohen (1991) presenta varios casos de muestras censuradas que proviene de conjuntos de datos para los cuales la variable de interés sigue distribuciones probabilísticas distintas a la distribución normal. Específicamente, en su libro Cohen trabaja los casos para distribuciones lognormal, inversa gaussiana, gamma, exponencial, Rayleigh, Pareto, normales bivariadas y multivariadas, y distribuciones discretas. Queda pendiente la exploración del problema de estimación haciendo uso de otro tipo de modelos de regresión que siguen distribuciones distintas a la normal.
3. Bajo simulación se encontró una relación inversa entre el porcentaje de censura y la exactitud del estimador. Queda pendiente hacer una revisión analítica de esta relación y determinar su influencia en el estimador. Por ejemplo, si esta relación puede ser usada para construir una medida de precisión del estimador o para mejora su exactitud.

# Apéndice **A**

## Distribución de los Estimadores Simulados

En el presente apéndice se presenta un conjunto de histogramas de frecuencia que muestran la distribución de los 10 estimadores propuestos a lo largo de las cuatro ejecuciones de la simulación presentada en el Capítulo 2.

Para todo el conjunto de gráficas la línea roja representa el valor del total de la variable latente en la simulación que es igual a  $t_{y^*} = 6998.0904$  y la línea negra representa el valor del total de la variable censurada  $t_y = 7106.6347$ . El orden de las Figuras corresponde con el presentado en la Tabla 2.3 del Capítulo 2.

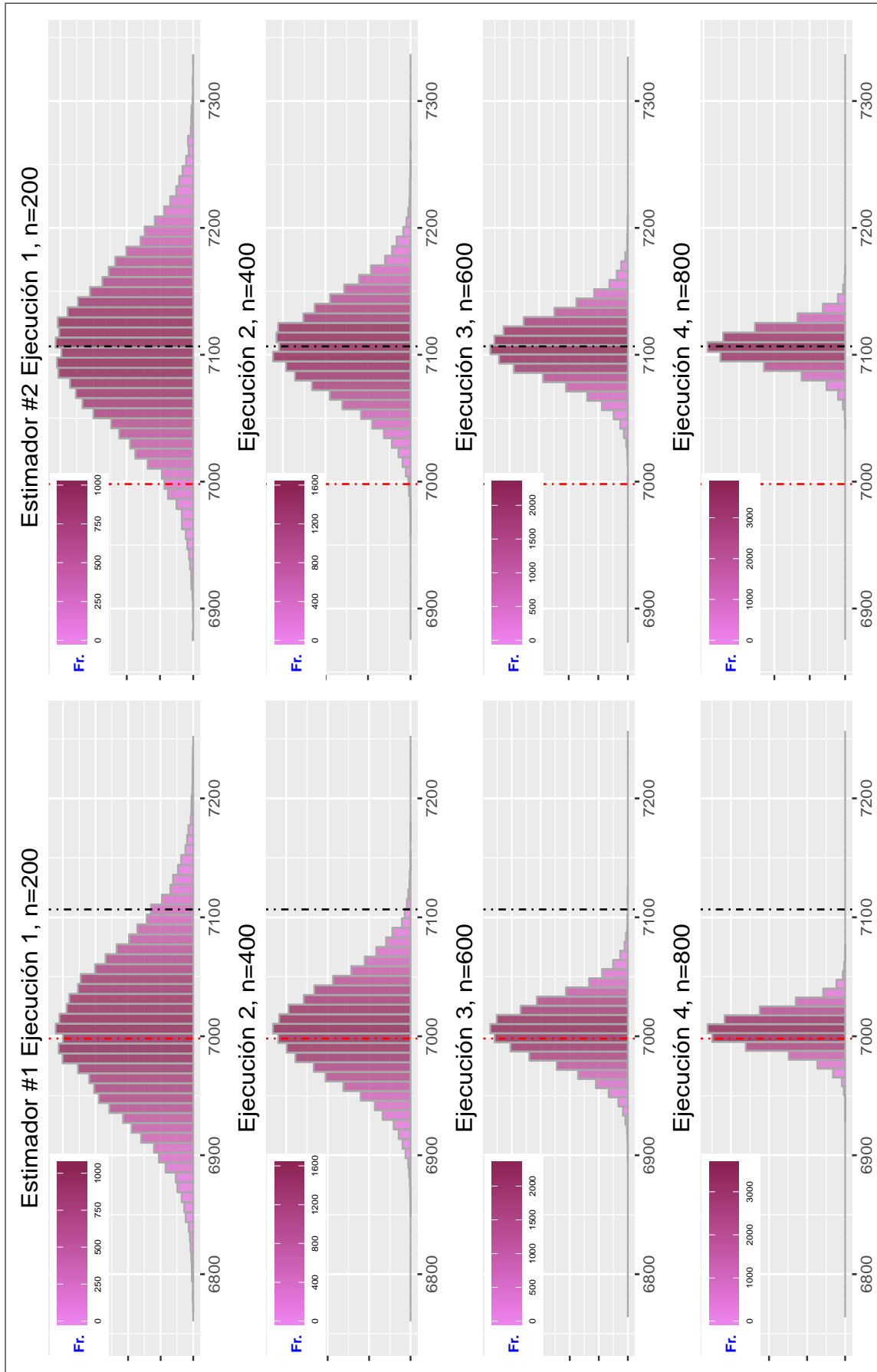


FIGURA A.1. Histograma de frecuencias para los estimadores (1) y (2) de la Tabla 2.3.

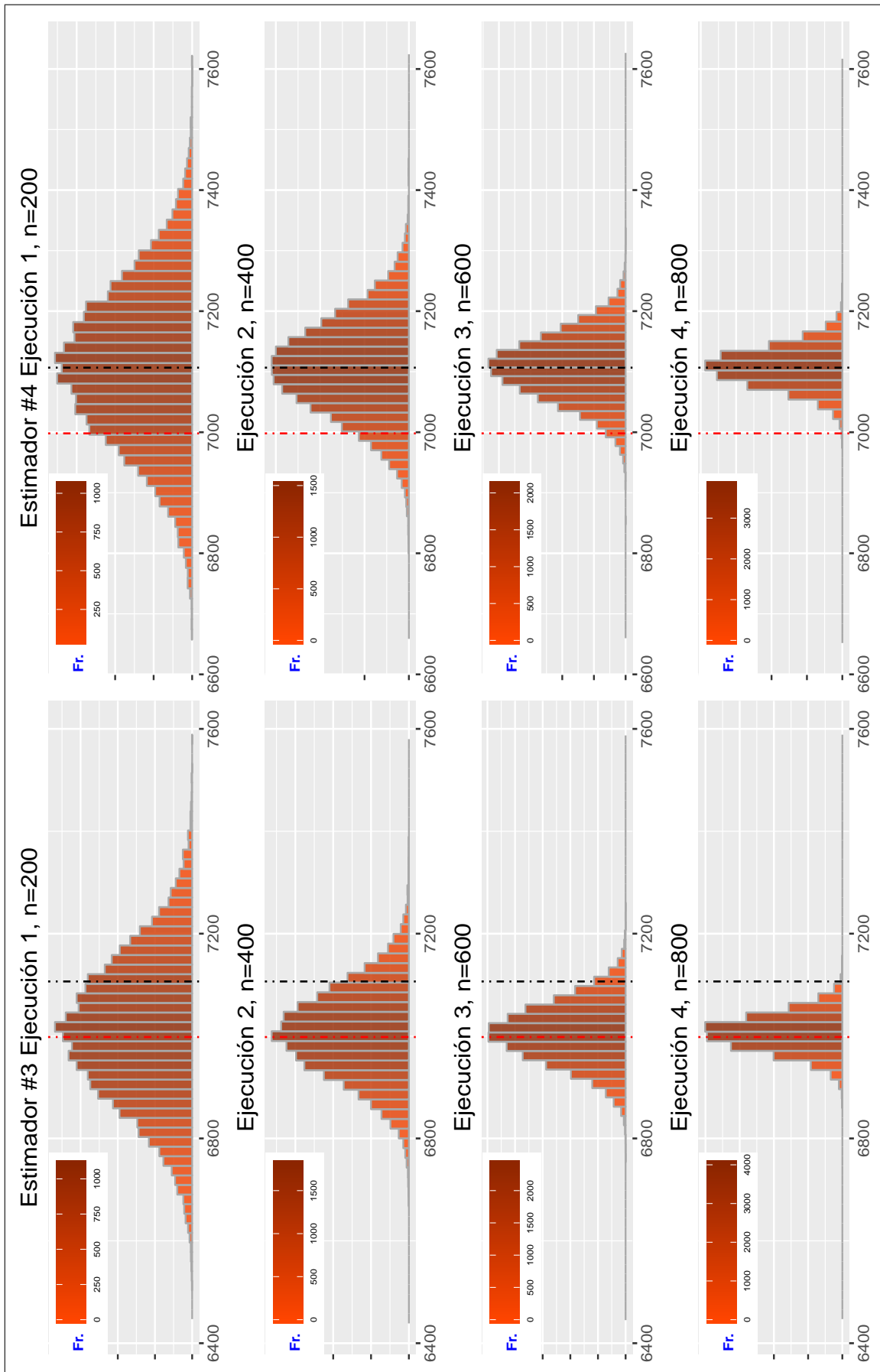


FIGURA A.2. Histograma de frecuencias para los estimadores (3) y (4) de la Tabla 2.3.

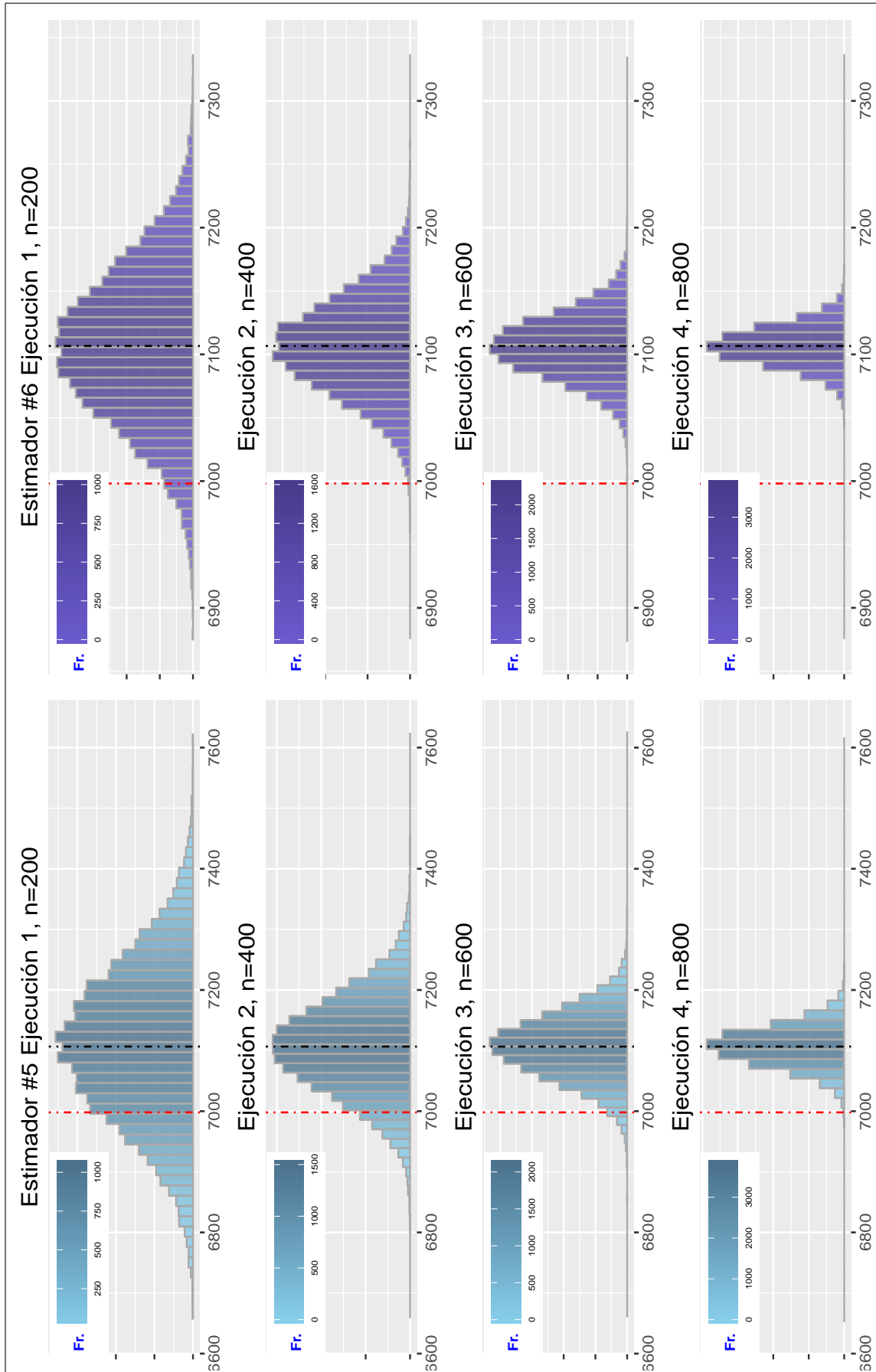


FIGURA A.3. Histograma de frecuencias para los estimadores (5) y (6) de la Tabla 2.3.

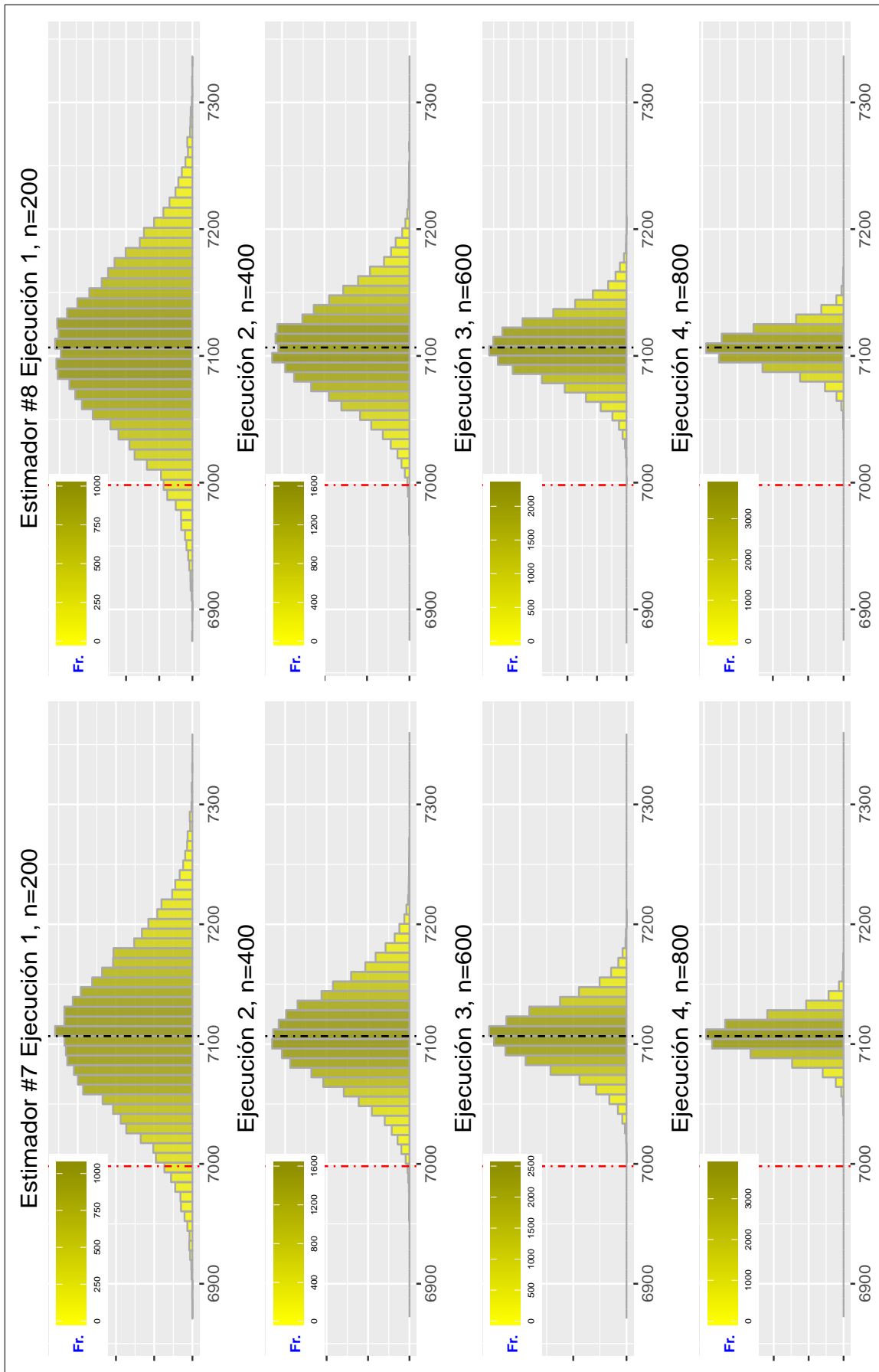


FIGURA A.4. Histograma de frecuencias para los estimadores (7) y (8) de la Tabla 2.3.

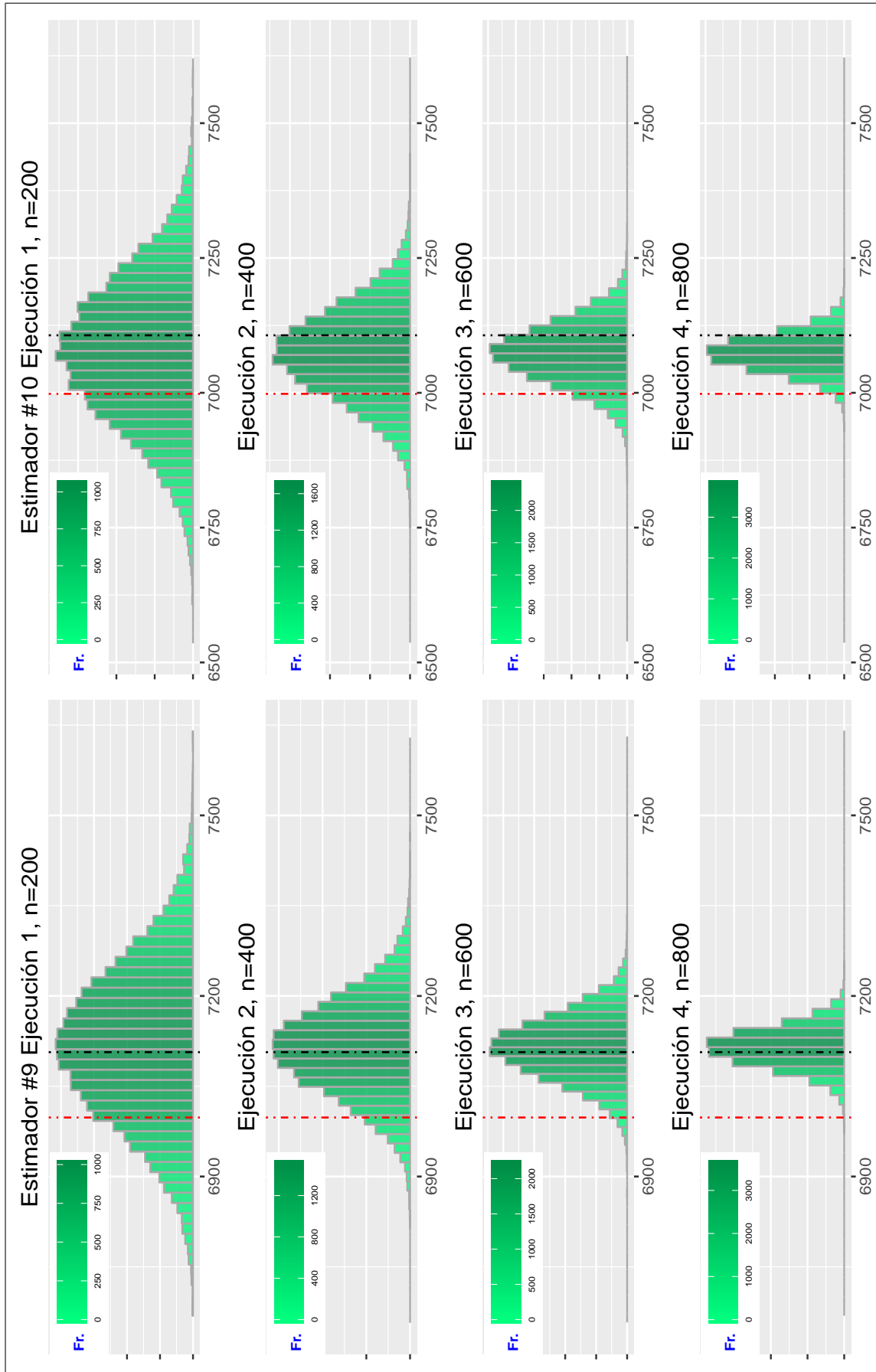


FIGURA A.5. Histograma de frecuencias para los estimadores (9) y (10) de la Tabla 2.3.

# Apéndice **B**

## Conjuntos De Datos

### B.1. ‘Población CO124’

**LABEL** Identificador de los individuos de 1 a 124.

**COUNTRY** Las abreviaciones, con algunas excepciones, son las de los nombres de los países del Comité Olímpico Internacional.

**P83** Población en 1983 (en millones).

**IMP** Importaciones en 1983 (en millones de dolares).

**EXP** Exportaciones en 1983 (en millones de dolares).

**GNP** Producto interno bruto en 1982 (en decenas de millones de dolares).

**MEX** Gastos militares en 1981 (en millones de dolares).

**P80** Población en 1980 (en millones)

**CONT** 1=África, 2=Norte y Centro América, 3=Sur América, 4=Asia(Sin la parte Soviética), 5=Europa (Sin la parte Soviética), 6=Oceanía y 7=Unión de Repúblicas Socialistas Soviéticas

LABEL	COUNTRY	P83	IMP	EXP	GNP	MEX	P80	CONT
1	ALG	20.5	10395	11163	4535	675	18.7	1
2	BEN	3.7	131	63	122	23	3.4	1
3	BUR	6.6	288	57	135	41	6.1	1
4	BRI	4.4	194	76	122	23	4.1	1
5	CMR	9.1	1217	940	715	82	8.5	1
6	CAF	2.4	127	109	76	12	2.3	1
7	CHA	4.7	109	58	36	62	4.5	1

---

8	CGO	1.6	807	977	234	68	1.5	1
9	EGY	45.9	10274	3215	3425	1650	42.3	1
10	ETH	33.6	875	403	465	485	31.1	1
11	GAB	1.1	724	2161	330	72	1	1
12	GHA	12.2	705	873	416	110	11.5	1
13	GUI	5.1	351	428	172	44	4.8	1
14	CIV	9.3	1808	2067	822	111	8.2	1
15	KEN	18.7	1274	876	698	183	16.7	1
16	LBR	2	422	464	99	16	1.8	1
17	LBA	3.3	8382	15576	2712	3670	3	1
18	MAD	9.7	540	316	299	71	8.7	1
19	MAW	6.4	312	230	136	22	6	1
20	MLI	7.5	344	167	123	46	7	1
21	MTN	1.8	227	305	81	82	1	1
22	MRI	.9	438	373	123	2	.9	1
23	MAR	21.1	3599	2062	1758	1005	20.1	1
24	MOZ	13.3	635	132	465	111	12.1	1
25	NIG	5.7	442	333	178	16	5.3	1
26	NGR	89	13440	11317	7722	2037	80.6	1
27	RWA	5.7	279	79	144	18	5.2	1
28	SLE	3.4	171	119	123	19	3.3	1
29	SOM	5.2	330	199	212	150	4.6	1
30	SAF	31.8	14528	9671	8234	2254	28.6	1
31	SUD	20.3	1354	624	935	470	18.7	1
32	TOG	2.7	391	177	95	22	2.5	1
33	TUN	6.8	3117	1872	922	214	6.4	1
34	UGA	14.6	293	345	325	852	13.2	1
35	TAN	20.3	4552	1900	535	285	18.6	1
36	ZAI	31.1	480	569	558	164	26.4	1
37	ZAM	6.2	690	831	386	290	5.8	1
38	ZIM	7.7	1430	1115	640	440	7.1	1
39	CAN	24.9	61325	73797	27909	4227	24	2
40	CRC	2.4	993	867	266	19	2.2	2
41	CUB	10	6293	5536	1740	1065	9.7	2
42	DOM	5.9	1282	811	767	92	5.4	2
43	ESA	5.2	891	735	356	86	4.7	2
44	GUA	7.9	1126	1184	870	95	6.9	2
45	HAI	5.3	461	154	158	22	5	2
46	HON	4	767	692	262	38	3.7	2
47	JAM	2.2	1531	745	300	29	2.2	2
48	MEX	75.1	8201	21012	19635	782	69.3	2
49	NCA	3	799	411	262	34	2.7	2
50	PAN	2	1412	304	415	22	2	2
51	TRI	1.1	2505	2379	772	12	1.1	2
52	USA	233.9	269878	200538	305690	134390	227.7	2
53	ARG	29.6	4486	7836	8949	2241	28.2	3

54	BOL	6	532	789	539	84	5.6	3
55	BRA	129.6	14494	26906	27474	1234	121.3	3
56	CHI	11.6	2754	3836	2516	225	11.1	3
57	COL	27.5	4968	3081	3819	229	27.1	3
58	ECU	9.2	1465	2203	1288	92	8.1	3
59	GUY	.9	283	256	47	27	.9	3
60	PAR	3.4	506	284	573	40	3.2	3
61	PER	18.7	2688	3015	2204	480	17.3	3
62	URU	2.9	788	1045	1001	150	2.9	3
63	VEN	15.1	6667	15002	6907	527	15	3
64	AFG	14.5	695	708	315	85	14.5	4
65	BRN	.4	3342	3200	375	115	.3	4
66	BAN	94.6	1716	690	1288	140	88.7	4
67	BIR	35.3	268	378	650	225	33.6	4
68	CHN	1024	21324	22151	30250	37200	1002.8	4
69	CYP	.6	1219	494	239	19	.6	4
70	PRK	19.1	1880	1520	1750	3424	17.9	4
71	YMD	2.1	1527	779	93	115	2	4
72	IND	730	13562	8304	18413	3991	663.6	4
73	INA	159.4	16352	21146	8945	1426	146.4	4
74	IRN	42	11539	19438	6970	5092	38.3	4
75	IRQ	14.6	20500	10230	2710	3759	13.2	4
76	ISR	4	8587	5112	2121	2750	3.9	4
77	JPN	117.1	126395	146676	119000	9461	116.8	4
78	JOR	3.2	3030	579	419	420	2.9	4
79	KUW	1.6	6980	11140	3044	2031	1.4	4
80	LAO	4.2	125	33	32	21	3.9	4
81	MAL	15.2	13987	13917	2681	1639	13.9	4
82	MGL	1.8	655	436	165	238	1.6	4
83	NEP	15.7	342	80	247	28	14	4
84	OMA	1.1	2492	4058	687	1444	1	4
85	PAK	92.9	5341	3149	3302	1307	82.6	4
86	PHI	51.9	7980	5005	4168	688	48.1	4
87	QAT	.2	1456	3384	595	893	.2	4
88	KOR	40.5	26192	24445	7509	3519	38.1	4
89	KSA	10.4	39206	46941	15638	22458	9.2	4
90	SIN	2.5	28712	24108	1478	556	2.4	4
91	SRI	15.4	1786	1123	491	35	14.7	4
92	SYR	10.4	4542	1900	1589	2166	8.7	4
93	THA	49.4	9159	6368	3853	1036	46.5	4
94	TUR	47.2	9348	5694	6315	3442	44.4	4
95	UAE	1.1	9414	17257	2917	1423	1	4
96	YAR	6.2	1521	39	371	320	5.8	4
97	ALB	2.8	250	200	266	127	2.6	5
98	AUT	7.5	19368	15428	7447	847	7.5	5
99	BEL	9.8	55269	51929	10388	3690	9.8	5

---

100	BUL	8.9	12164	12130	4603	964	8.9	5
101	TCH	15.4	16325	16522	8514	2900	15.3	5
102	DEN	5.1	16946	16221	6314	1546	5.1	5
103	FIN	4.8	12854	12530	5235	632	4.8	5
104	FRA	55	103734	93310	62731	23633	53.7	5
105	GDR	16.7	21254	23793	11727	4394	16.7	5
106	FRG	61	152899	169425	75709	25509	61.6	5
107	GRE	9.8	9632	4459	4089	2184	9.6	5
108	HUN	10.6	8503	8696	5542	810	10.7	5
109	IRL	3.5	9182	8612	1757	246	3.4	5
110	ITA	56.8	80367	72681	38223	8184	56.4	5
111	HOL	14.3	60743	64816	15428	4931	14.1	5
112	NOR	4.1	13890	18920	5884	1484	4.1	5
113	POL	36.5	10179	11478	14561	2467	35.6	5
114	POR	10.1	8134	4566	2472	779	9.9	5
115	ROM	22.7	9836	11714	7109	1285	22.2	5
116	ESP	38.2	28812	23544	20424	3682	37.4	5
117	SWE	8.3	25046	26313	11541	3175	8.3	5
118	SUI	6.4	29475	25865	10856	2000	6.4	5
119	GBR	55.7	105477	94562	53673	19901	55.9	5
120	YUG	22.8	11104	9038	7053	2936	22.3	5
121	AUS	15.3	19393	20594	16904	3508	14.7	6
122	FIJ	.6	484	240	128	4	.6	6
123	NZL	3.3	5283	5272	2539	393	3.1	6
124	URS	277.4	80410	91336	156300	118800	265.5	7

---

TABLA B.1. Datos del libro de (Särndal et al., 1992), *The CO124 Population*

## Código de Simulación

A continuación se presentará el código implementado para la realización de las simulaciones y la modificación de la función para calcular el modelo de regresión Tobit que incluye pesos muestrales.

### C.1. Información Técnica

**Lenguaje R** Todas las simulaciones fueron diseñadas bajo el lenguaje R (R Core Team, 2016) con las siguientes especificaciones de sistema

```
> sessionInfo()
R version 3.3.2 (2016-10-31)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 8 (jessie)
nickname      Sincere Pumpkin Patch
```

e implementadas a través de la interfaz gráfica de RStudio Version 1.0.44 | 2009-2016 RStudio, Inc. (RStudio Team, 2016).

#### Paquetes de R

##### 1. maxLik

```
> packageDescription("censReg")
Package: censReg
Version: 0.5-20
Date: 2013/08/20
Title: Censored Regression (Tobit) Models
Author: Arne Henningsen <arne.henningsen@gmail.com>
Description: Estimation of censored regression
```

(Tobit) models with cross-section and panel data

## 2. censReg

```
> packageDescription("maxLik")
Package: maxLik
Version: 1.3-4
Date: 2015-11-08
Title: Maximum Likelihood Estimation and Related
Tools
Author: Ott Toomet <otoomet@gmail.com>, Arne
Henningsen <arne.henningsen@gmail.com>, with
contributions from Spencer Graves and Yves
Croissant
Description: Functions for Maximum Likelihood (ML)
estimation and non-linear optimization, and
related tools. It includes a unified way to
call different optimizers, and classes and
methods to handle the results from the ML
viewpoint. It also includes a number of
convenience tools for testing and developing
your own models.
```

## 3. optimx

```
Package: optimx
Version: 2013.8.7
Date: 2013-08-07
Title: A Replacement and Extension of the optim()
Function
Author: John C Nash [aut, cre], Ravi Varadhan
[aut], Gabor Grothendieck [ctb]
Description: Provides a replacement and extension
of the optim() function to unify and
streamline optimization capabilities in R for
smooth, possibly box constrained functions of
several or many parameters. This is the CRAN
version of the package.
```

## 4. survey

```
Package: survey
Title: analysis of complex survey samples
Description: Summary statistics, two-sample tests,
rank tests, generalised linear models,
```

cumulative link models, Cox models, loglinear models, and general maximum pseudolikelihood estimation for multistage stratified, cluster-sampled, unequally weighted survey samples. Variances by Taylor series linearisation or replicate weights. Post-stratification, calibration, and raking. Two-phase subsampling designs. Graphics. PPS sampling without replacement. Principal components, factor analysis.  
Version: 3.30-3  
Author: Thomas Lumley  
Maintainer: "Thomas Lumley"  
<t.lumley@auckland.ac.nz>

## 5. **sampling**

Package: sampling  
Version: 2.7  
Date: 2015-06-30  
Title: Survey Sampling  
Author: Yves Tillé <yves.tille@unine.ch>, Alina Matei <alina.matei@unine.ch>  
Maintainer: Alina Matei <alina.matei@unine.ch>  
Description: Functions for drawing and calibrating samples.

## 6. **TeachingSampling**

Package: TeachingSampling  
Type: Package  
Title: Selection of Samples and Parameter Estimation in Finite Population  
Version: 3.2.2  
Date: 2015-04-28  
Author: Hugo Andres Gutierrez Rojas  
<hugogutierrez@usantotomas.edu.co>  
Maintainer: Hugo Andres Gutierrez Rojas  
<hugogutierrez@usantotomas.edu.co>  
Description: Allows the user to draw probabilistic samples and make inferences from a finite population based on several sampling designs.

## C.2. Código

### C.2.1. Función que Permite Calcular el Modelo Tobit

```

###(1)logaritmo de la función de verosimilitud con la
#transformación de olsen

lftvb.olsen<-function(start,x,y,n,cen_izq,sin_cen){

##parámetros
b <- start[-length(start)]; sig <- start[length(start)]
alf <- b/sig; h <- 1/sig

##datos
xa<-x%*%alf
hyxa<-h*y-xa
ll<-rep(NA,n)
grad<-matrix(NA, ncol = length(start), nrow = length(y))

##logaritmo de la función de verosimilitud
ll[cen_izq]<- pnorm(-xa[cen_izq],log.p = T)
ll[sin_cen]<- log(h) - (1/2) * ((hyxa[sin_cen])^2)
ll<-sum(ll)

##vector gradiente
grad[cen_izq] <- cbind(exp(dnorm(-xa[cen_izq], log = T) -
pnorm(-xa[cen_izq], log.p = T))*(-x[cen_izq, , drop=F]), 0)
grad[sin_cen] <- cbind(hyxa[sin_cen] * x[sin_cen, ,drop = F],
(1/h - hyxa[sin_cen] * y[sin_cen]))
attr(ll,"gradient")<- grad <- colSums(grad)
##salida
return(ll)
}

###(2)logaritmo de la función de verosimilitud con la transformación
#de olsen y pesos muestrales

wlfvtb.olsen<-function(start,x,y,n,cen_izq,sin_cen,pik){

##parámetros
b <- start[-length(start)]; sig <- start[length(start)]
alf <- b/sig; h <- 1/sig

##datos
xa<-x%*%alf

```

```

hyxa<-h*y-xa
ll<-rep(NA,n)
grad<-matrix(NA, ncol = length(start), nrow = length(y))

##logaritmo de la función de verosimilitud
ll[cen_izq]<- (1/pik)[cen_izq]*pnorm(-xa[cen_izq],log.p = T)
ll[sin_cen]<-((1/pik)[sin_cen]*log(h)-(1/(2*pik))[sin_cen]
*(hyxa[sin_cen])^2)
ll<-sum(ll)

##vector gradiente
grad[cen_izq]<-cbind((1/pik)[cen_izq]
*exp(dnorm(-xa[cen_izq], log = T)
- pnorm(-xa[cen_izq], log.p = T)) * (-x[cen_izq, , drop=F]), 0)
grad[sin_cen]<-cbind((1/pik)[sin_cen]*hyxa[sin_cen]
* x[sin_cen, ,drop = F], (1/(h*pik))[sin_cen] - (1/pik)[sin_cen]
*hyxa[sin_cen] * y[sin_cen])
attr(ll,"gradient")<- grad <- colSums(grad)
##salida
return(ll)
}

###(3)Logaritmo de la función de vorosimilitud

lfvtb<-function(start,x,y,n,cen_izq,sin_cen){

##parámetros
b <- start[-length(start)]; sig <- exp(start[length(start)])

##datos
xb<-x%*%b
yxb <- y-xb
ll<-rep(NA,n)
grad<-matrix(NA, ncol = length(start), nrow = length(y))

##logaritmo de la función de verosimilitud
ll[cen_izq] <- pnorm(-xb[cen_izq],log.p = T)
ll[sin_cen] <- (-1/2)*log(sig^2)-(1/(2*sig^2))*(yxb[sin_cen])^2
ll<-sum(ll)

##vector gradiente
grad[cen_izq] <- exp(dnorm(-xb[cen_izq]/sig, log = T)
- pnorm(-xb[cen_izq]/sig, log.p = T))
* cbind(-x[cen_izq, , drop=F]/sig, xb[cen_izq]/(2*sig^3))
grad[sin_cen] <- cbind((1/(sig^2))*yxb[sin_cen]*x[sin_cen, , drop=F],
(-1/(2*sig^2))+(1/(2*sig^4))*yxb[sin_cen]^2)

```

```

attr(ll,"gradient") <- grad <- colSums(grad)
##salida
return(ll)
}

###(4)logaritmo de la función de verosimilitud con pesos muestrales

wlfvtb<-function(start,x,y,n,cen_izq,sin_cen,pik){

##parámetros
b <- start[-length(start)]; sig <- exp(start[length(start)])

##datos
xb<-x%*%b
yxb <- y-xb
ll<-rep(NA,n)
grad<-matrix(NA, ncol = length(start), nrow = length(y))

##logaritmo de la función de verosimilitud
ll[cen_izq] <- (1/pik)[cen_izq] * pnorm(-xb[cen_izq]/sig,log.p=T)
ll[sin_cen] <- ((-1/(2*pik))[sin_cen]*log(sig^2)
- (1/(2*pik*sig^2))[sin_cen]*(yxb[sin_cen])^2)
ll<-sum(ll)

##vector gradiente
grad[cen_izq,]<- exp(dnorm(-xb[cen_izq]/sig,log=T)
-pnorm(-xb[cen_izq]/sig,log.p=T))
*cbind((-1/(sig*pik))[cen_izq]*
x[cen_izq, ,drop=F],(1/(2*pik*sig^3))[cen_izq]*xb[cen_izq])
grad[sin_cen,]<- cbind((1/(pik*sig^2))[sin_cen]
*yxb[sin_cen]*x[sin_cen, ,drop=F],(-1/(2*sig^2*pik))[sin_cen]
+(1/(2*sig^4*pik))[sin_cen]*(yxb[sin_cen])^2)

attr(ll,"gradient")<-grad <-colSums(grad)
return(ll)
}

#función para maximizar el modelo tobit

tob<-function(fml,dts,lc=0, weight=NULL ,wstart=FALSE,
fun.ver="lfvtb"){
#requerimientos
require(maxLik)
##condiciones de control
if(class(fml)!="formula"){

```

```
stop("El argumento 'formula' debe ser una formula")
}
if(is.data.frame(dts)!=TRUE){
warning("Preferiblemente ingrese un objeto del tipo 'data.frame'.
  Los datos automaticamente se transformaran en un objeto del tipo
  'data.frame'.")
dts<-as.data.frame(dts)
}
##variables internas o de trabajo

# if(is.null(weight)){
#   mf<-model.frame(fml,data=dts)
#   } else {
#   mf<-model.frame(fml,data=dts,weights=weight)
#   w<- model.weights(mf)
#   }
mf<-model.frame(fml,data = dts)

if(lc!=0){
y<-model.response(mf)-lc
} else {
y<-model.response(mf)
}

x<-model.matrix(attr(mf,"terms"),data=mf)
n<-dim(mf)[1]

c_izq <- y <= 0
sin_c <- !c_izq

##procedimientos

if(wstart==FALSE){
model<-lm.fit(x,y)
b<-model$coefficients
sig<-log(sqrt(sum(model$residuals^2)/length(model$residuals)))
para<-c(b, sigma = sig)
} else {
if(is.null(weight)){
stop("Para usar el argumento 'wstart=T' debe proporcionar el
arugumento 'weight'.")
} else {
model<-lm.wfit(x, y , w = weight)
b<-model$coefficients
sig<-log(sqrt(sum(model$residuals^2)/length(model$residuals)))
para<-c(b, sigma = sig)
```

```

}
}

#procesos
if(!fun.ver=="lfvtb"){
if(is.null(weight)){
stop("Para usar el parametro 'wlfvtb' debe proporcionar un
\part{ vector de pesos")
} else{}
max<-maxLik(wlfvtb,start = para,x=x,y=y,n=n,cen_izq=c_izq,
sin_cen=sin_c,pik=weight)
}
} else {
max<-maxLik(lfvtb,start = para,x=x,y=y,n=n,cen_izq=c_izq,
sin_cen=sin_c)

}

b<- c(max$estimate[1]+lc,max$estimate[-c(1,length(max$estimate))])
s<- max$estimate[length(max$estimate)]
z<-list(betas=b,sigma=s)
z$estimate<-max$estimate
z$code<-max$code
z$message<- max$message
z$betas<- b
z$sigma<- s
z$x<- x

if(lc!=0){
yesti<- x%*%b
z$y_esti <- yesti
z$residuals<- (y+lc) - yesti
z$y<- y+lc
} else {
yesti<- x%*%b
z$y_esti <- yesti
z$residuals<- y - yesti
z$y<- y
}

z$dt.cen<-sum(ifelse(y==0,1,0))
z$dt.sin.cen<-sum(ifelse(y==0,0,1))
z$start<-para
z$maximum<-max$maximum
z$iterations<-max$iterations
z$call<-match.call()
z
}

```

## C.2.2. Función que Permite Ejecutar la Simulación de Montecarlo (2)

```
sm<-function(formula,data,m,tm,lc=0,metodo,aux,esti=c(seq(1,11)),
guardar=TRUE,nombre="muestra",seed=NULL){
###verificación de parámetros,Paso de control
if(class(formula)!="formula"){
stop("El argumento 'formula' debe ser una formula")
}
if(is.data.frame(data)!=T){
warning("Preferiblemente ingrese un objeto del tipo 'data.frame'.
Los datos automaticamente se forzaran a ser un objeto del tipo
'data.frame'.")
data<-as.data.frame(data)
}
if(!is.numeric(m)){
stop("El argumento 'm' debe ser un número entero positivo.")
}
if(!is.numeric(tm)){
stop("El argumento 'tm' debe ser un número entero positivo.")
}
if(!is.numeric(lc)){
stop("El argumento 'lc' debe ser un número.")
}
if(!is.character(metodo)){
stop("El argumento 'metodo' debe ser una cadena de caracteres que
indica el método de muestreo a usar. 'mas','bernouli','pipt',
'estrat'.
Se usa por defecto 'mas'.")
}
if(metodo=="pipt"){
if(!is.vector(aux)){
!is.numeric(aux)
}
}
if(missing(nombre)){
warning("El argumento 'nombre' está vacío, se usara por defecto el
nombre 'simulacion' como prefijo para los archivos que se generen.")
}

#llama funciones
require(TeachingSampling)
require(sampling)

##variables auxiliares
N <- dim(data)[1]
```

```

mf <- model.frame(formula, data = data)
y <- model.response(mf, type = "any")
x <- model.matrix(attr(mf, "terms"), data=mf)

#creación de matrices para almacenar resultados
nombres <- c("tsin", "wtsin", "pitsin", "wpitsin", "ht", "greg",
"wgreg", "regtob", "wregtob", "parti", "wparti")
estimaciones <- matrix(NA, ncol = length(estimaciones)+1, nrow = m)
colnames(estimaciones) <- c("censura", nombres[estimaciones])
auxiliar <- matrix(NA, ncol = dim(x)[2]*4, nrow = m)
colnames(auxiliar) <- c(paste("A.", colnames(x), sep = ""),
paste("B.", colnames(x), sep = ""),
paste("C.", colnames(x), sep = ""),
paste("D.", colnames(x), sep = ""))

#for que aplica los estimadores
ptm <- proc.time()
for(i in 1 : m) {
#setseed
if(!is.null(seed)) set.seed(seed)
#selección de muestras de acuerdo a método de muestreo
#-----
if(metodo=="mas"){
s <- sample(1:N, tm, replace = F)
muestra <- mf[s,]
piks <- c(rep(tm/N, tm))
xs <- model.matrix(attr(muestra, "terms"), data=muestra)
ys <- model.response(muestra)
muestra<-cbind(muestra, piks)
} else if (metodo=="pipt"){
s <- S.piPS(tm, data[,aux])
muestra <- mf[s[,1],]
piks <- s[,2]
xs <- model.matrix(attr(muestra, "terms"), data=muestra)
ys <- model.response(muestra)
muestra<-cbind(muestra, piks)
} else if (metodo=="bernouli"){
s <- S.BE(N, tm/N)
muestra <- mf[s,]
piks <- c(rep(tm/N, dim(muestra)[1]))
xs <- model.matrix(attr(muestra, "terms"), data=muestra)
ys <- model.response(muestra)
muestra<-cbind(muestra, piks)
}

#cálculo de modelos

```

```

lineal <- lm(formula, data = muestra)
wlineal <- lm(formula, data = muestra, weights = piks)
modtob <- tob(fml=formula, dts=muestra, lc=lc, weight=NULL ,
wstart=FALSE, fun.ver="lfvtb")
wmodtob <- tob(fml=formula, dts=muestra, lc=lc, weight=piks ,
wstart=FALSE, fun.ver="wlfvtb")

auxiliar[i,] <- c(lineal$coefficients, wlineal$coefficients,
modtob$betas, wmodtob$betas)

estimaciones[i,1] <- modtob$dt.cen/tm

#definición de contador
j<-2
#-----

if(any(estí==1)){
tsin <- sum(x%*%modtob$betas)
estimaciones[i,j] <- tsin
j <- j+1
}
if (any(estí==2)){
wtsin <- sum(x%*%wmodtob$betas)
estimaciones[i,j] <- wtsin
j <- j+1
}
if (any(estí==3)){
pitsin <- sum(modtob$y_estí/piks)
estimaciones[i,j] <- pitsin
j <- j+1
}
if (any(estí==4)){
wpitsin <- sum(wmodtob$y_estí/piks)
estimaciones[i,j] <- wpitsin
j <- j+1
}
if (any(estí==5)) {
ht <- E.SI(N,tm,ys)
estimaciones[i,j] <- ht[1,2]
j <- j+1
}
if (any(estí==6)){
greg <- sum(x%*%lineal$coefficients) + sum(lineal$residuals/piks)
estimaciones[i,j] <- greg
j <- j+1
}

```

```

if (any(estí==7)){
wgreg <- sum(x%*%wlineal$coefficients) + sum(wlineal$residuals/piks)
estimaciones[i,j] <- wgreg
j <- j+1
}
if (any(estí==8)){
regtob <- sum(x%*%modtob$betas) + sum(modtob$residuals/piks)
estimaciones[i,j] <- regtob
j <- j+1
}
if (any(estí==9)){
wregtob <- sum(x%*%wmodtob$betas) + sum(wmodtob$residuals/piks)
estimaciones[i,j] <- wregtob
j <- j+1
}
if (any(estí==10)){
ci <- ys <= 0
sc <- !ci
parti <- sum(ys[sc]/piks[sc])+sum(modtob$y_estí[ci]/piks[ci])
estimaciones[i,j] <- parti
j <- j+1
}
if (any(estí==11)){
ci <- ys <= 0
sc <- !ci
wparti<- sum(ys[sc]/piks[sc]) + sum(wmodtob$y_estí[ci]/piks[ci])
estimaciones[i,j] <- wparti
j <- j+1
}
}#fin del for
time<-proc.time()-ptm
print(time)
if(guardar==TRUE){
estimaciones<-as.data.frame(round(estimaciones,4),row.names = NULL)
write.table(estimaciones, paste(nombre,".txt",sep = ""),
sep = "\t",row.names = F)
auxiliar<-as.data.frame(round(auxiliar,4),row.names = NULL)
write.table(auxiliar, paste(nombre,"auxiliar.txt",sep = "_"),
sep = "\t",row.names = F)
}
z<-list(estí=estimaciones,aux=auxiliar)
z$tamaño<-tm
z$muestras<-m
z$censura<-lc
z
}#fin de la función

```

# Índice alfabético

- $\pi$ -estimador, 6, 21, 27
  - Esperanza de, 6
  - Sesgo de, 6
  - Varianza de, 6
- $\pi$ -expansión, 5
- Censura, 12
  - Censura a derecha, 13, 29
  - Censura a izquierda, 13, 17, 29
  - Estimador del Total
    - $\pi$ -estimador, 27
    - $\pi$ -estimador Tobit sintético, 28
  - Estimador de regresión, 27
  - Estimador de regresión Tobit, 27
  - Estimador mixto, 27
  - Estimador Tobit sintético, 28
- Función de pseudoverosimilitud, 20
  - Logaritmo de la..., 21
- Función de verosimilitud, 17
  - Estimadores máximo verosímiles, 18
  - Logaritmo de la..., 17
- Límite de..., 12, 13
- Modelo de regresión censurada, 15–17, 27
- Muestra censurada, 13, 15, 25, 27
- Observaciones censuradas, 26
- Variable
  - Censurada, 15, 27
  - Indicadora de censura, 15
  - Latente, 13, 15
  - Pseudovariable latente, 27
  - Pseudovariable mixta, 27
- Diseño muestral, 2
- Distribución normal
  - Acumulada, 15
  - Densidad, 15
  - Distribución normal censurada, 16
- Estadístico
  - Covarianza de, 3
  - Definición de, 3
  - Esperanza de, 3
  - Varianza de, 3
- Estimador, 4
  - Definición de, 4
  - Error Cuadrático Medio de, 5
  - Esperanza de, 4
  - Estimación, 5
  - Sesgo de, 4
  - Varianza de, 4
- Estimador de regresión, 7
  - Linealización de Taylor, 9
  - Modelo que asiste, 10, 12
  - Varianza Aproximada, 9
- GEREG, 11
- Información auxiliar, 7
  - Matriz de diseño, 7
- Matriz de varianzas-covarianzas, 22
  - Matriz de Información de Fisher, 22
- Media, 2, 4
- Modelo
  - De regresión lineal, 10
  - De regresión Tobit, 15–17, 27
  - Estimación máxima verosimilitud, 11

- Modelo lineal generalizado, 10
- Muestra
  - Aleatoria, 2
  - Censurada, 13, 25, 27
  - Normal censurada, 15
- Parámetro, 2, 4
  - Estimador de, 4
  - Poblacional, 2
- Pesos muestrales, 11, 20
- Población objetivo, 1
- Probabilidad, 2
  - de inclusión de primer orden, 2
  - de inclusión de segundo orden, 2
- Total, 2, 4
- Variable censurada, 29
- Variable latente, 29
- Universo, 1
- Variable, 1
  - Aleatoria, 2
  - Censurada, 15, 27
  - de estudio, 1
  - de interés, 1
  - Indicadora de censura, 15
  - indicadora de pertenencia, 2
  - Latente, 13, 15
  - Pseudovariante latente, 27, 30
  - Pseudovariante mixta, 27

# Índice de Autores

- Aboueissa & Stoline (2004), 17  
Amemiya (1973), XI, 1, 18, 22, 43  
Amemiya (1984), XI, 1, 16, 18, 22  
Arabmazar & Schmidt (1981), 66  
Arellano-Valle et al. (2012), 22  
Bolfarine et al. (2013), 1, 13  
Burkhauser et al. (2011), 26  
Carson & Sun (2007), 17  
Chambers & Skinner (2003), 20  
Cohen (1991), XI, 13, 15, 66  
Dooley (1983), 66  
Greene (2007), 1, 13  
Groves et al. (2004), 1  
Helsel (2012), 25  
Henningsen & Toomet (2011), 20  
Hurd (1979), 66  
Kleiber & Zeileis (2008), 20  
Lohr (1999), 1  
Michalek et al. (1998), 17  
Millard & Neerchal (2000), 1, 12, 13  
Olsen (1978), XI, 1, 18  
Rondón et al. (2012), 1, 11, 20, 30  
Särndal et al. (1992), 1, 3, 31, 33, 53,  
55  
Therneau & Grambsch (2000), 20  
Therneau (2015), 20  
Thompson & Nelson (2003), 17, 25  
Tobin (1958), XI, 1, 16–18



# Índice de Símbolos

$Cov(Q_1, Q_2)$ , 3	$\hat{\theta}(s)$ , 5
$E[Q]$ , 3	$\hat{\theta} = \hat{\theta}(S)$ , 4
$E[\hat{\theta}]$ , 4	$\hat{t}_k$ , 6
$E[\hat{t}_\pi]$ , 6	$\hat{t}_{yr}$ , 7
$I_k$ , 2, 6	$\bar{y}_U$ , 2
$N$ , 1	$U$ , 1
$Q(s)$ , 3	$\mathbb{X}$ , 7
$Q = Q(S)$ , 3	$\mathcal{S}$ , 2
$S$ , 2	$\phi$ , 15
$T$ , 12, 13	$\pi_k$ , 2
$V(\hat{t}_\pi)$ , 6	$\pi_{kl}$ , 2
$V[Q]$ , 3	$\psi$ , 16
$V[\hat{\theta}]$ , 4	$\sigma^2$ , 10
$V_\xi[Y_k]$ , 10	$\theta$ , 2, 4
$\mathbf{AV}(\hat{t}_{yr})$ , 9	$\tilde{c}$ , 13
$\mathbf{B}(\hat{\theta})$ , 4	$\xi$ , 10
$\mathbf{E}_\xi[Y_k]$ , 10	$c$ , 13
$\Phi$ , 15	$h$ , 18
$\alpha$ , 18	$n_s$ , 2
$\bar{y}_U$ , 4	$p(\cdot)$ , 2
$\beta$ , 10	$p(s)$ , 2
$\hat{\beta}$ , 10	$s$ , 2, 13
$\mathcal{F}$ , 22	$t$ , 2, 4
$\mathbf{x}_{kj}$ , 7	$y^*$ , 13
$\check{y}_k$ , 5	$y_k$ , 1
$\mathbf{ECM}(\hat{\theta})$ , 5	
$\hat{\beta}_j$ , 10	$y$ , 1



## Bibliografía

- Aboueissa, A. E.-M. A. & Stoline, M. R. (2004). Estimation of the mean and standard deviation from normally distributed singly-censored samples., *Environmetrics* **15**(7): 659.
- Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal, *Econometrica* **41**(6): 9971016.
- Amemiya, T. (1984). Tobit models: A survey, *Journal of Econometrics* **24**(1-2): 361.
- Arabmazar, A. & Schmidt, P. (1981). Further evidence on the robustness of the Tobit estimator to heteroskedasticity, *Journal of Econometrics* **17**(2): 253258.  
**URL:** <http://www.sciencedirect.com/science/article/pii/0304407681900294>
- Arellano-Valle, R. B., Castro, L. M., González-Farías, G. & Muñoz-Gajardo, K. A. (2012). Student-t censored regression model: properties and inference., *Statistical Methods and Applications* **21**(4): 453473.
- Blanco, L. (2010). *Probabilidad*, segunda edn, Universidad Nacional de Colombia.
- Bolfarine, H., Santos, B., Correia, L., Martínez, G., Gómez, H., Bazan, J. & ABE- Associação Brasileira de Estatística (2013). *Modelos de Regressão com Respostas Limitadas e Censuradas*.
- Burkhauser, R. V., Feng, S., Jenkins, S. P. & Larrimore, J. (2011). Estimating trends in US income inequality using the Current Population Survey: the importance of controlling for censoring, *The Journal of Economic Inequality* **9**(3): 393415.  
**URL:** <http://dx.doi.org/10.1007/s10888-010-9131-6>
- Carson, R. T. & Sun, Y. (2007). The Tobit model with a non-zero threshold., *The Econometrics Journal* (3): 488.
- Chambers, R. & Skinner, C. (2003). *Analysis of Survey Data*, Wiley Series in Survey Methodology, Chichester: Wiley.
- Cohen, A. C. (1991). *Truncated and Censored Samples: Theory and Applications*, Statistics Textbooks and Monographs 119, New York: Marcel Dekker.

- Dooley, M. D. (1983). Estimation in censored samples when there is heteroskedasticity, *Economics Letters* **13**(4): 343349.  
**URL:** <http://www.sciencedirect.com/science/article/pii/0165176583901921>
- Gourieroux, C. (2000). *Econometrics of Qualitative Dependent Variables*, Themes in Modern Econometrics, Cambridge University Press.
- Greene, W. H. (2007). *Econometric Analysis*, 6th edn, Prentice Hall.
- Groves, R. M., Fowler-Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E. & Tourangeau, R. (2004). *Survey Methodology*, Wiley Series in Survey Methodology, Hoboken: John Wiley & Sons.
- Helsel, D. R. (2012). *Statistics for Censored Environmental Data Using Minitab and R (Statistics in Practice)*, 2 edn, Wiley.  
**URL:** <http://gen.lib.rus.ec/book/index.php?md5=A955153053D12C65F4E536C03C65FC12>
- Henningsen, A. (2016). *censReg: Censored Regression (Tobit) Models*. R package version 0.5-22.  
**URL:** <https://CRAN.R-project.org/package=censReg>
- Henningsen, A. & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R, *Computational Statistics* **26**(3): 443458.  
**URL:** <http://dx.doi.org/10.1007/s00180-010-0217-1>
- Hurd, M. (1979). Estimation in truncated samples when there is heteroscedasticity, *Journal of Econometrics* **11**(2): 247258.  
**URL:** <http://www.sciencedirect.com/science/article/pii/0304407679900393>
- Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R*, Springer-Verlag, New York. ISBN 978-0-387-77316-2.  
**URL:** <http://CRAN.R-project.org/package=AER>
- Lohr, S. L. (1999). *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press.
- Michalek, J. E., Gupta, P. L., Kulkarni, P. M., Tripathi, R. C. & Selvavel, K. (1998). Correction for bias introduced by truncation in pharmacokinetic studies of environmental contaminants., *Environmetrics* **9**(2): 165.
- Millard, S. P. & Neerchal, N. K. (2000). *Environmental Statistics with S-PLUS*, Chapman & Hall/CRC Applied Environmental Statistics, CRC Press.
- Olsen, R. J. (1978). Note on the uniqueness of the maximum likelihood estimator for the Tobit model, *Econometrica* **46**(5): 12111215.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.  
**URL:** <https://www.R-project.org/>

- 
- Rondón, L. M., Ferraz, C. & Vanegas, L. H. (2012). Finite population estimation under generalized linear model assistance, *Computational Statistics & Data Analysis* **56**(3): 680697.
- RStudio Team (2016). *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA.  
**URL:** <http://www.rstudio.com/>
- Short, K., Eargle, J. & Bureau, U. C. (1991). Reciprocity history and left-censored spells in AFDC SIPP , *Technical report*, U.S. Department of Commerce U.S. CENSUS BUREAU.
- Särndal, C., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer.
- Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.  
**URL:** <https://CRAN.R-project.org/package=survival>
- Therneau, T. M. & Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.
- Thompson, M. L. & Nelson, K. P. (2003). Linear regression with Type I interval- and left-censored response data., *Environmental & Ecological Statistics* **10**(2): 221.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables, *Econometrica* **26**(1): 2436.