Modelo de Censura Intervalar para datos positivos

Justo Andrés Manrique Urbina
Asesor: Cristian L. Bayes

October 26, 2020

Contents

Chapter 1

Introducción

Los modelos de regresión usualmente asumen que la variable respuesta es directamente observable. No obstante, en ciertos estudios, la variable de interés no lo es. Un caso concreto de ello es el sueldo: una persona dudaría indicarle su sueldo exacto a un encuestador pues es un tema personal; tema que solo se conversa con personas de su confianza. Ante ello, el encuestador le brinda opciones de escala salarial a la persona para obtener el dato, sin comprometer la privacidad de la persona. Ante esta situación, los modelos tienen que ser adaptados a esta nueva estructura de datos y estudios.

Por otro lado, los modelos de regresión estudiados modelan la media de la variable de interés, condicionada por otro conjunto de variables. Sin embargo, el interés del investigador puede recaer en otro objetivo: más allá de la respuesta media, el investigador busca los factores subyacentes que impactan a distintos cuantiles de la variable respuesta. Los factores relacionados a una persona con un gran sueldo son distintos a una persona que no percibe mucho. Para estudios de dicho corte, los modelos de regresión cuantílica brinda la flexibilidad requerida. Dicho modelo fue propuesto inicialmente por Koenker y Basset (1978) quienes, ante la situación en dónde la estimación de mínimos cuadrados es deficiente en modelos con errores no gaussianos, proponen una regresión de cuantiles que permiten modelar libremente los cuantiles de la variable respuesta en relación a las covariables.

La presente tesis propone utilizar los temas anteriormente expuestos para implementar un modelo de regresión cuantílica aplicado a datos con censura intervalar. Para efectos de la aplicación, los datos se modelarán bajo la distribución Weibull, la cual es de amplia aplicabilidad. Dicha distribución será reparametrizada para adecuarse al modelo de regresión. Asimismo, el método de estimación será el de máxima verosimilitud, siguiendo el marco de la inferencia clásica.

1.1 Objetivos de la tesis

El objetivo de la tesis, conforme indicado anteriormente, consiste en proponer un método de regresión cuantílica adaptado a datos con censura intervalar e implementar dicho modelo utilizando los datos de la Encuesta Nacional de Satisfacción de Usuarios en Salud. Para ello, asumimos que los datos subyacentes tienen una distribución Weibull. Los objetivos específicos son los siguientes:

- Revisar literatura académica relacionada a las propuestas de modelos de regresión con datos censurados intervalarmente.
- Identificar una estructura apropiada de la distribución Weibull para el modelo de regresión cuantílica vía una reparametrización del modelo. Posteriormente, estudiar el comportamiento de dicha estructura.
- Estimar los parámetros del modelo propuesto bajo inferencia clásica.
- Implementar el método de estimación para el modelo propuesto en el lenguaje R y aplicarlo en datos simulados.
- Aplicar el modelo propuesto en datos de la Encuesta Nacional de Satisfacción de Usuarios en Salud.

1.2 Organización del Trabajo

En el capítulo 2, se presenta una estructura de la distribución Weibull, apropiada para los datos con censura intervalar. Por ello, se realiza una parametrización alternativa y se estudia los

En el capítulo 3, se propone el modelo de regresión con datos censurados intervalarmente.

En el capítulo 4, se presenta la aplicación del modelo propuesto para determinar si existe diferencia entre los sueldos de enfermeras y enfermeros a lo largo de todos los cuantiles. Ello se realiza mediante inferencia clásica.

Finalmente, en el capítulo 5 se presentan las principales conclusiones obtenidas en la presente tesis así como los próximos pasos.

Chapter 2

Distribución Weibull

El presente capítulo tiene como objetivo estudiar las principales propiedades de la distribución Weibull vía una reparametrización del modelo original. Dicha distribución se utilizará en los capítulos futuros. Para esta reparametrización, se define su función de probabilidad y sus propiedades (esperanza y varianza).

2.1 Distribución Weibull

2.1.1 Función de densidad

La distribución Weibull, fue desarrollada por el ingeniero sueco Waloddi Weibull, es una distribución de amplia aplicabilidad (Weibull, 1951). Una variable aleatoria continua y, en dónde $y > 0$, tiene distribución Weibull con parámetro de forma α y dispersión σ respectivamente si su función de densidad es dada por la siguiente expresión:

$$f(y) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1}\exp\left(-\frac{y}{\sigma}\right)^{\alpha}$$

en dónde $\alpha > 0$ y $\sigma > 0$. La notación de una variable aleatoria u que sigue esta distribución se indica como $y \sim W(\alpha, \sigma)$. . Asimismo, la función acumulada de y corresponde a la siguiente expresión:

$$F(y) = 1 - \exp^{-(\frac{y}{\sigma})^{\alpha}}.$$

Y la función de cuantiles es dada por:

$$q_t = \sigma(-\log(1-t))^{\frac{1}{\alpha}}$$

para $0 < t < 1$.

Para dicha variable aleatoria y la media y varianza es de la siguiente forma:

$$E(y) = \sigma\Gamma\left(1 + \frac{1}{\alpha}\right).$$

$$V(y) = \sigma^2 \left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \left(\Gamma\left(1 + \frac{1}{\alpha}\right)\right)^2 \right].$$

2.1.2 Proposición de una nueva estructura de la distribución

Consideramos, para la distribución Weibull, una reparametrización en términos del cuantil t, q_t, dada por:

$$q_t = \sigma \left(-\log\left(1-t\right)\right)^{\frac{1}{\alpha}}.$$

Al respecto, cabe indicar que t será un valor conocido y se encuentra en el intervalo $[0, 1]$. En esta nueva esctructura, q_t y α tienen espacios paramétricos independientes tal que $(q_t, \alpha) \in (0, \infty)$ x $(0, \infty)$. Una variable aleatoria que sigue esta parametrización se denota como $Y \sim W_r(q_t, \alpha, t)$.

La función de densidad de dicha variable Y tiene la siguiente expresión:

$$f_y(y|q_t, \alpha, t) = \frac{\alpha c(t)}{q_t} \left(\frac{y}{q_t}\right)^{\alpha-1} \exp\left(-c(t)\left(\frac{y}{q_t}\right)^{\alpha}\right) \tag{2.1}$$

en dónde $c(t) = (-\log(1-t))^{\frac{1}{\alpha}}$. Los parámetros q_t y α caracterizan la función de densidad conforme se observa en el gráfico siguiente:

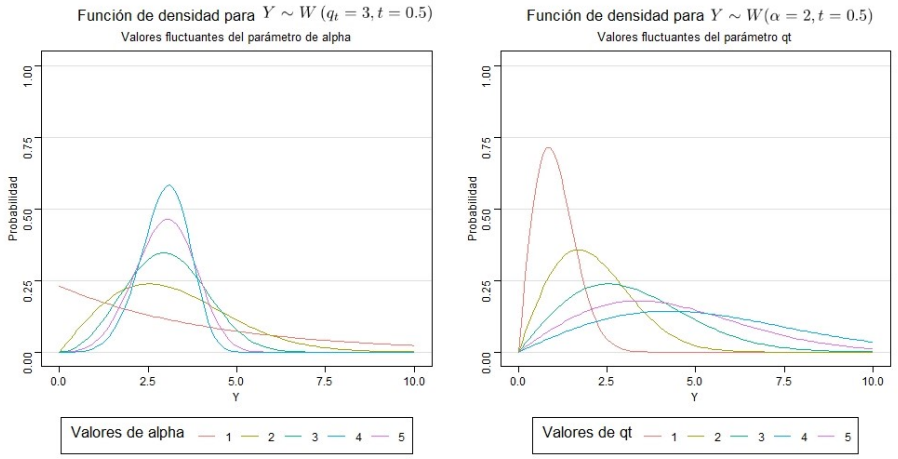

Figure 2.1: Función de densidad de una distribución Weibull bajo la reparametrización propuesta.

Se observa que en la medida que q_t aumenta, la distribución incrementa su asimetría hacia la derecha. Ello también sucede, aunque en menor grado, cuando α aumenta. No obstante, se observa que en la medida que α tiende a 0, incrementa la dispersión.

Reexpresando la función acumulada en los términos de la parametrización propuesta, esta tendría la siguiente forma:

$$F_y\left(y|q_t,\alpha,t\right)=1-\exp\left(-c(t)\left(\frac{y}{q_t}\right)^{\alpha}\right). \tag{2.2}$$

2.1.3 Estudio de la parametrización propuesta

La esperanza y varianza de una variable aleatoria bajo la parametrización Weibull propuesta están dadas bajo la siguiente expresión:

$$E(Y)=\frac{q_t}{c(t)^{\frac{1}{\alpha}}}\Gamma\left(1+\frac{1}{\alpha}\right) \tag{2.3}$$

$$Var(y)=\frac{q_t^2}{c(t)^{\frac{1}{\alpha}}}\left[\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma\left(1+\frac{1}{\alpha}\right)^2\right] \tag{2.4}$$

Bajo la parametrización propuesta, se observa que para un α fijo el valor esperado se comporta de forma lineal en la medida que aumente el parámetro q_t conforme se observa en el cuadro siguiente:

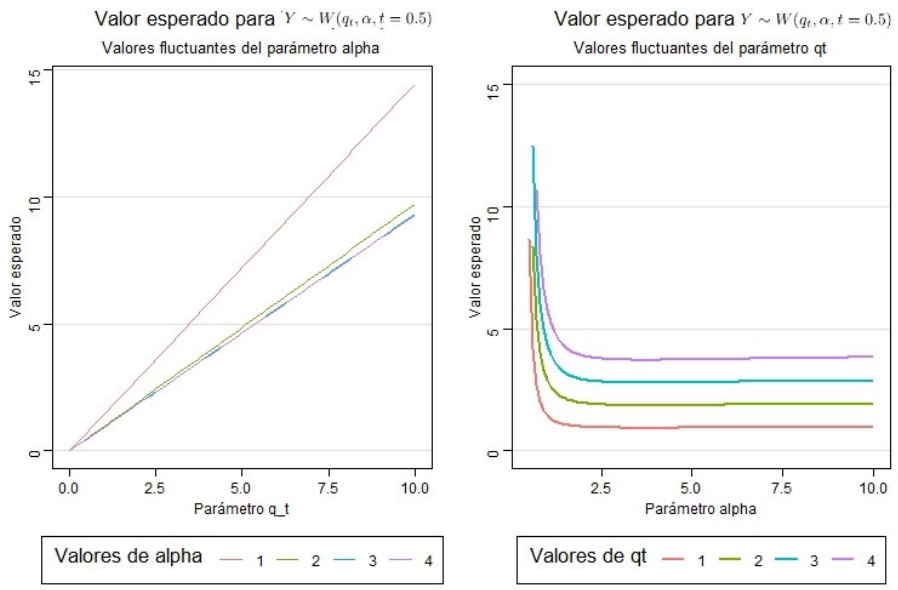

Figure 2.2: Valor esperado de una distribución Weibull bajo la parametrización propuesta.

No obstante, para un q_t fijo, lo mismo no se observa en la medida que aumente α. Se observa un comportamiento no lineal y asintótico: cuando α tiende a

0, el valor esperado tiende a infinito. Cuando α aumenta, el valor esperado se estabiliza.

En el caso de la varianza se observa que para un α fijo, en la medida que aumente el parámetro q_t la varianza aumenta de forma exponencial. No obstante, y como se puede apreciar cuando q_t está fijo, en la medida que los valores de α sean pequeños, la varianza incrementa drásticamente. Asimismo, como se aprecia en el cuadro adjunto, la varianza tiende a 0 en la medida que α aumente.

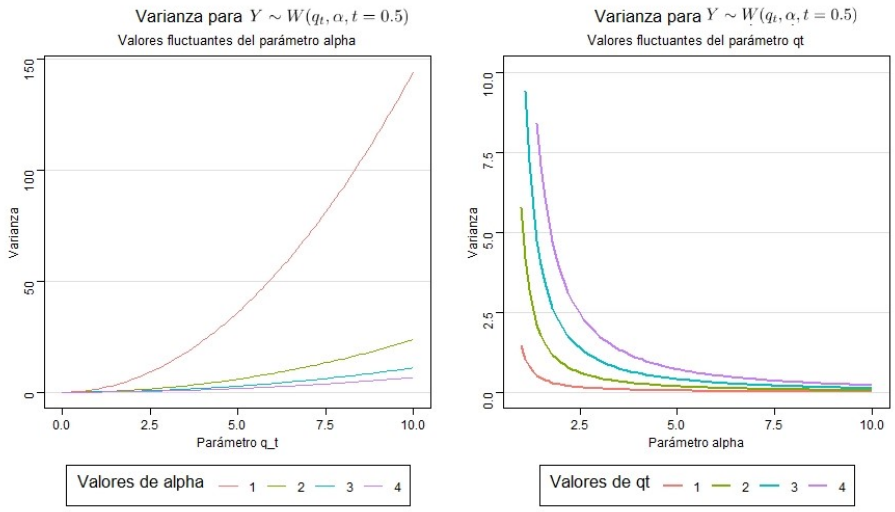

Figure 2.3: Varianza bajo la parametrización propuesta.

Chapter 3

Modelo de regresión cuantílica para datos positivos

El presente capítulo tiene como objetivo especificar el modelo de regresión cuantílica para datos positivos con censura intervalar. Asimismo, detallamos la estimación de los parámetros desde la perspectiva de inferencia clásica.

3.1 Datos positivos con censura intervalar

Conforme mencionado en la introducción, algunos características de la población pueden solo capturarse en un intervalo por multiplicidad de condiciones. Una de ellas es el derecho a privacidad de las personas, quienes solo desean revelar información sobre si mismas sin exponer mucha información. Caso concreto es el sueldo: una persona no desearía brindar su sueldo específico a alguien desconocido, no obstante puede indicar que su sueldo está en un rango.

Bajo ese contexto, podemos definir una variable aleatoria z como una variable que indica que la variable y se encuentra en el j-ésimo intervalo $[L_j, L_{j+1}]$. Se asume que solamente observamos la variable z mientras que y es una variable latente, que en el ejemplo correspone al sueldo específico de la persona (el cual no ha sido revelado). La variable aleatoria z es una variable cualitativa pues solo se indica el intervalo que la persona responde. Por lo tanto, la podemos definir mediante la siguiente expresión:

$$z = \begin{cases} 1, L_1 < y < L_2 \\ 2, L_2 \leq y < L_3 \\ 3, L_3 \leq y < L_4 \\ \vdots \\ K, L_K \leq y < L_{K+1} \end{cases} \tag{3.1}$$

En dónde $L_1 < L_2 < \cdots < L_{K+1}$, y corresponde a los límites del intervalo, con $L_1 = 0$ y $L_{K+1} = \infty$. La probabilidad de Z está definida bajo la siguiente expresión:

$$P\left(Z = j\right) = P\left(L_j \leq y < L_{j+1}\right)$$
$$P\left(Z = j\right) = F\left(L_j\right) - F\left(L_{j+1}\right)$$

dónde $F(\cdot)$ es la función de distribución acumulada de Y. La variable Z que sigue la distribución anteriormente mencionada está denotada por

$$Z \sim \text{Categórica}(\pi)$$

dónde $\pi = (\pi_1, \ldots, \pi_k)$y $\pi_j = P(Z = j)$.

3.2 Modelo de regresión para datos positivos con censura intervalar

El modelo de regresión cuantílica para datos positivos está dado por lo siguiente:

$$Y_i \sim W_r\left(q_{t_i}, \alpha, t\right).$$
$$g\left(q_{t_i}\right) = x_i^T \beta.$$

en dónde $\beta = [\beta_0, \beta_1, \ldots, \beta_p]^T$ y $x_i^T = [1, x_{i1}, x_{i2}, \ldots, x_{ip},]^T$. La función $g(\cdot)$ es una función de enlace estrictamente monótona y doblemente diferenciable. En el presente modelo, se utilizará la función de enlace logarítmica. El parámetro de forma α, el parámetro q_{t_i} y t está definido conforme la sección 2.1. La estimación de los parámetros β y α se realizará mediante el método de máxima verosimilitud.

3.2.1 Función de verosimilitud

Consideramos que solo conocemos que Y_i se encuentra en un intervalo de K posibles intervalos de la forma $[L_j, L_{j+1}]$ con $L_1 < L_2 < \cdots < L_{K+1}$ y que $Z_i = j$ denota que $Y_i \in [L_j, L_{j+1}]$. Por lo tanto, considerando los resultados de la sección 3.1, tenemos que

$$Z_i \sim \text{Categórica}(\pi_i).$$

con $\pi_i = (\pi_{i1}, \ldots, \pi_{ik})$ tal que

$$\pi_{ij} = F_y(L_j | q_{t_i}, \alpha, x) - F_y\left(L_{j+1} | q_{t_i}, \alpha, x\right) \tag{3.2}$$

Entonces la función de verosimilitud de las variables observadas $Z_1, Z_2, \ldots, Z_n$ es dada por lo siguiente:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_{ij}^{1(Z_i = j)}.$$

Luego, considerando $[\psi_{i_{inf}}, \psi_{i_{sup}}]$ como el intervalo dónde Y_i fue observado, podemos escribir la función de verosimilitud como:

$$L(\theta) = \prod_{i=1}^{n} \left(F(\psi_{i_{sup}} | q_{t_i}, \alpha, t) - F(\psi_{i_{inf}} | q_{t_i}, \alpha, t) \right)$$

$$L(\theta) = \sum_{i=1}^{n} \log \left(F(\psi_{i_{sup}} | q_{t_i}, \alpha, t) - F\left(\psi_{i_{inf}} | q_{t_i}, \alpha, t\right) \right)$$

$$L(\theta) = \sum_{i}^{n} \log \left(\exp \left(-c(t) \left(\frac{\psi_{i_{inf}}}{e^{x_i^T \beta}} \right)^{\alpha} - \exp \left(-c(t) \left(\frac{\psi_{i_{sup}}}{e^{x_i^T \beta}} \right)^{\alpha} \right) \right) \right)$$

en dónde $c(t) = (-log(1-t))^{\frac{1}{\alpha}}$.

Los estimadores de máxima verosimilitud para los parámetros α y β se encuentran maximizando la función anteriormente expuesta. Para ello, obtenemos las gradientes de α y β, se exponen a continuación (asumiendo que $g(\cdot)$ es la función logaritmo):

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n} \frac{c(t)}{(\gamma_i)^{\alpha}(\lambda_{i_2} - \lambda_{i_1})} \left((\psi_{i_{sup}})^{\alpha} \log \left(\frac{\psi_{i_{sup}}}{\gamma_i} \right) \lambda_{i_2} - (\psi_{i_{inf}})^{\alpha} \log \left(\frac{\psi_{i_{sup}}}{\gamma_i} \right) \lambda_{i_1} \right)$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{\alpha c(t) x_{ij}}{(\gamma_i)^{\alpha}(\lambda_{i_1} - \lambda_{i_2})} \right) \left((\psi_{i_{inf}})^{\alpha} \lambda_{i_1} - (\psi_{i_{sup}})^{\alpha} \lambda_{i_2} \right)$$

en dónde:

$$\gamma_i = \exp(\eta_i)$$

$$\eta_i = x_i^T \beta$$

$$\lambda_{i_1} = \exp \left(-c(t) \left(\frac{\psi_{i_{inf}}}{\gamma_i} \right)^{\alpha} \right)$$

$$\lambda_{i_2} = \exp \left(-c(t) \left(\frac{\psi_{i_{sup}}}{\gamma_i} \right)^{\alpha} \right)$$

La maximización de dicha la función de log-verosimilitud se realizará mediante métodos de optimización numérica a través del lenguaje de programación R.

Chapter 4

Estudio de Simulación

El presente capítulo tiene como objetivo realizar un estudio de simulación en el que se evalúe el adecuado rendimiento del modelo propuesto en los capítulos antecedentes. Esto comprende generar una base de datos en dónde se tenga una variable aleatoria $Y_i \sim W_r(q_t, \alpha, t)$, la cual está subyace la variable censurada Z que sigue lo denotado en la sección 3.1). Asimismo, dicha base de datos contiene otras variables simuladas, las cuales actuarán como variables independientes en un contexto de regresión. El objetivo principal del estudio de simulación es observar si el modelo de regresión planteado, así como la implementación del mismo, permite capturar adecuadamente los parámetros de regresión establecidos a priori. Los criterios sobre los cuales se analizará el rendimiento del modelo son: sesgo relativo, error cuadrático medio y cobertura.

El proceso de simulación consiste en generar 100 réplicas de la variable aleatoria $Y_i \sim W_r(q_t, \alpha, t)$ basados en los valores de $n = \{1000, 5000, 10000\}$ de las siguientes variables:

$$X_1 \sim Beta(2, 3)$$

$$X_2 \sim Normal(2, 0.5)$$

$$X_3 \sim Gamma(2, 25)$$

Conforme lo mencionado en la sección 3.2.1), la función de enlace para crear las réplicas está denotada por $q_t = \exp(x_i^T \beta)$, en dónde $\beta = [0.5, 0.3, 0.6, 0.8]^T$. Por otro lado, el parámetro de dispersión tiene el valor $\alpha = 2$. Finalmente, se realizará la evaluación por los cuantiles $t = [0.1, 0.2, \ldots, 0.9]$.

Cada réplica de la variable aleatoria $Y_i \sim W_r(q_t, \alpha, t)$ subyace la variable de censura intervalar Z, la cual particiona la variable Y_i en intervalos de igual amplitud, con la excepción del último intervalo, el cual tiene la estructura $[L_{inf}, \infty]$. Una vez generada dicha variable, se realiza el modelamiento de la variable de censura intervalar sobre las variables independientes creadas previamente. El objetivo final es, a través del modelo, obtener los coeficientes β definidos previamente.

4.1 Implementación del modelo

La implementación del modelo se realizó a través del lenguaje de programación R, tomando en consideración las fórmulas especificadas en el capítulo 3 de la presente tesis. El seudocódigo de la implementación es el siguiente:

```
Simulamos 10,000 valores de las siguientes distribuciones:

X1 ~ Beta(2,3)
X2 ~ Normal(2,0.5)
X3 ~ Gamma(2,25)

Definimos los siguientes valores:

B = [0.5, 0.3, 0.6, 0.8]
Sigma = 2
Qt = exp(B[1] + B[2]*X1 + B[3]*X2 + B[4]*X3)
t=[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
M = 100

Para cada cuantil en t:
        Para cada replica en M:
                1 Simular 10,000 valores de la distribucion
                        Y ~ W_r(Qt, Sigma, cuantil)
                2 Censurar la variable Y de forma intervalar tal que
                        Z ~ Categorica
                3 Obtener los limites inferiores y superiores de
                        cada categoria de Z
                4 Crear la base de datos simulada
                        df <- [L_inf, L_sup, X1, X2, X3]
                5 Ejecutar la regresion de censura intervalar
                6 Guardar los resultados
```

Una vez generados los resultados, se evaluó por cada cuantil lo siguiente:

$$\text{Sesgo relativo:} \frac{1}{M}(\hat{\theta_j} - \theta_j)$$

$$\text{ECM:} \frac{1}{M}\sum_{1}^{M}(\hat{\theta_j} - \theta_j)^2$$

$$\text{Cobertura:} \frac{1}{M}\sum_{1}^{M}(\hat{\theta_j} - \theta_j)^2$$

4.2 Resultados

En la tabla a continuación se muestra, para cada tamaño de muestra y parámetros $\Theta = [\beta_0, \beta_1, \beta_2, \beta_3, \sigma] = [0.5, 0.3, 0.6, 0.8, 2]$ los resultados de las métricas uti-

lizadas para evaluar el desempeño de los estimadores bajo la estructura propuesta. Para las 100 réplicas identificadas, se observa lo siguiente:

- En relación al sesgo relativo, se observa que cada parámetro permite capturar adecuadamente los parámetros previamente precisados. Existe, no obstante una ligera subestimación de los parámetros para determinados cuantiles, como el cuantil 0.1 y 0.9
- En relación a la cobertura, se observa que, en promedio, los intervalos de confianza contienen en un 95% el valor real del parámetro.
- En relación al error cuadrático medio, se observa que este es pequeño para todos los cuantiles y parámetros.

Ver tabla a continuación:

Cuantil	**Parámetros**	**n = 10000**		
		Sesgo Relativo	**Cobertura**	**Error Cuadrático Medio**
0.1	β_0	0.0847	98.00%	0.0091
	β_1	-0.0365	95.00%	0.0034
	β_2	-0.0292	98.00%	0.0034
	β_3	-0.0564	96.00%	0.0035
	σ	-0.0223	96.00%	0.0034
0.2	β_0	-0.0170	90.00%	0.0161
	β_1	0.0101	93.00%	0.0037
	β_2	0.0070	93.00%	0.0037
	β_3	-0.0305	96.00%	0.0038
	σ	-0.0122	91.00%	0.0037
0.3	β_0	-0.0263	98.00%	0.0074
	β_1	-0.0133	99.00%	0.0030
	β_2	0.0104	98.00%	0.0030
	β_3	0.0069	95.00%	0.0032
	σ	0.0028	97.00%	0.0030
0.4	β_0	-0.0229	91.00%	0.0141
	β_1	0.0182	97.00%	0.0026
	β_2	0.0024	95.00%	0.0026
	β_3	0.0840	90.00%	0.0026
	σ	0.0333	97.00%	0.0026
0.5	β_0	0.0137	96.00%	0.0117
	β_1	0.0556	92.00%	0.0042
	β_2	-0.0134	95.00%	0.0042
	β_3	0.0239	96.00%	0.0043
	σ	0.0097	97.00%	0.0042
0.6	β_0	0.0170	96.00%	0.0113
	β_1	-0.0105	96.00%	0.0033

Cuantil	**Parámetros**	**n = 10000**		
		Sesgo Relativo	**Cobertura**	**Error Cuadrático Medio**
	β_2	-0.0049	95.00%	0.0033
	β_3	-0.0185	93.00%	0.0035
	σ	-0.0073	96.00%	0.0033
0.7	β_0	-0.0107	96.00%	0.0108
	β_1	0.0207	94.00%	0.0044
	β_2	0.0011	93.00%	0.0044
	β_3	0.0334	96.00%	0.0050
	σ	0.0131	93.00%	0.0044
0.8	β_0	-0.0357	96.00%	0.0084
	β_1	0.0373	93.00%	0.0038
	β_2	0.0141	95.00%	0.0038
	β_3	-0.0552	93.00%	0.0038
	σ	-0.0217	88.00%	0.0038
0.9	β_0	0.0206	96.00%	0.0094
	β_1	0.0059	93.00%	0.0044
	β_2	-0.0108	96.00%	0.0045
	β_3	-0.0126	93.00%	0.0074
	σ	-0.0048	97.00%	0.0044

Asimismo, se visualizó la dispersión de cada uno de los parámetros en relación al valor precisado anteriormente (es decir, el parámetro centrado en su valor prefijado). Se observa que las 100 réplicas son simétricas en su eje y tienen relativamente poca dispersión, a excepción del parámetro bb_3.

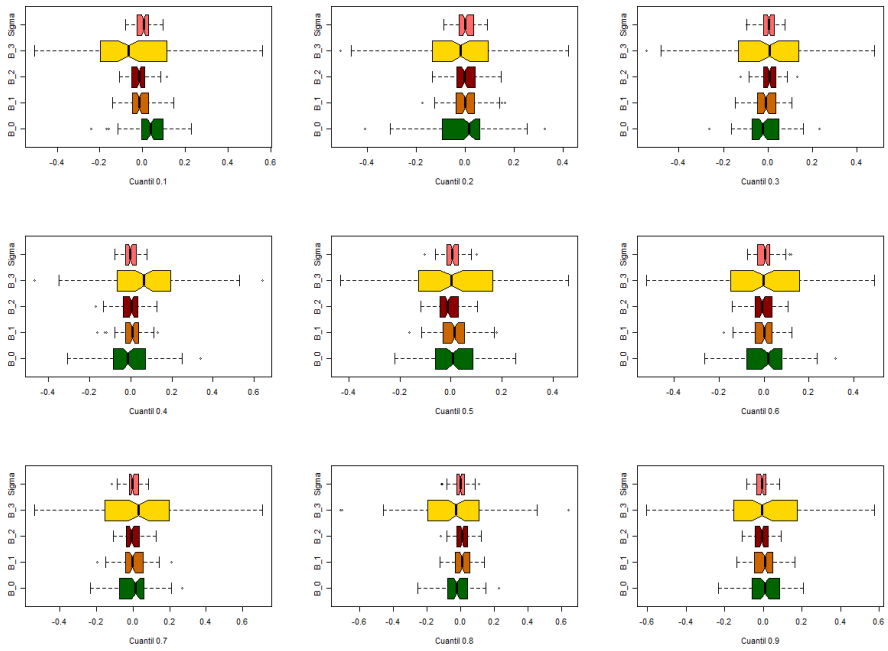

Figure 4.1: Valores de los parámetros (100 réplicas).

Por lo tanto, los resultados muestran que la implementación ejecutada permite capturar los parámetros pre-establecidos del modelo de regresión con censura intervalar.

Chapter 5

Aplicación en datos reales

5.1 ENSUSALUD 2015

La Encuesta Nacional de Satisfacción de Usuarios del Aseguramiento Universal en Salud (ENSUSALUD) es una investigación estadística realizada por el Instituto Nacional de Estadística e Informática del Perú, cuyo objetivo es "evaluar el grado de satisfacción de los usuarios internos y externos de los servicios de salud" (INEI, 2015). La investigación ENSUSALUD es una encuesta basada en un muestreo probabilístico polietápico. La unidad primaria de muestreo (UPM) está constituida por "los establecimientos de salud del MINSA-GR. EsSalud, clínicas privadas y sanidades de las Fuerzas Armadas y Policiales" (INEI, 2015). La unidad secundaria de muestreo (USM) está constituida por los usuarios elegibles dentro del establecimiento de salud: usuarios y profesionales (de Salud y administrativos). En el caso de la UPM, el método de selección fue proporcional al tamaño, tomando en consideración el número de atenciones del establecimiento. En el caso de la USM, la selección fue aleatoria sistemática. La investigación estadística tiene el siguiente alcance:

- **Cobertura geográfica:** Los 24 departamentos del Perú y 181 establecimientos de salud del MINSA, EsSalud, Sanidades y establecimientos privados.
- **Unidad de análisis:** La unidad muestral comprende a los siguientes:
  - Usuarios de Consulta Externa.
  - Usuarios en Boticas y Farmacias.
  - Usuarios en Unidades de Seguros.
  - Profesionales de la Salud.
- **Niveles de inferencia:** Nacional y dirigida a cada una de las unidades de análisis.

5.1.1 Base de datos

Para propósitos de la presente investigación, se utilizó la base de datos relacionada al personal médico y de enfermería. Dicha base de datos tiene el objetivo de "conocer características del personal relacionados a formación académica, actividad laboral, satisfacción con el trabajo, estrés laboral y conocimiento referido a la SUNASA" (INEI,2015). De dicha base de datos, se utilizaron las siguientes variables:

- Años de experiencia en el sector salud.
- Horas de trabajo semanales.
- Cantidad de personas que dependen económicamente del encuestado.
- Sexo del encuestado
- Límite inferior del sueldo percibido por el encuestado.
- Límite superior del sueldo percibido por el encuestado.

El modelo propuesto en el capítulo 3 será utilizado en la presente base de datos, tomando en consideración que el sueldo constituye una variable censurada de forma intervalar.

5.2 Resultados

Sección en construcción.

Chapter 6

Conclusiones

Este capítulo se entregará en posteriores entregas.

Chapter 7

Bibliografía

Weibull, Waloddi. "A statistical distribution function of wide applicability" *ASME Journal of Applied Mechanics.*.1951, pp. 293-297.

Fahrmeir, Ludwig. Kneib, Thomas. Lang, Stefan. Marx, Brian. *Regression: Models, Methods and Applications.* 2013. Springer.

Du, Xiuli. Feng, Yanqin. Zhou, Xiuqing. "Quantile regression for interval censored data". *Communications in Statistics - Theory and Methods.* 2015, pp. 3848-3863.

Koenker, Roger. Basset, Gilbert Jr. Regression Quantiles. 1978.

Sal y Rosas, Víctor. Moscoso-Porras, Miguel. Ormeño, Rubén. Artica, Fernando. Bayes, Cristian Luis. Miranda, Jaime. Gender Income Gap among physicians and nurses in Perú: a nationwide assessment. *The Lancet.* 2019.