PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE GRADUADOS

Modelo de censura intervalar para datos positivos

TESIS PARA OPTAR POR EL GRADO DE MAGISTER EN ESTADÍSTICA

Presentado por:

Justo Andrés Manrique Urbina

Asesor: Cristian Luis Bayes Rodríguez

Miembros del jurado:
Dr. Nombre completo jurado 1
Dr. Nombre completo jurado 2
Dr. Nombre completo jurado 3

Lima, Diciembre 2020

Agradecimientos

A lo largo de la escritura de esta tesis de maestría he recibido un inmejorable soporte y asistencia. En primer lugar, quisiera agradecer a mi asesor, el profesor Cristian Bayes, cuya supervisión y consejos han sido invaluables para la elaboración de este documento. La retroalimentación recibida mejoró la calidad de las ideas expuestas aquí. Asimismo, agradezco nuevamente al profesor Bayes y al profesor Giancarlo Sal y Rosas. Gracias a ambos, pude iniciar mi carrera académica en Estadística. Finalmente, agradezco a la plana docente de la maestría por su enseñanza y orientación durante los cursos llevados.

Resumen

La presente tesis propone un modelo de regresión cuantílica para datos positivos con censura intervalar, que permite modelar estructuras de datos en dónde la variable respuesta no es directamente observable, y la única información que se conoce sobre la misma es que se encuentra en un intervalo $[a_i, b_i]$. Para evaluar si el método propuesto captura adecuadamente los parámetros poblacionales desde el punto de vista de la inferencia clásica, se desarrolla un estudio de simulación. Se observa que el modelo propuesto sí captura estos parámetros establecidos. Finalmente, se aplica el modelo a los datos de la Encuesta Nacional de Satisfacción de Salud ejecutada el año 2015. La estructura del modelo permite evaluar los factores relacionados al sueldo de los profesionales en salud (el cual había sido censurado desde el proceso de recolección de datos). El presente modelo es una extensión al modelo de regresión de censura intervalar expuesto en Sal y Rosas et al. (2019), pues se evalúa los factores subyacentes a una variable respuesta a lo largo de sus cuantiles.

Palabras clave: censura intervalar, regresión con censura, inferencia clásica, regresión paramétrica.

Índice general

Lista de Símbolos

$Y \sim W_r(q_t, \alpha)$	Variable Y sigue una distribución Weibull reparametrizada.
q_t	Parámetro de forma.
α	Parámetro de escala.

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Por distintas razones, los datos recabados en una investigación de índole estadística carecen de precisión: existen discrepancias entre el valor real del objeto de medición y el valor obtenido. Este proceso puede ser sistémico: durante la administración de cuestionarios a una población objetivo, el encuestado puede omitir, rehúsar o incluso responder incorrectamente preguntas embarazosas o invasivas. Este dilema es conocido entre los encuestadores: sus encuestados, si bien están dispuestos a ofrecer la mejor ayuda posible, no están dispuestos a ofrecer información que posteriormente les pueda comprometer. Para obtener dichos datos, el encuestador usa todo su ingenio para equilibrar la privacidad del encuestado y los objetivos de su investigación. En un esfuerzo de aminorar el estrés del encuestado, el encuestador puede censurar los datos con el fin de obtener una respuesta.

Dicho tipo de datos se les denomina *datos censurados*, y han sido estudiados previamente en la literatura académica. Formalmente, y siguiendo las ideas plasmadas por Peto (1973), una variable C se le denota censurada cuando su valor c no es del todo observable y la única información sobre la misma es un intervalo I. Esta construcción permite definir tres tipos de datos censurados: datos censurados *hacia la izquierda* (en dónde el intervalo I se define de la forma $[-\infty, L_i]$), datos censurados *intervalares* (definido de la forma $[L_i, L_f]; L_i < L_f$), datos censurados *hacia la derecha* (definido de la forma $[L_f, \infty]$).

Este tipo de datos naturalmente generan retos en el proceso de modelamiento, pues los modelos estándares de regresión presumen que la variable respuesta es directamente observable. Situaciones como la precisada en el parráfo precedente han sido exploradas previamente: desde la determinación de la verosimilitud, la elaboración de modelos de regresión y su estimación bajo inferencia clásica y bayesiana. Gentleman y Geyer (1994) identificaron un método de máxima verosimilitud para este tipo de datos, asegurando su consistencia estadística e identificando métodos algorítmicos para su cómputo. Utilizado los puntos extremos del intervalo, L_i y L_f, era posible identificar la máxima verosímilitud a través de la diferencia de las funciones de distribución acumulada en dichos puntos. Tomando en consideración dicho método de estimación distintos autores propusieron modelos de regresión paramétricos bajo inferencia clásica y bayesiana, tales como Munoz y Xu (1996), quienes identificaron modelos paramétricos de supervivencia para este tipo de datos.

Los modelos anteriormente expuestos tienen como propósito modelar el valor esperado de la variable respuesta condicionada por un conjunto de variables, no obstante el investigador puede tener como objetivo identificar los factores subyacentes que impactan a distintos

cuantiles de la variable respuesta. Por ejemplo, los factores (y el efecto de los mismos) que modelen a una persona con un gran sueldo pueden ser muy distintos a una persona con un sueldo promedio o bajo. Bajo este contexto, Koenker y Bassett (1978) propuso un modelo que extiende esta idea a la estimación de modelos en los que los cuantiles de la distribución condicional de la variable respuesta son expresadas como funciones de un conjunto de covariables Koenker y Hallock (2001). Posteriormente, Zhou et al. (2016) propone un método de estimación para datos con censura intervalar y establece las propiedades asintóticas de los estimadores.

La presente tesis propone utilizar los temas y modelos anteriormente expuestos para implementar un modelo paramétrico de regresión cuantílica aplicado a datos con censura intervalar. Para efectos de la aplicación, los datos se modelarán bajo una distribución Weibull, la cual es de amplia aplicabilidad y permite modelar colas pesadas. Con el propósito de implementar la regresión cuantílica y, atendiendo a la estructura de los datos, dicha distribución será reparametrizada. Finalmente, el método de estimación será el de máxima verosimilitud, siguiendo el marco de la inferencia clásica.

1.1. Objetivos

El objetivo de la tesis consiste en proponer un modelo de regresión cuantílica adaptado a datos positivos con censura intervalar basado en la distribución Weibull. Para estudiar el modelo propuesto, aplicaremos este modelo en dos conjuntos de datos: uno simulado y otro real. La base de datos a utilizar será la Encuesta Nacional de Satisfacción de Usuarios en Salud elaborada por el Instituto Nacional de Estadística e Informática el año 2015. Los objetivos específicos de la tesis son los siguientes:

- Revisar la 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.
- Estimar los parámetros del modelo propuesto bajo inferencia clásica.
- Implementar el método de estimación para el modelo propuesto en el lenguaje R y realizar un estudio de simulación
- Aplicar el modelo propuesto a 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. Asimismo, se efectúa un estudio de simulación para evaluar si el modelo captura apropiadamente los parámetros poblacionales.

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.

Capítulo 2

Distribución Weibull

El presente capítulo tiene como objetivo principal proponer una reparametrización de la distribución Weibull para adaptarla al modelo de regresión cuantílica. Para dicha reparametrización, se definirá su función de densidad y función acumulada, y asimismo se examinará sus propiedades.

2.1. Distribución Weibull

La distribución Weibull fue presentada por Weibull (1951). En dicho artículo de investigación, Weibull menciona las características de una función de densidad suficientemente flexible para ser adaptada a diversas investigaciones, desde la rama de resistencia de materiales hasta el análisis de altura de hombres adultos radicados en las Islas Británicas. Una variable aleatoria continua Y, con soporte $Y \in (0, \infty)$, sigue una distribución Weibull si su función de densidad es dada por la siguiente expresión:

$$f(y) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1} \exp\left(-\frac{y}{\sigma}\right)^{\alpha} \tag{2.1}$$

en dónde α corresponde al parámetro de forma, con $\alpha > 0$, y σ corresponde al parámetro de escala, con $\sigma > 0$. La notación de una variable aleatoria Y que sigue esta distribución se indica como $Y \sim W(\alpha, \sigma)$. La función de distribución acumulada de Y tiene la siguiente expresión:

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

Asimismo, su función de cuantiles es dada por:

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

para $0 < t < 1$.

La flexibilidad denotada por Weibull puede observarse a través de la Figura 2.1. En dicha figura, observamos que una modificación del parámetro de escala σ modifica la tendencia central de la distribución; es decir, se observa que la distribución se mueve en la dirección que el parámetro σ se mueve. Cabe resaltar que, con un α fijo, la distribución tiende a ser más dispersa. Por otro lado, se observa que el aumento del parámetro de forma α contrae la distribución hacia el valor central de la misma. Esto es más pronunciado en la medida que dicho parámetro aumenta, manteniendo el parámetro σ constante. No obstante, cabe resaltar

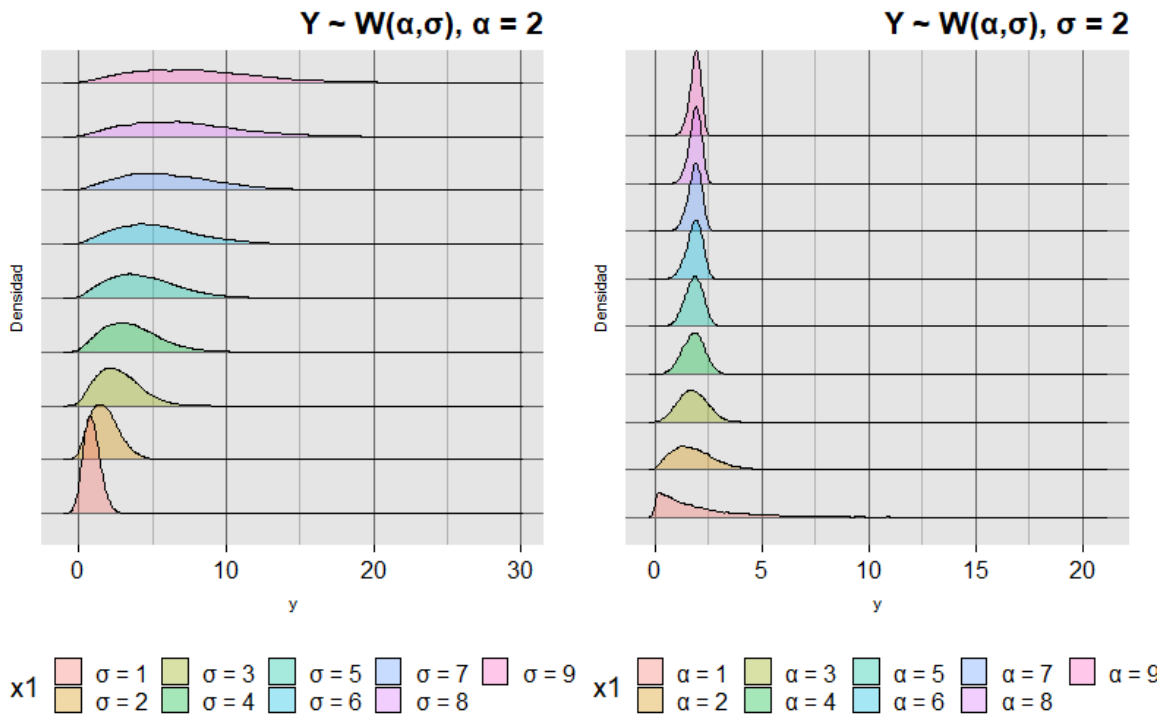


Figura 2.1: Función de densidad de una variable Y con distribución Weibull

que la distribución se torna asimétrica en tanto los valores de α son pequeños.

Para una variable $Y \sim W(\alpha, \sigma)$, la media y varianza se definen de la siguiente forma:

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

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

2.2. Proposición y estudio de una nueva estructura de la distribución

Consideramos una reparametrización del parámetro de forma σ en términos del cuantil t, q_t en los siguientes términos:

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

en dónde t será un valor conocido y se encuentra en el intervalo $[0, 1]$. En esta nueva estructura, q_t y α tienen espacios paramétricos independientes tal que $(q_t, \alpha) \in (0, \infty) \times (0, \infty)$. Una variable aleatoria Y que sigue una distribución Weibull bajo 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:

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

en dónde $c(t) = (-\log(1-t))$.

Los parámetros q_t y α, así como el nivel del cuantil t (el cual es conocido), caracterizan la función de densidad conforme se observa en la Figura 2.2. En dicha figura, se observa el efecto del nuevo parámetro q_t y el cuantil t. Podemos observar lo siguiente:

- Se observa que, para un mismo cuantil t, el nuevo parámetro q_t tiende a mover la tendencia central de la distribución en la misma dirección, manteniendo el parámetro de escala α constante. Asimismo, dicho aumento del parámetro genera valores extremos: la distribución se torna asimétrica, con valores extremos generados cada vez con mayor frecuencia.

- Se observa que, para un mismo parámetro q_t, la distribución se expande en la medida que se observe los cuantiles inferiores. Asimismo, en determinados casos pronuncia la asimetría identificada anteriormente.

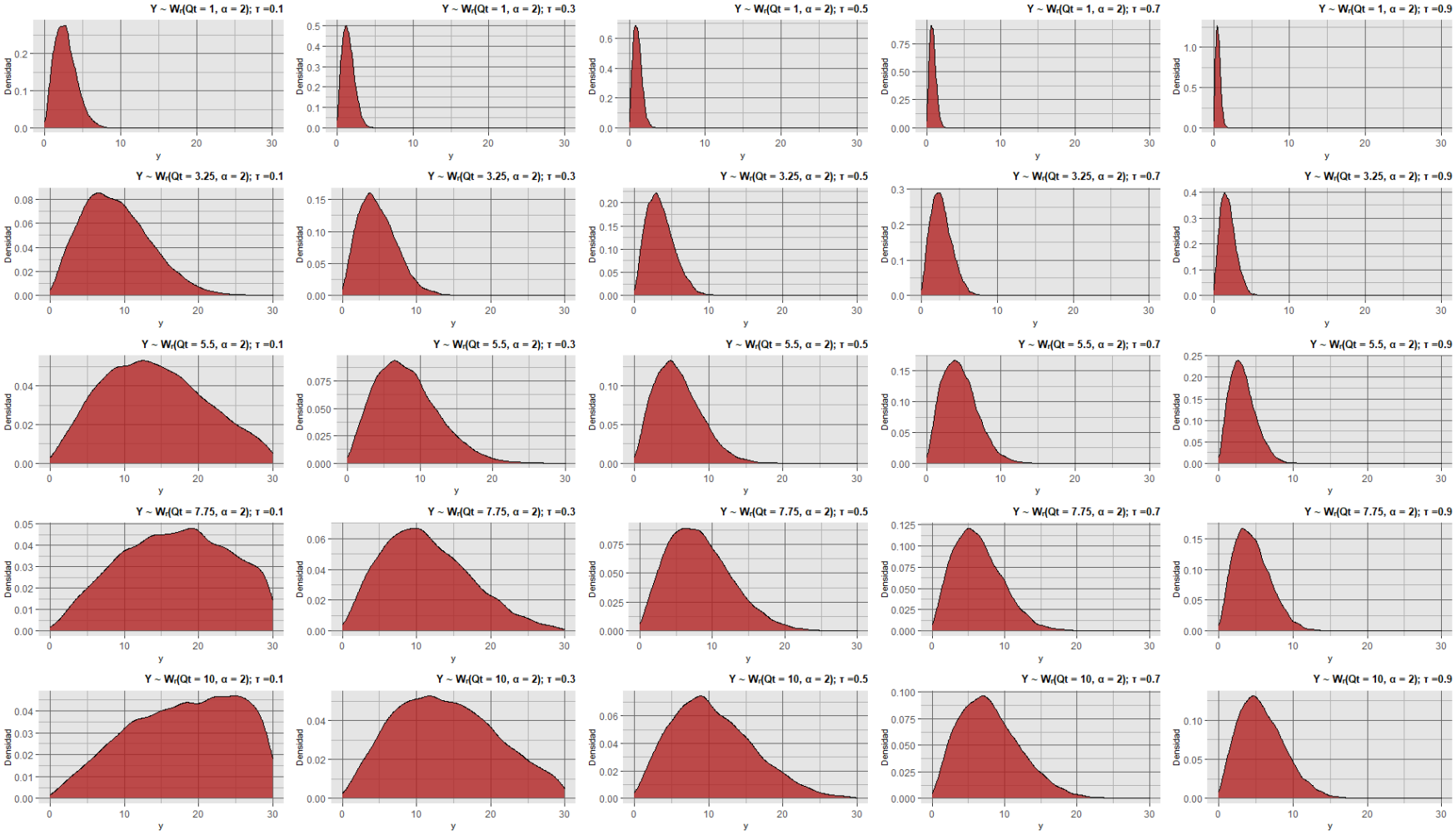


Figura 2.2: Estudio de la nueva parametrización

En base a la reparametrización propuesta, la función de distribución acumulada de $Y \sim W_r(q_t, \alpha, t)$ es de la forma:

$$F_Y\left(y|q_t, \alpha, t\right) = 1 - \exp\left(c(t)\left(\frac{y}{q_t}\right)^{\alpha}\right). \tag{2.3}$$

Asimismo, el valor esperado y varianza de dicha variable aleatoria está dada por:

$$E(Y) = \frac{q_t}{c(t)^{\frac{1}{\alpha}}}\Gamma\left(1 + \frac{1}{\alpha}\right) \tag{2.4}$$

$$Var(Y) = \frac{q_t^2}{c(t)^{\frac{1}{\alpha}}}\left[\Gamma\left(1 + \frac{2}{\alpha}\right) - \Gamma\left(1 + \frac{1}{\alpha}\right)^2\right] \tag{2.5}$$

Evaluamos las propiedades de la distribución por cada valor de los parámetros, conforme se observa en la figura 2.3. Se observa lo siguiente:

- **En relación al valor esperado:**
  - Para cada α fijo, el parámetro q_t mueve el valor esperado de la distribución hacia la dirección en la que dicho parámetro aumenta o disminuye. Ello se observa a lo largo de todos los posibles valores de α. No obstante, la fuerza en que afecta el valor esperado depende del valor de α, y se observa que el cambio en la esperanza disminuye en la medida que α aumente.
  - Para cada q_t fijo, el parámetro α dismuye el valor esperado de la distribución en la medida que dicho parámetro aumente. Este efecto es marginalmente menor por cada aumento de α, hasta tener una diferencia mínima cuando α considere valores cada vez más grandes. Asimismo, se observa que la esperanza es alta en la medida que α es pequeño, y esta propiedad aumenta rápidamente en la medida que α se acerque a 0.
- **En relación a la varianza:**
  - De forma similar que en la esperanza, el parámetro α disminuye considerablemente la varianza de la distribución. Esto tiene consistencia con la Figura 2.1, pues dicha variable no ha sido reexpresada en términos de la función cuantílica. Se observa, asimismo, que la varianza es alta en la medida que α es pequeño. Este efecto es mayor cuando el parámetro q_t aumenta conjuntamente.
  - El parámetro q_t aumenta la varianza de la distribución en la medida este aumente. No obstante, el efecto es modulado por el efecto del parámetro α: en la medida que α es alto, el efecto marginal en la varianza por cada aumento de q_t es pequeño. Por otro lado, cuando α es pequeño, el efecto marginal de la varianza por cada aumento de q_t es considerable e incrementa rápidamente.

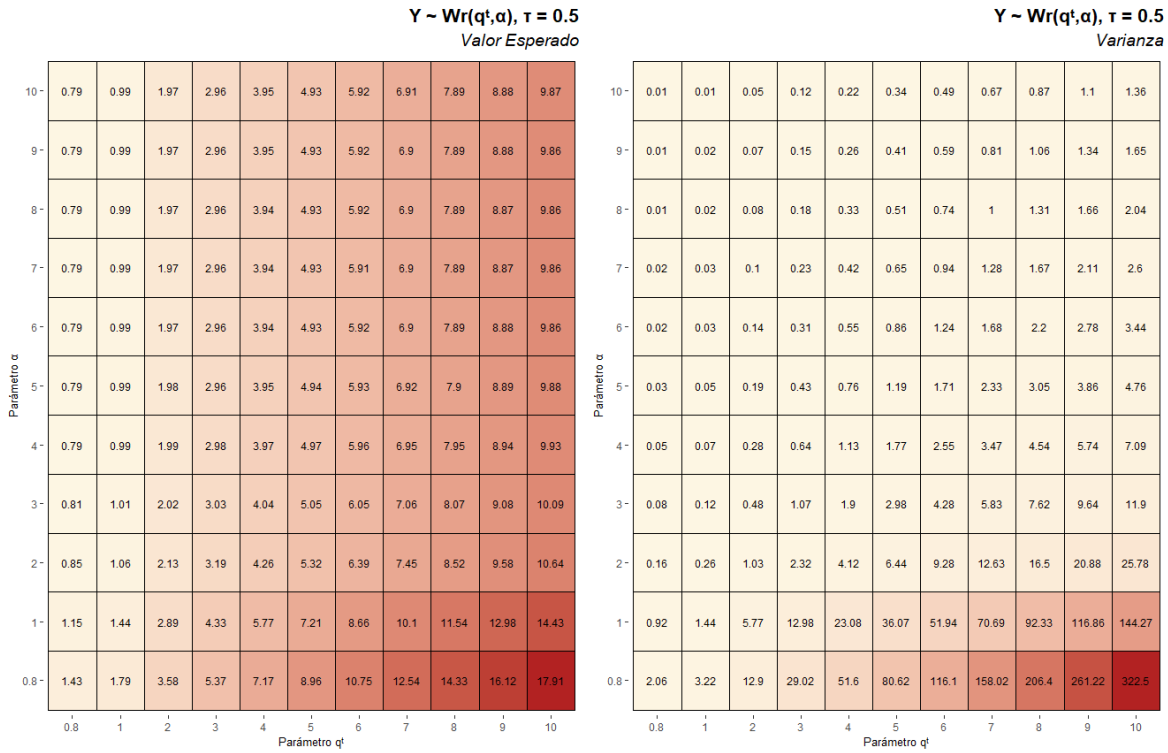


Figura 2.3: Valor esperado de una distribución Weibull bajo la parametrización propuesta.

Capítulo 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 la inferencia clásica.

3.1. Datos positivos con censura intervalar

Siguiendo la definición expuesta en Peto (1973), definimos a Y como una variable aleatoria con una **f.d.a.** $F_Y(y)$. Dicha variable se entiende como *censurada* si la única información que tenemos sobre ella es que Y yace en un intervalo I. Bajo este contexto, podemos definir una variable aleatoria Z como una variable indicadora que precisa el j-ésimo intervalo $[a_j, a_{j+1}]$, con $j = 1, \ldots, k$ en el que se encuentra la variable Y. Por lo tanto, durante el proceso de recolección de datos, observamos directamente la variable Z, mientras que la variable Y es una variable latente. Para ilustrar este proceso, imaginemos un proceso de administración de encuestas, en dónde el encuestador consulta a la persona en qué intervalo se encuentra su sueldo mensual. Esto requiere que la variable Z sea una variable categórica, pues la persona solo indica una opción. Entonces, podemos definir dicha variable mediante la siguiente expresión:

$$Z = \begin{cases} 1, a_1 < Y < a_2 \\ 2, a_2 \leq Y < a_3 \\ 3, a_3 \leq Y < a_4 \\ \vdots \\ k, a_k \leq Y < a_{k+1} \end{cases} \tag{3.1}$$

en dónde $a_1 < a_2 < \cdots < a_{k+1}$. Esto corresponde a los límites del intervalo I, con $a_1 = 0$ y $a_{K+1} = \infty$. La **f.d.p** de la variable observable Z está definida de la siguiente forma:

$$P(Z = j) = P(a_j \leq Y < a_{j+1}) = F_Y(a_{j+1}) - F_Y(a_j), j = 1, \ldots, k \tag{3.2}$$

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

$$Z \sim \text{Categórica}(\boldsymbol{\pi})$$

donde $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)^T$ y $\pi_j = P(Z = j)$.

Para efectos de la presente tesis, asumiremos que el proceso de censura de datos es independiente a la variable Y. Gomez et al. (2004) denomina esto como un proceso no informativo, pues ello indica que el conocimiento de que una observación se encuentra en el intervalo $[a_j, a_{j+1}]$ no precisa información adicional sobre la variable Y: solo indica que dicha variable está contenida entre esos límites. Conforme Self y Grossman (1986), esta suposición indica que dos valores específicos de Y, y_w y y_k, que se encuentren dentro del intervalo $[a_j, a_{j+1}]$, tienen la certeza de encontrarse en dicho intervalo.

3.2. Función de verosimilitud para datos positivos con censura intervalar

Bajo el contexto presentado anteriormente, y considerando las ideas plasmadas por Gentleman y Geyer (1994), el proceso de censura que deviene en la generación de la variable Z es independiente del proceso generador de datos de Y. Por lo tanto, la estimación de los parámetros que definen la distribución de Y, denotados por $\boldsymbol{\theta} = [q_t, \alpha]^t$, no es afectado por el proceso de censura. Bajo esta suposición, consideramos la verosímilitud de los datos con censura intervalar (es decir, los datos directamente observables) de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{\mathbb{I}(Z_i=j)} \tag{3.3}$$

Considerando los resultados identificados en la ecuación 3.2, la verosímilitud de la estructura observada de los datos es de la forma:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} (F_Y(l_i) - F_Y(u_i)) \tag{3.4}$$

en dónde l_i y u_i corresponden a los límites inferiores y superiores del intervalo en dónde se encuentra la i-ésima observación.

Bajo este criterio, la verosímilitud solo depende de los valores extremos del intervalo y de la **f.d.a.** de la variable latente Y.

3.3. Modelo de regresión para datos positivos con censura intervalar

Considerando la reparametrización expuesta en la sección 2, el modelo de regresión cuantílica, basado en la distribución Weibull, está dado por la siguiente expresión:

$$Y_i \sim W_r\left(q_{t_i}, \alpha, t\right).$$

$$g\left(q_{t_i}\right) = x_i^T \boldsymbol{\beta}.$$

en dónde $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$ 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 α, el parámetro q_{t_i} y t está definido conforme la sección 2.2. La estimación de los parámetros $\boldsymbol{\beta}$ y α se realizará mediante el método de máxima verosimilitud.

3.3.1. Función de verosimilitud

Consideramos que solo conocemos que Y_i se encuentra en un intervalo de K posibles intervalos de la forma $[a_j, a_{j+1}]$ con $a_1 < a_2 < \ldots < a_{k+1}$ y que $Z_i = j$ denota que $Y_i \in [a_j, a_{j+1}]$. Por lo tanto, considerando los resultados de la sección 3.1, tenemos que

$$Z_i \sim \text{Categórica}(\boldsymbol{\pi}_i).$$

con $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ik})$ tal que

$$\pi_{ij} = F_Y(a_{j+1}|q_{t_i}, \alpha, x) - F_Y\left(a_j|q_{t_i}, \alpha, x\right) \tag{3.5}$$

dónde $F_Y(\cdot|\cdot, \cdot, \cdot)$ es la **f.d.a** de la distribución Weibull reparametrizada dada en 2.2.Entonces la función de verosimilitud de las variables observadas $Z_1, Z_2, \ldots, Z_n$ es dada por lo siguiente:

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

Luego, considerando $[l_i, u_i]$ como el intervalo dónde Y_i fue observado, podemos escribir la función de verosimilitud como:

$$L\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{n} \left(F(u_i|q_{t_i}, \alpha, t) - F(l_i|q_{t_i}, \alpha, t)\right)$$

Luego, la función de log-verosímilitud es dada por:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(F(u_i|q_{t_i}, \alpha, t) - F\left(l_i|q_{t_i}, \alpha, t\right)\right)$$

$$l(\boldsymbol{\theta}) = \sum_{i}^{n} \log\left(\exp\left(-c(t)\left(\frac{u_i}{e^{x_i^T\boldsymbol{\beta}}}\right)^{\alpha} - \exp\left(-c(t)\left(\frac{l_i}{e^{x_i^T\boldsymbol{\beta}}}\right)^{\alpha}\right)\right)\right)$$

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

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

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \frac{c(t)}{(\gamma_i)^{\alpha}(\lambda_{i_2} - \lambda_{i_1})}\left((u_i)^{\alpha}\log\left(\frac{u_i}{\gamma_i}\right)\lambda_{i_2} - (l_i)^{\alpha}\log\left(\frac{l_i}{\gamma_i}\right)\lambda_{i_1}\right)$$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{\alpha c(t) x_{ij}}{(\gamma_i)^{\alpha}(\lambda_{i_1} - \lambda_{i_2})}\right)\left((a_{i_j})^{\alpha}\lambda_{i_1} - (a_{i_{j+1}})^{\alpha}\lambda_{i_2}\right)$$

en dónde:

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

$$\eta_i = x_i^T\boldsymbol{\beta}$$

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

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

Dichos estimadores de máxima verosimilitud, bajo ciertas condiciones de regularidad, son consistentes (es decir, que $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ cuando $n \to \infty$) y asíntoticamente normales con distribución:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\right)$$

cuando $n \to \infty$. $\mathcal{I}(\hat{\boldsymbol{\theta}})$ es la matriz de información de Fisher observada, la cual en nuestro modelo tiene la estructura:

$$I(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\partial l(\boldsymbol{\theta})}{\partial \alpha \partial \alpha^T} & \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \alpha} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \alpha} & \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \end{bmatrix} | \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$$

Los errores estándares de cada coeficiente se estiman a través de dicha matriz de información, denotada por $\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}$. Los errores estándares corresponden a la raíz cuadrada de cada elemento de la diagonal. Finalmente, en el marco de la inferencia clásica, los intervalos de confianza para cada parámetro está definido de la forma:

$$\hat{\theta_j} \pm z_{1-\frac{\alpha}{2}} C_{jj}.$$

dónde C_{jj} es la raíz del j-ésimo elemento diagonal de $I(\hat{\boldsymbol{\theta}})^{-1}$

3.4. Simulación de datos

En esta sección se presenta un estudio de simulación para evaluar si el método descrito en 3.3.1 permite recuperar los parámetros propuestos del modelo de censura intervalar para datos positivos. Para ello, se evaluará el desempeño de la simulación mediante tres criterios: el sesgo relativo, el error cuadrático medio y el ratio de cobertura.

3.4.1. Metodología para la simulación de datos

La presente sección tiene como objetivo realizar un estudio de simulación en el que se evalúe la adecuada estimación del modelo propuesto. Para ello, se generará un conjunto de datos dónde cada observación sigue la distribución $Y_i \sim W_r(q_t, \alpha, t)$. Luego, cada una de estas observaciones serán censuradas dando como resultado la variable Z_i, la cual sigue lo explicado 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 evaluar si el método de estimación planteado, permite recuperar adecuadamente los parámetros de regresión establecidos anteriormente. Los criterios sobre los cuales se analizará la estimación del modelo son: sesgo relativo, error cuadrático medio y cobertura.

El proceso de simulación consiste en generar 5,000 réplicas para cada tamaño de muestra de $n = \{100, 500, 1{,}000\}$. Simularemos la variable respuesta $Y_i \sim W_r(q_{ti}, \alpha, t)$ considerando 3 covariables X_{1i}, X_{2i}, X_{3i} que serán simuladas como:

$$X_{1i} \sim N(2, 0{,}25)$$

$$X_{2i} \sim Beta(2,3)$$

$$X_{3i} \sim Gamma(2,20)$$

Conforme lo mencionado en la sección 3.2.1, $q_{t_i} = \exp(x_i^T \boldsymbol{\beta})$, en dónde $\boldsymbol{\beta} = [7,0{,}3,0{,}84,2{,}5]^T$ y $x_i = (1, X_{1i}, X_{2i}, X_{3i})^T$. Por otro lado, el parámetro de dispersión tomará el valor $\alpha = 2$. Finalmente, se realizará la evaluación por los cuantiles $t = [0{,}1, 0{,}2, \ldots, 0{,}9]$.

Se asume que $Y_i \sim W_r(q_{ti}, \alpha, t)$ se observa con censura intervalar. En este estudio asumiremos que solo observamos una variable Z que particiona la variable Y_i en intervalos de igual amplitud, con la excepción del último intervalo, el cual tiene la estructura $[a_j, \infty)$. Una vez generada dicha variable, se realiza el modelamiento de la variable con censura intervalar sobre las variables independientes creadas previamente. El objetivo final es, a través del método de máxima verosímilitud, estimar los coeficientes $\boldsymbol{\beta}$ definidos previamente.

3.4.2. Implementación del modelo

La implementación del modelo se realizó a través del lenguaje de programación R, tomando en consideración las definiciones presentadas en el capítulo 3 de la presente tesis. El pseudocódigo de la implementación se encuentra en el Apéndice.

Una vez generadas las simulaciones, se evaluó para cada escenario (cuantil y tamaño de muestra) los siguientes indicadores:

$$\hat{\text{Sesgo relativo}}: \frac{1}{M} \sum_{j=1}^{M} \frac{(\hat{\theta}_j - \theta)}{\theta}$$

$$\hat{\text{ECM}}: \frac{1}{M} \sum_{j=1}^{M} (\hat{\theta}_j - \theta)^2$$

$$\hat{\text{Cobertura}}: \frac{1}{M} \sum_{j=1}^{M} \mathcal{I}(\boldsymbol{\theta} \in IC_j)$$

dónde θ es el verdadero valor del parámetro, $\hat{\theta}_j$ la estimación obtenida en la j-ésima réplica y M el número de réplicas. IC_j es el intervalo de confianza al 95 % obtenido en la j-ésima réplica.

3.4.3. Resultados

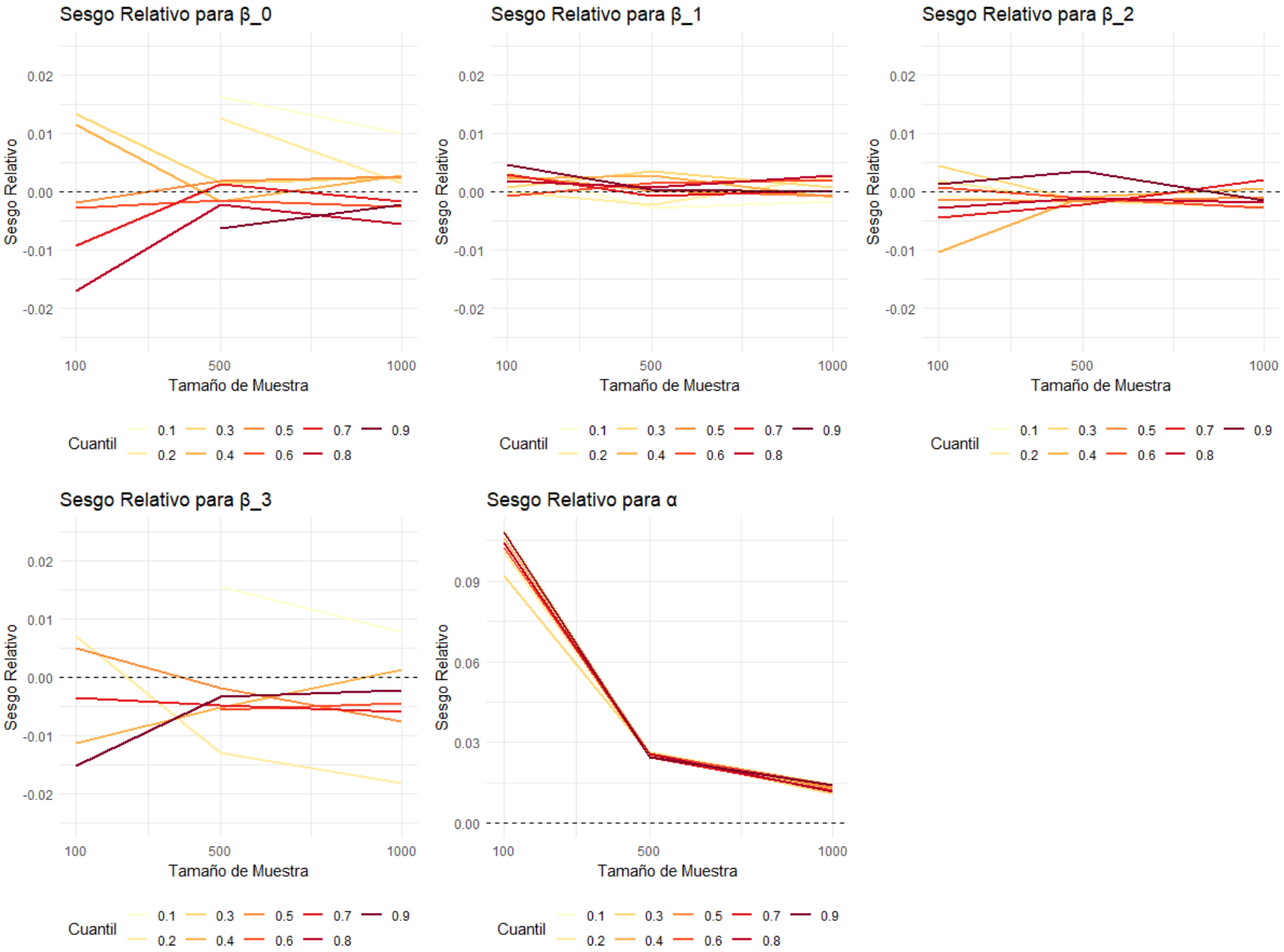
En las figuras 3.1, 3.2, y 3.3 se muestran la evaluación del rendimiento del modelo de regresión cuantílica con censura intervalar, de acuerdo a los criterios expuestos anteriormente. Observamos lo siguiente:

- En relación al sesgo relativo, se observa que este disminuye a lo largo de todos los parámetros en la medida que aumenta el tamaño de la muestra. Cabe resaltar que para tamaños de muestra pequeños, el parámetro α tiende a sobre-estimarse, no obstante esto disminuye considerablemente en la medida que el tamaño de muestra aumente.

- En relación a la cobertura, se observa que, para todos los tamaños de muestra, los parámetros establecidos en la sección precedente se encuentran aproximadamente el

95 % de las veces dentro del intervalo de confianza generado.

- En relación al error cuadrático medio, se observa que, para un tamaño de muestra pequeño, el error es considerable para todos los parámetros. No obstante, esto disminuye drásticamente en la medida que el tamaño de muestra aumenta.


Sesgo Relativo para β_0
Sesgo Relativo para β_1
Sesgo Relativo para β_2
Sesgo Relativo para β_3
Sesgo Relativo para α
Sesgo Relativo
Tamaño de Muestra
100
500
1000
0.02
0.01
0.00
-0.01
-0.02
0.09
0.06
0.03
Cuantil
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

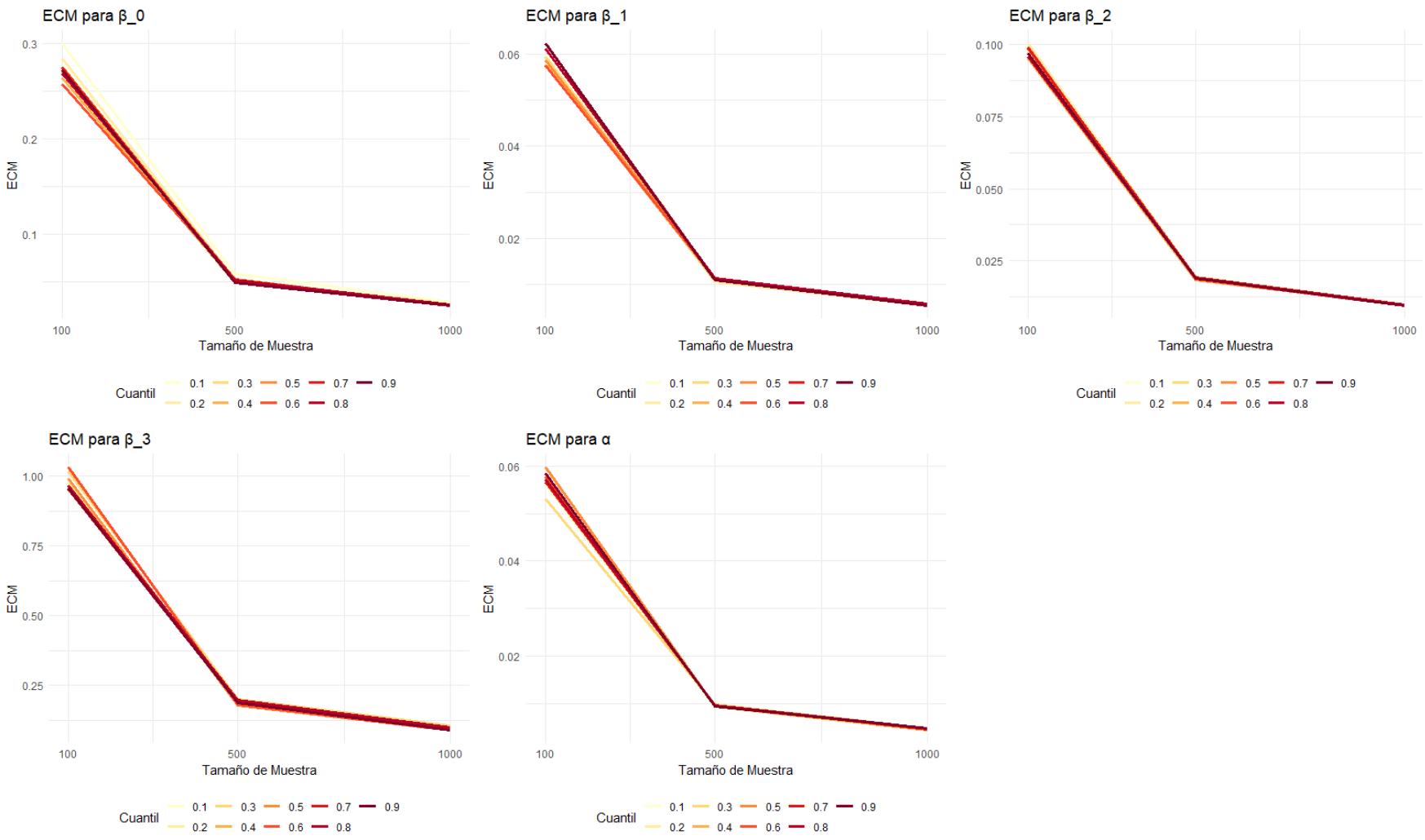


Figura 3.2: Estudio de Simulación: Análisis del error cuadrático medio

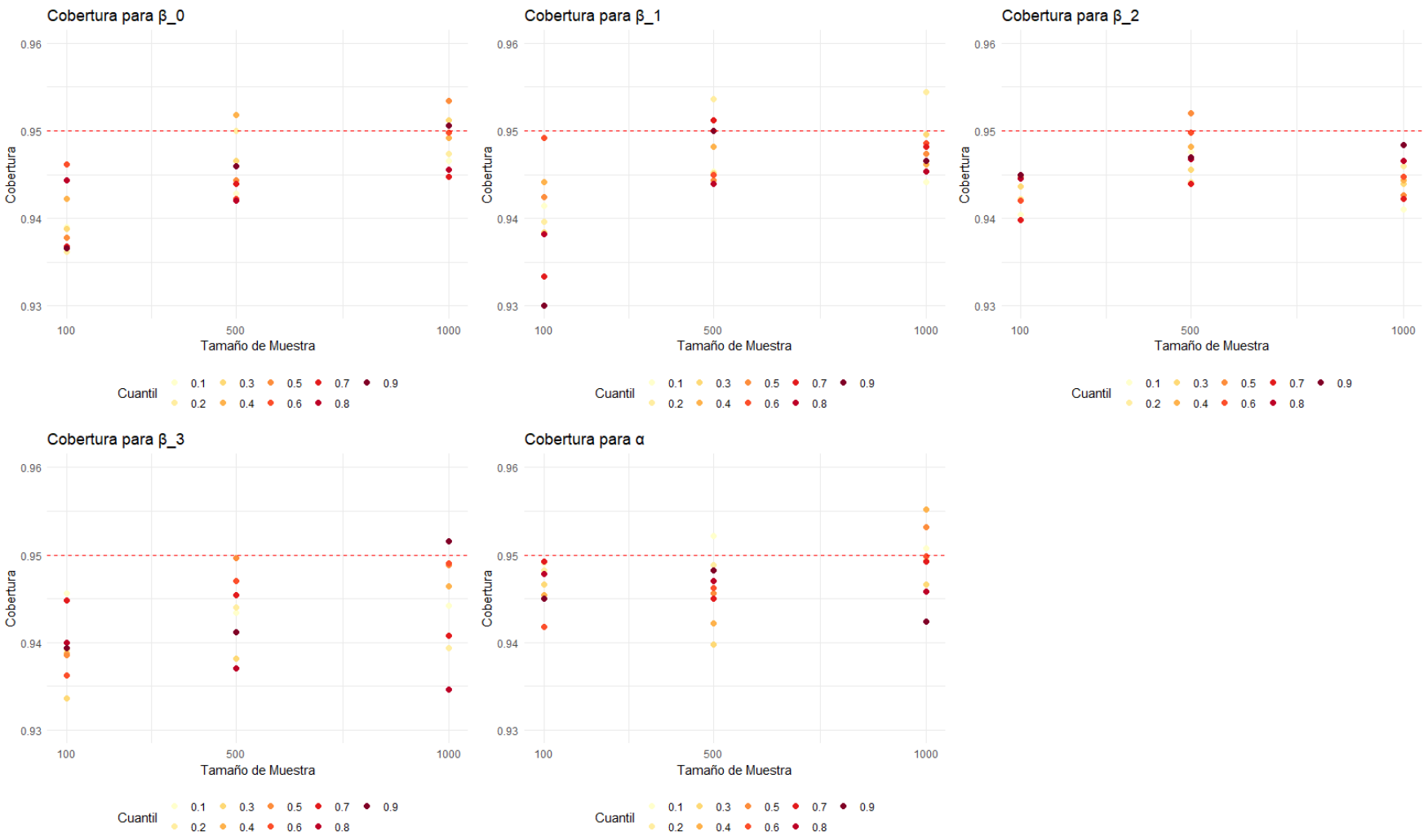


Figura 3.3: Estudio de Simulación: Análisis de la Cobertura

Capítulo 4

Aplicación en datos reales

El presente capítulo aplica el modelo propuesto en la sección 3.3 a encuestas de satisfacción de profesionales peruanos del sector salud. Se busca estimar los efectos de cada uno de los atributos en relación al sueldo reportado de dichos profesionales, con especial énfasis en el sexo de los mismos. Este énfasis está sustentado en estudios previos de instituciones del Gobierno del Perú y académicos. Al respecto, recientes investigaciones del INEI [1] identifican una brecha promedio de 29 % entre los sueldos de mujeres y hombres, siendo estos últimos quienes ganan más. Bajo dicha premisa, Sal y Rosas et al. (2019) también efectuaron el análisis a encuestas de profesionales de salud, identificando también dicha brecha. La presente aplicación tiene el objetivo de brindar una figura más completa de los efectos de las covariables por cada uno de los cuantiles del sueldo reportado.

4.1. Sobre los datos utilizados

Durante los años 2013 al 2015, instituciones gubernamentales del Perú, el INEI y la SUNASA [2] acordaron implementar y ejecutar encuestas nacionales de satisfacción en el sector salud. Ello comprende todas las ramas involucradas de dicho sector: los usuarios del servicio (de consulta externa, de boticas y farmacias, y de unidades de seguros) así como profesionales de la salud (médicos y enfermeros), por cada una de los conjuntos de establecimientos que existen. Estos comprenden establecimientos del Ministerio de Salud, Seguro Social de Salud, clínicas privadas, y establecimientos de Sanidad de las Fuerzas Armadas y Policiales. El objetivo final de la encuesta es «evaluar el grado de satisfacción de los usuarios internos y externos de los servicios de salud» INEI (2015).

Dicha encuesta formó parte de las investigaciones estadísticas realizadas por el INEI, cuyo diseño muestral fue probabilístico y polietápico. La primera unidad de muestreo constituyó el establecimiento de salud por cada uno de los conjuntos anteriormente expuestos, y la segunda unidad de muestreo constitió en los usuarios elegibles y profesionales de la salud. La encuesta tuvo alcance nacional, por los 24 departamentos del Perú. El nivel de inferencia indicado por el INEI es nacional y dirigida a cada una de las unidades de análisis.

Para própositos de la aplicación, se utilizó la encuesta ejecutada a profesionales de la salud. En esta, los datos obtenidos están relacionados con la formación académica, actividad laboral, satisfacción en el trabajo, estrés laboral y conocimiendo de la SUNASA INEI (2015).

De dicha encuesta, se utilizaron las siguientes variables:

[1] Instituto Nacional de Estadística e Informática del Perú.
[2] Superintendencia Nacional de Aseguramiento en Salud.

Grupo de Variable	Descripción de la variable	Código
Establecimiento de Salud	Tipo de institución del establecimiento	INSTITUCION
Formación del Profesional de Salud	¿El profesional cuenta con especialidad?	C2P13
Actividad Laboral del Profesional de salud	Años de experiencia en el sector salud	C2P21
	¿El profesional realiza labor asistencial en otra institución?	C2P24
	¿El profesional realiza labor docente remunerada?	C2P26
	Cantidad de horas laboradas semanalmente por el profesional de salud	C2P27
Atributos del Profesional de Salud	Sexo del profesional de salud	C2P4
	Rango Salarial	[li, lf]

Cuadro 4.1: Descripción de las variables dentro de la base de datos

4.1.1. Análisis descriptivo de los datos

En la figura 4.1 se puede observar que existe una mayor proporción de profesionales médicos varones en los intervalos salariales superiores que profesionales mujeres por cada una de las instituciones de la encuesta. Cabe resaltar que se observa una mayor proporción de encuestados en los establecimientos del Ministerio de Salud y ESSALUD, no obstante la proporción de varones en la banda salarial más alta se mantiene a lo largo de todos los establecimientos. Asimismo, en la figura 4.2 se puede observar que para las bandas salariales más bajas, las mujeres tienen más años de experiencia que los hombres (observando la mediana); no obstante se mantienen en la misma banda salarial. Lo anteriormente precisado sugiere que las mujeres tienen una menor probabilidad de pertenecer a las bandas salariales superiores. No obstante, esto se verificará a través de la aplicación del modelo de censura intervalar.

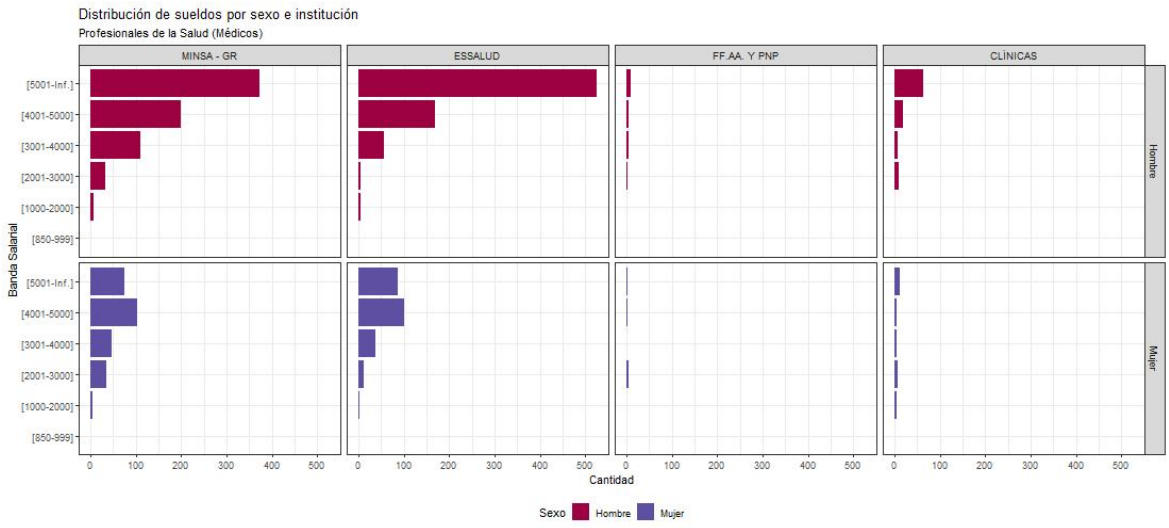


Figura 4.1: Distribución de sueldos por sexo e institución

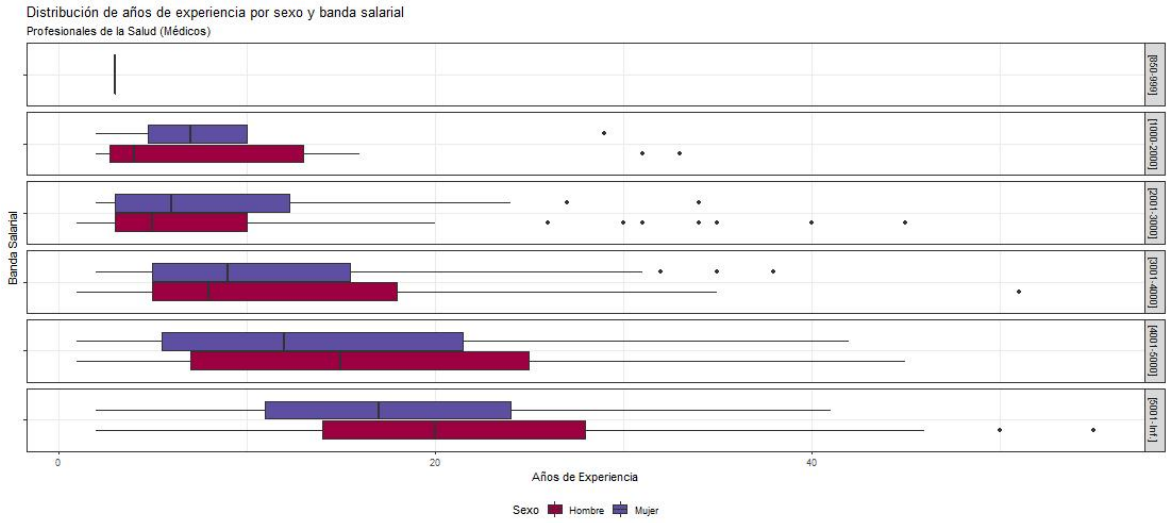


Figura 4.2: Años de experiencia por sexo y banda salarial

4.2. Resultados

Tomando como variable respuesta la banda salarial, se ajustó el modelo de regresión cuantílica para datos intervalares a los datos. Dado que existen variables categóricas en la base de datos, la categoría de referencia para el presente estudio es un médico hombre, con labor docente remunerada y estudios de especialización, y que trabaja en un establecimiento del MINSA. Adicionalmente, realizaría labor asistencial en otra institución. En ese sentido, el intervalo puede ser interpretado como una estimación del sueldo de dicha categoría de referencia a lo largo de los cuantiles de la variable respuesta. Para el ajuste del modelo, se consideró una función de enlace logarítmica para los parámetros $\boldsymbol{\beta}$ y α.

La figura 4.3 presenta un resumen de los resultados de la aplicación del modelo. Cada uno de los gráficos presentados tiene como eje horizontal el cuantil τ, y el eje vertical corresponde al efecto de la covariable (sin ajustar por la función de enlace). En dicho gráfico, se tiene tanto la estimación por el modelo de regresión cuantílica para datos con censura intervalar y un modelo de regresión censurada (el cual se realizó mediante el paquete GAMLSS). En azul, se tienen los efectos de cada covariable estimados por el modelo propuesto en la sección 3.3, así como los intervalos de confianza al 95 %. En negro y gris, se tiene la estimación mediante un modelo de regresión Weibull que estima la media.

En relación al efecto medio, se observa que un profesional de salud mujer tendría un decremento salarial en relación a la categoría de referencia. No obstante, el análisis se enriquece cuando tomamos en consideración el modelo de regresión cuantílica para datos intervalares, pues se observa que para los cuantiles inferiores existe un considerable efecto negativo para dicha variable. Esta disparidad entre el efecto medio y los efectos por cada cuantil se pueden apreciar adicionalmente en la cantidad de horas trabajadas (C2P27), pues se observa que el efecto positivo de las horas trabajadas es mayor en la medida que se analiza los cuantiles superiores. Por otro lado, el efecto negativo de no contar con una especialización (C2P13) es cada vez menor en la medida que se evalúan los cuantiles superiores.

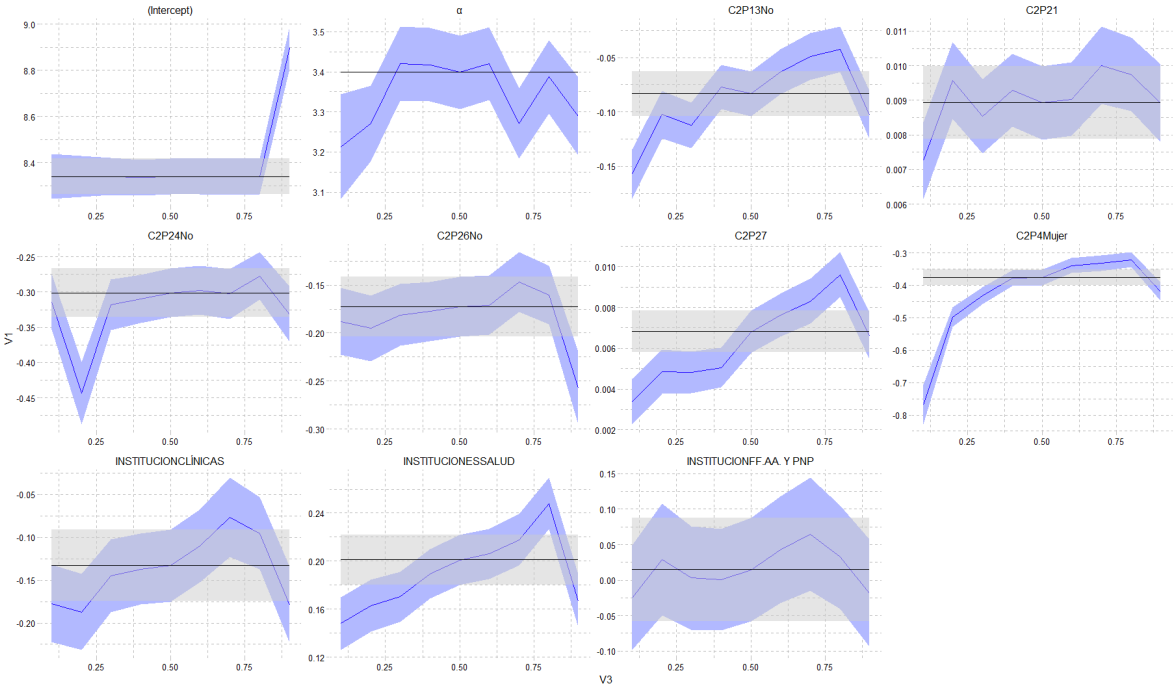


Figura 4.3: Efectos de las covariables sobre los cuantiles del sueldo de los profesionales de la salud.

Capítulo 5

Conclusiones

5.1. Conclusiones

Los datos con censura intervalar presentan retos en el proceso de modelamiento de datos, pues la no-observabilidad de los mismos requiere adaptar los procesos de inferencia clásica a esta estructura. Ante ello, la presente tesis estudió un modelo de regresión cuantílica para datos con censura intervalar, atendiendo los estudios realizados anteriormente por Peto (1973), Gentleman y Geyer (1994) y Koenker y Bassett (1978). Dicho modelo de regresión es paramétrico, asumiendo que la variable latente sigue una distribución Weibull, la cual fue reparametrizada para estudiar los efectos de las covariables en distintos cuantiles de la variable respuesta.

Para evaluar el modelo propuesto, se realizó un estudio de simulación para diversos cuantiles y distintos tamaños de muestras. Se observó que el estimador propuesto captura apropiadamente los parámetros poblacionales, y que el sesgo y error cuadrático medio se redució en la medida que aumentó el número de observaciones. La cobertura de los intervalos de confianza fue apropiada en todos los tamaños de muestra.

Finalmente, se aplicó el modelo de regresión a datos de la Encuesta Nacional de Satisfacción de Usuarios en Salud (ENSUSALUD) 2015. En dicha encuesta, el sueldo de los profesionales de salud (médicos/as y enfermeros/as) se censuró desde el proceso de recolección de datos. Atendiendo al estudio realizado por Sal y Rosas et al. (2019), la presente tesis extiende el modelo de regresión de censura intervalar expuesto a un modelo de regresión cuantílica. El presente modelo permitió analizar los factores de las covariables en relación al sueldo de dichos profesionales, por cada uno de los cuantiles de la variable respuesta.

5.2. Sugerencias para investigaciones futuras

- Establecer un método de verosimilitud que tome en cuenta los pesos muestrales de la encuesta realizada. Asimismo, proponer métodos para la estimación de los errores estándar atendiendo esta estructura.

- Proponer un modelo de regresión cuantílica con censura intervalar bajo inferencia bayesiana, tomando en consideración los métodos expuestos en la presente tesis.

Capítulo 6

Apéndice

6.1. Pseudocódigo de la simulación

```
Simulamos valores de las siguientes distribuciones:

Definimos los siguientes valores:
N = [100, 500, 1000]
B = [7, 0.3, 0.84, 2.5]
Sigma = 2
t=[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
M = 5000

Para cada cuantil en t:
        Para cada n en N:
                Para cada replica en M:
                1 Simular n valores de las siguientes distribuciones:
                        X1 ~ Beta(2,3)
                        X2 ~ Normal(2,0.5)
                        X3 ~ Gamma(2,25)
                2 Generar la función de enlace:
                        Qt = exp(B[1] + B[2]*X1 + B[3]*X2 + B[4]*X3)
                3 Para cada i en n:
                        Simular 1 valor de la siguiente distribucion:
                        Y[i] ~ W_r(Qt[i], Sigma, cuantil)
                4 Censurar la variable Y de forma intervalar tal que
                        Z ~ Categorica
                5 Obtener los limites inferiores y superiores de
                        cada categoria de Z
                6 Crear la base de datos simulada
                        df <- [L_inf, L_sup, X1, X2, X3]
                7 Ejecutar la regresion de censura intervalar
                8 Guardar los resultados
```

Bibliografía

Gentleman, R. y Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation, *Biometrika* **81**(3).

Gomez, G., Calle, L. y Oller, R. (2004). Interval censoring: model characterizations for the validity of the simplified likelihood, *Canadian Journal of Statistics* **32**(3).

INEI (2015). Encuesta nacional de satisfaccion de usuarios del aseguramiento universal en salud.

Koenker, R. y Bassett, G. (1978). Regression quantiles, *Econometrica* **46**(1).

Koenker, R. y Hallock, K. (2001). Quantile regression, *Journal of Economic Perspectives* **15**(4).

Munoz, l. y Xu, J. (1996). Models for the incubation of aids and variations according to age and period, *Statistics in Medicine* **15**(1).

Peto, R. (1973). Experimental survival curves for interval-censored data, *Journal of the Royal Statistical Society* **22**(1).

Sal y Rosas, V., Moscoso-Porras, M., Ormeno, R., Artica, F., Miranda, J. y Bayes, C. (2019). Gender income gap among physician and nurses in peru: a nationwide assessment, *The Lancet* **7**(4).

Self, S. G. y Grossman, E. A. (1986). Linear rank tests for interval-censored data with applications to pcb levels in adipose tissue of transformer repair workers, *Biometrics* **42**(3).

Weibull, W. (1951). A statistical distribution funcion of wide applicability, *Applied Mechanics Division* .

Zhou, X., Feng, Y. y Du, X. (2016). Quantile regression for interval censored data, *Communications in Statistics - Theory and Methods* .