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

Dedicatoria

Dedicatoria

Agradecimentos

A mi asesor Cristian Bayes y al profesor Giancarlo Sal y Rosas, quienes ofrecieron la

Resumen

Palabras clave: censura intervalar, regresión con censura.

Abstract

Abstract

Keywords: keyword1, keyword2, keyword3.

Índice general

Lista de Abreviaturas

fdp	Función de densidad de probabilidad.
pBF	Pseudo factor de Bayes(*Pseudo bayes factor*).

Lista de Símbolos

μ Media.

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.

Este tipo de datos han sido estudiados previamente. Siguiendo las ideas de Peto (1973), una variable C se le denota censurada cuando su valor c no se conoce y la única información sobre la misma es un intervalo no-cero 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 $[L_f, \infty]$). El presente estudio se enfoca en el segundo tipo.

Naturalmente, ello trae retos en el proceso de modelamiento de datos. Los modelos estándares de regresión presumen que la variable respuesta es directamente observable. No obstante, en situaciones como la precisada en el párrafo precedente dichos modelos tienen que adaptarse a la estructura de los datos. Estos modelos han sido explorados con anterioridad: Gentleman y Geyer (1994) investigaron cómo determinar la máxima verosimilitud de los datos censurados, asegurar su consistencia e identificar métodos algorítmicos para su cómputo. Utilizando los puntos de corte del dato, L_i y L_f, era posible identificar la máxima verosímilitud a través de la diferencia entre las funciones de distribución acumulada en dichos puntos. Posteriormente, métodos de regresión lineal atendiendo esta estructura fueron explorados por Lindsey (1998) de forma paramétrica. Un recuento de este tipo de métodos se encuentra en Gomez et al. (2004).

Cabe resaltar que dichos métodos de regresión lineal modelan la respuesta esperada de la variable respuesta condicionada por un 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

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.

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

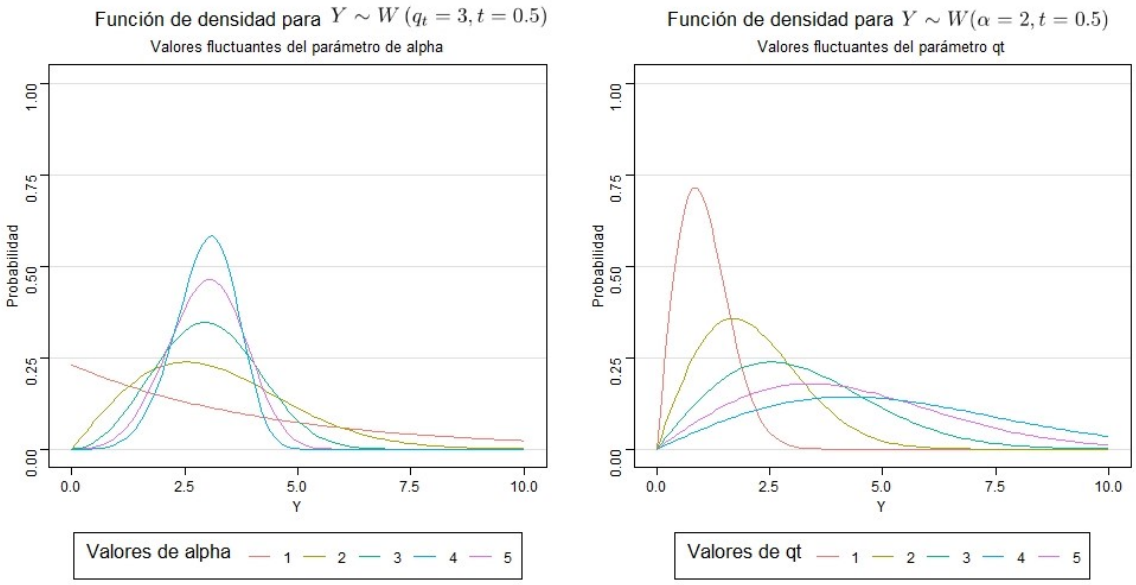

Figura 2.1: Función de densidad de una distribución Weibull bajo la reparametrización propuesta.

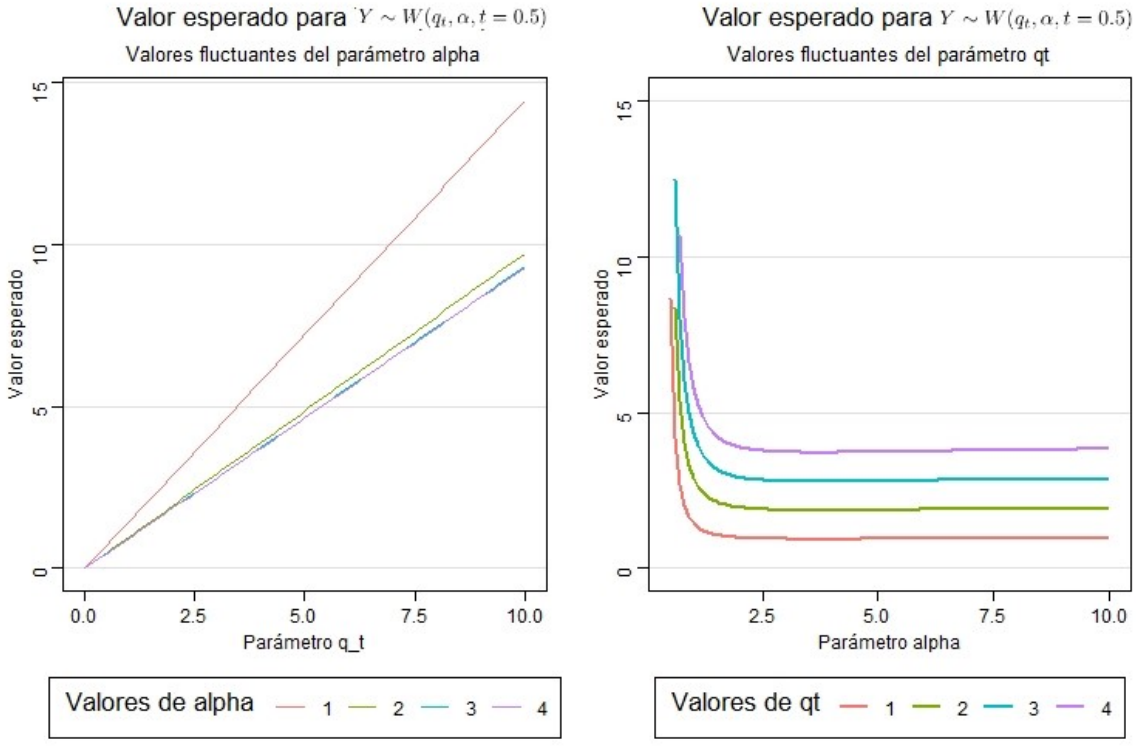

Figura 2.2: Valor esperado de una distribución Weibull bajo la parametrización propuesta.

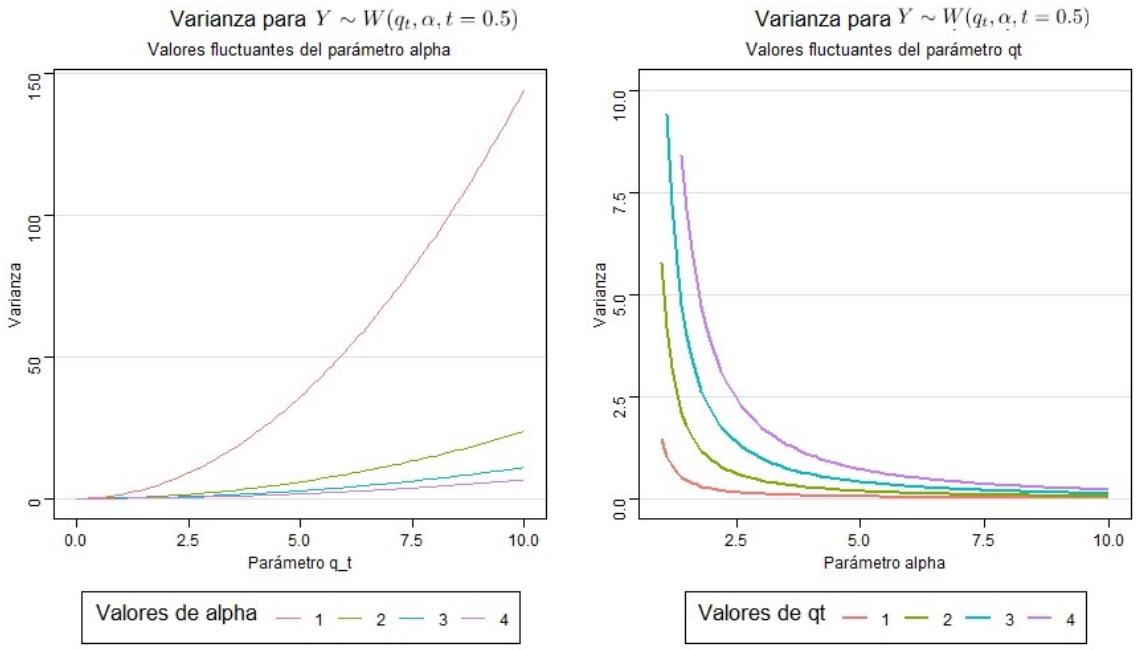

Figura 2.3: Varianza bajo la parametrización propuesta.

Apéndice A

Resultados teóricos

Bibliografía

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

Gomez, G., Calle, M. L. y Oller, R. (2004). Frequentist and bayesian approaches for interval-censored data, *Statistical Papers* **45**(1).

Lindsey, J. (1998). A study of interval censoring in parametric regression models, *Lifetime Data Analysis* **4**(4).

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