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 con el fin de obtener una respuesta. Dichos datos censurados 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 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 $[L_f, \infty]$). 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. Cabe resaltar que los modelos anteriormente expuestos identifican el valor esperado 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.

Pendiente: Texto explicando los estudios de regresión cuantílica con censura intervalar

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 método de regresión cuántiliza adaptado a datos con censura intervalar. Para identificar que el modelo propuesto es adecuado, aplicaremos la regresión 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 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 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 escala, con $\alpha > 0$, y σ corresponde al parámetro de forma, 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 densidad 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 1, en dónde se observa que el parámetro de forma permite ..... Asimismo, el parámetro de escala permite identificar lo junta que puede ser la distribución,

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

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

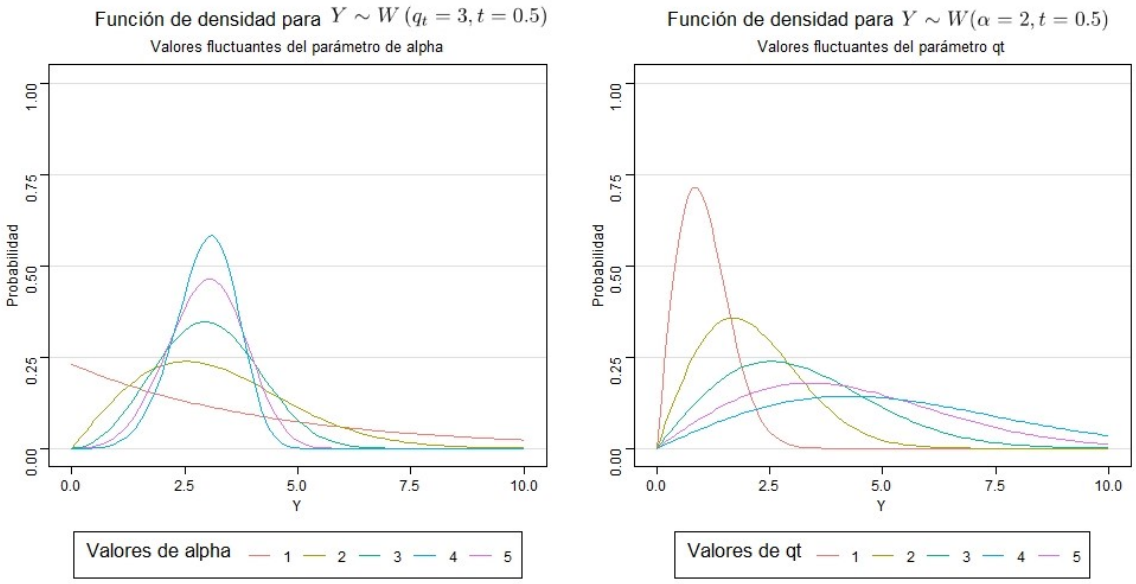

Figura 2.1: Función de densidad de una distribución Weibull bajo la reparametrización propuesta.

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

2.1.1. Proposición 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:

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

en dónde 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 $Y \sim W_r(q_t, \alpha)$.

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

Nota del profesor Valdivieso: Mejorar esta densidad.

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

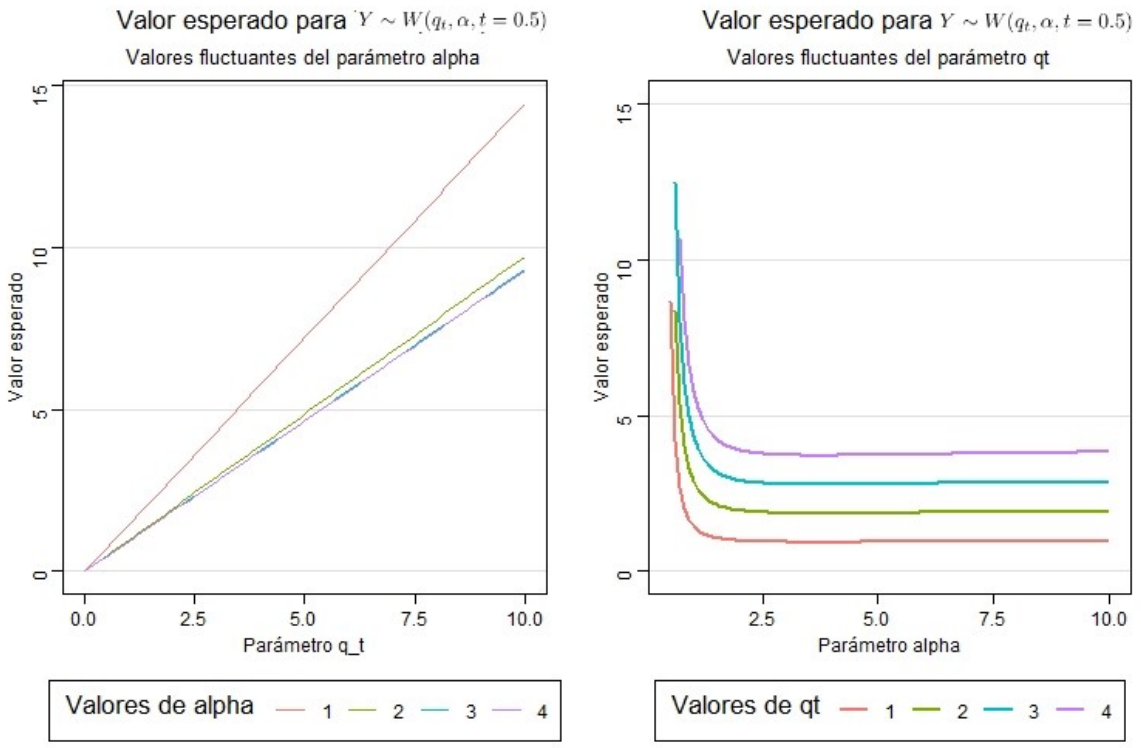

Figura 2.2: Valor esperado de una distribución Weibull bajo la parametrización propuesta.

tendría la siguiente forma:

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

2.1.2. Estudio de la parametrización propuesta

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

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

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

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:

Nota del profesor Valdivieso: Mejorar esta densidad.

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

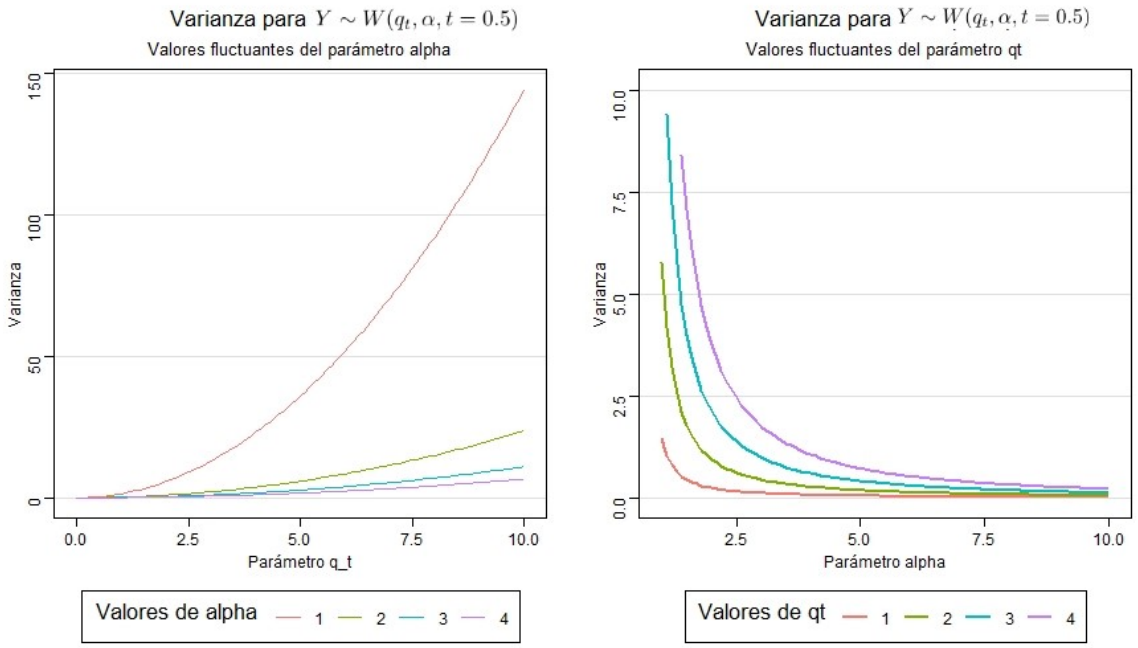

Figura 2.3: Varianza bajo la parametrización propuesta.

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.

Apéndice A

Resultados teóricos

Bibliografía

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

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).

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