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.

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 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.

Los modelos anteriormente expuestos tienen como propósito modelar el valor esperado due 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 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 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.

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 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 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

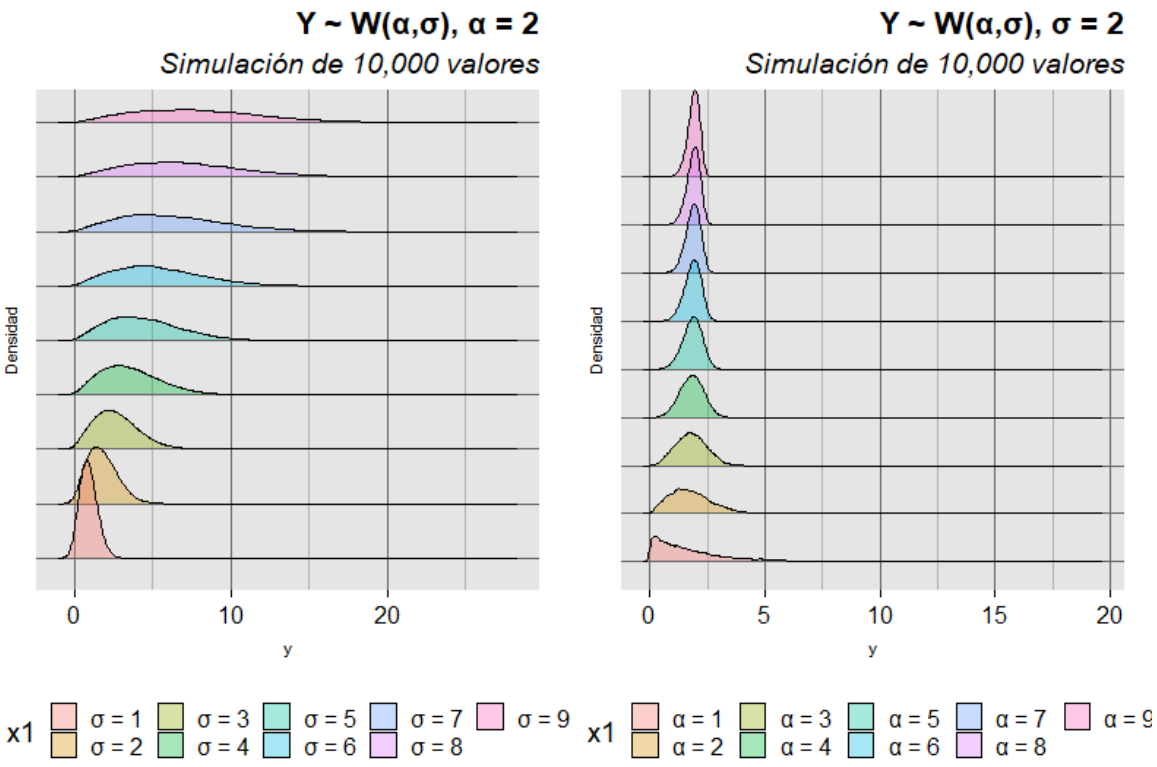

Figura 2.1: Estudio de una variable Y con distribución Weibull

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 define 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 $(q_t, \alpha) \in (0, \infty)$ 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 $c(t) = (-\log(1-t))$.

Los parámetros q_t y α, así como el 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.

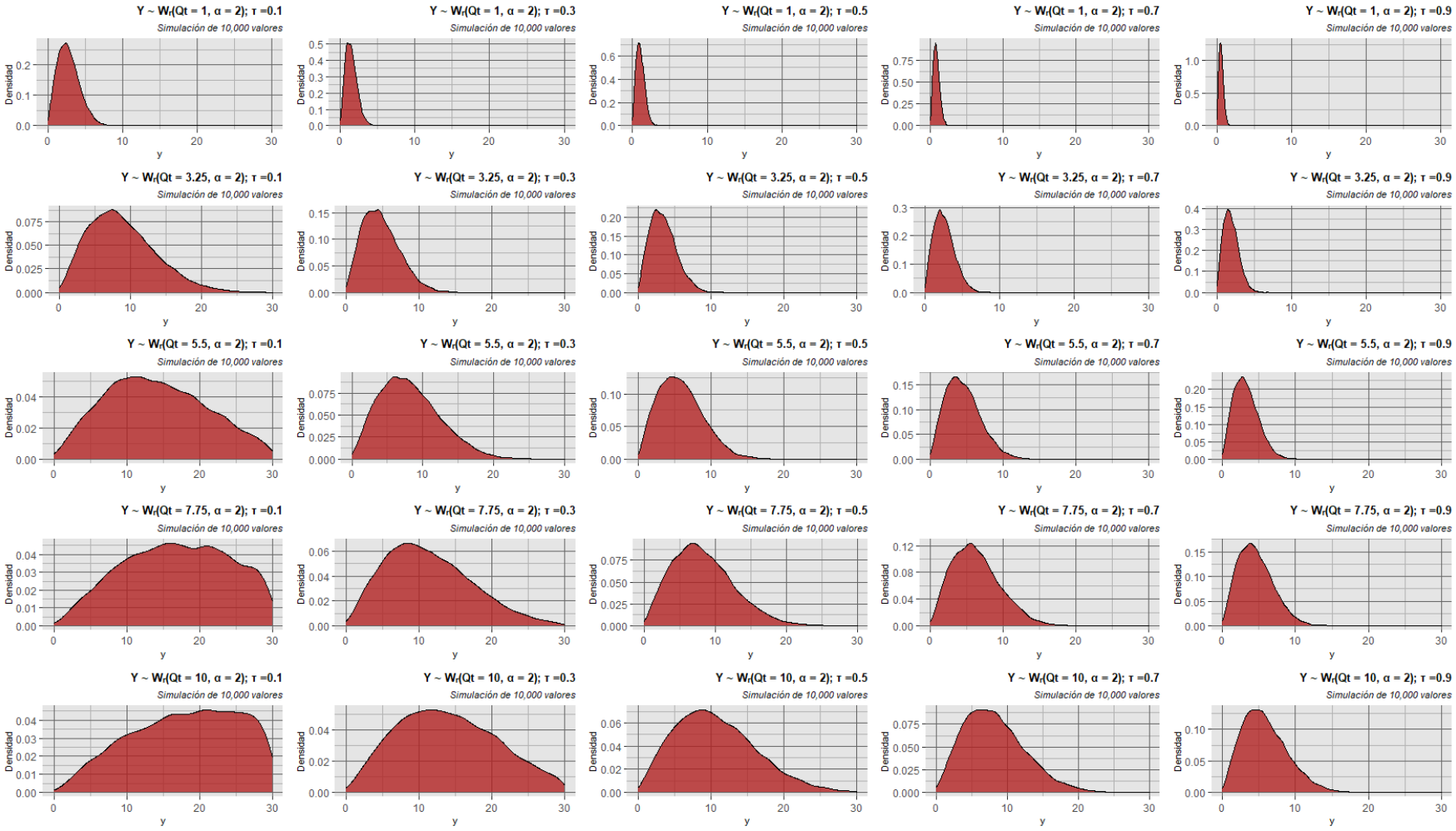

Figura 2.2: Estudio de la nueva parametrización

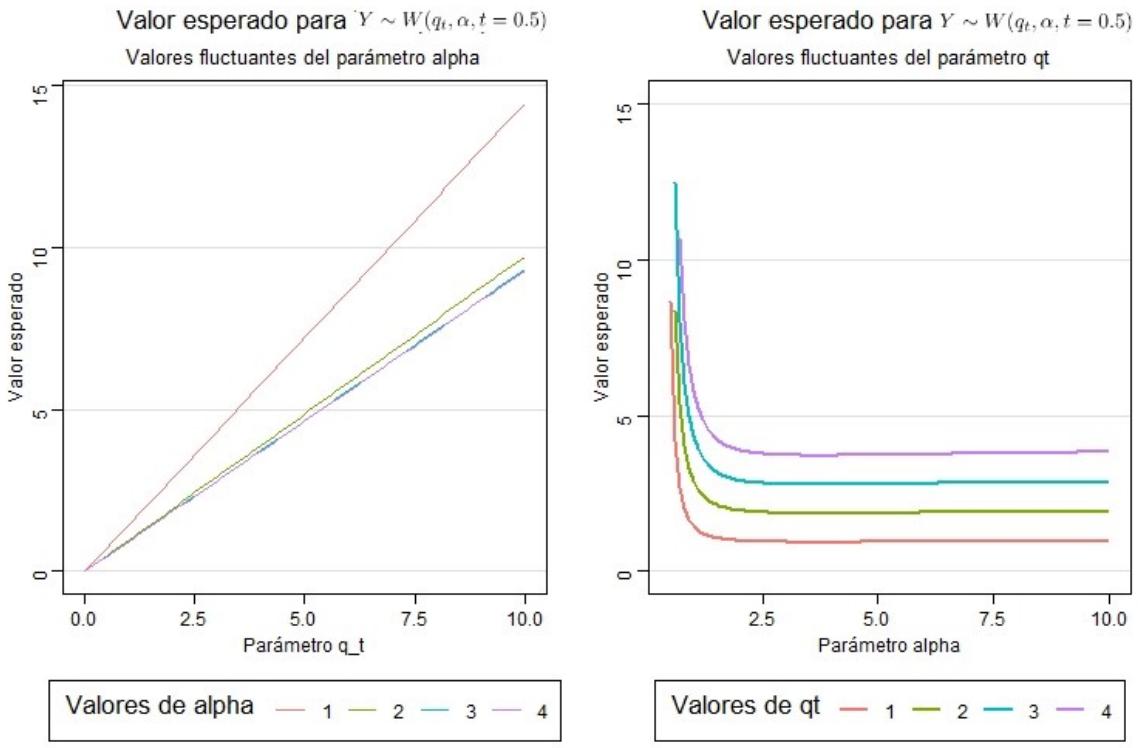

Figura 2.3: Valor esperado de una distribución Weibull bajo la parametrización propuesta.

En base a la reparametrización propuesta, la función acumulada de una variable Y que siga dicha distribución sería de la forma:

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

En torno a dicha reparametrización, la esperanza y varianza de una variable aleatoria Y está dada por las siguientes expresiones:

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

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

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:
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

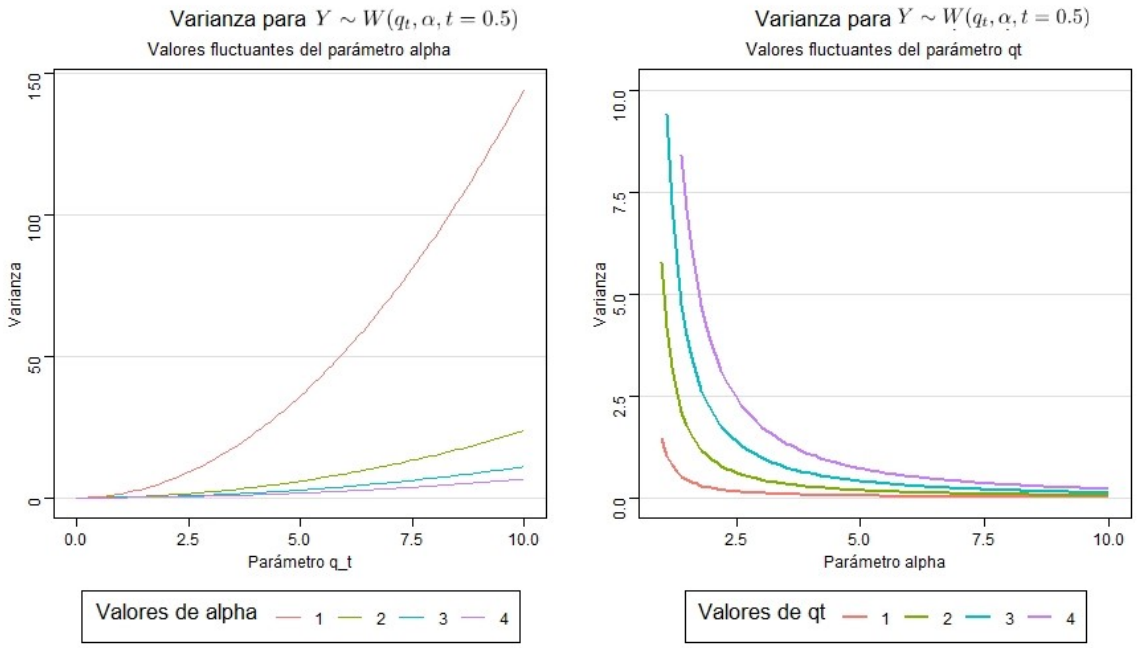

Figura 2.4: Varianza bajo la parametrización propuesta.

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).

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).

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* .