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 de forma intervalar.

Naturalmente, ello trae retos en el proceso de modelamiento de datos. Los modelos estándares de regresión presumen que el dato 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, asegurar su consistencia e identificar métodos algorítmicos para su cómputo. Utilizando los puntos de corte del dato (el inicio y final del intervalo), era posible identificar la máxima verosímilitud a través de la diferencia entre las funciones de distribución acumulada para dichos puntos. Posteriormente, métodos de regresión lineal atendiendo esta estructura fueron explorados por Zhang y Li (1996), quien (... pendiente)

(párrafo sobre regresión cuantílica)

1.1. Objetivos

El objetivo general de la tesis es estudiar propiedades, estimar y aplicar a conjuntos de datos reales el modelo de regresión beta desde el punto de vista de la estadística bayesiana. De manera específica:

- Revisar la literatura acerca de los diferentes propuestas de modelos de regresión para proporciones.

- Proponer estudiar, propiedades, e implementar la estimación del modelo de regresión beta desde la perspectiva bayesiana.

- Realizar estudios de simulación de computación intensiva aprovechando el uso de un GRID computacional (proyecto LEGION).

- Aplicar el modelo a conjunto de datos reales.

1.2. Organización del Trabajo

En el Capítulo 2, presentamos conceptos previos al desarrollo de ...

Finalmente, en el Capítulo 6 discutimos algunas conclusiones obtenidas en este trabajo. Analizamos la ventajas y desventajas de los métodos propuestos,...

En el anexo presentamos algunas pruebas de resultados en más detalles (Apéndice A) y también los programas utilizadas en las aplicaciones a conjuntos de datos reales.

Capítulo 2

Distribución Beta

Introducción al capítulo

2.1. Distribución Beta

Una clase de distribuciones que permite modelar variables continuas limitadas al intervalo $(0,1)$ es la distribución beta.

2.1.1. Función de densidad de probabilidad

la función de densidad de probabilidad de una variable aleatória Y que sigue una distribución beta es dada por

$$f_Y(y \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1. \tag{2.1}$$

Esta distribución es bastante flexible para modelar este tipo de datos debido a que su función de densidad de probabilidad puede tomar diversas formas dependiendo del valor de los parámetros que caracterizan a esta distribución (ver figura 2.1).

2.1.2. Propiedades

La media y la variancia de una distribución son expresadas por

$$E(y) = \frac{\alpha}{\alpha+\beta} \quad \text{y} \quad Var(y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{2.2}$$

2.2. Parametrización alternativa

Por este motivo, recientemente en la literatura han sido propuestos modelos de regresión basados en esta distribución, por ejemplo ver Ferrari y Cribari-Neto (2004). El modelo propuesto por estos autores permite la interpretación de los parámetros en términos de la respuesta en su escala original (variable respuesta sin transformar). Ellos consideran una reparametrización del modelo donde $\mu = \alpha/(\alpha+\beta)$ y $\phi = \alpha+\beta$, y de (2.2) obtenemos que

$$E(y) = \mu \quad \text{y} \quad Var(y) = \frac{V(\mu)}{1+\phi} \tag{2.3}$$

donde $V(\mu) = \mu(1-\mu)$, así tenemos que μ es la media de la variable respuesta y ϕ puede ser interpretado como un parámetro de precisión. Luego, en la nueva parametrización la densidad de la distribución beta puede ser escrita como

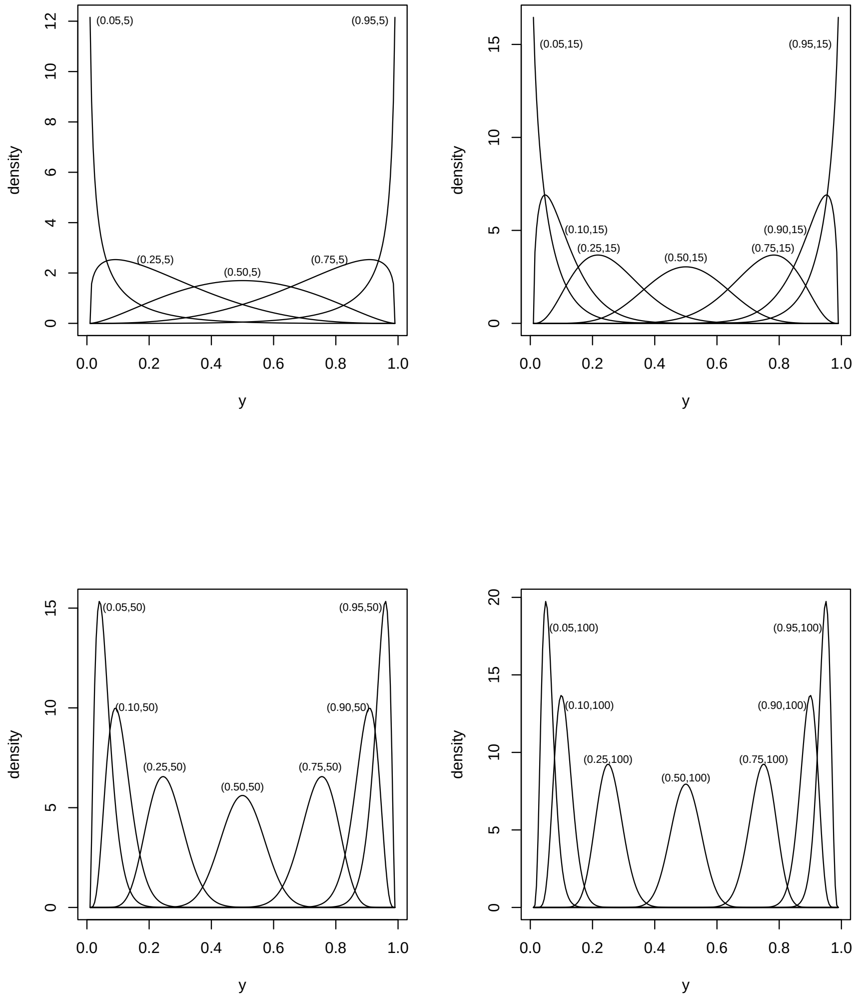

Figura 2.1: Densidades de la distribución beta para diferentes valores de los parámetros

$$f_Y(y \mid \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1. \tag{2.4}$$

donde $0 < \mu < 1$ y $\phi > 0$.

Capítulo 3

Modelo de Regresión Beta

El modelo de regresión para $Y_1, ..., Y_n$ variables aleatorias que siguen una distribución beta dada en (2.4) com parámetros μ_i y ϕ será obtenido cuando consideremos

$$g(\mu_i) = \sum_{i=1}^{k} x_{ij}\beta_j \tag{3.1}$$

donde $\beta = (\beta_1, ..., \beta_k)$ es el vector de parámetros de regresión, $x_{i1}, ..., x_{ik}$ son valores de k covariables y $g(.)$ es una función de enlace estrictamente monótona y dos veces diferenciable que va de $(0, 1)$ a $\Re$.

Desde un punto de vista clásico, Ferrari y Cribari-Neto (2004) consideran la estimación de los parámetros del modelo por el método de máxima verosimilitud. De este modo la inferencia en el modelo está basada en propiedades asintóticas de estos estimadores considerando una muestra suficientemente grande.
Sin embargo, en muchas situaciones prácticas el tamaño de muestra puede ser pequeño, o no siéndolo suele existir alguna información preliminar acerca de los parámetros de interés. Estas situaciones son perfectamente abordados desde la perspectiva bayesiana (ver Congdon, 2003) Para la regresión Beta este enfoque fue considerado por Branscum et al. (2007). Por este motivo, en la presente tesis se desarrollará el análisis bayesiano del modelo de regresión beta. Entre los tópicos a ser estudiados adicionalmente consideramos: la estimación del modelo utilizando MCMC, selección de modelos utilizando diferente tipos de criterios y aplicaciones a conjuntos de datos reales.

Capítulo 4

Estudio de Simulación

Capítulo 5

Aplicaciones

Capítulo 6

Conclusiones

6.1. Conclusiones

Conclusiones

6.2. Sugerencias para investigaciones futuras

- Item 1.
- Item 2.
- Item 3.
- ...

Apéndice A

Resultados teóricos

Bibliografía

Branscum, A. J., Johnson, W. O. y Thurmond, M. C. (2007). Bayesian beta regression; application to househol data and genetic distance between foot-and-mouth disease viruses, *Australian & New Zealand Journal of Statistics* **49**(3): 287–301.

Congdon, P. (2003). *Applied-Bayesian-Modelling* , John Wiley & Sons.

Ferrari, S. y Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**: 799–815.