PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Modelo de regresión robusta con censura intervalar

Tesis para obtener el grado académico de Magíster en Estadística que presenta:

Luis Carlos Aliaga Flores

Asesor:

Cristian Luis Bayes Rodriguez

Lima, 2022

Informe de Similitud

Yo, Cristian Luis Bayes Rodríguez, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de la tesis titulada *Modelo de regresión robusta con censura intervalar* del autor Luis Carlos Aliaga Flores, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 20 %, lo que está dentro del límite establecido. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 03/03/2022.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 12 de diciembre de 2022

Apellidos y nombres de asesor: Cristian Luis Bayes Rodríguez	
DNI: 40372640	Firma: 
ORCID: 0000-0003-0474-7921	

Dedicatoria

A mi mamá, mi papapa y mi esposa Gaby por su constante motivación, paciencia y amor durante el desarrollo de esta tesis. Son mi motor para dar siempre la milla extra en todos los proyectos de vida que me embarco.

Agradecimentos

A los profesores del curso de la maestría de Estadística de la PUCP y en especial al profesor Cristian Bayes, por su permamente apoyo, asesoría y enseñanzas durante el desarrollo de la presente tesis.

Resumen

El presente trabajo de tesis propone el modelo de regresion log t de Student, el cual permite modelar variables respuesta que presentan censura intervalar y se muestra robusto frente a la presencia de observaciones atípicas. Luego, se desarrolla aquí un estudio de simulacion clásico, con el fin de analizar la sensibilidad frente a distintos niveles de valores atípicos. Finalmente, se desarrolla la aplicacion del modelo para la estimación de las demoras en órdenes de compras de los proveedores de las empresas en el Perú, concluyendo que el modelo propuesto en esta tesis tiene un mejor ajuste a los datos en comparación con el modelo Log Normal.

Palabras-clave: censura intervalar, valores atípicos, modelo de regresión, robustez, estimación clásica.

Índice general

Índice de figuras VIII

Índice de cuadros IX

1. Introducción 1

1.1. Consideraciones Preliminares 1

1.2. Objetivos 3

1.3. Organización del Trabajo 3

2. Distribuciones t y log t de Student 4

2.1. Distribución t de Student 4

2.1.1. Función de densidad de probabilidad 4

2.1.2. Propiedades 5

2.2. Distribución log t de Student 6

2.2.1. Función de densidad de probabilidad 6

2.2.2. Propiedades 7

3. Modelos de regresión robusta con censura intervalar 9

3.1. Modelo con censura intervalar para la distribución log t-Student. 9

3.2. Estimación del Modelo bajo Máxima Verosimilitud 9

3.3. Residuales 10

3.4. Selección de modelos 11

4. Estudio de Simulación 13

4.1. Consideraciones para la simulación 13

4.1.1. Creando la censura intervalar 14
4.1.2. Generación de datos con presencia de outliers 15
4.2. Resultados 17

5. Aplicación **19**
5.1. Base de datos 19
5.2. Variables e indicadores 20
5.3. Resultados 21

6. Conclusiones **24**
6.1. Conclusiones 24

Bibliografía **25**
-

Índice de figuras

2.1. Función de densidad de la distribución t para $\mu = 0, \sigma = 1$ y diferentes grados de libertad. 5

2.2. Función de densidad de la distribución log t para $\mu = 0, \sigma = 1$ y diferentes grados de libertad. 7

4.1. Diagrama de dispersión e histograma de los datos simulados 14

4.2. Gráfico de Y con censura intervalar 14

4.3. Distintos escenarios sin y con outliers 16

4.4. Estimación puntual e intervalo de confianza al 95 % para β_0 y β_1 bajo modelos log Normal y log t para los distintos escenarios *contaminados* 17

4.5. AIC bajo el modelo log normal (linea roja) y log t (linea azul) 18

5.1. Gráfico Q-Q Plot de la estimación por el modelo Log t vs el modelo Log Normal 22

Índice de cuadros

5.1. Coeficientes estimados y errores estándar para todos los parámetros de los modelos de regresión a la media log Normal y log t de Student 22

5.2. Criterios de comparación AIC, BIC y de Devianza global para los modelos de regresión log t de Student y log Normal para el ajuste de los datos con presencia de censura intervalar 23

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

En muchos casos un investigador está interesado en estudiar el comportamiento de una variable dependiente dado un conjunto de variables explicativas. Sin embargo, puede ocurrir que algunas o todas las observaciones de la variable respuesta presenten censura intervalar, esto es que solamente conocemos que su valor pertenece a cierto intervalo conocido. Un ejemplo de esto se puede ver en Sal y Rosas, Moscoso-Porras, Ormeño, Artica, Bayes y Miranda (2019), donde se abordó la problemática de las brechas de ingresos por género que existen entre profesionales de la salud. Dentro de este estudio se usaron los datos de la *Encuesta nacional de satisfacción de usuarios en salud* ENSUSALUD (2015) donde la variable dependiente fue el reporte del monto de ingresos en USD, la cual no fue preguntada directamente como una variable continua en dicha encuesta sino que fue preguntada en intervalos, generandose de esta manera la censura intervalar. Por ejemplo estos intervalos fueron: [0, 314], [315, 629], entre otros.

Otro estudio donde se puede apreciar la censura de los datos es en los análisis de supervivencia (ver Lindsey y Ryan (1998)), donde la ocurrencia de un evento es observada durante un periodo de tiempo T. Si sabemos que el evento de interés ocurrió durante un intervalo de tiempo [L,R] donde $L \leqslant T \leqslant R$, se obtienen datos que presentan una censura intervalar. Por ejemplo esta casuistica puede ocurrir en una prueba médica, donde los pacientes son evaluados solo sobre visitas programadas. Si el evento no ha ocurrido durante la primera visita (tiempo L) pero sí ha ocurrido en la siguiente visita (tiempo R), por lo tanto sabemos que el evento de interés ha ocurrido entre [L,R], obteniendo así una censura intervalar dentro de este tipo de prueba médica (T).

Otro ejemplo se abarca en Bleda y Garces (2002), en donde se busca medir los valores de las concentraciones de mercurio medidas en orina en la población de Mataró. Esta medición, al ser obtenida mediante aparatos de medición con ciertos límites de detección inferiores y superiores, produce una censura en la variable respuesta.

Asimismo, un supuesto usual dentro del modelamiento de los datos es la normalidad de estos. Sin embargo, se sabe que la estimación de los parámetros del modelo puede verse

fuertemente afectadas ante la presencia de valores atípicos. Por este motivo, distribuciones de probabilidad con colas más pesadas que la distribución normal han sido propuestas en la literatura como una alternativa a la usual suposición de normalidad de los errores en modelos de regresión, las cuales tienen la ventaja de incorporar observaciones atípicas bajo el supuesto de normalidad.

Esta casuística descrita ya se ha estudiado antes por algunos autores, siendo los más representativos West (1984) y Lange et al. (1989). El primer autor estudia la distribución de los errores de modelos con colas pesadas para el manejo de valores atípicos bajo inferencia bayesiana. Dentro de este trabajo, el autor propone el uso del modelo t Student como distribución a priori dentro del enfoque bayesiano. Por otro lado, en Lange et al. (1989) se estudia algo similar pero bajo un enfoque de inferencia clásica. El artículo nos ilustra la habilidad de la distribución t de Student en manejar valores atípicos realizando la estimación del modelo bajo máxima verosimilitud y demostrando la robustez del modelo mediante los grados de libertad ν. Por lo tanto, como se puede apreciar, tanto West como Lange proponen el uso de una distribución t Student en presencia de valores atípicos ya que sus colas pesadas reducen la infuencia de estos dentro de la estimación de los parámetros, lo cual hace el análisis estadístico más confiable y robusto.

Por otro lado, dentro del mundo empresarial, la puntualidad en las entregas de los proveedores es fundamental dentro del ecosistema productivo de las compañias. Un retraso en la cadena de suministro no conlleva únicamente en la pérdida del valor total o parcial de la mercancía, sino también en el desprestigio de la marca y el valor de la empresa. Dado esta problemática, la presente tesis propone estudiar cuales son los factores que afectan los retrasos en las órdenes de compras de todas aquellas empresas formales con sectores relacionadas a la minería, manufactura, construcción, comercio y servicios, localizadas en el Perú y que en el año 2014 tuvieron ventas iguales o mayores a 20 Unidades Impositivas Tributarias. Para obtener los datos de estas empresas nos apoyaremos en la Encuesta Nacional de Empresas realizada por el INEI en el 2015 (ver INEI (2015)), donde la variable del tiempo de retraso no pudo ser medida exactamente sino que solamente fue recogida entre ciertos límites, ocurriendo de esta manera la censura intervalar. Finalmente, dada la naturaleza de la encuesta reflejando distintas realidades de las empresas a nivel nacional, se espera que existan varias observaciones atípicas en el retraso de las órdenes de compra, por lo que en la presente tesis se considerará una distribución log t de Student como modelo para esta variable.

Para estimar los parámetros del modelo se considera el método de máxima verosimilitud que será implementado usando métodos convencionales de optimización, tales como el de Newton-Raphson, Scoring de Fisher o el algoritmo EM. Otra opción es usar el paquete gamlss.cens de la libreria gamlss, el cual se menciona en Stasinopoulos et al. (2017) y tiene funciones que facilitan la implementación de este tipo de modelos.

1.2. Objetivos

El objetivo general de la tesis es estudiar las propiedades, estimar y aplicar a conjuntos de datos reales un modelo de regresión robusta con censura intervalar desde el punto de vista de la inferencia clásica. De manera específica:

- Revisar la literatura acerca de los diferentes propuestas de modelos de regresión para censura intervalar.
- Estudiar propiedades e implementar la estimación del modelo de regresión con censura intervalar desde la perspectiva clásica.
- Implementar métodos de inferencia clásica para este modelo como parte de la familia de los modelos GAMLSS (Generalized Additive Models for Location, Scale and Shape).
- Realizar estudios de simulación acerca de la regresión con censura intervalar considerando computación intensiva sobre diferentes escenarios.
- Aplicar el modelo a conjunto de datos reales.

1.3. Organización del Trabajo

En el Capítulo 2 se presenta conceptos y principales propiedades sobre las distribución t de Student y la distribución log t de Student.

En el Capítulo 3 se propone al modelo de regresión robusta con censura intervalar para el análisis de regresión de la media de una variable con censura intervalar. Asimismo, se detalla el método para la estimación de los parámetros del modelo desde la perspectiva frecuentista de la estadística.

Luego, en el capítulo 4, se presenta un estudio de simulación para evaluar qué tanto impactan los valores atípicos en una distribución log t usando una variable con censura intervalar como variable respuesta. Esto se logrará mediante un análisis de sensibilidad comparando los resultados de las estimaciones entre el modelo log t de Student y un modelo log Normal.

A continuación, en el Capítulo 5 se procederá a desarrollar la aplicación considerando el modelo propuesto y tomando como variable respuesta el tiempo de demora por parte de los proveedores en la entrega de suministros hacia las empresas en el Perú. Esta variable presenta censura intervalar.

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

Capítulo 2

Distribuciones t y log t de Student

Este capítulo se dedica al estudio de la distribución t de Student en su especificación básica, para luego analizar la distribución log t de Student. Cabe señalar que para cada distribución, se define su función de densidad de probabilidad y sus principales propiedades (esperanza y varianza).

2.1. Distribución t de Student

La distribución t de Student o distribución t define a una familia de distribuciones de probabilidad continua. Tiene la ventaja de tener un amplio número de aplicaciones dentro de la probabilidad, estadística y matemática, así como diversos campos que subyacen dentro de estos. Fue desarrollado por William S. Gosset en 1908 en su trabajo llamado 'The probable error of a mean' en Student (1908), publicado por él mismo bajo el pseudónimo de *Student*. Es una distribución simétrica usada comunmente para modelar variables con colas más pesadas que la distribución normal, lo cual hace que se pueda utilizar para modelar datos con valores atípicos. Con los grados de libertad se pueden controlar la dispersión y la kurtosis de la distribución (ver Ahsanullah et al. (2014)).

Una de las primeras personas que propuso usar la distribución t para modelar relaciones lineales entre variables fue Zellner (1976) ya que tradicionalmente, para desarrollar modelos de regresion lineal, se asumían que los errores seguian una distribución Normal. Sin embargo, en situaciones de la vida real no siempre se cumple el supuesto de normalidad, observandose usualmente colas pesadas. Es ahí donde la distribución t se presenta como una propuesta atractiva a poder ayudarnos mejor a estimar los parámetros del modelo.

2.1.1. Función de densidad de probabilidad

Una variable aleatória Y sigue una distribución t de Student, denotada por $Y \sim t(\mu, \sigma, \nu)$, si su función de densidad es dada por:

$$f(y;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}\sigma} \left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}, \; y \in \mathbb{R} \tag{2.1}$$

donde $\mu \in \mathbb{R}$ es un parámetro de localización, $\sigma > 0$ un parámetro de escala y $\nu > 0$ son los grados de libertad, el cual controla la curtosis de la distribución.

Como se puede observar en la figura 2.1, esta distribución tiene una forma similar a la distribución normal; ambas tienen forma de acampanada y son simétricas. La diferencia radica en que la distribución t de Student tiene las colas más pesadas por lo cual consigue modelar mejor la presencia de valores atípicos que una distribución Normal cuando ν <30. Sin embargo, cuando $\nu \to \infty$, la distribución t converge a una distribución Normal.

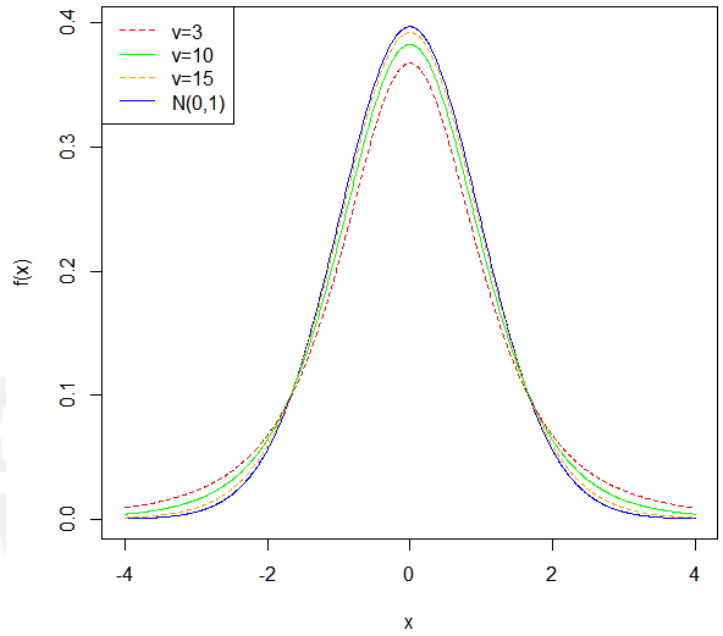


Figura 2.1: Función de densidad de la distribución t para $\mu = 0, \sigma = 1$ y diferentes grados de libertad.

2.1.2. Propiedades

Según Ahsanullah et al. (2014), la media y la variancia de una distribución son expresadas por

$$E(Y) = \mu \quad \text{cuando } \nu > 1; \quad \text{y } Var(Y) = \frac{\nu\sigma^2}{\nu - 2} \quad \text{cuando } \nu > 2. \tag{2.2}$$

Asimismo, la función de distribución acumulada (cdf) t puede ser escrita como:

$$F(y; \mu, \sigma, \nu) = \int_{-\infty}^{y} f(y; \mu, \sigma, \nu)du = 1 - \frac{1}{2} I_t \left(\frac{\nu}{2}, \frac{1}{2} \right), \text{ cuando } y > 0 \tag{2.3}$$

donde $t = \left(1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right)^{-1}$ e $I_t(.,.)$ es la función Beta incompleta regularizada, la cual se define como:

$$I_t(a,b) = \frac{\mathrm{B}(t;a,b)}{\mathrm{B}(a,b)} \tag{2.4}$$

donde $\mathrm{B}(t;a,b) = \int_0^t u^{a-1}(1-u)^{b-1}du$ es la función de Beta incompleta. Cuando $t = 1$ la función de Beta incompleta coincide con la función Beta.

Ademas, de acuerdo a Lange et al. (1989), si $Z \sim N(\mu,\sigma)$ y $W \sim Gamma(\frac{\nu}{2},\frac{\nu}{2})$ son variables aleatorias independientes, entonces:

$$Y = \mu + \sigma\frac{Z}{\sqrt{\frac{W}{\nu}}} \sim t(\mu,\sigma,\nu) \tag{2.5}$$

2.2. Distribución log t de Student

Asi como la distribución log-normal se define a partir de una distribución normal, la distribución *log t − Student* (o simplemente distribución *log t*), la cual fue introducida en Hogg y Klugman (1983), es un modelo derivado a partir de la distribución t-Student: si $log(y) \sim t(\mu,\sigma,\nu)$, entonces $y \sim Lt(\mu,\sigma,\nu)$.

Una aplicación donde se usa esta distribución log t de Student es en Vallejos y Steel (2015) donde buscan modelos robustos a la presencia de valores atípicos para datos positivos y analizan los modelos de mixturas de la distribución log-normal. Por otro lado, en Barroso et al. (2019) se analiza la relevancia en usar la distribución log t para modelar la distribución de ingresos, aplicandolo con datos de 25 paises europeos durante periodos antes, durante y despues de la gran recesión del 2008. El autor comprueba la efectividad de usar el modelo log t destacando su naturaleza bimodal y su capacidad de modelar valores cercanos al 0 con un número reducido de parámetros.

2.2.1. Función de densidad de probabilidad

La función de densidad de probabilidad de una variable aleatória $Y \sim Lt(\mu,\sigma,\nu)$ que sigue una distribución log t-Student, es dada de la siguiente manera:

$$g(y;\mu,\sigma,\nu) = \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\frac{1}{\sqrt{\pi\sigma^2\nu}}\frac{1}{y}\left[1+\frac{(\log(y)-\mu)^2}{\sigma^2\nu}\right]^{-\frac{\nu+1}{2}}, \nu > 0 \tag{2.6}$$

donde e^{μ} es la mediana, la media no existe y los parámetros σ y ν controlan la dispersión y la asimetría de la distribución: mientrás más grande es el valor de σ, más asimétrica será la distribución. Por el contrario, mientras más grande sea el parámetro ν más simétrica será la distribución (ver Barroso et al. (2019)).

Asimismo, como se puede observar en la la figura 2.2, la distribución de la log t se asemeja a la forma de una distribución log Normal, donde el parámetro ν marca el grosor de las colas de la distribución: cuando el valor de ν se hace más pequeño, las colas se hacen más pesadas.

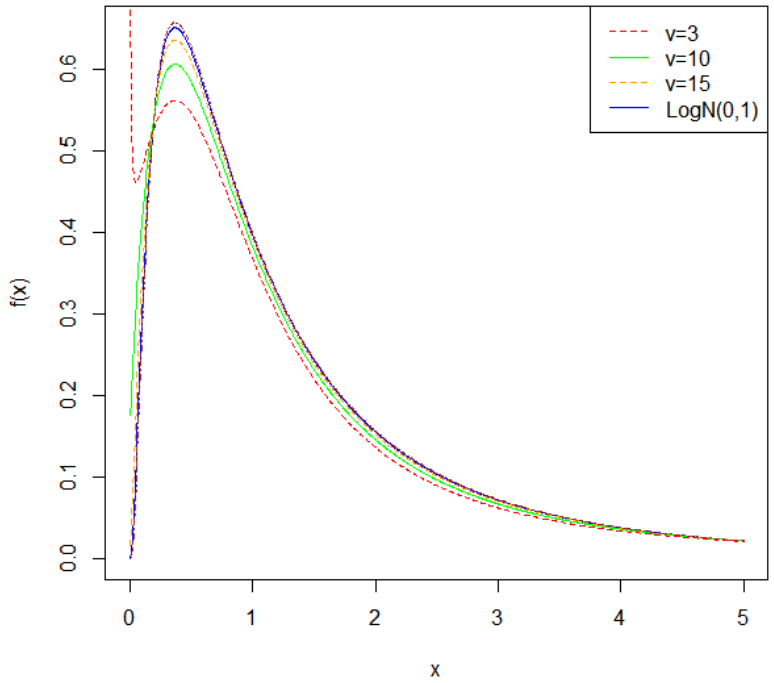


Figura 2.2: Función de densidad de la distribución log t para $\mu = 0, \sigma = 1$ y diferentes grados de libertad.

2.2.2. Propiedades

La función de distribución acumulada de una distribución *log t* es dada por:

$$G(y; \mu, \sigma, \nu) = \int_{-\infty}^{y} g(u; \mu, \sigma, \nu) du \tag{2.7}$$

donde dicha función no puede ser expresada de una forma cerrada, aunque puede ser obtenida aplicando la función de distribución acumulada en (2.3), con una variable aleatoria con distribución t de Student y ν grados de libertad.

$$G(y; \mu, \sigma, \nu) = F\left(log\left(y\right); \mu, \sigma, \nu\right). \tag{2.8}$$

Finalmente, según Barroso et al. (2019), la distribución log t-Student cuenta con las siguientes propiedades importantes:

1. A pesar que esta distribución puede acumular una cantidad importante de probabilidad alrededor del cero, su moda existe.
2. Cuando $\nu \to \infty$ la distribución log t-Student converge en una distribución log Normal.
3. Cuando $\nu = 1$ la distribución log t-Student converge en una distribución log Cauchy.

Asimismo, esta distribución no tiene una función generadora de momentos pero cuando v es muy grande, la log t-Student tiende a una distribución lognormal la cual tiene todos sus momentos.

Capítulo 3

Modelos de regresión robusta con censura intervalar

En este capítulo se presentará un modelo de regresión para una variable respuesta que presente censura intervalar. La estimación de los parámetros se realizarán por el método de máxima verosimilitud. Asimismo, en este capítulo revisaremos cómo se usan los residuales quantílicos para la verificación de supuestos en una regresión con censura intervalar.

3.1. Modelo con censura intervalar para la distribución log t-Student.

El modelo probabilístico de la distribución log t con su función de densidad definida en (2.6) permite una estructura adecuada para un análisis de regresión robusta de la mediana de una variable positiva con censura.

Consideremos Y_i como variable respuesta de la i-esima observación tal que su especificación es la siguiente:

$$\begin{aligned} Y_i &\sim Lt(\mu_i, \sigma, \nu) \\ \mu_i &= \boldsymbol{x_i}^{\top}\boldsymbol{\beta} \end{aligned} \tag{3.1}$$

donde de acuerdo a lo presentado en la sección (2.2.1), e^{μ_i} es la mediana de y_i, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\top}$ con $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$ y los otros parámetros de la distribución log t de Student son: σ el parámetro de escala de la distribución y ν el parámetro de grados de libertad, los cuales a su vez controlan la dispersión y la asimetría de la distribución. Finalmente, $\boldsymbol{x}_i = [1, x_i, \ldots, x_{ik}]^{\top}$ son vectores de las covariables por observación.

Asimismo, la estimación de los parámetros $\boldsymbol{\theta} = \left[\boldsymbol{\beta}^{\top}, \sigma, \nu\right]^{\top}$ se realizará usando el método de máxima verosimilitud.

3.2. Estimación del Modelo bajo Máxima Verosimilitud

La función de verosimilitud que depende del vector de parámetros desconocidos $\boldsymbol{\theta}$ en caso de observar la variable respuesta Y_i es dada por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(y_i; \mu_i, \sigma, \nu) \tag{3.2}$$

donde y_i es el valor observado de la variable respuesta. Sin embargo, en el caso de censura intervalar, cuando Y_i no es observado y sólo se conoce que $Y_i \in [y_{1i}, y_{2i}]$ donde $y_{1i} < y_{2i}$, entonces de acuerdo a Stasinopoulos et al. (2017), la función de verosimilitud es dada por:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[G\left(y_{2i}; \mu_i, \sigma, \nu\right) - G\left(y_{1i}; \mu_i, \sigma, \nu\right) \right] \\ &= \prod_{i=1}^{n} \left[F\left(log(y_{2i}); \mu_i, \sigma, \nu\right) - F\left(log(y_{1i}); \mu_i, \sigma, \nu\right) \right] \end{aligned} \tag{3.3}$$

donde G es la función de distribución acumulada de una log t de Student y F es la función de distribución acumulada de una t-Student dada en (2.8).

El estimador de máxima verosimilitud (EMV) de $\boldsymbol{\theta}$ es el valor de $\hat{\boldsymbol{\theta}}$ que maximiza la función de verosimilitud. Dado que la ecuación en (3.3) no brinda una forma analítica para el EMV, se implementará usando métodos convencionales de optimización numérica y el software de R (R Core Team (2013)), o usando el paquete **gamlss.cens** (Stasinopoulos et al. (2018)) dentro de la misma herramienta.

Los errores estándar serán calculados a partir de la raiz cuadrada de la diagonal de la inversa de la matriz de información de Fisher observada, siendo esta evaluada en el estimador de máxima verosimilitud (EMV). La estimación de esta matriz será hecha por métodos numéricos.

3.3. Residuales

Dentro de la literatura estadística, un análisis usual y recomendado para analizar la bondad de ajuste del modelo en estudio y poder diagnosticar sus fortalezas y debilidades, es examinar la distribución que siguen los residuales cuantílicos estandarizados del modelo escogido. La principal ventaja de estudiar los residuales es que cualquiera que sea la distribución de la variable respuesta, cuando el modelo escogido es el correcto deberiamos observar que siguen una cierta distribución. Este análisis es usualmente representado en los gráficos cuantil-cuantil o *QQ Plot*; en caso se vea que los residuos estandarizados siguan una linea recta, podemos decir que el modelo presenta un ajuste adecuado.

Según Stasinopoulos et al. (2017), en un modelo de regresión lineal simple los residuales simples son definidos como la diferencia entre el valor observado y el valor estimado del modelo. Por lo que, considerando que la distribución $g(y \mid \boldsymbol{\theta})$ es estimada para las observaciones y_i para $i = 1, ..n$, estos residuales $\hat{\varepsilon}_i$ son definidos como:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \tag{3.4}$$

El problema con los residuales simples es que son dificiles para generalizar a otras distri-

buciones distintas a las normales; por lo que se usa los residuales estandarizados para evitar estos problemas, los cuales para un modelo de regresión lineal simple son definidos como:

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{(1-h_{ii})}} \tag{3.5}$$

donde h_{ii} son los valores de la diagonal de la matriz sombrero H (o matriz de proyección) $H = X\left(X^TX\right)^{-1}X^T$. Por otro lado, los residuales cuantílicos $\hat{r}_i$ según Dunn y Smyth (1997) son definidos como:

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i) = \Phi^{-1}(G(y_i \mid \hat{\boldsymbol{\theta}})) = \Phi^{-1}(G\left(y_i; \mu, \sigma, \nu\right)) \tag{3.6}$$

donde $\Phi^{-1}()$ es la función de distribución acumulada (cdf) inversa de una distribución normal estándar. Si conocieramos la verdadera distribución $G(y_i \mid \hat{\boldsymbol{\theta}})$, al ser evaluada con la variable respuesta observada debería tener una distribución uniforme (0,1). Luego, los autores consideran evaluar esto en $\Phi^{-1}()$ para que los residuales sigan una distribución normal estandar (0,1), si el modelo estimado es adecuado.

Para el caso cuando y_i sea una observación con censura intervalar entre $[y_{1i}, y_{2i}]$, entonces $\hat{u}_i$ en (3.6) es definido como un valor aleatorio con distribución uniforme en el intervalo $[u_{1i}, u_{2i}] = [G(y_{1i} \mid \hat{\boldsymbol{\theta}}), G(y_{2i} \mid \hat{\boldsymbol{\theta}})]$.

Por lo tanto, para obtener los residuales cuantílicos estandarizados reales de la observación i-ésima con censura intervalar $[y_{1i}, y_{2i}]$ se selecciona aleatoriamente $\hat{u}_i$ del intervalo $[u_{1i}, u_{2i}]$ para luego transformarla en el residual cuantílico estandarizado $\hat{r}_i = \Phi^{-1}(\hat{u}_i)$, donde $\hat{r}_i$ debería seguir aproximadamente una distribución normal estandar.

Asimismo, según Stasinopoulos et al. (2017) para obtener los valores de los residuales estandarizados estimados se puede usar la función **resid()** que se encuentra dentro del paquete **gamlss** (ver Rigby y Stasinopoulos (2005)) en el software R, así como otras funciones para representar estos residuos de manera gráfica, tales como la función **plot()** o **wp()** dentro del mismo paquete.

3.4. Selección de modelos

Para la comparación de los modelos usaremos los criterios de información de Akaike (AIC, por sus siglas en inglés) y el criterio de información bayesiano (BIC, por sus siglas en inglés), donde el modelo a elegir será quien tenga menor valor del criterio de información. Tanto el AIC como el BIC premiam el ajuste del modelo a los datos e introducen un término de penalización por el número de parámetros del modelo.

El AIC es definido como:

$$\text{AIC} = 2p - 2\ln\left(L(\hat{\boldsymbol{\theta}})\right) \tag{3.7}$$

donde L es la función verosimilitud del modelo, p es el número de parámetros y $\hat{\boldsymbol{\theta}}$ es el estimador de máxima verosimilitud (EMV). Finalmente, el BIC es definido como:

$$\text{BIC} = \ln(n)p - 2\ln\left(L(\hat{\boldsymbol{\theta}})\right) \tag{3.8}$$

donde n es el número de observaciones o tamaño de la muestra.

Capítulo 4

Estudio de Simulación

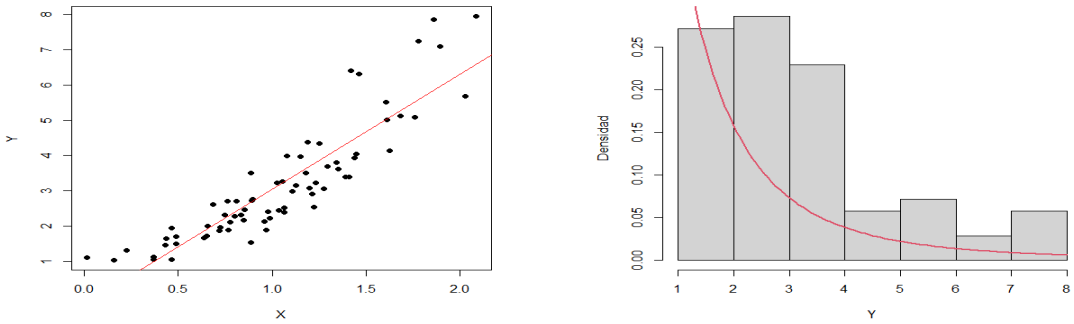
Este capítulo presenta un estudio de simulación con el fin de evaluar si el modelo log t de Student es más robusto que el modelo Log-Normal en presencia de valores atípicos, para lo cual crearemos valores atípicos artificiales perturbando los datos. A continuación, ajustaremos los datos de los distintos escenarios *contaminados* al modelo log Normal y al modelo log t. Finalmente, se evaluará un estudio de sensibilidad en la estimación de los coeficientes de la regresión y en el criterio de información de Akaike (AIC) para la comparación de los modelos en los diferentes escenarios.

4.1. Consideraciones para la simulación

Para el estudio de sensibilidad de los modelos Log t y Log Normal en presencia de valores atípicos, se tomó en consideración un dataset con Y_i como variable respuesta de la i-esima observación tal que siga una distribución Log Normal, considerando:

$$\begin{aligned} X_i &\sim N(1,0{,}5) \quad \text{donde } i = 1, ..., n \\ \mu_i &= \beta_0 + \beta_1 X_i \\ Y_i &\sim LN(\mu_i, 0{,}2) \end{aligned} \tag{4.1}$$

donde $n = 70$. Cabe resaltar que esta data no tiene presencia de valores atípicos aún. Para transformar estos datos en unos datos con valores atípicos lo que haremos es definir un proceso de *contaminación* de esta información.



(a) Relación entre X e Y (b) Histograma de Y

Figura 4.1: Diagrama de dispersión e histograma de los datos simulados

Como podemos observar en la figura 4.1a, existe una relación lineal representada con la linea roja entre la covariable X y la variable dependiente Y simulada. Finalmente, la forma logarítmica de esta variable dependiente Y es representada en la figura 4.1b, donde vemos su histograma.

4.1.1. Creando la censura intervalar

Una vez simuladas la covariable y la variable respuesta, procedemos a realizar la censura intervalar de la variable Y. Para esto vamos a crear 7 intervalos del mismo tamaño que van desde el [1,2) hasta el [8,9). En la la figura 4.2 se observa la censura que se ha creado para los valores de la variable dependiente, donde los puntos representan a los valores simulados y la linea vertical el intervalo:

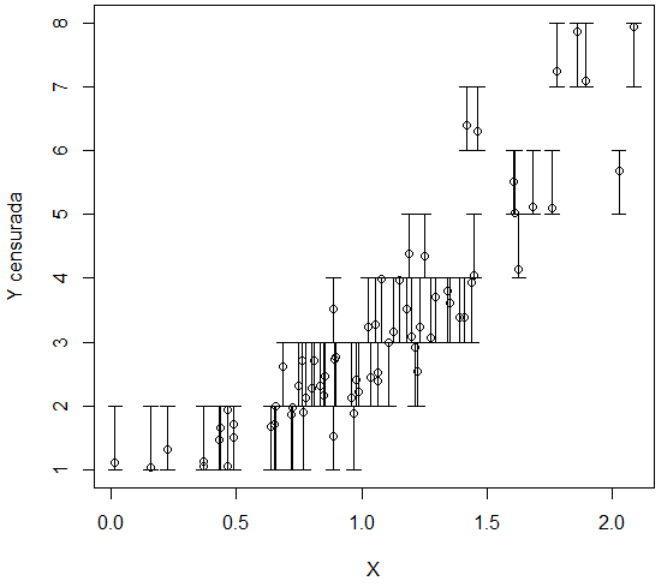


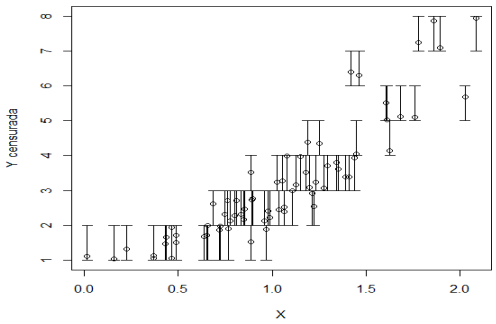
Figura 4.2: Gráfico de Y con censura intervalar

Como podemos apreciar en la figura 4.2, la relación lineal entre la variable Y simulada ahora con la censura intervalar se mantiene, donde la mayoría de intervalos generados se concentran entre los valores [1,4] siguiendo el histograma de la variable Y mostrada en la figura 4.1b.

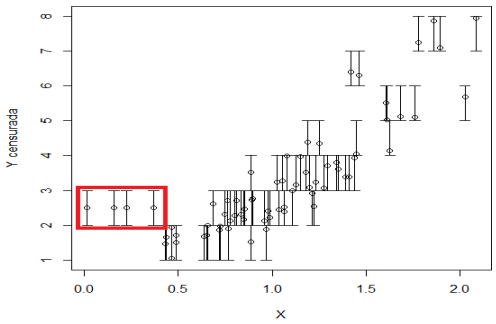
4.1.2. Generación de datos con presencia de outliers

Con el objetivo de poder comparar la robustez del modelo Log t versus el modelo Log Normal en presencia de valores outliers, procederemos a realizar un análisis de sensibilidad en las estimaciones de los coeficientes de regresión y el efecto en los criterios de comparación de ambos modelos mediante el criterio de Akaike (AIC).

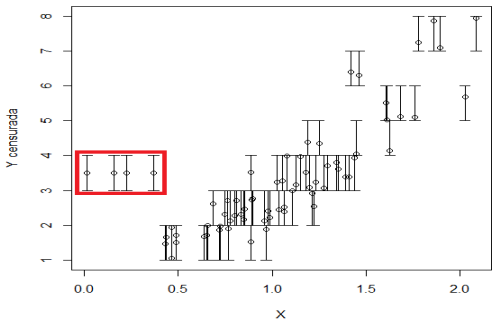
Para lograr ambos análisis, una vez obtenida la simulación de los datos ya con la censura intervalar procedemos a crear 7 distintos escenarios *contaminados* con valores atípicos en distintos niveles y reestimar ambos modelos. Cada uno de los escenarios contaminados se obtuvo reemplazando 5 observaciones del dataset con 5 observaciones outliers (señalados con las cajitas rojas), dándose lugar de esta manera a la *contaminación* de los datos. Cada escenario de análisis tiene un nivel de *contaminación* superior al anterior.



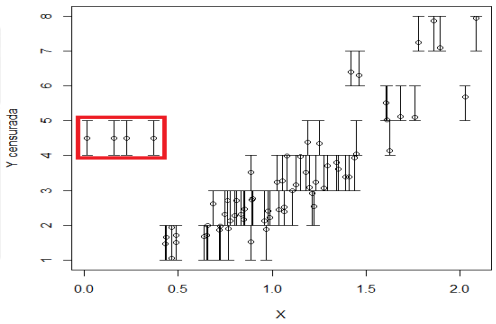
(a) Escenario 1: Sin outliers



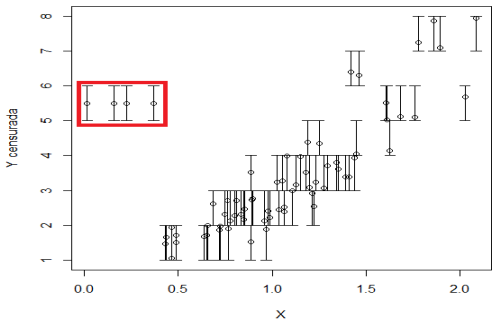
(b) Escenario 2: Con outliers



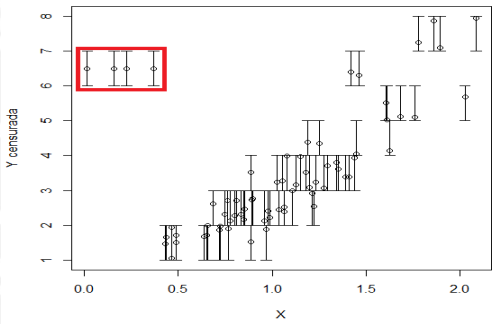
(c) Escenario 3: Con outliers



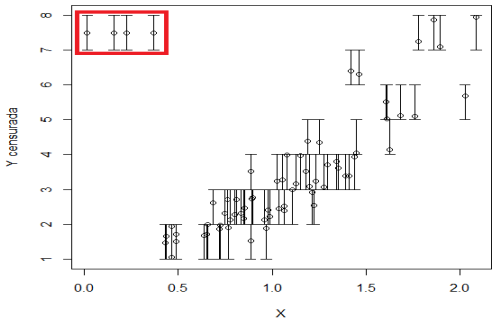
(d) Escenario 4: Con outliers



(e) Escenario 5: Con outliers



(f) Escenario 6: Con outliers



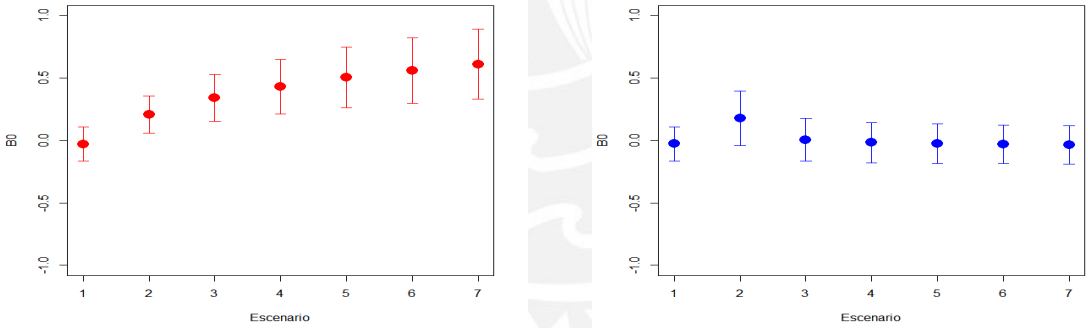
(g) Escenario 7: Con outliers

Figura 4.3: Distintos escenarios sin y con outliers

Como podemos observar en la figura 4.3, en el escenario 1 se consideran los datos sin contaminar que fueron los mostrados luego de generar la censura intervalar en la figura 4.2. Luego, en los escenarios 2 al 7 ya se genera la *contaminación* de las observaciones simuladas censuradas, donde se consideran las 5 observaciones con menor valor de la variable respuesta censurada (señalados en las cajitas rojas) y se reemplazan por 5 observaciones *contaminadas*. A medida que avanzamos en los escenarios, el nivel de *contaminación* de estas 5 observaciones se hace cada vez más fuerte.

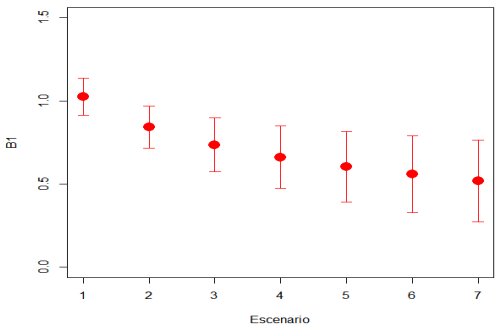
4.2. Resultados

Para cada uno de los conjuntos de datos de los 7 escenarios de análisis descritos previamente, se estimó el modelo log Normal y log t de Student, estimandose por máxima verósimilitud los parámetros de ambos modelos. De los resultados del análisis de sensibilidad entre los modelos Log Normal y Log t podemos notar claramente que conforme vamos introduciendo distintos niveles de valores atípicos en cada escenario, los intervalos de confianza de los β_0 y β_1 bajo el modelo Log Normal van haciendose cada vez más grandes. Por otro lado, la estimación de los β usando el modelo Log t se mantiene estable conforme vamos introduciendo los distintos niveles de escenarios *contaminados* con valores atípicos. Lo mismo ocurre para los intervalos de confianza de los coeficientes β_0 y β_1 los cuales permanecen estables.

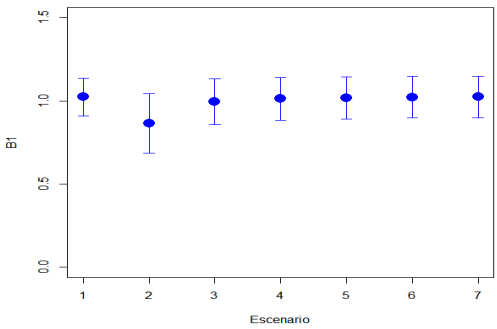


(a) β_0 - Modelo log Normal

(b) β_0 - Modelo log t de Student



(c) β_1 - Modelo log Normal



(d) β_1 - Modelo log t de Student

Figura 4.4: Estimación puntual e intervalo de confianza al 95 % para β_0 y β_1 bajo modelos log Normal y log t para los distintos escenarios *contaminados*

Como se puede observar en la figura 4.4, existe un sesgo en la estimación de β_0 y β_1 por el modelo Log Normal a medida que se van introduciendo los valores atípicos en los datos, inclusive ocasionando que el verdadero valor de β_0 y β_1 no se encuentre dentro del intervalo de confianza a media que la *contaminación* se hace cada vez más fuerte. Esto no ocurre en el ajuste a los datos con censura intervalar bajo el modelo log t, lo cual demuestra su robustez frente a la presencia de observaciones outliers.

En la figura 4.5 se presenta que cuando el nivel de *contaminación* es nulo (sin valores atípicos) o bajo, el AIC distingue correctamente que el modelo log Normal es mejor que el modelo log t. Sin embargo, a medida que la *contaminación* de los datos aumenta, vemos que el AIC indica que el modelo el log t es un mejor modelo que el log Normal, lo cual era de esperarse al ser más robusto que el modelo log Normal.

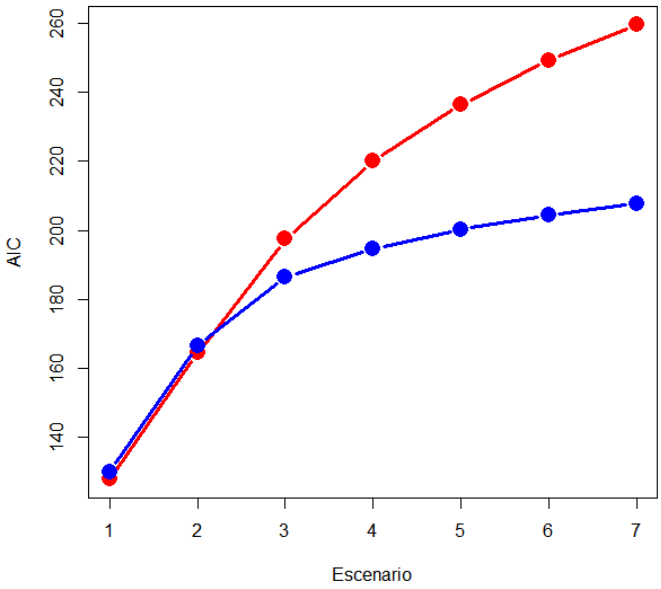


Figura 4.5: AIC bajo el modelo log normal (linea roja) y log t (linea azul)

Capítulo 5

Aplicación

Dentro del ciclo de vida de los productos o servicios que ofrecen las empresas, es decir, desde que se conciben hasta que se ponen a disposición de los clientes finales en el momento adecuado, la gestión y buena elección de proveedores es una pieza fundamental en lograr una cadena de suministro eficaz y óptima dado que con ellos se inician el proceso de venta de cualquier bien o servicio. Sin una correcta coordinación con los proveedores, no solo se puede perder el valor total o parcial del producto o servicio ofrecido debido a retrasos, sino también el desprestigio de la marca y la valorización de la compañía.

Dada la problemática descrita, para esta sección se ha considerado la información de la Encuesta Nacional de Empresas hecha por el Instituto Nacional de Estadística e Informática (INEI) en el año 2015 la cual brinda información de las empresas sobre características de organización, gestión de los productos e insumos, percepción sobre las regulaciones gubernamentales y de indicadores de productividad. Dicha encuesta se ejecutó a nivel nacional y estudió las actividades relacionadas a la minería, manufactura, construcción, comercio y servicios.

5.1. Base de datos

La población objetivo está conformada por todas aquellas empresas formales con sectores relacionadas a la minería, manufactura, construcción, comercio y servicios, localizadas en el Perú y que en el año 2014 tuvieron ventas iguales o mayores a 20 Unidades Impositivas Tributarias (UIT). Esta definición contiene a las grandes, medianas, pequeñas y parte de las micro empresas dado que el corte inferior es de 20 UIT. Finalmente, no fue del alcance de la encuesta las empresas de actividades agrícolas y pecuarias, administración pública, actividades de organizaciones y otras empresas peruanas fuera de Perú. Una vez definida esta población, se procedió con la elaboración de la muestra, por lo que se consideró un muestreo estratificado, unietápico e independiente a nivel de división de la Clasificación Internacional Industrial (CIIU). El tamaño de la muestra fue de 19,204 empresas.

La información de la Encuesta Nacional de Empresas 2015 está compuesta por más de 100 preguntas (ver ENSUSALUD (2015)), las cuales representan el set completo de variables del dataset. Estas buscan conocer características de las empresas dentro de su alcance en los

siguientes temas:

- Identificación y ubicación de la empresa.
- Características de la empresa.
- Recursos humanos.
- Prácticas de gestión.
- Tecnologías de información y comunicaciones.
- Insumos complementarios.
- Financiamiento.
- Percepción sobre regulaciones o clima de negocios.
- Principales productos, principales insumos.
- Ventas y gastos anuales.

Una vez extraida la información de la encuesta, se procedió a un procedimiento de limpieza de los datos, por lo que nos quedamos con aquellas empresas que indican haber presentado demoras en el tiempo de entrega durante el 2014 en más del un 5 % del total de ordenes de compras a proveedores nacionales.

5.2. Variables e indicadores

La variable respuesta del modelo será el tiempo de retraso promedio de las órdenes de compra o pedidos nacionales respecto al tiempo de entrega pactado, la cual se obtuvo a partir de las respuestas a la pregunta 6 de la encuesta bajo estudio. Esta pregunta fue: *¿Cuál fue el tiempo de retraso promedio de las órdenes de compra o pedidos nacionales respecto al tiempo de entrega pactado?*. Donde las posibles respuestas a dicha pregunta fueron:

1. Menos de un día
2. De un día a 2 días
3. De 2 días a 7 días
4. De 7 días a 15 días
5. De 15 días a más

Por otro lado, luego de un proceso de selección las covariables que se consideraron dentro del modelamiento son:

- **x1:** Principal modalidad que utilizó para realizar compras en la empresa (compra directa al productor o compra a un distribuidor mayorista).
- **x2:** Porcentaje de las órdenes de compras o pedidos a proveedores nacionales que presentaron demoras en el tiempo de entrega
- **x3:** La empresa realizó o no importaciones a través de despacho anticipado.
- **x4:** La empresa realizó o no sus compras por internet durante el 2014.
- **x5:** Porcentaje de trabajadores que utilizan equipos informáticos por lo menos una vez por semana dentro de la empresa.

Finalmente, se presentan otras covariables que se probaron dentro del ajuste de los datos al modelo pero que no fueron seleccionadas:

- **x6:** Porcentaje de las órdenes de compras o pedidos a proveedores nacionales presentaron problemas de cantidad, especificaciones o daños durante el 2014.
- **x7:** Porcentaje que representan las compras por internet del total de las órdenas de compras realizadas durante el 2014.
- **x8:** Porcentaje que representan las ventas por internet del total de las órdenas de ventas realizadas durante el 2014.
- **x9:** Número de trabajadores capacitados en idiomas, gestión empresarial, salud ocupacional, tecnologías de la información, habilidades socio-emocionales o técnicas de marketing y estratégia de ventas.
- **x10:** Porcentaje aproximado de las importaciones que fueron realizadas a través del despacho anticipado.
- **x11:** Utilización del internet (número de trabajadores que usan el servicio de internet entre el total de trabajadores)
- **x12:** Tipo de registro utiliza para las órdenes de compra o pedidos que realizan (apuntes en cuaderno, programas informáticos, software de gestión, otros)

5.3. Resultados

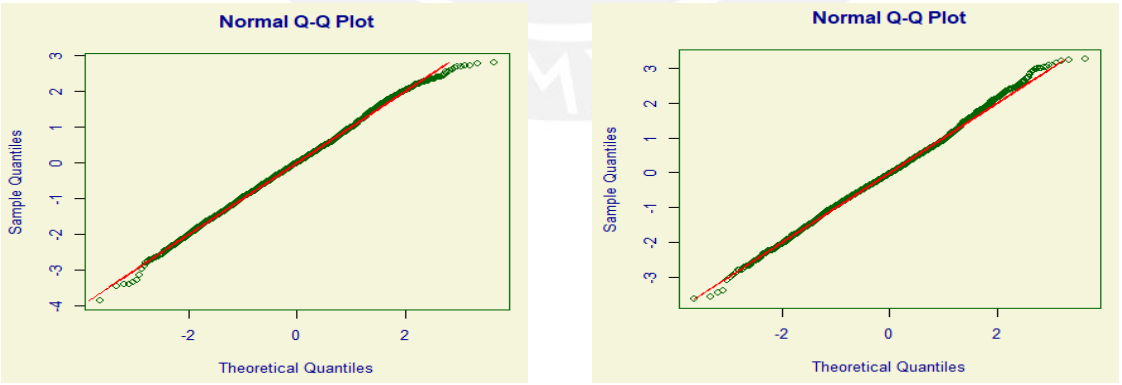
Con el objetivo de poder medir el desempeño del modelo log t de Student en la regresión robusta para el ajuste del tiempo de retraso promedio de las órdenes de compra de las empresas en el Perú en 2014, la cual presenta una censura intervalar, también se consideró conveniente compararlo contra el ajuste de los datos bajo el modelo log Normal. Para realizar el ajuste a los datos de las regresiones de ambos modelos se tomó la función de enlace identidad para estimar μ y una función de enlace logística para estimar σ. Los resultados de las regresiones se presentan en el cuadro 5.1.

		Log Normal			Log t Student		
Parámetro	**Covariable**	**Estimado**	**Error Estandar**	**P Valor**	**Estimado**	**Error Estandar**	**P Valor**
β	Intercepto	0.8246	0.1399	4.00e-09	0.8132	0.1399	6.59e-09
	x1_2	0.3396	0.1275	0.00775	0.3343	0.1274	0.00871
	x1_3	0.3632	0.1280	0.00456	0.3596	0.1279	0.00496
	x2	2.1496	0.1362	<2e-16	2.2477	0.1398	<2e-16
	x3_2	-0.2189	0.0463	2.35e-06	-0.2261	0.0470	1.52e-06
	x4_2	-0.0994	0.0309	0.00130	-0.1025	0.0313	0.00106
	x5	0.0984	0.0343	0.00411	0.1047	0.0347	0.00257
σ	N/A	-0.1712	0.0115	<2e-16	-0.2789	0.0233	<2e-16
ν	N/A	N/A	N/A	N/A	2.0905	0.1865	<2e-16

Cuadro 5.1: Coeficientes estimados y errores estándar para todos los parámetros de los modelos de regresión a la media log Normal y log t de Student

Se puede observar que existe una relación positiva entre la mediana del tiempo de retraso mediano de las órdenes y casi todas las covariables. Por ejemplo, podemos observar que aquellas empresas que no hayan realizado compras de proveedores por internet o que no hayan realizado importaciones por despacho anticipado, no les afecta tanto el posible retraso de las órdenes. Es decir, que es más probable que ocurran retrasos en las entregas cuando las empresas compran por internet. Asimismo, mientras más se incremente el porcentaje de las órdenes de compras o pedidos a proveedores nacionales que presentan demoras en el tiempo de entrega, es más probable que el tiempo de este retraso sea mayor. Finalmente, si la principal modalidad que se utilizó para realizar compras es directamente al productor del insumo, se incrementan las posibilidades de que tome un mayor tiempo de retraso en entregar la órden, mientras que comprando a un distribuidor mayorista o minorista, ocurre todo lo contrario.

Asimismo podemos ver los resultados de los residuales cuantílicos estandarizados obtenidos en el ajuste de los datos en la figura 5.1.



(a) Q-Q plot del modelo Log Normal (b) Q-Q plot del modelo Log t

Figura 5.1: Gráfico Q-Q Plot de la estimación por el modelo Log t vs el modelo Log Normal

Como se aprecia en los gráficos Q-Q plot de los residuales estandarizados, en el modelo Log Normal hay un desajuste más pronunciado en las colas, demostrando la sensibilidad en el ajuste de los datos del modelo log Normal en presencia de valores atípicos. Finalmente, en

el cuadro 5.2 se compara ambos modelos mediante el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC).

Criterio	**Log t**	**Log Normal**
AIC	12,446.9	12,475.7
BIC	12,505.6	12,527.9
Devianza global	12,428.9	12,459.7

Cuadro 5.2: Criterios de comparación AIC, BIC y de Devianza global para los modelos de regresión log t de Student y log Normal para el ajuste de los datos con presencia de censura intervalar

Según estos resultados, se puede notar claramente un mejor ajuste de los datos considerados en la presente aplicación para el modelo de regresión log t en comparación al ajuste obtenido bajo el modelo Log Normal.

Capítulo 6

Conclusiones

6.1. Conclusiones

Cuando estamos hablando de modelos robustos a valores atípicos la literatura nos sugiere considerar distribuciones de probabilidad con colas más pesadas que la distribución normal, como lo puede ser la distribución t de Student ya que con sus parámetros de escala y de grados de libertad podemos controlar la dispersión de los datos del modelamiento; y finalmente lograr un mejor ajuste del modelo que no se vea afectado por estos valores atípicos.

Sin embargo, para el caso de una variable respuesta que presenta una censura intervalar, el presente es el primer trabajo que estudia si los valores atípicos en este tipo de datos también influyen en el ajuste a los datos del modelo, y si finalmente modelar con distribuciones con colas pesadas como el modelo t de Student sigue siendo una mejor opción para una adecuada estimación de los parámetros del modelo.

Mediante el estudio de simulación hemos demostrado que frente a datos que presentan censura intervalar, el modelo log t de Student también se sigue mostrando robusto frente a la presencia de observaciones atípicas en comparación con el modelo log Normal. Esto quedó evidenciado en el sesgo que presenta el modelo Log Normal en la estimación de los coeficientes de regresión a medida que se van introduciendo los valores atípicos en los datos, inclusive ocasionando que el verdadero valor de estos coeficientes se salgan del intervalo de confianza en cada uno de los escenarios *contaminados*. Por el contrario, con el modelo log t de Student este sesgo no se presenta, por lo que este comportamiento no ocurre.

Finalmente, mediante la aplicación del modelo para la estimación de las demoras en órdenes de compras de los proveedores de las empresas en el Perú, se pudo corroborar que el modelo log t logra un mejor ajuste a estos datos y es más robusto que usando el modelo log Normal para variables con censura intervalar.

Bibliografía

Ahsanullah, M., Kibria, B. G. y Shakil, M. (2014). *Normal Distribution. In Normal and Student's t Distributions and Their Applications*, Atlantis Press, Paris.

Barroso, F. J. C., García-Pérez, C. y Prieto-Alaiz, M. (2019). Modelling income distribution using the log student distribution: New evidence for european union countries, *Economic Modelling* .

Bleda, M. y Garces, A. T. (2002). Aplicación de los modelos de regresión tobit en la modelización de variables epidemiológicas censuradas, *Gaceta Sanitaria* **16**(2): 188–195.

Dunn, P. y Smyth, G. (1997). Randomized quantile residuals, *Journal of Computational and Graphical Statistics* **5**.

ENSUSALUD (2015). Superintendencia nacional de salud. encuesta nacional de satisfacción de usuarios en salud. `http://portales.susalud.gob.pe/web/portal/239`.

Hogg, R. V. y Klugman, S. A. (1983). On the estimation of long tailed skewed distributions with actuarial applications, *Journal of Econometrics* **23**(1): 91 – 102.
URL: *http://www.sciencedirect.com/science/article/pii/0304407683900775*

INEI (2015). Encuesta nacional de empresas. `http://demi.produce.gob.pe/Content/files/doc_05/Informe%20Tecnico%20de%20Encuesta%20de%20Empresas.pdf`.

Lange, K. L., Little, R. J. A. y Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution, *Journal of the American Statistical Association* **84**(408): 881–896.
URL: *http://www.jstor.org/stable/2290063*

Lindsey, J. C. y Ryan, L. M. (1998). Methods for interval-censored data, *Statistics in Medicine* **17**(2): 219–238.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rigby, R. A. y Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion), *Applied Statistics* **54**: 507–554.

Sal y Rosas, V. G., Moscoso-Porras, M., Ormeño, R., Artica, F., Bayes, C. L. y Miranda, J. J. (2019). Gender income gap among physicians and nurses in peru: a nationwide assessment, *he Lancet Global Health* **7**(4): e412–e413.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. y De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*, CRC Press.

Stasinopoulos, M., Rigby, B. y Mortan, N. (2018). *gamlss.cens: Fitting an Interval Response Variable Using 'gamlss.family' Distributions*. R package version 5.0-1.
URL: *https://CRAN.R-project.org/package=gamlss.cens*

Student (1908). The probable error of a mean, *Biometrika* **6**(1): 1–25.
URL: *http://www.jstor.org/stable/2331554*

Vallejos, C. A. y Steel, M. F. J. (2015). Objective bayesian survival analysis using shape mixtures of log-normal distributions, *Journal of the American Statistical Association* **110**(510): 697–710.
URL: *https://doi.org/10.1080/01621459.2014.923316*

West, M. (1984). Outlier models and prior distributions in bayesian linear regression, *Journal of the Royal Statistical Society. Series B (Methodological)* **46**(3): 431–439.
URL: *http://www.jstor.org/stable/2345685*

Zellner, A. (1976). Bayesian and non-bayesian analysis of the regression model with multivariate student-t error terms, *Journal of the American Statistical Association* **71**(354): 400–405.
URL: *http://www.jstor.org/stable/2285322*