UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

COMPARACIÓN DE ENFOQUES PARA SOLVENTAR LA VIOLACIÓN AL SUPUESTO DE IGUALDAD DE VARIANCIAS DEL ERROR EN EL MODELO LINEAL GAUSSIANO.

Tesis sometida a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Académica en Estadística

RUBÉN MORALES AGUILAR

Ciudad Universitaria Rodrigo Facio, Costa Rica

2020

Dedicatoria

A Dios, mi familia, profesores en general, colegas y amigos(as).

Agradecimientos

El principal agradecimiento es para Dios por brindarme la oportunidad de vivir, a mis padres y hermana por su incondicional apoyo hacia mi persona. A mis profesores por su dedicación, y por acompañarme aportando su experiencia y conocimientos. Finalmente a mis colegas, y amigos(as) por compartir cada uno(a) los dones que Dios les dio conmigo y así conseguir la elaboración de este trabajo.

"Esta tesis fue aceptada por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Académica en **Estadística**"

OSCAR JAVIER CENTENO MORA (FIRMA)
Firmado digitalmente por OSCAR JAVIER CENTENO MORA (FIRMA)
Fecha: 2021.03.01 13:50:10 -06'00'

M.Sc. Oscar Centeno Mora

Representante del Decano

Sistema de Estudios de Posgrado

GILBERT BRENES CAMACHO (FIRMA)
Firmado digitalmente por GILBERT BRENES CAMACHO (FIRMA)
Fecha: 2021.02.10 18:42:01 -06'00'

Dr. Gilbert Brenes Camacho

Director de Tesis

Ricardo Alvarado Barrantes
Firmado digitalmente por Ricardo Alvarado Barrantes
Fecha: 2021.02.11 21:38:50 -06'00'

Dr. Ricardo Alvarado Barrantes

Asesor

JOHNNY ANGEL MADRIGAL PANA (FIRMA)
Digitally signed by JOHNNY ANGEL MADRIGAL PANA (FIRMA)
Date: 2021.02.12 11:56:57 -06'00'

M.Sc. Johnny Madrigal Pana

Asesor

Guaner Rojas Rojas
Firmado digitalmente por Guaner Rojas Rojas
Fecha: 2021.02.10 11:26:21 -06'00'

Dr. Guaner Rojas Rojas

Representante
Programa de Posgrado en Estadística

RUBEN DE JESUS MORALES AGUILAR (FIRMA)
Firmado digitalmente por RUBEN DE JESUS MORALES AGUILAR (FIRMA)
Fecha: 2021.02.10 09:12:21 -06'00'

Rubén de Jesús Morales Aguilar

Sustentante

Índice

Resumen

Este estudio evalúa una gama de propuestas metodológicas alternas al análisis de regresión lineal estimado mediante mínimos cuadrados ordinarios, que es uno de los más conocidos y utilizado para estudiar la relación entre variables. Bajo el cumplimiento de determinados supuestos, a sus estimadores se les atribuyen importantes propiedades estadísticas. Pero, en específico ante la violación del supuesto de varianza constante del error, pueden llegar a ser altamente ineficientes.

Por este motivo, se busca determinar cuál modelo es más robusto ante distintos orígenes de **heterocedasticidad**, como la naturaleza de las variables predictoras, la presencia de valores extremos y el sesgo de variable omitida.

Los datos empleados para el análisis del caso empírico provienen de la Encuesta de Ingresos y Gastos 2013 realizada por el Instituto Nacional de Estadística y Censos de Costa Rica. De este, para explicar el **Gasto** mensual en alimentos y bebidas no alcohólicas consumidas en el hogar, se utilizan las siguientes variables: el **Ingreso** monetario corriente neto sin valor locativo, el número de miembros del hogar, la edad del jefe del hogar y su escolaridad. Se realizó un estudio de simulación Monte Carlo, que está basado en las variables y el tamaño de muestra utilizados (n=5687, hogares) en el estudio empírico. Considera las tres diferentes causas de **heterocedasticidad** mencionadas y una cantidad total de 10 200 repeticiones para cada escenario propiciado por ellas. Los modelos analizados mediante el ejercicio de simulación presentan diferentes desempeños, según el origen de la **heterocedasticidad**. En concreto, la regresión cuantílica y el planteamiento bayesiano propuesto no ofrecen mayores ventajas en cuanto a reducción del error estándar se refiere. Los resultados sugieren que el modelo Lineal Doble Generalizado muestra el mejor desempeño general, seguido del modelo Mínimos Cuadrados Ponderados. Para extender los resultados obtenidos en la presente investigación, se considera apropiado en futuros estudios incluir una mayor cantidad de técnicas en el conjunto de modelos por evaluar, diferentes porcentajes de contaminación de valores extremos en las variables, tipos de sesgo y distribución de las variables…además de diferentes tamaños de muestra.

Palabras clave: Modelo de regresión lineal múltiple (mínimos cuadrados ordinarios), modelo de regresión lineal general (mínimos cuadrados ponderados), modelo de regresión cuantílica, modelo lineal doble generalizado, modelo lineal doble generalizado bayesiano, **heterocedasticidad** residual.

Índice de cuadros

Índice de figuras

Lista de abreviaturas

Abreviatura	Significado
BDGLM	Bayesian Double Generalized Linear Model/Modelo Lineal Doble Generalizado Bayesiano
CME	Cuadrado Medio de Error
CMR	Cuadrados Medio de Regresión
DGLM	Double Generalized Linear Model /Modelo Lineal Doble Generalizado
EAM	Error Absoluto Medio
ECM	Error Cuadrático Medio
ENIGH	Encuesta Nacional de Ingresos y Gastos de los Hogares
GL	Grados de Libertad
GLS/MCP	(General Least Squares) / (Mínimos Cuadrados Generales "Ponderados")
HCCME	Heteroscedasticity Consistent Covariance Matrix Estimator/Estimadores de la Matriz de Covarianza Heteroscedásticamente Consistentes
HPD	Highest Posterior Density/ Densidad Posterior más Probable
INEC	Instituto Nacional de Estadística y Censos
MCMC	Markov Chain Monte Carlo Methods/ Métodos de Cadena de Markov y Monte Carlo
OLS/MCO	Ordinary Least Squares/Mínimos Cuadrados Ordinarios
MELI	Mejores Estimadores Lineales Insesgados
NID	Normal e Idénticamente Distribuido
NCV	Non Constant Variance/ Varianza No Constante
PGD	Proceso Generador de Datos
QQ	Quantile-Quantile/ Cuanti-Cuantil
QREG	Quantile Regression/ Regresión Cuantílica
RECM	Raíz del Error Cuadrático Medio
RWLS	Robust Weighted Least Squares/Mínimos Cuadrados Ponderados Robustos
SCE	Suma de Cuadrados de Error
SCR	Suma de Cuadrados de Regresión
SCT	Suma de cuadrados total
UPM	Unidad primaria de muestreo

UNIVERSIDAD DE COSTA RICA

SEP Sistema de Estudios de Posgrado

Autorización para digitalización y comunicación pública de Trabajos Finales de Graduación del Sistema de Estudios de Posgrado en el Repositorio Institucional de la Universidad de Costa Rica.

Yo, Rubén de Jesús Morales Aguilar, con cédula de identidad 113260454, en mi condición de autor del TFG titulado Comparación de enfoques para solventar la violación al supuesto de igualdad de varianzas del error en el modelo lineal gaussiano.

Autorizo a la Universidad de Costa Rica para digitalizar y hacer divulgación pública de forma gratuita de dicho TFG a través del Repositorio Institucional u otro medio electrónico, para ser puesto a disposición del público según lo que establezca el Sistema de Estudios de Posgrado. **SI** [X] **NO** *[ ]

***En caso de la negativa favor indicar el tiempo de restricción: _____ año (s).**

Este Trabajo Final de Graduación será publicado en formato PDF, o en el formato que en el momento se establezca, de tal forma que el acceso al mismo sea libre, con el fin de permitir la consulta e impresión, pero no su modificación.

Manifiesto que mi Trabajo Final de Graduación fue debidamente subido al sistema digital Kerwá y su contenido corresponde al documento original que sirvió para la obtención de mi título, y que su información no infringe ni violenta ningún derecho a terceros. El TFG además cuenta con el visto bueno de mi Director (a) de Tesis o Tutor (a) y cumplió con lo establecido en la revisión del Formato por parte del Sistema de Estudios de Posgrado.

INFORMACIÓN DEL ESTUDIANTE:

Nombre Completo: Rubén de Jesús Morales Aguilar.

Número de Carné: A53735 Número de cédula: 113260454.

Correo Electrónico: rdjmoralesa1@gmail.com.

Fecha: 07/02/2021. Número de teléfono: 88960543.

Nombre del Director (a) de Tesis o Tutor (a): Gilbert Brenes Camacho.

RUBEN DE JESUS MORALES AGUILAR (FIRMA)

Firmado digitalmente por RUBEN DE JESUS MORALES AGUILAR (FIRMA)
Fecha: 2021.02.07 23:07:08 -06'00'

FIRMA ESTUDIANTE

Nota: El presente documento constituye una declaración jurada, cuyos alcances aseguran a la Universidad, que su contenido sea tomado como cierto. Su importancia radica en que permite abreviar procedimientos administrativos, y al mismo tiempo genera una responsabilidad legal para que quien declare contrario a la verdad de lo que manifiesta, puede como consecuencia, enfrentar un proceso penal por delito de perjurio, tipificado en el artículo 318 de nuestro Código Penal. Lo anterior implica que el estudiante se vea forzado a realizar su mayor esfuerzo para que no sólo incluya información veraz en la Licencia de Publicación, sino que también realice diligentemente la gestión de subir el documento correcto en la plataforma digital Kerwá.

CAPÍTULO 1: INTRODUCCIÓN

Motivó a realizar esta experiencia la percepción de ciertas distorsiones en el análisis de datos y la posibilidad de ofrecer alternativas para superarlas. Es evidente la incremental necesidad de que las decisiones trascendentes tengan un enfoque científico puesto que están basadas en datos. Si la validez de los análisis suscita dudas, no solo pueden llevar a decisiones incorrectas sino que van conduciendo a que se pierda credibilidad en los datos como base de las decisiones.

Ante este tipo de inconvenientes que se pueden presentar en la práctica, históricamente se han desarrollado y se siguen ampliando los métodos. Propiamente para los fines investigativos de este trabajo se estudiarán, en relación a la **heterocedasticidad**, los mínimos cuadrados ordinarios, los mínimos cuadrados ponderados, la regresión cuantílica, los modelos lineales generalizados de dispersión variable y su versión bayesiana planteada. Estimar así dichas relaciones permitirá inferir de una manera más acertada.

Considerando la tesis de Damodar Gujarati (Gujarati, 2010):

> El análisis de regresión es ampliamente utilizado para predicción y pronóstico en diversos campos pues ayuda a establecer la relación estadística entre variables. Quizás el modelo clásico de regresión lineal sea uno de los más conocidos. Su fundamento es el método de mínimos cuadrados y se le atribuyen importantes propiedades numéricas y estadísticas bajo el cumplimiento de supuestos determinados, pues así garantiza estimadores insesgados, consistentes y de varianza mínima (Gujarati, 2010, p. 78).
>
> Ante la violación de uno o más supuestos, los estimadores de mínimos cuadrados de los coeficientes de regresión pueden o no dejar de ser considerados como mejores estimadores lineales insesgados (MELI), en específico en el caso de la varianza del error no constante, siguen siendo consistentes e insesgados, pero pueden llegar a ser altamente ineficientes. En ocasiones esta falta se origina en la naturaleza teórica de las variables en estudio, la presencia de observaciones atípicas, la omisión de variables importantes en el modelo, entre otras. (Gujarati, 2010, p. 366).

Se desarrolla la investigación para generar conjuntos de datos que consigan manifestar la característica expuesta en el análisis residual de un modelo de regresión lineal. Para el efecto, se utiliza el método de simulación Monte Carlo para componer la variable respuesta a partir de parámetros hipotéticos fijos, distribuciones uniformes pseudoaleatorias en adición de un error creado con una distribución normal de media cero y varianza no constante o bien de varianza constante, pero absorbiendo la variabilidad no explicada por una variable omitida en el modelo.

Los estudios de simulación están además estrechamente relacionados con el estudio del caso empírico en donde se modela la variable **Gasto** en consumo de alimentos en función del **Ingreso** monetario corriente neto del hogar, la cantidad de miembros , la edad del jefe y su escolaridad (datos provenientes de ENIGH 2013).

Se busca determinar cuál de las técnicas o métodos de estimación para atender la violación del supuesto de igualdad de varianzas residual realiza las estimaciones más adecuadas y con errores estándar de menor magnitud según los orígenes de **heterocedasticidad** bajo estudio. Por lo tanto, este análisis se enfoca en la variable que tiene el mayor peso, utiliza medidas de ajuste (RECM; EAM) y conocidos estadísticos de tendencia central y variabilidad.

Es importante considerar que en este estudio se utilizan algunas de las técnicas sobresalientes en el tratamiento del error heterocedástico en los modelos lineales y que existe una amplia variedad de motivos por los cuales el error puede tener este comportamiento más allá de los que en el presente se analizan.

Problema de investigación

Al contar con una gama de propuestas metodológicas para abordar la violación del supuesto de igualdad de varianzas del error en el modelo lineal gaussiano múltiple (OLS), se busca determinar cuál de ellas es más robusta ante distintas causas de **heterocedasticidad**.

Objetivo general

El objetivo general es comparar la estimación y la matriz de variancias y covariancias de los coeficientes del modelo lineal gaussiano múltiple (OLS) cuando se viola el supuesto de igualdad de varianzas del error, frente a modelos alternativos que incorporan en su estimación medidas que contrarrestan esta falta, valorando distintos orígenes de **heterocedasticidad**.

Objetivos específicos

a) Verificar la estimación y comparar la varianza de los coeficientes para los modelos estadísticos que relajan el supuesto de igualdad de varianzas del error en el modelo lineal gaussiano (OLS), cuando existe **heterocedasticidad** y esta procede de la naturaleza de las variables predictoras.

b) Confirmar la estimación y confrontar la varianza de los coeficientes para los modelos estadísticos que suavizan el supuesto de igualdad de varianzas del error en el modelo lineal gaussiano (OLS), cuando existe **heterocedasticidad** y esta se origina del sesgo de variable omitida.

c) Comprobar la estimación y contrastar la varianza de los coeficientes para los modelos que flexibilizan el supuesto de igualdad de varianzas del error en el modelo lineal gaussiano (OLS), cuando existe **heterocedasticidad** y esta surge de la presencia de valores extremos.

CAPTULO 2: MARCO TEÓRICO

El análisis de regresión tiene sus orígenes en la última parte del siglo XIX con un famoso ensayo de Francis Galton donde estudiaba la relación entre la altura de los padres y sus hijos. Él notó que la altura de los hijos, tanto de padres altos como bajos, tendía a dirigirse a la estatura promedio de la población, y consideró esta tendencia como un retorno a la mediocridad. Posteriormente desarrolló el modelo matemático precursor de los modelos de regresión. (Neter J. M., 1990)

La interpretación moderna de regresión afirma lo siguiente: "El análisis de regresión trata del estudio de la dependencia de una variable (variable respuesta) respecto de una o más variables (variables explicativas) con el objetivo de estimar o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las segundas." (Gujarati, 2010, pág. 15)

De acuerdo con Neter et al. (1990), un modelo de regresión es un medio formal de expresar los dos ingredientes esenciales de una relación estadística: "Una tendencia de la variable respuesta **(Y)** a variar con la variable explicativa **(X)** de una forma sistemática y una dispersión de puntos alrededor de la curva de la relación estadística." (Neter et al., 1990, pág. 6). Ambas características se incorporan al modelo de regresión cuando se postula que: "existe una distribución de probabilidad de **(Y)** para cada nivel de **(X)** y que los promedios de estas distribuciones varían de una manera sistemática con **(X)**." (Neter et al., 1990, pág. 6)

El objetivo principal del análisis de regresión es estimar la función de regresión poblacional con base en una función de regresión muestral, pues en la mayoría de los casos el análisis se basa en una sola muestra tomada de una población. (Gujarati, 2010)

2.1. Modelo de regresión lineal múltiple

Para analizar la relación entre variables la mayoría de las ocasiones es impráctico realizarlo mediante el uso de una variable respuesta y una explicativa solamente (regresión lineal simple), debido a que en diversas situaciones la teoría o la experiencia previa corrobora que existe más de una variable **(X)** relacionada con la variable **(Y)**, por lo que el modelo de regresión lineal múltiple viene a solventar este detalle.

Según Neter et al. (1990), en general el modelo de regresión lineal múltiple se puede expresar así:

(1)
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

donde:

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ son parámetros, cada coeficiente (i-ésimo) está sujeto a una variable explicativa y mide el cambio en el valor de la media de (Y, variable respuesta), por unidad de incremento en la variable explicativa (i-ésima) manteniendo las demás variables incluidas en el modelo constantes (Gujarati, 2010).
- $X_1, X_2, \ldots, X_{p-1}$ representa p-1 distintas variables predictoras.
- *p* es la cantidad de parámetros a estimar en el modelo.
- $\boldsymbol{\varepsilon}$ es el término de error del modelo o perturbación aleatoria. Para la estimación a partir de una muestra, cada $\varepsilon_i = Y_i - E(Y|X_i)$. Lo cual se traduce en la desviación de Y_i de la recta de regresión verdadera, por lo que es desconocido.

Si el modelo poblacional **(1)** se escribe para todas las observaciones se obtiene el siguiente sistema

(2)
$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_{p-1} X_{1,p-1} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_{p-1} X_{2,p-1} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_{p-1} X_{n,p-1} + \varepsilon_n \end{aligned}$$

Este sistema de ecuaciones se puede expresar de una manera más compacta utilizando notación matricial, para lo que es necesario definir las siguientes matrices (Neter et al., 1990):

(3)
$$\underset{\boldsymbol{n\,x1}}{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{\boldsymbol{n\,xp}}{\boldsymbol{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{\boldsymbol{p\ x1}}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{\boldsymbol{n\ x1}}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

El modelo también puede expresarse matricialmente como sigue:

(4)
$$\underset{nx1}{\boldsymbol{Y}} = \underset{nxp}{\boldsymbol{X}}\ \underset{px1}{\boldsymbol{\beta}} + \underset{nx1}{\boldsymbol{\varepsilon}}$$

2.1.1. Supuestos del modelo de regresión lineal múltiple

Este modelo (Gujarati, 2010) plantea los supuestos que a continuación se detallan:

1. El modelo es lineal en los parámetros, "debido a que ningún parámetro aparece como exponente, multiplicado o dividido por otro parámetro." (Neter et al., 1990, pág. 10)
2. Los valores fijos o regresores $\boldsymbol{X_j}$ son independientes del término de error ε, lo que significa:

(5)
$$Cov\{\varepsilon, \boldsymbol{X}\} = 0$$

3. El valor esperado de la perturbación ε es igual a cero:

(6)
$$E\{\varepsilon|X\} = 0\,.$$

4. La varianza de las perturbaciones ε_i se supone constante:

(7)
$$Var\{\varepsilon|X\} = \sigma^2$$

5. La correlación entre las perturbaciones ε_i y ε_j para $i \neq j$ es cero:

(8)
$$Cov\{\varepsilon_{i,}\varepsilon_j\} = 0$$

6. El número de observaciones **(n)** debe ser mayor al de los parámetros **(p)** por estimar:
   $\mathrm{n} > p$
7. Debe haber variabilidad a lo interno de cada variable independiente.

8. No existe relación lineal exacta entre dos o más variables explicativas, lo que significa que ninguna de las regresoras puede escribirse como combinación lineal exacta de las otras.
9. No hay sesgo de especificación, lo que quiere decir que en el modelo no se omiten variables importantes.

2.1.2. Estimación de los coeficientes del modelo

El objetivo principal del modelo de regresión es la estimación de los coeficientes **β** a partir de una muestra dada. La función de regresión muestral es una estimación de la función de regresión poblacional y está dada por:

(9)
$$\underset{nx1}{\widehat{\boldsymbol{Y}}} = \underset{nx1}{\boldsymbol{X}}\ \underset{nx1}{\widehat{\boldsymbol{\beta}}}$$

donde:

- $\widehat{\boldsymbol{Y}}$ es el valor estimado de la función de regresión o *valor ajustado*. Dicho de otro modo $\widehat{\boldsymbol{Y}}$ es el estimador puntual insesgado de $E(Y/X)$.
- $\boldsymbol{X}$ es la matriz de variables predictoras.
- $\widehat{\boldsymbol{\beta}}$ son estimaciones puntuales de los parámetros poblacionales y se denotan de forma matricial de la siguiente manera (Neter et al., 1990):

(10)
$$\underset{\boldsymbol{px1}}{\widehat{\boldsymbol{\beta}}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$$

Con el fin de estimar los parámetros de la manera más precisa posible, se emplea el método de mínimos cuadrados; este selecciona los estimadores $\hat{\beta}_k$ con $(k = 0, \dots, p-1)$ que minimizan la expresión **Q** para una muestra dada (Neter et al., 1990):

(11)
$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

donde:

- A la desviación entre $\boldsymbol{Y}$ y $\widehat{\boldsymbol{Y}}$ se le conoce como residual:

(12)
$$\underset{\boldsymbol{nx1}}{\boldsymbol{e}} = \underset{\boldsymbol{nx1}}{\boldsymbol{Y}} - \underset{\boldsymbol{nx1}}{\widehat{\boldsymbol{Y}}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

- Por lo tanto, **Q** resulta en la suma de cuadrados residual.

De manera que los estimadores $\hat{\beta}$ que minimizan la expresión **Q** resultan de la siguiente expresión (Gujarati, 2010):

(13)
$$\underset{\boldsymbol{px1}}{\widehat{\boldsymbol{\beta}}} = \underset{pxp}{(\boldsymbol{X}'\boldsymbol{X})^{-1}} \underset{px1}{(\boldsymbol{X}'\boldsymbol{Y})}$$

2.1.3. Propiedades de los estimadores de mínimos cuadrados

***Teorema de Gauss-Markov*:**

El teorema de Gauss-Markov describe que: "bajo los supuestos del modelo de regresión los estimadores de mínimos cuadrados son insesgados y tienen varianza mínima dentro de la clase de estimadores lineales insesgados." (Neter et al., 1990, pág. 20). No hace condicionamiento respecto a la distribución de probabilidad del término de error y en consecuencia tampoco de Y. Por lo tanto, en la medida que se satisfagan los supuestos del modelo de regresión lineal, el teorema será válido. Lo anterior cubre las propiedades de los estimadores en repetidas muestras pequeñas o bien finitas. Cuando la muestra es grande, se ha demostrado que estos estimadores alcanzan la propiedad de consistencia, lo que significa que su varianza tiende a cero al aumentar el tamaño de la muestra y que la estimación tenderá al verdadero valor del parámetro. (Gujarati, 2010)

2.1.4. Modelo de regresión lineal múltiple con errores normales

En el análisis de regresión el objetivo no solo es estimar la función de regresión muestral sino también emplearla para tener inferencias de la función de regresión poblacional. Para esto es necesario hacer estimaciones por intervalo y pruebas estadísticas de hipótesis. Debido a que los errores (ε_i) no están condicionados a una forma distribucional particular, se supone por lo general que cada término de error (ε_i) proviene de una distribución Normal (Gujarati, 2010), lo cual se puede expresar así:

(14) $$\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$$

El supuesto de normalidad para el término de error es justificable puesto que representa el efecto de las variables no contempladas en el modelo que afectan la variable respuesta y se reflejan en él de manera compuesta. (Neter et al., 1990). Gracias al teorema del límite central se puede decir que, si existe un gran número de variables aleatorias independientes con idéntica distribución probabilística, la mayoría de las veces, a medida que el número de estas variables tiende a infinito, la distribución de la suma de ellas tiende a ser normal. Esto permite derivar las distribuciones de probabilidad muestral de $\hat{\beta}_k$ pues cualquier función lineal de variables normalmente distribuidas estará también normalmente distribuida. La suposición de normalidad desempeña además un papel fundamental en la utilización de las pruebas estadísticas $t, F\ y\ \chi^2$, las cuales impactan el proceso de inferencia. (Gujarati, 2010)

Al conocerse que para cualesquiera dos variables normalmente distribuidas una covarianza o correlación de cero se traduce como independencia entre ellas, esto se puede extender para cada ε_i y ε_j con $i \neq j$ (un par cualquiera de errores aleatorios), para afirmar que cada término de error se distribuye de manera normal y se encuentra independientemente distribuido. (Gujarati, 2010)

Debido a que se especifica la forma funcional de la distribución de probabilidad del error, los estimadores de los parámetros también se pueden obtener por el método de máxima verosimilitud. (Neter et al., 1990). Los estimadores de máxima verosimilitud para β coinciden con los estimadores de mínimos cuadrados. En el caso de σ^2 el estimador máximo verosímil resulta ser sesgado contra el que se obtiene por mínimos cuadrados. Pero, al aumentar el tamaño de la muestra, el primero alcanza al estimador de mínimos cuadrados. (Gujarati, 2010)

2.1.5. Inferencia en el modelo de regresión lineal múltiple

Los estimadores $\hat{\beta}$ obtenidos por mínimos cuadrados son insesgados (Neter et al., 1990), por lo tanto:

(15) $$E\{\hat{\beta}\} = \beta$$

Su matriz de varianzas y covarianzas $\sigma^2\{\hat{\beta}\}$ está dada por:

$$\underset{pxp}{\sigma^2\{\hat{\beta}\}} = \sigma^2(X'X)^{-1} \tag{16}$$

donde:

A. $\boldsymbol{\sigma^2}$ representa la varianza de los términos de error.

La matriz de varianzas y covarianzas se estima mediante la matriz $s^2\{\hat{\beta}\}$ (Neter et al., 1990):

$$\underset{pxp}{s^2\{\hat{\beta}\}} = CME(X'X)^{-1} \tag{17}$$

donde:

$$CME = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p} \tag{18}$$

Al respecto se ha demostrado que el error cuadrático medio o cuadrado medio de error (CME) es un estimador insesgado de $\boldsymbol{\sigma^2}$. (Neter et al., 1990)

Las pruebas de hipótesis para β_k se establecen de la manera usual (Neter et al., 1990):

$$\begin{aligned} H_0&: \beta_k = 0 \\ H_A&: \beta_k \neq 0 \end{aligned} \tag{19}$$

Para el modelo de regresión de errores normales, se tiene entonces que el estadístico de prueba:

$$t^* = \frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \tag{20}$$

Se distribuye como una T de Student con $(n-p)$ grados de libertad y **k** = 0,1,…,p-1.

Y la regla de decisión para controlar el error tipo I está dada por:

$$|t^*| \leq t(1 - {}^{\alpha}/_{2}\,; n-p) \tag{21}$$

Los límites de confianza $(1-\alpha)$ para β_k se pueden calcular mediante (Neter et al., 1990):

$$\hat{\beta}_k \pm t(1-\alpha/2\,; n-p)s\{\hat{\beta}_k\} \tag{22}$$

2.2. Heterocedasticidad en el modelo de regresión lineal múltiple

En la práctica del análisis de regresión lineal algunas veces se incurre en violaciones a los supuestos del modelo. Aquí se enfatiza en la violación del supuesto de igualdad de varianzas de las perturbaciones aleatorias ε_i, que simbólicamente se puede expresar:

(23) $$Var\{\varepsilon|X\} = \sigma_i{}^2$$

2.2.1. Naturaleza del problema

La existencia de **heterocedasticidad** en el modelo lineal gaussiano puede surgir por varias razones, entre ellas principalmente (Studenmund, 2000) (Gujarati, 2010):

- La presencia de datos atípicos que, si el tamaño de la muestra es pequeño, pueden alterar sustancialmente los resultados del análisis de regresión. Estos valores son aquellas observaciones que son o muy grandes o muy pequeñas con relación a las demás observaciones de la muestra.
- La naturaleza teórica del problema puede indicar **heterocedasticidad**, en algunas variables es común esperar que cuanto mayores son sus valores mayor será la dispersión absoluta del modelo. Un ejemplo clásico de lo anterior es cuando se está interesado en regresar el ahorro sobre el nivel de ingresos de un hogar, pues la teoría indica que a mayor ingreso más grande será la varianza del ahorro.
- La omisión de una variable en el modelo, debido a que la omisión de la porción omitida no representada por las demás variables explicativas es absorbida por el término de error. Si esta variable posee un efecto heteroscedástico, este también se reflejará en el error.
- La asimetría en la distribución de una o más variables regresoras incluidas en el modelo. Los ejemplos más frecuentes son, entre otras: el Ingreso, las variables que miden el bienestar, la escolaridad, la edad.

- La reducción del error en pruebas o evaluaciones aplicadas en diferentes momentos temporales, aplicable al campo de psicometría en los análisis basados en modelos de aprendizaje de errores.
- El desconocimiento, al momento de realizar el análisis, de la existencia de dos o más niveles de análisis en los datos, conocido como sesgo de agregación.
- La especificación de la forma funcional incorrecta entre predictando y predictores.
- La transformación incorrecta de los datos.

2.2.2. Consecuencias

En presencia de **heterocedasticidad,** las estimaciones $\hat{\beta}$ hechas mediante mínimos cuadrados ordinarios siguen siendo lineales, insesgadas y consistentes, pero ya no son de varianza mínima o eficientes. (Gujarati, 2010). Esto último debido a que se viola una de las condiciones necesarias para que se cumpla el teorema de Gauss Markov, que supone varianza constante del error.

La violación implica que la varianza del término de error corresponde a:

(24)

$$\underset{n\,xn}{\sigma^2\{\varepsilon\}} = \Omega = \begin{bmatrix} {\sigma^2}_1 & 0 & \cdots & 0 \\ 0 & {\sigma^2}_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & {\sigma^2}_n \end{bmatrix}$$

Por lo que la fórmula habitual de la varianza de los estimadores expresada en la ecuación 16 aproxima incorrectamente la matriz de varianzas y covarianzas de los estimadores de los parámetros β. Alternativamente debe utilizarse la siguiente expresión:

(25)

$$\underset{pxp}{\sigma^2\{\hat{\beta}\}} = (X'X)^{-1}\, X'\Omega\, X(X'X)^{-1}$$

Lo antes mencionado sobre $\sigma^2\{\hat{\beta}\}$ da origen a dos situaciones:

1. **Utilizar OLS considerando la correcta aproximación de $\boldsymbol{\sigma^2\{\beta\}}$.**

Si se utiliza la versión corregida de la matriz $\sigma^2\{\hat{\beta}\}$ se sigue enfrentando el problema de eficiencia, por lo que los contrastes usuales t, F tenderían a ser más exigentes y a la vez los intervalos de confianza más amplios. (Verbeek, 2004)

2. Utilizar OLS considerando la incorrecta aproximación de $\boldsymbol{\sigma^2\{\beta\}}$.

Se conoce que en un modelo de OLS homoscedástico:

$$E(CME) = \sigma^2$$

Pero, si se ignora la **heterocedasticidad** y se toma el CME como estimador de σ^2, esto causará un sesgo, pues en general no se puede saber si ese valor (CME) sobreestima o subestima la varianza real de las perturbaciones aleatorias ε_i. Verbeek (2004) afirma entonces que las varianzas y covarianzas de las estimaciones por mínimos cuadrados de los coeficientes β_k son sesgadas cuando la **heterocedasticidad** está presente pero es ignorada, por lo que también las pruebas de hipótesis e intervalos de confianza no serán válidas.

Se detalla a continuación cómo se ven afectados los intervalos de confianza y la prueba de significancia individual o prueba t:

1. Para el caso de la prueba de hipótesis de un coeficiente de regresión $\boldsymbol{k}$ o prueba t, bajo el supuesto de homoscedasticidad, se tiene que el estadístico de prueba sería:

(26)
$$t = \frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 * [(X'X)^{-1}]_{kk}}}$$

Donde k responde al $k - ésimo$ elemento de la diagonal de la matriz $s^2\{\hat{\beta}\}$.

Si existe **heterocedasticidad**, entonces el estadístico de prueba sería diferente para cada observación, como se muestra seguidamente:

(27)
$$t_i = \frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}_i^2 * [(X'X)^{-1}]_{kk}}}$$

2. Para el caso de los intervalos de confianza para β_k bajo el supuesto de homoscedasticidad, estos se obtienen mediante:

(28) $$\hat{\beta}_k \pm t(1-\alpha/2\,;n-p) * \sqrt{(\hat{\sigma}^2 * [(X'X)^{-1}]_{kk})}$$

En presencia de **heterocedasticidad**, existiría un intervalo de confianza por cada observación (i) para cada variable explicativa (k) en el modelo.

(29) $$\hat{\beta}_k \pm t(1-\alpha/2\,;n-p) * \sqrt{(\hat{\sigma}_i^2 * [(X'X)^{-1}]_{kk})}$$

2.2.3. Cómo detectar la heterocedasticidad

En la práctica, usualmente el análisis de **heterocedasticidad** se lleva a cabo una vez ajustado el modelo de regresión lineal y, para evidenciar la presencia de **heterocedasticidad,** se recurre a las gráficas y las pruebas formales. (Gujarati, 2010)

En cuanto a las gráficas, es bien sabido que los gráficos del valor absoluto de los residuales, o de los residuales al cuadrado contra la variable X, o contra los valores ajustados, son útiles para diagnosticar varianza no constante en el término de error (Neter et al., 1990), especialmente cuando en estos se evidencia variación creciente o decreciente con la variable X o con el valor ajustado. Este método puede resultar útil si la **heterocedasticidad** proviene de una sola variable regresora. Sin embargo, si la variable regresora se origina de la combinación lineal de todas las variables incluidas en el modelo, estas serán insuficientes.

Por lo anterior se recomienda complementar el análisis con las pruebas formales. Entre las pruebas formales para detectar **heterocedasticidad** se pueden mencionar: la prueba de Park, la prueba de Glejser, la prueba de Goldfeld Quandt, la prueba general de **heterocedasticidad** de White, la prueba de Koenker-Basset, la prueba de Levene modificada, la prueba de Breusch-Pagan-Godfrey. (Arce & Mahía, 2008). Si en conjunto las pruebas formales y la descripción gráfica presentan evidencia de **heterocedasticidad,** es preciso remediar el modelo o recurrir a otros que garanticen una inferencia no errada.

2.2.4. Modelos correctivos para obtener estimadores deseables

Cuando se tiene evidencia de que los residuos poseen heterocedasticidad, existen varias formas de corregir el problema, entre las cuales se mencionan: modelo de mínimos cuadrados generales o

mínimos cuadrados ponderados, modelo de regresión cuantílica, modelos lineales generalizados con dispersión variable y modelos lineales generalizados bayesianos con dispersión variable.

2.3. Modelo de regresión lineal múltiple general

Cuando de manera aislada existe presencia de **heterocedasticidad** en el modelo de regresión lineal, el enfoque de los mínimos cuadrados generales o también llamados los mínimos cuadrados ponderados (Neter et al., 1990) es uno de los más utilizados como medida correctiva para obtener las estimaciones de los parámetros.

2.3.1. Mínimos cuadrados ponderados

Si las varianzas del término de error son desiguales, el modelo de regresión general para cada observación puede ser expresado como sigue (Neter et al., 1990):

(30) $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

donde:

- $\boldsymbol{\varepsilon_i}$ es el término de error, distribuido $\boldsymbol{N(0, \sigma^2{}_i)}$.

La matriz de varianzas y covarianzas para el término de error se expresa en la ecuación (24), (Neter et al., 1990). Los estimadores de mínimos cuadrados de los parámetros poblacionales se pueden obtener al minimizar la siguiente expresión (Neter et al., 1990):

(31) $$Q_w = \sum_{i=1}^{n} w_i \left(Y_i - \hat{Y}\right)^2$$

Se puede observar que el criterio de mínimos cuadrados ponderados generaliza el de mínimos cuadrados ordinarios pues este último es el caso cuando $w_i = 1$. Siendo W una matriz diagonal que contiene los pesos w_i (Neter et al., 1990):

(32) $$\underset{n\,xn}{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

Para obtener estimadores de mínimos cuadrados de los parámetros poblacionales se consideran tres escenarios:

1. ${\sigma^2}_i$ conocidas.
2. ${\sigma^2}_i$ conocidas debidas a una constante de proporcionalidad.
3. ${\sigma^2}_i$ desconocidas.

Por motivo de que muy probablemente en la práctica no se conozcan las varianzas, en este estudio solo se detallará el caso en donde las varianzas son desconocidas.

2.3.2. Caso de varianzas desconocidas

Como en este caso es necesario estimar las varianzas, es bien sabido que la varianza del término de error ε_i denotada por σ_i^2 puede ser expresada como sigue:

(33)
$$\sigma_i^2 = E\{\varepsilon_i^2\} - \big(E(\varepsilon_i)\big)^2$$

Como $E(\varepsilon_i) = 0$ de acuerdo con los supuestos del modelo de regresión se obtiene:

(34)
$$\sigma_i^2 = E\{\varepsilon_i^2\}$$

Se conoce que el residual al cuadrado e_i^2 es un estimador de σ_i^2, además el valor absoluto de los residuales $|e_i|$ es un estimador de σ_i la desviación estándar (Neter et al., 1990).

El proceso de estimación de las varianzas se da según Neter et al. (1990) como sigue:

1. Se ajusta el modelo de regresión lineal mediante el método de mínimos cuadrados ordinarios y se examinan los residuales.

2. Se estima la función de varianza mediante una regresión de los residuales al cuadrado o la función de desviación estándar mediante una regresión del valor absoluto de los residuales contra el predictor apropiado, según lo decide el investigador.

3. Se usan los valores ajustados de la función de varianzas $\hat{v}_i$ o desviaciones estándar estimadas $\hat{s}_i$ para obtener los pesos w_i como el inverso de la varianza.

(35)

$$w_i = \frac{1}{\hat{v}_i} = \frac{1}{(\hat{s}_i)^2}$$

4. La matriz W se forma con los pesos w_i obtenidos.

5. Se estiman los coeficientes de regresión mediante:

(36)

$$\underset{px1}{\hat{\beta}_w} = \underset{pxp}{(\boldsymbol{X'WX})^{-1}}\underset{px1}{(\boldsymbol{X'WY})}$$

2.3.3. Inferencia por mínimos cuadrados ponderados

Los procedimientos para inferencia requieren de la matriz de varianzas y covarianzas:

(37)

$$\underset{pxp}{\sigma^2\{\hat{\beta}_w\}} = k(X'WX)^{-1}$$

donde:

$$k\left(\frac{1}{\sigma^2{}_i}\right) = w_i$$

la cual se estima a partir de:

(38)

$$\underset{pxp}{s^2\{\hat{\beta}_w\}} = CME_w(X'WX)^{-1}$$

donde: ECM_w es el estimador de la constante de proporcionalidad *k* es obtenido de la siguiente manera:

(39)

$$CME_w = \frac{\sum w_i e_i^2}{n-p}$$

Los intervalos de confianza para los coeficientes de regresión se obtienen por la ecuación:

(40)

$$\hat{\beta}_w \pm t(1-\alpha/2\,;n-p)s\{\hat{\beta}_{wk}\}$$

Estos al igual que para las pruebas de hipótesis, utilizan las $s\{\hat{\beta}_{wk}\}$ obtenidas de la matriz de varianzas y covarianzas $s^2\{\hat{\beta}_w\}$.

2.4. Modelo de regresión cuantílica paramétrica multidimensional

El concepto de regresión cuantílica fue introducido por Koenker & Bassett (1978) como un método para estimar relaciones funcionales entre variables para todas las porciones de una distribución de probabilidad. Este método es particularmente útil cuando la distribución de los datos no tiene varianzas iguales, pues esto implica que existe más de una pendiente o tasa de cambio que describe la relación entre la variable respuesta y la predictora. Además prescinde de especificar cómo los cambios de varianza están vinculados con la media y no se restringe a la familia exponencial de distribuciones. La regresión cuantílica estima múltiples pendientes desde la respuesta mínima hasta la máxima, lo que provee una visión más completa de la relación entre las variables. Puede verse también como complemento de la regresión lineal ordinaria. (Cade & Noon, 2003). También es bastante robusta en casos de no normalidad del término de error. (Davino, 2014). Por último, una de sus propiedades más fácilmente advertibles es que es menos sensible a los valores extremos.

2.4.1. Definición de cuantil

Dado un número $\tau \in (0,1)$ y una variable aleatoria "Y" cuya distribución de probabilidad $F(y) = P(Y \leq y)$, el cuantil τ se define (Martínez, 2010) como:

(41) $$Q(\tau) = inf\{y: F(y) \geq \tau\}$$

Si se tiene una muestra con observaciones Y_i independientes e igualmente distribuidas, que provienen de la función de distribución $F(y)$, el cuantil (τ) de la muestra (Martínez, 2010), se define como:

(42) $$\hat{Q}(\tau) = inf\{y: \hat{F}(y) \geq \tau\}$$

donde:

$\widehat{\boldsymbol{F}}$: es la función de distribución muestral.

La solución de $\hat{Q}(\tau)$ se obtiene (Martínez, 2010) mediante la siguiente optimización:

(43)

$$\underset{\delta_\tau \, \epsilon \, \mathbb{R}}{\operatorname{argmin}} \left[\sum_{Y_i \geq \delta_\tau} \tau |Y_i - \delta_\tau| + \sum_{Y_i < \delta_\tau} (1-\tau) |Y_i - \delta_\tau| \right]$$

donde:

δ_τ representa la estimación del cuantil.

2.4.2. Modelo de regresión cuantílica

Si el concepto de cuantil es trasladado a una recta de regresión, entonces se obtiene la regresión cuantílica lineal. El modelo poblacional (Martínez, 2010) se expresa de la siguiente manera:

(44)

$$Y = \beta_{0,\tau} + \beta_{1,\tau} X_1 + \cdots + \beta_{p-1,\tau} X_{p-1} + \varepsilon_\tau$$

donde:

- $\boldsymbol{Y}$: representa la variable respuesta.
- $\boldsymbol{\beta_{0,\tau}, \beta_{1,\tau}, \ldots, \beta_{p-1,\tau}}$: los coeficientes del modelo.
- $\boldsymbol{X_1, \ldots, X_{p-1}}$: las variables explicativas.
- $\boldsymbol{\varepsilon_\tau}$: el error.
- $\boldsymbol{\tau}$: indicador del cuantil, $\tau \in (0,1)$.
- $\boldsymbol{p-1}$**:** Cantidad de variables explicativas en el modelo.

Dada una muestra aleatoria el modelo quedaría expresado de la siguiente manera $\forall i \in \{1, \ldots, n\}$ (Martínez, 2010):

(45)

$$Y_i = \beta_{0,\tau} + \beta_{1,\tau} X_{i,1} + \cdots + \beta_{p-1,\tau} X_{i,p-1} + \varepsilon_{i,\tau}$$

El modelo anterior supone que: "el valor esperado condicional no es necesariamente cero, pero que el τ-ésimo cuantil del error con respecto a la variable regresora sí es cero, $(Q_\tau(\varepsilon_{i.\tau})|X) = 0$". (Martínez, 2010, pág. 14) Por consiguiente, el τ-ésimo cuantil de Y_i con respecto a X se puede escribir como:

$$Q_\tau(Y_i|X) = \beta_{0,\tau} + \beta_{1,\tau} X_{i,1} + \cdots + \beta_{p-1,\tau} X_{i,p-1}$$

(46)

Considerando la definición anterior de cuantil, los estimadores $\hat{\beta}_{0,\tau}, \hat{\beta}_{1,\tau}, \dots, \hat{\beta}_{p-1,\tau}$ de los coeficientes se obtienen mediante la minimización de la siguiente ecuación. (Martínez, 2010)

$$\hat{\beta}_{\tau} = \underset{\beta_{\tau} \epsilon\, R^{p}}{\operatorname{argmin}} \left\{ \sum_{Y_i \geq Q_{\tau}(Y_i|X)} \tau \left| Y_i - \beta_{0,\tau} - \beta_{1,\tau} X_{i,1} - \cdots - \beta_{p-1,\tau} X_{i,p-1} \right| + \sum_{Y_i < Q_{\tau}(Y_i|X)} (1-\tau) \left| Y_i - \beta_{0,\tau} - \beta_{1,\tau} X_{i,1} - \cdots - \beta_{p-1,\tau} X_{i,p-1} \right| \right\} \tag{47}$$

donde:

$\beta_{\tau} = (\beta_{0,\tau}, \beta_{1,\tau}, \dots, \beta_{p-1,\tau})$.

Cabe destacar, además, que la minimización anterior se aborda vía programación lineal, la cual permite obtener la recta de regresión para un determinado cuantil. (Martínez, 2010). Algunos de los métodos existentes más destacados se mencionan a continuación. (Koenker R. , 2015)

- En caso de tamaños de muestra moderados: variante de la Barrodale y Roberts, algoritmo simple descrito en Koenker y D'Orey (1987).
- En caso de problemas con más de unos pocos miles de observaciones: algoritmo de Newton-Frisch descrito en Portnoy y Koenker (1997).
- Para problemas con cantidades de datos grandes: el método de observaciones plausiblemente intercambiables que implementa una versión del algoritmo de Newton-Frisch.

2.4.3. Inferencia en el modelo de regresión cuantílica múltiple

La normalidad asintótica para $\hat{\beta}_{\tau}$ fue establecida por Koenker y Bassett (1978) bajo el supuesto de errores independientes e idénticamente distribuidos (iid). En un modelo con errores independientes e idénticamente distribuidos (Kocherginsky, He, & Mu, 2005), la matriz de varianzas y covarianzas para $\hat{\beta}_{\tau}$ se obtiene mediante:

$$V_{\tau} = (\tau(1-\tau)/f^{2}(0))\ (X'X)^{-1} \tag{48}$$

donde:

- $f(0)$ es la densidad del error común evaluada en cero.

La matriz V_τ bajo errores (iid) corresponde de manera muy cercana al comportamiento del estimador de mínimos cuadrados si se tiene en cuenta que $w^2 = (\tau(1-\tau)/f^2(0))$ y w^2 es remplazado por σ^2 la varianza de la distribución del error (Koenker & Hallock, 2000).

Es importante rescatar que para el caso anterior se ha demostrado que los resultados obtenidos no son robustos ante la presencia de errores heterocedásticos. Por esta razón es mejor recurrir al cálculo de la matriz de varianzas y covarianzas asumiendo que los errores $\varepsilon_{i,\tau}$ pueden no ser independientes o idénticamente distribuidos y tienen funciones de densidad f_i:

(49) $$V_\tau = \big(\tau(1-\tau)\big)(X'FX)^{-1}(X'X)(X'FX)^{-1}$$

donde:

- **X**: es una matriz n x p cuyas columnas son x_i
- **F**: diagonal$\{f_1(0), \dots, f_n(0)\}$

La matriz V_τ bajo errores no independientes e idénticamente distribuidos también puede ser entendida como una versión de la fórmula (sandwhich) Huber-Eicker- White para limitar la matriz de varianzas y covarianzas (Koenker & Hallock, 2000).

En la regresión cuantílica existen varios métodos para estimar los intervalos de confianza para β_τ, entre ellos: dispersión (sparcity), rango y re-muestreo (bootstrap) (Koenker R., 1994). El método de inversión rango descrito en Koenker (1994) es frecuentemente el más utilizado y puede ser calculado bajo errores iid o independientes pero no idénticamente distribuidos (inid). (Koenker R. , 2016)

El método de puntuación de rango (rank score) evita la estimación directa de la matriz de covarianzas asintótica de los coeficientes estimados. Surge naturalmente de las técnicas de programación lineal usadas para encontrar las estimaciones de los coeficientes de regresión cuantílica (Tarr, 2011). Koenker y Machado (1999) no solamente admiten errores (iid) sino que también permiten aquellos que son (inid) y consideran el siguiente modelo (location-scale):

(50) $$y_i = x_i^T\beta + \sigma_i u_i$$

donde $\sigma_i = x_i^T \gamma$ y los u_i son asumidos (iid) con función de distribución F.

Bajo el método de inversión de rango, las estimaciones de los intervalos de confianza para un parámetro pueden ser halladas mediante el proceso de inversión de la prueba estadística apropiada moviéndose de un pivote simplex al siguiente para obtener un intervalo. En este, la prueba estadística es aquella cuya hipótesis H_0: $\beta_{j,\tau} = b$ no es rechazada. El intervalo de confianza resultante no es necesariamente simétrico (Tarr, 2011). Para mayor detalle sobre el método bajo errores independientes e idénticamente distribuidos se puede consultar Koenker (1994), mientras que para errores heterocedásticos se puede revisar Koenker y Machado. (1999)

2.5. Modelo lineal generalizado con dispersión variable

2.5.1. Modelos lineales generalizados

Los modelos lineales generalizados extienden la regresión lineal ordinaria para abarcar distribuciones de respuesta no normales y posiblemente no lineales de la media. Tienen tres componentes. (Agresti, 2015)

I. **Componente aleatorio**. En donde se especifica la variable respuesta y su distribución de probabilidad. Además, las observaciones Y_i con $(i = 1, \dots, n)$ se tratan como independientes. Estas comparten la misma distribución dentro de la familia exponencial. En el caso de la regresión lineal múltiple con errores normales:

(51) $$Y_i \sim \boldsymbol{NID}(\mu_i, \sigma^2)$$

II. **Predictor lineal**: Para el vector de parámetros **β**= $(\beta_1, \beta_2, \dots, \beta_{p-1})^T$ y la matriz de modelo $\boldsymbol{X}$ que contiene valores de ***p-1*** variables explicativas para las ***n*** observaciones, el predictor lineal $\boldsymbol{\eta}$ en forma matricial es $\boldsymbol{X\beta}$. El predictor lineal de un modelo lineal generalizado relaciona los parámetros $\boldsymbol{\eta_i}$ pertinentes a $\boldsymbol{E}(Y|\boldsymbol{X_i})$ con las variables explicativas utilizando una combinación lineal de ellas:

(52)
$$\eta_i = \sum_{j=0}^{p-1} \beta_j x_{ij}, \qquad i = 1, \dots, n \ \ donde\ x_{10} \equiv 1$$

Es preciso recordar que en el modelo de regresión lineal múltiple:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

III. **Función de enlace**: Esta es una función que relaciona el componente aleatorio con el predictor lineal. Si $\mu_i = E(Y|\boldsymbol{X_i})$ los modelos lineales generalizados enlazan a η_i con μ_i mediante $g(\mu_i) = \eta_i$, donde la función de enlace $g(\cdot)$ es monotónica y derivable.

Mediante el uso de la función de enlace, el modelo se puede expresar como:

(53)
$$g(\mu_i) = \sum_{j=0}^{p-1} \beta_j x_{ij} \ \{con\ \ i = 1, \dots, n\ \ y\ donde\ x_{i0} \equiv 1\}$$

Para la distribución normal, si se hace $\mu_i = \eta_i$, la función $\boldsymbol{g}(\mu_i) = \mu_i$ y esta es conocida como función de enlace identidad, por lo tanto:

(54)
$$\mu_i = \sum_{j=0}^{p-1} \beta_j x_{ij}, \qquad i = 1, \dots, n \ \ donde\ x_{i0} \equiv 1$$

Lo anterior se conoce como *modelo lineal* pues iguala el predictor lineal a la media y asume varianza constante de las observaciones. Una forma alternativa de expresarlo es la siguiente:

(55)
$$Y_i = \sum_{j=0}^{p-1} \beta_j x_{ij} + \varepsilon_i$$

donde:

- $x_{i0} \equiv 1$.
- ε tiene $E\{\varepsilon\} = 0$ y $\text{Var}\{\varepsilon\} = \sigma^2$, lo cual es natural para la función de enlace identidad y respuesta normal, pero no para todos los modelos lineales generalizados.

2.5.2. Familia de dispersión exponencial para el componente aleatorio

Si las observaciones de la variable respuesta, provienen de la familia exponencial, poseen función de densidad de probabilidad (Agresti, 2015):

(56)
$$f(Y_i) = exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi)\right\}$$

donde:

- θ_i es el parámetro natural.
- ϕ es el parámetro de dispersión.
- $a(\phi)$, $b(\theta_i)$ y $c(Y_i, \phi)$ son funciones conocidas.

Se ha demostrado que la media y la varianza de la distribución (Agresti, 2015) están dadas por

(57)
$$Var(Y|X_i) = b''(\theta_i)a(\phi)$$

donde:

- $a(\phi) = \frac{\phi}{w_i}$.
- w_i es una ponderación conocida.

La relación existente entre la media y la varianza (Smyth, 1989) está dada por:

(58)
$$Var(Y|\boldsymbol{X_i}) = \sigma_i^2 = V(\mu_i)\phi w_i^{-1}$$

donde:

- $\boldsymbol{V(\mu_i)} = b''(\theta_i) = b''(b'^{-1}(\mu_i))$ pues $\theta_i = b'^{-1}(\mu_i)$.
- V es un escalar, no negativo cuya función de varianza está determinada por la distribución de probabilidad, los w_i son ponderaciones conocidas y ϕ es un parámetro de dispersión.

En el caso específico de la ***distribución normal***, la función de densidad se expresa como sigue (Agresti, 2015):

(59)

$$f(Y_i) = exp\left\{\frac{Y_i\mu_i - \frac{{\mu_i}^2}{2}}{\sigma^2} - \frac{{Y_i}^2}{2\sigma^2} - \frac{1}{2}\ln 2\pi\sigma^2\right\}$$

Donde para el modelo homocedástico:

- $\theta_i = \mu_i$ parámetro natural.
- $\phi = \sigma^2$.
- $a(\phi) = \frac{\phi}{w_i}$ en este caso $w_i = 1$ por lo que $a(\phi) = \phi$.
- $b(\theta_i) = \frac{{\mu_i}^2}{2}$.
- $c(Y_i, \phi) = -\frac{1}{2}\left[\frac{Y^2}{\phi^2} + \ln(2\pi\phi)\right]$.

Además:

(60)

$$E(Y|\boldsymbol{X_i}) = b'(\theta_i) = \theta_i = \mu_i$$

(61)

$$Var(Y|\boldsymbol{X_i}) = b''(\theta_i)a(\phi) = 1 * a(\phi) = \phi = \sigma^2$$

2.5.3. Modelo con predictor lineal para la dispersión

Los modelos lineales generalizados pueden además incluir un predictor lineal para la dispersión $Var(Y|\boldsymbol{X_i})$ así como para la media $E(Y|\boldsymbol{X_i})$ y ser llevados a un enfoque más general, configurando la estructura de la media y la dispersión, por separado, para casos en donde exista evidencia de **heterocedasticidad**.

Esto se logra si en lugar de especificar la varianza como función de la media y una constante multiplicativa (parámetro de dispersión ϕ) en la ecuación $Var(Y|\boldsymbol{X_i}) = V(\mu_i)\phi w_i^{-1}$, se le permite a este último depender de las covariables y parámetros desconocidos de la misma forma que a la media (Smyth, 1989), se puede entonces generalizar a:

(62)

$$Var(Y|\boldsymbol{X_i}) = \sigma_i^2 = V(\mu_i)\phi_i w_i^{-1}$$

En la cual las dispersiones ϕ_i pueden ser modeladas por:

(63) $$h(\phi_i) = z_i'\gamma$$

donde:

- $h(\phi_i)$ es otra función de enlace.
- z_i es un vector de covariables.
- γ otro vector de parámetros desconocidos.

2.5.4. Método de estimación

Como la media y la dispersión son ortogonales se puede estimar $\boldsymbol{\beta}$ y $\boldsymbol{\gamma}$ de forma independiente, mediante el método de máxima verosimilitud. En este tipo de modelos esto se cumple para las distribuciones: Normal, Gamma y Normal inversa (Smyth, 1989).

Si las observaciones (y_i, x_i, z_i) con i=1,…, n; siguen el modelo $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ con $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Omega})$ y para el parámetro de dispersión $\phi_i = \sigma_i^2$ se utiliza la función de enlace logarítmica $h() = log$ de manera que $log(\sigma_i^2) = z_i'\gamma$. La función de máxima verosimilitud para los parámetros "***β" y "γ"*** es (Cepeda & Gamerman, 2000):

(64) $$L(\beta, \gamma) = \prod_{i=1}^{n} \frac{1}{\sigma_i} exp\left\{\frac{-1}{2\sigma_i^2}(Y_i - X_i'\beta)^2\right\}$$

En muestras pequeñas, las estimaciones realizadas por máxima verosimilitud se pueden desempeñar pobremente, por lo que se sugiere utilizar máxima verosimilitud restringida para obtener estimadores de la varianza con mayor precisión (Western & Bloome, 2009).

2.5.5. Inferencia en los modelos lineales generalizados de dispersión variable

La matriz de información observada completa obtenida de la función de máxima verosimilitud para el vector de parámetros (β, γ) está dada por (Cepeda & Gamerman, 2000):

(65) $$J = \begin{bmatrix} X'W_{11}X & X'W_{12}Z \\ Z'W_{12}X & Z'W_{12}Z \end{bmatrix}$$

donde:

- $Z' = (z_1, \ldots, z_n)$
- $W_{11} = diag(1/\hat{o}_i^2)$

- $W_{12} = diag(\hat{e}_i/\hat{\sigma}_i^2)$
- $W_{22} = \frac{1}{2} diag(\hat{e}_i/\hat{\sigma}_i^2)$
- $\hat{e}_i = Y_i - x_i'\hat{\beta}$
- $\hat{\sigma}_i^2 = h^{-1}(z_i'\hat{\gamma})$
- Además $\hat{\beta}$ y $\hat{\gamma}$ son los estimadores máximos verosímiles de β , γ.

A la esperanza matemática de la matriz J se le conoce como matriz de información esperada estimada $\hat{I} = E(J)$. (Cepeda & Gamerman, 2000)

(66)

$$\hat{I} = \begin{bmatrix} X'W_{11}X & 0 \\ 0 & \frac{1}{2}Z'Z \end{bmatrix}$$

El algoritmo de puntuación de Fisher iterado hasta la convergencia simultáneamente provee las estimaciones máximo verosímiles $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ para $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ y también para $\hat{\mathbf{I}}$. La teoría asintótica proporciona la aproximación de las distribuciones de los estimadores de máxima verosimilitud (Cepeda & Gamerman, 2000), la cual es:

(67)

$$\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix} \sim N \begin{bmatrix} \beta \\ \gamma \end{bmatrix}, \quad \hat{I}^{-1}$$

Esta distribución aproximada es utilizada para construir intervalos de confianza o regiones para funciones de (β, γ). Por ejemplo: los intervalos de confianza **100(1-α) %** para $\boldsymbol{\beta_k}$ tienen límites $\widehat{\boldsymbol{\beta}}_{\mathbf{k}} \pm \mathbf{z}_{\boldsymbol{\alpha/2}}\mathbf{i}^{\mathbf{kk}}$ donde $\mathbf{i}^{\mathbf{kk}}$ es el k-ésimo elemento de la diagonal de la matriz $\hat{\mathbf{I}}^{-\mathbf{1}}$ **.** Similarmente se construyen los intervalos de confianza para cualquiera de los elementos de $\boldsymbol{\gamma}$ (Cepeda & Gamerman, 2000). Por último, las pruebas de hipótesis de t calculadas de la manera usual también requieren de los elementos $\mathbf{i}^{\mathbf{kk}}$.

2.6. Aproximación Bayesiana al problema de heterocedasticidad en el modelo de regresión lineal múltiple

En el modelo de regresión lineal múltiple Bayesiano se producen distribuciones a posteriori de los parámetros teniendo en consideración tanto los datos como las densidades a priori de los parámetros. De esta forma θ es el parámetro de interés y $\pi(\theta)$ es la distribución a priori de los

parámetros, que resume la incertidumbre sobre ellos. Una vez obtenidos los datos, la distribución a posteriori de los parámetros, según el paradigma Bayesiano (Correa, 2005), viene dada por:

(68) $$\pi(\theta|Datos) \propto \pi(\theta)L(\theta|Datos)$$

donde:

- $\propto$ es el símbolo de proporcionalidad, que indica que la densidad se divide por una constante para que sea una densidad propia.
- $L()$ es la función de verosimilitud.

Retomando el modelo frecuentista donde la media, $E(Y|X_i)$, y la varianza, σ_i^2 , pueden ser escritas en función de las covariables, para todo i: $(i = 1, .., n)$ (Western & Bloome, 2009):

(69) $$E(Y|X_i) = x_i'\beta$$
$$log(\sigma_i^2) = z_i'\gamma$$

donde:

- x_i' es un vector de covariables para estimar la media y z_i' para estimar la varianza.

El enfoque bayesiano puede ser ventajoso respecto a su contraparte frecuentista calculada mediante máxima verosimilitud o bien máxima verosimilitud restringida. En muestras pequeñas, los coeficientes $\boldsymbol{\gamma}$ de la ecuación de varianza pueden estar sesgados y la inferencia basada en una distribución normal sería inapropiada. La no normalidad de la distribución vendría a reflejarse en la distribución a posteriori de los datos y con esto se corregiría la inferencia. (Western & Bloome, 2009)

2.6.1. Modelo lineal generalizado Bayesiano con dispersión variable

La *perspectiva Bayesiana* combina una verosimilitud normal para y_i, junto con una distribución previa para los coeficientes $\boldsymbol{\beta}$ y una previa jerárquica para los coeficientes de la varianza $\boldsymbol{\gamma}$**.** Por consiguiente, para la variable dependiente y_i con predictores $\boldsymbol{x_i'}$ para la media y $\boldsymbol{z_i'}$ para la varianza, el modelo Bayesiano (Western & Bloome, 2009) se escribe como sigue:

(70) $$y_i \sim N(E(Y|X_i), \sigma_i^2)$$

$$E(Y|X_i) = x_i'\beta$$

$$log(\sigma_i^2) = z_i'\gamma$$

Cabe destacar que lo anterior define la verosimilitud del mismo. Mientras que las distribuciones a priori de los parámetros estarían dadas por:

(71)

$$\beta \sim N(b, V)$$

$$\gamma \sim N(g, U)$$

$$U_{jj} \sim Gamma^{-1}(u_0, u_1)$$

- β y γ tienen una distribución a priori no informativa (en este caso: una distribución normal) y se establece la media para los vectores b y g en cero.
- La matriz diagonal de varianzas y covarianzas $\boldsymbol{V}$ se especifica con varianzas grandes (por ejemplo: 10^6).
- Para asegurar que los datos en la muestra dominen la estimación de los coeficientes de la varianza, $\boldsymbol{\gamma}$ recibe una distribución a priori jerárquica cuya matriz diagonal $\boldsymbol{JxJ}$ de varianzas y covarianzas $\boldsymbol{U}$, sigue una distribución a priori gamma inversa con hiper-parámetros (u_0, u_1) pequeños (dígase : $u_0 = 0.001$ y $u_1 = 0.001$). (Western & Bloome, 2009)

Cuadro # 1 Modelo Bayesiano lineal doble generalizado distribuciones de probabilidad a priori

Distribuciones a priori			
Parámetro	**Distribución a priori**	**Hiperparámetros**	
$\boldsymbol{\beta}$	$\boldsymbol{N(b, V)}$	**b=0**	$\boldsymbol{V_{ii} = 10^{-6}}$
$\boldsymbol{\gamma}$	$\boldsymbol{N(g, U)}$	**g=0**	$\boldsymbol{U_{jj} \sim Gamma^{-1}(u_0, u_1)}$
	$\boldsymbol{Gamma^{-1}(u_0, u_1)}$	$\boldsymbol{u_0 = 0.001}$	$\boldsymbol{u_1 = 0.001}$

Fuente: Elaboración propia.

2.6.2. Inferencia en el modelo Bayesiano

Para realizar inferencia sobre los parámetros, en algunos casos resumir descriptivamente la distribución a posteriori puede ser difícil o imposible. Como solución al problema, es posible utilizar un algoritmo de cómputo para muestrear observaciones de la distribución posterior conjunta. Si se tiene una muestra suficientemente grande de la distribución a posteriori conjunta, entonces la distribución de dicha muestra debe aproximarse muy bien a la antedicha. Esto implica que se puede resumir la muestra para poder realizar inferencias de los parámetros en vez de utilizar la distribución a posteriori directamente.

Algunos de los algoritmos que hacen posible este muestreo (Johnson, 2012) son los siguientes:

- Muestreo de Gibbs (generalmente utilizado con distribuciones a priori conjugadas).
- Algoritmo Metrópolis-Hastings (generalmente utilizado cuando no se pueden conjugar o hallar las distribuciones a priori que faciliten el muestreo de la distribución a posteriori).

Todos estos algoritmos son parte de los llamados Cadenas de Markov Monte Carlo, métodos conocidos por sus siglas en inglés como MCMC: Markov Chain Monte Carlo.

Los resúmenes de la distribución a posteriori desempeñan un papel importante en el análisis Bayesiano. Los métodos estadísticos frecuentistas ofrecen en sus informes estimaciones de los parámetros y los errores estándar. El análogo Bayesiano de esta práctica es reportar los momentos de las distribuciones marginales de parámetros tales como la media y la desviación estándar de la distribución a posteriori. (Rossi, Allenby, & McCulloch, 2006)

Los intervalos de confianza Bayesianos son llamados intervalos de credibilidad, estos en vez de la cuidadosa interpretación clásica en términos del hipotético muestreo repetido, tienen una interpretación natural pues en un intervalo $[a, b]$ la probabilidad a posteriori de que dicho intervalo contenga al parámetro es $(1 - \alpha)$ (Rencher & Schaalje, 2008).

Existen dos formas popularmente conocidas de hallar los intervalos de credibilidad según Gelman, Carlin y Stern (2014):

1.1 Alta densidad posterior (HPD por sus siglas en inglés: highest posterior density) donde el intervalo representa el intervalo más pequeño con probabilidad (1-α).

2.1 Colas de igual probabilidad, el cual utiliza los valores (cuantiles de la distribución a posteriori) que acumulan la probabilidad $(\alpha/2)$ y $(1-\alpha/2)$ como sus límites. Se prefiere no utilizarlos cuando la distribución a posteriori es sesgada o multimodal.

2.7. Marco teórico para el caso de aplicación

2.7.1. Antecedentes

En 1857 Ernst Engel publicó un artículo sobre las condiciones de consumo en donde formulaba una ley empírica concerniente a la relación entre Ingreso y Gasto en alimentación de los hogares (lo que después se conocería como la ley de Engel). Esta mantiene que la proporción del ingreso que se invierte en alimentación (gasto en alimentación), disminuye conforme aumenta el ingreso del hogar, aún cuando es probable que el gasto aumente en términos absolutos. Desde esa fecha la ley de Engel ha sido corroborada en muchas otras encuestas de ingresos y gastos de los hogares (Houthakker, 1957).

La **heterocedasticidad** en la relación del gasto en alimentación e ingreso, puede surgir porque:

- Las familias pobres tienen menos flexibilidad en su nivel de gasto en alimentación, por lo que existe poca dispersión en su gasto de alimentación. Por el contrario, cabría esperar que algunas familias ricas gasten mucho en alimentación mientras que otras familias ricas con preferencias diferentes gasten mucho menos en alimentación, destinando su renta a otros usos. (Borrego & Carro, 2012)
- Las familias con diferentes ingresos tienden a gastar en diferentes tipos de alimentos y, más importante aún, el gusto difiere entre y dentro de las familias. Así pues, la calidad de la alimentación también varia con el ingreso (Mhlongo & Daniels, 2013).
- Generalmente la variabilidad incrementa a la par del ingreso (Kindleberger, 1997), lo que puede interpretarse como que la diferencia entre la observación real y el gasto estimado tiende a aumentar con el ingreso.

2.7.2. El modelo empírico

La técnica de regresión lineal múltiple con errores normales se plantea con la finalidad de relacionar la cantidad de dinero que los hogares costarricenses gastan en alimentos y bebidas no alcohólicas como variable dependiente del ingreso de los hogares y otras variables acorde con la teoría económica y evidencia empírica pertinente.

En la postulación de Engel, la teoría tradicional de la demanda del consumidor no incluye características socio-económicas. Comúnmente, esta teoría económica se basa en el supuesto de que el consumidor maximiza la utilidad de los bienes comprados en el mercado en un único período y sujeto a la restricción del ingreso. La literatura empírica presenta evidencia muy fuerte que apoya la teoría de que el ingreso es uno de los principales determinantes de los gastos en alimentación del hogar. No obstante, en diversos estudios se ha evidenciado que también otras características socio-económicas de los hogares son importantes determinantes del gasto alimentario. (Davis, 1982)

En específico las variables socio-económicas utilizadas en la predicción del gasto para los estudios consultados son numerosas, como se detallan a continuación:

- El nivel educativo del jefe de hogar, tamaño del hogar, la zona de residencia, etapa del ciclo de vida del jefe (edad del jefe), la composición de edad-raza-sexo del hogar (Davis, 1982).
- Número de personas del hogar, sexo del jefe, ciclo de vida de los miembros del hogar (cantidad de miembros según grupos de edades), nivel educativo del jefe y la ciudad de residencia (Muñoz, 2004).
- Tamaño de la familia, estacionalidad, región, religión, raza y características del jefe de familia (Villezca, 2005).
- Edad del jefe del hogar, estado civil del jefe de hogar, años de educación formal del jefe de hogar, tamaño del hogar, composición del hogar con respecto a dependientes y grupos de riesgo (infantes y mujeres embarazadas) (Babalola & Isitor, 2014).
- Tamaño del hogar, zona de residencia y región de planificación (Geurts, Jansen, & Tilburg, 1997).

- Educación, género, condiciones étnicas y la zona de residencia de la persona, además de la proporción de miembros mayores de 55 años, la proporción de miembros menores de 18 años y la cantidad de miembros del hogar (Zúñiga, Saborío, Ulate, Linares, & Hernández, 2004).
- Tamaño del hogar y zona de residencia (Vargas & Elizondo, 2015).

Es importante rescatar que todas estas variables socio-económicas se consideran potencialmente importantes en el comportamiento de consumo alimenticio, pero su trascendencia en un modelo de regresión múltiple está sujeta a factores como la disponibilidad de la información, la evidencia estadística que demuestren en relación con la variable respuesta, características (económicas-sociales- culturales) propias de la muestra o población en estudio, e inclusive la finalidad de la investigación y el criterio del investigador.

No existe un consenso sobre la forma funcional de estimar la relación entre el gasto en alimentación y el ingreso u otras variables. Sin embargo, gracias a la literatura consultada, se han logrado evidenciar ciertas tendencias. Los autores recomiendan que el modelo debe poseer un balance entre el pragmatismo estadístico y la teoría económica. Algunos ejemplos de esto último:

- Brady y Barber (1948) utilizan una segmentación por ingresos para contra-restar el efecto de la variabilidad en las curvas del gasto y la cantidad de miembros del hogar. Para ello, en cada grupo de ingreso realizan regresiones de mínimos cuadrados con transformación doble logarítmica.
- Stuvel y James (1950) relacionan la forma en que el gasto en alimentación se ve afectado por el tamaño del hogar y el ingreso mediante dos modelos: el primero de ellos de mínimos cuadrados ordinarios y el segundo doble-logarítmico.Pero antes de realizar los modelos controlan la variabilidad por la ocupación del jefe, la región y el vecindario del hogar. Destacan que el modelo doble-logarítmico es de esencial interés cuando se quiere hablar de elasticidades.
- Davis (1982) se inclina por una relación logarítmica doble, donde se especifica el gasto en alimentación, el ingreso y el tamaño del hogar de manera logarítmica y las demás variables sin esta transformación. Se utilizan los mínimos cuadrados ordinarios y los coeficientes del ingreso y el tamaño del hogar con el objeto de interpretar elasticidades. En el modelo no se evalúan los supuestos clásicos de los mínimos cuadrados ordinarios, por lo que no se menciona la igualdad

de varianzas del error. Sí se menciona que este modelo es una alternativa importante a la transformación de Box-Cox, por lo que se puede estar asumiendo que soluciona la falta de normalidad o bien **heterocedasticidad** (Neter et al. (1990).

- En el análisis de Geurts et al. (1997) para predecir el gasto de alimentación total se utiliza el modelo de regresión de mínimos cuadrados ordinarios para demostrar la dependencia del gasto de la cantidad de miembros del hogar y las variables geográficas. El modelo no posee transformaciones y no se evidenció **heterocedasticidad** en las pruebas de White y Glejser, lo cual también ayuda a evidenciar que la variable Ingreso posee buena parte de la responsabilidad en la desigualdad de varianzas del error.

- Dawoud (2013) por su parte analiza los cambios del gasto en alimentación en el tiempo para Egipto, con especial énfasis entre las zonas de residencia urbana y rural. Para ello, estima las curvas de Engel mediante el modelo doble-logarítmico utilizando el método de mínimos cuadrados ponderados. Explica que utiliza esta técnica debido a la forma en que trataron (obteniendo y agrupando) los datos. Acota además que los mínimos cuadrados ponderados les dan más importancia a las observaciones asociadas con clases de ingreso en proporciones mayores de población, mientras que los mínimos cuadrados en este caso tratarían las observaciones por igual.

- Babalola e Isitor (2014), en su análisis de determinantes de los patrones de consumo alimenticio, estiman también la relación entre el gasto en alimentación y sus diferentes variables predictoras mediante un modelo doble logarítmico de mínimos cuadrados (escogido entre varias especificaciones). Para ello se utilizaron los criterios del coeficiente de determinación y la teoría económica pertinente.

- Por su parte, Vargas y Elizondo (2015) estiman elasticidades utilizando las curvas de Engel mediante la estimación por mínimos cuadrados. Como variable respuesta recurren al gasto corriente per cápita en el alimento (x) y como variables predictoras utilizan el precio del alimento, el índice de precios de consumidor, el ingreso corriente per cápita mensual del hogar sin valor locativo, la zona de residencia y la cantidad de miembros en el hogar. Cabe destacar que tanto variables respuestas como predictoras fueron transformadas mediante el logaritmo natural a excepción de la zona de residencia. Se afirma que para llegar al modelo se ensayaron varias especificaciones conceptuales y que los resultados son robustos.

CAPTULO 3: MARCO METODOLÓGICO

Para determinar si los diferentes modelos de regresión anteriormente detallados estiman adecuadamente el parámetro de interés y ofrecen la menor inflación de su error estándar, es necesario definir un modelo teórico con parámetros hipotéticos fijos, además de contar con un conjunto de datos y un respectivo error.

Por ello, mediante el uso de simulación Montecarlo, se generan conjuntos de datos en donde para cada uno se define la variable respuesta en términos de la adición de variables independientes pseudoaleatorias. De un error también pseudoaleatorio de media cero, cuya varianza se especifica **heterocedástica** procedente de los orígenes bajo estudio, se definen: la naturaleza teórica de las variables, la omisión de una variable independiente de importancia en el modelo para explicar el comportamiento de la variable respuesta y ,finalmente, la presencia de valores extremos en las variables independientes.

Teniendo certeza de que el error posee varianza no constante, se estiman entonces los parámetros del modelo teórico mediante el método de mínimos cuadrados ordinarios y los diferentes modelos estadísticos que contemplan heteroscedasticidad o que reducirían sus efectos en las estimaciones y los errores estándar correspondientes.

Posteriormente se repite el proceso de generación de las variables pseudoaleatorias y el error un gran número de veces manteniendo los parámetros hipotéticos constantes. Se sintetiza el desempeño individual de cada modelo para el parámetro de interés gracias a medidas para evaluar el sesgo, la precisión y exactitud de la estimación y su respectivo error estándar, evaluando la tendencia central, dispersión y, en general, su distribución.

Además se considera importante mencionar que el estudio empírico se efectúa previo al de simulación, lo que también facilita el entendimiento del comportamiento heterocedástico residual y por esta razón los valores de los coeficientes obtenidos en los modelos empíricos se utilizan para diseñar los coeficientes de los modelos hipotéticos simulados.

Todo lo anterior se realiza utilizando el software estadístico R 3.5.1 y paquetes programados para este entorno, tales como: dglm, qreg, quantreg, R2jags, entre otros típicos del análisis de regresión. En el caso del paquete R2jags, se aclara que el mismo permite la conexión entre el entorno R y JAGS 4.3.0 previamente instalado en el ordenador.

3.1 Simulaciones Monte Carlo

El uso de números aleatorios en estadística se ha expandido más allá del muestreo aleatorio o la asignación aleatoria de tratamientos a las unidades experimentales. Actualmente, los usos más comunes están en estudios de simulación de procesos estocásticos, expresiones matemáticas analíticamente complejas o el estudio de una población por re-muestreo de una muestra perteneciente a esa población. Estas tres áreas generales de aplicación se conocen comúnmente como: simulación estocástica, método de Monte Carlo y re-muestreo. (Gentle, 2003, pág. 1)

En el presente estudio se utiliza el método de Monte Carlo. Es importante mencionar que para este método no se consensa una definición única; sin embargo, las definiciones existentes tienen en común el uso del muestreo aleatorio para calcular el resultado. Los algoritmos utilizados para el proceso de muestreo se basan en números pseudo-aleatorios generados por computadora que imitan números aleatorios verdaderos para generar un posible resultado de un proceso. No todos los resultados tienen por qué ser igualmente probables. Repitiendo el procedimiento con diferentes números aleatorios como entrada se pueden reunir los datos correspondientes al proceso de modelado. Con esta información se consigue entonces realizar un análisis estadístico para responder diferentes preguntas sobre el proceso. (Hellander, 2009)

Se consideran otras características del método de Monte Carlo:

- Las observaciones en el método Monte Carlo, como regla, son independientes en el tiempo por lo que no poseen correlación serial.
- En el método de Monte Carlo, es posible expresar la variable respuesta como función bastante simple de las variables estocásticas de entrada (Reuven, 1981, pág. 12).

Para comprobar la robustez de los diferentes enfoques se estudia, para las posibles causantes de **heterocedasticidad,** cuánto cambian las estimaciones de los coeficientes respecto a los parámetros poblacionales provenientes de un modelo poblacional hipotético y el tamaño de los respectivos errores estándar, tomando en cuenta:

I. La naturaleza de las variables predictoras.

II. La presencia de valores extremos.

III. El sesgo de variable omitida.

Para ello se aplica el método Monte Carlo, que consta (Tafalla, 2014) de los siguientes pasos:

a. Aplicar el proceso generador de datos (PGD) para un modelo de regresión múltiple, lo que implica especificar la distribución del término de error, la distribución de las variables explicativas, los valores de los coeficientes del modelo y el tamaño de la muestra. Además, permite inducir la **heterocedasticidad** según sea el escenario (I, II, III) sobre el término de error.

b. Generar un conjunto de datos usando este PGD.

c. Calcular para el conjunto de datos el modelo de mínimos cuadrados y los modelos para abordar la violación del supuesto de igualdad de varianzas del error.

d. Repetir las etapas b y c un elevado número de veces. En este estudio se repiten 10 000 veces para cada escenario.

3.2 Metodología de las simulaciones

Al aplicar el método Monte Carlo, este estudio considera un tamaño de muestra grande. Para generar las variables explicativas del modelo de regresión a partir valores aleatorios, se utiliza la Distribución Uniforme que supone igualmente probable cualquier valor dentro del intervalo **[a; b]**. La variable respuesta, por lo tanto, es el resultante de la sumarización de las mismas y el término de error proveniente de una distribución normal.

(72) $$\varepsilon_i \sim N(0, \sigma_i)$$

A partir de estas variables y las características propias de cada escenario se calculan los modelos:

1. Mínimos cuadrados ordinarios.
2. Mínimos cuadrados ponderados.
3. Modelos lineales doble generalizados.
4. Regresión cuantílica.
5. Modelos lineales doble generalizados bayesianos.

El proceso se repite para los modelos del 1 al 4 en cada escenario una cantidad de 10 000 veces, por lo que en total se ejecutan 30 000 corridas de simulación Monte Carlo para estos modelos,

generalmente cada uno de ellos toma un tiempo de estimación de fracciones de segundo como se aprecia en la celda tiempo transcurrido en segundos en el cuadro (**Cuadro #2**) obtenido mediante

el uso del paquete rbenchmark. Para el caso específico del modelo bayesiano, cada modelo toma un tiempo de estimación cercano a una hora (53 minutos) lo que puede variar bajo las especificaciones del equipo utilizado.

Cuadro # 2 Tiempos de estimación de los modelos

Tiempos de estimación de los modelos (segundos)					
Modelos	**Replicas**	**Tiempo transcurrido en segundos (elapsed)**	**Relativo**	**Tiempo de proceso en segundos (user.self)**	**Tiempo empleado por sistema (sys.self)**
Mod_DGLMB	1	3178,06	1	3170,03	0,79
Mod_DGLM	1	0,1	1	0,09	0
Mod_MCP	1	0,02	1	0,01	0
Mod_MCO	1	0,02	1	0,01	0
Mod_QREG	1	0,01	1	0,01	0

Fuente: Elaboración propia, mediante uso del paquete "rbenchmark".

Por lo que se realizan 200 simulaciones para cada escenario y, a lo interno de cada una, 6500 iteraciones del método de Markov, estas últimas se componen así:

1. BI= Pasos de Burn-In = 1500.
2. NC= Número de cadenas = 3.
3. A= Adelgazamiento (Thin-steps) = 5.
4. NPG= Número de pasos guardados = 3000.
5. Cantidad de iteraciones = (BI + (NPG * A) /NC) = 6500.

3.3 Simulación de heterocedasticidad proveniente de la naturaleza de las variables explicativas

En el caso de la **heterocedasticidad** generada por la naturaleza de las variables explicativas, para generar el conjunto de datos se decide simular utilizando el modelo doble generalizado. Las variables independientes pseudoaleatorias fueron creadas mediante la Distribución Uniforme.

La ecuación de la varianza del error residual se genera mediante:

$$\sigma_i^2 = e^{(z_i'\gamma)} \tag{73}$$

Utilizando en el cálculo del mismo las mismas variables pseudoaleatorias que para el modelo de la media. Para simular un error heteroscedástico con distribución Gaussiana se obtiene la raíz cuadrada de la expresión anterior de varianza (σ_i^2) y se generan n valores aleatorios provenientes de una distribución normal según la ecuación (73).

Posteriormente se calcula la variable dependiente tal y como se describe en la ecuación (30).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

3.4 Simulación de heterocedasticidad proveniente del sesgo de variable omitida

Para este escenario los conjuntos de datos se generan mediante una ***variable omitida*** obtenida mediante una variable con Distribución Uniforme $[a, b]$ ponderada por $\frac{1}{2}$ a la cual se le agrega otra variable generada a partir de una Distribución Uniforme $[a, b]$ con el mismo intervalo que la primera y también ponderada por $\frac{1}{2}$, esto con el objetivo de forzar la correlación de la variable omitida con la que vaya a estar presente cuando se estime el modelo.

Variable omitida:

(74)
$$Xomitida_{i1} = \frac{1}{2} * Xuniforme_{i1} + \frac{1}{2} * Xuniforme_{i2}$$

El término de error se obtiene a partir de una distribución normal con media igual a cero y desviación estándar relacionada con la variable omitida:

(75)
$$\varepsilon_i \sim N(0, (Xomitida))$$

Finalmente, se obtiene la variable dependiente.

(76)
$$Y_i = \beta_0 + \beta_1 Xomitida_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

3.5 Simulación de heterocedasticidad proveniente de la presencia de valores extremos

Para simular **heterocedasticidad** proveniente de la presencia de valores extremos y generar los conjuntos de datos se utilizan Distribuciones Uniformes para generar las variables independientes. A continuación, se ordena de menor a mayor valor una de las variables independientes (X_{i1}) y se contamina la misma, conservando el 99% de los valores originales mientras que la parte restante, el 1%, se selecciona mediante valores aleatorios provenientes de una distribución de Bernoulli. Se remplaza el valor seleccionado por uno proveniente de una Distribución Uniforme de valores $[a, b]$ más grandes que la original.

(77)
$$\boldsymbol{Xcontaminado}_{i1} = Xoriginal(99\%)_{i1} \cup Xextremo(1\%)_{i1}$$

La variable contaminada sustituye a la original en la estimación de los modelos y permite representar valores extremos cuyo valor aumenta conforme aumenta la variable dependiente y también genera **heterocedasticidad** residual creciente al ser estimada la variable dependiente original. Sin embargo, para garantizar la estimación de los modelos y consolidar un error aún más heteroscedástico, se agregan a la variable dependiente n valores aleatorios provenientes de una distribución normal ecuación (73), cuya desviación estándar proviene de la varianza expresada en la ecuación (74) formada a partir de las variables uniformes utilizadas para el modelo de la media. Finalmente, se obtiene la variable dependiente como se expresa en el modelo de la ecuación (30).

3.6 Evaluación del desempeño de los estimadores

Para evaluar el desempeño de los estimadores es necesario mencionar los conceptos de sesgo, precisión y exactitud:

- El **sesgo** es la diferencia entre el promedio de la muestra de coeficientes obtenidos mediante el proceso de simulación y el valor verdadero.
- La **precisión** es una medida estadística de varianza del proceso de estimación, atribuible a la variabilidad presente en la muestra de coeficientes obtenidos mediante el proceso de simulación.
- La **exactitud** combina el sesgo y la precisión para definir el desempeño de un estimador. Entre más sesgado y menos preciso el estimador, peor es la estimación puntual. Por lo tanto, la exactitud se define como la distancia total entre los valores observados y el valor verdadero.

Estos conceptos son cualitativos y por lo tanto para cuantificar el desempeño de los estimadores es necesario precisarlos mediante medidas cuantitativas (Bruno A. & Moore, 2005).

El análisis de las estimaciones generadas mediante el proceso de simulación Montecarlo para cada modelo se enfoca en la variable Ingreso y su relación con el Gasto, pues es la de mayor peso en la **heterocedasticidad** residual. Se emplean para ello como **medidas del sesgo: el promedio y la mediana** para determinar cuánto se alejan los mismos del valor real del parámetro del Ingreso. Como **medida de precisión** para caracterizar la variabilidad de los estimadores se utiliza la desviación estándar.

(78)

$$Promedio = \frac{1}{n}\sum_{i=1}^{n} O_i$$

(79)

$$Desviación\ estándar = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(O_i - \overline{O_i})^2}$$

donde**:**

O_i : Es el valor observado para el coeficiente de interés en la corrida i.

$\overline{O_i}$: Es el valor promedio de los coeficientes observados en la simulación.

n: La cantidad de simulaciones realizadas.

Como ***medida de exactitud*** se emplea el error cuadrático medio (ECM), que es la media de las diferencias al cuadrado e indica cuán cerca se encuentra el estimador del valor real. Esta medida incorpora los conceptos de sesgo y precisión. Al elevar al cuadrado todas las diferencias, el ECM no se encuentra en las unidades de medida originales, por lo que, para regresar a la escala original, se utiliza la raíz cuadrada del ECM. A esto último se le conoce como raíz del error cuadrático medio (RECM).

(80)

$$RECM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - R)^2}$$

Donde**:**

O_i : Es el valor observado para el coeficiente de interés en la corrida i.

R: Es el valor real del coeficiente de interés, fijado antes de realizar la simulación.

n: La cantidad de simulaciones realizadas.

El RECM se complementa con el error absoluto medio (EAM) calculado mediante el promedio de las diferencias absolutas entre el valor estimado y el real.

(81)

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|O_i - R|$$

CAPTULO 4: ESTUDIO DE SIMULACIÓN

El estudio de simulación busca evaluar la exactitud y precisión de los estimadores de los coeficientes de regresión de los cuatro modelos planteados en esta investigación. De acuerdo con la revisión bibliográfica, se esperaría que el modelo con el menor sesgo y error aleatorio sea el modelo de tipo lineal doble generalizado porque permite considerar la estructura de la media y la dispersión por separado y depender de las covariables, siendo la variable dependiente del submodelo de dispersión el componente de desviación del submodelo de la media, por lo tanto cada residual de este último proviene de una distribución Normal con varianza distinta para cada individuo.

Para el desarrollo del estudio Monte Carlo se consideran diferentes causas de **heterocedasticidad** y un tamaño de muestra de 5687 observaciones, congruente con el caso empírico. Para generar las variables explicativas a partir valores aleatorios se utiliza la Distribución Uniforme, en concreto se utilizan los valores mínimos y máximos de cada variable independiente del caso empírico en la unidad de medida correspondiente para sustituir los parámetros **a** y **b**. Las distribuciones utilizadas para cada variable se detallan en el siguiente Cuadro (**Cuadro #3**):

Cuadro # 3: Simulación MCMC, distribuciones utilizadas para generar las variables independientes

Distribuciones utilizadas para generar las variables independientes	
Variable	**Distribución utilizada**
Ingreso	~ Uniforme (a=0, b=3500)
Miembros	~ Uniforme (a=1, b=8)
Escolaridad	~ Uniforme (a=1, b=18)
Edad	~ Uniforme (a=21, b=86)

4.1. Resultados de la evaluación del desempeño

4.1.1. Escenario de heterocedasticidad producida por naturaleza de las variables

En el caso de la **heterocedasticidad** generada por la naturaleza de las variables explicativas se decide seguir los coeficientes obtenidos por el modelo doble generalizado calculado para los datos del caso empírico. Al usar las mismas ecuaciones que en el caso empírico, pero utilizando variables pseudoaleatorias con Distribución Uniforme, posiblemente la variancia sea mayor a la observada en el caso empírico.

La ecuación de la varianza del error residual se genera mediante:

(82) $$\sigma_i^2 = e^{(21.56+0.0008*Ingreso+0.14*Miembros+0.04*Escolaridad+0.009*Edad)}$$

Para simular un error heteroscedástico con distribución Gaussiana se obtiene la raíz cuadrada de la expresión anterior de varianza (σ_i^2) y se generan 5687 valores aleatorios provenientes de una distribución normal. Posteriormente se calcula la variable dependiente como sigue:

(83) $$Gasto_{nv} = -1000 + 60 * Ingreso + 18000 * Miembros + 1300 * Escolaridad + 500 * Edad + e_i$$

Léase la variable Ingreso en miles de colones.

Se puede observar en el **Cuadro (#4 y #5)** que todos los modelos estiman adecuadamente el valor del coeficiente del Ingreso $(\beta_{Ingreso} = 60)$ a excepción del modelo bayesiano doble generalizado propuesto **(ver Cuadro #4)**. Los indicadores de desempeño revelan que el modelo lineal doble generalizado **(DGLM)** es el que mejor lo estima. Esto se refleja en la mediana y el promedio de los coeficientes, pues son los más cercanos al valor original. El modelo (DGLM) a la vez presenta los valores más pequeños de desviación estándar, coeficiente de variación, raíz cuadrada del error cuadrático medio y error absoluto medio.

Cuadro # 4: Resultados de la simulación NV-10000

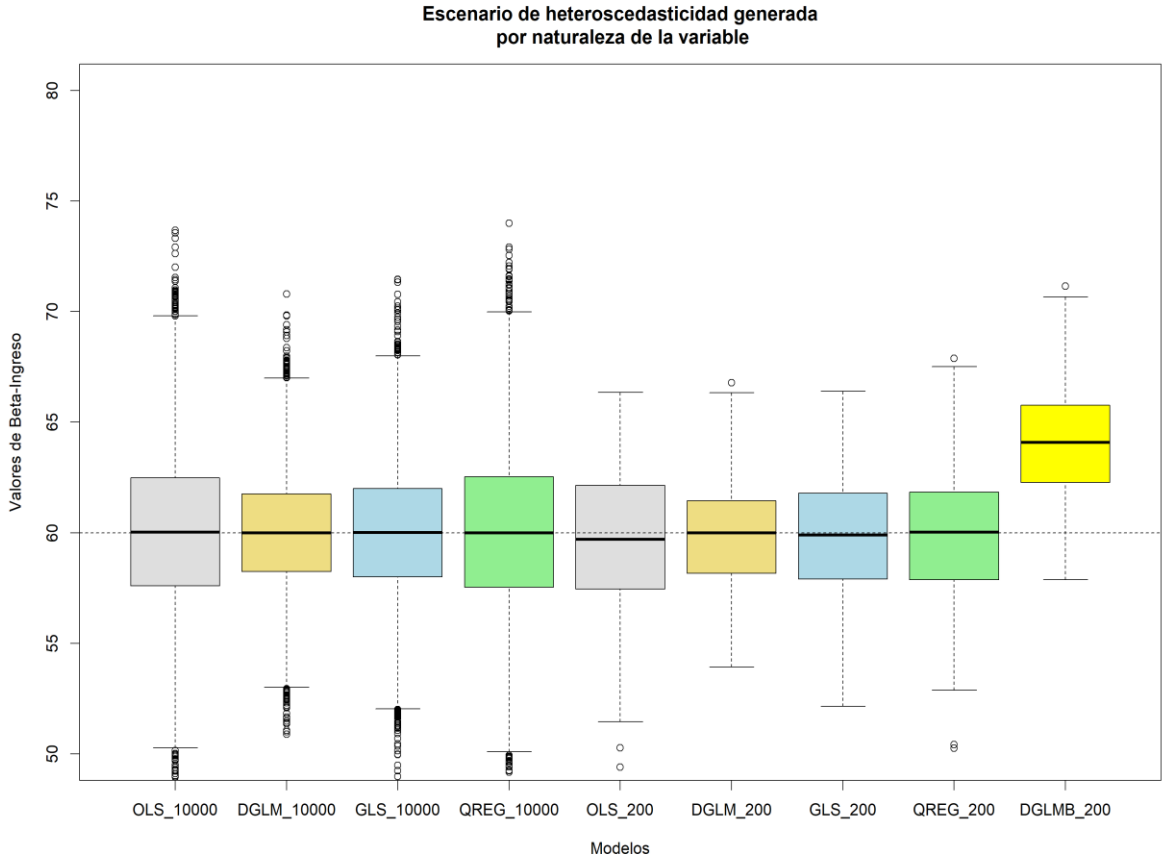
Resultados de la simulación (i=10000)								
Simulación heteroscedasticidad producida por naturaleza de las variables	OLS Beta	OLS se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se
Mediana	60,0207	3,2391	60,0021	2,6385	60,0077	2,6722	59,9918	3,2976
Media	60,0392	3,2389	60,0096	2,6384	60,0253	2,6721	60,0050	3,2975
Desviación estándar	3,5648	0,0468	2,6325	0,0309	2,9613	0,0350	3,6486	0,1483
Coeficiente de variación	5,94%	1,45%	4,39%	1,17%	4,93%	1,31%	6,08%	4,50%
Raíz del error cuadrático medio	4,7271		3,4906		3,9268		4,8379	
Error absoluto medio	2,2380		1,9190		2,0397		2,2669	

Cuadro # 5: Resultados de la simulación NV-200

Resultados de la simulación (i=200)										
Simulación heteroscedasticidad producida por naturaleza de las variables	MCO Beta	MCO se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se	DGLMB Beta	DGLMB se
Mediana	59,704	3,242	59,996	2,644	59,893	2,676	60,030	3,313	64,086	2,521
Media	59,647	3,243	59,924	2,639	59,760	2,674	59,751	3,297	92,880	104,500
Desviación estándar	3,362	0,050	2,506	0,034	2,748	0,039	3,351	0,159	360,071	502,285
Coeficiente de variación	5,64%	1,54%	4,18%	1,30%	4,60%	1,46%	5,61%	4,84%	387,67%	480,66%
Raíz del error cuadrático medio	0,632		0,469		0,516		0,629		67,637	
Error absoluto medio	0,304		0,264		0,278		0,298		1,627	

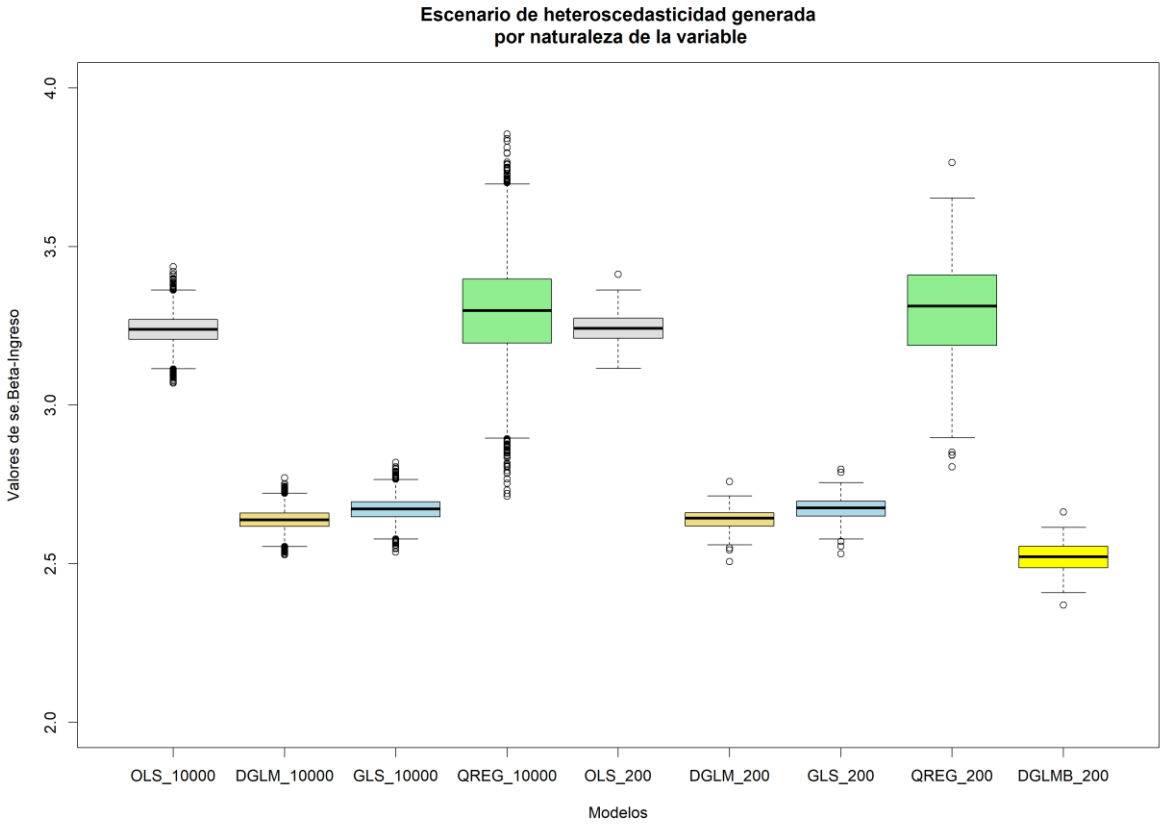
El gráfico de cajas **(Figura 1)** evidencia con mayor claridad lo antes expuesto; muestra como la mayoría de las medianas se encuentran a un mismo nivel y se aprecia la dispersión de las estimaciones de cada modelo:

Figura 1: Comparación de coeficientes del Ingreso, escenario (NV)



Debido a que los efectos de la **heterocedasticidad** se ven reflejados en la inflación del error estándar, se analiza el mismo para el coeficiente del Ingreso y nuevamente el modelo (DGLM) brinda los menores y menos variables, tal y como lo demuestra el siguiente gráfico de cajas (**Figura 2**):

Figura 2. Comparación de errores estándar del Ingreso, escenario (NV).



Mediante el **Figura 2** se evidencia que se están modelando errores estándar muy similares para el modelo DGLM y **GLS**. El modelo QREG muestra un error estándar medio más grande y de amplia variabilidad, lo que genera una mayor repartición en las pruebas de hipótesis que se rechazan o que no se rechazan. Finalmente la valoración para el modelo bayesiano es que su error estándar en su mayoría es bajo. Sin embargo presenta valores atípicos que elevan su valor medio.

4.1.2. Escenario de heterocedasticidad producida por sesgo de variable omitida

Para construir la ***variable omitida,*** en primer lugar se genera una variable proveniente de una Distribución Uniforme $[0,3500]$ se pondera por ½ y se le agrega a la variable Ingreso generada a partir de otra Distribución Uniforme $[0,3500]$, también ponderada por ½, con el objetivo de forzar la correlación de la variable omitida con el Ingreso.

Cuadro # 6: Distribuciones utilizadas para generar la variable omitida

Distribuciones utilizadas para generar la variable omitida	
Variable	**Distribución**
Ingreso	~Uniforme (a=0, b=3500)
Nueva variable	~Uniforme (a=0, b=3500)

Variable omitida:

(84) $$Xomitida = ½ * Ingreso + ½ * Nueva\_variable$$

Posteriormente el término de error se obtiene a partir de una distribución normal con media igual a 0 y desviación estándar de 60 veces la variable omitida, esto último debido a que la variable omitida, al estar expresada en unidades de miles de colones, arrojaría una desviación estándar muy pequeña:

(85) $$error \sim N\{0, [60 * (Xomitida)]\}$$

Finalmente, la variable dependiente (Gasto) se obtiene a partir de la siguiente ecuación:

(86) $$\boldsymbol{Gasto_{vo}} = -1000 + 60 * Ingreso + 18000 * Miembros + 1300 * Escolaridad + 500 * Edad + 60 * Xomitida + error$$

Léase la variable Ingreso en miles de colones.

Para las simulaciones de este escenario todos los modelos sobrestiman el valor real del parámetro Ingreso, pues son mayores al valor hipotético de 60. Este comportamiento es el esperado pues manifiesta que la pendiente del Ingreso es una combinación lineal formada por la adición de la variable Ingreso y una fracción (1/2) de la variable Omitida generada. El modelo que obtiene las estimaciones más cercanas al parámetro establecido para el Ingreso es el de mínimos cuadrados generalizados **(GLS)**. Lo anterior se refleja en ambos conjuntos de simulaciones (i=200 e i=10000) en los **Cuadros (#7 y #8)**.

Cuadro # 7: Resultados de la simulación VO-10000

Resultados de la simulación (i=10000)								
Simulación heteroscedasticidad producida por sesgo de variable omitida	**OLS Beta**	**OLS se**	**DGLM Beta**	**DGLM se**	**GLS Beta**	**GLS se**	**QREG Beta**	**QREG se**
Mediana	90,0234	1,5412	90,0371	1,4671	90,0189	1,4211	93,0202	1,6570
Media	90,0031	1,5415	90,0176	1,4674	89,9978	1,4213	93,0097	1,6582
Desviación estándar	1,6334	0,0271	1,4926	0,0220	1,5377	0,0230	1,7523	0,0653
Coeficiente de variación	1,81%	1,76%	1,66%	1,50%	1,71%	1,62%	1,88%	3,94%
Raíz del error cuadrático medio	39,8444		39,8539		39,8306		43,8339	
Error absoluto medio	7,2635		7,2653		7,2629		7,6188	

Cuadro # 8: Resultados de la simulación VO-200

Resultados de la simulación (i=200)										
Simulación heteroscedasticidad producida por sesgo de variable omitida	MCO Beta	MCO se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se	DGLMB Beta	DGLMB se
Mediana	90,061	1,542	89,988	1,464	90,045	1,421	93,067	1,655	96,111	1,408
Media	90,033	1,542	90,012	1,467	89,992	1,422	93,140	1,657	89,862	58,141
Desviación estándar	1,557	0,021	1,445	0,019	1,488	0,020	1,582	0,067	269,174	376,440
Coeficiente de variación	1,73%	1,36%	1,60%	1,27%	1,65%	1,43%	1,70%	4,04%	299,54%	647,46%
Raíz del error cuadrático medio	5,640		5,635		5,631		6,222		50,663	
Error absoluto medio	1,028		1,027		1,027		1,080		1,625	

El diagrama de cajas y bigotes **(Figura 3)** presenta las distribuciones del modelo DGLM y **GLS** muy similares. Sin embargo se conoce en los Cuadros **(#6 y #7)** que la raíz del error cuadrático medio y el error absoluto medio, favorecen al modelo de mínimos cuadrados generalizados.

Figura 3: Comparación de coeficientes del Ingreso, escenario (VO).

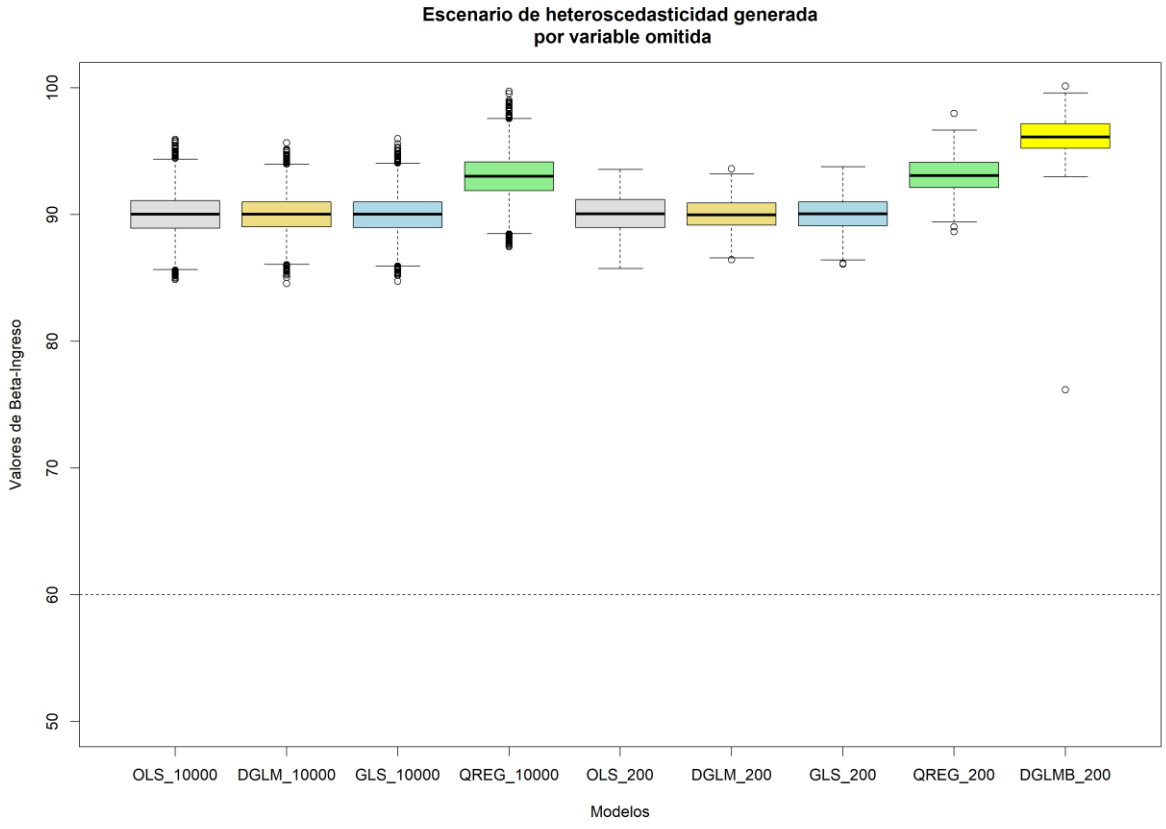
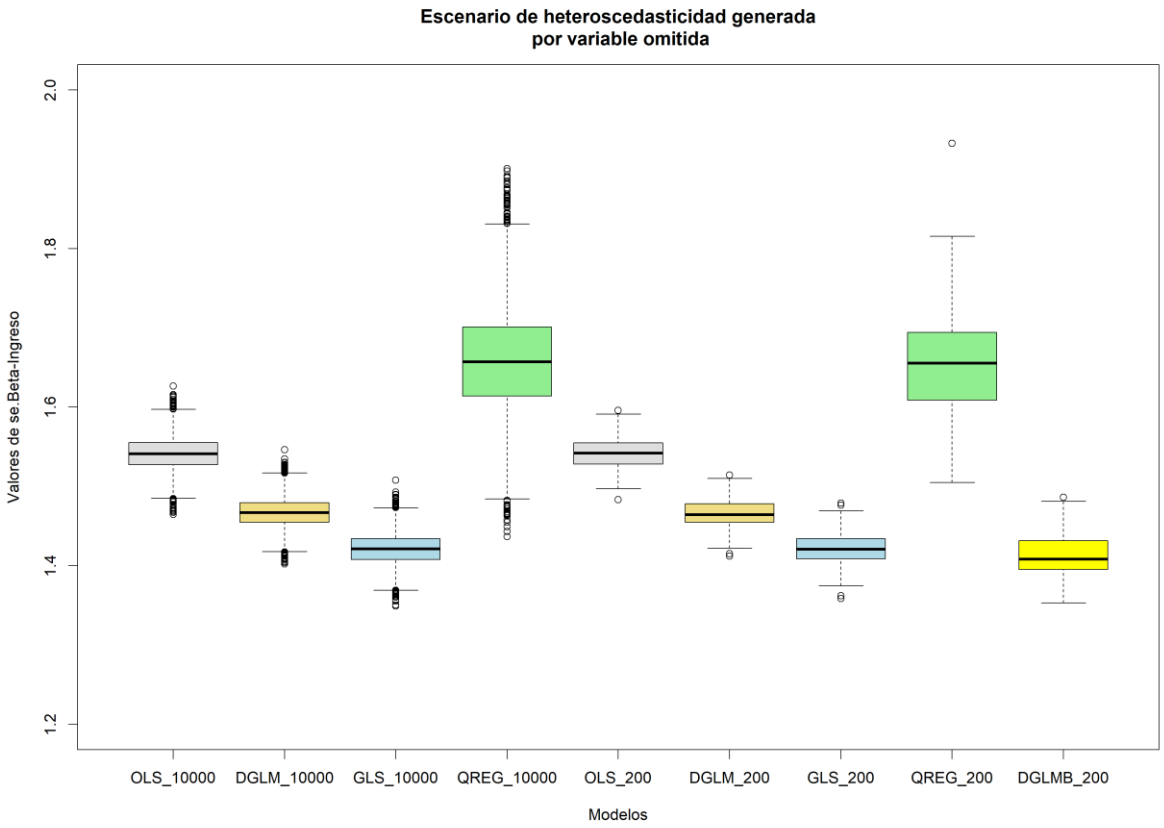


Figura 4: Comparación de errores estándar del Ingreso, escenario (VO)



El gráfico **(Figura 4)** apoya la evidencia de que el modelo de mínimos cuadrados generalizados **(GLS)** presenta errores estándares de poca variabilidad y los de valor numérico menor para este escenario.

Si se incluye la variable omitida en cada uno de los modelos como variable explicativa del Gasto, el error de sobrestimación de los coeficientes desaparece y el modelo que mejor representa el coeficiente del Ingreso vuelve a ser el DGLM. Esto se aprecia en los **Cuadros #9 y #10**.

Cuadro # 9: Resultados de la simulación sin VO-10000

Resultados de la simulación (i=10000)								
Simulación sin sesgo de variable omitida	**OLS Beta**	**OLS se**	**DGLM Beta**	**DGLM se**	**GLS Beta**	**GLS se**	**QREG Beta**	**QREG se**
Mediana	60,0163	2,1051	60,0444	1,6300	60,0205	1,8697	59,9653	2,2294
Media	59,9925	2,1058	60,0047	1,6304	59,9988	1,8700	59,9592	2,2310
Desviación estándar	2,0142	0,0315	1,7247	0,0214	1,7980	0,0251	2,2440	0,0880
Coeficiente de variación	3,36%	1,50%	2,87%	1,31%	3,00%	1,34%	3,74%	3,95%
Raíz del error cuadrático medio	212,7274		182,2242		189,9935		236,3388	
Error absoluto medio	2,8209		2,4164		2,5194		3,1340	

Cuadro # 10: Resultados de la simulación sin VO-200

Resultados de la simulación (i=200)										
Simulación sin sesgo de variable omitida	MCO Beta	MCO se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se	DGLMB Beta	DGLMB se
Mediana	60,064	2,106	59,953	1,628	60,105	1,872	60,149	2,224	59,980	1,641
Media	60,148	2,107	60,023	1,630	60,089	1,870	60,152	2,222	43,063	35,247
Desviación estándar	2,073	0,030	1,743	0,021	1,840	0,025	2,183	0,086	162,067	235,626
Coeficiente de variación	3,45%	1,40%	2,90%	1,31%	3,06%	1,33%	3,63%	3,89%	376,35%	668,51%
Raíz del error cuadrático medio	2,073		1,739		1,838		2,183		162,547	
Error absoluto medio	1,658		1,418		1,480		1,768		20,969	

Las **figuras 5 y 6,** por su parte, detallan que los modelos estiman adecuadamente el parámetro del Ingreso. Sin embargo, el modelo DGLM es el que lo hace con mayor precisión respecto al valor del parámetro y ofrece errores estándar más bajos.

Figura 5: Comparación de coeficientes del Ingreso, escenario sin (VO)

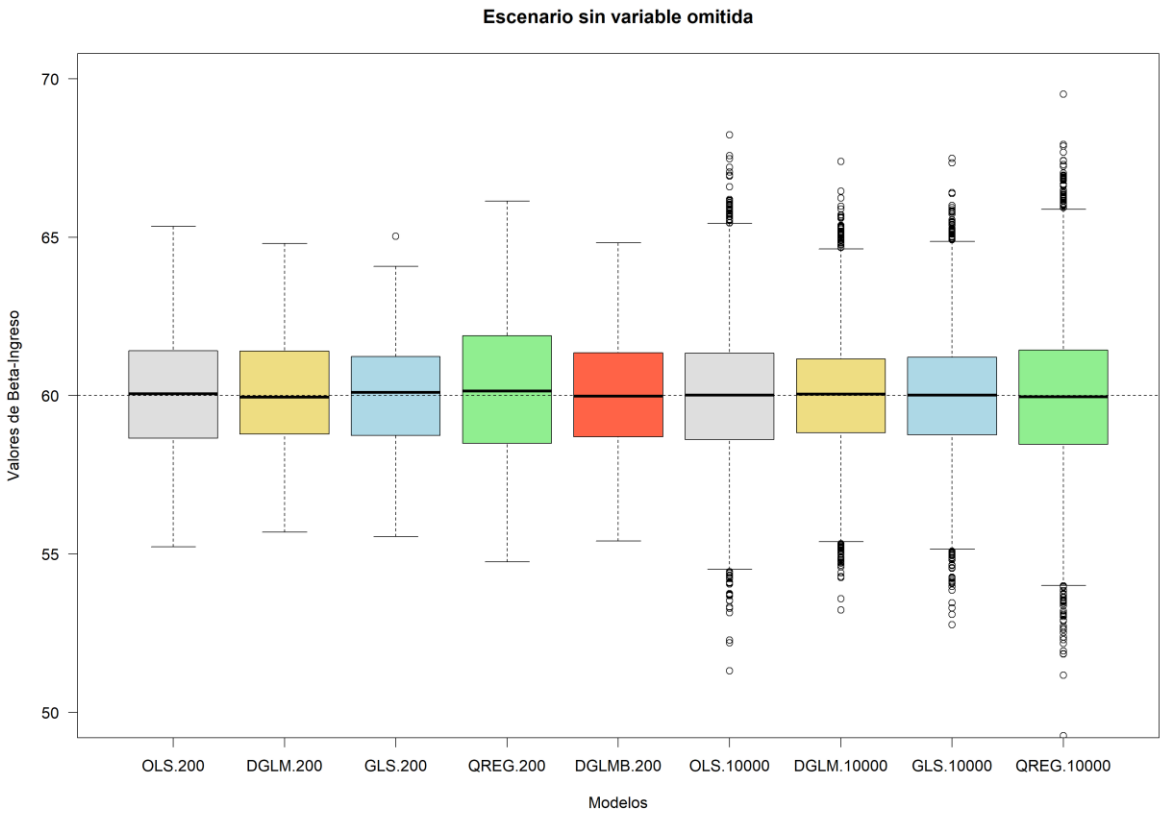
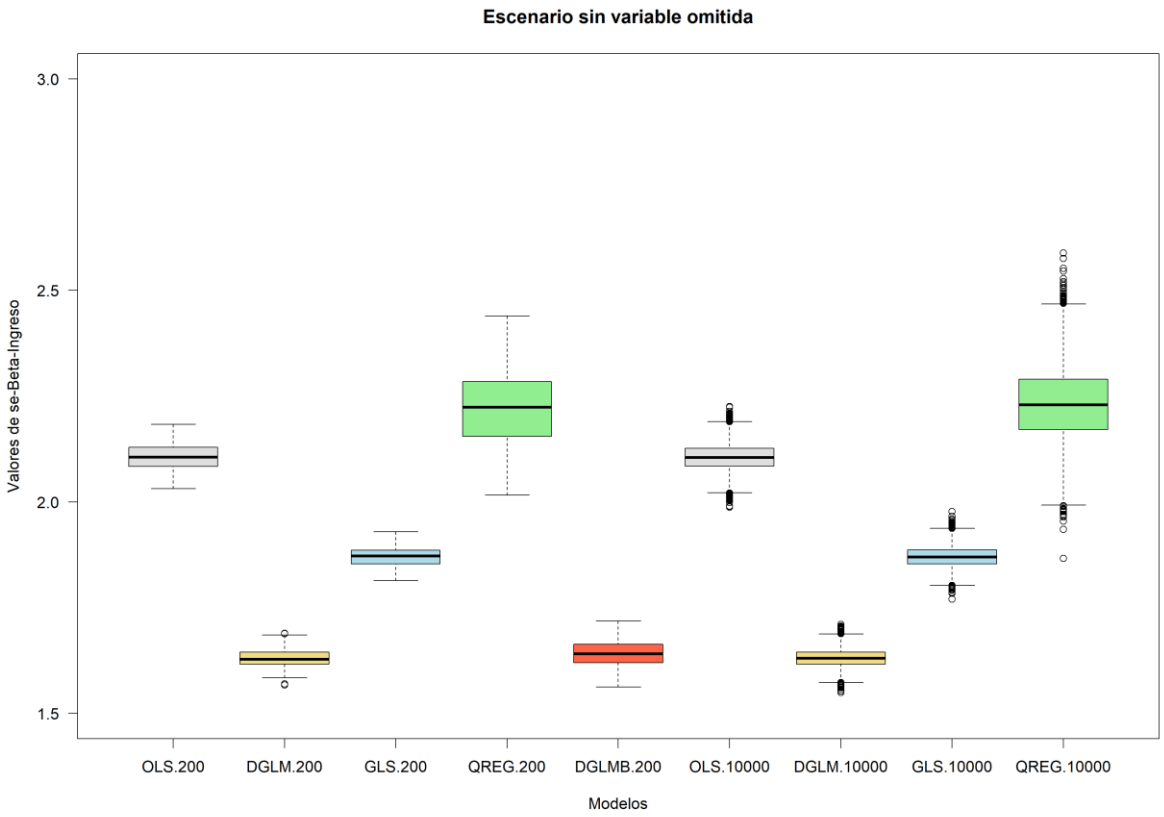


Figura 6: Comparación de errores estándar del Ingreso, escenario sin (VO)



4.1.3. Escenario de heterocedasticidad producida por presencia de valores extremos

Para simular valores extremos se toma en consideración que el Ingreso es la variable que los exhibe en mayor proporción en el caso empírico. Por esta razón, primero se utilizan Distribuciones Uniformes para generar las variables predictoras, como se detalla en el Cuadro (**Cuadro #3)**:

A continuación, se ordena la variable Ingreso de menor a mayor valor y se crea una variable Ingreso contaminada que conserva 5630 (99%) valores originales de la variable Ingreso y se contaminan los 57 valores restantes (1%) seleccionados aleatoriamente mediante una distribución de Bernoulli. Para que, del total de casos (5687) el 1% representen Ingresos extremos, el proceso se lleva a cabo de la siguiente manera:

(87) $$\boldsymbol{Ingreso\ contaminado} = Ingreso\ original(99\%) \cup Ingreso\ extremo(1\%)$$

La variable Ingreso extremo se genera tal y como se detalla en la Simulación VE (Cuadro #10):

Cuadro # 11: Simulación VE, distribuciones utilizadas para generar el Ingreso contaminado

Distribuciones utilizadas para generar el Ingreso contaminado			
Si [Bernoulli (99%)]=0	***Ingreso contaminado=* Ingreso original**	***Si [Bernoulli (1%)]=1***	***Ingreso contaminado= Ingreso extremo***
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [0,499]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[501,1000]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [500,999]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[1001,1500]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [1000,1499]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[1501,2000]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [1500,1999]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[2001,2500]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [2000,2499]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[2501,3000]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [2500,2999]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[3001,3500]$
$\text{Bernoulli}_i = 0$	$\text{Ingreso}_i \in [3000,3500]$	$\text{Bernoulli}_i = 1$	$\text{Ingreso}_i \in U[3501,4000]$

Seguidamente, se calcula la variable Gasto sin el término de error, mediante la siguiente ecuación:

(88) $$Gasto_{ve} = -1000 + 60 * Ingreso + 18000 * Miembros + 1300 * Escolaridad + 500 * Edad$$

Léase el coeficiente para la variable Ingreso en miles de colones.

Finalmente son agregados al $Gasto_{ve}$, 5687 valores aleatorios provenientes de una distribución normal $e_i \sim N(0, \sigma_i)$ cuya desviación estándar proviene de:

(89) $$\sigma_i^2 = e^{(21.56+0.0008*Ingreso+0.14*Miembros+0.04*Escolaridad+0.009*Edad)}$$

Para el escenario de valores extremos en promedio, todos los modelos subestiman el valor del parámetro del Ingreso. Al evaluar los estadísticos de ajuste se aprecia que el modelo DGLM presenta los mejores tanto para la raíz del error cuadrático medio como para el error absoluto medio. Además, si se analiza el error estándar promedio para las 10000 iteraciones, el modelo DGLM arroja los más pequeños, indicando que es el menos afectado por los valores extremos de la variable Ingreso. Lo anterior se respalda en los Cuadros (#12 y #13).

Cuadro # 12: Resultados de la simulación VE-10000

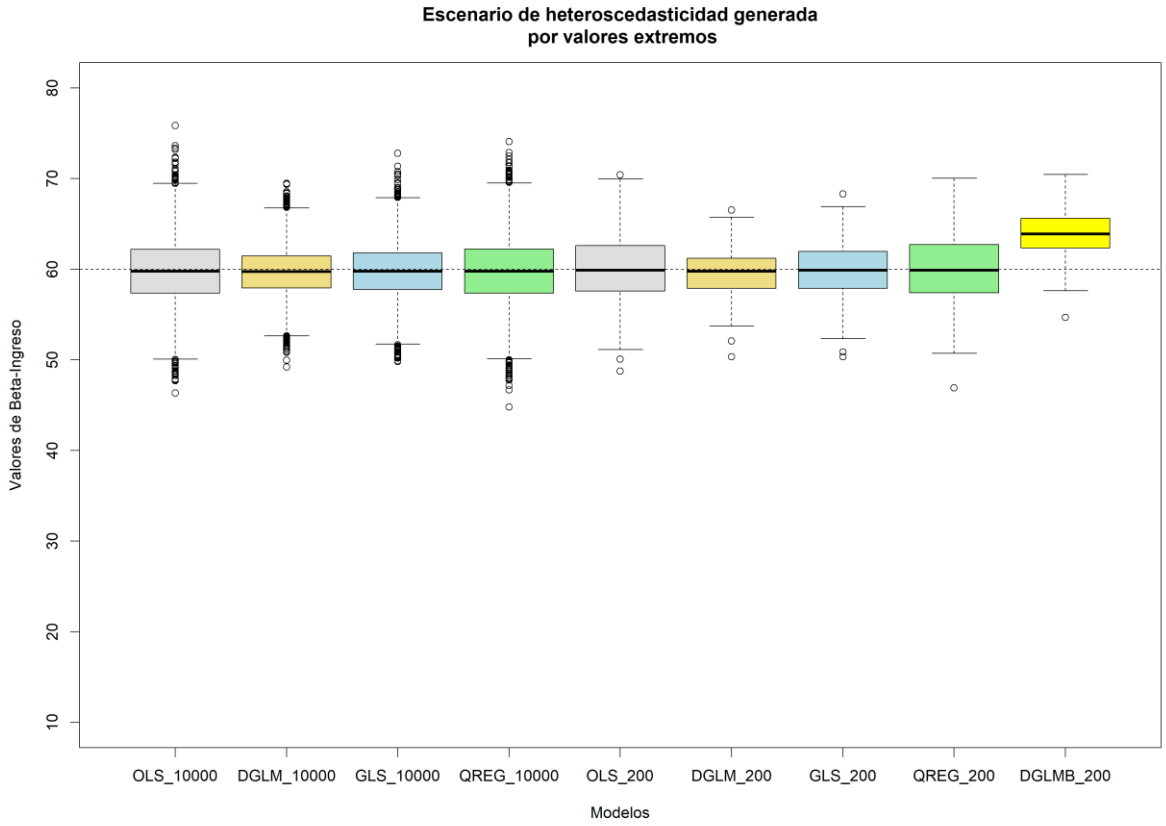
Simulación heteroscedasticidad producida por presencia de valores extremos	OLS Beta	OLS se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se
Mediana	59,7891	3,2341	59,7249	2,6332	59,7784	2,6684	59,7746	3,2938
Media	59,8072	3,2353	59,7384	2,6335	59,8157	2,6689	59,8026	3,2932
Desviación estándar	3,5941	0,0466	2,6550	0,0315	2,9816	0,0351	3,6258	0,1470
Coeficiente de variación	6,010%	1,441%	4,444%	1,196%	4,985%	1,315%	6,063%	4,464%
Raíz del error cuadrático medio	4,7726		3,5375		3,9610		4,8148	
Error absoluto medio	2,2487		1,9342		2,0489		2,2577	

Cuadro # 13: Resultados de la simulación VE-200 Cuadro

Simulación heteroscedasticidad producida por presencia de valores extremos	OLS Beta	OLS se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se
Mediana	59,7891	3,2341	59,7249	2,6332	59,7784	2,6684	59,7746	3,2938
Media	59,8072	3,2353	59,7384	2,6335	59,8157	2,6689	59,8026	3,2932
Desviación estándar	3,5941	0,0466	2,6550	0,0315	2,9816	0,0351	3,6258	0,1470
Coeficiente de variación	6,010%	1,441%	4,444%	1,196%	4,985%	1,315%	6,063%	4,464%
Raíz del error cuadrático medio	4,7726		3,5375		3,9610		4,8148	
Error absoluto medio	2,2487		1,9342		2,0489		2,2577	

Es importante destacar que la estimación del modelo Bayesiano (BDGLM) del parámetro de interés no es tan buena. El gráfico **(Figura 7)** apoya las conclusiones obtenidas a partir de los estimadores de desempeño:

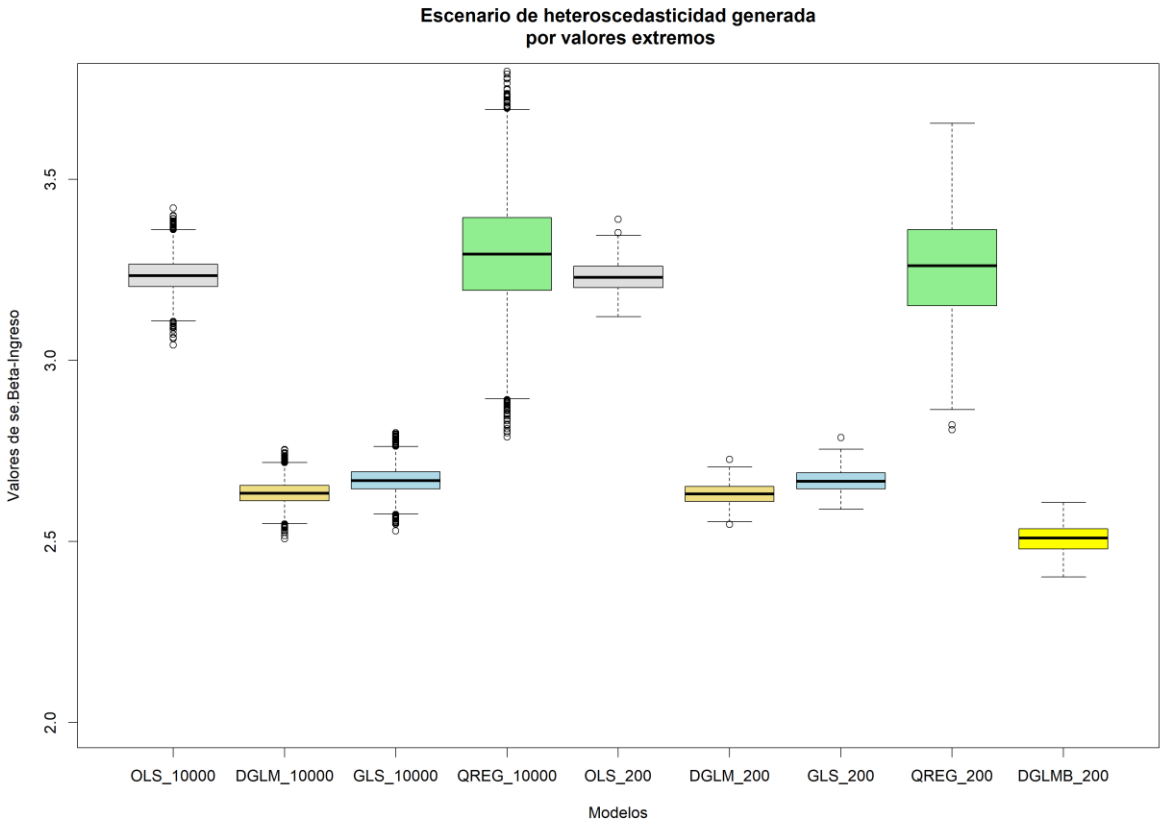
Figura 7: Comparación de coeficientes del Ingreso, escenario (VE)



Cuando se analiza la estimación del error estándar de los coeficientes, el modelo lineal doble generalizado (**DGLM**) incluye los errores más pequeños. Por lo tanto, en este escenario, el modelo que mejor aproxima el coeficiente y el que genera los menores errores estándares es el antes mencionado. El modelo bayesiano presenta algunos valores del error estándar muy altos, lo que

contribuye a elevar el error promedio, aunque no la mediana ni la gran mayoría de sus datos. El siguiente gráfico (**Figura 8**) ilustra lo antes descrito:

Figura 8: Comparación de errores estándar del Ingreso, escenario (VE)



4.1.4. Comparación entre escenarios

Cuando se comparan las estimaciones del ejercicio de 10000 simulaciones y las restantes 200 que incluyen el modelo bayesiano, se observa que se mantiene el mismo comportamiento en los modelos presentes en ambas, aunque el modelo DGLM presenta las menos dispersas y el modelo bayesiano presenta una mediana en general mayor que los demás modelos en la estimación del parámetro. Finalmente, los diagramas de cajas y bigotes (**Figura 9 y 10**) muestran distribuciones simétricas en los estimadores que se pueden deber a que el tamaño de la muestra es bastante grande, por lo que este comportamiento apoya el supuesto de normalidad asintótica.

Figura 9: Comparación de coeficientes del Ingreso, escenario n=10000

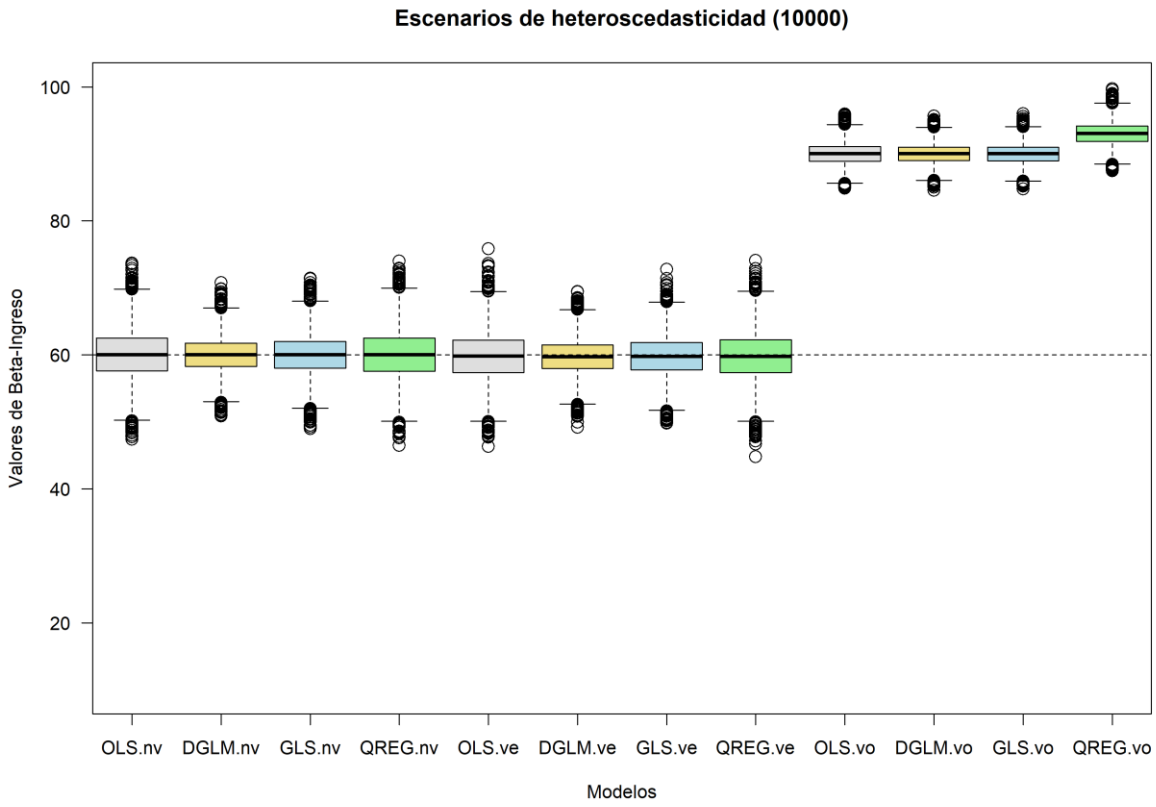
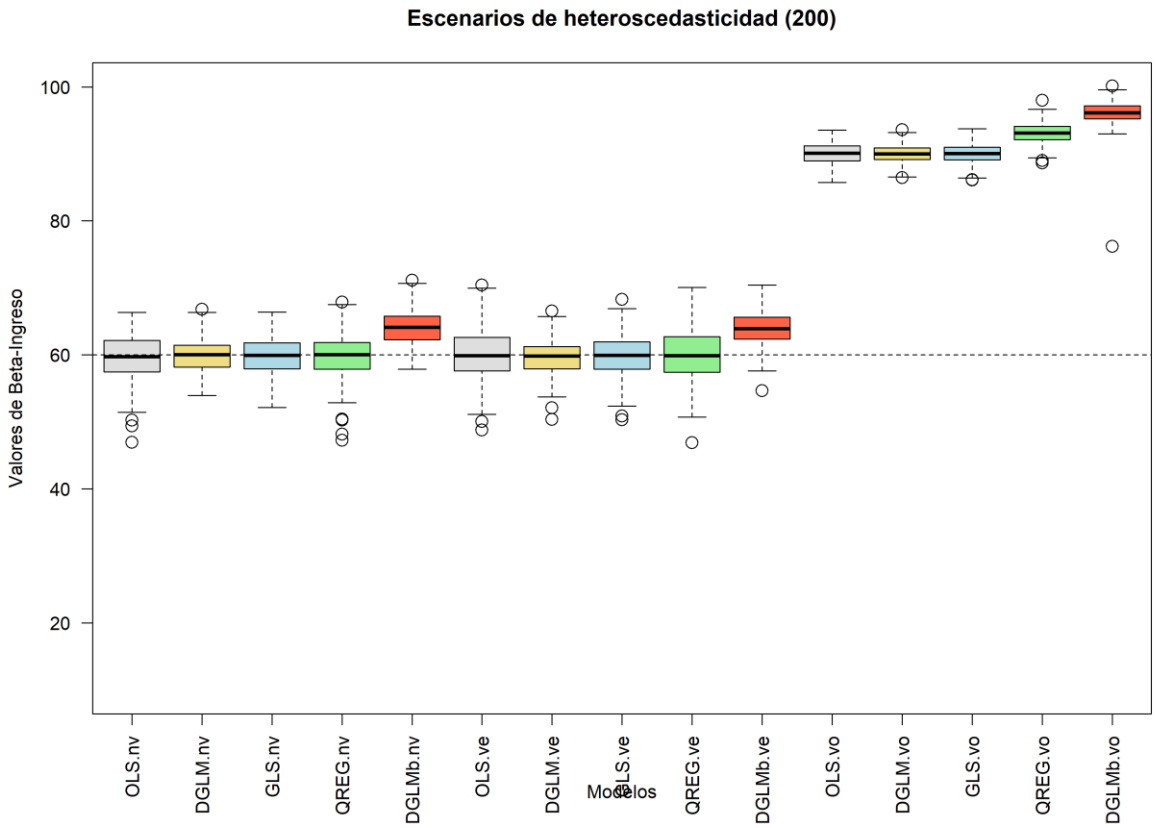


Figura 10: Comparación de coeficientes del Ingreso, escenario n=200



En los errores estándar de los coeficientes del Ingreso para la simulación (i=10000) (**Figura 11**), se observa que el modelo lineal doble generalizado presenta el mejor desempeño conjunto para los escenarios, con la salvedad en el escenario de **heterocedasticidad** generada por variable omitida en donde el modelo de mínimos cuadrados ponderados genera los errores estándar más cercanos a cero. También se puede apreciar que la regresión cuantílica para todos los escenarios presenta mayores valores medios y de dispersión (**Cuadro #14** y **Cuadro #15**).

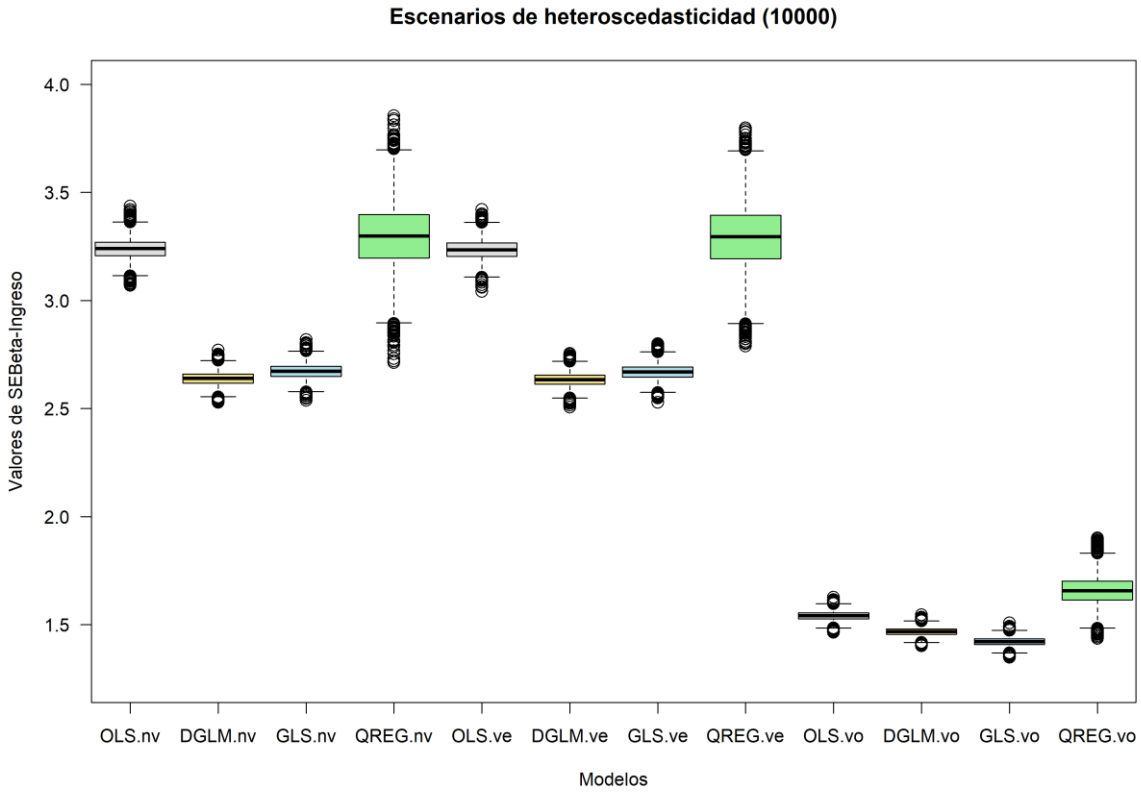
Cuadro # 14: Resultados de la simulación VE-10000

Simulación heteroscedasticidad producida por presencia de valores extremos	OLS Beta	OLS se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se
Mediana	59,7891	3,2341	59,7249	2,6332	59,7784	2,6684	59,7746	3,2938
Media	59,8072	3,2353	59,7384	2,6335	59,8157	2,6689	59,8026	3,2932
Desviación estándar	3,5941	0,0466	2,6550	0,0315	2,9816	0,0351	3,6258	0,1470
Coeficiente de variación	6,010%	1,441%	4,444%	1,196%	4,985%	1,315%	6,063%	4,464%
Raíz del error cuadrático medio	4,7726		3,5375		3,9610		4,8148	
Error absoluto medio	2,2487		1,9342		2,0489		2,2577	

Cuadro # 15: Resultados de la simulación VE-200

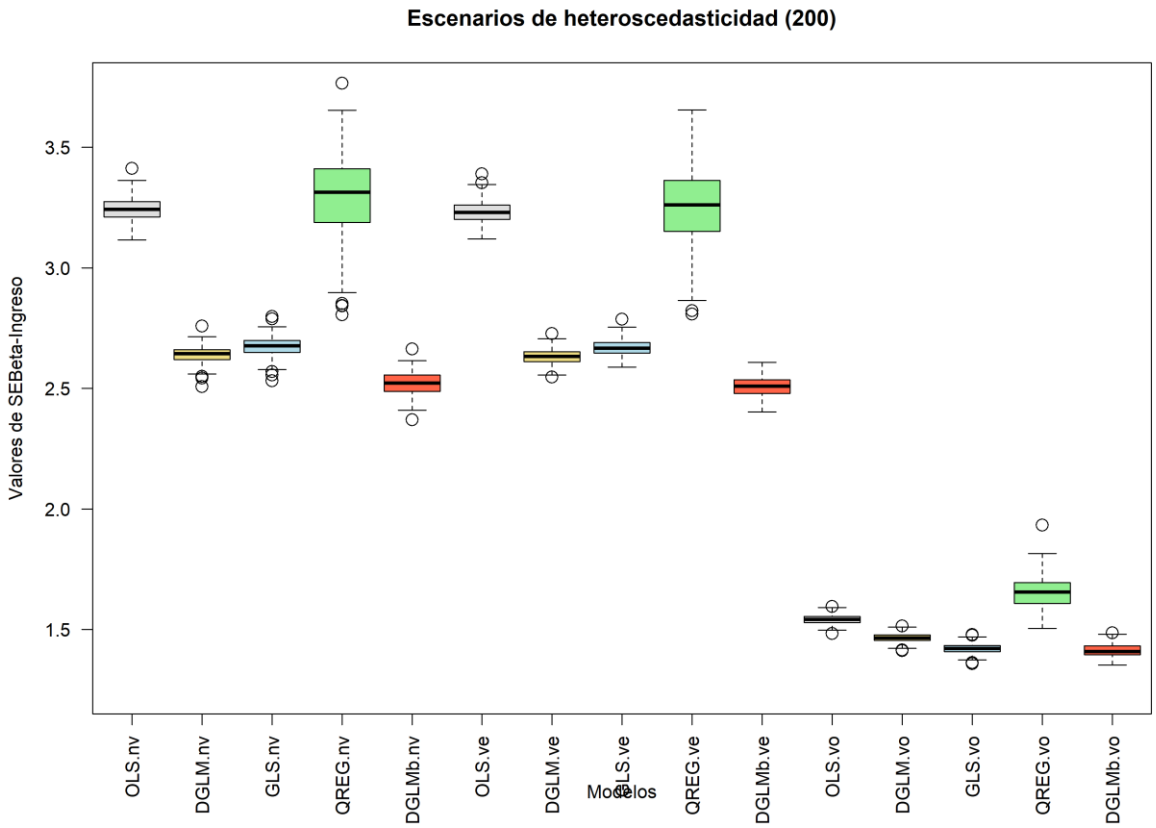
Resultados de la simulación (i=200)										
Simulación heterocedasticidad producida por presencia de valores extremos	OLS Beta	OLS se	DGLM Beta	DGLM se	GLS Beta	GLS se	QREG Beta	QREG se	DGLM Beta	DGLM se
Mediana	59,8733	3,2298	59,7788	2,6314	59,8944	2,6657	59,8787	3,2609	63,8850	2,5093
Media	60,0570	3,2310	59,6693	2,6328	59,9479	2,6675	59,9159	3,2629	55,0398	18,2720
Desviación estándar	3,7841	0,0470	2,6293	0,0302	3,1409	0,0317	3,8668	0,1687	122,3245	172,4025
Coeficiente de variación	6,30%	1,45%	4,41%	1,15%	5,24%	1,19%	6,45%	5,17%	222,25%	943,53%
Raíz del error cuadrático medio	5,0183		3,5138		4,1654		5,1285		162,3331	
Error absoluto medio	2,2985		1,9091		2,0885		2,3339		5,1663	

Figura 11: Comparación de errores estándar del Ingreso, escenario n=10000



Una vez que se incorporó el modelo bayesiano (i=200), las anteriores conclusiones para el error estándar (i=10000) se mantienen debido a que la dispersión que presenta este modelo en la estimación del error estándar es muy alta respecto a los demás (debido a valores positivos muy altos). Esto último no se refleja en el gráfico (**Figura 12**) por un ajuste de escala, sin embargo, puede ser comprobado en los cuadros que presentan los estadísticos de ajuste respectivos.

Figura 12: Comparación de errores estándar del Ingreso, escenario n=200.



CAPTULO 5: ANÁLISIS DEL CASO EMPÍRICO

Para la aplicación empírica se toman en cuenta los datos de la Encuesta de Ingresos y Gastos del año 2013, tomada del programa acelerado de datos del Instituto Nacional de Estadística y Censos INEC. La Encuesta de Ingresos y Gastos de los Hogares constituye una fuente de información fundamental para instituciones públicas, organismos no gubernamentales, universidades, e investigadores. Busca generar estadísticas sobre el monto, la distribución y la estructura del ingreso y gasto de los hogares. (INEC, 2015)

Esta encuesta es una encuesta por muestreo en donde la unidad de análisis es el hogar particular y residente en Costa Rica. Se entiende por hogar particular: "La persona sola o grupo de personas, residentes habituales en una vivienda individual, con vínculos familiares o sin ellos, que usan un fondo común para atender los gastos de alimentación y vivienda. Aquellas viviendas donde residían 6 o menos miembros no familiares se les dio el mismo tratamiento que un hogar particular. En el caso de una cantidad mayor a 6 miembros se consideró el hogar como colectivo, y quedó fuera del estudio." (INEC, 2015)

La encuesta cubre todo el territorio nacional de Costa Rica, segmentado en 14 dominios de estudio: Central Urbano Bajo, Central Urbano Medio, Central Urbano Alto, Central Rural, Chorotega Urbano, Chorotega Rural, Pacífico Central Urbano, Pacífico Central Rural, Brunca Urbano, Brunca Rural, Huetar Caribe Urbano, Huetar Caribe Rural, Huetar Norte Urbano y Huetar Norte Rural. (INEC, 2015)

El diseño muestral de la ENIGH 2013 corresponde a un diseño probabilístico de áreas, estratificado, bietápico y replicado. El marco muestral de viviendas se construyó a partir de la información de los censos nacionales de población y vivienda del 2011, compuesto por 10 381 UPM, y un total de 1 360 055 viviendas. Se visitaron 468 UPM y el tamaño de la muestra utilizado fue de 7 020 viviendas. (INEC, 2015)

5.1 Descripción de las variables

La técnica de regresión se plantea con la finalidad de relacionar la cantidad de dinero que los hogares costarricenses gastan en alimentos y bebidas no alcohólicas como variable dependiente del ingreso neto de los hogares y otras de índole socioeconómica que plantea la literatura.

5.1.1. Variable dependiente

Gasto mensual en alimentos y bebidas no alcohólicas consumidas en el hogar: Este gasto está conformado por la suma del gasto de pan y cereales, carne, pescado, leche-queso-huevos, aceites y grasas, frutas, vegetales, dulces, otros alimentos, café-té-cacao y bebidas no alcohólicas.

5.1.2. Variables independientes

Ingreso monetario corriente neto del hogar sin valor locativo*:* Esta variable corresponde a la suma del total del salario neto monetario del hogar (salario de ocupación principal con deducciones de ley, ganancias autónomas, renta y alquileres, transferencias y otros ingresos por trabajo) recibido en dinero y no incluye el valor locativo (imputación del ingreso mensual para las viviendas propias).

Número de miembros del hogar: Es una variable discreta y denota el número de miembros que residen en el hogar, excluyendo a las personas que se identifican como pensionistas y empleados domésticos. Puede incluir además del jefe de familia a: esposa(o) o compañera(o), hijo(a) o hijastro(a), yerno o nuera, nieto(a), padre o madre, suegro(a), hermano(a), cuñado(a), otro familiar y otros no familiares.

Edad del jefe: Variable discreta que toma la edad en años cumplidos del jefe del hogar.

Escolaridad del jefe: Variable discreta que representa los años de escolaridad del jefe del hogar, toma valores de entre 0 y 21 años.

Sexo del jefe: Variable nominal que denota el sexo del jefe del hogar. Toma los valores “Hombre” y “Mujer”.

Zona de residencia: Variable nominal que indica la zona de residencia. Toma los valores de “urbano” o “rural”.

5.2 Comportamiento del gasto en alimentación y su relación con otras variables

Parte de la importancia de estudiar el comportamiento del gasto en alimentación radica en que este informa sobre la capacidad para acceder a la canasta básica que poseen los hogares y también acerca de la composición, cantidad y calidad de la dieta. Se entiende cuáles preferencias de consumo de las familias están definidas como respuesta a su actividad maximizadora de

utilidad. Esto depende en cierto grado del ingreso, pues al aumentar el ingreso aumenta rápidamente el gasto en consumo. Además, juega un papel determinante en las preferencias alimenticias, individuales y familiares, pues abre la posibilidad de escoger entre mayor cantidad, calidad y variedad de alimentos. (Villezca Becerra & Máynez Cano, 2005)

Según Engel, la proporción del gasto empleada en alimentación está inversamente relacionada al ingreso total. Engel notó que existe una alta propensión de los hogares que experimentan ingresos crecientes en gastar una mayor proporción del presupuesto de alimentos en una dieta diversificada, mejorando así el estado nutricional de los miembros del hogar. (Babalola & Isitor, 2014)

Tambien existen otros factores de gran importancia que ejercen influencia cualitativa y cuantitativa en los patrones de consumo de los alimentos.

- Factores socioeconómicos tales como el tamaño de la familia, la edad de los miembros, el género, número de miembros con percepción monetaria, la escolaridad y la ocupación del jefe, permiten conocer las necesidades alimenticias de cada hogar (Villezca Becerra & Máynez Cano, 2005).
- Factores geográficos tales como la provincia, el cantón, la región, la zona de residencia, que permiten conocer las necesidades alimenticias asociadas al estilo de vida o hábitos culturales de consumo (García & Grande, 2010).

El **gasto en consumo de alimentos** ha sido bastante estudiado en relación a la variable **Ingreso**. Sin embargo, según la literatura recolectada, las variables socio-económicas y geográficas influyen cualitativa y cuantitativamente en los patrones de consumo pues reflejan desplazamientos en los gastos, debidos al ciclo de vida del jefe o los miembros (**edad del jefe(a)***), diferencias en la accesibilidad de los productos, diferencias en el clima (**zona de residencia***), cultura, gustos y preferencias (**años de escolaridad del jefe (a)***), estructura de las familias (**número de miembros del hogar y sexo del jefe (a)***). (Villezca Becerra & Máynez Cano, 2005)

Las variables encontradas en la Encuesta Nacional de Hogares que aproximan o representan el aspecto socioeconómico o geográfico que se desea incorporar en el modelo para caracterizar el gasto mensual en alimentos y bebidas no alcohólicas consumidas en el hogar comprenden:

- El ingreso monetario corriente neto del hogar.
- El número de miembros del hogar.

- Los años de escolaridad del jefe (a).
- La edad del jefe(a).
- El sexo del jefe (a).
- Y la zona de residencia.

5.3 Manejo previo de los datos

El archivo de datos se compone de 5705 registros de los cuales, al analizar previamente las variables, se descartó un registro por la ausencia de la variable años de escolaridad del jefe y otro por ingreso monetario negativo. Al final se utilizan 5703 casos para desarrollar el modelo.

Para la encuesta nacional de ingresos y gastos de los hogares se cuenta con el factor de expansión que corresponde a la ponderación de la muestra, necesario para generalizar los resultados a la población. Por lo tanto, cada uno de los modelos de regresión que se generen tomará en cuenta el factor de expansión como peso para la estimación de la siguiente manera:

(90)
$$Peso_i = \frac{FE_i}{\overline{\mathbf{FE}}}$$

Donde: FE corresponde al factor de expansión, $\overline{\mathbf{FE}}$ a la media aritmética de los FE_i. Esta es una forma de estandarizar los ponderadores, con el fin de que la muestra expandida tenga aproximadamente el mismo tamaño que la muestra original con la finalidad de no reducir artificialmente el error estándar.

5.4 Análisis descriptivo de las variables

- El gasto en alimentos presenta una distribución cuyo coeficiente de asimetría es de (2,126). Es una variable asimétrica hacia la derecha o positiva, lo cual indica que una pequeña cantidad de hogares destina sumas más altas a la alimentación. Se puede apreciar en el gráfico 1 de la **Figura 13** que el 80% de los casos se concentran debajo de ₡200 000; la presencia de valores en el gasto por arriba de ₡500 000 es apenas de un 1,5 %. La variable presenta 189 observaciones con un valor de 0, lo que equivale a un 3,3% de los registros que se utilizan en la estimación de los modelos. Si en la estimación de los parámetros no se contemplan estas observaciones, el método de mínimos cuadrados

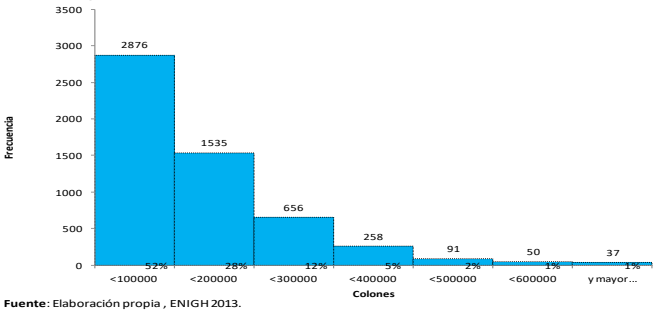
ordinarios o bien cualquiera de los otros métodos aquí presentados no rendirían estimaciones consistentes a causa del sesgo de selectividad, además de que se estaría perdiendo información. (Villezca Becerra & Máynez Cano, 2005)

- El ingreso monetario corriente neto del hogar presenta una distribución cuyo coeficiente de asimetría es de (8,794) el cual es mayor que en la variable Gasto. Es una variable asimétrica hacia la derecha o positiva, como se puede apreciar en el gráfico 2 de la **Figura 13**. Este a la vez indica que el 80% de los casos se concentran debajo de ₡ 1 000 000. El porcentaje de hogares que declara no tener ingreso es prácticamente nulo presentando un 0,2% de los casos. Por otra parte, la cantidad de salarios mayores a los ₡ 3 250 000 representa el 1,5% de los hogares entrevistados.
- La cantidad de miembros del hogar posee una distribución asimétrica cuyo coeficiente es de 0,825, un 80% de los hogares se componen de a lo sumo 4 miembros. Ver gráfico 3 de la **Figura 13.**
- Los años de escolaridad del jefe poseen una distribución multimodal con un coeficiente de asimetría de 0,437. Se destacan con mayor frecuencia los valores de 6, 9 y 11 años de estudio, correspondientes a la educación primaria completa (segundo ciclo), el tercer ciclo de la educación general básica y la educación secundaria o diversificada completa, cuya suma representa el 48% de los jefes de hogar. El 5% de los jefes de hogar manifiesta no poseer estudio y un 15% posee estudios después de finalizada la secundaria. Ver gráfico 4 de la **Figura 13**.
- La edad del jefe presenta una distribución aproximadamente simétrica con un coeficiente de 0,441; el 80% de los jefes de los hogares entrevistados se encuentra entre los 25 y 65 años. Ver gráfico 5 de la **Figura 13.**
- La zona de residencia del jefe se caracteriza por que el 65,8% de los jefes del hogar residen en la zona urbana y un 34,2% en la rural. Ver gráfico 6 de la **Figura 13.**
- Por último, la variable sexo del jefe, muestra que el porcentaje de hogares cuyo principal sostén es una mujer es del 34,5% mientras que el 65,5 % de los hogares entrevistados poseen un jefe de familia hombre. Ver gráfico 7 de la **Figura 13.**

La figura adjunta (**Figura 13**) ilustra la distribución de cada variable:

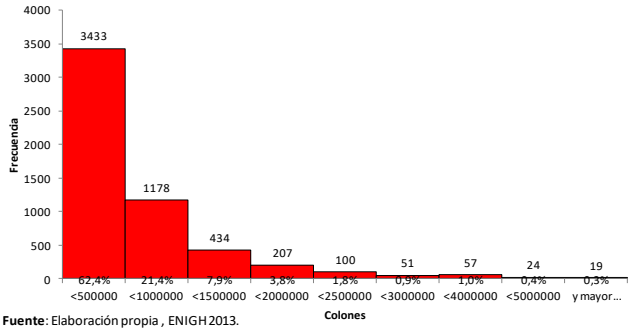
Figura 13: Histogramas de frecuencias para el gasto en consumo de alimentos y las variables explicativas asociadas.

1)



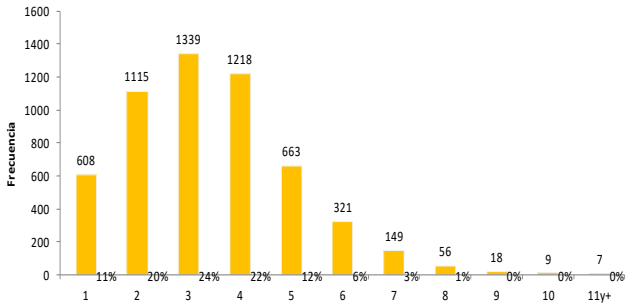
Fuente: Elaboración propia , ENIGH 2013.

2)



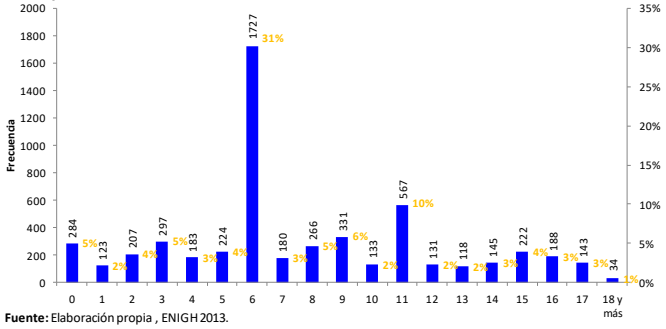
Fuente: Elaboración propia , ENIGH 2013.

3)



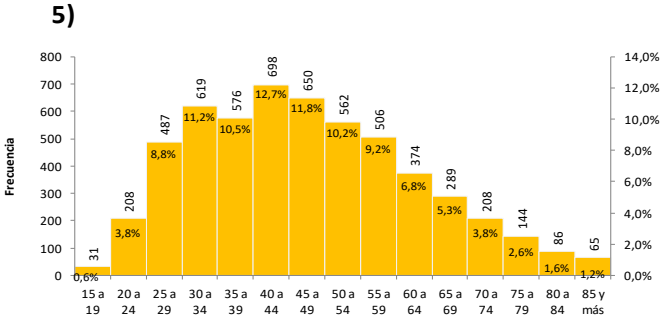
Fuente: Elaboración propia , ENIGH 2013.

4)



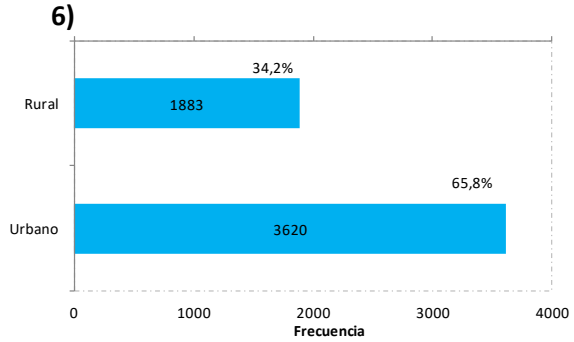
Fuente: Elaboración propia , ENIGH 2013.

5)



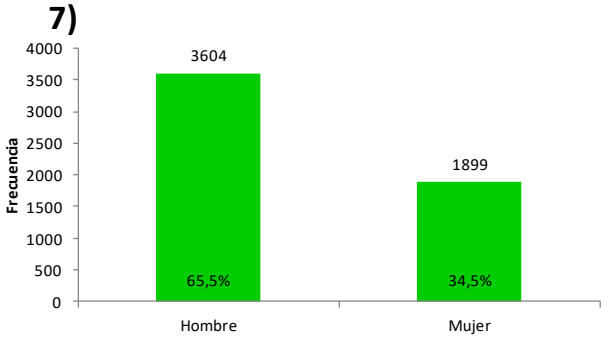
Fuente: Elaboración propia , ENIGH 2013.

6)



Fuente: Elaboración propia , ENIGH 2013.

7)



Fuente: Elaboración propia, ENIGH 2013.

5.5 Modelo de regresión por mínimos cuadrados ordinarios

5.5.1 Selección de las variables

Mediante las técnicas de selección de variables paso a paso, hacia adelante y hacia atrás, se determina que las variables más relevantes del conjunto evaluado en la explicación del gasto en alimentación de los hogares, utilizando el criterio de información de Akaike, son las siguientes:

- Ingreso monetario corriente neto del hogar.
- Número de miembros del hogar.
- Años de escolaridad del jefe (a).
- Edad del jefe (a).

El siguiente Cuadro (**Cuadro #16**) muestra el modelo de regresión estimado mediante mínimos cuadrados ordinarios utilizando las variables antes mencionadas y sin haber procedido a la evaluación de supuestos:

Cuadro # 16: Modelo de regresión por mínimos cuadrados ordinarios

Modelo de regresión por mínimos cuadrados ordinarios				
Error estándar residual	111500	**Estadístico F:**	292,6	
R cuadrado	0,1704	**Grados de libertad**	4	5698
R cuadrado ajustado	0,1698	**Valor p**	<2,2e-16	

	Estimación	**Error estándar**	**Valor t**	**Significancia**
Intercepto	-23598,0180	8022,3540	-2,9420	**
Ingreso(miles)	26,4580	1,7370	15,2350	***
Miembros	20969,8220	941,7100	22,2680	***
Edad	632,2250	100,9680	6,2620	***
Escolaridad	3641,6350	396,8260	9,1770	***

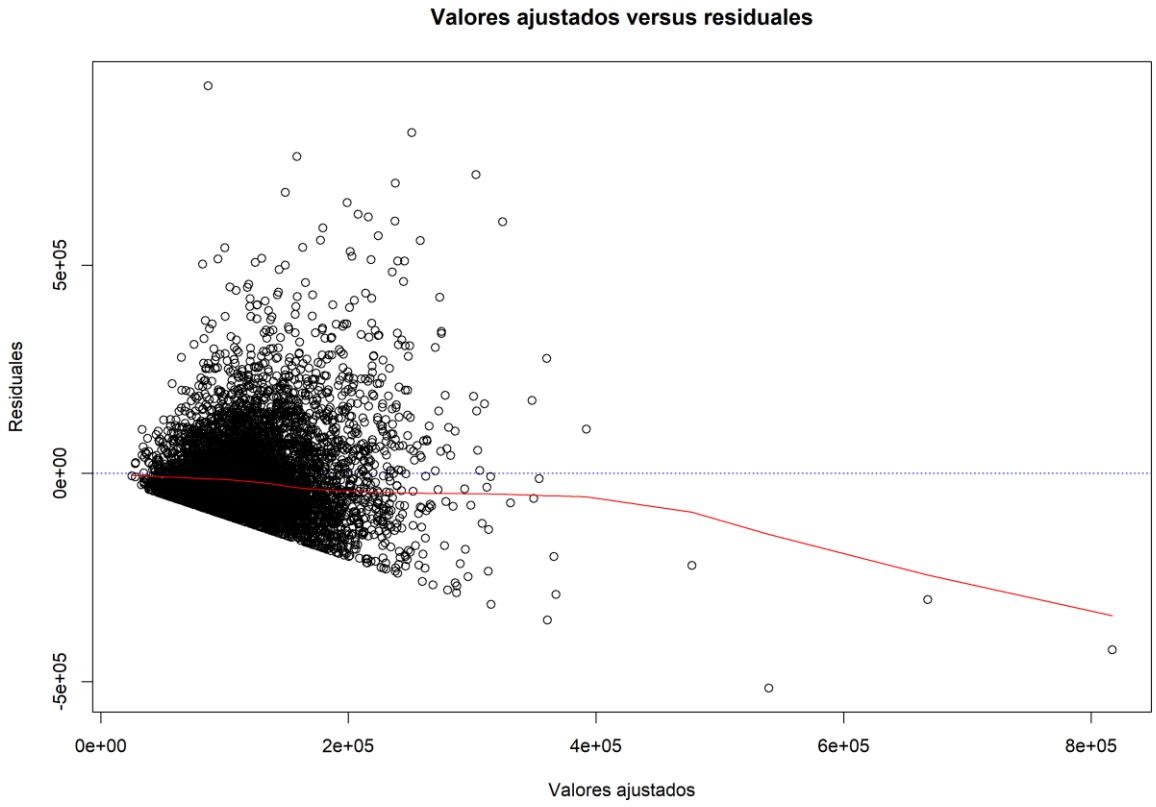
Códigos de significancia	0 '***' 0,001 '**' 0,01 '*' 0,05 ',' 0,1 ' ' 1

5.5.2 Linealidad de la variable respuesta

Para explorar la posible falta de linealidad se acude graficar los valores observados contra los residuales. La recta estimada a partir de los puntos no presenta una tendencia lineal pues parece

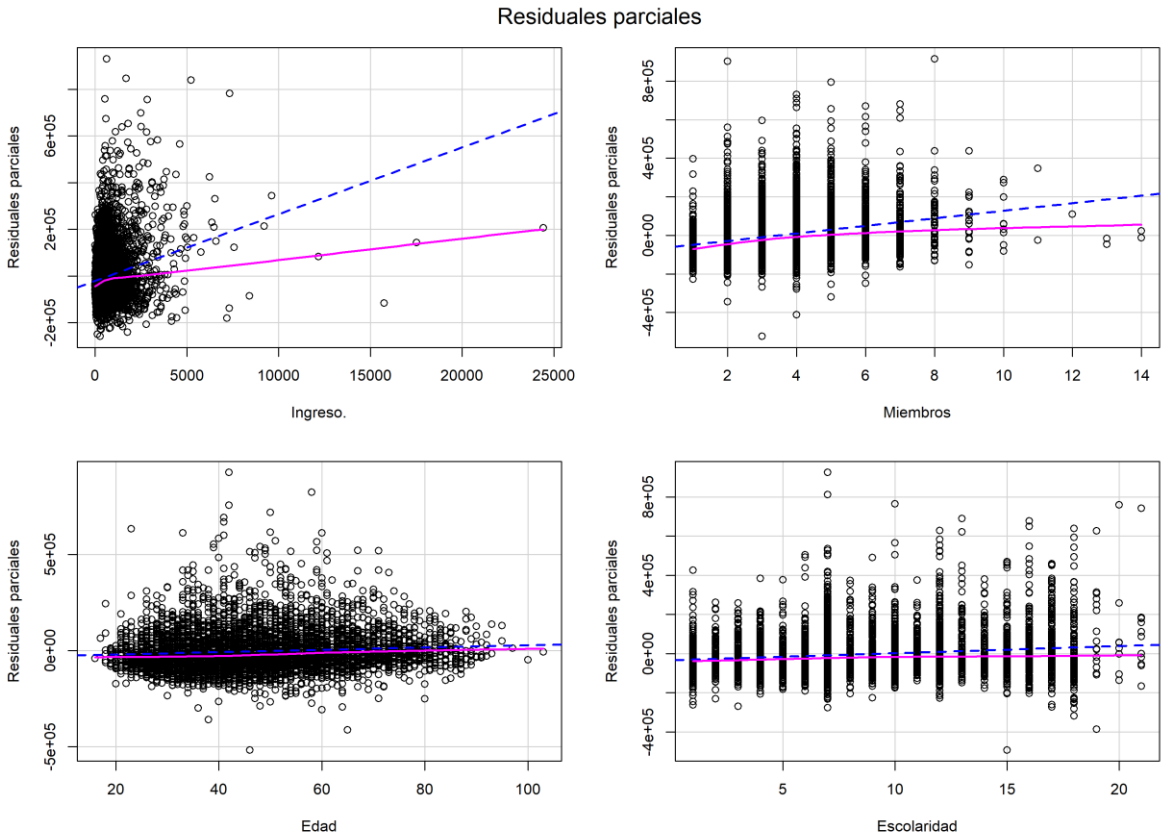
haber gran influencia de algunas observaciones en el modelo, tal y como se aprecia en la línea suavizada del gráfico de valores ajustados contra los residuales (**Figura 14**).

Figura 14: Valores ajustados versus residuales MCO.



Se puede observar en la gráfica de residuales parciales (**Figura 15**) que la influencia de los puntos tiene una mayor relación con la variable Ingreso pues, para los demás componentes del modelo, la relación no parece faltar a la linealidad. Esta violación al supuesto de linealidad será tomada en cuenta más adelante en este documento.

Figura 15: Residuales parciales MCO.



5.5.3 Multicolinealidad

En la matriz de correlaciones **(Cuadro #17)** se observa que el gasto correlaciona de manera moderada con el ingreso y la cantidad de miembros del hogar, también lo hace de manera leve con la educación del jefe y de manera muy débil con la edad. La matriz se detalla seguidamente:

Cuadro # 17: Matriz de correlaciones

Matriz de correlaciones					
	Gasto	**Ingreso**	**Miembros**	**Edad**	**Escolaridad**
Gasto	1				
Ingreso	0,3042	1			
Miembros	0,2728	0,1026	1		
Edad	0,0086	0,0128	-0,1343	1	
Escolaridad	0,1760	0,4518	-0,0853	-0,2513	1

Fuente: Elaboración propia

El factor de inflación de la varianza (**Cuadro #18**) no muestra evidencia fuerte de multicolinealidad, pues la mayoría de los valores son cercanos a 1. El valor unitario se interpreta como que no existe

multicolinealidad entre el k-ésimo predictor y los que se mantienen en el modelo, los valores mayores a 4 implicarían una multicolinealidad fuerte y mayores a 10 una multicolinealidad severa.

Cuadro # 18: Factor de inflación de la varianza

Factor de inflación de la varianza			
Ingreso	**Miembros**	**Edad**	**Escolaridad**
1,348168	1,077898	1,12064	1,421912

5.5.4 Valores influenciales

Al analizar detalladamente los valores influenciales (ver **Cuadro #19**): Df-betas, Df-fits, distancia de Cook y Leverages, se determina un total de 16 observaciones que se considera influyen en la calidad general del modelo, debido a que exceden los límites considerados seguros en 2 ó más de las características analizadas.

Cuadro # 19: Análisis de valores influenciales, mínimos cuadrados ordinarios

Análisis de valores influenciales

Tipo	Definición	Variable	Teoría indica	Estudio utiliza	Valores observados
Df-Betas	Los Df-betas miden cuanto ha afectado una observación la estimación de un coeficiente en particular, existe un Df-beta para cada coeficiente incluyendo el intercepto.	Ingreso	Df-betas > 2*√((k+1)/n) = 0,0265	Df-betas > 0,3	5702, 5701, 5703, 5310, 5671, 5697, 5694, 5700, 5696, 5679, 5664, 5660
		Intercepto			5310, 5701
		Miembros			5701, 5703
		Edad			5701
		Escolaridad			5701, 5702
Df-fits	Los Dffits miden cuanto cambian los valores estimados como resultado de obviar una observación	Todas	Df-fits > 2*√((k+1)/n) Df-fits > 0,065	Df-fits > 0,4	5702, 5701, 5703, 5697, 5694, 5700, 5671, 5310,5679, 5696,5879,5664, 5660
Distancia de cook	La distancia de Cook mide cuanto cambia la recta de regresión (coeficientes y valores ajustados) si cierta observación no es incluida en el análisis	Todas	D.Cook >4/(n-k-1) D.Cook >0,0007	D.Cook > 0,03	5700, 5703, 5697, 5671, 5310, 5701, 5702
Outliers	Generalmente los outliers son tomados como aquellas observaciones cuyos residuales estudentizados son mayores a 3	Todas	Residual estudentizado >3	Residual estudentizado >7 Residual estudentizado <-5	3309, 5310 5702, 5701
Leverages	Los leverages miden que tan inusual es una observación de otras en términos de las variables independientes	Todas	Leverages> 2(k+1)/n	Leverages>0,03	5702, 5701, 5703, 5696, 5681, 5698, 5699, 5700
Total, n=16	**3309, 5310, 5660, 5664, 5671, 5679, 5681, 5694, 5696, 5697, 5698, 5699, 5700, 5701, 5702, 5703.**				

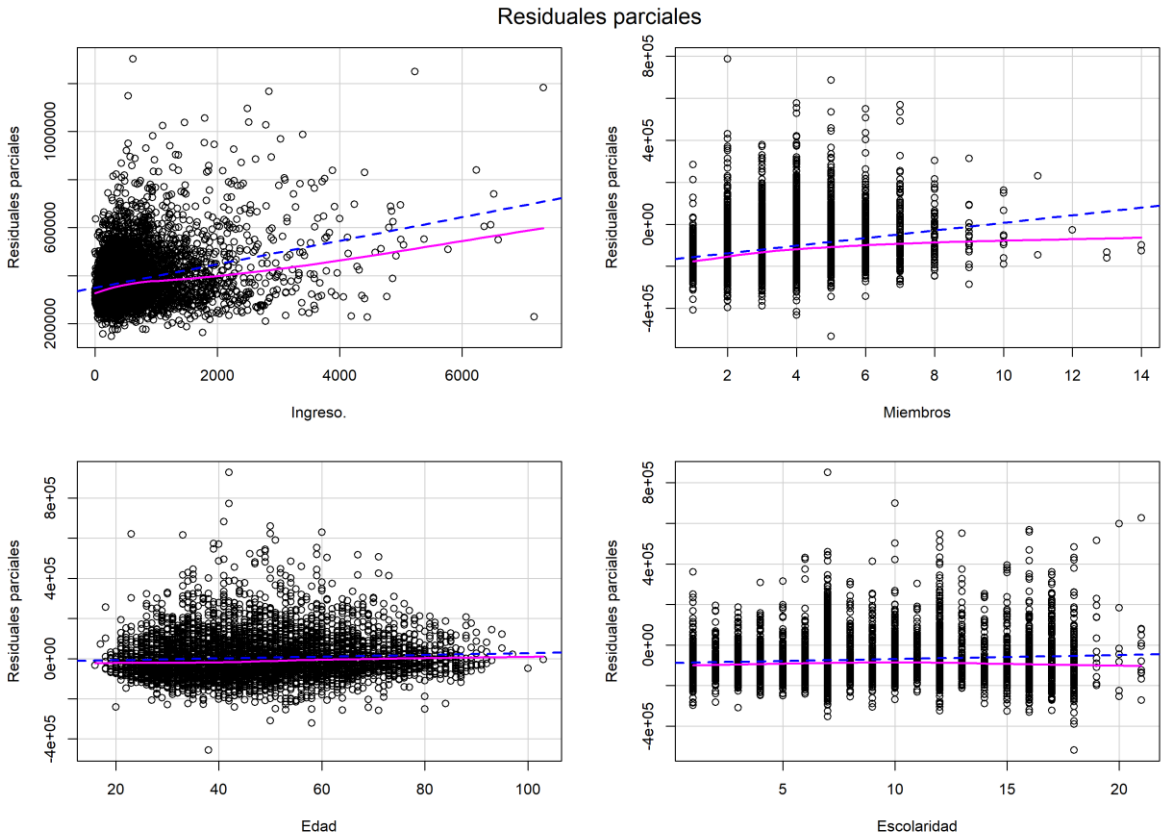
Se presenta a continuación el modelo resultante al excluir los valores de influencia (**Cuadro #20**):

Cuadro # 20: Comparación de modelos de regresión por mínimos cuadrados ordinarios

Comparación de modelos de regresión por mínimos cuadrados ordinarios						
	Modelo con toda la muestra			Modelo excluyendo valores influyentes		
	Estimación	Error estándar	Significancia	Estimación	Error estándar	Significancia
Intercepto	-23598,0180	8022,3540	**	-7266,8190	7825,9230	
Ingreso(miles)	26,4580	1,7370	***	45,9010	2,3730	***
Miembros	20969,8220	941,7100	***	18738,0090	920,6290	***
Edad	632,2250	100,9680	***	493,4170	98,0410	***
Escolaridad	3641,6350	396,8260	***	1972,6850	406,7040	***
Códigos de significancia	'***' p<0,001 '**' p<0,01 '*' p<0,05 ',' p<0,1 ' ' p<1					

Se observan cambios en el error estándar residual pues se reduce de 111500 a 107400, el R cuadrado aumenta de 0,1704 a 0,1877, el intercepto deja de ser significativo y las estimaciones de los coeficientes y errores estándar también varían. Por otra parte, el grado de correlación entre las variables no afecta los resultados del modelo (supuesto de multicolinealidad) y la linealidad mejora visualmente, el siguiente gráfico (Figura 16) así lo evidencia:

Figura 16: Residuales parciales MCO final



5.5.5 Normalidad de los residuos

La normalidad en los residuales es un importante requisito para obtener un modelo de regresión válido, pues implica que los errores residuales son ruido blanco y que el modelo lo ha capturado. Para analizar la normalidad del modelo se recurre a la prueba de Kolmogórov-Smirnov (**Cuadro #21**) y a los gráficos tales como el histograma y el QQ-plot (**Figuras 17 y 18**).

Cuadro # 21: Prueba KS para una muestra MCO

Prueba Kolmogorov-Smirnov para un muestra	
D = 0,61104	**p-value** = 2.2e-16

Según la prueba de Kolmogorov-Smirnov se rechazará la hipótesis nula de normalidad, pues el estadístico D posee una probabilidad asociada a la prueba por debajo de $\alpha = 0{,}05$. Lo anterior se explica de mejor manera si se observa que la distribución de los residuales estudentizados, presenta asimetría positiva, como evidencian el histograma y el 'QQ plot':

Figura 17: Histograma de los residuales estudentizados MCO

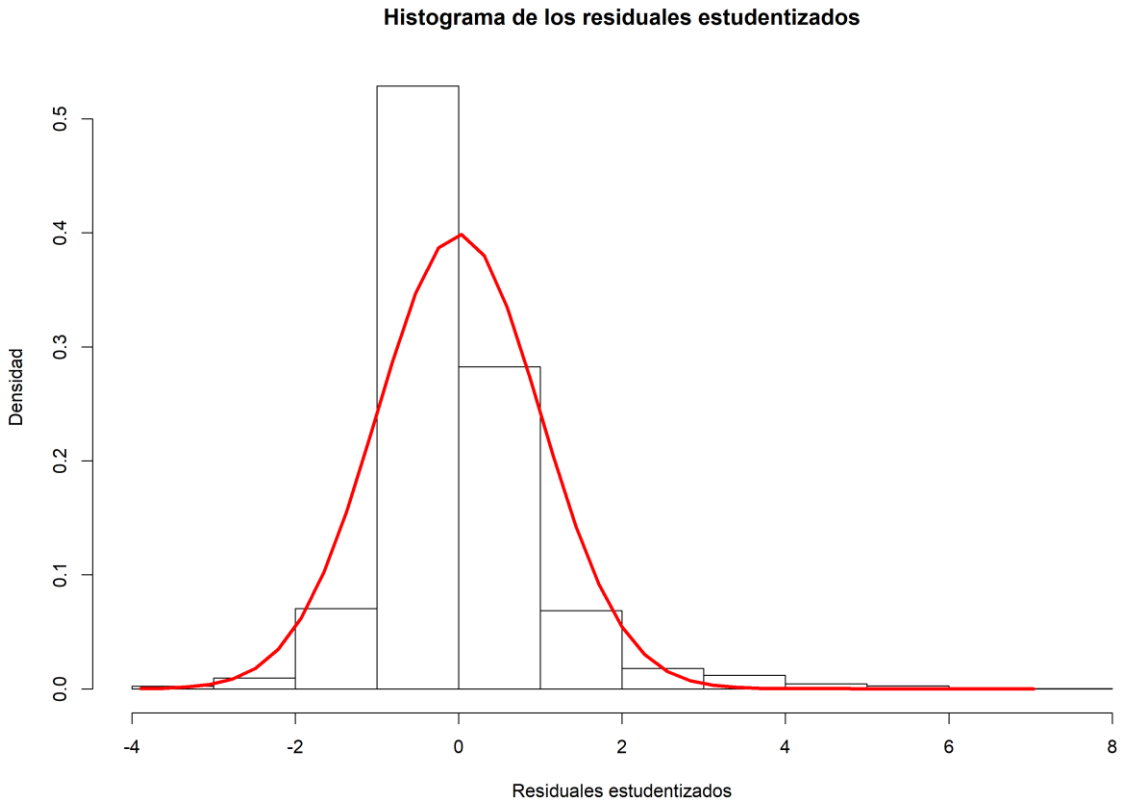
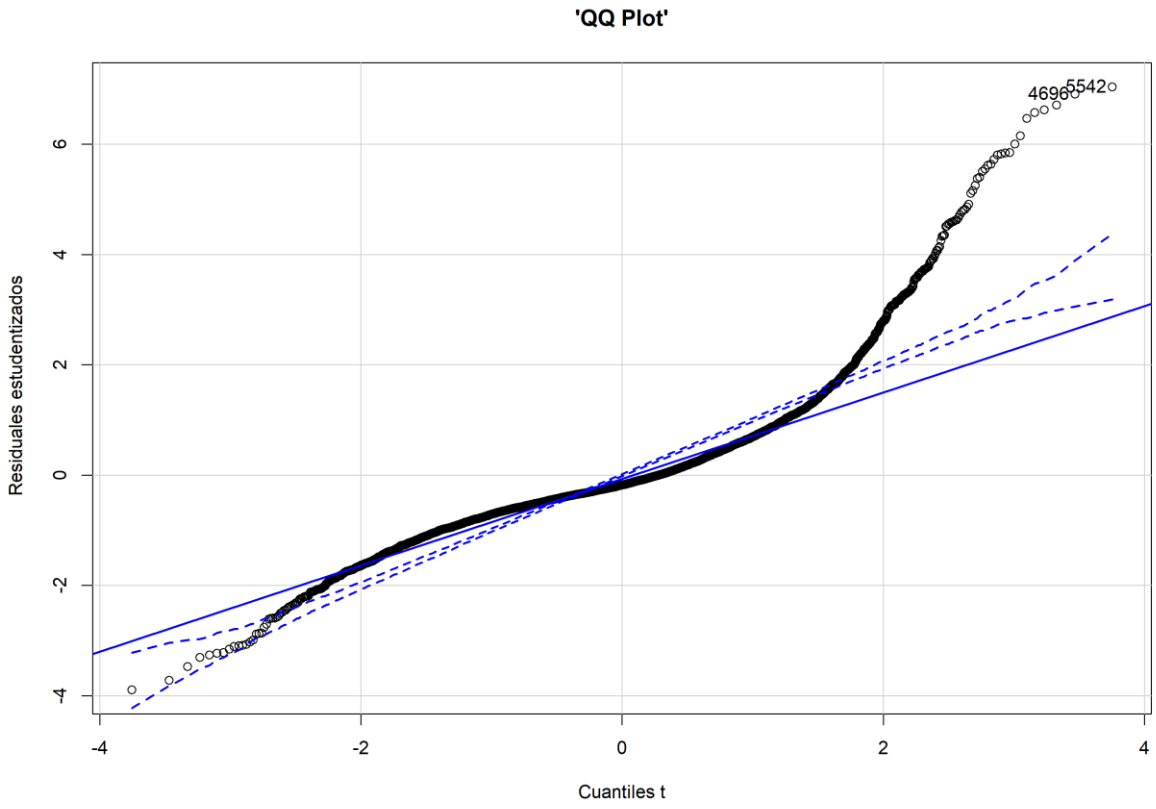


Figura 18: QQ-Plot MCO



Para tratar de obtener un mejor ajuste a este supuesto, se exploran las transformaciones de poder de Box-Cox. Sin embargo, los modelos para las transformaciones de poder sugeridas por el método de Box-Cox no alcanzan los estándares de normalidad que los gráficos de probabilidad normal y la prueba de Kolmogorov-Smirnov requieren. Varios de los autores consultados entre ellos Brady y Barber (1948), Stuvel y James (1950), Davis (1982), Dawoud (2013), Babalola e Isitor (2014), Vargas y Elizondo (2015), proponen la transformación logarítmica para las variables de Gasto y de Ingreso, sobre todo para lograr su interpretabilidad en términos relativos (elasticidades).

El supuesto de normalidad de los residuales es necesario pues de él se derivan con facilidad las distribuciones de probabilidad de los estimadores de mínimos cuadrados, ya que estos últimos son funciones lineales de los primeros. Si se trabaja con muestras pequeñas, la falta a la suposición de normalidad afecta las pruebas estadísticas t, F, χ^2 para los modelos de regresión. Cuando el tamaño de muestra es suficientemente grande, el supuesto se puede flexibilizar, gracias al teorema del límite central pues este ayuda a deducir que los estimadores seguirán teniendo las distribuciones

teóricas. Por último, también los estadísticos t y F tienden a la distribución normal conforme aumenta el tamaño de la muestra, por lo que aún estos serían válidos (Gujarati, 2010, pág. 318).

En el presente trabajo se desea analizar un modelo en el que los coeficientes se interpreten en términos propios de la unidad de medida original de los datos, en lugar de en términos relativos. Si a esto se le suma que la transformación de la variable respuesta no garantiza la normalidad y que se trabaja con un tamaño de muestra grande, entonces se puede preferir mantener la parsimonia del modelo.

5.5.6 Igualdad de varianzas

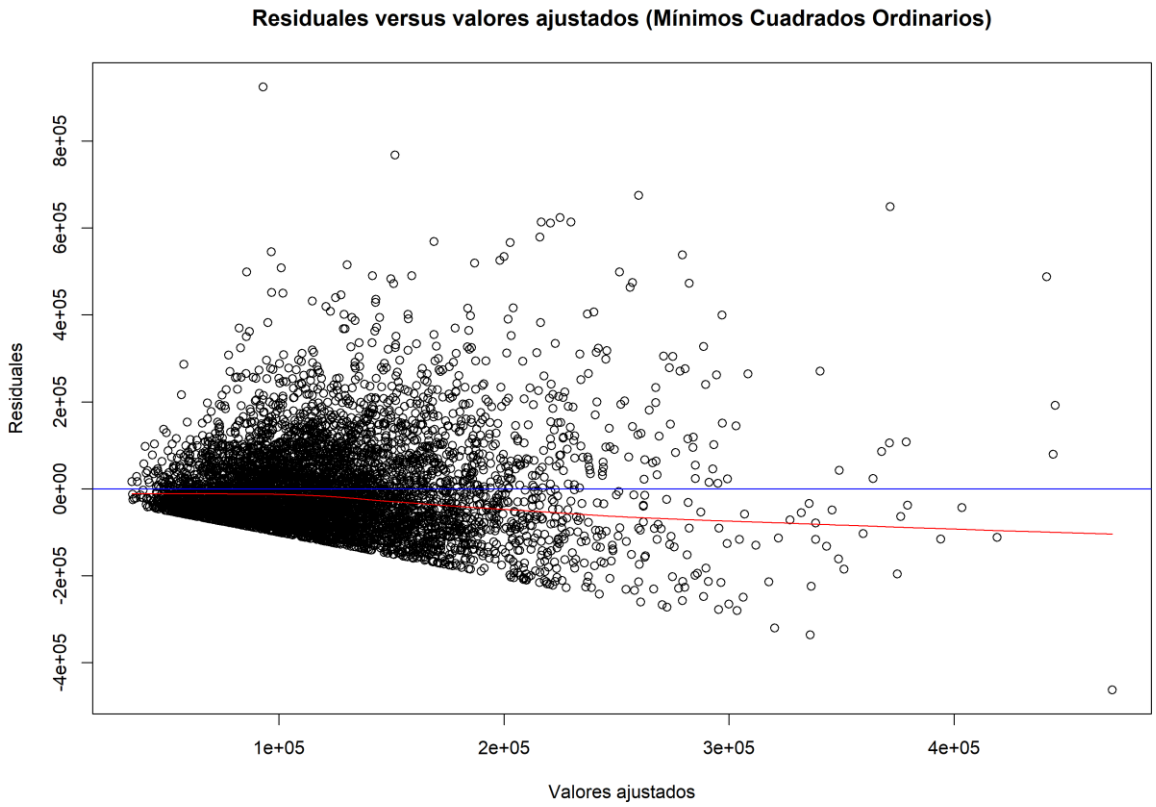
Para analizar las varianzas se utiliza la prueba de varianza no constante de Breusch-Pagan (**Cuadro #22**) y el método gráfico. La prueba formal asume en su hipótesis nula igualdad de varianzas residual y en su hipótesis alterna postula que la varianza residual cambia con respecto a los valores ajustados o bien con respecto a una combinación lineal de los predictores. Al observar la prueba se concluye que existe suficiente evidencia estadística para rechazar la hipótesis nula de igualdad de varianzas de los errores, manejando un nivel de significancia $\alpha = 0{,}05$.

Cuadro # 22: Prueba NCV para una muestra MCO final

Prueba de varianza residual no constante para una muestra			
Prueba	**Chi-Cuadrado**	**GL**	**p-value**
NCV test	1.635,0380	1	0,0000000
Breusch–Pagan test	325,728	4	3,05E-69

Por su parte la gráfica de residuales estudentizados versus valores ajustados (**Figura 19**) sugiere también presencia de **heterocedasticidad**, puesto que no se aprecia una uniformidad horizontal en la distribución, sino más bien que la dispersión aumenta conforme los valores ajustados (dispersión tipo embudo):

Figura 19: Residuales versus valores ajustados MCO final



Se concluye que, debido a la presencia de **heterocedasticidad** en el modelo, es necesario utilizar métodos u otros modelos que reduzcan el efecto de la misma en los errores estándar de los coeficientes con el fin de realizar inferencias fidedignas.

Nota: En el presente estudio el caso empírico no se corrige el modelo de regresión por la procedencia de la muestra de un diseño complejo, porque las conclusiones del mismo referentes a exactitud y precisión no cambian. En el siguiente cuadro (**Cuadro #23**) mediante el uso del paquete svyglm del software R, se corrige el modelo y se observa que pese a que el diseño complejo produce mayores errores estándar que el diseño irrestricto aleatorio simple, los valores del efecto del diseño son bajos.

Cuadro # 23: Comparación de modelos de regresión MCO, corregidos y no, por diseño muestral

Coeficientes	Modelo de regresión por mínimos cuadrados ordinarios (no corregido por diseño muestral complejo, n=5686)				Modelo de regresión por mínimos cuadrados ordinarios (corregido por diseño muestral complejo, n=5686)				Efecto del diseño
	Estimación	Error estándar	Valor t	Significado	Estimación	Error estándar	Valor t	Significado	Valor
Intercepto	-7266,819	7825,923	-0,929		-7.869,25	9.614,22	-0,82		1,509
Ingreso	45,901	2,373	19,346	***	45,61	3,95	11,53	***	2,776
Miembros	18738,009	920,629	20,353	***	18.824,49	1.188,73	15,84	***	1,667
Edad	493,417	98,041	5,033	***	502,22	112,85	4,45	***	1,325
Escolaridad	1972,685	406,704	4,850	***	1.969,72	525,46	3,75	***	1,669
Códigos de significancia	0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1								

5.6 Modelo de regresión lineal múltiple general

El modelo estimado mediante mínimos cuadrados ponderados (**Cuadro #24**) constituye una forma de corregir la **heterocedasticidad**, dejando intacta la estructura del modelo para la media. La siguiente ecuación muestra el modelo para la homogeneización de la varianza.

(91) $$\ln(\sigma_i^2) = 16{,}7525 + 0{,}3258 * Ingreso + 0{,}6910 * Miembros + 0{,}4353 * Escolaridad + 0{,}3131 * Edad$$

El anterior modelo contribuye a desarrollar el nuevo de mínimos cuadrados ponderados mediante la inclusión de sus valores ajustados como pesos ($1/\sqrt{\exp(\text{V. ajustados})}$). El modelo obtenido después de ello se muestra en la Cuadro #24:

Cuadro # 24: Modelo de regresión por mínimos cuadrados ordinarios ponderados

Modelo de regresión por mínimos cuadrados ordinarios ponderados

	Estimación	Error estándar	Valor t	Significancia
Intercepto	-7836,9400	6942,1600	-1,1290	
Ingreso	47,0000	2,5700	18,2850	***
Miembros	18902,7400	865,4700	21,8410	***
Edad	483,0600	85,7600	5,6330	***
Escolaridad	1816,3400	375,7500	4,8340	***
Códigos de significancia	'***' p<0,001 '**' p<0,01 '*' p<0,05 ',' p<0,1 ' ' p<1			

Al observar los estadísticos de las pruebas de varianza no constante y compararlos con los del modelo de mínimos cuadrados ordinarios, se evidencia en el modelo que se reducen sus valores y, en consecuencia, también la **heterocedasticidad**. Sin embargo, los mismos no alcanzan a ser significativos (**Cuadro #25**).

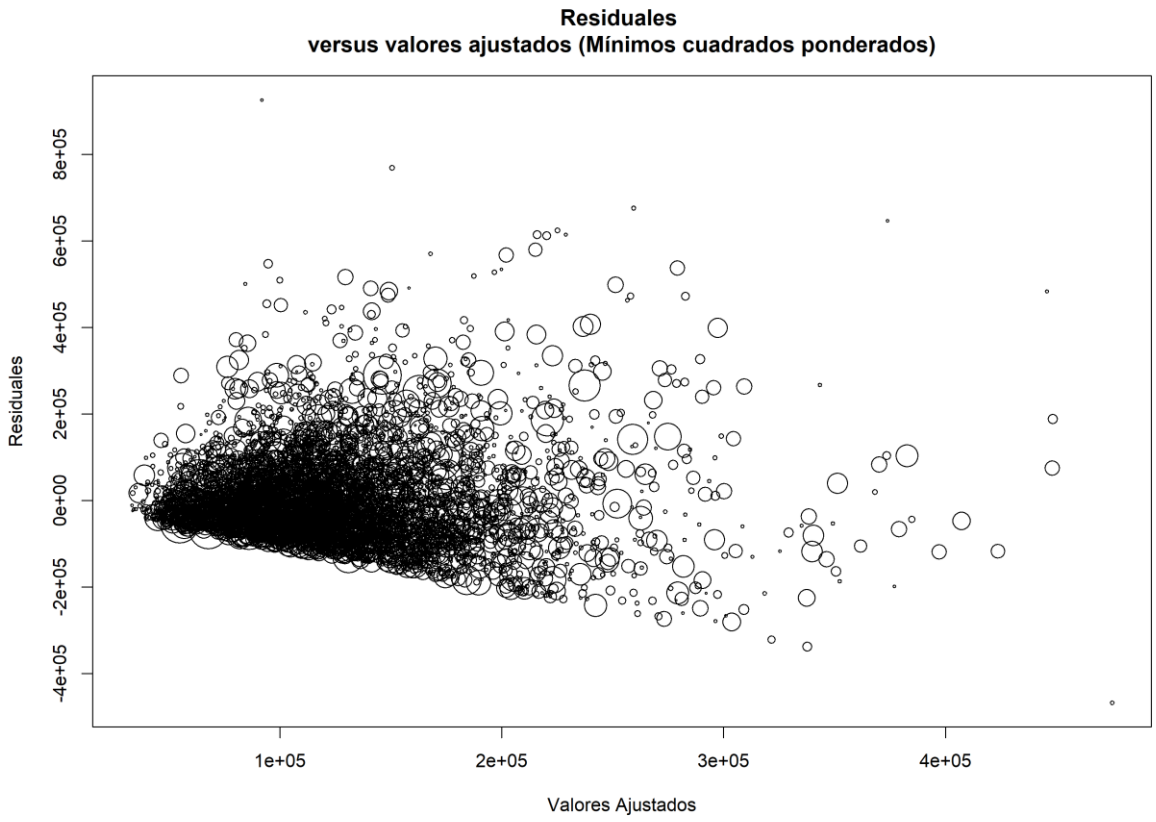
Cuadro # 25: Prueba NCV para una muestra MCP

Prueba de varianza residual no constante para una muestra (GLS)

Prueba	Chi-Cuadrado	GL	p-value
NCV test	513,2599	1	2,22e-16
Breusch–Pagan test	312,814	4	1,86E-66

Para evidenciar esta reducción en la prueba gráfica se utiliza el tamaño de las ponderaciones para representar cada punto de residuales estudentizados y valores ajustados en la gráfica (Figura 20).

Figura 20: Residuales versus valores ajustados **GLS**



5.7 Modelo de regresión cuantílica

Este tipo de regresión no solo brinda un valor mediano condicional del gasto en alimentación dadas las variables explicativas, sino que puede mostrar para cada percentil del gasto cual es la recta de regresión correspondiente. Cuando se está frente a un modelo homoscedástico es de esperarse que las pendientes se mantengan paralelas entre sí. El siguiente Cuadro (**Cuadro #26**) muestra que las pendientes para cada variable explicativa en este caso de estudio van en aumento para cada decil del gasto:

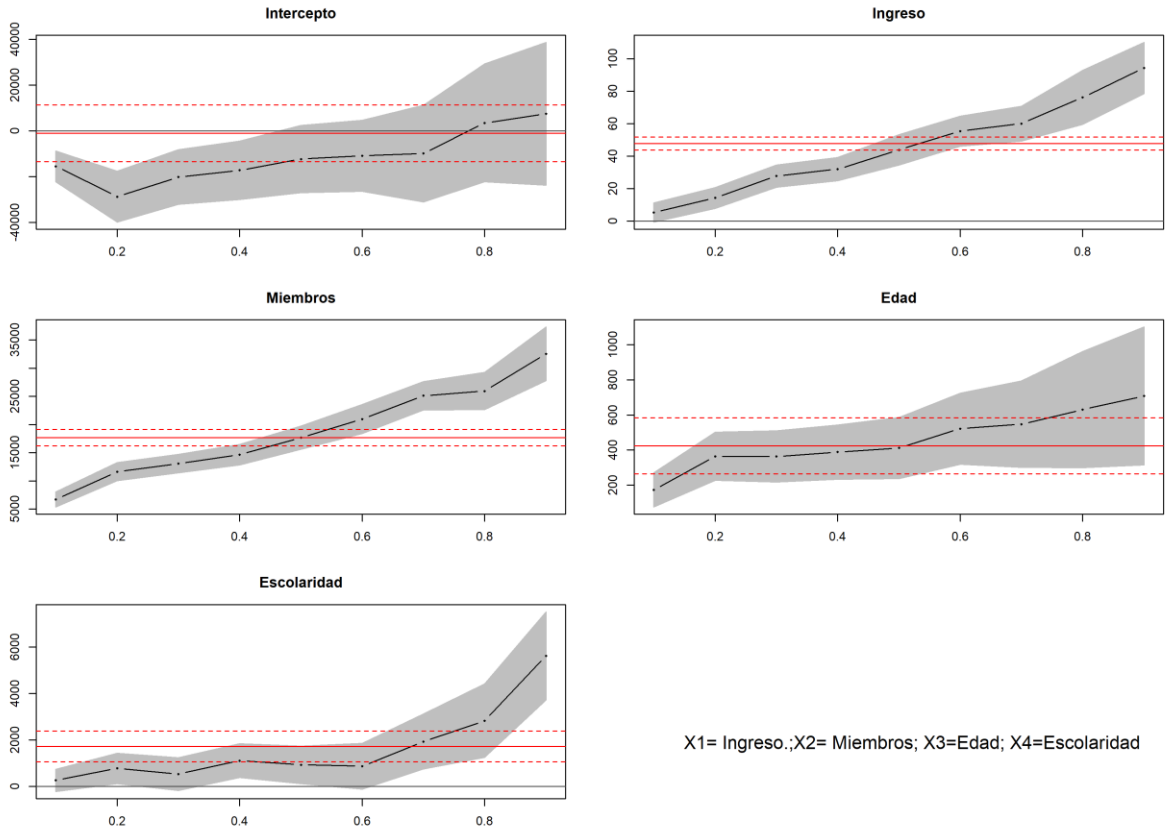
Cuadro # 26: Coeficientes modelo de regresión cuantílica

Coeficientes de modelos de regresión cuantílica para cada decil									
	Decil 1	**Decil 2**	**Decil 3**	**Decil 4**	**Decil 5**	**Decil 6**	**Decil 7**	**Decil 8**	**Decil 9**
(Intercept)	-15470,1007	-28708,6791	-20108,7901	-17172,1445	-12274,1657	-10810,8317	-9866,5527	3503,9395	7471,4797
Ingreso	5,2038	14,1795	27,7203	31,9832	43,8877	55,3726	60,0641	76,2407	94,3913
Miembros	6705,3035	11615,7020	13071,5495	14646,0383	17667,0133	20955,4373	25112,8627	25944,7887	32547,4643
Edad	173,3317	364,6452	363,5609	388,3418	411,8708	522,1051	547,2660	630,4594	707,5491
Escolaridad	259,9232	779,8810	531,7010	1108,4416	930,7057	870,6893	1930,7030	2829,1495	5625,0974

El siguiente gráfico (Figura 21) ilustra en líneas rojas los coeficientes e intervalos de confianza del (5% y 95%) para la regresión de mínimos cuadrados ordinarios, representando la línea continua la estimación y la línea punteada los intervalos; la línea negra discontinua de trazo y punto simboliza

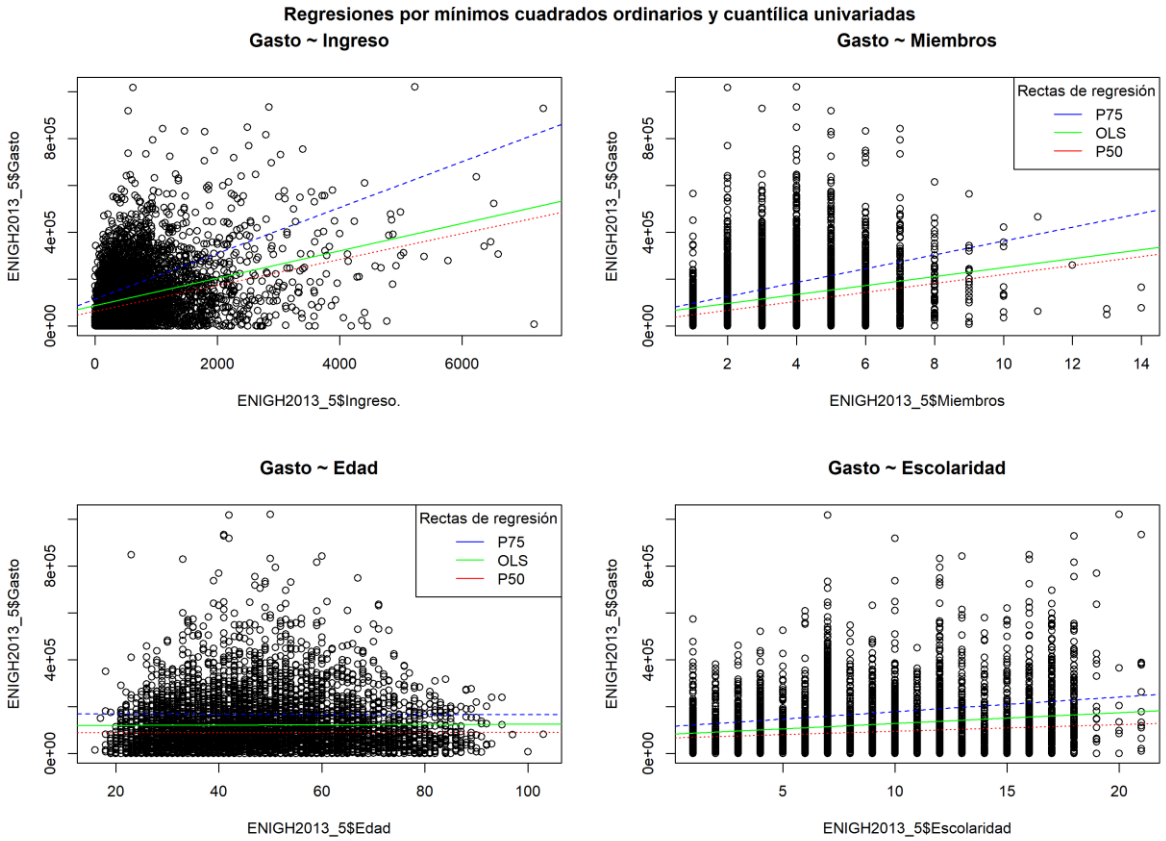
la evolución de la pendiente de regresión cuantílica para la variable explicativa(j) en cada percentil del gasto; la sombra en gris cómo evolucionan sus intervalos de confianza del (5% y 95%).

Figura 21: Coeficientes e intervalos de confianza OLS y QREG



En el mismo se aprecia que el ingreso, la cantidad de miembros y la escolaridad, poseen pendientes que difieren en mayor medida del análisis de mínimos cuadrados. La siguiente figura (Figura 22) ayuda a corroborar lo antes mencionado, representando mediante la línea roja de guiones, la recta de mínimos cuadrados ordinarios; la línea verde punteada, la recta de regresión cuantílica para la mediana del gasto; y las demás líneas continuas en gris los cuantiles: (5, 10, 25, 75, 90 y 95). De manera que se puede comprobar mediante la observación de las pendientes, que para la variable edad las mismas no presentan mayor variación.

Figura 22: Regresión univariada OLS y QREG



Debido a que la regresión cuantílica en teoría se ve menos afectada por los valores extremos, tal como refleja el anterior gráfico (Figura 22), al ser la mediana la medida de posición central más utilizada después de la media, como punto de comparación con los demás modelos se decide utilizar la recta estimada de regresión cuantílica para la mediana del gasto, el mismo se resume en el siguiente Cuadro (**Cuadro #27**):

Cuadro # 27: Modelo de regresión cuantílica, (tau=0,5)

Modelo de regresión cuantílica, tau= 0,5.

	Estimación	Error estándar	Valor t	Significancia
Intercepto	-12274,1657	8957,1367	-1,3703	
Ingreso	43,8877	5,6430	7,7774	***
Miembros	17667,0133	1263,3667	13,9841	***
Edad	411,8708	106,1209	3,8812	***
Escolaridad	930,7057	490,0514	1,8992	.
Códigos de significancia	0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1			

Si se compara este modelo (QREG) con el de (MCO), se aprecia que el coeficiente de escolaridad no es significativamente distinto de cero utilizando $\alpha = 0{,}05$, además, los errores estándar en general para todos los coeficientes son más grandes para el primero.

5.8 Modelo de regresión lineal doble generalizado

El modelo lineal doble generalizado permite describir la estructura de varianza de forma separada de la estructura para la media. Este tipo de modelo relaja los supuestos de la regresión lineal usual, pues la función de varianza se define heteroscedástica. Los coeficientes β_k tienen la misma interpretación que en la regresión usual, describiendo la diferencia promedio en la variable (Y) asociada a una unidad de cambio en la variable x_j. Los coeficientes de la varianza λ_j se interpretan como la diferencia en la log-varianza asociada a una unidad de cambio en z_j. El modelo utilizado para describir a los datos del Gasto en alimentación del hogar de la Encuesta Nacional de Hogares 2013 se detalla a continuación (**Cuadro #28**):

Cuadro # 28: Modelo de regresión lineal doble generalizado

Modelo de regresión doble generalizado				
Coeficientes para el modelo de la media				
	Estimación	**Error estándar**	**Valor t**	**Sig.**
(Intercepto)	-984,3718	6384,2278	-0,154	
Ingreso	58,8724	3,7320	15,775	***
Miembros	17482,5938	850,5639	20,554	***
Edad	431,5643	78,4859	5,499	***
Escolaridad	1288,5616	374,2363	3,443	***
Códigos de significancia	0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1			
Coeficientes para el modelo de la dispersión				
	Estimación	**Error estándar**	**Valor z**	**Sig.**
(Intercepto)	21,0043	0,1015	207,03	***
Ingreso	0,0007	0,0000	22,85	***
Miembros	0,1636	0,0119	13,78	***
Edad	0,0085	0,0013	6,56	***
Escolaridad	0,0576	0,0054	10,7424	***
Códigos de significancia	0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1			

En el caso del modelo de la varianza, las variables seleccionadas fueron nuevamente el ingreso, la cantidad de miembros la escolaridad y la edad. Se observa que la dispersión residual se ve afectada por todas las variables incorporadas, pues así lo revela la significancia de los coeficientes.

5.9 Modelo de regresión bayesiano lineal doble generalizado

La función de regresión doble generalizada puede desarrollarse en un marco bayesiano. El análisis bayesiano combina una verosimilitud normal con una distribución previa para β no informativa y una previa jerárquica para los coeficientes de varianza λ, tal y como se especifica en las ecuaciones 71 y 72. Se ejecutan tres cadenas MCMC con 6500 iteraciones cada una y descartan los primeros 1500 valores ("burn-in"), porque estos primeros valores dependen en gran medida de los valores iniciales elegidos y, por ende, no son una buena representación de la verosimilitud. Para reducir la correlación entre valores consecutivos en la cadena, solamente se guardan los valores cada cinco iteraciones y el resto se descarta ("adelgazamiento"). El modelo resultante se presenta en la **Cuadro #29.**

Cuadro # 29: Modelo de regresión lineal bayesiano doble generalizado

Modelo de regresión bayesiano doble generalizado					
Coeficientes para el modelo de la media					
		mu,vect	**sd,vect**	**2,50%**	**97,50%**
(Intercepto)	beta0	1473,381	992,947	-437,559	3477,98
Ingreso	beta1	62,694	4,513	54,062	71,524
Miembros	beta2	13480,197	679,592	12143,024	14811,561
Edad	beta3	787,885	53,967	683,806	896,945
Escolaridad	beta4	1908,167	340,113	1247,734	2558,993
Coeficientes para el modelo de la dispersión					
		mu,vect	**sd,vect**	**2,50%**	**97,50%**
(Intercepto)	lambda0	22,284	0,106	22,074	22,48
Ingreso	lambda1	0,001	0,000	0,001	0,001
Miembros	lambda2	0,214	0,013	0,19	0,239
Edad	lambda3	0,000	0,001	-0,003	0,002
Escolaridad	lambda4	0,015	0,005	0,005	0,026
(Intercepto)	tau,lambda0	0,002	0,003	0	0,01
Ingreso	tau,lambda1	494,213	685,728	0,563	2392,903
Miembros	tau,lambda2	20,858	30,225	0,026	107,732
Edad	tau,lambda3	496,779	681,6	0,376	2536,005
Escolaridad	tau,lambda4	447,346	609,644	0,467	2163,746

Este Cuadro muestra la media, la desviación estándar y los cuantiles de la distribución posterior marginal para cada uno de los parámetros del modelo. Además, el rango entre los cuantiles de 2.5% y 97.5% es el intervalo creíble del 95% para cada parámetro, por lo que existe una probabilidad del 95% de que el valor del parámetro verdadero se encuentre dentro de este rango.

5.10 Comparación modelos de regresión

Los predictores del gasto en consumo de alimentos del hogar que logran explicar mayor porcentaje de varianza son el Ingreso monetario corriente neto y el número de Miembros del hogar, seguidos por la Edad y los Años de escolaridad del jefe. Se puede rescatar además, para el modelo de la media (**Cuadro #30**), que conforme aumenta el valor de cada una de las variables independientes aumenta el valor del Gasto. Al comparar los errores estándar de los coeficientes y tomar como base el modelo de mínimo cuadrados ordinarios, tanto el modelo de mínimos cuadrados ponderados y el lineal doble generalizado frecuentista y bayesiano reducen los mismos. No así el de regresión cuantílica con tau= 0,5. Finalmente, en cuanto a la significancia de las variables, se observa que todas las variables explicativas son significativas ($\alpha = 0{,}05$) exceptuado los años de escolaridad del jefe en modelo cuantílico ($\alpha = 0{,}1$).

Cuadro # 30: Comparación de modelos para la media

Comparación de modelos para la media

	OLS	WLS	DGLM	BDGLM	QREG (0,5)
	Estimación	Estimación	Estimación	Estimación	Estimación
Intercepto	-7266,819	-7836,9400	-984,37177	1473,381	-12274,16571
Ingreso	45,901	47,0000	58,87242	62,694	43,88773
Miembros	18738,009	18902,7400	17482,59376	13480,197	17667,01333
Edad	493,417	483,0600	431,56425	787,885	411,87084
Escolaridad	1972,685	1816,3400	1288,56156	1908,167	930,70571
	OLS	WLS	DGLM	BDGLM	QREG (0,5)
	Error estándar	Error estándar	Error estándar	Error estándar	sd,vect
Intercepto	7825,923	6942,1600	6384,227806	992,947	8957,13671
Ingreso	2,373	2,5700	3,731965	4,513	5,64297
Miembros	920,629	865,4700	850,563854	679,592	1263,36667
Edad	98,041	85,7600	78,485878	53,967	106,12086
Escolaridad	406,704	375,7500	374,23629	340,113	490,05136
	OLS	WLS	DGLM	BDGLM	QREG (0,5)
	Pr(>|t|)	Pr(>|t|)	Pr(>|t|)	Pr(>|t|)	Pr(>|t|)
Intercepto	**				
Ingreso	***	***	***		***
Miembros	***	***	***		***
Edad	***	***	***		***
Escolaridad	***	***	***		.

Coeficientes para el modelo de la dispersión

	DGLM	DGLM	DGLM	DGLMB	DGLMB
	Estimación	Error estándar	Pr(>|z|)	mu,vect	sd,vect
(Intercepto)	21,0043	0,1015	***	22,284	0,106
Ingreso	0,0007	0,0000	***	0,001	0,000
Miembros	0,1636	0,0119	***	0,214	0,013
Edad	0,0085	0,0013	***	0,000	0,001
Escolaridad	0,0576	0,0054	***	0,015	0,005
Códigos de significancia	'***' p<0,001 '**' p<0,01 '*' p<0,05 ',' p<0,1 ' ' p<1				

Al comparar el modelo para la varianza del DGLM bayesiano y frecuentista se aprecia que tanto en estimaciones como en sus respectivos errores estándar son bastante similares; además todas las variables resultan ser significativas. Teóricamente estos últimos tipos de modelos presentan la mayor ventaja respecto a las otras técnicas evaluadas pues permiten a la distribución del error cambiar su varianza según las características de cada elemento u observación en la muestra y así ajustar de una mejor manera el comportamiento residual.

Respecto a las variables utilizadas para caracterizar el gasto en consumo de alimentos cabe mencionar que su relación con el ingreso ha sido bastante estudiada, pues se entiende que este

juega un papel determinante en las preferencias alimenticias, tanto individuales como familiares, ya que abre la posibilidad de escoger mayor cantidad, calidad y variedad de alimentos.

A la vez, la literatura recolectada y ya antes mencionada, también apunta a las características socio-económicas como un factor influyente en los patrones de consumo, pues reflejan cambios en los gastos que pueden estar asociados al ciclo de la vida del jefe (en este caso representado por la edad del jefe(a)), a gustos y preferencias de las personas (aquí ligados a los años de escolaridad del jefe) y a la estructura de las familias (capturada en el presente mediante el número de miembros del hogar).

Para lograr ofrecer una solución fiable de quien presenta mejor calidad de estimaciones y reducción de errores estándar es necesario considerar el comportamiento de cada modelo ante las diferentes causas de **heterocedasticidad,** ya demostrado durante el estudio de simulación**.**

CAPITULO 6: CONCLUSIONES

Las variables seleccionadas como predictoras contribuyen a explicar la variable de interés, que la teoría, otros estudios consultados y la experiencia obtenida del caso de estudio lo confirman. Se asociaron el gasto mensual en alimentos y bebidas no alcohólicas consumidas en el hogar con las variables independientes, en orden de importancia es de mayor a menor: primero con la variable Ingreso monetario corriente neto del hogar, luego con la cantidad de miembros del hogar, posteriormente la variable escolaridad del jefe y finalmente con la edad del jefe del hogar.

El estudio empírico presenta un caso típico de la violación del supuesto de homoscedasticidad residual en el modelo de regresión múltiple, pues manifiesta que cuanto mayor es el valor ajustado, mayor es la dispersión residual respecto a su propio valor medio. Para tratar con la violación al supuesto de normalidad de los errores y/o de homogeneidad de varianzas, frecuentemente en la práctica se emplea la transformación de la variable respuesta: Brady y Barber (1948), Stuvel y James (1950), Davis (1982), Dawoud (2013), Babalola e Isitor (2014), Vargas y Elizondo (2015). En específico cuando presenta asimetría de cola derecha, se pueden utilizar: logaritmos, raíces, y también recíprocos.

Si la transformación seleccionada es el logaritmo, una alternativa muy común para tratar la relación Gasto en Alimentación e Ingreso es: el doble logaritmo [ln(Y) = ln(X)]. Esta transformación permite interpretar elasticidades, entendiéndose lo que un 1% de cambio en la variable independiente (X) provoca de porcentaje de cambio en la variable dependiente (Y). Pese a las bondades que el enfoque de elasticidades presenta, los diferentes métodos de regresión fueron aplicados, con las unidades que el instrumento fue diseñado para medir y se reportan las mismas a los lectores. De esta manera se puede conocer qué porcentaje del Ingreso se utiliza en alimentación dentro del hogar y como varía este porcentaje con la incorporación de otras variables a la explicación del Gasto.

Entre las principales causas de este comportamiento, se pueden mencionar la naturaleza teórica del problema, la asimetría de cola derecha de las variables Gasto e Ingreso, la presencia de valores extremos en la variable Ingreso y finalmente que los residuales pueden estar almacenando información de una o más variables omitidas. Mediante los experimentos de simulación ejecutados, se trataron de controlar los posibles orígenes de **heterocedasticidad** y se obtuvo evidencia del grado de ajuste a un modelo teórico poblacional, previamente establecido para los distintos métodos en el parámetro de mayor asociación e interés: el Ingreso. Esto con el objeto de

precisar su comportamiento ante las diferentes causas, las cuales pueden incluso estar mezcladas en la práctica.

Para el caso en donde la **heterocedasticidad** es producida por la naturaleza de las variables, análisis previos como el realizado por (Olusola Et Al., 2016), en donde se trabaja con tamaños de muestra pequeños de entre 25 y 50 individuos, con distintos grados y orígenes de **heterocedasticidad**, provenientes de una variable o la interacción entre 2 variables explicativas en el modelo, establecen que el método OLS sobrestima el error estándar y el estimador **GLS** genera mejores resultados.

Otra investigación, llevada a cabo por (Vynck, 2016-2017), encuentra que el uso de **GLS** combinado con estimadores de la matriz de covarianza heteroscedasticamente consistentes (HCCME por sus siglas en inglés) a menudo reflejan un aumento de potencia y control del error tipo I respecto al enfoque clásico de emplear **GLS** para tratar la **heterocedasticidad**. Aunque se trate la **heterocedasticidad** mediante una ponderación de la manera más adecuada posible, el escenario ideal donde la función de varianza es conocida en general no es realista en la práctica.

En el ejercicio de simulación para el caso antes mencionado, las diferentes técnicas aciertan la estimación del coeficiente del Ingreso, pero la que presenta los mejores indicadores de desempeño y reduce en mayor cantidad el error estándar es el **modelo lineal doble generalizado**, seguido por el **modelo de mínimos cuadrados generalizados**.

En el escenario de **heterocedasticidad** generada por el sesgo de variable omitida, lo más recomendado en la práctica, de existir la disponibilidad, es identificar las variables importantes que se han dejado fuera y volver a ajustar el modelo con esas variables. Los resultados del estudio de simulación evidencian, para todos los métodos, que al existir variables omitidas no se estima correctamente el parámetro de la variable Ingreso, en concreto se sobreestima el mismo. En presencia de la variable omitida el **modelo de mínimos cuadrados generalizados** presenta los indicadores de desempeño menos desfavorables para el coeficiente del Ingreso y su error estándar, seguido por **modelo lineal doble generalizado**.

En el contexto de **heterocedasticidad** producida por valores extremos, Midi, Rana, & Imon (2009), desarrollan un método llamado mínimos cuadrados ponderados robustos (**RWLS** por sus siglas en inglés), con el fin de corregir el problema de errores heteroscedásticos en presencia de valores atípicos. Su estudio muestra que el método OLS y las estimaciones de **GLS** son fácilmente afectadas por los valores atípicos, por lo que las inferencias no son fiables. Los resultados del

mismo sugieren que las estimaciones de **RWLS** emergen ante estas circunstancias para ser visiblemente más eficientes y más confiables, ya que presentan los errores estándar más pequeños, y los valores t más grandes en comparación con el OLS y el **GLS**.

Finalmente, en este contexto para el ejercicio de simulación, los distintos métodos subestiman el parámetro del Ingreso. Respectivamente la técnica que mejor aproxima la correcta estimación del coeficiente y mantiene los errores estándar más bajos es el **modelo lineal doble generalizado**, mientras que el segundo en aproximarse al coeficiente y en reducir el error estándar de la estimación es la de **mínimos cuadrados generalizados.**

Las técnicas analizadas mediante el ejercicio de simulación presentan diferentes desempeños según el origen de la **heterocedasticidad**. Cabe mencionar que la regresión cuantílica y el planteamiento bayesiano propuesto no ofrecen mayores ventajas en cuanto a reducción del error estándar se refiere. Los resultados sugieren que el modelo DGLM y **GLS** muestran un buen desempeño general, sin embargo, no presentan una estimación exacta cuando existe **heterocedasticidad** generada a causa de valores extremos o bien ante la existencia de una variable omitida, por lo que, para futuros estudios, se plantea la posibilidad de incorporar técnicas tales como **RWLS** (Midi, Rana, & Imon, 2009) o HCCM (Vynck, 2016-2017), en el conjunto de modelos por evaluar.

Para el caso empírico, la teoría expuesta ayuda a inferir que la desigualdad de varianzas residual proviene principalmente de la asimetría de las variables económicas, pues la cantidad de valores extremos en el Ingreso representa una pequeña proporción de las observaciones. Además se trata de explicar la variable Gasto en alimentación con covariables que gran parte de la bibliografía apoya. Por su naturaleza, la variable Gasto mensual en alimentos y bebidas no alcohólicas consumidas en el hogar requiere una diversidad de variantes a nivel individual del jefe y del hogar para lograr ser explicado a cabalidad y no solamente de una variable adicional.

Por todo lo antes mencionado y por el desempeño mostrado en el estudio de simulación en cuanto a reducción de los errores estándar y eficaz estimación de los coeficientes, se decide utilizar el modelo DGLM para explicar dicho Gasto en el caso empírico. Este modelo ajusta la varianza de la distribución de cada residual generado por el modelo de la media acorde a las variables introducidas en el modelo de la dispersión. Esto permite obtener para cada individuo una mejor estimación del Gasto debido a que el término residual proviene de una distribución normal que permite diferentes varianzas para cada individuo, en vez de una que mantiene la misma para

todos. De este modo, emula el comportamiento real de los residuales del modelo y en consecuencia de las observaciones.

Según el (DGLM), el intercepto para el modelo de la media sería de -984 colones, de los cuales, en promedio, por cada 1 000 colones de aumento en el Ingreso monetario corriente neto del hogar se destinan 59 colones al Gasto en alimentación; por cada miembro adicional en el hogar. el monto destinado mensualmente al gasto en alimentación aumenta en promedio 17 483 colones; un año adicional en la edad del jefe(a) incrementa el gasto en promedio 431 colones y, finalmente, por cada año adicional de escolaridad del jefe(a), se destinan en promedio 1 288 colones adicionales al gasto en alimentación.

Acorde a este modelo y en términos sencillos, si un hogar posee un ingreso de 500 000 colones, su cantidad de miembros es de 2, el jefe posee una edad de 40 años y una cantidad de años de estudio de 11, el gasto en alimentación dentro del hogar se estima en 94 854 colones para ese hogar (o sea, alrededor del 19% de los ingresos a la alimentación). Mientras que un hogar con un ingreso de 1 000 000 de colones y las mismas características presentaría un gasto en alimentación de 124 290 colones, equivalente al 12,4% de sus ingresos. Por su parte el residual proveniente del primer caso pertenecería a una distribución Normal de media(μ) = 0 y log-varianza $(\sigma_i^2) = (22{,}68)$, mientras que en el segundo caso la log-varianza correspondería a: $(\sigma_i^2) = (23{,}06)$.

Bibliografía

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Hoboken, New Jersey: Wiley.

Arce, R., & Mahía, R. (03 de 2008). Recuperado el 24 de Noviembre de 2016, de http://tabarefernandez.tripod.com/dearce.pdf

Babalola, D. A., & Isitor, S. U. (2014). Analysis of the Determinants of Food Expenditure Patterns among Urban Households in Nigeria: Evidence from lagos States. *Journal of Agriculture and Veterinary Science , 7*, 71-75.

Borrego, C. A., & Carro, J. (Octubre de 2012). *Material de clase.* Recuperado el 22 de Junio de 2016, de Departamento de Economía Universidad Carlos III de Madrid , Fundamentos del Análisis Económico. Economía Aplicada- Oficina OCW OpenCourseWare: http://ocw.uc3m.es/economia/econometria/material-de-clase

Brady, D. S., & Barber, H. A. (1948). The Pattern of Food Expenditures. *The Review of Economics and Statistics , 30* (3), 198-206.

Bruno A., W., & Moore, J. (2005). The concepts of bias, precision and accuracy, and their use in testing. *Ecography 28* , 815-829.

Cade, B. S., & Noon, B. R. (2003). A gentle introduction to quantile regression for ecologist. *Frontiers in Ecology and the Environment* , 412-420.

Cepeda, E., & Gamerman, D. (2000). Bayesian modeling of variance heterogeneity in normal regression models. (I. o. Statistics, Ed.) *Brazilian Journal of Probability and Statistics , 14* (2), 207-221.

Correa, J. C. (2005). Una aproximación bayesiana al problema de heteroscedasticidad en el modelo lineal simple. *Revista Colombiana de Estadística , 28* (1), 17-21.

Davino, C. F. (2014). *Quantile Regression: Theory and Applications.* Oxford: John Wiley & Sons.

Davis, C. G. (1982). Linkages between socioeconomic characteristics, food expenditure patterns, and nutritional status of low income a critical review. *American Agricultural Economics Association , 64* (5), 1017-1025.

Dawoud, S. D. (2013). Econometric analysis of the changes in food consumption expenditure patterns in Egypt. *Journal of Development and Agricultural Economics , 6* (1), 1-11.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist , 63* (7), 591–601.

Feng, Y. G. (2002). *Heteroskedasticity.* Recuperado el 12 de Julio de 2016, de National Cheng Kung University: http://www.ncku.edu.tw/~account/chinese/course/eco91/lecture9.pdf

Fernández, T. (. (Junio de 2004). *Heterocedasticidad en el modelo de regresión lineal.* Recuperado el 15 de Julio de 2015, de Ficha No. 6.: http://tabarefernandez.tripod.com/hetero.pdf

García, T., & Grande, I. (2010). Determinants of food expenditure patterns among older consumers. The Spanish case. *Appetite , LIV* (1), 62-70.

Gelman, A., Carlin, J. B., & Stern, H. S. (2014). *Bayesian Data Analysis.* Boca Raton, Florida: Chapman & Hall.

Gentle, J. (2003). *Ramdom Number Generation and Montecarlo Methods.* Springer.

Geurts, J. A., Jansen, H. G., & Tilburg, A. v. (1997). *Domestic demand for food in Costa Rica: a double-hurdle analysis* (Vol. 286). Turrialba, Costa Rica: CATIE.

Green, W. H. (2012). *Econometric Analysis* (7 ed.). New York: Prentice Hall.

Gujarati, D. N. (2010). *Econometría* (4 ed.). (G. Arango Medina, Trad.) México D.F: McGraw Hill.

Habshah, M., Rana, S., & A.H.M.R, I. (Julio de 2009). *The Performance of Robust Weighted Least Squares in the Presence of Outliers.* Recuperado el Septiembre de 2016, de Department of Mathematical Sciences, Ball State University: http://www.wseas.us/e-library/transactions/mathematics/2009/29-388.pdf

Hayes, A. F. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression : An introduction and software implementation. *Behavior Research Methods , 39* (4), 709-722.

Hellander, A. (23 de Febrero de 2009). Stochastic Simulation and Monte Carlo Methods.

Houthakker, H. S. (1957). An International Comparison of Household Expenditure Patterns, Commemorating the Centenary of Engel's Law. *Econometrica , 25* (4), 532-551.

INEC. (8 de Enero de 2015). *Costa Rica, Encuesta Nacional de Ingresos y Gastos de los Hogares 2013.* Recuperado el 15 de Abril de 2016, de reporte generado del programa acelerado de datos Costa Rica: http://www.inec.go.cr/anda4/index.php

Johnson, N. (13 de Marzo de 2012). *Bayesian methods for regression in R.* Recuperado el 15 de Julio de 2016, de Department of Statistics at Virginia Tech Short Courses Bayesian Methods for Regression in R: http://www.lisa.stat.vt.edu/sites/default/files/Bayes%20Shortcourse%202012.pdf

Kindleberger, C. P. (1997). *Economic Laws and Economic History.* New York: Cambridge University Press.

Kleiber, C., & Zeileis, A. (2008). *Applied Econometrics with R.* New York: Springer Science Business Media.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical Confidence Intervals for Regression Quantiles. *Journal of Computational and Graphical Statistics , 14*, 41-55.

Koenker & Bassett, R. G. (1978). Regression quantiles. *Econometrica* , 33-50.

Koenker, R. (1994). Confidence Intervals for Regression Quantiles. En P. Mandl, *Asymptotic Statistics* (págs. 349-359). University of Illinois, Champaign-Urbana, USA: Springer-Verlag Berlin Heidelberg.

Koenker, R. (13 de Febrero de 2016). *Package 'quantreg'.* Recuperado el 1 de Mayo de 2016, de Quantile Regression: http://www.r-project.org

Koenker, R. (2005). *Quantile Regression.* New York: Cambridge University Press.

Koenker, R. (25 de Marzo de 2015). *Quantile regression in R: a vignette.* Recuperado el 18 de Julio de 2016, de The Comprehensive R Archive Network: https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf

Koenker, R., & Bassett, G. W. (1978). Regression quantiles. *Econometrica* , 33-50.

Koenker, R., & Hallock, K. F. (08 de Diciembre de 2000). *Quantile regression an introduction.* Recuperado el 15 de Junio de 2016, de University of Illlinois at urbana-champaing : http://www.econ.uiuc.edu/~roger/research/intro/rq3.pdf

Koenker, R., & Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association , 94* (448), 1296-1310.

Long, S. J., & Laurie, E. H. (22 de Septiembre de 1999). *Using Heteroscedasticity Consistent Standard Errors.* Recuperado el 01 de Julio de 2016, de http://www.indiana.edu/~jslsoc/files_research/testing_tests/hccm/99TAS.pdf

Martínez, I. S. (30 de Junio de 2010). *Trabajo fin de máster - Tema: Regresión Cuantil.* Recuperado el 8 de Mayo de 2016, de Universidad de Santiago de Compostela Departamento de Estadística e Investigación Operativa: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_404.pdf

Mhlongo, V., & Daniels, R. C. (2013). Food expenditure patterns in South Africa: Evidence from the NIDS. *NIDS Discussion Paper* (123), 27.

Midi, H., Rana, S., & Imon, R. (2009). The Performance of Robust Weighted Least Squares in the Presence of Outliers and Heteroscedastic Errors. *Wseas Transactions on Mathemathics , 8*, 351-361.

Muñoz, M. C. (2004). Determinantes del ingreso y del gasto corriente de los hogares. *Revista de economía institucional , 6* (10), 183-200.

Neter, J. M. (1990). *Models, Applied Linear Statistical* (4 ed.). Chicago: Irwin Publishing.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1990). *Applied Linear Statistical Models* (4 ed.). Chicago: Irwin Publishing.

Olusola, A. A., Obalowu, J., Titilayo, I. D., & Agboola, B. O. (Mayo de 2016). *Performances of Ordinary and Generalized Least Squares Estimators on Multiple Linear Regression Models with Heteroscedasticity.* (T. P. Technology, Ed.) Recuperado el Septiembre de 2016, de Department of Statistics, University of Ilorin, Ilorin, Nigeria & of Mathematics, Federal University Lafia, Nassarawa, Nigeria: http://www.akamaiuniversity.us/PJST17_1_68.pdf

Rencher, A. C., & Schaalje, B. (2008). *Linear Models in Statistics.* Hoboken, New Jersey: John Wiley & Sons.

Reuven, R. Y. (1981). *Simulation and the Monte Carlo Method.* Israel Institute of Technology: John Wiley & Sons.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2006). *Bayesian Statistics and Marketing.* The Atrium, Southern Gate, Chichester: John Wiley & Sons .

Sanford, W. (2014). *Applied Linear Regression.* Minneapolis, MN: Wiley.

Scott, J. L., & Laurie, E. H. (Septiembre de 1999). *Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model.* Recuperado el 20 de Septiembre de 2016, de Indiana University Bloomington: http://www.indiana.edu/~jslsoc/files_research/testing_tests/hccm/99TAS.pdf

Smyth, G. K. (1989). Generalized Linear Models with Varying Dispersion. *Royal Statistical Society , 51* (1), 47-60.

Studenmund, A. H. (2000). *Using Econometrics: A Practical Guide.* Boston: Pearson.

Stuvel, G., & James, S. F. (1950). Household Expenditure on Food in Holland. *Journal of the Royal Statistical Society , 113* (1), 59-80.

Tafalla, A. S. (2014). *Universidad de Zaragoza.* Recuperado el Septiembre de 2016, de Repositorio Institucional de Documentos: https://zaguan.unizar.es/record/15689/files/TAZ-TFG-2014-1329.pdf

Tarr, G. (2011). Small sample performance of quantile regression confidence intervals. *Journal of Statistical Computation and Simulation , 82* (1), 81-94.

Vargas, J. R., & Elizondo, A. (Marzo de 2015). *Estimación de la elasticidad precio e ingreso para alimentos: revisión a partir de los datos de la ENIGH 2013.* Recuperado el 25 de Agosto de 2016, de INEC: http://www.inec.go.cr/sites/default/files/documentos/pobreza_y_presupuesto_de_hogares/gastos_de_los_hogares/metodologias/documentos_metodologicos/mepobrezasimposioenig2013-2014-01.pdf

Verbeek, M. (2004). *A Guide to modern econometrics* (2 ed.). New York: Wiley.

Villezca Becerra, P. A., & Máynez Cano, M. (2005). Uso de funciones de ingreso gasto para el análisis del consumo de verduras en el área metropolitana de Monterrey. *Ensayos Revista de Economia , XXIV* (1), 21-52.

Villezca, P. A. (2005). Uso de funciones de ingreso gasto para el análisis del consumo de verduras en el área metropolitana de Monterrey. *Ensayos Revista de Economia , 24* (1), 21-52.

Vynck, M. (2016-2017). *Heteroscedasticity in linear models: an empirical comparison of estimation methods.* Recuperado el 15 de Enero de 2019, de Ghent University: https://lib.ugent.be/fulltxt/RUG01/002/376/288/RUG01-002376288_2017_0001_AC.pdf

Western, B., & Bloome, D. (2009). Variance Function Regressions for Studying Inequality. *Sociological Methodology , 39*, 293-326.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica , 48* (4), 817-838.

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software , 11* (10), 1-17.

Zúñiga, P., Saborío, M., Ulate, A., Linares, S., & Hernández, A. (2004). *Una mirada al bienestar de las familias costarricenses a través del consumo.* Recuperado el 13 de Junio de 2016, de INEC, Encuesta Ingresos

y Gastos, Simposio de Ingresos y Gastos, Resumenes y Ponencias: http://www.inec.go.cr/Web/Home/GeneradorPagina.aspx

ANEXOS

A. Cuadro ANOVA para el modelo de regresión lineal múltiple

Cuadro ANOVA para el modelo de regresión lineal general			
Fuente de variación	Suma de cuadrados (SC)	Grados de libertad (GL)	Cuadrado medio
Regresión(R)	$SCR = \hat{\beta}'X'Y - \left(\frac{1}{n}\right)Y'JY$	$p-1$	$CMR=\frac{SCR}{p-1}$
Error(E)	$SCE = Y'Y - \left(\frac{1}{n}\right)Y'JY$	$n-p$	$CME=\frac{SCE}{n-p}$
Total(T)	$SCT = YY' - \left(\frac{1}{n}\right)Y'JY$	$n-1$	

donde:

B. J es una matriz cuadrada n x n con todos sus elementos iguales a 1.

(92)
$$\underset{n\ x\ n}{J} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \ldots & 1 \end{bmatrix}$$

B. Pruebas simultáneas para más de un coeficiente de regresión (F)

En la prueba F de la importancia global se compara un modelo sin predictores o sólo intercepto contra el modelo que se especifique. La prueba F, prueba si existe una relación entre la variable respuesta y el conjunto de predictoras, para ello plantea las siguientes hipótesis (Neter et al., 1990):

(93)
$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$
$$H_A: No\ todos\ los\ \beta_K (K = 1, \ldots, p-1) = 0$$

La misma utiliza el siguiente estadístico de prueba, bajo el supuesto de homoscedasticidad:

(94)
$$F^* = \frac{CMR}{CME} = \frac{CMR}{\hat{\sigma}^2}$$

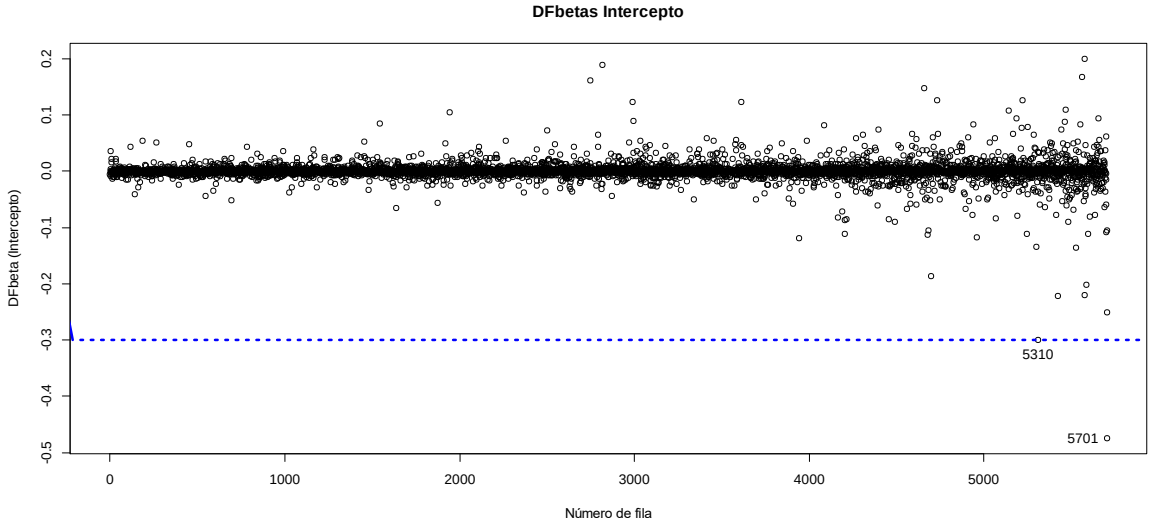
Por último, la prueba hace uso de la siguiente regla de decisión para controlar el error tipo I:

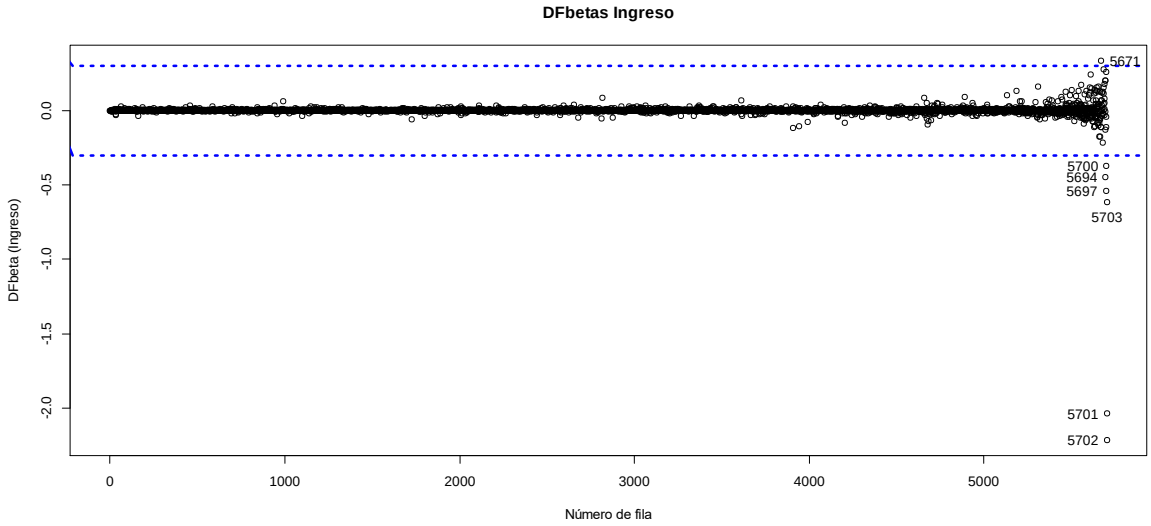
(95)
$$F^* \leq F(1-\alpha; p-1; n-p), no\ se\ rechaza\ H_0$$
$$F^* > F(1-\alpha; p-1; n-p), se\ rechaza\ H_0$$

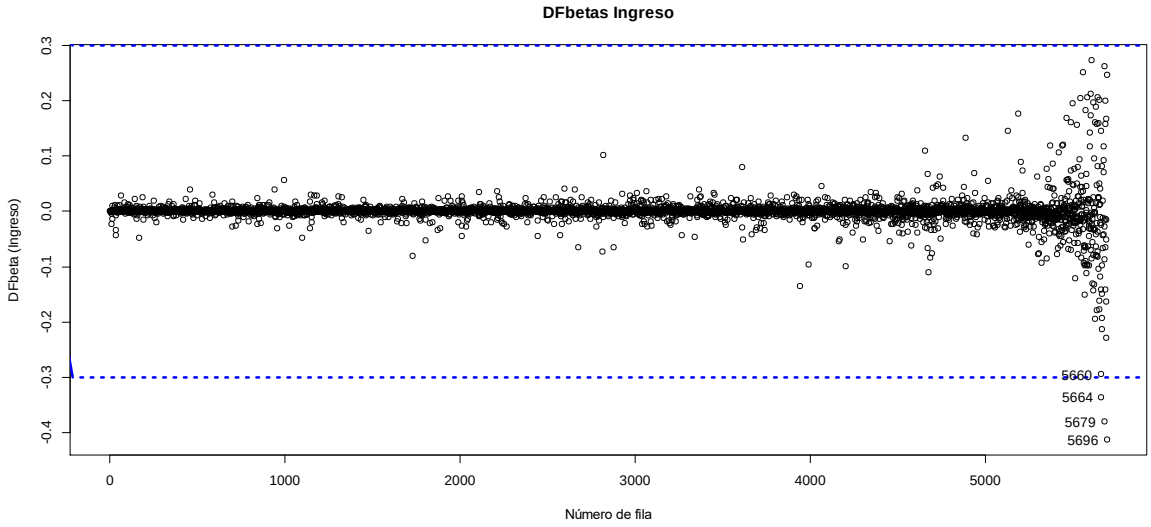
Nota: En condiciones de **heterocedasticidad**, no habría un solo estadístico F, sino que este cambiaría por cada observación, como se detalla a continuación:

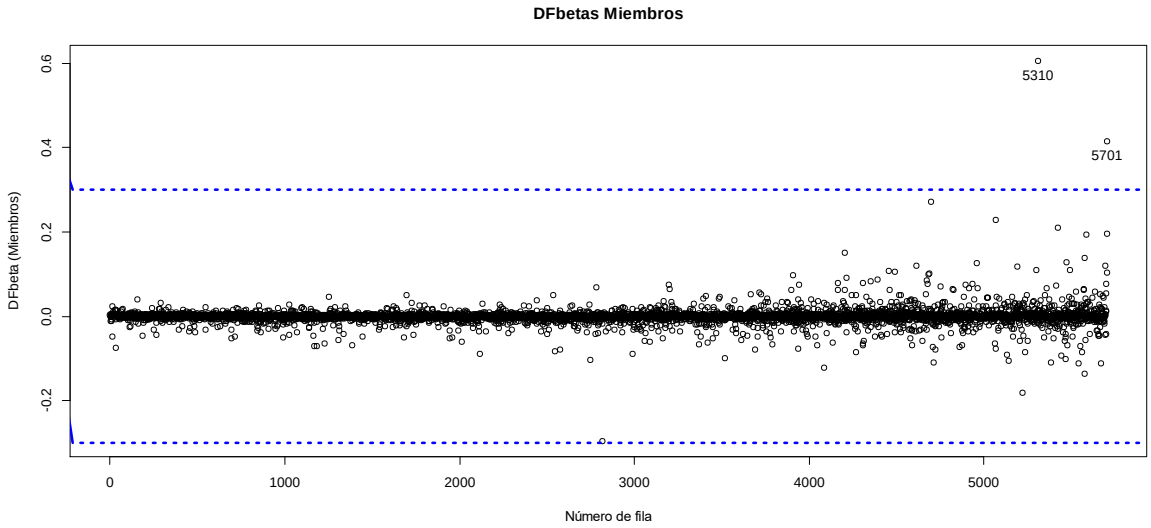
(96)
$$F^* = \frac{CMR}{\hat{\sigma}_i^2}$$

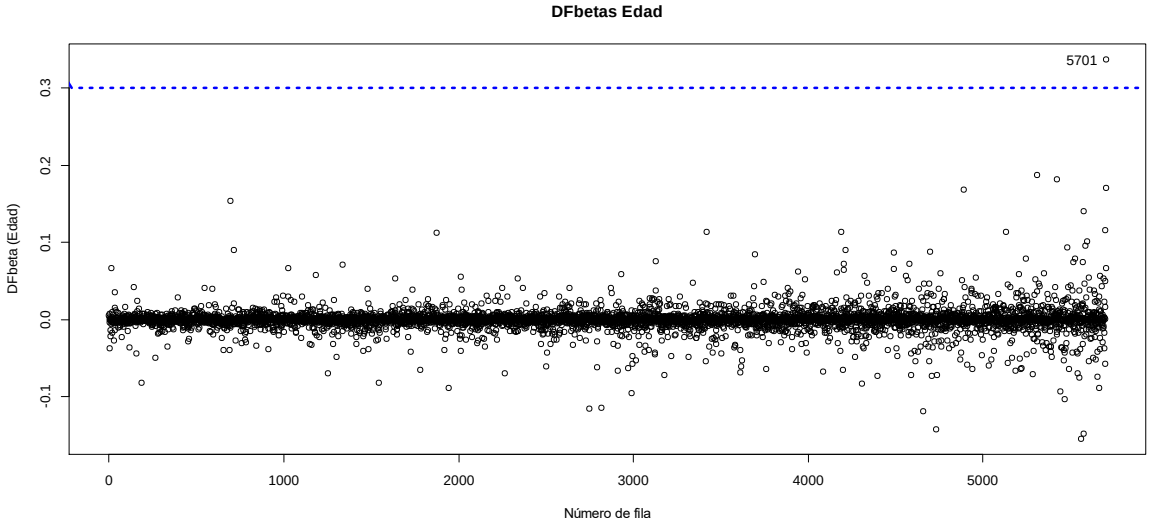
C. Caso de estudio empírico, modelo de regresión MCO: Df-betas

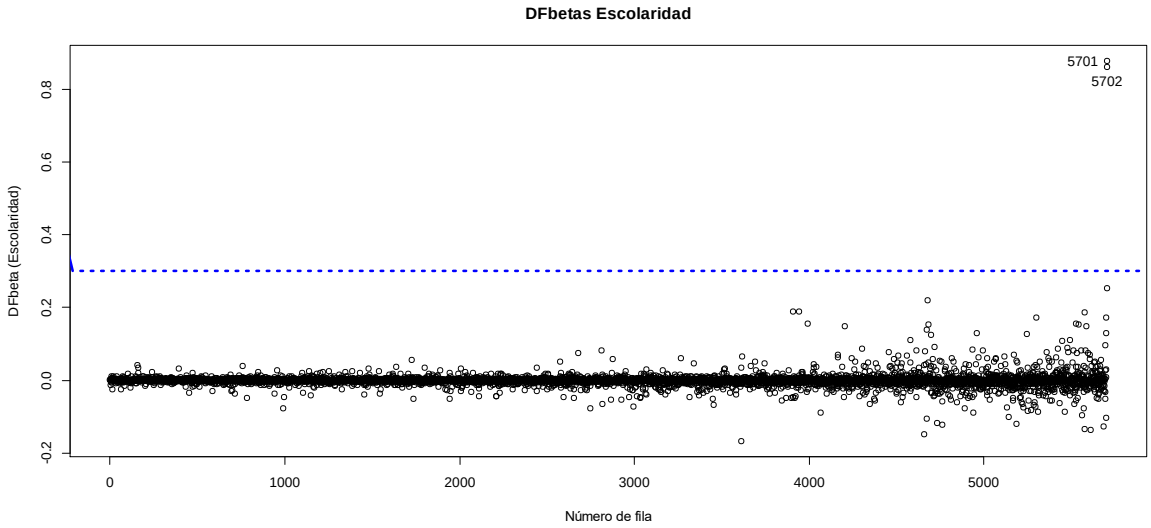



DFbetas Ingreso
DFbeta (Ingreso)
Número de fila
0.0
-0.5
-1.0
-1.5
-2.0
0
1000
2000
3000
4000
5000
5671
5700
5694
5697
5703
5701
5702

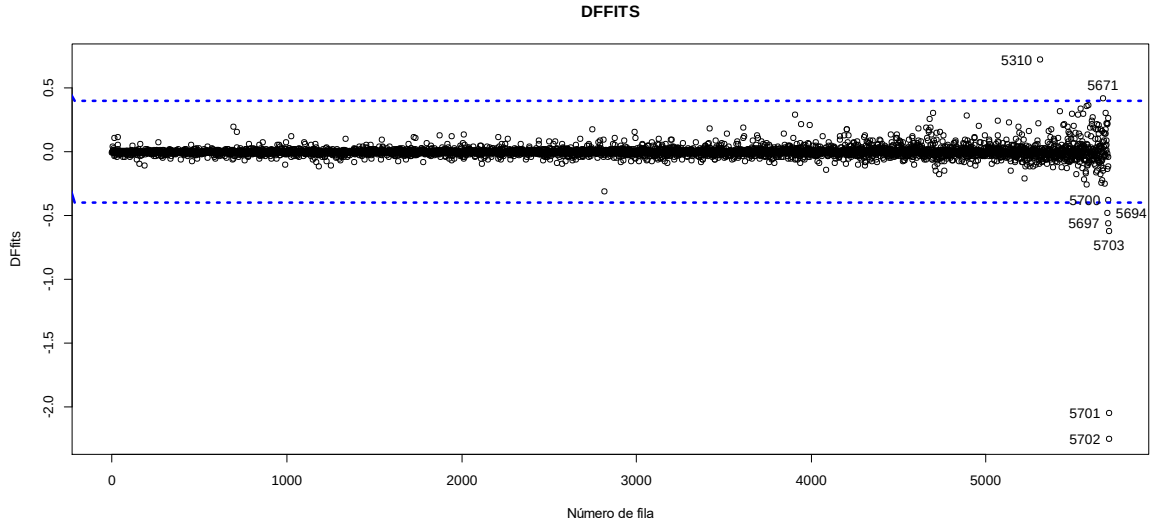

DFbetas Ingreso
DFbeta (Ingreso)
Número de fila
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
1000
2000
3000
4000
5000
5660
5664
5679
5696

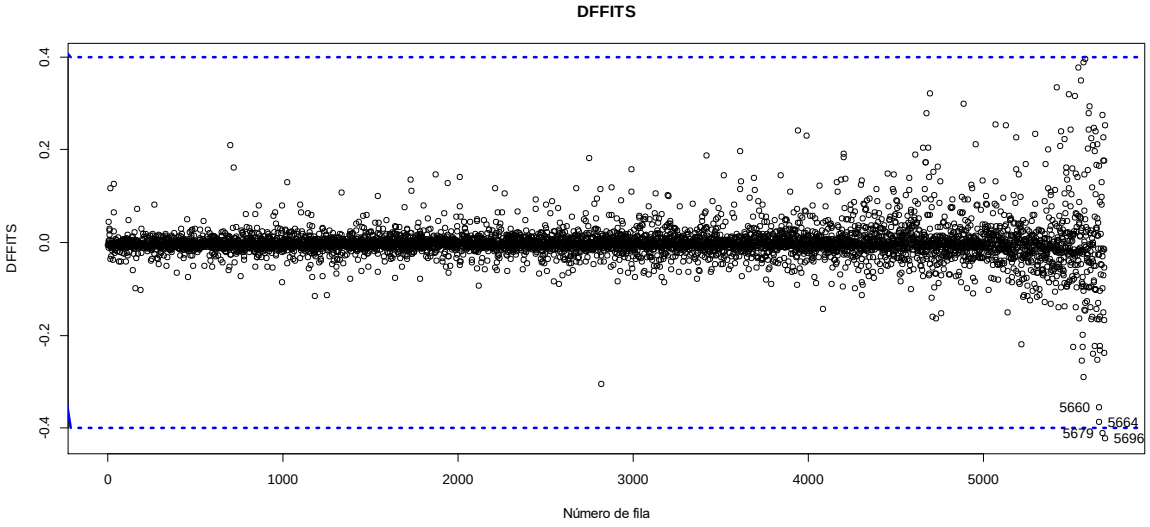

DFbetas Miembros
DFbeta (Miembros)
0.6
0.4
0.2
0.0
-0.2
5310
5701
0
1000
2000
3000
4000
5000
Número de fila


DFbetas Edad
DFbeta (Edad)
0.3
0.2
0.1
0.0
-0.1
5701
0
1000
2000
3000
4000
5000
Número de fila

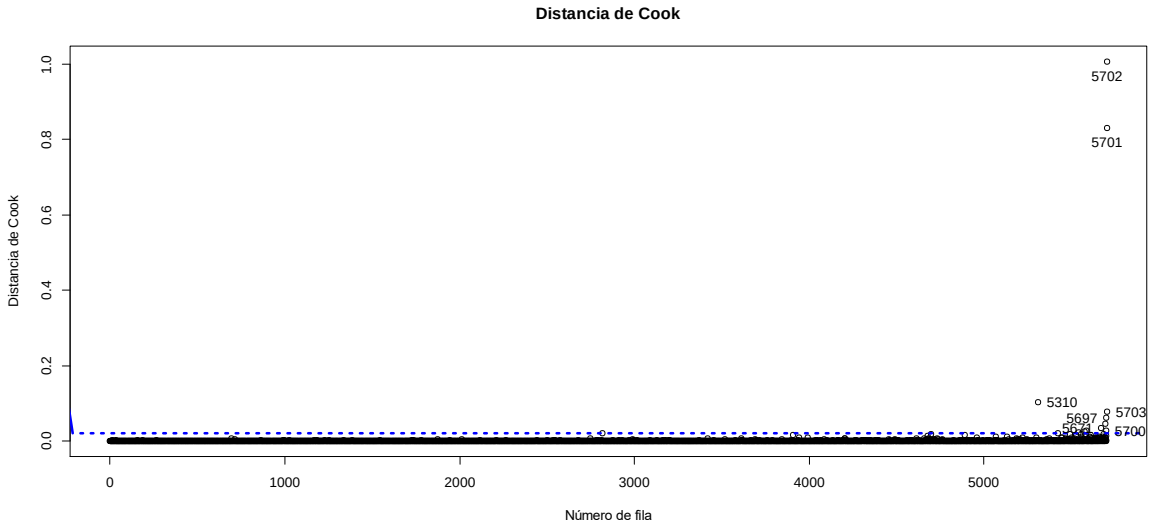

DFbetas Escolaridad
5701
5702
DFbeta (Escolaridad)
0.8
0.6
0.4
0.2
0.0
-0.2
0
1000
2000
3000
4000
5000
Número de fila

D. Caso de estudio empírico, modelo de regresión MCO: Df-fits

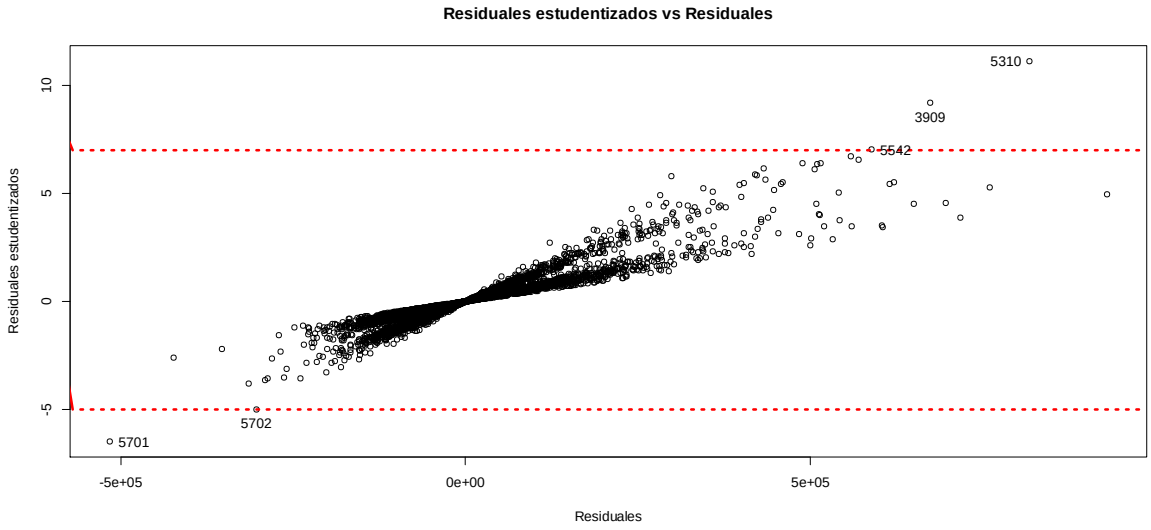




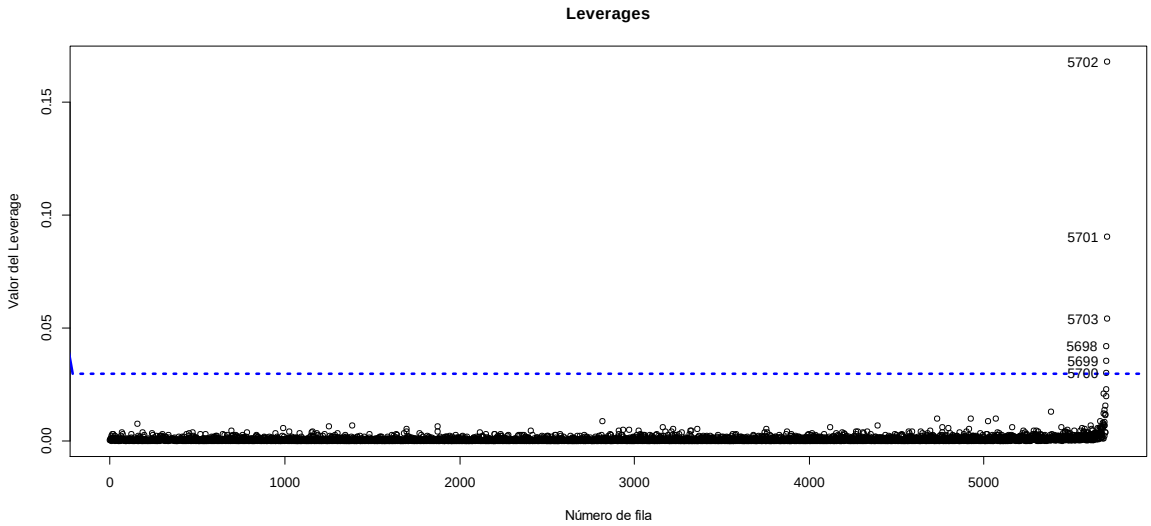
E. Caso de estudio empírico, modelo de regresión MCO: Distancia de Cook

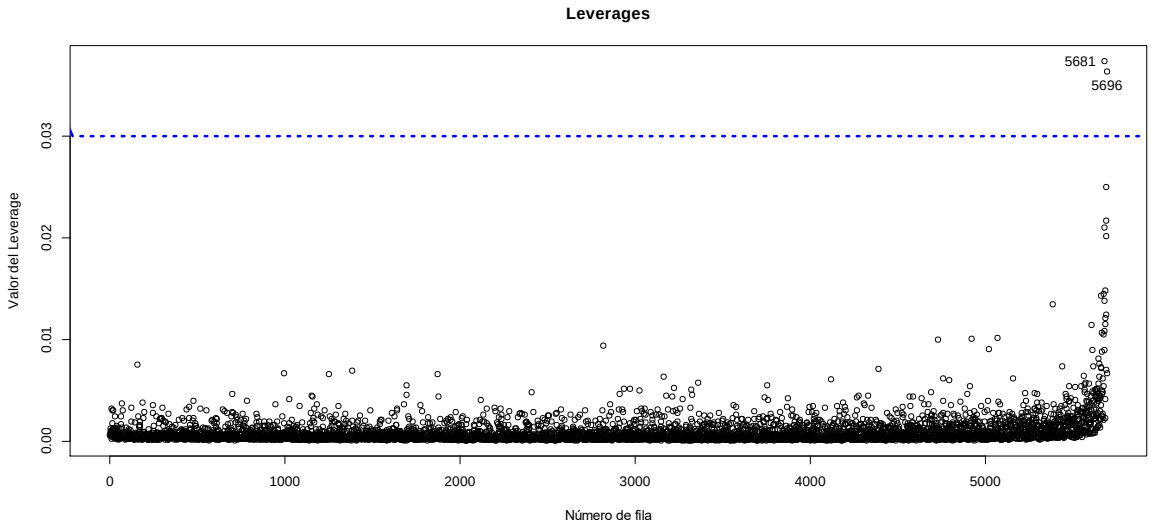


F. Caso de estudio empírico, modelo de regresión MCO: Valores extremos (Outliers)



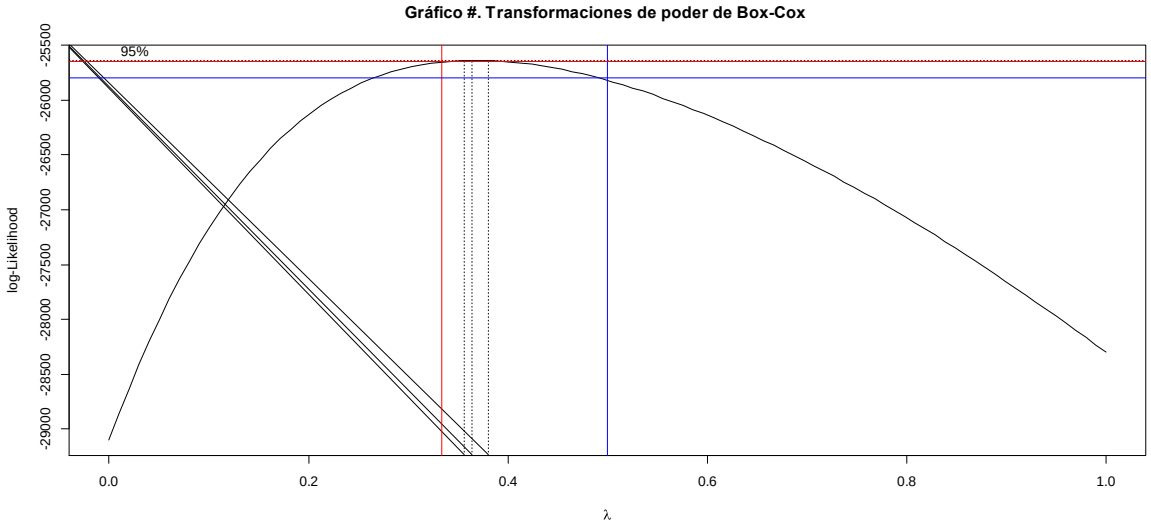
G. Caso de estudio empírico, modelo de regresión MCO: Leverages o valores Hat





H. Caso de estudio empírico, modelo de regresión MCO: Transformación de Box Cox

Para tratar de obtener un mejor ajuste a la distribución Normal se exploran las transformaciones de poder de Box-Cox que originan la siguiente gráfica:



Donde se aprecia que la verosimilitud se maximiza en $\lambda = 0,\overline{36}$ pero que incluso las transformaciones de $\lambda = 1/3$ o raíz cúbica de la variable respuesta y $\lambda = 1/2$ o raíz cuadrada de la variable respuesta ayudarían a estar bastante cerca del punto máximo de la verosimilitud el cual revela estar alrededor de -25650. Sin embargo los modelos para las transformaciones de poder sugeridas por el método de Box-Cox y las demás propuestas no alcanzaron los estándares de normalidad que los gráficos de probabilidad normal y la prueba de Kolmogorov-Smirnov requieren.

I. Desglose del modelo de mínimos cuadrados ponderados

La especificación general de la varianza, puede ser escrita como:

$$\text{var}(\hat{e}_i) = \sigma_i^2 = \sigma^2 x_i^{\gamma}$$

donde, si se toman logaritmos naturales en ambos lados de la ecuación, se obtiene:

$$\ln(\sigma_i^2) = \ln(\sigma^2) + \gamma \ln(x_i)$$

de manera que σ_i^2 se puede expresar

$$\sigma_i^2 = \exp(\ln(\sigma^2) + \gamma \ln(x_i)) = \exp(\alpha_1 + \alpha_2 z_i)$$

donde: $\alpha_1 = \ln(\sigma^2)$, $\alpha_2 = \gamma$, $z_i = \ln(x_i)$.

Se puede expresar la ecuación anterior de la siguiente manera:

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_i$$

*Como la varianza puede depender de más de una variable, la ecuación se puede extender a:

$$\ln(\sigma_i^2) = \alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{is}$$

Ambas pueden ser estimadas mediante mínimos cuadrados ordinarios, utilizando como observaciones $\hat{e}_i^2$:

$$\ln(\hat{e}_i^2) = \ln(\sigma_i^2) + v_i = \alpha_1 + \alpha_2 z_i + v_i$$

La ecuación de mínimos cuadrados para la función de varianza, es la siguiente:

$$\ln(\hat{\sigma}_i^2) = 16{,}780 + 0{,}2382 * Ingreso + 0{,}7 * Miembros + 0{,}46 * Escolaridad$$

$$\hat{\sigma}_i^2 = \exp(16{,}780 + 0{,}2382 * Ingreso + 0{,}7 * Miembros + 0{,}46 * Escolaridad)$$

Para obtener varianza constante en el modelo de mínimos cuadrados $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i$ es necesario dividir por el término $\hat{\sigma}_i$ que representa los valores ajustados de la función de varianza. De modo que:

$$var\left(\frac{e_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} var(e_i) = \frac{1}{\sigma_i^2} * \sigma_i^2 = 1$$

J. Código de simulaciones utilizando el lenguaje estadístico R

```
#Cargar librerías necesarias

library(dglm);library(quantreg);library(R2jags);library(Rlab)

#Cantidad de simulaciones

Qsimu=100

#Tamaño de muestra

Nsimu=5687

#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

#!Simulación de heterocedasticidad, causa: naturaleza de las variables

#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

#!Definición de Vectores

errorhete=rep(NA,Nsimu)

beta.ingreso.mco.nv=rep(NA,Qsimu)

beta.ingreso.dglm.nv=rep(NA,Qsimu)

beta.ingreso.gls.nv=rep(NA,Qsimu)

beta.ingreso.qreg.nv=rep(NA,Qsimu)

beta.ingreso.dglmbayes.nv=rep(NA,Qsimu)

sebeta.ingreso.mco.nv=rep(NA,Qsimu)

sebeta.ingreso.dglm.nv=rep(NA,Qsimu)

sebeta.ingreso.gls.nv=rep(NA,Qsimu)

sebeta.ingreso.qreg.nv=rep(NA,Qsimu)

sebeta.ingreso.dglmbayes.nv=rep(NA,Qsimu)
```

```
#Definición de listas
dflist<-list()
List.r2jags<-list()
#!_____________________________________________________________________
#####################Proceso generador de datos#######################
#!_____________________________________________________________________
#Se generan las variables aleatorias
for (i in 1:Qsimu) {
    Xingreso=runif(Nsimu,min =0, max = 3500)
    Xmiembros=runif(Nsimu,min =1, max = 8 )
    Xescolaridad=runif(Nsimu,min =1, max = 18 )
    Xedad=runif(Nsimu,min =21, max = 86)

#Se genera el error heterocedástico
error= exp(21.56+0.0008*Xingreso+0.14*Xmiembros+
  0.04*Xescolaridad+0.009*Xedad)

#Se genera la desviación estándar del error heterocedástico
error1=sqrt(error)

#Se genera una distribución Normal con error heteroscedástico
for (k in 1:Nsimu) {
errorhete[k]=rnorm(1,mean = 0, sd = error1[c(k)])
}
```

```
# Se genera la variable respuesta
ygasto=(-1000+60*Xingreso+18000*Xmiembros+1300*Xescolaridad+
  500*Xedad+errorhete)

# Se indexa el conjunto de datos
datos=data.frame(ygasto,Xingreso,Xmiembros,Xescolaridad,Xedad)
DATA <- paste("SIMU",i,sep = "_")
assign(DATA,datos)
df_ <- get(paste("SIMU", i, sep="_"))
dflist[[DATA]] <-df_
#!_____________________________________________________________________
####################Conjunto de modelos a evaluar#####################
#!_____________________________________________________________________

#Modelo mínimos cuadrados ordinarios
Mod_MCO= lm(ygasto~Xingreso+Xmiembros+Xescolaridad+Xedad)
#Modelo lineal doble generalizado
Mod_DGLM= dglm(ygasto ~Xingreso+ Xmiembros +Xedad+Xescolaridad,
  ~ Xingreso+ Xmiembros +Xedad+Xescolaridad,
  family=gaussian(link="identity"),dlink = "log")
#Modelo de mínimos cuadrados ponderados
wei.ols <- lm(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad)
wei.ols_res <- wei.ols$residuals
wei.ols_varf <- lm(log(wei.ols_res^2) ~ log(Xingreso+1)+
  log(Xmiembros) +log(Xescolaridad+1)+ log(Xedad))
varff_exp <- exp(wei.ols_varf$fitted.values)
```

```
Mod_GLS <- lm(ygasto ~ Xingreso + Xmiembros +Xedad+Xescolaridad,
  weights = 1/sqrt(varff_exp))
#Modelo regresión cuantílica
Mod_QREG= rq(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad,
 tau=.5,method="br", model = TRUE)
#Guardar coeficiente estimado de la variable ingreso
beta.ingreso.mco.nv[i]=Mod_MCO$coef[2]
beta.ingreso.dglm.nv[i]=Mod_DGLM$coef[2]
beta.ingreso.gls.nv[i]=Mod_GLS$coef[2]
beta.ingreso.qreg.nv[i]=Mod_QREG$coef[2]

#Error estándar del coeficiente estimado, variable ingreso

sebeta.ingreso.mco.nv[i]=coef(summary(Mod_MCO))[, "Std. Error"][2]
sebeta.ingreso.dglm.nv[i]=coef(summary(Mod_DGLM))[, "Std. Error"][2]
sebeta.ingreso.gls.nv[i]=coef(summary(Mod_GLS))[, "Std. Error"][2]
sebeta.ingreso.qreg.nv[i]=coef(summary(Mod_QREG))[, "Std. Error"][2]
}

#Definición del modelo lineal doble generalizado bayesiano
modelString = "
model {
#Verosimilitud
  for (i in 1:n) {
  y[i]~dnorm(mu[i],tau[i])
  mu[i] <- (beta0+beta1*x1[i]+beta2*x2[i]+beta3*x3[i]+beta4*x4[i])
```

```
  tau[i]<-1/(sigma2[i])
  log(sigma2[i])<-(lambda0+lambda1*x1[i]+lambda2*x2[i]+lambda3*x3[i]
  +lambda4*x4[i])
  }

  #Distribuciones previas
  beta0 ~ dnorm(0.01,1.0E-6)
  beta1 ~ dnorm(0,1.0E-6)
  beta2 ~ dnorm(0,1.0E-6)
  beta3 ~ dnorm(0,1.0E-6)
  beta4 ~ dnorm(0,1.0E-6)
  lambda0 ~ dnorm(0,tau.lambda0)
  lambda1 ~ dnorm(0,tau.lambda1)
  lambda2 ~ dnorm(0,tau.lambda2)
  lambda3 ~ dnorm(0,tau.lambda3)
  lambda4 ~ dnorm(0,tau.lambda4)
  tau.lambda0 ~ dgamma(0.001,0.001)
  tau.lambda1 ~ dgamma(0.001,0.001)
  tau.lambda2 ~ dgamma(0.001,0.001)
  tau.lambda3 ~ dgamma(0.001,0.001)
  tau.lambda4 ~ dgamma(0.001,0.001)
  } "
#Directorio para guardar modelo
writeLines(modelString, con = "./regression.txt")
```

```
#Estimación del modelo lineal doble generalizado bayesiano
for (j in 1:Qsimu) {
Datajags <- with(dflist[[j]], list(y = dflist[[j]]$ygasto,
 x1 = dflist[[j]]$Xingreso,
x2 = dflist[[j]]$Xmiembros, x3 = dflist[[j]]$Xescolaridad,
 x4 = dflist[[j]]$Xedad, n = nrow(dflist[[j]])))
DATAjags <- paste("JAGS",j,sep = "_")
assign(DATAjags,Datajags)
inits <- rep(list(list(beta0 = abs(rnorm(1)), beta1 =abs(rnorm(1)),
beta2 =abs(rnorm(1)),
beta3 =abs(rnorm(1)), beta4=abs(rnorm(1)), lambda0 = abs(rnorm(1)),
lambda1 =abs(rnorm(1)), lambda2 =abs(rnorm(1)),
lambda3 =abs(rnorm(1)), lambda4=abs(rnorm(1)),
tau.lambda0 =abs(rnorm(1)),
tau.lambda1 = abs(rnorm(1)),tau.lambda2 = abs(rnorm(1)),
tau.lambda3 = abs(rnorm(1)),
tau.lambda4 = abs(rnorm(1)), 3)))
params <- c("beta0","beta1","beta2","beta3","beta4","lambda0",
"lambda1","lambda2","lambda3","lambda4",
"tau.lambda0","tau.lambda1","tau.lambda2","tau.lambda3",
"tau.lambda4")
nChains = 3
burnInSteps = 1500
thinSteps = 5
numSavedSteps = 3000
nIter = ceiling(burnInSteps + (numSavedSteps * thinSteps)/nChains)
```

```
#Se indexa el conjunto de datos
data.r2jags  <- jags(data = get(paste("JAGS", j, sep="_")),
 inits = NULL, parameters.to.save = params,
 model.file ="./regression.txt",n.chains = nChains,
 n.iter = nIter, n.burnin = burnInSteps, n.thin = thinSteps)
 List.r2jags[[j]] =data.r2jags
#Guardar coeficiente estimado de la variable ingreso
 beta.ingreso.dglmbayes.nv[j]=
 List.r2jags[[j]][2]$BUGSoutput$mean$beta1
#Error estándar del coeficiente estimado, variable ingreso
sebeta.ingreso.dglmbayes.nv[j]=
 List.r2jags[[j]][2]$BUGSoutput$sd$beta1
}
#!_____________________________________________________________________
####################Calculo de medidas de ajuste#####################
#!_____________________________________________________________________

#Mediana de los coeficientes
mediana.mco.nv=median(beta.ingreso.mco.nv)
mediana.dglm.nv=median(beta.ingreso.dglm.nv)
mediana.gls.nv=median(beta.ingreso.gls.nv)
mediana.qreg.nv=median(beta.ingreso.qreg.nv)
mediana.dglmbayes.nv=median(beta.ingreso.dglmbayes.nv)
#Mediana de los errores estándar
se.mediana.mco.nv=median(sebeta.ingreso.mco.nv)
se.mediana.dglm.nv=median(sebeta.ingreso.dglm.nv)
```

```
se.mediana.gls.nv=median(sebeta.ingreso.gls.nv)

se.mediana.qreg.nv=median(sebeta.ingreso.qreg.nv)

se.mediana.dglmbayes.nv=median(sebeta.ingreso.dglmbayes.nv)

#Media de los coeficientes

media.mco.nv=mean(beta.ingreso.mco.nv)

media.dglm.nv=mean(beta.ingreso.dglm.nv)

media.gls.nv=mean(beta.ingreso.gls.nv)

media.qreg.nv=mean(beta.ingreso.qreg.nv)

media.dglmbayes.nv=mean(beta.ingreso.dglmbayes.nv)


#Media de los errores estándar

se.media.mco.nv=mean(sebeta.ingreso.mco.nv)

se.media.dglm.nv=mean(sebeta.ingreso.dglm.nv)

se.media.gls.nv=mean(sebeta.ingreso.gls.nv)

se.media.qreg.nv=mean(sebeta.ingreso.qreg.nv)

se.media.dglmbayes.nv=mean(sebeta.ingreso.dglmbayes.nv)


#Desviación estándar de los coeficientes

SD.mco.nv=sd(beta.ingreso.mco.nv)

SD.dglm.nv=sd(beta.ingreso.dglm.nv)

SD.gls.nv=sd(beta.ingreso.gls.nv)

SD.qreg.nv=sd(beta.ingreso.qreg.nv)

SD.dglmbayes.nv=sd(beta.ingreso.dglmbayes.nv)

#Desviación estándar de los errores estándar

se.SD.mco.nv=sd(sebeta.ingreso.mco.nv)

se.SD.dglm.nv=sd(sebeta.ingreso.dglm.nv)
```

```
se.SD.gls.nv=sd(sebeta.ingreso.gls.nv)
se.SD.qreg.nv=sd(sebeta.ingreso.qreg.nv)
se.SD.dglmbayes.nv=sd(sebeta.ingreso.dglmbayes.nv)
#Raíz del error cuadrático medio
RMS.ingreso.mco.nv=sqrt(sum((beta.ingreso.mco.nv-60)^2)/Nsimu)
RMS.ingreso.dglm.nv=sqrt(sum((beta.ingreso.dglm.nv-60)^2)/Nsimu)
RMS.ingreso.gls.nv=sqrt(sum((beta.ingreso.gls.nv-60)^2)/Nsimu)
RMS.ingreso.qreg.nv=sqrt(sum((beta.ingreso.qreg.nv-60)^2)/Nsimu)
RMS.ingreso.dglmbayes.nv=
sqrt(sum((beta.ingreso.dglmbayes.nv-60)^2)/Nsimu)
#Error absoluto medio
MAE.ingreso.mco.nv=sqrt(sum(abs(beta.ingreso.mco.nv-60))/Nsimu)
MAE.ingreso.dglm.nv=sqrt(sum(abs(beta.ingreso.dglm.nv-60))/Nsimu)
MAE.ingreso.gls.nv=sqrt(sum(abs(beta.ingreso.gls.nv-60))/Nsimu)
MAE.ingreso.qreg.nv=sqrt(sum(abs(beta.ingreso.qreg.nv-60))/Nsimu)
MAE.ingreso.dglmbayes.nv=
sqrt(sum(abs(beta.ingreso.dglmbayes.nv-60))/Nsimu)
#Resumen de medidas
nv0=c("MCO","DGLM","GLS","QREG","DGLMBAYES")
nv1=c(mediana.mco.nv,mediana.dglm.nv,mediana.gls.nv,
  mediana.qreg.nv,mediana.dglmbayes.nv)
nv2=c(se.mediana.mco.nv,se.mediana.dglm.nv,se.mediana.gls.nv,
  se.mediana.qreg.nv,se.mediana.dglmbayes.nv)
nv3=c(media.mco.nv,media.dglm.nv,media.gls.nv,media.qreg.nv,
  media.dglmbayes.nv)
nv4=c(se.media.mco.nv,se.media.dglm.nv,se.media.gls.nv,
```

```
  se.media.qreg.nv,se.media.dglmbayes.nv)
nv5=c(SD.mco.nv,SD.dglm.nv,SD.gls.nv,SD.qreg.nv,SD.dglmbayes.nv)
nv6=c(se.SD.mco.nv,se.SD.dglm.nv,se.SD.gls.nv,se.SD.qreg.nv,
  se.SD.dglmbayes.nv)
nv7=c(RMS.ingreso.mco.nv,RMS.ingreso.dglm.nv,RMS.ingreso.gls.nv,
  RMS.ingreso.qreg.nv,RMS.ingreso.dglmbayes.nv)
nv8=c(MAE.ingreso.mco.nv,MAE.ingreso.dglm.nv,MAE.ingreso.gls.nv,
  MAE.ingreso.qreg.nv,MAE.ingreso.dglmbayes.nv)

#Guardar en dataframe
beta.ingreso.nv=data.frame(beta.ingreso.mco.nv,beta.ingreso.dglm.nv
  ,beta.ingreso.gls.nv,beta.ingreso.qreg.nv,
  beta.ingreso.dglmbayes.nv)
sebeta.ingreso.nv=data.frame(sebeta.ingreso.mco.nv,
  sebeta.ingreso.dglm.nv, sebeta.ingreso.gls.nv,
  sebeta.ingreso.qreg.nv, sebeta.ingreso.dglmbayes.nv)
Estadisticos.nv=data.frame(nv0,nv1,nv2,nv3,nv4,nv5,nv6,nv7,nv8)
```

```
#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

 #!Simulación de heterocedasticidad generada por valores extremos

#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

#Definición de Vectores

errorhete=rep(NA,Nsimu)

Xingreso2=rep(NA,Nsimu)

ygasto=rep(NA,Nsimu)

beta.ingreso.mco.ve=rep(NA,Qsimu)

beta.ingreso.dglm.ve=rep(NA,Qsimu)

beta.ingreso.gls.ve=rep(NA,Qsimu)

beta.ingreso.qreg.ve=rep(NA,Qsimu)

beta.ingreso.dglmbayes.ve=rep(NA,Qsimu)

sebeta.ingreso.mco.ve=rep(NA,Qsimu)

sebeta.ingreso.dglm.ve=rep(NA,Qsimu)

sebeta.ingreso.gls.ve=rep(NA,Qsimu)

sebeta.ingreso.qreg.ve=rep(NA,Qsimu)

sebeta.ingreso.dglmbayes.ve=rep(NA,Qsimu)

#Definición de listas

dflist<-list()

List.r2jags<-list()
```

```
#!_____________________________________________________________________

#####################Proceso generador de datos####################

#!_____________________________________________________________________

 #Se generan las variables aleatorias

for (i in 1:Qsimu) {


  Xingreso=runif(Nsimu,min =0, max = 3500)

  Xmiembros=runif(Nsimu,min =1, max = 8 )

  Xescolaridad=runif(Nsimu,min =1, max = 18 )

  Xedad=runif(Nsimu,min =21, max = 86)

  Xingreso5001000=runif(Nsimu,min =501, max = 1000)

  Xingreso10001500=runif(Nsimu,min =1001, max = 1500)

  Xingreso15002000=runif(Nsimu,min =1501, max = 2000)

  Xingreso20002500=runif(Nsimu,min =2001, max = 2500)

  Xingreso25003000=runif(Nsimu,min =2501, max = 3000)

  Xingreso30003500=runif(Nsimu,min =3001, max = 3500)

  Xingreso35004000=runif(Nsimu,min =3501, max = 4000)

 #Se genera el error

  error= exp(21.56+0.0008*Xingreso+0.14*Xmiembros+0.04*Xescolaridad

    +0.009*Xedad)

#Se genera la desviación estándar del error

  error1=sqrt(error)

#Distribución Normal con error heteroscedástico

  for (k in 1:Nsimu) {

    errorhete[k]=rnorm(1,mean = 0, sd = error1[c(k)])

  }
```

```
#Generar variable respuesta
  ygasto=(-1000+60*Xingreso+18000*Xmiembros+1300*Xescolaridad+
    500*Xedad)+errorhete
# Generar conjunto de datos
  datos=data.frame(ygasto,Xingreso,Xmiembros,Xescolaridad,Xedad,
    Xingreso2,errorhete,Xingreso5001000,Xingreso10001500,
    Xingreso15002000,Xingreso20002500,Xingreso25003000,
    Xingreso30003500,Xingreso35004000)
  #Ordenar la variable ingreso de menor a mayor valor
  datos <- datos[order(datos$Xingreso),]
  #Se crea variable Bernoulli
  datos$bern=rbern(Nsimu,0.01)
  #Contaminar el ingreso
  for (m in 1:nrow(datos))
  {
    if (datos$bern[m]==1 && datos$Xingreso[m] <500) {
      datos$Xingreso2[m]=datos$Xingreso5001000[m]
    } else if ( datos$bern[m]==1 && datos$Xingreso[m] >=500
      && datos$Xingreso[m] <1000) {
      datos$Xingreso2[m]=datos$Xingreso10001500[m]
    } else if ( datos$bern[m]==1 && datos$Xingreso[m] >=1000
     && datos$Xingreso[m] <1500) {
      datos$Xingreso2[m]=datos$Xingreso15002000[m]
    }else if ( datos$bern[m]==1 && datos$Xingreso[m] >=1500
     && datos$Xingreso[m] <2000) {
      datos$Xingreso2[m]=datos$Xingreso20002500[m]
```

```
    }else if ( datos$bern[m]==1 && datos$Xingreso[m] >=2000
     && datos$Xingreso[m] <2500) {
      datos$Xingreso2[m]=datos$Xingreso25003000[m]
    }else if ( datos$bern[m]==1 && datos$Xingreso[m] >=2500
     && datos$Xingreso[m] <3000) {
      datos$Xingreso2[m]=datos$Xingreso30003500[m]
    }else if ( datos$bern[m]==1 && datos$Xingreso[m] >=3000)
     {      datos$Xingreso2[m]=datos$Xingreso35004000[m]
    }else {datos$Xingreso2[m]=datos$Xingreso[m]
    }}


  #Indexar el conjunto de datos
  DATA <- paste("SIMU",i,sep = "_")
  assign(DATA,datos)
  df_ <- get(paste("SIMU", i, sep="_"))
  dflist[[DATA]] <-df_
#!_____________________________________________________________________
####################Conjunto de modelos a evaluar####################
#!_____________________________________________________________________


  #Modelo  mínimos cuadrados ordinarios
  Mod_MCO= lm(ygasto~Xingreso2+Xmiembros+Xescolaridad+Xedad,
    data=datos)
 #Modelo lineal doble generalizado
  Mod_DGLM= dglm(ygasto ~Xingreso2+ Xmiembros + Xedad+ Xescolaridad,
    ~Xingreso2+ Xmiembros + Xedad+ Xescolaridad,
```

```
  family=gaussian(link="identity"),dlink = "log",data = datos)
 #Modelo  de mínimos cuadrados ponderados(wls)
 wei.ols <- lm(ygasto~ Xingreso2 + Xmiembros +Xedad+Xescolaridad,
  data=datos)
 wei.ols_res <- wei.ols$residuals
 wei.ols_varf <- lm(log(wei.ols_res^2) ~ log(Xingreso2+1)+
  log(Xmiembros) +log(Xescolaridad+1)+ log(Xedad),data=datos)
 varff_exp <- exp(wei.ols_varf$fitted.values)
 Mod_GLS <- lm(ygasto ~ Xingreso2 + Xmiembros +Xedad+Xescolaridad
  ,weights = 1/sqrt(varff_exp),data=datos)
#Modelo regresión cuantílica
 Mod_QREG= rq(ygasto~ Xingreso2 + Xmiembros +Xedad+Xescolaridad,
  tau=.5,method="br", model = TRUE,data = datos)
#Guardar coeficiente estimado de la variable ingreso
 beta.ingreso.mco.ve[i]=Mod_MCO$coef[2]
 beta.ingreso.dglm.ve[i]=Mod_DGLM$coef[2]
 beta.ingreso.gls.ve[i]=Mod_GLS$coef[2]
 beta.ingreso.qreg.ve[i]=Mod_QREG$coef[2]

 #Error estándar del coeficiente estimado, variable ingreso
 sebeta.ingreso.mco.ve[i]=coef(summary(Mod_MCO))[, "Std. Error"][2]
 sebeta.ingreso.dglm.ve[i]=coef(summary(Mod_DGLM))[, "Std. Error"][2]
 sebeta.ingreso.gls.ve[i]=coef(summary(Mod_GLS))[, "Std. Error"][2]
 sebeta.ingreso.qreg.ve[i]=coef(summary(Mod_QREG))[, "Std. Error"][2]}
```

```
#Definición del modelo lineal doble generalizado bayesiano

modelString = "

model {

#Verosimilitud

for (i in 1:n) {

y[i]~dnorm(mu[i],tau[i])

mu[i] <- (beta0+beta1*x1[i]+beta2*x2[i]+beta3*x3[i]+beta4*x4[i])

tau[i]<-1/(sigma2[i])

log(sigma2[i])<-(lambda0+lambda1*x1[i]+lambda2*x2[i]+

lambda3*x3[i]+lambda4*x4[i])}

#Distribuciones previas

beta0 ~ dnorm(0.01,1.0E-6)

beta1 ~ dnorm(0,1.0E-6)

beta2 ~ dnorm(0,1.0E-6)

beta3 ~ dnorm(0,1.0E-6)

beta4 ~ dnorm(0,1.0E-6)

lambda0 ~ dnorm(0,tau.lambda0)

lambda1 ~ dnorm(0,tau.lambda1)

lambda2 ~ dnorm(0,tau.lambda2)

lambda3 ~ dnorm(0,tau.lambda3)

lambda4 ~ dnorm(0,tau.lambda4)

tau.lambda0 ~ dgamma(0.001,0.001)

tau.lambda1 ~ dgamma(0.001,0.001)

tau.lambda2 ~ dgamma(0.001,0.001)

tau.lambda3 ~ dgamma(0.001,0.001)

tau.lambda4 ~ dgamma(0.001,0.001)} "
```

```
#Directorio para guardar modelo
 writeLines(modelString, con = "./regression.txt")
#Estimación del modelo lineal doble generalizado bayesiano
for (j in 1:Qsimu) {
  Datajags <- with(dflist[[j]], list(y = dflist[[j]]$ygasto,
   x1 = dflist[[j]]$Xingreso,
               x2 = dflist[[j]]$Xmiembros,
               x3 = dflist[[j]]$Xescolaridad, x4 = dflist[[j]]$Xedad,
                n = nrow(dflist[[j]])))
  DATAjags <- paste("JAGS",j,sep = "_")
  assign(DATAjags,Datajags)

  inits <- rep(list(list(beta0 = abs(rnorm(1)), beta1 =abs(rnorm(1)),
   beta2 =abs(rnorm(1)),
   beta3 =abs(rnorm(1)), beta4=abs(rnorm(1)), lambda0 = abs(rnorm(1)),
   lambda1 =abs(rnorm(1)), lambda2 =abs(rnorm(1)),
   lambda3 =abs(rnorm(1)), lambda4=abs(rnorm(1)),
   tau.lambda0 =abs(rnorm(1)),
   tau.lambda1 = abs(rnorm(1)),tau.lambda2 = abs(rnorm(1)),
   tau.lambda3 = abs(rnorm(1)),
   tau.lambda4 = abs(rnorm(1)), 3)))

  params <- c("beta0","beta1","beta2","beta3","beta4","lambda0",
  "lambda1","lambda2","lambda3","lambda4","tau.lambda0",
  "tau.lambda1","tau.lambda2","tau.lambda3","tau.lambda4")
  nChains = 3
```

```
  burnInSteps = 1500
  thinSteps = 5
  numSavedSteps = 3000
  nIter = ceiling(burnInSteps + (numSavedSteps * thinSteps)/nChains)
  nIter

  data.r2jags  <- jags(data = get(paste("JAGS", j, sep="_")),
  inits = NULL, parameters.to.save = params,
  model.file ="./regression.txt",n.chains = nChains, n.iter = nIter,
  n.burnin = burnInSteps, n.thin = thinSteps)
  #Se indexa el conjunto de datos
  List.r2jags[[j]] =data.r2jags
  #Guardar coeficiente estimado de la variable ingreso
  beta.ingreso.dglmbayes.ve[j]= List.r2jags[[j]][2]$BUGSoutput$mean$beta1
  #Error estándar del coeficiente estimado, variable ingreso
  sebeta.ingreso.dglmbayes.ve[j]= List.r2jags[[j]][2]$BUGSoutput$sd$beta1
}
#!_____________________________________________________________________
####################Calculo de medidas de ajuste#####################
#!_____________________________________________________________________
#Mediana
mediana.mco.ve=median(beta.ingreso.mco.ve)
mediana.dglm.ve=median(beta.ingreso.dglm.ve)
mediana.gls.ve=median(beta.ingreso.gls.ve)
mediana.qreg.ve=median(beta.ingreso.qreg.ve)
mediana.dglmbayes.ve=median(beta.ingreso.dglmbayes.ve)
```

```
se.mediana.mco.ve=median(sebeta.ingreso.mco.ve)
se.mediana.dglm.ve=median(sebeta.ingreso.dglm.ve)
se.mediana.gls.ve=median(sebeta.ingreso.gls.ve)
se.mediana.qreg.ve=median(sebeta.ingreso.qreg.ve)
se.mediana.dglmbayes.ve=median(sebeta.ingreso.dglmbayes.ve)

#Media
media.mco.ve=mean(beta.ingreso.mco.ve)
media.dglm.ve=mean(beta.ingreso.dglm.ve)
media.gls.ve=mean(beta.ingreso.gls.ve)
media.qreg.ve=mean(beta.ingreso.qreg.ve)
media.dglmbayes.ve=mean(beta.ingreso.dglmbayes.ve)
se.media.mco.ve=mean(sebeta.ingreso.mco.ve)
se.media.dglm.ve=mean(sebeta.ingreso.dglm.ve)
se.media.gls.ve=mean(sebeta.ingreso.gls.ve)
se.media.qreg.ve=mean(sebeta.ingreso.qreg.ve)
se.media.dglmbayes.ve=mean(sebeta.ingreso.dglmbayes.ve)

#Desviación #estándar
SD.mco.ve=sd(beta.ingreso.mco.ve)
SD.dglm.ve=sd(beta.ingreso.dglm.ve)
SD.gls.ve=sd(beta.ingreso.gls.ve)
SD.qreg.ve=sd(beta.ingreso.qreg.ve)
SD.dglmbayes.ve=sd(beta.ingreso.dglmbayes.ve)
se.SD.mco.ve=sd(sebeta.ingreso.mco.ve)
se.SD.dglm.ve=sd(sebeta.ingreso.dglm.ve)
```

```
se.SD.gls.ve=sd(sebeta.ingreso.gls.ve)
se.SD.qreg.ve=sd(sebeta.ingreso.qreg.ve)
se.SD.dglmbayes.ve=sd(sebeta.ingreso.dglmbayes.ve)

#Raíz del error cuadrático medio
RMS.ingreso.mco.ve=sqrt(sum((beta.ingreso.mco.ve-60)^2)/Nsimu)
RMS.ingreso.dglm.ve=sqrt(sum((beta.ingreso.dglm.ve-60)^2)/Nsimu)
RMS.ingreso.gls.ve=sqrt(sum((beta.ingreso.gls.ve-60)^2)/Nsimu)
RMS.ingreso.qreg.ve=sqrt(sum((beta.ingreso.qreg.ve-60)^2)/Nsimu)
RMS.ingreso.dglmbayes.ve=
sqrt(sum((beta.ingreso.dglmbayes.ve-60)^2)/Nsimu)

#Error absoluto medio
MAE.ingreso.mco.ve=sqrt(sum(abs(beta.ingreso.mco.ve-60))/Nsimu)
MAE.ingreso.dglm.ve=sqrt(sum(abs(beta.ingreso.dglm.ve-60))/Nsimu)
MAE.ingreso.gls.ve=sqrt(sum(abs(beta.ingreso.gls.ve-60))/Nsimu)
MAE.ingreso.qreg.ve=sqrt(sum(abs(beta.ingreso.qreg.ve-60))/Nsimu)
MAE.ingreso.dglmbayes.ve=
sqrt(sum(abs(beta.ingreso.dglmbayes.ve-60))/Nsimu)
#Resumen de medidas
ve0=c("MCO","DGLM","GLS","QREG","DGLMBAYES")
ve1=c(mediana.mco.ve,mediana.dglm.ve,mediana.gls.ve,mediana.qreg.ve,
  mediana.dglmbayes.ve)
ve2=c(se.mediana.mco.ve,se.mediana.dglm.ve,se.mediana.gls.ve,
  se.mediana.qreg.ve,se.mediana.dglmbayes.ve)
ve3=c(media.mco.ve,media.dglm.ve,media.gls.ve,media.qreg.ve,
```

```
  media.dglmbayes.ve)
ve4=c(se.media.mco.ve,se.media.dglm.ve,se.media.gls.ve,
  se.media.qreg.ve,se.media.dglmbayes.ve)
ve5=c(SD.mco.ve,SD.dglm.ve,SD.gls.ve,SD.qreg.ve,SD.dglmbayes.ve)
ve6=c(se.SD.mco.ve,se.SD.dglm.ve,se.SD.gls.ve,se.SD.qreg.ve,
  se.SD.dglmbayes.ve)
ve7=c(RMS.ingreso.mco.ve,RMS.ingreso.dglm.ve,RMS.ingreso.gls.ve,
  RMS.ingreso.qreg.ve,RMS.ingreso.dglmbayes.ve)
ve8=c(MAE.ingreso.mco.ve,MAE.ingreso.dglm.ve,MAE.ingreso.gls.ve,
  MAE.ingreso.qreg.ve,MAE.ingreso.dglmbayes.ve)
#Guardar en dataframe
beta.ingreso.ve=data.frame(beta.ingreso.mco.ve,beta.ingreso.dglm.ve,
  beta.ingreso.gls.ve,beta.ingreso.qreg.ve,beta.ingreso.dglmbayes.ve)
sebeta.ingreso.ve=data.frame(sebeta.ingreso.mco.ve,
  sebeta.ingreso.dglm.ve, sebeta.ingreso.gls.ve,
   sebeta.ingreso.qreg.ve,sebeta.ingreso.dglmbayes.ve)
Estadisticos.ve=data.frame(ve0,ve1,ve2,ve3,ve4,ve5,ve6,ve7,ve8)
```

```
#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

#Simulación de heterocedasticidad generada por variable omitida

#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________


#Definición de vectores

ygasto=rep(NA,Nsimu)

beta.ingreso.mco.vo=rep(NA,Qsimu)

beta.ingreso.dglm.vo=rep(NA,Qsimu)

beta.ingreso.gls.vo=rep(NA,Qsimu)

beta.ingreso.qreg.vo=rep(NA,Qsimu)

beta.ingreso.dglmbayes.vo=rep(NA,Qsimu)

sebeta.ingreso.mco.vo=rep(NA,Qsimu)

sebeta.ingreso.dglm.vo=rep(NA,Qsimu)

sebeta.ingreso.gls.vo=rep(NA,Qsimu)

sebeta.ingreso.qreg.vo=rep(NA,Qsimu)

sebeta.ingreso.dglmbayes.vo=rep(NA,Qsimu)

#Definición de listas

dflist<-list()

List.r2jags<-list()
```

```
#!_____________________________________________________________________

#####################Proceso generador de datos####################

#!_____________________________________________________________________

 #Se generan las variables aleatorias

for (i in 1:Qsimu) {

    Xingreso=runif(Nsimu,min =0, max = 3500 )

    Xmiembros=runif(Nsimu,min =1, max = 8 )

    Xescolaridad=runif(Nsimu,min =1, max = 18 )

    Xedad=runif(Nsimu,min =21, max = 86)


#La variable omitida se correlacionada 0,5 con la variable ingreso

    Xomitida=0.5*Xingreso+0.5*runif(Nsimu,min =0, max = 3500 )

    # Se genera el error

    error=rnorm(Nsimu,mean = 0, sd =Xomitida*60)

  # Se calcula la variable respuesta

    ygasto=(-1000+60*Xingreso+18000*Xmiembros+1300*Xescolaridad+

      500*Xedad+60*Xomitida+error)

 # Se indexan los conjuntos de datos

    datos=data.frame(ygasto,Xingreso,Xmiembros,Xescolaridad,Xedad)

      DATA <- paste("SIMU",i,sep = "_")

      assign(DATA,datos)

      df_ <- get(paste("SIMU", i, sep="_"))

      dflist[[DATA]] <-df_

  # Modelo mínimos cuadrados ordinarios

  Mod_MCO= lm(ygasto~Xingreso+Xmiembros+Xescolaridad+Xedad)
```

```
# Modelo lineal  doble generalizado
Mod_DGLM= dglm(ygasto ~Xingreso+ Xmiembros +Xedad+Xescolaridad,
  ~ Xingreso+ Xmiembros +Xedad+Xescolaridad,
  family=gaussian(link="identity"),dlink = "log")
# Modelo mínimos cuadrados ponderados(wls)
wei.ols <- lm(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad)
wei.ols_res <- wei.ols$residuals
wei.ols_varf <- lm(log(wei.ols_res^2) ~ log(Xingreso+1)+
log(Xmiembros) +log(Xescolaridad+1)+ log(Xedad))
varff_exp <- exp(wei.ols_varf$fitted.values)
Mod_GLS <- lm(ygasto ~ Xingreso + Xmiembros +Xedad+Xescolaridad,
  weights = 1/sqrt(varff_exp))
# Modelo de regresión cuantílica
Mod_QREG= rq(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad,
 tau=.5,method="br", model = TRUE)
# Guardar coeficiente estimado de la variable ingreso
beta.ingreso.mco.vo[i]=Mod_MCO$coef[2]
beta.ingreso.dglm.vo[i]=Mod_DGLM$coef[2]
beta.ingreso.gls.vo[i]=Mod_GLS$coef[2]
beta.ingreso.qreg.vo[i]=Mod_QREG$coef[2]
# Error estándar del coeficiente estimado, variable ingreso
sebeta.ingreso.mco.vo[i]=coef(summary(Mod_MCO))[, "Std. Error"][2]
sebeta.ingreso.dglm.vo[i]=coef(summary(Mod_DGLM))[, "Std. Error"][2]
sebeta.ingreso.gls.vo[i]=coef(summary(Mod_GLS))[, "Std. Error"][2]
sebeta.ingreso.qreg.vo[i]=coef(summary(Mod_QREG))[, "Std. Error"][2]
}
```

```
# Definición del modelo lineal doble generalizado bayesiano
modelString = "
  model {
  #Verosimilitud
  for (i in 1:n) {
  y[i]~dnorm(mu[i],tau[i])
  mu[i] <- (beta0+beta1*x1[i]+beta2*x2[i]+beta3*x3[i]+beta4*x4[i])
  tau[i]<-1/(sigma2[i])
  log(sigma2[i])<-(lambda0+lambda1*x1[i]+lambda2*x2[i]+lambda3*x3[i]
  +lambda4*x4[i])  }
  # Distribuciones previas
  beta0 ~ dnorm(0.01,1.0E-6)
  beta1 ~ dnorm(0,1.0E-6)
  beta2 ~ dnorm(0,1.0E-6)
  beta3 ~ dnorm(0,1.0E-6)
  beta4 ~ dnorm(0,1.0E-6)
  lambda0 ~ dnorm(0,tau.lambda0)
  lambda1 ~ dnorm(0,tau.lambda1)
  lambda2 ~ dnorm(0,tau.lambda2)
  lambda3 ~ dnorm(0,tau.lambda3)
  lambda4 ~ dnorm(0,tau.lambda4)
  tau.lambda0 ~ dgamma(0.001,0.001)
  tau.lambda1 ~ dgamma(0.001,0.001)
  tau.lambda2 ~ dgamma(0.001,0.001)
  tau.lambda3 ~ dgamma(0.001,0.001)
  tau.lambda4 ~ dgamma(0.001,0.001)
```

```
  }  "
# Directorio para guardar modelo
writeLines(modelString, con = "./regression.txt")
# Estimación del modelo lineal doble generalizado bayesiano
for (j in 1:Qsimu) {
Datajags <- with(dflist[[j]], list(y = dflist[[j]]$ygasto,
x1 = dflist[[j]]$Xingreso, x2 = dflist[[j]]$Xmiembros,
x3 = dflist[[j]]$Xescolaridad, x4 = dflist[[j]]$Xedad,
n = nrow(dflist[[j]])))
DATAjags <- paste("JAGS",j,sep = "_")
assign(DATAjags,Datajags)
inits <- rep(list(list(beta0 = abs(rnorm(1)), beta1 =abs(rnorm(1)),
beta2 =abs(rnorm(1)), beta3 =abs(rnorm(1)), beta4=abs(rnorm(1)),
lambda0 = abs(rnorm(1)), lambda1 =abs(rnorm(1)),
lambda2 =abs(rnorm(1)),lambda3 =abs(rnorm(1)), lambda4=abs(rnorm(1)),
tau.lambda0 =abs(rnorm(1)),tau.lambda1 = abs(rnorm(1)),
tau.lambda2 = abs(rnorm(1)), tau.lambda3 = abs(rnorm(1)),
tau.lambda4 = abs(rnorm(1)), 3)))

params <- c("beta0","beta1","beta2","beta3","beta4","lambda0",
"lambda1","lambda2","lambda3","lambda4", "tau.lambda0",
"tau.lambda1","tau.lambda2","tau.lambda3","tau.lambda4")
nChains = 3
burnInSteps = 1500
thinSteps = 5
numSavedSteps = 3000
```

```
nIter = ceiling(burnInSteps + (numSavedSteps * thinSteps)/nChains)
nIter


data.r2jags  <- jags(data = get(paste("JAGS", j, sep="_")),
 inits = NULL, parameters.to.save = params,
  model.file ="./regression.txt",n.chains = nChains, n.iter = nIter,
   n.burnin = burnInSteps, n.thin = thinSteps)
#Se indexa el conjunto de datos
 List.r2jags[[j]] =data.r2jags
#Guardar coeficiente estimado de la variable ingreso
 beta.ingreso.dglmbayes.vo[j]=
 List.r2jags[[j]][2]$BUGSoutput$mean$beta1
#Error estándar del coeficiente estimado, variable ingreso
  sebeta.ingreso.dglmbayes.vo[j]=
  List.r2jags[[j]][2]$BUGSoutput$sd$beta1
}
#!_____________________________________________________________________
####################Calculo de medidas de ajuste#####################
#!_____________________________________________________________________
#Mediana
mediana.mco.vo=median(beta.ingreso.mco.vo)
mediana.dglm.vo=median(beta.ingreso.dglm.vo)
mediana.gls.vo=median(beta.ingreso.gls.vo)
mediana.qreg.vo=median(beta.ingreso.qreg.vo)
mediana.dglmbayes.vo=median(beta.ingreso.dglmbayes.vo)
```

```
se.mediana.mco.vo=median(sebeta.ingreso.mco.vo)
se.mediana.dglm.vo=median(sebeta.ingreso.dglm.vo)
se.mediana.gls.vo=median(sebeta.ingreso.gls.vo)
se.mediana.qreg.vo=median(sebeta.ingreso.qreg.vo)
se.mediana.dglmbayes.vo=median(sebeta.ingreso.dglmbayes.vo)
#Media
media.mco.vo=mean(beta.ingreso.mco.vo)
media.dglm.vo=mean(beta.ingreso.dglm.vo)
media.gls.vo=mean(beta.ingreso.gls.vo)
media.qreg.vo=mean(beta.ingreso.qreg.vo)
media.dglmbayes.vo=mean(beta.ingreso.dglmbayes.vo)
se.media.mco.vo=mean(sebeta.ingreso.mco.vo)
se.media.dglm.vo=mean(sebeta.ingreso.dglm.vo)
se.media.gls.vo=mean(sebeta.ingreso.gls.vo)
se.media.qreg.vo=mean(sebeta.ingreso.qreg.vo)
se.media.dglmbayes.vo=mean(sebeta.ingreso.dglmbayes.vo)
#Desviación estándar
SD.mco.vo=sd(beta.ingreso.mco.vo)
SD.dglm.vo=sd(beta.ingreso.dglm.vo)
SD.gls.vo=sd(beta.ingreso.gls.vo)
SD.qreg.vo=sd(beta.ingreso.qreg.vo)
SD.dglmbayes.vo=sd(beta.ingreso.dglmbayes.vo)
se.SD.mco.vo=sd(sebeta.ingreso.mco.vo)
se.SD.dglm.vo=sd(sebeta.ingreso.dglm.vo)
se.SD.gls.vo=sd(sebeta.ingreso.gls.vo)
se.SD.qreg.vo=sd(sebeta.ingreso.qreg.vo)
```

```
se.SD.dglmbayes.vo=sd(sebeta.ingreso.dglmbayes.vo)


# Raíz del error cuadrático medio
RMS.ingreso.mco.vo=sqrt(sum((beta.ingreso.mco.vo-60)^2)/Nsimu)
RMS.ingreso.dglm.vo=sqrt(sum((beta.ingreso.dglm.vo-60)^2)/Nsimu)
RMS.ingreso.gls.vo=sqrt(sum((beta.ingreso.gls.vo-60)^2)/Nsimu)
RMS.ingreso.qreg.vo=sqrt(sum((beta.ingreso.qreg.vo-60)^2)/Nsimu)
RMS.ingreso.dglmbayes.vo=
sqrt(sum((beta.ingreso.dglmbayes.vo-60)^2)/Nsimu)


# Error absoluto medio
MAE.ingreso.mco.vo=sqrt(sum(abs(beta.ingreso.mco.vo-60))/Nsimu)
MAE.ingreso.dglm.vo=sqrt(sum(abs(beta.ingreso.dglm.vo-60))/Nsimu)
MAE.ingreso.gls.vo=sqrt(sum(abs(beta.ingreso.gls.vo-60))/Nsimu)
MAE.ingreso.qreg.vo=sqrt(sum(abs(beta.ingreso.qreg.vo-60))/Nsimu)
MAE.ingreso.dglmbayes.vo=
sqrt(sum(abs(beta.ingreso.dglmbayes.vo-60))/Nsimu)


# Resumen de medidas
vo0=c("MCO","DGLM","GLS","QREG","DGLMBAYES")
vo1=c(mediana.mco.vo,mediana.dglm.vo,mediana.gls.vo,
  mediana.qreg.vo,mediana.dglmbayes.vo)
vo2=c(se.mediana.mco.vo,se.mediana.dglm.vo,se.mediana.gls.vo,
  se.mediana.qreg.vo,se.mediana.dglmbayes.vo)
vo3=c(media.mco.vo,media.dglm.vo,media.gls.vo,
  media.qreg.vo,media.dglmbayes.vo)
```

```
vo4=c(se.media.mco.vo,se.media.dglm.vo,
  se.media.gls.vo,se.media.qreg.vo,se.media.dglmbayes.vo)
vo5=c(SD.mco.vo,SD.dglm.vo,SD.gls.vo,SD.qreg.vo,SD.dglmbayes.vo)
vo6=c(se.SD.mco.vo,se.SD.dglm.vo,se.SD.gls.vo,
  se.SD.qreg.vo,se.SD.dglmbayes.vo)
vo7=c(RMS.ingreso.mco.vo,RMS.ingreso.dglm.vo,RMS.ingreso.gls.vo,
  RMS.ingreso.qreg.vo,RMS.ingreso.dglmbayes.vo)
vo8=c(MAE.ingreso.mco.vo,MAE.ingreso.dglm.vo,MAE.ingreso.gls.vo,
  MAE.ingreso.qreg.vo,MAE.ingreso.dglmbayes.vo)
#Guardar en dataframe
beta.ingreso.vo=data.frame(beta.ingreso.mco.vo,beta.ingreso.dglm.vo,
beta.ingreso.gls.vo,beta.ingreso.qreg.vo,beta.ingreso.dglmbayes.vo)
sebeta.ingreso.vo=data.frame(sebeta.ingreso.mco.vo,
sebeta.ingreso.dglm.vo, sebeta.ingreso.gls.vo,
sebeta.ingreso.qreg.vo,sebeta.ingreso.dglmbayes.vo)
Estadisticos.vo=data.frame(vo0,vo1,vo2,vo3,vo4,vo5,vo6,vo7,vo8)
```

```
#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

  #Simulación sin variable omitida

#!_____________________________________________________________________

#!_____________________________________________________________________

#!_____________________________________________________________________

#Definición de Vectores

ygasto=rep(NA,Nsimu)

beta.ingreso.mco.vo=rep(NA,Qsimu)

beta.ingreso.dglm.vo=rep(NA,Qsimu)

beta.ingreso.gls.vo=rep(NA,Qsimu)

beta.ingreso.qreg.vo=rep(NA,Qsimu)

beta.ingreso.dglmbayes.vo=rep(NA,Qsimu)

sebeta.ingreso.mco.vo=rep(NA,Qsimu)

sebeta.ingreso.dglm.vo=rep(NA,Qsimu)

sebeta.ingreso.gls.vo=rep(NA,Qsimu)

sebeta.ingreso.qreg.vo=rep(NA,Qsimu)

sebeta.ingreso.dglmbayes.vo=rep(NA,Qsimu)


#Definición de listas

dflist<-list()

List.r2jags<-list()
```

```
#!_____________________________________________________________________

#####################Proceso generador de datos####################

#!_____________________________________________________________________


for (i in 1:Qsimu) {

    #Se generan las variables aleatorias

    Xingreso=runif(Nsimu,min =0, max = 3500 )

    Xmiembros=runif(Nsimu,min =1, max = 8 )

    Xescolaridad=runif(Nsimu,min =1, max = 18 )

    Xedad=runif(Nsimu,min =21, max = 86)

 #La variable omitida se correlaciona 0,5  con el ingreso

    Xomitida=0.5*Xingreso+0.5*runif(Nsimu,min =0, max = 3500 )

 #Se genera el error

    error=rnorm(Nsimu,mean = 0, sd =Xomitida*60)

 #Se genera la variable respuesta

    ygasto=(-1000+60*Xingreso+18000*Xmiembros+1300*Xescolaridad

      +500*Xedad+60*Xomitida+error)

 #Indexar el conjunto de datos

datos=data.frame(ygasto,Xingreso,Xmiembros,Xescolaridad,Xedad,

  Xomitida)

DATA <- paste("SIMU",i,sep = "_")

assign(DATA,datos)

df_ <- get(paste("SIMU", i, sep="_"))

  dflist[[DATA]] <-df_
```

```
#Modelo mínimos cuadrados ordinarios
Mod_MCO= lm(ygasto~Xingreso+Xmiembros+Xescolaridad+Xedad+Xomitida)
#Modelo lineal  doble generalizado
Mod_DGLM= dglm(ygasto ~Xingreso+ Xmiembros +Xedad+Xescolaridad+
Xomitida,~ Xingreso+ Xmiembros +Xedad+Xescolaridad+Xomitida,
family=gaussian(link="identity"),dlink = "log")
#Modelo mínimos cuadrados ponderados
wei.ols <- lm(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad+
  Xomitida)
wei.ols_res <- wei.ols$residuals
wei.ols_varf <- lm(log(wei.ols_res^2) ~ log(Xingreso+1)+
log(Xmiembros) +log(Xescolaridad+1)+ log(Xedad)+log(Xomitida+1))
varff_exp <- exp(wei.ols_varf$fitted.values)
Mod_GLS <- lm(ygasto ~ Xingreso + Xmiembros +Xedad+Xescolaridad+
  Xomitida,weights = 1/sqrt(varff_exp))
#Modelo regresión cuantílica
Mod_QREG= rq(ygasto~ Xingreso + Xmiembros +Xedad+Xescolaridad+
  Xomitida, tau=.5,method="br", model = TRUE)
#Guardar coeficiente estimado de la variable ingreso
beta.ingreso.mco.vo[i]=Mod_MCO$coef[2]
beta.ingreso.dglm.vo[i]=Mod_DGLM$coef[2]
beta.ingreso.gls.vo[i]=Mod_GLS$coef[2]
beta.ingreso.qreg.vo[i]=Mod_QREG$coef[2]
#Error estándar del coeficiente estimado, variable ingreso
sebeta.ingreso.mco.vo[i]=coef(summary(Mod_MCO))[, "Std. Error"][2]
sebeta.ingreso.dglm.vo[i]=coef(summary(Mod_DGLM))[, "Std. Error"][2]
```

```
sebeta.ingreso.gls.vo[i]=coef(summary(Mod_GLS))[, "Std. Error"][2]
sebeta.ingreso.qreg.vo[i]=coef(summary(Mod_QREG))[, "Std. Error"][2]
}
#Definición del modelo lineal doble generalizado bayesiano
modelString = "
  model {
  #Verosimilitud
  for (i in 1:n) {
  y[i]~dnorm(mu[i],tau[i])
  mu[i] <- (beta0+beta1*x1[i]+beta2*x2[i]+
  beta3*x3[i]+beta4*x4[i]+beta5*x5[i])
  tau[i]<-1/(sigma2[i])
  log(sigma2[i])<-(lambda0+lambda1*x1[i]+lambda2*x2[i]+
  lambda3*x3[i]+lambda4*x4[i]+lambda5*x5[i])
  }
  #Distribuciones previas
  beta0 ~ dnorm(0.01,1.0E-6)
  beta1 ~ dnorm(0,1.0E-6)
  beta2 ~ dnorm(0,1.0E-6)
  beta3 ~ dnorm(0,1.0E-6)
  beta4 ~ dnorm(0,1.0E-6)
  beta5 ~ dnorm(0,1.0E-6)
  lambda0 ~ dnorm(0,tau.lambda0)
  lambda1 ~ dnorm(0,tau.lambda1)
  lambda2 ~ dnorm(0,tau.lambda2)
  lambda3 ~ dnorm(0,tau.lambda3)
```

```
  lambda4 ~ dnorm(0,tau.lambda4)
  lambda5 ~ dnorm(0,tau.lambda5)
  tau.lambda0 ~ dgamma(0.001,0.001)
  tau.lambda1 ~ dgamma(0.001,0.001)
  tau.lambda2 ~ dgamma(0.001,0.001)
  tau.lambda3 ~ dgamma(0.001,0.001)
  tau.lambda4 ~ dgamma(0.001,0.001)
  tau.lambda5 ~ dgamma(0.001,0.001)} "
#Directorio para guardar modelo
writeLines(modelString, con = "./regression.txt")
#Estimación del modelo lineal doble generalizado bayesiano
for (j in 1:Qsimu) {
Datajags <- with(dflist[[j]], list(y = dflist[[j]]$ygasto,
 x1 = dflist[[j]]$Xingreso, x2 = dflist[[j]]$Xmiembros,
 x3 = dflist[[j]]$Xescolaridad, x4 = dflist[[j]]$Xedad,
 x5 = dflist[[j]]$Xomitida, n = nrow(dflist[[j]])))
DATAjags <- paste("JAGS",j,sep = "_")
assign(DATAjags,Datajags)
inits <- rep(list(list(beta0 = abs(rnorm(1)), beta1 =abs(rnorm(1)),
 beta2 =abs(rnorm(1)), beta3 =abs(rnorm(1)), beta4=abs(rnorm(1)),
 beta5=abs(rnorm(1)), lambda0 = abs(rnorm(1)), lambda1 =abs(rnorm(1)),
 lambda2 =abs(rnorm(1)),lambda3 =abs(rnorm(1)), lambda4=abs(rnorm(1)),
 lambda5=abs(rnorm(1)),tau.lambda0 =abs(rnorm(1)),
 tau.lambda1 = abs(rnorm(1)),tau.lambda2 = abs(rnorm(1)),
 tau.lambda3 = abs(rnorm(1)), tau.lambda4 = abs(rnorm(1)),
 tau.lambda5 = abs(rnorm(1)), 3)))
```

```
params <- c("beta0","beta1","beta2","beta3","beta4","beta5","lambda0",
"lambda1","lambda2","lambda3","lambda4","lambda5",
"tau.lambda0","tau.lambda1","tau.lambda2","tau.lambda3",
"tau.lambda4","tau.lambda5")
nChains = 3
burnInSteps = 1500
thinSteps = 5
numSavedSteps = 3000
nIter = ceiling(burnInSteps + (numSavedSteps * thinSteps)/nChains)
nIter
#Se indexa el conjunto de datos
data.r2jags  <- jags(data = get(paste("JAGS", j, sep="_")),
 inits = NULL, parameters.to.save = params,
  model.file ="./regression.txt",n.chains = nChains, n.iter = nIter,
   n.burnin = burnInSteps, n.thin = thinSteps)
 List.r2jags[[j]] =data.r2jags
#Guardar coeficiente estimado de la variable ingreso
 beta.ingreso.dglmbayes.vo[j]=
 List.r2jags[[j]][2]$BUGSoutput$mean$beta1
#Error estándar del coeficiente estimado, variable ingreso
 sebeta.ingreso.dglmbayes.vo[j]=
 List.r2jags[[j]][2]$BUGSoutput$sd$beta1}
```

```
#!____________________________________________________________________

###################Calculo de medidas de ajuste#####################

#!____________________________________________________________________

#Mediana

mediana.mco.vo=median(beta.ingreso.mco.vo)

mediana.dglm.vo=median(beta.ingreso.dglm.vo)

mediana.gls.vo=median(beta.ingreso.gls.vo)

mediana.qreg.vo=median(beta.ingreso.qreg.vo)

mediana.dglmbayes.vo=median(beta.ingreso.dglmbayes.vo)

se.mediana.mco.vo=median(sebeta.ingreso.mco.vo)

se.mediana.dglm.vo=median(sebeta.ingreso.dglm.vo)

se.mediana.gls.vo=median(sebeta.ingreso.gls.vo)

se.mediana.qreg.vo=median(sebeta.ingreso.qreg.vo)

se.mediana.dglmbayes.vo=median(sebeta.ingreso.dglmbayes.vo)

#Media

media.mco.vo=mean(beta.ingreso.mco.vo)

media.dglm.vo=mean(beta.ingreso.dglm.vo)

media.gls.vo=mean(beta.ingreso.gls.vo)

media.qreg.vo=mean(beta.ingreso.qreg.vo)

media.dglmbayes.vo=mean(beta.ingreso.dglmbayes.vo)

se.media.mco.vo=mean(sebeta.ingreso.mco.vo)

se.media.dglm.vo=mean(sebeta.ingreso.dglm.vo)

se.media.gls.vo=mean(sebeta.ingreso.gls.vo)

se.media.qreg.vo=mean(sebeta.ingreso.qreg.vo)

se.media.dglmbayes.vo=mean(sebeta.ingreso.dglmbayes.vo)
```

```
#Desviación estándar
SD.mco.vo=sd(beta.ingreso.mco.vo)
SD.dglm.vo=sd(beta.ingreso.dglm.vo)
SD.gls.vo=sd(beta.ingreso.gls.vo)
SD.qreg.vo=sd(beta.ingreso.qreg.vo)
SD.dglmbayes.vo=sd(beta.ingreso.dglmbayes.vo)
se.SD.mco.vo=sd(sebeta.ingreso.mco.vo)
se.SD.dglm.vo=sd(sebeta.ingreso.dglm.vo)
se.SD.gls.vo=sd(sebeta.ingreso.gls.vo)
se.SD.qreg.vo=sd(sebeta.ingreso.qreg.vo)
se.SD.dglmbayes.vo=sd(sebeta.ingreso.dglmbayes.vo)
#Raíz del error cuadrático medio
RMS.ingreso.mco.vo=sqrt(sum((beta.ingreso.mco.vo-60)^2)/Nsimu)
RMS.ingreso.dglm.vo=sqrt(sum((beta.ingreso.dglm.vo-60)^2)/Nsimu)
RMS.ingreso.gls.vo=sqrt(sum((beta.ingreso.gls.vo-60)^2)/Nsimu)
RMS.ingreso.qreg.vo=sqrt(sum((beta.ingreso.qreg.vo-60)^2)/Nsimu)
RMS.ingreso.dglmbayes.vo=
sqrt(sum((beta.ingreso.dglmbayes.vo-60)^2)/Nsimu)
#Error absoluto medio
MAE.ingreso.mco.vo=sqrt(sum(abs(beta.ingreso.mco.vo-60))/Nsimu)
MAE.ingreso.dglm.vo=sqrt(sum(abs(beta.ingreso.dglm.vo-60))/Nsimu)
MAE.ingreso.gls.vo=sqrt(sum(abs(beta.ingreso.gls.vo-60))/Nsimu)
MAE.ingreso.qreg.vo=sqrt(sum(abs(beta.ingreso.qreg.vo-60))/Nsimu)
MAE.ingreso.dglmbayes.vo=
sqrt(sum(abs(beta.ingreso.dglmbayes.vo-60))/Nsimu)
```

```
#Resumen de medidas
vo0=c("MCO","DGLM","GLS","QREG","DGLMBAYES")
vo1=c(mediana.mco.vo,mediana.dglm.vo,mediana.gls.vo,mediana.qreg.vo,
  mediana.dglmbayes.vo)
vo2=c(se.mediana.mco.vo,se.mediana.dglm.vo,se.mediana.gls.vo,
  se.mediana.qreg.vo,se.mediana.dglmbayes.vo)
vo3=c(media.mco.vo,media.dglm.vo,media.gls.vo,media.qreg.vo,
  media.dglmbayes.vo)
vo4=c(se.media.mco.vo,se.media.dglm.vo,se.media.gls.vo,
  se.media.qreg.vo,se.media.dglmbayes.vo)
vo5=c(SD.mco.vo,SD.dglm.vo,SD.gls.vo,SD.qreg.vo,SD.dglmbayes.vo)
vo6=c(se.SD.mco.vo,se.SD.dglm.vo,se.SD.gls.vo,se.SD.qreg.vo,
  se.SD.dglmbayes.vo)
vo7=c(RMS.ingreso.mco.vo,RMS.ingreso.dglm.vo,RMS.ingreso.gls.vo,
  RMS.ingreso.qreg.vo,RMS.ingreso.dglmbayes.vo)
vo8=c(MAE.ingreso.mco.vo,MAE.ingreso.dglm.vo,MAE.ingreso.gls.vo,
  MAE.ingreso.qreg.vo,MAE.ingreso.dglmbayes.vo)
#Guardar en dataframe
beta.ingreso.vo=data.frame(beta.ingreso.mco.vo,beta.ingreso.dglm.vo,
 beta.ingreso.gls.vo,beta.ingreso.qreg.vo,beta.ingreso.dglmbayes.vo)
 sebeta.ingreso.vo=data.frame(sebeta.ingreso.mco.vo,
 sebeta.ingreso.dglm.vo, sebeta.ingreso.gls.vo,
 sebeta.ingreso.qreg.vo,sebeta.ingreso.dglmbayes.vo)
Estadisticos.vo=data.frame(vo0,vo1,vo2,vo3,vo4,vo5,vo6,vo7,vo8)
```