UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

MODELOS LINEALES MIXTOS Y UNA APLICACIÓN EN LA ESTIMACIÓN DEL INGRESO PARA MÉXICO 2014

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

PRESENTA :

SHEILA CARBAJAL CHAVEZ

DIRECTOR DE TESIS:
DR. RICARDO RAMÍREZ ALDANA

Ciudad Universitaria, Cd. Mx., 2019

Hoja de información

1.Datos del alumno
Carbajal
Chavez
Sheila
5514259510
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
307087591

2. Datos del tutor
Dr.
Ricardo
Ramírez
Aldana

3. Datos del sinodal 1
Dra.
Lizbeth
Naranjo
Albarrán

4. Datos del sinodal 2
Dra.
Mónica
Tinajero
Bravo

5. Datos del sinodal 3
Act.
Isaías Manuel
Ramírez
Bañales

6. Datos del sinodal 4
Act.
Edna Gabriela
López
Estrada

7.Datos del trabajo escrito.
Modelos lineales mixtos y una aplicación en la estimación del ingreso para México 2014
85 páginas
2019

Agradecimientos

Dios, tu amor y tu bondad no tienen fin, me permites sonreír ante todos mis logros que son resultado de tu ayuda y mi esfuerzo, cuando caigo y me pones a prueba, aprendo de mis errores y me doy cuenta que los pones enfrente mío para que mejore como ser humano, y crezca de diversas maneras.

Gracias a la vida por este nuevo triunfo, gracias a todas las personas que me apoyaron y creyeron en la realización de esta tesis. No ha sido sencillo el camino hasta ahora, pero gracias a las personas que han contribuido al proceso y conclusión de este trabajo por medio de su apoyo, lo complicado de lograr esta meta se ha notado menos.

En primer lugar, quiero agradecer al Dr. Ricardo Ramirez Aldama, director de esta tesis por todo el apoyo, facilidades, paciencia y por haber compartido conmigo sus conocimientos.

Agradezco a la Universidad Nacional Autónoma de México y su Facultad de Ciencias, porque me permitieron formarme como profesionista.

A la Dra. Mónica Tinajero Bravo, por sus enseñanzas, por motivarme a seguir adelante en los momentos de desesperación y sobre todo por su amistad.

A Claudia Maria Gónzales Rivera por ser parte significativa de mi vida, por el apoyo recibido desde el día que la conocí, comprensión, amistad y cariño.

A los profesores Héctor Mendez Lango y Pedro Miramontes Vidal por confiar, alentarme y creer en mi durante mi etapa universitaria. Gracias por los conocimientos que me transmitieron.

Gracias al esfuerzo de mi madre y abuelos, gracias por siempre desear y anhelar lo mejor para mi vida, gracias por cada consejo y por cada una de sus palabras que me guiaron. Gracias por haber creído en mi hasta el último momento. ¡Ya soy Actuaria!

Finalmente agradezco a quien lee este apartado y más de mi tesis, por permitir a mis experiencias, investigaciones y conocimientos, incurrir dentro de su repertorio de información.

Introducción

Los modelos lineales mixtos (MLM) son modelos estadísticos para variables de respuesta continuas que consideran dos tipos de efectos: los efectos fijos (variables continuas o categóricas que modifican la media de la variable respuesta) y los efectos aleatorios (variables cuyos niveles son seleccionados aleatoriamente de una población), los estimadores de los parámetros para los efectos fijos tienen la misma expresión que aquellos asociados a los efectos fijos estimados como en un modelo de regresión lineal múltiple y los efectos aleatorios se estiman a partir de una esperanza condicional.

Son una extensión de los modelos lineales simples que se usan particularmente cuando no hay independencia en los datos, tal como surge de una estructura jerárquica. Los MLM son herramientas potentes y complejas sin embargo los avances en software han hecho que estas herramientas sean accesibles para los no expertos.

En este tipo de modelos se permite trabajar con datos agrupados, como estudiantes en aulas, o medidas longitudinales, en los que los sujetos se miden repetidamente a lo largo del tiempo.

El uso de estos modelos se ha extendido como una herramienta de interés para la modelización en diversas disciplinas, como por ejemplo, en la biología donde es muy frecuente que las preguntas de investigación se traten de resolver recogiendo información de variables en unidades agregadas en distintos niveles. Es por esta razón, que esta tesis está dirigida a proporcionar la teoría y herramientas disponibles para el desarrollo teórico-práctico de un modelo lineal mixto al ilustrar el desarrollo y análisis con conjuntos de datos reales.

El objetivo principal es familiarizar al lector con conceptos básicos del o de los modelos lineales mixtos que le permitan implementar de manera adecuada modelos de este tipo en situaciones prácticas propias de su quehacer cotidiano o como apoyo a nivel de licenciatura en la solución de problemas en diversas áreas del conocimiento empleando los procedimientos disponibles en los paquetes del software STATA.

Se presentan algunos desarrollos matemáticos de la teoría de los modelos lineales mixtos, de manera que, se asume que el lector está familiarizado con métodos estadísticos, especialmente aquellos relacionados con modelos lineales.

En el primer capítulo se encuentra un breve recuento de la teoría de la regresión lineal múltiple, una extensión del modelo de regresión lineal simple que considera más de una variable explicativa, esto con la finalidad de diferenciar los supuestos entre ambas técnicas, ya que, pueden verse comprometidos cuando no se cumple algún supuesto dentro de alguna de las técnicas señaladas.

En el segundo capítulo se aborda la teoría de los modelos lineales mixtos centrado en el caso de dos niveles, puesto que la generalización a un mayor número de niveles es más o menos inmediata.

El capítulo tres presenta la aplicación mediante el software estadístico STATA versión 14.1, la cual se realizó a través de datos reales generados por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), con información del Instituto Nacional de Estadística y Geografía (INEGI), específicamente el ingreso en escala logarítmica, esto ya que, la distribución de probabilidad sin tal tranformación no refleja un comportamiento normal y tranformada sí. Además, en este capítulo se presentan las fases que se siguieron para la implementación, verificación y selección del modelo que proporcionó el mejor ajuste a los datos observados.

Finalmente, se ofrecen algunas conclusiones y anexos con el propósito de dar mayor claridad sobre el tema.

Índice general

Capítulo 1

Regresión Lineal Múltiple

1.1. Introducción

El análisis de regresión es uno de los métodos más utilizados para hacer estimaciones, se emplea cuando existe relación entre dos o más variables.

En una regresión lineal simple se asocia una variable respuesta y con una variable explicativa x a través de una función lineal de la forma:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

para i =1,....,T, en donde T es el tamaño de la muestra; β_1 es la ordenada al origen; β_2 es la pendiente y ε_i es un error aleatorio con:

- $\mathbb{E}[\varepsilon_i] = 0$.
- $\text{Var}[\varepsilon_i] = \sigma^2$.
- $\text{Cov}(\varepsilon_i, \varepsilon_j)$ =0, para toda $i \neq j$.
- $\varepsilon_i \perp \varepsilon_j$, para toda $i \neq j$.

En la regresión lineal múltiple se usa más de una variable explicativa; esto nos ofrece la ventaja de utilizar más información en la construcción del modelo y, realizar estimaciones probablemente más precisas que en el caso de la regresión lineal simple. En este capítulo se presentan supuestos del modelo de regresión lineal múltiple, estimación de párametros, varianza residual, estimadores máximo verosímiles, sus propiedades, pruebas de hipótesis y medidas de bondad de ajuste a modo de repaso, sin profundizar en demostrar resultados, en vista de que, estos temas se estudian en cursos básicos de estadística.

1.2. El modelo de regresión lineal múltiple

Un modelo de regresión donde se permite más de una variable explicativa se llama modelo de regresión múltiple; siendo una generalización del modelo de regresión simple. Si se supone que se tienen k variables explicativas, entonces tenemos el modelo:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + \varepsilon_i,$$

para $i = 1, \ldots, T$ en donde:

x_{ji} es el valor de la variable explicativa j-ésima, $j = 1, \ldots, k$ para el individuo i.

ε_i es un error aleatorio con distribución $N(0, \sigma^2)$.

Si se expresa el modelo para los T individuos, se obtiene el siguiente sistema de ecuaciones:

$$y_1 = \beta_1 x_{11} + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_1 x_{12} + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + \varepsilon_2$$

$$\vdots$$

$$y_T = \beta_1 x_{1T} + \beta_2 x_{2T} + \beta_3 x_{3T} + \ldots + \beta_k x_{kT} + \varepsilon_T.$$

El anterior sistema de ecuaciones también puede expresarse usando la notación matricial:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}; X = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{k1} \\ x_{12} & x_{22} & \ldots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1T} & x_{2T} & \ldots & x_{kT} \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

De lo anterior, el modelo de regresión lineal múltiple se puede expresar de la siguiente forma:

$$Y = X\beta + \varepsilon,$$

donde Y es un vector $T \times 1$, X es una matriz $T \times k$, β es un vector $k \times 1$ y ε es un vector $T \times 1$.

1.3. Supuestos del modelo

Se deben cumplir los siguientes supuestos para utilizar la técnica de regresión lineal múltiple:

a) La relación entre la variable respuesta, las variables explicativas y los errores aleatorios es lineal (***supuesto de linealidad***). En caso de incumplimiento, se debe introducir en el modelo componentes no lineales; como transformaciones no lineales o interacciones entre dos o más variables explicativas.

b) Los parámetros $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ son constantes.

c) Los errores tienen una ***distribución normal***: $\varepsilon_i \backsim N(0, \sigma^2)$, $i = 1, \ldots, T$.

d) La media de los errores es cero: $\mathbb{E}(\varepsilon_i) = 0$, para $i = 1, \ldots, T$.

e) Los errores tienen varianza constante (***supuesto de homoscedasticidad***): $var(\varepsilon_i) = \sigma^2$, para $i = 1, \ldots, T$.

f) Los errores con diferentes subíndices no están correlacionadas entre sí (***supuesto de no correlación***): $Cov(\varepsilon_i, \varepsilon_j) = 0$, para toda $i \neq j$.

g) Los errores con diferentes subíndices son independientes entre sí (***supuesto de independencia***): $\varepsilon_i \perp \varepsilon_j$, para toda $i \neq j$.

h) Las variables explicativas $x_1, x_2, x_3, \ldots, x_k$ se distribuyen independientemente de los errores aleatorios, es decir; no hay multicolinealidad. Por otro lado, la matriz X contiene k columnas, correspondientes a las k variables explicativas, y T filas, correspondientes al número de observaciones, por lo que se tienen las siguientes implicaciones:

1. El número de observaciones, T, debe ser igual o mayor que el número de variables explicativas k ($T \geq k$), lo cual tiene sentido pues para estimar k parámetros, se necesita al menos k observaciones para una regresión lineal múltiple.
2. Cada variable explicativa debe ser linealmente independiente, lo que implica que no existen relaciones lineales exactas entre ellas. Si una variable explicativa es una combinación lineal exacta de otras, entonces se dice que hay *multicolinealidad perfecta*, y el modelo no puede estimarse. Cabe mencionar que si existe una relación aproximada se puede estimar los párametros aunque el modelo será menos fiable, en este caso se dice que existe una *multicolinealidad no perfecta*.

1.4. Estimación de párametros por mínimos cuadrados

Al igual que en el modelo de regresión lineal simple, se puede aplicar el método de mínimos cuadrados para estimar los coeficientes β_j o en términos vectoriales β al minimizar la suma de cuadrados residual (SS_{RES}):

$$SS_{RES} = \sum_{i=1}^{T} \hat{\varepsilon}_i^2 = \sum_{i=1}^{T} (y_i - \hat{y}_i)^2 ,$$

donde:

$\hat{\varepsilon}_i = y_i - \hat{y}_i$ son los residuales o error estimado y

$\hat{y}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki}$ son los valores ajustados bajo el modelo.

Al ser más cómodo manejar modelos de regresión múltiple en notación matricial, se presenta la forma compacta:

$$\hat{Y} = X\hat{\beta},$$
$$\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\beta},$$

con $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ el vector de estimadores de mínimos cuadrados de los coeficientes de regresión $\beta_1, \beta_2, ..., \beta_k$.

Obteniendo, la siguiente suma residual a minimizar:

$$SS_{RES} = \sum_{i=1}^{T} \hat{\varepsilon}_i^2 = \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = \hat{\varepsilon}'\hat{\varepsilon},$$

donde:

$\left(Y - X\hat{\beta}\right)'$ y $\hat{\varepsilon}'$ son vectores transpuestos. De igual forma, la suma residual se puede expresar como:

$$SS_{RES} = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta = Y'Y - 2\beta'X'Y + \beta'X'X\beta,$$

ya que $\beta'X'Y$ es una matriz de 1x1 y su transpuesta es igual a $Y'X\beta$.

Los estimadores de mínimos cuadrados deben satisfacer:

$$\frac{\partial SS_{RES}}{\partial \beta} \mid_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0.$$

Una vez minimizada SS_{RES}, se obtiene el llamado sistema de ecuaciones normales de mínimos cuadrados, en notación matricial:

$$(X'X)\,\hat{\beta} = X'Y.$$

Con la finalidad de resolver el sistema respecto a $\hat{\beta}$, es preciso que el rango de la matriz $X'X$ sea igual a k lo cual es consecuencia de que no exista multicolinealidad para X. Si esto se cumple, $X'X$ es invertible y se multiplican ambos lados por la inversa de $X'X$:

$$(X'X)^{-1}\,(X'X)\,\hat{\beta} = (X'X)^{-1}\,X'Y.$$

Y como $(X'X)^{-1}\,(X'X) = I$, se obtiene la expresión del vector de estimadores de mínimos cuadrados ordinarios:

$$\hat{\beta} = (X'X)^{-1}\,X'Y.$$

Con las siguientes propiedades:

- $\hat{\beta}$ es un estimador insesgado β, es decir, $\mathbb{E}[\hat{\beta}] = \beta$

Prueba:

$\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathrm{X'X})^{-1}\,\mathrm{X'Y}] = \mathbb{E}[(\mathrm{X'X})^{-1}\mathrm{X'}(\mathrm{X}\beta + \varepsilon)] = \mathbb{E}[(\mathrm{X'X})^{-1}\mathrm{X'X}\beta + (\mathrm{X'X})\mathrm{X'}\varepsilon]$

con $\mathbb{E}(\varepsilon) = 0$ y $(\mathrm{X'X})^{-1}X'X = I$. Entonces, $\mathbb{E}[\hat{\beta}] = \beta$.

- La varianza $\hat{\beta}$ se expresa con la matriz de covarianza dada por $\mathrm{Cov}(\hat{\beta}) = \mathbb{E}[[\hat{\beta} - \mathbb{E}(\hat{\beta})][\hat{\beta} - \mathbb{E}(\hat{\beta})]']$.

Prueba:

$$\mathrm{Cov}\left(\hat{\beta}\right) = \mathbb{E}[[\hat{\beta} - \mathbb{E}(\hat{\beta})][\hat{\beta} - \mathbb{E}(\hat{\beta})]'] = \mathbb{E}[[\hat{\beta} - \beta][\hat{\beta} - \beta]'] = \mathbb{E}[[\mathrm{X'X}]^{-1}\mathrm{X'}\varepsilon\varepsilon'\mathrm{X}[\mathrm{X'X}]^{-1}]$$

$$= [\mathrm{X'X}]^{-1}\mathrm{X'}\mathbb{E}(\varepsilon\varepsilon')\mathrm{X}[\mathrm{X'X}]^{-1} = [\mathrm{X'X}]^{-1}\mathrm{X'}\mathbb{E}(\sigma^2\mathrm{I})\mathrm{X}[\mathrm{X'X}]^{-1} = \boldsymbol{\sigma^2\,(\mathrm{X'X})^{-1}}.$$

Finalmente, es necesario señalar que, el teorema de Gauss-Markov establece al estimador de mínimos cuadrados $\hat{\beta}$ como el mejor estimador lineal insesgado; es decir, es el estimador lineal que tiene la mínima varianza (ver Apéndice A).

1.5. Varianza residual

Se puede obtener un estimador de σ^2 a partir de la suma de cuadrados residual de la variable Y respecto a su media o mejor dicho de los residuales, la cual al estimar k parámetros en el modelo de regresión asocia $T-k$ grados de libertada a la suma de cuadrados. No se omite señalar que, el estimador de σ^2 depende del modelo ajustado. Siendo así, el cuadrado medio residual es:

$$CMr = \frac{SS_{RES}}{T-k} = \frac{\sum_{i=1}^{T} \hat{\varepsilon}_i^2}{T-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{T-k}.$$

Ahora, mediante el siguiente proceso se observa que, el valor esperado de CMr es σ^2, lo cual deriva que el estimador insesgado de σ^2 es $\hat{\sigma^2} = CMr$.

Prueba:

$$\begin{aligned}
\mathbb{E}[SS_{RES}] &= \mathbb{E}[(Y-\hat{Y})'(Y-\hat{Y})] \\
&= \mathbb{E}[(Y - X(X'X)^{-1}X'Y)'(Y - X(X'X)^{-1}X'Y)] \\
&= \mathbb{E}(Y'[I - X(X'X)^{-1}X']Y) \\
&= Tr(I - X(X'X)^{-1}X']\sigma^2 I) + \mathbb{E}(Y)'[I - X(X'X)^{-1}X']\mathbb{E}(Y) \\
&= (T-k)\sigma^2,
\end{aligned}$$

donde Tr es el operador traza.

Como resultado $\mathbb{E}(\text{CMr}) = \mathbb{E}(\frac{SS_{RES}}{T-k}) = \sigma^2$, por lo que un estimador insesgado de σ^2 es:

$$\hat{\sigma^2} = CMr.$$

1.6. Prueba de hipótesis

Una vez estimados los parámetros del modelo, se puede obtener la significancia conjunta de todos los parámetros de la regresión y/o cuáles regresores resultan de interés. Esto es posible al obtener distribuciones asociadas a funciones de $\hat{\beta}$, lo cual nos permite hacer inferencia a través de pruebas de hipótesis; tales pruebas requieren que los errores sean independientes y tengan una distribución normal.

1.6.1. Prueba de la significancia de la regresión.

La prueba de significancia de la regresión es una de las pruebas más usadas, determina si existe una relación lineal entre la variable respuesta Y y alguna(s) de las variables regresoras. Las hipótesis son:

$$H_0 : \beta_2 = ... = \beta_k = 0$$

$$vs$$

$$H_1 : \beta_j \neq 0 \ \textit{para al menos una } j$$

Al rechazar la hipótesis nula se obtiene evidencia para pensar que al menos una de las variables regresoras contribuye significativamente al modelo. Ahora:

- La suma residual o variabilidad no explicada por el modelo está dada por:
$$SS_{RES} = \hat{\varepsilon}'\hat{\varepsilon} = \left(y - X\hat{\beta}\right)' \left(y - X\hat{\beta}\right) = y'y - \hat{\beta}'X'y.$$

- La variabilidad explicada por el modelo está dada por:
$$SS_R = \hat{\beta}'X'y - \frac{(\sum_{i=1}^{T} y_i)^2}{T}.$$

En este caso:

$$F = \frac{SS_R/k}{SS_{RES}/(T-k-1)},$$

es el estadístico asociado a la prueba cuya distribución asociada si H_0 es cierta es $F_{k,T-k-1}$, por tanto se rechaza H_0 si $F > F_{\alpha,k,T-k-1}$ con un nivel de significancia α. Para poder ver en forma más detallada la expansión de cada uno de los elementos integrantes de la ecuación se recomienda consultar Douglas C. Montgomery (2006). *Introducción al Análisis de Regresión Lineal*. Compañía Editorial Continental. Tercera Edición. (páginas 75 y 80).

Tradicionalmente, la información de la prueba antes presentada se asocia con la tabla de análisis de varianza (tabla ANOVA en el caso de regresores categóricos y tabla ANCOVA en el caso de regresores continuos y categóricos), siendo no indispensable para realizar la prueba de hipótesis, sin embargo la mayoría de los paquetes estadísticos la producen.

1.6.2. Prueba sobre un parámetro de la regresión.

Los coeficientes individuales en una regresión lineal múltiple resultan de gran interés, ya que el modelo podría ser más eficiente (en el sentido de utilidad del modelo) con la inclusión de regresores adicionales que tengan valor para explicar la respuesta, o quizás con la omisión de uno o más que no ayuden a explicar la respuesta. Las hipótesis para probar dicha significancia son:

$$H_0 : \beta_j = 0$$

$$vs$$

$$H_1 : \beta_j \neq 0$$

para $j = 1, ..., k$.

El estadístico de prueba es:

$$t = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 b_{jj}}},$$

donde b_{jj} es el elemento j-ésimo en la diagonal de la matriz $(X'X)^{-1}$. Se rechaza H_0 si $|t| > t_{\frac{\alpha}{2},T-k-1}$, donde $t_{\frac{\alpha}{2},T-k-1}$ es el cuantil asociado a una distribución t con $T - k - 1$ grados de libertad y que acumula una probabilidad de $\frac{\alpha}{2}$, por otro lado, si H_0 no se rechaza, entonces los datos dan evidencia para que parezca adecuado eliminar x_j del modelo.

1.6.3. Prueba de hipótesis general.

La teoría general de las pruebas de hipótesis antes mencionadas radica en la necesidad de decir algo sobre los coeficientes β, usando los coeficientes estimados $\hat{\beta}$, por lo que, el procedimiento general supone las siguientes hipótesis:

$$H_0 : R\beta = a \ \ vs \ \ H_1 : R\beta \neq a,$$

donde R es una matriz de constantes $r \times k$, con rango r, es decir, las r restricciones lineales son linealmente independientes y para la cual, se presenta el siguiente estadístico de prueba:

$$F = \frac{\left[(R\hat{\beta} - a)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - a)\right] /r}{SS_{RES}/(T-k)},$$

se rechazará H_0 si $F > F_{\alpha,r,T-k}$. Para poder profundizar en esta prueba de hipótesis se recomienda consultar Douglas C. Montgomery (2006).

Introducción al Análisis de Regresión Lineal. Compañía Editorial Continental. Tercera Edición. (páginas 89-92).

1.7. Medidas de bondad de ajuste

Es conveniente contar con criterios de bondad de ajuste, capaces de permitirnos comparar distintos modelos ajustados a una misma muestra, con la finalidad de seleccionar el modelo más apropiado.

1.7.1. Coeficiente de determinación.

El coeficiente de determinación es una medida con la que se mide el grado de variación de y modelada o explicada mediante el empleo de los regresores $x_1, x_2, ..., x_k$. Este es definido como:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{RES}}{SS_T} \text{ , donde } SS_T = SS_{RES} + SS_R,$$

al igual que en la regresión lineal simple $0 \leq R^2 \leq 1$, los valores cercanos a 1 implican que la mayor parte de la variabilidad está siendo explicada por el modelo de regresión. En general, R^2 aumenta a medida que aumenta la dispersión de las variables regresoras y disminuye cuando disminuye dicha dispersión, siempre y cuando sea correcta la forma supuesta del modelo.

Al interpretar R^2 se deben tener en cuenta lo siguiente:

1. El coeficiente de determinación mantiene o aumenta su valor al añadir nuevos regresores.
2. Si el modelo no tiene término independiente, el coeficiente no tiene interpretación clara.
3. El coeficiente no puede ser usado para comparar modelos con diferentes variables dependientes.

1.7.2. Coeficiente de determinación ajustado.

Se define el coeficiente de determinación ajustado de la siguiente manera:

$$\bar{R}_k{}^2 = 1 - (1 - R^2)\,\frac{T-1}{T-k}.$$

Este coeficiente impone una penalización al añadir otros regresores al modelo, por lo que, incrementa con la introducción de regresores, si, y solo si, el estadístico t asociado al constraste de significancia de dicho regresor satisface $t > 1$.

1.7.3. Criterio de información de Akaike (AIC) y criterio de Schwarz (SC).

El AIC (*Akaike's Information Criterion*) y el criterio de Schwarz son medios frecuentes para la selección de modelo. El AIC es un indicador que permite seleccionar un modelo estadístico basándose en la teoría de información y en las propiedades del método de máxima verosimilitud, para calcularlo se requiere uno de los supuestos de la regresión lineal múltiple, y es que, los errores tengan una distribución normal. Se expresa de la siguiente forma:

$$AIC = 2k - 2In(l),$$

donde, k es el número de parámetros y l es el máximo valor de la función de verosimilitud para el modelo estimado. La idea fundamental es tomar como estimación del parámetro estudiado el valor que haga máxima la probabilidad de obtener la muestra observada. Por otro lado, el criterio de Schwarz se basa en la función de probabilidad y está muy relacionado con el AIC, siendo expresado como sigue:

$$SC = kln(T) - 2In(l).$$

Ambos indicadores señalan que hay mejores ajustes mientras más bajos sean los valores de estos. Además no tienen cotas a diferencia de la R^2.

Penalizan la introducción de nuevos regresores, se pueden aplicar a modelos estadísticos sin término independiente, no son medidas relativas, como lo son los coeficientes de determinación, y se pueden aplicar para comparar modelos donde las variables endógenas (variable cuyo valor está determinado por las relaciones establecidas dentro del modelo en el que está incluida) son diferentes. Conceptos y derivaciones del AIC y SC se puede consutar en G. Kitagawa y Sadanori Konishi (2008). *Information Criteria and Statistical Modeling*. Springer.

Capítulo 2

Modelo Lineal Mixto

2.1. Introducción

Los modelos lineales mixtos (MLM), como mencionan Eugene Demidenko (2004). *Mixed Model*s y Galecki (2007). *Linear Mixed Models*, son modelos estadísticos que permiten analizar datos agrupados o medidas longitudinales o repetidas. Proporcionan una herramienta de interés para la modelización al considerar una variable respuesta y variables explicativas que pueden corresponder a efectos fijos (variables continuas o categóricas que modifican la media de la variable respuesta) o aleatorios (variables cuyos niveles son seleccionados aleatoriamente de una población).

Una de las principales características del éxito de estos modelos, es que, pueden combinar efectos fijos y efectos aleatorios de manera simultánea y son particularmente adecuados por ser flexibles para combinar con eficacia las distintas fuentes de información y modelizar adecuadamente las distintas fuentes de error.

Los MLM tienen una amplia aplicación que se ha extendido en diversas disciplinas, como por ejemplo en biología, entre las que se encuentran los estudios en ecología del comportamiento y ámbito de la ciencia forestal, para más detalle se recomienda consultar el documento ¿Modelos mixtos (lineales)? Una introducción para el usuario temeroso, elaborado por J.Seoane del grupo de ecología terrestre de la Univerdidad Autónoma de Madrid. Este capítulo introduce, para que sea claro para el lector, la teoría para el cálculo de los estimadores de un modelo lineal mixto de dos niveles, pero esta se puede generalizar para más niveles, con la debida precaución ya que la notación algebraica suele ser más compleja.

2.2. Tipos de datos

Existen diferentes tipos de datos: medidas repetidas, datos longitudinales y datos agrupados.

Los **datos de medidas repetidas** son aquellos donde la variable respuesta se mide más de una vez en la misma unidad de análisis. Los **datos longitudinales** son aquellos donde la respuesta es observada varias veces en diferentes tiempos para cada unidad involucrada en el estudio y donde no se observan necesariamente en los mismos tiempos, ni se observan necesariamente el mismo número de veces. Así por ejemplo, se pueden registrar diferentes medidas de presión arterial para cada paciente en diferentes días.

Otra clase de datos, son los **datos agrupados**, en los que existe un diseño jerárquico y donde la variable dependiente se mide una vez para cada unidad de análisis, y las unidades de análisis se agrupan dentro de grupos de unidades. Por ejemplo, un conjunto de datos agrupados de dos niveles es el ingreso de una persona (unidad de análisis), anidado dentro de hogares (grupos de unidades).

2.3. Efectos fijos y aleatorios

Robinson (2008). *IcebreakeR* comenta que los modeladores pueden estar en desacuerdo si los efectos deben ser fijos o aleatorios, y el mismo efecto puede cambiar dependiendo de las circunstancias. Ciertamente, los estadísticos no están de acuerdo en una estrategia. Algunos dicen que depende completamente de la inferencia, y algunos que depende completamente del diseño. Por lo que, la distinción entre factores fijos y aleatorios relacionados en una variable dependiente resulta ser de suma importancia para el contexto de los MLM.

a) **Efecto fijo:** Son variables categóricas o continuas que modifican la media de la variable respuesta Y; como el peso o género de una persona. También llamadas coeficientes de regresión o parámetros de efectos fijos y desconocidas en un MLM siendo estimadas en función de los datos.

b) **Efecto aleatorio:** Variables aleatorias asociadas con los niveles de una población. Estos efectos influencian la variabilidad sobre la respuesta Y y representan la desviación con respecto a las relaciones dadas por los parámetros de efectos fijos.

Es importante entender que los papeles de los efectos fijos y los efectos aleatorios son diferentes: los efectos fijos ***explican*** variabilidad y los efectos aleatorios ***organizan*** variabilidad, para ilustrar esto, considere el caso de un experimento para comparar los tiempos de secado de una pintura, la pintura proviene de diferentes marcas y se cuenta con información sobre la temperatura ambiente y el porcentaje de humedad; estas dos últimas variables serían los efectos aleatorios y las marcas de la pintura, los efectos fijos. Siendo así, la marca explica la variabilidad en la variable respuesta (el secado) y tanto la temperatura, como la humedad, organizan tal variabilidad.

2.4. MLM de dos niveles

Los datos agrupados se definen como conjuntos de datos en los que la variable dependiente se mide una vez para cada observación (unidad de análisis) y las unidades de análisis se agrupan o anidan dentro de grupos de unidades, también conocidos como datos jerárquicos, ya que las observaciones se pueden colocar en niveles de una jerarquía.

Para el conjunto de datos agrupados que se usará en esta tesis, el nivel 1 se refiere al hogar (observaciones en el nivel más detallado/unidades de análisis) y el nivel 2 a las entidades federativas (representa el siguiente nivel de la jerarquía/grupos de unidades). Por lo tanto, las especificaciones presentadas en esta sección se refieren a un modelo para un conjunto de datos agrupados de dos niveles.

La variable dependiente, que se mide en cada unidad de análisis, es siempre de nivel 1, en tanto que, las variables explicativas pueden ser de nivel 1 o 2 al medir características de los grupos o las unidades de análisis y se considera intercepción aleatoria que incluyen solo un efecto aleatorio único asociado con la intercepción de cada grupo.

2.4.1. Especificación general del MLM para una observación.

Se define la fórmula general del modelo lineal mixto (MLM), la cual indica como el modelo se puede escribir al nivel de una observación individual en el contexto de un conjunto de datos de dos niveles:

$$Y_{ji} = X_{ji}\beta + Z_{ji}b_i + \varepsilon_{ji},$$

donde:

$i = 1, \ldots, m$, con m el número de conglomerados.

$j = 1, \ldots, n_i$, con n_i el número de individuos en el conglomerado i.

Y_{ji} la variable respuesta del j ésimo miembro del conglomerado i.

$X_{ji}\beta$ es el elemento fijo del modelo.

$Z_{ji}b_i$ es el elemento aleatorio del modelo.

ε_{ji} es el error.

Todo lo anterior se puede aterrizar matricialmente de la siguiente manera:

$$X_{ji} = \begin{bmatrix} X_{ji}^1 \\ \vdots \\ X_{ji}^p \end{bmatrix};\ Z_{ji} = \begin{bmatrix} Z_{ji}^1 \\ \vdots \\ Z_{ji}^q \end{bmatrix};\ b_i = \begin{bmatrix} b_i^1 \\ \vdots \\ b_i^q \end{bmatrix};\ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Con X el vector de covariables correspondiente a los p efectos fijos (características invariables en el tiempo o variables según el tiempo para cada medición) y Z el vector de covariables correspondiente a los q efectos aleatorios; pueden ser continuas o categóricas, además, se supone que los efectos aleatorios y los errores son variables aleatorias independientes entre sí.

2.4.2. Especificación general del MLM para datos agrupados.

A partir de las fórmulas para observaciones individuales mencionadas en el apartado anterior, se define la especificación general de un MLM para un conglomerado i:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

con $i = 1, \ldots, m$, y la siguiente representación matricial:

$$Y_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{bmatrix},$$

Y_i : vector de dimensión n_i que representa las respuestas para el conglomerado i.

$$X_i = \begin{bmatrix} X_{i1}^1 & X_{i1}^2 & \ldots & X_{i1}^p \\ \vdots & \vdots & \vdots & \vdots \\ X_{in_i}^1 & X_{in_i}^2 & \ldots & X_{in_i}^p \end{bmatrix}; \ \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

$X_i\beta$: elemento fijo de la fórmula, compuesto por la matriz X_i que representa los p regresores para cada uno de los individuos n_i en el conglomerado (no se debe olvidar que n_i puede variar entre conglomerados y ninguna de las columnas es una combinación lineal de las restantes, es decir, no hay multicolinealidad).

En $X_i\beta$, β es un vector de parámetros de efectos fijos desconocidos asociados a los p regresores.

$$Z_i = \begin{bmatrix} Z_{i1}^1 & Z_{i1}^2 & \ldots & Z_{i1}^q \\ \vdots & \vdots & \vdots & \vdots \\ Z_{in_i}^1 & Z_{in_i}^2 & \ldots & Z_{in_i}^q \end{bmatrix}; \ b_i = \begin{bmatrix} b_{1i} \\ \vdots \\ b_{qi} \end{bmatrix},$$

$Z_i b_i$: elemento aleatorio, compuesto por la matriz Z_i donde las columnas representan valores observados para los regresores en cada conglomerado que tienen efectos en la variable respuesta.

En $Z_i b_i$, b_i representa un vector de efectos aleatorios asociados a las q covariables. Finalmente, la matriz Z_i tiene menos columnas que X_i .

$$\varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix},$$

ε_i : vector conformado con n_i errores, cabe señalar el vector puede tener diferente tamaño para los distintos conglomerados, ya que, cada conglomerado puede tener más o menos unidades de análisis.

2.5. Supuestos del MLM

Se supone que el MLM, cumple con los siguientes supuestos:

Primero: los efectos aleatorios son independientes entre conglomerados y siguen una distribución normal multivariada con media 0 y varianza D, es decir, $b_i \sim N(0, D)$, donde:

$$D = Var(b_i) = \begin{bmatrix} Var(b_{i1}) & cov(b_{i1}, b_{i2}) & \dots & cov(b_{i1}, b_{iq}) \\ cov(b_{i1}, b_{i2}) & Var(b_{i2}) & \dots & cov(b_{i2}, b_{iq}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(b_{i1}, b_{iq}) & cov(b_{i2}, b_{iq}) & \dots & Var(b_{iq}) \end{bmatrix}.$$

Segundo: se supone que los errores $\varepsilon_i, ..., \varepsilon_m$ son independientes entre conglomerados y satisfacen $\varepsilon_i \sim N(0, R_i)$ con:

$$R_i = Var(\varepsilon_i) = \begin{bmatrix} Var(\varepsilon_{i1}) & cov(\varepsilon_{i1}, \varepsilon_{i2}) & \dots & cov(\varepsilon_{i1}, \varepsilon_{in_i}) \\ cov(\varepsilon_{i1}, \varepsilon_{i2}) & Var(\varepsilon_{i2}) & \dots & cov(\varepsilon_{i2}, \varepsilon_{in_i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\varepsilon_{i1}, \varepsilon_{in_i}) & cov(\varepsilon_{i2}, \varepsilon_{in_i}) & \dots & Var(\varepsilon_{in_i}) \end{bmatrix}.$$

Tercero: como ya se había mencionado, los vectores ε_i y b_i son independientes.

2.5.1. Estructuras para la matriz D.

La matriz D se conoce como una matriz no estructurada, ya que no cuenta con restricciones adicionales a la de ser definida positiva y simétrica sobre sus elementos. La matriz D es de dimensión $q \times q$, lo cual implica, al ser simétrica, que se tienen por estimar $\frac{q(q+1)}{2}$ parámetros, los cuales se guardarán en un vector denotado por θ_D, se supone que es la misma para cada conglomerado. Por ejemplo, para una matriz D de dimensión 2×2 se tendrá el vector θ_D con 3 parámetros:

$$D = Var(b_i) = \begin{bmatrix} \sigma^2_{b1} & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma^2_{b2} \end{bmatrix}, \text{ entonces } \theta_D = \begin{bmatrix} \sigma^2_{b1} \\ \sigma_{b1,b2} \\ \sigma^2_{b2} \end{bmatrix}.$$

También se puede definir otra estructura, llamada estructura de componentes de varianza, en la que cada efecto aleatorio tiene su propia varianza y las covarianzas se definen como ceros. Siguiendo el ejemplo anterior, se tendría:

$$D = Var(b_i) = \begin{bmatrix} \sigma^2_{b1} & 0 \\ 0 & \sigma^2_{b2} \end{bmatrix}, \text{ entonces } \theta_D = \begin{bmatrix} \sigma^2_{b1} \\ \sigma^2_{b2} \end{bmatrix}.$$

Esto implica que los efectos aleatorios $b_i, ..., b_m$, asociados a los distintos conglomerados, son no correlacionados y con igual varianza. Además, se tiene homoscedasticidad ya que se cuenta con la misma estructura para la matriz D en cada conglomerado.

2.5.2. Estructuras para la matriz R_i.

Se presentan tres tipos de estructuras para la matriz asociada al error, R_i. La estructura más simple es la diagonal, en la que los residuos asociados con las observaciones sobre el mismo conglomerado no están correlacionadas y tienen una varianza igual. Donde θ_R es el conjunto de parámetros asociados a R escritos en forma vectorial, por lo que se tiene:

$$R_i = Var(\varepsilon_i) = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}, \text{ entonces } \theta_R = \begin{bmatrix} \sigma^2 \end{bmatrix}.$$

Los errores asociados a un mismo conglomerado son no correlacionados y con igual varianza. Además, si la misma estructura para la matriz se supone en todos los conglomerados, se tiene homoscedasticidad dentro y entre conglomerados.

Por otro lado, se tiene la forma general de la estructura de simetría compuesta donde se supone que los errores asociados con los valores de respuesta para el conglomerado i tienen covarianza constante σ_1 y una varianza constante $\sigma^2 + \sigma_1$. Además, en comparación con la estructura diagonal, se estiman dos parámetros:

$$R_i = Var(\varepsilon_i) = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \dots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \dots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \dots & \sigma^2 + \sigma_1 \end{bmatrix},$$

$$\text{entonces } \theta_R = \begin{bmatrix} \sigma^2 \\ \sigma_1 \end{bmatrix}.$$

Mientras que, la tercer estructura es autorregresiva de primer orden, está es generalmente usada en datos longitudinales y es conocida como $AR(1)$ con dos parámetros, un parámetro de varianza constante $\sigma^2 > 0$ y un parámetro de autocorrelación $-1 \leq \rho \leq 1$:

$$R_i = Var(\varepsilon_i) = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \dots & \sigma^2 \end{bmatrix}, \text{ entonces}$$

$$\theta_R = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}.$$

Para finalizar, no se omite señalar que también se pueden permitir varianzas heterogéneas, como por ejemplo, para diferentes dosis de medicamento (alta o baja). Suponiendo las mismas estructuras para cada una de las matrices de cada grupo (con diferentes parámetros). De acuerdo a lo anterior, se muestra la matriz R_i heterogénea para un conjunto de datos agrupados de dos niveles que provienen de un estudio en el que se asignaron al azar 30 ratas hembras para recibir una dosis de medicamento (alta o baja) y donde el objetivo fue comparar el peso al nacer de sus crías.

Las varianzas residuales heterogéneas para cada uno de los grupos de tratamiento, es decir, la varianza residual de las observaciones del peso al nacer, difieren para cada nivel de tratamiento (alta o baja) y se presenta como:

$$\theta_R = \begin{bmatrix} \sigma^2_{alta} \\ \sigma^2_{baja} \end{bmatrix}.$$

2.6. Relación con el modelo lineal marginal

La relación entre el MLM y el modelo marginal, radica en la presencia o ausencia de efectos aleatorios. Los efectos aleatorios se usan explícitamente en los MLM para explicar la variación entre sujetos o entre grupos, pero no se usan en la especificación de modelos marginales. Esta diferencia implica que el MLM permite la inferencia de sujetos específicos, mientras que el modelo marginal no. Por la misma razón, a menudo se hace referencia a los MLM como modelos específicos del sujeto, y los modelos marginales se denominan modelos de promedio poblacional.

- Modelo marginal general

La forma matricial para el conglomerado i es:

$$Y_i = X_i\beta + \varepsilon_i^*,$$

con la matriz X_i y el vector β de la misma forma que en un MLM. Donde $\varepsilon_i^* \sim N(0, V_i^*)$, con $\varepsilon_i^* = Z_i b_i + \varepsilon_i$ representando los residuos marginales y con las estructuras para la matriz R_i de los MLM válidas para $V_i^* = Var(Z_i b_i + \varepsilon_i)$. Es importante recordar que el modelo marginal no involucra efectos aleatorios por lo que, la única parte aleatoria se refiere a los residuos marginales.

- Modelo marginal implícito por un MLM

A continuación se presenta un ejemplo del modelo enunciado anteriormente, asociado al MLM especificado en la subsección 2.4.2:

$$Y_i = X_i\beta + \varepsilon_i^*,$$

donde $\varepsilon_i^* \sim N(0, V_i)$.

Ya que $\varepsilon_i^* = Z_i b + \varepsilon_i$, se calcula la $Var(Z_i b + \varepsilon_i)$, de la cual se obtiene que el modelo se define mediante la distribución marginal del vector Y_i como

$Y_i \sim N(X_i\beta, Z_iDZ_i' + R_i)$. No se omite señalar que el modelo cuenta con el mismo conjunto de parámetros de covarianzas que un MLM pero con menos restricciones, por ejemplo, los elementos diagonales de las matrices D y R_i no forzosamente deben ser positivas, es decir, el único requisito en el modelo marginal es que la matriz V_i sea positiva definida.

El concepto del modelo implícito marginal es importante por dos razones:

1. La estimación de los parámetros de efectos fijos y covarianzas en el MLM podría realizarse en el marco del modelo marginal implícito.
2. Cuando un MLM genera una estimación no positiva de la matriz D se puede aplicar un modelo marginal (ya que tiene menos restricciones).

2.7. Estimación de los parámetros fijos de un MLM

En los MLM se estima el parámetro β junto con los parámetros θ_R y θ_D (mejor especificado como vector $\theta = \begin{pmatrix} \theta_R \\ \theta_D \end{pmatrix}$) por medio de la estimación de máxima verosimilitud (ML) y la verosimilitud restringida (REML)..

2.7.1. Máxima verosimilitud (ML).

Se trata probablemente del método más empleado. Básicamente obtiene estimaciones de parámetros, para ello, la función de verosimilitud suele ser el producto de las densidades, vista como función del parámetro en el modelo especificado, con base en supuestos de distribución. Por lo que, las estimaciones de los parámetros son los valores de los argumentos que maximizan la función de verosimilitud.

A continuación se presenta para un MLM la función de densidad de probabilidad normal multivariada de β y θ, refiriéndose a la distribución marginal de la variable dependiente para el conglomerado i:

$$f\left(Y_i \mid \beta, \theta\right) = (2\pi)^{\frac{-n_i}{2}} det\left(V_i\right)^{-\frac{1}{2}} exp\left(-0.5(Y_i - X_i\beta)'V_i^{-1}(Y_i - X_i\beta)\right),$$

de la cual se deriva la función de verosimilitud para el conjunto de conglomerados:

$$\begin{aligned} L\left(\beta, \theta\right) &= \Pi_i \left(2\pi\right)^{\frac{-n_i}{2}} det\left(V_i\right)^{-\frac{1}{2}} exp\left(-0.5(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)\right) \\ &= \Pi_i L_i\left(\beta, \theta\right) \end{aligned}$$

y la función log-verosimil definida por:

$l(\beta,\theta) = lnL(\beta,\theta) = -0.5n \times ln(2\pi) - 0.5\sum_i ln(det(V_i)) - 0.5 \times \sum_i (y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta)$.

De lo anterior se desprenden dos casos para la estimación del parámetro β de efectos fijos:

a) **Cuando θ *es conocido*:** al suponer θ conocida, por consecuencia V_i también es conocida y los únicos parámetros que se estiman son los de efectos fijos, conocidos como β. Por lo que la función log-verosímil se reduce a:

$$q(\beta) = -0.5 \times \sum_i (y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta),$$

la cual, al optimizarla genera el mejor estimador de β que depende de los parámetros dados en θ :

$$\hat{\beta} = \Big(\sum_i X_i' V_i^{-1} X_i\Big)^{-1} \sum_i X_i' V_i^{-1} y_i.$$

Lo anterior, se puede estimar por medio de mínimos cuadrados generalizados (GLS) ya que, $q(\beta)$ se parece mucho a la fórmula matricial para la suma de los errores cuadrados que se minimiza en el modelo lineal estándar y la técnica GLS es comúnmente usada para la estimación de los parámetros desconocidos en un modelo de regresión lineal cuando se presenta heteroscedasticidad residual o correlación entre residuos (ver Apéndice B).

b) **Cuando θ *es desconocido*:** en este caso, se supone θ desconocida, por lo que, se considera la estimación de ambos parámetros (θ y β). De esta manera, al reemplazar el parámetro β por la expresión de $\hat{\beta}$ que se dio a conocer en el caso anterior (cuando θ es conocida) se obtiene la siguiente función de probabilidad:

$l_{ML}(\theta) = -0.5n \times ln(2\pi) - 0.5 \times \sum_i ln(det(V_i)) - 0.5 \times \sum_i (y_i - X_i\hat{\beta})' V_i^{-1}(y_i - X_i\hat{\beta})$.

Al tratarse de una optimización no lineal con restricciones de desigualdad se recurre a iteraciones computacionales para obtener las estimaciones de los parámetros de θ, es decir las estimaciones de las varianzas y covarianzas en D y R_i. Una vez que se cuenta con $\hat{D}$ y $\hat{R}_i$ respectivamente, como primer paso para calcular $\hat{\beta}$ se reemplaza en

$Var(Y_i) = V_i = Z_i D Z_i' + R_i$ (matriz de covarianzas del modelo marginal) las matrices D y R_i por sus estimadores, con el objetivo de calcular $\hat{V}_i$, obteniendo:

$$\hat{V}_i = Z_i \hat{D} Z_i' + \hat{R}_i.$$

Una vez calculada $\hat{V}_i$ se obtiene el mejor estimador empírico lineal insesgado (EBLUE) de β aplicando el método de mínimos cuadrados generalizados (GLS):

$$\hat{\beta} = (\sum_i X_i' \hat{V}_i^{-1} X_i)^{-1} \sum_i X_i' \hat{V}_i^{-1} y_i.$$

Información adicional a este cálculo es la varianza del estimador $\hat{\beta}$, dada por:

$$Var(\hat{\beta}) = (\sum_i X_i' \hat{V}_i^{-1} X_i)^{-1},$$

la cual depende de que $\hat{V}$ sea muy cercana al valor real de V y resulta ser sesgada al no considerar la incertidumbre que conlleva el reemplazar V por su estimador $\hat{V}$.

2.7.2. Máxima verosimilitud restringida (REML).

También conocida como máxima verosimilitud residual, se usa con frecuencia para eliminar el sesgo en las estimaciones de ML de los parámetros de covarianza al tener en cuenta la pérdida de grados de libertad que resulta de estimar los efectos fijos. Se basa en la función de densidad de la variable Y en términos de θ, como se muestra en seguida:

$f(Y \mid \theta) = \int f(Y \mid \beta, \theta)\, d\beta = \int L(\beta, \theta)\, d\beta,$

en consecuencia,

$$\int L(\beta,\theta)\, d\beta = \int \Pi_i (2\pi)^{\frac{-n_i}{2}} det(V_i)^{-\frac{1}{2}} exp\left(-0.5(y_i - X_i\beta)' V_i^{-1}(y_i - X_i\beta)\right) d\beta,$$

por lo que, después de algunos cálculos algebraicos se obtiene (ver Douglas M. Bate (2010). *Mixed-effects modeling with R* (página 109)):

$$\int L(\beta,\theta)\, d\beta = \Pi_i (2\pi)^{\frac{-n_i}{2}} det(V_i)^{-\frac{1}{2}} exp\left(-0.5(y_i - X_i\hat{\beta})' V_i^{-1}(y_i - X_i\hat{\beta})\right) \times (2\pi)^{\frac{n_i}{2}} det\left(A_i^{-1}\right)^{\frac{1}{2}},$$

con $\hat{\beta}$ dada en la subsección 3.7.1 y $A_i = X_i' V_i^{-1} X_i$. Para finalmente, al aplicar el logaritmo al resultado anterior obtener la log-verosimilitud

marginal que se debe maximizar para obtener el estimador de β, dada por:

$$l_{REML}(\theta) = ln(\int L(\beta, \theta)\, d\beta) = -0.5n \times ln(2\pi) - 0.5 \times (\sum_i ln(det(V_i) + (y_i - X_i\hat{\beta})' V_i^{-1}(y_i - X_i\hat{\beta})) - 0.5 \times ln(det(A_i) + .05n \times ln(2\pi)).$$

La cual es la verosimilitud generada mediante ML más una función dependiente de θ, siendo así, puede reescribirse como:

$$l_{REML}(\theta) = l_{ML}(\theta) - 0.5 \times ln(det(A_i) + .05n \times ln(2\pi)).$$

Tres notas importantes son:

1. Bajo ML o REML, como primer paso se debe estimar numéricamente las matrices D y R_i (estas estimaciones no son iguales para REML y para ML, ya que incluyen una función extra) para cálcular $\hat{V}_i$ conforme a la fórmula $\hat{V}_i = Z_i \hat{D} Z_i' + \hat{R}_i$, y así, estimar el parámetro β.
2. Las estimaciones bajo ML son preferidas cuando se quiere hacer inferencia sobre los efectos fijos, ya que en el procedimiento se maximiza la función adecuada para estimar estos parámetros.
3. Sobre los efectos aleatorios se prefiere usar REML, ya que los estimadores bajo este método son insesgados para θ.

2.7.3. Métodos de maximización para la verosimilitud.

Los métodos para maximizar la verosimilitud tanto para un ML como un REML son:

- Expected Maximization (EM): este algoritmo inicia aproximando los parámetros de las distribuciones y los usa para calcular las probabilidades de que cada unidad pertenezca a un grupo y usa esas probabilidades para re-estimar los parámetros de las probabilidades, hasta converger. Para estimar los parámetros, se debe considerar que se tiene únicamente las probabilidades de pertenecer a cada grupo y no los grupos en sí. Aunque EM garantiza convergencia, ésta puede ser a un máximo local, por lo que se recomienda repetir el proceso varias veces y por ello, se considera un algoritmo lento.
- Newton-Raphson (N-R) : consiste en tomar una aproximación inicial de los parámetros de covarianza, y a continuación obtener una aproximación más refinada. Es decir, se trata de acercarnos a la raíz de la ecuación log-verosimilitud según el método (ML o REML) por medio de una fórmula recursiva. Converge con menos iteraciones pero cada iteración tarda más tiempo.

- Puntaje de Fisher: este último es una versión modificada de N-R, la principal diferencia es que la puntuación de Fisher utiliza la matriz esperada del Hessiano en lugar de la observada y que suele ser más estable numéricamente.

Cualquiera de los antes mencionados pueden tener problemas a la hora de ser implementarlo en un paquete estadístico. Por ello, se sugiere lo siguiente: elegir valores de inicio alternativos para las estimaciones del parámetro de covarianza o usar EM (en caso de que el algoritmo no converga a valores óptimos); rescalar las covariables (en caso de que, la matriz se vuelva definida no positiva); eliminar los efectos aleatorios (esto simplificaría el método), y, por último; considerar ajustar desde el inicio un modelo marginal implícito (de este modo se ajusta un modelo más simple).

2.8. Estimación de los parámetros aleatorios de un MLM

Los valores del vector b no son fijos (son aleatorios), por esta razón, se realiza una predicción en lugar de una estimación. Para ello, es importante tener en cuenta que Y es una combinación lineal de b y ε con esperanza:

$$\mathbb{E}(Y) = X\beta + Z\mathbb{E}(b) + \mathbb{E}(\varepsilon) = X\beta$$

y matriz de covarianzas:

$$Var(Y) = V(Zb + \varepsilon) = V = ZDZ^{'} + R.$$

Bajo el supuesto de normalidad e independencia del efecto aleatorio, se tiene una distribución conjunta multivariada de $Y \sim N(X\beta, ZDZ^{'} + R)$ y $b \sim N(0, D)$ con:

$Cov(Y, b) = Cov(X\beta + Zb + \varepsilon, b) = Cov(X\beta, b) + Cov(Zb, b) + Cov(\varepsilon, b) =$

$0 + ZVar(b) + 0 = ZD.$

Aplicando, $(ZD)' = D'Z' = DZ'$ y la simetría en la matriz de varianzas y covarianzas, se tiene:

$$\begin{pmatrix} Y \\ b \end{pmatrix} \sim N\left(\begin{pmatrix} X\beta \\ 0_{mq*1} \end{pmatrix} \begin{pmatrix} V & ZD \\ DZ' & D \end{pmatrix}\right),$$

de donde se obtiene la esperanza para la distribución condicional de b dado Y, dada por:

$$\mathbb{E}(b \mid Y) = DZ^{'}V^{-1}(Y - X\beta).$$

Finalmente se obtiene el estimador del efecto aleatorio $\hat{b}$, conocido como el mejor predictor lineal insesgado (BLUP):.

$$\hat{b_i} = (DZ_i'V_i^{-1})(Y_i - X_i\hat{\beta}).$$

2.9. Pruebas de hipótesis

Las pruebas de hipótesis son herramientas útiles para tomar decisiones sobre qué modelo elegir y se especifican al proporcionar una hipótesis nula (H_0) y una alternativa (H_1) sobre los parámetros en cuestión. Estas hipótesis se pueden formular en el contexto de dos modelos que tienen en particular una relación de anidación, de esta forma; un modelo se anida dentro de otro modelo si se puede obtener un conjunto de efectos fijos y/o parámetros de covarianza en un modelo anidado imponiendo restricciones a los parámetros en un modelo más general, por ejemplo, restringiendo ciertos parámetros para que sean iguales a cero o iguales entre sí.

Prueba de razón de verosimilitud (LRT): se basa en la comparación de los valores de las funciones de probabilidad para dos modelos y pueden emplearse para probar hipótesis sobre parámetros de covarianza o parámetros de efectos fijos en el contexto de los MLM. En general, requieren que tanto el modelo anidado (nulo) y el modelo de referencia se ajusten al mismo subconjunto de datos. También se pueden obtener pruebas LRT bajo estimación REML, en cuyo caso la función L_{ML} se reemplaza por L_{REML}. En ambos casos se tiene que el LRT está definido como:

$$LRT = -2ln\left(\frac{L_{MLA}(\hat{\beta}_A,\hat{\theta}_A)}{L_{MLB}(\hat{\beta}_B,\hat{\theta}_B)}\right) = -2ln(L_{MLA}(\hat{\beta},\hat{\theta})) + 2ln(L_{MLB}(\hat{\beta},\hat{\theta})),$$

donde:

L_{MLA} representa la verosimilitud del modelo anidado.

L_{MLB} representa la verosimilitud del modelo de referencia.

El estadístico LRT sigue una distribución χ^2_{gl} en la que el número de grados de libertad, se obtienen restando al número de parámetros en el modelo anidado el número de parámetros en el modelo de referencia (suponiendo que los parámetros corresponden a variables sin multicolinealidad). Si el estadístico LRT es suficientemente grande, hay evidencia en contra del modelo de hipótesis nula y en favor del modelo de referencia. Si los valores de probabilidad de los dos modelos son muy cercanos y el estadístico LRT resultante es pequeño, se tiene evidencia a favor de la hipótesis nula.

a) Hipótesis sobre parámetros de efectos fijos.

Para este tipo de hipótesis, se tiene el modelo A (modelo anidado) y el modelo B (modelo de referencia). Se basan en la estimación de ML con las matrices de covarianzas iguales y grados de libertad igual al número de parámetros de efectos fijos que pertenecen al modelo B y que no pertenecen al modelo A. Se rechaza la hipótesis nula cuando:

$$LRT > \chi^2_{gl},$$

donde χ^2_{gl} acumula $1-\alpha$ de probabilidad.

b) Hipótesis sobre parámetros de efectos aleatorios (covarianzas).

Se basa en la estimación REML ya que reduce el sesgo inherente en las estimaciones de ML de los parámetros de covarianza y supone que los modelos anidados y de referencia tienen el mismo conjunto de parámetros de efectos fijos con diferentes conjuntos de parámetros de covarianza.

Caso 1

La hipótesis nula implica probar si algún parámetro se encuentra dentro del límite del espacio de parámetros y el estadístico de prueba está distribuido asintóticamente como un χ^2_{gl} con los grados de libertad calculados restando el número de parámetros de covarianza (las entradas de θ) del modelo de referencia al modelo anidado. Un ejemplo, es cuando se quiere comparar un modelo anidado con varianza homogénea de un modelo de referencia con varianza heterogénea, como lo es en el caso de varianzas distintas entre adultos mayores (adultos de 65 o más años de edad) y menores de 18 años de edad, donde los parámetros del modelo anidado son $\theta = \sigma^2$ y los parámetros del modelo de referencia son $\theta = \begin{pmatrix} \sigma^2_1 \\ \sigma^2_2 \end{pmatrix}$, donde σ^2_1 y σ^2_2 son las varianzas de adultos mayores y menores de 18 años respectivamente. Siendo así, es fácil ver que se está probando que el conjunto de las varianzas son iguales (homogénea) contra la alternativa de que no.

Caso 2

En este caso los parámetros de covarianza que satisfacen la hipótesis nula se encuentran en el límite del espacio de parámetros y se prueba si un efecto aleatorio dado debe mantenerse en un modelo o no. Es decir, si las varianzas y covarianzas correspondientes son iguales a cero o no. En el caso en el que se tenga un solo efecto aleatorio en un modelo se podría probar la hipótesis nula de que la varianza asociada al único efecto aleatorio es cero, es decir:

$$H_0 : El\ efecto\ aleatorio\ puede\ omitirse,$$

para la que se usa una combinación de una Ji cuadrada con cero grados de libertad, χ^2_0 y una χ^2_1 con un grado de libertad, cada una con peso $\frac{1}{2}$. Sin embargo, como la distribución «χ^2_0» acumula toda su masa en cero, sí se supone que el valor del estadístico LRT es de 50, el valo p asociado sería:

$$p = \frac{1}{2}P\left[\chi_0^2 > 50\right] + \frac{1}{2}P\left[\chi_1^2 > 50\right] = 0 + \frac{1}{2}P\left[\chi_1^2 > 50\right],$$

cuando se tienen dos efectos aleatorios y se quiere comparar si conviene agregar uno de ellos o no se usará la combinación de χ_2^2 y χ_1^2.

Para probar si hay efectos fijos, existen las siguientes pruebas alternativas.

Estadístico t: es una prueba para parámetros de efectos fijos que requiere el ajuste de solo el modelo de referencia, con las siguientes hipótesis:

$$H_0 : \beta = 0 \ \ vs \ \ H_1 : \beta \neq 0,$$

con:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})},$$

donde $SE(\hat{\beta})$ es el error estándar asociado al parámetro, en otras palabras es el elemento asociado a la variable en la diagonal de la matriz de varianzas y covarianzas que se indica en la subsección 2.7.1 inciso b. Se rechaza a un nivel α si $t < -t_l^{1-\frac{\alpha}{2}}$ o $t > t_l^{1-\frac{\alpha}{2}}$, donde $t_l^{1-\frac{\alpha}{2}}$ es el cuantil asociado a una distribución t con l grados de libertad y que acumula una probabilidad de $1 - \frac{\alpha}{2}$ (prueba de dos colas). Cabe mencionar que, para esta prueba se utilizan métodos aproximados para estimar los grados de libertad (l), por esta razón pueden variar según el software estadístico que se ocupe.

Estadístico F: este tipo de prueba se utiliza para probar hipótesis lineales sobre múltiples efectos fijos, dicho de otra manera, se cuenta con distintas combinaciones lineales independientes de los parámetros y se busca probar que dichas combinaciones valen cero. Siendo así, la prueba está dada por:

$$H_0 : R\beta = 0 \ \ vs \ \ H_1 : R\beta \neq 0.$$

Con R una matriz de dimensión $q \times p$, con $ran(R) = q$, $q \leq p$ (donde q es el número de efectos aleatorios y p el número de efectos fijos), con todas sus entradas conocidas y fijas y donde 0 es un vector de dimensión q.

El estadístico de prueba es:

$$F = \frac{(\hat{\beta}R)'(R(\sum_i X_i'\hat{V}_i^{-1}X_i)^{-1}R')^{-1}(R\hat{\beta})}{ran(R)},$$

el cual corresponde a la generalización del estadístico F que se muestra en la subsección 1.6.3 de la regresión lineal múltiple (caso particular) usando $a = 0$, como se muestra en seguida, tomando en cuenta que para la regresión lineal $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ y $\hat{Var}(\hat{\beta}) = \hat{\sigma^2}(X'X)^{-1}$, mientras que, para un MLM $Var(\hat{\beta}) = X_i'\hat{V}_i^{-1}X_i$. Siendo así, el estadístico mencionado en la subsección 1.6.3 está dado por:

$$F = \frac{\left[(R\hat{\beta}-a)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-a)\right]/q}{SS_{RES}/(T-k)} = \frac{\left[(R\hat{\beta}-a)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-a)\right]/q}{\hat{\varepsilon}'\hat{\varepsilon}/(T-k)}$$

$$= \frac{\left[(R\hat{\beta}-a)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-a)\right]/q}{CMr} = \frac{\left[(R\hat{\beta}-a)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-a)\right]/q}{\hat{\sigma^2}}$$

$$= \frac{\left[(R\hat{\beta}-a)'\left[R\hat{\sigma^2}(X'X)^{-1}R'\right]^{-1}(R\hat{\beta}-a)\right]}{q} = \frac{\left[(R\hat{\beta}-a)'\left[R\hat{V}(\hat{\beta})R'\right]^{-1}(R\hat{\beta}-a)\right]}{ran(R)}.$$

La hipótesis nula se rechaza a un nivel α si $F < F^{1-\alpha}_{ran(P),l}$, donde $F^{1-\alpha}_{ran(P),l}$ es el cuantil asociado a una distribución F que acumula $1-\alpha$ de probabilidad. Si se quiere probar que un solo efecto fijo es signicativo, la matriz P tiene que estar formada por ceros excepto en la entrada correspondiente a ese efecto, donde se da el valor 1.

Estadístico de Wald: se basa en la normalidad asintótica de los estimadores de máxima verosimilitud y máxima verosimilitud restringida, es una prueba para parámetros de covarianza símil a la prueba de hipótesis para el estadístico F. El estadístico de prueba está dado por el numerador del estadístico F:

$$W = \left(\hat{\beta}P\right)'\left(P\left(\sum_i X_i'\hat{V}_i^{-1}X_i\right)^{-1}P'\right)^{-1}\left(P\hat{\beta}\right),$$

cuya distribución es $\chi^2_{ran(P)}$ y se rechaza la hipótesis nula a un nivel α si $W > \chi^2_{ran(P)}$ que acumula $1-\alpha$ de probabilidad.

2.10. Criterios de información

Los criterios de información constituyen herramientas básicas para la selección de modelos. Los modelos no necesitan estar anidados y su objetivo es calcular una medida que indique qué tan próximos están los modelos alternativos al verdadero modelo generador (un valor más

pequeño del criterio indica un mejor ajuste). Dos de los criterios más usados son:

Criterio de Akaike (AIC): no es una prueba de hipótesis, es un criterio comparativo entre modelos, basado en ML o REML que penaliza la sobreparametrización. La fórmula es la siguiente:

$$AIC = -2 \times L(\hat{\beta}, \hat{\theta}) + 2p,$$

donde p que representa los parámetros que se estiman en el modelo para los efectos fijos y aleatorios.

Criterio de Bayes (BIC): al igual que el AIC, es un criterio paramétrico para selección de modelos, que aplica una mayor penalización para los modelos con más parámetros, a comparación del AIC. El BIC solo dice cuál de los modelos comparados es el mejor, pero no puede decir cuál es el mejor modelo para explicar los datos. Se define como:

$$BIC = -2 \times L(\hat{\beta}, \hat{\theta}) + p \times ln(n),$$

donde n es el número de observaciones utilizadas en la estimación de modelo.

2.11. Verificación de los supuestos

Si bien, la verificación de supuestos es compleja, resulta de suma importancia verificar la normalidad o en otras palabras, la distribución de los residuales, para lo cual se presentan diversas técnicas. Por ejemplo, se pueden realizar gráficos de los residuales contra los valores predichos, con la finalidad de observar un patrón, en caso de existir outliers. **a) Residual condicional.**

$$\hat{\varepsilon}_i = Y - Xi\hat{\beta} - Z\hat{b}_i,$$

es la diferencia entre el valor observado y el valor estimado.

b) Residual estandarizado.

En este método se divide los residuos por la desviación estándar estimada, obteniendo los llamados residuos de Pearson. Idealmente, se escala por sus verdaderas desviaciones estándar para obtener residuos estandarizados, pero normalmente este dato no se tiene.

Una vez que se cuenta con los residuos, para verificar la normalidad se suele apoyar de los siguientes métodos:

- Gráficos QQ: nos permiten observar qué tan cerca está la distribución de un conjunto de datos a alguna distribución ideal o comparar la distribución de dos conjuntos de datos.
- Gráficos de dispersión: es una herramienta gráfica que ayuda a identificar la posible relación entre dos variables y hace más fácil visualizar e interpretar los datos.
- Diagrama de caja: son una presentación visual que describe varias características importantes al mismo tiempo, tales como la dispersión y simetría. En ellos se puede ver claramente los datos divididos en cuatro cuartiles, donde se encuentra la mediana de los datos, los valores mínimo y máximo y los outliers, esto mediante un rectángulo, alineado horizontal o verticalmente.

2.12. Recomendaciones para la construcción del MLM

Algunas recomendaciones para la construcción de un modelo que incluya efectos aleatorios son las siguientes:

- Escoja un conjunto mínimo de efectos fijos y aleatorios para el modelo. Estos efectos deben ser tales que si ellos no están en el modelo, el modelo pierde sentido.
- Cuando los diagnósticos del MLM sugieran un modelo razonable, considere adicionar otros efectos fijos. En cada etapa reexamine los diagnósticos para asegurar una buena concordancia entre datos, modelo y supuestos.

Al finalizar la construcción del modelo, se propone evaluar la bondad del ajuste:

1. Realizar una prueba de hipótesis como verificación previa para ver si es necesario modelar los efectos aleatorios, por medio de una prueba de razón de verosimilitud (LRT) vista en la sección 2.9 inciso b caso 2.
2. Verificar el supuesto de distribución normal de los residuos (errores) como se menciona en la en la sección 2.11.
3. De ser posible, verificar normalidad de los efectos aleatorios en vista de que son estimados con predictores.

Capítulo 3

Aplicación

3.1. Introducción

A través de los modelos lineales mixtos (MLM) y el paquete de software estadístico STATA, en este capítulo se estimará variables socioeconómicas, en particular los ingresos por hogar, de una región integrada por Aguascalientes, Coahuila, Jalisco y Nuevo León, entidades federativas con mejores condiciones socioeconómicas de la República Mexicana (para mayor información sobre la contrucción de esta regiones ver «http://sc.inegi.gob.mx/niveles/datosnbi/reg_soc_mexico.pdf»). Esta regionalización fue elaborada por el Instituto Nacional de Estadística y Geografía (INEGI) en su página de internet:

«http://sc.inegi.gob.mx/niveles/index.jsp».

El objetivo específico reside en aplicar la teoría proporcionada en capítulos previos, con la finalidad de generar un modelo lineal mixto y compararlo contra un modelo de regresión lineal múltiple, para finalmente, determinar cual de estas técnicas estadísticas provee un mejor ajuste. Para ello, se dispone de los datos publicados por el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), con base en el Módulo de Condiciones Socieconómicas de la Encuesta Nacional de Ingresos y Gastos de los Hogares (MCS-ENIGH) 2014. Los datos se encuentran disponibles a nivel hogar, por lo que se realizará un modelo de dos niveles, donde el primer nivel corresponde a los hogares (observaciones) y el segundo nivel a las entidades (conglomerados).

El capítulo consta de cinco partes. En la primera se presentan las fuentes de información usadas para esta aplicación, así como su análisis descriptivo. En la segunda, se encuentran las especificaciones de los modelos, es decir, la información detallada correspondiente a definir a la variable respuesta, variables regresoras, número de conglomerados, número de individuos en cada conglomerado, entre otras. En la tercera, se presenta la estimación de los modelos, donde el objetivo es verificar si el efecto aleatorio es significativo y estimar los párametros, con el fin de obtener la estimación de la variable respuesta, el ingreso. En la cuarta, se busca validar los supuestos para cada modelo, a fin de determinar si cada modelo es adecuado. Si el modelo no es adecuado, representará incorrectamente los datos. En el último apartado, se presentan los resultados del proceso de la generación del MLM, con el propósito de dar paso a las conclusiones en el próximo capítulo.

3.2. Fuentes de información

El MCS-ENIGH es una encuesta realizada cada dos años por el INEGI en coordinación con el CONEVAL, que provee información que permite conocer el nivel de ingreso de la población y está diseñada para dar resultados a nivel nacional y de cada una de las entidades del país.

El INEGI anunció que derivado de los resultados del MCS-ENIGH 2015, que publicó en julio del 2016, donde se presentaba una discontinuidad en el ingreso de los hogares en relación con los datos del MCS-ENIGH 2014, ya que mostraba variaciones inconsistentes, respecto de lo que cabía esperar dada la dinámica de la actividad económica y del empleo ocurrida entre el 2014 y 2015. No era posible dar continuidad a la serie histórica que ya se conocía, es por ello que se dio a conocer el "Modelo Estadístico 2015 para la continuidad del MCS-ENIGH" y el "Modelo Estadístico 2016 para la continuidad del MCS-ENIGH", con el objetivo de proveer los insumos necesarios para la medición de pobreza que realiza el CONEVAL, de tal manera que se pueda dar continuidad a los resultados de la serie 2008-2014, tanto a nivel nacional como por entidad federativa.

La fuente de información que se decidió utilizar para la aplicación que se presenta en esta tesis corresponde a los generados por el CONEVAL, a través del MCS-ENIGH 2014, los cuales preservan una continuidad histórica y cuentan con variables que contribuyen a la estimación del ingreso; este es uno de los indicadores que consideran para la medición de la pobreza. La información puede ser descargada de:

«https://www.coneval.org.mx/Medicion/Paginas/Programas_BD_segunda.aspx».

3.2.1. Variables predictoras.

Como ya se señaló, la variable a predecir será el ingreso, que se define como la cantidad total de dinero que recibe una persona o un hogar, derivados del trabajo, de la renta de la propiedad, de los alquileres y de las transferencias (prestaciones sociales, seguro de desempleo, etc.) que se reciben del gobierno.

Para la estimación nos apoyaremos en factores a nivel individual y a nivel hogar como lo son:

- Laborales: prestaciones laborales y sector económico.
- Sociodemográficas: escolaridad, sexo y habla de lengua indígena.
- Características del hogar: servicios básicos y calidad y espacios de la vivienda.
- Otros ingresos en el hogar: programas de gobiernos, remesas y apoyo de otro hogar.

La lista de las variables involucradas para estimar los modelos de ingresos se presenta en las siguientes tablas, mientras que, las categorías de cada variable, e.g. indicador de personas jubiladas o pensionadas (1=Sí, 0=No) se encuentran en el Apéndice C.. LABORALES

Variable	Descripción	Tipo
jpea	Condición de actividad del jefe del hogar	categórica
cyafore	El cónyuge recibe afore como prestación laboral	categórica
cyaguin	El cónyuge recibe aguinaldo como prestación laboral	categórica
cyuplab	El cónyuge recibe utilidades como prestación laboral	categórica
jafore	El jefe del hogar recibe afore como prestación laboral	categórica
jaguin	El jefe del hogar recibe aguinaldo como prestación laboral	categórica
juplab	El jefe del hogar recibe utilidades como prestación laboral	categórica

SOCIODEMOGRÁFICAS

Variable	Descripción	Tipo
jaesc	Años de escolaridad	numérica
jsexo	Sexo del jefe del hogar	categórica

CARACTERÍSTICAS DEL HOGAR

Variable	Descripción	Tipo
ic_cv	Indicador de carencia de calidad y espacios de la vivienda	categórica
ic_sbv	Indicador de carencia de servicios básicos de la vivienda	categórica
combus2_mod	Tipo de combustible para cocinar	categórica
tam_hog	Tamaño del hogar	numérica

OTROS INGRESOS

Variable	Descripción	Tipo
ayuotr	Condición de recepción de ingresos por apoyo de otro hogar	categórica
bengob	Condición de recepción de ingresos por programas de gobierno	categórica
remesas	Condición de recepción de ingresos por remesas	categórica
jubi	Indicador de personas jubiladas o pensionadas	categórica
tenencia_viv	Tenencia de la vivienda	categórica

OTRAS

Variable	Descripción	Tipo
tamloc	Tamaño de localidad	categórica
ins_ali	Escala mexicana de seguridad alimentaria	categórica

Se decidió representar las variables categóricas con variables *dummies*; las variables *dummies* son variables binarias con valores cero y uno, tantas categorías tenga la variable de origen menos una (si no habría multicolinealidad). Por ejemplo, para la variable condición de actividad del jefe del hogar se tienen las categorías: ocupado, desempleado y población no económicamente activa, así que se crean dos variables *dummies*. Suponiendo que ocupado es la categoría de referencia, se debe generar una variable dummy para desempleado y otra para población económicamente activa. Así, la variable dummy para desempleado vale uno si el individuo está desempleado y cero en otro caso.

En nuestro caso, al usar STATA, no es necesario crear las variables *dummies* antes de incorporarlas al modelo ya que Stata las genera automáticamente al asignar el tipo de variable como factor. La forma de incorporar en Stata variables *dummies* en el modelado es a través de la siguiente instrucción (aplicada a la variable jpea en este ejemplo):

i.jpea

Cuando se generan variables *dummies* a partir de variables categóricas se debe omitir una de las categorías, este software omite la variable *dummie* para la categoría de mayor valor numérico.

3.2.2. Análisis exploratorio de datos.

Uno de los objetivos del análisis exploratorio o análisis descriptivo es que el investigador tenga una primera aproximación y pueda plantear un modelo inicial al identificar las variables más adecuadas en función del impacto en la variable a modelar, en nuestro caso, el ingreso en escala logarítmica (lnictpc).

En primer lugar se debe aprender a usar las bases de datos en STATA; las bases descargadas de la página del CONEVAL ya cuentan con el formato *.dta* reconocido por STATA; enseguida se muestra cómo usar el comando *use* para acceder a la base y empezar a trabajar:

use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Ejercicio MLM\Bases generadas\mcs14hog_reg6.dta", clear

Una vez que se abrió la base, se genera un resumen de datos de la variable respuesta, ingreso en escala logarítmico «lnictpc», así como dos histogramas para mostrar cuál es su comportamiento contra el comportamiento de la variable sin transformar (véase la figura 3.2.1 y 3.2.2).

```
. sum lnictpc

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     lnictpc |      6,928    8.050672    .9771274          0   12.76706
```

Para analizar la relación entre el ingreso y cada variable categórica, se generan diagramas de caja para el ingreso asociados a cada categoría (una gráfica por variable), los cuales presentan la siguiente información:

Variables laborales

El patrón de menores ingresos para la categoría de desempleados parece ser más evidente que en el caso de población ocupada y no económicamente activa. La distribución del ingreso parece ser simétrica en cada categoría. También se observan valores atípicos con ingresos bajos para la categoría de ocupada y no económicamente activa, así, como valores atípicos con ingresos altos en la categoría ocupada (véase la figura 3.2.3).

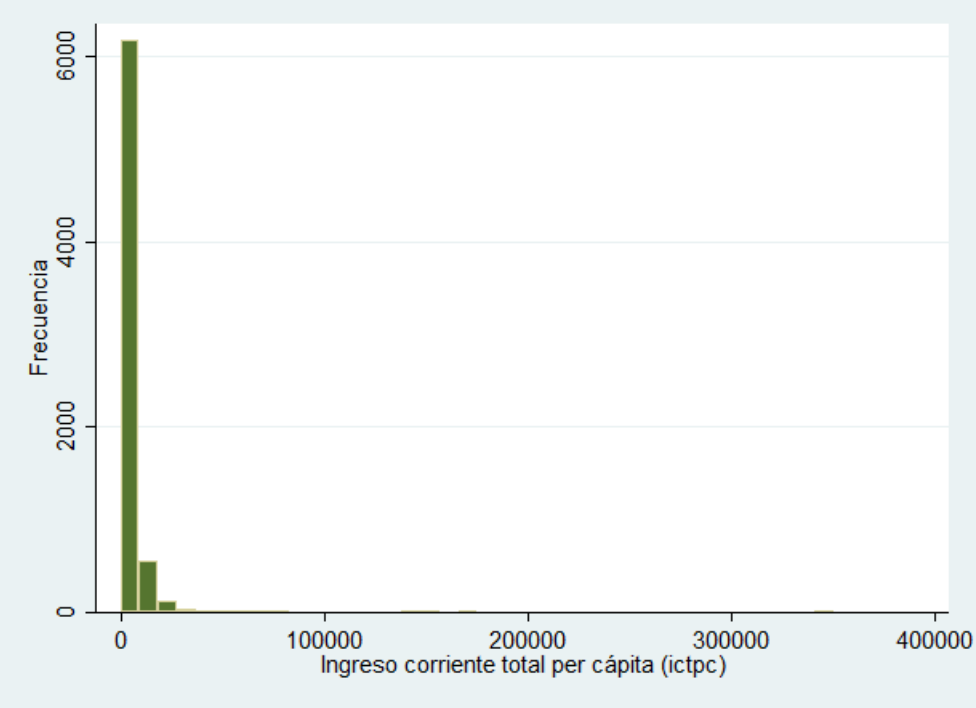

FIGURA 3.2.1. Histograma del ingreso sin tranformación logarítmica.

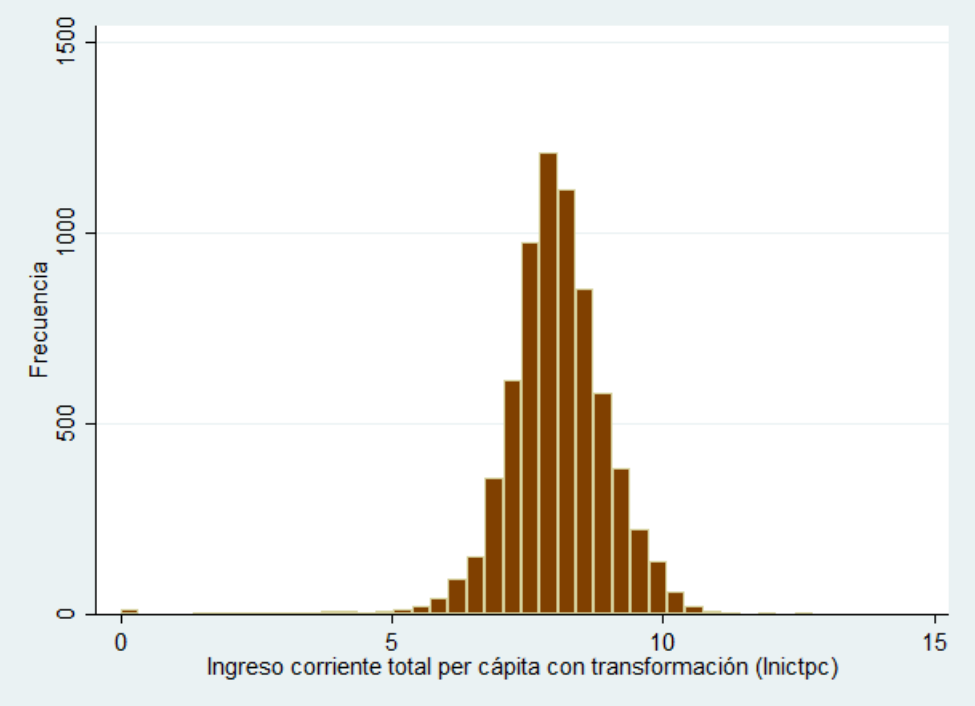

FIGURA 3.2.2. Histograma del ingreso en escala logarítmica.

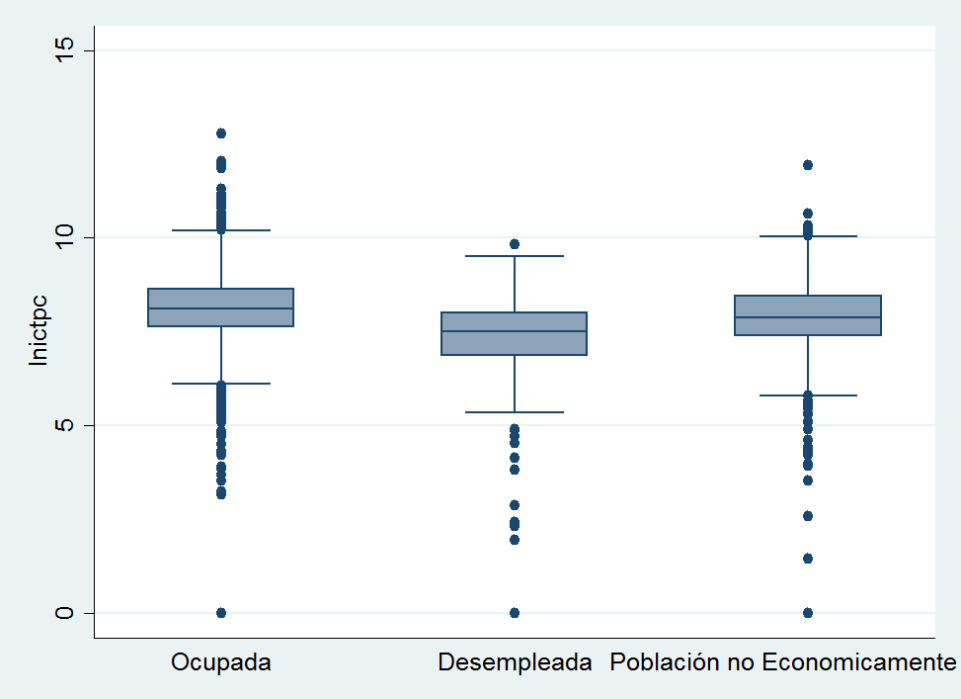

FIGURA 3.2.3. Diagrama de caja del ingreso según la categoría de condición de actividad del jefe del hogar.

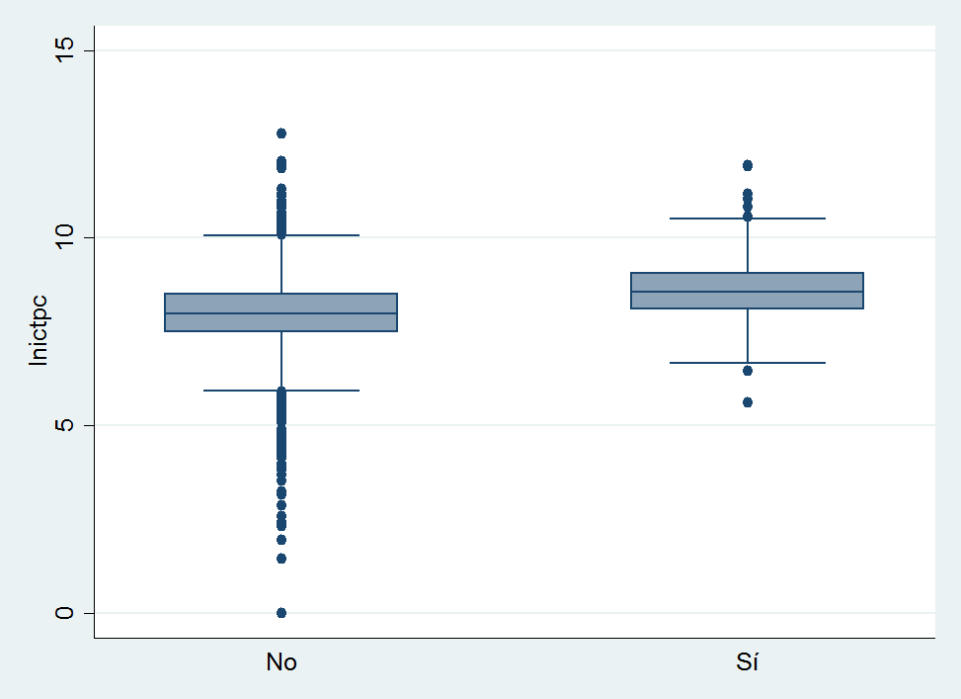

FIGURA 3.2.4. Diagrama de caja del ingreso según la categoría del cónyuge recibe afore como prestación laboral.

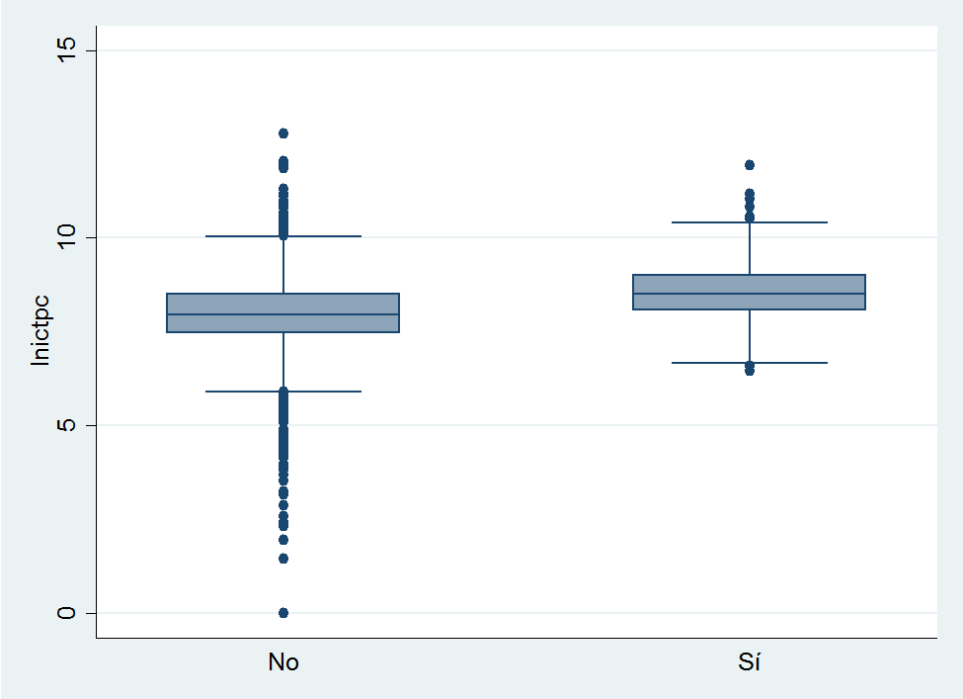

FIGURA 3.2.5. Diagrama de caja del ingreso según la categoría del cónyuge recibe aguinaldo como prestación laboral.

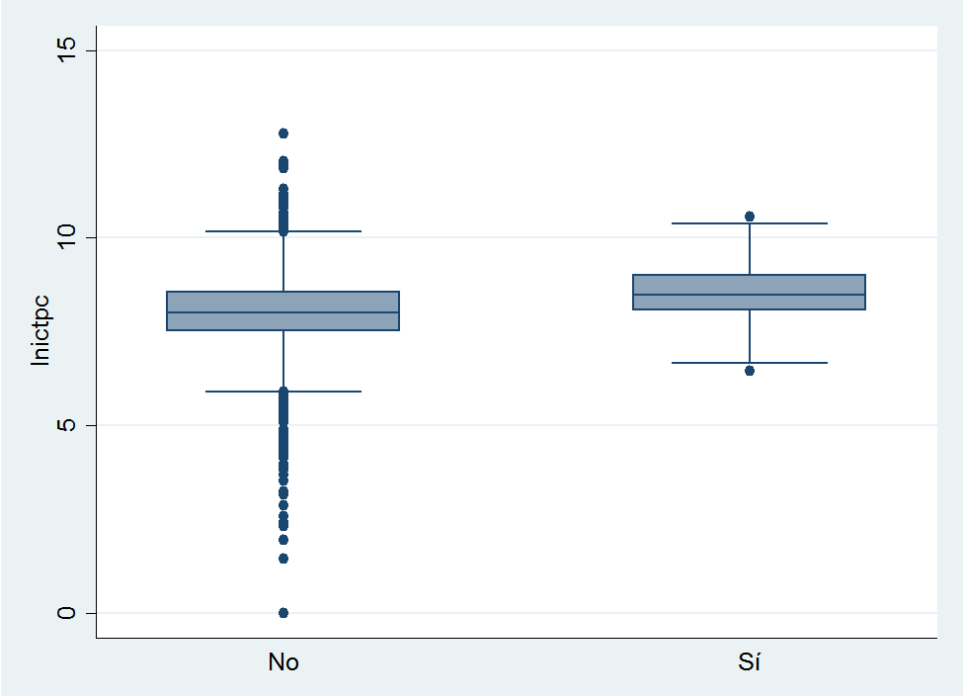

FIGURA 3.2.6. Diagrama de caja del ingreso según la categoría del cónyuge recibe utilidades como prestación laboral.

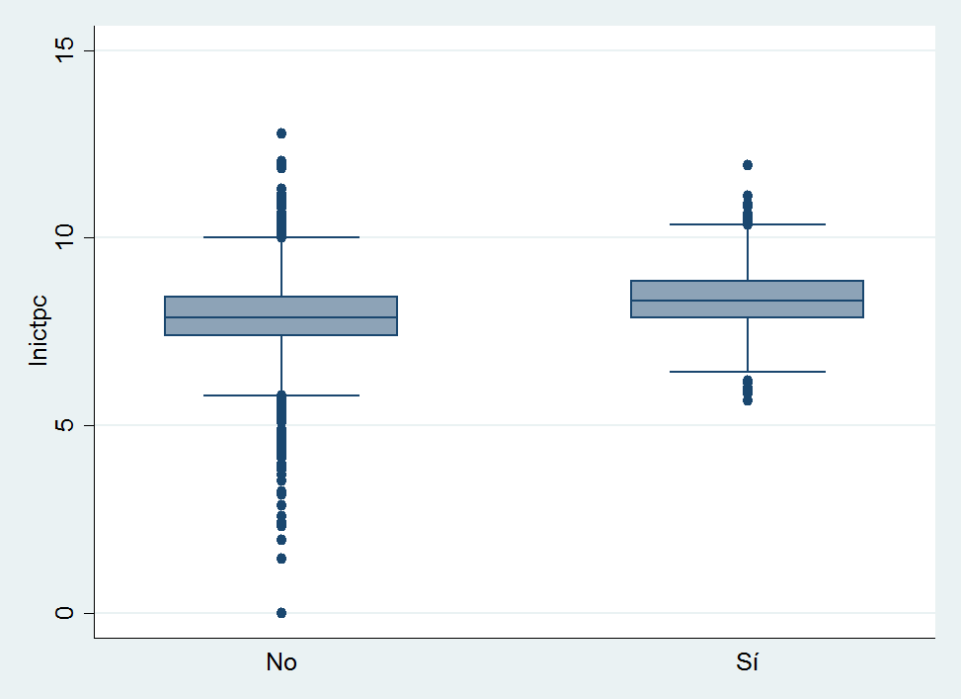

FIGURA 3.2.7. Diagrama de caja del ingreso según la categoría del jefe del hogar recibe afore como prestación laboral.

Acerca de las prestaciones laborales del cónyuge del hogar, se puede observar en las figuras 3.2.4, 3.2.5 y 3.2.6, mayor número de valores atípicos para las categorías no cuenta con afore, no cuenta con aguinaldo y no cuenta con utilidades. Además de un ingreso mayor para la categoría que sí cuenta con las prestaciones. Se puede pensar que, el contar con alguna o ninguna de estas prestaciones laborales por medio del cónyuge, impacta en la variable respuesta.

Respecto a las prestaciones laborales del jefe del hogar, se tiene el mismo comportamiento que en el caso de la variable que contempla al cónyuge; ambas variables presentan datos simétricos, además el ingreso es ligeramente mayor cuando se cuenta con prestaciones.

Variables sociodemográficas

Por otro lado, una variable que parece marcar una brecha importante entre ingresos (véase figura 3.2.10) es la de sexo del jefe del hogar, donde se ve grandes cúmulos de datos atípicos tanto de ingresos bajos, como ingresos altos en la categoría de hombre, también se puede observar simetría en los datos y mismo rango de ingreso para las observaciones que están dentro de la caja.

En el modelo se contemplan dos variables de rango continuo, una de ellas se presenta en la figura 3.2.11, la cual corresponde a un diagrama de dispersión entre ingreso y la variable años de escolaridad del jefe del hogar, aquí se observa un menor ingreso en hogares con jefes del hogar que presentan menos años de escolaridad.

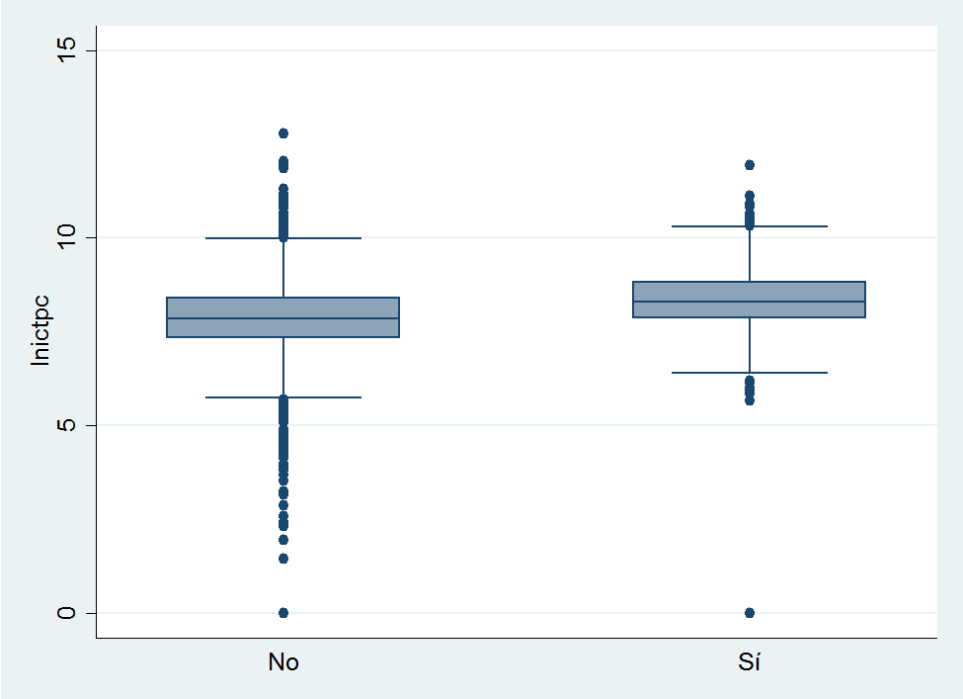

FIGURA 3.2.8. Diagrama de caja del ingreso según la categoría del jefe del hogar recibe aguinaldo como prestación laboral.

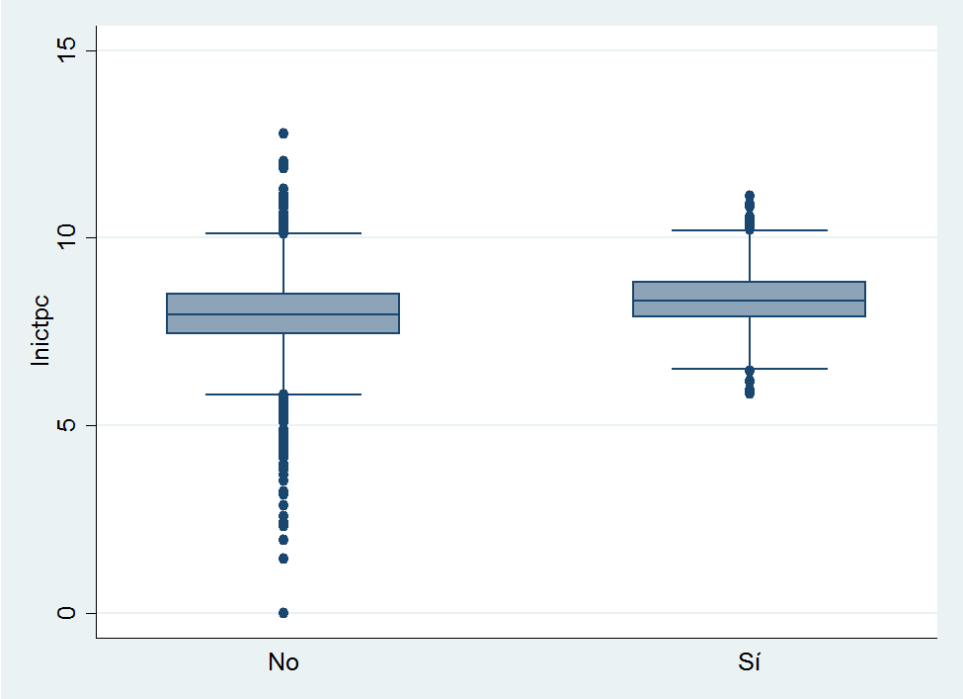

FIGURA 3.2.9. Diagrama de caja del ingreso según la categoría del jefe del hogar recibe utilidades como prestación laboral.

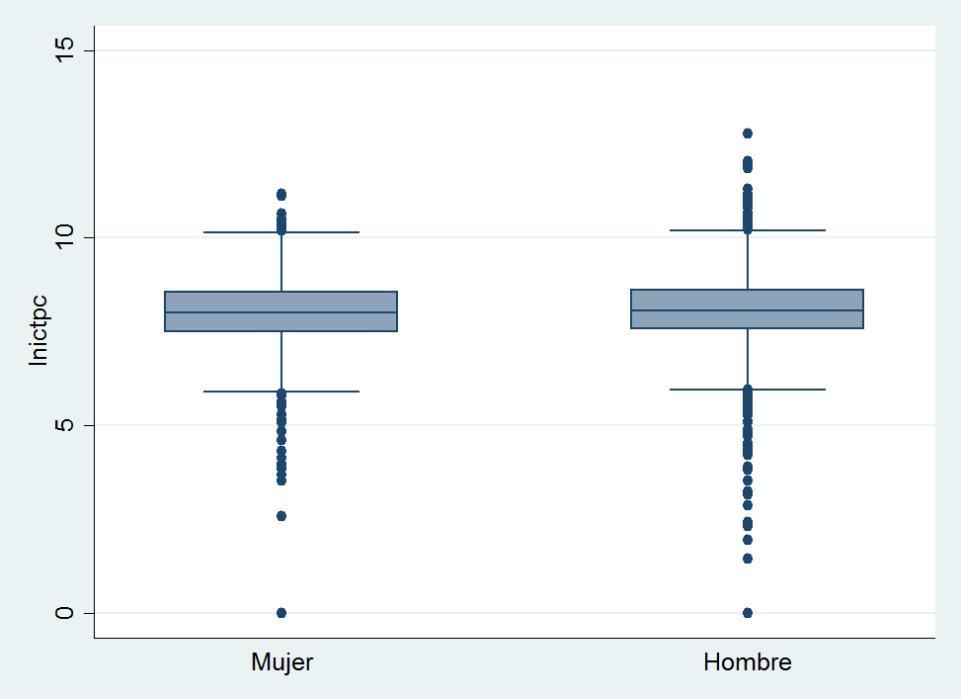

FIGURA 3.2.10. Diagrama de caja del ingreso según la categoría del sexo del jefe del hogar.

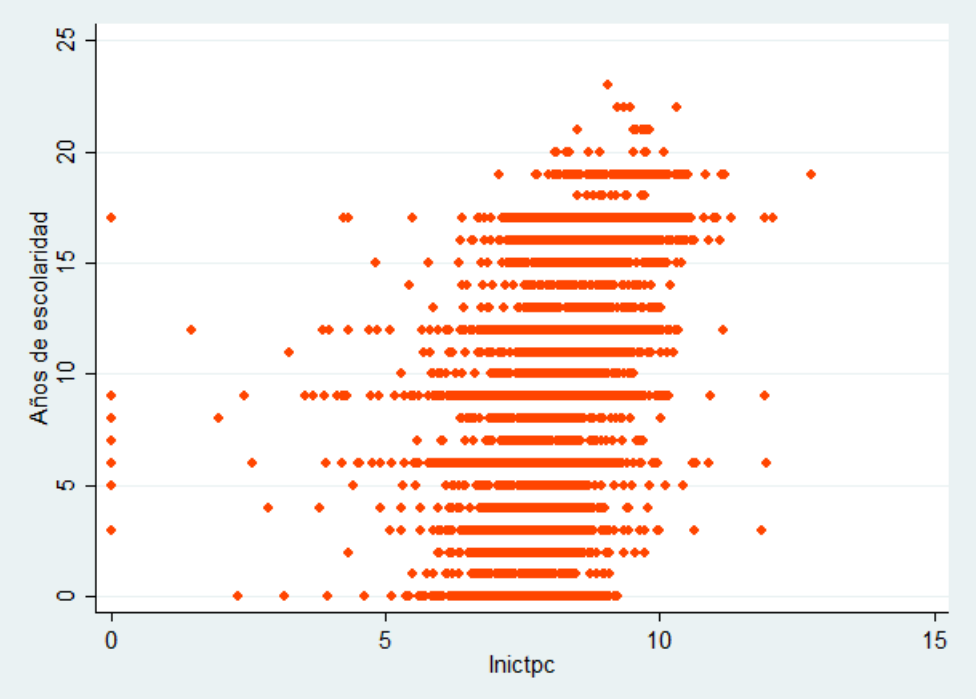

FIGURA 3.2.11. Diagrama de dispersión del ingreso versus los años de escolaridad del jefe del hogar.

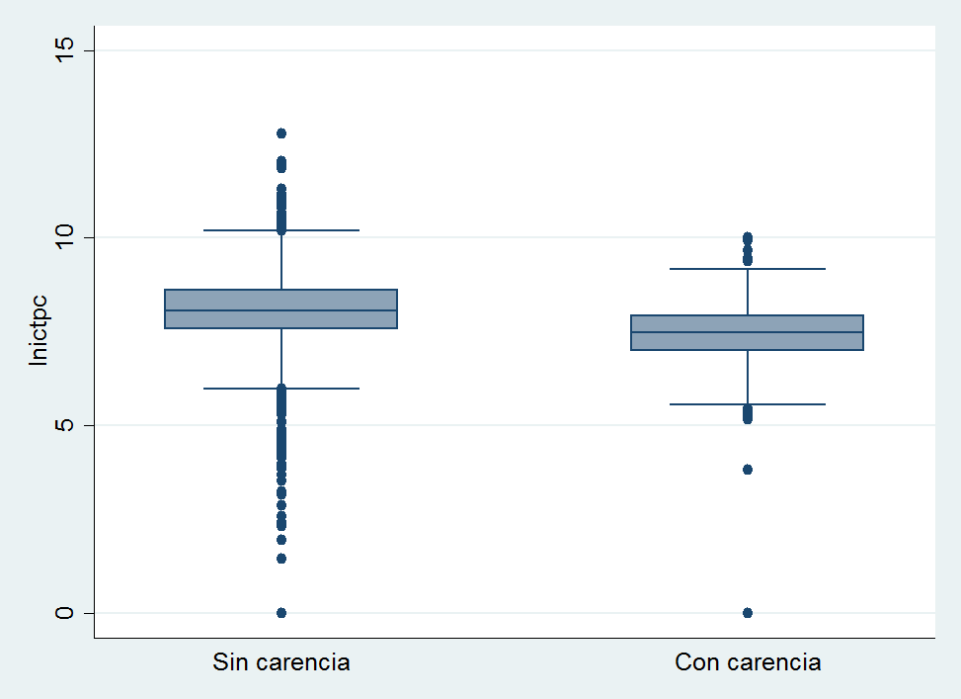

FIGURA 3.2.12. Diagrama de caja del ingreso según la categoría del indicador de carencia por acceso a los servicio de calidad y espacios de la vivienda.

Variables de características del hogar

Los hogares que presentan carencia por acceso a los servicios de calidad y espacios de la vivienda (material de contrucción de la vivienda y sus espacios) y carencia por acceso a los servicios básicos de la vivienda (agua, drenaje y luz) muestran ingresos muy por debajo de lo que se muestra en el caso contrario, véanse figuras 3.2.12 y 3.2.13. Lo mismo se puede observar en la variable, tipo de combustible para cocinar (véase figura 3.2.14). Cabe señalar que, para las tres variables antes mencionada (carencia de calidad y espacios de la vivienda, carencia por servicios básicos de la vivienda y combustible para cocinar), se observan cúmulos importantes de datos atípicos en los hogares que muestran la carencia o en su caso que utilizan leña o carbón como combustible para cocinar.

De la figura 3.2.15, correspondiente a un diagrama de dispersión, se puede ver que los hogares que presentan menor ingreso son los de menor tamaño, es decir, donde el número de personas pertenecientes al hogar están entre 1 y 6 personas.

Variables de otros ingresos

El ingreso de un hogar también puede estar conformado por ingresos no laborales, por ejemplo, por apoyo de otros hogares, programas de gobierno, remesas y por montos generados de jubilaciones. En las figuras 3.2.16, 3.2.17, 3.2.18 y 3.2.19, se observa que la recepción de ingresos por otro hogar no impacta en el ingreso significativamente y

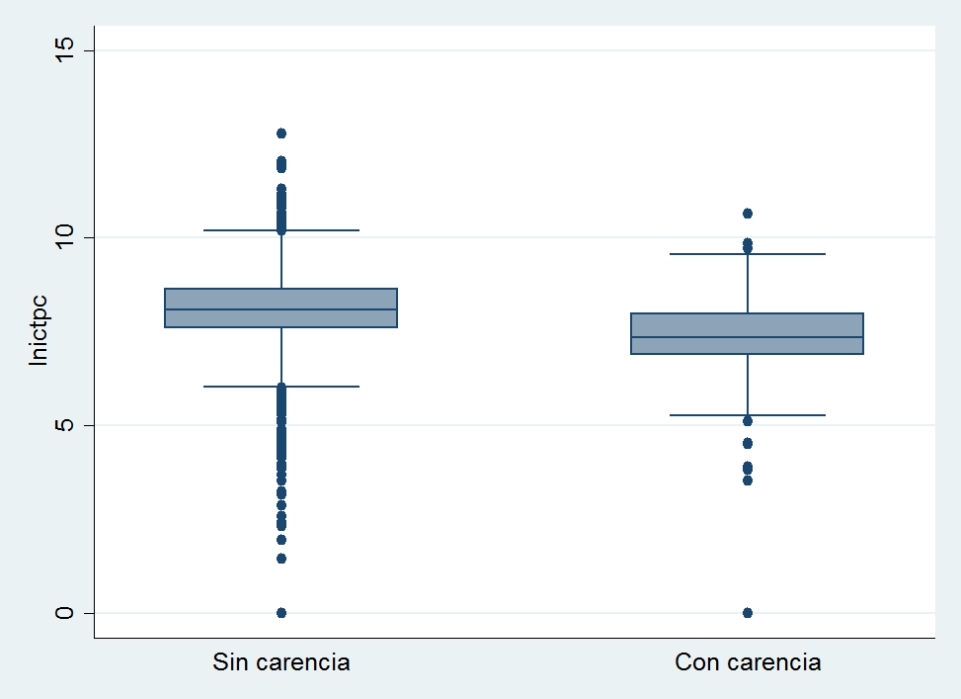

FIGURA 3.2.13. Diagrama de caja del ingreso según la categoría del indicador de carencia por acceso a los servicios básicos de la vivienda.

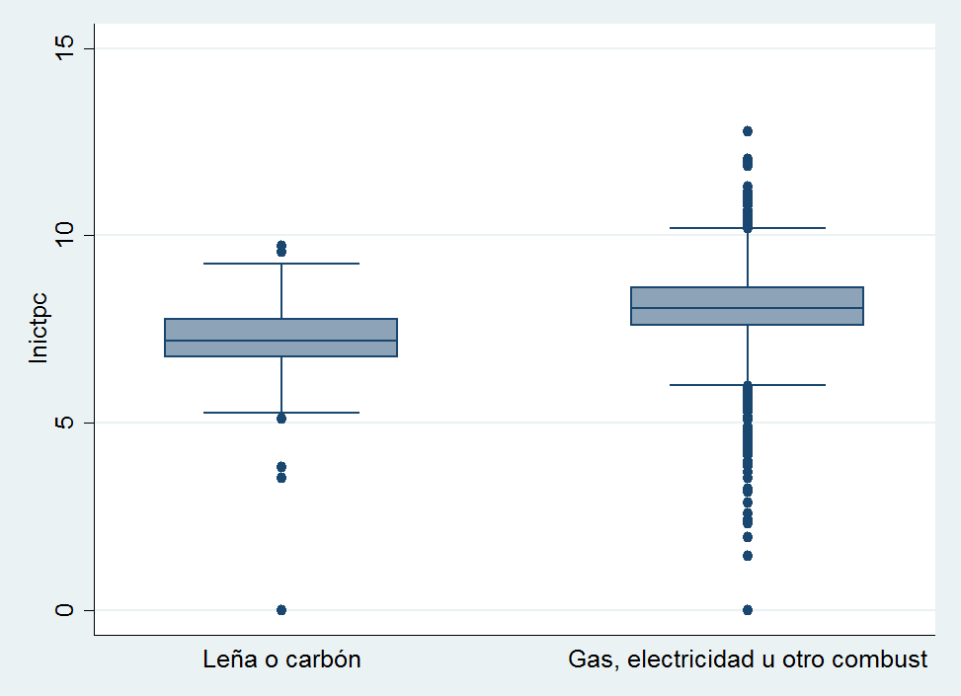

FIGURA 3.2.14. Diagrama de caja del ingreso según la categoría del tipo de combustible para cocinar.

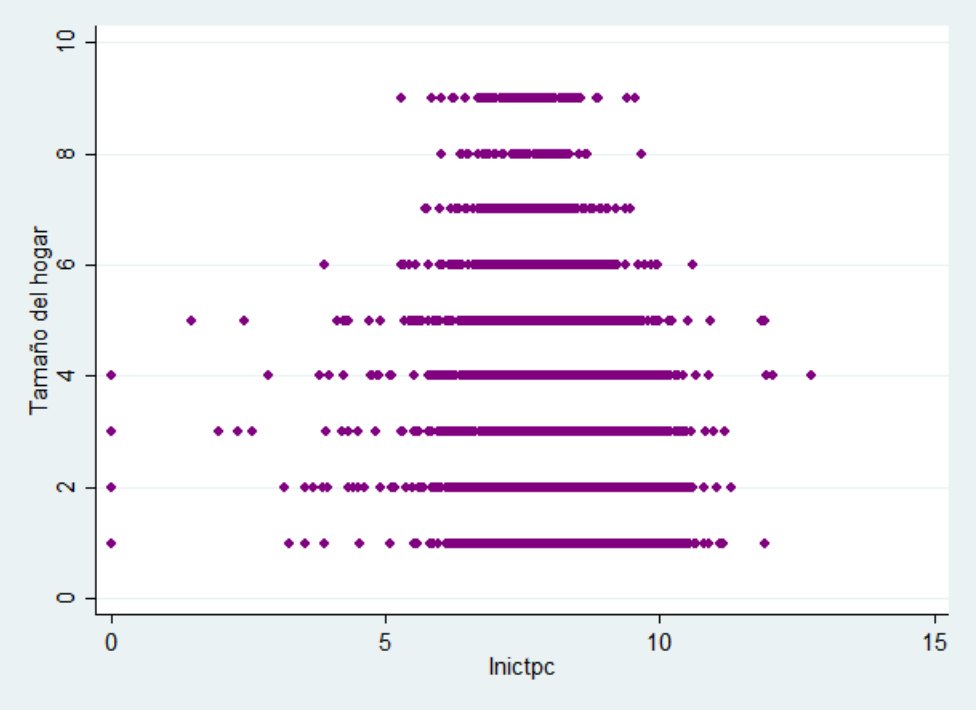

FIGURA 3.2.15. Diagrama de dispersión del ingreso versus el tamaño del hogar.

que los hogares con ingresos por programas de gobierno o que perciben ingresos por remesas presentan un menor ingreso; lo anterior hace pensar que los hogares que perciben menos ingresos son los que cuentan con apoyos, aún así estos no son suficientes para alcanzar el ingreso de un hogar que no tiene apoyo.

Una última variable categórica es la de tenencia de la vivienda, esta variable nos presenta la relación entre los residentes y la propiedad de la vivienda, es decir, si es propia o se está pagando, si es rentada o si se encuentra en otra situación. En la figura 3.2.20 se puede ver que los hogares con vivienda propia y que se está pagando presentan la misma distribución de ingresos que la categoría rentada, mientras que, la categoría otra situación revela ingresos menores.

Otras variables

La variable alusiva al tamaño de localidad resulta interesante, ya que, conforme la localidad cuenta con mayor cantidad de habitantes, el ingreso es mayor, además en la categoría de mayor número de habitantes, se observa gran cantidad de outliers en ambos sentidos, véase figura 3.2.21; de igual manera para la variable que presenta la escala mexicana de seguridad alimentaria, entre mejor sea la seguridad alimentaria mejor es el ingreso, por lo que esta variable está muy relacionada con la variable respuesta, ingreso. Por otro lado, como se puede ver en la figura 3.2.22, en la mayoría de las categorías se muestran datos simetrícos, excepto en la seguridad alimentaria donde poco más del 50 % de los datos se cargan de lado de mayor ingreso, además de contar con la mayor cantidad de datos atípicos.

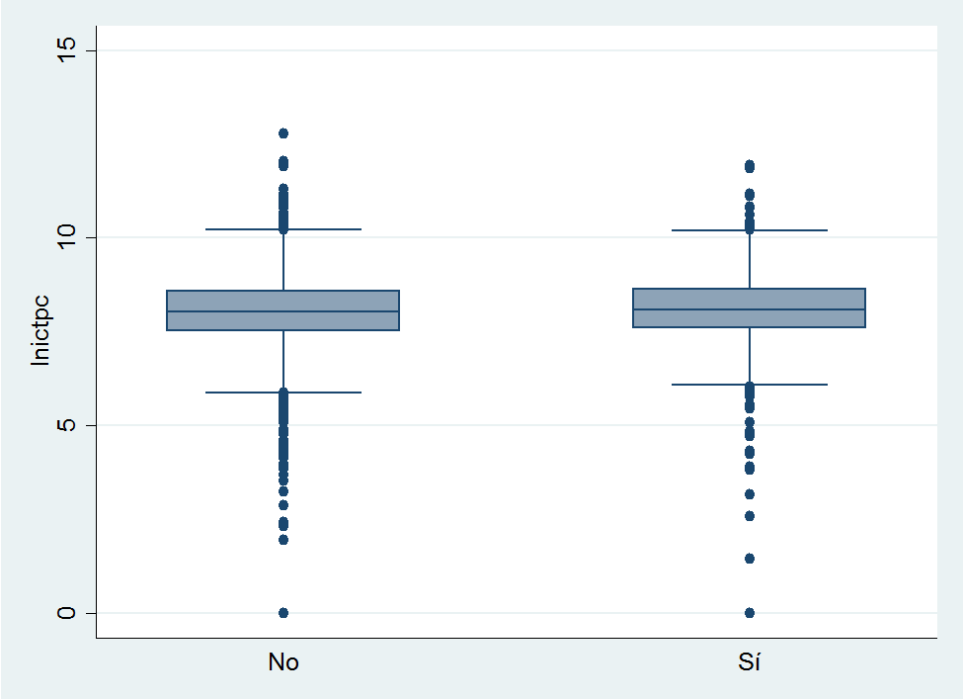

FIGURA 3.2.16. Diagrama de caja del ingreso según la categoría condición de recepción de ingresos por apoyo de otro hogar.

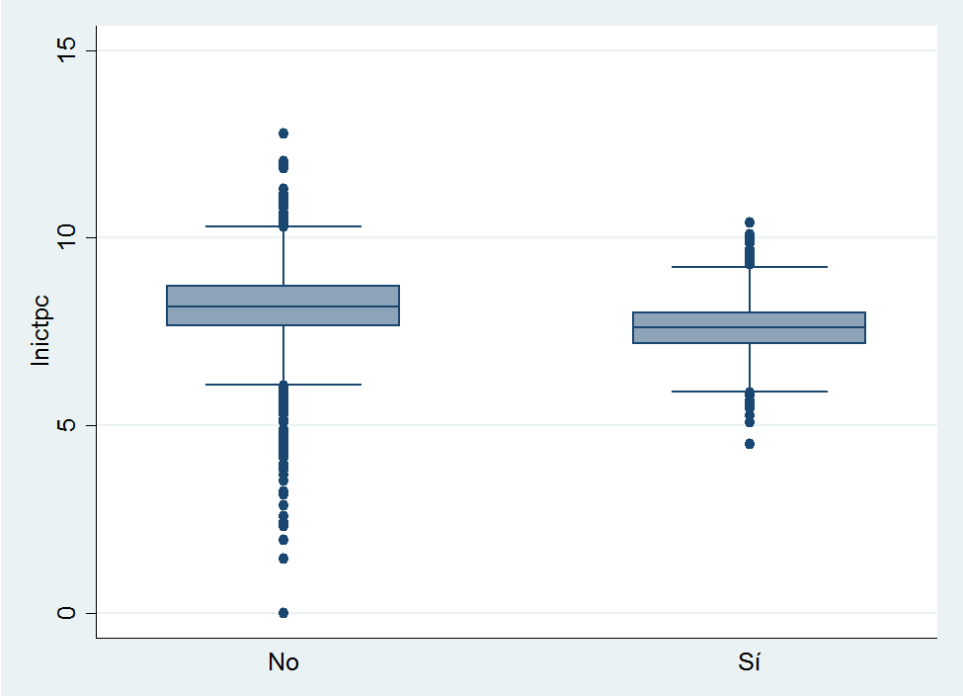

FIGURA 3.2.17. Diagrama de caja del ingreso según la categoría condición de recepción de ingresos por programas de gobierno.

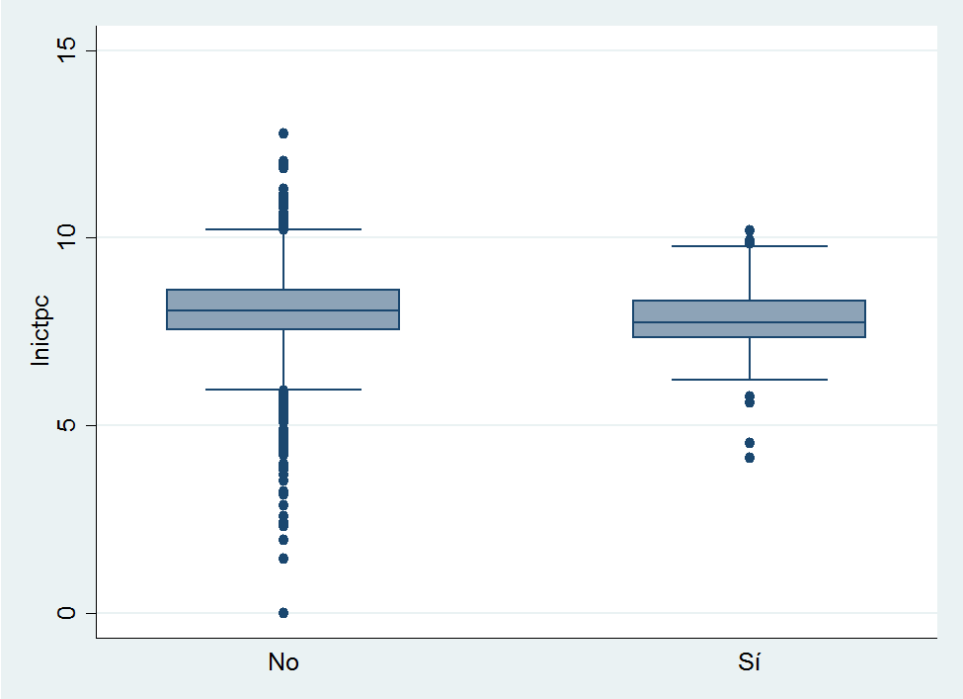

FIGURA 3.2.18. Diagrama de caja del ingreso según la categoría condición de recepción de ingresos por remesas.

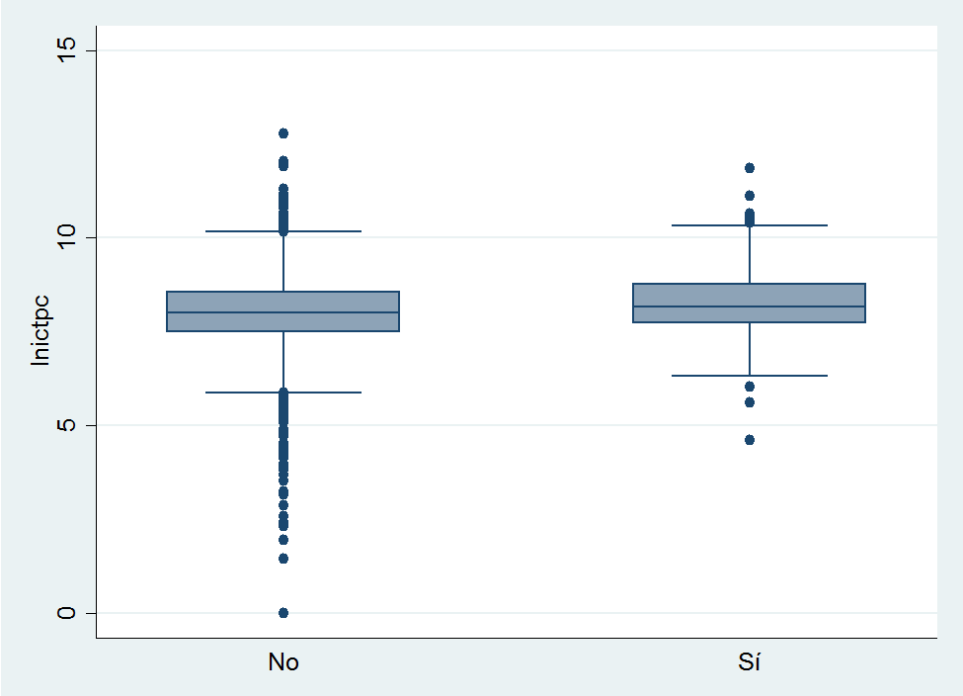

FIGURA 3.2.19. Diagrama de caja del ingreso según la categoría del indicador de personas jubiladas o pensionadas.

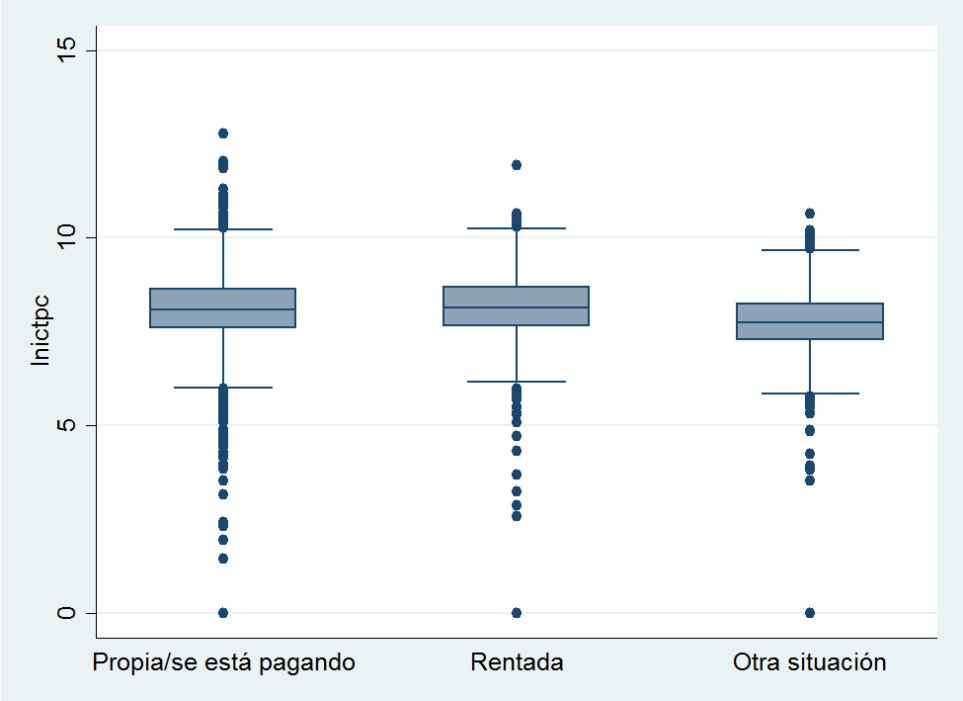

FIGURA 3.2.20. Diagrama de caja del ingreso según la categoría de tenencia de la vivienda.

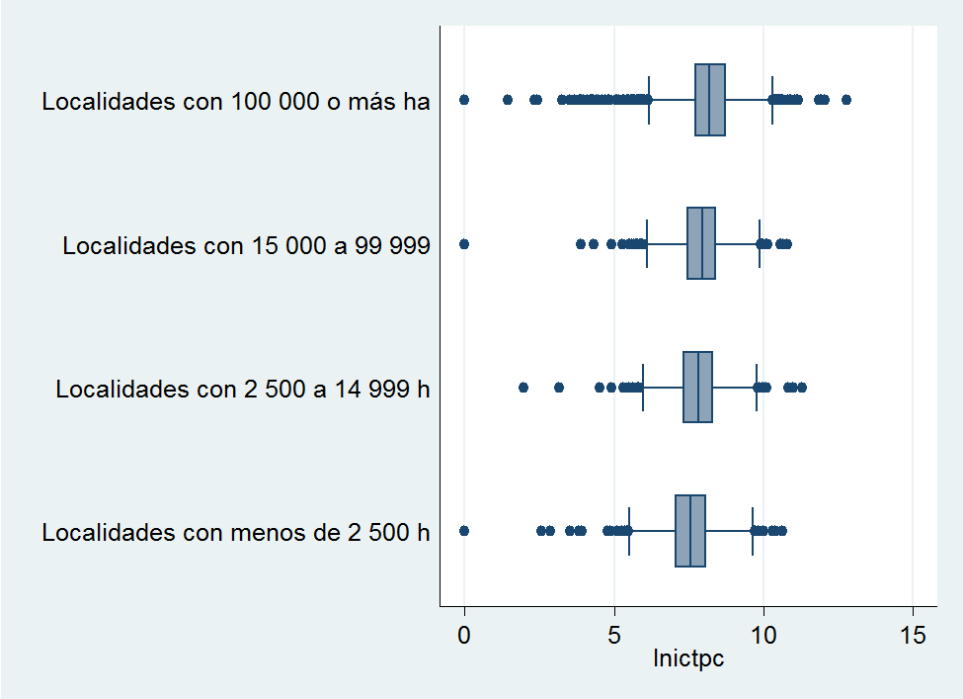

FIGURA 3.2.21. Diagrama de caja del ingreso para las categorías de la variable tamaño de la localidad.

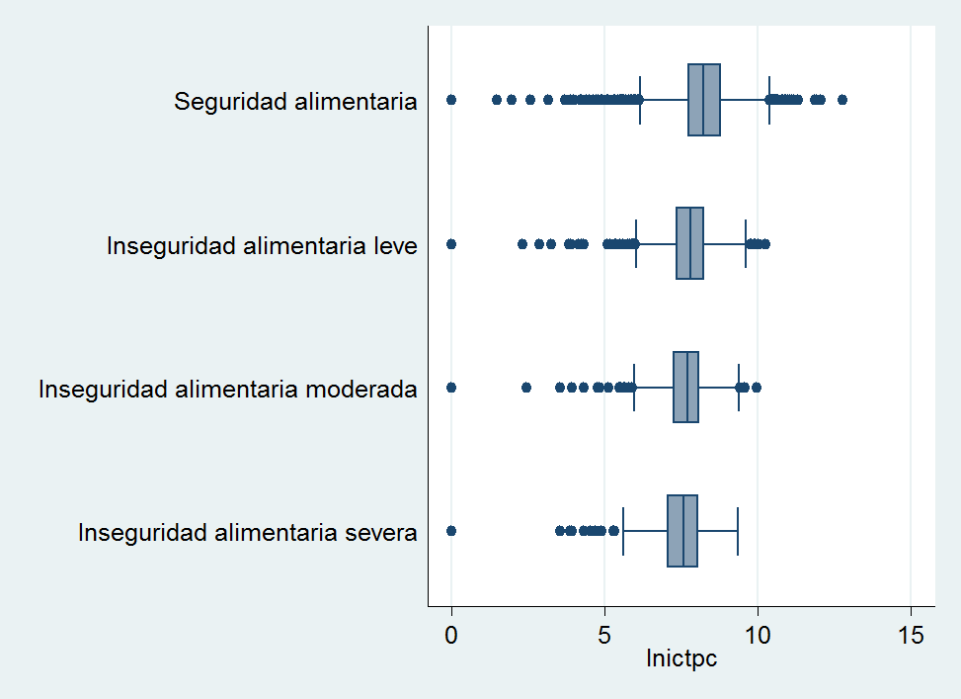

FIGURA 3.2.22. Diagrama de caja del ingreso para las categorías de la variable escala mexicana de seguridad alimentaria.

3.3. Especificación del modelo

Para el análisis de los datos de ingreso se tomará una región compuesta por las entidades federativas de Aguascalientes, Coahuila, Jalisco y Nuevo León, esta región representa un conjunto de datos agrupados de dos niveles donde la entidad representa al grupo (nivel 2) y los hogares que conforman al grupo representan la unidad de análisis (nivel 1), con el ingreso en escala logarítmica como variable respuesta.

El MLM que se propone incluye un solo intercepto aleatorio, es decir, un efecto aleatorio correspondiente al intercepto de cada grupo, en nuestro caso la entidad, lo que implica que las observaciones en la misma entidad están correlacionadas, es decir, en la notación usada en la sección 2.4.1 $q = 1$ y $Z_{ji}^1 = 1$ para toda i y j, con $p = 20$ que denota el número de efectos fijos dados por las variables regresoras que se mencionan en la subsección 3.2.1. De manera que, se tiene el modelo inicial:

$$ingreso_{ji} = \beta_0 + jpea2_i\beta_1 + jpea3_i\beta_2 + cyafore1_i\beta_3 + cyaguin1_i\beta_4 + ... + b_i + \varepsilon_{ji},$$

donde:

$jpea2_i$ la variable actividad del jefe del hogar asociada a la entidad i, y a la categoría de desempleado, con la categoría ocupado como referencia.

$jpea3_i$ la variable actividad del jefe del hogar asociada a la entidad i, y a la categoría de de población no económicamente activa, con la categoría ocupado como referencia.

$cyafore1_i$ la variable dummy afore prestación laboral del cónyuge asociada a la entidad i, usando la categoría, no cuenta como referencia.

$cyaguin1_i$ la variable dummy aguinaldo prestación laboral del cónyuge asociada a la entidad i, usando la categoría, no cuenta como referencia.

$i = 1, \ldots, m$, con $m = 4$, el número de conglomerados ya que se tiene 4 entidades en la región de estudio.

$j = 1, \ldots, n_i$ con n_i el número de individuos en el conglomerado i.

b_i es el efecto aleatorio del modelo.

ε_{ji} es el error.

En este modelo:

β_0 corresponde al valor esperado del ingreso en las categorías de referencia para las variables $jpea$, $cyafore$, $cyaguin$, etc.

β_1, β_2 son los efectos fijos sobre la respuesta que se tiene al cambiar de la categoría desempleada o población no economicamente activa con respecto a la categoría de referencia (ocupada).

β_3 es el efecto fijos sobre la respuesta al cambiar de si cuenta con prestaciones laborales de AFORE el conyugue a no cuenta con prestaciones laborales de AFORE el conyugue.

β_4 es el efecto fijos sobre la respuesta al cambiar de si cuenta con prestaciones laborales de aguinaldo el conyugue a no cuenta con prestaciones laborales de aguinaldo el conyugue.

De igual manera es para las siguientes 17 variables regresoras, las cuales por practicidad de escritura no se reportan en la fórmula.

OBSERVACIÓN:

1. Se supone homoscedasticidad dentro de grupos, es decir:
   - $b_i \sim N(0, D)$ con la matriz $D = \sigma^2_{entidad}$, ya que solo existe un efecto aleatorio.
   - $\varepsilon_{ji} \sim N(0, R_i)$ con $R_i = \sigma^2$, es decir, el vector de errores asociado a cada entidad $\varepsilon_i \sim N(0, I\sigma^2)$ para toda i.

3.4. Estimación del modelo

En esta sección se presenta de manera detallada el ajuste del MLM por medio del software STATA; no se omite señalar que, los tamaños de las entidades varían, desde 1,667 hasta 1,814 hogares.

- Ajuste de modelo:

Con ayuda del comando *xtmixed,* se ajusta un primer modelo, en adelante mencionado como «modelo 1», el cual supone varianza homogénea para todas las entidades, puesto que la versión actual del comando no permite ajustar un modelo con varianza heterogénea. El modelo se especifica de la siguiente forma:

xi : xtmixed lnictpc i.jpea i.cyafore i.cyaguin i.cyuplab i.jafore i.jaguin i.juplab jaesc i.jsexo i.ic_cv i.ic_sbv i.combus2_mod tam_hog i.tamloc i.ins_ali i.ayuotr i.bengob i.remesas i.jubi i.tenencia_viv || ent , covariance(identity) variance:

```
i.jpea            _Ijpea_1-3          (naturally coded; _Ijpea_1 omitted)
i.cyafore         _Icyafore_0-1       (naturally coded; _Icyafore_0 omitted)
i.cyaguin         _Icyaguin_0-1       (naturally coded; _Icyaguin_0 omitted)
i.cyuplab         _Icyuplab_0-1       (naturally coded; _Icyuplab_0 omitted)
i.jafore          _Ijafore_0-1        (naturally coded; _Ijafore_0 omitted)
i.jaguin          _Ijaguin_0-1        (naturally coded; _Ijaguin_0 omitted)
i.juplab          _Ijuplab_0-1        (naturally coded; _Ijuplab_0 omitted)
i.jsexo           _Ijsexo_0-1         (naturally coded; _Ijsexo_0 omitted)
i.ic_cv           _Iic_cv_0-1         (naturally coded; _Iic_cv_0 omitted)
i.ic_sbv          _Iic_sbv_0-1        (naturally coded; _Iic_sbv_0 omitted)
i.combus2_mod     _Icombus2_m_1-2     (naturally coded; _Icombus2_m_1 omitted)
i.tamloc          _Itamloc_1-4        (naturally coded; _Itamloc_1 omitted)
i.ins_ali         _Iins_ali_0-3       (naturally coded; _Iins_ali_0 omitted)
i.ayuotr          _Iayuotr_0-1        (naturally coded; _Iayuotr_0 omitted)
i.bengob          _Ibengob_0-1        (naturally coded; _Ibengob_0 omitted)
i.remesas         _Iremesas_0-1       (naturally coded; _Iremesas_0 omitted)
i.jubi            _Ijubi_0-1          (naturally coded; _Ijubi_0 omitted)
i.tenencia_viv    _Itenencia__1-3     (naturally coded; _Itenencia__1 omitted)
```

Donde *xi* , antes de invocar el comando *xtmixed* crea variables *dummies* para las categorías de los factores fijos, lo cual se puede observar en esta salida.

Posterior al comando *xtmixed*, la primera parte especifica los efectos fijos, la segunda parte especifica los efectos aleatorios y la tercer parte especifica la estructura de covarianza para los efectos aleatorios; en particular, la primer variable corresponde a la variable dependiente, mientras que, las siguientes variables corresponden a los efectos fijos asociados al modelo, las dos barras verticales (||) preceden a la variable que define los grupos de observaciones (entidades), la ausencia de variables adicionales al final de los dos puntos indica que habrá solo un efecto aleatorio asociado con la intercepción de cada entidad en el modelo, finalmente, se encuentra la opción de covarianza que define la estructura de covarianzas para los efectos aleatorios (en este caso se puede omitir al ser un caso simple) y la opción de varianza que solicita se muestren en la salida las desviaciones estimadas de los efectos aleatorios y los errores en lugar de las desviaciones estándar estimadas, que están predeterminadas por STATA. No se omite señalar que, por default el modelo se ajusta utilizando la máxima verosimilitud restringida (REML).

Con relación a la salida generada por STATA para el «modelo 1»: la primera tabla reporta los valores estimados para los efectos fijos β y la segunda tabla proporciona los componentes de varianza estimada, donde *ent:Identity* denota que se utiliza la identidad como patrón de varianzas y covarianzas en el efecto aleatorio.

```
Computing standard errors:

Mixed-effects ML regression                     Number of obs     =      6,924
Group variable: ent                             Number of groups  =          4

                                                Obs per group:
                                                              min =      1,666
                                                              avg =    1,731.0
                                                              max =      1,812

                                                Wald chi2(26)     =    3480.08
Log likelihood = -8198.7763                     Prob > chi2       =     0.0000
```

lnictpc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ijpea_2	-.5950028	.0633379	-9.39	0.000	-.7191428	-.4708628
_Ijpea_3	-.1960142	.0308092	-6.36	0.000	-.256399	-.1356293
_Icyafore_1	.1334847	.0601853	2.22	0.027	.0155236	.2514458
_Icyaguin_1	.3220179	.0534954	6.02	0.000	.2171688	.4268669
_Icyuplab_1	-.0343596	.0537406	-0.64	0.523	-.1396892	.0709699
_Ijafore_1	.0369231	.0375002	0.98	0.325	-.0365759	.1104221
_Ijaguin_1	.1904983	.0367969	5.18	0.000	.1183776	.262619
_Ijuplab_1	.0720364	.0312116	2.31	0.021	.0108629	.13321
jaesc	.0560132	.0023842	23.49	0.000	.0513402	.0606861
_Ijsexo_1	-.0842823	.0243253	-3.46	0.001	-.1319591	-.0366055
_Iic_cv_1	-.1955844	.0502337	-3.89	0.000	-.2940405	-.0971282
_Iic_sbv_1	-.0788108	.0533672	-1.48	0.140	-.1834086	.025787
_Icombus2_m_2	.1365631	.0642616	2.13	0.034	.0106128	.2625134
tam_hog	-.0884952	.0057466	-15.40	0.000	-.0997584	-.077232
_Itamloc_2	-.128551	.0278501	-4.62	0.000	-.1831363	-.0739658
_Itamloc_3	-.1414782	.0363693	-3.89	0.000	-.2127608	-.0701957
_Itamloc_4	-.1830201	.0344478	-5.31	0.000	-.2505365	-.1155036
_Iins_ali_1	-.2379892	.0279953	-8.50	0.000	-.2928591	-.1831194
_Iins_ali_2	-.35514	.0340471	-10.43	0.000	-.421871	-.288409
_Iins_ali_3	-.3479661	.0361922	-9.61	0.000	-.4189016	-.2770306
_Iayuotr_1	.0912828	.0212417	4.30	0.000	.0496498	.1329158
_Ibengob_1	-.0181266	.02774	-0.65	0.513	-.072496	.0362429
_Iremesas_1	.1928945	.0569604	3.39	0.001	.0812543	.3045348
_Ijubi_1	.4232655	.0289875	14.60	0.000	.366451	.4800801
_Itenencia__2	-.0352702	.0275082	-1.28	0.200	-.0891853	.0186449
_Itenencia__3	-.2028954	.0277135	-7.32	0.000	-.2572129	-.1485779
_cons	7.779191	.0882295	88.17	0.000	7.606265	7.952118

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
ent: Identity				
var(_cons)	.0068056	.0050788	.0015763	.0293823
var(Residual)	.6241377	.0106107	.6036838	.6452846

LR test vs. linear model: chibar2(01) = 60.08 Prob >= chibar2 = 0.0000

- Significancia del efecto aleatorio:

Una vez ajustado el «modelo 1», se procede a verificar si el efecto aleatorio asociado con la intersección de cada entidad debe omitirse, para ello, se compara un modelo de regresión lineal ordinaria de un nivel, el cual omite los efectos aleatorios, además, se realiza una prueba de razón de verosimilitud, con esto se prueba si la varianza del efecto aleatorio de la entidad es cero frente a la alternativa de que la varianza sea mayor que cero. En seguida se muestra el resultado generado automáticamente por el comando *xtmixed*.

```
LR test vs. linear model: chibar2(01) = 60.08        Prob >= chibar2 = 0.0000
```

La relación de probabilidad informada por el comando, es una prueba general de los parámetros de covarianza asociados con todos los efectos aleatorios en el modelo. En los modelos con un solo efecto aleatorio para cada grupo, como en el «modelo 1», es apropiado utilizar esta prueba para decidir si ese efecto aleatorio debe incluirse en el modelo. En términos de los resultados de la prueba, se rechaza la hipótesis nula de que la varianza asociada al efecto aleatorio sea cero y dado que la hipótesis nula involucra el borde del espacio parametral se hace el ajuste en el $p-value$ como se indica en la sección 2.9 inciso b.

Siendo así, el estadístico de prueba resulta significativo (p-value $<.05$) por lo que se conserva el efecto aleatorio. En otras palabras:

$p-value = 0.5 \times P(\chi^2_0 > 60.08) + .05 \times P(\chi^2_1 > 60.08) = 0 + .05 \times P(\chi^2_1 > 60.08) < .001.$

La distribución de χ^2_0 está concentrada en cero, por ello su contribución al $p-value$ es cero y se omite del cálculo.

- Significancia de los efectos fijos:

La idea de este paso, es reducir el modelo eliminando efectos fijos no significativos. Las pruebas relacionadas a los efectos fijos no son generadas por el comando *xtmixed*, no obstante, se puede realizar la prueba del estadístico de Wald; por ejemplo, para probar la importancia general de los efectos de la variable jpea se puede utilizar el siguiente comando, usando las variables ficticias creadas por STATA:

$$test \_Ijpea\_2 \_Ijpea\_3$$

el cual tiene como hipótesis nula que los dos efectos fijos son iguales a cero, es decir, las medias de jpea son todas iguales. Siendo así, la prueba está dada por:

$$H_0 : P\beta = 0 \ \ vs \ \ H_1 : P\beta \neq 0,$$

con P una matriz de dimensión 20×2 y β de 2×1.

La siguiente información generada por STATA muestra que el estadístico de prueba es significativo (p-value <0.05) por lo que se rechaza la hipótesis nula de que los parámetros asociados son simultáneamente cero y se conservan los efectos.

```
 ( 1)  [lnictpc]_Ijpea_2 = 0
 ( 2)  [lnictpc]_Ijpea_3 = 0

           chi2(  2) =  116.51
         Prob > chi2 =    0.0000
```

Esto se sigue para cada una de las variables regresoras que se ingresaron al modelo, vistas en la subsección 3.2.1:

```
( 1)  [lnictpc]_Iins_ali_3 = 0
( 2)  [lnictpc]_Iins_ali_2 = 0
( 3)  [lnictpc]_Iins_ali_1 = 0

          chi2(  3) =  205.30
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Iic_sbv_1 = 0

          chi2(  1) =    2.18
        Prob > chi2 =    0.1397

( 1)  [lnictpc]_Iic_cv_1 = 0

          chi2(  1) =   15.16
        Prob > chi2 =    0.0001

 ( 1)  [lnictpc]_Icyuplab_1 = 0

          chi2(  1) =    0.41
        Prob > chi2 =    0.5226
```

```
( 1)  [lnictpc]_Icyaguin_1 = 0

          chi2(  1) =   36.23
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Icyafore_1 = 0

          chi2(  1) =    4.92
        Prob > chi2 =    0.0266

( 1)  [lnictpc]_Icombus2_m_2 = 0

          chi2(  1) =    4.52
        Prob > chi2 =    0.0336

 ( 1)  [lnictpc]_Ibengob_1 = 0

          chi2(  1) =    0.43
        Prob > chi2 =    0.513
```

```
( 1)  [lnictpc]_Itenencia__3 = 0
( 2)  [lnictpc]_Itenencia__2 = 0

          chi2(  2) =   53.64
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Itamloc_4 = 0
( 2)  [lnictpc]_Itamloc_3 = 0
( 3)  [lnictpc]_Itamloc_2 = 0

          chi2(  3) =   47.47
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Iremesas_1 = 0

          chi2(  1) =   11.47
        Prob > chi2 =    0.0007

( 1)  [lnictpc]_Ijuplab_1 = 0

          chi2(  1) =    5.33
        Prob > chi2 =    0.0210
```

```
( 1)  [lnictpc]_Ijubi_1 = 0

          chi2(  1) =  213.21
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Ijsexo_1 = 0

          chi2(  1) =   12.00
        Prob > chi2 =    0.0005

( 1)  [lnictpc]_Ijaguin_1 = 0

          chi2(  1) =   26.80
        Prob > chi2 =    0.0000

( 1)  [lnictpc]_Ijafore_1 = 0

          chi2(  1) =    0.97
        Prob > chi2 =    0.3248
```

```
( 1)  [lnictpc]_Iayuotr_1 = 0

          chi2(  1) =   18.47
        Prob > chi2 =    0.0000


( 1)  [lnictpc]jaesc = 0

          chi2(  1) =  551.95
        Prob > chi2 =    0.0000

( 1)  [lnictpc]tam_hog = 0

          chi2(  1) =  237.14
        Prob > chi2 =    0.0000
```

Utilizando estas pruebas, se eliminan variables que no son significativas y se llega a la construcción del «modelo 2» , donde se excluye las variables: jafore, ic_sbv, cyuplab y bengob. Cabe señalar que, en este nuevo modelo el efecto aleatorio sigue siendo significativo; las estimaciones asociadas al modelo corresponden a:

```
Computing standard errors:

Mixed-effects ML regression                     Number of obs     =      6,924
Group variable: ent                             Number of groups  =          4

                                                Obs per group:
                                                              min =      1,666
                                                              avg =    1,731.0
                                                              max =      1,812

                                                Wald chi2(22)     =    3473.87
Log likelihood = -8200.8539                     Prob > chi2       =     0.0000
```

lnictpc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ijpea_2	-.5938763	.0633209	-9.38	0.000	-.717983	-.4697695
_Ijpea_3	-.1974279	.0305636	-6.46	0.000	-.2573314	-.1375244
_Icyafore_1	.1241991	.0573131	2.17	0.030	.0118675	.2365307
_Icyaguin_1	.3175317	.05263	6.03	0.000	.2143789	.4206845
_Ijaguin_1	.2157687	.0279395	7.72	0.000	.1610083	.2705291
_Ijuplab_1	.0802986	.0298017	2.69	0.007	.0218882	.1387089
jaesc	.0567839	.0023203	24.47	0.000	.0522361	.0613316
_Ijsexo_1	-.0837254	.024323	-3.44	0.001	-.1313977	-.0360531
_Iic_cv_1	-.2022112	.0500425	-4.04	0.000	-.3002927	-.1041296
_Icombus2_m_2	.1844847	.0566181	3.26	0.001	.0735153	.2954541
tam_hog	-.0883467	.0057227	-15.44	0.000	-.0995629	-.0771305
_Itamloc_2	-.129805	.02784	-4.66	0.000	-.1843704	-.0752395
_Itamloc_3	-.1465819	.0361933	-4.05	0.000	-.2175194	-.0756444
_Itamloc_4	-.1995853	.0331534	-6.02	0.000	-.2645648	-.1346058
_Iins_ali_1	-.2383387	.0279915	-8.51	0.000	-.2932011	-.1834763
_Iins_ali_2	-.3554801	.0340401	-10.44	0.000	-.4221975	-.2887626
_Iins_ali_3	-.3507855	.0361664	-9.70	0.000	-.4216703	-.2799007
_Iayuotr_1	.0899997	.0212333	4.24	0.000	.0483831	.1316162
_Iremesas_1	.194051	.0569745	3.41	0.001	.082383	.305719
_Ijubi_1	.4243289	.0289536	14.66	0.000	.3675809	.4810768
_Itenencia__2	-.0354946	.0273923	-1.30	0.195	-.0891824	.0181933
_Itenencia__3	-.2046937	.0277001	-7.39	0.000	-.2589849	-.1504024
_cons	7.721563	.0807124	95.67	0.000	7.56337	7.879757

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
ent: Identity				
var(_cons)	.0067337	.0050275	.0015586	.0290922
var(Residual)	.6245162	.0106171	.60405	.645676

LR test vs. linear model: chibar2(01) = 59.47 Prob >= chibar2 = 0.0000

- Criterios de información:

Al comparar dos modelos, «modelo 1» frente al «modelo 2» los criterios de información AIC y BIC se pueden obtener en Stata usando el siguiente comando:

estat ic

lo que nos generara la siguiente salida para el «modelo 1»:

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,924	.	-8198.776	29	16455.55	16653.99

Note: N=Obs used in calculating BIC; see [R] BIC note.

Y la siguiente salida para el «modelo 2»:

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,924	.	-8200.854	25	16451.71	16622.78

Note: N=Obs used in calculating BIC; see [R] BIC note.

Aquí se debe recordar que un AIC y/o BIC pequeño refleja un mejor modelo, dicho de otro modo, ajusta mejor a los datos. En este caso, se elegió el «modelo 2» ya que ajusta mejor a los datos.

3.5. Verificación de supuestos

Los residuos son discrepancias entre el valor estimado y el valor observado, o visto en otra forma la variabilidad que no puede explicarse mediante el modelo. Es por eso que se observan los residuos para saber si se cumple el supuesto de normalidad del modelo. Los residuos estandarizados son más apropiados para examinar este supuesto y detectar valores atípicos, para ello, se obtiene las predicciones y los residuales estandarizados, mediante el comando *predict* de la siguiente forma:

- Para obtener el valor estimado del ingreso:
  predict double lctpc_ estimado, xb

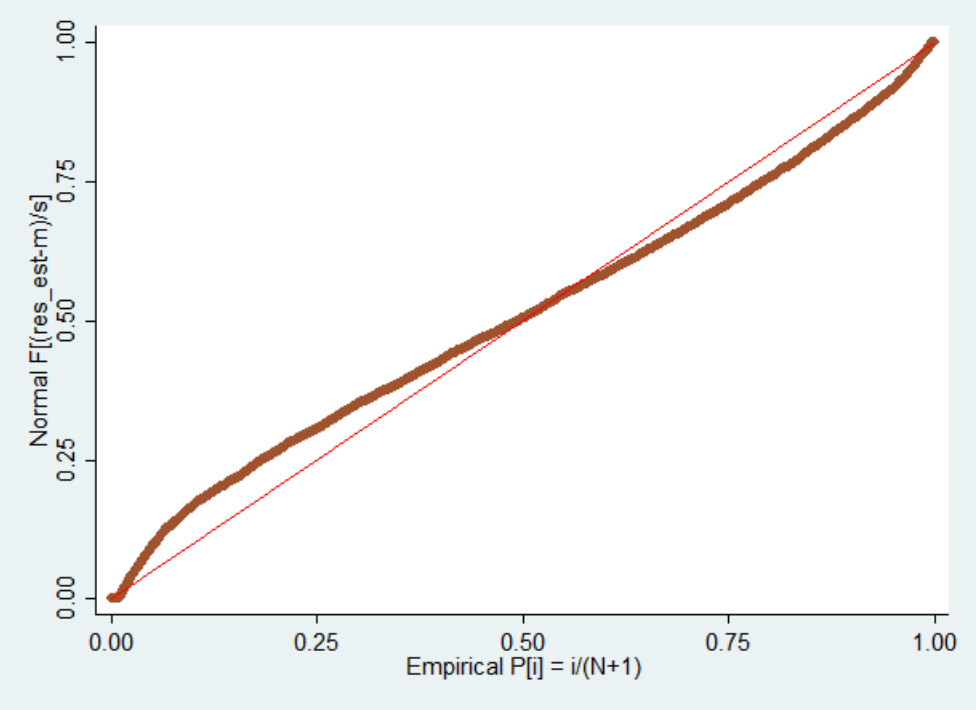

FIGURA 3.5.1. Gráfica de probabilidad normal de los residuos estandarizados.

- Para obtener los residuales estandarizados:
  predict double res_ est, rstandard

Ahora, para verificar el supuesto de normalidad se obtienen gráficos de probabilidad normal y gráficos de dispersión de los residuos estandarizados.

Las gráficas en las figuras 3.5.1 y 3.5.2 muestran que hay algunos valores atípicos en el extremo inferior de la distribución normal de los residuos y distribuciones potencialmente sesgadas negativas. Sin embargo, se puede ver especialmente en la figura 3.5.2 que la asimetría no es severa y no se observa algún patrón definido por lo que no se descarta la relación lineal con los efectos fijos. Cabe señalar que, esto ocurre cuando no se eliminan las observaciones que numéricamente están distantes del resto de los datos de ingreso (valores atípicos), por esta razón en el Apéndice D se muestra el ejercicio omitiendo los valores atípicos en la variable ingreso.

En la figura 3.5.3 se observa una distribución de los residuos homogénea en las entidades, lo cual brinda confianza a la estimación del «modelo 2».

3.6. Resultados

Con la finalidad de verificar la consistencia de la estimación del ingreso que se genera a partir del MLM, se generó un gráfico de dispersión y diagramas de caja.

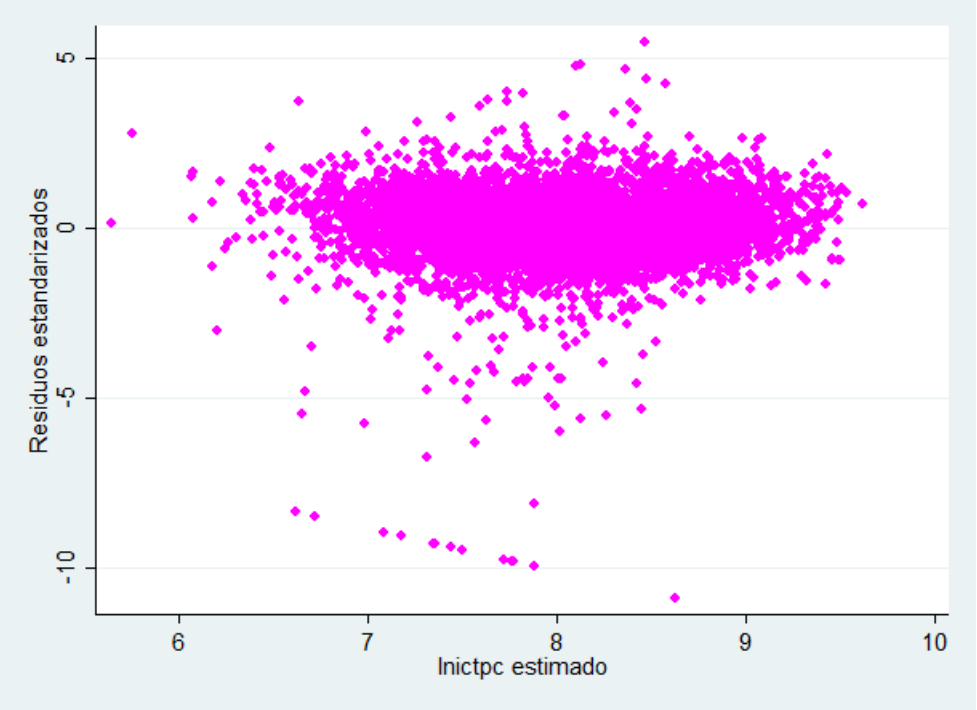

FIGURA 3.5.2. Gráfica de dispersión, residuos estandarizados vs lnictpc estimado.

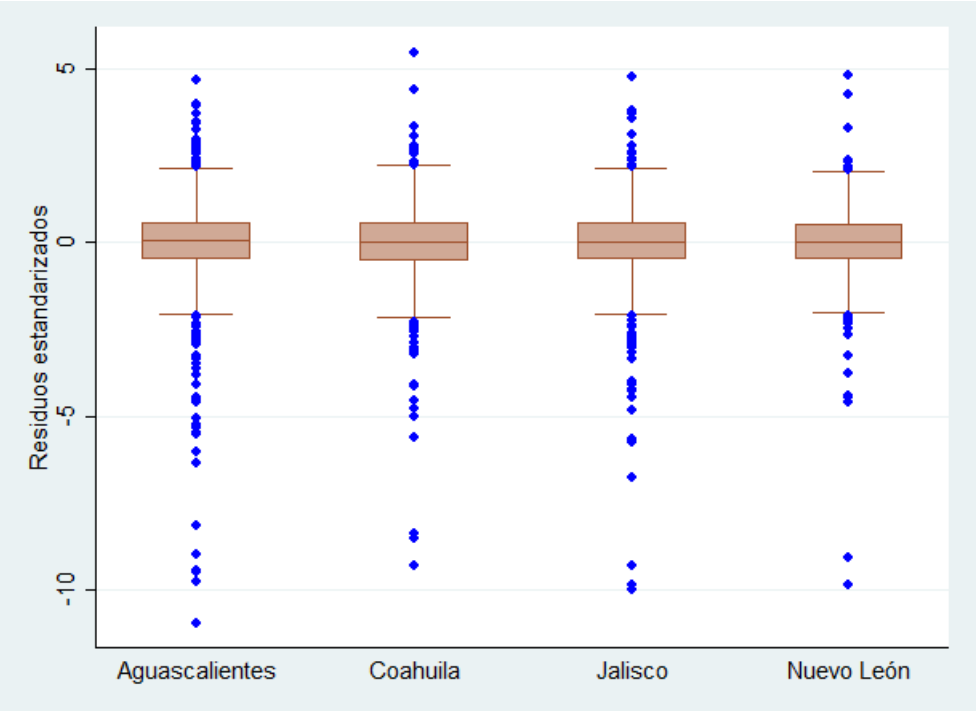

FIGURA 3.5.3. Diagrama de caja de los residuos estandarizados según entidad.

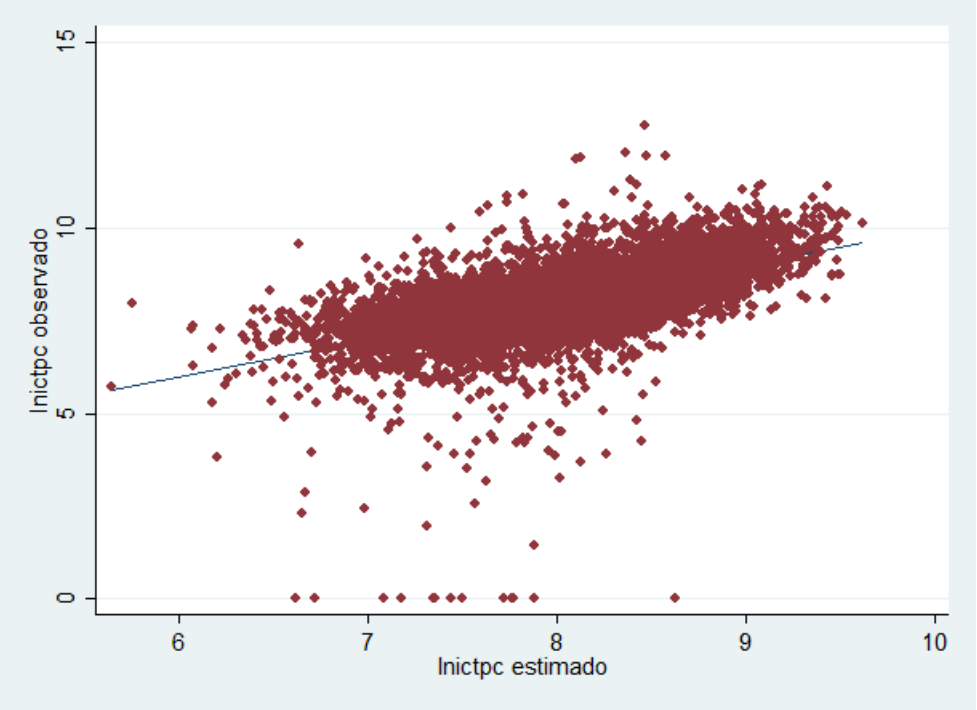

FIGURA 3.6.1. Diagrama de dispersión. lnictpc vs lnictpc estimado.

De la figura 3.6.1 la cual corresponde a un diagrama de dispersión entre los valores reales y los valores estimados, ambos en escala logarítmica, se confirma que el modelo subestima la variable observada «lnictpc», esto se observa al guiarnos con la recta de 45°, ya que la estimación perfecta debiera ir alrededor de ese valor. No obstante, existe una relación positiva entre el valor observado y el estimado, lo que nos indica que el modelo ajustó de manera correcta a los datos.

Respecto a los diagramas de caja presentados en la figura 3.6.2 y 3.6.3 se observa que el modelo no captura el rango de valores que toma el ingreso «lnictpc». NOTA: El código completo se encuentra en el Apéndice E.

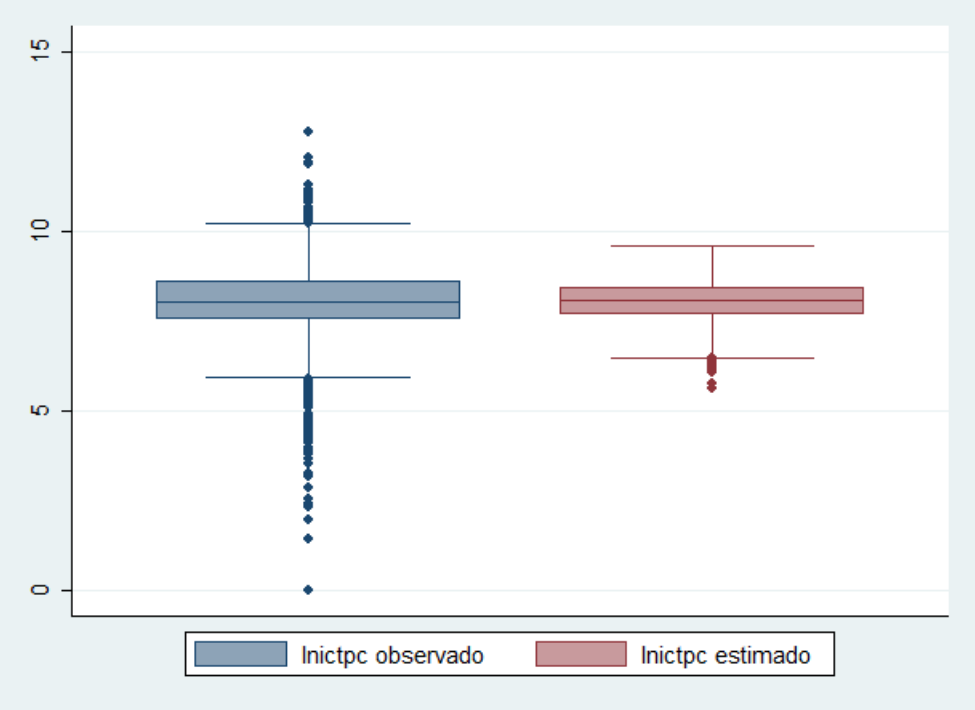

FIGURA 3.6.2. Diagrama de caja. lnictpc e lnictpc estimado.

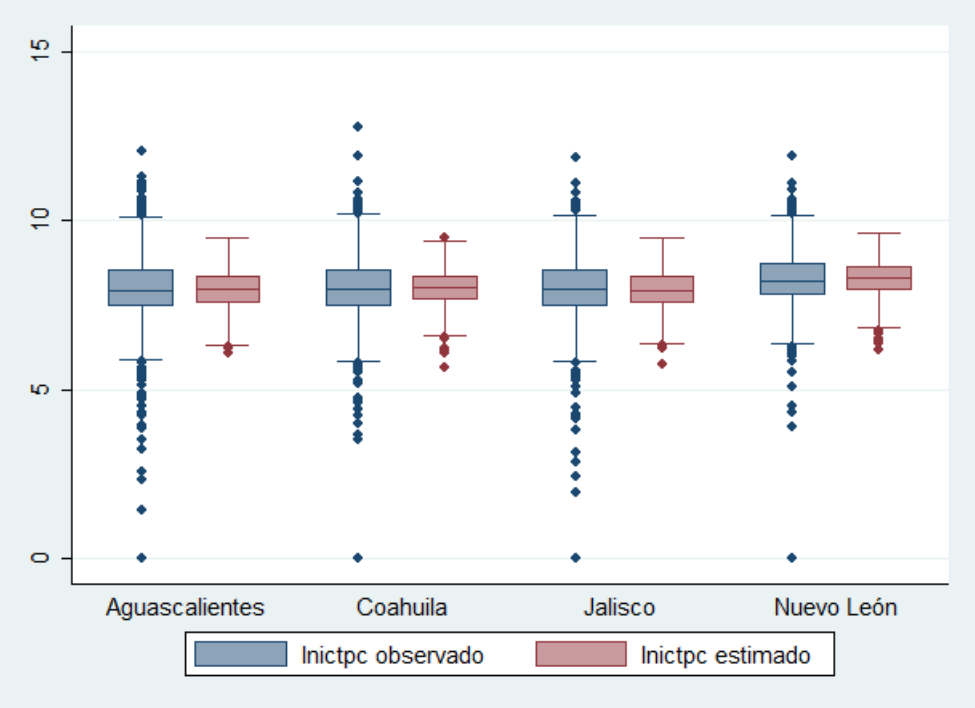

FIGURA 3.6.3. Diagrama de caja. lnictpc e lnictpc estimado por entidad.

Capítulo 4

Conclusiones

En este trabajo se ha estimado el ingreso para cada hogar de las entidades federativas con mejores condiciones socioeconómicas de la república mexicana (Aguascalientes, Coahuila, Jalisco y Nuevo León). Para ello se utilizó la técnica de modelos lineales mixtos (MLM), en concreto se realizó un modelo de dos niveles donde el primer nivel corresponde a los hogares y el segundo nivel a las entidades, ya que, el efecto aleatorio es significativo y se sospecha de falta de independencia entre las observaciones dentro de la misma entidad.

Se ajustó un primer modelo «modelo 1», que contempla el total de variables predictoras; posteriormente a través de una prueba de razón de verosimilitud se determinó que el efecto aleatorio asociado con la intersección de cada entidad se debía incluir en el modelo, ya que, resultó significativo. A posteriori, para seleccionar aquellas variables que se introducen en el modelo final «modelo 2», se utilizó el estadístico de Wald, excluyendo las variables: jafore, ic_sbv, cyuplab y bengob.

Para la selección del modelo que ajusta mejor a los datos se usaron los criterios de información AIC y BIC, resultando como mejor modelo, el «modelo 2», que excluye las variables mencionadas en el párrafo anterior y donde de igual forma resultan significativos los efectos aleatorios. Cabe señalar que, el «modelo 2» no resultó por mucho, el mejor modelo, por lo que ambos modelos son alternativas para la estimación del ingreso. Las variaciones entre las estimaciones de uno u otro modelo serán mínimas.

En relación con los resultados, se puede observar que el modelo subestima el ingreso, por esta razón, se sugiere incorporar más variables con la finalidad de predecir de manera más cercana la realidad, así como análizar la construcción de nuevas variables. No obstante, existe una relación positiva entre el ingreso observado y el estimado. Siendo así, se puede concluir que el modelo lineal mixto de dos niveles se aplicó de manera correcta.

Para efectos prácticos hay que recordar lo importante que resulta trabajar con variables de resultado con distribución normal, es por esa razón que se trabajó con el ingreso a escala logarítmica.

Para efectos metodológicos se puede concluir que los MLM resultan ser una alternativa al ajuste por regresión mendiante mínimos cuadrados ordinarios cuando las observaciones no son independientes entre sí.

Adicionalmente, estos modelos son útiles en las aplicaciones porque:

- Permiten la estimación eficiente de los parámetros fijos aplicando técnicas de Mínimos Cuadrados Generalizados (MCG), esto significa que pueden incluir heteroscedasticidad (cuando la varianza de los errores no es constante en todas las observaciones realizadas).
- Realizan predicciones eficientes de los párametros aleatorios al usar la teoría de los mejores predictores lineales insesgados.
- Permite la estimación de los componentes de la varianza que definen las matrices de varianza estimadas.

Referencias bibliográficas

[1] Brady T. West, Katheleen B. Welch and Andrzej T. Galecki (2007). LINEAL MIXED MODELS a Practical Guide Using Statistical Software. Chapman & Hall/CRC.

[2] Douglas M. Bates and Donal G. Watts (1988). Nonlinear Regressión Analysis and Its Applications. WILEY.

[3] Eugene Demidenko (2004). Mixed Models Theory and Applications. WILEY-INTERSCIENCE.

[4] Julian J. Faraway (2006). Extending the Linear Model with R. Chapman & Hall/CRC Taylor & Francis Group.

[5] Andrzej Galecki and Tomasz Burzykowski, (2013). Linear Mixed-Effects Models Using R. Springer.

[6] Lang Wu (2010). Mixed Effects Models for Complex Data. CRC Press Taylor & Francis Group A CHAPMAN & HALL BOOK.

[7] Douglas M. Bates (2010). Mixed-effects modeling with R. Springer.

[8] Juan Carlos Correa Morales y Juan Carlos Salazar Uribe, (2016). Introducción a los modelos mixtos. Universidad Nacional de Colombia.

[9] Roberto G. Gutierrez. Linear mixed models in STATA. StataCorp LP.

[10] John Fox (2002). Linear mixed models. The web appendix contains information about using S (R and S-PLUS).

[11] Douglas M. Bates (2011). Computational methods for mixed models. University of Wisconsin.

[12] Alain F. Zuur, Elena N. Ieno (2009). Mixed effects models and extensions in ecology with R. Springer.

[13] Montgomery, Douglas Peck, Elizabet (2006). Introducción al análisis de regresión lineal / 3ed. Compañía editorial continental.

[14] Sanford Weisberg (2005). Applied linear regression/ 3ed. Wiley.

[15] Michael H. Kutner, Christopher J. Nachtshei, y John Neter (2005). Applied linear statistical models/ 3ed. McGraw-Hill.

Apéndice A

Teorema de Gauss-Markov

Este teorema establece que el estimador $\hat{\beta} = (X'X)^{-1}X'y$, es óptimo entre la familia de estimadores lineales e insesgados. Es decir, no es posible encontrar otro estimador de β que siendo lineal e insesgado tenga una varianza menor que el. Para demostrar este teorema, se considera a $\widetilde{\beta} = [(X'X)^{-1}X' + B]y + b_0$ como otro estimador insesgado de β, resultado de un combinación lineal de los datos. Donde, B es una matriz de $p \times n$ y b_0 un vector de constantes $p \times 1$, que ajusta en forma adecuada para formar el estimado alternativo.

Primero se ve la esperanza del estimador:

$\mathbb{E}(\widetilde{\beta}) = \mathbb{E}([(X'X)^{-1}X' + B]y + b_0)$
$= [(X'X)^{-1}X' + B]\mathbb{E}(y) + b_0$
$= [(X'X)^{-1}X' + B]X\beta + b_0$
$= (X'X)^{-1}X'X\beta + BX\beta + b_0$
$= \beta + BX\beta + b_0,$
de lo cual se deriva que $\widetilde{\beta}$ es insesgado siempre y cuando $b_0 = 0$ y $BX = 0$.

Ahora se ve la varianza del estimador:

$Var(\widetilde{\beta}) = Var([(X'X)^{-1}X' + B]y)$
$= [(X'X)^{-1}X' + B]Var(y)[(X'X)^{-1}X' + B]'$
$= [(X'X)^{-1}X' + B]\sigma^2 I[(X'X)^{-1}X' + B]'$
$= \sigma^2[(X'X)^{-1}X' + B][X(X'X)^{-1}X' + B']$
$= \sigma^2[(X'X)^{-1} + BB'],$
donde se observa que $(BX)' = X'B' = 0$, puesto que, $BX = 0$.

Del cual sigue:

$$
\begin{aligned}
&Var(l'\widetilde{\beta}) = l'Var(\widetilde{\beta})l \\
&= l'(\sigma^2[(X'X)^{-1} + BB'])l \\
&= \sigma^2 l'(X'X)^{-1}l + \sigma^2 l'BB'l \\
&= Var(l'\widetilde{\beta}) + \sigma^2 l'BB'l,
\end{aligned}
$$

siendo que BB' es una matriz positiva definida $\sigma^2 l'BB'l \geq 0$ y definiendo $l^* = B'l$**:**

$$l'BB'l = l^{*'}l^* = \sum_{i=1}^{p} l_i^{*2} > 0 \ \textit{para cierta} \ l \neq 0.$$

Por lo tanto:

$$Var(l'\hat{\beta}) \geq \sigma^2 l'(XX')^{-1}l,$$

con al menos una l , tal que:

$$Var(l'\hat{\beta}) > \sigma^2 l'(XX')^{-1}l.$$

Por lo anterior, el estimador $\hat{\beta}$, es el mejor estimador lineal insesgado.

Apéndice B

Mínimos Cuadrados Generalizados (MCG)

Es una técnica aplicada con gran frecuencia para la estimación de los párametros desconocidos en un modelo de regresión lineal. Es aplicada cuando se tiene heterocedasticidad (varianzas desiguales) o cuando existe un cierto grado de correlación entre las observaciones. En otras palabras, este método consiste en transformar un modelo con heterocedasticidad en otro con varianza constante, de forma que al aplicarle a este último el método de mínimos cuadrados ordinarios (MCO) se obtenga un estimador lineal, insesgado y óptimo.

Partiendo del modelo de regresión lineal en forma matricial:

$$Y = X\beta + \varepsilon,$$

en donde,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; X = \begin{bmatrix} x_{11} & x_{21} & \dots & x_{k1} \\ x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},$$

con $\mathbb{E}[\varepsilon] = 0$ y $Var[\varepsilon] = \mathbb{E}[\varepsilon\varepsilon'] = \sigma^2 V$.

Dado que V es una matriz simétrica definida positiva puede ser factorizada como $V = KK'$, con la siguiente propiedad $K' = K^{-1}$.

Si se estima el parámetro β por mínimos cuadrados ordinarios (MCO) el estimador dado por $\hat{\beta} = (X'X)^{-1}X'y$ no será óptimo. Es por esta razón, que se considera la siguiente transformación, premultiplicando cada uno de los componentes del modelo por K^{-1}:

$$Y^* = K^{-1}Y,$$

$$X^* = K^{-1}X,$$

$$\varepsilon^* = K^{-1}\varepsilon.$$

Entonces el modelo tranformado es:

$$Y^* = X^*\beta + \varepsilon^*,$$

el cual cumple con los supuestos basicos del Teorema de Gauss-Markov, es decir:

- El modelo es lineal, dado que Y^*es funcion lineal de β.
- X^* es una matriz de rango completo, $rango(X^*) = k$.
- $\mathbb{E}[\varepsilon^*] = 0$, ya que $\mathbb{E}[\varepsilon^*] = \mathbb{E}[K^{-1}\varepsilon] = K^{-1}\mathbb{E}[\varepsilon] = 0$.
- $Var[\varepsilon^*] = \mathbb{E}[\varepsilon^*\varepsilon^{*'}]$
  $= \mathbb{E}[K^{-1}\varepsilon(K^{-1}\varepsilon)']$
  $= \mathbb{E}[K^{-1}\varepsilon\varepsilon'(K^{-1})']$
  $= K^{-1}\mathbb{E}[\varepsilon\varepsilon'](K')^{-1}$
  $= K^{-1}\sigma^2 V(K')^{-1}$
  $= \sigma^2 K^{-1}V(K')^{-1}$
  $= \sigma^2 K^{-1}KK'(K')^{-1}$
  $= \sigma^2 II$
  $= \sigma^2 I.$

Siendo así, el estimador de mínimos cuadrados generalizados, es decir, el estimador de mínimos cuadrados ordinarios de β en el modelo transformado está dado por:

$$\hat{\beta}_{MCG} = (X^{*'}X^*)X^{*'}y^*$$
$$= ((K^{-1}X)'(K^{-1}X))^{-1}(K^{-1}X)'(K^{-1}y)$$
$$= (X'(K^{-1})'(K^{-1})X))^{-1}(X'(K^{-1})'(K^{-1})y)$$
$$= (X'V^{-1}X)^{-1}(X'V^{-1}y).$$

Con las siguientes propiedades:

1. El estimador es insesgado
   $\mathbb{E}[\hat{\beta}_{MCG}] = \mathbb{E}[(X'V^{-1}X)^{-1}(X'V^{-1}y)]$
   $= \mathbb{E}[(X'V^{-1}X)^{-1}(X'V^{-1}(X\beta + \varepsilon))]$
   $= \mathbb{E}[(X'V^{-1}X)^{-1}(X'V^{-1}X)\beta + (X'V^{-1}X)^{-1}(X'V^{-1}\varepsilon)]$
   $= \mathbb{E}[I\beta + (X'V^{-1}X)^{-1}(X'V^{-1}\varepsilon)]$
   $= \mathbb{E}[\beta + (X'V^{-1}X)^{-1}(X'V^{-1}\varepsilon)]$
   $= \mathbb{E}[\beta] + (X'V^{-1}X)^{-1}(X'V^{-1})\mathbb{E}[\varepsilon]$
   $= \beta + 0$
   $= \beta.$

- Con matriz de varianza y covarianza

$$
\begin{aligned}
&Var[\hat{\beta}_{MCG}] = \mathbb{E}[(\hat{\beta}_{MCG} - \mathbb{E}(\hat{\beta}_{MCG}))(\hat{\beta}_{MCG} - \mathbb{E}(\hat{\beta}_{MCG}))'] \\
&= \mathbb{E}[(\hat{\beta}_{MCG} - \beta)(\hat{\beta}_{MCG} - \beta)'] \\
&= \mathbb{E}[((X'V^{-1}X)^{-1}X'V^{-1}\varepsilon)((X'V^{-1}X)^{-1}X'V^{-1}\varepsilon)'] \\
&= \mathbb{E}[(X'V^{-1}X)^{-1}X'V^{-1}\varepsilon\varepsilon'V^{-1}X(X'V^{-1}X)^{-1}] \\
&= (X'V^{-1}X)^{-1}X'V^{-1}\mathbb{E}[\varepsilon\varepsilon']V^{-1}X(X'V^{-1}X)^{-1} \\
&= (X'V^{-1}X)^{-1}X'V^{-1}\sigma^2 VV^{-1}X(X'V^{-1}X)^{-1} \\
&= \sigma^2(X'V^{-1}X)^{-1}X'V^{-1}VV^{-1}X(X'V^{-1}X)^{-1} \\
&= \sigma^2(X'V^{-1}X)^{-1}X'IX(X'V^{-1}X)^{-1} \\
&= \sigma^2(X'V^{-1}X)^{-1}I \\
&= \sigma^2(X'V^{-1}X)^{-1}.
\end{aligned}
$$

Apéndice C

Listado de variables en el modelo

LABORALES

Variable	Categoría
jpea	1= Ocupada, 2= Desempleada y 3=Población económicamente activa
cyafore	0= No y 1= Sí
cyaguin	0= No y 1= Sí
cyuplab	0= No y 1= Sí
jafore	0= No y 1= Sí
jaguin	0= No y 1= Sí
juplab	0= No y 1= Sí

SOCIODEMOGRÁFICAS

Variable	Categoría
jaesc	[0,23]
jsexo	0= Mujer y 1= Hombre

CARACTERÍSTICAS DEL HOGAR

Variable	Categoría
ic_cv	0= Si carencia y 1= Con carencia
ic_sbv	0= Si carencia y 1= Con carencia
combus2_mod	1= Leña o carbón y 2= Gas, electricidad u otro combustible
tam_hog	[1,9]

OTROS INGRESOS

Variable	Categoría
ayuotr	0= No y 1= Sí
bengob	0= No y 1= Sí
remesas	0= No y 1= Sí
jubi	0= No y 1= Sí
tenencia_viv	0= No y 1= Sí

OTRAS

Variable	Categoría
tamloc	1= Localidades con 100,000 o más habitantes, 2= Localidades con 15,000 a 99,999 habitantes, 3= Localidades con 2,500 a 14,999 habitantes y 4= Localidades con menos de 2,500 habitantes
ins_ali	0= Seguridad alimentaria , 1=Inseguridad alimentaria leve, 2=Inseguridad alimentaria moderada y 3=Inseguridad alimentaria severa

Apéndice D

Resultados omitiendo valores atípicos

En el ejercicio original se decidió mantener en el ingreso las observaciones con un valor que no parece corresponder con el resto de los valores en el grupo de datos, ya que no se puede corroborar que estos valores atípicos se deben a un error al construir la base de datos o en la medición de la variable. Eliminarlos podría no ser la solución ya que puede modificar las inferencias que se realicen, debido a que introduce un sesgo, a que disminuye el tamaño muestral y a que puede afectar tanto a la distribución como a las varianzas. Es decir, la variabilidad (diferencias en el comportamiento de un fenómeno) debe explicarse no eliminarse.

Por otro lado, este tipo de datos pueden generar más influencia que los demás en los cálculos, razón por la cual, se presenta el ejercicio sin valores atípicos en el ingreso.

D.1. Análisis exploratorio de la variable ingreso

La variable a estimar, es el ingreso corriente total per cápita en escala logarítmica (lnictpc). Con la finalidad de identificar de la manera más sencilla los valores atípicos, se presenta la figura D.1.1, donde los puntos color rosa representan a las observaciones que se encuentran lejanas del grupo de observaciones, datos son los valores atípicos que se eliminarán de la base.

- Ajuste de modelo y significancia del efecto aleatorio:

Una vez que los datos atípicos fueron excluidos de la base (389 de 6,924 observaciones), se ajusta el primer modelo («modelo 1»), de la misma manera que en la sección 3.4.

En seguida se muestra la salida generada por STATA, en la primer tabla se observan los valores estimados para los efectos fijos, mientras que, en la segunda tabla se observa que el estadístico de la prueba de razón de verosimilitud resulta significativo ($p - value < .05$), por lo que se conserva el efecto aleatorio.

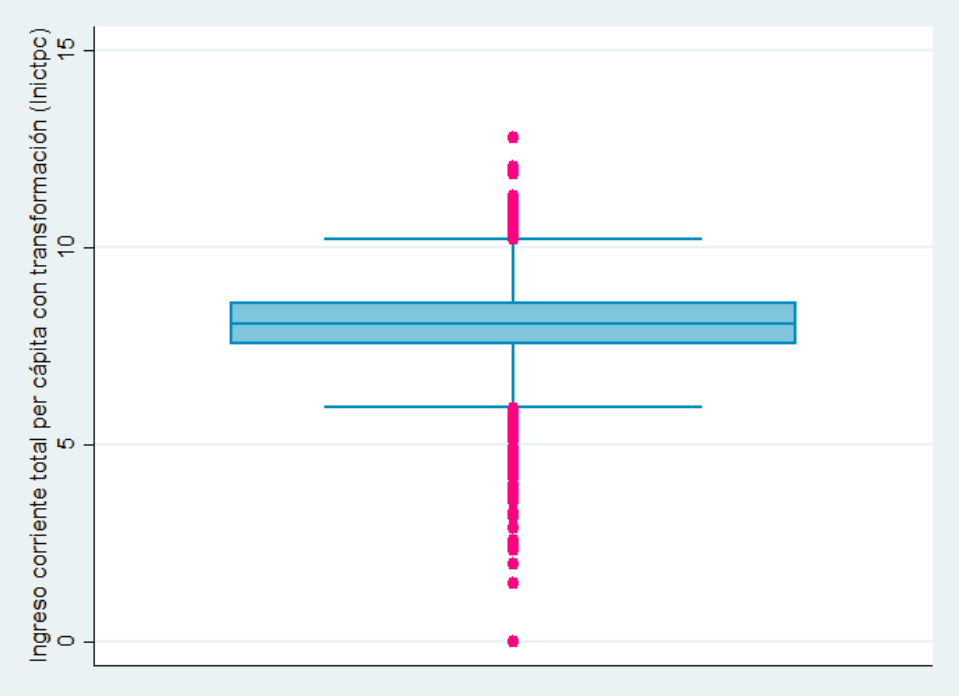

FIGURA D.1.1. Diagrama de caja del ingreso a escala logarítmica.

lnictpc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ijpea_2	-.1913264	.0501854	-3.81	0.000	-.289688	-.0929649
_Ijpea_3	-.0647083	.0230711	-2.80	0.005	-.1099268	-.0194899
_Icyafore_1	.1024275	.0439655	2.33	0.020	.0162568	.1885982
_Icyaguin_1	.2688255	.03892	6.91	0.000	.1925436	.3451073
_Icyuplab_1	-.0248617	.0393118	-0.63	0.527	-.1019115	.052188
_Ijafore_1	.0122453	.0272489	0.45	0.653	-.0411615	.0656522
_Ijaguin_1	.1489374	.0267939	5.56	0.000	.0964223	.2014526
_Ijuplab_1	.0577515	.0227856	2.53	0.011	.0130926	.1024104
jaesc	.0486105	.001772	27.43	0.000	.0451375	.0520836
_Ijsexo_1	-.0637586	.018027	-3.54	0.000	-.0990909	-.0284264
_Iic_cv_1	-.1112869	.0382944	-2.91	0.004	-.1863425	-.0362313
_Iic_sbv_1	-.0047154	.0404278	-0.12	0.907	-.0839524	.0745216
_Icombus2_m_2	.0505826	.0497624	1.02	0.309	-.0469499	.1481151
tam_hog	-.090281	.0042817	-21.09	0.000	-.0986729	-.0818891
_Itamloc_2	-.1215229	.020493	-5.93	0.000	-.1616885	-.0813574
_Itamloc_3	-.1447588	.0269078	-5.38	0.000	-.1974971	-.0920204
_Itamloc_4	-.1728419	.0257386	-6.72	0.000	-.2232887	-.1223951
_Iins_ali_1	-.1719716	.0207149	-8.30	0.000	-.2125721	-.1313711
_Iins_ali_2	-.2911897	.025219	-11.55	0.000	-.3406181	-.2417613
_Iins_ali_3	-.2882125	.0269958	-10.68	0.000	-.3411233	-.2353018
_Iayuotr_1	.0752034	.0156793	4.80	0.000	.0444726	.1059343
_Ibengob_1	-.0924444	.0205735	-4.49	0.000	-.1327677	-.0521212
_Iremesas_1	.1428916	.0422687	3.38	0.001	.0600463	.2257368
_Ijubi_1	.2479599	.0214311	11.57	0.000	.2059558	.2899641
_Itenencia__2	-.0486985	.0204034	-2.39	0.017	-.0886885	-.0087086
_Itenencia__3	-.1947079	.0205786	-9.46	0.000	-.2350412	-.1543747
_cons	7.99757	.0673909	118.67	0.000	7.865486	8.129653

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
ent: Identity				
var(_cons)	.0040379	.0030014	.0009407	.0173323
var(Residual)	.3229044	.0056507	.312017	.3341716

LR test vs. linear model: chibar2(01) = 65.59 Prob >= chibar2 = 0.0000

- Significancia de los efectos fijos:

En este paso, por medio de la prueba del estadístico de Wald (como se realizó en la sección 3.4) las variables, jafore, ic_sbv y cyuplab resultan no signicativas. Siendo así, se eliminan y se llega a la construcción del «modelo 2».

Computing standard errors:

Mixed-effects ML regression — Number of obs = 6,535
Group variable: ent — Number of groups = 4

Obs per group:
min = 1,578
avg = 1,633.8
max = 1,687

Wald chi2(23) = 4048.21
Log likelihood = -5585.6275 — Prob > chi2 = 0.0000

lnictpc	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ijpea_2	-.1916258	.0501697	-3.82	0.000	-.2899566	-.0932951
_Ijpea_3	-.0647689	.0230516	-2.81	0.005	-.1099492	-.0195886
_Icyafore_1	.0948541	.0418143	2.27	0.023	.0128995	.1768086
_Icyaguin_1	.2642974	.0383135	6.90	0.000	.1892044	.3393905
_Ijaguin_1	.1573097	.0204974	7.67	0.000	.1171356	.1974838
_Ijuplab_1	.0592851	.0217663	2.72	0.006	.0166239	.1019463
jaesc	.0487191	.0017644	27.61	0.000	.0452608	.0521773
_Ijsexo_1	-.0634323	.0180195	-3.52	0.000	-.09875	-.0281147
_Iic_cv_1	-.1113638	.0381124	-2.92	0.003	-.1860627	-.0366649
_Icombus2_m_2	.0526409	.0448446	1.17	0.240	-.0352529	.1405346
tam_hog	-.0902028	.0042782	-21.08	0.000	-.098588	-.0818177
_Itamloc_2	-.1216399	.0204922	-5.94	0.000	-.1618038	-.081476
_Itamloc_3	-.1450721	.0268824	-5.40	0.000	-.1977606	-.0923836
_Itamloc_4	-.1741626	.0248909	-7.00	0.000	-.2229478	-.1253774
_Iins_ali_1	-.1719686	.0207152	-8.30	0.000	-.2125696	-.1313675
_Iins_ali_2	-.2916168	.0252082	-11.57	0.000	-.3410239	-.2422097
_Iins_ali_3	-.2888228	.0269785	-10.71	0.000	-.3416996	-.235946
_Iayuotr_1	.0748589	.0156736	4.78	0.000	.0441393	.1055785
_Ibengob_1	-.0927981	.0205606	-4.51	0.000	-.1330961	-.0525002
_Iremesas_1	.1431769	.0422632	3.39	0.001	.0603426	.2260112
_Ijubi_1	.2477611	.0214164	11.57	0.000	.2057858	.2897364
_Itenencia__2	-.0495179	.0203503	-2.43	0.015	-.0894037	-.009632
_Itenencia__3	-.195223	.0205623	-9.49	0.000	-.2355243	-.1549216
_cons	7.995004	.0630936	126.72	0.000	7.871343	8.118665

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
ent: Identity				
var(_cons)	.0040457	.0030066	.0009428	.0173612
var(Residual)	.3229351	.0056512	.3120467	.3342034

LR test vs. linear model: chibar2(01) = 65.86 Prob >= chibar2 = 0.0000

- Criterios de información:

El «modelo 2» como se puede ver en las salidas de STATA, ajusta mejor a los datos al tener un AIC y un BIC más pequeño.

Para el «modelo 1»:

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,535	.	-5585.313	29	11228.63	11425.39

Para el «modelo 2»:

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	6,535	.	-5585.628	26	11223.26	11399.66

D.2. Verificación de supuestos

Para verificar el supuesto de normalidad se muestra en la figura D.2.1 el gráfico de probabilidad normal, con el cuál se puede determinar que los residuos efectivamente siguen una distribución normal, ya que se ubican alrededor de la recta normal. Asimismo, en la figura D.2.2, el gráfico de dispersión de los residuos estandarizados muestran algúnos valores atípicos, sin embargo, existe una fuerte relación lineal con los efectos fijos.

Por útimo, en la figura D.2.3 se muestra una distribución de los residuos homogénea en todas las entidades y destaca Aguascalientes por ser la entidad con menos valores atípicos.

D.3. Resultados

En el diagrama de dispersión que se presenta en la figura D.3.1 entre los valores reales y los valores estimados (ambos en escala logarítmica) se puede observar que el modelo subestima el «lnictpc», no obstante existe una relación positiva entre el valor observado y el estimado, por lo que el modelo ajusto de manera correcta a los datos.

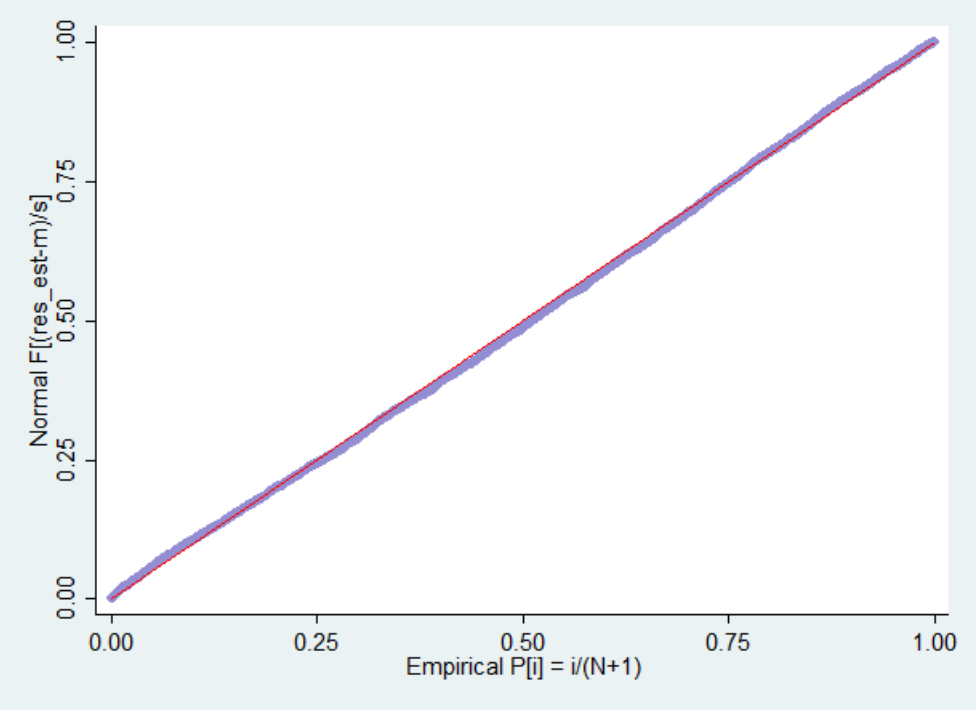

FIGURA D.2.1. Gráfica de probabilidad normal de los residuos estandarizados.

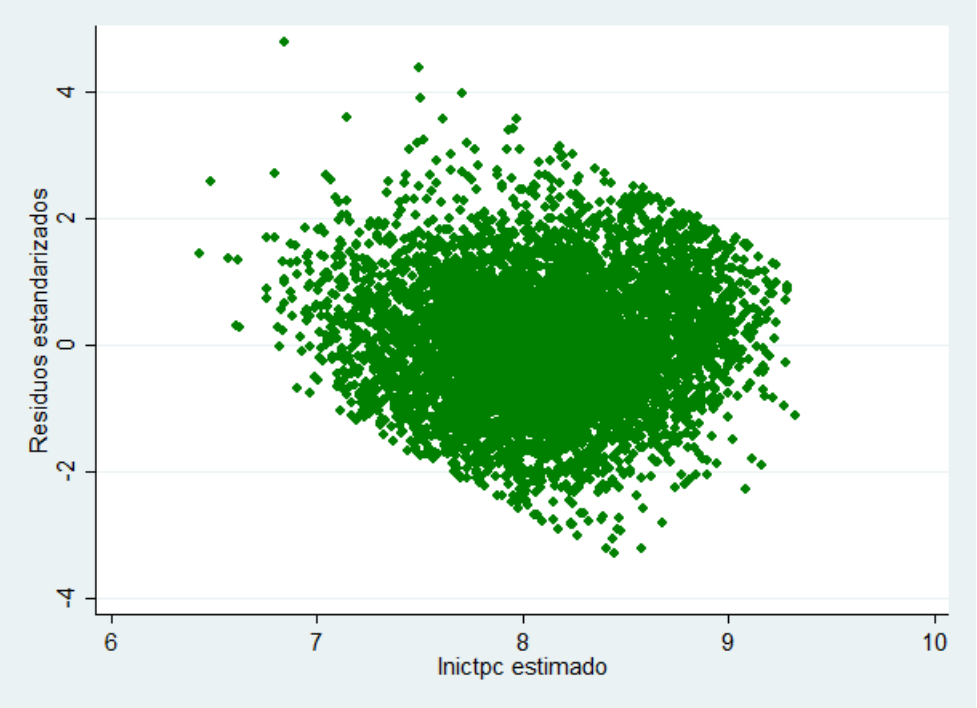

FIGURA D.2.2. Gráfica de dispersión, residuos estandarizados vs Inictpc estimado.

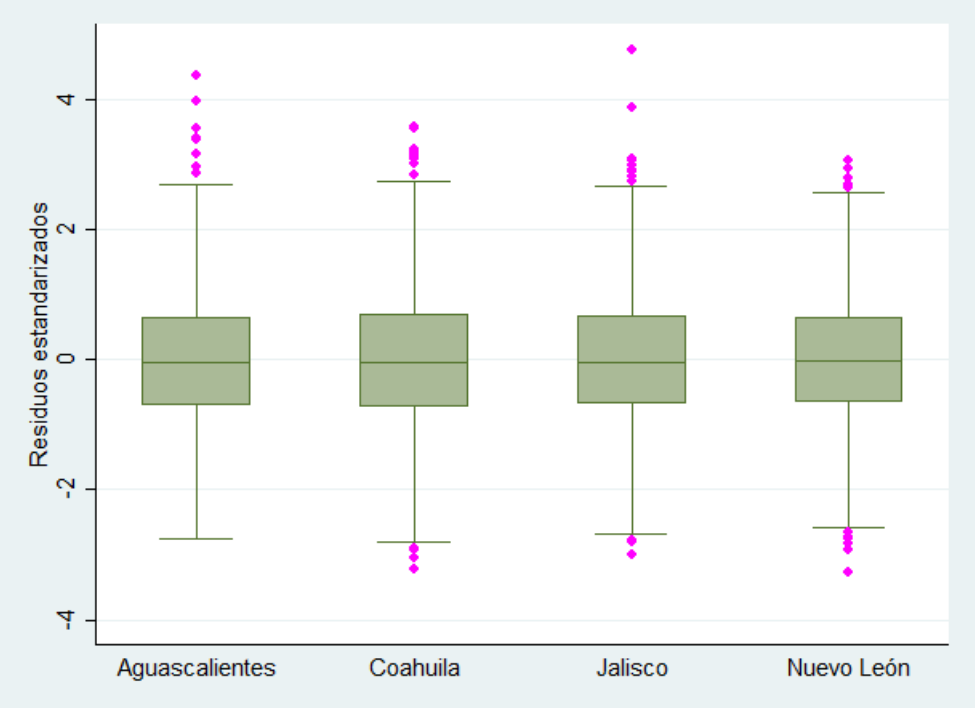

FIGURA D.2.3. Diagrama de caja de los residuos estandarizados según entidad.

Para finalizar, respecto a la figura D.3.2 y D.3.3 se puede observar nuevamente y confirmar que el ingreso estimado esta por debajo del observado, es decir, el modelo subestima y no capta el rango de valores que toma el «lnictpc».

D.4. Conclusiones

El efecto aleatorio resultó significativo, tanto para el «modelo 1» que contempla el total de variables predictoras, como para el «modelo 2» que solamente incluye las variables con efectos fijos significativos (ver sección D.1). Lo anterior y la sospecha de falta de independencia entre las observaciones, es la razón por la que se aplicó un modelo lineal mixto (MLM) de dos niveles, donde el nivel uno corresponde a los hogares y el nivel dos a las entidades.

La selección del mejor modelo se realizó mediante los criterios de información AIC y BIC, resultando ganador el «modelo 2», no se omite señalar, que ambos modelos son alternativas eficientes para la estimación del ingreso.

Acerca de los resultados, se observa que existe una fuerte relación positiva entre el ingreso observado y el estimado, lo cuál implica que el modelo lineal mixto de dos niveles se aplicó de manera correcta.

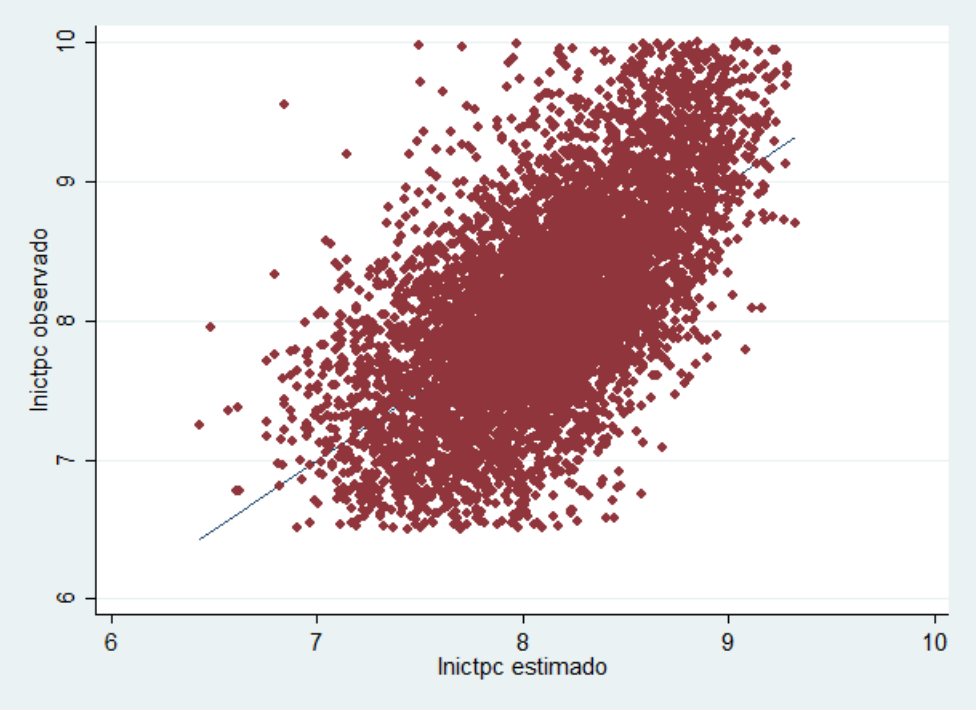

FIGURA D.3.1. Diagrama de dispersión. lnictpc vs lnictpc estimado.

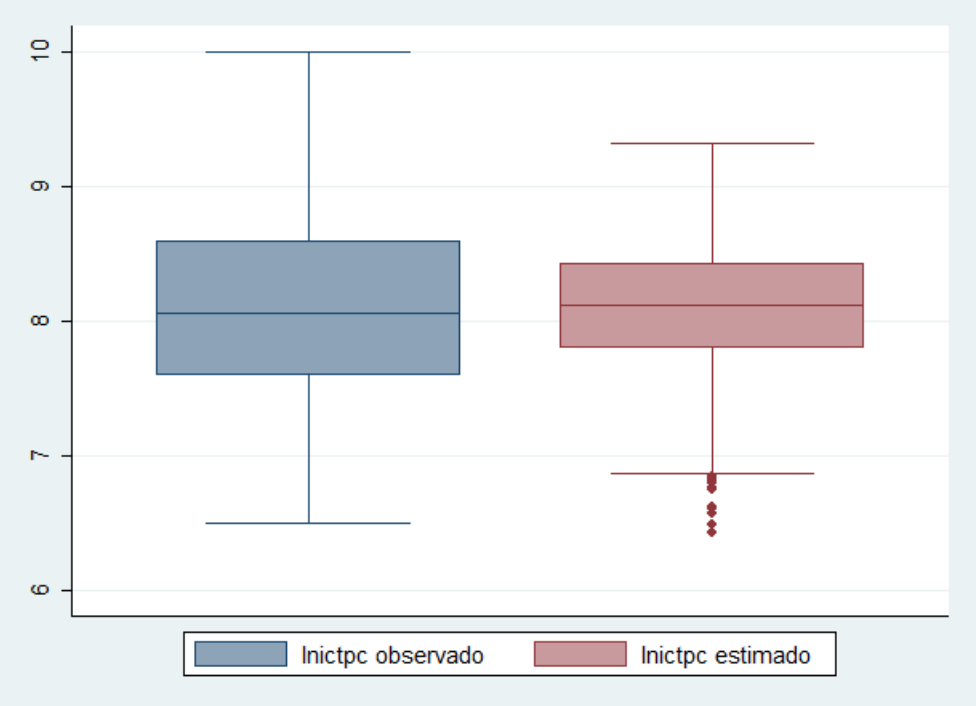

FIGURA D.3.2. Diagrama de caja. lnictpc e lnictpc estimado.

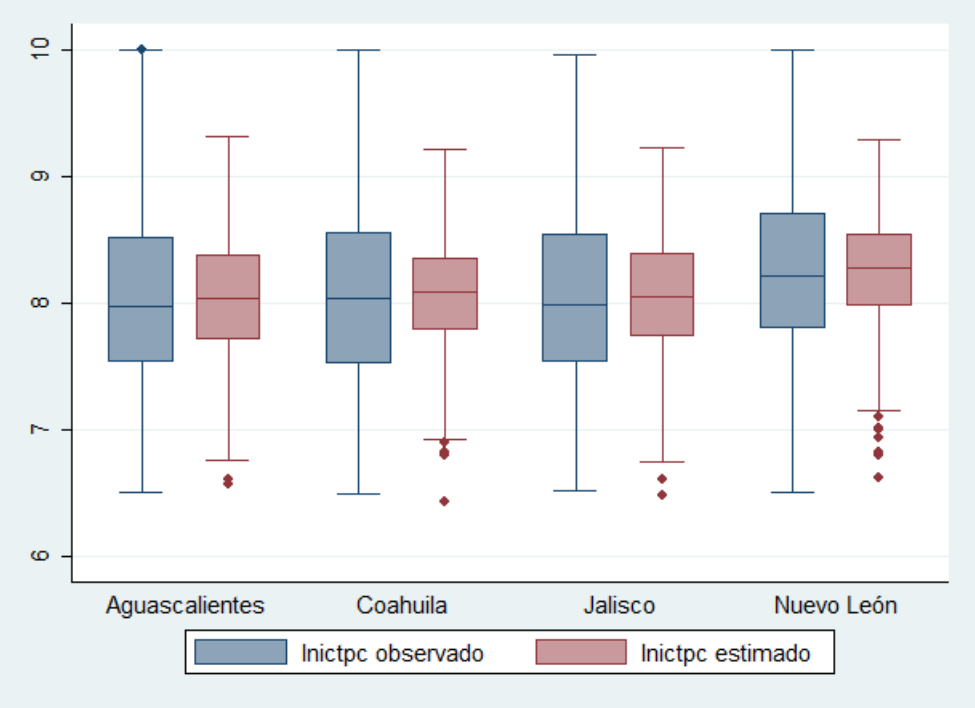

FIGURA D.3.3. Diagrama de caja. lnictpc e lnictpc estimado por entidad.

Por último y a efectos de mejora se sugiere la construcción de nuevas variables y la incorporación de un mayor número de variables predictoras esto apoyaría a predecir de manera mas exacta la realidad, puesto que el modelo propuesto subestima el ingreso.

Apéndice E

Código en STATA, con valores atípicos

```
set more off
clear
cap clear
gl varscompletas "jpea cyafore cyaguin cyuplab jafore jaguin juplab jaesc jsexo ic_cv ic_sbv combus2_mod tam_hog tamloc ins_ali ayuotr bengob remesas jubi tenencia_viv"
gl varscateg "jpea cyafore cyaguin cyuplab jafore jaguin juplab jsexo ic_cv ic_sbv combus2_mod tamloc ins_ali ayuotr bengob remesas jubi tenencia_viv"
*****************************************
***Almaceno las entidades de mi interes***
*****************************************
foreach x in rur urb {
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases coneval\
mcs14hog_reg_‘x’01.dta" , clear
drop ent
rename ent2 ent
keep if ent=="05" || ent=="19"
save "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’01.dta" , replace
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases coneval\
mcs14hog_reg_‘x’02.dta" , clear
drop ent
rename ent2 ent
keep if ent=="14"
save "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’02.dta" , replace
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases coneval\
mcs14hog_reg_‘x’03.dta" , clear
drop ent
rename ent2 ent
keep if ent=="01"
save "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
```

```
mcs14hog_reg_‘x’03.dta" , replace
}
*********************************
***Unión de bases según rururb***
*********************************
foreach x in rur urb {
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’01.dta" , clear
append using "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’02.dta"
append using "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’03.dta"
save "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’.dta", replace
erase "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’01.dta"
erase "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’02.dta"
erase "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_‘x’03.dta"
}
*********************************
***Genero base única de interes***
*********************************
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_rur.dta"
append using "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_urb.dta"
save "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg6.dta", replace
erase "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_rur.dta"
erase "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg_urb.dta"
```

```
**********
***Análisis de datos***
**********
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg6.dta", clear
gen nom_ent="Aguascalientes"
replace nom_ent="Coahuila" if ent=="05"
replace nom_ent="Jalisco" if ent=="14"
replace nom_ent="Nuevo León" if ent=="19"
sum lnictpc
histogram lnictpc
histogram ictpc
pwcorr lnictpc ${varscompletas}
foreach x in $varscateg {
graph box lnictpc , over(${varscateg})
graph export "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Tesis_Sheila_Carbajal\
Imagenes\diag_caja_${varscateg}.png", as(png) replace
}
twoway (scatter jaesc lnictpc, sort)
graph export "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Tesis_Sheila_Carbajal\
Imagenes\disper_jaes.png", as(png) replace
twoway (scatter tam_hog lnictpc, sort)
graph export "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Tesis_Sheila_Carbajal\
Imagenes\disper_tamhog.png", as(png) replace
**********
***Modelo Lineal mixto con una intercepción aleatoria***
**********
#delimit;
*Modelo inicial: modelo 1*;
xi:xtmixed lnictpc i.jpea i.cyafore i.cyaguin i.cyuplab i.jafore i.jaguin i.juplab jaesc
i.jsexo i.ic_cv i.ic_sbv i.combus2_mod
tam_hog i.tamloc i.ins_ali i.ayuotr i.bengob i.remesas i.jubi i.tenencia_viv || ent
:, covariance(identity) variance;
#delimit cr
*Criterios de información AIC y BIC*
estat ic
*Prueba de significancia de los efectos fijos*
test _Itenencia__3 _Itenencia__2
test _Itamloc_4 _Itamloc_3 _Itamloc_2
```

```
test _Iremesas_1
test _Ijuplab_1
test _Ijubi_1
test _Ijsexo_1
test _Ijpea_3 _Ijpea_2
test _Ijaguin_1
test _Ijafore_1
test _Iins_ali_3 _Iins_ali_2 _Iins_ali_1
test _Iic_sbv_1
test _Iic_cv_1
test _Icyuplab_1
test _Icyaguin_1
test _Icyafore_1
test _Icombus2_m_2
test _Ibengob_1
test _Iayuotr_1
test jaesc
test tam_hog
#delimit;
*Nuevo modelo con reducción en los efectos fijos: modelo 2*;
xi:xtmixed lnictpc i.jpea i.cyafore i.cyaguin i.jaguin i.juplab jaesc i.jsexo i.ic_cv
i.combus2_mod tam_hog i.tamloc i.ins_ali
i.ayuotr i.remesas i.jubi i.tenencia_viv || ent :, covariance(identity) variance;
#delimit cr
*Criterios de información AIC y BIC*
estat ic
*Estimaciones*
predict double lctpc_estimado, fitted
*Análisis de residual estandarizado*
predict double res_est, rstandard
pnorm res_est
twoway (scatter res_est lctpc_estimado)
graph box res_est, over(nom_ent)
*Análisis de Resultados*
twoway (scatter lnictpc lctpc_estimado)
twoway (lfit lnictpc lctpc_estimado)(scatter lnictpc lctpc_estimado)
graph box lnictpc lctpc_estimado
graph box lnictpc lctpc_estimado , over(nom_ent)
```

Apéndice F

Código en STATA, sin valores atípicos

```
set more off
clear
cap clear
**********************
***Análisis de datos***
**********************
use "C:\Users\scarbajal\Dropbox\Tesis\Mi tesis\Aplicación\Bases generadas\
mcs14hog_reg6.dta", clear
gen nom_ent="Aguascalientes"
replace nom_ent="Coahuila" if ent=="05"
replace nom_ent="Jalisco" if ent=="14"
replace nom_ent="Nuevo León" if ent=="19"
sum lnictpc
graph box lnictpc
drop if lnictpc<6.5
drop if lnictpc>10
graph box lnictpc
*****************************************************
***Modelo Lineal mixto con una intercepción aleatoria***
*****************************************************
#delimit;
*Modelo inicial: modelo 1*;
xi:xtmixed lnictpc i.jpea i.cyafore i.cyaguin i.cyuplab i.jafore i.jaguin i.juplab jaesc
i.jsexo i.ic_cv i.ic_sbv i.combus2_mod
tam_hog i.tamloc i.ins_ali i.ayuotr i.bengob i.remesas i.jubi i.tenencia_viv || ent
:, covariance(identity) variance;
#delimit cr
*Criterios de información AIC y BIC*
estat ic
*Prueba de significancia de los efectos fijos*
test _Itenencia__3 _Itenencia__2
test _Itamloc_4 _Itamloc_3 _Itamloc_2
```

test _Iremesas_1

test _Ijuplab_1

test _Ijubi_1

test _Ijsexo_1

test _Ijpea_3 _Ijpea_2

test _Ijaguin_1

test _Ijafore_1

test _Iins_ali_3 _Iins_ali_2 _Iins_ali_1

test _Iic_sbv_1

test _Iic_cv_1

test _Icyuplab_1

test _Icyaguin_1

test _Icyafore_1

test _Icombus2_m_2

test _Ibengob_1

test _Iayuotr_1

test jaesc

test tam_hog

#delimit;

Nuevo modelo con reducción en los efectos fijos: modelo 2;

xi:xtmixed lnictpc i.jpea i.cyafore i.cyaguin i.jaguin i.juplab jaesc i.jsexo i.ic_cv i.combus2_mod tam_hog i.tamloc i.ins_ali

i.ayuotr i.bengob i.remesas i.jubi i.tenencia_viv || ent :, covariance(identity) variance;

#delimit cr

Criterios de información AIC y BIC

estat ic

Estimaciones

predict double lctpc_estimado, fitted

Análisis de residual estandarizado

predict double res_est, rstandard

pnorm res_est

twoway (scatter res_est lctpc_estimado)

graph box res_est, over(nom_ent)

Análisis de Resultados

twoway (scatter lnictpc lctpc_estimado)

twoway (lfit lnictpc lctpc_estimado)(scatter lnictpc lctpc_estimado)

graph box lnictpc lctpc_estimado

graph box lnictpc lctpc_estimado , over(nom_ent)