INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

ESTADÍSTICA APLICADA II

EL INGRESO EN LA CIUDAD DE MÉXICO:

UNA INTROSPECCIÓN DE SU DISTRIBUCIÓN EN ÁLVARO OBREGÓN

TRABAJO FINAL

Equipo 6

Daniela Ruíz Martínez
Ismael Solano Ramírez
José Luis Cordero Rodríguez
María Fernanda Vázquez Hernández
Santiago Ayala Moreno
Tonantzin Real Rojas

Índice

1. Introducción

Cuando se habla acerca de los aspectos económicos de la Ciudad de México se suele resaltar su dinamismo en la economía global, su fortaleza como centro financiero del país y de Latinoamérica o su gran disponibilidad de activos financieros. Sin embargo, cuando se discute el tema del ingreso por habitante de la ciudad, poca atención se pone en las peculiaridades del individuo. Por ello, hemos decidido abordar el tema desde un punto de vista microeconómico; es decir, nuestro análisis se enfoca principalmente en los ingresos percibidos por habitantes de la Ciudad de México. Para poder entender con más claridad cómo fluctúa el ingreso entre los distintos ciudadanos, partimos del reconocimiento de diferentes variables que pudieran ser eficientes al tratar de explicar su comportamiento.

A partir de esto, asumimos la naturaleza multifactorial del fenómeno mediante la incorporación de cuatro variables explicativas al modelo. Nuestra intención es poder corroborar dicho supuesto por medio del análisis de estos cuatro elementos y su relación con el ingreso. Para lograr formular conclusiones relevantes y oportunas, realizamos un estudio sobre la relación existente entre el gasto, años escolarizados, erogaciones financieras y la edad de los habitantes con su respectivo ingreso. Aunque no pretendemos que el siguiente trabajo de investigación sea exhaustivo, sí buscaremos esclarecer las posibles asociaciones que nos motivaron a la realización de dicho reporte.

Para este trabajo se utilizó la *Encuesta Nacional de Ingresos y Gastos de los Hogares 2018* (ENIGH) elaborada por el Instituto Nacional de Estadística y Geografía (INEGI). La encuesta bianual realizada en agosto 2018 recaba información del trimestre inmediato anterior a la fecha de aplicación. La base de datos original constaba de un tamaño de muestra de 74,647 hogares; sin embargo, para el propósito del proyecto decidimos trabajar únicamente con una proporción de la muestra. Para ello, tomamos como punto de referencia a la alcaldía Álvaro Obregón en la Ciudad de México. Gracias a esta decisión, logramos reducir el número de datos significativamente, quedándonos únicamente con 144 de ellos. Para la justificación de nuestro trabajo también utilizamos información recabada del *Estudio básico de comunidad objetivo 2018* llevado a cabo por el Centro de Integración Juvenil.

2. Objetivo del modelo

Nuestro trabajo tiene como propósito principal construir un modelo de regresión lineal múltiple que permita explicar el ingreso de los habitantes de la alcaldía Álvaro Obregón en la Ciudad de México a partir de los gastos de los hogares, la procedencia, y las características sociodemográficas y ocupacionales de los integrantes del hogar. De igual manera, se pretende ahondar en las diferencias del ingreso según las características socioeconómicas de las familias y analizar los resultados correspondientes. Los datos proporcionados por el Centro de Integración Juvenil[1] sugieren que la muestra obtenida para el análisis no es representativa de la población nacional; no obstante, el propósito del siguiente trabajo es poder plantear y atender la interrogante sobre la extensión y aplicabilidad de los resultados en un contexto estatal.

3. Variables del modelo

3.1 Variable dependiente

La variable dependiente de nuestro modelo es el **ingreso**; por eso, consideramos que es pertinente contextualizar dicho componente. Esta variable está constituida por la suma de los ingresos por trabajo, rentas y transferencias privadas y/o públicas. De acuerdo con la información presentada en el *Estudio básico de comunidad objetivo 2018*, los niveles de ingreso de la alcaldía Álvaro Obregón se encuentran por arriba de los indicadores nacionales; lo que significa que existe una proporción más grande de los habitantes de este municipio que reciben ingresos más altos. Por otro lado, las tasas de participación económica correspondientes a esta alcaldía son del 45.29% y 71.21% para mujeres y hombres respectivamente. Esto quiere decir que, frente a las tasas de participación nacional que son del 33.46% y 68.48%, las medidas son ligeramente superiores.

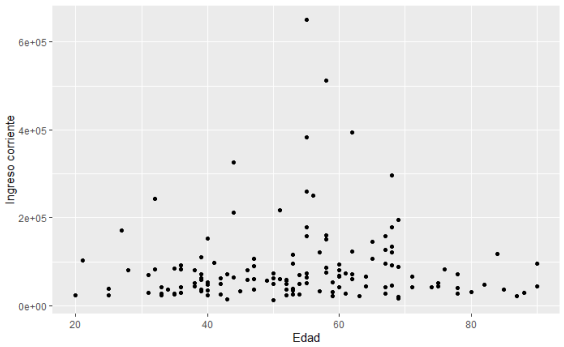
3.2 Variables explicativas y su contexto

Buscamos explicar los niveles de ingreso no solo con variables económicas, sino también sociodemográficas y financieras como lo son los años de escolaridad del jefe de familia y las inversiones de capital, respectivamente.

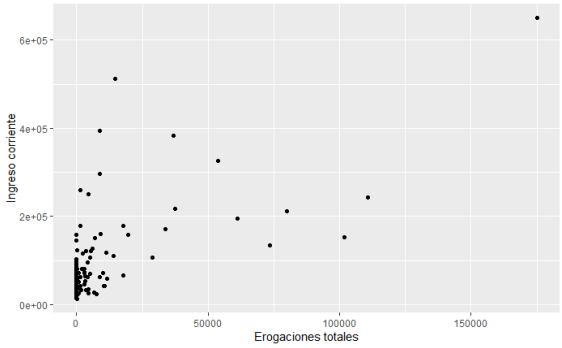
[1] *Estudio básico de comunidad objetivo 2018,* realizado por el Centro de Integración Juvenil.

En México desde una muy temprana edad hasta una muy avanzada se busca conseguir un medio para generar ingresos y en la Ciudad de México esto no es excepción. Por ello, la primera variable que se considera es la **edad**, la cual tiene un valor mínimo de 20 años, un máximo de 90 y una media de 53 años.

El acceso a mercados financieros es algo relativamente común en la Ciudad de México. Consideramos que incluir las erogaciones financieras y de capital tanto monetarias como no monetarias, era importante; por lo tanto, incluimos la variable de **erogaciones totales**. Estas se componen de la suma de depósitos de ahorro, pago por tarjeta de crédito y pago de deudas.



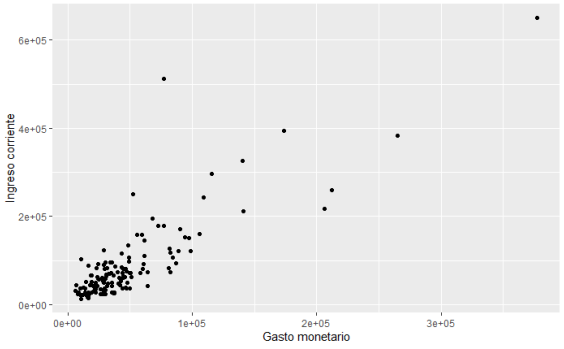
Gráfica 1: Ingreso corriente vs. Edad



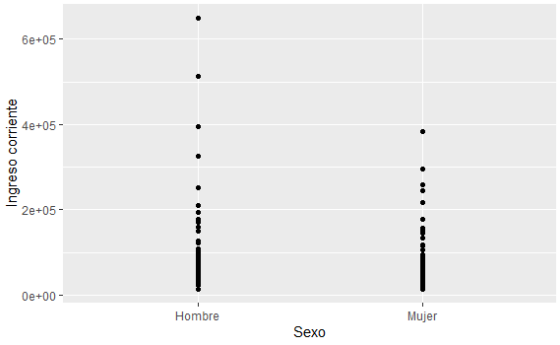
Gráfica 2: Ingreso corriente vs. Erogaciones totales

Si bien la Ciudad de México es uno de los estados con mayor nivel de ingresos del país, el costo de los bienes y servicios también es más alto que en otros lugares. De hecho, la Ciudad de México cuenta con uno de los mercados inmobiliarios más caros del país. Para entender mejor la dinámica del ingreso en los habitantes de la CDMX, decidimos incluir la variable de **gastos monetarios**. Se define como la suma de los gastos regulares que hacen los hogares en bienes y servicios para su consumo.

Una problemática común es la de la desigualdad y violencia de género. De acuerdo con estimaciones del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL), en muchos hogares mexicanos la mujer no solo se ocupa del cuidado del hogar y de los hijos, sino que también debe encontrar un sustento para apoyar económicamente a su familia. Por esto otra variable que se considera es la de **sexo**, vale la pena mencionar que en numerosos estudios se ha encontrado una brecha en el ingreso promedio mensual del 16% entre hombres y mujeres.



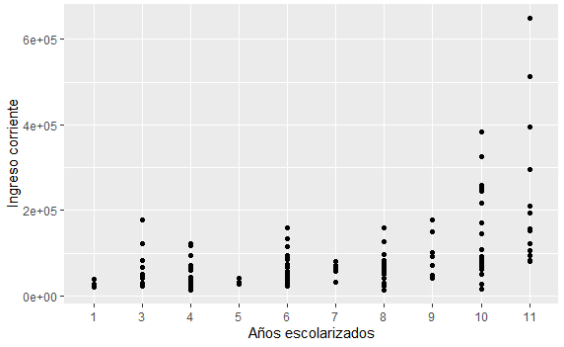
Gráfica 3: Ingreso corriente vs. Gasto monetario



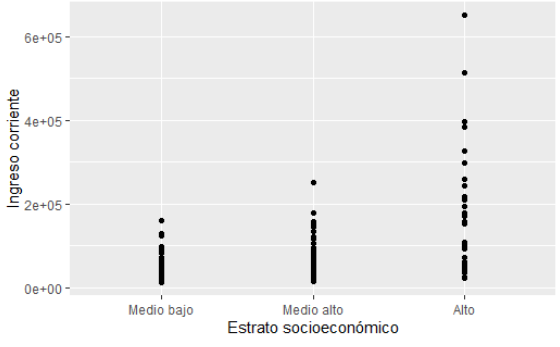
Gráfica 4: Ingreso corriente vs. Sexo

Precisamente como en México se empieza a buscar generar ingresos desde temprana edad, las tasas de deserción escolar suelen ser altas. De acuerdo con el diagnóstico del Derecho a la Educación del CONEVAL, la tasa de escolarización en preparatoria es del 62%. También son bien sabidas las altas tasas de economía informal que hay en todo el país las cuales se calculan ser alrededor de la mitad de la economía nacional. Por esta razón, otra variable que se considera es la de **años escolarizados**, la cual representa el grado escolar máximo aprobado por el jefe del hogar en un rango del 1 al 11.

La última variable que nos pareció importante incorporar es la del **estrato socioeconómico** ya que a pesar de que en la Ciudad de México los ingresos suelen ser más altos que en otros estados del país, la desigualdad y la pobreza son problemas inherentes a la economía de la ciudad. La variable estrato socioeconómico clasifica a las viviendas de acuerdo a sus características físicas y equipamiento de las mismas en 4 etiquetas: bajo, medio bajo, medio alto y alto. Cabe enfatizar que la delegación Álvaro Obregón no cuenta con viviendas de estrato socioeconómico bajo.



Gráfica 5: Ingreso corriente vs. Años escolarizados



Gráfica 6: Ingreso corriente vs. Estrato socioeconómico

4. Construcción del modelo

Como se mencionó anteriormente, nuestro modelo busca explicar el ingreso mediante las 6 variables discutidas en la sección anterior, tentativamente. De esta manera obtenemos la siguiente información:

Error residual estándar	44,800 con 137 gl
$\boldsymbol{R^2}$	0.765
$\boldsymbol{R^2}$ **ajustada**	0.754
Estadístico F	74.31 con 6 y 137 gl
Valor p	2.200×10^{-16}

Tabla 1: Resultados de la prueba F con todas las variables

Observemos que en la tabla 1, al incluir todas las variables en el modelo, la prueba F rechaza la hipótesis nula H_0: $\beta_i = 0 \quad \forall \ i \, \epsilon \, \{1, \ldots, 6\}$, por lo que se propone un primer modelo dado por :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

4.1 Revisiones del modelo

Tras graficar el ingreso corriente vs. todas las variables explicativas, nos quedó una idea más clara acerca de cómo era la correlación que podía haber entre las variables independientes con la dependiente. Sin embargo, para seleccionar las variables que mejor explicaban el modelo decidimos llevar a cabo pruebas t individuales. Los resultados que obtuvimos fueron los siguientes:

	Estadístico t	**Valor p**
Ordenada	-2.894	$4.425*10^{-3}$
Edad	3.609	$4.290*10^{-4}$
Erogaciones totales	2.805	$5.767*10^{-3}$
Gasto monetario	10.548	$< 2*10^{-16}$
Sexo	-1.556	0.122
Años escolarizados	3.063	$2.642*10^{-3}$
Estrato socioeconómico	1.420	0.158

Tabla 2: Resultados de las pruebas t con todas las variables

Notemos que la tabla 2 muestra que las variables que tienen mayor valor p son el sexo y el estrato socioeconómico, por lo que las descartamos. Tras descartar esas dos variables, el nuevo modelo consta de 4 variables explicativas. Al correr nuevamente una regresión lineal pero ahora sobre las 4 variables independientes seleccionadas, obtenemos la siguiente información:

	Coeficiente estimado	**Error estándar**	**Estadístico t**	**Valor p**
Ordenada	-74452.911	$2.146*10^{4}$	-3.469	$6.960*10^{-4}$
Edad (X_1)	998.405	$2.720*10^{2}$	3.671	$3.440*10^{-4}$
Erogaciones totales (X_2)	0.673	$2.286*10^{-1}$	2.943	$3.814*10^{-3}$
Gasto monetario (X_3)	1.228	$1.146*10^{-1}$	10.716	$< 2*10^{-16}$
Años escolarizados (X_4)	6390.161	$1.724*10^{3}$	3.706	$3.030*10^{-4}$

Tabla 3: Resultados de las pruebas t del modelo final

Error residual estándar	45,170 con 139 gl
$\boldsymbol{R^2}$	0.758
$\boldsymbol{R^2}$ **ajustada**	0.751
Estadístico F	108.6 con 4 y 139 gl
Valor p	$< 2.2*10^{-16}$

Tabla 4: Resultados de la prueba F del modelo final

Para este modelo, gracias a las pruebas t (tabla 3) y F (tabla 4) tenemos suficiente evidencia estadística de que las últimas variables seleccionadas explican al modelo. De esta manera, nuestro modelo y la interpretación de los parámetros está dada por:

$$\hat{Y} = -74452.911 + 998.405X_1 + 0.673X_2 + 1.228X_3 + 6390.161X_4$$

donde

Por cada año adicional de vida (X_1), el ingreso aumenta en \$998.405.

Por cada cambio en una unidad monetaria de erogaciones totales (X_2), el ingreso aumenta en \$0.673

Por cada cambio en una unidad de gasto monetario (X_3), el ingreso aumenta en \$1.228

Por cada año escolarizado adicional (X_4), el ingreso aumenta en \$6390.161

Observemos que $b_0 < 0$; sin embargo, la interpretación de este valor no significa que cuando todas las variables explicativas de un individuo son nulas, el ingreso de este sujeto es negativo pues dicho escenario no es realista y los valores reales de las variables explicativas se encuentran en un rango que permite la interpretación de una $b_0' \geq 0$. Es decir, para la interpretación de este parámetro necesitamos considerar el rango de valores de las $X_i \ \forall i \in \{1, \dots, 4\}$ puesto que al considerar la combinación de valores mínimos de dichas variables, se obtiene una nueva ordenada al origen b_0' cuya interpretación ya corresponderá al escenario en el que un individuo genera el mínimo ingreso posible, el cual será no negativo. De hecho, el ingreso mínimo de la muestra es de $13,226.55.

5. Validación de supuestos

5.1 Media del error

Al utilizar el método de Mínimos Cuadrados Ordinarios (MCO) e incluir en el modelo el intercepto β_0 sabemos que, por construcción, la suma de los residuos debe ser cero. Por lo tanto, se cumple el supuesto de media del error igual a cero.

5.2 Autocorrelación

Debido a que nuestra base de datos no es una serie de tiempo ni tenemos datos ordenados, no es posible hacer un análisis de autocorrelación de los errores.

5.3 Linealidad

Por construcción, nuestro modelo es de Regresión Lineal Múltiple (RLM) lo cual hace que sea lineal en los parámetros y en las variables explicativas. Sin embargo, para reforzar esta afirmación observemos lo siguiente:

$$\frac{\partial Y}{\partial \beta_0} = 1 \qquad \frac{\partial Y}{\partial \beta_j} = X_j \ , \quad \forall \ j = \{1, \dots, 4\}$$

Lo cual indica que la derivada parcial del modelo respecto a β_j no depende de $\beta_i \ \ \forall i \neq j$

De igual manera cabe destacar que las variables independientes X_j entran en el modelo como combinación lineal de los parámetros como se muestra a continuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Por lo tanto, podemos decir que efectivamente nuestro modelo es lineal tanto en los parámetros como en las variables explicativas.

5.4 Colinealidad

Para detectar la colinealidad utilizamos la descomposición espectral de la matriz X′X, donde X corresponde a nuestra matriz de datos, cuyos eigenvalores son los siguientes:

$$\lambda_1 = 6.959 \cdot 10^{11} \qquad \lambda_2 = 4.079 \cdot 10^{10} \qquad \lambda_3 = 2.215 \cdot 10^{5} \qquad \lambda_4 = 1.545 \cdot 10^{3}$$

Notemos que los cuatro valores propios distan de cero, lo cual es un primer indicio de que no hay colinealidad. Para cerciorarnos de esto, calculamos el Factor de Inflación de la Varianza (FIV), el cual está dado por:

$$FIV_j = \frac{1}{1 - R_j^2}, \quad \forall\, j = \{1, \dots, 4\}$$

Donde R_j^2 es el coeficiente de determinación múltiple del modelo

$$X_j = \beta_0 + \sum_{i \neq j} \beta_i X_i + \varepsilon_j, \quad \forall\, j = \{1, \dots, 4\}$$

Obtuvimos los siguientes resultados para cada variable explicativa $X_j\ con\ j = \{1, \dots, 4\}$:

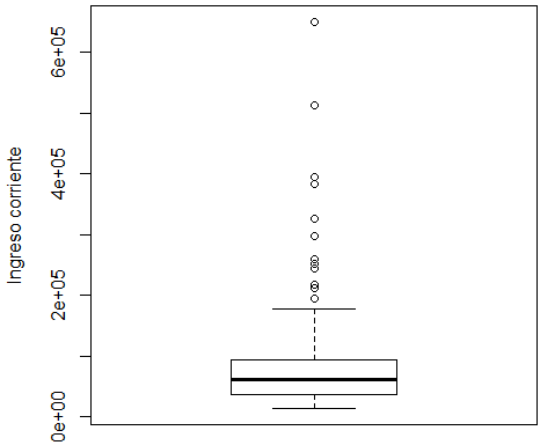
	FIV	$\boldsymbol{R^2}$
X_1	1.229	0.187
X_2	1.823	0.452
X_3	2.120	0.528
X_4	1.536	0.349

Tabla 5:FIV de las variables del modelo

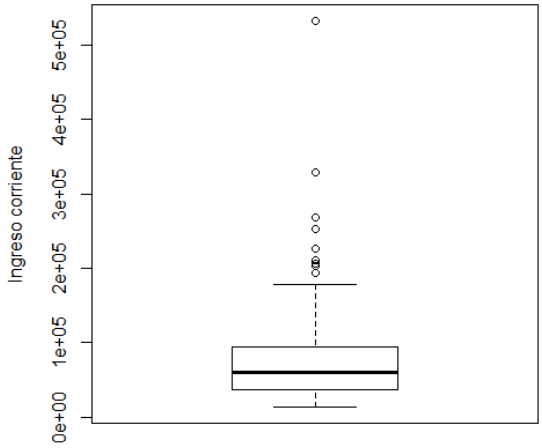
La tabla 5 muestra que $R_j^2 > 0.9$ y en consecuencia, ningún $FIV_j > 10$. Por lo tanto, no hay colinealidad en el modelo.

5.5 Observaciones atípicas

Sabemos que las observaciones atípicas pueden afectar de manera distinta al modelo, en este caso dichas observaciones invalidaban al modelo debido a que afectaban al supuesto de normalidad en los errores, por ello se decidió estimar los datos aberrantes con el modelo propuesto y así corregir el error antes mencionado.



Gráfica 7: Diagrama de caja y brazos con datos originales

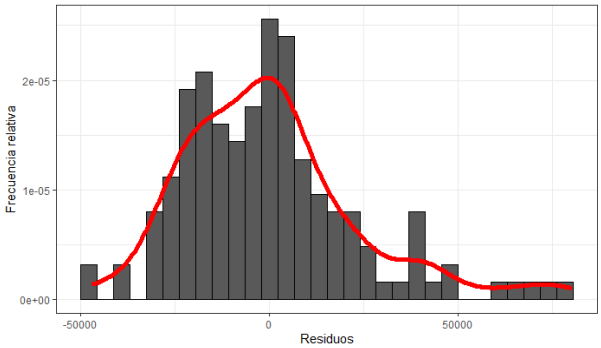


Gráfica 8: Diagrama de caja y brazos con datos atípicos estimados

Además, averiguamos la causa de dichos valores atípicos; de acuerdo con la estimación de la organización Oxfam, en México el 1% de la población recibe alrededor del 21% del ingreso de todo el país. Asimismo, El Economista menciona que México forma parte del 25% de los países con mayores niveles de desigualdad y que, según datos del Banco de México, posee un coeficiente de Gini de 0.48, dicho coeficiente mide la desigualdad salarial donde 0 indica la máxima igualdad y 1 la máxima desigualdad. Por lo anterior consideramos más adecuado estimar los datos aberrantes para tener un modelo que sea válido para todas las observaciones.

5.6 Normalidad

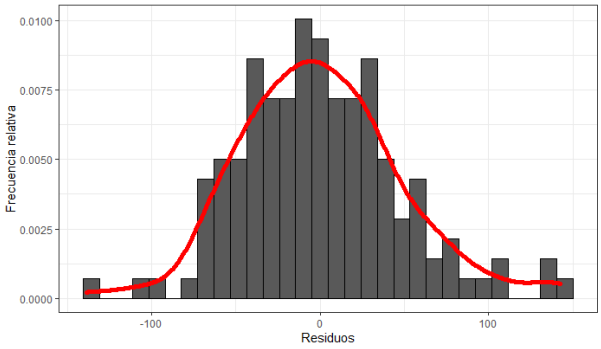
Para la comprobación de este supuesto primero hacemos una inspección del histograma de los residuos con la corrección de los datos atípicos.



Gráfica 9: Histograma de residuos, modelo con datos atípicos estimados

Podemos apreciar que la gráfica está sesgada a la derecha, lo que da indicios de no seguir una distribución normal; para comprobarlo utilizamos la prueba de Jarque-Bera. El estadístico de Jarque-Bera del modelo es: $JB = 36.76 > \chi^2_{(2),0.95} = 5.99$, por lo que confirmamos que hay no normalidad en los errores.

Para corregir este problema aplicamos una transformación potencia de los datos; utilizamos la transformación $\sqrt{Y_i}$ con los datos atípicos sustituidos por los estimados: $Y_i^{1/2}$. Esto es válido debido a que el ingreso es una variable positiva. Con esta transformación obtenemos el siguiente histograma:

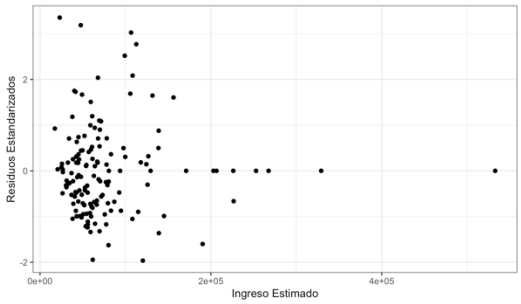


Gráfica 10: Histograma de residuos, modelo transformado

Notemos que el sesgo ha desaparecido; este nuevo histograma ya parece seguir una distribución normal. Para estar seguros volvemos a calcular el estadístico de Jarque-Bera $JB = 5.69 < 5.99 = \chi^2_{(2),0.95}$ lo cual nos conduce a no rechazar la hipótesis nula de que los errores siguen una distribución normal con un nivel de confianza del 95%.

5.7 Heteroscedasticidad

Para validar este supuesto comenzamos con un análisis gráfico del modelo con la corrección de los datos atípicos. En el siguiente diagrama no notamos indicios de violación al supuesto de varianza constante.



Gráfica 11: Ingreso estimado vs. Residuos estandarizados

Para asegurarnos de esto, realizamos la prueba de White, la cual se plantea como $H_0: \sigma_i^2 = \sigma^2 \ \forall\, i = \{1, \dots, n\} \ \ vs. \ \ H_A: \sigma_i^2 \neq \sigma^2 \ \ p.a. \ \ i = \{1, \dots, n\}$ y corremos una regresión auxiliar de los residuos al cuadrado sobre los regresores y sus productos cruzados. Obtuvimos que el estadístico de la prueba de White es $nR^2 = 144 * 0.068 = 9.796$ donde $nR^2 \sim \chi^2_{(14)}$ entonces $nR^2 = 9.796 \ < \chi^2_{(14),0.95} = 23.685$ y, como resultado de la prueba, no se rechaza la hipótesis de homoscedasticidad tomando un nivel de confianza del 95%.

También es importante destacar que la transformación potencia utilizada en el supuesto de normalidad no afecta al supuesto de heteroscedasticidad. Para corroborarlo, en el anexo del presente trabajo se realiza nuevamente un análisis gráfico, así como la prueba formal de White.

6. Conclusión

Dado el análisis preliminar de las variables explicativas, podemos concluir que las variables que tienen mayor impacto en el ingreso son las de edad y preparación escolar, lo cual posiblemente se deba a que, si bien México es un país con casi la mitad de su economía informal, la mayor parte de los trabajos con mayor salario son aquellos que requieren una mayor especificidad o preparación académica, situación que incrementa la ya mencionada desigualdad social en todo el país. Además, de acuerdo con los resultados obtenidos, podemos corroborar que dichas variables explican con gran claridad las diferencias del ingreso entre las distintas familias que habitan en la alcaldía. Por otro lado, también concluimos que las variables con menor peso en el ingreso individual son aquellas económicas o monetarias, específicamente las erogaciones totales y el gasto monetario.

Como limitaciones de nuestro trabajo consideramos; por un lado, que contamos con un número muy reducido de observaciones debido a que solo consideramos los resultados del ENIGH 2018 y por lo mismo no fue pertinente estudiar el supuesto de autocorrelación. Por otro lado, nos parece importante mencionar que para realizar un estudio acerca del ingreso en la Ciudad de México, valdría la pena considerar las diferentes alcaldías para encontrar un contraste entre las diferentes zonas de la ciudad porque, aunque existen similitudes sociodemográficas entre las distintas demarcaciones, no es adecuado generalizar nuestros resultados. Por último, las variables explicativas que consideramos puede que para un estudio que aborde con mayor profundidad el tema de la distribución del ingreso en Álvaro Obregón, sean insuficientes y se requiera incluir otras.

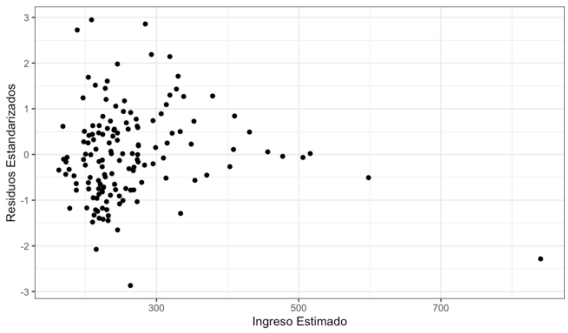
7. Referencias

- García A. (2020). 5 gráficos sobre la desigualdad en México. Recuperado el día 16 de mayo de 2020 de: https://www.eleconomista.com.mx/economia/5-graficos-sobre-la-desigualdad-en-Mexico-20200223-0001.html
- Encuesta Nacional de Ingresos y Gastos de los Hogares 2018. INEGI (2019). Recuperado el día 20 de abril de 2020 de: https://www.inegi.org.mx/programas/enigh/nc/2018/
- Estudio Básico de Comunidad Objetivo 2018. Centro de Integración Juvenil. Consultado el día 15 de mayo de 2020 de: http://www.cij.gob.mx/ebco2018-2024/9460/9460CSD.html
- Quintana, L. et al.. (2016). Econometría aplicada utilizando R. Ciudad de México, México: SAARE.
- Torres A (2016). Coeficiente de Gini, el detector de la desigualdad salarial. Recuperado el día 15 de mayo de 2020 de: https://www.bbva.com/es/coeficiente-gini-detector-la-desigualdad-salarial/
- Usla H. (2019). Desigualdad, la fractura de México. Recuperado el día 15 de mayo de 2020 de: https://www.elfinanciero.com.mx/bloomberg-businessweek/desigualdad-la-fractura-de-mexico

8. Anexo

Supuesto de heteroscedasticidad con transformación potencia

El siguiente diagrama parece sugerir indicios de violación al supuesto de varianza constante.



Gráfica 12: Ingreso estimado vs. Residuos estandarizados transformados

Sin embargo, en este caso el estadístico de prueba de White es $nR^2 = 144 * 0.125 = 17.957$ donde $nR^2 \sim \chi^2_{(14)}$ entonces $nR^2 = 17.957 < \chi^2_{(14),0.95} = 23.685$ y, como resultado, no se rechaza la hipótesis de homoscedasticidad tomando un nivel de confianza al 95%.