Benemérita Universidad Autónoma de Puebla

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Análisis econométrico del gasto en los hogares mexicanos, ENIGH (2022)

T E S I S

para obtener el grado de:

Licenciatura en Actuaría

presenta:

Rosario Juan Francisco

Directora de tesis:

Mtra. Rosalba Mercado Ortiz

Puebla Septiembre, 2024

Agradecimientos

Cada etapa en la vida tiene su propio final, y este, en particular, marca el cierre de un capítulo significativo en mi camino académico. Al llegar a este punto, me gustaría expresar mi sincero agradecimiento a quienes han sido parte fundamental de este viaje.

Agradezco a Dios por su guía constante y sus bendiciones, que me han acompañado en cada paso de esta etapa y me han permitido llegar hasta aquí.

A mis padres, Margarita Francisco Lobato y Carlos Juan Estanislao, quienes han sido mi mayor pilar de apoyo. Gracias por todo lo que me han brindado a lo largo de este camino y por sus consejos. Su amor incondicional y su confianza en mí han sido la base de cada uno de mis logros.

De igual forma, a mis hermanas, Elizabeth y Karla, quienes han sido un apoyo fundamental en este proceso. Gracias por su comprensión, sus palabras de aliento y por estar siempre presentes cuando más lo necesitaba.

A mis amigos, quienes han estado a mi lado a lo largo de este largo camino, les agradezco profundamente por su compañía y apoyo constante. En especial, a Anet, quien ha sido un apoyo invaluable en todo momento.Su amistad y apoyo han sido un faro de luz durante las etapas más difíciles.

A mi directora de tesis Mtra. Rosalba Mercado Ortiz, le agradezco profundamente su invaluable ayuda para que este trabajo pudiera llegar a su fin. Su paciencia, apoyo y comprensión a lo largo de todo este proceso fueron clave para avanzar. Aprecio enormemente su constante disponibilidad para atender mis dudas y orientarme cuando lo necesité. Gracias por sus con-

sejos, tanto académicos como personales, por brindarme su ayuda de manera incondicional, y por darme la confianza necesaria para expresarme con seguridad en diferentes ámbitos. Su guía ha sido fundamental en esta etapa.

A mis sinodales, M.C. Brenda Zavala López, Dr. José Asunción Hernández y Dr. Juan Reyes Álvarez, les agradezco sinceramente por tomarse el tiempo de leer este trabajo y por sus valiosos comentarios y observaciones. Su retroalimentación ha sido fundamental para enriquecer este trabajo.

Índice

Índice de ilustraciones

Índice de Tablas

Introducción

Este proyecto tiene como objetivo proponer un modelo econométrico para explicar las variables que influyen en el gasto de los hogares mexicanos, utilizando los datos de la Encuesta Nacional de Ingresos y Gastos del Hogar (ENIGH) de 2022. Comprender los patrones de gasto es fundamental debido a su impacto en la planificación económica, y el bienestar de los hogares. Según la ENIGH 2022, los hogares mexicanos destinan la mayor parte de su gasto a alimentos, transporte y educación, los cuales representan el 66.8 % del gasto total. En este contexto, el gasto corriente, definido por la ENIGH como los gastos regulares del hogar en su canasta de consumo, incluye rubros como alimentos, vivienda, salud y transporte, entre otros.

La hipótesis del estudio plantea que el gasto de los hogares mexicanos está significativamente influido por las categorías de alimentos, transporte y educación, dado que representan una parte considerable del presupuesto familiar. Se espera que las variaciones en estos gastos principales tengan un impacto notable en el gasto total del hogar.

Para analizar el gasto de los hogares mexicanos, inicialmente se emplearon modelos de regresión lineal múltiple con el objetivo de obtener un análisis detallado de los datos. Sin embargo, este enfoque resultó inadecuado debido a la presencia de heterocedasticidad y la no normalidad de los errores. Se intento corregir mediante el uso de deciles del gasto, sin embargo, no se resolvieron estos problemas, sugiriendo una forma funcional incorrecta. Posteriormente, se consideró un modelo de propensión marginal, pero la persistencia de la heterocedasticidad

llevó a la adopción del modelo Tobit, que finalmente proporcionó resultados más consistentes y ajustados.

Con la ayuda de los efectos marginales del modelo Tobit se tiene que un cambio en el estatus socioeconómico reduce el gasto del hogar en 218.98 unidades, mientras que un mayor nivel educativo del jefe del hogar lo incrementa en 316.79 unidades. Además, el gasto en categorías específicas como alimentos, vivienda, salud, comunicaciones, educación y transporte contribuye de manera significativa al gasto total del hogar. Un aumento en el ingreso corriente incrementa ligeramente el gasto total, reflejando la relación entre ingresos y consumo. Asimismo, se realizaron análisis adicionales en hogares representativos de la media de los deciles y en cuatro hogares de cada estrato socioeconómico.

Este estudio proporciona información valiosa que puede ser utilizada para diseñar políticas económicas más efectivas y entender mejor los comportamientos de gasto en diferentes segmentos de la población. No obstante, es importante considerar que los resultados se basan en datos de una única encuesta, y la complejidad de los datos requiere un análisis continuo para una comprensión completa y precisa del gasto de los hogares.

El análisis mediante el modelo Tobit revela cómo factores como el estatus socioeconómico y el nivel educativo del jefe del hogar influyen en el gasto total del hogar. Este enfoque es particularmente útil en Actuaría, ya que permite modelar fenómenos donde las variables dependientes están limitadas o censuradas, como en el caso del gasto, que no puede ser negativo. Los hallazgos obtenidos de este modelo son cruciales para la modelización del riesgo y la planificación financiera.

Desde una perspectiva actuarial, este tipo de análisis ayuda a comprender las variaciones en el gasto y su relación con los ingresos, permitiendo identificar cómo las restricciones presupuestarias y las preferencias de consumo varían entre diferentes segmentos de la población. Estas herramientas cuantitativas permiten a los actuarios ajustar las primas o beneficios de productos financieros y seguros en función de factores económicos, mejorando la precisión en la estimación de riesgos asociados a distintos grupos poblacionales.

Además, este análisis tiene implicaciones más allá del ámbito financiero y asegurador. En el contexto de políticas públicas, proporciona información clave para prever el impacto de cambios económicos, como las fluctuaciones de la inflación, sobre la capacidad de consumo y ahorro de los hogares. Al tener una mejor comprensión de los patrones de gasto, se promueve la creación de modelos más precisos para gestionar riesgos financieros y mejorar la estabilidad económica.

En resumen, el uso de modelos como el Tobit en la Actuaría no solo optimiza la gestión del riesgo financiero, sino que también contribuye a la toma de decisiones informadas tanto en el sector privado como en la formulación de políticas públicas, fomentando así la estabilidad económica y financiera a largo plazo.

Capítulo 1: "El gasto y sus componentes en los hogares mexicanos."

1.1 Relación entre gastos e ingresos en hogares mexicanos

El interés de estudiar los gastos en los hogares mexicanos nace de observar como en el año 2023 el precio de la canasta básica había incrementado por la inflación, sin embargo, el salario mínimo no había aumentado, lo cual nos lleva a cuestionar si dicho salario es suficiente para cubrir los gastos de los hogares, dado a que es importante conocer la relación existente entre el gasto y los ingresos, ya que los ingresos dentro de un hogar varían dependiendo al trabajo del jefe del hogar y la actividad económica a la que se dediquen sus integrantes, puesto que "las diferencias salariales en México son relativamente marcadas, tanto desde la perspectiva de la escolaridad como por género, sindicalización, tipo de contrato y territorio." (Varela Llamas et al., 2009).

Es fundamental mantener el gasto por debajo del ingreso para garantizar una estabilidad financiera. Esto cobra especial relevancia, ya que un desequilibrio entre ingresos y gastos puede desencadenar problemas socioeconómicos significativos, como señala Arellano Hidalgo (2021), en su estudio, encontró que durante el período 2016-2017, los ingresos per cápita disminuyeron en un 2.5 %, lo que resultó en un aumento en los niveles de pobreza. De hecho, un alarmante 41.0 % de la población se encontraba por debajo del costo de la canasta básica.

Por lo que, es fundamental comprender la composición de los gastos en los hogares para identificar las disparidades entre los niveles de ingresos y los gastos en México.

1.2 Gasto en los hogares mexicanos

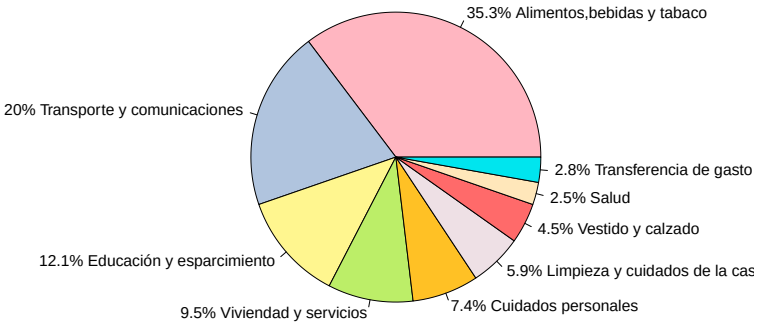
El gasto se conoce como el pago destinado para el consumo de bienes y servicios dentro de un hogar en un cierto periodo para cubrir las necesidades de los integrantes del hogar, además debemos destacar que el gasto de los hogares es una variable importante para la economía mexicana dado a que "el consumo de los hogares representa una importante proporción respecto del Producto Interno Bruto (más del 70 %), el tener un grupo tan amplio de la población en condiciones precarias de capacidad de consumo" (Martínez Solares, 2023).

Para conocer sobre los ingresos y gastos en México el Instituto Nacional de Estadística y Geografía (INEGI) realiza la Encuesta Nacional de Ingresos y Gastos en los Hogares (ENIGH), la cual tiene como "objetivo exponer un panorama estadístico de la conformación de ingresos y gastos dentro de los hogares mexicanos, en cuanto a su monto, distribución y procedencia, además de brindar características sociodemográficas y ocupacionales de cada uno de los integrantes, las características sobre el hogar y su equipamiento" (INEGI, 2022). Por otro lado, proporciona las variaciones que tienen los ingresos y gastos de los hogares de forma bienal, lo que nos permite obtener datos actualizados de las fuentes de ingresos y los gastos de los hogares de México.

De acuerdo con la ENIGH del año 2018 el gasto corriente monetario promedio trimestral era de 31,913 pesos, de los cuales el 35.3 % se destinó al gastó en alimentos, bebidas y tabaco,

siendo así el mayor gasto dentro de los hogares mexicanos, podemos observar por otro lado, el 20 % se destinó para el transporte y comunicaciones, mientras que a la educación el 12.1 % siendo el tercer más importante de los gastos lo cual se debe a que ayuda al crecimiento y el desarrollo de capacidades individuales de los integrantes de los hogares, sin embargo, los hogares mexicanos dirigían el 2.6 % a su salud siendo este el último de las nueve categorías, como se observa a continuación:

Gráfico 1: ENIGH 2018



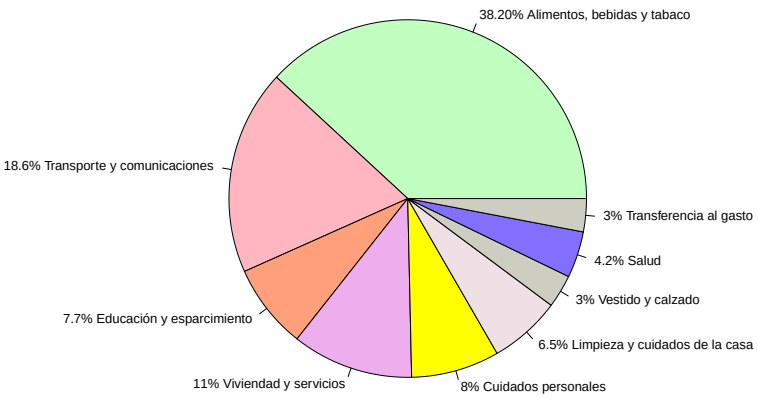
Fuente: Elaboración propia con datos extraídos de la ENIGH 2018.

Por otro lado, en la ENIGH del 2020 se observó que el gasto de los hogares se vio afectado por la pandemia del COVID-19, por lo que gasto corriente monetario promedio trimestral fue de 29,910 pesos, es decir el gasto disminuyo el 8.47 %, de igual forma la distribución de los gastos se vio afectada por una ligera disminución en los rubros, por lo que el 38 % se destinó para alimentos y bebidas, el 18.6 % a transporte y comunicaciones, para ese año se gastaba el 11 % en la vivienda y los servicios, dejando a la educación con el 7.7 % lo cual se debe a la modificación de clases presenciales a clases en línea, por lo que ya no se gastaba en esparcimiento y el gasto en artículos educativos disminuyo debido a la contingencia.

Es importante destacar que para dicho año los gastos en la salud aumentaron siendo del

4.2 %, lo cual es debido a la contingencia sanitaria, dejándolo en el número siete de las nueve categorías.

Gráfico 2: ENIGH 2020



Fuente: Elaboración propia con datos extraídos de la ENIGH (2020).

Finalmente, en la ENIGH de 2022 el gasto corriente monetario promedio trimestral fue de 39 965 pesos, el cual se distribuyó de misma forma en rubros del 2018, para alimentos y bebidas se gastó el 37.7 %, para transporte y comunicaciones el 19.3 %, teniendo un aumento del 0.7 % en comparación del 2018, siendo un efecto del COVID-19, debido a que por la pandemia la dinámica de la población cambio, por lo que "los servicios de telecomunicaciones presentaron una alta demanda para hacer frente a nuevas y variadas dinámicas en las que se vio inmersa la población, desde la atención médica a distancia hasta la búsqueda de entretenimiento para sobrellevar el confinamiento" (Ortíz, 2024).

1.3 Base de datos *concentradohogar*

La base de datos que se usará será 'concentradohogar', obtenida de la página oficial del INEGI a través de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2022 - Nueva

Serie. Esta base, disponible en la sección de microdatos, está conformada por 126 variables distribuidas en 9 categorías principales y 49 subcategorías, representando diversas influencias en el gasto de los hogares mexicanos (ver Anexos B: Tabla 47). Para comprender el concepto de cada variable, se consultó la descripción detallada proporcionada en la red nacional de metadatos DDI en la sección de infraestructura de la información, datos y metadatos del INEGI, lo que nos permitió entender su significado y aplicación.

Es importante destacar que no se incluirán las nueve categorías en las que se divide el gasto, ya que todas ellas lo conforman en su totalidad. Esto podría causar un problema de colinealidad, resultando en un análisis erróneo. La intención no es sólo determinar el peso de las categorías, sino también revisar las relaciones entre el gasto y sus principales componentes, así como otras variables fundamentales.

Para nuestro análisis, seleccionamos inicialmente 17 variables, de las cuales 8 pertenecen a las características del hogar, una dimensión fundamental para comprender la estructura del hogar y observar las necesidades de la población. Estas variables son: ubicación geográfica (ubica_geo), estrato socioeconómico (est_socio), clase del hogar (clase_hog), sexo del jefe del hogar (sexo_jefe), edad del jefe del hogar (edad_jefe), educación del jefe del hogar (educa_jefe), número de integrantes (tot_integ) y número de integrantes menores de 11 años (menores).

Las otras 4 variables se seleccionaron en base a los resultados de la ENIGH (2022) sobre las áreas de mayor gasto en los hogares mexicanos: alimentos, comunicación, educación y transporte. Además, se incluyeron 2 variables adicionales que, desde mi punto de vista,

deberían ser igual de importantes: salud y vivienda. Finalmente, la variable gasto corriente monetario (gasto_mon) es nuestra variable de interés, y se seleccionó la variable ingreso corriente (ing_cor) por ser fundamental para comprender y analizar los hábitos de gasto y el bienestar económico de los hogares.

De las que se excluyeron dos variables: ali_dentro (alimentos consumidos dentro del hogar) dado a que estos son considerados en la variable 'alimentos' y menores (integrantes menores de 11 años), ya que se encontraban incluidos en 'tot_integ'. Por otro lado, hemos separado la variable 'comunicaciones' de la variable 'transporte' porque la primera estaba incluida en la segunda. De esta manera, podemos analizar el gasto destinado únicamente al transporte, mientras que consideramos las comunicaciones de manera independiente. Esto se hizo para evitar sesgos al realizar el modelo econométrico.

Cabe destacar que, de las variables seleccionadas, siete influyen directamente en el gasto, mientras que las restantes son características demográficas y del jefe de hogar, lo cual observa en la tabla 48 en los Anexos B.

1.4 Componentes del gasto

Para empezar a trabajar los datos hemos seleccionado únicamente a las variables que influyen directamente en el gasto corriente monetario de acuerdo con la ENIGH(2022), por lo que tomamos en cuenta a las variables gasto_mon, alimentos, educacion, salud, tot_integ y transporte, ya que de acuerdo con la ENIGH (2022) los alimentos son el primer rubro en el que gastan más los hogares mexicanos con el 37.7 %. Además, consideramos la variable de

"transporte" que representan el segundo gasto más importante con el 19.3 %, y la variable de "educación" que ocupan el tercer lugar en el gasto corriente de los hogares con el 9.8 %.

Por otro lado, la variable salud fue seleccionada debido a que es el rubro en que los hogares mexicanos gasta menos y la variable "tot_integ" también fue incluida para tener en cuenta el número de integrantes en cada hogar, ya que este factor influye en el gasto.

Es importante resaltar que es fundamental comprender cómo se conforma cada una de las variables mencionadas anteriormente. Por lo tanto, se obtuvieron las estadísticas descriptivas de cada variable, como se observa a continuación:

Tabla 1: Estadísticos descriptivos de las variables

gasto_mon	alimentos	educacion	salud	tot_integ	trans
Min. : 0	Min. : 0	Min. : 0	Min. : 0.0	Min. : 1.000	Min. : 0.0
1st Qu.: 18561	1st Qu.: 7483	1st Qu.: 0	1st Qu.: 0.0	1st Qu.: 2.000	1st Qu.: 925.7
Median : 29679	Median : 11957	Median : 0	Median : 146.7	Median : 3.000	Median : 3193.5
Mean : 37615	Mean : 14046	Mean : 2467	Mean : 1270.6	Mean : 3.435	Mean : 5739.4
3rd Qu.: 45901	3rd Qu.: 17961	3rd Qu.: 2177	3rd Qu.: 841.3	3rd Qu.: 4.000	3rd Qu.: 6930.0
Max. :1703575	Max. :849840	Max. :451161	Max. :324547.7	Max. :19.000	Max. :489130.4

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

De las estadistícas notamos el gasto promedio de los hogares es de 37,615.00 y de los gastos destinados en alimentos es de 11,957.00 lo cual nos representa el 31.788 % del gasto promedio, ademas de que los integrantes de los hogares va de 1 a como máximo 19 integrantes dentro de un hogar, siendo el promedio de 3 integrantes por hogar, es importante resaltar que las variables *gasto_mon*, *alimentos*,*educación*,*salud* y *trans*, contienen en los números mínimos ceros, lo cual sera importante para la modelación más adelante

Es importante resaltar que la variable *gasto_mon* será la variable endógena y cuantitativa, mientras que las demás variables serán las explicativas y de igual forma son cuantitativas.

1.4.1 Alimentos

La alimentación es un rubro significativo para medir los gastos de los hogares mexicanos dado a que es el que tiene más importancia dentro de la economía de un hogar, puesto que de "cada 100 pesos de gasto, 38 pesos son destinados a la compra de alimentos" («Cuánto gastamos y en qué gastamos», 2023) lo cual se ha observado a través de la ENIGH en donde al medir en que gastan los hogares, se ha observado que el primer rubro es alimentación, bebidas y tabaco. Siendo este mayor al treinta por ciento en las cuatro encuestas realizadas cada dos años siendo la primera en 2016 y la más reciente en 2022, por lo que se ha observado que el mayor gasto dentro de un trimestre de este rubro es en los alimentos y bebidas consumidas dentro del hogar, con ello hacemos referencia a la compra de los productos de la canasta básica.

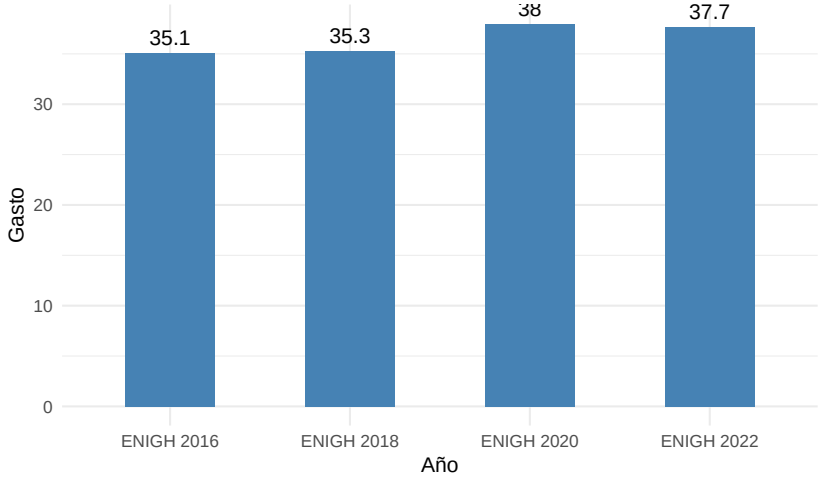
De acuerdo con Saldívar(2023), para el año 2022 los gastos que tenían los hogares del primer decil eran de 14,722 pesos en un trimestre, de los cuales el 51.1 % se destinó a los alimentos, en términos monetarios alrededor de 7,508 pesos que se gastaron en este rubro, lo cual es muy diferente en el décimo decil el cual esta conformado por hogares con mayores ingresos y en donde se excluyen a los multimillonarios,dichos hogares tenían un gasto trimestral de 102,241 pesos, y de los cuales destinaron 28.3 % de este a comprar alimentos, es decir, la distribución del gasto destinado alimentos va descendiendo conforme los hogares tienen mayores ingresos.

Además, teniendo encuenta al colectivo México ¿Cómo Vamos? (2022) se tiene que para el año 2022 se tuvieron tasas de inflación superiores al 7 %, lo que causo que los precios de los alimentos aumentaran a tasas de doble dígito durante el año, lo que conlleva a a una

reducción la capacidad de compra para los hogares de menores ingresos se pronuncien debido al mayor peso que tiene el gasto en alimentos en su ingreso corriente. León Bon (2020) nos dice que la inflación de la canasta alimentaria en los últimos años ha sido mucho mayor que el aumento en los salarios, lo que trae una pérdida en el poder adquisitivo real para los hogares en condición de pobreza.

De igual forma es importante destacar que dicho gasto ha ido creciendo durante los años, para el año 2016 los hogares gastaban 13,502 pesos mexicanos en la compra de los productos, mientras que para el año 2022 fue de 15,059 pesos mexicanos, en términos porcentuales el gasto incremento el 10.34 %, el cual se debe a que en el agosto de 2022 justo en el momento de levantamiento de la encuesta se observó que la inflación era 13.6 % anual, la cual causaba incremento en los precios de los productos como carne, tortillas, frutas y verduras, por mencionar algunos, a continuación se presenta un gráfico para observar el crecimiento del gasto.

Gráfico 3: Gasto porcentual en alimentos ENIGH por año



Fuente: Elaboración propia con datos de ENIGH (2022)

Por otro lado, están los alimentos y bebidas consumidas fuera del hogar, el hecho de que

las personas gasten en comprar comida, es porque con el paso del tiempo se ha dado la situación de que por el trabajo las personas tienen menos tiempo para comer, lo que los lleva al consumir alimentos fuera de sus hogares. Para el año 2020 los hogares gastaban 1,733 pesos dado a que por cuestiones de la pandemia no se tenía permitido comer en lugares públicos o fuera del hogar para evitar contagios, sin embargo, para el año 2022 al reducirse las restricciones el gasto para este rubro fue de 2,957 pesos, es decir el gasto tuvo una variación del 70.6 %.

Además, el gasto corriente monetario promedio trimestral por deciles de ingreso por hogar nos dice que "en los hogares del primer al sexto decil (el 60 % de los hogares con menores ingresos del país), más del 40 % del gasto se destina a los alimentos. Es decir, seis de cada 10 hogares destinan al menos cuatro de cada 10 pesos a alimentos, bebidas y tabaco." («Cuánto gastamos y en qué gastamos», 2023).

Por lo que, los hogares que perciben menos ingresos y ante la necesidad de gastar para los alimentos destinan menos dinero a otros bienes y servicios.

1.4.2 Educación

La educación es un punto importante dentro de los hogares mexicanos, dado a que se tiene la mentalidad de que el tener más años de educación mejorará la calidad de vida porque los ingresos generados serán más altos, sin embargo, la ENIGH de 2022 muestra que de un gasto corriente monetario promedio trimestral de 39,965 solo se le destinan 3,921 a la educación que representa el 9.81 % del gasto de los hogares, lo cual nos quiere decir que los hogares no invierten lo suficiente en la educación, lo cual se puede deber a diferente factores uno

de ellos es el incremento que tuvieron los alimentos mientras que los hogares con ingresos altos no tuvieron problemas para sobrellevar el incremento, sin embargo, en los hogares con ingresos limitados se vieron en la necesidad de disminuir el gasto hacia la educación, dado a que "los hogares de mayores ingresos gastan en educación 16 veces lo que gastan los hogares de menores ingresos" (Hernández, 2023).

Cabe aclarar que no es que no se quiera invertir en la educación solo que se volvió más caro invertir en ella, dado a que "mientras que los hogares de menores ingresos destinan alrededor de 242 pesos mensuales a la educación, los hogares con mayor disponibilidad de ingresos pueden dedicar 4 mil 702 pesos mensuales a ello, en promedio." («Cuánto gastamos y en qué gastamos», 2023). Junto a la disminución general de estudiantes en las escuelas, según Hernández (2023), en el año 2022 hubo 700,164 estudiantes menos en comparación con 2018. Esta reducción afectó de manera significativa a los niños que asisten a preescolar, con una caída de 433,000 estudiantes en ese nivel educativo.

1.4.3 Salud

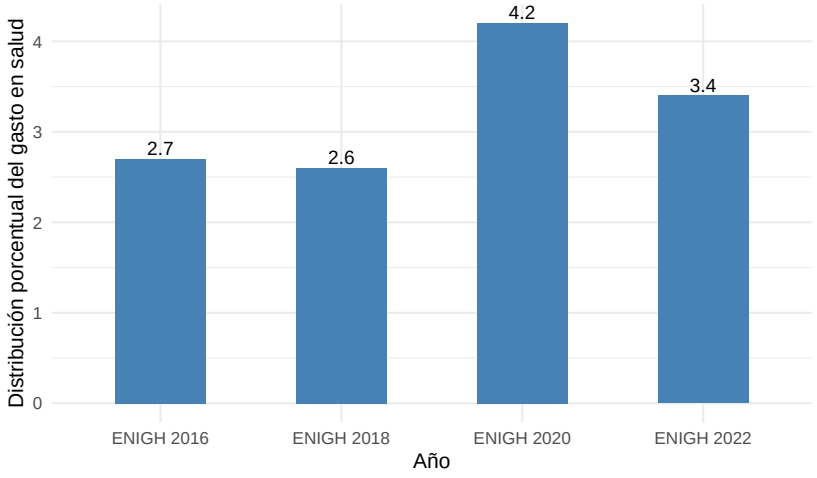
El rubro de la salud en los hogares mexicanos es importante, debido a que se vuelve una determinante de bienestar, además de mostrar el crecimiento económico de los hogares o en ocasiones la decadencia dentro de su economía, ya que aumentar el dinero destinado en la salud causa una reducción en el gasto corriente de los hogares.

Es importante destacar que México de acuerdo con Mendéz Méndez (2022) la insuficiencia presupuestaria y la falta de eficiencia en el gasto destinado para el sector de salud, refleja una caída de la población que se tiene afiliada a servicios de salud, es decir en la tasa de

atención pública y el aumento en el gasto de bolsillo.

Además con el paso de los años, la población ha ido incrementando la cantidad de dinero trimestral destinada para la salud. Observamos que el gasto en salud para el año 2016 fue de 2.7 %, mientras que para el año 2022 los hogares destinaron solo el 3.4 % de sus gastos en salud, y como podemos darnos cuenta el incremento porcentual fue muy pequeño, aunque pasaron 6 años; lo que nos lleva a decir que el dinero destinado de manera trimestral para la salud incremento 0.1167 % de manera trimestral durante ese tiempo, es importante destacar el hecho de que en el año 2020 dinero gastado a en salud fue de 4.2 % del gasto corriente monetario total trimestral , sin embargo, para el año 2022 disminuye nuevamente. Es importante resaltar que el rubro de salud es el menos significativo para el gasto de los hogares dado a que casi siempre ocupa el lugar número ocho de las nueve categorías existentes, a excepción del año 2020 que ocupo el lugar número siete, aunque sigue siendo poco significativo para los hogares.

Gráfico 4: Gasto porcentual en salud ENIGH por año



Fuente: Elaboración propia con datos de ENIGH (2022)

1.4.4 Número de integrantes en el hogar

Los hogares mexicanos han disminuido el número de integrantes dado a que la ENIGH ha registrado que en el año 2018 los hogares en promedio se conformaban de 3.6 integrantes por hogar, sin embargo para el año 2020 aunque el número de hogares había aumentado 3.9 % en comparación al año 2018. Los habitantes por hogar bajaron a 3.5 integrantes. En el año 2022, el promedio de personas por hogar aumentó un 5.1 % en comparación con 2020. A pesar de este crecimiento en el número de integrantes por hogar, la cantidad total de personas viviendo en hogares disminuyó un 3.2 %, lo que llevó a un promedio de 3.4 personas por hogar.

Aunque el número de hogares aumentó en México, sus integrantes han disminuido, dado que contar con más integrantes en los hogares se vuelve más caro para cubrir las necesidades de estos.

1.4.5 Transporte

El transporte es uno de los rubros que tiene más influencia en el gasto de los hogares mexicanos, dado a que es utilizado por la población todos los días para llevar a cabo sus actividades como lo son su trabajo, la escuela o simplemente el hecho de salir a comprar sus alimentos, necesitan pagar microbuses o en todo caso gasolina para sus automóviles, de acuerdo con Martínez (2019) se destinan “21 mil 816 pesos al transporte, ese monto es todavía mayor para la gente que vive en zonas remotas a los centros urbanos a donde deben trasladarse a diario para realizar sus actividades”, por ello en ocasiones necesitan tomar más de un trasporte y eso conlleva a gastar más en este rubro.

A través de la ENIGH obtenemos que ocupa el segundo lugar dentro de los gastos de los hogares mexicanos, dado a que para el año 2022 los hogares gastaron el 19.3 % del gasto corriente monetario total trimestral, lo cual es importante resaltar dado a que “conforme va aumentando el ingreso, es mayor lo que se dedica al transporte” (Alicia Gutiérrez, 2016).

Existen diferentes factores detrás de dicho incremento. Por ejemplo, los hogares con mayores ingresos invierten más dinero en el mantenimiento de sus automóviles. Además, los gastos generados son mayores debido al pago de gasolina y casetas.

1.4.6 Ingreso corriente total

Los ingresos en los hogares mexicanos son de principal importancia dado a que dependiendo al nivel de ingresos depende la calidad de vida y el poder adquisitivo con el que cuenta el hogar. De acuerdo con la pagina oficial del INEGI el ingreso corriente total es el dinero que se recibe durante el período de referencia como compensación por el trabajo asalariado en una empresa, institución o bajo la dirección de un empleador, incluyendo el ingreso en efectivo, además de que el ingreso es un elemento central para el estudio y evaluación de los has condiciones de vida de los hogares mexicanos.

En base a la ENIGH (2022) en los hogares mexicanos el ingreso corriente promedio trimestral por hogar fue de 63,695 pesos, el cual se conforma por el ingreso del trabajo el cual hace referencia al ingreso obtenido por trabajo subordinado, ingreso por renta de propiedad, transferencias el cual se refiere al efectivo que reciben los integrantes del hogar como pensiones, jubilaciones o becas, estimación del alquiler en la vivienda este ingreso hace referencia al valor estimado que tendría su casa en el mercado por la renta de su vivienda y otros ingresos

corrientes.

A través de los años el ingreso a tenido variaciones significativas en el año 2016 el ingreso corriente total fue de 63,565 pesos, sin embargo, para el año 2018 se presentó una disminución de 2,649 pesos, al igual que para el año de 2020 el ingreso en comparación con el ingreso de 2018 su disminución fue de 3,546 pesos esto debido a los gastos que ocasionó la pandemia, pero para el 2022 el ingreso tuvo un aumento de 6,325 pesos, sin embargo del año 2016 a 2022 el aumento porcentual del gasto corriente total fue de 0.2 %.

El ingreso es una variable fundamental para comprender y analizar los hábitos de gasto y el bienestar económico de los hogares, permitiendo un análisis integral y detallado de la situación económica. Al considerar el ingreso, se identifican las restricciones presupuestarias que enfrentan los hogares, así como sus capacidades para ahorrar e invertir.

1.4.7 Estrato socioeconómico.

El estrato socioeconómico de acuerdo con INEGI es una categorización que agrupa a la población en varios niveles según criterios como ingresos, nivel educativo, tipo de vivienda, servicios disponibles y posesión de bienes, en donde se asigna un número de acuerdo a dichas características.

De acuerdo con la ENIGH (2022) y para fines de la encuenta el estrato socio economico se define como la clasificación de las viviendas en el país según diversas características socioeconómicas de sus habitantes, así como sus características físicas y equipamiento, la clasificación es esencial para reconocer las disparidades en las condiciones de vida y el acceso a recursos

entre diferentes grupos sociales.

Es importante resaltar que Contreras (2023) nos dice que de acuerdo con los datos del INEGI el 56.6 % de la población mexicana esta en la clase baja, el 42.2 % en la clase media y solo el 1.2 % en la clase alta.

1.4.8 Educación del jefe del hogar

Como menciona el INEGI, la educación del jefe del hogar se refiere a la educación formal que posee dicha persona. Este factor es importante porque influye significativamente en los ingresos que se generan en los hogares, además de acuerdo con Marina Clemente, Gerónimo Antonio y Pérez Abarca (2017) el nivel educativo del jefe del hogar ayuda a incrementar la probabilidad de asistencia escolar de sus hijos, es decir, es parte clave para romper el círculo intergeneracional de la pobreza.

1.5 Conclusiones del capítulo

A lo largo de este capítulo, hemos analizado detalladamente la evolución del gasto en los hogares mexicanos, desglosando las principales variables que influyen en estos patrones. Desde la influencia de factores económicos como el ingreso disponible hasta el impacto de variables sociodemográficas como el tamaño del hogar y la urbanización, los datos presentados ofrecen una visión comprensiva de las dinámicas económicas en el contexto de los hogares en México.

Se ha observado que la distribución del gasto se ha mantenido constante en las categorías que componen al gasto. Sin embargo, existe una disminución en el gasto de los hogares en

alimentos, y se observó un aumento en el rubro de la salud en el año 2020 debido a la pandemia.

Además, se ha notado una tendencia de crecimiento en el gasto de los hogares a lo largo de las últimas décadas, influenciado por eventos económicos y políticos. Es importante destacar que el ingreso juega un papel crucial en el gasto de los hogares mexicanos, subrayando la necesidad de políticas económicas que fortalezcan el poder adquisitivo.

Conocer más sobre los patrones históricos y las variables asociadas es esencial para predecir futuras tendencias y diseñar políticas públicas eficaces. En el próximo capítulo, profundizaremos los conceptos detrás de los modelos econométricos necesarios para realizar un modelo econométrico que nos muestre si las variables mencionadas influyen en el gasto.

Capítulo 2: Marco metodológico

Este capítulo se enfoca en presentar y analizar el marco metodológico que respalda esta tesis. La metodología econométrica seleccionada proporciona una base sólida para comprender los fenómenos estudiados e interpretar los resultados obtenidos. A lo largo de este capítulo, se revisarán diversas técnicas econométricas y conceptos fundamentales.

Es relevante señalar que la elección de las técnicas de análisis de datos para estudiar las relaciones entre variables se basa en una revisión exhaustiva de la literatura existente y en su pertinencia para el ámbito de estudio. Además de contextualizar la investigación, se busca resaltar la importancia de un enfoque bien fundamentado para el éxito de este estudio académico.

2.1 ¿Qué es la econometría?

Según Elizalde Ángeles (2012), la econometría tiene sus raíces en los inicios de la década de 1930, cuando emergió con el propósito de evaluar los ciclos económicos. Este impulso se debió a la frecuente ocurrencia de fases recesivas que se habían observado desde finales del siglo XIX.

De acuerdo con Iglesias Ibarra y Fernández Rangel (2022) la econometría es la disciplina dentro de la economía que busca validar las teorías económicas utilizando métodos cuantitativos y estadísticos, con el objetivo de probar hipótesis para determinar la validez de los principios económicos, además de estimar y predecir diversos fenómenos económicos.

Por otro lado, para Spanos (1986 y1988) la econometría se define como el "estudio sistemático del fenómeno de interés usando los datos" (Citado por Herrera Yáñez, 2015) lo cual hace alusión a que se debe contar con observaciones de una población es decir encuestas, datos históricos, etc. Con la finalidad de que, con ayuda de estos, se pueda estudiar un tema de interés en específico en un punto del tiempo en específico con ayuda de una modelización econométrica. Es así que, la econometría se basa en el análisis cuantitativo de datos para comprender y explicar fenómenos económicos.

2.2 Modelo econométrico

Un modelo econométrico de acuerdo con Herrera Yáñez (2015) es el punto de partida para realizar un análisis econométrico lo cual sucede al identificar a la variable tanto endógena y las variables explicativas que influyen en el modelo. Además, se consideran los parámetros estructurales que se encuentran dentro de las variables, las ecuaciones y su formulación matemática, así como los datos estadísticos necesarios. El proceso de construcción de un modelo econométrico se divide en los siguientes pasos

1. Planteamiento de la hipótesis
2. Especificación del modelo
3. Adquisición de datos
4. Cálculo de parámetros del modelo
5. Prueba de hipótesis
6. Predicción

7. Aplicación del modelo para propósitos de control o política.

Es importante resaltar que, la creación de modelos econométricos no es un proceso lineal; a menudo, es necesario regresar pasos anteriores. Además, en algunos casos existen metodologías específicas para la creación de modelos, como en el caso de los modelos de series de tiempo.

2.3 Modelo de regresión lineal múltiple

Según Eva (2019), el propósito del modelo de regresión múltiple es describir el comportamiento de una variable dependiente utilizando los valores de un conjunto de variables explicativas. Por lo cual, H.Stock y Mark (2012) nos dicen que para el modelo se considera la posibilidad de explicar una variable y (variable endógena) utilizando diversas variables explicativas $X_{1i}, \cdots, X_{ki}$. De acuerdo con Meneses (2019) dado a que con este enfoque nos aproxima mejor a la realidad económica que se pretende modelar, dado que muchas relaciones económicas son de naturaleza multivariante, la expresión para representar al modelo es la siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i \tag{1}$$

Donde:

β_0: es el término que representa al intercepto.

β_k: es el término que representa el cambio que ocurre en y de un cambio en x_{ik}.

e_i: es el término del error.

2.3.1 Supuestos del modelo de regresión lineal múltiple

De acuerdo con Carter Hill(2017) el propósito de los supuestos es crear un conjunto de normas para poder calcular los parámetros desconocidos β_k, desarrollar las características del estimador para β_k y también validar las hipótesis relacionadas con los coeficientes desconocidos, los supuestos son los siguientes:

1.**Modelo económetrico**: Las observaciones en $(y_i, x_i) = (y_i, x_{i2}, x_{i3}, \cdots, x_{ik})$ satisfacen la relación poblacional siguiente:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \tag{2}$$

2.**Exogenedidad estricta**: La esperanza condicional del error aleatorio e_i considerando todas las observaciones de las variables explicativas $X = \{x_i, i = 1, 2, \cdots, N\}$ es cero

$$E(e_i|X) = 0 \tag{3}$$

Este supuesto implica que $E(e_i) = 0$ y la $\text{cov}(e_i, x_{jk}) = 0$ para $k = 1, 2, \cdots, k$.

Por lo que, cada error aleatorio sigue una distribución de probabilidad con una media de cero. Aunque algunos errores pueden ser positivos y otros negativos, su promedio será cero en un conjunto grande de observaciones. Además, las variables explicativas no están correlacionadas con el error; por lo tanto, conocer los valores de estas variables no ayuda a predecir el valor

del error. Es así que, otra implicación derivada del supuesto de exogeneidad estricta es que la forma de la función de regresión múltiple se define de la siguiente forma:

$$E(y_i|x) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \tag{4}$$

3.**Homocedasticidad condicional**:La varianza del término del error, condicionada a X es constante

$$\text{var}(e_i|x) = \sigma^2 \tag{5}$$

Esta suposición nos dice que $\text{var}(y_i|x) = \sigma^2$, por lo que la variabilidad de y_i alrededor de su función media condicional $E(y_i|X) = \beta_0 + \beta_1 x_i 1 + \beta_2 x_i 2 + \cdots + \beta_k x_i k$ no está influenciada por X. Además, no hay una probabilidad mayor o menor de que los errores sean más grandes para ciertos valores de las variables explicativas en comparación con otros. Estos errores se denominan homocedásticos.

4.**Erorres condicionales no correlacionados**: La covarianza entre los diferentes términos del error e_i y e_j condicionada a X, es cero

$$\text{cov}(e_i, e_j|x) = 0 \text{ para } i \neq j \tag{6}$$

Por lo cual, no existe correlación entre los errores para cada par de observaciones. La covarianza entre los errores aleatorios de dos observaciones distintas es cero para todos los valores de X. Es así que, no hay comovimiento entre los errores, lo que significa que el tama-

ño de un error en una observación no afecta la probabilidad de tamaño de un error en otra observación.

5.**Relación lineal no exacta entre variables explicativas**: No es posible expresar una de las variables explicativas como una función lineal exacta de las demás. Matemáticamente, escribimos esta suposición los únicos valores de $c_1, c_2, \cdots, c_k$ para los cuales

$$c_1 x_{i1} + c_2 x_{i2} + \cdots + c_k x_{ik=} 0 \text{ para todas las observaciones } i = 1, \cdots, N \tag{7}$$

6.**Normalidad en el error**: Condicionado a X, los errores se distribuyen normalmente

$$e_i | X \backsim N(0, \sigma^2) \tag{8}$$

Por lo que este supuesto implica que la distribución condicional de y también se distribuye normalmente, es decir:

$$y_i | X \backsim N(E(y_i|x), \sigma^2) \tag{9}$$

2.3.2 Estimación de minimos cuadrados ordinarios

Citando a H.Stock y Mark (2012) el estimador de Mínimos Cuadrados Ordinarios (MCO) selecciona los coeficientes de regresión de manera que la recta de regresión estimada sea lo más cercana posible a los datos observados. Esta proximidad se mide mediante la suma de los errores al cuadrado que se cometen al predecir Y dado X.

Por lo que, Elizalde Ángeles (2012) nos plantea que el problema de la predicción lineal se

reduce a ajustar una línea recta a un conjunto de puntos en un diagrama de dispersión. Este diagrama ayuda a determinar el tipo de curva que mejor se ajusta a los datos. Si la curva es una línea recta, se denomina recta de ajuste, la cual es una línea que minimiza la suma de las desviaciones de cada punto con respecto a dicha línea. Esta línea se conoce como la recta de mínimos cuadrados.

Para obtener las estimaciones mediante mínimos cuadrados ordinarios, de acuerdo con Wooldrigdge (2010) inicialmente se considera la estimación del modelo con dos variables independientes, es así que la ecuación estimada mediante MCO se formula de manera similar a la de la regresión simple, por lo que tenemos:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{10}$$

Donde:

$\hat{\beta}_0$: es la estimación de β_0

$\hat{\beta}_1$: es la estimación de β_1

$\hat{\beta}_2$: es la estimación de β_2

Las obtención de $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ es así que al utilizar el método de mínimos cuadrados ordinarios selecciona las estimaciones que minimizan la suma de los residuos al cuadrado. En otras palabras, al tener n observaciones sobre y, x_1 y x_2 para que las estimaciones sean elegidas de manera simultánea, con el propósito de que la siguiente expresión sea lo mas pequeña posible.

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \tag{11}$$

Es importante destacar que la $Ec.11$ el uso del subíndice i, cual representa al número de observación, de tal forma que la suma de dicha ecuación corresponda a todas las observaciones de 1 hasta n, mientras que con el otro subíndice sirve para distinguir a las variables independientes. De manera general, al tener el caso de k variables independientes, se tendrán que calcular las estimaciones $\hat{\beta}_0$,$\hat{\beta}_1$,$\cdots$,$\hat{\beta}_k$ en la siguiente expresión:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{12}$$

La ecuación se le conoce como línea de regresión de MCO o función de regresión muestral, ademas de que a $\hat{\beta}_0$ es la estimación del intercepto de MCO y las $\hat{\beta}_1$,$\cdots$, $\hat{\beta}_k$ son las estimaciones de las pendientes de MCO.

Es asi que, las $k+1$ estimaciones de los MCO serán elegidas respecto a que minimicen la suma la suma de los errores al cuadrado de las n observaciones:

$$\sum_{i=1}^{n}(y_i + \hat{B}_0 + \hat{B}_1 x_{i1} + \cdots + \hat{B}_k x_{ik})^2 \tag{13}$$

Para realizar la minimización se utiliza el cálculo multivariante, lo que conlleva a $k+1$ ecuaciones lineales en $k+1$ incógnitas $\hat{\beta}_0$,$\hat{\beta}_1$,$\cdots$,$\hat{\beta}_k$, por lo que tenemos lo siguiente:

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik})^2 = 0 \\
\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 = 0 \\
\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 = 0 \\
\vdots \\
\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 = 0
\end{aligned}
\tag{14}
$$

Las ecuaciones anteriores se llaman condiciones de primer orden de MCO, las cuales se obtienen con el método de momentos y son las contrapartes muestrales de los momentos poblaciones, aunque no se toma en cuenta a la división entre el tamaño de la muestra.

Para resolverlas, es necesario usar un software, ya que hacer el proceso manualmente es muy complicado, aunque tengamos el caso cuando n y k tienen un tamaño considerable. Además, el software permite resolverlas mucho más rápido, independientemente de que n y k sean muy grandes.

2.4 Criterios de selección

2.4.1 R^2

R-cuadrado o también conocida como coeficiente de determinación representa "la proporción de la varianza total de la variable explicada por la regresión" (Sanjuán, 2022) , es decir, nos muestra que tan cerca están los datos de la línea de regresión ajustada, de tal manera que refleja la bondad del ajuste de un modelo econométrico a la variable que se pretende explicar.

De acuerdo con H.Stock y Mark (2012) matemáticamente R^2 se puede calcular dividiendo la suma explicada de los cuadrados (SE) entre la sumatoria total de los cuadrados (ST). La suma explicada representa la suma de las desviaciones al cuadrado de los valores predichos de Y_i, $\hat{Y}_i$, en relación a su media y la suma total (ST) representa la suma de los cuadrados de las desviaciones entre Y_i y su media:

$$SE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{15}$$

$$ST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{16}$$

Por lo que R^2 se calcula de la siguiente forma:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \tag{17}$$

Sanjuán(2022) nos dice que es importante tener en cuenta que el coeficiente determinación se encuentra dentro de un rango de 0 a 1, en donde un valor cercano a cero nos indica que el modelo será menos ajustado y por ello menos fiable será, por otro lado, si es coeficiente es más cercano a uno el ajuste el modelo será mayor y por ende será más confiable.

2.4.2 R^2 ajustada

La R-cuadrada ajustada o también conocida como coeficiente de determinación ajustado, de acuerdo con Sanjuan (2022) es utilizado para observar la efectividad que tiene las variables

independientes en explicar a la variable dependiente de un modelo econométrico, además no toma en cuenta el impacto que tiene todas las variables independientes, sino solo las que impactan la variación de la variable dependiente del modelo, es decir penaliza la inclusión de variables independientes.

H.Stock & Mark (2012) nos indican que el $\bar{R}^2$ es una adaptación del R^2 estándar que no necesariamente aumenta con la adición de un nuevo regresor, por lo que se calcula con la siguiente expresión:

$$\bar{R^2} = 1 - \frac{n-1}{n-k-1} \times \frac{SR}{ST} = 1 - \frac{s^2_{\hat{u}}}{s^2_Y} \tag{18}$$

En donde n el número de observaciones de la muestra y k el de variables independientes.

Además de que nos plantean que es importante saber tres cosas sobre el r cuadrado ajustado, el primero es que la expresión de $(n-1)/(n-k-1)$ será siempre mayor a 1, por lo cual r cuadrado será siempre mayor a r cuadrado ajustado.

Por otro lado, la inclusión de un nuevo regresor tiene un doble efecto sobre $\bar{R}^2$, el primero es que SR disminuye lo que conlleva a un aumento del $\bar{R}^2$ y por otro lado, el factor $(n-1)/(n-k-1)$ aumenta, cabe resaltar que el hecho de que aumenta o disminuya dependerá de cuál de los dos efectos sea más dominante. Por último, $\bar{R}^2$ puede ser negativo lo cual sucede cuando todos los regresores se analizan en conjunto, por lo que, disminuyen la suma de los cuadrados de los residuos en una cantidad tan insignificante que esta reducción no puede compensar el efecto del factor $(n-1)/(n-k-1)$.

2.4.3 Raíz del error cuadrático medio

La raíz del error cuadrático medio o RMSE (Root Mean Square Error), representa la desviación estándar de los residuales, por lo que, de acuerdo con la nota técnica preparada por Asturias Corporación Universitaria (2015) mide el nivel de dispersión de los datos observados respecto a la línea de regresión, es decir, la cantidad de error que hay entre el valor predicho y el valor conocido, con la finalidad de indicarnos la concentración existente de los datos en la línea de mejor ajuste.

La expresión para calcular el valor de RMSE es el siguiente:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{)} \; n} \tag{19}$$

Donde $\hat{y}_1$,$\hat{y}_2$,$\hat{y}_3$,$\cdots$,$\hat{y}_n$ son los valores predichos, mientras que $y_1, y_2, y_3, \cdots, y_n$ son los valores conocidos y n es el número de observaciones.

En resumen, el RMSE representa el tamaño promedio de los residuos y se emplea como una estimación de la desviación estándar $SD(u_i)$, dado que el valor real de la desviación estándar es desconocido.

Teniendo en cuenta a Cloud (2024) es importante resaltar que mientras más bajo sea el valor de RMSE nos representa un mejor ajuste del modelo, dado a que se expresa la precisión del modelo.

2.4.4 Criterio de información Akaike

Como señala Zajic (2022) el criterio de información Akaike realiza una comparación entre modelos para obtener al mejor modelo dentro del grupo, es importante resaltar que, aunque el AIC elegirá el mejor modelo dentro de un conjunto, sin embargo, eso no quiere decir que sea el mejor, sino más bien solo será el mejor de ellos, es así que si tenemos un conjunto de modelos malos se escoge el mejor dentro de ellos, y para comprobar que sea el mejor se necesita realizar una prueba de hipótesis.

Por lo cual, la expresión para calcular el AIC es la siguiente:

$$\text{AIC} = -2\ln(L) + 2k \tag{20}$$

Donde k es el número de variables en el modelo más intercepto y ln(L) es el logaritmo de a máxima verosimilitud, la cual si es más grande el ajuste será mejor.

Por otro lado, de acuerdo con R. Carter Hill (2017) es importante resaltar que el AIC pone una penalización mayor que $\bar{R}^2$ sobre la inclusión de variables, además de que al comprar los modelos se escogerá el que tenga el valor más pequeño del AIC.

2.4.5 Criterio de información bayesiano

Teniendo en cuenta a Mohamad (2016) el criterio de información bayesiano, también conocido como criterio de Schwarz, es una medida de la calidad del ajuste de un modelo. Se utiliza como un estándar para elegir el mejor modelo entre un conjunto limitado de modelos disponibles.

Es asi que, la expresión para calcular el BCI es la siguiente:

$$\text{BCI} = k \times \ln(n) - 2\ln(L) \tag{21}$$

Donde k es el número de parámetros del modelo y ln(L) es la función de log-verosimilitud del modelo estadístico.

Se centra en la función de probabilidad logarítmica, además de estar relacionado con el criterio de información Akaike, por lo que de igual forma introduce un término de penalización por el número de parámetros presentes en el modelo, pero la penalización es mayor en comparación con el AIC.

Es importante tener en cuenta que tener un valor más bajo en BIC, dado a que nos indica que se tiene un mejor ajuste o se tiene un número menor de variables explicativas.

2.5 Normalidad: Prueba Jarque Bera

Es importante recordar que uno de los supuestos que debe cumplir un modelo de regresión lineal múltiple es la normalidad de los residuos, es decir:

$$e_i \mid X \sim N(0, \sigma^2) \tag{22}$$

De acuerdo con Vela Peón (2010) el cumplir con el supuesto de normalidad proporciona la base teórica necesaria para utilizar pruebas estadísticas que involucran las distribuciones t, f y chi-cuadrada, las cuales son ampliamente utilizadas en la parte inferencial de los modelos

estadísticos.

Por lo que una prueba para comprobar si nuestros errores cumplen con dicho supuesto es la prueba de Jarque Bera, de acuerdo con la explicación de Quispe Llanos la prueba de Jarque Bera nos permite conocer si los errores cumplen con el supuesto de normalidad, por lo que se realiza una prueba de hipótesis, donde:

H_0: Los datos siguen una distribución normal

H_1: Los datos no siguen una distribución normal

En donde el estadístico de prueba de acuerdo con Gonzáles Borja y Nieto Sánchez (2018) se calcula con la siguiente expresión:

$$JB = n\left(\frac{SK^2}{6} + \frac{(KU-3)^2}{24}\right) \tag{23}$$

Donde:

n es el número de observaciones.

SK es una medida de asimetría que se calcula con la siguiente expresión.

$$SK = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2)^{3/2}}$$

KU es una medida de curtosis, la cual se calcula con la siguiente expresión

$$KU = \frac{\hat{\mu}_4}{\hat{\mu}_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2)^2}$$

Por lo tanto, si se cumple la hipótesis nula, nos indicará que los datos siguen una distribución normal y su histograma formará una campana simétrica. Cabe resaltar que la probabilidad de rechazar la hipótesis nula suele establecerse en un nivel de significancia del 5 %.

2.6 Heterocedasticidad

Es importante tener en cuenta el supuesto sobre homocedasticidad mencionado con anterioridad, de acuerdo con H. Stock y M (2012) la homocedasticidad se presenta cuando el término del error u_i donde la varianza de la distribución condicional de u_i dado X_i es constante para cada $i = 1, ..., n$ y en particular no depende de X_i, es decir, todos los términos de los errores se distribuyen uniformemente alrededor de la recta de regresión, manteniendo la misma varianza.

Además, como menciona Elizalde Ángeles (2012) es importante que resaltemos que la homocedasticidad es necesaria para justificar el uso de las pruebas t y F, así como los intervalos de confianza para las estimaciones por mínimos cuadrados ordinarios (MCO) en el modelo de regresión lineal.

Sin embargo, cuando la varianza no es constante en los errores de un modelo de regresión lineal se conoce como heterocedasticidad, es decir, la dispersión en los errores es diferente para los diferentes valores de la variable dependiente.

Algunas de las causas de la heterocedasticidad son las siguientes:

- Cuando se omite una variable en la especificación del modelo, su efecto se refleja parcialmente en los errores aleatorios. Antes de analizar la heterocedasticidad, es crucial

asegurarse de que todas las variables relevantes están incluidas en el modelo, ya que la omisión de una variable importante puede resultar en estimadores sesgados y varianzas ineficientes

- Los datos disponibles se organizan por agentes o unidades económicas (datos de panel o longitudinales), lo que implica la existencia de dos dimensiones en los datos. Esto significa que los datos pueden ser agrupados, agregados o promediados en torno a un conjunto de individuos, empresas o sectores, presentando características individuales y variabilidad diferente. Por lo tanto, es posible visualizar un dispersión distinta en los datos.

- La forma funcional del modelo sea incorrecta, por ejemplo, utilizar una función lineal en lugar de una logarítmica hace que la calidad de la regresión varíe según los valores de la variable exógena. Es decir, se ajustará bien a los valores pequeños, pero mal a los valores grandes, resultando en errores mayores y más dispersos en las zonas de peor ajuste.

- La distribución de las variables independientes, lo cual se da cuando los datos proporcionados están alejados de la media ya sea a la derecha o izquierda, por lo que nos lleva a una transformación en las variables y así corregirlo, con la finalidad de que se encuentren de manera uniforme alrededor de la media.

2.6.1 Estimador de mínimos cuadrados ponderados

Los mínimos cuadrados ponderados de acuerdo con Faster Capital (2024) son una técnica que trata de corregir a la heterocedasticidad signando pesos más altos en las observaciones que tienen menor varianza y pesos más bajos a las observaciones que tienen mayor varianza. Citando a H. Stock y Mark (2012) el método conocido como mínimos cuadrados ponderados (MCP), pondera la i-ésima observación con la inversa de la raíz cuadrada de la varianza condicional de u_i dado X_i. Gracias a esta ponderación, los errores de la regresión ponderada son homocedásticos, lo que hace que los estimadores de mínimos cuadrados ordinarios (MCO) aplicados a los datos ponderados sean insesgados y eficientes.

Por otro lado, R. Carter Hill (2017) nos plantea que una de las formas para observar al estimador GLS es como un estimador de mínimos cuadrados ponderados, por lo que para ello debemos tener en cuenta que las estimaciones de MCO son los valores de β_1 y β_2 los cuales minimizan a la suma de los errores al cuadrado, por lo que tenemos:

$$S(\beta_1\beta_2 \mid y_i, x_i) = \sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_{i2})^2 \tag{24}$$

Tomando en cuenta la expresión

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^* \tag{25}$$

tenemos que la función de suma de cuadrados es la siguiente:

$$
\begin{aligned}
S(\beta_1\beta_2 \mid y_i, x_i) &= \sum_{i=1}^{N}(y_i^* - \beta_1 x_{i1}^* - \beta_2 x_{i2}^*)^2 \\
&= \sum_{i=1}^{N}\left(\frac{y_i}{\sqrt{x_i}} - \beta_1\frac{1}{\sqrt{x_i}} - \beta_2\frac{x_{i2}}{\sqrt{x_i}}\right)^2 \\
&= \sum_{i=1}^{N}\left[\frac{1}{\sqrt{x_i}}(y_i - \beta_1 - \beta_2 x_{i2})\right]^2 \\
&= \sum_{i=1}^{N}\frac{(y_i - \beta_1 - \beta_2 x_{i2})^2}{x_i}
\end{aligned}
\tag{26}
$$

Los errores al cuadrado se ponderan por $1/x$. Según nuestro supuesto, la varianza es var($e_i \mid x_i$) = $\sigma^2 x_i$. Cuando x_i es menor, asumimos que la varianza del error es menor y los datos están más cerca de la función de regresión. Estos datos proporcionan más información sobre la ubicación de $E(y_i \mid x_i) = \beta_1 + \beta_2 x_i$.

Cuando x_i es mayor, asumimos que la varianza del error es mayor y los datos pueden desviarse más de la función de regresión. Éstos son menos informativos sobre la ubicación de $E(y_i \mid x_i) = \beta_1 + \beta_2 x_i$. Intuitivamente, tiene sentido "reducir el peso" de las observaciones con menos información y dar más peso a las observaciones con más información. Esto es precisamente lo que logra la función de suma ponderada de cuadrados. Cuando x_i es pequeño, los datos contienen más información sobre la función de regresión y las observaciones reciben una ponderación alta. Cuando x_i es grande, los datos contienen menos información y las observaciones reciben una ponderación baja. De esta manera, aprovechamos la heterocedasticidad para mejorar la estimación de parámetros.

2.6.2 Prueba Breusch Pagan

Citando a Totumat (2021) la prueba de Breusch-Pagan se basa en la premisa de que la media de los residuos es igual a cero y que la varianza es independiente de la variable independiente. En tal caso, la varianza puede estimarse utilizando el promedio de los cuadrados de los residuos.

Por lo cual, Gujarati y Poter (2010), nos dicen que para ilustrar dicha prueba se debe considera un modelo de regresión lineal con k variables, como el que vemos a continuación:

$$Y_i = \beta_0 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{27}$$

Debemos suponer que la varianza del error es la siguiente:

$$\sigma_i^2 = f(\alpha_0 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}) \tag{28}$$

En otras palabras, σ_i^2 es una función de las variables Z no estocásticas; algunas o todas las variables X pueden actuar como Z. Específicamente, supongamos que $\sigma_i^2 = \alpha_0 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}$ Es decir si tenemos que σ_i es una función lineal de las Z y si $\sigma_1 = \sigma_2 = \cdots = \sigma_m$, $\sigma_i^2 = \sigma 0$ es constante. Se tendría que probar que σ_i^2 es homoscedástica„ por lo que la prueba de hipótesis sería la siguiente:

$$H_0 : \sigma_1 = \cdots = \sigma_m \; vs \; H_1 : \sigma_1 \neq \cdots \neq \sigma_m$$

Es así que, Totumat (2021) dice que la hipótesis nula plantea la existencia de homocedasticidad, mientras que la hipótesis alternativa indica la presencia de heterocedasticidad. Esta prueba utiliza una distribución chi-cuadrado: es decir el estadístico de prueba sigue una distribución χ^2 con k grados de libertad. Si el p-valor del estadístico de prueba es menor que un límite apropiado, se rechaza la hipótesis nula de homocedasticidad y se concluye que hay heterocedasticidad.

2.7 Correlación

Citando a Vinuesa (2016) la correlación es una medida de la relación lineal entre dos variables cuantitativas continuas (x, y). La forma más simple de verificar si dos variables están correlacionadas es observar si varían conjuntamente. Es crucial destacar que esta covariación no implica necesariamente causalidad; la correlación puede ser unánime. Por otro lado, de acuerdo con H. Stock y Mark (2012) la correlación es una medida alternativa de la dependencia entre X e Y que elimina el problema de las "unidades" presente en la covarianza. Específicamente, la correlación entre X e Y se calcula dividiendo la covarianza entre X e Y por sus respectivas desviaciones estándar. Por lo cual se calcula de la siguiente manera:

$$\text{Corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{29}$$

Vinuesa (2016) nos dice que esta medida de correlación, representada por r, puede variar entre -1 y +1, donde ambos extremos indican correlaciones perfectas, negativa y positiva respectivamente. Un valor de r igual a 0 señala que no hay relación lineal entre las dos variables.

Una correlación positiva muestra que ambas variables varían en la misma dirección, mientras que una correlación negativa indica que varían en direcciones opuestas. Lo interesante del índice de correlación es que r también sirve como una medida del tamaño del efecto, como señala Amat Rodrigo (2016) nos dice que los tamaños de efecto son los siguientes:

- 0 es una asociación nula
- 0.1 una asociación pequeña
- 0.3 una asociación mediana
- 0.5 una asociación moderada
- 0.7 una asociación alta
- 0.9 una asociación muy alta

Cabe resaltar que el tener variables que tiene una correlación muy alta podría causar un problema de colinealidad dado a que nos indicaría una relación lineal positiva perfecta.

2.8 Prueba de error de especificación de la regresión (RESET)

Citando a Sosa Pérez (2021) para conocer los errores de especificación tenemos a la prueba RESET (Regression Equation Specification Error Test), realizada por Ramsey (1969) la cual nos ayuda a detectar errores en la especificación del modelo los cuales son ocasionados por la omisión de variables independientes o una forma funcional incorrecta del modelo econométrico. De acuerdo con R. Carter Hill (2017), para trabajar la prueba RESET primero debemos especificar y estimar un modelo de regresión, como el siguiente:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 X_3 + e \tag{30}$$

Donde (b_1, b_2, b_3) son los mínimos cuadrados estimados y sea

$$\hat{y} = b_1 + b_2 x_2 + b_3 X_3 + e \tag{31}$$

Tengamos en cuenta al siguiente modelo especificado:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 X_3 + \gamma_1 \hat{y}^2 + ey = \beta_1 + \beta_2 x_2 + \beta_3 X_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^2 + e \tag{32}$$

La prueba de hipótesis a probar es la siguiente:

$$H_0 : \gamma_1 = 0, \gamma_2 = 0 \; vs \; H_1 : \gamma_1 \neq 0, \gamma_2 \neq 0$$

La hipótesis nula plantea que el modelo tiene una forma funcional correcta, mientras que la hipótesis alternativa sugiere que el modelo no tiene una forma funcional correcta. Según Sosa Pérez (2021), esta prueba sigue una distribución F y se lleva a cabo una vez que el modelo ha sido estimado. Es importante destacar que, si se rechaza la hipótesis nula, esto implica que el modelo no tiene una forma funcional correcta.

2.9 Modelos probabilísticos

Citando a Ross (2014) un modelo probabilístico es un tipo de modelo matemático que utiliza la teoría de la probabilidad para representar y analizar fenómenos inciertos o aleatorios. Estos modelos describen la relación entre variables observadas y eventos aleatorios mediante distribuciones de probabilidad y son fundamentales en el análisis estadístico y la inferencia.

De acuerdo con Devore (2011) los modelos probabilísticos tienen la capacidad de gestionar y representar la variabilidad intrínseca de los datos, lo que permite realizar análisis más sólidos y realistas de los fenómenos estudiados. Estos modelos son ampliamente aplicados en disciplinas como la estadística, la economía, la ingeniería y las ciencias sociales, debido a su eficacia para captar la aleatoriedad presente en los datos y ofrecer una base firme para la inferencia estadística.

Según Greene (2018), los modelos Logit y Probit son ejemplos específicos de modelos probabilísticos utilizados principalmente en el análisis de regresión cuando la variable de interés es binaria. Estos modelos facilitan la estimación de la probabilidad de que ocurra un evento particular, como éxito o fracaso, basándose en un conjunto de variables explicativas.

2.9.1 Modelo Probit

De acuerdo con Westreicher y Coll Morales (2021), el modelo Probit es una clase de modelo econométrico utilizado para decisiones binarias, lo que significa que se usa para situaciones donde hay que elegir entre dos opciones. Su particularidad radica en que se fundamenta en la distribución acumulada normal estándar.

Conforme a R. Carter Hill (2017) el modelo probit expresa la probabilidad de $p(x_i)$ de que se elija una alternativa $y_i = 1$, por lo que tenemos:

$$\begin{aligned} P(y_i = 1 \mid x_i) &= P\left[Z \leq \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}\right] \\ &= \Phi(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}) \end{aligned} \tag{33}$$

En donde $\Phi(z)$ representa a la función de distribución acumulativa de una normal estándar, el modelo Probit se considera no lineal, porque la ecuación anterior es una función no lineal de parámetros $\beta_1 + \cdots + \beta_K$. Si se conociera el valor de dichos parámetros, se utilizaría la ecuación para determinar la probabilidad de que se elija la alternativa uno para cualquier conjunto de valores predictores $x_i = (x_{i1} = 1, x_{i2}, ..., xiK)$.

Es importante destacar que interpretar el modelo Probit requiere cierto esfuerzo, dado a que la manera de evaluar el impacto de cualquier variable x_{ik} varía según sea continua o discreta, como una variable indicadora. Si una variable explicativa es continua, podemos analizar el efecto marginal de un cambio en su valor sobre la probabilidad $p(x_i)$. Si la variable explicativa es una variable indicadora, podemos calcular la diferencia en la probabilidad $p(xi)$ cuando x_ik toma los valores de 0 y 1. En ambos casos, debemos considerar que las magnitudes de los efectos dependen tanto de los valores de los parámetros, $\beta_1, \cdots, \beta_K$, como de los valores de las variables explicativas, $x_i = (x_{i1} = 1, x_{i2}, \cdots, x_{iK})$.

2.9.2 Modelo Logit

Según Westreicher y Coll Morales (2021) el modelo Logit es un modelo utilizado para analizar decisiones binarias, que se fundamenta en una función logística estándar acumulada.

De acuerdo con R. Carter Hill (2017) en un modelo Logit con una única variable explicativa x, entonces la probabilidad de $p(x)$ de que el valor observado y tome el valor de 1 se define como:

$$\begin{aligned} p(x) &= P\left[L \leq \gamma_1 + \gamma_2 x\right] \\ &= \Lambda(\gamma_1 + \gamma_2 x) \\ &= \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} \end{aligned} \tag{34}$$

La probabilidad de que $y = 0$ es:

$$1 - p(x) = \frac{1}{1 + e^{-(\gamma_1 + \gamma_2 x)}} \tag{35}$$

2.10 Muestras censuradas y regresión

Citando a Wooldridge (2010) las muestras censuradas se refieren a conjuntos de datos en los que las observaciones de la variable dependiente están incompletas debido a restricciones o truncamientos durante la medición.

De acuerdo con R. Carter Hill (2017), existen varias estrategias que se pueden utilizar, sin embargo, él solo nos presenta cuatro estrategias para manejar muestras censuradas. De estas

cuatro, dos no son adecuadas para tratar los datos censurados, mientras que las otras dos son las correctas para este tipo de análisis, las cuales son las siguientes:

- Estrategia 1: Eliminación de las observaciones límite y aplicación de OLS

Citando a R. Carter Hill (2017) una estrategia sencilla consiste en eliminar de la muestra las observaciones con $y_i = 0$ y continuar con el análisis. Sin embargo, esta estrategia no es efectiva. El modelo OLS estándar para $y_i > 0$ se expresa como $(y_i = \beta_1 + \beta_2 x_i + u_i)$, donde (u_i) es un término de error.

Sin embargo, u_i no es fácil de trabajar, dado a que se este se correlaciona con x_i, lo que implica que los mínimos cuadrados ordinarios estén sesgados e inconsistentes.

- Estrategia 2: Conservar todas las observaciones y aplicar mínimos cuadrados ordinarios.

Citando a Wooldridge (2010) utilizar mínimos cuadrados ordinarios (OLS) en muestras censuradas no es apropiado, ya que OLS asume que la variable dependiente se observa sin restricciones y que los errores tienen una distribución normal con media cero y varianza constante. En una muestra censurada, estas suposiciones no se cumplen, lo que produce estimaciones sesgadas e ineficientes. Específicamente, OLS no puede ajustar adecuadamente el modelo debido a la falta de información completa sobre las observaciones censuradas, lo que da como resultado coeficientes que no representan con precisión la relación entre las variables.

- Estrategia 3: Estimador de dos pasos de Heckman

Citando a Greene (2018) el estimador de dos pasos de Heckman es una técnica empleada para corregir el sesgo de selección en muestras censuradas. Este método trata los problemas que aparecen cuando la muestra disponible no representa adecuadamente a la población completa debido a un proceso de selección.

De acuerdo con Heckman (1979) explica que el procedimiento de Heckman se divide en dos etapas, las cuales son:

1. Primera etapa (Modelo Probit de selección): En esta parte se realiza la estimación de un modelo Probit para determinar la probabilidad de ser seleccionado. La ecuación que representa la selección es la siguiente:

$$s_i^* = \gamma_0 + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \cdots + \gamma_k z_{ki} + u_i \tag{36}$$

En donde s_i^* es una variable latente que indica la probabilidad de que una observación sea elegida, z_{1i},z_{2i},$\cdots$,z_{ki} son variables explicativas y u_i es el término de error. Se observa $s_i = 1$ si $s_i^* > 0$ y $s_i = 0$ en caso contrario. De esta estimación se obtiene el inverso de Mills, que se utiliza en la segunda etapa.

2. Segunda etapa (Modelo de regresión ajustado): En esta fase, se añade el inverso de Mills obtenido en la primera etapa como una variable adicional en el modelo de regresión, con la finalidad de corregir el sesgo de selección. La ecuación del modelo ajustado es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \lambda \hat{\rho}_i + e_i \tag{37}$$

Donde y_i es la variable dependiente, x_{1i},x_{2i},$\cdots$,x_{ki} son las variables independientes, $\widehat{\rho}_i$ es el inverso de Mills, y λ es el coeficiente que mide el impacto del sesgo de selección. Este método permite obtener estimaciones precisas y libres de sesgo de los parámetros del modelo que se está analizando.

- Estrategia 4: Estimación de máxima verosimilitud: Tobit

De acuerdo con R. Carter Hill (2017) el estimador de dos pasos de Heckman es consistente, sin embargo, no es eficiente por lo que es mejor utilizar un procedimiento de estimación de máxima verosimilitud, como lo es un modelo Tobit, el se explica en el apartado 2.10.2. de esta sección.

2.10.1 Ratio de Mills

Como señala Greene (2018) el ratio de Mills, también conocido como inverso de Mills o razón de Mills, es una herramienta importante en econometría para corregir el sesgo de selección en modelos de regresión, especialmente cuando se tienen muestras censuradas o truncadas. Esta función se define matemáticamente como la relación entre la función de densidad y la función de distribución acumulativa de una distribución normal estándar. Para una variable aleatoria Z distribuida normal estándar, el inverso de Mills $\lambda(Z)$ se calcula como:

$$\lambda(Z) = \frac{\phi(Z)}{\Phi(Z)} \tag{38}$$

Donde $\phi(Z)$ representa la función de densidad normal estándar y $\Phi(Z)$ es su función de

distribución acumulativa, que indica la probabilidad de que una variable normal estándar sea menor o igual a Z.

Este concepto es esencial en modelos como el Tobit y en la corrección de sesgo de selección de Heckman, donde el inverso de Mills se incorpora como una variable adicional en la segunda etapa del procedimiento de dos etapas. Después de estimar la probabilidad de selección en la primera etapa con un modelo Probit, se calcula el inverso de Mills para cada observación. Esta variable ajustada se utiliza luego en la segunda etapa de regresión para corregir el sesgo introducido por la selección de la muestra, permitiendo obtener estimaciones robustas y consistentes de los parámetros del modelo.

2.10.2 Modelo TOBIT

De acuerdo con Gujarati y Porter (2010) el modelo de regresión Tobit es un modelo, el cual fue desarrollado por James Tobin en el año de 1958, el cual es un modelo de regresión múltiple en donde la variable dependiente esta limita, es decir solo se representa a una parte de la población, por lo que se manejan dos tipos de datos, los cuales de acuerdo con Amat Rodrigo (2018), son los siguientes:

- Datos truncados: Las situaciones con datos truncados ocurren cuando, a partir de un cierto límite en una población, no existen observaciones. La diferencia fundamental respecto a los datos censurados es que, en estos últimos, las observaciones sí existen en la población latente, pero no se pueden captar en el muestreo.
- Datos censurados: se consideran censurados cuando la variable respuesta tiene un límite

(superior, inferior o ambos) a partir del cual todas las observaciones se asignan a un mismo valor. La característica principal de un escenario censurado es que existe una población subyacente con observaciones fuera de los límites de censura, pero debido a la incapacidad de detectarlas o seleccionarlas en el muestreo, parece que la población observada no las contiene. También nos resalta que, aunque la diferencia pueda parecer mínima, es crucial tenerla en cuenta porque, si el objetivo final de la inferencia es obtener información sobre la población real, en el caso de escenarios censurados, es necesario incluir de alguna manera esos eventos que existen, pero no se observan.

Por que cual el modelo de regresión Tobit también es conocido como modelo de regresión con variable dependiente limitada dado a que se tiene la restricción impuesta sobre los valores que toma la variable dependiente.

De acuerdo con Cartell R. Hill (2017) el modelo Tobit es un método de máxima verosimilitud que toma en cuenta la presencia de dos tipos de datos: observaciones limitadas (donde $y = 0$) y observaciones no limitadas (donde $y > 0$). Estos dos tipos de observaciones, las limitadas y las positivas, son el resultado de una variable latente y^* que supera o no supera un umbral de cero. La probabilidad (probit) de que $y = 0$ se basa en este cruce del umbral, lo cual se muestra a continuación:

$$P(y = 0|x) = P(y^* \leq 0|x) = 1 - \Phi[(\beta_1 + \beta_2 x)/\sigma] \tag{39}$$

En donde, P nos represeta la probabilidad de que se presenten las dos clases de observaciones que se identifican: las observaciones límite y las positivas, el resultado es una variable latente

y^* que supera o no supera el umbral cero.

Cuando y_i es positivo, el término que contribuye a la función de verosimilitud corresponde a la función de densidad de probabilidad (fdp) de una distribución normal con media $\beta_1 + \beta_2 x_i$ y varianza σ^2. La verosimilitud total se calcula como el producto de las probabilidades de las observaciones límite y las fdp de las observaciones positivas que no son límite, utilizando la notación de "pi grande" para representar la multiplicación. Como observamos a continuación:

$$\begin{aligned} L(\beta_1, \beta_2, \sigma \mid x, y) = & \prod_{y_i=0} \left\{ 1 - \Phi \left(\frac{\beta_1 + \beta_2 x_i}{\sigma} \right) \right\} \\ & \times \prod_{y_i>0} \left\{ (2\pi\sigma^2) \exp \left(-\frac{1}{2\sigma^2} (y_i - \beta_1 - \beta_2 x_i) \right) \right\} \end{aligned} \tag{40}$$

Esta compleja función de verosimilitud se maximiza numéricamente mediante el uso de software econométrico. El estimador de máxima verosimilitud resultante es consistente y sigue una distribución normal asintótica, con una matriz de covarianza que es conocida.

2.10.3 Efectos marginales

De acuerdo con Cartell R. Hill (2017) en el modelo Tobit, los coeficientes β_1 y β_2 representan la intersección y la pendiente respectivamente en el modelo de variable latente . En la práctica, nos interesa el efecto marginal de un cambio en x sobre la función de regresión de los datos observados $E(y \mid x)$, o sobre la función de regresión condicional $E(y \mid x, y > 0)$, estas funciones no son lineales y la pendiente de cada una varía con cada valor de x. La pendiente de $E(y \mid x)$ tiene una forma relativamente simple, siendo un factor de escala multiplicado por el valor del coeficiente, como se observa a continuación:

$$\frac{\partial E(y \mid x)}{\partial x} = \beta_2 \Phi \left(\frac{\beta_1 + \beta_2 x}{\sigma} \right) \tag{41}$$

La función Φ representa la función de distribución acumulativa (CDF) de una variable aleatoria normal estándar evaluada en las estimaciones y un valor particular x. Dado que los valores de la CDF son positivos, el signo del coeficiente indica la dirección del efecto marginal. Sin embargo, la magnitud del efecto marginal depende tanto del coeficiente como de la CDF. Cuando $\beta_2 > 0$, a medida que x aumenta, la función CDF se acerca a uno, lo que hace que la pendiente de la función de regresión también se aproxime a la de la variable latente.

Además, de que el efecto marginal se puede descomponer en dos factores llamados descomposición "McDonald-Moffitt", como se muestra a continuación:

$$\frac{\partial E(y \mid x)}{\partial x} = \text{Prob}(y > 0) \frac{\partial E(y \mid x, y > 0)}{\partial x} + E(y|x, y > 0) \frac{\partial \text{Prob}(y > 0)}{\partial x} \tag{42}$$

El primer factor describe cómo cambia la variable dependiente y para aquellos individuos cuyos datos ya se han observado, en respuesta a un cambio en x. El segundo factor explica cómo varía la proporción de la población que transita de la categoría de y no observada a la categoría de y observada cuando x experimenta cambios.

Conclusiones del capítulo

En este capítulo, hemos explorado las bases metodologicas de la econometría que sustentan nuestro análisis del gasto en los hogares mexicanos, centrándonos en dar a conocer el

modelo de regresión lineal múltiple y el modelo Tobit. Estos modelos nos proporcionan herramientas poderosas para analizar y entender las relaciones entre las variables económicas y sociodemográficas.

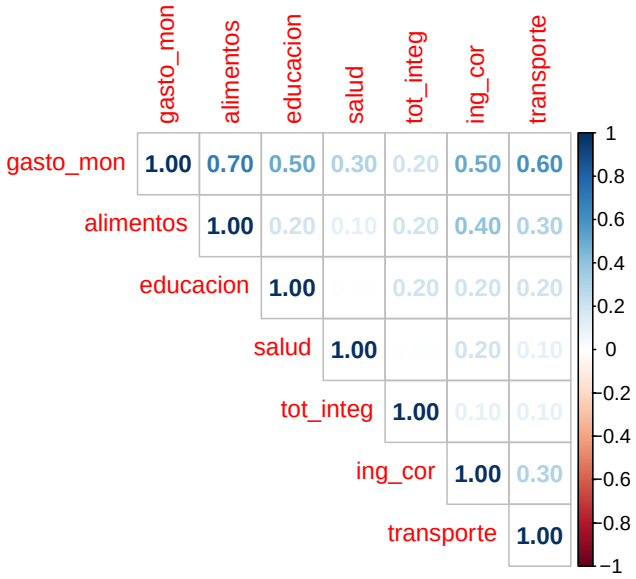
La comprensión y aplicación de estos modelos econométricos son esenciales para nuestro análisis empírico del gasto en los hogares mexicanos. El modelo de regresión lineal múltiple nos permitirá evaluar relaciones directas y cuantificar efectos, mientras que el modelo Tobit nos ayudará a manejar adecuadamente los datos censurados, proporcionando estimaciones más precisas y robustas.

Capítulo 3: Modelo

3.1 Modelo de regresión lineal múltiple

Inicialmente, para trabajar el modelo de regresión lineal múltiple, se realizó una matriz de correlación para identificar cuáles variables proporcionarían un buen modelo. Observamos que las variables que tienen una correlación que no es tan significativa, es decir, que no exista dependencia alta entre las variables con el gasto_mon, las cuales se observan a continuación:

Grafico 5: Matriz de correlaciones



Fuente: Elaboración propia en R studio con datos de la base 'concentradohogar'.

Se observa que las variables tienen las siguientes correlaciones:

- Educación con una correlación de 0.50

- Salud con una correlación de 0.30
- Tot_integ con una correlación de 0.20
- Transporte con una correlación de 0.60
- Alimentos con una correlación de 0.70, siendo esta la correlación más alta.

Es importante recordar que una correlación superior a 0.80 indica una dependencia alta y podría presentar multicolinealidad; sin embargo, este no es el caso en nuestro análisis.

El primer modelo propuesto se explica el gasto corriente monetario en función del gasto que hacen los hogares mexicanos en alimentos, educación, salud y transporte, además los integrantes que conforman el hogar y el ingreso, por lo que tenemos:

$$gasto\_mon = \beta_0 + \beta_1 alimentos + \beta_2 educacion + \beta_3 salud + \beta_5 tot\_integ + \beta_6 ing\_cor + \beta_7 trans \tag{43}$$

Se realizó la estimación de los coeficientes por el método de mínimos cuadrados ordinarios, los resultados de la regresión se muestran a continuación:

Tabla 2: Resumen del Modelo 1

term	estimate	std.error	statistic	p.value
(Intercept)	2887.5805462	97.6293659	29.57697	0
alimentos	1.5354825	0.0041912	366.36086	0
educacion	1.2132582	0.0058184	208.51987	0
salud	1.1830197	0.0074091	159.67043	0
tot_integ	-672.8197477	24.0934098	-27.92547	0
transporte	1.1934367	0.0037725	316.35122	0
ing_cor	0.0670954	0.0005896	113.79187	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

En la Tabla 2 se muestran los valor p obtenidos para cada variable y se demuestra que los coeficientes de cada variable son estadísticamente significativos, dado que tiene un valor p menor a 0.05, además de que sus estadísticos son los siguientes:

Tabla 3: Resultados del Modelo de Regresión 1

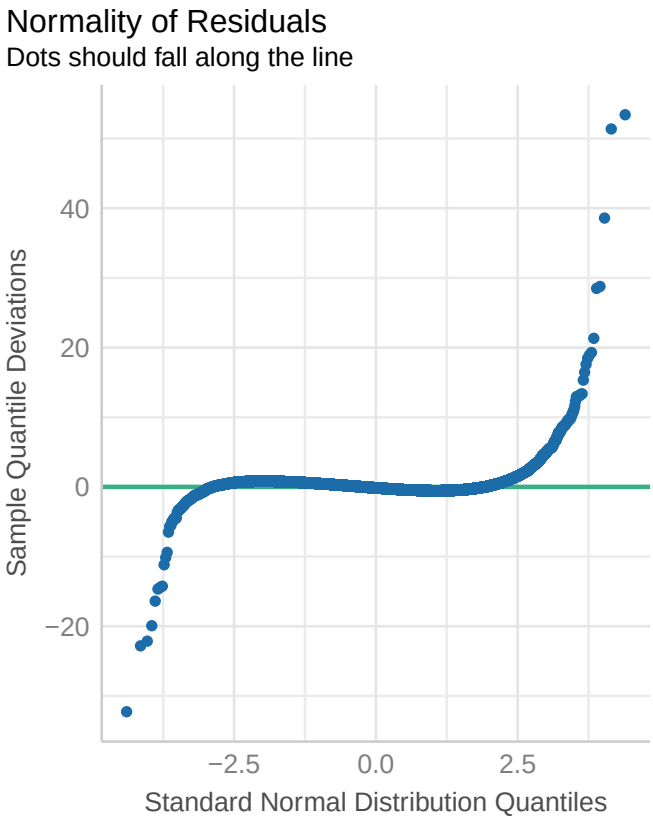
	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	12381.13	0.8723253	0.8723168	102594.3	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Como se observa el valor de $R^2 = 0.8723$ lo cual indica que el modelo de regresión lineal múltiple propuesto tiene un ajuste aceptable, explicando el $87.23\,\%$ de la variabilidad en el gasto. Otro punto por destacar es que el valor de F es grande, lo que indica que la varianza explicada por el modelo es mayor en comparación con la varianza no explicada. Es decir, al menos una de las variables independientes en el modelo tiene un efecto significativo en la variable dependiente. En conjunto con el valor p, que es menor a 2.2×10^{-16}, se concluye que el modelo en su totalidad es significativo.

Sin embargo, se debe verificar que el modelo cumple con las condiciones necesarias para ser considerado adecuado. El primer punto por verificar es si se cumple con la normalidad de los residuos, para ello se realizó una gráfica con la finalidad de observar el comportamiento de los datos, la cual se muestra a continuación:

Grafico 6: Normalidad de los errores del Modelo 1.

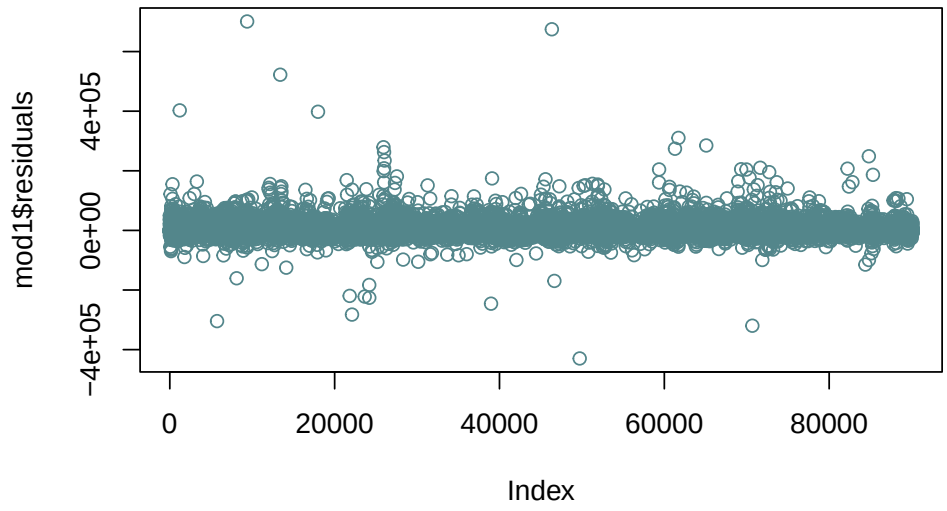


Fuente: Elaboración propia en R studio de la base de datos 'concentradohogar'.

Se realizó la prueba de Jarque-Bera, de la que se obtuvo un $valor - p < 2.2 \times 10^{-16}$, el cual es extremadamente pequeño. Por lo tanto, de ambos se concluye que los residuos del modelo no siguen una distribución normal.

Además, es importante comprobar si el modelo cumple con la homocedasticidad de los errores. Por lo cual, se graficaron los residuos de los datos para observar si presenta heteroscedasticidad, como se observa a continuación los datos no presentan un patrón, sin embargo, la mayoría de ellos se encuentran en el centro:

Grafico 7: Residuales del modelo 1



Fuente: Elaboración propia en R studio de la base de datos 'concentradohogar'.

Para saber si se presenta la heterocedasticidad, se realizó la prueba de Breusch-Pagan, la cual arrojó un $valor - p < 2.2 \times 10^{-16}$. Dado que este valor es muy pequeño y considerando un nivel de significancia del 5 %, se concluye que hay evidencia suficiente para afirmar que no se cumple la homocedasticidad de los errores. Por lo tanto, el modelo presenta heterocedasticidad.

Primero se corregirá, el problema de heterocedasticidad, para ello se aplicó al modelo, mínimo cuadrados ponderados siendo este un procedimiento donde se da más peso a las observaciones con menor varianza porque estas observaciones proporcionan información más confiable sobre la función de regresión que aquellas con grandes varianzas el cual se dio más peso a las observaciones con menor varianza en este caso tomaremos al tot_integ.

Sin embargo, al realizar este procedimiento, el problema de heterocedasticidad no se corrigió, dado a que al realizar la prueba de hipótesis seguimos obteniendo un valor p pequeño. Por

lo tanto, se cambió la forma funcional del modelo para abordar el problema, utilizando una forma cuadrática. Por lo que su forma funcional es de la siguiente forma:

$$\begin{aligned} gasto\_mon = \beta_0 + \beta_1 alimentos^2 + \beta_2 educacion^2 + \beta_3 salud^2 \\ + \beta_4 tot\_integ + \beta_5 transporte^2 + \beta_6 ing\_cor^2 + e \end{aligned} \tag{44}$$

Al estimarlo obtenemos los siguientes estimadores:

Tabla 4: Resumen del Modelo 3

term	estimate	std.error	statistic	p.value
(Intercept)	2.032627e+04	170.8388279	118.97920	0
I(alimentos^2)	3.100000e-06	0.0000000	118.15062	0
I(educacion^2)	6.200000e-06	0.0000001	70.94990	0
I(salud^2)	6.600000e-06	0.0000001	59.19058	0
tot_integ	4.321135e+03	58.3457225	74.06087	0
I(transporte^2)	4.500000e-06	0.0000000	125.71409	0
I(ing_cor^2)	0.000000e+00	0.0000000	26.66658	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Notamos que todas las variables vuelven a ser significativas, pero al realizar nuevamente las pruebas de hipótesis, el modelo no cumple con los supuestos de normalidad y homocedasticidad. (Véase en Anexo A)

3.2 Estudio de variables

Para conocer más acerca de las variables, realizamos unos modelos conformados por el gasto y cada una de ellas, con la finalidad de conocer como explican el gasto de manera individual, de lo que se observó que la variable que explica más al gasto son los alimentos dado a que su $R^2 = 0.5206$, siendo el R^2 más alto de todas las variables, el comportamiento de las demás variables se pueden observar en el Anexo A.

Además, se realizaron los histogramas de cada una de las variables para conocer el comportamiento de las mismas, de lo que se obtuvo que todas tiene un sesgo hacia la izquierda lo cual se debe a que en todas las categorías excepto *tot_integ*, tiene valores en cero. Esto se debe a que algunos hogares no declararon cuánto gastan en cada variable o, en otros casos, no tienen gasto en dicha variable,además de que algunos hogares no declararon su ingreso, y no se pueden eliminar dichos valores que tienen ceros, ya que dicha acción causaría un sesgo al momento de estimar el modelo, es importante resaltar que existe una gran variabilidad en los datos dado a que todas las variables excepto *tot_integ* tienen como mínimo 0 y como máximo a partir de 324547.7.

3.3 Modelo con deciles

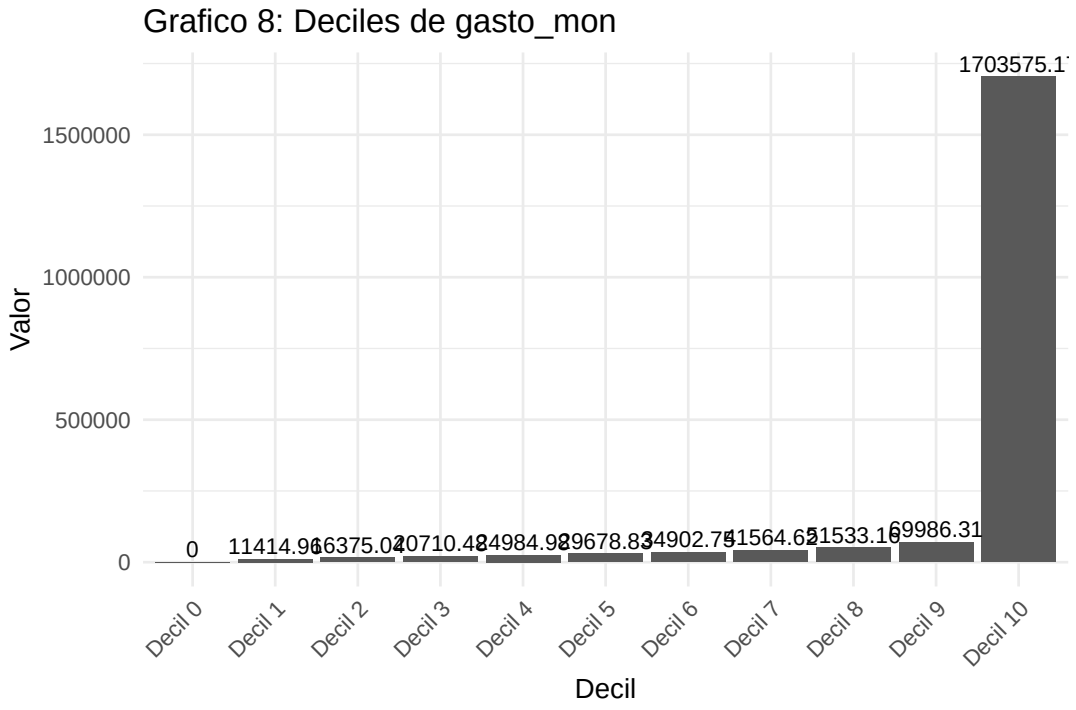
En el apartado anterior, se observó una gran variabilidad en los datos debido a factores económicos, ya que no toda la población tiene el mismo nivel de gasto e ingreso lo que causa heterocedasticidad. Por lo tanto, el gasto destinado a cada una de las variables es diferente en cada hogar. Para abordar este problema y analizar de manera más precisa la distribución del gasto, se realizaron cortes en la base de datos 'concentrado hogar' con el fin de seccionar la muestra en deciles de acuerdo con la variable *gasto_mon*. Trabajar con deciles permite dividir la población en diez grupos iguales, cada uno representando un 10 % del total, ordenados de menor a mayor gasto mensual. Esta metodología es útil, según Rus Arias, (2021) porque:

- Permite comparar el gasto en los diferentes segmentos de la población y observar las

diferencias entre los distintos niveles de ingreso.

- Facilita la evaluación de cómo se distribuye el gasto entre los deciles más bajos y más altos, permitiendo así evaluar la concentración del gasto y la equidad económica.
- Ayuda a identificar las necesidades específicas de los diferentes grupos, ya que los deciles más bajos pueden tener necesidades y prioridades diferentes en comparación con los deciles más altos.

A continuación se presenta un gráfico en donde se muestra los números en donde se inicia cada decil, para observar la variabilidad que existe, por lo que tenemos:



Fuente: Elaboración propia con R estudio con base de datos 'concentradohogar'.

Con el primer decil trabajamos con el modelo siguiente:

$$gasto\_mon = \beta_0 + \beta_1 \text{ alimentos} + \beta_2 \text{ educacion} + \beta_3 \text{ salud} + \beta_4 \text{ tot\_integ} + \beta_5 \text{ transporte} + \beta_6 \text{ ing\_cor} + e \tag{45}$$

Se realizó la estimación de los coeficientes por el método de mínimos cuadrados ordinarios, los resultados de la regresión se muestran a continuación:

$$gasto\_mon = 2516.81 + 0.9102 \text{ alimentos} + 0.8768 \text{ educacion} + 0.9037 \text{ salud} + 91.7666 \text{ tot\_integ} + 0.9840 \text{ transporte} + 0.0254 \text{ ing\_cor} + e \tag{46}$$

El modelo tiene un $R^2 = 0.6382$, lo cual es un valor aceptable. Sin embargo, al realizar las pruebas, se observó que dicho modelo no cumple con los supuestos de normalidad y homocedasticidad de los errores. A pesar de que los errores son más dispersos, siguen presentando estos problemas.

Por ello, se construyeron varios modelos con diferentes formas funcionales con la finalidad de encontrar al menos un modelo que explique adecuadamente el gasto corriente monetario. Es importante resaltar que no se consideró un modelo logarítmico debido a la presencia de ceros en los datos. Por esta razón, se realizaron combinaciones con variables y formas exponenciales, para después realizar una comparación entre los modelos y escoger el mejor. (Véase en Anexos A)

A continuación se presentaran los resultados del 'performance' para observar cual es el mejor modelo de los realizados:

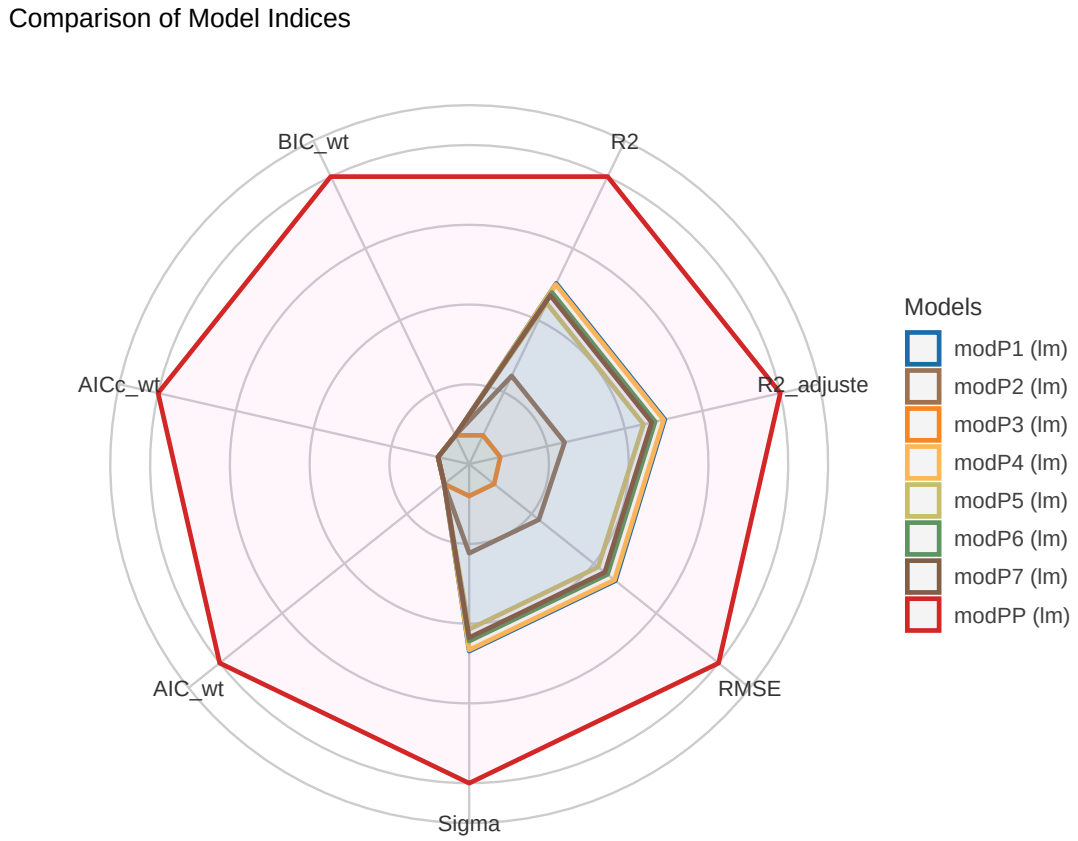
Tabla 5: Comparación de Modelos

Name	Model	R2	R2_adjusted	RMSE	Sigma	AIC_wt	AICc_wt	BIC_wt	Performance_Score
modPP	lm	0.64	0.64	1590.40	1591.03	1	1	1	1.00
modP1	lm	0.46	0.46	1934.92	1935.67	0	0	0	0.32
modP4	lm	0.46	0.46	1938.13	1938.78	0	0	0	0.32
modP6	lm	0.45	0.45	1961.24	1961.79	0	0	0	0.30
modP7	lm	0.44	0.44	1970.41	1970.97	0	0	0	0.30
modP5	lm	0.43	0.43	1992.33	1992.77	0	0	0	0.28
modP2	lm	0.31	0.31	2190.38	2191.24	0	0	0	0.12
modP3	lm	0.22	0.22	2339.69	2340.60	0	0	0	0.00

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Además se muestra un gráfico para tener mayor visualización de los criterios utilizados para encontrar al mejor modelo, por lo que tenemos:

Grafico 9 : Comparación de índices de los modelos.



Fuente: Elaboración propia en R studio con datos de la base 'concentradohogar'.

De ello se obtuvo que el mejor modelo es el que hemos planteado ($Ec.43$) para los deciles, sin embargo, al no cumplir con los supuestos de normalidad y homocedasticidad, se realizó una prueba RESET para comprobar si tenemos una forma funcional correcta en el modelo, de la cual se concluyó que el modelo está mal especificado para los datos, por lo que deberíamos buscar una forma funcional correcta.

Antes de ello, se comprobó si la forma funcional era adecuada para alguno de los nueve deciles, con la finalidad de observar si en alguno de ellos se corregían los errores que se presentaban. Sin embargo, no se corrigió el problema en ninguno de los nueve deciles, y para ninguno de ellos la forma funcional era correcta. (Véase en Anexos A)

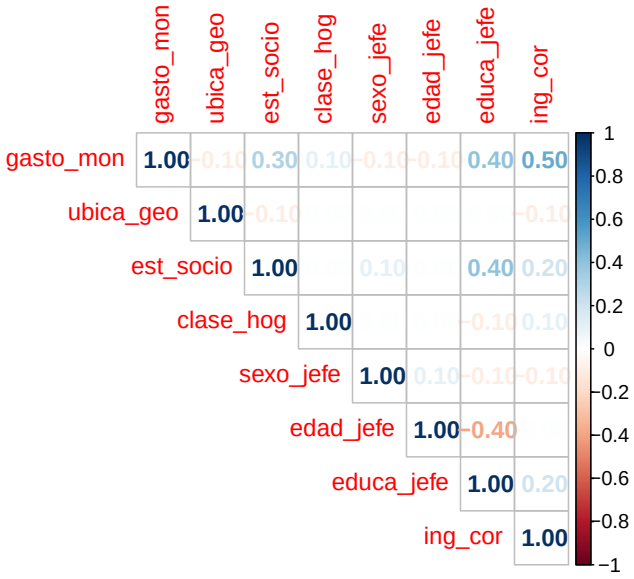
3.4 Reformulación

Como sabemos la construccion de un modelo econométrico efectivo es un proceso desafiante como lo hemos visto a lo largo de esta investigación, en donde se presentan obstáculos significativos. A pesar de los esfuerzos dedicados y metodológicamente sólidos para encontrar un modelo adecuado, los resultados pueden no ser satisfactorios. Por ello, el agregar más variables se convierte en una estrategia crucial para mejorar la capacidad predictiva y la explicación del fenómeno económico que se esta estudiado, en nuestro caso agregaremos a las siguientes variables $ubica\_geo$, est_socio,$clase\_hog$, $sexo\_jefe$, $edad\_jefe$,$educa\_jefe$,$comunica$ y $vivienda$.

Para empezar a trabajar se realizaron dos matrices de correlaciones con 14 variables, con la finalidad de observar la correlación existente con el $gasto\_mon$ y no omitir variables en un

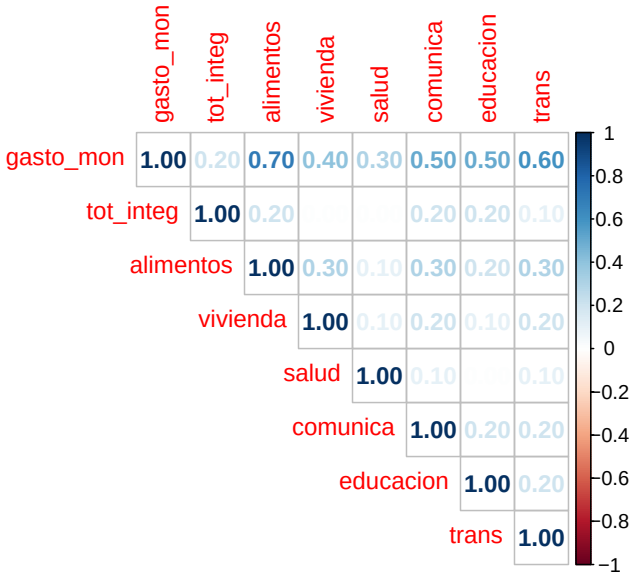
nuevo modelo, por lo que tenemos:

Grafico 10 : Matriz de correlaciones pt.1



Fuente: Elaboración propia en R studio con datos de la base 'concentradohogar'.

Grafico 11 : Matriz de correlaciones pt.2



Fuente: Elaboración propia en R studio con datos de la base 'concentradohogar'.

Al observar las correlaciones, las variables que seleccionadas fueron las siguientes: **gasto_mon**,**alimentos**,**educacion**,**salud**,**tot_integ**,**trans**,**vivienda**,**comunica**,**educa_jefe** y **est_socio**, con las cuales se creo el modelo siguiente:

$$\begin{aligned} gasto\_mon = \beta_0 &+ \beta_1 \text{alimentos} + \beta_2 \text{educacion} + \beta_3 \text{salud} \\ &+ \beta_4 \text{tot\_integ} + \beta_5 \text{trans} + \beta_6 \text{vivienda} \\ &+ \beta_7 \text{comunica} + \beta_8 \text{educa\_jefe} + \beta_9 \text{est\_socio} + \beta_9 \text{ing\_cor} + e_i \end{aligned} \tag{47}$$

Sus estimadores se presentan a continuación:

Tabla 6: Resumen del Modelo con todas las variables

term	estimate	std.error	statistic	p.value
(Intercept)	-1103.9108129	129.7726148	-8.506500	0e+00
est_socio	-238.8468809	47.0507977	-5.076362	4e-07
educa_jefe	301.7476705	14.9279243	20.213639	0e+00
tot_integ	-489.9312388	19.9809910	-24.519867	0e+00
alimentos	1.3518735	0.0035110	385.040055	0e+00
vivienda	1.1713882	0.0066459	176.256848	0e+00
salud	1.1561631	0.0059785	193.386756	0e+00
comunica	1.5861301	0.0164033	96.695903	0e+00
educacion	1.0907496	0.0047543	229.425590	0e+00
trans	1.1276003	0.0030730	366.939306	0e+00
ing_cor	0.0444704	0.0004923	90.330104	0e+00

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Tabla 7: Resultados del Modelo de Regresión con todas las variables

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	9987.488	0.9169237	0.9169145	99434.57	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

El modelo presenta un R cuadrada de 0.9169 lo cual puede indicar que existe colinealdad entre las variables, por ello primero observemos lo siguiente:

Tabla 8: Resultados de VIF

	x
est_socio	1.331171
educa_jefe	1.321332
tot_integ	1.140459
alimentos	1.349837
vivienda	1.176069
salud	1.035084
comunica	1.310478
educacion	1.134536
trans	1.169882
ing_cor	1.343052

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

De lo anterior, se observa que no se presenta colinealidad entre las variables por lo que, ahora veamos si presenta heterocedasticidad, por lo que tenemos:

Tabla 9: Resultados del Test de Breusch-Pagan

	Statistic	Degrees.of.Freedom	p.value	Method
BP	6456.143	10	0	studentized Breusch-Pagan test

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

De lo que se obtuvo, que el modelo presenta heterocedasticidad, para corregir dicho problema se realizaron pruebas con todas las variables para determinar una variable para corregir dicho problema, sin embargo, se obtenía el valor p de uno para las variables; **alimentos**,**educacion**,**salud**,**trans**,**vivienda**,**comunica** y **ing_cor**, mientras que para las variables; **tot_integ**,**educa_jefe** y **est_socio** se obtuvieron valor p muy pequeños, es así que se aplicarón los errores estándar robustos, sin embargo, al emplearlos, no se observó una reducción en los valores de los errores, lo cual puede ser indicativo de que estos no están mejorando la precisión del modelo de manera efectiva.(Vease e Anexos A)

3.5 Modelo de propensión marginal

Se construyó un modelo de propensión marginal al consumo, el cual mide la porción del ingreso destinada al consumo cuando la renta aumenta en una unidad. Este modelo tiene como finalidad ofrecer una interpretación más intuitiva y económica de los efectos de las variables explicativas sobre la variable dependiente en relación con el ingreso. La forma funcional del modelo es la siguiente:

$$\frac{\text{gasto\_mon}}{\text{ing\_cor}} = \beta_0 + \beta_1 \text{est\_socio} + \beta_2 \text{educa\_jefe} + \beta_3 \text{tot\_integ} + \beta_4 \frac{\text{alimentos}}{\text{ing\_cor}} + \beta_5 \frac{\text{vivienda}}{\text{ing\_cor}} + \beta_6 \frac{\text{salud}}{\text{ing\_cor}} + \beta_7 \frac{\text{comunica}}{\text{ing\_cor}} + \beta_8 \frac{\text{educacion}}{\text{ing\_cor}} + \beta_9 \frac{\text{transporte}}{\text{ing\_cor}} + e \tag{48}$$

La base de datos presenta valores cero, por lo que se estableció una condición: si había división entre cero o un cero dividido por el ingreso, se asignaba un cero. Esto permitió usar todas las observaciones y evitar omitir datos en la estimación.

Los valores estimados son los siguientes:

Tabla 10: Resumen del Modelo de propensión marginal

term	estimate	std.error	statistic	p.value
(Intercept)	0.0635493	0.0020013	31.75332	0
est_socio	-0.0088919	0.0006717	-13.23840	0
educa_jefe	0.0052596	0.0002118	24.83797	0
a	0.9920477	0.0031353	316.41484	0
b	1.6225603	0.0138800	116.89900	0
c	1.1324526	0.0028213	401.39161	0
d	1.1720471	0.0016668	703.17701	0
e	1.1011377	0.0050958	216.08660	0
f	1.0253525	0.0022120	463.53624	0
tot_integ	-0.0028864	0.0002842	-10.15770	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

En donde, a representa la división de vivienda entre ingreso, b los gastos del hogar entre ingreso, c los gastos de transporte entre ingreso, d los gastos de alimentos entre ingreso, e los gastos de educación entre ingreso, f los gastos de salud entre ingreso y g el gasto entre ingreso Y sus estadísticos son los siguientes:

Tabla 11: Resultados del Modelo de propensión marginal

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.1465197	0.9512451	0.9512402	195306.9	0

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Tenemos un modelo con una $R^2 = 0.9512$ lo que nos indica que aproximadamente el 95.12 % de la variabilidad en la variable dependiente es explicada por las variables independientes incluidad en el modelo. Además, un valor de F grande con un valor p asociado pequeño sugiere que el modelo en su conjunto es significativo.

Sin embargo, al presentar una R cuadrada grande, se puede presentar colinealidad, por lo que veamos;

Tabla 12: Resultados de VIF modelo de propensión marginal

	x
est_socio	1.260492
educa_jefe	1.235388
a	1.150783
b	1.173591
c	1.078841
d	1.215541
e	1.101634
f	1.009326
tot_integ	1.071734

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Con VIF tan bajos, podemos concluir que no existe una multicolinealidad significativa en las variables predictoras de este modelo. Esto es positivo, ya que asegura que las estimaciones

de los coeficientes de regresión no están infladas y son fiables. Por lo tanto, no se necesitan ajustes adicionales para abordar la multicolinealidad en este modelo. Sin embargo, es necesario verificar si el modelo presenta heterocedasticidad. Por lo tanto, tenemos:

Tabla 13: Resultados del Test de Breusch-Pagan modelo de propensión marginal

	Statistic	Degrees.of.Freedom	p.value	Method
BP	2741.585	9	0	studentized Breusch-Pagan test

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

El modelo presenta heterocedasticidad, y los intentos de corregirla no fueron exitosos debido a que las pruebas de hipótesis arrojaron valores p muy pequeños. Se intentó utilizar errores estándar robustos, pero esto resultó en un aumento incorrecto de los mismos. Además, el modelo enfrenta el problema de que sus errores no siguen una distribución normal (véase Anexo A).

3.6 Modelo TOBIT

El modelo TOBIT se presenta como una solución viable debido a las limitaciones encontradas en el modelo con deciles y el modelo con todas las variables. Como se mostró anteriormente, los datos exhiben un sesgo hacia la izquierda, lo que sugiere la presencia de valores truncados o censurados. Ante esta situación, el modelo TOBIT se convierte en una herramienta estadística adecuada para abordar tales casos.

Al considerar la naturaleza censurada de los datos, el modelo TOBIT puede ajustarse para manejar eficazmente esta limitación y proporcionar estimaciones precisas de los parámetros de interés. Además, al tener en cuenta tanto las observaciones registradas como las no registradas, el TOBIT aprovecha al máximo la información disponible, evitando la pérdida de datos valiosos. Es decir, el modelo TOBIT ofrece una solución que permite una estimación más precisa y realista de los efectos y relaciones subyacentes en el análisis de datos económicos.

El primer modelo propuesto es el siguiente:

$$\begin{cases} gasto\_mon^* & \text{si } gasto\_mon_i^* > 0 \\ 0 & \text{si } gasto\_mon_i^* \leq 0 \end{cases} \tag{49}$$

$$gasto\_mon^* = \beta_0 + \beta_1 \text{ est\_socio} + \beta_2 \text{ educa\_jefe} + \beta_3 \text{ tot\_integ} + \beta 4 \text{ alimentos} + \beta_5 \text{ vivienda} + \beta_6 \text{ salud} + \beta_7 \text{ comunica} + \beta_8 \text{ educacion} + \beta_9 \text{ transporte} + \beta_{10} \text{ ing\_cor} + e \tag{50}$$

Dicho modelo presenta buenos estimadores, y nos muestra que tenemos 69 datos censurados en la izquierda, los valores estimados se presentan a continuación:

Tabla 14: Resultados del Model TOBIT

	Coeficiente	Desviación.estándar	valorp
(Intercept)	-1197.8887300	124.1053131	0.0e+00
est_socio	-218.9980965	44.9514959	1.1e-06
educa_jefe	316.8059466	14.0143230	0.0e+00
tot_integ	-473.2598033	18.8182231	0.0e+00
alimentos	1.3508869	0.0011870	0.0e+00
vivienda	1.1751880	0.0063491	0.0e+00
salud	1.1592480	0.0056968	0.0e+00
comunica	1.5914330	0.0157507	0.0e+00
educacion	1.0934622	0.0045128	0.0e+00
transporte	1.1300791	0.0029049	0.0e+00
ing_cor	0.0421818	0.0001053	0.0e+00
Log(scale)	9.1703441	0.0023566	0.0e+00

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

En el Anexo A se observa el número de iteraciones del método de Newton-Raphson. Dado que este número es 30, esto indica que el modelo no necesitó muchas iteraciones para alcanzar una solución óptima durante el ajuste del modelo TOBIT.

Por otro lado, el valor Log verosimilitud con el valor de -953800 es interpretado como que el modelo tiene un buen ajuste.

Además, el valor p es muy pequeño lo que indica que al menos algunos de los coeficientes son estadísticamente significativos para explicar la variable dependiente.

3.6.1 Efectos marginales del modelo Tobit

Es importante resaltar que los estimadores en un modelo econométrico no representan directamente los efectos en las variables de interés. En su lugar, se utilizan los efectos marginales

y los signos de estos efectos para interpretar cómo cada variable afecta al resultado final.

Los efectos marginales nos proporcionan una medida precisa de cuánto cambia la variable dependiente cuando una variable explicativa cambia, manteniendo todo lo demás constante. Esta metodología nos permite entender mejor la verdadera influencia de cada variable en nuestros datos y tomar decisiones informadas basadas en estas interpretaciones detalladas.

Para el modelo, se calcularon los efectos marginales considerando que las variables monetarias están en sus valores medios, donde tenemos un hogar que pertenece al estrato socioeconómico medio bajo (estrato 2), con un jefe de hogar que tiene secundaria completa y está constituido por tres integrantes.

Esto nos proporciona una visión clara de cómo cambia la variable dependiente cuando las variables explicativas varían típicamente en su entorno promedio.

Este enfoque nos permite evaluar el impacto relativo de cada variable en el resultado del modelo, ofreciendo una interpretación más precisa y contextualizada de los resultados obtenidos, es así que tenemos:

Tabla 15: Efectos marginales modelo TOBIT

Concepto	Valor
Int	-1197.8405169
Est_socio	-218.9892822
Educa_jefe	316.7931957
Tot_integ	-473.2407554
Alimentos	1.3508326
Vivienda	1.1751407
Salud	1.1592013
Comunicaciones	1.5913690
Educación	1.0934182
Transporte	1.1300336
Ing_cor	0.0421801

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Los resultados muestran que si los valores del resto de las variables se encuentran en la media y los valores específicos, entonces:

- Un aumento en el estatus socioeconómico en una unidad disminuye el gasto en aproximadamente 218.98 unidades, lo que indica que un mejor estatus socioeconómico se asocia con un menor gasto en el hogar.
- Un aumento en la educación del jefe del hogar en una unidad incrementa el gasto en aproximadamente 316.79 unidades, lo cual sugiere que un mayor nivel educativo del jefe del hogar se asocia con un mayor gasto, posiblemente debido a mayores ingresos.
- Un aumento en el gasto en alimentos en una unidad incrementa el gasto total aproximadamente 1.35 unidades, lo cual se asocia con un incremento en el gasto total del hogar. Este resultado es esperado, ya que los alimentos constituyen una parte significativa del gasto total en la mayoría de los hogares.
- Un aumento en una unidad en el gasto en la vivienda incrementa el gasto total en 1.18 unidades, lo cual nos indica la alta proporción del gasto total del hogar destinada a la vivienda.
- Un aumento en una unidad en el gasto en salud incrementa el gasto total en 1.16 unidades, indicando la relevancia que tienen los gastos médicos en el gasto total del hogar.
- Un aumento en una unidad en el gasto en comunicaciones aumenta el gasto total en 1.59 unidades, destacando la importancia de las telecomunicaciones en el gasto total del hogar.

- Un aumento en una unidad en el gasto en educación aumenta el gasto en 1.09 unidades, resaltando la inversión en la formación académica dentro del gasto total del hogar.
- Un aumento en una unidad en el gasto en transporte aumenta el gasto total en 1.13 unidades, lo que nos indica los costos de movilidad dentro del gasto total del hogar.
- Un aumento en una unidad en el ingreso corriente aumentará el gasto total en 0.042 unidades, lo cual indica que al tener mayores ingresos existe un aumento pequeño pero positivo en el gasto total del hogar.

Estos efectos marginales nos proporcionan una visión detallada de como las diferentes variables influyen en el gasto total del hogar, permitiendo identificar podrían tener impacto significativo.

Además, se seleccionó un hogar representativo en las medias de cada decil con el fin de evaluar cómo se comporta el gasto en cada uno de estos. Este enfoque nos permite no solo entender cómo las variables influencian el resultado del modelo, sino también explorar cómo estas influencias se manifiestan en diferentes segmentos de la población, representados por los deciles y por estrato socieconómico, este análisis se detallará en el capítulo 4.

Capítulo 4. Conclusiones

En el siguiente capítulo se presentarán las conclusiones derivadas de la investigación, las cuales están basadas en el análisis de los datos presentados y en el cumplimiento de los objetivos planteados en el presente estudio. Además, se discutirán las implicaciones prácticas y metodológicas de los hallazgos.

El objetivo principal de esta investigación fue proponer un modelo econométrico con la finalidad de explicar las variables que influyen en el gasto de los hogares mexicanos en base a la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) del año 2022. La variable estudiada se ha relacionado con variables económicas, características de los hogares y del jefe del hogar. Lo que permite observar la relación entre cada una de ellas y los efectos causados por cada una.

Para alcanzar este objetivo, se realizaron tres tipos de modelos econométricos, los cuales fueron: modelos econométricos de regresión lineal múltiple, un modelo de propensión marginal y un modelo Tobit de los cuales se obtuvieron varios resultados interesantes respecto a la variable estudiada.

4.1 Análisis de los modelos de regresión lineal múltiple.

En el análisis econométrico, la regresión lineal múltiple es fundamental para entender cómo múltiples variables independientes influyen en una variable dependiente. Sin embargo, en el primer análisis, el modelo inicial, basado en los principales rubros de gasto de los hogares

mexicanos, presentó problemas de heterocedasticidad que afectaron la validez de las estimaciones. A pesar de los esfuerzos por corregir estos problemas, como la modificación de la forma funcional del modelo, no se logró mejorar su precisión. Diversas técnicas fueron exploradas, pero el problema persistió, lo que destacó la necesidad de reconsiderar el enfoque para obtener resultados más fiables.

Como se mencionó anteriormente, para evaluar la capacidad explicativa de cada variable sobre el gasto, se desarrollaron modelos individuales. Los resultados mostraron que la variable de alimentos tiene el mayor R^2 de 0.5206, lo que indica que es la que mejor explica el gasto de manera individual.

Asimismo, se analizaron los histogramas de cada variable, lo que reveló un sesgo hacia la izquierda, especialmente en aquellas categorías que incluyen valores de cero. Este sesgo se debe a que algunos hogares no declararon su gasto en ciertas variables o, en algunos casos, no tuvieron gasto en ellas. No se eliminaron estos valores de cero para evitar sesgos en la estimación del modelo. Este análisis gráfico sugirió que podrían ser necesarias transformaciones adicionales para normalizar las variables y mejorar el ajuste del modelo.

Se observó una gran variabilidad en los datos, lo cual podría deberse a factores económicos, ya que no toda la población tiene el mismo nivel de gasto e ingreso, lo que podría estar causando heterocedasticidad. Para abordar este problema y analizar la distribución del gasto de manera más precisa, se dividió la muestra en deciles según la variable gasto_total. Este enfoque segmenta la muestra en grupos más homogéneos en términos de nivel de gasto, lo que ayuda a reducir la heterocedasticidad y ofrecer una visión más clara de cómo las variables

influyen en el gasto en diferentes niveles de ingreso.

Sin embargo, los resultados continuaron siendo poco confiables debido a que el modelo no cumplía con los supuestos de normalidad y homocedasticidad de los errores, a pesar de que los errores eran más dispersos.

Para encontrar un modelo que explicara adecuadamente el gasto corriente monetario, se construyeron ocho modelos con diferentes formas funcionales. No se consideró un modelo logarítmico debido a la presencia de ceros en los datos. Aunque el modelo inicial parecía ser el más adecuado, la prueba RESET indicó que la forma funcional era incorrecta. Además, al probar el modelo en los diferentes deciles, se observó que la forma funcional seguía siendo incorrecta en cada uno de ellos. Esto reforzó la necesidad de explorar métodos adicionales para encontrar un modelo que cumpliera con los supuestos necesarios y proporcionara estimaciones confiables.

Por lo tanto, se consideró necesario incluir más variables en el análisis, utilizando todos los datos sin partirlos en deciles. Esto se basó en la hipótesis de que la omisión de variables relevantes podría estar contribuyendo a la falta de ajuste adecuado y a los problemas de heterocedasticidad y no normalidad de los errores.

Incorporar variables adicionales permitió capturar mejor la complejidad del gasto corriente monetario y reflejar de manera más precisa las diversas influencias sobre el gasto de los hogares. Así, se construyó un nuevo modelo de regresión lineal múltiple, incluyendo variables que podrían tener un impacto significativo en el gasto total. Esta estrategia buscó mejorar la explicación de la variabilidad del gasto y cumplir con los supuestos necesarios para obtener

estimaciones confiables y robustas.

Las variables del nuevo modelo son: est_socio, educ_jefe, tot_integ, alimentos, vivienda, salud, comunica, educación, trans e ing_cor. Aunque este modelo presentó un R^2 muy alto, indicando una buena capacidad de ajuste, el problema de heterocedasticidad persistió. Para abordarlo, se utilizaron errores estándar robustos, pero no se observó una reducción significativa en los valores de los errores, lo que sugiere que esta solución no mejoró la precisión del modelo de manera efectiva.

Otro método que se consideró fue utilizar solo los valores mayores a cero en todas las variables para aplicar logaritmos. Sin embargo, dado que la muestra original constaba de 90,102 observaciones, este enfoque solo consideraba menos de un tercio de los datos. Esta reducción en el tamaño de la muestra no solo limitaba la cantidad de información disponible, sino que también introducía un sesgo significativo. Al excluir los valores cero, se perdía una parte crucial de los datos que podría estar relacionada con patrones específicos de gasto, afectando así la representatividad y precisión de los resultados. Por lo tanto, este método resultó ser inviable como una solución efectiva para el análisis.

En un esfuerzo por encontrar una solución alternativa, se construyó un modelo de propensión marginal para evaluar su viabilidad. Este enfoque se exploró para ver si proporcionaría una mejora en la capacidad de ajuste del modelo y manejar de manera más efectiva la heterocedasticidad. Sin embargo, el análisis del modelo de propensión marginal también reveló limitaciones, y el problema de heterocedasticidad persistió.

Otro de los métodos que se utilizaron fue aplicar un modelo de reescalación. Este modelo

permitió ajustar las variables a una escala común, lo que facilitó la comparación entre diferentes magnitudes y mejoró la precisión en los análisis subsecuentes. Sin embargo, el modelo presentó heterocedasticidad, lo que indica que la varianza de los errores no fue constante a lo largo de las observaciones. Esto puede afectar la eficiencia y confiabilidad de las estimaciones, ya que las predicciones pueden volverse menos precisas en ciertos rangos de valores. En conclusión, aunque la reescalación facilitó el análisis, la presencia de heterocedasticidad debe abordarse con métodos adicionales para corregir este problema y mejorar la robustez del modelo.

4.2 Análisis del modelo Tobit

Dado que las variables presentarón un sesgo a la izquierda y hay hogares con un gasto reportado de cero, se utilizó un modelo Tobit para analizar el gasto de los hogares mexicanos. Este modelo se implementó para corregir el sesgo introducido por los datos censurados, proporcionando estimaciones más precisas y significativas de los coeficientes de las variables explicativas.

El análisis con el modelo Tobit permitió identificar y cuantificar las variables que influyen en el gasto de los hogares, adaptándose mejor a los datos censurados y ofreciendo estimaciones más confiables. Esto es crucial para comprender de manera más precisa los patrones de consumo de cada hogar mexicano.

El análisis econométrico realizado con el modelo TOBIT reveló que un aumento en el estatus socioeconómico está asociado con una disminución del gasto del hogar en aproximadamente

218.98 unidades, sugiriendo que un mejor estatus se correlaciona con un menor gasto, probablemente debido a un menor nivel de necesidad de gasto. Por otro lado, un mayor nivel educativo del jefe del hogar incrementa el gasto en 316.79 unidades, lo que refleja mayores ingresos relacionados con una educación superior. En cuanto a los componentes específicos del gasto, se observó que un aumento en el gasto en alimentos eleva el gasto total en 1.35 unidades, subrayando la importancia de este rubro en el presupuesto del hogar. Asimismo, incrementos en el gasto en vivienda y salud resultan en aumentos de 1.18 y 1.16 unidades en el gasto total, respectivamente. El gasto en comunicaciones incrementa el gasto total en 1.59 unidades, destacando su relevancia, mientras que un aumento en el gasto en educación incrementa el gasto total en 1.09 unidades, resaltando la importancia de la inversión en formación académica. El gasto en transporte también eleva el gasto total en 1.13 unidades, reflejando los costos asociados con la movilidad. Finalmente, un aumento en el ingreso corriente eleva el gasto total en 0.042 unidades, indicando que mayores ingresos tienen un pequeño pero positivo impacto en el gasto total del hogar.

4.3 Análisis por deciles

Se realizó un análisis de los deciles del gasto, donde se calcularon los efectos marginales utilizando hogares representativos en la media de cada uno de los deciles. Esto nos permite comparar los hogares que gastan más con los que gastan menos, de lo que se obtuvo lo siguiente:

Tabla 16: Efectos marginales de hogares representativos en las medias de deciles

Concepto	Decil1	Decil2	Decil3	Decil4	Decil5	Decil6
Int	-940.1293536	-1115.3995859	-1170.2251371	-1189.6846703	-1195.8565043	-1197.5719077
Est_socio	-171.8745103	-203.9174257	-213.9406366	-217.4982298	-218.6265649	-218.9401750
Educa_jefe	248.6362567	294.9900210	309.4897489	316.7931957	316.2684833	316.7221565
Tot_integ	-371.4246756	-440.6701352	-462.3305190	-470.0185576	-472.4569151	-473.1346337
Alimentos	1.0602057	1.2578620	1.3196901	1.3416350	1.3485952	1.3505297
Vivienda	0.9223133	1.0942620	1.1480486	1.1671394	1.1731942	1.1748771
Salud	0.9098033	1.0794197	1.1324768	1.1513086	1.1572813	1.1589414
Comunicaciones	1.2489915	1.4818436	1.5546811	1.5805337	1.5887331	1.5910121
Educación	0.8581731	1.0181641	1.0682102	1.0859733	1.0916071	1.0931730
Transporte	0.8869109	1.0522595	1.1039815	1.1223394	1.1281619	1.1297802
Ing_cor	0.0331052	0.0392771	0.0412077	0.0418929	0.0421102	0.0421706

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

De lo expuesto en la tabla, nos damos cuenta de que, conforme avanzamos en los deciles con menor gasto a los deciles de mayor gasto, los hogares incrementan sus gastos en todas las variables. Se observa una diferencia más marcada del decil uno al decil cinco, ya que a partir de este, el incremento en cada variable se vuelve de decimales.

Esto sugiere que los hogares en los deciles inferiores experimentan un aumento más significativo en su consumo a medida que sus ingresos crecen, probablemente porque satisfacen necesidades básicas y esenciales que no pueden cubrir con ingresos más bajos. Por otro lado, los hogares en los deciles superiores muestran incrementos más modestos, reflejando una saturación en el consumo de bienes esenciales y un posible cambio hacia gastos más discrecionales o de lujo. Lo cual se debe a que los hogares con mayores ingresos disponibles tienden a gastar más en cada una de las variables, ya que, al tener ingresos más altos, se permiten gastar más en alimentos, vivienda, salud, educación, transporte y otras necesidades. Dicho comportamiento refleja la capacidad de cada hogar para invertir en una mejor calidad de vida y acceder a bienes y servicios adicionales que los hogares con menores ingresos no pueden

costear.

4.4 Análisis por estrato socioeconómico

Se llevó a cabo un análisis por estrato socioeconómico para identificar las diferencias en el consumo entre cada uno de ellos. Éste es fundamental para entender cómo varían los patrones de gasto según el nivel socioeconómico, lo que proporciona información valiosa para la formulación de políticas públicas, el desarrollo de estrategias de mercado y la implementación de programas sociales más efectivos, asegurando así que las intervenciones estén adecuadamente dirigidas y sean más equitativos, para ello se utilizaron los efectos marginales donde se escogió un hogar perteneciente a cada estrato socioeconómico, las características de los hogares se presentan a continuación:

- Estrato socioeconómico bajo: se escogió un hogar donde el jefe del hogar cuenta con primaria incompleta, el hogar está conformado por cuatro integrantes, los gastos por trimestre que tiene son: en alimentos es de 5,323.78 pesos, en vivienda de 2,072.61 pesos, en salud es de 176.08 pesos, en comunicaciones es de 580.64 pesos, en educación son de 5,225.79 pesos, en transporte es de 244.56 pesos y el ingreso es de 36,859.61 pesos.
- Estrato socioeconómico medio bajo: se escogió un hogar donde el jefe del hogar cuenta con primaria completa, el hogar lo conforman cuatro integrantes, los gastos por trimestre que tiene son: en alimentos es de 10967.08 pesos, en vivienda es de 1045.64 pesos, en salud es de 2054.34 pesos, en comunicaciones es de 4122.58 pesos, en educación son de 11612.90 pesos, en transporte es de 2322.58 pesos y el ingreso es de 93,373.09 pesos.

- Estrato socioeconómico medio alto: se escogió un hogar donde el jefe del hogar cuenta con preparatoria incompleta, el hogar lo conforman cuatro integrantes, los gastos por trimestre que tiene son: en alimentos es de 16,495.52 pesos, en vivienda es de 7875 pesos, en salud es de 542.91 pesos, en comunicaciones es de 2700 pesos, en educación son de 2100 pesos, en transporte es de 4475.4 pesos y el ingreso es de 102,737.70 pesos.

- Estrato socioeconómico alto: se escogió un hogar donde el jefe del hogar cuenta con posgrado, el hogar lo conforman cuatro integrantes, los gastos por trimestre que tiene son: en alimentos es de 57,503.35 pesos, en vivienda es de 9,625 pesos, en salud es de 1829.33 pesos, en comunicaciones es de 5940 pesos, en educación son de 7920.65 pesos, en transporte es de 21480 pesos y el ingreso es de 235,819.62 pesos.

De los cuales se obtuvieron los efectos marginales:

Tabla 17: Efectos marginales de hogares representativos por estrato socioeconómico

	Concepto	Estrato1	Estrato2	Estrato3	Estrato4
7	Int	-1138.1869755	-1197.8812872	-1197.8879995	-1197.8887300
5	Est_socio	-208.0834178	-218.9967358	-218.9979630	-218.9980965
3	Educa_jefe	301.0166079	316.8039782	316.8057534	316.8059466
9	Tot_integ	-449.6729376	-473.2568628	-473.2595147	-473.2598033
1	Alimentos	1.2835599	1.3508785	1.3508861	1.3508869
11	Vivienda	1.1166176	1.1751807	1.1751872	1.1751880
8	Salud	1.1014721	1.1592408	1.1592473	1.1592480
2	Comunicaciones	1.5121173	1.5914231	1.5914320	1.5914330
4	Educación	1.0389650	1.0934554	1.0934615	1.0934622
10	Transporte	1.0737569	1.1300721	1.1300784	1.1300791
6	Ing_cor	0.0400795	0.0421815	0.0421818	0.0421818

Fuente: Elaboración propia utilizando R Studio con datos de la base 'concentradohogar'.

Como se observa, existe un cambio significativo entre el estrato socioeconómico bajo (estrato 1) y el estrato socioeconómico medio bajo (estrato 2). Esto ocurre porque los hogares del estrato 1 suelen enfrentar necesidades básicas insatisfechas. Por lo tanto, cualquier cambio

en sus condiciones socioeconómicas tiene un impacto notable en su gasto. Esto se refleja en los hogares seleccionados para el análisis: en el estrato 1, el jefe del hogar tiene un nivel educativo de primaria incompleta, mientras que en el estrato 2, el nivel educativo es primaria completa. Es decir, un grado adicional de educación impacta en las demás variables y en el dinero destinado a cada una de ellas. Por lo que, con mayores ingresos, los hogares cubren mejor sus gastos, lo que reduce la variabilidad en el gasto adicional.

Además, es importante destacar que solo se observan cambios significativos en los efectos marginales en el estrato socioeconómico bajo (estrato 1) y medio bajo (estrato 2) mientras para los dos estratos socioeconómicos restantes solo hay diferencias en decimales con el estrato socioeconómico medio bajo, lo cual se debe a que los patrones de consumo en los estratos socioeconómicos medios y altos se pueden volver más homogéneos, es decir, una vez que se alcanza un cierto nivel de ingreso las proporciones de gasto en distintos rubros se estabiliza como se observa en la tabla, por lo que la diferencia en gasto entre los estratos son menos pronunciadas.

4.5 Problemas encontrados.

Al desarrollar el modelo econométrico Tobit, se seleccionaron variables clave que se consideraron relevantes para explicar el gasto total de los hogares mexicanos. Específicamente, se encontraron dificultades son el intercepto y la variable tot_integ.

En el análisis de regresión reveló que algunas observaciones influyen significativamente en los valores ajustados y en los coeficientes estimados, como lo indica la presencia de valores

extremos en DFFITS y DFbetas (Vease en Anexo A). Estos puntos influyentes pueden distorsionar los resultados del modelo, lo que explica, por ejemplo, un intercepto negativo o coeficientes inesperadamente negativos.

En nuestro caso, un intercepto negativo en el modelo Tobit puede estar influenciado tanto por observaciones influyentes como por el sesgo hacia la izquierda de las variables independientes. En un modelo Tobit, que se usa para datos censurados, el intercepto refleja el valor esperado de la variable dependiente cuando todas las variables independientes son cero. Sin embargo, la censura en la variable dependiente y el sesgo hacia valores bajos en las variables independientes lleva a una estimación sesgada del intercepto. Específicamente, si las variables independientes están sesgadas hacia valores bajos, el modelo podría ajustar el intercepto negativo para compensar el efecto de la censura en la variable dependiente. En otras palabras, el intercepto negativo podría reflejar una tendencia a que los valores observados de la variable dependiente se concentren en el límite inferior debido a la censura.

Por otro lado, en el modelo Tobit, la variable *tot_integ* también muestra un valor negativo, lo cual es inesperado y puede estar influenciado por observaciones influyentes. Es sus estadísticas tenemos que su media aritmética es de 3.4535, con un valor mínimo de 1 y un máximo de 19. En el histograma de la variable muestra un sesgo hacia la izquierda, lo que indica que la mayoría de los valores se concentran en el extremo inferior. Este sesgo, combinado con la censura en la variable dependiente, puede contribuir a la estimación negativa del intercepto en el modelo Tobit.

La combinación de censura, sesgo en la distribución y posibles interacciones entre variables

influyen en cómo se ajusta el intercepto y en la interpretación de los coeficientes en el modelo Tobit.

4.6 Conclusiones generales

A lo largo de esta investigación, se ha observado que el trabajo para encontrar un modelo econométrico que explique las variables que influyen en el gasto de los hogares mexicanos ha sido extenso y complicado debido a la naturaleza de los datos. Inicialmente, se comenzó con un modelo de regresión lineal múltiple, el cual no logró capturar adecuadamente la relación entre las variables debido a la especificidad de los datos. Se probaron varias formas funcionales adicionales sin éxito, reflejando la complejidad inherente al fenómeno estudiado.

En un intento adicional, se exploró un modelo de propensión marginal, que tampoco resultó ser el más adecuado para describir las particularidades del gasto de los hogares. Finalmente, se optó por utilizar un modelo Tobit, que se adaptó de manera más precisa a la naturaleza censurada de los datos, ofreciendo una mejora significativa en el ajuste del modelo y representando una aproximación más adecuada para capturar la estructura de los datos disponibles.

Este proceso de selección y ajuste de modelos resalta la importancia de adaptar el modelo a la naturaleza específica de los datos y las limitaciones inherentes a cada tipo de análisis. A pesar de los desafíos encontrados, el modelo Tobit proporciona una base sólida para comprender mejor las variables que influyen en el gasto de los hogares mexicanos. No obstante, es esencial seguir explorando y refinando los modelos en futuras investigaciones para lograr

una comprensión más completa.

Además, es importante destacar que la econometría y la actuaría comparten una base común en el uso de modelos matemáticos y estadísticos para tomar decisiones informadas. La econometría proporciona las herramientas y técnicas necesarias para analizar datos económicos y modelar relaciones complejas, mientras que la actuaría se enfoca en la evaluación y gestión de riesgos financieros y seguros. Ambas disciplinas requieren un entendimiento profundo de los datos y la capacidad de aplicar modelos adecuados para prever y manejar incertidumbres.

En particular, el estudio de patrones de gasto, como el analizado en esta investigación, ofrece conocimientos valiosos para la planificación financiera y la gestión de riesgos en el campo actuarial. La capacidad de predecir y analizar el comportamiento del gasto de los hogares es crucial para diseñar estrategias de seguros, productos financieros y políticas económicas que respondan efectivamente a las necesidades y comportamientos de los consumidores. Así, este análisis no solo contribuye al conocimiento académico sino que también proporciona herramientas prácticas que pueden ser aplicadas en el ámbito actuarial para mejorar la toma de decisiones y la planificación financiera.

Bibliografía

1. Amat Rodrigo, J. (2018, abril). *Ciencia de datos net*. Recuperado de https://cienciadedatos.net/documentos/40_tobit_regression_modelos_lineales_para_datos_censurados

2. Amat Rodrigo, J. (2016, junio). *Ciencia de datos net*. Recuperado de https://cienciadedatos.net/documentos/24_correlacion_y_regresion_lineal

3. AppsFlyer. (2023, 21 septiembre). *Modelado probabilístico*. Recuperado de https://www.appsflyer.com/es/glossary/probabilistic-modeling/

4. Arellano Hidalgo, L. I. (2021). *Los salarios por sector de actividad económica en la economía de México, 1994-2019* [Tesis licenciatura]. Universidad Autónoma del Estado de México Centro Universitario UAEM Texcoco.

5. Barcelona, E. d. (s.f.). *RPbs by RStudio*. Recuperado de https://rpubs.com/aozoro/JarqueBera

6. Benites, L. (2021, 12 septiembre). *Criterio de información de Akaike: definición, fórmulas*. Statologos. Recuperado de https://statologos.com/criterio-de-informacion-de-akaikes/

7. Calzada, H. (2019, 27 diciembre). *¿Qué es el R cuadrado ajustado?* Rankia. Recuperado de https://www.rankia.mx/blog/como-comenzar-invertir-bolsa/4439147-que-r-cuadrado-ajustado

8. Cloud, S. A. (2024). *Ayuda para SAP Analytics Cloud*. Recuperado de https://help.sap.com/docs/SAP_ANALYTICS_CLOUD/00f68c2e08b941f081002fd3691d86a7/cd897576c3344475a208c2f7a52f151e.html?locale=es-ES

9. Contreras, J. (2023, 28 junio). *Alta, media o baja: ¿a qué clase social perteneces según el INEGI?* Infobae. Recuperado de https://www.infobae.com/mexico/2023/06/28/alta-media-o-baja-a-que-clase-social-perteneces-segun-el-inegi/#:~:text=Seg%C3%BAn%20datos%20del%20Instituto%20Nacional,1.2%25%20a%20la%20clase%20alta

10. Del Barrio Castro, T., Clar López, M., & Suriñach Caralt, J. (2013). *Modelo de regresión lineal múltiple: especificación, estimación y contraste*. España: Universitat Oberta de Catalunya.

11. Instituto Nacional de Estadística y Geografía (INEGI). (2022). *Encuesta Nacional de Ingresos y Gastos de los Hogares 2022, Nueva serie*. Recuperado de https://www.inegi.org.mx/programas/enigh/nc/2022/,2022

12. Elizalde Ángeles, E. N. (2012). *Econometría*. México: RED TERCER MILENIO.

13. Eva. (2019, 23 abril). *Tema 2b. El modelo de regresión múltiple*. Estadisticaparatodos.com. Recuperado de https://estadisticaparatodos.com/modelo-de-regresion-multiple/

14. Faster Capital. (2024, 9 abril). *Mínimos cuadrados ponderados: manejo de la heterocedasticidad con precisión*. Recuperado de https://fastercapital.com/es/contenido/Minimos-cuadrados-ponderados--manejo-de-la-heterocedasticidad-con-

precision.html#:~:text=M%C3%ADnimos%20cuadrados%20ponderados%20(WLS)%3A,las%20observaciones%20con%20mayor%20varianza

15. García, A. K. (2023, 12 octubre). *La canasta alimentaria se encarece por encima de la inflación; subió 6 % en septiembre*. El Economista. Recuperado de https://www.eleconomista.com.mx/economia/La-canasta-alimentaria-encarece-por-encima-de-la-inflacion-subio-6-en-septiembre-20231012-0043.html

16. Gonzáles Borja, J., & Nieto Sánchez, F. H. (2008). *Distribución de la estadística de Jarque y Bera para la prueba de normalidad en una serie temporal estacionaria con datos faltantes*. Bogotá.

17. Gómez, S. V., Guerra, C. F. V., & Teodoro, L. A. R. (s.f.). *Capítulo 2 Datos truncados | Modelos de supervivencia*. Recuperado de https://carlosfernandovg.github.io/supervivencia_y_series_FC2021-1/datos-truncados.html

18. Gutiérrez, A. (2016, 6 junio). *¿Cuánto gastan los mexicanos en transporte?* Propiedades.com. Recuperado 15 de febrero de 2024, de https://propiedades.com/blog/arquitecura-yurbanismo/cuanto-gastan-los-mexicanos-transporte

19. H. Stock, J., & Mark, M. (2012). *Introducción a la econometría* (3a ed.). Madrid: Pearson Educación, S.A.

20. Herrera Yáñez, R. (2015). *Análisis de la educación y el crecimiento económico en* [Tesis]. Universidad Autónoma del Estado de México.

21. Herrera, J. (2022, 11 mayo). *Que el transporte no acabe con tu bolsillo*. El Econo-

mista. Recuperado de https://www.eleconomista.com.mx/finanzaspersonales/Que-el-transporte-no-acabe-con-tu-bolsillo-20220511-0092.html

22. Hernández, D. (2023, 1 septiembre). *La apuesta por la educación*. IMCO. Recuperado de https://imco.org.mx/la-apuesta-por-la-educacion/#:~:text=La%20Encuesta%20Nacional%20de%20Ingresos,a%20la%20educaci%C3%B3n%20durante%202022.

23. Iglesias Ibarra, Á. J., & Fernández Rangel, J. A. (2022). *Introducción a la econometría: Teoría y aplicaciones usando STATA 17*. Bogotá, Colombia: Fundación Universitaria del Área Andina.

24. Instituto Nacional de Estadística y Geografía (INEGI). (2023). *Encuesta Nacional de Ingresos y Gastos de los Hogares 2022, Nueva serie*. Recuperado de https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/889463910589.pdf

25. Instituto Nacional de Estadística y Geografía (INEGI). (2023). *Red Nacional de Metadatos: Encuesta Nacional de Ingresos y Gastos de los Hogares 2022, Nueva serie*. Recuperado de https://www.inegi.org.mx/rnm/index.php/catalog/901/data-dictionary/F55?file_name=ingresos

26. Deloitte México. (s.f.). *¿Cómo gastan los mexicanos en telecomunicaciones?* Recuperado de https://www2.deloitte.com/mx/es/pages/consumer-business/articles/como-gastan-los-mexicanos-en-telecomunicaciones.html

27. León Bon, T. S. (2020). *Impacto de la inflación de los alimentos en el bienestar*. Tijuana: El Colegio de la Frontera Norte.

28. López, J. F. (2021, 19 febrero). *Coeficiente de determinación (R cuadrado)*. Economipedia. Recuperado de https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html

29. Lozano, R. (2023, 27 junio). *Datos anuales sobre la salud de los mexicanos: oportunos, pero sin buenas noticias*. El Economista. Recuperado de https://www.eleconomista.com.mx/opinion/Datos-anuales-sobre-la-salud-de-los-mexicanos-oportunos-pero-sin-buenas-noticias-20230627-0020.html

30. Lustig, N. (2007, 5 marzo). *Salud y desarrollo económico: El caso de México*. Recuperado de https://www.eltrimestreeconomico.com.mx/index.php/te/article/view/383/580#info

31. Marina Clemente, J. A., Gerónimo Antonio, V. M., & Pérez Abarca, J. M. (2017, 13 diciembre). *Universidad DeLaSalle*. Recuperado de \url{https://www.redalyc.org/journal/2033/203358383026/html/}

32. Martínez, E. (2019, 20 febrero). *México, el país que más gasta en alimentos*. Recuperado de https://www.eleconomista.com.mx/arteseideas/Mexico-el-pais-que-mas-gasta-en-alimentos-20190220-0124.html

33. McDonald, G. (2019, 22 septiembre). *Todo sobre la regresión Tobit*. Statology. Recuperado de https://www.statology.org/tobit-regression/

34. Moreno, M. (2023, 19 mayo). *Esto es lo que se necesita para elevar la educación en México*. IMCO. Recuperado de https://imco.org.mx/esto-es-lo-que-se-necesita-para-

elevar-la-educacion-en-mexico/#:~:text=Para%20el%20IMCO%2C%20si%20el,con%20mejores%20ingresos%20a%20futuro.

35. Navarro, E. M. (2021). *Los servicios de salud y la calidad de vida en los adultos mayores del centro occidente del Estado de México* [Tesis]. Universidad Autónoma del Estado de México.

36. Navarro, M. A. (2017). *El gasto de los hogares en el transporte público: efectos sobre la pobreza y el bienestar de los hogares urbanos en México*. Ciudad de México: Universidad Nacional Autónoma de México.

37. Negrete Medina, M. N. (2015). *Pobreza y feminización en el Estado de México*. Toluca, México: Tesis de la Universidad Autónoma del Estado de México.

38. Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619-632.

39. Pérez, J. J. (2022, 23 febrero). *Análisis del crecimiento de la economía mexicana: Una revisión de la literatura*. Economía Informa. Recuperado de https://www.revistas.unam.mx/index.php/ei/article/view/83795/76615

40. Petkova, E. (s.f.). *Truncated regression model*. Recuperado de https://stats.oarc.ucla.edu/stata/dae/truncated-regression-model/

41. Ramírez García, N. C., Ramírez García, E. P., & Vázquez Olguín, C. (2021). *Econometría Básica*. Toluca, México: Universidad Autónoma del Estado de México.

42. Ruiz, G. (2022, 20 agosto). *¿Cómo calcular el R cuadrado ajustado en Excel?* Excel y VBA. Recuperado de https://excelyvba.com/blog/como-calcular-el-r-cuadrado-ajustado-en-excel

43. Sansonetti, P. (2023, 12 septiembre). *Gasto en educación y el futuro de México.* IMCO. Recuperado de https://imco.org.mx/el-gasto-en-educacion-y-el-futuro-de-mexico/#:~:text=El%20gasto%20total%20en%20educaci%C3%B3n,Presupuesto%20de%20Egresos%20de%20la%20Federaci%C3%B3n.

44. Schneider, M. (2022, 5 junio). *El método de máxima verosimilitud: definición y ejemplo.* Economipedia. Recuperado de https://economipedia.com/definiciones/metodo-de-maxima-verosimilitud.html

45. Smith, T. (2018, 23 noviembre). *Prueba Jarque Bera: definición y usos.* Economipedia. Recuperado de https://economipedia.com/definiciones/prueba-jarque-bera.html

46. Sour, L. (2023, 18 julio). *¿Cómo afecta la inflación a la pobreza en México?* IMCO. Recuperado de https://imco.org.mx/como-afecta-la-inflacion-a-la-pobreza-en-mexico/#:~:text=La%20inflaci%C3%B3n%20ha%20aumentado%20de,de%20pobreza%20por%20ingresos%20laborales.

47. Spence, M. W. (2020). *Econometría intermedia.* Madrid, España: Pearson.

48. Torrico, P. M. (2022, 19 agosto). *El coeficiente de determinación R2.* Economipedia. Recuperado de https://economipedia.com/definiciones/coeficiente-de-determinacion-r2.html

49. Varela, M. A. (2023, 3 septiembre). *Impacto de los programas sociales en México*. Nexos. Recuperado de https://www.nexos.com.mx/?p=66292#:~:text=En%20M%C3%A9xico%20exist%C3%ADan%20en%202020,%E2%80%9D%2C%20abund%C3%B3%20Mar%C3%ADa%20del%20Carmen%20Garc%C3%ADa%20Naranjo.

50. Vázquez García, M. E., & Ramírez García, E. P. (2007). *Educación, salud y calidad de vida en el Estado de México*. Toluca, México: Tesis de la Universidad Autónoma del Estado de México.

51. Wooldridge, J. M. (2016). *Introducción a la econometría: Un enfoque moderno*. México: CENGAGE Learning.

52. X-Rates. (2023, 30 junio). *Tasa de inflación en México*. Recuperado de https://www.x-rates.com/graph/?from=MXN&to=USD&amount=1

53. Zavala, J. (2023, 23 agosto). *¿Qué tan alto es el gasto de los hogares mexicanos en educación?* Deloitte México. Recuperado de https://www2.deloitte.com/mx/es/pages/consumer-business/articles/que-tan-alto-es-el-gasto-de-los-hogares-mexicanos-en-educacion.html

ANEXO A

A.1 Modelo de regresión lineal múltiple

Para desarollar el modelo de regresión lineal múltiple, utilizaremos código en R studio.

Para trabajar el modelo, utilizaremos diversas bibliotecas las cuales se usarán en diferentes fases del análisis, por lo que las librerías son las siguientes:

```
library(pacman)
p_load(dbplyr, knitr, xtable,printr,rmarkdown,openxlsx,readxl,
       tidyr, effects, car, AER, broom, stats,lmtest, stargazer,
       ggplot2, bookdown, sandwich, broom, performance,
       Hmisc,corrplot,gridExtra,systemfit,
       tseries,kableExtra,see)
```

A continuación se cargara la base de datos con la que trabajaremos, por lo que tenemos:

```
hogar <- read_excel("~/Tesis/concentradohogar.xlsx")
attach(hogar)
```

Crearemos una base que solo contenga 17 variables con las categorías mostradas en los Anexos B, las características del jefe del hogar y las características sociodemográficas, es así que tenemos lo siguiente:

```
trans <- hogar[["transporte"]] - hogar[["comunica"]]
categorias <-hogar[c("ubica_geo", "est_socio", "clase_hog",
                  "sexo_jefe", "edad_jefe", "educa_jefe",
                  "tot_integ", "menores", "gasto_mon",
                  "alimentos", "ali_dentro", "vivienda",
```

```
                "salud", "comunica", "educa_espa",
                "educacion","ing_cor")]
hogar1 <- cbind(categorias, trans = trans)
```

De las cuales solo seleccionaremos a las variables **gasto_mon, alimentos, educacion, salud, tot_integ, ing_cor** y **trans**, dado a que dichas variables de acuerdo a la ENIGH 2022 alimentos, transporte y educación son los 3 gastos principales de los hogares mexicanos mientras que salud es el rubro en el que menos se gasta, además de incluir el **tot_integ** para observar si dicha variable influye en el gasto, además de incluir a la variable **ing_cor** por la relación económica existente entre gasto e ingreso, dicho lo anterior agruparemos a las variables en **hogar1**, por lo que tenemos:

```
seleccion <-hogar[c("gasto_mon", "alimentos", "educacion",
                "salud", "tot_integ", "ing_cor")]
hogar2<-cbind(seleccion, trans=trans)
```

A.1.1 Estadísticas descriptivas

Primero es importante conocer como estan estructurados los datos, por lo que tenemos:

```
ncol(hogar2)
```

```
## [1] 7
```

```
nrow(hogar2)
```

```
## [1] 90102
```

De análisis anterior sabemos que la tabla esta compuesta por 6 columnas y 90,102 filas.

Es importante tener una visualización de los datos se trabajaran, por lo que observemoslo en la siguiente tabla:

```
head(hogar2)
```

gasto_mon	alimentos	educacion	salud	tot_integ	ing_cor	trans
35091.17	9514.19	2903.22	2641.29	3	56123.75	5217.50
78670.73	17524.25	0.00	0.00	2	108048.87	4354.83
101647.27	18321.36	0.00	0.00	3	133852.88	11612.90
46702.31	14759.90	0.00	0.00	4	105054.15	20322.58
26927.85	12458.47	0.00	0.00	1	24211.95	4064.51
51176.07	6351.40	6967.74	1007.60	4	121649.86	17709.67

A continuación realizaremos los estadísticos descriptivos con la finalidad de conocer como se comportan los datos:

```
summary(hogar2)
```

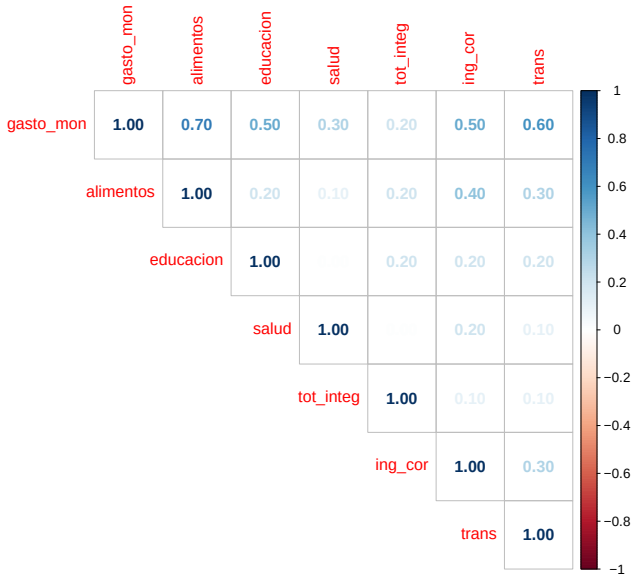
gasto_mon	alimentos	educacion	salud	tot_integ	ing_cor	trans
Min. : 0	Min. : 0	Min. : 0	Min. : 0.0	Min. : 1.000	Min. : 0	Min. : 0.0

gasto_mon	alimentos	educacion	salud	tot_integ	ing_cor	trans
1st Qu.:	1st Qu.:	1st Qu.: 0	1st Qu.:	1st Qu.:	1st Qu.:	1st Qu.:
18561	7483		0.0	2.000	28386	925.7
Median :	Median :	Median :	Median :	Median :	Median :	Median :
29679	11957	0	146.7	3.000	46074	3193.5
Mean :	Mean :	Mean :	Mean :	Mean :	Mean :	Mean :
37615	14046	2467	1270.6	3.435	61490	5739.4
3rd Qu.:	3rd Qu.:	3rd Qu.:	3rd Qu.:	3rd Qu.:	3rd Qu.:	3rd Qu.:
45901	17961	2177	841.3	4.000	74344	6930.0
Max.	Max.	Max.	Max.	Max.	Max.	Max.
:1703575	:849840	:451161	:324547.7	:19.000	:7153770	:489130.4

Visualizamos que el gasto medio de los hogares es de $\$37,615.00$, y la media de su gasto en alimentos es de $\$11,957.00$, siendo así el primer rubro en el que gastan más.

Para crear nuestro modelo, crearemos una matriz de correlación con la finalidad de proponer el mejor modelo:

```
correlacion<-round(cor(hogar2), 1)
corrplot(correlacion, method="number", type="upper" )
```



Es importante tener en cuenta que los datos superiores al 0.8 implican una dependencia alta y pueden significar la presencia de multicolinealidad.

En particular que el gasto_mon, tiene una correlación de 0.70 con los alimentos, mientras que con educación es del 0.50, el gasto en salud es del 0.30, con el total de integrantes dentro del hogar es de 0.20, los gastos en transporte de 0.70 y los gastos en vivienda son del 0.40.

Con lo anterior tenemos que el primer modelo a trabajar es el siguiente:

Modelo 1

$$gasto_mon = \beta_0 + \beta_1 \cdot alimentos + \beta_2 \cdot educacin + \beta_3 \cdot salud + \beta_4 \cdot tot\_integrantes + \beta_5 \cdot trans + \beta_6 \cdot ing\_cor$$

```
mod1<-lm(gasto_mon~alimentos+educacion+salud+tot_integ+trans+ing_cor,
         data=hogar2)
kable(tidy(mod1),caption="Resumen del Modelo 1")
```

Tabla 20: Resumen del Modelo 1

term	estimate	std.error	statistic	p.value
(Intercept)	2887.5805462	97.6293659	29.57697	0
alimentos	1.5354825	0.0041912	366.36086	0
educacion	1.2132582	0.0058184	208.51987	0
salud	1.1830197	0.0074091	159.67043	0
tot_integ	-672.8197477	24.0934098	-27.92547	0
trans	1.1934367	0.0037725	316.35122	0
ing_cor	0.0670954	0.0005896	113.79187	0

```
# Extraer los valores
modelo1 <- summary(mod1)
residual_standard_error <- modelo1$sigma
multiple_r_squared <- modelo1$r.squared
adjusted_r_squared <- modelo1$adj.r.squared
f_statistic <- modelo1$fstatistic[1]
p_value <- pf(modelo1$fstatistic[1], modelo1$fstatistic[2],
              modelo1$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultados<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                        "p-value" = p_value)
# Crear la tabla con kable
tablar <- kable(resultados, caption = "Resultados del Modelo de
                          Regresión",format = "latex",
                          booktabs = TRUE, align = "c") %>%
```

```
kable_styling(latex_options = c("striped", "hold_position"))
tablar
```

Tabla 21: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	12381.13	0.8723253	0.8723168	102594.3	0

En la tabla anterior el p-value nos lo marca como 0 dado a que su valor real es muy pequeño, para ser exactos es menor a 2.2e-16.

Con lo anterior decimos que el modelo de regresión lineal múltiple muestra un ajuste aceptable de 68.2 % de la variabilidad en la variable dependiente, además observamos que el valor de F es grande nos indica que la varianza explicada por el modelo es más grande en comparación de la varianza no explicada, lo que nos quiere decir que al menos una de las variables independientes en el modelo tiene un efecto significativo en la variable dependiente y en conjunto con el valor p muy pequeño nos indica que el modelo en su conjunto es significativo.

A.1.2 Validación de supuestos

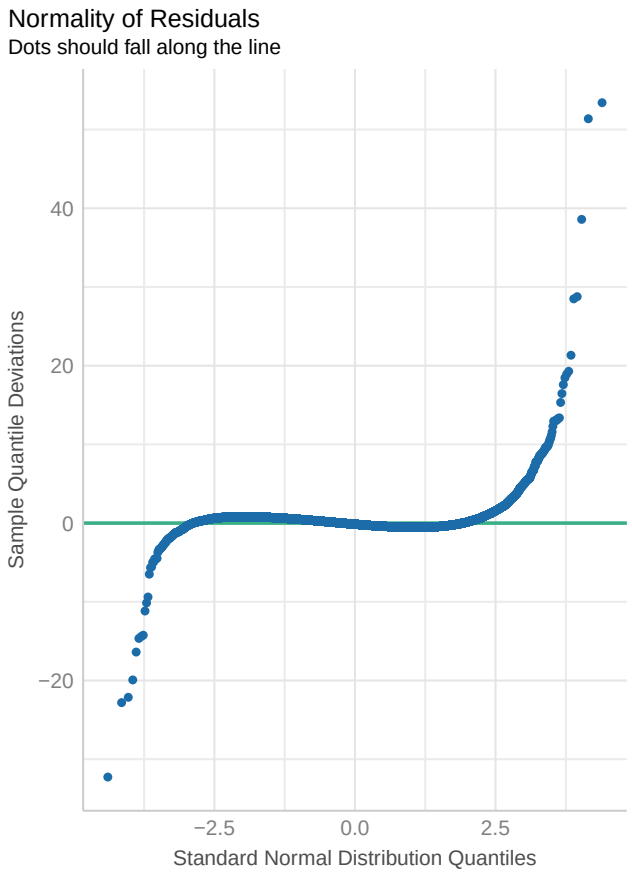
Verificaremos la regresión lineal múltiple en el modelo que estamos proponiendo, los datos deben cumplir con los 5 supuestos, los cuales son: linealidad, independencia, homocedasticidad, normalidad y no colinealidad.

Esto es de gran importancia ya que nos estimará un buen modelo que nos explique que variables influyen en el gasto de los hogares.

Normalidad

Realizaremos un gráfico con la finalidad de tener una visualización de los residuos, por lo que tenemos:

```
library(see)
ehat<-resid(mod1)
ebar<-mean(ehat)
sde<-sd(ehat)
check_model(mod1, check = "qq")
```



En la gráfica tenemos que los residuos no se ajustan a una distribución normal, por lo que realizaremos la prueba de jarque-bera para comprobar si existe normalidad, es así que tenemos:

```r
jarque.bera.test(mod1$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  mod1$residuals
## X-squared = 464112978, df = 2, p-value < 2.2e-16
```

La prueba *jarque-bera* nos plantea la hipótesis nula que nos dice que los datos provienen de una distribución normal, mientras que la hipótesis alternativa es que los datos no provienen de una distribución normal, cuando se observa un valor p muy pequeño, decimos que los errores no siguen una distribución normal.

Heterocedasticidad

Ahora con este modelo comprobemos el supuesto de homocedasticidad, por lo que para ello hagamos una prueba Breusch-Pagan Test, es así que nuestra prueba de hipótesis es:

$$H_0 : \text{Homocedasticidad en el modelo vs } H_a : \text{Sin homocedasticidad}$$

```r
bptest(mod1)
```

```
##
##  studentized Breusch-Pagan test
```

```
##
## data:  mod1
## BP = 9755.2, df = 6, p-value < 2.2e-16
```

Se observa que el valor p es menor que el nivel de significancia usual de 5 %, por lo tanto, hay evidencias para decir que no se cumple la homocedasticidad de los errores.

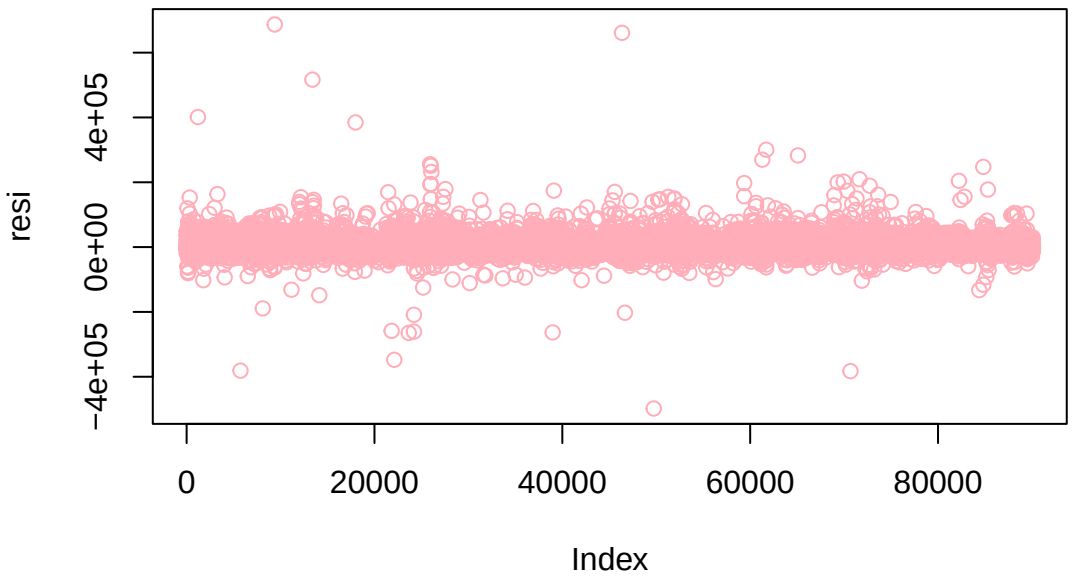
Para corregir este problema de heterocedasticidad utilizaremos Mínimos Cuadrados Ponderados, para corregir la heterocedasticidad. En un procedimiento donde se da más peso a las observaciones con menor varianza porque estas observaciones proporcionan información más confiable sobre la función de regresión que aquellas con grandes varianzas.Es así que tenemos:

```
mod2<-lm(gasto_mon~alimentos+educacion+salud+tot_integ+trans+ing_cor,
         weights = 1/tot_integ, data=hogar2)
bptest(mod2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 4990.1, df = 6, p-value < 2.2e-16
```

Ahora el gráfico de los errores, es:

```
resi<- residuals(mod2)
plot(resi, col="#FFAEB9")
```



Aunque hemos utilizado los mínimos cuadrados ponderados, vemos que aún el valor p es menor que el nivel de significancia usual de 5 %, por lo tanto, sigue existiendo evidencia para decir que no se cumple la homocedasticidad de los errores.

Por lo que intentaremos cambiar la forma funcional, para corregir este error, para ello utilizaremos una forma exponencial dado a que lo que queremos es reflejar la relación no lineal entre el gasto y sus componentes.

Por ellos cambiemos la forma funcional de las variables, por lo que veamos:

```
mod3<-lm(gasto_mon~I(alimentos^2)+I(educacion^2)+I(salud^2)+
           I(tot_integ^2)+I(trans^2)+I(ing_cor^2),
         weights = 1/tot_integ, data=hogar2)
kable(tidy(mod3),caption="Resumen del Modelo 3.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 22: Resumen del Modelo 3.

term	estimate	std.error	statistic	p.value
(Intercept)	2.683760e+04	114.2511714	234.89998	0
I(alimentos^2)	3.100000e-06	0.0000000	117.45540	0
I(educacion^2)	6.200000e-06	0.0000001	71.08606	0
I(salud^2)	6.600000e-06	0.0000001	58.79957	0
I(tot_integ^2)	4.979853e+02	8.1499570	61.10281	0
I(trans^2)	4.500000e-06	0.0000000	124.92639	0
I(ing_cor^2)	0.000000e+00	0.0000000	26.65190	0

Realizemos las pruebas, para saber si este modelo cumple con normalidad y heterocedasticidad, por lo que tenemos:

```
jarque.bera.test(mod3$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  mod3$residuals
## X-squared = 15735617, df = 2, p-value < 2.2e-16
```

```
bptest(mod3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod3
## BP = 10469, df = 6, p-value < 2.2e-16
```

Podemos observar que aunque hemos cambiado la forma funcional del modelo, sigue sin existir la normalidad en los erorres y además de segui presente la heterocedasticidad.

Por lo que, para saber el comportamientos de las variables creemos modelos con cada una de las variabless y el gasto monetario, por lo que veamos lo siguiente:

```
## GRÁFICO DE CADA VARIABLE CON GASTO_MON
modGA<-lm(gasto_mon~alimentos, data=hogar2)
## Valores numéricos del modelo
kable(tidy(modGA),
      caption="Resumen del Modelo de regresión lineal
      de gasto con alimentos.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 23: Resumen del Modelo de regresión lineal de gasto con alimentos.

term	estimate	std.error	statistic	p.value
(Intercept)	5721.82907	129.553441	44.16578	0

alimentos	2.27061	0.007259	312.79728	0

```
## Extracción de datos del summary del modelo
modeloGA <- summary(modGA)
residual_standard_error <- modeloGA$sigma
multiple_r_squared <- modeloGA$r.squared
adjusted_r_squared <- modeloGA$adj.r.squared
f_statistic <- modeloGA$fstatistic[1]
p_value <- pf(modeloGA$fstatistic[1], modeloGA$fstatistic[2],
              modeloGA$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosGA<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                           "p-value" = p_value)
# Crear la tabla con kable
tablaGA <- kable(resultadosGA, caption = "Resultados del Modelo de
                           Regresión.",format = "latex",
                           booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaGA
```

Tabla 24: Resultados del Modelo de Regresión.

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	23990.89	0.5205971	0.5205918	97842.14	0

```
modGE<-lm(gasto_mon~educacion, data=hogar2)
## Valores numéricos del modelo
kable(tidy(modGE),
      caption="Resumen del Modelo de regresión lineal de
      gasto con educación.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 25: Resumen del Modelo de regresión lineal de gasto con educación.

term	estimate	std.error	statistic	p.value
(Intercept)	32210.998012	107.2380495	300.3691	0
educacion	2.190681	0.0136574	160.4027	0

```
## Extracción de datos del summary del modelo
modeloGE <- summary(modGE)
residual_standard_error <- modeloGE$sigma
multiple_r_squared <- modeloGE$r.squared
adjusted_r_squared <- modeloGE$adj.r.squared
f_statistic <- modeloGE$fstatistic[1]
p_value <- pf(modeloGE$fstatistic[1], modeloGE$fstatistic[2],
              modeloGE$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosGE<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                           "p-value" = p_value)
# Crear la tabla con kable
tablaGE <- kable(resultadosGE, caption = "Resultados del Modelo de
                           Regresión.",format = "latex",
                           booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaGE
```

Tabla 26: Resultados del Modelo de Regresión.

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	30559.74	0.2221294	0.2221208	25729.04	0

```
modGS<-lm(gasto_mon~salud, data=hogar2)
## Valores números del modelo
kable(tidy(modGS),
      caption="Resumen del Modelo de regresión
      lineal de gasto con salud.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 27: Resumen del Modelo de regresión lineal de gasto con salud.

term	estimate	std.error	statistic	p.value
(Intercept)	35011.854690	111.4742988	314.0801	0
salud	2.048975	0.0192098	106.6630	0

```
## Extracción de datos del summary del modelo
modeloGS <- summary(modGS)
residual_standard_error <- modeloGS$sigma
multiple_r_squared <- modeloGS$r.squared
adjusted_r_squared <- modeloGS$adj.r.squared
f_statistic <- modeloGS$fstatistic[1]
p_value <- pf(modeloGS$fstatistic[1], modeloGS$fstatistic[2],
              modeloGS$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosGS<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                           "p-value" = p_value)
# Crear la tabla con kable
tablaGS<- kable(resultadosGS, caption = "Resultados del Modelo de
                           Regresión.",format = "latex",
                           booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaGS
```

Tabla 28: Resultados del Modelo de Regresión.

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	32649.35	0.112114	0.1121041	11376.99	0

```r
modGTI<-lm(gasto_mon~tot_integ, data=hogar2)
## Valores númericos del modelo
kable(tidy(modGTI),
      caption="Resumen del Modelo de regresión
      lineal de gasto con total de integrantes.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 29: Resumen del Modelo de regresión lineal de gasto con total de integrantes.

term	estimate	std.error	statistic	p.value
(Intercept)	25150.747	246.7065	101.94603	0
tot_integ	3628.262	63.7754	56.89124	0

```r
## Extracción de datos del summary del modelo
modeloGTI <- summary(modGTI)
residual_standard_error <- modeloGTI$sigma
multiple_r_squared <- modeloGTI$r.squared
adjusted_r_squared <- modeloGTI$adj.r.squared
f_statistic <- modeloGTI$fstatistic[1]
p_value <- pf(modeloGTI$fstatistic[1], modeloGTI$fstatistic[2],
              modeloGTI$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosGTI<- data.frame("Residual Standard Error" = residual_standard_error,
                           "Multiple R-squared" = multiple_r_squared,
                           "Adjusted R-squared" = adjusted_r_squared,
                           "F-statistic" = f_statistic,
                          "p-value" = p_value)
# Crear la tabla con kable
tablaGTI<- kable(resultadosGTI, caption = "Resultados del Modelo de
                         Regresión.",format = "latex",
                         booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaGS
```

De los modelos tenemos que el gasto se explica bien con los alimentos dado a que su R cuadrada es de 0.5206 siendo una r significativa.

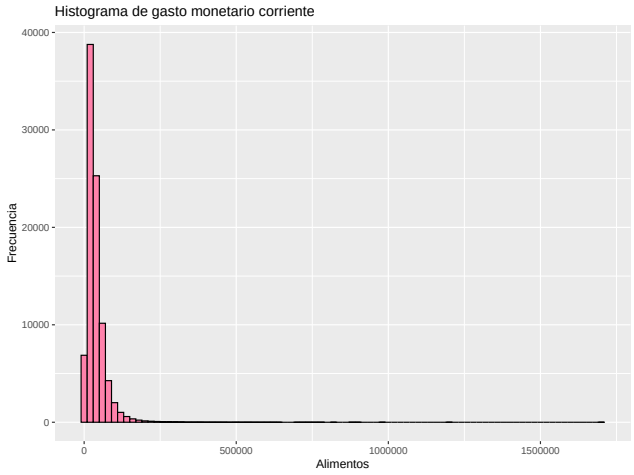
Tabla 30: Resultados del Modelo de Regresión.

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	32649.35	0.112114	0.1121041	11376.99	0

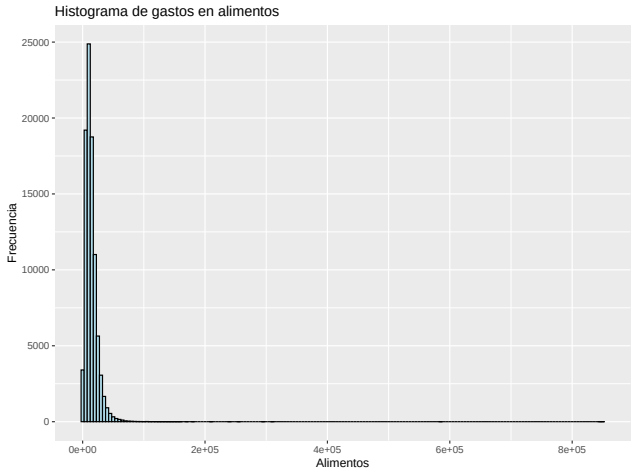
Histogramas de las variables

Realizaremos los histogramas de cada una de las variables con la finalidad de visualizar el comportamiento de cada una de ellas. Por lo que tenemos:

```
ggplot(data = hogar1, aes(x = gasto_mon)) +
  geom_histogram(binwidth = 20000 , fill = "#FF82AB", color = "black") +
  labs(title = "Histograma de gasto monetario corriente",
       x = "Alimentos", y = "Frecuencia")
```




```
ggplot(data = hogar1, aes(x = alimentos)) +
  geom_histogram(binwidth = 5000 , fill = "lightblue", color = "black") +
  labs(title = "Histograma de gastos en alimentos",
       x = "Alimentos", y = "Frecuencia")
```



A la vez, realicemos una tabla de frecuencias con la finalidad de ver mejor como se comportan los datos:

```
hogar1_filtrado <- subset(hogar1, gasto_mon>0)
hogarin <- cut(hogar1_filtrado$gasto_mon, breaks =30)
# Crear una tabla de frecuencias para los intervalos
tabla_frecuencias <- table(hogarin)
# Mostrar la tabla de frecuencias con intervalos
kable(tabla_frecuencias,caption="Tabla de frecuencias") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

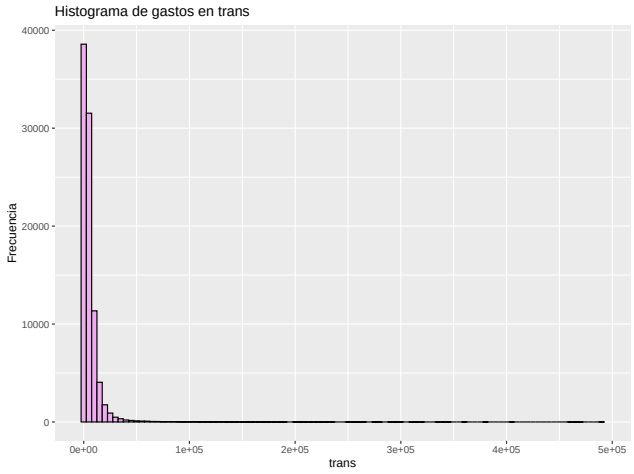
Tabla 31: Tabla de frecuencias

hogarin	Freq
(-1.68e+03,5.68e+04]	75347
(5.68e+04,1.14e+05]	12178
(1.14e+05,1.7e+05]	1738
(1.7e+05,2.27e+05]	441
(2.27e+05,2.84e+05]	165

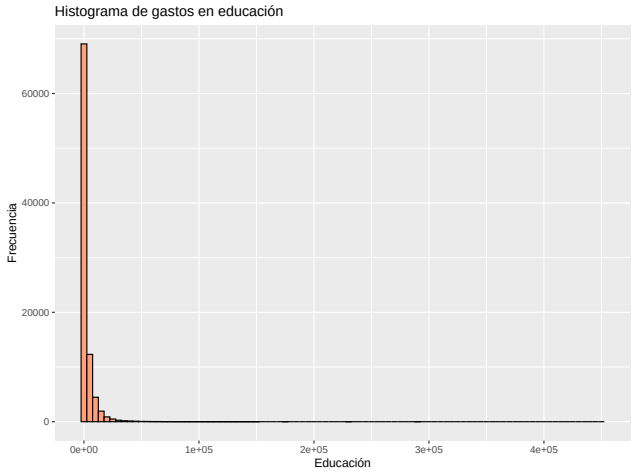
(2.84e+05,3.41e+05]	69
(3.41e+05,3.98e+05]	32
(3.98e+05,4.54e+05]	24
(4.54e+05,5.11e+05]	11
(5.11e+05,5.68e+05]	8
(5.68e+05,6.25e+05]	5
(6.25e+05,6.81e+05]	1
(6.81e+05,7.38e+05]	3
(7.38e+05,7.95e+05]	4
(7.95e+05,8.52e+05]	1
(8.52e+05,9.09e+05]	2
(9.09e+05,9.65e+05]	0
(9.65e+05,1.02e+06]	1
(1.02e+06,1.08e+06]	0
(1.08e+06,1.14e+06]	0
(1.14e+06,1.19e+06]	0
(1.19e+06,1.25e+06]	1
(1.25e+06,1.31e+06]	0
(1.31e+06,1.36e+06]	0
(1.36e+06,1.42e+06]	0
(1.42e+06,1.48e+06]	0

(1.48e+06,1.53e+06]	0
(1.53e+06,1.59e+06]	0
(1.59e+06,1.65e+06]	0
(1.65e+06,1.71e+06]	2

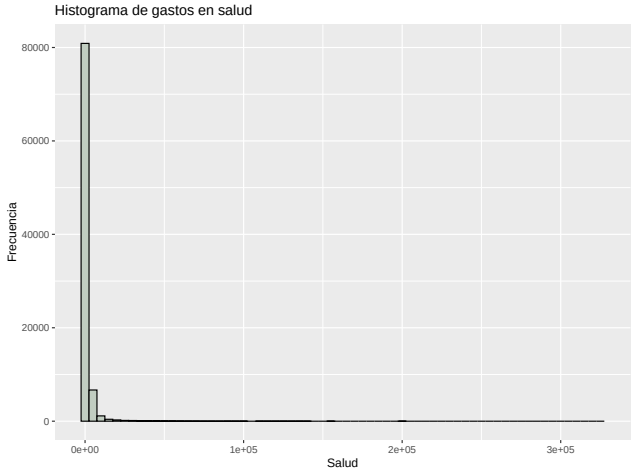
```
ggplot(data = hogar1, aes(x = trans)) +
  geom_histogram(binwidth = 5000 , fill = "#EEAEEE", color = "black") +
  labs(title = "Histograma de gastos en trans",
       x = "trans", y = "Frecuencia")
```



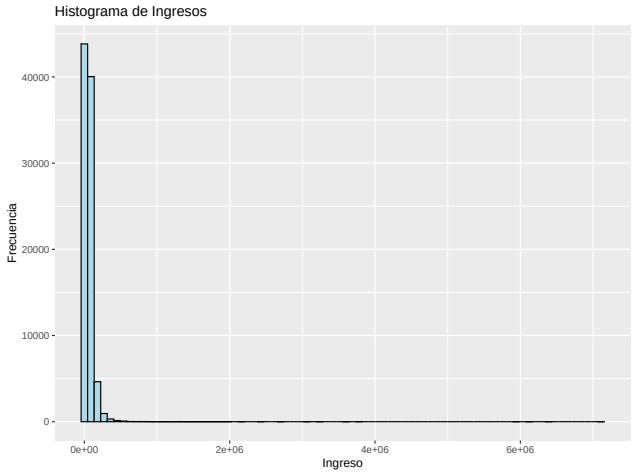

```
ggplot(data = hogar1, aes(x = educacion)) +
  geom_histogram(binwidth = 5000 , fill = "#FFA07A", color = "black") +
  labs(title = "Histograma de gastos en educación",
       x = "Educación", y = "Frecuencia")
```




```
ggplot(data = hogar1, aes(x = salud)) +
  geom_histogram(binwidth = 5000 , fill = "#C1CDC1", color = "black") +
  labs(title = "Histograma de gastos en salud",
       x = "Salud", y = "Frecuencia")
```

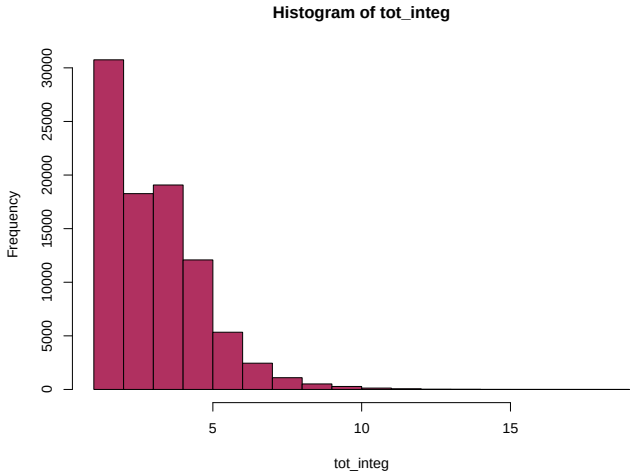



```
ggplot(data = hogar1, aes(x = ing_cor)) +
  geom_histogram(binwidth = 90000 , fill = "lightblue", color = "black") +
  labs(title = "Histograma de Ingresos",
       x = "Ingreso", y = "Frecuencia")
```



NOTA: La expresión de 2e+06 se refiere a 2 millones

```
hist(tot_integ, col='#B03060')
```



Los histogramas de las variables tienen un sesgo hacia la izquierda lo cual se debe a que la mayoría tiene como mínimo el 0 y como máximo apartir de 324547.7, por lo que para abordar este problema cortaremos a la muestra en deciles, en base a la variable *gasto_mon*, es así que primero obtenemos los deciles del gasto corriente monetario, para así cortar la muestra en 10 partes dependiendo del valor de sus deciles, por lo que tenemos:

```r
hogaro <- gasto_mon[order(gasto_mon)]
deci <- quantile(hogaro, probs = seq(0, 1, by = 0.1))
# Crear un dataframe con los resultados
resultados <- data.frame(Decil = paste0("Decil ", seq(0, 1, by = 0.1) * 10),
                         Valor = deci)
# Presentar los resultados en una tabla
kable(resultados, caption = "Deciles de gasto_mon")
```

Tabla 32: Deciles de gasto_mon

	Decil	Valor
0 %	Decil 0	0.00
10 %	Decil 1	11414.95
20 %	Decil 2	16375.04
30 %	Decil 3	20710.47
40 %	Decil 4	24984.98
50 %	Decil 5	29678.83
60 %	Decil 6	34902.75
70 %	Decil 7	41564.62
80 %	Decil 8	51533.16
90 %	Decil 9	69986.31
100 %	Decil 10	1703575.17

```r
##Cortado de muestras por decil
hogarD1<-subset(hogar1,gasto_mon>0 & gasto_mon<=11414.95)
hogarD2<-subset(hogar1, gasto_mon>11414.95 & gasto_mon<=16375.04)
hogarD3<-subset(hogar1, gasto_mon>16375.04 & gasto_mon<=20710.48)
hogarD4<-subset(hogar1, gasto_mon>20710.48 & gasto_mon<=24894.98)
```

```r
hogarD5<-subset(hogar1, gasto_mon>24894.98 & gasto_mon<=29678.83)
hogarD6<-subset(hogar1, gasto_mon>29678.83 & gasto_mon<= 34902.75)
hogarD7<-subset(hogar1, gasto_mon> 34902.75 & gasto_mon<= 41564.62)
hogarD8<-subset(hogar1, gasto_mon> 41564.62 & gasto_mon<=51533.16)
hogarD9<-subset(hogar1, gasto_mon>51533.16 & gasto_mon<= 69986.31)
hogarD10<-subset(hogar1, gasto_mon> 69986.31)
```

Ahora crearemos modelos con el primer decil, por lo que tenemos:

```r
modPP<-lm(gasto_mon~alimentos+educacion+salud+tot_integ+trans+ing_cor,
          data=hogarD1)
## Valores números del modelo
kable(tidy(modPP),
      caption="Resumen del Modelo de regresión lineal
      de gasto con el primer decil.")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 33: Resumen del Modelo de regresión lineal de gasto con el primer decil.

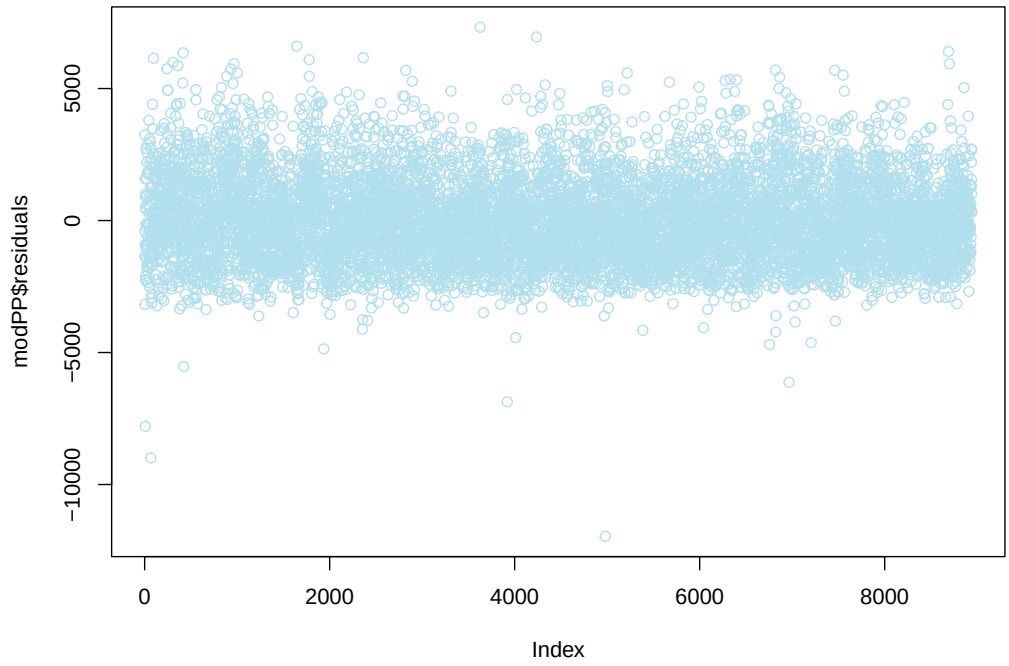
term	estimate	std.error	statistic	p.value
(Intercept)	2516.8070315	50.4297983	49.907140	0
alimentos	0.9102374	0.0082670	110.104989	0
educacion	0.8768495	0.0408786	21.450092	0
salud	0.9037234	0.0291760	30.974902	0
tot_integ	91.7665889	12.5976661	7.284412	0
trans	0.9840709	0.0184622	53.301946	0
ing_cor	0.0253804	0.0010442	24.305647	0

```r
## Extracción de datos del summary del modelo
modeloPP <- summary(modPP)
residual_standard_error <- modeloPP$sigma
multiple_r_squared <- modeloPP$r.squared
adjusted_r_squared <- modeloPP$adj.r.squared
f_statistic <- modeloPP$fstatistic[1]
p_value <- pf(modeloPP$fstatistic[1], modeloPP$fstatistic[2],
              modeloPP$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosPP<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                            "p-value" = p_value)
# Crear la tabla con kable
tablaPP<- kable(resultadosPP, caption = "Resultados del Modelo de
                            Regresión.",format = "latex",
                            booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaPP
```

Tabla 34: Resultados del Modelo de Regresión.

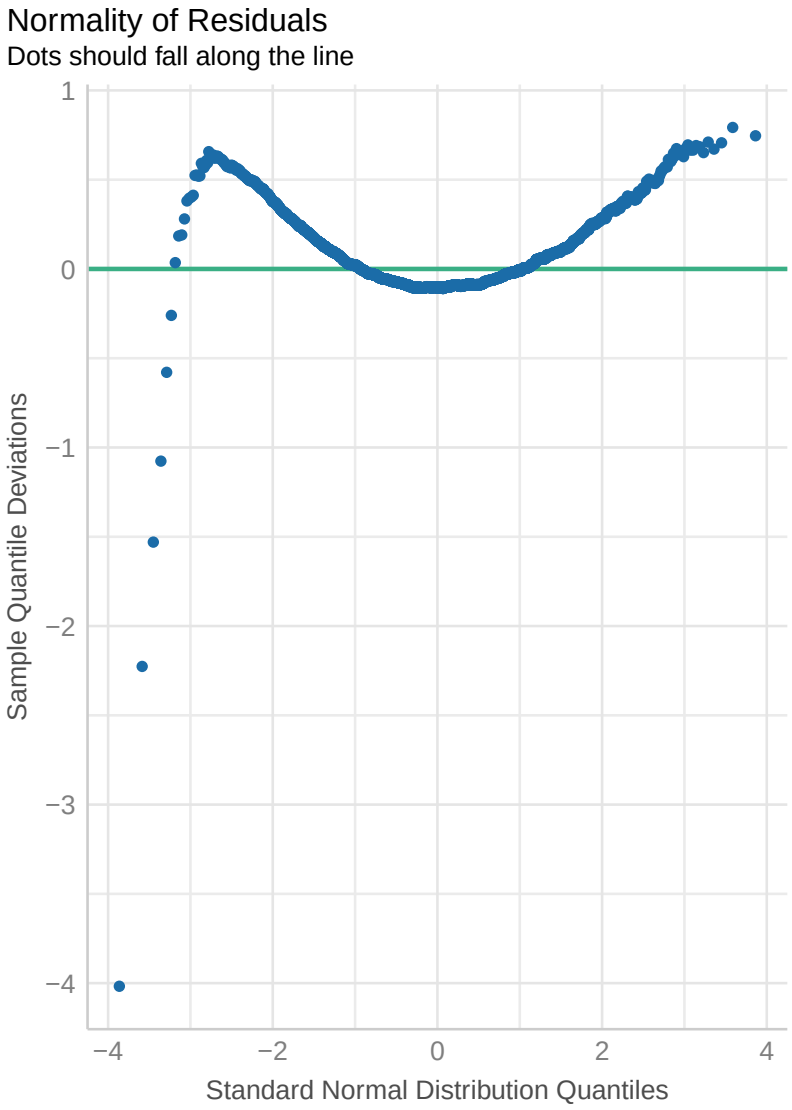
	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	1591.027	0.6382079	0.6379649	2626.917	0

```r
plot(modPP$residuals, col= '#B2DFEE')
```



Ahora veamos si cumple con la normalidad de los errores:

```
ehat1<-resid(modPP)
ebar1<-mean(ehat1)
sde<-sd(ehat1)
check_model(modPP, check = "qq")
```




```
jarque.bera.test(modPP$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP$residuals
## X-squared = 628.03, df = 2, p-value < 2.2e-16
```

El modelo aun no cumple, con la normalidad en los errores, por lo que realizaremos más modelos con la finalidad de encontrar el que mejor explica al gasto monetario corriente en el primer decil, por lo que:

```
modP1<-lm(gasto_mon~I(alimentos^2)+I(educacion^2)+I(salud^2)+
            I(tot_integ^2)+I(trans^2)+I(ing_cor^2), data=hogarD1)
modP2<-lm(gasto_mon~I(alimentos^3)+I(educacion^3)+I(salud^3)+
            I(tot_integ^3)++I(ing_cor^3)+I(trans^3), data=hogarD1)
modP3<-lm(gasto_mon~I(alimentos^4)+I(educacion^4)+I(salud^4)+
            I(tot_integ^4)++I(ing_cor^4)+I(trans^4), data=hogarD1)
modP4<-lm(gasto_mon~I(alimentos^2)+I(trans^2)+I(educacion^2)+
            I(salud^2)+I(tot_integ^2), data=hogarD1)
modP5<-lm(gasto_mon~I(alimentos^2)+I(trans^2)+I(tot_integ^2),
          data=hogarD1)
modP6<-lm(gasto_mon~I(alimentos^2)+I(trans^2)+I(tot_integ^2)
          +I(salud^2), data=hogarD1)
modP7<-lm(gasto_mon~I(alimentos^2)+I(trans^2)+I(tot_integ^2)
          +I(educacion^2), data=hogarD1)
```

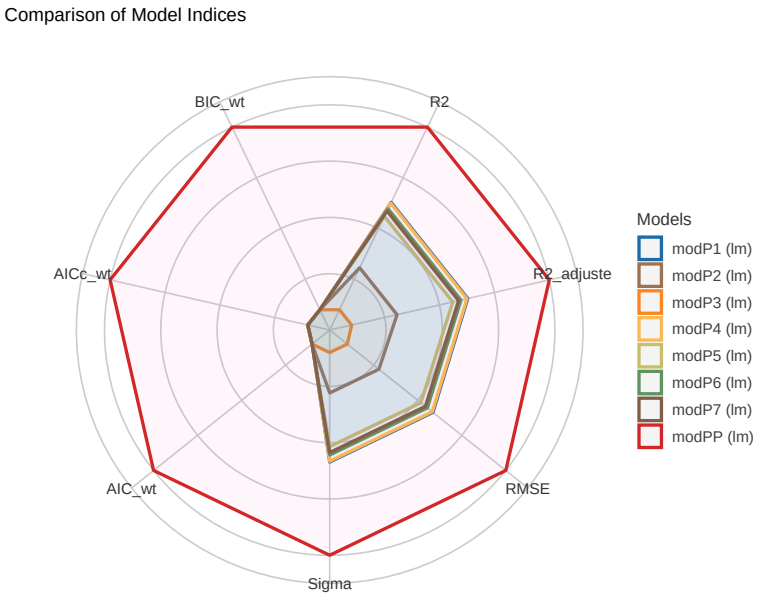
Para determinar cual es el mejor modelo

```
compare_performance(modPP, modP1,modP2,modP3,modP4,modP5,
                    modP6,modP7, rank=TRUE)
```

Name	Model	R2	R2_adjusted	RMSE	Sigma	AIC_wt	AICc_wt	BIC_wt	Performance_Score
modPP	lm	0.6382079	0.6379649	1590.404	1591.027	1	1	1	1.0000000
modP1	lm	0.4644886	0.4641290	1934.917	1935.675	0	0	0	0.3222217
modP4	lm	0.4627077	0.4624071	1938.132	1938.782	0	0	0	0.3198290
modP6	lm	0.4498185	0.4495723	1961.241	1961.790	0	0	0	0.3023165
modP7	lm	0.4446598	0.4444113	1970.414	1970.965	0	0	0	0.2953199
modP5	lm	0.4322397	0.4320492	1992.326	1992.772	0	0	0	0.2785837

Name	Model	R2	R2_adjusted	RMSE	Sigma	AIC_wt	AICc_wt	BIC_wt	Performance_Score
modP2	lm	0.3137484	0.3132875	2190.381	2191.239	0	0	0	0.1225580
modP3	lm	0.2170033	0.2164775	2339.688	2340.604	0	0	0	0.0000000

```
plot(compare_performance(modPP, modP1,modP2,modP3,modP4,modP5,
                         modP6,modP7, rank=TRUE))
```



El mejor modelo es el *modPP*, ahora veamos si su forma funcional es correcta:

```
resettest(modPP)
```

```
##
##  RESET test
##
## data:  modPP
```

```
## RESET = 459.1, df1 = 2, df2 = 8933, p-value < 2.2e-16
```

La prueba RESET tiene como finalidad saber si el modelo tiene una forma funcional correcta, dado que el valor p es menor a 0.05, por lo que decimos que el modelo esta mal especificado, es así que busquemos una forma funcional correcta.

Sin embargo, primero probaremos dicha forma funcional con los demás deciles, por lo que tenemos:

```
modPP1<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ, data=hogarD1)
## Valores númericos del modelo
kable(tidy(modPP1),
      caption="Resumen del Modelo de regresión lineal de gasto
      con el segundo decil")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 36: Resumen del Modelo de regresión lineal de gasto con el segundo decil

term	estimate	std.error	statistic	p.value
(Intercept)	3059.6310092	46.6831644	65.540352	0
alimentos	0.8951510	0.0085114	105.171444	0
educacion	0.8521832	0.0421930	20.197268	0
salud	0.9323508	0.0300988	30.976302	0
trans	1.0245133	0.0189841	53.966827	0
tot_integ	118.7475387	12.9561447	9.165345	0

```r
## Extracción de datos del summary del modelo
modeloPP1 <- summary(modPP1)
residual_standard_error <- modeloPP1$sigma
multiple_r_squared <- modeloPP1$r.squared
adjusted_r_squared <- modeloPP1$adj.r.squared
f_statistic <- modeloPP1$fstatistic[1]
p_value <- pf(modeloPP1$fstatistic[1], modeloPP1$fstatistic[2],
              modeloPP1$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosPP1<- data.frame("Residual Standard Error" = residual_standard_error,
                           "Multiple R-squared" = multiple_r_squared,
                           "Adjusted R-squared" = adjusted_r_squared,
                           "F-statistic" = f_statistic,"p-value" = p_value)
# Crear la tabla con kable
tablaPP1<- kable(resultadosPP1, caption = "Resultados del Modelo de
                         Regresión",format = "latex",
                         booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablaPP1
```

Tabla 37: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	1642.691	0.6142869	0.6140711	2846.296	0

```r
jarque.bera.test(modPP1$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP1$residuals
## X-squared = 644.85, df = 2, p-value < 2.2e-16
```

```
resettest(modPP1, power=2,type="fitted")
```

```
##
##  RESET test
##
## data:  modPP1
## RESET = 448.55, df1 = 1, df2 = 8935, p-value < 2.2e-16
```

```
modPP2<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD2)
jarque.bera.test(modPP2$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP2$residuals
## X-squared = 233.3, df = 2, p-value < 2.2e-16
```

```
resettest(modPP2, power=2,type="fitted")
```

```
##
##  RESET test
##
## data:  modPP2
```

```
## RESET = 359.25, df1 = 1, df2 = 9003, p-value < 2.2e-16
```

```
modPP3<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD3)
jarque.bera.test(modPP3$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP3$residuals
## X-squared = 385.8, df = 2, p-value < 2.2e-16
```

```
resettest(modPP3)
```

```
##
##  RESET test
##
## data:  modPP3
## RESET = 121.18, df1 = 2, df2 = 9002, p-value < 2.2e-16
```

```
modPP4<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD4)
jarque.bera.test(modPP4$residuals)
```

```
##
##  Jarque Bera Test
```

```
##
## data:  modPP4$residuals
## X-squared = 417.85, df = 2, p-value < 2.2e-16
```

```
resettest(modPP4)
```

```
##
##  RESET test
##
## data:  modPP4
## RESET = 119.54, df1 = 2, df2 = 8829, p-value < 2.2e-16
```

```
modPP5<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD5)
jarque.bera.test(modPP5$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP5$residuals
## X-squared = 426.23, df = 2, p-value < 2.2e-16
```

```
resettest(modPP5)
```

```
##
```

```
##  RESET test

##

## data:  modPP5

## RESET = 133.48, df1 = 2, df2 = 9175, p-value < 2.2e-16
```

```
modPP6<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD6)
jarque.bera.test(modPP6$residuals)
```

```
##

##  Jarque Bera Test

##

## data:  modPP6$residuals

## X-squared = 435.3, df = 2, p-value < 2.2e-16
```

```
resettest(modPP6)
```

```
##

##  RESET test

##

## data:  modPP6

## RESET = 140.37, df1 = 2, df2 = 9002, p-value < 2.2e-16
```

```
modPP7<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD7)
jarque.bera.test(modPP7$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP7$residuals
## X-squared = 425.7, df = 2, p-value < 2.2e-16
```

```
resettest(modPP7)
```

```
##
##  RESET test
##
## data:  modPP7
## RESET = 123.66, df1 = 2, df2 = 9002, p-value < 2.2e-16
```

```
modPP8<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
           data=hogarD8)
jarque.bera.test(modPP8$residuals)
```

```
##
##  Jarque Bera Test
##
```

```
## data:  modPP8$residuals
## X-squared = 416.16, df = 2, p-value < 2.2e-16
```

```
resettest(modPP8)
```

```
##
##  RESET test
##
## data:  modPP8
## RESET = 137.61, df1 = 2, df2 = 9002, p-value < 2.2e-16
```

```
modPP9<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ, data=hogarD9)
jarque.bera.test(modPP9$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP9$residuals
## X-squared = 371.98, df = 2, p-value < 2.2e-16
```

```
resettest(modPP9)
```

```
##
##  RESET test
```

```
##
## data:  modPP9
## RESET = 177.67, df1 = 2, df2 = 9002, p-value < 2.2e-16
```

```
modPP10<-lm(gasto_mon~alimentos+educacion+salud+trans+tot_integ,
            data=hogarD10)
jarque.bera.test(modPP10$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modPP10$residuals
## X-squared = 2641393, df = 2, p-value < 2.2e-16
```

```
resettest(modPP10)
```

```
##
##  RESET test
##
## data:  modPP10
## RESET = 206.29, df1 = 2, df2 = 9003, p-value < 2.2e-16
```

Se tiene que ningún modelo cumple con normalidad y su forma funcional es incorrecta, además se observo que el decil que no fue explicado con las variables del modelo fue el séptimo decil dado a que la variable dependiente se explica a un 2 % con las variables independientes.

Por otro lado, en el decimo decil el modelo explica a la variable dependiente al 48.12 %, sin embargo, para ninguno de los deciles tiene la fórmula funcional correcta y los errores siguen sin cumplir con normalidad.

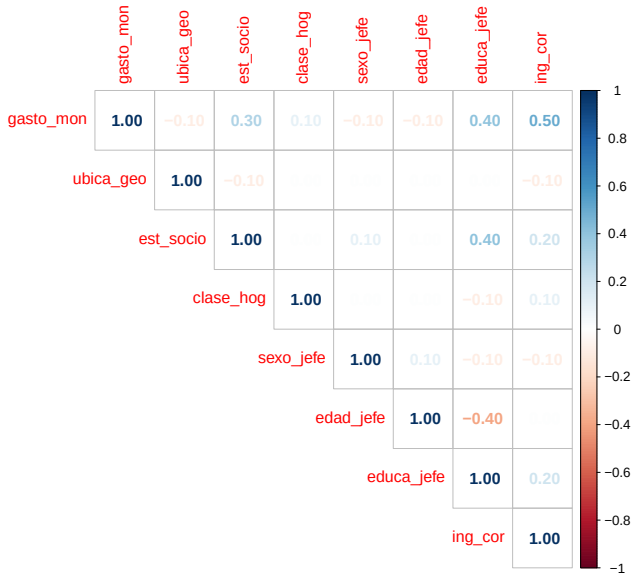
A.2 Reformulación

A continuación estudiaremos a las variables para construir un buen modelo, por ello primero observemos si estamos omitiendo variables importantes, es así que para mostrar mejor los datos separemos a los datos en dos variables, es así que tenemos:

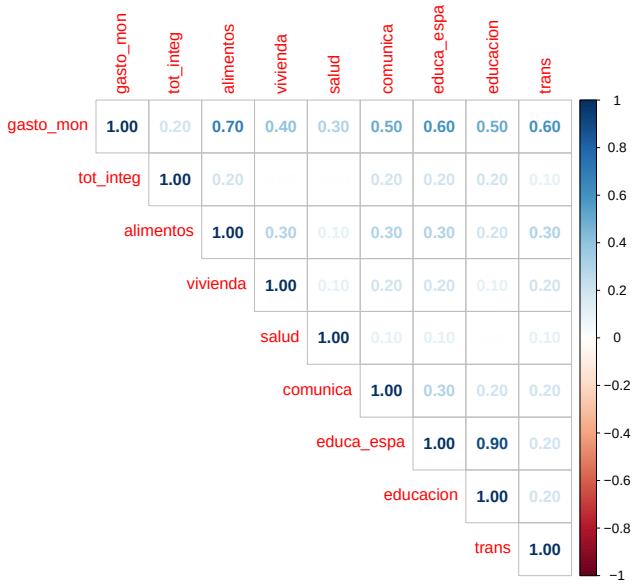
```
trans <- hogar[["transporte"]] - hogar[["comunica"]]
hogar4 <-hogar[c("gasto_mon","ubica_geo", "est_socio", "clase_hog",
                 "sexo_jefe", "edad_jefe", "educa_jefe", "ing_cor")]
partedos <-hogar[c("gasto_mon","tot_integ", "alimentos", "vivienda",
                   "salud","comunica", "educa_espa", "educacion")]
hogar5 <- cbind(partedos, trans = trans)
```

Se realizarán matrices de correlaciones con las variables, es así que tenemos:

```
correlacion<-round(cor(hogar4), 1)
corrplot(correlacion, method="number", type="upper" )
```




```
correlacion<-round(cor(hogar5), 1)
corrplot(correlacion, method="number", type="upper" )
```



Al observar las correlaciones, las variables que seleccionaremos son: **gasto_mon**,**alimentos**,**educacion**,**sal**

y **est_socio**, con las cuales crearemos el modelo siguiente:

```r
mrmtv<- lm(gasto_mon~est_socio +educa_jefe+tot_integ+
                    alimentos+vivienda+salud+comunica+educacion+
                    trans+ing_cor,data=hogar1)

## Valores numéricos del modelo
kable(tidy(mrmtv),
      caption="Resumen del Modelo de regresión lineal
      de gasto con el primer decil")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 38: Resumen del Modelo de regresión lineal de gasto con el primer decil

term	estimate	std.error	statistic	p.value
(Intercept)	-1103.9108129	129.7726148	-8.506500	0e+00
est_socio	-238.8468809	47.0507977	-5.076362	4e-07
educa_jefe	301.7476705	14.9279243	20.213639	0e+00
tot_integ	-489.9312388	19.9809910	-24.519867	0e+00
alimentos	1.3518735	0.0035110	385.040055	0e+00
vivienda	1.1713882	0.0066459	176.256848	0e+00
salud	1.1561631	0.0059785	193.386756	0e+00
comunica	1.5861301	0.0164033	96.695903	0e+00
educacion	1.0907496	0.0047543	229.425590	0e+00
trans	1.1276003	0.0030730	366.939306	0e+00
ing_cor	0.0444704	0.0004923	90.330104	0e+00

```r
## Extracción de datos del summary del modelo
modelo <- summary(mrmtv)
residual_standard_error <- modelo$sigma
multiple_r_squared <- modelo$r.squared
adjusted_r_squared <- modelo$adj.r.squared
f_statistic <- modelo$fstatistic[1]
p_value <- pf(modelo$fstatistic[1], modelo$fstatistic[2],
              modelo$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultados<- data.frame("Residual Standard Error" = residual_standard_error,
                                "Multiple R-squared" = multiple_r_squared,
                                "Adjusted R-squared" = adjusted_r_squared,
                                "F-statistic" = f_statistic,
                           "p-value" = p_value)
# Crear la tabla con kable
tablam<- kable(resultados, caption = "Resultados del Modelo de
                            Regresión",format = "latex",
                            booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablam
```

Tabla 39: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	9987.488	0.9169237	0.9169145	99434.57	0

Primero podemos observar que la r cuadrada es de 0.9169 lo cual puede indicar que exista colinealdad entre las variables, por ello primero observemos lo siguiente:

```r
vif(mrmtv)
```

```
##  est_socio educa_jefe  tot_integ  alimentos   vivienda      salud   comunica
##   1.331171   1.321332   1.140459   1.349837   1.176069   1.035084   1.310478
##  educacion      trans    ing_cor
##   1.134536   1.169882   1.343052
```

De lo anterior podemos observar que no se presenta colinealidad entre ls variables por lo que,

ahora veamos si nuestro modelo presenta heterocedasticidad, por lo que tenemos:

```
bptest(mrmtv)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mrmtv
## BP = 6456.1, df = 10, p-value < 2.2e-16
```

Podemos observar que si presenta heterocedasticidad, por lo que realizaremos nuestro analisis a traves de errores estándar robustos, por lo que tenemos:

```
cov1 <- hccm(mrmtv, type="hc1")
hogar.HC1 <- coeftest(mrmtv, vcov.=cov1)
kable(tidy(hogar.HC1),caption="Errores robustos")
```

Tabla 40: Errores robustos

term	estimate	std.error	statistic	p.value
(Intercept)	-1103.9108129	183.9124383	-6.002372	0.0000000
est_socio	-238.8468809	97.5592243	-2.448224	0.0143581
educa_jefe	301.7476705	31.5269669	9.571097	0.0000000
tot_integ	-489.9312388	54.5519994	-8.980995	0.0000000
alimentos	1.3518735	0.0340715	39.677590	0.0000000

term	estimate	std.error	statistic	p.value
vivienda	1.1713882	0.0340649	34.386995	0.0000000
salud	1.1561631	0.0189932	60.872577	0.0000000
comunica	1.5861301	0.0592772	26.757832	0.0000000
educacion	1.0907496	0.0166594	65.473478	0.0000000
trans	1.1276003	0.0130557	86.368126	0.0000000
ing_cor	0.0444704	0.0073611	6.041308	0.0000000

Sin embargo, los errores robustos muestran aun aumento en los valores de los errores, por lo que no es una técnica apropiada a utilizar.

De igual forma analicemos DFBETAS y DFFITS con la finalidad de observas si tenemos observaciones innfluyentes, por lo que tenemos:

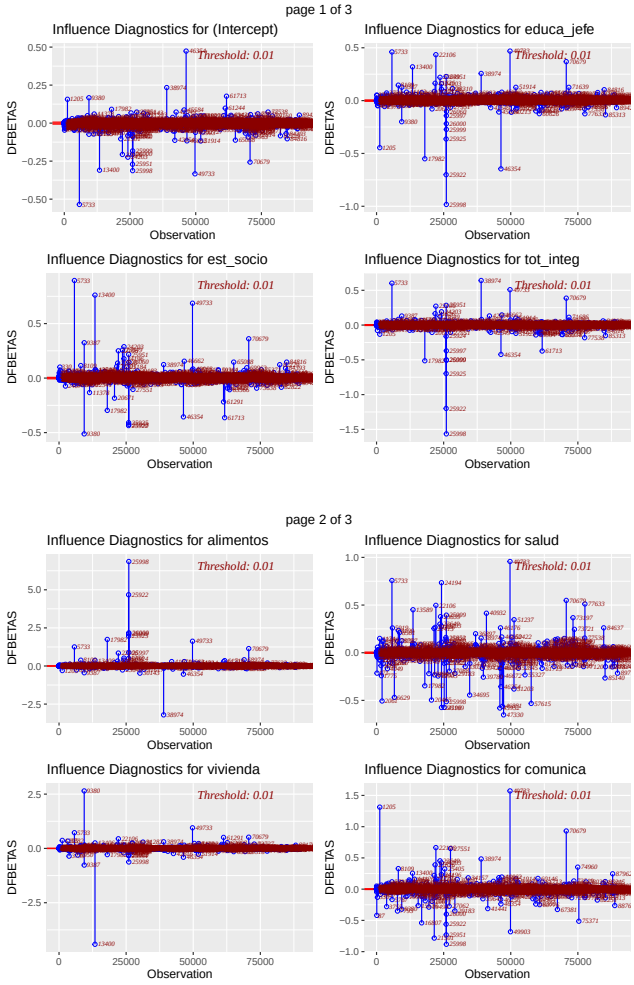
```
library(olsrr)
```

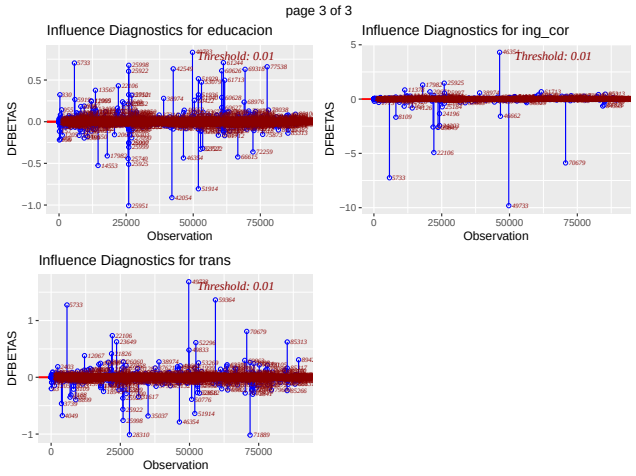
```
##
## Adjuntando el paquete: 'olsrr'
## The following object is masked from 'package:MASS':
##
##     cement
## The following object is masked from 'package:datasets':
```

##

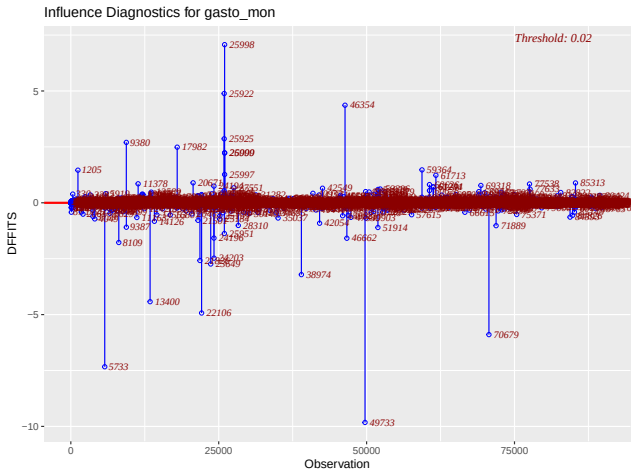
rivers

```
ols_plot_dfbetas(mrmtv)
```






```
ols_plot_dffits(mrmtv)
```



A.3 Modelo de propension marginal

Para trabajar el modelo de propensión marginal consideramos asignar cero cuando se da el caso de divisiones entre cero o se intente dividir cero entre una cantidad, para que así podamos correr el modelo, por lo que tenemos:

```r
viv <- hogar1$vivienda
ing <- hogar1$ing_cor
a <- viv/ing
a[is.infinite(a) | is.nan(a)] <- 0
#####
com <- hogar$comunica
b <- com/ing
b[is.infinite(b) | is.nan(b)] <- 0
#####
tra <- hogar1$trans
c<-tra/ing
c[is.infinite(c) | is.nan(c)] <- 0
#####
ali <- hogar1$alimentos
d <- ali/ing
d[is.infinite(d) | is.nan(d)] <- 0
####
edu <- hogar1$educacion
e<- edu/ing
e[is.infinite(e) | is.nan(e)] <- 0
#####
sal<-hogar1$salud
f <- sal/ing
f[is.infinite(f) | is.nan(f)] <- 0

##

gasto<-hogar1$gasto_mon
g <- gasto/ing
g[is.infinite(g) | is.nan(g)] <- 0

mpm<-lm(g~est_socio+educa_jefe+
             a+b+c+d+e+f+
             tot_integ, data=hogar1)

## Valores números del modelo
kable(tidy(mpm),
      caption="Resumen del Modelo de regresión lineal de gasto con el primer decil")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 41: Resumen del Modelo de regresión lineal de gasto con el primer decil

term	estimate	std.error	statistic	p.value

(Intercept)	0.0635493	0.0020013	31.75332	0
est_socio	-0.0088919	0.0006717	-13.23840	0
educa_jefe	0.0052596	0.0002118	24.83797	0
a	0.9920477	0.0031353	316.41484	0
b	1.6225603	0.0138800	116.89900	0
c	1.1324526	0.0028213	401.39161	0
d	1.1720471	0.0016668	703.17701	0
e	1.1011377	0.0050958	216.08660	0
f	1.0253525	0.0022120	463.53624	0
tot_integ	-0.0028864	0.0002842	-10.15770	0

```
## Extracción de datos del summary del modelo
modelomp <- summary(mpm)
residual_standard_error <- modelomp$sigma
multiple_r_squared <- modelomp$r.squared
adjusted_r_squared <- modelomp$adj.r.squared
f_statistic <- modelomp$fstatistic[1]
p_value <- pf(modelomp$fstatistic[1], modelomp$fstatistic[2],
              modelomp$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosmp<- data.frame("Residual Standard Error" = residual_standard_error,
                          "Multiple R-squared" = multiple_r_squared,
                          "Adjusted R-squared" = adjusted_r_squared,
                          "F-statistic" = f_statistic,"p-value" = p_value)
# Crear la tabla con kable
tablamp<- kable(resultadosmp, caption = "Resultados del Modelo de
                        Regresión",format = "latex",
                        booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablamp
```

Como se puede observar tenemos una r cuadrada muy grade por lo que veamos que las

Tabla 42: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	0.1465197	0.9512451	0.9512402	195306.9	0

variables no presenten colinealidad, de lo que se tiene:

```
vif(mpm)
```

```
##  est_socio educa_jefe          a          b          c          d          e
##   1.260492   1.235388   1.150783   1.173591   1.078841   1.215541   1.101634
##          f  tot_integ
##   1.009326   1.071734
```

De lo que se observa que no hay presencia de colinealidad, po rlo que ahora observemos si presenta heterocedasticidad el modelo y la normalidad de sus errores, por lo que tenemos:

```
bptest(mpm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mpm
## BP = 2741.6, df = 9, p-value < 2.2e-16
```

```r
jarque.bera.test(mpm$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  mpm$residuals
## X-squared = 4824376756, df = 2, p-value < 2.2e-16
```

Podemos obsevar que hay presencia de heterocedasticidad y no hay normalidad en los errores por lo que intentemos corregirla:

```r
bptest(mpm)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mpm
## BP = 2741.6, df = 9, p-value < 2.2e-16
```

```r
mpm2h <- lm(g~est_socio+educa_jefe+
            a+b+c+d+e+f+
            tot_integ, weights = 1/tot_integ, data=hogar1)
bptest(mpm2h)
```

```
##
```

```
##  studentized Breusch-Pagan test

##

## data:  mpm2h

## BP = 1698.8, df = 9, p-value < 2.2e-16
```

podemos observar que no se corrijió dicho problema dado a que su valor p sigue siendo muy pequeño, pero intentemos aplicar los errores robustos estandar.

```
cov1 <- hccm(mpm, type="hc1") #needs package car
hogar.HC1 <- coeftest(mpm, vcov.=cov1)
kable(tidy(hogar.HC1),caption="Errores robustos")
```

Tabla 43: Errores robustos

term	estimate	std.error	statistic	p.value
(Intercept)	0.0635493	0.0084944	7.481302	0
est_socio	-0.0088919	0.0012624	-7.043393	0
educa_jefe	0.0052596	0.0002694	19.525220	0
a	0.9920477	0.0343414	28.887816	0
b	1.6225603	0.1010471	16.057461	0
c	1.1324526	0.0159896	70.824154	0
d	1.1720471	0.0173339	67.616015	0
e	1.1011377	0.0244928	44.957635	0
f	1.0253525	0.0085893	119.374878	0

term	estimate	std.error	statistic	p.value
tot_integ	-0.0028864	0.0004116	-7.012036	0

De igual forma no es una manera funcional para obtener estimaciones más precisas.

A.3.1 Modelo de reescalación

La aplicación de un modelo de reescalación busca mejorar la coherencia y el desempeño de los datos en el análisis o en el modelo predictivo para ver si esos problemas iniciales se corrigen, por ello tenemos:

```
modr<- lm(gasto_mon~est_socio+ educa_jefe+tot_integ+
            I(alimentos/100)+I(vivienda/100)+I(salud/100)+
            I(comunica/100)+I(educacion/100)+I(trans/100)+
            I(ing_cor/100),data=hogar1)
## Valores números del modelo
kable(tidy(modr),
      caption="Resumen del Modelo de reescalación")%>%
kable_styling(latex_options = c("striped", "hold_position"))
```

Tabla 44: Resumen del Modelo de reescalación

term	estimate	std.error	statistic	p.value
(Intercept)	-1103.910813	129.7726148	-8.506500	0e+00
est_socio	-238.846881	47.0507977	-5.076362	4e-07
educa_jefe	301.747671	14.9279243	20.213639	0e+00
tot_integ	-489.931239	19.9809910	-24.519867	0e+00

I(alimentos/100)	135.187346	0.3510994	385.040055	0e+00
I(vivienda/100)	117.138817	0.6645916	176.256848	0e+00
I(salud/100)	115.616306	0.5978502	193.386756	0e+00
I(comunica/100)	158.613010	1.6403281	96.695903	0e+00
I(educacion/100)	109.074963	0.4754263	229.425590	0e+00
I(trans/100)	112.760033	0.3072989	366.939306	0e+00
I(ing_cor/100)	4.447038	0.0492310	90.330104	0e+00

```
## Extracción de datos del summary del modelo
modelomr <- summary(modr)
residual_standard_error <- modelomr$sigma
multiple_r_squared <- modelomr$r.squared
adjusted_r_squared <- modelomr$adj.r.squared
f_statistic <- modelomr$fstatistic[1]
p_value <- pf(modelomr$fstatistic[1], modelomr$fstatistic[2],
              modelomr$fstatistic[3],lower.tail = FALSE)
# Crear una tabla con los resultados
resultadosmr<- data.frame("Residual Standard Error" = residual_standard_error,
                          "Multiple R-squared" = multiple_r_squared,
                          "Adjusted R-squared" = adjusted_r_squared,
                          "F-statistic" = f_statistic,"p-value" = p_value)
# Crear la tabla con kable
tablamr<- kable(resultadosmr, caption = "Resultados del Modelo de
                        Regresión",format = "latex",
                        booktabs = TRUE, align = "c") %>%
kable_styling(latex_options = c("striped", "hold_position"))
tablamr
```

Tabla 45: Resultados del Modelo de Regresión

	Residual.Standard.Error	Multiple.R.squared	Adjusted.R.squared	F.statistic	p.value
value	9987.488	0.9169237	0.9169145	99434.57	0

Se realizaron las pruebas para comprobar si con la técnica de reescalación se corregían los errores presentados en modelos anteriores, obteniendo lo siguiente:

```r
bptest(modr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modr
## BP = 6456.1, df = 10, p-value < 2.2e-16
```

```r
jarque.bera.test(modr$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  modr$residuals
## X-squared = 697262560, df = 2, p-value < 2.2e-16
```

El modelo presentó heterocedasticidad, al igual que los modelos anteriores, y tampoco cumple con el supuesto de normalidad en los errores.

A.4 Modelo Tobit

Como hemos observado a lo largo de dicho documento, podemos darnos cuenta que no hemos logrado encontrar un modelo suficiente bueno,es asi que al observar otra vez los graficos de

las vairables podemos observar que presentan un sesgo hacia la izquierda,por lo que, podemos utilizar un modelo TOBIT, es asi que tenemos:

```
h2.tobit <- tobit(gasto_mon~est_socio +educa_jefe+tot_integ+
                  alimentos+vivienda+salud+comunica+educacion+
                  trans+ing_cor,data=hogar1)

summary(h2.tobit)
```

```
##
## Call:
## tobit(formula = gasto_mon ~ est_socio + educa_jefe + tot_integ +
##     alimentos + vivienda + salud + comunica + educacion + trans +
##     ing_cor, data = hogar1)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##          90102             69          90033              0
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -1.198e+03  1.241e+02   -9.652  < 2e-16 ***
## est_socio   -2.190e+02  4.495e+01   -4.872 1.11e-06 ***
## educa_jefe   3.168e+02  1.401e+01   22.606  < 2e-16 ***
## tot_integ   -4.733e+02  1.882e+01  -25.149  < 2e-16 ***
```

```
## alimentos    1.351e+00  1.187e-03 1138.059  < 2e-16 ***
## vivienda     1.175e+00  6.349e-03  185.095  < 2e-16 ***
## salud        1.159e+00  5.697e-03  203.490  < 2e-16 ***
## comunica     1.591e+00  1.575e-02  101.039  < 2e-16 ***
## educacion    1.093e+00  4.513e-03  242.301  < 2e-16 ***
## trans        1.130e+00  2.905e-03  389.024  < 2e-16 ***
## ing_cor      4.218e-02  1.053e-04  400.681  < 2e-16 ***
## Log(scale)   9.170e+00  2.357e-03 3891.305  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 9608
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 30
## Log-likelihood: -9.538e+05 on 12 Df
## Wald-statistic: 8.817e+06 on 10 Df, p-value: < 2.22e-16
```

Nota: Los 69 datos censurados a la izquierda no tienen nada de datos en ningun rubro respecto al gasto en los hogares

A.4.1 Efectos marginales con un hogar representativo en las medias

Se calcularan los efectos marginales en un hogar representativo en la media de los datos, por lo que tenemos:

```
mean(hogar1$est_socio)
```

```
## [1] 2.07092
```

```
xestsoc <- 2
mean(hogar1$educa_jefe)
```

```
## [1] 5.803856
```

```
xeduc<-6
mean(hogar1$tot_integ)
```

```
## [1] 3.435373
```

```
xtoti<-3
xalim<-mean(hogar1$alimentos)
xviv<-mean(hogar1$vivienda)
xsalu<-mean(hogar1$salud)
xcomu<-mean(hogar1$comunica)
xedu<-mean(hogar1$educacion)
xtran<-mean(hogar1$trans)
xing<-mean(hogar1$ing_cor)
bInt <- coef(h2.tobit)[[1]]
best<-coef(h2.tobit)[[2]]
beduj<-coef(h2.tobit)[[3]]
btot<-coef(h2.tobit)[[4]]
bali<-coef(h2.tobit)[[5]]
bviv<-coef(h2.tobit)[[6]]
```

```
bsal<-coef(h2.tobit)[[7]]
bcom<-coef(h2.tobit)[[8]]
beduc<-coef(h2.tobit)[[9]]
btran<-coef(h2.tobit)[[10]]
bing<-coef(h2.tobit)[[11]]
bSigma <- h2.tobit$scale
Phactor <- pnorm((bInt+best*xestsoc+beduj*xeduc+btot*xtoti+
                    bali*xalim+bviv*xviv+bsal*xsalu+
                    bcom*xcomu+beduc*xedu+btran*xtran+
                    bing*xing)/bSigma)
EMint <- bInt*Phactor
EMest <- best*Phactor
EMedj <- beduj*Phactor
EMtoti<-btot*Phactor
EMali<-bali*Phactor
EMviv<-bviv*Phactor
EMsal<-bsal*Phactor
EMcoc<-bcom*Phactor
EMeduc<-beduc*Phactor
EMtrans<-btran*Phactor
EMing<-bing*Phactor

# Crea un data frame con los valores calculados
data <- data.frame(
  Concepto = c("Int", "Est_socio", "Educa_jefe", "Tot_integ",
               "Alimentos", "Vivienda", "Salud", "Comunicaciones", "Educación",
               "Transporte", "Ing_cor"),
  Valor = c(EMint, EMest, EMedj, EMtoti, EMali, EMviv,
            EMsal, EMcoc, EMeduc, EMtrans, EMing))
kable(data, caption = "Efectos marginale modelo TOBIT")
```

Tabla 46: Efectos marginale modelo TOBIT

Concepto	Valor
Int	-1197.8405169
Est_socio	-218.9892822
Educa_jefe	316.7931957
Tot_integ	-473.2407554
Alimentos	1.3508326

Concepto	Valor
Vivienda	1.1751407
Salud	1.1592013
Comunicaciones	1.5913690
Educación	1.0934182
Transporte	1.1300336
Ing_cor	0.0421801

A.4.2 Efectos marginales con hogares representativos en los deciles

Se calcularán los efectos marginales con hogares representativos en la media de cada uno de los deciles, para conocer mas a fondo sobre los hogares:

DECIL 1

```
mean(hogarD1$est_socio)
```

```
## [1] 1.68743
```

```
xestsoc1 <- 2
mean(hogarD1$educa_jefe)
```

```
## [1] 3.882241
```

```r
xeduc1<-4
mean(hogarD1$tot_integ)
```

```
## [1] 2.12391
```

```r
xtoti1<-2
xalim1<-mean(hogarD1$alimentos)
xviv1<-mean(hogarD1$vivienda)
xsalu1<-mean(hogarD1$salud)
xcomu1<-mean(hogarD1$comunica)
xedu1<-mean(hogarD1$educacion)
xtran1<-mean(hogarD1$trans)
xing1<-mean(hogarD1$ing_cor)
Phactor1 <- pnorm((bInt+best*xestsoc1+beduj*xeduc1+btot*xtoti1+
                   bali*xalim1+bviv*xviv1+bsal*xsalu1+
                   bcom*xcomu1+beduc*xedu1+btran*xtran1+
                   bing*xing1)/bSigma)
EMint1<- bInt*Phactor1
EMest1 <- best*Phactor1
EMedj1 <- beduj*Phactor1
EMtoti1<-btot*Phactor1
EMali1<-bali*Phactor1
EMviv1<-bviv*Phactor1
EMsal1<-bsal*Phactor1
EMcoc1<-bcom*Phactor1
EMeduc1<-beduc*Phactor1
EMtrans1<-btran*Phactor1
EMing1<-bing*Phactor1
data1 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud","Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil1 = c(EMint1, EMest1, EMedj1, EMtoti1, EMali1,
             EMviv1, EMsal1,EMcoc1, EMeduc1, EMtrans1,EMing1))
```

DECIL 2

```r
mean(hogarD2$est_socio)
```

```
## [1] 1.781687
```

```r
xestsoc2 <- 2
mean(hogarD2$educa_jefe)
```

```
## [1] 4.645061
```

```r
xeduc2<-5
mean(hogar2$tot_integ)
```

```
## [1] 3.435373
```

```r
xtoti2<-3
xalim2<-mean(hogarD2$alimentos)
xviv2<-mean(hogarD2$vivienda)
xsalu2<-mean(hogarD2$salud)
xcomu2<-mean(hogarD2$comunica)
xedu2<-mean(hogarD2$educacion)
xtran2<-mean(hogarD2$trans)
xing2<-mean(hogarD2$ing_cor)
Phactor2 <- pnorm((bInt+best*xestsoc2+beduj*xeduc2+btot*xtoti2+
                   bali*xalim2+bviv*xviv2+bsal*xsalu2+
                   bcom*xcomu2+beduc*xedu2+btran*xtran2+
                   bing*xing2)/bSigma)
EMint2<- bInt*Phactor2
EMest2 <- best*Phactor2
EMedj2 <- beduj*Phactor2
EMtoti2<-btot*Phactor2
EMali2<-bali*Phactor2
EMviv2<-bviv*Phactor2
EMsal2<-bsal*Phactor2
EMcoc2<-bcom*Phactor2
EMeduc2<-beduc*Phactor2
EMtrans2<-btran*Phactor2
EMing2<-bing*Phactor2
data2 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil2 = c(EMint2, EMest2, EMedj2, EMtoti2, EMali2,
             EMviv2, EMsal2,EMcoc2, EMeduc2, EMtrans2,EMing2))
```

DECIL 3

```r
mean(hogarD3$est_socio)
```

```
## [1] 1.856715
```

```r
xestsoc3 <- 2
mean(hogarD3$educa_jefe)
```

```
## [1] 5.097447
```

```r
xeduc3<-5
mean(hogarD3$tot_integ)
```

```
## [1] 3.073363
```

```r
xtoti3<-3
xalim3<-mean(hogarD3$alimentos)
xviv3<-mean(hogarD3$vivienda)
xsalu3<-mean(hogarD3$salud)
xcomu3<-mean(hogarD3$comunica)
xedu3<-mean(hogarD3$educacion)
xtran3<-mean(hogarD3$trans)
xing3<-mean(hogarD3$ing_cor)
Phactor3 <- pnorm((bInt+best*xestsoc3+beduj*xeduc3+btot*xtoti3+
                   bali*xalim3+bviv*xviv3+bsal*xsalu3+
                   bcom*xcomu3+beduc*xedu3+btran*xtran3+
                   bing*xing3)/bSigma)
EMint3<- bInt*Phactor3
EMest3 <- best*Phactor3
EMedj3 <- beduj*Phactor3
EMtoti3<-btot*Phactor3
EMali3<-bali*Phactor3
EMviv3<-bviv*Phactor3
EMsal3<-bsal*Phactor3
EMcoc3<-bcom*Phactor3
EMeduc3<-beduc*Phactor3
EMtrans3<-btran*Phactor3
EMing3<-bing*Phactor3
```

```
data3 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil3 = c(EMint3, EMest3, EMedj3, EMtoti3, EMali3,
             EMviv3, EMsal3,EMcoc3, EMeduc3, EMtrans3,EMing3))
```

DECIL 4

```
mean(hogarD4$est_socio)
```

```
## [1] 1.915469
```

```
xestsoc4 <- 2
mean(hogarD4$educa_jefe)
```

```
## [1] 5.340048
```

```
xeduc4<-5
mean(hogarD4$tot_integ)
```

```
## [1] 3.317189
```

```
xtoti4<-3
xalim4<-mean(hogarD4$alimentos)
xviv4<-mean(hogarD4$vivienda)
xsalu4<-mean(hogarD4$salud)
xcomu4<-mean(hogarD4$comunica)
xedu4<-mean(hogarD4$educacion)
xtran4<-mean(hogarD4$trans)
xing4<-mean(hogarD4$ing_cor)
Phactor4 <- pnorm((bInt+best*xestsoc4+beduj*xeduc4+btot*xtoti4+
                   bali*xalim4+bviv*xviv4+bsal*xsalu4+
                   bcom*xcomu4+beduc*xedu4+btran*xtran4+
                   bing*xing4)/bSigma)
```

```
EMint4<- bInt*Phactor4
EMest4 <- best*Phactor4
EMedj4 <- beduj*Phactor
EMtoti4<-btot*Phactor4
EMali4<-bali*Phactor4
EMviv4<-bviv*Phactor4
EMsal4<-bsal*Phactor4
EMcoc4<-bcom*Phactor4
EMeduc4<-beduc*Phactor4
EMtrans4<-btran*Phactor4
EMing4<-bing*Phactor4
# Crea un data frame con los valores calculados
data4 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil4 = c(EMint4, EMest4, EMedj4, EMtoti4, EMali4, EMviv4, EMsal4,
             EMcoc4, EMeduc4, EMtrans4,EMing4))
```

DECIL 5

```
mean(hogarD5$est_socio)
```

```
## [1] 1.99009
```

```
xestsoc5 <- 2
mean(hogarD5$educa_jefe)
```

```
## [1] 5.623652
```

```
xeduc5<-6
mean(hogarD5$tot_integ)
```

```
## [1] 3.52815
```

```
xtoti5<-4
xalim5<-mean(hogarD5$alimentos)
xviv5<-mean(hogarD5$vivienda)
xsalu5<-mean(hogarD5$salud)
xcomu5<-mean(hogarD5$comunica)
xedu5<-mean(hogarD5$educacion)
xtran5<-mean(hogarD5$trans)
xing5<-mean(hogarD5$ing_cor)
Phactor5 <- pnorm((bInt+best*xestsoc5+beduj*xeduc5+btot*xtoti5+
                   bali*xalim5+bviv*xviv5+bsal*xsalu5+
                   bcom*xcomu5+beduc*xedu5+btran*xtran5+
                   bing*xing5)/bSigma)
EMint5<- bInt*Phactor5
EMest5 <- best*Phactor5
EMedj5 <- beduj*Phactor5
EMtoti5<-btot*Phactor5
EMali5<-bali*Phactor5
EMviv5<-bviv*Phactor5
EMsal5<-bsal*Phactor5
EMcoc5<-bcom*Phactor5
EMeduc5<-beduc*Phactor5
EMtrans5<-btran*Phactor5
EMing5<-bing*Phactor5
# Crea un data frame con los valores calculados
data5 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil5 = c(EMint5, EMest5, EMedj5, EMtoti5, EMali5, EMviv5,
             EMsal5,EMcoc5, EMeduc5, EMtrans5,EMing5))
```

DECIL 6

```
mean(hogarD6$est_socio)
```

```
## [1] 2.080244
```

```
xestsoc6 <- 2
mean(hogarD6$educa_jefe)
```

```
## [1] 5.904772
```

```
xeduc6<-6
mean(hogarD6$tot_integ)
```

```
## [1] 3.7
```

```
xtoti6<-4
xalim6<-mean(hogarD6$alimentos)
xviv6<-mean(hogarD6$vivienda)
xsalu6<-mean(hogarD6$salud)
xcomu6<-mean(hogarD6$comunica)
xedu6<-mean(hogarD6$educacion)
xtran6<-mean(hogarD6$trans)
xing6<-mean(hogarD6$ing_cor)
Phactor6 <- pnorm((bInt+best*xestsoc6+beduj*xeduc6+btot*xtoti6+
                     bali*xalim6+bviv*xviv6+bsal*xsalu6+
                     bcom*xcomu6+beduc*xedu6+btran*xtran6+
                     bing*xing6)/bSigma)
EMint6<- bInt*Phactor6
EMest6 <- best*Phactor6
EMedj6 <- beduj*Phactor6
EMtoti6<-btot*Phactor6
EMali6<-bali*Phactor6
EMviv6<-bviv*Phactor6
EMsal6<-bsal*Phactor6
EMcoc6<-bcom*Phactor6
EMeduc6<-beduc*Phactor6
EMtrans6<-btran*Phactor6
EMing6<-bing*Phactor6
# Crea un data frame con los valores calculados
data6 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil6 = c(EMint6, EMest6, EMedj6, EMtoti6, EMali6, EMviv6,
             EMsal6,EMcoc6, EMeduc6, EMtrans6,EMing6))
```

DECIL 7

```
mean(hogarD7$est_socio)
```

```
## [1] 2.137736
```

```r
xestsoc7 <- 2
mean(hogarD7$educa_jefe)
```

```
## [1] 6.147281
```

```r
xeduc7<-6
mean(hogarD7$tot_integ)
```

```
## [1] 3.831632
```

```r
xtoti7<-4
xalim7<-mean(hogarD7$alimentos)
xviv7<-mean(hogarD7$vivienda)
xsalu7<-mean(hogarD7$salud)
xcomu7<-mean(hogarD7$comunica)
xedu7<-mean(hogarD7$educacion)
xtran7<-mean(hogarD7$trans)
xing7<-mean(hogarD7$ing_cor)
Phactor7 <- pnorm((bInt+best*xestsoc7+beduj*xeduc7+btot*xtoti7+
                   bali*xalim7+bviv*xviv7+bsal*xsalu7+
                   bcom*xcomu7+beduc*xedu7+btran*xtran7+
                   bing*xing7)/bSigma)
EMint7<- bInt*Phactor7
EMest7 <- best*Phactor7
EMedj7 <- beduj*Phactor7
EMtoti7<-btot*Phactor7
EMali7<-bali*Phactor7
EMviv7<-bviv*Phactor7
EMsal7<-bsal*Phactor7
EMcoc7<-bcom*Phactor7
EMeduc7<-beduc*Phactor7
EMtrans7<-btran*Phactor7
EMing7<-bing*Phactor7
# Crea un data frame con los valores calculados
data7 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil7 = c(EMint7, EMest7, EMedj7, EMtoti7, EMali7, EMviv7,
             EMsal7,EMcoc7, EMeduc7, EMtrans7,EMing7))
```

DECIL 8

```r
mean(hogarD8$est_socio)
```

```
## [1] 2.229412
```

```r
xestsoc8 <- 2
mean(hogarD8$educa_jefe)
```

```
## [1] 6.472586
```

```r
xeduc8<-6
mean(hogarD8$tot_integ)
```

```
## [1] 3.996448
```

```r
xtoti8<-4
xalim8<-mean(hogarD8$alimentos)
xviv8<-mean(hogarD8$vivienda)
xsalu8<-mean(hogarD8$salud)
xcomu8<-mean(hogarD8$comunica)
xedu8<-mean(hogarD8$educacion)
xtran8<-mean(hogarD8$trans)
xing8<-mean(hogarD8$ing_cor)
Phactor8 <- pnorm((bInt+best*xestsoc8+beduj*xeduc8+btot*xtoti8+
                   bali*xalim8+bviv*xviv8+bsal*xsalu8+
                   bcom*xcomu8+beduc*xedu8+btran*xtran8+
                   bing*xing8)/bSigma)
EMint8<- bInt*Phactor8
EMest8 <- best*Phactor8
EMedj8 <- beduj*Phactor8
EMtoti8<-btot*Phactor8
EMali8<-bali*Phactor8
EMviv8<-bviv*Phactor8
EMsal8<-bsal*Phactor8
EMcoc8<-bcom*Phactor8
EMeduc8<-beduc*Phactor8
EMtrans8<-btran*Phactor8
EMing8<-bing*Phactor8
# Crea un data frame con los valores calculados
```

```
data8 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil8 = c(EMint8, EMest8, EMedj8, EMtoti8, EMali8,
             EMviv8, EMsal8, EMcoc8, EMeduc8, EMtrans8,EMing8))
```

DECIL 9

```
mean(hogarD9$est_socio)
```

```
## [1] 2.378912
```

```
xestsoc9 <- 2
mean(hogarD9$educa_jefe)
```

```
## [1] 6.990788
```

```
xeduc9<-7
mean(hogarD9$tot_integ)
```

```
## [1] 4.09889
```

```
xtoti9<-4
xalim9<-mean(hogarD9$alimentos)
xviv9<-mean(hogarD9$vivienda)
xsalu9<-mean(hogarD9$salud)
xcomu9<-mean(hogarD9$comunica)
xedu9<-mean(hogarD9$educacion)
xtran9<-mean(hogarD9$trans)
xing9<-mean(hogarD9$ing_cor)
Phactor9 <- pnorm((bInt+best*xestsoc9+beduj*xeduc9+btot*xtoti9+
                   bali*xalim9+bviv*xviv9+bsal*xsalu9+
                   bcom*xcomu9+beduc*xedu9+btran*xtran9+
                   bing*xing9)/bSigma)
```

```
EMint9<- bInt*Phactor9
EMest9 <- best*Phactor9
EMedj9 <- beduj*Phactor9
EMtoti9<-btot*Phactor9
EMali9<-bali*Phactor9
EMviv9<-bviv*Phactor9
EMsal9<-bsal*Phactor9
EMcoc9<-bcom*Phactor9
EMeduc9<-beduc*Phactor9
EMtrans9<-btran*Phactor9
EMing9<-bing*Phactor9
# Crea un data frame con los valores calculados
data9 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                 "Tot_integ", "Alimentos", "Vivienda",
                                 "Salud", "Comunicaciones", "Educación",
                                 "Transporte", "Ing_cor"),
  Decil9 = c(EMint9, EMest9, EMedj9, EMtoti9, EMali9, EMviv9,
             EMsal9,EMcoc9, EMeduc9, EMtrans9,EMing9))
```

DECIL 10

```
mean(hogarD10$est_socio)
```

```
## [1] 2.647653
```

```
xestsoc10 <- 3
mean(hogarD10$educa_jefe)
```

```
## [1] 7.932194
```

```
xeduc10<-8
mean(hogarD10$tot_integ)
```

```
## [1] 3.961048
```

```
xtoti10<-4
xalim10<-mean(hogarD10$alimentos)
xviv10<-mean(hogarD10$vivienda)
xsalu10<-mean(hogarD10$salud)
xcomu10<-mean(hogarD10$comunica)
xedu10<-mean(hogarD10$educacion)
xtran10<-mean(hogarD10$trans)
xing10<-mean(hogarD10$ing_cor)
Phactor10 <- pnorm((bInt+best*xestsoc10+beduj*xeduc10+btot*xtoti10+
                    bali*xalim10+bviv*xviv10+bsal*xsalu10+
                    bcom*xcomu10+beduc*xedu10+btran*xtran10+
                    bing*xing10)/bSigma)
EMint10<- bInt*Phactor10
EMest10 <- best*Phactor10
EMedj10 <- beduj*Phactor10
EMtoti10<-btot*Phactor10
EMali10<-bali*Phactor10
EMviv10<-bviv*Phactor10
EMsal10<-bsal*Phactor10
EMcoc10<-bcom*Phactor10
EMeduc10<-beduc*Phactor10
EMtrans10<-btran*Phactor10
EMing10<-bing*Phactor10
# Crea un data frame con los valores calculados
data10 <- data.frame(Concepto = c("Int", "Est_socio", "Educa_jefe",
                                  "Tot_integ", "Alimentos", "Vivienda",
                                  "Salud", "Comunicaciones", "Educación",
                                  "Transporte", "Ing_cor"),
  Decil10 = c(EMint10, EMest10, EMedj10, EMtoti10, EMali10,
              EMviv10, EMsal10,EMcoc10, EMeduc10, EMtrans10,EMing10))
```

```
# Niveles de Concepto
conceptoNls <- c("Int", "Est_socio", "Educa_jefe",
                 "Tot_integ", "Alimentos", "Vivienda",
                 "Salud", "Comunicaciones", "Educación",
                 "Transporte", "Ing_cor")
datos <- Reduce(function(x, y) merge(x, y, by = "Concepto", all = TRUE),
                list(data1, data2, data3, data4,data5,data6))

# Ordena los datos por la columna Concepto según los niveles especificados
datos$Concepto <- factor(datos$Concepto, levels = conceptoNls)
datos1 <- datos[order(datos$Concepto), ]
tablaEMD <- kable(datos1, caption = "Efectos marginales de hogares
representativos en las medias de deciles", format = "latex",
booktabs = TRUE, align = "c", row.names = FALSE) %>%
  kable_styling(latex_options = c("striped", "hold_position"))
tablaEMD
```

Tabla 47: Efectos marginales de hogares representativos en las medias de deciles

Concepto	Decil1	Decil2	Decil3	Decil4	Decil5	Decil6
Int	-940.1293536	-1115.3995859	-1170.2251371	-1189.6846703	-1195.8565043	-1197.5719077
Est_socio	-171.8745103	-203.9174257	-213.9406366	-217.4982298	-218.6265649	-218.9401750
Educa_jefe	248.6362567	294.9900210	309.4897489	316.7931957	316.2684833	316.7221565
Tot_integ	-371.4246756	-440.6701352	-462.3305190	-470.0185576	-472.4569151	-473.1346337
Alimentos	1.0602057	1.2578620	1.3196901	1.3416350	1.3485952	1.3505297
Vivienda	0.9223133	1.0942620	1.1480486	1.1671394	1.1731942	1.1748771
Salud	0.9098033	1.0794197	1.1324768	1.1513086	1.1572813	1.1589414
Comunicaciones	1.2489915	1.4818436	1.5546811	1.5805337	1.5887331	1.5910121
Educación	0.8581731	1.0181641	1.0682102	1.0859733	1.0916071	1.0931730
Transporte	0.8869109	1.0522595	1.1039815	1.1223394	1.1281619	1.1297802
Ing_cor	0.0331052	0.0392771	0.0412077	0.0418929	0.0421102	0.0421706

A.4.3 Comparación de modelos

Se realizara la comparación del modelo de regresión lineal múltiple y el modelo tobit para saber si su comportamiento es similar.

```
stargazer(mrmtv,h2.tobit, type='text',
          title = 'Modelos',df=F,digits = 4)
```

```
##
## Modelos
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                gasto_mon
##                          OLS             Tobit
##                          (1)              (2)
```

```
## ------------------------------------------------
## est_socio            -238.8469***    -218.9981***
##                       (47.0508)       (44.9515)
##
## educa_jefe           301.7477***      316.8059***
##                       (14.9279)       (14.0143)
##
## tot_integ            -489.9312***    -473.2598***
##                       (19.9810)       (18.8182)
##
## alimentos             1.3519***       1.3509***
##                       (0.0035)        (0.0012)
##
## vivienda              1.1714***       1.1752***
##                       (0.0066)        (0.0063)
##
## salud                 1.1562***       1.1592***
##                       (0.0060)        (0.0057)
##
## comunica              1.5861***       1.5914***
##                       (0.0164)        (0.0158)
##
```

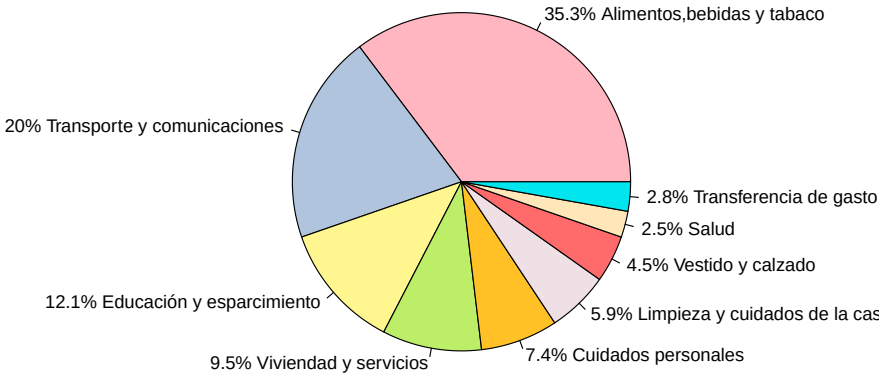
```
## educacion             1.0907***        1.0935***
##                        (0.0048)        (0.0045)
##
## trans                 1.1276***        1.1301***
##                        (0.0031)        (0.0029)
##
## ing_cor               0.0445***        0.0422***
##                        (0.0005)        (0.0001)
##
## Constant           -1,103.9110***  -1,197.8890***
##                      (129.7726)      (124.1053)
##
## ------------------------------------------------
## Observations           90,102          90,102
## R2                     0.9169
## Adjusted R2            0.9169
## Log Likelihood                      -953,812.0000
## Residual Std. Error   9,987.4880
## F Statistic         99,434.5700***
## Wald Test                          8,816,810.0000***
## ================================================
## Note:                   *p<0.1; **p<0.05; ***p<0.01
```

ANEXO B

```
library(grid)
valores <- c(35.3,20.0,12.1,9.5,7.4,5.9,4.5,2.5,2.8)
nombres <- c("35.3% Alimentos,bebidas y tabaco",
             "20% Transporte y comunicaciones",
             "12.1% Educación y esparcimiento",
             "9.5% Viviendad y servicios",
             "7.4% Cuidados personales",
             "5.9% Limpieza y cuidados de la casa",
             "4.5% Vestido y calzado","2.5% Salud",
             "2.8% Transferencia de gasto")
mi_paleta <- colorRampPalette(c("#FFB6C1", "#B0C4DE", "#FFF68F",
                                "#BCEE68","#FFC125",
                                "#EEE0E5", "#FF6A6A",
                                "#FFE7BA", "#00E5EE" ))
pie(valores, labels=nombres,col=mi_paleta(9),
    main = "Gráfico 1: ENIGH 2018")

grid.text("Fuente: Elaboración propia con datos extraídos de la ENIGH 2018. ",
          x = 0.1, y = 0.1, just = "left",
          gp = gpar(fontsize = 12, col = "black"))
```

Gráfico 1: ENIGH 2018

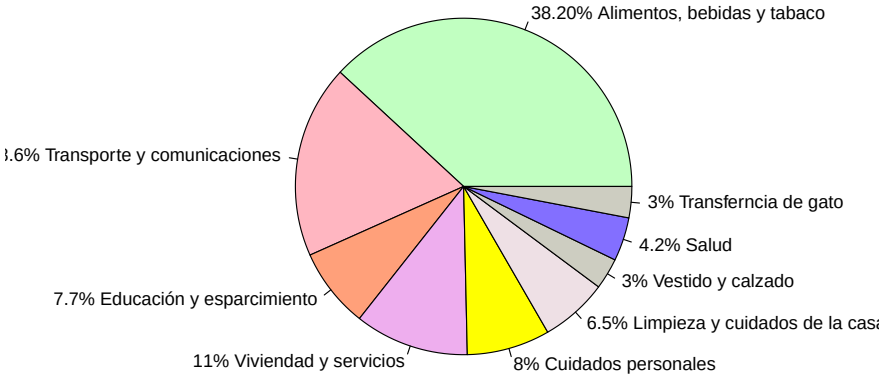


Fuente: Elaboración propia con datos extraídos de la ENIGH 2018.

```r
valores2 <- c(38.20,18.6,7.7,11.0,8.0,6.5,3.0,4.2,3.0)
nombres2 <- c("38.20% Alimentos, bebidas y tabaco",
              "18.6% Transporte y comunicaciones",
             "7.7% Educación y esparcimiento",
             "11% Viviendad y servicios",
             "8% Cuidados personales",
             "6.5% Limpieza y cuidados de la casa",
             "3% Vestido y calzado",
             "4.2% Salud", "3% Transferncia de gato")
mi_paleta2 <- colorRampPalette(c("#C1FFC1", "#FFB6C1", "#FFA07A",
                                 "#EEAEEE","#FFFF00",
                                "#EEE0E5", "#CDCDC1",
                                "#836FFF", "#CDCDC1" ))
pie(valores2, labels=nombres2, main="Gráfico 2: ENIGH 2020",
    col=mi_paleta2(9))

grid.text("Fuente: Elaboración propia con datos extraídos de la ENIGH (2020). ",
          x = 0.1, y = 0.1, just = "left",
          gp = gpar(fontsize = 12, col = "black"))
```

Gráfico 2: ENIGH 2020



Fuente: Elaboración propia con datos extraídos de la ENIGH (2020).

Tabla 48: Cuadro de variables

Categorías	Subcategorias
Alimentos, bebidas y tabaco	Alimentos, Alimentos dentro del hogar, Cereales, Carnes, Pescados y marisco,Leche y derivados, Huevo, Aceite y grasas, Tubérculos, Verduras, Frutas, Azúcar y mieles, Café, té y chocolate, Especias y aderezos, Otros alimentos diversos, Bebidas, Alimentos fuera del hogar y Tabaco.
Vestido y calzado	Vestido y calzado, Vestido, Calzado y su reparación.
Vivienda y servicios de conservación	Vivienda, Alquileres brutos, Predial y cuotas, Agua.
Energía eléctrica y combustibles	Electricidad y combustibles
Cuidados de la casa	Limpieza, Cuidados de la casa, Utensilios y Enseres domésticos.
Salud	Cuidado de la salud, Atención primaria o ambulatoria, Atención hospitalaria y Medicamentos sin receta.
Transporte	Transporte, Transporte público, Transporte foráneo, Adquisición de vehículos, Mantenimiento de vehículos, Refacciones para vehículos, Combustible para vehículos y Comunicaciones.
Educación y esparcimiento	Educación y esparcimiento, Educación, Esparcimiento y Paquetes turísticos.
Cuidados personales	Personales, Cuidados personales, Accesorios personales

Tabla 49: Cuadro de variables

Variable	Definición	Categorías
Ubica_geo	Ubicación geográfica	Se muestra un folio para representar a cada uno de los municipios de los 32 estados de la República Mexicana
Est_socio	Estrato socioeconómico	Se muestra un valor para representar el estrato socioeconómico al que pertenece cada hogar: 01 bajo, 02 medio bajo, 03 medio alto y 04 alto
Clase_hog	Clase de hogar	Se muestra un valor para representar el tipo de familia al que pertenece cada hogar: 01 unipersonal, 02 nuclear, 03 ampliado, 04 compuesto y 05 corresidente
Sexo_jefe	Sexo del jefe del hogar	Se muestra un valor para representar el sexo del jefe del hogar: 01 hombre y 02 mujer
Educa_jefe	Educación del jefe del hogar	Se muestra un valor para representar el nivel de estudios del jefe del hogar: 01 sin intrucción, 02 Preescolar, 03 primaria incompleta, 04 primaria completa, 05 secundaria incompleta, 06 secundaria completa, 07 preparatoria incompleta, 08 preparatoria completa, 09 profesional incompleta, 10 profesional completa y 11 posgrado
Edad_jefe	Edad del jefe del hogar	
Tot_integ	Número de integrantes del hogar	
Gasto_mon	Gasto corriente monetario	
Alimentos	Son los gastos en bienes de consumo no duradero que realizan día a día los integrantes del hogar en alimentos, bebidas y tabaco	
Vivienda	Es el gasto en vivienda, servicios de conservación, energía eléctrica y combustibles.	
Salud	Son los gastos en cuidados de la salud	
Comunica	Gasto en comunicaciones	
Educacion	Es el gasto en artículos y servicios de educación	
Transporte	Es el gasto en transporte; adquisición, mantenimiento, accesorios y servicios para vehículos; comunicaciones.	
Esparci	Gasto en artículos y servicios de esparcimiento	