Hogares mexicanos con automóviles: Modelo Lineal de Probabilidad, Probit y Logit

Roberto López Baldomero

Resumen: Con información de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2020 se desarrolló un modelo econométrico sobre la factibilidad de que un hogar mexicano cuente con al menos un automóvil. La probabilidad del modelo se basa en características del hogar como la educación del jefe y factores económicos como ingresos y gastos trimestrales. Para la estimación de la situación se aplicaron tres métodos: Modelo Lineal de Probabilidad, Probit y Logit. Tras un rebalanceo con 53,760 registros finales, el modelo explicó hasta el 61% de los casos, siendo la educación del jefe la variable con mayor impacto.

Palabras clave: automóviles, hogares, regresión, MLP, probit, logit, ENIGH.

I. INTRODUCCIÓN

Durante los últimos años, el automóvil ha tomado gran importancia sobre la vida cotidiana. Algunos pueden considerarlo como "una adquisición utilitaria […] indispensable para el traslado y tus labores". (Tovar, 2016) Para otros más, puede ser útil para hacer eficiente "sus tiempos de traslado, viajar más seguros y tener un acompañante que los llevará a donde quieran sin depender de horarios o rutas de transporte público". (El Universal, 2021) Pero por más beneficios que un automóvil pueda ofrecer a una familia "al fin de cuentas, tener un auto nos costará cerca de $32,000 pesos al año esto sin contar enganche y mensualidades." (El Universal, 2021) Ciertamente el poseer un vehículo no es para todos. Considérese que el salario mínimo en México, vigente a partir del 1° de enero de 2021 asciende a $213.39 y $141.70 pesos mexicanos diarios para la Zona Libre de la Frontera Norte y el resto del país, respectivamente. (CONASAMI[1], 2021). Suponiendo que un hogar familiar recibe solo un ingreso: el salario del jefe y que este trabaja 365 días al año, el contar con un vehículo tomaría 41.08% y el 61.87% del ingreso percibido para un hogar de la frontera y de uno del resto del país, nuevamente sin contar costos como mensualidades y bajo un supuesto débil en el que no todos quieren o pueden trabajar todos los días durante todo un año.

Asimismo, esta disparidad entre los mexicanos sobre la posesión de un automóvil se ve reflejada en la última Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) edición 2020. Dentro de su muestra representativa con información completa para 88,926 hogares[2] poco más del 30% de los hogares poseen al menos un vehículo.

Esta información se convirtió en un móvil para el desarrollo de este escrito. Mismo que busca definir un modelo econométrico el cual incorpora algunas de las posibles circunstancias, factores o características de los hogares mexicanos con el fin de pronosticar la probabilidad de que uno cuente con al menos un automóvil. Para la predicción de la factibilidad se emplearán tres modelados diferentes: Modelo Lineal de Probabilidad (MLP); Probit y Logit.

II. PLANTEAMIENTO

Reconocer la factibilidad de que un hogar mexicano cuente con al menos un vehículo puede ser de suma importancia en la toma de decisiones tanto para el sector privado como para el estado. Para las firmas, este conocimiento puede ofrecerle las herramientas para mejorar la productividad de sus trabajadores, en particular para aquellos que se transportan largas distancias de sus hogares a las instalaciones de las empresas. Al identificar las características de los trabajadores y reconocer si cuentan o no con un vehículo, pueden tomarse decisiones como la adquisición de transporte para los empleados, o en el caso contrario, ofrecer lugares designados para trabajadores con automóviles. Estas decisiones permitirían mejorar la productividad, de manera que, se reduciría el tiempo y energías perdidas en la toma de diferentes medios de transporte para llegar a las instalaciones, o en la búsqueda de un lugar donde aparcar.

Para el estado, el contar con información como la proporción de hogares mexicanos con automóviles toma relevancia al desarrollar e implementar políticas públicas. De esta manera, puede hacerse mejor uso de los recursos públicos y destinarse a programas que beneficien a la mayoría de la población, ya sea mediante la inversión en transporte público, mediante subsidios sobre combustibles u otros planes afines. Asimismo, puede utilizarse no solo en pro de los consumidores,

[1] Comisión Nacional de los Salarios Mínimos

[2] Registros filtrados para contar solo con aquellos que tuvieran registros sobre la posesión de al menos un automóvil. Metodología más adelante.

sino que también el sector público puede verse beneficiado al identificar áreas de oportunidad donde se puedan implementar nuevos mecanismos de fiscalización como la tenencia vehicular.[3]

III. MODELO

Para el desarrollo de este modelo se usó información de la ENIGH 2020. Esta encuesta tiene como objetivo "proporcionar un panorama estadístico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribución". De igual manera, proporciona información sobre "características ocupacionales, sociodemográficas, acceso a la alimentación […] así como las características de la infraestructura de la vivienda y equipamiento del hogar". (INEGI, 2021)

Para este trabajo se utilizaron cuatro bases de datos incorporadas dentro de los datos libres de la encuesta: ***gastos de los hogares; hogares; ingresos y población.***[4] Un punto por destacar es que, al ser fuentes relativamente separadas, es decir, que las variables de interés para el modelo no se encontraban en una misma base de datos, se optó por unir las bases en una misma según las variables de relevancia[5] de cada una. Para efectuar esta conexión entre registros se implementó la idea de un identificador de hogar único. Este distintivo se basa en los valores de **folioviv**[6] y **foliohog**[7]**,** dos variables presentes en cada una de las fuentes y que de forma conjunta permiten identificar un mismo hogar dentro de los miles de registros.

Del mismo modo, se establece como periodo temporal para las variables una base trimestral. Es decir, para aquellas variables que se encontraban en meses son multiplicados por un factor de tres. Asimismo, se destaca que la base de datos combinada por las variables de interés y según el identificador de hogar único cuenta con 88,926 registros.[8] Sin embargo, al evaluar la capacidad predictiva de los modelos, se consideró necesario un rebalanceo, siendo que los datos para aquellos hogares (62,046) sin vehículo eran poco más del doble que aquellos registros con vehículo (26,880). Para evitar sesgos sobre el modelo, se tomó una muestra aleatoria de los registros sin hogares y se igualaron a la cantidad de aquellos con vehículos (26,800), ofreciendo así una base de datos re balanceada con 53,760 filas, misma que será utilizada para el desarrollo de este trabajo.

Las variables utilizadas para este trabajo son las siguientes:

---

[3] Gravamen que surge en 1962 en México […] este tributo se cobraría a todos los que tuvieran un automóvil a su nombre. (BBVA, S.F)

[4] Para este trabajo, se optó por acortar el nombre de las bases de datos por la longitud del nombre original, sin embargo, mantiene el nombre principal de las originales dentro de los datos libres de la ENIGH 2020.

[5] Se entiende como variable de relevancia toda aquella variable o columna que fuera fundamental para el desarrollo de este escrito, por ejemplo: número de vehículos, escolaridad del jefe, etc.

[6] Identificador de la vivienda compuesto por dos dígitos con la clave de la entidad federativa, uno con el ámbito (urbano, código diferente a 6; rural, código 6), cuatro dígitos del número consecutivo de la upm, un dígito de la decena de levantamiento y dos dígitos con un número consecutivo para la vivienda seleccionada.

[7] Identificador del hogar, el código 1 identifica al hogar principal y del 2 al 5 los hogares adicionales.

[8] Se destaca que la base de ***hogares*** cuenta con 89,007 registros totales, sin embargo, al hacer la filtración de estos con base en las variables de interés se "perdieron" 80 observaciones las cuales no contaban con datos necesarios para el desarrollo de este escrito.

Tabla 1. Variables del modelo econométrico

Nombre	Tipo de Variable por Función	Tipo de Variable por Valores	Origen (BDD[9])	Descripción
tiene_auto	Endógena	Dummy	Modelación a partir de la columna *"num_auto"*[10] de ***hogares.***	Determina si el hogar cuenta con automóvil. • 1 = El hogar cuenta **con al menos** un automóvil • 0 = El hogar **NO** cuenta con automóvil
termino_sec	Exógena	Dummy	Variable modelada a partir de la columna *"nivelaprob"*[11] cuyo registro para "*parentesco"* fuera el código 101 (jefe de hogar) dentro de ***población.***	Determina si el **jefe del hogar**[12] terminó la secundaria: • 1 = El jefe del hogar terminó la secundaria • 0 = El jefe del hogar **NO** terminó la secundaria
gastot_tri	Exógena	Cuantitativa	Variable fabricada a partir de la suma de todos los tipos de gasto trimestral de un hogar de *"gasto_tri"* en ***gasto de los hogares.***	Es el acumulado de gasto trimestral del hogar.
est_trans_tri	Exógena	Cuantitativa	Variable fabricada a partir de la transformación de *"est_trans"* de monto mensual a trimestral en ***hogares.***	Es la estimación del gasto trimestral en transporte público del hogar.
ingtot_tri	Exógena	Cuantitativa	Variable fabricada a partir de la suma de todos los tipos de ingreso trimestral de *"ing_tri"* en ***ingresos.***	Es el ingreso total trimestral del hogar.

[9] Dentro de este escrito se considera a BDD como Base de Datos.
[10] Número de automóviles en el hogar. Se presenta como variable cuantitativa dentro de ***hogares.***
[11] Nivel escolar aprobado. Dentro de ***población,*** *"nivelaprob"* se encuentra como variable cualitativa modelada de manera ascendente, es decir, mayor el número, mayor el grado aprobado. Secundaria corresponde al número 3 dentro de esta modelación. Por lo que sí *"nivelaprob"* mayor o igual a 3, el jefe del hogar terminó la secundaria con valor 1 para *"termino_sec"***.**
[12] Jefe del hogar incorpora tanto a hombres como mujeres.

tasa_alim_tri	Exógena	Cuantitativa	Variable fabricada a partir de *"est_alim"* en ***hogares.*** Dicha variable se pasó de monto mensual a trimestral y por último se dividió entre *"ingtot_tri".*	Es la proporción de la estimación de gasto trimestral en alimentos con respecto al ingreso total por hogar.

Conforme a la pertinencia del modelo y sus variables, se considera que este modelo, además de utilizar registros de una fuente confiable y con recaudación de datos continúa como lo es la ENIGH, también se basa en factores sociales y económicos que pueden ofrecer una buena explicación teórica al momento de pronosticar una situación como la presente.

> La educación es necesaria en todos los sentidos. Para alcanzar mejores niveles de bienestar social y de crecimiento económico; para nivelar las desigualdades económicas y sociales; para propiciar la movilidad social de las personas; para acceder a mejores niveles de empleo. (UNAM, S.F)

Con base en lo anterior se considera a *"termino_sec"* dentro del modelo puesto que, como menciona la UNAM, la educación es un factor influyente sobre cuestiones económicas como la movilidad social de las personas, el empleo y sus beneficios directos como: salarios, prestamos, beneficios de salud, etc. Estos aspectos positivos pueden convertirse determinante al momento de poseer y mantener un vehículo, de manera que un mayor salario e ingresos (representados en *"ingtot_tri")* podrían garantizar las necesidades básicas de las familias como alimentación, vivienda, vestido, y una vez cubiertas, facilitan el tener un vehículo sin necesidad de sacrificar el bienestar de la familia. Por lo tanto, tanto para *"termino_sec"* como *"ingtot_tri"* se espera una relación directa sobre que un hogar cuente con al menos un vehículo.

De forma similar, se contempla una relación positiva entre *"gastot_tri"* y la variable endógena de manera que, un monto superior de gasto total es un indicio de contar con gran cantidad de recursos disponibles para gastar. Dentro de estos gastos además de productos y servicios básicos, también se comprenden gasolina, reparaciones, mantenimiento, limpieza del automóvil, etc.

Por el lado contrario, es de esperar una relación inversa con respecto a *"est_trans_tri"*. Esto debido a que un gasto estimado trimestral en transporte público puede indicar la falta de un automóvil en el hogar, aunque no es una regla general si consideramos que hay momentos en los que vehículos no están disponibles por accidentes, mantenimiento, reparaciones, decretos estatales como "hoy no circula",[13] etc. Estas últimas situaciones pueden guiar a la familia del hogar a usar y gastar en medios de transporte alternativos como el presente.

Finalmente, se contempla como resultado una relación inversa entre *"tasa_alim"tri"* y la variable dependiente. Una característica de las familias con mayor ingreso y riqueza es que la proporción de su ingreso gastado en la alimentación es muy inferior a comparación de aquellas con muchos menos recursos económicos. Según Portella (2018): "El grupo de mexicanos más pobres destinó el 50% de

[13] Es un decreto en el entonces distrito Federal, ahora Ciudad de México, que tiene como objetivo "Establecer medidas aplicables a la circulación vehicular de fuentes móviles o vehículos automotores, con el objetivo de prevenir, minimizar y controlar la emisión de contaminantes […] en la Ciudad de México…" (Secretaría del Medio Ambiente SEDEMA, S. F)

sus ingresos a alimentarse a comparación con el 25% del presupuesto familiar que destina el grupo más rico". Con base en lo anterior, es posible intuir que una mayor proporción del ingreso gastado en alimentos impide la adquisición de otros bienes y servicios como lo sería un automóvil y sus complementos.

En resumen, las relaciones esperadas de las variables exógenas con la característica de que un hogar cuente con al menos un vehículo se ven representadas en la siguiente tabla. (Tabla 2)

Tabla 2. Resumen de las relaciones exógenas con la endógena

Variable	Relación esperada con la variable endógena
termino_sec	Directa (Positiva)
gastot_tri	Directa (Positiva)
ingtot_tri	Directa (Positiva)
est_trans_tri	Inversa (Negativa)
tasa_alim_tri	Inversa (Negativa)

IV. ESTIMACIÓN Y DESARROLLO

Como se ha mencionado con anterioridad, el propósito de este trabajo es pronosticar la probabilidad de que un hogar mexicano cuente con al menos un vehículo dadas las variables exógenas establecidas como el ingreso, el gasto y la educación del jefe.

Un punto importante por destacar es que, para cada uno de los ejercicios se establece un nivel de significancia $\alpha = 0.05$ para definir la situación en la que, tanto una variable como el modelo en conjunto, se considere como estadísticamente significativo. A su vez, se contempla como conjunto de hipótesis para la determinación de significancia de las variables el siguiente:

$$B_i \in B \text{ dónde } B \text{ es el conjunto de coeficientes de las variables}$$
$$H_0: B_i = 0$$
$$H_a: B_i \neq 0$$

Es decir, se busca que el parámetro de la variable exógena posea un impacto estadístico diferente de 0 sobre la endógena.

Del mismo modo, es relevante destacar que la operación para estimar las pendientes no lineales para el modelo Probit y Logit es la siguiente:

$$F(Z) * Bi$$
$$\text{dónde } F(\cdot) \text{ es la distribución del modelo; } Z \text{ el valor estimado}$$
$$\text{y } B_i \text{ el coeficiente de la variable } i$$

MODELO LINEAL DE PROBABILIDAD (MLP)

Según Almilla-López y Camargo (2009) el modelo lineal de probabilidad es aquel en donde "la variable independiente es dicotómica[14] y es función de las variables explicativas". A su vez, "se

[14] Una variable dicotómica es aquella que solo puede tomar dos valores. Estos valores, habitualmente son cero, como ausencia, o uno, como presencia. (Arias, S. F)

puede interpretar en términos probabilísticos, en el sentido que un valor concreto de la recta de regresión mide la probabilidad de que ocurra el hecho objetivo de estudio". (*Ibid*, 2009, p. 2)

La ecuación teórica para el modelo de probabilidad lineal es:

$$\widehat{tiene\_auto} = \widehat{\beta_0} + \widehat{\beta_1} termino\_sec + \widehat{\beta_2} gastot\_tri + \widehat{\beta_3} est\_trans\_tri + \widehat{\beta_4} ingtot\_tri + \widehat{\beta_5} tasa\_alim\_tri + \varepsilon \quad (1)$$

Tabla 3. Resultados del Modelo Lineal de Probabilidad

Dependent Variable: TIENE_AUTO
Method: Least Squares
Sample (adjusted): 1 53760
Included observations: 53760 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.292052	0.004684	62.34439	0
TERMINO_SEC	0.215201	0.004229	50.8821	0
GASTOT_TRI	4.55E-06	8.47E-08	53.77072	0
INGTOT_TRI	1.82E-07	2.62E-08	6.953278	0
EST_TRANS_TRI	-4.96E-05	1.36E-06	-36.4842	0
TASA_ALIM_TRI	-0.115174	0.005568	-20.68419	0

R-squared	0.176775	Mean dependent var	0.5
Adjusted R-squared	0.176698	S.D. dependent var	0.500005
S.E. of regression	0.453684	Akaike info criterion	1.257281
Sum squared resid	11064.15	Schwarz criterion	1.258273
Log likelihood	-33789.7	Hannan-Quinn criter.	1.25759
F-statistic	2308.566	Durbin-Watson stat	0.317394
Prob(F-statistic)	0		

Dadas las variables establecidas y los resultados del modelo lineal de probabilidad, la ecuación estimada es la siguiente:

$$E(\varepsilon) = 0 \quad (2)$$
$$\widehat{tiene\_auto} = 0.292 + 0.215 * termino\_sec + (4.55E-06) * gastot\_tri + (1.82E-07) * ingtot\_tri + (-4.96E-05) * est\_trans\_tri + (-0.115) * tasa\_alim\_tri$$

Conforme a la significancia estadística de las variables, individualmente y en su conjunto, es posible apreciar dentro de la tabla 2 que todas las variables mantienen un p-value[15] de 0, por lo tanto, son estadísticamente significativos al este último ser un valor inferior a la prueba de significancia de $\alpha = 0.05$. Asimismo, se destaca que la varianza del error[16] obtuvo un valor de 0.206.

[15] Se encuentra representado dentro de la columna Prob.

[16] Obtenida a partir de $\frac{SSR}{n-k} = \frac{11064.15}{53760-6}$

Con base en los coeficientes de las variables, es posible observar que, en cuanto a la educación del jefe del hogar, el terminar la secundaria es la variable que toma una gran importancia sobre la probabilidad de que un hogar tenga un auto. A su vez, puesto que se está trabajando con modelo de regresión lineal, las pendientes son constantes y toman los valores de los coeficientes representados para cada una de las variables exógenas dentro de la tabla 3. Del mismo modo, tanto el gasto como el ingreso total trimestral también parecen ser determinantes y presentan una relación directa con la probabilidad de contar con un auto. Asimismo, el gasto estimado en transporte público y la proporción del gasto en alimentos respecto al ingreso total mantienen la relación inversa esperada.

Por último, de acuerdo con el R^2, el modelo estimado tiene la capacidad de explicar la variación en el 17.68% de los hogares. Un valor realmente bajo si es que se quiere aplicar el modelo econométrico para una toma de decisiones. Asimismo, una característica del modelo de probabilidad lineal es que puede colapsar la probabilidad al introducir datos que pronostiquen un valor fuera de los limites posibles [0, 1].

Dadas estas circunstancias, se concluye que este primer modelo no es óptimo ante el objetivo de este trabajo.

PROBIT

"Dadas las dificultades asociadas con el modelo lineal de probabilidad, es natural transformar el modelo original de tal forma que las predicciones caigan en el intervalo [0,1]." Uno de esos modelos que transforman al modelo original es el Probit. Este se caracteriza por usar una función de distribución normal acumulada. (Almilla-López & Camargo, 2009, p. 4)

La ecuación teórica para el modelo probit es:

$$\widehat{tiene\_auto} = \widehat{\beta_0} + \widehat{\beta_1} termino\_sec + \widehat{\beta_2} gastot\_tri + \widehat{\beta_3} est\_trans\_tri + \widehat{\beta_4} ingtot\_tri + \widehat{\beta_5} tasa\_alim\_tri + \varepsilon = Z \quad (3)$$

$$donde\ F(Z) = P_i;\ tal\ que\ F(\cdot)\ es\ una\ distribución\ normal\ acumulada$$

Dentro de este apartado, se expondrá la modelación probit de la base de datos original y la rebalanceada para demostrar su necesidad en cuestión de poder de predicción. A continuación, se exponen los resultados de los modelos probit estimados y las expectativas de estimación con base en un valor de éxito de 0.5.[17]

Tabla 4. Comparación de resultados BDD original-rebalanceada

Base de datos	Variable	Valor	Prob
Original	Observaciones	88926	
	C	-1.102377	0
	TERMINO_SEC	0.532868	0
	GASTOT_TRI	1.60E-05	0
	INGTOT_TRI	1.02E-06	0
	EST_TRANS_TRI	-0.000153	0

[17] El valor de éxito es un indicador en el que, sí un valor sobrepasa su valor, se generaliza a que ese registro sea tomado con la característica de la variable dependiente. En este caso el hogar tendría un automóvil y definiría la variable *"tiene_auto"* como 1.

	TASA_ALIM_TRI	-0.381163	0
	McFadden R-squared	0.147844	
	Prob(LR statistic)	0	
	Obs Dep=0	62046	
	Obs Dep=1	26880	
	% Correct Dep=0	75.47	
	% Correct Dep=1	42.26	
Rebalanceada	Observaciones	53760	
	C	-0.722307	0
	TERMINO_SEC	0.546102	0
	GASTOT_TRI	1.81E-05	0
	INGTOT_TRI	1.54E-06	0
	EST_TRANS_TRI	-0.000152	0
	TASA_ALIM_TRI	-0.310749	0
	McFadden R-squared	0.156401	
	Prob(LR statistic)	0	
	Obs Dep=0	26880	
	Obs Dep=1	26880	
	% Correct Dep=0	60.27	
	% Correct Dep=1	59.76	

De acuerdo con la tabla anterior, es posible identificar que el modelo Probit con la base original considera 88,926 observaciones de las cuales 62,046 son registros de hogares sin vehículos y solo 26,880 de aquellos que cuentan con al menos uno. Conforme a los valores estadísticos, tanto las variables como el modelo en su conjunto mostraron ser estadísticamente significativos, con un p-value para las primeras y una probabilidad de LR statistic[18] para el segundo, inferior a un $\alpha = 0.05$. Según el $R^2\ de\ McFadden$, esta primera modelación explica la variación del 14.78% de las observaciones. De forma similar, se demuestra que solo se puede predecir correctamente el 42.26% de los hogares que cuentan con vehículo. Debido a que este último punto es el más importante para este trabajo, se optó por rebalancear la muestra para evitar sesgos sobre la variable endógena y tener una base mucho más equilibrada.

Al efectuar el rebalanceo los resultados mejoraron considerablemente. Este segundo modelo Probit considera la base con 53,760 observaciones, la cuales son repartidas igualitariamente entre hogares con automóvil y sin. Dentro de las estadísticas del modelo nuevamente todas las variables y el modelo en conjunto se perciben como estadísticamente significativos. Así pues, el poder de predicción incrementó de 42.26% a 59.76%. De forma similar, esta nueva modelación con rebalanceo explica una variación en la variable dependiente superior a la original, incrementando de 14.78% a 15.64% (Véase en los valores para $R^2 de\ McFadden$ en la tabla 4). Por lo tanto, se

[18] La probabilidad de LR statistic de los modelos probit y logit se asemeja al valor de la probabilidad F statistic para un modelo de regresión lineal. De manera que, para que el modelo sea significativo en su conjunto, es necesario que el valor de probabilidad para estos sea inferior al nivel de significancia, en este caso $\alpha = 0.05$.

confirma una necesidad de rebalanceo sobre la base de datos y se concluye que, para este ejercicio, la BDD rebalanceada es la óptima para predecir con mayor precisión la probabilidad de que un hogar cuente con al menos un vehículo.

Así pues, la ecuación estimada para el modelo rebalanceado es el siguiente:

$$\widehat{tiene\_auto} = (-0.722) + (0.546) * termino\_sec + (1.81E-05) * gastot\_tri + (1.54E-06) * ingtot\_tri + (-0.0001) * est\_trans\_tri + (-0.311) * tasa\_alim\_tri = Z \quad (4)$$

$$donde\ F(Z) = P_i;\ tal\ que\ F(\cdot)\ es\ una\ distribución\ normal\ acumulada$$

Al igual que el modelo lineal de probabilidad, las variables mantienen las relaciones esperadas. De manera que haber terminado la secundaria (jefe del hogar), el gasto e ingreso total trimestral muestran una relación positiva entre estas y la probabilidad de contar con al menos un automóvil en el hogar. Del mismo modo, se conservan las relaciones inversas para el gasto estimado trimestral en transporte público y la proporción del ingreso total gasto en alimentos. La diferencia es que, además de ser un procedimiento diferente, el MLP solo era capaz de predecir correctamente el 16.67% de las observaciones, mientras que este segundo modelo pronóstica satisfactoriamente casi el 60%.

Cabe señalar que, a diferencia del MLP, el modelo Probit cuenta con pendientes no lineales, es decir, estas cambian con base en los datos de entrada. En la siguiente tabla se representan las razones de cambio de las variables exógenas a partir de registros de 10 hogares.[19]

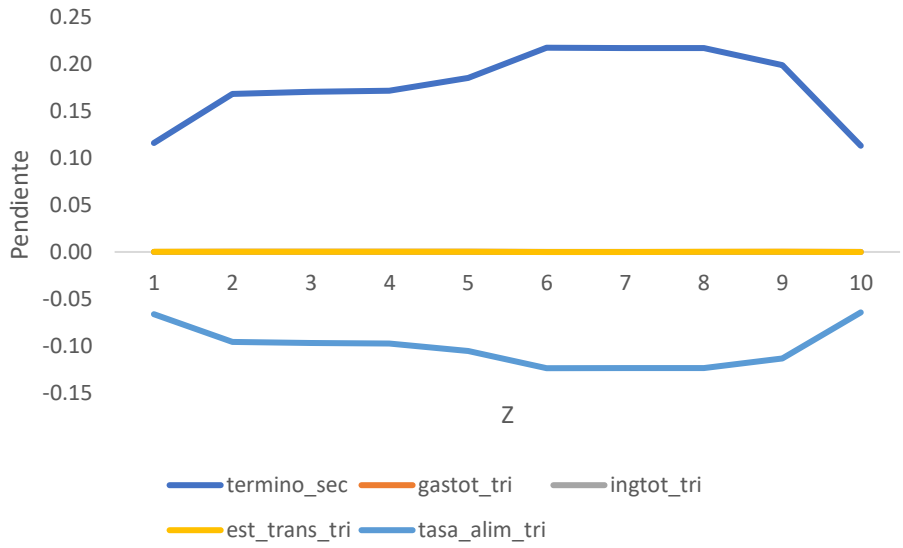
Tabla 5. Pendientes del modelo Probit

folioviv	foliohog	Z	F(Z)=Pi	termino_sec	gastot_tri	ingtot_tri	est_trans_tri	tasa_alim_tri
3061023409	1	-1.12	0.1311	0.12	3.85E-06	3.28E-07	-3.23E-05	-0.07
2660021414	1	-0.72	0.2355	0.17	5.57E-06	4.74E-07	-4.68E-05	-0.10
2660577904	1	-0.70	0.2424	0.17	5.66E-06	4.81E-07	-4.75E-05	-0.10
2062588003	1	-0.69	0.2453	0.17	5.69E-06	4.84E-07	-4.78E-05	-0.10
2504883315	2	-0.57	0.284	0.19	6.13E-06	5.22E-07	-5.15E-05	-0.11
505013303	1	0.05	0.5218	0.22	7.21E-06	6.13E-07	-6.05E-05	-0.12
860048822	1	0.07	0.526	0.22	7.21E-06	6.13E-07	-6.05E-05	-0.12
1560616912	1	0.09	0.5352	0.22	7.19E-06	6.12E-07	-6.04E-05	-0.12
2760272923	1	0.43	0.6647	0.20	6.60E-06	5.61E-07	-5.54E-05	-0.11
3200210904	1	1.14	0.8739	0.11	3.75E-06	3.19E-07	-3.15E-05	-0.06

De acuerdo con la tabla anterior, se aprecia que *"termino_sec"* cuenta con la pendiente más grande en términos absolutos y tras esta se destaca *"tasa_alim_tri"*, seguidas por las razones de cambio de *"est_trans_tri"*, *"gastot_tri"* y *"ingtot_tri"*. Así pues, dichas pendientes se visualizan de la siguiente manera:

[19] Es una muestra con 10 registros de la base original. Será ocupada para estimar las pendientes del Probit y Logit.

Gráfica 1. Pendientes de las Variables: Probit



Por último, este modelo cuenta con una predicción relativamente baja del 60%, sin embargo, muestra tener el poder como para predecir con más certeza que dejarlo en el azar de una moneda. De igual manera, se establece que puede ser mejorado al introducir variables que expliquen aún más las condiciones que favorezca el que un hogar mexicano cuente con un hogar.

LOGIT

Similar al Probit, el modelo Logit lineal también es una herramienta que transforma el modelo original de tal manera que las probabilidades esperadas caigan dentro de un intervalo de [0,1]. Esta modelación se caracteriza por usar una distribución logística. (Almilla-López & Camargo, 2009, p. 5)

La ecuación teórica para el modelo logit es:

$$\widehat{tiene\_auto} = \widehat{\beta_0} + \widehat{\beta_1} termino\_sec + \widehat{\beta_2} gastot\_tri + \widehat{\beta_3} est\_trans\_tri + \widehat{\beta_4} ingtot\_tri + \widehat{\beta_5} tasa\_alim\_tri + \varepsilon = Z \quad (5)$$

$$donde\ F(Z) = P_i;\ tal\ que\ F(\cdot)\ es\ una\ distribución\ logística\ acumulada$$

Bajo logit se mantiene la necesidad de rebalanceo al igual que para el probit. Del mismo modo, para determinar el poder de predicción del modelo se mantiene el valor de éxito de 0.5 Los resultados de este procedimiento son los siguientes:

Tabla 6. Resultados del Modelo Probit con base rebalanceada

Dependent Variable: TIENE_AUTO
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Sample (adjusted): 1 53760
Included observations: 53760 after adjustments
Convergence achieved after 5 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-1.435668	0.030818	-46.58565	0
TERMINO_SEC	8.41E-01	2.09E-02	40.33364	0
GASTOT_TRI	3.20E-05	8.15E-07	39.27125	0
INGTOT_TRI	7.95E-06	4.51E-07	17.64018	0
EST_TRANS_TRI	-0.000293	8.01E-06	-36.61106	0
TASA_ALIM_TRI	-0.373073	0.040302	-9.256998	0

McFadden R-squared	0.170246	Mean dependent var	0.5
S.D. dependent var	0.500005	S.E. of regression	0.436611
Akaike info criterion	1.150507	Sum squared resid	10247.1
Schwarz criterion	1.1515	Log likelihood	-30919.63
Hannan-Quinn criter.	1.150817	Restr. deviance	74527.18
Restr. log likelihood	-37263.59	LR statistic	12687.92
Avg. log likelihood	-0.575142	Prob(LR statistic)	0

Obs with Dep=0	26880	Total obs	53760
Obs with Dep=1	26880		

Con base en la tabla anterior, se reconoce que las variables de forma individual y conjunta son estadísticamente significativas. Individualmente, los factores exógenos presentan una probabilidad inferior al nivel de significancia $\alpha = 0.05$ y conforme al modelo en sí, (conjunto de variables) el valor para la probabilidad de LR statistic es de 0, por lo tanto, se confirma lo establecido. Así pues, según él $R^2 de\ McFadden$ el modelo puede explicar 17.02% de la variación en la variable endógena. Asimismo, la ecuación estimada del modelo es la siguiente:

$$\widehat{tiene\_auto} = (-1.436) + (8.41E-01) * termino\_sec + (3.20E-05) * gastot\_tri + (7.95E-06) * ingtotw\_tri + (-0.0003) * est\_trans\_tri + (-0.373) * tasa\_alim\_tri = Z \quad (6)$$

$$donde\ F(Z) = P_i;\ tal\ que\ F(\cdot)\ es\ una\ distribución\ logística\ acumulada$$

Conforme a la evaluación de expectativas de predicción, también denominada como matriz de confusión, los resultados se representan dentro de la siguiente tabla. (Tabla 7)

Tabla 7. Evaluación de Expectativas de Predicción/Matriz de confusión

Expectation-Prediction Evaluation for Binary Specification

Date: 09/09/21 Time: 19:02

Success cutoff: C = 0.5

	Estimated Equation	Constant Probability

	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	20269	8285	28554	26880	26880	53760
P(Dep=1)>C	6611	18595	25206	0	0	0
Total	26880	26880	53760	26880	26880	53760
Correct	20269	18595	38864	26880	0	26880
% Correct	75.41	69.18	72.29	100	0	50
% Incorrect	24.59	30.82	27.71	0	100	50
Total Gain*	-24.59	69.18	22.29			
Percent Gain**	NA	69.18	44.58			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(# of Dep=0)	16453.98	10426.02	26880	13440	13440	26880
E(# of Dep=1)	10426.02	16453.98	26880	13440	13440	26880
Total	26880	26880	53760	26880	26880	53760
Correct	16453.98	16453.98	32907.97	13440	13440	26880
% Correct	61.21	61.21	61.21	50	50	50
% Incorrect	38.79	38.79	38.79	50	50	50
Total Gain*	11.21	11.21	11.21			
Percent Gain**	22.43	22.43	22.43			

Con la tabla anterior como referencia, se percibe que el modelo, bajo un valor de éxito del 0.5, es capaz de predecir correctamente la posesión de un vehículo en 61.21% de los hogares. Esto representa una mejora en relación con el modelo Probit para la estimación de los casos con al menos un automóvil, donde solo se alcanzaba un 59.76% de poder de predicción.

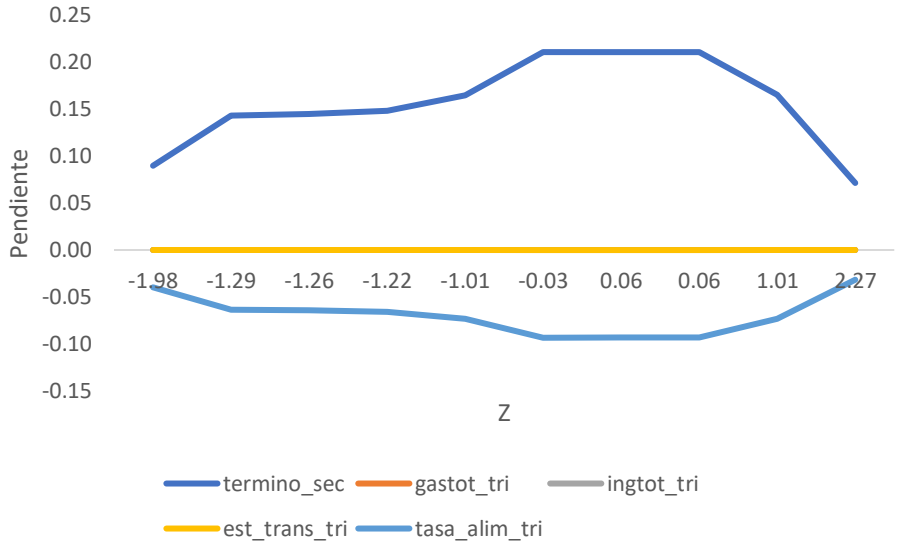
Conforme a las pendientes, al igual que el modelo Probit, el modelo Logit también cuenta con unas de tipo no lineal. Estas se presentan dentro de la siguiente tabla:

Tabla 8. Pendientes del modelo Logit

folioviv	foliohog	Z	F(Z)=Pi	termino_sec	gastot_tri	ingtot_tri	est_trans_tri	tasa_alim_tri
3061023409	1	-1.98	0.12127	0.09	3.41E-06	8.47E-07	-3.12E-05	-0.04
2660021414	1	-1.29	0.21647	0.14	5.43E-06	1.35E-06	-4.97E-05	-0.06
2062588003	1	-1.26	0.22013	0.14	5.49E-06	1.36E-06	-5.03E-05	-0.06
2660577904	1	-1.22	0.22746	0.15	5.62E-06	1.40E-06	-5.15E-05	-0.07
2504883315	2	-1.01	0.26642	0.16	6.25E-06	1.55E-06	-5.73E-05	-0.07
860048822	1	-0.03	0.4927	0.21	8.00E-06	1.99E-06	-7.32E-05	-0.09
1560616912	1	0.06	0.51397	0.21	7.99E-06	1.99E-06	-7.32E-05	-0.09
505013303	1	0.06	0.51501	0.21	7.99E-06	1.99E-06	-7.32E-05	-0.09
2760272923	1	1.01	0.73206	0.17	6.28E-06	1.56E-06	-5.75E-05	-0.07
3200210904	1	2.27	0.90662	0.07	2.71E-06	6.73E-07	-2.48E-05	-0.03

De acuerdo con la tabla 8, se aprecia que el *"termino_sec"* vuelve a contar con una pendiente de alto impacto sobre la variable endógena, seguida por la "*tasa_alim_tri"*. Asimismo, las pendientes se visualizan dentro de la siguiente gráfica:

Gráfica 2. Pendientes de las variables: Logit



Aunque este nuevo modelo cuenta con el poder de pronosticar acertadamente poco mas del 61% de los casos, lo cierto es que este podría considerarse no suficiente para una toma de decisiones crítica.

V. RESULTADOS Y CONCLUSIONES

Durante este escrito se estableció la situación de posesión de al menos un vehículo por parte de los hogares mexicanos. Posteriormente, se construyó un modelo a partir de que el jefe haya terminado la secundaria; el ingreso y gasto total trimestral del hogar; así como algunos estimados sobre gasto en transporte público y la proporción del ingreso gastado en alimento. Este modelo econométrico paso por diferentes transformaciones: Modelo Lineal de Probabilidad (MLP), Probit y Logit. A su vez, se destaca que a partir de una primera evaluación de predicción con Probit sobre la base de datos original (88,926) se consideró necesario un rebalanceo para aproximarse a la máxima capacidad de pronóstico dadas las variables ofrecidas. Estos datos re balanceados con 53,760 registros y repartidos de forma igualitaria entre hogares con vehículos y aquellos sin, fueron la piedra angular de este escrito.

Conforme a los resultados, se observaron mejoras variaciones conforme se aplicaban los diferentes modelos. En el primer momento, la ecuación estimada con MLP fue capaz de explicar el 17.68% de la variación en cuanto a la posesión de al menos un vehículo de los hogares, mientras que el Probit 15.64% y el Logit 17.02%.

Similarmente, la evolución del poder de predicción limitado a un valor de éxito de 0.5 fue positiva. Con la base original y el modelo Probit, la ecuación estimada era capaz de predecir solo el 42.26%. Tras el rebalanceo y usando el mismo modelo anterior, la predicción de expectativa incrementó 17.5 puntos porcentuales con respecto al primero, ofreciendo un poder de pronóstico del 59.76%. Por último, el Logit mostró poder predecir correctamente hasta el 61.21% de los casos.

Asimismo, un punto a destacar es que, en todos los modelos, tanto las variables de forma individual como en conjunto mostraron ser estadísticamente significativas. A su vez, parecen haber cumplido con las relaciones esperadas con respecto a la variable dependiente. De modo que, las pendientes de las variables *"termino_sec", "gastot_tri", "ingtot_tri"* mantuvieron una relación directa con la probabilidad de que un hogar cuente con al menos un vehículo, mientras que, *"est_trans_tri"* y *"tasa_alim_tri"* insistieron en ser inversas. Del mismo modo, se destaca que la variable con mayor impacto sobre la estimación de la factibilidad fue aquella que indicaba si el jefe del hogar había terminado o no la secundaria.

Sin duda, el poder predecir poco mas de 60 de cada 100 casos es considerablemente bueno, pero no los suficiente para ser empleado en una toma de decisiones crítica. Por tanto, se considera que hay un amplio margen de mejora sobre el modelo y que puede no ser el óptimo para emplear en situaciones reales con gran impacto, pero si como una base y punto de partida para modelaciones similares.

VI. REFERENCIAS

Alamilla-López, N., E. & Camargo, S., A. (2009). Limitaciones del modelo lineal de probabilidad y alternativas de modelación microeconométrica. *Temas de Ciencia y Tecnología.* https://www.utm.mx/edi_anteriores/Temas39/1ENSAYO%2039-1.pdf

Arias, E., R. (S.F) *Variable dicotómica.* Economipedia. https://economipedia.com/definiciones/variable-dicotomica.html

BBVA. (S.F) *¿Qué es la tenencia vehicular?* BBVA. https://www.bbva.mx/educacion-financiera/blog/que-es-la-tenencia-vehicular.html

CONASAMI. (2021). *SALARIOS MÍNIMOS 2021.* Gob.mx. https://www.gob.mx/cms/uploads/attachment/file/602096/Tabla_de_salarios_m_nimos_vigente_a_partir_de_2021.pdf

El Universal. (2021). *Cuánto cuesta tener un auto en México.* El Universal. https://www.eluniversal.com.mx/autopistas/cuanto-cuesta-tener-un-auto-en-mexico

INEGI. (2021) *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020 Nueva serie. https://www.inegi.org.mx/programas/enigh/nc/2020/#Documentacion*

INEGI. (2021). *Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH). 2020 Nueva serie* [Conjunto de datos]. ENIGH 2020. https://www.inegi.org.mx/programas/enigh/nc/2020/#Datos_abiertos

Portella, A. (2018) *1 de cada 5 mexicanos no tiene dinero suficiente para alimentarse: Coneval.* Forbes. https://www.forbes.com.mx/1-de-cada-5-mexicanos-no-tiene-dinero-suficiente-para-alimentarse-coneval/

SEDEMA. (S.F) *Hoy No Circula.* Sedema.cdmx.gob.mx. https://sedema.cdmx.gob.mx/programas/programa/hoy-no-circula

Tovar, A. (2016). *Compra de un auto, ¿inversión, lujo o necesidad?* El financiero. https://www.elfinanciero.com.mx/opinion/alberto-tovar/compra-de-auto-inversion-lujo-o-necesidad/

UNAM. (S.F) *Importancia de la educación para el desarrollo.* Plan Educativo Nacional. http://www.planeducativonacional.unam.mx/CAP_00/Text/00_05a.html#:~:text=Adem%C3%A1s%20de%20proveer%20conocimientos%2C%20la,nos%20caracteriza%20como%20seres%20humanos.&text=En%20la%20actualidad%2C%20el%20conocimiento,y%20prioritaria%20en%20lo%20social