Departamento de
Ingeniería Industrial
Universidad de Concepción

UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGRENIERÍA INDUSTRIAL

ESTIMACIÓN DE INGRESOS Y DE LA POSESIÓN DE AUTOMÓVILES EN LOS HOGARES DEL GRAN CONCEPCIÓN

POR

Gerardo Martín Carreño Hurtado

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título de Ingeniero Civil Industrial

Profesor Guía
Sebastián Astroza Tagle, Ph.D.

Julio, 2025

Concepción, Chile.

Dedicatoria

A mi padre, Gerardo, y madre, Andrea, que me dieron la confianza, cariño y apoyo incondicional a lo largo de todo este viaje de conocimiento y desarrollo personal. Independientemente de las decisiones y caminos que quisiera tomar, siempre me dieron la libertad de actuar y expresarme como quisiera, pero siendo responsable y consecuente con ello.

A mis hermanas Francisca y Connie, y a mi hermano Benjamín, cada uno enseñándome a ser mejor y un buen hermano a lo largo de toda mi vida, y siendo un apoyo, cada uno a su manera (y dándome unos sobrinos que tanto me hacen recordar a nosotros mismos cuando éramos pequeños).

A mi pareja Francisca, que estuvo conmigo y me acompañó física y emocionalmente cuando más lo necesité, viajando cientos de kilómetros para estar juntos y haciendo todo lo posible por verme feliz, dándome la oportunidad de ser cada día una mejor pareja y un mejor hombre.

A la vida, por permitirme vivir y aprender de cada persona, experiencia y lugar en el que pude estar. Estoy agradecido por todo lo vivido, bueno y malo, que me ha generado una enseñanza y me ha permitido ser quien soy. Y también por regalarme a Suki, una bendición que no esperaba y que me ha hecho sentirme afortunado y profundamente feliz de tener la familia y la vida que tengo.

Agradecimientos

Agradezco al profesor Sebastián Astroza por la oportunidad y confianza de realizar mi memoria de título con el como guía, por dedicar su tiempo en reunirse con sus memoristas una vez a la semana y ser un gran apoyo en el desarrollo de ella.

Agradezco también a los profesores de la Facultad de Ingeniería y del Departamento de Ingeniería Industrial que se dedicaron no solamente a enseñar, sino a formar profesionales.

Resumen

La caracterización de los ingresos de los hogares y su relación con el acceso a bienes como los automóviles ha sido tradicionalmente abordada mediante análisis agregados y desarticulados de la realidad territorial. Limitando la comprensión de las desigualdades socioeconómicas y espaciales. Esta memoria propone una metodología para estimar ingresos y posesión de automóviles de los hogares en el Gran Concepción, integrando distintas fuentes de datos y técnicas de aprendizaje estadístico para ofrecer una visión detallada y coherente de la estructura socioeconómica de la región.

El objetivo principal de esta memoria es estimar los ingresos de los hogares y la posesión de automóviles en la provincia de concepción, utilizando datos de fuentes oficiales —Casen 2017, Censo 2017 y Encuesta Origen-Destino 2015— e implementando modelos estadísticos que permitan caracterizar la heterogeneidad socioeconómica de los hogares de la zona. Para ello se desarrollaron tres modelos: uno para estimar el ingreso autónomo del hogar, y los otros dos para estimar la posesión de automóviles, ambos aplicados sobre la información censal para obtener estimaciones a nivel hogar.

Siguiendo ese orden, se estimó el ingreso del hogar mediante un modelo de regresión basado en Random Forest, entrenado sobre los datos Casen, y aplicado a los registros del Censo. Este modelo consideró variables sociodemográficas, físicas de la vivienda y características del hogar, mostrando un buen ajuste a la muestra de entrenamiento y un desempeño moderado en los datos de prueba ($R^2 = 0{,}28$), pero consistente con la alta variabilidad del ingreso en la población. La distribución del ingreso estimado respeta la estructura jerárquica comunal observada, reproduciendo correctamente las diferencias entre comunas y los deciles de ingreso. Las variables más relevantes resultaron ser la edad, tamaño de hogar y nivel educacional.

Posteriormente, se estimó la posesión de automóviles mediante dos modelos complementarios, una regresión logística para predecir la probabilidad de poseer al menos un automóvil, y una regresión Poisson, modelando directamente el número de vehículos por hogar. Ambos modelos se alimentaron de la predicción de ingresos previa y otras variables censales.

Los resultados evidencian que los modelos desarrollados son capaces de captar patrones fundamentales en los ingresos y acceso a vehículos en el Gran Concepción, ofreciendo así una herramienta útil para el análisis territorial y la toma de decisiones públicas. Este estudio no solo refuerza la utilidad de las encuestas y censos en conjunto, sino que también propone un marco metodológico replicable para estimaciones en otras regiones y con bases de datos más recientes.

Summary

The characterization of household income and its relationship to access to goods such as automobiles has been approached through aggregated analyses disconnected from the territorial realities, limiting the understanding of socioeconomic and spatial inequalities. This thesis proposes a methodology to estimate household income and car ownership in the province of Concepción, integrating different data sources and statistical learning techniques to provide a detailed and consistent view of the region's socioeconomic structure.

The main objective of this thesis is to estimate household income and car ownership in the province of Concepción, using data from official sources—Casen 2017, Censo 2017, and Encuesta Origen-Destino 2015—and implementing statistical models that allow characterizing the socioeconomic heterogeneity of households in the area. To this end, three models were developed: one to estimate household autonomous income and the other two to estimate car ownership, both applied to census data to obtain estimates at the household level.

First, household income was estimated using a Random Forest regression model trained on Casen survey data and applied to census records. This model considered sociodemographic, housing, and household characteristics, showing a good fit in the training sample and moderate but consistent performance in the test data ($R^2 = 0{,}28$), in line with the high variability of income in the population. The estimated income distribution respects the observed communal hierarchy, correctly reproducing differences between communes and income deciles. The most relevant variables were the age, household size, and education level.

Subsequently, car ownership was estimated using two complementary models: a logistic regression to predict the probability of owning at least one car, and a Poisson regression to directly model the number of vehicles per household. Both models were fed with the previously predicted income and other census variables.

The results show that the developed models can capture key patterns of inequality in income and access to vehicles in Gran Concepción, providing a useful tool for territorial analysis and public decision-making. This study not only reinforces the utility of combining surveys and censuses but also proposes a replicable methodological framework for similar estimates in other regions.

Tabla de contenidos

Lista de tablas

Lista de tablas

1. Introducción

El creciente incremento en la posesión de vehículos a nivel global en economías emergentes como Chile, constituye un acontecimiento multifacético que genera significativas implicancias socioeconómicas, ambientales y territoriales. Durante la última década, Chile ha experimentado una notable expansión de su parque automotriz, posicionándose como el país con mayor capacidad de consumo vehicular per cápita en la región latinoamericana (CAVEM, 2014). Esta tendencia al alza se refleja en las proyecciones que indican un crecimiento sostenido en la propiedad de vehículos de pasajeros, con una tasa de crecimiento anual compuesta del 1,35% proyectada para el periodo 2024-2028 (ReportLinker, 2024).

Si bien el crecimiento del parque vehicular ha generado beneficios tangibles en términos de conectividad y movilidad personal, simultáneamente ha generado externalidades negativas considerables, incluyendo la congestión vial y emisiones contaminantes que afectan particularmente a las áreas urbanas densamente pobladas (OECD, 2022). La dualidad de este fenómeno se evidencia en que, mientras el sector automotriz ha contribuido al desarrollo económico nacional mediante la estabilidad macroeconómica y al acceso a financiamiento, también ha intensificado los desafíos ambientales urbanos, especialmente en los centros metropolitanos como Santiago (OECD, 2022).

En este contexto, la estimación precisa de la posesión automóvil emerge como una herramienta fundamental para comprender y cuantificar los efectos multidimensionales de este fenómeno. Tal análisis resulta esencial para evaluar de manera integral su impacto en los patrones de movilidad urbana, la economía tanto familiar como nacional, y para fundamentar el diseño e implementación de políticas públicas efectivas que logren equilibrar los objetivos de accesibilidad, equidad social y mitigación ambiental en el contexto del desarrollo urbano sostenible (CAVEM, 2014; OECD, 2022; ReportLinker, 2024)

En particular, los modelos de estimación cumplen un rol clave en los sistemas de simulación de viajes y actividades, ya que permiten predecir la demanda de transporte y evaluar el impacto de distintas políticas o escenarios urbanos sobre los patrones de movilidad observados (Ortúzar & Willumsen, 2011; McNally & Rindt, 2007). Desde esta perspectiva, contar con estimaciones consistentes de variables como ingreso o tenencia de vehículos se vuelve un insumo esencial para alimentar modelos de transporte más amplios, comprender la movilidad urbana, anticipar la generación de actividades y

vincular los desplazamientos con las características sociodemográficas y territoriales. Esto proporciona una base cuantitativa sólida para evaluar escenarios futuros y respaldar decisiones en materia de políticas de movilidad.

En la actualidad existe una necesidad de comprender los determinantes detrás de posesión y uso de vehículos particulares, ya que aproximadamente el 50% de los viajes urbanos en todo el mundo se realiza en vehículos privados, y se estima que el número de viajes en automóvil alcanzará los 6.200 millones en 2025 (Soltani, 2017). La motorización ha ido en aumento incluso en países en desarrollo, donde el crecimiento económico, la urbanización acelerada y la mejora del poder adquisitivo han incentivado la adquisición de vehículos (Verma et al., 2016). En este contexto, es válido cuestionarse cómo distintos factores como el nivel de ingreso, ubicación geográfica, infraestructura urbana o características del hogar inciden en la decisión de poseer un automóvil. Comprender estas dinámicas es especialmente relevante en áreas metropolitanas como el Gran Concepción, donde la posesión de vehículos no solo refleja aspectos económicos, sino también impacta la movilidad, planificación urbana y la equidad en acceso a oportunidades.

La heterogeneidad en los niveles de ingreso constituye un determinante critico en las capacidades de movilidad urbana, generando disparidades socioespaciales que se intensifican progresivamente en las periferias metropolitanas. Investigaciones recientes demuestran que la variabilidad económica entre hogares actúa como catalizador de desigualdades sistémicas, limitando el acceso a opciones de transporte y condicionando las oportunidades laborales y educativas (Arellana, 2021). Este acontecimiento se evidencia en el estudio comparativo de Arellana (2021) realizado en Bogotá y Barranquilla, donde se identificó que los sectores de mayores ingresos disponen de una cobertura optima de transporte público, infraestructura vial priorizada y proximidad a centros de empleo. En contraste, las zonas periféricas de menor nivel socioeconómico enfrentan una marcada segregación espacial, con sistemas de transporte fragmentados, tiempos de desplazamiento prolongados y costos que consumen hasta el 30% del presupuesto familiar, perpetuando así ciclos de exclusión urbana.

Finalmente, es de destacar la importancia de la estimación precisa de los ingresos de los hogares, considerando posibles sesgos estadísticos existentes en otros estudios debido a la ausencia de variables socioeconómicas (Agostini et al., 2016). En encuestas que serán utilizadas en esta memoria de título para la formulación de modelos como: Censo Nacional de Población, Hogares y Viviendas del 2017, la encuesta de Caracterización Socioeconómica Nacional del mismo año y la encuesta origen destino del Gran Concepción 2015.

Esta memoria es parte del proyecto 1221724 del Fondo Nacional de Desarrollo Científico y Tecnológico (FONDECYT).

1.2 Objetivos

1.2.1 Objetivo General

Realizar una predicción del ingreso mediante un modelo econométrico para analizar su impacto y estimación en la posesión de vehículos, con el fin de contribuir en la formulación modelos predictivos para el ingreso y posesión de automóviles en los hogares del Gran Concepción

1.2.2 Objetivos específicos

1. Descripción de datos socioeconómicos respecto del ingreso y de movilidad urbana para la construcción de la base de datos de fuentes como: Censo, Encuesta de Caracterización Socioeconómica Nacional y Encuesta Origen-Destino
2. Formular un modelo econométrico para predecir el ingreso de los hogares.
3. Formular un modelo econométrico para predecir la cantidad de vehículos de los hogares.
4. Estimar la cantidad de vehículo por hogar en el Gran Concepción en función de la estimación de los ingresos, considerando también otras variables relevantes.
5. Analizar y concluir respecto de la relación entre los ingresos y otras variables relevantes con la posesión de vehículos.

1.2.3 Descripción de secciones de la memoria de titulo

La estructura de la memoria se compone de 5 capítulos principales. El Capítulo 1 introductorio y de definición de objetivos de la memoria, el capítulo 2 de marco teórico que desarrolla los fundamentos conceptuales y empíricos que sustentan la investigación. Revisando la literatura sobre movilidad urbana, posesión vehicular y sus determinantes económicos, así como las metodologías para estimar ingresos, presentando al final de este capítulo la relación ingreso-motorización que justifica e introduce los enfoques estadísticos utilizados. El capítulo 3 presenta el caso de estudio, describiendo el contexto territorial, las fuentes a utilizar, preparación de datos y especificación de modelos desarrollados. El capítulo 4 presenta los resultados obtenidos de los modelos. Y en el capítulo 5 se resumen los principales hallazgos de la investigación, evaluando la validez y robustez de la metodología propuesta y las contribuciones del estudio.

2. Marco Teórico

En este capítulo se desarrollan los fundamentos conceptuales y empíricos que sustentan el estudio. Se aborda la movilidad urbana y su vínculo con la tenencia de vehículos, revisando la literatura sobre sus determinantes económicos, sociodemográficos y territoriales. Además, se analizan las metodologías existentes para la estimación de ingresos como variable latente, incluyendo enfoques econométricos y de aprendizaje automático. Se discute la relación ingreso-motorización desde una perspectiva teórica y empírica, y se presenta el enfoque metodológico que será implementado posteriormente. El objetivo de este capítulo es entregar el marco analítico que orienta la construcción de los modelos de estimación utilizados.

2.1 Movilidad urbana y tenencia de vehículos particulares

La movilidad urbana no solo representa el conjunto de desplazamientos cotidianos de personas y mercancías dentro de una ciudad, ya sea por medios públicos o privados (Ferrovial, 2025), sino que constituye también una expresión tangible de las estructuras sociales, económicas y territoriales que configuran el entorno urbano. En contextos metropolitanos como el Gran Concepción, las dinámicas de movilidad están estrechamente vinculadas al acceso a recursos, servicios y oportunidades, lo que convierte su estudio en un componente central para entender la distribución del bienestar urbano.

La movilidad puede manifestarse en modos no motorizados, como caminar o utilizar bicicleta, o en modos motorizados, donde predominan automóviles, buses y camiones. Sin embargo, la disponibilidad de sistemas más complejos como los trenes urbanos o sistemas de metros, suele ser limitada fuera de la capital, intensificando la dependencia del vehículo particular en zonas intermedias. Por ello, comprender la movilidad urbana y su vínculo con la tenencia de vehículos resulta clave para analizar los patrones de desplazamiento, sus determinantes socioeconómicos y las implicancias para la planificación urbana.

En este marco, resulta pertinente examinar el rol específico que ha adquirido el vehículo particular en ciudades intermedias como Concepción, donde su uso refleja carencias estructurales en los sistemas de transporte, así como transformaciones en las lógicas de urbanización.

2.1.1 Rol del vehículo particular en ciudades intermedias

Las ciudades intermedias tienen dos características principales que las identifica: ejercen un reconocido rol de intermediación entre los núcleos más pequeños y las grandes áreas metropolitanas en términos económicos y sociales (Michelini, J. y C. Davies, 2009). Por otro lado, presentan

dinámicas urbanas menos densas que las grandes metrópolis como Santiago de Chile, lo que condiciona y diferencia la forma en la que las personas se movilizan.

Concepción es un gran ejemplo de una ciudad intermedia, caracterizada por servir de nodo entre pequeños centros urbanos y grandes metrópolis, y por una densidad urbana menor que la de la capital. Estas particularidades territoriales condicionan la forma en la que las personas se movilizan. En este contexto, el vehículo particular ha adquirido un rol creciente y estratégico en la movilidad urbana. Sobre todo, en Concepción y sus comunas aledañas, donde su expansión periférica ha generado desarrollo urbano más disperso, con una alta dependencia del automóvil para acceder a zonas residenciales, laborales y comerciales. Esto empeora cuando la infraestructura de transporte público es insuficiente, lo que hace del automóvil una necesidad más que una alternativa.

A pesar de que la provincia posee transporte público como el Biotrén y una red de microbuses urbanos, su cobertura y frecuencia no siempre se ajusta a las necesidades de todos los sectores y comunas. Sin ir más lejos, Biobío Chile publica en sus noticias que los mismos usuarios del Biotrén han reportado retrasos de hasta 1 hora en su recorrido más demandado el cual une Coronel y Concepción por problemas en uno de sus trenes (Friz y Belmar, 2025). La suspensión de la línea es algo que suele ocurrir, por lo que, para gente de estas comunas periféricas, se vuelve algo poco confiable, lo que genera que las familias opten por el uso de vehículos particulares. Afortunadamente se han gestionado mejoras a estos servicios como sumar más trenes a los servicios para este año 2025 (Vera y Arthur, 2024). También la implementación nuevas flotas de microbuses, mejoras en el servicio de pago con la implementación de pago con tarjeta y extensión en los horarios a las grandes comunas de la provincia (Reyes y Belmar, 2025). Sin embargo, en el corto plazo el automóvil continúa suplantando las deficiencias del transporte público en la región.

El crecimiento del parque automotriz ha generado también grandes problemas en la ciudad propiamente tal. Un claro ejemplo es la congestión vehicular que se da especialmente en accesos clave como las entradas de Concepción y el puente Llacolén desde y hacia San Pedro de la Paz. A partir de aquello, se genera una desigualdad territorial, privilegiando planificación y modelos de desarrollo urbano orientados al uso del automóvil antes de solucionar los problemas con la congestión actual y promover alternativas sostenibles.

Lo mencionado anteriormente son solo unas cuantas tendencias y desafíos que enfrenta la ciudad de Concepción respecto del rol de los vehículos particulares en ciudades intermedias. Es necesario

promover modos sostenibles de frenar la dependencia del auto particular, mejorar las intermodalidades a través de mejoras en las ciclovías, en los servicios de transporte público y peatonal. Integrando políticas de planificación urbana y transporte que favorezcan este tipo de movilidad accesible y equitativa.

2.1.2 Propiedad de vehículos y factores socioeconómicos incidentes

La posesión de vehículos particulares en los hogares es un acontecimiento estrechamente relacionado a las condiciones socioeconómicas de las familias, y se expresa tanto como una respuesta funcional a las necesidades de movilidad, como una manifestación del capital y de las preferencias individuales de consumo. Estudios previos han demostrado que factores como el ingreso del hogar, el nivel educacional, ocupación laboral, tamaño del hogar y el grado de urbanización son determinantes clave en la probabilidad de poseer uno o más vehículos. (Bocarejo & Oviedo, 2012; Cervero & Kockelman, 1997).

En particular, la disponibilidad y calidad del transporte público, la distancia a los centros urbanos y la seguridad del entorno son elementos que influyen de forma significativa en la decisión de adquirir un vehículo. En contextos donde el transporte público presenta deficiencias o no cubre adecuadamente las necesidades cotidianas, los hogares con mayor poder adquisitivo tienden a resolver su necesidad de movilidad mediante la compra de un automóvil propio (CEPAL, 2018; Rodríguez & Mojica, 2009).

De este modo, la propiedad vehicular resulta altamente correlacionada con el nivel socioeconómico, a la vez que está condicionada por aspectos del entorno físico y social. Esta relación plantea interrogantes relevantes para la planificación urbana y el diseño de políticas de transporte, por ejemplo ¿Qué factores efectivamente explican la tenencia de vehículos en distintos territorios?; ¿Cómo se puede modelar esta tenencia a partir de la información disponible?; ¿Contamos con los datos lo suficientemente precisos para ello?

Estas inquietudes motivan el análisis que se desarrolla en las secciones siguientes, centrada en los factores determinantes de la posesión de vehículos—con especial énfasis en los hogares del Gran Concepción— y las variables socioeconómicas y territoriales observables en encuestas aplicadas como son la encuesta de Caracterización Socioeconómica Nacional CASEN y la Encuesta Origen-Destino (SECTRA, 2021)

2.2 Factores determinantes de la tenencia de vehículos.

La decisión de un hogar de poseer uno o más vehículos no depende únicamente de su capacidad económica, sino que es una respuesta a una interacción compuesta entre variables individuales, sociales y territoriales, que responden a su vez a la capacidad adquisitiva y las necesidades prácticas de movilidad de cada familia. La literatura internacional y latinoamericana ha abordado esta temática desde enfoques multidisciplinarios, integrando elementos de la economía, geografía urbana y la planificación del transporte (Bhat & Pulugurta, 1998; Giuliano & Dargay, 2006).

En el caso chileno, la expansión de ciudades intermedias como el Gran Concepción, ha intensificado las demandas por movilidad motorizada. Al mismo tiempo, las condiciones del transporte público, la expansión periférica y la segregación residencial marcan diferencias en el acceso a oportunidades, lo que ha revelado distintos patrones de tenencia vehicular según varias clasificaciones como el nivel de ingreso, localización residencial y características de las viviendas. (SECTRA, 2021; MINVU, 2019).

A partir de esta perspectiva, es posible agrupar los factores determinantes de la tenencia de vehículos en tres categorías principales: factores económicos, factores sociodemográficos y factores territoriales. A continuación, se describen cada uno de ellos, junto con los hallazgos de estudios de referencia que han analizado la propiedad vehicular en distintos contextos.

2.2.1 Factores económicos

Los factores económicos constituyen tradicionalmente los principales predictores en la tenencia de vehículos de un hogar. En primer lugar, el nivel de ingreso determina no solo la posibilidad de adquirir un automóvil, sino también la capacidad de mantenerlo en el tiempo (combustible, mantenciones, seguros, estacionamientos), lo cual influye especialmente en la decisión de tener más de un vehículo por hogar (Bhat & Pulugurta, 1998). Así, el ingreso actúa como un requisito básico para la motorización: hogares de mayor renta pueden costear la compra y uso regular de vehículos, a diferencia de hogares de menor ingresos para quienes el costo fijo y variable de un auto representa una proporción considerable de su presupuesto.

Ligado a lo anterior, el nivel educativo del jefe de hogar suele estar directamente relacionado con el ingreso, lo que puede generar decisiones distintas respecto de la movilidad, planificación y consumo. Mayor educación típicamente se traduce en ingresos más altos y en una conciencia distinta respecto del uso del automóvil y alternativas de transporte. A su vez, la situación laboral incide en la necesidad de contar con medios de transporte flexibles y confiables. Trabajadores por turnos o con horarios

atípicos o en zonas mal conectadas, pueden necesitar indispensablemente vehículos particulares ante la limitada disponibilidad de transporte público en ciertos horarios y sectores. (Cervero & Kockelman, 1997). En resumen, un ingreso suficiente y estable, acompañado de un perfil socioeconómico acorde, tiende a aumentar la propensión a la posesión vehicular.

2.2.2 Factores sociodemográficos

Las características demográficas y del hogar son otro conjunto de factores clave. Hogares con mayor cantidad de integrantes adultos activos laboralmente suelen requerir más de un vehículo para satisfacer distintas demandas simultaneas de movilidad. El tamaño del hogar, la presencia de hijos e hijas y el ciclo de vida familiar influyen en las decisiones de compra de vehículos según las responsabilidades y necesidades que enfrentan.

Por otra parte, la localización residencial marca otro punto a considerar. Vivir en zonas periurbanas o rurales, con menor densidad urbana, típicamente aumenta la dependencia del vehículo particular debido a menores cobertura de transporte público, mayores distancias a servicios básicos y menor oferta local de empleo y/o educación. Estudios en América Latina, los hogares periféricos presentan tasas de motorización más altas que aquellos ubicados en áreas centrales con buena accesibilidad a transporte público. (Oviedo et al., 2016). En zonas aisladas, el automóvil se convierte prácticamente en un bien necesario para acceder a oportunidades que, en contextos urbanos centrales, podrían resolverse a pie o transporte colectivo.

2.2.3 Factores territoriales

El entorno físico y urbano incide directamente en las decisiones de movilidad. La cobertura, frecuencia y sobre todo la calidad del transporte público influye en el uso del automóvil: a mayor accesibilidad al transporte colectivo, menor dependencia del vehículo particular (Giuliano & Dargay, 2006).

Otro aspecto territorial relevante es la infraestructura vial y condiciones de desplazamiento. Además de lo mencionado anteriormente respecto de la residencia en sectores periféricos o mal conectados, la existencia de la infraestructura vial, carreteras, vías rápidas, ciclovías seguras, estacionamientos, el riesgo, existencia de la congestión vial, condicionan el uso y propiedad de automóviles. En el caso del Gran Concepción, la desconexión de algunas comunas con el centro urbano, como Hualpén o San Pedro de la Paz, genera patrones de movilidad orientados al uso del automóvil. (SECTRA, 2021). En

síntesis, la decisión de poseer un vehículo es tanto más fácil cuanto más favorable sea el entorno urbano para el auto y menos accesible sean las alternativas.

2.2.4 Estudios de referencia

Diversos estudios a nivel nacional e internacional han abordado y analizado la relación entre las condiciones socioeconómicas y propiedad vehicular, aplicando modelos estadísticos para identificar patrones comunes. En general, se encuentra una fuerte asociación positiva entre ingreso familiar y probabilidad de tener automóvil, junto con la influencia de otras variables ya mencionadas con anterioridad.

Metodológicamente, muchos trabajos emplean modelos de regresión logística binaria o modelos logit multinomiales/ordenados para predecir la probabilidad de poseer vehículos en función de dichas variables explicativas. Por ejemplo, Bhat y Pulugurta (1998) desarrollaron un modelo logit multinomial para la tendencia de 0, 1 o más de 2 autos en hogares, encontrando que el ingreso, el número de trabajadores en el hogar y disponibilidad de transporte alternativo eran factores críticos. Cervero y Kockelman (1997) estudiaron como variables del entorno construido junto con características socioeconómicas afectan la propiedad de autos en ciudades norteamericanas, hallando efectos significativos de la densidad urbana.

En el caso chileno, estudios aplicados a Santiago, Valparaíso y Temuco reflejan tendencias similares: a medida que mejora el nivel de ingresos de los hogares, tienden a aumentar los niveles de motorización; pero también, en zonas mal conectadas, hogares de ingresos medios o bajos realizan grandes esfuerzos para adquirir al menos un automóvil. (CEPAL, 2018; Tovar & Rodríguez, 2020).

La literatura evidencia un conjunto consistente de factores determinantes: ingreso y capacidad económica como habilitantes primarios; características del hogar modulando la necesidad de vehículos; y el contexto territorial, facilitando o restringiendo la dependencia de automóvil. Este marco teórico sirve de base para el presente estudio, donde se busca modelar la posesión de automóviles en función de variables observables en el Gran Concepción.

2.3 Ingreso como variable latente y métodos de estimación

Para comprender la relación entre el nivel socioeconómico y la tenencia de vehículos es fundamental disponer de medidas fiables de ingreso de los hogares. Sin embargo, la estimación precisa de los ingresos enfrenta importantes limitaciones prácticas y metodológicas. Por un lado, las encuestas, censos y encuestas de movilidad frecuentemente omiten consultas directas sobre ingresos

o enfrenten una alta tasa de no-respuesta y subdeclaración (Ministerio de Desarrollo Social, 2020). En Chile, en contextos de informalidad laboral, el 27.5% de los trabajadores no declara sus ingresos (INE, 2021) lo que genera un problema de información incompleta. Por otro lado, los censos nacionales —que cubren a toda la población— típicamente no incluyen información de ingreso para evitar sesgos y reducir la carga respondiente.

Frente a estas limitaciones, han cobrado relevancia los métodos indirectos de estimación basado en modelos estadísticos y de aprendizaje automático, que permiten inferir o predecir ingresos a partir de variables proxy disponibles. La idea central es aprovechar fuentes de datos complementarias que sí posean ingresos declarados para entrenar un modelo de predicción, y luego aplicar dicho modelo a registros censales. Este enfoque general, conocido en la literatura como estimación por áreas pequeñas, combina la granularidad de las encuestas con la cobertura exhaustiva del censo para obtener estimaciones locales más precisas. (Elbers et al., 2003). El método desarrollado por Elbers, Lanjouw & Lanjouw (2003), por ejemplo, es un precursor en combinar datos de encuesta y censo para estimar el bienestar a nivel micro-territorial, asumiendo que un conjunto de variables comunes permite vincular ambas fuentes.

A continuación, se revisan los principales enfoques metodológicos para la estimación de ingresos: desde modelos econométricos tradicionales hasta técnicas modernas de aprendizaje automático como el desarrollado en esta memoria, destacando sus características, supuestos y pertinencia para el problema de estudio.

2.3.1 Modelos econométricos

2.3.1.1 Regresión tradicional

El enfoque econométrico clásico para estimar ingresos se apoya en la regresión lineal múltiple, donde la variable dependiente se modela como una combinación lineal de varios predictores. (Granados, 2016). Este modelo asume que existe una relación aproximadamente lineal entre los factores explicativos y el ingreso en este caso, buscando estimar coeficientes β_j que minimicen el error cuadrático medio y adopta la siguiente formulación general:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \qquad (1)$$

Donde Y_i es la variable dependiente o explicada. X_{ij} son las variables independientes o explicativas. β_j son los coeficientes por estimar y ϵ_i representa el término del error aleatorio.

Los coeficientes β se estiman típicamente por mínimos cuadrados ordinarios (OLS) bajo supuestos de linealidad, homocedasticidad de varianza y no fuerte colinealidad (Wooldridge, 2019).

Una ventaja de este modelo es la fácil interpretación económica de los β como efectos marginales: por ejemplo, cuando aumenta en promedio el ingreso cada año adicional de educación del jefe de hogar, manteniendo los demás factores constantes (*ceteris paribus)*.

Además, en este modelo, si se aplica una transformación logarítmica, pueden interpretarse los coeficientes como elasticidades de ingreso (Wooldridge, 2019), es decir, cambios porcentuales en el ingreso ante cambios porcentuales en los predictores. Esta variante *log-linear* resulta útil cuando la distribución de ingresos es asimétrica y se desea reducir la influencia de valores atípicos (Rodríguez, 2001).

No obstante, los modelos lineales presentan limitaciones para este tipo de problemas. El ingreso suele tener una distribución altamente sesgada, y la relación con los predictores podría ser no lineal o presentar interacciones complejas. Además, la omisión de variables relevantes puede sesgar estas estimaciones. A pesar de ello, la regresión lineal múltiple continúa siendo una line base metodológica en estudios de ingresos, especialmente útil cuando el tamaño muestral es pequeño o se requiere de inferencia estadística sobre la influencia de cada factor (Angrist & Pischke, 2009). Su uso en estimaciones de ingreso es apropiado siempre que se tomen precauciones ante supuestos incumplidos y se complementen con análisis de diagnóstico.

2.3.1.2 Métodos de aprendizaje automático aplicado

En años recientes, los modelos de aprendizaje automático (*machine learning*) han emergido como herramientas efectivas para abordar estimación de ingresos, especialmente cuando se trata de relaciones no lineales e interacciones complejas, dinámicas o no convencionales entre variables (Hastie et al., 2009). A diferencia de los modelos paramétricos tradicionales, estos métodos pueden capturar estas interacciones sin requerir especificación previa. Entre estos métodos, los árboles de decisión constituyen una primera aproximación dentro de este conjunto metodológico, operando mediante particiones recursivas del espacio de características a través de reglas binarias. En su versión de regresión, la predicción en cada nodo terminal corresponde al valor promedio de las observaciones

contenidas en dicho nodo. (James et al., 2013) Si bien un solo árbol de decisión es intuitivo y flexible, suele presentar alta varianza, y por tanto riesgo de sobreajuste.

Para superar estas limitaciones, Breiman (2001) introdujo el método *Random Forest*, el cual representa una evolución sustancial a los árboles individuales, combinando dos principios fundamentales: el *bagging (boostrap aggregating)* y la selección aleatoria de características. Este es un ensamble de muchos arboles de decisión sobre diferentes muestras *boostrap* del conjunto de datos y considerando solo un subconjunto aleatorio de predictores en cada división. Al promediar las predicciones de múltiples arboles no correlacionados, *Random Forest* logra reducir significativamente la varianza del modelo, mejorando la precisión predictiva sin aumentar demasiado el sesgo. La predicción $\hat{Y}_{RF}$ de ingreso para un hogar se obtiene promediando las predicciones individuales de B arboles

$$\hat{Y}_{RF}(i) = \frac{1}{B}\sum_{b=1}^{B} \hat{y}_b(i) \tag{2}$$

Donde $\hat{y}_b(i)$ es la predicción del árbol b-ésimo para la observación i. Este modelo a demostrado buen desempeño en estimación de ingresos y otras variables económicas al capturar relaciones no lineales que serían difíciles de modelar con otro tipo de regresiones.

Otro conjunto poderoso de técnicas son los métodos de *boosting*, en específico implementaciones como *XGBoost (Extreme Gradient Boosting)* (Chen & Guestrin, 2016). El *boosting* construye iterativamente un ensamble aditivo de árboles, donde cada nuevo árbol se entrena para corregir los errores del ensamble acumulado hasta el momento. Este algoritmo optimiza una función objetivo compuesta por una medida de perdida, más términos de regularización que penalizan la complejidad del modelo, evitando así sobreajuste. Este y otros algoritmos de *boosting* pueden lograr precisión al modelar interacciones complejas, al costo de una menor interpretabilidad.

En síntesis, los métodos de *machine learning* ofrecen una alta capacidad predictiva y flexibilidad. No obstante, suelen requerir conjuntos de datos amplios y representativos para alcanzar su máximo potencial. Asimismo, privilegian la predicción sobre la interpretabilidad: es decir, son ideales cuando el objetivo es minimizar el error de predicción, pero menos útiles si se busca entender el efecto causal de cada variable. (Berk, 2016). Dado que en este estudio el interés principal es obtener una estimación

consistente del ingreso de hogares para el insumo del modelo de estimación vehicular, estos modelos aparecen como candidatos atractivos.

2.3.2 Evaluación comparativa de metodologías y selección

La selección del método de estimación más adecuado requiere considerar diversos aspectos metodológicos y prácticos (Molinaro et al., 2005). El tamaño y calidad de los datos disponibles constituyen un factor determinante, con muestras pequeñas, un modelo lineal parsimonioso puede ser más estable, mientras que con bases de datos grandes y muchas variables, métodos de *Machine Learning* pueden explotar mejor la información disponible. También influye la finalidad del análisis: si el objetivo es explicativo, los modelos tradicionales permiten extraer coeficientes interpretables y probar hipótesis; en cambio, en objetivos predictivos, modelos de *Machine Learning* suelen rendir mejor en términos de error de predicción (Angrist & Pischke, 2009; Molinaro et al., 2005)

En este estudio, donde se busca imputar y predecir ingresos faltantes en el Censo a partir de la encuesta Casen, el énfasis está en la precisión global de la predicción más que en la interpretación de cada coeficiente. Por ello, se optará por un enfoque de aprendizaje automático, validando su desempeño frente a métodos econométricos base. No obstante, es fundamental llevar a cabo evaluaciones rigurosas de cada modelo. Se contemplan comparaciones de desempeño mediante evaluación cruzada y conjuntos de prueba, análisis de error, así como la inspección de la importancia de las variables en el método seleccionado. Estas evaluaciones aseguran que la metodología supera las limitaciones de medición directa y ofrecen un marco robusto para estimar el nivel socioeconómico de los hogares con datos incompletos. (Ferreira et al., 2016 & Chen & Guestrin, 2016).

En efecto, esta aproximación adquiere especial relevancia al analizar fenómenos como la propiedad de vehículos, donde la disponibilidad de información financiera directa es limitada. La evidencia empírica internacional sugiere una relación estrecha entre ingresos y tenencia vehicular: estudios de la Organización para la Cooperación y el Desarrollo Económico (OCDE, 2019) indican que los hogares indican que los hogares del quintil superior de ingresos tienen 3.2 mayor probabilidad de poseer al menos un automóvil comparado con el quintil inferior.

En Chile la encuesta Casen (Ministerio de Desarrollo Social, 2020) revela que el 84% de los hogares del decil con mayores ingresos posee vehículo, frente a solo un 9% del decil con menores ingresos, evidenciando un gradiente socioeconómico marcado en el acceso a vehículos. Desde una perspectiva teórica la relación ingreso-motorización se sustenta en la teoría microeconómica del consumidor,

donde el vehículo actúa como bien normal superior cuya demanda crece más que proporcionalmente al ingreso disponible (Varian, 2014; Deaton & Muellbauer, 1980). Es decir, a mayor ingreso, mucho mayor es la probabilidad de comprar un auto y de incrementar su uso.

Este marco conceptual, junto con los métodos de estimación revisados, justifica plenamente el uso de proxis de ingreso para predecir patrones de tenencia vehicular en ausencia de datos directos, particularmente en áreas urbanas como el Gran Concepción, donde la movilidad está fuertemente estratificada (CEPAL, 2016). Un modelo entrenado en nuestro caso con datos de encuesta permitirá aproximar el nivel de ingreso de cada hogar censado, lo que habilita explorar su relación con la probabilidad de poseer automóvil a pequeña escala.

2.4 Relación entre ingreso y propiedad de vehículos

Este análisis presenta la relación existente entre ingresos y tenencia vehicular, abordando los fundamentos teóricos, evidencia empírica y metodologías de estimación. Investigaciones revela una fuerte correlación positiva entre niveles de ingreso y propiedad de vehículos, con notables disparidades entre quintiles socioeconómicos tanto en cantidad como en la calidad del transporte. Esta relación tiene importantes implicaciones para políticas públicas de transporte, planificación urbana e inclusión social. Es importante mencionar que, la elasticidad y forma de la relación entre estos factores puede variar según el contexto y otros condicionantes. En esta sección se abordan: (1) los fundamentos teóricos que explican el vínculo ingreso-motorización. (2) evidencia empírica comparada a nivel global y local; (3) metodologías para estimar dicha relación cuando no se dispone de datos directos; (4) un caso de estudio ilustrativo del uso de modelos predictivos; y (5) las implicancias de estos hallazgos para políticas públicas.

2.4.1 Fundamentos teóricos de la relación ingreso-motorización

La teoría microeconómica clásica establece que los vehículos son bienes normales superiores, conocidos también como bienes de lujo, cuya demanda aumenta más que proporcionalmente con el ingreso disponible (Varian, 2014). Este comportamiento se explica mediante la elasticidad ingreso de la demanda, definida como el cambio porcentual en la cantidad demandada ante la variación del 1% en el ingreso. Estudios empíricos han estimado que las magnitudes de las elasticidades ingreso y precio se modifican de acuerdo con las características y evolución de la flota vehicular (Galindo et al., 2015). Esta diferencia en las elasticidades precio de la demanda de gasolinas puede explicarse, entre otros factores, por la mayor disponibilidad de sustitutos del transporte privado en países desarrollados.

Adicionalmente, la teoría del capital social (Bourdieu, 1986) sugiere que la propiedad de bienes durables como automóviles opera como un indicador de estatus económico, reforzando ciclos de movilidad ascendente. En contextos urbanos fragmentados el acceso a vehículos particulares facilita la conexión con centros laborales, educativos y de servicios, generando un efecto acumulativo en la generación de ingresos. Según un estudio publicado en 2022 que utilizo datos del panel *Study of Income Dynamics*, en comparación con los jóvenes que tuvieron acceso continuo a un automóvil en la infancia, aquellos que no tuvieron este beneficio estudiaron menos, obtuvieron ingresos más bajos y tuvieron tasas de desempleo más altas. (Ralph, 2022). Los resultados de este estudio sugieren que la desventaja del transporte contribuye a los bajos niveles de movilidad económica intergeneracional.

2.4.2 Evidencia empírica a escala global y local

La relación positiva entre ingresos y tenencia vehicular se ha estudiado y verificado en múltiples contextos geográficos. En América Latina, el rápido crecimiento económico y urbano ha ido de la mano con un aumento acelerado de la motorización. Por ejemplo, América Latina experimento entre 1995 y 2009 un fuerte crecimiento poblacional urbano, pasando de 475 a 575 millones de habitantes, lo cual tensionó la oferta de servicios públicos en las ciudades. En paralelo, se observó un incremento significativo en la tasa de motorización regional, impulsado por el aumento de ingresos medios y la falta de transporte publico suficiente (CAF, 2011).

En Chile, la encuesta Casen 2020 (Ministerio de Desarrollo Social, 2020) revela ciertas disparidades significativas en la distribución del parque vehicular. El quintil 5 concentra un 63% del parque vehicular nacional, con un promedio de 1,8 vehículos por hogar, en tanto el quintil 1 concentra apenas un 4% del parque, con 0,1 vehículos por hogar. Esta brecha se amplía al considerar la calidad y antigüedad de los vehículos: el quintil superior posee vehículos con un valor promedio de 15 millones de pesos chilenos frente al quintil inferior con un promedio de 2.5 millones de pesos. (INE, 2021)

La encuesta Casen 2020 enfrento desafíos únicos debido a la pandemia, implementando un diseño mixto entre presencial y telefónico para sus encuestas con tal de garantizar representatividad. A pesar de esto, los datos sobre tenencia vehicular se consideran robustos con un error de muestreo máximo de 0,4% a nivel nacional y 1,6% a nivel regional (Ministerio de Desarrollo Social, 2020)

Otro detalle a nivel territorial local de alta relevancia es que, en regiones como La Araucanía y Ñuble, la tenencia vehicular en el decil más bajo no supera el 5%, mientras que en la región metropolitana

alcanza el 11%, reflejando desigualdades intrarregionales en acceso a movilidad (Subsecretaría de Desarrollo Regional, 2021)

La evidencia empírica muestra un patrón robusto: a mayor nivel socioeconómico, mayor la probabilidad de poseer vehículos, pero la pendiente de esa relación puede variar según la estructura de la ciudad, políticas públicas vigentes y otras variables. Esto refuerza la importancia de estudiar localmente la interacción ingreso-automóvil.

2.4.3 Metodologías para estimar la relación en ausencia de datos directos

Cuando los datos de ingreso son incompletos o inexistentes, se emplean técnicas de estimación indirectas basadas en variables proxy y modelos estadísticos avanzados. Estas metodologías permiten aproximar la relación entre ingresos y posesión vehicular incluso cuando no se dispone de información directa sobre ingresos familiares. Diversos indicadores observables pueden actuar como sustitutos del nivel de ingreso o riqueza de un hogar. Entre los más utilizados están: el nivel educacional del jefe de hogar; las características de la vivienda, acceso a servicios básicos y la posesión de otros bienes durables. Estas variables, disponibles en censos y encuestas, funcionan como indicadores indirectos del estrato socioeconómico del hogar, permitiendo estimaciones cuando no se cuenta con datos de ingresos directos.

2.4.3.1 Modelo de regresión Poisson

En el contexto de la estimación de posesión de automóviles, resulta pertinente considerar modelos estadísticos que respeten la naturaleza discreta y no negativa de la variable de interés. Si bien una regresión logística es adecuada cuando se modela la probabilidad de poseer al menos un vehículo, cuando el objetivo es modelar directamente el número de automóviles por hogar, se vuelve necesario emplear un modelo capaz de capturar esa característica. De todas formas, se evaluarán ambos modelos de manera complementaria y se detallara más adelante.

El modelo de regresión Poisson pertenece a la familia de los modelos lineales generalizados y está especialmente diseñado para variables de conteo que asumen valores enteros no negativos. Este modelo supone que la variable dependiente Y_i, número de vehículos en el hogar i, sigue una distribución Poisson con media condicional μ_i, de modo que:

$$Y_i \sim Poisson(\mu_i), P(Y_i = y) = \frac{e^{-\mu_i}{\mu_i}^y}{y!} \quad (2)$$

Donde la media μ_i esta relacionada linealmente en escala logarítmica con las variables explicativas:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \, x_{ij} \qquad (3)$$

La función de enlace logarítmica asegura que las predicciones sean siempre positivas y continuas en el rango de conteo posibles

La literatura respalda el uso del modelo Poisson para este tipo de fenómenos cuando los conteos son relativamente bajos y la varianza no difiere considerablemente de la media (equidispersión), condiciones que se evaluaron en su aplicación.

2.4.3.2 Modelo de regresión logística

Los modelos de regresión logística permiten estimar la probabilidad de tenencia vehicular basándose en las variables proxy mencionadas. La estructura del modelo es la siguiente:

$$P(Vehiculo = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \sum_k \beta_k \cdot X_k)}} \qquad (5)$$

Donde X_k incluyen las proxies socioeconómicas y otras variables de control relevantes. Los coeficientes β estimados nos indican la influencia de cada factor en la probabilidad de tener vehículo. Estos modelos son particularmente útiles en contextos con datos incompletos, ya que permiten incorporar múltiples variables sustitutas en ausencia del ingreso real, y aun así obtener una buena predicción de que hogares tienen automóvil. Es importante, no obstante, considerar posibles sesgos de omisión: la ausencia de alguna variable correlacionada tanto con ingreso como con tenencia de vehículos podría distorsionar la estimación.

2.4.4 Caso destacado: Modelos Predictivos con Machine Learning

Un caso reciente desatacado que ilustra el uso de modelos de predicción con machine learning en el ámbito de transporte es el estudio de Morales (2020) de la Universidad de Chile, titulado: "Uso de modelos de predicción y estimación de tiempos de traslado entre dos puntos utilizando datos de GPS". En este trabajo se emplearon modelos de *Random Forest* para predecir tiempos de viaje en la ciudad de Santiago a partir de datos masivos de GPS. Si bien el objetivo difiere del nuestro, la metodología es análoga: Usar grandes volúmenes de datos con múltiples variables para realizar predicciones precisas de un fenómeno de movilidad. El modelo de *Random Forest* permite incorporar múltiples variables predictoras y capturar relaciones no lineales entre ellas, lo que resulta

particularmente útil para modelar fenómenos complejos como la adquisición de vehículos en diferentes contextos socioeconómicos.

2.4.5 Implicancias para políticas publicas

La estrecha relación entre ingresos y motorización tiene repercusiones críticas en la planificación urbana, debido a que el diseño de redes de transporte público debe estar enfocado en la accesibilidad a grupos de menores ingresos. Las políticas ambientales mediante impuestos verdes diferenciados por nivel socioeconómico y programas de acceso a vehículos eléctricos con subsidios para hogares vulnerables deben considerar estas diferencias en el acceso a movilidad.

El fuerte crecimiento de las principales urbes latinoamericanas ha tenido un impacto importante en los sistemas viales, congestión vehicular, el estado del transporte y los servicios públicos (CAF, 2011). Las políticas públicas deben abordar no solo aspectos relacionados a movilidad, sino también a la equidad en el acceso a oportunidades que proporcionan los diferentes medios de transporte, especialmente considerando las brechas identificadas entre grupos socioeconómicos.

La estrecha relación entre ingresos y motorización en el Gran Concepción tiene profundas implicancias para la planificación urbana y la movilidad. El diseño de redes de transporte debe priorizar la accesibilidad de los grupos de menores ingresos, promoviendo sistemas integrados y eficientes que reduzcan la dependencia del automóvil privado y favorezcan la equidad del acceso a oportunidades urbanas (Ministerio de Desarrollo Social, 2020; Campos Miranda, 2004). Además, la evidencia muestra que la distribución de la propiedad vehicular refleja y amplifica las desigualdades socioeconómicas, por lo que medidas como subsidios al transporte público, incentivos para la adopción de vehículos menos contaminantes y políticas diferenciadas por estrato resultan fundamentales para avanzar hacia una movilidad más inclusiva y sostenible (INE, 2021)

En conclusión, la evidencia revisada confirma que la tenencia de vehículos está fuertemente condicionada por el nivel de ingresos. Esta correlación no solo impacta la movilidad cotidiana, sino que también incide en la calidad de vida y en el acceso a servicios y oportunidades. Por ello, comprender y modelar adecuadamente esta relación resulta esencial para el diseño de intervenciones urbanas y de transporte que respondan a las necesidades reales de la población y contribuyan a reducir las brechas de desigualdad. En la siguiente sección se presentan enfoques metodológicos que permiten estimar y analizar la relación entre ingreso y propiedad de vehículos.

2.5 Enfoque metodológico propuesto

A partir del marco teórico desarrollado, se define una metodología de estimación de dos etapas para el caso de estudio de la provincia de Concepción. En la primera etapa se implementó un modelo predictivo de ingreso autónomo del hogar empleando técnicas de aprendizaje automático, entrenado con datos de la encuesta Casen 2017. En la segunda etapa, se utilizaron las estimaciones de ingreso junto a otras variables para modelar la posesión de automóviles a nivel hogar.

Para esta segunda etapa se proponen dos modelos complementarios, con el objetivo de abordar la problemática desde dos perspectivas: (1) un modelo de regresión logística binaria para estimar la probabilidad de que un hogar posea al menos un vehículo, y (2) un modelo de regresión Poisson para estimar directamente el número de vehículos por hogar. El modelo logístico es apropiado para una respuesta dicotómica, mientras que el Poisson es especialmente adecuado para variables de conteo, ya que asume valores enteros no negativos y permite capturar la distribución completa de la variable dependiente. Ambos modelos fueron validados para asegurar que los supuestos requeridos se cumplieran y que las predicciones fueran coherentes con la teoría y evidencia empírica.

Esta estrategia metodológica combinada permite aprovechar las fortalezas de cada modelo, facilitando un análisis integral y robusto del fenómeno de la tendencia de automóviles en los hogares de la región.

3. Caso de estudio: Provincia de Concepción

Este capítulo describe el caso de estudio seleccionado, detalla las fuentes de datos empleadas y expone la metodología implementada para construir los modelos de estimación de ingresos y posesión de vehículos. Se explican los criterios de preparación y homologación de variables entre encuestas, se especifican y validan los modelos aplicados y se describe como se realizó su aplicación sobre los datos censales. El propósito de este capítulo es documentar el proceso técnico-metodológico que permite vincular las variables observadas con los resultados obtenidos en la etapa siguiente.

3.1 Objetivo y contexto del caso

El objetivo es desarrollar modelos de estimación socioeconómica aplicados a la provincia de Concepción, utilizando como base la Encuesta de Caracterización Socioeconómica nacional Casen 2017 y el Censo Nacional de Población y Vivienda 2017. Se busca en primer lugar, estimar el ingreso autónomo del hogar a partir de variables observables, y, en segundo lugar, emplear dicha estimación como insumo en un modelo explicativo sobre la posesión de automóviles.

La elección de la provincia de Concepción obedece a su importancia territorial en Chile, por su composición urbana, densidad poblacional, y la disponibilidad de datos desagregados. El estudio se enmarca en el análisis de condiciones de vida urbana, movilidad y acceso a bienes durables, siendo el ingreso una variable clave para aproximarse a estas dimensiones.

La metodología general combina ambas encuestas aprovechando las fortalezas de cada fuente, donde el Censo proporciona cobertura de los hogares de las 12 comunas del área metropolitana, pero carece información directa sobre ingresos. Por su parte la encuesta Casen incluye datos detallados de ingresos, pero con una muestra limitada.

3.2 Fuentes de datos

3.2.1 Encuesta Casen 2017

La encuesta de Caracterización Socioeconómica Nacional Casen es una encuesta representativa a nivel nacional y regional, de tipo muestral, que recoge información sobre educación, salud, ingresos, vivienda y empleo. Su unidad de observación es la persona, con identificación del hogar al que pertenece. Es la única fuente que entrega ingresos declarados, por lo que se utiliza como base de entrenamiento para estimar ingresos.

Cabe mencionar que, si bien la Encuesta Casen 2017 es representativa a nivel nacional y regional, no posee representatividad estadística a nivel comunal. En este estudio, su utilización con desagregación

comunal se justifica con fines exploratorios y metodológicos, especialmente para el uso de entrenamiento de modelos de predicción que posteriormente son aplicados sobre los registros censales representativos, cuya cobertura exhaustiva sí permite realizar análisis y descripciones más detalladas a nivel comunal. No obstante, debe considerarse que las estimaciones obtenidas a partir del modelo siguen siendo predicciones, y no valores directamente observados.

3.2.2 Censo 2017

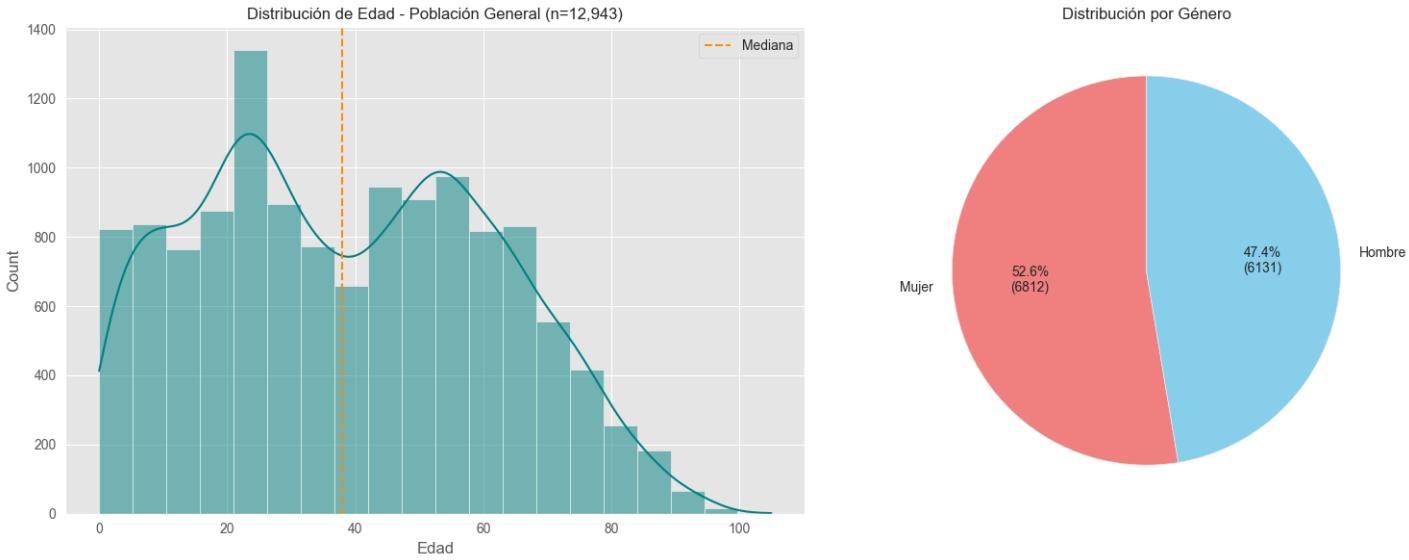
El Censo de Población y Vivienda 2017 es un levantamiento exhaustivo de tipo censal, cuya cobertura incluye todas las viviendas y personas residentes del país. Si bien no contiene variables económicas como ingreso, si permite acceder a una amplia gama de indicadores sociodemográficos y de vivienda. Su ventaja radica en la alta cobertura territorial, lo que permite realizar inferencias a mayor escala.

3.3 Preparación y descripción bases de datos

Se comienza con la preparación y armonización de ambas bases de datos, obteniéndolas directamente de las páginas oficiales, Pagina del Ministerio del Desarrollo Social y Familia para le encuesta Casen 2017, donde se accede a la base de datos general y la lectura se complementa con el Libro de Códigos Base de Datos Casen 2017 que permite comprender los códigos utilizados en esa base. Por otro lado, la página del Instituto Nacional de Estadísticas se utiliza para la obtención de la base de datos del Censo separadas en 2 archivos de Viviendas y Personas, además de un Manual de Usuario que tiene el mismo propósito del libro de códigos anteriormente mencionado.

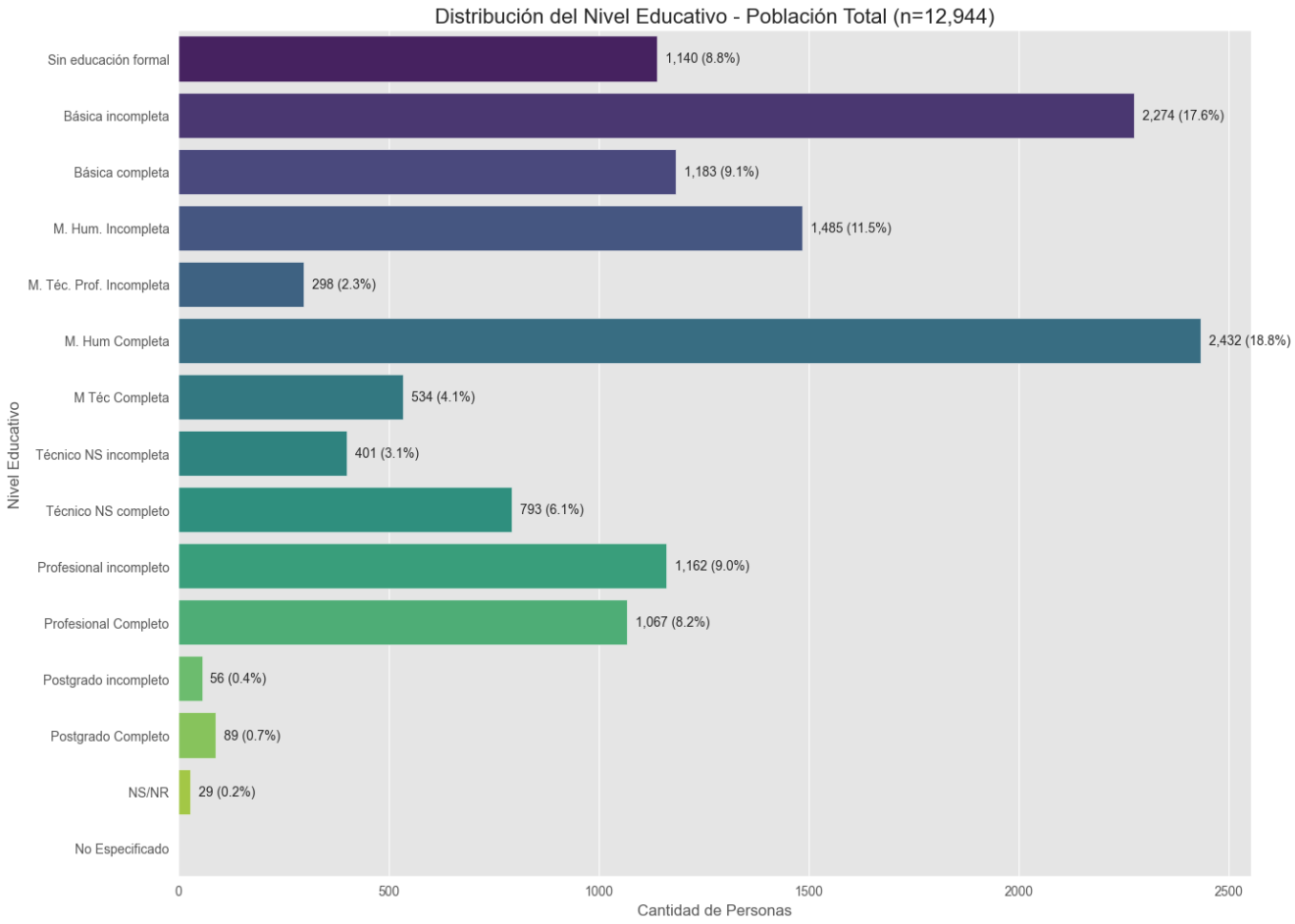
3.3.1 Datos encuesta Casen

Se realiza una descripción de los datos mediante la realización de gráficos que permiten conocer las características demográficas de la población como edad, sexo y nivel educacional según la base de datos de la encuesta Casen.



Fuente: Base de datos Casen. Elaboración Propia

Figura 3.1: Distribución de edad y distribución por género Casen



Fuente: Base de datos Casen. Elaboración Propia

Figura 3.2: Distribución del nivel educativo Casen

Los ingresos son la parte más importante de esta encuesta, siendo el ingreso autónomo la variable escogida a utilizar para su análisis. Esta corresponde a:

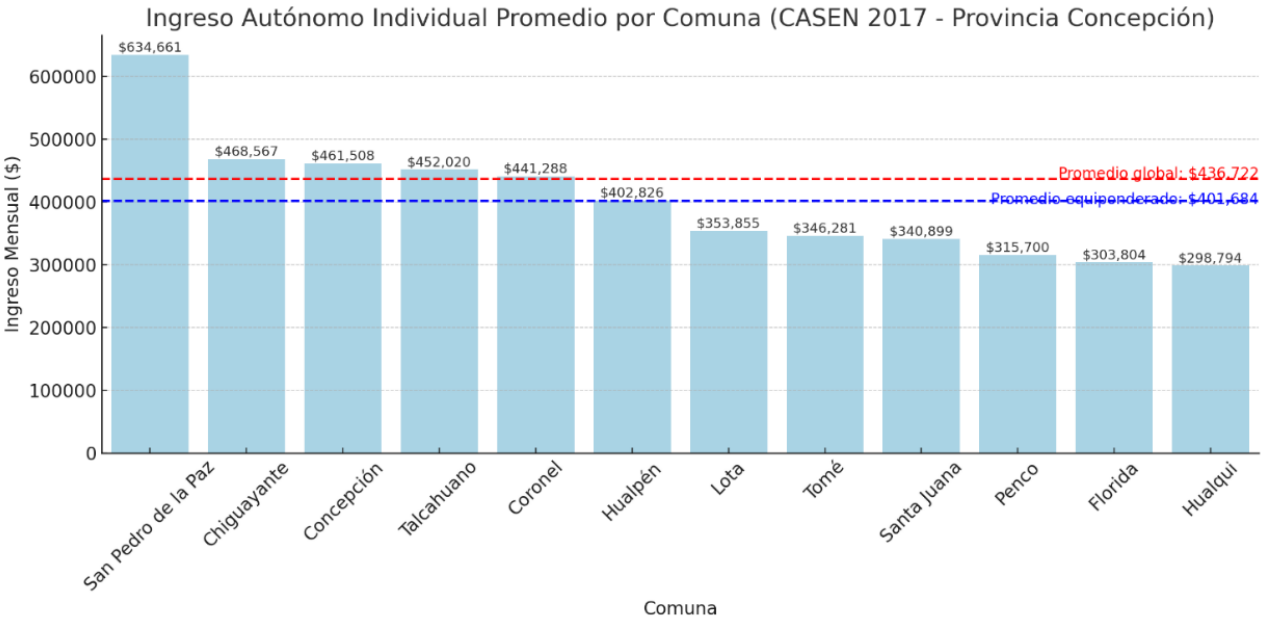
> La suma de todos los pagos que obtiene el miembro del hogar, excluido el servicio doméstico puertas adentro, provenientes del trabajo como de la propiedad de los activos. Estos incluyen sueldos y salarios, monetarios y en especies, ganancias provenientes del trabajo independiente, la auto provisión de bienes producidos por el hogar, rentas, intereses, dividendos y retiro de utilidades, jubilación, pensiones o montepíos, y transferencias corrientes. (Ministerio del Desarrollo Social, 2018)

La base de datos utiliza la variable definida como "*yaut*" para hacer alusión al ingreso autónomo. Se utilizará la variable "*yautcorh*" que se define como ingreso autónomo corregido por hogar.

“Corregido” se refiere a que la variable original de ingreso se ha sometido a procesos de imputación y ajustes metodológicos con el fin de mejorar la calidad y representatividad de la información. Correcciones como imputación de ingresos faltantes o erróneos, consistencia entre tipo de ocupación e ingreso declarado y ajustes por subdeclaración sistemática. Estas correcciones están documentadas por el Ministerio y buscan que el ingreso refleje de mejor forma la realidad económica de las personas y los hogares.

El uso de la variable de ingreso autónomo ante otras como “ingreso monetario”, que corresponde a los ingresos autónomos en adición a subsidios monetarios percibidos, se justifica en la obtención de una mejor aproximación al ingreso real de la población. Al excluir las transferencias del estado, se mide la capacidad de generación de los ingresos propia de las personas u hogares. Lo que permite estimar modelos predictivos con menos interferencias externas a las propias familias, además, las transferencias del estado, bonos o subsidios son generalmente variables por gobierno, temporales y focalizadas.

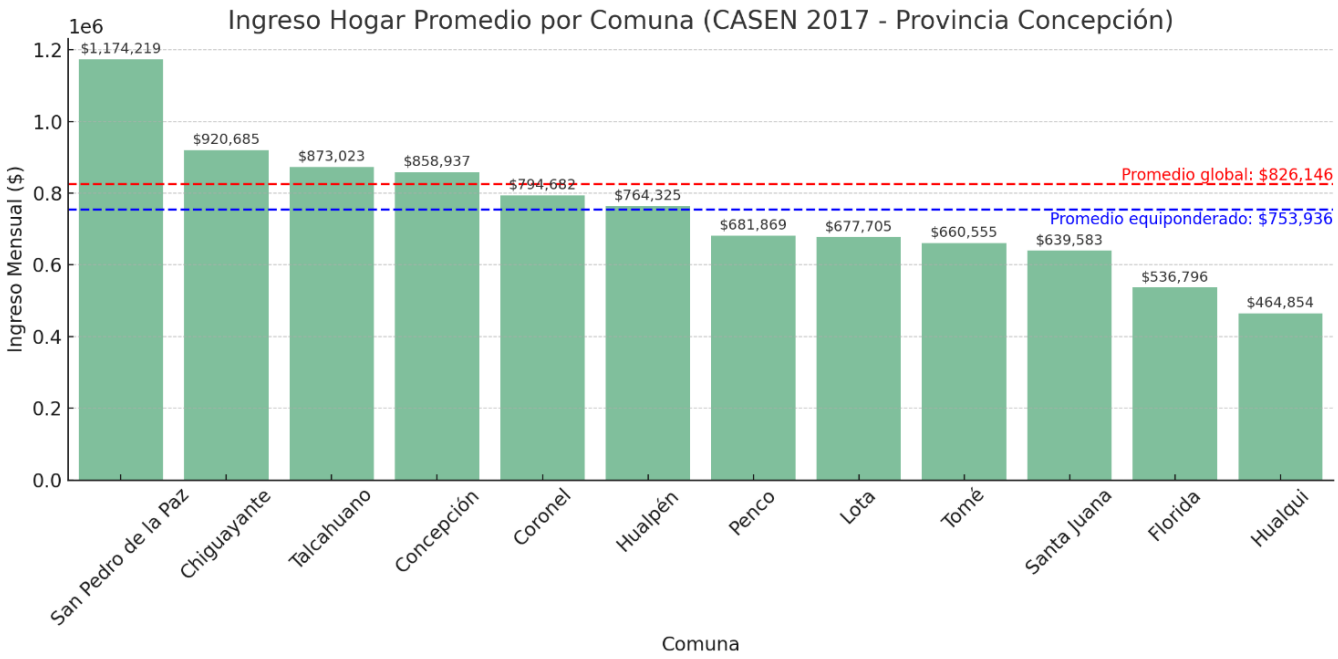
Se presenta en la figura 3.3 el ingreso autónomo individual promedio por comuna, donde se cuentan con 7.218 registros, que representan 55,8% del total. Se entiende que solo ese porcentaje tiene un ingreso autónomo individual declarado, lo que se explica por la cantidad de personas que no tienen ingreso real y fue no declarado en la encuesta.



Fuente: Base de datos Casen. Elaboración Propia

Figura 3.3: Ingreso individual promedio por comuna Casen

Finalmente, de manera agregada se tiene el ingreso autónomo del hogar corregido, con 4.015 hogares con registro, que representan el 93,6% del total de datos. Presentados en la figura 3.4 a continuación.



Fuente: Base de datos Casen. Elaboración Propia

Figura 3.4: Ingreso del hogar promedio por comuna Casen

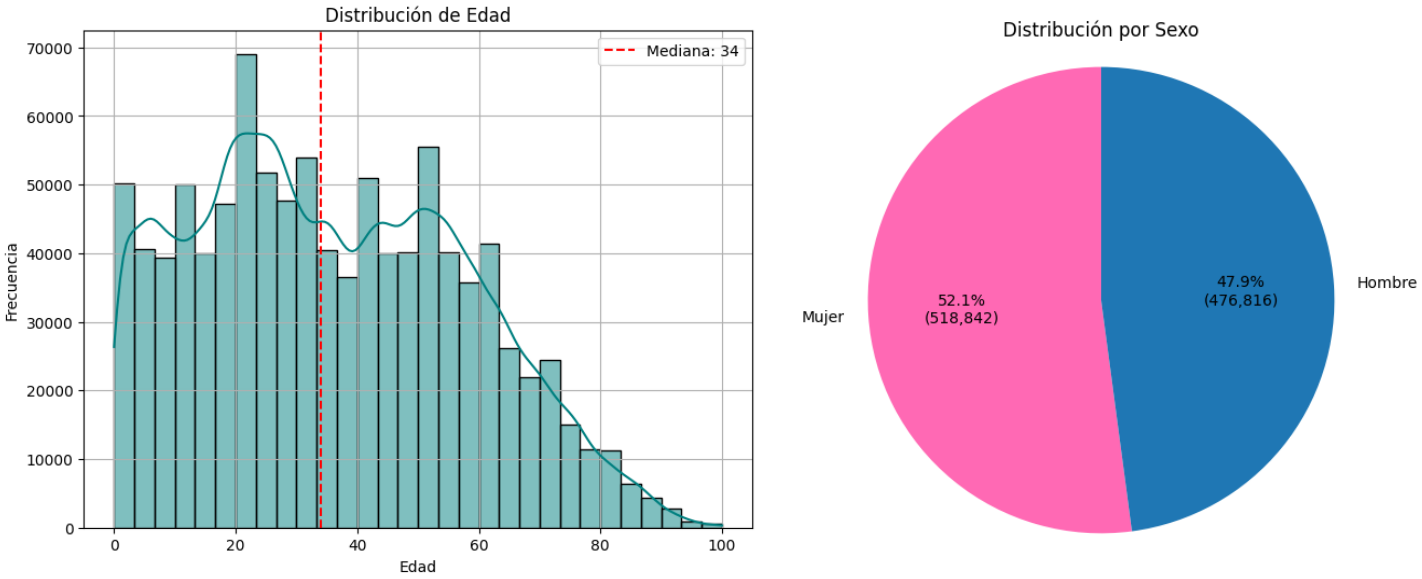
En la Síntesis de Resultados de Ingresos de los Hogares de la encuesta Casen (Ministerio de Desarrollo Social, 2017) se reporta el ingreso autónomo promedio del hogar para la región de Biobío es de $696.965. La figura 3.4 nos muestra un resultado de promedio global para este apartado de $826.146. El cual representa únicamente el ingreso autónomo del hogar en la provincia de Concepción, excluyendo las provincias de Bio Bío y Arauco las cuales conforman la región.

3.3.2 Datos Censo

La encuesta Censo se diseñó con el fin contar y caracterizar a todos los habitantes y viviendas de Chile. Tiene un tamaño muestral prácticamente igual a la población lo que permite una mayor precisión en términos de estimaciones agregadas y desagregadas.

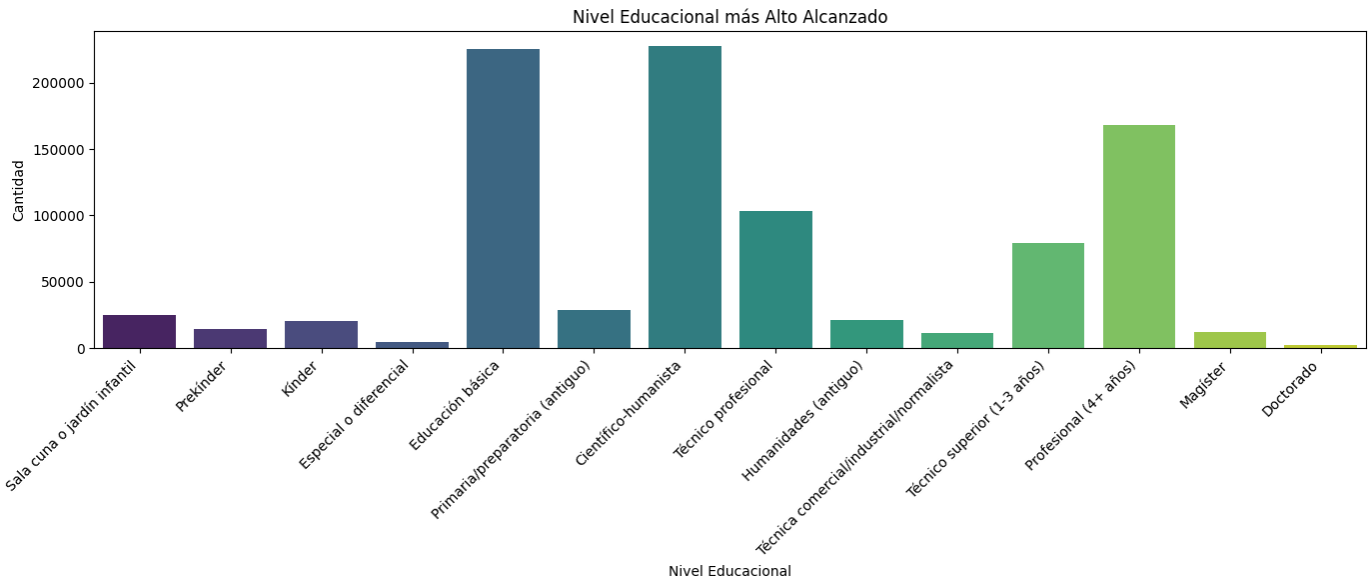
A pesar de tener características similares, el censo busca información estructural del país, como el tamaño, distribución, composición y características generales de la población y vivienda. En cambio, la encuesta Casen tiene como objetivo medición de la pobreza, desigualdad y evaluación de políticas sociales. Es así como se busca complementar ambas encuestas para tener un análisis más robusto.

La base de datos asociadas a personas del censo nos permite acceder a información semejante a la descrita en la sección anterior para la encuesta Casen, como la edad y sexo en la figura 3.5, y educación en la figura 3.6. La cual será descrita a continuación con el fin de comparar los resultados entregados entre ambas encuestas.



Fuente: Base de datos Censo. Elaboración Propia

Figura 3.5: Distribución de edad y distribución por sexo Censo



Fuente: Base de datos Censo. Elaboración Propia

Figura 3.6: Nivel educacional más alto alcanzado Censo

En conjunto con estos datos, otras variables relevantes asociadas a la capacidad económica de los hogares se encuentran presentes en ambas encuestas. Sin embargo, existen diferencias en la forma que están estructuradas las preguntas en algunos casos, así también las respuestas. En la figura 3.7 se ejemplifica con la pregunta en la encuesta del censo asociada al nivel del curso más alto aprobado, su equivalencia con la encuesta casen se presenta en la figura 3.8, donde se pregunta respecto del nivel educacional más alto alcanzado o nivel educacional actual. La pregunta puede ser directamente homologada, no obstante, las respuestas a pesar de tener un formato similar en orden ascendente de educación no son comparables por su discrepancia entre valores numéricos.

P15	Nivel del curso más alto aprobado	1	Sala cuna o jardín infantil	(1-14) 98 No aplica 99 Missing	Número entero
		2	Prekínder		
		3	Kínder		
		4	Especial o diferencial		
		5	Educación básica		
		6	Primaria o preparatorio (sistema antiguo)		
		7	Científico-humanista		
		8	Técnica profesional		
		9	Humanidades (sistema antiguo)		
		10	Técnica comercial, industrial/normalista (sistema antiguo)		
		11	Técnico superior (1-3 años)		
		12	Profesional (4 o más años)		
		13	Magíster		
		14	Doctorado		

Fuente: Manual de usuario Censo. Elaboración Instituto Nacional de Estadísticas

Figura 3.7: Pregunta de educación Censo

e6a	**e6a. ¿Cuál fue el nivel educacional más alto alcanzado o el nivel educacional actual?**	1 Nunca asistió
		2 Sala cuna
		3 Jardín Infantil (Medio menor y Medio mayor)
		4 Prekinder/Kinder (Transición menor y Transición Mayor)
		5 Educación Especial (Diferencial)
		6 Primaria o Preparatoria (Sistema antiguo)
		7 Educación Básica
		8 Humanidades (Sistema Antiguo)
		9 Educación Media Científico-Humanista
		10 Técnica, Comercial, Industrial o Normalista (Sistema Antiguo)
		11 Educación Media Técnica Profesional
		12 Técnico Nivel Superior Incompleto (Carreras 1 a 3 años)
		13 Técnico Nivel Superior Completo (Carreras 1 a 3 años)
		14 Profesional Incompleto (Carreras 4 o más años)
		15 Profesional Completo (Carreras 4 o más años)
		16 Postgrado Incompleto
		17 Postgrado Completo
		99 No sabe/no responde
		Total

Fuente: Libro de Códigos Casen. Elaboración Ministerio de Desarrollo Social

Figura 3.8: Pregunta de educación Casen

De esta manera, es necesario hacer ciertas modificaciones para que las variables sean equivalentes. Este proceso será detallado más adelante en la formulación del modelo de estimación.

3.4 Homologación de variables

Dado que ambas encuestan poseen estructuras y codificaciones distintas, fue necesario realizar un proceso de homologación entre variables comunes. Esto permitió entrenar modelos en Casen y aplicarlos sobre el censo.

La tabla 3.1 muestra las variables y su equivalencia entre encuestas que será utilizada para la estimación de los ingresos, identificando cada una con sus códigos de su respectivo libro manual.

Tabla 3.1: Equivalencia de variables entre encuestas Casen y Censo.

Variable	Sección	Casen	Censo
Sexo	Personas	sexo	p08
Edad	Personas	edad	p09
Nivel Educacional más alto alcanzado	Personas	e6a	p15
Tipo de vivienda	Viviendas	v1	p01
Tamaño del hogar	Viviendas	numper	cant_per
Numero de dormitorios	Viviendas	v27a	p04
Material predominante muro exterior	Viviendas	v2	p03a
Material predominante techo	Viviendas	v6	p03b
Material predominante piso	Viviendas	v4	p03c
Origen del agua	Viviendas	v20	P05

Se eliminaron registros con valores "No sabe/ No responde" (códigos 98, 99) y se transformaron variables cualitativas a formato categórico (*one-hot encoding*) para su uso en modelos. Además, se calcularon variables derivadas como escolaridad promedio del hogar y hacinamiento.

3.5 Modelo de estimación de ingresos

3.5.1 Especificación del modelo

Para abordar la ausencia de datos de ingresos en la base censal y la necesidad de contar con un indicador socioeconómico consistente, se propone un modelo de estimación del ingreso autónomo corregido del hogar (yautcorh) mediante técnicas de aprendizaje supervisado.

El modelo base corresponde a un algoritmo *Random Forest Regressor*, elegido por su capacidad de capturar relaciones no lineales entre variables y su robustez frente a valores atípicos (*outliers*) y multicolinealidad. La variable dependiente fue transformada mediante logaritmo natural para

estabilizar la varianza y mejorar el ajuste del modelo, el detalle del código del modelo se encuentra en el Anexo A

3.5.2 Entrenamiento del modelo

Se utilizo la base de datos Casen 2017 filtrada para la provincia de Concepción (n=12.943), con posterior limpieza y depuración, resultando en 9.610 registros completos. La variable *yautcorh* fue transformada mediante logaritmo y se eliminaron registros con ingreso nulo o negativo.

Para asegurar la robustez del modelo y prevenir problemas de sobreajuste, el modelo *Random Forest* fue entrenado utilizando validación cruzada y partición de entrenamiento/test (80/20). El preprocesamiento incluyo codificación *one-hot* de variables categóricas y estandarización de variables numéricas.

3.5.3 Aplicación al Censo

Una vez entrenado y validado el modelo de predicción de ingresos de la sección anterior, se procedió a la aplicación sobre los datos del Censo 2017. Esta etapa permitió extender la estimación del ingreso a una mayor cantidad de hogares del territorio, aprovechando la cobertura de esta encuesta.

3.5.3.1 Preparación base censal

La información censal utilizada provino de los microdatos de las bases de personas y viviendas, las cuales fueron integradas para generar un único registro por hogar mediante un identificador compuesto denominado hogar_id, construido a partir de las variables comuna, distrito (dc), zona/localidad (zc_loc), numero de vivienda (nviv) y numero de hogar (nhogar).

Las variables predictoras para la estimación como sexo, edad, nivel educativo, características de la vivienda y hacinamiento fueron derivadas o renombradas para ser compatibles con las utilizadas en el modelo entrenado. También se calcularon las variables derivadas personas_por_dormitorio y escolaridad_aprox.

3.5.3.2 Limpieza e imputación de datos.

Previo a la predicción, los datos fueron sometidos a una etapa de limpieza e imputación:

- Se eliminaron hogares con valores atípicos respecto del número de personas por hogar, en específico más de 20 personas por hogar, o viviendas con 0 dormitorios, ya que no son considerados como viviendas para este análisis.
- Se reemplazaron los códigos de No Sabe/No Responde (98/99) por valores faltantes (NaN)

- Las variables categóricas (v1, v2, v4, v6, v20, v27a) fueron imputadas por su moda.
- Las variables (personas_por_dormitorio y escolaridad_aprox) fueron imputadas, únicamente si presentaban valores faltantes, mediante un Imputador Iterativo con un *Random Forest Regressor* como estimador.

3.5.3.3 Aplicación del modelo

El modelo entrenado con Casen fue aplicado al conjunto de datos predictores del Censo preparado previamente. Dado que el modelo predice el logaritmo del ingreso, las predicciones fueron posteriormente transformadas exponencialmente para obtener las estimaciones en pesos chilenos del 2017. Estas estimaciones fueron incorporadas en un archivo que exporta el programa con la variable *ingreso_estimado*.

De manera preliminar, se logró estimar el ingreso de 341.697 hogares en la provincia de Concepción, cubriendo una proporción significativa de los hogares y garantizando valores coherentes y consistentes con las características socioeconómicas de la población.

3.6 Modelo de estimación de posesión de automóviles

En esta sección se presentan dos modelos econométricos desarrollados para estimar la posesión de automóviles en los hogares del Gran Concepción, a partir de los datos de la Encuesta Origen-Destino (EOD) 2015. Se empleó como enfoque principal un modelo de regresión Poisson, adecuado para modelar el conteo y capaz de capturar las diferencias en cantidad de vehículos entre los hogares en función de características socioeconómicas y demográficas.

De manera complementaria, se estimó un modelo de regresión logística con el objetivo de analizar la probabilidad de que un hogar posea al menos un automóvil. Este enfoque binario, si bien menos detallado y atractivo que el conteo, resulta útil para contrastar resultados y para interpretar de forma más sencilla la incidencia de las variables sobre la simple tenencia o no de vehículos. La combinación de ambos modelos permite abordar la problemática desde dos perspectivas complementarias y fortaleciendo la robustez del análisis.

3.6.1 Especificación de modelo de regresión Poisson

Para el modelo Poisson, la variable dependiente *NumeroVehiculos* se considera como el numero declarado de automóviles por hogar a partir de la EOD. Este modelo busca explicar la variación en el número de vehículos a partir de variables socioeconómicas, demográficas y territoriales del hogar. Se puede ver el detalle del código del modelo en el Anexo B

3.6.1.1 Variables predictoras

Las variables explicativas fueron incluidas considerando su relevancia teórica y empírica, además de su disponibilidad en la encuesta, con tal de explicar la tenencia vehicular, estas son:

- *TramoIngresoHogar*, clasificado en tres niveles (bajo, medio y alto), con el tramo más bajo como categoría de referencia. La especificación de los tramos es tal que: tramo bajo (0-$400.000), tramo medio ($400.000-$1.200.000) y tramo alto (+$1.200.000) (CLP). Estos tramos son los definidos por la propia EOD.
- *TamanoFamiliar* como el número total de integrantes del hogar (numérica)
- *Tipovivienda* con tres categorías principales (casa, departamento, otro)
- *DirecciónComuna* como la comuna de residencia del hogar con 11 categorías (10 comunas y una categoría base. Se excluye por limitaciones de la base de datos por no presencia las comunas de Florida y Santa Juana.)

3.6.2 Especificación de modelo de regresión logística

Para el modelo logístico, la variable dependiente *posee_auto* es binaria, que toma el valor 1 si el hogar reporta poseer al menos un vehículo, y 0 si no posee ninguno. Los datos de la encuesta muestran que el 56,8% de los hogares no posee vehículos, el 35,3% posee uno, el 6,8% posee dos y menos del 2% posee tres o más. Esto evidencia una distribución fuertemente concentrada en los hogares con 0 o 1 vehículo, lo que respalda indagar en el análisis de este modelo.

3.6.2.1 Variables predictoras

Las variables explicativas incluidas en los modelos fueron seleccionadas por su relevancia teórica y disponibilidad en la base de datos. Se consideraron las siguientes:

- Tramo de ingreso del hogar, clasificado en bajo, medio y alto.
- Tamaño del hogar agrupado en 3 categorías: 1-2, 3-4 y 5+
- Tipo de vivienda con tres categorías principales (casa, departamento, otro)
- Situación de la vivienda con 9 categorías (propia pagada, propia pagándose, arrendada, y de la 4 a la 9 son otras especificaciones)
- Acceso a internet en 3 categorías (Internet personal, Internet compartido, Sin internet)

Estas variables capturan distintas dimensiones socioeconómicas, demográficas y territoriales que influyen en la probabilidad de vehículo en los hogares, seleccionadas por su relevancia teórica y empírica en la literatura, como también por su disponibilidad en la base de datos.

3.6.2.2 Imputación del tramo de ingreso

En la base de datos, algunos hogares registran el tramo de ingreso del hogar como 9, lo cual entra en la categoría de dato faltante o no informado. Para imputar este valor se utilizó un modelo de clasificación *random forest* sobre los hogares en los tramos validos (bajo, medio, alto) utilizando las mismas variables explicativas. Este modelo permitió asignar a cada hogar con tramo 9 al tramo más probable en función de sus características en ambos modelos de estimación de automóviles.

3.6.2.3 Modelo de regresión logística

El modelo logístico fue estimado utilizando *statsmodels* sobre una muestra de hogares con información completa, el detalle del código de este modelo se encuentra en el Anexo D.

4.Resultados y discusión

En este capítulo se presentan y analizan los resultados a partir de los modelos estimados. Se reporta el desempeño del modelo de ingreso, se examina la distribución y coherencia territorial de las estimaciones, y se identifican las variables explicativas más relevantes. Posteriormente, se discuten los resultados de los modelos de posesión de automóviles —Poisson y Logístico—, evaluando su ajuste y capacidad predictiva, así como su aplicación sobre la estimación de ingresos previa. El análisis se complementa con comparaciones entre datos observados y estimados, a nivel comunal y por tramos de ingreso.

4.1 Resultados modelo estimación ingresos

En esta sección se presentan los resultados obtenidos a partir del modelo de estimación de ingresos aplicado a los hogares de la provincia de Concepción. El modelo se implementó mediante un *Random Forest Regressor*, ajustado a una muestra de los datos provenientes de la encuesta CASEN 2017, con la variable dependiente definida como el logaritmo del ingreso autónomo del hogar corregido.

4.1.1 Validación estadística

Dado que el modelo principal utilizado para la estimación de ingresos es un *Random Forest Regressor*, no se requiere verificar los supuestos clásicos de modelos lineales como homocedasticidad, normalidad de los residuos o independencia de errores. Este tipo de modelos no paramétrico está orientado a la predicción más que a la inferencia, y su desempeño se evalúa a través de métricas como el coeficiente de determinación y error cuadrático medio, además de técnicas de validación cruzada.

Por lo tanto, no se consideran necesarias estas pruebas estadísticas, ya que no son aplicables ni requeridas en el marco teórico del aprendizaje automático basado en arboles de decisión

4.1.2 Desempeño del modelo

El desempeño del modelo con una validación simple *hold-out* (una sola partición) mediante R cuadrado obtuvieron los siguientes resultados:

$$R^2\ de\ entrenamiento\ (80\%): 0{,}891$$

$$R^2\ de\ prueba\ (20\%): 0{,}278$$

El $R^2\ de\ entrenamiento$ explica aproximadamente el 89,1% de la variabilidad del ingreso sobre los datos de entrenamiento. Este valor elevado es consistente con la flexibilidad del modelo random forest, capaz de adaptarse bien a los patrones de muestra. Sin embargo, al evaluar sobre los datos de prueba, el R^2 disminuyo, reflejando su capacidad de generalización hacia datos no observados. Esta diferencia es esperable, dado que los ingresos presentan una alta heterogeneidad y factores no observables.

Para dar mayor robustez al nivel de explicación del modelo, se aplicó validación cruzada del tipo *k-folds* con 5 particiones (k=5). Esta metodología consiste en dividir aleatoriamente el conjunto de datos en cinco subconjuntos de igual tamaño, utilizando cuatro de ellos para el entrenamiento y uno para la validación en cada iteración.

Los resultados obtenidos fueron los siguientes:

$$Valores\ individuales\ por\ fold: [0{,}278\ 0{,}322\ 0{,}326\ 0{,}343\ 0{,}302]$$

$$R^2\ promedio: 0{,}314$$

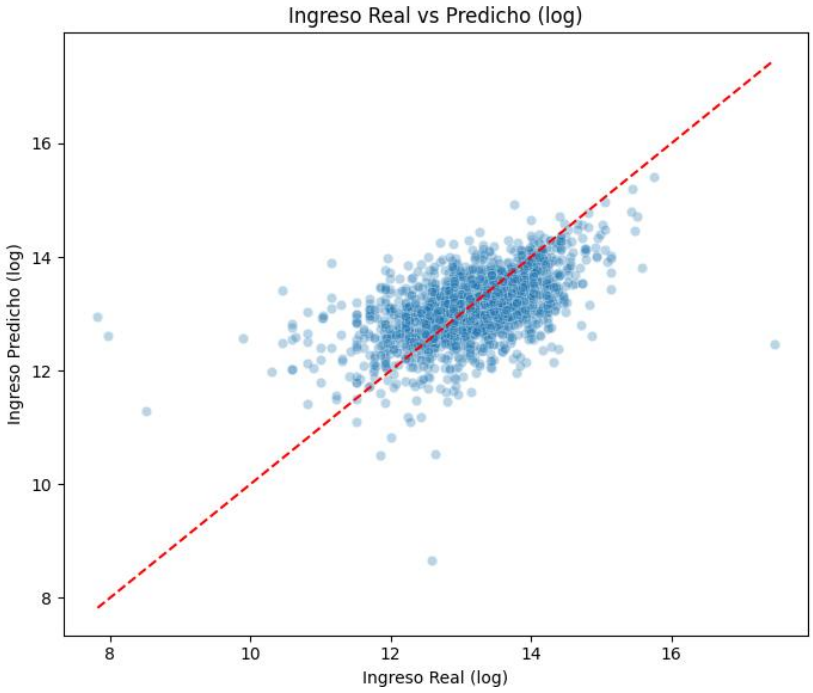
Lo anterior evidencia una capacidad predictiva consistente del modelo y respalda su validez para estimar el ingreso de los hogares. Esta aproximación permite modelar el ingreso con un enfoque orientado a la predicción, aceptando cierto margen de error, pero garantizando utilidad práctica para su uso posterior en el análisis de tenencia vehicular.

Además del coeficiente de determinación, se evaluó el desempeño del modelo mediante el Error de Raíz Cuadrada Media (RMSE), el cual mide el error absoluto promedio en la predicción del logaritmo del ingreso autónomo del hogar. Obteniendo lo siguiente:

$$RMSE\ (\log): 0{,}677$$

Este valor implica que, en promedio, las predicciones del modelo presentan un error relativo de un 97% respecto del ingreso real. Aunque este nivel de error puede parecer elevado, es consistente con la complejidad inherente a la estimación de ingresos mediante variables proxy censales, además, es importante notar que el RMSE en log-ingresos no se traduce linealmente en porcentajes de error absoluto. Por tanto, se considera que el modelo posee una capacidad predictiva razonable y adecuada para ser utilizado como insumo del análisis posterior.

El grafico de dispersión de ingresos reales vs. Predichos en escala logarítmica de la figura 4.1 muestra una clara tendencia positiva, con concentración en torno a la diagonal de igualdad, aunque con leve dispersión en los extremos. Esto sugiere que el modelo captura bien las tendencias generales, pero presenta menor precisión en valores muy bajos o altos de ingreso.

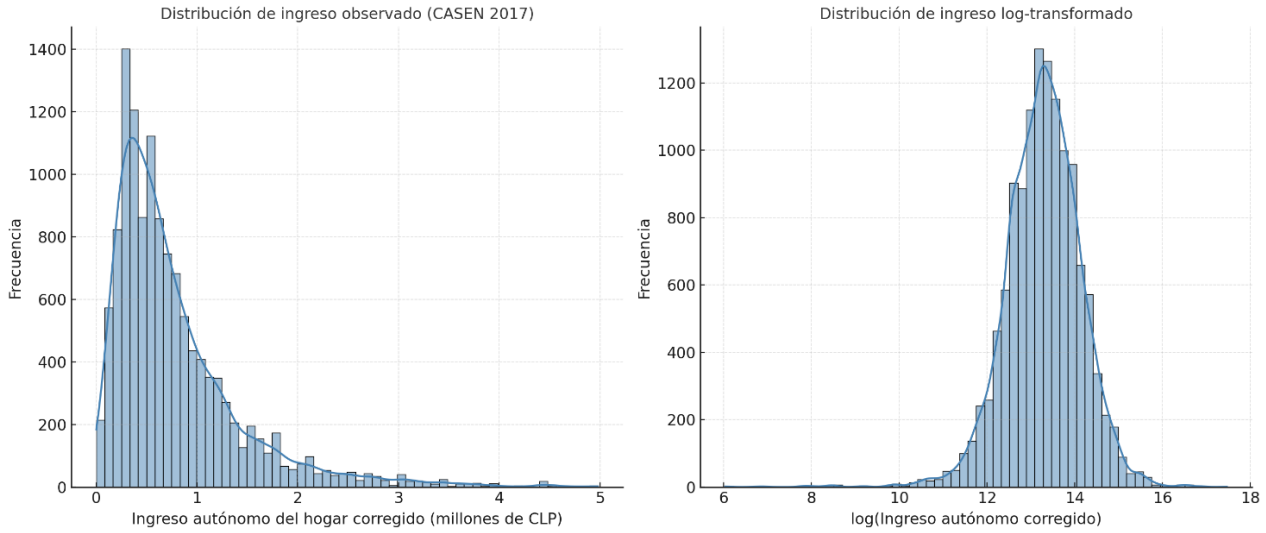


Fuente y elaboración: propia

Figura 4.1: Grafico de dispersión de ingresos reales y predichos.

4.1.3 Distribución del ingreso estimado

La transformación logarítmica aplicada a los ingresos originales corrigió la fuerte asimetría de la distribución, resultado en una forma aproximadamente normal, observable en la figura 4.2



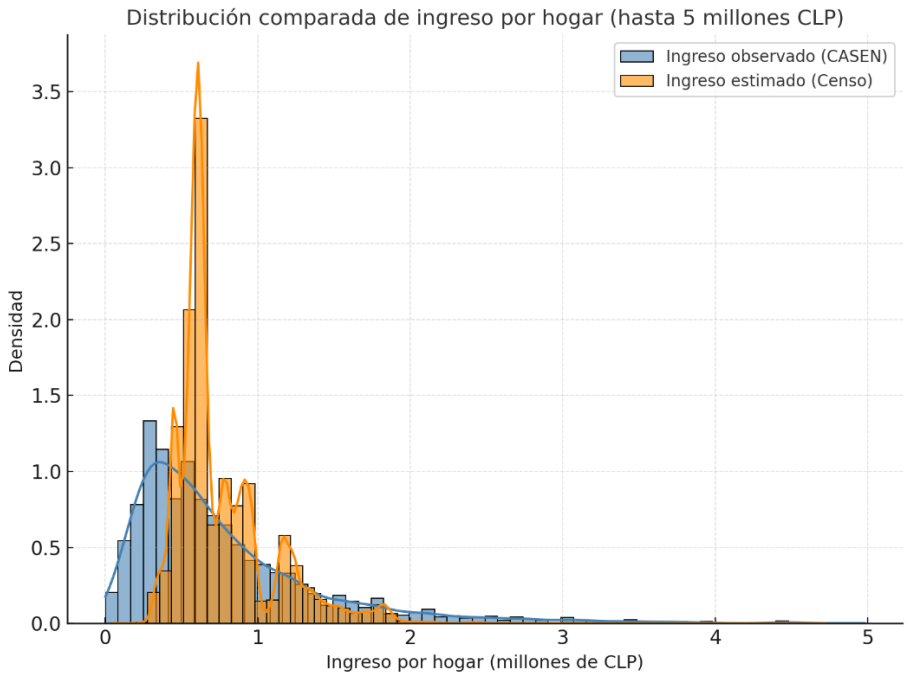
Fuente y elaboración: propia

Figura 4.2: distribución de ingreso original y su transformación logarítmica

Al aplicar el modelo al conjunto de hogares del censo, la distribución estimada del ingreso autónomo del hogar presenta una moda en torno a los 700.000-800.000 CLP y una cola alargada hacia los ingresos superiores observable en la figura 4.3 en color naranjo. En azul se observa la distribución de ingresos reales para comparar, el grafico se encuentra corregido utilizando densidades normalizadas debido a la diferencia de la frecuencia de datos. El ingreso estimado, desviación estándar y otras medidas de la estimación se detallan en la tabla 4.1

Tabla 4.1: Resumen estimación por hogar

Número de hogares estimado	341.697
Promedio	$773.571
Desviación Estándar	$330.317
Mínimo	$80.470
Máximo	$5.027.383



Fuente y elaboración: propia

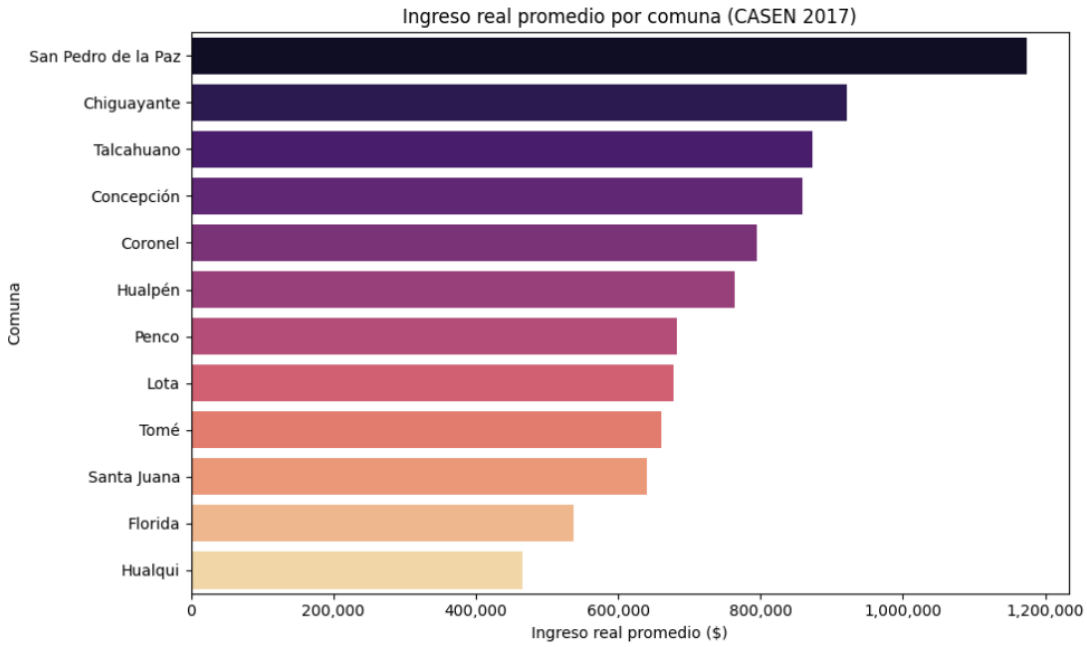
Figura 4.3: Distribución de ingresos observado y estimado por hogar

El análisis por deciles de ingreso estimado muestra una progresión consistente, desde un promedio de $415.327 en el primer decil, hasta $1.529.177 en el décimo, reflejando heterogeneidad socioeconómica de la población. (Anexo G)

Para evaluar la capacidad del modelo de estimación de ingresos para replicar patrones socioeconómicos, se realizó una comparación entre los ingresos promedios observados por comuna en la muestra CASEN y los ingresos promedios estimados para las mismas comunas. Esta comparación permite verificar si el modelo preserva las diferencias territoriales observadas empíricamente y, por tanto, si es adecuado para describir la heterogeneidad socioeconómica a nivel comunal.

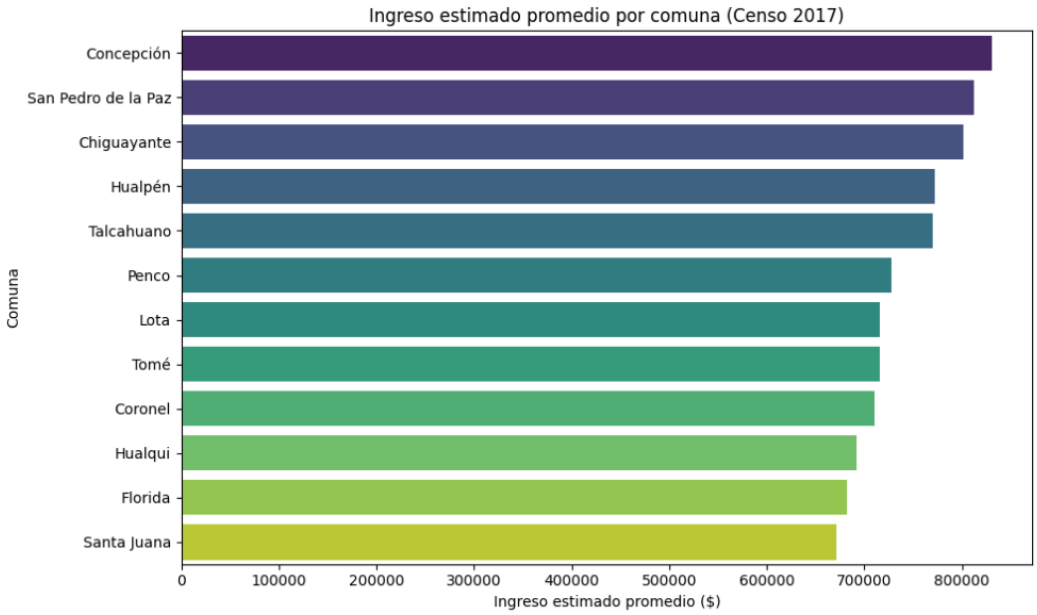
En la figura 4.4 se presentan los ingresos promedio por comuna observados ordenados de mayor a menor. En general se aprecia que las comunas de mayores ingresos en la muestra CASEN —como San Pedro de la Paz, Chiguayante y Talcahuano— también aparecen entre las de mayor ingreso en la estimación basada en el modelo en la figura 4.5. De igual manera, comunas como Hualqui y Florida se ubican consistentemente en los niveles inferiores. Esta consistencia sugiere que el modelo captura

adecuadamente la jerarquía relativa entre comunas, lo que es indicativo de una buena capacidad descriptiva a nivel agregado.



Fuente: base de datos Casen. Elaboración: propia

Figura 4.4: Ingreso observado Casen



Fuente y elaboración: propia

Figura 4.5: Ingreso estimado por comuna

No obstante, también se observan algunas discrepancias en la magnitud absoluta de los ingresos estimados, lo cual puede atribuirse a diferencias en las características de las poblaciones censadas y encuestadas, presencia de factores no incluidos en el modelo y a las limitaciones propias de las bases de datos. A pesar de ello, las diferencias relativas entre comunas se mantienen en gran medida, validado parcialmente el uso del modelo para análisis.

4.1.4 Importancia de variables explicativas

Las 5 variables más relevantes para el modelo se presentan en la tabla 4.2. En el Anexo H se detalla las otras variables con su importancia respectiva, pero que son menor al 3% individualmente.

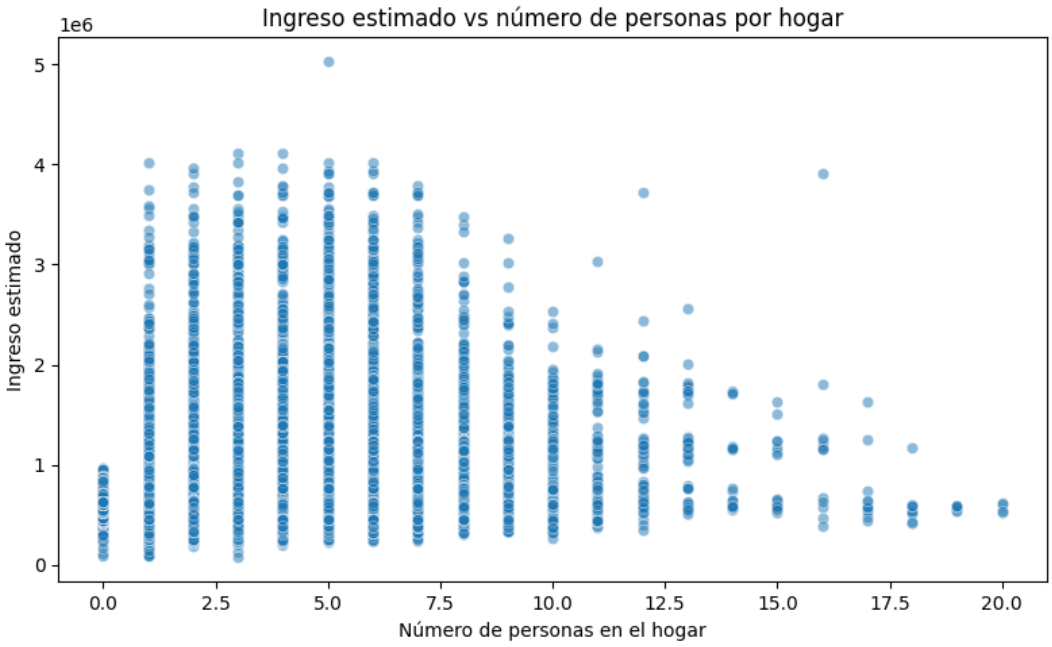
Tabla 4.2: Variables más importantes del modelo.

Variable	Importancia (%)
Edad	22,18%
Tamaño del hogar	17,00%
Número de dormitorios	7,00%
Personas por dormitorio	5,93%
Escolaridad aproximada	4,27%

Estas variables concuerdan con la literatura, ya que tanto la composición del hogar, educación y edad son determinantes conocidos del ingreso.

4.1.5 Relación entre ingreso estimado y otras variables

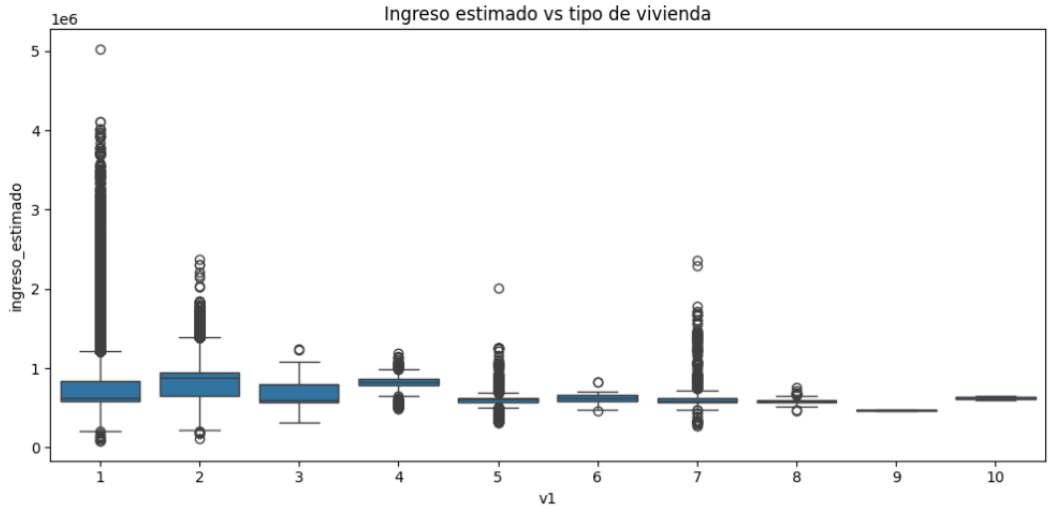
El análisis de la relación entre el ingreso y el tamaño del hogar muestra un patrón decreciente a partir de aproximadamente 4 personas por hogar en la figura 4.6. Esto refleja el efecto del mayor número de dependientes y la posible menor capacidad de ingreso per cápita en hogares grandes.



Fuente y elaboración: propia

Figura 4.6: Ingreso estimado vs número de personas por hogar.

Por otro lado, la relación entre ingreso estimado y tipo de vivienda en la figura 4.7 muestra mayores ingresos asociados a viviendas más consolidadas y formales, mientras que los ingresos más bajos se concentran en tipos de vivienda más precarios.



Fuente y elaboración: propia

Figura 4.7: Ingreso estimado vs tipo de vivienda (v1)

4.1.6. Síntesis de los resultados de estimación de ingresos

En síntesis, el modelo logro capturar las tendencias generales del ingreso de los hogares de la provincia, identificando variables explicativas coherentes con la teoría y arrojando una distribución del ingreso consistente con la estructura socioeconómica observada. Si bien la capacidad predictiva sobre los datos de prueba es moderada, las métricas y gráficos sugieren un modelo adecuado para estimar ingresos a nivel agregado.

4.2 Resultados modelo estimación posesión automóviles

4.2.1 Resultados modelo regresión Poisson

En esta sección se presentan y analizan los resultados obtenidos al estimar la posesión de automóviles en los hogares del Gran Concepción mediante un modelo de regresión Poisson. El análisis se estructura en cuatro partes: (1) desempeño del modelo y significancia de variables, (2) distribución de vehículos estimados, (3) relación entre vehículos y tramo de ingreso, y (4) distribución de vehículos estimados por comuna.

4.2.1.1 Validación del modelo de regresión Poisson

Antes de interpretar los resultados del modelo de regresión para la estimación de posesión de automóviles en los hogares del Gran Concepción, es fundamental verificar el cumplimiento de los supuestos básicos que sustentas este tipo de modelos. A continuación, se exponen los principales supuestos y la forma en las que fueron evaluados.

4.2.1.1.1 Variable dependiente como conteo no negativo

Este modelo requiere que la variable dependiente sea un conteo de eventos no negativo y discreto. En este caso, la variable analizada corresponde al número de automóviles en el hogar, la cual cumple con este criterio, con 0 como valor mínimo y 5 como máximo valor observado en la base de entrenamiento. No se identificaron valores negativos ni no enteros en la base de datos.

4.2.1.1.2 Equidispersión

Uno de los supuestos claves de este tipo de regresión, es que la media condicional de la variable dependiente sea igual a su varianza condicional. Para verificar este supuesto se calcularon la media y varianza muestral de la variable *NumeroVehiculos*, obteniéndose una media de 0.524 y una varianza de 0.466. La cercanía de estos valores respalda la ausencia de sobredispersión relevante en los datos.

Adicionalmente se calculó el índice de dispersión (ID), que se define como la razón entre la Desviación de Pearson y los grados de libertad residuales del modelo, dado por la expresión:

$$ID = \frac{\chi^2_{Pearson}}{gl} = \frac{5450}{6909} \approx 0{,}788 \quad (4)$$

El valor cercano a 1 indica que la varianza observada no excede significativamente a la media condicional y que no existe evidencia sustancial de sobredispersión. Por tanto, el supuesto de equidispersión se considera cumplido, y no es necesario recurrir a otros modelos alternativos.

4.2.1.1.3 Independencia de las observaciones

Para este estudio, cada registro corresponde a un hogar distinto identificado de manera única en la base de datos con un número de folio, sin duplicaciones ni agrupamientos. En consecuencia, la independencia entre observaciones es adecuada.

4.2.1.1.4 Correcta especificación del modelo

La correcta especificación del modelo se evaluó mediante el análisis de los signos y significancia de los coeficientes estimados, los cuales resultaron consistentes con la teoría económica y social. Un mayor ingreso del hogar se asocia con posesión mayor de vehículos. Asimismo, las predicciones obtenidas aumentan en función del tramo de ingreso y presentan diferencias esperadas por comuna y tipo de vivienda. Gráficos de residuos y de predicción versus valores reales observados no mostraron patrones sistemáticos que sugieran mala especificación. Por tanto, el supuesto de linealidad en el logaritmo de la media condicional se considera cumplido. Mas detalles del análisis de coeficientes y gráficos observados vs estimados serán revisados en la sección de resultados de este modelo.

4.2.1.1.5 Colinealidad

La presencia de alta colinealidad entre las variables independientes puede distorsionar la estabilidad de las estimaciones. Para evaluar este aspecto, se calcularon los factores de inflación de la varianza (VIF) de los predictores. Obteniéndose valores máximos de 1.563453 (Anexo C), muy por debajo del umbral critico de 5. Por tanto, no se evidenció colinealidad preocupante entre las variables explicativas incluidas en el modelo.

4.2.1.1.6 Presencia de ceros

Finalmente, para evaluar el supuesto de presencia excesiva de ceros, se calculó la proporción de hogares con cero vehículos en los datos observados. En la base de datos, un 56,83% de los hogares (3935 de 6924) no poseía vehículos. Lo cual es considerado razonable y compatible con la distribución

Poisson. Los datos no presentan una concentración anómala de ceros que justifique el uso de otros modelos adaptados para esas condiciones como un modelo inflado en ceros (ZIP).

4.2.1.2 Desempeño del modelo y significancia de las variables

El modelo de Poisson estimado, con 6.924 observaciones, muestra un $Pseudo - R^2$(Cox-Snell) de 0,1349, lo que indicaría que las variables incluidas explican aproximadamente un 13,49% de la variabilidad en la cantidad de vehículos por hogar. Esto es consistente con la naturaleza discreta y dispersa de la variable dependiente, donde se espera una capacidad explicativa moderada.

En cuanto a los coeficientes estimados, estos se detallan en la tabla 4.3.

Tabla 4.3: Coeficientes modelo Poisson

Variable	Coef.	Std. err.	z	P>|z|	[0.025	0.975]
Intercept	-1,2436	0,056	-22,323	0,000	-1,353	-1,134
Tramo de ingreso 2	0,4848	0,038	12,735	0,000	0,410	0,559
Tramo de ingreso 3	1,0407	0,050	20,913	0,000	0,943	1,138
Comuna [CORONEL]	-0,3200	0,066	-4,856	0,000	-0,449	-0,191
Comuna [CHIGUAYANTE]	-0,1410	0,067	-2,099	0,036	-0,273	-0,009
Comuna [HUALQUI]	-0,1520	0,130	-1,169	0,242	-0,407	0,103
Comuna [LOTA]	-0,7806	0,100	-7,819	0,000	-0,976	-0,585
Comuna [PENCO]	-0,4829	0,104	-4,628	0,000	-0,687	-0,278
Comuna [SAN PEDRO DE LA PAZ]	0,2027	0,052	3,891	0,000	0,101	0,305
Comuna [TALCAHUANO]	-0,0223	0,054	-0,413	0,680	-0,128	0,084
Comuna [TOME]	-0,4231	0,094	-4,509	0,000	-0,607	-0,239
Comuna [HUALPEN]	-0,1154	0,063	-1,838	0,066	-0,238	0,008
Vivienda [Departamento]	-0,3603	0,068	-5,263	0,000	-0,495	-0,226
Vivienda [Otro]	-0,3385	0,501	-0,676	0,499	-1,321	-0,644
Tamaño Familiar	0,1175	0,012	10,119	0,000	0,095	0,140

Los efectos de las variables predictoras para cada categoría se describen a continuación:

- Tramo de ingreso del hogar

  Hogares del tramo 2 (ingreso medio) presentan un numero esperado de vehículos aproximadamente un 62% mayor que los hogares de referencia (tramo 1, ingresos más bajos) ($e^{0,4848} \approx 1,62$). Y hogares del tramo 3 (ingreso alto) tienen un numero esperado más del 180% mayor que los hogares del tramo 1 ($e^{1,0407} \approx 2,83$). Ambos significativos estadísticamente

- Comunas de referencia

  Estas comunas se comparan con la categoría base elegida como Concepción, permitiendo ver como las comunas se comportan en base a su diferencia con esta base de referencia.

  San Pedro de la Paz, con un coeficiente de 0,2027 tiene un aumento significativo de aproximadamente 22% respecto de Concepción. Y por su contraparte, Lota, con un coeficiente de -0,7806 tiene una expectativa de aproximadamente 54% menor que Concepción. Ambos significativos.

  Hualpén, Hualqui y Talcahuano resultaron con diferencias no significativas respecto a la base.

- Tipo de vivienda

  Con respecto a la categoría Casa, las viviendas categoría Departamentos tienen un 30% menos vehículos que las casas. La categoría Otros resulto no significativos.

- Tamaño del hogar

  Con un coeficiente de 0,1175, cada persona adicional en el hogar se asocia con un incremento esperado del 12% en la cantidad de vehículos, altamente significativo.

4.2.1.3 Aplicación del modelo sobre archivo de estimación de ingresos

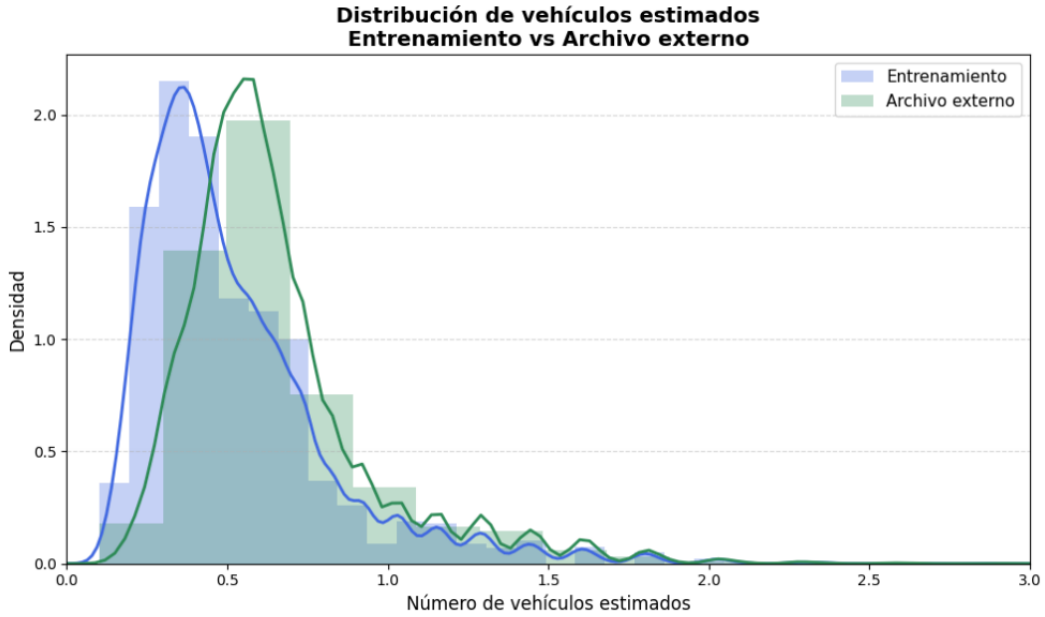
Una vez entrenado y validado el modelo de regresión Poisson para estimar el número de vehículos por hogar, se procedió a aplicar dicho modelo sobre el archivo que exporta la estimación de ingresos de la primera parte. En donde a partir de algunas de las columnas que exporta, como *hogar_id*, para identificar las comunas, *cant_per* para el tamaño del hogar, *v1*, para el tipo de vivienda, e ingreso estimado para clasificarlo dentro de los tramos de ingreso. Se homologó de manera correcta permitiendo una aplicación adecuada del modelo propuesto.

Una consideración importante es que, al comparar la distribución de tramos de ingresos entre la EOD y el Censo estimado, se observa una concentración marcada en el tramo medio (T2: 85,3%) en el Censo, en contraste con la distribución más equilibrada de la EOD (T1: 51,7%, T2: 37,5% y T3: 7,9%).

Esta diferencia refleja una reducción en la variabilidad del ingreso estimado, lo que disminuye la capacidad del modelo para discriminar entre niveles socioeconómicos en la predicción del número de vehículos a partir de la estimación de ingresos, lo que implica que se pueden atenuar las diferencias entre niveles socioeconómicos en las predicciones de tenencia vehicular.

4.2.1.3.1 Distribución de vehículos estimados

El primer análisis compara la distribución de los vehículos observados en la base de datos de la EOD, y el archivo externo de la estimación de ingresos aplicada la estimación de vehículos. La figura 4.8 muestra la densidad y comportamiento de ambas distribuciones:



Fuente entrenamiento: EOD. Elaboración: propia

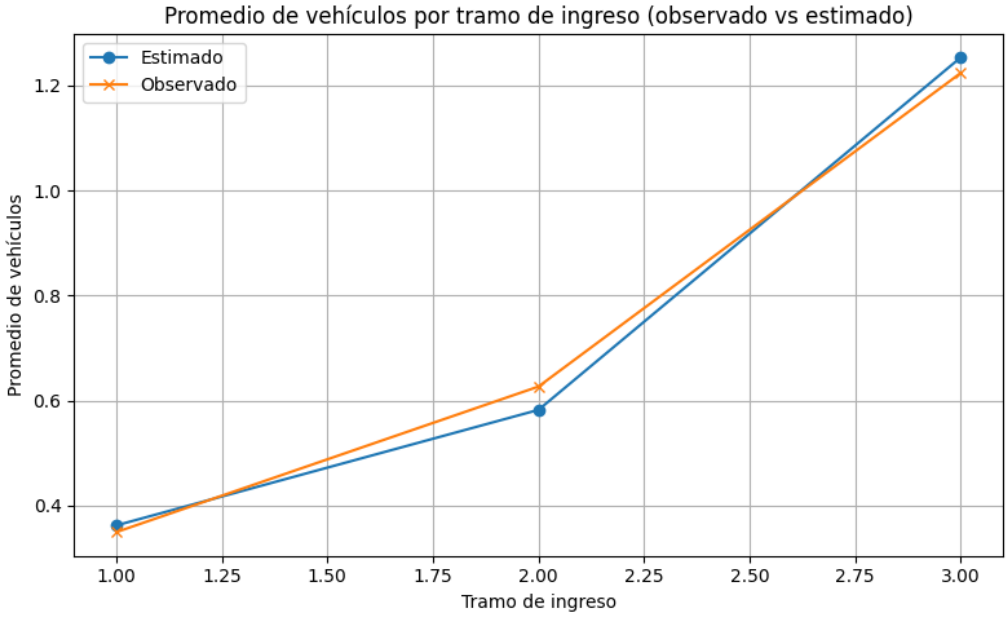
Figura 4.8: Distribución de vehículos observado y estimado

Se observa que ambas distribuciones son similares en forma, con una moda cercana a 0,4 vehículos por hogar. En el conjunto de entrenamiento, la densidad máxima ocurre alrededor de 0,35-0,4 vehículos, mientras que en el archivo externo se desplaza levemente hacia valores más altos, alcanzando su máximo alrededor de 0,5. Esta ligera diferencia puede explicarse por las características de los ingresos estimados de los hogares censales, que presentan en promedio, ingresos ligeramente superiores a los observados en la muestra de entrenamiento. En ambos casos, la distribución presenta asimetría positiva, con colas hacia valores más altos, consistentes con el hecho de que gran porcentaje de los hogares pose uno o ningún vehículo.

El modelo Poisson predice la media esperada de un conteo, que en la practica puede tomar valores no enteros como los mencionados. Para efectos de visualización gráfica y análisis agregado se utiliza el valor continúo estimado. En caso de ser necesario para distintas aplicaciones como simulaciones, estos valores pueden trabajarse mediante redondeo o muestrearse desde una distribución Poisson con media igual al valor estimado para obtener y utilizar valores enteros.

4.2.1.3.2 Promedio de vehículos por tramo de ingreso

En segundo lugar, se examina el promedio de vehículos estimados por tramo de ingreso del hogar, considerando los valores por tramo observados como base. La comparación grafica se observa en la figura 4.9 y el detalle número se especifica en la tabla 4.4.



Fuente observado: EOD. Elaboración: propia

Figura 4.9: Promedio de vehículos por tramo de ingreso observado y estimado

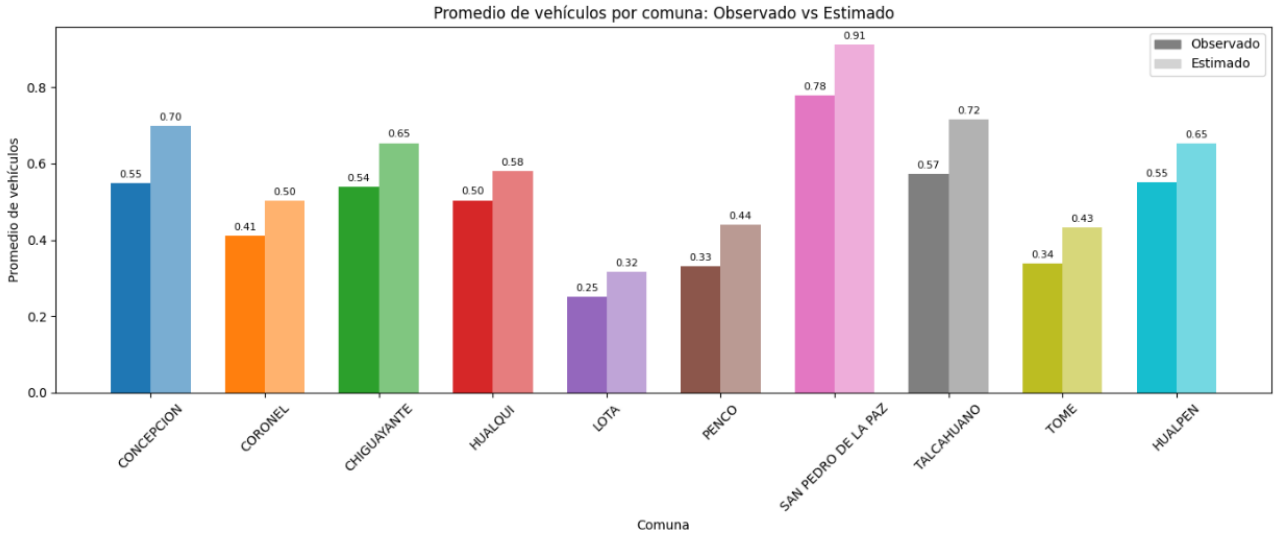
Tabla 4.4: Promedio de vehículos por tramo

Tramo	Vehículos observados (EOD)	Vehículos estimados
1	0.349253	0.362034
2	0.626570	0.582176
3	1.224044	1.253673

El modelo captura correctamente la tendencia creciente del número de vehículos conforme aumenta el tramo de ingreso del hogar, con una diferencia mínima en cada tramo.

4.2.1.3.3 Promedio de vehículos por comuna

Finalmente, se examinan las diferencias entre las predicciones y las observaciones reales a nivel comunal. La siguiente figura 4.10 muestra las barras correspondientes al promedio de vehículos por comuna, diferenciando observados con un color más oscuro, y estimados con un color más claro para cada una:



Fuente y elaboración: propia

Figura 4.10: Promedio de vehículo por comuna

Las predicciones del modelo respetan las diferencias estructurales entre comunas, destacando a San Pedro de la Paz y Talcahuano con una mayor dotación promedio de vehículos, tanto en los datos observados como en las estimaciones. Y en comunas como Lota o Tomé, los promedios son más bajos, reflejando contextos socioeconómicos con menor capacidad de adquisición. Aunque el modelo tiende a sobreestimar ligeramente en todas las comunas, mantiene correctamente la jerarquía relativa entre ellas, siendo especialmente preciso en comunas con menor dispersión de datos.

4.2.2 Resultados modelo regresión logística

El modelo logístico estima los efectos de las características de las variables predictoras sobre la probabilidad de poseer al menos un automóvil. Los coeficientes del modelo representan cambios en el *log-odds* relativos a la categoría de referencia, por lo que se interpretan mediante sus

exponenciales (e^{β}), conocidos como *odds ratio*. Estos indican cuantas veces más (o menos) probable es que un hogar en esa categoría posea un automóvil en comparación con la categoría base.

4.2.2.1 Validación del modelo de regresión logística

El ajuste del modelo fue significativo sobre 6.924 observaciones, a partir de una convergencia alcanzada de 6 iteraciones, considerando 16 parámetros y utilizando máxima verosimilitud (MLE). Se obtuvo un R^2 de 0,143. Este valor indica que el modelo explica aproximadamente un 14,2% de la variabilidad en la variable dependiente. Se realizaron pruebas para evaluar la adecuación y robustez del modelo, cumpliendo los supuestos básicos que deben cumplirse para este tipo de regresión. Independencia de los errores, linealidad en el logit para las variables continuas, ausencia de multicolinealidad y ausencia de valores atípicos con gran influencia.

4.2.2.1.1 Independencia de los errores

Al tratarse de datos que provienen de una encuesta transversal, registrados solo una vez, sin panel ni repetición, no hay razones para sospechar dependencia entre las observaciones.

4.2.2.1.2 Linealidad en el logit para variables continuas

Todas las variables fueron consideradas como variables categóricas, por lo que no existe análisis de linealidad para variables continuas

4.2.2.1.3 Multicolinealidad

Se evaluaron los factores de inflación de la varianza (VIF) para las variables independientes, encontrándose todas por debajo de 2.4 (Anexo E), lo que indicaría que no existe multicolinealidad significativa.

4.2.2.1.4 Valores típicos influyentes

Los valores de Leverage y la distancia de Cook fueron analizados para identificar posibles observaciones con influencia excesiva sobre el modelo. El Leverage máximo observado fue 0,1097, mayor que el umbral sugerido (0,0052), lo que indicaría la presencia de algunas observaciones con valores relativamente altos. Sin embargo, la distancia de Cook máxima fue apenas 0,0260, por debajo del valor de referencia de 1. Por lo tanto, no se detectaron observaciones con influencia excesiva que comprometieran la validez de los resultados, en el anexo F se detalla el grafico analizado.

4.2.2.2 Resumen del modelo y variables

El resumen desplegado por el programa se presenta en la tabla 4.5 a continuación, en conjunto con el análisis e interpretación de las variables predictoras sobre el logit de la probabilidad.

Tabla 4.5: Coeficientes modelo logístico

	Coef	Std.Err.	z	P > | z |	[0.025	0.975]
Intercept	-0,0092	0,1115	-0,0828	0,9340	-0,2279	0,2094
Tramo de ingreso 2	0,7772	0,0565	13,7527	0,0000	0,6665	0,8880
Tramo de ingreso 3	1,8707	0,1273	14,6962	0,0000	1,6212	2,1202
Tamaño de hogar (3–4)	0,6386	0,0620	10,2924	0,0000	0,5214	0,7602
Tamaño de hogar (5+)	0,6798	0,0808	8,4121	0,0000	0,5214	0,8382
Tipo de vivienda 2	0,6439	0,2962	2,1741	0,0297	0,0634	1,2244
Tipo de vivienda 3	-0,6210	0,1026	-6,0514	0,0000	-0,8221	-0,4198
Tipo de vivienda 4	0,0172	0,6089	0,0282	0,9775	-1,1762	1,2106
Situación vivienda 2	0,6815	0,0946	7,2036	0,0000	0,4961	0,8670
Situación vivienda 3	-0,0074	0,0837	-0,0883	0,9296	-0,1715	0,1567
Situación vivienda 4	0,2973	0,3192	0,9314	0,3517	-0,3283	0,9228
Situación vivienda 5	-0,3562	0,1288	-2,7663	0,0057	-0,6086	-0,1038
Situación vivienda 6	-0,2081	0,4413	-0,4716	0,6372	-1,0730	0,6567
Situación vivienda 7	-0,1094	0,4608	-0,2375	0,8123	-1,0125	0,7937
Situación vivienda 8	-0,8978	0,3942	-2,2777	0,0227	-1,6703	-0,1252
Internet compartido	-0,5126	0,1231	-4,1630	0,0000	-0,7540	-0,2713
Sin internet	-1,3806	0,1052	-13,1297	0,0000	-1,5867	-1,1745

Los efectos de las variables predictoras para cada categoría se describen a continuación.

- Tramo de ingreso

  En comparación con los hogares de ingreso bajo (T.1), los hogares de ingreso medio (T.2) y alto (T.3) presentan mayores probabilidades de poseer automóviles, con coeficientes de 0.7772 y 1.8307 significativos respectivamente. Lo que representa aproximadamente 2.18 veces ($e^{0,7772} \approx 2,18$) mayores probabilidades para el tramo medio y 6.49 veces ($e^{1,8307} \approx 6,49$)

mayores probabilidades de poseer un automóvil. Esto refleja la fuerte influencia del ingreso en la capacidad de adquisición de un automóvil.

- Tamaño del hogar

  En comparación con los hogares de 1-2 personas, los hogares de 3-4 personas tienen aproximadamente 1.89 veces mayores probabilidades de poseer automóvil y de 5 o más la probabilidad es de 1.97 veces mayor.

  Esto indica que los hogares más numerosos tienen una mayor necesidad para disponer de un automóvil, sin embargo, el cambio no es tan considerable como lo es el ingreso.

- Tipo de vivienda

  La categoría de tipo 2 (Casa en condominio) de tipo de vivienda respecto de la categoría base muestra un aumento de aproximadamente 1.9 veces. La categoría tipo 3 (Departamento) en cambio, las probabilidades son menores, siendo 0.54 veces las probabilidades de la categoría base.

- Situación de vivienda

  Con relación a la situación 1, la situación 2 (Casa pagándose) aumenta las probabilidades en 1.98 veces. Las situaciones 5 y 8 (Otras) en este caso las reducen a 0.7 y 0.64 veces las probabilidades respecto de la base.

- Acceso a internet

  En comparación con el nivel 1, el nivel 2 (Internet compartido) reduce las probabilidades en 0,6 veces y el nivel 3 (Sin internet) aún más, con aproximadamente 0,25 veces. Esto puede reflejar características socioeconómicas o de ubicación que estén correlacionadas con un menor acceso a internet se traduce en menor probabilidad de poseer un automóvil.

Finalmente, el intercepto no resulto significativo, lo que refleja que las categorías base por sí sola no explican diferencias significativas.

4.2.2.3 Desempeño del modelo

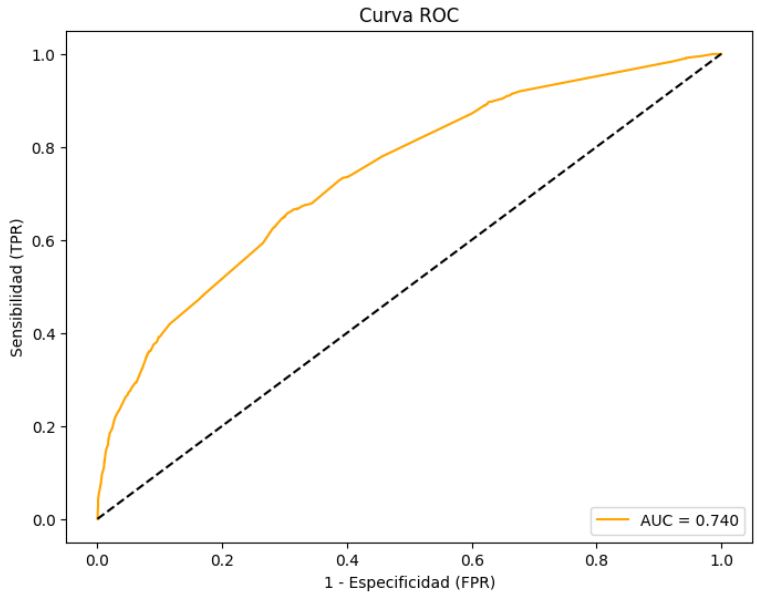
El desempeño del modelo se evaluó a través de varias métricas y gráficos estándar en modelos de regresión logística, con el objetivo de medir su capacidad predictiva y ajuste.

El modelo fue ajustado por máxima verosimilitud y presento un $pseudo - R^2$ de McFadden de 0,143. Este mide la proporción de mejora del *log-likelihood* del modelo respecto a un modelo sin predictores. Si bien no es directamente comparable con el R^2 de un modelo de regresion lineal, valores entre 0,2

y 0,4 se consideran aceptables en contextos sociales; aquí, un valor de 0,143 indica un ajuste razonable, considerando la complejidad del fenómeno y limitante de la base de datos.

4.2.2.4 Capacidad de discriminación

Para evaluar que tan bien el modelo distingue entre hogares que poseen y no poseen automóvil, se calculó el área bajo la curva ROC (AUC), que fue de 0,740 como se observa en la figura 4.11.



Elaboración y fuente: propia

Figura 4.11: Curva ROC

Esta curva traza la sensibilidad contra 1-especificidad (FPR) para todos los posibles umbrales de decisión. La línea punteada diagonal (valor 0,5) indica que el valor no discrimina mejor que el azar, y un valor de 1 indicaría una discriminación perfecta. En este caso, un AUC de 0,740 sugiere que el modelo tiene una capacidad discriminativa aceptable para fenómenos sociales.

De manera complementaria, a través de una matriz de confusión desplegada en la tabla 4.6 se observa el comportamiento del modelo a partir de los datos de la encuesta.

Tabla 4.6: Tabla con datos de matriz de confusión

	Estimación: NO	Estimación: SI
Observado: NO	2842	1093
Observado: SI	1143	1846

Es posible medir exactitud, precisión y sensibilidad a partir de esta matriz.

- Exactitud: proporción de predicciones correctas sobre el total.

$$Exactitud = \frac{Correctas(no) + Correctas(si)}{Total} \approx 67{,}7\% \quad (7)$$

- Precisión: proporción de predicciones positivas que fueron correctas.

$$Precision = \frac{Correctas(si)}{Correctas(si) + Incorrectas(no)} \approx 62{,}8\% \quad (8)$$

- Sensibilidad: Proporción de casos reales positivos que fueron correctamente predichos.

$$Sensibilidad = \frac{Correctas(si)}{Correctas(si) + Incorrectas(si)} \approx 61{,}8\% \quad (9)$$

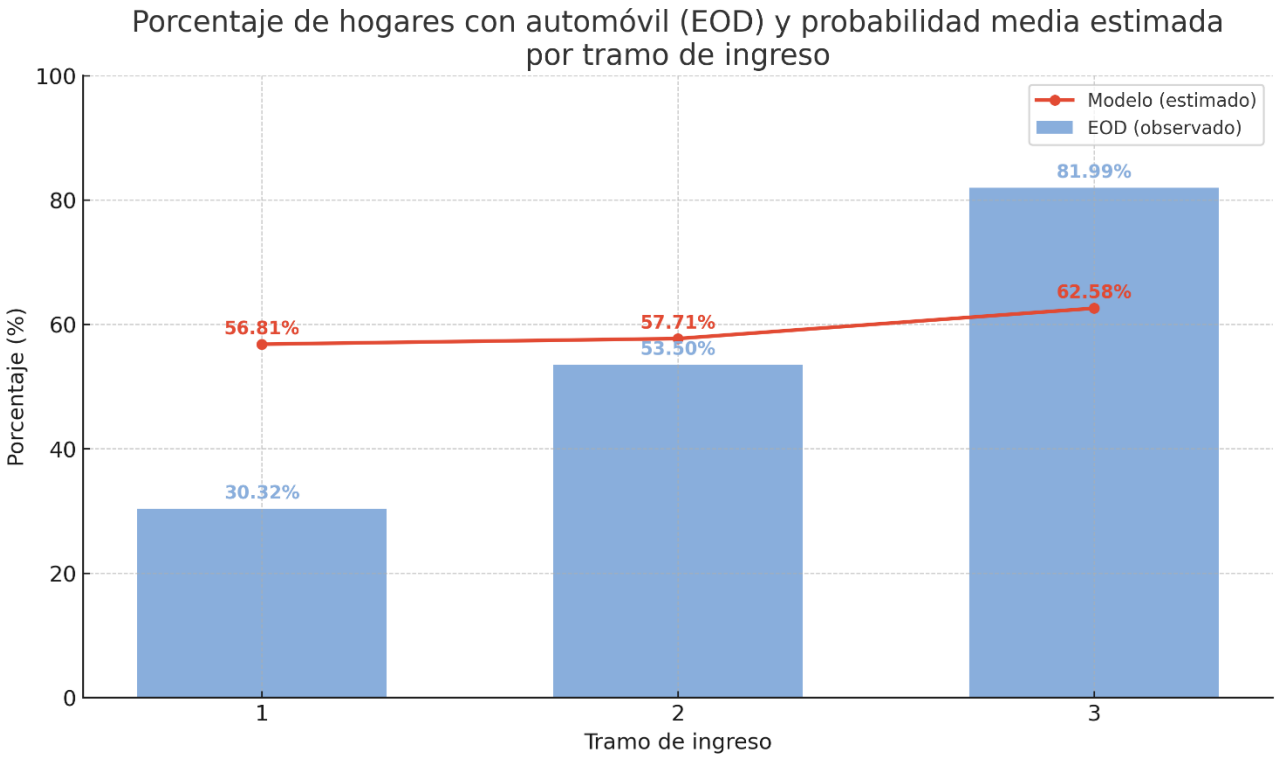
Estos valores muestran que el modelo tiene un balance adecuado entre identificar correctamente los hogares con automóvil (sensibilidad) y no sobre predecir (precisión).

4.2.3 Aplicación del modelo sobre archivo de estimación de ingresos

El modelo entrenado previamente con la encuesta origen destino, con las variables *TramoIngresoHogar, Tamaño_Hogar_cat, TipoVivienda, SituacionVivienda y CuentasInternet* fue aplicado al archivo que exporta el modelo anterior de estimación de ingresos. Para mantener coherencia metodológica, los tramos de ingreso fueron definidos utilizando el mismo formato que la EOD original. (Tramo 1: entre 0 y 400.000, tramo 2: 400.000 a 1.200.000 y tramo 3 sobre 1.200.000).

Dado que el archivo no incluía información homologable directamente salvo el tamaño de hogar con cantidad de personas y los ingresos estimados que se ajustan a los tramos, el resto de las variables fueron fijadas en sus categorías base para todos los hogares, para observar el comportamiento del modelo. Considerando teóricamente que el ingreso y el tamaño del hogar fueran capaz de influir de gran manera en las probabilidades de poseer un auto, manteniendo las otras variables como constantes.

En la figura 4.12, se observa el porcentaje de hogares que poseen un automóvil por tramo de ingreso de la base de datos de la encuesta original como barras, y se observa el comportamiento para la estimación por tramo en el trazo, con sus respectivos porcentajes.



Fuente: Base de datos EOD. Elaboración: propia

Figura 4.12: Porcentaje real vs probabilidad media estimada por tramo

La probabilidad media se concentró entre 0,57 y 0,63 para los tres tramos, con una leve tendencia creciente al pasar del tramo 2 al 3. Sin embargo, las diferencias entre tramos fueron menos marcadas que las observadas en la EOD, lo que se explica principalmente por las siguientes razones:

- El 85% de los hogares se encuentran dentro del tramo 2 (entre 400.000 y 1.200.000), lo que genera escasa representación en los extremos.
- Los ingresos estimados presentaron baja variabilidad y estuvieron concentrados en rango medio, limitando la capacidad del modelo para discriminar entre tramos.
- La ausencia de otras variables predictoras significativas del modelo original redujo la capacidad explicativa del modelo.

En conjunto, estos resultados muestran que, aunque la aplicación del modelo completo sobre los datos externos respeta la tendencia general prevista, donde mayor ingreso este asociado a mayor

probabilidad de poseer automóvil, las diferencias resultaron atenuadas por las limitaciones de los datos disponibles en la encuesta.

Tabla 4.7: Proporción media estimada de hogares con automóvil por tramo de ingreso

Tramo	Probabilidad media	Numero Hogares
1	56,81%	118.642
2	57,71%	165.052
3	62,58%	57.606

5. Conclusiones

Este estudio logró estimar con éxito el ingreso autónomo de los hogares y satisfactoriamente la posesión de automóviles en la provincia de Concepción a partir de datos censales y encuestas representativas, combinados con modelos de aprendizaje estadístico. Los principales hallazgos permiten concluir que la metodología planteada —que integra fuentes de datos heterogéneas y técnicas de regresión avanzadas— constituye una aproximación válida, robusta y replicable para caracterizar desigualdades socioeconómicas a nivel territorial.

En el caso del ingreso, el modelo *Random Forest* demostró alta capacidad explicativa en los datos de entrenamiento y un desempeño satisfactorio en prueba, considerando la complejidad y dispersión de la variable ingreso. Asimismo, las estimaciones respetan la jerarquía comunal observada, evidenciado mayores ingresos en comunas urbanas consolidadas y menores en zonas rurales, confirmando la coherencia espacial de las predicciones. La identificación de variables clave como la edad, tamaño del hogar y nivel educacional refuerza su pertinencia en el contexto socioeconómico local.

En cuanto a la posesión de automóviles, la combinación de modelos de regresión logística y Poisson permitió identificar factores significativos asociados al acceso y cantidad de vehículos. El modelo logístico fue útil para analizar la probabilidad de tener al menos un vehículo, mientras que el modelo Poisson permitió capturar de forma más precisa la distribución completa de números de vehículos por hogar. Ambos modelos arrojaron resultados coherentes con las tendencias observadas en la literatura y realidad local, reforzando la validez de la aproximación metodológica.

Estos resultados confirman la relación positiva entre nivel de ingreso y posesión de automóviles, en línea con la evidencia empírica revisada para otras ciudades latinoamericanas y chilenas (CEPAL, 2018; Tovar & Rodríguez, 2020). Al igual que en Santiago, Valparaíso y Temuco, en el Gran Concepción los hogares de mayores ingresos presentan tasas de motorización más altas, y las comunas con menor dotación vehicular corresponden a territorios de menores ingresos y/o peor conectividad. Sin embargo, a diferencia de estudios en grandes áreas metropolitanas, aquí la variabilidad entre comunas es menos pronunciada, lo que puede atribuirse tanto a la menor dispersión territorial de ingresos estimados como a la estructura intermedia del sistema urbano local. De manera similar, este patrón coincide con lo descrito por Arellana et al (2021) para ciudades intermedias, donde las diferencias de acceso al automóvil existen, pero son moduladas por el contexto urbano y la cobertura del transporte público.

Si bien el desempeño de los datos aplicados fue moderado, las métricas obtenidas resultan coherentes con fenómenos socioeconómicos de alta variabilidad, considerando las restricciones inherentes a la disponibilidad de información. Entre las principales limitaciones se encuentra la baja variabilidad del ingreso estimado en el Censo, lo que provocó una concentración de hogares en el tramo medio y redujo la sensibilidad de los modelos de la aplicación posterior para diferenciar patrones de motorización. Además, el análisis evidenció que parte importante de estas limitaciones se origina en las propias encuestas utilizadas como insumo: su diseño, cobertura y forma de recolección de datos condicionan los resultados y plantean desafíos estructurales para la formulación de modelos.

Una línea de trabajo futuro que permitiría mejorar la precisión y robustez de los modelos de estimación de tenencia vehicular es la incorporación de variables territoriales a nivel comunal que capturen las condiciones físicas, urbanas y de infraestructura que influyen en los patrones de motorización. Para ello, es posible complementar la información con bases de datos secundarias provenientes de instituciones públicas. El Instituto Nacional de Estadísticas, por ejemplo, entrega información sobre densidad poblacional, superficie comunal y tasas de urbanización; mientras que el Ministerio de Transportes y Telecomunicaciones dispone de métricas sobre longitud y jerarquía de vías, cobertura de transporte público y patrones de movilidad observados. Por su parte, el Ministerio de Vivienda y Urbanismo ofrece datos de zonificación urbana, uso de suelo y acceso a servicios urbanos, útiles para caracterizar la estructura territorial. Otras fuentes relevantes incluyen la Subsecretaría de Desarrollo Regional, con indicadores como el Índice de Desarrollo Comunal y la inversión en infraestructura, así como datos del Registro Civil, que contienen información sobre el parque vehicular por comuna y tasas de motorización. La integración de estas variables permitiría capturar factores contextuales no considerados, lo que podría mejorar la capacidad explicativa de los modelos y abrir paso a enfoques multiescalares o jerárquicos que aborden simultáneamente las dimensiones individuales y territoriales del fenómeno estudiado.

De forma complementaria, se recomienda explorar mejoras metodológicas que permitan abordar simultáneamente las dimensiones individuales y territoriales del fenómeno estudiado. Esto incluye la implementación de enfoques multiescalares o modelos jerárquicos, integración de datos observados de movilidad y la calibración con encuestas y censos más recientes (Censo 2024, Casen 2024-2025 e última encuesta origen destino disponible). Asimismo, replicar y adaptar este marco metodológico en otras regiones y países permitiría contrastar contextos urbanos y políticas de transporte diversas, evaluando su aplicabilidad y robustez. En este sentido, la propuesta no solo representa un aporte

técnico, sino también una herramienta estratégica para el análisis territorial y socioeconómico, con potencial para orientar políticas públicas y toma de decisiones.

Referencias

Agostini, C. A., Hojman, D., Román, A., & Valenzuela, L. (2016). Segregación residencial de ingresos en el Gran Santiago, 1992-2002: una estimación robusta. *Eure (Santiago)*.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arellana, J., Oviedo, D., Guzman, L. A., & Alvarez, V. (2021). Urban transport planning and access inequalities: A tale of two Colombian cities. *Research in Transportation Business & Management*.

Basu, R., & Ferreira, J. (2020). Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia*.

Berk, R. (2016). *Statistical learning from a regression perspective*. Springer.

Berk, R. A. (2016). *Statistical learning from a regression perspective* (Vol. 14). New York: Springer.

Bhat, C. R., & Pulugurta, V. (1998). A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. *Transportation Research Part B: Methodological*.

Bocarejo S, J. P., & Oviedo H, D. R. (2012). Transport accessibility and social inequities: a tool for identification of mobility needs and evaluation of transport investments. *Journal of transport geography*.

Bourdieu, P. (1986). The forms of capital. Cultural Theory: An Anthology, 81-93.

Breiman, L. (2001). Random forests. *Machine learning.*

CAF (Banco de desarrollo de América Latina). (2011). Desarrollo urbano y movilidad en América Latina. Observatorio de Movilidad Urbana.

Campos Miranda, H. (2004). Una nueva relación urbana para el Gran Concepción. Urbano, (9), 63-71.

Canal 9 Biobío televisión. (11 de noviembre de 2024). extenderán horarios de micros en Concepción, Talcahuano y Chiguayante. *Canal 9 Biobío televisión.* https://www.canal9.cl/episodios/noticias/2024/11/11/extenderan-horarios-de-micros-en-concepcion-talcahuano-y-chiguayante

CAVEM. (2014). El mercado automotriz chileno ante el nuevo escenario global y regional. Centro de Análisis de Vehículos de Motor.

CEPAL. (2016). La matriz de la desigualdad social en América Latina. Naciones Unidas.

CEPAL. (2018). Movilidad urbana sostenible en América Latina y el Caribe: desafíos y oportunidades. Comisión Económica para América Latina y el Caribe.

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. Transportation research part D: Transport and environment.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Davies, C. (2021). El rol de las ciudades intermedias en los planes estratégicos subnacionales. El caso de la provincia de Santa Fe. *Papeles: Revista del Centro de Investigaciones de la Facultad de Ciencias Jurídicas y Sociales de la Universidad Nacional del Litoral*.

Deaton, A. & Muellbauer, J. (1980). Economics and consumer behavior. Cambridge University Press.

Dote, S. (2024). Chile inaugura el tren más rápido de Sudamérica con un recorrido entre Santiago y Curicó. El País. https://elpais.com/chile/2024-01-19/chile-inaugura-el-tren-mas-rapido-de-sudamerica-con-un-recorrido-entre-santiago-y-curico.html

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, *71*(1), 355-364.

Ferrovial. (2025). *Movilidad.* Ferrovial. https://www.ferrovial.com/es/recursos/movilidad/#:~:text=La%20movilidad%20urbana%20es%20el,en%20transporte%20p%C3%BAblico%20o%20privado.

Friz, G., Belmar, V. (10 de marzo de 2025). Hasta 1 hora de retraso: la caótica mañana del Biotren en verdadero "súper lunes" en Concepción. *Biobío Chile.* https://www.biobiochile.cl/noticias/nacional/region-del-bio-bio/2025/03/10/hasta-1-hora-de-retraso-la-caotica-manana-del-biotren-en-verdadero-super-lunes-en-concepcion.shtml

Galindo, L. M., Samaniego, J., Alatorre, J. E., Ferrer, J., & Reyes, O. (2015). Meta-análisis de las elasticidades ingreso y precio de la demanda de gasolina: implicaciones de política pública para América Latina. Revista CEPAL, 117. Repositorio de la CEPAL.

Gobierno de Chile. (15 de julio de 2024). Más de 14 mil personas serán beneficiadas con nuevos buses entre Santa Juana y Concepción. *Gobierno de Chile.* https://www.gob.cl/noticias/nuevo-servicio-transporte-publico-santa-juana-concepcion-buses-caracteristicas-tarifas-rebajadas/

González, P. B. G. (2014). El mejoramiento del sistema de transporte y el espacio público en ciudades intermedias. Estudio de la oportunidad de implantación de un tranvía en Antofagasta. *Santiago*.

Granados, R. M. (2016). Modelos de regresión lineal múltiple. *Granada, España: Departamento de Economía Aplicada, Universidad de Granada*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Instituto Nacional de Estadísticas (INE). (2021). Estadísticas de vehículos en circulación. Santiago de Chile: INE.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

McNally, M. G., & Rindt, C. (2007). The activity-based approach. En D. A. Hensher & K. J. Button (Eds.), Handbook of Transport Modelling (Vol. 1, pp. 53–69). Elsevier.

Ministerio de Desarrollo Social. (2020). *Encuesta de Caracterización Socioeconómica Nacional (CASEN)*. Gobierno de Chile.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics.*

Morales Benítez, J. F. (2020). Modelo de predicción y estimación de tiempos de traslado entre dos puntos utilizando datos de GPS.

OCDE. (2019). *Under pressure: The squeezed middle class*. OECD Publishing.

OECD. (2022). Tackling air pollution in dense urban areas: The case of Santiago, Chile. OECD Publishing.

Ortúzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport (4th ed.). Wiley.

Oviedo, D., Guzmán, L. A., & Rivera, C. (2016). Accessibility and socioeconomic status in a context of transport system change: A Bogotá case study. *Journal of Transport Geography*.

Pastene, A. N. (2016). Centros tradicionales, nuevas centralidades y descentralización en metrópolis intermedias latinoamericanas: caso del Gran Concepción, Chile. *Cuaderno urbano*.

Ralph, K. M. (2022). Childhood car access: Long-term consequences for education, employment, and earnings. *Journal of Planning Education and Research*, *42*(1), 36-46.

ReportLinker. (2024). Forecast: Ownership of passenger cars in Chile. ReportLinker Dataset.

Reyes, V., Belmar, V. (15 de marzo de 2025). ¡Las fechas estimadas para que en el Gran Concepción se pague la micro con tarjeta al estilo Bip! *Biobío Chile.* https://www.biobiochile.cl/noticias/nacional/region-del-bio-bio/2025/03/15/las-fechas-estimadas-para-que-en-el-gran-concepcion-se-pague-la-micro-con-tarjeta-al-estilo-bip.shtml

Rodríguez, D. A., & Mojica, C. H. (2009). Capitalization of BRT network expansions effects into prices of non-expansion areas. *Transportation Research Part A: Policy and Practice*.

Rodríguez-Jaume, M. J., & Mora Catalá, R. (2001). Análisis de tablas de contingencia: modelos Log-lineales.

SECTRA. (2021). Encuesta Origen Destino Gran Concepción 2021. Ministerio de Transportes y Telecomunicaciones de Chile.

Soltani, A. (2017). Social and urban form determinants of vehicle ownership; evidence from a developing country. *Transportation Research Part A: Policy and Practice*.

Tovar, M. A., & Rodríguez, D. A. (2020). Car ownership in Latin American cities: Socioeconomic drivers and policy implications. Transportation Research Record.

Varian, H. R. (2014). Intermediate microeconomics: A modern approach (9th ed.). W.W. Norton & Company.

Vera, P., Arthur, P. (07 de noviembre de 2024). Dos trenes se sumarán al servicio de Biotren en el Gran Concepción para 2025: vienen desde China. *Biobío Chile*.https://www.biobiochile.cl/noticias/nacional/region-del-bio-bio/2024/11/07/dos-trenes-se-sumaran-al-servicio-de-biotren-en-el-gran-concepcion-para-2025-vienen-desde-china.shtml

Verma, M., Manoj, M., & Verma, A. (2016). Analysis of the influences of attitudinal factors on car ownership decisions among urban young adults in a developing country like India. *Transportation research part F: traffic psychology and behaviour*.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach* (7th ed).

Anexos

Anexo A

Estimación de Ingresos en el Censo 2017

Este notebook aplica un modelo entrenado en la Encuesta CASEN 2017 para estimar ingresos autónomos del hogar en la base del Censo 2017.

```
import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import make_scorer, mean_squared_error
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.ensemble import RandomForestRegressor
import matplotlib.pyplot as plt
import warnings
import seaborn as sns
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
import statsmodels.api as sm
from sklearn.metrics import r2_score, mean_squared_error
```

1. Entrenamiento del modelo con datos CASEN

```
# Suprimir warnings específicos
warnings.filterwarnings("ignore", category=FutureWarning)
# 1. Carga y preparación inicial de datos
```

```
df = pd.read_csv("…/casen_concepcion.csv",low_memory=False)


# 2. Selección y limpieza de columnas
columnas = ['sexo', 'edad', 'e6a', 'v1', 'v2', 'v6', 'v4', 'v27a', 'v20',
'numper', 'yautcorh']
df = df[columnas].replace(99, np.nan).dropna()


# 3. Conversión de tipos y filtrado
df['yautcorh'] = pd.to_numeric(df['yautcorh'], errors='coerce')
df = df[df['yautcorh'] > 0]


# 4. Creación de variables derivadas
df['personas_por_dormitorio'] = df['numper'] / df['v27a']
df['escolaridad_aprox'] = df['e6a'].map({
    0: 0, 1: 3, 2: 8, 3: 10, 4: 10, 5: 12, 6: 12,
    7: 13, 8: 14, 9: 14, 10: 15, 11: 16, 12: 18
})


# 5. Preparación de datos para el modelo
X = df[['sexo', 'edad', 'e6a', 'v1', 'v2', 'v6', 'v4',
        'v27a', 'v20', 'numper', 'personas_por_dormitorio',
'escolaridad_aprox']]
y = np.log(df['yautcorh'])


# 6. Eliminación de valores faltantes
X = X.dropna()
y = y[X.index]


# 7. Definición de variables categóricas
cat_vars = ['sexo', 'e6a', 'v1', 'v2', 'v6', 'v4', 'v20']
```

```
# 8. Pipeline con ColumnTransformer actualizado
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore', sparse_output=False), cat_vars)
    ],
    remainder='passthrough',
    force_int_remainder_cols=False  # Elimina la advertencia
)

pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=1000, random_state=42, n_jobs=-1))
])

# 9. División y entrenamiento del modelo
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
pipeline.fit(X_train, y_train)

# 10. Evaluación del modelo (opcional)
print(f"Score en entrenamiento: {pipeline.score(X_train, y_train):.3f}")
print(f"Score en prueba: {pipeline.score(X_test, y_test):.3f}")
Score en entrenamiento: 0.891
Score en prueba: 0.278
```

2. Evaluación del Modelo

```
#R2 y RMSE
y_pred = pipeline.predict(X_test)
print("R²:", r2_score(y_test, y_pred))
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
```

```
print("RMSE:", rmse)


# Validación cruzada (k-fold)
kfold = KFold(n_splits=5, shuffle=True, random_state=42)
# Evaluación con R²
r2_scores = cross_val_score(pipeline, X, y, cv=kfold, scoring='r2')
print("R² por fold:", np.round(r2_scores, 3))
print("R² promedio:", np.round(r2_scores.mean(), 3))
# Evaluación con RMSE
rmse_scorer = make_scorer(lambda y_true, y_pred:
np.sqrt(mean_squared_error(y_true, y_pred)), greater_is_better=False)
rmse_scores = cross_val_score(pipeline, X, y, cv=kfold, scoring=rmse_scorer)
rmse_scores = -rmse_scores  # sklearn retorna negativos por defecto
print("RMSE por fold:", np.round(rmse_scores, 3))
print("RMSE promedio:", np.round(rmse_scores.mean(), 3))
```

3. Aplicación al Censo 2017 (muestra representativa)

```
# Cargar muestras
df_personas = pd.read_csv("…/CENSO/csv-personas-censo-2017/microdato_censo2017-
personas/censo_personas.csv", usecols=["nviv", "nhogar", "p08", "p09", "p15"])
df_personas = df_personas.sample(n=300000) #muestra representativa aleatoria
df_viviendas = pd.read_csv("…/CENSO/csv-viviendas-censo-
2017/microdato_censo2017-viviendas/censo_viviendas.csv", usecols=["comuna",
"dc", "area", "zc_loc", "id_zona_loc", "nviv", "p01", "p03a", "p03b", "p03c",
"p04", "p05", "cant_per"])


# Preparación
df_personas = df_personas.dropna()
df_personas["hogar_id"] = df_personas["nviv"].astype(str) + "_" +
df_personas["nhogar"].astype(str)
hogar_size = df_personas.groupby("hogar_id").size().reset_index(name="numper")
hogar_jefe = df_personas.groupby("hogar_id").first().reset_index()
```

```
df_viviendas["hogar_id"] = df_viviendas["nviv"].astype(str) + "_1"

df_viviendas = df_viviendas.rename(columns={"p01": "v1", "p03a": "v2", "p03b":
"v6",
                                            "p03c": "v4", "p04": "v27a", "p05":
"v20"})

# Unión

df_hogares = hogar_jefe.merge(hogar_size,
on="hogar_id").merge(df_viviendas.drop(columns="nviv"), on="hogar_id")

df_hogares = df_hogares.rename(columns={"p08": "sexo", "p09": "edad", "p15":
"e6a"})

df_hogares["personas_por_dormitorio"] = df_hogares["numper"] /
df_hogares["v27a"]

df_hogares["escolaridad_aprox"] = df_hogares["e6a"].map({

    1: 1, 2: 2, 3: 3, 4: 4, 5: 8, 6: 8, 7: 10, 8: 12,

    9: 10, 10: 12, 11: 14, 12: 16, 13: 17, 14: 18

})


# Filtrar valores válidos

df_hogares = df_hogares[(df_hogares["v27a"] > 0) & (df_hogares["cant_per"] <=
20)]


# Reemplazar códigos 98/99 por NaN ANTES de imputar

variables_categoricas = ['v1', 'v2', 'v6', 'v4', 'v20', 'v27a']

for var in variables_categoricas:

    df_hogares[var] = df_hogares[var].replace({98: np.nan, 99: np.nan})


# Variables categóricas (solo admiten enteros)

categ_vars = ['v1', 'v2', 'v4', 'v6', 'v20', 'v27a']

# Variables numéricas continuas

num_vars = ['personas_por_dormitorio', 'escolaridad_aprox']


# Imputar categóricas con moda
```

```
for col in categ_vars:
    if df_hogares[col].isna().any():
        moda = df_hogares[col].mode()[0]
        df_hogares[col] = df_hogares[col].fillna(moda)

# Imputar numéricas solo si tienen NaN
df_hogares[num_vars] = df_hogares[num_vars].replace([np.inf, -np.inf], np.nan)

if df_hogares[num_vars].isna().any().any():
    imputador_rf = IterativeImputer(
        estimator=RandomForestRegressor(n_estimators=10, random_state=0),
        max_iter=5,
        random_state=0
    )
    df_hogares[num_vars] = imputador_rf.fit_transform(df_hogares[num_vars])
    print("Imputación numérica realizada.")
else:
    print("No hay valores faltantes en numéricas. Imputación no necesaria.")

print("Imputación finalizada correctamente.")
# Eliminar registros con valores extremos después del merge
df_hogares = df_hogares[df_hogares["cant_per"] <= 20]

# Preparar datos para predicción
X_pred = df_hogares[['sexo', 'edad', 'e6a', 'v1', 'v2', 'v6', 'v4',
                     'v27a', 'v20', 'numper', 'personas_por_dormitorio',
'escolaridad_aprox']]


# Aplicar el modelo (pipeline debe estar definido previamente)
```

```
log_pred = pipeline.predict(X_pred)
df_hogares["ingreso_estimado"] = np.exp(log_pred)


# Limpieza de valores 98 y 99 en variables categóricas antes de predicción #
variables_categoricas = ['v1', 'v2', 'v6', 'v4', 'v20', 'v27a']
for var in variables_categoricas:
    df_hogares = df_hogares[~df_hogares[var].isin([98, 99])]


#  Recalcular variables derivadas con datos ya limpios  #
df_hogares["personas_por_dormitorio"] = df_hogares["cant_per"] /
df_hogares["v27a"]
df_hogares["escolaridad_aprox"] = df_hogares["e6a"].map({
    1: 1, 2: 2, 3: 3, 4: 4, 5: 8, 6: 8, 7: 10, 8: 12,
    9: 10, 10: 12, 11: 14, 12: 16, 13: 17, 14: 18
})
df_hogares = df_hogares.dropna()


# Crear dataframe con columnas clave y derivadas
columnas_exportar = ['hogar_id', 'cant_per', 'v1', 'v2', 'v6', 'v4', 'v20',
'v27a',
                     'personas_por_dormitorio', 'escolaridad_aprox',
'ingreso_estimado']
df_hogares_final =
df_hogares[columnas_exportar].drop_duplicates(subset='hogar_id')


df_viviendas['cant_per'] = df_viviendas['cant_per'].replace({98: np.nan, 99:
np.nan})
df_viviendas = df_viviendas[df_viviendas['cant_per'] <= 20]



print("Hogares con datos completos:", df_hogares.shape[0])
```

```
# 4. Validación y análisis del modelo de estimación de ingresos
# Asegurar que y_test y y_pred estén definidos
y_real = np.exp(y_test)
y_pred_real = np.exp(y_pred)
residuos = y_test - y_pred

# Generar el identificador de hogar
df_hogares["hogar_id"] = (
    df_hogares["comuna"].astype(str) +
    df_hogares["dc"].astype(str) +
    df_hogares["zc_loc"].astype(str) +
    df_hogares["nviv"].astype(str)
)



# R² y RMSE
print("R²:", r2_score(y_test, y_pred))
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))

# Gráfico 1: Ingreso real vs predicho (log)
plt.figure(figsize=(7, 6))
sns.scatterplot(x=y_test, y=y_pred, alpha=0.3)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], linestyle='--',
color='red')
plt.xlabel("Ingreso Real (log)")
plt.ylabel("Ingreso Predicho (log)")
plt.title("Ingreso Real vs Predicho (log)")
plt.tight_layout()
plt.show()
```

```
# Distribución del ingreso antes y después del logaritmo
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
sns.histplot(df['yautcorh'], bins=50, kde=True)
plt.title("Distribución de ingreso original")
plt.xlabel("Ingreso autónomo del hogar corregido")

plt.subplot(1, 2, 2)
sns.histplot(np.log(df['yautcorh']), bins=50, kde=True)
plt.title("Distribución de ingreso log-transformado")
plt.xlabel("log(Ingreso autónomo corregido)")

plt.tight_layout()
plt.show()

#Ingreso estimado vs número de personas por hogar (gráfico de dispersión)
plt.figure(figsize=(8, 5))
sns.scatterplot(data=df_hogares, x="cant_per", y="ingreso_estimado", alpha=0.5)
plt.title("Ingreso estimado vs número de personas por hogar")
plt.xlabel("Número de personas en el hogar")
plt.ylabel("Ingreso estimado")
plt.tight_layout()
plt.show()

# Importancia de variables
importances = pipeline.named_steps['regressor'].feature_importances_
features = pipeline.named_steps['preprocessor'].get_feature_names_out()
df_importancias = pd.DataFrame({'Variable': features, 'Importancia':
importances})
```

```
df_importancias["Importancia_pct"] = 100 * df_importancias["Importancia"] /
df_importancias["Importancia"].sum()

df_importancias = df_importancias.sort_values(by="Importancia_pct",
ascending=False)


# Gráfico 5 Importancia Variables

colors = sns.color_palette("hsv", len(df_importancias.head(20)))

plt.figure(figsize=(10, 6))

sns.barplot(data=df_importancias.head(20), x="Importancia", y="Variable",
hue="Variable")

plt.xlabel("Importancia")

plt.title("Top 20 variables más importantes")

plt.tight_layout()

plt.show()


# Tabla con % de importancia de variables

df_tabla_formateada = df_importancias.head(20).copy()

df_tabla_formateada["Importancia_pct"] =
df_tabla_formateada["Importancia_pct"].apply(lambda x: f"{x:.2f} %")

df_tabla_formateada["Importancia"] =
df_tabla_formateada["Importancia"].apply(lambda x: round(x, 6))

display(df_tabla_formateada)



# Diccionario de códigos

comunas_dict = {

    "8101": "Concepción",

    "8102": "Coronel",

    "8103": "Chiguayante",

    "8104": "Florida",

    "8105": "Hualqui",

    "8106": "Lota",
```

```
    "8107": "Penco",
    "8108": "San Pedro de la Paz",
    "8109": "Santa Juana",
    "8110": "Talcahuano",
    "8111": "Tomé",
    "8112": "Hualpén"
}

# Extraer los 4 primeros dígitos como código de comuna
df_hogares["comuna_codigo"] = df_hogares["comuna"].astype(str).str[:4]
df_hogares["comuna_nombre"] = df_hogares["comuna_codigo"].map(comunas_dict)

# Verificar que haya comunas válidas
if df_hogares["comuna_nombre"].isnull().all():
    raise ValueError("No se encontraron comunas válidas después de extraer los 4 primeros dígitos.")

# Calcular ingreso promedio por comuna
ingreso_promedio = (
    df_hogares
    .groupby("comuna_nombre")["ingreso_estimado"]
    .mean()
    .dropna()
    .sort_values(ascending=False)
)

# Graficar
plt.figure(figsize=(10, 6))
sns.barplot(
    x=ingreso_promedio.values,
```

```
    y=ingreso_promedio.index,
    palette="viridis"
)
plt.xlabel("Ingreso estimado promedio ($)")
plt.ylabel("Comuna")
plt.title("Ingreso estimado promedio por comuna (Censo 2017)")
plt.tight_layout()
plt.show()

# Análisis cruzado con df_hogares
if 'ingreso_estimado' in df_hogares.columns:
    # Distribución de ingreso estimado
    plt.figure(figsize=(8, 5))
    sns.histplot(df_hogares["ingreso_estimado"], kde=True, bins=40)
    plt.title("Distribución de ingreso estimado por hogar (Censo)")
    plt.xlabel("Ingreso estimado")
    plt.tight_layout()
    plt.show()

    # Ingreso vs número de personas por hogar
    plt.figure(figsize=(8, 5))
    sns.boxplot(x=df_hogares["cant_per"], y=df_hogares["ingreso_estimado"])
    plt.title("Ingreso estimado vs número de personas por hogar (cant_per)")
    plt.tight_layout()
    plt.show()

    # Ingreso vs tipo de vivienda
    plt.figure(figsize=(10, 5))
    sns.boxplot(x=df_hogares["v1"], y=df_hogares["ingreso_estimado"])
    plt.title("Ingreso estimado vs tipo de vivienda")
```

```
    plt.tight_layout()
    plt.show()


    # Deciles de ingreso
    df_hogares["decil"] = pd.qcut(df_hogares["ingreso_estimado"], 10,
labels=False) + 1
    print("Ingreso promedio por decil:")
    print(df_hogares.groupby("decil")["ingreso_estimado"].mean().apply(lambda x:
f"${x:,.0f}"))


    # RESUMEN FINAL DE INGRESOS ESTIMADOS
    resumen = df_hogares["ingreso_estimado"].describe().apply(lambda x:
f"${x:,.0f}")
    print("Resumen de ingreso estimado por hogar")
    print(resumen)


    # Exportar a Excel
    columnas_exportar = [
    'hogar_id', 'cant_per',
    'v1', 'v2', 'v6', 'v4', 'v20', 'v27a', 'escolaridad_aprox',
    'personas_por_dormitorio','ingreso_estimado'
    ]


    # Filtrar duplicados por hogar_id
    df_hogares_exportar =
df_hogares[columnas_exportar].drop_duplicates(subset='hogar_id')


    # Exportar
    #df_hogares_exportar.to_excel("estimaciones_censo_FINAL.xlsx", index=False)
    print("Archivo exportado como: estimaciones_censo_FINAL.xlsx")
else:
```

```
    print("df_hogares no contiene ingreso_estimado. Asegúrate de haber ejecutado
la predicción correctamente.")
```

Anexo B

```
# MODELO POISSON

import pandas as pd

import statsmodels.formula.api as smf

import statsmodels.api as sm

from statsmodels.stats.outliers_influence import variance_inflation_factor

from patsy import dmatrix

from sklearn.ensemble import RandomForestClassifier

import matplotlib.pyplot as plt


# Cargar datos nuevos

df_nuevo = pd.read_excel("…/estimaciones_censo_FINAL.xlsx")


# Mapear comuna desde hogar_id

df_nuevo['comuna_cod'] = df_nuevo['hogar_id'].astype(str).str[:4]

comunas = {

    "8101": "CONCEPCION", "8102": "CORONEL", "8103": "CHIGUAYANTE",

    "8105": "HUALQUI", "8106": "LOTA", "8107": "PENCO", "8108": "SAN PEDRO DE LA
PAZ",

    "8110": "TALCAHUANO", "8111": "TOME", "8112": "HUALPEN"

}

orden_comunas = list(comunas.values())

df_nuevo['DireccionComuna'] = pd.Categorical(

    df_nuevo['comuna_cod'].map(comunas),

    categories=orden_comunas,

    ordered=False

)
```

```
# Clasificar ingreso estimado
def clasificar_tramo(ingreso):
    if ingreso <= 400_000:
        return 1
    elif ingreso <= 1_200_000:
        return 2
    else:
        return 3

df_nuevo['TramoIngresoHogar'] =
df_nuevo['ingreso_estimado'].apply(clasificar_tramo)

# Recodificar tipo de vivienda en df_nuevo
def recodificar_tipo_vivienda_externo(x):
    if x in [1, 2, 3]:
        return 'Casa'
    elif x in [4, 5]:
        return 'Departamento'
    else:
        return 'Otro'

df_nuevo['TipoVivienda'] =
df_nuevo['v1'].apply(recodificar_tipo_vivienda_externo)



# Cargar datos de entrenamiento
df_entrenamiento = pd.read_excel(
    "…/BD_EOD_GranConcepcion.xlsx",
    sheet_name="HOGAR_LAB"
)
```

```
df_entrenamiento['DireccionComuna'] = pd.Categorical(
    df_entrenamiento['DireccionComuna'],
    categories=orden_comunas,
    ordered=False
)

# Recodificar tipo de vivienda en entrenamiento
def recodificar_tipo_vivienda_entrenamiento(x):
    if x == 1:
        return 'Casa'
    if x == 2:
        return 'Casa'
    elif x == 3:
        return 'Departamento'
    elif x == 4:
        return 'Otro'
    else:
        return 'Otro'

df_entrenamiento['TipoVivienda'] =
df_entrenamiento['TipoVivienda'].apply(recodificar_tipo_vivienda_entrenamiento)

# Imputar TramoIngresoHogar en entrenamiento si es necesario
df_train = df_entrenamiento[df_entrenamiento["TramoIngresoHogar"].isin([1, 2,
3])].copy()
df_pred = df_entrenamiento[df_entrenamiento["TramoIngresoHogar"] == 9].copy()

if not df_pred.empty:
    print(f"Imputando {len(df_pred)} observaciones con TramoIngresoHogar=9...")
    X_train = pd.get_dummies(df_train.drop(columns=["TramoIngresoHogar",
"NumeroVehiculos"]), drop_first=True)
```

```
    y_train = df_train["TramoIngresoHogar"]


    clf = RandomForestClassifier(n_estimators=1000, random_state=42)
    clf.fit(X_train, y_train)


    X_pred = pd.get_dummies(df_pred.drop(columns=["TramoIngresoHogar",
"NumeroVehiculos"]), drop_first=True)
    X_pred = X_pred.reindex(columns=X_train.columns, fill_value=0)


    tramos_imputados = clf.predict(X_pred)
    df_entrenamiento.loc[df_pred.index, "TramoIngresoHogar"] = tramos_imputados


assert 'TamanoFamiliar' in df_entrenamiento.columns, "Falta TamanoFamiliar en
EOD"


# Entrenar el modelo
formula = """
NumeroVehiculos ~ C(TramoIngresoHogar) + C(DireccionComuna) + C(TipoVivienda) +
TamanoFamiliar
"""


poisson_model = smf.glm(
    formula=formula,
    data=df_entrenamiento,
    family=sm.families.Poisson()
).fit()


print(poisson_model.summary())


# Ajustar niveles en df_nuevo
niveles_comuna = df_entrenamiento['DireccionComuna'].dropna().unique()
```

```
df_nuevo.loc[~df_nuevo['DireccionComuna'].isin(niveles_comuna),
'DireccionComuna'] = 'CONCEPCION'


niveles_tipo = df_entrenamiento['TipoVivienda'].dropna().unique()

df_nuevo.loc[~df_nuevo['TipoVivienda'].isin(niveles_tipo), 'TipoVivienda'] =
'Casa'


niveles_tramo = df_entrenamiento['TramoIngresoHogar'].dropna().unique()

df_nuevo.loc[~df_nuevo['TramoIngresoHogar'].isin(niveles_tramo),
'TramoIngresoHogar'] = 1


df_nuevo['TamanoFamiliar'] = df_nuevo['cant_per']


# Predecir

df_nuevo['vehiculos_estimados'] = poisson_model.predict(df_nuevo)


# Exportar resultados

df_nuevo[['hogar_id', 'TramoIngresoHogar', 'DireccionComuna', 'TamanoFamiliar',
'vehiculos_estimados']].to_excel(

    "predicciones_vehiculos_FINAL.xlsx",

    index=False

)


# VIF

formula_vif = "C(TramoIngresoHogar) + C(DireccionComuna) + C(TipoVivienda) +
TamanoFamiliar"

X = dmatrix(formula_vif, df_entrenamiento, return_type='dataframe')

vif = pd.DataFrame()

vif["Variable"] = X.columns

vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

print(vif.sort_values(by='VIF', ascending=False))
```

```
# Equidispersión

media = df_entrenamiento['NumeroVehiculos'].mean()

varianza = df_entrenamiento['NumeroVehiculos'].var()

print(f"Media: {media:.3f}, Varianza: {varianza:.3f}")

print(" Predicciones guardadas en: predicciones_vehiculos_FINAL.xlsx")
```

#Graficos Modelo Poisson

```
#  Cargar archivo externo

df = pd.read_excel("…/predicciones_vehiculos_FINAL.xlsx")


# Histograma de predicciones

plt.figure(figsize=(8,5))

sns.histplot(df['vehiculos_estimados'], bins=30, color='blue', alpha=0.5,
kde=True, label='Estimado')

plt.title("Distribución de vehículos estimados en archivo externo")

plt.xlabel("Vehículos estimados")

plt.ylabel("Frecuencia")

plt.tight_layout()

plt.show()


# DISTRIBUCIÓN REAL vs ESTIMADA EN EL CONJUNTO DE ENTRENAMIENTO

# predicciones en entrenamiento

df_entrenamiento['vehiculos_estimados'] =
poisson_model.predict(df_entrenamiento)

plt.figure(figsize=(10,6))

sns.histplot(df_entrenamiento['NumeroVehiculos'], bins=30, color='orange',
alpha=0.5, kde=True, label='Observado')

sns.histplot(df_entrenamiento['vehiculos_estimados'], bins=30, color='blue',
alpha=0.5, kde=True, label='Estimado')

plt.title("Distribución de vehículos: observado vs estimado (entrenamiento)")

plt.xlabel("Número de vehículos")
```

```
plt.ylabel("Frecuencia")
plt.legend()
plt.tight_layout()
plt.show()

# DISTRIBUCIÓN REAL vs ESTIMADA EN EL CONJUNTO DE ENTRENAMIENTO
plt.figure(figsize=(10,6))

# Histograma estimado en entrenamiento
sns.histplot(df_entrenamiento['vehiculos_estimados'],
             bins=30, color='royalblue', alpha=0.3, stat='density',
             edgecolor=None, label='Entrenamiento')

sns.kdeplot(df_entrenamiento['vehiculos_estimados'], color='royalblue',
linewidth=2, bw_adjust=0.75)

# Histograma estimado en archivo externo
sns.histplot(df['vehiculos_estimados'],
             bins=30, color='seagreen', alpha=0.3, stat='density',
             edgecolor=None, label='Archivo externo')

sns.kdeplot(df['vehiculos_estimados'], color='seagreen', linewidth=2,
bw_adjust=1.5)

# Títulos y ejes
plt.title("Distribución de vehículos estimados\nEntrenamiento vs Archivo
externo", fontsize=14, weight='bold')
plt.xlabel("Número de vehículos estimados", fontsize=12)
plt.ylabel("Densidad", fontsize=12)

plt.xlim(0, 3)  # ajusta el rango si quieres
```

```
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.legend(title='', fontsize=11)
plt.tight_layout()
plt.show()




# === Promedio de predicciones por tramo de ingreso ===
promedios_tramo =
df.groupby("TramoIngresoHogar")['vehiculos_estimados'].mean().reset_index()


plt.figure(figsize=(8,5))
sns.barplot(x="TramoIngresoHogar", y="vehiculos_estimados",
data=promedios_tramo, palette="viridis")
plt.title("Promedio de vehículos estimados por tramo de ingreso")
plt.xlabel("Tramo de ingreso")
plt.ylabel("Vehículos estimados (promedio)")
plt.ylim(0, promedios_tramo["vehiculos_estimados"].max()*1.1)


# Agregar valores sobre las barras
for i, row in promedios_tramo.iterrows():
    plt.text(i, row['vehiculos_estimados'] + 0.02,
f"{row['vehiculos_estimados']:.2f}", ha='center', va='bottom')


plt.tight_layout()
plt.show()


print("\nPromedio de vehículos estimados por tramo de ingreso:")
print(promedios_tramo)


# === Promedio de predicciones por comuna ===
```

```
promedios_comuna =
df.groupby("DireccionComuna")['vehiculos_estimados'].mean().reset_index()


plt.figure(figsize=(12,6))

ax = sns.barplot(x="DireccionComuna", y="vehiculos_estimados",
data=promedios_comuna, palette="coolwarm")

plt.title("Promedio de vehículos estimados por comuna")

plt.xlabel("Comuna")

plt.ylabel("Vehículos estimados (promedio)")

plt.xticks(rotation=45)


# Agregar valores sobre las barras

for i, row in promedios_comuna.iterrows():

    ax.text(i, row['vehiculos_estimados'] + 0.02,
f"{row['vehiculos_estimados']:.2f}", ha='center', va='bottom')


plt.tight_layout()

plt.show()


print("\nPromedio de vehículos estimados por comuna:")

print(promedios_comuna)


# === Boxplot de la distribución de predicciones por comuna ===

plt.figure(figsize=(12,6))

sns.boxplot(x='DireccionComuna', y='vehiculos_estimados', data=df)

plt.xticks(rotation=45)

plt.title("Distribución de vehículos estimados por comuna")

plt.ylabel("Vehículos estimados")

plt.tight_layout()

plt.show()
```

```
# === Observado vs Estimado por Tramo ===


# calcular promedio observado por tramo desde df_entrenamiento

observados_tramo =
df_entrenamiento.groupby("TramoIngresoHogar")["NumeroVehiculos"].mean().reset_in
dex()

observados_tramo.columns = ["TramoIngresoHogar", "vehiculos_observados"]


# calcular promedio estimado por tramo desde df (archivo externo)

promedios_tramo =
df.groupby("TramoIngresoHogar")["vehiculos_estimados"].mean().reset_index()


# combinar ambos

comparacion = pd.merge(promedios_tramo, observados_tramo,
on="TramoIngresoHogar")


# graficar

plt.figure(figsize=(8,5))

plt.plot(comparacion['TramoIngresoHogar'], comparacion['vehiculos_estimados'],
marker='o', label='Estimado')

plt.plot(comparacion['TramoIngresoHogar'], comparacion['vehiculos_observados'],
marker='x', label='Observado')

plt.title("Promedio de vehículos por tramo de ingreso (observado vs estimado)")

plt.xlabel("Tramo de ingreso")

plt.ylabel("Promedio de vehículos")

plt.legend()

plt.grid(True)

plt.tight_layout()

plt.show()


print("\nComparación estimado vs observado por tramo:")

print(comparacion)
```

```
#grafico promedio vehículos observados vs estimados por comuna
comparacion_long = comparacion_comuna.melt(
    id_vars="DireccionComuna",
    value_vars=["Vehiculos_Observados", "Vehiculos_Estimados"],
    var_name="Tipo",
    value_name="Promedio"
)

# colores base distintos por comuna
comunas = comparacion_comuna['DireccionComuna'].unique()
base_colors = sns.color_palette("tab10", len(comunas))

# mapa: comuna → color
color_map = {comuna: base_colors[i] for i, comuna in enumerate(comunas)}

plt.figure(figsize=(14,6))

# ancho de las barras
bar_width = 0.35
x = np.arange(len(comunas))

fig, ax = plt.subplots(figsize=(14,6))

for i, comuna in enumerate(comunas):
    obs = comparacion_comuna.loc[comparacion_comuna['DireccionComuna'] ==
comuna, 'Vehiculos_Observados'].values[0]
    est = comparacion_comuna.loc[comparacion_comuna['DireccionComuna'] ==
comuna, 'Vehiculos_Estimados'].values[0]

    # color base para la comuna
```

```
    base = np.array(color_map[comuna])


    # para estimado: tono diferente (ligeramente rotado)
    est_color = tuple(np.clip(base * 0.6 + 0.4, 0, 1))  # más oscuro pero con otro contraste


    ax.bar(x[i] - bar_width/2, obs, bar_width, label=f"{comuna} Observado", color=base)
    ax.bar(x[i] + bar_width/2, est, bar_width, label=f"{comuna} Estimado", color=est_color)


    # agregar texto
    ax.text(x[i] - bar_width/2, obs + 0.01, f"{obs:.2f}", ha='center', va='bottom', fontsize=8)
    ax.text(x[i] + bar_width/2, est + 0.01, f"{est:.2f}", ha='center', va='bottom', fontsize=8)


ax.set_xlabel("Comuna")
ax.set_ylabel("Promedio de vehículos")
ax.set_title("Promedio de vehículos por comuna: Observado vs Estimado")
ax.set_xticks(x)
ax.set_xticklabels(comunas, rotation=45)


# leyenda solo con observados y estimados
handles = [plt.Rectangle((0,0),1,1,color="grey"), plt.Rectangle((0,0),1,1,color="lightgrey")]
ax.legend(handles, ['Observado', 'Estimado'], loc='upper right')


plt.tight_layout()
plt.show()
```

Anexo C

	Variable	VIF
0	Intercept	10.497890
9	C(DireccionComuna)[T.TALCAHUANO]	1.563453
8	C(DireccionComuna)[T.SAN PEDRO DE LA PAZ]	1.473345
3	C(DireccionComuna)[T.CORONEL]	1.459872
11	C(DireccionComuna)[T.HUALPEN]	1.360503
4	C(DireccionComuna)[T.CHIGUAYANTE]	1.323995
6	C(DireccionComuna)[T.LOTA]	1.278331
10	C(DireccionComuna)[T.TOME]	1.246979
7	C(DireccionComuna)[T.PENCO]	1.181396
1	C(TramoIngresoHogar)[T.2]	1.137431
14	TamanoFamiliar	1.133307
2	C(TramoIngresoHogar)[T.3]	1.122001
12	C(TipoVivienda)[T.Departamento]	1.095551
5	C(DireccionComuna)[T.HUALQUI]	1.087335
13	C(TipoVivienda)[T.Otro]	1.004498

Anexo D

```
#Modelo Logistico

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import train_test_split

from sklearn.metrics import classification_report

import statsmodels.api as sm

from patsy import dmatrices

from sklearn.metrics import roc_curve, roc_auc_score, confusion_matrix,
accuracy_score, precision_score, recall_score

from statsmodels.stats.outliers_influence import variance_inflation_factor

# Cargar base de datos

df = pd.read_excel("…/BD_EOD_GranConcepcion.xlsx", sheet_name="HOGAR_LAB")

# Modelo logístico

# 1. Variables a usar
```

```
vars_modelo = [
    "TramoIngresoHogar", "TipoVivienda", "SituacionVivienda", "DireccionComuna",
    "Tamaño Hogar", "CuentasInternet", "NumeroVehiculos"
]

df_modelo = df[vars_modelo].dropna()



# 2. Imputación del tramo de ingreso

df_train = df_modelo[df_modelo["TramoIngresoHogar"].isin([1, 2, 3])].copy()
df_pred = df_modelo[df_modelo["TramoIngresoHogar"] == 9].copy()

X_train = pd.get_dummies(df_train.drop(columns="TramoIngresoHogar"),
drop_first=True)
y_train = df_train["TramoIngresoHogar"]

clf = RandomForestClassifier(n_estimators=1000, random_state=42)
clf.fit(X_train, y_train)

X_pred = pd.get_dummies(df_pred.drop(columns="TramoIngresoHogar"),
drop_first=True)
X_pred = X_pred.reindex(columns=X_train.columns, fill_value=0)

tramos_imputados = clf.predict(X_pred)

df_imputado = df.copy()
df_imputado.loc[df_pred.index, "TramoIngresoHogar"] = tramos_imputados



# 3. Preparar base para modelo logístico
```

```
df_modelo_logit = df_imputado[[
    "TramoIngresoHogar", "Tamaño Hogar", "NumeroVehiculos",
    "TipoVivienda", "SituacionVivienda", "CuentasInternet"
]].dropna()

df_modelo_logit["posee_auto"] = (df_modelo_logit["NumeroVehiculos"] >
0).astype(int)

# Agrupar Tamaño Hogar en 3 categorías
df_modelo_logit["Tamaño_Hogar_cat"] = pd.cut(
    df_modelo_logit["Tamaño Hogar"],
    bins=[0,2,4,15],
    labels=["1–2","3–4","5+"]
)



# 4. Fórmula extendida

formula = """
posee_auto ~ C(TramoIngresoHogar) + C(Tamaño_Hogar_cat) +
             C(TipoVivienda) + C(SituacionVivienda) +
             C(CuentasInternet)
"""



# 5. Crear matrices

y, X = dmatrices(formula, df_modelo_logit, return_type="dataframe")
```

```
# 6. Ajustar modelo logístico

modelo_logit_ext = sm.Logit(y, X).fit()

# Mostrar y exportar resumen
print(modelo_logit_ext.summary2())
with open("resumen_modelo_logit_categorizado.txt", "w") as f:
    f.write(modelo_logit_ext.summary2().as_text())
print(" Resumen exportado: resumen_modelo_logit.txt")



# 7. Validaciones adicionales

y_true = y.iloc[:, 0]
y_pred_prob = modelo_logit_ext.predict(X)
y_pred_class = (y_pred_prob >= 0.5).astype(int)

# AUC y curva ROC
auc = roc_auc_score(y_true, y_pred_prob)
fpr, tpr, thresholds = roc_curve(y_true, y_pred_prob)

plt.figure(figsize=(8,6))
plt.plot(fpr, tpr, color='orange', label=f"AUC = {auc:.3f}")
plt.plot([0,1], [0,1], 'k--')
plt.xlabel("1 - Especificidad (FPR)")
plt.ylabel("Sensibilidad (TPR)")
plt.title("Curva ROC")
plt.legend(loc="lower right")
plt.show()
```

```
print(f"\n AUC: {auc:.3f}")

# Matriz de confusión y métricas
cm = confusion_matrix(y_true, y_pred_class)
accuracy = accuracy_score(y_true, y_pred_class)
precision = precision_score(y_true, y_pred_class)
recall = recall_score(y_true, y_pred_class)

print("\n Matriz de Confusión:")
print(cm)
print(f"\n Exactitud (Accuracy): {accuracy:.3f}")
print(f" Precisión: {precision:.3f}")
print(f" Sensibilidad (Recall): {recall:.3f}")



# 8. VIF para multicolinealidad

vif_logit = pd.DataFrame()
vif_logit["variable"] = X.columns
vif_logit["VIF"] = [variance_inflation_factor(X.values, i) for i in
range(X.shape[1])]

print("\n VIF para las variables independientes:")
print(vif_logit)

# Agregar predicción al dataframe
df_modelo_logit["probabilidad_auto"] = modelo_logit_ext.predict(X)
df_modelo_logit.to_excel("hogares_con_probabilidad_EODFINAL.xlsx", index=False)

print(" Modelo exportado y aplicado correctamente.")
```

```
#  Verificación de outliers influyentes: leverage y Cook’s distance

influence = modelo_logit_ext.get_influence()

# leverage (hat values)
leverage = influence.hat_matrix_diag

# Cook’s distance
cooks_d = influence.cooks_distance[0]

# Gráfico
plt.figure(figsize=(10,6))
plt.scatter(leverage, cooks_d, alpha=0.5)
plt.xlabel("Leverage")
plt.ylabel("Cook's distance")
plt.title("Leverage vs Cook's Distance")
plt.show()

# Valores típicos de referencia:
# – leverage alto: > 2*(k+1)/n
# – Cook’s distance preocupante: > 1
n, k = X.shape
leverage_threshold = 2 * (k+1) / n

print(f"Leverage medio: {np.mean(leverage):.4f}")
print(f"Leverage máximo: {np.max(leverage):.4f}")
print(f"Leverage umbral sugerido: {leverage_threshold:.4f}")
print(f"Máximo Cook's distance: {np.max(cooks_d):.4f}")
```

```
# Figura: Porcentaje de hogares con auto por tramo REAL
por_tramo = df_modelo_logit.groupby("TramoIngresoHogar")["posee_auto"].mean().reset_index()
por_tramo["porcentaje"] = (por_tramo["posee_auto"] * 100).round(2)


plt.figure(figsize=(8, 6))
sns.barplot(x="TramoIngresoHogar", y="porcentaje", data=por_tramo)
plt.title("Figura 2: Porcentaje de hogares que poseen automóvil por tramo de ingreso")
plt.xlabel("Tramo de ingreso")
plt.ylabel("Porcentaje de hogares con automóvil")
plt.ylim(0, 100)
for i, row in por_tramo.iterrows():
    plt.text(i, row["porcentaje"] + 2, f"{row['porcentaje']}%", ha='center')
plt.show()
```

#APLICACION SOBRE ESTIMACION

```
#  Cargar archivo con ingresos estimados
archivo_est = "…/hogares_con_probabilidad_auto_actualizadoFINAL.xlsx"
df_est = pd.read_excel(archivo_est)


#  Clasificar tramo de ingreso
def clasificar_tramo(ingreso):
    if ingreso <= 600_000:
        return 1
    elif ingreso <= 1_000_000:
        return 2
    else:
        return 3
```

```
df_est["tramo_ingreso"] =
df_est["ingreso_estimado"].apply(clasificar_tramo).astype(int)


#  Agrupar tamaño hogar
df_est.rename(columns={"cant_per": "tamano_hogar"}, inplace=True)


def clasificar_tamano_hogar(cant):
    if cant <= 2:
        return "1–2"
    elif cant <= 4:
        return "3–4"
    else:
        return "5+"


df_est["Tamaño_Hogar_cat"] =
df_est["tamano_hogar"].apply(clasificar_tamano_hogar)


#  Asumir valores base para otras variables
df_est["TipoVivienda"] = "Tipo1"  # categoría base
df_est["SituacionVivienda"] = "Situacion1"  # categoría base
df_est["CuentasInternet"] = "Nivel1"  # categoría base


# Preparar matrices para el modelo
formula = """
C(tramo_ingreso) + C(Tamaño_Hogar_cat) +
C(TipoVivienda) + C(SituacionVivienda) +
C(CuentasInternet)
"""


X_nuevo = dmatrix(formula, df_est, return_type='dataframe')
```

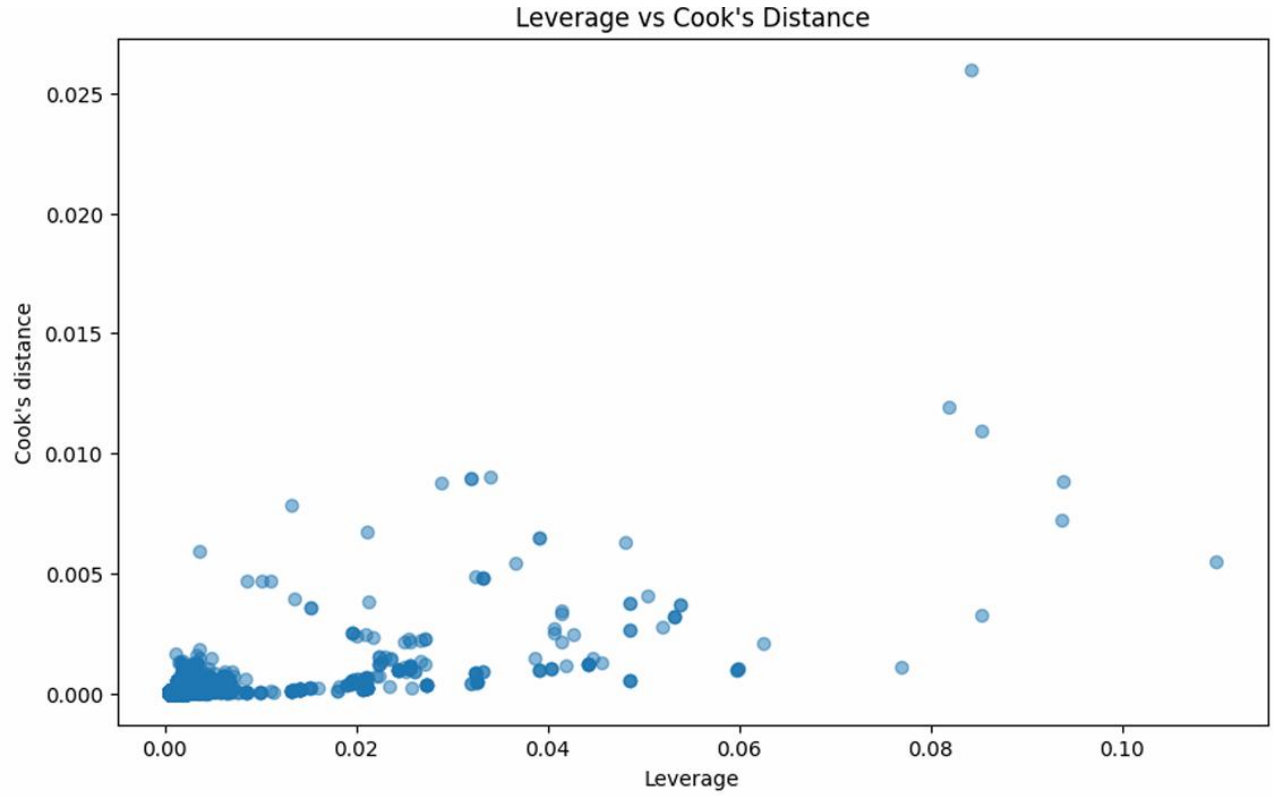
```
#  Alinear columnas con las del modelo entrenado
X_nuevo = X_nuevo.reindex(columns=modelo_logit_ext.model.exog_names,
fill_value=0)


#  Predecir con el modelo entrenado
df_est["probabilidad_auto"] = modelo_logit_ext.predict(X_nuevo)


# Resumen y figura
resumen = (
    df_est.groupby("tramo_ingreso")
    .agg(media_probabilidad=("probabilidad_auto", "mean"),
         n_hogares=("probabilidad_auto", "count"))
    .reset_index()
)


print(resumen)


# Figura: distribución de probabilidades por tramo
sns.set(style="whitegrid")
plt.figure(figsize=(10, 6))
sns.boxplot(x="tramo_ingreso", y="probabilidad_auto", data=df_est)
plt.title("Figura: Distribución de la probabilidad de poseer automóvil según
tramo de ingreso")
plt.xlabel("Tramo de ingreso")
plt.ylabel("Probabilidad de poseer automóvil")
plt.tight_layout()
plt.show()


# Eliminar las columnas fijas antes de exportar
df_est.drop(columns=["TipoVivienda", "SituacionVivienda", "CuentasInternet"],
inplace=True)
```

```
#  Exportar resultados
df_est.to_excel("hogares_con_probabilidad_auto_actualizado.xlsx", index=False)
print(" Archivo exportado: hogares_con_probabilidad_auto_actualizado.xlsx")
```

Anexo E

```
 VIF para las variables independientes:
                   variable        VIF
                  Intercept  15.783167
  C(TramoIngresoHogar)[T.2]   1.147659
  C(TramoIngresoHogar)[T.3]   1.224393
 C(Tamaño_Hogar_cat)[T.3-4]   1.310405
  C(Tamaño_Hogar_cat)[T.5+]   1.324522
       C(TipoVivienda)[T.2]   1.022520
       C(TipoVivienda)[T.3]   1.094108
       C(TipoVivienda)[T.4]   1.012217
  C(SituacionVivienda)[T.2]   1.125268
  C(SituacionVivienda)[T.3]   1.119975
  C(SituacionVivienda)[T.4]   1.019736
  C(SituacionVivienda)[T.5]   1.023421
  C(SituacionVivienda)[T.6]   1.005195
  C(SituacionVivienda)[T.7]   1.012129
  C(SituacionVivienda)[T.8]   1.003087
    C(CuentasInternet)[T.2]   2.103037
    C(CuentasInternet)[T.3]   2.349205
```

Anexo F

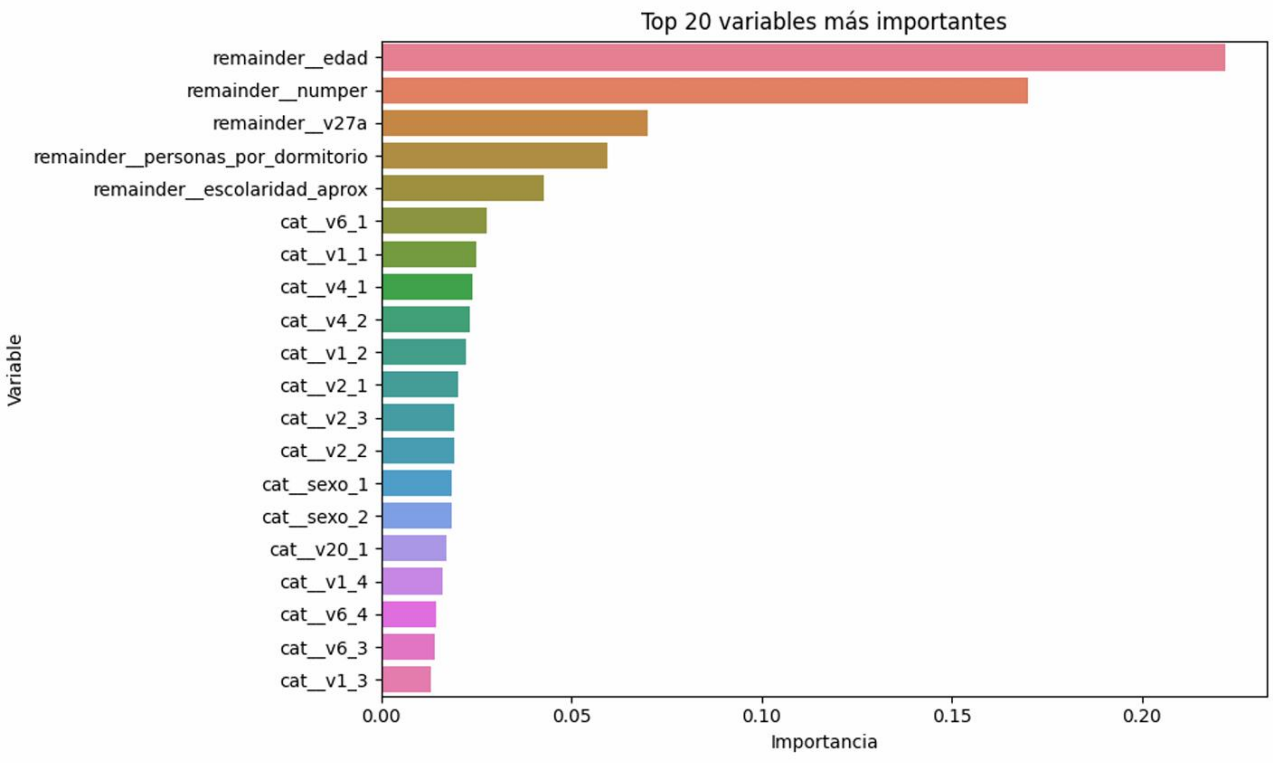



```
Leverage medio: 0.0025
Leverage máximo: 0.1097
Leverage umbral sugerido: 0.0052
Máximo Cook's distance: 0.0260
```

Anexo G

```
Ingreso promedio por decil:
decil
1       $415,327
2       $519,553
3       $586,551
4       $607,086
5       $627,115
6       $664,768
7       $782,228
8       $902,480
9     $1,101,757
10    $1,529,177
Name: ingreso_estimado, dtype: object
```

Anexo H


Top 20 variables más importantes
remainder__edad
remainder__numper
remainder__v27a
remainder__personas_por_dormitorio
remainder__escolaridad_aprox
cat__v6_1
cat__v1_1
cat__v4_1
cat__v4_2
cat__v1_2
cat__v2_1
cat__v2_3
cat__v2_2
cat__sexo_1
cat__sexo_2
cat__v20_1
cat__v1_4
cat__v6_4
cat__v6_3
cat__v1_3
Variable
0.00
0.05
0.10
0.15
0.20
Importancia