Universidad de Concepción
Facultad de Ingeniería
Departamento de Ingeniería Industrial

Valor estadístico de la vida: Regresión lineal versus regresión simbólica

POR

Tomás David Vergara Obreque

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

Profesores Guía:

Marcela Parada Contzen

Carlos Contreras Bolton

Concepción, julio de 2025

Sumario

El presente estudio se basa en la aplicación del enfoque de salarios hedónicos para estimar el valor estadístico de la vida en Chile, utilizando información proveniente de la Encuesta de Caracterización Socioeconómica Nacional (CASEN), junto con datos de las tasas de accidentabilidad fatal y no fatal, obtenidos por medio de los boletines de la Superintendencia de Seguridad Social (SUSESO) correspondientes al año 2017 y 2022. Para la estimación del valor estadístico de la vida se aplicaron dos metodologías de estimación, el método de mínimos cuadrados ordinarios mediante regresión lineal y la regresión simbólica, aplicada con dos variantes, un enfoque tradicional que busca expresiones no lineales para optimizar la capacidad predictiva del modelo y un segundo enfoque que utiliza como base la regresión lineal, para encontrar expresiones no lineales sobre sus residuos. En base a los resultados obtenidos, la regresión simbólica sobre residuos demuestra mayor capacidad predictiva y flexibilidad al revelar relaciones complejas, en tanto el método de mínimos cuadrado ordinarios resulta muy competitivo cuando se prioriza la rapidez y sencillez del ajuste.

Abstract

This study applies the hedonic-wage approach to estimate the Value of a Statistical Life in Chile, using data from the National Socio-Economic Characterization Survey (CASEN) combined with fatal and non-fatal accident rates reported in the 2017 and 2022 bulletins of the Superintendence of Social Security (SUSESO). Two estimation techniques are employed: ordinary least squares (OLS) linear regression and symbolic regression. The latter is implemented in two variants: a traditional approach that searches for nonlinear expressions to maximize predictive accuracy, and a residual-based approach that starts from the linear model and seeks nonlinear patterns within its residuals. The results show that residual-based symbolic regression offers superior predictive performance and greater flexibility by uncovering complex relationships, whereas OLS remains highly competitive when speed and simplicity of estimation are prioritized.

Índice

Índice de Tablas

Índice de Figuras

1. Introducción

El presente capítulo introduce el enfoque y la relevancia del estudio en torno a la estimación del valor estadístico de la vida en Chile, comparando métodos de regresión lineal y simbólica. Se exponen objetivos generales y específicos, junto con los principales alcances y limitaciones de la investigación. Finalmente, se describe la estructura del documento.

1.1 Motivación

Las personas revelan sus preferencias por la seguridad a través de los riesgos que están dispuestas a asumir, como fumar cigarrillos, conducir automóviles o elegir trabajos peligrosos (Viscusi, 1993)**.** Si bien los riesgos para la salud son indeseables, las personas aceptan ciertos niveles de riesgos cuando se acompañan de beneficios, como mayores ingresos o comodidad, lo que permite observar preferencias frente al riesgo (Viscusi & Aldy, 2003). El valor estadístico de la vida se define como la disposición a pagar para reducir el riesgo de fatalidad o, alternativamente, como la disposición a aceptar una compensación por un mayor riesgo de mortalidad (Parada-Contzen, Riquelme-Won, & Vasquez-Lavin, 2012). Esta cuantificación permite monetizar los beneficios de disminuir los riesgos y facilitar la comparación entre costos y beneficios, facilitando a los responsables de formular políticas públicas, una medida que refleje dicho valor asociado a la reducción de riesgos, para evaluar la disposición a pagar de la sociedad por mejoras en la salud (Kniesner & Viscusi, 2019).

Dentro del enfoque de reducción de riesgo para el cálculo del valor estadístico de la vida predominan dos metodologías: los modelos de salarios hedónicos y los modelos de valoración contingente. El modelo de salarios hedónicos se basa en la idea de que los trabajadores eligen entre distintas combinaciones de salario y atributos del empleo, revelando de este modo sus preferencias respecto a condiciones laborales específicas (Kniesner, Viscusi, & Ziliak, 2010). A partir de estas elecciones, es posible estimar el valor monetario que asignan al riesgo (Kniesner & Leeth, 2010). Por otro lado, el método de valoración contingente se aplica cuando no existen mercados con precios observables para ciertos riesgos o bienes (Masterman & Viscusi, 2020). Por lo que este método utiliza encuestas que construyen un mercado hipotético, en las cuales los individuos declaran su disposición a pagar por una mejora o reducción en un riesgo, lo que permite estimar el valor económico asociado a dicho cambio (Carson & Hanemann, 2005).

La metodología de salarios hedónicos presenta una serie de limitaciones que inciden en la precisión de la estimación (Parada-Contzen, Riquelme-Won, & Vasquez-Lavin, 2012). Entre estas limitaciones,

se destaca el sesgo por especificación incorrecta (o misspecification bias en inglés), que ocurre cuando el modelo utilizado para analizar los datos no representa adecuadamente la relación entre sus variables, lo que puede generar estimadores sesgados (Blackwell & Olson, 2022). Lo anterior se ve reforzado por Hainmueller & Hazlett (2014), quienes advierten que "los modelos mal especificados pueden generar sesgos, ineficiencia, un control incompleto de las variables, inferencias incorrectas y resultados frágiles dependientes del modelo". En este contexto, la regresión simbólica ofrece una alternativa prometedora al descubrir nuevas relaciones funcionales no lineales y potenciales relaciones causales, manteniendo un alto grado de interpretabilidad (Makke & Chawla, 2024).

A partir de lo anterior, esta memoria de título estima el valor estadístico de la vida aplicando metodologías de regresión lineal y simbólica, comparando los resultados obtenidos. El enfoque principal de esta investigación es la implementación de la regresión simbólica para mejorar la precisión en la estimación. Incluyendo la recopilación y la elaboración de los datos, la aplicación de ambas regresiones y la evaluación de la robustez de las estimaciones generadas. Finalmente, se lleva a cabo la comparación entre el método convencional y el método sugerido.

Esta memoria de título contribuye en la innovación para la formulación de políticas públicas efectivas y focalizadas en la mitigación de riesgos laborales, optimizando la asignación de los recursos públicos y el bienestar social, por medio de la estimación del valor estadístico de la vida. Se propone una técnica innovadora para superar las limitaciones y los sesgos presentes en las metodologías convencionales, mejorando la precisión y la eficiencia en las estimaciones proporcionando una base sólida para capacitar a nuevos profesionales en el uso de técnicas avanzadas de modelamiento computacional.

Para esta memoria se utilizaron datos de la Asociación Chilena de Seguridad (ACHS) y datos de la encuesta de caracterización socioeconómica nacional (CASEN). Posteriormente, se prepararon los datos y se empleó el método de salarios hedónicos para estimar el valor estadístico de la vida. Luego, se aplicaron las metodologías de regresión lineal y simbólica para la estimación, evaluando los resultados y comparando la precisión y robustez de ambos métodos.

1.2 Objetivo del tema

1.2.1 Objetivo general

Comparar la precisión de los resultados al utilizar métodos de salarios hedónicos, a través de modelos de regresión lineal y modelos de regresión simbólica, para la estimación del valor estadístico de la vida para Chile.

1.2.2 Objetivos específicos

1. Estimar el valor estadístico de la vida utilizando la metodología de salarios hedónicos mediante modelos de regresión lineal, a través de datos provenientes de las Encuesta de Caracterización Socioeconómica Nacional 2017 y 2022 y de la Superintendencia de Seguridad Social 2017 y 2022.
2. Evaluar el ajuste del modelo de regresión simbólica en el mismo conjunto de datos utilizados en la regresión lineal, mediante métricas estadísticas de desempeño, para asegurar la consistencia de los resultados.
3. Estimar el valor estadístico de la vida aplicando el método de regresión simbólica, optimizando la metodología en función del mismo conjunto de datos utilizado en la regresión lineal.
4. Comparar los resultados de las estimaciones generadas por ambos modelos mediante métricas estadísticas e interpretar los resultados.

1.3 Alcances y limitaciones

La presente investigación estima el valor estadístico de la vida para Chile mediante el enfoque de salarios hedónicos, combinando datos de las encuestas CASEN con las tasas de accidentabilidad publicadas por la SUSESO en los años 2017 y 2022. Aplicando métodos de regresión simbólica y comparándolo con métodos basado en regresiones lineales.

La aplicación de la regresión simbólica enfrenta varias limitaciones. Dado que su algoritmo evolutivo explora un espacio inmenso de ecuaciones, los resultados pueden depender sensiblemente de la semilla aleatoria y de la forma en que se acote la búsqueda. Para contener esa complejidad se entrena el modelo sobre los residuos de la regresión lineal, decisión que facilita la convergencia, pero restringe la libertad del algoritmo para descubrir patrones completamente nuevos desde el inicio. Además, los limites computacionales y los tiempos de cómputo obligan a fijar valores máximos para población, iteraciones y profundidad en la búsqueda. A lo anterior, se suma que las tasas de accidentabilidad se asignan de forma sectorial, cualquier heterogeneidad interna afecta en el coeficiente de riesgo, y con ello, en la estimación del valor estadístico de la vida. En conjunto, los beneficios de la regresión simbólica quedan condicionados por la sensibilidad de la parametrización, las restricciones computacionales y la precisión de los datos disponibles.

1.4 Organización del documento

La presente memoria de título se organiza de la siguiente forma. El capítulo 2 presenta una revisión de la literatura, en la cual se revisan artículos relacionados a la estimación del valor estadístico de la vida y utilización de la regresión simbólica. El capítulo 3 se expone las metodologías de estimación seleccionadas y aplicadas es el presente estudio. El capítulo 4 describe las fuentes de datos utilizadas y sus principales características. El capítulo 5 analiza y presentan los resultados obtenidos a partir de las metodologías aplicadas. Finalmente, el capítulo 6 sintetiza las conclusiones y principales hallazgos del estudio.

2. Revisión de la literatura

El presente capítulo revisa los principales enfoques teóricos y empíricos para estimar el valor estadístico de la vida, incluyendo métodos de preferencias reveladas y declaradas. Además, se exponen estudios relevantes aplicados en Chile y se analiza la regresión simbólica como una metodología emergente de con alto potencial explicativo e interpretativo

2.1 El valor estadístico de la vida

El valor estadístico de la vida es la tasa de compensación local entre el riesgo de mortalidad y el dinero (Kniesner & Viscusi, 2019). Cuando dichos valores de compensación se obtienen a partir de elecciones en contextos de mercado, el valor estadístico de la vida, refleja tanto la disposición a pagar de la población por reducciones marginales en el riesgo de muerte, como el costo marginal de mejorar la seguridad (Hammitt, 2000). La estimación de los beneficios asociados a la reducción del riesgo de mortalidad suele ser el criterio determinante para evaluar si una política regulatoria resulta eficiente o justificable (Escobar, 2007). Este enfoque ha sido adoptado por diversos estados para fundamentar decisiones sobre políticas públicas en áreas como protección ambiental y la seguridad en el transporte, donde la cuantificación de vidas salvadas permite priorizar intervenciones según su costo-efectividad (Kniesner & Viscusi, 2019).

Una de las aplicaciones del valor estadístico de la vida más comunes es en el ámbito de seguridad vial, donde las agencias gubernamentales utilizan dicha estimación para priorizar proyectos de infraestructura según su costo por vida salvada (Viscusi, 2015). Dicha lógica aplica también en la evaluación de normativas ambientales, por medio de la utilización del valor estadístico de la vida se puede determinar si los beneficios esperados por la reducción de riesgos justifican los costos asociados (Viscusi, 2000). En el campo de la salud, el valor estadístico de la vida se transforma en medidas utilizadas para comparar intervenciones medicas en términos de costo-efectividad (Ashenfelter, 2006). Sin embargo, según Viscusi (2021), el uso actual del valor estadístico de la vida es limitado y debería ampliarse a otras aplicaciones, como la determinación de sanciones por violaciones regulatorias, decisiones sobre gasto militar, valoración de daños en casos judiciales relaciones con muertes evitables y la limpieza de sitios contaminados.

2.1.1 Metodologías para la estimación del valor estadístico de la vida

Desde que se desarrolló el concepto de la disposición a pagar, el valor estadístico de la vida se ha estimado en múltiples estudios. Como la reducción del riesgo de mortalidad no es comercializable de

manera directa en el mercado, se utilizan métodos alternativos para la estimación (Viscusi & Aldy, 2003). Estos métodos se dividen en dos grupos, los métodos de preferencias reveladas, que se basan en decisiones reales observadas en mercados existentes y los métodos de preferencias declaradas o de valoración contingente, que utilizan escenarios hipotéticos para la recolección de datos (Andersson & Treich, 2011).

2.1.1.1 Métodos de preferencias declaradas

La metodología de preferencias declaradas o valoración contingente para la estimación del valor estadístico de la vida se basa en encuestas que capturan la disposición a pagar de los individuos para reducir riesgos de mortalidad (Carson & Louviere, 2011). Al depender de escenarios hipotéticos, la metodología de preferencias declaradas es susceptible a sesgos como el sesgo hipotético, que puede llevar a una sobreestimación sistemática del valor declarado respecto al comportamiento real (Murphy, Allen, Stevens, & Weatherhead, 2005). Por otro lado, dicha metodología permite una mayor flexibilidad para adaptarse a contextos específicos, lo cual puede ser ventajoso en ausencia de datos laborales (Masterman & Viscusi, 2020). Además, permite capturar valoraciones elevadas en grupos vulnerables que tienden a mostrar una alta disposición a pagar por seguridad, pese a no estar reflejada en sus condiciones laborales reales (Lanoie, Pedro, & Latour, 1995).

2.1.1.2 Métodos preferencias reveladas

La metodología de preferencias reveladas para la estimación del valor estadístico de la vida se obtiene a partir del análisis del comportamiento real de los individuos en contextos de mercado, lo que permite estimar el valor que el individuo le asigna a una reducción del riesgo de mortalidad (Viscusi, 1993). A diferencia de los métodos de preferencias declaradas, estos se basan en decisiones observables, como las elecciones laborales relacionadas a la exposición de los riesgos o el valor implícito en los precios de propiedades afectadas por factores ambientales (Bateman & Kling, 2020). Dichos métodos son útiles para la evaluación de políticas ambientales y de salud pública, ya que proporcionan estimaciones más precisas y menos sesgadas al depender de información empírica (Viscusi, 1993). Sin embargo, requiere de una elección precisa, tanto de las variables para evitar sesgos en los resultados como de su metodología aplicada, ya que es crucial capturar de manera correcta la relación entre sus variables (Bateman & Kling, 2020).

Una de las técnicas más utilizadas dentro de este enfoque es la de salarios hedónicos, que permite calcular la prima salarial asociada al riesgo, y con ello estimar el valor estadístico de la vida (Majumder & Madheswaran, 2018). Este enfoque considera a los empleos como un conjunto de

características, como condiciones laborales y los niveles de riesgo de lesiones. Los trabajadores deciden la cantidad de dinero que requieren como compensación para diferentes niveles de riesgo y las empresas deciden el monto que están dispuestas a ofrecer para la aceptación del riesgo (Kniesner, Viscusi, & Ziliak, 2010). Por lo tanto, el salario real representa una serie de precios hedónicos para diversos atributos del trabajo, incluyendo el riesgo de accidentes y otros precios para las características del trabajador (Shanmugam & Madheswaran, 2011).

El enfoque de salarios hedónicos se fundamenta en la teoría de los diferenciales compensatorios de salarios, propuesto inicialmente por Adam Smith en el año 1776 y posteriormente modernizada por Rosen (1974). La idea central radica en que los trabajadores requieren una compensación adicional para aceptar empleos más peligrosos (Biddle & Zarkin, 1988). Por lo tanto, según señalan Majumder & Madheswaran (2018), el diferencial salarial actúa como una prima de riesgo que refleja el valor económico asociado al peligro inherente en el trabajo. Los mismos autores, advierten que los trabajadores toman decisiones laborales considerando tanto su ingreso, como el riesgo asociado a su empleo, partiendo de la premisa de que los trabajadores tienen preferencias por salarios más altos, por lo que su utilidad marginal es positiva, como se puede observar en la Ecuación (1).

$$U'(w) > 0 \;\; y \;\; v'(w) > 0 \tag{1}$$

Donde $U'(w)$ representa la utilidad marginal respecto al ingreso w en el estado de buena salud y $v'(w)$, por el contrario, representa la utilidad marginal respecto al ingreso w en el estado de mala salud (Majumder & Madheswaran, 2018). Además, se asume que el estado de buena salud proporciona mayor utilidad que el estado de mala salud, como se puede observar en la Ecuación (2).

$$u(w) > v(w) \tag{2}$$

La utilidad esperada del trabajador se expresa en la Ecuación (3).

$$Z = (1 - p) \times u(w) + p \times v(w) \tag{3}$$

Donde p representa la probabilidad de que ocurra un evento adverso, como una lesión fatal, $u(w)$ es la utilidad percibida cuando el trabajador está en buen estado de salud, $v(w)$ corresponde a la utilidad percibida en caso de sufrir una lesión y Z denota la utilidad esperada (Majumder & Madheswaran, 2018). En la Ecuación (3), la utilidad esperada es el resultado de ponderar la utilidad en ambos estados

(sano y lesionado) por la probabilidad de ocurrencia de cada uno. El trabajador busca maximizar su utilidad esperada considerando el riesgo laboral. Para determinar la elección óptima, se resuelve la condición de primer orden derivada de la maximización de la utilidad esperada, como se muestra en la Ecuación (4).

$$\frac{dw}{dp} = -\frac{Zp}{Zw} = \frac{u(w) - v(w)}{(1-p)u'(w) + pv'(w)} > 0 \tag{4}$$

La Ecuación (4) muestra que el cambio en el salario a causa de un incremento en el riesgo es equivalente a la diferencia de las utilidades en cada estado (buena salud y mala salud) dividida por la suma de utilidades marginales en cada estado ponderada por su probabilidad de ocurrencia. Prácticamente, indica que un mayor riesgo debe ser compensado por un aumento salarial para mantener constante la utilidad esperada del trabajador (Majumder & Madheswaran, 2018).

El enfoque de salarios hedónicos utiliza la función de ganancias de Mincer (1974) adaptada para considerar riesgos laborales. Se considera que el salario de un trabajador depende tanto de su productividad como de los atributos no pecuniarios del trabajo, que se expresa en la Ecuación (5).

$$\ln w_i = \alpha + \beta_1 p_i + \beta_2 q_i + \sum_{k=1}^{K} \gamma_k X_{ki} + \epsilon_i \tag{5}$$

La Ecuación (5) genera el intercambio de salario-riesgo de los trabajadores, el cual es utilizado para estimar el valor estadístico de la vida. Donde, w_i representa el salario por hora para el trabajador i, α es el término constante que refleja el valor base del salario, p_i es el riesgo de lesión falta asociado al trabajador i, q_i representa el riesgo de lesión no fatal relacionado con el trabajo, X_{ki} es un vector que agrupa las k características individuales y laborales del trabajador i, como la edad, nivel educativo, experiencia laboral, horas de trabajo y condiciones laborales, mientras que γ_k captura el efecto de cada característica individual, ϵ_i es el término de error aleatorio, y por último, β_1 y β_2 son los coeficientes que capturan los efectos el efecto de los riesgos laborales sobre el salario del trabajador (Majumder & Madheswaran, 2018).

Una vez obtenidos los coeficientes a partir de la regresión, el valor estadístico de la vida se estima a partir de la formula en la Ecuación (6).

$$VSL = \hat{\beta}_1 \times \bar{W} \times 2.160 \times 10.000 \tag{6}$$

En la Ecuación (6), $\hat{\beta}_1$ corresponde al coeficiente estimado del riesgo de lesión fatal, $\overline{W}$ representa el salario promedio por hora, 2.160 corresponde a una jornada laboral completa anual estandarizada (aunque este valor puede variar según el país o el tipo de empleo) y el valor de 10.000 estandariza el riesgo de muerte por cada 10.000 trabajadores (Majumder & Madheswaran, 2018).

2.1.2 El valor estadístico de la vida en Chile

El valor estadístico de la vida varía según el ingreso del país (Miller, 2000). Estudios internacionales han demostrado que la elasticidad del ingreso del valor estadístico de la vida es mayor en países con menos ingresos, mientras que, en países con ingresos más altos, como Estados Unidos, cuya elasticidad se estima entre 0,4 y 0,6 (Hammitt & Robinson, 2011). Esto implica que en países con economías más desarrolladas, el valor es mayor que en países en vías de desarrollo (Viscusi & Aldy, 2003). El valor en economías de altos ingresos es de aproximadamente $6.4 millones, mientras que en economías de bajos ingresos puede ser de $107,000 (Viscusi & Masterman, 2017). El estudio de Miller (2000) analiza como el valor estadístico de la vida varía a partir de estudios realizados en 13 países. Se estima que el valor es aproximadamente 120 veces el PIB per cápita y aumenta casi linealmente con el ingreso, con una elasticidad cercana a 1 en los países desarrollados.

En Chile, Parada-Contzen, Riquelme-Won y Vasquez-Lavin (2012) estimaron el valor estadístico de la vida utilizando el método de salarios hedónicos con datos de la Encuesta de Caracterización Nacional (2006), complementados con estadísticas de riesgo laboral de la Asociación Chilena de Seguridad (ACHS). Los resultados entregaron un valor estadístico de la vida de US$ 4,63 millones sin corrección por endogeneidad, sin embargo, dicha cifra aumento a US$ 12,83 millones al aplicar variables instrumentales, reflejando el impacto de la endogeneidad en la estimación. Ajustando estos valores a dólares constantes de marzo de 2025, se obtienen aproximadamente US$ 7,3 millones y US$ 20,4 millones, respectivamente (U.S. Bureau of Labor Statistics, 2025). El estudio de Mardones y Riquelme (2018), ofrece una estimación actualizada del valor estadístico de la vida en Chile, también utilizando método de salarios hedónicos y estadísticas de las ACHS, pero con datos de la CASEN 2013. A diferencia de lo realizado por Parada-Contzen et al., los autores aplican un modelo Heckit 2SLS con instrumentos validados, logrando corregir de forma más rigurosa los problemas de endogeneidad y sesgo de selección. El valor estimado del valor estadístico de la vida fue de US$ 3,73 millones, que al ajustarse a dólares de marzo de 2025 equivale a aproximadamente US$ 5,1 millones (U.S. Bureau of Labor Statistics, 2025). Este valor más bajo se atribuye a una mejora en la calidad de los datos y en la especificación del modelo.

Además de los enfoques basados en salarios hedónicos, en Chile también se han utilizados métodos de preferencias declaradas para la estimación del valor estadístico de la vida. El estudio de Rizzi y Ortuzar (2003) evaluó la disposición a pagar por la reducción de riesgos de accidentes fatales en autopistas, a través de un instrumento de encuesta en el cual las personas debían elegir rutas con diferentes niveles de riesgo, tiempo de viaje y costo. Los resultados entregaron valores en el valor estadístico de la vida entre US$ 650.000 y US$ 1.300.000 para individuos con comportamiento lexicográfico (que priorizan absolutamente un atributo, en este caso exclusivamente la seguridad), entre US$ 350.000 y US$ 460.000 para aquellos con comportamiento compensatorio (que consideran un equilibrio entre varios atributos como tiempo, costo y riesgo) y entre US$ 149,000 y US$ 285,000 para los individuos quienes perciben los accidentes fatales no solo como la muerte en sí, si no, como un indicador general de seguridad vial. A partir de lo anterior, se destaca la importancia del factor comportamental en la valorización económica del riesgo. Ajustando estos valores a dólares de marzo de 2025, se obtienen rangos aproximados de US$ 1.130.000 a US$ 2.260.000 para el comportamiento lexicográfico, US$ 608.000 a US$ 800.000 para el comportamiento compensatorio y US$ 259.000 a US$ 495.000 para la percepción general de seguridad vial (U.S. Bureau of Labor Statistics, 2025). A partir de lo anterior, se destaca la importancia del factor comportamental en la valorización económica del riesgo.

Otro enfoque utilizado para la estimación del valor estadístico de la vida en Chile, es el de preferencias declaradas aplicado en el estudio de Cifuentes, Prieto y Escobari (2008), cuyo objetivo era evaluar la disposición a pagar para reducir el riesgo de mortalidad en diferentes rangos temporales. Se estimó el valor estadístico de la vida a través de encuestas aplicadas a estudiantes de Ingeniería de la Pontificia Universidad Católica de Chile para diferentes escenarios de reducción de riesgo (1 en 1.000 y 5 en 1.000). Los resultados arrojaron un valor de US$ 413.703 para una reducción de 1 en 1.000 en los próximos 10 años, mientras que para una reducción de 5 en 1.000 el valor fue de US$ 310.277. En el caso de una reducción de 5 en 1.000 en personas de 70 a 80 años, el valor estimado disminuyó considerablemente a US$ 13.043. Ajustando estos valores a dólares de marzo de 2025, se obtienen aproximadamente US$ 614.500, US$ 460.900 y US$ 19.400, respectivamente (U.S. Bureau of Labor Statistics, 2025).

Todos los valores del valor estadístico de la vida en estudios chilenos han sido ajustados a dólares constantes de marzo de 2025 son presentados en las Tabla 1. Se utiliza el índice de precios al consumidor (CPI) de Estados Unidos, que alcanzó un nivel de 319,799 en dicho mes, con el objetivo

de asegurar la comparabilidad entre estudios con diferentes años base (U.S. Bureau of Labor Statistics, 2025).

Tabla 1: Valor estadístico de la vida ajustado a dólares de marzo de 2025.

Estudio	Método	Año del estudio	CPI año del estudio	Valor original (US$)	Valor ajustado a 2025 (US$)	Comentarios sobre especificación y ajuste del modelo
Parada-Contzen et al. (2012)	Salarios hedónicos (OLS)	2006	201,6	4.630.000	7.344.590,13	Sin corrección por endogeneidad
Parada-Contzen et al. (2012)	Salarios hedónicos (2SLS)	2006	201,6	12.830.000	20.352.287,55	Con corrección por endogeneidad
Mardones y Riquelme (2018)	Salarios hedónicos (Heckit 2SLS)	2013	232,957	3.730.000	5.120.474,04	Incluye corrección por sesgo de selección y endogeneidad
Rizzi y Ortúzar (2003) "Lexicográfico"	Preferencias declaradas	2003	184,000	1.300.000	2.259.449,46	Individuos que priorizan exclusivamente la seguridad
Rizzi y Ortúzar (2003) "Compensatorio"	Preferencias declaradas	2003	184,000	460.000	799.497,50	Individuos que equilibran tiempo, costo y riesgo
Rizzi y Ortúzar (2003) "Seguridad vial"	Preferencias declaradas	2003	184,000	285.000	495.340,84	Individuos que interpretan riesgo como percepción general
Cifuentes et al. (1 en 1.000, 10 años)	Preferencias declaradas	2008	215,300	413.703	614.499,79	Riesgo reducido de 1 en 1.000 en 10 años
Cifuentes et al. (5 en 1.000)	Preferencias declaradas	2008	215,300	310.277	460.874,47	Riesgo reducido de 5 en 1.000
Cifuentes et al. (5 en 1.000, personas de 70–80 años)	Preferencias declaradas	2008	215,300	13.043	19.373,61	Riesgo reducido de 5 en 1.000 en adultos mayores

* CPI año 2025: 319,779

Fuente: Elaboración propia a partir de los valores originales reportados en cada estudio y el Índice de Precios al Consumidor (CPI) de Estados Unidos (U.S. Bureau of Labor Statistics, 2025)

2.2 Regresión simbólica

El aprendizaje automático es una rama clave de la inteligencia artificial dedicada al desarrollo de técnicas que permiten a los sistemas informáticos identificar patrones y generar predicciones a partir de datos (Sarker, 2021). Las técnicas más comunes incluyen regresiones lineales, arboles de

decisiones, redes neuronales y modelos probabilísticos, entre otros, destacando la capacidad de adaptación y aprendizaje (Shinde & Shah, 2018). Sin embargo, como señala Khosravi et al. (2023), las relaciones entre variables suelen presentar relaciones no lineales o patrones complejos que métodos tradicionales no siempre pueden capturar adecuadamente. En este contexto, surge la regresión simbólica, una técnica emergente dentro del aprendizaje automático que permite inferir expresiones matemáticas interpretables a partir de datos (Makke & Chawla, 2024). En el mismo estudio, los autores destacan que, a diferencia de los modelos de caja negra, como las redes neuronales, los modelos simbólicos permiten comprender de manera clara como las entradas se traducen en salidas, lo que la posiciona como una técnica muy útil en disciplinas científicas que requieren interpretabilidad.

2.2.1 Metodologías de regresión simbólica

Generalmente, el proceso de regresión simbólica emplea algoritmos evolutivos, destacando particularmente la programación genética (Zeng et al., 2023). Dicho proceso, empieza con la generación de una población inicial de expresiones matemáticas creadas aleatoriamente (Mundhenk et al., 2021). Posteriormente, cada una de estas expresiones es evaluada mediante una función específica de aptitud que determina su capacidad para ajustar adecuadamente los datos disponibles (Kronberger, Burlacu, Kommenda, Winkler, & Affenzeller, 2024). Luego, según Mundhenk et al. (2021), se seleccionan aquellas que ofrecen el mejor ajuste según la evaluación previa. Según el mismo autor, estas expresiones se convierten en "padres" que darán origen a la siguiente generación mediante la aplicación de operaciones genéticas, incluyendo mutaciones y cruzamientos, donde se realizan modificaciones aleatorias sobre las expresiones ya existentes e intercambios de subexpresiones entre diferentes soluciones seleccionadas. El ciclo iterativo descrito sigue repitiéndose, con la evaluación, selección y aplicación de operaciones genéticas hasta alcanzar un criterio específico para finalizar el proceso, como podría ser un número máximo de generaciones o un nivel satisfactorio de ajuste de las expresiones grandes (Kronberger, Burlacu, Kommenda, Winkler, & Affenzeller, 2024).

Además de la programación genética, existen otras técnicas alternativas para llevar a cabo la regresión simbólica, que igualmente buscan descubrir relacionales matemáticas explícitas a partir de los datos (Udrescu & Tegmark, 2020). Algunos métodos, como el propuesto por Rad, Feng, & Iba (2018), combinan funciones generadas mediante programación genética con técnicas de optimización generalizada, como máquinas de vectores de relevancia. Estas seleccionan de manera automática las

funciones más prometedoras, penalizando las menos relevantes, buscando así, un balance óptimo entre simplicidad estructural y predicción predictiva. Otra alternativa son los enfoques probabilísticos bayesianos, como el propuesto por Jin, Fu, Kang, Guo, & Guo (2019), en el cual las expresiones simbólicas se representan mediante árboles y se explora el espacio de modelos posibles utilizando algoritmos de muestreo basado en cadenas de Markov Monte Carlo que permiten muestrear modelos desde su distribución posterior.

Además, recientemente Schnur & Chawla (2023), aplicaron métodos avanzados basados en redes neuronales profundas y aprendizaje por refuerzo, optimizando criterios como la precisión predictiva y la simplicidad del modelo mediante recompensas y penalizaciones asociadas a cada decisión durante el proceso de aprendizaje. Según los mismos autores, estas alternativas ofrecen ventajas específicas según el contexto, la complejidad de datos y los requisitos particulares del análisis.

El estudio realizado por La Cava et al. (2021) compara los diversos métodos contemporáneos de regresión simbólica mediante una plataforma abierta y reproducible denominada SRBench. Evaluando estos métodos tanto en conjuntos de datos del mundo real, como en problemas sintéticos con soluciones conocidas. Se concluye de la evaluación que los métodos basados en programación genética, especialmente Operon, FEAT y SBP-GP, ofrecen modelos precisos y sencillos en escenarios reales, superando a los tradicionales como bosques aleatorios o redes neuronales. En condiciones ideales con bajo ruido, enfoques específicos como AlFeyman destacan por su capacidad para recuperar ecuaciones exactas, aunque pueden volverse sensibles conforme aumenta el ruido en los datos.

Una ventaja esencial de la regresión simbólica es su habilidad para producir modelos altamente interpretables, que explícitamente revelan como se relacionan las variables estudiadas (Makke & Chawla, 2024). Al no estar restringida a formas funcionales específicas, la regresión simbólica puede descubrir relaciones complejas y no lineales, lo que es difícil de detectar mediante los métodos convencionales (Kronberger, Burlacu, Kommenda, Winkler, & Affenzeller, 2024). Además, gracias a la explicitud de las expresiones matemáticas generadas, esta técnica permite desglosar claramente como cada variable influye en los resultados del modelo, facilitando análisis como el de sensibilidad, modularidad y descomposición (Schnur & Chawla, 2023).

2.2.2 Aplicaciones de regresión simbólica

Existen diversas investigaciones que han implementado la regresión simbólica como método analítico para abordar problemas complejos. Recientemente, Abdellaoui y Mehrkanoon (2021) aplicaron dicha técnica mediante el enfoque de Equation Learner para la predicción de la velocidad del viento en tres ciudades danesas. Los autores demostraron que dicho modelo genera ecuaciones analíticas compactas, destacándose por su rapidez en la inferencia en comparación de modelos basados en redes neuronales convolucionales 3D. Los autores señalan que, aunque la precisión del modelo puede ser ligeramente menor, su ventaja radica en la transparencia y comprensibilidad de los resultados, lo que es relevante cuando se requiere comprender la relación directa entre variables. Además, Valsaraj, Thumba, Asokan, & Kumar. (2020) aplicaron regresión simbólica para la medir la velocidad del viento a alturas elevadas, por medio de la extrapolación vertical su velocidad en diferentes ubicaciones en Kerela, India. Por medio de la herramienta Eureqa, los mismos autores generaron funciones analíticas capaces de estimar la velocidad del viento a alturas mayores partiendo de mediciones en torres meteorológicas más bajas. El resultado de dicho estudio mostró una reducción significativa del error medio en comparación con métodos empíricos tradicionales como la ley de la potencia, alcanzando una mejora máxima del 61,04%. Aunque la precisión de estas funciones puede disminuir levemente en ciertos escenarios, su principal ventaja radica en la capacidad de producir modelos más interpretables y efectivos, incluso cuando se disponde de conjuntos de datos limitados (Wilstrup & Kasak, 2021).

2.3 Aplicación de regresión simbólica para la estimación del valor estadístico de la vida

Como se mencionó anteriormente, las metodologías para la estimación del valor estadístico de la vida presentan diversas limitaciones que afectan la exactitud de las estimaciones (Parada-Contzen, Riquelme-Won, & Vasquez-Lavin, 2012). Algunos de los problemas más relevantes en la estimación del valor estadístico de la vida se relacionan con el sesgo por especificación incorrecta (misspecification bias). Esto ocurre cuando el modelo utilizado no representa adecuadamente la relación real entre las variables, generando estimaciones sesgadas y, por tanto, posibles conclusiones erróneas respecto a los efectos causales (Blackwell & Olson, 2022). Sin embargo, la regresión simbólica se presenta como alternativa para superar dicha limitación presente, ya que permite identificar estructuras no lineales y descubrir nuevas relaciones causales (Makke & Chawla, 2024). Además, su habilidad para producir modelos altamente interpretables permite comprender como las variables se relacionan entre sí, aportando así mayor robustez y precisión a las estimaciones del valor estadístico de la vida (Schnur & Chawla, 2023).

En la Tabla 2 se presenta un resumen de los principales estudios revisados, detallando sus objetivos, metodologías aplicadas, países estudiados, variables dependientes utilizadas y los principales hallazgos en relación con la estimación del valor estadístico de la vida, incluyendo tanto enfoques tradicionales como métodos emergentes de regresión simbólica.

Tabla 2: Resumen de revisión de literatura.

Autores (año)	Objetivo	Método de estimación	Variable Dependiente	País	Hallazgo
Kniesner y Viscusi (2019)	Revisar conceptos, métodos y aplicaciones del Valor de una Vida Estadística	Preferencias reveladas y declaradas (laboral y encuestas)	Valor de una Vida Estadística	EE. UU. e internacional	Valor estadístico de la vida en EE. UU. cercano a $10 millones (2017), menor en otros países por diferencias de ingreso. Preferencia por datos laborales (CFOI).
Ashenfelter (2006)	Analizar problemas teóricos y empíricos en la estimación del Valor de la Vida Estadística .	Preferencias reveladas (cambios en límites de velocidad)	Valor de una Vida Estadística	EE. UU.	Estima el valor estadístico de la vida entre 1,6 y 6 millones de dólares
Rosen (1974)	Desarrollar un modelo teórico de precios hedónicos en mercados competitivos con diferenciación de productos.	Precios hedónicos (preferencias reveladas)	Precio implícito de atributos del producto	EE. UU.	Propone que los precios hedónicos son resultados del equilibrio espacial, donde atributos del producto determinan las decisiones de consumidores y productores.
Majumder y Madheswaran (2018)	Estimar el Valor de la Vida Estadística en trabajadores industriales de India usando diferencias salariales compensatorias.	Precios hedónicos (Preferencias reveladas)	Valor de una Vida Estadística	India	El valor estadístico de la vida estimado es de 44,69 millones INR (0,64 millones USD). Se concluye que los trabajadores reciben una compensación salarial por los riesgos laborales asumidos.
Viscusi y Masterman (2017)	Estimar elasticidades ingreso y calcular valor estadístico de la vida para distintos países, usando estimaciones desde EE.UU.	Metaanálisis (preferencias reveladas)	Valor de una Vida Estadística	Internacional (189 países)	La elasticidad ingreso del valor estadístico de la vida varía entre 0,5-0,7 en EE. UU. y cercana a 1,0 en otros países. Sugieren ajustar el valor estadístico de la vida por ingresos para evitar sesgos al extrapolar internacionalmente

(continuación) Tabla 2: Resumen de revisión de literatura.

Autores (año)	Objetivo	Método de estimación	Variable Dependiente	País	Hallazgo
Miller (2000)	Analizar diferencias del Valor de la Vida Estadística entre países y proponer modelos de transferencia para estimar valor estadístico de la vida internacionalmente.	Metaanálisis (Varios métodos)	Valor de una Vida Estadística	13 países estudiados	El valor estadístico de la vida promedio es alrededor de 120 veces el PIB per cápita; elasticidad ingreso del valor estadístico de la vida entre 0.85 y 1.0 según nivel de agregación.
Parada-Contzen, Riquelme-Won y Vásquez-Lavín (2013)	Estimar el valor estadístico de la vida y VSI para trabajadores chilenos utilizando diferencias salariales compensatorias.	Precios hedónicos (Preferencias reveladas) corregidos por endogeneidad	Valor de una Vida Estadística y lesión estadística	Chile	Valor estadístico de la vida estimado de USD 12,8 millones tras corregir endogeneidad, siendo considerablemente mayor que las estimaciones previas indirectas para Chile.
Mardones y Riquelme (2018)	Estimar el valor estadístico de la vida para Chile y extrapolar resultados a países de América Latina utilizando diferencias salariales compensatorias.	Precios hedónicos (Preferencias reveladas), corrección por sesgo de selección y endogeneidad	Valor de una Vida Estadística	Chile y América Latina	Estiman un valor estadístico de la vida de USD 3,7 millones para Chile.
Rizzi y Ortúzar (2003)	Estimar el valor estadístico de la vida en carreteras interurbanas en Chile mediante preferencias declaradas.	Preferencias declaradas (encuestas)	Valor de una Vida Estadística	Chile	El valor estadístico de la vida estimado oscila entre USD 350.000 y 460.000 al considerar respuestas compensatorias; al incluir respuestas lexicográficas aumenta significativamente.
Cifuentes, Prieto y Escobari (2000)	Estimar el valor estadístico de la vida en Chile considerando reducciones de riesgo de mortalidad presente y futura.	Valoración contingente	Valor de una Vida Estadística	Chile	El valor estadístico de la vida estimado es de aproximadamente USD 413.700 para riesgos presentes y USD 13.043 para riesgos futuros.

(continuación) Tabla 2: Resumen de revisión de literatura.

Autores (año)	Objetivo	Método de estimación	Variable Dependiente	País	Hallazgo
Schnur y Chawla (2023)	Revisar el uso de la regresión simbólica para la fusión de información en el contexto de la salud humana.	Regresión simbólica (QLattice)	Porcentaje de grasa corporal total	Estados Unidos	La regresión simbólica permite generar modelos matemáticos interpretables para estimar el porcentaje de grasa corporal a partir de medidas antropométricas, mejorando el rendimiento predictivo respecto a métodos lineales tradicionales.
La Cava et al. (2024)	Evaluar el desempeño de métodos contemporáneos de regresión simbólica (SR) mediante un benchmark abierto y reproducible.	Regresión simbólica (varios métodos)	Precisión y simplicidad del modelo	Internacional	Los métodos basados en GP, como Operon, superan a otros enfoques en problemas de regresión reales, mientras que AIFeynman es más efectivo en problemas físicos con poco ruido.
Abdellaoui y Mehrkanoon (2021)	Aplicar la regresión simbólica para predecir la velocidad del viento en ciudades danesas.	Regresión simbólica (EQL) y modelos basados en redes neuronales	Velocidad del viento	Dinamarca	El modelo EQL logra predicciones precisas con menor tiempo de inferencia en comparación con redes neuronales 3D-CNN. La regresión simbólica ofrece mayor interpretabilidad.
Valsaraj et al. (2020)	Proponer un método mejorado basado en regresión simbólica para extrapolar la velocidad del viento a mayores altitudes.	Regresión simbólica vs. Ley de Potencia	Velocidad del viento a mayores altitudes	India	La regresión simbólica reduce el RMSE hasta en un 61% respecto a la Ley de Potencia y ofrece mayor precisión en predicciones a corto plazo.

Fuente: Elaboración propia a partir de lo reportado en cada estudio

3. Metodología

El presente capítulo describe las metodologías aplicadas para estimar el valor estadístico de la vida en Chile, combinando el enfoque de salarios hedónicos con técnicas de regresión lineal y regresión simbólica. Se detallan los fundamentos del modelo empírico, el procedimiento de estimación mediante mínimos cuadrados ordinarios, la implementación de la regresión simbólica y las métricas estadísticas utilizadas para evaluar la calidad del ajuste del modelo.

3.1 Modelo empírico

Para la estimación del valor estadístico de la vida, se utiliza el método de salarios hedónicos ya que permite analizar la relación del salario con diversas características laborales, demográficas y de exposición de riesgo (Majumder & Madheswaran, 2018). Dicho modelo empieza del supuesto de que el salario no se ve reflejado exclusivamente por su productividad, sino también por diversos aspectos del entorno laboral, incluyendo los niveles de riesgo de mortalidad asociados (Shanmugam & Madheswaran, 2011). Como enfoque alternativo, la regresión simbólica es implementada sobre los residuos del modelo lineal. Tras calcular las diferencias entre los salarios observados y los predichos por la regresión lineal, se ajusta una ecuación simbólica a esos residuos para capturar relacionales no lineales que el modelo lineal no detecta, siguiendo un procedimiento análogo al de descomposición simbólica escalonada descrito por Ricketts (2013).

Esta memoria de título estima el valor estadístico de la vida por dos métodos distintos. El primero corresponde al método de mínimos cuadrados ordinarios, basado en la propuesta desarrollada por Montgomery & Runger (2004). Posteriormente, se aplica la metodología de regresión simbólica para la identificación de una expresión funcional interpretable que relacione el salario con las variables explicativas, sin asumir una forma estructural previa y explorando combinaciones de operaciones matemáticas mediante programación genética (Wang et al. 2022).

3.2 Mínimos cuadrados ordinarios

La regresión lineal permite determinar de forma estadística la relación de variables (Schneider, Hommel, & Blettner, 2010). Los componentes para considerar son: una variable dependiente y_t, variables independientes x, coeficientes de la regresión β y la constante de error aleatorio ε. Siendo β, el valor que cambia Y por un cambio en una unidad de x_i, manteniendo a todas las variables constantes (Montgomery & Runger, 2004). La Ecuación (7) muestra la forma general de la regresión lineal.

$$y_t = \beta_0 + \sum_{i=1}^{N} \beta_i x_i + \varepsilon \tag{7}$$

Si N solo tiene un elemento se considera una regresión lineal simple, en caso contrario es una regresión lineal múltiple, es decir hay más de una variable independiente. Para determinar el valor de los coeficientes de la regresión, se utiliza el método de los mínimos cuadrados ordinarios, cuyo objetivo es minimizar la suma de los cuadrados de los errores. Una de las consideraciones para utilizar este método es tener t observaciones, y k coeficientes de la regresión, siendo $t > k$.

Las variables dependientes tienen la siguiente notación: x_i^t, donde $i = 1, \dots, N$ representa cada variable en $t = 1, \dots, T$ observaciones. De esta forma la Ecuación (7) se desglosa para cada observación, como se muestra en la Ecuación (8).

$$y_t = \beta_0 + \sum_{i=1}^{N} \beta_i x_i^t + \varepsilon_t \quad t = 1, \dots, T \tag{8}$$

Luego, se aplica la minimización de los errores al cuadrado, según la Ecuación (9).

$$MCO = \sum_{t=1}^{T} \left(y_t - \beta_0 + \sum_{i=1}^{N} \beta_i x_i^t \right)^2 \tag{9}$$

3.3 Regresión simbólica

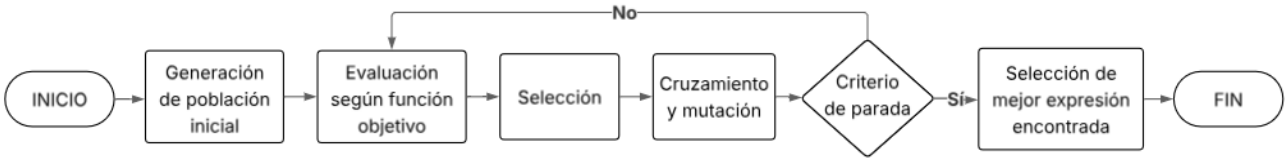
La regresión simbólica se inicia de un conjunto de datos $D = \{x_i, y_i\}_{i=1}^{n}$, donde $x_i \in \mathbb{R}^d$ y $y_i \in \mathbb{R}^d$. El objetivo es encontrar una función simbólica $f \in F$ que minimice la raíz del error cuadrático medio (Kronberger, Burlacu, Kommenda, Winkler, & Affenzeller, 2024).

3.3.1 Regresión simbólica por programación genética

La regresión simbólica basada en programación genética comienza con la generación aleatoria de una población inicial de expresiones matemáticas, construidas combinando operadores aritméticos y funciones establecidas. Cada expresión es evaluada mediante la función objetivo (minimización de la raíz del error cuadrático medio) y las expresiones con mejor desempeño son seleccionadas para producir una nueva generación. Se aplican operadores genéticos como mutación y cruzamiento, en función de probabilidades relativas. Además, se incorpora un parámetro que actúa como límite estructural para la regresión simbólica, denominado tamaño máximo, que restringe el número total de operadores de cada expresión para evitar que las expresiones crezcan sin aportar valor, mejorando

tanto la interpretabilidad como la eficiencia de la evaluación. Por último, el proceso se detiene cuando se alcanza un número máximo de generaciones (Kronberger, Burlacu, Kommenda, Winkler, & Affenzeller, 2024), como se puede observar en la Figura 1.

Figura 1 : Diagrama de flujo de regresión simbólica por programación genética.



Fuente: Elaboración propia

3.4 Pruebas y métricas estadísticas

3.4.1 Prueba F

La prueba F se utiliza para determinar si el modelo en su conjunto es significativo, específicamente, para verificar si al menos una de las variables explicativas se asocia significativamente con la variable dependiente. Bajo la hipótesis nula, se asume que todos los coeficientes del modelo son iguales a cero, como se puede observar en la Ecuación (10). Luego, la hipótesis alternativa plantea que existe al menos un coeficiente distinto de cero, como se puede apreciar en la Ecuación (11). La estadística F compara la proporción de la varianza explicada por el modelo con la varianza no explicada. En la práctica, si $F > 2{,}5$, se rechaza H_0, concluyendo que al menos existe un cociente distinto de cero.

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_j = 0, \text{para algún } j \tag{10}$$

$$H_1 = \beta_j \neq 0, \text{para algún } j \tag{11}$$

3.4.2 Coeficiente de determinación (R^2)

El coeficiente de determinación $\boldsymbol{R^2}$ mide que fracción de la variabilidad total de la variable dependiente logra explicar el modelo. Sus valores fluctúan desde 0, que sugiere que el modelo no aporta ninguna capacidad explicativa y 1, que indica que el modelo predice perfectamente los datos.

3.4.3 Coeficiente de determinación ajustado (R^2 ajustado)

El coeficiente de determinación ajustado $\boldsymbol{R^2}$ corrige el $\boldsymbol{R^2}$ estándar por el número de variables independientes y la cantidad de observaciones. De este modo, penaliza la inclusión de variable que no aportan información relevante, ofreciendo una medida más realista de la bondad de ajuste cuando se comparan modelo con distinto números de regresores.

3.4.4 Raíz del error cuadrático medio (RMSE)

La raíz del error cuadrático medio o RMSE es una métrica que indica el promedio de la magnitud de los errores de predicción, expresado en las mismas unidades que la variable dependiente. Se obtiene a partir de la raíz cuadrada del promedio de los cuadrados de las diferencias entre los valores observados y los predichos, lo que permite interpretar el ajuste del modelo, según la Ecuación (12).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2} \tag{12}$$

3.4.5 Prueba t

La prueba t evalúa que un coeficiente particular $\beta_j$sea igual a cero, como se puede observar en la Ecuación (13). Si H_0 se cumple, significa que la variable explicativa β_j no contribuye en la predicción del modelo. En la práctica, si el estadístico t es mayor que 2 en valor absoluto, se rechaza H_0, indicando que β_j es estadísticamente distinto de cero, como se puede apreciar en la Ecuación (14).

$$H_0: \beta_j = 0, \text{ para algún } j \tag{13}$$

$$H_1: \beta_j \neq 0, \text{ para algún } j \tag{14}$$

3.4.6 Nivel de significación

El valor de significación p indica la probabilidad de que el resultado observado se deba al azar, bajo la suposición de que H_0 es verdadera. Si el valor p es menor que el nivel de significación preestablecido (por defecto, 0,05) se considera que el resultado es estadísticamente significativo y se rechaza la hipótesis nula.

4. Datos

El presente capítulo describe las fuentes, características y tratamiento de los datos utilizados en la estimación del valor estadístico de la vida. Se detallan las bases de datos provenientes de la SUSESO y la encuesta CASEN, así como la construcción de la muestra y las principales variables consideradas. Además, se presentan descriptivas que permiten caracterizar la población analizada.

4.1 Fuentes de datos

Para este estudio se utilizarán dos fuentes de datos en particular. La primera es el Boletín Estadístico de Seguridad Social. La cual se encuentra disponible en el sitio web de la Superintendencia de Seguridad Social (SUSESO). Dicho boletín proporciona datos anuales sobre variables vinculadas a los sistemas de prevención y asistencia que fiscaliza la SUSESO y consta de diez capítulos y 163 tablas, de los cuales se utilizarán exclusivamente los del primer capítulo "Régimen de Accidentes del Trabajo y Enfermedades Profesionales" (Tablas N°s 1 al 63). Este presenta estadísticas detalladas del régimen de accidentes laborales (Ley N°16.774), incluyendo accidentes, trabajadores protegidos, enfermedades profesionales y mortalidad laboral en el año 2022, contemplando un total de 7.178.925 trabajadores. Además, informa sobre pensiones, subsidios, indemnizaciones y bonificaciones asociadas. Los datos provienen directamente de registros administrativos, diferenciándose así de los boletines mensuales y presenta información de tres mutualidades: Asociación Chilena de Seguridad (ACHS), Mutual de Seguridad de la Cámara Chilena de la Construcción (MUSEG), y por último del Instituto de Seguridad del Trabajo (IST). En la presente investigación solo se consideran los datos de la ACHS para poder complementarla con otras fuentes. Del total de trabajadores, 2.826.657 pertenecen a la ACHS, siendo 1.586.278 hombres y 1.240.378 mujeres (Superintendencia de Seguridad Social, 2024).

La segunda fuente de datos es la Encuesta de Caracterización Socioeconómica Nacional (CASEN), cuyo propósito es entregar un diagnóstico de la situación socioeconómica de los hogares en Chile. Esta estima los niveles de pobreza, desigualdad de ingresos y cobertura de políticas sociales, con el objetivo de identificar necesidades prioritarias en grupos específicos y orientar la gestión de programas sociales en el país. La encuesta CASEN se realizó entre el 1 de noviembre de 2022 y el 2 de febrero de 2023, completando una muestra de 72.056 hogares, que ocupan 70.751 viviendas en 335 comunas de las 16 regiones del país, y obteniendo información respecto de 202.231 personas (Ministerio de Desarrollo Social y Familia, 2023).

4.2 Muestra de datos

Para la estimación del modelo de salarios hedónicos se diseñó una muestra que incorpora tanto características laborales, como sociodemográficas de los entrevistados. En primer lugar, se incluyeron variables como el ingreso salarial y el tipo de trabajo. Adicionalmente, siguiendo las recomendaciones de Kniesner & Leeth (2010), se integraron variables utilizadas en la literatura como el estado civil, la ubicación geográfica y la edad. Específicamente, se crea una variable que recoge el cuadrado de la edad del encuestado, para capturar efectos no lineales del ciclo de vida. Adicionalmente, se incorpora una variable que identifica si la pareja del entrevistado se encuentra actualmente trabajando. Además, se generaron variables binarias para cada una de las regiones de residencia, con el fin de controlar las diferencias territoriales en las compensaciones salariales. En base a la literatura, se incluye una variable que identifica a aquellos hogares con mujeres mayores a 15 años en los cuales habitan niños menores a 6 años. Esta inclusión se fundamenta en estudios que evidencian el impacto de los hijos pequeños en las trayectorias laborales femeninas. Según Cukrowska-Torzewska & Matysiak (2020), "las madres tienden a percibir salarios más bajos porque eligen empleos que, por término medio, pagan menos, pero son más compatibles con el cuidado de los hijos". Por su parte, Looze (2017) sostiene que, "los niños en edad preescolar reducen la movilidad laboral de las mujeres, que está asociada al crecimiento salarial".

Al analizar la relevancia del tamaño de la empresa donde se desempeña el trabajador como factor determinante, se opta por incluir una variable específica que captura el tamaño promedio de las empresas del sector al que pertenece el encuestado, permitiendo así evaluar empíricamente su influencia en el contexto analizado. Para incorporar la información del nivel educacional en el análisis, se crearon variables binarias adicionales, la cuales permiten una mejor captura y representación de estos datos. Estas variables abarcan desde aquellas personas quienes no tuvieron una educación formal hasta aquellas quienes completaron o están cursando la educación superior. La encuesta CASEN 2022 presenta dos variables específicas para registrar esta información, la cuales son las personas que no asisten a ningún establecimiento educacional y por el contrario, las que si asisten. Sin embargo, para efectos prácticos del estudio, se decide agrupar en una misma categoría tanto a las personas quienes alcanzaron dicho nivel educativo en el pasado, como quienes actualmente cursan dicho nivel. Por ejemplo, alguien que estudió solo hasta la educación básica y otra persona actualmente está en educación básica fueron agrupados en la misma categoría. De esta manera, se establecieron cinco categorías educativas en total.

Otro aspecto considerado en la construcción de la muestra es la cantidad de personas que conforman el hogar del entrevistado, ya que esta variable puede incidir en las decisiones laborales y en los ingresos disponibles. También, se incluye una variable que identifica si la persona cuenta con contrato de trabajo, lo que permite diferenciar entre empleo formal e informal. Además, se agregaron variables binarias que permiten identificar el tipo de trabajo en el cual se desempeña cada encuestado. Igualmente, se considera la distinción entre jornada laboral completa y parcial, debido a su influencia en los niveles de ingreso.

Finalmente, para capturar el riesgo inherente a las distintas ocupaciones, se recurre al uso de tasas de accidentes laborales. Ya que, no existe una variable directa que represente de manera precisa el riesgo en todos los casos. Se utiliza el número de accidentes fatales como una aproximación válida, utilizando como referencia lo planteado por Andersson (2005). Los datos fueron obtenidos desde los boletines informativos de la SUSESO, en los que se presentan tasas calculadas dividiendo el número de accidentes fatales entre el total de trabajadores protegidos por actividad económica. En base a lo anterior, se crearon dos variables que miden la tasa de accidentes fatales y no fatales para cada rubro (Tabla 3). Finalmente, se determina la variable dependiente del modelo como el logaritmo natural del salario mensual, aplicando la función logarítmica para mejorar la homocedasticidad.

Tabla 3: Datos accidentabilidad año 2017 y 2022.

Industria	Accidentes fatales		Tasa de accidentes fatales (1/10.000)		Accidentes no fatales		Tasa de accidentes no fatales (1/100)	
	2017	2022	2017	2022	2017	2022	2017	2022
Agricultura, ganadería, caza y silvicultura	36	10	1,053	0,544	14.368	6.239	4,20	3,39
Pesca	6	4	1,549	1,305	1.556	1.197	4,02	3,91
Explotación de minas y canteras	4	2	0,683	0,565	730	340	1,25	0,96
Industrias Manufactureras	23	4	0,455	0,149	23.757	9.830	4,70	3,66
Suministro de electricidad, gas y agua	1	1	0,333	0,550	485	224	1,61	1,23
Construcción	44	6	0,745	0,307	24.345	7.242	4,12	3,70
Comercio, reparación de vehículos y otros	8	4	0,112	0,104	26.381	8.953	3,68	2,33
Hoteles y restaurantes	6	0	0,275	0,000	10.235	4.201	4,70	3,39
Transporte, almacenamiento y comunicaciones	57	17	1,613	0,947	16.291	5.133	4,61	2,86
Intermediación financiera	3	0	0,168	0,000	2.093	699	1,17	0,91
Actividades inmobiliarias, empresariales y de alquiler	20	2	0,257	0,041	20.727	9.412	2,66	1,95
Administración pública y defensa	4	1	0,105	0,044	9.292	4.222	2,44	1,84
Enseñanza	0	1	0,000	0,040	7.049	4.184	1,85	1,69
Servicios sociales y de salud	2	0	0,108	0,000	4.263	2.391	2,3	1,75
Otras actividades de servicios comunitarios	7	2	0,281	0,110	7.398	4.353	2,97	2,39
Hogares privados con servicio doméstico	0	1	0,000	0,424	859	353	2,0	1,5
Organizaciones y órganos extraterritoriales	0	0	0,000	0,000	13	0	1,84	0,00

Fuente: Ministerio de Desarrollo Social y Familia (2023)

4.3 Descripción de datos

El modelo incluye variables clasificadas en siete categorías. En dimensión salarial, se utiliza la variable "Salario (en logaritmo)" como dependiente. En el ámbito empresarial, se incorporan variables sobre el tamaño de la empresa desde "1 a 5 trabajadores" hasta "201 a 499 trabajadores", la existencia de contrato laboral, distinguida entre "Trabajo con contrato firmado" y "Trabajo sin contrato", el tipo de jornada, diferenciando "Trabajo con jornada completo" y "Trabajo con jornada parcial", el "Tamaño promedio de empresas del sector" y el rol del encuestado, representado por las variables binarias "Trabajador independiente", "Empleado público", "Empleado doméstico", "Fuerzas armadas" y "Trabajo familiar no remunerado".

Respecto a la formación, se consideran los niveles educativos "Sin educación", "Educación básica", "Educación media" y "Educación técnica. En cuanto a las características individuales, se incluye las variables "Mujer", la "Edad (elevada al cuadrado)" y los "Ingresos no laborales".

A nivel de hogar, se utilizan las variables "Número de hijos menores de 6 años", "Número de personas en el hogar", además, la variable binaria que representa el estado civil, según si el encuestado esta "Casado", los "Años de escolaridad del cónyuge" y si "El cónyuge trabaja".

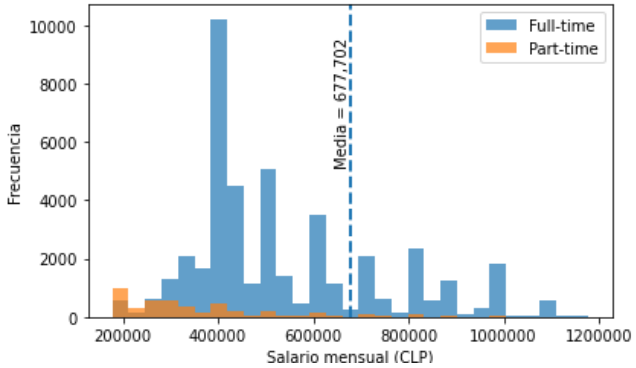
Las variables geográficas corresponden las regiones de Chile: "Tarapacá", "Antofagasta", "Atacama", "Coquimbo", "Valparaíso", "O'Higgins", "Maule", "Biobío", "Araucanía", "Los Lagos", "Aysén", "Magallanes", "Metropolitana", "Los Ríos", "Arica" y "Ñuble".

Tras aplicar criterios de depuración y ajuste en la base datos, el tamaño de muestra se redujo de 202.231 a 83.994 observaciones, dicha disminución se debe a la exclusión de menores de 15 años y de adultos mayores a 92, quienes, por lo general, no participan de manera activa en el mercado laboral o no perciben ingresos provenientes del trabajo.

La Figura 2, presenta el histograma que representa el salario, centrado en un rango de $500.000 alrededor de la media, la cual es aproximadamente $677.700, donde se aprecia una alta concentración de frecuencias bajo el valor promedio, evidenciando que, la mayoría de los salarios se encuentran cercanos a los $400.000. La Figura 3, presenta un gráfico de dispersión del salario de cada una de las observaciones con dicha variable declarada. Se puede apreciar que la mayor cantidad de los valores se agrupan bajo los $5.000.000, pero se pueden observar casos extremos que alcanzan los $15.000.000, incluso los $25.000.000, evidenciando la heterogeneidad salarial dentro de la muestra. Para profundizar en el comportamiento en los tramos menores, se segmentó la muestra en las

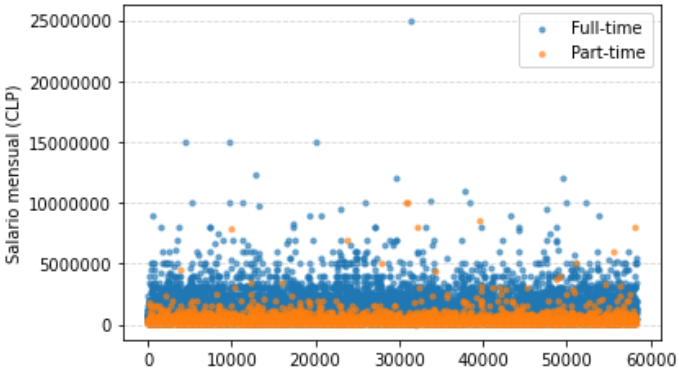
observaciones que trabajan con jornada completa (full-time) o parcial (part-time), lo que permitió analizar la existencia de sueldos por debajo del mínimo legal de 2022, el cual según el Gobierno de Chile (2025), llegaba a $350.000. En cambio, la Figura 4, se presenta el histograma de densidad de la variable “Salario (en logaritmo)” para el año 2022, donde se aprecia una alta concentración alrededor del promedio, adoptando una distribución similar a la normal, lo que facilita la interpretación y mejora el ajuste de modelos estadísticos al homogenizar la dispersión.

Figura 2: Distribución del salario (centrado en la media) año 2022.



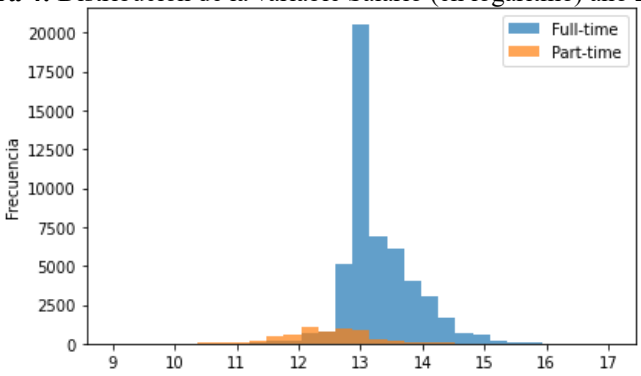
Fuente: Elaboración propia basada en datos de CASEN 2022

Figura 3: Dispersión del salario año 2022.



Fuente: Elaboración propia basada en datos de CASEN 2022

Figura 4: Distribución de la variable Salario (en logaritmo) año 2022.



Fuente: Elaboración propia basada en datos de CASEN 2022

La Tabla 4 presenta un resumen estadístico de las variables utilizadas. Se muestra el número de observaciones, el promedio, la desviación estándar y los valores máximos y mínimos para cada variable considerada.

Tabla 4: Descripción de variables año 2022.

Variable	Observaciones	Promedio	Desv. estándar	Mínimo	Máximo
Tasa de fatalidad	83.994	0,2250	0,2911	0,0000	1,131
Región					
Tarapacá	83.994	0,0443	0,2059	0,0000	1,0
Antofagasta	83.994	0,0442	0,2055	0,0000	1,0
Atacama	83.994	0,0435	0,2039	0,0000	1,0
Coquimbo	83.994	0,0381	0,1915	0,0000	1,0
Valparaíso	83.994	0,1015	0,3020	0,0000	1,0
O´Higgins	83.994	0,0682	0,2521	0,0000	1,0
Maule	83.994	0,0659	0,2481	0,0000	1,0
Biobío	83.994	0,0899	0,2860	0,0000	1,0
Araucanía	83.994	0,0579	0,2335	0,0000	1,0
Los Lagos	83.994	0,0522	0,2225	0,0000	1,0
Aysén	83.994	0,0218	0,1460	0,0000	1,0
Magallanes	83.994	0,0286	0,1667	0,0000	1,0
Los Ríos	83.994	0,0524	0,2228	0,0000	1,0
Arica	83.994	0,0405	0,1971	0,0000	1,0
Ñuble	83.994	0,0366	0,1878	0,0000	1,0
Educación (hasta último nivel alcanzado)					
Sin educación	83.994	0,0053	0,0726	0,0000	1,0
Educación básica	83.994	0,1745	0,3796	0,0000	1,0
Educación media	83.994	0,4481	0,4973	0,0000	1,0
Educación técnica	83.994	0,1277	0,3337	0,0000	1,0
Mujer	83.994	0,4396	0,4963	0,0000	1,0
Edad (elevada al cuadrado)	83.994	286,08	1.281,99	225,0	8.464,0
Número de hijos menores de 6 años	83.975	0,0981	0,3378	0,0000	4,0
Tamaño promedio de empresas del sector	83.994	85,3031	71,0074	0,0000	621,8132
Tasa de accidentes no fatales	83.994	2,5365	0,8511	0,0000	3,9100
Trabajo con jornada completa	83.994	0,6202	0,4853	0,0000	1,0
Trabajo con contrato firmado	83.994	0,5995	0,4900	0,0000	1,0
Trabajo sin contrato	83.994	0,0069	0,0829	0,0000	1,0
Casado	83.994	0,3098	0,4624	0,0000	1,0
Años de escolaridad del cónyuge	64.872	8,4521	6,4375	0,0000	29,0
Número de personas en el hogar	83.994	3,4023	1,5680	0,0000	13,0
Ingresos no laborales	83.994	62.579,33	101.901,3	0,0000	1.849.188,0
Tamaño de la empresa					
1 a 5 trabajadores	83.994	0,2244	0,4172	0,0000	1,0
6 a 10 trabajadores	83.994	0,1537	0,3606	0,0000	1,0
11 a 50 trabajadores	83.994	0,0623	0,2417	0,0000	1,0
51 a 200 trabajadores	83.994	0,1615	0,3680	0,0000	1,0
201 a 499 trabajadores	83.994	0,1402	0,3472	0,0000	1,0
El cónyuge trabaja	83.994	0,3200	0,4665	0,0000	1,0
Trabajador independiente	83.994	0,2849	0,4514	0,0000	1,0
Empleado público	83.994	0,1368	0,3436	0,0000	1,0
Empleado doméstico	83.994	0,0257	0,1582	0,0000	1,0
Trabajo con jornada parcial	83.994	0,0751	0,2635	0,0000	1,0
Fuerzas armadas	83.994	0,0072	0,0844	0,0000	1,0
Trabajo familiar no remunerado	83.994	0,0045	0,0668	0,0000	1,0
Salario (en logaritmo)	58.208	13,192	0,6495	8,9872	17,0344

Fuente: Superintendencia de Seguridad Social (2023)

5. Resultados

El presente capítulo presenta los resultados obtenidos mediante los métodos de regresión lineal y regresión simbólica aplicados para la estimación del valor estadístico de la vida. Se analizan las métricas de desempeño de cada enfoque, se comparan sus ventajas y limitaciones. Finalmente, se exponen las implicancias prácticas de dichos resultados. Para el desarrollo de los modelos lineales se utilizó el software estadístico Stata, mientras que la implementación de la regresión simbólica se realizó por medio de lenguaje de programación Python.

5.1 Mínimos cuadrados ordinarios mediante regresión lineal

Para determinar la cantidad de variables para incluir en el modelo de regresión, se llevaron a cabo diversas pruebas incrementales. Inicialmente se utilizaron dos variables, aumentando de manera progresiva hasta 44 variables, aplicando indicadores estadísticos de rendimiento, como el coeficiente de determinación R^2, el estadístico de la Prueba F y su correspondiente valor de significación p. La Tabla 5 muestra cómo evolucionan los valores de los indicadores, el tamaño de la muestra y la cantidad de variables incluidas en cada iteración del modelo. Además, en el conjunto de datos se excluyeron las observaciones con datos faltantes, la muestra finalmente utilizada para las estimaciones quedó con 45.591 observaciones para el año 2022.

Tabla 5: Prueba incremental de variables año 2022.

Número de Variables	Observaciones	R^2	Prueba F	p
2	45.591	0,0011	24,78	0,000
4	45.591	0,0076	87,55	0,000
6	45.591	0,0086	65,64	0,000
8	45.591	0,0141	81,43	0,000
10	45.591	0,0210	97,61	0,000
12	45.591	0,0234	91,07	0,000
14	45.591	0,0304	102,07	0,000
16	45.591	0,0387	114,55	0,000
18	45.591	0,0773	212,22	0,000
20	45.591	0,2297	679,47	0,000
22	45.591	0,2895	844,04	0,000
24	45.591	0,3110	857,15	0,000
26	45.591	0,4107	1221,45	0,000
28	45.591	0,4391	1274,28	0,000
30	45.591	0,4826	1416,78	0,000
32	45.591	0,4848	1339,77	0,000
34	45.591	0,4918	1296,77	0,000
36	45.591	0,4983	1256,63	0,000
38	45.591	0,5000	1198,50	0,000
40	45.591	0,5001	1168,64	0,000
42	45.591	0,5097	1154,75	0,000
44	45.591	0,5104	1130,63	0,000

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

Como se puede observar en la Tabla 5, el valor R^2 aumenta al incorporar más variables, lo que indica una mayor proporción de variabilidad de los salarios explicada por el modelo. No obstante, es importante considerar que el valor R^2 tiende a incrementarse naturalmente al añadir mas variables, por lo que su crecimiento no siempre implica una mejora sustancial del modelo.

Finalmente, se opta por construir un modelo con 44 variables explicativas, de las 44 propuestas inicialmente, como se puede apreciar en la Tabla 6, se omitieron las variables "Trabajador independiente" y "Trabajo familiar no remunerado" para el año 2022 y 2017, además de la variable "Trabajo con jornada parcial" que se omitió exclusivamente en el año 2017. A partir de este modelo, se estima un sueldo base de $624.631 y, utilizando una tasa de cambio de $859 por dólar (Banco Central de Chile, 2022), se calcula el valor estadístico de la vida para 2022 de 8,93 millones de dólares aproximadamente. En comparación, el modelo estimado con datos del año 2017, con un tipo de cambio de $615 por dólar y un sueldo promedio de $473.692, se estima un valor estadístico de la vida de 2,75 millones de dólares aproximadamente, como se puede apreciar en la Tabla 6.

El valor del valor estadístico de la vida 2022 es relativamente alto en comparación la estimación obtenida en el año 2017. Además, es relevante observar los coeficientes de regresión reportados. Por ejemplo, en el caso de las regiones, la mayoría de las regresiones presentan coeficientes negativos respectos a la región Metropolitana (variable omitida), lo cual implica que sus salarios son más bajos en comparación. En el caso de "Biobío", el coeficiente es de -0,1400, lo que sugiere que los salarios son en promedio un 13,06% menores que en la región Metropolitana, según la Ecuación (15).

$$(e^{-0,1400} - 1) * 100\% = -13,06\% \quad (15)$$

En cuanto al nivel educacional, el nivel profesional es utilizado como categoría base. Así, los coeficientes negativos para los otros niveles reflejan una penalización salarial. Por ejemplo, para personas si "Educación técnica", el coeficiente es -0,3870, lo que se traduce en un salario promedio 32,09% menor que quienes posee formación profesional

Respecto al género, la categoría base es hombre, por lo tanto, el coeficiente de -0,1610 para mujeres indica que estas ganan, en promedio, un 14,87% menos que los hombres.

Tabla 6: Resultados metodología MCO años 2022 y 2017.

Variable	Coeficiente (2022)	Error estándar (2022)	Coeficiente (2017)	Error estándar (2017)
Tasa de fatalidad	0,1023***	(0,008)	0,0297	(0,027)
Región				
Tarapacá	–0,0784***	(0,011)	–0,1551***	(0,049)
Antofagasta	0,0377***	(0,011)	–0,1403***	(0,053)
Atacama	–0,0347***	(0,011)	–0,1439**	(0,060)
Coquimbo	–0,0840***	(0,012)	–0,0577	(0,054)
Valparaíso	–0,1330***	(0,008)	–0,4343***	(0,039)
O´Higgins	–0,1048***	(0,009)	–0,2808***	(0,041)
Maule	–0,1561***	(0,009)	–0,3208***	(0,043)
Biobío	–0,1403***	(0,008)	–0,0844**	(0,038)
Araucanía	–0,1777***	(0,010)	–0,1907***	(0,045)
Los Lagos	–0,0829***	(0,010)	–0,3601***	(0,046)
Aysén	0,0329**	(0,016)	–0,2698***	(0,065)
Magallanes	0,0630***	(0,013)	–0,2059***	(0,054)
Los Ríos	–0,1491***	(0,010)	–0,2604***	(0,050)
Arica	–0,1053***	(0,012)	–0,1036*	(0,060)
Ñuble	–0,1537***	(0,012)	–0,6149***	(0,057)
Educación (hasta último nivel alcanzado)				
Sin educación	–0,6140***	(0,035)	–0,6666***	(0,128)
Educación básica	–0,6322***	(0,008)	–0,3939***	(0,042)
Educación media	–0,5150***	(0,006)	–0,4368***	(0,032)
Educación técnica	–0,3869***	(0,007)	–0,3291***	(0,031)
Mujer	–0,1607***	(0,005)	–0,2620***	(0,025)
Edad (elevada al cuadrado)	0,00003***	(0,000)	0,0000***	(0,000)
Número de hijos menores de 6 años	–0,0103	(0,007)	0,1033***	(0,030)
Tamaño promedio de empresas del sector	0,0006***	(0,000)	0,0000***	(0,000)
Tasa de accidentes no fatales	–0,0417***	(0,003)	0,0196	(0,012)
Trabajo con jornada completa	0,0972***	(0,015)	0,0484	(0,037)
Trabajo con contrato firmado	0,2403***	(0,007)	–0,0357	(0,028)
Trabajo sin contrato	0,1161***	(0,024)	–0,0201	(0,075)
Casado	0,0349***	(0,005)	0,1116***	(0,022)
Años de escolaridad del cónyuge	0,0210***	(0,000)	0,0196***	(0,003)
Número de personas en el hogar	0,0030*	(0,002)	–0,0620***	(0,007)
Ingresos no laborales	0,0000***	(0,000)	0,0000	(0,000)
Tamaño de la empresa				
1 a 5 trabajadores	–0,2397***	(0,014)	–0,2293***	(0,057)
6 a 10 trabajadores	–0,1902***	(0,008)	–0,1338***	(0,037)
11 a 50 trabajadores	–0,1565***	(0,009)	–0,0520	(0,041)
51 a 200 trabajadores	–0,1133***	(0,006)	0,0536**	(0,027)
201 a 499 trabajadores	–0,0499***	(0,006)	0,0861***	(0,028)
El cónyuge trabaja	–0,0424***	(0,006)	–0,7791***	(0,059)
Trabajador independiente	Omitida	Omitida	Omitida	Omitida
Empleado público	0,0321***	(0,006)	0,0052	(0,030)
Empleado doméstico	–0,0010***	(0,018)	–0,0430	(0,073)
Trabajo con jornada parcial	–0,4829***	(0,016)	Omitida	Omitida
Fuerzas armadas	0,1694***	(0,020)	–0,0973	(0,078)
Trabajo familiar no remunerado	Omitida	Omitida	Omitida	Omitida
Constante	13,3449***	(0,020)	13,0683***	(0,093)
Año	2022		2017	
Valor estadístico de la vida (millón de dólares)	8,9274		2,7471	

Nota: *** p< 0,01, ** p<0,05, * p<0,1

Fuente: Elaboración propia en base a resultados obtenidos en Stata

5.1.1 Análisis de robustez

Al analizar el modelo mediante unas métricas estadísticas de desempeño, se puede apreciar que el coeficiente de determinación R^2 es 0,5104 para el año 2022, lo que implica que el modelo explica el 51% de la variabilidad total de la variable dependiente, que en este caso es el logaritmo natural del salario. Aunque este porcentaje puede parecer moderado, resulta adecuado en el contexto de salarios hedónicos, dado que es probable que otros factores no incluidos en el modelo también influyan en el salario, o incluso, hay relaciones entre las variables, que la regresión lineal no es capaz de reconocer. Además, se puede apreciar que la diferencia con el coeficiente de determinación ajustado (R^2 ajustado) es mínima, con un valor de 0,5100, lo que sugiere que no existe un problema de sobreajuste, confirmando que las variables seleccionadas aportan información relevante. La prueba F del modelo alcanza un valor de 1130,63, lo que indica que el conjunto de variables explicativas mejora significativamente el ajuste en comparación con un modelo sin variables explicativas. Además, el valor de significación es 0,000, lo que permite rechazar la hipótesis nula de que todos los coeficientes sean iguales a cero. Finalmente, la raíz del error cuadrático medio es 0,4482 lo que lo que refleja una precisión razonable del ajuste, como se puede observar en la Tabla 7.

Tabla 7 Análisis de robustez MCO 2022.

	SS	df	MS	F	p
Modelo	9.538,0633	42,0	227,0967	1130,6340	0,000
Residuos	9.148,6747	45548,0	0,2009		
Total	18.686,7380	45590,0			
	R^2	R^2 ajustado	RMSE		
Modelo	0.5104	0.5100	0.4482		

Fuente: Elaboración propia en base a resultados obtenidos en Stata

Por otro lado, el Factor de Inflación de Varianza (VIF) mide cuánto se incrementa la varianza de un coeficiente cuando una variable explicativa está correlacionada con las demás. Un VIF próximo a 1 indica ausencia de colinealidad entre variables, un valor entre 1 y 5 sugiere una correlación moderada, generalmente sin implicar problemas graves. Mientras que un VIF superior a 10 puede evidenciar una alta correlación que afecta a la estabilidad de los coeficientes estimados.

Como se puede apreciar en la Tabla 8, las variables relacionadas con el tipo de jornada laboral muestran valores de VIF más elevados, con valores cercanos a 5. Por lo tanto, no representan un problema significativo de multicolinealidad. Por otro lado, las variables “Sin educación” y

“Trabajador sin contrato” presentan valores más bajos, lo que indica una correlación mínima con el resto de los regresores. El VIF promedio es 1,54 para el año 2022 y 1,40 para el año 2017, aproximadamente, lo que refuerza la conclusión de que, en términos generales, no existe una multicolinealidad severa en el modelo. Si bien algunas variables presentan cierto grado de correlación, este no compromete la estabilidad ni la confiabilidad de las estimaciones.

Tabla 8: Análisis multicolinealidad MCO 2022 y 2017.

Variable	VIF 2022	VIF 2017
Tasa de fatalidad	1,3787	1,8071
Tasa de accidentes no fatales	1,6450	2,2082
Educación (hasta último nivel alcanzado)		
Sin educación	1,0431	1,1176
Educación básica	1,8763	2,8665
Educación media	1,7820	2,4447
Educación técnica	1,3392	1,7967
Mujer	1,4680	1,4959
Edad (elevada al cuadrado)	1,6537	1,4733
Número de hijos menores de 6 años	1,2470	1,3444
Tamaño promedio de empresas del sector	1,4148	1,7352
Trabajo con jornada completa	5,1710	1,0567
Trabajo con jornada parcial	5,2298	-
Trabajo con contrato firmado	1,3321	1,2105
Trabajo sin contrato	1,0628	1,1084
Casado	1,5428	1,2150
Años de escolaridad del cónyuge	2,0599	1,9354
Número de personas en el hogar	1,1851	1,2527
Ingresos no laborales	1,1930	1,1481
Tamaño de la empresa		
1 a 5 trabajadores	1,9305	1,8129
6 a 10 trabajadores	1,3512	1,1668
11 a 50 trabajadores	1,1850	1,1151
51 a 200 trabajadores	1,3239	1,2210
201 a 499 trabajadores	1,2753	1,1837
El cónyuge trabaja	1,7416	1,0990
Empleado público	1,2959	1,4098
Empleado doméstico	1,9175	1,7740
Fuerzas armadas	1,0727	1,1754
Región		
Tarapacá	1,1632	1,1979
Antofagasta	1,1744	1,1674
Atacama	1,1786	1,1331
Coquimbo	1,1424	1,1632
Valparaíso	1,3350	1,2988
O´Higgins	1,2685	1,3080
Maule	1,2598	1,2838
Biobío	1,3255	1,3317
Araucanía	1,2065	1,2449
Los Lagos	1,2091	1,2202
Aysén	1,0770	1,1143
Magallanes	1,1248	1,1451
Los Ríos	1,1915	1,1713
Arica	1,1480	1,1266
Ñuble	1,1502	1,1507

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

5.2 Regresión simbólica por programación genética

La metodología de regresión simbólica busca descubrir de forma directa una expresión que prediga el logaritmo natural del salario a partir del conjunto de variables explicativas, midiendo su comportamiento y ajuste en los datos. Para ello, se elabora un modelo en Python, utilizando principalmente la biblioteca PySRRegressor, que recurre a algoritmos genéticos en una búsqueda evolutiva de expresiones analíticas. El modelo de regresión simbólica se entrenó utilizando el 80% de la muestra original, mientras que el 20% restante se reservó para validar su desempeño y así evaluar la robustez del modelo. Durante el proceso de entrenamiento se exploraron diversas combinaciones de operadores aritméticos básicos como suma, resta, multiplicación y división, junto con funciones no lineales como exponenciales, raíces cuadradas, logaritmos, valores absolutos y funciones trigonométricas.

Para estudiar desempeño y estabilidad del método, se aplicaron varias instancias de entrenamiento, variando parámetros clave. El objetivo principal en esta exploración es evaluar el comportamiento de la regresión simbólica en la especificación del modelo, evaluado por medio de las métricas de desempeño (la raíz del error cuadrático medio y coeficiente de determinación ajustado) sobre la variable dependiente en el conjunto de validación.

La Tabla 9 resume cinco combinaciones distintas de parámetros utilizados en el entrenamiento del modelo de regresión simbólica. Las columnas representan cada una de las combinaciones evaluadas, mientras que las filas indican los valores asignados a los principales parámetros del modelo, los cuales son el número de iteraciones, el tamaño de la población y el tamaño máximo. Finalmente, se incluye el tiempo total de entrenamiento registrado para cada ejecución.

Tabla 9: Parametrización de combinaciones de regresión simbólica.

Parámetros	Combinación 1	Combinación 2	Combinación 3	Combinación 4	Combinación 5
Número de iteraciones	60	100	200	250	300
Tamaño de la Población	50	50	60	70	80
Tamaño máximo	50	90	90	100	150
Tiempo de entrenamiento	66,3 s	45,3 s	139,5 s	140,2 s	277,8 s

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

La “Combinación 1” de parámetros resultó ser la más efectiva, con la expresión simbólica resultante, la cual incluye tan solo cinco variables, como muestra la Ecuación (16). Esta variante alcanzó un valor de la raíz del error cuadrático medio de 0,4800 y un coeficiente de determinación ajustado de 0,4259

en validación, con un tiempo de entrenamiento de 66,3 segundos. Dicha expresión revela interacciones relevantes entre las variables "Años de escolaridad del cónyuge", "Educación media", "Trabajo con contrato firmado", "Trabajo con jornada parcial" y "Tasa de accidentes no fatales". Además, se puede apreciar que la mejor expresión encontrada utilizó mayoritariamente operadores aritméticos.

$$\left(\left|\frac{\text{Años de escolaridad del cónyuge} - 5{,}2833886}{-1{,}529195 \; - \; \text{Educación media}}\right| + \right.$$

$$\left.\frac{(\text{Trabajo con contrato firmado} - 0{,}6147644) - 2 \times (\text{Trabajo con jornada parcial})}{0{,}8160713}\right) \times \qquad (16)$$

$$1{,}939894 \times (0{,}16443613 - \text{Tasa de accidentes no fatales}) + 12{,}656367$$

Se puede apreciar que, al comparar los resultados con los obtenidos mediante la regresión lineal, la raíz del error cuadrático medio disminuye de 0,4802 a 0,4800. Sin embargo, el coeficiente de determinación ajustado no alcanza los valores obtenidos mediante los métodos lineales, el cual es de 0,5104 contra 0,4259 obtenidos mediante la regresión simbólica.

Tabla 10: Análisis de robustez de regresión simbólica año 2022.

Métricas de desempeño	Combinación 1	Combinación 2	Combinación 3	Combinación 4	Combinación 5
Raíz del error cuadrático medio	0,4800	0,5153	0,5052	0,5196	0,4867
Coeficiente de determinación ajustado	0,4259	0,3354	0,3691	0,3321	0,4175

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

Este enfoque utilizado para la regresión simbólica no se forzó la presencia de la variable independiente "Tasa de fatalidad" en la ecuación final, por lo que no se estimó un valor estadístico de la vida. En este caso se optó por una exploración libre, para evaluar el ajuste del modelo de regresión simbólica mediante las métricas estadísticas de desempeño. El modelo se ajustó para preferir que incluyan la mayor cantidad de variables explicativas y que a su vez, optimicen los resultados, buscando la minimización de la raíz del error cuadrático medio.

5.3 Regresión simbólica sobre los residuos de la regresión lineal

Para mejorar los resultados obtenidos tanto de la regresión simbólica ejecutada anteriormente, como del método de mínimos cuadrados ordinarios, con el objetivo de asegurar la permanencia de la variable "Tasa de fatalidad" en la expresión final, para posteriormente poder estimar el valor estadístico de la

vida, se aplica la regresión simbólica sobre los residuos de la regresión lineal, la cual opera sobre la fracción no explicada por el modelo lineal.

La técnica de regresión simbólica sobre residuos se desarrolla a partir de un modelo de mínimos cuadrados ordinarios, ajustado sobre el 80% de los datos seleccionados aleatoriamente para entrenamiento. A partir de este modelo lineal inicial se obtuvieron las predicciones de la variable dependiente, o sea el logaritmo natural del salario, y se calcularon los residuos como la diferencia entre los valores observados y los predichos. Estos residuos representan la parte de la variabilidad salarial no explicada por la relación lineal.

Para capturar no linealidades e interacciones omitidas, se utilizó nuevamente la biblioteca PySRRegressor, que recurre a algoritmos genéticos en una búsqueda evolutiva de expresiones analíticas. En esta etapa, al igual que en la regresión simbólica ejecutada con anterioridad, se incluyeron los operadores aritméticos (suma, resta, multiplicación y división) junto a funciones no lineales (exponenciales, raíces cuadradas, logaritmos, valores absolutos, seno, coseno y tangente). Para garantizar un amplio espacio de búsqueda, se exploraron diferentes combinaciones de parámetros variando el número de iteraciones, el tamaño de la población, las probabilidades de cruce y mutación, entre otros, como se puede observar en la Tabla 11.

Tabla 11: Parametrización de combinaciones de regresión simbólica de residuos.

Parámetros	Combinación 1	Combinación 2	Combinación 3	Combinación 4	Combinación 5	Combinación 6
Número de iteraciones	100	200	100	50	80	100
Tamaño de la Población	40	40	50	25	50	20
Tamaño máximo	80	50	44	25	10	15
Probabilidad cruce	10%	20%	20%	20%	20%	20%
Probabilidad de mutación	90%	80%	80%	80%	80%	80%
Tiempo de ejecución (segundos)	84,49	158,96	413,03	405,23	403,74	412,41

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

Una vez finalizado el entrenamiento simbólico sobre los residuos del modelo lineal, como se puede apreciar en la Tabla 12. Se seleccionaron dos combinaciones que reflejaron el mismo rendimiento, la "Combinación 5" y la "Combinación 6". Sin embargo, se opta por la "Combinación 5", ya que, está presenta un menor tiempo de cómputo. Posteriormente, la expresión simbólica seleccionada (17), es

aplicada a la muestra completa para generar un nuevo regresor que, junto a las variables originales, fue incorporado en una reestimación global de los coeficientes de cada parámetro del modelo. Esta reestimación permitió cuantificar el cambio en el coeficiente "Tasa de fatalidad" tras incluir el término no lineal.

$$\text{Tamaño promedio de empresas del sector} \times (0{,}0494 - 0{,}1497 \times \sin(0{,}1954 \times \text{Casado})) \quad (17)$$

Tabla 12: Análisis de robustez de combinaciones de regresión simbólica de residuos años 2022.

Métricas de desempeño	Combinación 1	Combinación 2	Combinación 3	Combinación 4	Combinación 5	Combinación 6
Raíz del error cuadrático medio	0,4482	0,4481	0,4482	0,4481	0,4478	0,4478
Coeficiente de determinación ajustado	0,5100	0,5101	0,5100	0,5100	0,5109	0,5109

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

Finalmente, para estimar el valor estadístico de la vida, al igual que en la regresión lineal, se utiliza el salario promedio anual, la tasa de conversión de peso chileno por dólar y se deriva la función completa respecto al regresor "Tasa de fatalidad", escalando este valor para 10.000 individuos. El valor estadístico de la vida obtenido en la "Combinación 6" es de 9,1581 millones de dólares en el año 2022. Sin embargo, las métricas estadísticas presentan resultados muy similares entre sí, además de que los resultados obtenidos en la totalidad de las iteraciones son bastante similares. Por lo tanto, se puede concluir que el valor estadístico de la vida obtenido mediante la regresión simbólica oscila entre 8,8664 y 9,2467 millones de dólares en el año 2022 y entre 2,8274 y 3,0233 millones de dólares para el año 2017.

De manera complementaria, se evaluó la variabilidad del coeficiente de la variable "Tasa de fatalidad" obtenido en las estimaciones de las 6 combinaciones del año 2022. La desviación estándar de este coeficiente fue de 0,002 lo que equivale a un coeficiente de variación del 1,75%. Este valor indica una variabilidad relativa baja entre las estimaciones, lo que refuerza la estabilidad de la regresión simbólica sobre residuos.

Tabla 13: Estimación del valor estadístico de la vida año mediante regresión simbólica.

Año	Combinación 1	Combinación 2	Combinación 3	Combinación 4	Combinación 5	Combinación 6
2022	8,8679	8,8806	8,8691	9,2467	9,1581	9,1581
2017	2,8274	3,0233	2,8205	2,8317	2,8361	2,8704

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

En la Tabla 14, se pueden apreciar los coeficientes reajustados de cada variable resultante del año 2022. Se consideraron las cuatro combinaciones que reportaron un menor valor de la raíz del error cuadrático medio, las cuales fueron: “Combinación 2”, “Combinación 4”, “Combinación 5” y “Combinación 6”.

Además, los coeficientes estimados de la variable “Tasa de fatalidad” de las cuatro combinaciones de la Tabla 14, se compararon con el intervalo de confianza al 95% del mismo coeficiente en la regresión lineal por mínimos cuadrados ordinarios. Según lo reportado en la Tabla 6, para el año 2022 el modelo lineal estimó un coeficiente de la variable “Tasa de fatalidad” equivalente a 0,1023, con un error estándar de 0,008, lo que se traduce en un intervalo de confianza de {0,0866;0,1180}. En todos los casos, los valores obtenidos mediante la regresión simbólica sobre residuos se ubicaron dentro del intervalo de confianza del modelo lineal, lo que evidencia una alta coherencia entre ambos métodos y respalda la validez del enfoque simbólico como alternativa consistente al método tradicional.

Tabla 14: Coeficientes regresión simbólica de residuos año 2022.

Variable	Combinación 2	Combinación 4	Combinación 5	Combinación 6
Tasa de fatalidad	0,1025	0,1027	0,1057	0,1057
Región				
Tarapacá	-0,0783	-0,0783	-0,0779	-0,0779
Antofagasta	0,0379	0,0378	0,0368	0,0368
Atacama	-0,0347	-0,0345	-0,0347	-0,0347
Coquimbo	-0,0840	-0,0843	-0,0850	-0,0850
Valparaíso	-0,1330	-0,1328	-0,1329	-0,1329
O´Higgins	-0,1048	-0,1047	-0,1052	-0,1052
Maule	-0,1561	-0,1553	-0,1554	-0,1554
Biobío	-0,1403	-0,1403	-0,1400	-0,1400
Araucanía	-0,1779	-0,1773	-0,1774	-0,1774
Los Lagos	-0,0830	-0,0825	-0,0831	-0,0831
Aysén	0,0322	0,0316	0,0317	0,0317
Magallanes	0,0629	0,0635	0,0627	0,0627
Los Ríos	-0,1492	-0,1491	-0,1486	-0,1486
Arica	-0,1055	-0,1050	-0,1058	-0,1058
Ñuble	-0,1538	-0,1530	-0,1536	-0,1536
Educación (hasta último nivel alcanzado)				
Sin educación	-0,6145	-0,6152	-0,6137	-0,6137
Educación básica	-0,6322	-0,6148	-0,6300	-0,6300
Educación media	-0,5149	-0,5148	-0,5154	-0,5154
Educación técnica	-0,3779	-0,3864	-0,3865	-0,3865
Mujer	-0,1607	-0,1606	-0,1599	-0,1599
Edad (elevada al cuadrado)	0,0000	0,0000	0,0000	0,0000
Número de hijos menores de 6 años	-0,0107	-0,0090	-0,0108	-0,0108
Tamaño promedio de empresas del sector	0,0006	0,0006	0,0013	0,0015
Tasa de accidentes no fatales	-0,0417	-0,0417	-0,0420	-0,0420
Trabajo con jornada completa	0,0975	0,0976	0,0977	0,0977
Trabajo con contrato firmado	0,2408	0,2409	0,2417	0,2417
Trabajo sin contrato	0,1166	0,1166	0,1164	0,1164
Casado	0,3366	0,0571	-0,0169	-0,0169
Años de escolaridad del cónyuge	0,0207	0,0205	0,0206	0,0206
Número de personas en el hogar	0,0029	-0,0004	0,0023	0,0023
Ingresos no laborales	-0,0000	-0,0000	-0,0000	-0,0000
Tamaño de la empresa				
1 a 5 trabajadores	-0,2400	-0,2398	-0,2387	-0,2387
6 a 10 trabajadores	-0,1804	-0,1901	-0,1894	-0,1894
11 a 50 trabajadores	-0,1563	-0,1561	-0,1550	-0,1550
51 a 200 trabajadores	-0,1133	-0,1130	-0,1123	-0,1123
201 a 499 trabajadores	-0,0501	-0,0500	-0,0490	-0,0490
El cónyuge trabaja	-0,0422	-0,0428	-0,0418	-0,0418
Trabajador independiente	omitida	omitida	omitida	omitida
Empleado público	0,0320	0,0322	0,0322	0,0322
Empleado doméstico	-0,0011	-0,0022	-0,0003	-0,0003
Trabajo con jornada parcial	0,0975	0,0976	0,0977	0,0977
Fuerzas armadas	0,1716	0,1686	0,1642	0,1642
Trabajo familiar no remunerado	omitida	omitida	omitida	omitida

Fuente: Elaboración propia en base a resultados obtenidos en Jupyter Notebook

Al comparar las métricas estadísticas obtenidas en el año 2022 (Tabla 13) por ambos enfoques de regresión simbólica se puede apreciar que, analizando la mejor configuración de cada uno, ambos alcanzan valores bastante similares de la raíz del error cuadrático medio (0,4478 contra 0,4480), pero la regresión simbólica sobre residuos es bastante superior respecto a el coeficiente de determinación ajustado (0,5109 contra 0,4259).

Al analizar las métricas estadísticas de desempeño medidas en todas las combinaciones (Tabla 15). El método sobre residuos mantiene una variabilidad muy baja ante distintos parámetros, con una media de valor de la raíz del error cuadrático medio de 0,4480 y una desviación estándar de 0,0002. Mientras que el enfoque tradicional oscila con mayor amplitud con un valor promedio de 0,5104 y una desviación estándar de 0,0174. Esta consistencia en las estimaciones, junto con la posibilidad de estimar el valor estadístico de la vida, consolida la regresión simbólica sobre residuos como una alternativa más estable y confiable para la aplicación de la metodología de salarios hedónicos.

El enfoque sobre residuos al incorporar las variables explicativas obtenidas por medio de la regresión lineal, incluyendo relaciones nuevas, mantiene siempre la totalidad de las variables posibles, lo que tiende a ampliar la capacidad explicativa del modelo. Sin embargo, la regresión simbólica tradicional casi iguala el valor de la raíz del error cuadrático medio en su mejor combinación, utilizando exclusivamente cuatro variables, demostrando que es posible lograr resultados aceptables con configuraciones más parsimoniosas.

Tabla 15: Comparación de ambos enfoques de regresión simbólica año 2022.

	Enfoque sobre residuos		Enfoque tradicional	
Métricas de desempeño	Media	Desviación estándar	Media	Desviación estándar
Raíz del error cuadrático medio	0,4880	0,0002	0,5014	0,0174
Coeficiente de determinación ajustado	0,5103	0,0004	0,3760	0,0443

Fuente: Elaboración propia

5.4 Comparación de resultados

El presente estudio aplica diversos métodos para la estimación del valor estadístico de la vida en Chile. Entre estos se encuentran el método de mínimos cuadrados ordinarios, mediante la regresión lineal y el método de regresión simbólica aplicado sobre los residuos de la regresión lineal anteriormente mencionada. Además, también son comparados por los resultados obtenidos por Retamal (2024), en

su memoria de título “Estimación del valor estadístico de la vida en Chile comparando métodos de regresión lineal”.

Al aplicar la regresión lineal sobre las 42 variables explicativas definidas sobre el enfoque de salarios hedónicos con datos de la CASEN y la SUSESO, se obtuvo un valor estadístico de la vida de 8,93 millones de dólares. Dicha metodología ajusta los datos de manera directa, un coeficiente de determinación ajustado de 0,5100 y una raíz del error cuadrático medio de 0,4482. Su gran ventaja radica en la sencillez en su aplicación y en el mínimo ajuste o restricción para la aplicación del modelo, lo que genera la obtención de resultados con relativa rapidez. Sin embargo, puede verse afectado por el sesgo de especificación incorrecta, debido a que se limita a expresiones lineales y no es capaz de reconocer interacciones complejas entre las variables. Por su parte, por medio de la regresión simbólica sobre residuos, se obtuvo un valor estadístico de la vida de 9,16 millones de dólares. Esto mejora levemente los resultados estadísticos, reportando un valor de la raíz del error cuadrático medio de 0,4478 y coeficiente de determinación ajustado de 0,5109. Esto es gracias a que permite explorar de forma autónoma combinaciones de operadores aritméticos y funciones no lineales. Sin embargo, requiere de una correcta manipulación de los parámetros para la optimización de los resultados, además presenta un grado de incertidumbre al presentar variabilidad entre las iteraciones.

Finalmente, se compararon los resultados obtenidos por Retamal (2024), quién implementó la metodología de penalización LASSO con validación cruzada, que tiene como objetivo controlar el sobreajuste y manejar la multicolinealidad, obteniendo 41 variables, un coeficiente de determinación de 0,5093 y un valor estadístico de la vida de 8,66 millones de dólares. Aunque LASSO reduce eficazmente la complejidad del modelo, algunos coeficientes quedan forzados a cero, lo que puede dificultar la interpretación individual de los efectos.

En conjunto, al comparar las métricas estadísticas, la regresión simbólica demuestra ser la mejorar si se busca mejorar la especificación del modelo y si se busca capturar patrones relevantes, mientras que MCO y LASSO siguen siendo opciones muy competitivas cuando la velocidad y la simplicidad son prioritarias.

La Tabla 16 resume y compara los tres métodos utilizados para la estimación del valor estadístico de la vida en el año 2022, el método de mínimos cuadrados ordinarios mediante la regresión lineal, el método LASSO y la regresión simbólica sobre residuos. Se destacan sus principales ventajas y

desventajas, junto con sus métricas de desempeño, el valor estadístico de la vida y el tiempo aproximado de ejecución requerido.

Tabla 16: Comparación métodos de estimación año 2022.

Método	Ventajas	Desventajas	RMSE	R^2 ajustado	Valor estadístico de la vida (Millones USD)	Tiempo de ejecución
MCO	Implementación sencilla y rápida con coeficientes fáciles de interpretar.	Asume linealidad estricta y es sensible a correlaciones entre regresores.	0,4482	0,5100	8,93	≈ 0,5 s
LASSO	Falta de regularización	Poca o nula penalización	0,4490	0,5093	8,66	≈ 20 s
Regresión simbólica sobre residuos	Captura no linealidades faltantes y presenta baja desviación entre iteraciones	Alto costo computacional.	0,4478	0,5109	9,16	≈ 404 s

Fuente: Elaboración propia en conjunto con datos provenientes de la memoria de título de Retamal (2024)

5.5 Implicancias en políticas públicas

La estimación del valor estadístico de la vida a partir de una muestra nacional de más de 45.000 trabajadores (CASEN 2022), complementada con tasas de accidentabilidad reportadas por la SUSESO, permite no solo cuantificar este valor con base empírica, sino también identificar relaciones no lineales relevantes mediante el uso de la regresión simbólica, A diferencia de los métodos tradicionales, este enfoque mejora la robustez de las estimaciones al capturar interacciones complejas que los modelos lineales no logran detectar.

Con esta información, las políticas públicas pueden dirigirse con mayor eficiencia. Si las decisiones introducen incentivos monetarios (subsidios, multas, campañas) en las industrias de mayor sensibilidad entre riesgo y salario, se puede maximizar la reducción de mortalidad por peso presupuestario, evitando aplicar medidas que subutilizan recursos o dejan vacíos en áreas críticas.

Otro aporte clave de la regresión simbólica, es la capacidad de actualización continua. Una vez implementado el modelo, se puede recalibrar en otro periodo, con nuevos datos, permitiendo la introducción de nuevas variables futuras, de modo que las políticas se mantengan siempre alineadas

con los contextos recientes y no queden desfasadas ante variaciones socioeconómicas o tecnológicas. Además, por medio de un valor estadístico de la vida más preciso, los organismos estatales pueden asignar recursos de prevención con mucha mayor precisión en relación con el contexto nacional, de esta forma se puede identificar cuanto invertir en equipamiento de seguridad, capacitaciones o inspecciones.

En definitiva, contar con valor estadístico de la vida calibrado a la realidad chilena, brinda a las autoridades una herramienta de orientación cuantitativa para maximizar el bienestar social, permite destinar cada peso público de manera eficiente para salvar más vidas y al mismo tiempo, transparenta las decisiones ante la ciudadanía al basarlas en evidencia objetiva y actualizada.

6. Conclusión

En el presente estudio se estimó el valor estadístico de la vida siguiendo dos enfoques complementarios, el método de mínimos cuadrados ordinarios y la regresión simbólica. Con el método lineal de mínimos cuadrados ordinarios se obtuvo un valor estadístico de la vida de 8,93 millones de dólares. Posteriormente, se aplicó regresión simbólica obtenida mediante programación genética que explora combinaciones de operadores aritméticos y funciones no lineales, generando así, nuevas relaciones entre variables que no fueron capturadas por la regresión lineal.

La incorporación de este término simbólico y tras la reestimación global del modelo produjo un valor estadístico de la vida de 9,16 millones de dólares, lo cual refleja una cifra mayor al valor obtenido originalmente. Este aumento se respalda de un ajuste correcto del modelo, verificado con un set de datos aleatorios correspondientes a un 20% del total de la muestra, con el objetivo de demostrar que el modelo predice con credibilidad y alineado con el contexto real de los datos, evitando así un modelo sobre ajustado. La robustez del modelo es corroborada por medio de métricas estadísticas reportando un decrecimiento de la raíz del error cuadrático medio y un aumento del coeficiente de determinación ajustado, lo que demuestra que el modelo aumenta su capacidad predictiva.

En términos prácticos, mientras que el MCO conserva la ventaja de una implementación inmediata y una interpretación directa de los coeficientes, la regresión simbólica se perfila como una herramienta poderosa cuando existe sospecha de no linealidades o interacciones omitidas. Su capacidad para generar expresiones analíticas explicitas, a través de grandes volúmenes de datos, permite disponer de estimaciones precisas del valor estadístico de la vida y adaptadas a la realidad del contexto laboral chileno. Asimismo, el modelo de regresión simbólica es altamente adaptable en futuras ocasiones, ya que, puede reentrenarse con nuevos datos, incorporar o eliminar variables según cambien los escenarios económicos, sociales o tecnológicos y ajustare sin perder la interpretabilidad de sus resultados. De este modo, ofrece una base metodológica robusta y flexible para la para la evaluación costo-beneficio en la gestión de riesgos laborales y para la asignación eficiente de recursos en prevención y seguridad.

A pesar de las mejoras relativas que ofrece la regresión simbólica frente al modelo lineal, el margen de optimización sigue siendo considerable. Es por ello, que se sugiere continuar integrando nuevas variables, técnicas de selección y transformaciones funcionales del modelo, así como evaluar distintos esquemas de entrenamiento y validación, con el fin de potenciar aún más la capacidad predictiva y su ajuste a la complejidad del entorno laboral.

7. Referencias

Abdellaoui, I. A., & Mehrkanoon, S. (2021). Symbolic regression for scientific discovery: an application to wind speed forecasting. *2021 IEEE symposium series on computational intelligence (SSCI)* (págs. 01-08). IEEE.

Andersson, H. (2005). The value of safety as revealed in the Swedish car market: an application of the hedonic pricing approach. *Journal of Risk and Uncertainty, 30*, 211–239.

Andersson, H., & Treich, N. (2011). The value of a statistical life. En *A handbook of transport economics.* Edward Elgar Publishing.

Ashenfelter, O. (2006). Measuring the value of a statistical life: problems and prospects. *The Economic Journal, 116(510)*, C10-C23.

Banco Central de Chile. (2022). *Tipo de cambio histórico del dólar observado*. Obtenido de Banco Central de Chile.

Bateman, I. J., & Kling, C. L. (2020). Revealed preference methods for nonmarket valuation: An introduction to best practices . *Review of Environmental Economics and Policy*.

Biddle, J. E., & Zarkin, G. A. (1988). Worker preference and market compensation for job risk. *The Review of Economics and Statistics, 70*(4), 660–667.

Blackwell, M., & Olson, M. P. (2022). Reducing model misspecification and bias in the estimation of interactions. *Political Analysis, 30*(4), 495–514.

Carson, R. T., & Hanemann, W. M. (2005). Contingent valuation. *Handbook of Environmental Economics, 2*, 821–936.

Carson, R. T., & Louviere, J. J. (2011). A common nomenclature for stated preference elicitation approaches. *Environmental and Resource Economics, 49*, 539–559.

Cifuentes, L. A., Prieto, J. J., & Escobari, J. (2008). *Valuing mortality risk reductions at present and at advanced age: preliminary results from a contingent valuation study in Chile.* Pontificia Universidad Católica de Chile, Departamento de Ingeniería Industrial y de Sistemas.

Cukrowska-Torzewska, E., & Matysiak, A. (2020). The motherhood wage penalty: A meta-analysis. *Social Science Research, 88–89*(102416).

Escobar, C. P. (2007). Cost benefit analysis, value of a statistical life and culture: Challenges for regulation. *Vniversitas, 56*(113), 235–258.

Gobierno de Chile. (Junio de 2025). *¿Cuánto subirá el sueldo mínimo tras aprobación en el Congreso?* Obtenido de Gob.cl: https://www.gob.cl/noticias/conoce-cuanto-aumenta-sueldo-minimo/#:~:text=En%202022%2C%20cuando%20inici%C3%B3%20sus,ya%20se%20encontraba%20en%20%24510.636.

Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis, 22*(2), 143–168.

Hammitt, J. K. (2000). Valuing mortality risk: Theory and practice. *Environmental Science & Technology, 34*(8), 1396–1400.

Hammitt, J. K., & Robinson, L. A. (2011). The income elasticity of the value per statistical life: Transferring estimates between high and low income populations. *Journal of Benefit-Cost Analysis, 2*(1).

Jin, Y., Fu, W., Kang, J., Guo, J., & Guo, J. (2019). Bayesian symbolic regression. *arXiv preprint*, arXiv:1910.08892.

Khosravi, B., Weston, A. D., Nugen, F., Mickley, J. P., Kremers, H. M., Wyles, C. C., & Taunton, M. J. (2023). Demystifying statistics and machine learning in analysis of structured tabular data. *The Journal of Arthroplasty, 38*(10), 1943–1947.

Kniesner, T. J., & Leeth, J. D. (2010). Hedonic wage equilibrium: Theory, evidence and policy. *Foundations and Trends® in Microeconomics, 5*(4), 229-299.

Kniesner, T. J., & Viscusi, W. K. (2019). The value of a statistical life. *Forthcoming, Oxford Research Encyclopedia of Economics and Finance, Vanderbilt Law Research Paper*(19-15).

Kniesner, T. J., Viscusi, W. K., & Ziliak, J. P. (2010). Policy relevant heterogeneity in the value of statistical life: New evidence from panel data quantile regressions. *Journal of Risk and Uncertainty, 40*(1), 15–31.

Kronberger, G., Burlacu, B., Kommenda, M., Winkler, S. M., & Affenzeller, M. (2024). *Symbolic Regression.* CRC Press.

La Cava, W., Orzechowski, P., Burlacu, B., de França, F. O., Virgolin, M., Jin, Y., Kommenda, M., & Moore, J. H. (2021). Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems, 2021(DB1)*, (pág. 1).

Lanoie, P., Pedro, C., & Latour, R. (1995). The value of a statistical life: a comparison of two approaches. *Journal of Risk and Uncertainty, 10(3)*, 235-257.

Looze, J. (2017). Why Do (n't) they Leave?: Motherhood and women's Job Mobility. *Social Science Research, 65*, 47–59.

Majumder, A., & Madheswaran, S. (2018). Value of statistical life in India: A hedonic wage approach. *Institute for Social and Economic Change*.

Makke, N., & Chawla, S. (2024). Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review, 57(1)*, 2.

Mardones, C., & Riquelme, M. (2018). Estimation of the value of statistical life in Chile and extrapolation to other Latin American countries. *Latin American Research Review, 53(4)*, 815-830.

Masterman, C. J., & Viscusi, W. K. (2020). Publication selection biases in stated preference estimates of the value of a statistical life . *Journal of Benefit-Cost Analysis, 11(3)*, 357-379.

Miller, T. R. (2000). Variations between countries in values of statistical life. *Journal of transport economics and policy*, 169-188.

Mincer, J. (1974). Schooling, experience, and earnings. Human behavior & social institutions no. 2.

Ministerio de Desarrollo Social y Familia. (2023). *Encuesta de Caracterización Socioeconómica Nacional (Casen) 2022*. Obtenido de Observatorio Social del Ministerio de Desarrollo Social y Familia: https://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-2022

Montgomery, D. C., & Runger, G. C. (2004). *Probabilidad y estadística aplicadas a la ingeniería* (2.a ed. ed.). Editorial Limusa.

Mundhenk, T. N., Landajuela, M., Glatt, R., Santiago, C. P., Faissol, D. M., & Petersen, B. K. (2021). Symbolic regression via neural-guided genetic programming population seeding. *arXiv preprint*, arXiv:2111.00053.

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics, 30*, 313–325.

Parada-Contzen, M., Riquelme-Won, A., & Vasquez-Lavin, F. (2012). The value of a statistical life in Chile. *Empirical Economics, 45*, 1073-1087.

Rad, H. I., Feng, J., & Iba, H. (2018). GP-RVM: Genetic programing-based symbolic regression using relevance vector machine. *arXiv preprint*, arXiv:1806.02502.

Retamal Fuentes, L. F. (2024). Estimación del valor estadístico de la vida en Chile: comparando métodos de regresión lineal y técnicas de aprendizaje estadístico.

Ricketts, J. H. (2013). Using genetic programming for symbolic regression to detect climate change signatures. *Proceedings of the 20th International Congress on Modelling and Simulation*.

Rizzi, L. I., & de Dios Ortúzar, J. (2003). Stated preference in the valuation of interurban road safety. *Accident Analysis & Prevention, 35(1)*, 9-22.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy, 82(1)*, 34-55.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160.

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International, 107*(44), 776–782.

Schnur, J. J., & Chawla, N. V. (2023). Information fusion via symbolic regression: A tutorial in the context of human health. *Information Fusion, 92*, 326-335.

Shanmugam, K. R., & Madheswaran, S. (2011). The value of statistical life. *Environmental Valuation in South Asia*, 412-437.

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. *2018 Fourth international conference on computing communication control and automation (ICCUBEA)* (págs. 1-6). IEEE.

Superintendencia de Seguridad Social. (2024). *Estadísticas de la Seguridad Social 2023*. Obtenido de Superintendencia de Seguridad Social: https://www.suseso.cl/607/w3-propertyvalue-10362.html

U.S. Bureau of Labor Statistics. (2025). *Consumer Price Index (CPI)*. Obtenido de U.S. Department of Labor: https://www.bls.gov/

Udrescu, S. M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances, 6*(16), eaay2631.

Valsaraj, P., Thumba, D. A., Asokan, K., & Kumar, K. S. (2020). Symbolic regression-based improved method for wind speed extrapolation from lower to higher altitudes for wind energy applications. *Applied Energy, 260*, 114270.

Viscusi. (1993). The Value of Risks to Life and Health. *Journal of Economic Literature, 31*, 1912-1946.

Viscusi. (2000). The value of life in legal contexts: Survey and critique. *American Law and Economics Review, 2*(1), 195–210.

Viscusi. (2015). The role of publication selection bias in estimates of the value of a statistical life. *American Journal of Health Economics, 1*(1), 27–52.

Viscusi, W. K. (2021). *Extending the domain of the value of a statistical life. Journal of Benefit-Cost Analysis* (Vol. 12(1)). 1-23.

Viscusi, W. K., & Aldy, J. E. (2003). The value of a statistical life: a critical review of market estimates throughout the world. *Journal of risk and uncertainty, 27*, 5-76.

Viscusi, W. K., & Masterman, C. J. (2017). Income elasticities and global values of a statistical life. *Journal of Benefit-Cost Analysis, 8(2)*, 226-250.

Wang, C., Zhang, Y., Wen, C., Yang, M., Lookman, T., Su, Y., & Zhang, T.-Y. (2022). Symbolic regression in materials science via dimension-synchronous-computation. *Journal of Materials Science & Technology, 122*, 77-83.

Wilstrup, C., & Kasak, J. (2021). Symbolic regression outperforms other models for small data sets. *arXiv preprint*, arXiv:2103.15147.

Zeng, P., Song, X., Lensen, A., Ou, Y., Sun, Y., Zhang, M., & Lv, J. (2023). Differentiable genetic programming for high-dimensional symbolic regression. *arXiv preprint*, arXiv:2304.08915.