Machine Learning
Trabajo Práctico 1

Nicolás Abbate, Rocío Bisang, Lucio Garay Mendez

November 20, 2023

Introducción

La evasión fiscal significa un problema relevante para el campo de la investigación económica por sus implicancias para la estabilidad macroeconómica y la desigualdad. Este problema toma particular importancia en economías donde el déficit fiscal es moneda corriente y el margen para la mejora en términos de equidad aún es amplio.

En el caso de Colombia, el déficit fiscal alcanzó en 2022, el $5,3\%$ del PIB[1] y el índice de Gini se estima fue para el mismo año de $0,475$[2], por lo que la búsqueda de mecanismos para mejorar la capacidad recaudatoria es de particular relevancia.

La estimación de los ingresos a partir de modelos construidos en base a registros administrativos es una estrategia habitual para la detección de casos de potenciales evasión. Con eso en mente, y a partir de los datos que surgen del documento "Medición de Pobreza Monetaria y Desigualdad" para Bogotá, cuya información procede de la Gran Encuesta integrada de Hogares (GEIH) para el año 2018, presentamos una serie de modelos con diferente capacidad predictiva para la estimación de los mismos.

La construcción de los siguientes modelos se realiza desde la perspectiva de un estado que busca reducir la evasión fiscal, determinando si un trabajador se encuentra subdeclarando sus ingresos. En este sentido, solo se eligieron variables observables para el Estado. Por ejemplo, las horas trabajadas, variable altamente relevante para la determinación del ingreso total de una persona, no es considera debido a que es un inobservable para el Estado. Por la misma razón, el modelo se constuye únicamente para trabajadores formales: el modelo busca estimar (con la mayor precisión posible) el ingreso "real" de los trabajadores formales, permitiendo estimar la subdeclaración de esos trabajadores. Para los informales, la elección más lógica sería considerar un modelo que los identifique y que posteriormente estime el ingreso condicional a su condición. Ese ejercicio queda postegrado en pos de estimar correctamente el ingreso de los trabajadores formales.

[1] Departamento Administrativo Nacional de Estadística (DANE)

[2] Comisión Económica para América Latina y el Caribe (CEPAL), Panorama Social de América Latina y el Caribe, 2022 (LC/PUB.2022/15-P), Santiago, 2022.

Datos

Como se mencionó en la introducción, los datos utilizados para la estimación de los distintos modelos surgen del documento "Medición de Pobreza Monetaria y Desigualdad", elaborado a partir de datos de la GEIH para el año 2018 en Bogotá. Para una mejor especificación de nuestros modelos, limitamos los datos a personas empleadas formales mayores de dieciocho años de edad.

Para la extracción de los datos se realizó un scrap de la página , a partir del código proporcionado en clase.

La limpieza de los datos se realizó en varios pasos. En primer lugar se conservaron sólo las variables potencialmente observables por el Estado de forma independiente a la encuesta. Esto incluye múltiples subgrupos de variables:

- la variable ingreso total (se considera ingreso total y no ingreso horario porque se entiende la subdeclaración horaria como un mecanismo de evasión, siendo imposible para el Estado constatar la cantidad efectiva de horas trabajadas)
- variables que se utilizarán como filtro del empleo formal (el empleo informal es inobservable para el Estado)
- variables típicas de las ecuaciones de Mincer como edad, sexo, nivel educativo, etc.
- variables que captan los efectos fijos por región
- la variable mes para controlar por estacionalidad
- múltiples variables que pueden contribuir a la predicción como cotización al Sistema de Pensiones, ingresos por accidentes, ingresos por auxilio alimentario, entre otras

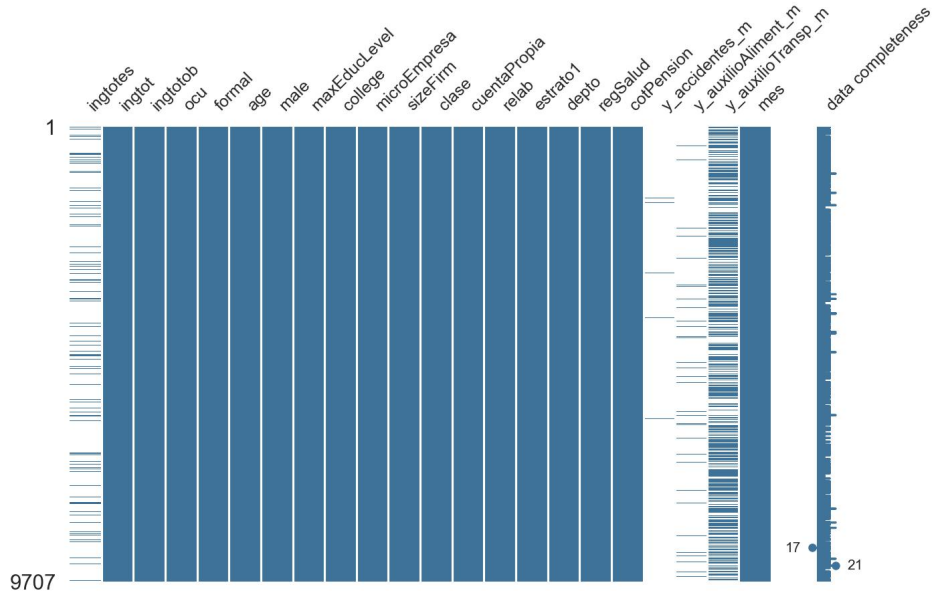
En segundo lugar, se convirtieron a numéricas las variables correspondientes. Se realizó un filtro, dejando sólo las observaciones para personas ocupadas formales mayores de 18 años y se reemplazaron las variables categóricas por binarias para convertirlas en variables dummy.

En tercer lugar, se consideraron las variables que contemplan los ingresos adicionales (auxilio de transporte, alimetario e ingresos por accidentes) y en aquellas observaciones donde se identificaron valores faltantes (NaN), se las reemplazó por cero. Se entiende que en este caso ambos son equivalentes. Para los valores fatantes de las variables binarias se generó una variable binaria extra para estimar promedio para esos valores nulos.

En este sentido, los valores faltantes se observan en la figura 1, donde las únicas variables con valores faltantes generalizados son los ingresos adicionales (las variables con prefijo "y") y el ingreso total imputado (*ingtotes*). Por esta razón, se utiliza como predictor la variable *ingtot*. Además, la selección de esta variable se refleja en la figura 2, donde se comparan los ingresos observados en relación al ingreso total calculado según la encuesta de hogares. Como se observa claramente, los ingresos observados son siempre inferiores a los finalmente imputados, en parte por ingresos indirectos como la renta, e ingresos reimputados.

Por último, se convirtieron a logarítmo las variables de ingreso a predecir. Además se incluyeron modificaciones menores de formato y limpieza, con el objetivo de obtener una base

Figure 1: Valores faltantes por variable



de datos sin variables innecesarias para el ejercicio de estimación.

Análisis descriptivo de las variables

Luego de filtrar las observaciones según los criterios explicados anteriormente la muestra resultante consta de 9707 observaciones. En cuanto a la edad media de los individuos es de 38 años en el universo que va entre 18 y 90 años. Por su parte, la mayoría de los ocupados formales cuenta con terciario como máximo nivel educativo alcanzado (56 por ciento del total), seguido de un 30 por ciento que alcanzó como máximo secundario completo. Al trabajar con una muestra de trabajadores registrados es esperable que la mayoría de ellos tenga relativamente el nivel educativo más alto.

Por su parte, la variable 'sizeFirm' muestra que la mayoría trabaja en establecimientos con más de once empleados. Las variables dummies indicativas del sector de construyeron con Obrero o Empleado de empresa particular como grupo base y estos representan un 75 por ciento de la muestra, siendo el segundo grupo más representativo el de los trabajadores por cuenta propia.

Otra de las variables que se utilizaron fueron los Estratos que indican el nivel de subsidios que recibe el individuo en su tarifa energética. El 76 por ciento de la muestra pertenece a los Estratos 2 y 3, es decir, que gran parte de la muestra recibe un nivel significativo de subsidios. Por otra parte, la información relativa a las transferencias del Estado muestra que en promedio, los ocupados formales en estudio reciben 1.499 pesos de ingresos por accidente, 4.386 por auxilio alimentario pesos y 36.371 pesos por auxilio al transporte.

Por otra parte, se observó el regimen de salud y de pensión de cada una de las observaciones,

Figure 2: Correlación entre Ingreso total y observado

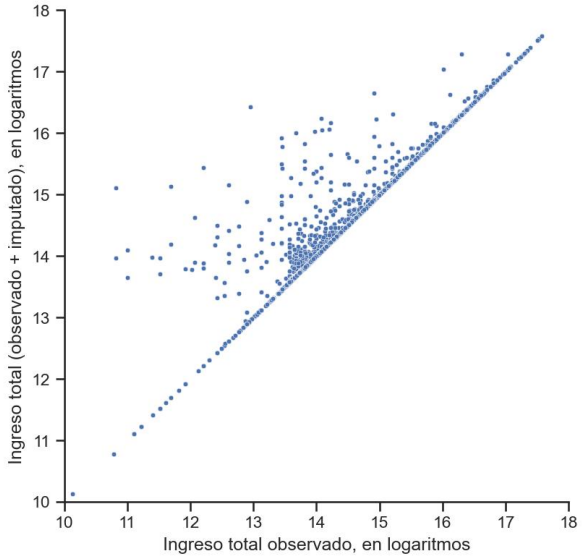


Figure 3: Correlograma de variables seleccionadas

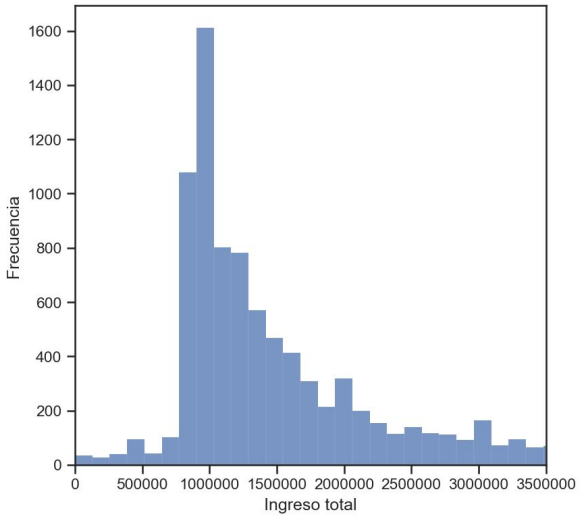


Figure 4: Variables binarias - Porcentajes

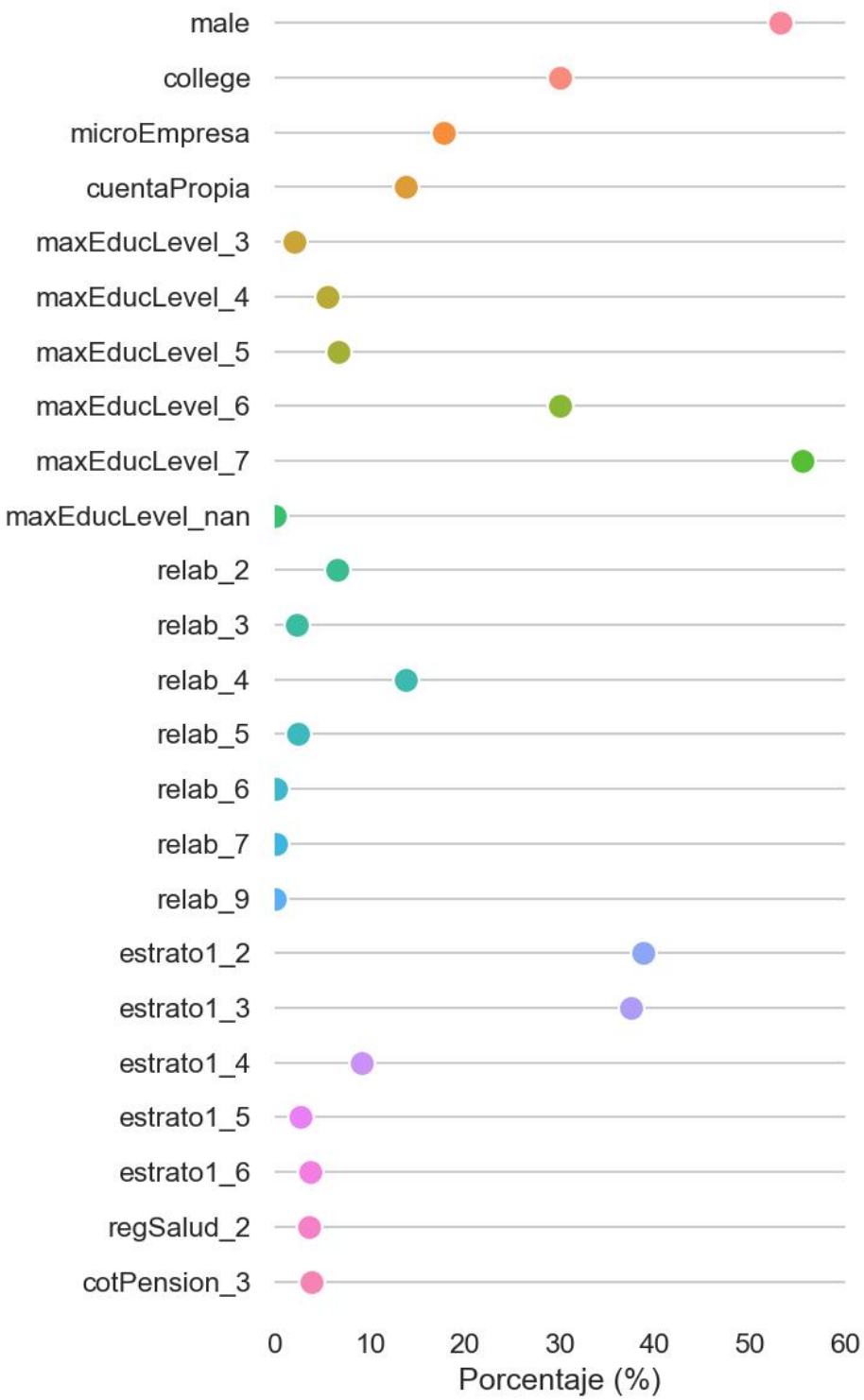


Tabla 1: Descripción de los datos

	Edad	Tamaño del establecimiento	Ingreso por accidente	Ingreso por auxilio alimentario	Ingreso por auxilio al transporte
Observaciones	9,707	9,707	9,707	9,707	9,707
Media	38	4	1,499	4,386	36,371
Desvío estándar	12	1	50,589	39,668	61,729
Mínimo	18	1	0	0	0
25%	28	4	0	0	0
50%	36	5	0	0	0
75%	47	5	0	0	88,211
Máximo	90	5	3,125,000	1,428,000	3,000,000

Tabla 2: Nivel educativo

Nivel educativo máximo alcanzado	Participación en %
Primaria incompleto	2%
Primaria completo	6%
Secundaria incompleto	7%
Secundaria completo	30%
Terciario	56%

Tabla 3: Tipo de empleo

Tipo de empleo	Participación en %
Obrero o empleado de empresa particular	75%
Obrero o empleado del Gobierno	7%
Empleado doméstico	2%
Trabajador por cuenta propia	14%
Patrón o empleador	2%
Trabajador familiar sin remuneración	0%
Trabajador sin remuneración en empresas o negocios de otros hogares	0%
Otro	0%

Tabla 4: Subsidios

Categoría de subsidios	Participación en %
Estrato 2	39%
Estrato 3	38%
Estrato 4	9%
Estrato 5	3%
Estrato 6	4%

Resumen de los ingresos			
	Ingreso total	Ingreso total imputado	Ingreso total observado
Observaciones	9,707	1,187	9,707
Media	2,350,538	1,949,997	2,112,087
Desvío estándar	3,224,866	3,085,289	3,084,789
Mínimo	0	83	0
25%	969,453	350,000	930,929
50%	1,338,211	1,000,000	1,231,544
75%	2,344,583	2,249,904	2,089,106
Máximo	85,833,330	50,000,000	85,833,330

dando como resultado que solamente el 4 por ciento esta afiliado a un regimen de salud especial y que también un 4 es pensionado.

En cuanto a la variable de interés, se observa que el ingreso promedio de los ocupados formales mayores de 18 años es de 2.350.538 pesos y que la mediana es de 1.338.211 pesos, por debajo de la media como es esperar en una distribución del ingreso típicamente sesgada a la derecha, como se puede ver en el histograma a continuación. Los valores que toma la variable ingreso alcanza como máximo a los 85,8 millones de pesos aproximadamente.

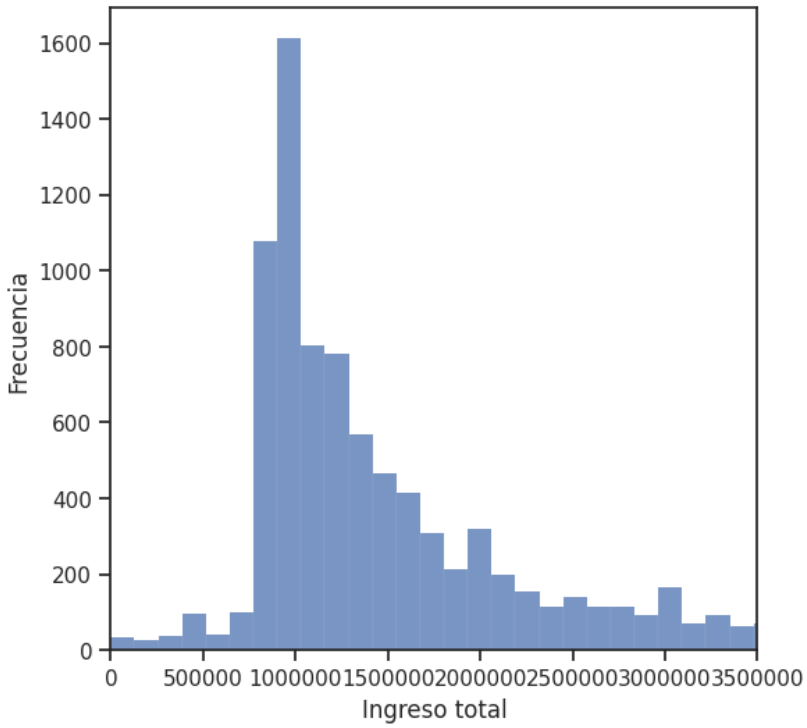
Predicción de los salarios

Para la predicción del salario total individual, se utilizaron 10 específicaciones diferentes considerando combinaciones de las covariables mencionadas en la sección previa, incorporando no linealidades, en forma de componentes polinómicos e interacciones entre variables. Las especificaciones buscan representar modelos de menor a mayor complejidad, intentando encontrar un "optimo" respecto al *trade-off* sesgo-varianza.

La especificación base (modelo 1) se trata de la media no condicionada del ingreso, ya que es el modelo predictivo de menor complejidad posible —una predicción constante para todas las personas. El modelo 2 es una especificación de Mincer básica, que tiene como variables la edad, el género y binarias de educación. El modelo 3 incorpora como variables el mes, el tamaño de la firma y binarias de universidad, micro emprendimiento, de régmien de salud y de pensiones. El modelo 4 además incorpora los ingresos por accidentes, auxilio alimentario y de transporte, binarias de tipo de relación laboral y de estrato; es decir, utiliza todas las variables descriptas en las secciones previas. El modelo 5 replica el modelo 2 considerando interacciones de grado 2 entre todas las variables, incluyendo interacciones entre las mismas. El modelo 6 replica el modelo 3, con interacciones de grado 2; y el modelo 7 replica el modelo 4 con interacciones de grado 2. Los modelos 8 y 9 replican la especificación de los modelos 3 y 4 con polinomios de grado 3, y, finalmente, el modelo 10 utiliza todas las variables con interacciones de grado 4.

La figura 7 presenta los resultados de las diferentes especificaciones descriptas. La métrica

Figure 5: Histograma de los ingresos de los ocupados formales



expuesta es la raíz cuadrada del error cuadrático medio, tanto para el conjunto de entrenamiento (70%) como del conjunto de prueba (30%). Las métricas de todos los modelos son sobre el mismo conjunto de prueba. Por un lado, en el conjunto de prueba se observa la tendencia esperable: a mayor complejidad del modelo, el poder predictivo en un principio crece, pero a partir de cierto punto —en este caso, el modelo 5 y el modelo 6— la varianza de la estimación es tan grande que el error crece, llegando a valor muy altos que salen de la escala del gráfico. Dos cuestiones más a destacar. Por un lado, en los primeros modelos el error de prueba es menor que el de entrenamiento, fenómeno posible debido a la partición aleatoria del conjunto. Por otro lado, es también destacable que a partir del modelo 8, la especificación, a pesar de crecer en complejidad. Esto puede ocurrir cuando la especificación seleccionada no se corresponde con la distribución observada de los datos, es decir, el modelo polinómico de grado alto no ajusta apropiadamente a lo observado.

En la figura 7, además, se muestra el error calculado por Leave-one-out-cross-validation (LOOCV) para los dos modelos de menor error en el conjunto de prueba original (modelos 5 y 6). Es destacable que los errores computados por LOOCV son muy similares a los del conjunto de prueba. Como conclusión general, el modelo 5, que incluye todas las variables descriptas en las secciones anteriores sin componentes polinómicos, es el que presenta la mejor performance de todos los presentados.

Figure 6: Poder predictivo de las diferetes especificaciones

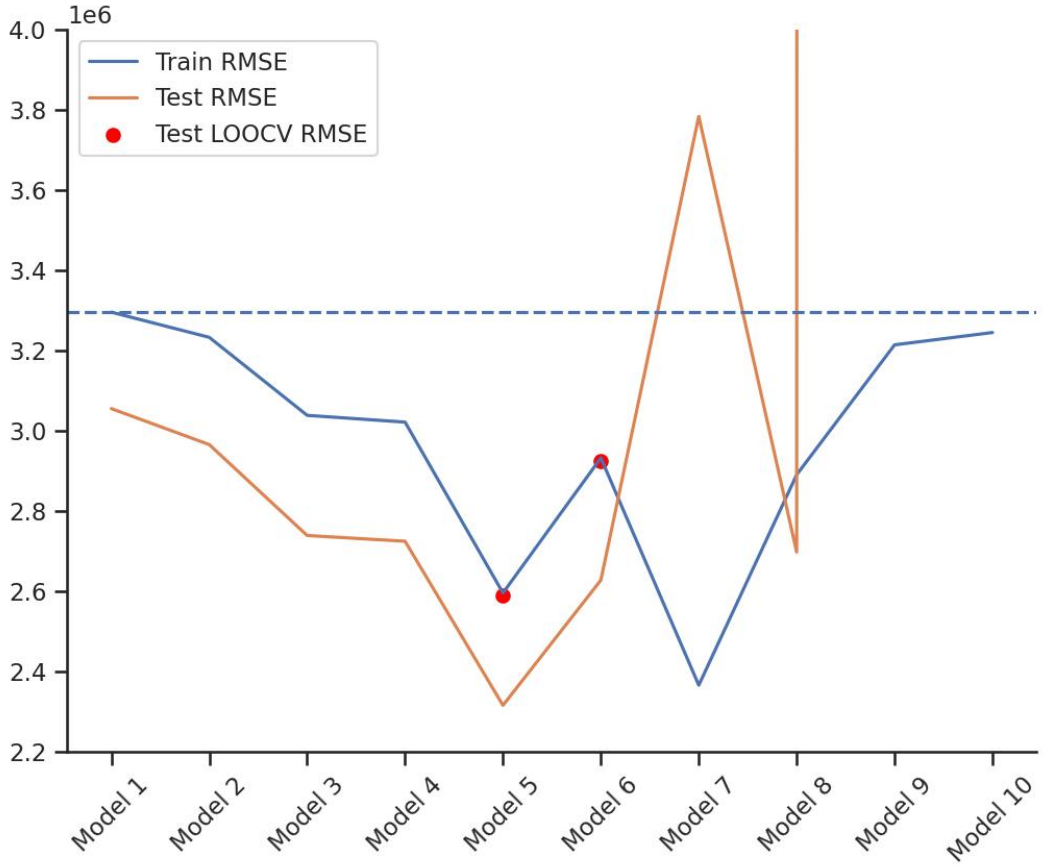
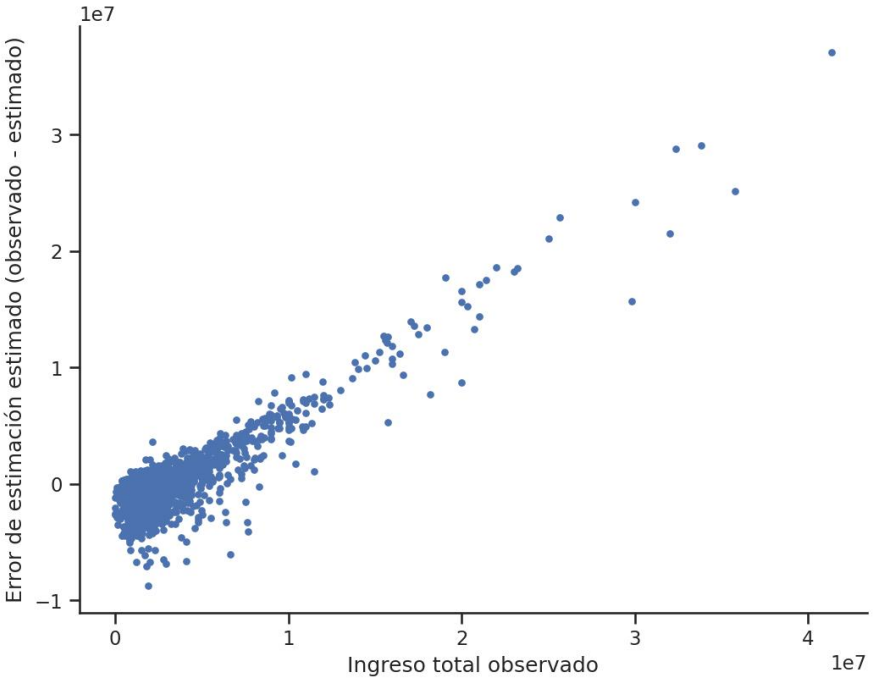


Figure 7: Error en la estimación vs ingreso observado



¿Qué tan bien predice el modelo?

Una vez estimado el modelo "óptimo", en base a las especificaciones definidas, el modelo arroja una raíz del error cuadrático medio es de aproximadamente 2.6 millones de pesos algo por encima del salario promedio. La figura **??** muestra el error de predicción en relación al salario observado. Se destacan dos patrones: por un lado, el error de estimación (como se espera de un modelo cuya función de pérdida es el Error cuadrático medio) no presenta grandes "outliers", observaciones con grandes errores de estimación. Sin embargo, emerge un patrón claro donde, en los ingresos bajos, el error de estimación es marcadamente negativo, es decir, que la estimación se encuentra por debajo del ingreso observado, mientras que para los casos de mayores ingresos el modelo subestima el ingreso. Esto es un patrón típico de los modelos de Mincer, donde, para los ingresos altos, los inobservables son muy relevantes para explicar la tendencia del modelo.