Problem Set 1: Prediciendo el Ingreso

Mario Andrés Mercado, Julian Delgado, Ana María, Juan David

March 4, 2025

1 Introduction

In the public sector, accurate reporting of individual income is critical for computing taxes. However, tax fraud of all kinds has always been a significant issue. According to the Internal Revenue Service (IRS), about 83.6% of taxes are paid voluntarily and on time in the US. [1] One of the causes of this gap is the under-reporting of incomes by individuals. An income predicting model could potentially assist in flagging cases of fraud that could lead to the reduction of the gap. Furthermore, an income prediction model can help identify vulnerable individuals and families that may need further assistance.

The objective of the problem set is to apply the concepts we learned using real world data. For that, we are going to scrape from the following website: https://ignaciomsarmiento.github.io/GEIH2018-sample/. This website contains data for Bogotá from the 2018 Medición de Pobreza Monetaria y Desigualdad Report that takes information from the GEIH

1.1 General Instructions

The main objective is to construct a model of individual hourly wages

$$w = f(X) + u \tag{1}$$

where w is the hourly wage, and X is a matrix that includes potential explanatory variables/predictors. In this problem set, we will focus on $f(X) = X\beta$.

The final document, in .pdf format, must contain the following sections:

1. *Introduction.* The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.

[1] https://www.irs.gov/newsroom/the-tax-gap

La subdeclaración de ingresos es un problema recurrente en la recaudación de impuestos, contribuyendo a la evasión fiscal. Por lo cual, se considera necesario un modelo de predicción de ingresos para detectar fraudes y asistir a personas vulnerables que necesiten focalización y desarrollo de políticas públicas. Por un lado, el nivel de ingresos en el mercado laboral es clave en la economía y en la formulación de políticas públicas, ya que influye en la distribución de la riqueza y en la calidad de vida de la población. Diversos estudios han analizado cómo factores como la edad, el género, el nivel educativo y el estrato socioeconómico afectan los ingresos de los trabajadores. Este trabajo busca explorar la relación entre estas variables y los salarios, con un énfasis particular en la estructura de los ingresos por hora. Para ello, se utiliza una amplia base de datos representativa de trabajadores en Bogotá, la cual permite examinar patrones de ingreso en función de la edad, el género, la educación, el estrato socioeconómico y la formalidad. Se emplean diversas herramientas estadísticas y visualizaciones para identificar tendencias y disparidades en los ingresos, y en particular, se modela el salario por hora en términos logarítmicos para evaluar la relación no lineal con la edad, considerando efectos decrecientes del salario a medida que aumenta la experiencia laboral. El estudio confirma la importancia de la educación y la experiencia en la determinación de los salarios y resalta las desigualdades estructurales existents en el mercado laboral. Estos hallazgos tienen implicaciones importantes para el diseño de políticas públicas orientadas a la reducción de la desigualdad y la mejora de las condiciones laborales. Por el otro lado, los modelos estadísticos buscan estimar cuánto debería ganar un individuo en función de variables como edad, educación, tipo de empleo y sector económico. Si el ingreso declarado por una persona es significativamente menor al valor predicho por el modelo, podría ser un indicio de subdeclaración de ingresos. De este modo, los contribuyentes que reportan ingresos inusualmente bajos dentro de su grupo de

referencia pueden ser marcados para auditoría (outliers). Con ayuda de modelos de predicción, análisis de outliers, comparación con bases de datos externas y modelos de machine learning se logra detectar el fraude fiscal.

2. *Data.*[2] We will use data for Bogotá from the 2018 Medición de Pobreza Monetaria y Desigualdad Report that takes information from the GEIH The data set contains all individuals sampled in Bogota and is available at the following website https://ignaciomsarmiento.github.io/GEIH2018-sample/. To obtain the data, you must scrape the website. In this problem set, we will focus only on employed individuals older than eighteen (18) years old. Restrict the data to these individuals and perform a descriptive analysis of the variables used in the problem set. Keep in mind that in the data, there are many observations with missing data or 0 wages. I leave it to you to find a way to handle this data. When writing this section up, you must:

   (a) Describe the data briefly, including its purpose, and any other relevant information.

   Una vez se combinan los datos y se obtiene la base completa, se evidencia que hay información que nos permite analizar el mercado laboral y variables ecónomicas de diferentes individuos. Se encuentra que el objetivo principal de la base de datos es analizar ingresos (salario) y comportamientos del mercado. Hay variables que nos permiten analizar características individuales como la edad, género, educación, departamento, entre otras y variables que se relacionan con la actividad laboral como los ingresos, horas trabajadas, tipo de empleo, sector económico, entre otras. La selección de las variables es clave puesto que

[2]This section is located here so the reader can understand your work, but probably it should be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here

permitirá evaluar los factores que influyen en el ingreso y analizar fenómenos como la brecha salarial de género. De igual forma, es importante mencionar que la base tiene valores faltantes y observaciones con ingreso igual a cero, lo que sugiere la necesidad de un proceso de limpieza antes del análisis. La base tiene una estructura de formato tabular, la cual facilita el procesamiento de los datos y el análisis de los mismos mediante descripciones estadísticas. Adicionalmente, se considera que los datos proporcionados son fundamentales para construir un modelo predictivo de ingresos por hora y detectar como variables sociodemográficas y laborales inciden en la remuneración de los individuos.

(b) Describe the process of acquiring the data and if there are any restrictions to accessing/scraping these data.

En primer lugar hay que tener presente que los datos no se pueden descargar como un archivo, sino que se encuentran en un página web https://ignaciomsarmiento.github.io/GEIH2018_sample/ por lo cual, es necesario hacer web scrapping para extraerlos. Los datos estaban distribuidos en diferentes sitios web, por lo se combinaron en un único conjunto de datos. De esta forma, se difinió la base de la URL, donde cada página se nombró siguiendo un patrón (geih_page_1.html, geih_page_2.html ..., etc). Con ayuda de la libreria rvest se leyó el contenido HTML de cada página (data chunk 1:10), se extrajeron las tablas y se combinaron los datos en un solo dataframe, almacenando el resultado en la variables datos_totales. Para evitar que el código fuese a generar error en caso de que una de las páginas no cargara correctamente, se utilizó tryCatch (manejo de errores). De esta forma, se buscaba que si las páginas cargaban correctamente, se extrayera las tablas y se añadieran a la variable datos_totales y si se llegaba a presentar algún problema de lectura, se captaría el error sin detener la ejecución

total. Por lo tanto, el código permite automatizar el proceso de recolección de datos que están distribuidores en diferentes sitios web y cuando se trabaja con grandes volúmenes de datos, aún si se presenta algún problema. No se encuentran restricciones dado que la página no tiene capchas ni bloqueos para la extracción de datos, pero tuvimos presente no hacer múltiples solicitudes en un corto período de tiempo para no sobrecargar el servidor. Por el otro lado, los datos son públicos y fueron diseñados para fines académicos, por lo que facilito en gran medida el scrapping.

(c) Describe the data cleaning process and

En primer lugar, se consideró necesario renombrar ciertas variables como edad, sexo, ingreso total para facilidad de lectura del grupo y se creó una variable ln_Salario_hora, a la cual se le aplicó logarítmo natural para facilitar el análisis de regresión posterior. De acuerdo al enunciado del problem set, se identificó que debíamos filtrar la base por aquellos individuos empleados mayores de 18 años. Por lo cual, se filtraron los datos para las muestras mayores a 18 años, donde el ingreso total fuese mayor a cero y tanto las horas trabajadas como el salario por hora fuese mayor a cero. Adicionalmente, se identificaron y eliminaron outliers extremos para las variables de horas trabajadas y salario por hora usando percentiles 1% y 99% con el fin de que dichos valores atípicos distorsionen las interpretaciones de los resultados. Posteriormente, se seleccionaron únicamente variables númericas y se elimnaron aquellas con desviación estándar igual a cero, puesto que no aporta variabilidad al modelo. Asimismo, las variables NA dado que no tienen información útil.

Teniendo en cuenta que la base completa era bastante amplia, se decidió calcular la correlación de cada variable numérica con Salario_hora y de esta forma, se se-

leccionaron las 25 variables con mayor correlación puesto que serían las más relevantes para analizar. Adicionalmente, se consideraron otras variables como la edad, el género, etc dado que estas variables han sido ampliamente estudiadas en investigaciones económicas y del mercado laboral, demostrando efectos significativos en los ingresos. De este modo, las variables relevantes para el análisis son de características individuales (edad, genero, educación (college, maxEducLeval)), características laborales (oficio, informalidad, horas trabajadas), características económicas (ingresos, estrato) y así se crea un nuevo dataframe con solo estas variables para reducir la dimensionalidad de la base.

(d) Descriptive the variables included in your analysis. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a dry list of ingredients.

La tabla 1 corresponde a las estadísticas descriptivas de las variables seleccionadas aquellas con una correlación significativa y aquellas ampliamente estudiadas). En primer lugar, se analizan las variables demográficas. Se evidencia que los individuos están entre los 19 y los 86 años de edad, con una media de 36.3 años lo que sugiere que la muestra poblacional está en edad laboral. La media de la variable de género (hombre) es de 0.5038, lo que indica que hay una distribución equilibrada entre hombres y mujeres. A partir de la variable maxEducLevel se evidencia que la mayoría de la población tiene al menos educación secundaria o técnica, dado que la media es 6 y a partir de college se evidencia que el alrededor del 35% de la muestra ha alcanzado la educación superior (universitaria), dado que la media es de 0.3474. De esta forma, se puede concluir

que la muestra incluye una población trabajadora en su mayoría joven-adulta, con igualdad de representación entre hombres y mujeres y un nivel educativo medio-alto.

En segundo lugar, se analizan las variables laborales y se encuentra a partir de la variable informal que el 21.41 % de la muestra está en el sector informal (informal = 0.2141), mientras que el 78.59% está en el sector formal (formal = 0.7859). Adicionalmente, se encuentra por medio de la variable hoursWorkUsual que la mayoría trabaja jornadas laborales completas puesto que la media es de 48.13 horas y que no hay individuos independientes dado que la media de la variable cuenta propia es cero.

En tercer lugar, se analizan las variables económicas y se encuentra la variable de Salario por hora, la cual tiene un rango de 1,215 a 58,333, con una mediana de 4,555 y una media de 7,291.La gran diferencia entre media y mediana sugiere que hay valores extremos que elevan el promedio. La variable de ingreso total(ingtotes) muestra una gran dispersión en los valores, con un máximo de 8 millones. La media (434,783) es significativamente mayor a la mediana (191,834), indicando la presencia de ingresos muy elevados en la parte alta de la distribución y, por medio de Logaritmo del salario (ln_Salario_hora) se evidencia que los ingresos siguen una distribución normal. Es decir, los ingresos presentan una fuerte dispersión y una distribución sesgada con valores muy altos en la parte superior. Esto refuerza la necesidad de usar el logaritmo del salario en los modelos de regresión para reducir heterocedasticidad.

Finalmente, se analizan las variables sociales y se encuentra a partir de la variable de estrato socioecon ómico que la población está en estratos medios-bajos pues la media es de 2.491. Sin embargo, se encuentra un máximo de 6, lo que refleja la diversidad de condiciones socioeconómicas en la muestra. Por medio de

la variable (cotPension), se evidencia que la mayoría de los trabajadores están afiliados a pensión pues la media de 1.232 indica que más del 50% cotiza. De esta forma, el acceso a la seguridad social está presente en una parte importante de la población, pero es posible que una proporción significativa no cotice. La distribución de estratos refleja la heterogeneidad económica en Bogotá. En conclusión, la población trabajadora se compone principalmente por adultos jóvenes con una distribución casi equitativa entre hombres y mujeres. La mayoría de los trabajadores están en el sector formal y con jornadas laborales completas, aunque hay presencia de informalidad y posibles casos de sobreempleo. Los ingresos presentan una alta dispersión, con valores extremos que sesgan la media hacia valores elevados.Existe una diversidad en niveles educativos y estratos socioeconómicos, lo que puede ser clave en la modelización del salario.

Para completar el análisis de las estadísticas descriptivas, se consideró pertinente ver la relación gráfica entre la variable salario y otras variables de interés como las mencionadas anteriormente. La primera relación que encontramos es la de la Figura 1. Age-Earnings Profile, en el que se muestra la relación entre la edad y el salario por hora estimado. Se observa una curva en forma de U cóncava, lo que sugiere que los ingresos tienden a aumentar con la edad hasta un punto máximo alrededor de los 45-50 años y luego disminuyen progresivamente. Esto puede explicarse por la acumulación de experiencia y educación en la primera mitad de la vida laboral, seguida por una posible reducción en la intensidad del trabajo, jubilación o cambios en el tipo de empleo en una edad mayor.

En la Figura 2 Ingresos promedio por grupo de edad y género, se evidencian los ingresos promedio por grupo de edad y se diferencia entre hombres (azul) y mujeres (rojo). Se analiza que en la mayoría de los grupos de edad, los hombres tienden a tener ingresos más altos que las mujeres, aunque en algunos rangos

(como 40-50 años) la diferencia es menor. Hay una tendencia similar al primer gráfico, donde los ingresos crecen con la edad hasta un punto máximo entre 40-60 años y luego disminuyen en el grupo de mayores de 70 años. El grupo de menores de 20 años, tiene los ingresos más bajos, lo cual tiene sentido puesto que están en etapas de educación o comenzando la carrera laboral. Ambos gráficos reflejan una relación clara entre la edad y los ingresos, con un punto máximo alrededor de los 40-50 años y una posterior disminución. También se observa una brecha de género en los ingresos, con los hombres ganando más en la mayoría de las edades. Esto puede estar influenciado por diferencias en acceso a empleos mejor remunerados, interrupciones laborales en el caso de las mujeres (embarazo) y otros factores estructurales del mercado laboral.

En la Figura 3 Ingresos promedio por estrato se muestra la relación entre el estrato socioeconómico y el ingreso promedio de la muestra poblacional. Se observa una relación positiva entre estrato e ingresos, puesto que a medida que el estrato aumenta, el ingreso promedio también crece. Esto es posible, ya que los estratos más altos suelen tener acceso a mejores oportunidades económicas y educativas. Sin embargo, se evidencia una gran diferencia entre el estrato 1 (el más bajo) y los estratos superiores. Por ejemplo, el estrato 1 tiene ingresos muy bajos en comparación con el estrato 2, que experimenta un fuerte incremento.Los estratos 5 y 6 tienen ingresos similares, por lo que no se observa un aumento significativo entre estos dos estrato. Esto indica que, a partir de cierto punto, el crecimiento del ingreso se estabiliza.Este gráfico refleja la desigualdad económica y cómo el acceso a recursos (vivienda, educación) influye en los ingresos.

La Figura 4 Ingresos promedio por nivel de educación muestra cómo el nivel educativo impacta en el ingreso promedio, diferenciando entre hombres (azul) y mujeres (rojo).Se evidencia una clara relación positiva entre el nivel educativo

y el ingreso promedio. Aquellos con mayor nivel de estudios obtienen ingresos significativamente más altos. Asimismo, se evidencia que en los niveles educativos más bajos, las diferencias salariales entre hombres y mujeres se perciben, mientras que a medida que aumenta el nivel educativo, las diferencias de género se reducen, especialmente en los niveles más altos donde los ingresos parecen igualarse. El salto en ingresos entre los niveles educativos bajos y altos es notable. La diferencia entre los que tienen poca educación y los que alcanzan niveles universitarios es muy marcada.Este gráfico refuerza la importancia de la educación como un determinante clave de los ingresos y sugiere que una mayor escolaridad puede ser una vÃa para mejorar las condiciones económicas.

En la Figura 6 Salario-hora por horas de trabajo se muestra la relación entre las horas de trabajo (eje X) y el salario por hora (eje Y), diferenciando entre trabajadores formales (azul) e informales (rojo).Principalmente, se observa que los trabajadores formales tienden a tener salarios por hora más altos, con una mayor dispersión de valores en la parte superior del gráfico, mientras que los trabajadores informales se concentran en los niveles más bajos de salario por hora, con pocos casos de ingresos altos. Asimismo, se observa que la mayoría de los trabajadores se agrupan entre 30 y 60 horas semanales. No hay una clara tendencia de aumento del salario por hora con más horas trabajadas, lo que indica que la cantidad de horas no garantiza mejores ingresos. Finalmente, se observa que hay trabajadores informales con techo de ingresos dado que casi todos los puntos rojos están en la parte baja del gráfico, lo que indica que presentan limitaciones salariales y no alcanzan los mismo niveles salariales que los trabajadores formales.Por último, se analiza

la Figura 9 Boxplot promedio de ingresos por estrato, la cual muestra la distribución de los ingresos promedio en función del estrato socioeconómico. Se

encuentra una tendencia creciente de ingresos con respecto al estrato. Los estratos más bajos (1, 2 y 3) tienen ingresos considerablemente menores y con menor dispersión, mientras que los estratos altos (4, 5 y 6) presentan ingresos más elevados y una mayor variabilidad. Por otro lado, se encuentra que en los estratos 1, 2 y 3, los ingresos están muy concentrados en valores bajos, con algunas excepciones representadas por los outliers. A partir del estrato 4, la mediana de los ingresos aumenta notablemente y la dispersión de los datos es mucho mayor. El estrato 6 tiene la distribución más amplia, lo que indica una mayor diferencia entre quienes ganan poco y quienes ganan mucho dentro de este grupo. Es importante mencionar que se observan outliers en todos los estratos, pero en los más bajos estos outliers son menos frecuentes y no tan extremos, mientras que en los estratos más altos, la presencia de valores atípicos con ingresos significativamente altos sugiere que hay grupos reducidos de personas con ingresos muy elevados.

De esta manera, se puede concluir que hay una fuerte relación entre variables socioeconómicas y los ingresos de la población. Se observa que el nivel de ingresos aumenta con la edad hasta cierto punto, mostrando un pico, con diferencias notables entre hombres y mujeres. Asimismo, el estrato socioeconómico y el nivel educativo tienen un impacto significativo en los ingresos, donde los estratos altos y las personas con mayor nivel educativo perciben salarios mÃ¡s elevados. Además, el mercado laboral presenta disparidades entre el empleo formal e informal, con trabajadores formales obteniendo mayores salarios por hora, aunque con una variabilidad considerable. Finalmente, la distribución de ingresos refleja altos niveles de desigualdad, especialmente en los estratos más altos, donde hay una mayor dispersión y presencia de outliers con ingresos significativamente altos. Estos resultados demuestran la importancia de desarrollar

polÃticas públicas que promuevan el acceso a la educación y la formalización del empleo como estrategias clave para reducir las brechas económicas y laborales.

3. *Age-wage profile.* A great deal of evidence in labor economics suggests that the typical workers age-wage profile has a predictable path: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50

   In this subsection we are going to estimate the Age-wage profile profile for the individuals in this sample:

$$log(w_i) = \beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + u \tag{2}$$

   (a) Regression table

   Ver 2

   (b) Interpretation

   β_1 *Intercept*. Represents the baseline value of log(wage) when Age is 0. β_2 *Age* Coefficient. Represents the change in log(wage) for each additional year of age. β_3 Age^2 Coefficient. Captures the non-linear effect of age on wages, accounting for the curvilinear relationship.

   (c) Discussion

   La primera tabla muestra la estimación de un modelo de regresión para explicar el logaritmo del salario por hora (ln_salario_hora) en función de la edad y la edad al cuadrado. En primer lugar, se analiza el coeficiente edad, donde edad = 0.054, p¡0.01. Se encuentra que por cada año adicional de edad, el salario por hora aumenta 5,4%, manteniendo constantes las demás variables. El error estándar es pequeño (0.003), lo que sugiere que la estimación es precisa. Dado

que el coeficiente es estadísticamente significativo a un nivel del 1% (p¡0.01), hay fuerte evidencia de que la edad tiene un impacto positivo en los salarios por hora. En segundo lugar, se analiza la edad al cuadrado, donde edad_squared = -0.001, p¡0.01. El coeficiente negativo sugiere que, a medida que aumenta la edad, el crecimiento del salario por hora se desacelera. Es decir, la relación entre edad e ingresos sigue una curva cuadrática (cóncava), lo que indica que hay un punto máximo a partir del cual los salarios dejan de crecer y comienzan a disminuir. También es significativo a un nivel del 1% (p¡0.01), lo que confirma que esta relación cuadrática es relevante en el modelo. La constante igual a 7.527 representa el valor de ln_salario_hora cuando edad = 0. Su valor elevado sugiere que, en términos logarítmicos, los salarios de referencia sin considerar la edad son relativamente altos.

Por otro lado, se considera necesario analizar los indicadores estadísticos del modelo. Por ejemplo, observamos que la muestra usada en la estimación del modelo incluye 9423 individuos. El R^2 indica que el modelo solo explica el 3.5% (0.035) de la variabilidad en los salarios por hora, lo cual nos indica que hay otros factores importantes que determinan el salario y que no están incluidos en el modelo como educación, experiencia o sector de empleo. El error estándar residual (0.634) muestra la dispersión de los errores del modelo, indicando qué tan lejos están los valores reales de los valores predichos en promedio. Finalmente, se logra concluir que el modelo es globalmente significativo dado que valor p del Estadístico F es menos a 0.01 se logra rechazar la hipótesis nula de que todos los coeficientes son iguales a cero.

Siendo así, el modelo sugiere que el salario por hora aumenta con la edad, pero a un ritmo decreciente, alcanzando un punto máximo antes de empezar a disminuir. Sin embargo, el bajo R^2 indica que la edad por sśola no es un buen

predictor del salario y que se necesitan otras variables (como nivel educativo, experiencia, género, sector laboral, etc.) para explicar mejor la variabilidad en los ingresos.

(d) plot of the estimated age-earnings

ver figura 2

4. *The gender earnings GAP.* Policymakers have long been concerned with the gender wage gap, and is going to be our focus in this subsection.

(a) Begin by estimating and discussing the unconditional wage gap:

$$log(w_i) = \beta_1 + \beta_2 Female_i + u \tag{3}$$

where F emale is an indicator that takes one if the individual in the sample is identified as female.

Vease table (4)

(b) *Equal Pay for Equal Work?* A common slogan is equal pay for equal work. One way to interpret this is that for employees with similar worker and job characteristics, no gender wage gap should exist. Estimate a conditional earnings gap incorporating control variables such as similar worker and job characteristics. In this section, estimate the conditional wage gap: First, using FWL Second, using FWL with bootstrap. Compare the estimates and the standard errors.

$$log(w_i) = \beta_1 + \beta_2 Female_i + X\gamma + u \tag{4}$$

(c) Next, plot the predicted age-wage profile and estimate the implied peak ages with the respective confidence intervals by gender.

Vease figura (12)

Vease tabla (3)

Las variables de control son escolaridad, edad, $edad^2$, cuentapropia, jefe de hogar, formal y ocupado. Las variables explicativas que pueden ser malos controles, son formal, ocupado o cuentapropista, la explicación se debe a que estas variables pueden estar correlacionas entre ellas, al igual que con la variable predictora (logaritmo de salarios), es el caso de ser formal, esta posiblemente esté fuertemente relacionada con la variable dependiente, o lo que es lo mismo presentar problemas de endogeneidad (econometría clásica).

Interpretación de los coeficientes:

Para el modelo de brecha salarial incondicional, observamos que el coeficiente es de -0.028 y es estadisticamente significativo al 95% de confianza. Este quiere decir que la mujer gana en promedio 2.8 pesos menos con respecto al hombre, por lo que la brecha salarial es considerable. Si se incluyen las variables de control, este valor se incrementa y llega a -0.067 es estadisticamente significativo al 99% de confianza. En este caso, la brecha se incrementa un poco y ahora la mujer gana en promedio 6.7 menos pesos que el hombre evidenciando asÃ la discriminación en el mercado de trabajo.

5. *Predicting earnings.* In the previous sections, you estimated some specifications with inference in mind. In this subsection, we will evaluate the predictive power of these specifications.

   (a) Split the sample into two: a training (70%) and a testing (30%) sample. (Dont forget to set a seed to achieve reproducibility. In R, for example you can use set.seed(10101), where 10101 is the seed.)

(b) Report and compare the predictive performance in terms of the RMSE of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity.

ver figuras (7, 5 y 6)

(c) In your discussion of the results, comment: About the overall performance of the models. About the specification with the lowest prediction error. For the specification with the lowest prediction error, explore those observations that seem to miss the mark. To do so, compute the prediction errors in the test sample, and examine its distribution. Are there any observations in the tails of the prediction error distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

- Desempeño de los modelos: En general, los modelos presentan un desempeño relativamente bueno, con RMSE que oscilan entre 0.626 y 0.634. Dado que ahora solo usamos las variables edad y género (mujer), la capacidad predictiva ha sido más conservadora en comparación con los modelos anteriores, lo que sugiere que la eliminación de variables como escolaridad ha reducido la complejidad del modelo, pero también su precisión.
- Modelos con mejor desempeño: El modelo con el menor RMSE es el Modelo 2.5 (0.626), seguido muy de cerca por el Modelo 2.3 (0.627). Estos modelos incluyen combinaciones no lineales de edad (como su cuadrado) e interacciones con género. Esto reafirma que, aunque las variables predictivas son limitadas, la inclusión de términos no lineales sigue aportando valor, al capturar relaciones complejas entre edad y salario.
- Distribución de errores: La distribución de los errores de predicción muestra una forma aproximadamente normal, con una fuerte concentración alrede-

dor de 0. Esto indica que, en general, los modelos tienden a predecir los salarios con una precisión razonable, sin sesgo sistemático claro.

- Asimetría positiva: Hay una asimetría positiva evidente, con valores más alejados hacia la derecha. Esto sugiere que algunos salarios fueron significativamente subestimados por los modelos, lo cual podría reflejar que existen individuos con ingresos muy altos que las variables edad y género no logran explicar adecuadamente.
- Outliers: El análisis de outliers (percentiles 1% y 99%) revela que hay observaciones extremas con errores considerables. Estos outliers podrían estar relacionados con personas que reciben salarios inusualmente bajos o altos, potencialmente debido a factores no contemplados por los modelos, como la ocupación, el nivel educativo o experiencia laboral.

(d) LOOCV. For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-out-cross-validation (LOOCV). Compare the results of the test error with those obtained with the validation set approach and explore the potential links with the influence statistic. (Note: when attempting this subsection, the calculations can take a long time, depending on your coding skills, plan accordingly!)

Tras el análisis realizado, se obtuvieron los siguientes resultados en cuanto al desempeño predictivo de los modelos estimados:

Desempeño en el conjunto de validación: El Modelo 2.5 presentó el menor RMSE, con un valor de 0.626. El Modelo 2.3 le siguió de cerca, con un RMSE de 0.627. Desempeño con validación cruzada (LOOCV): El RMSE LOOCV para el Modelo 2.5 fue de 0.6331. El RMSE LOOCV para el Modelo 2.3 fue de 0.6340.

Ambos modelos incluyen combinaciones no lineales de la variable edad, como su término cuadrático, y consideran las interacciones con el género (mujer), lo que permite capturar relaciones complejas entre estas variables y el salario por hora. A pesar de que el Modelo 2.5 obtuvo el menor RMSE tanto en validación como en LOOCV, la diferencia con el Modelo 2.3 es marginal. Esto sugiere que ambos modelos tienen un desempeño muy similar y que las mejoras adicionales logradas por el Modelo 2.5 son leves, aunque consistentes. El hecho de que las puntuaciones de RMSE en validación y LOOCV no difieran significativamente indica que los modelos no están sobreajustados y que su capacidad predictiva es relativamente estable al enfrentarse a nuevos datos.

Sin embargo, es importante destacar que, dado que solo se utilizaron las variables edad y género, las predicciones aún presentan limitaciones. La presencia de outliers y la asimetría positiva en los errores sugieren que hay factores adicionales como la escolaridad, la ocupación o la experiencia laboral que no fueron considerados y que podrían ayudar a mejorar el poder explicativo de los modelos.

En conclusión, el Modelo 2.5 se posiciona como el mejor modelo predictivo dentro de los estimados, aunque el Modelo 2.3 ofrece resultados muy similares. Para futuras investigaciones, se recomienda explorar la inclusión de nuevas variables y el uso de técnicas más avanzadas para robustecer las predicciones.

Tables and Figures

Table 1. *Estadísticas descriptivas*

	edad	hombre	college	maxEducLevel	informal	formal	r
X	Min. :19.0	0:4676	Min. :0.0000	Min. :1.000	Min. :0.0000	Min. :0.0000	M
X.1	1st Qu.:27.0	1:4747	1st Qu.:0.0000	1st Qu.:6.000	1st Qu.:0.0000	1st Qu.:1.0000	1
X.2	Median :34.0		Median :0.0000	Median :6.000	Median :0.0000	Median :1.0000	M
X.3	Mean :36.3		Mean :0.3474	Mean :6.104	Mean :0.2141	Mean :0.7859	M
X.4	3rd Qu.:45.0		3rd Qu.:1.0000	3rd Qu.:7.000	3rd Qu.:0.0000	3rd Qu.:1.0000	3
X.5	Max. :86.0		Max. :1.0000	Max. :7.000	Max. :1.0000	Max. :1.0000	M
X.6				NA's :1			

Table 2. *Estimación del modelo de salarios*

	Dependent variable:
	ln_salario_hora
edad	0.054***
	(0.003)
edad_squared	−0.001***
	(0.00004)
Constant	7.527***
	(0.065)
Observations	9,423
R^2	0.035
Adjusted R^2	0.035
Residual Std. Error	0.634 (df = 9420)
F Statistic	171.668*** (df = 2; 9420)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 1. Age-Earnings Profile

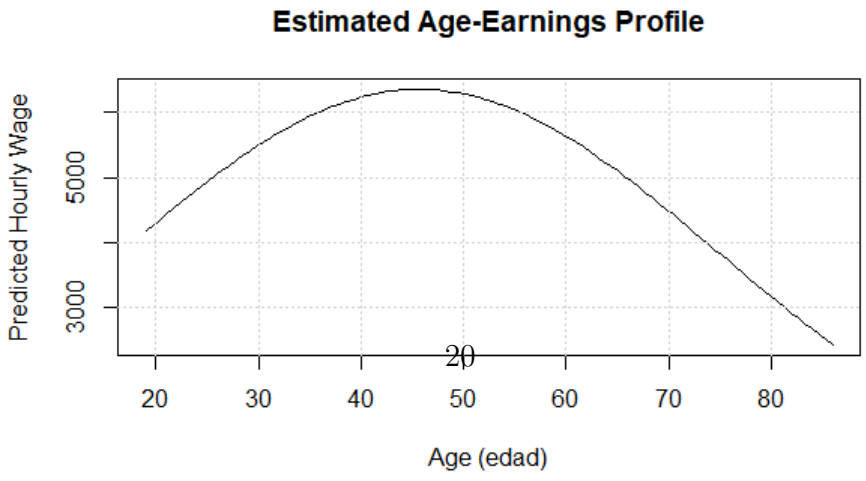


Table 3. *Estimación del modelo de brecha salarial incondicional*

	Dependent variable:
	ln_salario_hora
mujer	-0.028^{**}
	(0.013)
Constant	8.636^{***}
	(0.009)
Observations	9,423
R^2	0.0005
Adjusted R^2	0.0004
Residual Std. Error	0.645 (df = 9421)
F Statistic	4.400^{**} (df = 1; 9421)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4. *Estimación del modelo de brecha salarial incondicional vs condicional*

	Dependent variable:	
	ln_salario_hora	
	(1)	(2)
mujer	-0.028^{**}	-0.067^{***}
	(0.013)	(0.012)
Constant	8.636^{***}	5.854^{***}
	(0.009)	(0.066)
Observations	9,423	9,422
R^2	0.0005	0.293
Adjusted R^2	0.0004	0.293
Residual Std. Error	0.645 (df = 9421)	0.543 (df = 9415)
F Statistic	4.400^{**} (df = 1; 9421)	650.314^{***} (df = 6; 9415)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5. *Submodelos del modelo 1 de salarios*

	Dependent variable:				
	ln_salario_hora				
	(1)	(2)	(3)	(4)	(5)
edad	0.133***	0.057***	0.133***		0.058***
	(0.016)	(0.004)	(0.016)		(0.004)
I(edadˆ2)	−0.003***	−0.001***	−0.003***		−0.001***
	(0.0004)	(0.0001)	(0.0004)		(0.0001)
I(edadˆ3)	0.00002***				
	(0.00000)				
mujer		−0.032**			
		(0.016)			
edad:I(edadˆ2)			0.00002***		
			(0.00000)		
poly(edad, 3)1				6.961***	
				(0.635)	
poly(edad, 3)2				−7.691***	
				(0.635)	
poly(edad, 3)3				3.160***	
				(0.635)	
edad:mujer					−0.001***
					(0.0004)
Constant	6.533***	7.483***	6.533***	8.623***	7.464***
	(0.204)	(0.078)	(0.204)	(0.008)	(0.078)
Observations	6,596	6,596	6,596	6,596	6,596
R^2	0.042	0.039	0.042	0.042	0.040
Adjusted R^2	0.042	0.039	0.042	0.042	0.040
Residual Std. Error (df = 6592)	0.635	0.636	0.635	0.635	0.636
F Statistic (df = 3; 6592)	97.148***	90.050***	97.148***	97.148***	91.853***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 6. *Submodelos del modelo 2 de salarios*

	Dependent variable:		
	ln_salario_hora		
	(1)	(2)	(3)
mujer	−0.032**	0.131**	−0.032**
	(0.016)	(0.051)	(0.016)
edad	0.057***	0.009***	
	(0.004)	(0.001)	
I(edadˆ2)	−0.001***		
	(0.0001)		
mujer:edad		−0.004***	
		(0.001)	
poly(edad, 3)1			6.967***
			(0.635)
poly(edad, 3)2			−7.730***
			(0.635)
poly(edad, 3)3			3.157***
			(0.635)
poly(edad, 4)1			
poly(edad, 4)2			
poly(edad, 4)3			
poly(edad, 4)4			
Constant	7.483***	8.297***	8.638***
	(0.078)	(0.035)	(0.011)
Observations	6,596	6,596	6,596
R^2	0.039	0.019	0.043
Adjusted R^2	0.039	0.019	0.042
Residual Std. Error	0.636 (df = 6592)	0.643 (df = 6592)	0.635 (df = 6591)
F Statistic	90.050*** (df = 3; 6592)	43.546*** (df = 3; 6592)	73.960*** (df = 4; 6591)

Note:

Table 7. *Modelo 1 y 2 de salarios*

	Dependent variable:	
	ln_salario_hora	
	(1)	(2)
mujer		-0.027^{*}
		(0.016)
edad	0.056^{***}	0.007^{***}
	(0.004)	(0.001)
I(edadˆ2)	-0.001^{***}	
	(0.0001)	
Constant	7.472^{***}	8.372^{***}
	(0.078)	(0.027)
Observations	6,596	6,596
R^2	0.039	0.018
Adjusted R^2	0.038	0.018
Residual Std. Error (df = 6593)	0.636	0.643
F Statistic (df = 2; 6593)	132.870^{***}	59.996^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 2. Ingresos promedio por grupo de edad

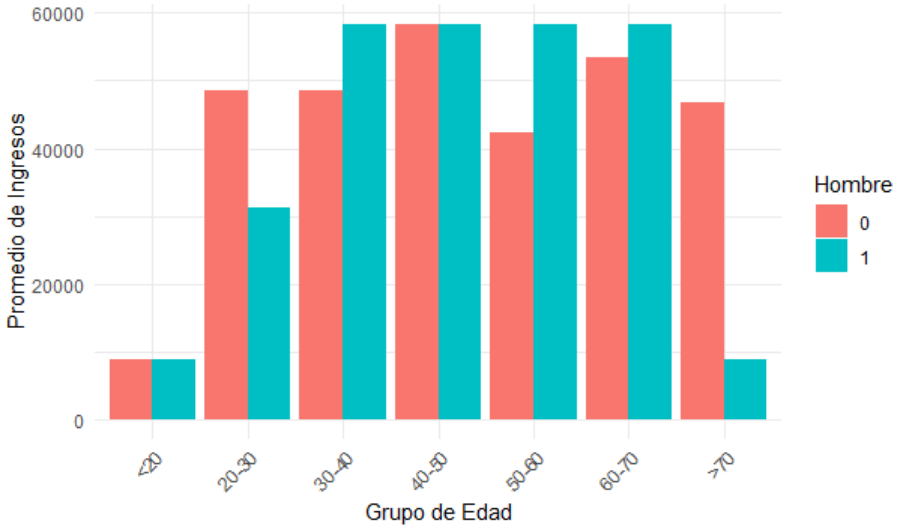


Figure 3. Ingresos promedio por estrato

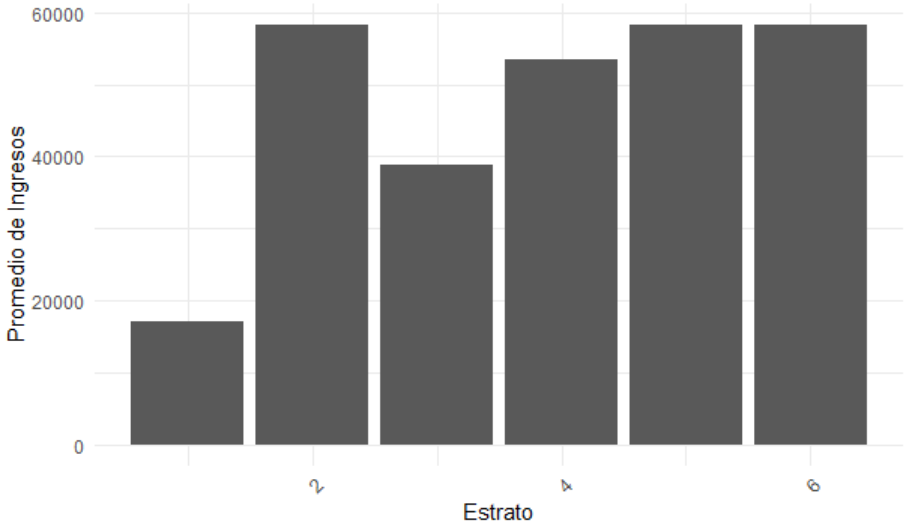


Figure 4. Ingresos promedio por nivel de educación

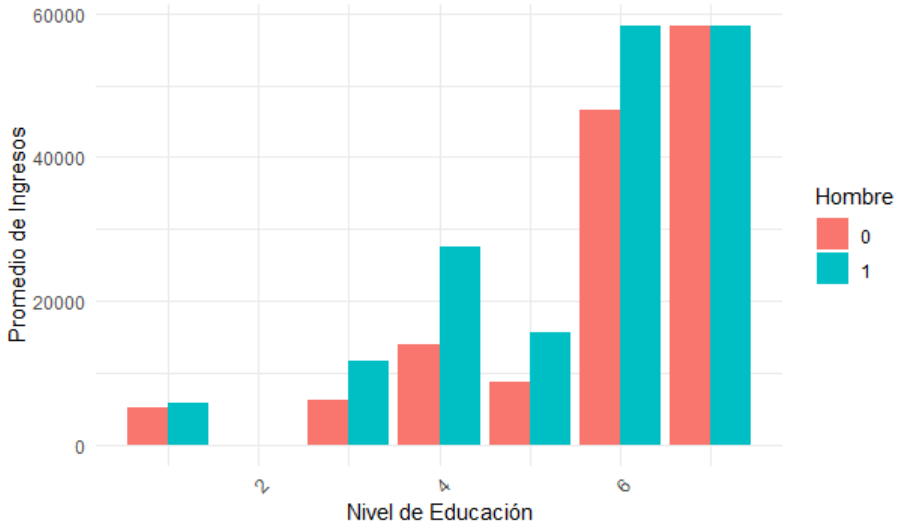


Figure 5. Ingresos promedio por oficio

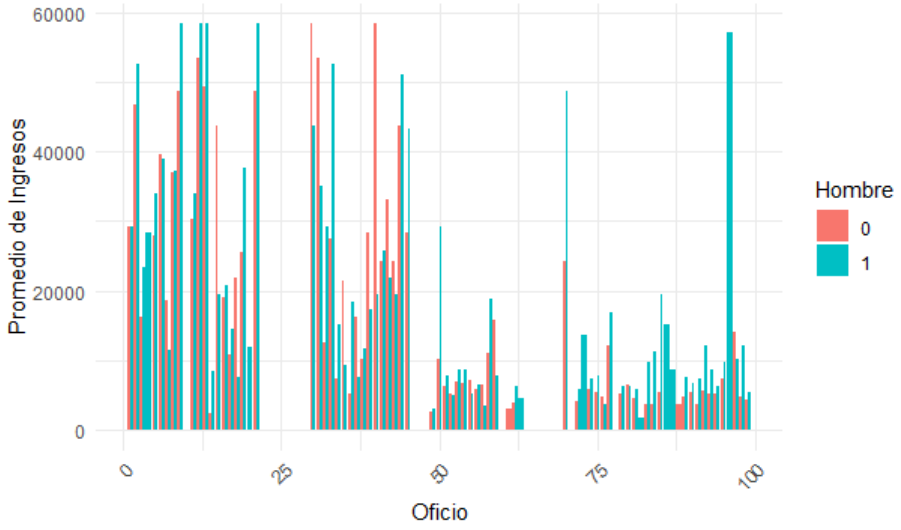


Figure 6. Salario-hora por horas de trabajo

Figure 7. Salario-hora por oficio

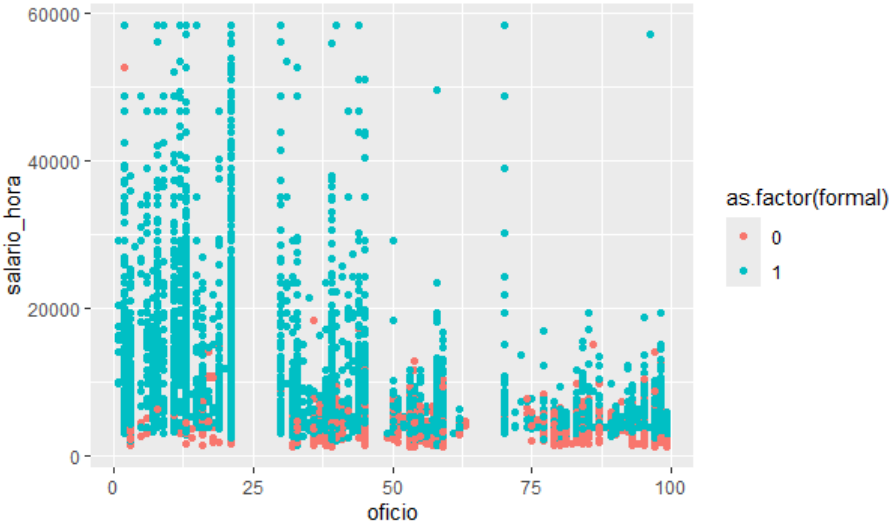


Figure 8. Salario-hora por estrato

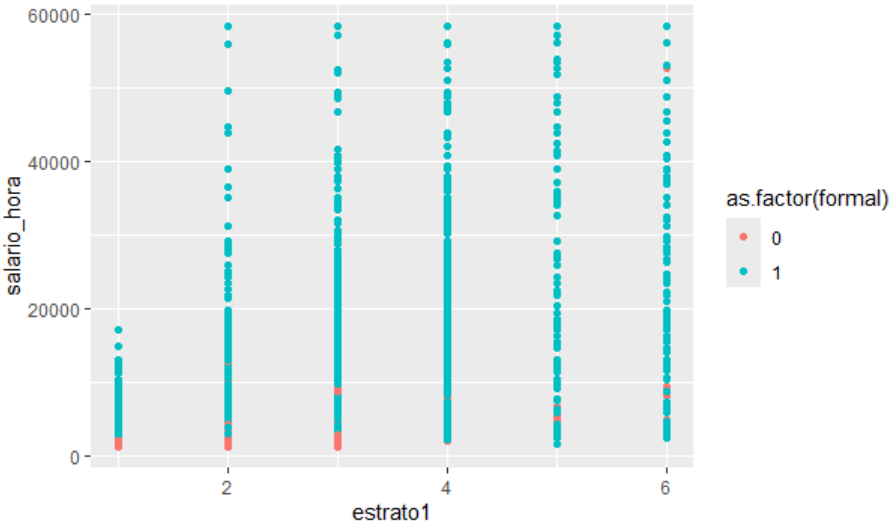


Figure 9. Boxplot promedio de ingresos por estrato

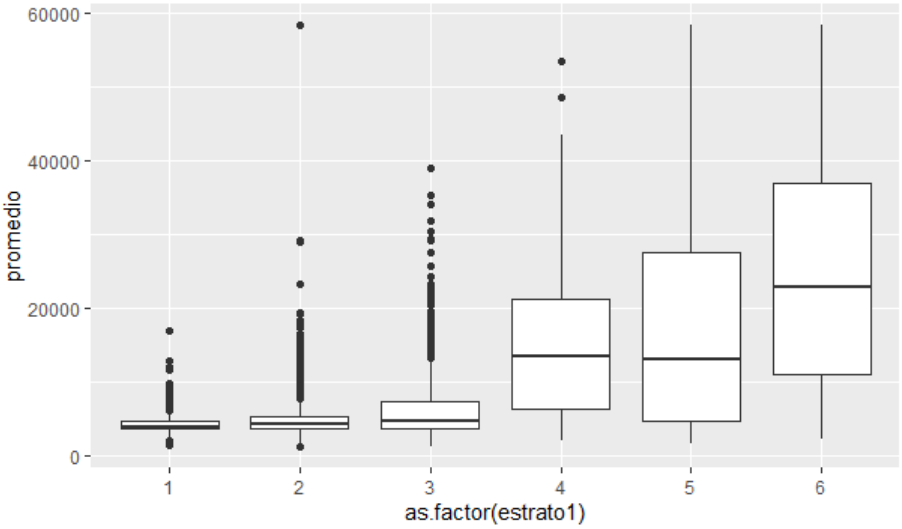


Figure 10. Distribución de errores de predición

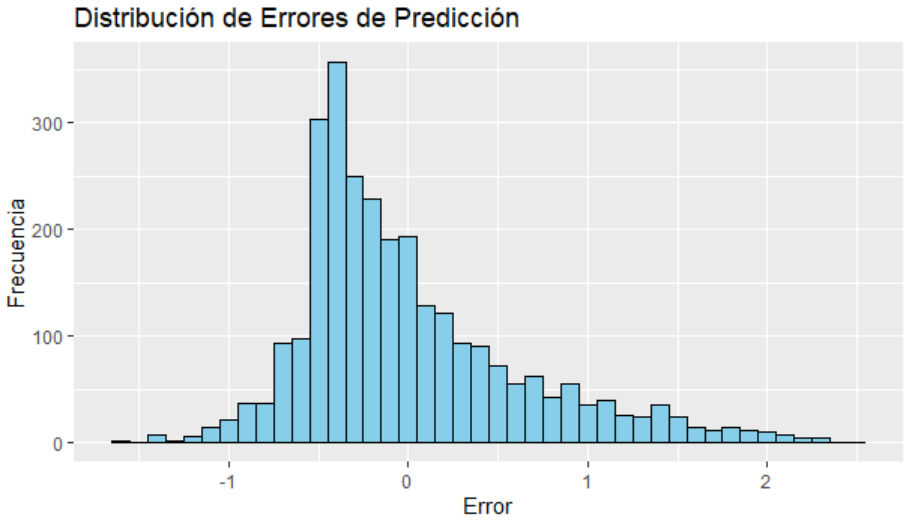


Figure 11. Outliers edad-salario (log)

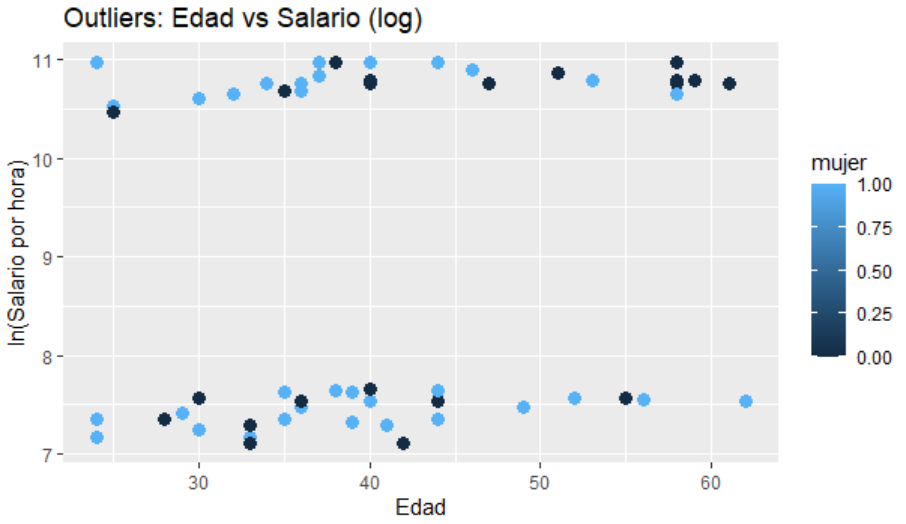
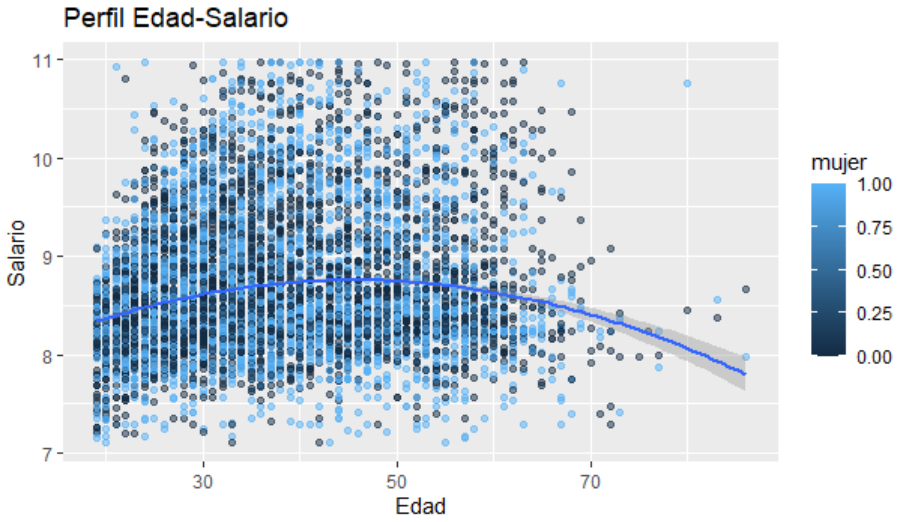


Figure 12. Predicted age-wage profile



.