# Modelización lineal del salario

Emanuelle Marsella, Maximiliano Saldaña

Julio 2020

# Índice


**1. Introducción** **3**

**2. Descripción de los datos** **3**

2.1. Variables Indicadoras . . . 3

**3. Exploracion inicial** **3**

3.1. Variables categóricas . . . 3

3.1.1. Región . . . 3

3.1.2. Rama de actividad . . . 4

3.1.3. Ocupación . . . 5

3.1.4. Otras variables categóricas . . . 5

3.2. Análisis de multicolinealidad . . . 5

**4. Modelos lineales múltiples** **6**

4.1. Primer acercamiento al problema . . . 6

4.2. Modelo final . . . 7

4.2.1. Planteo del modelo . . . 7

4.2.2. Tests de hipótesis iniciales . . . 8

4.2.3. Selección del mejor modelo . . . 9

4.2.4. Análisis de los supuestos . . . 11

4.2.5. Modelo logarítmico . . . 13

4.2.6. Análisis de los supuestos . . . 14

4.2.7. Interpretación de los coeficientes . . . 16

**5. Conclusiones** **18**

**6. Anexos** **19**

6.1. Anexo 1: Modelos para el salario no transformado . . . 19

6.1.1. Selección de variables . . . 19

6.1.2. Análisis de los supuestos . . . 20

6.1.3. Tratamiento de observaciones atípicas . . . 24

6.2. Anexo 2: Gráficos de dispersión del salario para las distintas regiones . . . 25

6.3. Anexo 3: Modelos lineales simples . . . 26

6.4. Anexo 4: Gráficos de regresión lineal simples . . . 26

**7. Bibliografía** **29**

# 1. Introducción

En este proyecto aplicaremos los conceptos y metodologías aprendidos en el curso Modelos Lineales para analizar los datos seleccionados, una base de datos sobre salarios en Estados Unidos extraída del libro *Introducción a la Econometría, Un enfoque moderno (2009)* de J.M. Wooldridge. Nuestra intención es tratar de explicar la variable principal, el salario por hora medido en dólares, a partir de las otras variables de la base. Para lograr esto haremos uso de modelos de regresión lineal y metodologías asociadas, que implementaremos mediante el software R.

# 2. Descripción de los datos

La base cuenta con observaciones de 526 personas y con las siguientes 22 variables: Salario (promedio por hora, medido en dólares), Años de Educación, Años de Experiencia, Antigüedad, Raza (variable indicadora, 1 si la persona no es de raza blanca), Sexo, Estado Civil (vale 1 si la persona está casada), Número de Dependientes, Región Metropolitana (vale 1 si la persona vive en dicha región), Región (dividida en tres variales indicadoras: Norte, Sur y Oeste), Rama de Actividad (dividida en tres variables indicadoras: Construcción, Comercio y Servicios), Ocupación (tres variables indicadoras: Profesional, Administrativos y Servicios), el logaritmo de la variable Salario y los cuadrados de las variables Experiencia y Antigüedad. Contábamos con una variable indicadora más, llamada *"profserv"* , sobre la cual no poseíamos información por lo que optamos por removerla.

## 2.1. Variables Indicadoras

Inicialmente modificamos la base para llevarla a una forma que nos parecía más práctica de trabajar, agrupando variables que refieren a una sola característica de los datos y estaban en una forma binaria. Estas variables son la Región, separada en la base original en Norte, Sur y Oeste; la Rama de Actividad, separada en Construcción, Servicios y Comercio; y la Ocupación, separada en Profesional, Administrativos y Servicios. Luego de consultar a las docentes optamos por no descartar las variables indicadoras ya que nos permiten su utilización a la hora de aplicar un modelo lineal multivariado y a su vez mantenemos la variable agrupada para poder usarla a la hora de manipular y graficar datos.

# 3. Exploracion inicial

Para familiarizarnos con la base utilizamos medidas de resumen. También realizamos algunos modelos lineales simples, que se encuentran en el Anexo.

## 3.1. Variables categóricas

Lo primero que hacemos es visualizar cómo se distribuye la población de acuerdo a las variables categóricas de las que disponemos (Región, Rama de actividad, Ocupación, Raza, Sexo y Estado civil). En cuanto a las primeras cuatro variables buscamos saber cuántos individuos pertenecen a cada rama, así como aplicar algunas medidas de resumen de la variable salario para cada categoría de las tres variables, como se puede ver en los siguientes cuadros:

### 3.1.1. Región

La variable región nos indica en qué región se indican los individuos. Las categorías de esta variable son Norte, Sur, Este y Oeste.

Tenemos 3 variables indicadoras para esta región, Norte-Centro, Sur y Oeste. Este es la categoría de referencia, la cual no cuenta con variable indicadora ya que en este caso la matriz de datos $\mathbb{X}$ no sería de rango completo y por lo tanto no sería invertible, lo que más adelante impediría la estimación única de los coeficientes. Esto ocurre análogamente para las otras variables cualitativas.

| | Región | Cantidad | Mínimo | Media | Máximo |
|---|---|---|---|---|---|
| 1 | Norte | 132 | 1.50 | 5.71 | 21.86 |
| 2 | Sur | 187 | 1.50 | 5.39 | 20.00 |
| 3 | Este | 118 | 1.43 | 6.37 | 24.98 |
| 4 | Oeste | 89 | 0.53 | 6.61 | 22.20 |
| 5 | Todas | 526 | 0.53 | 5.90 | 24.98 |

Cuadro 1: Algunas medidas de resumen para las distintas regiones.

En primer lugar, podemos observar que la región Oeste cuenta con un número considerablemente pequeño de observaciones y la región sur cuenta con muchas observaciones, en comparación con las demás.

En cuanto al salario por hora, vemos que las medias para las regiones Norte y Sur son menores a la media del total de observaciones, mientras que para las regiones Este y Oeste son mayores. Por otro lado, los máximos son similares, si bien apreciamos que el máximo total se observa en la región este.

Los mínimos también son similares para todas las regiones excepto para la región Oeste, para la cual podemos apreciar que el mínimo es casi 3 veces menor que para las demás.

Como parte de la exploración inicial, también realizamos algunos gráficos de dispersión del salario en función de algunas variables cuantitativas, con recta de ajuste lineal, diferenciados por la región. Estos resultados pueden encontrarse en el Anexo 2, en ellos podemos apreciar que no hay demasiada diferencia entre las regiones a la hora de explicar el salario a partir de las distintas variables explicativas individualmente.

### 3.1.2. Rama de actividad

| | Rama de Actividad | Cantidad | Mínimo | Media | Máximo |
|---|---|---|---|---|---|
| 1 | Construcción | 24 | 3.00 | 5.96 | 17.71 |
| 2 | Servicios | 53 | 0.53 | 4.34 | 12.50 |
| 3 | Comercio | 151 | 1.43 | 4.79 | 21.86 |
| 4 | Otros | 298 | 1.50 | 6.73 | 24.98 |
| 5 | Todas | 526 | 0.53 | 5.90 | 24.98 |

Cuadro 2: Algunas medidas de resumen para las distintas ramas de actividad.

Lo primero que observamos, es que la cantidad de observaciones para las distintas categorías varía mucho. En particular, para la categoría “Otros”, que es la de referencia, tenemos un gran número de individuos, más de la mitad. Consideramos que esto se debe a que la diferenciación de la que disponemos para las diferentes ramas de actividad no es exhaustiva (como sí sucedía para las regiones, que solo son 4). Por lo tanto, es lógico que suceda que muchos individuos no pertenezcan ni al sector de Construcción, ni al de Servicios, ni al de Comercio; sino que pertenecen a alguna otra categoría que no está especificada y queda incluída dentro de “Otros”.

Comparando las regiones en media, observamos que la media de la categoría de referencia es mayor que en las demás ramas de actividad. La categoría Servicios tiene la menor media, y en esta categoría se observa el mínimo total, y el menor de los máximos. La categoría Comercio es similar en media a Servicios, si bien tanto su mínimo como su máximo son mayores. Finalmente, la categoría Construcción tiene la mayor media de las 3 indicadoras de las que disponemos, y también el menor mínimo.

#### 3.1.3. Ocupación

| | Ocupación | Cantidad | Mínimo | Media | Máximo |
|---|---|---|---|---|---|
| 1 | Servicio | 74 | 0.53 | 3.59 | 7.81 |
| 2 | Administrativos | 88 | 2.65 | 4.74 | 12.50 |
| 3 | Profesional | 193 | 2.23 | 8.04 | 24.98 |
| 4 | Otros | 171 | 1.43 | 5.07 | 15.00 |
| 5 | Todas | 526 | 0.53 | 5.90 | 24.98 |

Cuadro 3: Algunas medidas de resumen para las distintas ocupaciones.

Podemos observar que para la variable Ocupación, la media de la categoría Profesional es considerablemente mayor respecto de las demás, así como también su máximo. Es razonable esperar que así sea, que las personas que se desempeñan en un empleo profesional perciban un mayor salario, si bien no tenemos fundamentos para afirmar que esta diferencia sea estadísticamente significativa, dado que aún no hemos realizado inferencia a partir de nustros datos.

Por otro lado, la categoría Administrativos tiene el mayor mínimo y el menor máximo, y una media menor que la de la categoría de referencia, lo que sugiere que los niveles de salario no varían demasiado, y en general son menores que los de la categoría de referencia.

Luego, la categoría Servicio tiene el menor mínimo, la menor media, y el menor máximo, lo que sugiere que los individuos que se encuentran en esta categoría perciben en promedio un menor salario.

#### 3.1.4. Otras variables categóricas

Con el resto de las variables econtramos que las proporciones entre las categorías son:

- Para la variable Sexo hay un 47.91 % de mujeres y 52.09 % de hombres. La proporción de la mitad cada sexo.
- Para la variable Estado Civil hay un 60.84 % de casados y 39.16 % de no casados.
- Para la variable Área Metropolitana hay un 72.24 % de personas que habitan en la misma y 27.76 % que no. Predomina bastante la población metropolitana.
- Para la variable Raza hay un 10.27 % de personas no blancas y 89.73 % blancas. El resultado es de esperarse dado que las personas no blancas son una minoría en los Estados Unidos.

### 3.2. Análisis de multicolinealidad

| | Variables | VIF |
|---|---|---|
| 1 | Educación | 22.5 |
| 2 | Experiencia | 16.1 |
| 3 | Antigüedad | 8 |
| 4 | Número de dependientes | 1.2 |
| 5 | Experiencia al cuadrado | 16.1 |
| 6 | Antigüedad al cuadrado | 7.2 |
| 7 | Educación al cuadrado | 21.9 |

Cuadro 4: Variables con mayor FIV.

Buscamos analizar la multicolinealidad entre las distintas variables cuantitativas, utilizando la medida del factor inflacionario de la varianza (FIV), el cual es una medida de qué tan correlacionada está una variable explicativa con las demás incluídas en el modelo. Valores altos del FIV para las variables explicativas pueden ocasionar un aumento en la varianza de los estimadores de los coeficientes de la regresión.

La variable Número de Dependientes tiene un FIV pequeño, de 1.2, por lo cual no parece haber problemas de mumlticolinealidad para esa variable en presencia de las demás. Por otro lado, las variables Experiencia, Educación, Antigüedad y sus transformaciones al cuadrado presentan valores más altos. En particular Educación y Experiencia y sus transformaciones al cuadrado presentan valores que superan a 10, el cual se utiliza como regla empírica para determinar si existen problemas de multicolinealidad. Esto es razonable de todas formas ya que es esperable que exista colinealidad entre una variable y su transformada al cuadrado.

A continuación presentamos la matriz de correlaciones entre las variables explicativas cuantitativas del modelo, dejando de lado la variable número de dependientes para la cual ya vimos que no existen problemas de multicolinealidad.

| | Educ | Educ. Cuadrado | Antig | Antig. Cuadrado | Exper. | Exper. Cuadrado |
|---|---|---|---|---|---|---|
| Educ | 1.00 | 0.98 | -0.06 | -0.07 | -0.30 | -0.33 |
| Educ. Cuadrado | 0.98 | 1.00 | -0.04 | -0.05 | -0.27 | -0.30 |
| Antig | -0.06 | -0.04 | 1.00 | 0.92 | 0.50 | 0.46 |
| Antig. Cuadrado | -0.07 | -0.05 | 0.92 | 1.00 | 0.42 | 0.41 |
| Exper. | -0.30 | -0.27 | 0.50 | 0.42 | 1.00 | 0.96 |
| Exper. Cuadrado | -0.33 | -0.30 | 0.46 | 0.41 | 0.96 | 1.00 |

Cuadro 5: Matriz de correlaciones para las variables con mayor VIF.

La matriz de correlaciones nos permite observar que efectivamente, para el conjunto de variables con mayor FIV, solo hay 3 correlaciones altas, y se dan entre cada variable y su transformada al cuadrado.

# 4. Modelos lineales múltiples

## 4.1. Primer acercamiento al problema

Para comenzar, planteamos un modelo lineal múltiple de la forma

$$Y = X\beta + \varepsilon$$

donde $\varepsilon$ es el vector aleatorio de errores, para el cual consideramos los siguientes supuestos clásicos:

- $E(\varepsilon_i) = 0 \; \forall \; i = 1, ..., n$
- $Var(\varepsilon_i) = \sigma^2 \; \forall \; i = 1, ..., n$
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall \; i \neq j$

Además, para realizar inferencia a través de estadísticos, hicimos la suposición que los errores siguen una distribución normal.

A la hora de seleccionar qué variables utilizaremos en nuestro modelo de regresión lineal múltiple, tenemos varios criterios a aplicar. En un principio podemos hacer uso de la bibliografía sobre cómo se relacionan las variables en cuestión, por ejemplo el Sexo y la Raza, y que nos puede llevar a tomar en cuenta variables que tal vez no sean significativas estadísticamente. También contamos con métodos de selección más generales, como el criterio del Akaike y contrastes de significación de las distintas variables. Hay que tener en cuenta

que puede ocurrir que la conclusión de los distintos criterios no sea la misma y haya que optar por uno u otro.

Inicialmente consideramos un modelo con todas las variables explicativas de la base de datos. El contraste de significación de dicho modelo, en el cual ponemos a prueba la hipótesis nula de que al menos una de las variables sea significativa para explicar el salario, nos dio un p-valor próximo a cero, por lo que rechazamos la hipótesis nula. Para aproximarnos a la selección de aquellas que incluiremos en nuestro modelo final realizamos para cada variable su contraste de significación.

Observamos que había varias variables que no resultaban significativas al 5 %. Como estas conclusiones estaban hechas en un contexto donde considerábamos cada variable en presencia de todas las demás, pueden cambiar al usar un proceso de selección secuencial. Por lo tanto decidimos utilizar el método de selección *backward* secuencial del mejor modelo, para eliminar las variables que menos aportaban al modelo, mediante el criterio de selección de Akaike (AIC). Dicho metodo va eliminando variables con el menor AIC de una en una, en presencia de las demás que todavía se consideran.

Llegamos a un modelo con 11 variables explicativas, que contaba con un $R_a^2$ (coeficiente de determinación ajustado) de 0.48.

El siguiente paso fue realizar el análisis de los supuestos de nuestro modelo.

Para ver si la varianza era homocedástica como suponíamos, realizamos el test de Breusch-Pagan, que nos llevó a concluir que la varianza de los errores no es constante. Esto se correspondía con lo que veíamos en el gráfico de los residuos en función de los valores predichos para las distintas observaciones, en el cual observábamos que el valor de los residuos aumentaba a medida que aumentaba el valor de $\hat{y}_i$.

Luego, decidimos poner a prueba el supuesto de normalidad de los errores. Realizamos tres tests de normalidad: el test de Lilliefors (Kolmogorov-Smirnov), el test de Jarque Bera, y el test de Shapiro-Wilk. Todos ellos nos llevaron a la conclusión de que la distribución de los errores no era normal. Sin embargo, realizamos algunos gráficos de los residuos en los cuales podíamos observar que su distribución se asemejaba bastante a una distribución normal. Sí notamos la presencia de algunas observaciones atípicas, que se alejaban bastante de las demás observaciones y de lo que uno esperaría de una distribución normal. Consideramos por lo tanto que ese podía ser el motivo de la falla de los tests.

Así, como nos fallaban tanto el supuesto de normalidad de los errores como el de homogeneidad de la varianza, decidimos realizar una transformación de Box-Cox de la variable explicada, que nos llevó a trabajar con el logaritmo del salario.

A partir de esta transformación, logramos que la varianza de los errores sea homogénea, de acuerdo al test de Breusch-Pagan. Sin embargo, no logramos corregir la normalidad, debido a que 2 de los 3 test que realizamos nos indicaban que la distribución de los errores no era normal. Aquí recordamos las observaciones atípicas que habíamos observado antes, y decidimos realizar un estudio de las mismas. Llegamos a la conclusión de que algunas observaciones nos estaban causando problemas, por lo cual decidimos eliminar una de ellas (la número 24, la más problemática), y crear una variable indicadora para las demás (las observaciones, 128, 440 y 381), de forma de darle a ellas un tratamiento especial. El detalle de todo este proceso se puede ver en el Anexo 1.

Por lo tanto, lo que haremos a continuación es trabajar desde el principio con un modelo que incluya esas indicadoras y no incluya la observación que removimos. Esto lo haremos debido a que los cambios realizados pueden afectar los tests de significación realizados, en particular, el test de eliminación secuencial *backward*, por lo cual es posible que la mejor elección de las variables ahora sea distinta.

## 4.2. Modelo final

### 4.2.1. Planteo del modelo

Consideramos nuevamente un modelo de la forma

$$Y = X\beta + \varepsilon$$

Y los mismos supuestos explicitados anteriormente. La diferencia con el modelo inicial anterior es la inclusión de variables indicadoras para representar un posible cambio en media y/o varianza en cada observación, que indique un posible apartamiento del comportamiento general de la población. Estas variables indicadoras son de la forma:

$$z_t = \begin{cases} 1 \text{ si } t = i \\ 0 \text{ si } t \neq i \end{cases}$$

Así, estamos considerando un modelo de la forma.

$$y_t = x_t'\beta + \Delta_i z_t + \varepsilon_t$$

Para verificar si corresponde darle un tratamiento especial a estas observaciones planteamos el contraste de significación de $z_t$:

$$H_0)\Delta_i = 0$$
$$H_1)\Delta_i \neq 0$$

*Observación: Dado que eliminamos la observación número 24 ahora tenemos 525 observaciones, por lo tanto los indices a partir de dicha observación serán una unidad menos (ej. la 128 será la 127, la 440 la 439). De todas formas nos referiremos con su subíndices iniciales para evitar confusiones.*

### 4.2.2. Tests de hipótesis iniciales

Inicialmente consideramos un modelo con todas las variables explicativas de la base de datos. Planteamos el siguiente contraste de significación de dicho modelo:

$$H_0)\beta_i = 0 \ \ \forall \ \ i = 1, ..., k$$
$$H_1)\beta_i \neq 0 \ \ con \ \ i \in \{1, ..., k\}$$

En el cual ponemos a prueba la hipótesis nula de que al menos una de las variables sea significativa para explicar el salario con el estadístico:

$$\frac{SCReg/k}{SCErr/(n-k-1)} \sim F_{k,n-k-1}$$

Dicho contraste nos da un p-valor próximo a cero, por lo que rechazamos la hipótesis nula. Al menos una de las variables del modelo planteado es significativa pra explicar el salario.

Para aproximarnos a la selección de aquellas que incluiremos en nuestro modelo final realizamos para cada variable su contraste de significación:

$$H_0)\beta_i = 0$$
$$H_1)\beta_i \neq 0$$

Utilizando el siguiente estadístico:

$$\frac{SCReg - SCReg_{-i}}{SCErr/(n-k-1)} \sim F_{1,n-k-1}$$

A continuación presentamos un cuadro con los distintos p-valores de dichos contrastes.

| | Variable | p-valor |
|---|---|---|
| 1 | Educación | 0.066 |
| 2 | Experiencia | 0 |
| 3 | Antigüedad | 0.002 |
| 4 | Raza | 0.751 |
| 5 | Sexo | 0 |
| 6 | Est. Civil | 0.877 |
| 7 | N° dependientes | 0.838 |
| 8 | Reg. metropolitana | 0.001 |
| 9 | Norte | 0.11 |
| 10 | Sur | 0.049 |
| 11 | Oeste | 0.216 |
| 12 | Construcción | 0.824 |
| 13 | Comercio | 0 |
| 14 | Servicios | 0.004 |
| 15 | Ocup. profesional | 0 |
| 16 | Ocup. administrativos | 0.943 |
| 17 | Ocup. servicios | 0.535 |
| 18 | Exper. al cuadrado | 0 |
| 19 | Antig. al cuadrado | 0.6 |
| 20 | Educ. al cuadrado | 0 |
| 21 | Indic. Obs. 128 | 0.059 |
| 22 | Indic. Obs. 381 | 0.005 |
| 23 | Indic. Obs. 440 | 0 |

Cuadro 6: Tabla de p-valores para el modelo con todas las variables explicativas.

En este modelo, el valor del coeficiente de determinación ajustado es de 0.51, eso indica que el 51 % de la variabilidad del salario es explicada por el modelo. Observamos que la educación no resulta significativa al 5 % para explicar el salario si bien la educación al cuadrado sí lo es. También vemos que no son significativas la variables Raza, Número de Dependientes y Estado Civil, con p-valores próximos a uno, lo que nos llama la atención, en particular para la variable Raza, para la cual esperábamos que fuera significativa para explicar el salario, dado lo que indican los estudios previos[1]. En este modelo la única región que se diferencia significativamente del resto para explicar el salario es la Sur.

En el caso de las ramas de actividad, no resulta significativo diferenciar entre la rama de actividad Construcción y la de referencia, pero si para las ramas de actividad Comercio y Servicios. En cuanto a las ocupaciones, solamente la Profesional es significativa, las otras dos no se logran diferenciar de la categoría de referencia.

Apreciamos que las variables indicadoras para las observaciones 440 y 381 son significativas al 5 %, mientras que la de la 128 no lo es. Mas allá de lo que pasa con las variables indicadoras, estos resultados se corresponden con los del primer modelo que realizamos, que se encuentra en el Anexo 1.

De todas formas, debemos tener en cuenta que todas estas conclusiones son hechas en un contexto donde consideramos cada variable en presencia de todas las demás y pueden cambiar al usar un proceso de selección secuencial, como aplicaremos a continuación.

#### 4.2.3. Selección del mejor modelo

Utilizamos el método de selección *backward* secuencial del mejor modelo, para eliminar las variables que menos aportan al modelo, mediante el criterio de selección de Akaike (AIC). Dicho método va eliminando

[1]Income and poverty in the United States: 2017. (2018) Fontenot, Kayla et al.

variables con el menor AIC de una en una, en presencia de las demás que todavía se consideran. En la tabla que presentamos a continuación se muestran las variables incluidas en el modelo al que llegamos, con sus respectivos p-valores:

| | Variable | Coef. estimado | p-valor |
|---|---|---|---|
| 1 | Educación | -0.373 | 0.057 |
| 2 | Experiencia | 0.2 | 0 |
| 3 | Antigüedad | 0.116 | 0 |
| 4 | Sexo | -1.604 | 0 |
| 5 | Reg. metropolitana | 0.896 | 0.001 |
| 6 | Norte | -0.721 | 0.013 |
| 7 | Sur | -0.824 | 0.002 |
| 8 | Comercio | -1.325 | 0 |
| 9 | Servicios | -1.236 | 0.002 |
| 10 | Ocup. profesional | 1.42 | 0 |
| 11 | Exper. al cuadrado | -0.004 | 0 |
| 12 | Educ. al cuadrado | 0.03 | 0 |
| 13 | Indic. Obs. 128 | -5.136 | 0.051 |
| 14 | Indic. Obs. 381 | 7.57 | 0.004 |
| 15 | Indic. Obs. 440 | 13.887 | 0 |

Cuadro 7: p-valores y coefs. estimados para el modelo obtenido a través del método de selección backward.

Observamos que hay variables categóricas para las cuales algunas de sus indicadoras son significativas y las demás no. Esto significa que pertenecer a una de esas categorías faltantes no implica un cambio en el salario con respecto a las demás que no están incluidas en el modelo . El único cambio significativo en el salario se observa al considerar una observación que está en el grupo de la variable que sí es significativa. Los p-valores que figuran como 0 en el Cuadro 7 corresponden a p-valores muy pequeños, no exactamente 0.

Observamos que en el modelo completo tenemos 9 variables significativas (sin contar las indicadores de las observaciones atípicas), mientras que luego de aplicar la selección secuencial de variables tenemos 11. Esto significa que la variable indicadora de la región norte pasa a ser significativa al considerar un modelo que no incluye las variables eliminadas con el metodo de eliminación backward.

Podemos observar que el método *backward* mantiene la variable educación a pesar de que esta no resulte significativa con el nivel del 5 % que venimos usando. Como el AIC nos indica mantener la variable educación y además la teoría (modelo de Mincer) [2] nos indica que es una variable que aporta a la explicación del salario optamos por mantenerla por ahora.

Además apreciamos que los coeficientes estimados de las indicadoras de Norte y Sur son similares por lo cual consideramos hacer el contraste de la restricción:

$$H_0) \beta_{norte} = \beta_{sur}$$
$$H_1) \beta_{norte} \neq \beta_{sur}$$

Este contraste usa el siguiente estadístico (donde r es la cantidad de restricciones, en este caso una) :

$$\frac{(SCErr_{restr} - SCErr_{comp})/r}{SCErr_{comp}/(n-k-1)} \sim F_{r,n-k-1}$$

El resultado de dicho contraste arroja un p-valor de 0.733, que nos lleva a no rechazar la hipotesis nula, por lo que tomamos la restricción. Esto se interpreta como que Norte y Sur se pueden considerar como un mismo grupo de la variable Región respecto a la de referencia, la cual agrupa Este y Oeste.

[2] The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings. Lemieux, Thomas. (2006). pág. 2.

Siguiendo un proceso similar llegamos a la conclusión de que podemos agrupar a las categorías de la variable Rama de Actividad comercio y servicios, no rechazando la hipótesis nula de que son iguales sus coeficientes con un p-valor de 0.8396.

Queremos ver si es viable incluir la variable raza que ha demostrado ser importante para explicar el salario [3]. Su contraste de significación en presencia de las demás variables del modelo luego de aplicar el método *backward* nos lleva a un p-valor del 0.78, el cual resulta muy alto para optar por incluirla de todas maneras al modelo.

Finalmente llegamos a un modelo con 7 variables, donde el $R_a^2$ (coeficiente de determinación ajustado), con el cual vemos la bondad del ajuste del modelo propuesto es de 0.52, lo que indica que un 52 % de la variabilidad total del salario está explicada por el modelo, el cual es un valor relativamente alto en la práctica. En comparación con el modelo completo considerado anteriormente, el $R_a^2$ aumenta en 1 %.

### 4.2.4. Análisis de los supuestos

Para comenzar con nuestro análisis de los supuestos, lo primero que haremos es un gráfico de los residuos en función de los valores predichos para las distintas observaciones.

Dado que suponemos que los errores tienen esperanza igual a 0 y varianza constante, si eso se cumple, esperamos que a simple vista los residuos se distribuyan aleatoriamente en torno al 0, sin seguir ningún patrón en particular.

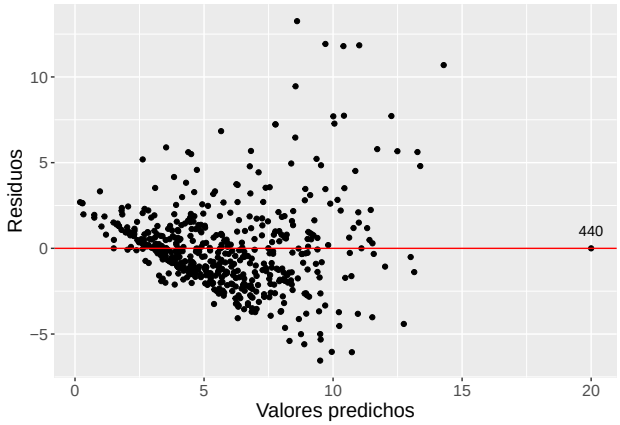



Figura 1: Gráfico de residuos en función de los valores predichos por el modelo.

Inicialmente notamos una observación notablemente separada de las demás, la que corresponde a la 440. Esta es una de las observaciones en las cuales consideramos que había un salto en media y cuyo valor predicho original (en el modelo inicial del anexo) era de 6.371 con un residuo de 13.629, lo que significa que nuestro modelo subestimaba el salario predicho en esta última cantidad. Ahora, en el nuevo modelo su valor predicho es de 20 con un residuo de -0.281, por lo cual el tratamiento especial a esta observación corrigió la subestimación.

Por lo demás, en este gráfico podemos ver que en principio los residuos parecen estar distribuidos en torno al 0. Sin embargo, al contrario de lo que suponíamos, la varianza parece aumentar a medida que aumenta el valor de $\hat{y}$. Por lo tanto, no se cumple nuestro supuesto de que la varianza es constante para todas las observaciones. Eso es un problema, ya que si no se cumple el supuesto de homocedasticidad, no estamos en las hipótesis del Teorema de Gauss-Markov, y por lo tanto los estimadores de los coeficientes de nuestro modelo obtenidos por Mínimos Cuadrados Ordinarios no son los mejores estimadores insesgados (es decir, no son los de mínima varianza).

[3] Ver cita pág. 9

Para confirmar que no se cumple la homocedasticidad utilizamos el test de Breusch-Pagan tomando en cuenta todas las variables seleccionadas para el modelo. Nuestra hipótesis nula es que el modelo es homocedástico y la rechazamos dado el valor-p próximo a 0 que arroja el contraste, por lo que concluiriamos que la varianza no es constante y tenemos que arreglarlo. Un posible camino a tomar para corregir esto es la transformación de Box-Cox de la variable explicada, que a su vez se usa para corregir la falla del supuesto de normalidad de los errores. Antes de llevar a cabo esta transformación veremos que pasa con dicho supuesto.

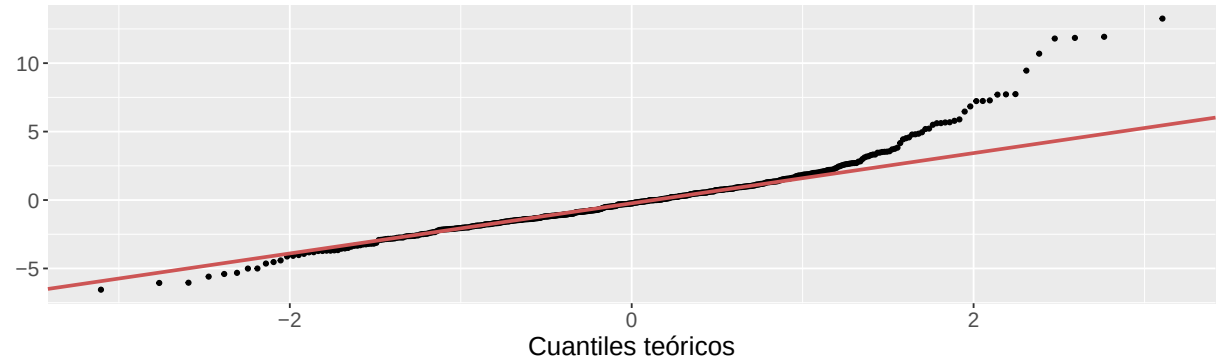



Figura 2: Gráfico QQ-plot

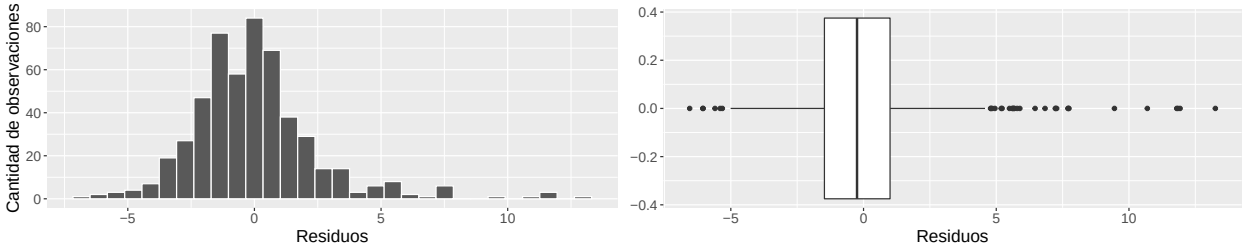



Figura 3: Histograma y boxplot de los residuos

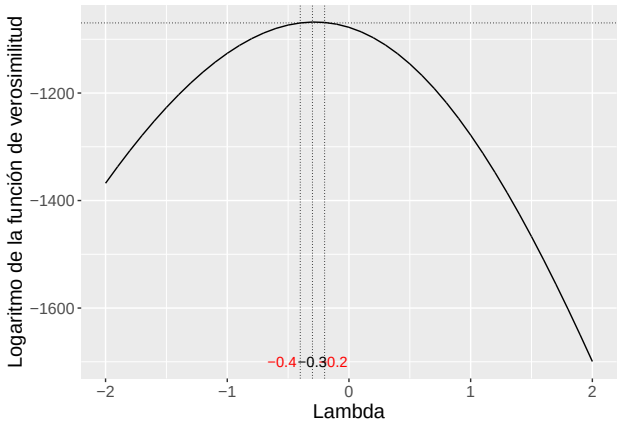
Como acercamiento a analizar el supuesto de normalidad de los errores realizamos un gráfico de cuantiles teóricos ( *QQ-plot* ), un histograma de los residuos y un *boxplot*. La primera de las figuras compara los cuantiles de la distribución de los residuos con los cuantiles teóricos de la distribución normal, donde se espera que la nube de puntos se ajuste a la recta de cuantiles teóricos en tanto la distribución de los residuos se asemeje a una normal. En el histograma podemos ver una representación de la distribución de los residuos, mientras que en el boxplot permite apreciar alguna medidas de resumen de los residuos (como la media y los cuantiles) lo cual ayuda a analizar si se cumplen las características de la distribución normal como su simetría.

En particular en nuestros gráficos podemos observar que a grandes rasgos la distribución de los residuos se asemeja a una normal, pero las observaciones con residuos altos distorsionan esta semejanza. En el *QQ-plot* podemos apreciar que la nube de puntos se aproxima bastante a la recta de comparación con la distribución normal teórica (de esperanza 0 y varianza $\hat{\sigma}^2$), excepto en las colas, donde a pesar de que se espera un distanciamiento mayor que en centro de los datos el que se presenta es muy grande. En el *boxplot* podemos notar numerosas observaciones que se alejan del resto por lo cual podríamos considerarlas como atípicas, asi como también distinguimos que la mediana esta próxima a 0 y sin ser por los atípicos hay cierta simetría, además de haber una acumulación de observaciones en torno al 0. En cuanto al histograma presentado se puede ver que la forma de la distribución de los residuos se asemeja bastante a la de una normal, a excepción de los atípicos ya mencionados.

Confirmamos la falla en el cumplimiento del supuesto mediante tests de hipótesis, en este caso el de

Kolgomorov-Smirnov (Lilliefors) que considera un estadístico que toma en cuenta la máxima distancia vertical entre la distribución teórica y la distribución empírica de los residuos, el de Jarque-Bera, que se basa en los coeficientes de simetría y curtosis de la muestra, y el de Shapiro-Wilk. Todos nos llevan a la misma conclusión, rechazar la hipótesis nula de que la distribución de los residuos es normal. Por lo tanto como también encontramos problemas con la normalidad de los errores es apropiado realizar la transformación Box-Cox de la variable explicada salario, lo que además se corresponde con la especificación del modelo de Mincer si la transformación apropiada es el logaritmo de la variable en cuestión.

La transformación de Box-Cox consiste en considerar la función de verosimilitud del salario y hayar su máximo para un conjunto dado de valores $\lambda$, generalmente entre -2 y 2. La transformación logaritmica la utilizamos en el caso de que $\lambda = 0$.




El intervalo de confianza al 95 % del verdadero valor maximal de $\lambda$ es $(-0,4; -0,2)$, por lo cual la transformación de Box-Cox no nos sugiere el uso del logaritmo para corregir la heterocedasticidad y la no normalidad. Sin embargo la teoría nos indica el uso del logaritmo (modelo de Mincer y el uso de esta transformación en la econometría[4]), teniendo una interpretación clara y además se justifica que lo utilicemos debido a que le hicimos un tratamiento especial a las observaciones problemáticas con el objetivo de corregir estos supuestos para el logaritmo del salario.

### 4.2.5. Modelo logarítmico

A continuación, vamos a plantear un modelo de la forma

$$log(Y) = X\beta + \varepsilon$$

donde una vez más hacemos los mismos supuestos sobre los errores: normalidad, esperanza nula, homocedasticidad e incorrrelación entre los mismos.

Nuestra variable explicada pasa ahora a ser el logaritmo del salario, que se puede interpretar de la siguiente forma: dado un aumento unitario en una de las variables explicativas $x_i$, la variación porcentual del salario equivale aproximadamente a $100\beta_i$, a niveles constantes de todas las demás variables [5], es decir:

$$\%\Delta y \simeq (100\beta_i)\Delta x$$

Partimos de un modelo para explicar la variación porcentual del salario a partir del mismo conjunto de variables explicativas al que habíamos llegado para explicar el salario sin transformar.

[4]Introducción a la econometría: Un enfoque moderno. Wooldrige, Jeffrey. (2009). pág. 43.
[5]Idem cita 4.

Para comenzar, realizamos el contraste de significación para todos los regresores del modelo. Lo primero que podemos observar es que aún se cumple que la educación no es significativa al 5 %, pero ahora su p-valor se separa mucho del valor prefijado de $\alpha$ , ya que aumentó hasta 0,26. Por lo tanto, optamos por eliminarla del modelo.

A continuación, se presenta la tabla de p-valores para las variables explicativas que mantenemos en el modelo:

| | Variable | p-valor |
|---|---|---|
| 1 | Experiencia | 0 |
| 2 | Antigüedad | 0 |
| 3 | Sexo | 0 |
| 4 | Reg. metropolitana | 0 |
| 5 | Norte/Sur | 0.005 |
| 6 | Comercio/Servicios | 0 |
| 7 | Ocup. profesional | 0 |
| 8 | Exper. al cuadrado | 0 |
| 9 | Educ. al cuadrado | 0 |
| 10 | Indic. Obs. 128 | 0 |
| 11 | Indic. Obs. 381 | 0.001 |
| 12 | Indic. Obs. 440 | 0 |

Cuadro 8: p-valores para el modelo que explica la variación porcentual del salario.

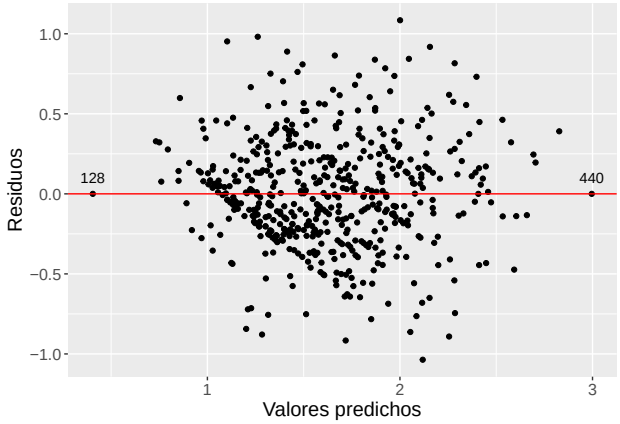
Podemos observar que la mayoría de los p-valores son tan pequeños que no es posible representarlos en dos cifras decimales. En particular, la indicadora de la observación 128, que no era significativa al 5 % en el modelo para explicar el salario sin transformar, ahora sí lo es.

Luego, se aprecia una diferencia significativa para explicar la variación porcentual del salario para las diferentes regiones, ocupaciones y ramas de actividad. La variable indicadora de las regiones Norte y Sur agrupadas es significativa, mientras que Oeste no lo es. Por lo tanto, consideramos que podemos agrupar a Oeste y Este dentro de la categoría de referencia, y así aquellas observaciones que pertenezcan al Norte o al Sur observaran un cambio en el valor esperado de la variable explicada. Lo mismo sucede con las otras variables categóricas: para las ramas de actividad, comercio y servicios se diferencian de la categoría de referencia, mientras que industria no; y para las ocupaciones, la ocupación profesional se diferencia de las demás, mientras que consideramos a las categorías administrativos, servicios y la de referencia dentro de un mismo grupo.

El modelo en su conjunto resulta significativo con un p-valor próximo a 0, y cuenta con un coeficiente de determinación ajustado $R_a^2$ igual a 0.566. Este supone un aumento del 5 % con respecto al modelo anterior.

#### 4.2.6. Análisis de los supuestos

Nos disponemos ahora a analizar nuevamente qué ocurre con el cumplimiento de los supuestos una vez realizada la transformación.




El gráfico de los residuos nos muestra que ahora ya no está tan marcado el patrón de la distribución que anteriormente veíamos. Notamos dos observaciones que se separan ligeramente del resto de la nube de puntos, la 128 y la 440, que son dos de las que recibieron un tratamiento especial en el modelo. El test de homocedasticidad de Breusch-Pagan nos da un p-valor de 0.12 nos lleva a no rechazar la hipótesis nula de que los residuos tienen igual varianza a un nivel de significación del 5 %.

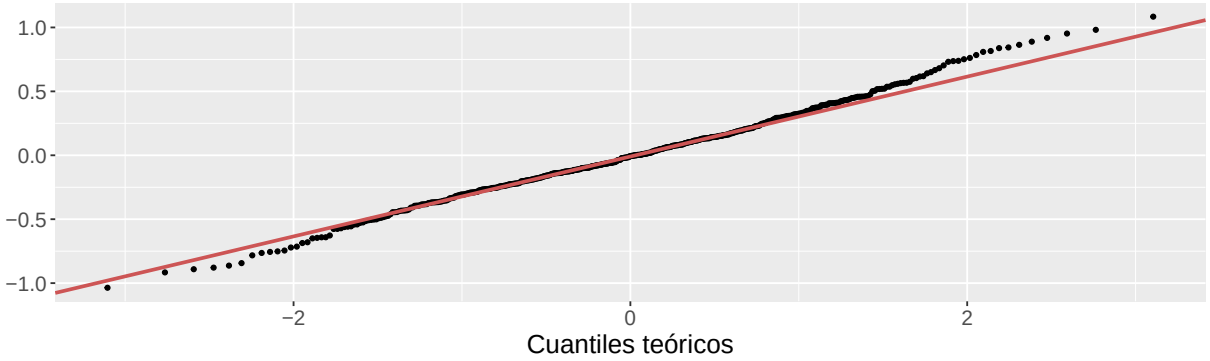



Figura 4: Gráfico QQ-plot para el modelo logarítmico

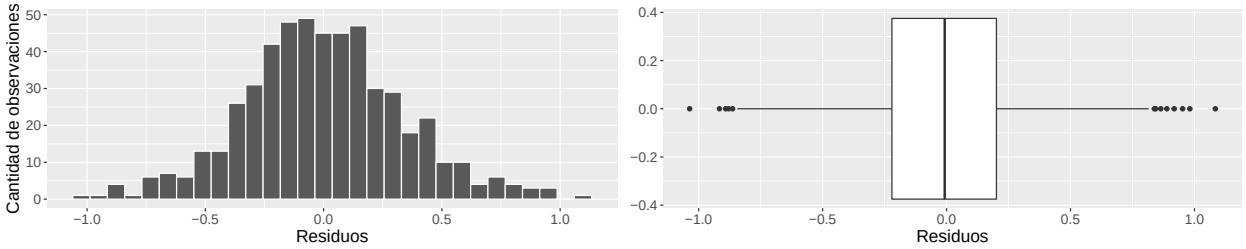



Figura 5: Histograma y boxplot de los residuos para el modelo logarítmico

En cuanto a la normalidad, en comparación con el modelo sin la transformación, podemos apreciar como en los tres gráficos las características de la distribución normal son más claras. En el QQ-plot los puntos se ajustan más a la recta de cuantiles teóricos, por lo que los cuantiles empíricos de los residuos se asemejan más a los de una normal, el histograma se asemeja más a la forma de la normal y el *boxplot* nos muestra una mayor simetría y una mediana más próxima a 0, así como una mayor concentración de las observaciones en

torno a ese valor. En el boxplot aún se pueden notar algunas observaciones atípicas, pero de forma mucho menos pronunciada que antes.

Efectivamente, al realizar los tests de normalidad obtenemos los siguientes p-valores:

| | Test | p-valor |
|---|---|---|
| 1 | Lilliefors (Kolmogorov-Smirnov) | 0.0839 |
| 2 | Jarque–Bera | 0.1 |
| 3 | Shapiro-Wilk | 0.0784 |

Cuadro 9: p-valores para los distintos tests de normalidad de los errores

Podemos ver que ahora, tras haber realizado la transformación logaritmica, los tests de Lilliefors, Jarque–Bera y Shapiro-Wilk nos indican que existe evidencia significativa al 5 % para afirmar que nuestra variable aleatoria de error sigue una distribución normal.

En conclusión, llegamos que nuestros errores siguen una distribución normal, con media 0 y varianza homocedástica. Por lo tanto, podemos afirmar que nuestros estimadores de los coeficientes de nuestro modelo son los mejores estimadores insesgados (son los de mínima varianza), y además estamos en condiciones de realizar inferencia sobre nuestro modelo, ya que la confirmación del supuesto de normalidad valida los contrastes de hipótesis realizados sobre el mismo.

#### 4.2.7. Interpretación de los coeficientes

Contamos con un modelo que cumple con los supuestos clásicos, cuyas variables explicativas fueron seleccionadas para conformar el mejor modelo, y cuya variable explicada cuenta con una interpretación clara. Por lo tanto, estamos en condiciones de tomar nuestro modelo como válido y realizar conclusiones a partir de él.

Primero que nada, recordemos que nuestra variable a explicar es el salario por hora promedio medido en dólares. Al transformarla logarítmicamente, teníamos la siguiente interpretación para los coeficientes:

$$\%\Delta y \simeq (100\beta_i)\Delta x$$

Así, nos interesa conocer el valor de los coeficientes al multiplicarlos por 100. En el caso de las variables cuantitativas, ese producto nos indica la variación porcentual en el salario dado un aumento unitario en la variable explicativa, manteniendo los demás regresores constantes. En el caso de las variables categóricas, ese producto se interpreta como la variación porcentual en el salario al pertenecer a una categoría u otra de la variable, respecto de la categoría de referencia.

**Variables cuantitativas**

| | Variable ($x_i$) | Parámetro ($\hat{\beta}_i$) | $100 \cdot \hat{\beta}_i$ |
|---|---|---|---|
| 1 | Experiencia | 0.0309 | 3.09 |
| 2 | Antigüedad | 0.0138 | 1.38 |
| 3 | Exper. al cuadrado | -0.000629 | -0.0629 |
| 4 | Educ. al cuadrado | 0.00218 | 0.218 |

Cuadro 10: Parámetros para las distintas variables explicativas cuantitativas.

En el Cuadro 10 presentamos los coeficientes estimados del modelo y los mismos multiplicados por 100. En primer lugar, si analizáramos el coeficiente de la experiencia por si solo, el aumento en la experiencia en un año manteniéndose todas las demás variables constantes se traduce en un aumento esperado del 3.1 %. Sin embargo, notemos que también está presente la experiencia al cuadrado, por lo cual la variación porcentual del salario esperada dado un aumento unitario en la experiencia depende de los dos coeficientes. No es

posible aislar el efecto de una u otra variable. Notemos también que el signo del coeficiente estimado para la experiencia al cuadrado es negativo, lo cual se interpreta como que a medida que aumenta la experiencia el crecimiento porcentual esperado en el salario será decreciente.

En el caso de la antigüedad, un aumento unitario de la variable, con las demás constantes se ve reflejado en un aumento porcentual del 1.38 % del salario esperado.

Por su parte un aumento en los años de educación tiene una relación cuadrática con el aumento porcentual en el salario (solo se encuentra presente la educación al cuadrado). Así cuanto mayor sea la cantidad de años de educación el incremento porcentual del salario se irá haciendo mayor a su vez, por el contrario de como sucedió con la experiencia. De esta manera, considerando un individuo con $\mu$ años de educación un aumento unitario en esta variable supone un aumento porcentual del salario de $0,218 \times \left((\mu+1)^2 - \mu^2\right)$, dadas todas las demás variables constantes.

**Variables cualitativas**

| | Variable ($x_i$) | Parámetro ($\hat{\beta}_i$) | $100 \cdot \hat{\beta}_i$ |
|---|---|---|---|
| 1 | Sexo | -0.277 | -27.7 |
| 2 | Reg. metropolitana | 0.158 | 15.8 |
| 3 | Norte/Sur | -0.0875 | -8.75 |
| 4 | Comercio/Servicios | -0.236 | -23.6 |
| 5 | Ocup. profesional | 0.196 | 19.6 |

Cuadro 11: Parámetros para las distintas variables explicativas cualitativas.

Cuando nos enfrentamos al analisis de las variables cualitativas, buscamos explicar la diferencia en la variación porcentual del salario al pertenecer a una categoría u otra de la variable, respecto de la de referencia. Lo primero que observamos es que se puede hacer una distinción en los ingresos esperados dependiendo del sexo de la persona analizada. Así, una mujer tendrá un salario esperado 27.7 % menor que el de un hombre, aunque ambos tengan el mismo nivel educativo, experiencia, antigüedad, trabajen en la misma área y en definitiva sean indistinguibles para todas las variables a excepción de su sexo.

Por otro lado, se aprecia una distinción entre las regiones metropolitana y no metropolitana, esperandose un salario 15.8 % mayor en el primer caso, a iguales niveles de las demás variables.

A continuación nos disponemos a analizar la influencia de la ubicación geográfica en el salario esperado. Los individuos que se encuentran en la región Norte o en la Sur se espera que perciban un salario 8.75 % menor que aquellos que se encuentran en las regiones Este u Oeste (las de referencia). Esto se corresponde con la estadística descriptiva realizada en la sección *Descripción de los datos* (Cuadro 1), en la que observábamos que el salario es similar en cuanto a su media para las regiones Norte y Sur, la cual es superior a la de las regiones Este y Oeste.

En cuanto a las ramas de actividad observamos que las personas que pertenecen a las ramas Comercio y Servicios se espera que perciban un salario 23.6 % menor que aquellas que están en la categoría de referencia, la que incluye la rama Construcción y otras no especificadas en la base. Ya observábamos indicios de esto en el Cuadro 2, donde la media para esas dos ramas eran visiblemente las más bajas.

La ocupación profesional se distingue del resto (servicios, administrativos y otras no especificadas), siendo el salario esperado para los que pertenecen a esa ocupación un 19.6 % mayor que para las demás. Nuevamente, esto es coherente con lo observado en el Cuadro 3.

**Variables indicadoras**

| | Variable ($x_i$) | Parámetro ($\hat{\beta}_i$) | $100 \cdot \hat{\beta}_i$ |
|---|---|---|---|
| 1 | Indic. Obs. 128 | -1.24 | -124 |
| 2 | Indic. Obs. 381 | 1.14 | 114 |
| 3 | Indic. Obs. 440 | 1.32 | 132 |

Cuadro 12: Parámetros para las distintas variables indicadoras de observaciones atípicas.

Consideramos estas tres observaciones que se alejan de lo predicho por el modelo, en particular, el salario esperado para la observación 128 es un 124 % menor a lo que se esperaría para otra persona con los mismos niveles de las variables. Por otro lado, para las observaciones 381 y 440 su salario esperado es 114 % y 132 % mayor respectivamente, respecto a observaciones con sus mismos niveles de las variables.

# 5. Conclusiones

Finalmente llegamos a un modelo explicativo del salario que toma como variable explicada su transformación logarítmica, interpretada como la variación porcentual del salario ante cambios en una de las variables explicativas, considerando las demás constantes.

Del conjunto de variables cualitativas, observamos la importancia del sexo de la persona a la hora de analizar su salario, concluyéndose una desventaja en para las mujeres incluso en iguales condiciones de las otras variables. Por otro lado, llegamos a la conclusión de que en el contexto estudiado la raza no resulta significativa para estudiar el salario, lo que indicaría que no hallamos discriminación en este aspecto. A su vez, no parecer aportar al análisis el estado civil o el número de dependientes que tienen las personas. Encontramos significativa la región donde viven las personas para la explicación, aunque debimos recurrir a una recategorización. Las personas que viven en el conjunto formado por las regiones Este y Oeste tienen una ventaja salarial sobre aquellas que viven en el Norte o el Sur, y también los habitantes de la región metropolitana aventajan a los que no lo son. En cuanto al área laboral en la que se desempeñan las personas, aquellos que trabajan en las ramas de actividad Comercio y Servicios enfrentan una pérdida porcentual de salario respecto al resto de las ramas (en la que se incluyen la de la construcción); mientras que ocurre lo contrario para aquellos que se ocupan como profesionales respecto al resto de las ocupaciones.

A la hora de tomar en cuenta las variables cuantitativas de la base, apreciamos que cuanto mayor sea la cantidad de años de educación, mayor será el salario y también el aumento porcentual relativo del salario. No es este el caso con la experiencia, para la cual se observa que el salario aumentará porcentualmente con ella pero lo hará a una tasa decreciente. Por último encontramos que la antigüedad de una persona en su empleo tendrá una influencia positiva en su salario percibido.

Fue clave para poder hacer el análisis reconocer la presencia de tres individuos que se apartan considerablemente del comportamiento predicho por el modelo. Una de esas observaciones contaba con un salario notablemente menor al predicho para alguien de sus características mientras que ocurría lo contrario para las otras dos.

# 6. Anexos

## 6.1. Anexo 1: Modelos para el salario no transformado

Plantearemos un modelo lineal múltiple de la forma

$$Y = X\beta + \varepsilon$$

donde $\varepsilon$ es el vector aleatorio de errores, considerando los siguientes supuestos clásicos:

- $E(\varepsilon_i) = 0 \; \forall \; i = 1, ..., n$
- $Var(\varepsilon_i) = \sigma^2 \; \forall \; i = 1, ..., n$
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall \; i \neq j$

Además, para realizar inferencia a través de estadísticos, suponemos que los errores siguen una distribución normal.

### 6.1.1. Selección de variables

Como planteábamos en el cuerpo del texto, inicialmente consideramos un modelo con todas las variables explicativas de la base de datos. El contraste de significación de dicho modelo nos da un p-valor próximo a cero. Para aproximarnos a la selección de aquellas que incluiremos en nuestro modelo final realizamos para cada variable su contraste de significación, a continuación presentamos una tabla con los distintos p-valores de dichos contrastes.

| | Variable | p-valor |
|---|---|---|
| 1 | Educación | 0.113 |
| 2 | Experiencia | 0 |
| 3 | Antigüedad | 0.002 |
| 4 | Raza | 0.731 |
| 5 | Sexo | 0 |
| 6 | Est. Civil | 0.82 |
| 7 | N° dependientes | 0.973 |
| 8 | Reg. metropolitana | 0.007 |
| 9 | Norte | 0.111 |
| 10 | Sur | 0.086 |
| 11 | Oeste | 0.258 |
| 12 | Construcción | 0.856 |
| 13 | Comercio | 0 |
| 14 | Servicios | 0.004 |
| 15 | Ocup. profesional | 0 |
| 16 | Ocup. administrativos | 0.746 |
| 17 | Ocup. servicios | 0.575 |
| 18 | Exper. al cuadrado | 0 |
| 19 | Antig. al cuadrado | 0.398 |
| 20 | Educ, al cuadrado | 0.001 |

Cuadro 13: Tabla de p-valores para el modelo con todas las variables.

En este modelo, observamos que la educación no resulta significativa al 5 % para explicar el salario si bien la educación al cuadrado sí lo es. También vemos que no son significativas la variables Raza, Número de

Dependientes y Estado Civil, con p-valores próximos a uno, lo que nos llama la atención, en particular para la variable Raza, para la cual esperábamos que fuera significativa para explicar el salario, dado lo que indican los estudios previos [6]. En este modelo tampoco se aprecia una diferencia significativa para explicar el salario entre las regiones.

En el caso de las ramas de actividad, no resulta significativo diferenciar entre la rama de actividad Contrucción y la de referencia, pero sí para las ramas de actividad Comercio y Servicios. En cuanto a las ocupaciones, solo la Profesional es significativa, las otras dos no se logran diferenciar de la categoría de referencia.

De todas formas, debemos tener en cuenta que todas estas conclusiones son hechas en un contexto donde consideramos cada variable en presencia de todas las demás y pueden cambiar al usar un proceso de selección secuencial, como aplicaremos a continuación.

Utilizamos el método de selección *backward* secuencial del mejor modelo, cuyo funcionamiento fue explicado en el cuerpo del texto. A continuación se presenta una tabla del modelo al cual llegamos, con las variables y sus correspondientes p-valores:

| | Variable | p-valor |
|---|---|---|
| 1 | Educación | 0.102 |
| 2 | Experiencia | 0 |
| 3 | Antigüedad | 0 |
| 4 | Sexo | 0 |
| 5 | Reg. metropolitana | 0.005 |
| 6 | Norte | 0.015 |
| 7 | Sur | 0.008 |
| 8 | Comercio | 0 |
| 9 | Servicios | 0.001 |
| 10 | Ocup. profesional | 0 |
| 11 | Exper. al cuadrado | 0 |
| 12 | Educ, al cuadrado | 0.001 |

Cuadro 14: Tabla de p-valores, modelo obtenido a través de método backward

Observamos que hay variables categóricas para las cuales algunas de sus indicadoras son significativas y las demás no. Como se explicó en el cuerpo del texto, esto significa que el único cambio significativo en el salario se observa al considerar una observación que esta en el grupo de la variable que sí es significativa.

A su vez, en el modelo completo tenemos 9 variables significativas, mientras que luego de aplicar la selección secuencial de variables tenemos 11. Esto significa que las variables indicadoras de la regiones norte y sur pasan a ser significativas al considerar un modelo que no incluye las variables eliminadas con el metodo de eliminación *backward*.

Al igual que en el modelo presentado en el cuerpo del texto, mantenemos la variable Educación a pesar de que no resulta significativa al 5 %; y que el p-valor de la variable Raza al incluirla en este modelo es de 0.774, por lo cual se opta por no incluirla.

El $R_a^2$ (coeficiente de determinación ajustado), es de 0.48. En comparación con el modelo completo considerado anteriormente, el $R_a^2$ aumenta en 2 %, lo que es esperable ya que este coeficiente penaliza por la incorporación de variables que no aportan mucho.

### 6.1.2. Análisis de los supuestos

Para comenzar con nuestro análisis de los supuestos, lo primero que haremos es un gráfico de los residuos en función de los valores predichos para las distintas observaciones.

[6] Ver cita pág. 9

Dado que suponemos que los errores tienen esperanza igual a 0 y varianza constante, si eso se cumple, esperamos que a simple vista los residuos se distribuyan aleatoriamente en torno al 0, sin seguir ningún patrón en particular.

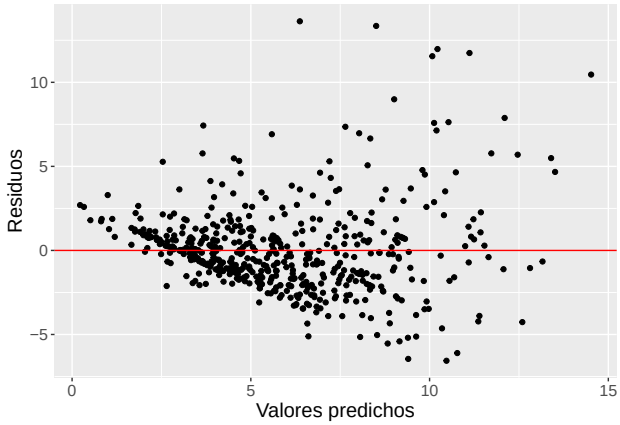



Figura 6: Gráfico de residuos en función de los valores predichos por el modelo.

En este gráfico podemos ver que en principio los residuos parecen estar distribuidos en torno al 0. Sin embargo, al contrario de lo que suponíamos, la varianza parece aumentar a medida que aumenta el valor de $\hat{y}$. Por lo tanto, no se cumple nuestro supuesto de que la varianza es constante para todas las observaciones.

Para confirmar que no se cumple la homocedasticidad utilizamos el test de Breusch-Pagan tomando en cuenta todas las variables seleccionadas para el modelo. Nuestra hipótesis nula es que el modelo es homocedástico y la rechazamos dado el valor-p de $1{,}0891457 \times 10^{-8}$ por lo que concluiríamos que la varianza no es constante y tenemos que arreglarlo. Un posible camino a tomar para corregir esto es la transformación de Box-Cox de la variable explicada, que a su vez se usa para corregir la falla del supuesto de normalidad de los errores. Antes de llevar a cabo esta transformación veremos que pasa con dicho supuesto.

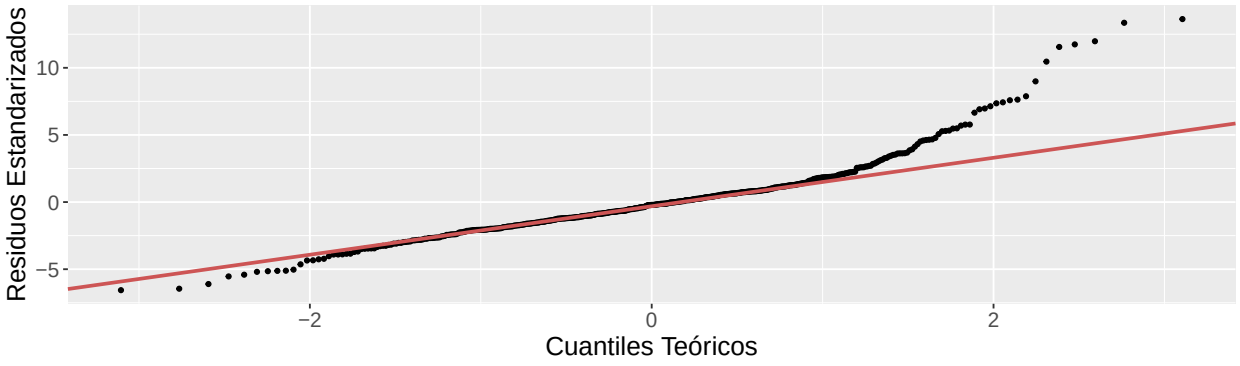



Figura 7: Gráfico QQ Plot.

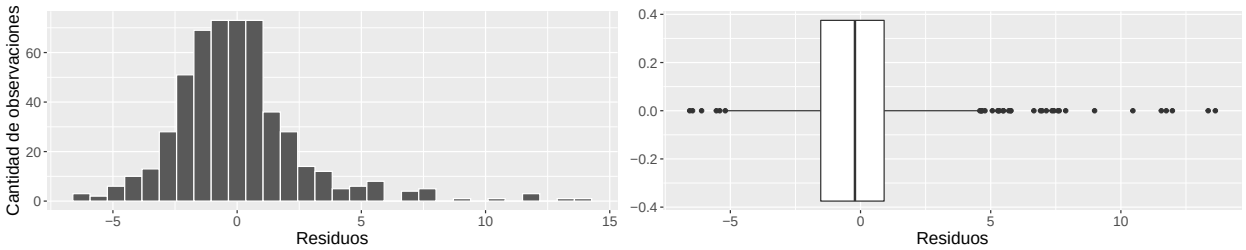



Figura 8: Histograma y boxplots de los residuos.

En nuestros gráficos podemos observar que la distribución de los residuos se asemeja bastante a una normal, pero existen observaciones con valores altos de los residuos distorsionan esta semejanza.

Confirmamos la falla en el cumplimiento del supuesto de normalidad mediante tests de hipótesis, los de Kolgomorov-Smirnov (Lilliefors), Jarque-Bera y el de Shapiro-Wilk. Todos los tests tienen un p-valor próximo a 0, por lo que nos llevan a la misma conclusión, rechazar la hipótesis nula de que la distribución de los residuos es normal. Por lo tanto como también encontramos problemas con la normalidad de los errores, nos resulta conveniente realizar la transformación Box-Cox de la variable explicada salario para enfrentar estos problemas.

La transformación de Box-Cox consiste en considerar la función de verosimilitud del salario y hayar su máximo para un conjunto dado de valores $\lambda$, generalmente entre -2 y 2. La transformación logaritmica (que sugiere el modelo de Mincer) la utilizamos en el caso de que $\lambda = 0$.

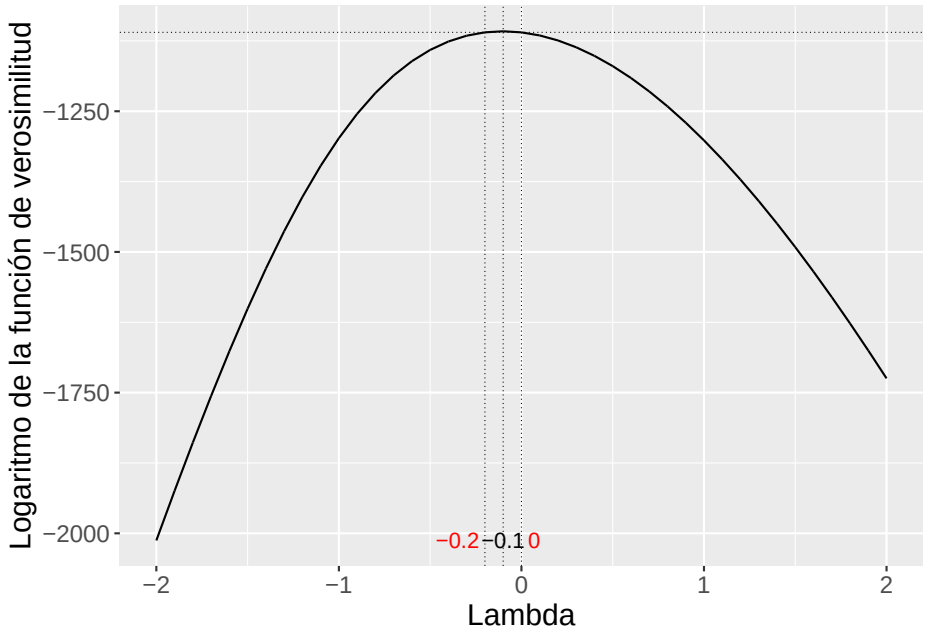



Figura 9: Gráfico de la transformación de boxcox, con intervalo de confianza para el valor maximo de lambda.

El intervalo de confianza al 95 % del verdadero valor maximal de $\lambda$ es $(-0{,}2; 0{,}0)$. Como el 0 se encuentra en dicho intervalo, podemos utilizar la transformación logarítmica especificada en la transformación de Box-Cox. Como además esto se corresponde con la especificación comunmente utilizada y respaldada por la teoría planteamos un nuevo modelo donde la variable explicada pasa a ser el logaritmo del salario.

Nos disponemos ahora a analizar nuevamente qué ocurre con los supuestos una vez realizada la transformación.

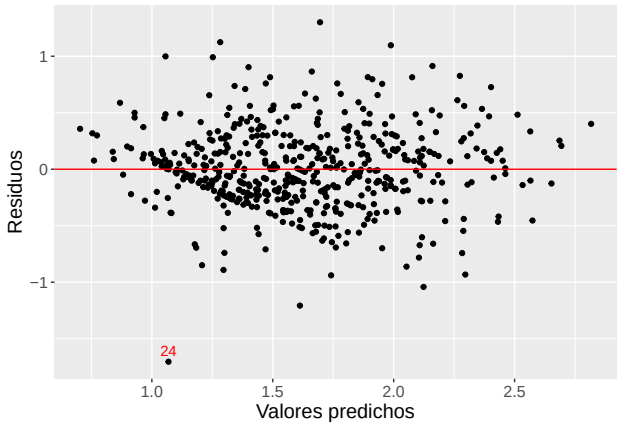



Figura 10: Gráfico de residuos en función de los valores predichos para el modelo logarítmico.

El gráfico de los residuos nos muestra que ahora ya no está tan marcado el patrón de la distribución que anteriormente veíamos. El test de homocedasticidad de Breusch-Pagan nos lleva a no rechazar la hipótesis nula de que los residuos tienen igual varianza con un nivel de significación del 5 %, con un p-valor de 0.055. Notamos que hay un residuo que se aparta de los demás, que corresponde a la observación número 24.

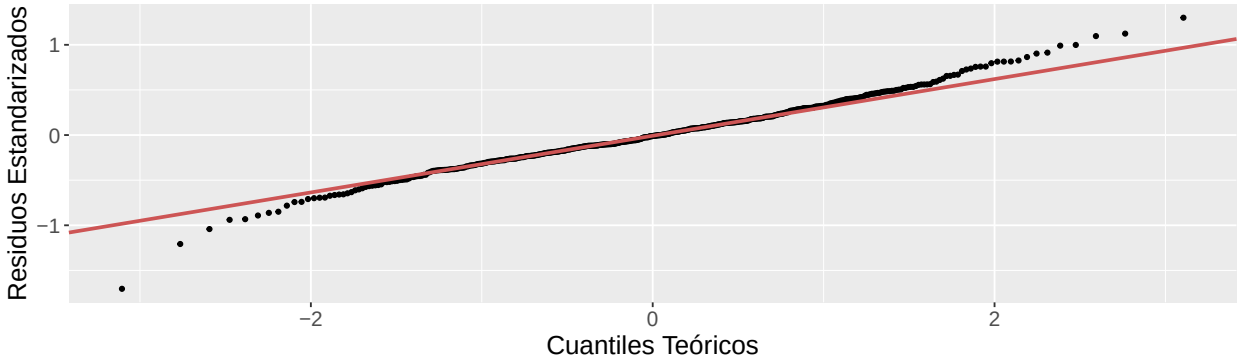



Figura 11: Gráfico QQ-Plot para el modelo logarítmico.

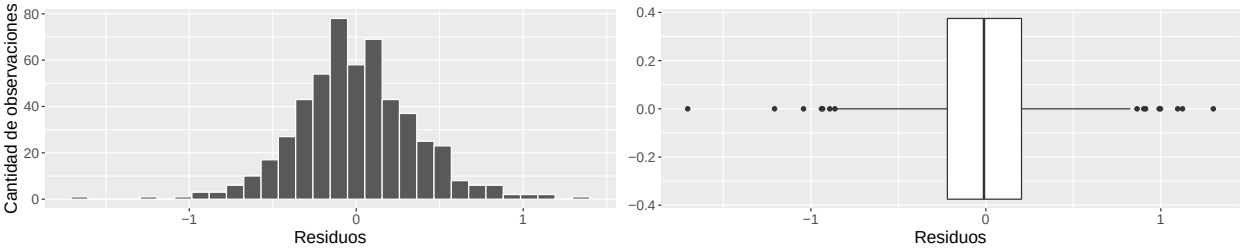



Figura 12: Histograma y boxplots de los residuos para el modelo logarítmico.

En cuanto a la normalidad, en comparación con el modelo sin la transformación, podemos apreciar como en los tres gráficos las características de la distribución normal son más claras. En el QQ-plot los puntos se ajustan más a la recta de cuantiles teóricos, los cuantiles empíricos de los residuos se asemejan más a los cuantiles teóricos de la normal, el histograma se asemeja más a la forma de la normal y el *boxplot* nos

muestra una mayor simetría y una concentración mayor en torno al 0. Sin embargo en los tres casos todavía podemos notar observaciones atípicas, que alejan la distribución de los residuos de la normal que buscamos.

Efectivamente, al realizar los tests de normalidad, obtenemos p-valores de 0.0278 para el test de Lilliefors, $3{,}86 \times 10^{-10}$ para el de Jarque-Bera y $1{,}38 \times 10^{-4}$ para el de Shapiro-Wilk, por lo que seguimos rechazando la hipótesis nula de que los residuos se distribuyen normal,. Graficamente esto lo asociamos a las observaciones atípicas, por lo que intentaremos trabajarlas para lograr el cumplimiento de los supuestos.

### 6.1.3. Tratamiento de observaciones atípicas

Como primer aproximación al problema, realizamos el test de Bonferroni para detectar outliers. El mismo nos indica que la observación 24 es un outlier con cambio en media, con un p-valor ajustado por el método de Bonferroni de $8{,}89 \times 10^{-4}$.

Eliminamos esta observación de la base, debido a que como habíamos observado antes, la observación 24 era la que presentaba un mucho mayor apartamiento de las demás observaciones en cuanto al valor de su residuo r-estudentizado, por lo cual considerábamos que era influyente. Al realizar nuevamente los tests de homocedasticidad y normalidad, vemos que los p-valores mejoran sustancialmente, si bien aún no estamos en condiciones de afirmar que se cumplen con un nivel de significación del 5 %. Por tal motivo, continuamos explorando para ver que sucede con otras observaciones.

Realizando nuevamente el test de Bonferroni, encontramos que no existen más outliers con cambio en media, según el resultado de este test. En este punto, a modo meramente exploratorio, optamos por eliminar sistemáticamente observaciones de la base en una en una de acuerdo al valor absoluto de su residuo r-student (eliminamos el que tenga valor más alto), y en cada paso realizando los tests de homocedasticidad y normalidad. Encontramos que aún eliminando muchos residuos, los tests no nos dan un resultado favorable de que se cumplan los supuestos, por lo cual optamos por ir por otro camino.

| | Observación | Residuo R-Student |
|---|---|---|
| 1 | 440 | 3.655 |
| 2 | 128 | 3.428 |
| 3 | 381 | 3.156 |
| 4 | 186 | 3.076 |
| 5 | 282 | 2.928 |

Cuadro 15: Observaciones con mayor residuo r-student en valor absoluto

En el cuadro anterior se puede apreciar las observaciones con mayor residuo r-studentizado, una vez eliminada la observación 24. Basándonos en ello, nos dispusimos a hacer algo diferente: en vez de eliminar observaciones de la base, les dimos un tratamiento especial, creando una variable indicadora para cada una de ellas que representara un salto en media para esa observación. Hicimos esto para las 3 observaciones con mayor residuo r-studentizado en valor absoluto: las 440, 128 y 381.

| | Test realizado | p-valor |
|---|---|---|
| 1 | Breusch-Pagan | 0.0525 |
| 2 | Lilliefors | 0.316 |
| 3 | Jarque-Bera | 0.0917 |
| 4 | Shapiro-Wilk | 0.102 |

Cuadro 16: p-valores para los diferentes test de normalidad y homocedasticidad

En la tabla anterior, se puede observar que el test de Breusch-Pagan no rechaza la hipótesis nula de que los errores son homocedásticos, y que ninguno de los los tests rechaza la hipótesis nula de que los errores se distribuyen normal. Por lo tanto, concluimos que eliminar la observación 24 y darle un tratamiento especial a

las observaciones 128, 381 y 440 nos permite que nuestros datos sean homocedásticos y sigan una distribución normal.

Optamos ahora por realizar toda la construcción del modelo desde el principio, porque creemos que la presencia de estas nuevas indicadoras puede afectar los tests de significación realizados. Esto se puede encontrar en el cuerpo del texto.

## 6.2. Anexo 2: Gráficos de dispersión del salario para las distintas regiones

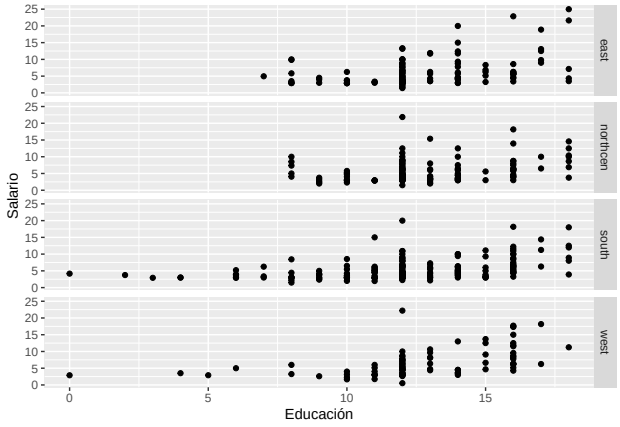



Figura 13: Gráficos de dispersión del salario en función de los años de educación, para cada región.

Observando los datos (y luego lo verificamos), podemos apreciar que mientras que para las regiones Sur y Oeste contamos con observaciones a lo largo de todos los niveles educativos, para este y nor-centro solo tenemos observaciones con un nivel educativo de a partir de 7 y 8 años de educación respectivamente. Esto nos hace dudar de si la muestra de la que disponemos realmente captura todas las características de las poblaciones de las regiones.

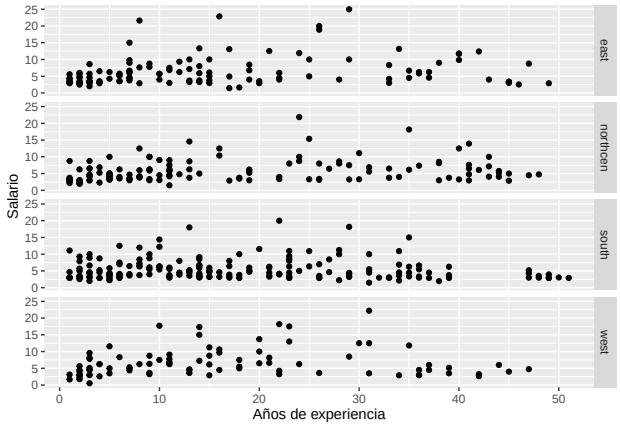



Figura 14: Gráficos de dispersión del salario en función de la experiencia, para cada región.

En este caso, podemos observar que para las distintas regiones, no cambia la dispersión de los datos del salario en función de los años de experiencia. La única diferencia que se podría llegar a apreciar es que para la región sur, la pendiente de la recta de ajuste parece ser negativa, mientras que para las demás regiones parece ser positiva.

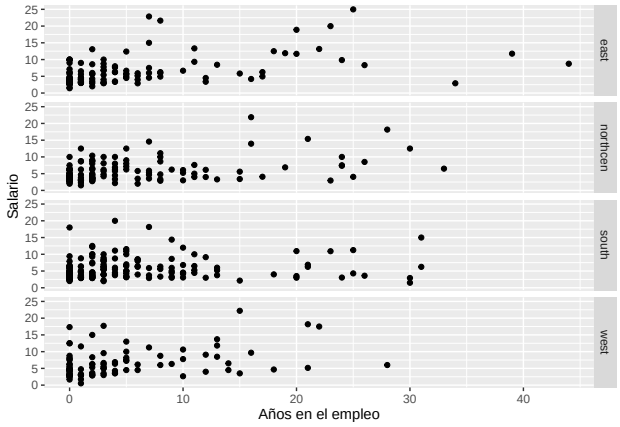



Figura 15: Gráficos de dispersión del salario en función de los años en el empleo, para cada región.

Nuevamente, observamos que algunos niveles de la variable explicativa no se observan para algunas regiones. En particular, no contamos con observaciones en la región Norte, Sur y Oeste que cuenten con más de 30-35 años en el empleo, mientras que para la región Este sí.

Por lo demás, no se observan diferencias en la dispersión.

## 6.3. Anexo 3: Modelos lineales simples

A la hora de ver una medida de la bondad del ajuste, vemos que los valores del $R^2$ son bajos, menores a 20 %, lo cual nos indica que el porcentaje de la variación muestral del salario explicada por los distintos modelos es menor al 20 %. Esto se corresponde con lo observado en las gráficas del Anexo 4, en las que vemos que las observaciones se ajustan pobremente a las rectas.

No obstante en todos los modelos tenemos p-valores muy bajos, muy próximos a 0, que indican que el contraste de significación de la variable explicativa nos lleva a rechazar la hipótesis nula de que la variable regresora no tiene nivel explicativo sobre la variabla regresada, para niveles de significación muy bajos. Podemos interpretar esto como que las variables en su conjunto pueden explicar de forma correcta al salario, pero no así de forma individual y por esto realizamos modelos múltiples.

## 6.4. Anexo 4: Gráficos de regresión lineal simples

Para comenzar, decidimos construir un gráfico de dispersión del salario en función de nuestras distintas variables explicativas: educación, experiencia, y años en el empleo. Además, le agregamos una recta de ajuste lineal

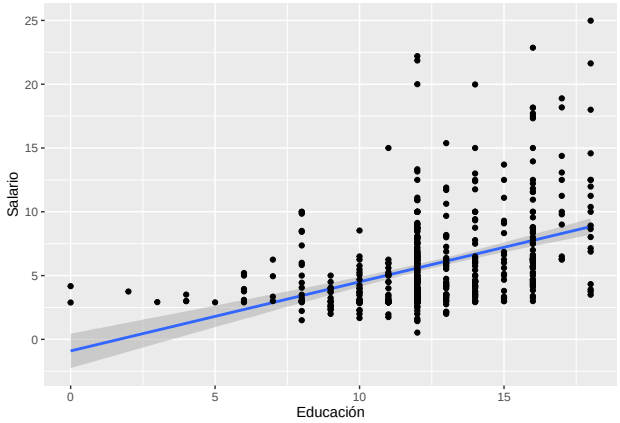



Figura 16: Gráfico de dispersión del salario en función de la educación, con recta de ajuste lineal.

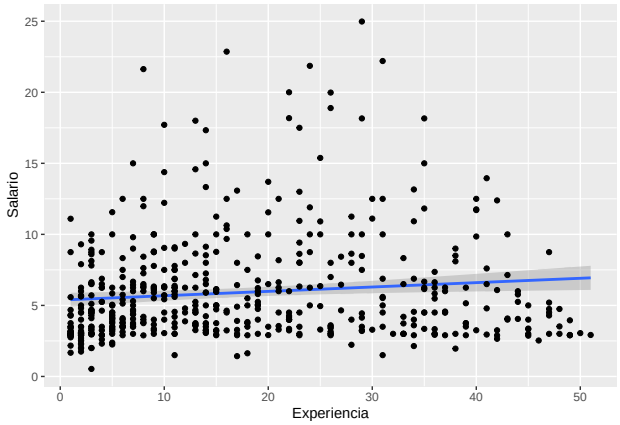



Figura 17: Gráfico de dispersión del salario en función de la experiencia, con recta de ajuste lineal.

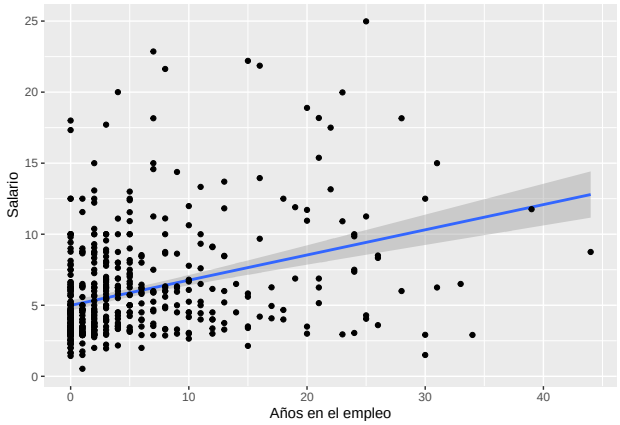



Figura 18: Gráfico de dispersión del salario en función de los años en el empleo, con recta de ajuste lineal.

En todo los casos, podemos observar que la dispersión de los datos es muy grande, y estos no parecen ajustarse a la recta. A simple vista parece haber problemas con el cumplimiento del supuesto de varianzas

de los errores constantes para todas las observaciones.

En el caso de la educación, vemos que la dispersión aumenta a medida que los años de educación aumentan. Intuitivamente, podemos interpretar esto como que una baja cantidad de años de educación implica un salario necesariamente bajo, y una cantidad alta de años de educación implica la posibilidad de tener un salario tanto alto como bajo, lo cual genera una variabilidad mayor del salario.

Para el gráfico de salario contra experiencia se puede apreciar como a niveles de experiencia mas bajos el salario toma valores que se acumulan en el tramo más bajo y a medida que aumenta la experiencia el salario experimenta mayor dispersión. Es algo a destacar que a los niveles mas altos de experiencia el salario parece tomar valores que se acumulan en el tramo menor. esto tengan incidencia las demás variables.

En el gráfico de salario contra años en el empleo se puede ver que los datos se acumulan para valores bajos de experiencia y salario y a medida que aumentan los años en el empleo, la dispersión aumenta. También se pueden notar dos observaciones influyentes, las cuales toman valores altos de la variables años en el empleo y afecta la recta de ajuste.

# 7. Bibliografía


- Income and poverty in the United States: 2017. Fontenot, Kayla et al. (2018). *Link*.
- The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings. Lemieux, Thomas. (2006). *Link*.
- Introducción a la econometría: Un enfoque moderno. Wooldrige, Jeffrey. (2009).
- Materiales del curso 2020 de Modelos Lineales