Contents

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, implementados 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, años en el empleo actual, no blanco (variable indicadora, 1 si la persona no es de raza blanca), sexo, estado matrimonial, número de dependientes, región metropolitana (vale 1 si la persona vive en dicha región), región (dividida en tres variales indicadoras, nor-centro, sur y oeste), rama de actividad (dividida en tres variables indicadoras, construcción, comercio y servicios), ocupación (tres variables indicadoras, profesional, cajero y servicio), el logaritmo de la variable salario y los cuadrados de las variables experiencia y años en el empleo actual. Contabamos con una variable indicadora más *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-centro, sur y oeste; la rama de actividad,

separada en construcción, servicios y comercio; y la ocupación, separada en profesional, cajeros 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 como 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 las siguientes tablas:

3.1.1 Región

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

Tenemos 3 variables indicadoras para esta región, Norte-Centro, Sur y Oeste. Este es la categoría de referencia.

	Región	Cantidad	Minimo	Media	Maximo
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

Table 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 no presentan una diferencia considerable entre ellas para las distintas regiones. Lo mismo sucede con los máximos, los cuales son similares.

Por otro lado, 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 1, y 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	Minimo	Media	Maximo
1	Construccion	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

Table 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 variable "Otros", que es la categoría 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 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 incluida 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. En particular, la categoría Servicios tiene la menor media, y tambíen el menor mínimo y el menor máximo. Luego, la categoría construcción tiene la mayor media de las 3 indicadoras de las que disponemos, y también el menor mínimo. Por otro lado, la categoría Comercio es bastante similar a la categoría de referencia para las 3 medidas de resumen.

3.1.3 Ocupación

	Ocupación	Cantidad	Minimo	Media	Maximo
1	Servicio	74	0.53	3.59	7.81
2	Cajero	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

Table 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 significativa, dado que aún no hemos realizado inferencia a partir de nustros datos.

Por otro lado, la categoría Cajero 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.
- 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.
- Para la variable raza hay un 10.27% de personas no blancas y 89.73% blancas.

4 Modelos lineales múltiples

Plantearemos un modelo lineal múltiple de la forma

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

donde ε es el vector aleatorio de errores, para el cual hacemos 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.

4.1 Selección de variables

4.1.1 Análisis de multicolinealidad

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

Table 4: Variables con mayor VIF

Utilizando la medida de el factor inflacionario de la varianza (FIV) buscamos analizar la multicolinealidad entre las distintas variables. La mayoría de las variables presentan valores del FIV pequeños, menores a 2, por lo cual no parecería haber problemas de mumlticolinealidad entre las variables. Por otro lado, las variables experiencia, educación y antigüedad y sus transformaciones al cuadrado presentan valores más altos. En particular educación y experiencia y sus transformaciones al cuadrado presentan valores que superan el 10, el cual se utiliza como regla empírica para determinar si existen problemas de multicolinealidad. Esto es esperable de todas formas ya que va a existir colinealidad entre una variable y su transformada al cuadrado.

La siguiente matriz de correlaciones nos permite visualizar que efectivamente la colinealidad solo existe (en cierto grado) de a pares entre las variables y sus transformaciones.

	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

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

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

4.1.2 Tests de hipótesis iniciales

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 como 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 consideraremos 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 da 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, a continuación presentamos una tabla con los distintos p-valores de dichos contrastes.

	.
Educación	0.11
Experiencia	0.00
Antigüedad	0.00
Raza	0.73
Sexo	0.00
Est. Civil	0.82
N° dependientes	0.97
Reg. metropolitana	0.01
Norte	0.11
Sur	0.09
Oeste	0.26
Construcción	0.86
Comercio	0.00
Servicios	0.00
Ocup. profesional	0.00
Ocup. administrativos	0.75
Ocup. servicios	0.57
Exper. al cuadrado	0.00
Antig. al cuadrado	0.40
Educ, al cuadrado	0.00

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 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 [survey 2017]. 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 si para las ramas de actividad comercio y servicios. En cuanto a las ocupaciones, sola 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 estás 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.

Utilizando 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 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 que tiene las siguientes variables:

- Años de Educación
- Años de Educación al cuadrado
- Años de Experiencia
- Años de Experiencia al cuadrado
- Antigüedad
- Sexo
- Región metropolitana
- Indicadora de la región Norte
- Indicadora de la región Sur
- Comercio (variable indicadora de rama de actividad)
- Servicios (variable indicadora de rama de actividad)
- Indicadora de la Ocupación Profesional

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 esta en el grupo de la variable que sí es significativa.

Observamos que 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.

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) nos indica que es una variable que aporta a la explicación del salario optamos por mantenerla.

Queremos ver si es viable incluir la variable raza que en estudios ha demostrado ser importante para explicar el salario. 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 0.774, el cual resulta muy alto para optar por incluirla de todas maneras al modelo.

El R_a^2 (coeficiente de determinación ajustado), con el cual vemos la bondad del ajuste del modelo propuesto es de 0.48, lo que indica que un 48% 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 2%, lo que es esperable ya que este coeficiente penaliza por la incorporación de variables que no aportan mucho.

4.2 Análisis de los supuestos

Para comenzar con nuestro análisis de los residuos, 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.

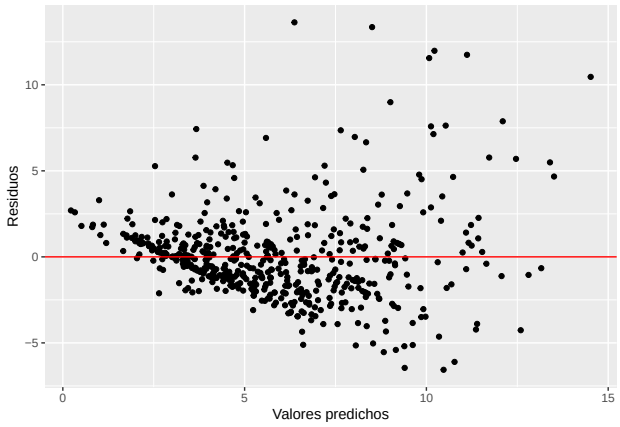
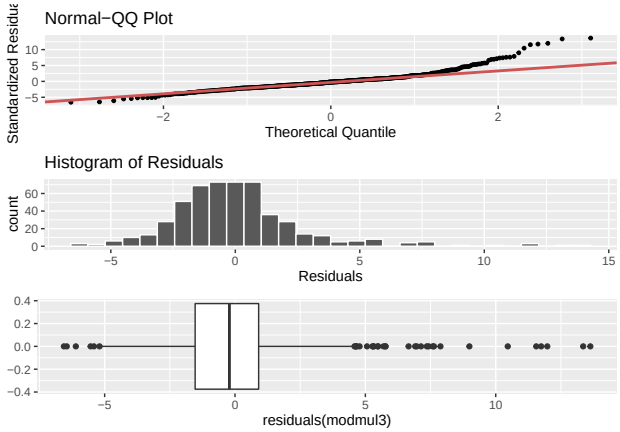


Figure 1: Gráfico de residuos en función de los valores predichos por el modelo.

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.

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).

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 1.0891457×10^{-8} 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.

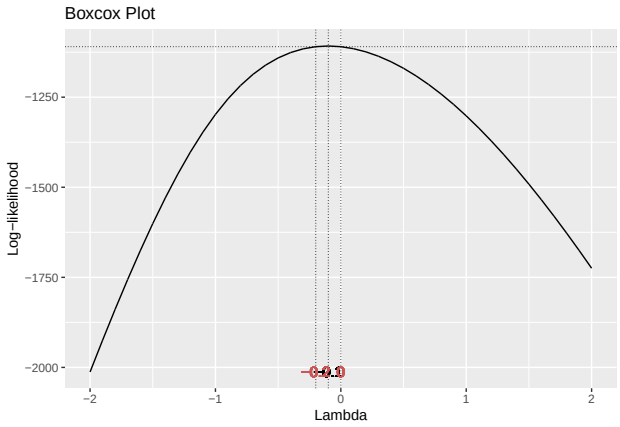


Como acercamiento a analizar el supuesto de normalidad de los errores realizamos un gráfico de cuantiles teóricos (QQ plot), un histograma de los residuos y un *boxplot* de los residuos. La primera de las figuras compara los cuantiles de la distribución de los residuos con los cuantiles teóricos de la 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 la distribución de los residuos se asemeja bastante a una normal, pero los residuos con valores altos que se alejan de la probabilidad acumulada que uno esperaría de una distribución normal. (DESARROLLAR)

Nuevamente 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 y el de Jarque-Bera, que se basa en los coeficientes de simetría y curtosis de la muestra. Ambos 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.

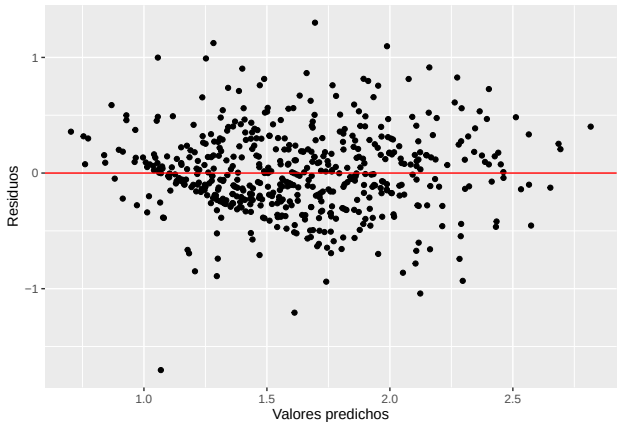
La transformación de Box-Cox consiste en considerar la función de verosimilitud del salario y hayar su máximo para un conjunto dado de valores λ, generalmente entre -2 y 2. La transformación logaritmica que sugiere el modelo de Mincer la utilizamos en el caso de que $\lambda = 0$.

```
## [1] -0.2  0.0
```

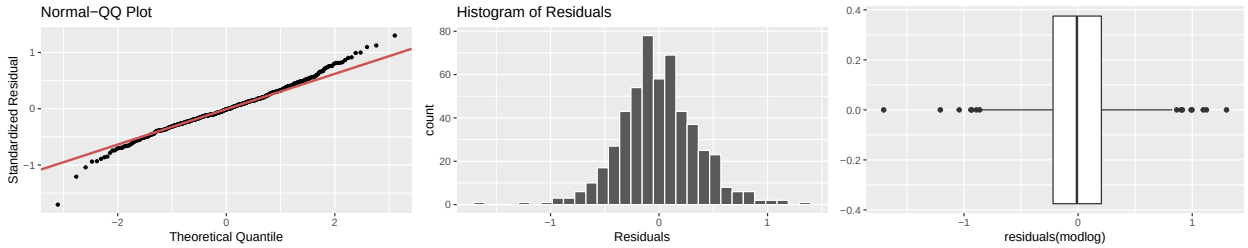


El intervalo de confianza al 95% del verdadero valor maximal de λ es $ [-0.2,0.0]$, como el 0 se encuentra en dicho intervalo, podemos utilizar las transformación logarítmica especificada en la transformación de Box-Cox. Por lo tanto planteamos un nuevo modelo donde la variable explicada pasa a ser el logaritmo del salario, la cual multiplicada por 100 es interpretada como la variación porcentual del salario, en lo que profundizaremos luego.

Nos disponemos ahora a analizar nuevamente que ocurre con 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. 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%. Notamos que hay un residuo que se aparta de los demás, que corresponde a la observación número 24.



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 vemos que seguimos rechazando la hipótesis nula de que los residuos se distribuyen normal. Graficamente esto lo asociamos a las observaciones atípicas, por lo que tendremos que trabajarlas para lograr el cumplimiento de los supuestos.

5 Anexos

5.1 Anexo 1: Gráficos de dispersión del salario para las distintas regiones

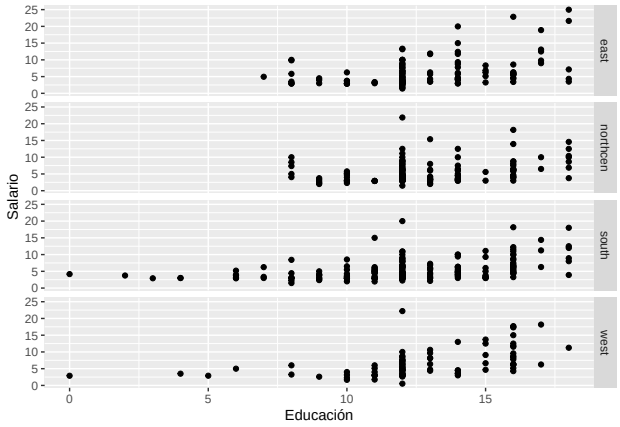


Figure 2: Gráficos de dispersión del salario en función de los años de educación, para cada región.

```
## [1] 7
```

```
## [1] 8
```

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 la población para esas regiones.

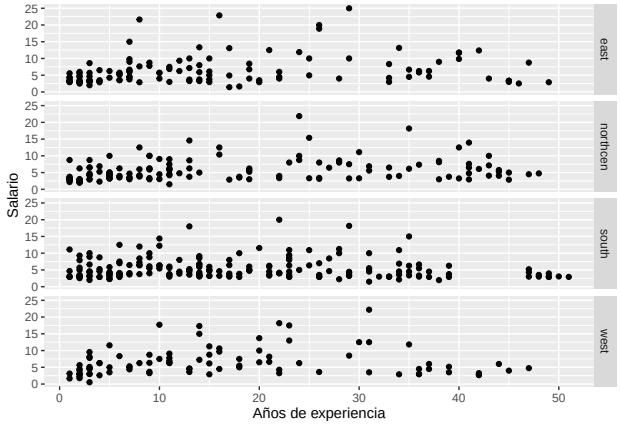


Figure 3: Gráficos de dispersión del salario en función de la experiencia, para cada región.

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.

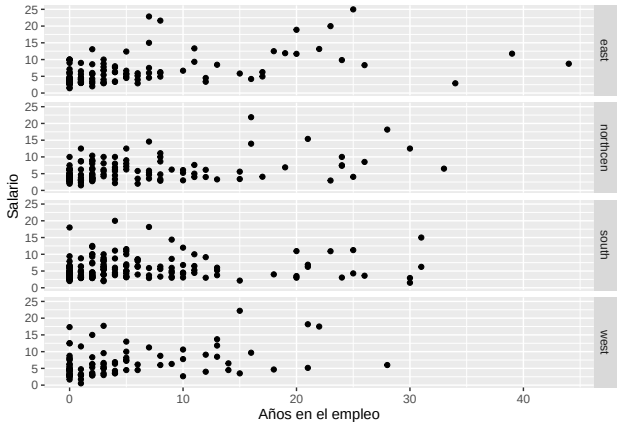


Figure 4: Gráficos de dispersión del salario en función de los años en el empleo, para cada región.

Nuevamente, observamos que algunos niveles de la variable explicativa no ser observan para algunas regiones. En particular, no contamos con observaciones en la region nor-centro, 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.

5.2 Anexo 2: 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 el modelo es menor al 20%. Esto se corresponde con lo observado en las gráficas, 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.

Hay una discordancia entre los valor del R^2 y la conclusión resultante de los contraste de hipótesis, ya que por un lado tenemos que las observaciones no se ajustan muy bien a las rectas, pero por el otro, vemos que las variables regresoras son significativas para explicar el salario. Puede que ocurra que las variables no se relacionen linealmente y requieran algún tipo de transformación para ver el relacionamiento.

También podemos interpretar esto como que las variables en su conjunto pueden explicar de forma correcta al salario, pero no así de forma individual. Debemos seguir avanzando y realizar modelos múltiples para observar si esto efectivamente es así.

5.3 Modelos de regresión 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

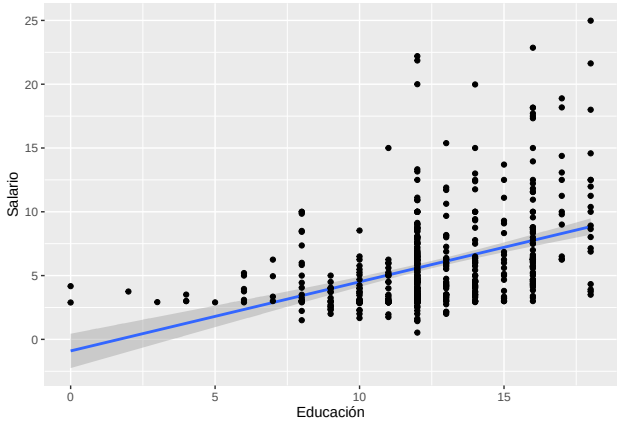


Figure 5: Gráfico de dispersión del salario en función de la educación, con recta de ajuste lineal.

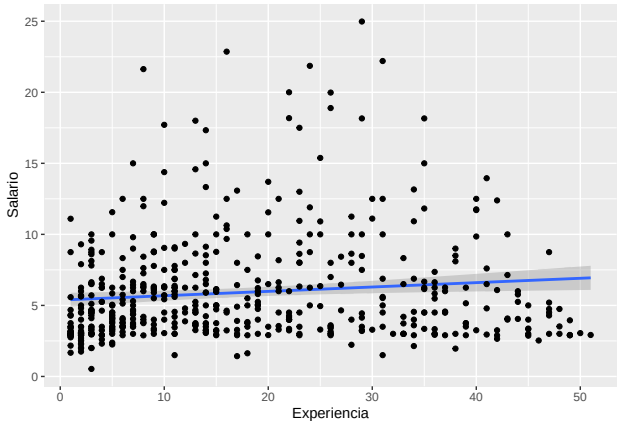


Figure 6: Gráfico de dispersión del salario en función de la experiencia, con recta de ajuste lineal.

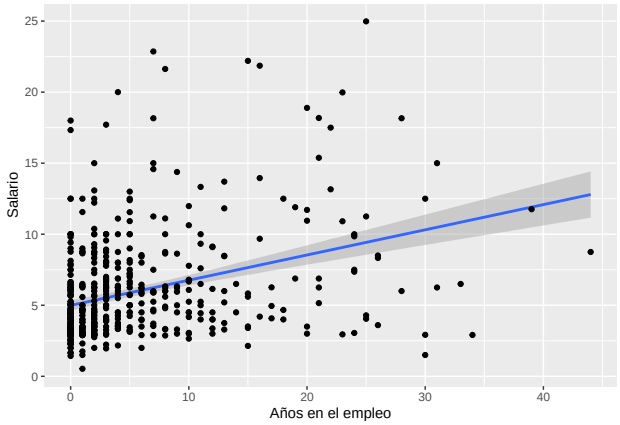


Figure 7: Gráfico de dispersión del salario en función de los años en el empleo, con recta de ajuste lineal.

En todo los casos, podemos observar que la dispersión de los datos es muy grande, y estos no parecen ajustarse a la recta. En particular, podemos ver que no se cumple el 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.

5.4 Anexo 3