Juan Sebastián Durán Durán

Pablo Cerezo Lesmes

Problem Set 1: Predicting Income MECA 4107 Due Date: June 27 at 23:59 on Bloque Neón

Link GITHUB: https://github.com/pcerezol/ML-TALLER1

1. Data acquisition

(a) Scrape the data that is available at the following website

https://ignaciomsarmiento.github.io/GEIH2018 sample/

(b) Are there any restrictions to accessing/scraping these data?

Al realizar el scraping si hay una serie de restricciones dependiendo si la página web es dinámica o tiene un formato estático. Es decir, si la página a la que se está intentando hacer scraping tiene un formato el cual no cambia y en donde todos los elementos vienen escritos en html entonces se diría que es una página estática. En ese sentido, realizar scraping se remite a identificar en la página de internet el texto, tabla o información que se quiere extraer y codificarlo de acuerdo con la url inicial para que r ejecute la acción de extraer la información. Por el otro lado, cuando la página es dinámica esta está escrita en diferentes lenguajes de programación y contiene diferentes elementos interactúan con el usuario. Lo anterior implica que, se genera un cambio en la visualización en la medida que el usuario ejecuta diversas acciones en la página. Por lo tanto, para hacer scraping no basta con ejecutar el comando, sino que, en principio, es necesario identificar dentro de la página el lugar específico donde se encuentra ubicada la información y posteriormente, identificar los elementos de ese sitio que se quieren extraer. En estas páginas, dada la cantidad de información, esta se almacena en servidores, una página dinámica hace solicitudes para ir trayendo esa información a medida que el usuario interactúa, precisamente este punto mencionado es lo que hace que una página sea dinámica.

(c) Using pseudocode describe your process of acquiring the data

1. Para extraer la información nos dirigimos a la página que contiene las tablas que vamos a extraer.
2. Se identifica si es una página de características estática o dinámica.
3. En este caso se identifica la url con la cual se va a hacer el scraping - "https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html
4. Como se desea unificar la información de las diferentes tablas se puede hacer el proceso de dos formas, ir creando una tabla y debajo de esa pegar la nueva base la cual se le está haciendo scraping o realizar un loop que itere el comando por las diferentes páginas que contienen la información y pegarlo, en el caso propio se escoge hacer un loop.
5. El primer paso que se realiza es crear una nueva tabla de datos que en principio está vacía dado que no se ha hecho scraping.
6. Para iniciar el loop se utiliza la función for (i in #:#) esto le indica a el programa que, en el elemento que se le va a señalar, repita una acción cuando encuentre el elemento i, esta acción se realizará a partir del número de veces que se señale.
7. Dentro del loop, llamamos url al elemento en donde se va a concatenar los elementos sin ningún separador, para eso se utiliza la función **paste0.**

8. De la mano al paso anterior, en la función paste0 se modificará un elemento de la url para que r sepa en donde va a hacer el loop. En ese sentido, dentro de la url, el número que se encuentra al final es el que está indicando la tabla que se está visualizando, por tanto, en vez del número 1 de la url se pondrá ", i , " esto le indicará a R, la ubicación de cada una de las tablas que va a ir extrayendo y concatenando la información.
9. En un archivo temporal "temp" le vamos a indicar a r que lea el elemento que definimos anteriormente como "url" y que ahí mismo – utilizando la función pipe %>% - convierta la tabla en un data frame de r.
10. Con la función del loop este elemento va a ir haciendo el scrapin por las diferentes tablas, y en nuestra base de datos creada inicialmente con el la función rbin va a ir pegando una debajo de otra las 10 tablas del ejercicio.

2. Data Cleaning

Para limpiar la base de datos fue necesario en primer lugar, qué contiene la base de datos y a partir de la intuición económica ir removiendo o generando nuevas variables.

En ese sentido, la base de datos inicial contiene 32,177 observaciones y 178 variables. Dado que, en el problema están especificando que se debe trabajar con personas mayores de 18 años y aquellas que se encuentren trabajando, lo primero que se debe remover son las observaciones que no cumplen estas condiciones.

Si creamos un nuevo conjunto de datos con estas especificaciones entonces resultamos con un conjunto de 16,542 observaciones. Dado que, en la base de datos hay 178 observaciones entonces para este trabajo debemos utilizar la lógica económica detrás de los modelos e incluir únicamente las variables relevantes. Por tanto, en este caso en donde se busca evaluar los ingresos que recibe una persona se dejaron las variables que se ven en la tabla 1, al remover variables que no se utilizaran y aquellos registros que no cuentan con información (NA) el número de observaciones desciende a 16,276.

Tabla 1. Variables Relevantes

Variable type: numeric

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1	sex	0	1	0.533	0.499	0	0	1	1	1	<U+2587>
2	ingtot	0	1	1798246.	2687773.	15000	800000	1069453	1750844.	85833333.	<U+2587>
3	age	0	1	39.4	13.4	18	28	38	50	93	<U+2587>
4	formal	0	1	0.596	0.491	0	0	1	1	1	<U+2587>
5	tipo_ocu	0	1	2.21	1.49	1	1	1	4	9	<U+2587>
6	urbano	0	1	1	0	1	1	1	1	1	<U+2587>
7	tiempo_tra	0	1	63.8	89.5	0	7	24	84	720	<U+2587>
8	educ	0	1	11.6	3.59	0	10	11	15	15	<U+2587>

En la ecuación de Mincer (1974) el ingreso aumenta a través de dos vías, la educación y la experiencia, pero a su vez se deben tener en cuenta otras variables que interactúan en este modelo. En ese sentido, a partir de la literatura, algunas de las variables que conservamos en la base de datos debido a que permiten explicar los ingresos son: la edad, el sexo, si el empleo es formal, el tipo de ocupación, si la persona se encuentra en la zona urbana, el tiempo de trabajo en esa empresa y finalmente el nivel de educación.

En el caso de género, esta variable se conserva en la base de datos ya que actualmente se ha observado que hay una discriminación negativa que afecta el género femenino. Lo anterior implica que, para mismos cargos, las mujeres pueden tener un ingreso inferior.

En el caso de la edad, esta variable se conserva pues se espera que a medida que aumente la edad la persona tenga mayores ingresos, sin embargo, se debe tener en cuenta que después de cierta edad la tendencia es que los ingresos tengan una tendencia decreciente y por lo tanto se crea una variable de edadˆ2 que siga esta forma concava.

Se incluye la variable de empleo formal y tipo de empleo pues estas características laborales determinan las condiciones laborales en que se encuentran las personas. En ese sentido, se espera que, por ejemplo, un empleado formal se espera que gane más que un empleado informal. Así mismo, un patrón o empleador debería tener un ingreso más alto que el de un trabajo doméstico.

Así mismo se incluyen la variable de tiempo de trabajo en una empresa, esto se puede convertir en un proxy de experiencia, si una persona está más tiempo en una empresa se esperaría que aumente su experiencia y el ingreso se incremente. La experiencia tiene la misma forma concava que la educación, por tanto, en principio es creciente, pero luego tiene rendimientos decrecientes.

Otra de las variables que se incluyeron fue la de zona urbana y rural, si bien se está trabajando en Bogotá que es una ciudad cuya población es en mayor medida urbana, puede ser que los salarios también presenten una discriminación por esta variable.

Finalmente, de acuerdo con el modelo de Mincer (1974) también se incluyó información sobre educación. En ese sentido, a mayor número de años cursados y completados, el ingreso debería presentar rendimientos crecientes.

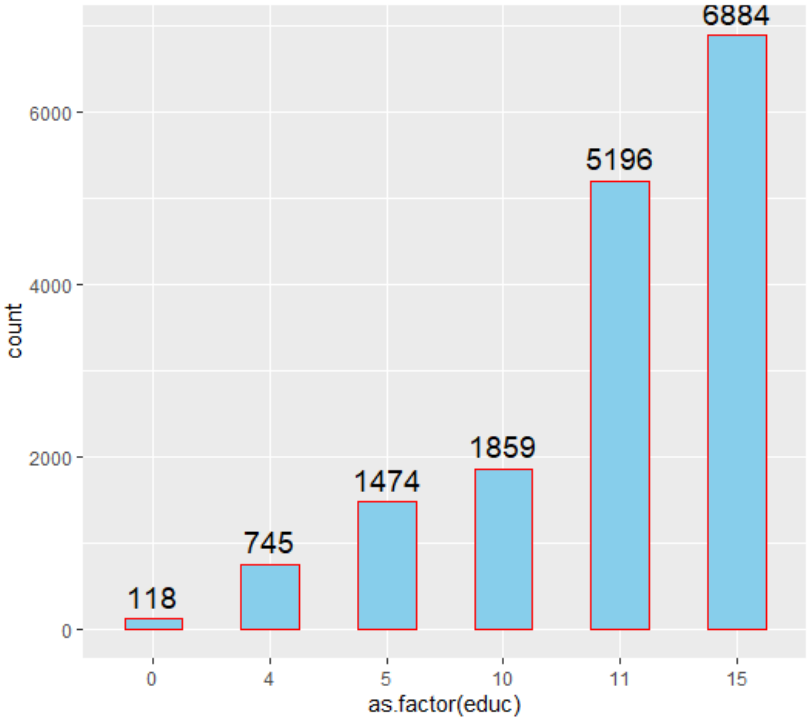
Las siguientes tablas e información enseñan la camntidad de hombres y mujeres por cada grado de educación:

Tabla cruzada grados de educación y género.

Cell Contents
|---|
Count
Column Percent
|---|

Educ	Sex 0	1	Total
0	61 0.8%	57 0.7%	118
4	309 4.1%	436 5.0%	745
5	631 8.3%	843 9.7%	1474
10	772 10.2%	1087 12.5%	1859
11	2277 30.0%	2919 33.6%	5196
15	3545 46.7%	3339 38.5%	6884
Total	7595 46.7%	8681 53.3%	16276

Gráfico de años de estudio

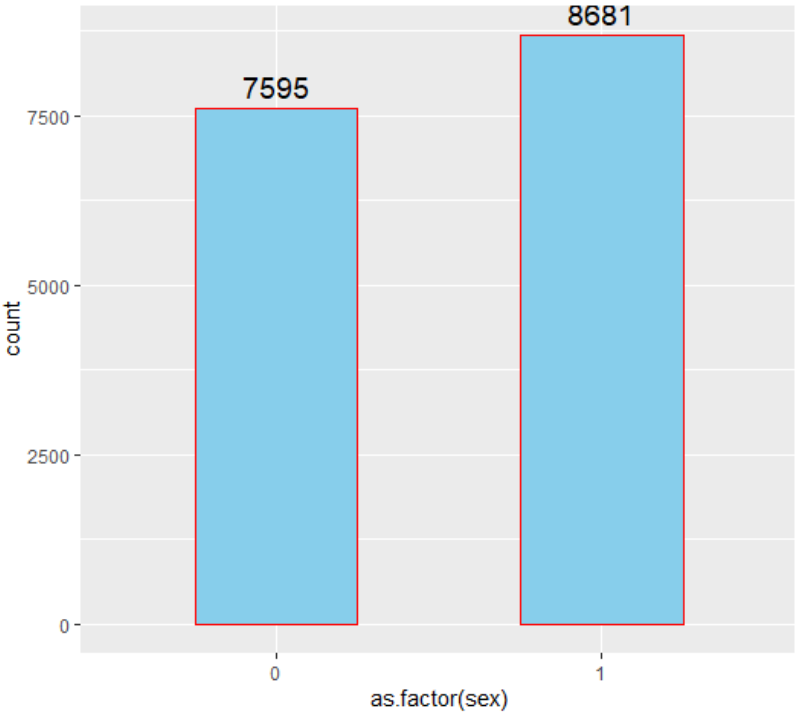


Etiqueta de dato de educación: 0 = Ninguno | 4 = Primaria incompleta

5 = Primaria completa | 10 = Secundaria incompleta |11 = Secundaria completa

15 = Educación superior)

Gráfico de Género



Género 0 = Hombre | 1 = Mujer

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

• In the data set, multiple variables describe income. Choose one that you believe is the most representative of the workers' total earnings, justifying your selection.

Los datos quedan justificados en el punto inmediatamente anterior.

• Based on this estimate using OLS the age-earnings profile equation:

$$Income_i = \beta_1 + \beta_2 Age_i + \beta_3 Age^2{}_i + u_i$$

	(1)
(Intercept)	-391868.524 *
	(181975.866)
age	89535.612 ***
	(9078.642)
age2	-772.020 ***
	(105.220)
N	16276
R2	0.018
logLik	-263903.683
AIC	527815.366

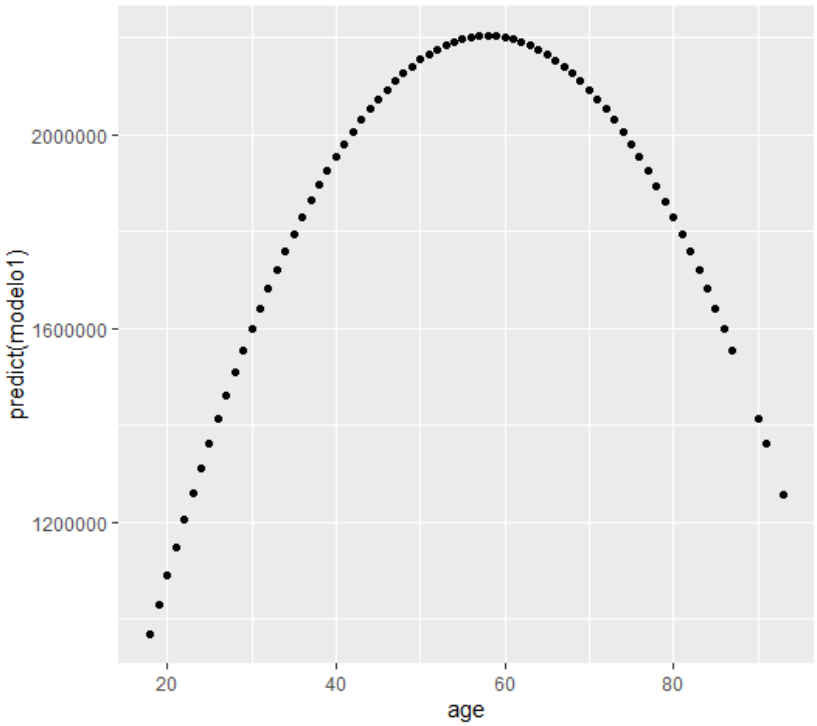
*** p < 0.001; ** p < 0.01; * p < 0.05.

• How good is this model in sample fit?

Actualmente se observa un mal modelo, las variables incluidas solo logran explicar un 1.8% del salario. Sin embargo, vale la pena resaltar que las variables siguen la forma esperada del modelo, es decir, en principio, un año de edad cambia aumenta positivamente los ingresos de una persona, sin embargo, a medida que la persona envejece llega a un punto de inflexión y comienza a decrecer, en otras palabras, un año de edad más, después de los 56 años muestran un decrecimiento en los ingresos de la persona.

• Plot the predicted age-earnings profile implied by the above equation.

Gráfico de valores predichos de salario - edad



• What is the "peak age" suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.

Para calcular la edad pico de ingresos se hará la derivada parcial de la regresión respecto a la edad y luego se igualará a cero para obtener le máximo de la función.

$$\frac{\partial Income}{\partial Age} = \beta_2 + 2\beta_3 Age = 0$$

$$\rightarrow Age = -\frac{\beta_2}{2\beta_3}$$

Al reemplazar los coeficientes por los obtenidos en la regresión obtenemos:

$$Age *= -\frac{89.516{,}9}{2 * -771{,}8} = 56$$

Con esta operación se encuentra que la edad en la que se maximizan los ingresos es 56 años.

Los resultados del ejercicio de Bootstrap son los siguientes:

Bootstrap Statistics :			
	original	bias	std. Error
t1*	-391.598	4.886,70	232.575,90
t2*	89.516,85	-248,565	13.310,12
t3*	-771,785	2,707	170,2

Para calcular el intervalo de confianza se tomará el estimador y se le sumará y restará el error estándar

Intervalo de confianza t2*

$$Int.Conf = 89.516 \pm 13.310{,}12$$
$$Int.Conf = 76.206{,}729 : 102.826{,}977$$

Intervalo de confianza t3*

$$Int.Conf = -771{,}785 \pm 170{,}2$$
$$Int.Conf = -941{,}985 : -601{,}585$$

4. The earnings GAP. Most empirical economic studies are interested in a single low dimensional parameter, but determining that parameter may require estimating additional "nuisance" parameters to estimate this coefficient consistently and avoid omitted variables bias. Policymakers have long been concerned with the gender earnings gap

4.1 Estimate the unconditional earnings gap:

$$\boldsymbol{log(Income)_i = \alpha_1 + \alpha_2 Female_i + u_i}$$

	(1)
(Intercept)	14.064 ***
	(0.009)
female	-0.193 ***
	(0.014)
N	16277
R2	0.012
logLik	-20871.957
AIC	41749.914

*** p < 0.001; ** p < 0.01; * p < 0.05.

4.2 How should we interpret the β2 coefficient? How good is this model in sample fit?

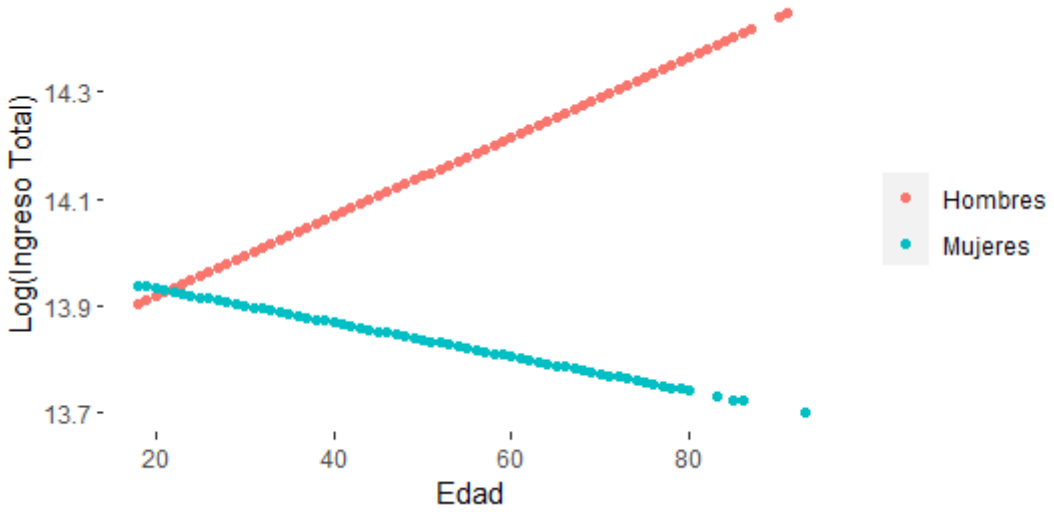
Al tener una única variable dummy, se puede interpretar como la diferencia de medias de salario entre hombres y mujeres. La regresión nos arroja que, en promedio, las mujeres ganan 19.3 % menos que los hombres. El *sample fit* es bastante bajo con un valor de 0,012. Es decir, la regresión se ajusta muy poco a los datos.

4.3 Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes?

	Hombres	Mujeres
(Intercept)	13.767 ***	13.996 ***
	(0.027)	(0.034)
age	0.007 ***	-0.003 ***
	(0.001)	(0.001)
N	8682	7595
R2	0.016	0.002
logLik	-10536.591	-10197.525
AIC	21079.181	20401.050

*** p < 0.001; ** p < 0.01; * p < 0.05.

Gráfico edad – Log(ingresos)



Las dos regresiones presentan diferentes pendientes e interceptos. Las mujeres tienen un intercepto mayor al de los hombres, 13,996 contra 13,767 respectivamente, sin embargo, la regresión para las mujeres presenta una pendiente negativa de 0,003 en comparación con la pendiente positiva para los hombres de 0,007. Esto indica que entre mayor sea la mujer, esta tendrá menores ingresos. En cambio, la regresión para los hombres tiene una pendiente positiva, lo que indica que entre mayor edad tiene, mayores serán sus ingresos.

4.4 What is the implied "peak age" by gender?. Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?

La edad pico de ingresos por género para las mujeres es al comienzo de su vida laboral, debido a la muestra que se tuvo, es a los 18 años. En cambio, la edad pico de ingreso de los hombres es al final de su vida laboral.

El intervalo de confianza para los hombres es el siguiente:

Bootstrap Statistics :			
	original	bias	std. Error
t1*	13,766	0,006	0,0284
t2*	0,007	-0,0001	0,0007

$$Int.Conf = 0.007 \pm 0{,}0007$$
$$Int.Conf = 0{,}0063 : 0{,}0077$$

El intervalo de la confianza de las mujeres es el siguiente:

Bootstrap Statistics :			
	original	bias	std. Error
t1*	13,995	0,0007	0,0323
t2*	-0,003	-0,00002	0,0008

$$Int.Conf = -0.003 \pm 0{,}0008$$
$$Int.Conf = 0{,}0022 : 0{,}0038$$

Los intervalos de confianza no se superponen.

4.5 Equal Pay for Equal Work? A common slogan is "equal pay for equal work". One way to interpret this is that for employees with similar worker and job characteristics, no gender earnings gap should exist. Estimate a conditional earnings gap that incorporates control variables such as similar worker and job characteristics (X).

(a) Estimate the conditional earnings gap $log(Income)_i = \delta_1 + \delta_2 Female_i + \theta X + u_i$

La regresión que se decidió correr fue la siguiente:

$$log(Income)_i = \delta_1 + \delta_2 Female_i + \delta_3 Edad_i + \delta_4 Relab_i + \delta_5 Formal_i + \delta_6 Educ_i + \delta_7 Tiempo\_tra_i + \delta_2\delta_2 + u_i$$

Dónde:

Female: es una variable dicótoma que toma el valor de 1 si el indivicuo es mujer y 0 en otro caso.

Relab: Es el tipo de ocupación del individuo.

Formal: es una variable dicótoma que toma el valor de 1 si el empleo de la persona es formal y 0 en otro caso.

Educ: Son los años de educación del individuo.

Tiempo_tra: Es cuántos meses lleva la persona trabajando.

	(1)
(Intercept)	12.376 ***
	(0.032)
female	-0.217 ***
	(0.011)
age	0.009 ***
	(0.001)
tipo_ocu	-0.020 ***
	(0.005)
formal	0.643 ***
	(0.014)
educ	0.079 ***
	(0.002)
tiempo_tra	0.001 ***
	(0.000)
N	16277
R2	0.342
logLik	-17567.918
AIC	35151.836

*** p < 0.001; ** p < 0.01; * p < 0.05.

(b) Use FWL to repeat the above estimation, where the interest lies on β2. Do you obtain the same estimates?

Aquí se muestran los resultados de la regresión del punto anterior y utilizando el teorema FWL. Vemos que los coeficientes de female para el modelo completo y female_resi para el modelo de FWL son los mismos, al igual que los residuales.

	Modelo completo	FWL
(Intercept)	12.376 ***	0.000
	(0.032)	(0.006)
female	-0.217 ***	
	(0.011)	
age	0.009 ***	
	(0.001)	
tipo_ocu	-0.020 ***	
	(0.005)	
formal	0.643 ***	
	(0.014)	
educ	0.079 ***	
	(0.002)	
tiempo_tra	0.001 ***	
	(0.000)	
female_resi		-0.217 ***
		(0.011)
N	16277	16277
R2	0.342	0.022
logLik	-17567.918	-17567.918
AIC	35151.836	35141.836

*** p < 0.001; ** p < 0.01; * p < 0.05.

(c) How should we interpret the β2 coefficient? How good is this model in sample fit? Is the gap reduced? Is this evidence that the gap is a selection problem and not a "discrimination problem"?

El coeficiente β_2 se interpreta como que, las mujeres en promedio ganan 21,7 % menos que los hombres, *ceteris paribus*, con una significación del 1 %. El sample fit es de 0,342. Aunque este ajuste es el doble al del primer modelo de este punto, este sigue siendo muy bajo e indica que el modelo no es bueno haciendo predicciones. El gap aumenta en comparación al primer modelo por 2 puntos porcentuales, lo que nos indica que el gap se debe a un tema de discriminación pues, al controlar por nivel educativo, edad, el tiempo en el trabajo, tipo de ocupación y si el trabajo es formal, ya identificamos cómo estas variables afectan el ingreso, lo que refuerza la intuición de que es un sesgo de género y no por características del trabajo ni del trabajador.

5. Predicting earnings.

(a) Split the sample into two samples: a training (70%) and a test (30%) sample. Don't forget to set a seed (in R, set.seed(10101), where 10101 is the seed.)

i. Estimate a model that only includes a constant. This will be the benchmark.

ii. Estimate again your previous models

iii. In the previous sections, the estimated models had different transformations of the dependent variable. At this point, explore other transformations of your independent variables also. For example, you can include polynomial terms of certain controls or interactions of these. Try at least five (5) models that are increasing in complexity.

De acuerdo a las instrucciones, se estimaron cinco modelos, estos son:

- Modelo4 <- lm(ingtot~female+age+age2, data=GEIH_clean)
- Modelo5 <- lm(ingtot~female+age+age2+ tipo_ocu, data=GEIH_clean)
- Modelo6 <- lm(ingtot~female+age+age2+tipo_ocu+formal, data=GEIH_clean)
- Modelo7 <- lm(ingtot~female+age+age2+tipo_ocu+formal+educ+educ2 + realb_sex, data=GEIH_clean)
- Modelo8 <- lm(ingtot~female+age+age2+tipo_ocu+formal+educ+educ2+ tiempo_tra+ tiempo_tra2 + formal_sex + realb_sex+age_sex + age_sex2, data=GEIH_clean)

En orden descendente se observa que cada modelo es más complejo que el anterior. Detrás de cada modelo se iba agregando un mayor número de elementos que incluían interacciones entre diferentes variables y formas funcionales diferentes que obedezcan la lógica económica, tales como educación, experiencia o edad.

Finalmente, el último modelo incluyólas varibles de género, edad, edad al cuadrado, el tipo de ocupación, variable dicotómica de si el empleo formal, tiempo de trabajo visto como una proxy de experiencia, así como el tiempo de trabajo al cuadrado para que siga la forma funcional, las interacciones de trabajo formal y género, tipo de ocupación y género, y finalmente, edad y género y género al cuadrado.

iv. Report and compare the average prediction error of all the models that you estimated before. Discuss the model with the lowest average prediction error.

Para los cinco modelos previamente mencionados, se calculó el error de predicción promedio y se observa que a medida que crece, el error de predicción promedio disminuye, lo que indica que, a mayor

conocimiento de variables relevantes incluidas en el modelo, menor error de predicción se obtendrá. A continuación, se presenta el modelo y su respectivo error de predicción promedio.

- Modelo4

  7.074752e+12
- Modelo5

  7.023515e+12
- Modelo6

  6.568404e+12
- Modelo7

  5.953627e+12
- Modelo8

  5.89523e+12

A partir de esta información se observan dos cosas, en primera medida, el error de predicción es alto lo que nos sugiere debería identificarse mejores variables o utilizar otro tipo de modelo que prediga mejor los resultados. En segunda mediada, vale recalcar que, a pesar de que hay un error de predicción alto, al ir incluyendo variables con interacciones se observa una disminución en el error de predicción promedio. En ese sentido, es necesario entender específicamente cada una de las variables que tienen relación con la variable dependiente – ingresos totales para así reducir al mínimo este error de predicción promedio.

v. For the model with the lowest average prediction error, compute the leverage statistic for each observation in the test sample. Are there any outliers, i.e., observations with high leverage driving the results? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

En la muestra test hay un total de 4882 observaciones, al computar el estadístico de apalancamiento el programa solo se ejecuta hasta la observación 3345, lo anterior implica que fuera de esta muestra (test) alcanzan a quedar un gran número de observaciones sin identificar. Cabe resaltar que, en este conjunto de observaciones se encontró que unos pocos valores que son entre diez a veinte veces más grande que el del resto de resultados. Dado que el modelo, no es perfecto, sino que se puede realizar ajustes para tener valores más precisos, estos valores que se mencionaron anteriormente no deberían ser un punto para que la DIAN se preocupe, sino que, por el contrario, debería ser el punto de partida para ver cómo se puede mejorar el modelo y así tener una mejor predicción de valores, una vez realizado este ejercicio, si los errores continúan, la DIAN si debería poner la lupa sobre este tipo de personas.

(b) Repeat the previous point but use K-fold cross-validation. Comment on similarities/differences of using this approach.

A partir de la validación cruzada K fold se encuentran los siguientes resultados para los modelos referenciados anteriormente:

Para el modelo 4

```
Linear Regression

4882 samples
  3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3905, 3905, 3906, 3907, 3905
Resampling results:

 RMSE     Rsquared    MAE
 2474733  0.01950355  1259269
```

Para el modelo 5

```
Linear Regression

4882 samples
  4 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3905, 3907, 3905, 3905, 3906
Resampling results:

 RMSE     Rsquared    MAE
 2454598  0.03119617  1258803
```

Para modelo 6

```
Linear Regression

4882 samples
  5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3906, 3905, 3906, 3905, 3906
Resampling results:

 RMSE     Rsquared    MAE
 2381899  0.09826452  1206912
```

Para modelo 7

```
Linear Regression

4882 samples
  8 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3907, 3904, 3906, 3906, 3905
Resampling results:

  RMSE     Rsquared   MAE
  2252317  0.1912958  1151787
```

Para modelo 8

```
Linear Regression

4882 samples
  13 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 3906, 3906, 3905, 3905, 3906
Resampling results:

  RMSE     Rsquared   MAE
  2244928  0.2004101 1146749
```

Lo que se puede intuir de esta validación cruzada es que a medida que se incluyen un mayor número de variables en los modelos, el root mean squared error (RMSE) comienza a disminuir. Lo que se busca específicamente con este indicador es que sea lo menor posible ya que, entre más bajo sea este indicador, más cercano va a ser las predicciones que se realiza este modelo. De forma análoga, el rsquared a medida que volvemos más complejo el modelo y se incluyen un mayor número de variables relevantes, entonces, el rsquared comienza a aumentar, y nos dice en qué medida el modelo si queda explicado por las variables incluidas. En un inicio, con el modelo más sencillo solo se explica alrededor de un 1.9%, sin embargo, en el último modelo que se incluyó alcanzamos a explicar el ingreso de las personas en un 20%. Cabe resaltar que, entre el último y penúltimo modelo presentado no hay una gran diferencia entre los rquareds, esto demuestra que, a pesar de que le incluimos cinco variables más, estas no logran ser de todos determinantes para explicar mejor el modelo. Finalmente, a partir del mean absolute error (MAE) se puede observar la diferencia entre los valores predichos y las observaciones, lo que se busca es que este valor también sea bajo, lo que se observa en este ejercicio es que a medida que complejizamos el modelo este valor se reduce.

(c) LOOCV. With your preferred predicted model (the one with the lowest average prediction error) perform the following exercise: i. Write a loop that does the following:

El modelo preferido es el número 8.

• Estimate the regression model using all but the i − th observation

• Calculate the prediction error for the i − th observation, i.e. (yi − yˆi)

• Calculate the average of the numbers obtained in the previous step to get the average mean square error. This is known as the Leave-One-Out Cross-Validation (LOOCV) statistic. ii. Compare the results to those obtained in the computation of the leverage statistic