Taller-1

2024-02-07

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::combine() masks gridExtra::combine()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ readr::guess\_encoding() masks rvest::guess\_encoding()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors  
##   
## Attaching package: 'zoo'  
##   
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

if (!requireNamespace("zoo", quietly = TRUE)) {  
 install.packages("zoo")}

# 1. Introduccion

El objetivo de este informe es encontrar los determinantes del salario de las personas encuestadas en Bogotá y realizar una predicción del mismo. Encontrar los determinantes del salario es importante porque proporciona una guía tanto en términos personales como de política económica cuando se busca aumentar los ingresos de las personas. Desde la perspectiva de la política económica, aumentar los ingresos reales de las personas puede contribuir al mejoramiento de su calidad de vida. Por lo tanto, si se conocen los factores que pueden aumentar su capacidad adquisitiva, es importante desarrollar políticas públicas orientadas en ese sentido. Por otro lado, el estudio de los determinantes del salario también es útil para las personas que ya tienen un empleo y aspiran a mejorar sus condiciones de vida. Si se encuentra, por ejemplo, que la educación es clave, es en este ámbito donde el trabajador debe centrar sus esfuerzos para mejorar.

Desde la teoría económica, la ecuación de Mincer es un punto de partida para estudiar los determinantes del salario laboral. Este modelo estima el impacto de un año adicional de estudios en las rentas laborales de los individuos (Mincer, 1974). La ecuación tradicional de Mincer se puede estimar mediante mínimos cuadrados ordinarios (MCO), utilizando como variable dependiente los ingresos y como variables independientes los años de educación, la experiencia laboral y el cuadrado de esta última.

Donde: - son los ingresos del individuo. - es el número de años de educación formal completada. - son los años de experiencia laboral. - es el término de perturbación aleatoria que se distribuye según una Normal .

Para este estudio, se utilizará información de la Gran Encuesta Integrada de Hogares (GEIH). Esta encuesta solicita información sobre las condiciones de empleo de las personas y las características generales de la población. Por lo tanto, tenemos los insumos necesarios para encontrar los determinantes del salario.

Con base en lo anterior, podemos iniciar el análisis de las siguientes variables:

* Edad (Age)
* Cuenta propia (Self-employed)
* Estrato (Socioeconomic stratum)
* Formal (Formal employment)
* Horas trabajadas la semana pasada (Actualhours worked previous week)
* Horas trabajadas habitualmente (Usual weeklyhours worked)
* Informal (Informal employment)
* Ingreso total (Totalincome)
* Educación (Max. Education level attained)
* Oficio (Occupation)
* Experiencia (Experience)
* Salario (Salary)
* Sexo (Sex)
* Horas trabajadas(Total hours worked)
* Tamaño de la empresa (Company size)

# 2. Datos

## Lectura de los datos

El proceso para leer los datos consiste en hacer un web scraping a la información que se encuentra en esta página web: <https://ignaciomsarmiento.github.io/GEIH2018sample/.> Los datos están particionados en 10 páginas diferentes, por lo cual se tiene que hacer un proceso iterativo. En primer lugar, se guarda en la variable **url\_base** la dirección web de la que se debe extraer la información. Hay un ligero cambio respecto a la URL general y este es que las tablas se encuentran un poco más anidadas, por lo que se debe completar la URL. Como la base de datos completa se encuentra separada en 10 partes, el proceso se debe repetir el mismo número de veces. Primero se crea una lista vacía llamada **tables\_list**, en esta lista se guarda la información extraída de la página en cada iteración. Dentro del bucle **for**, se va modificando la ruta de extracción para que haga un recorrido por todas las direcciones donde se encuentran los datos. Esto es relativamente sencillo debido a que lo único que cambia es el número de la página; por esto, el iterador es un entero.

Cuando se hace el proceso iterativo en las 10 páginas, estas quedan anidadas en la lista vacía **tables\_list**. Para darle un formato manejable en el proceso de machine learning, se concatenan en una sola sabana de datos con formato **tibble**. Como la cantidad de datos es considerable, este proceso dura aproximadamente 8 minutos.

t1<-Sys.time()  
  
url\_base<-my\_url <- "https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_"  
  
tables\_list <- list()  
for (i in 1:10){  
 bucle\_url<-paste0(url\_base, i, ".html")  
 my\_html=read\_html(bucle\_url)  
 table<-my\_html %>% html\_table()  
 tables\_list[[i]] <- table}  
  
data<-dplyr::bind\_rows(tables\_list)

## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## New names:  
## • `` -> `...1`

t2<-Sys.time()  
Tiempo\_ejecucion<-t2-t1  
Tiempo\_ejecucion

## Time difference of 4.583233 mins

## Preprocesamiento de datos

La estrategia propuesta para conservar la representatividad de la muestra que permite el factor de expansión consiste en transformar esta variable en un factor de ponderación de manera que conservemos su proporción con respecto al universo sin afectar el tamaño original de la muestra, de esta manera se ahorra recursos computacionales la ejecución de los algoritmos y el tratamiento de los datos.

base\_no\_ponderada<-length(data$fex\_c)  
universo\_representado<-sum(data$fex\_c)  
  
data<- data %>%  
 mutate(data, peso=fex\_c/universo\_representado) %>%   
 mutate(data,fponderacion=peso\*base\_no\_ponderada)  
  
base\_ponderada<-sum(data$fponderacion)  
unidades\_muestrales\_fex<-length(unique(data$fex\_c))  
unidades\_muestrales\_fpon<-length(unique(data$fponderacion))  
  
tabla1<-t(tibble(base\_no\_ponderada,base\_ponderada,unidades\_muestrales\_fex,unidades\_muestrales\_fpon,universo\_representado))  
tabla1

## [,1]  
## base\_no\_ponderada 32177  
## base\_ponderada 32177  
## unidades\_muestrales\_fex 3536  
## unidades\_muestrales\_fpon 3536  
## universo\_representado 8164164

Para trabajar con los datos ponderados se replica cada registro tantas unidades muestrales represente.

data\_ponderada<-uncount(data,weights = round(data$fponderacion))  
base\_ponderada<-length(data\_ponderada$directorio)  
base\_ponderada

## [1] 33466

base\_no\_ponderada

## [1] 32177

El preprocesamiento de los datos se realiza de forma simultánea con el análisis descriptivo. Esto se debe a que, generalmente, al analizar la distribución de cada variable de forma individual, se logra identificar registros vacíos o atípicos que merecen un tratamiento aparte. Sin embargo, aplicamos unas reglas de validación iniciales para garantizar que los modelos de regresión tengan resultados satisfactorios. Primero, se cambiaron los nombres de las variables que se van a usar para que su identificación sea más sencilla. También se les dio el formato adecuado de tal forma que se respetara la naturaleza cualitativa o cuantitativa de las mismas. Los filtros más importantes de la base de datos fueron para la población en edad de trabajar. Como sugerencia del ejercicio, solo hay registros para personas mayores de edad. También se filtran las personas que están activas en el mercado laboral, ya que son estas las únicas que perciben un salario.

# Limpieza de mayores de edad: Este chunk lo usamos solo como limpieza de datos   
  
  
## Falta filtrar solo por personas   
data\_ponderada <- data\_ponderada %>%  
 rename(  
 Edad= age,  
 Estrato=estrato1,  
 Ingreso=ingtot,  
 Educacion=p6210,  
 Experiencia= p6426,  
 Salario=y\_total\_m,  
 Sexo=sex,  
 Horas\_trabajadas=totalHoursWorked  
 )   
  
data\_ponderada <- data\_ponderada %>%  
 mutate(  
 across(c(cuentaPropia,Estrato,formal,informal,maxEducLevel,oficio,Educacion,Sexo,microEmpresa),as.character)   
 )  
### Filtro por población economica activa  
  
data\_ponderada <- data\_ponderada %>%  
 filter(Edad >18 & dsi !="1")   
  
  
data\_ponderada <-data\_ponderada %>%  
 filter(p6240==1)

## Análisis exploratorio de datos individual

Para que la manipulación de los datos sea más sencilla, se tomará del dataframe original solo las columnas que se establecieron en la introducción y se guardarán en el dataframe **datah**.

data\_h <- data\_ponderada %>%  
 select(Edad,cuentaPropia,Estrato,formal,informal,maxEducLevel,oficio,Educacion,Experiencia,Salario,Sexo,Horas\_trabajadas,directorio,microEmpresa)  
  
data\_h <- data\_h %>%  
 mutate(Sexo = ifelse(Sexo == 1, 0, ifelse(Sexo == 0, 1, Sexo)))  
  
  
data\_h$maxEducLevel <- na.aggregate(data\_h$maxEducLevel, FUN = function(x) {  
 moda <- names(sort(table(x), decreasing = TRUE))[1]  
 return(moda)  
})

Después de depurar los datos, revisamos cuántos valores faltantes hay en la variable salario. Evidenciamos una cantidad considerable de valores NA, por lo tanto, procedemos a realizar la imputación de estos con el promedio del salario del directorio de cada hogar.

summary(data\_h$Salario)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 97 839575 1002356 1741928 1625000 70000000 1285

data\_h = data\_h %>%   
 group\_by(directorio) %>%   
 mutate(mean\_salario = mean(Salario, na.rm=T))  
  
data\_h = data\_h %>%  
 mutate(Salario2 = ifelse(test = is.na(Salario)==T,  
 yes = mean\_salario,  
 no = Salario))  
table(is.na(data\_h$Salario2))

##   
## FALSE TRUE   
## 13786 637

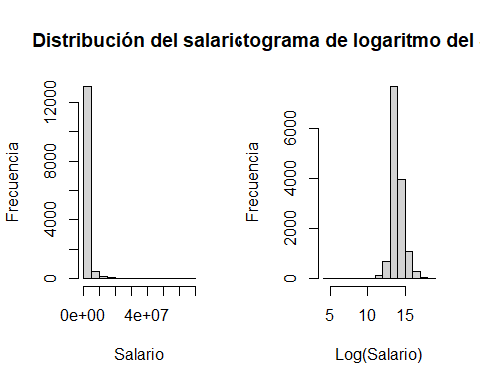
Se identifica que hay personas que reportaron ingresos iguales a cero a pesar de ser trabajadores y haber reportado horas laboradas, razón por la cual se excluyen estos registros de la muestra.

data\_h <-data\_h %>%  
filter(Salario2>0)  
  
table(is.na(data\_h$Salario2))

##   
## FALSE   
## 13786

Para visualizar la distribución individual de las variables con más precisión, se dividen en dos grupos tomando como referencia su naturaleza, es decir, si son cualitativas o cuantitativas. En primer lugar, se muestra una gráfica con la escala normal del salario y en base logarítmica. De esta forma, se observa una distribución más uniforme de la variable y la interpretación de los determinantes de naturaleza cuantitativa se hará desde las elasticidades.

data\_h<-data\_h %>% mutate(log\_salario=log(Salario2),  
 Edad2=Edad^2  
 )  
par(mfrow=c(1,2))   
hist(data\_h$Salario2, main="Distribución del salario", xlab="Salario", ylab="Frecuencia")  
hist(data\_h$log\_salario, main="Histograma de logaritmo del Salario", xlab="Log(Salario)", ylab="Frecuencia")



summary\_salario <- summary(data\_h$Salario)  
summary\_salario

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 97 839575 1002356 1741928 1625000 70000000 648

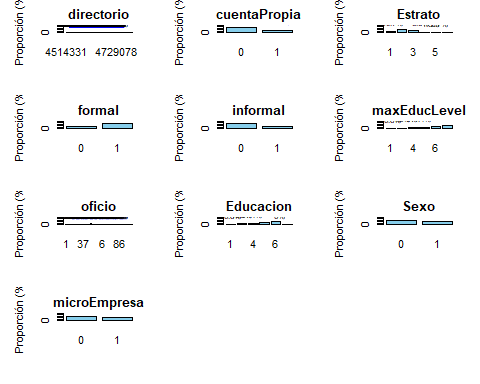
Para el año 2018, el salario mínimo en Colombia era de $781.242, por lo que la mayoría de las personas que hicieron parte de este estudio y que trabajan ganan al menos un salario mínimo. Históricamente, según el DANE, la tasa de informalidad en Bogotá ha estado ubicada alrededor del 40%. Comparando esta cifra con las personas que reportan el salario, asumiendo su representatividad, se puede ver que, en términos nominales, el salario del sector formal está muy cerca de la retribución en el mercado informal. Sin embargo, para este análisis se debe tener en cuenta que el salario mínimo solo es una parte de los costos laborales y que, a pesar de que estén cercanos los salarios con los beneficios de un salario integral, como la afiliación a EPS, pensión y cesantías, la brecha puede ser mayor. Para identificar el comportamiento de las variables que sirven para explicar el comportamiento del salario, se dividen en 2 categorías según su naturaleza de medición en cualitativas y cuantitativas. Los resultados más interesantes se ven en la variable del Estrato; en este caso, se observa cómo la mayoría de personas encuestadas se ubican en los estratos 2 y 3, con una representación del 42.9% y 37.2% respectivamente. La clasificación de las personas a través de los estratos tiene el propósito de identificar la capacidad adquisitiva. Por tanto, en esta encuesta predominan personas con ingresos medio-bajos siguiendo las definiciones del estrato socioeconómico. Las diferencias de representación por sexo no son significativas y siguen la regla general de las cifras poblacionales en Colombia.

Las cifras de las condiciones de las personas en el mercado laboral son consistentes, sobre todo en la condición de formalidad. Para el año 2018, el 63.3% reporta que se encuentra trabajando en el sector formal. La formalidad en este estudio se define como la persona que está afiliada a la seguridad social. Desde la teoría, se espera que las personas en el sector formal perciban salarios mayores a las personas en condiciones de informalidad.

qualitative\_factor\_vars <- data\_h %>%  
 select(where(function(x) is.factor(x) || is.character(x)))

## Adding missing grouping variables: `directorio`

plot\_proportion <- function(column, col\_name) {  
 proportion <- prop.table(table(column)) \* 100  
 barplot(proportion, main = col\_name, ylab = "Proporción (%)", col = "skyblue", ylim = c(0, 100))  
 text(x = 1:length(proportion), y = proportion,   
 label = paste0(round(proportion, 1), "%"), pos = 3, cex = 0.8, col = "black")  
}  
num\_vars <- ncol(qualitative\_factor\_vars)  
num\_rows <- ceiling(sqrt(num\_vars))  
num\_cols <- ceiling(num\_vars / num\_rows)  
par(mfrow = c(num\_rows, num\_cols), mar = c(5, 5, 2, 2))   
walk2(qualitative\_factor\_vars, names(qualitative\_factor\_vars), plot\_proportion)

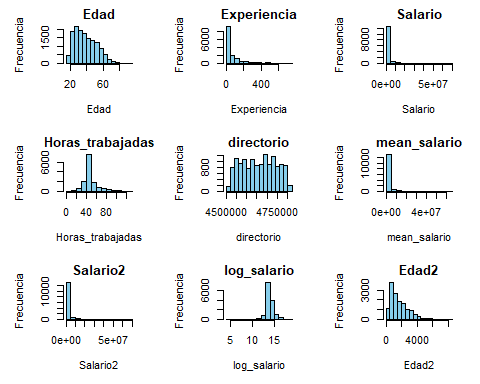


La distribución individual de las variables cuantitativas se muestra a continuación. El comportamiento de la edad es bastante uniforme y la mayor concentración se encuentra alrededor de los 37 años. Siguiendo la teoría del ciclo de vida, la mayoría de las personas encuestadas se encuentran en una edad adulta temprana. En el aspecto laboral, esto significa que ya se encuentran en un momento de estabilidad y donde son más productivas, y los salarios deberían ser más altos ya que se cuenta con más experiencia.

Este argumento se contrasta con los datos de experiencia en su trabajo, que se miden a partir de los meses que llevan trabajando en la misma empresa. Los datos muestran que, en promedio, las personas llevan trabajando en el mismo lugar 5 años. Sin embargo, la diferencia con la mediana es muy alta, ya que el valor medio de los meses trabajados es 2 años. Esta distorsión se puede explicar por la presencia de datos atípicos.

Por el lado de las horas trabajadas la semana pasada, sus valores están muy cercanos a las exigidas por la legislación laboral colombiana.

quantitative\_vars <- data\_h %>%  
 select(where(is.numeric))  
  
plot\_histogram <- function(column, col\_name) {  
 hist(column, main = col\_name, xlab = col\_name, ylab = "Frecuencia", col = "skyblue")  
}  
  
num\_vars <- ncol(quantitative\_vars)  
num\_rows <- ceiling(sqrt(num\_vars))  
num\_cols <- ceiling(num\_vars / num\_rows)  
  
par(mfrow = c(num\_rows, num\_cols), mar = c(5, 5, 2, 2))   
  
walk2(quantitative\_vars, names(quantitative\_vars), plot\_histogram)



## Análisis exploratorio bivariado

En este apartado se explorarán las diferencias que pueden existir entre las categorías analizadas en la sección anterior cuando se presentan a través del logaritmo del salario. Las diferencias más grandes se observan a través de la variable Estrato Socioeconómico y en el nivel de educación cuando el encuestado cuenta con grado de educación superior.

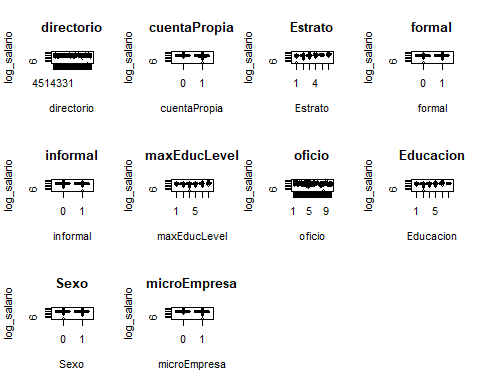
Las diferencias a través del estrato se intuyen por la naturaleza de esta variable, ya que mide, entre otras cosas, la capacidad adquisitiva del hogar. Mientras tanto, el tema del estudio se explica por la teoría del capital humano, que sostiene que la formación de los individuos son inversiones que aumentan su productividad e ingresos a lo largo del tiempo.

Se observan diferencias moderadas a través de variables como si el encuestado trabaja por cuenta propia o si está desempeñando su trabajo en condiciones de informalidad. Los datos muestran un ligero aumento del salario promedio si la persona trabaja como empleada. Además, los trabajadores que trabajan como empleados formales perciben unos ingresos ligeramente mayores que los trabajadores informales. No se observan diferencias en el caso del sexo, donde los salarios son uniformes para hombres y mujeres.

options(repr.plot.width = 20, repr.plot.height = 10)  
qualitative\_vars <- data\_h %>% select(where(is.factor) | where(is.character))

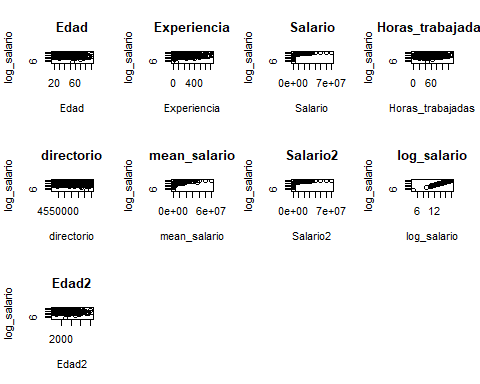
## Adding missing grouping variables: `directorio`

num\_cols <- 4  
num\_rows <- ceiling(ncol(qualitative\_vars) / num\_cols)  
par(mfrow = c(num\_rows, num\_cols))  
for (col in colnames(qualitative\_vars)) {  
 suppressWarnings(boxplot(log\_salario ~ data\_h[[col]], data = data\_h, main = col, xlab = col, ylab = "log\_salario"))  
}



En relación a las variables de naturaleza cuantitativa, no se observan patrones que ayuden a determinar cambios significativos en el ingreso laboral. Incluso, los niveles de correlación entre las variables no muestran señales de que puedan aportar cuando se inicie con el proceso de modelación.

options(repr.plot.width = 20, repr.plot.height = 10)  
  
# Seleccionar variables cuantitativas  
numeric\_vars <- data\_h %>% select(where(is.numeric))  
  
# Establecer el número de columnas para la disposición de la grilla  
num\_cols <- 4  
num\_rows <- ceiling(ncol(numeric\_vars) / num\_cols)  
  
# Establecer el tamaño de la gráfica  
par(mfrow = c(num\_rows, num\_cols))  
  
# Crear scatter plots para variables cuantitativas  
for (col in colnames(numeric\_vars)) {  
 suppressWarnings(plot(data\_h[[col]], data\_h$log\_salario, main = col, xlab = col, ylab = "log\_salario"))  
}



# 3. Análisis de regresion

EL propósito es estimar el siguiente modelo:

modelo1<-lm(log\_salario ~ Edad + Edad2,  
data=data\_h)  
stargazer::stargazer(modelo1,type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log\_salario   
## -----------------------------------------------  
## Edad 0.064\*\*\*   
## (0.003)   
##   
## Edad2 -0.001\*\*\*   
## (0.00004)   
##   
## Constant 12.786\*\*\*   
## (0.060)   
##   
## -----------------------------------------------  
## Observations 13,786   
## R2 0.032   
## Adjusted R2 0.032   
## Residual Std. Error 0.754 (df = 13783)   
## F Statistic 229.589\*\*\* (df = 2; 13783)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Los coeficientes de la regresión son significativos hasta a un nivel de confianza del 95%. A partir de ellos podemos rescatar la siguiente ecuación de salarios en función de la edad

Se evidencia que el ajuste del modelo dentro de la muestra no es muy grande ya que el R2 es de 0.032. Esta medida también la podemos complementar con el cálculo el error cuadrático medio MSE.

fit<-predict.lm(modelo1,newdata = data\_h)  
data\_h$log\_salario\_gorro<-fit  
MSE<-sum((data\_h$log\_salario-data\_h$log\_salario\_gorro)^2)/length(data\_h$log\_salario)  
MSE

## [1] 0.568164

Aumentando la complejidad del modelo es probable que tanto el R2 aumente como el MSE se reduzca, sin embargo, con esto estaremos aumentando la varianza de los estimadores, lo que permite disminuir el sesgo.

A partir de los coeficientes obtenidos es posible calcular la edad pico en el crecimiento de los ingresos de las personas.

Calculamos la edad pico, al tiempo que construimos la función que nos permitirá mediante el bootstrap estimar la varianza de los estimadores para la generación del intérvalo de confianza.

bfuncion<-function(data,index){  
   
modelo<-lm(log\_salario ~ Edad + Edad2, data=data\_h,subset = index)  
  
coeficientes<-modelo$coefficients  
  
b1<-coeficientes[2]  
b2<-coeficientes[3]  
  
edad\_pico<-b1/(-2\*b2)  
return(edad\_pico) #returns the second coefficient of the linear regression  
}  
  
edad\_pico<-bfuncion(data\_h,1:nrow(data\_h))  
modelo1$coefficients[2]/(-2\*modelo1$coefficients[3])

## Edad   
## 41.93591

edad\_pico

## Edad   
## 41.93591

Para poder realizar esta inferencia calculamos la Varianza utilizando bootstap.

set.seed(123)  
repeticiones<-1000  
boot(data\_h,bfuncion,R=repeticiones)

##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = data\_h, statistic = bfuncion, R = repeticiones)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 41.93591 -2.638615e-05 0.3462259

Con esto podemos establecer el intervalo de confianza para la edad pico

Es decir que para un nivel de confianza del 95%

n=nrow(x = data\_h)  
t=abs(qt(p = (1-0.95)/2,df =n-1 ))  
limite\_inferior=edad\_pico-(t\*0.3462259)  
limite\_superior=edad\_pico+(t\*0.3462259)  
limites<-c(limite\_inferior,limite\_superior)  
limites

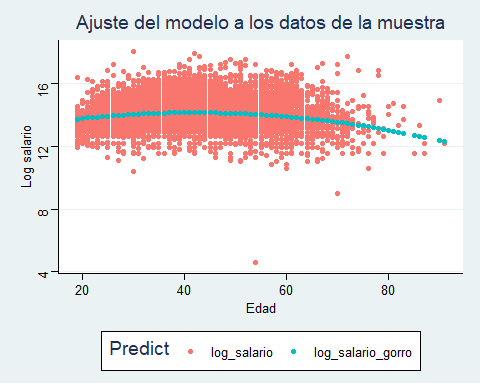
## Edad Edad   
## 41.25726 42.61456

Observamos de manera gráfica el ajuste del modelo a los datos.

data\_aux<-data\_h %>% select(log\_salario,log\_salario\_gorro,Edad)

## Adding missing grouping variables: `directorio`

data\_aux<-data\_aux %>% pivot\_longer(cols = log\_salario:log\_salario\_gorro,  
names\_to = "Predict",  
values\_to = "Valor")  
  
ggplot(data\_aux,aes(x=Edad,color=Predict)) +  
geom\_point(aes(y=Valor)) +  
labs(title = "Ajuste del modelo a los datos de la muestra",  
x="Edad",  
y="Log salario"  
) +  
theme\_stata() +  
scale\_fill\_stata()



# 4. Modelo de género

a) Estimación del modelo general.

modelS<-lm(log\_salario ~Sexo,data=data\_h)  
stargazer::stargazer(modelS,type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log\_salario   
## -----------------------------------------------  
## Sexo1 -0.096\*\*\*   
## (0.013)   
##   
## Constant 14.046\*\*\*   
## (0.009)   
##   
## -----------------------------------------------  
## Observations 13,786   
## R2 0.004   
## Adjusted R2 0.004   
## Residual Std. Error 0.765 (df = 13784)   
## F Statistic 53.854\*\*\* (df = 1; 13784)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

En el modelo general, al estimar el logaritmo natural del salario contra el sexo de la persona, se observa que las mujeres, en promedio, ganan menos que los hombres. La relación es estadísticamente significativa y el parámetro -0.068 indica que por ser mujer se espera que el logaritmo natural del salario disminuya 0.068 unidades en promedio. Esta relación es estadísticamente significativa y valida la hipótesis muy trabajada en el mercado laboral que muestra que hay diferencias significativas en las ganancias entre hombres y mujeres.

Aplicación del teorema FWL:

Para la aplicación del teorema de FWL se siguen dos pasos. En primer lugar, se estima un modelo donde la variable de interés, en este caso el género de la persona encuestada, se deja como variable independiente frente a las variables de control y se toman los residuales.

Luego se estima el segundo modelo donde la variable independiente es el logaritmo del salario frente a las variables de control y se toman los residuales. Por último, se estima la regresión de los dos residuales.

Se espera que el parámetro estimado en el primer modelo sea igual a la estimación del modelo en los residuales.

Para esto se incluyen las siguientes variables de control:

* Edad
* Oficio
* Si trabaja en microempresa
* Educación formal

El objetivo de esta estimación no es encontrar los determinantes del salario como se hizo en la primera parte, sino entender las diferencias en salario que se pueden presentar a través del sexo de las personas. Por esto se incluyeron estas variables que pueden controlar mejor el sesgo de solo incluir el sexo.

Paso 1: Estimar la ecuación del género de las personas contra las variables de control y tomar los residuales.

resid1 <- lm(Sexo ~Estrato+oficio+Edad+Edad2+informal+maxEducLevel+microEmpresa,data = data\_h)$residuals

Paso 2: Se estima el salario solo frente a las variables de control y se toman los residuales.

resid2 = lm(log\_salario ~Estrato+oficio+Edad+Edad2+informal+maxEducLevel+microEmpresa,data = data\_h)$residuals

general<-lm(log\_salario ~Estrato+oficio+Edad+Edad2+informal+maxEducLevel+microEmpresa+Sexo,data = data\_h)  
mod\_resid=lm(resid2~resid1)

stargazer::stargazer(general,mod\_resid,modelS,type = "text",digits= 3)

##   
## ====================================================================================================  
## Dependent variable:   
## --------------------------------------------------------------------------------  
## log\_salario resid2 log\_salario   
## (1) (2) (3)   
## ----------------------------------------------------------------------------------------------------  
## Estrato2 0.026\*   
## (0.016)   
##   
## Estrato3 0.137\*\*\*   
## (0.017)   
##   
## Estrato4 0.536\*\*\*   
## (0.025)   
##   
## Estrato5 0.763\*\*\*   
## (0.038)   
##   
## Estrato6 1.119\*\*\*   
## (0.034)   
##   
## oficio11 -0.267\*\*   
## (0.127)   
##   
## oficio12 -0.079   
## (0.127)   
##   
## oficio13 -0.388\*\*\*   
## (0.124)   
##   
## oficio14 -0.958\*\*\*   
## (0.267)   
##   
## oficio15 -0.329\*\*   
## (0.143)   
##   
## oficio16 -0.644\*\*\*   
## (0.131)   
##   
## oficio17 -0.599\*\*\*   
## (0.149)   
##   
## oficio18 -0.820\*\*\*   
## (0.149)   
##   
## oficio19 -0.458\*\*\*   
## (0.128)   
##   
## oficio2 -0.223\*   
## (0.126)   
##   
## oficio20 0.036   
## (0.544)   
##   
## oficio21 -0.176   
## (0.124)   
##   
## oficio3 -0.708\*\*\*   
## (0.124)   
##   
## oficio30 -0.310\*\*   
## (0.132)   
##   
## oficio31 0.063   
## (0.187)   
##   
## oficio32 -0.679\*\*\*   
## (0.132)   
##   
## oficio33 -0.693\*\*\*   
## (0.125)   
##   
## oficio34 -0.867\*\*\*   
## (0.142)   
##   
## oficio35 -0.760\*\*\*   
## (0.207)   
##   
## oficio36 -1.192\*\*\*   
## (0.151)   
##   
## oficio37 -0.910\*\*\*   
## (0.130)   
##   
## oficio38 -0.850\*\*\*   
## (0.127)   
##   
## oficio39 -0.711\*\*\*   
## (0.123)   
##   
## oficio4 0.517\*   
## (0.292)   
##   
## oficio40 -0.557\*\*\*   
## (0.146)   
##   
## oficio41 -0.648\*\*\*   
## (0.124)   
##   
## oficio42 -0.549\*\*\*   
## (0.142)   
##   
## oficio43 -0.201   
## (0.180)   
##   
## oficio44 -0.544\*\*\*   
## (0.127)   
##   
## oficio45 -0.790\*\*\*   
## (0.123)   
##   
## oficio49 -0.619\*\*   
## (0.267)   
##   
## oficio5 -0.161   
## (0.157)   
##   
## oficio50 -0.555\*\*\*   
## (0.156)   
##   
## oficio51 -0.555\*\*\*   
## (0.135)   
##   
## oficio52 -0.776\*\*\*   
## (0.235)   
##   
## oficio53 -0.804\*\*\*   
## (0.124)   
##   
## oficio54 -1.002\*\*\*   
## (0.125)   
##   
## oficio55 -0.899\*\*\*   
## (0.125)   
##   
## oficio56 -0.853\*\*\*   
## (0.151)   
##   
## oficio57 -0.792\*\*\*   
## (0.129)   
##   
## oficio58 -0.673\*\*\*   
## (0.124)   
##   
## oficio59 -0.775\*\*\*   
## (0.125)   
##   
## oficio6 -0.234\*   
## (0.131)   
##   
## oficio60 -2.516\*\*\*   
## (0.544)   
##   
## oficio61 -0.623\*\*\*   
## (0.180)   
##   
## oficio62 -0.809\*\*\*   
## (0.153)   
##   
## oficio63 -0.489   
## (0.395)   
##   
## oficio7 -0.416\*\*\*   
## (0.136)   
##   
## oficio70 -0.459\*\*\*   
## (0.141)   
##   
## oficio72 -0.717\*\*\*   
## (0.224)   
##   
## oficio73 -0.056   
## (0.395)   
##   
## oficio74 -0.708\*\*\*   
## (0.201)   
##   
## oficio75 -1.244\*\*\*   
## (0.168)   
##   
## oficio76 -0.698\*\*   
## (0.330)   
##   
## oficio77 -0.680\*\*\*   
## (0.129)   
##   
## oficio79 -0.836\*\*\*   
## (0.125)   
##   
## oficio8 -0.144   
## (0.129)   
##   
## oficio80 -0.967\*\*\*   
## (0.134)   
##   
## oficio81 -0.886\*\*\*   
## (0.135)   
##   
## oficio82 -1.136\*\*\*   
## (0.395)   
##   
## oficio83 -0.761\*\*\*   
## (0.133)   
##   
## oficio84 -0.720\*\*\*   
## (0.128)   
##   
## oficio85 -0.670\*\*\*   
## (0.129)   
##   
## oficio86 -0.528   
## (0.330)   
##   
## oficio87 -0.741\*\*\*   
## (0.130)   
##   
## oficio88 -0.851\*\*\*   
## (0.249)   
##   
## oficio89 -0.741\*\*\*   
## (0.170)   
##   
## oficio9 -0.093   
## (0.140)   
##   
## oficio90 -0.808\*\*\*   
## (0.141)   
##   
## oficio91 -0.809\*\*\*   
## (0.208)   
##   
## oficio92 -0.667\*\*\*   
## (0.145)   
##   
## oficio93 -0.776\*\*\*   
## (0.133)   
##   
## oficio94 -0.993\*\*\*   
## (0.163)   
##   
## oficio95 -0.671\*\*\*   
## (0.124)   
##   
## oficio96 -0.077   
## (0.395)   
##   
## oficio97 -0.870\*\*\*   
## (0.125)   
##   
## oficio98 -0.682\*\*\*   
## (0.124)   
##   
## oficio99 -0.859\*\*\*   
## (0.134)   
##   
## Edad 0.046\*\*\*   
## (0.002)   
##   
## Edad2 -0.001\*\*\*   
## (0.00003)   
##   
## informal1 -0.241\*\*\*   
## (0.013)   
##   
## maxEducLevel3 0.094   
## (0.063)   
##   
## maxEducLevel4 0.152\*\*   
## (0.061)   
##   
## maxEducLevel5 0.201\*\*\*   
## (0.061)   
##   
## maxEducLevel6 0.254\*\*\*   
## (0.061)   
##   
## maxEducLevel7 0.414\*\*\*   
## (0.061)   
##   
## microEmpresa1 -0.208\*\*\*   
## (0.013)   
##   
## Sexo1 -0.129\*\*\* -0.096\*\*\*   
## (0.011) (0.013)   
##   
## resid1 -0.129\*\*\*   
## (0.011)   
##   
## Constant 13.531\*\*\* -0.000 14.046\*\*\*   
## (0.144) (0.005) (0.009)   
##   
## ----------------------------------------------------------------------------------------------------  
## Observations 13,786 13,786 13,786   
## R2 0.524 0.010 0.004   
## Adjusted R2 0.521 0.010 0.004   
## Residual Std. Error 0.530 (df = 13692) 0.528 (df = 13784) 0.765 (df = 13784)   
## F Statistic 162.314\*\*\* (df = 93; 13692) 139.317\*\*\* (df = 1; 13784) 53.854\*\*\* (df = 1; 13784)  
## ====================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Luego de aplicar el teorema de FWL se observa cómo los coeficientes estimados son iguales, por lo que se valida utilizar las variables de control que se han venido trabajando.

A través de las tres especificaciones se ha comprobado que existen diferencias en el salario entre hombres y mujeres. Desde la estimación se observa que las mujeres, en promedio, ganan menos que los hombres en condiciones de empleo iguales. En el primer modelo general, cuando no se incluyen variables de control, se nota cómo en promedio las mujeres ganan 0.06 unidades menos que los hombres, y cuando se incluyen variables de control la diferencia es de 0.09 unidades en el logaritmo del salario.

# 5 Predicción

Las especificaciones que se van a trabajar para el proceso de predicción se van a construir tomando como referencia la ecuación de Mincer.

1. Tomando los elementos de la ecuacion de mince pero anadiendo las variables propuestas desde el punto de vista empirico
2. Modelo del punto 3 donde solo se tenia en cuenta la edad

Luego de estas estimaciones se estiman otros modelos derivados de los generales, aplicando alguna complejidad al modelo y poder estimar de mejor forma los determinantes del salario.

1. Particion de los datos

set.seed(10101)  
  
train\_size <- floor(0.7 \* nrow(data\_h))  
  
  
train\_indices <- sample(1:nrow(data\_h), size = train\_size, replace = FALSE)  
  
  
train <- data\_h[train\_indices, ]  
  
  
test <- data\_h %>%  
 filter(!row\_number() %in% train\_indices)  
  
  
niveles\_oficio <- unique(c(train$oficio, test$oficio))  
  
  
train$oficio <- factor(train$oficio, levels = niveles\_oficio)  
test$oficio <- factor(test$oficio, levels = niveles\_oficio)

1. Estimacion de los modelos

mod1 = lm(log\_salario~ maxEducLevel+Experiencia+Experiencia^2,data = train)  
mod2 = lm(log\_salario~ Edad+Edad^2+cuentaPropia+Estrato+informal+Sexo+Experiencia+Experiencia^2+microEmpresa+Horas\_trabajadas+oficio,data = train)  
mod3 = lm(log\_salario~ Edad+Edad^2,data = train)  
mod4 = lm(log\_salario~ Edad+Edad^2+Experiencia+Experiencia^2,data = train)  
mod5 =lm(log\_salario~ Edad+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas,data = train)  
mod6 =lm(log\_salario~ Edad+Experiencia+Horas\_trabajadas,data = train)  
mod7 =lm(log\_salario~ Edad+Edad^2+Experiencia+Horas\_trabajadas,data = train)  
mod8 =lm(log\_salario~ Edad+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas++cuentaPropia,data = train)  
mod9 =lm(log\_salario~ Edad+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas+informal,data = train)  
mod10 =lm(log\_salario~ Edad+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas+Sexo,data = train)  
  
mod11 =lm(log\_salario~ Edad\*maxEducLevel+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas+Sexo,data = train)  
  
mod12 =lm(log\_salario~ Edad+Edad^2+Experiencia\*maxEducLevel+Experiencia^2+Horas\_trabajadas+Sexo,data = train)  
  
mod13 =lm(log\_salario~ Edad\*maxEducLevel+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas+Sexo+oficio,data = train)  
  
mod14 =lm(log\_salario~ Edad+Edad^2+Experiencia\*maxEducLevel+Experiencia^2+Horas\_trabajadas+Sexo,data = train)  
  
mod15 =lm(log\_salario~ Edad+Edad^2+Experiencia\*maxEducLevel+oficio+Horas\_trabajadas+Sexo,data = train)  
  
mod16 = lm(log\_salario~ oficio,data = train)  
mod17 = lm(log\_salario~ oficio+Edad+Edad^2,data = train)  
mod18 = lm(log\_salario~ oficio+Edad+Edad^2+Sexo,data = train)

calculate\_metrics <- function(model, test\_data) {  
 metrics <- glance(model)  
 predictions <- predict(model, newdata = test\_data)  
 rmse <- sqrt(mean((predictions - test\_data$log\_salario)^2))  
 error\_train <- sqrt(mean(residuals(model)^2))  
 return(c(metrics$r.squared, rmse, error\_train))  
}  
  
models <- list(mod1, mod2, mod3,mod4,mod5,mod6,mod7,mod8,mod9,mod10,mod11,mod12,mod13,mod14,mod15,mod16,mod17,mod18)  
  
  
model\_names <- c("Modelo 1", "Modelo 2", "Modelo 3","Modelo 4","Modelo 5","Modelo 6","Modelo 7","Modelo 8","Modelo 9","Modelo 10","Modelo 11","Modelo 12","Modelo 13","Modelo 14","Modelo 15","Modelo 16","Modelo 17","Modelo 18")  
  
  
metrics\_list <- vector("list", length(models))  
  
  
for (i in seq\_along(models)) {  
 metrics\_list[[i]] <- calculate\_metrics(models[[i]], test)  
}  
  
  
metrics\_table <- tibble(  
 Modelo = model\_names,  
 R2 = sapply(metrics\_list, "[[", 1),  
 Error\_Test = sapply(metrics\_list, "[[", 2),  
 Error\_Train = sapply(metrics\_list, "[[", 3)  
)

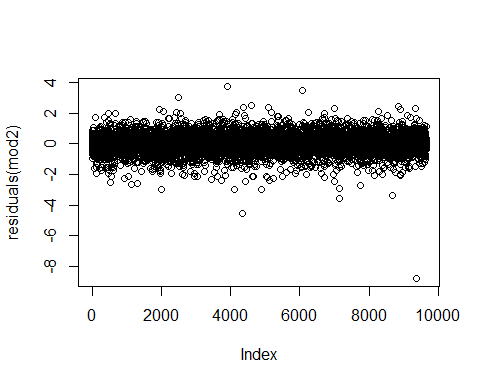
metrics\_table %>% arrange(Error\_Train)

## # A tibble: 18 × 4  
## Modelo R2 Error\_Test Error\_Train  
## <chr> <dbl> <dbl> <dbl>  
## 1 Modelo 2 0.518 NaN 0.537  
## 2 Modelo 13 0.422 NaN 0.587  
## 3 Modelo 15 0.417 NaN 0.590  
## 4 Modelo 18 0.356 NaN 0.620  
## 5 Modelo 17 0.353 NaN 0.622  
## 6 Modelo 16 0.351 NaN 0.622  
## 7 Modelo 11 0.284 NaN 0.654  
## 8 Modelo 12 0.275 NaN 0.658  
## 9 Modelo 14 0.275 NaN 0.658  
## 10 Modelo 1 0.244 NaN 0.672  
## 11 Modelo 9 0.205 NaN 0.689  
## 12 Modelo 8 0.0774 NaN 0.742  
## 13 Modelo 10 0.0257 NaN 0.763  
## 14 Modelo 5 0.0220 NaN 0.764  
## 15 Modelo 6 0.0220 NaN 0.764  
## 16 Modelo 7 0.0220 NaN 0.764  
## 17 Modelo 4 0.0218 NaN 0.764  
## 18 Modelo 3 0.000000462 NaN 0.773

1. Discusion de los resultados

El modelo con menor error en prediccion es el modelo mas general, el que se construyo a partir de las referencias teoricas y empiricas en los anteriores capitulos. La razon principal es porque tiene en cuenta todas estas variables y tambien porque captura muy bien todos los fenomenos analizados que pueden tener a la hora de explicar los salarios.

plot(residuals(mod2))



train$res = residuals(mod2)  
  
observaciones\_alejadas <- train %>%   
 arrange(desc(abs(res))) %>%  
 slice\_head(n = 10)  
  
observaciones\_alejadas

## # A tibble: 9,650 × 20  
## # Groups: directorio [6,522]  
## Edad cuentaPropia Estrato formal informal maxEducLevel oficio Educacion  
## <int> <chr> <chr> <chr> <chr> <chr> <fct> <chr>   
## 1 36 0 2 1 0 6 39 5   
## 2 51 0 2 1 0 7 85 6   
## 3 45 0 2 1 0 4 45 3   
## 4 61 0 2 0 1 3 53 3   
## 5 35 0 2 1 0 6 42 5   
## 6 56 0 2 1 0 4 55 3   
## 7 23 0 2 1 0 7 39 6   
## 8 30 0 2 1 0 6 53 5   
## 9 28 0 2 1 0 7 39 6   
## 10 50 1 2 0 1 7 45 6   
## # ℹ 9,640 more rows  
## # ℹ 12 more variables: Experiencia <int>, Salario <dbl>, Sexo <chr>,  
## # Horas\_trabajadas <int>, directorio <int>, microEmpresa <chr>,  
## # mean\_salario <dbl>, Salario2 <dbl>, log\_salario <dbl>, Edad2 <dbl>,  
## # log\_salario\_gorro <dbl>, res <dbl>

predicciones\_train <- predict(mod2, newdata = train)  
errores\_train <- train$log\_salario - predicciones\_train  
umbral\_train <- 1.5 \* IQR(errores\_train)   
valores\_atipicos\_train <- which(abs(errores\_train) > umbral\_train)  
  
print(train[valores\_atipicos\_train[1:5], ])

## # A tibble: 5 × 20  
## # Groups: directorio [5]  
## Edad cuentaPropia Estrato formal informal maxEducLevel oficio Educacion  
## <int> <chr> <chr> <chr> <chr> <chr> <fct> <chr>   
## 1 23 0 2 1 0 6 38 5   
## 2 47 1 3 1 0 6 41 5   
## 3 50 1 3 0 1 5 87 4   
## 4 47 0 4 1 0 7 19 6   
## 5 28 1 1 1 0 4 95 3   
## # ℹ 12 more variables: Experiencia <int>, Salario <dbl>, Sexo <chr>,  
## # Horas\_trabajadas <int>, directorio <int>, microEmpresa <chr>,  
## # mean\_salario <dbl>, Salario2 <dbl>, log\_salario <dbl>, Edad2 <dbl>,  
## # log\_salario\_gorro <dbl>, res <dbl>

Con este metodo identificamos las personas en los que el modelo se equivocp mas, para esto el modelo sobreestimo el salario , es decir que hizo la prediccion por encima de los esperado.

Cuando se analizan los 5 mas significativos se puede ver claramente que son personas que reportan que trabajan en condiciones precarias, el salario es muy bajo, son mujeres , trabajan en el secotr informal y su nivel de estudios es bajo.

Como al final el modelo de regresion, es un modelo sobre la media influyen mucho todos los demas, por lo que la sobreestimacion del salario que se da en este caso es normal. Es necesairo estimar con otro tipos de modelos mas flexibles que logren interpretar mucho mejor estos casos.

#ctrl <- trainControl(  
 #method = "LOOCV")  
  
#es3 <- log\_salario~ Edad+Edad^2+cuentaPropia+Estrato+informal+Sexo+Experiencia+Experiencia^2+microEmpresa+Horas\_trabajadas+oficio  
  
#odelo1c <- train(es3,  
 # data = data\_h,  
 # method = 'lm',   
 #trControl= ctrl)

Cuando se realiza la estimacion del error de test mediante LOOCV muestra un error de estimacion de 0.47 el cual es inferior cuando se hacer por medio de el primer metodo, es costoso en memoria pero da una clave para entender que el error de la prediccion puede mejorar.

#ctrl <- trainControl(  
#method = "LOOCV")  
  
#es13 <- log\_salario~ Edad\*maxEducLevel+Edad^2+Experiencia+Experiencia^2+Horas\_trabajadas+Sexo+oficio  
  
#odelo2c <- train(es13,  
 # data = data\_h,  
 # method = 'lm',   
 # trControl= ctrl)

# Bibliografía

Mincer, J (1974), Schooling, experience and earnings, national bureau of economic research, New York.