Catalina Esmeral Flórez

Ronny Johan Cruz

Juan Camilo Sánchez

Federico Ramírez

Link al repositorio GitHub: https://github.com/federicorr/Repositorio-Taller-1/blob/main/Taller%201.R

Problem Set I: Predicting income

1. Introducción

Este taller surge de la detección de una necesidad de construir un modelo de predicción de salarios de horarios individuales, donde citamos, tratamos, reunimos y procesamos las bases de datos en el DANE sobre la Gran Encuesta Integrada de Hogares - GEIH – 2018 que nos permitirá, analizar los comportamientos de varios grupos de habitantes en varias ciudades principales y municipios de Colombia, sin embargo es de aclarar que para este ejercicio y el modelo predictivo que se plantea hacer, los análisis de los datos serán enfocados en datos muéstrales tomados en la ciudad de Bogotá D.C.

$$w = f(X) + u$$

Donde w es el salario por hora, y (X) es una matriz que incluye posibles variables explicativas/predictores. En este conjunto de problemas, nos centraremos en $f(X) = X\beta$

La elección del universo de estudio ha sido propuesta por nuestro tutor principal de clase Santiago Barbieri, entendiendo la importancia de la información suministrada por el DANE y la aplicación a esta misma de un modelo predictivo que nos permita identificar posibles casos de fraude o un segundo objetivo específico, que sea poder identificar y ayudar a familias en posición de vulnerabilidad.

La conformación de la base de datos muestra personas mayores a 18 años que están trabajando actualmente, la variable utilizada para diferenciar a las personas en actividad es OCU, que nos ayuda a identificar a la gente que realiza una actividad no pagada o remunerada, esto nos ayuda a entender por qué tenemos ingresos con cero (0) como valor de ingreso

2. Datos

La Gran Encuesta Integrada de Hogares (GEIH)

La gran encuesta integrada de hogares es una metodología de recolección de datos, mediante la cual se solicita y suministra información sobre las condiciones de empleo de las personas (si estas actualmente trabajan, en qué trabajan, cuánto reciben por esta labor o ganan, si tienen seguridad social en salud o si están en un momento de busca de empleo), además de las características generales de la población como su sexo, la edad, su estado civil y su nivel educativo, se pregunta sobre sus fuentes de ingreso o ingresos. La GEIH proporciona al país información a todo nivel nacional, cabecera o municipio, regional, departamental, y para cada una de las capitales de los departamentos en el país.

Proceso de obtención de datos

El URL a tener como entrada de información y leer es https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html URL chunk 1, debemos tener en cuenta la carga de la página que pueda tomar unos minutos, aquí encontramos 10 particiones, de esta forma, dentro del código debemos usar una lógica para la unión de todos los archivos y crear una sola base de datos para poder realizar los modelos.

Proceso de limpieza de datos

Para la limpieza de datos tuvimos dos consideraciones importantes **1.** que el conjunto de problemas sea centrado únicamente en los trabajadores por mayores de dieciocho (18) años; y **2.** restringir los datos a estas categoría o edad de personas. Descrito lo anterior, utilizamos en el código dos variables > 18 AGE y OCU para tener individuos laborales activamente.

Análisis descriptivo

Las variables en observación como las horas de trabajo, el tipo de empleo, la cantidad de horas que demanda una jornada se indican que las personas objeto de estudio, pertenecen en su gran mayoría a un sector formal, donde trabajan como empleadores, esto facilita rastrear los ingresos y poder explicar las altas tasas de afiliación, en un régimen de salud (EPS) y de cotización de pensión o sistema general de pensiones, en primer lugar el oficio más frecuente sitúa a los vendedores ambulantes, domicilios , venta de loterías y periódicos, revistas o mercaderistas, esto mirando o agrupando los datos en el género hombre, si cambiamos al género mujeres, el trabajo más frecuente es el mismo, solo que para los hombres es de conductores de vehículos de transporte, taxistas o en su defecto conductores.

El análisis descriptivo se basó en la variable de ingreso y_ingLab_m_ha, esta corresponde al ingreso total por individuo incluyendo otros ingresos que puedan ser parte del salario y así tener un total. Esta variable queda definida como dependiente, dado que puede otorgar el mejor ajusta en el análisis sobre el modelo de predicción de ingresos. Esta variable de ingreso total incluye no solo el salario de la actividad principal, adiciona otro tipo de ingresos, como segunda actividad, subsidios otorgados por los entes correspondientes, rentabilidades o inversiones externas. Tener un modelo con la mayor precisión en el monto de tributación correspondiente que debe realizar cada por persona, es necesario que este tipo de ingresos sean adicionados.

Otra parte de la base de datos trae variables de ingresos que no son propiamente por una actividad laboral activa, que para el objetivo no son las indicadas a medir, analizar o tener como referencia principal.

Revisemos los análisis presentados a continuación:

Cantidad de personas por régimen contributivo en cada uno de los estratos sociales

>

	1	3	4	5	6	7
1	30	212	509	670	2816	4034
2	0	0	3	4	58	221
3	10	76	137	166	289	73

Cantidad de personas por edad en cada uno de los regímenes contributivos

>

1	2	3	
18	73	2	18
19	118	6	28
20	190	2	27
21	215	2	28
22	222	4	20
23	281	1	28
24	312	4	20
25	317	7	22
26	332	3	12
27	303	4	18
28	296	10	20
29	306	6	22
30	275	10	13
31	253	11	14
32	259	20	13
33	232	16	22
34	242	9	16
35	239	16	17
36	247	8	11
37	204	8	15
38	217	16	19
39	184	3	14
40	216	7	20
41	183	5	10
42	162	3	16
43	155	7	17
44	170	5	19
45	145	5	15
46	142	5	11
47	147	6	16
48	140	6	16
49	143	4	9
50	130	6	10
51	134	7	9
52	106	3	20
53	126	4	13
54	119	6	13
55	108	7	7
56	98	4	16
57	89	3	12
58	83	6	7
59	64	2	11
60	53	3	6
61	64	4	13
62	33	5	12
63	28	2	6
64	20	2	6
65	22	0	5
66	14	0	3
67	14	1	4
68	12	0	2
69	6	0	1
70	5	0	1
71	5	0	0
72	4	0	5
73	5	0	1
75	0	0	1
76	1	0	0
77	3	0	0
78	1	0	0
80	2	0	0
83	1	0	1
86	2	0	0

Cantidad de personas por sexo en cada uno de los regímenes contributivos

	0	1
1	4127	4145
2	123	163
3	435	316

Cantidad de personas por cotización en pensión en cada uno de los regímenes contributivos

>

	1	2	3
1	7234	947	91
2	252	19	15
3	23	728	0

Cantidad de personas en régimen contributivo por estrato social

>

1	3	4	5	6	7

```
  1   30  212  509  670 2816 4034
  2    0    0    3    4   58   221
  3   10   76  137  166 289   73
```

```
> 8822.229

> # Promedio de salario por cada uno de los géneros
>
> # por sexo
>
```

A tibble: 2 × 2

	<int>	<dbl>
1	0	8666.
2	1	8976.

```
> # Por edad – mostramos las primeras diez edades
```

	<int>	<dbl>
1	18	4617.
2	19	4428.
3	20	4470.
4	21	5213.
5	22	5202.
6	23	5691.
7	24	6258.
8	25	6681.
9	26	7406.
10	27	7560.

```
# … with 53 more rows
# ℹ Use `print(n = ...)` to see more rows


> # Por estrato social en orden
>
```

	<int>	<dbl>
1	1	4819.
2	2	5711.
3	3	8430.
4	4	22536.
5	5	27836.
6	6	43751.

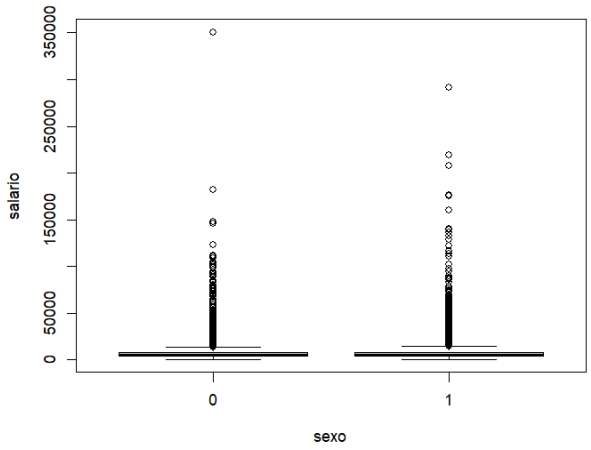
```
> #Varianzas de variables continuas
> df18 %>% var()
                             college maxEducLevel           age          age2      estrato1           sex regSalud
college                 2.261936e-01           NA    -0.6844334 -5.485623e+01   -0.10186736  0.012553626       NA
maxEducLevel                    NA           NA            NA            NA            NA           NA       NA
age                    -6.844334e-01           NA   144.5856361  1.152743e+04    1.65713451 -0.168329770       NA
age2                   -5.485623e+01           NA 11527.4260204  9.446573e+05  129.18628178 -10.735905447       NA
estrato1               -1.018674e-01           NA     1.6571345  1.291863e+02    0.95119147 -0.034499794       NA
sex                     1.255363e-02           NA    -0.1683298 -1.073591e+01   -0.03449979  0.250017825       NA
regSalud                        NA           NA            NA            NA            NA           NA       NA
cotPension              7.223540e-03           NA     0.3904178  5.245513e+01   -0.03729115 -0.008470329       NA
log_inglab_h           -9.749183e-02           NA     1.1645834  7.391560e+01    0.35528764  0.011161340       NA
y_ingLab_m_ha          -1.292986e+03           NA 21644.9597665  1.566157e+06 6036.58274414 77.498171237       NA
sizeFirm               -3.964186e-02           NA    -1.8007432 -1.706850e+02    0.11574588  0.063706757       NA
microEmpresa            1.038405e-02           NA     0.5191058  4.909600e+01   -0.03030569 -0.018514853       NA
oficio                  4.023701e+00           NA    24.0888002  2.110385e+03   -9.43421733  3.226368825       NA
hoursWorkActualSecondJob        NA           NA            NA            NA            NA           NA       NA
hoursWorkUsual          5.306330e-01           NA    -7.3926531 -8.116895e+02   -0.82840908  1.075264704       NA
informal                1.142902e-02           NA    -0.1739383 -2.091407e+00   -0.05492149 -0.007914043       NA
relab                  -1.404897e-02           NA     1.3524364  1.112389e+02    0.02204159 -0.051223439       NA
y_predicho             -5.299766e-03           NA     1.1645834  7.391560e+01    0.01551599 -0.003340818       NA
IC_alto                -5.802878e-03           NA     1.2666416  8.339259e+01    0.01655174 -0.003326982       NA
IC_bajo                -4.796653e-03           NA     1.0625253  6.443862e+01    0.01448024 -0.003354653       NA
                          cotPension  log_inglab_h y_ingLab_m_ha      sizeFirm  microEmpresa        oficio
college                 7.223540e-03  -0.09749183 -1.292986e+03 -3.964186e-02  1.038405e-02  4.023701e+00
maxEducLevel                      NA           NA            NA            NA            NA           NA
age                     3.904178e-01   1.16458343  2.164496e+04 -1.800743e+00  5.191058e-01  2.408880e+01
age2                    5.245513e+01  73.91560186  1.566157e+06 -1.706850e+02  4.909600e+01  2.110385e+03
estrato1               -3.729115e-02   0.35528764  6.036583e+03  1.157459e-01 -3.030569e-02 -9.434217e+00
sex                    -8.470329e-03   0.01116134  7.749817e+01  6.370676e-02 -1.851485e-02  3.226369e+00
regSalud                          NA           NA            NA            NA            NA           NA
cotPension              2.097096e-01  -0.11660551 -8.657989e+02 -3.455298e-01  1.023155e-01  2.147032e+00
log_inglab_h           -1.166055e-01   0.52943318  7.387253e+03  4.099228e-01 -1.112251e-01 -8.877834e+00
y_ingLab_m_ha          -8.657989e+02 7387.25252108  1.660531e+08  3.904761e+03 -9.948329e+02 -1.126706e+05
sizeFirm               -3.455298e-01   0.40992285  3.904761e+03  1.777658e+00 -4.908860e-01 -7.755382e+00
microEmpresa            1.023155e-01  -0.11122511 -9.948329e+02 -4.908860e-01  1.755172e-01  2.086351e+00
oficio                  2.147032e+00  -8.87783361 -1.126706e+05 -7.755382e+00  2.086351e+00  7.523431e+02
hoursWorkActualSecondJob          NA           NA            NA            NA            NA           NA
hoursWorkUsual         -5.859091e-01  -1.66805990 -1.970526e+04  9.017333e+01 -2.680536e-01  4.625183e+01
informal                1.713115e-01  -0.12716913 -1.060990e+03 -3.417356e-01  1.011809e-01  2.285828e+00
relab                   3.104895e-02  -0.01307449 -6.422374e+01 -2.600325e-01  7.543253e-02 -2.599896e-01
y_predicho             -1.262543e-02   0.02337979  2.923044e+02  5.545740e-03 -1.518916e-03  5.365145e-02
IC_alto                -1.103712e-02   0.02324741  3.009258e+02  2.714595e-03 -6.999037e-04  8.075014e-02
IC_bajo                -1.421374e-02   0.02351216  2.836831e+02  8.376885e-03 -2.337928e-03  2.655276e-02
```

```
                        hoursWorkActualSecondJob hoursWorkUsual      informal         relab    y_predicho       IC_alto
college                                       NA   5.306330e-01  1.142902e-02  -0.014048969  -0.005299766 -5.802878e-03
maxEducLevel                                  NA             NA            NA            NA            NA           NA
age                                           NA  -7.392653e+00 -1.739383e-01   1.352436422   1.164583426  1.266642e+00
age2                                          NA  -8.116895e+02 -2.091407e+00 111.238908960  73.915601859  8.339259e+01
estrato1                                      NA  -8.284091e-01 -5.492149e-02   0.022041588   0.015515989  1.655174e-02
sex                                           NA   1.075265e+00 -7.914043e-03  -0.051223439  -0.003340818 -3.326982e-03
regSalud                                      NA             NA            NA            NA            NA           NA
cotPension                                    NA  -5.859091e-01  1.713115e-01   0.031048947  -0.012625434 -1.103712e-02
log_inglab_h                                  NA  -1.668060e+00 -1.271691e-01  -0.013074487   0.023379785  2.324741e-02
y_ingLab_m_ha                                 NA  -1.970526e+04 -1.060990e+03 -64.223739335 292.304442102  3.009258e+02
sizeFirm                                      NA   9.017333e-01 -3.417356e-01  -0.260032494   0.005545740  2.714595e-03
microEmpresa                                  NA  -2.680536e-01  1.011809e-01   0.075432526  -0.001518916 -6.999037e-04
oficio                                        NA   4.625183e+01  2.285828e+00  -0.259989649   0.053651451  8.075014e-02
hoursWorkActualSecondJob                      NA             NA            NA            NA            NA           NA
hoursWorkUsual                                NA   1.476702e+02 -4.320526e-01  -0.888296545   0.104819055  8.763613e-02
informal                                      NA  -4.320526e-01  1.784677e-01   0.029400571  -0.010108386 -9.563024e-03
relab                                         NA  -8.882965e-01  2.940057e-02   0.260688328   0.008369959  9.492115e-03
y_predicho                                    NA   1.048191e-01 -1.010839e-02   0.008369959   0.023379785  2.321091e-02
IC_alto                                       NA   8.763613e-02 -9.563024e-03   0.009492115   0.023210907  2.320147e-02
IC_bajo                                       NA   1.220020e-01 -1.065375e-02   0.007247803   0.023548664  2.322034e-02
```

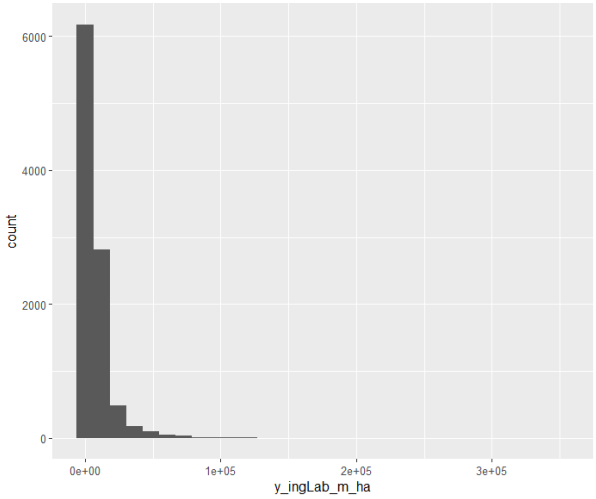
IC_bajo

```
college                    -0.004796653
maxEducLevel                NA
age                        1.062525271
age2                       64.438617152
estrato1                   0.014480238
sex                        -0.003354653
regSalud                   NA
cotPension                 -0.014213743
log_inglab_h               0.023512157
y_ingLab_m_ha              283.683075483
sizeFirm                   0.008376885
microEmpresa               -0.002337928
oficio                     0.026552761
hoursWorkActualSecondJob        NA
hoursWorkUsual             0.122001980
informal                   -0.010653749
relab                      0.007247803
y_predicho                 0.023548664
IC_alto                    0.023220345
IC_bajo                    0.023876982
> var(df18$y_ingLab_m_ha)
[1] 166053134
> var(df18$age)
[1] 144.5856
> var(df18$hoursWorkUsual)
[1] 147.6702
> var(df18$sex)
[1] 0.2500178

> # Diferencia de medias de wage entre edades <57> y sexo
>
        Welch Two Sample t-test

t = -1.1967, df = 9855.7, p-value = 0.2315
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -817.7140  197.7728
sample estimates:
mean in group 0 mean in group 1
      8666.398        8976.369
```

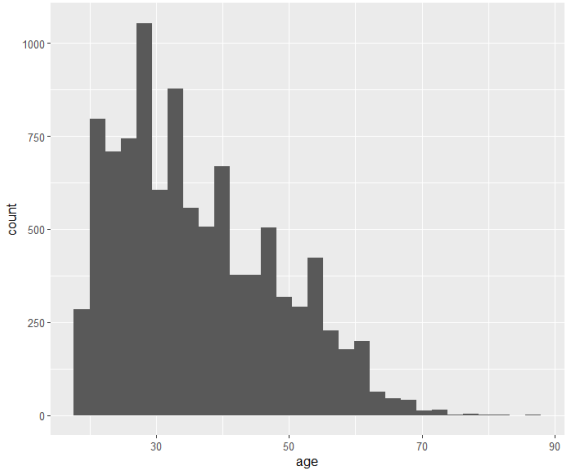
Grafico de Diferencia de Medias



Histograma de Salario



Histograma de Edad



										output_corr
1		college	maxEducLevel	age	age2	estrato1	sex	regSalud	cotPension	
2	college	1.00000000	NA	-0.11968188	-0.118672102	-0.21961443	0.05278901	NA	0.03316658	
3	maxEducLevel	NA	1	NA	NA	NA	NA	NA	NA	
4	age	-0.11968188	NA	1.00000000	0.986353561	0.14130620	-0.02799709	NA	0.07090190	
5	age2	-0.11867210	NA	0.98635356	1.000000000	0.13628413	-0.02209103	NA	0.11785326	
6	estrato1	-0.21961443	NA	0.14130620	0.136284126	1.00000000	-0.07074521	NA	-0.08349541	
7	sex	0.05278901	NA	-0.02799709	-0.022091034	-0.07074521	1.00000000	NA	-0.03699181	
8	regSalud	NA	NA	NA	NA	NA	NA	1	NA	
9	cotPension	0.03316658	NA	0.07090190	0.117853263	-0.08349541	-0.03699181	NA	1.00000000	
10	log_inglab_h	-0.28172310	NA	0.13310748	0.104518553	0.50065729	0.03067788	NA	-0.34994865	
11	y_ingLab_m_ha	-0.21097450	NA	0.13969172	0.125047417	0.48032329	0.01202770	NA	-0.14671832	
12	sizeFirm	-0.06251580	NA	-0.11232209	-0.131714666	0.08901179	0.09555996	NA	-0.56591635	
13	microEmpresa	0.05211548	NA	0.10304655	0.120572739	-0.07417034	-0.08838426	NA	0.53330090	
14	oficio	0.30844486	NA	0.07303728	0.079161947	-0.35266627	0.23524508	NA	0.17093121	
15	hoursWorkActualSecondJob	NA	NA	NA	NA	NA	NA	NA	NA	
16	hoursWorkUsual	0.09181373	NA	-0.05059311	-0.068723673	-0.06989795	0.17696342	NA	-0.10528698	
17	informal	0.05688387	NA	-0.03424150	-0.005093561	-0.13329949	-0.03746563	NA	0.88551836	
18	relab	-0.05785534	NA	0.22028934	0.224160339	0.04426374	-0.20064228	NA	0.13279353	
19	y_predicho	-0.07287796	NA	0.63341417	0.497368994	0.10404596	-0.04369654	NA	-0.18030878	
20	IC_alto	-0.08010240	NA	0.69156568	0.563290672	0.11141713	-0.04368247	NA	-0.15823007	
21	IC_bajo	-0.06526921	NA	0.57185638	0.429061297	0.09608420	-0.04341826	NA	-0.20086746	

		log_inglab_h	y_ingLab_m_ha	sizeFirm	microEmpresa	oficio	hoursWorkActualSecondJob
22							
23	college	-0.28172310	-0.210974495	-0.06251580	0.05211548	0.308444857	NA
24	maxEducLevel	NA	NA	NA	NA	NA	NA
25	age	0.13310748	0.139691721	-0.11232209	0.10304655	0.073037284	NA
26	age2	0.10451855	0.125047417	-0.13171467	0.12057274	0.079161947	NA
27	estrato1	0.50065729	0.480323290	0.08901179	-0.07417034	-0.352666270	NA
28	sex	0.03067788	0.012027696	0.09555996	-0.08838426	0.235245079	NA
29	regSalud	NA	NA	NA	NA	NA	NA
30	cotPension	-0.34994865	-0.146718318	-0.56591635	0.53330090	0.170931205	NA
31	log_inglab_h	1.00000000	0.787868665	0.42254454	-0.36486937	-0.444829360	NA
32	y_ingLab_m_ha	0.78786866	1.000000000	0.22727246	-0.18427513	-0.318771304	NA
33	sizeFirm	0.42254454	0.227272465	1.00000000	-0.87881352	-0.212066047	NA
34	microEmpresa	-0.36486937	-0.184275128	-0.87881352	1.00000000	0.181559704	NA
35	oficio	-0.44482936	-0.318771304	-0.21206605	0.18155970	1.000000000	NA
36	hoursWorkActualSecondJob	NA	NA	NA	NA	NA	1
37	hoursWorkUsual	-0.18865124	-0.125838149	0.05565545	-0.05265207	0.138763283	NA
38	informal	-0.41371023	-0.194898163	-0.60671712	0.57168750	0.197267559	NA
39	relab	-0.03519316	-0.009761381	-0.38198189	0.35264521	-0.018564672	NA
40	y_predicho	0.21014288	0.148351390	0.02720290	-0.02371121	0.012792436	NA
41	IC_alto	0.20975451	0.153312712	0.01336667	-0.01096783	0.019327585	NA
42	IC_bajo	0.20912078	0.142468932	0.04066013	-0.03611450	0.006264869	NA

		hoursWorkUsual	informal	relab	y_predicho	IC_alto	IC_bajo
43							
44	college	0.09181373	0.056883865	-0.057855341	-0.07287796	-0.08010240	-0.065269212
45	maxEducLevel	NA	NA	NA	NA	NA	NA
46	age	-0.05059311	-0.034241501	0.220289344	0.63341417	0.69156568	0.571856385
47	age2	-0.06872367	-0.005093561	0.224160339	0.49736899	0.56329067	0.429061297
48	estrato1	-0.06989795	-0.133299488	0.044263740	0.10404596	0.11141713	0.096084205
49	sex	0.17696342	-0.037465632	-0.200642278	-0.04369654	-0.04368247	-0.043418259
50	regSalud	NA	NA	NA	NA	NA	NA
51	cotPension	-0.10528698	0.885518363	0.132793530	-0.18030878	-0.15823007	-0.200867459
52	log_inglab_h	-0.18865124	-0.413710226	-0.035193161	0.21014288	0.20975451	0.209120776
53	y_ingLab_m_ha	-0.12583815	-0.194898163	-0.009761381	0.14835139	0.15331271	0.142468932
54	sizeFirm	0.05565545	-0.606717120	-0.381981894	0.02720290	0.01336667	0.040660131
55	microEmpresa	-0.05265207	0.571687499	0.352645210	-0.02371121	-0.01096783	-0.036114500
56	oficio	0.13876328	0.197267559	-0.018564672	0.01279244	0.01932758	0.006264869
57	hoursWorkActualSecondJob	NA	NA	NA	NA	NA	NA
58	hoursWorkUsual	1.00000000	-0.084160896	-0.143169587	0.05641227	0.04734554	0.064972676
59	informal	-0.08416090	1.000000000	0.136306044	-0.15648824	-0.14861328	-0.163204784
60	relab	-0.14316959	0.136306044	1.000000000	0.10721172	0.12205188	0.091866218
61	y_predicho	0.05641227	-0.156488242	0.107211723	1.00000000	0.99658446	0.996681255
62	IC_alto	0.04734554	-0.148613283	0.122051881	0.99658446	1.00000000	0.986554805
63	IC_bajo	0.06497268	-0.163204784	0.091866218	0.99668126	0.98655481	1.000000000

3. Age-wage profile

Para este ejercicio, se estimará la relación entre el salario y el ingreso. Para ello, se sigue el modelo de la forma:

$$log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

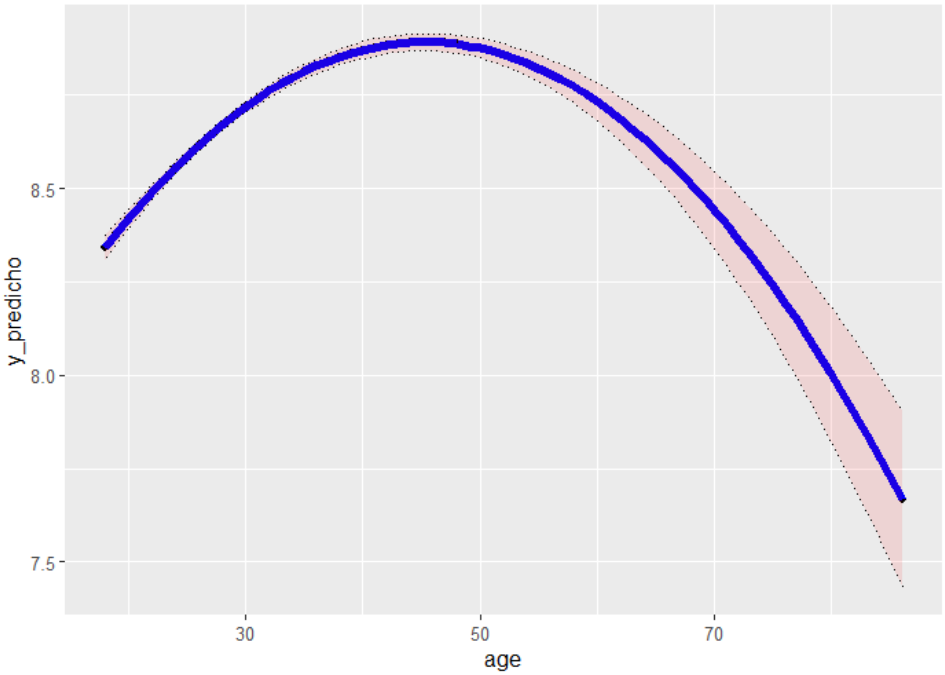
La variable escogida fue el logaritmo de los ingresos laborales por hora, dadas las instrucciones generales. A diferencia de las variables que consideraban ingresos totales, esta permitía dar una distinción certera entre ingreso y salario, considerando las actividades informales que son recogidas en la encuesta.

Regresión

	Dependent variable: log_inglab_ h
	Regresión 1
age	0.067***
	(0.004)
age2	-0.001***
	(0.00004)
Constant	7.374***
	(0.068)
Observations	9,892
R2	0.044
Adjusted R2	0.044
Residual Std. Error	0.711 (df = 9889)
F Statistic	228.437*** (df = 2; 9889)
Note:	*p<0.1; **p<0.05; ***p<0.01

Los resultados muestran que el salario por hora aumenta a medida que lo hace la edad (signo age), aunque este procedimiento sigue un comportamiento marginalmente decreciente (signo age2), alcanzando un punto máximo de salario por hora. La estimación muestra que un incremento de la edad, el salario por hora de la población ocupada incrementó en promedio un 6.7%, con un nivel de significancia del 0.01. Por otro lado, se observa que existe un efecto cuadrático que se recoge de la edad, teniendo en cuenta que su coeficiente es negativo y también significativo al 0.01, por lo que si la edad incrementa va decreciendo de manera que se tiene un punto máximo. Mediante los intervalos de confianza calculados por Bootstrap, se encuentra que el salario máximo por hora es de 8.89 y se alcanza a la edad de 45 años.

El R2 es de 0.044, lo que explica que el 4.4% del cambio en el salario promedio por hora se explica por el aumento marginalmente decreciente en la edad. Esto puede indicar problemas de sesgo por especificación al haber omitido variables posiblemente relevantes para el análisis. Sin embargo, en este modelo se observa que la edad es un elemento que debe hacer parte de análisis posteriores.



En la gráfica se puede observar que los intervalos de confianza se hacen más grandes conforme la edad se aleja del peakage de 45 años y el nivel de salario promedio por hora va descendiendo. Lo anterior puede estar asociado con los niveles educativos alcanzados a esa edad o los niveles de experiencia. Asímismo, el deceso en edades posteriores también estaría relacionado con el número de ocupados en esos rangos de edad, considerando las edades de retiro en Colombia.

4. The gender earnings GAP

Existe una brecha salarial entre mujeres y hombres, ya que en promedio las mujeres tienen menos ingresos laborales por hora que los hombres y es de esperarse que la edad de máximo salario (peak age) al igual que el salario máximo varíe en la misma dirección. Para comprobar este supuesto se estimaron dos modelos: el primero de ellos corresponde a la estimación de la semi-eslasticidad del salario con respecto al sexo y el segundo incluye las variables de la edad incluidas en el modelo anterior. El propósito de correr estos modelos es poder comparar sus especificaciones en cuánto a la predicción del ingreso máximo y su edad correspondiente diferenciada entre hombres y mujeres.

$$\log(\text{wage}) = \beta 1 + \beta 2 Female + v$$

$$\log(\text{wage}) = \beta 1 + \beta 2 Female + age + age\ 2 + e$$

$$\log(\text{wage}) = \beta 1 + \beta 2 Female + age + age\ 2 + age\_Female + \text{u}$$

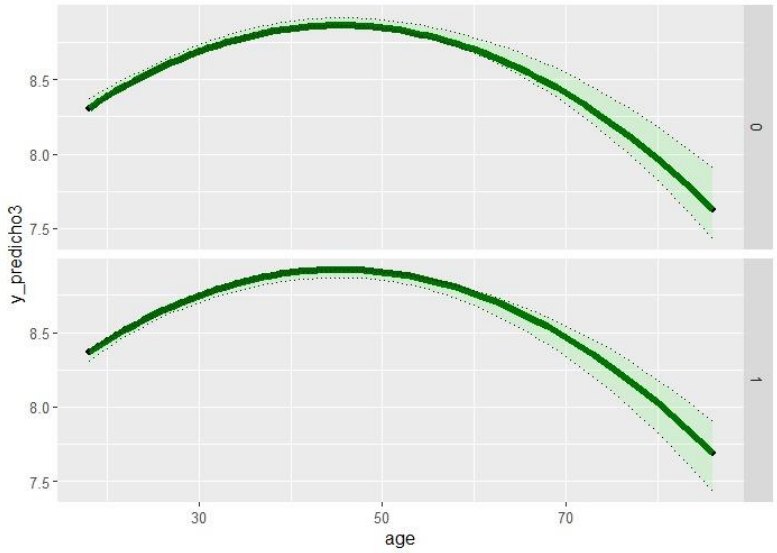
Los resultados de la estimación se encuentran a continuación:

==========

Dependent variable:

---

	log_inglab_h			NA
	(1)	(2)	(3)	(4)
sex	-0.219***	0.058***	0.045***	
	(0.045)	(0.014)	(0.015)	
age	0.064***	0.068***		0.067***
	(0.004)	(0.004)		(0.004)
age2	-0.001***	-0.001***		-0.001***
	(0.00004)	(0.00004)		(0.00004)
sex:age	0.008***			
	(0.001)			
Constant	7.474***	7.334***	8.702***	7.374***
	(0.072)	(0.069)	(0.010)	(0.068)
Observations	9,892	9,892	9,892	9,892
R2	0.050	0.046	0.001	0.044
Adjusted R2	0.049	0.045	0.001	0.044
Residual Std. Error	0.709 (df = 9887)	0.711 (df = 9888)	0.727 (df = 9890)	0.711 (df = 9889)
F Statistic	129.356*** (df = 4	158.028*** (df = 3	9.317*** (df = 1	228.437*** (df = 2
	9887)	9888)	9890)	98
				89
				)"
Note:		*p<0.1	**p<0.05	***p<0.01"



Esta brecha en los ingresos incondicionados, muestran que a pesar de que el poder predictivo de los modelos no es el mejor, hecho evidenciado en el R2 de todos los modelos en la tabla anterior, se observa como estas dos variables tienen un papel relevante en la estimación de los ingresos de los agentes. Sin embargo, con el propósito de limpiar aún más estas estimaciones y predecir de forma correcta el salario total, en la próxima sección se incluirán covariables de control.

A continuación, se estimaron los mismos modelos, pero incluyendo covariables.

	Dependent variable:	
	log_inglab_h	residuals
	(1)	(2)	(3)
sex	0.058***	0.124***	
	(0.014)	(0.010)	
maxEducLevel3		0.107	
		(0.079)	
maxEducLevel4		0.133*	
		(0.076)	

maxEducLevel5		0.158**
		(0.076)
maxEducLevel6		0.219***
		(0.075)
maxEducLevel7		0.581***
		(0.075)
age	0.068***	0.052***
	(0.004)	(0.003)
age2	-0.001***	-0.001***
	(0.00004)	(0.00003)
estrato12		0.028
		(0.017)
estrato13		0.137***
		(0.018)
estrato14		0.741***
		(0.026)
estrato15		0.969***
		(0.040)
estrato16		1.396***
		(0.036)
regSalud2		-0.119***

	(0.036)
regSalud3	-0.010
	(0.023)
cotPension2	-0.195**
	(0.099)
cotPension3	0.142***
	(0.050)
sizeFirm2	0.260***
	(0.052)
sizeFirm3	0.356***
	(0.055)
sizeFirm4	0.429***
	(0.054)
sizeFirm5	0.534***
	(0.054)
hoursWorkUsual	-0.012***
	(0.0004)
informal1	-0.061
	(0.101)
relab2	0.426***
	(0.027)

relab3		0.060	
		(0.051)	
relab8		0.176	
		(0.467)	
residuals			0.098***
			(0.011)
Constant	7.334***	7.247***	
	(0.069)	(0.105)	
---	---	---	---
Observations	9,892	9,308	9,308
R2	0.046	0.582	0.009
Adjusted R2	0.045	0.580	0.009
Residual Std. Error	0.711 (df = 9888)	0.466 (df = 9281)	0.427 (df = 9307)
F Statistic	158.028*** (df = 3	9281) 83.887*** (df = 1	9307)"
==========	==========	==========	==========
Note:		*p<0.1	***p<0.01"

5. Predicting earnings

Para lograr predecir el salario de una persona, según las variables observadas, tomamos nuestra muestra y la dividimos en dos, 70% en entrenamiento y 30% en testeo, de manera que no sobre ajustemos la muestro. De igual manera buscamos encontrar un modelo que logre predecir dicho salario de la mejor manera, por lo que realizamos la comparación del MSE (error cuadrático medio) entre modelos, explorando la no linealidad y la interacción que hay entre las variables, de manera que se logre llegar a la complejidad óptima.

Se escogió la metodología MSE dado que esta permite comparar y representar el balance entre sesgo y varianza en la estimación de un modelo y se permite proyectar la idea de la complejidad

del modelo. A pesar de la posibilidad de utilizar diferentes métricas existentes, se elige MSE dada la particularidad de los datos trabajados. El MSE puede presentar alta variabilidad, dado que utiliza mayor número de variables de entrenamiento y con menos observaciones.

Para la comparación realizamos 5 especificaciones, incluyendo las 2 realizadas anteriormente, esto con el propósito de encontrar y entender no linealidades e interacciones y percibir como la sobre especificación puede presentar igualmente problemas a la hora de encontrar el modelo con menor varianza. La especificación número 5 presento un error cuadrado de 0.01677, siendo esta la mejor predicción

```
Call:
lm(formula = y_ingLab_m_ha ~ age, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-13286  -4837  -2915   -202 340917

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3677.55     497.62    7.39 1.64e-13 ***
age           142.58      13.02   10.95  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13190 on 6976 degrees of freedom
  (10237 observations deleted due to missingness)
Multiple R-squared:  0.01691,   Adjusted R-squared:  0.01677
F-statistic:   120 on 1 and 6976 DF,  p-value: < 2.2e-16
```