Universidad de Los Andes
Maestría en Economía Aplicada
Big Data

26 de junio de 2022

Problem Set 1 – Big Data

Estudiantes: María Isabel Gómez, Paola Ortiz y Sofía Vanegas

El objetivo de este problem set es construir un modelo predictivo de la renta individual. Para ello, se realizó scrapping de los datos de la base GEIH 2018 que se encontraba en la página web de https://ignaciomsarmiento. github.io/GEIH2018 sample/.

A continuación, se presenta el link para dirigirse al repositorio de GITHUB:

- https://github.com/BigData-Gomez-Ortiz-Vanegas

1.1 General Instructions

1. Data acquisition

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

No. No hay ningún tipo de restricción a los datos. No existe un documento 'robots.txt', lo cual significa que hay libre acceso.

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

```
Inicializar contador en 1
Mientras contador sea menor o igual a 10
Crear un objeto llamado "datos_ignacio_[contador]" que contenga el código html de
la página correspondiente al contador

Poner el código html en formato tabla y guardarlo en un objeto llamado
"tabla_[contador]"

Agregar 1 al contador;

Inicializar contador en 2
Mientras contador sea menor o igual a 10
Crear un objeto llamado "tabla_final" que sea igual a "tabla_1"

Unir verticalmente "tabla_final" con "tabla_[contador]"

Agregar 1 al contador
```

2. Data Cleaning

Para lidiar con la cantidad de valores faltantes en la base, fue útil restringir la muestra al grupo de interés. Al excluir a las personas menores de 18 años, desocupadas y que vivían fuera de Bogotá, la cantidad de missing values se redujo considerablemente. En vista de ello, decidimos utilizar la base de datos en su estado original y no hacer ninguna imputación, pues en la discusión se concluyó que esa era la mejor manera de mantenernos en consonancia con los datos. Ahora bien, sí hicimos una modificación para el cuarto punto del problem set.

- ✓ Aclaración para el Punto 4: The earnings gap

En ánimos de mantener la transparencia del ejercicio, es importante mencionar que se hicieron modificaciones a la base de datos para ejecutar el cuarto punto del problem set. Dado que la ecuación del enunciado propone el logaritmo de la variable dependiente, la presencia de ceros en nuestra variable resultado (ingreso total observado) causaba un error en el cálculo. En vista de esto, hicimos el siguiente procedimiento para eliminar los ceros en la columna intentando mantener lo mejor posible la información de la variable:

1) Remplazar el ingreso total observado (ingtotob) por el ingreso total (ingtot) cuando ingtotob = 0 y ingtot≠ 0
2) Remplazar los que definitivamente quedan en 0 por NAs (alrededor de 200 observaciones)

Se debe reconocer que este procedimiento puede causar ruido en la variable, sin embargo, se procuró capturar la mayor información y mantenernos fieles a ella en la medida de lo posible.

A continuación, presentamos la justificación de la selección de las variables y las estadísticas descriptivas:

Selección de la variable dependiente: ingreso		
Variable	**Descripción**	**Justificación**
• Ingtotob	Ingreso total observado	Seleccionamos esta variable por que como su nombre lo indica, es el ingreso total observable de un individuo. Esta variable es empleada en diversas investigaciones socioeconómicas económicas para llevar a cabo la medición de pobreza monetaria y niveles de desigualdad, dado que brinda un mayor nivel de precisión de los niveles de ingresos percibidos, ha sido empleada por el DANE y por la CEPAL para la medición del índice de Gini. Sin embargo, cuando nos encontramos con un valor de 0 en esta categoría, tomamos la variable de ingtot para reemplazar esos datos faltantes.

Selección de las variables independientes		
Variable	**Descripción**	**Justificación**
• Age	Edad	En estudios socioeconómicos sobre ingresos, la edad es un factor importante dado que permite establecer tres periodos a lo largo de la vida: educación, trabajo y retiro, siendo los individuos productivos durante el periodo de trabajo, el cual normalmente presentan una forma de U invertida. Esto nos permite entender como pueden ser los ingresos en la etapa laboral.
• Sex	Sexo, mujer -hombre	El sexo constituye un determinante clave en el presupuesto de los hogares e ingreso de un individuo. La literatura ha demostrado una tendencia sobre las mujeres de percibir un menor salario en comparación con los hombres, incluso desempeñando las mismas funciones. Existe discriminación en el marcado laboral, dada las responsabilidades

		domésticas que socialmente tienen a cargo. En Colombia, la brecha salarial general entre hombres y mujeres, según la media, es de 12,9% para el año 2019, según la Gran Encuesta Integrada de Hogares (GEIH).
• Estrato		La clasificación en cualquiera de los seis estratos es una aproximación a la diferencia socioeconómica jerarquizada, léase pobreza a riqueza (teniendo en cuenta los niveles de ingresos de los individuos).
• MaxEducLevel	Máximo nivel educativo alcanzado	Bien es conocida la literatura que demuestra el impacto del nivel educativo sobre los ingresos. Urrutia -Sandoval concluye que "probablemente el instrumento de política más eficiente para contrarrestar los efectos sobre la distribución de los ingresos y la riqueza es la educación". Arias, (2019), en su investigación encontró que el mayor nivel de educación tiene efecto positivo en la reducción de la pobreza monetaria, a través del mecanismo de productividad y mejora de ingresos
• HoursWorkUsual	Horas de trabajo semanal	El tiempo que las personas dedican directamente a actividades productivas influye en sus niveles de ingresos. La literatura ha demostrado mayoritariamente una relación positiva entre el número de horas de trabajo y los ingresos percibidos.
• CotPension	Cotiza a pensión	Permite comprender la cobertura pensional, dado que las personas ocupadas en el país tienen mayor probabilidad de tener un nivel de ingresos que les permita estar afiliados a estos fondos ya sea empleado formal o independiente. La población con poco o muy bajos ingresos no cotiza a ningún fondo de pensiones.
• Formal	Vinculación laboral formal o informal	Posibilita la comprensión de las condiciones laborales de los individuos. Las personas vinculadas formalmente al entorno laboral perciben un ingreso mensual, por el contrario, la informalidad afecta a las personas más vulnerables, generando una diferencia salarial que impacta negativamente a estos últimos.
• Antigüedad_indu	Antigüedad laborando	El nivel de antigüedad laboral está relacionado directamente con los años de experiencia que pueda tener un individuo durante su etapa productiva, que lo ubica en una etapa dentro del mercado laboral, ya sea junior, senior, entre otros y está directamente relacionado con su nivel de ingresos.
• cuentaPropia	Si es independiente o empleado	El nivel de ingresos de los individuos puede variar dependiendo de si este es independiente o se encuentra empleado. Dichos ingresos difieren dependiendo del tipo de contrato, por ejemplo, por prestación de servicios o si tiene un contrato laboral a

		término fijo o indefinido. Según un artículo publicado en el tiempo, un trabajador independiente necesita ganar 60% más que un asalariado para tener el mismo nivel de ingresos.
• oficio	Ocupación	El resultado del proceso de fijación salarial se encuentra relacionado con el oficio o la ocupación de la persona. Por ejemplo, investigaciones han demostrado que el sector tecnología ofrece salarios comparativamente superiores a otros sectores.

A continuación, presentamos una tabla elaborada por el DANE sobre la brecha salarial de género para la población ocupada con ingresos laborales en Colombia para el total nacional de 2019, las cual nos permite entender el comportamiento de los niveles de ingresos y la justificación de algunas de las variables que seleccionamos para la estimación de los modelos.

Desagregaciones	Ingreso laboral promedio[a]			Ingreso laboral promedio por hora[a]			Número de personas (cifras en miles y %)			
	Hombres (miles)	Mujeres (miles)	Brecha (%) (H-M)	Hombres (miles)	Mujeres (miles)	brecha (%) (H-M)	Hombres (miles)	Mujeres (miles)	Total (miles)	Porcentaje de mujeres (%)
Total[b]	1.230	1.072	12,9	6,2	6,3	-2,3	12.757	8.696	21.453	40,5%
A) Dominio geográfico										
Urbano (cabeceras)	1.416	1.171	17,3	7,1	6,8	4,0	9.586	7.520	17.106	44,0%
Rural (centros poblados y rural disperso)	665	435	34,5	3,4	3,2	7,2	3.171	1.176	4.347	27,1%
B) Grupos etarios										
15 a 24 años	769	727	5,4	4,1	4,4	-6,1	1.870	1.213	3.082	39,3%
25 a 34 años	1.266	1.149	9,2	6,2	6,4	-3,5	3.275	2.334	5.609	41,6%
35 a 44 años	1.466	1.246	15,0	7,1	7,1	-0,3	2.846	2.109	4.955	42,6%
45 a 54 años	1.360	1.150	15,4	6,7	6,7	-0,5	2.442	1.656	4.098	40,4%
55 o más años	1.133	888	21,6	6,2	6,3	-0,4	2.298	1.375	3.673	37,4%

Nota: tabla elaborada por el DANE 2020. Disponible en https://www.dane.gov.co/files/investigaciones/notas-estadisticas/nov-2020-brecha-salarial-de-genero-colombia.pdf

Tabla 1. Estadísticas descriptivas

	18-30 (N=5273)	30-50 (N=7397)	>50 (N=3872)	Total (N=16542)
age				
Mean (SD)	24.8 (3.41)	39.8 (5.76)	58.6 (6.35)	39.4 (13.5)
Median [Min, Max]	25.0 [18.0, 30.0]	39.0 [31.0, 50.0]	57.0 [51.0, 94.0]	38.0 [18.0, 94.0]
factor(sex)				

0	2416 (45.8%)	3629 (49.1%)	1730 (44.7%)	7775 (47.0%)
1	2857 (54.2%)	3768 (50.9%)	2142 (55.3%)	8767 (53.0%)
ingtot				
Mean (SD)	1290000 (1250000)	1980000 (2800000)	2030000 (3600000)	1770000 (2680000)
Median [Min, Max]	1000000 [0, 30000000]	1160000 [0, 85800000]	1000000 [0, 70000000]	1050000 [0, 85800000]
estrato				
1	681 (12.9%)	767 (10.4%)	325 (8.4%)	1773 (10.7%)
2	2379 (45.1%)	3140 (42.4%)	1397 (36.1%)	6916 (41.8%)
3	1830 (34.7%)	2615 (35.4%)	1537 (39.7%)	5982 (36.2%)
4	254 (4.8%)	525 (7.1%)	347 (9.0%)	1126 (6.8%)
5	57 (1.1%)	143 (1.9%)	120 (3.1%)	320 (1.9%)
6	72 (1.4%)	207 (2.8%)	146 (3.8%)	425 (2.6%)
maxEducLevel				
1	11 (0.2%)	39 (0.5%)	74 (1.9%)	124 (0.7%)
3	50 (0.9%)	292 (3.9%)	427 (11.0%)	769 (4.6%)
4	108 (2.0%)	677 (9.2%)	725 (18.7%)	1510 (9.1%)
5	406 (7.7%)	826 (11.2%)	663 (17.1%)	1895 (11.5%)
6	2024 (38.4%)	2395 (32.4%)	865 (22.3%)	5284 (31.9%)
7	2674 (50.7%)	3167 (42.8%)	1118 (28.9%)	6959 (42.1%)
hoursWorkUsual				
Mean (SD)	46.8 (14.7)	48.6 (14.7)	44.3 (17.6)	47.0 (15.5)
Median [Min, Max]	48.0 [1.00, 119]	48.0 [1.00, 130]	48.0 [1.00, 126]	48.0 [1.00, 130]
cotPension				
1 = cotiza	3251 (61.7%)	4684 (63.3%)	1465 (37.8%)	9400 (56.8%)
2 = no cotiza	2022 (38.3%)	2690 (36.4%)	2050 (52.9%)	6762 (40.9%)
3 = pensionado	0 (0%)	23 (0.3%)	357 (9.2%)	380 (2.3%)
formal				
0	2033 (38.6%)	2720 (36.8%)	2082 (53.8%)	6835 (41.3%)
1	3240 (61.4%)	4677 (63.2%)	1790 (46.2%)	9707 (58.7%)
antiguedad_indu				
Mean (SD)	19.7 (23.5)	60.4 (69.5)	130 (131)	63.8 (89.5)
Median [Min, Max]	12.0 [0, 192]	36.0 [0, 480]	84.0 [0, 720]	24.0 [0, 720]

3. Age-earnings profile

a) 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.

En un inicio, la variable que mejor representa el valor de los ingresos totales de los trabajadores parecía ser ingreso total (*ingtot*). Esto tiene sentido pues, como bien dice su nombre, la variable muestra el ingreso total de cada individuo en la base de datos. Sin embargo, existe una variable adicional que muestra el ingreso total observado (*ingtotob*). Elegimos esta última variable pues consideramos relevante analizar la cantidad de ingreso total que el individuo efectivamente percibe, más allá del ingreso total en general. Así, la variable que elegimos para representar el ingreso es el ingreso total observado (*ingtotob*).

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

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

Tabla 2. Estimación mínimos cuadrados ordinarios - Ingreso total observado contra edad

	Estimates	CI	P-value
(Intercept)	-510304.70	-843847.00 – -176762.41	**0.003**
age	93605.59	76986.35 – 110224.84	**<0.001**
age2	-923.35	-1115.70 – -730.99	**<0.001**
Observations	16542		
R^2 / R^2 adjusted	0.012 / 0.011		

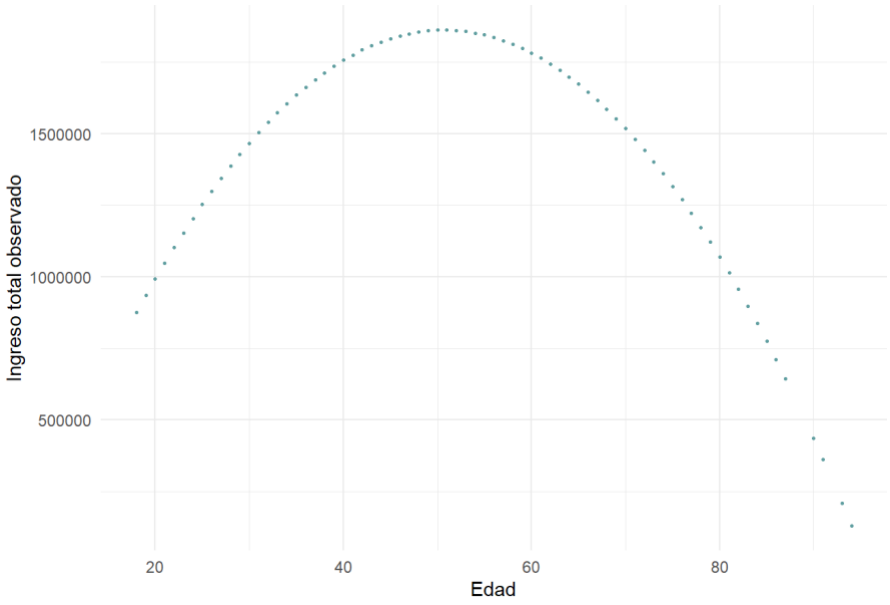
c) How good is this model in sample fit?

El R^2 y el R^2 ajustado son indicadores de la porción de la variación de la variable dependiente que se explica por el modelo. En este caso, el R^2 y el R^2 ajustado son una muestra de qué tanto se explica la variación del ingreso total observado por las variables independientes de manera colectiva. El valor de estos indicadores va de 0% hasta 100%, donde 100% quiere decir que la variable dependiente se explica totalmente por el modelo. En este ejercicio, los valores del R^2 y el R^2 ajustado son 1.2% y 1.1%, respectivamente. Esto quiere decir que el modelo explica muy poco la variación del ingreso total observado.

d) Plot the predicted age-earnings profile implied by the above equation. What is the "peak age" suggested by the above equation?

Al graficar el ingreso total observado predicho contra la edad, se obtiene el siguiente gráfico:

Gráfico 1. Valores predichos del ingreso total observado contra la edad



Esta gráfica coincide considerablemente con lo sugerido en la instrucción del problem set: se puede ver un crecimiento hasta la edad de los 50 años y, en los años siguientes, se aprecia una disminución.

e) Use bootstrap to calculate the standard errors and construct the confidence intervals.

Tabla 3. OLS vs. Bootstrap - EE

	OLS	Std.Error	Bootstrap	Std.Error
(Intercept)	-510.305	170.165	-433.220	
age	93.606	8.479	90.903	17.998
age2	-923	98	-796	208

Tabla 4. OLS vs. Bootstrap - CI

	OLS		Bootstrap	
	2.5 %	**97.5 %**	**2.5 %**	**97.5 %**
(Intercept)	-843.847	-176.762	-1.235.988	551.953
age	76.986	110.225	34.703	136.853
age2	-1.116	-731	-1.387	-765

4. The earnings GAP

a) Estimate the unconditional earnings gap

$$\log(Income) = \beta_1 + \beta_2 Female + u$$

Tabla 5. Estimación mínimos cuadrados ordinarios – logaritmo del ingreso total observado contra el sexo

	log wage		
Predictors	Estimates	CI	p
(Intercept)	13.840	13.82 – 13.86	**<0.001**
sex	0.198	0.17 – 0.23	**<0.001**
Observations	16276		
R^2 / R^2 adjusted	0.12 0.012		

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

El coeficiente es significativo con 99% de confianza y refleja que, en promedio, ser hombre está asociado con un incremento de 19.8% en el ingreso observado con respecto a las mujeres. Sin embargo, es importante considerar que el modelo solo cuenta con una variable independiente. Por tanto, es probable que el modelo esté sub-especificado, que sea exógeno y que el estimador sea sesgado. Ahora bien, el R^2 y el R^2 ajustado y es igual a 1.2%. Esto indica que el modelo explica el 1.2% de la variación del logaritmo del ingreso total observado. Así, parece que el modelo no es un muy bien *fit* para la muestra.

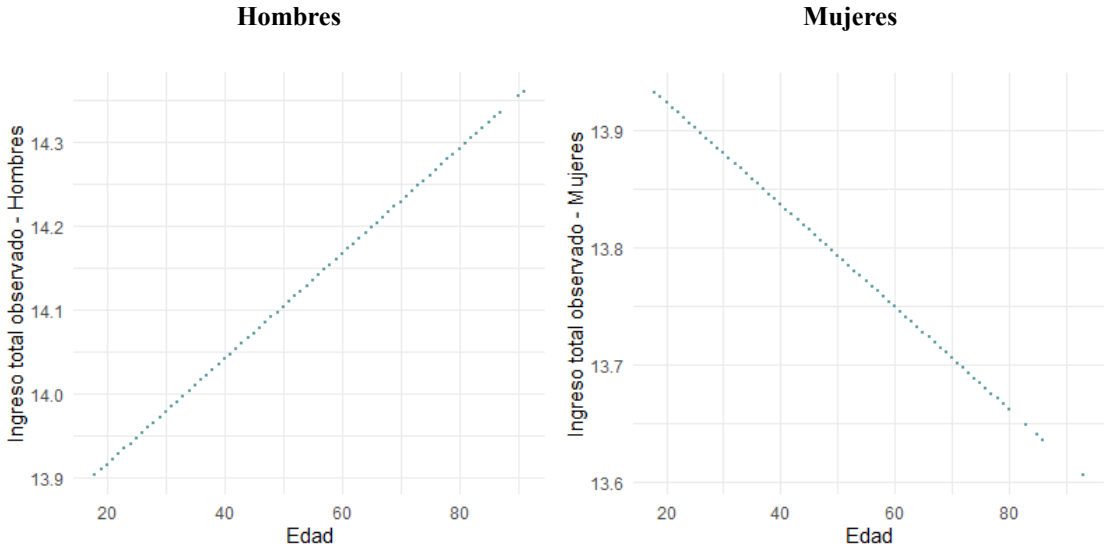
c) Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes? What is the implied "peak age" by gender?

Para empezar, se graficó una relación lineal del logaritmo del ingreso total observado contra la edad. Se puede ver que el intercepto en el eje y es mayor para los hombres y que su pendiente es negativa, mientras que la pendiente de la gráfica de las mujeres es positiva. Esto indica que, en promedio, los

hombres perciben ingresos más altos que las mujeres y que los hombres probablemente aumentan su salario a medida que crecen, mientras que las mujeres disminuyen su ingreso total observado.

$$\log(Income_i) = \beta_1 + \beta_2 age_i + u_i,\ where\ i = \{male, female\}$$

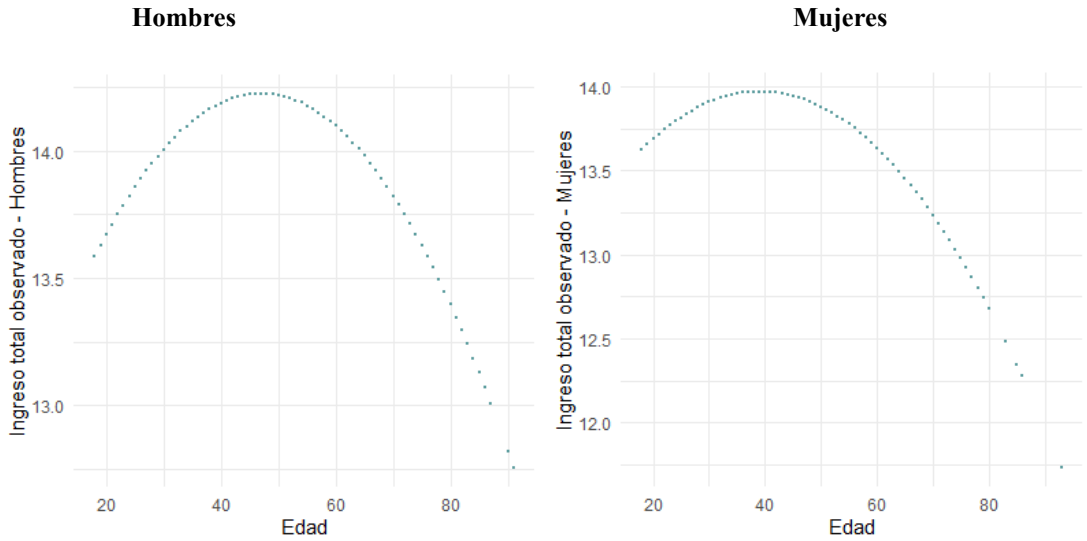
Gráfico 2. Relación lineal del logaritmo del ingreso total observado contra la edad por sexo



Adicionalmente, se graficó la relación cuadrática entre el logaritmo del ingreso total observado y la edad. Para los hombres, el pico del ingreso se ve aproximadamente en los 49 años, mientras que esta cifra para las mujeres está alrededor de los 40 años. Ahora bien, en este caso el intercepto con el eje y parece relativamente similar. Sin embargo, la tendencia de caída de la variable resultado (en promedio) comienza más pronto para las mujeres y se extiende por una mayor cantidad de años.

$$\log(Income_i) = \beta 1 + \beta_2 age_i\ + \beta_3 age_i^2 + u_i,\ where\ i = \{male, female\}$$

Gráfico 3. Relación cuadrática del logaritmo del ingreso total observado contra la edad por sexo



d) Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?

Tabla 6. Bootstrap: errores estándar e intervalos de confianza por sexo

Hombres

	Coef	SE	Bootstrap	SE
(Intercept)	-663.373	254.254	-534.250	
age	98.573	12.606	91.840	24.080
age2	-882	145	-673	276

	OLS		Bootstrap	
	2.5 %	**97.5 %**	**2.5 %**	**97.5 %**
(Intercept)	-1.161.769	-164.976	-1.660.343	875.867
age	73.863	123.283	9.237	157.346
age2	-1.166	-599	-1.492	329

Mujeres

	Coef	SE	Bootstrap	SE
(Intercept)	-567.674	216.827	-623.862	
age	101.698	10.890	107.928	20.922
age2	-1.152	128	-1.173	245

	OLS		Bootstrap	
	2.5 %	**97.5 %**	**2.5 %**	**97.5 %**
(Intercept)	-992.713	-142.635	-1.268.454	94.456
age	80.351	123.044	68.525	144.430
age2	-1.402	-902	-1.623	-697

Los intervalos de confianza del bootstrap (al igual que los de OLS) se sobreponen. En el caso del intervalo de confianza del error estándar del coeficiente de "edad", este resulta ser mucho más ancho para los hombres que para las mujeres y el de los hombres contiene completamente al de las mujeres. En el caso de edad al cuadrado no se sobreponen completamente pero sí en parte. Esto indica que la diferencia entre géneros no es estadísticamente significativa. No tenemos suficiente evidencia estadística para afirmar que el efecto de la edad sobre el ingreso es diferente entre hombres y mujeres.

e) ***Equal Pay for Equal Work?***

i. Estimate the conditional earnings gap.

Como controles que consideramos relevantes en la ecuación se incluyó la edad, el estrato, el máximo nivel de educación, las horas semanales trabajadas usualmente, el estado de formalidad, si está auto-empleado o no, el oficio de la persona, el tamaño de la firma que lo contrata y su antigüedad en ella. En la tabla a continuación no se reportan los coeficientes del control de oficio para hacer eficiente el uso del espacio, pero la tabla completa se puede ver al final del documento (Anexo 1). Se debe destacar que este modelo arroja un R^2 y el R^2 ajustado de 51.2% y 51.5%, respectivamente. Lo que quiere decir que más de la mitad de la variación del logaritmo del ingreso total observado se explica por el modelo. Este parece ser un *fit* considerablemente mayor al de otros modelos.

$$\log(Income) = \beta_1 + \beta_2 sex + \theta X + u$$

Tabla 7. Regresión logaritmo del ingreso total observado contra el sexo con controles

		log wage	
Predictors	**Coef**	**CI**	**p**
(Intercept)	12.64	12.35 – 12.94	**<0.001**
sex [1]	0.15	0.13 – 0.17	**<0.001**
age	0.01	0.00 – 0.01	**<0.001**
estrato1 [2]	0.06	0.02 – 0.09	**0.001**
estrato1 [3]	0.15	0.11 – 0.19	**<0.001**
estrato1 [4]	0.58	0.52 – 0.63	**<0.001**
estrato1 [5]	0.76	0.68 – 0.83	**<0.001**
estrato1 [6]	1.15	1.08 – 1.23	**<0.001**

maxEducLevel	0.09	0.08 – 0.10	**<0.001**
hoursWorkUsual	0.01	0.01 – 0.01	**<0.001**
formal [1]	0.33	0.30 – 0.36	**<0.001**
cuentaPropia [1]	-0.11	-0.14 – -0.08	**<0.001**
oficio			
sizeFirm	0.07	0.07 – 0.08	**<0.001**
antiguedad	0.00	0.00 – 0.00	**<0.001**
Observations	16276		
R^2 / R^2 adjusted	0.515 / 0.512		

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

Como controles que consideramos relevantes en la ecuación se incluyó la edad, el estrato, el máximo nivel de educación, las horas semanales trabajadas usualmente, el estado de formalidad, si está autoempleado o no, el oficio de la persona, el tamaño de la firma que lo contrata y su antigüedad en ella. En la tabla a continuación no se reportan los coeficientes del control de oficio para hacer eficiente el uso del espacio, pero la tabla completa se puede ver al final del documento (Anexo 2).

Sí, obtenemos los mismos resultados:

Tabla 8. Regresiones OLS y FWL del logaritmo del ingreso total observado contra el sexo

	OLS			**FWL**		
Predictors	Coef	CI	p	Coef	CI	p
(Intercept)	12.64	12.35 – 12.94	**<0.001**	0.00	-0.01 – 0.01	1.000
sex [1]	**0.15**	**0.13 – 0.17**	**<0.001**	**0.15**	**0.13 – 0.17**	**<0.001**
age	0.01	0.00 – 0.01	**<0.001**			
estrato1 [2]	0.06	0.02 – 0.09	**0.001**			
estrato1 [3]	0.15	0.11 – 0.19	**<0.001**			
estrato1 [4]	0.58	0.52 – 0.63	**<0.001**			
estrato1 [5]	0.76	0.68 – 0.83	**<0.001**			
estrato1 [6]	1.15	1.08 – 1.23	**<0.001**			
maxEducLevel	0.09	0.08 – 0.10	**<0.001**			
hoursWorkUsual	0.01	0.01 – 0.01	**<0.001**			
formal [1]	0.33	0.30 – 0.36	**<0.001**			
cuentaPropia [1]	-0.11	-0.14 – -0.08	**<0.001**			
oficio						
sizeFirm	0.07	0.07 – 0.08	**<0.001**			
antiguedad industria	0.00	0.00 – 0.00	**<0.001**			

Observations	16276	16276
R^2 / R^2 adjusted	0.515 / 0.512	0.010 / 0.010

iii. 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 refleja que, en promedio, los hombres ocupados de Bogotá tienen un incremento de 15% en sus ingresos totales observados con respecto a las mujeres ocupadas de Bogotá. Este valor es estadísticamente significativo con 99% de confianza y su valor es igual al del modelo anterior. De modo que el *gap* de ingresos entre hombres y mujeres no se vio reducido por esta nueva aproximación y no parece ser un problema de selección. Ahora, los valores de R^2 y el R^2 ajustado de este modelo son de 1%. Esto indica que el modelo explica la variación del logaritmo del ingreso observado en 1% y no es una proporción muy alta, por lo que es *fit* no parece ser muy grande.

5) Predicting earnings

* Decidimos no reportar en las tablas todos los niveles de las variables categóricas, sin embargo, estos sí fueron incluidos en las estimaciones.

a) Split the sample into two samples: a training (70%) and a test (30%) sample.

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

Tabla 9. Estimación del modelo con solo la constante para determinación del *benchmark*

		log wage	
Predictors	**Estimates**	**CI**	**p**
(Intercept)	13.95	13.93 – 13.97	**<0.001**
Observations	13021		
R^2 / R^2 adjusted	0.000 / 0.000		

Error de predicción del benchmark: 0.78.

ii. Estimate again previous models.

Entendemos que, por *previous models,* la instrucción se refiere al modelo de la edad (punto 3) y el modelo relacionado con el sexo (punto 4). De este modo, corrimos el modelo del punto **3.b)** y el punto **4.e)** (el que incluye controles). Se hizo una modificación al modelo **3.b)**: se aplicó el logaritmo a la variable dependiente – el ingreso total observado- para que los estimadores y errores estándar de los dos modelos fueran más comparables.

Tabla 10. Regresión modelos logaritmo del ingreso total observado

	Model 1				Model 2		
	log wage						
	Coef	CI	p		Coef	CI	p
(Intercept)	12.69	12.34 – 13.03	<**0.001**	(Intercept)	12.74	12.60 – 12.88	<**0.001**
sex [1]	0.15	0.12 – 0.18	<**0.001**				
age	0.00	0.00 – 0.01	<**0.001**	age	0.06	0.06 – 0.07	<**0.001**
				age2	-0.00	-0.00 – -0.00	<**0.001**
estrato1 [2]	0.05	0.01 – 0.09	**0.015**				
estrato1 [3]	0.14	0.10 – 0.18	<**0.001**				
estrato1 [4]	0.56	0.50 – 0.62	<**0.001**				
estrato1 [5]	0.71	0.62 – 0.80	<**0.001**				
estrato1 [6]	1.17	1.08 – 1.25	<**0.001**				
maxEducLevel	0.09	0.08 – 0.10	<**0.001**				
hoursWorkUsual	0.01	0.01 – 0.01	<**0.001**				
formal [1]	0.31	0.28 – 0.35	<**0.001**				
cuentaPropia [1]	-0.11	-0.14 – -0.07	<**0.001**				
oficio							
sizeFirm	0.08	0.07 – 0.09	<**0.001**				
antiguedad industria	0.00	0.00 – 0.00	<**0.001**				
Obs	11393			Obs	11393		
R^2 / R^2 adj	0.517 / 0.513			R^2 / R^2 adj	0.026 / 0.026		

iii. 5 models increasing in complexity.

Se proponen los siguientes cinco modelos que van incrementando su complejidad:

Modelo 1

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \theta X + u$$

Modelo 2

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \theta X + u$$

Modelo 3

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 sex \cdot age + \theta X + u$$

Modelo 4

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 age^3 + \beta_5 sex \cdot age + \theta X + u$$

Modelo 5

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 age^3 + \beta_5 sex \cdot age + \beta_6 sex \cdot \max E\,duc + \theta X + u$$

A continuación, se muestra la regresión de los modelos propuestos.

Tabla 11. Cinco modelos con aumento de complejidad

Model 1		Model 2		Model 3		Model 4		Model 5	
				log wage					
	Coef		Coef		Coef		Coef		Coef
Const.	12.69	Const.	12.17	Const.	12.25	Const.	11.61	Const.	11.64
sex	0.15	sex	0.16	sex	-0.02	sex	-0.01	sex	-0.05
age	0.00	age	0.03	age	0.03	age	0.08	age	0.08
		age2	-0.00	age2	-0.00	age2	-0.00	age2	-0.00
						age3	0.00	age3	0.00
			0.05	sex * age	0.00	sex * age	0.00	sex * age	0.00
								sex * maxEduc	0.01
estrato2	0.05	estrato2	0.14	estrato2	0.05	estrato2	0.05	estrato2	0.05
estrato3	0.14	estrato3	0.56	estrato3	0.14	estrato3	0.14	estrato3	0.14
estrato4	0.56	estrato4	0.72	estrato4	0.56	estrato4	0.57	estrato4	0.57
estrato5	0.71	estrato5	1.18	estrato5	0.72	estrato5	0.73	estrato5	0.73
estrato6	1.17	estrato6	0.08	estrato6	1.18	estrato6	1.18	estrato6	1.18
maxEduc	0.09	maxEduc	0.01	maxEduc	0.08	maxEduc	0.08	maxEduc	0.08
hWork	0.01	hWork	0.30	hWork	0.01	hWork	0.01	hWork	0.01
formal	0.31	formal	-0.11	formal	0.30	formal	0.30	formal	0.30
indep	-0.11	indep	-0.21	indep	-0.11	indep	-0.11	indep	-0.11
oficio		oficio		oficio		oficio		oficio	
sizeFirm	0.08	sizeFirm	0.00	sizeFirm	0.08	sizeFirm	0.08	sizeFirm	0.08
antigüedad	0.00	antigüedad		antigüedad	0.00	antigüedad	0.00	antigüedad	0.00
Observations	11393	Observations	11393	Observations	11393	Observations	11393	Observations	11393
R^2 / R^2 adjusted	0.517 / 0.513	R^2 / R^2 adjusted	0.523 / 0.519	R^2 / R^2 adjusted	0.524 / 0.520	R^2 / R^2 adjusted	0.526 / 0.522	R^2 / R^2 adjusted	0.526 / 0.521

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.

La siguiente tabla muestra los errores de predicción de cada uno de los modelos propuestos.

Tabla 12. Error de predicción de los cinco modelos propuestos

Modelo	Error pred.	Orden
1	0.3882793	5
2	0.3852743	4
3	0.3849555	3
4	**0.3839591**	**1**
5	0.3840328	2

Todos los modelos tienen un mejor desempeño que el benchmark. Como se evidencia, el modelo 4 es el que tiene un menor error de predicción. A continuación, se hace una descripción de este modelo y las posibles ventajas que tiene.

Modelo 4

$$\log(Income) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 age^3 + \beta_5 sex \cdot age + \theta X + u$$

Este modelo cuenta con varios componentes interesantes. En primer lugar, tiene una cantidad de controles que especifica la ecuación con mayor detalle. Los controles incluyen el sexo, la edad, el estrato, el máximo nivel de educación, las horas semanales trabajadas usualmente, el estado de formalidad, si está auto-empleado o no, el oficio de la persona, el tamaño de la firma que lo contrata y su antigüedad en ella. En segundo lugar, se evidencia un polinomio de grado 3 en la ecuación. Esta forma de la función puede permitir un mayor nivel de ajuste al comportamiento de los datos. En tercer lugar, es muy interesante la interacción de edad con sexo, pues permite capturar la variación que se debe a estas dos variables. Los coeficientes más interesantes en este caso, desde nuestra perspectiva, son los de sexo, edad y el de la interacción entre sexo y edad. Además, los valores de R^2 y el R^2 ajustado de este modelo son 52.6% y 52.2%, respectivamente. Esto indica que más de la mitad de la variación de la variable dependiente está siendo explicada por el modelo y no es un ajuste despreciable.

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?

El Leverage Statistic del modelo 4 se dsitribuye de la siguiente forma:

Gráfico 4. Histograma del estadístico *Leverage*

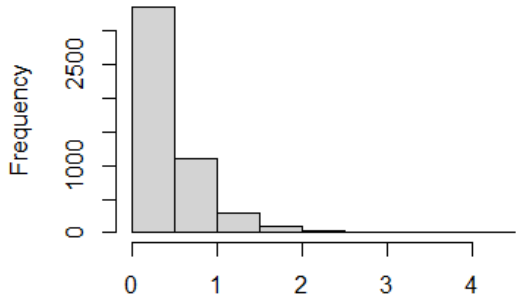


Tabla 13. Estadísticas descriptivas del estadístico *Leverage*

Min.	1st Q	Median	Mean	3rd Q	Max.
0.00018	0.13851	0.31333	0.43529	0.59000	4.41188

Vemos que, en general, el valor de este índice es bajo. Esto puede deberse a que el residual no es muy alto (el modelo es bueno), o que la influencia baja (no hay muchos outliers). No parece haber outliers que estén impactando los resultados del modelo. De todas formas valdría la pena revisar las observaciones en las que el indicador es suficientemente alto (e.g > percentil 80 = 0.6) para poder determinar si la razón de este problema viene de la muestra o del modelo.

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

Tabla 14. Error de predicción de los cinco modelos propuestos con *K-fold cross-validation*

Modelo	Error pred.	Orden
1	0.3930944	5
2	0.3888018	4
3	0.3872743	2
4	0.3874566	3
5	**0.3872713**	**1**

En general se puede observar que los errores predichos tienen un mejor desempeño que el *benchmark* y son relativamente similares a los valores obtenidos en el inciso anterior. No obstante, hay algunas diferencias en los órdenes de magnitud que vale la pena destacar. El modelo 1 y 2 siguen siendo los dos modelos con el error predicho más grande. El modelo cuatro pasó de ser el modelo con el error más pequeño a tener el tercero más grande, mientras que el modelo 3 pasó de tener el tercero más grande al segundo más pequeño. Por último, el modelo 5 contaba con el segundo error predicho más pequeño y ahora tiene el más pequeño. En conclusión, las similitudes residen en que sigue habiendo un mejor rendimiento que el *benchmark* y en que los valores son relativamente similares, mientras que la diferencia radica en el orden de las magnitudes obtenidas.

c) LOOCV

ii. Compare the results to those obtained in the computation of the leverage statistic.

El error de predicción obtenido de la estimación por LOOCV fue de 0.28. Una cifra significativamente menor a la media y la mediana del leverage statistic. Esto indica que este método de estimación tiene un mejor desempeño que el método tradicional (OLS) para un mismo modelo.

Anexos

Anexo 1: Tabla 7. Regresión logaritmo del ingreso total observado contra el sexo con controles

		log wage	
Predictors	**Coef**	**CI**	**p**
(Intercept)	12.64	12.35 – 12.94	**<0.001**
sex [1]	0.15	0.13 – 0.17	**<0.001**
age	0.01	0.00 – 0.01	**<0.001**
estrato1 [2]	0.06	0.02 – 0.09	**0.001**
estrato1 [3]	0.15	0.11 – 0.19	**<0.001**
estrato1 [4]	0.58	0.52 – 0.63	**<0.001**
estrato1 [5]	0.76	0.68 – 0.83	**<0.001**
estrato1 [6]	1.15	1.08 – 1.23	**<0.001**
maxEducLevel	0.09	0.08 – 0.10	**<0.001**
hoursWorkUsual	0.01	0.01 – 0.01	**<0.001**
formal [1]	0.33	0.30 – 0.36	**<0.001**
cuentaPropia [1]	-0.11	-0.14 – -0.08	**<0.001**
oficio [2]	-0.28	-0.56 – 0.01	0.062
oficio [3]	-0.70	-0.99 – -0.41	**<0.001**
oficio [4]	0.01	-0.56 – 0.58	0.973
oficio [5]	-0.27	-0.62 – 0.09	0.140
oficio [6]	-0.10	-0.40 – 0.20	0.507
oficio [7]	-0.35	-0.66 – -0.04	**0.026**
oficio [8]	-0.11	-0.41 – 0.18	0.456
oficio [9]	-0.06	-0.38 – 0.25	0.690
oficio [11]	-0.14	-0.43 – 0.15	0.333
oficio [12]	0.03	-0.26 – 0.32	0.834
oficio [13]	-0.36	-0.65 – -0.08	**0.013**
oficio [14]	-0.95	-1.52 – -0.37	**0.001**
oficio [15]	-0.35	-0.67 – -0.03	**0.034**
oficio [16]	-0.57	-0.87 – -0.27	**<0.001**
oficio [17]	-0.54	-0.86 – -0.23	**0.001**
oficio [18]	-0.62	-0.94 – -0.30	**<0.001**
oficio [19]	-0.41	-0.71 – -0.12	**0.006**
oficio [20]	-0.90	-1.81 – 0.01	0.053
oficio [21]	-0.19	-0.48 – 0.09	0.186
oficio [30]	-0.38	-0.68 – -0.08	**0.013**
oficio [31]	-0.06	-0.48 – 0.36	0.774
oficio [32]	-0.69	-0.99 – -0.38	**<0.001**
oficio [33]	-0.71	-1.00 – -0.42	**<0.001**
oficio [34]	-0.73	-1.05 – -0.41	**<0.001**
oficio [35]	-0.80	-1.28 – -0.32	**0.001**
oficio [36]	-0.92	-1.25 – -0.59	**<0.001**

oficio [37]	-0.94	-1.23 – -0.64	**<0.001**
oficio [38]	-0.88	-1.17 – -0.59	**<0.001**
oficio [39]	-0.73	-1.01 – -0.44	**<0.001**
oficio [40]	-0.62	-0.95 – -0.28	**<0.001**
oficio [41]	-0.70	-0.98 – -0.41	**<0.001**
oficio [42]	-0.49	-0.81 – -0.17	**0.003**
oficio [43]	-0.21	-0.60 – 0.19	0.304
oficio [44]	-0.48	-0.78 – -0.19	**0.001**
oficio [45]	-0.85	-1.13 – -0.57	**<0.001**
oficio [49]	-0.61	-1.18 – -0.03	**0.038**
oficio [50]	-0.51	-0.86 – -0.16	**0.004**
oficio [51]	-0.59	-0.89 – -0.28	**<0.001**
oficio [52]	-0.80	-1.37 – -0.23	**0.006**
oficio [53]	-0.77	-1.06 – -0.48	**<0.001**
oficio [54]	-0.86	-1.14 – -0.57	**<0.001**
oficio [55]	-0.84	-1.13 – -0.55	**<0.001**
oficio [56]	-0.81	-1.15 – -0.48	**<0.001**
oficio [57]	-0.79	-1.08 – -0.49	**<0.001**
oficio [58]	-0.84	-1.12 – -0.55	**<0.001**
oficio [59]	-0.81	-1.09 – -0.52	**<0.001**
oficio [60]	-0.83	-1.74 – 0.08	0.075
oficio [61]	-0.79	-1.16 – -0.42	**<0.001**
oficio [62]	-0.81	-1.16 – -0.47	**<0.001**
oficio [63]	-0.49	-1.40 – 0.42	0.295
oficio [70]	-0.58	-0.90 – -0.26	**<0.001**
oficio [72]	-0.86	-1.37 – -0.34	**0.001**
oficio [73]	-0.30	-1.21 – 0.61	0.521
oficio [74]	-0.88	-1.33 – -0.43	**<0.001**
oficio [75]	-1.30	-1.66 – -0.94	**<0.001**
oficio [76]	-0.58	-1.34 – 0.18	0.135
oficio [77]	-0.76	-1.05 – -0.46	**<0.001**
oficio [78]	-1.62	-2.88 – -0.37	**0.011**
oficio [79]	-0.84	-1.12 – -0.55	**<0.001**
oficio [80]	-0.96	-1.26 – -0.65	**<0.001**
oficio [81]	-0.81	-1.12 – -0.51	**<0.001**
oficio [82]	-1.07	-1.98 – -0.16	**0.022**
oficio [83]	-0.74	-1.05 – -0.44	**<0.001**
oficio [84]	-0.71	-1.00 – -0.42	**<0.001**
oficio [85]	-0.68	-0.98 – -0.39	**<0.001**
oficio [86]	-0.38	-0.95 – 0.20	0.196
oficio [87]	-0.76	-1.06 – -0.47	**<0.001**
oficio [88]	-0.91	-1.43 – -0.40	**0.001**
oficio [89]	-0.74	-1.13 – -0.35	**<0.001**
oficio [90]	-0.87	-1.19 – -0.54	**<0.001**
oficio [91]	-0.92	-1.35 – -0.49	**<0.001**
oficio [92]	-0.80	-1.13 – -0.48	**<0.001**

oficio [93]	-0.81	-1.11 – -0.51	**<0.001**
oficio [94]	-1.09	-1.43 – -0.76	**<0.001**
oficio [95]	-0.64	-0.93 – -0.35	**<0.001**
oficio [96]	-0.14	-1.05 – 0.77	0.769
oficio [97]	-0.92	-1.21 – -0.64	**<0.001**
oficio [98]	-0.74	-1.02 – -0.45	**<0.001**
oficio [99]	-0.96	-1.26 – -0.65	**<0.001**
sizeFirm	0.07	0.07 – 0.08	**<0.001**
antiguedad	0.00	0.00 – 0.00	**<0.001**
Observations	16276		
R^2 / R^2 adjusted	0.516 / 0.512		

Anexo 2: Tabla 8. Regresiones OLS y FWL del logaritmo del ingreso total observado contra el sexo

	OLS			FWL		
Predictors	Coef	CI	p	Coef	CI	p
(Intercept)	12.64	12.35 – 12.94	**<0.001**	0.00	-0.01 – 0.01	1.000
sex [1]	**0.15**	**0.13 – 0.17**	**<0.001**	**0.15**	**0.13 – 0.17**	**<0.001**
age	0.01	0.00 – 0.01	**<0.001**			
estrato1 [2]	0.06	0.02 – 0.09	**0.001**			
estrato1 [3]	0.15	0.11 – 0.19	**<0.001**			
estrato1 [4]	0.58	0.52 – 0.63	**<0.001**			
estrato1 [5]	0.76	0.68 – 0.83	**<0.001**			
estrato1 [6]	1.15	1.08 – 1.23	**<0.001**			
maxEducLevel	0.09	0.08 – 0.10	**<0.001**			
hoursWorkUsual	0.01	0.01 – 0.01	**<0.001**			
formal [1]	0.33	0.30 – 0.36	**<0.001**			
cuentaPropia [1]	-0.11	-0.14 – -0.08	**<0.001**			
oficio [2]	-0.28	-0.56 – 0.01	0.062			
oficio [3]	-0.70	-0.99 – -0.41	**<0.001**			
oficio [4]	0.01	-0.56 – 0.58	0.973			
oficio [5]	-0.27	-0.62 – 0.09	0.140			
oficio [6]	-0.10	-0.40 – 0.20	0.507			
oficio [7]	-0.35	-0.66 – -0.04	**0.026**			
oficio [8]	-0.11	-0.41 – 0.18	0.456			
oficio [9]	-0.06	-0.38 – 0.25	0.690			
oficio [11]	-0.14	-0.43 – 0.15	0.333			
oficio [12]	0.03	-0.26 – 0.32	0.834			
oficio [13]	-0.36	-0.65 – -0.08	**0.013**			
oficio [14]	-0.95	-1.52 – -0.37	**0.001**			

oficio [15]	-0.35	-0.67 – -0.03	**0.034**
oficio [16]	-0.57	-0.87 – -0.27	**<0.001**
oficio [17]	-0.54	-0.86 – -0.23	**0.001**
oficio [18]	-0.62	-0.94 – -0.30	**<0.001**
oficio [19]	-0.41	-0.71 – -0.12	**0.006**
oficio [20]	-0.90	-1.81 – 0.01	0.053
oficio [21]	-0.19	-0.48 – 0.09	0.186
oficio [30]	-0.38	-0.68 – -0.08	**0.013**
oficio [31]	-0.06	-0.48 – 0.36	0.774
oficio [32]	-0.69	-0.99 – -0.38	**<0.001**
oficio [33]	-0.71	-1.00 – -0.42	**<0.001**
oficio [34]	-0.73	-1.05 – -0.41	**<0.001**
oficio [35]	-0.80	-1.28 – -0.32	**0.001**
oficio [36]	-0.92	-1.25 – -0.59	**<0.001**
oficio [37]	-0.94	-1.23 – -0.64	**<0.001**
oficio [38]	-0.88	-1.17 – -0.59	**<0.001**
oficio [39]	-0.73	-1.01 – -0.44	**<0.001**
oficio [40]	-0.62	-0.95 – -0.28	**<0.001**
oficio [41]	-0.70	-0.98 – -0.41	**<0.001**
oficio [42]	-0.49	-0.81 – -0.17	**0.003**
oficio [43]	-0.21	-0.60 – 0.19	0.304
oficio [44]	-0.48	-0.78 – -0.19	**0.001**
oficio [45]	-0.85	-1.13 – -0.57	**<0.001**
oficio [49]	-0.61	-1.18 – -0.03	**0.038**
oficio [50]	-0.51	-0.86 – -0.16	**0.004**
oficio [51]	-0.59	-0.89 – -0.28	**<0.001**
oficio [52]	-0.80	-1.37 – -0.23	**0.006**
oficio [53]	-0.77	-1.06 – -0.48	**<0.001**
oficio [54]	-0.86	-1.14 – -0.57	**<0.001**
oficio [55]	-0.84	-1.13 – -0.55	**<0.001**
oficio [56]	-0.81	-1.15 – -0.48	**<0.001**
oficio [57]	-0.79	-1.08 – -0.49	**<0.001**
oficio [58]	-0.84	-1.12 – -0.55	**<0.001**
oficio [59]	-0.81	-1.09 – -0.52	**<0.001**
oficio [60]	-0.83	-1.74 – 0.08	0.075
oficio [61]	-0.79	-1.16 – -0.42	**<0.001**
oficio [62]	-0.81	-1.16 – -0.47	**<0.001**
oficio [63]	-0.49	-1.40 – 0.42	0.295
oficio [70]	-0.58	-0.90 – -0.26	**<0.001**
oficio [72]	-0.86	-1.37 – -0.34	**0.001**
oficio [73]	-0.30	-1.21 – 0.61	0.521
oficio [74]	-0.88	-1.33 – -0.43	**<0.001**
oficio [75]	-1.30	-1.66 – -0.94	**<0.001**
oficio [76]	-0.58	-1.34 – 0.18	0.135
oficio [77]	-0.76	-1.05 – -0.46	**<0.001**
oficio [78]	-1.62	-2.88 – -0.37	**0.011**

oficio [79]	-0.84	-1.12 – -0.55	**<0.001**
oficio [80]	-0.96	-1.26 – -0.65	**<0.001**
oficio [81]	-0.81	-1.12 – -0.51	**<0.001**
oficio [82]	-1.07	-1.98 – -0.16	**0.022**
oficio [83]	-0.74	-1.05 – -0.44	**<0.001**
oficio [84]	-0.71	-1.00 – -0.42	**<0.001**
oficio [85]	-0.68	-0.98 – -0.39	**<0.001**
oficio [86]	-0.38	-0.95 – 0.20	0.196
oficio [87]	-0.76	-1.06 – -0.47	**<0.001**
oficio [88]	-0.91	-1.43 – -0.40	**0.001**
oficio [89]	-0.74	-1.13 – -0.35	**<0.001**
oficio [90]	-0.87	-1.19 – -0.54	**<0.001**
oficio [91]	-0.92	-1.35 – -0.49	**<0.001**
oficio [92]	-0.80	-1.13 – -0.48	**<0.001**
oficio [93]	-0.81	-1.11 – -0.51	**<0.001**
oficio [94]	-1.09	-1.43 – -0.76	**<0.001**
oficio [95]	-0.64	-0.93 – -0.35	**<0.001**
oficio [96]	-0.14	-1.05 – 0.77	0.769
oficio [97]	-0.92	-1.21 – -0.64	**<0.001**
oficio [98]	-0.74	-1.02 – -0.45	**<0.001**
oficio [99]	-0.96	-1.26 – -0.65	**<0.001**
sizeFirm	0.07	0.07 – 0.08	**<0.001**
antiguedad industria	0.00	0.00 – 0.00	**<0.001**
Observations	16276		16276
R^2 / R^2 adjusted	0.515 / 0.512		0.010 / 0.010