Problem Set 1

Camilo Bonilla Rafael Santofimio Nicolás Velásquez

May 11, 2022

Abstract

Para el sector fiscal, la correcta implemetación de políticas públicas y reformas tributarias han sido temáticas importantes a la hora de poder formar naciones con sociedades crecientes en ingreso y con redistribuciones que pérmitan una equidad para los sectores marginales con igualdad de oportunidades. En un mundo ideal el gobierno tendría control total al observar los ingresos de todas las personas, sin embargo actualmente las personas suelen subreportar estas cifras con temor a ser segmentada bajo una tasa proporsional a sus ingresos. Con base en lo anterior, se vuelve de suma importancia obtener herramientas que nos permitan tener estimaciones de los ingresos de las personas y llegar a cuestionarnos si debido a un subreporte de los mismos se estan teniendo impuestos suboptimos. Más importante aún si dentro de los multiples modelos posibles, el gobierno y entidades reguladoras se debería enfocar en ciertas observaciones atipicas para llegar a una estimación cercana.

A partir de un ejercicio con la base de datos del GEIH del Dane del 2008, se pretende dar con un modelo de predicción de los ingresos, aprender también desde los datos, perfiles de ingresos segun la edad de los encuestados y el sesgo o brecha de genero. El siguiente escrito se dividirá de la siguiente manera. En la sección Data se explica la optención de la data, limpieza y criterios para la segmentación de variables, seguido de una breve descripción de las variables sociodemográficas entre otras. En la sección "Age Earning Profile", se estima el ingreso de las personas en Bogotá a través de su vida laboral (tiempo).

Posteriormente, en la sección 3 llamada "Gender Erning Gap" se aborda el perfil de ingresos para hombres y mujeres, verificando la hipótesis de la brecha salarial que se presenta en los primeros años laborales y su aumento a lo largo de la vida. Así mismo, se determina la edad cuando alcanza el máximo ingreso salarial para cada uno de los generos.

Finalmente, en la sección 4, "Predicting Earnings", se retomaron los modelos estimados anteriormente y se comparan con algunas especificaciones adicionales en las que se exploraron algunas no linealidades y metodologias LOOCV.

Los archivos, tablas, bases de datos y códigos totalmente replicables se pueden obtener en el siguiente link: ProblemSet1 el cual redirige al repositorio en Github.

Keywords: Earnings, Earning Gender GAP, Taxes, Predicting Model [1]ProblemSet1

[1]https://github.com/Nicolas-Velasquez-Oficial/Problem-Set1

1 Data

Para la extracción de la base de datos, se relizó un raspado de la pagina web GEIH2018Sample . De allí se obtuvo 10 tablas, estas se agruparon por nombre de columna, dando como resultado una base con 178 variables y 32.177 observaciones.Al realizar esta estracción se observó que las diferentes tablas se demoraban al cargar posiblemente por el peso y por la forma especifica que estaban alojadas en el sitio web. Debido a lo anterior,el código utiliza la url directamente de la dirección en la que esta viculada la data para hacer el proceso relativamente más rapido.

En la etapa de limpieza y filtración, la base de datos se segmentó para la población con edad mayor a 18 años y en situación laboral de empleados. Acto seguido, se definieron las varibles continuas y categoricas segun el diccionario del Dane. Se completaron dos varibles categoricas Regimen de Salud y maximo nivel de educación con la moda, las cuales tenian tasas de copletitud del 91.4% y 99.9%, porcentajes que son deseables para hacer imputaciones en los datos faltantes.

Finalmente se filtra la base para las variables más pertinetes para el estudio, lo que culmina con 17 variables y 16542 observaciones para los 12 meses del año 2008.

Table 1: Tabla datos descriptivos

Statistic	N	Mean	St. Dev.	Min	Max
mes	16,542	6.453	3.372	1	12
sex	16,542	0.530	0.499	0	1
age	16,542	39.436	13.483	18	94
ingtot	16,542	1,769,379.000	2,675,628.000	0.000	85,833,333.000
college	16,542	0.319	0.466	0	1
totalHoursWorked	16,542	47.403	15.662	1	130
formal	16,542	0.587	0.492	0	1
sizeFirm	16,542	3.152	1.649	1	5
y_ingLab_m	9,892	1,745,416.000	2,403,441.000	20,000.000	60,100,000.000
y_salary_m	16,542	1,556,200.000	1,661,599.000	10,000	34,000,000
y_total_m	14,764	1,617,551.000	2,431,319.000	84.000	70,000,000.000

Fuente: (Dane,2008)

Como se puede ver en la tabla #1, la base continene variables como edad, genero, estrato, college, entre otras variables sociodemográfias. los encuestados en esta muestra possen una media de edad de casi 40 años entre 18 y 94 años de edad. El 53% de la muestra son hombres y 32% de la muestra a completado la educación terciaria. También se observa una tasa de formalidad de 58.7% y una media de 47 horas trabajadas por persona.

Una de las discuciones más prominentes es la escogencia de la varible dependiente que se va a utilizar para estimar los salarios/ganancias de las personas en las siguientes secciones del trabajo. Como se observa en la tabla descriptiva hay diferentes variables como ingtot, ingLab, salary y ytotal. Segun las medias reportadas entre estas variables los salarios tendrian una media entre 1.56 y 1.76 millones de pesos colombianos con desviciones estandar relativamente parecidas entre 2.14 y 2.65 millones de pesos. La gran diferencia entre estas

variables depende del número de datos no reportados o faltantes y el comportamiento o distrubución de las mismas. Observando la media y desviación estandar pareciera que no varía mucho. Para poder rectificar en la figura 5.1 se observa un correlograma entre estas varibles. Como se observa es casi de 1 la correlación entre estas varibles por lo que escoger una de las 4 sería valido, sin embargo los datos que posee la variable ingtot puede diferenciarse respecto a salary o ingLab por los retornos al capital, es por esto que parece prudente utilizar la varible más completa y que segun nuestro criterio sería la que más se acercaría a los ingresos de las personas con todas las limitaciones que se comentaron anteriormente.

Para la sección 3 en donde se estiman los alarios, la variable escogida es salary, precisamente por lo que se quiere ver brechas en el salario, estimar esto con las demás varibles de la base de datos anteriormente nombradas, generaría impreciciones y ruido al trtar de estimar estos picos de edad tanto para hombres como para mujeres. En particular al tomar (y salary m) se tuvo que imputar la mediana como medida para lidiar con los datos missing, pues sin esto era imposible generar los respectivos modelos sin tener que perder observaciones.

En la Figura 5.3 se obserba el histograma de la variable IngTot, la cual posee una acumulación de la muestra en la cola izquierda y su forma no parece ser una distribución normal. Siembargo esto se podría corregir mediante alguna transformación logaritmica tradicional o normalización.

En la figura #5.2 en el Apendice, se observa la correlación entre las distintas varibles de la base de datos. Uno de los posibles predictores más fuertes para el Ingreso Total podría ser si es formal o no ya ue tiene una correlación relativamente alta (0.3), también se observa que la varible formal posee correlaciones relativamnente alta con tamaño de la firma, lo que podria indicar una posible multicolinealidad entre estas dos varibles al entrar como control en el modelo.

Por último la figura 5.4 y 5.5 se observan gráficos de barras para estrato y nivel máximo de educación. En la muestra se observa que la mayor parte de la población se encuentra en estrato 2 y 3.

2 Age Earning Profile

La variable utilizada para la estimación de las ganacias es ingreso total(Ingtot), esta variable es la que acumula todos los tipos de ingresos de diferentes variables segmentadas que hay en la base de datos original sin datos faltantes. Por lo anterior y razones ya dichas en la anterior sección, esta será la variable dependiente para estimar el perfil de ingresos de edad.

$$Earnings_{ij} = \beta_0 + \beta_1 Age_{ij} + \beta_2 Age_{ij}^2 + \epsilon_{ij} \tag{1}$$

La formula 1 muestra la especificación a estimar, donde β_0 es el intercepto de la función en el eje Y(Earnings). β_1 y β_2 es el coeficiente asociado a la edad y edad al cuadrado del encuestado y ϵ_{ij} es el error en la muestra.

Table 2: Age Earning Profile Model

	Dependent variable:
	Earnings
age	91,143.460***
	(8,886.416)
I(ageˆ2)	−799.261***
	(102.852)
Constant	−436,662.900**
	(178,347.200)
Observations	16,542
R^2	0.017
Adjusted R^2	0.017
Residual Std. Error	2,652,732.000 (df = 16539)
F Statistic	144.382*** (df = 2; 16539)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fuente: (Dane,2008)

Como se observa en la tabla #2 el modelo que estima el perfil de ingresos por edad, indica que una persona sin ningun año de edad posee un salario promedio de -$436,662.9 COP al 1% de significancia, dejando todo lo demás constante. Por otro lado un año adicional de edad genera un aumento promedio de $ 91,143.46 COP en el salario de las personas, este resultado es significativo a cualquier nivel de significancia tradicional (10%, 5% y 1%) dejando todo lo demás constante. Adicionalmente podemos observar que por cada año adicional de edad, el ingreso marginal de las personas en promedio decrece en -$799.26 COP, este resultado es significativo a cualquier nivel de significancia tradicional, dejando todo lo demás constante. Con base en lo anterior, podemos decir que el comportamiento del salario a medida que aumentan los años en las personas no tiene una relación lineal sino decreciente, siendo la función cóncava entre los salarios predichos y la edad.

El R^2 indica que este modelo explica el 1.7% de la variabilidad de los ingresos,pese a este resultado esta especificación no es muy buena al predecir los salarios en la muestra, este valor no es distinto al R^2 ajustado que penaliz por los regresores incluidos. También se obtuvo un F Statistic = 144.382, significativo a cuañquier nivel de significancia trdicional, lo que nos indica que al menos una variable del modelo es importante en la explicación de los salarios de las personas.

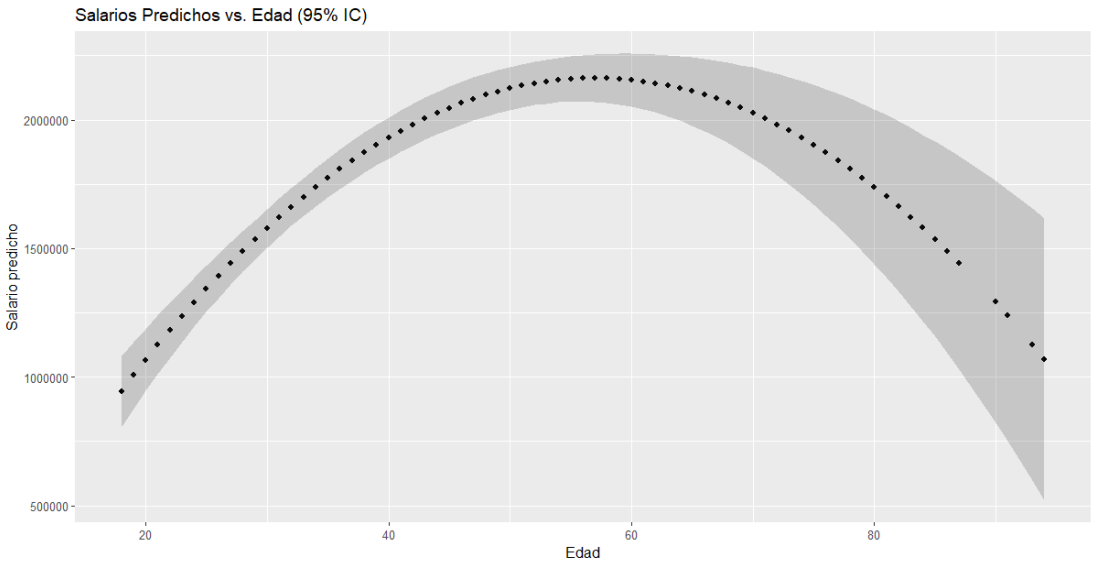
Los error estandar de los residuos es igual a $2,652,732 COP esta medida es la raiz del error cuadratico medio sobre los grados de libertad 16,539. el cual puede ser relativamente alto pese a las pocas varibles explicativas introducidas en el modelo.

En la figura 5.6 en el apendice, se muestra los valores predichos del modelo contra los residuales. Obsevando el comportamiento de la gráfica se puee inferir el buen ajuste que tienen las predicciones en la parte izquierda de la gráfica, a medida que aumentan los valores de las personas que más ganan en la muestra, el modelo parece aumentar en el error de sus

estimaciones. Esto mismo se puede ver en el gráfico 5.7 donde el error del modelo se dispara a medida que se aleja de la diagonal de 45 grados entre los cuantiles de las distribuciones de los residuos.

Un aspecto importante a resaltar sobre este modelo es su ajuste dentro de la muestra, debido a que este modelo solo posee dos variables explicativas su poder de predicción es muy bajo, y esto se puede deber precisamente a alguna variable omitida que sea de importancia para el modelo.

Figure 2.1: Predicción de Salario vs Edad



Note: Esta figura muestra la predicción del salario segun la edad del encuaestado.
Fuente: (Dane,2008)

Como se observa en la figura 2.1 el comportamiento de los ingresos respecto a la edad es similar a una función cóncava en la muestra de mayor edad, al haber menor muestra que en la primera parte el intervalo de confianza se hace más grande por una mayor varianza. De igual manera el salario de las personas aumenta pero empieza a decrecer marginalemnte al punto llegar a un máximo, este punto se refiere al "Peak age". Este punto en la curva se puede obtener de la siguiente derivada.

$$Earnings_{ij} = \beta_0 + \beta_1 Age_{ij} + \beta_2 Age_{ij}^2 + \epsilon_{ij}$$
$$\frac{\partial Earnings_{ij}}{\partial Age} = 0; \beta_1 + 2 * \beta_2 Age_{ij} = 0$$
$$Age = \frac{-\beta_1}{2 * \beta_2}; Age = -(91,143.460)/2 * (-799.261)$$
$$PeakAge = 57.01$$

Según la estimación del modelo de ingresos de la tabla 2, la edad promedio a la que se alcanza el salario máximo de la muestra es de 57 años, dejando todo lo demás constante. El anterior valor corresponde a unas ganacias promedio de $2'161.591 COP cuyo intervalo de confianza de este reporte esta entre $2'136.186 COP y $2'186.996 COP años de edad a un 95% de confianza dejando todo lo demás constante.

3 Gender Earning GAP

La brecha salarial de genero cada vez ha ido recobrando mayor relevancia, y en Colombia no ha sido la excepción para los creadores de política pública. Así entonces, en esta sección se toma la variable ingreso salarial mensual para determinar esta brecha. Se elige esta variable debido a que captura la remuneración percibida por labores desempeñadas en cerca de 96 oficios relacionados con el sector formal e informal de la Bogotá urbana. En cambio, si se mantuviera el ingreso total que se usó en el apartado anterior, se estuviera capturando efectos provenientes de otras actividades que no necesariamente corresponden a actividades laborales como lo son las ganancias ocasionales o el retorno al capital.

$$log(salary)_{ij} = \beta_0 + \beta_1 female_{ij} + \epsilon_{ij} \tag{2}$$

El primer modelo que se propone para establecer esta brecha salarial es relativamente sencillo, ya que no contempla variables de control sino únicamente el genero, y su variable dependiente (salario mensual) la cual se le aplica una trandformación logaritmica para obtener estabilidad en los regresores y suavizar la variable por la presencia de outliers para salarios más altos.

Table 3: Gender Earning GAP Model 1

	Dependent variable:
	Ln(income)
female	-0.149^{***}
	(0.015)
Constant	13.977^{***}
	(0.011)
Observations	9,892
R^2	0.010
Adjusted R^2	0.010
Residual Std. Error	0.751 (df = 9890)
F Statistic	97.364^{***} (df = 1; 9890)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaborado a partir de (Dane,2008)

Como se reporta en la tabla 3, este modelo de semi-elasticidad cuenta con una baja capacidad de predicción o más bien un bajo ajuste dado que 1% de los valores que toma Log(salario mensual) son predichos por la variable de genero, cuestión que es consecuente con el alto RSE reportado (0.751) y el cual va ser comparado con los otros dos modelos porpuestos en esta sección. Por otro lado, se observa que tanto el intercepto y el estimador correspondiente a female son significativos al 1%, esto implica que, las mujeres perciben en promedio 14.9% menos en comparación a los hombres, dejando todo lo demás constante. Con el animo de lograr mayor precisión y analizar el perfil salarial según la edad, se propone el siguiente modelo.

$$log(income)_{ij} = \beta_0 + \beta_1 female_{ij} + \beta_2 Age_{ij} + \beta_3 Age_{ij}^2 + \beta_4 female * Age_{ij} + \beta_5 female * Age_{ij}^2 + \epsilon_{ij} \quad (3)$$

En este nuevo modelo se introducen efectos hetereogeneos entre sexo y edad con el animo de cuantificar la relación entre las mismas, la dirección del efecto y concavidad o pico donde se alcanza el ingreso máximo.

Table 4: Gender Earning GAP Model 2

	Dependent variable:
	Ln(income)
female	0.082
	(0.139)
age	0.086***
	(0.005)
I(age^2)	−0.001***
	(0.0001)
female:age	−0.003
	(0.007)
female:I(age^2)	−0.0001
	(0.0001)
Constant	12.228***
	(0.097)
Observations	9,892
R^2	0.071
Adjusted R^2	0.070
Residual Std. Error	0.728 (df = 9886)
F Statistic	151.001*** (df = 5; 9886)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaborado a partir de (Dane,2008)

Con estas modificaciones se evidencia una pequeña mejora en el ajuste del modelo, pues pasó predecir el 1% a 7% del log(salario mensual), cuestión que resulta preocupante porque su RSE sigue sobre del orden del 0.7, es decir, el modelo aún sigue sin ajustarse fielmente a los resultados y por lo tanto, no se pueden realizar predicción con buena precisión. Por otro lado, la variable sexo (female) ya no es significativa al 5% ni al 10%, esto mismo ocurre con las interacciones. En la tabla 4 se observa que los estimadores asociados al cuadrado de la edad son negativos, esto indica la concavidad de la función de la variable dependiente, en otras palabras, el salario mensual percibido por un individuo, indistintamente si es hombre o mujer alcanza un pico máximo a una edad determinada. Así entonces, por cada año, el

ingreso de una persona en la capital incrementa en 8.6% hasta alcanzar un pico y luego disminuye con el mismo gradiente, todo esto bajo un nivel de significancia del 0.01.

Para calcular la edad en la que se alcanza el máximo salario para cada sexo, se deriva el modelo 2 respecto a la edad. Se establece que a las mujeres les toma 1.5 años menos alcanzar su pico, sin embargo, este es cerca del 2.8% menos que el logrado por los hombres. En otras palabras, las mujeres en promedio consiguen obtener su máximo salario mensual cuando llegan a los 41.5 años, lo que representa entre $ 950.743.89 y $ 960.299.04 el log (salario mensual). En cambio, los hombres en promedio llegan a devengar mensualmente entre $ 1.292.385.12 a $ 1.305.373.80 el log (salario mensual) cuando tienen 43 años. Todo esto con un nivel de confianza del 99%.

$$log(income)_{ij} = \beta_0 + \beta_1 female_{ij} + \beta_2 Age_{ij} + \beta_3 Age_{ij}^2 + \beta_4 female * Age_{ij} + \beta_5 female * Age_{ij}^2 + \epsilon_{ij}$$

$$\frac{\partial log(income)_{ij}}{\partial Age} = 0; \beta_2 + \beta_3 2 * Age_{ij} + \beta_4 female + \beta_5 2 female * Age_{ij} = 0$$

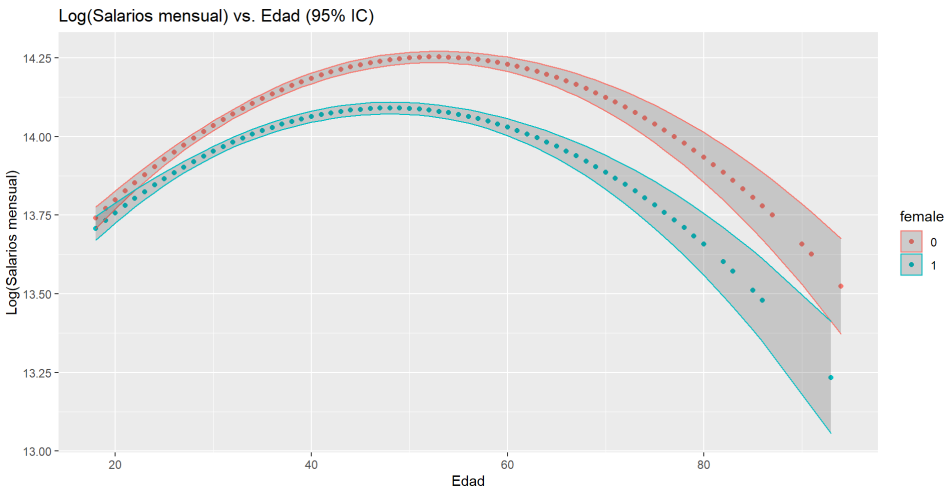
$$Age = \frac{-(\beta_2 + \beta_4 female)}{2 * (\beta_3 + \beta_5 female)}$$

$$Age_{female} = 41.5 años$$

$$Age_{male} = 43 años$$

En la Figura 2. se observa estos picos estimados a partir de la maximización con derivadas. Aunque el desplazamiento entre los picos no es evidente, el GAP o brecha si lo es. De hecho, los intervalos de confianza permiten evidenciar que no hay empate técnico en este punto de la gráfica, solo al inicio y fin de la vida laboral.

Figure 3.1: Log(Salario mensual) vs Edad por sexo



Note: Esta figura muestra la predicción del logaritmo del salario mensual segun la edad y el sexo del encuaestado.
Fuente: Elaborado a partir (Dane, 2008)

Ahora, para ahondar un poco más sobre la discusión acerca del trabajo e ingreso equitativos, se introducen una serie variables que permiten establecer o más bien realizar comparaciones

entre trabajadores y empleos con características similares. Así entonces, se introduce la variable oficio, tamaño de la firma, relación laboral, total horas trabajadas y nivel de educación.

$$log(income)_{ij} = \beta_0 + \beta_1 female + \beta_2 Age + \beta_3 Oficio + \beta_4 MaXEduLevel + \beta_5 sizeFirm +$$

$$\beta_6 relab + \beta_7 totalHoursWorked + \epsilon_{ij} (4)$$

Con estos nuevos controles, el modelo determina que por cada peso ganando por los hombres, las mujeres reciben 4.1% menos, esto con un nivel de significancia del 0.01 y manteniendo lo demás constante. A su vez, se evidencia una mejora significativa en el poder de predicción del modelo, pues pasó de un RSE de 0.71 a 0.46, lo que explica que 42% (R2) de los valores que toma la variable log(salario mensual) es explicado por covariables o controles. En el caso particular de una mujer de 25, que tiene educación terciaria completa y trabaja en el oficio de la arquitectura o ingeniería durante 40 horas a la semana, como empleada de una empresa particular, y la cual tiene entre 6 a 10 trabajadores, su ingreso esperado es de $2.138.467.48, y para un hombre con las mismas características de empleo se espera que reciba $2.227.075.83, es decir, $1.063.300.2 más por año que la mujer. Dado que el modelo de la tabla 6 contiene 108 parámetros de las estimaciones de este nuevo modelo, esta sin reportar los coeficientes de la variable oficio, sinembargo se puede consultar la tabla completa en la carpeta Views del repositorio o en el apéndice.

Para corroborar este resultado, se aplica el teorema FWL (Frisch-Waugh-Lovell Theorem) el cual permite en este caso estimar el efecto causal de la variable sexo sobre el ingreso mensual de los individuos, esta estimación numéricamente debe ser igual a la estimación del modelo 3 que se reportan en la tabla 5.

Table 5: Gender Earning GAP Model 2

	Dependent variable:	
	Ln(income)	FWL
	(1)	(2)
female1	-0.041^{***}	
	(0.009)	
mod_res2		-0.041^{***}
		(0.009)
Observations	16,542	16,542
R^2	0.423	0.001
Adjusted R^2	0.420	0.001
Residual Std. Error	0.464 (df = 16442)	0.463 (df = 16541)
F Statistic	121.781^{***} (df = 99; 16442)	21.778^{***} (df = 1; 16541)
Note:		*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaborado a partir de (Dane,2008)

Table 6: Desviaciones estándar usando FWL y boots

Método	s.d.
MCO controles	0.009
FWL	0.009
Boots	0.00922

Note que no solo el estimador sino también el error y RSE son iguales a los estimados por el modelo 3, lo que comprueba el teorema (Ver tabla 6). De nuevo resulta evidente que en promedio se presenta una brecha salarial del 4.1% bajo un nivel de significancia del 0.01.

A la luz de los resultados, la evidencia muestra que efectivamente en promedio no hay equidad de género a la hora de la remuneración salarial en Bogotá. De hecho, la brecha (GAP) según el modelo 3, corresponde a un 4,1% en promedio. Lo cual es un panorama desolador para las mujeres sin importar su formación, tipo de oficio o edad. Así entonces, la ciudad de Bogotá esta frente a un crudo escenario de discriminación laboral basado en genero que debe ser manejado no solo a nivel de empresa sino también a la hora de imputar los impuestos, pues pueden ser regresivos sino se le da el tracto adecuado a esta población.

4 Predicting earnings

Para comparar el desempeño de las distintas especificaciones exploradas hasta este punto y añadir a la comparación otras 5 especificaciones, se dividió de manera aleatoria la base de datos en una muestra de entrenamiento (70%) y una de prueba (30%). Las especificaciones que se exploraron se listan a continuación:

(1) $\log(Earnings_i) = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \epsilon_i,$

(2) $\log(Earnings_i) = \beta_0 + \beta_1 female * agei + \beta_2 Age_i^2 + \beta_3 female_i * Age_i^2 + \epsilon_i,$

(3) $\log(Earnings_i) = \beta_0 + \beta_1 Female_i + \beta_2 totalHoursWorked_i + \beta_3 sizeFirm_i + \beta_4 relabi + \beta_5 maxEducLeveli + \beta_6 oficioi + \epsilon_i$

(4) $\log(Earnings_i) = \beta_0 + \beta_1 femalei + \beta_2 sizeFirmi + \beta_3 totalHoursWorkedi + \beta_4 totalHoursWorked_i * sizeFirmi + \beta_5 totalHoursWorked_i^2 * sizeFirmi + \beta_6 totalHoursWorked_i^3 * sizeFirmi + \epsilon_i$

(5) $\log(Earnings_i) = \beta_0 + \beta_1 age_i + \beta_2 agei^2 + \beta_3 agei^3 + \beta_4 maxEducLevel_i + \epsilon_i$

(6) $\log(Earnings_i) = \beta_0 + \beta_1 log(age)_i + \beta_2 log(age)_i^2 + \beta_3 log(age)_i^3 + \beta_4 llog(totalHoursWorked)_i + \beta_5 log(totalHoursWorked)_i^2 + \beta_6 log(totalHoursWorked)_i^3 + \beta_7 log(age)_i * log(totalHoursWorked) + \beta_7 log(age)_i^2 * log(totalHoursWorked)^2 + \beta_7 log(age)_i^3 * log(totalHoursWorked)^3 + \epsilon_i$

Para evaluar el desempeño de los distintos modelos, se escogió el RMSE porque penaliza aquellas predicciones que se alejan demasiado del valor realmente observado. Además, esta métrica da una indicación de la distancia con respecto al valor real en las unidades en las que este valor se midió (en este caso, se trata del logaritmo del salario mensual).

Los desempeños para los modelos explorados, en términos del RMSE, se citan a continuación:

Table 7

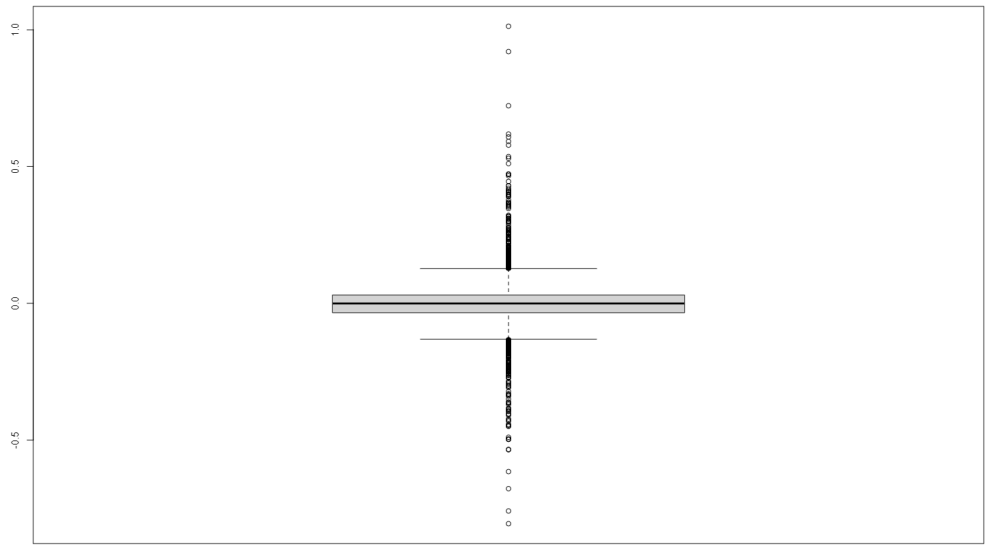
(Mod. 1)	(Mod. 2)	(Mod. 3)	(Mod. 4)	(Mod. 5)	(Mod. 6)
0.590	0.588	0.462	0.574	0.553	0.582

Como se puede ver, el modelo 3 y el modelo 5 fueron los de mejor desempeño, es decir, fueron los modelos que predijeron de mejor manera (con menos error) las observaciones de la muestra de prueba. Sin embargo, los demás modelos no tuvieron un desempeño tan alejado, lo cual puede deberse a que los tipos de complejidad explorados no se prestaban para que se generara un *overfitting* excesivo.

Con respecto al modelo 3, se calculó el estadístico de influencia para cada una de las observaciones de la muestra de prueba, con el fin de determinar qué tanto cambia la estimación al incorporar cada una de esas observaciones. Este estadístico de influencia usa los residuos y la matriz de proyección del modelo de regresión para asignar un valor de influencia estandarizado a cada una de las observaciones.

Como se puede ver en la figura 4.1, la mayoría de observaciones no alcanza a tener un grado de influencia que se pueda considerar relativamente alto. Al estandarizar las influencias de todas las observaciones en la muestra de prueba, se encuentra que ninguna de estas influencias excedió el umbral de una desviación estándar. Esto puede indicar que el modelo tiene un buen desmpeño fuera de muestra y que tiene un ajuste que, en promedio, es adecuado para las observaciones de la muestra de prueba.

Figure 4.1: Influencias dentro de la muestra de prueba

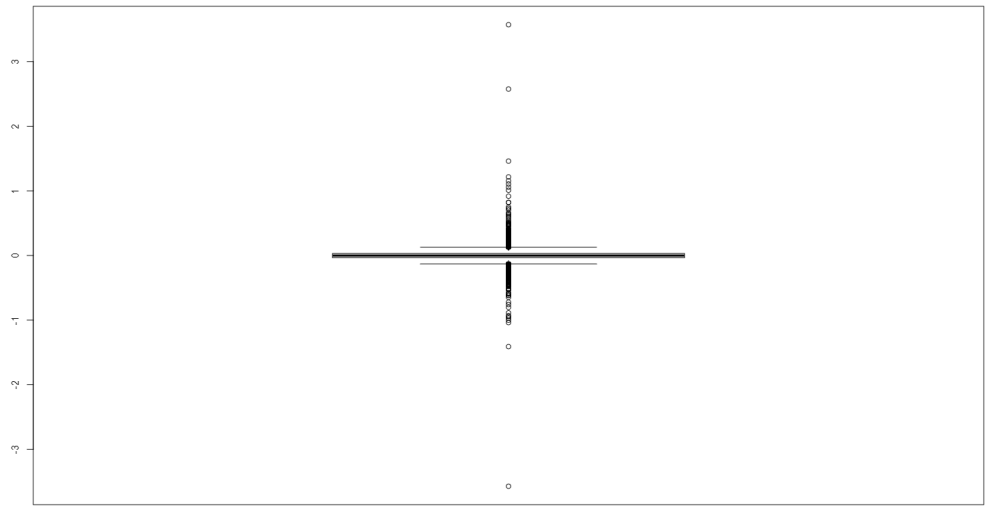


Fuente: (Dane,2008)

Sin embargo, realizar este mismo análisis dentro de la muestra de entrenamiento arroja otros resultados. Como se aprecia en la figura 4.2, hay algunas observaciones que tienen una influencia relativamente grande sobre el modelo estimado.

La existencia de estas observaciones puede estar asociada a personas que tienen ingresos anómalamente grandes. Por lo tanto, si se relacionan estos datos con el problema de subreporte que evidencia la DIAN, se podría sugerir que haya una auditoría fiscal especial para las personas identificadas en el gráfico que aparecen con ingresos demasiado altos.

Figure 4.2: Influencias dentro de la muestra de entrenamiento



Fuente: (Dane,2008)

Por último, se tomaron los dos modelos con los mejores desempeños (modelos 3 y 5) y se estableció el error de predicción a partir de un enfoque de validación cruzada estilo *LOOCV*. Los resultados se presentan a continuación:

Table 8: Desempeño del modelo 3

RMSE	MAE
0.466	0.315

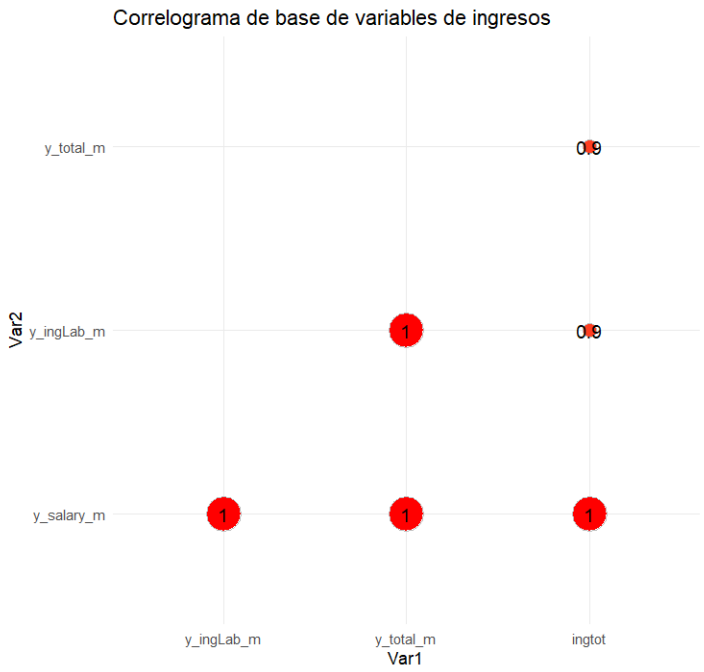
Table 9: Desempeño del modelo 5

RMSE	MAE
0.558	0.394

Los resultados que aparecen en estas tablas son similares a los que se había obtenido en la aproximación que usa el estadístico de influencia, ya que, como se recuerda de la figura 4.1, las distancias de las observaciones en la muestra de prueba no son grandes en promedio. De la misma manera, el RMSE y el MAE para los modelos 3 y 5 son pequeños, lo cual confirma que el ajuste de los datos fuera de muestra es adecuado.

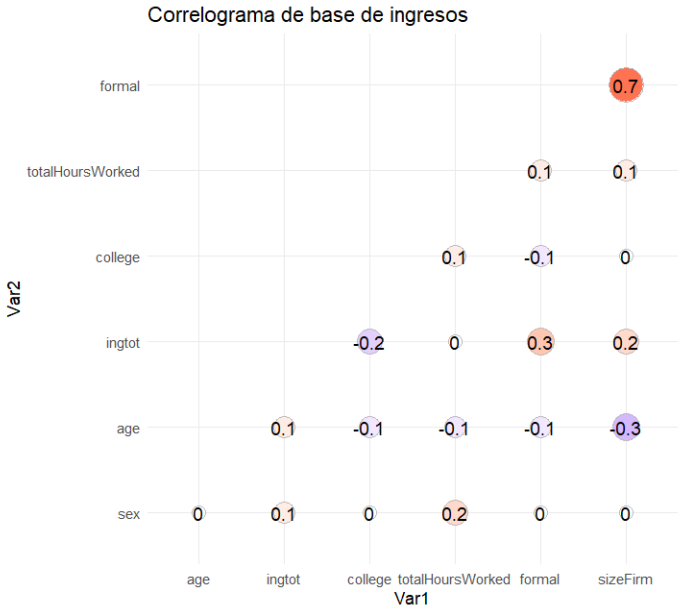
5 Appendix

Figure 5.1: Correlaciones entre variables de ingresos



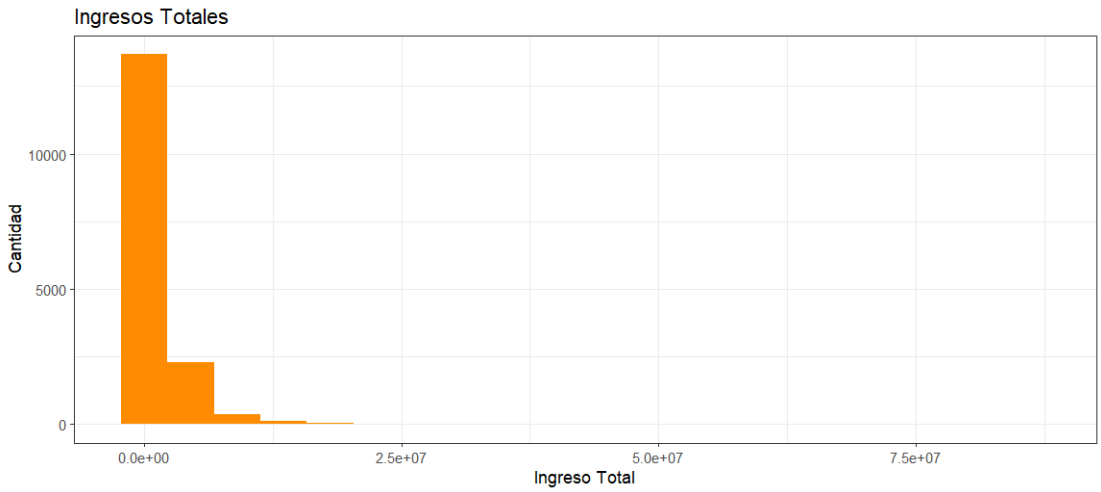
Note: Esta figura muestra la correlación entre las variables de ingresos o salarios.
Fuente: (Dane,2008)

Figure 5.2: Correlaciones panel de datos



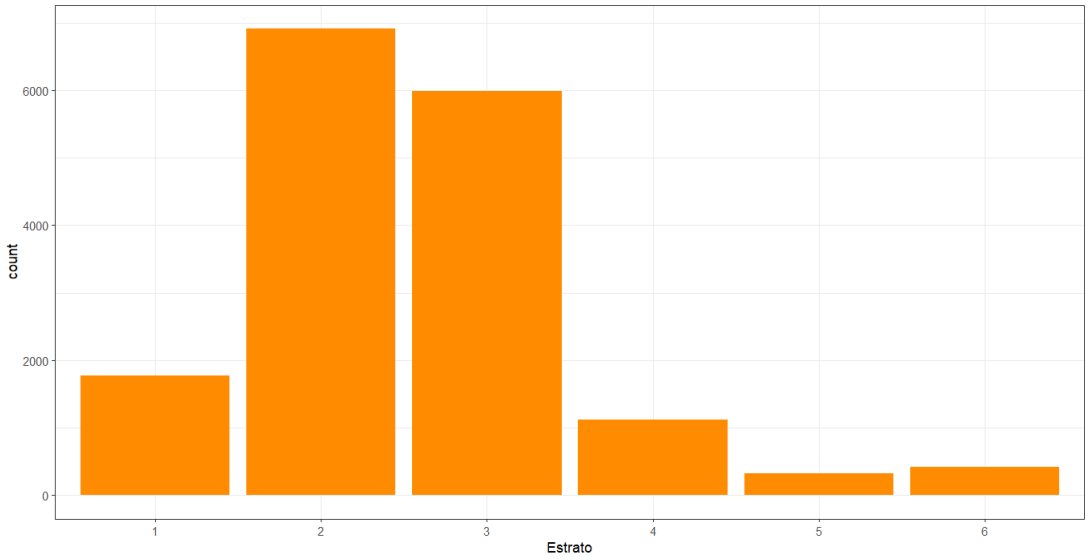
Note: Esta figura muestra correlación entre las variables filtradas de la base de datos.
Fuente: (Dane,2008)

Figure 5.3: Histograma de Ingresos Totales



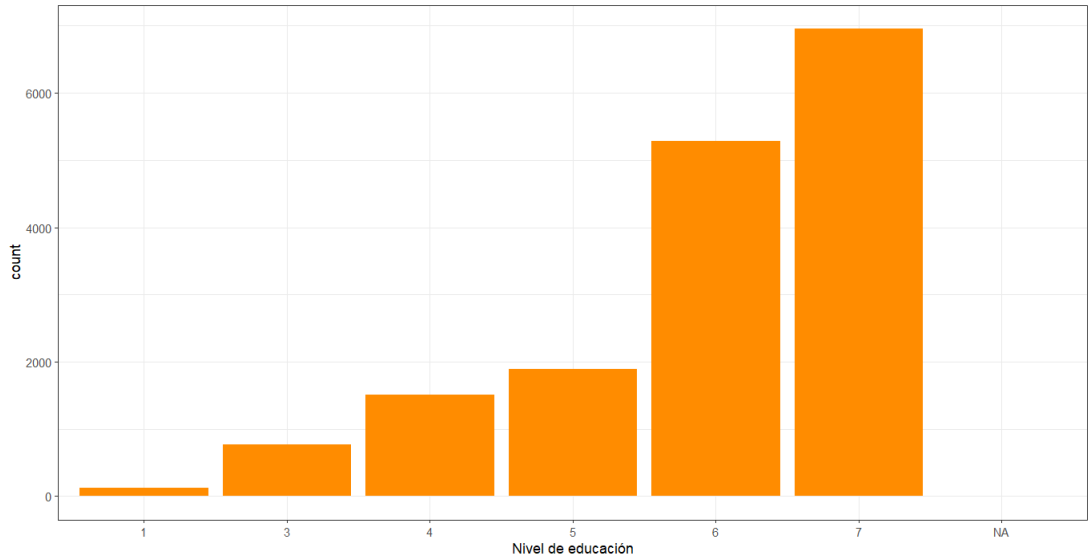
Note: Esta figura muestra la distribución de ingresos por observación.
Fuente: (Dane,2008)

Figure 5.4: Estrato



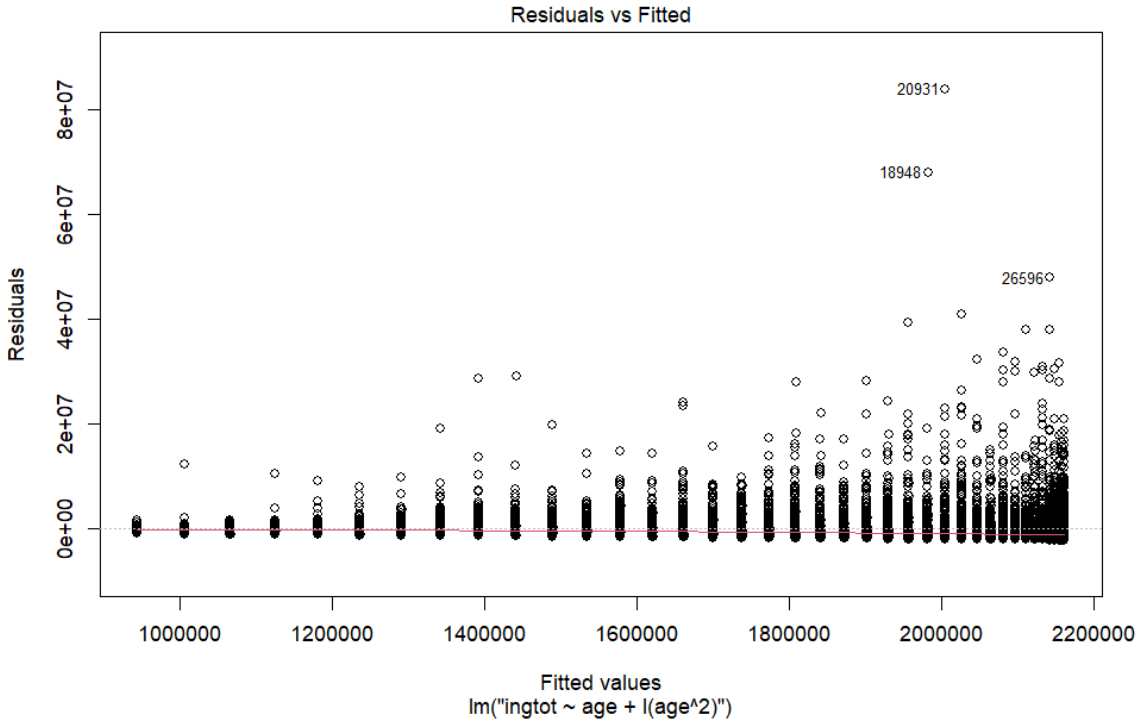
Note: Esta figura muestra la cantidad de encuestados por estrato.
Fuente: (Dane,2008)

Figure 5.5: Nivel Máximo de Educación



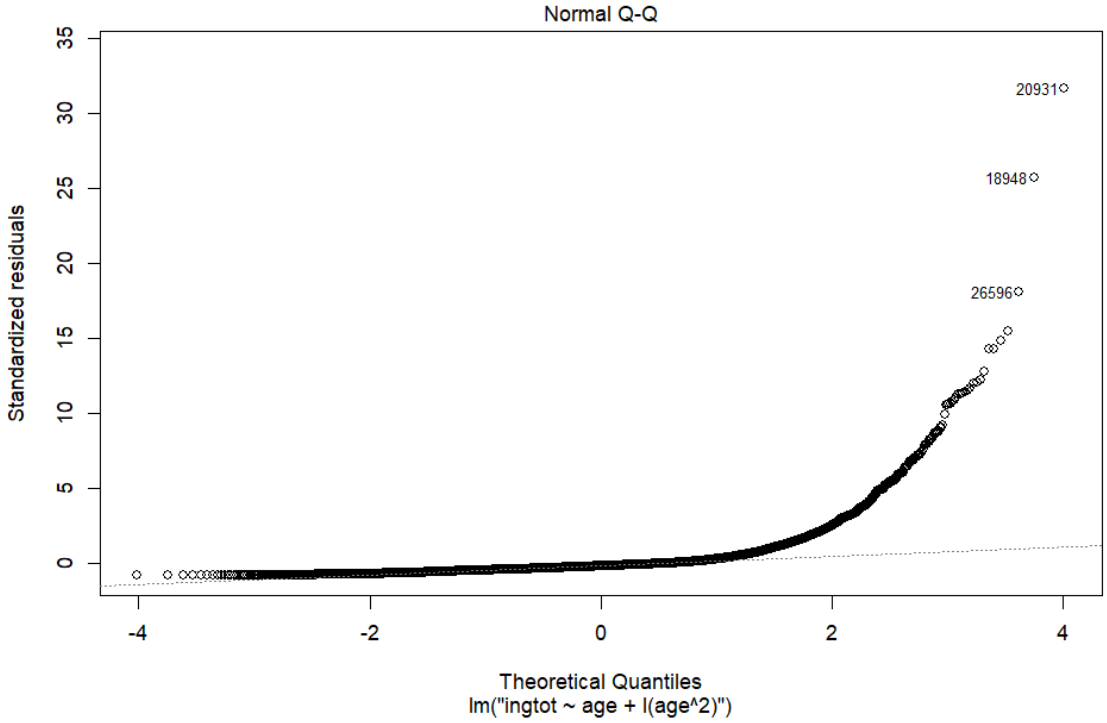
Note: Esta figura muestra la cantidad de encuestados por nivel educativo.
Fuente: (Dane,2008)

Figure 5.6: Ajuste en el Modelo de Edad



Note: Esta figura muestra los residuales contra los valores predichos de la muestra.
Fuente: (Dane,2008)

Figure 5.7: Ajuste en el Modelo de Edad



Note: Esta figura muestra los residuales estandarizados para comparar dos distribuciones de probabilidad al trazar sus cuantiles uno contra el otro.En este caso, lo ideal es que los puntos se acerquen a una recta diagonal.
Fuente: (Dane,2008)

Table 10: Gender Earning GAP Model 3

height	*Dependent variable:*
	Ln(income)
age	0.005***
	(0.0003)
totalHoursWorked	0.006***
	(0.0002)
female1	−0.041***
	(0.009)
sizeFirm2	−0.168***
	(0.015)
sizeFirm3	−0.104***
	(0.020)
sizeFirm4	0.019
	(0.018)
sizeFirm5	0.152***
	(0.016)
relab2	0.411***
	(0.021)
relab3	−0.548***
	(0.034)
relab4	0.379***
	(0.013)
relab5	0.305***
	(0.021)
relab6	0.681***
	(0.034)
relab7	0.670***
	(0.073)
relab8	−0.514
	(0.470)
relab9	0.638***
	(0.155)
maxEducLevel3	0.067
	(0.045)
maxEducLevel4	0.080*
	(0.043)
maxEducLevel5	0.053
	(0.043)
maxEducLevel6	0.071*
	(0.043)
maxEducLevel7	0.238***
	(0.043)
Constant	13.884***
	(0.118)
Observations	16,542
R^2	0.423
Adjusted R^2	0.420
Residual Std. Error	0.464 (df = 16442)
F Statistic	121.781*** (df = 99; 16442)
Note:	*p<0.1; **p<0.05; ***p<0.01