Problem set 1

1.a.i.

ii. La base de datos obtenida se trata de la encuesta GEIH del 2018 para la ciudad de Bogotá, esta es la gran encueta integrada de hogares y contiene información sobre las condiciones de empleo de las personas, además de sus fuentes de ingreso y características generales de la población. En cuanto al acceder a los datos y hacer el scrapping, hubo inconvenientes en la medida que la página ofrecida por el profesor no cargaba en Chrome, por lo que se utilizo el buscador Mozilla Firefox para acceder al url escondido en la sección de red (network) al momento de inspeccionar los chunks.

b.

Estadísticas descriptivas:

directorio	ingtot	sex	age	clase	college	cuentaPropia	depto	formal	maxEducLevel	oficio	relab	totalHoursWorked
Min. : 1	Min. : 0	Min. :1.00	Min. :18.00	Min. :1	Min. :1.000	Min. :1.000	Min. :1	Min. :1.000	Min. :1.000	Min. : 1.00	Min. :1.000	Min. : 1.0
1st Qu.: 2338	1st Qu.: 800000	1st Qu.: 1.00	1st Qu.: 28.00	1st Qu. :1	1st Qu.: 1.000	1st Qu.:1.000	1st Qu. :1	1st Qu.: 1.000	1st Qu.:4.000	1st Qu.: 24.00	1st Qu.: 1.000	1st Qu.: 40.0
Median :4636	Median : 1051160	Median :2.00	Median :38.00	Median :1	Median :1.000	Median :1.000	Median :1	Median :2.000	Median :5.000	Median :36.00	Median :1.000	Median : 48.0

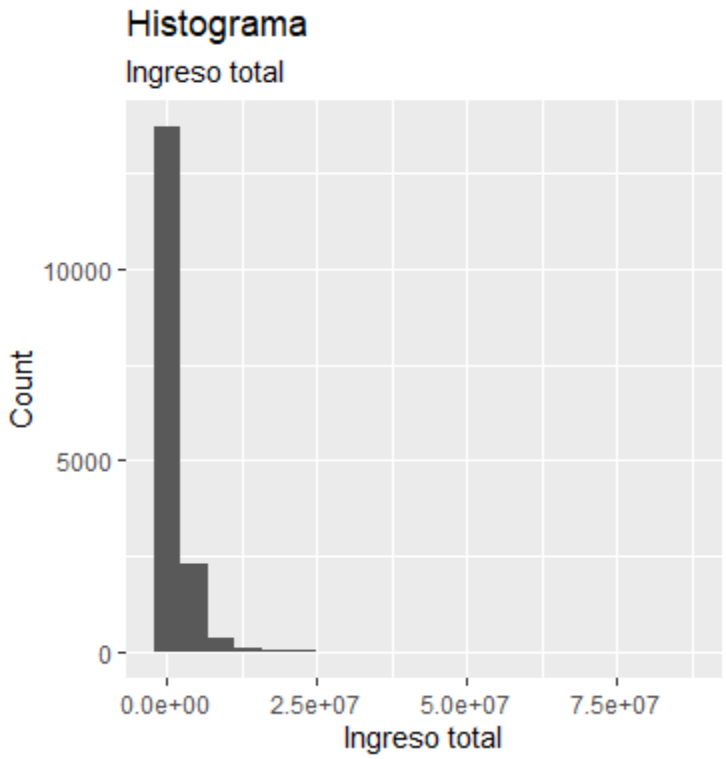
Mean :4624	Mean : 1769379	Mean :1.53	Mean :39.44	Mean :1	Mean :1.319	Mean :1.309	Mean :1	Mean :1.587	Mean :4.954	Mean :39.45	Mean :2.268	Mean : 47.4
3rd Qu.: 6933	3rd Qu.: 1723158	3rd Qu.: 2.00	3rd Qu.: 50.00	3rd Qu. :1	3rd Qu.: 2.000	3rd Qu.:2.000	3rd Qu. :1	3rd Qu.: 2.000	3rd Qu.:6.000	3rd Qu.: 52.00	3rd Qu.: 4.000	3rd Qu.: 50.0
Max. :9214	Max. :8583333 3	Max. :2.00	Max. :94.00	Max. :1	Max. :2.000	Max. :2.000	Max. :1	Max. :2.000	Max. :6.000	Max. :80.00	Max. :9.000	Max. :130.0
NA: 0	NA: 0	NA: 0	NA: 0	NA :0	NA: 0	NA: 0	NA :0	NA: 0	NA's :1	NA: 0	NA: 0	NA: 0

A partir de la tabla anterior se observan las principales características de las variables que se deciden utilizar para realizar el problema set. En primer lugar, la variable de ingtot hace referencia al ingreso total de la persona encuestada, se escogió esta como la medida apropiada del ingreso ya que recoge las diferentes formas y tipos de ingreso que una persona recibe, dando así una visión general de este rubro.

Continuando, esta variable nos indica que en promedio los bogotanos tienen un ingreso total de $1´769.379 de pesos colombianos teniendo como valor mínimo el 0 (se entiende por construcción de la base de datos que se presentan situaciones donde las personas son ocupados pero que no perciben ingresos por este trabajo) y como valor máximo $85´833.333. Continuando, nos encontramos con que las personas dentro de la muestra tienen en promedio 39.4 años de edad y que hay ligeramente más hombres encuestados que mujeres (media de 1.53, siendo el valor de 2 igual a hombre y de 1 igual a mujer). Junto a esto, se observa que en promedio las personas trabajaron 47.4 horas la semana anterior a ser encuestadas, lo cual se toma como el promedio trabajado cada semana.

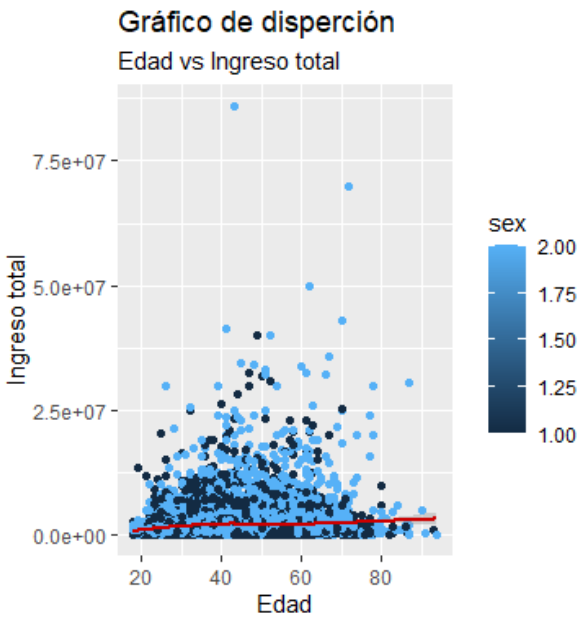
Además, se presentan variables que se considera pueden ser buenos predictores del ingreso como lo son: clase, la cual toma el valor de 1 si la persona encuestada reside en zona urbana y 0 si reside en zona rural (en las observaciones que se tienen solo hay residentes de la zona urbana); collegue que indica si la persona tiene educación terciaria (=1); cuentaPropia, que indica si la persona es autoempleada (=1) o no, formal, que toma el valor de 1 si la persona trabaja dentro del sector formal de la economía; maxEducLevel, que es el nivel de educación máximo que obtiene la persona; y oficio y relab que son, respectivamente, la ocupación y el tipo de ocupación de la persona.

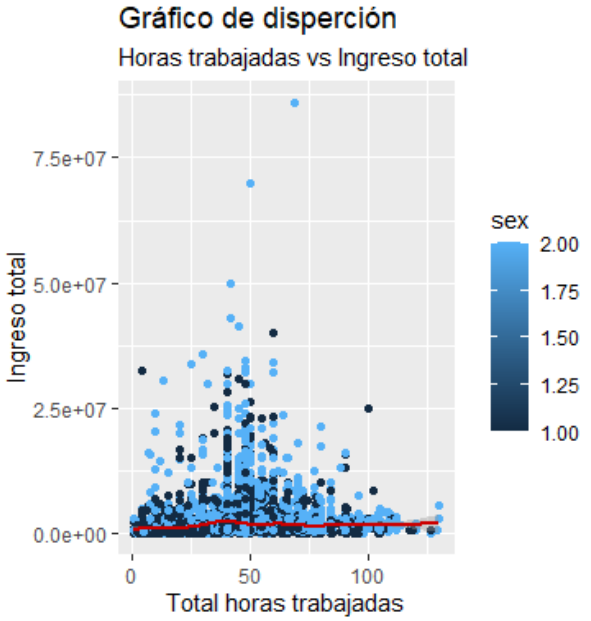
Según las siguientes gráficas, se observa que la distribución del ingreso total dentro de la muestra esta muy concentrada en la zona izquierda del plano, lo cual muestra que muchas personas ingresan montos mucho menores en relación con quienes expanden el histograma a la derecha, esto junto a los valores mínimo y máximo soportan la existencia de una fuerte inequidad en el ingreso.

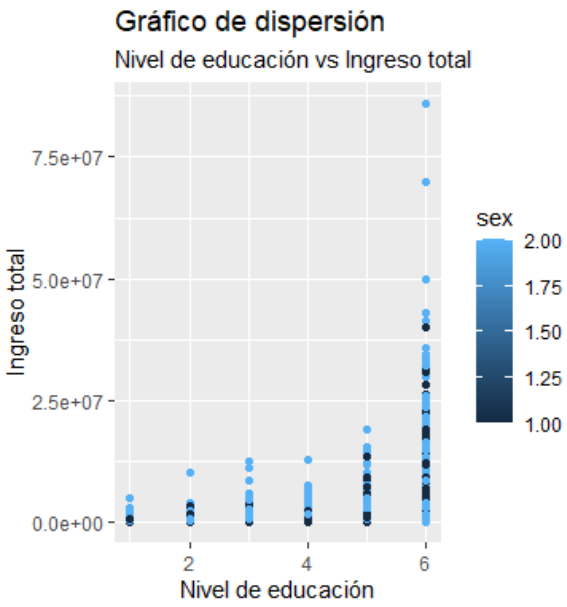


Respecto a los diagramas de dispersión, se observa como las horas trabajadas se correlacionan de forma positiva con el ingreso total, pero esta relación no es muy fuerte por lo que se puede decir que hasta cierto punto de las horas trabajas dan mayores ingresos. En cuanto a la relación de ingreso total y edad, estos también se relacionan positivamente, pero hay una tendencia más clara a que en los 50 años aproximadamente se encuentran las personas con mayores niveles de ingreso total. En cuanto al

histograma de nivel de educación máximo e ingreso total, se puede observar la clara tendencia de que a mayor sea el nivel de educación alcanzado, mayor es el ingreso total que se percibe. Por último, se debe mencionar que en los 3 gráficos de dispersión ya presentados se diferencian a los hombres de las mujeres por el color de los puntos, los hombres toman el color azul claro y las mujeres el azul oscuro, y con esta diferenciación es posible ver como los hombres tienden a tener mayores ingresos que las mujeres, así estas tengan el mismo nivel de educación, trabajen las mismas horas o tengan la misma edad.







2.

En primer lugar, la variable que se escoge para representar las earnings fue la referente al ingreso total. Esto debido a que creemos que ver el nivel agregado de todos los ingresos frente de la edad es más relevante puesto que, sean ingresos laborales o no laborales, a través del paso de los años la posibilidad de acceder a estos ingresos puede ser mayor. Por ejemplo, es más fácil comprar una casa y arrendarla en una edad mayor, por lo que, a pesar de ser un ingreso no laboral, tener más edad implica mayores posibilidades de tener ingresos.

Ahora bien, estos fueron los resultados del modelo:

Regresión Earning Age

==========

Variable dependiente:

---

ingtot

---

age 91,143.460***

(8,886.416)

age2 -799.261***

(102.852)

Constant -436,662.900**

(178,347.200)

Observations 16,542

R2 0.017

Adjusted R2 0.017

Residual Std. Error 2,652,732.000 (df = 16539)

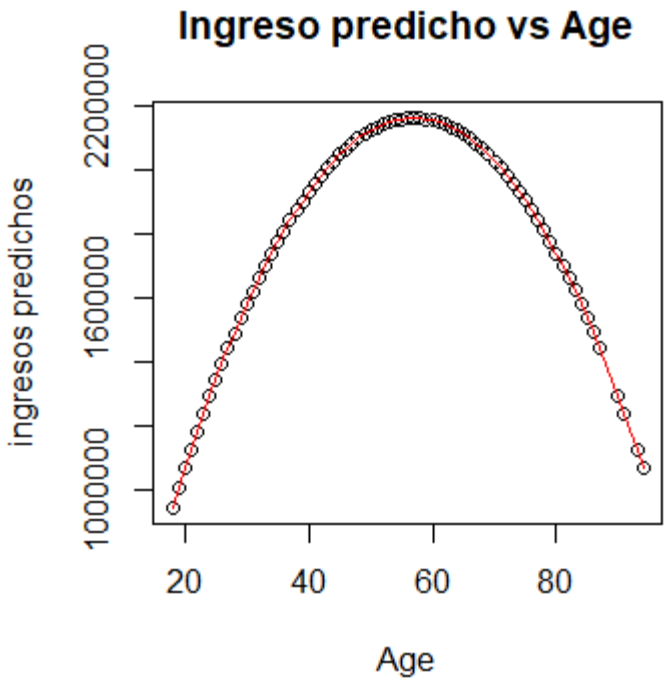
F Statistic 144.382*** (df = 2; 16539)

==========

Note: *p<0.1; **p<0.05; ***p<0.01

Como se puede observar en el modelo, tanto la variable age como la variable age^2 son estadísticamente significativas, no solo al 5%, sino a un nivel de significancia del 1%. Estos coeficientes en su interpretación no se pueden ver como cambios marginales puesto que, al ser una regresión cuadrática, los betas no nos muestran la marginalidad del cambio. Sin embargo, lo que se puede concluir, por medio de los signos y la significancia, es que la relación es creciente, pero decreciente en el margen hasta un punto de inflexión en donde se vuelve decreciente tanto en el margen como en la generalidad. Es decir, podemos estar hablando que hay un punto máximo donde las personas alcanzan sus ingresos más altos y que a partir de esa edad comienzan a disminuir. Se hace referencia a una parábola cóncava.

Asimismo, a pesar de encontrar significancia y una explicación lógica detrás, es importante hacer énfasis en el ajuste del modelo. En este caso tanto nuestro R^2 y nuestro R^2 ajustado son los mismos y nos dicen que el modelo tiene un poder de predicción del 1,7%. Es decir, a pesar de tener una gran significancia, el modelo pareciera ser que no es del todo bueno a la hora de predecir estimadores que representen la realidad. Aquí entra a la discusión el hecho de mejorar la R cuadrado, pero empeorar la varianza por medio de una especificación más grande en la modelación. Desde nuestro punto de vista y una primera impresión sobre los resultados, pensamos que si sería bueno sacrificar un poco de varianza para que el modelo tenga un poder de predicción mayor y poder hacer un análisis más correcto de la situación.



Viendo el plot anterior, podemos observar lo dicho anteriormente de la función cóncava con un máximo. Luego de un proceso de maximización, se encontró que la edad optima es de 57 años (exactitud de 57.01731) y que por cada año que pasa antes de esa se gana un plus de 28.103.87 pesos (luego de la edad optima se pierde). Es importante recalcar que el intervalo de confianza para esta elasticidad es (25.328, 30.872), es decir que el aumento (disminución) del salario cada año que pasa varía entre este rango en donde efectivamente se encuentra la elasticidad calculada. Por tanto, podemos ver que la varianza no es grande, que, si bien es cierto que la variación de la elasticidad puede cambiar nuestro nivel óptimo de edad, luego del proceso de bootstraping, vemos que estos cambios realmente serían mínimos y que la edad 57 es con proximidad acertada. Por último, es claro que nuestra visión inicial de los resultados no era precisamente correcta porque si bien es cierto que no se tenía un poder de predicción fuerte; verificando luego del boot no es necesario sacrificar varianza dado que los resultados parecen tener buena precisión.

3. a.

Regresión log-Earnings Mujer

==========

	Dependent variable:
	log_ingtot
mujer	-0.193***
	(0.014)
Constant	14.064***
	(0.009)
Observations	16,277
R2	0.012
Adjusted R2	0.012
Residual Std. Error	0.872 (df = 16275)
F Statistic	198.856*** (df = 1; 16275)
Note:	*p<0.1; **p<0.05; ***p<0.01

La brecha de genero dentro del anterior modelo es significativa a un nivel de significancia del 1% lo cual implica que ser mujer esta correlacionado de forma estrecha con tener menores ingresos, ahora, debido a la forma en que esta identificado el modelo (solo un regresor) no es posible decir que se presenta causalidad pero si es posible argumentar que a partir de la muestra el factor de ser mujer reduce considerablemente los ingresos que se perciben, soportando así el argumento de la brecha de género.

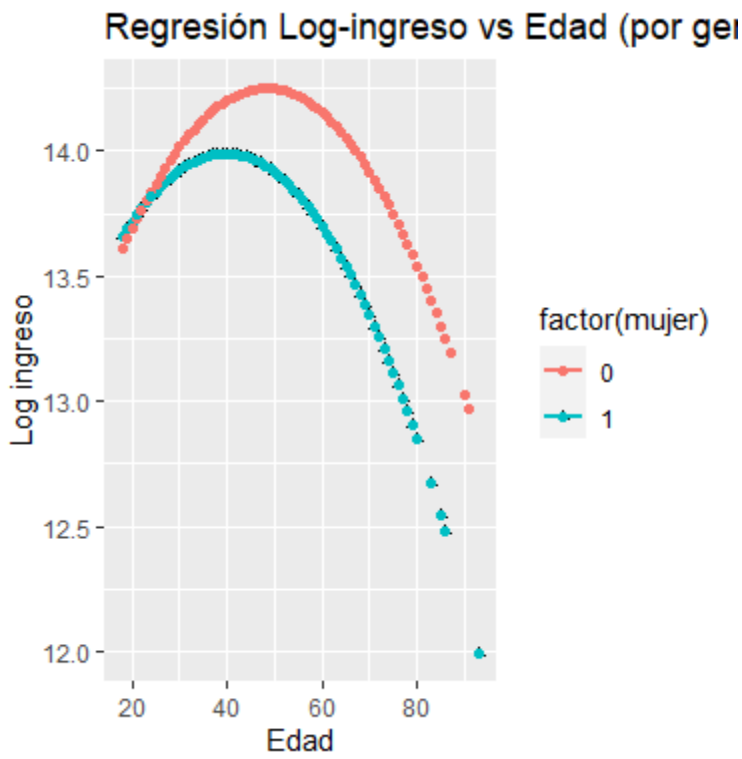
b)

Regresión log-Earnings Edad Mujer

	Dependent variable:
	log_ingtot
age	0.068***
	(0.003)
age2	-0.001***
	(0.00003)
mujer	0.266***
	(0.042)
age_mujer	-0.012***
	(0.001)
Constant	12.615***
	(0.062)
Observations	16,277
R2	0.046
Adjusted R2	0.045
Residual Std. Error	0.857 (df = 16272)
F Statistic	194.501*** (df = 4; 16272)

Note: *p<0.1; **p<0.05; ***p<0.01

A partir de la anterior regresión, se observa como la edad y el factor de ser mujer son estadísticamente significativos a un nivel de 1% al momento de percibir el ingreso, lo cual implica que son variables relevantes para tratar de explicar los ingresos totales de una persona. Ahora, se puede observar que la variable age2 la cual es edad al cuadrado tiene signo negativo al igual que la interacción entre edad y mujer, lo cual da a entender que en algún punto la edad va a influir negativamente en los ingresos, y con una magnitud mayor, en los ingresos totales de las mujeres. Para entender esto de forma más clara, se presenta la siguiente gráfica:



A partir de esto, se puede observar claramente como las mujeres a pesar de que parten (aproximadamente) del mismo punto que los hombres en cuanto a los ingresos, el comportamiento de su función llega a un máximo menor que el de los hombres. Además, también es importante mencionar que las mujeres llegan a la edad en la que optimizan su ingreso total antes que los hombres, específicamente esta edad para las mujeres son los 39.69 años mientras que para los hombres son los 48.24 años, mostrando así que los hombres tienen 9 años más de vida en los cuales su ingreso continúa aumentando en comparación con las mujeres.

Finalmente, al construir los intervalos de confianza por género se llega a que por cada año que aumente, con una confianza del 95% las mujeres varían su ingreso total en el intervalo de (-0.09%, 0.0.23%)

mientras que los hombres con el mismo nivel de confianza varían su ingreso en el intervalo (-0.16%, 0.0.19%).

c.

Al incluir la variable oficios, la cual contiene información sobre el oficio al que se dedica cada individuo dentro de la base de datos, se busca controlar por el tipo de ocupación de las personas. La regresión "larga" y la FWL se estimaron y los resultados son los siguientes:

Comparación Long y FWL

==========

==

Dependent variable:

---

log_ingtot r1

(1) (2)

---

mujer -0.353***

(0.013)

oficio -0.015***

(0.0003)

r2 -0.353***

(0.013)

Constant 14.739***

(0.016)

Observations	16,277	16,277
R2	0.150	0.043
Adjusted R2	0.149	0.043
Residual Std. Error	0.809 (df = 16274)	0.809 (df = 16276)
F Statistic	1,431.446*** (df = 2; 16274)	725.605*** (df = 1; 16276)

==========

==

Note: *p<0.1; **p<0.05; ***p<0.01

Según lo observado en la tabla, lo primero que se tiene que mencionar es que se siguió utilizando la misma variable de ingresos totales, esto debido a que se quiere mantener una misma línea con los puntos anteriores y con las estimaciones anteriores, a pesar de que se es consciente que al regresar por la variable oficio, esta variable influye directamente al ingreso que se obtiene a través del trabajo. Por lo tanto, al decidir mantener la variable de ingresos totales el efecto de la variable oficio puede verse mermado, pero aun así este es un costo que se esta dispuesto a aceptar con el fin de no modificar la forma de accionar que se tuvo desde un principio en el problema set.

Continuando, al hacer la comparación entre los estimadores obtenidos por la regresión “larga” y el FWL se observa que la estimación de la variable mujer es igual en ambos modelos, lo cual nos indica que el procedimiento se realizo adecuadamente. En cuanto a su interpretación, manteniendo lo demás constante el ser mujer implica obtener tener menores ingresos totales en comparación con ser hombre, específicamente el ser mujer disminuye los ingresos totales percibidos en 35.3%.

Ahora, para la estimación no se quisieron agregar las variables edad y edad al cuadrado debido a que se intentó evitar caer en overfitting del modelo, pero con el anterior punto se considera prudente mantener las mismas edades óptimas tanto para hombres como para mujeres ya que al calcular los óptimos la variable oficio no afecta el desarrollo matemático, por lo que se deja que la edad óptima para mujeres son los 39.69 y para los hombres son los 48.24 años.

Finalmente, entre la brecha condicional e incondicional se puede decir que si bien es claro que hay problemas de discriminación se considera que los cambios en los coeficientes pueden deberse, en parte, a problemas de selección debido a la forma en que se identifica el modelo. Aun así, en todas las estimaciones realizadas la variable mujer a sido significativa y de signo negativo, mostrando así que el problema de discriminación por género existe.