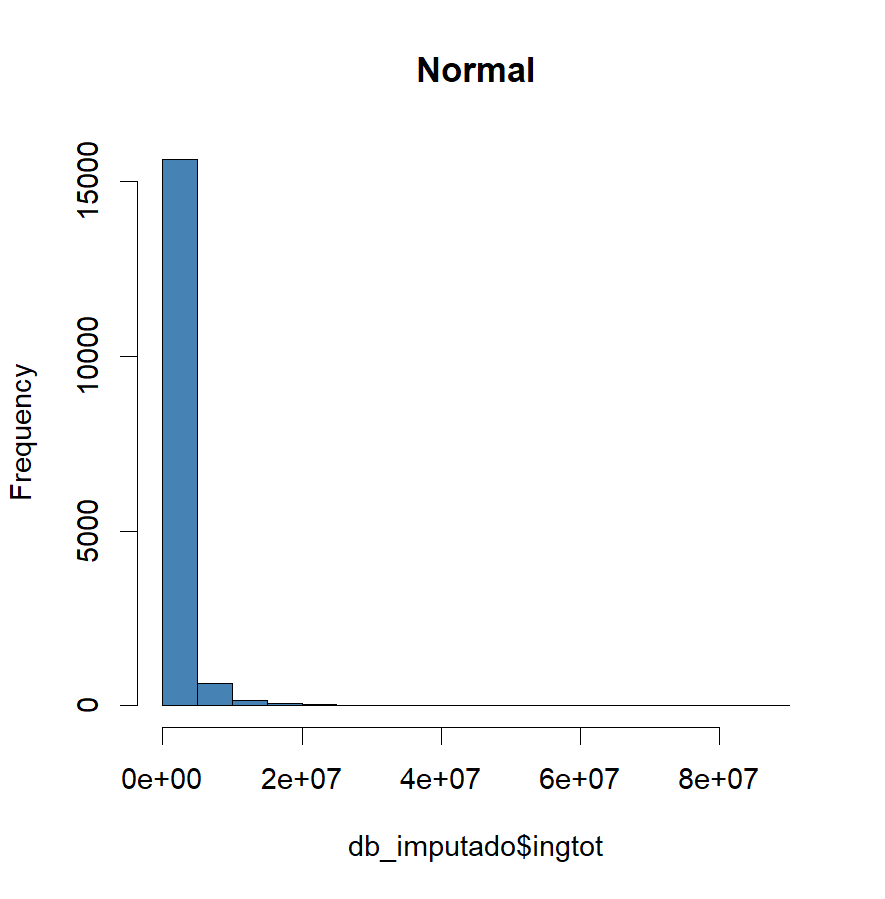
**3. Age-earnings profile.**

A great deal of evidence in Labor economics suggests that the typical worker’s age-earnings profile has a predictable path: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50.

• **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.**

La variable escogida para representar los ingresos totales de los trabajadores fue “ingtot” y representa el ingreso total de los mismos. En primer lugar, se escogió esta variable ya que, a comparación de las demás variables relacionadas con los ingresos, representa los ingresos totales de los individuos pues otras variables representan ingresos parciales o fraccionados del ingreso total.

Chart, scatter chart

Description automatically generated

|  |
| --- |
| **Estadísticas descriptivas variable ingtot**  Min. 1st Qu. Median Mean 3rd Qu. Max.  15.000 800.000 1.051.160 1.769.379 1.723.158 85.833.333 |

• **Based on this estimate using OLS the age-earnings profile equation:**

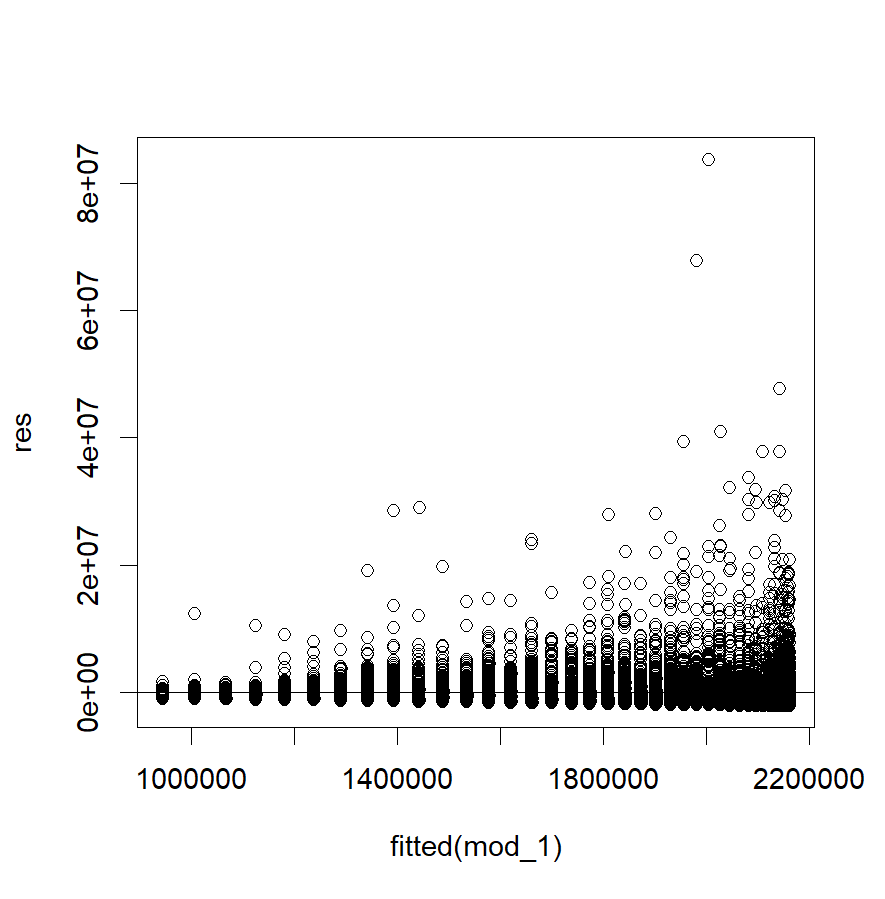
**Income = β1 + β2Age+ β3Age2 + u**

|  |
| --- |
| **Resultados regresión ingresos-edad**  Dependent variable: 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**

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0.01. Así mismo, El estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global. Se puede apreciar que uno de los coeficientes relacionados con la edad es positivo mientras que el otro es negativo, lo cual es coherente con la forma funcional cuadrática del modelo.

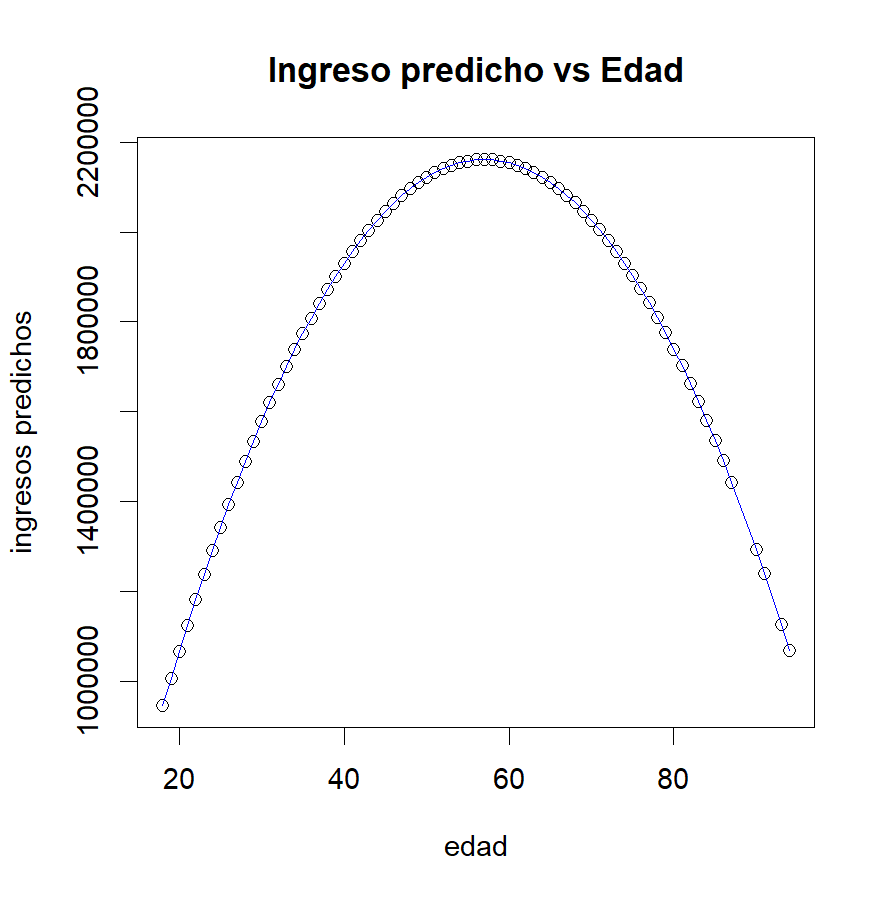
**• How good is this model in sample fit?**

****

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo. En esta gráfica se puede observar que el modelo ***Income = β1 + β2Age+ β3Age2 + u*** no se ajusta del todo bien a la muestra puesto que existen diferencias entre los residuales y valores predichos diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0,017, es decir, que las variables explicativas explican en 1,7% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas y mala especificación del modelo.

**• Plot the predicted age-earnings profile implied by the above equation.**

****

La gráfica superior nos muestra los valores predichos de los ingresos totales de los individuos de Bogotá frente a la edad de los mismos. Se puede analizar en la gráfica que la relación entre valores predichos de los ingresos y la edad es coherente con la forma funcional cuadrática del modelo, en la cual se espera que el ingreso total alcance un máximo a cierta edad y empiece a disminuir paulatinamente a medida que la edad aumenta.

**• What is the “peak age” suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.**

Para poder calcular “la edad pico” de ingresos en nuestro modelo se tuvo que encontrar la elasticidad ingreso-edad a partir de derivar nuestra función inicial e igualar a cero en busca de encontrar la edad máxima. A continuación, se muestra el procedimiento realizado:

=

De acuerdo con este procedimiento, “la edad pico” de ingresos en nuestro modelo es de 57 años estando acorde con lo que expone la literatura de una edad pico de ingresos cerca a los 50 años.

Ahora bien, dado que la forma funcional del modelo es cuadrática, los coeficientes estimados con respecto a la edad no se pueden interpretar directamente y es necesario encontrar el efecto marginal ingreso-edad por medio de derivar nuestro modelo. Así mismo, se va a utilizar la edad media de la muestra para poder encontrar este efecto marginal

Este efecto marginal se puede interpretar como que, manteniendo todas las demás variables constantes, en promedio un año más de edad implica que el ingreso aumente en 28.103,7 pesos.

Para encontrar los errores estándar asociados al efecto marginal ingreso-edad e intervalos de confianza, usamos el procedimiento Bootstrap y obtuvimos los siguientes resultados:

|  |
| --- |
| **Bootstrap Statistics :**  **original bias std. error**  **t1\* 28103.87 35.86716 1491.037** |

Estos resultados están acordes a nuestras estimaciones, puesto que se puede apreciar que el efecto marginal por medio de Bootstrap es el mismo que encontramos con anterioridad (28.103,87) y podemos ver que el error estándar es de 1491.037.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad como se puede observar a continuación:

Este intervalo de confianza se puede interpretar como en promedio un año más de edad de un individuo de Bogotá implica que sus ingresos aumenten en el rango del intervalo .

**4. The earnings GAP**

Most empirical economic studies are interested in a single low dimensional parameter, but determining that parameter may require estimating additional “nuisance” parameters to estimate this coefficient consistently and avoid omitted variables bias. Policymakers have long been concerned with the gender earnings gap.

**• Estimate the unconditional earnings gap: log(Income) = β1 + β2Female + u**

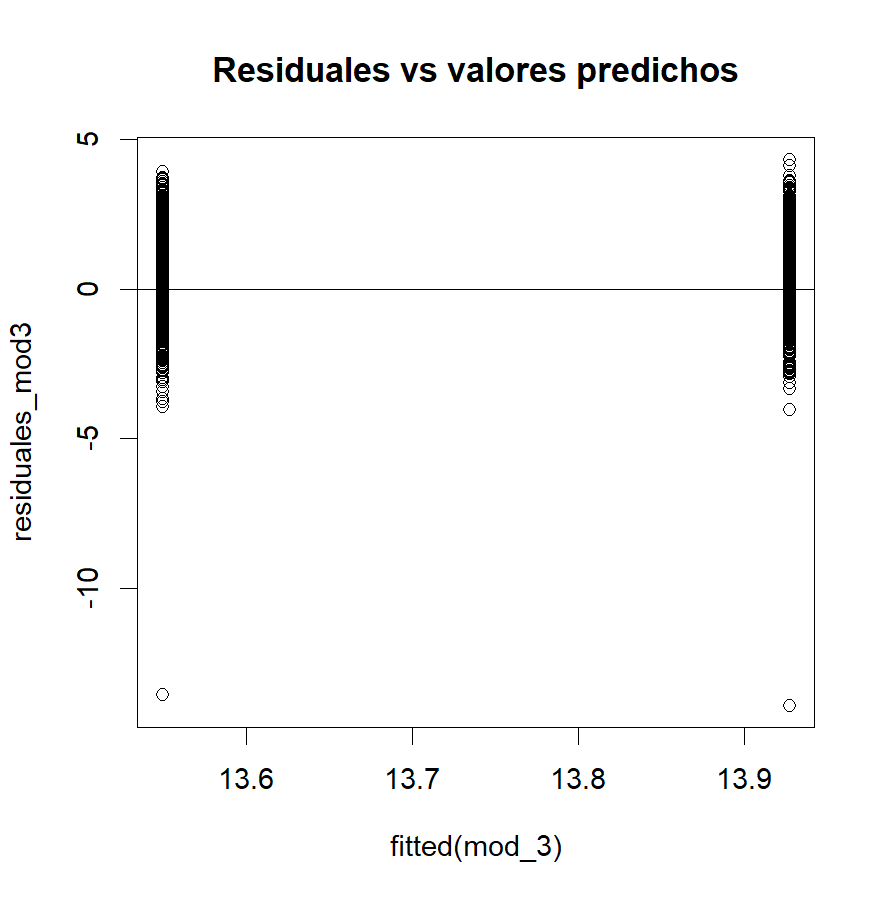
|  |
| --- |
| **Regresión Log ingresos-mujer**    Dependent variable: log\_ingreso  -----------------------------------------------  **mujer -0.378\*\*\***  **(0.030)**    **Constant 13.927\*\*\***  **(0.021)**  -----------------------------------------------  Observations 16,542  R2 0.009  Adjusted R2 0.009  Residual Std. Error 1.950 (df = 16540)  F Statistic 154.943\*\*\* (df = 1; 16540) |

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

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0.01. Así mismo, El estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global.

**• How should we interpret the β2 coefficient? How good is this model in sample fit?**

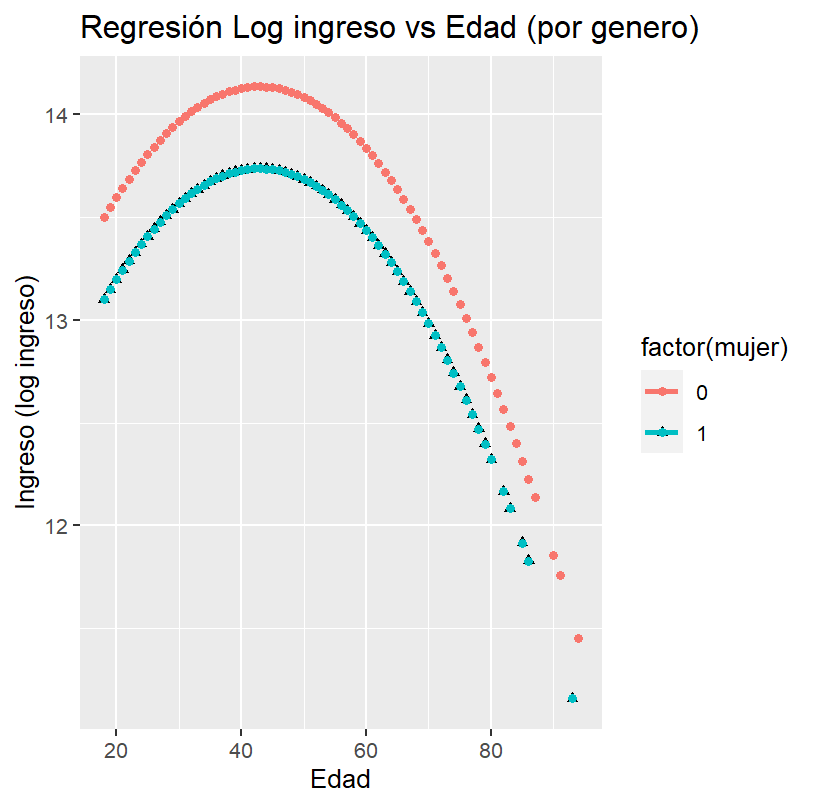
El coeficiente asociado a la variable dummy de mujer (toma el valor de 1 si el individuo es mujer y 0 si es hombre) se puede interpretar como que, manteniendo todas las demás variables constantes, en promedio las mujeres ganan 37,8% menos en ingresos que a comparación de los hombres.

****

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo ***log(Income) = β1 + β2Female + u*** . En esta gráfica se puede observar que el modelo no se ajusta bien a la muestra puesto que existen grandes diferencias entre los residuales y valores predichos diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0.009, es decir, que las variables explicativas explican en 0,9% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas y mala especificación del modelo.

**• Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes?**

****

De acuerdo con la gráfica del logaritmo natural de los ingresos predichos contra la edad de los individuos por género podemos analizar de que definitivamente el sexo del individuo afecta el ingreso de acuerdo con su edad. Específicamente, la gráfica ilustra que los valores predichos de los ingresos-edad es diferente entre hombres y mujeres, pues las pendientes e interceptos de las líneas de regresión difieren. En el caso de los hombres, la linea de regresión (línea rosada) tiene una pendiente y un intercepto mayor que el de las mujeres (linea azul), estando esto relacionado con que los ingresos que reciben los hombres son mayores a los de las mujeres.

**• What is the implied “peak age” by gender? Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?**

|  |
| --- |
| **Regresión lineal ingresos-edad-genero-edad\*genero**    Dependent variable: log\_ingreso  -----------------------------------------------  age 0.098\*\*\*  (0.007)    age2 -0.001\*\*\*  (0.0001)    mujer 0.244\*\*\*  (0.094)    mujer\_age -0.016\*\*\*  (0.002)    Constant 11.908\*\*\*  (0.138)  -----------------------------------------------  Observations 16,542  R2 0.023  Adjusted R2 0.023  Residual Std. Error 1.936 (df = 16537)  F Statistic 99.240\*\*\* (df = 4; 16537) |

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

Para poder calcular “la edad pico” de ingresos se construyó un modelo que contuviera la interacción entre genero y edad con el fin de encontrar la elasticidad ingreso-edad por género a partir de derivar nuestra función e igualar a cero en busca de encontrar la edad máxima. A continuación, se muestra el procedimiento realizado:

**En el caso de las mujeres:**

=

**En el caso de los hombres:**

=

Los picos de edad de ingresos para mujeres esa los 38,5 años mientras que el de los hombres es a los 46,2 años. Estos resultados están acordes con nuestra gráfica de los ingresos predichos con respecto a la edad por género y, así mismo, ilustra que los hombres tienen una mayor fuente de ingresos por más tiempo a comparación de las mujeres.

Para encontrar los errores estándar asociados al efecto marginal ingreso-edad por género e intervalos de confianza, usamos el procedimiento Bootstrap y obtuvimos los siguientes resultados:

**En el caso de las mujeres: (se utilizó la edad promedio de las mujeres en nuestra muestra)**

|  |
| --- |
| **Bootstrap Statistics :**  **original bias std. error**  **t1\* -0.00143758 -5.376217e-05 0.002163139** |

Estos resultados nos muestran que el efecto marginal ingreso-edad para las mujeres es de -0,00143, es decir, manteniendo las demás variables explicativas constantes, en promedio con un año más de edad el ingreso de las mujeres disminuye en 0,143% con respecto a los hombres.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad para las mujeres como se puede observar a continuación:

Este intervalo de confianza se puede interpretar como en promedio un año más de edad en las mujeres de Bogotá implica que sus ingresos cambien porcentualmente en el rango del intervalo frente a los hombres.

**En el caso de los hombres: (se utilizó la edad promedio de los hombres en nuestra muestra)**

|  |
| --- |
| **Bootstrap Statistics :**  original bias std. error  t1\* 0.01414982 4.626327e-06 0.001515723 |

Estos resultados nos muestran que el efecto marginal ingreso-edad para los hombres es de 0,00141, es decir, manteniendo las demás variables explicativas constantes, en promedio con un año más de edad el ingreso de los hombres aumenta en 1,41% con respecto a las mujeres.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad para los hombres como se puede observar a continuación:

Este intervalo de confianza se puede interpretar como en promedio un año más de edad en los hombres de Bogotá implica que sus ingresos cambien porcentualmente en el rango del intervalo frente a las mujeres.

Los intervalos de confianza entre hombres y mujeres no se superponen, siendo muestra de esto que en promedio la diferencia de ingresos al aumentar un año de edad entre hombres y mujeres es significativa.

• Equal Pay for Equal Work? A common slogan is “equal pay for equal work”. One way to interpret this is that for employees with similar worker and job characteristics, no gender earnings gap should exist. Estimate a conditional earnings gap that incorporates control variables such as similar worker and job characteristics (X).

1. **Estimate the conditional earnings gap log(Income) = β1 + β2Female + θX + u**

|  |
| --- |
| **Regresión Log ingresos-mujer-oficio**  Dependent variable: log\_ingreso  -----------------------------------------------  **mujer -0.358\*\*\***  **(0.034)**  **.**  **.**  **.**  **.**  **.**  **Constant 15.208\*\*\***  **(0.423)**  -----------------------------------------------  Observations 16,542  R2 0.117  Adjusted R2 0.113  Residual Std. Error 1.844 (df = 16461)  F Statistic 27.391\*\*\* (df = 80; 16461) |

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

El modelo planteado para este punto fue el siguiente:

La variable de control seleccionada fue “oficio”, la cual representa el trabajo en el que se estaba desenvolviendo cada individuo en el momento de realizar la encuesta.

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0.01. Así mismo, El estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global.

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

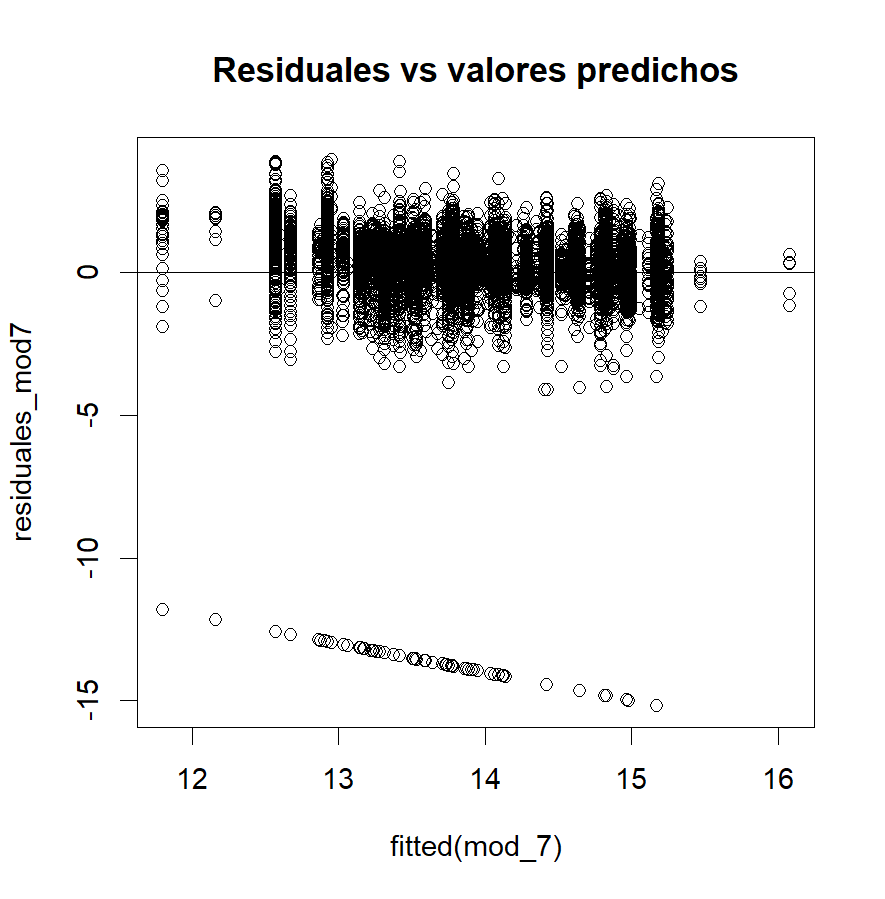
Para aplicar el Teorema FWL es necesario encontrar los residuales de la regresión entre el logaritmo del ingreso frente a la variable explicativa Oficio y, así mismo, se deben encontrar los residuales de la regresión entre el genero del individuo (female) y el Oficio.

Con la identificación de los residuales, realizamos la siguiente regresión para estimar el coeficiente B2.

Este coeficiente estimado por medio del teorema FWL es idéntico al estimado por medio de OLS.

**(c) 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 se interpreta como que, manteniendo las demás variables explicativas constantes, el ingreso de las mujeres es en promedio 35,8% menor en comparación al ingreso de los hombres. Si comparamos el efecto marginal ingreso-genero de este modelo (-35,8%) con interacción frente al modelo sin interacción (-37,8%) podemos analizar que la brecha disminuye ligeramente. Esto podría explicarse como que la brecha de ingresos entre genero si puede relacionarse con un problema de selección de las variables explicativas del modelo, puesto que puede que hayan variables omitidas. Sin embargo, estas brechas de ingresos entre hombres y mujeres siguen siendo un problema de discriminación porque en todas las regresiones realizadas las mujeres tienen un menor salario que los hombres.

****

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo. En esta gráfica se puede observar que el modelo ***log***(***Income) = β1 + β2Female+ β3Oficio + u*** no se ajusta del todo bien a la muestra puesto que existen diferencias entre los residuales y valores predichos diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0,117, es decir, que las variables explicativas explican en 11,7% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas, mala especificación del modelo y datos outliers.