Luisa Cuellar - 201613942

Isabella Riveros - 201923015

**Problem Set 1: Predicting Income**

1. The main objective is to construct a model of individual income:

* Describe the data set briefly, the process of acquiring the data, and if there any restrictions to accessing/scraping these data.
* descriptive analysis of the variables used in the problem set.
* Contains multiple income measures. Choose those that best suit your analysis.
* There are many observations with missing data. I leave it to you to find a way to handle these missing data.

Para el modelo se escogió la variable de *ingtot* (Ingreso total), pues es la que mejor se ajusta a lo que queremos analizar. En la base de datos había varias variables que median el ingreso de varias maneras. Se tenían variables que indicaban los subsidios o las primas, pero estas no eran ideales pues muestran solo una parte de todo el análisis. Por otro lado, habían otras variables como *impa*  o *isa* que solo mostraban el ingreso monetario de alguna de las dos actividades antes de imputación dejando de lado la otra actividad económica, lo cual sesgaría el análisis. Otras variables coo *y\_salary\_m* o *y\_ingLab\_m* solo observan los ingresos que se dan por el salario, dejando de lado los demás que también son importantes para el estudio. En el caso de la variable *p6500* esta solo toma en cuenta los ingresos del mes pasado en su empleo, sin contar subsidios u otro empleo. Y, por último, la variable *ingtotes* que es el ingreso total imputado pues deja de lado otros tipos de ingresos que no fueron imputados.

En el caso de los *missing values* ***encontramos*** uno en la variable *maxEducLevel,* y debido a que solo era una observación que no iba a afectar de forma significativa el resultado lo reemplazamos por 1. En la variable *regSalud* encontramos 1420 *missing values* , para estos revisamos las observaciones y nos dimos cuenta que estas observaciones no cotizan a pensión por lo cual asumimos que tampoco cotizaban a salud por lo cual los reemplazamos por 0 (categoría de no cotiza). Y en la variable *hoursWorkActualSecondJob* , encontramos 15980 *misiisng values*  los cuales reemplazamos por 0, pues supusimos que la información estaba vacía debido a que las personas no tenían un segundo trabajo.

2. In this subsection we are going to estimate the Age-earnings profile profile for the individuals in this sample:

• A discussion of the variable chosen as a measure of earnings.

• An interpretation of the coefficients and it’s significance.

• A discussion of the model’s in sample fit.

• A plot of the estimated age-earnings profile implied by the above equation. Including a discussion of the “peak age” with it’s respective confidence intervals. (Note: Use bootstrap to construct the confidence intervals.)

La variable ingresos totales se eligió como la variable que representa las ganancias, dado que esta agrupa todas las fuentes de ingresos del individuo durante el mes. Los resultados del modelo muestran que, tal como lo predice la teoría el ingreso aumenta con la edad (dado que el signo del β1 es positivo), pero lo hace a un ritmo decreciente (dado que el signo del β2 es negativo).

REGRESION AGE-AGE2

Así las cosas, el modelo estima que, partiendo desde los 18 años, por cada año extra de edad el ingreso de las personas aumentará en $91.143.5. Por otro lado, para determinar la edad máxima de ingresos se deriva con respecto a la edad y se iguala a cero. Esto nos da que los máximos ingresos se alcanzan cuando se llega a los 57 años (peak age).

Para el caso de edad (< 2e-16 \*\*\*) y edad al cuadrado (8.15e-15 \*\*\*), los p-value son estadísticamente significativos, mientras que el R-cuadrado es 0.01716 lo cual indica que solamente el 1,7 % del cambio en el ingreso es explicado por el aumento en la edad.

Este último dato puede dar un indicio de que la edad no es la medida más precisa a la hora de intentar explicar el ingreso de los trabajadores (el ajuste del modelo no es el mejor), por lo cual se agregan otras variables para tener modelos cuyos resultados sean más contundentes.

GRAFICO PEAK AGE CON INTERVALOS

El ingreso total predicho en la *peak age* es de 2.161.715. Se puede ver que el y\_predicho para la edad empieza a distanciarse cada vez más del que se estimó inicialmente a partir del peak age (57 años). Justo después de este, hay mucha más variación en el ingreso total de las personas. La varianza depende de age, por lo que a medida que aumenta la edad, también aumenta el intervalo de confianza.

Al llegar al punto del peak age se explota y hay mayor variabilidad del ingreso total después de esta edad. Esto ya se podía preveer pues, al hacer el análisis inicial de las estadísticas descriptivas, la gráfica de puntos del ingreso total por edad mostraba que la mayoría de la distribución estaba en edades más jóvenes, con ingresos más bajos y medianos, mientras que en las edades más grandes la dispersión aumenta considerablemente.

3. Begin by estimating and discussing the unconditional earnings gap:

* A discussion of the variable chosen as a measure of Income, if it is the same or different from the previous point.
* An interpretation of the“Female” coefficients, a comparison between the models, and the in-sample fit.

REGRESION FEMALE

El modelo, el cual solamente tiene en cuenta el sexo de la persona, estima que los hombres tienen un salario promedio 37,8 % más alto que en el caso de las mujeres.

Para el caso de sexo (< 2e-16 \*\*\*) el p-value son estadísticamente significativos, mientras que el R-cuadrado es 0.009221 lo cual indica que menos del 1 % del cambio en el ingreso es explicado por el sexo.

* Next, estimate, plot, and discuss the predicted age-earnings profile and the implied “peak ages” with the respective confidence intervals by gender.
* A discussion about the implied peak ages and their statistical similarity/difference

GRAFICO PEAK AGE FEMALE CON INTERVALOS

El máximo valor predicho para el salario de los hombres en el *peak age* es de 85.833.333, mientras que el máximo valor predicho para el salario de las mujeres en esta edad es 40.000.000

* To highlight that results are the same using the ”long” regression or FWL, include a comparison of both in your results. Include a third estimate using FWL and bootstrap to construct the standard error.

REGRESION LONG Y FWL

STANDARS ERROR

Se observa que al agregar más variables de control se mejora la significancia estadística de las estimaciones de los parámetros. El R cuadrado en este caso es de 0.67, lo cual indica que el 67 % del cambio en el ingreso estaría explicado por las variables.

Este modelo condicional predice que los hombres tienen un salario 14,6 % mayor que las mujeres, y es estadísticamente significativo.

* A thoughtful discussion about the unconditional and conditional earnings gap, seeking to answer if the changes in the coefficient are evidence of a selection problem, a discrimination problem,” or none of these issues.

4. Predicting earnings

* Report and compare the predictive performance of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity

ESPECIFICACIONES Y MSE

* In your discussion of the results, comment:

i. About the performance metric you have chosen and your rationale for choosing it.

ii. About the specification with the lowest prediction error.

iii. For the specification, explore those observations that seem to ”miss the mark.” To do so, compute the influence statistic for each observation in the test sample, and examine its distribution. Are observations in the tails of the influence statistic distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

INFLUENCIA DE LAS OBSERVACIONES Y DISTRIBUCION

* LOOCV.For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-outcross-validation (LOOCV). Compare the results of the test error with those obtained with the validation set approach and explore the potential links with the influence statistic. (Note: when attempting this subsection, the calculations can take a long time depending on your coding skills, plan accordingly!)

ERROR PREDICTIVO