Luisa Cuellar - 201613942
Isabella Riveros – 201923015
Link repositorio: https://github.com/MiguelContreras1/Taller-1-BDML.git
Link de repositorio pasado:

Problem Set 1: Predicting Income

1. **Análisis descriptivo**

En el sector público, la declaración precisa de los ingresos individuales es fundamental para calcular los impuestos. Sin embargo, el fraude fiscal de todo tipo siempre ha sido un problema constante. Según el Servicio de Impuestos Interno, alrededor del 83,6% de los impuestos se pagan voluntariamente y a tiempo en EE. UU. Una de las causas de esta brecha es la subdeclaración de ingresos por parte de las personas. Un modelo de predicción de ingresos podría potencialmente ayudar a señalar casos de fraude que podrían conducir a la reducción de la brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a las personas y familias vulnerables que pueden necesitar más ayuda. Este trabajo tiene como propósito analizar los datos obtenidos en la Gran Encuesta Integrada de Hogares (GEHI) para el año 2018, con el propósito de elegir el mejor modelo que prediga los ingresos que tiene una persona en Colombia dadas sus características personales y laborales, y con este obtener señales claras de las posibles obligaciones tributarias que debería asumir.

Dichos datos fueron obtenidos a partir de la metodología de scrapping propuesta por el economista Ignacio Sarmiento a través de su página web, donde se hallaban reposados 10 links correspondientes a diferentes secciones de la muestra. Para obtenerlos se utilizó un loop que nos permit pasar por todas las bases de datos. Realizar este loop fue la mayor limitación.

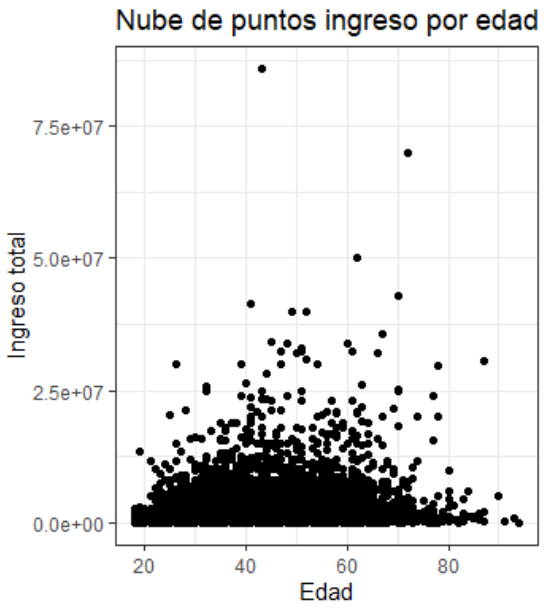
La base de datos utilizada en este trabajo está conformada por personas mayores de edad (mayores a 18 años) que actualmente se encuentren trabajando. Cabe resaltar que la variable utilizada para diferenciar a las personas ocupadas fue la varibale *ocu* la cual incluía como personas ocupadas a aquellas que realizan trabajos no remunerados, incluyendo a mujeres encargadas de la economía doméstica del hogar o a personas informales que apoyan actividades económicas familiares o externas, por este motivo es que se observarán un grupo de personas con ingreso 0 pero que están ocupadas. El resultado del filtro fue una muestra con un total de 16542 personas observadas para el año 2018.

En la tabla 1 se encuentran algunas estadísticas descriptivas de esta muestra, donde se observa que en términos de género se encuentra balaceada, siendo el 47% mujeres, pero se observa que las personas informales están representadas al ser el 41,3% de la muestra. Así mismo, la edad promedio de las personas observadas es aproximadamente de 39 años, siendo la edad promedio muy cercana entre hombres y mujeres.

Ahora bien, adentrándose en la variable de interés *ingtot*, se observa en la figura 1 su dispersión en la que se ve que la mayoría de los ingresos están condensados por debajo de 2’500.000, pero teniendo un ingreso promedio de 1.769.379 asociado a una varianza de 7.15e12. Esto, junto a los valores máximo (85.833.333) y mínimo (0), revelan la importante variabilidad que tienen los datos de ingreso, mostrando que se observan personas de todas las edades con ingresos dispares. Si se observa esta misma distribución dividida por género se observa que el comportamiento es casi que similar entre hombres y mujeres, con la excepción de que, para los hombres, la variabilidad es mayor hacia los ingresos altos, pues el ingreso promedio de hombres es de 1.915.848 con una varianza de

9.030043e+12, mientras que el ingreso promedio de las mujeres es de 1.604.223 con una varianza de 4.998615e+12. Esto podría indicar la importancia de diferenciar en el análisis el efecto sobre los ingresos del género.

Figura 1. Nube de puntos del ingreso



Así mismo, si se observan variables como la educación, las horas de trabajo o el tipo de empleo, se denota que las personas observadas pertenecen en su mayoría al sector formal, y trabajan como empleadores, lo cual facilitaría el rastreo de sus ingresos y además explica las altas tasas de afiliación a un régimen de salud y de cotización de pensión, además, el oficio más frecuente es el de vendedores, ambulantes, a domicilio, de loterías y periódicos o mercaderistas, cambiando de acuerdo al género, para las mujeres el oficio más frecuente es el mismo, pero para los hombres es de conductores de vehículos de transporte, taxistas o choferes. Así mismo, se observa consistencia en la muestra al observar que casi la totalidad de personas que cotizan pensión son del régimen contributivo en salud, al igual que las personas con mayores niveles educativos son los que mayores ingresos tienen. Si se observa el número de horas trabajadas por empleado se puede apreciar que en promedio una persona empleada de manera formal trabaja alrededor de 50 horas semanales, asociada a una varianza de 24.157, donde solo el 3,39% de la muestra trabaja horas extras en otro empleo. La distribución de estas horas laborales en un segundo trabajo muestra que en promedio las personas que las tienen trabajan entre 1 a 50 horas semanalmente, con una varianza de 7.13, lo cual muestra que hay personas que pueden trabajar hasta 8 horas más en un segundo trabajo.

Ahora bien, todo el análisis descriptivo se basó en la variable de ingreso *ingtot*, la cual corresponde al ingreso total por individuo incluyendo otros ingresos a parte del salario. Esta variable fue la elegida como variable dependiente, debido a que es la que mejor se ajusta al análisis sobre predicción de ingresos que se pretende realizar. Esta variable de ingreso total acota no solo el salario de los agentes sino también, otro tipo de ingresos, ya sean derivados de segundos trabajos, subsidios, o rentabilidades externas. Este tipo de ingresos son los que se quisieran predecir, para poder seleccionar correctamente el monto de tributación correspondiente por persona. En la encuesta GEIH para el año

2018, se encontraban también otras posibilidades de variables que median el ingreso, pero lo hacían de manera corta o incompleta para el propósito de este trabajo. Por ejemplo, se encontraban variables indicativas de ingresos por subsidios de vivienda, alimentación o transporte, o primas salariales. Estas, al igual que otras, solo muestran una parte de todo el ingreso a analizar. Por otro lado, había otras variables como *impa* o *isa* que solo representaban el ingreso monetario de una actividad económica (no necesariamente laboral) entre varias actividades antes de imputación, dejando de lado esas otras actividades que también generan ingresos y sesgando el análisis. Así mismo, las variables *y_salary_m* o *y_ingLab_m* solo observan los ingresos obtenidos por el salario, al igual que la variable *p6500, la cual* solo toma en cuenta los ingresos del mes pasado en su empleo, sin contar subsidios, dobles salarios o ingresos extras. Por último, la variable *ingtotes* solo tiene en cuenta los ingresos imputados y deja de lado otros tipos de ingresos que no fueron imputados. Por todo lo mencionado anteriormente, la variable ingreso total (*ingtot)* será la elegida para este trabajo.

Por último, cabe resaltar que, en el desarrollo del análisis descriptivo de la muestra, se encontraron varios missing values en algunas variables que serán utilizadas en modelos posteriores. La forma como se atendieron estos missings se presenta en la siguiente tabla:

Tabla 4. Missing values

Variable con missings	***Cantidad de Missings***	***Tratamiento***
Máximo nivel educativo alcanzado (maxEducLevel)	*1*	*Reemplazo por nivel de educación más repetido (nivel 3)*
Régimen de salud (regSalud)	*1420*	*Reemplazo por 0- nueva categoría asociada a no tiene régimen de salud*
Horas semanales trabajadas en segundo empleo (hoursWorkActualSecondJob)	*15980*	*Reemplazamos por 0 pues representa que estas personas no tienen segundos empleos*

Cabe aclarar que, para el caso de la variable asociada al régimen de salud, se observó que las personas sin información en esta variable estaban clasificadas como no cotizantes a pensión y no tenían un patrón definido en cuánto a la edad, ingresos u horas trabajadas en su primer empleo, por lo que se asume que no pertenecen a ningún régimen de salud. Así mismo, para la variable *de horas trabajadas en el segundo empleo (hoursWorkActualSecondJob)* se asumió que estas personas no tienen información en esta variable porque no cuentan con un segundo trabajo.

2. Estimación modelo edad-ingresos

Como primer ejercicio de acercamiento al modelo final de predicción del ingreso, se estimará la relación entre el ingreso total y la edad, pues como se ha especificado en la literatura, la edad tiene un efecto directo en el ingreso. Dicha relación no se planteará de manera lineal, sino que, como es sabido, existe una edad en la que el ingreso llega a su máximo y el objetivo en este caso es poder reconocer dicho pico para esta muestra.

En primer lugar, se plantea un modelo base de ingreso individual de la forma:

$$Earnings = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \qquad (1)$$

Donde la variable *ingtot* (ingresos totales) se eligió como la variable que representa las ganancias, dado que esta agrupa todas las fuentes de ingresos del individuo durante el último mes.

Tabla. Regresión del modelo edad-ingresos

	Estimación	***Error Estándar***	***t Valor***	***Pr(>|t|)***	***IC***
Intercepto	*-436662.9*	*178347.2*	*-2.448*	*0.0144 **	*[*
Edad	*91143.5*	*8886.4*	*10.256*	*2e-16 ****	*[*
Edad^2	*-799.3*	*102.9*	*-7.771*	*8.24e-15 ****	*[*
Observaciones	*15642*	*R2 ajustado 0.01716*		*IC estimación {*	

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), alcanzando una edad donde el ingreso es máximo.

Al estimarlo se obtuvieron los resultados en la tabla anterior, donde se observa que por cada año extra de edad el ingreso de las personas en 2018 aumentó en $91143.5 pesos, para un nivel de significancia del 1% y con un error estándar calculado por Bootstrap. Por otro lado, los resultados de la tabla muestran que existe un efecto cuadrático asociado a la edad, debido a que el estimado de la variable *age2* es negativo y significativo al 1%, mostrando que si bien a medida que aumenta la edad los ingresos totales aumentan, estos lo hacen de manera decreciente alcanzando un óptimo. Dicha edad máxima para este caso es de 57 años con un intervalo de confianza calculado por Bootstrap de *2.062.541-2.260.888* (peak age).

$$Ingreso\ total = \hat{\beta}0 + \hat{\beta}_1 age + \hat{\beta}_2 age^2$$

$$\frac{dy}{dage} = \hat{\beta}_1 + 2\hat{\beta}_2 age = 0$$

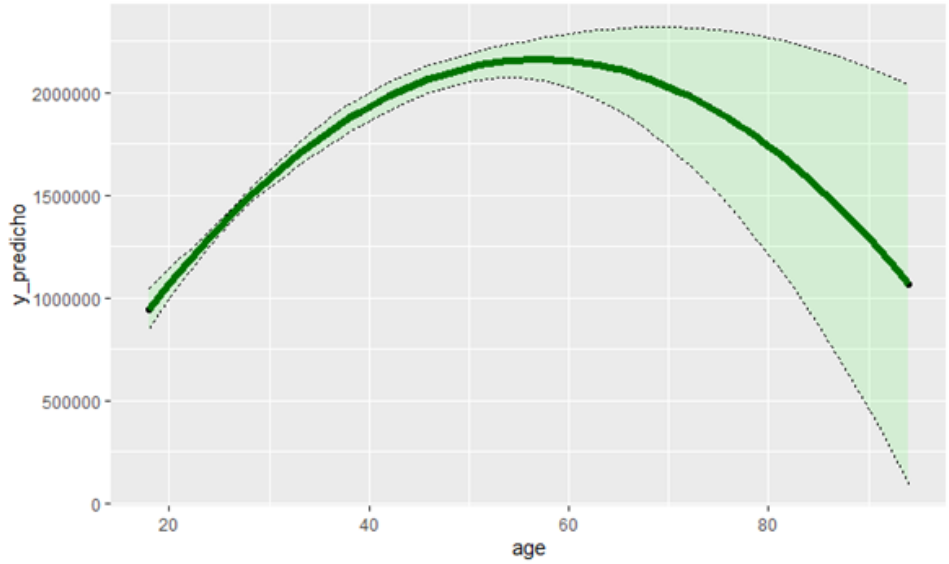
$$agemáx = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$$

$$agemáx = -\frac{91160.7}{2(-799.5)}$$

$$agemáx = \frac{91160.7}{1599}$$

$$agemáx = 57\ años$$

Figura. Estimación de peak age



El ingreso total predicho para la *peak age* de 57 años es de $ 2.161.715, con un intervalo de confianza por Bootstrap de [2.061.110 - 2.262.319]. En la figura 8 se observa dicha predicción, donde se observa particularmente como los intervalos de confianza empiezan a hacerse más granes justo después del *peak age*. [1], y en segundo lugar para niveles altos de edad, los datos presentan mayor variabilidad en el ingreso total. Este comportamiento 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 aumentaba considerablemente. En cuanto a la discusión de por qué la edad de ingreso máximo es a los 57 años, se podrían pensar varios escenarios, uno de ellos es que, en ese momento, la persona ya cuenta con todos los niveles educativos alcanzados y con un nivel de experiencia que le permite asumir cargos que le representan ingresos altos.

En cuanto al ajuste del modelo el R-cuadrado ajustado en este caso es de 0.01716 lo cual indica que solamente el 1,7 % del cambio en el ingreso es explicado por el aumento decreciente en la edad, lo cual, sumado a la alta significancia de los estimadores, podría indicar problemas de colinealidad y sesgo por especificación, sin embargo, al ser el primer acercamiento queda claro que la edad es un elemento relevante para el análisis, pero que necesita ser limpiado.

3. Estimación modelo sexo-ingresos y sexo-edad-ingresos

Después de realizada la discusión alrededor del modelo de edad-ingresos, en esta parte del trabajo se buscará ahondar en dichos resultados separando las estimaciones por sexo, pues como bien es sabido las mujeres en promedio tienen menos ingresos totales que los hombres y es de esperarse que la edad de máximo ingreso (*peak age)* al igual que el ingreso máximo varíe en la misma dirección. Como se observará más adelante, en esta muestra dicho comportamiento se cumple debido, entre otras cosas a

[1] *Método delta*:

que, en esta muestra se incluyeron en el grupo de ocupados a mujeres que realizan trabajos domésticos no remunerados, o acompañan actividades laborales familiares no remuneradas, por lo que como se vio en la tabla 1, las personas con ingreso 0 en su mayoría son mujeres y esto, además de las rigideces en el mercado laboral y hechos asociados a discriminación, el ingreso máximo alcanzado por las mujeres es inferior.

Para este propósito se estimaron dos modelos, el primero de ellos es el correspondiente a la estimación de la semi-eslasticidad del ingreso con respecto al sexo (2) y el segundo incluye las variables de la edad incluidas en el modelo anterior. El propósito de correr estos modelos es poder comparar sus especificaciones en cuánto a la predicción del ingreso máximo y su edad correspondiente diferenciada entre hombres y mujeres.

$$\log(Income) = \beta_1 + \beta_2 Female + v \qquad (2)$$

$$\log(Income) = \beta_1 + \beta_2 Female + age + age^2 + e \qquad (3)$$

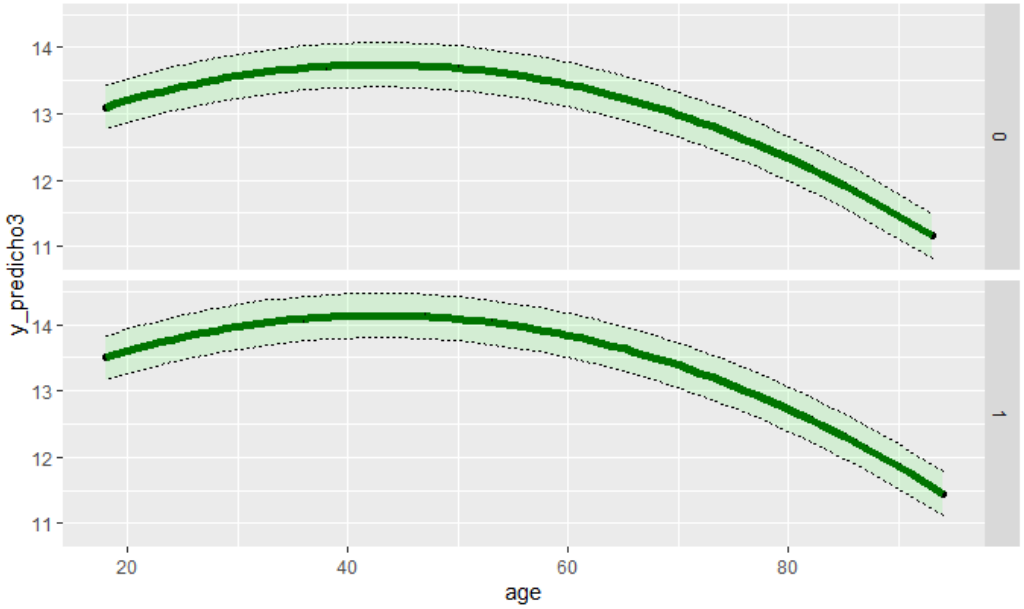
$$\log(Income) = \beta_1 + \beta_2 Female + age + age^2 + age\_Female + p \qquad (4)$$

Para estas especificaciones se decidió suavizar el ingreso dada la alta varianza observada en el modelo anterior y se mantuvo la misma variable de ingreso elegida previamente, pues esta reveló comportamientos esperados de la *peak age* y evidenció unos intervalos de confianza acordes con el tipo de estimación. Los resultados de esta parte, sumado a los resultados de la estimación del primer modelo se encuentran en la tabla 5.

	(modelo 1)	**(modelo 2)**	**(modelo 3)**	**(modelo 4)**
Age	91,143.460*** (8,886.416)		0.088*** (0.007)	0.082*** (0.007)
Age^2	-799.261*** (102.852)		-0.001*** (0.0001)	-0.001*** (0.0001)
Sex_age				0.016*** (0.002)
Sex		0.378*** (0.030)	0.399*** (0.030)	-0.244*** (0.094)
Constante	-436,662.900** (178,347.200)	13.549*** (0.022)	11.846*** (0.132)	12.151*** (0.138)
Observaciones	16,542	16,542	16,542	16,542
R^2	0.017	0.009	0.020	0.023
Peak ages	57		43	
IC	2.062.541- 2.260.888			
IC mujeres			13.19 -14.28	
IC hombres			13.59- 14.68	

En esta se observa que el modelo que 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, siendo esta variable significativa para un p-valor del 1%. Sin embargo, el R-cuadrado asociado sigue siendo muy bajo 0,92%, lo cual indica que menos del 1 % del cambio en el ingreso es explicado únicamente por el sexo. Para la estimación del modelo 2, agregando el efecto de la edad, se ve como el estimador del

sexo se ve disminuido, sin embargo, no es menos significativo. Para este caso, se observa que en promedio los hombres ganan XX% más que las mujeres y tienen un ingreso máximo predicho que duplica al de las mujeres (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) sin embargo, la peak age de ambos sexos es la misma, alrededor de 43 años, para unos intervalos de confianza de [13.59-14.68] para los hombres y [13.19-14.28] para las mujeres para un nivel de confianza del 95%. Esto demuestra como la edad de ingreso máximo es la misma, y hay cambios en la pendiente debido al género.



Esta brecha en los ingresos incondicionados, muestran que a pesar de que el poder predictivo de los modelos no es el mejor, hecho evidenciado en el R2 de todos los modelos en la tabla anterior, se observa como estas dos variables tienen un papel relevante en la estimación de los ingresos de los agentes, e incluso evidencian hechos estilizados de la economía asociados a la edad y al sexo.

Sin embargo, con el propósito de limpiar aún más estas estimaciones y predecir de forma correcta el ingreso total, en la próxima sección se incluirán covariables de control.

3.1 Estimación del modelo sexo-edad-ingresos con covariables

Como se mencionó previamente en este caso se estimará el mismo modelo (3) pero incluyendo covaribles asociadas a características socioeconómicas de las personas y a características del empleo donde trabajan, con el fin de limpiar el efecto asociado al sexo.

Las variables elegidas para incluir como controles en esta especificación son las siguientes:

Características del trabajador	***Características del trabajo***
Estrato: Es una variable ….	*Relab*
Maximo nivel educativo alcanzado	*Oficio*
Número de horas trabajadas a la semana (1er trabajo)	*Tamaño de la firma*

Número de horas trabajadas a la semana (2do trabajo)
Informalidad:

El modelo a estimar es el siguiente

$$Log(ingreso_i) = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \sum \beta_3 Max_{educ} + \sum \beta_4 estrato + \sum \beta_5 Regsalud + \sum \beta_6 CotPension + \sum \beta_7 oficio + \beta_8 HoursWorkUsual + \beta_9 HoursWorkUsualSecondJob + \beta_{10} iformal + \sum \beta_{11} relab + q \quad (5)$$

En este solo se analizan linealidades pues en la próxima sección se estudiará el poder predictivo de modelos más complejos.

	(2)	***(3)***	***(5)***	***FWL***
Age		*0.088*** (0.007)*	*0.027*	
Age^2		*-0.001*** (0.0001)*	*-0.00022*	
Sex_age				
Sex	*0.378*** (0.030)*	*0.399*** (0.030)*	*0.153*	
Res_a				*0.153*
Constante	*13.549*** (0.022)*	*11.846*** (0.132)*	*1.27 e 1*	
Observaciones	*16,542*	*16,542*	*16,542*	
R^2	*0.009*	*0.020*	*0.6635*	

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.

4. Prediciendo los ingresos

Una vez analizadas las especificaciones anteriores, donde se evidenció que el ingreso, la edad y demás covariables logran predecir los ingresos totales de una persona, en esta sección se encontrará el modelo que mejor prediga dichos ingresos a través de la comparación del MSE (error cuadrático medio) entre modelos con distintas especificaciones que explorarán no linealidades e interacciones entre variables. Esto con el propósito de llegar a la complejidad óptima de la especificación abordada en el ejercicio previo.

La razón por la que se eligió esta métrica de comparación (MSE) es porque esta permite comparar y representar el balance entre sesgo y varianza en la estimación de un modelo y permite generalizar la idea de complejidad, haciendo que el nivel óptimo de esta sea aquel que permita pronosticar con un sesgo y varianza asociadas a un MSE bajo.

A pesar de que existen otras métricas, se elige esta por la particularidad de los datos observados en este trabajo. Como se vio desde el principio el ingreso es una variable que tiende a tener colas grandes y outliers con influencia relevante en la estimación, dado que MSE no es tan sensible a los outliers en comparación con otras métricas, se elige esta como punto de comparación entre las especificaciones. Además de ello, es simple y fácil de implementar. Respecto al resto de métricas hay algunos problemas asociados a estas que no permitieron elegirlas. Por ejemplo, el RSS brinda una medida absoluta de ajuste de los datos, pero no es claro que constituye un buen RSS. Así mismo, el R-cuadrado proporciona la cantidad de varianza explicada por la estimación, pero esta es una métrica de causalidad más que de predicción.

No obstante, se deja en claro que el MSE no es la métrica perfecta, también tiene algunas desventajas: relacionadas a su alta variabilidad por el uso de un subconjunto de observaciones para ajustar la especificación (datos de entrenamiento). Y su sensibilidad a funcionar peor cuando se entrenan con pocas observaciones.

Ahora bien, para la comparación se estimaron en total 9 especificaciones, incluyendo las cuatro estimaciones anteriores. La idea con las 5 especificaciones adicionales es poder explorar no linealidades e interacciones y mostrar como la sobreespecificación también es un problema a la hora de encontrar el modelo con menor varianza. Particularmente las especificaciones exploradas fueron:

Tabla 8. Especificaciones utilizadas para comparación de MSE

	Variables e interacciones incluidas en la especificación													
Especificación	*age*	*Age2*	*sex*	*Max_educ*	*estrato*	*Reg_salud*	*Cotpensión*	*sizeFirm*	*oficio*	*HWU*	*HWUSJ*	*informal*	*Relab*	*Interacciones*
1	*X*	*X*												
2	*X*	*X*	*X*											
3	*X*	*X*	*X*											*Age_sex*
4	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	
5	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*Age_sex , sex_MaxEduc, sex_informal, sex oficio*
6	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*sex_MaxEduc, sex_Age, sex_informal, HWU_max educ*
7	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*XX*	*X*	*X*	*XX*	*X*	*Sex_oficio, sex_regsalud, sex_informal, age_sex, HWU_Maxeduc, HWU_informal, regsalud_Cotpension, oficio_relab, HWU^2*
8	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*HWU_Maxeduc HWUSJ_informal, HWUSJ_oficio, HWUSJ_relab, HWUSJ_relab^3 relab^3_oficio, HWU^3, regsalud_Cotpensio,*

														HWUSJ^3 _ age^2
9	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*X*	*HWU_Maxeduc HWUSJ_informal, HWUSJ_oficio, HWUSJ_relab, HWUSJ_relab^3 relab^3_oficio, HWU^3, regsalud_Cotpensio, HWUSJ^3 _ age^2 Oficio_HWU^3*

Los resultados obtenidos del MSE para cada modelo se observan en la tabla 10.

Tabla 10. Comparación de MSE para las 9 especificaciones elegidas

Especificación	MSE
1	5.766696
2	3.86366
3	3.794644
4	1.119276
5	1.12227
6	1.117673
7	1.172084
8	3.1105
9	3.561439

Con esto se evidencian dos resultados principales. El primero de ellos es el mejoramiento notorio del MSE al incluir los controles en el modelo 4, pues este disminuye de 3.79 a 1.12. El segundo es que la sobreespecificación del modelo lleva a que el MSE aumente, esto se evidencia en las especificaciones 7 en adelante, donde mayor cantidad de interacciones y no linealidades se incluyen. Para este caso los mejores modelos son los correspondientes a las especificaciones 5 y 6, siendo la 6 el modelo con menor MSE (1.117). En este se incluyeron interacciones entre la variable de sexo y el nivel educativo, si es formal o informal y las horas trabajadas por el nivel educativo.

$$Log(ingreso_i) = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \sum \beta_3 Max_{educ} + \sum \beta_4 estrato + \sum \beta_5 Regsalud + \sum \beta_6 CotPension + \sum \beta_7 oficio + \beta_8 HoursWorkUsual + \beta_9 HoursWorkUsualSecondJob + \beta_{10} iformal + \sum \beta_{11} relab + \beta_{12} age\_sex + \beta_{13} sex\_Maxeduc + \beta_{14} sex\_informal + \beta_{15} HoursWorkUsual\_MaxEduc + w$$

Cabe resaltar que estas interacciones se incluyeron con un sentido económico, dado que se piensa que puede existir un efecto diferenciado asociado al genero a través de la educación, el oficio y las horas trabajadas. Especialmente, se piensa que una mujer que trabaja la misma cantidad de horas que un hombre, o que tenga el mismo oficio o el mismo nivel educativo, su ingreso total promedio será menor en comparación. De hecho, estos resultados se respaldan por los obtenidos por la especificación 6, donde se observa que, si existen dichos efectos diferenciados, sin embargo variables como la edad al cuadrado y los niveles de educación pierden significancia. Este modelo no solo obtuvo el MSE más bajo sino el R-cuadrado ajustado más alto, lo cual es relevante para esta selección, pues se debe tener

en cuenta que para poder predecir correctamente el ingreso variable es necesario estimar de manera óptima los coeficientes.

Ahora bien, este modelo no tiene un ajuste perfecto ni predice de forma inequívoca cualquier observación, pues esto implicaría una sobreespecificación dañina fuera de muestra. Es por ello, que se observarán aquellos niveles de ingreso que el modelo no predice bien, los cuales están asociados a ingresos muy altos y a ingresos de cero (outliers). En la Figura 6, se observa como el ajuste del modelo no predice algunos ingresos altos, ni tampoco los ingresos que son cero. Frente a esto se optó por calcular cuál es la influencia de cada observación sobre la estimación del modelo, con el fin de identificar si son dichos outliers los que mayor influencia poseen.

En la Figura 7. Se observa la distribución de la influencia estadística de las observaciones utilizadas en la base de testeo (3308). En esta se observa como la influencia pareciera comportarse como una distribución normal, a excepción de un par de observaciones que pareciera no tienen influencia en la estimación. Para poder reconócelas se presenta la figura 8 donde se cruza la influencia con el logaritmo del salario y se ve como los salarios altos son los que influencian más la estimación en comparación con salarios más bajos. Sin embargo, en esta gráfica también se observa que algunas observaciones asociadas a salarios altos tampoco tienen influencia. Estos salarios altos corresponden a observaciones en la cola derecha de la distribución del ingreso. Con esto queda en evidencia que las observaciones con mayor influencia estadística son las correspondientes a salarios altos. Dado que el ingreso en este caso tiene tanta variabilidad en la cola derecha, el modelo está dándole mayor peso a estas personas, por lo que podría verse afectada la predicción.

Si observamos las características de la distribución de la influencia se observa que este resultado se respalda, pues esta tiene una influencia mínima de 0.52 y máxima de 15.97 la cual corresponde a una desviación de 1.7105 y una media de 13.60. Sin embargo, el 75% de los datos están por debajo de 14.00 y el 25% por debajo de 13.00. Frente a la motivación de este trabajo, este resultado evidencia como las personas con ingresos altos en efecto corresponden a observaciones con pesos relevantes en la estimación, y por ende a una posible sobrepredicción del ingreso, no obstante, estas personas en la cola derecha de la distribución son potenciales tributarios que la DIAN pues el modelo se encarga de predecir mejor dichos salaries altos frente a salaries más bajos.

Figura 6. Distribución de la influencia estadística

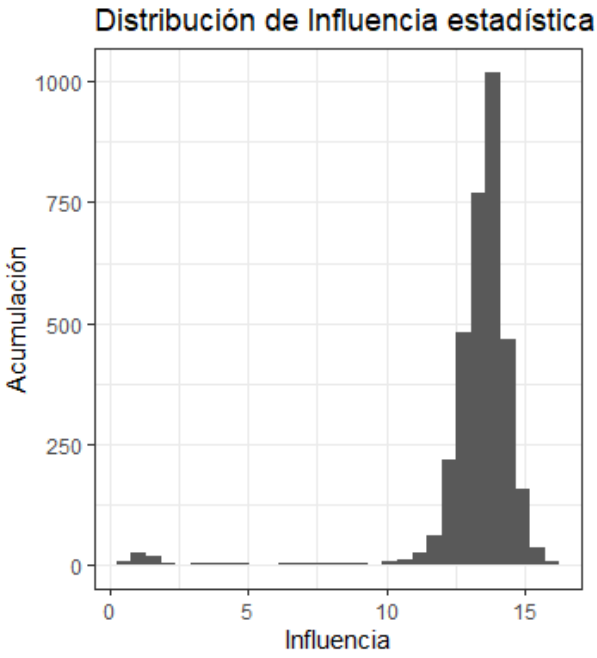


Figura 7. Distribución de la influencia estadística

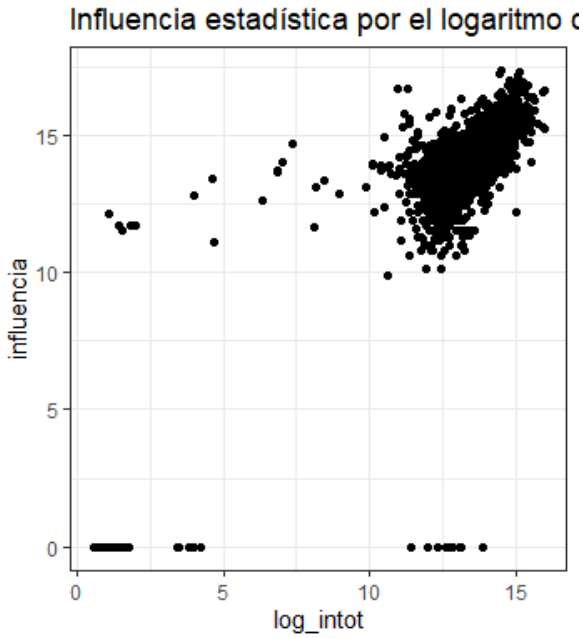
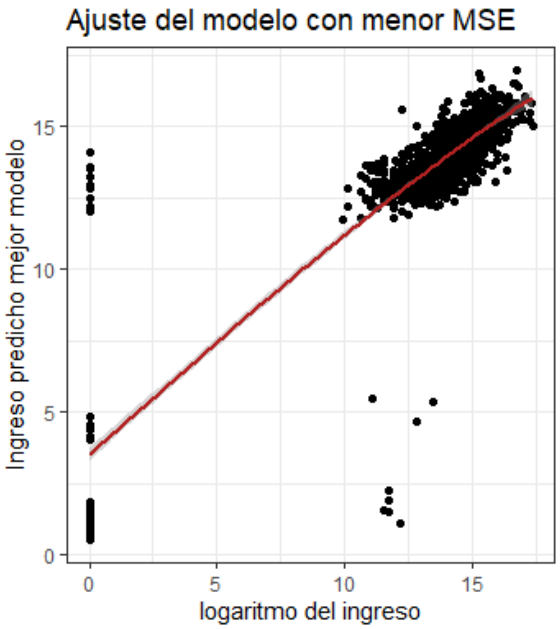


Figura 8. Ajuste modelo con menor MSE (especificación 6) contrastado con el logaritmo del ingreso



Por último, para la elección del modelo definitivo de predicción se tomarán en cuenta los dos modelos ganadores del filtro realizado a través del MSE (especificación 5 y 6). Frente a los cuales se hará uso de validación cruzada en K-partes para la elección del modelo definitivo[2]. En este caso se dividirán los datos en K partes (5 y 10) y luego se ajustará el modelo dejando afuera una de las partes, para posteriormente calcular el error de predicción en dicha parte extraída y con esto promediar los MSE y obtener el modelo que mejor predice. Cabe mencionar que el enfoque del conjunto de validación tiende a sobreestimar el error de predicción en la muestra de prueba y es sensible a la cantidad de datos.

Dicho ejercicio de validación cruzada es posible efectuarlo en n partes (LOOCV), donde se promedian los resultados de n modelos ajustados, cada uno entrenado en un conjunto casi idéntico de observaciones. Este último está más correlacionado que el método utilizado en este trabajo de hacerlo en K partes, pues este promedia la salida de k modelos ajustado que estarían algo menos correlacionados. Se ha demostrado empíricamente que estas validaciones en k partes producen estimaciones del error de predicción que no sufren ni de un sesgo excesivamente alto ni de una varianza muy alta. Es por ello que con el ejercicio previamente realizado de influencia y con la validación cruzada con k partes, se cubre el análisis justo para poder seleccionar el mejor modelo predictivo, pues dicho LOOCV (k=n), está correlacionado con el ejercicio de influencia estadística previamente explorado, pues lo que realiza el LOOCV es la comparación del MSE cuando se elimina una observación de la muestra, lo que casi equivale a identificar la relevancia estadística de esa observación sobre la estimación.

Siendo así en la siguiente tabla se presentan los resultados obtenidos por CV en 5 y 10 partes.

[2] No se realizó LOOCV porque ninguno de los computadores permitió correrlo debido a problemas de memoria.

Tabla. Resultados por validación cruzada para 5 y 10 partes especificaciones 5 y 6

	Especificación 5	Especificación 6
	K=5	
RMSE	1.1087	1.093937
R-cuadrado	0.6598	0.6877783
MAE	0.4889	0.5549926
	K=10	
RMSE	1.1108	1.071623
R-cuadrado	0.5996	0.678271
MAE	0.4598	0.5531

Con esto queda claro que el modelo elegido para predecir es el correspondiente a la especificación 6.