ESTIMACIÓN DE LOS RETORNOS A LA EDUCACIÓN A PARTIR DE ENCUESTAS DE HOGARES Y ALGUNOS RESULTADOS PARA COLOMBIA

JORGE LEONARDO DUARTE RODRÍGUEZ
CÓDIGO: 07215001

PROYECTO DE GRADO II

PROFESORA:
BLANCA CECILIA ZULUAGA DÍAZ

UNIVERSIDAD ICESI

FACULTAD DE CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS
PROGRAMA DE ECONOMÍA CON ÉNFASIS EN POLÍTICAS PÚBLICAS
SANTIAGO DE CALI
24 DE MAYO DE 2012

Tabla de Contenido

1. **Introducción**

Este documento es un tutorial que explica cómo calcular algunos retornos de la educación a partir de encuestas de hogares. El objetivo es ayudar al lector a pasar de las clases de econometría a la estimación con datos reales. Generalmente los objetivos de un curso básico de econometría se desarrollan con los datos listos para la estimación. Cuando se pasa a aplicar la teoría con datos reales, el investigador encuentra algunos aspectos que no necesariamente debería saber. Por medio de la estimación de una ecuación de oferta salarial para Colombia con la Gran Encuesta Integrada de Hogares (GEIH) del mes de septiembre de 2010[1], se intenta compartir una explicación útil que permita al lector replicar y entender el procedimiento. Al ser datos reales, se discuten los resultados de la estimación.

Para lograr el objetivo, el documento se divide en cinco partes. La primera es esta introducción. La segunda es una explicación sencilla de lo que se debería saber del funcionamiento de las encuestas de hogares, en especial del contenido y forma de las variables y de los factores de expansión. Como se va a estimar una ecuación particular y los economistas no estiman porque sí, la tercera parte es una corta revisión de la literatura de interés que brinda la intuición del trabajo a realizar. La cuarta parte es la estimación[2]. Se intenta corregir algunos problemas de endogeneidad. Y la quinta parte son las discusiones y conclusiones de los resultados.

Palabras claves: tutorial, encuestas de hogares, muestreo, factores de expansión, endogeneidad, ecuación de salarios, educación, Mincer.

[1] Los datos los comparte el Departamento Administrativo Nacional de Estadística (DANE) en su página de internet.
[2] El paquete estadístico para la estimación es Stata 12.0. Stata es un producto de StataCorp. Como el paquete es sólo una herramienta, la información sigue siendo útil para la estimación con otros programas.

2. Encuestas de Hogares

Gracias a los avances tecnológicos en computación, que permiten entrevistar personas y procesar datos a un bajo costo (comparado con años anteriores), las encuestas de hogares se han convertido en un instrumento importante para la medición demográfica, el comportamiento de los hogares y el efecto de los programas y políticas públicas. Si bien cada encuesta tiene sus propios objetivos (por ejemplo, la GEIH está particularmente interesada en el mercado laboral), varios países han recibido acompañamiento del Banco Mundial para desarrollar la metodología de las encuestas de hogares. Esto trae la ventaja de poder comparar resultados entre países y entre encuestas (y para el lector, la ventaja de que puede aprender por medio de este documento así no necesite usar la GEIH).

Es importante tener en cuenta que las observaciones de las encuestas de hogares (al no ser resultados de experimentos controlados) no muestran ninguna causalidad entre las variables ni la magnitud del efecto de programas sociales o políticas públicas. Para responder a problemáticas específicas es necesario asumir una metodología de investigación empírica. En el caso de los retornos a la educación, se usa la microeconometría para determinar el efecto de la educación (y otras variables sociodemográficas) en el salario de las personas.

La población objetivo de las encuestas de hogares son, como su nombre lo indica, los hogares. Es preciso hacer esta aclaración porque no hacen parte de la muestra las personas sin hogar, las personas de las fuerzas armadas o alguna otra clase de personas que trabajan por largo tiempo fuera del hogar. Se debe tener cuidado al interpretar los resultados de cualquier análisis con encuestas de hogares porque algunas personas se estas excluyendo de la muestra de forma no aleatoria.

2.1 Muestreo y factores de expansión

Para realizar una encuesta de hogares es necesario conocer una lista de los hogares de la población. Muchas encuestas se basan en censos o realizan actividades de reconocimiento de los hogares que se actualizan con la misma encuesta, para conocer a toda la población y la probabilidad que tiene cada hogar de ser seleccionado en la muestra. Si la muestra fuera completamente aleatoria, cada hogar tendría la misma probabilidad de ser seleccionado. Ésta sería la mejor muestra que podría tenerse, pero, por varios motivos, la recolección de datos requiere usualmente un muestreo un poco más complejo.

Por motivo de costos de recolección, generalmente se hacen muestras por conglomerados. En la práctica es útil hacer grandes conglomerados, que no están aleatoriamente distribuidos en el espacio pero están geográficamente agrupados. Esto puede generar algunos costos estadísticos y de precisión de los datos, pero hay que recordar que precisamente la ventaja de las encuestas es conocer a la población a un bajo costo económico. A la hora de hacer una investigación, es indispensable que el investigador se informe sobre la encuesta que va a utilizar para tener una idea de la precisión de sus resultados.

Las muestras por lo general también son estratificadas. Esto quiere decir que se divide a la población en diferentes subpoblaciones con características similares cada una. El caso más común es estratificar la muestra por algún nivel de riqueza o calidad de vida. Se debe garantizar entonces que cada una de las submuestras tenga suficiente información que permita estimar las características de cada una. La razón de hacer esto es que una muestra aleatoria puede subestimar algunas características de la población si la muestra no cuenta con suficiente información sobre lo que se desea saber. Por ejemplo, si se desea saber el gasto en salud de Colombia, la GEIH puede no tener suficiente información sobre dicho gasto, así que una encuesta apropiada sería una que tenga en cuenta a las personas que más gastan en salud y luego descuente el peso que éstas tienen en la población porque se sabe que tienen más probabilidad de ser escogidas que las que no

gastan. De esta manera se contaría con más información para estimar mejor el gasto en salud. Esta es la razón por la que los países cuentan con diferentes encuestas. Por ejemplo, en Colombia, algunas encuestas de hogares importantes con objetivos específicos son la GEIG, la Encuesta Nacional de Calidad de Vida, La Encuesta Nacional de Ingresos y Gastos, la Encuesta Nacional de Demografía y Salud y la Encuesta Nacional de Salud.

También hay razones estadísticas para la estratificación. Si se sabe que el ingreso rural es menor (o simplemente diferente) que el ingreso urbano y se sabe la proporción de familias rurales y urbanas, es mejor estratificar la muestra en estos dos componentes. La media del ingreso total es la media ponderada de cada media. Como son muestras independientes, la varianza es simplemente la suma de las varianzas de los dos componentes. Esta varianza es menor que la varianza de un muestreo aleatorio porque en el muestreo estratificado la varianza es sólo la de grupos con características similares y no la de un grupo muy diverso. Por esta razón, las características relevantes y conocidas de la población ayudan a elegir muestras más representativas haciendo estratificaciones (como urbano y rural, etnicidad, regional, departamental o cabecera y resto).

Si una muestra es aleatoria completamente, se sabe que la probabilidad de cada hogar de ser seleccionado en la muestra es la misma, y por tanto se sabe que cada hogar representa el mismo número de hogares en la población. Si se quisiera saber el número de hogares de la población sólo habría que multiplicar cada uno por el número de hogares que representa y sumar todos los de la muestra. Sin embargo, si el muestreo se ha hecho de alguna forma más compleja (o tiene en cuenta las entrevistas que no se pudieron realizar pero que estaban programadas), cada hogar no tiene que representar exactamente el mismo número de hogares en la población porque posiblemente la probabilidad de ser seleccionado fue distinta. Las encuestas tienen en cuesta estos pesos de los hogares en la población para hacer estimaciones representativas. Por esta razón, las encuestas tienen una variable que refleja este hecho. Esta variable es el *factor de expansión final* de cada hogar. Por lo general se presenta como la última

variable de un conjunto de datos y el investigador sólo tiene que tener en cuenta esta variable sin preocuparse por la forma del muestreo de la encuesta. De esta forma, si se quiere conocer el número de hogares de la población se debe multiplicar cada hogar por su factor de expansión, que en este caso no es el mismo para todos los hogares, y luego sumar los hogares de la muestra.

Los paquetes estadísticos facilitan el manejo de datos provenientes de encuestas. Por lo general, tienen una opción para informarle al programa que se está trabajando con encuestas y que deben tenerse en cuenta los factores de expansión. *Stata* tiene un comando (*svy*) para este propósito y también tiene las opciones *weight* que se ponen al final de cada entrada para hacer estadísticas descriptivas o regresiones, teniendo en cuenta los diferentes pesos muestrales (que no son más que promedios ponderados).

2.2 Contenido y forma de las variables

Las encuestas de hogares pueden tener objetivos muy específicos, por lo que cada una tiene sus propios conglomerados y estratificaciones, preguntas particulares y periodicidad. Sin embargo, como sería bastante costoso e ineficiente hacer una encuesta por cada tema que se quiera estudiar, varias encuestas reúnen grandes grupos de preguntas con el fin de servir para diferentes objetivos. Si bien se puede sacrificar muestreos muy específicos, se gana consistencia y credibilidad entre diferentes investigaciones. Un ejemplo de estas encuestas de hogares son las llamadas Encuestas de Medición de Condiciones de Vida (LSMS, Living Standard Measurement Surveys) apoyadas en varios países por el Banco Mundial.

A pesar de que los objetivos de las políticas públicas y de las mediciones de bienestar están dirigidos a los individuos, el hecho de que las encuestas sean dirigidas también hacia hogares tiene varias explicaciones. Las encuestas de hogares tienen preguntas del hogar que van dirigidas a algún miembro en capacidad de responder (preferiblemente el jefe de hogar) y tienen preguntas individuales. La diferencia entre cada tipo de preguntas refleja la intención de

facilitar la respuesta al encuestado y obtener la información más confiable posible. Los hogares son núcleos económicos que comparten ciertas decisiones y características. Los hogares generalmente comparten los gastos en alimentación o gastos en servicios públicos y redistribuyen los ingresos: en ocasiones es mejor conocer el gasto (o el ingreso) del conjunto del hogar que, por ejemplo, el del hijo que gasta en pasajes y utilices para la escuela pero no percibe ingreso, separado del de la madre que va al trabajo y no a la escuela. Las preguntas sobre las condiciones del hogar pueden complementarse con preguntas individuales que tendrán mejor respuesta al dirigirse a cada persona del hogar. Para efectos de estimación, es importante tener en cuenta el tipo de variable que se está observando.

A la hora de preguntar por ingresos y gastos, las encuestas de hogares preguntan por diferentes periodos de tiempo. Por ejemplo, preguntan cuál fue el gasto en carne o leche la semana pasada, cuál fue el salario del mes pasado o cuánto pagó por impuesto a la renta en el último año. Cuando se pregunta por un periodo muy corto de tiempo se corre el riesgo de generar una varianza muy alta en las estimaciones porque nada garantiza que el gasto de hace un día se comporte como el gasto de todo el año: las personas hacen gastos que no tienen la misma periodicidad. Preguntar por periodos muy largos genera sesgo porque las personas no recuerdan muy bien cuánto se gastaron en todo el año. Las encuestas tratan de escoger el periodo de tiempo que se va a preguntar de tal forma que se minimicen estos riesgos: por ejemplo, el consumo de alimentos se debería preguntar en un periodo corto de tiempo y el consumo de bienes durables en un periodo largo. Para efectos de estimación es importante tener en cuenta el periodo de tiempo de las variables para no caer en errores de agregar preguntas de forma incorrecta (por ejemplo, semanales con trimestrales).

En la estimación de retornos a la educación se necesita la variable continua *Años de educación*. Teniendo en cuenta lo anterior, el lector no puede esperar encontrarse en la encuesta con esta variable. Las personas no tienen por qué estar contando cuántos años estudian. Esta variable se debe construir por medio

de preguntas como cuál es el mayor nivel educativo alcanzado y cuántos años cursados tiene (o realizó) en este último nivel. Es común en Colombia contar los años de educación de tal manera que once años de educación representen secundaria completa y dieciséis representen universitaria completa. En este sentido hay que tener cuidado en, por ejemplo, rebajar a once años las personas que han tenido algún bachillerato europeo de doce o trece años, y dejar en dieciséis a profesionales que tardaron más de cinco años en terminar sus estudios.

Aunque las características particulares de cada encuesta las irá aprendiendo el lector con su propia experiencia, cabe una última anotación. Algunas preguntas de ingresos y gastos de las encuestas de hogares se preguntan como una estimación de transferencias de bienes y servicios. Por ejemplo, además de preguntar a una persona por la compra de tomates y por el costo de dichos tomates, se pregunta si obtuvo tomates de alguna otra forma y cuánto piensa que deben costar dichos tomates (se pudieron haber cogido del patio de la casa). Además de preguntar a una persona si tiene algún empleo y cuánto recibe por dicho empleo, se pregunta si recibió algún bien por el mismo concepto. Esto refleja la necesidad de mediar algunas condiciones de vida de las personas que no se expresan necesariamente en los precios del mercado, en especial en zonas rurales donde las personas pueden cultivar sus mismos alimentos o pueden recibir pagos no monetarios por su trabajo. Estas estimaciones de transferencias de bienes pueden ser consideradas como ingresos y gastos al mismo tiempo y, para efectos de estimación, el investigador debe decidir cuándo tenerlas en cuenta.

3. Literatura

Existe amplia literatura y abundantes estudios empíricos sobre el tema. La revisión de la literatura aquí es simplemente un pequeño esbozo que tiene como objetivo hacer explícita la intuición económica detrás de la estimación.

La ecuación a estimar intenta explicar el efecto de la educación en el salario de las personas. Esta forma de ver la educación es conocida en la literatura de la *Economía de la Educación* como la perspectiva microeconómica o privada, y en *Economía Laboral* como la ecuación de oferta salarial. No se tiene en cuenta aspectos de la educación como las externalidades positivas (o negativas) de ésta en la población. Las ecuaciones a estimar están interesadas particularmente en el retorno económico de la educación y no en la felicidad (o tristeza) que genera estudiar, ni en el bienestar y crecimiento económico que le puede generar a un país el hecho de tener una población con mayor nivel de educación. Es importante tener en cuenta que los resultados de las estimaciones a realizar no muestran completamente qué tan bueno (o malo) es estudiar, sino que sólo muestran una parte de la demanda de educación.

Los estudios económicos de la educación (y del valor de la educación) son relativamente recientes. Becker ha sido uno de los microeconomistas más influyentes en el tema desde 1964 con su libro *El Capital Humano*. Becker interpretó a la educación como una inversión de las personas, debido a que el mercado laboral recompensa económicamente mayores niveles de educación (también la interpretó como un bien de consumo cuando las personas estudian por placer o sin esperar retornos futuros). La Teoría del Capital Humano de Becker ha sido tema de gran cantidad de estudios posteriores y se han generado otras teorías referentes al papel de la educación.

La Teoría del Capital Humano debe sus aportes a Becker (1964) y Mincer (1974), y posiblemente a los aportes anteriores de Schultz (1961). A grandes rasgos, la teoría expone que las inversiones en capital humano soportan unos costos directos e indirectos: educarse genera gastos en pago de matriculas, útiles

escolares, transporte y otros; y genera un costo de oportunidad al ocupar tiempo que podría dedicarse a otras actividades. Estos costos tienen como beneficio la adquisición de ciertas capacidades y conocimientos que son recompensados económicamente por el mercado laboral. De esta manera, un rasgo importante de esta teoría es que la educación aumenta la productividad de las personas con consecuencias tanto privadas como sociales.

Algunas críticas posteriores a la Teoría del Capital Humano están dirigidas a algunos de sus supuestos, pero no atacan las bases más fuertes de la teoría. Una de éstas es la Teoría de la Señalización (conocida también como credencialista, del filtro, certificación o, en inglés, Sheepskin Effects). Esta teoría es defendida por autores como Arrow (1973), Spence (1973), Taubman y Wales (1973) y Stiglitz (1975). A pesar de que se sigue viendo a la educación como una forma de capital, la diferencia radica en que la educación no aumenta la productividad de las personas, sino que simplemente resuelve un proceso de información asimétrica. La intuición es que los empleadores no conocen la productividad de los trabajadores, así que se basan en los títulos educativos para escoger a los más eficientes. Se supone que las personas tiene diferentes niveles de productividad, así que serán los más dedicados o inteligentes los que logren obtener el título o los que decidan obtener más educación debido a que sus costos son relativamente más bajos.

Existen dos ramas de la Teoría de la Señalización. La versión fuerte, que argumenta que la educación no incrementa absolutamente en nada la productividad de las personas y sólo se limita a diferenciar los estudiantes hábiles. La versión débil argumenta que la educación además de señalizar contribuye a aumentar la productividad del individuo.

Se debe tener en cuenta que la Teoría del Capital Humano infiere que la educación es una variable apropiada para explicar los salarios. Como más años de educación representan aumentos en la productividad y la productividad se valora en el mercado laboral, es necesario conocer sólo los años educación (sin darle mucha importancia al título) para conocer la productividad de los

trabajadores y consecuentemente su impacto en el salario. Esta simpleza de la teoría es la que genera las críticas posteriores. La Teoría fuerte de la Señalización infiere que la educación no determina la productividad y los retornos económicos son explicados por los títulos y no por los años de educación. La versión débil de esta teoría infiere que los títulos generan aumentos en el salario sólo al principio de la contratación y que con el paso del tiempo se revela la información de la productividad del trabajador al empleador y los salarios se ajustan de acuerdo a la productividad de cada trabajador (que puede ser innata o aprendida).

4 Estimación

En la práctica, se necesita saber cómo usar los paquetes estadísticos para hacer las estimaciones, cuál es la teoría o hipótesis que se va a probar y los problemas estadísticos o econométricos que se quieren corregir. En esta sección se estiman varias ecuaciones de salarios por Mínimos Cuadrados Ordinario (MCO), Variables Instrumentales (VI) y corrección del sesgo de selección por el método de Heckman. Se va a dar la explicación teórica básica seguida de algunos comentarios de creación de variables y de comandos de *Stata*. El **Anexo 1** es el archivo de trabajo (***do-file***) de *Stata*, con algunas explicaciones importantes, que permite replicar los resultados[3].

4.1 Estadísticas descriptivas

A la hora de realizar cualquier investigación empírica, lo primero que se debe hacer son las estadísticas descripticas que muestras algunas características importantes de la población a analizar. Para efectos de estimación, también ayudan a organizar y crear las variables que se van a usar en las regresiones.

LA GEIH es una encuesta, con relativamente pocas preguntas, que realiza el DANE, con una periodicidad mensual, con el objetivo específico de obtener información sobre el mercado laboral. Es una encuesta apropiada para ecuaciones de oferta de salarios. Se va a usar los datos de septiembre por ser un mes estable y se va a usar la muestra estratificada por áreas. Esto último quiere decir que estamos interesados en las trece áreas principales de Colombia. Se debe tener cuidado a la hora de interpretar los resultados porque la población que se va a estudiar no es todo el país.

El DANE dividió la encuesta en ocho archivos. Esto permite manejar los datos de forma sencilla porque cada uno contiene un subgrupo organizado de población. De estos ocho archivos, se van a utilizar por lo menos cinco (no es de nuestro

[3] Se espera que el lector que desea trabajar con Stata tenga un pequeño conocimiento del programa. Si no, con un tutorial básico de Stata y este *do-file*, tiene todas las herramientas para aprender. Se recomienda leer esta sección acompaña del *do-file*.

interés mirar los datos de inactivos, desempleados ni los de otras actividades). Para empezar de forma organizada, se debe mirar cada archivo, escoger las variables de interés y, por último, combinar en una sola base de datos las observaciones finales.

El primer archivo contiene información sobre viviendas y hogares[4]. Está compuesto por 9088 observaciones. Este archivo muestra información importante sobre condiciones de vida. Nos interesa de este archivo identificar el área y conocer algunas condiciones de los hogares. En la **Tabla 1** se muestra la proporción de áreas tenidas en cuenta por la encuesta, estimada con y sin factor de expansión. Al comparar las dos proporciones se aprecia de manera clara que con el factor de expansión la representatividad de cada hogar en la población es diferente.

Tabla 1. Proporción de áreas.

Área	Con expansión	Sin expansión
Cali	0,12	0,08
Ibagué	0,03	0,07
Bucaramanga	0,05	0,07
Pereira	0,03	0,07
Cúcuta	0,04	0,06
Pasto	0,02	0,06
Villavicencio	0,02	0,08
Monteía	0,01	0,06
Manizales	0,02	0,08
Cartagena	0,04	0,07
Bogotá	0,38	0,10
Barranquilla	0,07	0,09
Medellín	0,17	0,11
Total	1,00	1,00

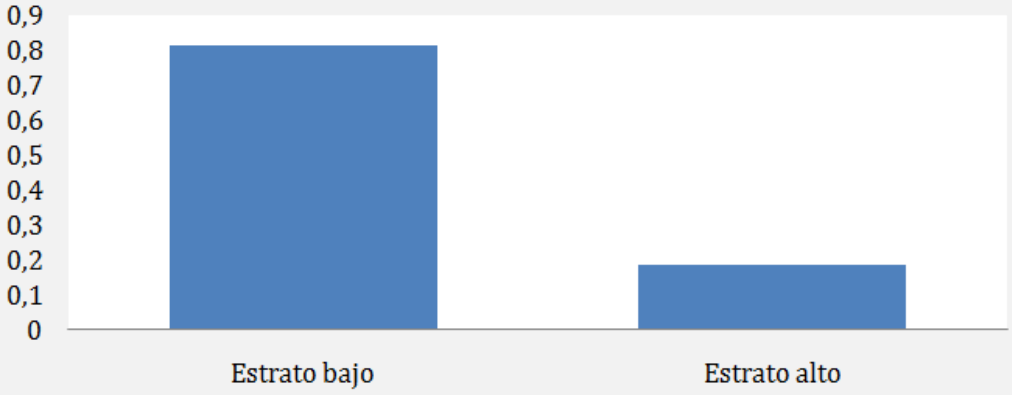
FUENTE: DANE. Cálculos propios.

La **Gráfica 1** muestra dos grupos de población por estrato[5]. Se han hecho de acuerdo al estrato de los servicios públicos (estrato alto si la vivienda pertenece

[4] Se debe recordar que por ser el archivo de hogares, es el archivo que contiene menos observaciones.
[5] Recomiendo al lector que diseñe las gráficas y tablas en el programa que conozca mejor y con el que pueda ser más eficiente.

por lo menos al estrato cuatro). Si bien ésta no es una medida directa de pobreza porque, por cualquier motivo, las personas pueden vivir en un estrato que no es consistente con su capacidad de pago, en términos generales podría ser una buena aproximación. La columna *Estrato bajo* de esta gráfica muestra que la gran mayoría de hogares viven en un estrato inferior al cuatro.

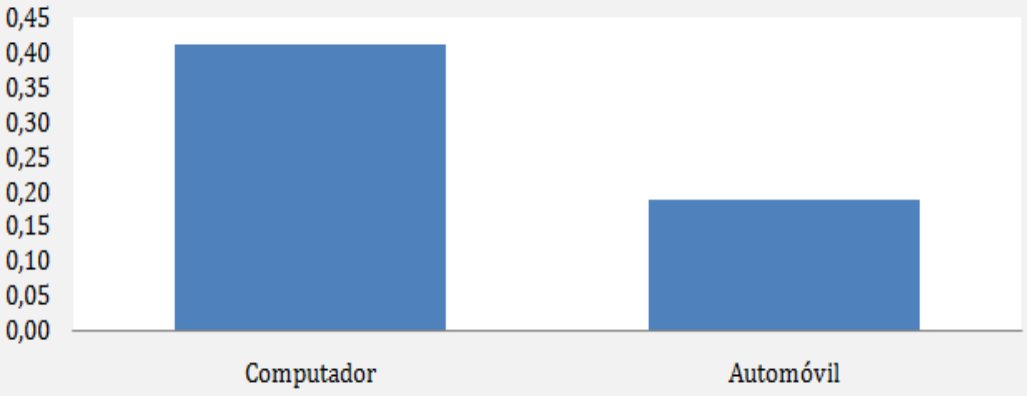
Gráfica 1. Proporción de hogares pertenecientes a estrato bajo y alto



FUENTE: DANE. Cálculos propios.

La **Gráfica 2** muestra la proporción de hogares con computador personal y con automóvil. Éstas pueden ser otras m**e**didas de bienestar.

Gráfica 2. Proporción de familias con computador y automóvil

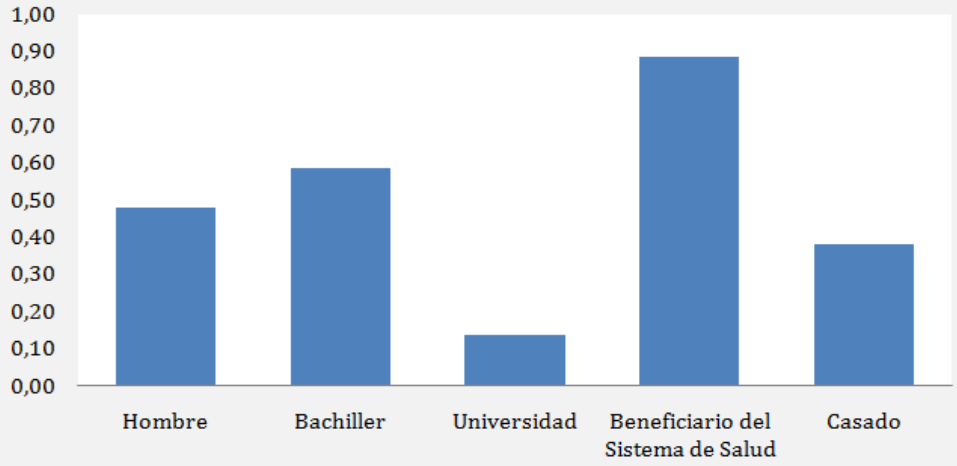


FUENTE: DANE. Cálculos propios.

El segundo archivo de datos contiene información sobre las características de las personas. Está compuesto por 31982 observaciones. Éste es el archivo con más datos porque contiene información de todas las personas del hogar. La **Gráfica 3**

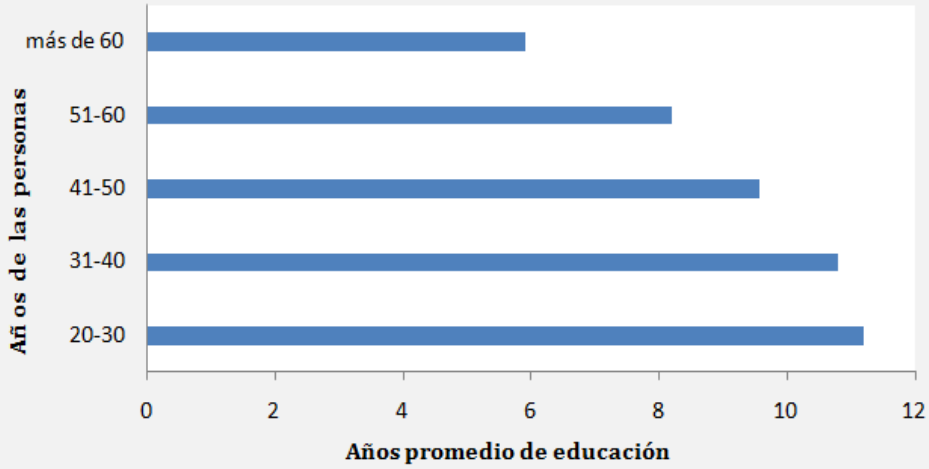
muestra algunas proporciones que se van a necesitar más adelante. Por el momento, note que la afiliación al sistema de salud es casi universal y que tan sólo un poco más del 10% de la población tiene educación universitaria. La **Gráfica 4** muestra los años de educación según tres rangos de edad. Se aprecia que la educación promedio de las personas ha aumentado con el tiempo.

Gráfica 3. Algunas proporciones para todas las personas



FUENTE: DANE. Cálculos propios.

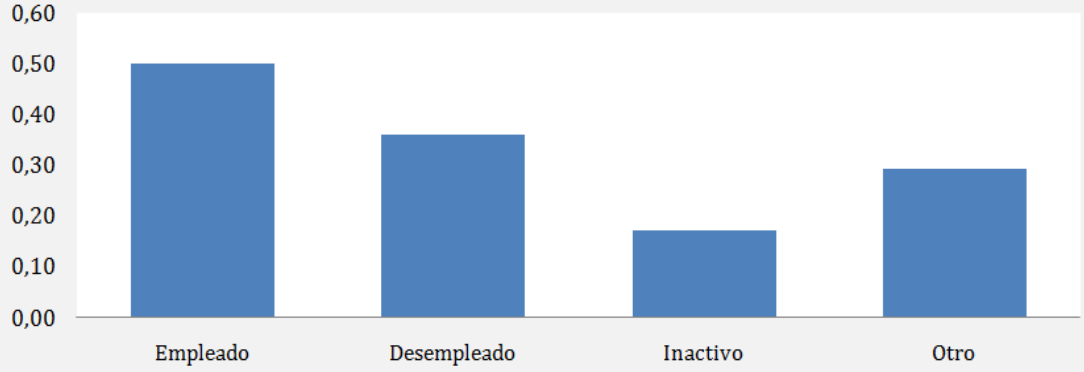
Gráfica 4. Promedio de años de educación por rangos de edad



FUENTE: DANE. Cálculos propios.

El tercer archivo de datos contiene información sobre la fuerza de trabajo. Contiene 25752 observaciones. La diferencia con el anterior es que éste excluye a las personas que no están en edad de trabajar (menores de 14 años). Va a ser útil para conocer la cantidad de empleados, desempleados e inactivos. La **Gráfica 5** muestra esta información. La columna *Otro* de esta gráfica se refiere a personas que ocuparon la mayor parte del tiempo la semana pasada a la encuesta en oficios del hogar, incapacitadas, pensionadas u otras actividades. Note que la mitad de las personas con edad para trabajar se encuentran efectivamente trabajando. (No confunda esto con la Tasa de Empleo).

Gráfica 5. Composición por actividad de la fuerza de trabajo

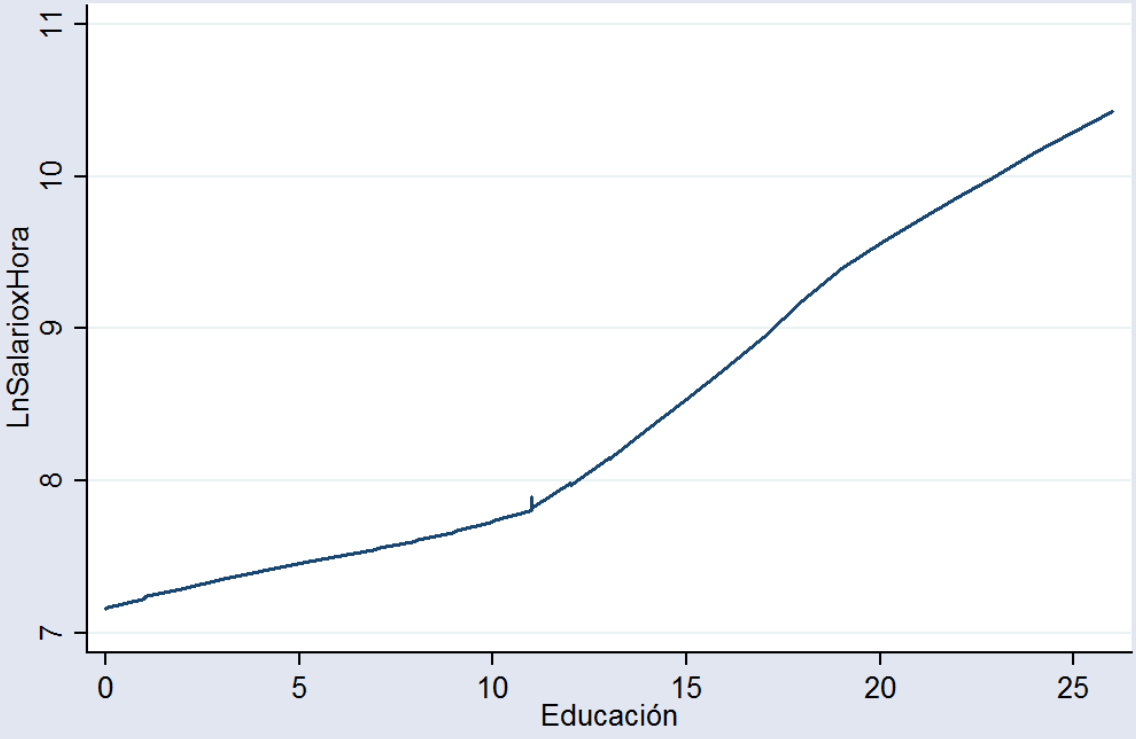


FUENTE: DANE. Cálculos propios.

El cuarto archivo de datos contiene información sólo sobre los ocupados. Contiene 14424 observaciones. Este archivo proporciona información sobre salario e ingresos, tipo y calidad del empleo y horas trabajadas. En este momento ya se puede saber algo sobre el objetivo a realizar, a saber, conocer el efecto de la educación sobre el salario. La Gráfica 6 muestra la correlación entre estas dos variables. Como se esperaba, la relación es positiva. No está de más recordar que para demostrar causalidad se debe tener un marco teórico muy fuerte y que la razón por la que se usan las regresiones es para separar los efectos de posibles variables que afectan la relación que muestra la **Gráfica 6.** Sin embargo, ésta

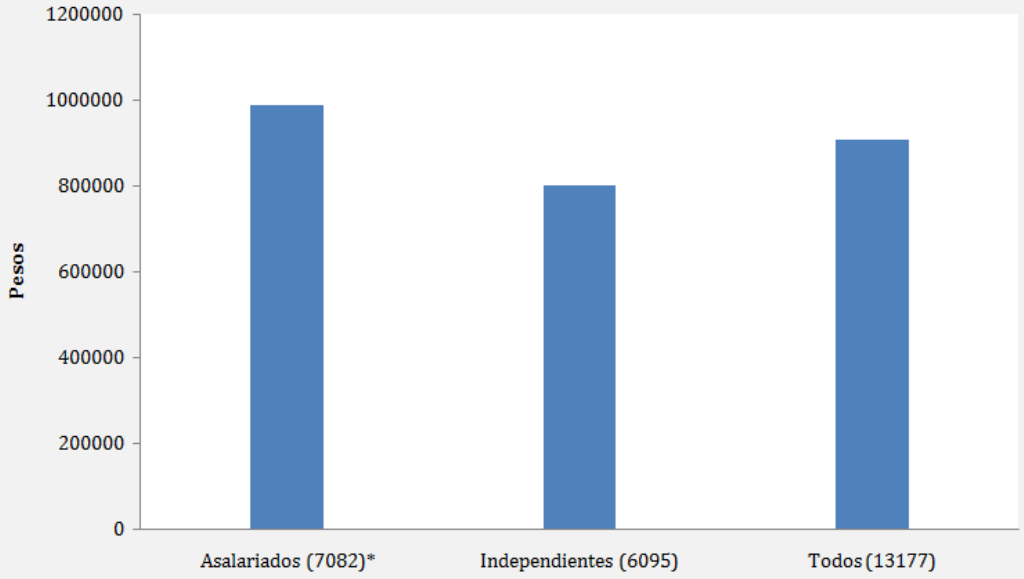
brinda un buen acercamiento. La **Gráfica 7** muestra que los asalariados ganan, en promedio, más que los independientes. Esto sugiere que entre los independientes debe existir una importante informalidad laboral. Note que el número de personas que reportan algún tipo de ganancia presenta una diferencia de 1247 con el número de observaciones de este archivo de datos (son personas que se consideran ocupadas pero no reciben compensación monetaria). Este archivo también permite conocer que del total de ocupados, el 60,2% tiene algún tipo de contrato, y de éstos, sólo el 64,4% tiene contrato escrito. También se encuentra que el salario por hora de todos los ocupados es 4732,2 pesos. Conocer el salario por hora es importante porque es una medida más exacta de la verdadera ganancia de las personas (puede ser mejor ganarse diez pesos en una hora que once en dos).

Gráfica 6. Correlación entre educación y Ln(Salario por hora)



FUENTE: DANE. Cálculos propios

Gráfica 7. Ganancia media de asalariados e independientes



*Número de personas
FUENTE: DANE. Cálculos propios.

Ya se tiene toda la información necesaria para conseguir el objetivo. Las otras bases de datos que contiene la GEIH son las que tienen información sobre desocupados, inactivos, otras actividades y otros ingresos. Se va a utilizar esta última base de datos para conocer cómo afecta la educación no sólo al salario sino a todos los otros ingresos. Este archivo tiene 25752 observaciones al igual que el tercer archivo. Esto quiere decir que contiene información sobre toda la fuerza de trabajo.

Sin tener en cuanta de dónde vienen los ingresos, se van a agregar todos con una periodicidad mensual. Sin incluir la ganancia del empleo principal, el promedio de los otros ingresos, para las 7156 personas que los reciben, es 709822 pesos. Esto muestra que es una cantidad importante.

Para terminar las estadísticas descriptivas, en el **Anexo 2** se muestra un resumen con los datos numéricos de todos los resultados anteriores, y en el **Anexo 3** se muestran las correlaciones de las variables que se han formado. Las correlaciones son importantes para saber cuáles *dummies* se pueden incluir en las regresiones

(no queremos modelos con problemas de multicolinealidad). Se presenta una alta correlación entre posesión de computador y automóvil, y entre éstas y los años de educación.

4.2 Regresiones

La estimación más sencilla que se podría hacer es una regresión lineal simple (MCO) que relacione el salario con la educación:

$$\boldsymbol{Ln(Wh_i) = \alpha_1 + \alpha_2 edu_i + \mu_i} \qquad \boldsymbol{(1)}$$

Donde $\boldsymbol{Ln(Wh_i)}$ representa el logaritmo natural del salario por hora de $\boldsymbol{i}$, $\boldsymbol{edu_i}$ representa los años de educación de $\boldsymbol{i}$. El subíndice $\boldsymbol{i}$ muestra que se trabaja con datos de corte transversal y que los datos corresponden al individuo $\boldsymbol{i}$, donde $\boldsymbol{i = 1, 2, \dots, n}$. Se utiliza el logaritmo del salario para que la interpretación de los coeficientes tenga un significado intuitivo a primera vista. Como se espera que el salario guarde una relación positiva con la educación, se espera que $\boldsymbol{\alpha_2}$ sea positivo. La ecuación (1) podría no ser una buena descripción porque posiblemente se esté sobreestimando $\boldsymbol{\alpha_2}$ si existen otras variables que ayuden a determinar el salario.

La ecuación más conocida de este tipo de investigaciones es la Ecuación de Mincer (1974) que se basa en la Teoría de Capital Humano:

$$\boldsymbol{Ln(Wh_i) = \alpha_1 + \alpha_2 edu_i + \alpha_3 edad_i + \alpha_4 edad_i^2 + \mu_i} \qquad \boldsymbol{(2)}$$

Donde, además de lo anterior, $\boldsymbol{edad_i}$ representa los años de las personas. La ecuación original de Mincer no usa la edad sino la experiencia potencial (la edad de la persona menos seis años menos los años de educación). Como la experiencia potencial sugiere que cuando las personas no son niños o se encuentran estudiando, están trabajando, y esto puede no ser verdad, para simplificar se usa simplemente la edad de la persona. De esta manera $\boldsymbol{\alpha_3}$ y $\boldsymbol{\alpha_4}$ recogen efectos que no tenía la ecuación (1). Se espera que $\boldsymbol{\alpha_3}$ sea positivo porque mayor edad puede significar mayor experiencia y ésta la debe valorar el

mercado laboral. Y se espera que $\boldsymbol{\alpha_4}$ sea negativo porque la tendencia de $\boldsymbol{\alpha_3}$ debe invertirse cuando la persona envejece.

Otra ecuación muy usada es la que intenta conocer el efecto de los títulos educativos sobre el salario. Se debe separar el efecto de los años acumulados de educación de los años en que se recibe título. Siguiendo la ecuación de Hungerford y Solon (1987)[6]:

$$\boldsymbol{Ln(Wh_i) = \alpha_1 + \alpha_2 edu_i + \alpha_3 edad_i + \alpha_4 edad_i^2}$$
$$\boldsymbol{+ \beta_1 S11_i + \beta_2 S11_i(edu_i - 11) + \beta_3 S16_i + \beta_4 S16_i(edu_i - 16) + \mu_i} \qquad \boldsymbol{(3)}$$

Lo nuevo en esta ecuación son los $\boldsymbol{\beta_i}$. Éstos son coeficientes de variables *dummy* que hacen explícito los años de educación en que se recibe título. De esta manera, $\boldsymbol{S11_i}$ es una *dummy* que toma el valor de uno si $\boldsymbol{i}$ tiene once años de educación, y cero en caso contrario. $\boldsymbol{S16_i}$ es una *dummy* que toma el valor de uno si $\boldsymbol{i}$ tiene más de quince años de educación, y cero en caso contrario. Si existe el efecto de los títulos, $\boldsymbol{\beta_1}$ y $\boldsymbol{\beta_3}$ deben ser positivos. El rendimiento estimado del bachillerato completo es $\boldsymbol{\alpha_2 + \beta_1}$, y el rendimiento estimado de la educación universitaria completa es $\boldsymbol{\alpha_2 + \beta_1 + \beta_3}$. La existencia de $\boldsymbol{\beta_2}$ y $\boldsymbol{\beta_4}$ recoge los efectos de los años de educación por medio de interacciones con la variable $\boldsymbol{edu_i}$. En la **Tabla 2** se muestran las estimaciones de las ecuaciones (1), (2) y (3). Las estimaciones sugieren que en Colombia podría cumplirse la Teoría fuerte de la Señalización. Note que al capturar el efecto de los años en que se obtiene título, pierde poder de explicación los años de educación. Además, el efecto de los títulos tiene una importante implicación en el salario, en especial el título universitario (tenerlo representa un aumento promedio alrededor de 25%).

[6] La ecuación de Hungerford y Solon para Estados Unidos planteó discontinuidades en los salarios en los años de educación 8, 12 y 16. En Colombia es interesante mirarlas en los años 11 y 16.

Tabla 2. Regresiones (1), (2) y (3)

Variable dependiente: Ln(Wh) (Estadístico t entre paréntesis)			
Variable	**(1)**	**(2)**	**(3)**
Constante	**6,8510**	**5,6161**	**6,1291**
	(397,36)	(92,38)	(39,85)
Edu	**0,1017**	**0,1075**	**0,0515**
	(62,55)	(65,27)	(30,23)
Años		**0,0515**	**0,0477**
		(16,05)	(5,32)
Años^2		**-0,0005**	**-0,0005**
		(-11,94)	(-2,89)
S11			**0,0630**
			(2,98)
S11(Edu-11)			**0,1231**
			(-4,74)
S16			**0,1486**
			(4,49)
S16(Edu-16)			**-0,1279**
			(-5,16)
N	12972	12972	12972
R^2	0,2582	0,2999	0,3375
F	3912,16	1586,75	848,36

Nota: Corrección de heteroscedasticidad con error estándar robusto deWhite
Todos los coeficientes significativos al 1%.
FUENTE: DANE. Cálculos propios

4.2.1 Método de Variables Instrumentales (VI)

El método de variables instrumentales se usa para corregir el problema de endogeneidad de una o varias variables explicativas. Particularmente se va a usar para generar estimadores consistentes en presencia de variables omitidas para la ecuación (2) o (3). En realidad, aquí ya se hace interesante empezar a usar

dummies que se crearon en la sección pasada para conocer los efectos sobre el salario de algunas variables socioeconómicas y separarlas del efecto de la educación. Una crítica común a las ecuaciones anteriores es que no tienen en cuenta otras variables que afecten el salario. Si sólo se explica éste por medio de la educación y la edad, pueden sobreestimarse sus efectos (como se vio con la ecuación (1)). Un determinante importante de la diferencia de los salarios entre las personas puede ser la habilidad innata de éstas. Se puede pensar que las personas tienen diferencias en dicha habilidad y que la educación ayuda un poco a desarrollarlas (Teoría débil de la Señalización). En este caso la educación puede actuar como señal para los empleadores, pero no es la única explicación para los salarios. Lastimosamente, la habilidad innata no se puede observar.

Se podría encontrar una *proxy* para la habilidad, por ejemplo, alguna medición del coeficiente intelectual de las personas. Sin embargo, habría que utilizar una encuesta muy especializada en el tema de educación que contenga esta medición y, aún teniéndola, puede no ser apropiada. El lector debe conocer personas no muy brillantes que, por su dedicación o esfuerzo, logran ser muy productivas. Otra opción interesante es usar datos de panel. Como la habilidad innata debería ser una característica que no cambia en el tiempo, se podrían usar primeras diferencias o utilizar el método de efectos fijos para separar el efecto de la educación. Este método no se va a usar porque no se cuenta con datos de panel[7]. Se remite al lector a Mora y Muro (2007) que, si bien no usan datos de panel, usan una metodología con datos de diferentes años para estimar dichos efectos.

Otra opción es no corregir el problema, pero tener en cuenta en la interpretación que los estimadores son sesgados e inconsistentes. Si se sabe que alguna variable está sesgada hacia cero, pero se encuentra que es positiva, no se puede confiar completamente en el número del coeficiente, pero la estimación estaría mostrando algo importante: que el efecto es positivo. Lastimosamente, éste no es el caso de las ecuaciones que se están tratando. Omitir la habilidad de las

[7] La desventaja de los datos de panel es que hacerlos aumenta considerablemente los costos de la encuesta. Ésta es la razón por la que en los países en desarrollo no se cuenta con muchos de estos datos. El DANE en la última Encuesta Nacional de Calidad de Vida (2010) hizo una submuestra con datos de panel (hogares encuestados en la encuesta de 2008).

personas como explicación de los salarios sobreestima la influencia de la educación. Como se espera que el coeficiente sea positivo, y si lo importante es conocer el valor del coeficiente, la estimación no enseñará mucho.

El método **VI** consiste en encontrar una variable exógena a la ecuación que no esté correlacionada con la variable no observada y que sí esté correlacionada con la variable que se cree que está sesgada. En este caso en particular, se necesita una variable $\boldsymbol{Z_i}$ tal que $\boldsymbol{Cov(Z_i, \mu_i) = 0}$ y $\boldsymbol{Cov(Z_i, \alpha_2) \neq 0}$**.** La primera de estas dos ecuaciones tiene sentido porque al no observarse la variable habilidad, se sabe que es recogida por el error. Un instrumento muy utilizado en este tema es la educación del padre (o de la madre). Ésta debería afectar la educación de $\boldsymbol{i}$ y no tener efecto parcial sobre $\boldsymbol{Wh_i}$ (o no estar correlacionada con la habilidad). Sin embargo, podría no ser un buen instrumento porque no es claro que la educación de los padres no tenga alguna influencia sobre la habilidad de los hijos (sea por razones genéticas o de crianza). Se va a usar la variable instrumental, $\boldsymbol{Z_i}$**,** utilizada por Zuluaga (2009). $\boldsymbol{Z_i}$ es una *dummy* que toma el volar de uno para personas que nacieron después del año 1951 y cero en caso contrario. La argumentación que ella brinda para utilizar esta variable como instrumento es que en ese año el Gobierno hizo obligatoria la educación primaria, y se vieron cambios significativos en la escolaridad, tanto en ese año como en los siguientes[8]. Recuerde que en la **Gráfica 4** se mostraba que las personas jóvenes tienen en promedio más años de educación. No existe una forma sencilla de mostrar que $\boldsymbol{Cov(Z_i, \mu_i) = 0}$. Se debe hacer con una argumentación teórica. Con esta variable instrumental la argumentación es sencilla porque es claro que la habilidad se distribuye en las personas independientemente del año en que nazcan. Para mostrar que $\boldsymbol{Cov(Z_i, \alpha_2) \neq 0}$ se debe estimar por MCO:

$$\boldsymbol{edu_i = \pi_1 + \pi_2 Z_i + \pi_3 edad_i + \pi_4 edad_i^2 + \varepsilon_i} \qquad \boldsymbol{(4)}$$

[8] Se debe mencionar que esta variable puede no ser un buen instrumento si el Estado no tiene la capacidad de controlar que efectivamente todos los niños estén yendo a la escuela. Aunque en Colombia el acceso a la educación (en especial primaria) ha mejorado increíblemente en los últimos años y está cerca de la universalización, el control podría ser difícil de implementar en zonas rurales o muy pobres.

Sí $\boldsymbol{\pi_2}$ es estadísticamente diferente de cero, entonces se demuestra que $\boldsymbol{Cov(Z_i, \alpha_2) \neq 0}$. Al hacerlo con varias *dummies* como variables explicativas, se genera un $\boldsymbol{\pi_2}$ positivo y significativo al 1% (el lector lo puede comprobar con el ***do-file***) Como $\boldsymbol{Z_i}$ cumple los dos requisitos para **VI**, se puede pensar que es un buen instrumento.

Las implicaciones estadísticas de **VI** son tenidas en cuenta por los programas econométricos al igual que lo hacen con **MCO**. *Stata* lo hace automáticamente con el comando *ivregress*.

En la **Tabla 3** se muestra una estimaciones **VI** utilizando $\boldsymbol{Z_i}$ como variable instrumental. Para comparar la estimación **VI**, la Tabla 3 muestra también una estimación **MCO** teniendo en cuenta el efecto señalización, las *dummies* de área, una *dummy* del género y una *dummy* del estado civil. Esta estimación **MCO** hace aún más pequeño el efecto de la educación (comparándolo con las estimaciones anteriores). Muestra que ser hombre representa un aumento en el salario, en promedio, de alrededor 26%. El signo negativo de las variables para las áreas son resultado de que la *dummy* omitida fue la de Bogotá. Como esta ciudad es la que presenta salarios en promedio más altos, su efecto es recogido por el intercepto (que es positivo) y los coeficientes de las otras ciudades se interpretan con respecto a ésta. De esta manera, Cali y Medellín son las ciudades en donde menos caen los salarios respecto a Bogotá.

La estimación **VI** hace no significativo el efecto de la educación y aumenta, respecto a la estimación **MCO**, la importancia del título universitario. **VI** no permite conocer el efecto de la habilidad, pero permite por medio del instrumento descontar el sesgo que podría generar no incluir dicha variable en la regresión. De esta manera, confirma la sospecha (de la ecuación (3)) de que los años de educación no tienen un efecto en el salario y de que, posiblemente, la habilidad de las personas que terminaron sus estudios universitarios es la que es recompensada por el mercado. Note que las variables que identifican las áreas y el estado civil prácticamente no cambiaron con **VI.** Esto tiene sentido porque estas

variables no están relacionadas con la habilidad. En cambio, el efecto de ser hombre se redujo de forma importante al incluir la variable instrumental.

Tabla 3. Regreses MCO y VI

Variable dependiente: Ln(Wh) (Error estándar entre paréntesis)		
Variable	**(MCO)**	**(VI)**
Constante	**6,1397** (0,0574225)	**5,9357** (0,6876872)
Edu	**0,0484** (0,0037683)	**0,0824** (0,0861725)*
Años	**0,0464** (0,0026132)	**0,0457** (0,0031789)
Años²	**-0,0005** (0,0000311)	**-0,0005** (0,0000641)
S11	**0,0923** (0,0242688)	**-0,0815** (0,4408898)*
S11(Edu-11)	**0,1269** (0,0055485)	**0,0921** (0,0884249
S16	**0,1851** (0,0272057)	**0,1850** (0,0272695)
S16(Edu-16)	**-0,1555** (0,0219748)	**-0,1547** (0,022115)
Cali	**-0,1034** (0,0255531)	**-0,1055** (0,0389443)
Ibagué	**-0,2868** (0,0275208)	**-0,2841** (0,0422787)
Bucaramanga	**-0,0030** (0,026312)*	**-0,0030** (0,0394615)*
Pereira	**-0,1265** (0,0267689)	**-0,1235** (0,0413827)
Cúcuta	**-0,1373** (0,0290577)	**-0,1260** (0,057757)*
Pasto	**-0,4193** (0,0290378)	**-0,4102** (0,0530881)
Monteía	**-0,3755** (0,02732)	**-0,3708** (0,0439382)
Manizales	**-0,1369** (0,0274516)	**-0,1393** (0,0419989)
Cartagena	**-0,1636** (0,0286947)	**-0,1631** (0,0430567)
Barranquilla	**-0,3701** (0,0255621)	**-0,3695** (0,0384001)
Medellín	**-0,1040** (0,0232904)	**-0,1016** (0,0358515)
Hombre	**0,2642** (0,012505)	**-0,1016** (0,0188075)
Estado Civil	**0,0828** (0,0131483)	**-0,0803** (0,0214005)
N	12971	12971
R^2	0,3883	
F	411,06	
Wald		3603,43

Nota: Corrección de heteroscedasticidad con error estándar robusto deWhite

*** no significativos al 10%.**

Educación instrumentado con D1951

FUENTE: DANE. Cálculos propios

4.2.2 Corrección del sesgo de selección por el método de Heckman.

Existe otro problema de endogeneidad que debería corregirse. Las estimaciones anteriores intentan conocer el efecto de la educación en el salario de todas las personas en edad de trabajar. Sin embargo, en las estimaciones se ha excluido de la muestra a los desempleados e inactivos. Esto se ha hecho porque estas personas no perciben salario y, por lo tanto, no tienen información para esta variable. Al hacer esto, se están haciendo estimaciones para un grupo de personas que no han sido seleccionadas aleatoriamente. Este problema se conoce como truncamiento incidental y es un caso particular de los problemas de selección muestral. Sólo se observa la variable $\boldsymbol{Wh_i}$ para un subgrupo de la población. Observar o no $\boldsymbol{Wh_i}$ no depende de las variables del modelo, sino de una variable externa (que es la participación en el mercado laboral).

El método de Heckman proporciona una forma de obtener estimaciones consistentes sólo con los datos observados. Este método convierte al sesgo de selección en un problema de omisión de una variable que se puede obtener. Esta variable es llamada el Cociente inverso de Mills. Para obtener estimaciones consistentes para la ecuación (2), se le debe agregar el Cociente inverso de Mills como variable explicativa.

Para conocer esta variable se debe estimar un modelo *probit* que es llamado la ecuación de selección. La variable dependiente de esta ecuación, $\boldsymbol{S_i}$, es una variable *dummy* que toma el valor de 1 si se observa $\boldsymbol{Wh_i}$ y cero si no. Esto quiere decir que la muestra no son sólo las personas que perciben salario, sino todas las personas en edad de trabajar (si las observaciones de $\boldsymbol{S_i}$ fueran las mismas de todas las ecuaciones anteriores, $\boldsymbol{S_i}$ estaría compuesta sólo de unos). Las variables independientes de la ecuación de selección, $\boldsymbol{Z_i}$, deben ser todas las de la ecuación (2) y otras variables que estén relacionadas con la decisión de pertenecer (o poder estar en) el mercado laboral y que no estén relacionadas con $\boldsymbol{Wh_i}$**.** Se necesita que las variables independientes de la ecuación de selección sean las mismas que la ecuación (2) para generar consistencia entre las dos

ecuaciones; y se necesitan las que no están en la ecuación (2) para no generar algún problema de multicolinealidad en la estimación de la ecuación que contiene el Cociente inverso de Mills. Estas otras variables de la Ecuación de selección pueden no ser muy fáciles de conseguir y se puede prescindir de ellas si no existe una fuerte correlación entre el Cociente inverso de Mills y las variables de la ecuación (2).

El Cociente inverso de Mills es simplemente $\boldsymbol{S_i}$ estimado, $\boldsymbol{\widehat{S_i}}$**.** Si al estimar la ecuación (2) con $\boldsymbol{\widehat{S_i}}$ se encuentra que el coeficiente que acompaña este último término, $\boldsymbol{\rho_1}$, es significativamente diferente de cero, se muestra que la ecuación efectivamente presentaba un problema de sesgo de selección y que era necesario usar el método de Heckman. Para resumir, los pasos son:

$$\boldsymbol{Paso\ 1}: \quad \boldsymbol{S_i = \Phi(Z\gamma)}$$

$$\boldsymbol{Paso\ 2}: \quad \boldsymbol{Ln(Wh_i) = \alpha_1 + \alpha_2 edu_i + \alpha_3 edad_i + \alpha_4 edad_i^2 + \rho_1 \widehat{S_i} + \mu_i} \qquad (\mathbf{5})$$

Corregir el sesgo de selección por este método es relativamente fácil. Sin embargo, al igual que con **VI**, los paquetes estadísticos hacen el procedimiento en un sólo paso. *Stata* tiene el comando ***heckman*** que realiza el procedimiento automáticamente. Si al final del comando se agrega la opción ***twostep***, además de estimar, brinda el Cociente del inverso de Mill que, como se acaba de ver, su significancia es una prueba de la existencia del sesgo de selección. En la **Tabla 4** se muestran la estimación corregida por el método de Heckman y, otra vez, la estimación **MCO.**

La **Tabla 4** muestra que el sesgo de selección existía y era necesario corregirlo. La estimación sin sesgo de selección tiene unos coeficientes relativamente parecidos a la estimación **MCO**. Se destaca que el coeficiente de los años adicionales de educación es un poco más grande en la estimación por Heckman que en la MCO y que pasa lo contrario con el título universitario.

Tabla 4. Regreses MCO y corregida por el Método de Heckman

Variable dependiente: Ln(Wh) (Error estándar entre paréntesis) Variable	(MCO)	(Heckman)
Constante	**6,1397** (0,0574225)	**6,6778** (0,3877564)
Edu	**0,0484** (0,0037683)	**0,0488** (0,0038041)
Años	**0,0464** (0,0026132)	**0,0255** (0,0151559)*
Años²	**-0,0005** (0,0000311)	**-0,0002** (0,0001787)*
S11	**0,0923** (0,0242688)	**0,0674** (0,030208)
S11(Edu-11)	**0,1269** (0,0055485)	**0,1256** (0,0056765)
S16	**0,1851** (0,0272057)	**0,1774** (0,0280458)
S16(Edu-16)	**-0,1555** (0,0219748)	**-0,1521** (0,0223389)
Cali	**-0,1034** (0,0255531)	**-0,0962** (0,0262704)
Ibagué	**-0,2868** (0,0275208)	**-0,2829** (0,0278823)
Bucaramanga	**-0,0030** (0,026312)*	**-0,0114** (0,0271944)*
Pereira	**-0,1265** (0,0267689)	**-0,1175** (0,0277282)
Cúcuta	**-0,1373** (0,0290577)	**-0,1140** (0,0336486)
Pasto	**-0,4193** (0,0290378)	**-0,3982** (0,032889)
Monteía	**-0,3755** (0,02732)	**-0,3740** (0,0275666)
Manizales	**-0,1369** (0,0274516)	**-0,1134** (0,0323144)
Cartagena	**-0,1636** (0,0286947)	**-0,1236** (0,0405902)
Barranquilla	**-0,3701** (0,0255621)	**-0,3326** (0,037125)
Medellín	**-0,1040** (0,0232904)	**-0,1038** (0,0232908)
Hombre	**0,2642** (0,012505)	**0,2124** (0,0390433)
Estado Civil	**0,0828** (0,0131483)	**0,0820** (0,0132814)
N	12971	25738
R^2	0,3883	
F	411,06	
Wald		5830,89
Mills		-0,1831631 (0,1304482)

Nota: Corrección de heteroscedasticidad con error estándar robusto deWhite

*** no significativos al 5%.**

FUENTE: DANE. Cálculos propios

4.2.3 Variable dependiente: Todos los ingresos

Como última estimación, para explorar un poco más los efectos de la educación, la **Tabla 5** muestra tres estimaciones con todos los ingresos mensuales como variable dependiente. Los resultados pueden ser interesantes porque se podría pensar que los efectos positivos de la educación están asociados sólo al empleo formal y que otro tipo de ingresos podrían no presentar la correlación positiva de las estimaciones anteriores.

Tabla 5. Regreses MCO, VI y Heckman

Variable dependiente: Ln(Ingresos Totales) (Error estándar entre paréntesis)			
Variable	**(MCO)**	**(VI)**	**(Heckman)**
Constante	**10,1192** (0,1033622)	**10,5653** (0,9126103)	**13,6121** (0,5064684)
Edu	**0,0662** (0,0072982)	**-0,0042** (0,0035465)*	**0,0568** (.0075621)
Años	**0,0862** (0,0045089)	**0,0862** (0,0000678)	**-0,0178** (.0153716)*
Años²	**-0,0008** (0,0000511)	**-0,0008** (0,746968)	**0,0002** (.0001453)*
S11	**0,1189** (0,0452016)	**0,4852** (0,4408898)*	**-0,1876** (0,0668507)
S11(Edu-11)	**0,0718** (0,0110327)	**0,1439** (0,1472672)*	**0,0705** (0,0120975)
S16	**0,1515** (00578867)	**0,1559** (0,0272695)	**0,1495** (0,0643065)
S16(Edu-16)	**-0,1677** (0,0578434)	**-0,1703** (0,0416931)	**-0,1581** (0,0509258)
Cali	**-0,3319** (0,0493824)	**-0,3222** (0,0515255)	**-0,3417** (0,0554017)
Ibagué	**-0,2776** (0,0371571)	**-0,2797** (0,0505918)	**-0,3365** (0,0595651)
Bucaramanga	**-0,0220** (0,0388583)*	**-0,0245** (0,0493439)*	**-0,2011** (0,0632551)
Pereira	**-0,1362** (0,0358175)	**-0,1482** (0,546)	**-0,1716** (0,0570612)
Cúcuta	**-0,3774** (0,0523716)	**-0,4078** (0,0820244)*	**-0,2227** (0,0651777)
Pasto	**-0,9049** (0,0742837)	**-0,9222** (0,0637733)	**-0,7963** (0,0631284)
Monteía	**-0,3303** (0,0366676)	**-0,3424** (0,0577745)	**-0,2790** (0,0609401)
Manizales	**-0,1421** (0,0338674)	**-0,1330** (0,0531093)	**-0,0536** (0,0591028)*
Cartagena	**-0,7032** (0,0730032)	**-0,7011** (00527529)	**-0,4076** (0,0731335)
Barranquilla	**-0,7356** (0,0593573)	**-0,7346** (0,0482995)	**-0,3562** (.0765559)
Medellín	**-0,1474** (0,0358362)	**-0,1509** (0,0443731)	**-0,1482** (0,0440776)
Hombre	**0,5082** (0,0228561)	**0,5081** (0,0230823)	**0,2148** (0,0496752)
Estado Civil	**0,1119** (0,0242559)	**0,1169** (0,0264481)	**0,1675** (0,0300965)
Mills			**-1,7741** (0,2473486)
N	16846	16846	25738
R^2	0,1797		
F	145,11		
Wald		3564,79	846,95

Nota: Corrección de heteroscedasticidad con error estándar robusto deWhite
*** no significativos al 5%.**
Educación instrumentado con D1951
FUENTE: DANE. Cálculos propios

La **Tabla 5** muestra que el título universitario también tiene efecto positivo en todos los ingresos y que los años adicionales de educación sin título pierden importancia en dicha explicación cuando se corrigen problemas de endogeneidad y se captura el efecto de los títulos educativos. Al igual que con el empleo principal, el título de bachillerato no parece tener mucha importancia. Llama la atención de las estimaciones con esta variable dependiente que ser hombre reporta un aumento en los ingresos mucho más grande que las estimaciones sólo para el salario.

5. Conclusiones

A pesar de que las estimaciones pueden tener sesgos y de que no se han corregido los problemas de endogeneidad en una misma regresión, se ha aprendido algo para Colombia:

(1) El título universitario repercute indudablemente de forma positiva, tanto en el salario como en los ingresos.

(2) Al capturar el efecto de los títulos educativos y corregir los problemas de endogeneidad, pierden importancia los retornos económicos de los años adicionales de educación. Parece probarse la teoría fuerte de la señalización.

(3) Existe una fuerte diferencia de género que perjudica a las mujeres en el mercado laboral.

(4) Vivir en pareja es bien visto por el mercado laboral.

(5) Bogotá, Cali y Medellín son las ciudades del país con los salarios más altos.

6. Referencias

- Deaton, A. (2000). *The analysis of Household Surveys: A Microeconometrical Approach to Development Policy*. World Bank. The Johns Hopkins University Press. Baltimore, Maryland, U.S.A.
- González, C. (2009). Desarrollos recientes sobre la demanda de educación y sus aplicaciones empíricas internacionales. *Borradores de Economía y Finanzas*. No. 19. Universidad Icesi.
- Mora, J; Muro, J. (2007). Sheepskin effecs by cohorts in Colombia. Emerald.
- Zuluaga Días, Blanca. (2010). Different Impact Channels of Education on Poverty. *Estudios Gerenciales.* Universidad Icesi.
- Wooldridge, J. (2001). *Introducción a la econometría. Un enfoque moderno*. Segunda edición. Thomson.
- Departamento Administrativo Nacional de Estadística: www.dane.gov.co *Consultado en 23/04/2012*.

7. Anexos

Anexo 1. Stata do-dile

```
*ESTIMACIÓN
clear
set more off
clear matrix
cap log close
set mem 700m
* cd sirve para no volver a escribir la dirección completa
cd "C:\Users\1144136015\Desktop\GEIH 2010\09\09\areas\dat\reg\"
*Abrir el archivo de hogares
use Sas01area1009.dta
*dpto: departamento
*P4030S1A1: estrato por tarifa
*P5210S22: posesión de automovil
*P5210S16: posesión de computador
*destring: convierte la variable que no es numérica a numérica
destring dpto, replace
prop dpto.
*Generar dummys para cada área y conocer las proporciones
*Note que la proporción de una dummy es la su media
gen Cali=1 if dpto==76
recode Cali (.=0)
*Tenga en cuenta el factor de expansión
sum Cali [aw=FEX_C]
*
gen Ibagué=1 if dpto==73
recode Ibagué (.=0)
sum Ibagué [aw=FEX_C]
*
gen Bucaramanga=1 if dpto==68
recode Bucaramanga (.=0)
sum Bucaramanga [aw=FEX_C]
*
gen Pereira=1 if dpto==66
recode Pereira (.=0)
sum Pereira [aw=FEX_C]
*
gen Cúcuta=1 if dpto==54
recode Cúcuta (.=0)
sum Cúcuta [aw=FEX_C]
*
gen Pasto=1 if dpto==52
recode Pasto (.=0)
sum Pasto [aw=FEX_C]
*
gen Villavicencio=1 if dpto==50
recode Villavicencio (.=0)
sum Villavicencio [aw=FEX_C]
*
gen Montería=1 if dpto==23
recode Montería (.=0)
sum Montería [aw=FEX_C]
*
gen Manizales=1 if dpto==17
recode Manizales (.=0)
sum Manizales [aw=FEX_C]
*
gen Cartagena=1 if dpto==13
recode Cartagena (.=0)
sum Cartagena [aw=FEX_C]
*
gen Bogotá=1 if dpto==11
recode Bogotá (.=0)
sum Bogotá [aw=FEX_C]
*
gen Barranquilla=1 if dpto==8
```

```
recode Barranquilla (.=0)
sum Barranquilla [aw=FEX_C]
*
gen Medellín=1 if dpto==5
recode Medellín (.=0)
sum Medellín [aw=FEX_C]
*
*Crear dummys de estratos
gen estrato_bajo=1 if P4030S1A1<4
recode estrato_bajo (.=0)
gen estrato_alto=1 if P4030S1A1>=4
recode estrato_alto (.=0)
sum estrato_bajo [aw=FEX_C]
sum estrato_alto [aw=FEX_C]
*Gen Dummy para posesión de
automóvil
gen automóvil=P5210S22
recode automóvil (2=0)
sum automóvil [aw=FEX_C]
*Gen dummy para posesión de
computador
gen computador=P5210S16
recode computador (2=0)
sum computador [aw=FEX_C]
*Crear un archivo sólo las variables
que nos interesan
keep   computador automóvil
estrato_alto estrato_bajo Medellín
Barranquilla Bogotá Cartagena
Montería Manizales Villavicencio
Pasto Cúcuta Pereira Bucaramanga
Ibagué Cali dpto FEX_C directorio
SECUENCIA_P
*ordenar por viviendas y familias
sort  directorio SECUENCIA_P
*Guardar la base de datos en un
archivo nuevo para no modificar el
original
save base1, replace
*Abril es archivo de características de
las personas
use Sas10area1009.dta
*Generar dummy para género
gen Hombre=1 if P6020==1
recode Hombre (.=0)
*Generar variable para años
gen Años=P6040
*Generar duumy 1 si vive en pareja
gen Casado=1 if P6070<4
recode Casado(.=0)
**Generar duumy 1 si está afiliado al
régimen de salud (subsidiado
también)
gen Seguro_salud=1 if P6090==1
replace Seguro_salud=0 if P6090==2
*Generar variable continua años de
educación
gen edu=P6210S1
replace edu=. if edu==99
replace edu=11 if  P6210==5 &
P6210S1>11
replace edu=11 if P6210==6 &
P6210S1==0
replace edu=12 if  P6210==6 &
P6210S1==1
replace edu=13 if  P6210==6 &
P6210S1==2
replace edu=14 if  P6210==6 &
P6210S1==3
replace edu=15 if  P6210==6 &
P6210S1==4
replace edu=16 if  P6210==6 &
P6210S1==5
replace edu=17 if  P6210==6 &
P6210S1==6
replace edu=18 if  P6210==6 &
P6210S1==7
replace edu=19 if  P6210==6 &
P6210S1==8
replace edu=20 if  P6210==6 &
P6210S1==9
replace edu=21 if  P6210==6 &
P6210S1==10
replace edu=22 if  P6210==6 &
P6210S1==11
replace edu=23 if  P6210==6 &
P6210S1==12
replace edu=24 if  P6210==6 &
P6210S1==13
replace edu=25 if  P6210==6 &
P6210S1==14
replace edu=26 if  P6210==6 &
P6210S1==15
*Genera Dummy para Bachiller y
```

```
Universitario
gen Bachiller=1 if P6220==2
recode Bachiller (.=0)
replace Bachiller=. if P6220==.
*
gen Universidad=1 if P6220==4
recode Universidad (.=0)
replace Universidad=. if P6220==.
*Estadísticas
sum Hombre [aw=FEX_C]
sum Bachiller [aw=FEX_C]
sum Universidad [aw=FEX_C]
sum Seguro_salud [aw=FEX_C]
sum Casado [aw=FEX_C]
sum Años [aw=FEX_C]
sum edu [aw=FEX_C]
sum edu if Años>19 & Años<=30
[aw=FEX_C]
sum edu if Años>30 & Años<=40
[aw=FEX_C]
sum edu if Años>40 & Años<=50
[aw=FEX_C]
sum edu if Años>50 & Años<=60
[aw=FEX_C]
sum edu if Años>60 [aw=FEX_C]
*Generar una dummy que identifique
en cuál trimestre del año nació la
persona
gen instrument1=1
replace instrument1=2 if
(P6030S1==4 | P6030S1==5 |
P6030S1==6)
replace instrument1=3 if
(P6030S1==7 | P6030S1==8 |
P6030S1==9)
replace instrument1=4 if
(P6030S1==10 | P6030S1==11 |
P6030S1==12)
*Borrar las variables que no se
necesitan y ordenar por vivienda,
hogar y orden de la persona


keep  Años instrument1 Universidad
Bachiller edu Seguro_salud Casado
Hombre dpto mes esc FEX_C area
directorio SECUENCIA_P orden
hogar regis
sort  directorio SECUENCIA_P orden
save base2, replace
**Usar datos de la fuerza laboral
(personas mayores de 12 años)
use Sas50area1009.dta
*
*P6240: ¿En qué ocupó la mayor
parte del tiempo la semana pasada?
*Generar dummys de emplo,
desempleo y inactividad
*
gen Empleado=1 if P6240==1
recode Empleado (.=0)
sum Empleado [aw=FEX_C]
*
gen Desempleado=1 if P6240==2
recode Desempleado (.=0)
sum Desempleado [aw=FEX_C]
*
gen Inactivo=1 if P6240==3
recode Inactivo (.=0)
sum Inactivo [aw=FEX_C]
*
gen Otro=1 if P6240>3
recode Otro (.=0)
sum Otro [aw=FEX_C]
*
keep  Otro Inactivo Desempleado ft
directorio SECUENCIA_P orden
hogar regis FEX_C
sort  directorio SECUENCIA_P orden
save base3, replace
*
*Usar datos de Ocupados
use Sas60area1009.dta
*
*Generar dummy: 1 si tiene contrato
*
gen Tiene_contrato=1 if P6440==1
recode Tiene_contrato (.=0)
sum Tiene_contrato [aw=FEX_C]
*
*P6450=¿el contrato es escrito?
gen contrato_escrito=1 if P6450==2
replace contrato_escrito=0 if
P6450==1
```

```
sum contrato_escrito [aw=FEX_C]
*Identificar las variables de ingresos
gen salario_mensual=P6500
gen ganancia_independiente=P6750
sum salario_mensual [aw=FEX_C]
sum ganancia_independiente
[aw=FEX_C]
*Identificar la variable de horas
trabajadas
gen horas_trabajadas=P6800
*
*Generar una sola variable de
ingresos tanto para empleados como
independientes
gen salario_todos=salario_mensual
replace
salario_todos=ganancia_independient
e if salario_mensual==.
sum salario_todos [aw=FEX_C]


*General la variable salario por hora
gen
salarioxhora=salario_todos/((52/12)*h
oras_trabajadas)
sum salarioxhora [aw=FEX_C]
***Gen Dummy para asalariados
gen asalariado=1 if
salario_mensual~=.
recode asalariado (.=0)
sum asalariado [aw=FEX_C]
*Sostener, ordenar y guardar
sort  directorio SECUENCIA_P orden
keep  directorio SECUENCIA_P
orden hogar regis area FEX_C
Tiene_contrato contrato_escrito
salario_mensual
ganancia_independiente
horas_trabajadas salario_todos
salarioxhora asalariado
save base4, replace
*Abrir base de datos de otros
ingresos.
use Sas95area1009.dta
*Identificar las variables que
contienen los ingresos
*Mensual
*P7500S3A1: pago por pensión
*P7500S2A1: pensión jubilación
*P7500S1A1: arriendo de casas
*Anual
*P7510S2A1: Transferencia de otros
hogares fuera del país
*P7510S6A1: Cesantías
*P7510S5A1: Intereses de prestamos
*P7510S3A1: Ayuda de instituciones
*P7510S1A1: Transferencia de otros
hogares dentro del país
*P7510S7A1 : Otros, como loterías
*Arreglarlas
recode P7500S3A1 (.=0)
recode P7500S2A1 (.=0)
recode P7500S1A1 (.=0)
*
recode P7510S2A1 (.=0)
replace P7510S2A1=P7510S2A1/12
recode P7510S6A1 (.=0)
replace P7510S6A1=P7510S6A1/12
recode P7510S5A1 (.=0)
replace P7510S5A1=P7510S5A1/12
recode P7510S3A1 (.=0)
replace P7510S3A1=P7510S3A1/12
recode P7510S1A1 (.=0)
replace P7510S1A1=P7510S1A1/12
recode P7510S7A1 (.=0)
replace P7510S7A1=P7510S7A1/12
*
gen
otros_ingresos=P7500S3A1+P7500S
2A1+P7500S1A1+P7510S2A1+P751
0S6A1+P7510S5A1+P7510S3A1+P7
510S1A1+P7510S7A1
recode otros_ingresos (0=.)
sum otros_ingresos [aw=FEX_C]
*Sostener, ordenar y guardar
sort  directorio SECUENCIA_P orden
keep  directorio SECUENCIA_P
orden hogar regis FEX_C
otros_ingresos
save base5, replace
***Guardar todo en una sóla base de
datos****
use base1
*Unir base1 + base2 = base6
```

```
merge directorio SECUENCIA_P
using base2
rename  _merge  _merge1
sort directorio SECUENCIA_P orden
save base6, replace
*Unir base1 + base2 + base4 = base7
merge directorio SECUENCIA_P
orden using base4
rename  _merge  _merge2
sort directorio SECUENCIA_P orden
save base7, replace
*Unir base1 + base2 + base4 +
baqse5= BaseFinal
merge directorio SECUENCIA_P
orden using base5
save BaseFinal, replace
*BASE DE DATOS FINAL PARA
REGRESIONES
*Arreglar variables
recode  Bachiller (.=0)
replace Bachiller=. if edu==.
recode   Universidad (.=0)
replace  Universidad=. if edu==.
*Borrar a los menores de 12 años
para dejar sólo la fuerza laboral
drop if Años<12
gen LnSalarioxHora=ln(salarioxhora)
gen Años2=Años*Años
sum  otros_ingresos [aw= FEX_C]
*Regresión 1
reg  LnSalarioxHora  edu, r
*Regresión2
reg  LnSalarioxHora edu Años Años2,
r
*Arreglar para Regresión 3
gen edu_menos_11=edu-11
gen edu_menos_16=edu-16
*La Dummy apropiada es la siguiente
y no Bachillerato porque Bachillerato
no tiene en cuenta que los
universitarios también tienen
bachillerato, entonces daría negativo
el coeficiente.
gen Secundaria=1 if edu>10
recode Secundaria (.=0)
gen
InteracciónSecundaria=Secundaria*e
du_menos_11
gen
InteracciónUniversidad=Universidad*
edu_menos_16
*Regresión3
reg  LnSalarioxHora edu Años Años2
Secundaria InteracciónSecundaria
Universidad InteracciónUniversidad, r
*Regresión VI
*gen dummy=1 si la persona es
menor de 59 años
gen D1951=1 if Años<59
recode D1951 (.=0)
*Mostrar si la VI es significativa.
reg edu  D1951  Años Años2 Cali
Ibagué Bucaramanga Pereira Cúcuta
Pasto Montería Manizales Cartagena
Barranquilla automóvil Hombre
Casado Seguro_salud Universidad
Tiene_contrato Secundaria
estrato_alto
*Correlaciones
correlate Años automóvil computador
Hombre Casado Seguro_salud edu
Universidad Tiene_contrato
Secundaria D1951 estrato_alto [aw=
FEX_C]
*Regrsiones VI
*La más sencilla
ivregress 2sls LnSalarioxHora Años
Años2 (edu=D1951)
*Con el efeto de los cartones
ivregress 2sls LnSalarioxHora Años
Años2  Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Medellín Hombre
Casado (edu=D1951)
**MCO teniendo en cuenta áreas,
Hombre y Estado Civi
reg LnSalarioxHora edu Años Años2
Cali Ibagué Bucaramanga Pereira
Cúcuta Pasto Montería Manizales
Cartagena Barranquilla automóvil
Hombre Casado Seguro_salud
```

```
reg LnSalarioxHora edu Años Años2
Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla  Medellín Hombre
Casado
*Regresión Heckman
heckman LnSalarioxHora edu Años
Años2 Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Medellín Hombre
Casado, select (edu Años Años2
Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Hombre Casado)
twostep
**Tener en cuenta los otros ingresos
(mensuales
*Sumarlos con el salario mensual de
todos
recode  salario_todos (.=0)
recode  otros_ingresos (.=0)
gen
IngresoTotal=salario_todos+otros_ing
resos
recode IngresoTotal (0=.)
gen LnIngresoTotal=ln(IngresoTotal)
*Correr las mismas regresiones con
variable dependiente=LnIngresoTotal
*Regresión 1
reg  LnIngresoTotal  edu, r
*Regresión2
reg  LnIngresoTotal edu Años Años2,
r
*Regresión3
reg  LnIngresoTotal edu Años Años2
Secundaria InteracciónSecundaria
Universidad InteracciónUniversidad, r
*Regresión IV
ivregress 2sls LnIngresoTotal Años
Años2  Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Medellín Hombre
Casado (edu=D1951)
*Regresión Heckman
heckman LnIngresoTotal edu Años
Años2 Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Medellín Hombre
Casado, select (edu Años Años2
Secundaria Universidad
InteracciónSecundaria
InteracciónUniversidad Cali Ibagué
Bucaramanga Pereira Cúcuta Pasto
Montería Manizales Cartagena
Barranquilla Hombre Casado)
twostep
*FIN
```

FUENTE: DANE. Stata do-dile

Anexo 2. Estadísticas descriptivas

Variables	Observaciones	Media	Desviación Estándar
Dummies			
Cali	9088	0,118755	0,323518
Ibagué	9088	0,025587	0,157909
Bucaramanga	9088	0,052555	0,223155
Pereira	9088	0,029476	0,169144
Cúcuta	9088	0,035376	0,184738
Pasto	9088	0,016981	0,129208
Villavicencio	9088	0,021659	0,145575
Montería	9088	0,013318	0,114640
Manizales	9088	0,021927	0,146455
Cartagena	9088	0,038275	0,191869
Bogotá	9088	0,381230	0,485716
Barranquilla	9088	0,072272	0,258951
Medellín	9088	0,172590	0,377913
Estrato bajo	9088	0,813531	0,389506
Estrato alto	9088	0,186469	0,389506
Automóvil	9088	0,187059	0,389980
Computador	9088	0,409530	0,491774
Hombre	31982	0,479633	0,499593
Bachiller	13027	0,584749	0,492784
Universidad	13027	0,139389	0,346365
Seguro_salud	31969	0,885967	0,317857
Casado	31982	0,383767	0,486310
Empleado	25752	0,499962	0,500010
Desempleado	25752	0,036062	0,186447
Inactivo	25752	0,171196	0,376687
Otra ocupación	25752	0,292781	0,455048
Tiene contrato	14424	0,602082	0,489485
Contrato escrito	8087	0,643512	0,478991
Asalariado	14424	0,517870	0,499698
Continuas			
Años	31982	31,37	20,45
Salario mensual	7082	986772,80	3068214,00
Ganancia (independientes)	6095	801609,00	1496033,00
Salario y ganancias	13177	906088,70	2509006,00
Salario por hora	13177	4732,17	12867,07
Otros ingresos	7156	709821,90	3662522,00

FUENTE: DANE. Cálculos propios.

Anexo 3. Correlación de variables

	Años	Automóvil	Computador	Hombre	Casado	Seguro Salud	Educación	Universidad	Tiene Contrato	Secundaria
Años	1,0000									
Automóvil	0,1002	1,0000								
Computador	0,0323	0,3945	1,0000							
Hombre	0,0172	-0,0118	-0,0441	1,0000						
Casado	0,2317	0,0802	0,0424	0,1202	1,0000					
Seguro Salud	0,1056	0,0800	0,1200	-0,0563	0,1085	1,0000				
Educación	-0,2257	0,3446	0,4112	-0,0380	-0,0435	0,0848	1,0000			
Universidad	0,0227	0,2973	0,2669	-0,0345	0,0119	0,0582	0,4665	1,0000		
Tiene Contra	-0,2822	0,0109	0,0843	-0,0176	-0,0831	0,1084	0,2221	0,0847	1,0000	
Secundaria	-0,2511	0,2433	0,3309	-0,0352	-0,0514	0,0656	0,8201	0,2787	0,2092	1,0000

FUENTE: DANE. Cálculos propios.