Taller 1

1. Data acquisition

(a) Scrape the data that is available at the following website https://ignaciomsarmiento. github.io/GEIH2018 sample/.

(b) Are there any restrictions to accessing/scraping these data?

Al realizar el scraping si hay una serie de restricciones dependiendo si la página web es dinámica o tiene un formato estático. Es decir, si la página a la que se está intentando hacer scraping tiene un formato el cual no cambia y en donde todos los elementos vienen escritos en html entonces se diría que es una página estática. En ese sentido, realizar scraping se remite a identificar en la página de internet el texto, tabla o información que se quiere extraer y codificarlo de acuerdo con la url inicial para que r ejecute la acción de extraer la información. Por el otro lado, cuando la página es dinámica esta está escrita en diferentes lenguajes de programación y contiene diferentes elementos interactúan con el usuario. Lo anterior implica que, se genera un cambio en la visualización en la medida que el usuario ejecuta diversas acciones en la página. Por lo tanto, para hacer scraping no basta con ejecutar el comando, sino que, en principio, es necesario identificar dentro de la página el lugar específico donde se encuentra ubicada la información y posteriormente, identificar los elementos de ese sitio que se quieren extraer. En estas páginas, dada la cantidad de información, esta se almacena en servidores, una página dinámica hace solicitudes para ir trayendo esa información a medida que el usuario interactúa, precisamente este punto mencionado es lo que hace que una página sea dinámica.

(c) Using pseudocode describe your process of acquiring the data

1. Para extraer la información nos dirigimos a la página que contiene las tablas que vamos a extraer.
2. Se identifica si es una página de características estática o dinámica.
3. En este caso se identifica la url con la cual se va a hacer el scraping - "https://ignaciomsarmiento.github.io/GEIH2018\_sample/pages/geih\_page\_1.html
4. Como se desea unificar la información de las diferentes tablas se puede hacer el proceso de dos formas, ir creando una tabla y debajo de esa pegar la nueva base la cual se le está haciendo scraping o realizar un loop que itere el comando por las diferentes páginas que contienen la información y pegarlo, en el caso propio se escoge hacer un loop.
5. El primer paso que se realiza es crear una nueva tabla de datos que en principio está vacía dado que no se ha hecho scraping.
6. Para iniciar el loop se utiliza la función for ( i in #:#) esto le indica a el programa que, en el elemento que se le va a señalar, repita una acción cuando encuentre el elemento i, esta acción se realizará a partir del número de veces que se señale.
7. Dentro del loop, llamamos url al elemento en donde se va a concatenar los elementos sin ningún separador, para eso se utiliza la función **paste0.**
8. De la mano al paso anterior, en la función paste0 se modificará un elemento de la url para que r sepa en donde va a hacer el loop. En ese sentido, dentro de la url, el número que se encuentra al final es el que está indicando la tabla que se está visualizando, por tanto, en vez del número 1 de la url se pondrá “, i , “ esto le indicará a R, la ubicación de cada una de las tablas que va a ir extrayendo y concatenando la información.
9. En un archivo temporal “temp” le vamos a indicar a r que lea el elemento que definimos anteriormente como “url” y que ahí mismo – utilizando la función pipe %>% - convierta la tabla en un data frame de r.
10. Con la función del loop este elemento va a ir haciendo el scrapin por las diferentes tablas, y en nuestra base de datos creada inicialmente con el la función rbin va a ir pegando una debajo de otra las 10 tablas del ejercicio.

2. Data Cleaning

Para limpiar la base de datos fue necesario en primer lugar, qué contiene la base de datos y a partir de la intuición económica ir removiendo o generando nuevas variables.

En ese sentido, la base de datos inicial contiene 32177 observaciones y 178 variables. Dado que, en el problema están especificando que se debe trabajar con personas mayores de 18 años y aquellas que se encuentren trabajando, lo primero que se debe remover son las observaciones que no cumplen estas condiciones.

Si creamos un nuevo conjunto de datos con estas especificaciones entonces resultamos con un conjunto de 16542 observaciones. Dado que, en la base de datos hay 178 observaciones entonces para este trabajo debemos utilizar la lógica económica detrás de los modelos e incluir únicamente las variables relevantes. Por tanto, en este caso en donde se busca evaluar los ingresos que recibe una persona se dejaron las variables que se ven en la tabla 1, al remover variables que no se utilizaran y aquellos registros que no cuentan con información (NA) el número de observaciones desciende a 16276.

Tabla 1. Variables Relevantes



En la ecuación de Mincer (1974) el ingreso aumenta a través de dos vías, la educación y la experiencia, pero a su vez se deben tener en cuenta otras variables que interactúan en este modelo. En ese sentido, a partir de la literatura, algunas de las variables que conservamos en la base de datos debido a que permiten explicar los ingresos son: la edad, el sexo, si el empleo es formal, el tipo de ocupación, si la persona se encuentra en la zona urbana, el tiempo de trabajo en esa empresa y finalmente el nivel de educación.

En el caso de género, esta variable se conserva en la base de datos ya que actualmente se ha observado que hay una discriminación negativa que afecta el género femenino. Lo anterior implica que, para mismos cargos, las mujeres pueden tener un ingreso inferior.

En el caso de la edad, esta variable se conserva pues se espera que a medida que aumente la edad la persona tenga mayores ingresos, sin embargo, se debe tener en cuenta que después de cierta edad la tendencia es que los ingresos tengan una tendencia decreciente y por lo tanto se crea una variable de edadˆ2 que siga esta forma concava.

Se incluye la variable de empleo formal y tipo de empleo pues estas características laborales determinan las condiciones laborales en que se encuentran las personas. En ese sentido, se espera que, por ejemplo, un empleado formal se espera que gane más que un empleado informal. Así mismo, un patrón o empleador debería tener un ingreso más alto que el de un trabajo doméstico.

Así mismo se incluyen la variable de tiempo de trabajo en una empresa, esto se puede convertir en una proxy de experiencia, si una persona está más tiempo en una empresa se esperaría que aumente su experiencia y el ingreso se incremente. La experiencia tiene la misma forma concava que la educación, por tanto, en principio es creciente, pero luego tiene rendimientos decrecientes.

Otra de las variables que se incluyeron fue la de zona urbana y rural, si bien se está trabajando en Bogotá que es una ciudad cuya población es en mayor medida urbana, puede ser que los salarios también presenten una discriminación por esta variable.

Finalmente, de acuerdo con el modelo de Mincer (1974) también se incluyó información sobre educación. En ese sentido, a mayor número de años cursados y completados, el ingreso debería presentar rendimientos crecientes.

Tablas cruzadas



Media de ingresos por tipo de empelados

también hay otros factores que con base en algunas de estas variables se construyó además, otro tipo de variables para que mostraran las características que toman las variables de forma cóncava o convexa. En ese sentido, se construye las variables de educación al cuadrado.