Examen Seminario de Instrumentos Computaciones

Pautas generales

Tienen hasta el día lunes 10/06 a las 23:59 para enviar el examen resuelto. Deben enviar en un mismo mail los archivos de Stata y de R. El examen es domiciliario e INDIVIDUAL. Seremos muy tajantes en caso de encontrar resoluciones gemelas

A. Parte de Stata

Para la resolución del apartado de Stata se pide trabajar con un "dofile master" que vaya ejecutando un dofile distinto para cada punto del examen. Se debe enviar un archivo comprimido que contenga carpetas incluyendo los datos utilizados, los dofiles y los resultados de los ejercicios.

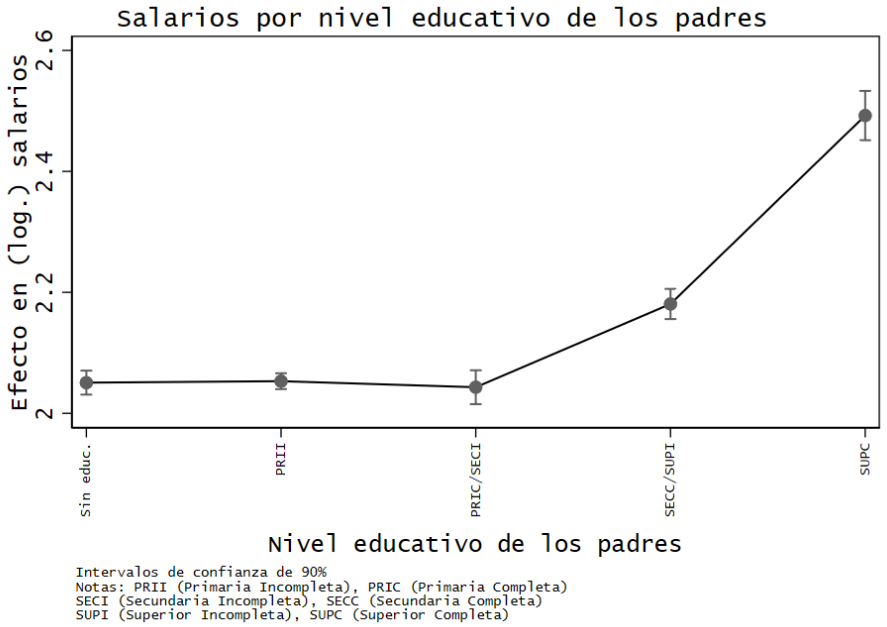
1) El objetivo de este ejercicio es evaluar en qué medida están correlacionados la dispersión de resultados de pruebas las pruebas PISA por nivel socioeconómico con la desigualdad de ingreso de algunos países.
    a. Utilizar los datos provenientes del archivo Excel "EDU-2019-4229-EN-T008", hoja "Table II.B1.3.1", donde se encuentra la información sobre puntajes promedio de pruebas pisa del año 2018 por nivel socioeconómico, y particularmente la diferencia entre el mayor respecto al menor ("Top-Bottom").[1]
    b. Elegir entre 20 países y recurrir a *World Development Indicators* del Banco Mundial para obtener información del Gini del ingreso per cápita y agregar esta información a los datos de PISA del inciso a). La base de datos final debe contener sólo a los países para los que se tiene información de ambas fuentes.
    c. Realizar cualquier tipo de gráfico que relacione la diferencia "top-bottom" con el índice de Gini por país

2) En este ejercicio vamos a utilizar la base de datos "eph_q32023" utilizada en clase. Pero también va a ser necesario descargarse y procesar otra EPH del mismo trimestre, pero diferente año. Puede ser cualquiera entre 2016 y 2019.
   Se pide, pensando en comparar los resultados de ambos años:
    a. Crear 12 muestras: por región (6) y género
    b. Identificar en cada una de ellas a las personas activas, ocupadas y desempleadas con tres variables binarias. Buscar el diccionario de la EPH que provee el INDEC para realizar esto.
    c. Utilizando la variable "p21" que identifica el ingreso promedio en la ocupación principal, obtener una tabla que identifique el promedio de ese ingreso, la tasa de actividad, ocupación y de desocupación por región para cada género.

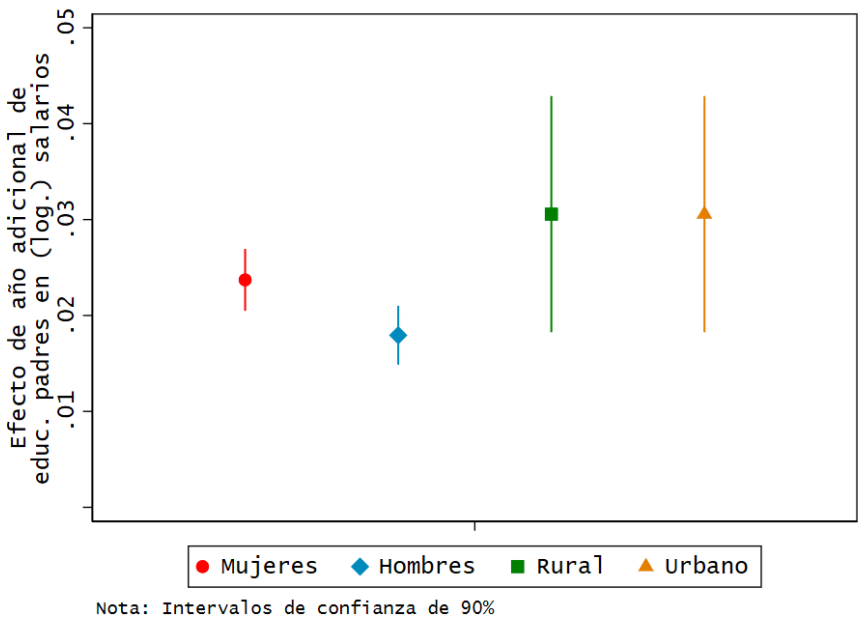
[1] Baku (Azerbaijan) y B-S-J-Z (China) pueden ser eliminados ya que no representan el total nacional.

d. Exportar esas tablas en un archivo Excel en hojas separadas para ambos años.

3) Utilizar la base de datos "pnad14_padres". La misma contiene información que comúnmente se encuentra en encuestas de hogares, así como preguntas retrospectivas sobre educación de los padres del entrevistado. A partir de un análisis de regresión que explique el logaritmo de los salarios horarios (*wage*) a partir del género, zona de residencia (rural/urbana), años de educación del entrevistado, si la persona tiene derecho a jubilarse en su trabajo, región de residencia (*reg_uf)*, y educación de los padres, hacer lo siguiente:
   a. En este inciso utilizamos la variable categórica de educación de los padres. Generar una nueva variable de nivel educativo de los padres con las siguientes categorías:
      i. Nunca asistió
      ii. Primaria Incompleta
      iii. Primaria Completa o Secundaria Incompleta
      iv. Secundaria Completa o Superior Incompleta
      v. Superior completa
   b. Replicar el siguiente gráfico. En lo posible, también su estilo:



   c. Ahora, utilizando la variable continua (en años) de educación de los padres, replicar el siguiente gráfico. En lo posible, también su estilo:



B. Parte de R

Para la resolución de este apartado deberá entregar el examen en un script de R, nombrándolo con su apellido (Examen_Apellido.R).

A lo largo del examen vamos a trabajar con datos para México. En los dos primeros puntos con la encuesta de hogares. Para ello debe descargarse los siguientes archivos (csv) con los microdatos correspondientes al año 2016 desde la página oficial del instituto de Estadísticas:[2]

- Características de las viviendas que habitan los integrantes del hogar
- Características sociodemográficas de los integrantes del hogar

Debe ubicar estos archivos en una carpeta llamada data, y definir los *path* correspondientes, junto con las librerías y cualquier otro *seteo* inicial relevante

1) En este ejercicio vamos a trabajar con los microdatos de la encuesta. **[3.5 pts]**
   a) Cargar como data frame la base de viviendas y explorar los nombres. Luego seleccionar desde la variable en el primer lugar hasta la variable llamada "renta", más la última variable llamada "ubica_geo" y "factor", que corresponden a la ubicación geográfica y al ponderador de la encuesta. Para ello no deben escribir todas las variables en el medio, sino buscar la forma de hacerlo eficientemente, según lo visto en clase
   b) Carguen el dataframe de la base de individuos. Pero antes deben

[2] En este link encontrará la documentación detallada con los cuestionarios de cada archivo

descomprimir el archivo descargado, no de forma manual para este caso, sino usando el comando *unzip()*. El archivo descomprimido debe estar en la misma carpeta de destino. Luego carguen la base y seleccionen desde la variable 1 hasta la variable en la posición 45

c) Cada uno de las bases de microdatos se relacionan entre sí mediante un atributo que tienen en común para poder vincularlas el cual identifica viviendas y personas. En el Excel auxiliar, se deja el detalle de la estructura. Identifique el atributo, analice el tipo de relación y en base a ello haga un *merge* de las dos bases cargadas en los a y b. Asegurarse de que tenga la cantidad de filas y columnas correctas.

d) En base al df del punto c, crear una variable categórica llamada "rango_edad" de tipo character indicando el rango de edad de la persona a partir de la variable numérica en la encuesta llamada edad. Las equivalencias se encuentra en la hoja edad del Excel auxiliar. Asegurarse que si hay valores hay valores que no caen en ninguno de los rangos sean asignados como NA.

e) Generar en un solo bloque de código una tabla con la cantidad de población en cada rango de edad por sexo usando dicha variable en la encuesta. Recuerde el ponderador.

f) En el siguiente bloque de código realizar los cambios necesarios para que la estructura de la tabla y los nombres de la variable sea idéntica a la que se encuentra en la hoja "tabla 1.f" del Excel auxiliar. Esto incluye agregar una columna con el total.

g) Replicar el análisis, pero con la distribución de edades, es decir con la proporción de personas en cada franja etaria sobre el total (sin desagregar por sexo). Note que la tabla resultante debe tener solo 2 columnas, una para el rango de edad y otra para el porcentaje (de 0 a 100). Hacerlo en un solo bloque de código

h) Replicar lo anterior calculando las proporciones por sexo. La tabla debe responder a la pregunta: sobre el total de mujeres y sobre el total de hombres, ¿qué porcentaje corresponde a cada franja etaria? La tabla resultante deberá estar en formato *wide* con los nombres apropiados.[3]

i) Realice un gráfico de torta o de dona, representando la estructura poblacional para cada género mediante el uso de *facet_wrap()*

2) Ahora trabajaremos con un análisis a nivel de estado o entidad federativa de México. **[2.5 pts]**

j) El código de la entidad está contenido en el primer dígito de la variable "ubicageo", cuando esta tiene 8 caracteres y en los primeros dos dígitos cuando esta tiene 9 caracteres.[4] Utilice la función *substr()* para obtener el primer o primer y segundo valor según el largo del character de la variable original de forma tal de generar una variable "entidad".[5]

k) Ahora calculemos el valor del alquiler por entidad. Realice un cruce de las variables "renta" y "tenencia" con *table()*. Filtre entonces solo los hogares donde la tenencia de la vivienda es rentada (valor 1) y calcule para cada

[3] Asegurense de que el total de los porcentajes calculados sumen 100 para cada sexo. Si esto no ocurre piense más detenidamente en cómo utilizar el group_by() para obtener proporciones *"within"* sexo.

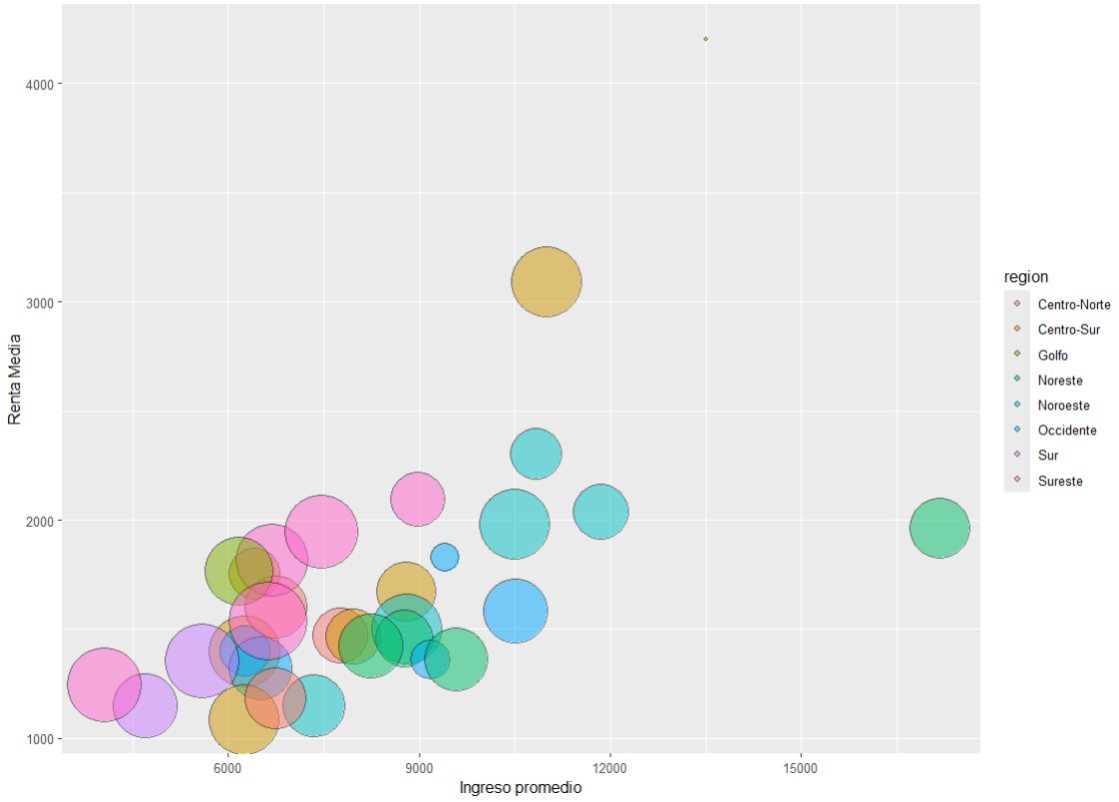
[4] Por ejemplo, el valor "151040000" tiene 9 caracteres, donde el 15 indica que el hogar pertenece a la entidad federativa "15". Un valor "60080000" tiene 8 caracteres e indica la unidad federativa 6. Hay 32 entidades en total.

[5] Tip: explorar el uso de nchar() para formular la condición del largo.

estado, el valor promedio, la mediana, el desvío, el mínimo y el máximo de renta. En ningún caso deben quedarle valores missings.

l) Por último, pegue a esta tabla los nombres de las entidades federativas, la región a la que corresponde, el ingreso promedio y el share de propietarios de viviendas, mediante el *join* que considere relevante con la hoja "clean" del archivo "región_inc_estados" que contiene dichos datos. Realice los cambios necesarios en la columna para poder realizar el *join*,

m) A partir de estos insumos y del valor de la renta promedio calculado antes replique este gráfico, donde el tamaño del circulo indica el share de propiedad en cada entidad.[6]



3) En este apartado trabajaremos con datos de cobertura de alumbrado púbico para cada entidad con sus subniveles de desagregación geográfica (municipio, localidad). **[4pts]**

a) Diríjase a este link. Elija cualquier estado a partir de Aguascalientes en adelante, seleccione y descargue los datos haciendo click en el botón xls. Si necesita vea el **print** de guía en la hoja del excel. Una vez descargado, abra el archivo en excel y explore su estructura. Cargue luego el archivo en R.

b) Seleccione las columnas que identifican numérica y textualmente al estado, municipio y localidad y además a la columna que indica la cobertura de alumbrado público, llamada"ALUMCOB". Renombre a minúsculas todas las variables.

c) Note que la columna de alumbrado, es del tipo character. A partir de ella queremos calcular el nivel de cobertura o intensidad de cobertura en el alumbrado. Para ello genero una variable numérica usando las

[6] Explore el use de range() dentro de scale_size(). Pruebe con valores entre 1 a 30.

correspondencias detalladas en la hoja del excel. Al finalizar controle con *table()* que la correspondencia fue la correcta

Suponga ahora que usted quiere hacer este análisis para cada estado. Podría descargar manualmente cada archivo y repetir el paso anterior con cada uno. Sin embargo en este inciso se pide aplicar lo visto en la clase 6, para generar códigos que le permitan replicar análisis eficientemente

d) Diríjase al botón xls que usó antes para descargar, haga click derecho y elija "copiar dirección del vínculo". Pegue el vínculo en su script. Haga lo mismo con los siguientes 3 estados. ¿Encuentra algún patrón que le permita automatizar el proceso de descarga?

e) Para agilizar el proceso, consideremos por ahora el caso de los 5 primeros estados. Formule entonces un bucle que itere adecuadamente y en cada iteración realice la descarga de la bases para estos 5 estados. En cada vuelta impriman un mensaje que diga "Descargando archivo entidad: X" y que X corresponda al estado en cada iteración. [7]

f) Agregar dentro del bucle los pasos del ejercicio b y c. Es decir, ahora no solo debe descargar en cada iteración los archivos sino que además debe limpiar y manipular la base de datos y guardar el df modificado como elemento de una lista. Para acortar el tiempo de ejecución los primeros 9 estados. [8]

g) Otra forma eficiente de hacerlo sería descargar primero todos los archivos y luego usar la función *list.files()* para generar un vector que contenga las bases descargadas en su carpeta de destino. Dado que ya realizó las descargas antes, genere directamente este vector. Luego, escriba un bucle que vaya iterando por cada una de estas bases, cargándolas en cada vuelta y realizando la transformación anterior. En vez de guardarla en una lista, haga una appendeo de los data frames. [9]

h) Realice lo mismo que antes, pero ahora a los pasos anteriores, agregue los pasos necesarios para calcular la media de la variable numérica alumbrado a nivel municipal dentro del bucle. Filtre previamente los valores NA u omitalos dentro de la función de la media. Al igual que antes realice un append de los dataframes resultantes de manera iterativa dentro del bucle. El df resultante del ejercicio debe tener las columnas identificadoras del estado, del municipio y el valor promedio obtenido. Responda por qué el total de filas del dataframe final es distinto al del ejercicio anterior.

Bonus: estos dos últimos puntos, de hacerse correctamente valen 0.5 ptos que podrá recuperar en caso de haber perdido algunos puntos en alguno de los incisos anteriores. En caso de no hacerlo no resta puntos.

Por último se incorpora el uso de condicionales dentro del bucle. De explorar los archivos originales que descargó, debió notar que hay 3 niveles de desagregación geográfica posible: localidad, municipio, estado. Antes calculamos la media de alumbrado a nivel municipal. Ahora lo haremos para los 3 niveles, dentro del mismo bucle.

---

[7] Tip I: En la clase 6 hemos visto un ejemplo cercano que le servirá para orientarse. Incluya dentro de download.file(), el argumento mode="wb" para asegurarse que el archivo se descargue en modo binario y R luego pueda leerlo.
Tip II: Debe especificar carpeta destino y recordar especificar la extensión correspondiente al archivo

[8] Corra esta linea de código al inicio "options(timeout=120)", para permitir descargas que demoren hasta 2 minutos.

[9] Para esto último, deberá generar inicialmente un df vacio antes del bucle y aplicar la función bind_rows() que hemos visto en clase.

i) Definir un elemento llamado nivel que puede tomar 1 de los 3 valores character: "estado"; "municipio"; "localidad". Debe incorporar el uso de condicionales en el bucle para que si “nivel” fue definido como "municipio", calcule la media para dicho nivel, si en cambio toma el valor "localidad" calcule la media para cada localidad y en cualquier otro caso a nivel estado simplemente. Los condicionales deben estar bien escritos de forma tal que se según el valor indicado en nivel, detecte el tipo el nivel de agregación del análisis. Cambie usted mismo, manualmente, el valor asignado a nivel, para asegurarse que el bucle formulado funciona bien bajo todos los casos.[10]

j) Por último, en vez de cambiar manualmente los valores usted mismo, genero el elemento nivel como un vector con los 3 valores posibles y cree un bucle que itere sobre ellos. Agregue su estructura de bucle del punto anterior dentro de este nuevo bucle. Si las condiciones fueron bien formuladas debe correr correctamente. Como resultado se precisan 3 df distintos, uno para cada nivel.

[10] No replique el código anterior completo dentro de cada condición, sino que realice las transformaciones en primer lugar y luego formule las condiciones para obtener la media de alumbrado en cada caso.