HW0

MMSS 311-2, Sarah B. Bouchat

Due: 14 April 2019

1. Create the following objects in R:
   (a) A vector with the numbers 1–5 in order
   (b) A scalar named `Mindy` that takes the value 12
   (c) A 2×3 matrix with the numbers 1–6 in order by rows
   (d) A 2×3 matrix with the numbers 1–6 in order by columns
   (e) A 10×10 matrix of 1's
   (f) A vector consisting of the words `THIS`, `IS`, `A`, `VECTOR` (each word a separate element)
   (g) A function that takes the sum of any three numbers
   (h) A function that takes one number as input, returns "Yes" if the number is less than or equal to 10 and "No" if the number is greater than 10
   (i) Generate synthetic data by taking 1,000 draws from a normal distribution with a mean of 10 and a standard deviation of 1. Save these data to an object `g`.
   (j) Create a separate object called `y` with 1,000 draws from a normal distribution with a mean of 5 and a standard deviation of 0.5.
   (k) Generate a variable `x` with 1,000 values, where each value is a mean of 10 samples from `g`, with replacement. *(Hint: use a for loop)*
   (l) Estimate a simple bivariate regression $y \sim x$ and print your results. What do your results show?
2. Pull the `pums_chicago.csv` data from the XXXXX repository on my GitHub account. This is a *50,000-person subset* of the US Census Bureau's Public-Use Microdata Sample for almost all of Chicago, covering the five years from 2013 to 2017 inclusive. Documentation is available in the `ACS2013_2017_PUMS_README` file, and a full data dictionary is available here starting at the heading `PERSON RECORD-PERSON VARIABLES`. *Note: ignore language about weights; weighting variables have been excluded and each row of the dataset represents one person.*
   (a) Create an R script file that sets your working directory and loads the data.
   (b) How many variables are there in the dataset?
   (c) What is the mean annual income, `PINCP` in this dataset?
   (d) Create a new variable in the PUMS dataframe called `PINCP_LOG` that is equal to the log of annual income. Were `NaN` values produced? Why?
   (e) Create a new variable `GRAD.DUMMY` that takes the value "grad" if the respondent has any post-high school education, and "no grad" otherwise. Use the `SCHL` variable.
   (f) Drop the variable `SERIALNO` from the dataset.
   (g) Save your new dataset to a `csv` file in the working directory.
   (h) Use the variable ESR, create 5 new dataframes: under 16, employed, unemployed, in the armed forces, and not in the labor force.
   (i) Create a new dataframe that combines employed people and people in the armed forces.
   (j) In your new `employed_af` dataframe, keep only the variables `AGEP`, `RAC1P`, and `PINCP_LOG`

(k) For the following questions, return to the full Chicago dataset.

(i) Find the mean, median, and 80th percentile of travel time to work, `JWMNP`

(ii) Find the correlation between travel time to work `JWMNP` and annual wages `WAGP`

(iii) Make a scatterplot of age and log income.

(iv) Export this graph to your working directory in pdf format.

(v) Create a crosstab of employment status `ESR` by race `RAC1P`

(vi) Estimate a linear regression of annual wages `WAGP` on hours worked per week `WKHP`

(vii) Plot the residuals from this regression against the fitted values. What does this show?

(l) Load the `mtcars` data in R.

(i) Estimate a linear regression of miles per gallon on weight
(ii) Estimate this regression separately for manual versus automatic transition
(iii) Estimate a regression of miles per gallon on the log of horsepower.

(m) Use `ggplot2` to evaluate the `mtcars` data

(i) Make a scatterplot of weight against miles per gallon.
(ii) Color the points in your graph according to the transmission of the vehicle.
(iii) Change the shape of the points to correspond to the number of forward gears in the vehicle.
(iv) Change the x and y labels on the plot to make full words.
(v) Change the background of the plot so that the panel background is not gray.