Assignment 2 Deadline: May 9, 2021

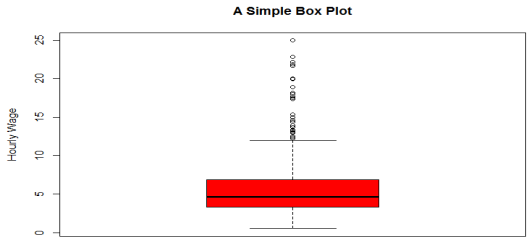
Hand in your hard copy, and submit the R scripts to AE8212@gmail.com

1. (a) Import WAGE1.dta (Stata format) and rename the dataset to "mydata".

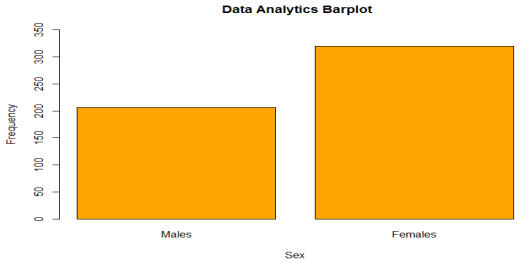
(b) Use the tail( ) function to print out the last 8 observations.

(c) Calculate the average, max, min and median for (hourly) wage.

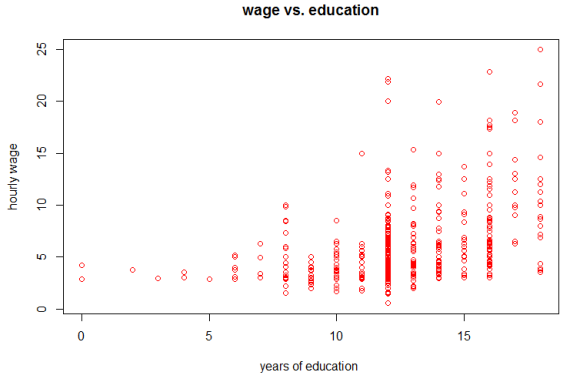
(d) Draw the following box plot for (hourly) wage and change the red color to any other colors you like. Explain your finding.



(e) Draw the following graph and change the orange color to any other colors you like.



(f) Draw a scatter plot like below to show how years of education are related to hourly wage. Change the red color to any other colors you like.



2. Refer to the lecture note about Binary Choice Models (in the Applied econometrics course) and use the MORZ.dta dataset to solve the following questions.

(a) What percentage of the married women were working?

(b)What percentage of married women received at least 12 years of education? How many percentage of them were working?

(c) How many percentage of the married women received at least 16 years of education and their husband's years of education is less than 16.

(d) Create a bar plot to compare the numbers of women in (c) and other married women.

(e) Run the linear probability model: $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 kidsslt6 + \beta_4 exper + \beta_5 exper^2 + u$. Explain your regression output.

(f) Find out $\widehat{inlf}$ for the case that nwifeinc 20, educ=12, kidsslt6=3, exper =6.

3.
The NBAsalary data set contains the information of 269 NBA basketball players. The variables include:

wage: annual salary, thousands $

exper: years as a professional player

age: age in years

coll: years playing at college

games: average games per year

minutes: minutes per season

guard: =1 if guard; forward:=1 if forward; center: =1 if center

points: points per game

rebounds: rebounds per game

assists: assists per game

allstar: all-star player

avgmin: minutes per game

black: =1 if black

children: =1 if has children

(a) Use the pair command to draw 4X4 scatter plots that show the relationships between wage, points, assists, avgmin. Explain your findings.

(b) Run the regression model: $wage = \beta_0 + \beta_1 points + u$, and explain the R-squared result.

(c) Calculate the 90% confidence interval for β_1 and explain your result.

(d) Draw a scatter plot for wage (y) and points (x), and include the estimated regression line from (b) in your scatter plot.

(e) Run the following regression model:

$$wage = \beta_0 + \beta_1 points + \beta_2 \text{agvmin} + \beta_3 forward + \beta_4 center + \beta_5 exper + \beta_6 black + u$$

(f) Can we include the variable "guard" in the (e) regression model? Why or why not?

(g) Is there a racial salary discrimination in the NBA?

(h) Apply AIC on the estimated model found in (e). What is your finding?

(i) Is the adjusted R-squared high for the finalized model in (h)? If not, why the adjusted R-squared is not high?

(j) Based on the finalized model in (h), predict the salary of the first observation in the data set and construct a 95% confidence interval for your predicted wage $\widehat{wage}_1$. How big is the residual?

4. The Admission data set contains a graduate school admission decisions on applications. Admission contains three features, namely GRE, GPA and rank where rank takes on the values from 1 to 4. If a student's undergraduate institution falls into rank 1, the institution has the highest prestige, while those with a rank of 4 have the lowest. The variable admit is a binary decision outcome (=1 is if *admitted*, 0 otherwise).

(a) Use the pair command to draw 4X4 scatter plots that show the relationships between variables. Explain your findings.

(b) Create a box plot for GRE for each rank. [Note: In this box plot, y represents GRE and X represents rank =1, 2, 3 and 4]

(c) How many applicants are from institutions ranked 1? How many percentage of them are rejected?

(d) How many applicants are from institutions ranked 4? How many percentage of them are rejected?

(e)Run the regression model: $admit = \beta_0 + \beta_1 GPA + u$, and explain your estimated slope coefficient.

(f) Draw a scatter plot (admit against GPA) and add the estimated regression line in (e) in this scatter plot. What is your finding?

(g) Consider a multiple regression model: $admit = \beta_0 + \beta_1 GPA + \beta_2 GRE + u$. Suppose GPA is 3.6 and GRE is 700. Calculate the probability of being accepted.

5. Use the Admission data set.

(a) Use the rpart function to find out how one could use GPA, rank and GRE to decide if the applicants should be accepted. Explain the structure of the decision tree you constructed. [Note: set cp=0.05 and minsplit =20]

(b) Suppose an applicant (GPA =3.6, GRE =580) is from an undergraduate institution ranked 2. Based on the decision tree, will s/he be accepted?

6.

(a) Explain in detail the significance of Gini Index and Entropy in decision trees .

(b) Install the"party" package and use the data set called readingSkills found in the package. Calculate the descriptive e statistics (e.g., mean, max, min, etc.) of the variables in readingSkills. [Note: install.packages("party"); library(party); data("readingSkills")]

(c) Draw 4X4 scatter plots for nativespeaker, age, shoeSize and score. Can you draw causal inferences from observational data and the plots? Why or why not?

(d) Use the rpart command to find out how age, ShoeSize and score can be used to decide if someone is a nativespeker. [Note: set cp=0.05 and minsplit =20]

(e) Use rpart.plot to draw and explain your decision tree.

7.

Suppose we would like to use Gender, Place and Education level to predict one's married status. Which feature (Gender or Place or Education) should be used as the first internal node? Use Gini index and Entropy to find out your answer. Do you have the same answer? Why or why not.

X1 (Gender)	X2 (Education)	X3 （Place）	Y （Married）
Female	Secondary	Suburban	Married
Female	No formal education	Suburban	Married
Female	No formal education	Suburban	Married
Female	Secondary	Urban	Married
Female	Degree	Urban	Single
Female	Degree	Urban	Single
Male	Degree	Suburban	Single
Male	Secondary	Suburban	Single
Male	No formal education	Urban	Married
Male	Degree	Urban	Married
Male	Secondary	Urban	Single
Male	Degree	Urban	Single