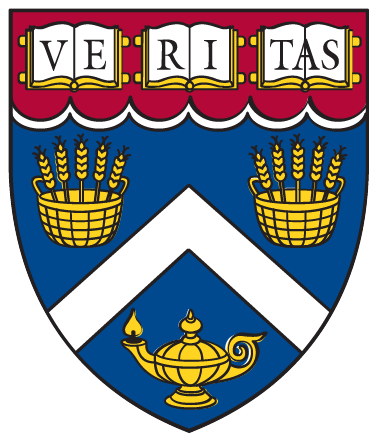
CSCI E-106: Data Modeling



Fall 2019

Dr.Hakan Gogtas

Assignment 1

Due: Monday, 09/23/19 at 7pm EST

**Instructions:** Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr for the .Rmd file submitted in (1) where appropriate. Please, use RStudio Cloud for your solutions.

1. For the regression model Yi=β0+εi, derive the least square estimation for β0?



1. The dataset teengamb (see below for the instructions in r) concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

library(faraway) # download the library

data.help("teengamb") # see the description of the data

You can always the check the scatter plots and boxplots and correlation matrix. These would give you information about the distribution of each variables, outliers, and correlations.

Gamble and Income are correlated. There are outliers in Gamble. While there is no gender impact in gamble, there is a gender difference in status variable.

> round(cor(teengamb),2)

sex status income verbal gamble

sex 1.00 -0.48 -0.12 -0.11 -0.41

status -0.48 1.00 -0.28 0.53 -0.05

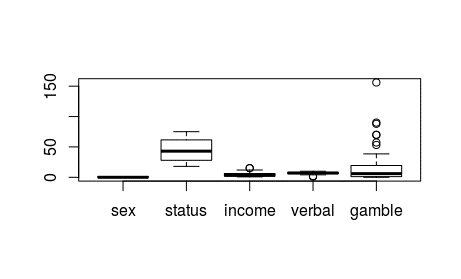
income -0.12 -0.28 1.00 -0.18 0.62

verbal -0.11 0.53 -0.18 1.00 -0.22

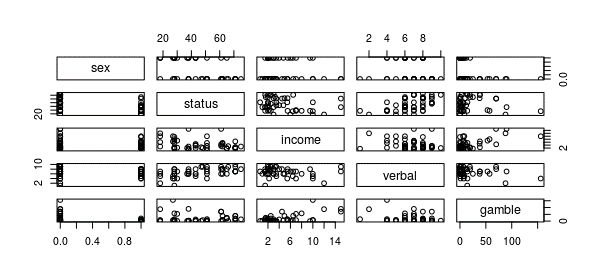
gamble -0.41 -0.05 0.62 -0.22 1.00

> par(mfrow=c(1,1))

> boxplot(teengamb)



plot(teengamb)



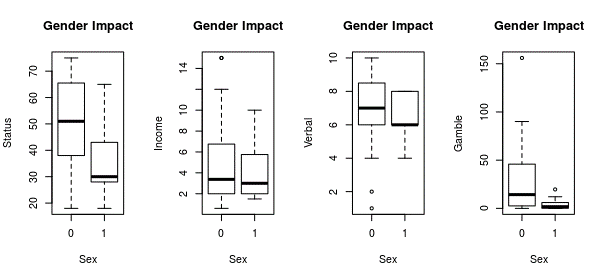
par(mfrow=c(1,4))

boxplot(status~sex,data=teengamb, main="Gender Impact", xlab="Sex", ylab="Status")

boxplot(income~sex,data=teengamb, main="Gender Impact", xlab="Sex", ylab="Income")

boxplot(verbal~sex,data=teengamb, main="Gender Impact", xlab="Sex", ylab="Verbal")

boxplot(gamble~sex,data=teengamb, main="Gender Impact", xlab="Sex", ylab="Gamble")



1. Refer to the CDI data set. The number of active physicians in a CDI (Y) is

expected to be related to total population, number of hospital beds, and total personal income.

* 1. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.
  2. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?
  3. Calculate MSE for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

First upload the data into R:

Y= -1.106e+02 + 2.795e-03\*Total Population. MSE=372204

Y= -95.9322 + 0.7431\*Number of hospital beds. MSE=310192

Y= -48.3948 + 0.1317\*Total personal income. MSE=324539

the linear relationship provides a good fit. Number of hospital beds has the highest predictability power as it has the lowest MSE.

f1<-lm(Number.of.active.physicians~Total.population,data = CDI)

f2<-lm(Number.of.active.physicians~Number.of.hospital.beds,data=CDI)

f3<-lm(Number.of.active.physicians~Total.personal.income,data=CDI)

par(mfrow=c(1,3))

plot(CDI$Total.population,CDI$Number.of.active.physicians)

abline(f1)

plot(CDI$Number.of.hospital.beds,CDI$Number.of.active.physicians)

abline(f2)

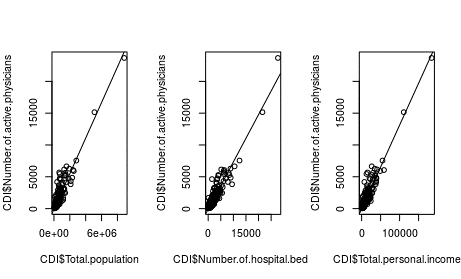
plot(CDI$Total.personal.income,CDI$Number.of.active.physicians)

abline(f3)

anova(f1)

anova(f2)

anova(f3)



1. Repeat question 3, by building the models on the development sample (a random sample of 70% of CDI data), and calculating MSE’s on the hold out sample (remainder 30% of the CDI data).

Y= -152.25793 + 0.00291\*Total Population. MSE=371464.1

Y= -35.3824 + 0.6897\*Number of hospital beds. MSE=412520.1

Y= -67.8265 + 0.1327\*Total personal income. MSE=357228.9

the linear relationship provides a good fit. Number of hospital beds has the highest predictability power as it has the lowest MSE.

set.seed(12345)

ind <- sample(1:nrow(CDI), size =nrow(CDI)\*0.70)

dev <- CDI[ind,]

holdout <- CDI[-ind,]

f11<-lm(Number.of.active.physicians~Total.population,data=dev)

f21<-lm(Number.of.active.physicians~Number.of.hospital.beds,data=dev)

f31<-lm(Number.of.active.physicians~Total.personal.income,data=dev)

par(mfrow=c(1,3))

plot(dev$Total.population,dev$Number.of.active.physicians)

abline(f11)

plot(dev$Number.of.hospital.beds,dev$Number.of.active.physicians)

abline(f21)

plot(dev$Total.personal.income,dev$Number.of.active.physicians)

abline(f31)

anova(f11)

anova(f21)

anova(f31)

MSE1<-sum((holdout$Number.of.active.physicians-predict(f11,holdout))^2)/( dim(holdout)[1]-2)

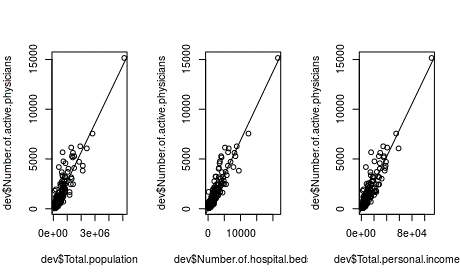
MSE2<-sum((holdout$Number.of.active.physicians-predict(f21,holdout))^2)/( dim(holdout)[1]-2)

MSE3<-sum((holdout$Number.of.active.physicians-predict(f31,holdout))^2)/( dim(holdout)[1]-2)

cbind(MSE1,MSE2,MSE3)

MSE1 MSE2 MSE3

[1,] 371464.1 412520.1 357228.9



1. The dataset teengamb concerns a study of teenage gambling in Britain.
   1. Regress the expenditure on gambling (Y) on income (X). State the estimated regression function. Compute the mean and median of the residuals.
   2. Which observation has the largest (positive) residual? Give the case number.

Gamble=-6.325 + 5.520\*income

summary(f$residuals)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-46.020 -11.874 -3.757 0.000 11.934 107.120

24th observation has the largest residual

res[which.min(res$res),]

gamble income res

17 0.1 9.5 -46.02005

res[which.max(res$res),]

gamble income res

24 156 10 107.1197

f<-lm(gamble~income,data=teengamb)

summary(f)

summary(f$residuals)

res<-f$residuals

res<-cbind(f$model,res)

res[which.min(res$res),]

res[which.max(res$res),]