# HW 4 Due Tuesday Sept 27, 2016. Upload R file to Moodle with name: HW4_490IDS_YOUR_CLASSID.R
# Notice we are using the new system with your unique class ID. You should have received an email with
# your unique class ID. Please make sure that ID is the only information on your hw that identifies you.
# Do not remove any of the comments. These are marked by #

### Part 1: Linear Regression Concepts
## These questions do not require coding but will explore some important concepts
## from lecture 5.

## "Regression" refers to the simple linear regression equation:
##    y = B0 + B1*x
## This homework will not discuss any multivariate regression.

## 1. (1 pt)
## What is the interpretation of the coefficient B1?
## (What meaning does it represent?)
## Your answer
**B1 represents the difference in the predicted value of y for each one-unit difference in x.**

## 2. (1 pt)
## If the residual sum of squares (RSS) of my regression is exactly 0, what does that mean about my model?
## Your answer
**It means that the observations fit the model.**

## 3. (2 pt)
## Outliers are problems for many statistical methods, but are particularly problematic
## for linear regression. Why is that? It may help to define what outlier means in this case.
## (Hint: Think of how residuals are calculated)
## Your answer
**The distance to the best-fit line is squared what calculating residuals, amplifying the influence of the farthest points.**

### Part 2: Sampling and Point Estimation

## The following problems will use the ggplot2movies data set and explore
## the average movie length of films in the year 2000.

## Load the data by running the following code
```
install.packages("ggplot2movies")
library(ggplot2movies)
data(movies)
```

## 4. (2 pts)
## Subset the data frame to ONLY include movies released in 2000.
movies<-movies[movies$year==2000,]

## Use the sample function to generate a vector of 1s and 2s that is the same
## length as the subsetted data frame. Use this vector to split
## the 'length' variable into two vectors, length1 and length2.

## IMPORTANT: Make sure to run the following seed function before you run your sample
## function. Run them back to back each time you want to run the sample function.

## Check: If you did this properly, you will have 1035 elements in length1 and 1013 elements
## in length2.
set.seed(1848)
# sample(...)
s<-sample(1:2,2048,replace=T)
l<-split(movies$length,s)
length1<-l[[1]]
length2<-l[[2]]

## 5. (3 pts)
## Calculate the mean and the standard deviation for each of the two
## vectors, length1 and length2. Use this information to create a 95%
##confidence interval for your sample means. Compare the confidence
## intervals -- do they seem to agree or disagree?
## Your answer here
length1:
mean(length1)
**78.33623**
sd(length1)
**40.07299**
**SE=40.07299/sqrt(1035)=1.245609**
**78.33623+1.96*1.245609=80.77762**
**78.33623-1.96*1.245609=75.89484**
**confidence intervals: (75.89484, 80.77762)**
length2:
mean(length2)
**80.02073**
sd(length2)
**39.3216**
**SE=39.3216/sqrt(1013)=1.235454**
**80.02073+1.96*1.235454=82.44222**
**80.02073-1.96*1.235454=77.59924**
**confidence intervals: (77.59924, 82.44222)**

**I think they seem to agree because the difference is little.**

```
## 6. (4 pts)
## Draw 100 observations from a standard normal distribution. Calculate the sample mean.
## Repeat this 100 times, storing each sample mean in a vector called mean_dist.
## Plot a histogram of mean_dist to display the sampling distribution.
## How closely does your histogram resemble the standard normal? Explain why it does or does not.
## Your answer here
set.seed(1848)
rnorm(100)
mean(rnorm(100))
-0.09746113
mean_dist<-replicate(100,mean(rnorm(100)))
hist(mean_dist)
```

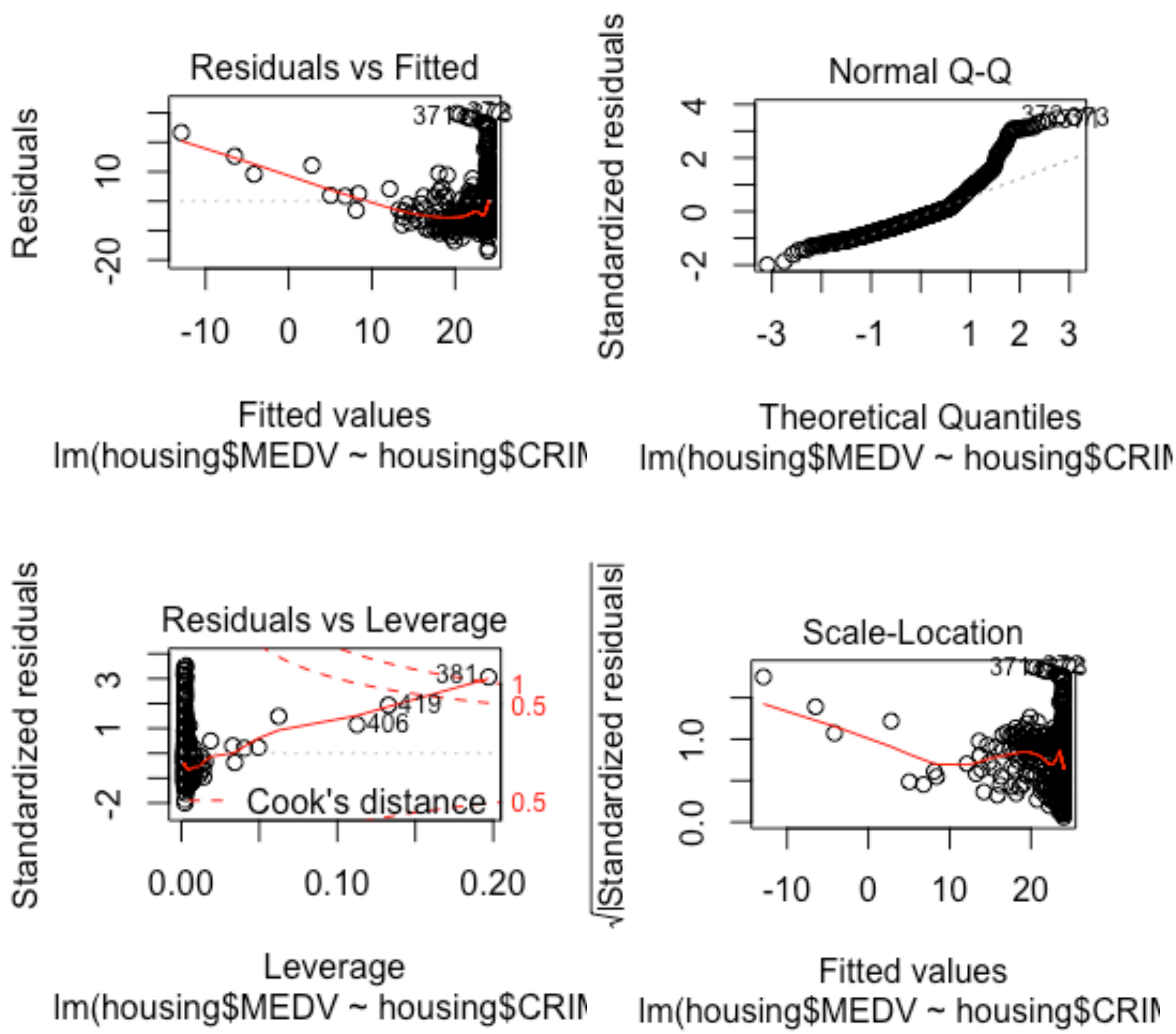
**I think the histogram does not resemble the standard normal very well because the sample is not large enough.**

```
## 7. (3 pts)
## Write a function that implements Q6.

## Your answer here
HW.Bootstrap=function(n,reps) {
 set.seed(1848)
 mean_dist<-replicate(reps,mean(rnorm(n)))
 hist(mean_dist)
 return(mean_dist)
}
```

```
### Part 3: Linear Regression
## This problem will use the Boston Housing data set.
## Before starting this problem, we will declare a null hypthosesis that the
## crime rate has no effect on the housing value for Boston suburbs.
## That is: H0: B1 = 0
##         HA: B1 =/= 0
## We will attempt to reject this hypothesis by using a linear regression


# Load the data
housing <- read.table(url("https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data"),sep="")
names(housing) <- c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B","LSTAT","MEDV")
```

## 7. (2 pt)
## Fit a linear regression using the housing data using CRIM (crime rate) to predict
## MEDV (median home value). Examine the model diagnostics using plot(). Would you consider this a good model or not? Explain.

```
lmfit<-lm(housing$MEDV~housing$CRIM)
plot(lmfit)
plot(housing$CRIM,housing$MEDV)
```




**I think this is not a good model. According to the Residuals vs Fitted plot, there is a trend in the plot, and the residuals are not randomly scattered. According to Q-Q plot, the residuals is not a normal distribution.   According to standardized residuals vs fitted plot, the variance is not constant and there are outliers.**

## 8. (2 pts)
## Using the information from summary( ) on your model, create a 95% confidence interval
## for the CRIM coefficient
summary(lmfit)

```
Call:
lm(formula = housing$MEDV ~ housing$CRIM)

Residuals:
    Min      1Q  Median      3Q     Max
-16.957  -5.449  -2.007   2.512  29.800

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.03311    0.40914   58.74   <2e-16 ***
housing$CRIM -0.41519    0.04389   -9.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.484 on 504 degrees of freedom
Multiple R-squared:  0.1508,	Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

**-0.41519-2*0.04389**
**-0.41519+2*0.04389**
**confidence interval: (-0.50297, -0.32741)**

## 9. (2 pts)
## Based on the result from question 8, would you reject the null hypothesis or not?
## (Assume a significance level of 0.05). Explain.
## Your answer
**As the p-value is much less than 0.05, we reject the null hypothesis H0 that B1 = 0. Hence there is a significant relationship between the crime rate and median value of owner-occupied homes in the linear regression model.**

## 10. (1 pt)
## Pretend that the null hypothesis is true. Based on your decision in the previous
## question, would you be committing a decision error? If so, which one?
## Your answer
**Yes. Type 1 error**

## 11. (1 pt)
## Use the variable definitions from this site:
## https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names

## Discuss what your regression results mean in the context of the data (using appropriate units)
## (Hint: Think back to Question 1)
## Your answer
**The intercept is 24.03311 and the coefficient for the crime rate is -0.41519. Therefore, the predicted Median value of owner-occupied homes will decrease by 0.41519 for one unit increase in the crime rate.**

## 12. (2 pt)
## Describe the LifeCycle of Data for Part 3 of this homework.
**First, we plan to use the Boston Housing data set to test a hypothesis.**
**Second, we acquire the data set using the URL, and store it in a data frame.**
**Then we select two variables of the data set, crime rate and median home value.**
**Third, the data of the two variables are used to model a linear regression.**
**We write codes to analyze the model.**
**Based on the analysis of the model, the problem in the planning could be answered.**
**Data was interpreted in the final phase.**