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 is the slope of the regression line. B1 > 0 means positive correlation between y and x.

B1 < 0 means negative correlation. B1 = 0 means no correlation between y and 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 there is no discrepancy between the data and the estimation model.

The model is a perfect fit of the data. There is no error.

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

Because extreme values of observed variables can distort estimates of regression coefficients.

A point which lies far from the line (and thus has a large residual value) is known as an outlier.

Such points may represent erroneous data, or may indicate a poorly fitting regression line.

Residuals are the deviations from the fitted line to the observed values.

If an outlier lies far from the other data especially in the horizontal direction, it may have a

significant impact on the slope of the linear regression line. So outliers are particularly

problematic for linear regression.

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.

```
> movies2000 = 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.

```
> dim(movies2000)
[1] 2048   24
> set.seed(1848)
> length.vector = sample(1:2, size = 2048, replace = TRUE)
> length1 = movies2000$length[which(length.vector==1)]
> length2 = movies2000$length[which(length.vector==2)]

> length(length1)
[1] 1035
> length(length2)
[1] 1013
```

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

```
> mean1 = mean(length1)
> sd1 = sd(length1)
> mean2 = mean(length2)
> sd2 = sd(length2)
> interval1 = c(mean1 - 1.96 * sd1/sqrt(length(length1)), mean1 + 1.96 * s
d1/sqrt(length(length1)))
> interval2 = c(mean2 - 1.96 * sd2/sqrt(length(length2)), mean2 + 1.96 * s
d2/sqrt(length(length2)))
> interval1
[1] 75.89484 80.77762
> interval2
[1] 77.59924 82.44222
```

The 95% confidence interval of length1 is [75.89484, 80.77762].

The 95% confidence interval of length2 is [77.59924, 82.44222].

The two confidence intervals seem to agree. Though there are about 2 minutes' difference

between them, they are mostly overlapping.

6. (4 pts)

Draw 100 observations from a standard normal distribution. Calculate the sample mean.

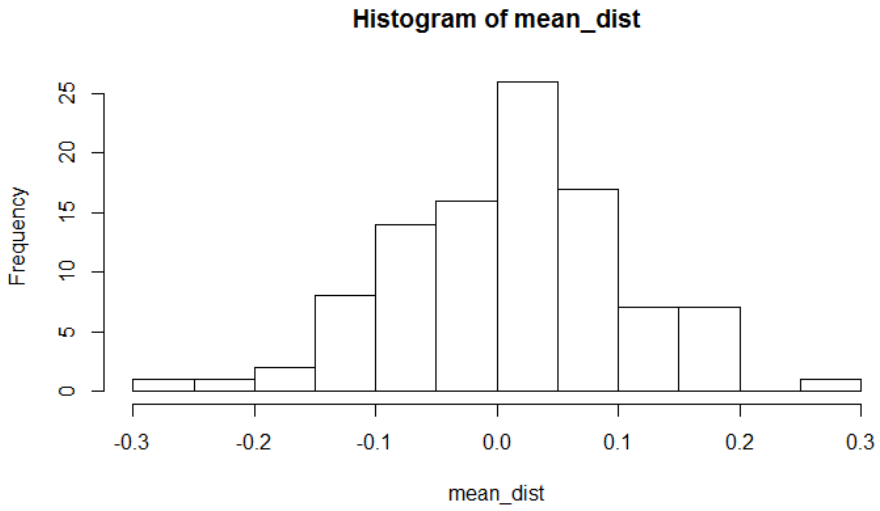
Repeat this 100 times, storing each sample mean in a vector called mean_dist.

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

```
mean_dist = rep(0, 100)
for (i in 1:100){
  mean_dist[i] = mean(rnorm(100))
}
hist(mean_dist)
```



My histogram basically resembles the standard normal. According to central limit theorem,

the distribution of sample means is approximately normal distribution when the sample size

is large enough. The mean of the sample means is equal to the mean of the original population, which is 0.

The standard deviation of the sample means is equal to 1/sqrt(100) in this case. So the

histogram of sample means resembles normal distribution, but not exactly standard. The mean is still 0,

but the standard deviation is 1/sqrt(100). It is obvious to get a normal distribution because of randomness.

7. (3 pts)

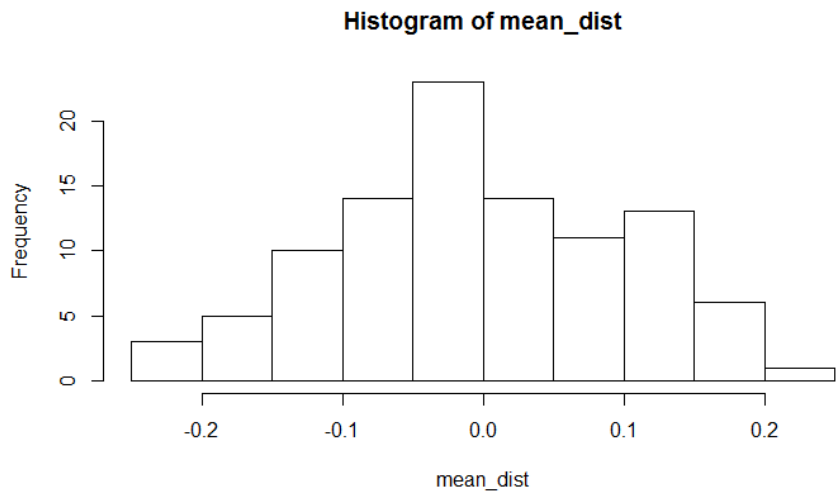
Write a function that implements Q6.

Your answer here

```
HW.Bootstrap=function(distn=rnorm,n,reps){
 set.seed(1848)

 #more lines here
 mean_dist = rep(0, reps)
 for (i in 1:reps){
  mean_dist[i] = mean(distn(n))
 }
 hist(mean_dist)
```

}

HW.Bootstrap(n=100, reps = 100)



Part 3: Linear Regression

This problem will use the Boston Housing data set.

Before starting this problem, we will declare a null hypothesis 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)

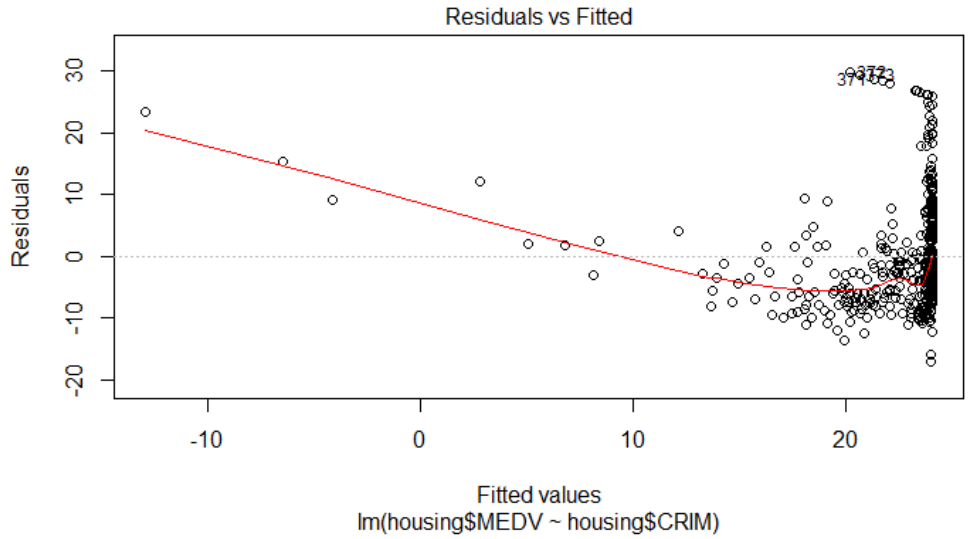
Fit a linear regression using the housing data using CRIM (crime rate) to predict

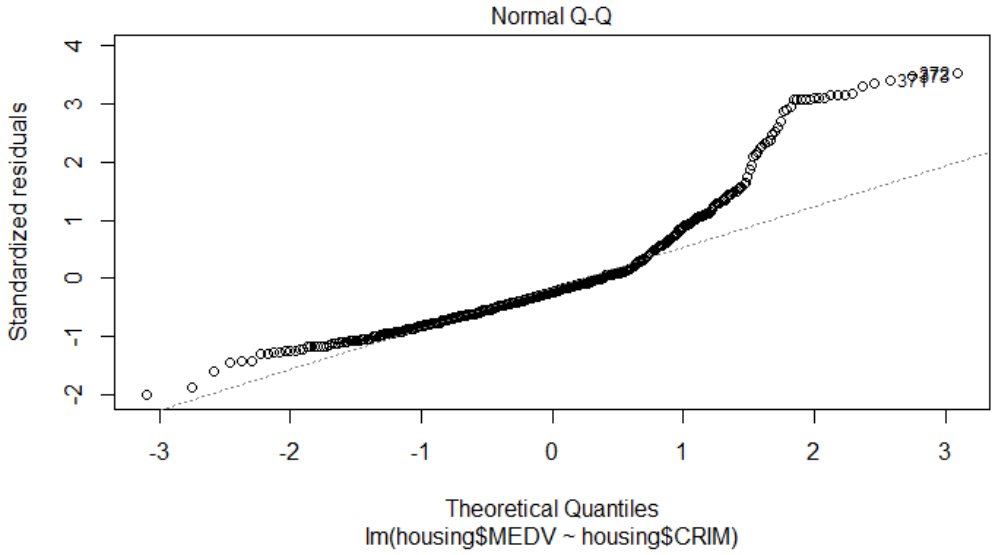
MEDV (median home value). Examine the model diagnostics using plot(). Would you consider this a good model or not? Explain.

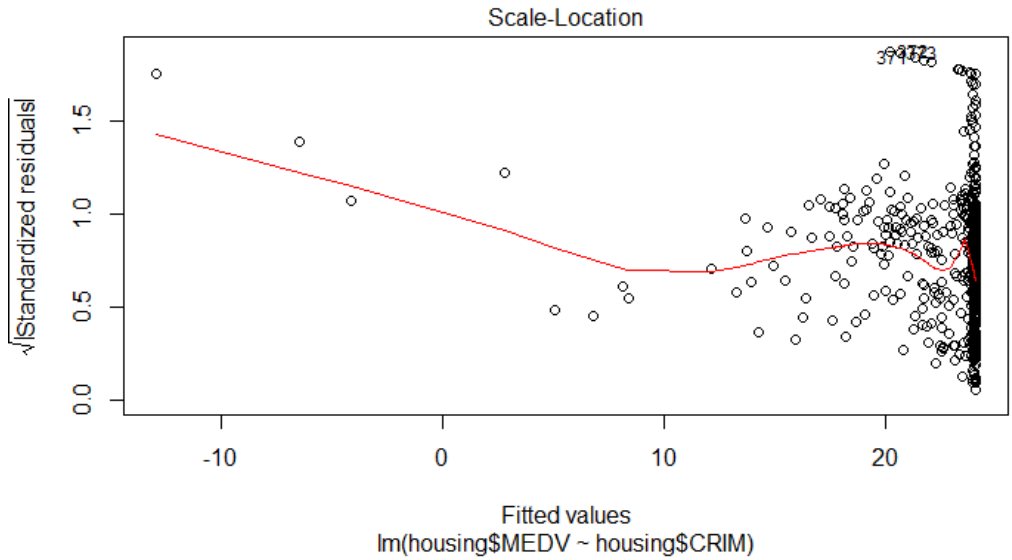
lm(formula = housing$MEDV ~ housing$CRIM, data = housing)

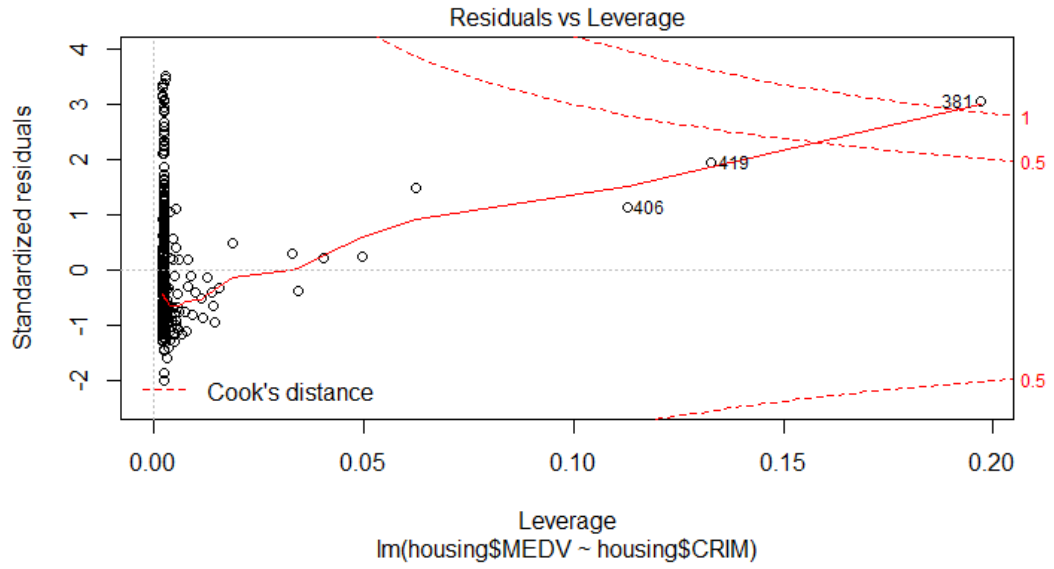
```
> lm(formula = housing$MEDV ~ housing$CRIM, data = housing)

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

Coefficients:
 (Intercept)  housing$CRIM
     24.0331       -0.4152
```

plot(lm(formula = housing$MEDV ~ housing$CRIM, data = housing))









The model is MEDV = 24.0331 - 0.4152*CRIM

I think it is not a good model. In the Residuals vs Fitted graph, the red line is not close to

the zero line and it is not parallel to the zero line. Some are obviously positive residuals

and others are obviously negative residuals. In the Normal Q-Q graph, half values of data

fit well with the line. But half obviously deviate from the line. In the Scale-Location graph,

most data gather in the right edge of the graph and disperse. So from the Residuals vs Fitted

graph and the Scale-Location graph, we can see that several values on the left might be regarded

as outliers. The model might be well generated when leaving out those outliers. And in the Residuals

vs Leverage graph, one value is beyond the 1 cook's distance line, which is not good.

8. (2 pts)

Using the information from summary() on your model, create a 95% confidence interval

for the CRIM coefficient

```
> summary(lm(formula = housing$MEDV ~ housing$CRIM, data = housing))

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

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


> CRIM.coefficient = -0.41519
> CRIM.sd = 0.04389
> CRIM.interval = c(CRIM.coefficient - 1.96 * CRIM.sd/sqrt(506), CRIM.coef
ficient + 1.96 * CRIM.sd/sqrt(506))
> CRIM.interval
[1] -0.4190143 -0.4113657
```

The 95% confidence interval of CRIM coefficient is [-0.4190143, -0.4113657].

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

I would reject the null hypothesis because the p-value is far less than 0.05. It means

a significant correlation. Besides, the 95% confidence interval of CRIM coefficient does

not have any overlap with zero.

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

If the null hypothesis is true, I would be committing a decision error. That's 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

My regression results reflect a significant negative correlation between MEDV and CRIM.

That means as the per capita crime rate by town goes up, the median value of

owner-occupied homes in $1000's goes down. Based on the regression line, for example, the per capita

crime rate rises from 0.5% to 1%, the median value of owner-occupied homes drops from

$23826 to $23618.

12. (2 pt)

Describe the LifeCycle of Data for Part 3 of this homework.

In a data lifecycle, we could first think about data collection. For Part 3, housing dataset

was collected in 1978 by Harrison, D. and Rubinfeld, D.L. Then after data cleaning and preprocessing,

Boston Housing Data was taken from the StatLib library of Carnegie Mellon University in 1993.

Loading the housing data to R Studio can be regarded as a data exploration. It has 506 observations

and 14 variables. Then we choose MEDV and CRIM two columns from the dataset to make mathematical

computation for the interest of the correlation between crime rate and median home value.

Then, to make data inference and prediction, we establish a linear regression model to find the

correlation and coefficients. Plot of the model is a data visualization and presentation part.

Then the summary, hypothesis testing and confidence interval are further data analysis.

Finally, the regression model can draw a qualitative and quantitative conclusion of the correlation

and make knowledge discovery. This correlation conclusion idea should be shared and re-used

for future data analysis in the whole lifecycle.