ILLINOIS TECH

Homework 1

EDUARDO GALEOTE ESCALERA

A20552496

CSP 571

Data Preparation and Analysis

September 2023

Contents

1 Recitation Exercises

1.1 Chapter 2

1.1.1 Excercise 1

For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

a) *The sample size n is extremely large, and the number of predictors p is small.* When the number of predictors is small, an inflexible model might work better than the flexible model because it simplifies the relationship between predictors and the response variable.

b) *The number of predictors p is extremely large, and the number of observations n is small.* With a large amount of predictors, the more flexible model can be prone to overfitting. We would expect an inflexible model to perform better.

c) *The relationship between the predictors and response is highly non-linear.* If the real f is a highly non-linear function we would need a flexible model for the predictions to be more accurate. Because the inflexible model would assume linear relationships.

d) *The variance of the error terms, i.e.* $\sigma^2 = Var(\epsilon)$*, is extremely high.* A model with high variance will perform better with a more inflexible model because it tends to smooth out the noise we don´t want.

1.1.2 Excercise 2

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a) *We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.*

- Regression.
- Inference.
- n: the 500 companies.
- p: 3(profit, number of employees, industry).

b) *We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.*

- Classification.
- Prediction.
- n: the 20 similar products.

- p: 13 (price, marketing budget, competition price, and ten other variables).

c) *We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.*

- Regression.
- Prediction.
- n: the 52 weeks in all 2012.
- p: 3 (changes in the US market, British market, and German market).

1.1.3 Excercise 4

You will now think of some real-life applications for statistical learning.

a) *Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.*

- A medical diagnose. If the patient will have a stroke or not based on previous infections such as COVID-19.
  - Response: The response variable represents if it is likely for a patient to develop a stroke, such as "stroke" (class 1) or "no stroke" (class 0).
  - Predictors: Predictors may include patient demographics, medical history, lab test results, and symptoms.
  - Goal: In the context of medical diagnosis, the primary goal is prediction. The system aims to predict whether a patient has a particular medical condition based on their symptoms and medical history.
- Whether a person is a good or bad client for a business.
  - Response: The response variable indicates how good a client is, often categorized as "good client" (class 1) or "bad client" (class 0).
  - Predictors: Predictors can include purchase history, money spent on website or physical stores, return history, etc.
  - Goal: In this application, inference is the main goal. To find relationships between certain variables to determine whether a client is good or bad.
- Email Spam Detection.
  - Response: The response variable is binary, indicating whether an email is spam (class 1) or not (class 0).
  - Predictors: Predictors may include various email attributes, such as sender information, subject line, body content, attachments, etc.
  - Goal: The primary goal here is prediction. The system aims to predict whether an incoming email is spam or not, allowing it to automatically filter out unwanted emails.

b) *Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.*

- A medical diagnose. Heart rate prediction during physical activity.
  - Response: The response variable will be the numerical value of the person´s heart rate.
  - Predictors: The predictors can be the patient´s age, physical condition, genre, and other body attributes.
  - Goal: The main goal is to predict with the best accuracy the heart rate of the patient when doing some type of sport.
- Real estate price prediction.
  - Response: The response variable may be the price of a property in a city center such as Madrid.
  - Predictors: Square feet, distance to the underground, neighborhood, number of bedrooms and bathrooms, etc.
  - Goal: To predict a price of a property in Madrid based on certain predictors related to other homes in Madrid and the property.
- Income of a person.
  - Response: The response variable would be the annual income of a person with a degree in engineering.
  - Predictors: Some examples of possible predictors may be person´s age, engineering degree, sector in the industry, years in the field, job position, etc.
  - Goal: The goal is to predict an engineer salary.

c) *Describe three real-life applications in which cluster analysis might be useful.*

- Supermarket costumers. One type of cluster analysis that me be useful could be dividing the recurrent costumers from a supermarket in three groups. Low spenders, medium spenders and big spenders. Use this division to send more offers and discounted products to some clients.
- Clustering the medical urgency of an illness. To filter all the people that come to an emergency room, it would be useful to cluster the ill in different cluster to analyze the urgency of their illness. Some may require urgent surgery and some others may not require immediate attention.
- Traffic Flow. Transportation authorities use cluster analysis to analyze traffic patterns and segment roads based on traffic density and congestion levels.

1.1.4 Excercise 6

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non- parametric approach)? What are its disadvantages?
Parametric methods involve a two-step model-based approach. First, we make an assumption about the functional form, or shape, of f. After a model has been selected, we need a procedure that uses the training data to fit or train the model. The model-based approach just described is referred to as parametric; it reduces the problem of estimating f down to one of estimating a set of parameters. Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters. The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f.
On the other hand, non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as

possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.
Some advantages of a parametric approach to regression or classification can be great interpretability, efficiency with limited Data as we've seen before and reduced risk of overfitting. The disadvantages of a parametric method may be the limitation of flexibility and model assumption sensitivity.

1.1.5 Excercise 7

a) *Compute the Euclidean distance between each observation and thetestpoint,X1 =X2 =X3 =0.*
We use the next equation to compute the Eucladian distance $d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$.

- Observation 1: d = 3.
- Observation 2: d = 2.
- Observation 3: d = $\sqrt{10}$.
- Observation 4: d = $\sqrt{5}$.
- Observation 5: d = $\sqrt{2}$.
- Observation 6: d = $\sqrt{3}$.

b) *What is our prediction with K = 1? Why?*
With K = 1, we select the nearest neighbor, which is Observation 5. The response for Observation 5 is "Green", so our prediction with K = 1 is "Green".

c) *What is our prediction with K = 3? Why?*
With K = 3, we select the three nearest neighbors, which are Observations 5, 6, and 4. Among these neighbors, there are two "Green" responses and one "Red" response. Since "Green" is the majority class among the three nearest neighbors, our prediction with K = 3 is "Green".

d) *If the Bayes decision boundary in this problem is highly non- linear, then would we expect the best value for K to be large or small? Why?*
If the Bayes decision boundary is highly nonlinear, it is likely that the data samples are not well-separated. In such cases, we would expect the best value for K to be small.

1.2 Chapter 3

1.2.1 Excercise 1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

- h_0 for TV: There is no statistically significant linear relationship between TV advertising spending and sales.
  Because the p-value is < 0.0001 we reject h_0.
- h_0 for radio: There is no statistically significant linear relationship between radio advertising spending and sales.
  Because the p-value is < 0.0001 we reject h_0.
- h_0 for newspaper: There is no statistically significant linear relationship between newspaper advertising spending and sales.
  The p-value for newspaper is approximately 0.8599, which is much larger than a typical significance level like 0.05. Therefore we accept h_0.

1.2.2 Excercise 3

Suppose we have a data set with five predictors, X_1 = GPA, X_1 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = 10$.

a) *Which answer is correct, and why?*
   ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.
   This statement is the correct one because $\hat{\beta}_3$ represents the difference in salary between high school graduates (Level = 0) and college graduates (Level = 1). Since $\hat{\beta}_3$ is positive (35), it implies that, on average, college graduates earn more than high school graduates.

b) *Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.*

$$Salary = 50 + 20 * 4.0 + 0.07 * 110 + 35 + 0.01 * (4.0 * 110)10 * (4.0 * 1) = 137.1k\$. \quad (1.1)$$

c) *True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.*
   False. The size of the coefficient $\hat{\beta}_4$ (0.01) for the GPA/IQ interaction term alone is not sufficient to determine the interaction effect. The significance of an interaction effect is assessed based on the p-value associated with the coefficient $\hat{\beta}_4$.

1.2.3 Excercise 4

I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \hat{\beta}_3 X^3 + \epsilon$.

a) *Suppose that the true relationship between X and Y is linear, i.e. $Y = \hat{\beta}_0 + \hat{\beta}_1 + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.*
   In this case, when the true relationship is linear, the linear regression model is the correct

model to use. The cubic regression model is overly complex. We would expect the training RSS for the linear regression model to be lower than the training RSS for the cubic regression model.

b) *Answer (a) using test rather than training RSS.*
The expectation remains the same when considering test RSS. Since the true relationship is linear, the linear model should generalize better to new, unseen data compared to the cubic one. Thus, we would expect the test RSS for the linear model to be lower than the test RSS for the cubic model.

c) *Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.*
The cubic regression model is more flexible and can capture nonlinear patterns in the data. The linear regression model is a subset of the cubic model (linear is a special case of cubic). In this scenario, we would expect the training RSS for the cubic regression model to be equal to or lower than the training RSS for the linear model. This is because the cubic model can fit both linear and nonlinear relationships, so it should be at least as good as the linear model.

d) *Answer (c) using test rather than training RSS.*
The expectation remains the same when considering test RSS. Since the cubic regression model can capture nonlinear patterns, it is expected to perform as well as or better than the linear regression model when dealing with test data, especially in cases where the true relationship is not linear. Therefore, we would expect the test RSS for the cubic model to be equal to or lower than the test RSS for the linear model.

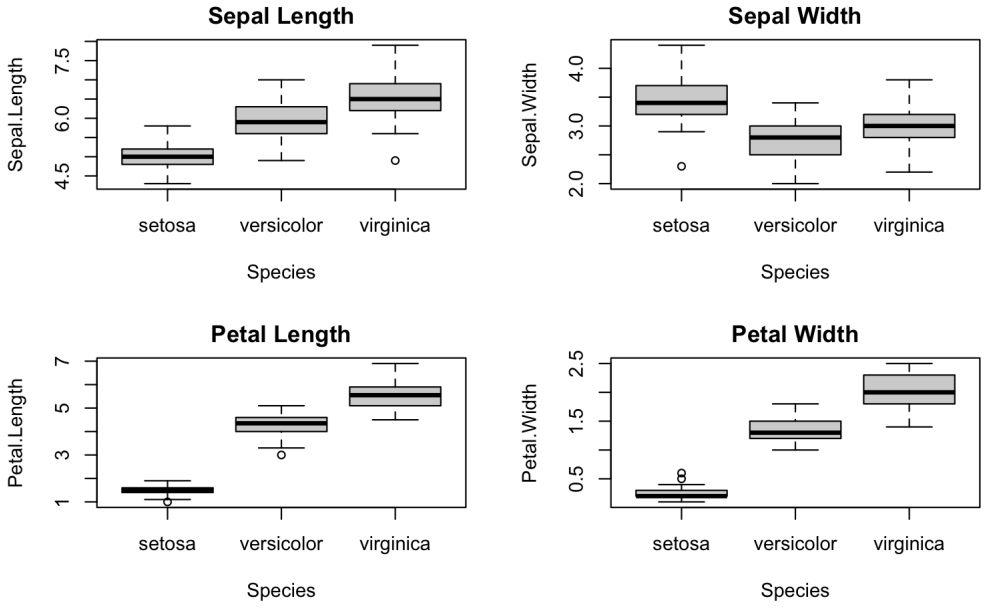
2 Practicum Problems

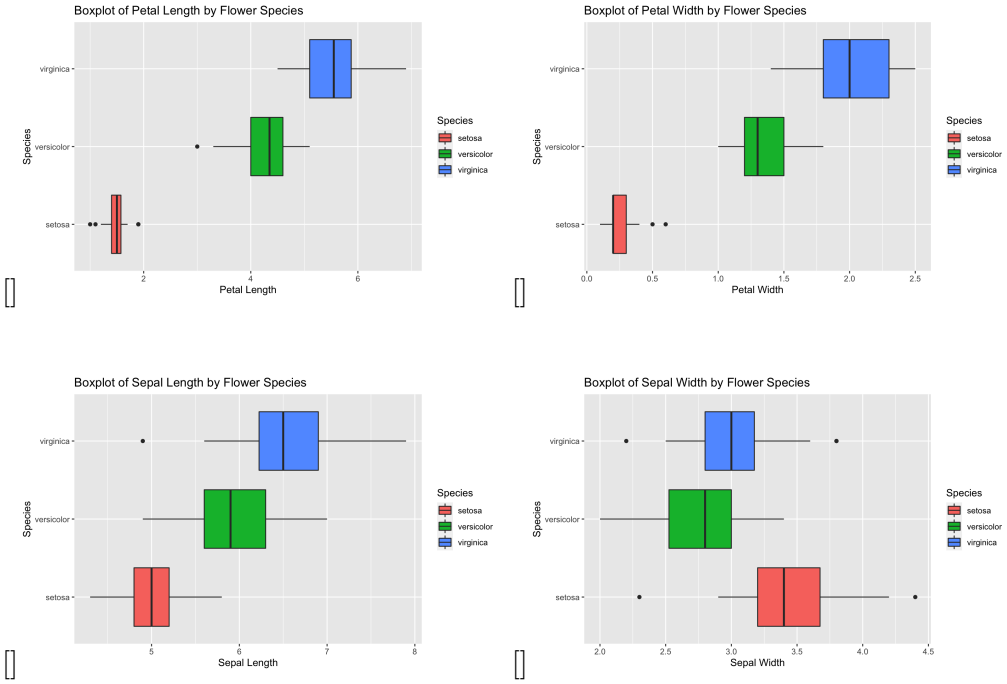
2.1 Problem 1

Load the iris sample dataset into R using a dataframe (it is a built-in dataset). Create a boxplot of each of the 4 features, and highlight the feature with the largest empirical IQR. Calculate the parametric standard deviation for each feature - do your results agree with the empirical values? Use the ggplot2 library from CRAN to create a colored boxplot for each feature, with a box-whisker per flower species. Which flower type exhibits a significantly different Petal Length/Width once it is separated from the other classes?
The feature with the largest empirical IQR is Petal Length.
We can compare the parametric standard deviations with the empirical IQR values. While parametric standard deviations and IQRs are different measures of dispersion, they both provide information about the spread of the data. However, they do not agree exactly, because of the data distribution having outliers.
Doing a couple linear models with the Petal Length and the Petal Width we can answer the last question. The p-values are less than a significance level of 0.05, it indicates that the each flower specie has a significantly different Petal Length and Width compared to the reference specie.

```
df=iris
par(mfrow = c(2, 2))
```

```r
boxplot(Sepal.Length ~ Species, data = df, main = "Sepal Length", xlab = "Species")
boxplot(Sepal.Width ~ Species, data = df, main = "Sepal Width", xlab = "Species")
boxplot(Petal.Length ~ Species, data = df, main = "Petal Length", xlab = "Species")
boxplot(Petal.Width ~ Species, data = df, main = "Petal Width", xlab = "Species")

sd1=sd(df$Sepal.Length)
sd2=sd(df$Sepal.Width)
sd3=sd(df$Petal.Length)
sd4=sd(df$Petal.Width)

ggplot(df, aes(x = Species, y = Petal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Petal Length by Flower Species", x = "Species", y = "Petal Length")+coord_flip()
ggplot(df, aes(x = Species, y = Petal.Width, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Petal Width by Flower Species", x = "Species", y = "Petal Width")+coord_flip()
ggplot(df, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Length by Flower Species", x = "Species", y = "Sepal Length")+coord_flip()
ggplot(df, aes(x = Species, y = Sepal.Width, fill = Species)) +
  geom_boxplot() +
  labs(title = "Boxplot of Sepal Width by Flower Species", x = "Species", y = "Sepal Width")+coord_flip()
IQR(iris$Sepal.Length)
IQR(iris$Sepal.Width)
IQR(iris$Petal.Length)
IQR(iris$Petal.Width)

model1 <- lm(Petal.Length ~ Species, data = iris)
summary(model1)

model2 <- lm(Petal.Width ~ Species, data = iris)
summary(model2)
```

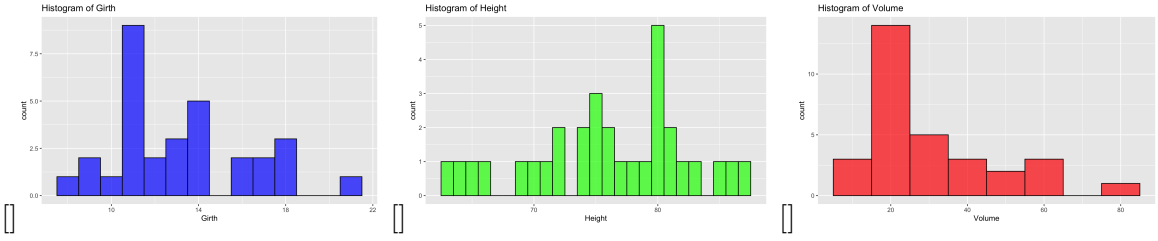




2.2 Problem 2

Load the trees sample dataset into R using a dataframe (it is a built-in dataset), and produce a 5-number summary of each feature. Create a histogram of each variable - which variables appear to be normally distributed based on visual inspection? Do any variables exhibit positive or negative skewness? Install the moments library from CRAN use the skewness function to calculate the skewness of each variable. Do the values agree with the visual inspection?

```
     Girth           Height       Volume     
 Min.   : 8.30   Min.   :63   Min.   :10.20  
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40  
 Median :12.90   Median :76   Median :24.20  
 Mean   :13.25   Mean   :76   Mean   :30.17  
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30  
 Max.   :20.60   Max.   :87   Max.   :77.00  
```



Non of them look similar to a normal distribution. The histogram of Girth and the histogram of Height do look more like a normal distribution than the histogram of Volume.

- Skewness of Girth: 0.5263163

- Skewness of Height: -0.374869
- Skewness of Volume: 1.064357

As we have discussed earlier non of them have Skewness close to 0 so the histogram does not have a shape like a normal distribution.

```
df=trees
ggplot(trees, aes(x = Girth)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Girth")

ggplot(trees, aes(x = Height)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Height")

ggplot(trees, aes(x = Volume)) +
  geom_histogram(binwidth = 10, fill = "red", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Volume")

install.packages("moments")
library(moments)

skew_girth <- skewness(trees$Girth)
skew_height <- skewness(trees$Height)
skew_volume <- skewness(trees$Volume)
```

2.3 Problem 3

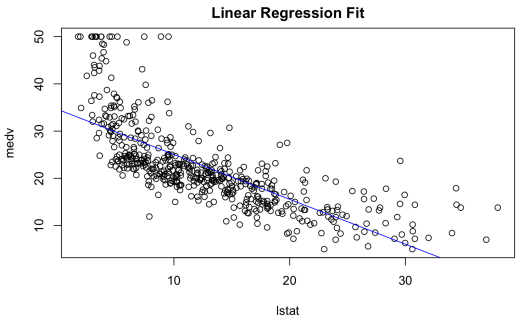
Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into R using a dataframe (Hint: You will need to use read.csv with url, and set the appropriate values for header,as.is, and sep). The horsepower feature has a few missing values with a ? - and will be treated as a string. Use the as.numeric casting function to obtain the column as a nu- meric vector, and replace all NA values with the median. How does this affect the value obtained for the mean vs the original mean when the records were ignored?
We obtain the same mean.

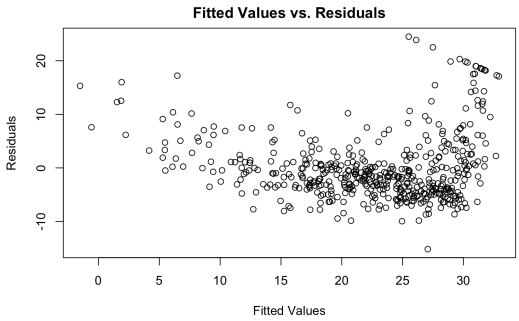
```
url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-
    mpg.data"

column_names <- c("mpg", "cylinders", "displacement", "horsepower", "weight", "
    acceleration", "model_year", "origin", "car_name")

auto_mpg <- read.csv(url, header = FALSE, sep = "", col.names = column_names)

# Horsepower column to a string
auto_mpg$horsepower <- as.character(auto_mpg$horsepower)

# Horsepower column to a numeric vector
auto_mpg$horsepower <- as.numeric(auto_mpg$horsepower)

# The median of the horsepower column
median_horsepower <- median(auto_mpg$horsepower, na.rm = TRUE)

# Replacing NA values with the median
auto_mpg$horsepower[is.na(auto_mpg$horsepower)] <- median_horsepower
```

2.4 Problem 4

Load the Boston sample dataset into R using a dataframe (it is part of the MASS package). Use lm to fit a regression between medv and lstat - plot the resulting fit and show a plot of fitted values vs. residuals. Is there a possible non-linear relationship between the predictor and response? Use the predict function to calculate values response values for lstat of 5, 10, and 15 - obtain confidence intervals as well as prediction intervals for the results - are they the same? Why or why not? Modify the regression to include lstat2 (as well lstat itself) and compare the R2 between the linear and non-linear fit - use ggplot2 and stat smooth to plot the relationship.

```
library(MASS)
data(Boston)
m1 <- lm(medv ~ lstat, data = Boston)
plot(Boston$lstat, Boston$medv, xlab = "lstat", ylab = "medv", main = "Linear
    Regression Fit")
abline(m1, col = "blue")
plot(fitted(m1), residuals(m1), xlab = "Fitted Values", ylab = "Residuals",
     main = "Fitted Values vs. Residuals")
```





We can see in the plot of fitted values vs. residuals the linear model might not be the most appropriate.

```
new_lstat <- data.frame(lstat = c(5, 10, 15))

# Prediction
predictions <- predict(m1, newdata = new_lstat, interval = "confidence")
conintervals <- as.data.frame(predictions)

predictions <- predict(m1, newdata = new_lstat, interval = "prediction")
predintervals <- as.data.frame(predictions)
```

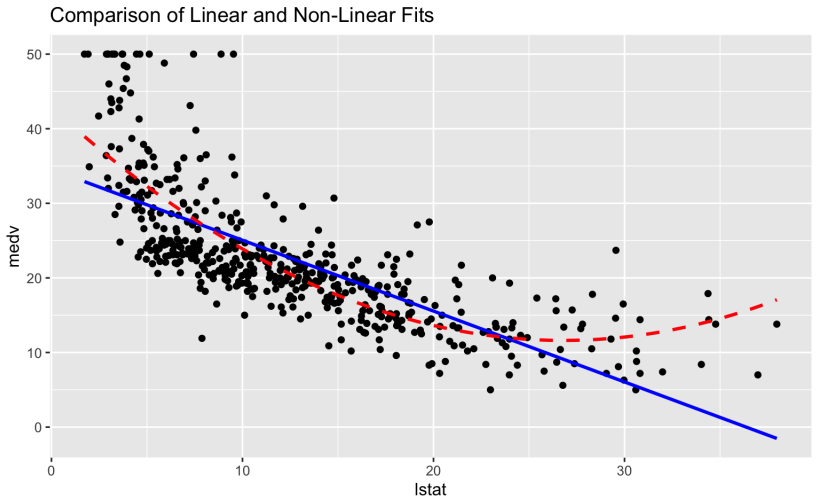
fit <dbl>	lwr <dbl>	upr <dbl>
29.80359	29.00741	30.59978
25.05335	24.47413	25.63256
20.30310	19.73159	20.87461

fit <dbl>	lwr <dbl>	upr <dbl>
29.80359	17.565675	42.04151
25.05335	12.827626	37.27907
20.30310	8.077742	32.52846

The intervals are not the same but the middle of the intervals are the same. They are not the same because prediction intervals account for the uncertainty in the model itself, in addition to the variability of the data.

```
m2 <- lm(medv ~ lstat + I(lstat^2), data = Boston)

summary(m1)$r.squared
summary(m2)$r.squared
```

```
5
6  ggplot(Boston, aes(x = lstat, y = medv)) +
7    geom_point() +
8    stat_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "blue",
       linetype = "solid", size = 1) +
9    stat_smooth(method = "lm", formula = y ~ poly(x, 2), se = FALSE, color = "red"
       , linetype = "dashed", size = 1) +
10   labs(title = "Comparison of Linear and Non-Linear Fits", x = "lstat", y = "
       medv")
```



The linear model has a R^2 of 0.5441463 and the non-linear model has a R^2 of 0.6407169. We can both see visually and analytically that the model improves if we fit a non-linear model.