KENTUCKY STATE UNIVERSITY, COMPUTER SCIENCE

Data Mining Midterm

Anil Varma B

October 11, 2019

***. Project: Select a dataset from internet and follow the instructions below: 1. Find missing data and replace/delete them 2. remove outliers 3. check data skewness... ideal data set should be normally distributed 4. build correlation table and find correlation variables 5 data partition – option 6. build model 7. assess the model**

Dataset Details: Advertising data set which has 3 dependent variable TV, radio, newspaper and one predictor variable sales. We use the Linear regression model to predict which advertisement dependent variables contribute to the increase in sales.

1 UNDERSTANDING THE DATA SET

1.1 Summary statistics of dataset

```
#Selecting the csv file
Advertising<-read.csv(file.choose(),head=TRUE)[-1]
summary(Advertising)
```

```
> summary(Advertising)
       TV              radio          newspaper          sales
 Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60
 1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38
 Median :149.75   Median :22.900   Median : 25.75   Median :12.90
 Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02
 3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40
 Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00
>
```

Figure 1.1: Dataset Summary

2 CHECKING FOR MISSING VALUES IN THE DATA SET

```
sum(is.na(Advertising))
```

```
> #Checking for missing values in the data set
> sum(is.na(Advertising))
[1] 0
>
```

Figure 2.1: Missing Values

No missing values found in the data set.

3 VISUALIZING THE DATA SET

3.1 Checking outliers with Boxpot

```
boxplot(Advertising)
```

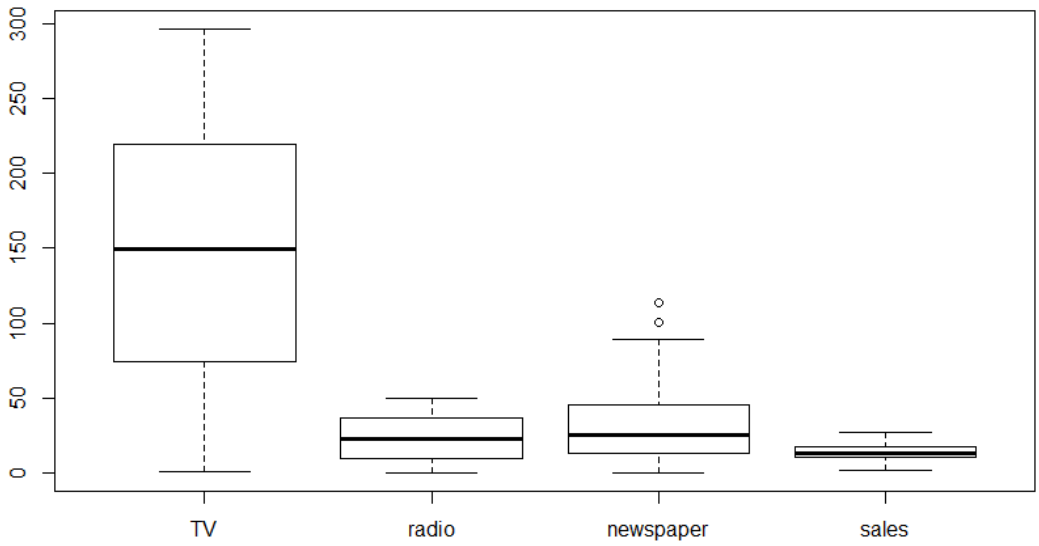

Figure 3.1: Boxplot

3.2 Density plots for checking skewness

```
library(e1071)
par(mfrow=c(2,2))
plot(density(Advertising$TV))
plot(density(Advertising$radio))
plot(density(Advertising$newspaper))
plot(density(Advertising$sales))
```

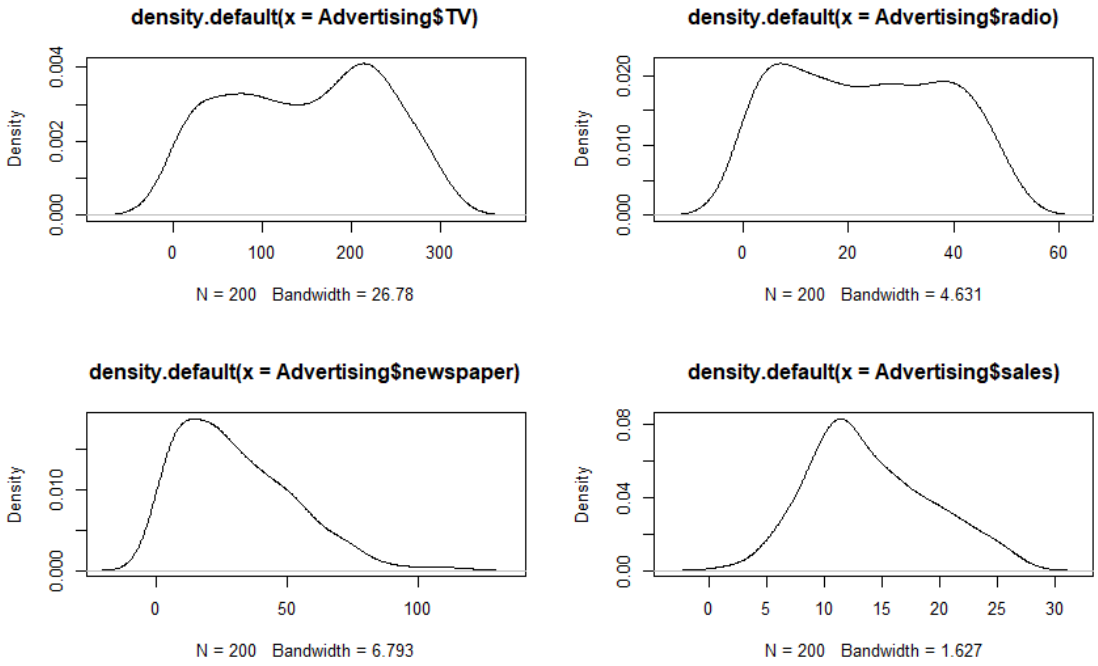

Figure 3.2: skew plots

3.3 Checking correlation and scatter plot using pairplots for checking the relationship

```
between variables
library(ggplot2)
library(GGally)
ggpairs(Advertising)
```

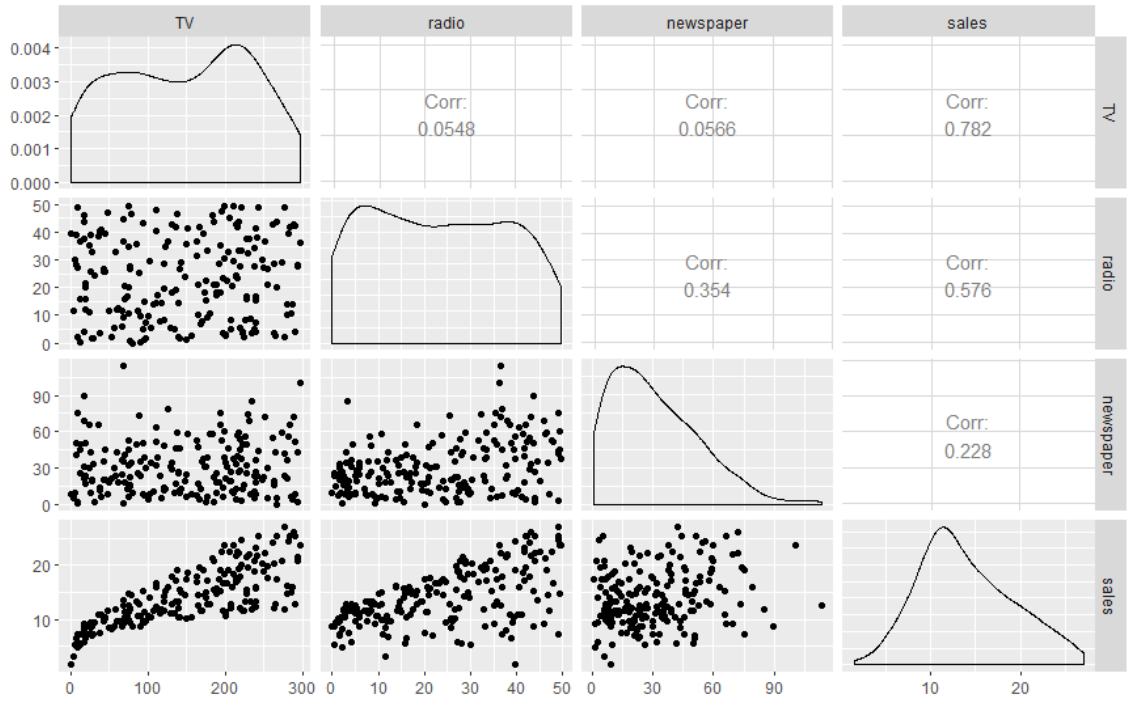

Figure 3.3: pairplots

From the above plot we can see that TV is having the highest correlation with Sales followed by radio and sales. So to begin with building a linear model, we will divide the data set into training set with 80% of the data and testing set with 20% of data. Before building the linear model using the training data we scale the data set.

4 DIVIDING THE DATASET INTO TRAIN AND TEST DATA

```
adRowCount <- floor(0.8 * nrow(Advertising))
adRowCount
set.seed(1) #Set seed to specify that all train and test set always have the same rows
adIndex <- sample(1:nrow(Advertising), adRowCount)
train <- Advertising[adIndex,]
test <- Advertising[-adIndex,]
```

4.1 Rescaling the values in train dataset for standardisation

```
train_scale <- scale(train,scale = TRUE)
```

4.2 Checking the scaled dataset

```
summary(train_scale)
boxplot(train_scale)
```

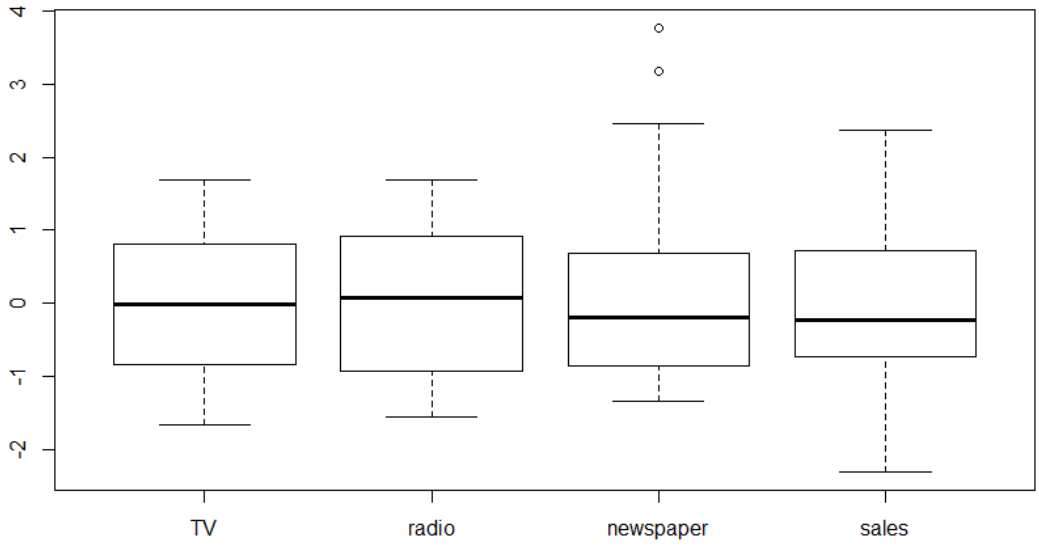

Figure 4.1: Boxplot for scaled training data set

We can see that the data set is scaled now and the data is in the same variance range.

```
#Checking the type of dataset
class(train_scale)
#converting the dataset to dataframe
df_train_scale <- data.frame(train_scale)
```

5 BUILDING THE REGRESSION MODELS

5.1 Building linear regression model - 1 with one variable TV

```
ad.lm <- lm(sales~TV, data = df_train_scale)
summary(ad.lm)
```

```
> #Building a linear regression model
> ad.lm <- lm(sales~TV, data = df_train_scale)
> summary(ad.lm)

Call:
lm(formula = sales ~ TV, data = df_train_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58041 -0.35977 -0.03881  0.39886  1.30181

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.271e-17  4.918e-02     0.0        1
TV           7.845e-01  4.934e-02    15.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6221 on 158 degrees of freedom
Multiple R-squared:  0.6154,    Adjusted R-squared:  0.613
F-statistic: 252.8 on 1 and 158 DF,  p-value: < 2.2e-16
```

Figure 5.1: Summary for Linear model

The linear regression model has high residual standard error and average R squared value of 0.61 which show that the model is not very good. So we will good ahead and create another model with adding the second variable.

5.2 Adding another variable, to build multiple Linear regression model - 2 with two variables TV and radio

```
ad.lm_2 <- lm(sales~TV+radio, data = df_train_scale)
summary(ad.lm_2)
```

```
> summary(ad.lm_2)

Call:
lm(formula = sales ~ TV + radio, data = df_train_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61314 -0.16827  0.04683  0.21956  0.53522

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.603e-17  2.538e-02    0.00        1
TV          7.503e-01  2.551e-02   29.41   <2e-16 ***
radio       5.329e-01  2.551e-02   20.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.321 on 157 degrees of freedom
Multiple R-squared:  0.8983,    Adjusted R-squared:  0.897
F-statistic:   693 on 2 and 157 DF,  p-value: < 2.2e-16
```

Figure 5.2: Summary for Linear model 2

The second linear regression model has lower residual standard error and a good R squared value of 0.9 which show that the model is very good. It also has a Fstatistic value and low P value which are good indicators of a model. Let us try to create one more model with adding the third variable.

5.3 Adding another variable, Linear regression model - 3 with three variables

```
ad.lm_3 <- lm(sales~TV+radio+newspaper, data = df_train_scale)
summary(ad.lm_3)
```

```
> ad.lm_3 <- lm(sales~TV+radio+newspaper, data = df_train_scale)
> summary(ad.lm_3)

Call:
lm(formula = sales ~ TV + radio + newspaper, data = df_train_scale)

Residuals:
     Min       1Q   Median       3Q      Max
-1.61868 -0.17139  0.04567  0.22534  0.53452

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.586e-17  2.546e-02   0.000    1.000
TV           7.506e-01  2.566e-02  29.248   <2e-16 ***
radio        5.344e-01  2.700e-02  19.795   <2e-16 ***
newspaper   -4.750e-03  2.706e-02  -0.176    0.861
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.322 on 156 degrees of freedom
Multiple R-squared:  0.8983,    Adjusted R-squared:  0.8963
F-statistic: 459.2 on 3 and 156 DF,  p-value: < 2.2e-16
```

Figure 5.3: Summary for Linear model 3

There is no change in the R squared value and standard error values on adding the third variable which shows that the third variable is not significant on the preictor variable.
So we will consider selection of the second linear model .

6 ASSESSING THE MODEL ACCURACY

6.1 Sales prediction using linear regression models

```
salespred1 <- predict(ad.lm,test)
salespred2 <- predict(ad.lm_2,test)
```

6.2 Calculating accuracy of prediction for linear model 1

```
# make actuals_predicteds dataframe.
actuals_preds <- data.frame(cbind(actuals=test$sales, predicteds=salespred1))
correlation_accuracy <- cor(actuals_preds)
correlation_accuracy #77\%
```

```
> correlation_accuracy #77%
              actuals predicteds
actuals     1.0000000  0.7735568
predicteds  0.7735568  1.0000000
```

Figure 6.1: Correlation accuracy for model 1

6.3 Calculating accuracy of prediction for linear model 2

```
actuals_preds2 <- data.frame(cbind(actuals=test$sales, predicteds=salespred2))
correlation_accuracy2 <- cor(actuals_preds2)
correlation_accuracy2 #83\%
```

```
> correlation_accuracy2 #83%
              actuals predicteds
actuals     1.0000000  0.8322338
predicteds  0.8322338  1.0000000
```

Figure 6.2: Correlation accuracy for model 2

Here we can see the prediction accuracy is very far for the first model. The second model is able to predict near to the model statistics and considered to be our model to be used for prediction.