Ansombe Datasets - Linear Regression Analysis

Eric B.

6/12/2022

Question 1: (50 points) As explained in Lesson 5, data exploration through visualization is important because statistics alone might not tell the entire story. This is best shown by the French statistician Francis Anscombe in 1973 when he presented four sets of data.

```
#reading in dataset from excel workout
library(readxl)
#excel workbook has multiple sheets in it that contain each of the four datasets used for this assignment
#each worksheet has a different name
sheet_names <- excel_sheets("C:/Users/esbro/Desktop/IE 575/Week 5/Anscombe.xlsx")
#generic function to get each dataset from each sheet in the workbook excel file
anscomData <- lapply(sheet_names, function(x) {          # Read all sheets to list
  as.data.frame(read_excel("C:/Users/esbro/Desktop/IE 575/Week 5/Anscombe.xlsx", sheet = x)) } )
anscomData
```

```
## [[1]]
##     x     y
## 1  10  8.04
## 2   8  6.95
## 3  13  7.58
## 4   9  8.81
## 5  11  8.33
## 6  14  9.96
## 7   6  7.24
## 8   4  4.26
## 9  12 10.84
## 10  7  4.82
## 11  5  5.68
##
## [[2]]
##     x    y
## 1  10 9.14
## 2   8 8.14
## 3  13 8.74
## 4   9 8.77
## 5  11 9.26
## 6  14 8.10
## 7   6 6.13
## 8   4 3.10
## 9  12 9.13
## 10  7 7.26
## 11  5 4.74
##
## [[3]]
##     x     y
## 1  10  7.46
## 2   8  6.77
## 3  13 12.74
## 4   9  7.11
## 5  11  7.81
## 6  14  8.84
## 7   6  6.08
## 8   4  5.39
## 9  12  8.15
## 10  7  6.42
## 11  5  5.73
##
## [[4]]
##     x     y
## 1   8  6.58
## 2   8  5.76
## 3   8  7.71
## 4   8  8.84
## 5   8  8.47
## 6   8  7.04
## 7   8  5.25
## 8  19 12.50
## 9   8  5.56
## 10  8  7.91
## 11  8  6.89
```

```
 names(anscomData)<-sheet_names #changing the default names for each dataset back to their orignial sheet name
```

Calculate the mean, variance, correlation, and a linear regression for each data set (No data partition). Using base R or ggplot2, create a visual representation of this data. What does this visualization show?

```
#Dataset 1 - calculations
anscomData$Data1
```

```
##     x     y
## 1  10  8.04
## 2   8  6.95
## 3  13  7.58
## 4   9  8.81
## 5  11  8.33
## 6  14  9.96
## 7   6  7.24
## 8   4  4.26
## 9  12 10.84
## 10  7  4.82
## 11  5  5.68
```

```
#mean calculations
Dat1x_mn<-mean(anscomData$Data1$x)
Dat1x_mn
```

```
## [1] 9
```

```
Dat1y_mn<-mean(anscomData$Data1$y)
Dat1y_mn
```

```
## [1] 7.500909
```

```
#variance calculations
Dat1x_var<-var(anscomData$Data1$x)
Dat1x_var
```

```
## [1] 11
```

```
Dat1y_var<-var(anscomData$Data1$y)
Dat1y_var
```

```
## [1] 4.127269
```

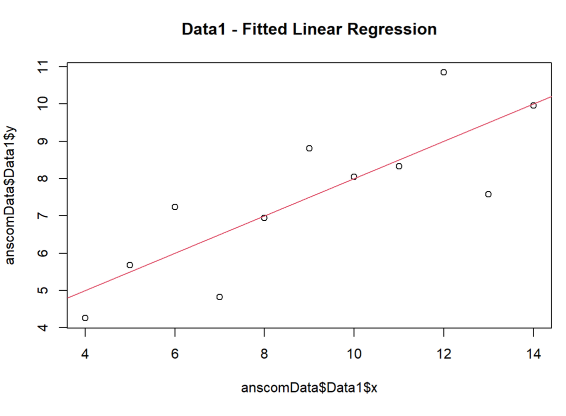
```
#correlation calculations
Dat1_corr<-cor(anscomData$Data1$x,anscomData$Data1$y)
Dat1_corr
```

```
## [1] 0.8164205
```

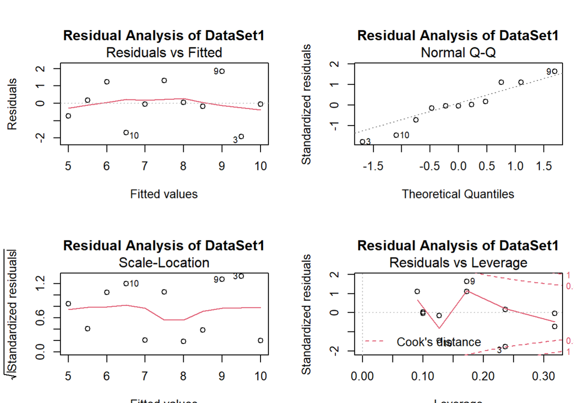
```
#creating linear regression model for dataset 1
Dat1_lreg<-lm(Data1$y~Data1$x,anscomData)
summary(Dat1_lreg)
```

```
##
## Call:
## lm(formula = Data1$y ~ Data1$x, data = anscomData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0001     1.1247   2.667  0.02573 *
## Data1$x       0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
#visualizations for dataset 1
plot(anscomData$Data1$x,anscomData$Data1$y,main="Data1 - Fitted Linear Regression")
abline(Dat1_lreg,col=10)
```




```
par(mfrow=c(2,2))
#residual analysis for dataset 1 - fitted regression model
plot(Dat1_lreg, main="Residual Analysis of DataSet1")
```




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

#Dataset 2 - calculations
anscomData$Data2
```

```
##     x    y
## 1  10 9.14
## 2   8 8.14
## 3  13 8.74
## 4   9 8.77
## 5  11 9.26
## 6  14 8.10
## 7   6 6.13
## 8   4 3.10
## 9  12 9.13
## 10  7 7.26
## 11  5 4.74
```

```
#mean calculations
Dat2x_mn<-mean(anscomData$Data2$x)
Dat2x_mn
```

```
## [1] 9
```

```
Dat2y_mn<-mean(anscomData$Data2$y)
Dat2y_mn
```

```
## [1] 7.500909
```

```
#variance calculations
Dat2x_var<-var(anscomData$Data2$x)
Dat2x_var
```

```
## [1] 11
```

```
Dat2y_var<-var(anscomData$Data2$y)
Dat2y_var
```

```
## [1] 4.127629
```

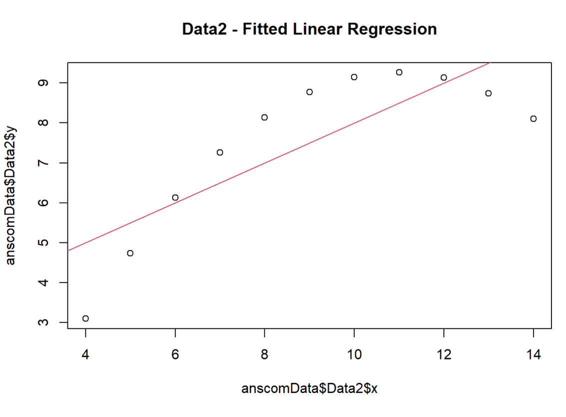
```
#correlation calculations
Dat2_corr<-cor(anscomData$Data2$x,anscomData$Data2$y)
Dat2_corr
```

```
## [1] 0.8162365
```

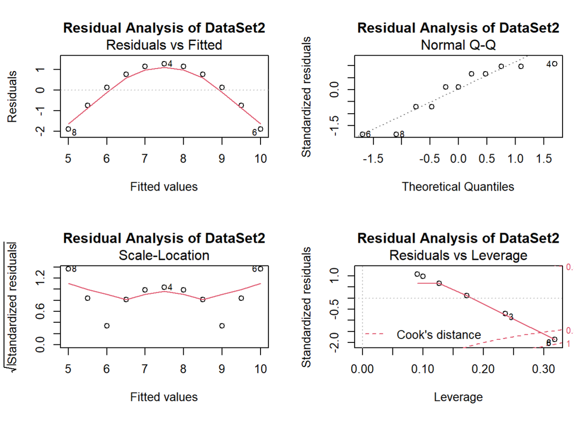
```
#creating linear regression model for dataset 2
Dat2_lreg<-lm(Data2$y~Data2$x,anscomData)
summary(Dat2_lreg)
```

```
##
## Call:
## lm(formula = Data2$y ~ Data2$x, data = anscomData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.001      1.125   2.667  0.02576 *
## Data2$x        0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```
#visualizations for dataset 2
plot(anscomData$Data2$x,anscomData$Data2$y,main="Data2 - Fitted Linear Regression")
abline(Dat2_lreg,col=10)
```




```
par(mfrow=c(2,2))
#residual analysis for dataset 2 - fitted regression model
plot(Dat2_lreg,main="Residual Analysis of DataSet2")
```




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

#Dataset 3 - calculations
anscomData$Data3
```

```
##     x     y
## 1  10  7.46
## 2   8  6.77
## 3  13 12.74
## 4   9  7.11
## 5  11  7.81
## 6  14  8.84
## 7   6  6.08
## 8   4  5.39
## 9  12  8.15
## 10  7  6.42
## 11  5  5.73
```

```
#mean calculations
Dat3x_mn<-mean(anscomData$Data3$x)
Dat3x_mn
```

```
## [1] 9
```

```
Dat3y_mn<-mean(anscomData$Data3$y)
Dat3y_mn
```

```
## [1] 7.5
```

```
#variance calculations
Dat3x_var<-var(anscomData$Data3$x)
Dat3x_var
```

```
## [1] 11
```

```
Dat3y_var<-var(anscomData$Data3$y)
Dat3y_var
```

```
## [1] 4.12262
```

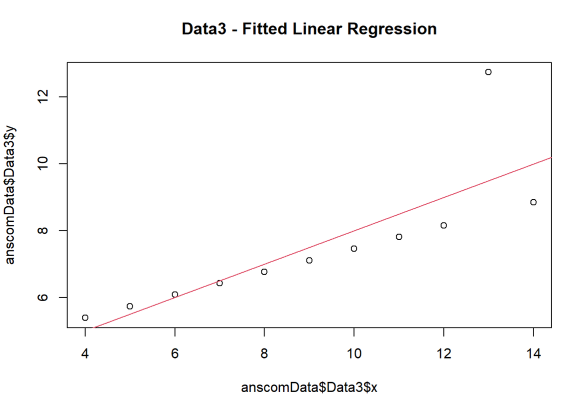
```
#correlation calculations
Dat3_corr<-cor(anscomData$Data3$x,anscomData$Data3$y)
Dat3_corr
```

```
## [1] 0.8162867
```

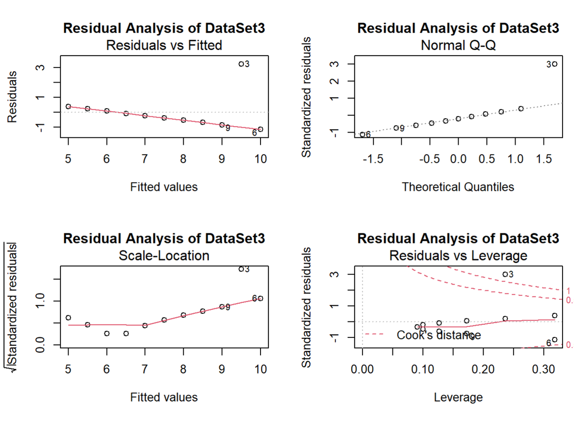
```
#creating linear regression model for dataset 3
Dat3_lreg<-lm(Data3$y~Data3$x,anscomData)
summary(Dat3_lreg)
```

```
##
## Call:
## lm(formula = Data3$y ~ Data3$x, data = anscomData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0025     1.1245   2.670  0.02562 *
## Data3$x       0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
#visualizations for dataset 3
plot(anscomData$Data3$x,anscomData$Data3$y,main="Data3 - Fitted Linear Regression")
abline(Dat3_lreg,col=10)
```




```
par(mfrow=c(2,2))
#residual analysis for dataset 3 - fitted regression model
plot(Dat3_lreg,main="Residual Analysis of DataSet3")
```




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

#Dataset 4 - calculations
anscomData$Data4
```

```
##     x     y
## 1   8  6.58
## 2   8  5.76
## 3   8  7.71
## 4   8  8.84
## 5   8  8.47
## 6   8  7.04
## 7   8  5.25
## 8  19 12.50
## 9   8  5.56
## 10  8  7.91
## 11  8  6.89
```

```
##mean calculations
Dat4x_mn<-mean(anscomData$Data4$x)
Dat4x_mn
```

```
## [1] 9
```

```
Dat4y_mn<-mean(anscomData$Data4$y)
Dat4y_mn
```

```
## [1] 7.500909
```

```
#variance calculations
Dat4x_var<-var(anscomData$Data4$x)
Dat4x_var
```

```
## [1] 11
```

```
Dat4y_var<-var(anscomData$Data4$y)
Dat4y_var
```

```
## [1] 4.123249
```

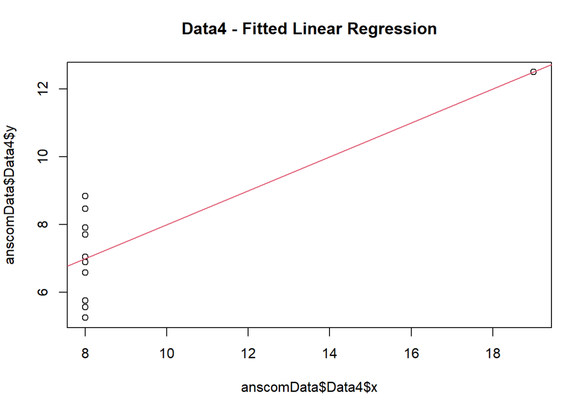
```
#correlation calculations
Dat4_corr<-cor(anscomData$Data4$x,anscomData$Data4$y)
Dat4_corr
```

```
## [1] 0.8165214
```

```
#creating linear regression model for dataset 4
Dat4_lreg<-lm(Data4$y~Data4$x,anscomData)
summary(Dat4_lreg)
```

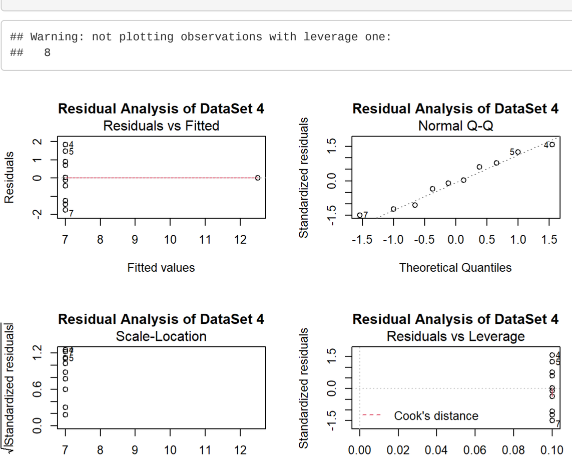
```
##
## Call:
## lm(formula = Data4$y ~ Data4$x, data = anscomData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0017     1.1239   2.671  0.02559 *
## Data4$x       0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

```
#visualizations for dataset 4
plot(anscomData$Data4$x,anscomData$Data4$y,main="Data4 - Fitted Linear Regression")
abline(Dat4_lreg,col=10)
```




```
par(mfrow=c(2,2))
#residual analysis for dataset 4 - fitted regression model
plot(Dat4_lreg,main="Residual Analysis of DataSet 4")
```

```
## Warning: not plotting observations with leverage one:
##   8
```




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

Response: The visualization shows that the Data 1 - dataset models a linear relationship between both variables (x & y). However, the Data 2 - dataset does not model a linear relationship. Since both the y and x variables do not seem to have a linear relationship between one other. The Data 3 - dataset, does meet the criteria of modelling a linear relationship, but one outlier data point does skew the fitted regression line. The outlier data point also drops the correlation coefficient for the linear model of the dataset by almost 0.2 points. Lastly, the Data 4 - dataset shows that there is no linear relationship between variables. However, there happens to be one extreme data point that raises the correlation coefficient to make it seem as though both variables are highly correlated to each other. As with regards, to the analysis of residuals, Dataset 1 the only dataset that resembles a linear model, does meet all residual assumptions. The Residuals vs Fitted plot indicates there is no nonlinear relationship, since all data points are equally distributed around the horizontal line without showing any pattern. The Normal Q-Q plot shows that the residuals are normally distributed. Then, the Scale-Location plot shows that the assumption of equal variance has not been violated. Lastly, the Residuals vs Leverage plot does not show any data points above the upper red dotted line or below the lower dotted line. Which means no outliers significantly affect the linear model of Dataset 1.

Question 2: (50 points) (25 points) How do you evaluate and compare the developed linear regression models in Question 1? Is there any issues with the models you built?

One could evaluate the developed linear regression models created in question 1 through different types of statistical calculations expressed through the summary function output shown for each model. The main statistical calculation that most statisticians use to evaluate linear regression models is the R^2-statistic and/or the adjusted R^2-statistic. Since the R^2-statistic shows "the amount of variance in output y that the developed linear regression model explains." In our example, Dataset 4's, linear model R^2-statistic happens to be the highest out of all four models with a value of 0.6667. Since, each model that we created does not have more than one feature/variable in its respective model, we do not have to worry about the adjusted R^2-statistic. Because the adjusted R^2 statistic is used mainly to compare models who have more the one feature in them. Though, there is a problem with the interpretation of the R-squared statistic value for dataset 4. Normally, we would say that 66.67% of the variability observed in y is explained by the regression model. However, as previously noted dataset 4 has an issue with an extreme outlier data point. This outlier data point happens to be raising the correlation coefficient to make it seem as though both variables are highly correlated to each other. Which as a result, happens to be raising the R-squared statistic of the model. To make the model, falsely more significant than the other three models. Consequently, out of the only two remaining datasets that model a linear relationship, dataset 1 & dataset 3, dataset 1 should really be considered the model with the highest variability explained in its regression model with a value of 66.65%.

(25 points) Is there any way to improve your linear regression models for Data I, II, and III?

Yes, there are ways to improve our linear regression models for the three datasets(I,II,III). One way we could help improve our linear regression model R-squared scores would be to remove outlier data points from the dataset. However, this method is not suggested when your dataset is based on a small sample size. Another way we could improve our linear regression models without decreasing the size of our current dataset, would be to floor capping outlier data values. This concept involves determining a percentile to cap data values at and changing all values that are above this percentile value to this percentile value. For example, we could establish the 75th percentile to be our highest percentile for our dataset. Thus, changing any data points above this percentile value to the same value as the 75th percentile value. Another way we could handle outlier data points to improve our linear regression models would be to change all outlier data points to our established median value of the dataset.