**Name: Nithin Das**

**CWID: 10422784**

**Date: 11/07/2019**

**BIA 652: FALL 2019**

**MIDTERM- FALL 2019**

**Problem 1 - (30 points)**

Use the MPG SAS dataset in canvas, build a multiple regression model for the purpose of predicting MPG using Weight, Horsepower and Age as predictors.

1. Is this a good regression model? Why and Why not? (Please be specific)
2. What are the 5 most influential points?
3. What are the 5 highest leverage point?
4. Which one of the three predictors has the highest impact on the prediction? Why?
5. Consider the car in the sample with: Age=3 Horsepower=100 and Weight=2420:
6. What is the actual MPG?
7. What is the prediction of MPG for a car with the same characteristics?
8. What is the 95% confidence interval for the MPG prediction?
9. What is the 95% confidence interval for averages of MPGs of cars with the same characteristics?

**SOLUTION: Refer Midterm Question\_1.R**

1. **Is this a good regression model? Why and Why not? (Please be specific)**

The summary of regression model is shown below:

Call:

lm(formula = MPG ~ weight + horsepower + age)

Residuals:

Min 1Q Median 3Q Max

-4.3594 -1.2462 0.0888 1.1499 3.7245

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 50.7258296 0.3224607 157.309 <2e-16 \*\*\*

weight -0.0060621 0.0001967 -30.816 <2e-16 \*\*\*

horsepower -0.0110112 0.0045408 -2.425 0.0158 \*

age -0.7691567 0.0250731 -30.677 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.65 on 388 degrees of freedom

Multiple R-squared: 0.9472, Adjusted R-squared: 0.9468

F-statistic: 2322 on 3 and 388 DF, p-value: < 2.2e-16

**R squared value**: 0.9472 or 94.72%. This means approximately 94% of variation in Miles per gallon can be explained using “age”, ”weight” and “horsepower”.

**Adjusted R-squared:** 0.9468 or 94.68%. R squared value adjusted for independent variables, this case 3. Adjust R squared value did not drop significantly from R squared value, showing this is a good regression fit.

**Residual standard error**: 1.65 on 388 degrees of freedom. This tells how far the observed MPG value is from fitted/predicted MPG values. Less residual standard error, better the model.

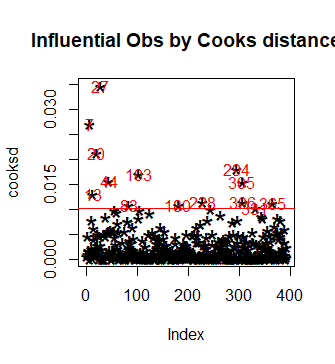
The ‘p-value’ for f statistic is also significant.

All the coefficients are also significant from the p values.

Therefore, this is a good regression model.

1. **What are the 5 most influential points?**

Influential points can be identified by applying cook’s distance method. Here, we set threshold as cook’s distance greater than 4 times the mean.



The 5 most influential points are :

ID cylinders horsepower weight acceleration age MPG Domestic

27 1027 8 200 4376 15.0 13 7.9 yes

7 1007 8 220 4354 9.0 13 14.7 yes

20 1020 4 46 1835 20.5 13 32.1 no

294 1294 4 80 1915 14.4 4 30.8 yes

103 1103 8 150 4997 14.0 10 13.7 yes

The regression equation with outliers is:

MPG= -( 0.006062\* weight) –(0.011011\* horsepower) –( 0.769157 \* age) + 50.725830

Let’s test each points:

(i)

ID cylinders horsepower weight acceleration age MPG Domestic

27 1027 8 200 4376 15.0 13 7.9 yes

Updated regression equation:

MPG= -(0.006086\* weight) –(0.009944\* horsepower) –( 0.767414 \* age) + 50.682437

Multiple R-squared: 0.9472, Adjusted R-squared: 0.9468

F-statistic: 2315 on 3 and 387 DF, p-value: < 2.2e-16

(ii)

ID cylinders horsepower weight acceleration age MPG Domestic

7 1007 8 220 4354 9.0 13 14.7 yes

Updated regression equation:

MPG=-(0.006029\*weight)- (0.012204 \* horsepower) –(0.769593 \* age) +50.747255

Multiple R-squared: 0.9473, Adjusted R-squared: 0.9469

F-statistic: 2319 on 3 and 387 DF, p-value: < 2.2e-16

(iii)

ID cylinders horsepower weight acceleration age MPG Domestic

20 1020 4 46 1835 20.5 13 32.1 no

Updated regression equation:

MPG= -(0.006066\*weight)-(0.010390\*horsepower)-(0.775045\*age)+50.706251

Multiple R-squared: 0.9476, Adjusted R-squared: 0.9472

F-statistic: 2333 on 3 and 387 DF, p-value: < 2.2e-16

(iv)

ID cylinders horsepower weight acceleration age MPG Domestic

294 1294 4 80 1915 14.4 4 30.8 yes

Updated regression equation:

MPG= -(0.006101\* weight) –(0.010347\* horsepower) –( 0.771765 \* age) + 50.802442

Multiple R-squared: 0.9481, Adjusted R-squared: 0.9477

F-statistic: 2358 on 3 and 387 DF, p-value: < 2.2e-16

(v)

ID cylinders horsepower weight acceleration age MPG Domestic

103 1103 8 150 4997 14.0 10 13.7 yes

Updated regression equation:

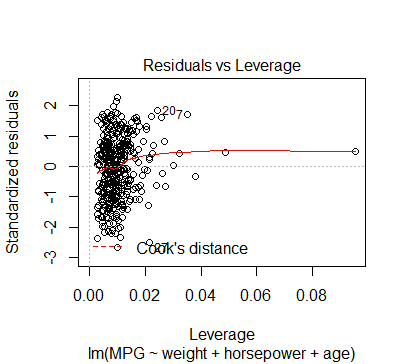
MPG= -(0.006107\* weight) –(0.010321\* horsepower) –( 0.770499 \* age) + 50.788576

Multiple R-squared: 0.9472, Adjusted R-squared: 0.9468

F-statistic: 2313 on 3 and 387 DF, p-value: < 2.2e-16

For all the regression equations above, we can observe that **slope changes every time** we remove the influential points.

**c)** **What are the 5 highest leverage point?**



The 5 highest leverage points are

MPG weight horsepower age

14 20.3 3086 225 13

116 15.3 4278 230 10

9 10.9 4425 225 13

7 14.7 4354 220 13

8 12.9 4312 215 13

The hat values for these points are high

**d)Which one of the three predictors has the highest impact on the prediction? Why?**

The coefficients of the regression equation cannot be directly used to find the important predictors. This is because the units and scales of the variables will be different for age, horsepower and weight.

Therefore, we standardize the variables by subtracting the mean of the variable from the value and then divide the result by variable’s standard deviation.

The updated regression equation is:

MPG= -(5.1492\* weight)-(0.4238\* horsepower)-(2.8334\* age) +26.1253

Here, the absolute coefficient value is higher for weight(5.1492). Therefore, weight is the important variable.

1. **Consider the car in the sample with: Age=3 Horsepower=100 and Weight=2420:**
2. What is the actual MPG?

Actual MPG= 33.6

1. What is the prediction of MPG for a car with the same characteristics?

MPG= -( 0.006062\* weight) –(0.011011\* horsepower) –( 0.769157 \* age) + 50.725830

Predicted MPG= -( 0.006062\* 2420) –(0.011011\* 100) –( 0.769157 \* 3) + 50.725830

* 32.6469

1. What is the 95% confidence interval for the MPG prediction?

fit lwr upr

332 32.64693 29.38645 35.90741

95% CI for MPG prediction is (29.28645,35.90741)

1. What is the 95% confidence interval for averages of MPGs of cars with the same characteristics?

**fit lwr upr**

**332 32.64693 32.32419 32.96967**

The 95% confidence interval for averages of MPGs is (32.32419,32.9696)

**Problem 2 - (20 points)**

Use the MPG SAS dataset in CANVAS; build the best multiple regression model for the purpose of predicting MPG using **only five predictors** and the following selection techniques. Specify whether the five predictor models are good models or not and why.

1. Forward
2. Stepwise
3. MaxR

SOLUTION: **SOLUTION: Refer Midterm Question\_2.R**

a)**Forward**

Subset selection object

Call: regsubsets.formula(MPG ~ cylinders + horsepower + weight + acceleration +

age + Domestic, data = data, nbest = 5, nvmax = 5, method = "forward")

6 Variables (and intercept)

Forced in Forced out

cylinders FALSE FALSE

horsepower FALSE FALSE

weight FALSE FALSE

acceleration FALSE FALSE

age FALSE FALSE

Domestic1 FALSE FALSE

5 subsets of each size up to 5

Selection Algorithm: forward

cylinders horsepower weight acceleration age Domestic1

1 ( 1 ) " " " " "\*" " " " " " "

1 ( 2 ) " " "\*" " " " " " " " "

1 ( 3 ) "\*" " " " " " " " " " "

1 ( 4 ) " " " " " " " " "\*" " "

1 ( 5 ) " " " " " " " " " " "\*"

2 ( 1 ) " " " " "\*" " " "\*" " "

2 ( 2 ) " " "\*" "\*" " " " " " "

2 ( 3 ) " " " " "\*" "\*" " " " "

2 ( 4 ) " " " " "\*" " " " " "\*"

2 ( 5 ) "\*" " " "\*" " " " " " "

3 ( 1 ) " " " " "\*" " " "\*" "\*"

3 ( 2 ) " " "\*" "\*" " " "\*" " "

3 ( 3 ) " " " " "\*" "\*" "\*" " "

3 ( 4 ) "\*" " " "\*" " " "\*" " "

3 ( 5 ) "\*" "\*" "\*" " " " " " "

4 ( 1 ) " " "\*" "\*" " " "\*" "\*"

4 ( 2 ) "\*" " " "\*" " " "\*" "\*"

4 ( 3 ) " " " " "\*" "\*" "\*" "\*"

4 ( 4 ) "\*" "\*" "\*" "\*" " " " "

5 ( 1 ) "\*" "\*" "\*" " " "\*" "\*"

5 ( 2 ) " " "\*" "\*" "\*" "\*" "\*"

5 ( 3 ) "\*" "\*" "\*" "\*" "\*" " "

From above, we can see that there are 3 subsets which has 5 important predictors. We must choose the best model among these 3 subsets.

The 3 subsets are:

1. MPG~ cylinders +horsepower+ weight +age+ Domestic

Rsq: 0.9601488

Adjusted R2: 0.9596326

Residual SS: 798.0579

BIC: -1227.4326

1. MPG~ horsepower+ weight + acceleration +age+ Domestic

Rsq : 0.9588984

Adjusted R2: 0.9583660

Residual SS: 823.0984

BIC: -1215.3220

1. MPG~ cylinders+ horsepower+ weight + acceleration +age

Rsq :0.9473446

Adjusted R2: 0.9466626

Residual SS: 1054.4736

BIC: -1118.2153

From the above metrics, it is clear that, MPG~ cylinders +horsepower+ weight +age+ Domestic is the best model with predictors **cylinders, horsepower, weight, age, Domestic**

Call:

lm(formula = MPG ~ cylinders + horsepower + weight + age + Domestic)

Residuals:

Min 1Q Median 3Q Max

-3.9792 -1.1673 0.0385 1.2068 3.7123

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 49.7995317 0.2931598 169.872 < 2e-16 \*\*\*

cylinders 0.3833604 0.1047847 3.659 0.000289 \*\*\*

horsepower -0.0177652 0.0041385 -4.293 2.24e-05 \*\*\*

weight -0.0057239 0.0002241 -25.537 < 2e-16 \*\*\*

age -0.7865398 0.0219547 -35.826 < 2e-16 \*\*\*

Domesticyes -2.1613802 0.1940807 -11.137 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.438 on 386 degrees of freedom

Multiple R-squared: 0.9601, Adjusted R-squared: 0.9596

F-statistic: 1860 on 5 and 386 DF, p-value: < 2.2e-16

This is a good model as the R squared and Adjusted r square is the maximum, and Residual sum of squares and BIC values are lower. Also, p value for F statistic is also significant. The p values of all coefficients are also significant.

**b) Stepwise**

Subset selection object

Call: regsubsets.formula(MPG ~ cylinders + horsepower + weight + acceleration +

age + Domestic, data = data, nbest = 5, nvmax = 5, method = "seqrep")

6 Variables (and intercept)

Forced in Forced out

cylinders FALSE FALSE

horsepower FALSE FALSE

weight FALSE FALSE

acceleration FALSE FALSE

age FALSE FALSE

Domestic1 FALSE FALSE

5 subsets of each size up to 5

Selection Algorithm: 'sequential replacement'

cylinders horsepower weight acceleration age Domestic1

1 ( 1 ) " " " " "\*" " " " " " "

1 ( 2 ) " " "\*" " " " " " " " "

1 ( 3 ) "\*" " " " " " " " " " "

1 ( 4 ) " " " " " " " " "\*" " "

1 ( 5 ) " " " " " " " " " " "\*"

2 ( 1 ) " " " " "\*" " " "\*" " "

2 ( 2 ) "\*" " " " " " " "\*" " "

2 ( 3 ) " " "\*" "\*" " " " " " "

2 ( 4 ) " " " " "\*" "\*" " " " "

2 ( 5 ) " " " " "\*" " " " " "\*"

3 ( 1 ) " " " " "\*" " " "\*" "\*"

3 ( 2 ) " " "\*" "\*" " " "\*" " "

3 ( 3 ) " " " " "\*" "\*" "\*" " "

3 ( 4 ) "\*" " " "\*" " " "\*" " "

3 ( 5 ) "\*" "\*" "\*" " " " " " "

4 ( 1 ) " " "\*" "\*" " " "\*" "\*"

4 ( 2 ) "\*" " " "\*" " " "\*" "\*"

4 ( 3 ) " " " " "\*" "\*" "\*" "\*"

4 ( 4 ) "\*" "\*" "\*" "\*" " " " "

5 ( 1 ) "\*" "\*" "\*" "\*" "\*" " "

The best model with 5 predictors is:

MPG~ cylinders+ horsepower+ weight + acceleration +age

Predictors : cylinders, horsepower, weight, acceleration, age

Rsq : 0.9473446

Adjusted R2: 0.9466626

Residual SS: 1054.4736

BIC: -1118.2153

Call:

lm(formula = MPG ~ cylinders + horsepower + weight + acceleration +

age)

Residuals:

Min 1Q Median 3Q Max

-4.3884 -1.1911 0.1211 1.1409 3.6348

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 50.7038745 0.9543779 53.128 <2e-16 \*\*\*

cylinders 0.1030777 0.1186119 0.869 0.3854

horsepower -0.0121157 0.0064667 -1.874 0.0617 .

weight -0.0062022 0.0002894 -21.434 <2e-16 \*\*\*

acceleration 0.0001405 0.0487900 0.003 0.9977

age -0.7708345 0.0252476 -30.531 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.653 on 386 degrees of freedom

Multiple R-squared: 0.9473, Adjusted R-squared: 0.9467

F-statistic: 1389 on 5 and 386 DF, p-value: < 2.2e-16

Though R squared and adjusted r square approximately 94% are good, p value for f statistic is significant, this is not a good model. As the p values of the coefficients of cylinders, horsepower, acceleration is greater than 0.05 and is not significant.

**c) maxR**

Subset selection object

Call: regsubsets.formula(MPG ~ cylinders + horsepower + weight + acceleration +

age + Domestic, data = data, nbest = 5, nvmax = 5, method = "exhaustive")

6 Variables (and intercept)

Forced in Forced out

cylinders FALSE FALSE

horsepower FALSE FALSE

weight FALSE FALSE

acceleration FALSE FALSE

age FALSE FALSE

Domestic1 FALSE FALSE

5 subsets of each size up to 5

Selection Algorithm: exhaustive

cylinders horsepower weight acceleration age Domestic1

1 ( 1 ) " " " " "\*" " " " " " "

1 ( 2 ) " " "\*" " " " " " " " "

1 ( 3 ) "\*" " " " " " " " " " "

1 ( 4 ) " " " " " " " " "\*" " "

1 ( 5 ) " " " " " " " " " " "\*"

2 ( 1 ) " " " " "\*" " " "\*" " "

2 ( 2 ) "\*" " " " " " " "\*" " "

2 ( 3 ) " " "\*" "\*" " " " " " "

2 ( 4 ) " " "\*" " " " " "\*" " "

2 ( 5 ) " " " " "\*" "\*" " " " "

3 ( 1 ) " " " " "\*" " " "\*" "\*"

3 ( 2 ) " " "\*" "\*" " " "\*" " "

3 ( 3 ) " " " " "\*" "\*" "\*" " "

3 ( 4 ) "\*" " " "\*" " " "\*" " "

3 ( 5 ) " " "\*" " " " " "\*" "\*"

4 ( 1 ) " " "\*" "\*" " " "\*" "\*"

4 ( 2 ) "\*" " " "\*" " " "\*" "\*"

4 ( 3 ) " " " " "\*" "\*" "\*" "\*"

4 ( 4 ) "\*" "\*" "\*" " " "\*" " "

4 ( 5 ) " " "\*" "\*" "\*" "\*" " "

5 ( 1 ) "\*" "\*" "\*" " " "\*" "\*"

5 ( 2 ) " " "\*" "\*" "\*" "\*" "\*"

5 ( 3 ) "\*" " " "\*" "\*" "\*" "\*"

5 ( 4 ) "\*" "\*" "\*" "\*" "\*" " "

5 ( 5 ) "\*" "\*" " " "\*" "\*" "\*"

which.max(b$rsq)

[1] 21

The model with maximum R squared value is:

MPG ~ cylinders+ horsepower+ weight+ age+ Domestic

Predictors are cylinders, horsepower, weight, age, Domestic

As seen in the forward selection, this is a good model with good r squared and adjusted r squared values. The p value of coefficients and f statistic are significant.

**Problem 3 - (20 points)**

Use the following ANOVA table of a regression model, produced by a software package, to answer the questions below.

|  |  |  |
| --- | --- | --- |
| **Source** | **DF** | **Sum of** |
| **Squares** |
| **Model** | 8 | 0.00218 |
| **Error** | 560 | 0.00294 |

1. How many observations are used for this model?
2. How many parameters are used in this model?
3. What is the R-square for this model?
4. What is the F value of the model?
5. Is the F value of the model significant at 5% level of significance? Why?

**Solution:**

1. **How many observations are used for this model?**

DF =Total records -1

Degree of freedom = 8+560

* 568

Total observations= 569

1. **How many parameters are used in this model?**

Total Parameters= 8+1 =>9

1. **What is the R-square for this model?**

R^2 = Sum of Squares(model) / Sum of Squares (total)

* (0.00218)/(0.00218+0.00294)
* 0.4257 or 42.57%

1. **What is the F value of the model?**

Mean Square(model)= Sum of squares(model)/DF

* 0.00218/8
* 0.0002725

Mean Square(error)= Sum of squares(error)/DF

* 0.00294/ 560
* 0.00000525

F value= Mean Square(model)/ Mean Square(error)

* 51.9047

1. **Is the F value of the model significant at 5% level of significance? Why?**

α = 0.05

from f distribution table,

F( 8,560,0.05)=1.954

F(calculated)>F(static)

The value is not significant at 5% level of significance.

**Problem 4 - (20 points)**

Using the lung dataset in canvas, perform the following regression analysis:

FEV on height and weight for the Mother.

1. Is this a good model? Why?
2. Are the residuals normally distributed? Why?
3. What are the influential observations? Why?
4. What are the high leverage observations? Why?
5. For predicting FEV of the mother, is height more important than weight? Why?

**SOLUTION: Refer Midterm Question\_4.R**

1. **Is this a good model? Why?**

Call:

lm(formula = MFEV1 ~ MHEIGHT + MWEIGHT, data = lungs\_data)

Residuals:

Min 1Q Median 3Q Max

-117.786 -29.401 0.148 25.896 142.212

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -279.70652 93.18014 -3.002 0.00315 \*\*

MHEIGHT 8.77729 1.51687 5.786 4.18e-08 \*\*\*

MWEIGHT 0.09834 0.12101 0.813 0.41771

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.36 on 147 degrees of freedom

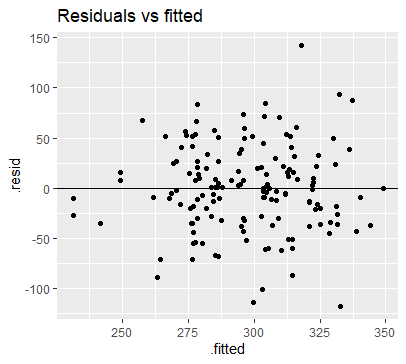
Multiple R-squared: 0.2193, Adjusted R-squared: 0.2087

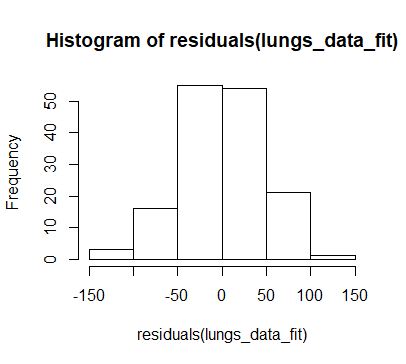
F-statistic: 20.65 on 2 and 147 DF, p-value: 1.251e-08

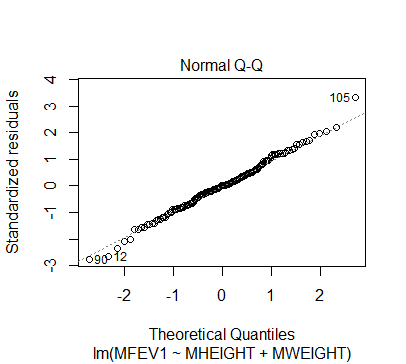
This is not a good model because Multiple R squared (21.97% )and Adjusted R squared value(20.87%) are very less. This means that height and weight variables can explain only 21.97% for variation in the target variable FEV1.

Also, the p-value for the coefficient of MWEIGHT is greater than 0.05 and is not significant.

**b)** **Are the residuals normally distributed? Why?**





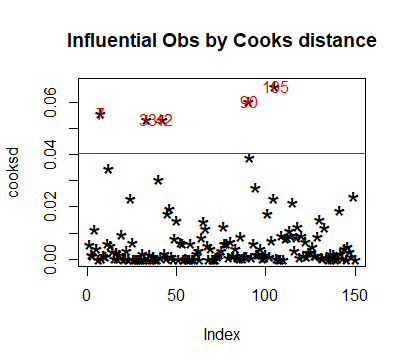


1.Histogram shows bell curve for residuals.

2. The Q-Q pot shows that almost all the points are fitted on the line.

Yes, the residuals are normally distributed because the residual error is random and do not follow any pattern in the residual v/s fitted plot.

**c) What are the influential observations? Why?**



We are setting the threshold as the cooks distance greater than 6 times the mean value.

The original regression equation is:

MFEV1= (8.77729\* MHEIGHT)+ (0.09834\* MWEIGHT) – 279.70652

ID MHEIGHT MWEIGHT MFEV1

7 7 68 206 425

33 33 60 168 175

42 42 65 241 254

90 90 68 159 215

105 105 66 185 460

These are the various influential points. Let us test one by one by removing from the data and fitting a new model

1. **Remove the record with ID 7 and fit the model**

ID MHEIGHT MWEIGHT MFEV1

7 7 68 206 425

The updated regression equation is:

**MFEV1= (8.5085\* MHEIGHT)+ (0.0675\* MWEIGHT) – 258.5555**

2. **Remove the record with ID 33 and fit the model**

ID MHEIGHT MWEIGHT MFEV1

33 33 60 168 175

The updated regression equation is:

**MFEV1= (8.258\* MHEIGHT)+ (0.125\* MWEIGHT) – 249.719**

3. **Remove the record with ID 42 and fit the model**

ID MHEIGHT MWEIGHT MFEV1

42 42 65 241 254

The updated regression equation is:

**MFEV1= (8.6596\* MHEIGHT)+ (0.1442\* MWEIGHT) – 278.4748**

4. **Remove the record with ID 90 and fit the model**

ID MHEIGHT MWEIGHT MFEV1

90 90 68 159 215

The updated regression equation is:

**MFEV1= (9.30908\* MHEIGHT)+ (0.09503\* MWEIGHT) – 312.499**

5. **Remove the record with ID 105 and fit the model**

ID MHEIGHT MWEIGHT MFEV1

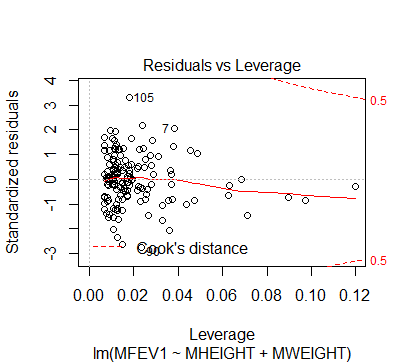
105 105 66 185 460

The updated regression equation is:

**MFEV1= (8.6101\* MHEIGHT)+ (0.06401\* MWEIGHT) – 264.91198**

**As we can see, the slope of the original regression equation changes every time we remove the outlier. Since these data points have impact on the slope, these are the high influential points.**

**d) What are the high leverage observations? Why?**



The sample 144,94,45 can be considered as high leverage points because these points has high hat values and can affect the regression line fit.

ID MHEIGHT MWEIGHT MFEV1

144 144 64 267 296

ID MHEIGHT MWEIGHT MFEV1

94 94 66 260 289

ID MHEIGHT MWEIGHT MFEV1

45 45 61 231 248

**e) For predicting FEV of the mother, is height more important than weight? Why?**

For identifying the important variable, first we will standardize the variable as height and weight are different quantity and units.

Once the variables are standardized, we can fit the model.

The updated regression equation is:

MFEV1= (21.676 \* MHEIGHT)+ (3.044\* MWEIGHT) +297.313

The absolute value of coefficient of Height is greater than that of Weight.

Therefore, Height variable is more important than Weight for predicting FEV1 OF mother.

|  |
| --- |
|  |

**Problem 5 - (10 points)**

1. Consider the following sample of 50 observations, is X normally distributed? Why?
2. What are the mean and variance of the X population?
3. If several samples of 36 observations are taken from X, what would be the expected mean and variance of the samples’ mean?

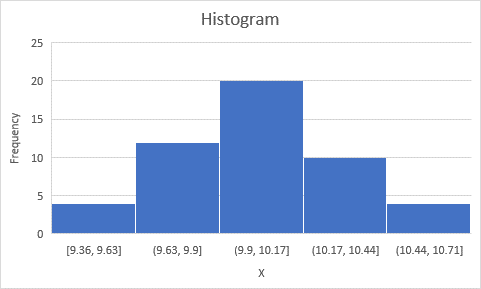
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Obs** | **X** |  | **Obs** | **X** |
| **1** | 9.8 |  | **26** | 10.56 |
| **2** | 10.2 |  | **27** | 10.47 |
| **3** | 10.27 |  | **28** | 9.42 |
| **4** | 9.7 |  | **29** | 10.44 |
| **5** | 9.76 |  | **30** | 10.16 |
| **6** | 10.11 |  | **31** | 10.11 |
| **7** | 10.24 |  | **32** | 10.36 |
| **8** | 10.2 |  | **33** | 9.94 |
| **9** | 10.24 |  | **34** | 9.77 |
| **10** | 9.63 |  | **35** | 9.36 |
| **11** | 9.99 |  | **36** | 9.89 |
| **12** | 9.78 |  | **37** | 9.62 |
| **13** | 10.1 |  | **38** | 10.05 |
| **14** | 10.21 |  | **39** | 9.72 |
| **15** | 10 |  | **40** | 9.82 |
| **16** | 9.96 |  | **41** | 9.99 |
| **17** | 9.79 |  | **42** | 10.16 |
| **18** | 10.08 |  | **43** | 10.58 |
| **19** | 9.79 |  | **44** | 10.7 |
| **20** | 10.06 |  | **45** | 9.54 |
| **21** | 10.1 |  | **46** | 10.31 |
| **22** | 9.95 |  | **47** | 10.07 |
| **23** | 9.84 |  | **48** | 10.33 |
| **24** | 10.11 |  | **49** | 9.98 |
| **25** | 9.93 |  | **50** | 10.15 |

**SOLUTION: (Refer Question\_5.xlsx)**

a) **Consider the following sample of 50 observations, is X normally distributed? Why?**

Normality can be tested using various methods:

1. Histogram



Histogram representation of X shows normal distribution curve or bell curve. If we fit line touching each bar, it will resemble that of a normal distribution curve.

Therefore, our data is normally distributed

1. Skew and Kurtosis

Skewness: this is a measure of symmetry

Kurtosis: this is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution

For normal distribution, skewness =0

For our data, skewness= -0.0651 approximately 0

Kurtosis= 0.03117

Therefore, our data is normally distributed

1. Z- value plot

Here, we arrange all observations in ascending order and create rank order accordingly. We use the formula, f= (i-0.375)/(n+0.25)

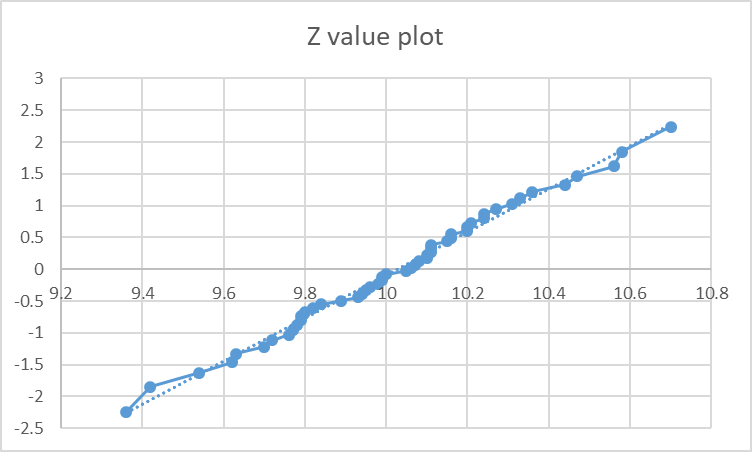
Where i is the rank

N is the number of observations

Now, calculate the z values for each f values.

Plot a scatter diagram with X values on x axis and Z values on y axis.

The plot shows that our data is normally distributed.



1. **What are the mean and variance of the X population?**

Mean x̄ = ( Σ xi ) / n

=10.0268

Variance = ∑ (X-Mean)^2/ N

=> 0.08181

1. **If several samples of 36 observations are taken from X, what would be the expected mean and variance of the samples’ mean?**

Since the population was normal, the sample distribution will also be normal with :

Mean(sample)=Mean(population)

Mean(sample)= 10.0268

Variance(sample)=Variance(population)/n

* 0.08181/36
* 0.0022725