STAT 441/541 Statistical Methods II

Homework Assignment 2 Simple Linear Regression

Submit a single pdf document to the Dropbox folder *Homework Assignment 2 Simple Linear Regression*.

Start each exercise on a new page.

Note: The scenarios and data are the same as in our text. Some questions have been modified.

Exercise 11.6

The R code file: Exercise 11-6 R Code.R

The dataset is Excel file: ex11-6.xlsx

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***x*** | 20 | 36 | 50 | 80 | 95 | 121 | 85 | 63 | 98 | 108 |
| ***y*** | 32 | 75 | 87 | 152 | 195 | 274 | 184 | 123 | 136 | 203 |

1. Construct, or paste plot from R output, the data on a scatterplot and describe the relationship between *x* and *y*.

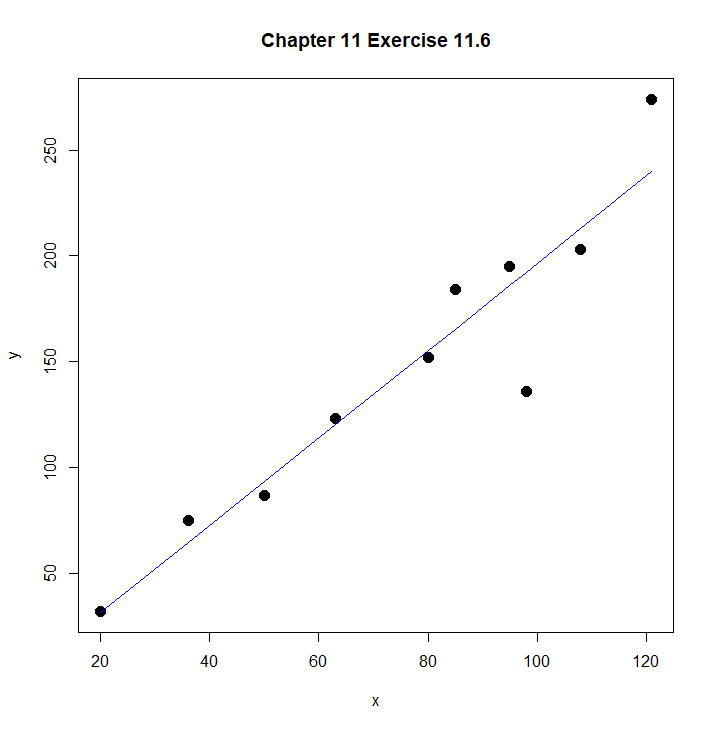


Figure above shows the relation between x and y. The relation between x and y is almost linear which means increase in x will lead to increase in y linearly and they make the resultant relation can be expressed as a straight line.

1. Paste the Coefficient Table from R output for the least-squares prediction and give the estimated linear regression model . (Rewrite this equation and plug in the estimated values of the intercept and slope.)

Coefficients:

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

(Intercept) -9.6627 20.9454 -0.461 0.657

x 2.0604 0.2564 8.035 4.23e-05 \*\*\*

.

(c) Using the Coefficient Table in part (b), test the hypothesis that there is a linear relationship between *y* and *x* at significance level . Perform this hypothesis test using the five-step method.

Hypotheses:

H0 : β1 = 0

Ha : β1 ≠ 0

Test Statistic:

To prove that β1 = 0 beyond a reasonable doubt we need to have p value of more than 0.05 as this is the threshold value as given in

P-value:

4.23e-05 = 0.0000423

Decision about the null hypothesis:

Given the p value very less than NULL hypothesis can not be rejected.

Conclusion:

The rejection of NULL hypothesis leads to adoption of alternate hypothesis which states β1 ≠ 0

Thus it can be inferred that there is a linear relationship

1. Paste the regression summary information below the Coefficient Table from R output and give the estimate of . What does this measure?

Residual standard error: 25.08 on 8 degrees of freedom

Multiple R-squared: 0.8898, Adjusted R-squared: 0.876

F-statistic: 64.56 on 1 and 8 DF, p-value: 4.231e-05

= 25.08 means the achieved tests scores. It is the average value by which the response variable will deviate from regression line.

1. Paste the R output for a 95% confidence interval on the true slope parameter for the independent variable *x* and interpret this confidence interval.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2.5 % | 97.5 % |  |
| Intercept | -57.962 | 38.637 |  |
| x | 1.469 | 2.651 |  |

CI for the slope estimate is 1.469 to 2.651. this means that we cay say with 95 percent confidence or we are 95 percent certain that value of of slope will be in this range. Likewise, we can also say that if different confidence intervals are taken for 95 percent of them will include true value of slope.

NOTE: Part of the next set of exercises asks for confidence and prediction bands for *y*. See the following website

[Predict in R: Model Predictions and Confidence Intervals](http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/)

The code at this site uses an R package called ggplot2. This will need to be installed before running the code provided on D2L.

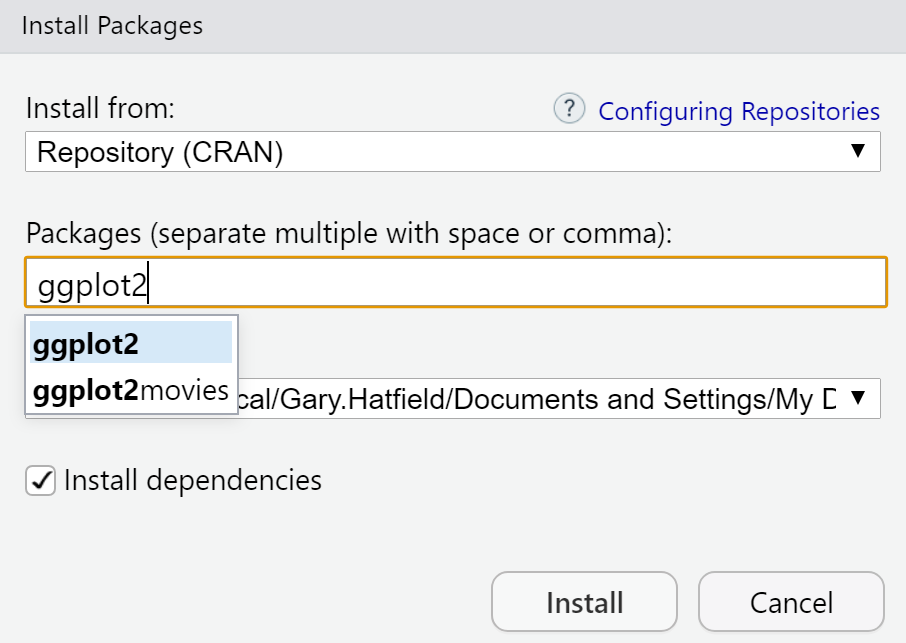
To install an R package using RStudio:

1. Run R Studio

2. Click on the Packages tab in the bottom-right section and then click on install.



3. A dialog box will appear. Type the package name to be installed and make sure the “Install dependencies” box is checked. Then click Install.



11.32 and 11.33

The R code file: Exercises 11-32 and 11-33 R Code.R

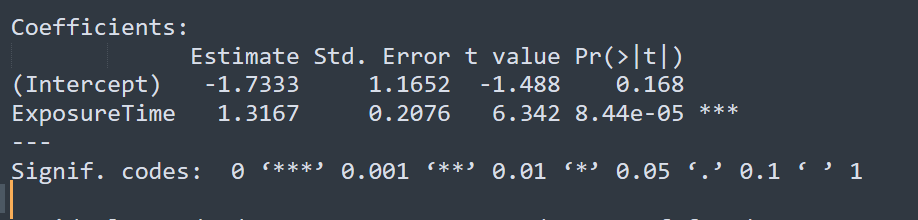
The dataset is Excel file: ex11-32.xlsx

A chemist is interested in determining the weight loss of a particular compound as a function of the amount of time the compound is exposed to air. The data give the weight losses associated with twelve settings of the independent variable, exposure time. So .

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Exposure Time, hours* | 4 | 5 | 6 | 7 | 4 | 5 | 6 | 7 | 4 | 5 | 6 | 7 |
| *Weight Loss, y, pounds* | 4.3 | 5.5 | 6.8 | 8.0 | 4.0 | 5.2 | 6.5 | 7.5 | 2.0 | 4.0 | 5.7 | 6.5 |

(a) Determine the least-squares prediction equation for the model .

Paste the Coefficient Table from R output for the least-squares prediction and give the estimated linear regression model . (Rewrite this equation and plug in the estimated values of the intercept and slope.)

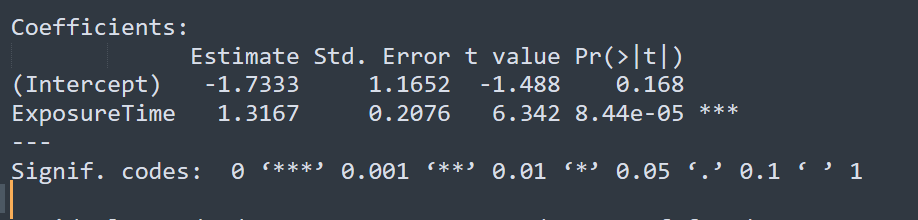


.

(b) Test ; give the p-value for at significance level and draw conclusions. Perform this hypothesis test using the five-step method. Note that this is a one-side test since we are only interested in a positive slope parameter.

Hypotheses:

Test Statistic:





P-value:

0.0000844

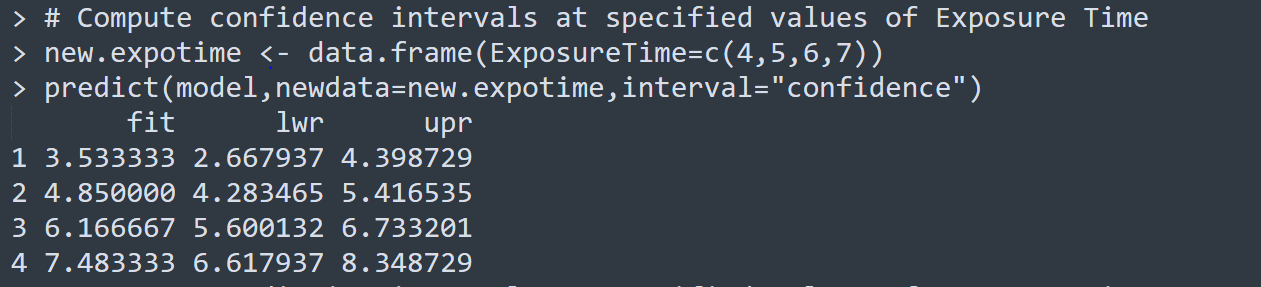
Decision about the null hypothesis:

Since p value is very very small as compared to value of . This indicates NULL hypothesis can be rejected based on p-value.

Conclusion:

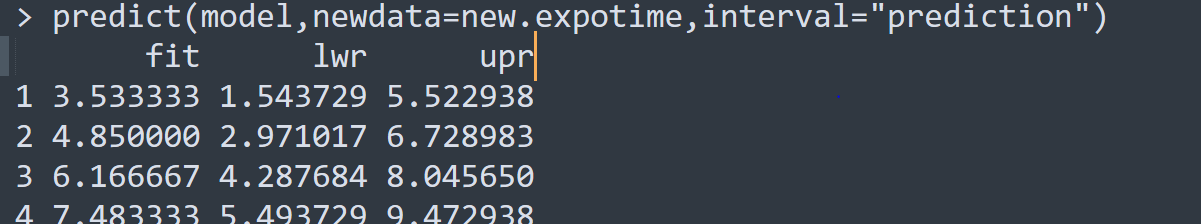
With rejection of NULL hypothesis, the alternate hypothesis is true which states

(c) Determine the 95% confidence intervals for when . Paste the R output for confidence intervals and interpret at least one of these confidence intervals.



Interpretation: for exposure time= 4, the fitted weight loss is 4.53. We are also 95 percent confident that fitted value lies in 2.66-4.39.

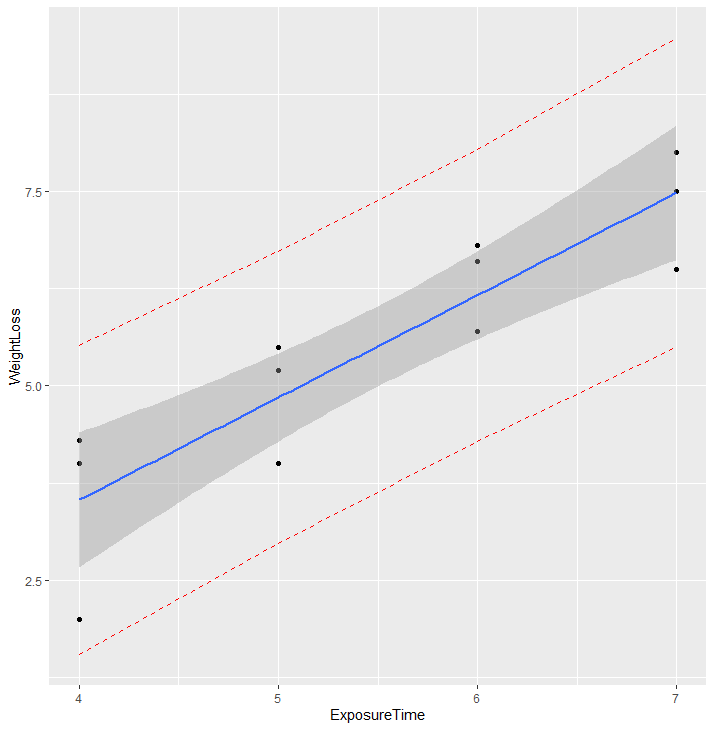
(d) Determine the 95% prediction intervals for when . Paste the R output for 95% prediction intervals and interpret at least one of these prediction intervals.



Interpretation: As we can see that the fit is same however the interval got wider. This additional margin in intervals is due to prediction standard error.

(e) Paste the plot from R output that give 95% confidence bands and 95% prediction bands

when . Comment on this plot and write a short essay explaining the difference between confidence bands and prediction bands.



Plot above depicts the weight loss with change in exposure time. The filled dark graph is showing the confidence intervals while the red dotted lines are showing prediction intervals. For exposure time of 4 hours with 95 percent confidence we can say that average weight loss is 2.66 to 4.39 pounds. The prediction interval as evident from the graph is always wider than the confidence interval. This is since it accounts for uncertainties in two types of estimations i.e., 1) population means 2) variation of individual values.

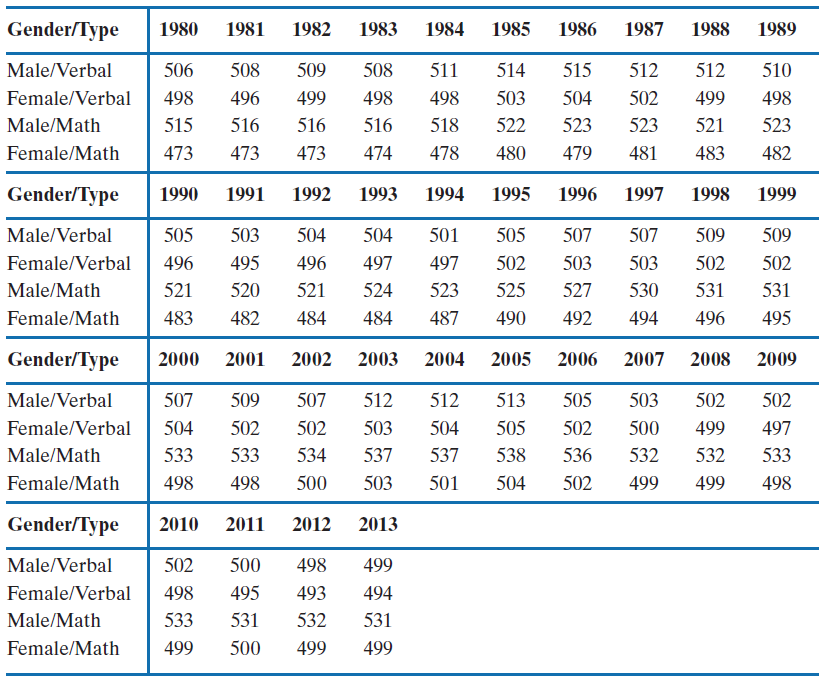
11.50

The R code file: Exercise 11-50 R Code.R

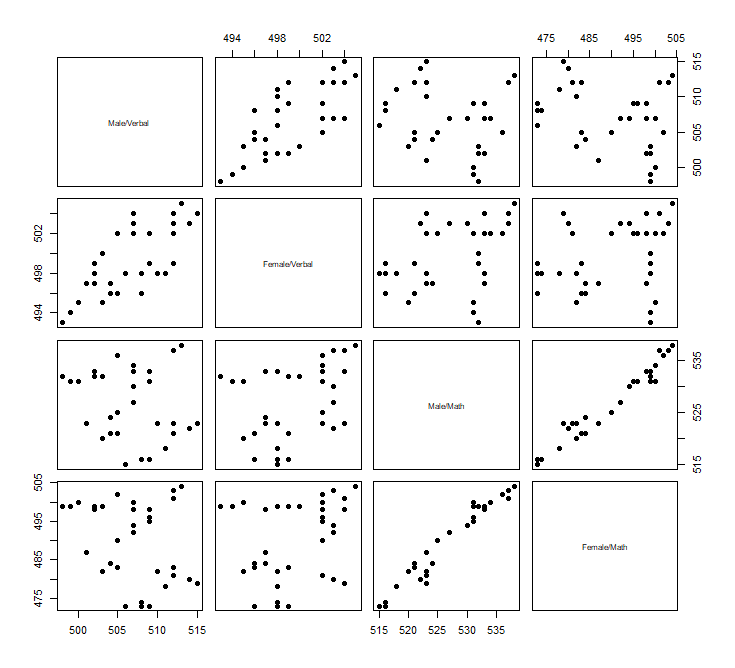
The dataset is Excel file: ex11-50.xlsx

NOTE: The code for this exercise uses an R package called Hmisc. This will need to be installed before running the code provided on D2L.

There has been an increasing emphasis in recent years on making sure that young women are given the same opportunities to develop their mathematical skills as young men are given in U.S. educational systems. The data provides SAT scores for male and female students over a 34-year period. There are scores for four Gender/Type variables: Male/Verbal, Female/Verbal, Male/Math, Female/Math.



1. Plot the six pairs of data values in a scatterplot matrix: Male/Verbal versus Female/Verbal, Male/Math versus Male/Verbal, and so on. Paste the plot of the scatterplot matrix from R output and comment on the nature of the relationship of the variables for each pair of variables. Note that there are six pairs.





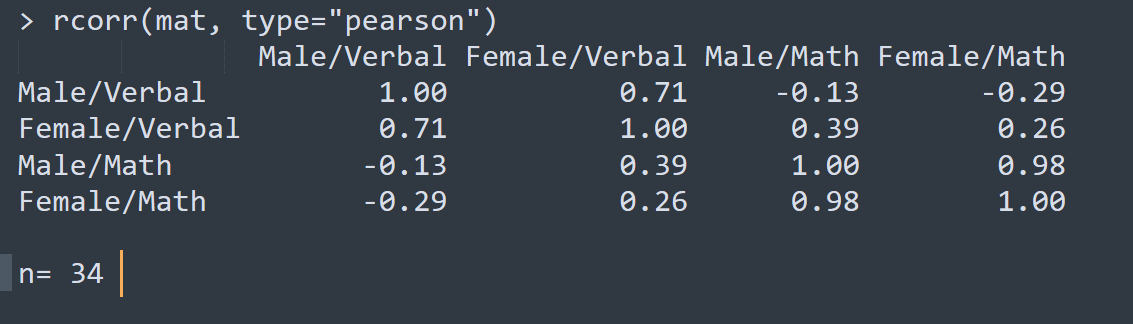
Note: For interpreting the graph also the correlations value were consulted in cases of weak or unclear relationships as sometime it was not easy to only interpret the graphs.

1. Male Verbal/Female Verbale: Strong positive linear relationship
2. Male Verbal/Male Math: Very weak negative or no relationship

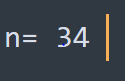


1. Male Verbal/Female Math: Weak negative relationship
2. Female Verbal/Male Math: Positive relationship but not strong
3. Female Verbal/Female Math: Weak positive relationships
4. Male Math/Female Math: Very strong positive relationship
5. Which, if any, of the six correlations are significantly different from 0 at the 5% level? Paste the R output that gives the correlation matrix, number of observations, and p-values for the corresponding correlation coefficients.

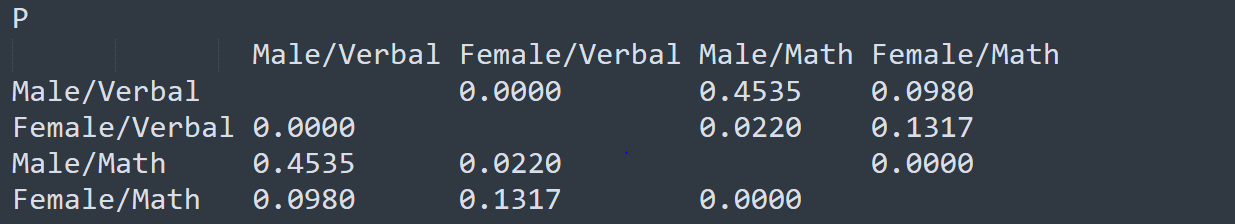
Correlation Matrix:



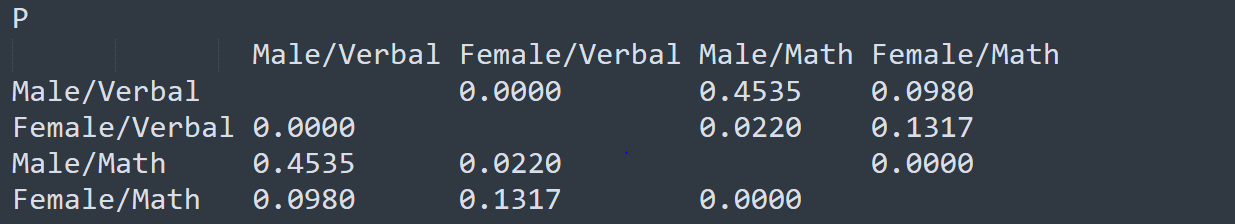
Number of observations :



P values :



Which pairs of variables have a correlation that is significant at the 5% level?





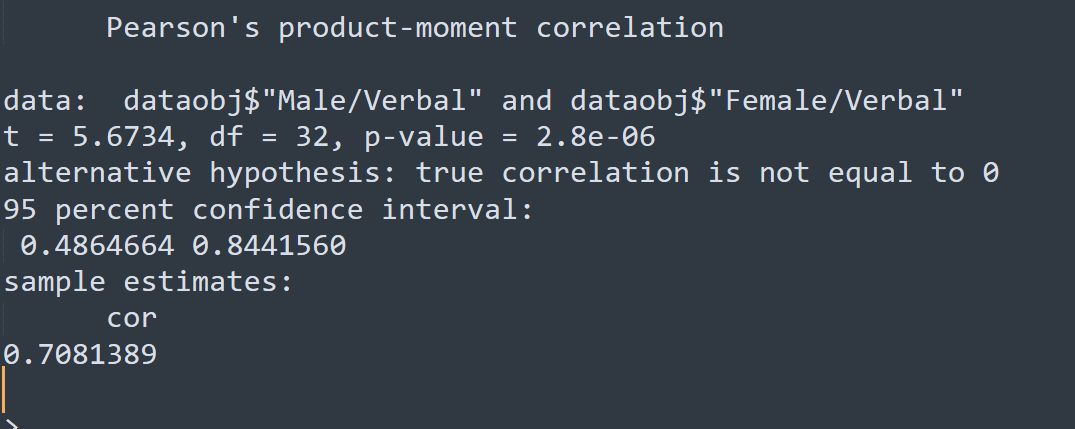
For one of the correlations that is significant at the 5% level, perform a hypothesis test of the population correlation coefficient using the five-step method.

1. Male Verbal/Female Verbale
2. Female Verbal/ Male Math
3. Male Math/Female Math

Note: It must be noted that the most significant at 5 percent also are the one who have a strongest relationship.

For one of the correlations that is significant at the 5% level, perform a hypothesis test of the population correlation coefficient using the five-step method.

Q. Test the hypothesis that there is a linear relationship between Male Verbal / Female Verbal variables at significance level .





Hypotheses:

H0 = p value is greater than or equal to 0.05

Ha = p value is less than 0.05

Test Statistic: 5.6734

P-value: 0.0000028

Decision about the null hypothesis:

Since p value is very less than 0.05 it stands rejected

Conclusion:

The positive linear relationship is statistically significant.

1. Do the plots reflect the sizes of the correlations between the pairs of variables?

Although in case of not very strong relationships the plots were not very clear however this was not the case in several other places such as where the relationship was strong enough. So the answer is yes, the plots reflects the size of the correlations specially in places where the relationship was very stronger such as Male Math/Female Math.