STAT 511A Homework 11

Kathleen Wendt

12/09/2019

Question 1: Running

Review problem 11.22 from Ott & Longnecker regarding treadmill "time to exhaustion" (X) and 10km race times (Y).

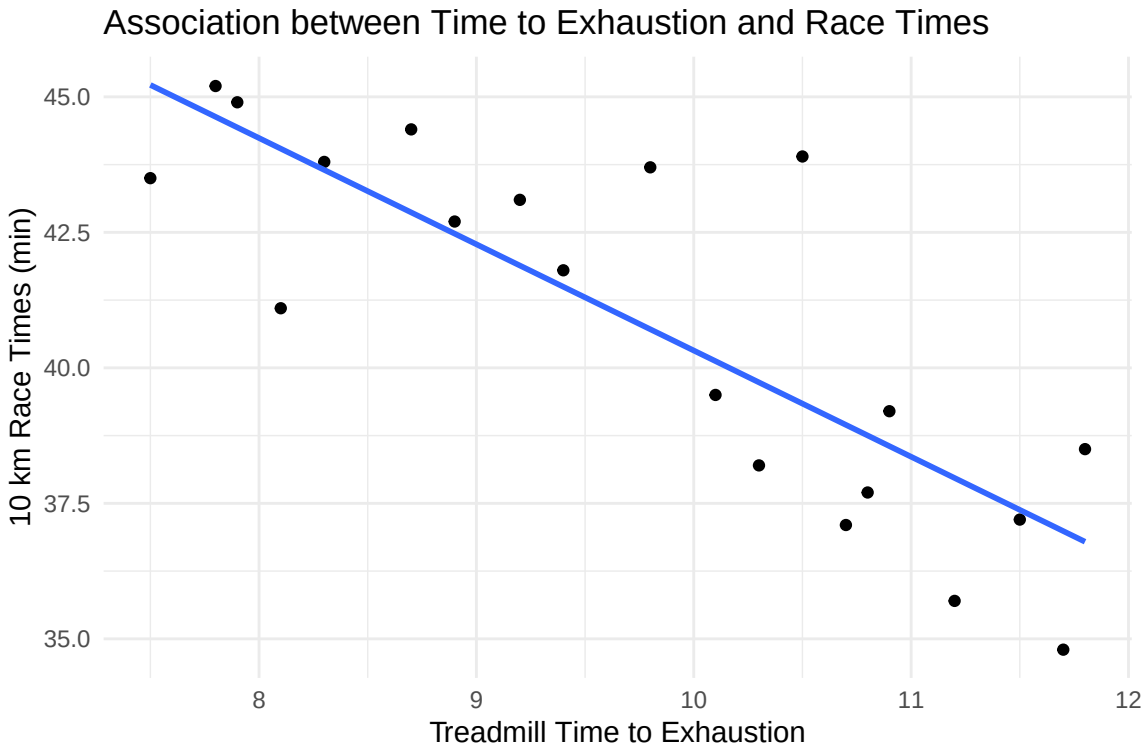
Part 1A: Linear Regression

Regress 10.K (Y) on Treadmill (X) and include the "summary" information in your assignment.

```
##
## Call:
## lm(formula = race_time ~ treadmill, data = run_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9440 -1.5788  0.1860  0.7863  4.5603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.9211     3.1166  19.226 1.90e-13 ***
## treadmill    -1.9601     0.3164  -6.194 7.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.921 on 18 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6629
## F-statistic: 38.37 on 1 and 18 DF,  p-value: 7.589e-06
```

Part 1B: Scatterplot

Create a scatterplot of 10-K vs Treadmill with fitted regression line overlaid.



Part 1C: Slope

Give the estimate, 95% confidence interval and interpretation of the slope. (4 pts)

Estimate

The estimate for the slope is -1.9601351.

95% Confidence Interval

The 95% confidence interval for the slope is (-2.6249573, -1.295313).

Interpretation

A difference of a one unit increase in treadmill time to exhaustion is, on average, associated with a -1.9601351 minute decrease in 10 km race time in this sample. There is a negative, linear association between treadmill time to exhaustion and 10 km race time.

Part 1D: R-squared

Give the R^2 value and interpretation in terms of this scenario.

$R^2 = 0.6807$

Interpretation

68% of the variance in 10 km run times can be attributed to the model relating treadmill time to exhaustion and 10 km run times.

Part 1E: Prediction

Give the predicted 10.K time for a runner with Treadmill = 11.

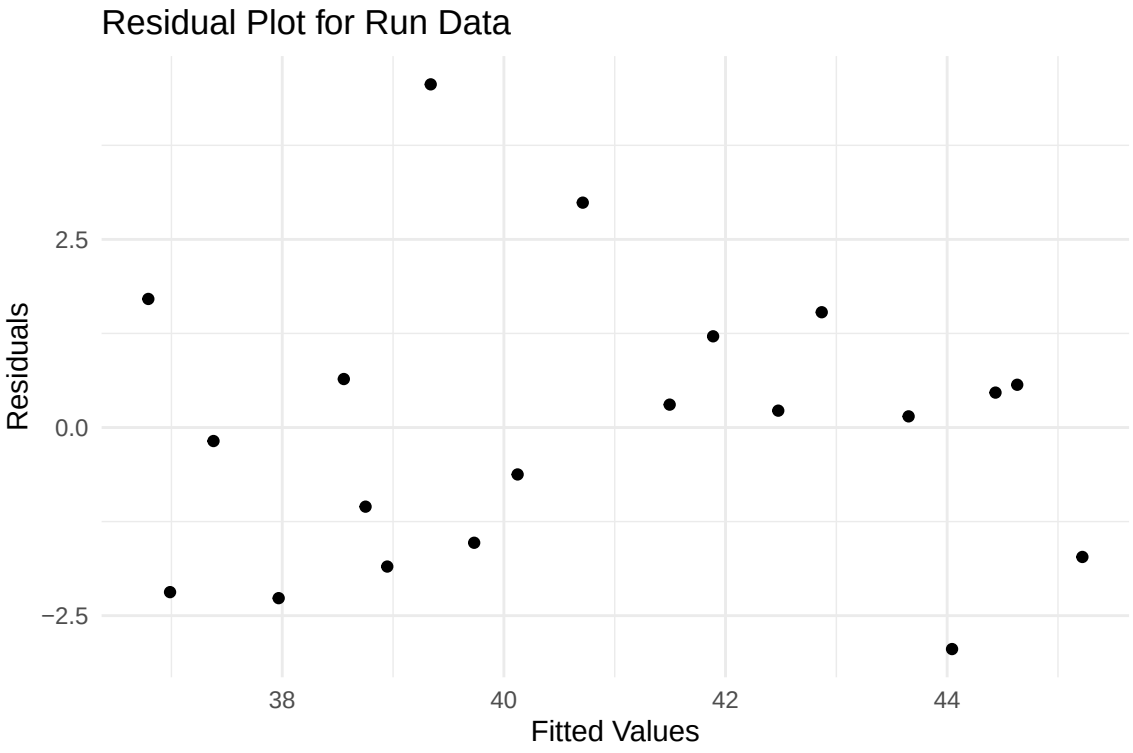
38.3596317 minutes

Also provide a corresponding prediction interval.

(37.1351131, 39.5841504)

Part 1F: Assumptions

Create the plots of (1) residuals vs fitted values and (2) qqplot of residuals.



Part 1G: Outlier

Based on the plots above, subject 13 appears to be a bit of an outlier. Run a formal outlier test for this observation. Provide the p-value and make a conclusion. Note that since we identified this observation after looking at the data, a Bonferonni adjustment is appropriate.

Using `car::outlierTest` to identify the observation with the most extreme residual, the Bonferonni p-value for observation 13 is $0.18867 > \alpha = 0.05$. Based on this, we fail to reject the null hypothesis that observation 13 is NOT an outlier.

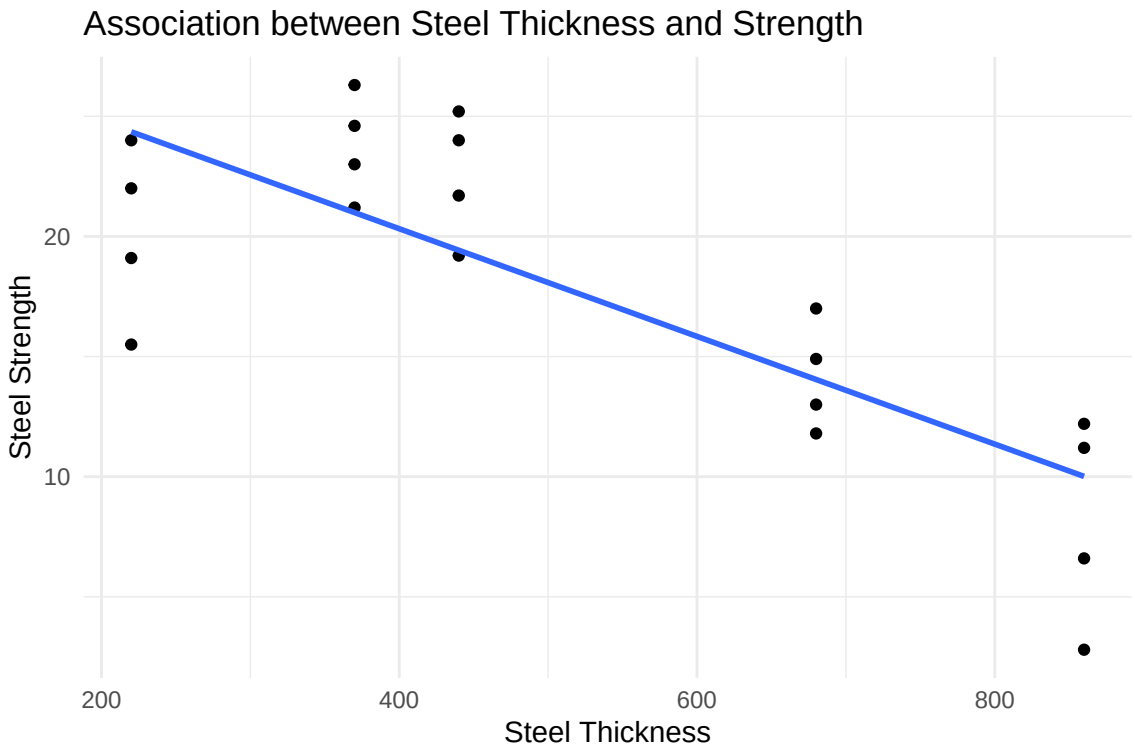
Question 2: Steel

Data on age in coating Thickness (X) and Strength (Y) from an experiment involving steel are available from Canvas as Steel.csv.

Part 2A

Regress Strength (Y) against Thick (X) and look at (1) the plot of Strength versus Thick (2) residuals versus predicted values and (3) qqplot of residuals. Include these plots in your assignment. Do the regression assumptions appear to be met? Discuss. (4 pts)

Steel Data

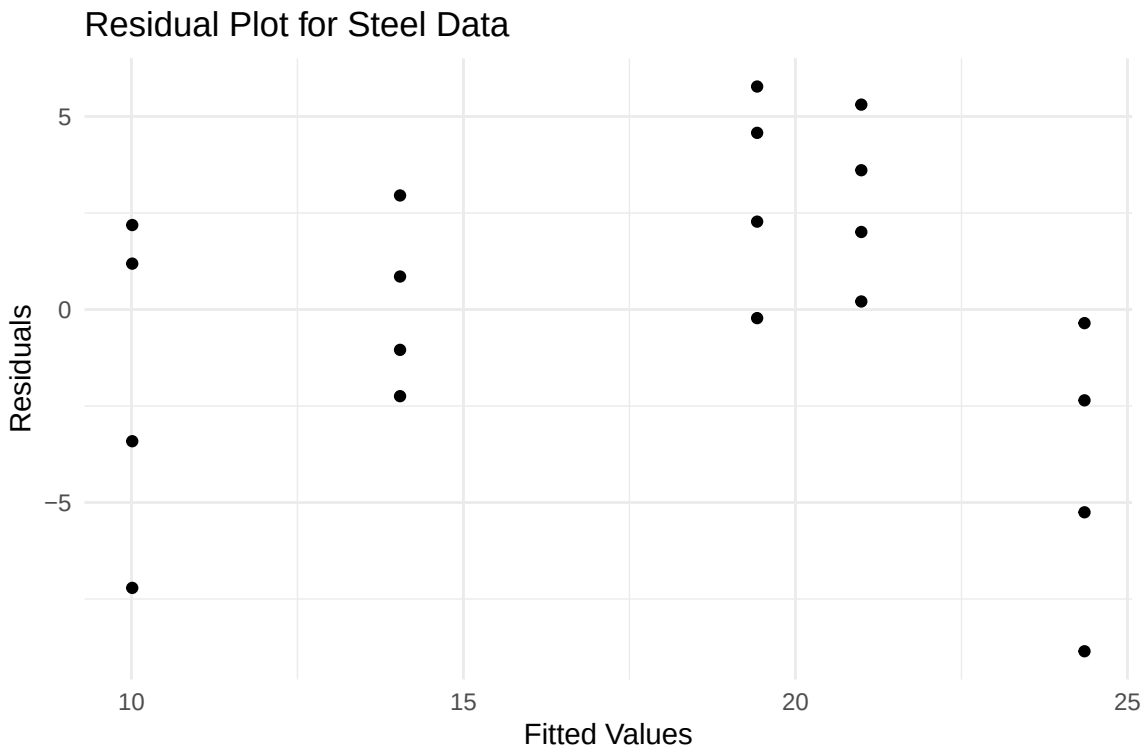


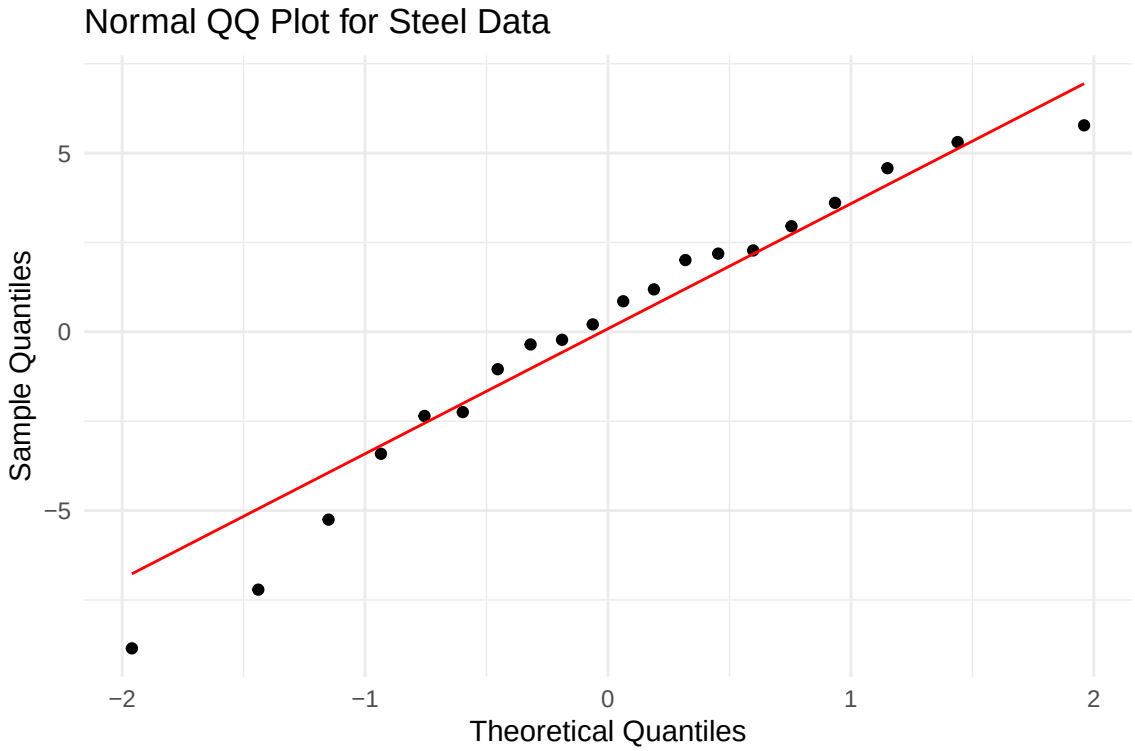
Linear Regression

```
##
## Call:
## lm(formula = strength ~ thick, data = steel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8530 -2.2722  0.5315  2.4463  5.7768
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.282737   2.258464  12.966 1.44e-10 ***
## thick       -0.022408   0.004016  -5.579 2.70e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.095 on 18 degrees of freedom
## Multiple R-squared:  0.6336, Adjusted R-squared:  0.6132
## F-statistic: 31.13 on 1 and 18 DF,  p-value: 2.699e-05
```

Check Assumptions





Based on the above plots and additional histograms, there is evidence against the assumptions for simple linear regression. Specifically, the assumption of linearity is challenged by the curvilinear/quadratic "trend" in the plot of residual vs. fitted values. Furthermore, the assumption of equal variances is challenged by the "categorical" pattern in the first plot. Regarding the Q-Q plot to assess normality of residuals, there is some evidence for heavy tails and against the assumption of normality. A simple linear regression is not an appropriate analytic technique for these data.

Part 2B: F-test for Fit

Perform an F-test for "lack of fit". Give your p-value and make a conclusion. (4 pts)

```
## Analysis of Variance Table
##
## Model 1: strength ~ thick
## Model 2: strength ~ as.factor(thick)
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1     18 301.90
## 2     15 148.57  3    153.33 5.16 0.01195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

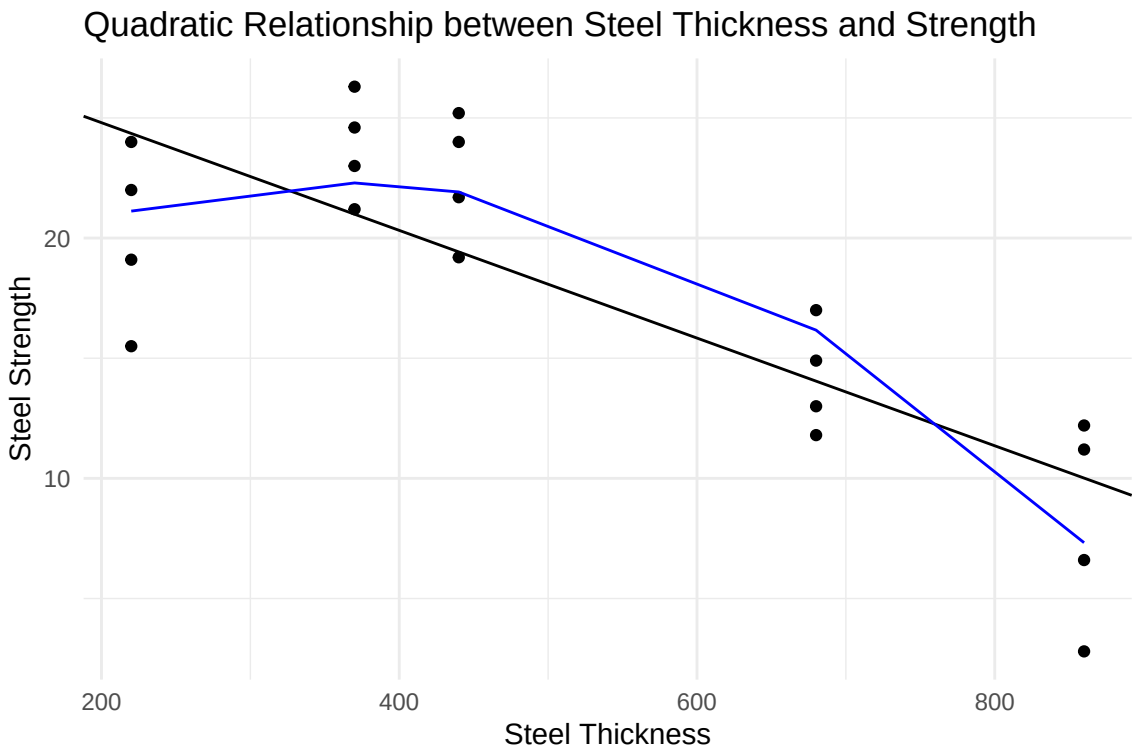
We reject the null hypothesis that the linear regression model is appropriate, $p = 0.0119463 < \alpha = 0.05$.

Part 2C

Now perform a quadratic regression and create a scatterplot with the fitted curve overlaid. Include the "summary" table and plot in your assignment. This can be done with code like the following. Note that b0,b1,b2 need to be replaced with estimates from the quadratic regression. (4 pts)

```
##
```

```
## Call:
## lm(formula = strength ~ thick + thick2, data = steel_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6222 -2.1960  0.2443  2.4491  4.8763
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.452e+01  4.752e+00   3.057  0.00713 **
## thick        4.318e-02  1.980e-02   2.181  0.04354 *
## thick2      -5.994e-05  1.786e-05  -3.357  0.00374 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.268 on 17 degrees of freedom
## Multiple R-squared:  0.7796, Adjusted R-squared:  0.7537
## F-statistic: 30.07 on 2 and 17 DF,  p-value: 2.609e-06
```

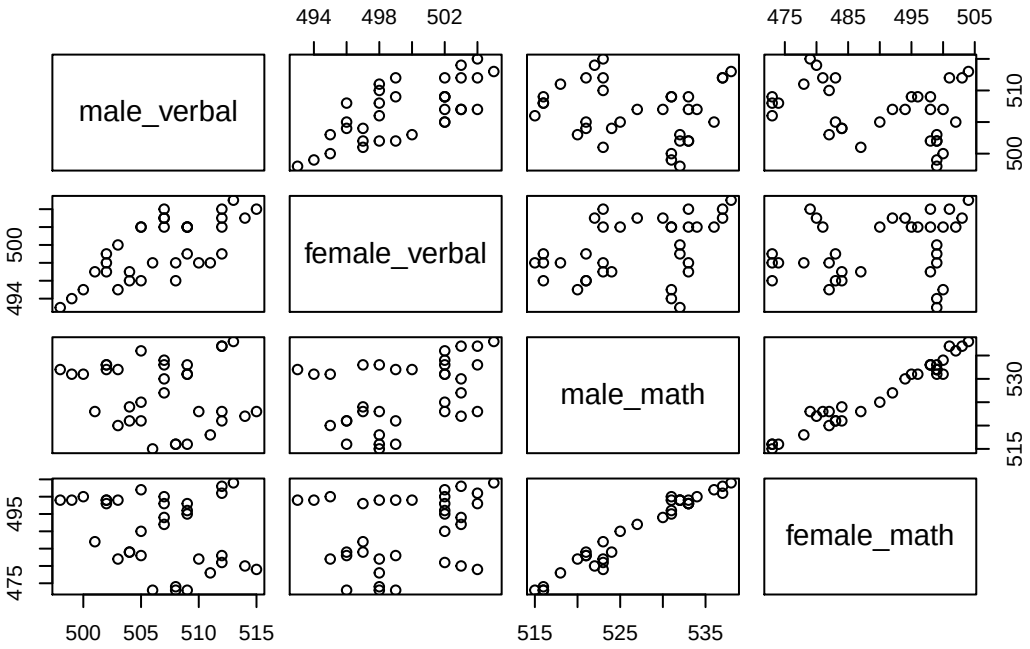


Question 3

Review problem 11.50 from Ott & Longnecker regarding SAT Scores.

Part 3A

Create pairwise scatterplots for all 4 variables (Male.Verbal, Female.Verbal, Male.Math, Female.Math)



Part 3B

Calculate pairwise (Pearson) correlations for all 4 variables. Which pair of variables has the strongest correlation? (4 pts)

rowname	male_verbal	female_verbal	male_math	female_math
male_verbal	NA	0.7081389	-0.1329501	-0.2884984
female_verbal	0.7081389	NA	0.3915856	0.2637590
male_math	-0.1329501	0.3915856	NA	0.9773392
female_math	-0.2884984	0.2637590	0.9773392	NA

`male_math`/`female_math` pair has the highest correlation (0.98).

Part 3C

Provide a test of the correlation for Female.Verbal vs Female.Math. Give the p-value and conclusion in your assignment.

The above correlation test between `female_verbal` and `female_math` yielded a p-value of 0.1317401, which is higher than $\alpha = 0.05$. There is no evidence to suggest a significant relationship between female verbal and female math scores.