Hw11

Amy Fox

12/4/2019

Read in packages

```
library(readr)
library(tidyverse)
```

Question 1

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

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

```
treadmill <- read.csv("../Data/OTT_Final/ASCII-comma/CH11/ex11-22.txt") %>%
  rename(`Treadmill` = 'X.Treadmill.',
         `10K` = `X.10.K.`)

treadmill_fit <- lm(`10K` ~ `Treadmill`, data = treadmill)
summary(treadmill_fit)
```

```
##
## Call:
## lm(formula = `10K` ~ Treadmill, data = treadmill)
##
## 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
```

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

```
plot(`10K` ~ `Treadmill`, data = treadmill, main = "10K vs. Treadmill with regression line")
abline(lm(`10K` ~ `Treadmill`, data = treadmill))
```

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

```
confint(treadmill_fit, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 53.373295 66.468942
## Treadmill   -2.624957 -1.295313
```

Slope estimate: -1.96

95% CI for slope: (-2.62, -1.30)

Slope interpretation: As the time to exhaustion (Treadmill) increases by 1, the race time (10K) decreases by 1.96.

D. Give the R2 value and interpretation in terms of this scenario.

R2: 0.6807

In this case, 68% of the variability in the race time (10K) is explained by the linear regression on the time to exhaustion (Treamill)

E. Give the predicted 10.K time for a runner with Treadmill = 11. Also provide a corresponding prediction interval.

```
predict(treadmill_fit, data.frame(Treadmill = 11), interval = "predict")
```

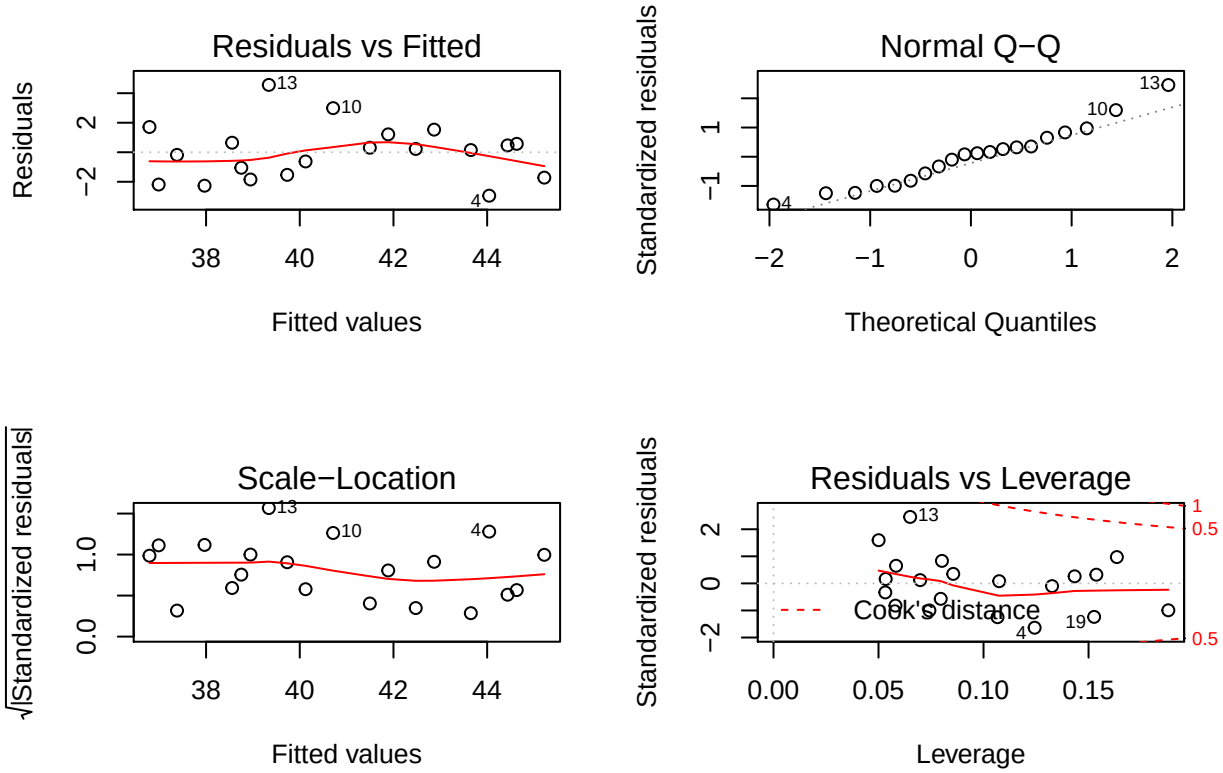
```
##        fit      lwr      upr
## 1 38.35963 34.14223 42.57704
```

The predicted 10K time is 38.36.

Prediction interval: (34.14, 42.58)

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

```
par(mfrow = c(2,2))
plot(treadmill_fit)
```



G. 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.

```
library(car)
outlierTest(treadmill_fit)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 13 2.925728          0.0094335      0.18867
```

Bonferroni p-value: 0.188

Because the p-value is $> 0.0.5$, we fail to reject H0 and cannot conclude that subject 13 is an outlier.

Question 2

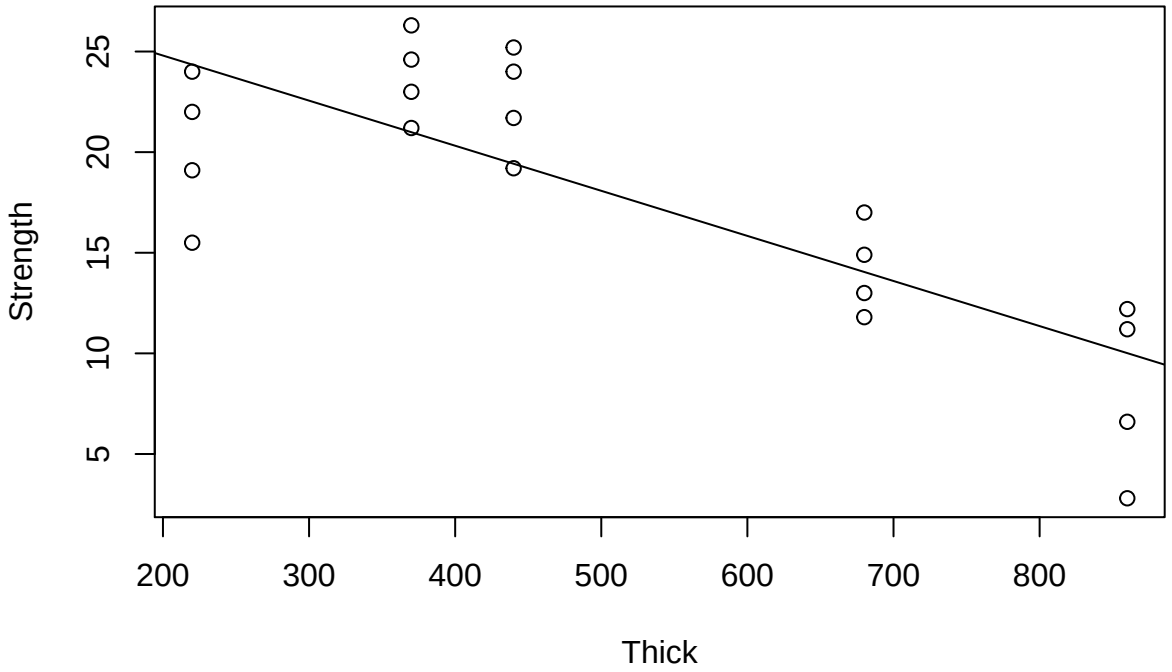
Data on age in coating Thickness (X) and Strength (Y) from an experiment involving steel are available from Canvas as Steel.csv.

```
steel <- read_csv("../Data/Steel.csv")
```

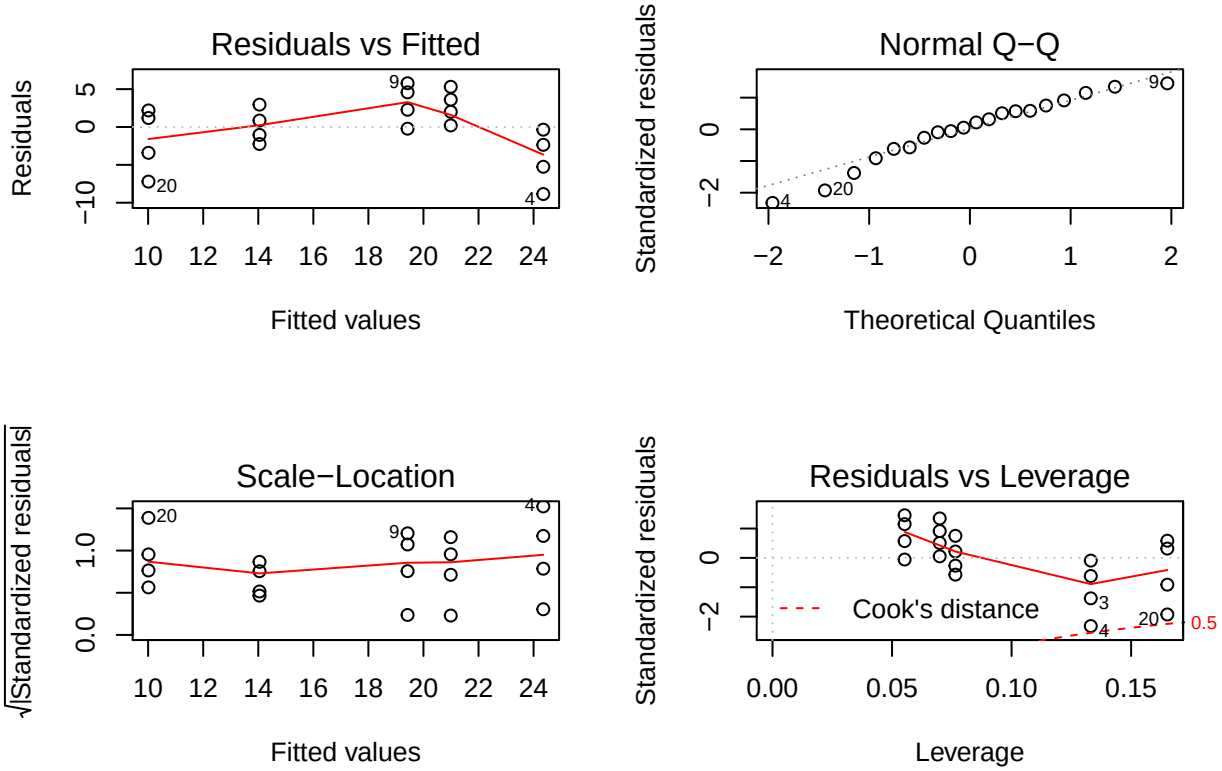
A. Regress Strength (Y) against Thick (X) and look at (1) the plot of Strength versus Thick (2) residuals versus predicted values and (3) qqplot of residuals. Include these plots in your assignment. Do the regression assumptions appear to be met? Discuss. (4 pts)

```r
steel_fit <- lm(`Strength` ~ `Thick`, data = steel)

plot(`Strength` ~ `Thick`, data = steel)
abline(steel_fit)
```




```r
par(mfrow = c(2,2))
plot(steel_fit)
```



While the QQ-plot of the residuals looks mostly okay (aka linear), the Residuals v. Fitted does not appear to show equal scatter (equal variance). The regressed plot is supposed to show a linear trend- it looks kind of okay, but not great.

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

```
anova_steel_fit <- lm(`Strength` ~ as.factor(`Thick`), data = steel)

anova(steel_fit, anova_steel_fit)
```

```
## 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
```

p-value: 0.01195

Conclusion: Because the p-value < 0.05, we reject H0 and assume there there is evidence for lack of fit for the steel regression.

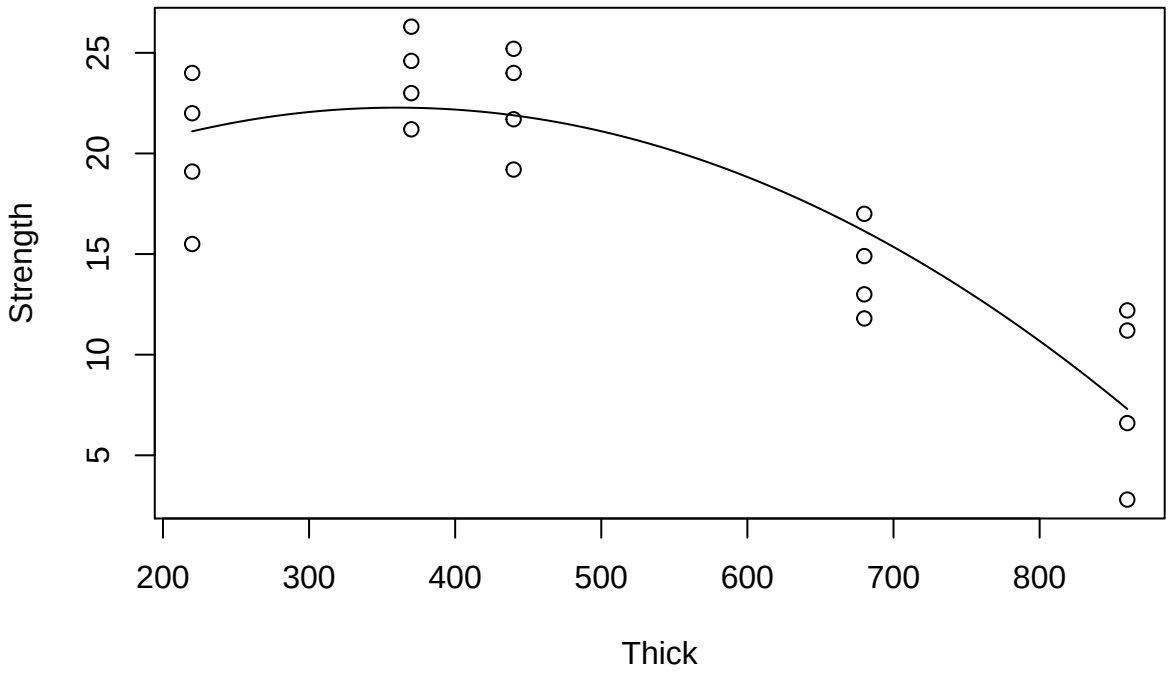
C. Now perform a quadratic regression and create a scatterplot with the fitted curve overlaid. Include the "summary" table and plot in your assignment. This can be done with code like the following. (4 pts)

Note that b0,b1,b2 need to be replaced with estimates from the quadratic regression.

```
QFit <- lm(Strength ~ Thick + I(Thick^2), data = steel)
summary(QFit)
```

```
##
## Call:
## lm(formula = Strength ~ Thick + I(Thick^2), data = steel)
##
## 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 *
## I(Thick^2)  -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
```

```
plot(Strength ~ Thick, data = steel)
curve(14.5 + 0.04318*x + -5.994e-5*x^2, add = TRUE)
```



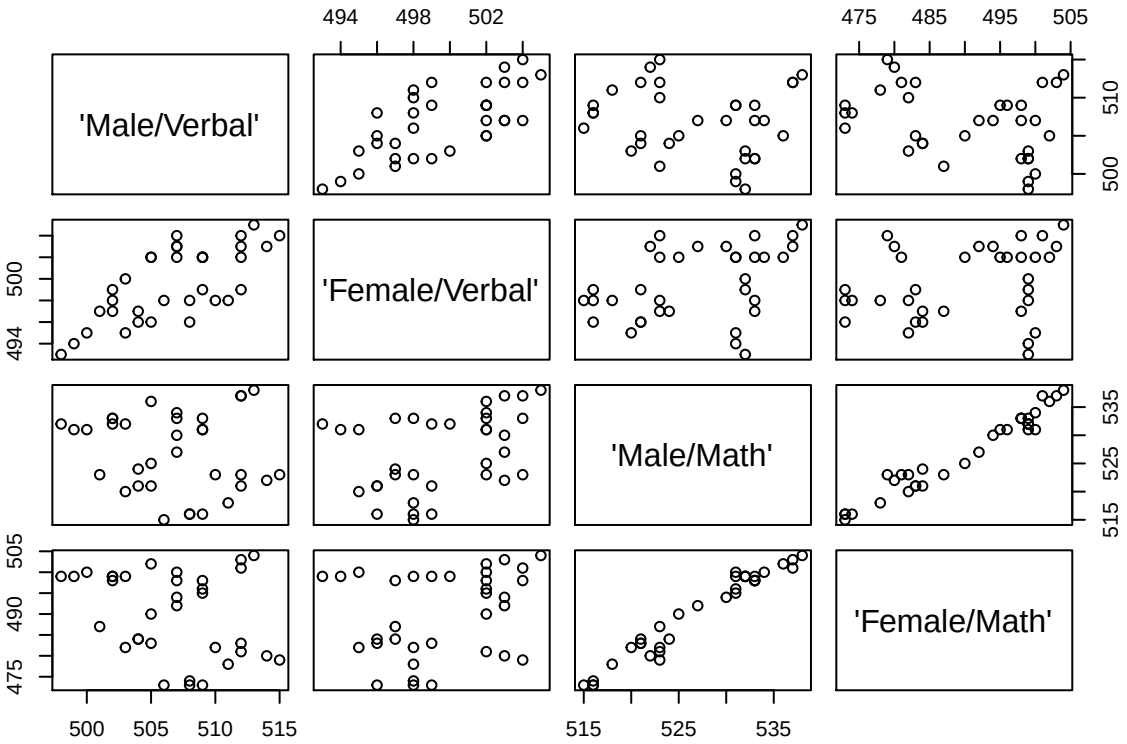
Question 3

Review problem 11.50 from Ott & Longnecker regarding SAT Scores.

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

```
SAT_scores <- read_csv("../Data/OTT_Final/ASCII-comma/CH11/ex11-50.txt") %>%
  column_to_rownames("'Gender/Type'")

pairs(SAT_scores)
```



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

```
cor(SAT_scores)
```

```
##                 'Male/Verbal' 'Female/Verbal' 'Male/Math' 'Female/Math'
## 'Male/Verbal'       1.0000000       0.7081389  -0.1329501    -0.2884984
## 'Female/Verbal'     0.7081389       1.0000000   0.3915856     0.2637590
## 'Male/Math'        -0.1329501       0.3915856   1.0000000     0.9773392
## 'Female/Math'      -0.2884984       0.2637590   0.9773392     1.0000000
```

The strongest correlation is Female/Math with Male/Math.

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

```
cor.test(SAT_scores$"'Female/Verbal'", SAT_scores$"'Female/Math'")
```

```
##
##  Pearson's product-moment correlation
##
## data:  SAT_scores$"'Female/Verbal'" and SAT_scores$"'Female/Math'"
## t = 1.5468, df = 32, p-value = 0.1317
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08169324  0.55263303
## sample estimates:
##      cor
## 0.263759
```

p-value: 0.1317

Conclusion: Because the p-value > 0.194, we fail to reject H0 and assume that there is no correlation between the two groups (aka the true correlation may be 0)

Note: if the data is non-normal, we can add in method = "spearman"

```
cor.test(SAT_scores$"'Female/Verbal'", SAT_scores$"'Female/Math'", method = "spearman")
```

```
## Warning in cor.test.default(SAT_scores$"'Female/Verbal'",
## SAT_scores$"'Female/Math'", : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  SAT_scores$"'Female/Verbal'" and SAT_scores$"'Female/Math'"
## S = 5050.6, p-value = 0.194
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2283323
```

In this case, the p-value is 0.194 and has the same conclusion as the pearson correlation above.