Assign 11

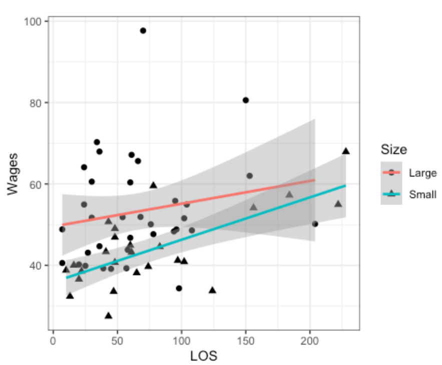
Natalie Schmer

1. Bank Salaries

```
library(ggplot2)
library(knitr)
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────────
tidyverse 1.3.0 ──

## ✓ tibble  3.0.0     ✓ dplyr   0.8.5
## ✓ tidyr   1.0.2     ✓ stringr 1.4.0
## ✓ readr   1.3.1     ✓ forcats 0.5.0
## ✓ purrr   0.3.3

## ── Conflicts ──────────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

saldata <-
read.csv("/Users/natalieschmer/Desktop/GitHub/stats_511/data/BankSalaries.csv
")
Large <- subset(saldata, Size=="Large")
Small <- subset(saldata, Size=="Small")
```

1A. Scatter Plot

```
ggplot(data = saldata, aes(x = LOS, y= Wages, shape = Size))+
  geom_point(size = 2)+
  geom_smooth(mapping = aes(LOS, Wages, color = Size), method = "lm", se =
T)+
  theme_bw()

## `geom_smooth()` using formula 'y ~ x'
```



1B Regression Models and Equation for line

```
lm_large <- lm(Wages ~ LOS, data = Large)
#y = 49.45 + 0.05595(LOS)

lm_small <- lm(Wages ~ LOS, data = Small)
#y = 35.87192 + 0.10416(LOS)
```

1C. Interpretation

For large banks, wages start higher but for one unit increase in LOS increases by 0.056, whereas for small banks wages start lower but for one unit increase in LOS increase by 0.10.

1D.Test for slope

```
summary(lm_large)

##
## Call:
## lm(formula = Wages ~ LOS, data = Large)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.688  -8.472  -3.691   5.767  44.218
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.54532    4.01305  12.346 6.46e-14 ***
## LOS          0.05595    0.05116   1.094    0.282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.02 on 33 degrees of freedom
## Multiple R-squared:  0.03498,    Adjusted R-squared:  0.005741
## F-statistic: 1.196 on 1 and 33 DF,  p-value: 0.282

summary(lm_small)

##
## Call:
## lm(formula = Wages ~ LOS, data = Small)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0716  -4.4861   0.3944   2.8101  15.5273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.87192    2.28194  15.720 8.53e-14 ***
## LOS          0.10416    0.02326   4.478 0.000171 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.021 on 23 degrees of freedom
## Multiple R-squared:  0.4657, Adjusted R-squared:  0.4425
## F-statistic: 20.05 on 1 and 23 DF,  p-value: 0.0001712
```

For large banks, p = 0.282 and for small banks, p = 0.0001712. For, large banks, there is not evidence that LOS is linearly related to wages, but there is evidence this is true for small banks.

1E CI for Intercept

```
confint(lm_large, level = 0.95)

##                   2.5 %     97.5 %
## (Intercept) 41.38071287 57.7099183
## LOS         -0.04812646  0.1600341

confint(lm_small, level = 0.95)

##                   2.5 %     97.5 %
## (Intercept) 31.15135828 40.5924796
## LOS          0.05603753  0.1522847
```

Ultimately, an employee would be better off at a large bank because the representing the starting wage, starts higher than a smaller bank.

1F. CI for Mean LOS=96

```
newdata <- data.frame(LOS = 96)
predict(lm_large, newdata, interval = "predict", level = 0.95)

##        fit      lwr      upr
## 1 54.91688 27.86905 81.96471

predict(lm_small, newdata, interval = "predict", level = 0.95)

##        fit      lwr     upr
## 1 45.87139 31.03197 60.7108

predict(lm_large, newdata, interval = "confidence", level = 0.95)

##        fit      lwr      upr
## 1 54.91688 49.43465 60.39911

predict(lm_small, newdata, interval = "confidence", level = 0.95)

##        fit      lwr      upr
## 1 45.87139 42.83057 48.91221
```

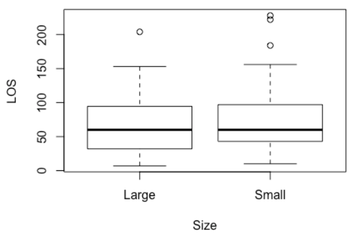
After 8 years, an emplyee would still be better off at a larger bank, as the range of possible wages as represented by the confidence interval is higher than for smaller banks.

1G. Outlier

```
car::outlierTest(lm_large)

##    rstudent unadjusted p-value Bonferroni p
## 15 4.242492         0.00017638    0.0061735

#row 15 of the large dataset, LOS is 70 and wage is 97.68
```

1H. T-test

```
boxplot(LOS~Size, data =saldata)
```



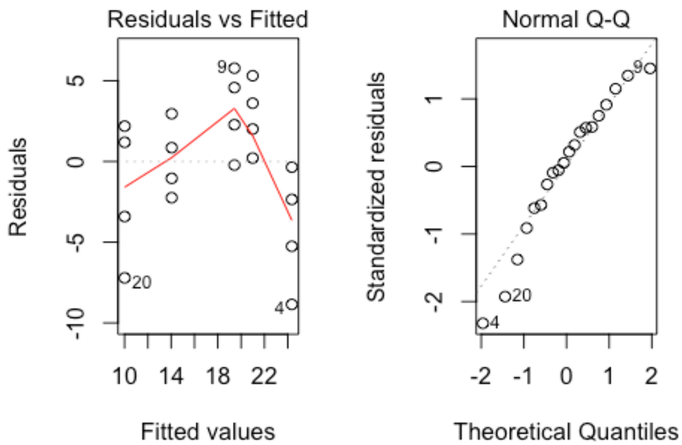

```
t.test(LOS~Size, data =saldata)

##
##  Welch Two Sample t-test
##
## data:  LOS by Size
## t = -0.81609, df = 40.606, p-value = 0.4192
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -40.73149  17.29149
## sample estimates:
## mean in group Large mean in group Small
##               65.60               77.32
```

Since the boxplots have a fair amount of overlap and p > 0.05, we cannot conclude that LOS is significantly different between bank sizes.

Q2 Steel Quadratic

```
Steel<-
read.csv("/Users/natalieschmer/Desktop/GitHub/stats_511/data/Steel.csv")
str(Steel)

## 'data.frame':    20 obs. of  2 variables:
##  $ Thick   : int  220 220 220 220 370 370 370 370 440 440 ...
##  $ Strength: num  24 22 19.1 15.5 26.3 24.6 23 21.2 25.2 24 ...
```

2A. (4pts)



Based on the plots, the regression assumptions do not seem to be well met. the Residuals vs fitted show deviations away from the 0 line, and the QQ plot is not very linear.

2B. (4pts)

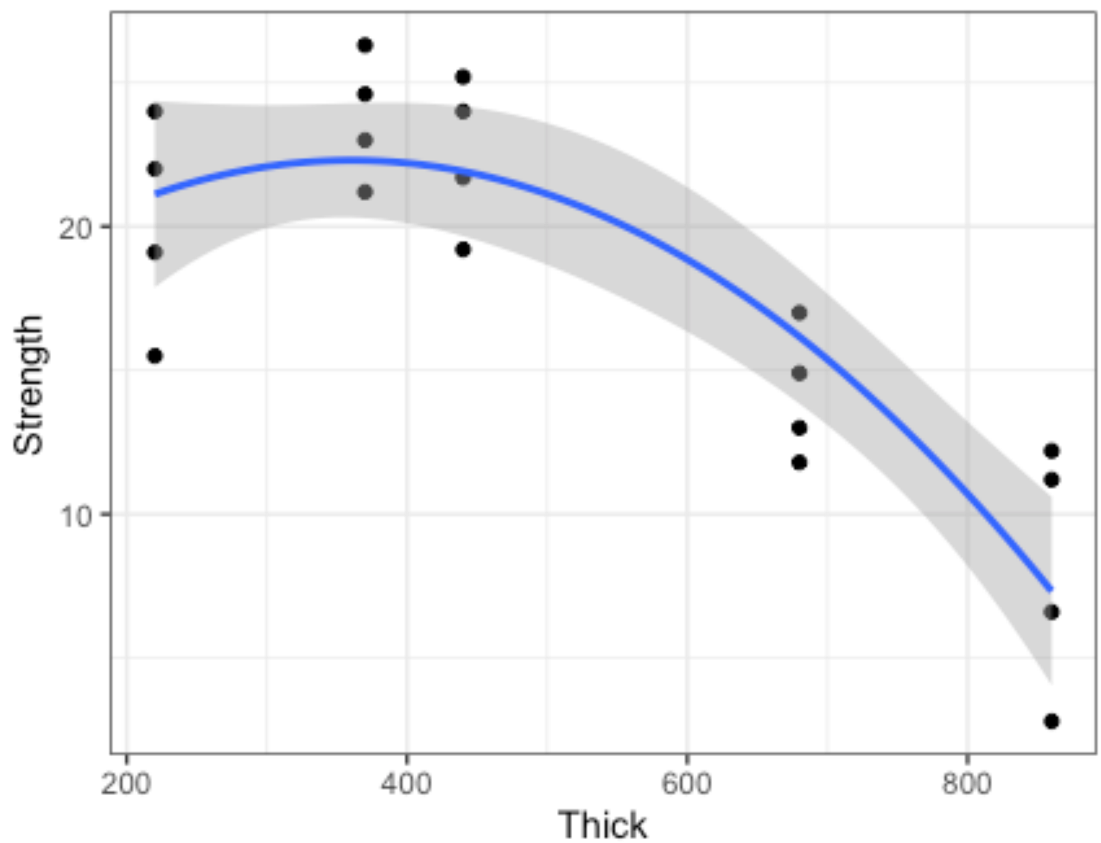
```
regfit2a <- anova(lm_2a)

ANOVA2a <- lm(Strength ~ as.factor(Thick), Steel)
ANOVAfit2a <- anova(ANOVA2a)

#lack of fit
anova(lm_2a, ANOVA2a)

## 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

#F
regfit2a$'Sum Sq'

## [1] 522.0448 301.9007

SSreg = regfit2a$'Sum Sq'[2]
SSanova = ANOVAfit2a$`Sum Sq`[2]
DFreg = regfit2a$Df[2]
DFanova = ANOVAfit2a$Df[2]
F = ((SSreg-SSanova)/(DFreg-DFanova))/ANOVAfit2a$`Mean Sq`[2]
```

```r
1-pf(F,(DFreg-DFanova),DFanova) # p-value

## [1] 0.01194633
```

Since p< 0.05, there is evidence of a lack of fit.

2C. (4 pts)

```r
ggplot(data = Steel, aes(x = Thick, y = Strength))+
  geom_point()+
  geom_smooth(method = "lm", formula = y ~ x + I(x^2))+
  theme_bw()
```



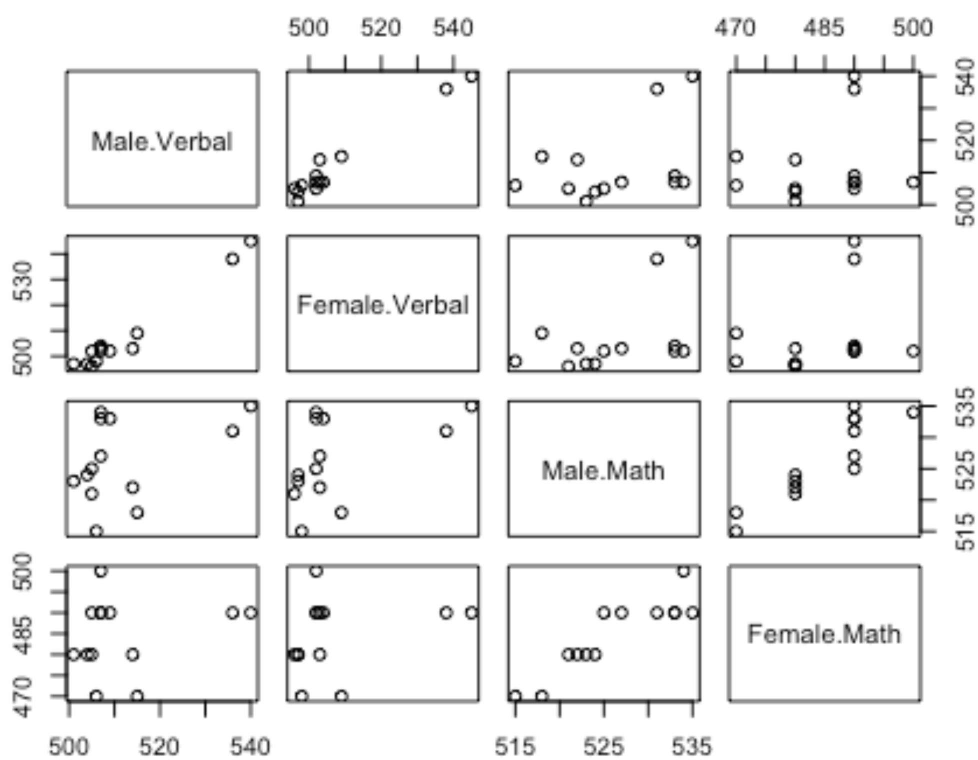

```r
summary(lm(Steel$Strength ~ Steel$Thick + I(Steel$Thick^2)))

##
## Call:
## lm(formula = Steel$Strength ~ Steel$Thick + I(Steel$Thick^2))
##
## 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 **
## Steel$Thick          4.318e-02  1.980e-02   2.181  0.04354 *
## I(Steel$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
```

Q3 SAT Scores

3A.

```
sat_cor <- sat %>%
              select(-year)

pairs(sat_cor)
```



{height=200 px}

3B.

```
cor(sat_cor, method = "pearson")
```

```
##                Male.Verbal Female.Verbal Male.Math Female.Math
## Male.Verbal     1.0000000     0.9814674 0.4167834   0.1952538
## Female.Verbal   0.9814674     1.0000000 0.4960357   0.2841566
## Male.Math       0.4167834     0.4960357 1.0000000   0.8995118
## Female.Math     0.1952538     0.2841566 0.8995118   1.0000000
```

The strongest correlation is between female verbal and male verbal

3C

```
cor.test(sat_cor$Female.Verbal, sat_cor$Female.Math, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  sat_cor$Female.Verbal and sat_cor$Female.Math
## t = 0.98296, df = 11, p-value = 0.3468
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3163599  0.7220875
## sample estimates:
##       cor
## 0.2841566
```

Since $p > 0.05$, and the correlation coefficient is very low, there is little to no correlation bewteen female verbal and female math scores.