BTRY 6020 Homework III

NAME: student name

NETID: student NetID

DUE DATE: 8:40 am Monday February 20

Question 1

Question 1. Patient satisfaction with their hospital stay is rapidly becoming more important to hospital administrators. In an effort to evaluate factors which influence patient satisfaction at a particular hospital, a survey of 46 randomly selected patients was conducted, and the following vaiables measured: Patient satisfaction (PatSat, an index), patient age (Age, in years), the severity of the patient's condition (Sev, an index) and the patient's level of anxiety (Anx, an index).

a) Regress patient satisfaction against the three predictors. What is the resulting regression equation?

```
library(readxl)
Hwk3Q1DatSp14 <- read_excel("Hwk3Q1DatSp17.xlsx")
lmPatSat <- lm(PatSat ~ Age + Sev + Anx, data=Hwk3Q1DatSp14)
lmPatSat
```

```
##
## Call:
## lm(formula = PatSat ~ Age + Sev + Anx, data = Hwk3Q1DatSp14)
##
## Coefficients:
## (Intercept)          Age          Sev          Anx
##     158.491       -1.142       -0.442      -13.470
```

The estimated regression equation is

$$\hat{PatSat} = 158.491 - 1.142Age - 0.442Sev - 13.470Anx.$$

b) Get the correlation coefficients for each pair of the three predictor variables. Does there appear that multicollinearity will be an issue? Explain briefly.

```
cor(Hwk3Q1DatSp14[2:4], method = "pearson")
```

```
##           Age       Sev       Anx
## Age 1.0000000 0.5679505 0.5696775
## Sev 0.5679505 1.0000000 0.6705287
## Anx 0.5696775 0.6705287 1.0000000
```

Yes. From the output we see that *Age* has positive correlation with *Sev* and *Anx*, and *Sev* has positive correlation with *Anx*. So multicollinearity could well be an issue.

c) Get the VIFs of the three predictors. Describe the degree of multicollinearity between these three predictors. Explain how this relates back to your answer to part b above.

```
library(car)
vif(lmPatSat)
```

```
##      Age      Sev      Anx 
## 1.632296 2.003235 2.009062
```

Given $VIF_{Age} = 1.63$, we know that the variance of Age coefficient is inflated by a factor of 1.63 because Age is correlated with at least one of the other two predictors. The explanations for the other two variables' VIF are similar. Since none of the variables have their VIFs exceed 4, the degree of multicollinearity among the three predictors is not a big issue.

d) Are there any predictors which appear non-significant in the presence of the other two? Explain briefly.

```
summary(lmPatSat)
```

```
## 
## Call:
## lm(formula = PatSat ~ Age + Sev + Anx, data = Hwk3Q1DatSp14)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -18.3524  -6.4230   0.5196   8.3715  17.1601 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## Age          -1.1416     0.2148  -5.315 3.81e-06 ***
## Sev          -0.4420     0.4920  -0.898   0.3741    
## Anx         -13.4702     7.0997  -1.897   0.0647 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595 
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

From the summary table, we notice both Age and Anx are significant at $\alpha = .10$ significance level. Only Sev is not significant with p-value being 0.3741.

e) Remove the least significant predictor from the regression and rerun the regression using the remaining two predictors. Has the significance of the remaining predictors changed since removing the least significant variable? Why has this happened (explain briefly)?

```
lmPatSat2 <- lm(PatSat ~ Age + Anx, data=Hwk3Q1DatSp14)
summary(lmPatSat2)
```

```
## 
## Call:
## lm(formula = PatSat ~ Age + Anx, data = Hwk3Q1DatSp14)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -19.4453  -7.3285   0.6733   8.5126  18.0534 
## 
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.9412    11.5251  12.663 4.21e-16 ***
## Age          -1.2005     0.2041  -5.882 5.43e-07 ***
## Anx         -16.7421     6.0808  -2.753  0.00861 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.04 on 43 degrees of freedom
## Multiple R-squared:  0.6761, Adjusted R-squared:  0.661
## F-statistic: 44.88 on 2 and 43 DF,  p-value: 2.98e-11
```

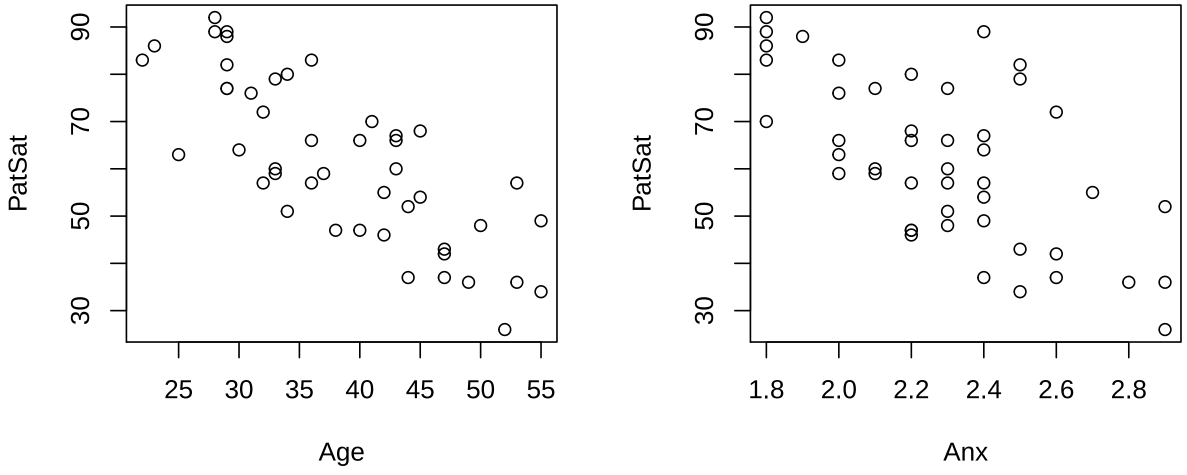
The p-values of the remaining two predictors (Age and Anx) decrease after Sev is removed. So the remaining predictors become more significant in explaining the variation in $PatSat$. This increase of model performance is due to the removal of a correlated predictor from the model.

f) Can you now interpret the point estimate of the coefficient of age in the presence of the other predictor? Briefly defend your answer-and then interpret, if you can.

For fixed level of anxiety, the patient satisfication index will decrease by 1.2005 for one year's increase in patient's age. The fact that both predictors in the model are significant and $R^2 = 0.661$ indicates our model performs moderately well in explaining the variation in patient satisfication.

g) Get the required diagnostic plots and use them to check to see if the assumptions for inference have been met. Be sure to examine the data for outliers and influential points.

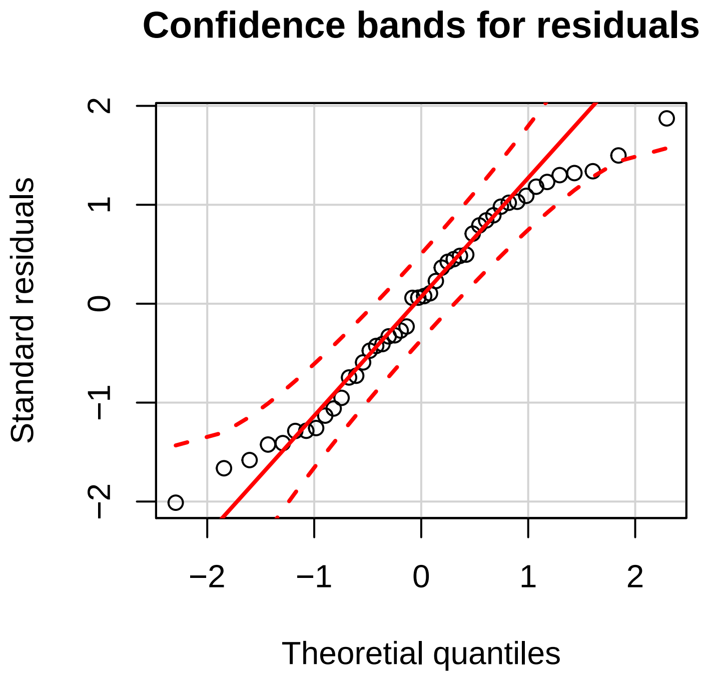
```r
old.par <- par(mfrow=c(1, 2))
plot(Hwk3Q1DatSp14$PatSat ~ Hwk3Q1DatSp14$Age, xlab="Age", ylab="PatSat")
plot(Hwk3Q1DatSp14$PatSat ~ Hwk3Q1DatSp14$Anx, xlab="Anx", ylab="PatSat")
```



```r
par(old.par)
```

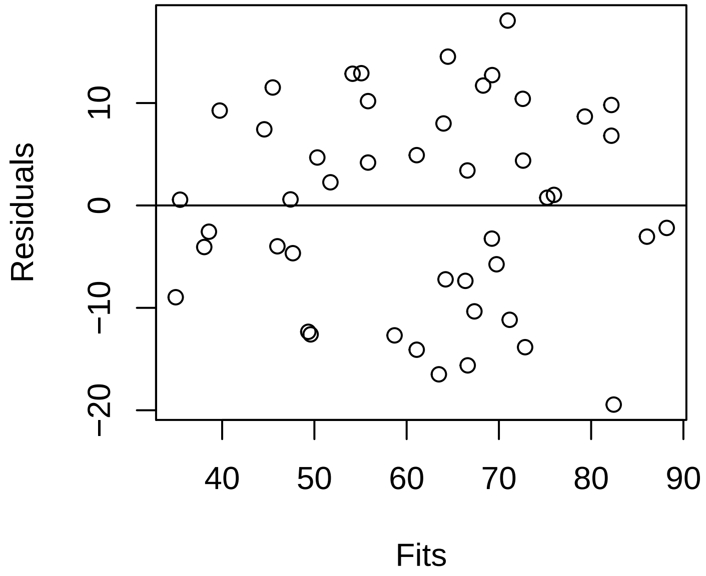
From the above scatterplots for $PatSat$ with each of the two predictors, we notice approximate linear relationship between them. So the linearity assumption is satisfied.

```r
qqPlot(rstandard(lmPatSat2), xlab="Theoretial quantiles", ylab="Standard residuals",
       main="Confidence bands for residuals")
```



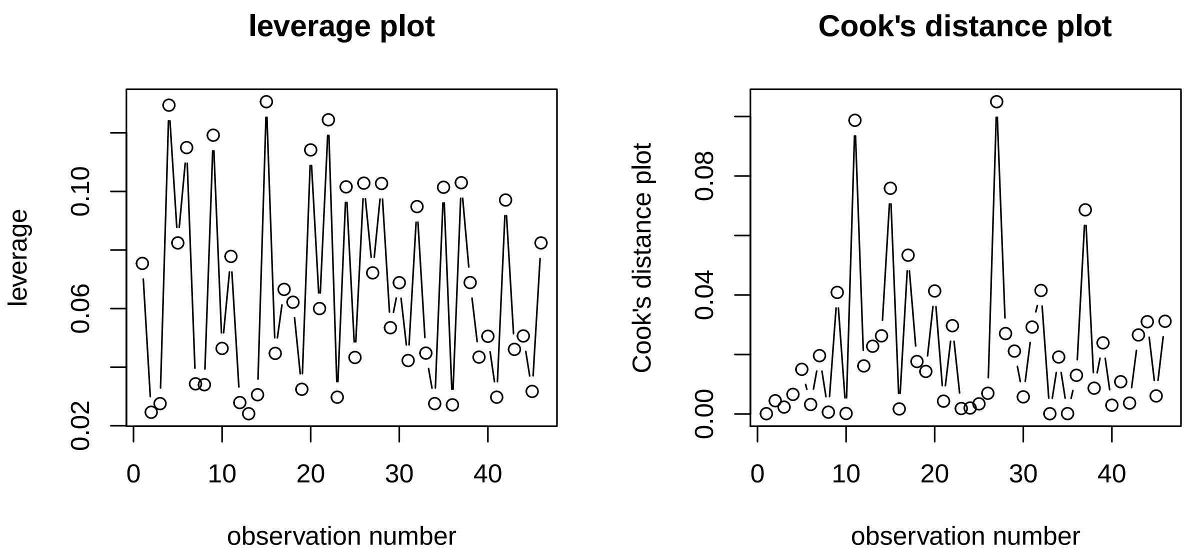
From the above qqplot for the standardized residuals, we verify that the normality assumptions for the random error is satisfied because all the points are within the 95% confidence bands in red dashed lines.

```r
plot(lmPatSat2$resid ~ lmPatSat2$fit, xlab="Fits", ylab="Residuals")
abline(h=0)
```



From the above residual plot, we verify that constant variance assumption is satisfied because all the points distribute roughly evenly above and below the zero line. We proceed to detect potential outliers by plotting the leverage plot and the cook's distance.

```
old.par <- par(mfrow=c(1, 2))
# leverage
lev = hat(model.matrix(lmPatSat2))
plot(lev, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
# Cook's distance
cook <- cooks.distance(lmPatSat2)
plot(cook, type = "b", xlab="observation number", ylab="Cook's distance plot",
     main="Cook's distance plot")
```



```
par(old.par)
```

The benchmark for the leverage value is $\frac{2p}{n} = \frac{2*3}{46} = 0.13$. From the above leverage plot, there are a few points above the boundary for a little bit. The benchmark for the Cook's distance is 1. From the above Cook's distance plot, all the points are within the boundary. Since all the deviations from the benchmarks are not severe, we would not worry about any points being influential outliers.

h) Assuming that the assumptions for inference have been met, test to see if the patient satisfaction index decrease by more than a half unit for each additional year of age, after controlling for the patient's anxiety level. State hypotheses, test statistic, p-value, and conclusions.

Our model for the regression is $PatSat = \beta_0 + \beta_1 Age + \beta_2 Anx + \epsilon$. Our hypotheses are

$$H_0 : \beta_1 \geq -0.5 \quad \text{and} \quad H_a : \beta_1 < -0.5.$$

From the output in part e), the estimated coefficient is $\hat{\beta}_1 = -1.2005$ and its standard error is $SE(\hat{\beta}_1) = 0.2041$.

```
# Compute test statistic
tstat = (-1.2005-(-0.5))/0.2041
# Compute p-value
1 - pt(abs(tstat), df = 43)
```

```
## [1] 0.0006676247
```

Since the pvalue 0.00067 is less than 0.05, we reject H_0 in favor of H_a. Therefore, we are conclude that the patient satisfaction index decreases by more than a half unit for each additional year of age, after controlling for the patient's anxiety level.

i) A 22 year old patient arrives with an anxiety index of 2.6. What will this patient's satisfaction level be? (Answer with an appropriate inferential procedure, not just a point estimate.)

We can construct the 95% prediction interval for this patient's satisfication level based on our model.

```
newdata <- data.frame(Age = 22, Anx = 2.6)
predict(lmPatSat2, newdata, interval="prediction")
```

```
##        fit      lwr     upr
## 1 76.00152 53.45684 98.5462
```

The 95% prediction interval for this patient's satisfication level is (53.46, 98.55), indicating the probability this patient's satisfaction level will be between 53.46 and 98.55 is .95..

Question 2.

A young businessman is considering getting an MBA. He evaluates the cost and decides that if he can expect his starting salary after graduation to be greater than $70,000, it will be worth his while. He finds a ratings report in US News on what are regarded as the 50 top business schools and the average starting salaries of their graduates (data appear in Hwk3Q2DatSp17). US News rates multiple criteria and combines them into one score. Although the school at which he was accepted does not appear in this top 50 list, he finds on the school's website that US News had rated it a 62.

Does this young businessman have enough evidence to support going to this school?

A) Formulation of the research question and choice of the appropriate statistical technique used to answer this question.

The question here is: Can this young scholar expect to have a starting salary greater than $70,000 after graduating from this particular business school. This is a yes-no question, so we'll do a hypothesis test with an alternative hypothesis that mean starting salaries for schools with a US News rating of 62 is greater than 70,000 dollars.

B) Notation for the random variable(s) and parameter(s) of interest; define these explicitly. Give the distributional assumptions for your random variable(s) and state all assumptions necessary for the statistical application you intend to use.

Let y_i be the mean starting salary of graduates from the ith business school, i = $1, 2, \ldots, 50$

our model is

$$y_i \sim ind\ N(\beta_o + \beta_1 x_i + \beta_2 x_i^2, \sigma^2)$$

Where: β_o is a parameter denoting the intercept β_1 and β_2 are coefficients of x_i and x_i^2, respectively.

We want to due inferences on $\mu_{y|x=62}$; specifically, to test to see if this mean exceeds $70,000 dollars.

This model includes the following explicit assumptions: i) Independence of obsevrations; ii) normality of y's at each x_i; iii) means quadratically related to x_i; iv) constant variance; v) (implicitly) outliers are not driving our conclusions.

C) Calculations for the analysis. For hypothesis and significance tests, formulate the null and the alternative hypotheses, calculate the value of your test statistic, and then calculate your p-value. For confidence intervals, show and apply the appropriate formula. Use ?? = .05 if not otherwise specified.

```r
library(readxl)
Hwk3Q2DatSp17 <- read_excel("~/BTRY6020Sp17/Homework/Hwk3/Hwk3Q2DatSp17.xlsx")

USNSq <- Hwk3Q2DatSp17$USNscor^2

Quad.lm <- lm(StartSal ~ USNscor + I(USNscor^2), data = Hwk3Q2DatSp17)
summary(Quad.lm)
```

```
##
## Call:
## lm(formula = StartSal ~ USNscor + I(USNscor^2), data = Hwk3Q2DatSp17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5982  -3.1894  -0.3122   2.7766   9.7158
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -52.152194  29.647623  -1.759 0.085075 .
## USNscor        2.785370   0.776566   3.587 0.000795 ***
## I(USNscor^2)  -0.012331   0.004943  -2.494 0.016188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.465 on 47 degrees of freedom
## Multiple R-squared:  0.8748, Adjusted R-squared:  0.8694
## F-statistic: 164.1 on 2 and 47 DF,  p-value: < 2.2e-16
```

```r
newdata<- data.frame(USNscor = 62)
predict(Quad.lm, newdata, se.fit=TRUE, interval="confidence")
```

```
## $fit
##        fit      lwr      upr
## 1 73.14022 71.24799 75.03245
##
## $se.fit
## [1] 0.9405956
##
## $df
## [1] 47
##
## $residual.scale
## [1] 4.464701
```

$$H_0 : \mu_y|x = 62 \leq 70 \quad \text{and} \quad H_a : \mu_y|x = 62 > 70$$

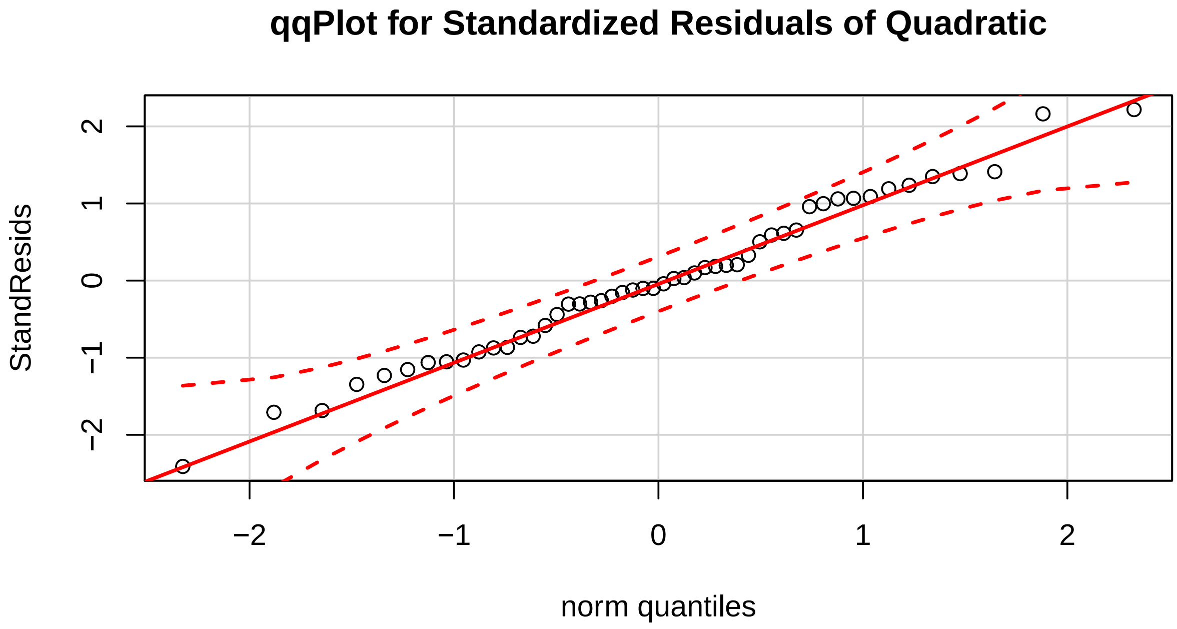
$$TS = (\hat{y}_i \; - \; \mu_o)/SE\,(\hat{y}_i) = (73.14 \; - \; 70)/.941 \; = \; 3.34)$$

```r
p = 1-pt(3.34,47)
```

D) Discuss whether the assumptions stated in Part B above are met sufficiently for the validity of the statistical inferences; use graphs and other tools where applicable.

i) Independence: Though not a random sample, there is no logical reason to think that one rating was related to another, so assuming independence appears reasonable

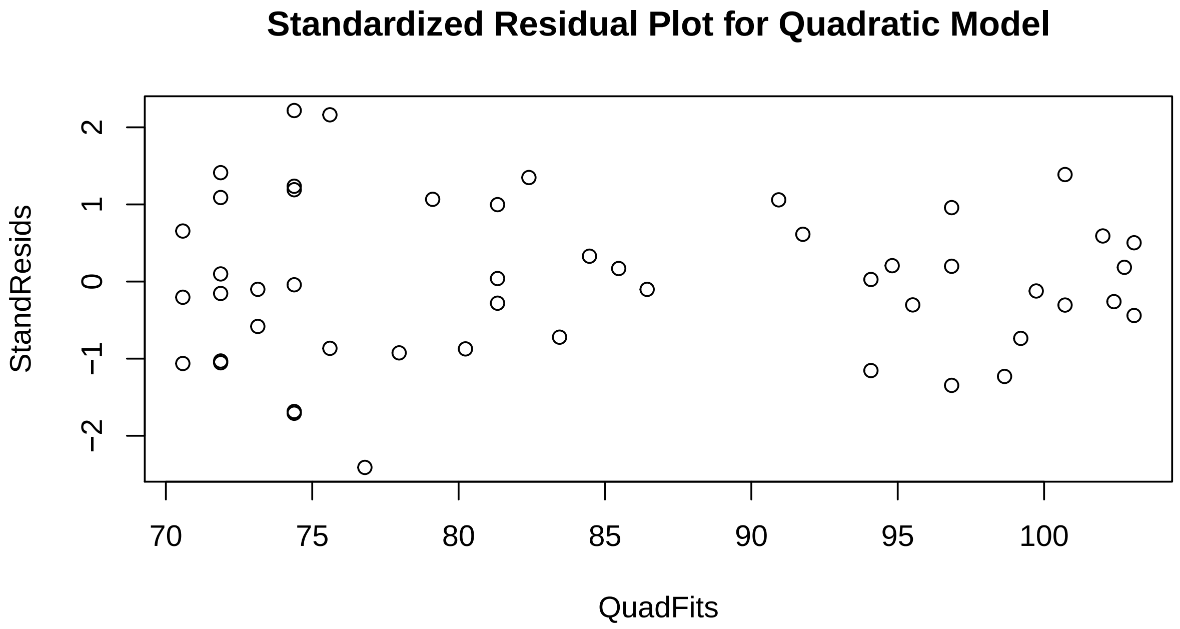
```
library(car)
StandResids <- rstandard(Quad.lm)
qqPlot(StandResids, main="qqPlot for Standardized Residuals of Quadratic")
```



Here we see that the standardized residuals appear normally distributed, so the normality assumption appears valid.

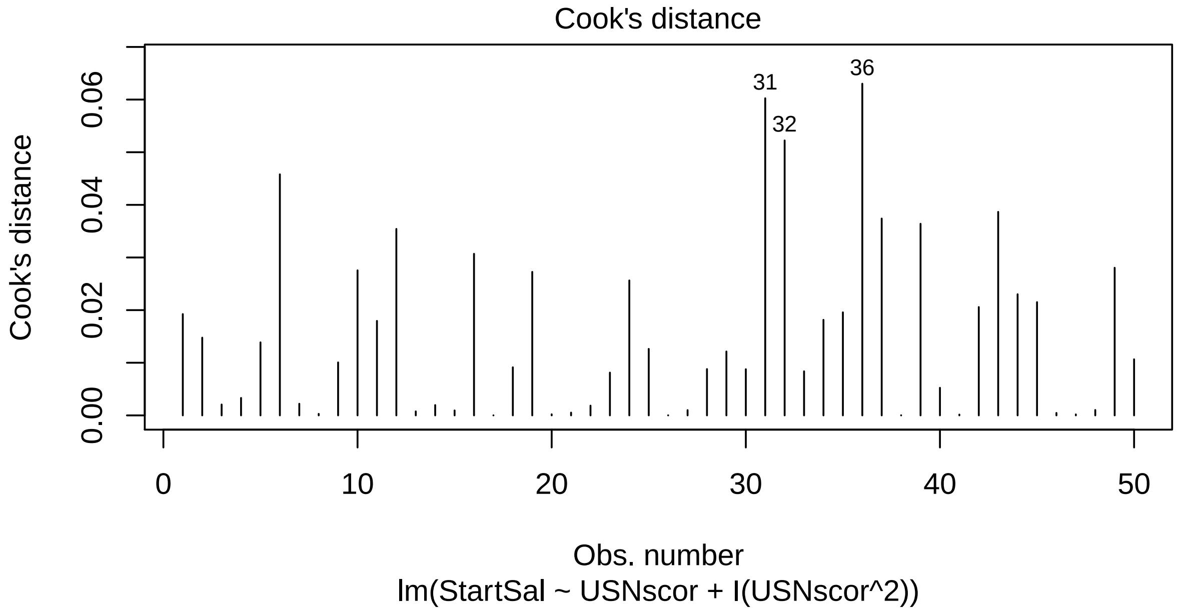
```
library(car)

StandResids <- rstandard(Quad.lm)
QuadFits<-fitted.values(Quad.lm)
plot(QuadFits, StandResids, main="Standardized Residual Plot for Quadratic Model")
```



qqPlot(StandResids, main="qqPlot for Standardized Residuals of Quadratic Model") The standardized residuals appear to have constant variance, so this assumption appears valid. And by the lack of curvilinearity in this residual plot, we see that the means are linearly related to this quadratic regression, so the linearity assumption aslo appears valid.

```
plot(Quad.lm, which = 4,)
```



From the above Cook's Distance plot, we see there are no outliers even close to being influential (all Cook's Distances < .10), so outliers are not driving our conclusions.

E) Discuss the sampling scheme and whether or not it is sufficient to meet the objective of the study. Be sure to include whether or not subjective inference is necessary and if so, defend whether or not you believe it is valid.

The sampling scheme is rather odd: The sample is what is commonly regarded as the 50 best MBA programs in the United States. Since this candidate's school is not a part of this sampled population, subjective inference is necessary, as his school is not a memeber of this population. However, since his schools US News Score does fall in the range of these data, I would consider the subjective inference to be valid, as his school is proably not much different that the ones at the lower end of this ratings scale, making this sampled population representative of his target population of schools in the lower end of the rating scale.

F) State the conclusions of the analysis. These should be practical conclusions from the context of the problem, but should also be backed up with statistical criteria (like a p-value, etc.). Include any considerations such as limitations of the sampling scheme, impact of outliers, etc., that you feel must be considered when you state your conclusions.

With p = .0008, we see that there is considerable evidence to reject H_o and conclude that the starting salary of this young man will exceed $70,000. All assumptions appear to be met, so this young man should go for it!

Question 3.

In a short, brief, one paragraph answer, describe how the overall F-test in a multiple regression is just a special case of the general linear test.

The general linear test (GLT) is used to test whether the coefficients of any subset of variables in a regression are simultaneously zero. In the overall F-test (OFT), this subset is the entire set of predictors. Essentially, when you would do the OFT, you simultaneously drop all predictors from the model. When you do this, you lose the entire sums of squares regression, since you no longer have any predictors left. Then the k from the GLT becomes the number of predictors, which is the same as the df for Regression, and the numerator of the GLT becomes SSR/dfReg, which is just MSReg. And since in the general linear test you divide through by the MSE of the full model, the test statistic of the GLT just boils down to MSR/MSE, the test statistic of the OFT. Therefore, the OFT just boils down to the GLT of all predictors.

Question 4.

Let's revisit the class example of creatinine clearance as a function of a patient's creatinine concentration (Conc), Age, and Weight. Data appears in Hwk3Q4DatSp17.

A) Run the multiple regression of the three predictors on creatinine clearance. What is the resulting regression equation?

```r
library(readxl)
Hwk3Q4DatSp17 <- read_excel("~/BTRY6020Sp17/Homework/Hwk3/Hwk3Q4DatSp17.xlsx")
lmq4 <- lm(CreatClear ~ Conc + Age + Weight, data = Hwk3Q4DatSp17)
summary(lmq4)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight, data = Hwk3Q4DatSp17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.668  -7.002   1.518   9.905  16.006
##
```

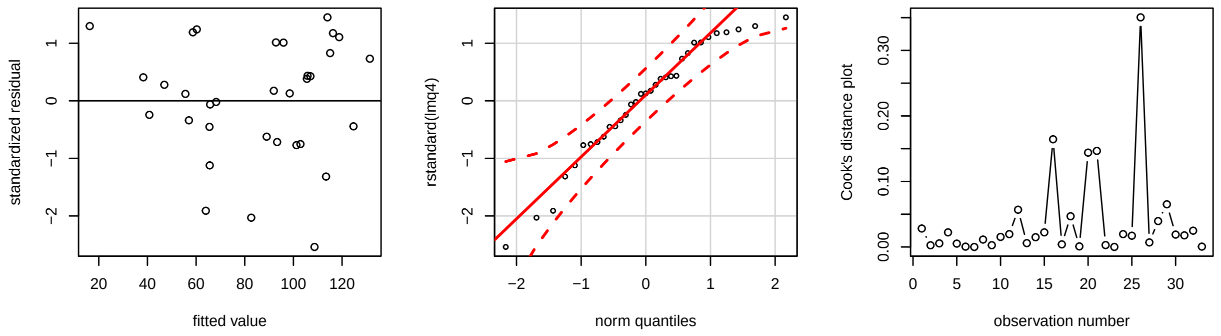
```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
## Conc        -39.9393     5.6000  -7.132 7.55e-08 ***
## Age          -0.7368     0.1414  -5.211 1.41e-05 ***
## Weight        0.7764     0.1719   4.517 9.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.46 on 29 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8398
## F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12
```

The estimated regression equation is

$$\widehat{CreatClear} = 120.0473 - 39.9393 \times Conc - 0.7368 \times Age + 0.7764 \times Weight.$$

B) Get a standardized residual plot (standardized residuals versus fitted values), a qqPlot of the standardized residuals, and a Cook's distance plot. Do you notice any problems with any of the diagnostics?

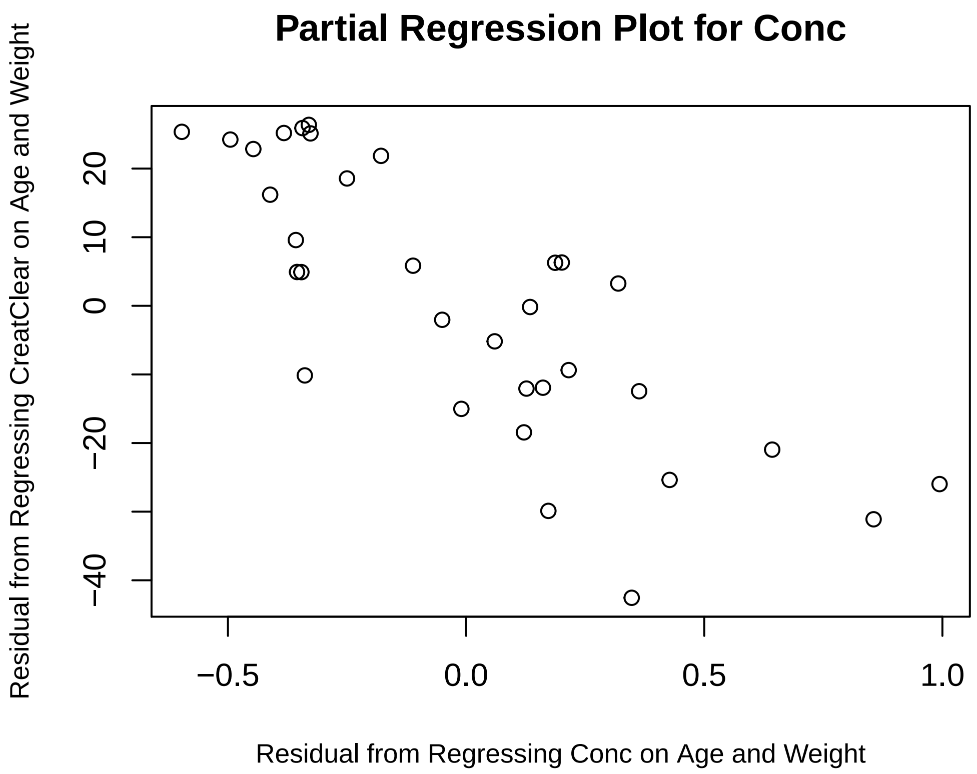
```
par(mfrow=c(1,3))
# Get standardized residual plot
plot(rstandard(lmq4) ~ lmq4$fit, xlab="fitted value", ylab="standardized residual")
abline(h=0)
# Get qqPlot
library(car)
qqPlot(rstandard(lmq4))
# Plot Cook's distance
cook <- cooks.distance(lmq4)
plot(cook, type = "b", xlab="observation number", ylab="Cook's distance plot")
```



The qqPlot looks good: all the points lie within the dashed confidence bands, indicating normality assumption is met. From the residual plot (residual vs fitted value), we notice variance appears to increase with fitted value which might be a problem, but don't worry about that since there are only four data points in the early part of the graph with reduced variance (anything can happen with 4 points). The Cook's distance plot shows that the 26th observation has the highest Cook's distance value, but it's Cook's distance is less than .40, indicating it is not an influencial outlier, so there appears to be no problems with outliers.

C) By "hand" get the partial regression plot for the first variable "Conc". To do this, regress creatinine clearance on Age and weight, and store the residuals. Then regress "Conc" on Age and Weight, storing those residuals. Use these two groups of residuals to create the partial regression plot for "Conc".

```
# Regress response variable on Age and Weight and obtain the residuals
lmq4_a <- lm(CreatClear ~ Age + Weight, data = Hwk3Q4DatSp17)
resid_a <- lmq4_a$resid
# Regress "Conc" on Age and Weight and obtain the residuals
lmq4_b <- lm(Conc ~ Age + Weight, data = Hwk3Q4DatSp17)
resid_b <- lmq4_b$resid
plot(resid_a ~ resid_b, xlab="Residual from Regressing Conc on Age and Weight",
     ylab="Residual from Regressing CreatClear on Age and Weight",
     main="Partial Regression Plot for Conc", cex.lab=0.8)
```



The partial regression plot shows a clear downward trend between the two group of residuals, indicating a significant negative slope for “Conc”..

D) Regress these two residuals on each other, being sure to use as your predictor variable the residuals from regression “Conc” on Age and Weight. What is the resulting regression equation?

```
lm_resid <- lm(resid_a ~ resid_b)
summary(lm_resid)
```

```
##
## Call:
## lm(formula = resid_a ~ resid_b)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.668  -7.002   1.518   9.905  16.006
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.544e-15  2.097e+00   0.000        1
## resid_b     -3.994e+01  5.416e+00  -7.374 2.65e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.05 on 31 degrees of freedom
## Multiple R-squared:  0.6369, Adjusted R-squared:  0.6252
## F-statistic: 54.37 on 1 and 31 DF,  p-value: 2.655e-08
```

Let res_Y be the residual from regressing the response variable on Age and Weight, and res_{Conc} be the residual from regressing Conc on Age and Weight. The estimated regression equation is

$$res_Y = -39.94 \times res_{Conc}.$$

E) How does the coefficient of "Conc" in the regression equation obtained in part D compare to the regression coefficient of "Conc" obtained in part A?

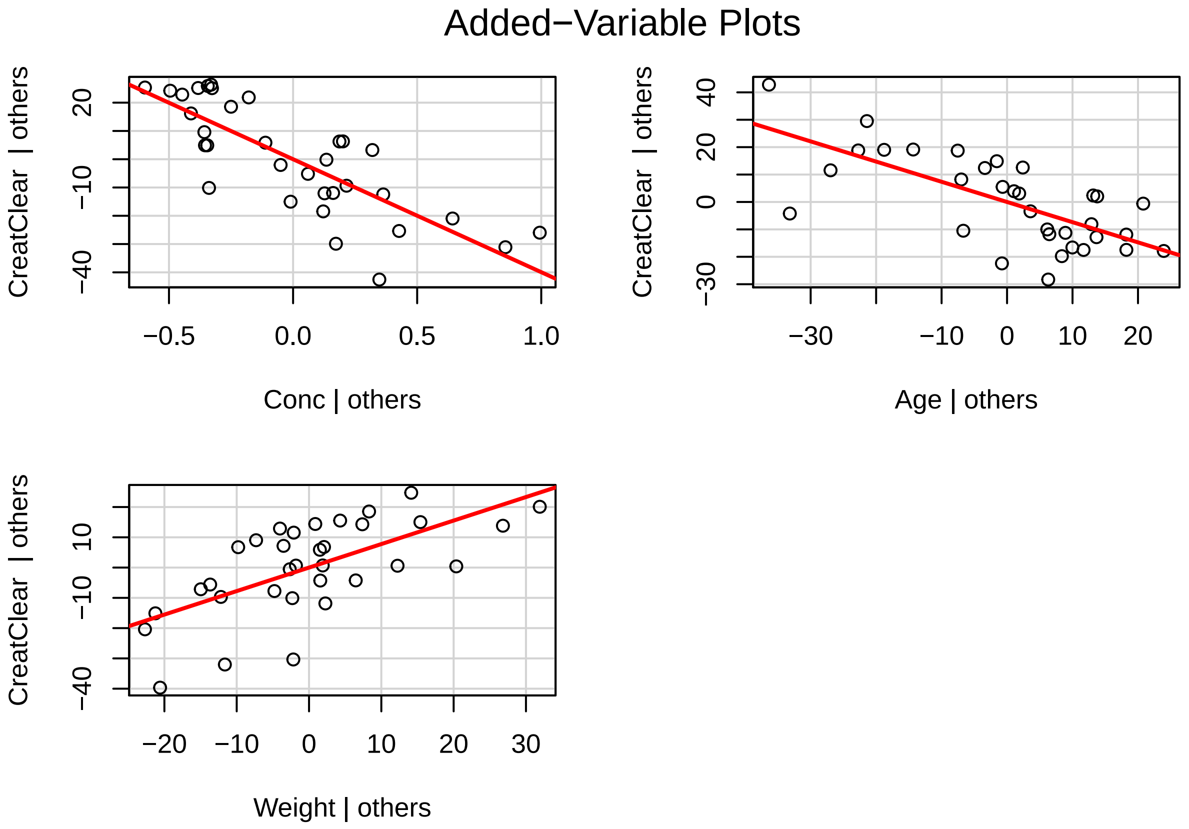
The estimated coefficient of "Conc" in the regression in part D is -39.94 which is roughly of the same amount as in the regression in part A.

F) Explain why your findings in Part E above make perfect sense from an intuitive perspective.

The similar estimated coefficient values of "Conc" between part A and part D tells us at least two things. First, there are variations in the response variable that cannot be fully accounted for by Age and Weight. Second, in terms of explaining variations in the reponse variable, little amount of "Conc" has can be accounted for by Age and Weight. The results make sense because a patient's creatinine concentration is likely to a more important factor in explaining the creatinine clearance than the other two predictors. The other two predictors (Age and Weight) explains a small amount of variations in creatinine clearance and "Conc". Therefore, it makes sense the estimated coefficient of "Conc" in part D is almost equal to that in part A.

G) Use the "library(car)" and avPlots(LinearModelName) to get the partial regression plots for all three variables. Do you see any curvilinearity in any of these plots?

```
avPlots(lmq4)
```



Two variables appear to have possible curvilinearity issues: Age, and to a lesser degree, weight.

H) Add some polynomial terms to test for the curvilinearity you saw in part G. If adding multiple terms, be sure to add them to your linear model last so you can simultaneously test them. Then do so. What are your conclusions? What model should you use?

```
#Check for curvilinearity in Age first
# Start by adding second order and third order terms of Age in the model
lmq4_c <- lm(CreatClear ~ Conc + Age + Weight + I(Age^2) + I(Age^3), data = Hwk3Q4DatSp17)
summary(lmq4_c)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight + I(Age^2) + I(Age^3),
##     data = Hwk3Q4DatSp17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.061  -6.846   3.197   6.670  20.095
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.266e+01  4.214e+01   2.199   0.0366 *
## Conc        -4.321e+01  6.167e+00  -7.007 1.57e-07 ***
## Age          1.885e+00  2.684e+00   0.702   0.4885
## Weight       8.280e-01  1.721e-01   4.811 5.06e-05 ***
## I(Age^2)    -6.995e-02  5.819e-02  -1.202   0.2397
## I(Age^3)     5.442e-04  3.975e-04   1.369   0.1823
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.27 on 27 degrees of freedom
## Multiple R-squared:  0.8688, Adjusted R-squared:  0.8445
## F-statistic: 35.75 on 5 and 27 DF,  p-value: 4.344e-11
```

```
# Both polynomial terms are not significant; reduce to quadratic model
lmq4_d <- lm(CreatClear ~ Conc + Age + Weight + I(Age^2), data = Hwk3Q4DatSp17)
summary(lmq4_d)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight + I(Age^2), data = Hwk3Q4DatSp17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.272  -7.739   2.013   9.573  16.798
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.649244  24.819415   5.627 5.02e-06 ***
## Conc        -42.660019   6.249244  -6.826 2.04e-07 ***
## Age          -1.591883   0.881169  -1.807   0.0816 .
## Weight        0.796592   0.173203   4.599 8.29e-05 ***
## I(Age^2)      0.008782   0.008932   0.983   0.3339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.46 on 28 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8396
## F-statistic: 42.88 on 4 and 28 DF,  p-value: 1.498e-11
```

Since we notice potential curvilinearity in Age, we start by adding second order and third order of Age into the model. From the first output above, we see neither of the two added polynomial terms are significant. So we reduce the order to quadratic by only keeping the second order term of Age. However, Age^2 is still not significant. Hence, we conclude curvilinearity is not an issue for the variable Age in our initial multiple linear model.

```
#Check for curvilinearity in Weight
# Start by adding second order and third order terms of Weight in the model
lmq4_h <- lm(CreatClear ~ Conc + Age + Weight + I(Weight^2) + I(Weight^3), data = Hwk3Q4DatSp17)
summary(lmq4_h)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight + I(Weight^2) +
##     I(Weight^3), data = Hwk3Q4DatSp17)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.871  -8.435   3.122   8.497  16.514
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.504e+01  2.660e+02  -0.357    0.724    
## Conc        -4.179e+01  5.644e+00  -7.403 5.79e-08 ***
## Age         -6.745e-01  1.450e-01  -4.653 7.73e-05 ***
## Weight       8.358e+00  1.071e+01   0.780    0.442    
## I(Weight^2) -8.516e-02  1.412e-01  -0.603    0.552    
## I(Weight^3)  3.009e-04  6.059e-04   0.497    0.623    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.3 on 27 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8438 
## F-statistic: 35.58 on 5 and 27 DF,  p-value: 4.596e-11
```

Both polynomial terms are not significant; reduce to quadratic model in Weight:

```
lmq4_h2 <- lm(CreatClear ~ Conc + Age + Weight + I(Weight^2), data = Hwk3Q4DatSp17)
summary(lmq4_h2)
```

```
## 
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight + I(Weight^2), 
##     data = Hwk3Q4DatSp17)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -28.350  -8.156   2.928   7.571  17.056 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  34.107359  55.539845   0.614   0.5441    
## Conc        -41.716086   5.566167  -7.495 3.66e-08 ***
## Age          -0.685050   0.141456  -4.843 4.26e-05 ***
## Weight        3.086307   1.451508   2.126   0.0424 *  
## I(Weight^2)  -0.015169   0.009469  -1.602   0.1204    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 12.13 on 28 degrees of freedom
## Multiple R-squared:  0.867,  Adjusted R-squared:  0.848 
## F-statistic: 45.64 on 4 and 28 DF,  p-value: 7.113e-12
```

We see that the quadratic term in Weight is nonsignificant. Thus, we still use the initial multiple linear model with the three predictors, and no polynomial terms.

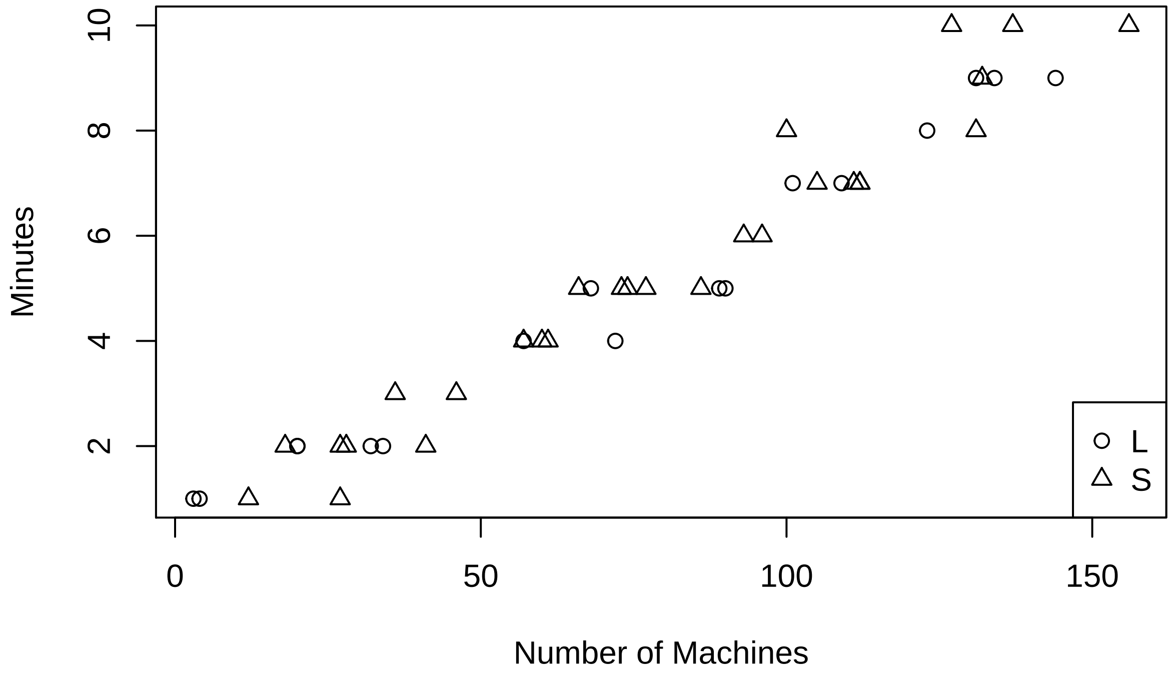
Question 5.

A company services copiers in large businesses and institutions. These businesses and institutions, although they have many copiers, have either the large or small model. A manager wants to relate the time it takes one of his technicians to make a service call on the number of copiers serviced and whether or not they are large or small. Data on a random sample of 45 service calls records the number of machines serviced and whether or nopt they were lage (L) or small (S). Data appears in Hwk3Q5DatSp17. (Note that below you may assume assumptions for inference are met).

A) Plot the relationship between service time and number of machines by machine type. What do you hypothesize for the relationships between service time and number of machines for the two types of machines?

```r
library(readxl)
Hwk3Q5DatSp17 <- read_excel("~/BTRY6020Sp17/Homework/Hwk3/Hwk3Q5DatSp17.xlsx")
Type.f<-factor(Hwk3Q5DatSp17$Type)
pchSelect = rep(1, length(Type.f))
pchSelect[which(Type.f == 'L')] = 2
plot(Hwk3Q5DatSp17$Mins, Hwk3Q5DatSp17$Machs, pch = pchSelect, xlab = "Number of Machines", ylab = "Minu

legend("bottomright", legend=levels(Type.f), pch=c(1:2))
```



With the symbols of both types apparently intermingled within the graph, it doesn't appear that the type of machine really effects service time.

B) Run a regression of service time as a function of number of machines serviced and type of machine. Interpret the coefficient of "Type".

```r
MachNoInt.lm<-lm(Mins ~ Machs + Type, data = Hwk3Q5DatSp17)
summary(MachNoInt.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs + Type, data = Hwk3Q5DatSp17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5390  -4.2515   0.5995   6.5995  14.9330
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.9225     3.0997  -0.298    0.767    
## Machs        15.0461     0.4900  30.706   <2e-16 ***
## TypeS         0.7587     2.7799   0.273    0.786    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.011 on 42 degrees of freedom
## Multiple R-squared:  0.9576, Adjusted R-squared:  0.9556 
## F-statistic: 473.9 on 2 and 42 DF,  p-value: < 2.2e-16
```

Here we see that $\hat{\beta}_2 = .7587$ for TypeS, so it appears for a fixed number of machines, it takes on average .7587 minutes more to fix the same number of small machines as it does large machines.

C) For a fixed number of machines, does the type make any difference? State hypotheses, p-value, and conclusions.

From the output above, we test $H_o : \beta_2 = 0 \ \ vs \ \ H_a : \beta_2 \neq 0$ The test statistic is given in the output above: t value = .273, with p = .786. So we have no evidence to conclude that the type of machine makes any difference in service time, for any fixed number of machines.

D) As the number of machines increases, does the amount of service time increase at the same rate for both types of machines? State hypotheses, p-value, and conclusions.

```r
MachInt.lm<-lm(Mins ~ Machs + Type + Machs:Type, data = Hwk3Q5DatSp17)
summary(MachInt.lm)
```

```
## 
## Call:
## lm(formula = Mins ~ Machs + Type + Machs:Type, data = Hwk3Q5DatSp17)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -19.2072  -6.7887  -0.1708   7.1504  14.7441 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   2.8131     3.6468   0.771   0.4449    
## Machs        14.3394     0.6146  23.333   <2e-16 ***
## TypeS        -8.1412     5.5801  -1.459   0.1522    
## Machs:TypeS   1.7774     0.9746   1.824   0.0755 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 8.771 on 41 degrees of freedom
## Multiple R-squared:  0.9608, Adjusted R-squared:  0.9579 
## F-statistic: 334.6 on 3 and 41 DF,  p-value: < 2.2e-16
```

The rate of increase in service time is given by the slopes of the lines of both types. The difference in slopes is given by the interaction term, Machs:Types, the third predictor in the model. To test if the slopes are the same, test to see if the coefficient of the interaction terms is the same: $H_o : \beta_3 = 0 \ vs \ H_a : \beta_3 \neq 0$. The p-value of this test is given in the linear model summary above: p = .0755. So we conclude (at $alpha \ = .05$) that we have insufficient evidence to say there is any difference in the rate of increase in service time between the two types of machines as the number of machines increases.

E) Is the linear relationship between number of machines serviced and the time it takes to service them the same for both types of machines? State hypotheses, test statistic, p-value, and conclusions.

```
anova(MachInt.lm)
```

```
## Analysis of Variance Table
##
## Response: Mins
##            Df Sum Sq Mean Sq   F value  Pr(>F)    
## Machs       1  76960   76960 1000.2987 < 2e-16 ***
## Type        1      6       6    0.0786 0.78059    
## Machs:Type  1    256     256    3.3260 0.07549 .  
## Residuals  41   3154      77                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we need to do a simultaneous test:

$$H_o : \beta_2 = \beta_3 = 0 \ vs \ H_a : Not \ H_o$$

Here our test statistic will be:

$$TS = F = \frac{\frac{\Delta SSR}{2}}{MSE(full)} = \frac{\frac{256+6}{2}}{77} = 1.701$$

and p = $P(F_{2,41} > 1.701) =$

```
1-pf(1.701,2,41)
```

```
## [1] 0.1951269
```

With p = .195, we conclude there is no evidence that both of these coefficients are not zero, and drop them from the model.

F) What model should this manager use to predict the service time reqyuired for his technicians going out on a service call?

Since the type of machine apparently doesn't make any difference, we can use a simple linear regression of Mins on Machs:

```
MachsSimp.lm <- lm(Mins ~ Machs, data=Hwk3Q5DatSp17)
summary(MachsSimp.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs, data = Hwk3Q5DatSp17)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -22.7723  -3.7371   0.3334   6.3334  15.4039 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.5802     2.8039  -0.207    0.837    
## Machs        15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

This yields the model: $\hat{y}_i \ = \ -.5802 + 15.0352\,(Machs)$