BTRY 6020 Homework III

NAME: ANDRES CASTANO

NETID: ac986

DUE DATE: February 27 2017, by 8:40 am

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).

Initially, we are going to load the data:

```
library(readxl)
data_satist = read_excel("Hwk3Q1DatSp17(1).xlsx")
head(data_satist)
```

```
##   PatSat Age Sev Anx
## 1     48  50  51 2.3
## 2     57  36  46 2.3
## 3     66  40  48 2.2
## 4     70  41  44 1.8
## 5     89  28  43 1.8
## 6     36  49  54 2.9
```

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

```
satisf.lm = lm(PatSat ~ Age + Sev + Anx, data = data_satist)
summary(satisf.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Age + Sev + Anx, data = data_satist)
##
## 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
```

The estimated regression equation is:

$$E(Y|X_{i1}, X_{i2}, X_{i3}) = \hat{Y} = 158.49 - 1.1416 * Age_i - 0.4420 * Sev_i - 13.4702 * Anx_i$$

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

```
# Age (Age) vs severity of condition (Sev)
cor(data_satist$Age, data_satist$Sev)
```

```
## [1] 0.5679505
```

```
# Age (Age) vs level of anxiety (Anx)
cor(data_satist$Age, data_satist$Anx)
```

```
## [1] 0.5696775
```

```
# Severity of condition (Sev) vs level of anxiety (Anx)
cor(data_satist$Sev, data_satist$Anx)
```

```
## [1] 0.6705287
```

```
# A more efficiente way to do it is:
cor(data_satist[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
```

Since the correlation is reasonable high among the explanatory variables, the multicollinearity may be a problem and therefore influence the precision of the estimated coefficients (higher sampling variances).

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.

The VIF for a particular variable X_j is defined as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where R_j^2 is the multiple correlation coefficient obtained from the regression of X_j against the other X's used in the problem of interest. In our case, we have 3 explanatory variables then we need to get three VIF'S. Those VIFs are the multiple correlation coefficients obtained for the following regressions:

$$Age_i = \beta_0 + \beta_1 Sev_i + \beta_2 Anx_i + \epsilon_i \quad (1)$$

$$Sev_i = \beta_0 + \beta_1 Age_i + \beta_2 Anx_i + \epsilon_i \quad (2)$$

$$Anx_i = \beta_0 + \beta_1 Age_i + \beta_2 Sev_i + \epsilon_i \tag{3}$$

In R, we run the regressions above as follows:

```
m1.lm = lm(Age ~  Sev + Anx, data = data_satist)
summary(m1.lm)
```

```
##
## Call:
## lm(formula = Age ~ Sev + Anx, data = data_satist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.8527  -4.6453   0.1767   4.5288  15.0583
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.2182    12.4938  -1.618   0.1129
## Sev           0.6985     0.3326   2.100   0.0416 *
## Anx          10.2224     4.7933   2.133   0.0387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.141 on 43 degrees of freedom
## Multiple R-squared:  0.3874, Adjusted R-squared:  0.3589
## F-statistic: 13.59 on 2 and 43 DF,  p-value: 2.66e-05
```

```
m2.lm = lm(Sev ~  Age + Anx, data = data_satist)
summary(m2.lm)
```

```
##
## Call:
## lm(formula = Sev ~ Age + Anx, data = data_satist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9010 -2.4312 -0.1823  2.1135  7.5820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.39344    3.58056   7.930 5.96e-10 ***
## Age          0.13317    0.06341   2.100 0.041628 *
## Anx          7.40239    1.88916   3.918 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.118 on 43 degrees of freedom
## Multiple R-squared:  0.5008, Adjusted R-squared:  0.4776
## F-statistic: 21.57 on 2 and 43 DF,  p-value: 3.257e-07
```

```
m3.lm = lm(Anx ~  Age + Sev, data = data_satist)
summary(m3.lm)
```

```
##
## Call:
```

```
## lm(formula = Anx ~ Age + Sev, data = data_satist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41358 -0.12441 -0.01547  0.13261  0.53048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.135068   0.388793   0.347 0.729983
## Age         0.009357   0.004388   2.133 0.038705 *
## Sev         0.035544   0.009071   3.918 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.216 on 43 degrees of freedom
## Multiple R-squared:  0.5023, Adjusted R-squared:  0.4791
## F-statistic: 21.69 on 2 and 43 DF,  p-value: 3.059e-07
```

Then we can see that the multiple linear correlation coefficients for each regression specified above are 0.3874 (equation 1), 0.5008 (equation 2) and 0.5023 (equation 3).

Then the VIf for each predictor is:

```
R_Age = 0.3874
R_Sev = 0.5008
R_Anx = 0.5023
VIF_Age = 1 / (1 - R_Age)
VIF_Sev = 1 / (1 - R_Sev)
VIF_ANX = 1 / (1 - R_Anx)
VIF_Age
```

```
## [1] 1.632387
```

```
VIF_Sev
```

```
## [1] 2.003205
```

```
VIF_ANX
```

```
## [1] 2.009243
```

The result of the VIF for the three variables shown that the multicollinearity may not be a huge problem after all (all the VIFs are quite less than 10, which is the the practical threshold to determine if the multicollinearity is a problem).

We can confirm our results with the following command:

```
library(car)
vif(satisf.lm)
```

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

Exactly the same than mannually.

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

To answer this question we need to run three simple linear regressions between Y and each X, then compare the results with the full model to determine which predictor(s) appear non-significant in the presence of the other two. In R, we can do this as follows:

```r
reg1.lm = lm(PatSat ~ Age, data = data_satist)
summary(reg1.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Age, data = data_satist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9281  -9.6808   0.7573  10.8913  17.7986
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 119.9432     7.0848  16.930  < 2e-16 ***
## Age          -1.5206     0.1799  -8.455 9.06e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.76 on 44 degrees of freedom
## Multiple R-squared:  0.619,  Adjusted R-squared:  0.6103
## F-statistic: 71.48 on 1 and 44 DF,  p-value: 9.058e-11
```

```r
reg2.lm = lm(PatSat ~ Sev, data = data_satist)
summary(reg2.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Sev, data = data_satist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.203 -10.839  -1.113  10.342  30.843
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 183.0770    24.3249   7.526 1.95e-09 ***
## Sev          -2.4093     0.4806  -5.013 9.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.91 on 44 degrees of freedom
## Multiple R-squared:  0.3635, Adjusted R-squared:  0.3491
## F-statistic: 25.13 on 1 and 44 DF,  p-value: 9.23e-06
```

```r
reg3.lm = lm(PatSat ~ Anx, data = data_satist)
summary(reg3.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Anx, data = data_satist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.369  -9.606  -1.946   9.212  31.631
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  146.449     15.304   9.569 2.55e-12 ***
## Anx          -37.117      6.637  -5.593 1.33e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.33 on 44 degrees of freedom
## Multiple R-squared:  0.4155, Adjusted R-squared:  0.4022
## F-statistic: 31.28 on 1 and 44 DF,  p-value: 1.335e-06
```

The results above show that individually each predictor has a statistical significant influence in the patient satisfaction. To determine which one loss significance, we compare the results above with the full model:

```
summary(satisf.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Age + Sev + Anx, data = data_satist)
##
## 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
```

We can see that the predictors Sev (patient's severity condition) and Anx (patient's level of anxiety) are not longer significant in the full model (includes all the predictors). We can find an explanation for these If we consider the correlation among the explanatory variables (see correlation matrix below). The variables patient's severity condition and patient's level of anxiety are highly correlated (coefficient of 0.67), which in turn is inflating the variance of each coefficient and affecting the t test for individual significance (the t test statistics has the standard error of the coefficient in the denominator), since the t statistic is smaller, it will be harder to reject the null hypothesis of each coefficient being equal to zero.

```
cor(data_satist[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
```

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)?

From the simple linear regressions (estimated above), we can see that the variables that individually explaing

a higher portion of the variance in the patient satisfaction scores are Age and level of anxiety. Then, fitting the regression with only these explanatory variables we get:

```
reg4.lm = lm(PatSat ~ Age + Anx, data = data_satist)
summary(reg4.lm)
```

```
##
## Call:
## lm(formula = PatSat ~ Age + Anx, data = data_satist)
##
## 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
```

As we expect the t test statistics for each coefficient have changed. This changes do not affect the significance of Age, but affect the significance of the predictor Anxiety which is now statistical significant. This result has happened because the coefficient for Anxiety is estimated with more precision (less sampling variance), compared with the full model in which the variance of the coefficient anxiety was inflated due to the high correlation between patient's anxiety and patient's severity condition. The improvement in the precision of the coefficient for anxiety also affects its t statistics which is now larger, which makes easier to reject the null hypothesis $H_0 : B_{anxiety} = 0$

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.

Since we have removed the variable tha was worsening the multicollinearity problem, now we can interpret the coefficient for age with more confidence, but still we need to be cautious. Now it makes more sense (in practical terms) the idea of keeping the other variables constant because age is not heavily correlated with anxiety. However, from a formal point of view age and anxiety are still moderated correlated, which means that if these variables are related in a structural or causal form it will be not reasonable to assume that we can keep constant one without affecting the other. I think if we view the relation of these variables as a descriptive or empirical one, it will be easier to defend the idea that we can keep one constant without affecting the other.

That being said, then we can cautiously interpret the coefficient as follows: keeping anxiety constant, an increased in one year in patient's age decreases the patient's satisfaction index by 1.2 points on average. Besides, 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.

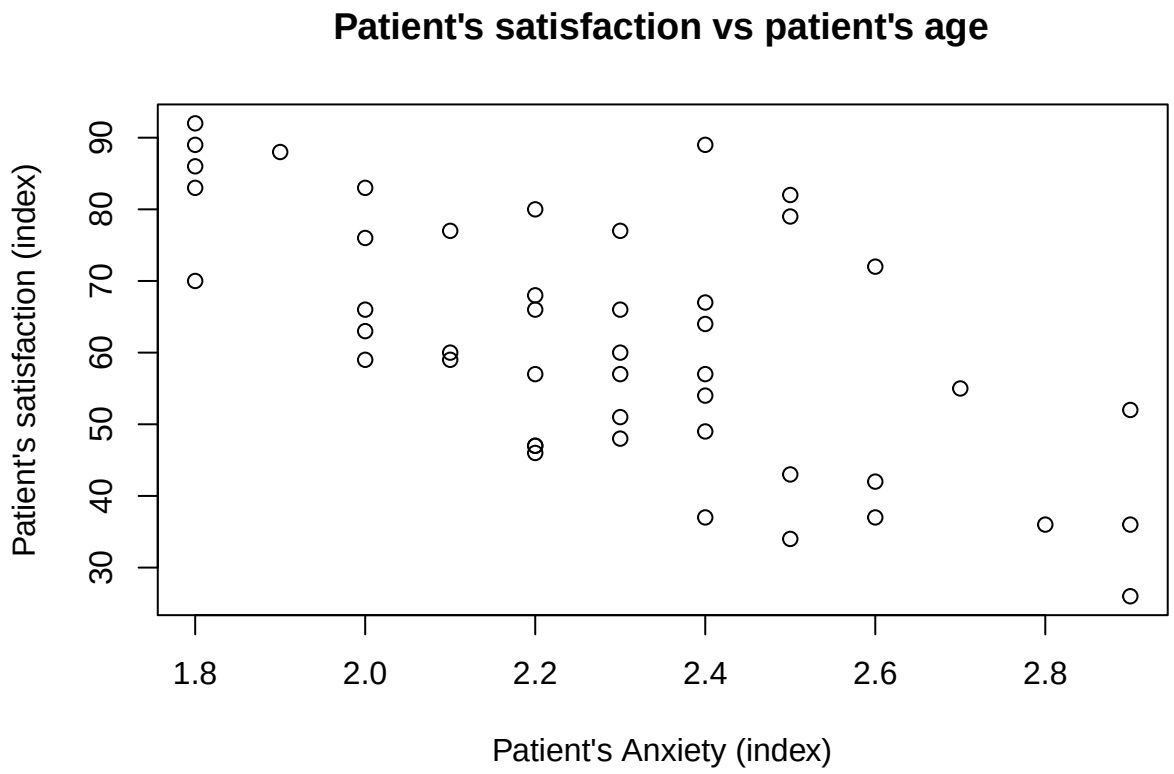
1) Since the observations came from a simple random sampling and posibly represent less than 10% of the population under study, we can defend the independence assumption.

2) To analyze curvature (linearity), we can plot the response variable against each predictor. We are going

to use the model with age and anxiety as explanatory variable.

```
plot(data_satist$Age, data_satist$PatSat, xlab = "Patient's age (years)",
     ylab = "Patient's satisfaction (index)", main = "Patient's satisfaction vs patient's age")
```

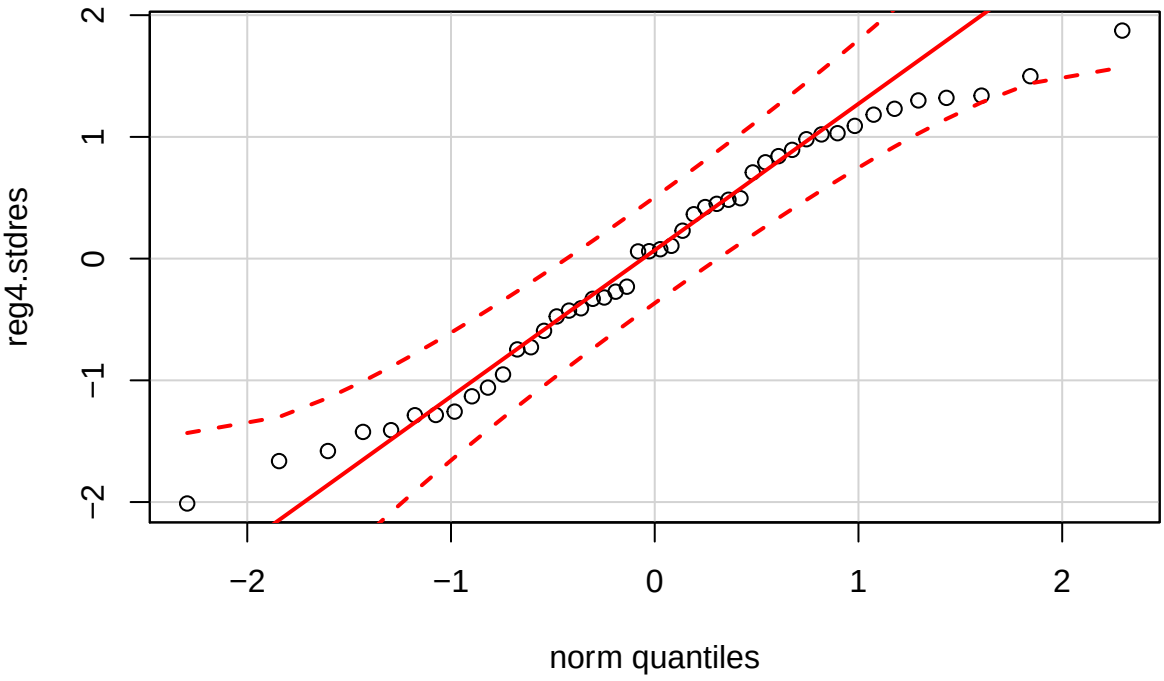

```
plot(data_satist$Anx, data_satist$PatSat, xlab = "Patient's Anxiety (index)",
     ylab = "Patient's satisfaction (index)", main = "Patient's satisfaction vs patient's age")
```



The relationship of the explanatory variables with the satisfaction index seems to be linear.

3) To assess normality of the residuals, we can make a quantile plot of the standarized residuals

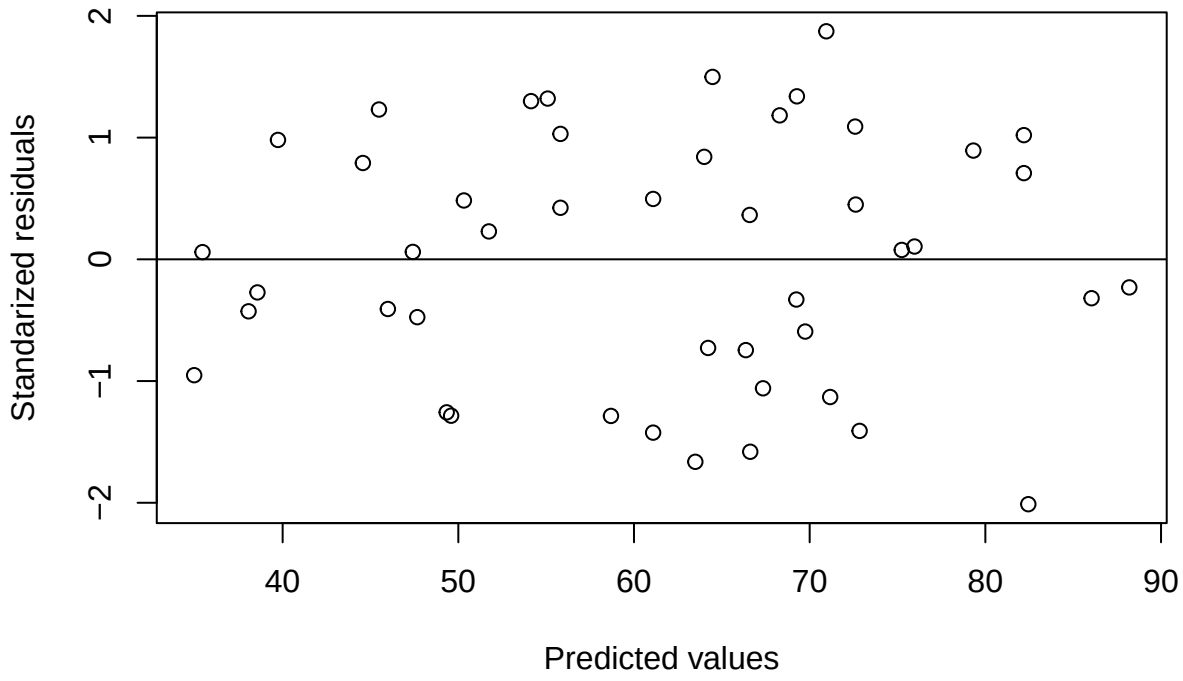
```
reg4.stdres=rstandard(reg4.lm)
library(car)
qqPlot(reg4.stdres)
```



The standarized residuals are roughly normal.

4) To assess constant variance we can plot the standardized residuals against the predicted values. Assess whether the assumption of equal variance is valid or not.

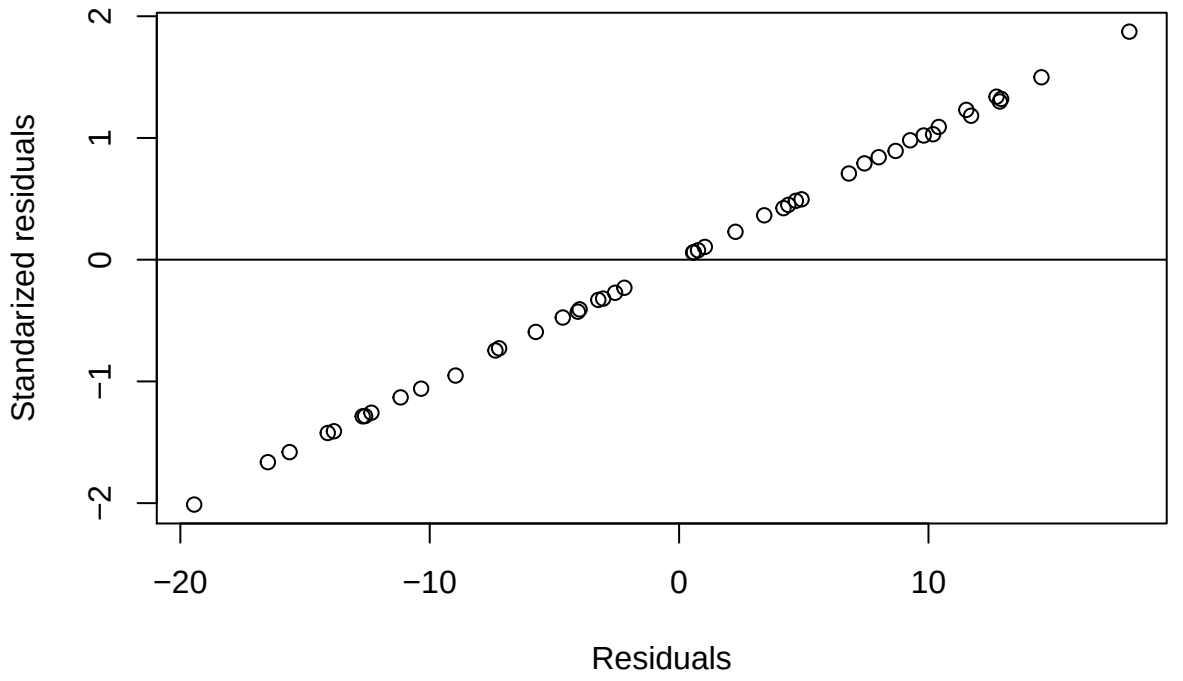
```
plot(reg4.lm$fitted.values, reg4.stdres, ylab="Standarized residuals",
     xlab="Predicted values", abline(0,0))
```



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.

5) To evaluate the presence of outliers, we can plot the residuals against the standardized residuals.

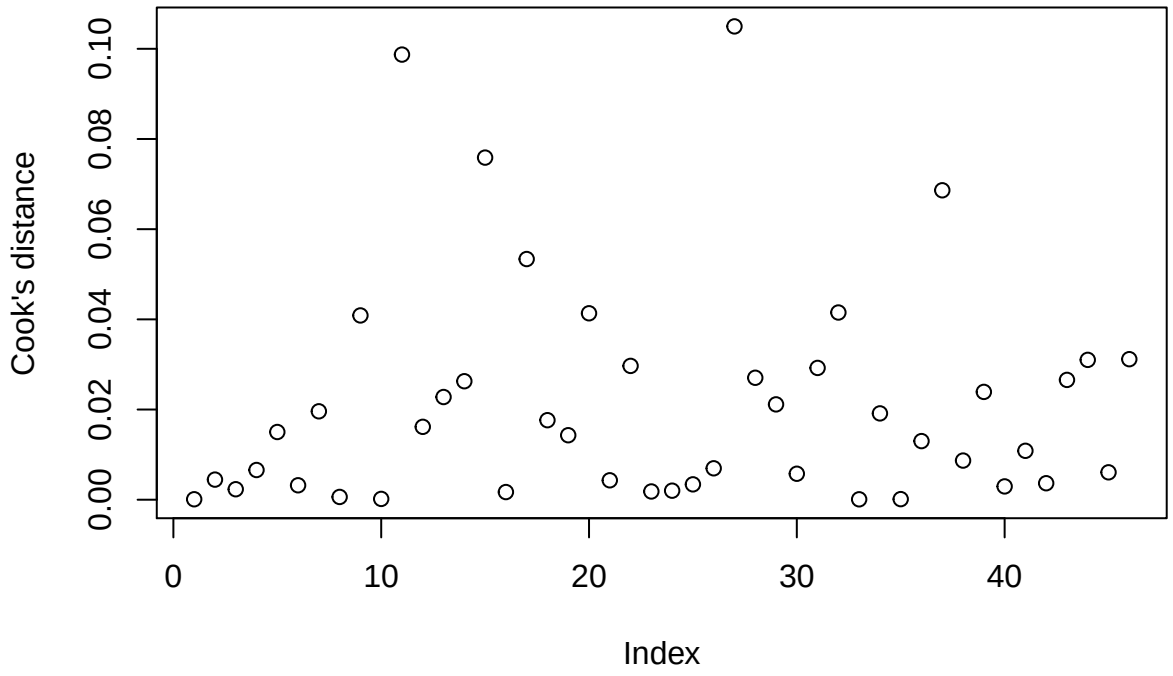
```
plot(reg4.lm$residuals, reg4.stdres, ylab="Standarized residuals", xlab="Residuals", abline(0,0))
```



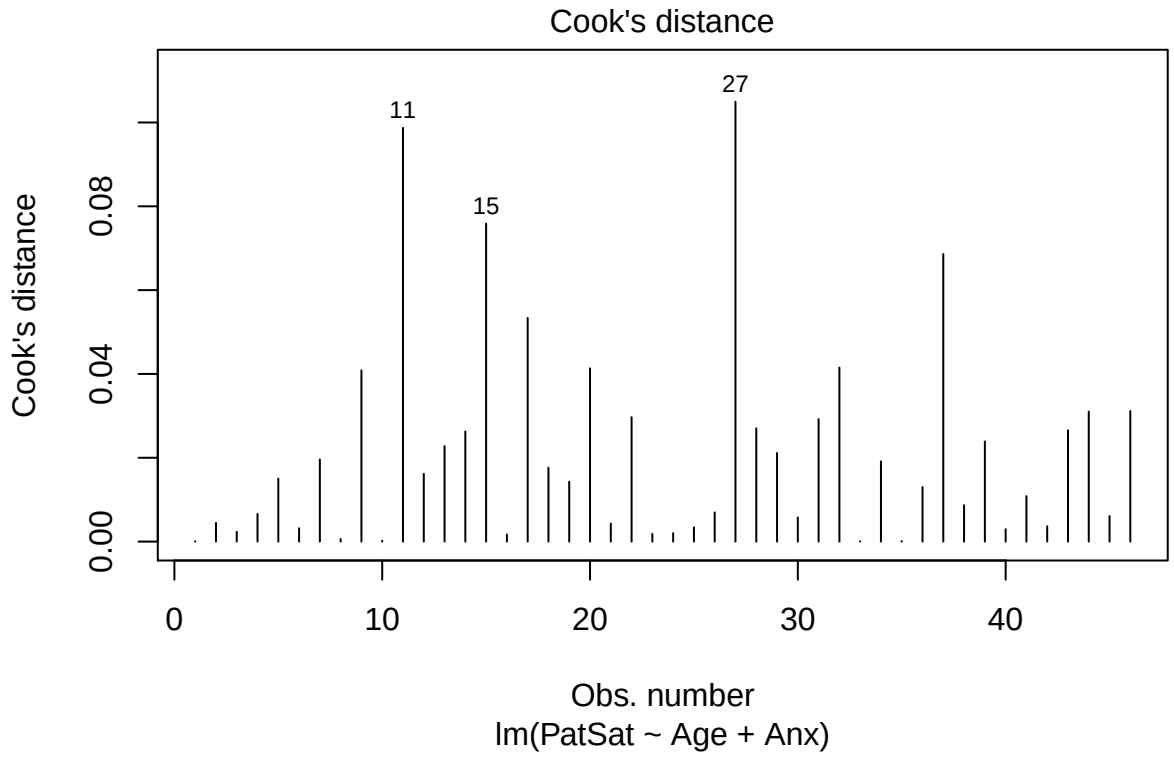
This plot is a good sign that we do not have outliers in our data. All the standarized residuals are between -2 and +2 standard deviations from the mean.

6) To assess point with high leverage and influential data points we can use a leverage plot and and a plot of the Cook's distance against observation number.

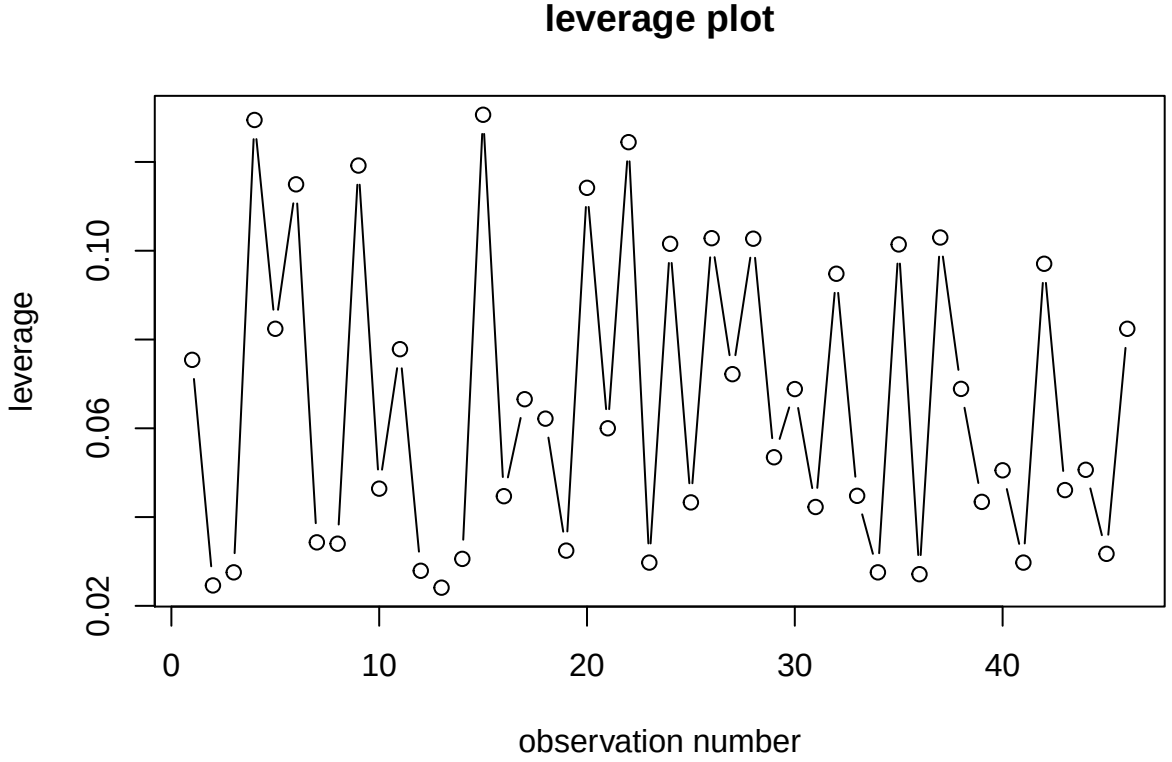
```
# cook's distance
cook_reg4 = cooks.distance(reg4.lm)
plot(cook_reg4, ylab = "Cook's distance")
```




```
# another way to get the cooks distance
plot(reg4.lm, which=4)
```




```
# Leverage plot
leverage_reg4 = hat(model.matrix(reg4.lm))
plot(leverage_reg4, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
```



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.

Here we are interested in determine wheter $\beta_1 > 0.5$, then we can formulate our hypothesis as follows:

$H_0 : \beta_1 = -0.5$

$H_A : \beta_1 > -0.5$

Our test statistics is:

$test = \frac{\hat{\beta_1} - \beta_1}{SE(\hat{\beta_1})} = \frac{-1.2005-(-0.5)}{0.2041} = -3.432$

In R, we can compute the p-value associate to this statistic as follows:

```
n=46
k=3
B1=-0.5
b1=-1.2005
SE_b1 = 0.2041
t=(b1-B1)/(SE_b1)
t
```

```
## [1] -3.432141
```

```
p_value= pt(t,n-k, lower.tail = TRUE) # pvalue related with one side test, we are only interested if b1
p_value
```

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

At significance level of 0.05, we have evidence to reject the null hypothesis (p-value<0.05). 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.)

Here we are interested in a prediction of the patient satisfaction for a specific anxiexity index and age, so in other words, we need to construct a prediction interval for this data point. We can compute the prediction interval in R as follows:

```
newdata = data.frame(Age=22, Anx=2.6)
pred_interval = predict(reg4.lm, newdata, interval="prediction", level = 0.95)
pred_interval
```

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

With these values, the prediction will 76.001. Then, for a 22 year old with an anxiety index of 2.6, the probability that the satisfaction index is between 53.45 and 98.54 is 0.95 (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 research question is ¿Does the average starting salary after graduation for a school rated 62 is greater than $70,000?

I think one of the alternatives to deal with these question is to use the simple linear regression. We can model how schools scores (X) are related with the starting salary after graduation (Y). Then, we can construct test statistic for the prediction of the mean starting salary for a school rated 62 $E(Y|X = 62)$. This prediction is a point estimate, so we can construct a t test in the traditional way to test wheter this point estimate is greater than $70,0000, and then decide wheter or not the businessman should go to this school.

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.

We are interested in estimate the equation:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Where: Y_i is starting salary for school i; X_i score of the school i; β_0 is the mean starting salary when the score of the school is 0; β_1 is the mean increase in starting salary when the score of the school increases by one unit; and ϵ_i is an error term. In this specification Y_i, β_0, β_1 and ϵ_i are random variables.

The random variables are distributed as follows:

$$Y_i \sim indepedentN(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$\epsilon_i \sim i.i.dN(0, \sigma_\epsilon^2)$$

$$\hat{\beta_0} \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2})$$

$$\hat{\beta_1} \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$$

To estimate the SLR, we assume:

1) Observations are independent ($\epsilon_i \neq \epsilon_j$)
2) The ϵ_i are normally distributed
3) The expected value of Y is a linear function of the variable X: $\mu_i = E(Y_i) = E(Y|X_i) = \beta_0 + \beta_1 X_i$
4) Constant variance of the errors.

5) Outliers not driving conclusions (implicit assumption).

The key random variable that we want to identify is the point estimate

$$E(Y|X = 62)$$

This point estimate will tell us what is the expected starting salary for a school rated 62. Then, we are going to construct a test statistic for this point in the traditional way:

$$Test = \frac{pointestimate - valueunderthenull}{SE(pointestimate)}$$

Where the point estimate will be $E(Y|X = 62)$. This test will help us to determine if the mean starting salary for a school rated 62 is greater than \$70,000, and then decide wheter or not the businessman should go to get his MBA.

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 $\alpha = .05$ if not otherwise specified.

Initially we load the data:

```
library(readxl)
data_wages = read_excel("Hwk3Q2DatSp17(1).xlsx")
head(data_wages)
```

```
##   StartSal USNscor
## 1    105.1     100
## 2    101.3     100
## 3    103.5      99
## 4    101.3      98
## 5    104.5      97
## 6    106.7      94
```

Now, we run the SLR to get:

```
wages.lm = lm(StartSal~USNscor, data = data_wages)
summary(wages.lm)
```

```
##
## Call:
## lm(formula = StartSal ~ USNscor, data = data_wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4314  -3.2987  -0.5798   3.3584   9.3204
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.25968    3.77696   5.629  9.2e-07 ***
## USNscor      0.85187    0.04999  17.042  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.701 on 48 degrees of freedom
## Multiple R-squared:  0.8582, Adjusted R-squared:  0.8552
## F-statistic: 290.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

We are interested in the point estimate $E(Y|X = 62)$. In R we can calculate this as follows:

```
x_0 = 62
b0 = 21.25968
b1 =  0.85187
E_wages = b0 + b1*x_0
E_wages
```

```
## [1] 74.07562
```

This point estimate can be also obtained:

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

```
## $fit
##        fit      lwr      upr
## 1 74.07581 72.24956 75.90206
##
## $se.fit
## [1] 0.9082936
##
## $df
## [1] 48
##
## $residual.scale
## [1] 4.701311
```

We can observe that $E(Y|X_i = 62) = 74.07562$. Then we can define our null and alternative hypothesis as follows:

$$H_0 : E(Y|X = 62) = 70$$
$$H_A : E(Y|X = 62) > 70$$

Now, to procede with our inference process, we can contruct a test statistic as follows:

$$Statistic = \frac{pointestimate - valueundernull}{SE(pointestimate)}$$

Where $SE(pointestimate)$ is the standard error for the point estimate for the expected value of Y when X is a 62. We can calculate this value as follows:

$$SE(E(Y|X = 62)) = S_e\sqrt{\frac{1}{50} + \frac{62 - \bar{X}}{(50 - 1) * (S_x^2)}}$$

We can compute this expression in R as follows:

```
x_0 = 62
SE_res=4.701
n=50
mean_x= mean(data_wages$USNscor)
var_x= var(data_wages$USNscor)
SE_point_estimate= SE_res* sqrt(((1/n) + (x_0-mean_x)/((n-1) * var_x)))
SE_point_estimate
```

```
## [1] 0.6411389
```

Then, our test statistic is:

$$Statistic = \frac{74.07562 - 70}{0.6411389} = 6.36$$

Now we can find the p-value associate with this test statistic as follows:

```
n=50
k=2
statistic = (74.07562 - 70) / (0.6411389)
statistic
```

```
## [1] 6.356844
```

```
p_value = pt(statistic, n-k, lower.tail = FALSE )
p_value
```

```
## [1] 3.56506e-08
```

$p - value < 0.05$, then we have evidence to say that the mean initial salary for a school with a score of 62 is greater than $70,000. We can conclude that the data shows that the Businessman should pursue his MBA.

On the other hand, we can calculate the 95% interval for the point estimate as follows:

$$74.07562 \pm (t_{0.025,48}) * 0.6411389$$

The quantile for the t distribution with 48 degress of freedon is:

```
qt(0.015, 48, lower.tail=FALSE)
```

```
## [1] 2.236518
```

Then our 95% confidence interval for the point estimate is:

$$74.07562 \pm (2.236518 * 0.6411389)$$

Then, the interval is:

```
lower= 74.07562 - (2.236518*0.6411389)
upper= 74.07562 + (2.236518*0.6411389)
c(lower,upper)
```

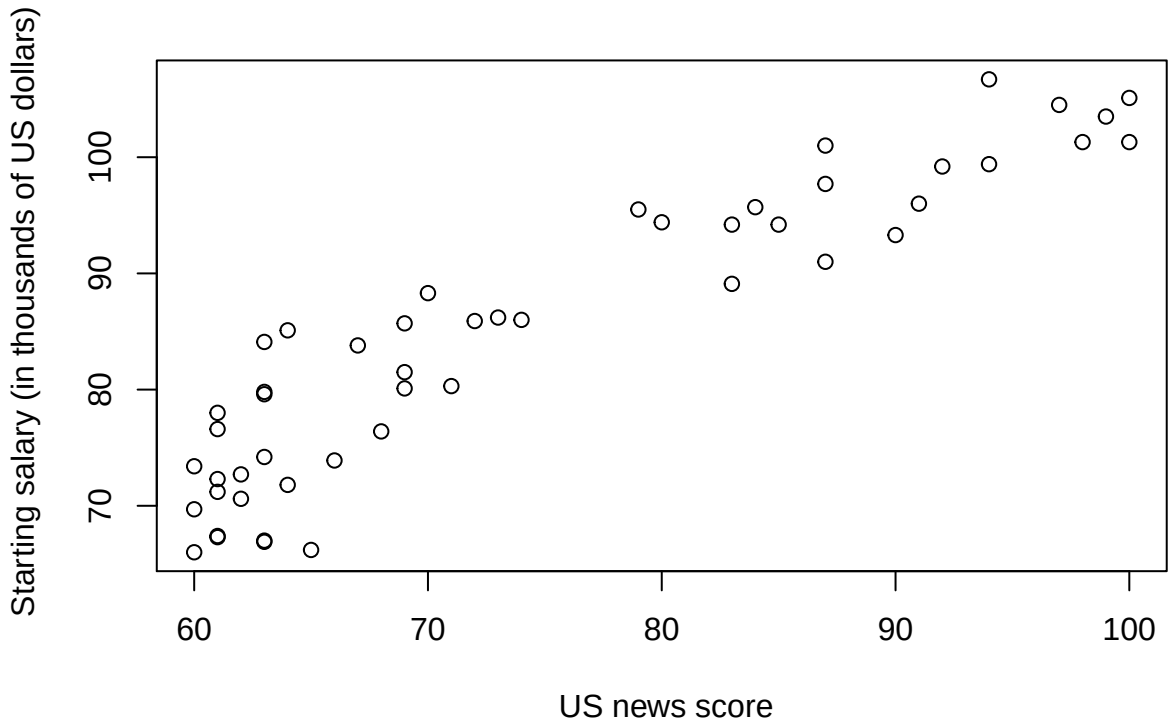
```
## [1] 72.64170 75.50954
```

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.

1) The assumption of idependence is not feasible given the data collection was not random. The database was conformed by the top 50 business schools. The violation of this assumpation can put in jeopardy the inference process.

2) To analyze curvature (linearity), we can plot the response variable against each predictor. We are going to use the model with age and anxiety as explanatory variable.

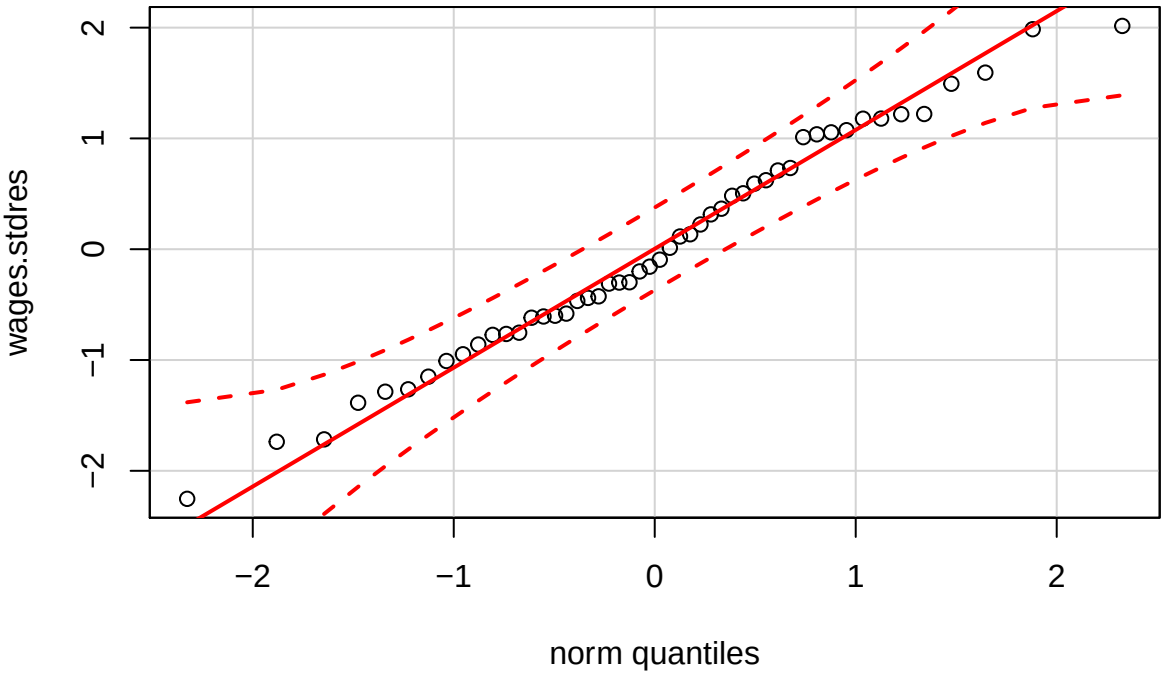
```
plot(data_wages$USNscor, data_wages$StartSal,
     ylab=" Starting salary (in thousands of US dollars)", xlab=" US news score")
```



The relationship between X and Y is roughly linear.

3) To assess normality of the residuals, we can make a quantile plot of the standarized residuals

```
wages.stdres=rstandard(wages.lm)
library(car)
qqPlot(wages.stdres)
```

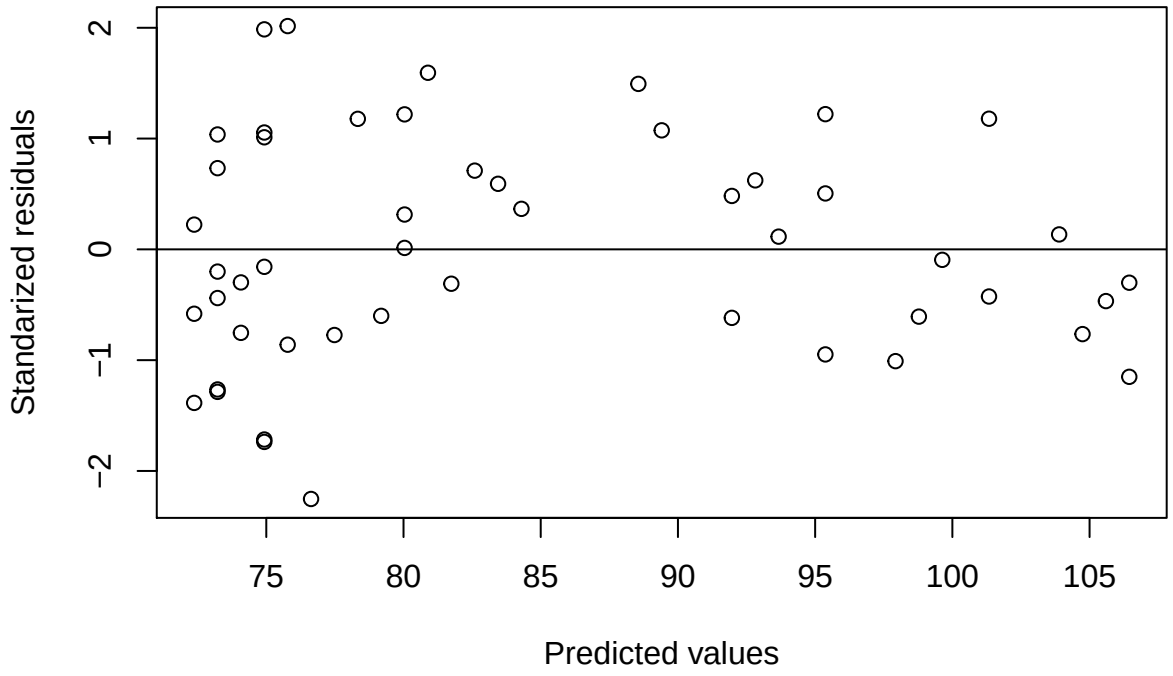


The standarized residuals are roughly normal.

4) To assess constant variance we can plot the standardized residuals against the predicted values. This

would help us to determine whether the assumption of equal variance is valid or not.

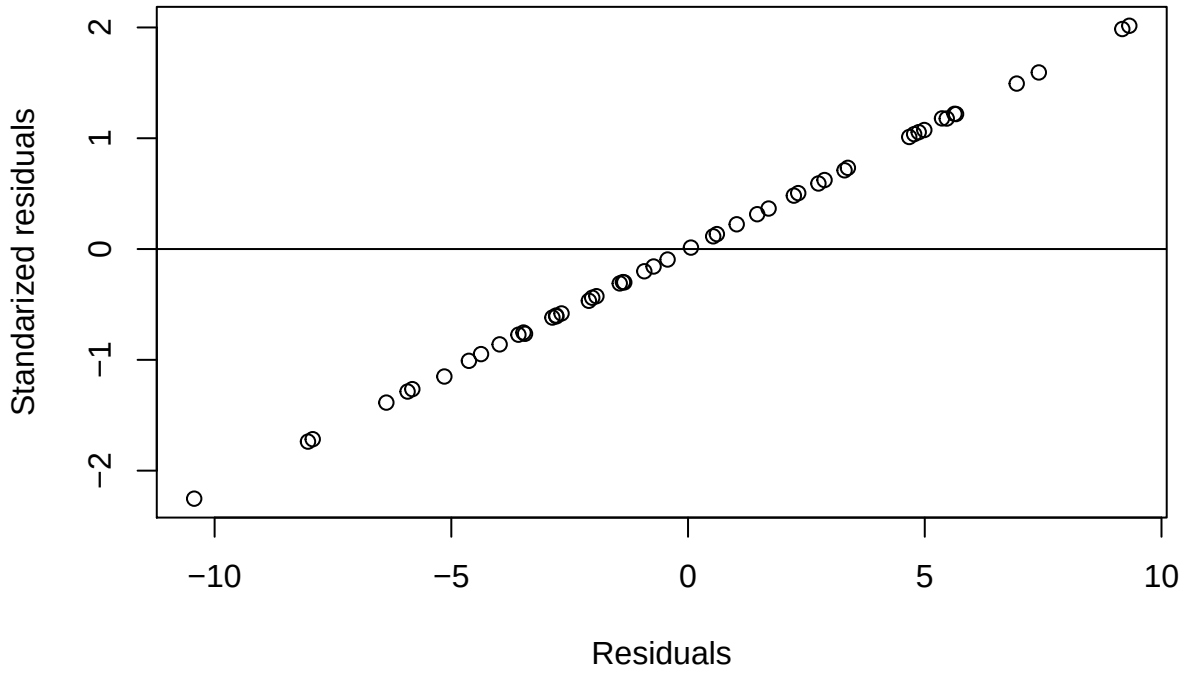
```
plot(wages.lm$fitted.values, wages.stdres, ylab="Standarized residuals", xlab="Predicted values", ablin
```



The plot seems to show a nonlinear (cuadratic) relationship between the standarized residuals and the predicted values, which is a violation of the constant variance assumption.

5) To evaluate the presence of outliers, we can plot the residuals against the standardized residuals.

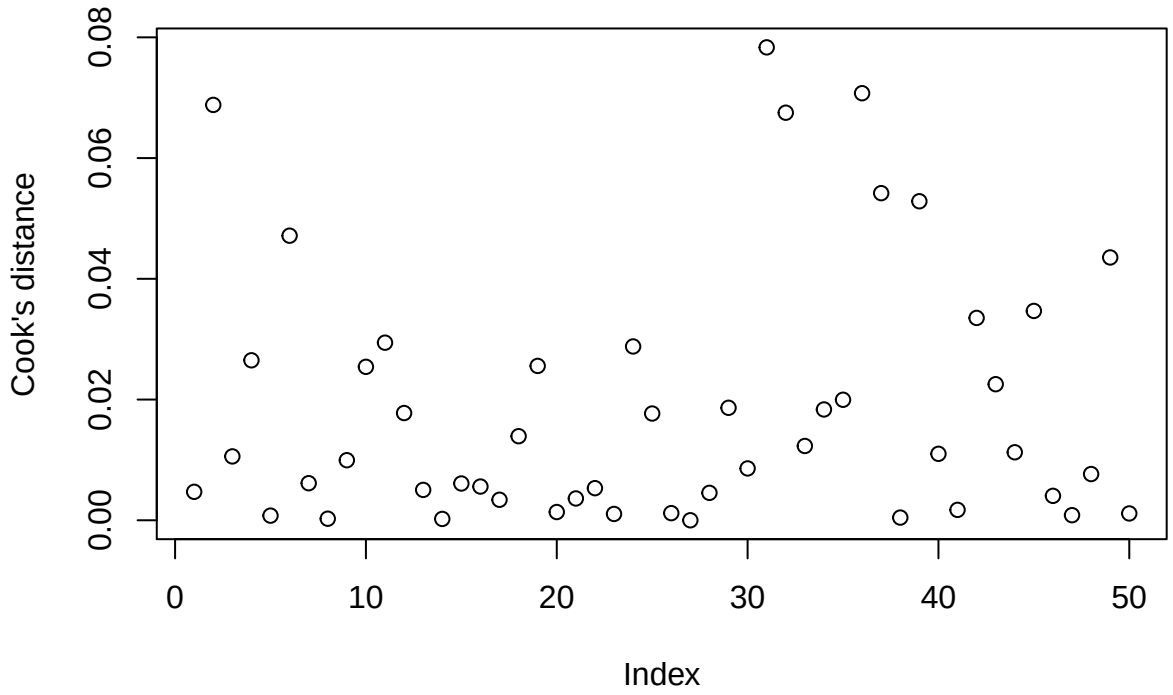
```
plot(wages.lm$residuals, wages.stdres, ylab="Standarized residuals", xlab="Residuals", abline(0,0))
```



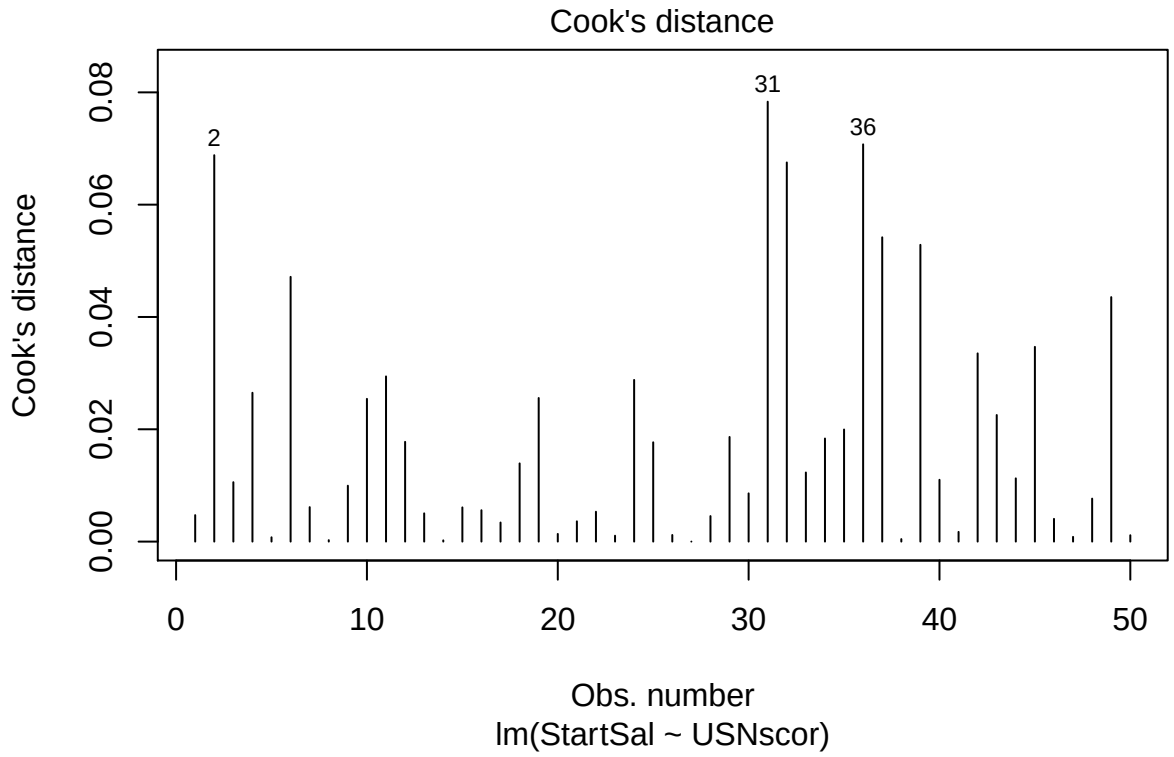
This plot is a good sign that we do not have outliers in our data. All the standarized residuals are between roughly between -2.5 and +2.5 standard deviations from the mean.

6) To assess point with high leverage and influential data points we can use a leverage plot and and a plot of the Cook's distance against observation number.

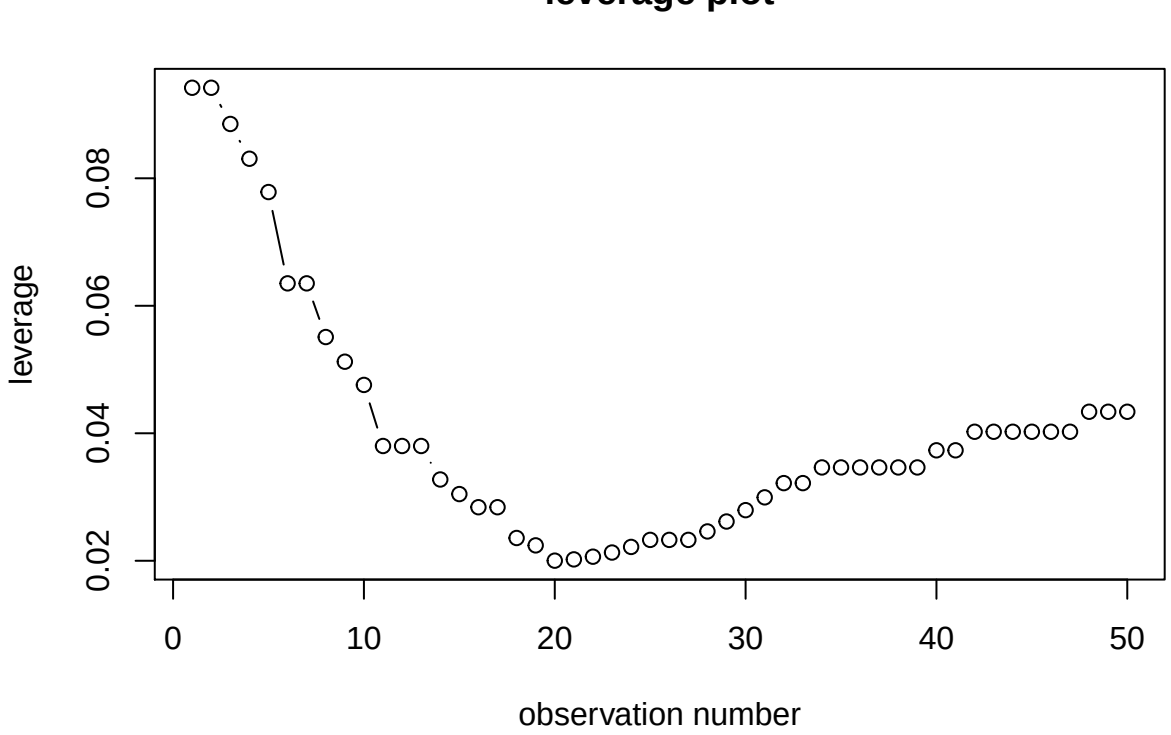
```
# cook's distance
cook_wages = cooks.distance(wages.lm)
plot(cook_wages, ylab = "Cook's distance")
```




```
# another way to cook's distance
plot(wages.lm, which = 4)
```




```
# Leverage plot
leverage_wages = hat(model.matrix(wages.lm))
plot(leverage_wages, type="b", xlab="observation number", ylab="leverage", main="leverage plot")
```



```r
mean(leverage_wages)
```

```
## [1] 0.04
```

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. However, the initial values of the explanatory variable are more influential than the others (observation number 11 and observation 27). On the other hand, the leverage plot shows that we do not have points with highly leverage compared to others.

Given the violation of the independence and constant variance it is fair to say that the assumpations are not met sufficiently for the credibility of the inference process.

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 sample scheme was not randon which causes that a pair of observations are not indepedent. The assumpation of independence is very important for the efficiency of the the least squares estimators. By efficiency we mean getting coefficient estimates with the smallest sample variance. In our case, since the lack of independence cause that the least square coefficient estimates are no fully efficient then our inference process which is heavily based on the variance of the coefficients will be compromised. I consider that the inference shown so far is not valid. We should use a estimation process that account for the heterocedasticity, that is, that take into account the no constant variance of the errors (robust regression).

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.

The statistical analysis shown that the mean starting salary for a school with a score of 62 is greater than 70,000 (p-value < 0.05), then we may recommend the businessman pursue his MBA. However, some of the assumptions of our statistical model are no met, and this affect the results of the inference. For instance, the residuals are not independent (the sampling was not random) and its variance is not constant (Heterocedasticity), this can certainly affect the efficiency of our estimation (in the sense that our estimators do not have minimun variance and we can improve our efficiency with another method). Given this considerations, I think we need to repeat the exercise starting by improving the sampling design. We should apply a simple random samplig of the schools that offer MBA and collect its US news scores, then repeat all the statistical analysis.

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.

Answer: The general linear test evalaute wheter or not a group of population slopes from the multiple regression model are 0 or not, that is $H_0 : \beta_1 = \beta_2 = \beta_q = 0$, for $1 \leq q \leq k$, whereas the F test is only a special case of this general linear test which evaluates if all population slopes are equal to cero ($H_0 : \beta_1 = \beta_2 = \beta_k = 0$). The latter is based on the Anova for the regression, whereas the former is based on the comparison of the regression summ squares (REGSS) for two models: the full model with all the variables and a null or restricted model that deletes the explanatory variables which slopes are being evaluated in the null hypothesis.

Feedback: 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.

```
library(readxl)
data_creati = read_excel("Hwk3Q4DatSp17.xlsx")
head(data_creati)
```

```
##   Obs Conc Age Weight CreatClear
## 1   1 0.71  38     71        132
## 2   2 1.48  78     69         53
## 3   3 2.21  69     85         50
## 4   4 1.43  70    100         82
## 5   5 0.68  45     59        110
## 6   6 0.76  65     73        100
```

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

```
creat.lm=lm(CreatClear~Conc + Age + Weight, data=data_creati)
summary(creat.lm)
```

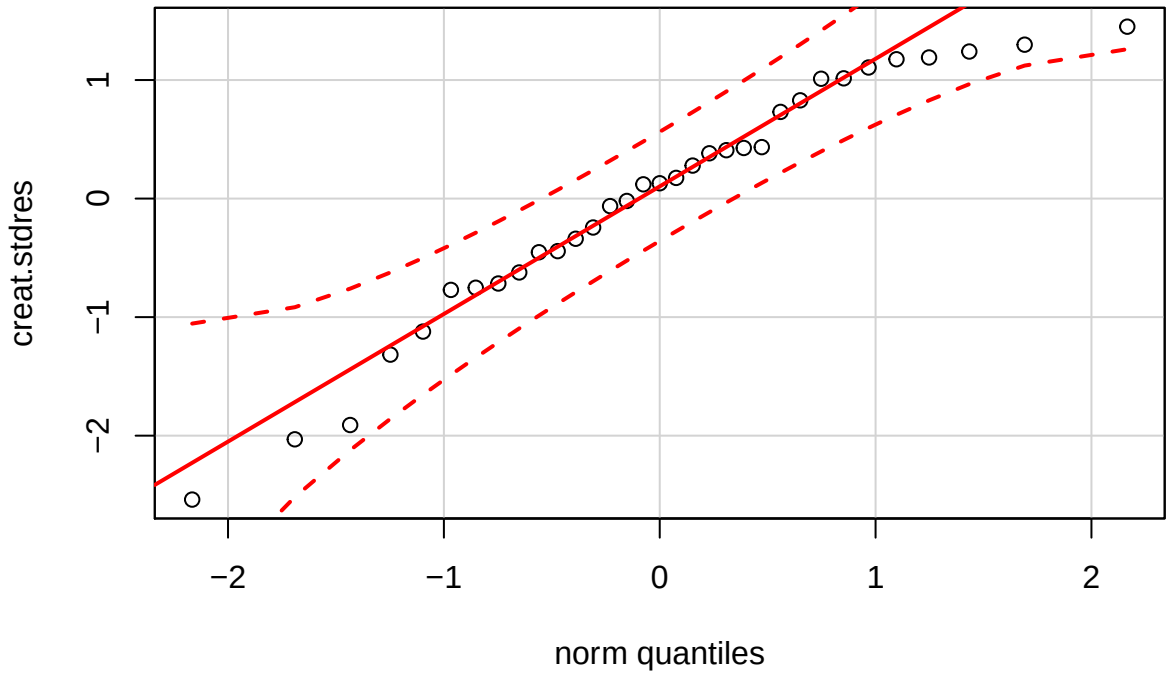
```
##
## Call:
## lm(formula = CreatClear ~ Conc + Age + Weight, data = data_creati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.668  -7.002   1.518   9.905  16.006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 120.0473    14.7737   8.126 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 resulting regression equation is:

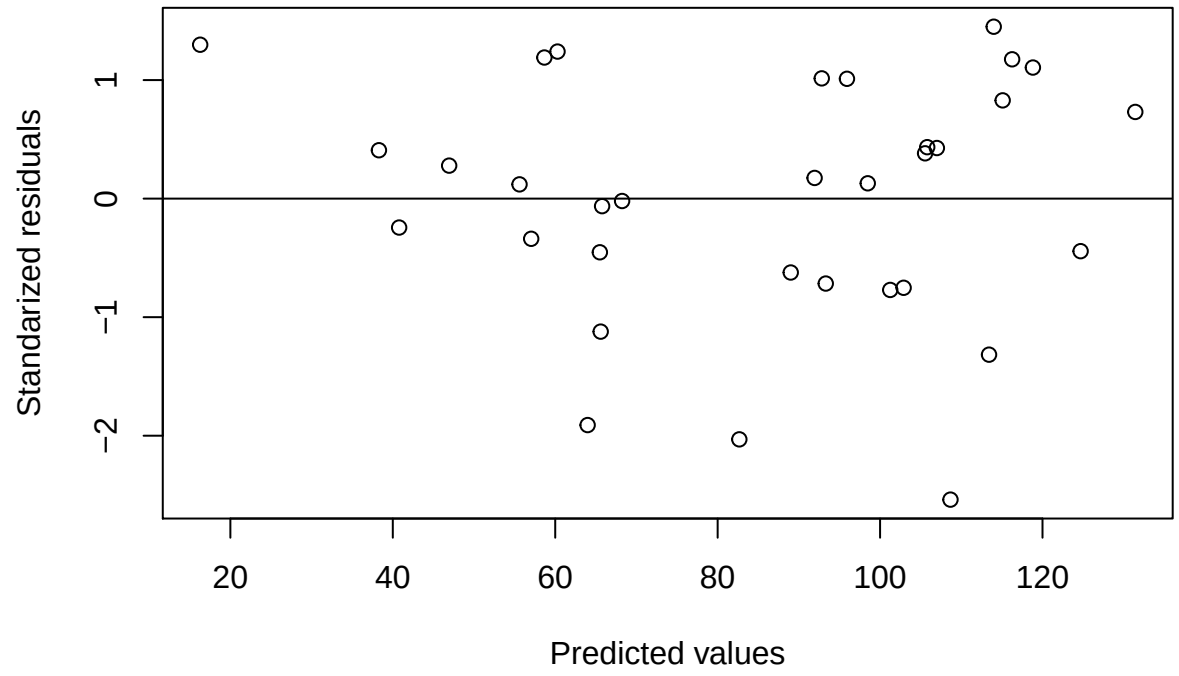
$$E(Y|Conc_i, Age_i, Weight_i) = \hat{Y}_i = 120.04 - 39.9393 * Conc_i - 0.7368 * Age_i + 0.7764 * Weight_i$$

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?

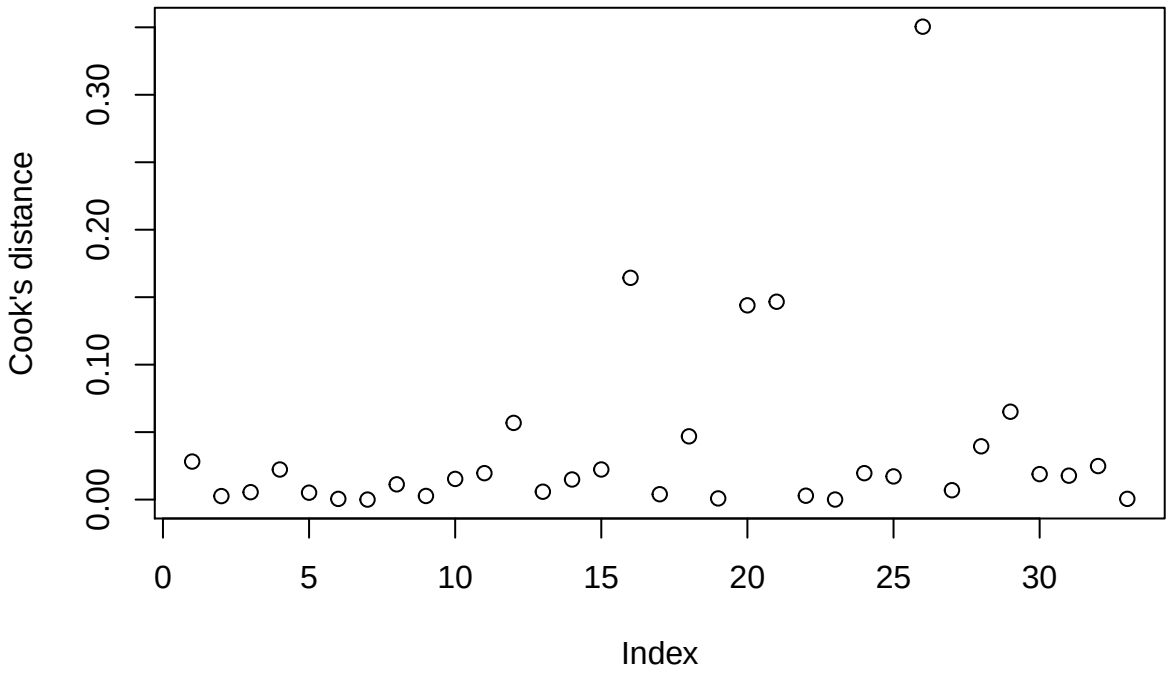
```
creat.stdres=rstandard(creat.lm)
#qqPlot standarized residuals
library(car)
qqPlot(creat.stdres)
```



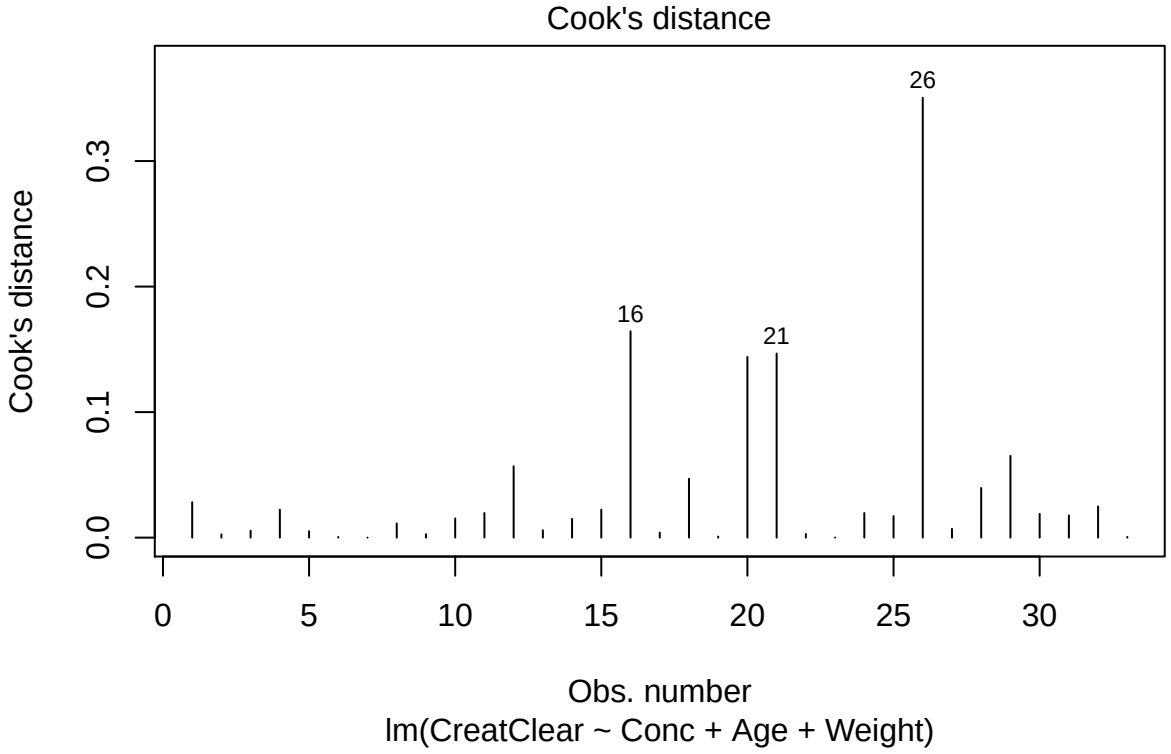

```
# standarized residual plot (standarized residuals vs fitted values)
plot(creat.lm$fitted.values, creat.stdres, ylab="Standarized residuals", xlab="Predicted values", ablin
```




```
# Cook's distance plot
cook_creat = cooks.distance(creat.lm)
plot(cook_creat, ylab = "Cook's distance")
```




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



- The standarized residuals are normal.
- the standarized residual plot shows that the variance is not constant along the predicted values (the plot suggest a nonlinear patter in the relation between the standarized residuals and the predicted values). This result put in jeopardy the constant variance assumpation. 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".

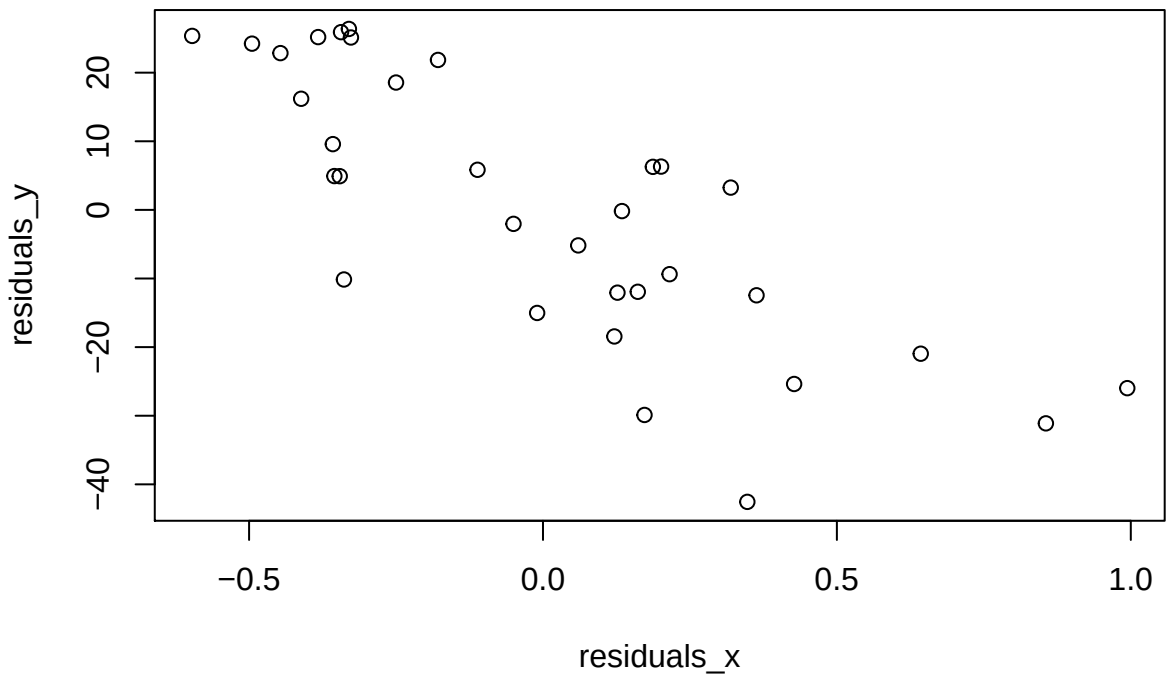
```
# Model fit without Conc as explanatory variables
creat1.lm=lm(CreatClear~Age + Weight, data=data_creati)
summary(creat1.lm)
```

```
##
## Call:
## lm(formula = CreatClear ~ Age + Weight, data = data_creati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.553 -12.451   3.248  18.560  26.368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  84.8316    22.7191   3.734 0.000789 ***
## Age          -1.2176     0.2028  -6.004 1.38e-06 ***
## Weight        0.9447     0.2778   3.400 0.001921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.32 on 30 degrees of freedom
## Multiple R-squared:  0.6002, Adjusted R-squared:  0.5735
## F-statistic: 22.52 on 2 and 30 DF,  p-value: 1.067e-06
```

```
residuals_y = creat1.lm$residuals
# Model fit for Conc=bo + b1 Age + b2 Weight
creat2.lm=lm(Conc~Age + Weight, data=data_creati)
summary(creat2.lm)
```

```
##
## Call:
## lm(formula = Conc ~ Age + Weight, data = data_creati)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59681 -0.34357 -0.00993  0.20094  0.99404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.881731   0.453966   1.942  0.06154 .
## Age          0.012040   0.004052   2.971  0.00579 **
## Weight      -0.004212   0.005551  -0.759  0.45386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4061 on 30 degrees of freedom
## Multiple R-squared:  0.2335, Adjusted R-squared:  0.1824
## F-statistic: 4.569 on 2 and 30 DF,  p-value: 0.01853
```

```
residuals_x = creat2.lm$residuals
# Partial regression plot for Conc
plot(residuals_x,residuals_y )
```



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?

```
creat_residuals.lm = lm(residuals_y ~ residuals_x)
summary(creat_residuals.lm)
```

```
##
## Call:
## lm(formula = residuals_y ~ residuals_x)
##
## 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
## residuals_x -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
```

The estimated equation is:

$$E(Residuals_{reg1}|Residuals_{reg2_i}) = -0.00000 - 39.94 Residuals_{reg2_i}$$

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 coefficients estimated are the same.

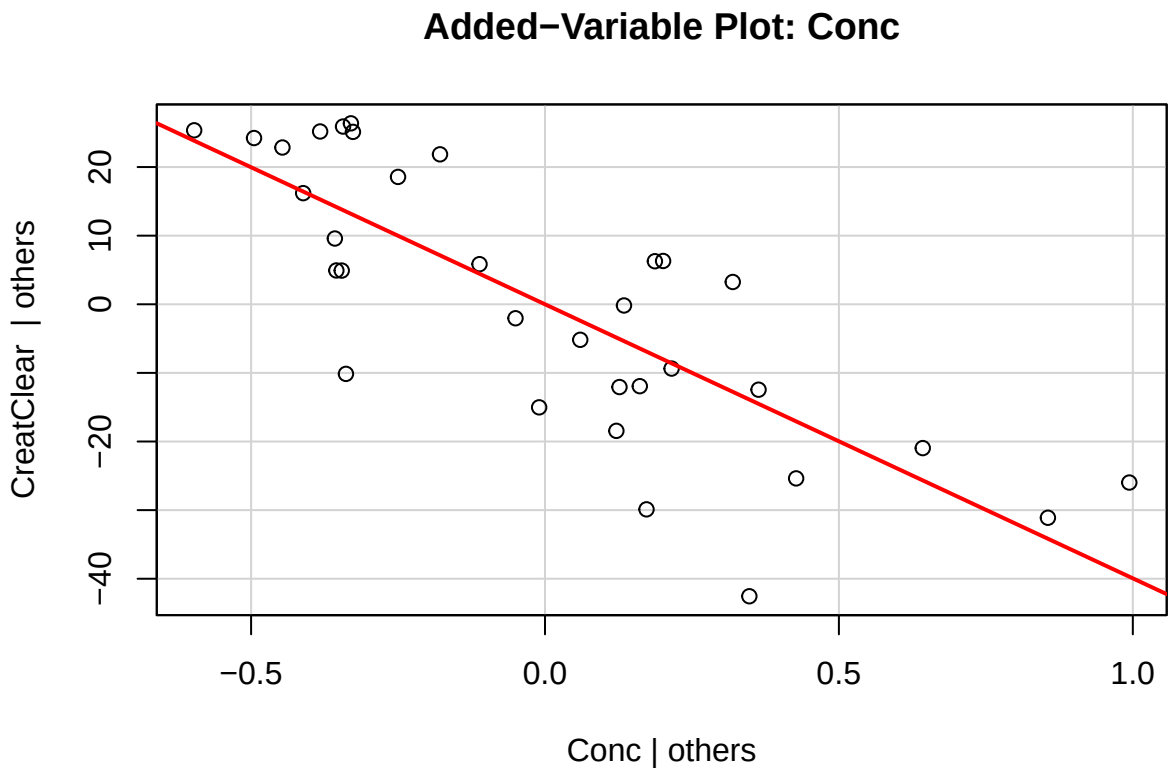
F) Explain why your findings in Part E above make perfect sense from an intuitive perspective.

Answer: The residuals of the regression 1 ($CreatClear_i = \beta_0 + \beta_2 Age_i + \beta_3 Weight_i + \epsilon_i$) represent the part of creatine clearance that is not explained by Age or Weight. On the other hand, the residuals from the regression 2 ($Conc_i = \beta_0 + \beta_1 Age_i + \beta_2 Weight + \epsilon_i$) represent the part ot patient's creatine concentration that is not explained by age or weight. Then, if we fit a model with residuals from regression 1 as response variable and residuals from regression 2 as explanatory variable ($Residuals_{reg1} = \beta_0 + \beta_1 Residuals_{reg2} + \epsilon_i$), we basically are trying to determine what is the effect to add patient's creatine concentration to a model that already includes age and weight as explanatory variables for patient's creatine clearance. So, it makes sense that the slope of this regression is equal to the slope of creatine concentration in the full model ($CreatClear_i = \beta_0 + \beta_1 Conc_i + \beta_2 Age_i + \beta_3 Weight_i + \epsilon_i$). In other words, the slope for the residuals regression ($Residuals_{reg1} = \beta_0 + \beta_1 Residuals_{reg2} + \epsilon_i$) β_1 is the change in patient's creatine clearance for a unit change in patient's creatine concentration, adjusted for fitting a model with Age and weight first.

Feedback: 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.

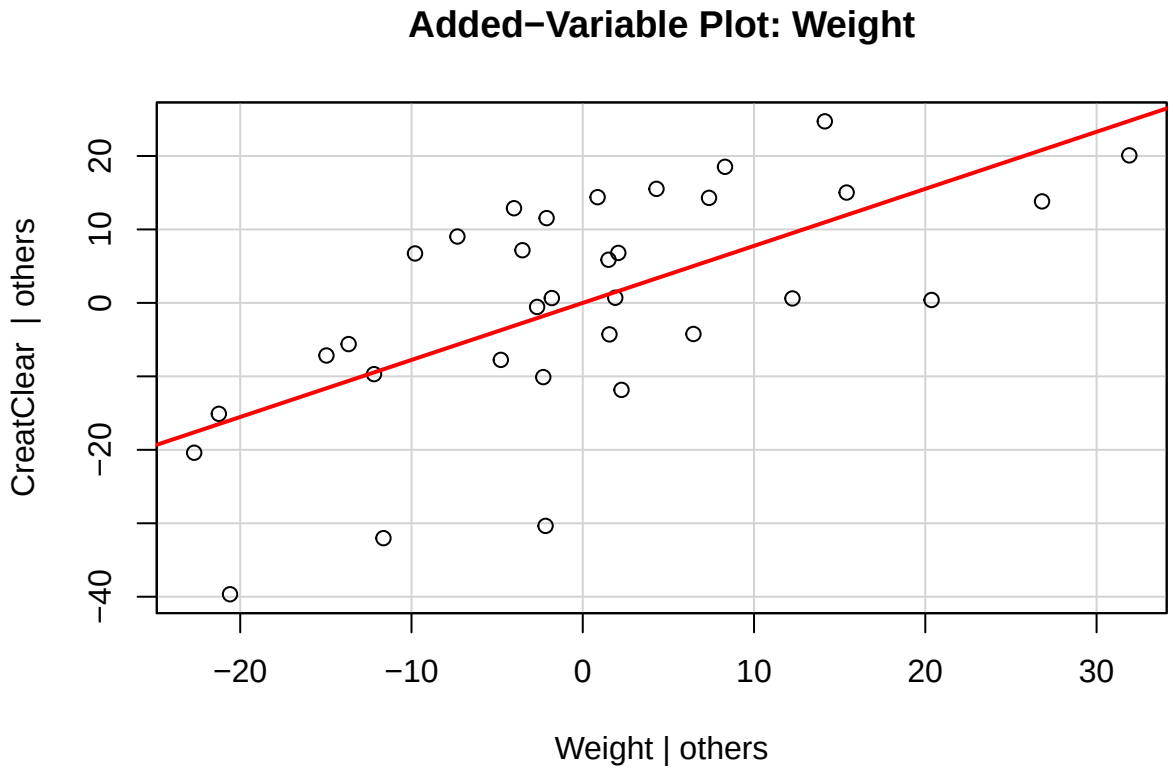
G) Use the "library(car)" and avPlots(LinearModelName) to get the partial regression plots for all three variables. Do you see any curvilinearity in any of these plots?

```
library(car)
avPlot(creat.lm, variable="Conc")
```




```
avPlot(creat.lm, variable="Age")
```


```
avPlot(creat.lm, variable="Weight")
```



Yes, the partial regression plot for Age seems to be curvilinear (cubic). Then, to improve the fit it would be reasonable to include a quadratic and cubic term for Age.

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?

Since the Partial regression plot for Age suggest a cubic fit, we are going to run a new regression overfitting, that is including as explanatory variables the square, cube and quartic of Age:

```
creat3.lm=lm(CreatClear~Conc + Weight + Age  + I(Age^2) + I(Age^3) + I(Age^4), data=data_creati)
summary(creat3.lm)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Weight + Age + I(Age^2) + I(Age^3) +
##     I(Age^4), data = data_creati)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.300  -6.731   3.404   6.724  20.082
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.177e+02  1.066e+02   1.105    0.279
## Conc        -4.279e+01  6.490e+00  -6.593 5.44e-07 ***
## Weight       8.387e-01  1.800e-01   4.659 8.28e-05 ***
## Age         -1.015e+00  1.161e+01  -0.087    0.931
## I(Age^2)     3.615e-02  4.171e-01   0.087    0.932
## I(Age^3)    -1.022e-03  6.108e-03  -0.167    0.868
## I(Age^4)     8.054e-06  3.134e-05   0.257    0.799
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.49 on 26 degrees of freedom
## Multiple R-squared:  0.8691, Adjusted R-squared:  0.8389
## F-statistic: 28.77 on 6 and 26 DF,  p-value: 2.684e-10
```

```
anova(creat3.lm)
```

```
## Analysis of Variance Table
##
## Response: CreatClear
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Conc       1 19927.0 19927.0 127.6995 1.568e-11 ***
## Weight     1  2355.3  2355.3  15.0937 0.0006305 ***
## Age        1  4213.2  4213.2  26.9996 2.004e-05 ***
## I(Age^2)   1   150.2   150.2   0.9623 0.3356373
## I(Age^3)   1   282.3   282.3   1.8090 0.1902383
## I(Age^4)   1    10.3    10.3   0.0660 0.7992163
## Residuals 26  4057.2   156.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the highest power is not significant, so we run a model with the square and cube of Age:

```
creat4.lm=lm(CreatClear~Conc + Weight + Age  + I(Age^2) + I(Age^3), data=data_creati)
summary(creat4.lm)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Weight + Age + I(Age^2) + I(Age^3),
##     data = data_creati)
##
## 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 ***
## Weight       8.280e-01  1.721e-01   4.811 5.06e-05 ***
## Age          1.885e+00  2.684e+00   0.702   0.4885
## 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
```

```
anova(creat4.lm)
```

```
## Analysis of Variance Table
##
## Response: CreatClear
```

```
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Conc       1 19927.0 19927.0 132.2751 6.516e-12 ***
## Weight     1  2355.3  2355.3  15.6345 0.0004997 ***
## Age        1  4213.2  4213.2  27.9670 1.404e-05 ***
## I(Age^2)   1   150.2   150.2   0.9968 0.3269438
## I(Age^3)   1   282.3   282.3   1.8739 0.1823147
## Residuals 27  4067.5   150.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The cube power is also not significant, then we run the regression with only the square power:

```
creat5.lm=lm(CreatClear~Conc + Weight + Age  + I(Age^2), data=data_creati)
summary(creat5.lm)
```

```
##
## Call:
## lm(formula = CreatClear ~ Conc + Weight + Age + I(Age^2), data = data_creati)
##
## 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 ***
## Weight        0.796592   0.173203   4.599 8.29e-05 ***
## Age          -1.591883   0.881169  -1.807   0.0816 .
## 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
```

```
anova(creat5.lm)
```

```
## Analysis of Variance Table
##
## Response: CreatClear
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## Conc       1 19927.0 19927.0 128.2718 5.756e-12 ***
## Weight     1  2355.3  2355.3  15.1614 0.0005583 ***
## Age        1  4213.2  4213.2  27.1206 1.572e-05 ***
## I(Age^2)   1   150.2   150.2   0.9667 0.3339348
## Residuals 28  4349.8   155.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The square power is also not significant. Then, the model without polynomial effects seems to be correct. Hence, we conclude curvilinearity is not an issue for the variable Age in our initial multiple linear model.

Question 5.

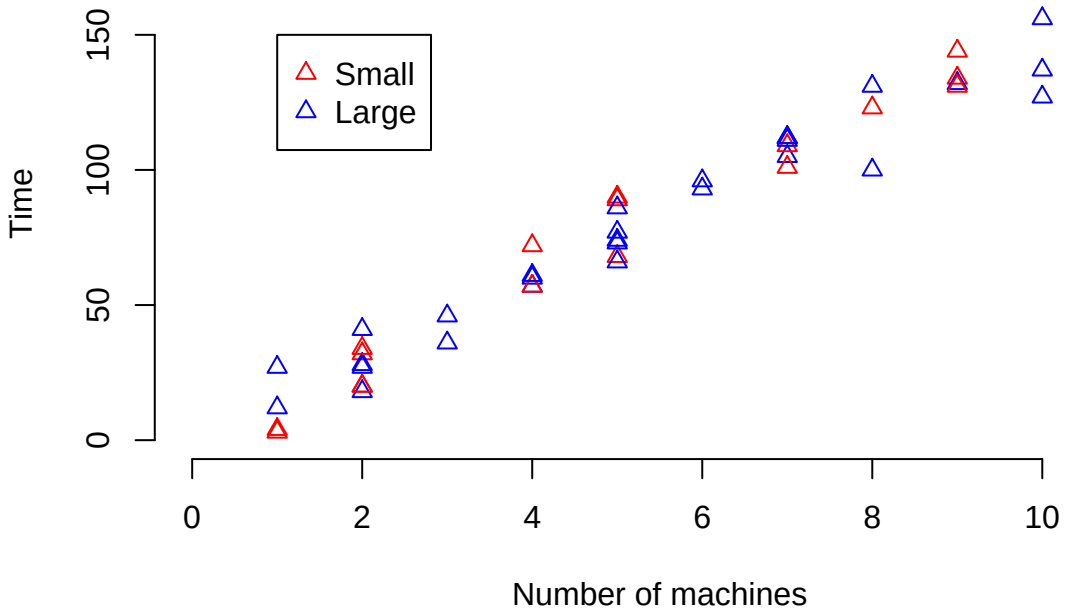
A company services copiers in large businesses and institutions. These businesses and institutions, although they have many copiers, have either the large or small model. A manager wants to relate the time it takes one of his technicians to make a service call on the number of copiers serviced and whether 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?

```
library(readxl)
data_copies = read_excel("Hwk3Q5DatSp17.xlsx")
head(data_copies)
```

```
##   Mins Machs Type
## 1   20     2    S
## 2   60     4    L
## 3   46     3    L
## 4   41     2    L
## 5   12     1    L
## 6  137    10    L
```

```
plot(data_copies$Machs, data_copies$Mins, xlab="Number of machines", ylab="Time",
     xlim=c(0,11), ylim=c(0,175), pch=2, frame.plot=FALSE,
     col=ifelse(data_copies$Type=="S", "red", "blue"), )
     legend(1, 150, pch=c(2,2), col=c("red", "blue"), c("Small", "Large"))
```



The graph suggest that neither the intercept or the slope of the regression would change by type of copy.

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

```
copies.lm=lm(Mins~Machs + Type, data=data_copies)
summary(copies.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs + Type, data = data_copies)
##
## 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
```

The coefficient 0.7585 means that keeping the number of machines constant, the time it takes the technicians to make a service call is 0.76 minutes greater for a small machine compared to a large machine. In other words, we are measuring the change in the intercept when we go from a large machine to a small machine.

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

Here we are interested in test (where β_2 is the coefficient for TypeS):

$$H_0 : \beta_2 = 0$$
$$H_A : \beta_2 \neq 0$$

We have a p-value of 0.786 ($p - value > 0.05$), then we do not reject the null hypothesis. We conclude that the type of machine does no make any difference in the time it takes to the technician to make a service call.

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.

Here we need to run a new regression including the interaction between number of machines and type of machines, our need regression to estimate is:

$$Time = \beta_0 + \beta_1 Machs + \beta_2 Type + \beta_3 Machs * Type + \epsilon$$

In R, we estimate this regression as follows:

```
copies1.lm=lm(Mins~Machs*Type, data=data_copies)
summary(copies1.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs * Type, data = data_copies)
##
## 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 coefficient for the interaction between Type and Machines tell us if the slope of the regression (Machs) increase at the same rate for both types of machines. In other words, it tell us if the change that experiment the time it takes to a technician to make a service call for a unit increase in the number of machines has the same effect for both types of machines. Then we are interested in test:

$$H_0 : \beta_3 = 0$$

$$H_A : \beta_3 \neq 0$$

The p-value for this is test is 0.0755 ($p - value > 0.05$), then we do not reject the null hypothesis. We conclude that amount of service time increase at the same rate for both types of machines. Another way to say it is: 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: $H_0 : \beta_3 = 0$ 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.

Here we are interested in determine if there is any linear effect of type of machine over the time it takes to the technician to make a service call, by any, we mean that the effect could be through the intercept or through the slope, then for the model:

$$Time = \beta_0 + \beta_1 Machs + \beta_2 Type + \beta_3 Machs * Type + \epsilon$$

We are interested to test:

$$H_0 : \beta_2 = \beta_3 = 0$$

$$H_A : At least one of \beta_2 or \beta_3 is not equal to 0$$

We can apply a simultaneous test using the general linear test as follows:

$$F_0 = \frac{\frac{SSREG_{full} - SSREG_{null}}{q}}{\frac{RSS_{full}}{n-k-1}}$$

Where: $SSREG_{full}$ is the regression sum squares of the full regression; $SSREG_{null}$ is the regression sum squares under the null regression ($Time = \beta_0 + \beta_1 Machs + \epsilon$); RSS_{full} is the residuals sum squares for the

full model; k is the number of parameter in the full model; q is the number of variables to be ommited in the null model; and n is the number of observations.

To get all the components of the F test, we are goint to run the full and null models separately as follows:

```
reduce_copies.lm=lm(Mins~Machs, data=data_copies)
summary(reduce_copies.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs, data = data_copies)
##
## 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
```

```
full_copies.lm=lm(Mins~Machs*Type, data=data_copies)
summary(full_copies.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs * Type, data = data_copies)
##
## 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
```

Then we can make the test in R with the following command:

```
anova(reduce_copies.lm,full_copies.lm,test="F")
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Mins ~ Machs
## Model 2: Mins ~ Machs * Type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     43 3416.4
## 2     41 3154.4  2    261.94 1.7023 0.1949
```

The p-value of the test is 0.1949. Since $0.1949 > 0.05$, we fail to reject the null hypothesis that the regression coefficients β_2 and β_3 are both equal to zero.

Another way to make this test in R is as follows:

```
copies1.lm=lm(Mins~Machs*Type, data=data_copies)
summary(copies1.lm)
```

```
##
## Call:
## lm(formula = Mins ~ Machs * Type, data = data_copies)
##
## 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
```

```
anova(copies1.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
```

The F test then should be:

$$F_0 = \frac{\frac{77222-76960}{2}}{\frac{3154}{45-3-1}} = 1.70$$

This value is exactly the same that we obtained calculating the test directly with the command `anova(reduce_copies.lm,full_copies.lm,test="F")`. Then we can calculated the p-value associated with the test above as:

```r
pf(1.7023, 2, 41, lower.tail = FALSE)
```

```
## [1] 0.1948928
```

Exactly the same obtained before. 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 required for his technicians going out on a service call?

Since the results support the idea that there is not linear effect of machines over time separated by type of machine and there is not linear effect of type of machine over time (that is, there is not a linear effect through the slope or through the intercept), then the manager should use the reduced model (without type and type*Machs) to predict the service time required for going out on a service call.