Homework 4

BANA 7052: Applied Linear Regression

Section 1: Wednesday 6:00 – 9:50

Group 5

Tess Newkold

Aabhaas Sethi

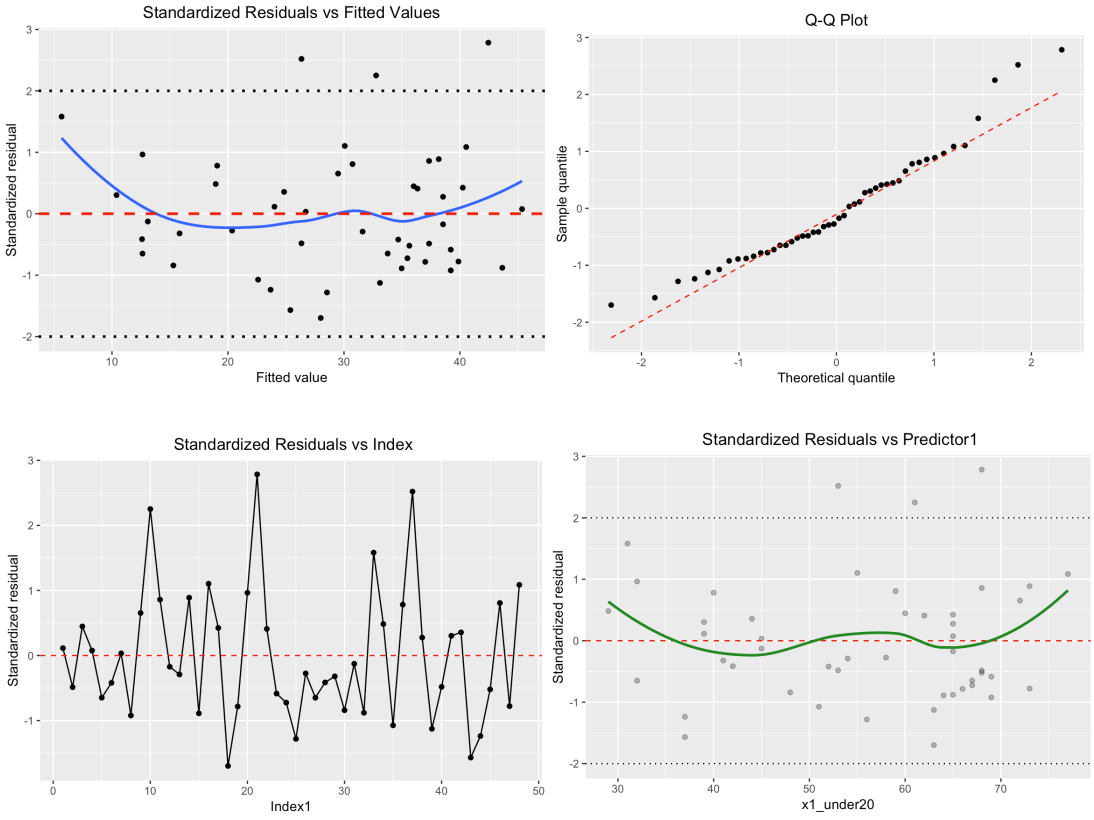
Anirudh Chekuri

James Brand

Question 1

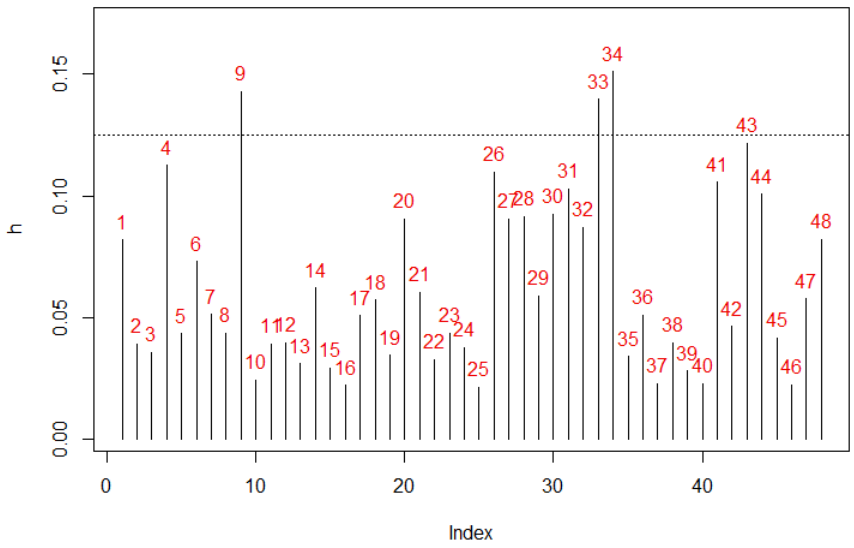
Question 1. (10 points) Alumni Donation Data (Multiple Linear Regression). Continue with the same data from homework 1 (https://bgreenwell.github.io/uc-bana7052/homework/bana7052-hw1) and fit a multiple linear regression model to the data, where the alumni giving rate is the response variable (Y), and the percentage of classes with fewer than 20 students (X_1) and Student/Faculty Ratio (X_2) as the predictors.

Explore various residual diagnostics and possible remedies, including but not limited to:



1. **Does the assumption of error normality appear to be violated?**
   It does not appear to be violated because the points on the QQ plot seem to be following the 45 degree line. However, this is subjective, so you could say they are not close enough to the straight line. The data seem to be skewed to the right as suggested by the upward bending at the end.
2. **Does the assumption of constant error variance appear to be violated?**
   It is not violated, the width for error variance at any area should be more or less the same for the whole plot, which it appears to be.

3. **Does there appear to be any Y or X outliers or influential points?**
   Yes, there appear to be 3 outliers that lie above the dashed line. Additionally, we can see that the 9th, 33rd, and 34th point have hat values > than 2p/n. Hence, these can be considered outliers as well.



4. **Is there any concern about multicollinearity?**
   There is no concern because both variance inflation factors are less than 10. They both are 2.612, which is well below the threshold of 10.

5. **What is the predicted alumni giving rate for an observation with (X_1 = 40, X_2 = 5)? Is there any concern about this prediction? Please explain.**
   Predicted Giving Rate: 37.79. There is concern about this prediction because the hat value for this point (0.3412287) is more than the maximum hat value of our original points. So, this is an influential point.

Question 2

- **Simulation Study (Simple Linear Regression). Assume mean function E(Y|X)=10+5X−2X^2. Generate data with X1∼N(μ=3,σ=0.5)X1∼N(μ=3,σ=0.5), sample size n=100, and error term ϵ∼N(μ=0,σ=0.5).**

```
set.seed(7052)

x1 <- rnorm(100,3,0.5)

x2 <- (x1)^2

y <- rnorm(100,10+5*x1-2*x2,0.5)

m.data <- data.frame(x1,y)

l.model<- lm(y~x1, m.data)

summary(l.model)
```

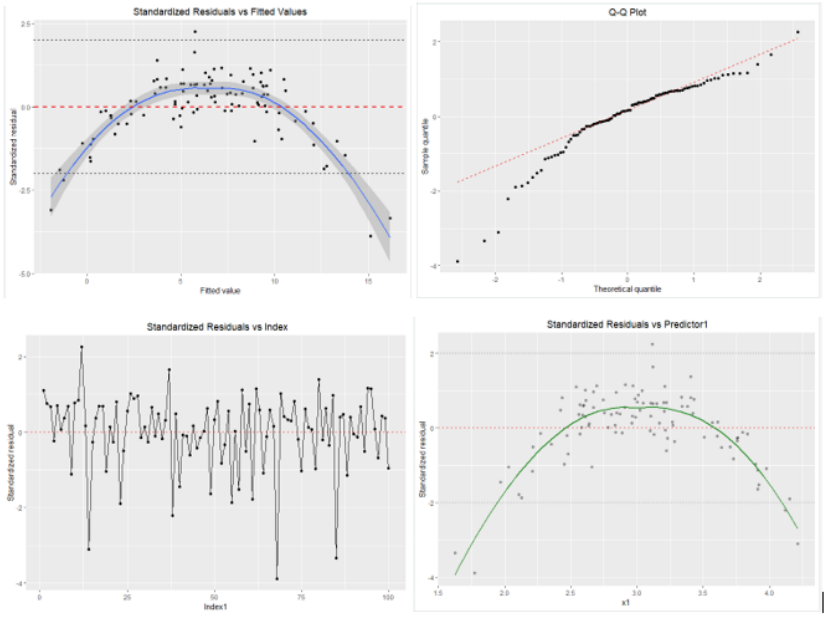
- **Fit a simple linear regression using just XX. What is the estimated regression equation? Please conduct model estimation, inference, and residual diagnostics. What do you conclude?**

n	Error Variance	Intercept	B1	Std. Error (B1)	T value	P value	MSE	Rsq
100	0.5	27.5188	-6.9848	0.1588	-43.99	<2e-16	0.7425235	0.9518
$\hat{Y} = 27.5188 - 6.9848X$								

alpha=0.05 **Hypotheses1: Ho: B1=0 ; H1: B1≠0**

Since the p Value for the hypotheses<0.05, hence, we reject the hypothesis that B1=0 at alpha=0.05 level. So, the predicted value increases by ~45.11 units for every unit increase in the predictor variable B1.

Residual Diagnostics: We can clearly see that the top left graph of stand. residuals vs fitted values is a non-linear plot indicating that there is a non linear term missing in our regression model. There do not seem to be any outliers. The Q-Q plot suggests that the data is left skewed. The stand. Residual vs index plot does not follow any trend and the points seem randomly scattered.



- **Update the model from part b) by adding a quadratic term. Conduct model estimation, inference, and residual diagnostics. What do you conclude? Does this model seem to fit the data better? Please explain.**

n	Error Variance	Intercept	B1	Std. Error (B1)	T value	P value	B2	Std. Error (B2)	T value	P value	MSE	Rsq
100	0.5	10.7438	4.4770	-0.7154	6.258	1.06e-08	-1.8949	0.1175	-16.131	<2e-16	0.2037066	0.9869
$\hat{Y} = 10.7438 + 4.4770X - 1.8949X\text{^}2$												

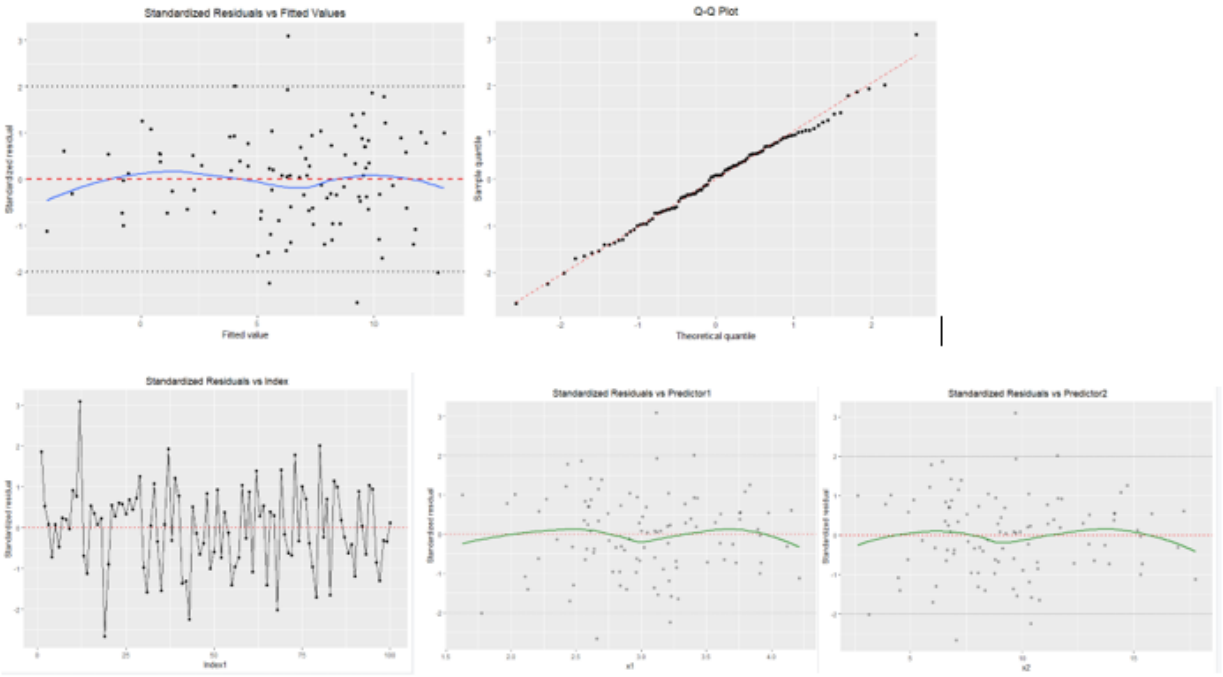
```
m.data <- data.frame(x1,x2,y)

l.model<- lm(y~x1+I(x2), m.data)
```

alpha=0.05 **Hypotheses1: Ho: B1=0 ; H1: B1≠0 and Ho: B2=0 ; H1: B2≠0**

Since the p Value for both the hypotheses<0.05, hence, we reject the hypothesis that B1=0 and B2=0 at alpha=0.05 level. All held constant, the predicted value increases by ~4.47 units for every unit increase in the predictor variable x1. All held constant, the predicted value decreases by ~1.89 units for every unit increase in the predictor variable x2.

Residual Diagnostics: We can see that the top left graph of stand. residuals vs fitted values has improved after including the quadratic term in our model. Even though it is not perfectly linear, but it is way better than the previous case. There do not seem to be any outliers. The Q-Q plot follows the theoretical quantile line at 45 degrees better than the previous plot. The R sq value has also increased from ~95% to 98.6%. **Hence, we can conclude that this model seems to fit data better.**



- **What is the variance inflation factor (VIF) for the quadratic model? Any concern of multicollinearity?**

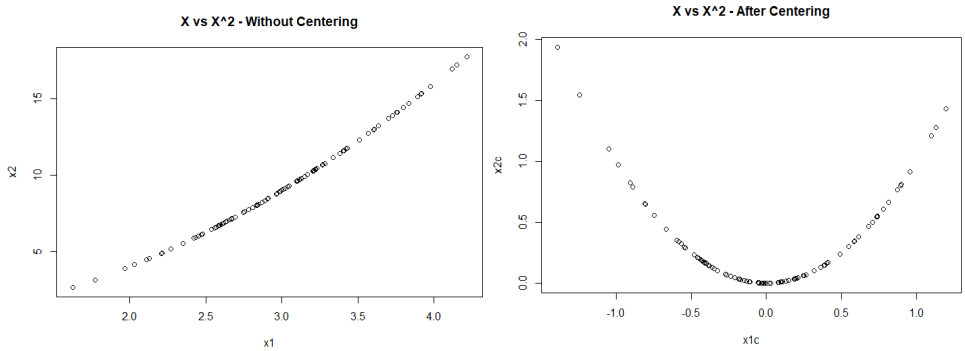
```
vif(l.model)
    x1   I(x2)
73.97863 73.97863
cor(m.data)
          x1        x2         y
x1  1.0000000  0.9932183 -0.9755971
x2  0.9932183  1.0000000 -0.9907689
y  -0.9755971 -0.9907689  1.0000000
```

Clearly, VIF is very high indicating multicollinearity which can be seen through the correlation matrix.

- **Now center the X variable and compare the VIF from d). What did you find? Which VIF is smaller? Please briefly explain the reason.**

```
fit_centered <- lm(y ~ I(x1 - mean(x1)) + I((x1 - mean(x1))^2), data = m.data)
 vif(fit_centered)
   I(x1 - mean(x1)) I((x1 - mean(x1))^2)
          1.000295           1.000295
```

The VIF after centering the X variable is significantly smaller than the VIF obtained in d.



We can see that the plot of x vs x^2 after centering is almost symmetric around the origin. Hence, the positive and negative terms cancel each other in the calculation of correlation. Thus, the VIF is very close to 1 for x and x^2 after centering.

Question 3

3. a.
- There is a linear relationship: We are using the correct model
- The observations are independent: $Cov(\varepsilon_i, \varepsilon_j) = 0, \; i \neq j$
- The errors have a constant variance: $Var(\varepsilon_i) = Cov(\varepsilon_i, \varepsilon_i) = \sigma^2$
- The errors are normally distributed: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

b. In general, the residuals e_i are not independent random variables because they are derived from the fitted $\hat{Y}_i$ which are all based on the same fitted regression model. The error terms ε_i are assumed to be independent, but in general, the residuals e_i are not uncorrelated. As n increases, this effect becomes small.
The error terms $e_i \overset{iid}{\sim} N(0, \sigma^2)$.
The mean of the residuals e_i, $\bar{e} = \frac{\sum e_i}{n} = 0$
The variance of the residuals e_i, $s_e^2 \approx \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = MSE$
Unfortunately, the standard deviation of each e_i is complex and varies for each residual, so we cannot summarize the distribution of the residuals as easily as the error terms.
$Var(e_i) = MSE$ is only an approximation. The variance of each individual e_i is actually $\sigma^2(1-h_{ii})$, which can be estimated with $MSE(1-h_{ii})$

c. We plot the residuals against the fitted values $\hat{Y}_i$ because if our assumptions are met, the residuals and the fitted values are uncorrelated, while the residuals and observed values are correlated.
$Cov(\hat{Y}_i, e_i) = 0$
$Cov(Y_i, e_i) = Cov(\hat{Y}_i + e_i, e_i) = Cov(\hat{Y}_i, e_i) + Cov(e_i, e_i) = \sigma_e^2$
So when we try to check our assumptions with residual plots, we don't want to plot e_i against Y_i since we know they are correlated.

d. Multicollinearity is when the predictor variables are highly correlated. Multicollinearity can cause (X^TX) to become near-singular, which can cause multiple computational issues. Multicollinearity can cause some of the estimated coefficients to become unstable. The standard errors and t-statistics can become unreliable. Multicollinearity can also make it more difficult to interpret the estimated coefficients. If our model contains two predictors that are highly correlated, it can be difficult to distinguish the effect one predictor has on the response from the effect the other predictor has on the response.