Unit 1 Assignment

Sri Seshadri

9/24/2017

Contents

Chapter 1. Question 1

a. Figure 1 shows the sctterplot of GPA vs SAT-Quant. There appears to be a linear relationship between the two variables. However the observations 2 and 5 in the plot appear to be influential points. If the two poitns are removed from the data, the straight line fit would have a steeper slope. Further analysis of the model is in below sections.

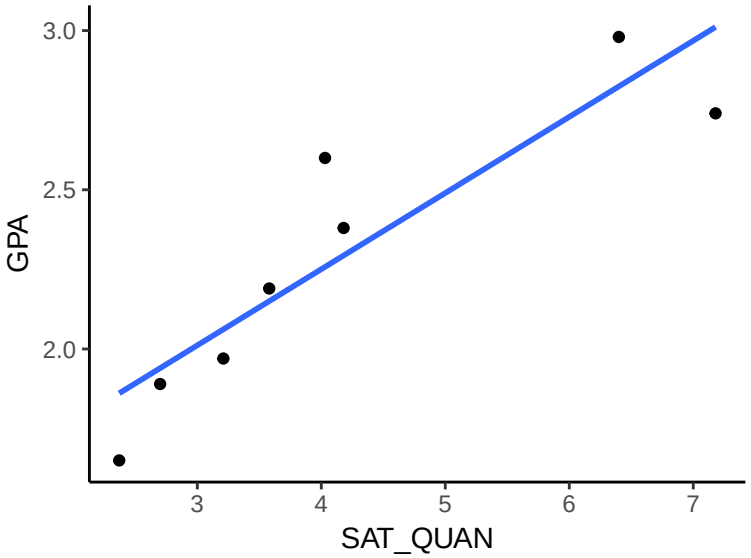


Figure 1: Scatterplot of GPA vs SAT-Quant

b. The GPA is modeled as :

GPA = 1.29 + 0.24 * SAT_QUANT

With Intercept as 1.29 and slope as 0.24

c. The Predicted values and the residuals are shown in the table below.

Table 1: Predicted values and residuals

GPA	SAT_QUAN	.fitted	.resid
1.97	3.21	2.061721	-0.0917210
2.74	7.18	3.011249	-0.2712494
2.19	3.58	2.150216	0.0397839
2.60	4.03	2.257845	0.3421548
2.98	6.40	2.824692	0.1553078
1.65	2.37	1.860813	-0.2108132
1.89	2.70	1.939741	-0.0497413
2.38	4.18	2.293722	0.0862784

The R Squared value is found to be 0.8093426 and SSE = 0.2791225

d. Figure 2 shows the residuals plot against the predicted values of GPA. The residuals appear random.

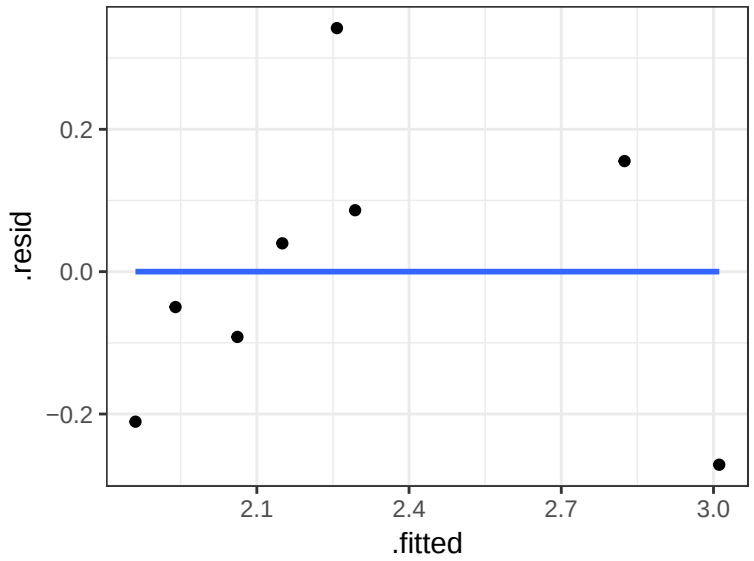


Figure 2: Residuals vs predicted GPA

Chapter 1. Question 2

Figure 3 shows the relationship between quantitative SAT score against GPA for the full data. Observation 2 now doesn't seem as bad as an influential point in the context of the full data.

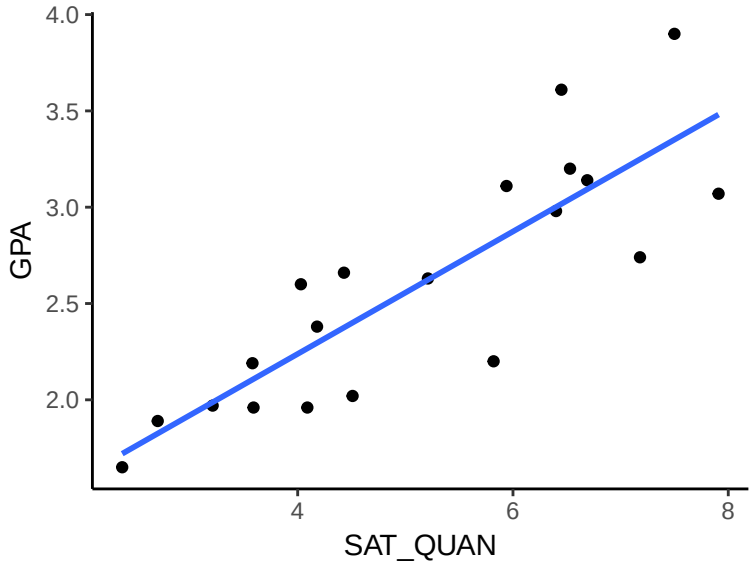


Figure 3: Scatterplot of GPA vs SAT-Quant - Full data

Below is the summary for the model :

GPA = 0.97 + 0.32 * SAT_QUANT

```
##
## Call:
## lm(formula = GPA ~ SAT_QUAN, data = GPA)
##
## Residuals:
```

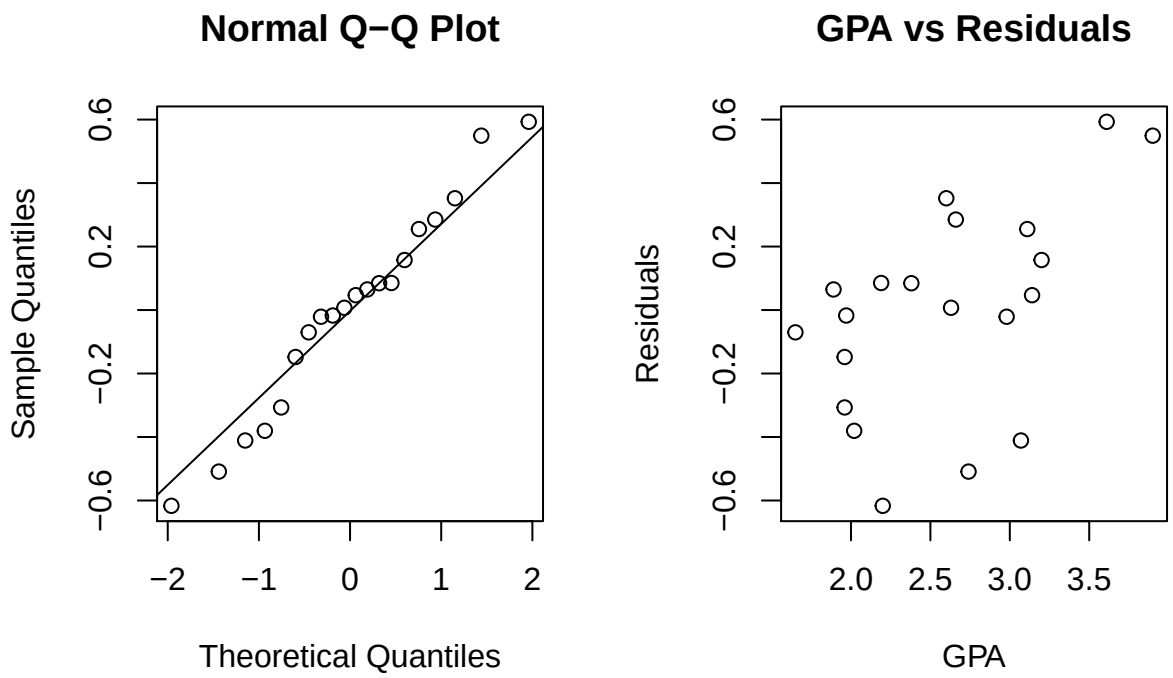


Figure 4: Residuals of model with full data

```
##      Min       1Q   Median       3Q      Max
## -0.61675 -0.18772  0.02693  0.18197  0.59302
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.96699    0.24963   3.874  0.00111 **
## SAT_QUAN     0.31783    0.04652   6.832 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.337 on 18 degrees of freedom
## Multiple R-squared:  0.7217, Adjusted R-squared:  0.7062
## F-statistic: 46.68 on 1 and 18 DF,  p-value: 2.146e-06
```

```
## Analysis of Variance Table
##
## Response: GPA
##           Df Sum Sq Mean Sq F value    Pr(>F)
## SAT_QUAN   1 5.3015  5.3015  46.681 2.146e-06 ***
## Residuals 18 2.0443  0.1136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4 shows the residual analysis of the model above. The residuals are fairly normally distributed and there is a hint of unequal variance in the residuals. May be we are missing another predictor. The next section will explore additional predictors.

Model comparison

The models in question 1 and 2 are cpmpared in the table below. We see that model for question 1 has a better R Squared value and MSE. However, the second model fits the data in question 1 better with much lower MSE. As mentioned above, the residuals for model 2 appear to have non-constant variance. May be another predictor is required.

```
##
## ------------------------------------------------------------------------------
##  Question               model               RSquared     MSE     First_8obs_MSE
## ---------- ------------------------------ ---------- --------- ----------------
##     1       GPA = 1.29 + 0.24 * SAT_QUANT    0.8093    0.03489      0.03489
##
##     2       GPA = 0.97 + 0.32 * SAT_QUANT    0.7217    0.1022      0.0008923
## ------------------------------------------------------------------------------
##
## Table: Model comparison
```

Chapter 1 Question 3

a. Figure 5 shows the relationship between GPA and Highschool English scores. There isn't a strong linear relationship.

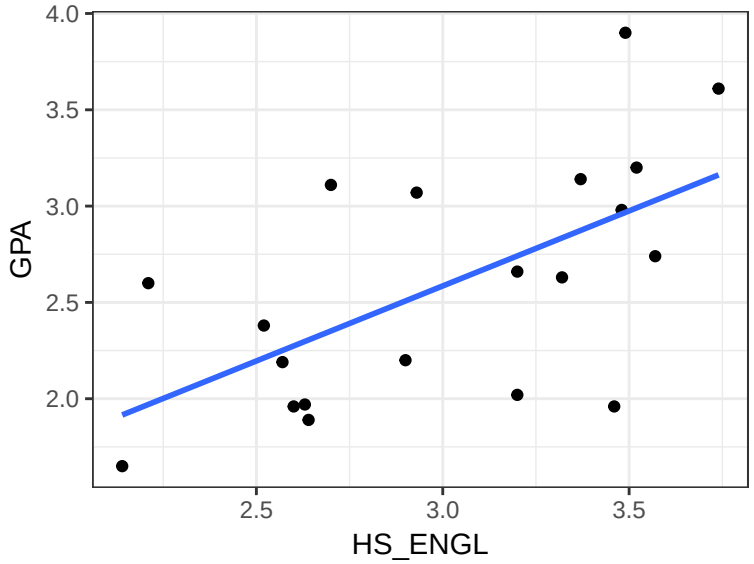


Figure 5: GPA vs HighSchool English

Below is the summary of the model:

GPA = 0.25 + 0.78 * HS_ENGL

```
##
## Call:
## lm(formula = GPA ~ HS_ENGL, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98392 -0.30928 -0.07102  0.31163  0.93271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2487     0.7332   0.339  0.73838
## HS_ENGL       0.7790     0.2407   3.236  0.00458 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5079 on 18 degrees of freedom
## Multiple R-squared:  0.3678, Adjusted R-squared:  0.3327
## F-statistic: 10.47 on 1 and 18 DF,  p-value: 0.004583
```

```
## Analysis of Variance Table
##
## Response: GPA
##           Df Sum Sq Mean Sq F value   Pr(>F)
## HS_ENGL    1 2.7019 2.70193  10.473 0.004583 **
## Residuals 18 4.6439 0.25799
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. Using Highschool english and SAT verbal scores as predictors

It'll be useful to understand the correlations that exist amongst the preditors. To better interpret the model in the context of variability of the regression coefficents (when the predictors are correlated). Figure 6 shows the correalations plot of GPA data. It's seen that the highschool english and verbal SAT scores are not highly correlated.

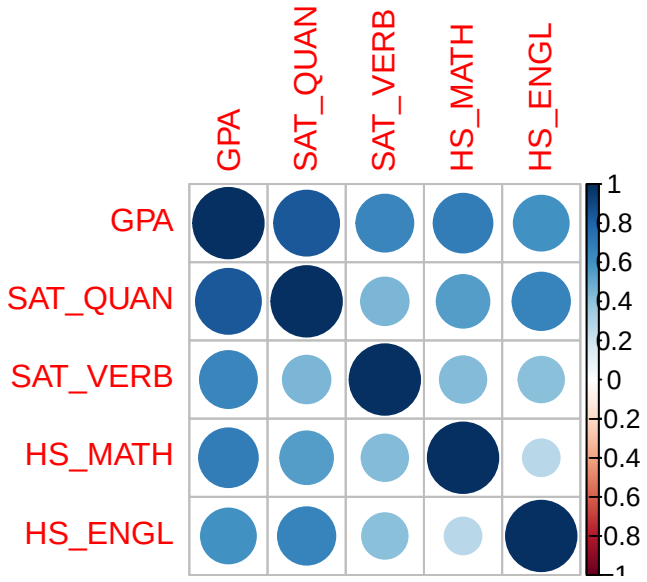


Figure 6: Correlation Plot

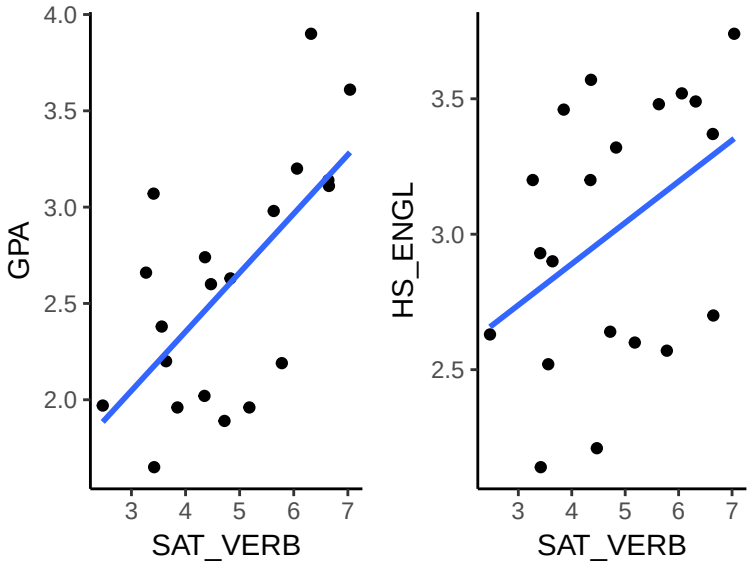


Figure 7: Relationship between predictors

Below is the summary for the model :

GPA = -0.06 + 0.52 * HS_ENGL + 0.23 * SAT_VERB

The model has an adjusted R squared of 51%, which is about half of the variation. Not a good explanatory model.The residuals vs actuals have a non-random relationship. The model might be missing a predictor.

```
##
## Call:
## lm(formula = GPA ~ HS_ENGL + SAT_VERB, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65511 -0.23661 -0.04908  0.26797  0.83021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.05721    0.63789  -0.090   0.9296
## HS_ENGL      0.51947    0.22682   2.290   0.0351 *
## SAT_VERB     0.22726    0.08280   2.745   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4351 on 17 degrees of freedom
## Multiple R-squared:  0.5619, Adjusted R-squared:  0.5104
## F-statistic:  10.9 on 2 and 17 DF,  p-value: 0.0008976
```

```
## Analysis of Variance Table
##
## Response: GPA
##           Df Sum Sq Mean Sq F value   Pr(>F)
## HS_ENGL    1 2.7019 2.70193 14.2739 0.001501 **
## SAT_VERB   1 1.4259 1.42594  7.5331 0.013822 *
## Residuals 17 3.2179 0.18929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

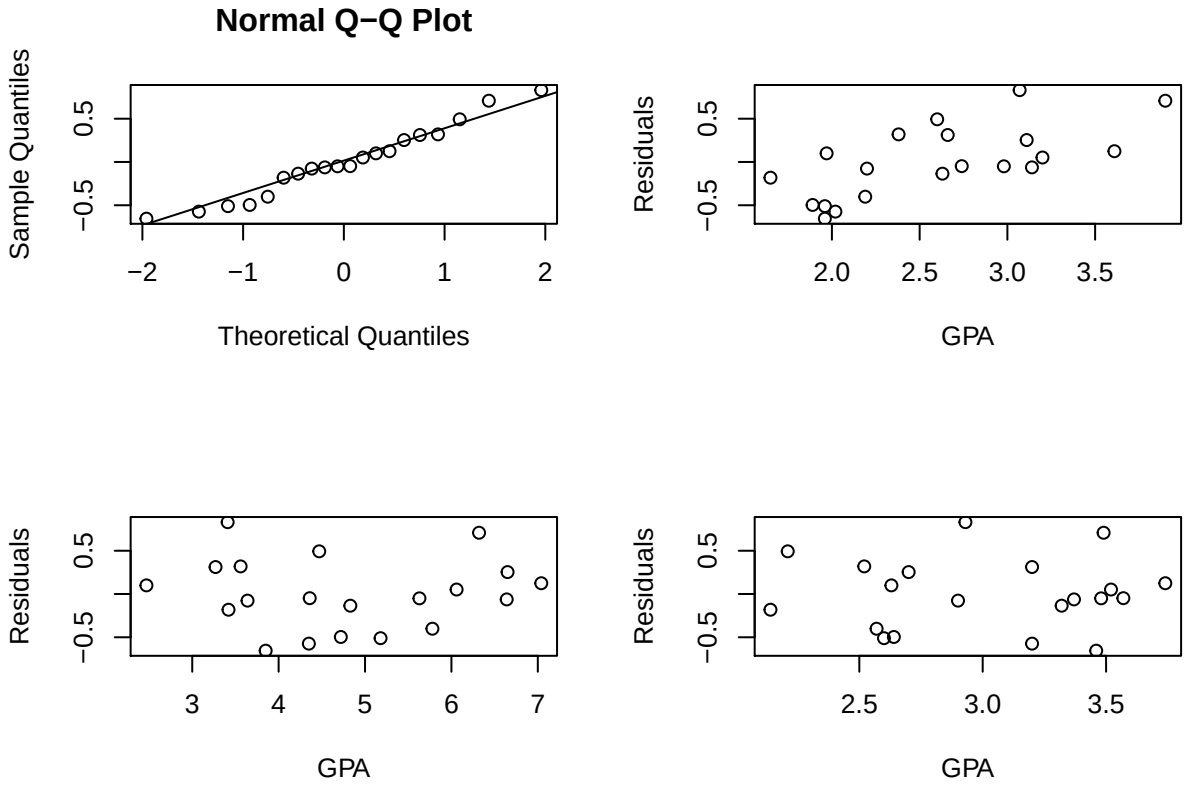


Figure 8: Residual analysis

c. Below is the summary for the model, with addition of SAT_QUAN as predictor to the above model in 3b.

GPA = 0.49 + 0.01 * HS_ENGL + 0.16 * SAT_VERB + 0.26 * SAT_QUAN

It is seen that the residuals are fairly random and normally distributed. The intercept and HS_ENGL scores from the t -test are statistically insignificant. The SAT scores are key contributors to the model.

```
##
## Call:
## lm(formula = GPA ~ HS_ENGL + SAT_VERB + SAT_QUAN, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42543 -0.15587 -0.01946  0.22889  0.47949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48633    0.44777   1.086   0.2935
## HS_ENGL      0.01115    0.18929   0.059   0.9538
## SAT_VERB     0.15683    0.05811   2.699   0.0158 *
## SAT_QUAN     0.25862    0.05631   4.593   0.0003 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2945 on 16 degrees of freedom
## Multiple R-squared:  0.811,  Adjusted R-squared:  0.7756
```

```
## F-statistic: 22.89 on 3 and 16 DF,  p-value: 4.95e-06
```

```
## Analysis of Variance Table
##
## Response: GPA
##           Df Sum Sq Mean Sq F value    Pr(>F)
## HS_ENGL    1 2.7019 2.70193  31.144  4.14e-05 ***
## SAT_VERB   1 1.4259 1.42594  16.436 0.0009210 ***
## SAT_QUAN   1 1.8299 1.82987  21.092 0.0003003 ***
## Residuals 16 1.3881 0.08676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

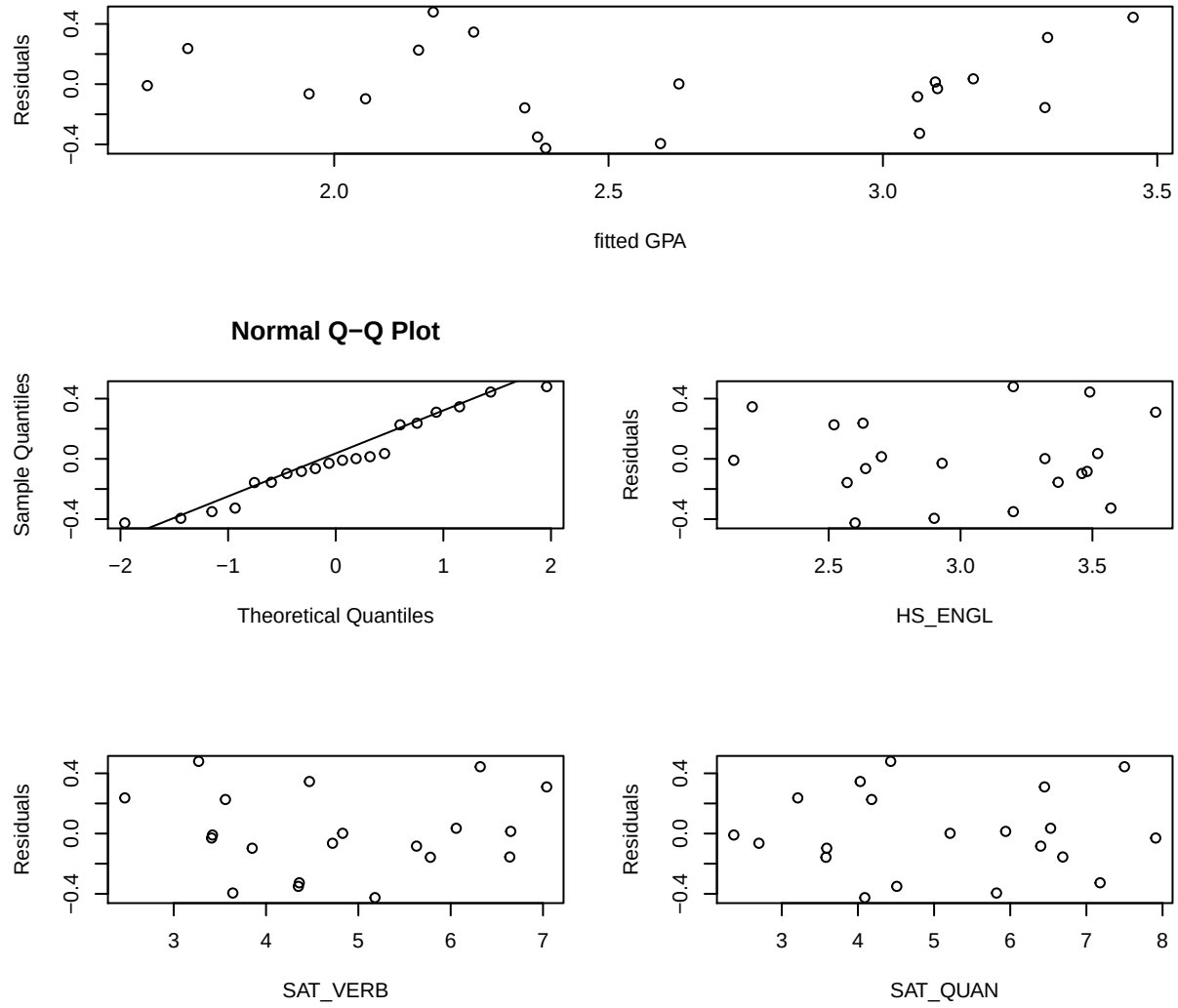


Figure 9: Model 3c residual analysis

Chapter 1 Question 4

a. 78% of the Highschool english score contributes to the overall GPA at the end of 1st year of college.
b. 26% of the SAT quantitative score contributes to the overall GPA at the end of 1st year of college.

Chapter 1 Question 5

The contribution of highschool english score to 1st year college GPA turned out to be negligible after the inclusion of the SAT scores. From test of individual regression coefficients' significance (t-test) in question 3c, it can be seen that the regression coefficient for HS_ENGL is not statistically different from zero. The t-test results is reproduced below:

```
##
## Call:
## lm(formula = GPA ~ HS_ENGL + SAT_VERB + SAT_QUAN, data = GPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42543 -0.15587 -0.01946  0.22889  0.47949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.48633    0.44777   1.086   0.2935
## HS_ENGL      0.01115    0.18929   0.059   0.9538
## SAT_VERB     0.15683    0.05811   2.699   0.0158 *
## SAT_QUAN     0.25862    0.05631   4.593   0.0003 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2945 on 16 degrees of freedom
## Multiple R-squared:  0.811,  Adjusted R-squared:  0.7756
## F-statistic: 22.89 on 3 and 16 DF,  p-value: 4.95e-06
```

Chapter 1 Question 6

a. Figure 9 shows the qqplot of residuals, though the residuals are a bit left leaning/skewed; for practical purposes they seem normal.

b. Figure 9 shows the raw residuals against fitted GPA. The residuals look to have a cyclical pattern. The same plot is reproduced below (Figure 10) with spline. However this may be purely random occurance, more sampling may help to confirm the pattern. The raw residuals can be misleading if there is an influential point that is influencing the regression coefficients to pull the fit closer to itself. Looking at studentized residuals can be helpful; shown in the right panel of figure 10.

The two plots in figure 10 are not different. This indicates that there aren't influential points. It'll be interesting to look at the Cook's distance.

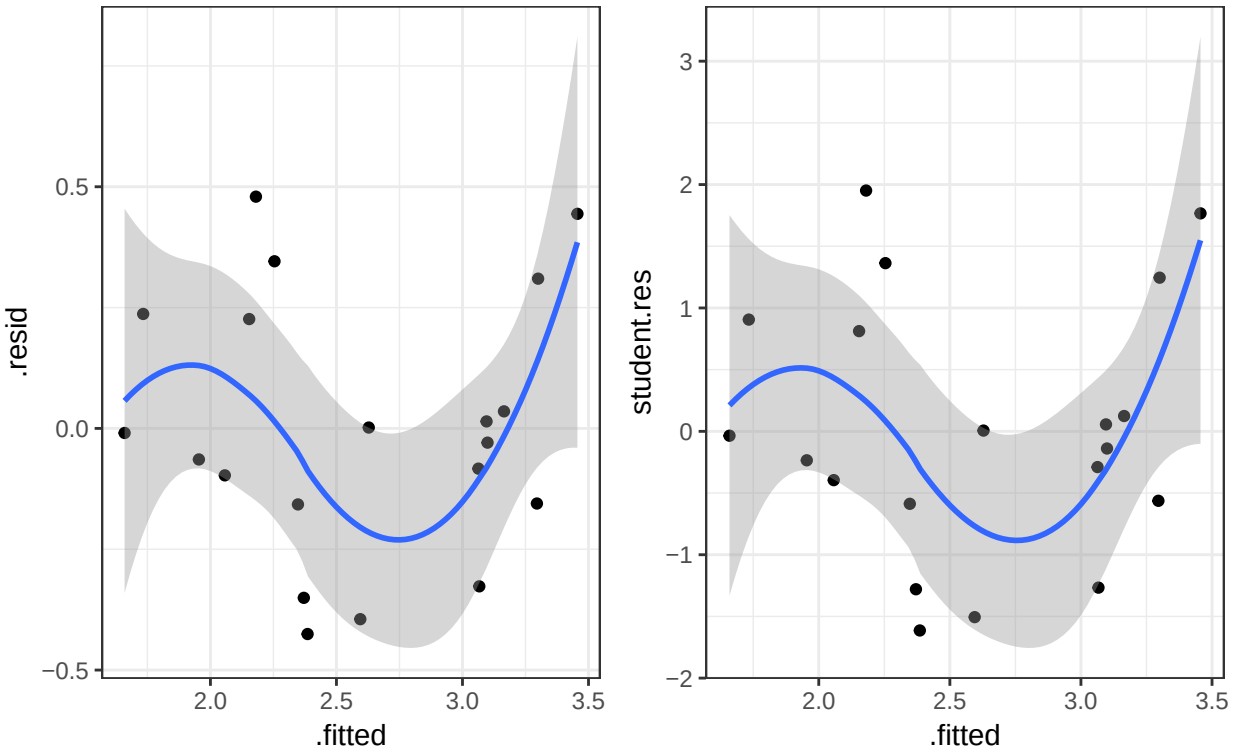
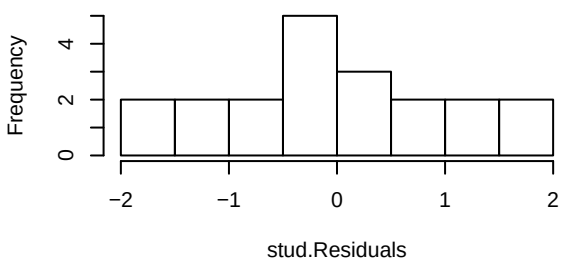
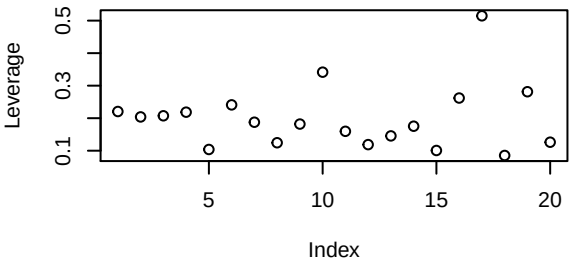


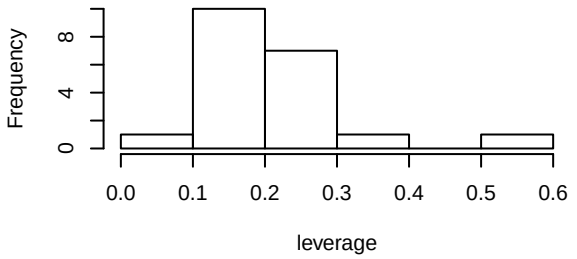
Figure 10: fitted vs residuals

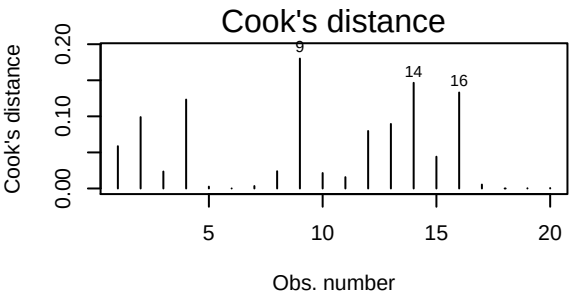
c. The distribution of the studentized residuals, leverage and cook's distance is shown below.

The studentized residuals do not show outliers. The residuals are within +/- 2 standard deviations. Observation 17 has leverage of 0.5 which is greater than 2p/n = 0.4. This shows that this point is outlier in the x space. However the Cook's distance show that there is no effect of beta coefficients. Obervation 17 is more a leverage point and not an influencial point.









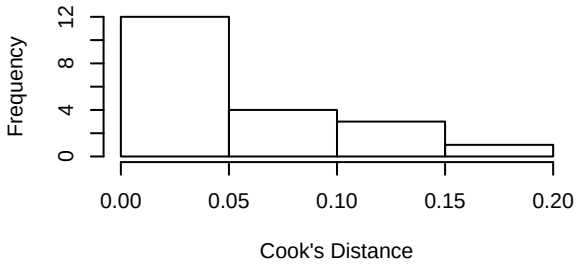


Figure 11: Leverage and Influence analysis

Chapter 2 Question 1

a. Poisson distribution
b. Binomial distribution
c. Normal distribution ?
d. Multinomial
e. Binomial
f. Multinomial
g. negative binomial

Chapter 2 Question 2

a. Poisson distribution - Natural logrithm link
b. Binomial distribution - Logit
c. Normal distribution - Identity
d. Multinomial - generalized logit
e. Binomial - logit
f. Multinomial - generalized logit
g. negative binomial - Logit

Chapter 2 Question 3

Question	What	Expected	Variance
a	Alcohol use	975.00	341.250
b.1	employed full time	975.00	49.500
b.2	employed part time	30.00	25.500
b.3	unemployed	20.00	18.000
b.4	non-participants	40.00	32.000
c.1	Low-Self Esteem	382.50	210.375
c.2	High-Self Esteem	467.50	210.375
d	Accidents per day	0.92	1.050

Chapter 2 Question 4

```
p <- 0.4
n <- 30
r <- 12
NBmean <- r/p
NBVar <- r*(1-p)/p^2

p30 <- choose(n-1,r-1)*p^r*(1-p)^(n-r)
```

The computation formula for mean, variance and probability can be see in the R Code above.

Mean	Variance	Prob.of.Success.by.reaching.30
30	45	0.06

Chapter 2 Question 5

p	probability.when.i.equals.2
0.1	0.10
0.2	0.25
0.3	0.32
0.4	0.31

p = 0.3 is the most likely value of p

Chapter 2 Question 6

```
##
## Call:
## glm(formula = VIOLRATE ~ 1, data = usdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -452.68  -184.62   -26.84   167.65   531.53
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   539.42      38.08   14.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 72487.54)
##
##     Null deviance: 3551890  on 49  degrees of freedom
## Residual deviance: 3551890  on 49  degrees of freedom
## AIC: 704.44
##
## Number of Fisher Scoring iterations: 2
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: VIOLRATE
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    49    3551890
```

```
##
## Call:
## glm(formula = VIOLRATE ~ UNEMPRAT + DENSITY + GSPROD, data = usdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -419.24  -148.27   -25.04   140.42   461.35
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.204e+01  1.380e+02   0.450  0.65506
## UNEMPRAT    7.206e+01  2.626e+01   2.744  0.00862 **
## DENSITY     4.212e-02  1.404e-01   0.300  0.76547
## GSPROD      6.587e-04  2.004e-04   3.286  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 50893.5)
##
##     Null deviance: 3551890  on 49  degrees of freedom
## Residual deviance: 2341101  on 46  degrees of freedom
## AIC: 689.6
##
## Number of Fisher Scoring iterations: 2
```