Unit 1 Assignment

Sri Seshadri

9/24/2017

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.

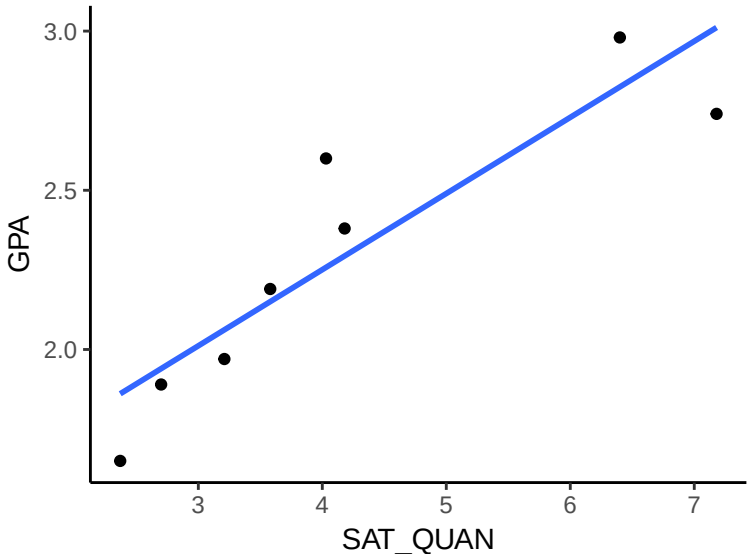

Figure 1: Scatterplot of GPA vs SAT-Quant

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

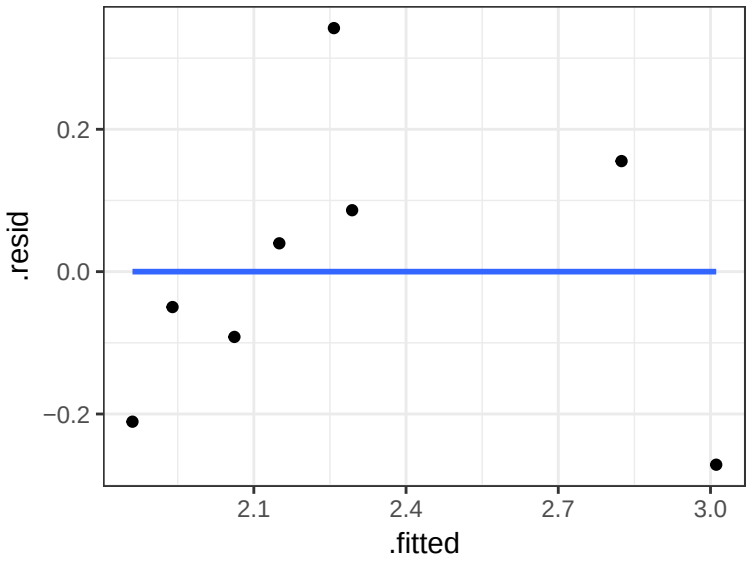

Figure 2: Residuals vs predicted GPA

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.

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.

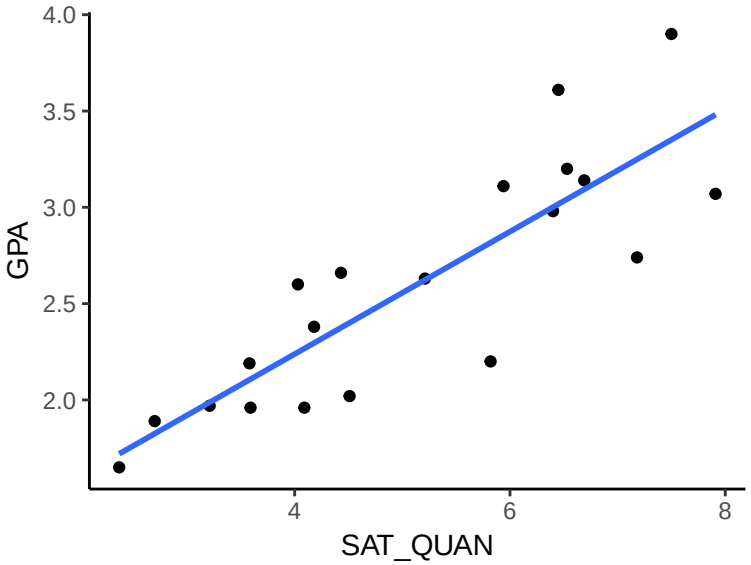

Figure 3: Scatterplot of GPA vs SAT-Quant - Full data

Below is the summary for the model :

GPA = 0.97 + 0.32 * SAT_QUANT

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

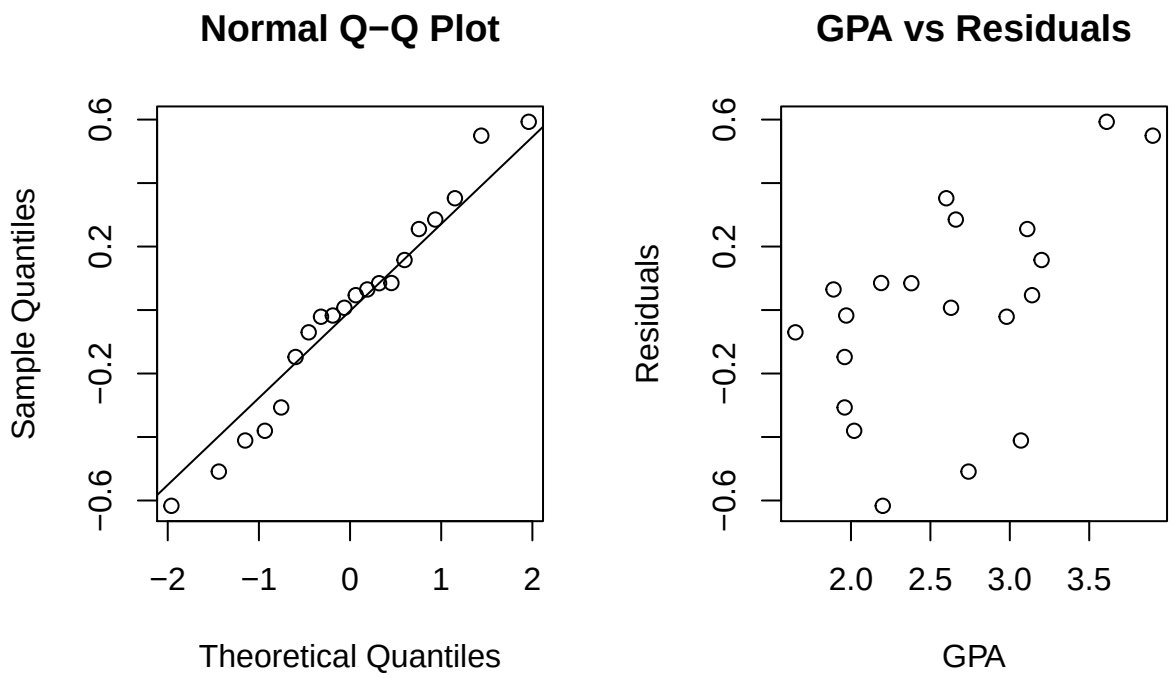

Figure 4: Residuals of model with full data

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.

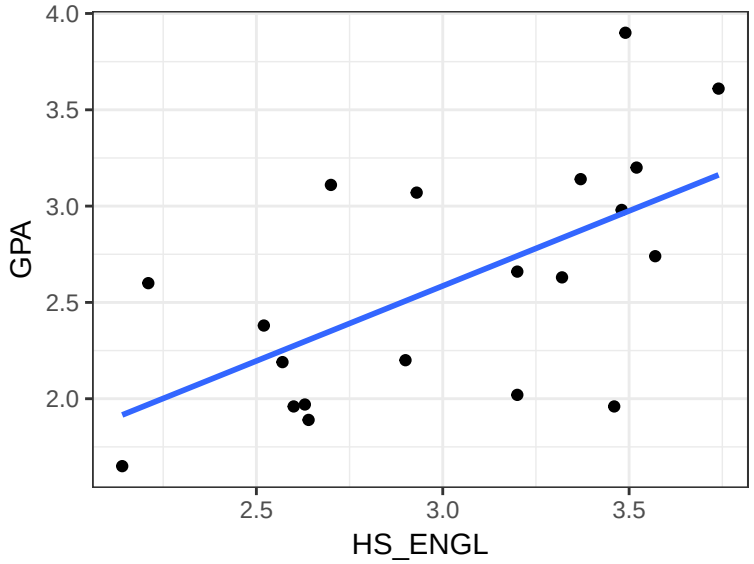

Figure 5: GPA vs HighSchool English

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.

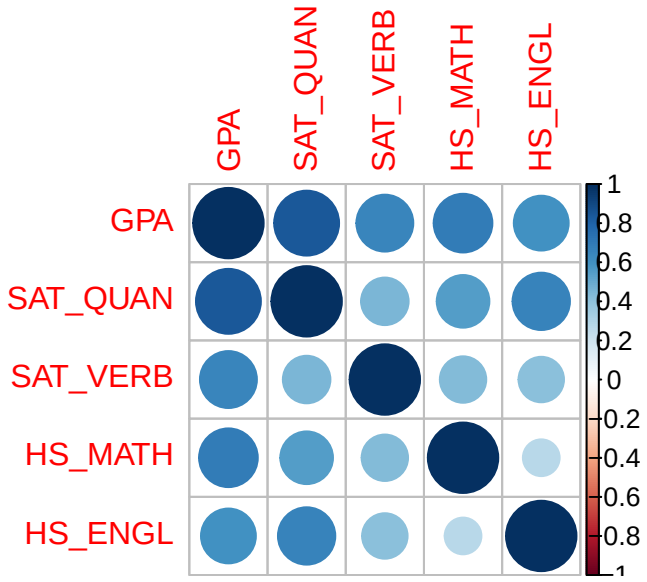

Figure 6: Correlation Plot

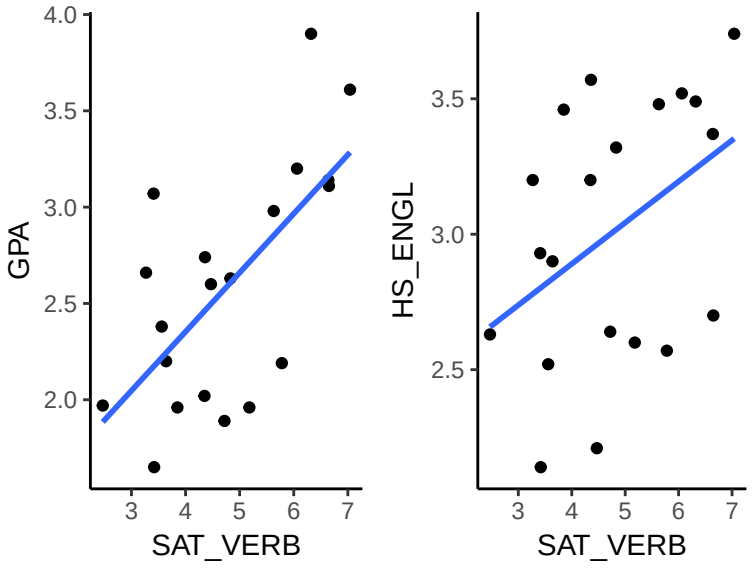

Figure 7: Relationship between predictors

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
```

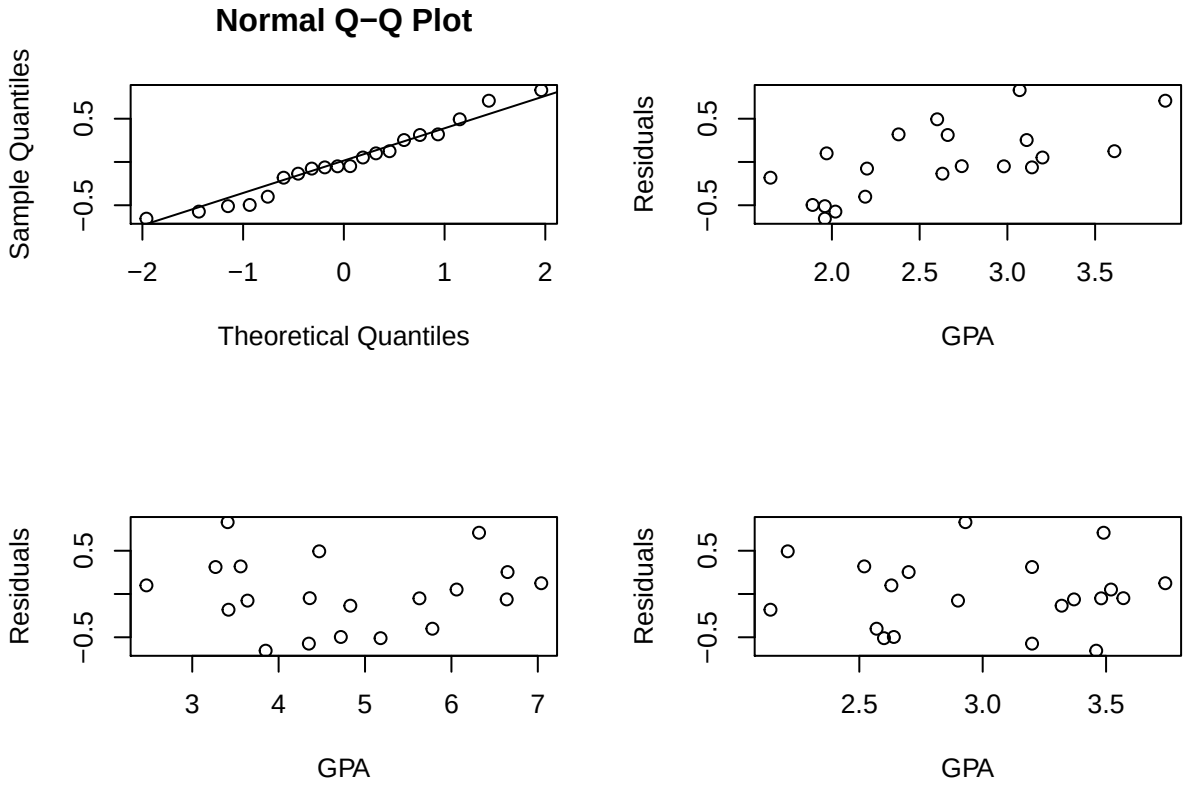

Figure 8: Residual analysis

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

```
##
## 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
```

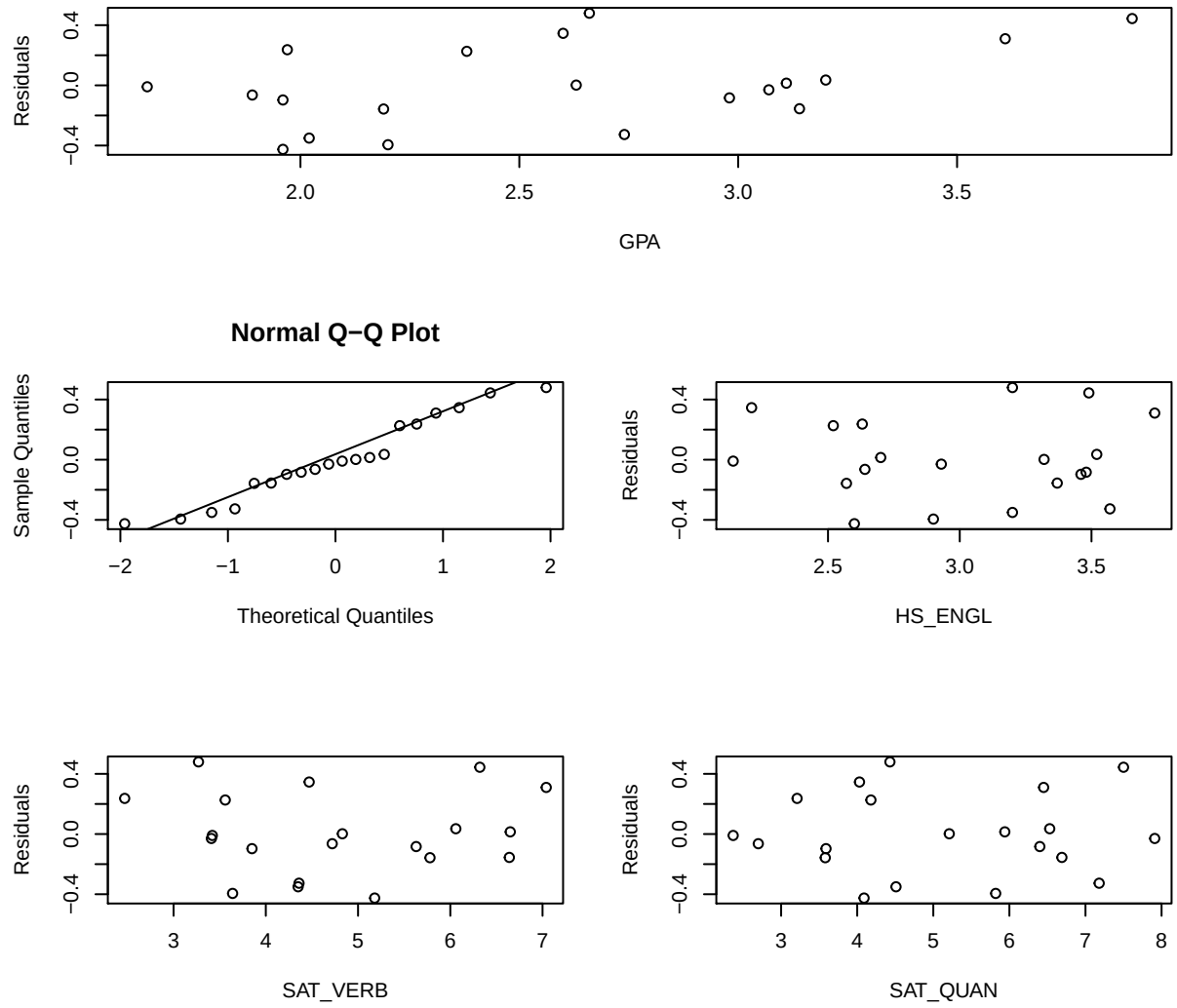

Figure 9: Model 3c residual analysis