Regression Analysis of Education Data
Stat 4043, Fall 2019 – Final Project
Yifei (Yvan) Li

Dataset Description

Education is a test used to detect the average SAT score. By summarizing data from 2005 for each state (51 states in total) on the average SAT score and several other variables thought to have an effect on the success in the SAT exam, this analysis attempts to study the relationship between the average SAT score and several effect socioeconomic variables. It is important for the policy makers to understand the effect socioeconomic variables on education.

The response in this analysis is the total SAT score in each state measured as the average across the state, where the full score is 1600. This analysis will create a regression model that will estimate the total SAT score based on certain socioeconomic measures.

The 7 variables in the dataset and their corresponding types in the regression model are showed as **Table 1.**

Table 1

Variables	Descriptions	Types
STATE	Name of the state	\
SATSCORE	Total SAT score (average across the state)	Response
TAKEPCT	Percent of high school seniors taking the SAT	Predictor
EXPEND	State per capita expenditure on instruction in elementary/secondary schools	Predictor
REDSTATE	Whether the state voted for Republican (1) or Democratic (2) in the 2004 presidential election	Predictor (Category)
POVRATE	Percentage of people below the poverty line, estimate for 2004-2006	Predictor
MEDINCOME	Median household income for 2005 (in dollars)	Predictor

Data Visualization

SAT score measurements were taken for each state in the study, and measurements ranged from 959 to 1227. These values are not evenly dispersed throughout this range, however. It is clear from **Figure 1** that the data for SAT score is non-symmetric without any significant shape. Roughly half of the SAT score locate below 1050 of 1600, and there are just 2 of 51 state have SAT score above 1200 of 1600. **Figure 2** shows that there is no potential outlier in the data for SAT score, given the fact that SAT is a standard academic test with official limited range.

Figure 1

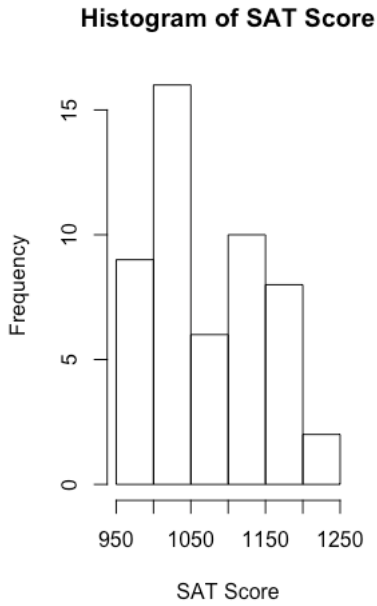
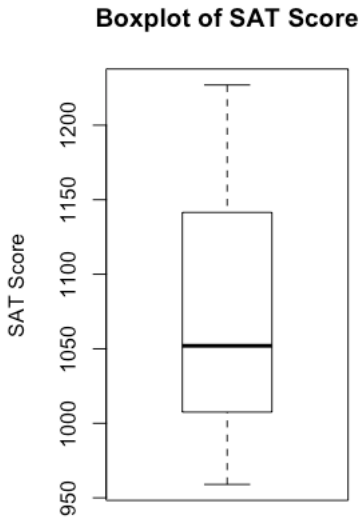
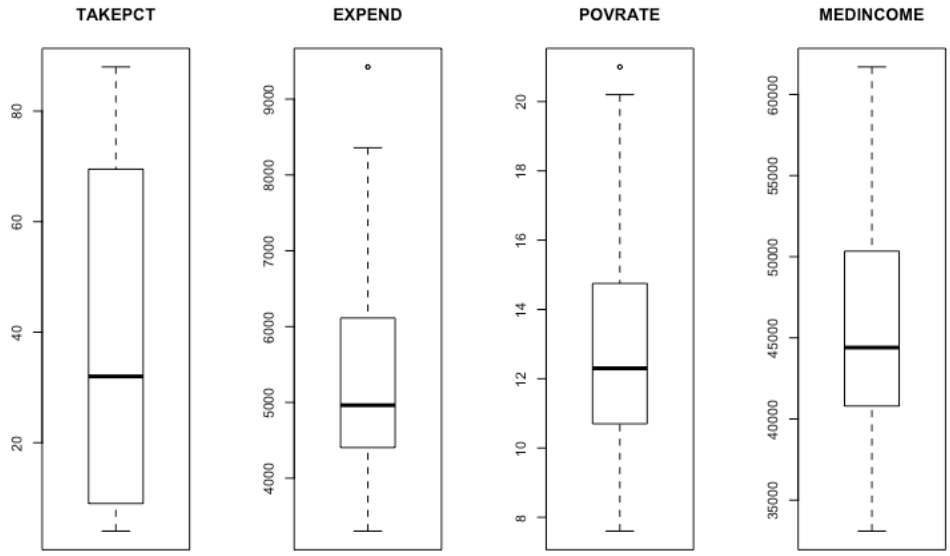


Figure 2



When creating boxplots for all the continuous predictors **(Figure 3)**, it appeared that both EXPEND and POVRATE contained outliners but without great significance. Due to these observations, I anticipated that once my model was fit, even if there is any influential observation, it would not be highly influential.

Figure 3



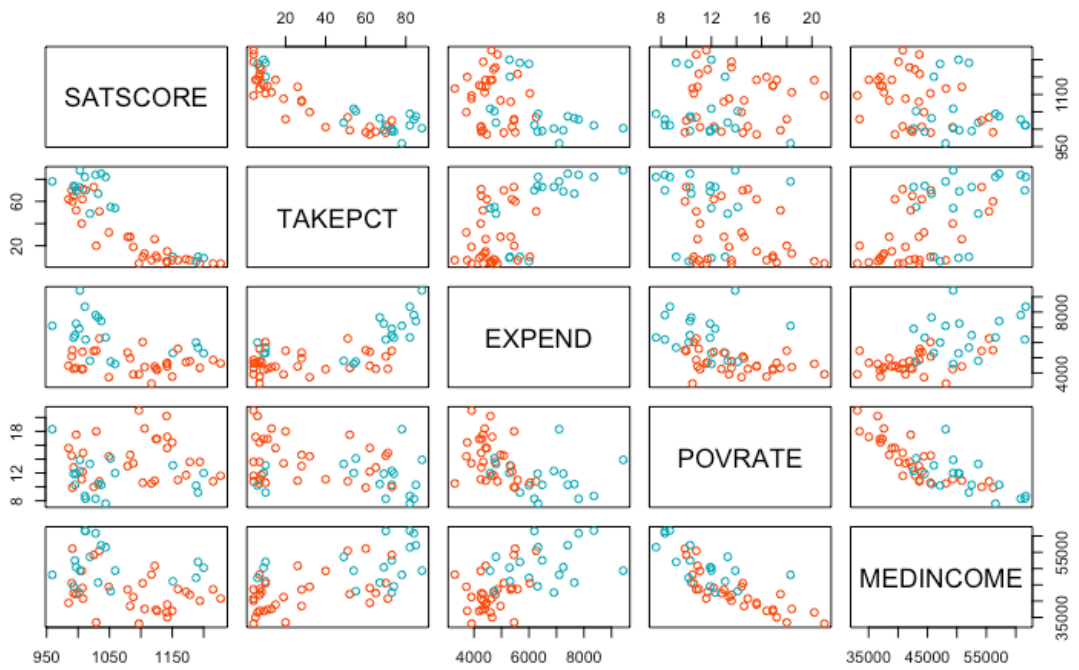
For the categorical predictor REDSTATE, **Table 2** was created to analyze the distribution of observations. These tables indicated that 32 of 51 states are read states (Republican) while others are blue states (Democratic).

Table 2

Red State	1	2
The Number of State	32	19

From the scatterplots of between all pairs of these measures **(Figure 4)**, there is a linear tendency between POVRATE and MEDINCOME, which is reasonable because both are descriptions of the wealth level in each state from different aspects. It seems that there is a curve-shape relationship between SATSCORE and TAKEPCT. There is no any clear relationship among other pairs of measures. It's obvious that comparing to the blue states (Democratic), the red states (Republican) tend to have a lower TAKEPCT, lower EXPEND, higher POVRATE and lower MEDINCOME.

Figure 4



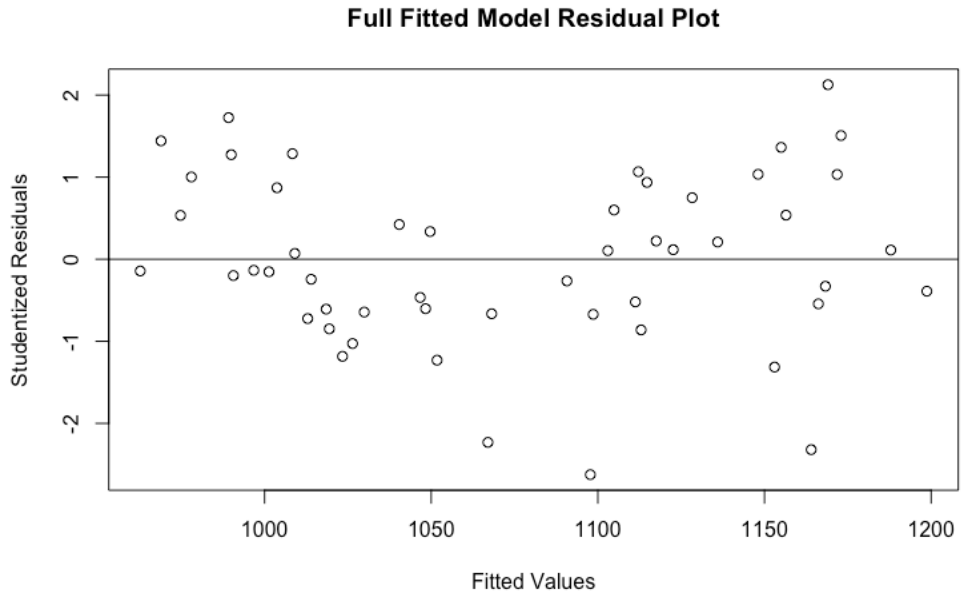
Blue point = Democratic state; Red point = Republican state

Full Model

I began the analysis by fitting a model that contained all five predictors. When I plotted the fitted values vs. studentized residuals **(Figure 5)**, there was no any clear pattern in the residuals. This indicated that the assumptions of linearity and homoskedasticity would be met by the

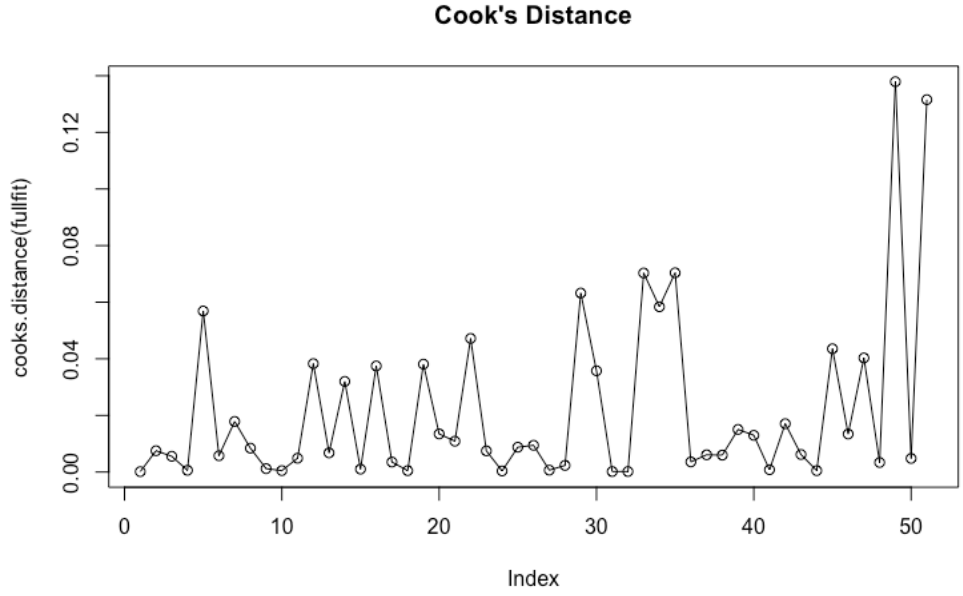
model. Therefore, we don't need any transformation such as log transformation of the response.

Figure 5



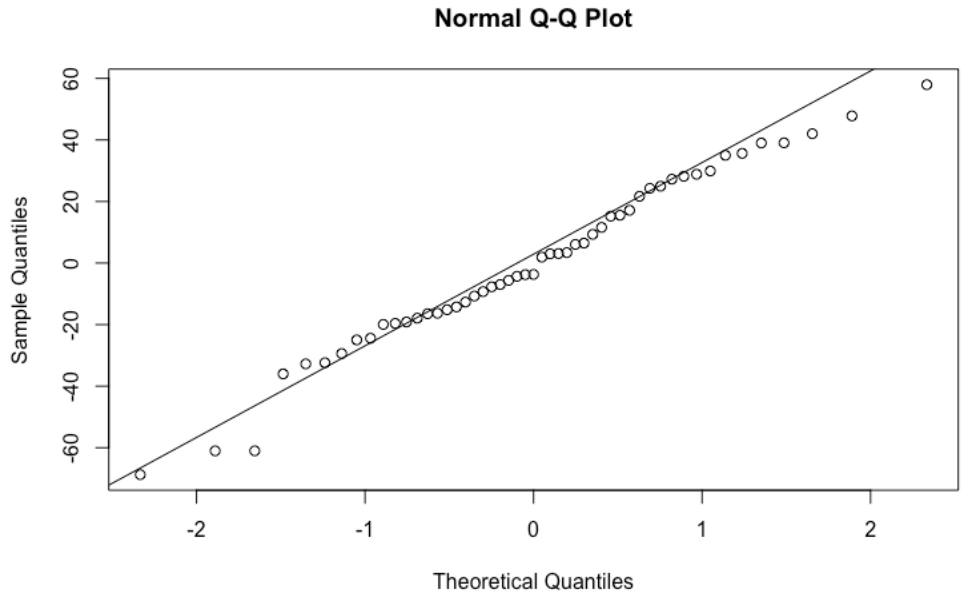
Although the residual plot did not indicate any outliers, Cook's Distance values were plotted to check for influential observations. From **Figure 6**, it is clear that there is no any measure with Cook's Distance greater than 1. That is, there is no any highly influential values.

Figure 6



A QQ plot **(Figure 7)** was created to check for normality. Although there is slight deviation at the left tail end, the correlation between the theoretical quantiles and sample quantiles is 0.9913, indicating that the assumption of normality holds. Finally, VIF values were calculated to test for multicollinearity. Between the VIF values and generalized VIF values used for categorical variables, the highest value was 4.439411, indicating there is no reason to worry about multicollinearity. Given all the assumptions for linear regression hold for the full model, I proceeded to model selection.

Figure 7



Model Selection

I used the AIC stepwise method for model selection, beginning with no predictors and building the model from there. After running the "step" function in R, the best model with only main effects based on AIC is given below:

$$SATSCORE = \beta_0 + \beta_1 TAKEPCT + \beta_2 POVRATE + \beta_3 REDSTATE(2) + \epsilon$$

Using this main effect model as a starting point, the AIC stepwise method was repeated for all possible two-way interactions of these predictors. The best model based on AIC is the same as above.

Final Model Diagnostics

After the model was selected, diagnostics were run to ensure all assumptions were met. The fitted values vs. studentized residuals plot **(Figure 8)** does not show any pattern, indicating that

the assumptions of linearity and homoskedasticity are met. The residual plots for each individual main effect predictor **(Figure 9)** also look good.

Figure 8

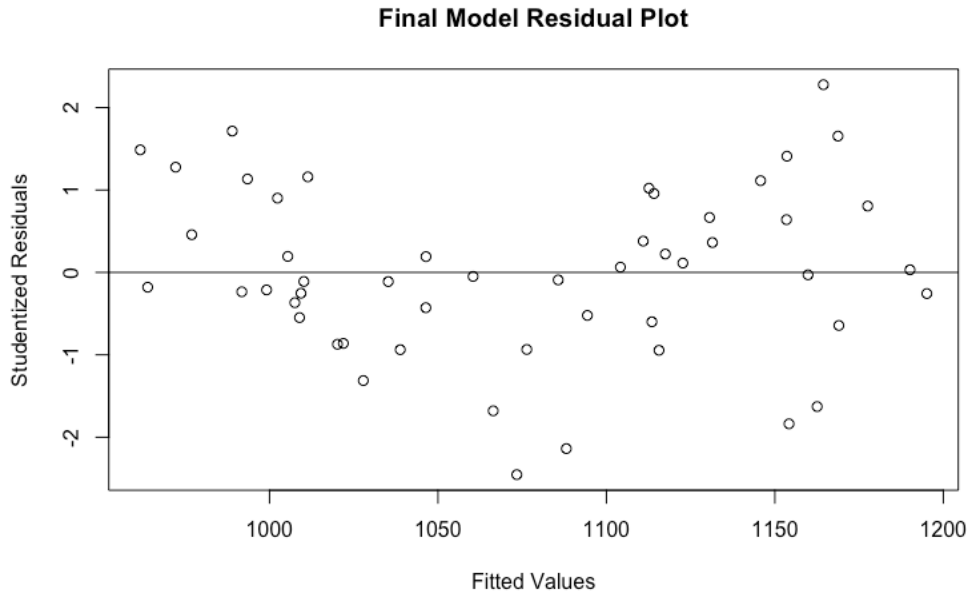
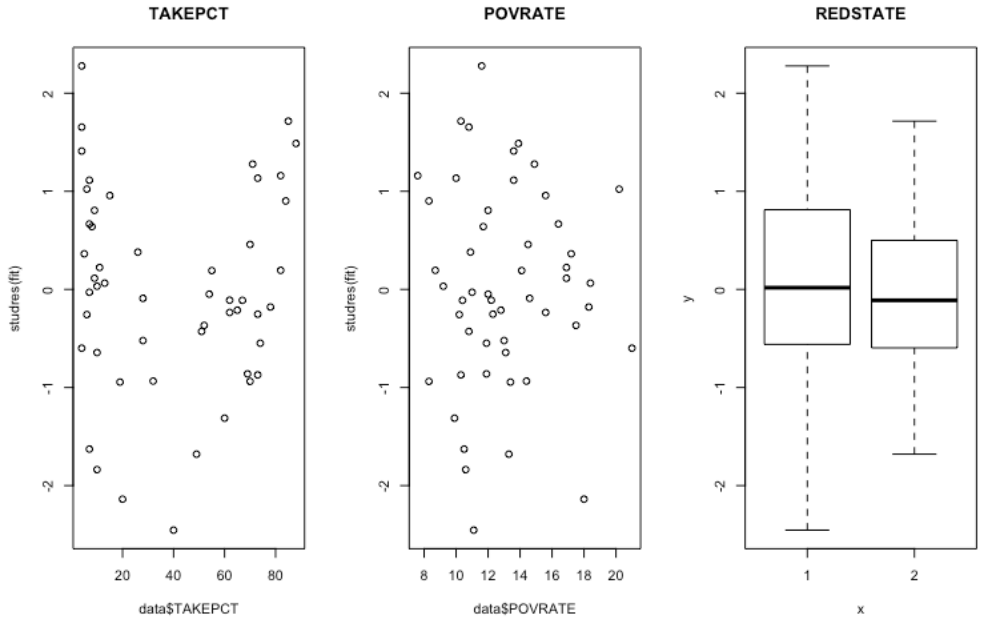


Figure 9



The QQ plot **(Figure 10)** indicates that the residuals do follow a normal distribution. A Cook's Distance plot **(Figure 11)** shows that there is no any highly influencer. Finally, VIF values were

checked to test for multicollinearity. Given that all of the VIF values are low, it does not appear multicollinearity is present.

Table 3

Predictor	TAKEPCT	POVRATE	factor(REDSTATE)
GVIF	1.391075	1.240344	1.479749

Figure 10

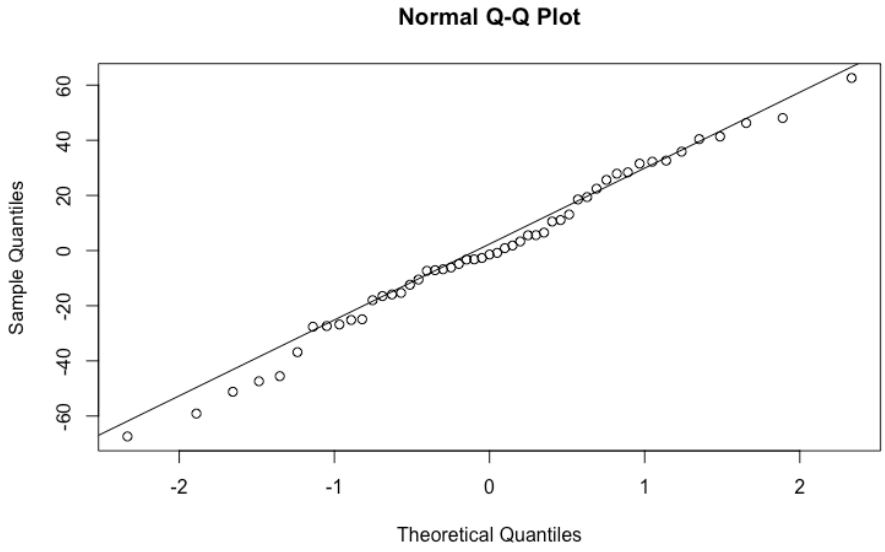
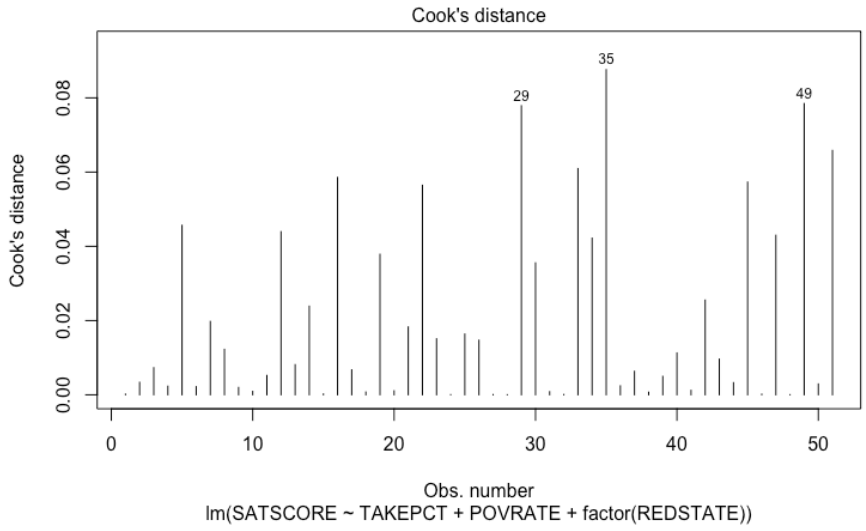


Figure 11



Given the above diagnostics, all assumptions of linear regression are met, and the model selected will be used for interpretation and analysis. The model that will be used is restated below:

$$SATSCORE = \beta_0 + \beta_1 TAKEPCT + \beta_2 POVRATE + \beta_3 REDSTATE(2) + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ and errors are independent, and

- SATSCORE (y) means the total SAT score, which is the average across the state
- TAKEPCT (x_1) means the percent of high school seniors taking the SAT
- POVRATE (x_2) means the percentage of people below the poverty line
- REDSTATE (x_3) means the political status of the state, where 1 means Republican while 2 means Democratic

The estimated model using these predictors is below:

$$\widehat{SATSCORE} = 1237.6298 - 2.6030 * TAKEPCT - 5.4149 * POVRATE + 28.3321 * REDSTATE(2)$$

where $\sigma^2 = 882.9657, R^2 = .8516$

Findings and Conclusion

All predictors were significant at the .05 level. The estimated model explains about 85.16% of the variation in the SAT score, which is high enough to draw meaningful insights from the model. The adjusted R^2 nearly remains the same for this model comparing to the full model with all 5 predictors (.8521 ≅ .8424), but the p-value of each predictor generally is much less than the one in the full model, indicating that the final model used is a better fit than just using all predictors.

We can get following conclusions drawn from the final estimated model:

1. Using the coefficient β_1, we can say that for every 1 percent increase in the percent of high school seniors taking the SAT, holding all other predictors constant, the total SAT score (average across the state) is predicted to decrease by 2.6030 points of 1600 total points.

2. Using the coefficient β_2, we can say that for every 1 percent increase in the percent of people below the poverty line, which is estimated for 2004-2006, holding all other predictors constant, the total SAT score (average across the state) is predicted to decrease by 5.4149 points of 1600 total points.

3. Using the coefficient β_3, we can say when holding all other predictors constant, a blue state (Democratic) has more 28.3321 points of the total SAT score (average across the state) on average than a read state (Republican).

4. The mean of each continuous predictor in the model is listed below:

   TAKEPCT = 39.17647 percentage POVRATE = 13.05294 percentage

   where TAKEPCT means the percent of high school seniors taking the SAT and POVRATE means the percentage of people below the poverty line

   4.1 A red state (Republican) which has these average characteristics is estimated to have a total SAT score of 1064.974 of 1600 points with a 95% C.I. of (1053.501, 1076.447).

   4.2 A blue state (Democratic) which has these average characteristics is estimated to have a total SAT score of 1093.306 of 1600 points with a 95% C.I. of (1077.664, 1108.949).

These results indicate that the presence of blue state (Democratic), a smaller percent of high school seniors taking the SAT, and a smaller percentage of people below the poverty line are all indicators of higher total SAT score of states.

The results of this study indicate that the percent of high school seniors taking the SAT, the percentage of people below the poverty line (estimate for 2004-2006), and the presence of the political status of the state can all be used to predict the total SAT score (average across the state) of the state. As presented in the examples above, the model estimated for this paper is helpful in predicting the total SAT score of each state as well as liking in how changes in the effect socioeconomic variables may affect the education level.

To have a further understanding about the influence of socioeconomic variables on the education, the researches in the future can cover more factors such as race, religion and so on to construct a more comprehensive analysis.

Appendix A
Code used for this analysis

```
################### Final Project ##########################

library('car')
library('MASS')

#setwd("/Users/yifeili/Library/CloudStorage/data_and_code")
data <- read.csv("Education.csv", header=T)
data$REDSTATE <- factor(data$REDSTATE)


#### Initial Data Analysis ######
range(data$SATSCORE)
hist(data$SATSCORE, main = 'Histogram of SAT Score', xlab="SAT Score")
boxplot(data$SATSCORE, main = 'Boxplot of SAT Score', ylab="SAT Score")

par(mfcol=c(1,4))
boxplot(data$TAKEPCT, main="TAKEPCT")
boxplot(data$EXPEND, main="EXPEND")
boxplot(data$POVRATE, main="POVRATE")
boxplot(data$MEDINCOME, main="MEDINCOME")
table(data$REDSTATE)

boxplot(SATSCORE ~ REDSTATE, data=data, main="SATSCORE by REDSTATE")

my_cols <- c("#FC4E07","#00AFBB")
pairs(data[,-c(1,5)],col=my_cols[data$REDSTATE])
cor(data[, -c(1,5)])
dev.off()

#### Full Model ####
fullfit <- lm(SATSCORE ~ TAKEPCT + EXPEND + factor(REDSTATE) + POVRATE + MEDINCOME,
data=data)
summary(fullfit)

plot(fullfit$fitted.values, studres(fullfit), xlab="Fitted Values", ylab="Studentized Residuals",
main="Full Fitted Model Residual Plot")
abline(h=c(0, 3, -3))
#heteroskedasticity

#data$logSATSCORE <- log(data$SATSCORE)
```

```
#fullfitlog <- lm(logSATSCORE ~ TAKEPCT + EXPEND + factor(REDSTATE) + POVRATE +
MEDINCOME, data=data)
#log(PSA) corrected residual plot

plot(fullfitlog$fitted.values, studres(fullfitlog), xlab="Fitted Values", ylab="Studentized
Residuals", main="Full Fitted Model Residual Plot")
abline(h=c(0, 3, -3))
#heteroskedasticity

qqnorm(fullfit$residuals)
qqline(fullfit$residuals)

#qqnorm(fullfitlog$residuals)
#qqline(fullfitlog$residuals)

hist(fullfit$residuals, main="Histogram of Residuals from Full Fitted Model", xlab="Residuals")
norm <- qqnorm(fullfit$residuals)
cor(norm$x, norm$y)
#looks approximately normal
#large enough sample size

plot(fullfit, 4)
plot(cooks.distance(fullfit), type="o", main="Cook's Distance")
identify(cooks.distance(fullfit))
order(cooks.distance(fullfit))
cooks.distance(fullfit)

#no need any change for current linear regression model

vif(fullfit)
# all fairly low, far less than 10


########### Model Selection ################
step(lm(SATSCORE ~ 1, data=data), SATSCORE ~ TAKEPCT + EXPEND + factor(REDSTATE) +
POVRATE + MEDINCOME, direction="both", trace=1)

step(lm(SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE), data=data), scope=. ~ .^2, trace
= 1, direction = "both")


################## Final Model ##################################
fit <- lm(SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE), data=data)
summary(fit)
```

```
summary(fit)$sigma^2

plot(fit$fitted.values, studres(fit), main="Final Model Residual Plot", xlab="Fitted
Values",ylab="Studentized Residuals")
abline(h=c(0,-3,3))
#good

par(mfcol = c(1,3))

plot(data$TAKEPCT, studres(fit), main="TAKEPCT")
plot(data$POVRATE, studres(fit), main="POVRATE")
plot(data$REDSTATE, studres(fit), main="REDSTATE")
#all look good

qqnorm(fit$residuals)
qqline(fit$residuals)
#good

plot(fit, 4)
#moderately high, not too concerned

vif(fit)
#good. Interactions are moderately high, but expected
#This is the best model


#### Interpretation ####
mean(data$TAKEPCT) #39.17647
mean(data$POVRATE) #13.05294
mean(data$SATSCORE)

predict(fit, data.frame(TAKEPCT=39.17647, POVRATE=13.05294,
REDSTATE=1),interval="confidence")
predict(fit, data.frame(TAKEPCT=39.17647, POVRATE=13.05294,
REDSTATE=2),interval="confidence")
```

Appendix B
Model Log

Main effects:

```
Start:  AIC=441.08
SATSCORE ~ 1

                   Df Sum of Sq    RSS    AIC
+ TAKEPCT           1    212414  67210 370.37
+ EXPEND            1     43393 236232 434.48
+ MEDINCOME         1     34177 245447 436.43
+ factor(REDSTATE)  1     17909 261716 439.70
<none>                          279625 441.08
+ POVRATE           1       757 278868 442.94

Step:  AIC=370.37
SATSCORE ~ TAKEPCT

                   Df Sum of Sq    RSS    AIC
+ POVRATE           1     19244  47966 355.17
+ factor(REDSTATE)  1     13850  53360 360.60
+ MEDINCOME         1     12236  54975 362.12
+ EXPEND            1     12077  55133 362.27
<none>                          67210 370.37
- TAKEPCT           1    212414 279625 441.08

Step:  AIC=355.17
SATSCORE ~ TAKEPCT + POVRATE

                   Df Sum of Sq    RSS    AIC
+ factor(REDSTATE)  1      6467  41499 349.78
+ EXPEND            1      5113  42853 351.42
<none>                          47966 355.17
+ MEDINCOME         1         5  47961 357.16
- POVRATE           1     19244  67210 370.37
- TAKEPCT           1    230901 278868 442.94

Step:  AIC=349.78
SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE)

                   Df Sum of Sq    RSS    AIC
<none>                          41499 349.78
+ EXPEND            1      1301  40198 350.16
+ MEDINCOME         1       484  41015 351.18
- factor(REDSTATE)  1      6467  47966 355.17
- POVRATE           1     11861  53360 360.60
- TAKEPCT           1    219292 260792 441.52

Call:
lm(formula = SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE),
    data = data)

Coefficients:
      (Intercept)            TAKEPCT            POVRATE  factor(REDSTATE)2
```

```
        1237.630            -2.603            -5.415            28.332
```

Interaction Effects:

```
Start:  AIC=349.78
SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE)

                           Df Sum of Sq    RSS    AIC
<none>                                   41499 349.78
+ TAKEPCT:factor(REDSTATE)  1      1238  40262 350.24
+ POVRATE:factor(REDSTATE)  1       474  41025 351.20
+ TAKEPCT:POVRATE           1       374  41125 351.32
- factor(REDSTATE)          1      6467  47966 355.17
- POVRATE                   1     11861  53360 360.60
- TAKEPCT                   1    219292 260792 441.52

Call:
lm(formula = SATSCORE ~ TAKEPCT + POVRATE + factor(REDSTATE),
    data = data)

Coefficients:
      (Intercept)            TAKEPCT            POVRATE  factor(REDSTATE)2
        1237.630             -2.603             -5.415             28.332
```