Expenditures and Test Scores

Daniel Roberts dir170130

5/7/2020

```r
# Caching
knitr::opts_chunk$set(cache = 1, echo = 1)

# Contains data set
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 3.6.3
```

```r
# Used for graphing and visuals
library(ggplot2)
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 3.6.3
```

```r
# Used for analyis
library(MASS)
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```r
library(MASS)
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

The Data

The data set 'sat' from the 'faraway' package consists of three potential response variables, four potential regressor variables and 50 observations. In this analysis the focus will be on predicting the total SAT score so the score breakdown is removed from the data set for easier use of the data set. The variables can be broken down as follows:

SAT Scores	Response and regressor variables
y	Total Score
x1	Expenditure per Student ($1000)
x2	Students/Teacher
x3	Teacher Salary ($1000)
x4	Percent of Test Takers

This data was gathered as averages from each state from the 1994 - 1995 school year, so each data point is a state average which is why there are 50. We can fit the full data model with the following:

```r
# Loading data
```

```r
summary(full)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## x1             4.4626    10.5465   0.423    0.674
## x2            -3.6242     3.2154  -1.127    0.266
## x3             1.6379     2.3872   0.686    0.496
## x4            -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

Analysis of the Full Model

Checking for Multicollinearity

Even though mulicollinearity does not occur often, it is important to check the model's VIF values to ensure that it is not effected by the issue.

```r
vif(full)
```

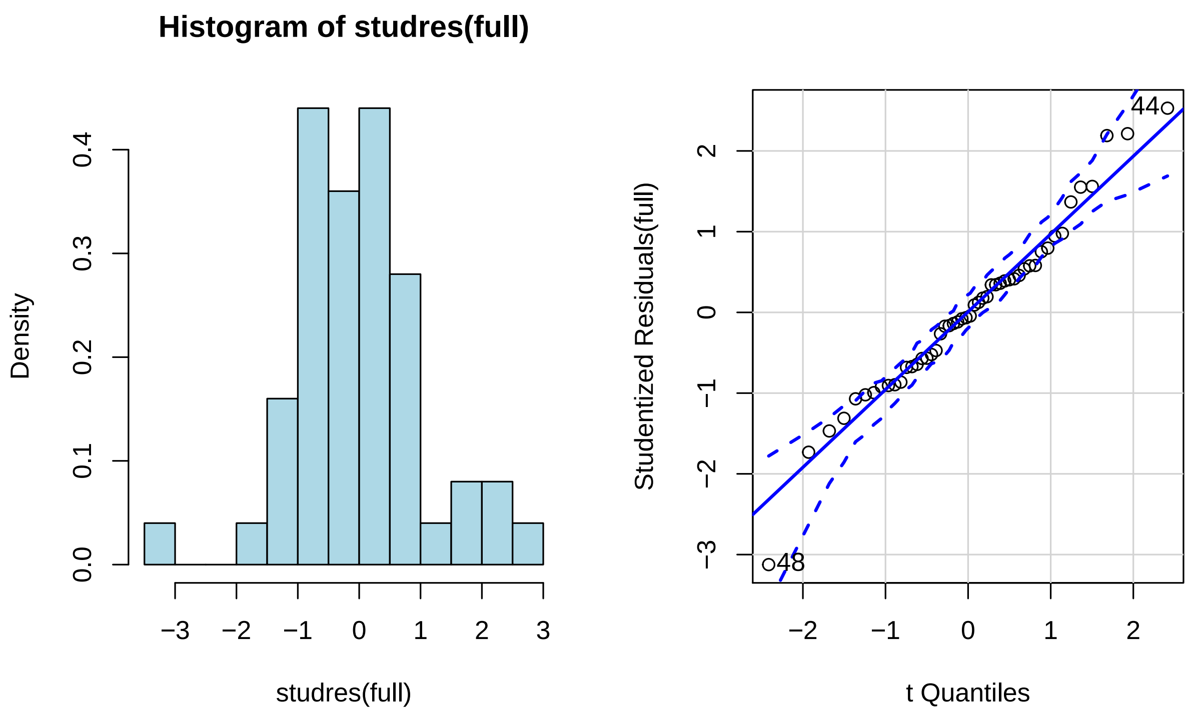
```
##       x1       x2       x3       x4
## 9.465320 2.433204 9.217237 1.755090
```

As seen in the summary since all VIF values are less than 10, even though expenditures and salary are close, we can say that there is no issues with multicollinearity in the full model.

Regression on Full Model

The analysis process starts with full model residual analysis to get a "big picture" idea of the data, regressors, and any potential outliers. This process can begin by looking at the histogram of studentized residuals and QQ-Plot of the studentized residuals.

```r
par(mfrow = c(1,2))
```

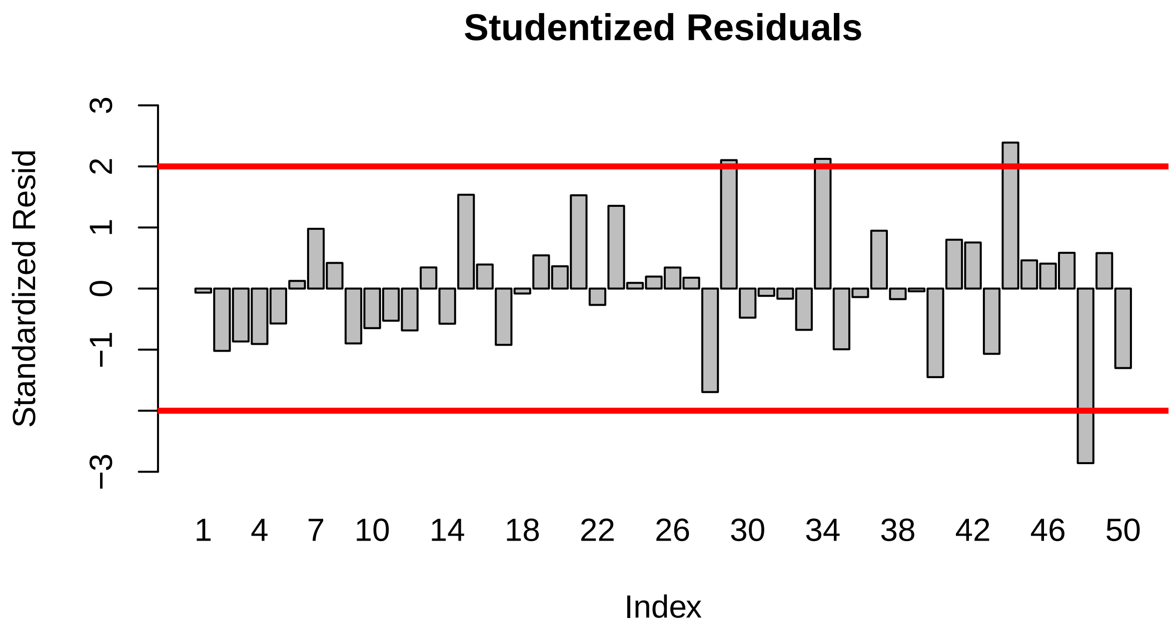



```
## [1] 44 48
```

We can see that the studentized residuals are normalized decently, but in the QQ-Plot analysis we see that points 44 and 48 made need further inspection.

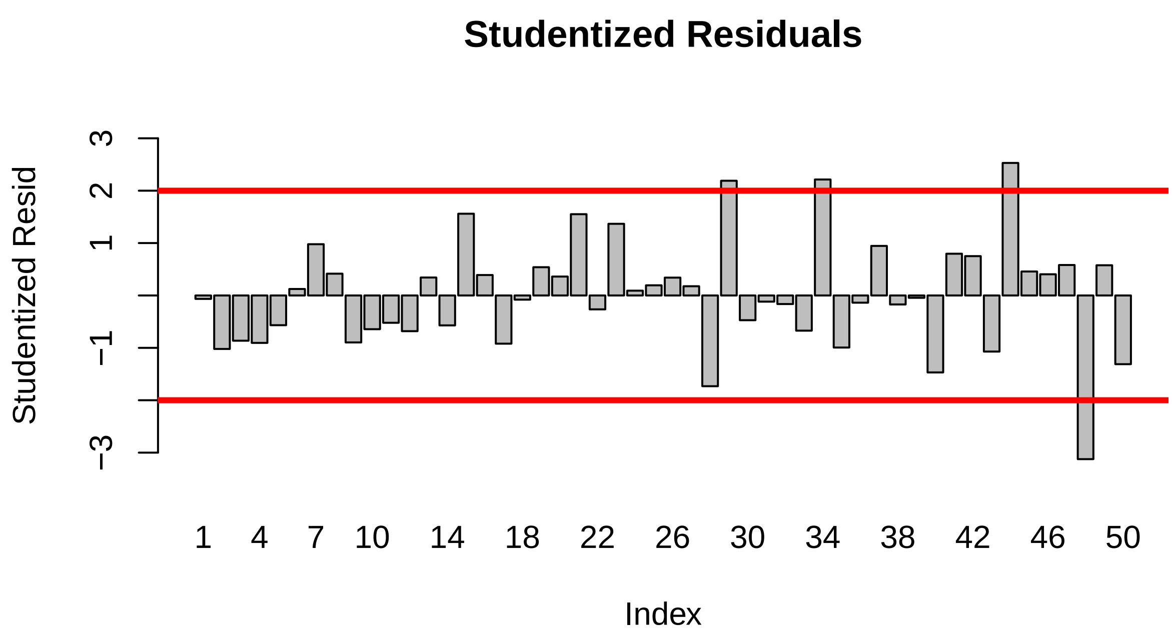
Next we further inspect the data by looking at the individual standardized and studentized residuals via barplots.

```
# Standardized
```




```
## numeric(0)
```

```
## 29 34 44 48
## 29 34 44 48
```




```
## numeric(0)
```

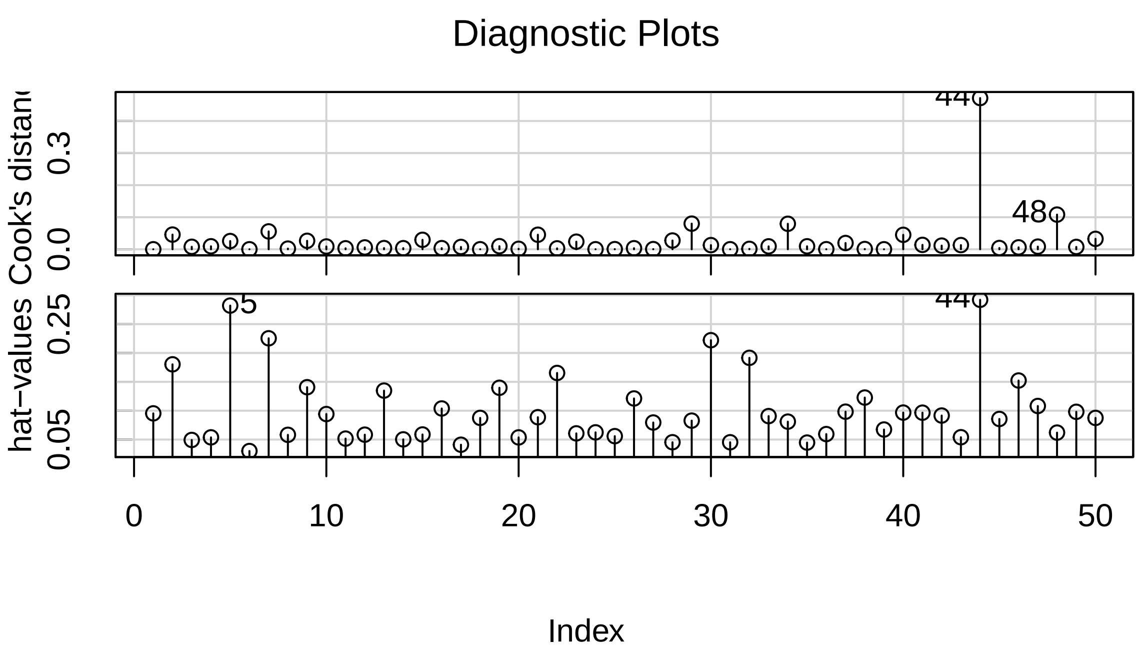
```
## 29 34 44 48
## 29 34 44 48
```

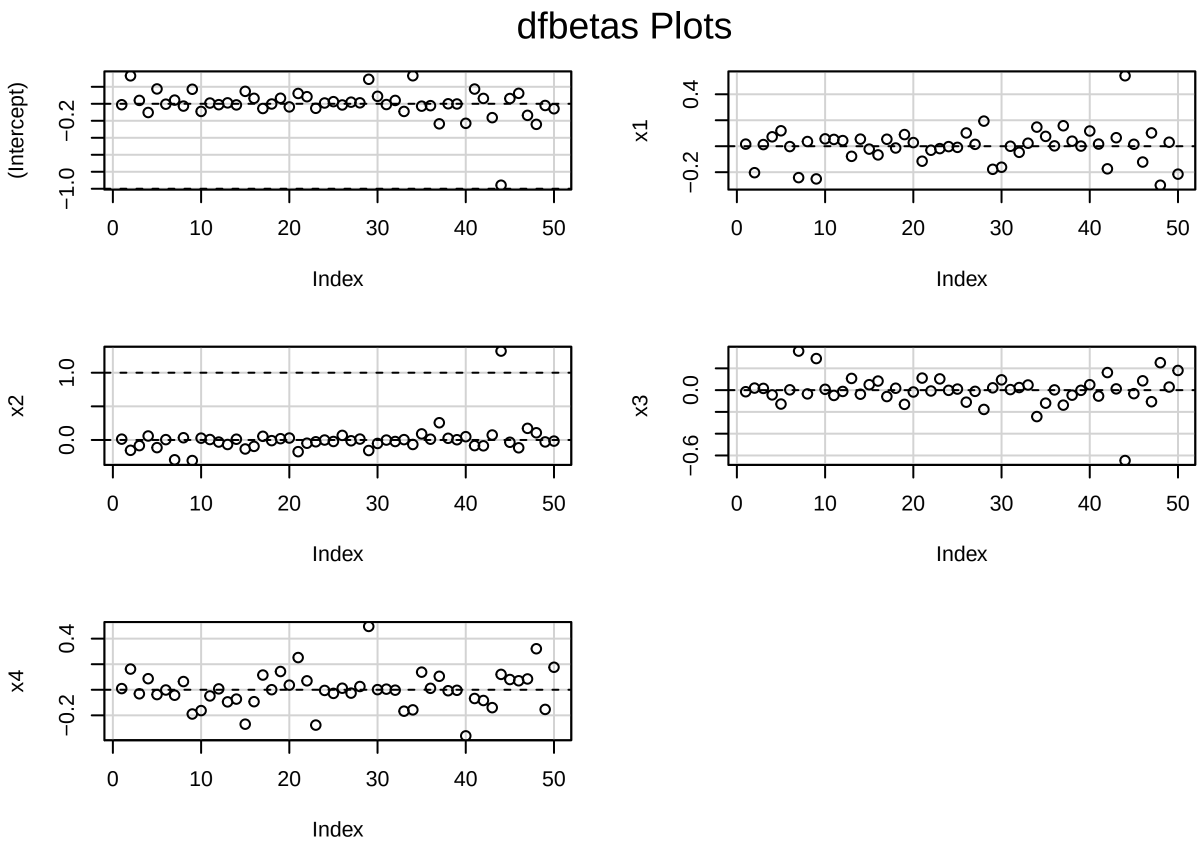
Based on the standardized and studentized residuals, points 44 and 48 are once again seen to be a potenial distortion of the model, but 29 and 34 may also require further inspection now.

We can continue the analysis via a look into the measures of influence by the points.

```
# Measures of influence summary
```

```
## Potentially influential observations of
##   lm(formula = y ~ x1 + x2 + x3 + x4) :
##
##    dfb.1_ dfb.x1 dfb.x2  dfb.x3 dfb.x4 dffit   cov.r   cook.d hat
## 5   0.17   0.12  -0.11   -0.13  -0.04  -0.36    1.50_*  0.03   0.28
## 30  0.09  -0.16  -0.05    0.09   0.00  -0.25    1.40_*  0.01   0.22
## 32  0.04  -0.05  -0.02    0.02   0.00  -0.08    1.38_*  0.00   0.19
## 44 -0.96   0.54   1.32_* -0.65   0.12   1.62_*  0.80    0.47   0.29
## 48 -0.24  -0.30   0.11    0.25   0.32  -0.80    0.44_*  0.11   0.06
```



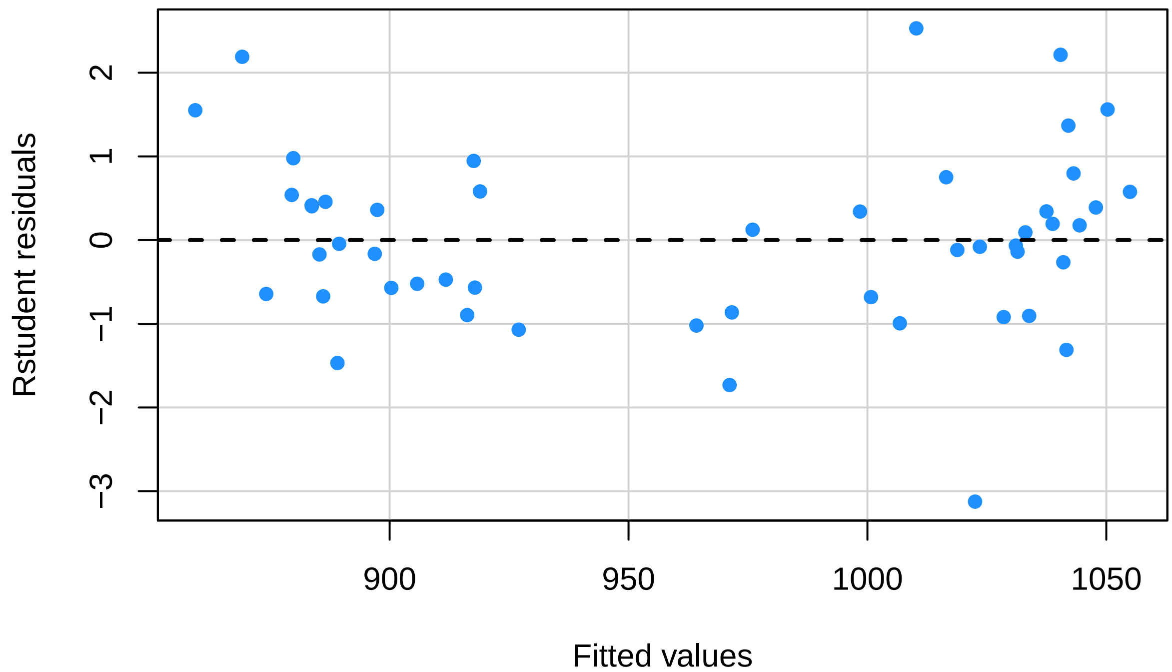



```
## [1]  5 30 32 44 48
```

From these measures of influence we see that points 44 and 48 continue to be an issue and may need to be removed. Aside from those two observations the rest have not been of trouble.

We conduct a final test of studentized residuals vs fitted values to conclude the residual testing.

```r
residualPlot(full, type = "rstudent", quadratic = F, col = "dodgerblue",
             pch = 16)
```



Analyzing this final plot we see that aside from point 48 the graph is fairly evenly spread.

Results of Initial Analysis

After initial analysis of the full model using all data we can conclude that the data does not need to be transformed but points 44 and 48 should be removed and the data should then be reanalyzed as a full model before moving onto model selection.

Residual Analysis with Modified Data

Since the full model has been anaylyzed once already we will take a look at all plots together rather than going step-by-step. We start by modifying our data.

```
# Resetting and removing
```

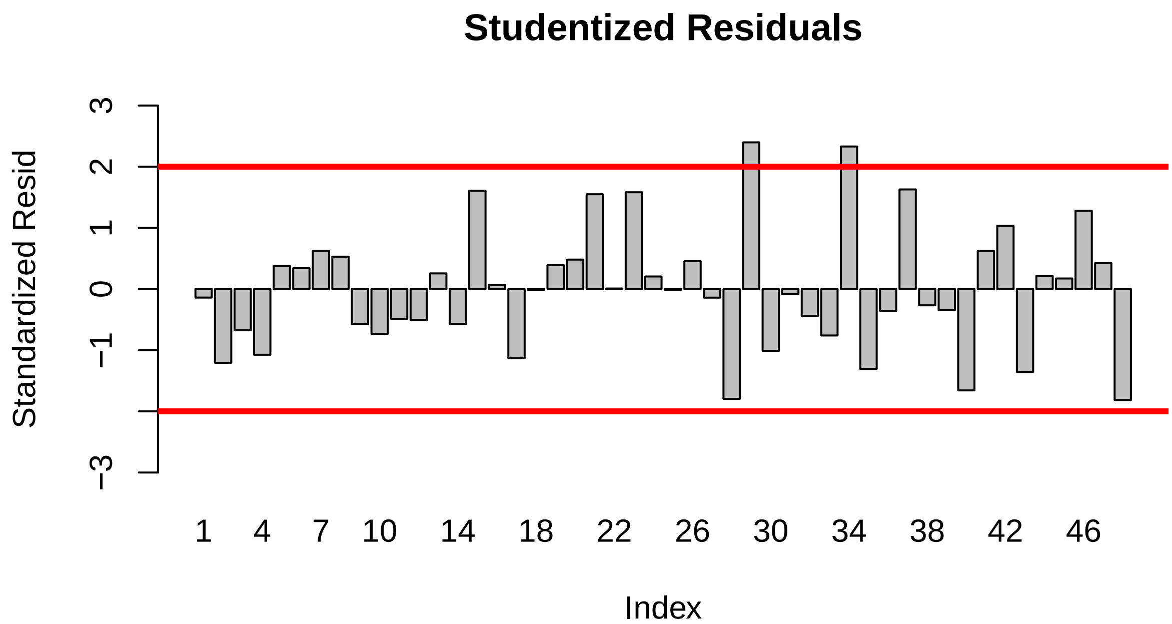
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.477 -16.066  -0.417  12.046  63.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1106.9838    48.0189  23.053   <2e-16 ***
## x1             1.8681     9.1687   0.204    0.840
## x2            -8.0628     3.0739  -2.623    0.012 *
## x3             2.5734     2.0917   1.230    0.225
## x4            -3.0006     0.1971 -15.226   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.69 on 43 degrees of freedom
## Multiple R-squared:  0.8737, Adjusted R-squared:  0.8619
## F-statistic: 74.34 on 4 and 43 DF,  p-value: < 2.2e-16
```

```
##       x1       x2       x3       x4
## 9.411090 2.359893 9.626851 1.709083
```

Modified Data Residual Analysis

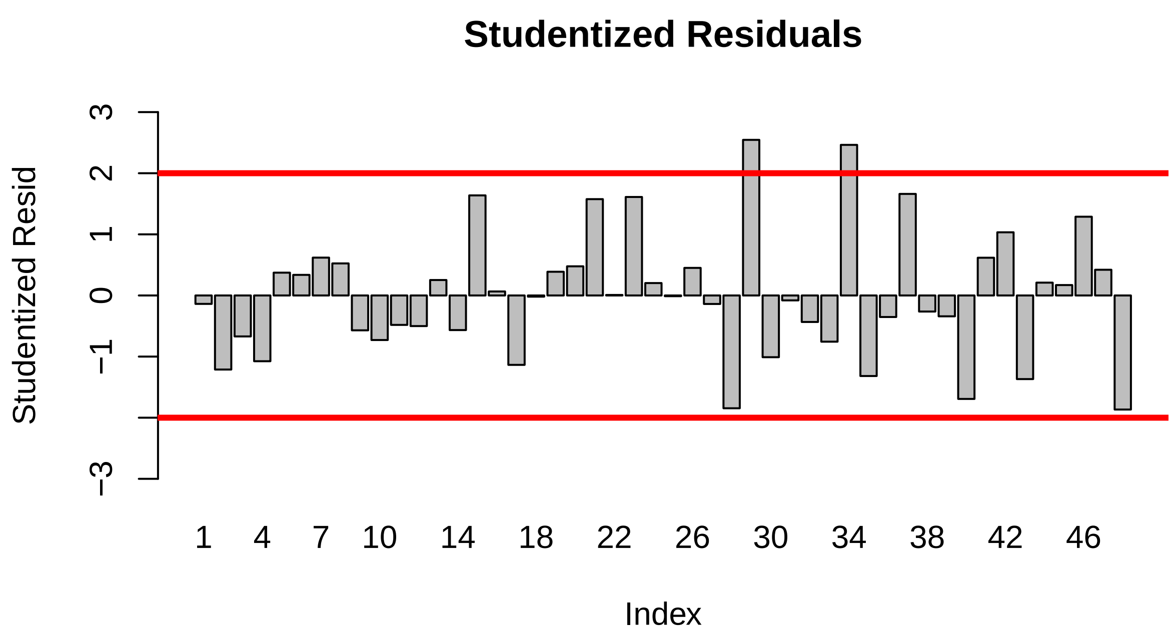
We see that there is still no issues with multicollinearity so we move onto residual analysis.

```
# Standardized
```




```
## numeric(0)
```

```
## 29 34
## 29 34
```

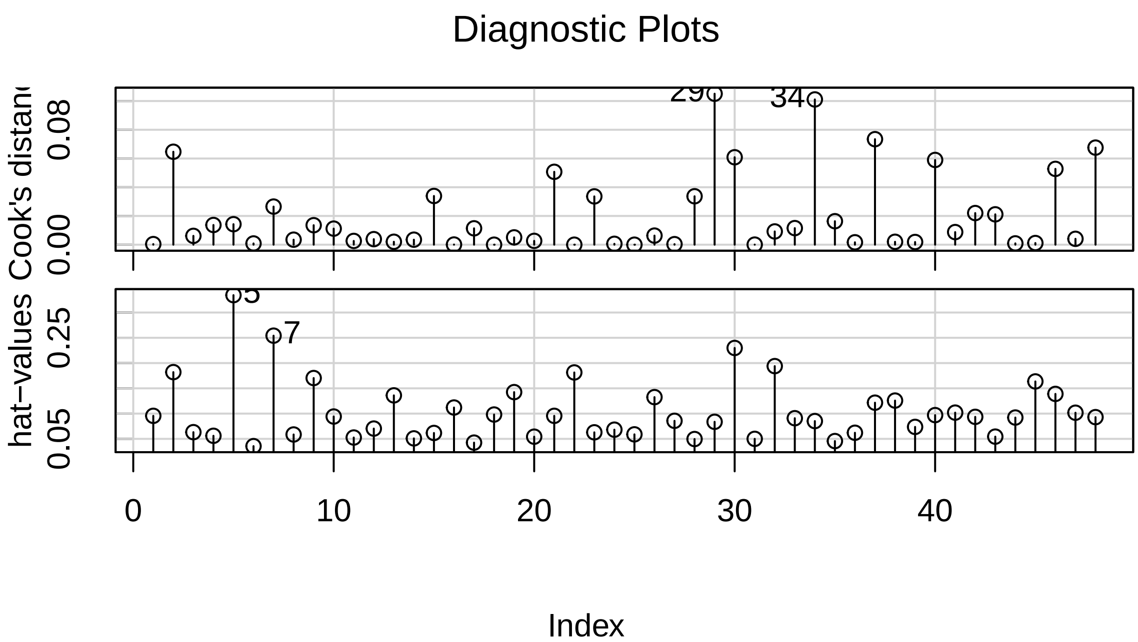



```
## numeric(0)
```

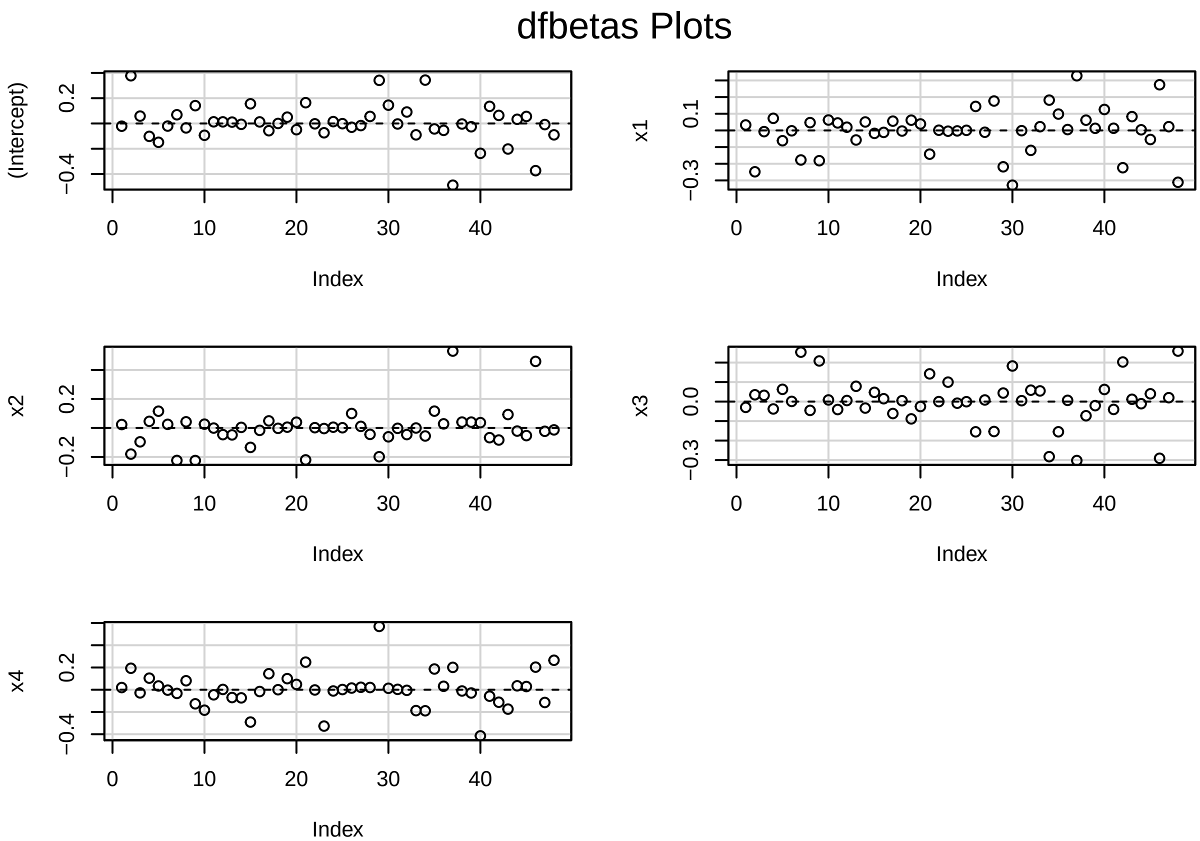
```
## 29 34
## 29 34

## Potentially influential observations of
##   lm(formula = y ~ x1 + x2 + x3 + x4) :
##
##    dfb.1_ dfb.x1 dfb.x2 dfb.x3 dfb.x4 dffit cov.r   cook.d hat
## 5  -0.15  -0.06   0.12   0.06   0.03   0.26  1.66_*  0.01   0.33_*
## 7   0.07  -0.18  -0.22   0.25  -0.03   0.36  1.44_*  0.03   0.25
## 22  0.00   0.00   0.00   0.00   0.00   0.00  1.37_*  0.00   0.18
## 29  0.34  -0.22  -0.20   0.04   0.57   0.77  0.60_*  0.11   0.08
## 32  0.09  -0.12  -0.05   0.06  -0.01  -0.21  1.36_*  0.01   0.19
## 34  0.34   0.18  -0.06  -0.28  -0.19   0.75  0.63_*  0.10   0.09
```

```r
# Plotted
```

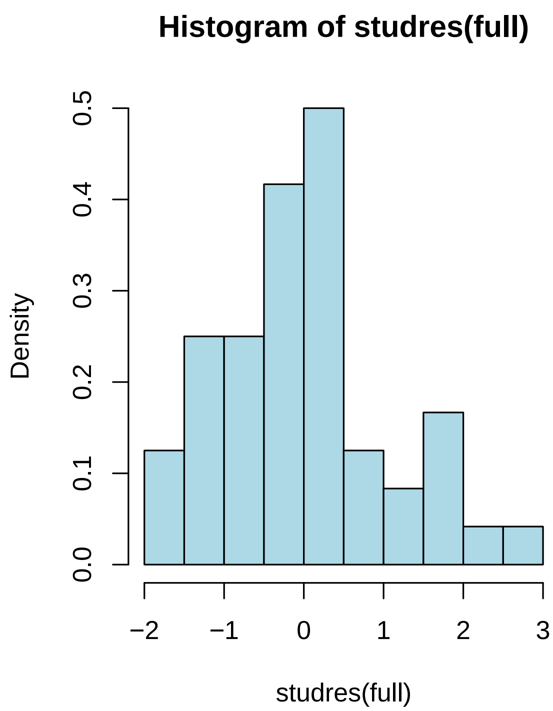


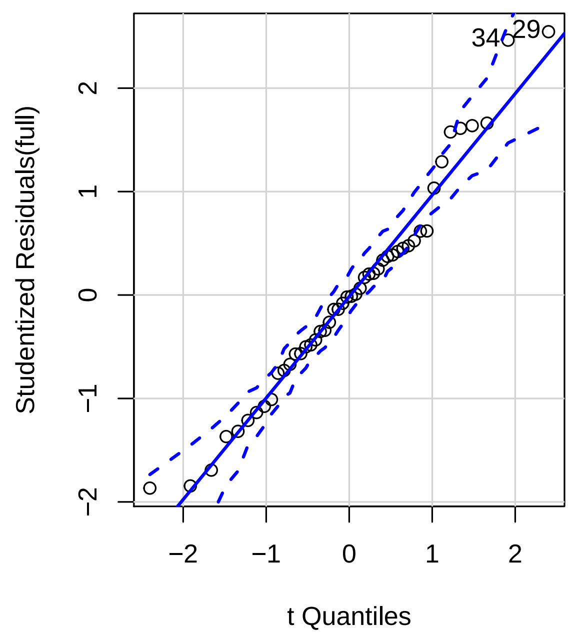
Diagnostic Plots
Cook's distanc
0.00
0.08
29
34
hat−values
0.05
0.25
5
7
0
10
20
30
40
Index



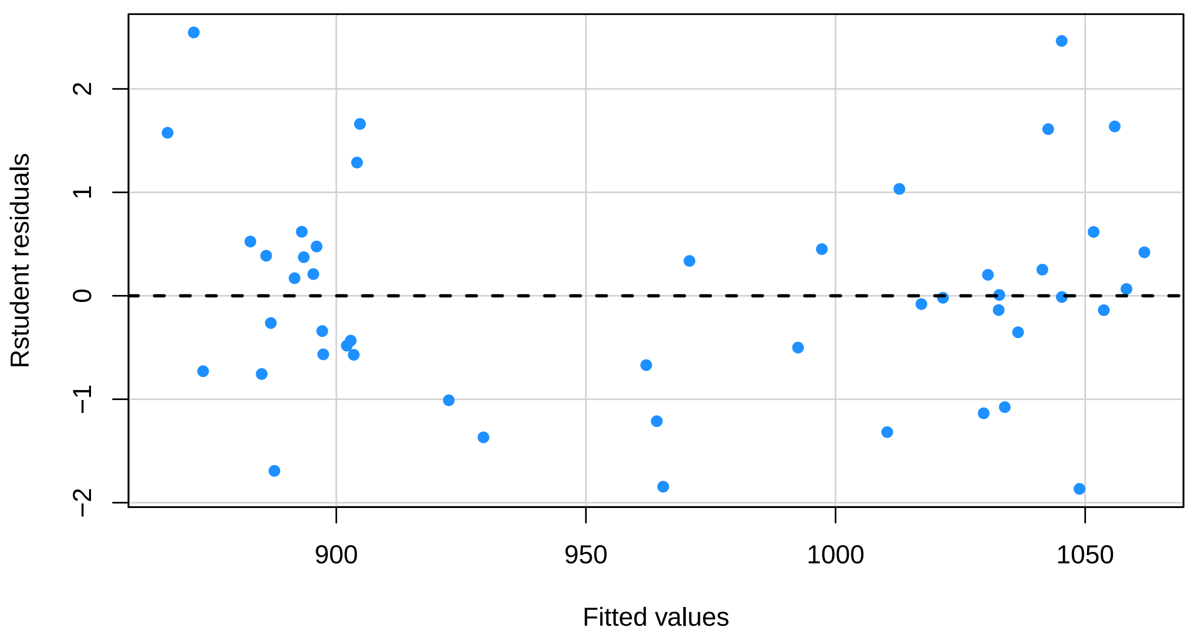
We can already see from the barplots of the standardized and studentized residuals that there is improvement after removing points 44 and 48.

```
par(mfrow = c(1,2))
```


Histogram of studres(full)
Density
0.0
0.1
0.2
0.3
0.4
0.5
−2
−1
0
1
2
3
studres(full)


Studentized Residuals(full)
−2
−1
0
1
2
34
29
t Quantiles

```
## [1] 29 34
```


Rstudent residuals
−2
−1
0
1
2
900
950
1000
1050
Fitted values

Results of Modified Data Analysis

After observing the histogram and QQ-Plot we see that the data has become slightly less normalized but not by enough to be concerning and the studentized residuals vs fitted values plot is evenly dispersed. Based on this result, the other residual analysis graphs and improvment in the adjusted R^2 value we can conclude that the model is helped by the removal and can move on to model fitting.

Model Fitting

Now that we have our final dataset we must see which combination of variables provides the best possible model. We do this by adding or removing variables from the model one at a time based on specific selection criteria. For this data we use a forward selection method.

Forward Selection

```
fwd <- regsubsets(total ~ ., method = "forward", data = data)
```

```
## Subset selection object
## Call: regsubsets.formula(total ~ ., method = "forward", data = data)
## 4 Variables  (and intercept)
##        Forced in Forced out
## expend     FALSE      FALSE
## ratio      FALSE      FALSE
## salary     FALSE      FALSE
## takers     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: forward
##          expend ratio salary takers
## 1  ( 1 ) " "    " "   " "    "*"
## 2  ( 1 ) "*"    " "   " "    "*"
## 3  ( 1 ) "*"    "*"   " "    "*"
## 4  ( 1 ) "*"    "*"   "*"    "*"
```

From this forward selection process we see that all regressors are candidates to be in the final model, so we now have to generate additional selection criteria and use this to determine the best possible model.

Selction Criteria

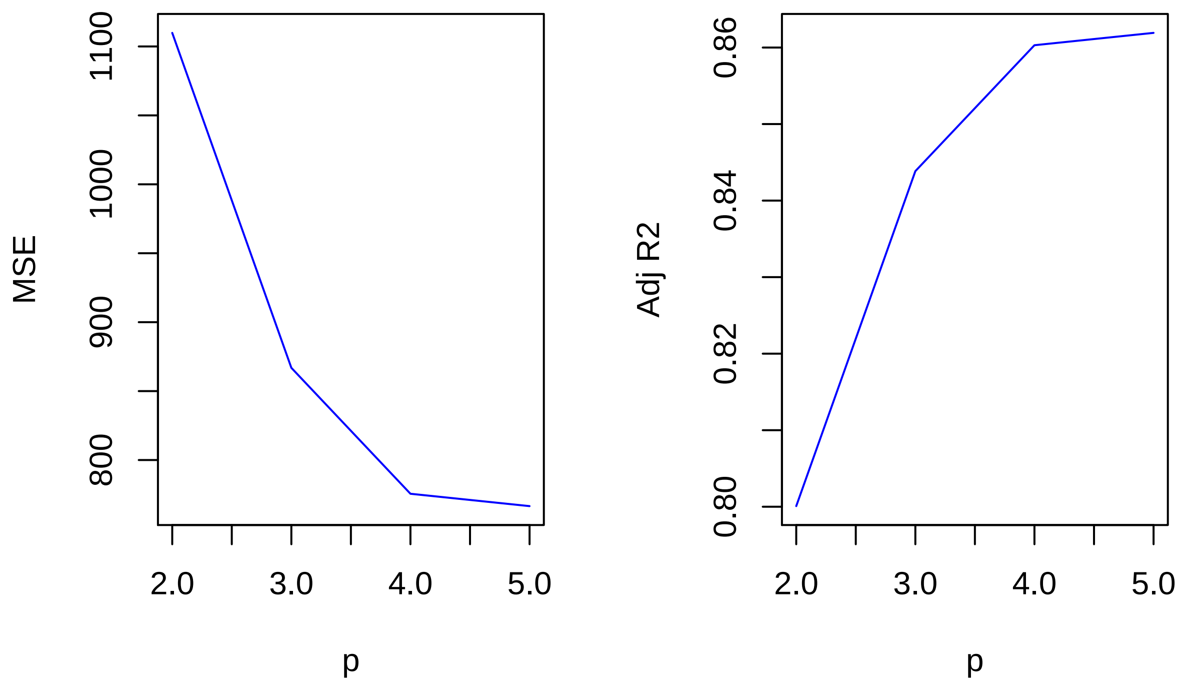
```
mse <- summary(fwd)$rss / (n - (2:5))
```

```
##         MSE    Adj R2        Cp       BIC
## 2 1109.8392 0.8000816 22.598376 -70.56250
## 3  866.8788 0.8438467  8.888151 -79.60571
## 4  775.5222 0.8603030  4.513597 -82.15861
## 5  766.5743 0.8619149  5.000000 -79.94795
```

From this table we can already see that the full four variable model may be the best option, but the data becomes much more visible when applied to a graph.

Graphing Selection Criteria

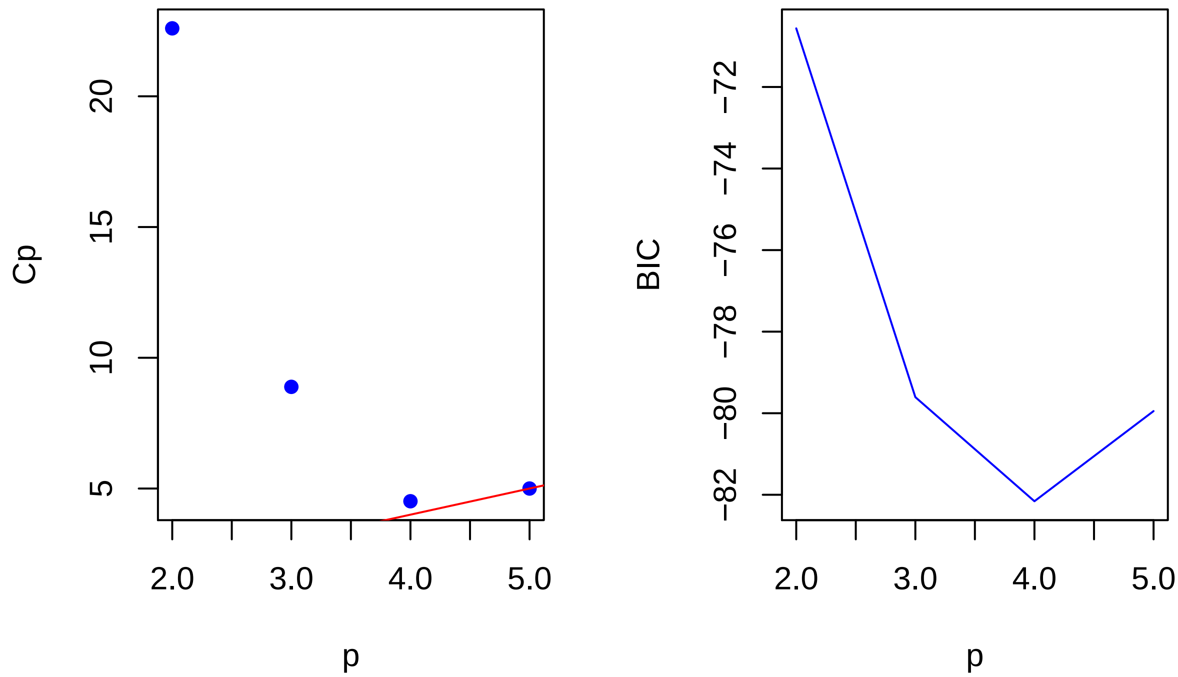
```
par(mfrow=c(1,2))
```



We do not graph the R^2 value here but rather the adjusted R^2 value becuase the R^2 increases with number of variables. This means it favors models with more variables and is not a good indicator for model selection. We continue by looking at the C_p and BIC values.

```
par(mfrow=c(1,2))
```

```
## integer(0)
```



We see from the C_p graph that the idea of choosing the full model as the final model continues to be a good descision, however it could also indicate that the three variable model is an option. We also see that the BIC graph indicates a three variable model, however this may be due to the fact that the data set being used has a relativley small number of observations which negatively effects the accuracy of the BIC values.

There is a final test to perform to confirm which model may be the best option for this data set, the stepwise selection method.

Stepwise Selection

```
intercept <- lm(total ~ 1, data = data)
```

```
## Start:  AIC=414.84
## total ~ 1
##
##         Df Sum of Sq    RSS    AIC
## + x4     1    209866  51053 338.53
## + x3     1     48375 212544 406.99
## + x1     1     31479 229440 410.67
## <none>               260919 414.84
## + x2     1       258 260660 416.79
##
## Step:  AIC=338.53
## total ~ x4
##
```

```
##        Df Sum of Sq   RSS    AIC
## + x1    1   12043.1 39010 327.62
## + x2    1    9113.6 41939 331.09
## + x3    1    5170.8 45882 335.41
## <none>              51053 338.53
##
## Step:  AIC=327.62
## total ~ x4 + x1
##
##        Df Sum of Sq   RSS    AIC
## + x2    1    4886.6 34123 323.19
## <none>              39010 327.62
## + x3    1     772.7 38237 328.66
##
## Step:  AIC=323.19
## total ~ x4 + x1 + x2
##
##        Df Sum of Sq   RSS    AIC
## <none>              34123 323.19
## + x3    1    1160.3 32963 323.53
```

From this stepwise selection we see that the three variable model may be just as good of an option as the full model, however the final AIC values that determines the three variable vs full model only have a difference of .34 so either model could be argued as best by this method. Due to the fact that the last test was partially inconclusive we can perform a backwards selection as an additional test.

Backward Selection

```
bwd <- regsubsets(total ~ ., method = "backward", data = data)
```

```
## Subset selection object
## Call: regsubsets.formula(total ~ ., method = "backward", data = data)
## 4 Variables  (and intercept)
##        Forced in Forced out
## expend     FALSE      FALSE
## ratio      FALSE      FALSE
## salary     FALSE      FALSE
## takers     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: backward
##          expend ratio salary takers
## 1  ( 1 ) " "    " "   " "    "*"
## 2  ( 1 ) " "    "*"   " "    "*"
## 3  ( 1 ) " "    "*"   "*"    "*"
## 4  ( 1 ) "*"    "*"   "*"    "*"
```

```
##         MSE    Adj R2        Cp       BIC
## 2 1109.8392 0.8000816 22.598376 -70.56250
## 3  931.9778 0.8321203 12.709635 -76.13004
## 4  749.8754 0.8649229  3.041513 -83.77283
## 5  766.5743 0.8619149  5.000000 -79.94795
```

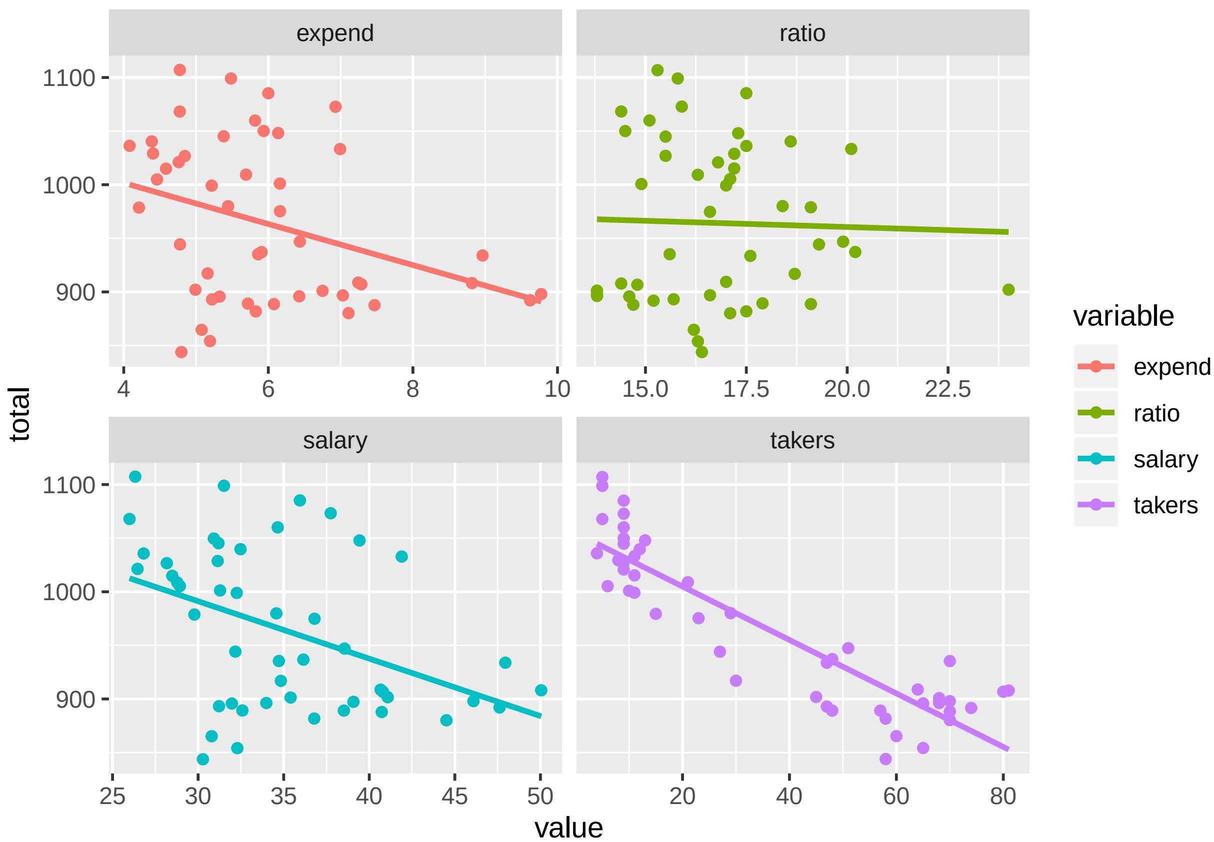
From both the model selection and most of the selection criteria we see that a full model is the best model for this data. The only contradictory evidence is BIC values which as stated earlier can be less accurate with less observations like in our case.

Model Fitting Results

After perfroming intensive model fitting tests we have determined that a full model with all four regressors is the best option for regression. Now that we have determined that a full model is the best option and since we have already run residual testing on the model we can take a look at the regressors plotted individually against the predicted SAT score values and then make recomendations.

Regressors vs Predicted Values

```r
data2 <- melt(data, id.vars = 'total')
```



Conclusion

From this regression model we can see that surprisingly that the more money spent per student and the higher a teacher's salary, the worse a student performs on the SAT overall. This may be due to a percieved laziness in more weathly areas such as suburbs, whereas students in lower income areas, and therefore less money to schools, are more driven to succeed to leave the area they are in.

Not as surprisingly we see that the higher the percentage of test takers in that are in a state the worse the states overall average is. This point is less interesting because it comes mostly as common sense that the more people who take an exam the worse the overall average will be.

Finally we see that the student to teacher ratio does not have a strong effect on the total SAT score. This comes as more of a surprise since you would think more one on one time with a teacher would improve overall retention of a taught subject.

Reflection on Analysis

Looking back on the overall anaylsis of the data set, having more data over several school years would improve the overall accuracy of the regression. If data over multiple years was not available the data broken down by county or city averages rather than by state could also be a good way to add additional data while still being obtainable.

Having more variables to analyze could potentially help the overall regression depending on what variables could be available for analysis. While a subject matter expert would be the best one to determine additional variables I believe that having the average student GPA would also be useful as it is a good determination of a student's knowledge.

I believe the regression might have been skewed by the fact that most of the data exists on the lower end of the regressor variable's range. This is primarily seen in expenditures and teacher salary as in both catagories the majority of data points are in the bottom 50% of the the value range. Seeing the data from higher end neighborhoods or private schools could provide more data on the higher end of the value range.