# NCSU ST 503 Discussion 6

Probem 6.1 Faraway, Julian J. Linear Models with R CRC Press.

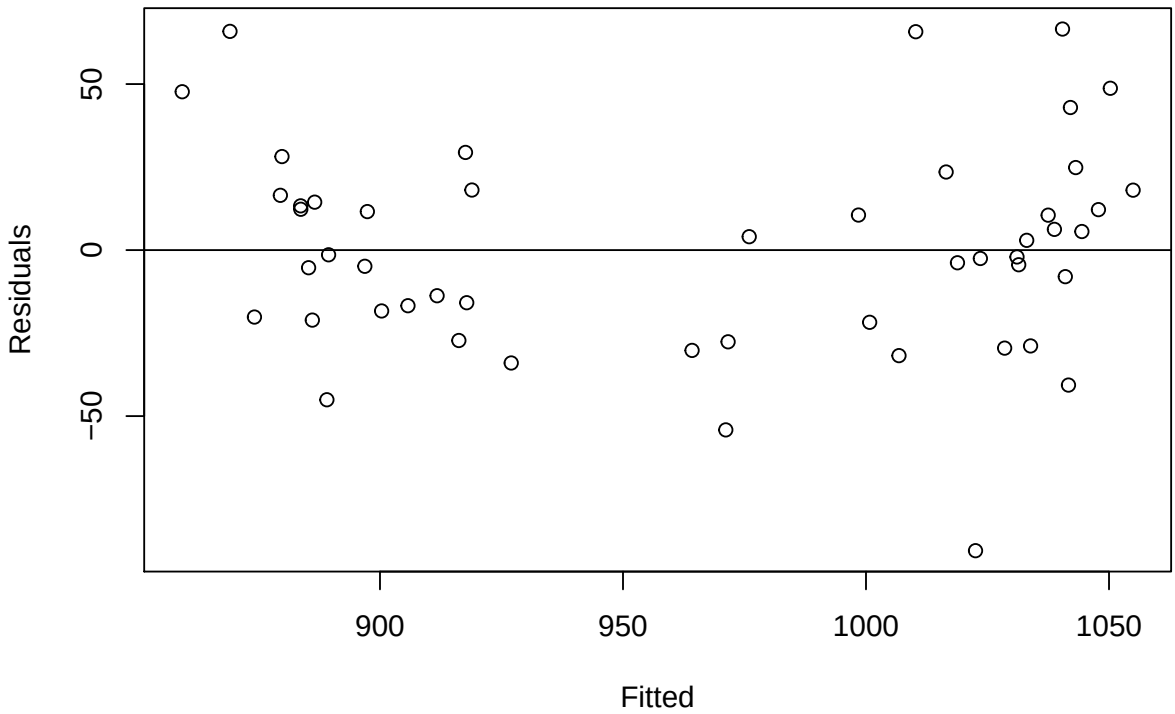
*Bruce Campbell*

---

## Regression diagnostics with the SAT data set.

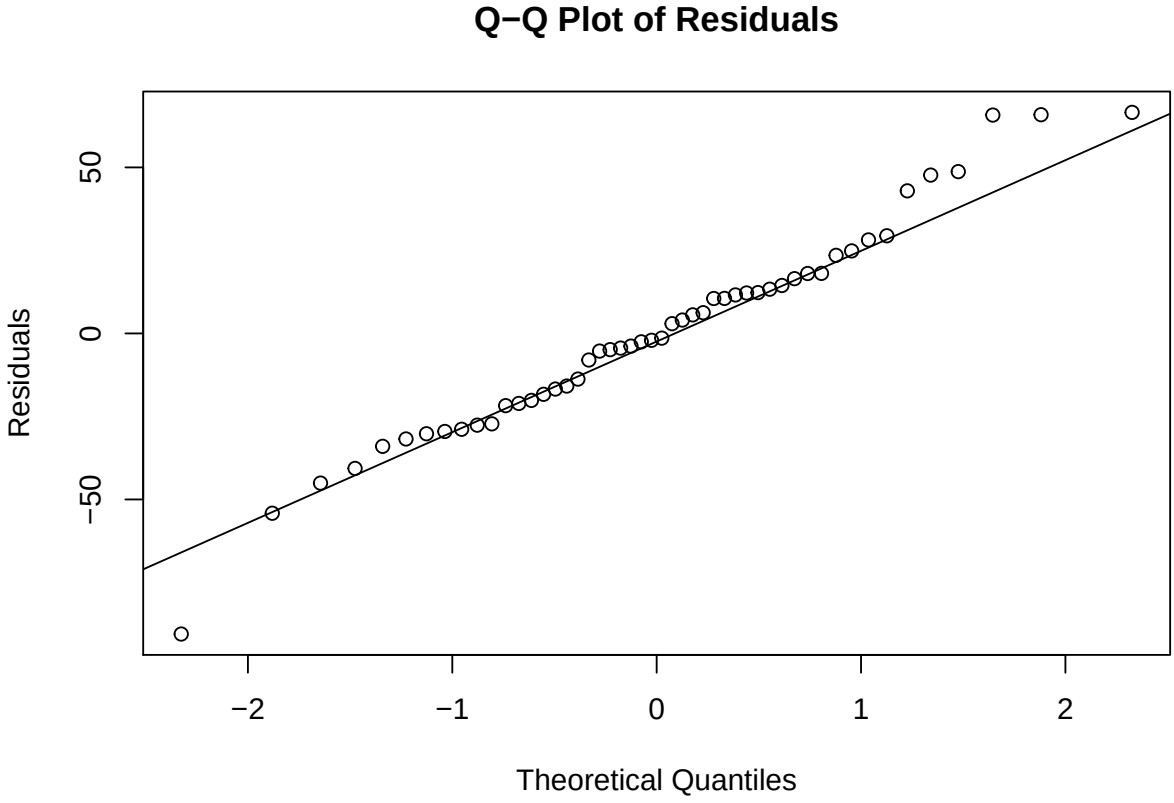
*Using the sat dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.*

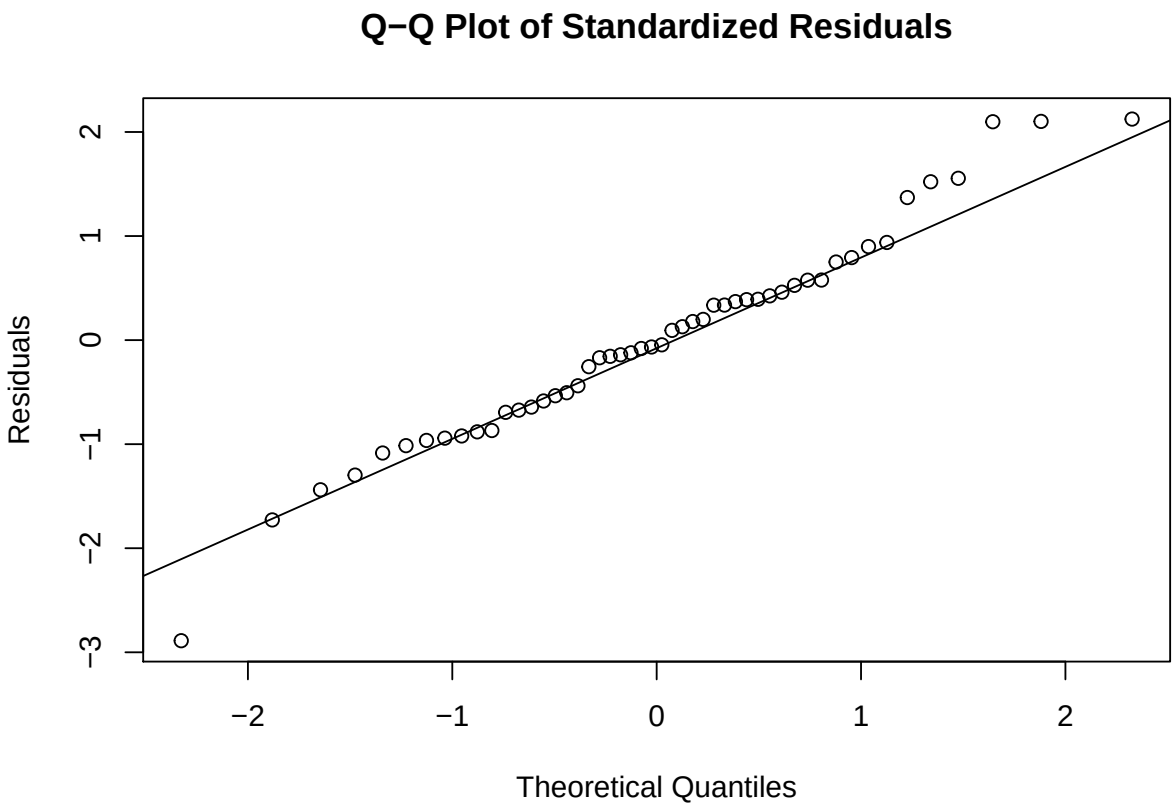
### (a) Check the constant variance assumption for the errors.



To check the assumption of constant variance we plot fitted values against the residuals - looking for any structure in the distribution of values about the theoretical mean value line $E[\epsilon] = 0$. There is nothing alarming with this plot, the variance seems relatively constant along the range of the fitted values.

**(b) Check the normality assumption.**






Generally the residuals appear normally distributed in the middle of the range. The empirical distribution is slightly right skewed and there's a single point on the lower quantile that deviates from the theoretical distribution.

**(c) Check for large leverage points.**

Table 1: High Leverage Data Elements

| | expend | ratio | salary | takers | verbal | math | total |
|---|---|---|---|---|---|---|---|
| **California** | 4.992 | 24 | 41.08 | 45 | 417 | 485 | 902 |
| **Connecticut** | 8.817 | 14.4 | 50.05 | 81 | 431 | 477 | 908 |
| **New Jersey** | 9.774 | 13.8 | 46.09 | 70 | 420 | 478 | 898 |
| **Utah** | 3.656 | 24.3 | 29.08 | 4 | 513 | 563 | 1076 |

We've used the rule of thumb that points with a leverage greater than $\frac{2p}{n}$ should be looked at.

**(d) Check for outliers.**

Table 2: Range of Studentized residuals

| range.residuals.left | range.residuals.right |
|---|---|
| -3.124 | 2.53 |

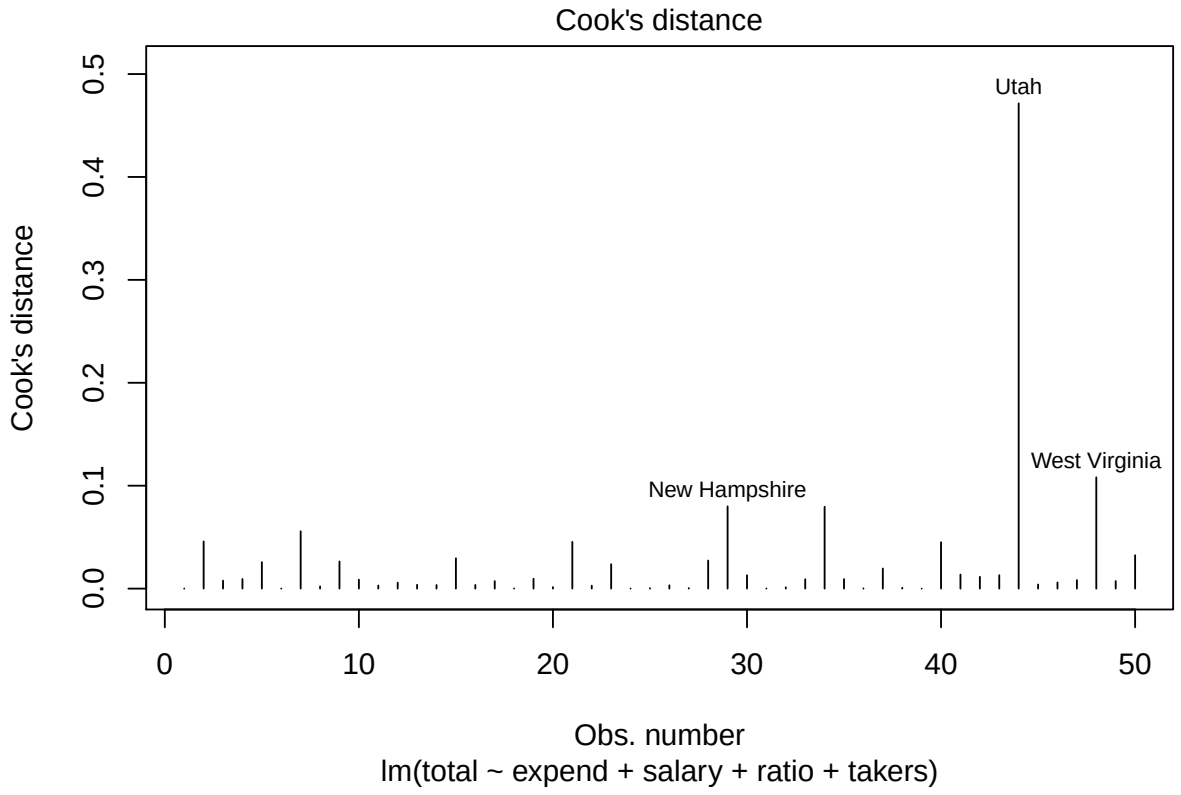
Table 3: Bonferroni corrected t-value

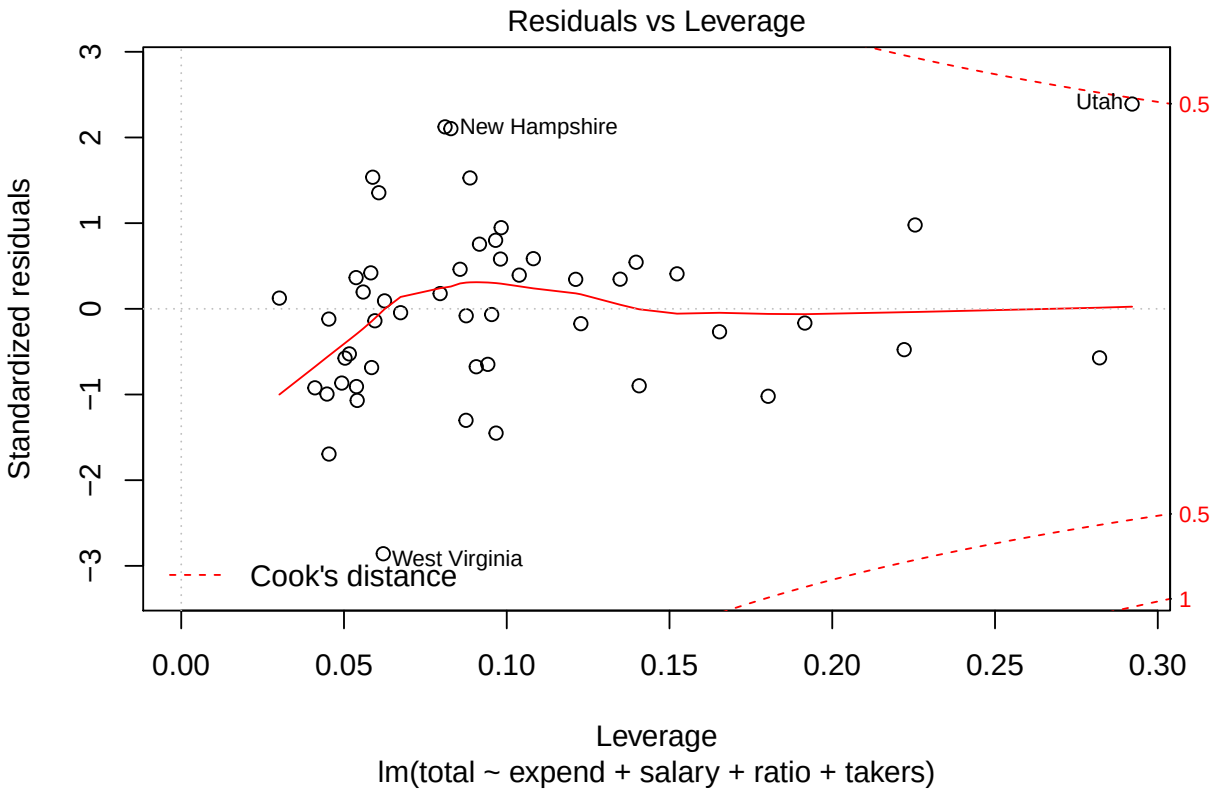
| t.val.alpha |
|---|
| -3.526 |

Since none of the studentized residuals fall outside the interval given by the Bonferroni corrected t-values we claim there are no outliers in the dataset.

**(e) Check for influential points.**

We plot the Cook's distances and the residual-leverage plot with level set contours of the Cook distance.






We see the Utah, New Hampshire, and West Virginia are candidate influential points. The book does not discuss a criteria for selecting influential points from the Cook distances.

Some guidelines for selecting influential points; * points with a Cook distance more than three times the mean Cook distance
* points with a Cook distance greater than 4/n * points with a cook distance greater than 1

Here we select points with a Cook distance more than three times the mean Cook distance.

Table 4: Mean Cook distance

| mean.cooks.distance |
|---|
| 0.02575 |

Table 5: Points with Cook distance greater than three times the mean Cook distance.

| | cook.distance |
|---|---|
| **New Hampshire** | 0.07989 |
| **North Dakota** | 0.07954 |
| **Utah** | 0.4715 |

| | cook.distance |
|---|---|
| **West Virginia** | 0.1081 |