Analysis of SAT Dataset

Ian McLeod

October 16, 2017

Purpose

First. This report provides an illustrative example of iterative regression model building including the use of important regression plots and diagnostics:

- Scatterplot Matrix
- Variance Inflation Factor Barchart
- Visualization of Correlation Matrix
- Residuals vs Fitted plot
- Scale-Location plot
- Normal Q-Q plot
- Residual vs Leverage plot
- Residual dependency plots
- Overfitting model diagnostic check using polynomial regression

Second. It is shown how to use RMD (R markdown) to produce a beautiful, well organized PDF report. In order to make this report more readable, I have suppressed the R code. Please see the associated RMD file for the complete script to produce this document. I would like students to adopt this style with their Assignments and Projects. Specifically:

- Strive to make your report readable and nicely formatted.
- Upload your Assignment as a PDF but also upload the Rmd file used to create the ouput

When compiling for PDF rather than HTML, an additional challenge in the quest to produce a well-organized presentation is that figures will *float* and so are not necessarily in close proximity to the relevant parts of the text. To overcome this you can:

- use the *chunk options* **fig.height** and **fig.width**
- use the latex command **newpage** to start a fresh page

Other style suggestions include:

- number all figures and tables
- include titles for all figures and tables
- no raw computer output - use only tables and figures
- can also use inline R output in text

Preamble

Before running the Rmd-script for this document, you will need to install some CRAN packages by running the following commands:

```
install.packages("tibbleverse")
```

I use 'tibbleverse' scripting to generate the *residual dependency plots*. Also I include a script for my function **vifx()** which computes the variance inflation factors for a given design matrix.

Introduction to SAT Dataset

We input the spreadsheet in CSV format to R using the function **read.csv()**. Table 1 shows the beautifully formatted output using stargazer to summarize the dataframe.

Table 1: Dataframe Summary.

Statistic	N	Mean	St. Dev.	Min	Max
cost	50	5.905	1.363	3.656	9.774
ratio	50	16.858	2.266	13.800	24.300
salary	50	34.829	5.941	25.994	50.045
percent	50	35.240	26.762	4	81
verbal	50	457.140	35.176	401	516
math	50	508.780	40.205	443	592
sat	50	965.920	74.821	844	1,107

The question of interest for our analysis is what are the variables that are useful in predicting the total SAT score from among the possible *predictor variables*: **cost**, **ratio**, **salary** and **percent**.

Scatterplot Matrix

It is most important to read across the rows to see the y vs. x relationship for each variable!

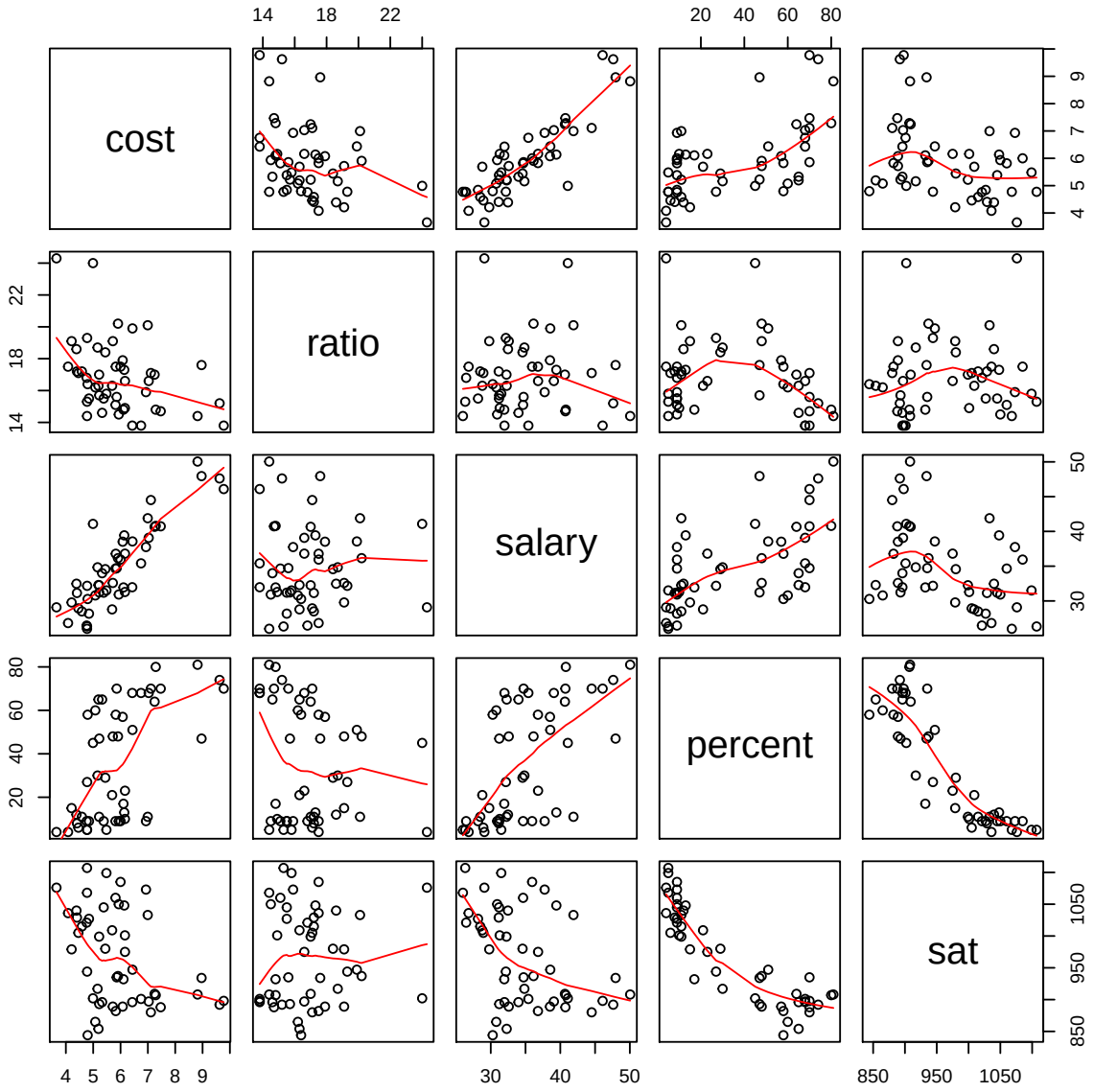

Figure 1: Scatterplot Matrix

From the scatterplot matrix we see:

- reading along the bottom row, we see **sat** vs each predictor.
- percent seems the most important predictor since the points are more tightly cluster around the loess curve. Note that percent is negatively associated with **sat**, so low **percent** translates into high **sat**.
- **ratio** seems like the least important predictor
- panel (5,3), top row and 3rd panel from right, shows **cost** vs **salary** and we see the points are tightly clustered around the loess curve. This indicates a strong relationship between these variables.
- panel (2,3), **percent** vs **salary** shows also that these variables are positively associated.

In summary, the scatterplot matrix suggests **sat** is most closely predicted by **percent** since the data more tightly cluster around the loess curve and that **ratio** is the least important predictor variable.

Variance Inflation Factor (VIF)

The VIF for the design matrix indicates which variables contribute to multicollinearity. Ideally, as in randomized experimental designs, the variables are orthogonal so the X^primeX is a diagonal matrix. In practice this never happens with observation data such as we face with the SAT scores dataset. The VIF for the j-th variable, $j = 1, \ldots, p$, in terms of its coefficient of determination, R_j^2 when it is regressed against all the other input variables. Then the VIF for this variable is $VIF_j = 1/(1 - R_j^2)$. My function **vifx()** defined in the **Rmd** source file provides a simple elegant method for VIF computation. An empoirical rule-of-thumb is that when $VIF_j > 10$ for any $j = 1, \ldots, p$, near multicollinearity is an important consideration. The problem with near multicollinearity is more care is needed in drawing conclusions about which variables are really important.

It is helpful to use a barchart to visualize these VIF's. This is especially useful when there are larger number of input variables but even in this case with only four inputs it is a good idea. The barchart shown below indicates that as might be expected from the scatterplot matrix that **cost** and **salary** are closely related.

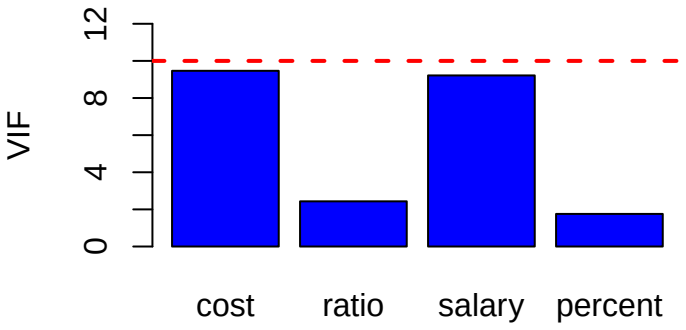

Figure 2: VIF for SAT input variables.

Multicollinearity, especially when only two variables are involved, can sometimes be detected from the correlation matrix. Inspection of the correlation matrix in the table below indicates that **cost** and **salary** are have a correlation 0.8698 which would be considered moderately strong but not hugely impressive since the coefficient of determination for the regression of **cost** on **salary** is only 75.66%.

Table 2: Correlation Matrix

	cost	ratio	salary	percent
cost	1	-0.371	0.870	0.593
ratio	-0.371	1	-0.001	-0.213
salary	0.870	-0.001	1	0.617
percent	0.593	-0.213	0.617	1

Visualizing the correlation matrix as shown in Figure 3 below is often helpful since it makes patterns more apparent. This is especially true when there are more explanatory variables. From Figure 4, the correlation between **cost** and **salary** stands out.

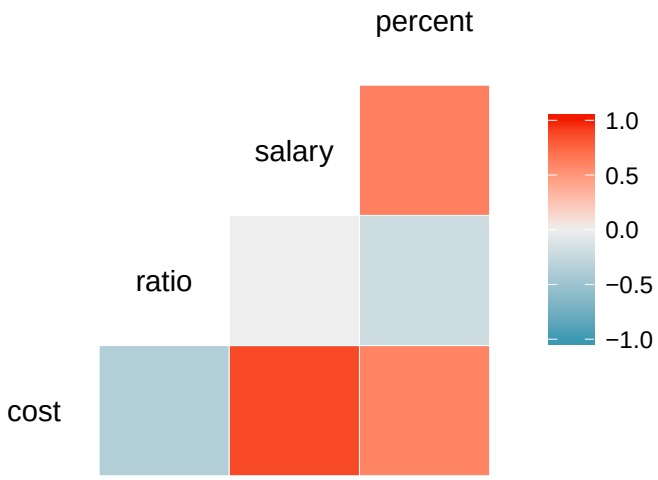

Figure 3: Visualization of the Correlation Matirx

Proposed Model

Statisticians often work closely with subject matter experts. We will suppose that a school administrator suggests that only **cost** and **percent** are really important since **salary** is a closely related to the annual operating cost and **ratio** is not likely important to be important. If we follow this suggestion then we are looking at a regression of **sat** on **cost** and **percent**.

Table 3 summarizes the fitted linear regression sat ~ cost + percent.

Table 3: Model Summary

	Dependent variable:
	sat
cost	12.287^{***}
	(4.224)
percent	-2.851^{***}
	(0.215)
Constant	993.832^{***}
	(21.833)
Observations	50
R^2	0.819
Adjusted R^2	0.812
Residual Std. Error	32.459 (df = 47)
F Statistic	106.674^{***} (df = 2; 47)
Note:	*p<0.1; **p<0.05; ***p<0.01

Both variables are significant at less than 1% and $R^2 = 81.9\%$ so provided the model diagnostic checks are OK this model could be useful for prediction. Model diagnostic checks include the residual diagnostics plus if appropriate the overfitting model diagnostic check.

Basic diagnostic checks

The basic diagnostic checks for the model sat ~ cost + percent are shown in Figure 4.

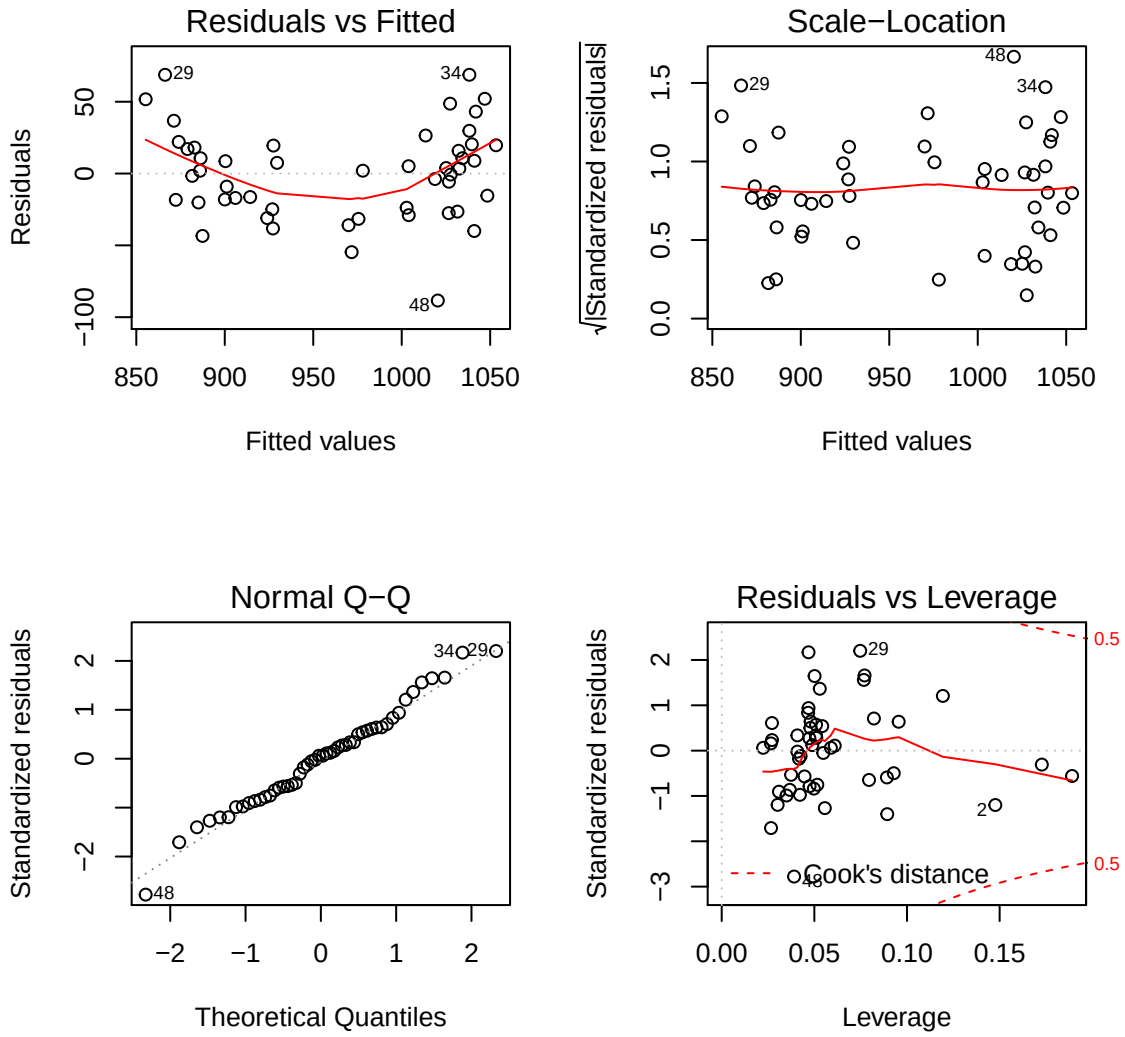

Figure 4: Basic Regression Diagnostic Checks

A problem is indicated in the *Residuals vs Fitted* diagnostic plot since the loess trend is not flat. The curve suggests possible non-linearity due to interaction and some nonlinearity present in the inputs.

Residual dependency checks

We use tidyverse scripting with ggplot2 to produce the residuals dependency plots - see Rmd file for an R script you can use.

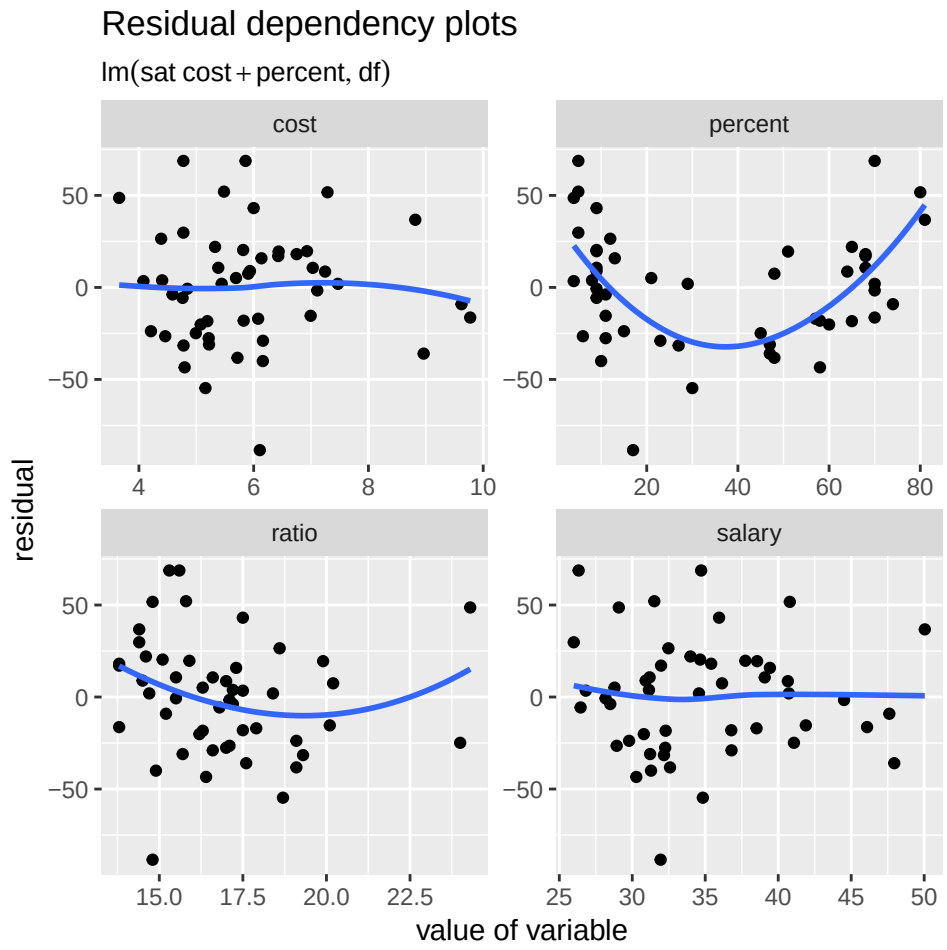

Figure 5: Residual Dependency Diagnostic Checks

The residual dependence plot for **percent** shown in Figure 4 indicates that indeed **percent** exhibits a nonlinear effect.

Improved Model

The diagnostic checks suggest that the previous model may be improved by including a quadratic term with percent so the regression equation could be represented

$$sat = \beta_0 + \beta_1 cost + \beta_2 percent + \beta_3 percent^2 + error.$$

Table 4 summarizes the fitted linear regression. We see that the quadratic term is significant.

Table 4: Model Summary

	Dependent variable:
	sat
cost	7.914**
	(3.498)
poly(percent, 2)1	−509.363***
	(32.734)
poly(percent, 2)2	139.109***
	(26.880)
Constant	919.188***
	(20.985)
Observations	50
R^2	0.886
Adjusted R^2	0.878
Residual Std. Error	26.084 (df = 46)
F Statistic	119.055*** (df = 3; 46)
Note:	*p<0.1; **p<0.05; ***p<0.01

Model is sat ~ cost + poly(percent, 2). The basic regression plots are OK. The model appears to be statistically speaking adequate as far as these plots go.

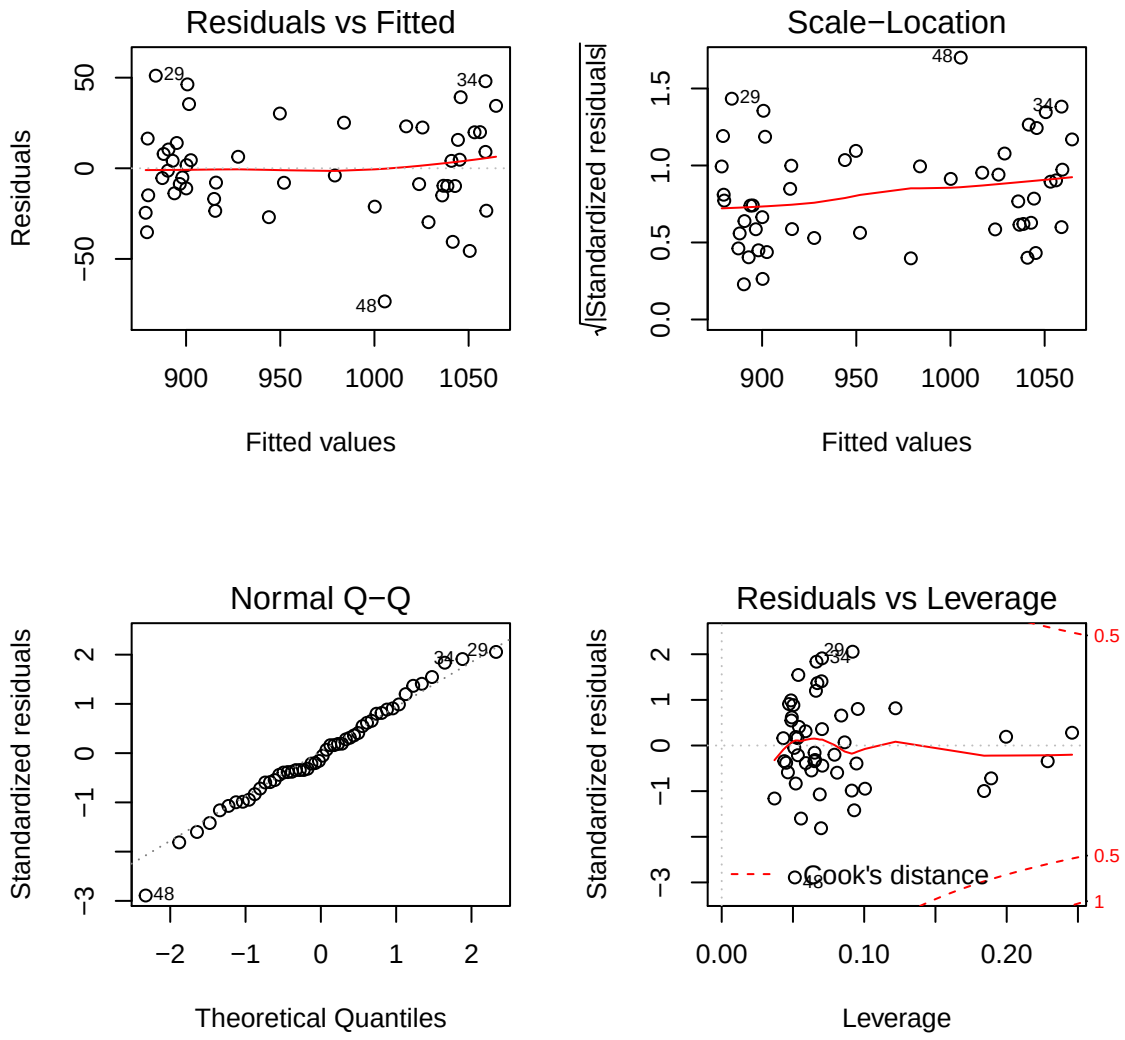

Figure 6: Basic Regression Diagnostic Checks, Revised Model.

But in the revised dependency plot their appears still some curvature in with cost as well salary. But the curvature is weaker and seems influenced by the endpoints – this is a well-known problem with all smoothers. We will consider an enlarged model which allows for quadratic dependence on **salary** in the next section.

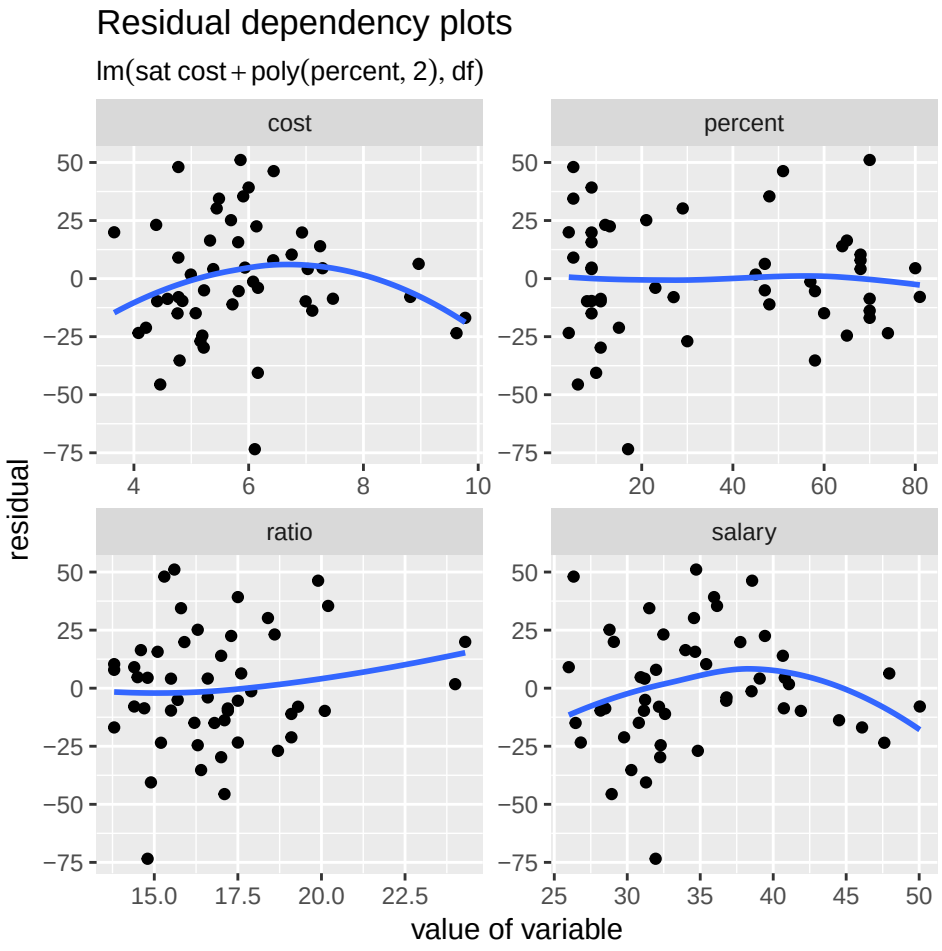

Figure 7: Residual Dependency Diagnostic Checks, Revised Model.

Overfitting lack-of-fit test

As a final check on the model, we try overfitting the model.

First we overfit simply by adding in the inputs **salary** and **ratio**. The enlarged model is shown in the Table 5 and we see that the new variables are not significant at 10%.

Table 5: Model Summary

	Dependent variable:
	sat
cost	9.112
	(8.545)
poly(percent, 2)1	-506.909^{***}
	(35.671)
poly(percent, 2)2	152.085^{***}
	(30.203)
salary	-0.078
	(1.953)
ratio	2.140
	(2.832)
Constant	878.745^{***}
	(45.192)
Observations	50
R^2	0.889
Adjusted R^2	0.876
Residual Std. Error	26.341 (df = 44)
F Statistic	70.266^{***} (df = 5; 44)
Note:	*p<0.1; **p<0.05; ***p<0.01

We may use **extra sum of squares principle** to compare the model in Table 5 (the full model) with the reduced model in Table 4. The resulting value of the F-statistic is 0.5528975 on (2, 46) DF which implies a two-sided p-value about 57.9% so it is not significant at 10%. We accept the simpler model in Table 4.

Next we overfit by also including a quadratic term in **salary** so our fitted model is sat ~ cost + poly(percent, 2) + poly(salary, 2) + ratio. The ANOVA lack-of-fit test comparing this model with the simpler model sat ~ cost + poly(percent, 2) + poly(salary, 2) + ratio has a two-sided p-value about 30.3%. So the simpler model is not rejected at 10%. We may attribute the apparent curvature in the plots shown in Figure 7 to randomness since the evidence suggests it is not important.

Conclusion

This example shows the iterative model building technique for linear regression. This technique is widely used in statistical practice. It contrasts with the Machine Learning approach which focuses less on obtain a statistical or mathematical model and more on finding a reasonable prediction algorithm. Both the algorithmic and model building have their advantages and disadvantages so neither method is in a general sense better than the other.

The report also demonstrates how a beautiful PDF report may be produced using R/RStudio. Sometimes the reports like this are called **dynamic documents** since the report is easy to update if new data becomes available. Another important advantage of this type of report is reproducibility. Since it is easy for the reader to verify and check the computations and exact assumptions made in the claims reported. It is easy for errors, either blunders or more subtle biases/assumptions, to creep in and so reproducibility is often critical in the worlds of Science and Business.