Multiple Linear Regression Diagnostics and Transformations

Laura Barnes & Julianne Quinn

University of Virginia
Charlottesville, VA

Agenda

1. Review of Multiple Regression
2. Model Diagnostics
   - Graphical Diagnostics
   - Analytical Diagnostics
3. Transformation of Response Variables and Predictors

Review of Multiple Regression

- Multiple Regression: A method for measuring and modeling the relationship between sets of variables.
- Multiple Linear regression :

$$y = f(X) + \epsilon = X\beta + \epsilon$$

- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of square errors:

$$\text{minimize } (X\beta - y)^T(X\beta - y)$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\hat{\sigma}^2 = \frac{(X\hat{\beta} - y)^T(X\hat{\beta} - y)}{n-k-1} = \frac{\hat{\epsilon}^T\hat{\epsilon}}{n-k-1}$$

Sum of Squares Decomposition, F-Test, t-tests

- The sum of squares has a convenient decomposition:

$$\text{Total S.S.} = \text{Residual S.S} + \text{Model S.S}$$

- Model Utility Test:

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$

$$H_A : \beta_i \neq 0, i \in \{1, \cdots, k\}$$

- t-tests:

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

Metrics of Models

- Multiple Coefficient of Determination: R^2
- Adjusted-R^2
- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Mallows C_p

Variable Selection

- Importance of variable selection
- Automated Selection:
  - Forward selection
  - Backward selection
  - Stepwise regression
- Partial F Test:

$$H_0 : \beta_i = \cdots = \beta_n = 0$$

$$H_A : \beta_i \neq 0, i \in \{i, \cdots, n\}$$

Model Comparison

- Test sets:
  - In using test sets we sample a subset (without replacement) of the data and do not use it to build the model. We use this subset for testing.
  - What criterion do we use to choose among the models? Predicted Mean Squared Error (PMSE)
  - Because test sets are generated by sampling, you will see different PMSE on different runs. Therefore, it is better to generate paired groups of PMSE and then use statistical tests to see whether the difference between models is significant.

Model Comparison

- Cross-Validation:
  - In cross validation we start by sampling without replacement to get k parts, called folds.
  - We build a model on $k-1$ of the folds and test on the remaining fold. We repeat this using each fold as a test set for the other $k-1$ folds.
  - We use all the data points for training and testing.
  - Example:
    - Folds $k = 10$; models in comparison:

$$EQPDMG \sim TEMP + TRNSPD + TONS + CARS$$
$$EQPDMG \sim TEMP + TRNSPD$$

    - CV Results:

	MSE
Model 1	54788955566
Model 2	558409204733

Regression Assumptions

- Recall our goal: determine if there is relationship between variables.
- Regression allows us to test hypotheses about relationships between the predictor variables and the response variable. What tests do we use?
- Regression allows for association tests while controlling for the values of other variables in the equation (e.g., ACCDMG vs. TONS and TRNSPD)
- To gain these powerful results regression requires certain assumptions:
  - Independent, identically distributed Gaussian errors with zero mean and constant variance;
  - Linearly independent predictor variables; and
  - Correct linear model for the response.

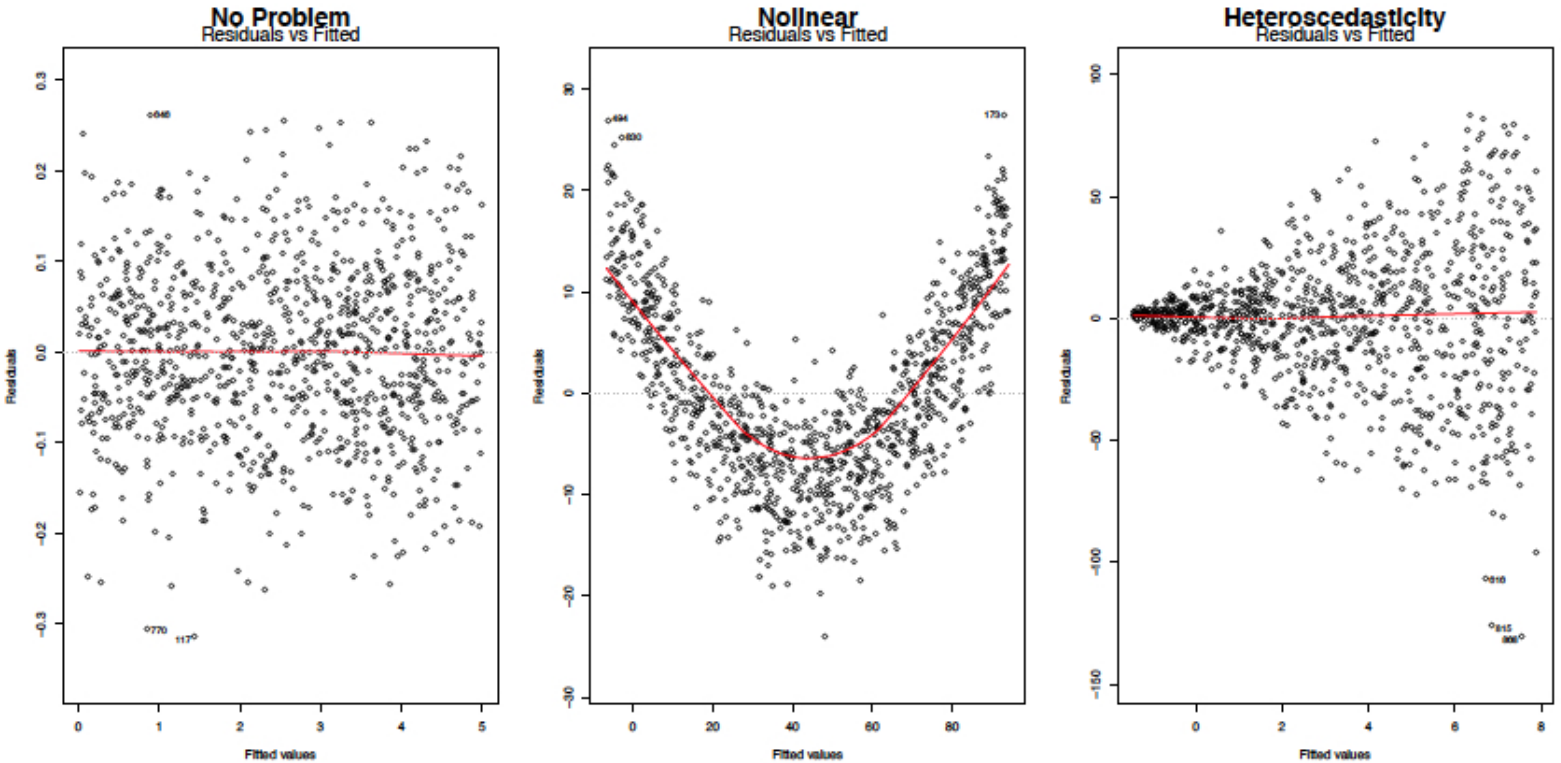
Diagnostic Plots

- We can use diagnostic plots to exam how well those assumptions are satisfied.
- Four common diagnostic plots:
  - Residuals vs. fitted
  - Scale-Location plot of square root of absolute standardized residuals vs. fitted
  - QQ plot of standardized residuals
  - Residual-Leverage plot

Residuals vs. Fitted

- The most important and useful plot.
- You look for
  - Absence of patterns in a symmetric display around zero.
  - Constant variance.
- Problem if you see:
  - A relationship between the residuals and the fitted values. This indicates a lack of fit for the model.
  - Changing variance which means you need to transform your response or explicitly account for the variance in the model.
  - Extreme, influential, or outlying points. Investigate these and make a decision on their influence. Could be lack of fit, heteroscedasticity, both, or neither.

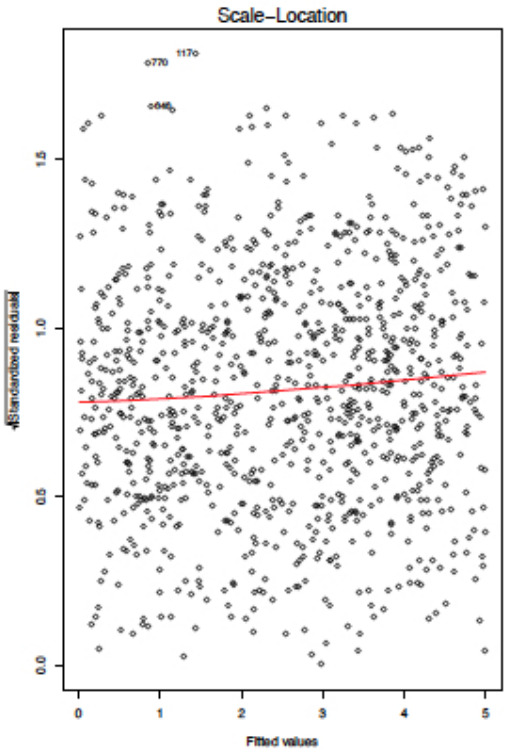
Examples of Residual vs. Fitted Plots

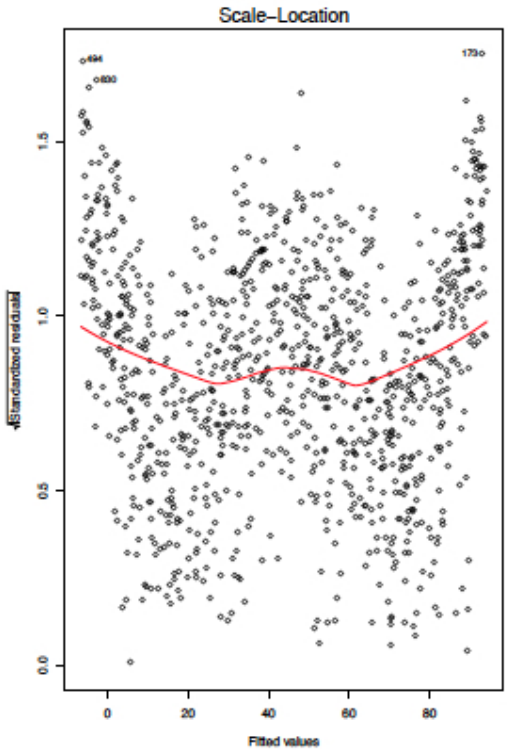


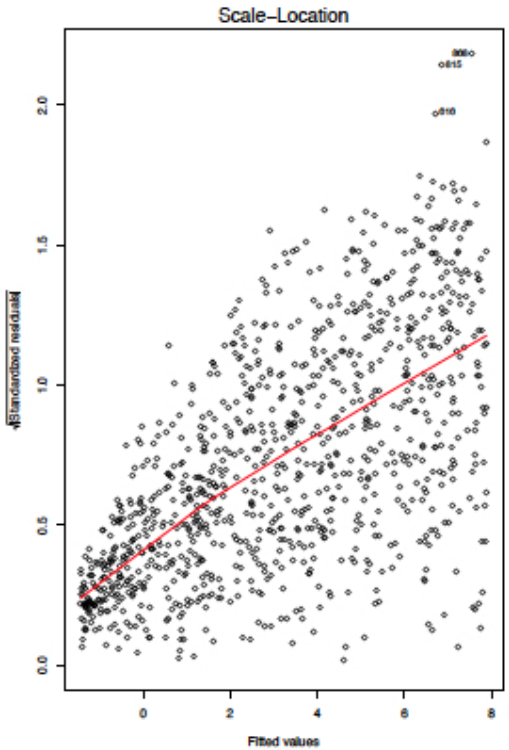
Scale-Location Plot

- Similar to residual vs. fitted, so look for the same things.
- S-L plots are less affected by skew.

Examples of Scale-Location Plots



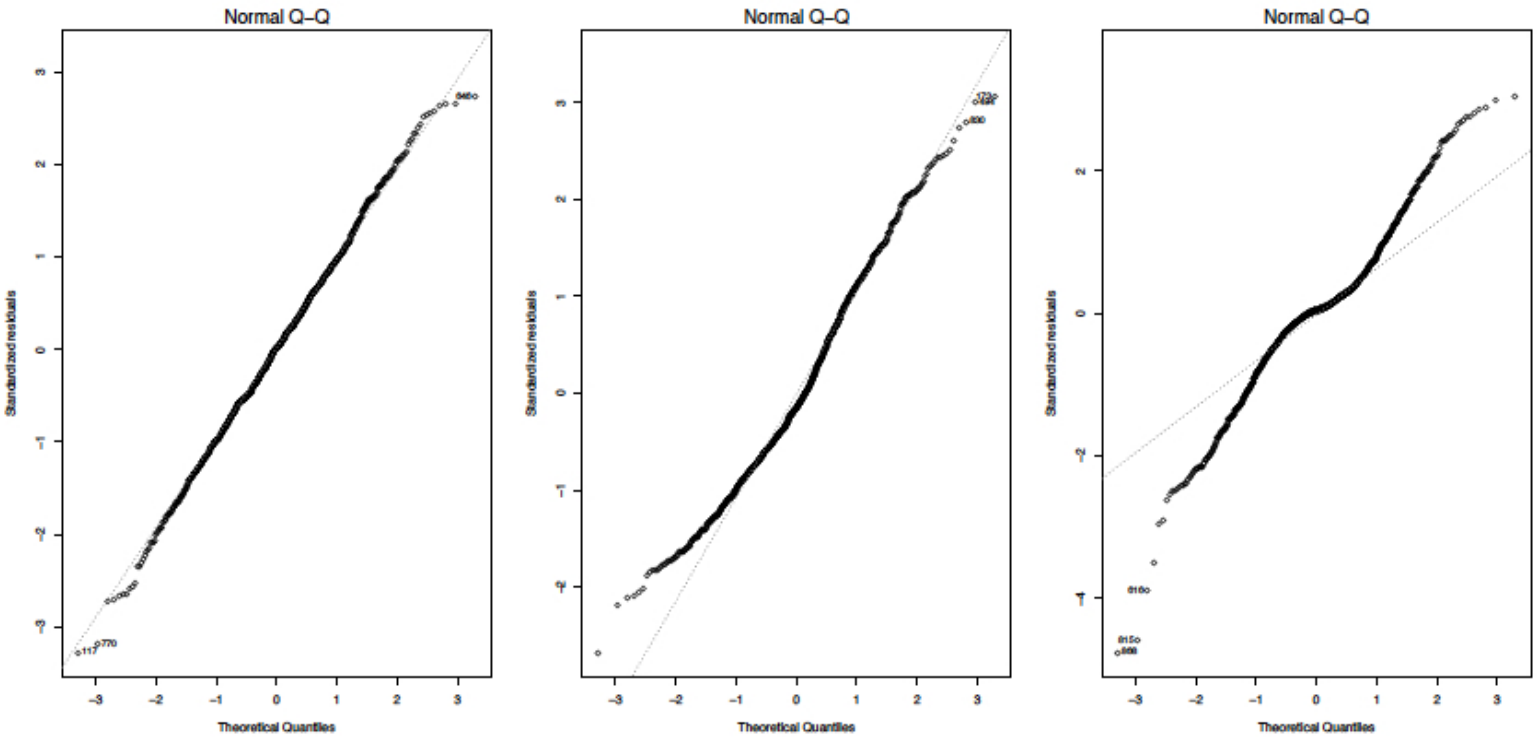




QQ Plot of Standardized residuals

- What is QQ plot?
- QQ diagnose failure of the Gaussian assumption for the error term.
- Residuals are the estimated errors.
- Failure is indicated by points far from the line.

Examples of QQ Plots



Influential Points

- Recall that $\hat{\beta} = (X^T X)^{-1} X^T y$
- Some data points might influence $\hat{\beta}$ more than the others.
- Influential points can change the values of parameter estimates and other results by their removal.
- Outliers do not fit the model well.
- Observations can be either, neither, or both.
- Influential observations can come from a lack of fit, heteroscedasticity, data entry errors, etc.
- To discover influential points and outliers we need a measures of influence: combine leverage and standardized residuals.

Leverage and Cook's Distance

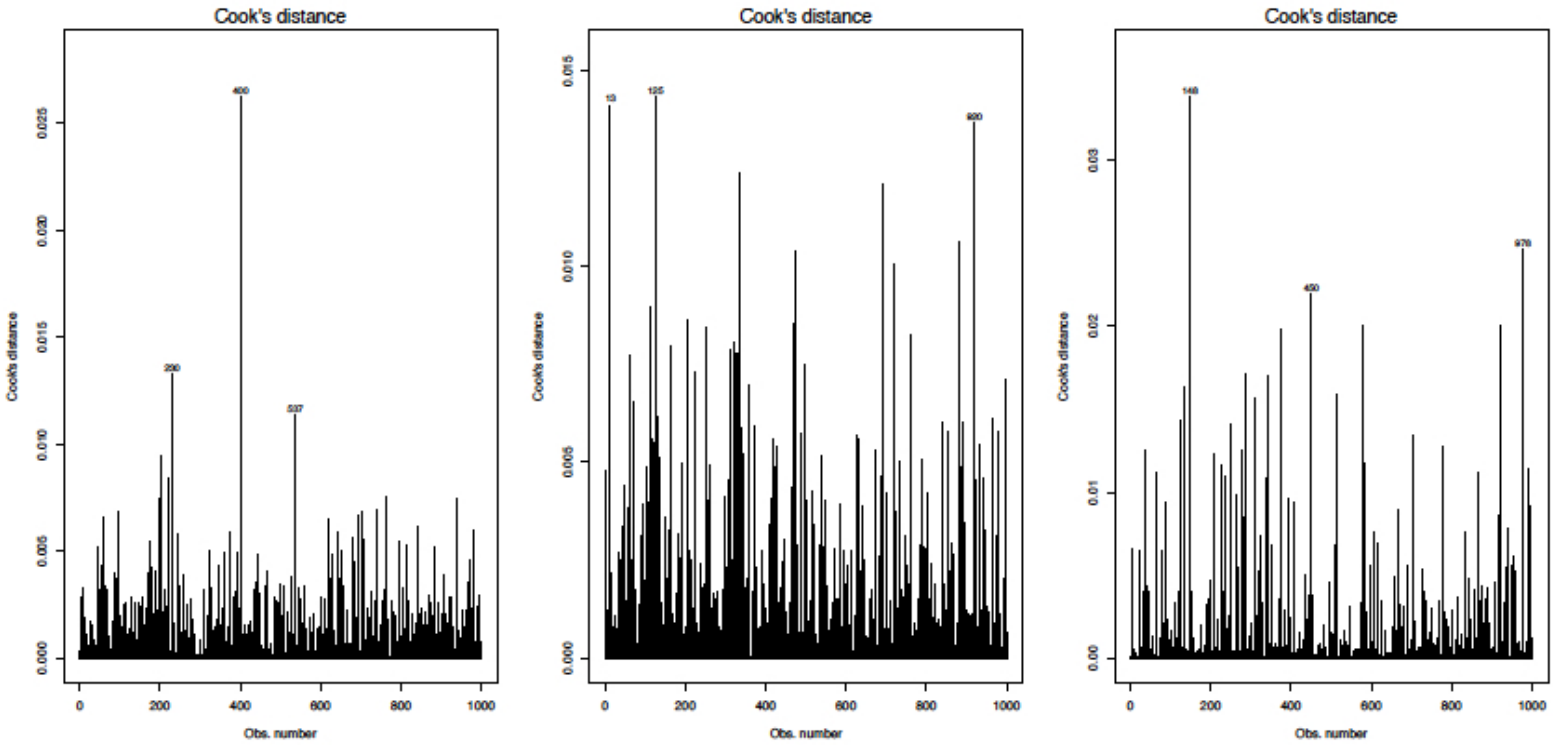
- Leverage are the diagonal entries in H or $H_{ii} = h_i$
- Observations with high leverage are far away from the average of predictor values.
- Cook's distance attempts to identify influential observations using leverage.
- Cook's distance is calculated from the leverage and standardized residual of an observation:

$$C_i = r_i^2 \frac{h_i}{p(1 - h_i)}$$

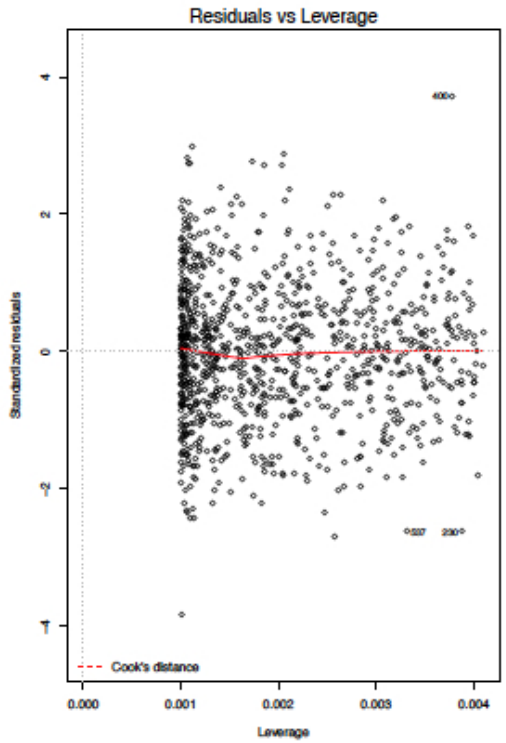
where r_i is the standardized residual, and p is the number of parameters.

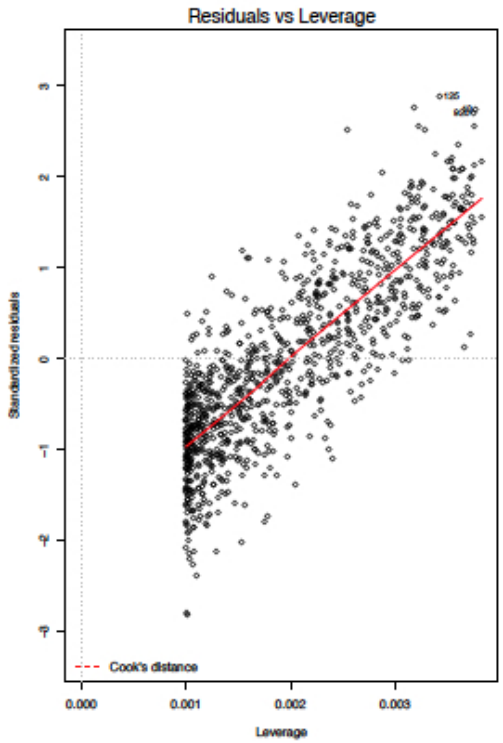
- What does a large Cook's distance mean?

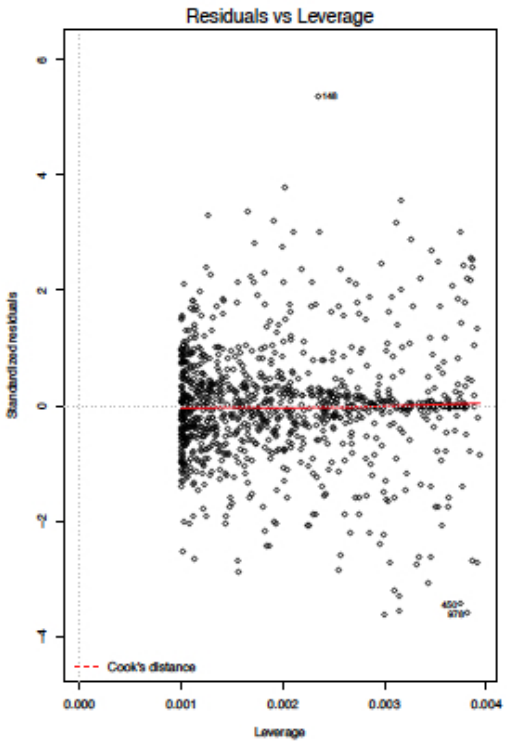
Examples of Cook's Distance Plots



Examples of Residual-Leverage plots

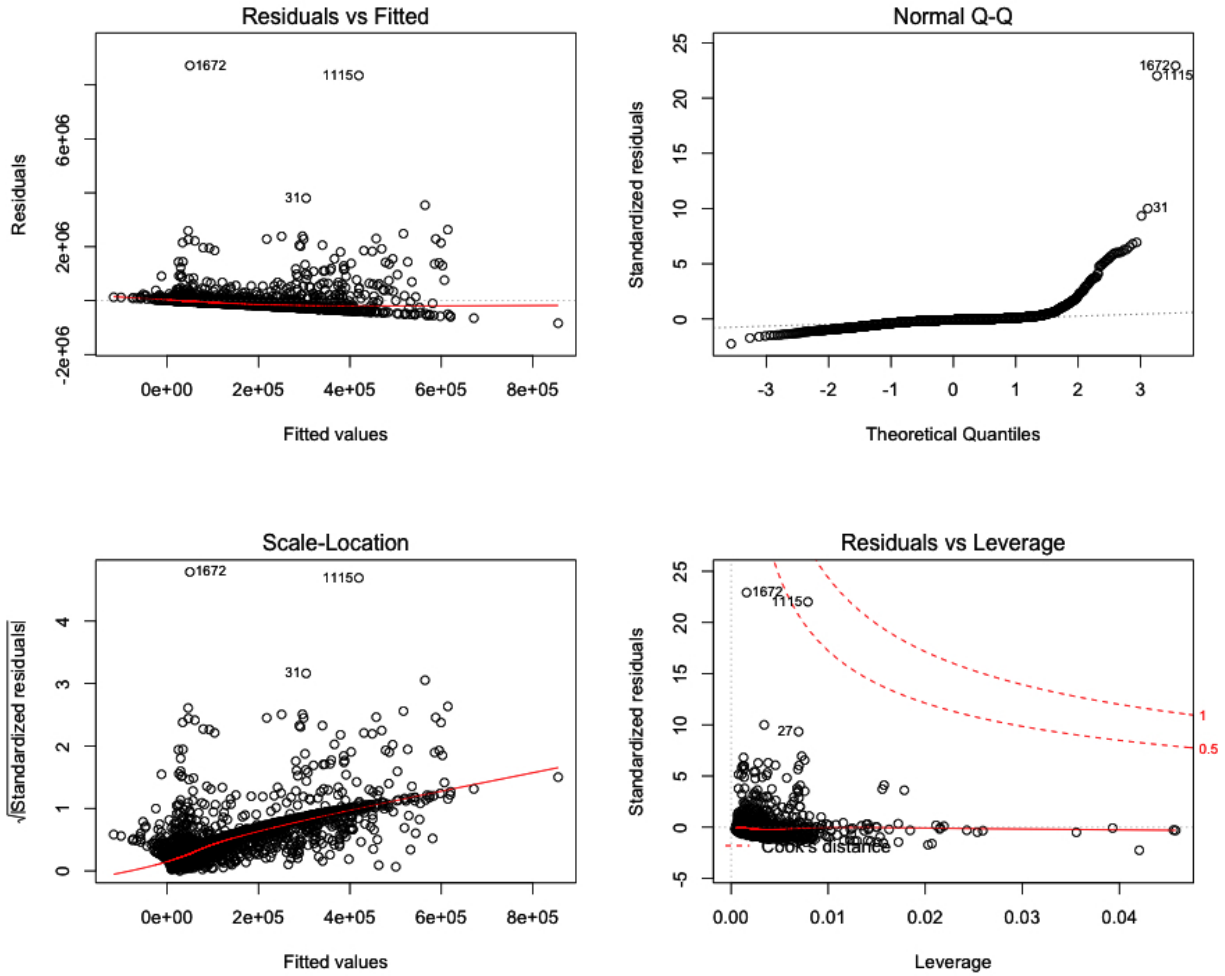






Example of Diagnostic plots

Model: ACCDMG $\sim$ TEMP+TRNSPD+TONS+CARS+HEADEND1



Diagnosis from Regression Results

- We can also diagnose problems in the regression by analyzing the results.
- Coefficient values can reveal problems.
- Correlations between predictors and between residuals can reveal problems.
- Performance tests for the model can reveal problems.

Diagnosing Multicollinearity with Coefficient Values

- Multicollinearity means the assumption of linearly independent predictors is violated.
- If it is violated the resulting regression function is unstable. This means small changes to values of the observations will cause large changes in the response.
- Multicollinearity is apparent in the flipped signs for coefficients and for large changes in coefficient values with small changes to values in the observations.

Simpson's Paradox: 1973 Berkely Gender Bias Case

Gender	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Simpson's Paradox: 1973 Berkely Gender Bias Case

- A closer look...

Dept.	Men Applicants	Men Admitted	Women Applicants	Women Admitted
1	825	62%	108	82%
2	560	63%	25	68%
3	325	37%	593	34%
4	417	33%	375	35%
5	191	28%	393	24%
6	272	6%	341	7%

Diagnosing Simpson's Paradox with Coefficient Values

- Coefficients show contradictory patterns.
- Adding variables or replacing variables causes major changes to coefficient values.

When the Tests Fail

- Transform the response;
- Transform the predictors;
- Add new variables (i.e., new models) for Simpson's paradox and lack of fit;
- Remove or combine variables for multicollinearity;
- Principal components regression or ridge regression for multicollinearity; and
- Use robust methods for regression.

Response Transformations: Box-Cox Plots

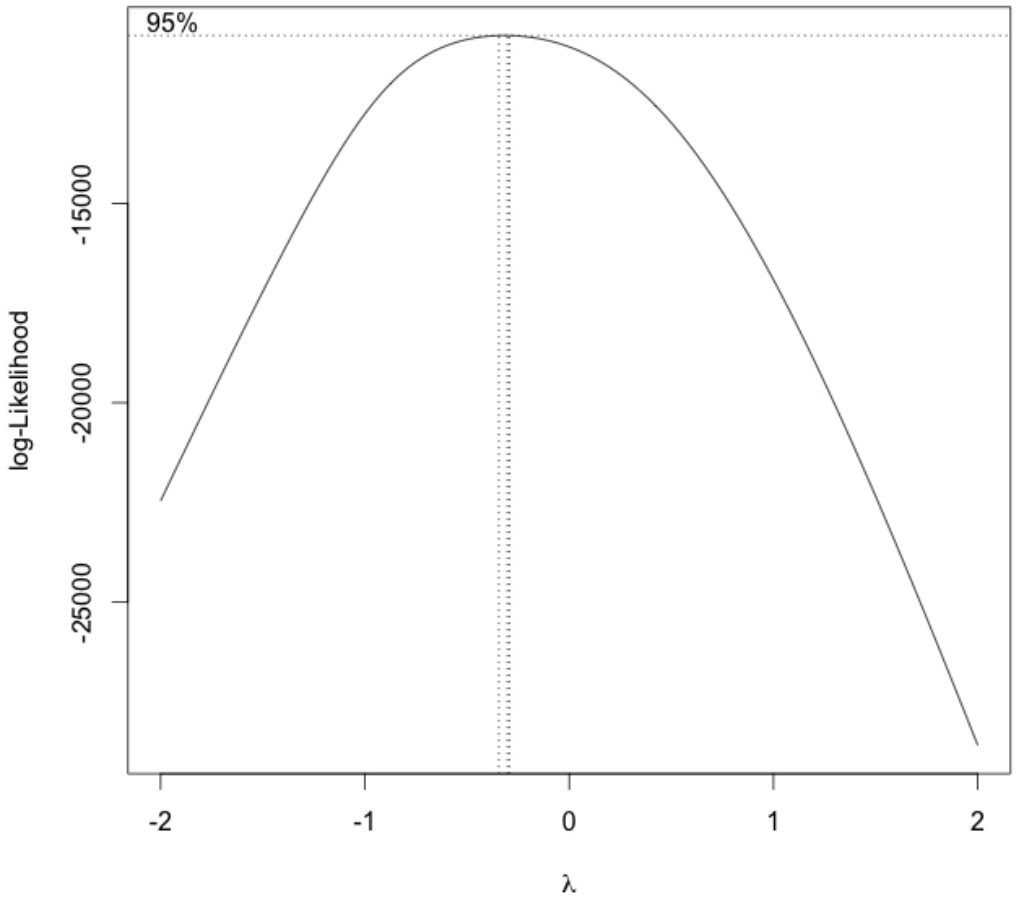
- When lack of Gaussian errors or lack of fit are detected consider transformation of the response. This can be helped with Box-Cox plots.
- Box-Cox plots allow us to consider power transformations of the form:

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$

- The plot shows the 95% confidence interval for the λ. Look for the integers that this interval brackets. If it is around 1 what do we do? No transformation.
- If it is around zero what do we do? Log transformation.

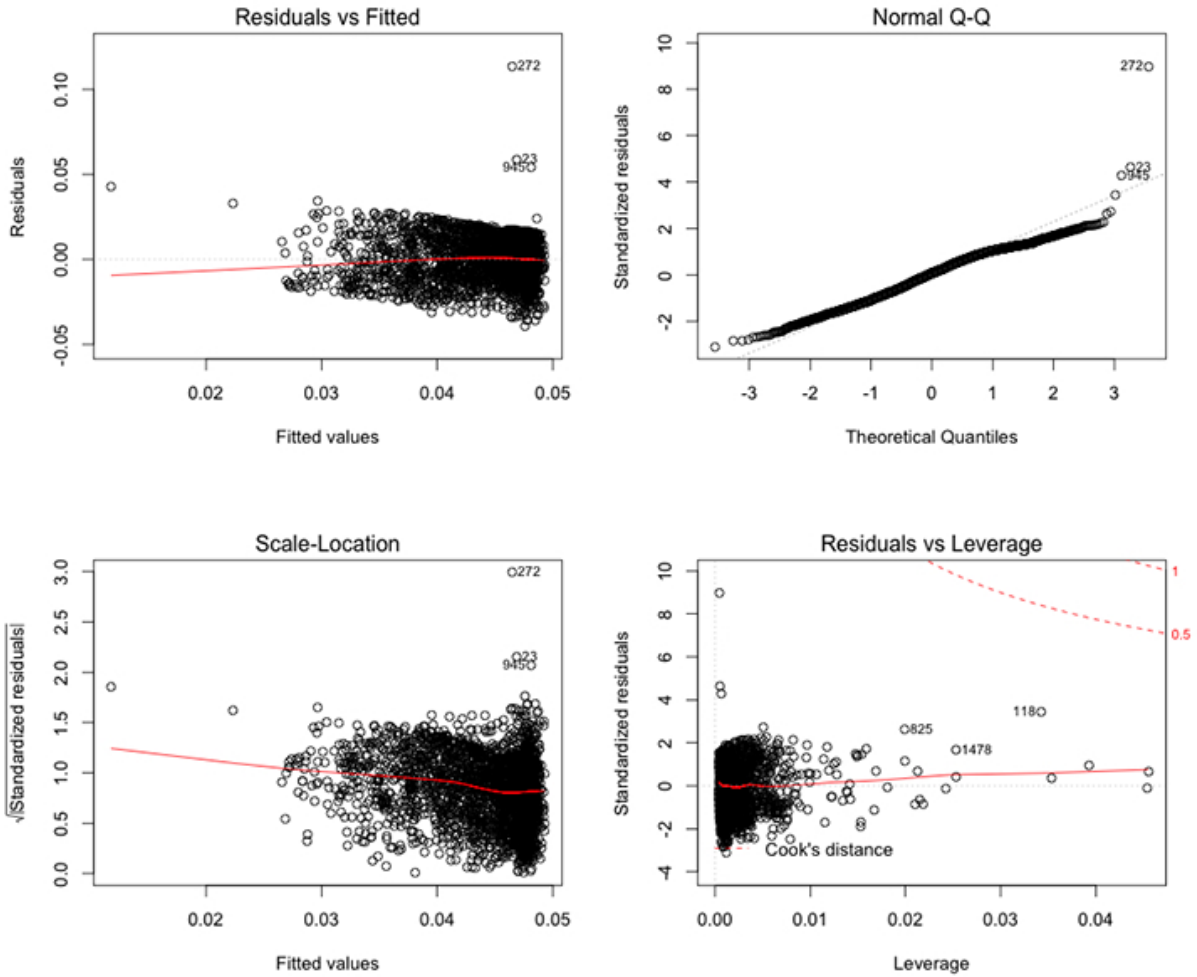
Example of Box-Cox Plot

Model: ACCDMG TEMP+TRNSPD+TONS+CARS+HEADEND1



Example of Diagnostic Plots: Log transformation

Model: log(ACCDMG) $\sim$ TEMP+TRNSPD+TONS+CARS+HEADEND1



Predictor Transformations

- When lack of fit is detected the consider transformations of the predictors.
- Predictors need to cover the range of desired values.
- Log and other transformations can help reduce the effects of skewness.
- What kinds of transformations can we do on predictors? Any