Model Diagnostics

Prof. Lauren Perry

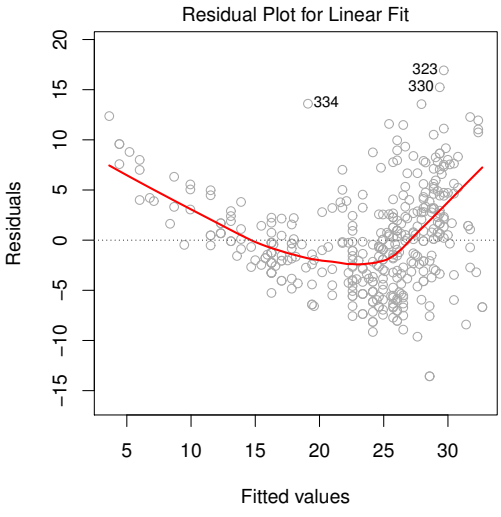
Potential Problems

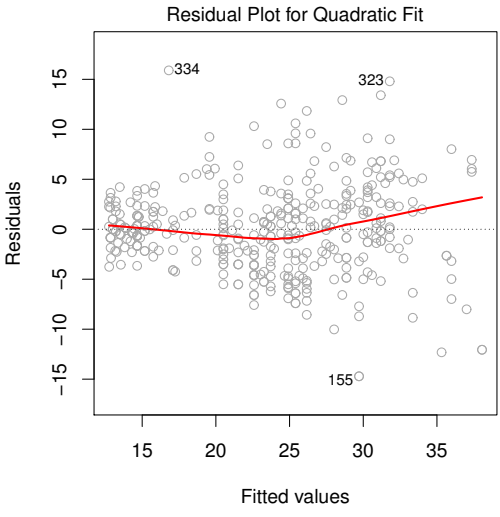
1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers and high-leverage points.
5. Collinearity.

1. Non-linearity of the response-predictor relationships.

- We can examine non-linearity using *residual plots*.
- Ideally, these will show no discernible pattern (random scatter).
- We can work on fixing this problem by transforming the predictors:
  - Ex: $\log X$, $\sqrt{(X)}$, X^2, etc.

Example Residual Plots Showing Non-Linearity





2. Correlation of Error Terms

Assumption: error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are uncorrelated.

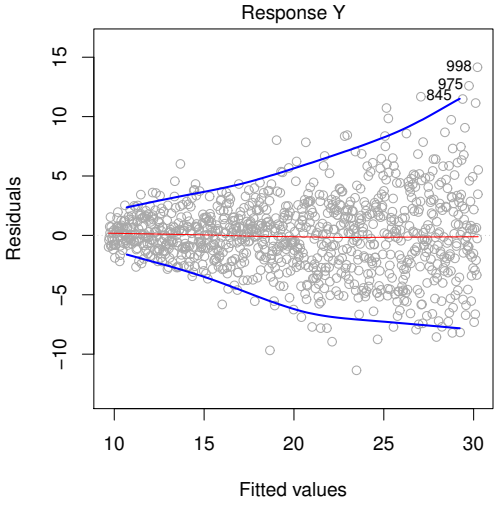
- That is, knowing something about ϵ_i, doesn't tell us anything about ϵ_{i+1}.
- Our standard error calculations rely on this.
  - Violations tend to result in std error being underestimated.
  - This causes erroneously narrow confidence/prediction intervals.
- These correlations can occur for data that is *time dependent*.
  - We should use different modeling techniques for this type of data.

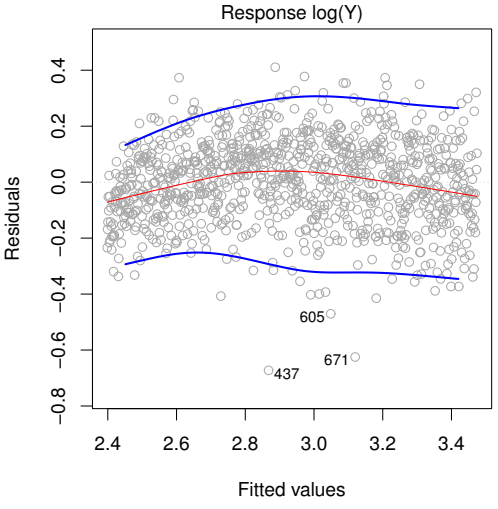
3. Non-constant variance of error terms.

Assumption: error terms have constant variance, $\text{Var}(\epsilon_i) = \sigma^2$.

- We can check for homoscedasticity using residual plots.
- There should be no discernible pattern in the variability.
- Standard errors rely on this assumption.
- This assumption is often violated, but we can usually fix (or at least improve) it!
- We work on fixing this problem by transforming the outcome variable:
  - Ex: $\log Y$, $\sqrt{(Y)}$, Y^2, etc.

Example Residual Plot - Before and After log(Y) Transformation





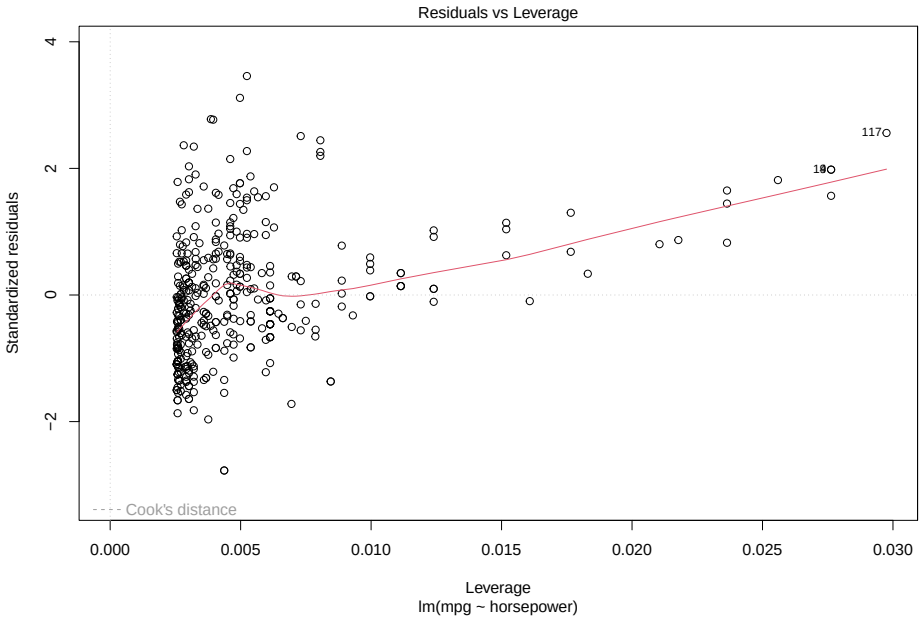
4. Outliers and High-Leverage Points

An *outlier* is a point for which y_i is far from the value predicted by the model.

- If we think the outlier resulted from an error in data collection, we can remove it.
- ... but there is nothing inherently wrong with outliers.

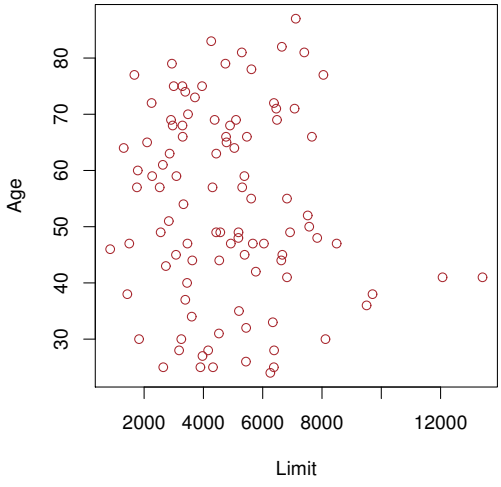
From a model fitting perspective, we are much more interested in *high-leverage points*.

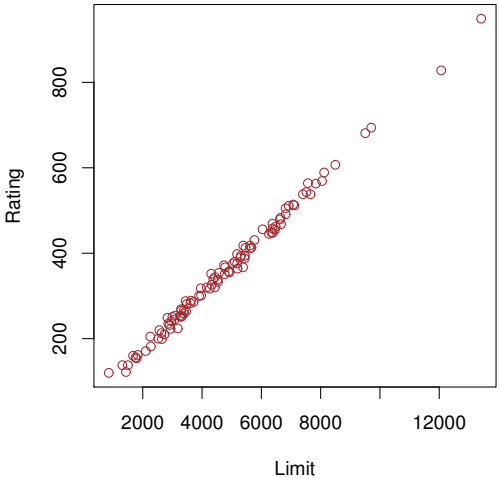
- These are observations which have a significant individual impact on the regression line.
  - We can examine this by removing a point from the data and refitting the model, and then examining how much the regression line changed.


Residuals vs Leverage
Standardized residuals
4
2
0
−2
117
14
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(mpg ~ horsepower)

6. Collinearity

Collinearity is the situation in which one or more predictor variables are closely related to one another.





Collinearity

When two variables are *collinear*:

- It can be difficult to separate out their individual effects on the response.
- The accuracy of regression coefficient estimates is decreased.
- Standard error is increased, which shrinks the test statistics (toward 0).
  - This results in larger p-values and potentially a failure to reject H_0.

Dealing with Collinearity

- Examine the correlation matrix for the predictors.

.	mpg	cyl	disp	hp	wt	acc	yr	orgn
mpg	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cyl	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35	-0.57
disp	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37	-0.61
hp	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42	-0.46
wt	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31	-0.59
acc	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29	0.21
yr	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00	0.18
orgn	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1.00

Multicollinearity

Sometimes, we can run into collinearity between three or more variables that will not appear in the two-way correlations shown in the correlation matrix.

- To examine possible multicollinearity, we compute the *variance inflation factor* (VIF).
  - This is the ratio of (variance of $\hat{\beta}_j$ when fitting the full model) to (the variance of $\hat{\beta}_j$ if fit on its own).
  - The minimum value for VIF is 1.
  - There are different ideas for what constitutes a "high" VIF, but people often use 5 or 10.

Multicollinearity

```
cylinders    displacement    horsepower
10.737535    21.836792       9.943693
weight       acceleration    year        origin
10.831260    2.625806        1.244952    1.772386
```

- Now what?

Multicollinearity

Let's try removing the variable with the highest VIF:

cylinders	horsepower	weight
6.008253	9.088413	9.219674
acceleration	year	origin
2.598356	1.239409	1.594220

- Notice that removing `displacement` also slightly improved the VIF for all the other variables!
- At this point, we can stop (if we're using 10) or try removing another variable.

Multicollinearity

Let's try removing one more variable (`displacement`):

```
##    cylinders   horsepower acceleration         year       origin
##     4.155143     5.323311     1.996560     1.209909     1.495100
```

- That made a big difference!

The Final Model

```
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)    -7.87876    5.05154  -1.560     0.12
cylinders      -1.22202    0.22524  -5.425 1.02e-07 ***
horsepower     -0.08815    0.01130  -7.802 5.75e-14 ***
acceleration   -0.40305    0.09654  -4.175 3.69e-05 ***
year            0.66601    0.05628  11.833  < 2e-16 ***
origin          1.82772    0.28612   6.388 4.84e-10 ***
---
Residual standard error: 3.727 on 386 degrees of freedom
Multiple R-squared:  0.7749,   Adjusted R-squared:  0.772
F-statistic: 265.7 on 5 and 386 DF,  p-value: < 2.2e-16
```