STAT 420: Methods of Applied Statistics

Model Diagnostics — Transformation

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

Department of Statistics
University of Illinois at Urbana-Champaign
March 25, 2017

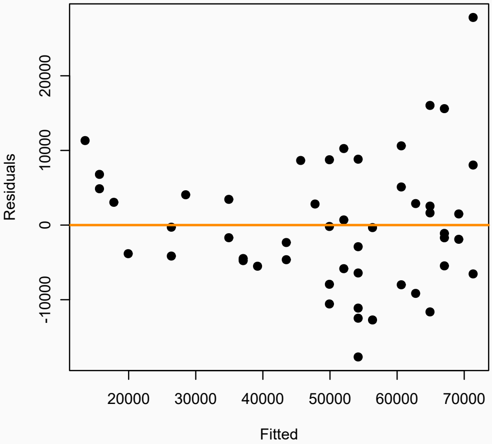
Improving the model fitting

- We discussed three model diagnostics: Normality, Constant variance and Outliers.
- In the next stage, we will discuss
  - Linear relationship between predictors and dependent variables
  - Highly correlated predictors

Transformation

- Recall that when we used the BP test or the white test, the purpose was to test whether the variance are constant.
- In many real data examples, the variance increases (or decreases) as the fitted value changes.
- Load the initech data and fit linear regression

A picture


Residuals
-10000
0
10000
20000
20000
30000
40000
50000
60000
70000
Fitted

Variance stabilizing transformations

- The constant variance reassumption requires that

$$\text{Var}[Y|X = x] = \sigma^2$$

  regardless of the value of x.
- However, in the initech data, we see that the variance is a function of the mean

$$\text{Var}[Y|X = x] = h(\mu)$$

- It looks like h is some increasing function here. In order to correct for this, we would like to find some function of Y, namely $g(Y)$, such that

$$\text{Var}[g(Y)|X = x] = c$$

$\log$ transformation

- Usually when the outcome $\mathbf{y}$ are all positive, the $\log$ transformation can help to stabilize the variance
- In $\mathbb{R}$, the function $\log()$ is the natural log by default, and I will also use this instead of $\ln()$.
- Consider the new model:

$$\begin{aligned} & \log(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i \\ \Longrightarrow \qquad & Y_i = \exp(\beta_0 + \beta_1 X_i) \cdot \exp(\epsilon_i) \end{aligned}$$

- Hence the scale of Y is larger when the mean is larger
- Note: when the outcomes contains 0 or negative values, we could add a constant to all subjects and force them to be positive. Usually we add 1 if the minimum of y's is 0.

log transformation

```
> logfit = lm(log(salary) ~ years, data = initech)
> summary(logfit)

Call:
lm(formula = log(salary) ~ years, data = initech)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.841325   0.056355  174.63   <2e-16 ***
years       0.049978   0.002868   17.43   <2e-16 ***
---
Signif. codes:  0 ?***?0.001 ?**?0.01 ?*?0.05 ?.?0.1 ? ?1

Residual standard error: 0.1541 on 48 degrees of freedom
Multiple R-squared:  0.8635, Adjusted R-squared:  0.8607
F-statistic: 303.6 on 1 and 48 DF,  p-value: < 2.2e-16
```

Interpretation of Coefficients after ln Transformation

- After the transformation, the $\boldsymbol{\beta}$ parameters have different meaning.
- Consider one unit change in X, we have

$$
\begin{aligned}
Y^* &= \exp(\beta_0 + \beta_1(X_i + 1)) \cdot \exp(\epsilon_i) \\
&= \exp(\beta_1) \cdot \exp(\beta_0 + \beta_1 X_i) \cdot \exp(\epsilon_i) \\
&= \exp(\beta_1) Y
\end{aligned}
$$

- So $\exp(\beta_1)$ represents the factor by which Y increases for 1 unit increase in X.

Box-Cox Transformation

- Choosing the transformation can be subjective. Box and Cox (1964) proposed a method for finding the appropriate y transformation:

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

- We want to choose λ to maximize the log-likelihood:

$$L(\lambda) = -\frac{n}{2} \log(\sigma_\lambda^2) + (\lambda - 1) \sum_i \log(y_i)$$

  where σ_λ^2 is the MSE if the transformation g_λ is used.
- The second term is derived from the Jacobian of the transformation so that the log-likelihood is comparable to the original scale.

Calculating the Box-Cox yourself

```
> # box-cox transformation
> library(MASS)
> bc = boxcox(fit, plotit = T)
> # calculate your own box-cox transformation
> y = initech$salary; x = initech$years; n = nrow(initech)

> y = y/exp(mean(log(y))) # this is for numerical stability

> lambda = 0.38383838 # The 60th lambda value in the boxcox output
> gy = (y^lambda - 1) / lambda # the transformation

> # the second term is not needed since now sum(log(y)) = 0
> LL = -n/2*log(sum(lm(gy ~ x)$residuals^2))
> LL
[1] -4.570907
> # compare this to the boxcox function
> bc$y[60]
[1] -4.570908
```

Box-Cox Transformation

- Compare the log-likelihood of λ at 0.5 and -0.5, which one is better?

A : $\lambda = 0.5$ B : $\lambda = -0.5$ C : They are the same.

Transforming the X variables

- Transforming the X variables is even more common, and it also deals with the issue that some observations are two influential due to extreme X values (recall the Cook's distance)
- In a lot of situations, the $\log$ transformation is used because it effectively takes care of the extreme values.
- For example, the ATT data. See the R code.
- The $\log$ transformation explicitly requires the original variable to be positive. Sometimes a constant is added to the original scale to force them to be positive. This applies to both X and Y.

Transforming the X variables

- In some other situations, transforming the X variables (or including higher order terms) will allow us to model non-linear relationships and interactions.
- For example we would like to fit this model to the marketing dataset

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

  which will describe the decreasing effect of advertisement on the sales.
- See R code. Some useful functions: poly , polym
- Be careful that using high order polynomial may overfit the data.