

Lecture 8: Variable Transformations

Nancy R. Zhang

Statistics 191, Stanford University

February 6, 2008

Announcements

- Homework 1 Graded.
- Grading questions: This week only Sunny Kim OH Thus 4-5 PM Sequoia Hall Girshick Library.
- Show all of your R code and output in homeworks!
- Last year's midterm online.
- Next Monday: Review.
- Additional OH W 4:30-5:30 PM.

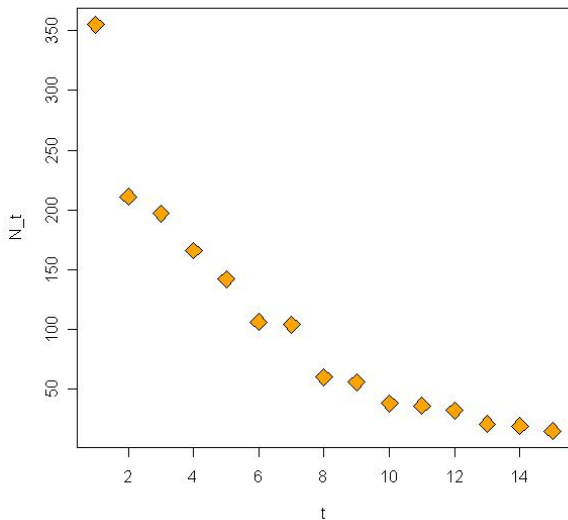
Linear regression (ANOVA) model:

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_p X_p + \text{error},$$
$$\text{error} \sim N(0, \sigma^2).$$

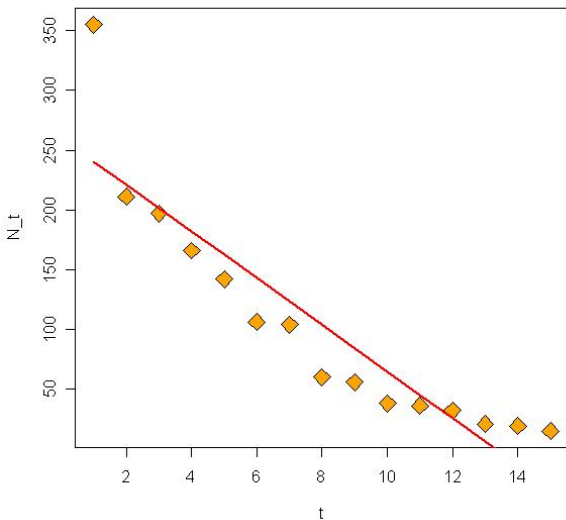
- 1 Mean depends on predictors in a *linear* way.
- 2 Error is Gaussian.
- 3 Variance is constant.
- 4 Variance is independent.

When these assumptions are violated, linear Gaussian models can *sometimes* still apply after transforming the variables.

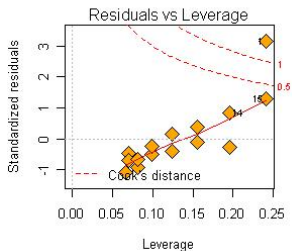
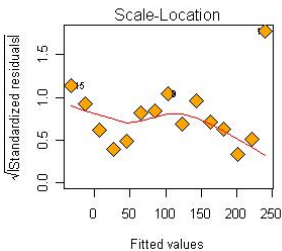
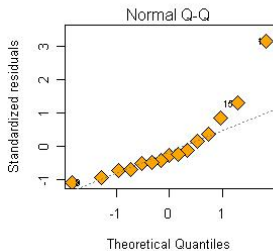
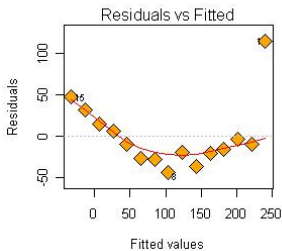
Experiment: Number of surviving marine bacteria following exposure to X-rays.



Experiment: Number of surviving marine bacteria following exposure to X-rays.



Trend visible in residual



plots.

Exponential growth (decay) model

- Suppose the expected number of cells grows like

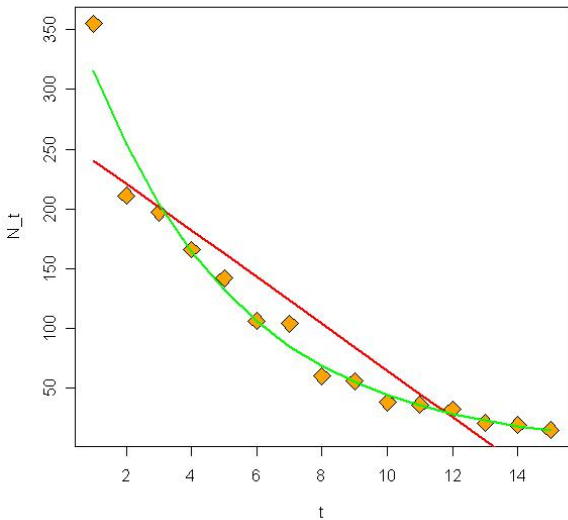
$$E(n_t) = n_0 e^{\beta_1 t}, \quad t = 1, 2, 3, \dots$$

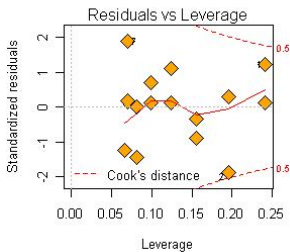
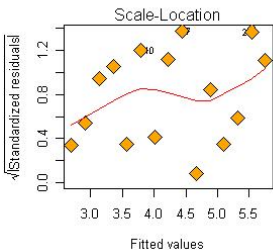
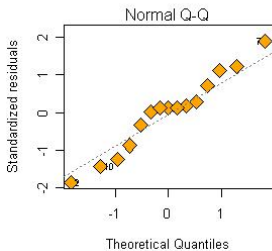
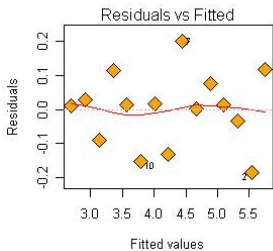
- If we take logs of both sides

$$\log E(n_t) = \log n_0 + \beta_1 t.$$

- (Reasonable ?) model:

$$\log n_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2) \text{ independent}$$



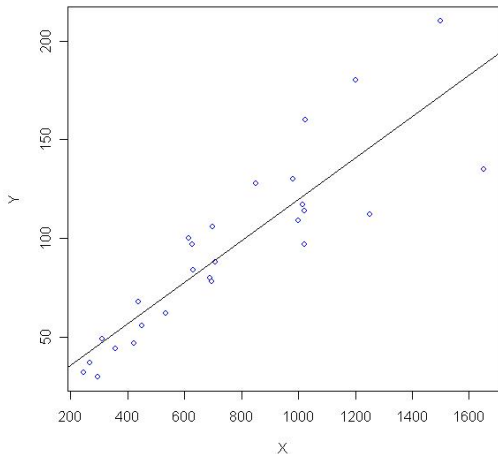


Some models that can be linearized

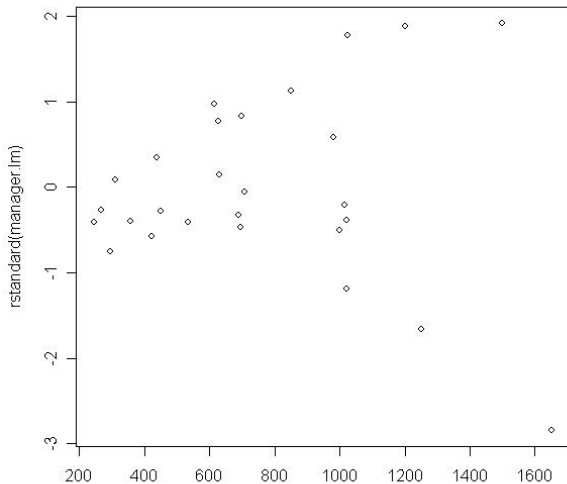
- $y = \alpha x^\beta$, use $\tilde{y} = \log(y)$, $\tilde{x} = \log(x)$;
- $y = \alpha e^{\beta x}$, use $\tilde{y} = \log(y)$;
- $y = x/(\alpha x - \beta)$, use $\tilde{y} = 1/y$, $\tilde{x} = 1/x$.
- More examples in chapter 6 of the textbook.

Nonconstant Variance

In a study of 27 companies, the number of workers (X) and the number of supervisors (Y) were recorded.



Transformations for Stabilizing Variance - Manager Example



How do you test for nonconstant variance?

- Divide data into two groups: high X and low X .
- Let n_1 and n_2 be the number of data points in groups 1 and 2.
- e_{ij} be the i -th residual of group j , $j = 1, 2$.
- $\tilde{e}_j = \text{median}\{e_{i,j} : i = 1, \dots, n_j\}$.
- $d_{ij} = |e_{ij} - \tilde{e}_j|$.

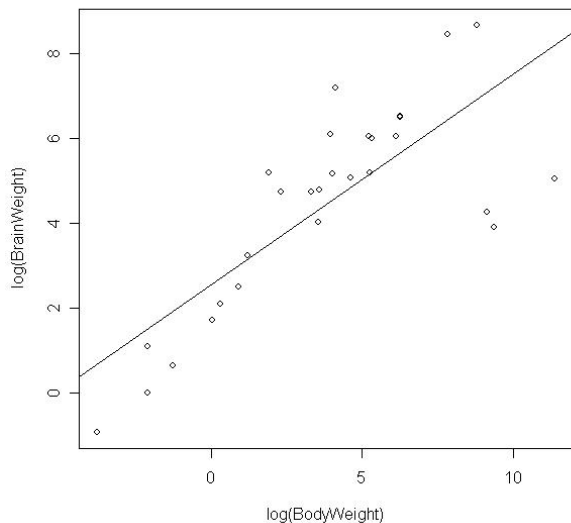
$$t^* = \frac{\bar{d}_1 - \bar{d}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where $\bar{d}_j = \text{mean}\{d_{ij} : i = 1, \dots, n_j\}$, and

$$s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n - 2}.$$

This is the *Brown-Forsythe Test*. Note that it does not assume Gaussianity of errors.

Highly asymmetric data - log transformation



Look at R script...

Summary of Common Transformations

- $\text{Var}(\epsilon) \propto X^2$, then

$$Y' = \frac{Y}{X}, \quad X' = \frac{1}{X}.$$

- $\text{Var}(\epsilon) \propto X$, then

$$Y' = \sqrt{Y}, \quad X' = X.$$

- Either Y or X has large, asymmetric variation (e.g. Brain data),

$$Y' = \log(Y), \quad X' = \log(X).$$

There is often more than one solution. The best approach is to use empirical evidence and domain knowledge.

Variance Stabilizing Transformations

Suppose $E(y) = \mu$, and $Var(Y) = f(\mu)$. Seek transformation $g(Y)$ such that $Var[g(Y)]$ does not rely on μ :

$$g(Y) \approx g(\mu) + g'(\mu)(Y - \mu).$$

$$Var[g(Y)] = [g'(\mu)]^2 Var(Y),$$

thus, we can pick $g(\cdot)$ such that

$$[g'(\mu)]^2 = \frac{1}{f(\mu)}.$$

or

$$g(y) = \int_0^y \frac{1}{\sqrt{f(\mu)}} d(\mu).$$

Example: $Y \text{ Poisson}(\mu)$, $Var(Y) = \mu = E(Y)$, thus a good transformation would be $\text{sqrt}(Y)$.