Lecture 9: Transformations, Weighted ANOVA

Nancy R. Zhang

Statistics 203, Stanford University

February 9, 2010

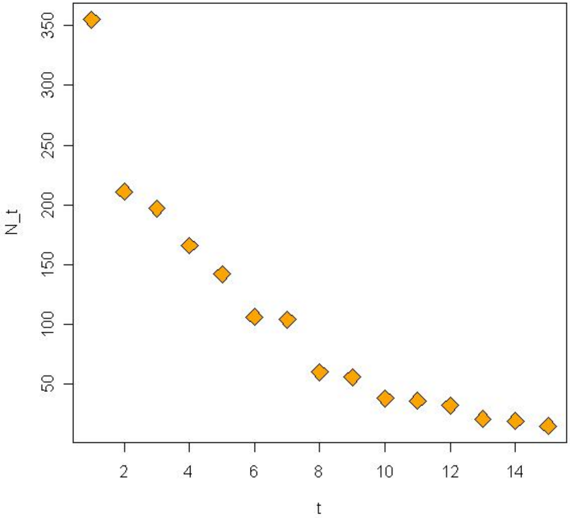
Linear regression (ANOVA) model:

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

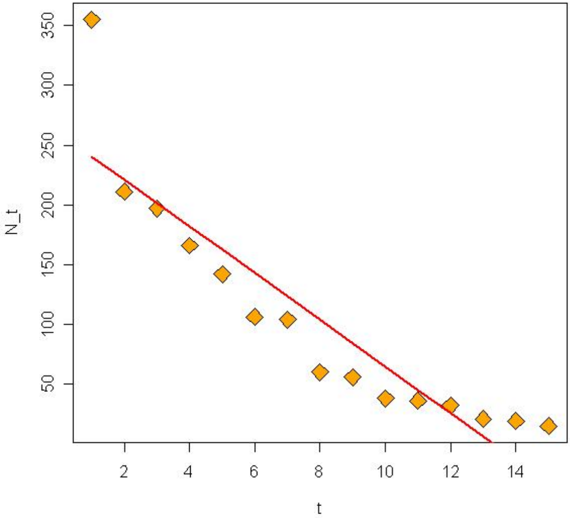
1. Mean depends on predictors in a *linear* way.
2. Error is Gaussian.
3. Variance is constant.
4. Variance is independent.

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

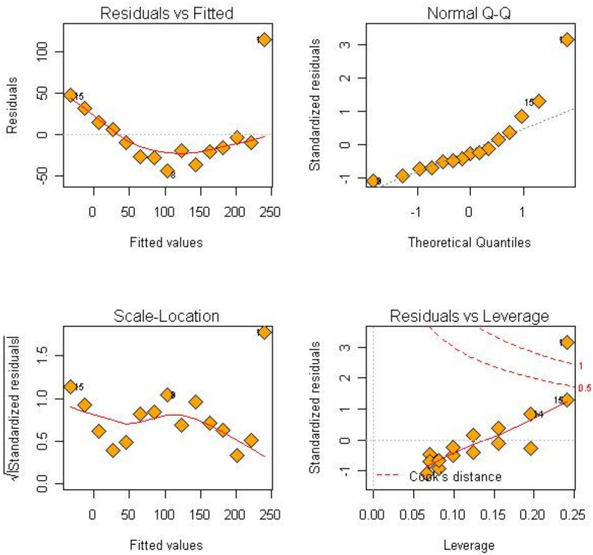
Experiment: Number of surviving marine bacteria following exposure to X-rays.



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

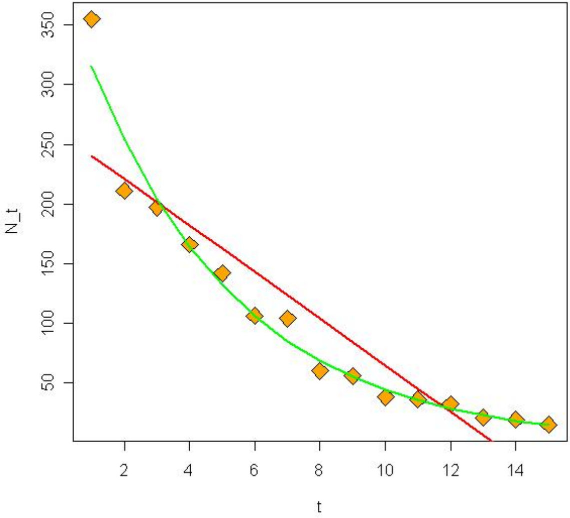
$$E(n_t) = n_0 e^{\beta_1 t}, \qquad t = 1, 2, 3, \ldots$$

- If we take logs of both sides

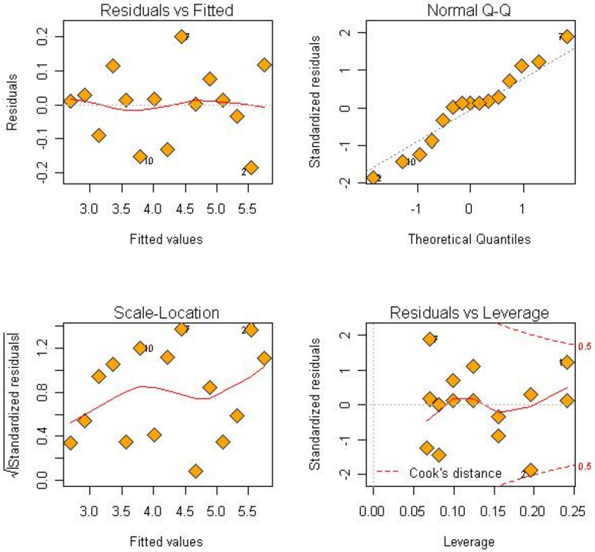
$$\log E(n_t) = \log n_0 + \beta_1 t.$$

- (Reasonable ?) model:

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



N_t
50
100
150
200
250
300
350
2
4
6
8
10
12
14
t

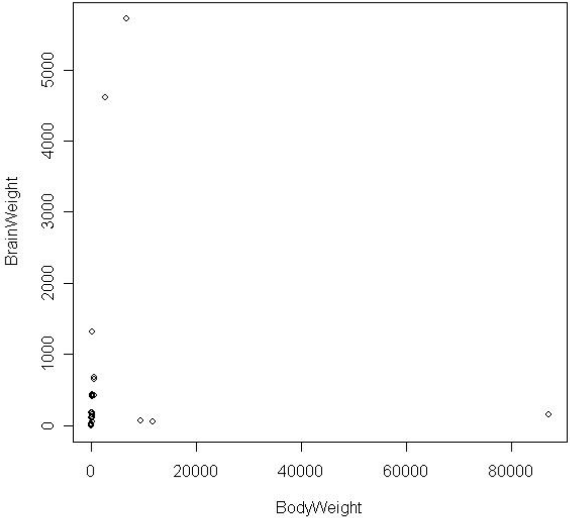


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

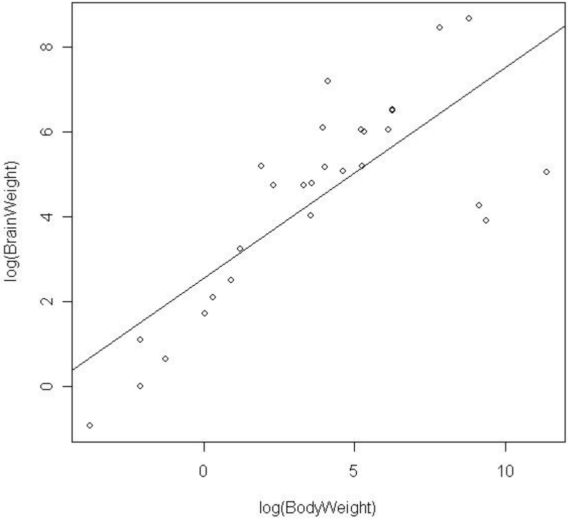
Some models that can be linearized

- $y = \alpha x^{\beta}$, 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.

Highly asymmetric data - Brain Example



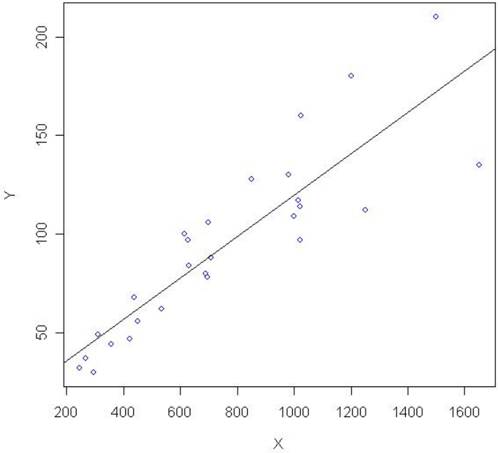
Highly asymmetric data - log transformation



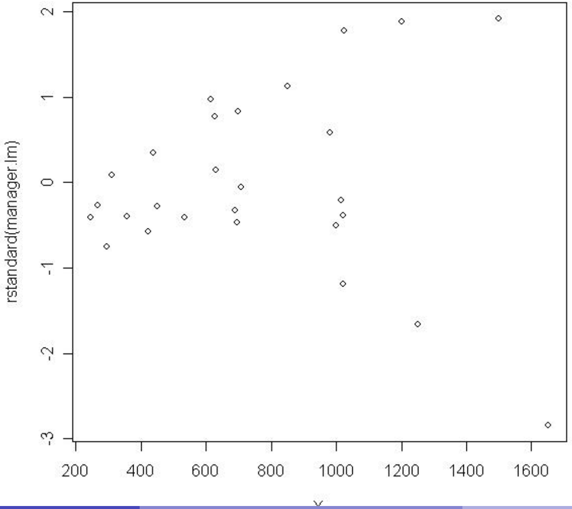
Look at R script...

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



Summary of Common Transformations

- $Var(\epsilon) \propto X^2$, then
$$Y' = \frac{Y}{X}, \quad X' = \frac{1}{X}.$$

- $Var(\epsilon) \propto X$, 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 $Poisson(\mu)$, $Var(Y) = \mu = E(Y)$, thus a good transformation would be $sqrt(Y)$.

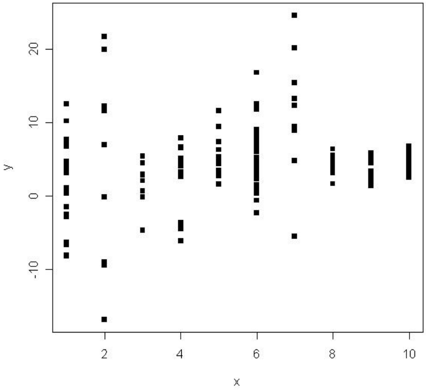
Another example: College expenses

What determines total annual expense for college students?

- Y: *Average* annual expense over students surveyed in the institution.
- Size of city where the school is located.
- Size of student body
- ...

Each data point is an average over sampling units taken over pre-defined groups. The error variance of the observations decrease over group size. Weigh observations by $\sqrt{n_i}$, the size of group i.

Another example: Hypothetical lab experiment



Data at each x can be used to estimate σ_x, weigh observations by σ_x^{-1}.

Solving Weighted Least Squares

Minimize:

$$L_w(\beta) = \sum_{i=1}^{n} w_i (Y_i - \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p})^2.$$

In matrix form:

$$L_w(\beta) = (Y - X\beta)' W (Y - X\beta),$$

where W is diagonal matrix with entries $w_1, \ldots, w_n$. The solution to the above remains linear in Y:

$$\hat{\beta} = (X'WX)^{-1} X'WY.$$

As expected, this is the same as rescaling row i of the data by $\sqrt{w_i}$.

Note that W does not have to be diagonal. Weighted least squares is a special case of *generalized least squares*.

Generalized Least Squares

When W is any symmetric positive-definite square matrix, then solutions to

$$L_w(\beta) = (Y - X\beta)' W (Y - X\beta),$$

are called generalized least squares solutions. Let

$$W = LL', \quad L \text{ lower triangular}$$

be a Cholesky decomposition of W. Then the above is equivalent to least squares on the transformed data,

$$X' = L'X, \quad Y' = L'Y.$$

If we assume Gaussian errors, then this corrresponds to maximum likelihood of a multivariate Gaussian density with covariance marix W^{-1}.

When the error variance structure is not known.

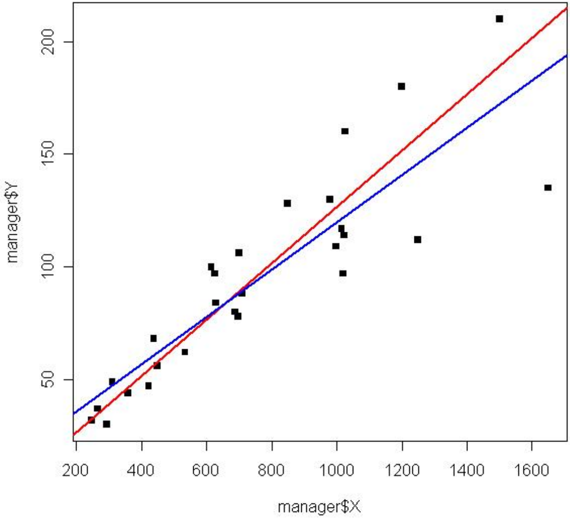
Assume that variance is a function of X:

$$\sigma_i = f(X_i).$$

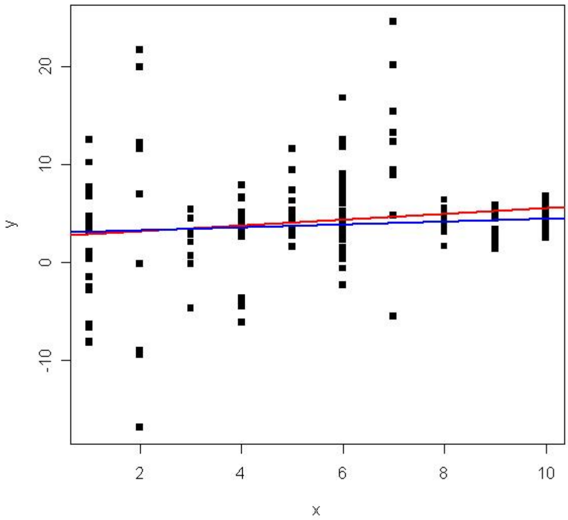
Multiple predictors: rely on prior knowledge to choose X. Relationship should be graphically obvious.

Iterative re-weighted least squares:

1. Fit unweighted least squares,
2. Estimate $\hat{s}_i = f(X_i)$ from absolute residuals (assuming an appropriate functional form),
3. Let $w_i = 1/\hat{s}_i^2$.
4. Repeat the above two steps until convergence.



200
150
100
50
manager$Y
200
400
600
800
1000
1200
1400
1600
manager$X


y
20
10
0
-10
2
4
6
8
10
x