Data Transformations

BST 226
Statistical Methods for Bioinformatics
David M. Rocke

Assumptions

- Consider a two-sample t-test between two random variables X and Y with samples $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_m\}$.
- Assumptions under which we do the math are as follows:
  - The values of X are statistically independent
  - The values of Y are statistically independent
  - The values of X and Y are statistically independent
  - Each value of X has the same variance σ_X^2.
  - Each value of Y has the same variance σ_Y^2.
  - The values of X are normally distributed
  - The values of Y are normally distributed
  - Possibly $\sigma_X^2 = \sigma_Y^2$.

Assumptions

- If we transform the variables to f(*X*) and f(*Y*) then these assumptions are still true or false as with *X* and *Y*
  - The values of X are statistically independent
  - The values of Y are statistically independent
  - The values of X and Y are statistically independent
- But these may change with the transformation
  - Each value of X has the same variance σ_X^2.
  - Each value of Y has the same variance σ_Y^2.
  - The values of X are normally distributed
  - The values of Y are normally distributed

Transformations in Regression

- Transforming X or Y or both (for example to logs) can affect linearity, additivity, non-constant variance, and normality.
- Often logs are useful with measured data at levels well above 0
- Often square roots are useful for count data.
- The generalized logarithm can be used for measured data that has both low and high level observations.

The Delta Method

$E(X) = \mu$

$V(X) = \sigma^2$

$Y = a + bX$

$E(Y) = a + b\mu$

$V(Y) = b^2\mu^2$

$Y = f(X)$

Taylor's theorem says that if f is smooth, then

$f(X) = f(\theta) + f'(\theta)(X - \theta) + f''(\theta)(X - \theta)^2 + f^{(3)}(\theta)(X - \theta)^3 + \cdots$

for points close to θ. We pick $\theta = \mu$ and for points close enough to μ

$f(X) \approx f(\mu) + f'(\mu)(X - \mu)$ so that

$V(f(X)) \approx [f'(\mu)]^2 V(X)$

Variance-Stabilizing Transformations

Suppose that we have a collection of random variables $X_1, X_2, \ldots$ such that

$E(X_i) = \mu_i$

$V(X_i) = a^2 \mu_i^2$

These are random variables with constant CV *a*.

$\ln(X_i) \approx \ln(\mu_i) + \mu_i^{-1}(X_i - \mu_i)$ so long as μ_i is well bounded away from 0.

$\mathrm{V}(\ln(X_i)) \approx \mu_i^{-2} V(X_i) = \mu_i^{-2} a^2 \mu_i^2 = a^2$

so the log results in a variance that is approximately constant for values not too close to 0.

And the variance on the log scale is the same as the square CV on the original scale.

Variance-Stabilizing Transformations

Suppose the X_i are Poisson random variables with parameter λ_i

$E(X_i) = \lambda_i$

$V(X_i) = \lambda_i$

Find a variance-stabilizing transformation

Let $f(x) = x^\alpha$

$f(X_i) \approx \lambda_i^\alpha + \alpha\lambda_i^{\alpha-1}(X_i - \lambda_i)$

$V(X_i) \approx \alpha^2\lambda_i^{2\alpha-2}V(X_i) = \alpha^2\lambda_i^{2\alpha-2}\lambda_i = \alpha^2\lambda_i^{2\alpha-2+1} = \alpha^2\lambda_i^{2\alpha-1}$

This does not vary with λ_i only if $2\alpha - 1 = 0$ or $\alpha = 0.5$

The square root transformation stabilizes the variance of Poisson random variables

Variance-Stabilizing Transformations

Suppose that we have a collection of random variables $X_1, X_2, \ldots$ such that

$E(X_i) = \mu_i$

$V(X_i) = a^2 + b^2\mu_i^2$

These are random variables with constant CV b at high levels and constant standard deviation a at low levels.

If we have a transformation $Y = f(X)$, then

$Y_i \approx \mu_i + f'(\mu_i)(X_i - \mu_i)$

and the variance of Y is approximately

$V(Y_i) \approx [f'(\mu_i)]^2 V(X_i) = [f'(\mu_i)]^2 (a^2 + b^2\mu_i^2)$

so the variance is approximately constant when

$$f'(x) = \frac{k}{\sqrt{a^2 + b^2x^2}}$$

$$f(x) = \int \frac{k}{\sqrt{a^2 + b^2x^2}} dx = \left(\frac{k}{b}\right) \int \frac{1}{\sqrt{a^2/b^2 + x^2}} dx$$

$$f(x) = \left(\frac{k}{b}\right) \ln\left(x + \sqrt{x^2 + a^2/b^2}\right)$$

If we choose $k = b$ and consider a single parameter $\lambda = a^2/b^2$ then the transformation is

$$f(x) = \ln\left(x + \sqrt{x^2 + \lambda^2}\right)$$

Variance-Stabilizing Transformations

If we have uncalibrated values (or pre-calibrated) and

$$E(X_i) = \alpha + \beta\mu_i$$

$$V(X_i) = a^2 + b^2\mu_i^2$$

then we have to subtract α from the X_i before transformation

so that the mean and variance work correctly

This means our transformation is

$$f(x) = \ln\left(x - \alpha + \sqrt{(x-\alpha)^2 + \lambda^2}\right)$$

we do not have to separately account for β since it is absorbed into b^2

Transformations vs. Weighting

- Suppose we have a regression with heteroscedasticity.
- We can transform y and/or x so that the variance is more nearly constant.
- We could also conduct a weighted least squares analysis with weights equal to the inverse estimated variance of each observation.
- These will often yield results that are similar, but sometimes one method may be better than the other, depending on context.