Variance-stabilizing Transformations and Weighted Least Squares[1]

STA442/2101 Fall 2016

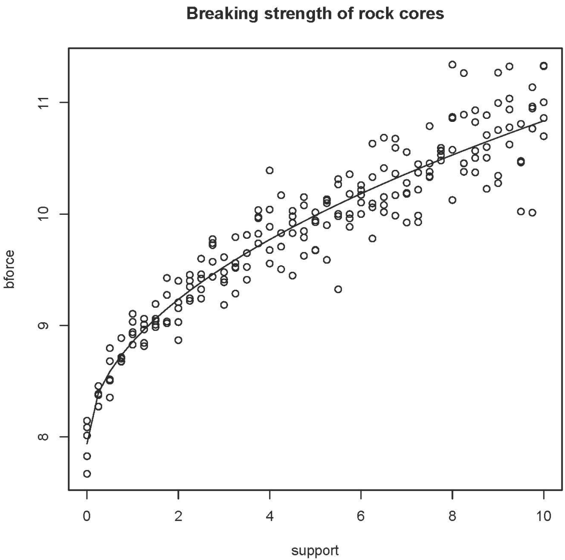
[1] See last slide for copyright information.

Overview

1. Unequal Variance
2. Delta Method
3. Weighted Least Squares

Unequal Variance

Can you say “heteroscedasticity?”



Why is unequal variance a problem?

Not just because the model is wrong – let's be more specific.

- Normal distribution theory depends on cancelling σ^2 in numerator and denominator.
- There is some robustness. Tests have approximately the right Type I error probability when the number of observations at each combination of x values is large and roughly equal.
- $\widehat{\boldsymbol{\beta}}$ is still unbiased, but no longer minimum variance.
- Intuitively, observations where the variance is smaller should count more.
- If the variance depends on x, prediction intervals should be wider for x values with larger variance.

Two solutions

- Variance-stabilizing transformations: If the variance depends on $E(Y_i)$, transform the response variable.
- Weighted least squares: If the variance is proportional to some known constant, transform both $\mathbf{X}$ and $\mathbf{y}$.

The Delta Method

The univariate version of the delta method says that if

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} T$$

then

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow{d} g'(\theta)\,T.$$

If $T \sim N(0, \sigma^2)$, it says

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow{d} Y \sim N\left(0, g'(\theta)^2\,\sigma^2\right).$$

Taylor's Theorem
Basis of the Delta Method

For the function $g(x)$, let the nth derivative $g^{(n)}$ be continuous in $[a, b]$ and differentiable in (a, b), with x and x_0 in (a, b). Then there exists a point ξ between x and x_0 such that

$$\begin{aligned} g(x) &= g(x_0) + g'(x_0)(x - x_0) + \frac{g''(x_0)(x - x_0)^2}{2!} + \dots \\ &+ \frac{g^{(n)}(x_0)(x - x_0)^n}{n!} + \frac{g^{(n+1)}(\xi)(x - x_0)^{n+1}}{(n+1)!}, \end{aligned}$$

where $R_n = \frac{g^{(n+1)}(\xi)(x-x_0)^{n+1}}{(n+1)!}$ is called the *remainder term*.

If $R_n \to 0$ as $n \to \infty$, the resulting infinite series is called the *Taylor Series* for $g(x)$.

Two terms of a Taylor Series

Plus remainder

$$g(x) = g(x_0) \, + \, g'(x_0) \, (x - x_0) \, + \, \frac{g''(\xi)(x - x_0)^2}{2!}$$

Proof of the Delta Method

Using $g(x) = g(x_0) + g'(x_0)\,(x - x_0) + \frac{g''(\xi)(x-x_0)^2}{2!}$

Suppose $\sqrt{n}\,(T_n - \theta) \xrightarrow{d} T$. Then expanding $g(x)$ about θ,

$$
\begin{aligned}
\sqrt{n}\,(g(T_n) - g(\theta)) &= \sqrt{n}\left(g(\theta) + g'(\theta)\,(T_n - \theta) + \frac{g''(\xi_n)(T_n - \theta)^2}{2!} - g(\theta)\right) \\
&= g'(\theta)\,\sqrt{n}\,(T_n - \theta) \; + \\
&\qquad \underset{\substack{p\,\downarrow \\ 1/2\,\theta}}{\frac{1}{2} g''(\xi_n)} \cdot \underset{\substack{d\,\downarrow \\ T}}{\sqrt{n}\,(T_n - \theta)} \cdot \underset{\substack{p\,\downarrow \\ 0}}{(T_n - \theta)} \\
&\xrightarrow{d} g'(\theta)T \; + \; 0
\end{aligned}
$$

A variance-stabilizing transformation

An application of the delta method

- Because the Poisson process is such a good model, count data often have approximate Poisson distributions.
- Let $X_1, \ldots, X_n \stackrel{i.i.d}{\sim}$ Poisson(λ)
- $E(X_i) = Var(X_i) = \lambda$
- CLT says $\sqrt{n}(\overline{X}_n - \lambda) \stackrel{d}{\rightarrow} T \sim N(0, \lambda)$.
- Delta method says $\sqrt{n}\left(g(\overline{X}_n) - g(\lambda)\right) \stackrel{d}{\rightarrow} g'(\lambda)\, T = Y \sim N\left(0, g'(\lambda)^2\, \lambda\right)$
- If $g'(\lambda) = \frac{1}{\sqrt{\lambda}}$, then $Y \sim N(0, 1)$.

An elementary differential equation: $g'(x) = \frac{1}{\sqrt{x}}$

Solve by separation of variables

$$
\begin{aligned}
& \frac{dg}{dx} = x^{-1/2} \\
\Rightarrow \;\; & dg = x^{-1/2}\,dx \\
\Rightarrow \;\; & \int dg = \int x^{-1/2}\,dx \\
\Rightarrow \;\; & g(x) = \frac{x^{1/2}}{1/2} + c = 2x^{1/2} + c
\end{aligned}
$$

We have found

$$\begin{aligned}\sqrt{n}\left(g(\overline{X}_n) - g(\lambda)\right) &= \sqrt{n}\left(2\overline{X}_n^{1/2} - 2\lambda^{1/2}\right) \\ &\stackrel{d}{\rightarrow} Z \sim N(0,1)\end{aligned}$$

- We could say that $\sqrt{\overline{X}_n}$ is asymptotically normal, with mean $\sqrt{\lambda}$ and variance $\frac{1}{4n}$.
- This is because $\overline{X}_n^{1/2} = \frac{Z}{2\sqrt{n}} + \sqrt{\lambda}$.
- Notice that the variance no longer depends on λ.
- This calculation could justify a square root transformation for count data.
- Because if $\overline{X}_n$ is asymptotically normal, so is $\sum_{i=1}^n X_i$
- And the sum of independent Poissons is Poisson.

Sometimes it can be pretty loose

Just drop the remainder term in $g(x) = g(x_0) + g'(x_0)(x - x_0) + R$

If $Var(X) = \sigma^2$, then

$$\begin{aligned} Var(g(X)) &\approx Var\left(g(x_0) + g'(x_0)(X - x_0)\right) \\ &= Var\left(g'(x_0)X\right) \\ &= g'(x_0)^2 Var(X) \\ &= g'(x_0)^2 \sigma^2 \end{aligned}$$

Call it "linearization."

The approximation $g(x) = g(x_0) + g'(x_0)(x - x_0)$ is good, for x close to x_0.

The arcsin-square root transformation for proportions

This is careful again.

Sometimes, variable values consist of proportions, one for each case.

- For example, cases could be hospitals.
- The variable of interest is the proportion of patients who came down with something *unrelated* to their reason for admission – hospital-acquired infection.
- This is an example of *aggregated data*.

The advice you often get

When a proportion is the response variable in a regression, use the *arcsin square root* transformation.

That is, if the proportions are $P_1, \ldots, P_n$, let

$$Y_i = 2\sin^{-1}(\sqrt{P_i})$$

and use the Y_i values in your regression.

Why?

It's a variance-stabilizing transformation.

- The proportions are little sample means: $P_i = \frac{1}{m}\sum_{j=1}^{m} X_{i,j}$
- Drop the i for now.
- $X_1, \ldots, X_m$ may not be independent, but let's pretend.
- $P = \overline{X}_m$
- Approximately, $\overline{X}_m \sim N\left(\theta, \frac{\theta(1-\theta)}{m}\right)$
- Normality is good.
- Variance that depends on the mean θ is not so good.

Apply the delta method

Central Limit Theorem says

$$\sqrt{m}(\overline{X}_m - \theta) \xrightarrow{d} T \sim N\left(0, \theta(1-\theta)\right)$$

Delta method says

$$\sqrt{m}\left(g(\overline{X}_m) - g(\theta)\right) \xrightarrow{d} Y \sim N\left(0, g'(\theta)^2\,\theta(1-\theta)\right).$$

Want a function $g(x)$ with

$$g'(x) = \frac{1}{\sqrt{x(1-x)}}$$

Try $g(x) = 2\sin^{-1}(\sqrt{x})$.

Chain rule to get $\frac{d}{dx}\sin^{-1}\left(\sqrt{x}\right)$

"Recall" that $\frac{d}{dx}\sin^{-1}(x) = \frac{1}{\sqrt{1-x^2}}$. Then,

$$
\begin{aligned}
\frac{d}{dx} 2\sin^{-1}\left(\sqrt{x}\right) &= 2\frac{1}{\sqrt{1-\sqrt{x}^2}} \cdot \frac{1}{2}x^{-1/2} \\
&= \frac{1}{\sqrt{x(1-x)}} \text{ For } 0 < x < 1.
\end{aligned}
$$

Conclusion:

$$
\sqrt{m}\left(2\sin^{-1}\sqrt{\overline{X}_m} - 2\sin^{-1}\sqrt{\theta}\right) \stackrel{d}{\rightarrow} Y \sim N(0,1)
$$

So the arcsin-square root transformation stabilizes the variance

Because $\sqrt{m}\left(2\sin^{-1}\sqrt{\overline{X}_m} - 2\sin^{-1}\sqrt{\theta}\right) \stackrel{d}{\rightarrow} Y \sim N(0,1)$

- If we want to do a regression on aggregated data, the point we have reached is that approximately,

$$Y_i \sim N\left(2\sin^{-1}\sqrt{\theta_i}\,,\,\frac{1}{m_i}\right)$$

- The variance no longer depends on the probability that the proportion is estimating.
- Y is meaningful because the function $g(x)$ is increasing.
- But the variance still depends on the number of patients in the hospital.

Weighted Least Squares

- Suppose that the variances of $Y_1, \ldots, Y_n$ are unequal, but proportional to known constants.
- Aggregated data fit this pattern. Means are usually based on different sample sizes.
- Generalize it: In the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, with $\mathbf{V}$ a *known* symmetric positive definite matrix.

Transform the Data

Have $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$.

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \Rightarrow \mathbf{V}^{-1/2}\mathbf{y} &= \mathbf{V}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{V}^{-1/2}\boldsymbol{\epsilon} \\ \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \end{aligned}$$

So that

- $cov(\boldsymbol{\epsilon}^*) = \sigma^2\mathbf{I}_n$
- Note that the transformed model has the same $\boldsymbol{\beta}$.

You don't have to literally transform the data

Just transform the estimates, tests and intervals

- $\widehat{\boldsymbol{\beta}}_{wls} = (\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y}$ and so on.
- The most common case is where the variances are proportional to known constants and the errors are independent. That is, $\mathbf{V}$ is diagonal.
- Most software will allow you to supply the diagonal elements of $\mathbf{V}^{-1}$.
- These are called the "weights."
- In the case of aggregated data where $Var(Y_i) = \frac{\sigma^2}{m_i}$, the weights are just $m_1, \ldots m_n$.
- in `help(lm)`, R's help says

  Non-NULL weights can be used to indicate that different observations have different variances (with the values in weights being inversely proportional to the variances); or equivalently, when the elements of `weights` are positive integers w_i, that each response y_i is the mean of w_i unit-weight observations (including the case that there are w_i observations equal to y_i and the data have been summarized).

Sometimes weighted least squares is used loosely

Is this an abuse?

- Residual plots suggest that variance might be proportional to x_{ij}.
- So pretend it's known, and use weights $\frac{1}{x_{1j}}, \ldots, \frac{1}{x_{nj}}$.
- This has been studied. The Wikipedia article has references.

Copyright Information