Linear Regression Models
P8111

Lecture 19

Jeff Goldsmith
April 5, 2016

THE DEPARTMENT OF BIOSTATISTICS
Columbia University
MAILMAN SCHOOL
OF PUBLIC HEALTH

Today’s Lecture

- Weighted least squares
- Generalized least squares

Multiple regression model

We typically pose a model of the form

$$y_i|\boldsymbol{x}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$$

and assume $Var(\epsilon_i) = \sigma^2$

- Today we're concerned with $Var(\epsilon_i) = \frac{\sigma^2}{w_i}$
- More generally, we'll look at $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{W}$ or $Var(\boldsymbol{\epsilon}) = \Sigma$
- Contexts include non-constant variance, sampling data (survey weights), proportional data (sample size in groups), meta-analysis (variance of effects in each study)

Weighted least squares

- One way to handle non-constant variance is a variance stabilizing transformation, which works well if the variance depends on the mean
- Weighted least squares builds the weighting terms directly into the criterion to be minimized
- Let $\boldsymbol{W}$ be the matrix with $(i, i)^{th}$ entry $\frac{1}{w_i}$ and 0 elsewhere
- Then $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}$

Weighted least squares

- For weighted least squares, we minimize the RSS with terms weighted according to their variance

$$\begin{aligned} RSS_W(\boldsymbol{\beta}) &= \sum w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \\ &= (y - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}^{-1} (y - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

- We weight more heavily terms with low variance (small $\frac{\sigma^2}{w_i}$) and less heavily terms with high variance (big $\frac{\sigma^2}{w_i}$)
- Basic plan – differentiate $RSS_W(\boldsymbol{\beta})$ wrt $\boldsymbol{\beta}$ and find the minimum

Weighted least squares estimator

$$RSS_W(\boldsymbol{\beta}) = (y - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W}^{-1} (y - \boldsymbol{X}\boldsymbol{\beta})$$

A note about MLE

We have the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}$.

- Additionally, assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{W})$
- Put differently, we're imposing the model

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{W})$$

- $\boldsymbol{y}$ is multivariate Normal

Maximum likelihood estimation

Using matrix notation:

Pre-whitening data

- Let $\boldsymbol{W}^{1/2}$ be the diagonal matrix with $(i,i)^{th}$ $\frac{1}{\sqrt{w_i}}$ and 0 elsewhere
- So $\boldsymbol{W}^{-1/2} \stackrel{def}{=} \left(\boldsymbol{W}^{1/2}\right)^{-1}$ is a diagonal matrix with $\sqrt{w_i}$ on the main diagonal and 0 elsewhere
- Note $\boldsymbol{W} = \boldsymbol{W}^{1/2}(\boldsymbol{W}^{1/2})^T$ and $\boldsymbol{W}^{1/2}\boldsymbol{W}^{-1/2} = I$
- So $Var(\boldsymbol{W}^{-1/2}\boldsymbol{\epsilon}) =$

Pre-whitening data

- Let's pre-multiply everything by $\boldsymbol{W}^{-1/2}$:
  - $\boldsymbol{z} = \boldsymbol{W}^{-1/2}\boldsymbol{y}$
  - $\boldsymbol{M} = \boldsymbol{W}^{-1/2}\boldsymbol{X}$
  - $\delta = \boldsymbol{W}^{-1/2}\boldsymbol{\epsilon}$
- Our model is now

$$\boldsymbol{z} = \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{\delta}$$

- The OLS estimate of $\boldsymbol{\beta}$ is

$$(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{z}$$

WLS example

- Data from a physics experiment, available as `physics` from the library `alr3`
- y: scattering cross-section, s: square of total energy, $x = s^{-1/2}$
- Theoretical model:
  $E(y|s) = \beta_0 + \beta_1 s^{-1/2}$ + relatively small terms
- Regression model: $y = \beta_0 + \beta_1 x + \epsilon$
- $SD = \sqrt{Var(y|x)}$ are known from the experiment

WLS example

```
> library(alr3)
> data(physics)
> physics
       x   y SD
1  0.345 367 17
2  0.287 311  9
3  0.251 295  9
4  0.225 268  7
5  0.207 253  7
6  0.186 239  6
7  0.161 220  6
8  0.132 213  6
9  0.084 193  5
10 0.060 192  5
```

WLS example

```
> lm.physics.wls <- lm(y~x, weights=1/SD^2,data=physics)
> summary(lm.physics.wls)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.473      8.079   18.38 7.91e-08 ***
x            530.835     47.550   11.16 3.71e-06 ***

Residual standard error: 1.657 on 8 degrees of freedom
Multiple R-squared: 0.9397,     Adjusted R-squared: 0.9321
F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.710e-06
```

WLS example

```
> lm.physics.ols <- lm(y~x, data=physics)
> summary(lm.physics.ols)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   135.00      10.08    13.4 9.21e-07 ***
x             619.71      47.68    13.0 1.16e-06 ***

Residual standard error: 12.69 on 8 degrees of freedom
Multiple R-squared: 0.9548,Adjusted R-squared: 0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.165e-06
```

WLS in practice

- Real life is rarely nice enough to give you the right weight
- Try to obtain an estimate of $var(\epsilon_i)$, plug that into $\boldsymbol{W}$ ...

Generalized least squares

- Weighted least squares can help a lot, but what if errors are correlated?
- That is, suppose our model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\Sigma$
- (By analogy with WLS, suppose Σ is known but σ^2 is not; in general, one usually writes $Var(\boldsymbol{\epsilon}) = \Sigma$)
- Note, in terms of generality, GLS $>$ WLS $>$ OLS

Generalized least squares

- Writing out $RSS_G(\boldsymbol{\beta})$ as a sum is hard; possible using vector notation.
- Possibilities:
  - MLE (equivalent to minimizing RSS)
  - Pre-whiten

MLE

We have the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $E(\boldsymbol{\epsilon}) = 0$ and $Var(\boldsymbol{\epsilon}) = \sigma^2\Sigma$.

- Additionally, assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2\Sigma)$
- Put differently, we're imposing the model

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\Sigma)$$

- $\boldsymbol{y}$ is multivariate Normal

MLE

Using matrix notation:

Pre-whitening data

- Let $\Sigma = SS^T$ be the *Cholesky decomposition* of Σ
- Let's pre-multiply everything by S^{-1}:
  - $\boldsymbol{z} = \boldsymbol{W}^{-1/2}\boldsymbol{y}$
  - $\boldsymbol{M} = \boldsymbol{W}^{-1/2}\boldsymbol{X}$
  - $\delta = \boldsymbol{W}^{-1/2}\boldsymbol{\epsilon}$
- Our model is now

$$\boldsymbol{z} = \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{\delta}$$

- The OLS estimate of $\boldsymbol{\beta}$ is

$$(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{M}^T\boldsymbol{z}$$

Some useful notes on GLS

Using $\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{y}$, it turns out that

- $E(\hat{\boldsymbol{\beta}}_{GLS}) = \boldsymbol{\beta}$
- $Var(\hat{\boldsymbol{\beta}}_{GLS}) = \sigma^2(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}$

Some less useful notes on GLS

- Typically we don't really know Σ and have to estimate it too
- A common approach is to parameterize Σ using a small number of parameters
- Comes up a lot for longitudinal and multilevel data

Today's big ideas

- Weighted and generalized least squares

- Suggested reading: Ch. 5