

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

$$\text{Var}(\epsilon) = \sigma^2 I$$

OLS

PenLS

- Weighted least squares
- Generalized least squares

Multiple regression model

We typically pose a model of the form

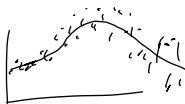
$$y_i | x_i = x_i \beta + \epsilon_i$$

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 $Var(\epsilon) = \sigma^2 \mathbf{W}$ or $Var(\epsilon) = \Sigma$
- 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 \mathbf{W} be the matrix with $(i, i)^{th}$ entry $\frac{1}{w_i}$ and 0 elsewhere
- Then $Var(\epsilon) = \sigma^2 \mathbf{W}$

Weighted least squares

$$\frac{1}{w_i}$$

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

$$\begin{aligned} \underline{RSS}_W(\beta) &= \sum w_i (y_i - x_i^T \beta)^2 \quad \checkmark \quad \text{wt'd mean!} \\ &= (y - X\beta)^T \underline{W}^{-1} (y - X\beta) \end{aligned}$$

- 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 $RSS_W(\beta)$ wrt β and find the minimum

Weighted least squares estimator

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

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

$$y^T W^{-1} y - \beta^T X^T W^{-1} y - y^T W^{-1} X \beta + \beta^T X^T W^{-1} X \beta$$

$$= -2 \beta^T X^T W^{-1} y + \beta^T X^T W^{-1} X \beta$$

$$\frac{\partial}{\partial \beta} = -2 X^T W^{-1} y + 2 X^T W^{-1} X \beta = 0$$

$$X^T W^{-1} X \beta = X^T W^{-1} y \Rightarrow \beta_{WLS} = (X^T W^{-1} X)^{-1} X^T W^{-1} y$$

A note about MLE

We have the model

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

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

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

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

- \mathbf{y} is multivariate Normal

Maximum likelihood estimation

Using matrix notation:

$$L(\beta; y) \propto \exp \left\{ \frac{-1}{2\sigma^2} \underbrace{(\underline{y} - X\underline{\beta})^T \mathbf{W}^{-1} (\underline{y} - X\underline{\beta})}_{\text{RSS}(\beta)} \right\}$$

Pre-whitening data

$$\begin{bmatrix} \frac{1}{w_1} & & \\ & \ddots & \\ & & \frac{1}{w_n} \end{bmatrix}$$

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

$$\begin{aligned} & (\underline{W}^{-1/2})^T Var(\epsilon) \underline{W}^{-1/2} \\ &= \sigma^2 \underbrace{(\underline{W}^{-1/2})^T (\underline{W}^{1/2} \underline{W}^{1/2})}_{I} \underline{W}^{-1/2} \\ &= \sigma^2 I \end{aligned}$$

Pre-whitening data

- Let's pre-multiply everything by $W^{-1/2}$:

$$\left\{ \begin{array}{l} \triangleright z = W^{-1/2} y \\ \triangleright M = \overline{W}^{-1/2} X \\ \triangleright \underline{\delta} = \underline{W}^{-1/2} \epsilon \end{array} \right.$$

$$y = (X\beta + \epsilon) \quad \checkmark$$

$\epsilon \sim (0, \sigma^2 \underline{W})$

- Our model is now

$$z = M\beta + \delta$$

- The OLS estimate of β is

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

$$\underline{\delta} \sim (0, \sigma^2 \underline{I})$$

$$\begin{array}{l} (X^T W^{-1/2^T} W^{-1/2} X) X^T W^{-1/2^T} W^{1/2} y \\ (X^T W^{-1} X)^T X^T W^{-1} y \end{array}$$

WLS example

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

↙

$$SD(\epsilon_i) =$$

$$\text{Var}(\epsilon_i) = \frac{\sigma^2}{w_i}$$

$$\Rightarrow w_i \propto \frac{1}{SD(\epsilon_i)^2}$$

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	230.835	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

$$530 \pm 90$$

$$(440, 620)$$

WLS example

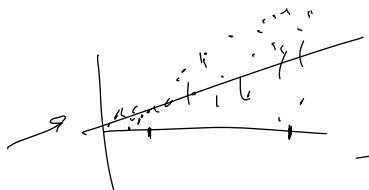
```
> lm.physics.ols <- lm(y~x, data=physics)
> summary(lm.physics.ols)
```

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



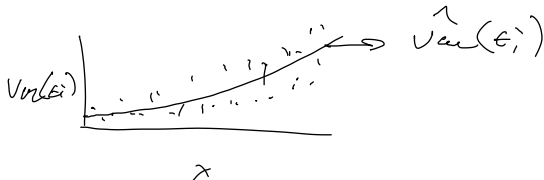
- ▶ Real life is rarely nice enough to give you the right weight
- ▶ Try to obtain an estimate of $var(\epsilon_i)$, plug that into W ...

① OLS

$$var(\epsilon_i) \approx \frac{\sum \hat{\epsilon}_i^2}{n}$$

② $\hat{\epsilon}_i$

$$\frac{\hat{\epsilon}_i^2}{\underbrace{\quad}}$$



Generalized least squares

$$\sigma^2 \underline{\epsilon} \Rightarrow \sigma^2 \underbrace{w}_{\text{diag}} \Rightarrow \underline{\sigma^2 \Sigma}$$

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

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

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

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

Generalized least squares

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

MLE

We have the model

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

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

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

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

- \mathbf{y} is multivariate Normal

MLE

Using matrix notation:

$$L(\beta; y) \propto \exp\left\{-\frac{1}{2\sigma^2} (y - X\beta)^T Z^{-1} (y - X\beta)\right\}$$

$$\frac{\partial}{\partial \beta} = \dots = 0$$

$$\Rightarrow \hat{\beta}_{GLS} = (X^T Z^{-1} X)^{-1} X^T Z^{-1} y$$

Pre-whitening data

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

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

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

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

Some useful notes on GLS

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

- $E(\hat{\beta}_{GLS}) = \beta$

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

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