

BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 5: Weighted least squares

Version: 09/17/2023

TABLE OF CONTENTS

1 Weighted least squares

2 Iteratively re-weighted least squares

WEIGHTED LEAST SQUARES

Recall: Ordinary least squares

- For this set of notes, assume \mathbf{X} is fixed and of full rank (though key results do not depend upon this).
- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.
- OLS minimizes sum of squared errors (convex optimization problem):

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^K}{\text{minimize}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

- Leads to the normal equations:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}.$$

- Closed-form expression for least squares estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- Closed-form expression for variance: $\text{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$

WEIGHTED LEAST SQUARES

New setup: Diagonal variance

- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$.
- Specifically, $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$ with $\sum_{i=1}^N \mathbf{V}_{ii} = 1$ (indeed, $\mathbf{V}_{ii} > 0$, and $\mathbf{V}_{ij} = 0$ for $i \neq j$, at least in BIOS 7345).
- The OLS estimate is unbiased, but OLS-based variance is incorrect.
- Solution: Let $\mathbf{K} = \mathbf{V}^{-1/2}$. Then:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \iff \mathbf{K}\mathbf{y} = \mathbf{K}\mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\epsilon}.$$

- Let $\mathbf{y}^* = \mathbf{K}\mathbf{y}$, $\mathbf{X}^* = \mathbf{K}\mathbf{X}$, and $\boldsymbol{\epsilon}^* = \mathbf{K}\boldsymbol{\epsilon}$.
- Then, $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$, with:
 - ▶ $E[\boldsymbol{\epsilon}^*] = E[\mathbf{K}\boldsymbol{\epsilon}] = \mathbf{K}E[\boldsymbol{\epsilon}] = \mathbf{0}$.
 - ▶ $\text{Cov}[\boldsymbol{\epsilon}^*] = \text{Cov}[\mathbf{K}\boldsymbol{\epsilon}] = \mathbf{K}\text{Cov}[\boldsymbol{\epsilon}]\mathbf{K}^T = \mathbf{V}^{-1/2}\sigma^2\mathbf{V}\mathbf{V}^{-1/2} = \sigma^2\mathbf{I}$.
- Applying OLS on the transformed data gives:

$$\hat{\boldsymbol{\beta}}^* = ((\mathbf{X}^*)^T\mathbf{X}^*)^{-1}(\mathbf{X}^*)^T\mathbf{y}^* = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}.$$

Properties:

- Straightforward to show:
 - ▶ $\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}$ is an oblique projection matrix
 - ★ Idempotent but not necessarily symmetric.
 - ▶ $E[\hat{\boldsymbol{\beta}}^*] = \boldsymbol{\beta}$.
 - ▶ $\text{Cov}[\hat{\boldsymbol{\beta}}^*] = \sigma^2 (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$.
 - ▶ $E[\text{RSS}] = \sigma^2 (N - K)$, where $\text{RSS} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^*)$.
 - ★ This suggests estimating σ^2 as:

$$\hat{\sigma}^2 = \frac{1}{N - K} \sum_{i=1}^N \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^*)^2}{V_i}$$

- If $\hat{\boldsymbol{\beta}}$ denotes the OLS estimate based on the original data, we also have $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\text{Cov}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$.
- Since $\hat{\boldsymbol{\beta}}$ is unbiased, what is the motivation for $\hat{\boldsymbol{\beta}}^*$?

Theorem 5.1: The Gauss-Markov theorem (version 4)

Let \mathbf{X} be of full rank, and suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\text{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$. Then, for any constant vector \mathbf{a} , $\mathbf{a}^T \hat{\boldsymbol{\beta}}^*$ is the BLUE of $\mathbf{a}^T \boldsymbol{\beta}$.

- The proof of this version of the Gauss-Markov theorem relies on the version for OLS.

TABLE OF CONTENTS

1 Weighted least squares

2 Iteratively re-weighted least squares

Choosing weights:

- The premise of the previous results is, in some sense, a little bit cheap in the sense that we almost never know the form of \mathbf{V} exactly.
 - ▶ To be fair, OLS can be thought of as making the specific choice $\mathbf{V} = \mathbf{I}$, which we also almost never know to be true. Specifically having to correctly choose a \mathbf{V} that has a potentially unique entry along each entry of the main diagonal seems like even more of a stretch.
- The slides that follow suggest a procedure on how to deal with the case where you are willing to posit (and estimate) a form for the mean-variance relationship.

Unknown variance:

- Consider $\mathbf{V} = \mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}); \boldsymbol{\theta})$, so the mean is a function of $\boldsymbol{\mu} = E[\mathbf{y}]$ that depends upon $\boldsymbol{\beta}$ and possibly other unknown parameters $\boldsymbol{\theta}$.
 - ▶ We often focus on the case where $\mathbf{V} = \mathbf{V}(\boldsymbol{\mu})$.
- The structure of \mathbf{V} is assumed known, specified by the user.
- Beginning with some particular estimate of $\boldsymbol{\beta}$ (e.g., from OLS) allows us to subsequently *estimate* $\text{Cov}[\boldsymbol{\epsilon}]$ and then use that to inform our weights on the next round.
- But why stop there? We can cycle through that process again (until apparent convergence). This procedure is known as iteratively re-weighted least squares (IRLS).
 - ▶ If your model for $\text{Cov}[\boldsymbol{\epsilon}]$ is correct, resulting $\hat{\boldsymbol{\beta}}$ from IRLS is asymptotically BLUE for $\boldsymbol{\beta}$.

IRLS: Estimating $\boldsymbol{\beta}$ when $\text{Cov}[\boldsymbol{\epsilon}] = \mathbf{V}(\boldsymbol{\mu})$.

- 1 Set convergence threshold, $\delta_{\text{stop}} > 0$ (small).
- 2 If you initialize the value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^{(0)}$ (say, from OLS), the next step in a first-order Taylor approximation is given by:

$$\begin{aligned}
 \boldsymbol{\beta}^{(j+1)} &= \boldsymbol{\beta}^{(j)} + (\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(j)}) \\
 &= (\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{X})^{-1} \left(\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{X} \boldsymbol{\beta}^{(j)} + \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(j)}) \right) \\
 &= (\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \left(\mathbf{X} \boldsymbol{\beta}^{(j)} + (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^{(j)}) \right) \\
 &= (\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(j)}) \mathbf{y},
 \end{aligned}$$

where $\mathbf{W}(\boldsymbol{\beta}^{(j)}) = \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(j)}))$.

- 3 Iterate step 2 until convergence: $\|\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}\|^2 < \delta_{\text{stop}}$
(Note: omitting details I will later show you when we discuss GLMs).

IRLS: Example

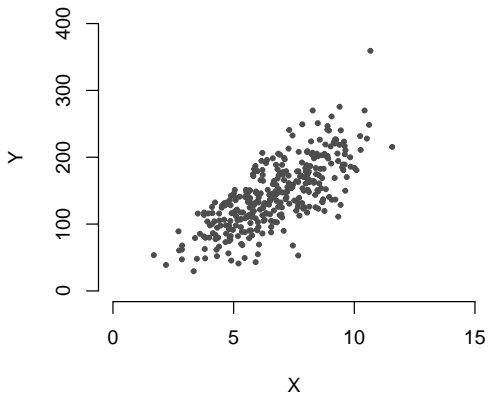
- Suppose $X \sim \mathcal{N}(\mu = 6.5, \sigma^2 = 4)$.
- Let $\boldsymbol{\mu} = 10 + 20\mathbf{x}$
- Suppose $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\epsilon_i \sim \mathcal{N}(\mu = 0, \sigma_i^2 = 8\mu_i)$.
- Put another way, $\mathbf{V}(\boldsymbol{\mu}) = \text{diag}(8\boldsymbol{\mu})$.

Example: Coding setup of example in R

```
set.seed(7345)
n <- 350
X <- rnorm(n, 6.5, 2)
mu <- 10 + 20*X
Y <- mu + rnorm(n, 0, sqrt(8*mu))
```

ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: Scatter plot



ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: IRLS in R

```
## Set threshold and initialize
iter <- 0
tol <- 1e-16
delta <- 1

zz <- lm(Y ~ X)
betaj <- as.numeric(zz$coefficients)
zz0 <- zz

while (delta > tol)
{
  ## Store previous iteration
  prior.betaj <- betaj

  ## Form weight matrix
  W <- diag(1/predict(zz))

  ## Run WLS
  zz <- lm(Y ~ X, weights = diag(W))

  ## Extract coefficient
  betaj <- as.numeric(zz$coefficients)

  ## Check for convergence and track number of iterations
  delta <- sum((betaj - prior.betaj)^2)
  iter <- iter + 1
}
```

ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: Results

```
## Report number of iterations
> print(iter)
[1] 7
```

```
## Verify convergence
> print(delta)
[1] 3.586288e-18
```

```
> summary(zz0)
```

Coefficients:

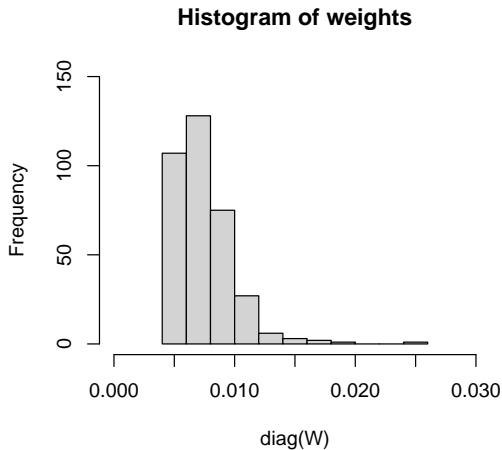
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.6414	6.9195	0.671	0.503	
X	20.7031	0.9971	20.764	<2e-16	***

```
> summary(zz)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	5.878	5.965	0.985	0.325	
X	20.519	0.924	22.207	<2e-16	***

ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: Check to make sure weights are sensible



IRLS: More sophisticated functions in R

- With the `glm()` function in R, you can specify more complex variance functions.
- Example: $\text{Var}[\epsilon] = (\theta_1 + |\mu|^{\theta_2})^2$, where θ_1 and θ_2 are also unknown parameters.
- θ_1 and θ_2 are also estimated iteratively (details omitted).

Example:

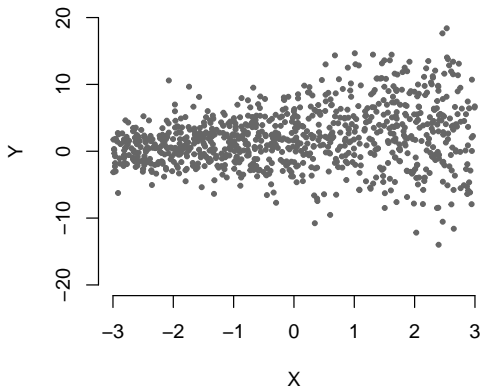
- Let $X \sim \text{Uniform}(-3, 3)$.
- Let $Y = 2 + 0.7X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 2 + |\mu|^{1.1})$.
 - ▶ Mean model: $Y = \beta_0 + \beta_1 x + \epsilon$.
 - ▶ Variance model: $\text{Var}[\epsilon] = (\theta_1 + |\mu|^{\theta_2})^2$, where θ_1 and θ_2 are the unknown parameters of the variance model.
- IRLS is based on $W(\mathbf{x}) = (\theta_1 + |\hat{Y}(\mathbf{x})|^{\theta_2})^{-2}$, where θ_1 , θ_2 , and β are estimated iteratively.

Example: Coding setup of example in R

```
set.seed(7345)
n <- 900
X <- runif(n, -3, 3)
Y <- 2 + 0.7*X + rnorm(n, 0, 2 + abs(2 + 0.7*X)^(1.1))
```

ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: Scatter plot



ITERATIVELY RE-WEIGHTED LEAST SQUARES

Example: GLS (generalized least squares)

```
## Required library for GLS
library(nlme)

zz1 <- gls(Y~X, weights = varConstPower())

## Salient part of R output

Variance function:
  Structure: Constant plus power of variance covariate
  Formula: ~fitted(.)
  Parameter estimates:
      const      power
2.902585 1.458620

Coefficients:
                Value  Std.Error  t-value p-value
(Intercept) 1.9545581 0.14316666 13.652327    0
X            0.6223332 0.07673958  8.109676    0
```

Example: Compare to results from `lm()`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.93307	0.14587	13.252	< 2e-16	***
X	0.57323	0.08525	6.724	3.14e-11	***

Notes from examples:

- WLS (whether fit through IRLS through as in the first example or the `gls()` function as in the second example) produced lower variance, which is not surprising as the variance function was correctly specified.
- If you fit a model using the `gls()` function in R, you'll notice it's using something called *restricted maximum likelihood* (REML) to obtain estimates. You'll talk about this in BIOS 7346, so I'm electing to avoid an in-depth discussion on it in this course.
- Keep in mind that $\text{Cov}[\hat{\boldsymbol{\beta}}]$ can be estimated in a plug-in fashion:

$$\begin{aligned} \text{Cov}[\hat{\boldsymbol{\beta}}] &= \hat{\sigma}^2 (\mathbf{X}^T \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\beta}}) \mathbf{X})^{-1} \\ &= \left(\frac{1}{N-K} \sum_{i=1}^N \frac{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2}{\hat{V}_i(\hat{\boldsymbol{\beta}})} \right) (\mathbf{X}^T \hat{\mathbf{V}}^{-1} (\hat{\boldsymbol{\beta}}) \mathbf{X})^{-1} \end{aligned}$$

Things to come:

- We have not yet considered what happens when the weights are “misspecified” (meaning that they are *not* actually inversely proportional to the variance), or how to correct this issue. We will eventually do so!
- Further, these ideas will generalize next semester, in which \mathbf{V} will be block-diagonal (rather than diagonal) to account for clusters.

So far:

- Weighted least squares.

Up next:

- Hypothesis testing.