BIOS 7345: Advanced Regression for Independent Data

Andrew J. Spieker, Ph.D.

Associate Professor of Biostatistics
Vanderbilt University

Set 4: Weighted least squares

Version: 09/12/2025

Table of Contents

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:

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

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

Weighted least squares

New setup: Diagonal variance

- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$.
- Specifically, $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$ 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:

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

- 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:
  - $\mathrm{E}[\boldsymbol{\epsilon}^*] = \mathrm{E}[\mathbf{K}\boldsymbol{\epsilon}] = \mathbf{K}\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
  - $\mathrm{Cov}[\boldsymbol{\epsilon}^*] = \mathrm{Cov}[\mathbf{K}\boldsymbol{\epsilon}] = \mathbf{K}\mathrm{Cov}[\boldsymbol{\epsilon}]\mathbf{K}^\intercal = \mathbf{V}^{-1/2}\sigma^2\mathbf{V}\mathbf{V}^{-1/2} = \sigma^2\mathbf{I}$.
- Applying OLS on the transformed data gives:

$$\widehat{\boldsymbol{\beta}}^* = ((\mathbf{X}^*)^\intercal\mathbf{X}^*)^{-1}(\mathbf{X}^*)^\intercal\mathbf{y}^* \quad = \quad (\mathbf{X}^\intercal\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{V}^{-1}\mathbf{y}.$$

Properties:

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

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

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

Theorem 4.1: The Gauss-Markov theorem (version 4)

Let $\mathbf{X}$ be of full rank, and suppose $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$. Then, for any constant vector $\mathbf{a}$, $\mathbf{a}^\intercal\widehat{\boldsymbol{\beta}}^*$ is the unique BLUE of $\mathbf{a}^\intercal\boldsymbol{\beta}$ (where $\widehat{\boldsymbol{\beta}}^*$ is the WLS estimator having weights given by $\mathbf{W} \propto \mathbf{V}^{-1}$).

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

Table of Contents

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.

Iteratively re-weighted least squares

Unknown variance:

- Consider $\mathbf{V} = \mathbf{V}(\boldsymbol{\mu}(\boldsymbol{\beta}); \boldsymbol{\theta})$, so the mean is a function of $\boldsymbol{\mu} = \mathrm{E}[\mathbf{y}]$ 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 initializer allows us to *estimate* $\mathrm{Cov}[\boldsymbol{\epsilon}]$, use that to inform our weights on the next round, and cycle through until apparent convergence. This procedure is known as iteratively re-weighted least squares (IRLS).
- If your model for $\mathrm{Cov}[\boldsymbol{\epsilon}]$ is correct, resulting $\widehat{\boldsymbol{\beta}}$ from IRLS is asymptotically BLUE for $\boldsymbol{\beta}$.

Iteratively re-weighted least squares

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 $\boldsymbol{\beta}^{(0)}$, the next step in a first-order Taylor approximation is given by:

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

   where $\mathbf{W}(\boldsymbol{\beta}^{(j)}) = \mathbf{V}^{-1}(\boldsymbol{\mu}(\boldsymbol{\beta}^{(j)}))$.
3. Iterate Step 2 until convergence: $||\boldsymbol{\beta}^{(j+1)} - \boldsymbol{\beta}^{(j)}||^2 < \delta_{\text{stop}}$
   - If $\boldsymbol{\beta}^{(0)}$ is initialized as a consistent estimator of $\boldsymbol{\beta}$ (e.g., OLS), you only need to take a single step in order to achieve an asymptotically efficient estimator. You'll see this again in BIOS 7346.

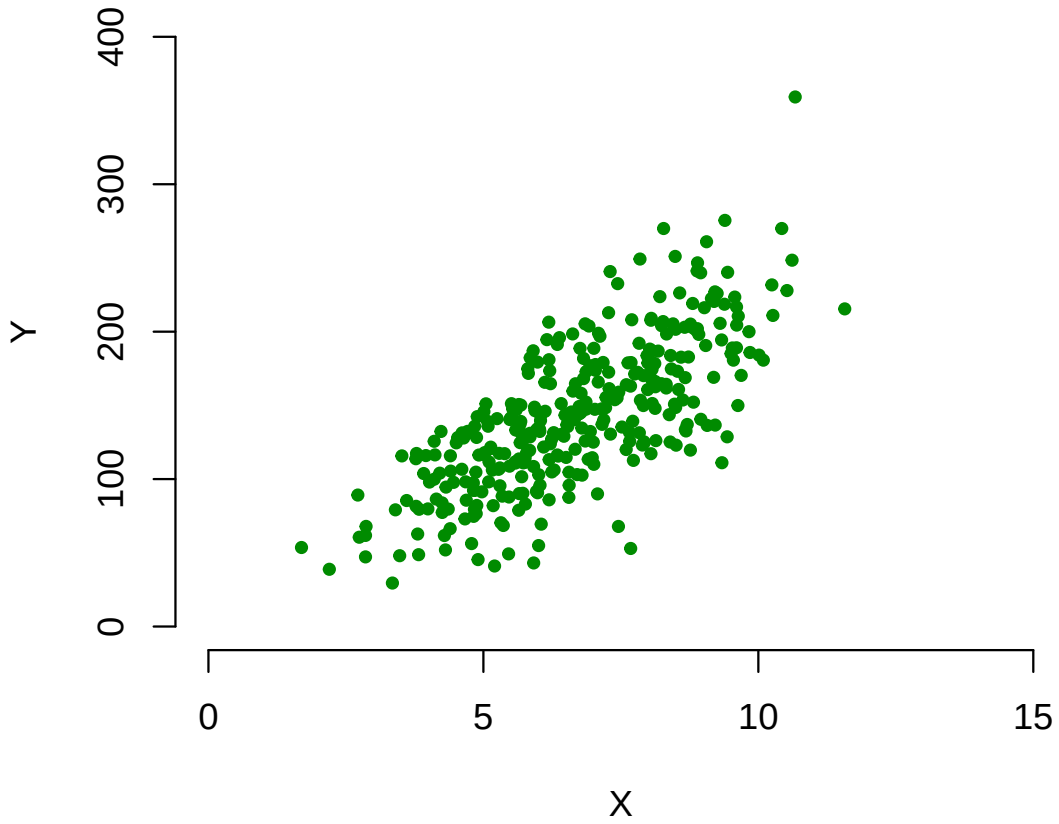
Example 4.1: IRLS

- 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 4.1: Coding setup of example in R

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

Example 4.1: Scatter plot



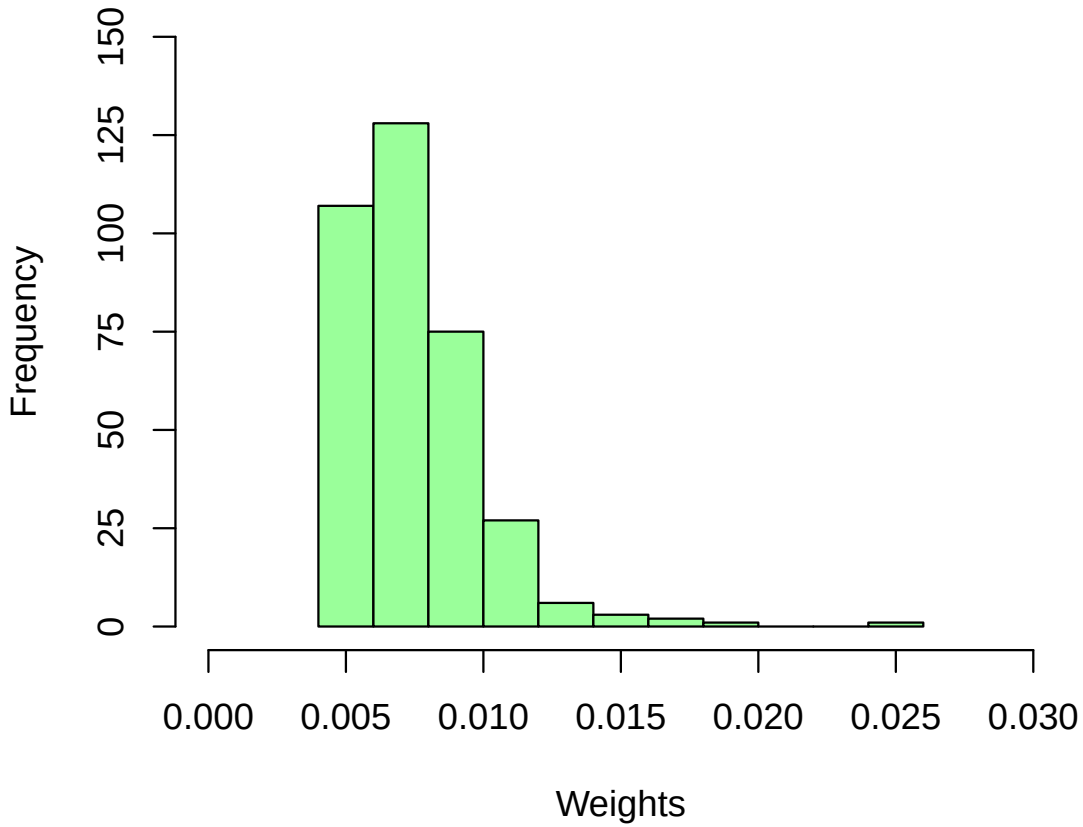
Example 4.1: IRLS in R

```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

  ## Weights
  W <- 1/predict(zz)

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

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

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

Example 4.1: Results

```
## Report number of iterations
> print(iter)
[1] 6

## Verify convergence
> print(delta)
[1] 3.849917e-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 ***
```

Example 4.1: Check to make sure weights are sensible



Iteratively re-weighted least squares

IRLS: More sophisticated functions in R

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

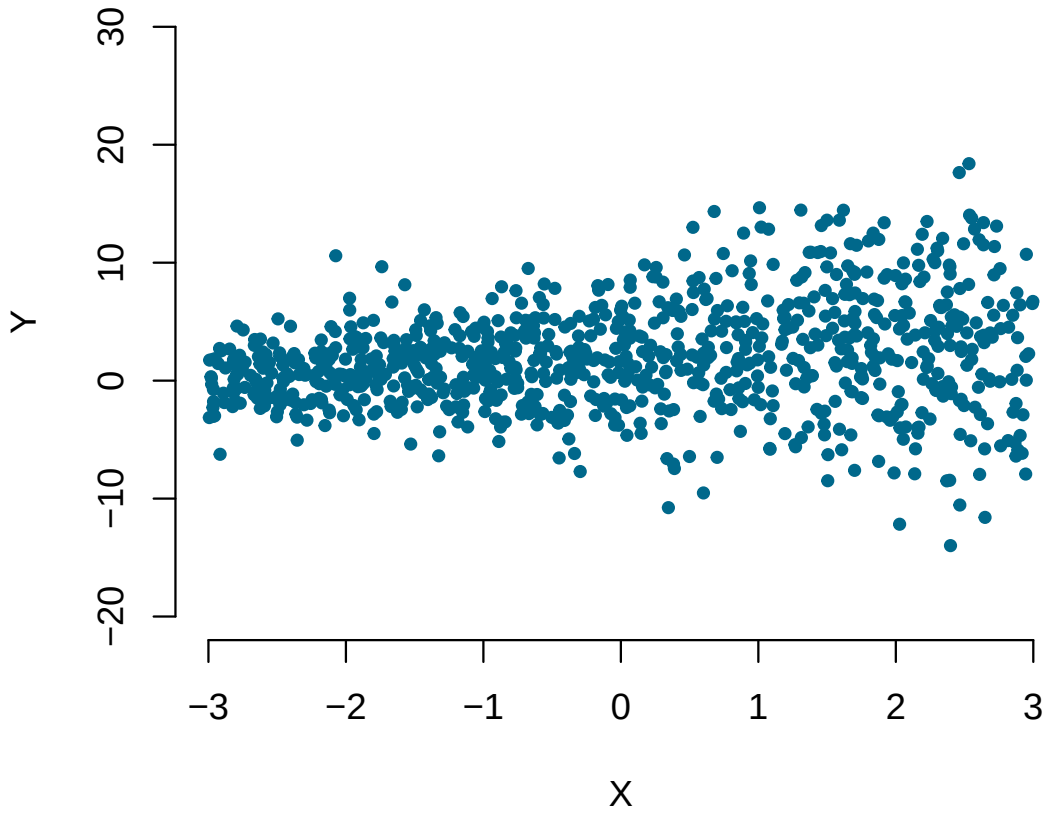
Example 4.2:

- 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: $\text{Var}[\epsilon_i] = (\theta_1 + |\boldsymbol{\mu}_i|_2^{\theta_2})^2$, where θ_1 and θ_2 are the unknown parameters of the variance model.
- One possibility: IRLS based on $W(\mathbf{x}) = (\theta_1 + |\widehat{Y}(\mathbf{x})|_2^{\theta_2})^{-2}$, where θ_1, θ_2, and $\boldsymbol{\beta}$ are estimated iteratively (details omitted).
- Another possibility: Maximum likelihood (although variance estiamte is biased).
- Most common: REML (restricted maximum likelihood).

Example 4.2: Coding setup of example in R

```
## Set seed for reproducibility
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))
```

Example 4.2: Scatter plot



Example 4.2: 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 4.2: 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 $\text{Cov}[\widehat{\boldsymbol{\beta}}]$ can be estimated in a plug-in fashion:

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

Further thoughts

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.

Summary

This unit:

- Weighted least squares.
- Gauss-Markov theorem for weighted least squares.
- Iteratively re-weighted least squares.

Summary: So far

- Random vectors and matrices; multivariate normal theory.
- Ordinary least squares.
- Hypothesis testing and ANOVA.
- Weighted least squares.

Summary: Coming up

- Misspecification.
- Confidence regions and prediction.
- Diagnostics.
- Regularization.
- Bayesian regression.
- Exponential families.
- Generalized linear models.
- Sandwich and bootstrap.
- Quasi-likelihood.
- Hypothesis testing for GLMs.
- Diagnostics for GLMs.
- Further considerations for binary outcomes.
- Nonlinear least squares.