BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 6: Weighted Least Squares and the Gauss-Markov Theorem

Version: 02/08/2023

Table of Contents

1. Weighted least squares
2. The Gauss-Markov theorem
3. Re-emphasizing the importance of the sandwich variance
4. Example: Height and FEV

Weighted least squares

Weighting:

- First, it's important to understand that the term "weighting" refers to a number of classes of approaches, only one of which we are going to discuss in this set of notes.
- I'll briefly summarize the kinds of weighting approaches we will *not* be discussing in detail at this time:
    1. Frequency weights.
    2. Inverse probability weights.
- We will instead be discussing the theory of *analytic* weights.

Weighted least squares

Frequency weights:

- Frequency weights are a essentially a matter of conveniently organizing data, allowing a single row in a data set to represent what should be additional (identical) rows in the data set.
- Associating a frequency weight of, say, $w_f = 2$ to an observation accomplishes the same thing as duplicating that row in the data set, thereby altering the relevant sample size.
- That's essentially all I have to say about frequency weights.

Weighted least squares

Inverse probability weights:

- Inverse probability weights are used to handle a number of challenges that arise in trying to use information from the sample that does not fully represent the target population.
    - Non-random treatment.
    - Missingness.
    - Censoring.
    - Selection bias.
- Weights are used to create some form of balance by giving observations additional weight if they were, by some metric, “less likely to be observed” than other observations in the sample.
- We will briefly cover this later in the course; you are not expected to understand this yet.

Weighted least squares

Analytic weights:

- We will be discussing analytic weights (sometimes called variance weights or regression weights).
- Usually, the idea is to give extra relative weight to observations for which the outcome variance is smaller (i.e., *without* actually changing the sample size).

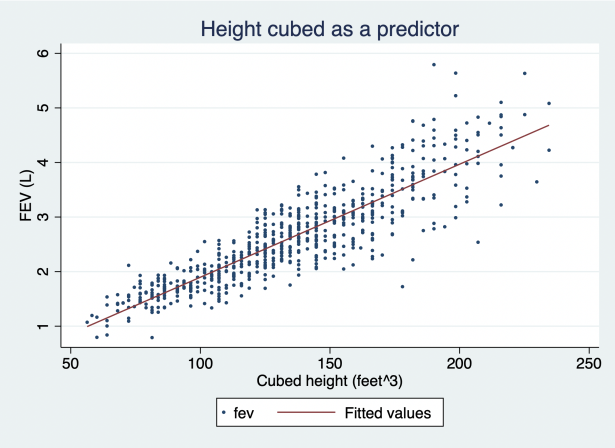
Weighted least squares

FEV: Cubed height and FEV

- Let us return to a previous example from the FEV data set in order to motivate the problem.
  - X: height cubed (cubic feet).
  - Y: FEV (L).
- Simple linear regression model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.

Weighted least squares

Example: Cubed height and FEV



Weighted least squares

FEV: Cubed height and FEV

- Note: the average value of cubed height is roughly 135 ft^3.
- Imagine that you hypothetically had the opportunity to sample one of the two additional observations to be added to this data set.
  - An individual with a cubed height of 115 ft^3.
  - An individual with a cubed height of 155 ft^3.
- Without any further information about this subject, which observation would you prefer to sample and why?
  - Note that these two observations have the same leverage (or at least approximately).

Weighted least squares

FEV: Cubed height and FEV

- Conceptually, observations with lower variance provide more information about an association (all else being equal). It therefore stands to reason that if observations with lower variance were given *higher* weight compared to observations with higher variance, we may be able to estimate the association more precisely.

Weighted least squares

Linear model: $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta}$

- Review of assumptions:
  - $\mathrm{E}[\epsilon|\mathbf{X} = \mathbf{x}] = 0$ (linearity).
  - Pairwise independent errors.
  - $\mathrm{Var}[Y|\mathbf{X} = \mathbf{x}] < \infty$.
- Recall that ordinary least squares (OLS) involves the minimization of sum of squared errors:

$$\text{minimize } ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \quad = \quad \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

$$\overset{\text{(math)}}{\Rightarrow} \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Weighted least squares

Optimization problem:

- With weighted least squares (WLS), we instead seek to minimize the weighted sum of squared errors:

$$\text{minimize } \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \overset{\text{(math)}}{\Rightarrow} \mathbf{X}^T\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$\Rightarrow \quad \mathbf{X}^T\mathbf{W}\mathbf{y} - \mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$\Rightarrow \quad \widehat{\boldsymbol{\beta}}_{\mathbf{W}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

- Note that **W** is being used to represent a *diagonal* matrix of weights.

Weighted least squares

Optimization problem:

- Notation:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{NK} \end{pmatrix}, \mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & w_N \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

- Weighted least squares formula:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{W}} \;=\; (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

- If all observations receive equal weight (as they have so far this semester), then **W** is the identity matrix and this formula reduces to the ordinary least squares formula from Slide 455.

Weighted least squares

Properties of WLS estimators:

- As was the case with OLS estimation, a WLS estimator is unbiased:

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] &= \mathrm{E}[\mathrm{E}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}|\mathbf{X}]] \\
&= \mathrm{E}[\mathrm{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}|\mathbf{X}]] \\
&= \mathrm{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathrm{E}[\mathbf{y}|\mathbf{X}]] \\
&= \mathrm{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta}] \\
&= \mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}.
\end{aligned}$$

Weighted least squares

WLS: General properties under a diagonal weight matrix, $\mathbf{W}$

- Suppose $\text{Var}[\mathbf{y}|\mathbf{X}] = \mathbf{V}$ for some diagonal matrix $\mathbf{V}$.
  - Please understand the distinction between $\mathbf{V}$ and $\mathbf{W}$.
  - $\mathbf{V}$ is a matrix that represents the true, unknown error variance.
  - $\mathbf{W}$ is a user-specified weight matrix.
- The variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$ is given by:

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \quad = \quad \text{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}].$$

- I have not (yet) provided a way to estimate $\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}]$.
  - Recall that $\mathbf{V}$ is unknown.
- As a foreshadowing of things to come (soon), is there a choice of $\mathbf{W}$ that makes this expression reduce down to something nice?
  - The mathematically convenient choice happens to have further theoretical justification. I love when that happens!

Table of Contents

1. Weighted least squares
2. The Gauss-Markov theorem
3. Re-emphasizing the importance of the sandwich variance
4. Example: Height and FEV

The Gauss-Markov Theorem

Theorem: Optimality of WLS (under the right choice of weights)

- Suppose that each of the following conditions is satisfied:
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}|\mathbf{X}] = \mathbf{0}$.
  - $\mathbf{X}$ has full rank ($N > K + 1$ and no collinearity).
  - $\mathrm{Var}[\mathbf{y}|\mathbf{X}] = \mathbf{V}$ (for a diagonal $\mathbf{V}$).
- Let $\mathbf{W}$ denote a weight matrix such that $\mathbf{W} \propto \mathbf{V}^{-1}$.
- Then, $\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$ has minimum variance among all unbiased linear estimators of $\boldsymbol{a}^T\boldsymbol{\beta}$.
  - Note: $\boldsymbol{a}^T\boldsymbol{\beta}$ represents any linear combination of the coefficients of $\boldsymbol{\beta}$ (e.g., β_1, $\beta_3 - \beta_2$, $\beta_1 + 2\beta_3$).
- This is referred to as the **Gauss-Markov** theorem.

The Gauss-Markov Theorem

Unpacking the meaning of BLUE

- BLUE: **B**est **L**inear **U**nbiased **E**stimator
- More specifically:
  - Best: "Lowest variance."
  - Linear: "Linear in $\mathbf{y}$."
  - Unbiased: "Unbiased for $\boldsymbol{\beta}$."
- Putting it together: When the weight matrix is chosen to be proportional to the inverse variance, no other linear (in $\mathbf{y}$) and unbiased (for $\boldsymbol{\beta}$) estimator has smaller variance than $\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$.
- For those in biostatistics, this is similar to the concept of MVUE (minimum-variance unbiased estimation), but we're only allowing *linear* estimators of $\boldsymbol{\beta}$.

The Gauss-Markov Theorem

Specific case: OLS is BLUE when $\mathbf{V} = \sigma^2\mathbf{I}$ (i.e., homoscedasticity)

- Here is the proof. Let $\mathbf{d}^T\mathbf{y}$ denote an unbiased linear estimator of $\mathbf{a}^T\boldsymbol{\beta}$. By unbiasedness, $\mathrm{E}[\mathbf{d}^T\mathbf{y}] = \mathbf{a}^T\boldsymbol{\beta} \Rightarrow \mathbf{d}^T\mathbf{X} = \mathbf{a}^T$.
- Further, $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) = \sigma^2\mathbf{d}^T\mathbf{d}$, and
- $\mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\sigma^2 = \mathbf{d}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}\sigma^2$.
- So, $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) - \mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{d}^T(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{d} \geq 0$.
- We have proven the minimum variance claim. What's more,
- $\mathrm{Var}(\mathbf{d}^T\mathbf{y}) = \mathrm{Var}(\mathbf{a}^T\widehat{\boldsymbol{\beta}}) \Leftrightarrow (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{d} = 0$.
  - Put another way, $\mathbf{d}^T = \mathbf{d}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and hence $\mathbf{d}^T\mathbf{y} = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{a}^T\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$.
- Therefore, $\mathbf{a}^T\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is the *unique* unbiased estimator having the minimum variance property.

The Gauss-Markov Theorem

Key idea: Optimality of OLS under homoscedasticity

- As we have previously discussed, some will simplistically say that OLS estimation relies on the assumption of homoscedasticity.
- Instead, the following more nuanced statements are more accurate:
  - The OLS estimator is the best (i.e., minimum variance) linear (in $\mathbf{y}$) unbiased (for $\boldsymbol{\beta}$) estimator of $\boldsymbol{\beta}$ under the assumptions of the Gauss-Markov theorm if the errors are homoscedastic.
  - Correctly estimating the variance of the OLS estimator relies on error homoscedasticity unless the sandwich variance estimator is used.
- The following statements are also correct:
  - The OLS estimator is unbiased even if homoscedasticity does not hold.
  - The OLS estimator is less efficient than WLS (with $\mathbf{W} \propto \mathbf{V}^{-1}$) if homoscedasticity does not hold.
  - The sandwich variance is agnostic to the mean-variance relationship.

Table of Contents

1. Weighted least squares
2. The Gauss-Markov theorem
3. Re-emphasizing the importance of the sandwich variance
4. Example: Height and FEV

Variance estimation

WLS: Expressing the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$:

- Suppose $\text{Var}[\mathbf{y}|\mathbf{X}] = \mathbf{V}$, where $\mathbf{V}$ is a diagonal matrix.
- As displayed on a prior slide,

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \;=\; \text{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}].$$

- Suppose that the optimal (in the Gauss-Markov sense) weights are chosen (that is, that $\alpha\mathbf{W} = \mathbf{V}^{-1}$). Then, this can be reduced:

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \;=\; \text{E}[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}].$$

- Notice the cancellation of α.

Variance estimation

WLS: Estimating the variance of $\widehat{\boldsymbol{\beta}}_{\mathbf{W}}$:

- Leveraging the assumption that $\alpha\mathbf{W} = \mathbf{V}^{-1}$,

$$\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \;=\; \text{E}[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}] = \text{E}[(\mathbf{X}^T\alpha\mathbf{W}\mathbf{X})^{-1}].$$

- That gets us a little closer to estimating $\text{Var}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}]$, as we know $\mathbf{W}$.
- Typically, α is estimated by method of moments:

$$\widehat{\alpha} \;=\; \frac{1}{N-(K+1)}\sum_{i=1}^{N} w_i(y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}})^2.$$

- This lends itself to the model-based (non-robust) variance estimator:

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \;=\; \widehat{\alpha}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}.$$

Variance estimation

Further notes:

- Suppose you propose a weight matrix, $\mathbf{W}$ in the hopes of improving efficiency, but are unwilling to go so far as to say that it perfectly captures the mean-variance relationship.
- Never fear! Sandwich standard errors can be used:

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}_{\mathbf{W}}] \;=\; (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{D}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1},$$

  where $\mathbf{D} = \text{diag}(\widehat{\epsilon}_i^2)$ is a diagonal matrix of squared residuals.
- This will be valid even when $\mathbf{W} \not\propto \mathbf{V}^{-1}$.

Variance estimation

Characterizing assumptions:

- Suppose the following:
  - $\text{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbf{x}^T\boldsymbol{\beta}$.
  - $\text{Var}[\mathbf{y}|\mathbf{X} = \mathbf{x}] = \mathbf{V}$ for a diagonal $\mathbf{V}$.
- The table below summarizes the properties of WLS under different estimation circumstances/choices.

$\mathbf{W} \propto \mathbf{V}^{-1}$	Variance	$\widehat{\boldsymbol{\beta}}_{\text{W}}$ unbiased?	$\widehat{\boldsymbol{\beta}}_{\text{W}}$ BLUE?	$\widehat{\text{SE}}(\widehat{\boldsymbol{\beta}}_{\text{W}})$ valid?
Yes	Model	**YES**	**YES**	**YES**
Yes	Sandwich	**YES**	**YES**	**YES**
No	Model	**YES**	**NO**	**NO**
No	Sandwich	**YES**	**NO**	**YES**

Variance estimation

Further notes:

- Choice of robust vs. non-robust standard error does not impact point estimate.
- Point estimates change depending upon weighting scheme (but the bias does not).
- If you *choose* to assume a particular variance structure, you can change the variability (efficiency) of an estimator.

Table of Contents

1. Weighted least squares

2. The Gauss-Markov theorem

3. Re-emphasizing the importance of the sandwich variance

4. Example: Height and FEV

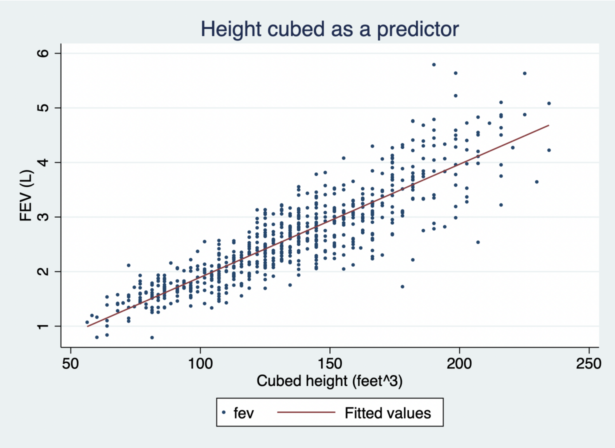
Example

FEV: Cubed height and FEV

- Recall the motivating example from the FEV data set.
  - X: height cubed (cubic feet).
  - Y: FEV (L).
- Simple linear regression model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.
- Consider two weighting schemes:
  1. Unweighted (or, in particular, $\mathbf{W} = \mathbf{I}$).
  2. Weighted (how do we decide weights?)
     - ⋆ Just for the purposes of illustration, I'm going to "eyeball" the conditional standard deviation of FEV at different heights and construct weights accordingly.
- For good measure, consider two variance estimators:
  1. Model-based (non-robust).
  2. Sandwich.

Example

Example: Cubed height and FEV



Example

FEV: Cubed height and FEV

- First, it appears that the standard deviation grows approximately linearly with height cubed. I therefore need to approximate the standard deviation at two points to come up with a suitable variance function (and in turn, a weight function).
- I'll make an educated visual guess that the standard deviation of FEV is about 0.3 at $X = 75$ and 0.8 at $X = 200$.
- This is to say that I am guessing the following function for the variance: $\text{Var}[Y|X = x] = (x/250)^2$.
- This suggests use of the following weights:

$$\mathbf{W}(x) = \frac{1}{(x/250)^2} \propto x^{-2}.$$

Example

FEV: Cubed height and FEV

- In Stata, you can generate the weights first and then incorporate them into a fitted regression model:
  - `gen htcb = (height/12)ˆ3`
  - `gen wts = 1/(htcbˆ2)`
- To incorporate weights, state so before including the options. For instance:
  - `regress fev htcb [weight = wts], robust`
- Before fitting the models, we should be able to take a lot of educated guesses right off the bat.

Further thoughts

Re-weighting: Practical issues

- Remember that we are assuming linearity to hold throughout. When linearity is violated, weighting can have unintended consequences if it's not done well.
    - Example: suppose $\mathrm{E}[Y|X=x]=x^2$ and $\mathrm{Var}[Y|X=x]\propto 1/x$. How might this have consequences for estimation if a linear model is assumed?
- Note that our focus has been on linear unbiased estimators. Don't be fooled! It is possible to achieve greater efficiency using a different kind of estimator (i.e., non-linear, biased).

Summary

Notes: Topics in this unit

- Interpretation and assumptions.
- Categorical predictors.
- Interaction and effect modification.
- Subgroup-specific effects.
- Variable transformations, basis expansions, and nonlinearity.
- Prediction intervals and diagnostics
- Weighted least squares and the Gauss-Markov theorem.

Summary

Notes: Next unit

- Binary outcome regression.
- Regression with categorical outcomes.
- Regression with ordinal outcomes.
- Regression with count outcomes.