

# Linear Models Lecture 6: Finite Sample Properties

Robert Gulotty

University of Chicago

February 26, 2026

## Why use $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ in my work?

- Motive 1: best approximation to the conditional expectation function (CEF).
  - While the conditional mean  $E(Y|\mathbf{X})$  is the best predictor of  $Y$ , its form is typically unknown.
  - The linear model is an approximation to the conditional mean that has the lowest mean squared error among linear predictors.
- Motive 2: Sampling Properties.
  - When the underlying CEF is linear and the residual variance is constant, OLS is optimal.
  - OLS is the Best Linear Unbiased Estimator, here meaning the minimum variance (most efficient) linear estimator for the parameter vector  $\beta$ .

# Why Study Finite Sample Properties?

- Finite sample results hold *exactly* for any sample size  $n$ —they do not require  $n \rightarrow \infty$ .
- The finite sample framework answers four questions:
  - 1 Is  $\hat{\beta}$  centered on  $\beta$ ?  $\rightarrow$  **Unbiasedness**.
  - 2 How much does  $\hat{\beta}$  vary across samples?  $\rightarrow$  **Variance**.
  - 3 Can any other estimator do better?  $\rightarrow$  **Efficiency (Gauss-Markov)**.
  - 4 What if  $\text{var}[\mathbf{e}|\mathbf{X}] \neq \sigma^2 \mathbf{I}$ ?  $\rightarrow$  **GLS**.
- These results provide the theoretical foundation for standard errors, confidence intervals, and hypothesis tests.
- They also reveal *when* OLS is and is not optimal, motivating the alternatives we study in this lecture and beyond.

## Finite Sample Assumptions of OLS estimator

- Assumption 1: The random variables  $\{(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)\}$  are independent and identically distributed.
- Assumption 2:  $Y = \mathbf{X}'\beta + e$ , where  $\mathbb{E}(e|\mathbf{X}) = 0$ .
- Assumption 3:  $\mathbb{E}(\mathbf{X}\mathbf{X}') > 0$  is invertible (with probability 1).
- Assumption 4\*:  $\mathbb{E}[e^2|\mathbf{X}] = \sigma^2(\mathbf{X}) = \sigma^2$ .

## When Are These Assumptions Reasonable?

- **A1 (i.i.d.):** Holds well for cross-sectional surveys with random sampling. Fails with time-series data (serial correlation), clustered data (students within schools), or panel data.
- **A2 ( $\mathbb{E}[e|\mathbf{X}] = 0$ ):** Requires that no omitted variable is correlated with  $\mathbf{X}$ . Fails when there is selection bias, simultaneity, or measurement error in  $\mathbf{X}$ . This is the assumption most contested in applied work.
- **A3 (rank condition):** Fails with perfect multicollinearity (e.g. including a dummy for every category plus an intercept). Nearly violated when regressors are highly collinear.
- **A4\* (homoskedasticity):** Rarely holds exactly. Variance of earnings typically grows with education; variance of GDP growth differs across countries. Violation does *not* bias  $\hat{\beta}$ , but makes OLS inefficient and standard errors wrong.

## Goals for Sampling Properties.

- **unbiased:** The (conditional) expectation of the estimator  $\hat{\beta}$  of  $\beta$  is equal to the parameter.  $\mathbb{E}[\hat{\beta}] = \beta$
- **efficient:** The estimator  $\hat{\beta}$  of  $\beta$  has a lower variance matrix than other estimators.
- The latter efficiency result is called the Gauss Markov Theorem and depends on Assumption 4\*.

# Unbiasedness

- We will show that  $\mathbb{E}[\hat{\beta}|\mathbf{X}] = \beta$  using three methods: summation, matrix, and decomposition.
- Each relies heavily on the conditioning theorem:

$$\mathbb{E}[g(\mathbf{X})Y|\mathbf{X}] = g(\mathbf{X})\mathbb{E}[Y|\mathbf{X}]$$

and

$$\mathbb{E}[g(\mathbf{X})Y] = \mathbb{E}[g(\mathbf{X})\mathbb{E}[Y|\mathbf{X}]]$$

# Unbiasedness of the OLS slope estimator: Version 1

- Assume  $\mathbf{x}_i$  is a  $k \times 1$  vector representing observation  $i$
- $\mathbb{E}[Y_i | \mathbf{x}_1, \dots, \mathbf{x}_n] = \mathbb{E}[Y_i | \mathbf{x}_i]$  because of independence of observations across  $i$ .

## Unbiasedness (shown in summation operation)

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\mathbf{x}_1, \dots, \mathbf{x}_n] &= \mathbb{E}\left[\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i Y_i\right) \middle| \mathbf{x}_1, \dots, \mathbf{x}_n\right] \\&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbb{E}\left[\left(\sum_{i=1}^n \mathbf{x}_i Y_i\right) \middle| \mathbf{x}_1, \dots, \mathbf{x}_n\right] && \text{(Conditioning Theorem)} \\&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^n \mathbb{E}[\mathbf{x}_i Y_i | \mathbf{x}_1, \dots, \mathbf{x}_n] && \text{(Linearity of Expectations)} \\&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbb{E}[Y_i | \mathbf{x}_i] && \text{(Conditioning Theorem, and independence)} \\&= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \beta && \text{(Linear conditional expectation)} \\&= \beta && \text{(Inverse)}\end{aligned}$$

## Unbiasedness of the OLS slope estimator: Version 2

Using Matrix notation,  $\mathbb{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ ,  $\mathbb{E}[\mathbf{e}|\mathbf{X}] = 0$  (by assumption 2)

$$\begin{aligned}\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{y}|\mathbf{X}] && \text{(Conditioning Theorem)} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta && \text{(Independence)} \\ &= \beta && \text{(Inverse)}\end{aligned}$$

## Unbiasedness of the OLS slope estimator: Version 3

Decomposition, noting  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) && \text{(plug in)} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} && \text{(distribute)} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} && \text{(Inverse)}\end{aligned}$$

So now we can just check that  $\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] = 0$  by applying assumption 2.

$$\mathbb{E}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|\mathbf{X}] = \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{e}|\mathbf{X}] = 0$$

## Variance of Least Squares Estimator

We can write the variance of the OLS estimator in terms of

$$\mathbf{D} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Where  $\sigma_i^2$  gives us the variation in the regression error for observation  $i$ .

## Derivation of Variance of $\hat{\beta}$

$$\begin{aligned}\text{var}[\hat{\beta}|\mathbf{X}] &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}] \\ &= E[((\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}) - \beta)((\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}) - \beta)'|\mathbf{X}] \\ &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e})'|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Note that  $\mathbf{X}'\mathbf{D}\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\sigma_i^2$

## The Sandwich Formula in Practice

- The variance  $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$  is called the **sandwich formula**:  $(\mathbf{X}'\mathbf{X})^{-1}$  is the “bread” and  $\mathbf{X}'\mathbf{D}\mathbf{X}$  is the “meat.”
- This is the formula behind every “robust standard error” you see in applied papers.
- In R: `vcovHC(lm_model, type="HC2")` computes this variance using  $\hat{e}_i^2$  to estimate the diagonal of  $\mathbf{D}$ .
- Under homoskedasticity ( $\mathbf{D} = \sigma^2\mathbf{I}$ ), the sandwich simplifies to  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ —the classical formula. Robust SEs and classical SEs agree.
- When they diverge, it signals heteroskedasticity in your data.

## Special case, $\sigma_i^2 = \sigma_j^2 = \sigma^2$

- If  $\mathbf{D} = \mathbf{I}_n \sigma^2$ , (assumption 4\*), then we call the error homoskedastic.

$$\begin{aligned} \text{var}[\hat{\beta}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{I}_n \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

- Homoskedasticity is a convenient assumption, but it also has played an important role in offering intellectual support for the linear model.

# Proof of Gauss Markov Theorem

Formal statement: In the homoskedastic linear model, if  $\tilde{\beta}$  is a linear unbiased estimator of  $\beta$ , then

$$\text{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

## Proof (4 steps):

- 1 Let  $\mathbf{A}$  be any  $n \times k$  linear function of  $\mathbf{X}$  such that  $\mathbf{A}'\mathbf{X} = \mathbf{I}_k$ .  
By (A2) and (A3),  $\mathbf{A}'\mathbf{Y} = \tilde{\beta}$  is an *unbiased* estimator of  $\beta$ , as  
 $E[\tilde{\beta}|\mathbf{X}] = E[\mathbf{A}'\mathbf{Y}|\mathbf{X}] = \mathbf{A}'E[\mathbf{Y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\beta = \beta$ .
- 2 Under (A1) and (A4), the variance of  $\hat{\beta}$  is  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$  and the variance of  $\tilde{\beta}$  is  $\sigma^2\mathbf{A}'\mathbf{A}$ .
- 3 Evaluate  $\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}$  to show that it is positive semi-definite.

## Proof of Gauss Markov Theorem

Set  $\mathbf{C} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ , and note that  $\mathbf{X}'\mathbf{C} = 0$ ,

$$\begin{aligned}\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= (\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})'(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) - (\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{C}'\mathbf{C} + \mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\ &= \mathbf{C}'\mathbf{C}\end{aligned}$$

4 Any matrix that can be written as a product  $\mathbf{C}'\mathbf{C}$  is positive semi-definite:

If  $\mathbf{M} = \mathbf{C}'\mathbf{C}$ , then  $\mathbf{x}'\mathbf{M}\mathbf{x} = \mathbf{x}'\mathbf{C}'\mathbf{C}\mathbf{x} = \|\mathbf{C}\mathbf{x}\|^2 \geq 0$ .

So  $\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}$  is positive semidefinite and

$$\text{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \square$$

No linear estimator will get a lower variance! OLS is the **best** among **linear unbiased** estimators (BLUE).

## What Does BLUE Mean for Applied Work?

- Under homoskedasticity, OLS gives you the *tightest possible* confidence intervals among linear unbiased estimators.
- Practical implication: if  $\text{var}[\mathbf{e}|\mathbf{X}] = \sigma^2 \mathbf{I}$ , there is no reason to search for a cleverer estimator—OLS is already optimal.
- What BLUE does *not* guarantee:
  - If errors are heteroskedastic, OLS is no longer efficient  $\rightarrow$  GLS can do better.
  - If  $\mathbb{E}[\mathbf{e}|\mathbf{X}] \neq 0$  (omitted variables, simultaneity), OLS is biased regardless of efficiency.
  - BLUE says nothing about nonlinear estimators (MLE may dominate if you know the error distribution).
- Bottom line: Gauss-Markov tells you *when you can stop looking* for a better estimator, and when you cannot.

## Gauss-Markov in Action: OLS vs. Alternatives

```
set.seed(42); B <- 5000; n <- 50
b_ols <- b_split <- numeric(B)
for (i in 1:B) {
  x <- rnorm(n); e <- rnorm(n, 0, 2)
  y <- 2 + 3*x + e
  b_ols[i] <- coef(lm(y ~ x))[2]
  h1 <- 1:(n/2); h2 <- (n/2+1):n
  b_split[i] <- (coef(lm(y[h1] ~ x[h1]))[2]
    + coef(lm(y[h2] ~ x[h2]))[2])/2
}
c(var(b_ols), var(b_split))
# ~ 0.08 ~ 0.16
```

- DGP:  $Y = 2 + 3X + e$ ,  $e \sim N(0, 4)$ .
- Split-sample: run OLS on each half, average the slopes. Linear, unbiased—but wastes information.
- OLS variance  $\approx$  half the split-sample variance.

Under homoskedasticity, OLS uses all the information in the data. Any other linear unbiased estimator wastes some.

## Hansen's Gauss Markov Theorem (Technical)

In the homoskedastic linear model, if  $\tilde{\beta}$  is ~~a linear~~ *any* unbiased estimator of  $\beta$ , then

$$\text{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

This result is derived using the Cramér-Rao bound.

## Theoretical Criteria for Estimators

- Gauss Markov: OLS is "best" by reference to any alternative unbiased linear estimator.
- Cramér-Rao bound: the precision (inverse of the variance) of any unbiased estimator is bounded by the Fisher information of the estimator.
- We will define "information" more formally when we get to Maximum Likelihood Estimation.

## Method of Moments Estimation

- A **method of moments estimator** (MME) sets sample moments equal to population moments and solves for the parameter.
- Population moment condition:  $\mathbb{E}[g(\mathbf{x}_i, Y_i, \theta)] = 0$
- Sample analog:  $\frac{1}{n} \sum_{i=1}^n g(\mathbf{x}_i, Y_i, \hat{\theta}) = 0$
- OLS is a method of moments estimator:
  - Population:  $\mathbb{E}[\mathbf{x}_i(Y_i - \mathbf{x}_i'\beta)] = 0$  (from  $\mathbb{E}[\mathbf{x}_i e_i] = 0$ )
  - Sample:  $\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i(Y_i - \mathbf{x}_i'\hat{\beta}) = 0 \Rightarrow \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0$
  - Solving gives  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ —the OLS estimator.

## Method of Moments for $\sigma^2$

- Under homoskedasticity, the population moment is  $\mathbb{E}[e_i^2] = \sigma^2$ .
- The MME replaces  $e_i$  with  $\hat{e}_i$  and averages:

$$\hat{\sigma}_{MM}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2$$

- This is biased:  $\mathbb{E}[\hat{\sigma}_{MM}^2 | \mathbf{X}] = \frac{n-k}{n} \sigma^2$  (shown later in the residuals section).
- The bias-corrected version  $s^2 = \frac{1}{n-k} \sum \hat{e}_i^2$  is the standard estimator in practice.
- Method of moments generalizes to **GMM** (Generalized Method of Moments) when we have more moment conditions than parameters.

# Spherical Errors

- The Gauss Markov Theorem assumed that  $E(\mathbf{ee}') = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$
- In quadratic form,  $\sigma^2 \mathbf{I}$  is the formula for a sphere

$$\begin{aligned} \mathbf{e}'(\sigma^2 \mathbf{I})\mathbf{e} &= \sigma^2 e_1^2 + \sigma^2 e_2^2 + \cdots + \sigma^2 e_n^2 = q \\ &= e_1^2 + e_2^2 + \cdots + e_n^2 = q/\sigma^2 \end{aligned}$$

## Non-Spherical Errors

- More generally we think that errors may exhibit variation: If

$$E(\mathbf{ee}') = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}, \sigma_i \neq \sigma_j, \text{ then we say we have heteroskedasticity.}$$

- Heteroskedasticity can occur when cases are aggregates of uneven size.

$$\text{■ If } E(\mathbf{ee}') = \begin{bmatrix} \sigma^2 & a & \cdots & b \\ a & \sigma^2 & \cdots & d \\ \vdots & \vdots & \cdots & \vdots \\ b & d & \cdots & \sigma^2 \end{bmatrix}, a \neq b \neq d \neq 0, \text{ then we say we have autocorrelation.}$$

- More generally, we will allow for both.

# Where Do Non-Spherical Errors Arise?

## ■ Heteroskedasticity:

- Cross-country regressions: richer countries often have more variable outcomes (GDP growth, trade flows).
- Earnings regressions: variance of wages increases with education and experience.
- Any regression where the outcome is bounded below (e.g. expenditure  $\geq 0$ ): variance shrinks near the bound.

## ■ Autocorrelation:

- Time series: an economic shock in one quarter persists into the next.
  - Panel data: unobserved country or individual traits make errors within a unit correlated.
  - Spatial data: neighboring counties share unobserved shocks (weather, policy spillovers).
- In all these cases, OLS  $\hat{\beta}$  is still unbiased, but its variance is *not*  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ —classical standard errors are wrong.

## Generalized Least Squares

- Consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- Allow for heteroskedasticity or correlation in the errors:

$$\mathbb{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$$

$$\text{var}[\mathbf{e}|\mathbf{X}] = \Sigma\sigma^2$$

- Here  $\Sigma$  is  $n \times n$  and may be a function of  $\mathbf{X}$ ,  $\sigma^2$  is the common component to the diagonals (could be 1).
- Recall, Gauss-Markov proves we can do no better than OLS when  $\text{var}[\mathbf{e}|\mathbf{X}] = \sigma^2\mathbf{I}$ . But now we can do better.

## Derivation of Variance (again)

$$\begin{aligned} \text{var}[\hat{\beta}] &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[((\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}) - \beta)((\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}) - \beta)'] \\ &= E[((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e})'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

## Lower Bound on Variance of Estimators

Theorem: Given assumptions of linear model, if  $\tilde{\beta}$  is an unbiased estimator of  $\beta$ , then

$$\text{var}[\tilde{\beta}] \geq \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$$

All we need is to know  $\Sigma^{-1}$ , then correct for it.

## Proof: Efficiency Lower Bound for Linear Estimators (Hansen Ex. 4.6)

**Theorem:** If  $\tilde{\beta} = \mathbf{A}'\mathbf{Y}$  is linear and unbiased, then  $\text{var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$ .

**Proof:** Transform to the spherical model via  $\Sigma^{-1/2}$ :

- 1 Define  $\tilde{\mathbf{Y}} = \Sigma^{-1/2}\mathbf{Y}$ ,  $\tilde{\mathbf{X}} = \Sigma^{-1/2}\mathbf{X}$ ,  $\tilde{\mathbf{e}} = \Sigma^{-1/2}\mathbf{e}$ , so  $\text{var}[\tilde{\mathbf{e}}|\mathbf{X}] = \sigma^2\mathbf{I}_n$ .
- 2 Write  $\tilde{\beta} = \mathbf{A}'\mathbf{Y} = \mathbf{A}'\Sigma^{1/2}\tilde{\mathbf{Y}} \equiv \tilde{\mathbf{A}}'\tilde{\mathbf{Y}}$ .
- 3 Unbiasedness:  $\tilde{\mathbf{A}}'\tilde{\mathbf{X}} = \mathbf{A}'\Sigma^{1/2}\Sigma^{-1/2}\mathbf{X} = \mathbf{A}'\mathbf{X} = \mathbf{I}_k$ .
- 4 Apply Gauss-Markov to the transformed (spherical) model:

$$\text{var}[\tilde{\beta}|\mathbf{X}] = \sigma^2\tilde{\mathbf{A}}'\tilde{\mathbf{A}} \geq \sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} \quad \square$$

**Intuition:**  $\Sigma^{-1/2}$  converts any non-spherical model into a spherical one. In the spherical world, OLS (= GLS on original data) is already BLUE.

## Aside: Matrix Decomposition

- Any matrix  $\Sigma$  that is positive definite and symmetric can be factored:

$$\Sigma = \mathbf{C}\mathbf{\Lambda}\mathbf{C}'$$

- The columns of  $\mathbf{C}$  are the eigenvectors of  $\Sigma$ .
- $\mathbf{\Lambda}$  is a diagonal matrix of the eigenvalues of  $\Sigma$ .
- $\mathbf{\Lambda}^{1/2}$  is a diagonal matrix of the square roots of the eigenvalues of  $\Sigma$ .

$$\Sigma = \mathbf{C}\mathbf{\Lambda}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{C}' = \mathbf{T}\mathbf{T}'$$

$$\Sigma^{-1} = \mathbf{C}\mathbf{\Lambda}^{-1}\mathbf{C}' = \mathbf{C}\mathbf{\Lambda}^{-1/2}\mathbf{\Lambda}^{-1/2}\mathbf{C}' = \Sigma^{-1/2}(\Sigma^{-1/2})'$$

## Aitken (1935) Generalized Least Squares

- If we know  $\Sigma$ , we can do no better than to premultiply our linear model with  $\Sigma^{-1/2}$ ,

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \Sigma^{-1/2}\mathbf{y} &= \Sigma^{-1/2}\mathbf{X}\boldsymbol{\beta} + \Sigma^{-1/2}\mathbf{e} \\ \tilde{\mathbf{y}} &= \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{e}} \\ \tilde{\boldsymbol{\beta}}_{GLS} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= ((\Sigma^{-1/2}\mathbf{X})'(\Sigma^{-1/2}\mathbf{X}))^{-1}(\Sigma^{-1/2}\mathbf{X})'(\Sigma^{-1/2}\mathbf{y}) \\ &= (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}\end{aligned}$$

Here we have  $E[\tilde{\boldsymbol{\beta}}_{GLS}] = \boldsymbol{\beta}$  and  $var(\tilde{\boldsymbol{\beta}}_{GLS}) = \sigma^2(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}$

## GLS is BLUE

- Suppose  $b$  is an alternative linear unbiased estimator that differs by  $A$

$$\begin{aligned}b &= [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A] Y \\&= [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A] (X \beta + e) \\&= [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A] X \beta + [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A] e\end{aligned}$$

$$AX = 0 \quad (\text{Because } b \text{ is unbiased.})$$

$$\begin{aligned}\text{var}(b) &= \sigma^2 [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A] \Sigma [(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} + A]' \\&= \sigma^2 [(X' \Sigma^{-1} X)^{-1} + A \Sigma A' + (X' \Sigma^{-1} X)^{-1} X' A' + AX (X' \Sigma^{-1} X)^{-1}] \\&= \sigma^2 [(X' \Sigma^{-1} X)^{-1} + A \Sigma A' + 0 + 0] \\&= \sigma^2 (X' \Sigma^{-1} X)^{-1} + \sigma^2 A \Sigma A'\end{aligned}$$

- $\Sigma$  is positive definite, so  $\sigma^2 A \Sigma A' \geq 0$ ; minimum variance requires  $A = 0$



## GLS Properties

- GLS estimators are unbiased and efficient
- Small difference,  $P_* = \mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}$  is not symmetric, but otherwise, it is OLS.
- However, it requires knowing  $\Sigma$ , which makes it infeasible.
- Later we will show how to use a two-step procedure to estimate  $\Sigma$ .

## When GLS Beats OLS: A Two-Group Example

- Survey data from two groups: Group A (rich,  $n_A = 50$ ,  $\sigma_A^2 = 100$ ) and Group B (poor,  $n_B = 50$ ,  $\sigma_B^2 = 1$ ).
- OLS gives *equal weight* to every observation  $\rightarrow$  noisy Group A observations inflate variance.
- GLS (= WLS here) weights by  $w_i = 1/\sigma_i^2$ : Group A gets weight 1/100, Group B gets weight 1.

**Variance comparison** (scalar case, equal group sizes):

$$\text{var}(\hat{\beta}_{OLS}) \propto \frac{\sigma_A^2 + \sigma_B^2}{2} = \frac{101}{2} \quad \text{var}(\hat{\beta}_{GLS}) \propto \frac{1}{1/\sigma_A^2 + 1/\sigma_B^2} = \frac{100}{101} \approx 1$$

GLS is just common sense: trust precise observations more. When you know which observations are noisier, use that information.

## Comparing OLS and WLS: R Example

```
set.seed(99); B <- 5000; n <- 100
b_ols <- b_wls <- numeric(B)

for (i in 1:B) {
  x <- rnorm(n)
  # Two groups: first 50 noisy, last 50 precise
  sigma <- c(rep(10, 50), rep(1, 50))
  y <- 1 + 2*x + rnorm(n, 0, sigma)

  b_ols[i] <- coef(lm(y ~ x))[2]
  b_wls[i] <- coef(lm(y ~ x, weights = 1/sigma^2))[2]
}

c(sd(b_ols), sd(b_wls)) # WLS SE is smaller
# ~ 1.01 ~ 0.14
```

- Both estimators are unbiased ( $\bar{\hat{\beta}} \approx 2$ ), but WLS standard errors are  $\sim 7\times$  smaller.
- In practice, use `lm(..., weights = 1/sigma_hat^2)` when group variances are known or estimable.

# Residuals

- We will be using estimates of the residuals to construct covariance matrix estimators that do not require homoskedasticity.
- The residuals  $\hat{e}_i = Y_i - \mathbf{x}_i'\hat{\beta}$  can be written in vector notation as  $\hat{\mathbf{e}} = \mathbf{M}\mathbf{e}$ .
- Recall  $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Given  $\mathbf{X}$ , the expectation of the residuals is zero :

$$\mathbb{E}[\hat{\mathbf{e}}|\mathbf{X}] = \mathbb{E}[\mathbf{M}\mathbf{e}|\mathbf{X}] = \mathbf{M}\mathbb{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$$

- Given  $\mathbf{X}$ , the variance of the residuals is:

$$\text{var}[\hat{\mathbf{e}}|\mathbf{X}] = \text{var}[\mathbf{M}\mathbf{e}|\mathbf{X}] = \mathbf{M}\text{var}[\mathbf{e}|\mathbf{X}]\mathbf{M} = \mathbf{MDM}$$

## Homoskedastic Errors

- If we assume homoskedasticity,  $\mathbf{E}[e^2|X] = \sigma^2$

$$\text{var}[\hat{\mathbf{e}}|\mathbf{X}] = \text{var}[\mathbf{M}\mathbf{e}|\mathbf{X}] = \mathbf{M}\text{var}[\mathbf{e}|\mathbf{X}]\mathbf{M} = \mathbf{M}\mathbf{I}\sigma^2\mathbf{M} = \mathbf{M}\sigma^2$$

- Note that the  $i$ th diagonal element of  $\mathbf{M}$  is  $1 - h_{ii}$ , so

$$\text{var}[\hat{e}_i|\mathbf{X}] = \mathbb{E}[\hat{e}_i^2|\mathbb{X}] = (1 - h_{ii})\sigma^2 \neq \sigma^2$$

- $\hat{e}_i^2$  is a biased estimator and it is heteroskedastic.

## Prediction Errors

- The prediction errors  $\tilde{e}_i = Y_i - \mathbf{x}_i \hat{\beta}_{-i} = (1 - h_{ii})^{-1} \hat{e}_i$ .
- We defined  $\mathbf{M}^* = \text{diag}\{(1 - h_{11})^{-1}, (1 - h_{22})^{-1}, \dots, (1 - h_{nn})^{-1}\}$ .
- $\tilde{\mathbf{e}} = \mathbf{M}^* \hat{\mathbf{e}} = \mathbf{M}^* \mathbf{M} \mathbf{e}$
- Given  $\mathbf{X}$ , the expectation of the prediction errors is zero :

$$\mathbb{E}[\tilde{\mathbf{e}}|\mathbf{X}] = \mathbf{M}^* \mathbf{M} \mathbb{E}[\mathbf{e}|\mathbf{X}] = \mathbf{0}$$

- Given  $\mathbf{X}$ , the variance of the prediction errors is:

$$\text{var}[\tilde{\mathbf{e}}|\mathbf{X}] = \text{var}[\mathbf{M}^* \mathbf{M} \mathbf{e}|\mathbf{X}] = \mathbf{M}^* \mathbf{M} \text{var}[\mathbf{e}|\mathbf{X}] \mathbf{M} \mathbf{M}^* = \mathbf{M}^* \mathbf{M} \mathbf{D} \mathbf{M} \mathbf{M}^*$$

## Prediction Errors under homoskedasticity

- Under homoskedasticity, the variance of the prediction errors is:

$$\text{var}[\tilde{\mathbf{e}}|\mathbf{X}] = \text{var}[\mathbf{M}^* \mathbf{M} \mathbf{e}|\mathbf{X}] = \mathbf{M}^* \mathbf{M} \text{var}[\mathbf{e}|\mathbf{X}] \mathbf{M} \mathbf{M}^* = \mathbf{M}^* \mathbf{M} \sigma^2 \mathbf{M} \mathbf{M}^* = \mathbf{M}^* \mathbf{M} \mathbf{M}^* \sigma^2$$

- The variance of the  $i$ th prediction error is then:

$$\begin{aligned} \text{var}[\tilde{e}_i|\mathbf{X}] &= \mathbb{E}[\tilde{e}_i^2|\mathbf{X}] \\ &= (1 - h_{ii})^{-1} (1 - h_{ii}) (1 - h_{ii})^{-1} \sigma^2 \\ &= (1 - h_{ii})^{-1} \sigma^2 \end{aligned}$$

- $\sum \tilde{e}_i^2 = \sum ((1 - h_{ii})^{-1} \hat{e}_i)^2 = \text{PRESS}$ , "predictive error".

## Standardized Residuals

- $\text{var}[\hat{\mathbf{e}}|\mathbf{X}] = (1 - h_{ii})^{-1}\sigma^2$  varies with  $\mathbf{X}$
- To make it constant, we can scale the residuals by  $(1 - h_{ii})^{-1/2}$

$$\bar{e}_i = (1 - h_{ii})^{-1/2} \hat{e}_i$$

- $\bar{\mathbf{e}} = \mathbf{M}^{*1/2} \mathbf{M} \mathbf{e}$
- $\text{var}[\bar{e}_i|\mathbf{X}] = \mathbb{E}[\bar{e}_i^2|\mathbf{X}] = \sigma^2$
- If the error is homoskedastic,  $\bar{e}_i$  has the same bias and variance as the original errors.
- If the error is heteroskedastic, these standardized residuals are not.
- In R, these are recovered by “`rstandard(lmmod)`”: sometimes compared to  $-2$ ,  $2$ .

## Estimating $\hat{\sigma}^2$ , $s^2$ , $\bar{\sigma}^2$

- The error variance  $\sigma^2$  measures the unexplained part of the regression.
- Three estimators:
  - 1 The method of moments estimator is  $\hat{\sigma}^2 = \frac{1}{n} \sum \hat{e}_i^2$  (Biased)
  - 2 The bias-corrected estimator is  $s^2 = \frac{1}{n-k} \sum \hat{e}_i^2$  (sigma in R.)
  - 3 The standardized estimator  $\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \bar{e}_i^2 = \frac{1}{n} \sum_{i=1}^n (1 - h_{ii})^{-1} \hat{e}_i^2$

## Estimating $\hat{\sigma}^2$

- Estimator (1)  $\hat{\sigma}^2$  takes the average of the squared residuals:  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}^2$
- The expectation of the estimator uses the following fact:

$$\hat{\sigma}^2 = \frac{1}{n} \mathbf{e}' \mathbf{M} \mathbf{e} = \frac{1}{n} \text{tr}(\mathbf{e}' \mathbf{M} \mathbf{e}) = \text{tr}(\mathbf{M} \mathbf{e} \mathbf{e}')$$

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2 | \mathbf{X}] &= \frac{1}{n} \text{tr}(\mathbb{E}[\mathbf{M} \mathbf{e} \mathbf{e}' | \mathbf{X}]) \\ &= \frac{1}{n} \text{tr}(\mathbf{M} \mathbb{E}[\mathbf{e} \mathbf{e}' | \mathbf{X}]) \\ &= \frac{1}{n} \text{tr}(\mathbf{M} \mathbf{D}) \\ &= \frac{1}{n} \sum_{i=1}^n (1 - h_{ii}) \sigma_i^2 \end{aligned}$$

# Estimating $\hat{\sigma}^2$

- Under conditional homoskedasticity

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2|\mathbf{X}] &= \frac{1}{n} \text{tr}(\mathbb{E}[\mathbf{M}\mathbf{e}\mathbf{e}'|\mathbf{X}]) \\ &= \frac{1}{n} \text{tr}(\mathbf{M}\mathbb{E}[\mathbf{e}\mathbf{e}'|\mathbf{X}]) \\ &= \frac{1}{n} \text{tr}(\mathbf{M}\mathbf{I}_n\sigma^2) \\ &= \sigma^2 \frac{n-k}{n}\end{aligned}$$

## What should I report?

- The bias-corrected error variance estimator  $s^2$  is used throughout applied work.
- It is used to calculate standard errors, F-tests, t-test, and confidence intervals.
- It is typically reported as the  $\text{RMSE} = \sqrt{s^2}$
- However,  $s^2$  assumes homoskedasticity:
  - You will use robust sandwich estimators like HC2 and HC3 (more on that later),
  - you can report  $\bar{\sigma}^2$ , which is unbiased under heteroskedasticity.