Properties of OLS Estimators and the Gauss-Markov Theorem

Why Should We Trust These Estimates?

Jake Anderson

Econ 103, Lecture 4

Outline

1. The Problem: Sampling Variation
2. Preparing the Proof: b_2 as a Weighted Sum
3. Unbiasedness: $\mathrm{E}(b_2) = \beta_2$
4. Variance of b_2
5. The Gauss-Markov Theorem
6. Summary

Where We Stand

Last time, we estimated the food expenditure model ($N = 40$ Australian households):

$$\hat{y}_i = 83.42 + 10.21\, x_i$$

We got $b_2 = 10.21$: an extra \$100 in weekly income $\implies$ \$10.21 more food spending.

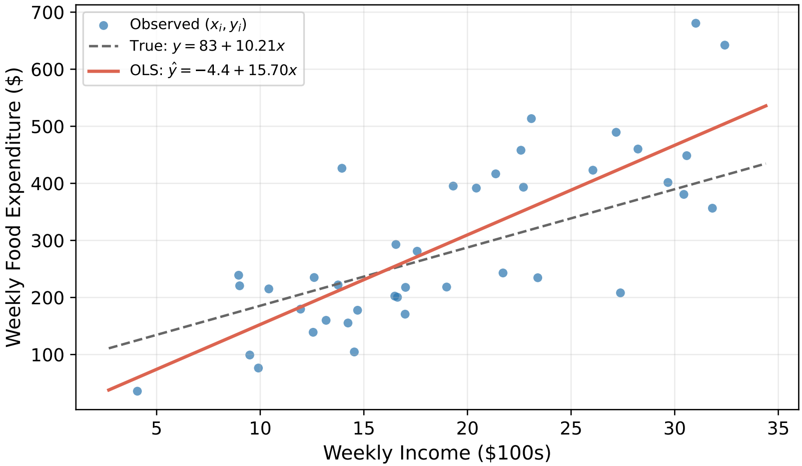
But here is the question:

If you went back to Australia and surveyed a *different* group of 40 households, would you get 10.21 again?

Almost certainly not. The estimates depend on the specific households in the sample.

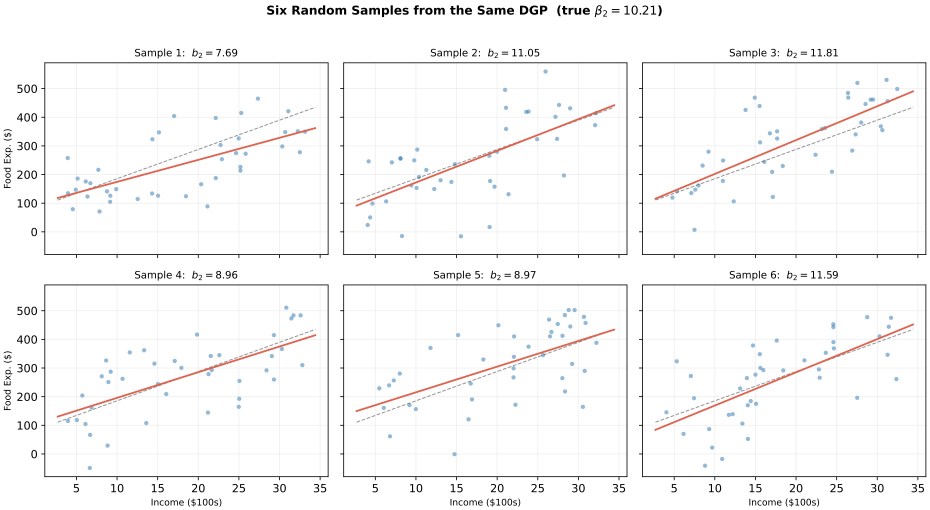
$\implies$ b_2 is a **random variable**. It has a distribution, a mean, and a variance.

One Sample, One Estimate



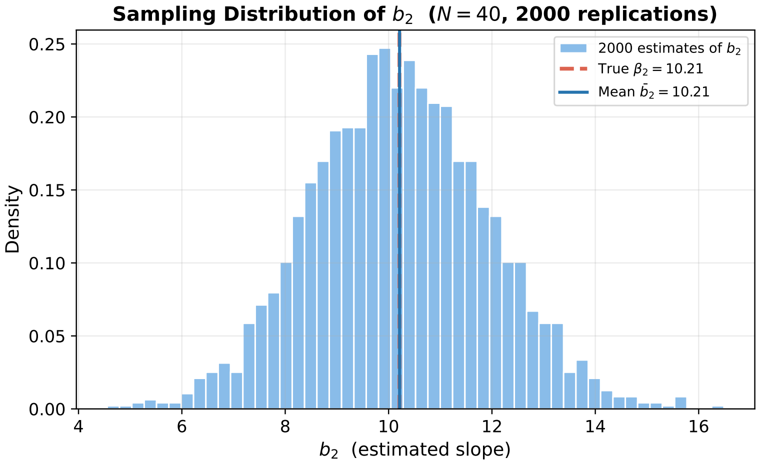
The gray dashed line is the truth. The red line is our estimate from one sample. They are close but not the same.

Different Samples, Different Estimates



Suppose the true slope is $\beta_2 = 10.21$. Same DGP, same β_2, but each sample gives a different b_2. This

The Sampling Distribution



Draw 2,000 samples of $N = 40$, compute b_2 each time, and plot the histogram.

Two observations:

Rewriting b_2 as a Weighted Sum of the y_i's

To prove b_2 is unbiased, we first need to rewrite it in a form where we can take expectations. Start from the formula you already know:

$$b_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Define weights that depend only on the x-values:

$$w_i = \frac{x_i - \bar{x}}{\sum_{j=1}^{N}(x_j - \bar{x})^2}$$

Then (shown in the textbook appendix):

$$\boxed{b_2 = \sum_{i=1}^{N} w_i \, y_i}$$

b_2 is a **weighted sum** of the y_i values. The weights w_i are constants (they depend only on the x-values, which are treated as fixed).

$\Longrightarrow$ b_2 is a **linear estimator**: it is a linear function of the data $y_1, y_2, \ldots, y_N$.

Properties of the Weights

The weights $w_i = \dfrac{x_i - \bar{x}}{\sum(x_j - \bar{x})^2}$ have two useful properties:

Property 1: $\displaystyle\sum_{i=1}^{N} w_i = 0$

Because $\sum(x_i - \bar{x}) = 0$ in the numerator.

Property 2: $\displaystyle\sum_{i=1}^{N} w_i x_i = 1$

Because $\sum w_i x_i = \sum \frac{(x_i - \bar{x})x_i}{\sum(x_j - \bar{x})^2} = \frac{\sum(x_i - \bar{x})x_i}{\sum(x_j - \bar{x})^2} = 1$.

We will use both of these in the unbiasedness proof.

Decomposing b_2: Signal + Noise

Substitute the model $y_i = \beta_1 + \beta_2 x_i + e_i$ into $b_2 = \sum w_i y_i$:

$$
\begin{aligned}
b_2 &= \sum w_i(\beta_1 + \beta_2 x_i + e_i) \\
&= \beta_1 \underbrace{\sum w_i}_{=0} + \beta_2 \underbrace{\sum w_i x_i}_{=1} + \sum w_i e_i \\
&= \beta_2 + \sum_{i=1}^{N} w_i e_i
\end{aligned}
$$

$$\boxed{b_2 = \beta_2 + \sum_{i=1}^{N} w_i e_i}$$

$\implies$ The estimator equals the **true parameter** plus a **weighted sum of random errors**. All the randomness in b_2 comes from the e_i's.

What Does Unbiasedness Mean?

An estimator $\hat{\theta}$ is **unbiased** if:

$$\mathsf{E}(\hat{\theta}) = \theta$$

In words: if we could repeat the experiment infinitely many times and average all the estimates, the average would equal the true parameter.

What unbiasedness does not mean:

- It does *not* mean any single estimate is close to β_2
- It does *not* mean $b_2 = \beta_2$ in your sample
- It is a property of the *procedure*, not of any one number

$\Longrightarrow$ The simulation histogram was centered on $\beta_2 = 10.21$. That is exactly what unbiasedness predicts.

Proof: $\mathrm{E}(b_2) = \beta_2$ Under SR1–SR2

We showed that $b_2 = \beta_2 + \sum w_i e_i$. Now take the expected value:

$$\mathrm{E}(b_2) = \mathrm{E}\left(\beta_2 + \sum_{i=1}^{N} w_i e_i\right)$$

$$= \beta_2 + \sum_{i=1}^{N} w_i \; \mathrm{E}(e_i)$$

By **SR2**: $\mathrm{E}(e_i) = 0$ for every i.

$$\mathrm{E}(b_2) = \beta_2 + \sum_{i=1}^{N} w_i \cdot 0$$

$$= \beta_2 \qquad \blacksquare$$

$\implies$ OLS is **unbiased** for β_2. On average, across all possible samples, b_2 hits the true slope. The

When Does Unbiasedness Fail?

The proof used $\mathrm{E}(e_i|x_i) = 0$ (assumption SR2). What if this fails?

Suppose we omit an important variable that is correlated with x. Then $\mathrm{E}(e_i|x_i) \neq 0$, and:

$$\mathrm{E}(b_2|x) = \beta_2 + \sum w_i \underbrace{\mathrm{E}(e_i|x)}_{\neq 0} \neq \beta_2$$

This is **omitted variable bias**. The OLS estimator systematically overshoots or undershoots β_2.

Example: regressing wages on education alone. The error contains ability, which is correlated with education. So $\mathrm{E}(e_i|x_i) \neq 0$ and b_2 is biased upward.

$\implies$ Unbiasedness is not automatic. It depends on the assumptions being correct.

Why Variance Tells Us About Precision

We have proven the centering: $\mathrm{E}(b_2) = \beta_2$. Now let's explain the spread we saw in the sampling distribution histogram.

- Small variance $\implies$ estimates cluster tightly around β_2 $\implies$ precise
- Large variance $\implies$ estimates spread widely $\implies$ imprecise

Two researchers can both have unbiased estimators, but the one with smaller variance will get more reliable results.

$\implies$ We want to know: what determines $\mathrm{Var}(b_2)$?

Deriving Var(b_2)

From $b_2 = \beta_2 + \sum w_i e_i$:

$$\text{Var}(b_2) = \text{Var}\left(\sum_{i=1}^{N} w_i e_i\right)$$

$$= \sum_{i=1}^{N} w_i^2 \ \text{Var}(e_i) + \underbrace{\sum_{i \neq j} w_i w_j \ \text{Cov}(e_i, e_j)}_{\text{SR4: } = 0}$$

By **SR3** (homoskedasticity): $\text{Var}(e_i) = \sigma^2$ for all i.

Substitute $w_i = \dfrac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2}$:

$$\text{Var}(b_2) = \sigma^2 \sum_{i=1}^{N} w_i^2 = \sigma^2 \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{\left[\sum (x_j - \bar{x})^2\right]^2}$$

Three Factors That Control Precision

$$\text{Var}(b_2) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Factor 1: Error variance σ^2 (numerator)

- More noise in the DGP $\implies$ harder to detect the signal $\implies$ larger $\text{Var}(b_2)$
- You cannot control σ^2 (it is a feature of the population)

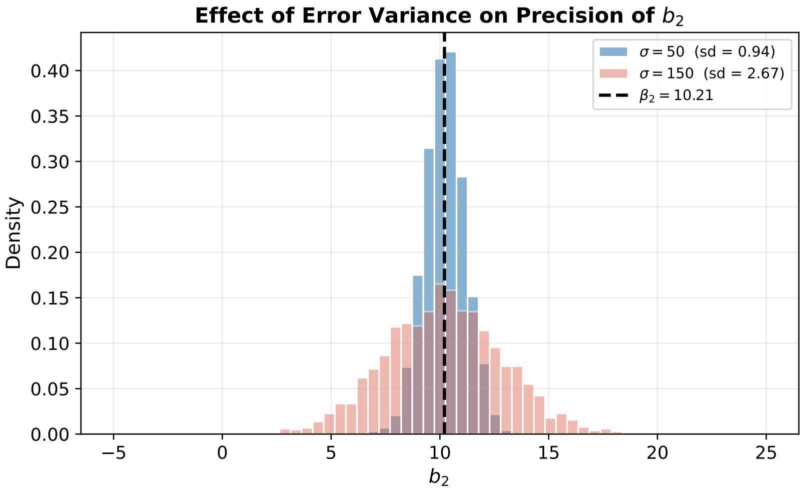
Factor 2: Spread of x**-values** (denominator)

- $\sum(x_i - \bar{x})^2$ measures total variation in the explanatory variable
- More spread in x $\implies$ larger denominator $\implies$ smaller $\text{Var}(b_2)$
- Intuition: a line is easier to estimate when the x-values span a wide range

Factor 3: Sample size N

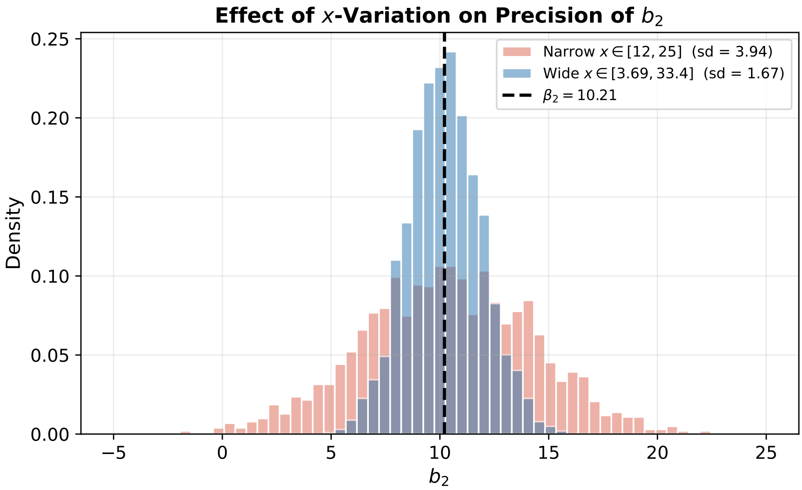
- More observations $\implies$ more terms in $\sum(x_i - \bar{x})^2$ $\implies$ denominator grows
- $\implies$ more data reduces $\text{Var}(b_2)$

Visualizing Factor 1: Error Variance σ^2



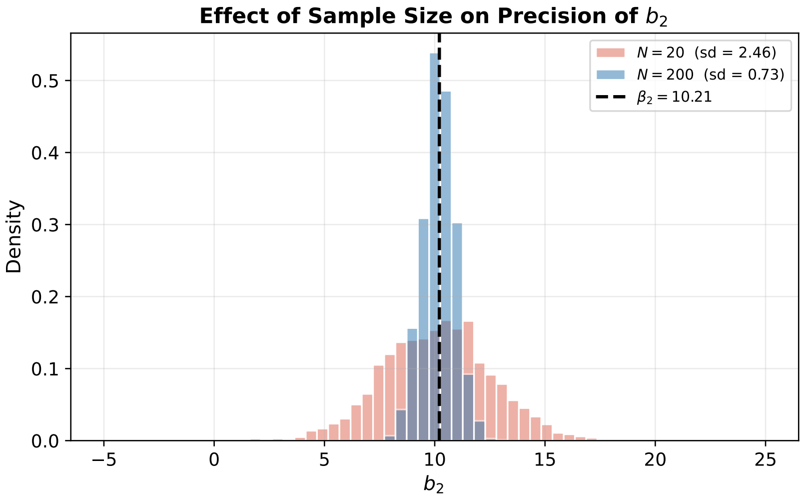
Same N, same x-range. Tripling σ triples the standard deviation of b_2.

Visualizing Factor 2: Spread of x



Same N, same σ. Narrowing the x-range concentrates x near its mean, making $\sum(x_i - \bar{x})^2$ smaller

Visualizing Factor 3: Sample Size N



Same σ, same x-range. Going from $N = 20$ to $N = 200$ shrinks the spread dramatically.

What Can You Control?

$$\mathrm{Var}(b_2) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Factor	**Effect on** $\mathrm{Var}(b_2)$	**Researcher controls?**
Larger σ^2	Increases	No
Wider x-spread	Decreases	Sometimes (experimental design)
Larger N	Decreases	Yes (collect more data)

$\Longrightarrow$ The most reliable way to improve precision is to collect more data. If you can design the study, spreading the x-values over a wide range also helps.

Setting Up the Question

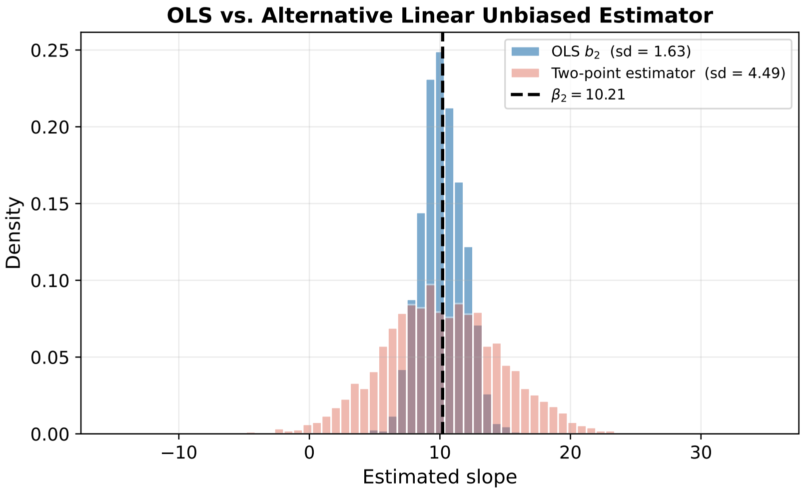
We have shown:

- b_2 is a **linear** estimator (weighted sum of y_i's)
- b_2 is **unbiased**: $\mathrm{E}(b_2) = \beta_2$
- b_2 has variance $\mathrm{Var}(b_2) = \sigma^2 / \sum(x_i - \bar{x})^2$

But OLS is not the only linear unbiased estimator. For example, you could use just the two observations with the smallest and largest x-values to draw a line through the data.

Question: Among all possible linear unbiased estimators of β_2, is there one with a *smaller* variance than OLS?

A Concrete Competitor: The Two-Point Estimator



The two-point estimator uses only the observations with the smallest and largest x-values to

The Gauss-Markov Theorem

Gauss-Markov Theorem

Under assumptions SR1–SR4, the OLS estimators b_1 and b_2 have the **smallest variance** of all linear and unbiased estimators of β_1 and β_2.

Normality (SR5) is not needed for Gauss-Markov. We will use it later for inference.

OLS is the **B**est **L**inear **U**nbiased **E**stimator: **BLUE**.

- **Best** = smallest variance (most precise)
- **Linear** = weighted sum of y_i's
- **Unbiased** = $\mathrm{E}(b_2) = \beta_2$
- **Estimator** = a rule applied to data (random variable, not a number)

$\Longrightarrow$ You cannot do better than OLS without either (a) giving up linearity, (b) accepting bias, or (c) violating one of SR1–SR4.

Proof Sketch: Setup

Consider any other linear unbiased estimator $\tilde{b}_2 = \sum c_i y_i$ where the c_i are constants (possibly different from the OLS weights w_i).

Write $c_i = w_i + d_i$, where d_i is the "extra" part beyond OLS.

For $\tilde{b}_2$ to be unbiased, the d_i's must satisfy the same constraints as the w_i's:

$$\sum d_i = 0 \qquad \text{and} \qquad \sum d_i x_i = 0$$

$\implies$ If $d_i = 0$ for all i, we are back to OLS. Any nonzero d_i is a *departure* from OLS.

Proof Sketch: The Algebra

The variance of the alternative estimator is:

$$\mathrm{Var}(\tilde{b}_2) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (w_i + d_i)^2$$

Expanding the square:

$$= \sigma^2 \left(\sum w_i^2 + 2 \sum w_i d_i + \sum d_i^2 \right)$$

The cross term $\sum w_i d_i = 0$ (using the constraints on d_i). So:

$$\mathrm{Var}(\tilde{b}_2) = \underbrace{\sigma^2 \sum w_i^2}_{\mathrm{Var}(b_2)} + \underbrace{\sigma^2 \sum d_i^2}_{\geq 0} \geq \mathrm{Var}(b_2) \qquad \blacksquare$$

$\implies$ Any departure from OLS weights adds variance. OLS is the minimum.

What Gauss-Markov Does and Does Not Say

What it says:

- Among *linear and unbiased* estimators, OLS has the smallest variance
- This holds under SR1–SR4
- It applies to the *estimators* (procedures), not to any particular *estimate* (number)

What it does not say:

- OLS is the best of *all possible* estimators
  - A biased estimator with lower variance might have a smaller mean squared error
  - A nonlinear estimator could outperform OLS
- OLS works well when the assumptions are violated
  - Heteroskedasticity (SR3 fails) $\Longrightarrow$ OLS is not BLUE; GLS (a different estimator we will cover later) is better
  - Serial correlation (errors in consecutive observations are correlated; common in time-series data; SR4 fails) $\Longrightarrow$ OLS is not BLUE

$\Longrightarrow$ Gauss-Markov gives you a strong reason to use OLS *when the assumptions hold*.

Putting It All Together

Property	Result	Requires
Linearity	$b_2 = \sum w_i y_i$	Definition of OLS
Decomposition	$b_2 = \beta_2 + \sum w_i e_i$	SR1
Unbiasedness	$\mathrm{E}(b_2) = \beta_2$	SR1, SR2
Variance formula	$\mathrm{Var}(b_2) = \dfrac{\sigma^2}{\sum(x_i - \bar{x})^2}$	SR1–SR4
BLUE	Smallest variance among linear unbiased estimators	SR1–SR4

Next time: how to *use* these results. We will estimate σ^2, compute standard errors, and build confidence intervals for β_2.

Thank you!

jakeanderson@g.ucla.edu