Interpreting Multiple Regression & the MR Assumptions

What "Holding All Else Constant" Actually Means

Jake Anderson

March 21, 2026

Outline

Outline

A Wage Equation

Suppose we estimate the following model with a large cross-section:

$$\widehat{\log(\text{wage})} = 0.50 + 0.08\,\text{educ} + 0.03\,\text{exper} + 0.25\,\text{female}$$

What does the coefficient 0.08 on education mean?

"Each additional year of education is associated with an 8% increase in wages, **holding experience and gender constant**."

$\Longrightarrow$ This is the ceteris paribus interpretation. It only works because experience and gender are *in the model*.

Coefficients as Partial Effects

In the general model:

$$E(y \mid x_2, x_3, \ldots, x_K) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K$$

Each slope coefficient is a **partial derivative**:

$$\beta_k = \frac{\partial\, E(y \mid x_2, \ldots, x_K)}{\partial\, x_k}$$

The change in the expected value of y when x_k increases by one unit, **holding all other explanatory variables fixed**.

Compare to simple regression, where $\beta_2 = dE(y)/dx$. The partial derivative notation is the formal version of "ceteris paribus."

Why “Holding Constant” Requires the Variable in the Model

Imagine we *omit* experience and estimate a shorter model:

$$\widehat{\log(\text{wage})} = 0.84 + 0.11\,\text{educ} + 0.20\,\text{female}$$

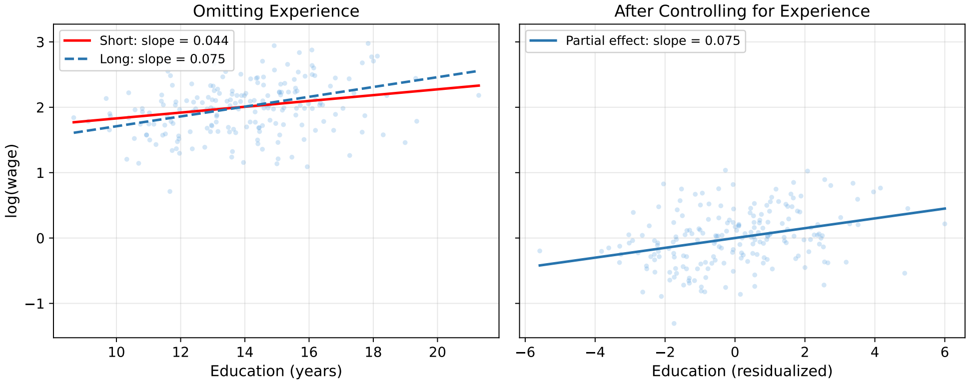
The education coefficient jumped from 0.08 to 0.11. Why?

Education and experience are **correlated** in the population:

- More-educated workers tend to have less experience (they spent more time in school)
- Experience raises wages, so omitting it creates *omitted variable bias*

$\Longrightarrow$ Without experience in the model, you *cannot* interpret $\hat{\beta}_{\text{educ}}$ as “holding experience constant.” The coefficient absorbs part of the experience effect.

The Omission in Pictures



Left: Omitting experience inflates the education coefficient.
Right: Including experience isolates each variable's partial effect.

Example: Big Andy's Burger Barn

From the textbook (Example 5.2), using data from 75 franchise locations:

$$\widehat{\text{SALES}} = 118.91 - 7.91\,\text{PRICE} + 1.86\,\text{ADVERT}$$

Interpretations:

- $b_2 = -7.91$: a \$1 price increase is associated with a \$7,910 *decrease* in monthly sales revenue, **holding advertising constant**
- $b_3 = 1.86$: a \$1,000 increase in advertising is associated with a \$1,860 *increase* in monthly sales revenue, **holding price constant**

$\Longrightarrow$ Each coefficient tells you the effect of *one* variable, as if you could change it in isolation.

The Intercept

In the Burger Barn model: $b_1 = 118.91$.

Literal interpretation: predicted sales when PRICE $= 0$ and ADVERT $= 0$.

Is that meaningful? **No.** Free burgers with zero advertising never happens in the data.

The intercept anchors the regression plane. It is needed for unbiased residuals ($\sum \hat{e}_i = 0$) and correct predictions *within the data range*.

$\implies$ Always include an intercept unless you have a specific economic reason not to.

Outline

Frisch–Waugh–Lovell: The Mechanics of "Holding Constant"

How does OLS actually hold other variables constant? The FWL theorem gives a precise answer.

Goal: Find b_3 (the ADVERT coefficient) in:

$$\text{SALES}_i = \beta_1 + \beta_2\,\text{PRICE}_i + \beta_3\,\text{ADVERT}_i + e_i$$

FWL recipe:

1. Regress SALES on PRICE (and a constant). Save residuals $\widetilde{\text{SALES}}_i$.
2. Regress ADVERT on PRICE (and a constant). Save residuals $\widetilde{\text{ADVERT}}_i$.
3. Regress $\widetilde{\text{SALES}}_i$ on $\widetilde{\text{ADVERT}}_i$ (no constant).

Result: The slope from step 3 is *exactly* b_3 from the full regression.

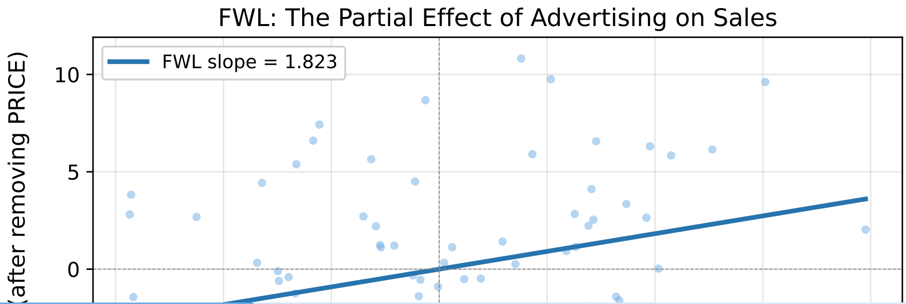
FWL: What the Residuals Represent

$\widetilde{\text{SALES}}_i$ = the part of SALES that **cannot be predicted by PRICE**

$\widetilde{\text{ADVERT}}_i$ = the part of ADVERT that **cannot be predicted by PRICE**

$\Longrightarrow$ Step 3 asks: once we strip out everything that PRICE explains about both variables, does the leftover variation in ADVERT still predict the leftover variation in SALES?

That is what "holding PRICE constant" means computationally. You remove its influence from both sides, then look at what remains.



FWL: A Caution

The FWL regression gives the **correct coefficient** but the **wrong standard errors**.

- The partialled-out regression does not account for parameters estimated in Steps 1 and 2, so its standard errors are too small
- The full regression correctly uses $N - K$ degrees of freedom
- For inference, always use the full model

$\Longrightarrow$ FWL is a conceptual tool for understanding what MR does. It is not a shortcut for estimation.

Outline

From SR to MR: What Changes?

You already know SR1–SR6 for simple regression. The MR assumptions are nearly identical, with one important addition.

#	SR Version	MR Version	Change?
1	Linear model, 2 params	Linear model, K params	Generalized
2	$\text{E}(e_i) = 0$	$\text{E}(e_i \mid x_i) = 0$	Same
3	$\text{Var}(e_i) = \sigma^2$	$\text{Var}(e_i) = \sigma^2$	Same
4	$\text{Cov}(e_i, e_j) = 0$	$\text{Cov}(e_i, e_j) = 0$	Same
5	x_i not random	x_{ik} not random + **no perf. collin.**	**New!**
6	$e_i \sim N(0, \sigma^2)$	$e_i \sim N(0, \sigma^2)$	Same

$\Longrightarrow$ Five assumptions carry over directly. The only genuinely new requirement is the "no perfect multicollinearity" condition in MR5.

MR1: The Model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_K x_{iK} + e_i, \quad i = 1, \ldots, N$$

The data are generated by a **linear function** of K parameters plus a random error.

This is the same as SR1, extended to K explanatory variables. “Linear in parameters” means the β’s enter linearly. The x’s can be logs, squares, interactions, etc.

What this rules out: models like $y = \beta_1 x^{\beta_2} e$ (nonlinear in parameters, unless you take logs).

MR2: Strict Exogeneity

$$E(e_i \mid X) = 0$$

Knowing the explanatory variables does not help predict the error. This gives us the regression function:

$$E(y_i \mid X) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

MR2 implies:

- $E(e_i) = 0$ (errors average to zero)
- $\text{Cov}(e_i, x_{jk}) = 0$ for all i, j, k (no correlation between errors and any regressor)

$\implies$ This is the assumption that makes the ceteris paribus interpretation valid. If an omitted variable is correlated with a regressor, MR2 fails and coefficients are biased.

MR3: Homoskedasticity and MR4: No Serial Correlation

MR3:

$$\text{Var}(e_i \mid X) = \sigma^2 \quad \text{for all } i$$

The spread of the errors is the same regardless of the x-values. Identical to SR3.

MR4:

$$\text{Cov}(e_i, e_j \mid X) = 0 \quad \text{for } i \neq j$$

One observation's error tells you nothing about another's. Identical to SR4.

When these fail:

- OLS is still **unbiased** (that only needs MR1–MR2)
- But OLS is no longer **efficient**, and standard errors are wrong
- We address MR3 violations in Chapter 8, MR4 violations in Chapter 9

MR5: No Perfect Multicollinearity (The New Assumption)

No explanatory variable is an exact linear function of the others.

Formally:

$$c_1 x_{i1} + c_2 x_{i2} + \cdots + c_K x_{iK} = 0 \text{ for all } i \implies c_1 = c_2 = \cdots = c_K = 0$$

This also requires that each x_k has some variation across observations (not all the same value).

Why is this new? In simple regression, there is only one x, so there is nothing for it to be collinear *with*. With multiple regressors, exact linear relationships can arise.

What happens if MR5 fails? OLS *cannot be computed*. There is no unique solution to the normal equations.

Examples of Perfect Multicollinearity

Example 1: Proportional variables

Model: $\text{wage}_i = \beta_1 + \beta_2\,\text{educ}_i + \beta_3\,\text{educ\_months}_i + e_i$

But $\text{educ\_months} = 12 \times \text{educ}$. These carry the same information.

Example 2: Exhaustive dummies

Model: $y_i = \beta_1 + \beta_2\,\text{male}_i + \beta_3\,\text{female}_i + e_i$

But $\text{male} + \text{female} = 1 = x_{i1}$ (the constant). Three variables, one exact linear relationship.

This is the **dummy variable trap**: include $J - 1$ dummies for J categories, not J.

Example 3: Budget shares

If $x_2 + x_3 + x_4 = 1$ for all observations (shares of income spent on food, housing, and other), then including all three plus an intercept creates an exact linear dependency.

MR6: Normality of Errors (Optional)

$$e_i \mid X \sim N(0, \sigma^2)$$

Identical to SR6. Needed for **exact** t- and F-distributions in finite samples.

Without MR6, the Central Limit Theorem still gives us approximate normality in large samples, so t-tests and confidence intervals remain valid asymptotically.

When does MR6 really help?

- Small samples where CLT approximation is poor
- Deriving the exact distribution of t-statistics as $t_{(N-K)}$

$\Longrightarrow$ MR6 is a convenience, not a necessity.

The Gauss–Markov Theorem (Revisited)

Under MR1–MR5, the OLS estimators are **BLUE**:

B	Best: smallest variance
L	Linear: linear functions of the y_i
U	Unbiased: $\mathrm{E}(b_k) = \beta_k$
E	Estimators

This is the same result you saw in simple regression. It extends directly to the multiple regression case.

With MR6 added:

- $b_k \sim N(\beta_k, \mathrm{Var}(b_k))$
- $t = (b_k - \beta_k)/\,\mathrm{se}(b_k) \sim t_{(N-K)}$
- Degrees of freedom: $N - K$ (not $N - 2$ as in simple regression)

Outline

Two Kinds of Multicollinearity

Perfect multicollinearity: One regressor is an *exact* linear function of others.

- Violates MR5
- OLS cannot be computed (software throws an error)
- Always fixable by dropping the redundant variable

Near multicollinearity: Regressors are *highly* correlated but not perfectly.

- Does *not* violate MR5
- OLS is still BLUE
- But estimator variances become very large

$\Longrightarrow$ Perfect collinearity is a showstopper. Near collinearity is a practical nuisance.

Near Collinearity and Estimator Variance

Recall the variance of b_2 in the $K = 3$ model:

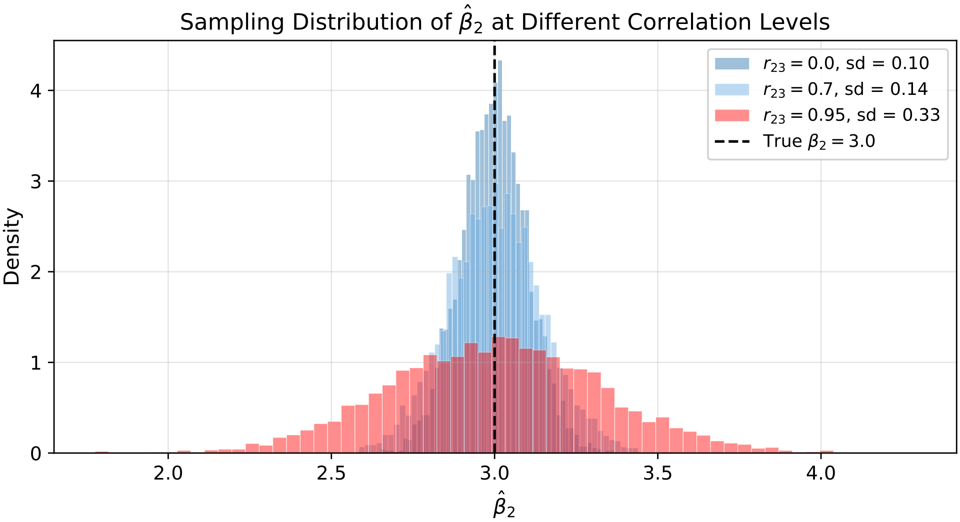
$$\mathrm{Var}(b_2 \mid \mathrm{X}) = \frac{\sigma^2}{\underbrace{(1 - r_{23}^2)}_{\text{collinearity factor}} \sum_{i=1}^{N} (x_{i2} - \bar{x}_2)^2}$$

where r_{23} is the sample correlation between x_2 and x_3.

$\lvert r_{23} \rvert$	$1 - r_{23}^2$	**Effect on** $\mathrm{Var}(b_2)$
0	1.00	Baseline
0.5	0.75	$1.3\times$ larger
0.9	0.19	$5.3\times$ larger
0.99	0.02	$50\times$ larger
1	0	$\to \infty$ (cannot compute)

$\Longrightarrow$ As $\lvert r_{23} \rvert \to 1$, the denominator vanishes and $\mathrm{Var}(b_2)$ explodes.

Visualizing Near Collinearity



Higher correlation between regressors $\rightarrow$ wider sampling distribution for $\hat{\beta}_2$. The estimate

Living with Near Collinearity

When regressors are highly correlated:

- **Large standard errors:** coefficients are imprecisely estimated
- **Wide confidence intervals:** hard to draw conclusions about individual effects
- **Insignificant t-tests:** coefficients may appear "insignificant" even though the variables genuinely affect y
- **Sensitive estimates:** adding or removing one observation can change $\hat{\beta}$ substantially

What near collinearity does *not* do:

- It does not bias OLS (still unbiased, still BLUE)
- It does not violate any assumption

$\Longrightarrow$ The problem is not with the estimator. The problem is that the data do not contain enough independent variation to separate the effects of correlated variables. We cover diagnosis and remedies in Chapter 6.

Outline

What Each Assumption Buys You

Assumptions	Property	Result
MR1–MR2	Unbiasedness	$E(b_k) = \beta_k$
MR1–MR5	BLUE	OLS has smallest variance
MR1–MR5 + MR6	Exact inference	$t = (b_k - \beta_k)/\,\text{se}(b_k) \sim t_{(N-K)}$

Each layer of assumptions gives you something stronger:

- MR1–MR2 alone: OLS is unbiased, but you have no guarantee about efficiency
- Add MR3–MR5: OLS is the best you can do among linear unbiased estimators
- Add MR6: you can construct exact t-tests and confidence intervals with $N - K$ degrees of freedom

Degrees of Freedom: $N - K$

In simple regression: df $= N - 2$ (estimate two parameters: β_1, β_2).

In multiple regression: df $= N - K$ (estimate K parameters: $\beta_1, \ldots, \beta_K$).

This affects:

- The unbiased error variance estimator: $\hat{\sigma}^2 = SSE/(N - K)$
- The t-distribution for hypothesis tests: $t_{(N-K)}$
- Critical values for confidence intervals

Intuition: Each estimated parameter "uses up" one degree of freedom. With K parameters estimated from N observations, you have $N - K$ pieces of independent information left to estimate σ^2.

Back to the Wage Equation

$$\widehat{\log(\text{wage})} = 0.50 + 0.08\,\text{educ} + 0.03\,\text{exper} + 0.25\,\text{female}$$

For this model to deliver unbiased, ceteris paribus estimates, we need:

- **MR1:** The true relationship is linear in these parameters ✓
- **MR2:** No omitted variables correlated with educ, exper, or female
- **MR3:** Wage variability is the same at all education/experience levels
- **MR4:** One person's wage tells you nothing about another's (cross-section, plausible)
- **MR5:** educ, exper, female are not perfectly collinear (they are not)
- **MR6:** Errors are normally distributed (for exact t-tests with $N - 4$ df)

$\Longrightarrow$ The hardest assumption to defend is usually MR2. Are there omitted variables that correlate with the regressors?

Takeaways

1. Every MR coefficient is a **partial effect**: the change in $E(y)$ when one variable changes, holding all others fixed
2. FWL shows what “holding constant” does mechanically: strip out the other variables’ influence, then regress the residuals
3. Assumptions MR1–MR6 carry over from SLR almost unchanged; the genuinely new requirement is **MR5** (no perfect multicollinearity)
4. Near collinearity does not bias OLS, but it inflates standard errors and makes individual coefficients hard to pin down

Thank you!

jakeanderson@g.ucla.edu