

450A Political Methodology I

Lecture 5: Regression with Two Regressors

2. Omitted Variables and Multicollinearity

Stanford University

Ordinary Least Squares (OLS)

- So far we have learned about:
 - Intuition behind multiple regression
 - Derivation of the OLS estimator
 - Goodness of fit
- Next we learn:
 - Assumptions and properties of multiple OLS
 - Consequences of violating the assumptions
- We will do this for the case of two explanatory variables
- For the more general case of k explanatory variables, we will use matrices

Multiple OLS Assumptions

Omitted Variables Bias

Multicollinearity

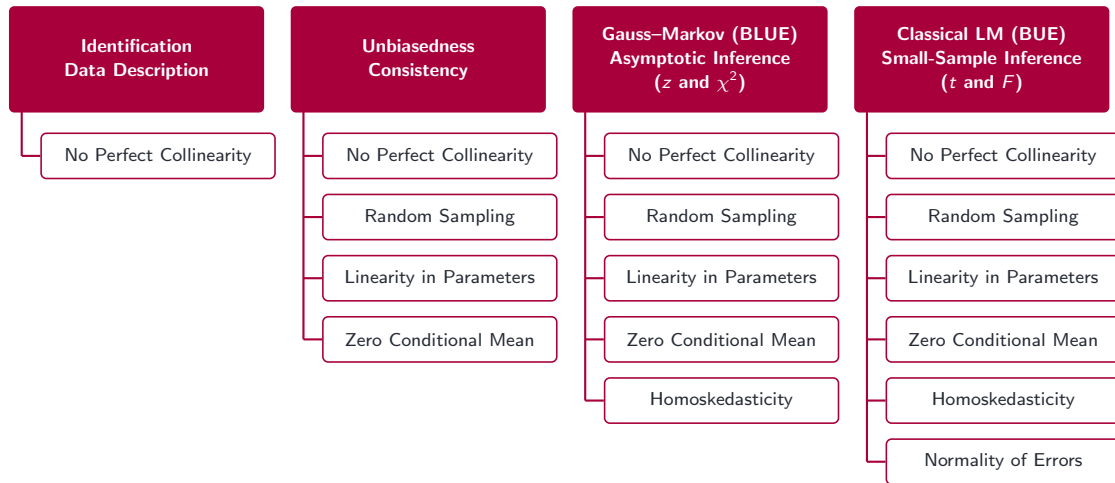
Including an Irrelevant Variable

Multiple OLS Assumptions

We make essentially the same set of assumptions as in the simple linear regression case:

- ① **Linearity in Parameters:** The population model is linear in its parameters and correctly specified
- ② **Random Sampling:** The observed data represent a random sample from the population described by the model.
- ③ **No perfect collinearity:** There is variation in **all explanatory variables** and **no predictor is a linear combination of other predictors**
- ④ **Zero conditional mean:** Expected value of the error term is zero conditional on all values of the explanatory variables
- ⑤ **Homoskedasticity:** The error term has the same variance conditional on all values of the explanatory variables.
- ⑥ **Normality:** The error term is independent of the explanatory variables and normally distributed.

Hierarchy of OLS Assumptions



Assumption (I. Linearity)

The population regression model is linear in its parameters (it can be non-linear in the variables) and correctly specified. In the population, Y is related to $X_1, X_2, \dots, X_k$ and the unobserved error term u as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$$

where $\beta_0, \beta_1, \dots, \beta_k$ represent the fixed and unknown regression parameters.

The random variable u (error term) captures all other factors which influence Y other than the regressors $X_1, X_2, \dots, X_k$, so its meaning changes once we add more explanatory variables.

Examples consistent with this assumption:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 \log(X_2) + \beta_4 (X_1 \cdot \log(X_2)) + u$

Violations:

- $Y = \beta_0 + \beta_1^2 X_1 + \exp(\beta_2) X_2 + u$
- Functional form misspecification (e.g. population model is not linear)

Assumption (II. Random Sampling)

The observed data:

$$(y_i, x_{1,i}, x_{2,i}, \dots, x_{k,i}) \text{ for } i = 1, \dots, n \text{ and } j = 1, \dots, k$$

represent an i.i.d. random sample of size n following the population model. This implies stochastic regressors.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)

OLS Assumption III

OLS assumptions are essentially identical (with slight modification to the notation) to the simple regression case, except

Assumption III:

Assumption (III. No Perfect Collinearity)

In the sample, none of the independent variables is constant and there are no exact linear relationships among the independent variables.

Examples consistent with this assumption:

- $X_1 = Z, X_2 = Z^2$
- $X_1 = Z, X_2 = V, X_3 = Z \cdot V$

Violations:

- $X_1 = Z, X_2 = 2 \cdot Z$

Q. Is this a violation: $X_1 = \log(Z), X_2 = \log(Z^2)$?

A. Yes, because $\log(Z^2) = 2 \log(Z)$.

Assumption (IV. Zero Conditional Mean)

The error term has zero conditional mean, i.e. the expected value of the error term is zero conditional on any values of the explanatory variables:

$$E[u|X_1, X_2, \dots, X_k] = 0$$

This implies $E[u] = 0$ and $\text{Cov}[X_j, u] = 0$ for all j . The explanatory variables are uncorrelated with the error term.

- What does $E[u|X_1, X_2, \dots, X_k] = 0$ mean for this stylized example based on Cochran's (1968) data (Lecture 5-1):
Death Rate = $\beta_0 + \beta_1$ Cigarette Smoker + β_2 Age + u
- It requires that, at each age, the other causes of death in u average out the same for cigarette smokers and for cigar or pipe smokers, as if smoking type were randomly assigned within age groups.
- Fails if u still contains omitted variables that are correlated with an included X_j and that affect Y . Notice: the meaning of u changes when we add variables.
- Given random sampling, $E[u|X_1, X_2, \dots, X_k] = 0$ implies $E[u_i|x_{1,i}, x_{2,i}, \dots, x_{k,i}] = 0$ for all sampled i . Together these assumptions allow us to derive the statistical properties of the OLS estimator conditional on the sample values $x_{1,i}, x_{2,i}, \dots, x_{k,i}$

Multiple OLS Assumptions

Omitted Variables Bias

Multicollinearity

Including an Irrelevant Variable

Theorem (Unbiasedness of OLS)

Given OLS Assumptions I–IV:

$$E[\hat{\beta}_j] = \beta_j$$

for all $j = 0, 1, 2$. The sampling distributions of the OLS estimators $\hat{\beta}_j$ are centered about the true population parameter values β_j .

Proof.

Proof is similar to the simple regression case:

- 1 Start with the “partialling” out representation of the estimator $\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{1,i} y_i}{\sum_i^n \hat{r}_{1,i}^2}$ where $\hat{r}_1$ is residual from regression of X_1 on X_2 and a constant.
- 2 Plug in for y_i and rearrange to $\hat{\beta}_1 = \beta_1 + \frac{\sum_i^n \hat{r}_{1,i} u_i}{\sum_i^n \hat{r}_{1,i}^2}$
- 3 Take expectations $E[\hat{\beta}_1|X] = \beta_1 + \frac{\sum_i^n \hat{r}_{1,i} E[u_i|X]}{\sum_i^n \hat{r}_{1,i}^2} = \beta_1$.

And so forth. (See Wooldridge 3A.3. for a complete treatment)



Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Wants Less Immigration} = \beta_0 + \beta_1 \text{ Watch Fox News} + \beta_2 \text{ Strong Republican} + u$$

Underspecified Model that we use:

$$\widehat{\text{Wants Less Immigration}} = \tilde{\beta}_0 + \tilde{\beta}_1 \text{ Watch Fox News}$$

Q: Which statement is correct?

- ❶ $E[\tilde{\beta}_1] > \beta_1$
- ❷ $E[\tilde{\beta}_1] < \beta_1$
- ❸ $E[\tilde{\beta}_1] = \beta_1$
- ❹ Can't tell

A: strong Republicans watch more Fox News ($\tilde{\delta} > 0$) and want less immigration ($\beta_2 > 0$), so $E[\tilde{\beta}_1] > \beta_1$.

Source: Pew Research Center (2025): 57% of Republicans vs. 18% of Democrats regularly get news from Fox News.

Omitted Variable Bias: Simple Case

True Population Model:

$$\text{Survival} = \beta_0 + \beta_1 \text{ Hospitalized} + \beta_2 \text{ Prior Health} + u$$

Underspecified Model that we use:

$$\widehat{\text{Survival}} = \tilde{\beta}_0 + \tilde{\beta}_1 \text{ Hospitalized}$$

Q: Which statement is correct?

- ① $E[\tilde{\beta}_1] > \beta_1$
- ② $E[\tilde{\beta}_1] < \beta_1$
- ③ $E[\tilde{\beta}_1] = \beta_1$
- ④ Can't tell

Answer: people who are hospitalized tend to be in worse prior health (the slope $\tilde{\delta}$ of Prior Health on Hospitalized is negative), and at the same hospitalization status, people in better prior health are more likely to survive ($\beta_2 > 0$). So $\tilde{\beta}_1$ is biased downward: $E[\tilde{\beta}_1] < \beta_1$.

Source: Adapted from the “hospital allegory” in Angrist & Pischke (2009), *Mostly Harmless Econometrics*, Sec. 2.1, where the outcome is self-reported health.

Omitted Variable Bias: Simple Case

True Population Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Underspecified Model that we use:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

In any sample, if both regressions are fit by OLS to the same observations with an intercept, the short and long slopes are related exactly by:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$$

- $\tilde{\delta}$ is the slope coefficient of a regression of x_2 on x_1 . If $\tilde{\delta} > 0$ then $\text{Cor}(x_1, x_2) > 0$ and if $\tilde{\delta} < 0$ then $\text{Cor}(x_1, x_2) < 0$.
- $\hat{\beta}_2$ is from the long regression (y on x_1 and x_2) and measures the relationship between x_2 and y , holding x_1 fixed.

Q. When will $\tilde{\beta}_1 = \hat{\beta}_1$?

A. If $\tilde{\delta} = 0$ or $\hat{\beta}_2 = 0$.

Omitted Variable Bias: Executive Corruption and Court Independence

Lecture 5-1's example: executive corruption (0–100) on high court independence (0 = never, 4 = always), 168 countries in 2024. What happens when we leave out democracy?

Short regression	Corruption on Courts	$\tilde{\beta}_1 = -17.6637$
Long regression	Corruption on Courts and Democracy	$\hat{\beta}_1 = -10.6301, \hat{\beta}_2 = -21.5101$
Auxiliary regression	Democracy on Courts	$\tilde{\delta} = 0.326989$

Check the identity.

$$\hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta} = -10.6301 + (-21.5101)(0.326989) = -10.6301 + (-7.0336) = -17.6637 = \tilde{\beta}_1$$

- $\tilde{\delta} > 0$: democracies have more independent courts (an average of 2.9, against 1.3 in autocracies).
- $\hat{\beta}_2 < 0$: at the same court independence, democracies have 22 points less executive corruption.
- So $\hat{\beta}_2 \cdot \tilde{\delta} < 0$. Leaving democracy out makes court independence look more strongly tied to clean government. About 40% of the short slope (-7.0 of -17.7) runs through democracy.

Source: V-Dem v16, 2024. Democracy: V-Dem's electoral democracy index above 0.5.

Omitted Variable Bias: The College Gap in the 2016 Trump Vote

White voters in 2016: vote for Trump (1) or Clinton (0) on a college degree (1 = four-year degree). Linear probability model, survey-weighted OLS, $n = 1,648$.

Also in the model	College
nothing	-0.178
hostile sexism	-0.090
racism denial	-0.044
both	-0.031

Racism denial is a 0–1 scale from three items, e.g. agreeing that racial problems in the U.S. are “rare, isolated situations.” Leave it out, and $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \tilde{\delta}$ reads

$$-0.1778 = -0.0443 + (1.434)(-0.0931) = -0.0443 + (-0.1335)$$

- $\tilde{\delta} < 0$: graduates score lower on racism denial (0.33 vs. 0.42).
- $\hat{\beta}_2 > 0$: at the same education, racism denial goes with a Trump vote.
- So $\hat{\beta}_2 \cdot \tilde{\delta} < 0$. Leaving racism denial out makes the college gap look larger. Racism denial accounts for about 75% of the gap (-0.133 of -0.178).

Source: Schaffner, MacWilliams & Nteta (2018), *Political Science Quarterly* 133(1), Table 4; data: CCES 2016.

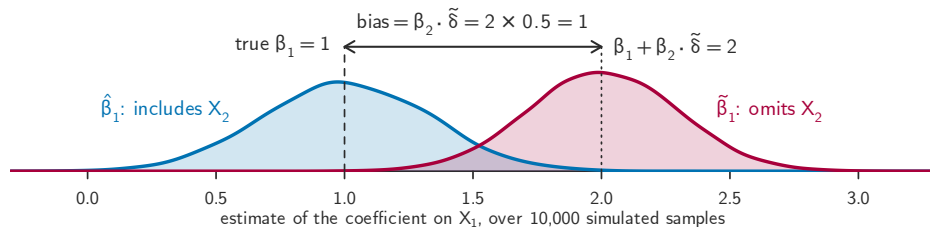
Omitted Variable Bias: Simple Case

If the long model satisfies Assumptions I–IV, take expectations given X ($\tilde{\delta}$ is fixed given X):

$$BIAS[\tilde{\beta}_1 | X] = E[\tilde{\beta}_1 | X] - \beta_1 = \beta_2 \cdot \tilde{\delta}$$

So the bias depends on the relationship between x_2 and x_1 , our $\tilde{\delta}$, and the relationship between x_2 and y , our β_2 .

An omitted variable that is correlated with x_1 ($\tilde{\delta} \neq 0$) and affects Y holding x_1 fixed ($\beta_2 \neq 0$) is called a **confounder**.



Source: simulation with $\beta_1 = 1$, $\beta_2 = 2$, $\tilde{\delta} = 0.5$, $n = 50$; X held fixed across 10,000 samples.

Omitted Variable Bias: Simple Case

Direction of the bias of $\tilde{\beta}_1$ compared to β_1 is given by:

	$\text{Cor}[x_1, x_2] > 0$	$\text{Cor}[x_1, x_2] < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Further points:

- Magnitude of the bias matters too
- If you miss an important confounder, your estimates are biased and inconsistent.
- In the more general case with $k \geq 3$ the bias is more difficult to discern. It depends on all the pairwise correlations.

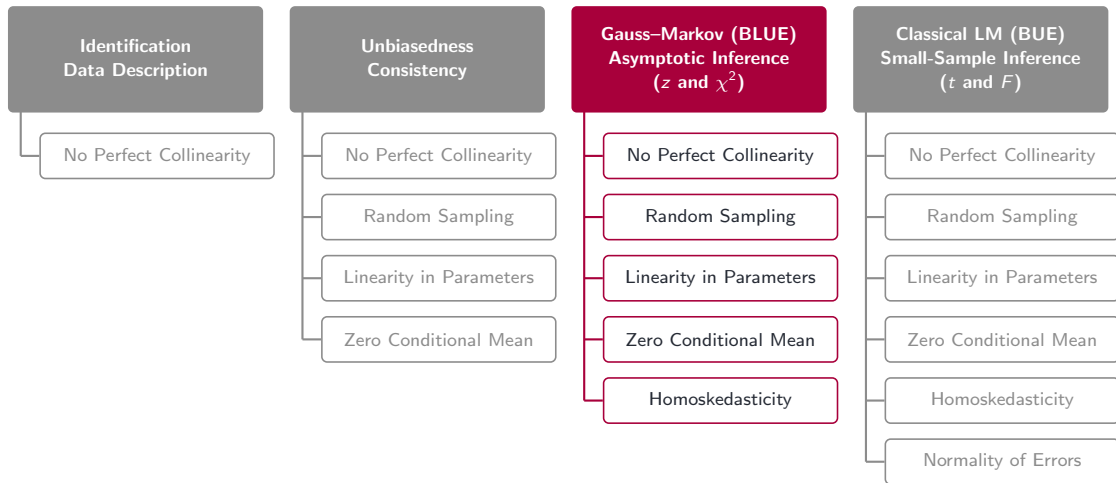
Multiple OLS Assumptions

Omitted Variables Bias

Multicollinearity

Including an Irrelevant Variable

Hierarchy of OLS Assumptions



Variance of OLS Estimators

Given unbiasedness, we know the means of the sampling distributions of our OLS estimators $\hat{\beta}_j$, but we also want to know their variances.

We could compute the variance under Assumptions I–IV, but we'll make an additional assumption to keep things tractable.

Assumption (V. Homoskedasticity)

The error term has the same (conditional) variance given any values of the explanatory variables:

$$V[u|X_1, X_2, \dots, X_k] = \sigma_u^2$$

This implies $V[u] = \sigma_u^2$. The constant error variance σ_u^2 measures the unseen scatter around the population regression (hyper-)plane.

Together the Gauss–Markov Assumptions I–V imply:

$$E[Y|X_1, \dots, X_k] = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k, \quad V[Y|X_1, \dots, X_k] = \sigma_u^2$$

Violation: $V[u|X_1 = 1] \neq V[u|X_1 = 0]$ called heteroskedasticity.

Variance of OLS Estimators

The variance expression is also similar to the simple regression case.

Theorem (Variance of OLS Estimators)

Given OLS Assumptions I–V (Gauss-Markov Assumptions):

$$\text{Var}[\hat{\beta}_j | X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 (1 - R_j^2)} = \frac{\sigma_u^2}{SST_j (1 - R_j^2)}$$

for $j = 1, 2$, where R_j^2 is the R-squared from regression of X_j on the other X variable and a constant.

Proof.

Analogous to the proof for the simple regression case, but algebraically cumbersome. See Wooldridge Chapter 3 Appendix for details.

(We'll cover a more general case of K regressors using matrices, which makes the derivation much simpler.)



Estimating σ^2 to get Standard Errors

To estimate the OLS variance, we first need to estimate σ_u^2 , as in the simple regression case.

Theorem (Unbiased Estimation of the Error Variance)

Given OLS Assumptions I–V, to estimate the error variance σ_u^2 we use our unbiased estimator $\hat{\sigma}_u^2$:

$$\hat{\sigma}_u^2 = \frac{1}{n-3} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-3}$$

Proof.

Analogous to the simple regression case. We'll revisit the proof using matrices. □

- $\hat{\sigma}_u$ is often called the **residual standard error** of the regression.
- $n - 3$ is the **degrees of freedom**.
- When $k > 2$, the degrees of freedom = $n - k - 1$, or (# of observations) – (# of estimated parameters).

Estimating the Variance of OLS Estimators

Given our unbiased estimator for the error variance σ^2 :

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-3} = SSR/(n-3)$$

we can estimate the **standard errors** of our regression coefficients as

$$\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{V}[\hat{\beta}_j]} = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 (1 - R_j^2)}}$$

The Gauss-Markov Theorem: Given OLS Assumptions I–V, the multiple OLS estimator is **BLUE**.

Example: Executive Corruption on Court Independence and Democracy

R Code

```
> fit <- lm(execorr100 ~ courtind + democracy, data = d)
> summary(fit)

~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   78.324      3.593   21.799 < 2e-16 ***
courtind     -10.630      2.113   -5.031 1.27e-06 ***
democracy     -21.510      4.672   -4.604 8.22e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 20.92 on 165 degrees of freedom
Multiple R-squared: 0.5039,    Adjusted R-squared: 0.4978
F-statistic: 83.78 on 2 and 165 DF,  p-value: < 2.2e-16
```

R Code

```
> DF <- nrow(d) - length(coef(fit)) # = DF of the regression
> DF
[1] 165
>
> sqrt(sum(residuals(fit)^2) / DF) # = residual standard error
[1] 20.91745
```

Source: V-Dem v16, 2024, 168 countries, as in Lecture 5-1. execorr100: V-Dem's executive corruption index times 100 (0 = clean); courtind: high court independence (0 = never, 4 = always); democracy: electoral democracy index above 0.5.

Three Components of the OLS Variance Estimators

$$V[\hat{\beta}_j | X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 (1 - R_j^2)} = \frac{\sigma_u^2}{SST_j (1 - R_j^2)}$$

When will the variance be large?

Notice the three components of the variance:

- ① σ_u^2 is the **error variance**, the unseen scatter around the population regression plane. The OLS variance is increasing in σ_u^2 .
- ② SST_j is the **total sample variation in x_j** , which equals:

$$SST_j = n \cdot \frac{1}{n} \sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 = n \cdot MSD(x_j).$$

So the OLS variance is decreasing in sample size and spread of x_j .

- ③ R_j^2 measures the **strength of linear relationship between the explanatory variables**. The OLS variance is increasing in R_j^2 .

Multicollinearity and Collinearity

- Multicollinearity occurs if two or more explanatory variables are strongly correlated. That is, when R_j^2 is large.
- We know that (perfect) **collinearity** violates our regression assumptions and OLS breaks down. What does collinearity imply about R_j^2 ? $R_j^2 = 1$, which implies

$$V[\hat{\beta}_j | X] = \frac{\sigma_u^2}{SST_j(1 - R_j^2)} = \frac{\sigma_u^2}{SST_j(1 - 1)} = \frac{\sigma_u^2}{0}$$

- Therefore, OLS breaks down when $R_j^2 = 1$ (perfect collinearity).
- If $R_j^2 \approx 1$ (multicollinearity), OLS does NOT break down; it just has a very high standard error.

How Do We Detect Multicollinearity?

- The best practice is to directly compute $\text{Cor}(X_1, X_2)$ before running your regression.
- Warning signs in the regression output (none proves multicollinearity; the first is also what omitted variable bias looks like):
 - Large changes in the estimated regression coefficients when a predictor variable is added or deleted
 - Lack of statistical significance despite high R^2
 - Estimated regression coefficients have an opposite sign from predicted
- A more formal indicator is the **variance inflation factor (VIF)**:

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}$$

which measures how much $V[\hat{\beta}_j | X]$ is inflated compared to (hypothetical) uncorrelated data. In R, `vif()` in the `car` package.

Multicollinearity: Simulation Example

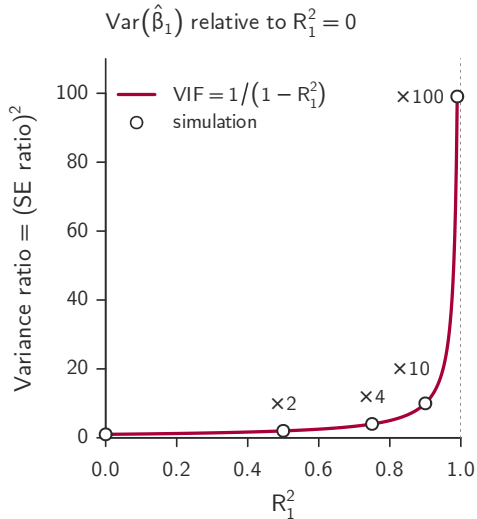
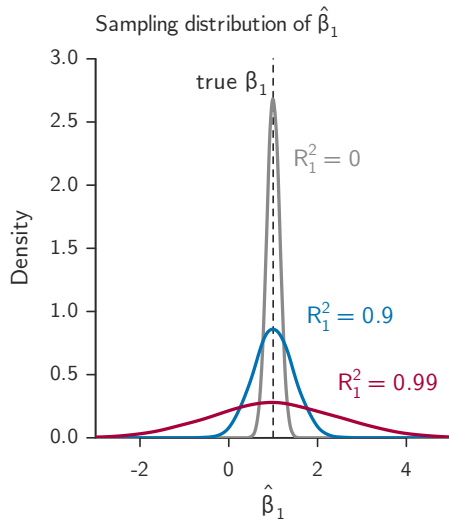
- So how serious is the multicollinearity problem in practice? Let's find out by simulation.
- The model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
- Vary R_1^2 , the squared correlation between X_1 and X_2 . Fix n at 50.
- See what happens to the standard error of $\hat{\beta}_1$

R_1^2	0	.5	.75	.9	.99
$SE(\hat{\beta}_1)$	0.144	0.204	0.288	0.454	1.429
SE ratio to $R_1^2 = 0$	1.00	1.42	2.01	3.16	9.95
$\sqrt{VIF} = 1/\sqrt{1 - R_1^2}$	1.00	1.41	2.00	3.16	10.00

So the standard error doubles at $R_1^2 = .75$, more than triples (about 3.2 times) at .9, and grows tenfold at .99: modest correlations cost little, but the cost explodes as $R_1^2 \rightarrow 1$.

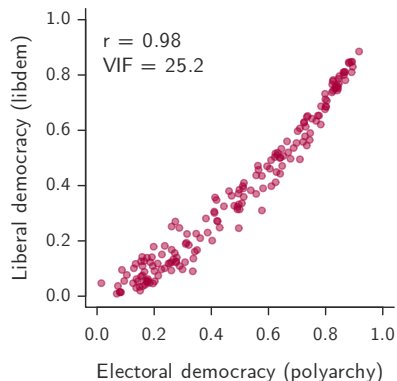
Source: simulation with $\beta_1 = \beta_2 = 1$, $u \sim N(0, 1)$. Each column: 10,000 samples, with X fixed and only u redrawn. Columns differ only in x_2 , which sets R_1^2 .

Variance Inflation due to Multicollinearity ($k = 2$)



Multicollinearity: Two Measures of Democracy

Executive corruption (0–100) on V-Dem's electoral democracy index (polyarchy) and liberal democracy index (libdem), 179 countries in 2024.



	alone	alone	together
polyarchy	–82.6 (5.9)		+130.5 (24.5)
libdem		–89.2 (5.2)	–216.6 (24.4)

Standard errors in parentheses.

- $\text{Cor}(\text{polyarchy}, \text{libdem}) = .980$, so $R_j^2 = .96$ and $VIF = 25.2$.
- Together, the standard errors grow about 4 to 5 times ($\sqrt{VIF} = 5.0$, partly offset by a smaller $\hat{\sigma}_u$). The signs split. Polyarchy turns positive.
- Both slopes are still significant ($t = 5.3$ and -8.9). The warning sign here is the sign flip.

Why? V-Dem builds libdem from polyarchy plus a liberal component. The data have almost no countries that differ in one index at the same value of the other, so OLS cannot pull the two apart.

Source: V-Dem v16, 2024. The codebook says the liberal democracy index “also takes the level of electoral democracy into account.”

So How Should I Think about Multicollinearity?

- Multicollinearity does NOT lead to bias; estimates will be unbiased and consistent.
- Multicollinearity should in fact be seen as a problem of **micronumerosity** (Goldberger 1991), that is, too little data. You can't ask the OLS estimator to distinguish the partial effects of X_1 and X_2 if they are essentially the same.
- If X_1 and X_2 are almost the same, why would you want a unique β_1 and a unique β_2 ? Think about how you would interpret that?
- Relax, you got way more important things to worry about!
- If possible, get more data
- Drop one of the variables (if its true coefficient is not zero, this brings back omitted variable bias), or combine them into one index
- Or maybe linear regression is not the right tool

Multiple OLS Assumptions

Omitted Variables Bias

Multicollinearity

Including an Irrelevant Variable

Including an Irrelevant Variable: Simple Case

True Population Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \quad \text{where} \quad \beta_2 = 0$$

and Assumptions I–IV hold.

Overspecified Model that we use:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Q: Which statement is correct?

- ❶ $E[\hat{\beta}_1] > \beta_1$
- ❷ $E[\hat{\beta}_1] < \beta_1$
- ❸ $E[\hat{\beta}_1] = \beta_1$
- ❹ Can't tell

Including an Irrelevant Variable: Simple Case

Recall: Given Assumptions I–IV, we have:

$$E[\hat{\beta}_j] = \beta_j$$

for **all** values of β_j . So, if $\beta_2 = 0$, we get

$$E[\hat{\beta}_0] = \beta_0, E[\hat{\beta}_1] = \beta_1, E[\hat{\beta}_2] = 0$$

and thus including the irrelevant variable does not affect the unbiasedness. The sampling distribution of $\hat{\beta}_2$ will be centered about zero.

But under Assumption V (homoskedasticity), including an irrelevant variable increases the variance of $\hat{\beta}_1$ (unless x_2 is uncorrelated with x_1 in the sample, so that $R_1^2 = 0$), because

$$V[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{SST_1(1 - R_1^2)} \geq \frac{\sigma_u^2}{SST_1} = V[\tilde{\beta}_1 | X], \quad R_1^2 \equiv R^2 \text{ from regressing } x_1 \text{ on } (1, x_2),$$

where $\tilde{\beta}_1$ is the slope from the correct simple regression of y on x_1 .