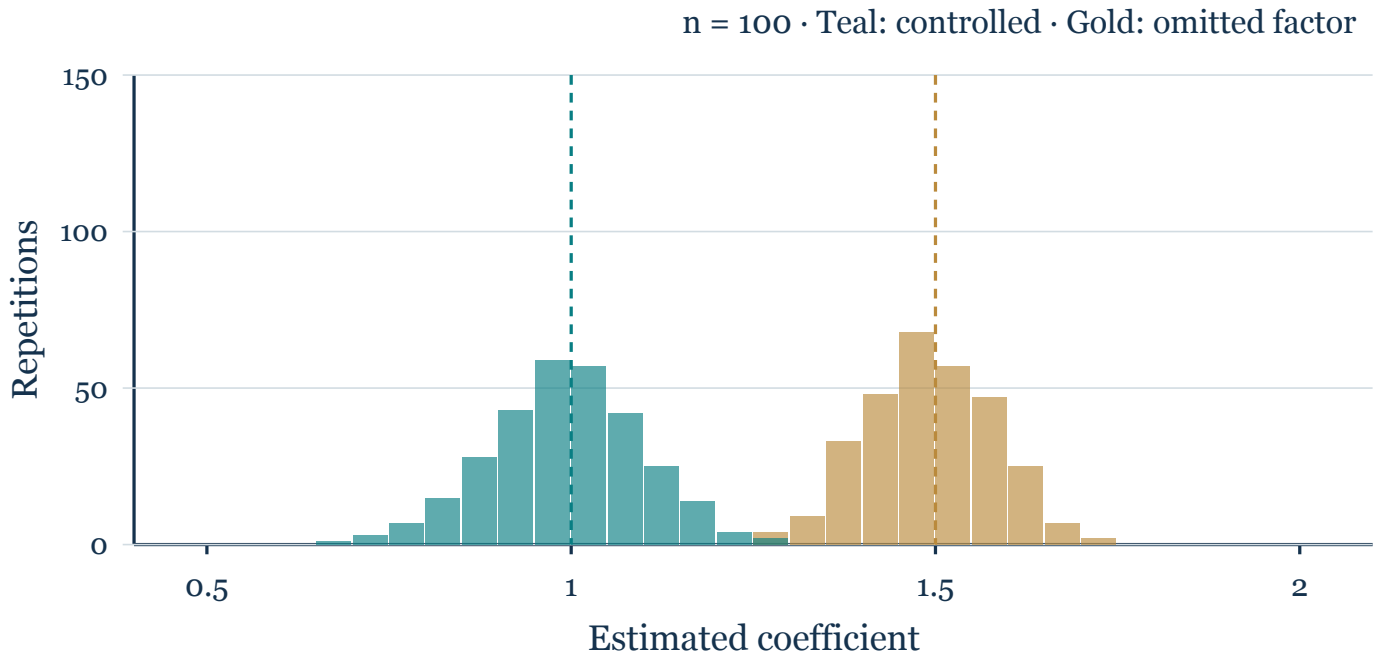
Can a larger dataset rescue a biased design?

Use a known population to separate bias from sampling noise.



Actual Stata Monte Carlo estimates: 300 repetitions at each sample size. The data-generating staffing coefficient is 1; the omitted-variable target is 1.5.

Define the simulation as an economic thought experiment

The generated truth is known by construction.

> $X = A + V$

> $Y = 1 + X + A + U$

> A, V, U independent, with mean zero and variance one

Unbiasedness and consistency ask different questions

One concerns an average at a sample size; the other concerns a limit.

> Unbiasedness: correct repeated-sample mean at a fixed sample size.

> Estimator consistency: concentration near the target as sample size grows.

> Causal consistency: observed/potential-outcome link; a different use of the word.

The law of large numbers stabilises averages

More data can reveal a stable but unwanted relationship.

> Law of large numbers: sample averages approach population means under stated conditions.

> Consistent estimator: concentrates near its target as the sample grows.

Measure co-movement before calculating the limit

Covariance records whether centered variables tend to move together.

> $\mathrm{Cov}(X, A) = E[(X - EX)(A - EA)]$

> $X = A + V: \quad \mathrm{Var}(X) = 2, \quad \mathrm{Cov}(X, A) = 1$

Calculate the wrong limit before simulating it

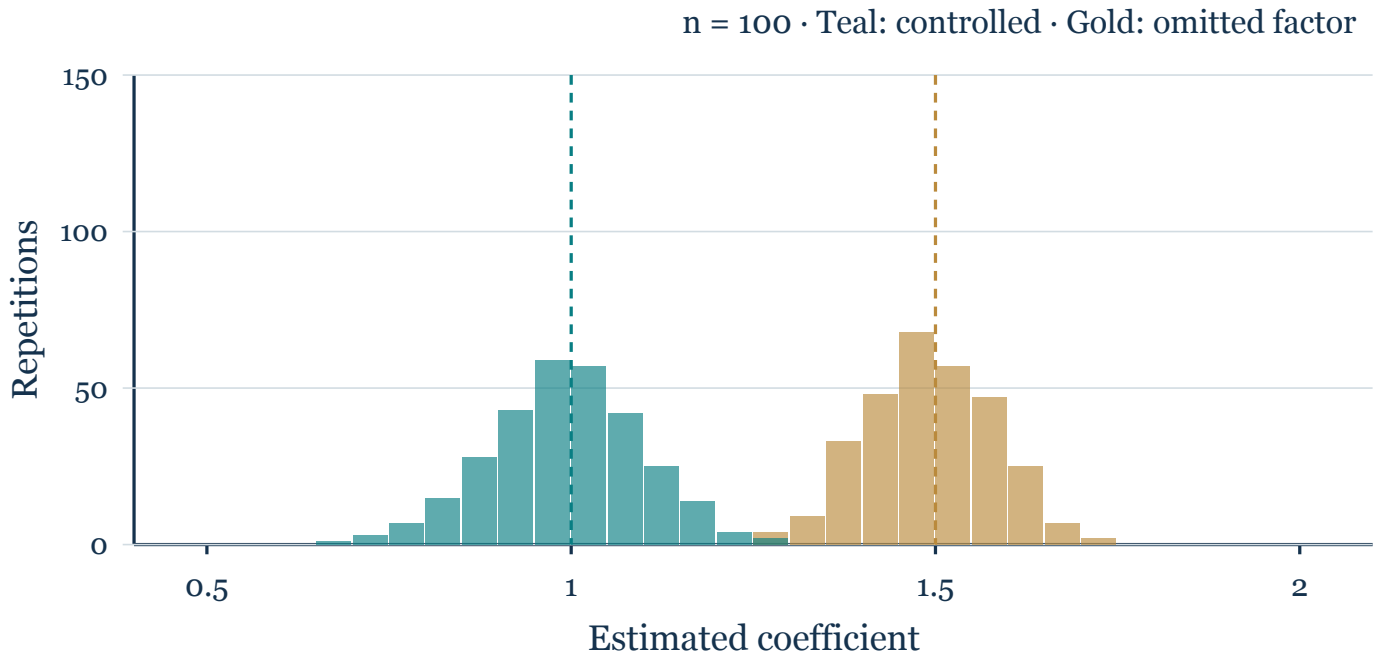
The short slope tends to one and a half.

$$\delta_1 = \frac{\operatorname{Cov}(X, A)}{\operatorname{Var}(X)} = \frac{1}{2}$$

$$\operatorname{plim} \hat{b}_{\text{short}} = 1 + \frac{1}{2} = 1.5$$

Compare repeated samples at two sizes

Concentration alone is not evidence of a correct target.



Actual Stata Monte Carlo estimates: 300 repetitions at each sample size. The data-generating staffing coefficient is 1; the omitted-variable target is 1.5.

Read one large generated sample

A small SE can accompany substantial omitted-variable bias.

```
. regress y x, vce(robust)

------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
           x |   1.516317   .0139733   108.51   0.000     1.488921    1.543712
       _cons |   .9947365   .0194616    51.11   0.000     .9565808    1.032892
------------------------------------------------------------------------------
```

The central limit theorem describes scaled fluctuations

The square-root rate concerns sampling variation around a limit.

> Average of independent zero-or-one draws, each with mean 1/2: $SD(\overline{Y}) = 0.5/\sqrt{n}$

> Coefficient analogue: $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$

Square-root sample-size gains require a comparable sampling model

A stable average in the denominator sets the scale for sampling fluctuations.

> $$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{n^{-1/2}\sum_i (X_i - \bar{X})u_i}{n^{-1}\sum_i (X_i - \bar{X})^2}.$$

> Positive denominator limit: $Q = \mathrm{Var}(X) > 0$.

> With $S = \mathrm{Var}[(X - EX)u]$, approximate slope variance is $S/(nQ^2)$.

A restricted-model LM test asks what the omitted columns still explain

The auxiliary statistic and reference distribution must match the null.

> $H_0 : \gamma_1 = \cdots = \gamma_q = 0$ for the q additional columns.

> Conventional homoskedastic auxiliary test: $LM = nR^2_{aux} \xrightarrow{d} \chi^2_q$.

> Estimate under the null, then test the information left in the omitted columns.

Compare sample size and precision in the district data

Try the comparison before opening the explanation.

Practice

Use overleaf_technical.do to fit score on str and el for a fixed random 105-district subset and all 420 districts. Independently reproduce both robust fits. Compare sample sizes, staffing variation, coefficients and standard errors. Does the observed SE ratio have to equal two?

Compare sample size and precision in the district data: explanation

Follow the calculation and interpret its units.

> The smaller sample has 105 observations and the full sample 420, but its regressor mix and realised residuals differ. The observed ratio need not equal the square-root size ratio of two. The 105-district staffing estimate is about -0.176454 with robust SE 0.689649; the 420-district estimate is about -1.101296 with SE 0.432847. Their observed SE ratio is about 1.593286. Report values from your own successful run. The rate approximation requires comparable information and a valid sampling model; neither fit establishes a causal staffing effect.

Two simple estimators separate bias from concentration

The two properties answer different repeated-sample questions.

> $\widehat{\mu}_n = Y_1$: $E\widehat{\mu}_n = \mu$, $\mathrm{Var}(\widehat{\mu}_n) = \sigma^2 > 0$.

> $\widetilde{\mu}_n = \overline{Y}_n + 1/n$: bias $= 1/n$, but $\widetilde{\mu}_n \xrightarrow{p} \mu$.

More data, same bias

Try the comparison before opening the explanation.

Practice

An estimator converges to 1.4 while the causal target is 1.0. What happens to the bias problem as n grows?

More data, same bias: explanation

Follow the calculation and interpret its units.

> It remains: the estimator is inconsistent for the causal target. Sampling variation may shrink around the wrong limit. Narrow intervals around 1.4 do not repair the design.

The OLS large-sample argument separates a stable denominator from random scores

Define the moments needed for consistency and normal approximation.

> $Q = E[x_i x_i\prime]$, $E[x_i u_i] = 0$, $S = E[u_i^2 x_i x_i\prime]$.

> $\sqrt{n}(\widehat{\beta} - \beta) = [n^{-1} \sum_i x_i x_i\prime]^{-1} n^{-1/2} \sum_i x_i u_i$.

> $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, Q^{-1} S Q^{-1})$; approximate covariance $Q^{-1} S Q^{-1}/n$.

Count the restrictions before choosing an LM reference

Try the comparison before opening the explanation.

Practice

A conventional homoskedastic LM exercise has n=500, auxiliary R squared 0.012 and two excluded-coefficient restrictions. Calculate the statistic and compare it with the 5% chi-squared critical value 5.991 for two degrees of freedom. Would a three-degree-of-freedom critical value answer the stated question?

Count the restrictions before choosing an LM reference: explanation

Follow the calculation and interpret its units.

> LM=500*0.012=6. With q=2, compare it with 5.991, so the conventional test narrowly rejects at 5% under its assumptions. A three-degree-of-freedom critical value tests a different restriction count and is incorrect here. Rounding and the reference assumptions matter near the boundary.