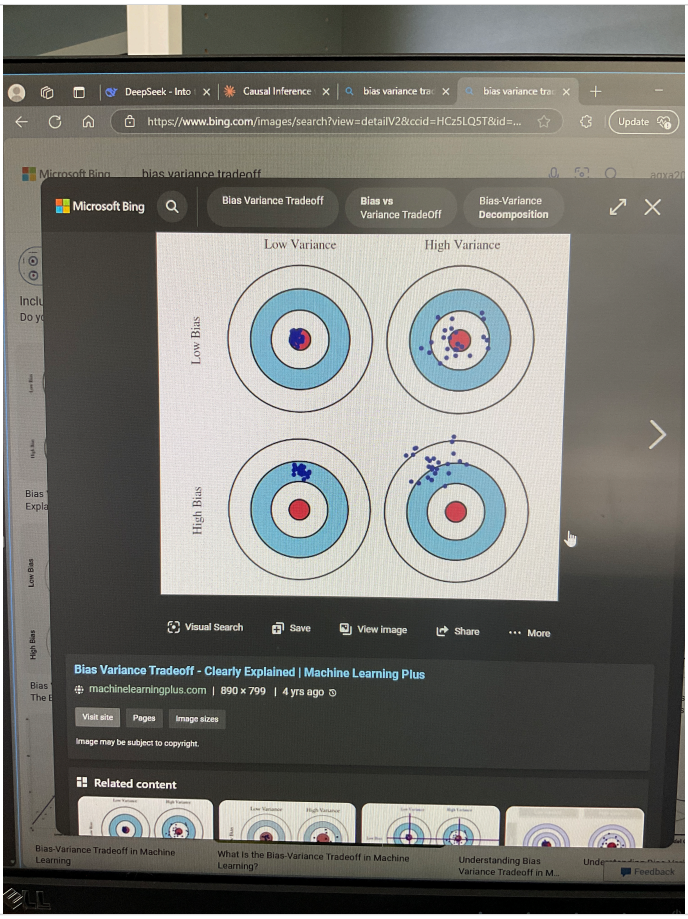
Introduction

This lecture will cover the effects of mis specification (i.e, omitting variables, including irrelevant ones, or failing linearity straight up), multicollinearity (i.e, regressors correlated with each other), and how to choose a model when multiple candidates are available.

Treatments and the level of concern for each topic present or core difference between the econometrics and data science methodologies. So let me start by giving a few comments on the distinction.

As we already know, econometrics is ultimately interested in finding causal relationships. Even at the cost of efficiency. In contrast, Data science prioritizes prediction accuracy obtained by very high efficiency (low variance).
Unfortunately, unbiasedness and efficiency are complementary. That is, there is a natural trade-off between bias and variance.

Bias-Variance tradeoff



In general, the lower the variance the higher the prediction accuracy (or confidence of the model) but the higher the bias. That is, bias and variance are inversely related.

Unbiasedness is a necessary but not sufficient property for causal identification. That is, even if our OLS estimates are BLUE we still cannot claim a causal relationship. For our model to gain causal interpretation we need at least the following four conditions:

1) Exogeneity ($E[\varepsilon_i | X] = 0$) must hold deterministically, not just as a statistical property. Although finding evidence in data is often good enough.

2) No omitted variables. We will see why today. In other words, our

model must be perfectly specified: no missing nor redundant variables!

3) No reverse causality or simultaneity. That is, $X \rightarrow Y$ not $X \leftrightarrow Y$

4) No measurement error.

We can test for 1-3 but 4 is notoriously annoying. We might be able to return to this towards the end of the semester.

A lot of people out there don't pay enough attention to this and end up making causal claims willy-nilly.

Don't be one of those people.

After me: "BLUE OLS is not a causal finding".

Data Science approaches ignore this completely, hence their estimates are often biased by design but highly efficient. It is a simpler but less rigorous world. They are just starting to realize how important it is to produce unbiased estimates, think of high-dimensional A/B testing in marketing or product development for example.

I predict that demand for

econometricians with data science skills will surge in the next five years because of this.

Let's formalize this trade off to understand the distinction between the objectives of both approaches.

We can quantify the total error of a model with the Mean Squared Error (MSE) which decomposes the errors into bias and variance.

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Recall $E[x^2] = Var(x) + E[x]^2$

$$\Rightarrow E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta} - \theta) + E[\hat{\theta} - \theta]^2$$

$$\Rightarrow \underbrace{Var(\hat{\theta})}_{\text{Variance}} + \underbrace{E[\hat{\theta}-\theta]^2}_{\text{Bias}} + IE$$

often we include an irreducible error term to account for the random noise generated by our model.

Then, given $MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta},\theta)^2$ metrics and DS distinguish themselves in the optimization procedure.

(Econometrics) We require unbiasedness (i.e, $E[\hat{\theta}-\theta] = 0$) so we optimize MSE under this constraint. Hence,

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + 0$$

In an OLS setting,

$$\hat{\theta} = (X^T X)^{-1} X^T y \quad , \quad E[\hat{\theta}] = \theta$$

$$Var(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$$

Hence, $MSE(\hat{\theta}) = \sigma^2 (X^T X)^{-1}$

Intuitively, this constraint limits the possible estimators we can use.

(Data science) No constraints required, the goal is to minimize both terms jointly. So we consider the full expression

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$$

Techiques used here are called

regularization (like Ridge or Lasso regr.) which introduce bias on purpose to decrease the variance, ensembles or black-box NNs which "average out" noise at the cost of interpretability.

I think we will finish the core contents of ECON 167 early, so we can talk about these more towards the end of the semester if you are interested.

In short, as econometricians we rather keep our models lean and unbiased for the sake of truth. The following table, created with Deep Seek, summarizes

these differences nicely:

Econometric Over-Specification vs. Data Science Regularization

Aspect	Econometric Over-Specification	Data Science Regularization
Bias	Unbiased	Biased (by design)
Variance	Higher	Lower
Goal	Parameter interpretability	Prediction accuracy
Example	Adding irrelevant Z to $y = X\beta$	LASSO shrinking coefficients

Great. Let's start studying the various ways in which the OLS assumptions can break.

Gauss Markov Assumptions (GMAs)

1. (Linearity in the parameters)

$$y = X\beta + \varepsilon$$

2. (Full Rank) We can't have perfect multicollinearity between regressors

$$\text{Rank}(X) = K$$ where K is the # of regressors

3. (Exogenity) Errors should be independent of the regressors X

$$E[\varepsilon | X] = 0$$

4. (Homoskedasticity) Given the data for our regressor, the errors should have constant variance

$$Var(\varepsilon | x) = \sigma^2 I$$

5. (No autocorrelation) Errors should be independent of each other across observations

$$Cov(\varepsilon_i, \varepsilon_j | x) = 0 \quad , \forall\, i \neq j$$

4 and 5 together are often called "spherical errors or disturbances" requirement.

Under GMAs the OLS estimator $(x^T x)^{-1} x^T y$ is BLUE.

Multicollinearity

This happens when two or more regressors are correlated (i.e., linear

combinations). The consequence is that these will move together, making it difficult to figure out their individual effects on the dependent variable

There are two flavors:

1) **Perfect multicollinearity**: This is extreme, and hence rare. It happens when two regressors are perfectly correlated (i.e., $corr(x_1, x_2) = 1$ or -1). For example,
$X_1 = 2X_2 + 1$ or $X_3 = X_1 + 3X_2$

2) **High multicollinearity**: Much more common.
Often correlations between |0.7| and |0.9|. Income and education may exhibit

this

for example.

Anything below 0.7 is often not a big deal because OLS is robust. Remember our simulations from last week.

Fortunately, the OLS estimator remains unbiased (unless we have perfect multic. because the inverse won't exist) but it becomes less efficient. That is, the variance is much larger. This means larger SE, wider confidence intervals, and unreliable t-statistics. Like trying to hear to people

talking over each other, hard to tell who is saying what.

There are 3 direct effects for

regression:

1) Inflated SE: A variable might be important but OLS is unable to distinguish its effect from the one(s) is correlated with. This leads to small t-stats

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$ which would suggest otherwise.

Hence the probability of Type II error increases.

2) Unstable coefficient estimates: Small changes to the data couses big changes to the estimates. Hence OLS is not stable in repeated sampling.

3) Tough interpretation: The ceteris paribus assumption ("other things constant") doesn't make sense anymore. That is, if x_1 and x_2 are correlated in a regression $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ then $\hat{\beta}_1$ will not be "the change in y from a unit change in x_1"

There are a few methods to catch (test for) multicollinearity:

1) Check the rank of the data matrix X.

If $\det(X^T X)$ is 0 then the matrix is singular and we have perfect multic. For big datasets, you can use the condition number $\kappa = \sqrt{\lambda_{max}}$

λ_{min}

- If $\kappa > 30$ there is severe multic.
- If $\kappa < 10$ you are fine

2) Correlation matrix: Plot a heatmap or correlation plot of all your regressors.
   - If $\text{Corr}(x_i, x_j) > |0.8|$ you are in trouble

3) Variance Inflation factor (VIF): A go-to metric. It measures, for each regressor x_j, how much its variance is inflated due to correlation with other regressors.

$$VIF_j = \frac{1}{1 - R^2}$$

where R_j^2 is the R^2 goodness of fit from $x_j \sim \Sigma x_i$, $\forall i \neq j$

- If VIF < 1 there is no multicollinearity
- If VIF < 5 you are fine
- If VIF > 5 you should look into it more. Suggest moderate multic.
- If VIF > 10 you are in trouble.

There are a couple of solutions we can experiment with in case of multicollinearity:

1) If x_i and x_j are highly correlated, then

drop one. They explain the same thing. But you lose information and risk OVB.

2) Get more data. Multic. often decreases with sample size. But this might not be possible, specially in macro...

Ex: $y = X\beta + \varepsilon$, $\varepsilon \sim N(0,1)$

$$X = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 2 & 5 \\ 1 & 3 & 7 \end{bmatrix} \Rightarrow x_2 = 2x_1 + 1$$

$$y = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$X^TX = \begin{bmatrix} 3 & 6 & 15 \\ 6 & 14 & 34 \\ 15 & 34 & 83 \end{bmatrix}, \quad \det(X^TX) = 0$$

OLS can't solve this!

Proof: Sub in X_3 into the model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 (2X_1 + 1) + \varepsilon$$

$$y = \beta_0 + \beta_1 X_1 + 2\beta_2 X_1 + \beta_2 + \varepsilon$$

$$y = \underbrace{(\beta_0 + \beta_2)}_{\omega_1} + X_1 \underbrace{(\beta_1 + 2\beta_2)}_{\omega_2} + \varepsilon$$

We can estimate ω_1 and ω_2 but not $\beta_0, \beta_1, \beta_2$ individually. So we can't separate the individual effects of

each variable.
We have infinitely many solutions!

Takeaway: Multicollinearity reduces precision but not bias. Inflated SE affect inference by reducing t-stats. You can test for it with VIF, correlation matrices or rank(X^TX). OLS is robust to low multicollinearity, specially with large sample sizes, but it is a problem for identifying individual effects (required for causal inference).

Misspecification

Two latent assumptions implied by GMA are that 1) the model is correct (i.e., no missing nor redundant variables) and 2) the parameter are actually linear.

The "correctness" of the model is informed by theory. But we don't have a theory for everything and often, because theory is context dependent, is wrong. So we need a protocol to build our models from the data.

One approach that has become more popular with cheap compute is "general-to-simple". Start by specifying a very complicated

regression. Add in squared and cubed vars, logged vars, and the whole "kitchen sink".

Then, trim down the model by testing multiple restrictions and filtering out insignificant variables step by step.

Under-specified model (OVB)

Omitted Variable Bias (OVB) occurs when we fail to include a variable that in fact belongs in the true model. You are leaving out something that systematically affects the dependent variable.

The statistical consequence is that the effect of this missing variable

will be captured by the error term, and hence exogeneity is violated $E[\varepsilon | X] \neq 0$. Hence, OLS will be biased. But where is this bias coming from?

Suppose the true model is

$$Y = X\beta + Z\gamma + \varepsilon$$

X is the matrix of K regressors included

Z is the matrix of m regressors omitted

Exogeneity and Homoskedasticity hold in this model $E[\varepsilon | X, Z] = 0$ and $Var(\varepsilon | X, Z) = \sigma^2 I$

Because we omit, either by ignorance

or lack of data, the matrix Z, we estimate

$$Y = X\beta + u$$

where the new error u captures what we are missing

$$u = Z\gamma + \varepsilon$$

Run OLS to get $\hat{\beta} = (X^TX)^{-1}X^Ty$

Substitute the true model

$$\hat{\beta} = (X^TX)^{-1}X^T(X\beta + Z\gamma + \varepsilon)$$

$$= \beta + (X^TX)^{-1}X^TZ\gamma + (X^TX)^{-1}X^T\varepsilon$$

Take conditional expectation

$$E[\hat{\beta} \mid X] = \beta + (X^TX)^{-1}X^TZ\gamma$$

$$E[\beta \mid X, Z] = \beta + (X'X)\, X'Z\gamma$$

$$+ E[(X^TX)^{-1}X^T\varepsilon \mid X, Z]$$

↳ exogeneity applies with true error ε

$$E[\hat{\beta} \mid X, Z] = \beta + \underbrace{(X^TX)^{-1}X^TZ\gamma}_{\text{Bias}}$$

$$\text{Bias}(\hat{\beta}) = (X^TX)^{-1}X^TZ\gamma$$

The OLS will be unbiased under OVB only if X and Z are orthogonal ($X^TZ = 0$) or all the omitted terms are irrelevant ($\gamma = 0$).

The direction of the bias will depend on the sign of the correlation between X and Z (given by X^TZ) as well as

the effect of z on y (via γ). Also, note that these values don't depend on the sample size. So more data will not help. OLS is inconsistent under OVB.

Ex. Fire up your R to do the math!

(True model) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

$y :=$ final grade,

$x_1 :=$ Homeworks avg., $x_2 :=$ Effort (z)

If effort correlates with HW avg. and $\gamma > 0$ then $\hat{\beta}_1$ will be overestimated.

Suppose the full data is

$$X = \begin{bmatrix} 1 & 80 & 5 \\ 1 & 70 & 3 \\ \vdots & & \end{bmatrix} \qquad y = \begin{bmatrix} 85 \\ 75 \\ 95 \end{bmatrix}$$

$\begin{bmatrix} 1 & 90 & 7 \end{bmatrix}$ ← effort level

↳ HW avg.

and that the true parameters are

$$\beta = \begin{bmatrix} 10 \\ 0.5 \\ 2 \end{bmatrix}$$

Because I can't measure effort, I estimate with $X_1 = \begin{bmatrix} 1 & 80 \\ 1 & 70 \\ 1 & 90 \end{bmatrix}$

$$\hat{\beta} = (X_1^T X_1)^{-1} X_1^T y = \begin{bmatrix} \frac{97}{3} & -\frac{2}{5} \\ -\frac{2}{5} & \frac{1}{200} \end{bmatrix} \begin{bmatrix} 255 \\ 20.500 \end{bmatrix}$$

$$= \begin{bmatrix} 45 \\ \frac{1}{2} \end{bmatrix}$$

$\hat{\beta}_1$ is correct but $\hat{\beta}_0$ is way off.

Takeaway: OVB systematically skews your estimates in the direction of the coefficient of the omitted variables and their correlation with the included ones.

Very hard to avoid. But applying the "general-to-simple" approach can help. Otherwise, consider a different estimation procedure like IV (we might get to cover this one).

One way to test for OVB is via sensitivity analysis, which is a simulation procedure. See Cinelli and

Hazlett (2023) "An OVB framework for sensitivity analysis of Instrumental Variables" for a reference.

Over-Specified model

Including irrelevant variables is not a huge deal. OLS will be unbiased but the variance will be larger. This means larger SE, wider confidence intervals, and less statistical power for inference.

Suppose the true model is

$$Y = X\beta + \varepsilon$$

But we estimate

$$Y = X\beta + Z\gamma + \varepsilon$$

where all the variables in Z are irrelevant

Recall the three important matrices we've seen in the partitioned regression lecture.

Projection $(P) = X(X^TX)^{-1}X^T$

→ from $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty$

Annihilator $(M) = I - P$

→ Aka residual maker

Let's create a partitioned regression

$$\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \left(\begin{bmatrix} X^T \\ Z^T \end{bmatrix} [X \; Z] \right)^{-1} \begin{bmatrix} X^T \\ Z^T \end{bmatrix} y$$

Then,

$$\hat{\beta} = (X^T M_Z X)^{-1} X^T M_Z y$$

where $M_Z = I - Z(Z^T Z)^{-1} Z^T$ projects X and y to the column space of Z

Plug in true model

$$\hat{\beta} = (X^T M_Z X)^{-1} X^T M_Z (X\beta + \varepsilon)$$

$$= (X^T M_Z X)^{-1} X^T M_Z X \beta + (X^T M_Z X)^{-1} X^T M_Z \varepsilon$$

$$= \beta + (X^T M_Z X)^{-1} X^T M_Z \varepsilon$$

* see the lecture notes 9.2 for another derivation.

$$E[\hat{\beta}] = E[\beta] + E[(X^T M_Z X)^{-1} X^T M_Z \varepsilon]$$

$$= \boxed{\beta}$$ unbiased under exogeneity if the Z variables are irrelevant (i.e., $\gamma = 0$)

what about the variance?

$$Var(\hat{\beta} \mid X, Z) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T \mid X, Z]$$

$$= (X^T M_Z X)^{-1} X^T M_Z \, E[\varepsilon \varepsilon^T \mid X, Z] \, M_Z X (X^T M_Z X)^{-1}$$

since M_Z is idempotent $M_Z^2 = M_Z$

$$Var(\hat{\beta} \mid X, Z) = \sigma^2 (X^T M_Z X)^{-1}$$

Now, the BLUE OLS has variance

$$Var(\hat{\beta}_{true}) = \sigma^2 (X^T X)^{-1}$$

since M_z is idempotent $M_z^{-1} = M_z$

Hence $(x^T M_z x) \leq (x^T x)$ and so

$$(x^T M_z x)^{-1} \geq (x^T x)^{-1}$$

Same logic as dividing by a smaller scalar

$$Var(\hat{\beta}_{over}) \geq Var(\hat{\beta}_{true})$$

As a consequence of this inefficiency, we'll get lower t-stats (since SEs are larger) and therefore overaccept the null.

Takeaway: Start with a "kitchen sink" regression understanding that the SEs will be inflated. Start trimming down by eliminating least significant vars. Then re estimate and keep trying. We

can use a few criteria to compare two models. We will see these at the end of the lecture.

Functional form Mis specification

Sometimes we fully fail to capture the true form of the DGP. Two ways in which this arises:

1) Nonlinearity in regressors:
   (true) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$
   (estimated) $y = \beta_0 + \beta_1 x_1 + u$
   - u would contain $\beta_2 x_2^2 + \varepsilon$ and so exogeneity is broken and we get biased estimates

2) Nonlinearity in parameters:

Either the true DGP is inherently non-linear like $y = \beta_0 + \beta_1 e^{\beta_2 x_1} + \varepsilon$ or inherently linear like $y = AK^{\alpha}L^{\beta}$ which can be transformed into a linear function

We have two ways to test 1):

- Residuals vs fitted plot: Any systematic pattern (e.g., U-shaped or positive slope) suggests non-linearity.

- RESET* (Ramsey) test: Run $y \sim x$ to get $\hat{y}$, then add powers of $\hat{y}$ ($\hat{y}^2, \hat{y}^3$) to the model. If these terms are significant then the functional form might be misspecified.

* Regression Equation Specification Error Test

Dealing with 2) is harder and, unless the equation can be linearized, it invalidates OLS.

Non-linear Least Squares

Same optimization concept (min SSR) but without assuming linearity.

$$y_i = g(x_i; \beta) + \varepsilon \quad , \quad x_i \in X$$

↳ vector col of X

The NLS estimator minimizes SSR

$$\hat{\beta}_{NLS} = \arg\min_{\beta} \sum (y_i - g(x_i; \beta))^2$$

The problem is that we'll probably don't have a closed form solution because there may exist multiple global minima. Also, our optimization algorithm may get stuck in local minima. To solve this optimization problem we have to rely on numerical solutions and iterative optimization like Gradient Descent.

For example, $g(\cdot)$ could be a Cobb-Douglas $g(K, L) = AK^{\alpha}L^{\beta}$ which requires NLS unless you are fine linearizing it to get elasticities.

The challenge with iterative methods, and this gets to the heart of ML, is

picking good initial values for β (to avoid local minima), convergence (poor data or model fit can prevent it), and computational complexity (much slower optimization).