ECON 6200: Econometrics II

Welcome

Welcome to ECON 6200!

First things first:

- The course directly builds on ECON 6190.
  This material will not be revised.
- If you need to catch up on background regarding probability and statistics, I can recommend Bruce Hansen's and Richard Durrett's textbooks.
- See syllabus for assessment etc.

Welcome

Very rough outline of first weeks:

- We first analyze the linear model in some detail.
- We then generalize to IV, TSLS, and pretty rapidly to Generalized Method of Moments (GMM) and extremum (or m-) estimation.
- I also expect to cover nonparametrics as well as bootstrap.

The Linear Model

You will have encountered the **Ordinary Least Squares** estimator:

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i.
\end{aligned}
$$

The Linear Model

You will have encountered the **Ordinary Least Squares** estimator:

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i.
\end{aligned}$$

The interpretation of $\hat{\beta}$ depends on context:

1. In any given sample, it just projects $\boldsymbol{Y}$ onto $\boldsymbol{X}$.
2. Under weak assumptions, it converges to the population analog $\beta^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}XY$, which is the population projection coefficient and characterizes the best linear predictor under square loss.
3. Under stronger assumptions, it estimates a causal effect of X on Y.

The Linear Model

You will have encountered the **Ordinary Least Squares** estimator:

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i .
\end{aligned}
$$

The interpretation of $\hat{\beta}$ depends on context:

1. In any given sample, it just projects $\boldsymbol{Y}$ onto $\boldsymbol{X}$.
2. Under weak assumptions, it estimates the population projection coefficient.
3. Under much stronger assumptions, it estimates a causal effect of X on Y.

We will elaborate these points in this order and then develop the classic theory of Least Squares estimation.

The Linear Model

You will have encountered the **Ordinary Least Squares** estimator:

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i Y_i .
\end{aligned}$$

Some notational conventions:

- The random variable X has realization X_i which may take value $x \in \boldsymbol{R}^K$.
- $\mathbb{E}_n(\cdot)$ is a sample average.
- I do not in general boldface vectors, but I will use boldface to indicate data matrix notation as in the first line above.

Ordinary Least Squares (OLS) as In-Sample Projection

Recall that $\hat{\beta}$ can be derived as minimization (in b) of

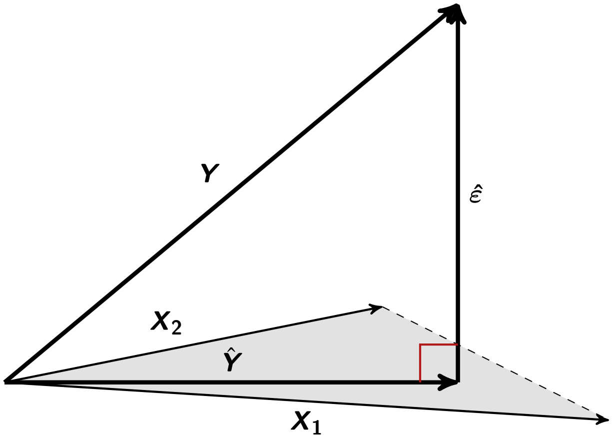
$$\sum_{i=1}^{n}(Y_i - X_i'b)^2 = (\boldsymbol{Y} - \boldsymbol{X}b)'(\boldsymbol{Y} - \boldsymbol{X}b).$$

Of course, this is where its name came from.

But recall also that (as especially obvious in the second expression) this minimization defines b s.t. $\boldsymbol{X}b$ is the point in the span of $\boldsymbol{X}$ that is closest to $\boldsymbol{Y}$ in Euclidean distance.

That is to say, we projected $\boldsymbol{Y}$ onto $\boldsymbol{X}$.

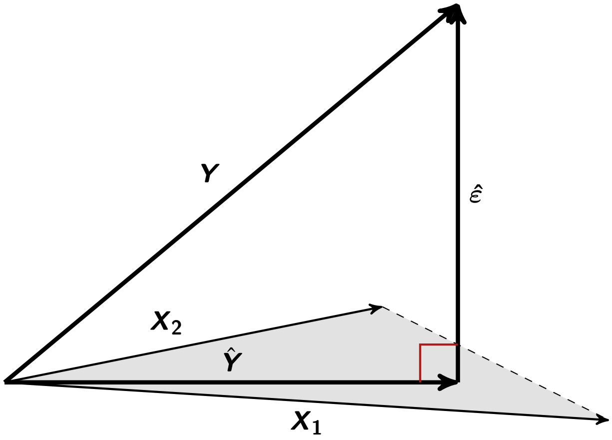
Ordinary Least Squares (OLS) as In-Sample Projection



Illustration

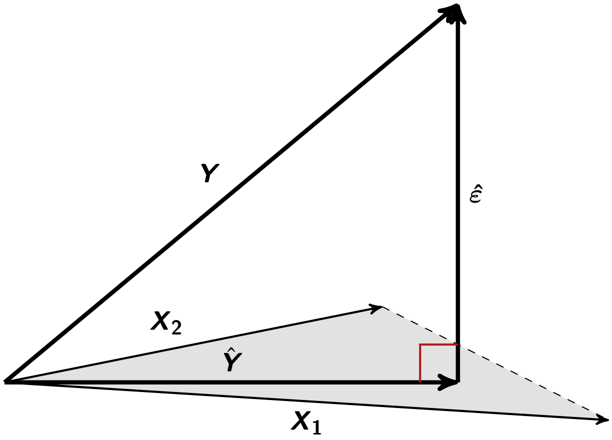
This uses demeaned vectors; alternatively, $\boldsymbol{X_1} = \text{constant}$.

Ordinary Least Squares (OLS) as In-Sample Projection



The illustration also defines the projection $\hat{\boldsymbol{Y}} \equiv \boldsymbol{X}\beta = \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2$ and the residual $\hat{\varepsilon} \equiv \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.

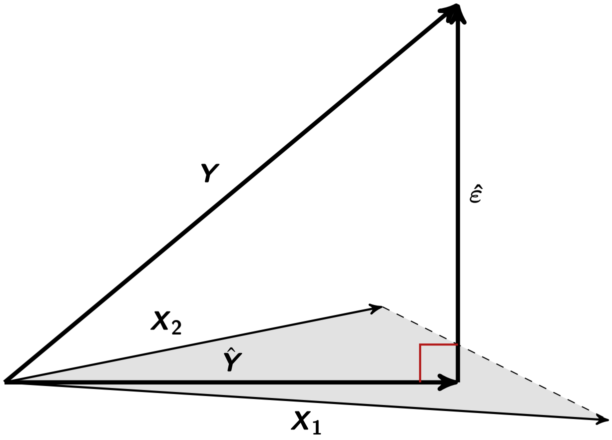
Ordinary Least Squares (OLS) as In-Sample Projection



Corollary: Sum of Squares Decomposition

By Pythagoras, we have $\boldsymbol{Y}'\boldsymbol{Y} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\varepsilon}'\hat{\varepsilon}$.

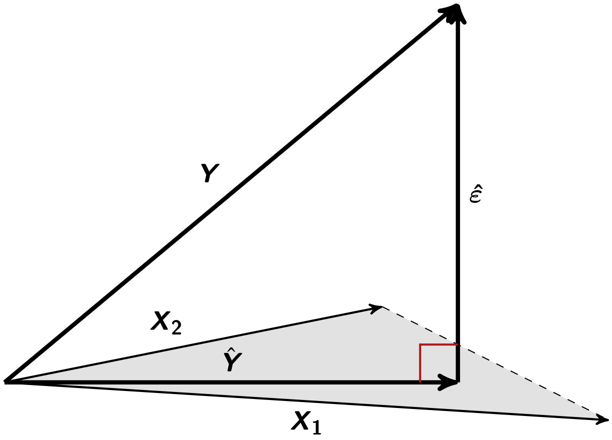
Ordinary Least Squares (OLS) as In-Sample Projection



Corollary: Sum of Squares Decomposition

Equivalently, $\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ or $SST = SSE + SSR$.

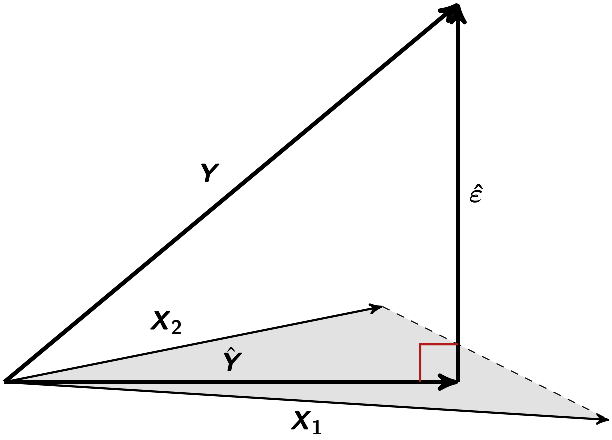
Ordinary Least Squares (OLS) as In-Sample Projection



It follows immediately that $R^2 \equiv SSE\,/\,SST \in [0, 1]$.

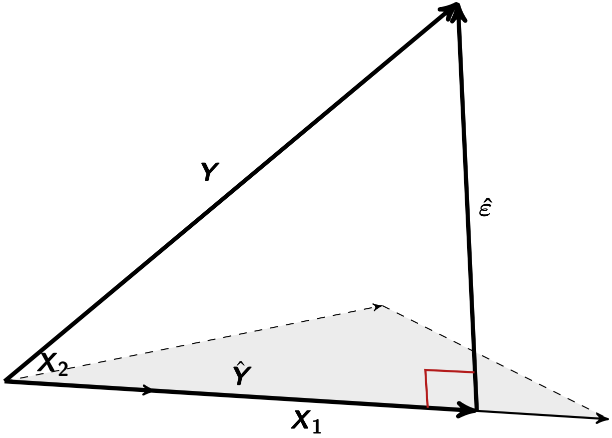
The extreme values of R^2 correspond to $\hat{\boldsymbol{Y}} = \mathbf{0}$ respectively $\hat{\varepsilon} = \mathbf{0}$.

Ordinary Least Squares (OLS) as In-Sample Projection



It also follows that $\hat{\varepsilon}$ is *by construction* orthogonal to $\hat{\boldsymbol{Y}}$.
We will recap this and related basic facts in the first homework.

What Happens with Collinear $\boldsymbol{X}$?



With collinear $\boldsymbol{X}$, the set on which we project is of lower dimension.

The projection $\hat{\boldsymbol{Y}}$ is still unique. The projection coefficients are not.

The Linear Model: Algebra of Projection

The projection coefficient $\hat{\beta}$ is defined as

$$\hat{\beta} \equiv \arg\min_{\beta} \sum_{i} (Y_i - X_i'\beta)^2 = \arg\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)' (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

and can be characterized by FOC (the SOC is obvious)

$$\begin{aligned} \frac{d}{d\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)' (\boldsymbol{Y} - \boldsymbol{X}\beta) &= -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\beta \stackrel{!}{=} \mathbf{0} \\ \implies \hat{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Y}. \end{aligned}$$

The fitted values and residuals equal

$$\begin{aligned} \hat{\boldsymbol{Y}} &= \boldsymbol{X}\hat{\beta} = \underbrace{\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'}_{\boldsymbol{P}_X\text{, the projection matrix}} \boldsymbol{Y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} \\ \hat{\boldsymbol{\varepsilon}} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \underbrace{\left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)}_{\text{annihilator matrix}} \boldsymbol{Y}. \end{aligned}$$

The Linear Model: Decomposing the Projection

Application: Frisch-Waugh(-Lovell)

The projection of Y on X can be decomposed, giving rise to some important results.

To fix ideas, consider projecting Y

- on the scalar X_1 (plus a constant), getting slope coefficient $\tilde{\beta}_1$, versus
- on the scalars (X_1, X_2) (plus a constant), getting slope coefficients $(\hat{\beta}_1, \hat{\beta}_2)$.

Can we interestingly relate $\tilde{\beta}_1$ to $\hat{\beta}_1$?

The Linear Model: Decomposing the Projection

Application: Frisch-Waugh(-Lovell)

The projection of Y on X can be decomposed, giving rise to some important results.

To fix ideas, consider projecting Y

- on the scalar X_1 (plus a constant), getting slope coefficient $\tilde{\beta}_1$, versus
- on the scalars (X_1, X_2) (plus a constant), getting slope coefficients $(\hat{\beta}_1, \hat{\beta}_2)$.

Can we interestingly relate $\tilde{\beta}_1$ to $\hat{\beta}_1$?

Yes! To do so, consider also projecting X_1 on X_2, getting slope coefficient $\hat{\gamma}$. To simplify expressions, also assume all variables are demeaned. Across regressions, this leads to FOC's on next slide.

The Linear Model: Decomposing the Projection

$$
\begin{aligned}
0 &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\
0 &= \mathbb{E}_n\big(X_2(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\
0 &= \mathbb{E}_n\big(X_2(X_1 - \hat{\gamma} X_2)\big)
\end{aligned}
$$

Combine the first two, then use the third one to find

$$
\begin{aligned}
0 &= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2) \\
&= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)(Y - \hat{\beta}_1 X_1 + \hat{\beta}_1 \hat{\gamma} X_2) \\
&= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)\big(Y - \hat{\beta}_1 (X_1 - \hat{\gamma} X_2)\big).
\end{aligned}
$$

The Linear Model: Decomposing the Projection

$$\begin{aligned} 0 &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\ 0 &= \mathbb{E}_n\big(X_2(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\ 0 &= \mathbb{E}_n\big(X_2(X_1 - \hat{\gamma} X_2)\big) \end{aligned}$$

Combine the first two, then use the third one to find

$$\begin{aligned} 0 &= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2) \\ &= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)(Y - \hat{\beta}_1 X_1 + \hat{\beta}_1 \hat{\gamma} X_2) \\ &= \mathbb{E}_n(X_1 - \hat{\gamma} X_2)\big(Y - \hat{\beta}_1 (X_1 - \hat{\gamma} X_2)\big). \end{aligned}$$

The last line is the FOC from regressing Y on the "X_1 on X_2" residuals.

Thus, $\hat{\beta}_1$ is the slope coefficient from that regression.

Can extend the argument to show that Y may be residualized as well.

Omitted Variable Bias

In scalar case, we can similarly characterize projection of X_2 on X_1 by

$$\begin{aligned} 0 &= \mathbb{E}_n\big(X_1(X_2 - \tilde{\gamma}X_1)\big) \\ &= \mathbb{E}_n\big(X_1(\hat{\beta}_2 X_2 - \hat{\beta}_2\tilde{\gamma}X_1)\big). \end{aligned}$$

We first recall and then substitute into the first FOC from preceding slide

$$\begin{aligned} 0 &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\ &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2\tilde{\gamma}X_1)\big), \end{aligned}$$

recovering the FOC from regressing Y on X_1 only.

Omitted Variable Bias

In scalar case, we can similarly characterize projection of X_2 on X_1 by

$$\begin{aligned} 0 &= \mathbb{E}_n\big(X_1(X_2 - \tilde{\gamma}X_1)\big) \\ &= \mathbb{E}_n\big(X_1(\hat{\beta}_2 X_2 - \hat{\beta}_2\tilde{\gamma}X_1)\big). \end{aligned}$$

We first recall and then substitute into the first FOC from preceding slide

$$\begin{aligned} 0 &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)\big) \\ &= \mathbb{E}_n\big(X_1(Y - \hat{\beta}_1 X_1 - \hat{\beta}_2\tilde{\gamma}X_1)\big), \end{aligned}$$

recovering the FOC from regressing Y on X_1 only.

We conclude that the slope coefficient $\tilde{\beta}_1$ from regressing Y on *only* X_1 equals

$$\tilde{\beta}_1 = \hat{\beta}_1 + \tilde{\gamma}\hat{\beta}_2.$$

If a causal interpretation of the projection of Y on (X_1, X_2) is appropriate, then the difference term $\tilde{\gamma}\hat{\beta}_2$ is (the sample analog of) the omitted variable bias incurred by omitting X_2 from the regression.

The Linear Model: Decomposing the Projection

Frisch-Waugh-Lovell: General Statement

We now do the same thing again but more generally. Partition

$$\begin{aligned}
X &= \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \\
\mathbb{E}XX' &\equiv \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1X_1' & \mathbb{E}X_1X_2' \\ \mathbb{E}X_2X_1' & \mathbb{E}X_2X_2' \end{pmatrix} \\
\mathbb{E}XY &\equiv \boldsymbol{Q}_{XY} = \begin{pmatrix} \boldsymbol{Q}_{1Y} \\ \boldsymbol{Q}_{2Y} \end{pmatrix} = \begin{pmatrix} \mathbb{E}X_1Y \\ \mathbb{E}X_2Y \end{pmatrix}
\end{aligned}$$

and use notation $\hat{\boldsymbol{Q}}_{11}$ etc. for sample analogs. Then it can be shown that

$$\hat{\beta} = \begin{pmatrix} (\hat{\boldsymbol{Q}}_{11} - \hat{\boldsymbol{Q}}_{12}\hat{\boldsymbol{Q}}_{22}^{-1}\hat{\boldsymbol{Q}}_{21})^{-1}(\hat{\boldsymbol{Q}}_{1Y} - \hat{\boldsymbol{Q}}_{12}\hat{\boldsymbol{Q}}_{22}^{-1}\hat{\boldsymbol{Q}}_{2Y}) \\ (\hat{\boldsymbol{Q}}_{22} - \hat{\boldsymbol{Q}}_{21}\hat{\boldsymbol{Q}}_{11}^{-1}\hat{\boldsymbol{Q}}_{12})^{-1}(\hat{\boldsymbol{Q}}_{2Y} - \hat{\boldsymbol{Q}}_{21}\hat{\boldsymbol{Q}}_{11}^{-1}\hat{\boldsymbol{Q}}_{1Y}) \end{pmatrix}.$$

The Linear Model: Decomposing the Projection

Regressing X_1 on X_2 would yield

- coefficients $\hat{\gamma} \equiv \hat{\boldsymbol{Q}}_{22}^{-1}\hat{\boldsymbol{Q}}_{21}$
- and residuals $\hat{\eta} \equiv X_1 - X_2\hat{\boldsymbol{Q}}_{22}^{-1}\hat{\boldsymbol{Q}}_{21}$.

The Linear Model: Decomposing the Projection

Regressing X_1 on X_2 would yield

- coefficients $\hat{\gamma} \equiv \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}$
- and residuals $\hat{\eta} \equiv X_1 - X_2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}$.

Hence,

$$\begin{aligned} \mathbb{E}_n \hat{\eta}^2 &= \mathbb{E}_n X_1^2 + \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \mathbb{E}_n X_2^2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21} - 2 \mathbb{E}_n X_1 X_2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21} \\ &= \hat{\boldsymbol{Q}}_{11} - \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}. \end{aligned}$$

The Linear Model: Decomposing the Projection

Regressing X_1 on X_2 would yield

- coefficients $\hat{\gamma} \equiv \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}$
- and residuals $\hat{\eta} \equiv X_1 - X_2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}$.

Hence,

$$\begin{aligned} \mathbb{E}_n \hat{\eta}^2 &= \mathbb{E}_n X_1^2 + \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \mathbb{E}_n X_2^2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21} - 2 \mathbb{E}_n X_1 X_2 \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21} \\ &= \hat{\boldsymbol{Q}}_{11} - \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21}. \end{aligned}$$

By similar algebra,

$$\mathbb{E}_n \hat{\eta} Y = \hat{\boldsymbol{Q}}_{1Y} - \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{2Y}.$$

It follows that if we projected Y on the residual from regressing X_1 on X_2, we'd get coefficient $\tilde{\beta}$, where

$$\tilde{\beta}_1 = \left(\hat{\boldsymbol{Q}}_{11} - \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{21} \right)^{-1} \left(\hat{\boldsymbol{Q}}_{1Y} - \hat{\boldsymbol{Q}}_{12} \hat{\boldsymbol{Q}}_{22}^{-1} \hat{\boldsymbol{Q}}_{2Y} \right) = \hat{\beta}_1.$$

The Linear Model: Decomposing the Projection

Interpretation

- Verbally, the multivariate OLS coefficient on X_1 is the coefficient one would get by regressing Y on the residual from regressing X_1 on all other covariates.
- Can show: The statement remains true if we also replace Y with the residual $Y - \mathcal{P}_{X_2}(Y)$.
- May look like a curiosity now, but is an important starting point for, e.g., partially linear models.
- An immediate payoff is that if you recall it, you'll never again forget the own-variance formula for multivariate regression coefficients (coming up in a few slides).

Interpreting the Linear Model

There are two ways to state/interpret OLS. Neither of them is uniquely "right," but you should always be clear about which one you are appealing to.

Best Linear Prediction

- Is an interpretation that "makes sense" under extremely general conditions.
- Is the notion of linear model that is generalized in most predictive (notably, data science/statistical learning) applications.

Causal (or Structural) Linear Model

- Is more demanding but allows for causal interpretation.
- Is the notion of linear model that is generalized in most causal (e.g., Instrumental Variables et al.) applications.

NB: This does not preclude predictive application of linear models as a component of causal inference. A salient example is the "first stage" in IV regression.

Interpreting the Linear Model

1. Best Linear Prediction

Write $Y = m(X) + \varepsilon$, where $m(x) \equiv \mathbb{E}(Y \mid X = x)$.

That $\mathbb{E}(\varepsilon \mid X) = 0$ is then a tautology.

Can show: $b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}(XY)$, if it exists, minimizes $\mathbb{E}(Y - Xb)^2$.

That is, $\hat{Y} \equiv \mathcal{P}_X Y \equiv X'b^*$ is the Best Linear Predictor Under Square Loss.

Furthermore, under those conditions, $\mathbb{E}Xe = 0$, where $e \equiv Y - \mathcal{P}_X(Y)$.
That is, the projection error e is not correlated with X.

Interpreting the Linear Model

1. Best Linear Prediction

Write $Y = m(X) + \varepsilon$, where $m(x) \equiv \mathbb{E}(Y \mid X = x)$.

That $\mathbb{E}(\varepsilon \mid X) = 0$ is then a tautology.

Can show: $b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}(XY)$, if it exists, minimizes $\mathbb{E}(Y - Xb)^2$.

That is, $\hat{Y} \equiv \mathcal{P}_X Y \equiv X'b^*$ is the Best Linear Predictor Under Square Loss.

Furthermore, under those conditions, $\mathbb{E}Xe = 0$, where $e \equiv Y - \mathcal{P}_X(Y)$.
That is, the projection error e is not correlated with X.

Theorem:
If a WLLN applies to both $\frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ and $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$ and $\mathbb{E}XX'$ is nonsingular, then b^* is uniquely defined and

$$\hat{\beta} \xrightarrow{p} b^*.$$

OLS estimates the population linear projection under weak assumptions.

Interpreting the Linear Model

1. Best Linear Prediction

Write $Y = m(X) + \varepsilon$, where $m(x) \equiv \mathbb{E}(Y \mid X = x)$.

That $\mathbb{E}(\varepsilon \mid X) = 0$ is then a tautology.

Can show: $b^* \equiv (\mathbb{E}XX')^{-1}\mathbb{E}(XY)$, if it exists, minimizes $\mathbb{E}(Y - Xb)^2$.

That is, $\hat{Y} \equiv \mathcal{P}_X Y \equiv X'b^*$ is the Best Linear Predictor Under Square Loss.

Furthermore, under those conditions, $\mathbb{E}Xe = 0$, where $e \equiv Y - \mathcal{P}_X(Y)$.
That is, the projection error e is not correlated with X.

Theorem:
If a WLLN applies to both $\frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ and $\frac{1}{n}\sum_{i=1}^{n} X_i Y_i$ and $\mathbb{E}XX'$ is nonsingular, then b^* is uniquely defined and

$$\hat{\beta} \xrightarrow{p} b^*.$$

Fact: The Best Linear Predictor $\hat{Y}$ is uniquely defined even if $\mathbb{E}XX'$ is singular.
It's just that the coefficient b^* is then not unique.

This will be important for prediction from high-dimensional covariates (i.e., statistical learning/"big data" methods).

Interpreting the Linear Model

2. The (Causal/Structural) Linear Model

Write $Y = X'\beta + \varepsilon$, where $\mathbb{E}(\varepsilon \mid X) = 0$.

Equivalently, $m(x) = \mathbb{E}(Y \mid X = x) = x'\beta$.

In this version, $\mathbb{E}(\varepsilon \mid X) = 0$ is **not** tautological!

The assumptions therefore became much stronger.

The benefits are:

- This model allows for causal interpretation of the estimand:
  In expectation, a change ΔX causes a corresponding change $\Delta X'\beta$ in Y.
- But the difference is not just about interpretation:
  Some important results are only available under the stronger assumption.

Interpreting the Linear Model

2. The (Causal/Structural) Linear Model (ctd.)

Reminder: Within limits, a linear statistical model can capture nonlinear substantive models.

Examples:

- Polynomial expansion (and other series):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \varepsilon.$$

- Log or Log-log-regression:

$$\log Y = \log A + \alpha \log K + (1 - \alpha) \log L + \varepsilon$$

  (but note that taking logs changes the necessary assumption on ε!).

- Treatment effects with interactions:

$$Y = \beta + \delta \,\text{treatment} + \gamma \,\text{female} \cdot \text{treatment} + \cdots + \varepsilon.$$

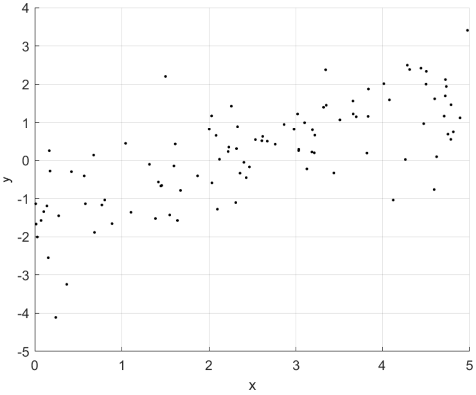
Indeed, many "big data" models are high-dimensional but linear.

The OLS Estimator as a Random Variable

What can we say about this random variable?

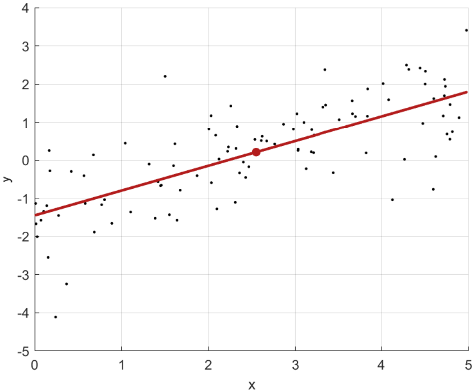
Are there conditions under which it has desirable properties, notably if our objective is to learn about β (or possibly the population $\mathcal{P}_X(Y)$)?

OLS: The Estimator as a Random Variable



As a brief reminder, here is one possible sample...

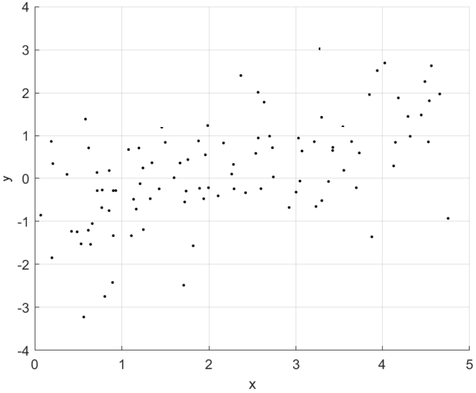
OLS: The Estimator as a Random Variable



As a brief reminder, here is one possible sample...

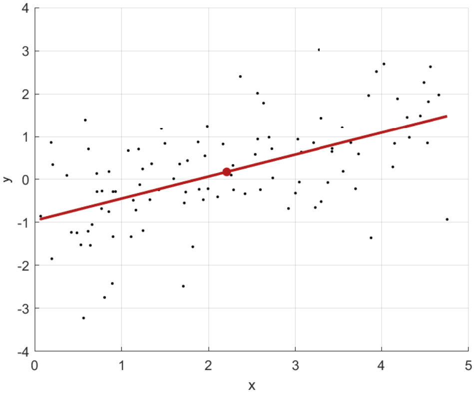
...with the OLS fit overlaid.

OLS: The Estimator as a Random Variable



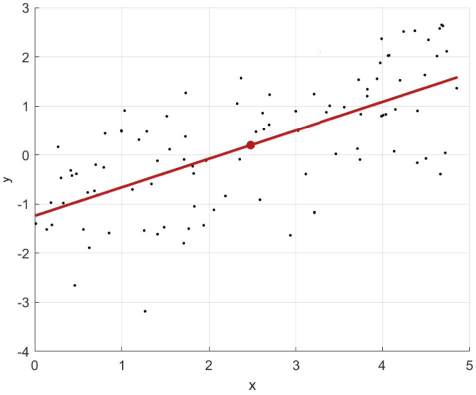
Consider doing the same thing on a new sample...

OLS: The Estimator as a Random Variable



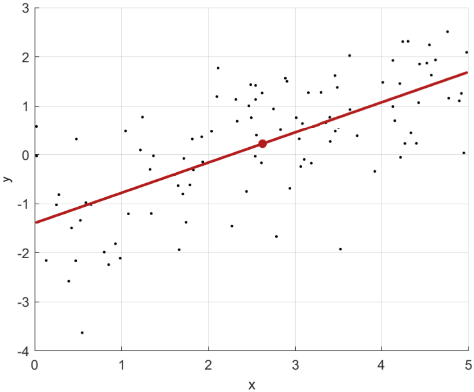
Consider doing the same thing on a new sample...
...and finding the OLS fit again...

OLS: The Estimator as a Random Variable



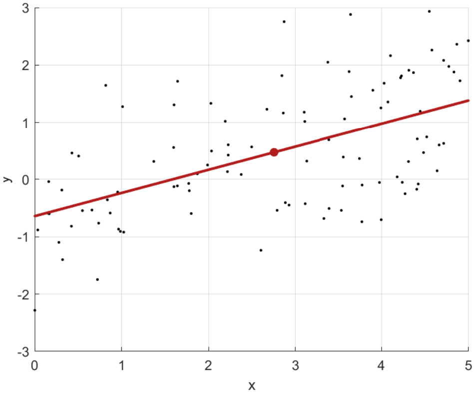
...and again...

OLS: The Estimator as a Random Variable



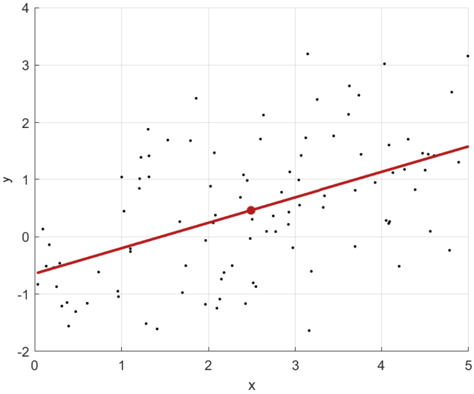
...and again...

OLS: The Estimator as a Random Variable



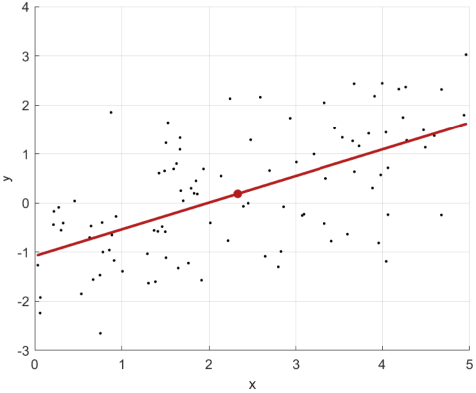
...and again...

OLS: The Estimator as a Random Variable



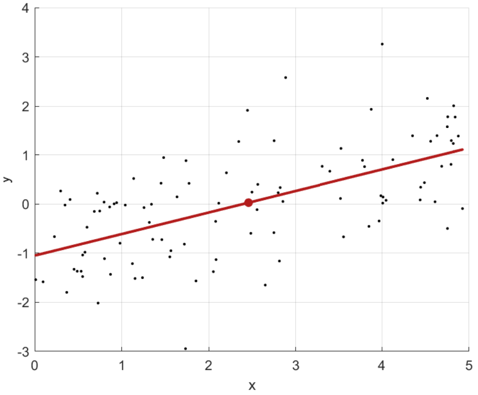
...and again...

OLS: The Estimator as a Random Variable



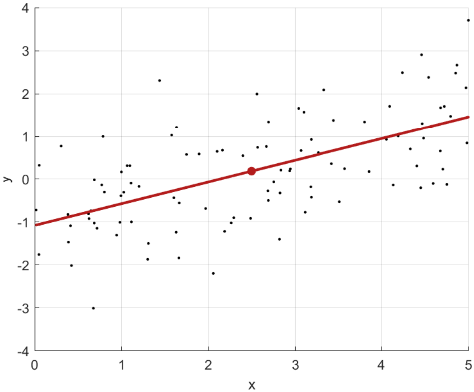
...and again...

OLS: The Estimator as a Random Variable



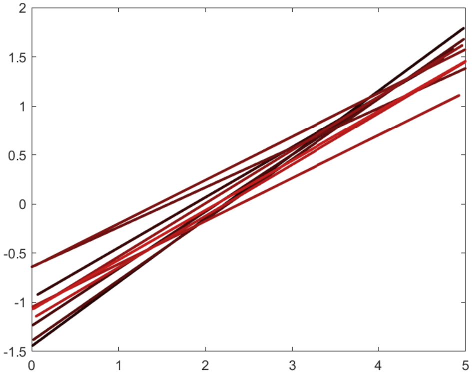
...and again...

OLS: The Estimator as a Random Variable



...and again...

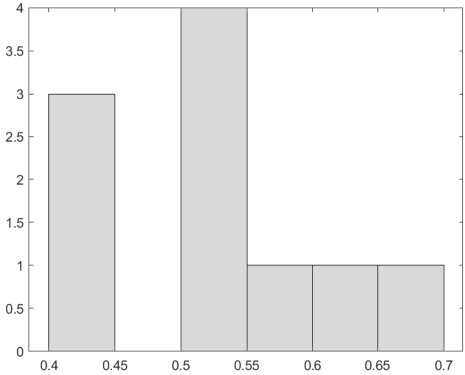
OLS: The Estimator as a Random Variable



Here are the 10 fitted lines again.

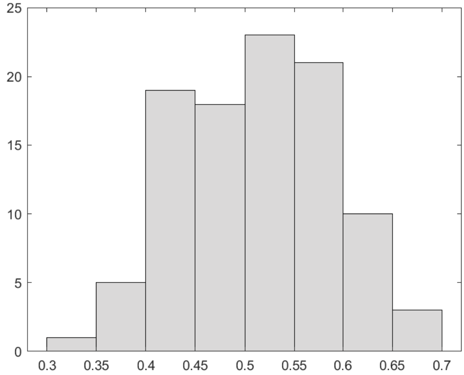
Of course, we have better ways to summarize this.

OLS: The Estimator as a Random Variable



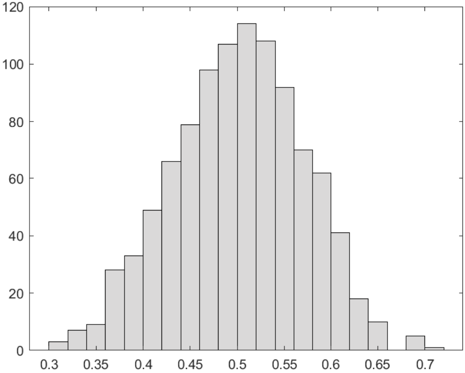
This is a histogram of the 10 realizations of estimated slope coefficient $\hat{\beta}_1$ from the preceding slides.

OLS: The Estimator as a Random Variable



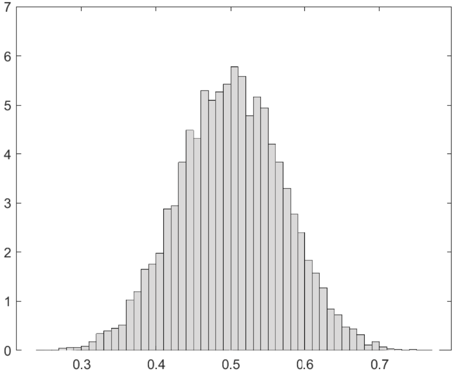
As we increase the number of realizations to 100,...

OLS: The Estimator as a Random Variable



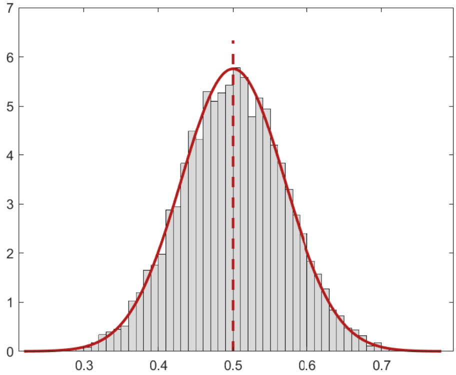
As we increase the number of realizations to 100, 1000,...

OLS: The Estimator as a Random Variable



As we increase the number of realizations to 100, 1000, 10000,...

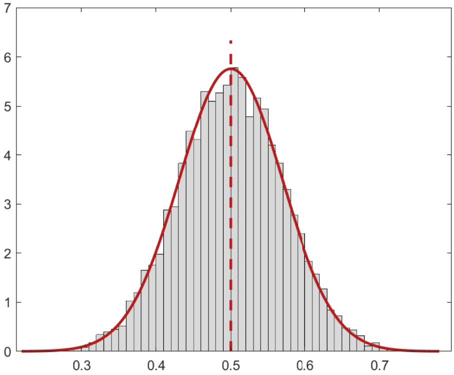
OLS: The Estimator as a Random Variable



As we increase the number of realizations to 100, 1000, 10000,...

...a picture emerges.

OLS: The Estimator as a Random Variable



(Clarifying note: I here sent number of Monte Carlos to ∞, not n. The normality reflects that error terms were normal in my fake data, not a CLT. In words, I illustrated the exact behavior of OLS under the strongest assumptions we will see, though you are probably aware that it is also the approximate behavior of OLS for large n under weaker assumptions.)

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$ ("strong exogeneity").
3. $\text{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition").
   Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

The Linear Model: Finite-Sample Theory

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{0}$ ("strong exogeneity").
3. $\operatorname{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition"). Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

- The first two assumptions together imply a causal linear model. Indeed, assumptions that are natural in a "Best Linear Predictor" interpretation do not suffice to claim unbiasedness of $\hat{\beta}$ (even for b^*).
- Assumptions further imply that ε is mean independent of "past and future" covariates, which is restrictive and not essential for causal interpretation.
- The second assumption is implied by $\mathbb{E}(\varepsilon \mid X) = 0$ (i.e., conditioning on the contemporaneous X only) if the data are assumed to be i.i.d.

The Linear Model: Finite-Sample Theory

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$ ("strong exogeneity").
3. $\text{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition"). Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

- Assumption 3 is an identification condition. Intuitively, if it fails, we cannot disentangle the effect of some components of X.
- The assumption can be verified in a given sample, and algebra will assume that $\boldsymbol{X}$ fulfils it.
- If it fails, there is a set of observationally equivalent "true" coefficients that form a linear subspace of $\boldsymbol{R}^K$ and all induce the same $\hat{\boldsymbol{Y}}$.
- Indeed, conditions needed to successfully (in a sense to be defined) approximate $\hat{\boldsymbol{Y}}$ are weaker than the rank condition. This is relevant for high-dimensional extensions of OLS.

The Linear Model: Finite-Sample Theory

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$ ("strong exogeneity").
3. $\operatorname{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition"). Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

- Assumption 4 combines conditional uncorrelatedness and homoskedasticity of errors. The latter only makes sense in the causal model because, even if the true regression error $\varepsilon = Y - m(X)$ is homoskedastic, the projection error

$$e = Y - \mathcal{P}_X(Y) = Y - m(X) + m(X) - \mathcal{P}_X(Y) = \varepsilon + m(X) - \mathcal{P}_X(Y)$$

is not.

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$ ("strong exogeneity").
3. $\text{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition"). Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

- There is a hidden assumption that $\mathbb{E}||\boldsymbol{X}||^2 < \infty$. Without it, $\mathbb{E}\hat{\beta}$ may not exist and $\mathbb{E}(\hat{\beta} \mid \boldsymbol{X})$ then not be well-defined in the sense of measure theoretic probability.
- If we think of $\boldsymbol{X}$ as nonstochastic, all results claimed in this section hold without the hidden assumption. This corresponds to the historic development of OLS.
- I will be cavalier about such hidden assumptions in this lecture, but this is why you may see assumptions like "all r.v.'s have moments as needed."

The Linear Model: Finite-Sample Theory

Assumptions

1. $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ ("linearity").
2. $\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \mathbf{0}$ ("strong exogeneity").
3. $\text{rank}(\boldsymbol{X}) = K$ a.s., where $\boldsymbol{X} \in \boldsymbol{R}^{n \times K}$ ("rank condition"). Equivalently, $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular a.s.
4. $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n$ ("spherical error").

Theorem

Under the above assumptions, we have:

1. $\mathbb{E}(\hat{\beta} \mid \boldsymbol{X}) = \beta$ ("$\hat{\beta}$ is unbiased").
2. $\text{var}(\hat{\beta} \mid \boldsymbol{X}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
3. If an estimator $\tilde{\beta}$ is linear (in $\boldsymbol{Y}$) and unbiased, then $\text{var}(\tilde{\beta} \mid \boldsymbol{X}) \geq \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ ("$\hat{\beta}$ is BLUE / Gauss-Markov Theorem").

Proof: Bias and Variance

We first observe that

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \varepsilon) \\ &= \beta + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon}_{=\text{estimation error}}.\end{aligned}$$

Proof: Bias and Variance

We first observe that

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \varepsilon) \\
&= \beta + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon}_{=\text{estimation error}}.
\end{aligned}$$

Therefore, the first two claims really are that

$$\begin{aligned}
\mathbb{E}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) &= \mathbf{0}, \\
\text{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Proof: Bias and Variance

We first observe that

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \varepsilon) \\
&= \beta + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon}_{=\text{estimation error}}.
\end{aligned}$$

Therefore, the first two claims really are that

$$\begin{aligned}
\mathbb{E}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) &= \mathbf{0}, \\
\operatorname{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Indeed we have

$$\mathbb{E}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\varepsilon \mid \boldsymbol{X}) = 0,$$

$$\operatorname{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\varepsilon \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Proof: Gauss-Markov

We assume that $\tilde{\beta}$ is linear in Y, i.e. $\tilde{\beta} = \boldsymbol{C}\boldsymbol{Y}$, where the matrix $\boldsymbol{C}$ may depend on $\boldsymbol{X}$. Define $\boldsymbol{D} \equiv \boldsymbol{C} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and write

$$\begin{aligned}
\beta &= \mathbb{E}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})\boldsymbol{Y} \mid \boldsymbol{X}\right) \\
&= \beta + \mathbb{E}\left(\boldsymbol{D}\boldsymbol{Y} \mid \boldsymbol{X}\right) \\
&= \beta + \mathbb{E}\left(\boldsymbol{D}(\boldsymbol{X}\beta + \varepsilon) \mid \boldsymbol{X}\right) \\
&= \beta + \mathbb{E}(\boldsymbol{D}\boldsymbol{X}\beta \mid \boldsymbol{X}) + \boldsymbol{D}\underbrace{\mathbb{E}(\varepsilon \mid \boldsymbol{X})}_{=\mathbf{0}},
\end{aligned}$$

therefore we have the identity $\mathbb{E}(\boldsymbol{D}\boldsymbol{X}\beta \mid \boldsymbol{X}) = \mathbf{0}$ irrespective of the value taken by β; in addition, conditionally on $\boldsymbol{X}$ the expression $\boldsymbol{D}\boldsymbol{X}\beta$ is not stochastic. This is only possible if $\boldsymbol{D}\boldsymbol{X} = \mathbf{0}$.

Proof: Gauss-Markov

To wrap up, write

$$\begin{aligned}
& \text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})\boldsymbol{Y} \mid \boldsymbol{X}\right) \\
= \; & \text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})(\boldsymbol{X}\beta_0 + \varepsilon) \mid \boldsymbol{X}\right) \\
= \; & \text{var}\left(((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})\varepsilon \mid \boldsymbol{X}\right) \\
= \; & \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{D})' \\
= \; & \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{D}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{D}' + \boldsymbol{D}\boldsymbol{D}') \\
\geq \; & \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}$$

where the last step used cancellation, the fact that $\boldsymbol{D}\boldsymbol{X} = \boldsymbol{0}$, and the fact that $\boldsymbol{D}\boldsymbol{D}'$ is positive semidefinite.

Is homoskedasticity necessary for this result?

The Linear Model: Finite-Sample Theory

Is homoskedasticity necessary for this result?

Yes.

Consider the case where $\mathbb{E}\varepsilon\varepsilon' = \mathbf{\Omega}$ is known and diagonal but its diagonal entries are not the same.

(Think of heteroskedasticity where the variance of ε is a known function of X, considering that our analysis conditions on $\boldsymbol{X}$.)

Then the Gauss-Markov assumptions apply to the transformed model

$$\mathbf{\Omega}^{-1/2}\boldsymbol{Y} = \mathbf{\Omega}^{-1/2}\boldsymbol{X} + \mathbf{\Omega}^{-1/2}\boldsymbol{\varepsilon}$$

and therefore the estimator

$$\hat{\beta}_{WLS} = (\boldsymbol{X}'\mathbf{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{\Omega}^{-1}\boldsymbol{Y}$$

is BLUE. But $\hat{\beta}_{WLS} = \hat{\beta}$ only if $\mathbf{\Omega} = \sigma^2\boldsymbol{I}_n$ for some $\sigma^2 > 0$.

Indeed, this is the Weighted Least Squares estimator. It can be equivalently derived by reweighting observations in the Least Squares objective function.

Some important closed-form expressions

Consider simple linear regression: $Y = \alpha + \beta X + \varepsilon$.

Then

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} \\ \mathrm{var}(\hat{\beta} \mid \boldsymbol{X}) &= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.\end{aligned}$$

Some important closed-form expressions

Consider simple linear regression: $Y = \alpha + \beta X + \varepsilon$.

Then

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} \\
\operatorname{var}(\hat{\beta} \mid \boldsymbol{X}) &= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.
\end{aligned}$$

In multivariate regression, the closed-form variance of the k'th component of $\hat{\beta}$ is

$$\operatorname{var}(\hat{\beta}_k \mid \boldsymbol{X}) = \frac{\sigma^2}{(1 - R_k^2)\sum_{i=1}^{n}(X_{ki} - \overline{X}_k)^2},$$

where R_k^2 is the R^2 from the regression of X_k on the other covariates.

Why is this obvious?

Some important closed-form expressions

Consider simple linear regression: $Y = \alpha + \beta X + \varepsilon$.

Then

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})X_i} \\ \text{var}(\hat{\beta} \mid \boldsymbol{X}) &= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.\end{aligned}$$

In multivariate regression, the closed-form variance of the k'th component of $\hat{\beta}$ is

$$\text{var}(\hat{\beta}_k \mid \boldsymbol{X}) = \frac{\sigma^2}{(1 - R_k^2)\sum_{i=1}^{n}(X_{ki} - \overline{X}_k)^2},$$

where R_k^2 is the R^2 from the regression of X_k on the other covariates.

Why is this obvious?

- Think back to Frisch-Waugh-Lovell.
- Aside: $1/(1 - R_k^2)$ is sometimes called "variance inflation factor" (VIF).

Estimating the Variance

The sample analog and also method-of-moments estimator of σ^2 is

$$\hat{\sigma}^2 \equiv \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Estimating the Variance

The sample analog and also method-of-moments estimator of σ^2 is

$$\hat{\sigma}^2 \equiv \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

However, it can be shown that $\mathbb{E}(\hat{\sigma}^2 \mid \boldsymbol{X}) = \frac{n-K}{n} \cdot \sigma^2$.

(Heuristically, the r.v. $\hat{\varepsilon}$ has only $(n-K)$ degrees of freedom because its sample space is constrained by K equations $\boldsymbol{X}'\hat{\varepsilon} = \mathbf{0}$.)

Estimating the Variance

The sample analog and also method-of-moments estimator of σ^2 is

$$\hat{\sigma}^2 \equiv \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

However, it can be shown that $\mathbb{E}(\hat{\sigma}^2 \mid \boldsymbol{X}) = \frac{n-K}{n} \cdot \sigma^2$.
(Heuristically, the r.v. $\hat{\varepsilon}$ has only $(n-K)$ degrees of freedom because its sample space is constrained by K equations $\boldsymbol{X}'\hat{\varepsilon} = \mathbf{0}$.)

Therefore, it is more common to use the unbiased

$$s^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Estimating the Variance (ctd.)

Estimators of standard deviations will become important.

We will call them standard errors.

- $\sqrt{s^2}$ is the standard error of the regression.
- $SE(\hat{\beta}) \equiv \left(s^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{kk}\right)^{1/2}$ is the standard error of $\hat{\beta}_k$.

Estimating the Variance (ctd.)

Estimators of standard deviations will become important.

We will call them standard errors.

- $\sqrt{s^2}$ is the standard error of the regression.
- $SE(\hat{\beta}) \equiv \left(s^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{kk}\right)^{1/2}$ is the standard error of $\hat{\beta}_k$.

Disclaimers:

- We **cannot** claim that these estimators are unbiased.
- Beware: The usage "standard error" for estimated standard deviation is dominant in econometrics (cf. the Goldberger, Hansen, Hayashi, Stock/Watson, and Wooldridge textbooks) but not in other fields, where "standard error" and "[sampling] standard deviation" might be synonyms. In that case, the above are *estimated* standard errors (cf. Imbens/Rubin and numerous statistics textbooks).

Heteroskedasticity

The spherical error assumption was only (fully) used for the variance expressions.

Recall algebra:

$$\begin{aligned} \hat{\beta} &= \beta + \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}}_{=\text{estimation error}} \\ \Longrightarrow \operatorname{var}(\hat{\beta} \mid \boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X})}_{\equiv \boldsymbol{D}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

Spherical error ($\boldsymbol{D} = \sigma^2 \boldsymbol{I}_n$) leads to simplification, but as long as we can estimate $\boldsymbol{D}$, we are good either way.

Recall (with slight rewriting)

$$\mathrm{var}(\hat{\beta} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Assuming "only" heteroskedasicity, i.e. maintaining that $\boldsymbol{D}$ is diagonal, we have:

- The "oracle estimator"

  $$\widehat{\mathrm{var}}_{\mathrm{oracle}}(\hat{\beta} \mid \boldsymbol{X}) \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  is unbiased but is not available.

The Linear Model: Finite-Sample Theory

Recall (with slight rewriting)

$$\mathrm{var}(\hat{\beta} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Assuming "only" heteroskedasicity, i.e. maintaining that $\boldsymbol{D}$ is diagonal, we have:

- The "oracle estimator"

  $$\widehat{\mathrm{var}}_{\mathrm{oracle}}(\hat{\beta} \mid \boldsymbol{X}) \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  is unbiased but is not available.
- Plugging in $\hat{\varepsilon}_i$ leads to plausible estimator

  $$\widehat{\mathrm{var}}_{HC0}(\hat{\beta}) \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

The Linear Model: Finite-Sample Theory

Recall (with slight rewriting)

$$\text{var}(\hat{\beta} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Assuming "only" heteroskedasicity, i.e. maintaining that $\boldsymbol{D}$ is diagonal, we have:

- The "oracle estimator"

$$\widehat{\text{var}}_{\text{oracle}}(\hat{\beta} \mid \boldsymbol{X}) \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \varepsilon_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

  is unbiased but is not available.
- Plugging in $\hat{\varepsilon}_i$ leads to plausible estimator

$$\widehat{\text{var}}_{HC0}(\hat{\beta}) \equiv (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

- This estimator is biased, motivating ad hoc "d.f. adjustment"

$$\widehat{\text{var}}_{HC1}(\hat{\beta}) \equiv \frac{n}{n-K} (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{\varepsilon}_i^2 \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Some informal remarks on dropping spherical error.

- HC0 is the "original" (Eicker-White) heteroskedasticity robust variance estimate.
- HC1 is the "industry standard," e.g. it is the STATA default.
- Neither is obviously best. See Hansen's textbook for other options.
- It is dominant applied practice to use these variance estimates/standard errors because homoskedasticity is rarely considered plausible.
- At the same time, though the idea of WLS can be adapted to pre-estimating heteroskedasticity ("Feasible Generalized Least Squares"), this is *not* common in applied practice.
- While we will omit it for now, the topic of clustered standard errors takes the same point of departure.

Exact distribution and hypothesis tests under Normality

Next, we also impose that ε has a normal distribution:

$$(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n).$$

Here, only the $N(\cdot)$ part is really new, the parameters of the distribution then follow from previous assumptions. Then we have:

Theorem

Define $s^2 \equiv \frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{n-K}$ and let the matrix $\underset{r\times K}{\boldsymbol{R}}$ have maximal rank $r \leq K$.
Under the previous assumptions and normality, we then have:

$$(\hat{\beta} - \beta) \mid \boldsymbol{X} \sim N(\mathbf{0}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1})$$

$$t\text{-ratio} = t \quad \equiv \quad \frac{\hat{\beta}_k - \beta_k}{\left(s^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{kk}\right)^{1/2}} \sim t_{n-K}$$

$$F\text{-statistic} = F \quad \equiv \quad \frac{\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{R}\beta\right)' \left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1} \left(\boldsymbol{R}\hat{\beta} - \boldsymbol{R}\beta\right)}{s^2 \ r} \sim F_{r,n-K},$$

Theorem

Define $s^2 \equiv \frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\beta})}{n-K}$ and let the matrix $\underset{r \times K}{\boldsymbol{R}}$ have maximal rank $r \leq K$. Under the previous assumptions and normality, we then have:

$$
\begin{aligned}
(\hat{\beta}-\beta) \mid \boldsymbol{X} &\sim N(\mathbf{0}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
t\text{-ratio} = t \quad &\equiv \quad \frac{\hat{\beta}_k - \beta_k}{\left(s^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{kk}\right)^{1/2}} \sim t_{n-K} \\
F\text{-statistic} = F \quad &\equiv \quad \frac{\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{R}\beta\right)' \left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1} \left(\boldsymbol{R}\hat{\beta} - \boldsymbol{R}\beta\right)}{s^2 \; r} \sim F_{r,n-K},
\end{aligned}
$$

Thus, the null hypothesis $H_0 : \boldsymbol{R}\beta = \boldsymbol{r}$ can be tested with exact size control by comparing

$$
\frac{\left(\boldsymbol{R}\hat{\beta} - \boldsymbol{r}\right)' \left(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right)^{-1} \left(\boldsymbol{R}\hat{\beta} - \boldsymbol{r}\right)}{s^2 \; r}
$$

to the relevant quantile of $F_{r,n-K}$ etc.

The Linear Model: Finite-Sample Theory

You saw a proof of this result before.

For a quick intuition regarding the t-statistic, recall that

$$(\hat{\beta} - \beta) \mid \boldsymbol{X} \sim N\big(\mathbf{0}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\big)$$

implies

$$\frac{\hat{\beta}_k - \beta_k}{\left(\sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{kk}\right)^{1/2}} \sim N(0, 1).$$

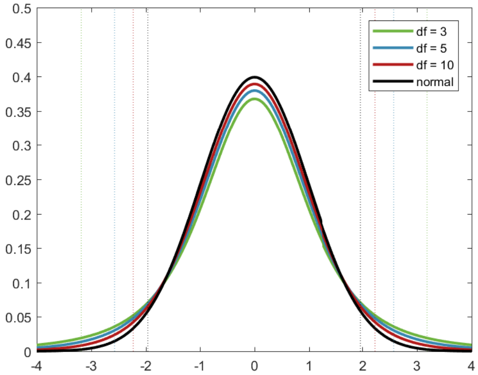
The difference is that the estimator s^2 stands in for σ^2.

It can be shown that (under current assumptions)

$$(n - K)\frac{s^2}{\sigma^2} \mid \hat{\beta} \sim \chi^2_{n-K}.$$

The result then follows from how the t-distribution is defined.

The Linear Model: Finite-Sample Theory



The t-distribution for different (small) degrees of freedom. We see:

- rapid convergence to standard normal, but also...
- considerable difference in critical values if degrees of freedom are very small.