Linear Regression with Exogenous 'Treatment'

Erich Battistin
University of Maryland, CEPR, FBK-IRVAPP and IZA

AREC 623
Applied Econometrics I

Population Regression Functions

Population Regression and Bivariate Basics

- Recall that any random variable Y can be written as a piece that is explained by X (the CEF) and a piece left over (ε) that is uncorrelated with any function of X:

$$Y = E(Y|X = x) + \varepsilon.$$

- **Regression functions** are tightly related to the CEF, and defined as the solution to the population **best linear prediction** problem:

$$\underset{b_0, b_1}{\operatorname{argmin}} E\left[(Y - b_0 - b_1 X)^2\right].$$

- Using the first-order conditions, the following two equations in two unknowns are defined:

$$E[Y - \beta_0 - \beta_1 X] = 0,$$
$$E[X(Y - \beta_0 - \beta_1 X)] = 0,$$

where β_0 and β_1 are the solutions.

- By arranging terms we get:

$$\begin{aligned}\beta_0 &= E[Y] - \beta_1 E[X], \\ \beta_1 &= \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} = \frac{Cov[X,Y]}{Var[X]}.\end{aligned}$$

- Moment conditions define a correspondence that holds in the population and nails down a relationship between the observables (X, Y) and the unknown parameters β_0 and β_1.
- The above relationship is equivalent to:

$$Y = \beta_0 + \beta_1 X + e,$$

  where because of the first-order conditions the population residual, $e \equiv Y - \beta_0 - \beta_1 X$, is uncorrelated with X and centered at zero.
- This error term has no life of its own: e owes its existence to, and derives its meaning from, β_0 and β_1. It doesn't follow from any assumptions about an underlying economic relationship.

Population Regression and Bivariate Basics

- Just as the CEF is the best predictor of Y in the class of all functions of X (see the previous set of slides), the regression function is the best we can do in the class of linear functions.
- The regression function also provides the best linear approximation to the CEF, because the solution to:

$$\underset{b_0, b_1}{\operatorname{argmin}} E\left[\left(E\left(Y|X=x\right) - b_0 - b_1 X\right)^2\right],$$

  yields the same values β_0 and β_1.
- This result follows from LIE, which implies:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y], \\ &= E\left[E\left(Y|X=x\right)X\right] - E[X]E\left[E\left(Y|X=x\right)\right], \\ &= Cov\left[E\left(Y|X=x\right), X\right]. \end{aligned}$$

- Regression always makes sense: regression always fits or approximates the CEF regardless of the nature of the variable Y (e.g., employment, earnings, number of patents).

Population Regression for Dummies

- If the CEF is linear, then the population regression function must be it (e.g., you want to predict a line with a line).
- When X takes on two values (**dummy variable**), the CEF also takes on two values and the population regression function fits it perfectly:

$$\begin{aligned} E(Y|X=0) &= \beta_0, \\ E(Y|X=1) &= \beta_0 + \beta_1. \end{aligned}$$

- This also implies:

$$Y = \underbrace{E(Y|X=0)}_{\beta_0} + \left[\underbrace{E(Y|X=1) - E(Y|X=0)}_{\beta_1}\right] X + e.$$

- The dummy variable representation is convenient any time we want to think of the effects of a binary "treatment" X on the outcome Y.

Population Regression for Dummies

- Let (Y_1, Y_0) be the potential outcomes from having $X = 1$ ("treatment") and $X = 0$ ("control"), respectively. We can write:

$$\begin{aligned} Y &= E(Y_0|X=0) + [E(Y_1|X=1) - E(Y_0|X=0)]X + e, \\ &= E(Y_0|X=0) + [E(Y_1 - Y_0|X=1)]X \\ &+ [E(Y_0|X=1) - E(Y_0|X=0)]X + e. \end{aligned}$$

- It follows that:

$$\begin{aligned} \beta_1 &= E(Y|X=1) - E(Y|X=0), \\ &= \underbrace{E(Y_1 - Y_0|X=1)}_{\text{treatment effect}} + \underbrace{[E(Y_0|X=1) - E(Y_0|X=0)]}_{\text{selection bias}}, \end{aligned}$$

  because $E(e|X=1) = E(e|X=0) = 0$ (see next slide).

- Exploring regressions in the data will in general yield statistical patterns: β_1 (correlation) does not reveal $E(Y_1 - Y_0|X=1)$ (causation).

A Comment on "Residuals"

- Errors arising from the least squares approximation to the CEF, e, are different from errors defined from the CEF decomposition, ε.
- We can always write:

$$\begin{aligned} Y &= E(Y|X=x) + \varepsilon, \\ &= \beta_0 + \beta_1 X + [E(Y|X=x) - \beta_0 - \beta_1 X] + \varepsilon. \end{aligned}$$

- This expression implies:

$$\underbrace{e}_{\text{regression error}} = \underbrace{[E(Y|X=x) - \beta_0 - \beta_1 X]}_{\text{CEF approximation error}} + \underbrace{\varepsilon.}_{\text{CEF decomposition error}}$$

- It follows that $E(e|X=x) = 0$ only if the CEF is linear, because by construction $E(\varepsilon|X=x) = 0$ (see previous set of slides), but otherwise the first order conditions only imply $E(eX) = 0$.

Multivariate Population Regression

- The CEF defined by conditioning on multiple variables can be approximated by a population regression using the same reasoning:

$$\beta = \underset{b}{\text{argmin}}\, E\left[\left(Y - X'b\right)^2\right],$$

where the following $K \times 1$ vectors are defined:

$$\underset{K\times 1}{X} \equiv (1, X_1, X_2, \ldots, X_{K-1})', \qquad \underset{K\times 1}{\beta} \equiv (\beta_0, \beta_1, \ldots, \beta_{K-1})'.$$

- Using the first-order conditions:

$$\begin{aligned} \underset{K\times 1}{0} &= E\left[X\left(Y - X'\beta\right)\right], \\ &= E[\underset{K\times 1}{XY}] - E[\underset{K\times K}{XX'}]\beta, \end{aligned}$$

which imply (provided that the matrix is invertible):

$$\beta = E[XX']^{-1}E[XY].$$

The Frisch-Waugh-Lovell Theorem

- It turns out that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after **partialling out** all the other variables.
- Consider the long regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e,$$

where by construction the residual e is uncorrelated with X_1 and X_2.

- Consider the auxiliary regression of X_1 on X_2:

$$X_1 = \delta_0 + \delta_1 X_2 + \tilde{X}_1,$$

where by construction the residual $\tilde{X}_1$ is uncorrelated with X_2, and:

$$Cov[\tilde{X}_1, X_1] = Var[\tilde{X}_1].$$

- We have that:

$$\frac{Cov[\tilde{X}_1, Y]}{Var[\tilde{X}_1]} = \beta_1 \overbrace{\frac{Cov[\tilde{X}_1, X_1]}{Var[\tilde{X}_1]}}^{=Var[\tilde{X}_1]} + \beta_2 \overbrace{\frac{Cov[\tilde{X}_1, X_2]}{Var[\tilde{X}_1]}}^{=0} + \overbrace{\frac{Cov[\tilde{X}_1, e]}{Var[\tilde{X}_1]}}^{=0}.$$

The Frisch-Waugh-Lovell Theorem

- This implies that the coefficient on X_1 in the long regression is the coefficient on $\tilde{X}_1$ in the bivariate regression of Y on $\tilde{X}_1$:

$$\beta_1 = \frac{Cov[\tilde{X}_1, Y]}{Var[\tilde{X}_1]}.$$

- The **regression anatomy formula** is very useful in empirical work, as it allows the correlation of interest (e.g., between Y and X_1) to be analyzed conditional on the effect of other variables (e.g., X_2).
- It also shows that identification of β_1 rests upon the variability in X_1 net of X_2, ruling out any multicollinearity problems. This connects to the rank condition for invertibility of the matrix $E[XX']$.
- The flip side of this result is an important formula describing the relationship between regressions with an increasingly large number of variables.

Long vs Short Population Regressions

- Go long, and consider again the regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e,$$

  where by construction the residual e is uncorrelated with X_1 and X_2.

- Leave out X_2, and write:

$$\overbrace{\frac{Cov[X_1, Y]}{Var[X_1]}}^{\text{short}} = \beta_1 \overbrace{\frac{Cov[X_1, X_1]}{Var[X_1]}}^{=Var[X_1]} + \beta_2 \frac{Cov[X_1, X_2]}{Var[X_1]} + \overbrace{\frac{Cov[X_1, e]}{Var[X_1]}}^{=0},$$

$$= \underbrace{\beta_1}_{\text{long}} + \underbrace{\beta_2}_{\text{effect of omitted}} \underbrace{\frac{Cov[X_1, X_2]}{Var[X_1]}}_{\text{omitted on included}}.$$

- It follows that short equals long plus the effect of omitted times the regression of omitted on included.
- Short equals long when omitted and included are uncorrelated.

The Perils of Regression Mechanics

Nonlinear CEF vs Linear (Probability) Approximation

- Let Y be a binary dependent variable:

$$E\left(Y|X_1 = x_1, X_2 = x_2\right) = P\left(Y = 1|X_1 = x_1, X_2 = x_2\right).$$

- The (randomized) treatment X_1 is also binary, and X_2 is an additional covariate. Assume the the true CEF is:

$$E\left(Y|X_1 = x_1, X_2 = x_2\right) = \Phi\left(\tau_0 + \tau_1 x_1 + \tau_2 x_2\right).$$

- The average treatment effect is:

$$\int \left[\Phi\left(\tau_0 + \tau_1 + \tau_2 x_2\right) - \Phi\left(\tau_0 + \tau_2 x_2\right)\right] f_{X_2}\left(x_2\right).$$

- The **linear probability model (LPM)** uses the following approximation to the CEF above:

$$E\left(Y|X_1 = x_1, X_2 = x_2\right) \simeq \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

This probability is linear in (X_1, X_2), hence the name LPM.

Nonlinear CEF vs Linear (Probability) Approximation

- Without X_2, the coefficients β_0 and β_1 of a linear regression are in a one-to-one correspondence with the unknown parameters τ_0 and τ_1.
- The latter is a **saturated model**: there is a parameter for each value of the covariate X_1 so that the regression fits the CEF perfectly.
- This is also true when X_2 is binary and we add the interaction X_1X_2.
- More in general, the error e in the LPM cannot be independent of any regressors because, given (X_1, X_2), one must have:

$$e = 1 - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \quad \textit{or} \quad e = -\beta_0 - \beta_1 x_1 - \beta_2 x_2.$$

- LPM predicts probabilities that may take on impossible values.
  - No regressor can have a normal distribution (or any distribution that extends to plus or minus infinity).
  - Any regressor that can take on a large range of values must have a very small coefficient.

Nonlinear CEF vs Linear (Probability) Approximation

- The usual counter argument is to claim that the LPM is intended to approximate true probabilities.
  - Not always true: think of a straight line approximation to the "S shape" of most distribution functions.
- Examples can be made of non-linear models where the true treatment effect is far from the derivative of a linear approximation (i.e., the coefficient of the LPM) and the LPM yields the wrong sign.

Contamination Bias in Linear Regressions

- Let Y be a dependent variable, X_1 a binary treatment, and X_3 a binary control variable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + e.$$

- Assume that treatment is randomized conditional on X_3 (e.g., stratified RCT).
- We wish to interpret the coefficient β_1 in terms of the causal effects of X_1 on Y. For this, use the potential outcomes (Y_0, Y_1).
- The Frisch-Waugh-Lovell theorem can be used to show that β is a **convex weighted average** of treatment effects:

$$\beta = \vartheta\tau(0) + (1-\vartheta)\tau(1),$$

where:

$$\begin{aligned}\tau(x_3) &\equiv E(Y_1 - Y_0 | X_3 = x_3), \\ \vartheta &\equiv \frac{Var(X_1|X_3=0)P(X_3=0)}{\sum_{x_3=0}^{1} Var(X_1|X_3=x_3)P(X_3=x_3)}.\end{aligned}$$

Contamination Bias in Linear Regressions

- This reasoning fails when an additional treatment arm is included!
- Consider a regression on indicators for two treatment arms (X_1, X_2) (think of Project STAR or a stratified multi-armed RCT):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$

- Through the Frisch-Waugh-Lovell theorem, one uses the residual variability in X_1 netting off X_2 and X_3.
- Unlike before, this residual is not mean-independent of (X_2, X_3).
  - If $X_2 = 1$, X_1 must be zero regardless of the value of X_3 (because they are mutually exclusive).
  - If $X_2 = 0$, the mean of X_1 depends on X_3 unless the treatment assignment is completely random.
- Such **multiple-treatment regression generally fails to identify convex weighted average of heterogeneous treatment effects**.

Two Examples

Class Size Effects and Cheating Teachers

- Standardized tests provide the yardstick by which school quality is most often assessed and compared.
- As testing regimes have proliferated, so has the temptation to cut corners or cheat.
- We will use the 2009 – 2011 student censuses for second and fifth graders in Italy to show how cheating hampers the reliability of standardized assessments for studying the effects of class size.
- Italy is characterized by a sharp North-South divide along many dimensions, which motivates public interventions to improve schools.
- The South is also distinguished by widespread manipulation on standardized tests at primary school. Dishonest grading from local teachers has been found to be the reason for this.
- Moral hazard in honesty is an unwelcome input here, and our journey will document its surprising interaction with class size.

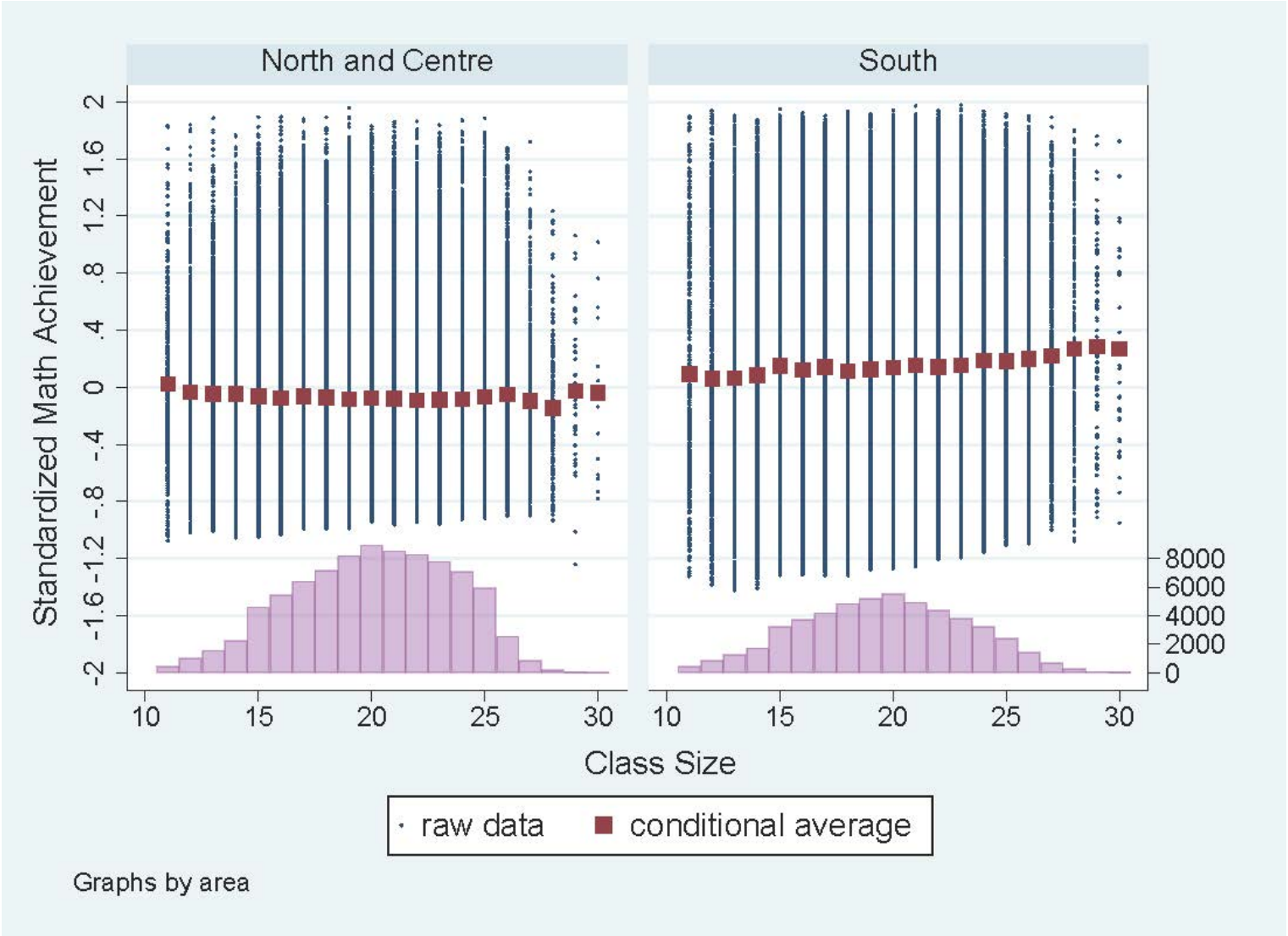
Class Size Effects and Cheating Teachers

- About 2.6 million students, but class is the **statistical unit**:

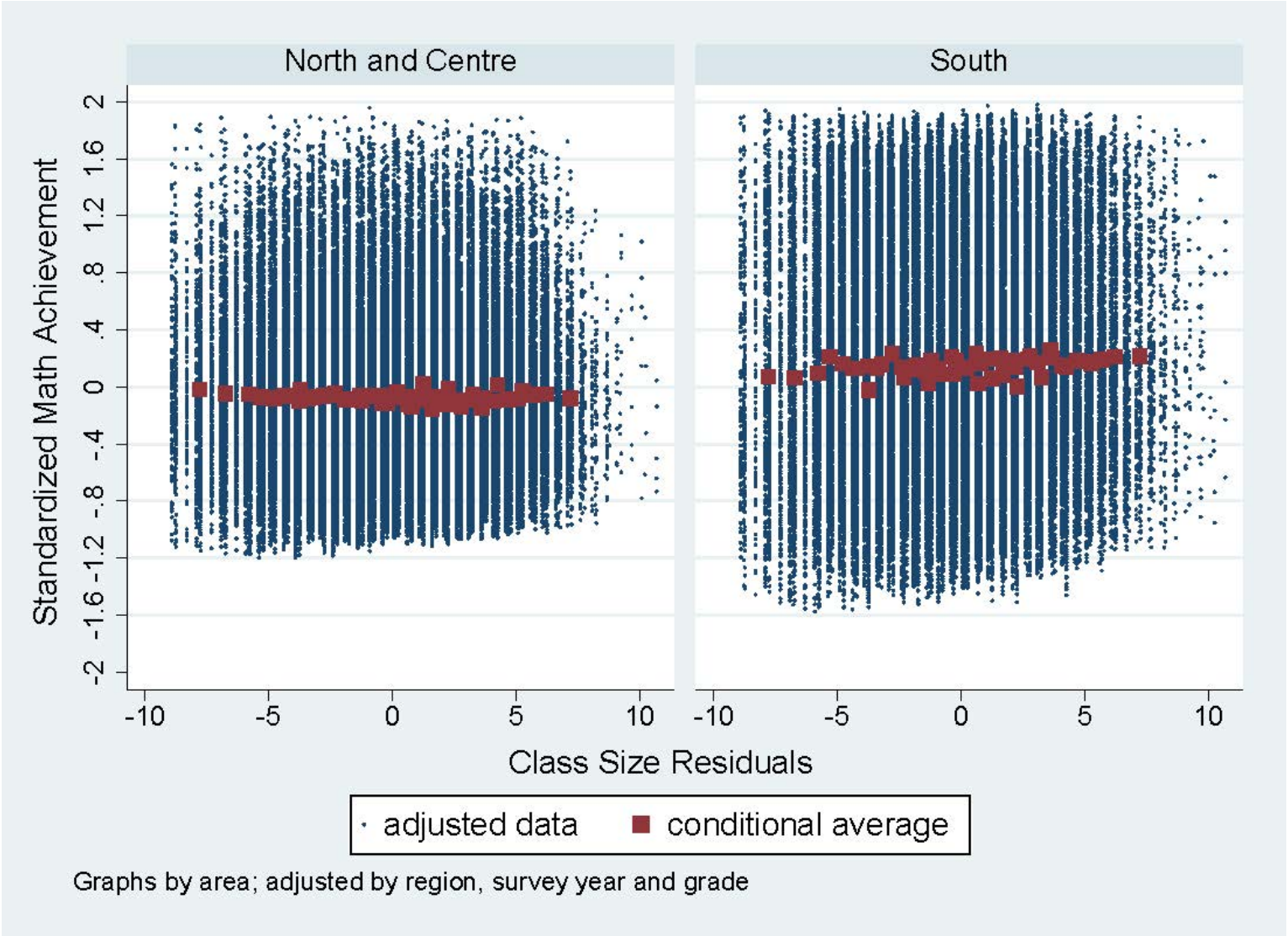
$$Y = \beta_0 + \beta_1 X_1 + X_2' \beta_2 + e.$$

- Class-level **outcomes** Y.
  - Average score standardized by grade, census year and subject.
  - Indicator for having compromised scores (as in the map).
- Class-level **scalar treatment** X_1.
  - Number of students (from administrative records).
  - Presence of an external monitor. About 20% of schools are randomly assigned external monitors, who supervise test administration and grading.
- Class-level **vector of controls** X_2.
  - Student demographics: include gender, citizenship, and information on parents' employment status and education.
  - Sampling strata (used in the monitoring experiment).
  - Indicators for census year, grade and region.

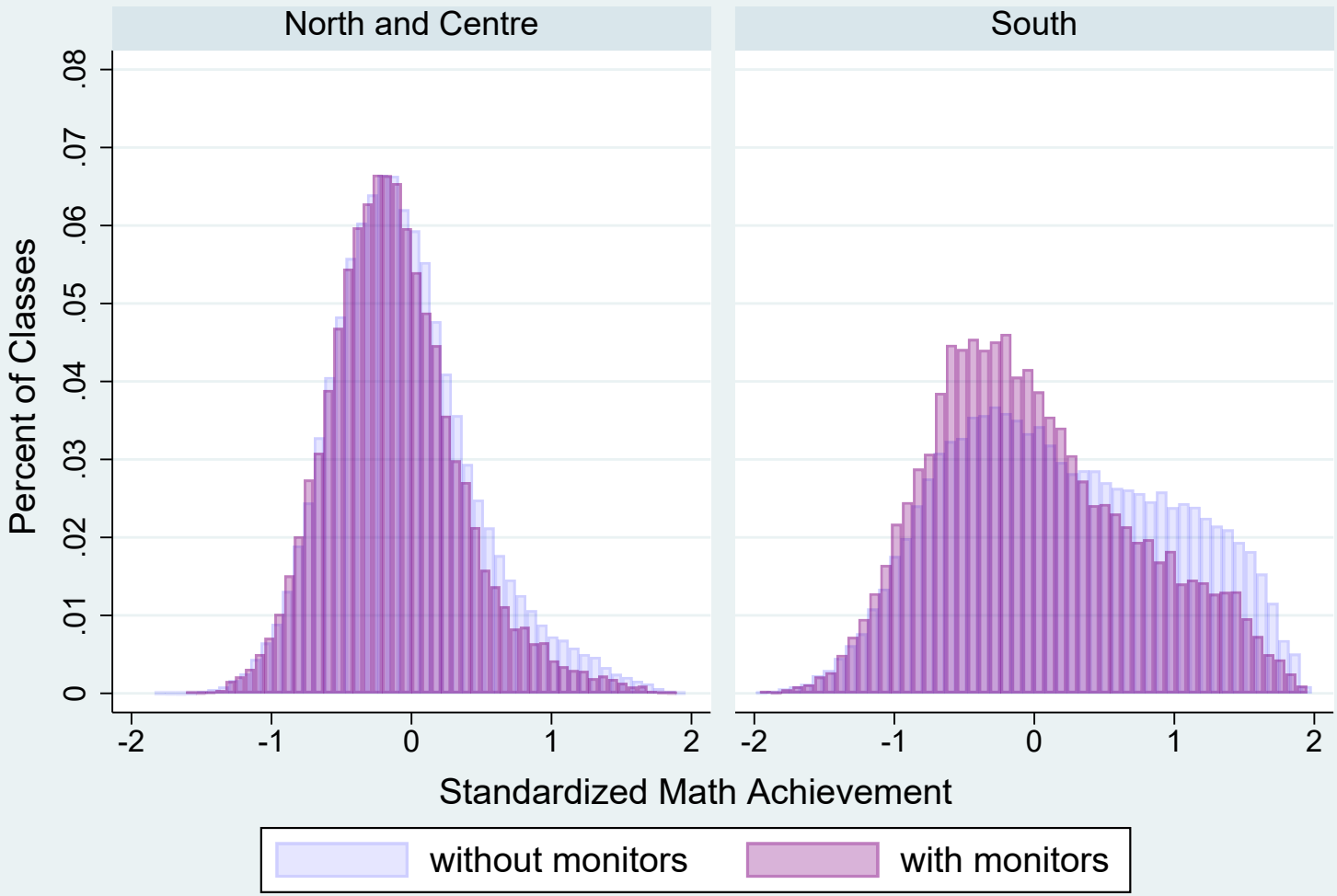
Visualize Achievement and Class Size



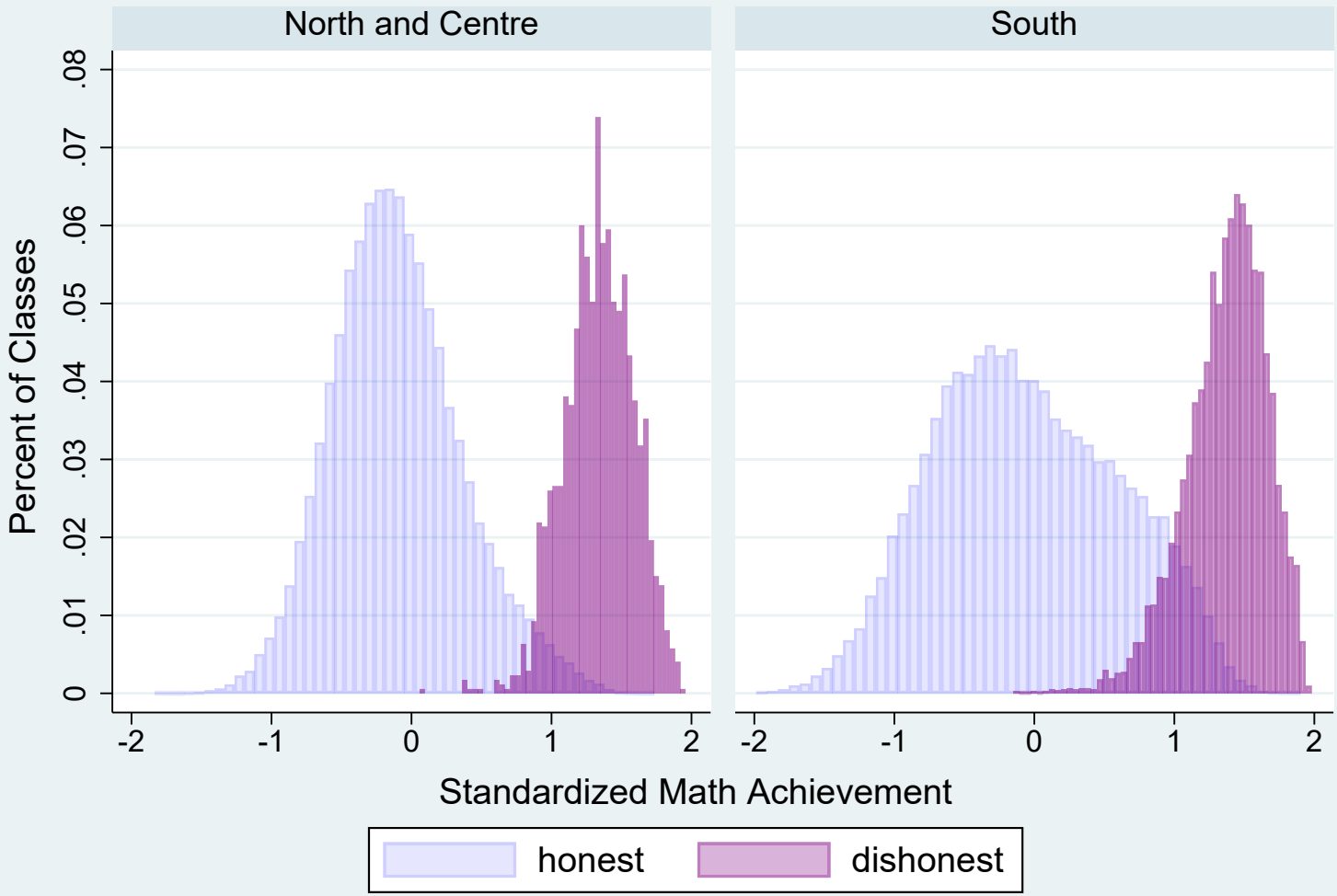
Visualize Achievement and Class Size



Visualize Achievement and Monitoring



Visualize Achievement and Cheating



Regressions of Achievement on Class Size

	(1) Italy	(2) Italy	(3) North/Centre	(4) North/Centre	(5) South	(6) South
			Panel A. Standardized Math Achievement			
Class size	-0.0004	-0.0005	-0.0022***	-0.0016***	0.0073***	0.0056***
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0010)	(0.0010)
			Panel B. Standardized Language Achievement			
Class size	0.0033***	0.0032***	-0.0002	0.0002	0.0095***	0.0086***
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0008)	(0.0008)
Observations	140,010	140,010	87,498	87,498	52,512	52,512
R-squared	0.000	0.000	0.000	0.007	0.001	0.009
Method	OLS	OLS	OLS	OLS	OLS	OLS
Controls	NO	YES	NO	YES	NO	YES
Clustered SE	NO	NO	NO	NO	NO	NO

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regressions of Achievement on Monitoring

	(1) Italy	(2) Italy	(3) North/Centre	(4) North/Centre	(5) South	(6) South
			Panel A. Standardized Math Achievement			
Monitor	-0.1179***	-0.1123***	-0.0638***	-0.0740***	-0.1974***	-0.1804***
	(0.0039)	(0.0043)	(0.0039)	(0.0043)	(0.0082)	(0.0089)
			Panel B. Standardized Language Achievement			
Monitor	-0.0671***	-0.0817***	-0.0446***	-0.0529***	-0.1041***	-0.1330***
	(0.0032)	(0.0036)	(0.0033)	(0.0037)	(0.0067)	(0.0073)
Observations	140,010	140,010	87,498	87,498	52,512	52,512
R-squared	0.000	0.000	0.000	0.007	0.001	0.009
Method	OLS	OLS	OLS	OLS	OLS	OLS
Controls	NO	YES	NO	YES	NO	YES
Clustered SE	NO	NO	NO	NO	NO	NO

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regressions of Cheating on Monitoring

	(1) Italy	(2) Italy	(3) North/Centre	(4) North/Centre	(5) South	(6) South
			Panel A. Math Score Manipulation			
Monitor	-0.0399***	-0.0291***	-0.0107***	-0.0101***	-0.0825***	-0.0624***
	(0.0015)	(0.0016)	(0.0011)	(0.0012)	(0.0036)	(0.0039)
			Panel B. Language Score Manipulation			
Monitor	-0.0319***	-0.0247***	-0.0127***	-0.0119***	-0.0592***	-0.0472***
	(0.0014)	(0.0015)	(0.0012)	(0.0013)	(0.0032)	(0.0035)
Observations	140,010	140,010	87,498	87,498	52,512	52,512
R-squared	0.000	0.000	0.000	0.007	0.001	0.009
Method	OLS	OLS	OLS	OLS	OLS	OLS
Controls	NO	YES	NO	YES	NO	YES
Clustered SE	NO	NO	NO	NO	NO	NO

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Estimation

Inference and the Analogy Principle

- In practice we don't usually know the CEF or the population regression vector, and learn about these quantities using **samples**.
- Assume that a sample of n independent and identically distributed units is available, so that we have n realizations of values for the variables Y and X.
- This is a quite common position to take in micro-econometric work where data may arise by random sampling of households, students or firms from some population.
- We might have data recording responses of many members of a number of households. Then we might expect dependence among responses from household members, but perhaps independence among responses from different households.
- As another example, we might have students grouped in the same class or school.
- At times households in the same village share a common value of X, for example when this denotes a treatment assigned across villages.

Inference and the Analogy Principle

- As the regression equation must hold for all units in the sample, variables are now indexed to unit i:

$$Y_i = X_i'\beta + e_i,$$

where the following $K \times 1$ vector is defined:

$$\underset{K\times 1}{X_i} \equiv (1, X_{1i}, X_{2i}, \ldots, X_{K-1i})'.$$

- Also recall that the population regression yields:

$$\beta = E[X_i X_i']^{-1} E[X_i Y_i].$$

- One way to proceed is by what I will loosely call the **analogy principle**. This involves expressing the parameter to be estimated (here β) as a function of expected values of random variables, and replacing expected values by sample data based analogs of them.

Inference and the Analogy Principle

- It is tedious but indeed useful to take a closer look at the matrices involved in the last expression. For example there is:

$$\underset{K\times K}{E[X_i X_i']} = \begin{pmatrix} 1 & E[X_{1i}] & E[X_{2i}] & \cdots & E[X_{K-1i}] \\ & E[X_{1i}^2] & E[X_{1i}X_{2i}] & \cdots & E[X_{1i}X_{K-1i}] \\ & & \ddots & \vdots & \vdots \\ & & & \ddots & \vdots \\ & & & & E[X_{K-1i}^2] \end{pmatrix},$$

  which is a $K \times K$ symmetric matrix.

- A natural estimator for any of these moments is the **sample analogue** (which is motivated by the law of large numbers):

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i'.$$

Inference and the Analogy Principle

- The same reasoning applies to the $K \times 1$ vector $E[X_i Y_i]$, for which the following estimator seems plausible:

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i.$$

- This analogy principle yields the following plug-in estimator:

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right).$$

- It is called the **Ordinary Least Squares (OLS)** estimator of β because it solves the sample analog of the least-squares problem defined at the beginning:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2.$$

Inference

- The OLS estimator is a statistic $\hat{\beta}_n$ (i.e., a function of the data) computed from n observations:

$$\begin{aligned}\hat{\beta}_n &\equiv \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right), \\ &= \beta + \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i e_i\right).\end{aligned}$$

- Consider how the values of this statistic may change across **repeated samples** (recall our discussion on sampling).
- Note that:

$$E\left[\hat{\beta}_n\right] = \beta + E\left[\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i e_i\right)\right],$$

so that LIE implies unbiasedness of the estimator only if the CEF is linear (because in this case $E(e_i|X_i = x) = 0$).

- To see this, consider again the case of a simple univariate regression:

$$\beta_1 = \frac{Cov[X, Y]}{Var[X]}.$$

- The empirical analog of this expression is:

$$\hat{\beta}_{1n} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^{n} (X_i - \bar{X}) e_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

- It follows that the last term is not zero in general since:

$$E\left[\frac{\sum_{i=1}^{n} (X_i - \bar{X}) e_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right] \neq \frac{E\left[(X_i - \bar{X}) e_i\right]}{E\left[(X_i - \bar{X})^2\right]} = 0.$$

- This result is true asymptotically however, because of the analogy principle, or in finite samples if $E(e_i|X_i = x) = 0$ (using LIE).

Conditional Inference

- Holding X_i fixed in the derivation ensures that the distribution of $\hat{\beta}_n$ is always centered at β, regardless of the sample size n.
- Since with fixed regressors the only source of variability is e_i:

$$E\left[\hat{\beta}_n\right] = \beta + \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i \underbrace{E\left[e_i\right]}_{=0}\right) = \beta,$$

  the result following from the first order conditions that define e_i.
- This **principle of conditionality** tells us that inferences about the regression function should be made conditional on realizations of X_i.
- Rationale for conditioning on X_i: when interested in features of the conditional distribution of Y_i given $X_i = x$ the information relevant for statistical inference is not contained in the marginal distribution of X_i.
- My proofs will avoid this conditioning, as it is not needed to discuss the properties of $\hat{\beta}_n$ when the sample gets bigger and bigger.

On the Road to Asymptopia

Large Sample Theory

- We can also re-arrange terms to write:

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right).$$

- I will study the **large sample** (or **asymptotic**, i.e. as $n \to \infty$) approximation to the distribution of the estimator.
- I will consider the properties of this distribution as n gets larger and larger. The limit distribution is often referred to as limiting distribution of the estimator.
- This distribution will be used to draw **approximate inference** on the parameters the estimator is constructed for.
- What we will do next is to review the core terms and concepts of statistical theory needed for general asymptotic distribution theory.
- Computer age alternatives to approximate inference (e.g., the bootstrap) will be considered later in this course.

Modes of Convergence

- The **scalar** sequence $\mathcal{S}_n$ here is indexed to n, and convergence is investigated as $n \to \infty$.
- With **convergence in distribution** we expect the next element in the sequence $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots$ becoming "better and better" described by a certain distribution (e.g., a normal random variable). This is the weakest mode of convergence.
- The sequence $\mathcal{S}_n$ converges in law (or weakly) to the variable Θ if:

$$\lim_{n \to \infty} Pr[\mathcal{S}_n \leq s] = Pr[\Theta \leq s],$$

  at every point s at which the distribution of Θ is continuous (i.e., does not have a jump).
- In other words, we require that the random variables $\mathcal{S}_n$ and Θ have the same distribution in a large enough sample.
- If this condition is verified we write $\mathcal{S}_n \xrightarrow{D} \Theta$.

Modes of Convergence

- Clearly two identically distributed random variables $\mathcal{S}_n$ and Θ are not the same variable in general, but only have realizations from the same probability law.
- **Convergence in probability** embodies the idea that far out in the sequence $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \ldots$ each realization of $\mathcal{S}_n$ is "very close" to each realization of Θ.
- Formally, we say that the sequence $\mathcal{S}_n$ converges in probability to the variable Θ if for every positive ξ:

$$\lim_{n \to \infty} Pr[|\mathcal{S}_n - \Theta| > \xi] = 0.$$

- In other words, the probability of large differences between realizations from $\mathcal{S}_n$ and Θ can be made as small as you like in a large enough sample.
- If this condition is verified we write $\mathcal{S}_n \xrightarrow{P} \Theta$.

Modes of Convergence

- It follows that $\mathcal{S}_n \xrightarrow{P} \Theta$ implies $\mathcal{S}_n \xrightarrow{D} \Theta$, but not the opposite.
- Clearly Θ can degenerate at the point θ, in which case $\mathcal{S}_n$ converges to the constant θ. Only in this case $\mathcal{S}_n \xrightarrow{D} \theta$ implies $\mathcal{S}_n \xrightarrow{P} \Theta$.
- Convergence in probability passes through continuous transformations, since $\mathcal{S}_n \xrightarrow{P} \Theta$ implies $g(\mathcal{S}_n) \xrightarrow{P} g(\Theta)$ for any continuous function $g(\cdot)$ (**continuous mapping theorem**).
- Alternative concepts of convergence use different definitions of "proximity" between $\mathcal{S}_n$ and Θ. **Convergence in r-th mean** requires that the r-th power of the absolute difference of their realizations is small on average:

$$\lim_{n \to \infty} E\left[|\mathcal{S}_n - \Theta|^r\right] = 0.$$

- If this condition is verified we write $\mathcal{S}_n \xrightarrow{r} \Theta$.

Modes of Convergence

- $\mathcal{S}_n \xrightarrow{r} \Theta$ implies $\mathcal{S}_n \xrightarrow{P} \Theta$.
- $\mathcal{S}_n \xrightarrow{r_1} \Theta$ implies $\mathcal{S}_n \xrightarrow{r_2} \Theta$ for $r_1 > r_2$.
- $\mathcal{S}_n \xrightarrow{2} \Theta$ implies $E[\mathcal{S}_n] \to E[\Theta]$ and $E\left[(\mathcal{S}_n)^2\right] \to E\left[(\Theta)^2\right]$, which in turn implies $Var[\mathcal{S}_n] \to Var[\Theta]$.
- $Var[\mathcal{S}_n] \to 0$ and $E[\mathcal{S}_n] \to \theta$ imply $\mathcal{S}_n \xrightarrow{2} \theta$ (a result that we use to show the consistency of estimators).
- The **Slutzky's Theorem** states that the sum and product of two variables, one of which converges in distribution and the other in probability to a constant, have asymptotic distributions unaffected by replacing the one that converges to a constant by this constant.
- More formally, $\mathcal{S}_{1n} \xrightarrow{D} \Theta_1$ and $\mathcal{S}_{2n} \xrightarrow{P} \theta_2$ imply:

$$\mathcal{S}_{1n} + \mathcal{S}_{2n} \xrightarrow{D} \Theta_1 + \theta_2, \qquad \mathcal{S}_{1n}\mathcal{S}_{2n} \xrightarrow{D} \theta_2\Theta_1,$$

a result that we use for approximate inference.

Law of Large Numbers

- It turns out that the **scalar** sequences $\mathcal{S}_n$ we will be interested in take the form of (possibly weighted) sums of random variables:

$$\mathcal{S}_n = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

- This connects to the idea of performing the same experiment a large number of times (coin tossing, die rolling, sampling), in which case the W_i's are independent and identically distributed (i.i.d.).
- In the i.i.d. case, **Khintchine's law of large numbers (LLN)** states that $\mathcal{S}_n \xrightarrow{P} E[W_i]$ if this moment is finite (curiously enough with no restriction on the variance).
- This embodies the idea that sample moments converge to the corresponding population moments as more trials are performed ($n \to \infty$). This is the foundation to the analogy principle.

Law of Large Numbers

- **Chebychev's law** extends the idea to variables independent but not identically distributed, stating that $\mathcal{S}_n \xrightarrow{P} n^{-1} \sum_{i=1}^{n} E[W_i]$ if all $E[W_i]$ and $Var[W_i]$ are finite and $Var[\mathcal{S}_n] \to 0$.
- This embodies the idea that the law works only if $\mathcal{S}_n$ becomes increasingly more precise as the sample gets bigger (which is obvious in the i.i.d. case when $Var[W_i]$ is finite).
- LLNs are not informative about how large n must be for $\mathcal{S}_n$ to be "not too far" from the probability limit. The answer to this question is obtained by showing that the distribution of $\mathcal{S}_n$ is well approximated in large samples by a normal distribution. We will see this operation done with the OLS estimator shortly.
- In this case we will say that the sequence $\mathcal{S}_n$ converges in distribution to a normal random variable, or simply that **asymptotic normality** holds.

Central Limit Theorems

- The rationale for referring to asymptotic normality builds upon the existence of **central limit theorems (CLT)**.
- The Lindberg-Levy theorem gives the limiting distribution of a mean of n i.i.d. random variables (i.e., identical statistical experiments):

$$\mathcal{S}_n = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

- In particular if $E[W_i]$ and $Var[W_i]$ are both finite we have that:

$$P\left[\frac{\mathcal{S}_n - E[W_i]}{\sqrt{\frac{Var[W_i]}{n}}} \leq s\right] \rightarrow \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{s} e^{-\frac{1}{2}x^2}dx.$$

or more compactly:

$$\frac{\mathcal{S}_n - E[W_i]}{\sqrt{\frac{Var[W_i]}{n}}} \xrightarrow{D} \mathcal{N}(0,1).$$

Central Limit Theorems

- Using the properties of the normal random variable, the result can be stated equivalently as:

$$\begin{aligned} \sqrt{n}\left(\mathcal{S}_n - E[W_i]\right) &\xrightarrow{D} \mathcal{N}\left(0, Var[W_i]\right), \\ \mathcal{S}_n &\xrightarrow{D} \mathcal{N}\left(E[W_i], \frac{Var[W_i]}{n}\right). \end{aligned}$$

- The first expression embodies the idea of **root-n consistency**, as this is the rate at which $\mathcal{S}_n \simeq E[W_i]$ when sample size grows.
- There is no sense in which we ever think of the sample size actually becoming "infinite". Sometimes the use of large sample approximations is criticised by saying that "the sample isn't large", which is an ignorant comment.
- These large sample approximations are just... approximations. They are as good or bad as the size of the error incurred in using the approximation, which depends on the sample size but on other factors as well. This is studied in Monte Carlo simulations.

- This result can be used to make **approximate probability statements** (i.e. statements that hold as long as n is large enough). For example, since:

$$P\left[-1.96 \leq \frac{\mathcal{S}_n - E[W_i]}{\sqrt{\frac{Var[W_i]}{n}}} \leq 1.96\right] \simeq 0.95,$$

there is also:

$$\mathcal{S}_n \in \left[E[W_i] - 1.96\sqrt{\frac{Var[W_i]}{n}}, E[W_i] + 1.96\sqrt{\frac{Var[W_i]}{n}}\right],$$

with approximately 95% probability.

Central Limit Theorems

- The definitions discussed apply element by element to sequences of random **vectors** or **matrices**. For example, we write:

$$[\mathcal{S}_{1n}, \ldots, \mathcal{S}_{Kn}]' \xrightarrow{D} [0, \ldots, 0]',$$

  if each element of the $K \times 1$ random vector converges to zero. The same idea extends to the elements of matrices.

- Note that "distance" here is the Euclidean length, where for any $K \times 1$ vector $|X| = \sqrt{X'X}$. For any $K \times K$ matrix we have $|X| = \sqrt{tr(X'X)}$, which uses the trace of a square matrix.

- The Lindberg-Levy CLT extends to a **multivariate normal distribution** very intuitively:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underset{K\times 1}{W_i} \xrightarrow{D} \mathcal{N}\left(\underset{K\times 1}{0}, \underset{K\times K}{\Sigma_W} \right),$$

  where Σ_W is the $K \times K$ variance covariance matrix (which simplifies to $E[W_i W_i']$ if $E[W_i] = 0$).

Approximate Inference

Asymptotic OLS Inference

- Go back the this expression for the OLS estimator $\hat{\beta}_n$ re-centred at β and scaled by $\sqrt{n}$ (recall that I assumed i.i.d. units for now, we will see how to relax this shortly):

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right),$$

  and consider the probability limits of each component.

- The LLN implies:

$$\underset{K\times K}{\frac{1}{n}\sum_{i=1}^{n} X_i X_i'} \xrightarrow{P} E[X_i X_i'].$$

- Recall that $E[X_i e_i] = 0$ because of the first-order conditions yielding $\hat{\beta}_n$, so that the $K \times 1$ vector $X_i e_i$ is centred at zero. We have that:

$$\underset{K\times K}{Var\,(X_i e_i)} = E[X_i X_i' e_i^2].$$

Asymptotic OLS Inference

- The multivariate Lindberg-Levy's CLT implies:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i e_i\right) \xrightarrow{D} \mathcal{N}\left(0, E[X_i X_i' e_i^2]\right).$$

- Using the Slutzky theorem and re-arranging terms to maximize beauty we get (as long as the matrix of expected squares and cross-products is non-singular):

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta\right) &\xrightarrow{D} [E(X_i X_i')]^{-1}\,\mathcal{N}\left(0, E[X_i X_i' e_i^2]\right), \\ &\overset{D}{\sim} \mathcal{N}\left(0, [E(X_i X_i')]^{-1} E[X_i X_i' e_i^2] [E(X_i X_i')]^{-1}\right).\end{aligned}$$

- It follows that $\hat{\beta}_n \xrightarrow{2} \beta$ because:

$$\lim_{n\to\infty} Var\left(\hat{\beta}_n\right) = \lim_{n\to\infty} \frac{1}{n}[E(X_i X_i')]^{-1} E[X_i X_i' e_i^2] [E(X_i X_i')]^{-1} = 0,$$

from which $\hat{\beta}_n \xrightarrow{P} \beta$ also follows.

Asymptotic OLS Inference

- In practice the terms in the variance of $\hat{\beta}_n$ are unknown and we will have to replace them by estimates using the estimated residuals $\hat{e}_i \equiv Y_i - X_i'\hat{\beta}$:

$$\frac{1}{n}\left[\underbrace{E(X_iX_i')}_{\frac{1}{n}\sum_{i=1}^n X_iX_i'}\right]^{-1} \underbrace{E[X_iX_i'e_i^2]}_{\frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 X_iX_i'} \left[\underbrace{E(X_iX_i')}_{\frac{1}{n}\sum_{i=1}^n X_iX_i'}\right]^{-1}$$

- A straightforward application of the analogue principle ensures that we can estimate the variance of $\hat{\beta}_n$.
- It is a surprising and crucial aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy...

Taking Stock

- The OLS estimator $\hat{\beta}_n$ is asymptotically unbiased or **consistent** (recall our discussion on finite sample unbiasedness).
- The value of $\hat{\beta}_n$ in the data is a realization of this random variable, and is called estimate.
- How accurate is this number? The **standard errors** (that we will use soon to construct t-statistics) are the square roots of the diagonal elements of the $K \times K$ matrix:

$$\widehat{Var}\left(\hat{\beta}_n\right) = \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1} \sum_{i=1}^{n} \hat{e}_i^2 X_i X_i' \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1}.$$

- This matrix computed from the data is an estimate of the precision on an estimate...
- Asymptotic standard errors computed in this way are known as **heteroskedasticity-consistent** or **robust** standard errors. It is also known as the **sandwich estimator** of variance (because of how the calculation formula physically appears).

Taking Stock

- Default standard errors are usually derived under a **homoskedasticity** assumption:

$$\begin{aligned} E[X_i X_i' e_i^2] &= E\left(X_i X_i' E[e_i^2|X_i]\right) = \sigma_e^2 E[X_i X_i'], \\ \lim_{n\to\infty} Var\left(\hat{\beta}_n\right) &= \lim_{n\to\infty} \frac{1}{n}\sigma_e^2 E[X_i X_i']^{-1} = 0. \end{aligned}$$

- By letting:

$$\hat{\sigma}_e^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n} \hat{e}_i^2,$$

the diagonal elements of the following $K \times K$ matrix are usually reported unless you request otherwise:

$$\widehat{Var\left(\hat{\beta}_n\right)} = \hat{\sigma}_e^2 \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1}.$$

Heteroskedasticity for Dummies

- When the outcome variable Y_i is a dummy (e.g., employment status) the CEF describes a probability:

$$E\left(Y_i|X_i = x\right) = P\left(Y_i = 1|X_i = x\right).$$

- In the **linear probability model**:

$$Y_i = X_i'\beta + e_i,$$

  we have $Var\left(Y_i|X_i = x\right) = Var\left(e_i|X_i = x\right)$, and this variance is equal to $P\left(Y_i = 1|X_i = x\right)\left[1 - P\left(Y_i = 1|X_i = x\right)\right]$ because Y_i is a realization from a Bernoulli trial.

- The problem remains if the CEF is linear.
- More in general, for any outcome Y_i we can write:

$$\begin{aligned} Var\left(Y_i|X_i = x\right) &= E\left[e_i^2|X_i = x\right] - \left[E\left(e_i|X_i = x\right)\right]^2, \\ &= E\left[\left(Y_i - X_i'\beta\right)^2|X_i = x\right] - \left[E\left(Y_i|X_i = x\right) - X_i'\beta\right]^2. \end{aligned}$$

Heteroskedasticity for Dummies

- The last expression implies:

$$E\left[(Y_i - X_i'\beta)^2 | X_i = x\right] = Var\left(Y_i | X_i = x\right) + \left[E\left(Y_i | X_i = x\right) - X_i'\beta\right]^2,$$

  meaning that the residual variation may increase with the square of the gap between the regression line and the CEF even if $Var\left(Y_i | X_i = x\right)$ doesn't depend on X_i.
- Bottom line: always consider heteroskedasticity-consistent standard errors to be on the safe side.
- However, Monte Carlo simulations show that with moderate heteroskedasticity the robust estimator is biased downwards. To avoid the use of technical fixes to this problem, as a rule-of-thumb use the maximum of old-fashioned and robust standard errors to avoid gross misjudgments of precision.

Example of Stata Output

Source	SS	df	MS	Number of obs =	87,498
				F(4, 87493) =	149.36
Model	149.391479	4	37.3478699	Prob > F =	0.0000
Residual	21877.9484	87,493	.2500537	R-squared =	0.0068
				Adj R-squared =	0.0067
Total	22027.3398	87,497	.251749658	Root MSE =	.50005

answers_ma~d	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
clsize_snv	-.0015838	.0004817	-3.29	0.001	-.0025279	-.0006396
survey						
2010	.0230391	.0041351	5.57	0.000	.0149343	.0311439
2011	-.0215401	.0041422	-5.20	0.000	-.0296588	-.0134214
1.grade	-.0723657	.0033874	-21.36	0.000	-.079005	-.0657263
_cons	-.0074777	.0099945	-0.75	0.454	-.0270668	.0121113

Example of Stata Output

- **Coef.** is the $K \times 1$ vector:

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right).$$

- **Std. Err.** is the square root of the K diagonal elements of:

$$\widehat{Var}_{standard} \left(\hat{\beta}_n \right) = \hat{\sigma}_e^2 \left[\sum_{i=1}^{n} X_i X_i' \right]^{-1}.$$

- The **ANOVA table** is obtained from the sample analogue of:

$$\underbrace{Var(Y_i)}_{\text{Total}} = \underbrace{Var(X_i'\beta)}_{\text{Model}} + \underbrace{Var(e_i)}_{\text{Residual}},$$

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (X_i'\hat{\beta} - \bar{Y})^2 + \sum_{i=1}^{n} \hat{e}_i^2,$$

because $E\left(X_i'\hat{\beta}\right) = E\left(Y_i\right)$ and $\bar{Y} \xrightarrow{P} E\left(Y_i\right)$.

Example of Stata Output

- **df** are the degrees of freedom (values in the calculation that are "free to vary"): $n-1=87,497$ in total and $n-K-1=87,493$ residual after fitting the model (K includes the constant).
- **Root MSE** is ($n-K=87,493$):

$$\sqrt{\hat{\sigma}_e^2 \equiv \frac{1}{n-K}\sum_{i=1}^{n}\hat{e}_i^2}.$$

- **R-squared** is:

$$1-\frac{\sum_{i=1}^{n}\hat{e}_i^2}{\sum_{i=1}^{n}Y_i^2}.$$

- **95% Conf. Interval** are obtained using the approximation:

$$\sqrt{n}\left(\hat{\beta}_n-\beta\right)\xrightarrow{D}\mathcal{N}\left(0,\hat{\sigma}_e^2\left[\sum_{i=1}^{n}X_iX_i'\right]^{-1}\right).$$

Example of Stata Output

```
Linear regression                               Number of obs     =        87,498
                                                F(4, 87493)       =        150.65
                                                Prob > F          =        0.0000
                                                R-squared         =        0.0068
                                                Root MSE          =        .50005

------------------------------------------------------------------------------
             |               Robust
answers_ma~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    clsize_snv |  -.0015838   .0004918    -3.22   0.001    -.0025478   -.0006198
               |
      survey |
        2010 |   .0230391   .0040532     5.68   0.000     .0150948    .0309834
        2011 |  -.0215401   .0041647    -5.17   0.000    -.0297029   -.0133772
               |
     1.grade |  -.0723657   .0033867   -21.37   0.000    -.0790035   -.0657278
       _cons |  -.0074777   .0103115    -0.73   0.468    -.0276881    .0127326
------------------------------------------------------------------------------
```

Example of Stata Output

- **Robust Std. Err.** is the square root of the K diagonal elements of:

$$\widehat{Var}_{robust}\left(\hat{\beta}_n\right) = \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1} \sum_{i=1}^{n} \hat{e}_i^2 X_i X_i' \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1}.$$

- **95% Conf. Interval** are obtained using the approximation:

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{D} \mathcal{N}\left(0, \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1} \sum_{i=1}^{n} \hat{e}_i^2 X_i X_i' \left[\sum_{i=1}^{n} X_i X_i'\right]^{-1}\right).$$

Asymptotic Fix-Ups

Nonformulaic Approach to Standard Errors

- An algorithm (e.g., the average $\bar{x}$) has produced an estimate of the corresponding population parameter. How accurate is the estimate?
- The standard error of an estimate is the standard deviation one would observe by repeatedly obtaining new samples from the population distribution that generated the data. Of course this is impossible because the distribution is unknown:

$$F \overset{i.i.d.}{\rightarrow} \underbrace{\mathbf{X}}_{data} \rightarrow \bar{x}.$$

- Explicit standard error formulas exist:
  - for algorithms that can be written in some forms of averaging, such as linear regressions;
  - after "linearizing" (i.e., Taylor series approximations for smooth functions of averages).
- The bootstrap replaces the unknown population distribution with its analog estimate, and then simulates data:

$$\hat{F} \overset{i.i.d.}{\rightarrow} \underbrace{\mathbf{X}^*}_{pseudo-data} \rightarrow \bar{x}^*.$$

Formulaic Approach to Standard Errors

- Computer power is being substituted for theoretical calculations!
- Taylor expansions for the transformation of a vector of statistics $\mathbf{S}$ (local linear approximation, sometimes known as the **delta method**):

$$G(\mathbf{S}) \approx G(\mu_S) + \nabla G(\mu_S)'(\mathbf{S} - \mu_S).$$

- We can then take the variance of this approximation to estimate the variance of $G(\mathbf{S})$ and thus the standard error:

$$Var(G(\mathbf{S})) \approx \nabla G(\mu_S)' Cov(\mathbf{S}) \nabla G(\mu_S).$$

- For example, consider $G(\mathbf{S}) = \hat{\beta}_n$ and:

$$\mathbf{S} = \left[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \sum_{i=1}^{n}(X_i - \bar{X})^2\right]'.$$

The Nonparametric Bootstrap

- Motivation begins by noticing that $\hat{\beta}_n$ is obtained in two steps. First, observations (Y_i, X_i) are generated by sampling from a population distribution $F_{Y_i X_i}$, so that a sample is obtained:

$$(\mathbf{Y}, \mathbf{X}) \equiv \{(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\}.$$

- Second, the OLS "algorithm" $\hat{\beta}_n$ yields an estimate using this sample.
- The population distribution is unknown but can be estimated from the observed data (the **empirical probability distribution** $\hat{F}_{Y_i X_i}$).
- Some large number B of samples are independently drawn from this distribution as I discussed in the previous set of slides:

$$(\mathbf{Y}^*, \mathbf{X}^*) \equiv \{(Y_1^*, X_1^*), (Y_2^*, X_2^*), \ldots, (Y_n^*, X_n^*)\}.$$

- Each pseudo-sample provides a bootstrap replication of the statistic of interest, $\hat{\beta}_n^{*b}$, for $b = 1, \ldots, B$, and obviously $\hat{\beta}_n^{*b} \overset{d}{\sim} \hat{\beta}_n$.

The Nonparametric Bootstrap

- In practice each (Y_i^*, X_i^*) is drawn randomly with equal probability and with replacement from $(\mathbf{Y}, \mathbf{X})$.
- Think of the $n \times 1$ resampling vector $R^* \equiv [R_1^*, R_2^*, \ldots, R_n^*]'$ whose elements count the number of times each observation (Y_i, X_i) appears in the sample, and the $n \times 1$ vector $R_{obs} \equiv [1, 1, \ldots, 1]'$.
- If $n = 3$: $[3,0,0]'$, $[0,3,0]'$, $[0,0,3]'$, $[2,1,0]'$, $[2,0,1]'$, $[1,2,0]'$, $[0,2,1]'$, $[1,0,2]'$, $[0,1,2]'$, $[1,1,1]'$.
- We have R^* is distributed according to a Multinomial distribution, $R^* \overset{d}{\sim} \mathcal{M}\left(n, \frac{1}{n} R_{obs}\right)$, which is a generalization of the Binomial distribution.
- Note that $\hat{\beta}_n^{*b}$ has the same value for any permutation of the vector elements: there are $\binom{2n-1}{n}$ possible distinct samples ($n = 10$ yields $92,378$ samples).
- When $n = 20$ and $B = 2000$, the probability is greater than 95% that none of the bootstrap samples will repeat.

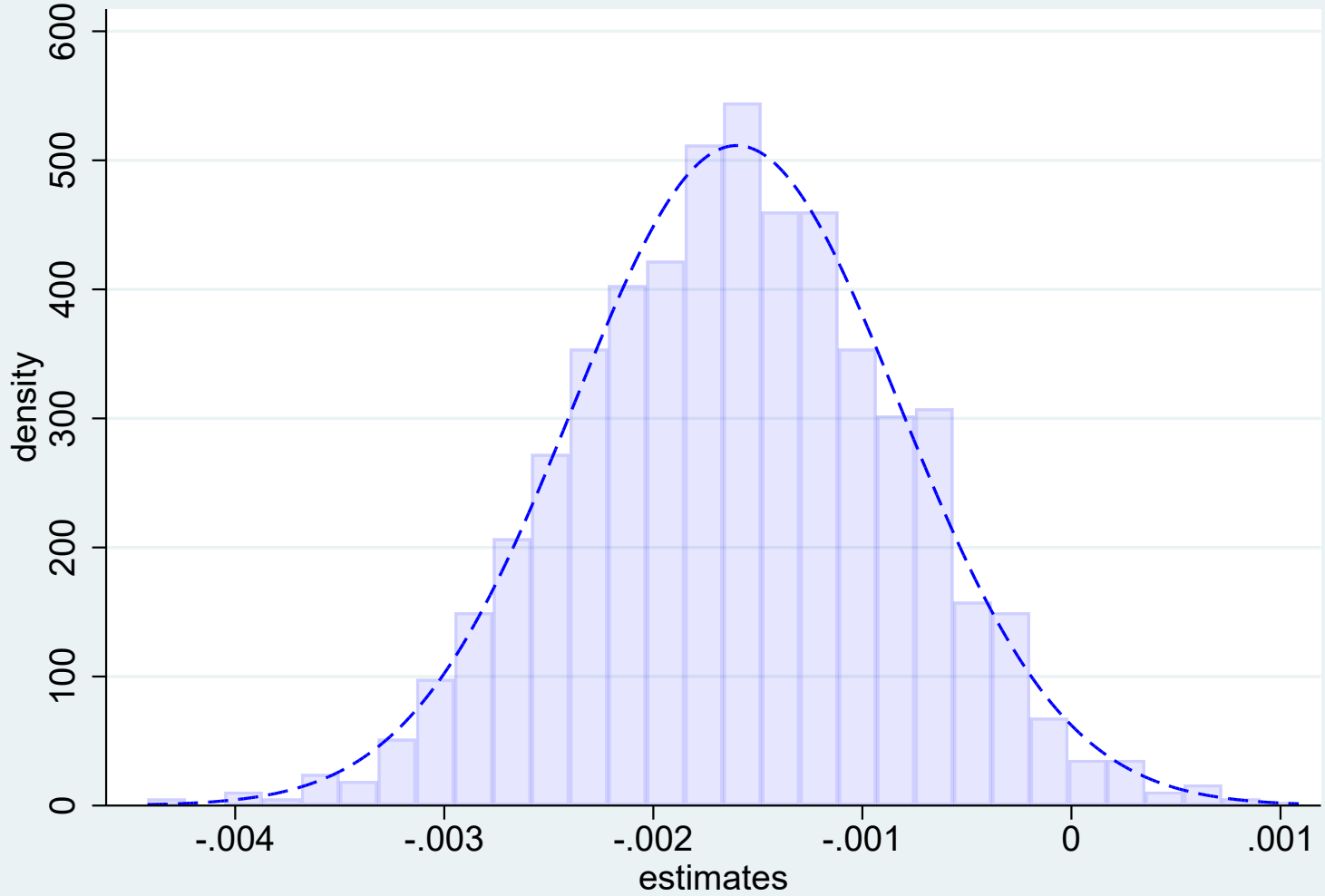
The Nonparametric Bootstrap

- The resampling plan operates by holding the original data fixed, and seeing how the statistic of interest $\hat{\beta}_n$ changes as the resampling vector R^* varies:

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_n^{*b} = \frac{\sum_{i=1}^n R_i^* (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n R_i^* (X_i - \bar{X})^2}.$$

- Examining the resampling surface: $B = 200$ is usually sufficient for evaluating the variance. Larger values, 1000 or 2000, will be required for bootstrap confidence intervals discussed further below.
- The sampling scheme can be generalized to allow for stratification or clustering (e.g., resampling entire schools instead of individual students).

Example of Stata Output



B=2000 bootstrap replications of the class size coefficient

Example of Stata Output

Linear regression

Number of strata = **13** Number of obs = **87,498**
Replications = **2,000**
Wald chi2(**4**) = **510.84**
Prob > chi2 = **0.0000**
R-squared = **0.0068**
Adj R-squared = **0.0067**
Root MSE = **0.5001**

(Replications based on **3,530** clusters in schoolid)

answers_ma~d	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
clsize_snv	**-.0015838**	**.00078**	**-2.03**	**0.042**	**-.0031126**	**-.000055**
survey						
2010	**.0230391**	**.0053895**	**4.27**	**0.000**	**.0124759**	**.0336023**
2011	**-.0215401**	**.00539**	**-4.00**	**0.000**	**-.0321042**	**-.0109759**
1.grade	**-.0723657**	**.0035571**	**-20.34**	**0.000**	**-.0793374**	**-.0653939**
_cons	**-.0074777**	**.0159239**	**-0.47**	**0.639**	**-.038688**	**.0237326**

Example of Stata Output

- **Bootstrap Std. Err.** for $\hat{\beta}_n$ is the standard deviation of the $\hat{\beta}_n^{*b}$ values:

$$\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_n^{*b} - \hat{\beta}_n^{*\cdot}\right)\left(\hat{\beta}_n^{*b} - \hat{\beta}_n^{*\cdot}\right)',$$

  where $\hat{\beta}_n^{*\cdot}$ is the average estimate across boostrap replications.

- **95% Conf. Interval** are obtained using the approximation:

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_n^{*b} - \hat{\beta}_n^{*\cdot}\right)\left(\hat{\beta}_n^{*b} - \hat{\beta}_n^{*\cdot}\right)'\right).$$

The Parametric Bootstrap

- Suppose one is willing to assume that the observed data vector $\mathbf{X}$ comes from a parametric family of distributions:

$$F(\vartheta) \overset{i.i.d.}{\rightarrow} \underbrace{\mathbf{X}}_{data} \rightarrow \bar{x}.$$

- One can resample from an estimate of this distribution:

$$F\left(\hat{\vartheta}\right) \overset{i.i.d.}{\rightarrow} \underbrace{\mathbf{X}^*}_{pseudo-data} \rightarrow \bar{x}^*.$$

- Approximate the data generating process for Y_i with:

$$Y_i = X_i'\beta + e_i.$$

- The **parametric bootstrap** trusts the model to be completely correct: it trusts that the CEF is correctly specified and that we have the right distribution for the noise e_i, and then generates numbers from this distribution.

The Parametric Bootstrap

- Another natural way to resample from this model is the **residual bootstrap**. We condition on X_i, $\hat{\beta}$, and the distribution of the estimated residuals $\hat{e}_i$ to simulate a new Y_i^{*b} for $b = 1, \ldots, B$:

$$Y_i^{*b} \equiv X_i' \hat{\beta} + \hat{e}_i^{*b}.$$

- Resampling residuals assumes that the CEF is correctly specified but doesn't make any assumption about how the residuals are distributed.
- It mimics a sample drawn with fixed regressors (the conditionality principle), and makes X_i and e_i independent by construction.
- Whenever error terms are not identically distributed, this resampling algorithm is no longer valid.

The Parametric Bootstrap

- A variant of the residual bootstrap, called **wild bootstrap**, draws $X_i'\hat{\beta} + \hat{e}_i^{*b}$ with probability 0.50 and $X_i'\hat{\beta} - \hat{e}_i^{*b}$ otherwise (Rademacher transformation of residuals).
- Calculations (not discussed here) show that this transformation preserves the relationship between residual variances and X_i in the original sample.
- The standard error on the class size coefficient calculated using the wild bootstrap with $B = 2,000$ is 0.0007924. The 95% confidence interval is $[-0.003123, -0.00004568]$.

Confidence Intervals

Example of Stata Output

Linear regression
Number of strata = **13** Number of obs = **87,498**
Replications = **2000**

(Replications based on **3,530** clusters in schoolid)

answers_ma~d	Observed Coef.	Bias	Bootstrap Std. Err.	[95% Conf.	Interval]	
clsize_snv	**-.00158376**	**-.0000174**	**.00078001**	**-.0031126**	**-.000055**	(N)
				-.003081	**-.0001143**	(P)
				-.0030328	**-.0000274**	(BC)
2009b.survey	**0**	**0**	**0**	**0**	**0**	(N)
				0	**0**	(P)
				.	**.**	(BC)
2010.survey	**.0230391**	**-.0001723**	**.00538948**	**.0124759**	**.0336023**	(N)
				.0124251	**.0333125**	(P)
				.0124862	**.0334109**	(BC)
2011.survey	**-.02154007**	**-.0000551**	**.00538997**	**-.0321042**	**-.0109759**	(N)
				-.032418	**-.010967**	(P)
				-.0324477	**-.0109772**	(BC)
0b.grade	**0**	**0**	**0**	**0**	**0**	(N)
				0	**0**	(P)
				.	**.**	(BC)
1.grade	**-.07236565**	**.0000379**	**.00355708**	**-.0793374**	**-.0653939**	(N)
				-.0791648	**-.0650526**	(P)
				-.0793517	**-.0651361**	(BC)
_cons	**-.00747773**	**.0005038**	**.01592391**	**-.038688**	**.0237326**	(N)
				-.0372383	**.0238598**	(P)
				-.0390089	**.0223916**	(BC)

(N) normal confidence interval
(P) percentile confidence interval
(BC) bias-corrected confidence interval

Normal Intervals

- Consider the case of a bivariate regression to ease notation.
- The simplest approach takes literally asymptotic normality, and yields **normal confidence intervals (CI)** with $1-\alpha$ coverage:

$$\left[\hat{\beta}_n - z_{\alpha/2}\sqrt{\widehat{Var}_{standard}\left(\hat{\beta}_n\right)}, \hat{\beta}_n + z_{1-\alpha/2}\sqrt{\widehat{Var}_{standard}\left(\hat{\beta}_n\right)}\right],$$

$$\left[\hat{\beta}_n - z_{\alpha/2}\sqrt{\widehat{Var}_{robust}\left(\hat{\beta}_n\right)}, \hat{\beta}_n + z_{1-\alpha/2}\sqrt{\widehat{Var}_{robust}\left(\hat{\beta}_n\right)}\right],$$

$$\left[\hat{\beta}_n - z_{\alpha/2}\sqrt{\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right)}, \hat{\beta}_n + z_{1-\alpha/2}\sqrt{\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right)}\right],$$

where z_τ is the τ-th percentile of a standard normal distribution.

- **Interpretation**: if we could repeatedly calculate confidence intervals for β using independent samples from the population, $100(1-\alpha)\%$ of these intervals will include the unknown value β (which is a constant!).

Bootstrap Percentile Confidence Intervals

- We can use the shape of the bootstrap distribution to improve upon normal CIs. Define the τ-th quantile from replications:

$$F_{\hat{\beta}_n^*}(t) \equiv \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_n^{*b} \leq t), \qquad \underbrace{\hat{\beta}_n^*(\tau)}_{\tau-th\ quantile} \equiv F^{-1}_{\hat{\beta}_n^*}(\tau).$$

- The **percentile confidence interval** with $1-\alpha$ coverage is:

$$\left[\hat{\beta}_n^*(\alpha/2), \hat{\beta}_n^*(1-\alpha/2)\right].$$

- Consider a transformation $\theta \equiv g(\beta)$ through a monotone function $g(\cdot)$, e.g., a linear transformation, so that we have $\hat{\theta}_n \equiv g(\hat{\beta}_n)$ and bootstrap replications $\hat{\theta}_n^{*b} \equiv g(\hat{\beta}_n^{*b})$.
- The percentile CIs are transformation invariant: the τ-th quantile of the transformed replications is $\hat{\theta}_n^*(\tau) \equiv g(\hat{\beta}_n^*(\tau))$, so that the percentile CI for θ with $1-\alpha$ coverage is:

$$\left[g\left(\hat{\beta}_n^*(\alpha/2)\right), g\left(\hat{\beta}_n^*(1-\alpha/2)\right)\right].$$

Bootstrap Percentile Confidence Intervals

- A percentile CI works well if there exists a monotone $g(\cdot)$ such that this transformation yields an estimator of θ which is unbiased, normally distributed and with constant variance:

$$\hat{\theta}_n \overset{D}{\sim} \mathcal{N}(\theta, \sigma^2_\theta).$$

- The percentile CI does not require actually knowing the transformation $g(\cdot)$, it only assumes its existence.
- It would then be true that:

$$\hat{\theta}^{*b}_n \overset{D}{\sim} \mathcal{N}(\hat{\theta}_n, \sigma^2_\theta).$$

- Then it must be that the percentile CI for θ:

$$\left[\hat{\theta}^*_n(\alpha/2), \hat{\theta}^*_n(1-\alpha/2)\right],$$

has **exact** $1-\alpha$ coverage.

- But, because of transformation invariance, the percentile intervals:

$$\left[g^{-1}\left(\hat{\theta}^*_n\left(\alpha/2\right)\right), g^{-1}\left(\hat{\theta}^*_n\left(1-\alpha/2\right)\right)\right],$$

for our original parameter β would also have **exact** $1-\alpha$ coverage.

Bootstrap Bias-Corrected Confidence Intervals

- Note that the existence of $g(\cdot)$ also implies median unbiasedness:

$$P\left[\hat{\theta}_n^{*b} \leq \hat{\theta}_n\right] = 0.50 \quad \Longrightarrow \quad P\left[\hat{\beta}_n^{*b} \leq \hat{\beta}_n\right] = 0.50.$$

- Median unbiasedness can be checked using the bootstrap replications. For example, using $B = 2,000$ and our class size coefficient we have:

$$F_{\hat{\beta}_n^*}\left(\hat{\beta}_n\right) \simeq P\left[\hat{\beta}_n^{*b} \leq \hat{\beta}_n\right] = 0.519,$$

  which is far enough from 0.50 to have a small impact on proper inference.

- It suggests that the median of replications $\hat{\beta}_n^{*b}$ is biased downward relative to $\hat{\beta}_n$, as 51.9% of the bootstrap probability lies below $\hat{\beta}_n$.
- If one takes the median of replications $\hat{\beta}_n^{*b}$ as an estimate of the bias in $\hat{\beta}_n$, by implication $\hat{\beta}_n$ is biased downward for estimating β. Accordingly, the percentile CI should be adjusted a little bit upward.

Bootstrap Bias-Corrected Confidence Intervals

- The **bias-corrected (BC) confidence interval** is a data-based algorithm for making such adjustments:

$$\left[\hat{\beta}_n^*(\alpha_1), \hat{\beta}_n^*(\alpha_2)\right],$$

where:

$$\alpha_1 \equiv \Phi\left(2\tilde{b} + z_{\alpha/2}\right), \qquad \alpha_2 \equiv \Phi\left(2\tilde{b} + z_{1-\alpha/2}\right),$$

and:

$$\tilde{b} \equiv \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_n^{*b} \leq \hat{\beta}_n)\right).$$

- In our case $\tilde{b} = \Phi^{-1}(0.519) = 0.04764396$.
- The basic idea is to replace $\alpha/2$ and $1-\alpha/2$ used in the percentile CI with "adjusted" quantiles α_1 and α_2.

Additional Fix-Ups

- Other fix-ups are available to account for changing variance (other than bias) in the distribution of $\hat{\theta}_n$ (**accelerated bias-corrected (BCa) confidence intervals**).
- Takeaway message: bootstrap CIs work better with statistics that are truly approximately normal, centred at zero and with constant variance. They provide a coverage probability which is closer to the nominal confidence level.
- An example is the t-statistic, which is asymptotically distributed as a standard normal and asymptotically pivotal (its asymptotic distribution does not depend on any unknown parameters).
- Regression coefficients are not asymptotically pivotal, as they have an asymptotic distribution which depends on the unknown residual variance.

Additional Fix-Ups

- **Bootstrap-t confidence intervals** are obtained by computing:

$$t_n^{*b} \equiv \frac{\hat{\beta}_n^{*b} - \hat{\beta}_n}{\sqrt{\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right)}},$$

where replications provide estimated percentiles $t_n^*(\tau)$ and corresponding confidence limits with $1-\alpha$ coverage:

$$\left[\hat{\beta}_n + t_n^*(\alpha/2)\sqrt{\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right)}, \hat{\beta}_n + t_n^*(1-\alpha/2)\sqrt{\widehat{Var}_{bootstrap}\left(\hat{\beta}_n\right)}\right].$$

- In our case we have $[-0.00308103, -0.00011428]$ for the coefficient on class size.

Clustering

The Clustering Problem in a Nutshell

- The clustering problem can be easily put across by considering a bivariate regression in data with a group structure:

$$Y_{ig} = \beta_0 + \beta_1 X_{ig} + e_{ig},$$

  where Y_{ig} and X_{ig} are defined for unit i in **cluster** g.

- For example, our class size application uses scores for students in the same group g, with $g = 1, \ldots, G$.
- In this setting we are concerned that data may not be independent across observations.
  - Test scores of students in the same class may be correlated because of common background characteristics and teachers.
  - The regressor of interest (class size) varies only at the group level, X_g, so that all students face the same input.
  - The random sampling model is unrealistic, as all students are typically sampled when the class is sampled.

Moulton (1986)

Econometricians frequently fit regression models to micro data that are drawn from populations with a grouped structure. Examples of grouping factors would include geographical location, industry, occupation, and years of schooling. It is usually necessary to take account of the grouping either in the specification of the regressors or in the stochastic structure of the errors.

Confusion:

- Why do we cluster by state, but not by age or gender?
- How can one justify the clustering adjustment for one but not for the other?

Misconceptions

- Clustering depends on a common unobserved random shock that leads to within-group correlation of outcomes or errors.
  - This makes it difficult to justify clustering on some partitioning of the population, but not on others.
- "We should cluster if it makes a difference."
- "We should keep clustering at higher levels until the standard errors do not change much."

The Clustering Problem in a Nutshell

- To understand the implications of clustering, I will take a stand on the origin of the within-cluster correlation structure:

$$\frac{Cov\left(e_{ig}, e_{jg}\right)}{\sqrt{Var\left(e_{ig}\right) Var\left(e_{jg}\right)}}, \quad \forall i \neq j \quad \forall g.$$

- I assume that (**Moulton's structure**):
  - errors have constant variance, σ_e^2, across clusters;
  - error pairs are equicorrelated within clusters, but uncorrelated across clusters (reasonable when observations are "exchangeable", e.g., members of the same household);
  - the within-cluster correlation is the same across clusters.
- The common correlation assumption implies that the dependence on g can be suppressed, so that the **intra-cluster correlation of residuals** becomes:

$$\rho_e \equiv \frac{Cov\left(e_{ig}, e_{jg}\right)}{\sigma_e^2}, \quad \forall i \neq j \quad \forall g.$$

The Clustering Problem in a Nutshell

- The variance-covariance matrix of residuals in the cluster is:

$$\sigma_e^2 \begin{pmatrix} 1 & \rho_e & \rho_e & \cdots & \rho_e \\ & 1 & \rho_e & \cdots & \rho_e \\ & & \ddots & \vdots & \\ & & & & 1 \end{pmatrix}.$$

- There are $n_g(n_g - 1)$ elements off the main diagonal, n_g is the size of cluster g, $n = \sum_{g=1}^{G} n_g$ and $\bar{n}$ is the average cluster size.
- Calculations show that:

$$\frac{Var_{cluster}\left(\hat{\beta}_1\right)}{Var_{standard}\left(\hat{\beta}_1\right)} = 1 + \left[\frac{Var(n_g)}{\bar{n}} + (\bar{n} - 1)\right]\rho_e\rho_X,$$

  where ρ_X is the **intra-cluster correlation of the regressor** X_{ig}.

- Covariances in ρ_e and ρ_X can be estimated from $\frac{1}{2}\sum_{g=1}^{G} n_g(n_g - 1)$ observations (pairwise estimator), and variances from n observations.

Taking Stock

- What can we learn from this formula?
  - Clustering has a bigger impact on standard errors when group size varies.
  - Use as many clusters as possible (e.g., think of $G \to n$).
  - We should worry most about clustering when the regressor of interest is fixed within groups ($\rho_X = 1$).
- If X_{ig} varies at the cluster level of aggregation (e.g., class, school or family), use **cluster averages** instead of micro data: this makes clear that asymptotics are based on the number of clusters (class, not student, is the unit of observation in my example).
- Because cluster averages are close to normally distributed with modest group sizes, you can expect the good finite-sample properties of regression with normal errors to kick in.
- The standard errors that come out of grouped estimation are therefore likely to be more reliable than clustered standard errors in samples with few clusters.

Clustering and Random Effects

- The correlation within groups is often modeled using an additive **random effects** structure (also known as the components of variance model):

$$e_{ig} = \alpha_g + u_{ig},$$

  where α_g is a mean-zero random component specific to class g which is uncorrelated with the mean-zero idiosyncratic student-level component u_{ig}.

- This implies:

$$\rho_e \equiv \frac{Var(\alpha_g)}{Var(\alpha_g) + Var(u_{ig})}.$$

- This quantity is easy to compute using ANOVA.

Example of Stata Output

One-way Analysis of Variance for ehat: Residuals

Number of obs = **87,498**
R-squared = **0.2409**

Source	SS	df	MS	F	Prob > F
Between pippo	**5305.0722**	**3,529**	**1.5032792**	**7.55**	**0.0000**
Within pippo	**16716.938**	**83,968**	**.19908701**		
Total	**22022.01**	**87,497**	**.25168874**		

Intraclass correlation	Asy. S.E.	[95% Conf.	Interval]
0.20905	**0.00496**	**0.19932**	**0.21878**

Estimated SD of pippo effect **.2293885**
Estimated SD within pippo **.4461917**
Est. reliability of a pippo mean **0.86756**
(evaluated at n= **24.79**)

Example of Stata Output

One-way Analysis of Variance for clsize_snv: Class size

Number of obs = **87,498**
R-squared = **0.2630**

Source	SS	df	MS	F	Prob > F
Between schoolid	**285798.12**	**3,529**	**80.985583**	**8.49**	**0.0000**
Within schoolid	**800738.78**	**83,968**	**9.5362374**		
Total	**1086536.9**	**87,497**	**12.41799**		

Intraclass correlation	Asy. S.E.	[95% Conf.	Interval]
0.23212	**0.00527**	**0.22179**	**0.24246**

Estimated SD of schoolid effect **1.697853**
Estimated SD within schoolid **3.08808**
Est. reliability of a schoolid mean **0.88225**
(evaluated at n= **24.79**)

Cluster-Robust Standard Errors

- When schools are used as clusters in my example, the data show that $Var(n_g) = 11.33391$ and $\bar{n} = 27.98736$ (note that clustering here is across grades and over time).
- Combined with $\rho_e = 0.20905$ and $\rho_X = 0.23212$, the formula discussed above yields a value of 2.329204 for the inflation of variances, so that conventional standard errors should be expected to change by a factor of $1.526173 = \sqrt{2.329204}$.
- **Cluster-robust standard errors** are an easy way to account for possible issues related to clustered data if you do not want to model intra-cluster correlation and heteroskedasticity (**Liang and Zeger's structure**): variance-covariance matrix is block diagonal, blocks correspond to clusters, no restrictions on blocks.

Example of Stata Output

```
Linear regression                               Number of obs     =        87,498
                                                F(4, 3529)        =        126.90
                                                Prob > F          =        0.0000
                                                R-squared         =        0.0068
                                                Root MSE          =        .50005

                                   (Std. Err. adjusted for    3,530 clusters in schoolid)
------------------------------------------------------------------------------
             |               Robust
answers_ma~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  clsize_snv |  -.0015838   .0007827    -2.02   0.043    -.0031184   -.0000491
             |
      survey |
        2010 |   .0230391   .0053217     4.33   0.000     .0126052    .033473
        2011 |  -.0215401   .0054278    -3.97   0.000     -.032182   -.0108981
             |
     1.grade |  -.0723657   .0036318   -19.93   0.000    -.0794862   -.0652451
       _cons |  -.0074777   .0163333    -0.46   0.647    -.0395014    .024546
------------------------------------------------------------------------------
```

Cluster-Robust Standard Errors

- **Cluster-robust Std. Err.** is the square root of the K diagonal elements of:

$$\widehat{Var}_{cluster\text{-}robust}\left(\hat{\beta}_n\right) = q\left[\sum_{i=1}^{n} X_i X_i'\right]^{-1}\left(\sum_{g=1}^{G} \mathbf{X}_g \mathbf{e}_g \mathbf{e}_g' \mathbf{X}_g\right)\left[\sum_{i=1}^{n} X_i X_i'\right]^{-1},$$

  where $\mathbf{X}_g$ is the $k \times n_g$ matrix of regressors for units in cluster g (i.e., the matrix with columns the vectors X_{ig}'s), $\mathbf{e}_g$ is the $n_g \times 1$ vector of residuals for units in cluster g, and q is a degrees of freedom correction because now any approximation holds when G grows large.

- The clustered variance estimator is **consistent as the number of groups gets large** under any within-group correlation structure (not just the random effects model).
- Clustered standard errors are therefore unlikely to be reliable with few clusters: **use wild bootstrap in this case**.

When Should You Cluster Standard Errors?

The clustering of standard errors follows from the **design**.

- You should assess whether the sampling process is clustered or not, and whether the assignment mechanism is clustered.
- Research design reason? Arises if assignment to "treatment" is clustered (e.g., class size is assigned at the class level, and lives off this variability). In my example, clustering on schools is also reasonable because of monitors.
- Sampling design reason? Arises if units in groups (e.g., schools, villages) are sampled using clustered sampling, and we want to say something about the broader population (e.g., there are schools or villages in the population of interest beyond those seen in the sample).
- If the answer to both is no, you should not adjust.

When Should You Cluster Standard Errors?

- In the regression model considered earlier, consider the case of one binary X_{ig} (e.g., treatment status).
- The quantities needed to decide about clustering are:
  - $\bar{X}_g$: share of treated units within clusters. This quantity depends on clustering in assignment.
  - q: share of clusters observed in the sample. This quantity comes from clustering in sampling.
- The quantity that matters is:

$$\lambda \underbrace{\widehat{Var}_{\text{cluster-robust}}\left(\hat{\beta}_n\right)}_{Liang-Zeger} + (1-\lambda) \underbrace{\widehat{Var}_{\text{robust}}\left(\hat{\beta}_n\right)}_{Sandwich},$$

where:

$$\lambda = 1 - q\frac{\left(E\left[\bar{X}_g(1-\bar{X}_g)\right]\right)^2}{E\left[\bar{X}_g^2(1-\bar{X}_g)^2\right]}.$$

When Should You Cluster Standard Errors?

% Sampled Clusters	Treatment Assignment		
	Random	**Clustered**	**Partially Clustered**
	($\bar{X}_g$ constant)	($\bar{X}_g \in \{0, 1\}$)	($\bar{X}_g \in (0, 1)$)
$\mathbf{q = 1}$	Sandwich	Liang-Zeger	Liang-Zeger
$\mathbf{q \to 0}$	Liang-Zeger	Liang-Zeger	
$\mathbf{q \in (0, 1)}$		Liang-Zeger	

Takeaways

- Uncertainty comes at least partly, and at times entirely, from the **assignment process** rather than from **sampling**.
- The econometrics literature traditionally (mistakenly) has focussed on sampling based uncertainty, which leads to confusion and incorrect standard errors.
- Random sampling and random assignment? Should not cluster and use robust standard errors instead. Clustering standard errors can be unnecessarily over-conservative.
- Random sampling and clustered assignment (fixed within clusters)? Should cluster.
- What if assignment is not perfectly correlated within clusters? Both robust and clustered standard errors are wrong.
- In panel data any periods after the first period do not have random sampling and, often, treatment remains the same for the unit over time.

Testing Statistical Hypotheses in the Twenty-First Century

Simple Hypotheses and Empirical Nulls

- Consider the classical single-test situation on the parameters of a multivariate regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i.$$

- In our example scores (Y_i) are explained by class size (X_{1i}), controlling for census year (X_{2i} and X_{3i}) and grade (X_{4i}) dummies.
- Our primary interest lies in testing the null hypothesis on the **significance** class size parameter:

$$H_0 : \beta_1 = 0.$$

- We want to see if class size determines scores Y_i net of the stratification defined by grade and census year (i.e., holding these two variables fixed).
- $\hat{\beta}_1$ is the OLS estimate of the class size coefficient.

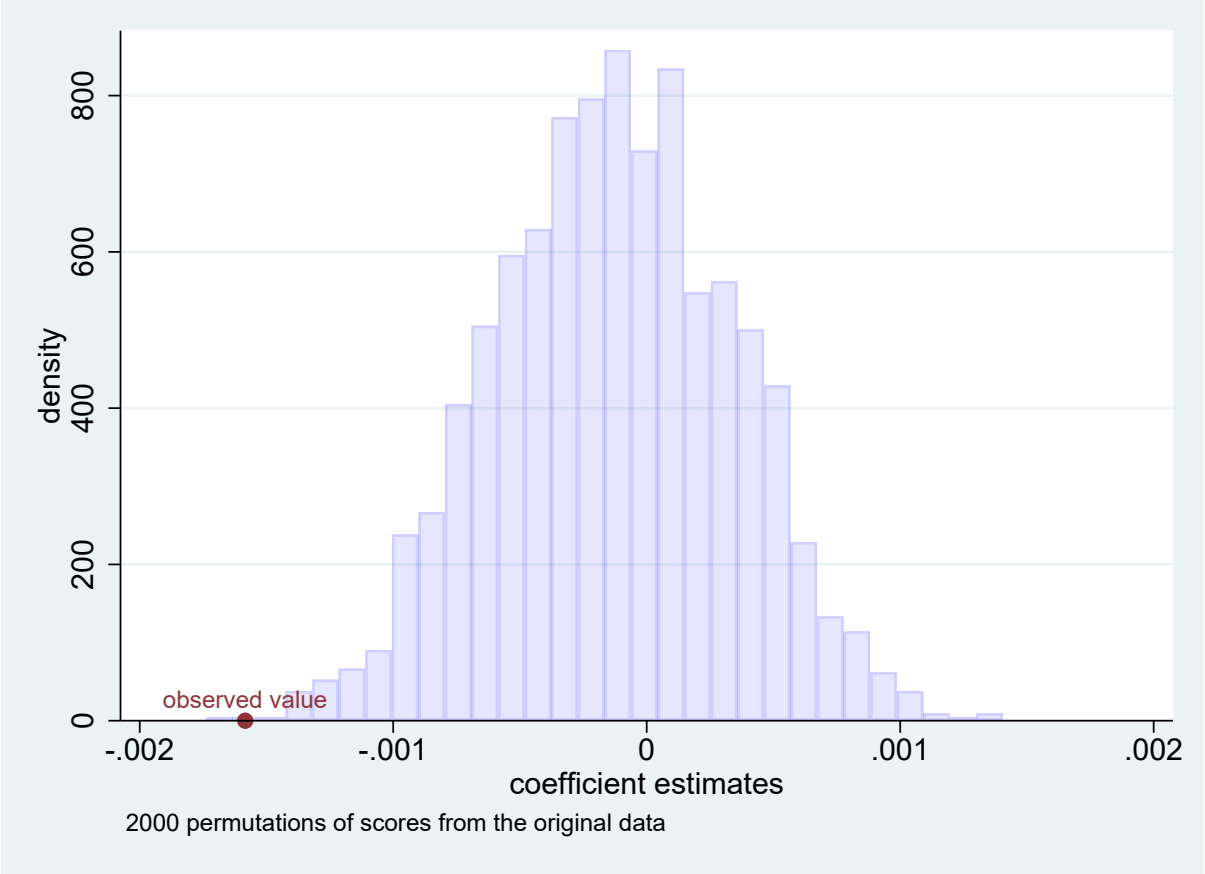
Simple Hypotheses and Empirical Nulls

- One idea is to fit this model:

$$Y_i^* = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i,$$

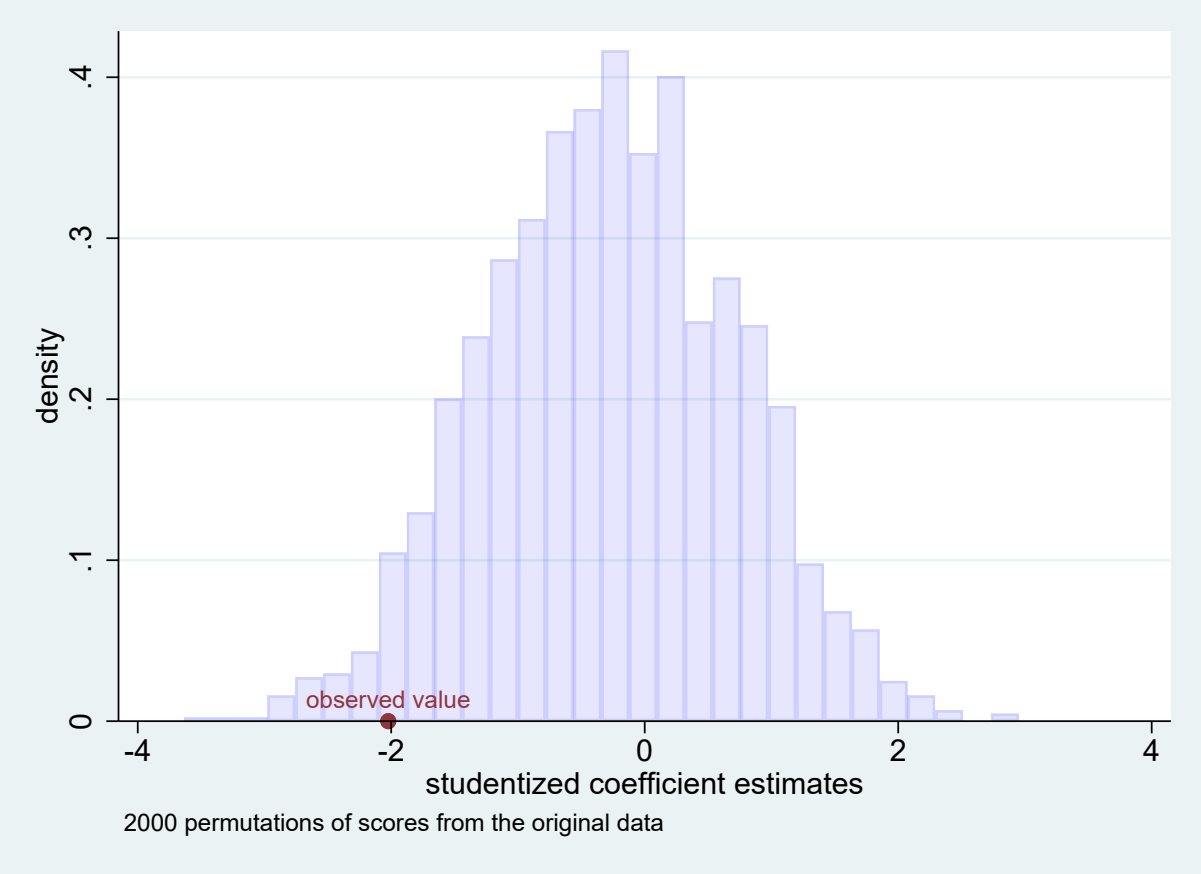
  where Y_i^* are permutations of the original Y_i's within strata defined by grade and census year (**permutation inference** is used in much empirical work).
- This generates data from the **sharp null (or constrained) model**: for any combination defined by values of X_{2i}, X_{3i} and X_{4i}, scores are by construction matched to class size at random.
- Note that the possible correlation between Y_i and (X_{2i}, X_{3i}, X_{4i}) is preserved in this simulation
- **Empirical nulls**, illustrated in the next figures, use the study's own data to estimate an appropriate null distribution of estimates $\hat{\beta}_1^*$ obtained across permutation samples.
- They represent the sampling distribution of the test statistic when the null hypothesis is true, and play the role of devil's advocate.

Example of Stata Output



Distribution of $\hat{\beta}_1^*$ from $2,000$ permutations. The share of values for which $|\hat{\beta}_1^*| \geq \hat{\beta}_1$ is 0.0010.

Example of Stata Output



Distribution of $t^* \equiv \hat{\beta}_1^* / \sqrt{\hat{Var}(\hat{\beta}_1^*)}$ from $2,000$ permutations. The share of values for which $|t^*| \geq \hat{\beta}_1 / \sqrt{\hat{Var}(\hat{\beta}_1)}$ is 0.0385.

P-Values

- If the null hypothesis is true, estimates from the shuffled data should look like those from real data apart from sampling variability.
- The ranking of the real test statistic among the shuffled test statistics gives the **p-value**. This summarizes the strength or weakness of the empirical evidence against the null hypothesis.
- **Interpretation**: the p-value is the probability of observing across repeated samples a value of the statistic as extreme as we did if the null hypothesis is true:

$$P_{H_0}\left(|\hat{\beta}_1^*| \geq \hat{\beta}_1\right), \qquad P_{H_0}\left(|\hat{\beta}_1^*/\sqrt{\hat{Var}(\hat{\beta}_1^*)}| \geq \hat{\beta}_1/\sqrt{\hat{Var}(\hat{\beta}_1)}\right).$$

- This means that small p-values are evidence against the null, and large p-values provide little evidence against the null.
- P-values serve as a universal language for hypothesis testing: if α denotes the **significance level** of the test, then H_0 is rejected if the p-value is lower than α; otherwise, H_0 is not rejected at the $100\alpha\%$ level.

Simple Hypotheses and Approximate Nulls

- Permutation tests and empirical nulls are useful when we do not know how to compute the distribution of a test statistic.
- A null distribution is not something one estimates in classic hypothesis testing theory: normal approximations provide the null, which we must use for better or worse.
- For the $K \times 1$ vector of estimates we know that:

$$\hat{\beta} \xrightarrow{D} \mathcal{N}\left(\beta, \frac{1}{n} Var(\hat{\beta})\right).$$

- Suppose we are interested in a particular linear combination of the elements of β, say $c'\beta$, where c is a $K \times 1$ vector. For example, the second element of β (the class size coefficient) is obtained from:

$$c \equiv [0, 1, 0, 0, 0]', \qquad c'\beta = \beta_1.$$

- Testing the significance of the class size parameter is equivalent to:

$$H_0 : c'\beta = 0.$$

Simple Hypotheses and Approximate Nulls

- The properties of normal random variables ensure that:

$$c'\hat{\beta} \xrightarrow{D} \mathcal{N}\left(c'\beta, \frac{1}{n}c'Var(\hat{\beta})c\right),$$

$$\frac{c'\hat{\beta} - c'\beta}{\sqrt{\frac{1}{n}c'Var(\hat{\beta})c}} \xrightarrow{D} \mathcal{N}(0,1).$$

- This defines the **t statistic**, whose asymptotic distribution does not depend on any unknown parameters (asymptotically pivotal).
- The **approximate p-value ($P > |t|$)** for the null considered is:

$$P_{H_0}\left(|\mathcal{N}(0,1)| \geq \frac{c'\hat{\beta}}{\sqrt{\frac{1}{n}c'Var(\hat{\beta})c}}\right).$$

- Theoretical null derivations like that for the t-statistic are gems of the statistical literature, as well as pillars of applied practice.

Example of Stata Output

```
Linear regression                               Number of obs     =        87,498
                                                F(4, 3529)        =        126.90
                                                Prob > F          =        0.0000
                                                R-squared         =        0.0068
                                                Root MSE          =        .50005

                                   (Std. Err. adjusted for    3,530 clusters in schoolid)
------------------------------------------------------------------------------
             |               Robust
answers_ma~d |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  clsize_snv |  -.0015838   .0007827    -2.02   0.043    -.0031184   -.0000491
             |
      survey |
        2010 |   .0230391   .0053217     4.33   0.000     .0126052    .033473
        2011 |  -.0215401   .0054278    -3.97   0.000     -.032182   -.0108981
             |
     1.grade |  -.0723657   .0036318   -19.93   0.000    -.0794862   -.0652451
       _cons |  -.0074777   .0163333    -0.46   0.647    -.0395014    .024546
------------------------------------------------------------------------------
```

Linear Combinations and Approximate Nulls

- Consider testing the restriction (η is a known value):

$$H_0 : c'\beta = \eta.$$

- The expressions above can be used to write the following p-value:

$$P_{H_0}\left(|\mathcal{N}(0,1)| \geq \frac{c'\hat{\beta} - \eta}{\sqrt{\frac{1}{n}c' Var(\hat{\beta})c}} \right).$$

- Any linear combination of elements of β is also obtained by setting a particular vector c. For example, the difference between the coefficients on survey year dummies, $\beta_2 - \beta_3$, is obtained from:

$$c \equiv [0, 0, 1, -1, 0]'.$$

- Computation of p-values is analogous to the calculations above.

Linear Combinations and Approximate Nulls

- Add and subtract $\beta_3 X_{2i}$ from the equation above to write:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + e_i, \\ &= \beta_0 + \beta_1 X_{1i} + (\beta_2 - \beta_3) X_{2i} + \beta_3 (X_{2i} + X_{3i}) + \beta_4 X_{4i} + e_i. \end{aligned}$$

- This means that the restriction $\beta_2 = \beta_3$ can be tested by looking at the significance of the coefficient on X_{2i} in this manipulated regression.
- **Predictions** and **standard errors of predictions** also can be obtained using suitable manipulations. The predicted value of Y_i at:

$$(X_{1i} = x_1, X_{2i} = x_2, X_{3i} = x_3, X_{4i} = x_4),$$

  can be obtained by setting $c \equiv [1, x_1, x_2, x_3, x_4]'$.
- It is easy to see that the prediction coincides with the intercept of a manipulated regression obtained by adding and subtracting $c'\beta$.

Example of Stata Output

Linear regression

Number of obs	= **87,498**
F(4, 3529)	= **126.90**
Prob > F	= **0.0000**
R-squared	= **0.0068**
Root MSE	= **.50005**

(Std. Err. adjusted for **3,530** clusters in schoolid)

answers_ma~d	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
clsize_snv	**-.0015838**	**.0007827**	**-2.02**	**0.043**	**-.0031184**	**-.0000491**
survey_2010	**.0445792**	**.005049**	**8.83**	**0.000**	**.0346799**	**.0544784**
survey_new	**-.0215401**	**.0054278**	**-3.97**	**0.000**	**-.032182**	**-.0108981**
1.grade	**-.0723657**	**.0036318**	**-19.93**	**0.000**	**-.0794862**	**-.0652451**
_cons	**-.0074777**	**.0163333**	**-0.46**	**0.647**	**-.0395014**	**.024546**

Multiple Linear Restrictions

- Frequently we wish to test hypotheses that involve multiple restrictions on the regression parameters.
- The **overall significance of the regression** is one notable example (seldom of interest if we are after the causal effect of one variable):

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0.$$

- Four restrictions must hold jointly under the null:

$$H_0 : \beta_1 = 0, \beta_1 = \beta_2, \beta_1 = \beta_3, \beta_1 = \beta_4.$$

- What test statistic can be used to detect rejection? Consider the $K \times Q$ matrix for the four restrictions:

$$\underset{K \times Q}{C} \equiv \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

The Wald Statistic

- The constraints under the null are equivalent to:

$$H_0 : C'\beta = \underset{Q\times 1}{0}.$$

- The properties of normal random variables ensure that:

$$\left(\frac{1}{n}CVar(\hat{\beta})C'\right)^{-1/2}\left(C'\hat{\beta} - C'\beta\right) \xrightarrow{D} \mathcal{N}\left(0, \underset{Q\times Q}{I}\right),$$

$$\underbrace{\left(C'\hat{\beta} - C'\beta\right)'\left(\frac{1}{n}CVar(\hat{\beta})C'\right)^{-1}\left(C'\hat{\beta} - C'\beta\right)}_{Wald\ \ Statistic} \xrightarrow{D} \chi^2_Q,$$

where χ^2_Q is a **Chi-squared random variable** with Q degrees of freedom.

- The last quantity is known as **Wald statistic** ($\mathcal{W}$), which is a quadratic form of the $Q \times 1$ vector $C'\hat{\beta}$ that depends only on estimates from the unrestricted model.

The Wald Statistic

- The important message here is that the statistic $\mathcal{W}$ is valid under heteroskedasticity or clustering since it re-weights estimates $C'\hat{\beta}$ using the appropriate OLS variance covariance matrix.
- The idea is to decide whether the difference between estimates from the unrestricted model ($C'\hat{\beta}$) and the restricted model ($C'\beta$) is large enough to warrant rejection.
- Different linear constraints can be included in the matrix C, meaning that this approach is not limited to the overall significance of the regression. For example, the restriction that all controls other than the class size coefficient are not significant can be tested using:

$$\underset{K\times Q}{C} \equiv \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

The F Statistic

- If a command in Stata reports significance levels using t statistics with H degrees of freedom, the **F statistic** is computed:

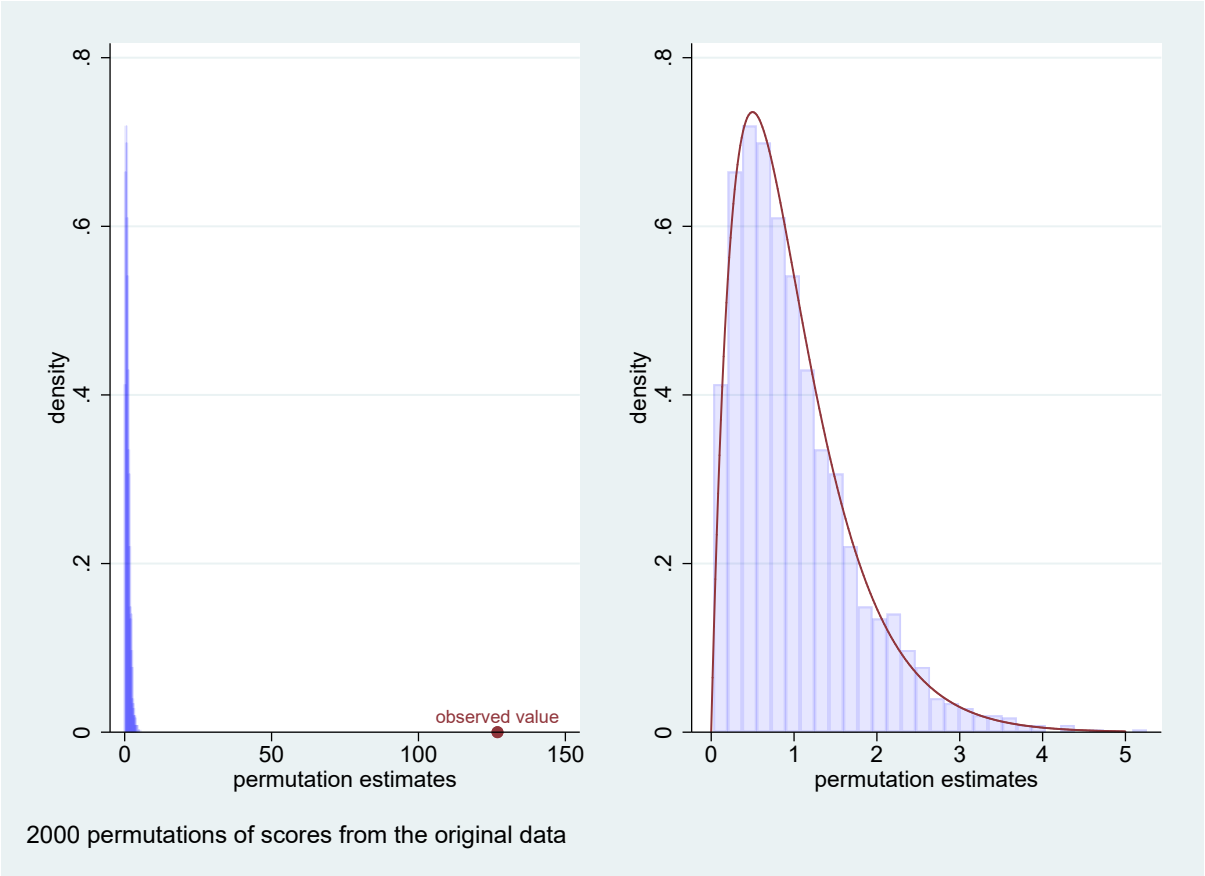
$$\mathcal{F} \equiv \frac{1}{Q}\mathcal{W}.$$

- To use the $\mathcal{F}$ statistic, we must know its sampling distribution under the null in order to set critical values and rejection rules.
- For example, the empirical null distribution for the overall significance of the regression can be obtained by permutations of the original Y_i's independently of all other variables:

$$P_{H_0}\left(|\mathcal{F}^*| \geq \mathcal{F}\right) = P_{H_0}\left(\mathcal{F}^* \geq \mathcal{F}\right).$$

- Nevertheless, it is appropriate and common to treat the $\mathcal{F}$ statistic as having an approximate **F random variable** with Q numerator degrees of freedom and H denominator degrees of freedom.

Example of Stata Output



Distribution of $\mathcal{F}^*$ from $2,000$ permutations. The share of values for which $|\mathcal{F}^*| \geq \mathcal{F}$ is 0.0000. The continuous line is the density of a F random variable with 4 and $3,529$ degrees of freedom.

The F Statistic

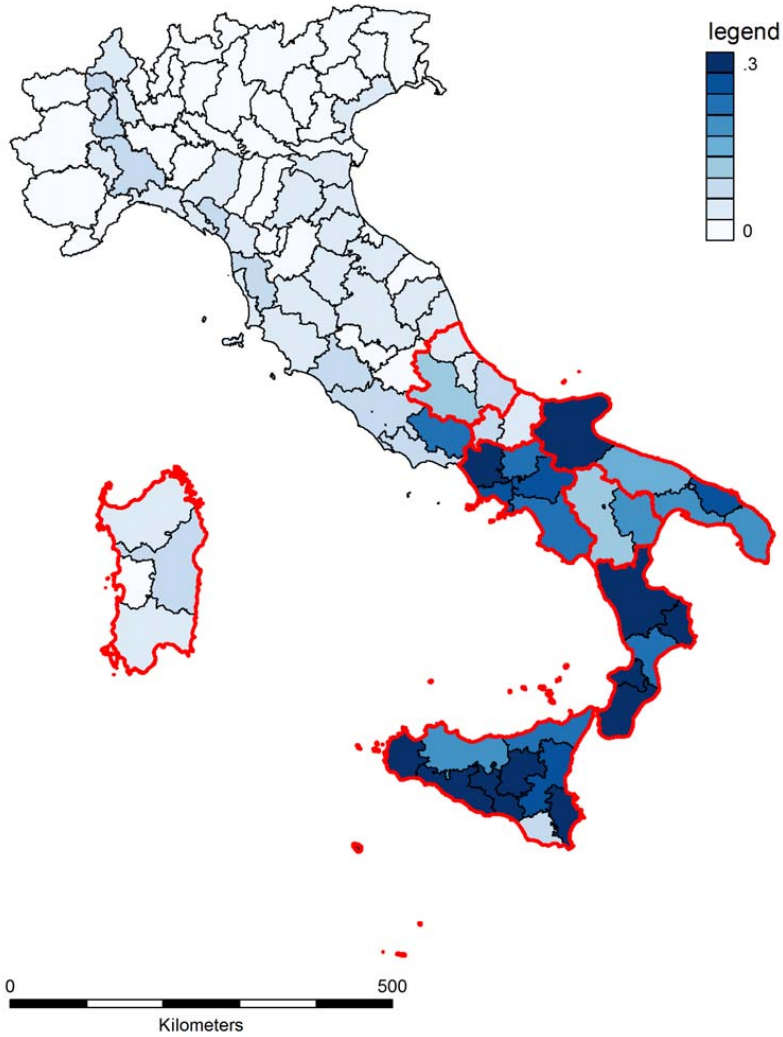
- You may recall that the statistic $\mathcal{F}$ is related to the difference of the R-squares or the sums of squared errors in the restricted and unrestricted models.
- This representation is old-fashioned and meaningful **only under homoskedasticity**.
- The omission of the ANOVA table when using robust or clustered standard errors is intentional in Stata.
- This is done because the sums of squares are no longer appropriate for use in the usual hypothesis tests, even though computationally the sums of squares remain the same.
- For example when you specify robust in a regression the meaning you might be tempted to give those sums is no longer relevant.
- The root MSE can no longer be used as an estimate for the variance because there is no longer a single variance to estimate: the variance of the residual varies observation by observation.

The F Statistic

- If homoskedasticity is violated the statistic $\mathcal{F}$ is no longer based on sums of squares but is rather a scaled $\mathcal{W}$ statistic based on the robustly estimated variance matrix.
- Under homoskedasticity, the quantity $\mathcal{F} = \frac{1}{Q}\mathcal{W}$ is numerically identical to the definition obtained using R-squares or sums of squared errors.

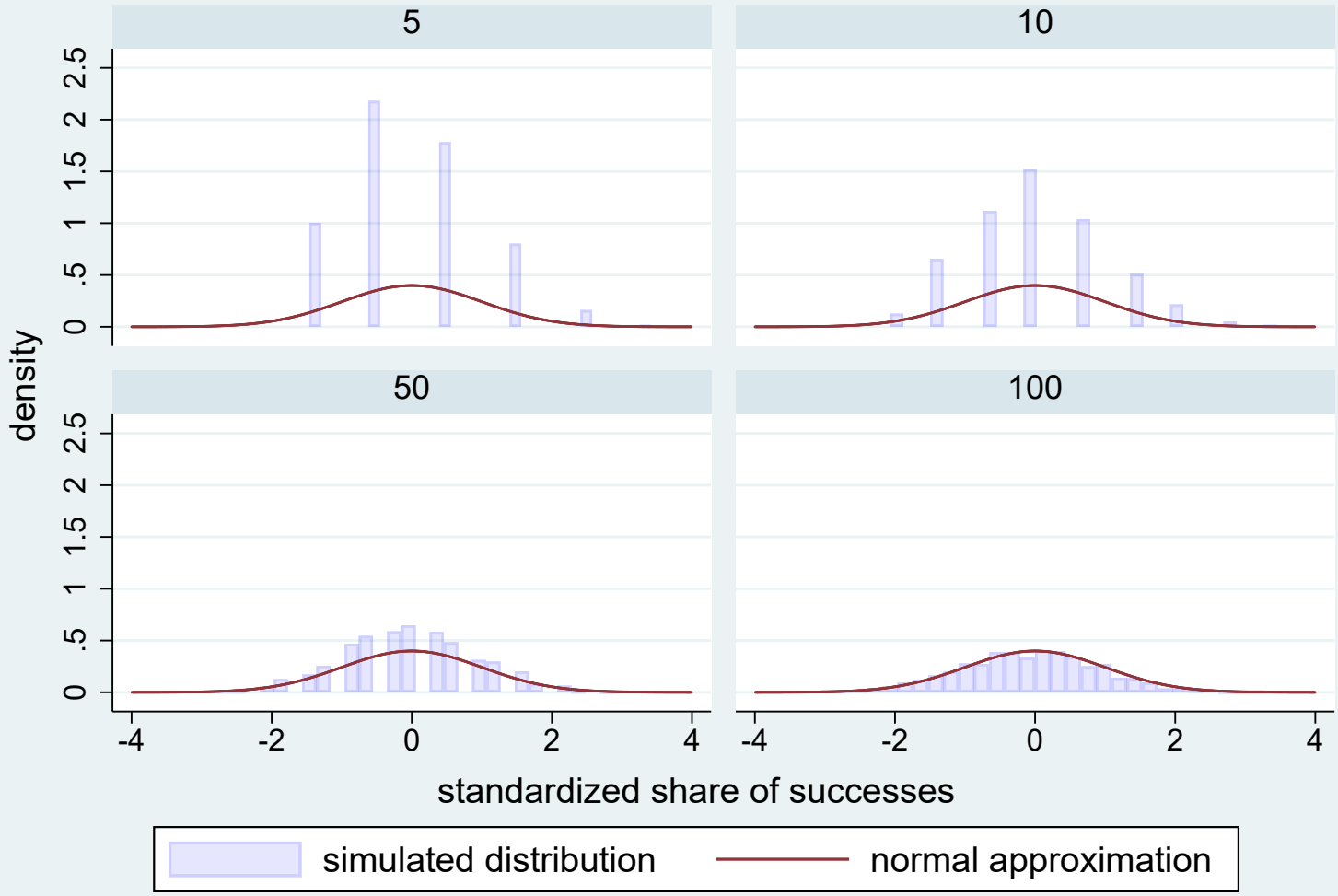
Additional Tables and Figures

Score Manipulation in Italy



Back

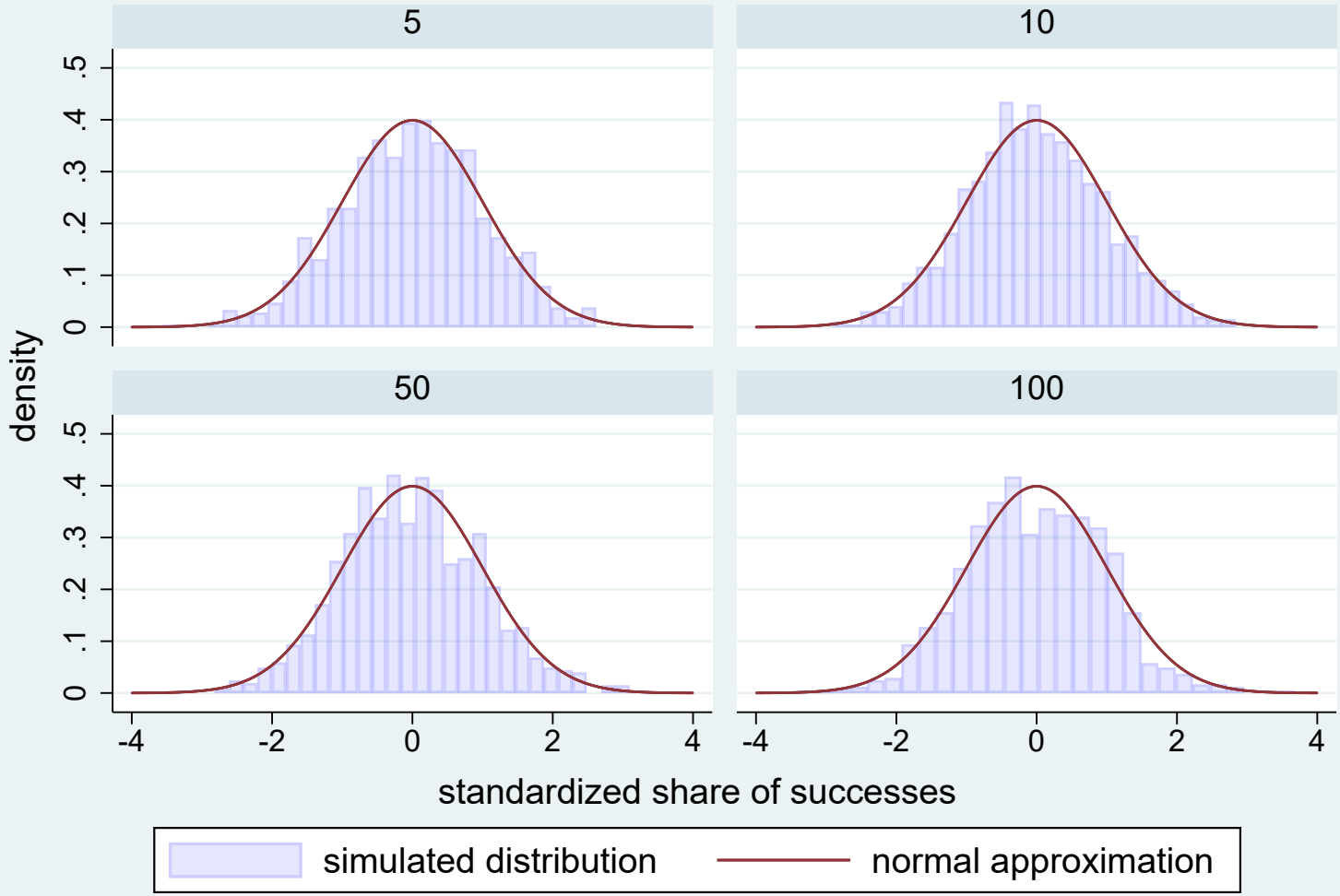
Monte Carlo Simulation from $\mathcal{B}(n, p)$



Graphs by number of trials (1,000 simulations)

Back

Monte Carlo Simulation from $\mathcal{N}(\mu, \sigma^2)$



Back

Convergence in Distribution

- The requirement that only the continuity points of Θ should be considered is essential. For example if $\mathcal{S}_n$ are uniform random variables over $(0, \frac{1}{n})$, we have that:

$$Pr[\mathcal{S}_n \leq s] = (ns)\, \mathbb{1}(s \in (0, \frac{1}{n})) + \mathbb{1}(s \in [\frac{1}{n}, +\infty)).$$

- This implies $Pr[\mathcal{S}_n \leq 0] = 0$ for all n.
- This sequence converges in distribution to a degenerate random variable $\Theta = 0$, for which $Pr[\Theta \leq 0] = 1$. Thus the convergence fails at the point $s = 0$ where Θ is discontinuous.
- This example shows why the formal definition requires convergence only at continuity points of Θ.

Back

Exact Within-Cluster Correspondence

- This happens when X_{ig} varies only across clusters.
- In this case we have $X_{ig} = X_g$ for all i's, and LIE implies:

$$\begin{aligned} Cov\left(X_{ig}, X_{jg}\right) &= E\left(X_{ig}X_{jg}\right) - E\left(X_{ig}\right)^2, \\ &= E\left[E\left(X_g^2 | G = g\right)\right] - E\left[E\left(X_g | G = g\right)\right]^2, \\ &= E\left(X_g^2\right) - E\left(X_g\right)^2, \\ &= Var\left(X_g\right), \\ Var\left(X_{ig}\right) &= E\left[Var\left(X_{ig} | G = g\right)\right] + Var\left[E\left(X_g | G = g\right)\right], \\ &= Var\left(X_g\right). \end{aligned}$$

- It follows that:

$$\rho_X \equiv \frac{Cov\left(X_{ig}, X_{jg}\right)}{\sqrt{Var\left(X_{ig}\right) Var\left(X_{jg}\right)}} = \frac{Cov\left(X_{ig}, X_{jg}\right)}{Var\left(X_{ig}\right)},$$

must be equal to one.

Back