Introduction to Econometrics
Third Edition
James H. Stock
Mark W. Watson

The statistical analysis of economic (and related) data

Brief Overview of the Course

Economics suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.

- What is the *quantitative* effect of reducing class size on student achievement?
- How does another year of education change earnings?
- What is the price elasticity of cigarettes?
- What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?
- What is the effect on housing prices of environmental improvements?

This course is about using data to measure causal effects.

- Ideally, we would like an experiment
  - What would be an experiment to estimate the effect of class size on standardized test scores?
- But almost always we only have observational (nonexperimental) data.
  - returns to education
  - cigarette prices
  - monetary policy
- Most of the course deals with difficulties arising from using observational to estimate causal effects
  - confounding effects (omitted factors)
  - simultaneous causality
  - "correlation does not imply causation"

In this course you will:

- Learn methods for estimating causal effects using observational data
- Learn some tools that can be used for other purposes; for example, forecasting using time series data;
- Focus on applications – theory is used only as needed to understand the whys of the methods;
- Learn to evaluate the regression analysis of others – this means you will be able to read/understand empirical economics papers in other econ courses;
- Get some hands-on experience with regression analysis in your problem sets.

Review of Probability and Statistics

(SW Chapters 2, 3)

Empirical problem: Class size and educational output

- Policy question: What is the effect on test scores (or some other outcome measure) of reducing class size by one student per class? by 8 students/class?
- We must use data to find out (is there any way to answer this *without* data?)

The California Test Score Data Set

All K-6 and K-8 California school districts (n = 420)

Variables:

- 5^{th} grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers

Initial look at the data:

(*You should already know how to interpret this table*)

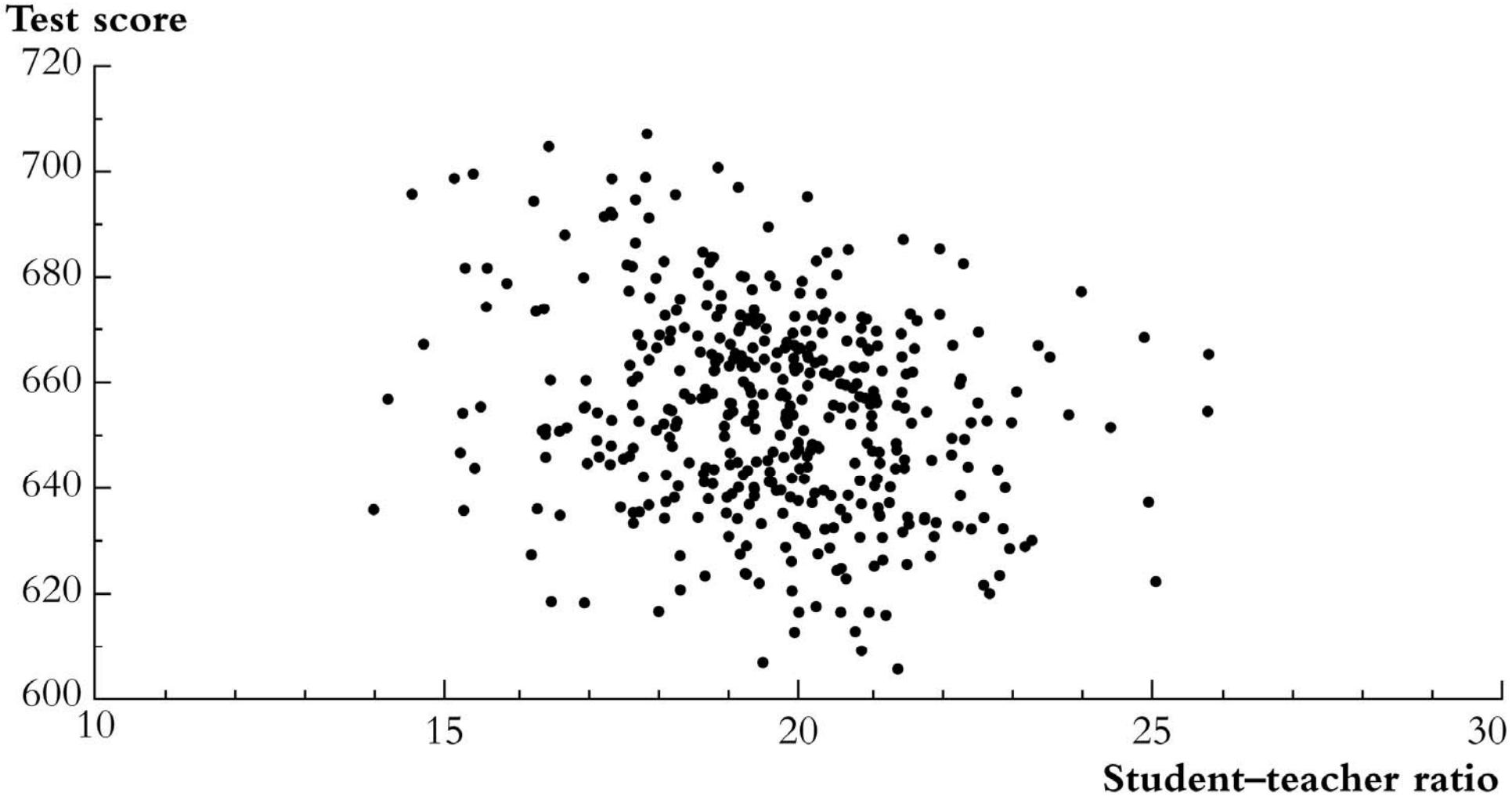
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			Percentile						
	Average	Standard Deviation	10%	25%	40%	50% (median)	60%	75%	90%
Student–teacher ratio	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
Test score	665.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

This table doesn't tell us anything about the relationship between test scores and the *STR*.

Do districts with smaller classes have higher test scores?

Scatterplot of test score v. student-teacher ratio



What does this figure show?

We need to get some numerical evidence on whether districts with low STRs have higher test scores – but how?

1. Compare average test scores in districts with low STRs to those with high STRs ("***estimation***")

2. Test the "null" hypothesis that the mean test scores in the two types of districts are the same, against the "alternative" hypothesis that they differ ("***hypothesis testing***")

3. Estimate an interval for the difference in the mean test scores, high v. low STR districts ("***confidence interval***")

Initial data analysis: Compare districts with "small" (STR < 20) and "large" (STR ≥ 20) class sizes:

Class Size	Average score ($\bar{Y}$)	Standard deviation (s_Y)	n
Small	657.4	19.4	238
Large	650.0	17.9	182

1. ***Estimation*** of Δ = difference between group means
2. ***Test the hypothesis*** that $\Delta = 0$
3. Construct a ***confidence interval*** for Δ

1. Estimation

$$\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = \frac{1}{n_{\text{small}}} \sum_{i=1}^{n_{\text{small}}} Y_i - \frac{1}{n_{\text{large}}} \sum_{i=1}^{n_{\text{large}}} Y_i$$

$$= 657.4 - 650.0$$

$$= 7.4$$

Is this a large difference in a real-world sense?

- Standard deviation across districts = 19.1
- Difference between 60$^{\text{th}}$ and 75$^{\text{th}}$ percentiles of test score distribution is 667.6 – 659.4 = 8.2
- This is a big enough difference to be important for school reform discussions, for parents, or for a school committee?

2. Hypothesis testing

Difference-in-means test: compute the *t*-statistic,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)} \qquad \text{(remember this?)}$$

where $SE(\bar{Y}_s - \bar{Y}_l)$ is the "standard error" of $\bar{Y}_s - \bar{Y}_l$, the subscripts *s* and *l* refer to "small" and "large" STR districts, and $s_s^2 = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2$ (etc.)

Compute the difference-of-means *t*-statistic:

Size	$\bar{Y}$	s_{Y}	n
small	657.4	19.4	238
large	650.0	17.9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

|t| > 1.96, so reject (at the 5% significance level) the null hypothesis that the two means are the same.

3. Confidence interval

A 95% confidence interval for the difference between the means is,

$$(\bar{Y}_s - \bar{Y}_l) \pm 1.96 \times SE(\bar{Y}_s - \bar{Y}_l)$$

$$= 7.4 \pm 1.96 \times 1.83 = (3.8, 11.0)$$

Two equivalent statements:

1. The 95% confidence interval for Δ doesn't include 0;
2. The hypothesis that $\Delta = 0$ is rejected at the 5% level.

What comes next…

- The mechanics of estimation, hypothesis testing, and confidence intervals should be familiar
- These concepts extend directly to regression and its variants
- Before turning to regression, however, we will review some of the underlying theory of estimation, hypothesis testing, and confidence intervals:
  - Why do these procedures work, and why use these rather than others?
  - We will review the intellectual foundations of statistics and econometrics

Review of Statistical Theory

1. **The probability framework for statistical inference**
2. Estimation
3. Testing
4. Confidence Intervals

The probability framework for statistical inference

(a) Population, random variable, and distribution

(b) Moments of a distribution (mean, variance, standard deviation, covariance, correlation)

(c) Conditional distributions and conditional means

(d) Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

(a) Population, random variable, and distribution

Population

- The group or collection of all possible entities of interest (school districts)
- We will think of populations as infinitely large (∞ is an approximation to "very big")

Random variable Y

- Numerical summary of a random outcome (district average test score, district STR)

Population distribution of Y

- The probabilities of different values of Y that occur in the population, for ex. Pr[$Y = 650$] (when Y is discrete)
- or: The probabilities of sets of these values, for ex. Pr[$640 \leq Y \leq 660$] (when Y is continuous).

(b) Moments of a population distribution: mean, variance, standard deviation, covariance, correlation

mean = expected value (expectation) of Y

$= E(Y)$

$= \mu_Y$

= long-run average value of Y over repeated realizations of Y

variance $= E(Y - \mu_Y)^2$

$= \sigma_Y^2$

= measure of the squared spread of the distribution

standard deviation $= \sqrt{\text{variance}} = \sigma_Y$

Moments, ctd.

$$\textbf{\textit{skewness}} = \frac{E\left[\left(Y - \mu_Y\right)^3\right]}{\sigma_Y^3}$$

= measure of asymmetry of a distribution

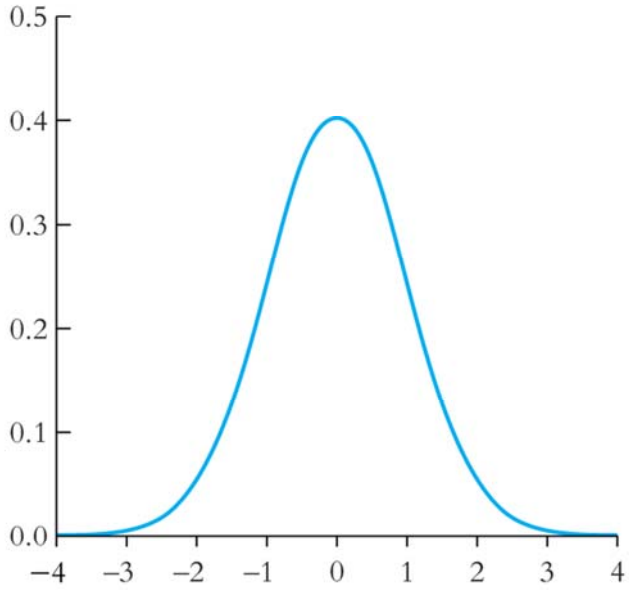
- *skewness* = 0: distribution is symmetric
- *skewness* > (<) 0: distribution has long right (left) tail

$$\textbf{\textit{kurtosis}} = \frac{E\left[\left(Y - \mu_Y\right)^4\right]}{\sigma_Y^4}$$

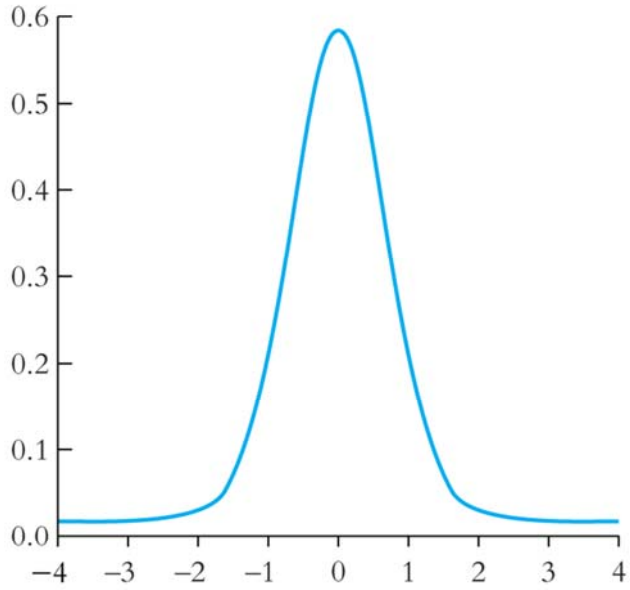
= measure of mass in tails

= measure of probability of large values

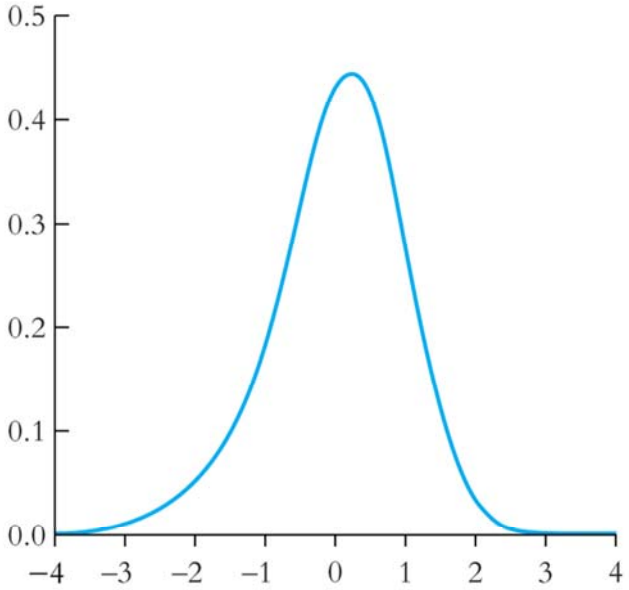
- *kurtosis* = 3: normal distribution
- *skewness* > 3: heavy tails ("***leptokurtotic*")**



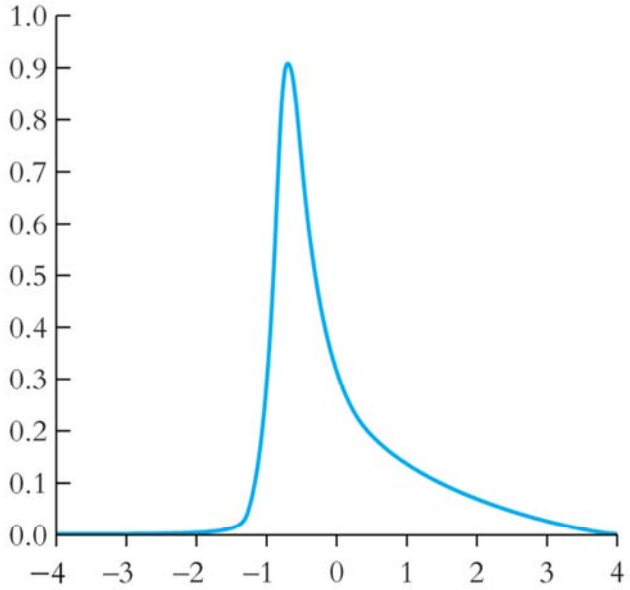
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = −0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

2 random variables: joint distributions and covariance

- Random variables X and Z have a ***joint distribution***
- The ***covariance*** between X and Z is

$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

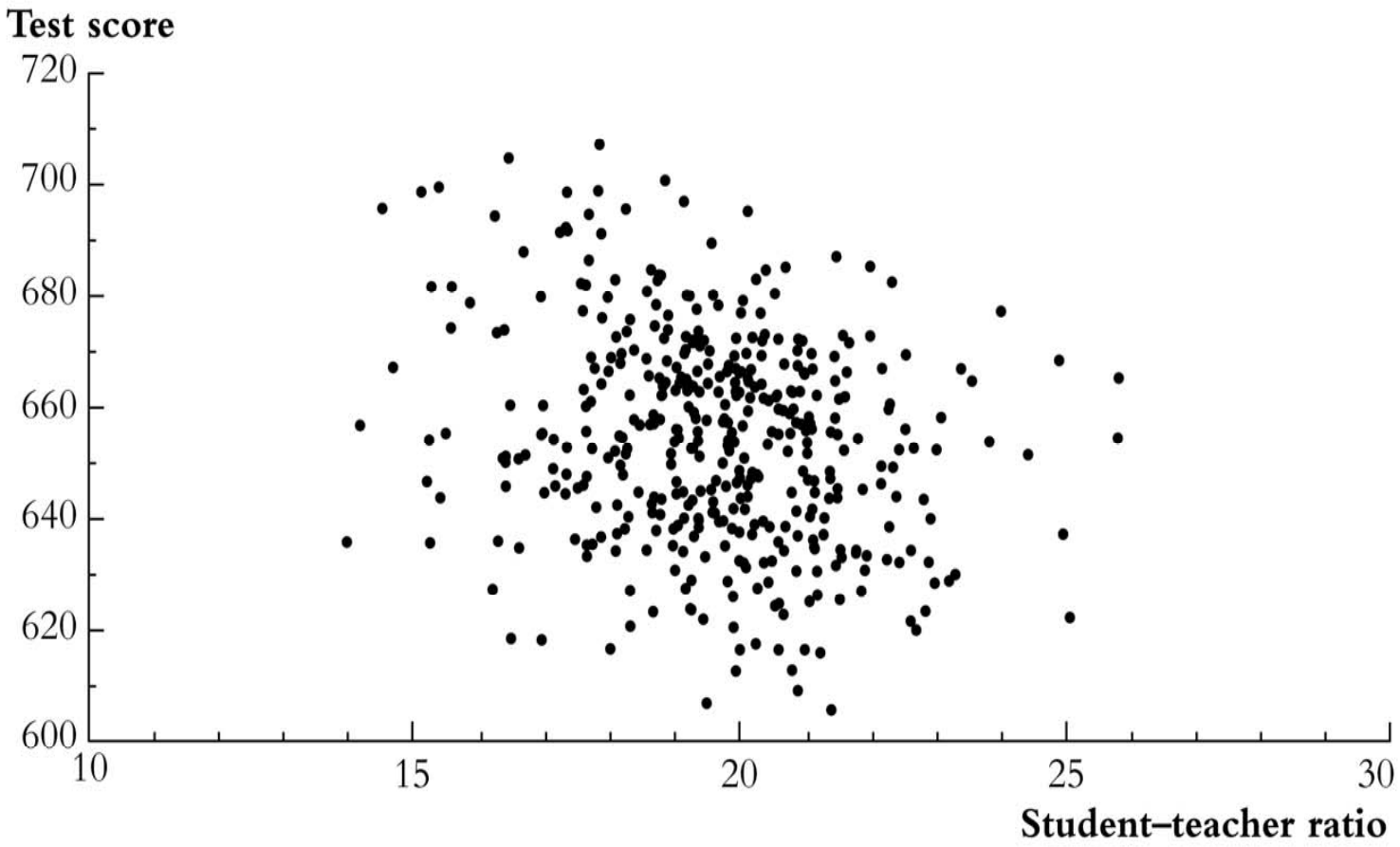
- The covariance is a measure of the linear association between X and Z; its units are units of X × units of Z
- $\text{cov}(X,Z) > 0$ means a positive relation between X and Z
- If X and Z are independently distributed, then $\text{cov}(X,Z) = 0$ (but not vice versa!!)
- The covariance of a r.v. with itself is its variance:

$$\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$$

The covariance between *Test Score* and *STR* is negative:

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



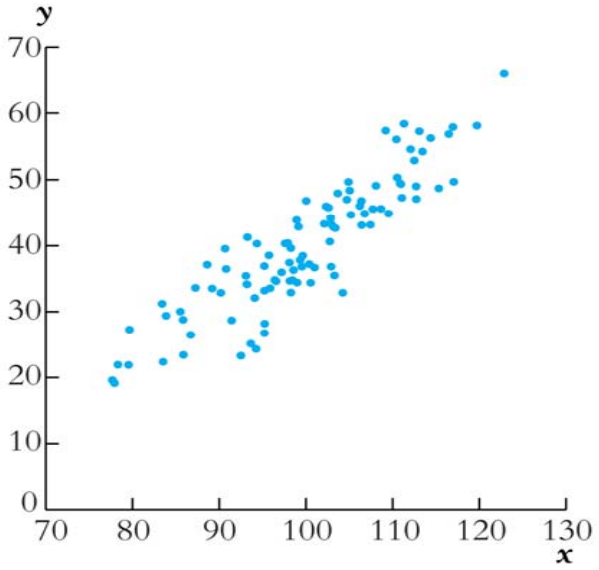
So is the ***correlation***…

The ***correlation coefficient*** is defined in terms of the covariance:

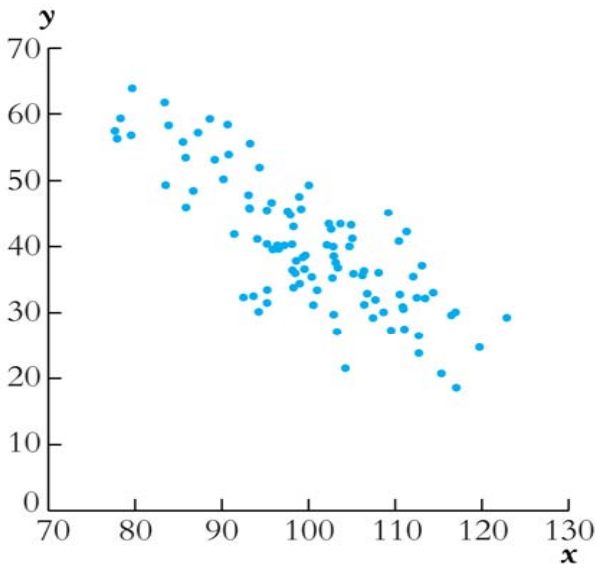
$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}.$$

- $-1 \le \text{corr}(X,Z) \le 1$
- $\text{corr}(X,Z) = 1$ mean perfect positive linear association
- $\text{corr}(X,Z) = -1$ means perfect negative linear association
- $\text{corr}(X,Z) = 0$ means no linear association

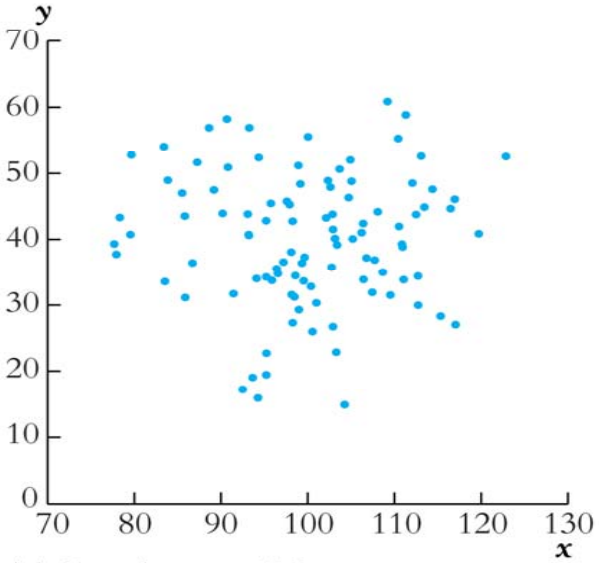
The correlation coefficient measures linear association



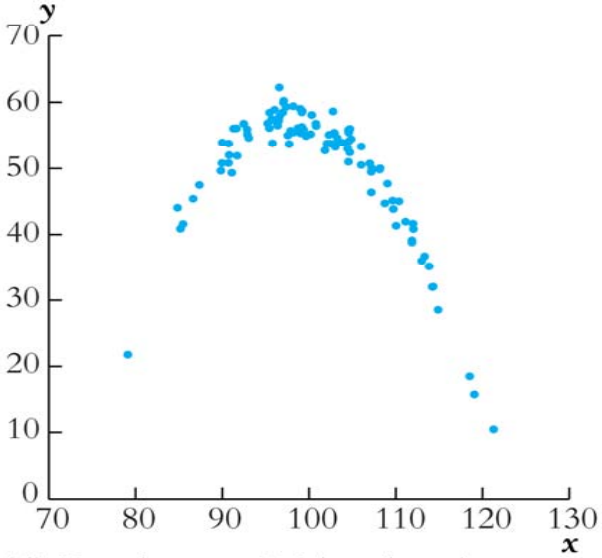
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

(c) Conditional distributions and conditional means

Conditional distributions

- The distribution of Y, given value(s) of some other random variable, X
- Ex: the distribution of test scores, given that STR < 20

Conditional expectations and conditional moments

- *conditional mean* = mean of conditional distribution
  $= E(Y|X = x)$ (***important concept and notation***)
- *conditional variance* = variance of conditional distribution
- *Example*: $E(Test\ scores|STR < 20)$ = the mean of test scores among districts with small class sizes

The difference in means is the difference between the means of two conditional distributions:

Conditional mean, ctd.

$\Delta = E(Test\ scores|STR < 20) - E(Test\ scores|STR \geq 20)$

Other examples of conditional means:

- Wages of all female workers (Y = wages, X = gender)
- Mortality rate of those given an experimental treatment (Y = live/die; X = treated/not treated)
- If $E(X|Z)$ = const, then corr(X,Z) = 0 (not necessarily vice versa however)

The conditional mean is a (possibly new) term for the familiar idea of the group mean

(d) Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

We will assume simple random sampling

- Choose and individual (district, entity) at random from the population

Randomness and data

- Prior to sample selection, the value of Y is random because the individual selected is random
- Once the individual is selected and the value of Y is observed, then Y is just a number – not random
- The data set is (Y_1, Y_2,…, Y_n), where Y_i = value of Y for the i^{th} individual (district, entity) sampled

Distribution of $Y_1,\ldots, Y_n$ under simple random sampling

- Because individuals #1 and #2 are selected at random, the value of Y_1 has no information content for Y_2. Thus:
  - Y_1 and Y_2 are ***independently distributed***
  - Y_1 and Y_2 come from the same distribution, that is, Y_1, Y_2 are ***identically distributed***
  - That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (***i.i.d.***).
  - More generally, under simple random sampling, $\{Y_i\}$, $i = 1,\ldots, n$, are i.i.d.

This framework allows rigorous statistical inferences about moments of population distributions using a sample of data from that population …

1. The probability framework for statistical inference
2. **Estimation**
3. Testing
4. Confidence Intervals

Estimation

$\bar{Y}$ is the natural estimator of the mean. But:

(a) What are the properties of $\bar{Y}$?

(b) Why should we use $\bar{Y}$ rather than some other estimator?

- Y_1 (the first observation)
- maybe unequal weights – not simple average
- median(Y_1,…, Y_n)

The starting point is the sampling distribution of $\bar{Y}$ …

(a) The sampling distribution of $\bar{Y}$

$\bar{Y}$ is a random variable, and its properties are determined by the ***sampling distribution*** of $\bar{Y}$

- The individuals in the sample are drawn at random.
- Thus the values of $(Y_1,\ldots, Y_n)$ are random
- Thus functions of $(Y_1,\ldots, Y_n)$, such as $\bar{Y}$, are random: had a different sample been drawn, they would have taken on a different value
- The distribution of $\bar{Y}$ over different possible samples of size n is called the ***sampling distribution*** of $\bar{Y}$.
- The mean and variance of $\bar{Y}$ are the mean and variance of its sampling distribution, $E(\bar{Y})$ and var($\bar{Y}$).
- The concept of the sampling distribution underpins all of econometrics.

The sampling distribution of $\bar{Y}$, ctd.

Example: Suppose Y takes on 0 or 1 (a ***Bernoulli*** random variable) with the probability distribution,

$$\Pr[Y = 0] = .22,\ \Pr(Y = 1) = .78$$

Then

$$E(Y) = p \times 1 + (1 - p) \times 0 = p = .78$$

$$\sigma_Y^2 = E[Y - E(Y)]^2 = p(1 - p) \quad \textit{[remember this?]}$$

$$= .78 \times (1 - .78) = 0.1716$$

The sampling distribution of $\bar{Y}$ depends on n.

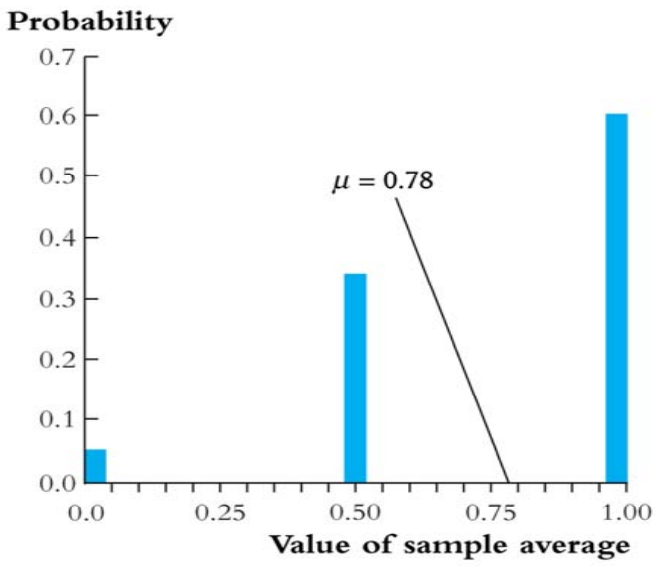
Consider $n = 2$. The sampling distribution of $\bar{Y}$ is,

$$\Pr(\bar{Y} = 0) = .22^2 = .0484$$

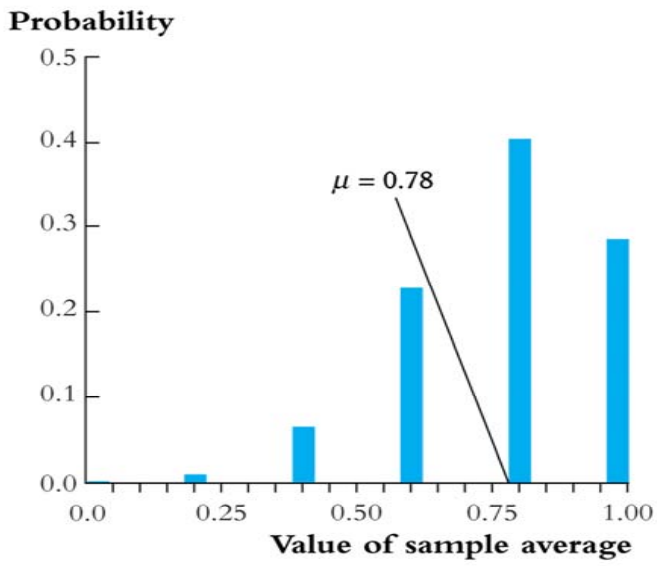
$$\Pr(\bar{Y} = ½) = 2 \times .22 \times .78 = .3432$$

$$\Pr(\bar{Y} = 1) = .78^2 = .6084$$

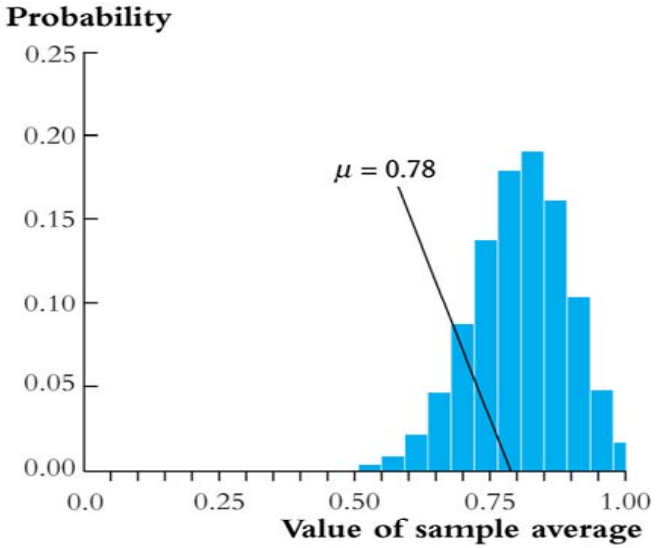
The sampling distribution of $\bar{Y}$ when Y is Bernoulli ($p = .78$):



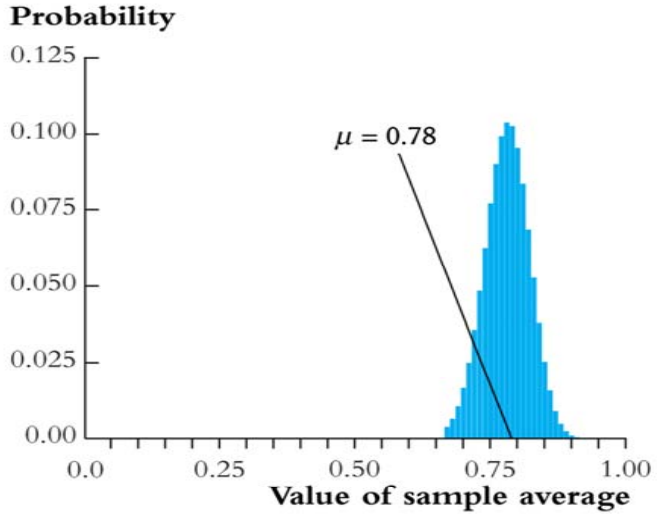
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

Things we want to know about the sampling distribution:

- What is the mean of $\bar{Y}$?
  - If $E(\bar{Y})$ = true μ = .78, then $\bar{Y}$ is an ***unbiased*** estimator of μ
- What is the variance of $\bar{Y}$?
  - How does var($\bar{Y}$) depend on n (famous $1/n$ formula)
- Does $\bar{Y}$ become close to μ when n is large?
  - Law of large numbers: $\bar{Y}$ is a ***consistent*** estimator of μ
- $\bar{Y} - \mu$ appears bell shaped for n large…is this generally true?
  - In fact, $\bar{Y} - \mu$ is approximately normally distributed for n large (Central Limit Theorem)

The mean and variance of the sampling distribution of $\bar{Y}$

General case – that is, for Y_i i.i.d. from any distribution, not just Bernoulli:

mean: $E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_Y = \mu_Y$

Variance: $\text{var}(\bar{Y}) = E[\bar{Y} - E(\bar{Y})]^2$

$$= E[\bar{Y} - \mu_Y]^2$$

$$= E\left[\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \mu_Y\right]^2$$

$$= E\left[\frac{1}{n}\sum_{i=1}^{n} (Y_i - \mu_Y)\right]^2$$

so

$$\text{var}(\bar{Y}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]^2$$

$$= E\left\{\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_Y)\right]\times\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu_Y)\right]\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}E\left[(Y_i - \mu_Y)(Y_j - \mu_Y)\right]$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{cov}(Y_i, Y_j)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2$$

$$= \frac{\sigma_Y^2}{n}$$

Mean and variance of sampling distribution of $\bar{Y}$, ctd.

$$E(\bar{Y}) = \mu_Y$$

$$\text{var}(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

Implications:

1. $\bar{Y}$ is an *unbiased* estimator of μ_Y (that is, $E(\bar{Y}) = \mu_Y$)
2. $\text{var}(\bar{Y})$ is inversely proportional to n
   - the spread of the sampling distribution is proportional to $1/\sqrt{n}$
   - Thus the sampling uncertainty associated with $\bar{Y}$ is proportional to $1/\sqrt{n}$ (larger samples, less uncertainty, but square-root law)

The sampling distribution of $\bar{Y}$ when *n* is large

For small sample sizes, the distribution of $\bar{Y}$ is complicated, but if *n* is large, the sampling distribution is simple!

1. As *n* increases, the distribution of $\bar{Y}$ becomes more tightly centered around μ_Y (the *Law of Large Numbers*)
2. Moreover, the distribution of $\bar{Y} - \mu_Y$ becomes normal (the *Central Limit Theorem*)

The *Law of Large Numbers*:

An estimator is ***consistent*** if the probability that its falls within an interval of the true population value tends to one as the sample size increases.

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $\sigma_Y^2 < \infty$, then $\bar{Y}$ is a consistent estimator of μ_Y, that is,

$$\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1 \text{ as } n \to \infty$$

which can be written, $\bar{Y} \xrightarrow{p} \mu_Y$

("$\bar{Y} \xrightarrow{p} \mu_Y$" means "$\bar{Y}$ converges in probability to μ_Y").

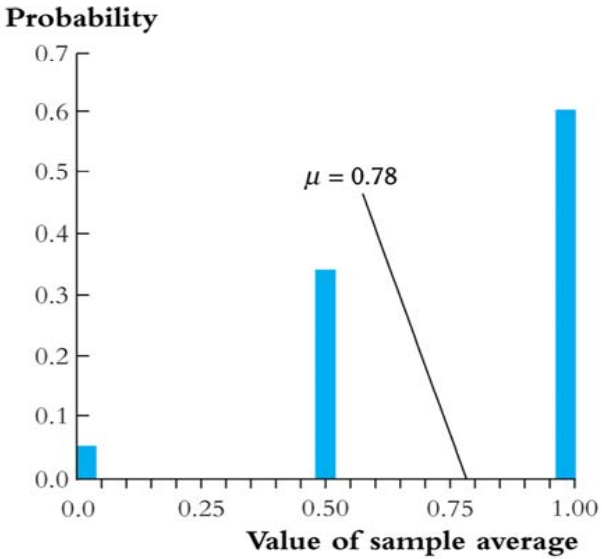
(*the math*: as $n \to \infty$, $\text{var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n} \to 0$, which implies that $\Pr[|\bar{Y} - \mu_Y| < \varepsilon] \to 1$.)

The *Central Limit Theorem* (CLT):

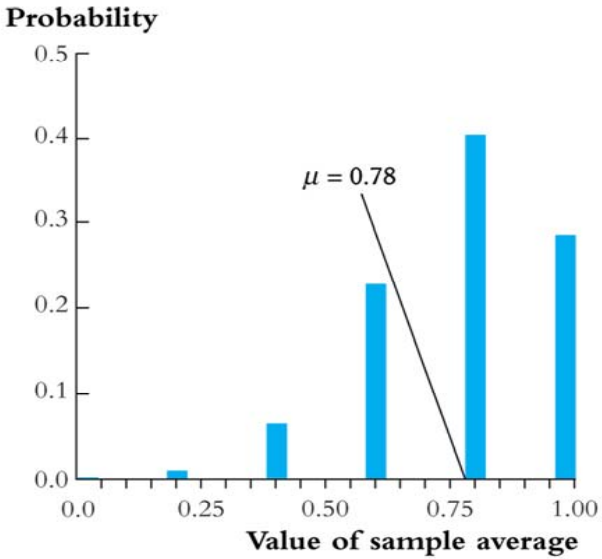
If $(Y_1,\ldots,Y_n)$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then when n is large the distribution of $\overline{Y}$ is well approximated by a normal distribution.

- $\overline{Y}$ is approximately distributed $N(\mu_Y, \frac{\sigma_Y^2}{n})$ ("normal distribution with mean μ_Y and variance σ_Y^2/n")
- $\sqrt{n}(\overline{Y} - \mu_Y)/\sigma_Y$ is approximately distributed $N(0,1)$ (standard normal)
- **That is, "standardized"** $\overline{Y} = \frac{\overline{Y} - E(\overline{Y})}{\sqrt{\text{var}(\overline{Y})}} = \frac{\overline{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$ **is approximately distributed as $N(0,1)$**
- **The larger is n, the better is the approximation.**

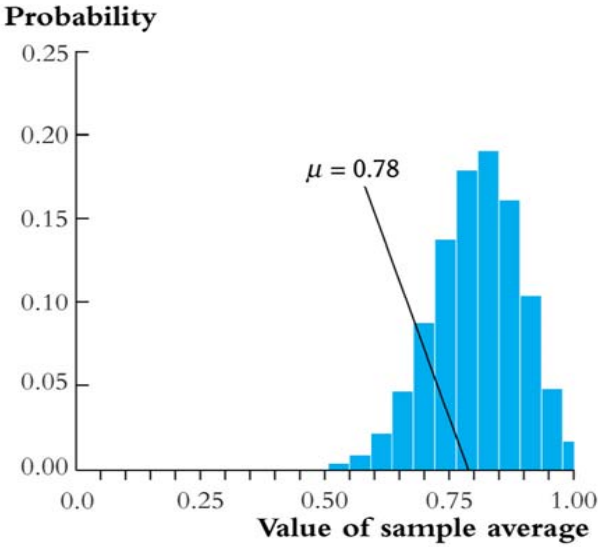
Sampling distribution of $\overline{Y}$ when Y is Bernoulli, $p = 0.78$:



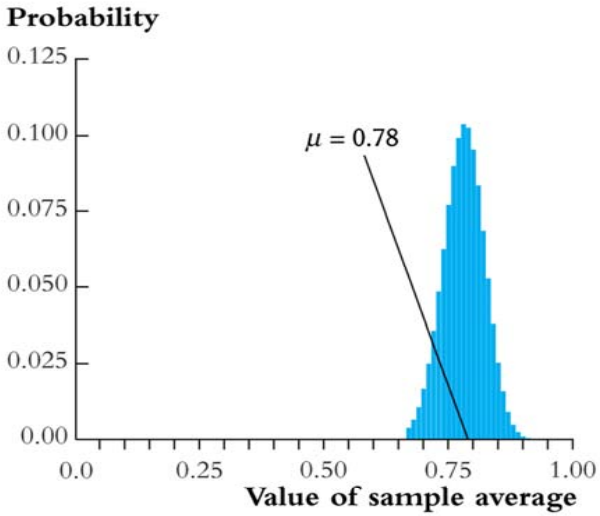
(a) $n = 2$



(b) $n = 5$

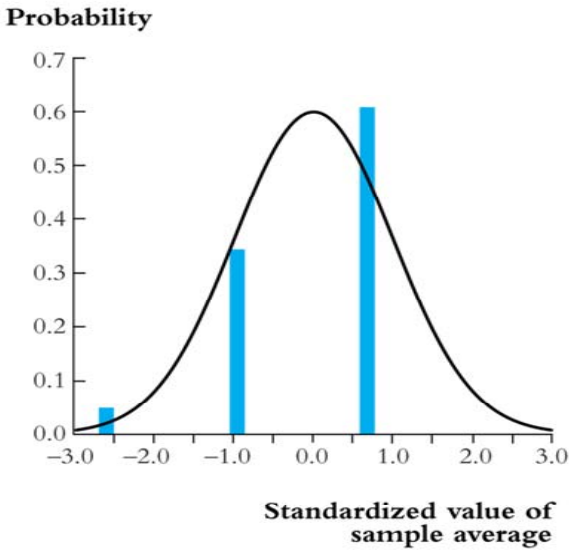


(c) $n = 25$

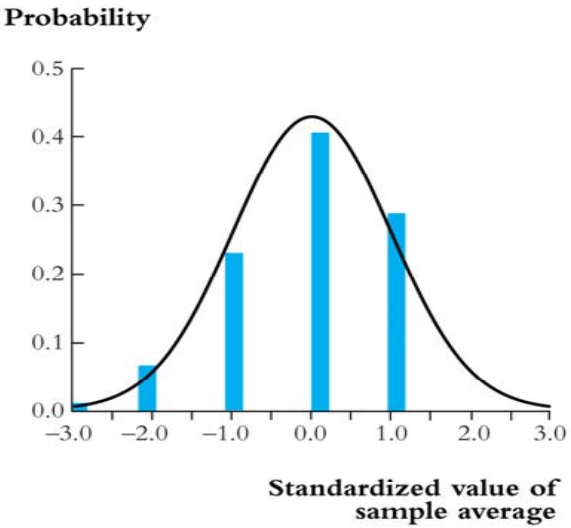


(d) $n = 100$

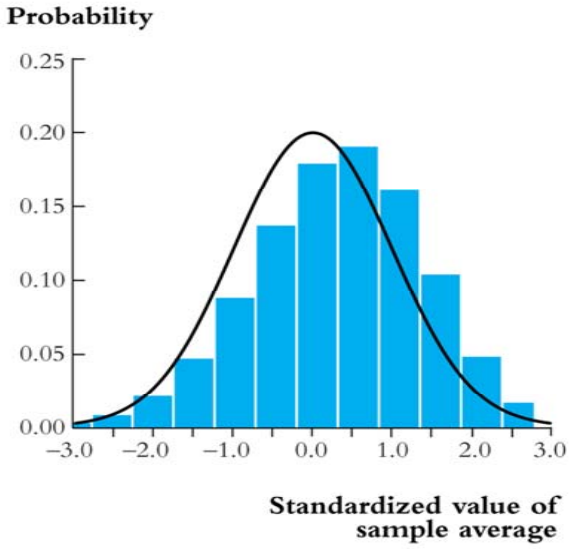
Same example: sampling distribution of $\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}}$:



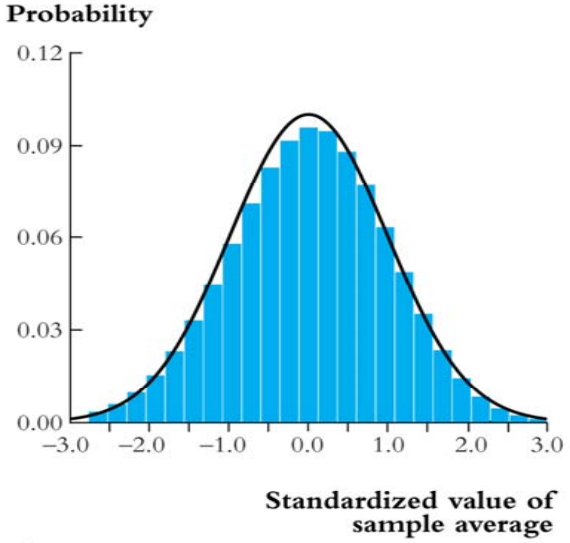
(a) $n = 2$



(b) $n = 5$



(c) $n = 25$



(d) $n = 100$

Summary: The Sampling Distribution of $\bar{Y}$

For $Y_1,\ldots,Y_n$ i.i.d. with $0 < \sigma_Y^2 < \infty$,

- The exact (finite sample) sampling distribution of $\bar{Y}$ has mean μ_Y ("$\bar{Y}$ is an unbiased estimator of μ_Y") and variance σ_Y^2/n
- Other than its mean and variance, the exact distribution of $\bar{Y}$ is complicated and depends on the distribution of Y (the population distribution)
- When n is large, the sampling distribution simplifies:
  - $\boxed{\bar{Y} \xrightarrow{p} \mu_Y}$ (Law of large numbers)
  - $\boxed{\dfrac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}} \text{ is approximately } N(0,1)}$ (CLT)

(b) Why Use $\bar{Y}$ To Estimate μ_Y?

- $\bar{Y}$ is unbiased: $E(\bar{Y}) = \mu_Y$
- $\bar{Y}$ is consistent: $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y}$ is the "least squares" estimator of μ_Y; $\bar{Y}$ solves,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

so, $\bar{Y}$ minimizes the sum of squared "residuals"

optional derivation (also see App. 3.2)

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i - m)^2 = \sum_{i=1}^{n}\frac{d}{dm}(Y_i - m)^2 = 2\sum_{i=1}^{n}(Y_i - m)$$

Set derivative to zero and denote optimal value of m by $\hat{m}$:

$$\sum_{i=1}^{n} Y = \sum_{i=1}^{n}\hat{m} = n\hat{m} \text{ or } \hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Why Use $\bar{Y}$ To Estimate μ_Y, ctd.

- $\bar{Y}$ has a smaller variance than all other *linear unbiased* estimators: consider the estimator, $\hat{\mu}_Y = \frac{1}{n}\sum_{i=1}^{n} a_i Y_i$, where $\{a_i\}$ are such that $\hat{\mu}_Y$ is unbiased; then $\text{var}(\bar{Y}) \leq \text{var}(\hat{\mu}_Y)$ (proof: SW, Ch. 17)
- $\bar{Y}$ isn't the only estimator of μ_Y – can you think of a time you might want to use the median instead?

1. The probability framework for statistical inference
2. Estimation
3. **Hypothesis Testing**
4. Confidence intervals

Hypothesis Testing

The ***hypothesis testing*** problem (for the mean): make a provisional decision based on the evidence at hand whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (1-sided, >)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (1-sided, <)

H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (2-sided)

Some terminology for testing statistical hypotheses:

p-value = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

The ***significance level*** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on $\bar{Y}$:

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]$$

where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom)

Calculating the p-value, ctd.

- To compute the *p*-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if *n* is small.
- If *n* is large, you can use the normal approximation (CLT):

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

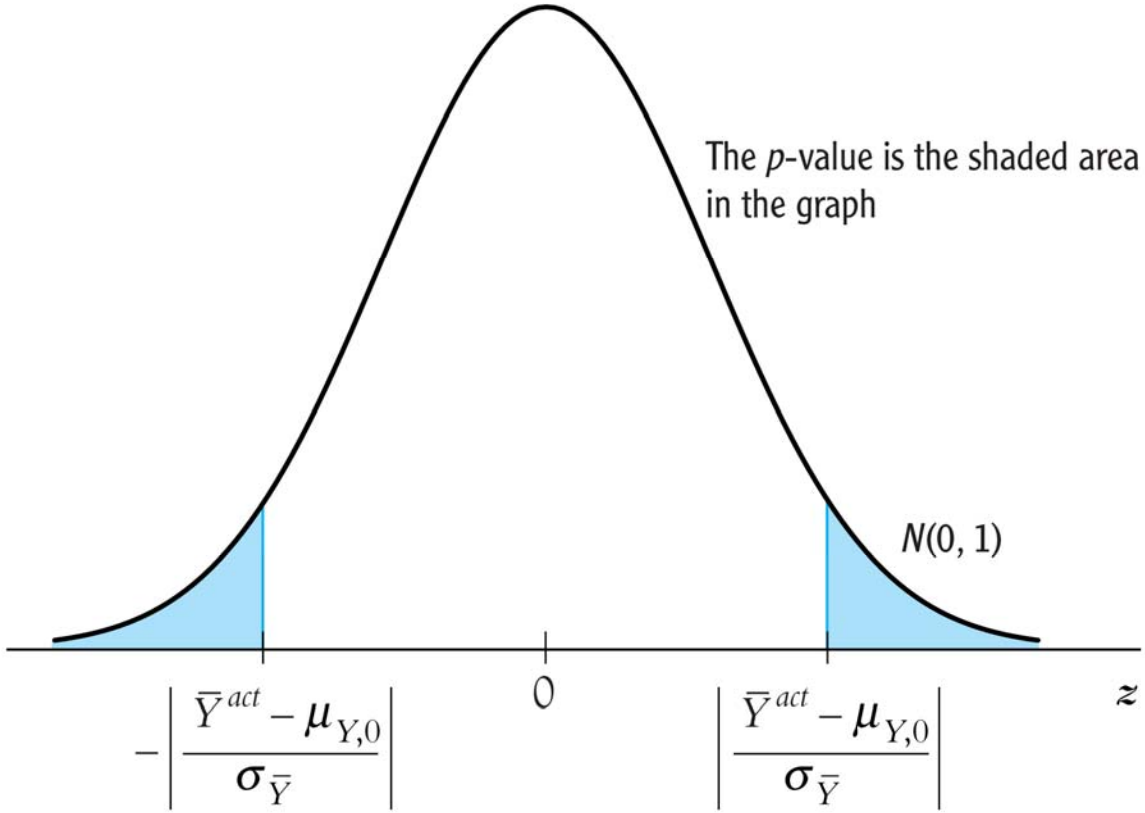
$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right]$$

$$\cong \text{probability under left+right } N(0,1) \text{ tails}$$

where $\sigma_{\bar{Y}}$ = std. dev. of the distribution of $\bar{Y} = \sigma_Y/\sqrt{n}$.

Calculating the p-value with σ_Y known:



- For large n, p-value = the probability that a $N(0,1)$ random variable falls outside $|(\bar{Y}^{act}-\mu_{Y,0})/\sigma_{\bar{Y}}|$
- In practice, $\sigma_{\bar{Y}}$ is unknown – it must be estimated

***Estimator of the variance of Y*:**

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“sample variance of } Y\text{”}$$

Fact:

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $E(Y^4) < \infty$, then $s_Y^2 \xrightarrow{p} \sigma_Y^2$

Why does the law of large numbers apply?

- Because s_Y^2 is a sample average; see Appendix 3.3
- Technical note: we assume $E(Y^4) < \infty$ because here the average is not of Y_i, but of its square; see App. 3.3

Computing the p-value with σ_Y^2 estimated:

$$p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|],$$

$$= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right]$$

$$\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right] \text{ (large } n\text{)}$$

so

$$\boxed{p\text{-value} = \Pr_{H_0}[|t| > |t^{act}|]} \quad \textbf{(}\sigma_Y^2 \textbf{ estimated)}$$

$\cong$ probability under normal tails outside $|t^{act}|$

where $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$ (the usual t-statistic)

What is the link between the *p*-value and the significance level?

The significance level is prespecified. For example, if the prespecified significance level is 5%,

- you reject the null hypothesis if $|t| \geq 1.96$.
- Equivalently, you reject if $p \leq 0.05$.
- The *p*-value is sometimes called the ***marginal significance level***.
- Often, it is better to communicate the *p*-value than simply whether a test rejects or not – the *p*-value contains more information than the "yes/no" statement about whether the test rejects.

At this point, you might be wondering,...

What happened to the *t*-table and the degrees of freedom?

Digression: the Student *t* distribution

If Y_i, $i = 1,\ldots, n$ is i.i.d. $N(\mu_Y,\sigma_Y^2)$, then the *t*-statistic has the Student *t*-distribution with $n - 1$ degrees of freedom.

The critical values of the Student *t*-distribution is tabulated in the back of all statistics books. Remember the recipe?

1. Compute the *t*-statistic
2. Compute the degrees of freedom, which is $n - 1$
3. Look up the 5% critical value
4. If the *t*-statistic exceeds (in absolute value) this critical value, reject the null hypothesis.

Comments on this recipe and the Student *t*-distribution

1. The theory of the *t*-distribution was one of the early triumphs of mathematical statistics. It is astounding, really: if *Y* is i.i.d. normal, then you can know the *exact*, *finite-sample* distribution of the *t*-statistic – it is the Student *t*. So, you can construct confidence intervals (using the Student *t* critical value) that have *exactly* the right coverage rate, no matter what the sample size. This result was really useful in times when "computer" was a job title, data collection was expensive, and the number of observations was perhaps a dozen. It is also a conceptually beautiful result, and the math is beautiful too – which is probably why stats profs love to teach the *t*-distribution. But….

Comments on Student t distribution, ctd.

2. If the sample size is moderate (several dozen) or large (hundreds or more), the difference between the *t*-distribution and N(0,1) critical values is negligible. Here are some 5% critical values for 2-sided tests:

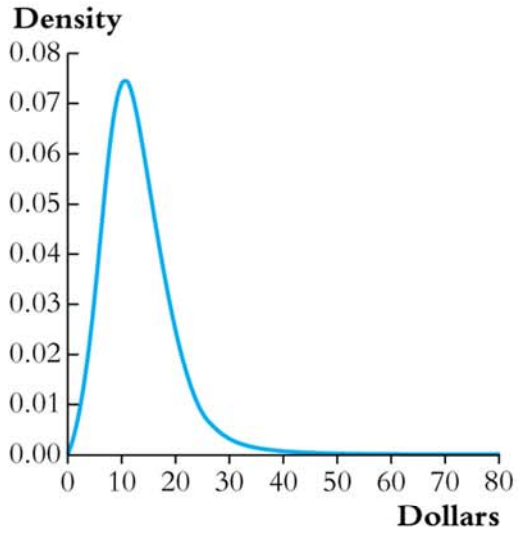
degrees of freedom ($n-1$)	5% *t*-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

Comments on Student t distribution, ctd.

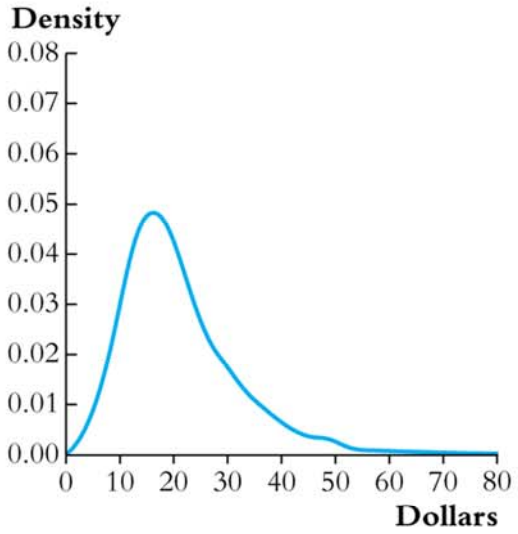
3. So, the Student-*t* distribution is only relevant when the sample size is very small; but in that case, for it to be correct, you must be sure that the population distribution of *Y* is normal. In economic data, the normality assumption is rarely credible. Here are the distributions of some economic data.
    - Do you think earnings are normally distributed?
    - Suppose you have a sample of $n = 10$ observations from one of these distributions – would you feel comfortable using the Student *t* distribution?

FIGURE 2.4 **Conditional Distribution of Average Hourly Earnings of U.S. Full-Time Workers in 2004, Given Education Level and Gender**

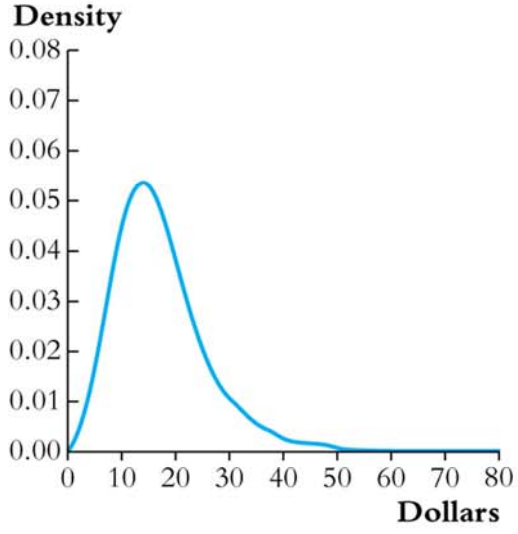
The four distributions of earnings are for women and men, for those with only a high school diploma (a and c) and those whose highest degree is from a four-year college (b and d).



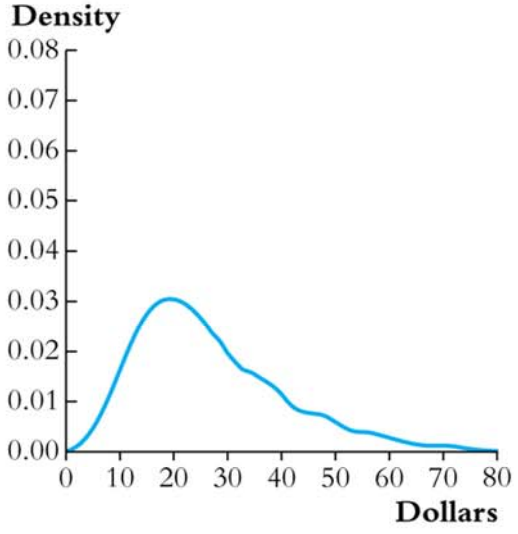
(a) Women with a high school diploma



(b) Women with a college degree



(c) Men with a high school diploma



(d) Men with a college degree

Comments on Student t distribution, ctd.

4. You might not know this. Consider the *t*-statistic testing the hypothesis that two means (groups *s*, *l*) are equal:

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

Even if the population distribution of *Y* in the two groups is normal, this statistic doesn't have a Student *t* distribution!

There is a statistic testing this hypothesis that has a normal distribution, the "pooled variance" *t*-statistic – see SW (Section 3.6) – however the pooled variance *t*-statistic is only valid if the variances of the normal distributions are the same in the two groups. Would you expect this to be true, say, for men's v. women's wages?

The Student-t distribution – Summary

- The assumption that Y is distributed $N(\mu_Y, \sigma_Y^2)$ is rarely plausible in practice (Income? Number of children?)
- For $n > 30$, the t-distribution and $N(0,1)$ are very close (as n grows large, the t_{n-1} distribution converges to $N(0,1)$)
- The t-distribution is an artifact from days when sample sizes were small and "computers" were people
- For historical reasons, statistical software typically uses the t-distribution to compute p-values – but this is irrelevant when the sample size is moderate or large.
- For these reasons, in this class we will focus on the large-n approximation given by the CLT

1. The probability framework for statistical inference
2. Estimation
3. Testing
4. **Confidence intervals**

Confidence Intervals

A 95% ***confidence interval*** for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.

Digression: What is random here? The values of $Y_1,\ldots,Y_n$ and thus any functions of them – including the confidence interval. The confidence interval will differ from one sample to the next. The population parameter, μ_Y, is not random; we just don't know it.

Confidence intervals, ctd.

A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.

$$\{\mu_Y: \left|\frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}}\right| \leq 1.96\} = \{\mu_Y: -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1.96\}$$

$$= \{\mu_Y: -1.96\frac{s_Y}{\sqrt{n}} \leq \bar{Y} - \mu_Y \leq 1.96\frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y \in (\bar{Y} - 1.96\frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96\frac{s_Y}{\sqrt{n}})\}$$

This confidence interval relies on the large-n results that $\bar{Y}$ *is approximately normally distributed and* $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

Summary:

From the two assumptions of:

(1) simple random sampling of a population, that is, $\{Y_i, i = 1,\ldots,n\}$ are i.i.d.

(2) $0 < E(Y^4) < \infty$

we developed, for large samples (large n):

- Theory of estimation (sampling distribution of $\overline{Y}$)
- Theory of hypothesis testing (large-n distribution of t-statistic and computation of the p-value)
- Theory of confidence intervals (constructed by inverting the test statistic)

Are assumptions (1) & (2) plausible in practice? **Yes**

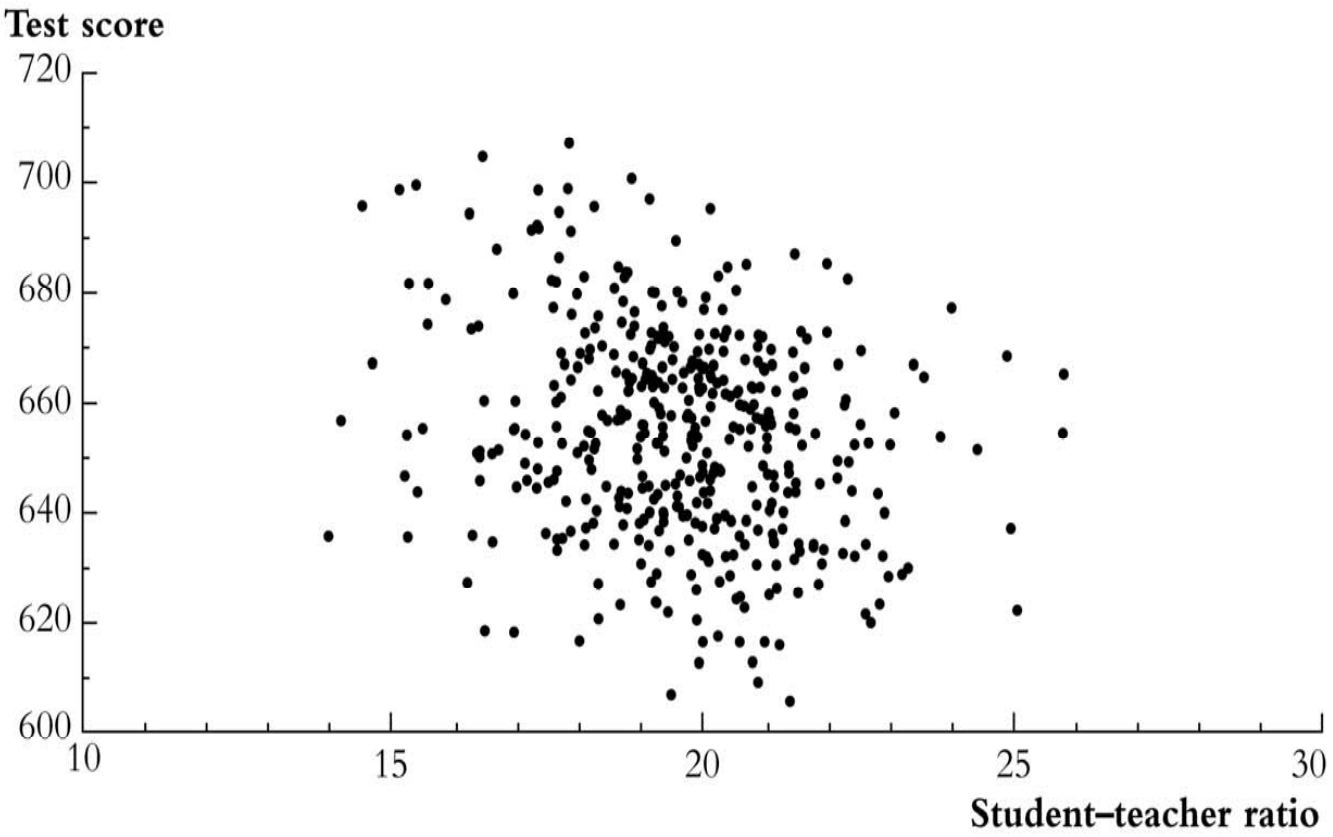
Let's go back to the original policy question:

What is the effect on test scores of reducing STR by one student/class?

Have we answered this question?

FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student–teacher ratio and test scores: The sample correlation is −0.23.



Linear Regression with One Regressor

(Stock/Watson Chapter 4)

Outline

1. The population linear regression model
2. The ordinary least squares (OLS) estimator and the sample regression line
3. Measures of fit of the sample regression
4. The least squares assumptions
5. The sampling distribution of the OLS estimator

Linear regression lets us estimate the slope of the population regression line.

- The slope of the population regression line is the expected effect on Y of a unit change in X.

- Ultimately our aim is to estimate the causal effect on Y of a unit change in X – but for now, just think of the problem of fitting a straight line to data on two variables, Y and X.

The problem of statistical inference for linear regression is, at a general level, the same as for estimation of the mean or of the differences between two means. Statistical, or econometric, inference about the slope entails:

- Estimation:
  - How should we draw a line through the data to estimate the population slope?
    - Answer: ordinary least squares (OLS).
  - What are advantages and disadvantages of OLS?
- Hypothesis testing:
  - How to test if the slope is zero?
- Confidence intervals:
  - How to construct a confidence interval for the slope?

The Linear Regression Model
(SW Section 4.1)

The ***population regression line*:**

$$Test\ Score = \beta_0 + \beta_1 STR$$

β_1 = slope of population regression line

$$= \frac{\Delta \text{Test score}}{\Delta STR}$$

= change in test score for a unit change in *STR*

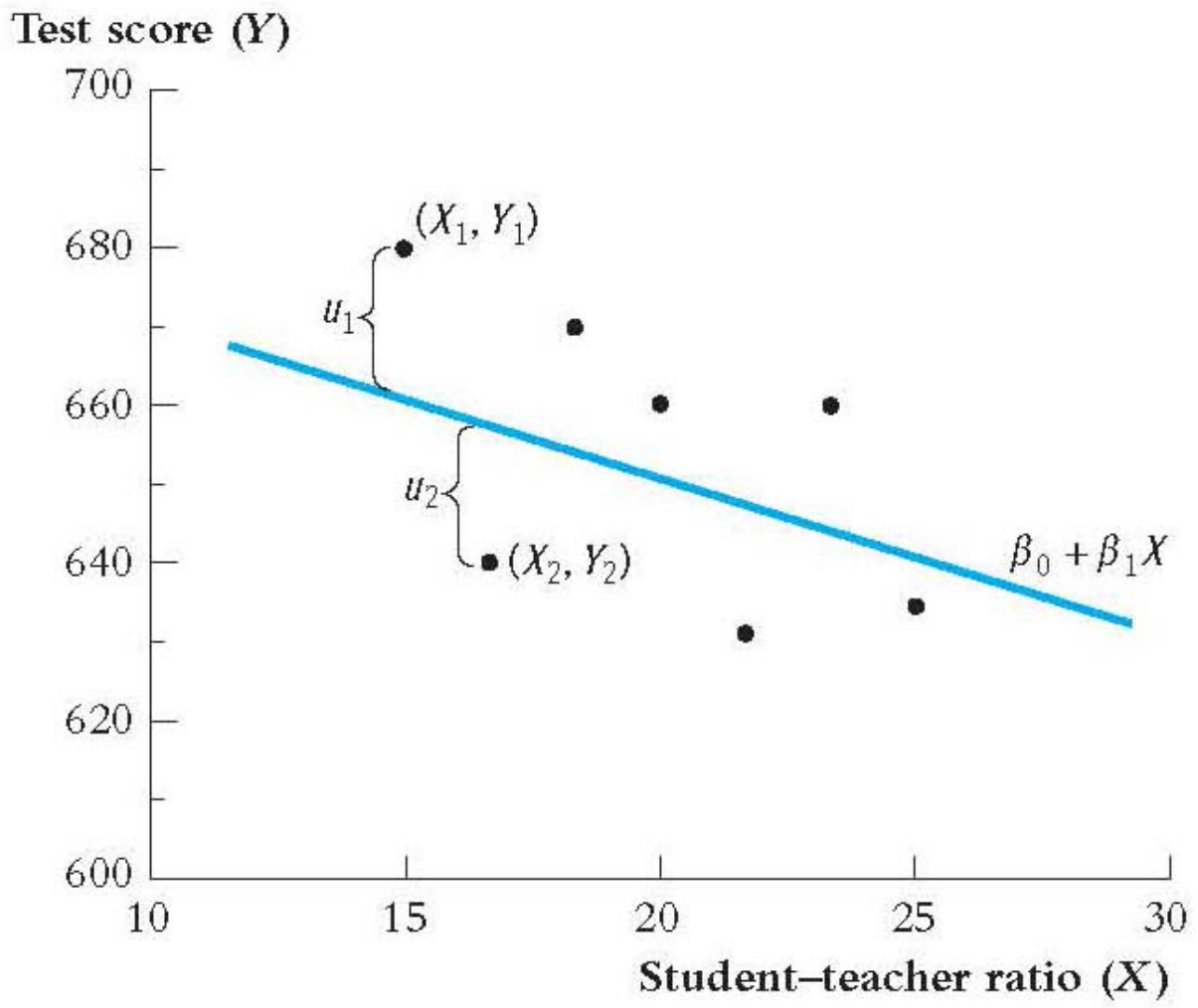
- *Why are β_0 and β_1 "population" parameters*?
- We would like to know the population value of β_1.
- We don't know β_1, so must estimate it using data.

The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

- We have n observations, $(X_i,\ Y_i)$, $i = 1,..,\ n$.
- X is the ***independent variable*** or ***regressor***
- Y is the ***dependent variable***
- β_0 = ***intercept***
- β_1 = ***slope***
- u_i = the regression ***error***
- The regression error consists of omitted factors. In general, these omitted factors are other factors that influence Y, other than the variable X. The regression error also includes error in the measurement of Y.

The population regression model in a picture: Observations on Y and X ($n = 7$); the population regression line; and the regression error (the "error term"):



The Ordinary Least Squares Estimator (SW Section 4.2)

How can we estimate β_0 and β_1 from data?

Recall that $\bar{Y}$ was the least squares estimator of μ_Y: $\bar{Y}$ solves,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

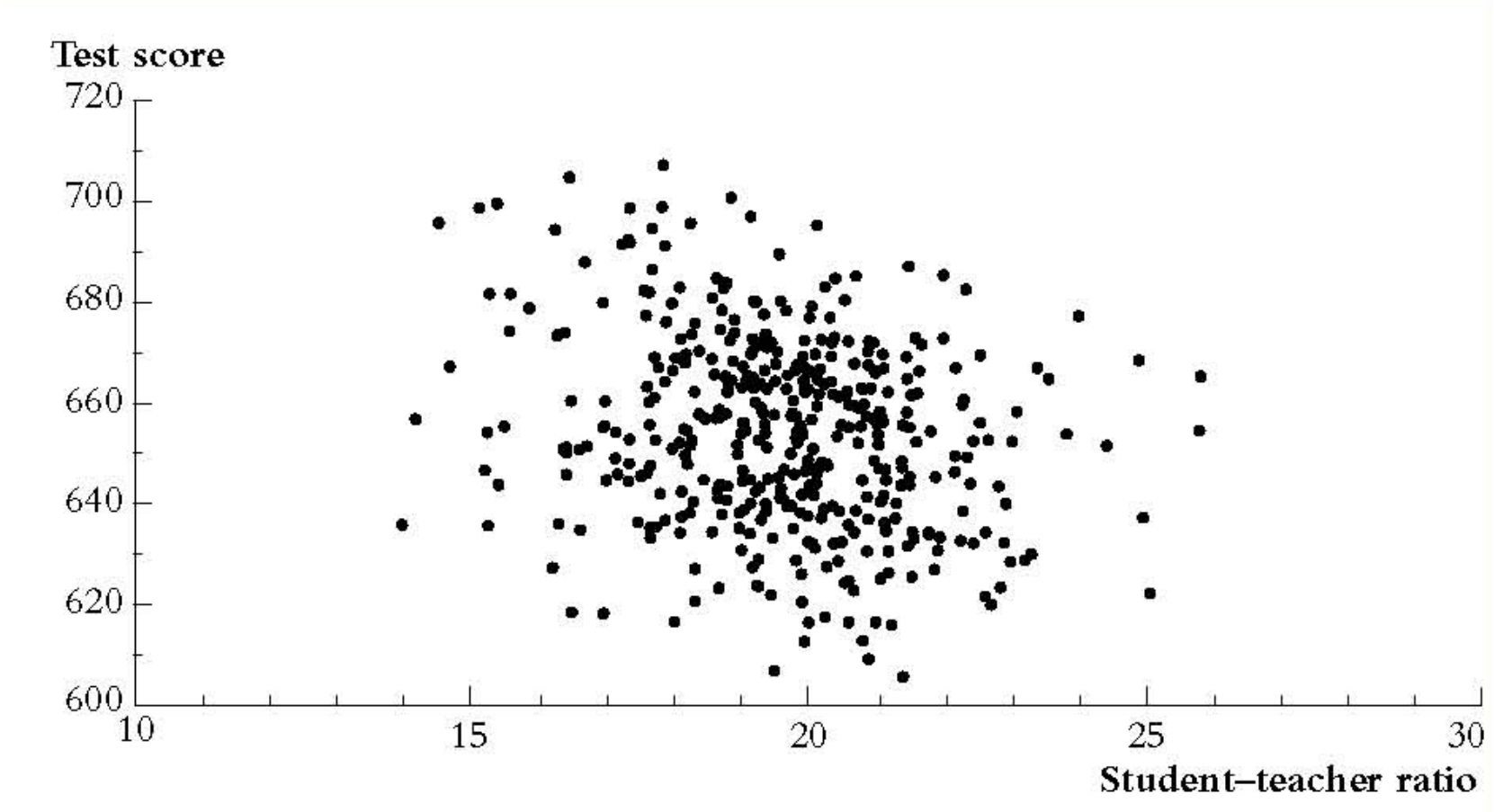
By analogy, **we will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters β_0 and β_1.** The OLS estimator solves,

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

Mechanics of OLS

The population regression line: $Test\ Score = \beta_0 + \beta_1 STR$

$$\beta_1 = \frac{\Delta \text{Test score}}{\Delta STR} = ??$$



The OLS estimator solves: $\min_{b_0,b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction ("predicted value") based on the estimated line.

- This minimization problem can be solved using calculus (App. 4.2).

- **The result is the OLS estimators of β_0 and β_1.**

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4.8}$$

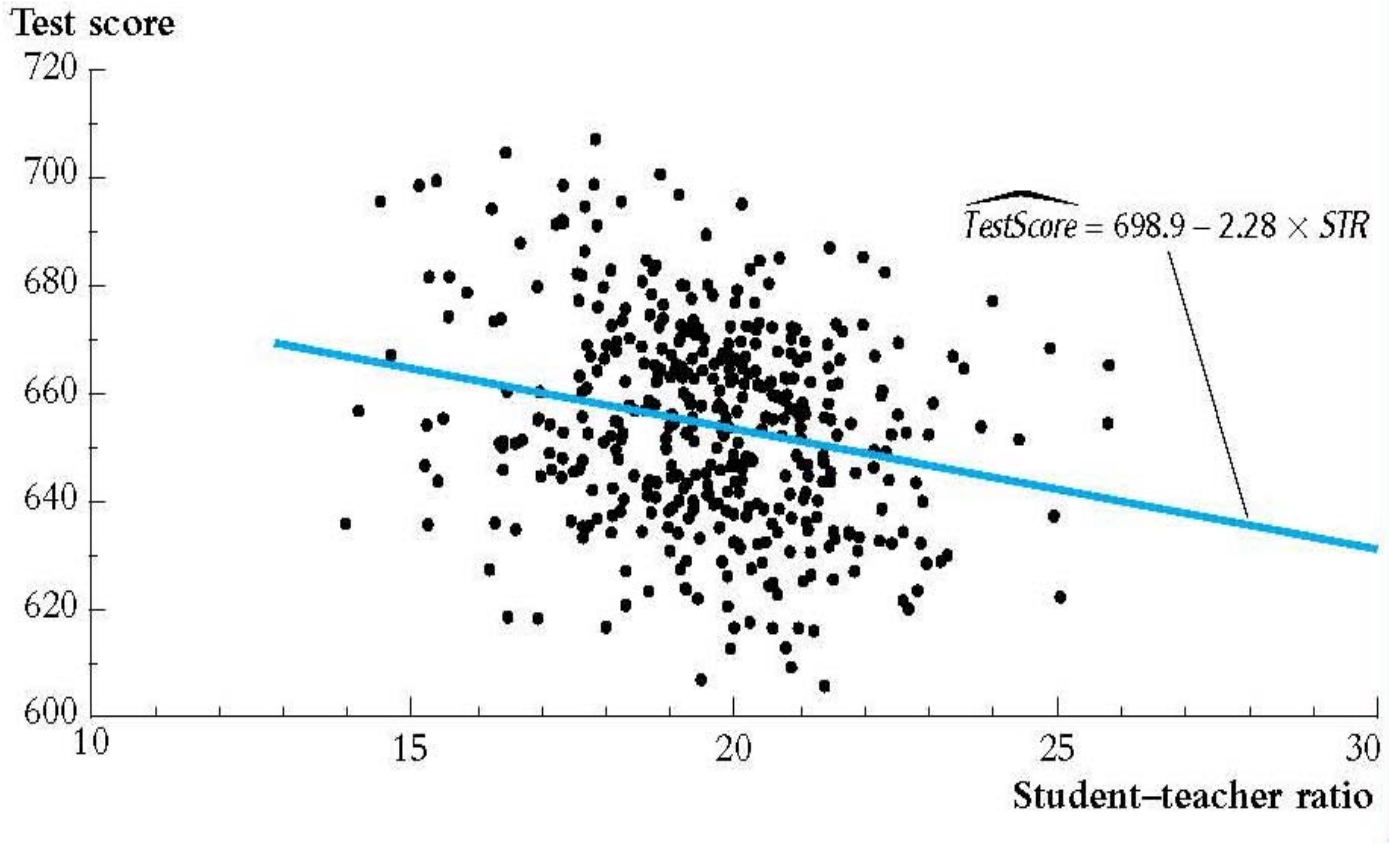
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept $(\hat{\beta}_0)$, slope $(\hat{\beta}_1)$, and residual $(\hat{u}_i)$ are computed from a sample of n observations of X_i and $Y_i, i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

Application to the California *Test Score – Class Size* data



Estimated slope $= \hat{\beta}_1 = -2.28$

Estimated intercept $= \hat{\beta}_0 = 698.9$

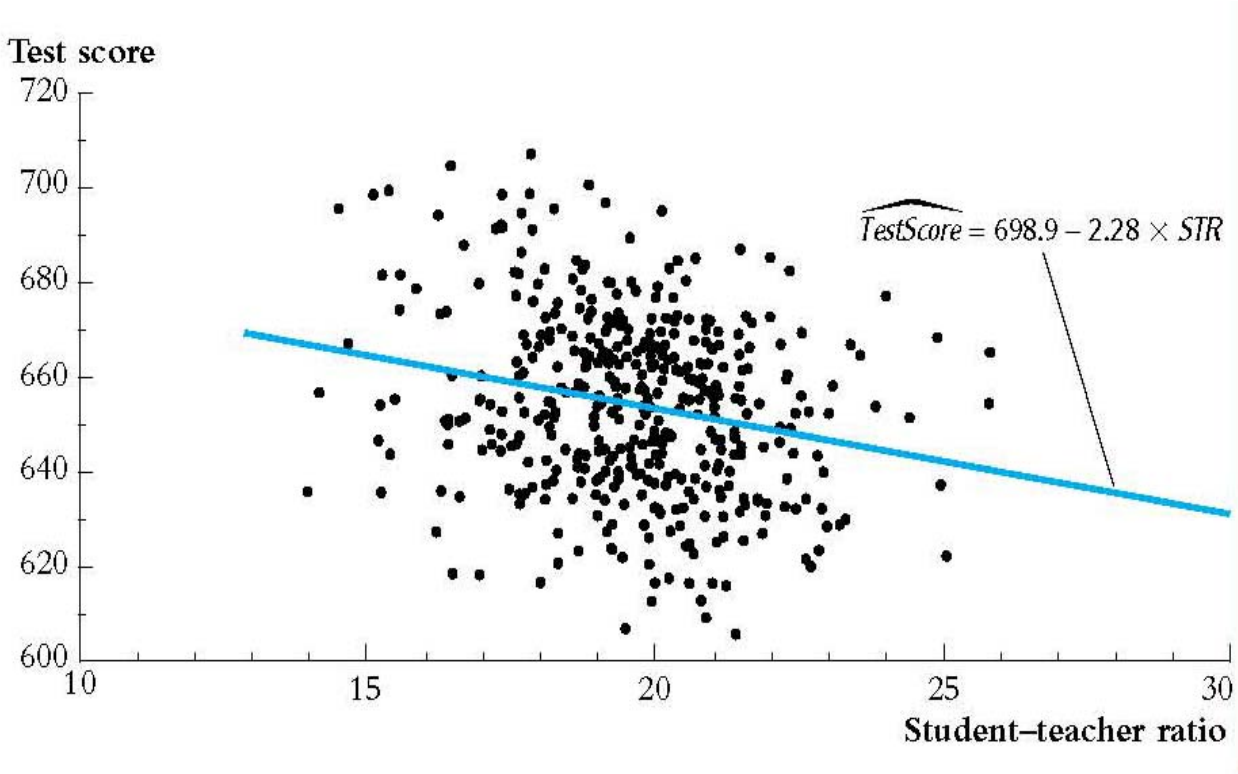
Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Interpretation of the estimated slope and intercept

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

- Districts with one more student per teacher on average have test scores that are 2.28 points lower.
- That is, $\dfrac{\Delta \text{Test score}}{\Delta STR} = -2.28$
- The intercept (taken literally) means that, according to this estimated line, districts with zero students per teacher would have a (predicted) test score of 698.9. But this interpretation of the intercept makes no sense – it extrapolates the line outside the range of the data – here, the intercept is not economically meaningful.

Predicted values & residuals:



One of the districts in the data set is Antelope, CA, for which *STR* = 19.33 and *Test Score* = 657.8

predicted value: $\hat{Y}_{Antelope} = 698.9 - 2.28 \times 19.33 = 654.8$

residual: $\hat{u}_{Antelope} = 657.8 - 654.8 = 3.0$

OLS regression: STATA output

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581

------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |   698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

(We'll discuss the rest of this output later.)

Measures of Fit
(Section 4.3)

Two regression statistics provide complementary measures of how well the regression line "fits" or explains the data:

- The ***regression R^2*** measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit)

- The ***standard error of the regression (SER)*** measures the magnitude of a typical regression residual in the units of Y.

The *regression* R^2 is the fraction of the sample variance of Y_i "explained" by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual

$\Rightarrow$ sample var (Y) = sample var($\hat{Y}_i$) + sample var($\hat{u}_i$) (*why?*)

$\Rightarrow$ total sum of squares = "explained" SS + "residual" SS

Definition of R^2:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \le R^2 \le 1$
- For regression with a single X, R^2 = the square of the correlation coefficient between X and Y

The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of *u*. The *SER* is (almost) the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$.

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The *SER*:

- has the units of *u*, which are the units of *Y*
- measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)
- The ***root mean squared error*** (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

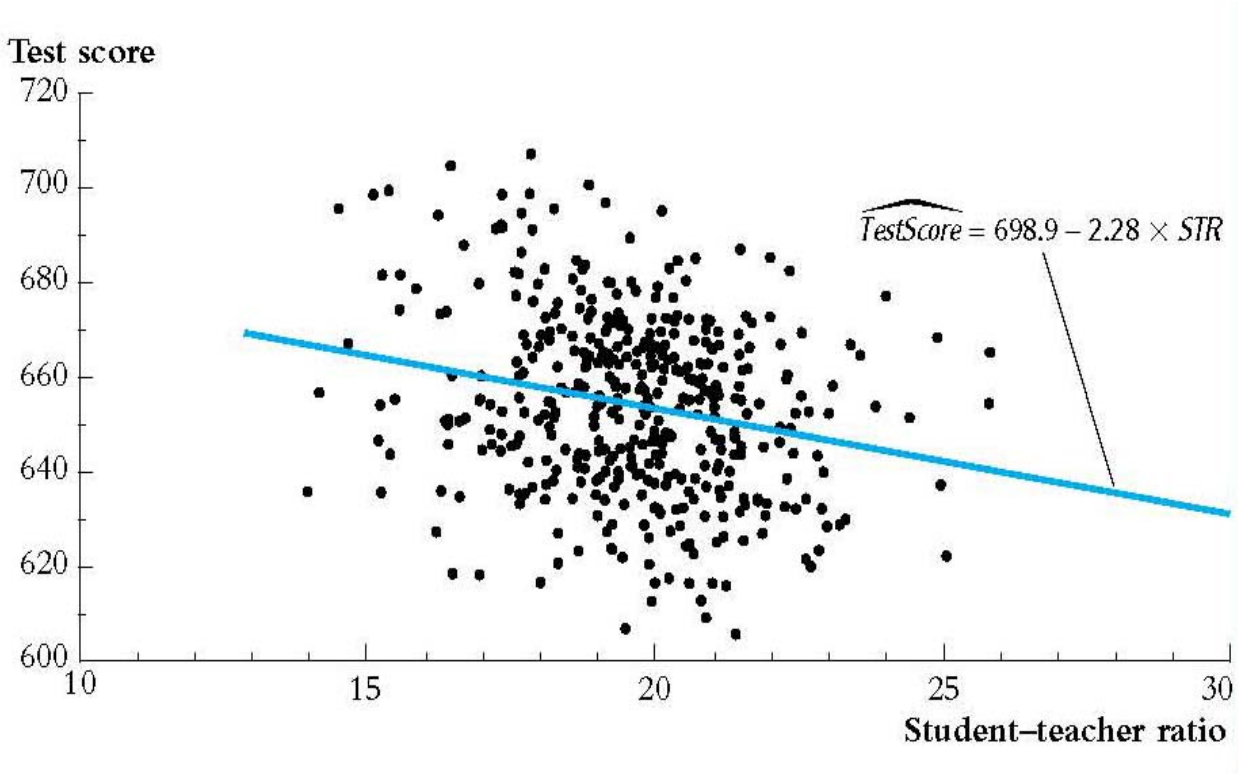
This measures the same thing as the *SER* – the minor difference is division by $1/n$ instead of $1/(n-2)$.

Technical note: why divide by n–2 instead of n–1?

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- Division by n–2 is a "degrees of freedom" correction – just like division by n–1 in s_Y^2, except that for the *SER*, two parameters have been estimated (β_0 and β_1, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in s_Y^2 only one has been estimated (μ_Y, by $\overline{Y}$).
- When n is large, it doesn't matter whether n, n–1, or n–2 are used – although the conventional formula uses n–2 when there is a single regressor.
- For details, see Section 17.4

Example of the R^2 and the *SER*



$\widehat{TestScore} = 698.9 - 2.28 \times STR$, $R^2 = .05$, $SER = 18.6$

STR explains only a small fraction of the variation in test scores. Does this make sense? Does this mean the STR is unimportant in a policy sense?

The Least Squares Assumptions
(SW Section 4.4)

What, in a precise sense, are the properties of the sampling distribution of the OLS estimator? When will $\hat{\beta}_1$ be unbiased? What is its variance?

To answer these questions, we need to make some assumptions about how *Y* and *X* are related to each other, and about how they are collected (the sampling scheme)

These assumptions – there are three – are known as the Least Squares Assumptions.

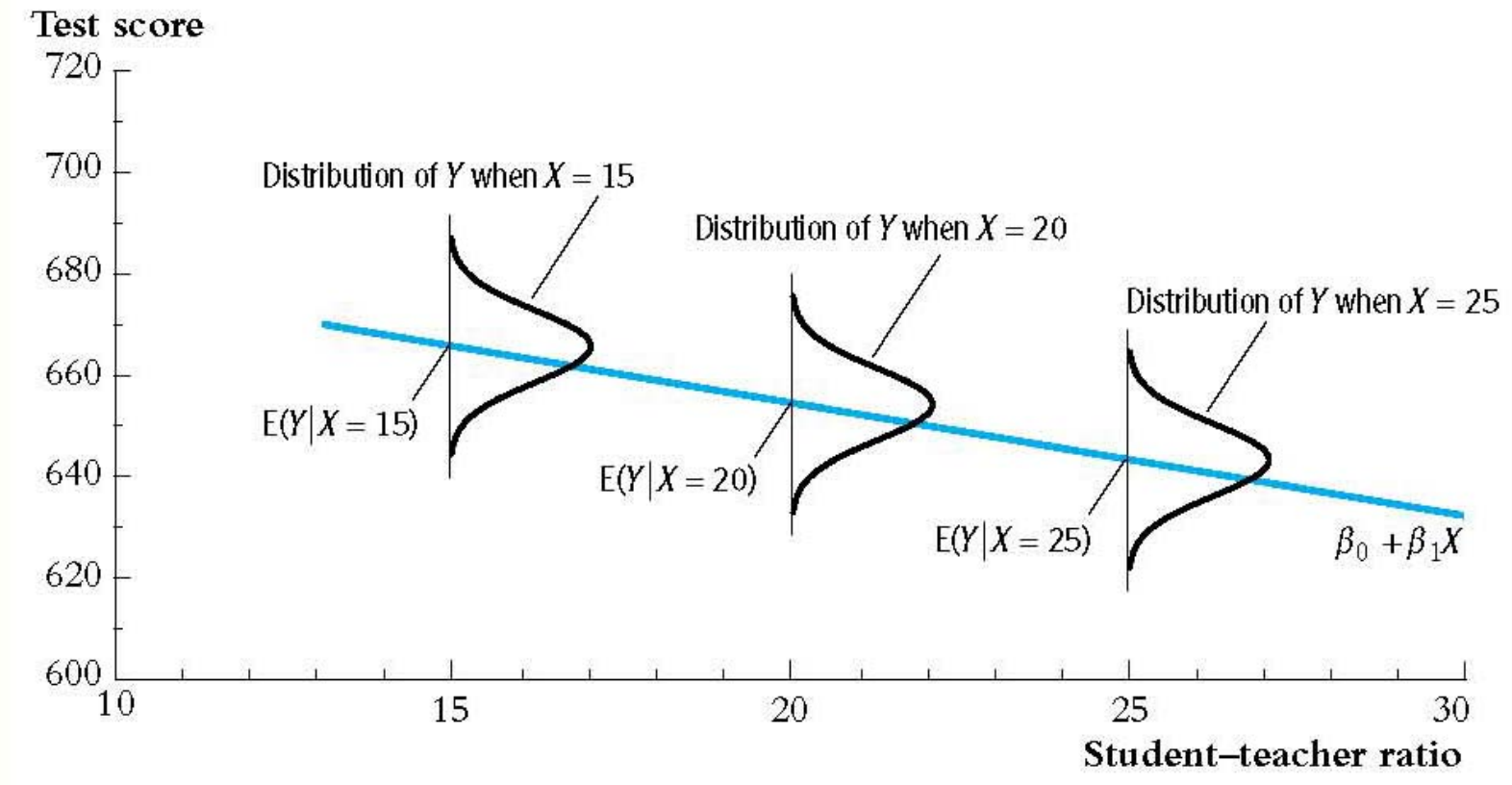
The Least Squares Assumptions

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

1. The conditional distribution of u given X has mean zero, that is, $E(u|X = x) = 0$.
    - *This implies that $\hat{\beta}_1$ is unbiased*
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
    - *This is true if (X, Y) are collected by simple random sampling*
    - *This delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$*
3. Large outliers in X and/or Y are rare.
    - *Technically, X and Y have finite fourth moments*
    - *Outliers can result in meaningless values of $\hat{\beta}_1$*

Least squares assumption #1: $E(u|X = x) = 0$.

For any given value of X, the mean of u is zero:



Example: $Test\ Score_i = \beta_0 + \beta_1 STR_i + u_i$, u_i = other factors

- What are some of these "other factors"?
- Is $E(u|X=x) = 0$ plausible for these other factors?

Least squares assumption #1, ctd.

A benchmark for thinking about this assumption is to consider an ideal randomized controlled experiment:

- X is randomly assigned to people (students randomly assigned to different size classes; patients randomly assigned to medical treatments). Randomization is done by computer – using no information about the individual.
- Because X is assigned randomly, all other individual characteristics – the things that make up u – are distributed independently of X, so u and X are independent
- Thus, in an ideal randomized controlled experiment, $\boldsymbol{E(u|X = x) = 0}$ (that is, LSA #1 holds)
- In actual experiments, or with observational data, we will need to think hard about whether $E(u|X = x) = 0$ holds.

Least squares assumption #2: (X_i,Y_i), $i = 1,\ldots,n$ are i.i.d.

This arises automatically if the entity (individual, district) is sampled by simple random sampling:

- The entities are selected from the same population, so (X_i, Y_i) are *identically distributed* for all $i = 1,\ldots, n$.
- The entities are selected at random, so the values of (X, Y) for different entities are *independently distributed*.

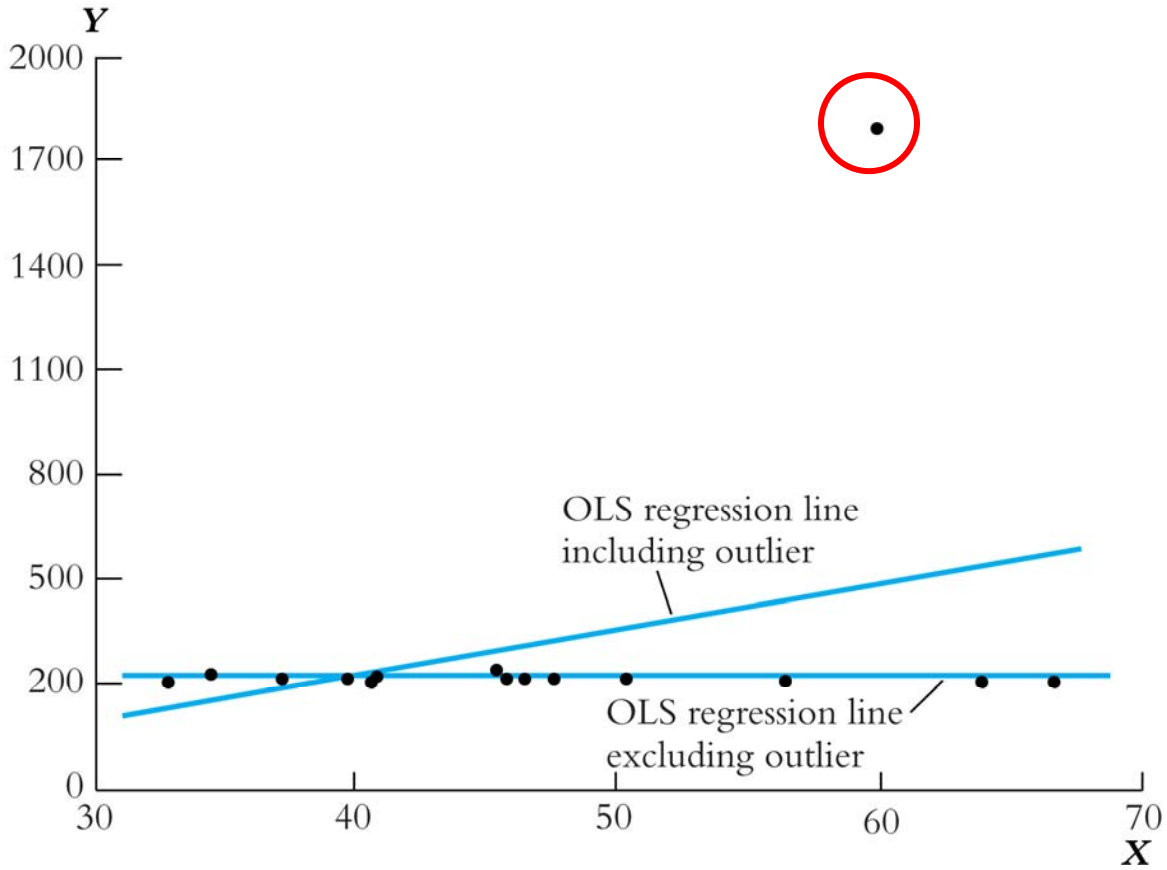
The main place we will encounter non-i.i.d. sampling is when data are recorded over time for the same entity (panel data and time series data) – we will deal with that complication when we cover panel data.

Least squares assumption #3: *Large outliers are rare*

Technical statement: $E(X^4) < \infty$ **and** $E(Y^4) < \infty$

- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; *STR*, family income, etc. satisfy this too.)
- The substance of this assumption is that a large outlier can strongly influence the results – so we need to rule out large outliers.
- Look at your data! If you have a large outlier, is it a typo? Does it belong in your data set? Why is it an outlier?

OLS can be sensitive to an outlier:



- *Is the lone point an outlier in X or Y?*
- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. Plot your data!

The Sampling Distribution of the OLS Estimator (SW Section 4.5)

The OLS estimator is computed from a sample of data. A different sample yields a different value of $\hat{\beta}_1$. This is the source of the "sampling uncertainty" of $\hat{\beta}_1$. We want to:

- quantify the sampling uncertainty associated with $\hat{\beta}_1$
- use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$
- construct a confidence interval for β_1
- All these require figuring out the sampling distribution of the OLS estimator. Two steps to get there…
  - Probability framework for linear regression
  - Distribution of the OLS estimator

Probability Framework for Linear Regression

The probability framework for linear regression is summarized by the three least squares assumptions.

Population

- The group of interest (ex: all possible school districts)

Random variables*: *Y*, *X

- Ex: (*Test Score*, *STR*)

***Joint distribution of* (*Y*, *X*). We assume:**

- The population regression function is linear
- $E(u|X) = 0$ (1^{st} Least Squares Assumption)
- X, Y have nonzero finite fourth moments (3^{rd} L.S.A.)

Data Collection by simple random sampling implies:

- $\{(X_i, Y_i)\}$, $i = 1,\ldots, n$, are i.i.d. (2^{nd} L.S.A.)

The Sampling Distribution of $\hat{\beta}_1$

Like $\bar{Y}$, $\hat{\beta}_1$ has a sampling distribution.

- What is $E(\hat{\beta}_1)$?
  - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
- What is var($\hat{\beta}_1$)? (measure of sampling uncertainty)
  - We need to derive a formula so we can compute the standard error of $\hat{\beta}_1$.
- What is the distribution of $\hat{\beta}_1$ in small samples?
  - It is very complicated in general
- What is the distribution of $\hat{\beta}_1$ in large samples?
  - In large samples, $\hat{\beta}_1$ is normally distributed.

The mean and variance of the sampling distribution of $\hat{\beta}_1$

Some preliminary algebra:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$\overline{Y} = \beta_0 + \beta_1 \overline{X} + \overline{u}$$

so $$Y_i - \overline{Y} = \beta_1 (X_i - \overline{X}) + (u_i - \overline{u})$$

Thus,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \overline{X})[\beta_1 (X_i - \overline{X}) + (u_i - \overline{u})]}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$\hat{\beta}_1 = \beta_1 \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so $$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now $$\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\sum_{i=1}^{n}(X_i - \bar{X})\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i - \left[\left(\sum_{i=1}^{n}X_i\right) - n\bar{X}\right]\bar{u}$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})u_i$$

Substitute $\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u}) = \sum_{i=1}^{n}(X_i - \bar{X})u_i$ into the expression for $\hat{\beta}_1 - \beta_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

so

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Now we can calculate $E(\hat{\beta}_1)$ and var($\hat{\beta}_1$):

$$E(\hat{\beta}_1) - \beta_1 = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

$$= E\left\{E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \middle| X_1, \ldots, X_n\right\}$$

$= 0$ **because $E(u_i|X_i=x) = 0$ by LSA #1**

- Thus LSA #1 implies that $E(\hat{\beta}_1) = \beta_1$
- That is, $\hat{\beta}_1$ **is an unbiased estimator of β_1.**
- For details see App. 4.3

Next calculate var($\hat{\beta}_1$):

write

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i$. If n is large, $s_X^2 \approx \sigma_X^2$ and $\frac{n-1}{n} \approx 1$, so

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2},$$

where $v_i = (X_i - \bar{X})u_i$ (see App. 4.3). Thus,

$$\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}$$

so $\text{var}(\hat{\beta}_1 - \beta_1) = \text{var}(\hat{\beta}_1)$

$$= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} v_i\right) \Big/ (\sigma_X^2)^2 = \frac{\text{var}(v_i)/n}{(\sigma_X^2)^2}$$

where the final equality uses assumption 2. Thus,

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2} \quad .$$

Summary so far

1. $\hat{\beta}_1$ is unbiased: $E(\hat{\beta}_1) = \beta_1$ – just like $\bar{Y}$!
2. $\text{var}(\hat{\beta}_1)$ is inversely proportional to n – just like $\bar{Y}$!

What is the sampling distribution of $\hat{\beta}_1$?

The exact sampling distribution is complicated – it depends on the population distribution of (Y, X) – but when n is large we get some simple (and good) approximations:

(1) Because $\text{var}(\hat{\beta}_1) \propto 1/n$ and $E(\hat{\beta}_1) = \beta_1$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$

(2) When n is large, the sampling distribution of $\hat{\beta}_1$ is well approximated by a normal distribution (CLT)

*Recall the **CLT***: suppose $\{v_i\}$, $i = 1,\ldots, n$ is i.i.d. with $E(v) = 0$ and $\text{var}(v) = \sigma^2$. Then, when n is large, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0, \sigma_v^2 / n)$.

Large-*n* approximation to the distribution of $\hat{\beta}_1$:

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\sigma_X^2}, \text{ where } v_i = (X_i - \bar{X})u_i$$

- When n is large, $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$, which is i.i.d. (*why*?) and $\text{var}(v_i) < \infty$ (*why*?). So, by the CLT, $\frac{1}{n}\sum_{i=1}^{n} v_i$ is approximately distributed $N(0, \sigma_v^2 / n)$.
- Thus, for n large, $\hat{\beta}_1$ is approximately distributed

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ where } v_i = (X_i - \mu_X)u_i$$

The larger the variance of *X*, the smaller the variance of $\hat{\beta}_1$

The math

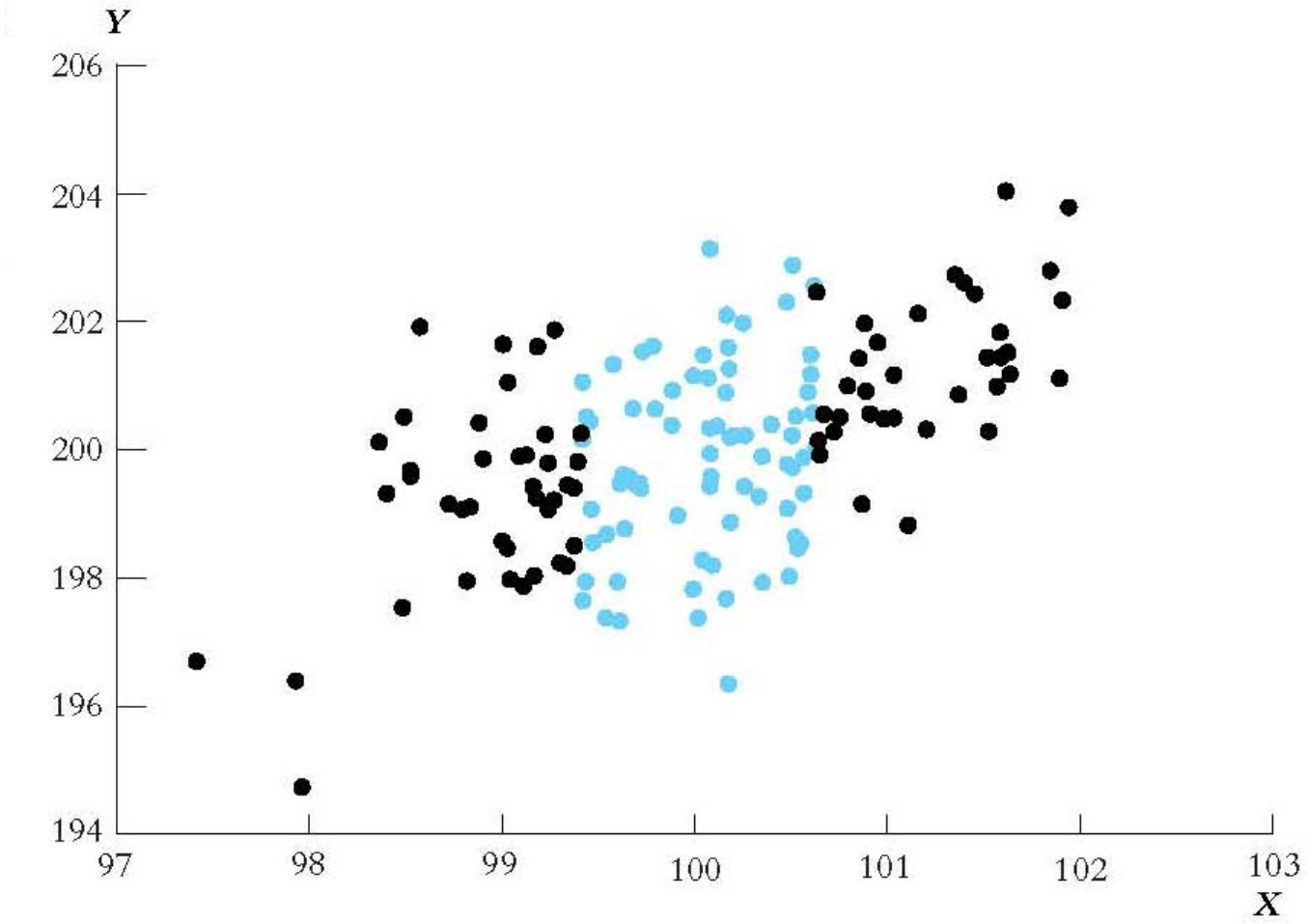
$$\text{var}(\hat{\beta}_1 - \beta_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{(\sigma_X^2)^2}$$

where $\sigma_X^2 = \text{var}(X_i)$. The variance of X appears (squared) in the denominator – so increasing the spread of X decreases the variance of β_1.

The intuition

If there is more variation in X, then there is more information in the data that you can use to fit the regression line. This is most easily seen in a figure…

The larger the variance of X, the smaller the variance of $\hat{\beta}_1$



The number of black and blue dots is the same. Using which would you get a more accurate regression line?

Summary of the sampling distribution of $\hat{\beta}_1$:

If the three Least Squares Assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:
  - $E(\hat{\beta}_1) = \beta_1$ (that is, $\hat{\beta}_1$ is unbiased)
  - $\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4} \propto \frac{1}{n}$.
- Other than its mean and variance, the exact distribution of $\hat{\beta}_1$ is complicated and depends on the distribution of (X, u)
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (that is, $\hat{\beta}_1$ is consistent)
- When n is large, $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0,1)$ (CLT)
- *This parallels the sampling distribution of $\overline{Y}$.*

Large-Sample Distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

If the least squares assumptions in Key Concept 4.3 hold, then in large samples $\hat{\beta}_0$ and $\hat{\beta}_1$ have a jointly normal sampling distribution. The large-sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$, where the variance of this distribution, $\sigma^2_{\hat{\beta}_1}$, is

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\text{var}[(X_i - \mu_X)u_i]}{[\text{var}(X_i)]^2}. \quad (4.21)$$

The large-sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$, where

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{\text{var}(H_i u_i)}{[E(H_i^2)]^2}, \text{ where } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right]X_i. \quad (4.22)$$

We are now ready to turn to hypothesis tests & confidence intervals…

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals

(SW Chapter 5)

Outline

1. The standard error of $\hat{\beta}_1$
2. Hypothesis tests concerning β_1
3. Confidence intervals for β_1
4. Regression when X is binary
5. Heteroskedasticity and homoskedasticity
6. Efficiency of OLS and the Student t distribution

A big picture review of where we are going…

We want to learn about the slope of the population regression line. We have data from a sample, so there is sampling uncertainty. There are five steps towards this goal:

1. State the population object of interest
2. Provide an estimator of this population object
3. Derive the sampling distribution of the estimator (this requires certain assumptions). In large samples this sampling distribution will be normal by the CLT.
4. The square root of the estimated variance of the sampling distribution is the standard error (SE) of the estimator
5. Use the *SE* to construct *t*-statistics (for hypothesis tests) and confidence intervals.

Object of interest: β_1 in,

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots,\ n$$

$\beta_1 = \Delta Y/\Delta X$, for an autonomous change in X (***causal effect***)

Estimator: the OLS estimator $\hat{\beta}_1$.

The Sampling Distribution of $\hat{\beta}_1$:

To derive the large-sample distribution of $\hat{\beta}_1$, we make the following assumptions:

The Least Squares Assumptions:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare ($E(X^4) < \infty$, $E(Y^4) < \infty$.

The Sampling Distribution of $\hat{\beta}_1$, ctd.

Under the Least Squares Assumptions, for *n* large, $\hat{\beta}_1$ is approximately distributed,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n(\sigma_X^2)^2}\right), \text{ where } v_i = (X_i - \mu_X)u_i$$

Hypothesis Testing and the Standard Error of $\hat{\beta}_1$ (Section 5.1)

The objective is to test a hypothesis, like $\beta_1 = 0$, using data – to reach a tentative conclusion whether the (null) hypothesis is correct or incorrect.

General setup

Null hypothesis and **two-sided** alternative:

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 \neq \beta_{1,0}$$

where $\beta_{1,0}$ is the hypothesized value under the null.

Null hypothesis and **one-sided** alternative:

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 < \beta_{1,0}$$

General approach: construct *t*-statistic, and compute *p*-value (or compare to the $N(0,1)$ critical value)

- ***In general***: $$t = \frac{\text{estimator - hypothesized value}}{\text{standard error of the estimator}}$$

  where the *SE* of the estimator is the square root of an estimator of the variance of the estimator.

- ***For testing the mean of Y***: $$t = \frac{\overline{Y} - \mu_{Y,0}}{s_Y / \sqrt{n}}$$

- ***For testing β_1***, $$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)},$$

  where $SE(\hat{\beta}_1)$ = the square root of an estimator of the variance of the sampling distribution of $\hat{\beta}_1$

Formula for SE($\hat{\beta}_1$)

Recall the expression for the variance of $\hat{\beta}_1$ (large n):

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n(\sigma_X^2)^2}, \text{ where } v_i = (X_i - \mu_X)u_i.$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_v^2 and σ_X^2 by estimators constructed from the data:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{estimator of } \sigma_v^2}{(\text{estimator of } \sigma_X^2)^2} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

where $\hat{v}_i = (X_i - \bar{X})\hat{u}_i$.

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{v}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}, \text{ where } \hat{v}_i = (X_i - \bar{X})\hat{u}_i .$$

$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$ = the standard error of $\hat{\beta}_1$

This is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates [var(X)]^2.
- Why the degrees-of-freedom adjustment $n - 2$? Because two coefficients have been estimated (β_0 and β_1).
- $SE(\hat{\beta}_1)$ is computed by regression software
- Your regression software has memorized this formula so you don't need to.

Summary: To test H_0: $\beta_1 = \beta_{1,0}$ v. H_1: $\beta_1 \neq \beta_{1,0}$,

- Construct the *t*-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$

- Reject at 5% significance level if |t| > 1.96
- The *p*-value is $p = \Pr[|t| > |t^{act}|]$ = probability in tails of normal outside $|t^{act}|$; you reject at the 5% significance level if the *p*-value is < 5%.
- This procedure relies on the large-*n* approximation that $\hat{\beta}_1$ is normally distributed; typically $n = 50$ is large enough for the approximation to be excellent.

Example: *Test Scores* and *STR*, California data

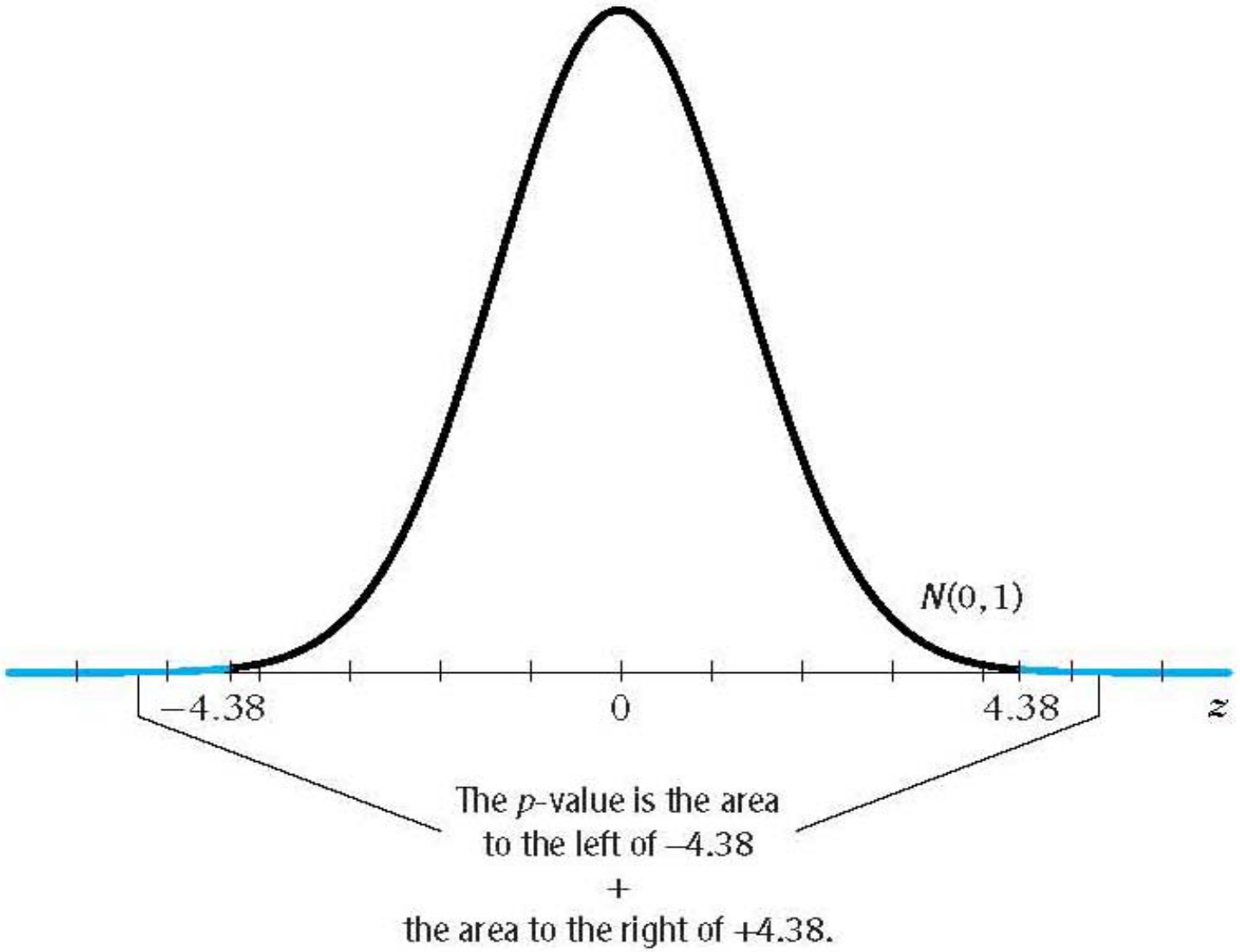
Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

Regression software reports the standard errors:

$$SE(\hat{\beta}_0) = 10.4 \qquad SE(\hat{\beta}_1) = 0.52$$

t-statistic testing $\beta_{1,0} = 0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} = \dfrac{-2.28 - 0}{0.52} = -4.38$

- The 1% 2-sided significance level is 2.58, so we reject the null at the 1% significance level.
- Alternatively, we can compute the p-value…



The *p*-value based on the large-*n* standard normal approximation to the *t*-statistic is 0.00001 (10^{-5})

Confidence Intervals for β_1 (Section 5.2)

Recall that a 95% confidence is, equivalently:

- The set of points that cannot be rejected at the 5% significance level;
- A set-valued function of the data (an interval that is a function of the data) that contains the true parameter value 95% of the time in repeated samples.

Because the *t*-statistic for β_1 is N(0,1) in large samples, construction of a 95% confidence for β_1 is just like the case of the sample mean:

95% confidence interval for $\beta_1 = \{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$

Confidence interval example: *Test Scores* and *STR*

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$$SE(\hat{\beta}_0) = 10.4 \qquad SE(\hat{\beta}_1) = 0.52$$

95% confidence interval for $\hat{\beta}_1$:

$$\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\} = \{-2.28 \pm 1.96 \times 0.52\}$$
$$= (-3.30, -1.26)$$

The following two statements are equivalent *(why?)*

- The 95% confidence interval does not include zero;
- The hypothesis $\beta_1 = 0$ is rejected at the 5% level

A concise (and conventional) way to report regressions:
Put standard errors in parentheses below the estimated coefficients to which they apply.

$$\widehat{TestScore} = 698.9 - 2.28 \times STR,\ R^2 = .05,\ SER = 18.6$$
$$(10.4)\ \ (0.52)$$

<u>This expression gives a lot of information</u>

- The estimated regression line is
$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$
- The standard error of $\hat{\beta}_0$ is 10.4
- The standard error of $\hat{\beta}_1$ is 0.52
- The R^2 is .05; the standard error of the regression is 18.6

OLS regression: reading STATA output

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.38   0.000    -3.300945   -1.258671
       _cons |   698.933   10.36436     67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

so:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \; , \; R^2 = .05, \; SER = 18.6$$
$$(10.4) \; (0.52)$$

$t\,(\beta_1 = 0) = -4.38$, p-value $= 0.000$ (2-sided)

95% 2-sided conf. interval for β_1 is $(-3.30, -1.26)$

Summary of statistical inference about β_0 and β_1

Estimation:

- OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ have approximately normal sampling distributions in large samples

Testing:

- H_0: $\beta_1 = \beta_{1,0}$ v. $\beta_1 \neq \beta_{1,0}$ ($\beta_{1,0}$ is the value of β_1 under H_0)
- $t = (\hat{\beta}_1 - \beta_{1,0})/SE(\hat{\beta}_1)$
- p-value = area under standard normal outside t^{act} (large n)

Confidence Intervals:

- 95% confidence interval for β_1 is $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$
- This is the set of β_1 that is not rejected at the 5% level
- The 95% CI contains the true β_1 in 95% of all samples.

Regression when X is Binary (Section 5.3)

Sometimes a regressor is binary:

- $X = 1$ if small class size, = 0 if not
- $X = 1$ if female, = 0 if male
- $X = 1$ if treated (experimental drug), = 0 if not

Binary regressors are sometimes called "dummy" variables.

So far, β_1 has been called a "slope," but that doesn't make sense if X is binary.

How do we interpret regression with a binary regressor?

Interpreting regressions with a binary regressor

$Y_i = \beta_0 + \beta_1 X_i + u_i$, where X is binary ($X_i = 0$ or 1):

When $X_i = 0$, $Y_i = \beta_0 + u_i$

- the mean of Y_i is β_0
- that is, $E(Y_i|X_i=0) = \beta_0$

When $X_i = 1$, $Y_i = \beta_0 + \beta_1 + u_i$

- the mean of Y_i is $\beta_0 + \beta_1$
- that is, $E(Y_i|X_i=1) = \beta_0 + \beta_1$

so:

$$\beta_1 = E(Y_i|X_i=1) - E(Y_i|X_i=0)$$

= population difference in group means

Example: Let $D_i = \begin{cases} 1 \text{ if } STR_i \le 20 \\ 0 \text{ if } STR_i > 20 \end{cases}$

OLS regression: $\widehat{TestScore} = \mathbf{650.0} + \mathbf{7.4} \times D$

(1.3) (**1.8**)

Tabulation of group means:

Class Size	**Average score ($\bar{Y}$)**	**Std. dev. (s_Y)**	***N***
Small ($STR > 20$)	657.4	19.4	238
Large ($STR \ge 20$)	**650.0**	17.9	182

Difference in means: $\bar{Y}_{\text{small}} - \bar{Y}_{\text{large}} = 657.4 - 650.0 = \mathbf{7.4}$

Standard error: $SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}} = \sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}} = \mathbf{1.8}$

Summary: regression when X_i is binary (0/1)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- β_0 = mean of Y when $X = 0$
- $\beta_0 + \beta_1$ = mean of Y when $X = 1$
- β_1 = difference in group means, $X = 1$ minus $X = 0$
- SE($\hat{\beta}_1$) has the usual interpretation
- t-statistics, confidence intervals constructed as usual
- This is another way (an easy way) to do difference-in-means analysis
- The regression formulation is especially useful when we have additional regressors (*as we will very soon*)

Heteroskedasticity and Homoskedasticity, and Homoskedasticity-Only Standard Errors (Section 5.4)

1. What…?
2. Consequences of homoskedasticity
3. Implication for computing standard errors

What do these two terms mean?

If var($u|X=x$) is constant – that is, if the variance of the conditional distribution of u given X does not depend on X – then u is said to be ***homoskedastic***. Otherwise, u is ***heteroskedastic***.

Example: hetero/homoskedasticity in the case of a binary regressor (that is, the comparison of means)

- Standard error when group variances are ***unequal***:

$$SE = \sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}$$

- Standard error when group variances are ***equal***:

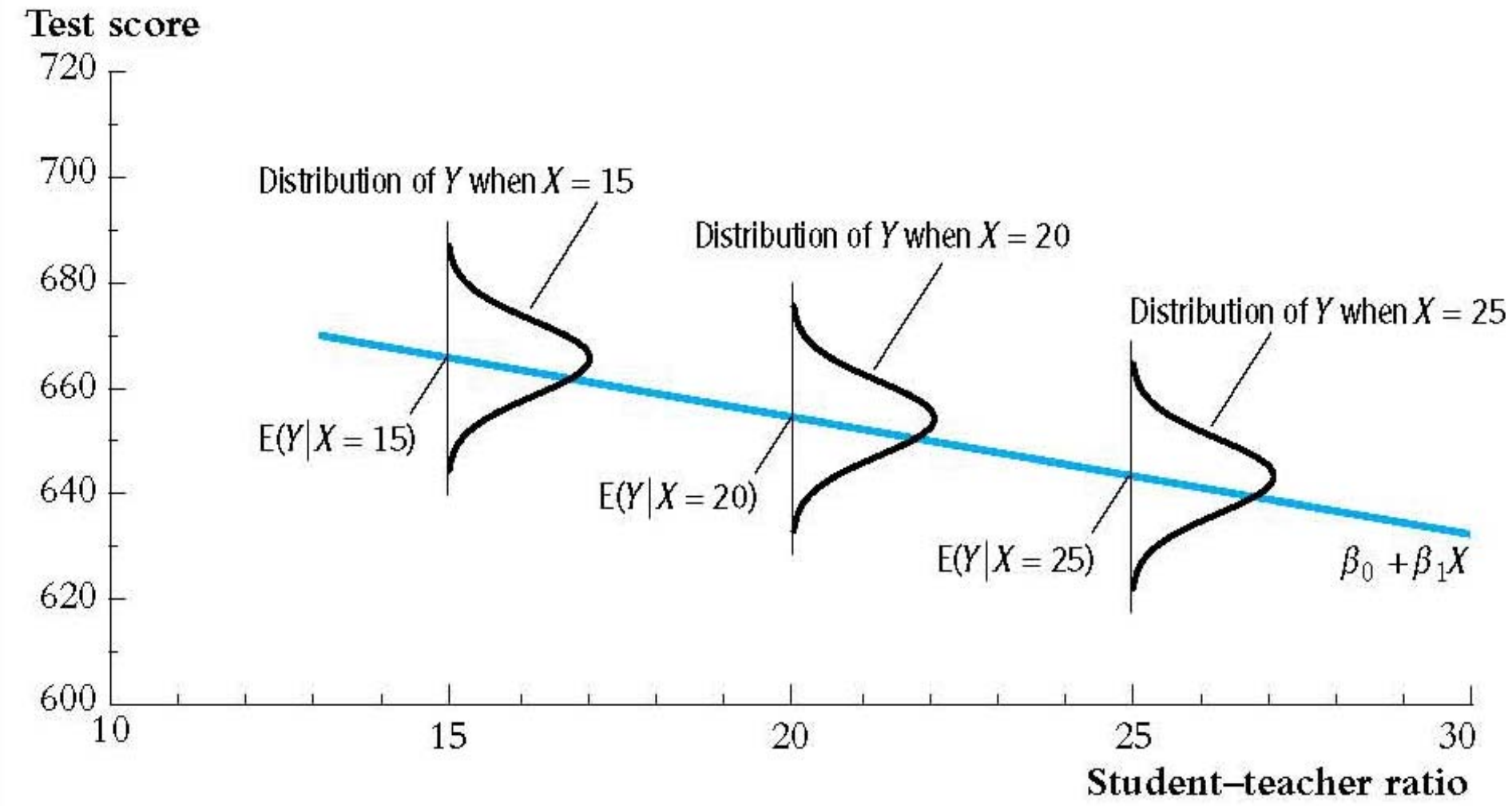
$$SE = s_p \sqrt{\frac{1}{n_s} + \frac{1}{n_l}}$$

where $s_p^2 = \dfrac{(n_s - 1)s_s^2 + (n_l - 1)s_l^2}{n_s + n_l - 2}$ (SW, Sect 3.6)

s_p = "pooled estimator of σ^2" when $\sigma_l^2 = \sigma_s^2$

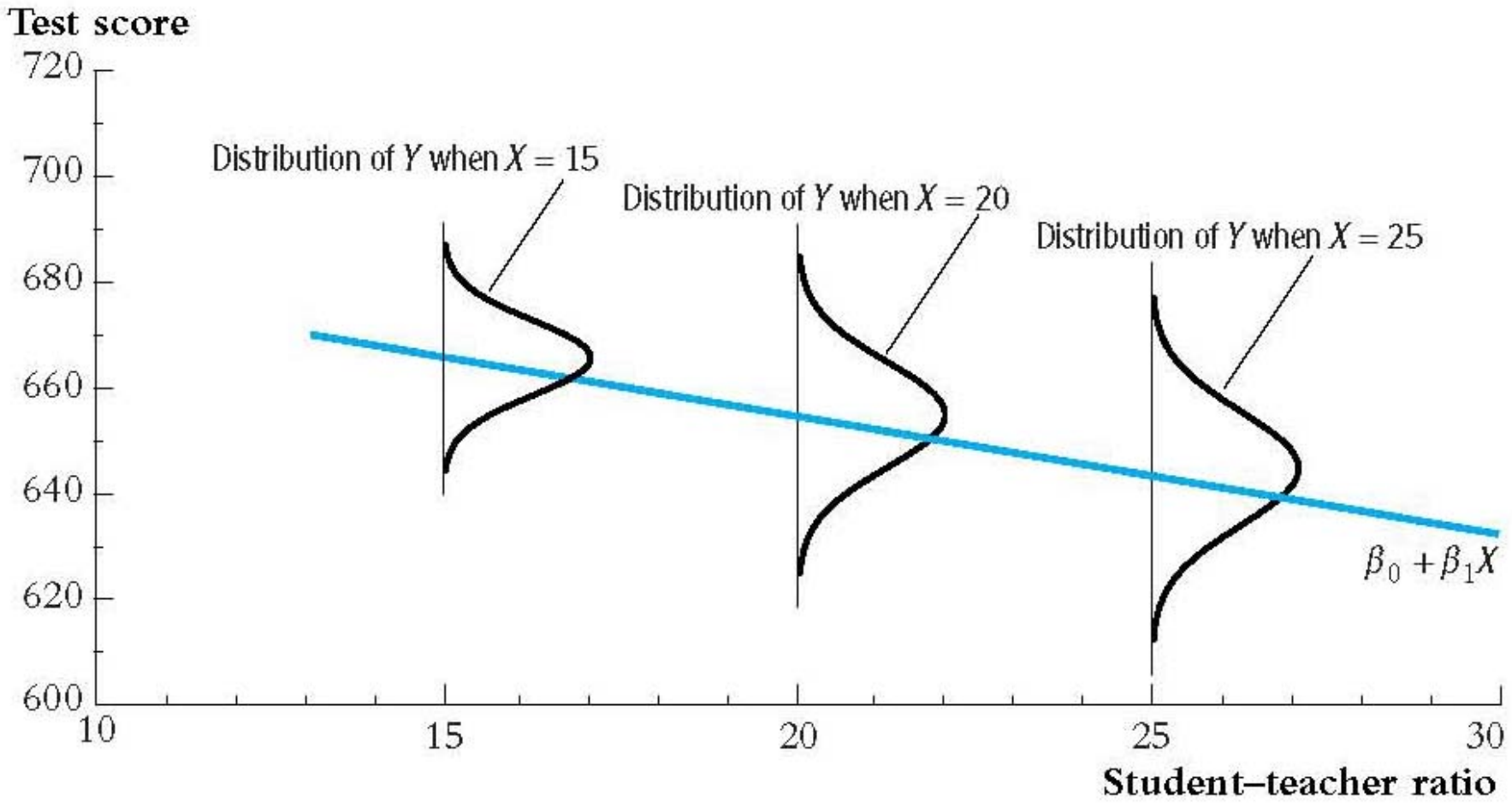
- **Equal** group variances = **homo**skedasticity
- **Unequal** group variances = **hetero**skedasticity

Homoskedasticity *in a picture:*



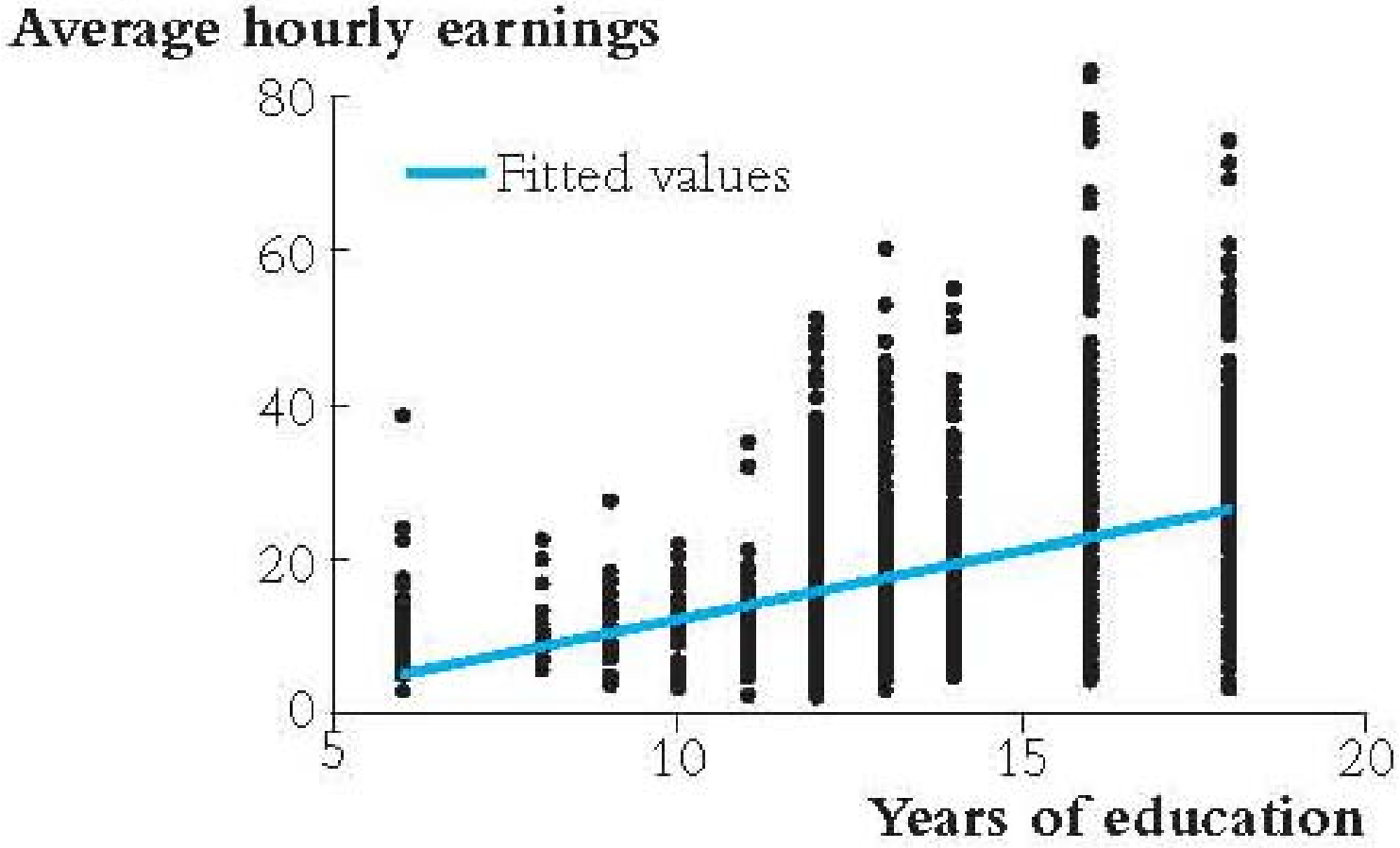
- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u ***does not*** depend on x

Heteroskedasticity *in a picture:*



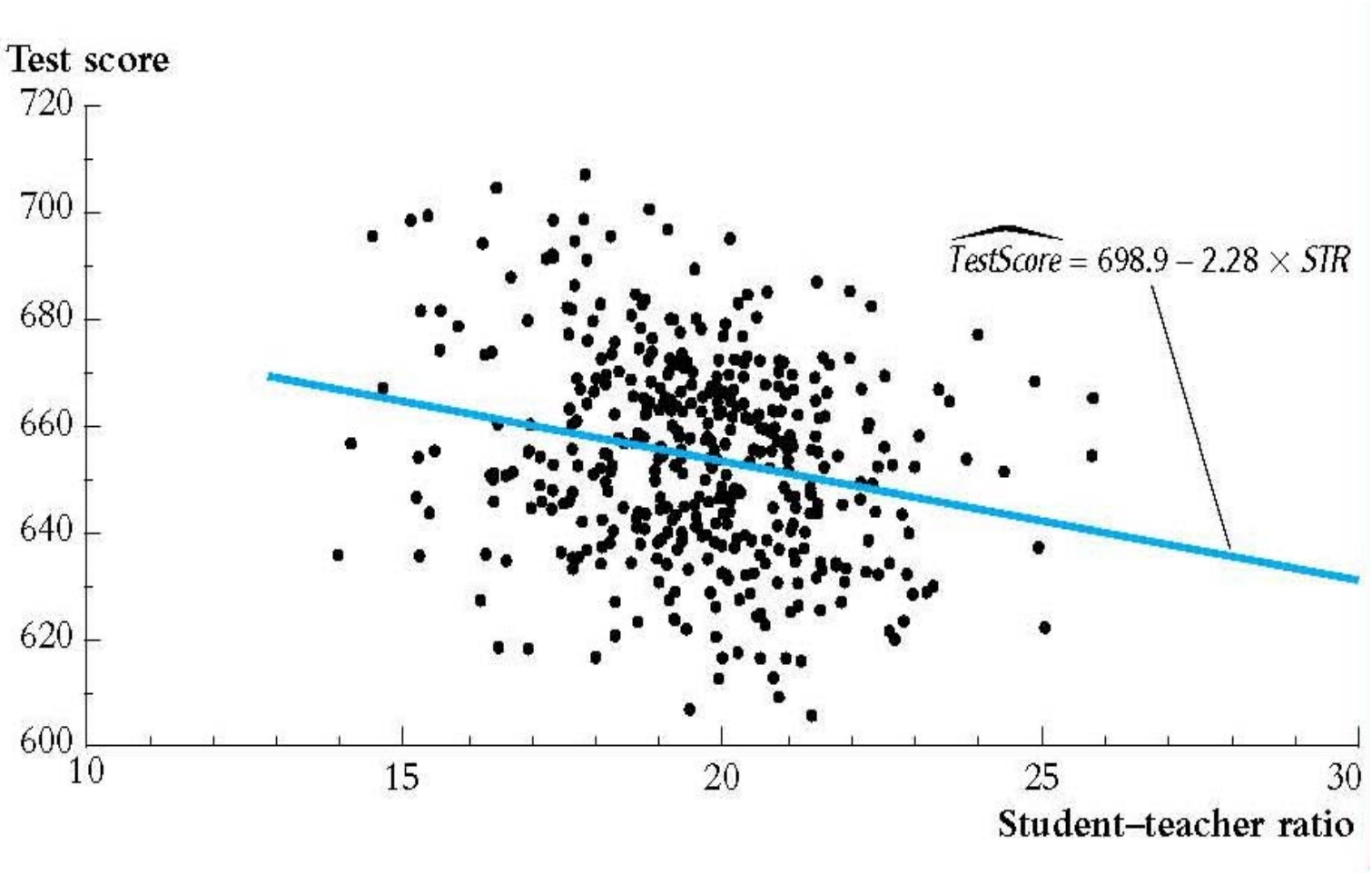
- $E(u|X=x) = 0$ (u satisfies Least Squares Assumption #1)
- The variance of u ***does*** depend on x: u is heteroskedastic.

A real-data example from labor economics: average hourly earnings vs. years of education (data source: Current Population Survey):



Heteroskedastic or homoskedastic?

The class size data:



Heteroskedastic or homoskedastic?

So far we have (without saying so) assumed that u might be heteroskedastic.

Recall the three least squares assumptions:

1. $E(u|X = x) = 0$
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare

Heteroskedasticity and homoskedasticity concern var($u|X=x$). Because we have not explicitly assumed homoskedastic errors, we have implicitly allowed for heteroskedasticity.

What if the errors are in fact homoskedastic?

- You can prove that OLS has the lowest variance among estimators that are linear in Y… a result called the Gauss-Markov theorem that we will return to shortly.
- The formula for the variance of $\hat{\beta}_1$ and the OLS standard error simplifies: If $\text{var}(u_i|X_i=x) = \sigma_u^2$, then

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}[(X_i - \mu_x)u_i]}{n(\sigma_X^2)^2} \quad \text{(general formula)}$$

$$= \frac{\sigma_u^2}{n\sigma_X^2} \quad \text{(simplification if } u \text{ is homoscedastic)}$$

Note: $\text{var}(\hat{\beta}_1)$ is inversely proportional to $\text{var}(X)$: more spread in X means more information about $\hat{\beta}_1$ – we discussed this earlier but it is clearer from this formula.

- Along with this homoskedasticity-only formula for the variance of $\hat{\beta}_1$, we have homoskedasticity-only standard errors:

  Homoskedasticity-only standard error formula:

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

  Some people (e.g. Excel programmers) find the homoskedasticity-only formula simpler – but it is wrong unless the errors really are homoskedastic.

We now have two formulas for standard errors for $\hat{\beta}_1$.

- ***Homoskedasticity-only standard errors*** – these are valid only if the errors are homoskedastic.
- The usual standard errors – to differentiate the two, it is conventional to call these ***heteroskedasticity – robust standard errors***, because they are valid whether or not the errors are heteroskedastic.
- The main advantage of the homoskedasticity-only standard errors is that the formula is simpler. But the disadvantage is that the formula is only correct if the errors are homoskedastic.

Practical implications…

- The homoskedasticity-only formula for the standard error of $\hat{\beta}_1$ and the "heteroskedasticity-robust" formula differ – so in general, *you get different standard errors using the different formulas*.
- Homoskedasticity-only standard errors are the default setting in regression software – sometimes the only setting (e.g. Excel). To get the general "heteroskedasticity-robust" standard errors you must override the default.
- **If you don't override the default and there is in fact heteroskedasticity, your standard errors (and *t*-statistics and confidence intervals) will be wrong – typically, homoskedasticity-only *SE*s are too small.**

Heteroskedasticity-robust standard errors in STATA

```
regress testscr str, robust

Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

- If you use the "**, robust**" option, STATA computes heteroskedasticity-robust standard errors
- Otherwise, STATA computes homoskedasticity-only standard errors

The bottom line:

- If the errors are either homoskedastic or heteroskedastic and you use heteroskedastic-robust standard errors, you are OK
- If the errors are heteroskedastic and you use the homoskedasticity-only formula for standard errors, your standard errors will be wrong (the homoskedasticity-only estimator of the variance of $\hat{\beta}_1$ is inconsistent if there is heteroskedasticity).
- The two formulas coincide (when n is large) in the special case of homoskedasticity
- So, you should always use heteroskedasticity-robust standard errors.

Some Additional Theoretical Foundations of OLS (Section 5.5)

We have already learned a very great deal about OLS: OLS is unbiased and consistent; we have a formula for heteroskedasticity-robust standard errors; and we can construct confidence intervals and test statistics.

Also, a very good reason to use OLS is that everyone else does – so by using it, others will understand what you are doing. In effect, OLS is the language of regression analysis, and if you use a different estimator, you will be speaking a different language.

Still, you may wonder…

- Is this really a good reason to use OLS? Aren't there other estimators that might be better – in particular, ones that might have a smaller variance?
- Also, what happened to our old friend, the Student *t* distribution?

So we will now answer these questions – but to do so we will need to make some stronger assumptions than the three least squares assumptions already presented.

The Extended Least Squares Assumptions

These consist of the three LS assumptions, plus two more:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u is homoskedastic
5. u is distributed $N(0,\sigma^2)$

- Assumptions 4 and 5 are more restrictive – so they apply to fewer cases in practice. However, if you make these assumptions, then certain mathematical calculations simplify and you can prove strong results – results that hold if these additional assumptions are true.
- We start with a discussion of the efficiency of OLS

Efficiency of OLS, part I: The Gauss-Markov Theorem

Under extended LS assumptions 1-4 (the basic three, plus homoskedasticity), $\hat{\beta}_1$ has the smallest variance among *all linear estimators* (estimators that are linear functions of $Y_1,\ldots, Y_n$). This is the ***Gauss-Markov theorem***.

Comments

- The GM theorem is proven in SW Appendix 5.2

The Gauss-Markov Theorem, ctd.

- $\hat{\beta}_1$ is a linear estimator, that is, it can be written as a linear function of $Y_1,\ldots, Y_n$:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{n}\sum_{i=1}^{n} w_i u_i ,$$

where $w_i = \dfrac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

- The G-M theorem says that among all possible choices of $\{w_i\}$, the OLS weights yield the smallest var($\hat{\beta}_1$)

Efficiency of OLS, part II:

- Under <u>all five</u> extended LS assumptions – including normally distributed errors – $\hat{\beta}_1$ has the smallest variance of <u>*all*</u> consistent estimators (linear *or* nonlinear functions of $Y_1,\ldots,Y_n$), as $n \to \infty$.
- This is a pretty amazing result – it says that, if (in addition to LSA 1-3) the errors are homoskedastic and normally distributed, then OLS is a better choice than any other consistent estimator. And because an estimator that isn't consistent is a poor choice, this says that OLS really is the best you can do – if all five extended LS assumptions hold. (The proof of this result is beyond the scope of this course and isn't in SW – it is typically done in graduate courses.)

Some not-so-good thing about OLS

The foregoing results are impressive, but these results – and the OLS estimator – have important limitations.

1. The GM theorem really isn't that compelling:
   - The condition of homoskedasticity often doesn't hold (homoskedasticity is special)
   - The result is only for linear estimators – only a small subset of estimators (more on this in a moment)

2. The strongest optimality result ("part II" above) requires homoskedastic normal errors – not plausible in applications (think about the hourly earnings data!)

Limitations of OLS, ctd.

3. OLS is more sensitive to outliers than some other estimators. In the case of estimating the population mean, if there are big outliers, then the median is preferred to the mean because the median is less sensitive to outliers – it has a smaller variance than OLS when there are outliers. Similarly, in regression, OLS can be sensitive to outliers, and if there are big outliers other estimators can be more efficient (have a smaller variance). One such estimator is the least absolute deviations (LAD) estimator:

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left| Y_i - (b_0 + b_1 X_i) \right|$$

In virtually all applied regression analysis, OLS is used – and that is what we will do in this course too.

Inference if *u* is homoskedastic and normally distributed: the Student *t* distribution (Section 5.6)

Recall the five extended LS assumptions:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare ($E(Y^4) < \infty$, $E(X^4) < \infty$).
4. u is homoskedastic
5. u is distributed $N(0,\sigma^2)$

If all five assumptions hold, then:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed *for all n* (!)
- the t-statistic has a Student t distribution with $n - 2$ degrees of freedom – this holds exactly *for all n* (!)

Normality of the sampling distribution of $\hat{\beta}_1$ under 1–5:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} w_i u_i \text{, where } w_i = \frac{(X_i - \bar{X})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

What is the distribution of a weighted average of normals?

Under assumptions 1 – 5:

$$\hat{\beta}_1 - \beta_1 \sim N\left(0, \frac{1}{n^2}\left(\sum_{i=1}^{n} w_i^2\right)\sigma_u^2\right) \qquad (*)$$

Substituting w_i into (*) yields the homoskedasticity-only variance formula.

In addition, under assumptions 1 – 5, under the null hypothesis the t statistic has a Student t distribution with $n - 2$ degrees of freedom

- Why $n - 2$? because we estimated 2 parameters, β_0 and β_1
- For $n < 30$, the t critical values can be a fair bit larger than the $N(0,1)$ critical values
- For $n > 50$ or so, the difference in t_{n-2} and $N(0,1)$ distributions is negligible. Recall the Student t table:

degrees of freedom	5% t-distribution critical value
10	2.23
20	2.09
30	2.04
60	2.00
∞	1.96

Practical implication:

- If $n < 50$ ***and*** you really believe that, for your application, u is homoskedastic and normally distributed, then use the t_{n-2} instead of the $N(0,1)$ critical values for hypothesis tests and confidence intervals.
- In most econometric applications, there is no reason to believe that u is homoskedastic and normal – usually, there are good reasons to believe that neither assumption holds.
- Fortunately, in modern applications, $n > 50$, so we can rely on the large-n results presented earlier, based on the CLT, to perform hypothesis tests and construct confidence intervals using the large-n normal approximation.

Summary and Assessment (Section 5.7)

- The initial policy question:

  Suppose new teachers are hired so the student-teacher ratio falls by one student per class. What is the effect of this policy intervention ("treatment") on test scores?

- Does our regression analysis using the California data set answer this convincingly?

  Not really – districts with low *STR* tend to be ones with lots of other resources and higher income families, which provide kids with more learning opportunities outside school…this suggests that corr(u_i, STR_i) > 0, so $E(u_i|X_i)\neq 0$.

- It seems that we have omitted some factors, or variables, from our analysis, and this has biased our results...

Linear Regression with Multiple Regressors
(SW Chapter 6)

Outline

1. Omitted variable bias
2. Causality and regression analysis
3. Multiple regression and OLS
4. Measures of fit
5. Sampling distribution of the OLS estimator

Omitted Variable Bias
(SW Section 6.1)

The error u arises because of factors, or variables, that influence Y but are not included in the regression function. There are always omitted variables.

Sometimes, the omission of those variables can lead to bias in the OLS estimator.

Omitted variable bias, ctd.

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called ***omitted variable*** bias. For omitted variable bias to occur, the omitted variable "*Z*" must satisfy two conditions:

<u>The two conditions for omitted variable bias</u>

(1) *Z* is a determinant of *Y* (i.e. *Z* is part of *u*); **and**

(2) *Z* is correlated with the regressor *X* (*i.e.* corr(*Z*,*X*) ≠ 0)

Both *conditions must hold for the omission of Z to result in omitted variable bias*.

Omitted variable bias, ctd.

In the test score example:

1. English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
2. Immigrant communities tend to be less affluent and thus have smaller school budgets and higher STR: Z is correlated with X.

Accordingly, $\hat{\beta}_1$ is biased. What is the direction of this bias?

- *What does common sense suggest?*
- If common sense fails you, there is a formula…

Omitted variable bias, ctd.

A formula for omitted variable bias: recall the equation,

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right)s_X^2}$$

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$. Under Least Squares Assumption #1,

$$E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = 0.$$

But what if $E[(X_i - \mu_X)u_i] = \text{cov}(X_i, u_i) = \sigma_{Xu} \neq 0$?

Omitted variable bias, ctd.

Under LSA #2 and #3 (that is, even if LSA #1 is not true),

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

$$= \left(\frac{\sigma_u}{\sigma_X}\right) \times \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu},$$

where $\rho_{Xu} = \text{corr}(X,u)$. If assumption #1 is correct, then $\rho_{Xu} = 0$, but if not we have….

The omitted variable bias formula:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right)\rho_{Xu}$$

- If an omitted variable Z is ***both***:
  (1) a determinant of Y (that is, it is contained in u); ***and***
  (2) correlated with X,

  then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.

- For example, districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect. *Is this is actually going on in the CA data*?

TABLE 6.1 Differences in Test Scores for California School Districts with Low and High Student–Teacher Ratios, by the Percentage of English Learners in the District

	Student–Teacher Ratio < 20		Student–Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All districts	657.4	238	650.0	182	7.4	4.04
Percentage of English learners						
< 1.9%	664.5	76	665.4	27	−0.9	−0.30
1.9–8.8%	665.2	64	661.8	44	3.3	1.13
8.8–23.0%	654.9	54	649.7	50	5.2	1.72
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes
- Among districts with comparable *PctEL*, the effect of class size is small (recall overall "test score gap" = 7.4)

Causality and regression analysis

The test score/*STR*/fraction English Learners example shows that, if an omitted variable satisfies the two conditions for omitted variable bias, then the OLS estimator in the regression omitting that variable is biased and inconsistent. So, even if n is large, $\hat{\beta}_1$ will not be close to β_1.

This raises a deeper question: how do we define β_1? That is, what precisely do we want to estimate when we run a regression?

What precisely do we want to estimate when we run a regression?

There are (at least) three possible answers to this question:

1. We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning.

   This can be useful at times, but isn't very interesting intellectually and isn't what this course is about.

2. We want to make forecasts, or predictions, of the value of *Y* for an entity not in the data set, for which we know the value of *X*.

Forecasting is an important job for economists, and excellent forecasts are possible using regression methods without needing to know causal effects. We will return to forecasting later in the course.

3. We want to estimate the causal effect on *Y* of a change in *X*.

This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect. Except when we discuss forecasting, the aim of this course is the estimation of causal effects using regression methods.

What, precisely, is a causal effect?

- "Causality" is a complex concept!

- In this course, we take a practical approach to defining causality:

  A causal effect is defined to be the effect measured in an ideal randomized controlled experiment.

Ideal Randomized Controlled Experiment

- *Ideal*: subjects all follow the treatment protocol – perfect compliance, no errors in reporting, etc.!
- *Randomized*: subjects from the population of interest are randomly assigned to a treatment or control group (so there are no confounding factors)
- *Controlled*: having a control group permits measuring the differential effect of the treatment
- *Experiment*: the treatment is assigned as part of the experiment: the subjects have no choice, so there is no "reverse causality" in which subjects choose the treatment they think will work best.

Back to class size:

Imagine an ideal randomized controlled experiment for measuring the effect on *Test Score* of reducing *STR*…

- In that experiment, students would be randomly assigned to classes, which would have different sizes.

- Because they are randomly assigned, all student characteristics (and thus u_i) would be distributed independently of STR_i.

- Thus, $E(u_i|STR_i) = 0$ – that is, LSA #1 holds in a randomized controlled experiment.

How does our observational data differ from this ideal?

- The treatment is not randomly assigned

- Consider *PctEL* – percent English learners – in the district. It plausibly satisfies the two criteria for omitted variable bias: $Z = PctEL$ is:
  (1) a determinant of Y; ***and***
  (2) correlated with the regressor X.

- Thus, the "control" and "treatment" groups differ in a systematic way, so corr(*STR*,*PctEL*) ≠ 0

- Randomization + control group means that any differences between the treatment and control groups are random – not systematically related to the treatment

- We can eliminate the difference in *PctEL* between the large (control) and small (treatment) groups by examining the effect of class size among districts with the same *PctEL*.
  - If the only systematic difference between the large and small class size groups is in *PctEL*, then we are back to the randomized controlled experiment – within each *PctEL* group.
  - This is one way to "control" for the effect of *PctEL* when estimating the effect of *STR*.

Return to omitted variable bias

Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment (*STR*) is randomly assigned: then *PctEL* is still a determinant of *TestScore*, but *PctEL* is uncorrelated with *STR*. (*This solution to OV bias is rarely feasible.*)
2. Adopt the "cross tabulation" approach, with finer gradations of *STR* and *PctEL* – within each group, all classes have the same *PctEL*, so we control for *PctEL* (*But soon you will run out of data, and what about other determinants like family income and parental education*?)
3. Use a regression in which the omitted variable (*PctEL*) is no longer omitted: include *PctEL* as an additional regressor in a multiple regression.

The Population Multiple Regression Model (SW Section 6.2)

Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1,\ldots,n$$

- Y is the *dependent variable*
- X_1, X_2 are the two *independent variables* (*regressors*)
- (Y_i, X_{1i}, X_{2i}) denote the i^{th} observation on Y, X_1, and X_2.
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding X_2 constant
- β_2 = effect on Y of a change in X_2, holding X_1 constant
- u_i = the regression error (omitted factors)

Interpretation of coefficients in multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1,\ldots,n$$

Consider changing X_1 by ΔX_1 while holding X_2 constant:
Population regression line ***before*** the change:

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

Population regression line, ***after*** the change:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$$

Before: $Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

After: $Y + \Delta Y = \beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2$

Difference: $\Delta Y = \beta_1 \Delta X_1$

So:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \textbf{ holding } \boldsymbol{X_2} \textbf{ constant}$$

$$\beta_2 = \frac{\Delta Y}{\Delta X_2}, \textbf{ holding } \boldsymbol{X_1} \textbf{ constant}$$

β_0 = predicted value of Y when $X_1 = X_2 = 0$.

The OLS Estimator in Multiple Regression (SW Section 6.3)

With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **This yields the OLS estimators of β_0 and β_1.**

Example: the California test score data

Regression of *TestScore* against *STR*:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

Now include percent English Learners in the district (*PctEL*):

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

- What happens to the coefficient on *STR*?
- Why? (*Note*: corr(*STR*, *PctEL*) = 0.19)

Multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$$

More on this printout later…

Measures of Fit for Multiple Regression (SW Section 6.4)

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$

SER = std. deviation of $\hat{u}_i$ (with d.f. correction)

RMSE = std. deviation of $\hat{u}_i$ (without d.f. correction)

R^2 = fraction of variance of Y explained by X

$\overline{R}^2$ = "adjusted R^2" = R^2 with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\overline{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the *SER* and the *RMSE* are measures of the spread of the *Y*s around the regression line:

$$SER = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

R^2 and $\bar{R}^2$ (adjusted R^2)

The R^2 is the fraction of the variance explained – same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

where $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2$, $SSR = \sum_{i=1}^{n}\hat{u}_i^2$, $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

- The R^2 always increases when you add another regressor (*why?)* – a bit of a problem for a measure of "fit"

R^2 and $\bar{R}^2$, ctd.

The $\bar{R}^2$ (the "adjusted R^2") corrects this problem by "penalizing" you for including another regressor – the $\bar{R}^2$ does not necessarily increase when you add another regressor.

$$\textit{Adjusted } R^2\text{: } \bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS}$$

Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.

Measures of fit, ctd.

Test score example:

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR$,

$R^2 = .05$, $SER = 18.6$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 PctEL$,

$R^2 = .426$, $\overline{R}^2 = .424$, $SER = 14.5$

- *What – precisely – does this tell you about the fit of regression (2) compared with regression (1)?*
- *Why are the R^2 and the $\overline{R}^2$ so close in (2)?*

The Least Squares Assumptions for Multiple Regression (SW Section 6.5)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

1. The conditional distribution of u given the X's has mean zero, that is, $E(u_i|X_{1i} = x_1,\ldots, X_{ki} = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki}, Y_i)$, $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are unlikely: $X_1,\ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty,\ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

Assumption #1: the conditional mean of *u* given the included *X*s is zero.

$$E(u|X_1 = x_1, \ldots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to omitted variable bias, specifically, if an omitted variable
  (1) belongs in the equation (so is in u) ***and***
  (2) is correlated with an included X
  then this condition fails and there is OV bias.
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (discussed in Ch. 7)

Assumption #2: $(X_{1i},\ldots,X_{ki},Y_i)$, $i = 1,\ldots,n$, are i.i.d.

This is satisfied automatically if the data are collected by simple random sampling.

Assumption #3: large outliers are rare (finite fourth moments)

This is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).

Assumption #4: There is no perfect multicollinearity

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Example: Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
Regression with robust standard errors                Number of obs =     420
                                                      F(  1,   418) =   19.26
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.0512
                                                      Root MSE      =  18.581
------------------------------------------------------------------------------
             |               Robust
    testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -2.279808   .5194892    -4.39   0.000    -3.300945   -1.258671
         str |  (dropped)
       _cons |    698.933   10.36436    67.44   0.000     678.5602    719.3057
------------------------------------------------------------------------------
```

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression, β_1 is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples…

With these least squares assumptions in hand, we now can derive the sampling distribution of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.

The Sampling Distribution of the OLS Estimator (SW Section 6.6)

Under the four Least Squares Assumptions,

- The sampling distribution of $\hat{\beta}_1$ has mean β_1
- var($\hat{\beta}_1$) is inversely proportional to n.
- Other than its mean and variance, the exact (finite-n) distribution of $\hat{\beta}_1$ is very complicated; but for large n…
  - $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (law of large numbers)
  - $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT)
  - These statements hold for $\hat{\beta}_1, \ldots, \hat{\beta}_k$

Conceptually, there is nothing new here!

Multicollinearity, Perfect and Imperfect (SW Section 6.7)

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

Some more examples of perfect multicollinearity

1. The example from before: you include *STR* twice,
2. Regress *TestScore* on a constant, *D*, and *B*, where: D_i = 1 if $STR \leq 20$, = 0 otherwise; B_i = 1 if $STR > 20$, = 0 otherwise, so $B_i = 1 - D_i$ and there is perfect multicollinearity.
3. Would there be perfect multicollinearity if the intercept (constant) were excluded from this regression? This example is a special case of…

The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables *and* a constant, you will have perfect multicollinearity – this is sometimes called ***the dummy variable trap***.

- *Why is there perfect multicollinearity here*?
- *Solutions to the dummy variable trap*:
  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept
- *What are the implications of (1) or (2) for the interpretation of the coefficients?*

Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know – either by crashing or giving an error message or by "dropping" one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names.

Imperfect multicollinearity occurs when two or more regressors are very highly correlated.

- Why the term "multicollinearity"? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line – they are "co-linear" – but unless the correlation is exactly ±1, that collinearity is imperfect.

Imperfect multicollinearity, ctd.

Imperfect multicollinearity implies that one or more of the regression coefficients will be imprecisely estimated.

- The idea: the coefficient on X_1 is the effect of X_1 holding X_2 constant; but if X_1 and X_2 are highly correlated, there is very little variation in X_1 once X_2 is held constant – so the data don't contain much information about what happens when X_1 changes but X_2 doesn't. If so, the variance of the OLS estimator of the coefficient on X_1 will be large.
- Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.
- The math? See SW, App. 6.2

Next topic: hypothesis tests and confidence intervals…

Hypothesis Tests and Confidence Intervals in Multiple Regression

(SW Chapter 7)

Outline

1. Hypothesis tests and confidence intervals for one coefficient
2. Joint hypothesis tests on multiple coefficients
3. Other types of hypotheses involving multiple coefficients
4. Variables of interest, control variables, and how to decide which variables to include in a regression model

Hypothesis Tests and Confidence Intervals for a Single Coefficient
(SW Section 7.1)

Hypothesis tests and confidence intervals for a single coefficient in multiple regression follow the same logic and recipe as for the slope coefficient in a single-regressor model.

- $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT).
- Thus hypotheses on β_1 can be tested using the usual t-statistic, and confidence intervals are constructed as $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$.
- So too for $\beta_2, \ldots, \beta_k$.

***Example*:** The California class size data

(1) $\widehat{TestScore} = 698.9 - 2.28 \times STR$

(10.4) (0.52)

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$

(8.7) (0.43) (0.031)

- The coefficient on *STR* in (2) is the effect on *TestScores* of a unit change in *STR*, holding constant the percentage of English Learners in the district
- The coefficient on *STR* falls by one-half
- The 95% confidence interval for coefficient on *STR* in (2) is $\{-1.10 \pm 1.96 \times 0.43\} = (-1.95, -0.26)$
- The *t*-statistic testing $\beta_{STR} = 0$ is $t = -1.10/0.43 = -2.54$, so we reject the hypothesis at the 5% significance level

Standard errors in multiple regression in STATA

```
reg testscr str pctel, robust;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  223.82
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4264
                                                      Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754    703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$$
$$(8.7) \quad (0.43) \qquad (0.031)$$

We use heteroskedasticity-robust standard errors – for exactly the same reason as in the case of a single regressor.

Tests of Joint Hypotheses
(SW Section 7.2)

Let *Expn* = expenditures per pupil and consider the population regression model:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

The null hypothesis that "school resources don't matter," and the alternative that they do, corresponds to:

$$H_0: \beta_1 = 0 \textbf{ and } \beta_2 = 0$$

$$\text{vs. } H_1: \textbf{either } \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \textbf{ or both}$$

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Tests of joint hypotheses, ctd.

$$H_0: \beta_1 = 0 \textbf{ and } \beta_2 = 0$$

$$\text{vs. } H_1: \textbf{either } \beta_1 \neq 0 \textbf{ or } \beta_2 \neq 0 \textbf{ or both}$$

- A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this "one at a time" test isn't valid: the resulting test rejects too often under the null hypothesis (more than 5%)!

Why can't we just test the coefficients one at a time?

Because the rejection rate under the null isn't 5%. We'll calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual *t*-statistics. To simplify the calculation, suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are independently distributed (this isn't true in general – just in this example). Let t_1 and t_2 be the *t*-statistics:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \text{ and } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

The "one at time" test is:

reject H_0: $\beta_1 = \beta_2 = 0$ if $|t_1| > 1.96$ and/or $|t_2| > 1.96$

What is the probability that this "one at a time" test rejects H_0, when H_0 is actually true? (It *should* be 5%.)

***Suppose t_1 and t_2 are independent* (for this example).**

The probability of incorrectly rejecting the null hypothesis using the "one at a time" test

$= \Pr_{H_0}[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96]$

$= 1 - \Pr_{H_0}[|t_1| \le 1.96 \text{ and } |t_2| \le 1.96]$

$= 1 - \Pr_{H_0}[|t_1| \le 1.96] \times \Pr_{H_0}[|t_2| \le 1.96]$

(because t_1 and t_2 are independent by assumption)

$= 1 - (.95)^2$

$= .0975 = 9.75\%$ – which is ***not*** the desired 5%!!

The ***size*** of a test is the actual rejection rate under the null hypothesis.

- The size of the "common sense" test isn't 5%!
- In fact, its size depends on the correlation between t_1 and t_2 (and thus on the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$).

Two Solutions:

- Use a different critical value in this procedure – not 1.96 (this is the "Bonferroni method – see SW App. 7.1) (this method is rarely used in practice however)
- Use a different test statistic designed to test *both* β_1 and β_2 at once: the F-statistic (this is common practice)

The *F*-statistic

The *F*-statistic tests all parts of a joint hypothesis at once.

Formula for the special case of the joint hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ in a regression with two regressors:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t_1 and t_2.

Reject when *F* is large (how large?)

The F-statistic testing β_1 and β_2:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

- The F-statistic is large when t_1 and/or t_2 is large
- The F-statistic corrects (in just the right way) for the correlation between t_1 and t_2.
- The formula for more than two β's is nasty unless you use matrix algebra.
- This gives the F-statistic a nice large-sample approximate distribution, which is…

Large-sample distribution of the *F*-statistic

Consider the *special case* that t_1 and t_2 are independent, so $\hat{\rho}_{t_1,t_2} \xrightarrow{p} 0$; in large samples the formula becomes

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- Under the null, t_1 and t_2 have standard normal distributions that, in this special case, are independent
- The large-sample distribution of the *F*-statistic is the distribution of the average of two independently distributed squared standard normal random variables.

The chi-squared distribution

The ***chi-squared*** distribution with q degrees of freedom (χ^2_q) is defined to be the distribution of the sum of q independent squared standard normal random variables.

In large samples, F is distributed as χ^2_q/q.

Selected large-sample critical values of χ^2_q/q

q	5% critical value	
1	3.84	(*why*?)
2	3.00	(the case q=2 above)
3	2.60	
4	2.37	
5	2.21	

***Computing the p-value using the F-statistic*:**

p-value = tail probability of the χ^2_q/q distribution beyond the *F*-statistic actually computed.

Implementation in STATA

Use the "test" command after the regression

Example: Test the joint hypothesis that the population coefficients on *STR* and expenditures per pupil (*expn_stu*) are both zero, against the alternative that at least one of the population coefficients is nonzero.

F-test example, California class size data:

```
reg testscr str expn_stu pctel, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  147.20
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4366
                                                      Root MSE      =  14.353

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------

test str expn_stu;

 ( 1)  str = 0.0
 ( 2)  expn_stu = 0.0

       F(  2,   416) =    5.43
      Prob > F =    0.0047
```

NOTE

The test command follows the regression

There are q=2 restrictions being tested

The 5% critical value for q=2 is 3.00

Stata computes the p-value for you

More on *F*-statistics.

There is a simple formula for the F-statistic that holds only under homoskedasticity (so it isn't very useful) but which nevertheless might help you understand what the F-statistic is doing.

The homoskedasticity-only *F*-statistic

When the errors are homoskedastic, there is a simple formula for computing the "homoskedasticity-only" *F*-statistic:

- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- Compare the fits of the regressions – the R^2s – if the "unrestricted" model fits sufficiently better, reject the null

The "restricted" and "unrestricted" regressions

Example: are the coefficients on *STR* and *Expn* zero?

Unrestricted population regression (under H_1):

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

Restricted population regression (that is, under H_0):

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i \qquad (why?)$$

- The number of restrictions under H_0 is $q = 2$ (*why*?).
- The fit will be better (R^2 will be higher) in the unrestricted regression (*why*?)

By how much must the R^2 increase for the coefficients on *Expn* and *PctEL* to be judged statistically significant?

Simple formula for the homoskedasticity-only F-statistic:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

where:

$R^2_{restricted}$ = the R^2 for the restricted regression

$R^2_{unrestricted}$ = the R^2 for the unrestricted regression

q = the number of restrictions under the null

$k_{unrestricted}$ = the number of regressors in the unrestricted regression.

- The bigger the difference between the restricted and unrestricted R^2s – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only F.

Example:

Restricted regression:

$$\widehat{TestScore} = 644.7 - 0.671 PctEL, \quad R^2_{restricted} = 0.4149$$
$$(1.0) \quad (0.032)$$

Unrestricted regression:

$$\widehat{TestScore} = 649.6 - 0.29 STR + 3.87 Expn - 0.656 PctEL$$
$$(15.5) \quad (0.48) \qquad (1.59) \qquad (0.032)$$

$$R^2_{unrestricted} = 0.4366, \; k_{unrestricted} = 3, \; q = 2$$

So $$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

$$= \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = \mathbf{8.01}$$

Note: Heteroskedasticity-robust $F = \mathbf{5.43}\ldots$

The homoskedasticity-only F-statistic – summary

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

- The homoskedasticity-only F-statistic rejects when adding the two variables increased the R^2 by "enough" – that is, when adding the two variables improves the fit of the regression by "enough"
- If the errors are homoskedastic, then the homoskedasticity-only F-statistic has a large-sample distribution that is χ^2_q/q.
- But if the errors are heteroskedastic, the large-sample distribution of the homoskedasticity-only F-statistic is <u>not</u> χ^2_q/q

The *F* distribution

Your regression printouts might refer to the "*F*" distribution.

If the four multiple regression LS assumptions hold ***and if***:

5. u_i is homoskedastic, that is, var($u|X_1,\ldots,X_k$) does not depend on X's
6. $u_1,\ldots,u_n$ are normally distributed

then the homoskedasticity-only F-statistic has the "$F_{q,n-k-1}$" distribution, where q = the number of restrictions and k = the number of regressors under the alternative (the unrestricted model).

- **The *F* distribution is to the χ^2_q/q distribution what the t_{n-1} distribution is to the $N(0,1)$ distribution**

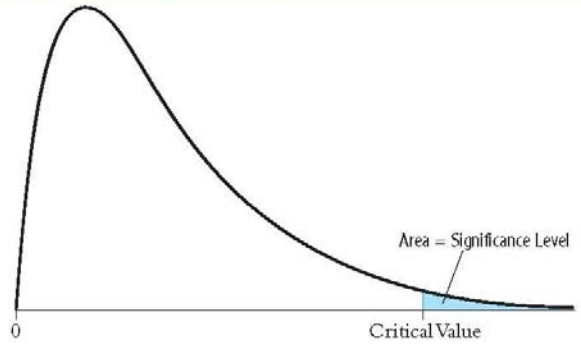
The $F_{q,n-k-1}$ distribution:

- The F distribution is tabulated many places
- As $n \to \infty$, the $F_{q,n-k-1}$ distribution asymptotes to the χ^2_q/q distribution:

  The $F_{q,\infty}$ and χ^2_q/q distributions are the same.

- For q not too big and $n \geq 100$, the $F_{q,n-k-1}$ distribution and the χ^2_q/q distribution are essentially identical.
- Many regression packages (including STATA) compute p-values of F-statistics using the F distribution
- You will encounter the F distribution in published empirical work.

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

Another digression: A little history of statistics…

- The theory of the homoskedasticity-only F-statistic and the $F_{q,n-k-1}$ distributions rests on implausibly strong assumptions (are earnings normally distributed?)
- These statistics date to the early 20^{th} century… the days when data sets were small and computers were people…
- The F-statistic and $F_{q,n-k-1}$ distribution were major breakthroughs: an easily computed formula; a single set of tables that could be published once, then applied in many settings; and a precise, mathematically elegant justification.

A little history of statistics, ctd…

- The strong assumptions were a minor price for this breakthrough.
- But with modern computers and large samples we can use the heteroskedasticity-robust F-statistic and the $F_{q,\infty}$ distribution, which only require the four least squares assumptions (not assumptions #5 and #6)
- This historical legacy persists in modern software, in which homoskedasticity-only standard errors (and F-statistics) are the default, and in which p-values are computed using the $F_{q,n-k-1}$ distribution.

Summary: the homoskedasticity-only F-statistic and the F distribution

- These are justified only under very strong conditions – stronger than are realistic in practice.
- *You* should use the heteroskedasticity-robust F-statistic, with χ^2_q/q (that is, $F_{q,\infty}$) critical values.
- For $n \geq 100$, the F-distribution essentially is the χ^2_q/q distribution.
- For small n, sometimes researchers use the F distribution because it has larger critical values and in this sense is more conservative.

Summary: testing joint hypotheses

- The "one at a time" approach of rejecting if either of the *t*-statistics exceeds 1.96 rejects more than 5% of the time under the null (the size exceeds the desired significance level)
- The heteroskedasticity-robust *F*-statistic is built in to STATA ("test" command); this tests all *q* restrictions at once.
- For *n* large, the *F*-statistic is distributed χ^2_q/q $(= F_{q,\infty})$
- The homoskedasticity-only *F*-statistic is important historically (and thus in practice), and can help intuition, but isn't valid when there is heteroskedasticity

Testing Single Restrictions on Multiple Coefficients (SW Section 7.3)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1,\ldots,n$$

Consider the null and alternative hypothesis,

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

This null imposes a *single* restriction ($q = 1$) on *multiple* coefficients – it is not a joint hypothesis with multiple restrictions (compare with $\beta_1 = 0$ and $\beta_2 = 0$).

Testing single restrictions on multiple coefficients, ctd.

Here are two methods for testing single restrictions on multiple coefficients:

1. ***Rearrange ("transform") the regression***

   Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression; or,

2. ***Perform the test directly***

   Some software, including STATA, lets you test restrictions using multiple coefficients directly

Method 1: Rearrange ("transform") the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

Add and subtract $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i$$

or

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

where

$$\gamma_1 = \beta_1 - \beta_2$$

$$W_i = X_{1i} + X_{2i}$$

Rearrange the regression, ctd.

(a) Original equation:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

(b) Rearranged ("transformed") equation:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$

so

$$H_0: \gamma_1 = 0 \quad \text{vs.} \quad H_1: \gamma_1 \neq 0$$

- These two regressions ((a) and (b)) have the same R^2, the same predicted values, and the same residuals.
- The testing problem is now a simple one: test whether $\gamma_1 = 0$ in regression (b).

Method 2: Perform the test directly

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

Example:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

In STATA, to test $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ (two-sided):

```
regress testscore str expn pctel, r
test str=expn
```

The details of implementing this method are software-specific.

Confidence Sets for Multiple Coefficients (SW Section 7.4)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

What is a *joint* confidence set for β_1 and β_2?

A 95% ***joint confidence set*** is:

- A set-valued function of the data that contains the true coefficient(s) in 95% of hypothetical repeated samples.
- Equivalently, the set of coefficient values that cannot be rejected at the 5% significance level.

You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an F-test (*why not just combine the two 95% confidence intervals?*).

Joint confidence sets ctd.

Let $F(\beta_{1,0}, \beta_{2,0})$ be the (heteroskedasticity-robust) F-statistic testing the hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$:

95% confidence set = $\{\beta_{1,0}, \beta_{2,0}: \ F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$

- 3.00 is the 5% critical value of the $F_{2,\infty}$ distribution
- This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%

  5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

The confidence set based on the F-statistic is an ellipse:

$$\{\beta_1, \beta_2:\ F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \leq 3.00\}$$

Now

$$F = \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)} \times \left[t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2\right]$$

$$= \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)} \times$$

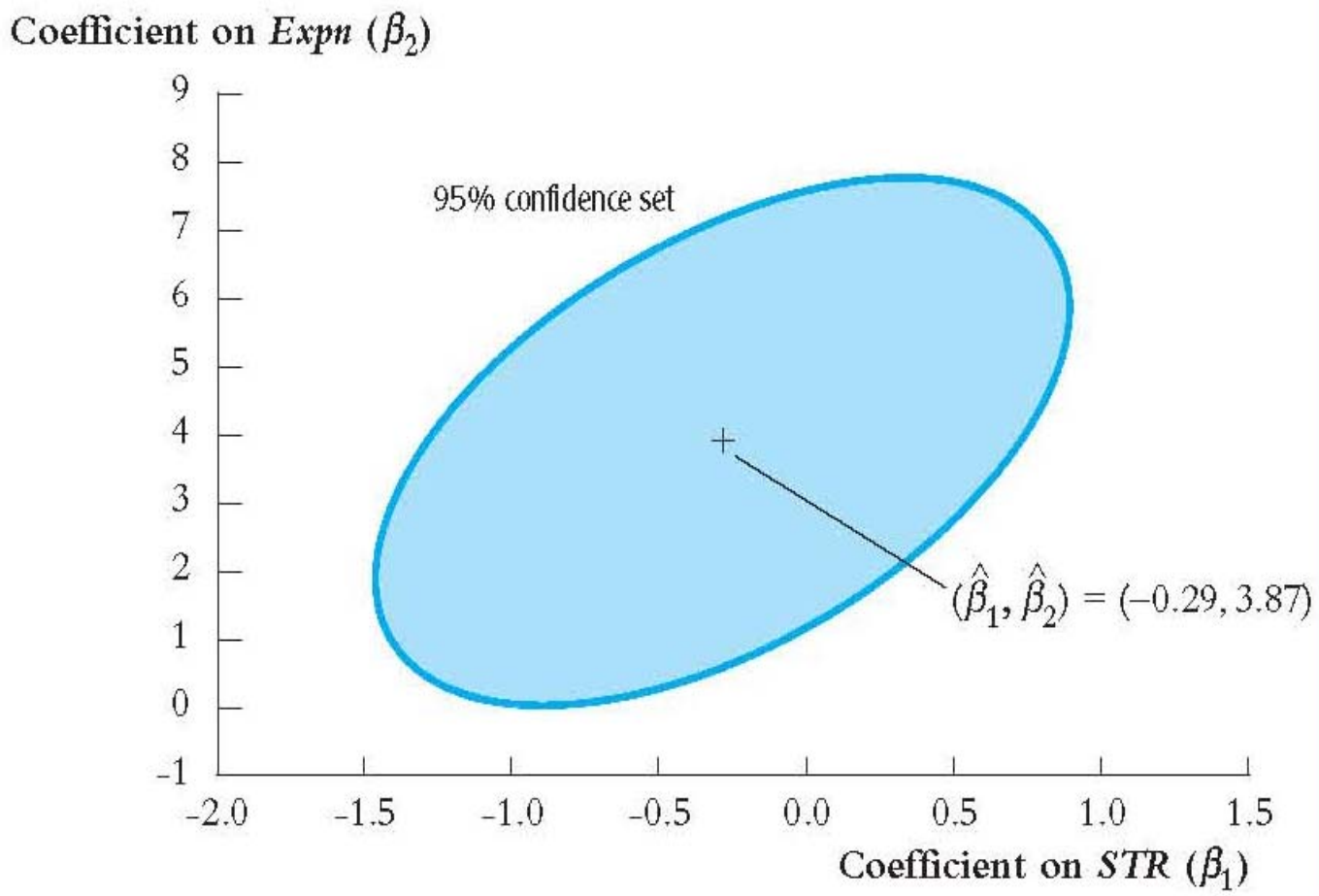
$$\left[\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

This is a quadratic form in $\beta_{1,0}$ and $\beta_{2,0}$ – thus the boundary of the set $F = 3.00$ is an ellipse.

Confidence set based on inverting the F-statistic

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



Regression Specification: variables of interest, control variables, and conditional mean independence (SW Section 7.5)

We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant factors outside the school committee's control – such as outside learning opportunities (museums, etc), parental involvement in education (reading with mom at home?), etc.

If we could run an experiment, we would randomly assign students (and teachers) to different sized classes. Then STR_i would be independent of all the things that go into u_i, so $E(u_i|STR_i) = 0$ and the OLS slope estimator in the regression of $TestScore_i$ on STR_i will be an unbiased estimator of the desired causal effect.

But with observational data, u_i depends on additional factors (museums, parental involvement, knowledge of English etc).

- If you can observe those factors (e.g. *PctEL*), then include them in the regression.
- But usually you can't observe all these omitted causal factors (e.g. parental involvement in homework). ***In this case, you can include "control variables" which are correlated with these omitted causal factors, but which themselves are not causal***.

Control variables in multiple regression

A **control variable** ***W*** is a variable that is correlated with, and controls for, an omitted causal factor in the regression of *Y* on *X*, but which itself does not necessarily have a causal effect on *Y*.

Control variables: an example from the California test score data

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\ \ (0.27)\qquad (.033)\qquad\qquad (.024)$$

PctEL = percent English Learners in the school district
LchPct = percent of students receiving a free/subsidized lunch (only students from low-income families are eligible)

- Which variable is the variable of interest?
- Which variables are control variables? Do they have causal components? What do they control for?

Control variables example, ctd.

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 = 0.773$$
$$(5.6)\ \ (0.27) \qquad (.033) \qquad (.024)$$

- *STR* is the variable of interest
- *PctEL* probably has a direct causal effect (school is tougher if you are learning English!). But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and *PctEL* is correlated with those omitted causal variables. *PctEL is both a possible causal variable and a control variable.*
- *LchPct* might have a causal effect (eating lunch helps learning); it also is correlated with and controls for income-related outside learning opportunities. *LchPct is both a possible causal variable and a control variable.*

Control variables, ctd.

1. **Three interchangeable statements about what makes an effective control variable:**
   i. An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.
   ii. Holding constant the control variable(s), the variable of interest is "as if" randomly assigned.
   iii. Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of Y

Control variables, ctd.

2. Control variables need not be causal, and their coefficients generally do not have a causal interpretation. For example:

$$\widehat{TestScore} = 700.2 - 1.00STR - 0.122PctEL - 0.547LchPct,\ \bar{R}^2 0.773$$
$$(5.6)\ \ (0.27) \qquad (.033) \qquad (.024)$$

- Does the coefficient on *LchPct* have a causal interpretation? If so, then we should be able to boost test scores (by a lot! Do the math!) by simply eliminating the school lunch program, so that *LchPct* = 0! (Eliminating the school lunch program has a well-defined causal effect: we could construct a randomized experiment to measure the causal effect of this intervention.)

The math of control variables: conditional mean independence.

- Because the coefficient on a control variable can be biased, LSA #1 ($E(u_i|X_{1i},\ldots,X_{ki}) = 0$) must not hold. For example, the coefficient on *LchPct* is correlated with unmeasured determinants of test scores such as outside learning opportunities, so is subject to OV bias. But the fact that *LchPct* is correlated with these omitted variables is precisely what makes it a good control variable!
- If LSA #1 doesn't hold, then what does?
- We need a mathematical statement of what makes an effective control variable. This condition is **conditional mean independence**: given the control variable, the mean of u_i doesn't depend on the variable of interest

Conditional mean independence, ctd.

Let X_i denote the variable of interest and W_i denote the control variable(s). W is an effective control variable if conditional mean independence holds:

$$E(u_i|X_i, W_i) = E(u_i|W_i) \text{ (conditional mean independence)}$$

If W is a control variable, then conditional mean independence replaces LSA #1 – it is the version of LSA #1 which is relevant for control variables.

Conditional mean independence, ctd.

Consider the regression model,

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

where X is the variable of interest and W is an effective control variable so that conditional mean independence holds:

$$E(u_i|X_i, W_i) = E(u_i|W_i).$$

In addition, suppose that LSA #2, #3, and #4 hold. Then:

1. β_1 has a causal interpretation.
2. $\hat{\beta}_1$ is unbiased
3. The coefficient on the control variable, $\hat{\beta}_2$, is in general biased.

The math of conditional mean independence

Under conditional mean independence:

1. β_1 has a causal interpretation.

The math: The expected change in Y resulting from a change in X, holding (a single) W constant, is:

$$
\begin{aligned}
&E(Y|X = x+\Delta x, W=w) - E(Y|X = x, W=w) \\
&\quad = [\beta_0 + \beta_1(x+\Delta x) + \beta_2 w + E(u|X = x+\Delta x, W=w)] \\
&\qquad - [\beta_0 + \beta_1 x + \beta_2 w + E(u|X = x, W=w)] \\
&\quad = \beta_1 \Delta x + [E(u|X = x+\Delta x, W=w) - E(u|X = x, W=w)] \\
&\quad = \beta_1 \Delta x
\end{aligned}
$$

where the final line follows from conditional mean independence: under conditional mean independence, $E(u|X = x+\Delta x, W=w) = E(u|X = x, W=w) = E(u|W=w)$.

The math of conditional mean independence, ctd.

Under conditional mean independence:

2. $\hat{\beta}_1$ is unbiased
3. $\hat{\beta}_2$ is in general biased

The math: Consider the regression model,

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

where u satisfies the conditional mean independence assumption. For convenience, suppose that $E(u|W) = \gamma_0 + \gamma_2 W$ (that is, that $E(u|W)$ is linear in W). Thus, under conditional mean independence,

The math of conditional mean independence, ctd.

$$E(u|X, W) = E(u|W) = \gamma_0 + \gamma_2 W. \qquad (*)$$

Let

$$v = u - E(u|X, W) \qquad (**)$$

so that $E(v|X, W) = 0$. Combining (*) and (**) yields,

$$\begin{aligned} u &= E(u|X, W) + v \\ &= \gamma_0 + \gamma_2 W + v, \text{ where } E(v|X, W) = 0 \end{aligned} \qquad (***)$$

Now substitute (***) into the regression,

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u \qquad (+)$$

So that

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \beta_2 W + u && (+)\\
&= \beta_0 + \beta_1 X + \beta_2 W + \gamma_0 + \gamma_2 W + v \quad \text{from (***)}\\
&= (\beta_0 + \gamma_0) + \beta_1 X + (\beta_2 + \gamma_2) W + v\\
&= \delta_0 + \beta_1 X + \delta_2 W + v && (++)
\end{aligned}
$$

- Because $E(v|X, W) = 0$, equation (++) satisfies LSA#1 so the OLS estimators of δ_0, β_1, and δ_2 in (++) are unbiased.
- Because the regressors in (+) and (++) are the same, the OLS coefficients in regression (+) satisfy, $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$ in general.

$$E(\hat{\beta}_1) = \beta_1$$

and

$$E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$$

In summary, if *W* is such that conditional mean independence is satisfied, then:

- The OLS estimator of the effect of interest, $\hat{\beta}_1$, is unbiased.
- The OLS estimator of the coefficient on the control variable, $\hat{\beta}_2$, is biased. This bias stems from the fact that the control variable is correlated with omitted variables in the error term, so that $\hat{\beta}_2$ is subject to omitted variable bias.

Implications for variable selection and "*model specification*"

1. Identify the variable of interest

2. Think of the omitted causal effects that could result in omitted variable bias

3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables. The control variables are effective if the conditional mean independence assumption plausibly holds (if u is uncorrelated with STR once the control variables are included). This results in a "base" or "benchmark" model.

Model specification, ctd.

4. Also specify a range of plausible alternative models, which include additional candidate variables.

5. Estimate your base model and plausible alternative specifications ("sensitivity checks").
   - Does a candidate variable change the coefficient of interest (β_1)?
   - Is a candidate variable statistically significant?
   - Use judgment, not a mechanical recipe…
   - Don't just try to maximize R^2!

Digression about measures of fit…

It is easy to fall into the trap of maximizing the R^2 and $\overline{R}^2$, but this loses sight of our real objective, an unbiased estimator of the class size effect.

- A high R^2 (or $\overline{R}^2$) means that the regressors explain the variation in Y.
- A high R^2 (or $\overline{R}^2$) does *not* mean that you have eliminated omitted variable bias.
- A high R^2 (or $\overline{R}^2$) does *not* mean that you have an unbiased estimator of a causal effect (β_1).
- A high R^2 (or $\overline{R}^2$) does *not* mean that the included variables are statistically significant – this must be determined using hypotheses tests.

Analysis of the Test Score Data Set (SW Section 7.6)

1. Identify the variable of interest:

   STR

2. Think of the omitted causal effects that could result in omitted variable bias

   Whether the students know English; outside learning opportunities; parental involvement; teacher quality (if teacher salary is correlated with district wealth) – there is a long list!

3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables. The control variables are effective if the conditional mean independence assumption plausibly holds (if *u* is uncorrelated with *STR* once the control variables are included). This results in a "base" or "benchmark" model.

   Many of the omitted causal variables are hard to measure, so we need to find control variables. These include PctEL (both a control variable and an omitted causal factor) and measures of district wealth.

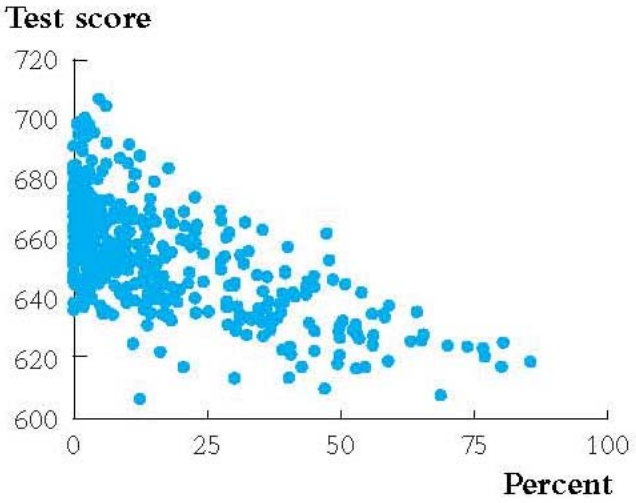
4. Also specify a range of plausible alternative models, which include additional candidate variables.

   It isn't clear which of the income-related variables will best control for the many omitted causal factors such as outside learning opportunities, so the alternative specifications include regressions with different income variables. The alternative specifications considered here are just a starting point, not the final word!

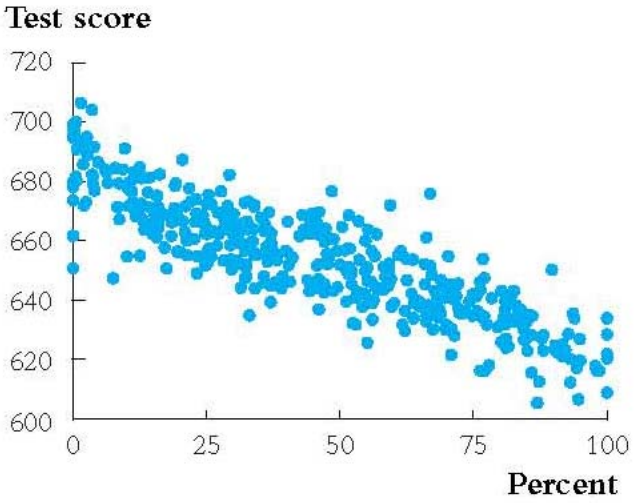
5. Estimate your base model and plausible alternative specifications ("sensitivity checks").

Test scores and California socioeconomic data…

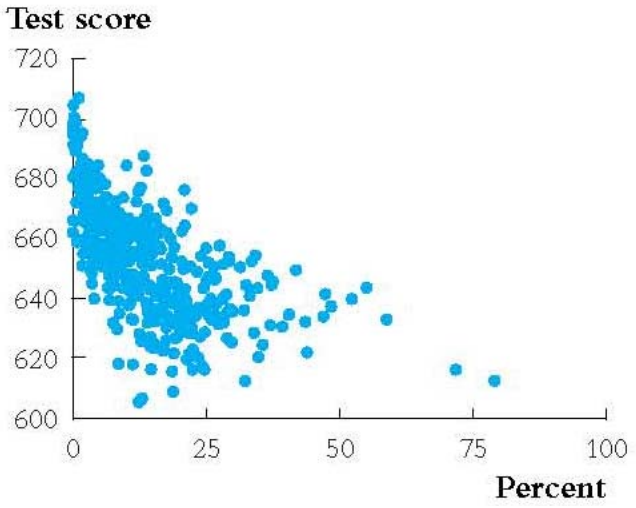
FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



Digression on presentation of regression results

- We have a number of regressions and we want to report them. It is awkward and difficult to read regressions written out in equation form, so instead it is conventional to report them in a table.
- A table of regression results should include:
  - estimated regression coefficients
  - standard errors
  - measures of fit
  - number of observations
  - relevant *F*-statistics, if any
  - Any other pertinent information.

Find this information in the following table:

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on Y of a change in X_1, holding other included variables constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- If you can't measure the omitted variable, you still might be able to control for its effect by including a control variable.
- There is no simple recipe for deciding which variables belong in a regression – you must exercise judgment.
- One approach is to specify a base model – relying on *a-priori* reasoning – then explore the sensitivity of the key estimate(s) in alternative specifications.

Nonlinear Regression Functions
(SW Chapter 8)

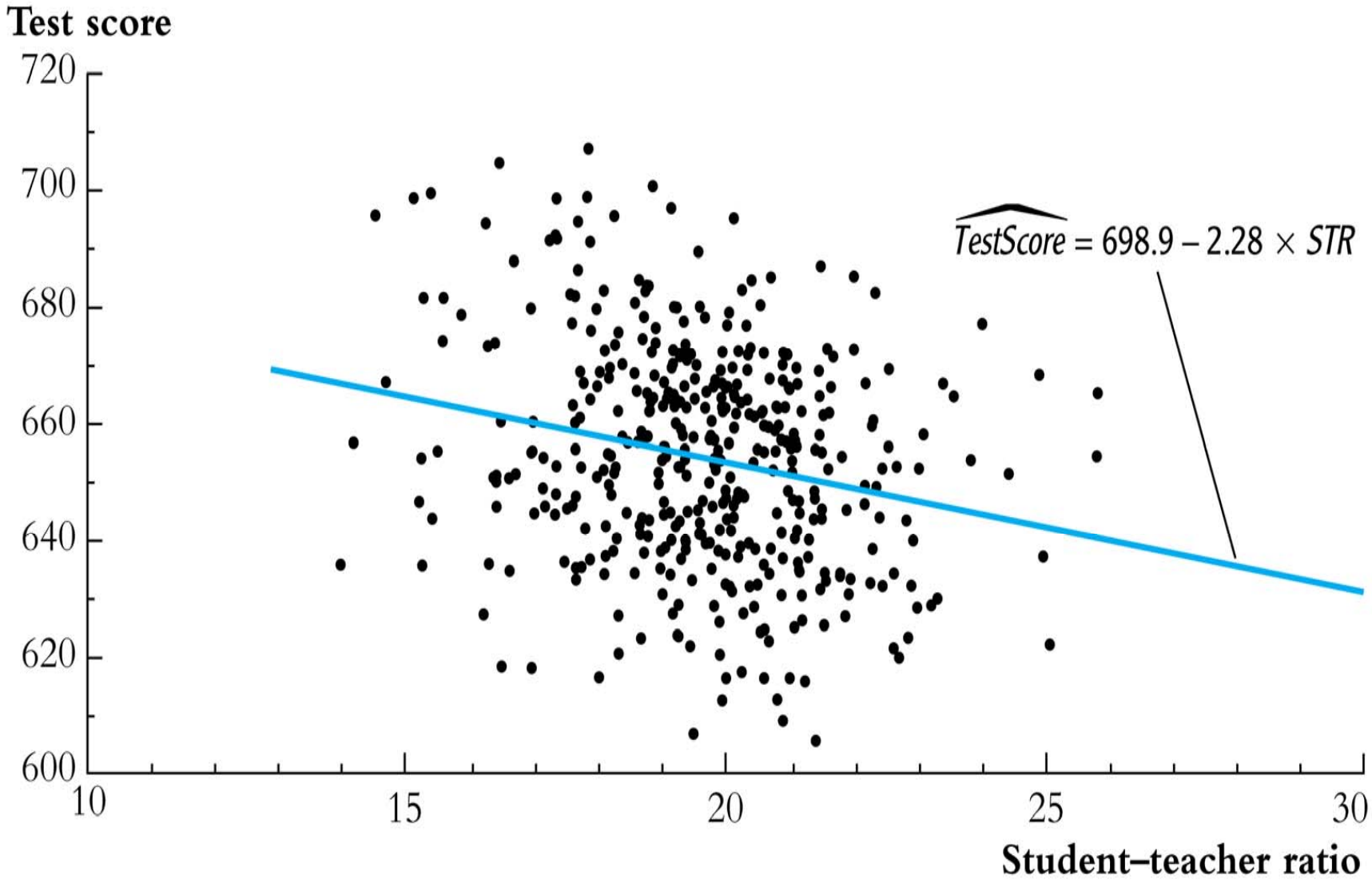
Outline

1. Nonlinear regression functions – general comments
2. Nonlinear functions of one variable
3. Nonlinear functions of two variables: interactions
4. Application to the California Test Score data set

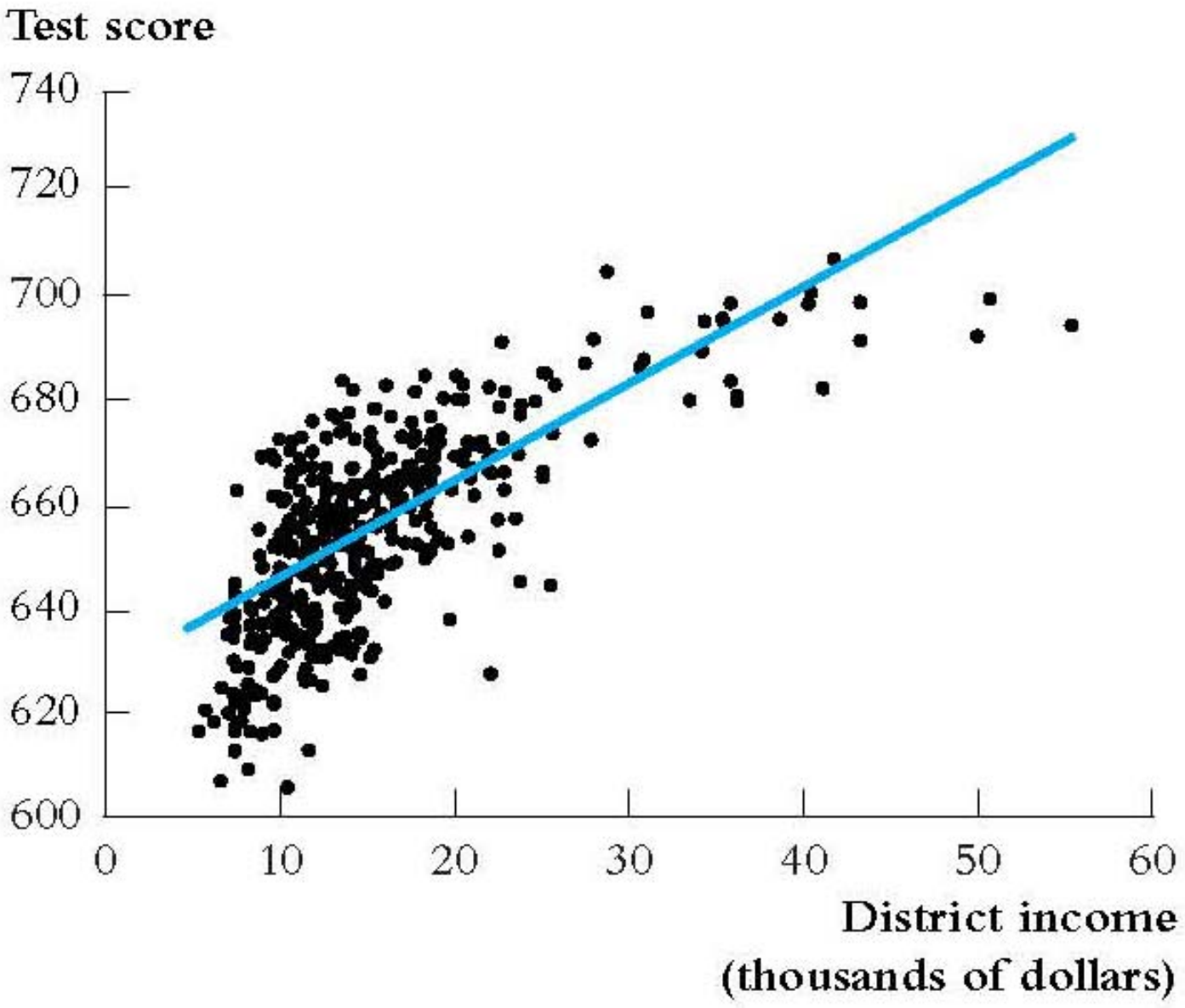
Nonlinear regression functions

- The regression functions so far have been linear in the X's
- But the linear approximation is not always a good one
- The multiple regression model can handle regression functions that are nonlinear in one or more X.

The *TestScore* – *STR* relation looks linear (maybe)…



But the *TestScore* – *Income* relation looks nonlinear...



Nonlinear Regression Population Regression Functions – General Ideas (SW Section 8.1)

If a relation between Y and X is **nonlinear**:

- The effect on Y of a change in X depends on the value of X – that is, the marginal effect of X is not constant
- A linear regression is mis-specified: the functional form is wrong
- The estimator of the effect on Y of X is biased: in general it isn't even right on average.
- The solution is to estimate a regression function that is nonlinear in X

The general nonlinear population regression function

$$Y_i = f(X_{1i}, X_{2i}, \ldots, X_{ki}) + u_i,\ i = 1, \ldots, n$$

Assumptions

1. $E(u_i | X_{1i}, X_{2i}, \ldots, X_{ki}) = 0$ (same); implies that f is the conditional expectation of Y given the X's.
2. $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d. (same).
3. Big outliers are rare (same idea; the precise mathematical condition depends on the specific f).
4. No perfect multicollinearity (same idea; the precise statement depends on the specific f).

The change in Y associated with a change in X_1, holding $X_2, \ldots, X_k$ constant is:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$$

The Expected Effect on Y of a Change in X_1 in the Nonlinear Regression Model (8.3)

KEY CONCEPT 8.1

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \quad (8.4)$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \quad (8.5)$$

Nonlinear Functions of a Single Independent Variable (SW Section 8.2)

We'll look at two complementary approaches:

1. Polynomials in X

   The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial

2. Logarithmic transformations
   - Y and/or X is transformed by taking its logarithm
   - this gives a "percentages" interpretation that makes sense in many applications

1. Polynomials in X

Approximate the population regression function by a polynomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- This is just the linear multiple regression model – except that the regressors are powers of X!
- Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS
- The coefficients are difficult to interpret, but the regression function itself is interpretable

Example: the *TestScore – Income* relation

$Income_i$ = average district income in the i^{th} district
(thousands of dollars per capita)

Quadratic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$$

Cubic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$$

Estimation of the quadratic specification in STATA

```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  2,   417) =  428.52
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5562
                                                      Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

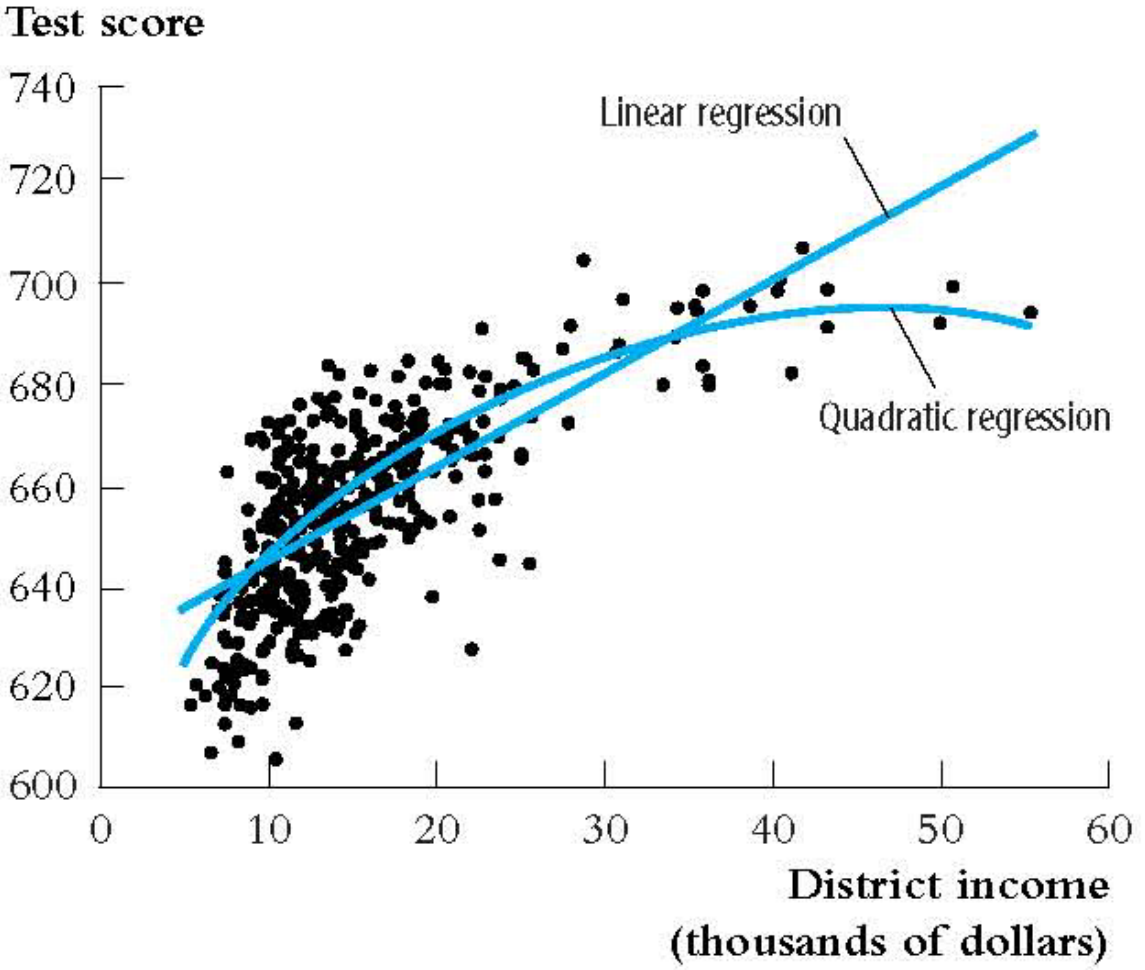
Test the null hypothesis of linearity against the alternative that the regression function is a quadratic….

Interpreting the estimated regression function:

(a) Plot the predicted values

$$\widehat{TestScore} = 607.3 + 3.85 Income_i - 0.0423 (Income_i)^2$$

(2.9) (0.27) (0.0048)



***Interpreting the estimated regression function, ctd*:**

(b) Compute "effects" for different values of X

$$\widehat{TestScore} = 607.3 + 3.85 Income_i - 0.0423 (Income_i)^2$$
$$\qquad\qquad\quad (2.9)\ \ (0.27)\qquad\qquad (0.0048)$$

Predicted change in *TestScore* for a change in income from \$5,000 per capita to \$6,000 per capita:

$$\Delta\widehat{TestScore} = 607.3 + 3.85\times 6 - 0.0423\times 6^2$$
$$\qquad\qquad\qquad\qquad - (607.3 + 3.85\times 5 - 0.0423\times 5^2)$$
$$\qquad\qquad = 3.4$$

$$\widehat{TestScore} = 607.3 + 3.85Income_i - 0.0423(Income_i)^2$$

Predicted "effects" for different values of X:

Change in *Income* ($1000 per capita)	$\Delta\widehat{TestScore}$
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

The "effect" of a change in income is greater at low than high income levels (perhaps, a declining marginal benefit of an increase in school budgets?)

Caution! What is the effect of a change from 65 to 66?

Don't extrapolate outside the range of the data!

Estimation of a cubic specification in STATA

```
gen avginc3 = avginc*avginc2;                    Create the cubic regressor
reg testscr avginc avginc2 avginc3, r;

Regression with robust standard errors                Number of obs =     420
                                                      F(  3,   416) =  270.18
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5584
                                                      Root MSE      =  12.707

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   5.018677   .7073505     7.10   0.000     3.628251    6.409104
     avginc2 |  -.0958052   .0289537    -3.31   0.001    -.1527191   -.0388913
     avginc3 |   .0006855   .0003471     1.98   0.049     3.27e-06    .0013677
       _cons |    600.079   5.102062   117.61   0.000     590.0499     610.108
------------------------------------------------------------------------------
```

Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

H_0: population coefficients on $Income^2$ and $Income^3 = 0$

H_1: at least one of these coefficients is nonzero.

```
test avginc2 avginc3;  Execute the test command after running the regression

 ( 1)  avginc2 = 0.0
 ( 2)  avginc3 = 0.0

       F(  2,   416) =   37.69
       Prob > F =    0.0000
```

The hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

Summary: polynomial regression functions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- Estimation: by OLS after defining new regressors
- Coefficients have complicated interpretations
- To interpret the estimated regression function:
  - plot predicted values as a function of x
  - compute predicted $\Delta Y/\Delta X$ at different values of x
- Hypotheses concerning degree r can be tested by t- and F-tests on the appropriate (blocks of) variable(s).
- Choice of degree r
  - plot the data; t- and F-tests, check sensitivity of estimated effects; judgment.
  - *Or use model selection criteria* (*later*)

2. Logarithmic functions of *Y* and/or *X*

- $\ln(X)$ = the natural logarithm of X
- Logarithmic transforms permit modeling relations in "percentage" terms (like elasticities), rather than linearly.

Here's why: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x}$

(calculus: $\frac{d\ln(x)}{dx} = \frac{1}{x}$)

Numerically:

$\ln(1.01) = .00995 \cong .01;$

$\ln(1.10) = .0953 \cong .10$ (sort of)

The three log regression specifications:

Case	Population regression function
I. linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
II. log-linear	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- The interpretation of the slope coefficient differs in each case.
- The interpretation is found by applying the general "before and after" rule: "figure out the change in Y for a given change in X."
- Each case has a natural interpretation (for small changes in X)

I. Linear-log population regression function

Compute Y "before" and "after" changing X:

$$Y = \beta_0 + \beta_1 \ln(X) \qquad \text{("before")}$$

Now change X: $\quad Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X) \qquad$ ("after")

Subtract ("after") – ("before"): $\quad \Delta Y = \beta_1 [\ln(X + \Delta X) - \ln(X)]$

now $$\ln(X + \Delta X) - \ln(X) \cong \frac{\Delta X}{X},$$

so $$\Delta Y \cong \beta_1 \frac{\Delta X}{X}$$

or $$\beta_1 \cong \frac{\Delta Y}{\Delta X / X} \quad \text{(small } \Delta X\text{)}$$

Linear-log case, continued

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***a 1% increase in X (multiplying X by* 1.01*) is associated with a* .01β_1 *change in Y.***

(1% increase in $X \Rightarrow$ .01 increase in $\ln(X)$

$\Rightarrow .01\beta_1$ increase in Y)

Example: TestScore vs. ln(Income)

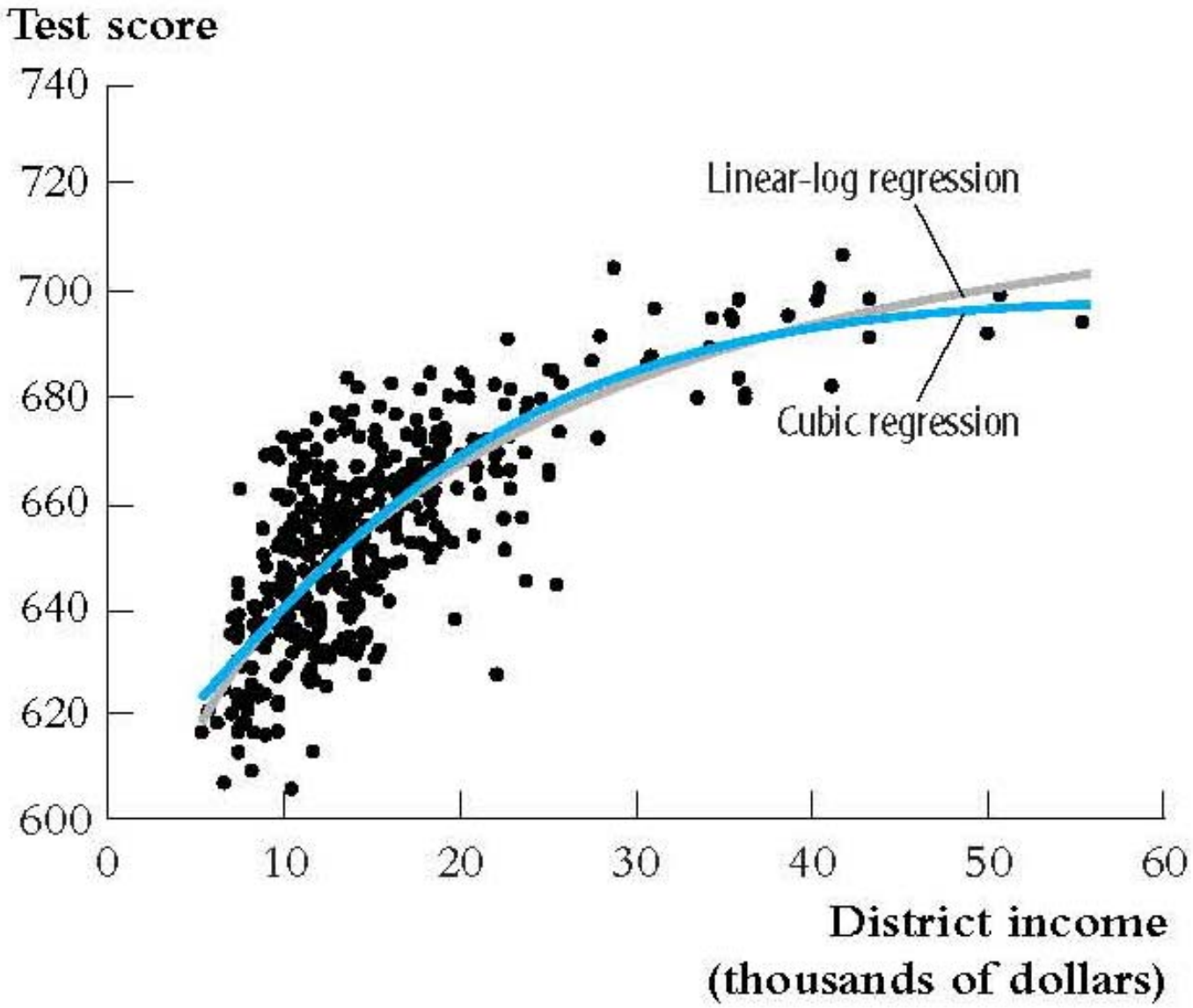
- First defining the new regressor, ln(*Income*)
- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

$$\widehat{TestScore} = 557.8 + 36.42 \times \ln(Income_i)$$
$$(3.8) \quad (1.40)$$

so a 1% increase in *Income* is associated with an increase in *TestScore* of 0.36 points on the test.

- Standard errors, confidence intervals, R^2 – all the usual tools of regression apply here.
- How does this compare to the cubic model?

The linear-log and cubic regression functions



II. Log-linear population regression function

$$\ln(Y) = \beta_0 + \beta_1 X \qquad \text{(b)}$$

Now change X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$ (a)

Subtract (a) – (b): $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X} \text{ (small } \Delta X)$$

Log-linear case, continued

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

for small ΔX, $\quad \beta_1 \cong \dfrac{\Delta Y / Y}{\Delta X}$

- Now $100 \times \dfrac{\Delta Y}{Y}$ = percentage change in Y, so ***a change in X by one unit*** **($\Delta X = 1$)** ***is associated with a*** **$100\beta_1$%** ***change in Y.***
- 1 unit increase in $X \Rightarrow \beta_1$ increase in $\ln(Y)$

  $\Rightarrow 100\beta_1\%$ increase in Y
- *Note*: What are the units of u_i and the SER?
  - fractional (proportional) deviations
  - for example, $SER = .2$ means…

III. Log-log population regression function

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \qquad \text{(b)}$$

Now change X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$ (a)

Subtract: $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 [\ln(X + \Delta X) - \ln(X)]$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X} \text{ (small } \Delta X)$$

Log-log case, continued

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta Y}{Y}$ = percentage change in Y, and $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***a 1% change in X is associated with a β_1% change in Y***.

- ***In the log-log specification, β_1 has the interpretation of an elasticity***.

Example: ln(TestScore) vs. ln(Income)

- First defining a new dependent variable, ln(*TestScore*), ***and*** the new regressor, ln(*Income*)
- The model is now a linear regression of ln(*TestScore*) against ln(*Income*), which can be estimated by OLS:

$$\widehat{\ln(TestScore)} = 6.336 + 0.0554 \times \ln(Income_i)$$
$$(0.006)\ \ (0.0021)$$

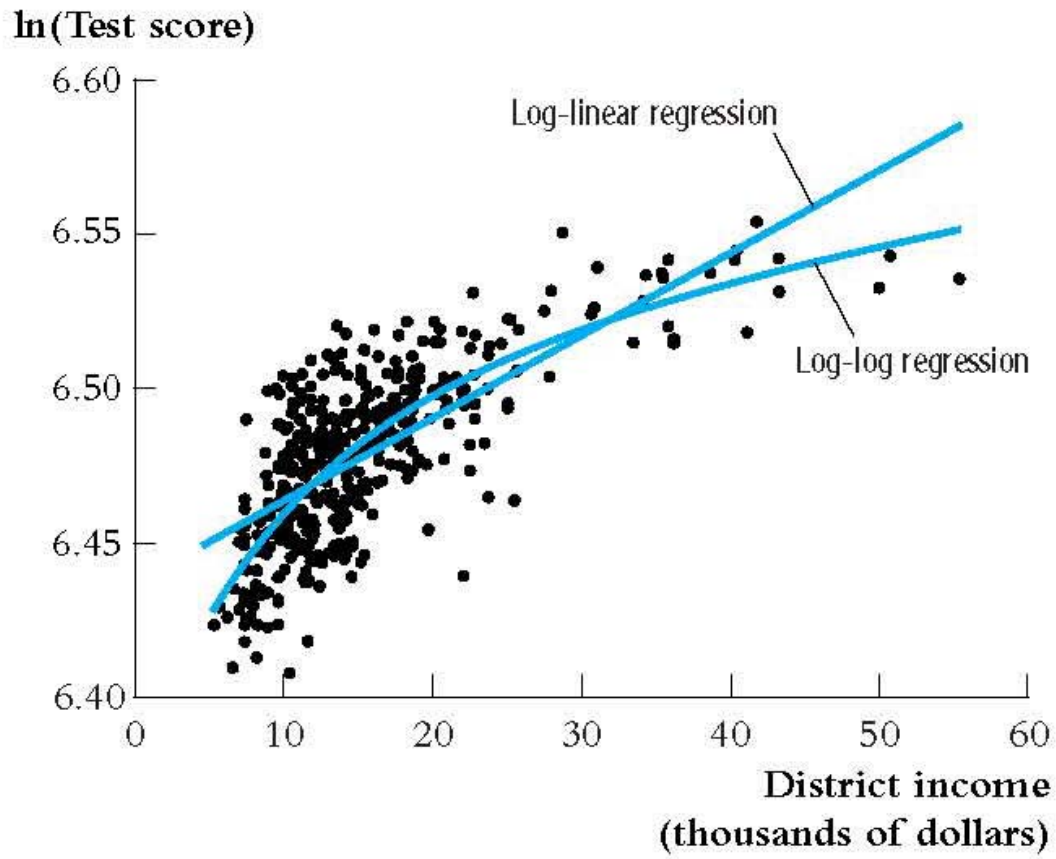
An 1% increase in *Income* is associated with an increase of .0554% in *TestScore* (*Income* up by a factor of 1.01, *TestScore* up by a factor of 1.000554)

Example: ln(TestScore) vs. ln(Income), ctd.

$$\widehat{\ln(TestScore)} = 6.336 + 0.0554 \times \ln(Income_i)$$
$$(0.006)\ \ (0.0021)$$

- For example, suppose income increases from $10,000 to $11,000, or by 10%. Then *TestScore* increases by approximately .0554×10% = .554%. If *TestScore* = 650, this corresponds to an increase of .00554×650 = 3.6 points.
- How does this compare to the log-linear model?

The log-linear and log-log specifications:



- *Note vertical axis*
- *Neither seems to fit as well as the cubic or linear-log, at least based on visual inspection (formal comparison is difficult because the dependent variables differ)*

Summary: Logarithmic transformations

- Three cases, differing in whether Y and/or X is transformed by taking logarithms.
- The regression is linear in the new variable(s) ln(Y) and/or ln(X), and the coefficients can be estimated by OLS.
- Hypothesis tests and confidence intervals are now implemented and interpreted "as usual."
- The interpretation of β_1 differs from case to case.

The choice of specification (functional form) should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values

Other nonlinear functions (and nonlinear least squares) (SW Appendix 8.1)

The foregoing regression functions have limitations…

- Polynomial: test score can decrease with income
- Linear-log: test score increases with income, but without bound
- Here is a nonlinear function in which *Y* always increases with *X and* there is a maximum (asymptote) value of *Y*:

$$Y = \beta_0 - \alpha e^{-\beta_1 X}$$

β_0, β_1, and α are unknown parameters. This is called a negative exponential growth curve. The asymptote as $X \to \infty$ is β_0.

Negative exponential growth

We want to estimate the parameters of,

$$Y_i = \beta_0 - \alpha e^{-\beta_1 X_i} + u_i$$

or

$$Y_i = \beta_0 \left[1 - e^{-\beta_1 (X_i - \beta_2)}\right] + u_i \qquad (*)$$

where $\alpha = \beta_0 e^{\beta_2}$ (why would you do this???)

Compare model (*) to linear-log or cubic models:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_2 X_i^3 + u_i$$

The linear-log and polynomial models are *linear in the parameters* β_0 and β_1 – but the model (*) is not.

Nonlinear Least Squares

- Models that are linear in the parameters can be estimated by OLS.
- Models that are nonlinear in one or more parameters can be estimated by nonlinear least squares (NLS) (but not by OLS)
- The NLS problem for the proposed specification:

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} \left\{ Y_i - \beta_0 \left[1 - e^{-\beta_1 (X_i - \beta_2)} \right] \right\}^2$$

  This is a nonlinear minimization problem (a "hill-climbing" problem). How could you solve this?
  - Guess and check
  - There are better ways…
  - Implementation in STATA…

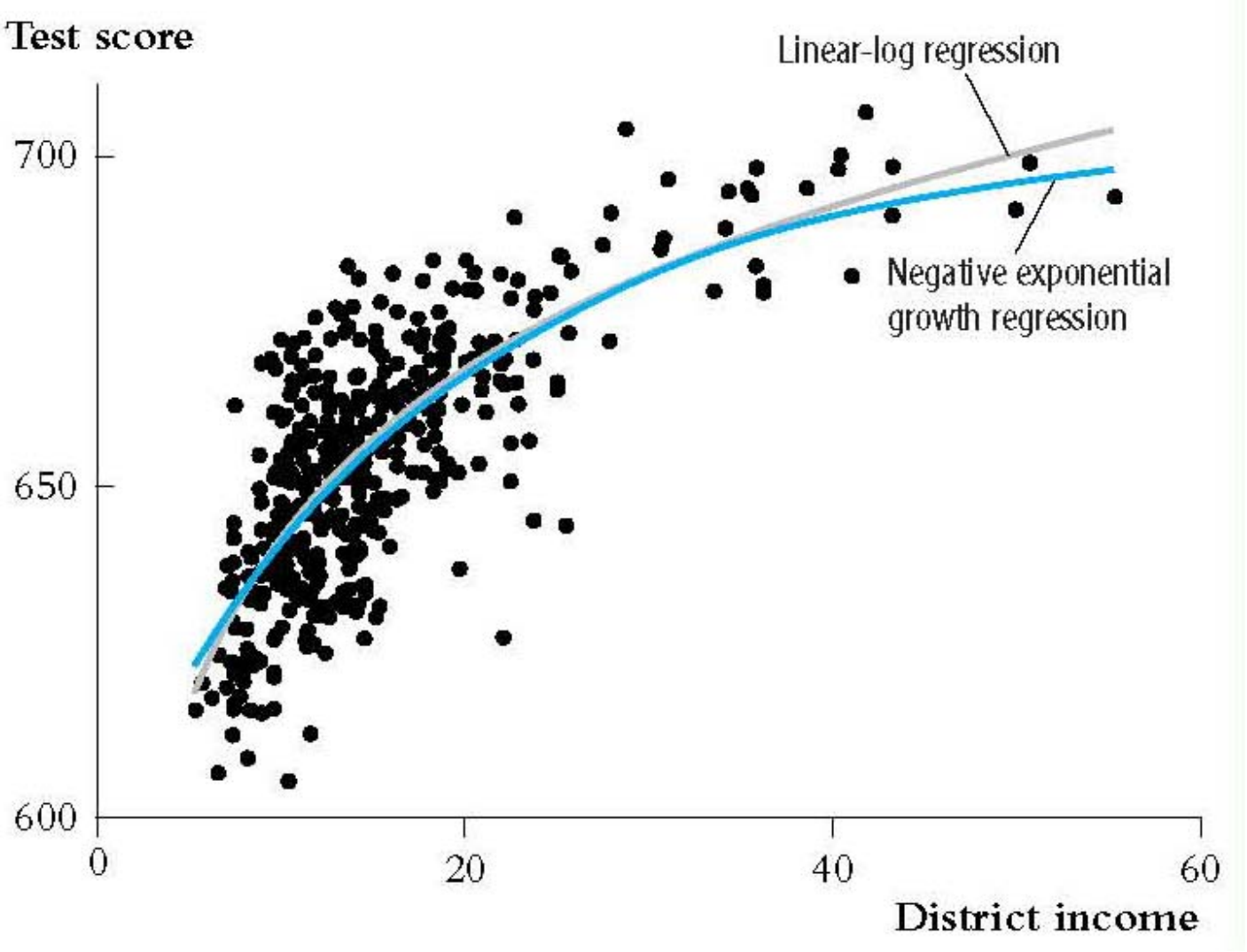
```
. nl (testscr = {b0=720}*(1 - exp(-1*{b1}*(avginc-{b2})))), r

(obs = 420)
Iteration 0:  residual SS =  1.80e+08                    .
Iteration 1:  residual SS =  3.84e+07                    .
Iteration 2:  residual SS =   4637400                    .
Iteration 3:  residual SS =  300290.9      STATA is "climbing the hill"
Iteration 4:  residual SS =  70672.13      (actually, minimizing the SSR)
Iteration 5:  residual SS =  66990.31                    .
Iteration 6:  residual SS =   66988.4                    .
Iteration 7:  residual SS =   66988.4                    .
Iteration 8:  residual SS =   66988.4

Nonlinear regression with robust standard errors      Number of obs =       420
                                                      F(  3,   417) = 687015.55
                                                      Prob > F      =    0.0000
                                                      R-squared     =    0.9996
                                                      Root MSE      =  12.67453
                                                      Res. dev.     =  3322.157
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          b0 |   703.2222   4.438003   158.45   0.000     694.4986    711.9459
          b1 |   .0552339   .0068214     8.10   0.000     .0418253    .0686425
          b2 |  -34.00364    4.47778    -7.59   0.000    -42.80547    -25.2018
------------------------------------------------------------------------------
 (SEs, P values, CIs, and correlations are asymptotic approximations)
```

Negative exponential growth; *RMSE* = 12.675

Linear-log; *RMSE* = 12.618 (oh well…)



Interactions Between Independent Variables (SW Section 8.3)

- Perhaps a class size reduction is more effective in some circumstances than in others…
- Perhaps smaller classes help more if there are many English learners, who need individual attention
- That is, $\frac{\Delta TestScore}{\Delta STR}$ might depend on *PctEL*
- More generally, $\frac{\Delta Y}{\Delta X_1}$ might depend on X_2
- How to model such "interactions" between X_1 and X_2?
- We first consider binary X's, then continuous X's

(a) Interactions between two binary variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

- D_{1i}, D_{2i} are binary
- β_1 is the effect of changing D_1=0 to D_1=1. In this specification, *this effect doesn't depend on the value of* D_2.
- To allow the effect of changing D_1 to depend on D_2, include the "interaction term" $D_{1i} \times D_{2i}$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

Interpreting the coefficients

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

General rule: compare the various cases

$$E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_0 + \beta_2 d_2 \qquad \text{(b)}$$

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \qquad \text{(a)}$$

subtract (a) – (b):

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) - E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_1 + \beta_3 d_2$$

- The effect of D_1 depends on d_2 (what we wanted)
- β_3 = increment to the effect of D_1, when $D_2 = 1$

Example: *TestScore, STR, English learners*

Let

$$HiSTR = \begin{cases} 1 \text{ if } STR \geq 20 \\ 0 \text{ if } STR < 20 \end{cases} \quad \text{and} \quad HiEL = \begin{cases} 1 \text{ if } PctEL \geq 10 \\ 0 \text{ if } PctEL < 10 \end{cases}$$

$$\widehat{TestScore} = 664.1 - 18.2HiEL - 1.9HiSTR - 3.5(HiSTR \times HiEL)$$
$$(1.4) \quad (2.3) \qquad (1.9) \qquad (3.1)$$

- "Effect" of *HiSTR* when *HiEL* = 0 is –1.9
- "Effect" of *HiSTR* when *HiEL* = 1 is –1.9 – 3.5 = –5.4
- Class size reduction is estimated to have a bigger effect when the percent of English learners is large
- This interaction isn't statistically significant: $t = 3.5/3.1$

Example: *TestScore, STR, English learners, ctd.*

Let

$$HiSTR = \begin{cases} 1 \text{ if } STR \geq 20 \\ 0 \text{ if } STR < 20 \end{cases} \quad \text{and} \quad HiEL = \begin{cases} 1 \text{ if } PctEL \geq 10 \\ 0 \text{ if } PctEL < 10 \end{cases}$$

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(2.3)}{18.2}HiEL - \underset{(1.9)}{1.9}HiSTR - \underset{(3.1)}{3.5}(HiSTR \times HiEL)$$

- Can you relate these coefficients to the following table of group ("cell") means?

	Low STR	***High STR***
Low EL	664.1	662.2
High EL	645.9	640.5

(b) Interactions between continuous and binary variables

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- D_i is binary, X is continuous
- As specified above, the effect on Y of X (holding constant D) = β_2, which does not depend on D
- To allow the effect of X to depend on D, include the "interaction term" $D_i \times X_i$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

Binary-continuous interactions: the two regression lines

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

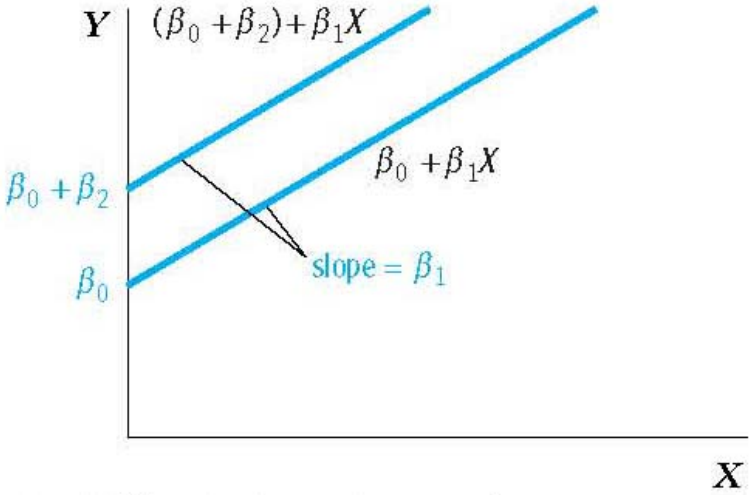
Observations with $D_i = 0$ (the "$D = 0$" group):

$$Y_i = \beta_0 + \beta_2 X_i + u_i \quad \textbf{\textit{The D=0 regression line}}$$

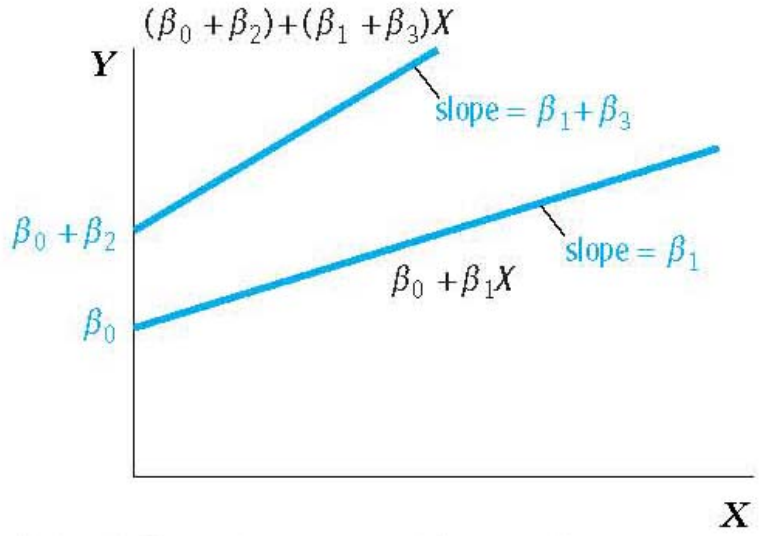
Observations with $D_i = 1$ (the "$D = 1$" group):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + u_i \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + u_i \end{aligned} \quad \textbf{\textit{The D=1 regression line}}$$

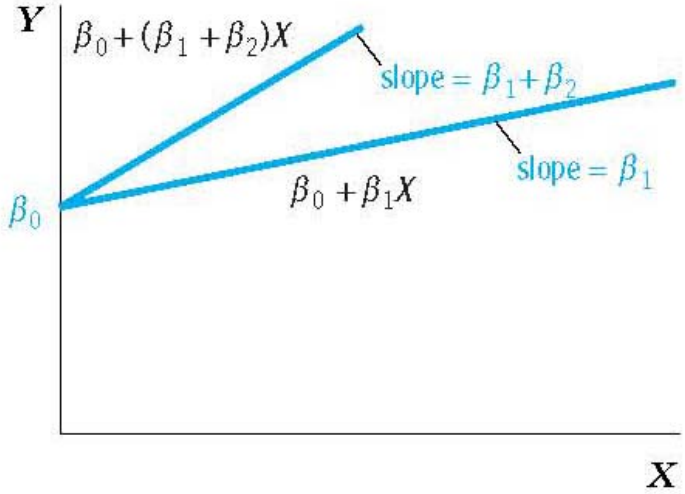
Binary-continuous interactions, ctd.



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interpreting the coefficients

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$$

General rule: compare the various cases

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) \qquad \text{(b)}$$

Now change X:

$$Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 [D \times (X + \Delta X)] \text{(a)}$$

subtract (a) – (b):

$$\Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \text{ or } \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$$

- The effect of X depends on D (what we wanted)
- β_3 = increment to the effect of X, when $D = 1$

Example: *TestScore*, *STR*, *HiEL* (=1 if *PctEL* ≥ 10)

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL)$$
$$(11.9)\ (0.59) \qquad (19.5) \qquad (0.97)$$

- When $HiEL = 0$:

$$\widehat{TestScore} = 682.2 - 0.97STR$$

- When $HiEL = 1$,

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6 - 1.28STR$$
$$= 687.8 - 2.25STR$$

- Two regression lines: one for each *HiSTR* group.
- Class size reduction is estimated to have a larger effect when the percent of English learners is large.

***Example, ctd**: Testing hypotheses*

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6HiEL - 1.28(STR \times HiEL)$$
$$(11.9)\ \ (0.59) \qquad (19.5) \qquad (0.97)$$

- The two regression lines have the same **slope** ⇔ the coefficient on $STR \times HiEL$ is zero: $t = -1.28/0.97 = -1.32$
- The two regression lines have the same **intercept** ⇔ the coefficient on $HiEL$ is zero: $t = -5.6/19.5 = 0.29$
- The two regression **lines** are the same ⇔ population coefficient on $HiEL = 0$ ***and*** population coefficient on $STR \times HiEL = 0$: $F = 89.94$ (p-value < .001) **!!**
- We reject the joint hypothesis but neither individual hypothesis (*how can this be?*)

(c) Interactions between two continuous variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- X_1, X_2 are continuous
- As specified, the effect of X_1 doesn't depend on X_2
- As specified, the effect of X_2 doesn't depend on X_1
- To allow the effect of X_1 to depend on X_2, include the "interaction term" $X_{1i} \times X_{2i}$ as a regressor:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

Interpreting the coefficients:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$$

General rule: compare the various cases

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) \qquad \text{(b)}$$

Now change X_1:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \beta_3 [(X_1 + \Delta X_1) \times X_2] \qquad \text{(a)}$$

subtract (a) – (b):

$$\Delta Y = \beta_1 \Delta X_1 + \beta_3 X_2 \Delta X_1 \text{ or } \frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- The effect of X_1 depends on X_2 (what we wanted)
- β_3 = increment to the effect of X_1 from a unit change in X_2

Example: *TestScore*, *STR*, *PctEL*

$$\widehat{TestScore} = 686.3 - 1.12STR - 0.67PctEL + .0012(STR \times PctEL),$$
$$(11.8) \quad (0.59) \qquad (0.37) \qquad (0.019)$$

The estimated effect of class size reduction is nonlinear because the size of the effect itself depends on *PctEL*:

$$\frac{\Delta TestScore}{\Delta STR} = -1.12 + .0012PctEL$$

PctEL	$\frac{\Delta TestScore}{\Delta STR}$
0	–1.12
20%	–1.12+.0012×20 = –1.10

***Example, ctd**: hypothesis tests*

$$\widehat{TestScore} = 686.3 - 1.12STR - 0.67PctEL + .0012(STR \times PctEL),$$
$$(11.8) \quad (0.59) \qquad (0.37) \qquad (0.019)$$

- Does population coefficient on $STR \times PctEL = 0$?

  $t = .0012/.019 = .06 \Rightarrow$ can't reject null at 5% level
- Does population coefficient on $STR = 0$?

  $t = -1.12/0.59 = -1.90 \Rightarrow$ can't reject null at 5% level
- Do the coefficients on ***both*** STR ***and*** $STR \times PctEL = 0$?

  $F = 3.89$ (p-value = .021) $\Rightarrow$ reject null at 5% level(!!)

  (Why? high but imperfect multicollinearity)

Application: Nonlinear Effects on Test Scores of the Student-Teacher Ratio (SW Section 8.4)

Nonlinear specifications let us examine more nuanced questions about the Test score – *STR* relation, such as:

1. Are there nonlinear effects of class size reduction on test scores? (Does a reduction from 35 to 30 have same effect as a reduction from 20 to 15?)
2. Are there nonlinear interactions between *PctEL* and *STR*? (Are small classes more effective when there are many English learners?)

Strategy for Question #1 (different effects for different *STR*?)

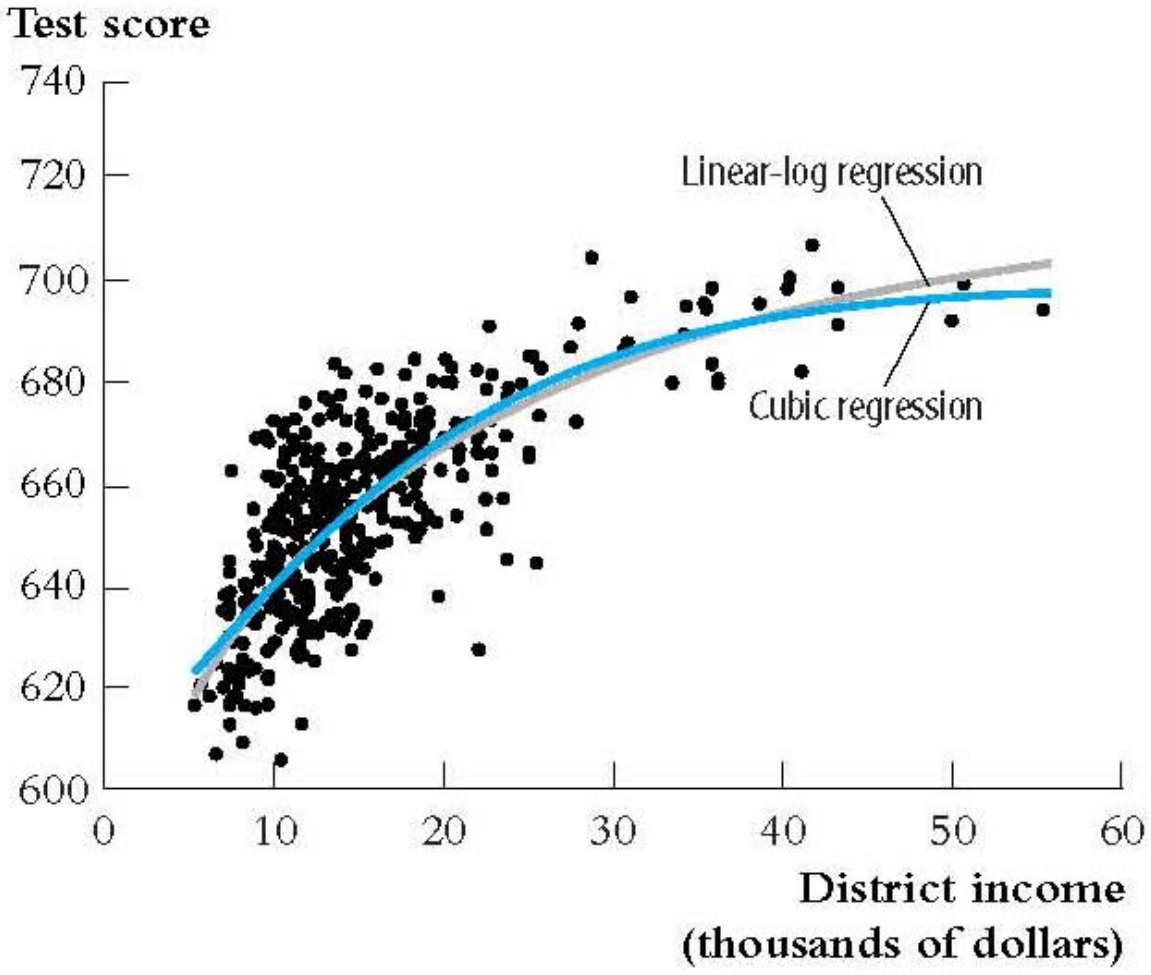
- Estimate linear and nonlinear functions of *STR*, holding constant relevant demographic variables
  - *PctEL*
  - *Income* (remember the nonlinear *TestScore*-*Income* relation!)
  - *LunchPCT* (fraction on free/subsidized lunch)
- See whether adding the nonlinear terms makes an "economically important" quantitative difference ("economic" or "real-world" importance is different than statistically significant)
- Test for whether the nonlinear terms are significant

Strategy for Question #2 (interactions between *PctEL* and *STR***?)**

- Estimate linear and nonlinear functions of *STR*, interacted with *PctEL*.
- If the specification is nonlinear (with *STR*, STR^2, STR^3), then you need to add interactions with all the terms so that the entire functional form can be different, depending on the level of *PctEL*.
- We will use a binary-continuous interaction specification by adding $HiEL \times STR$, $HiEL \times STR^2$, and $HiEL \times STR^3$.

What is a good "base" specification?

The *TestScore – Income* relation:



The logarithmic specification is better behaved near the extremes of the sample, especially for large values of income.

TABLE 8.3 Nonlinear Regression Models of Test Scores

Dependent variable: average test score in district; 420 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (STR)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English learners ≥ 10%? (Binary, $HiEL$)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.2** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

Tests of joint hypotheses:

F-Statistics and p-Values on Joint Hypotheses

(a) All *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (< 0.001)	4.96 (< 0.001)	5.91 (0.001)
(b) $STR^2, STR^3 = 0$					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR, HiEL \times STR^2, HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

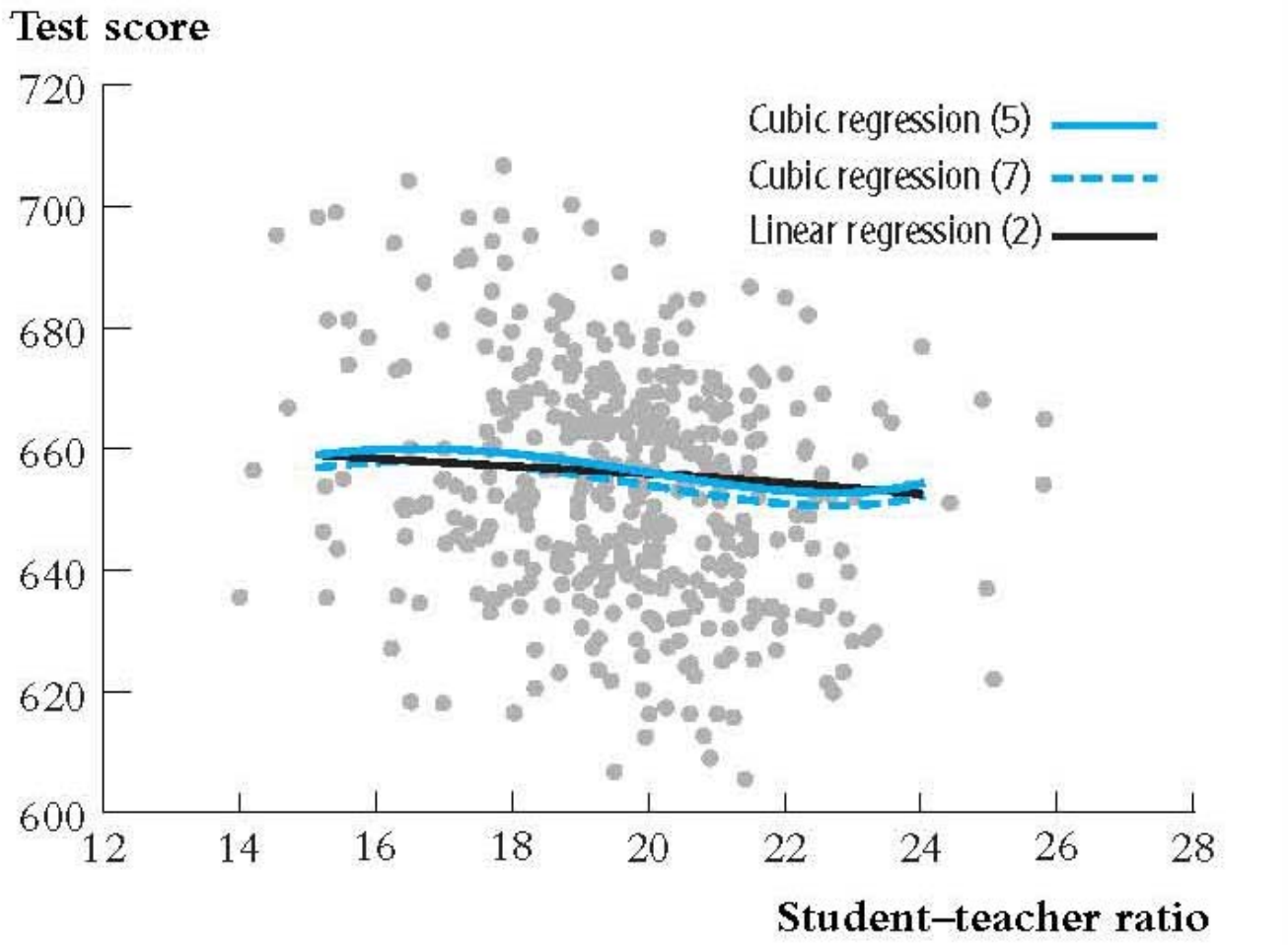
These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

What can you conclude about question #1?

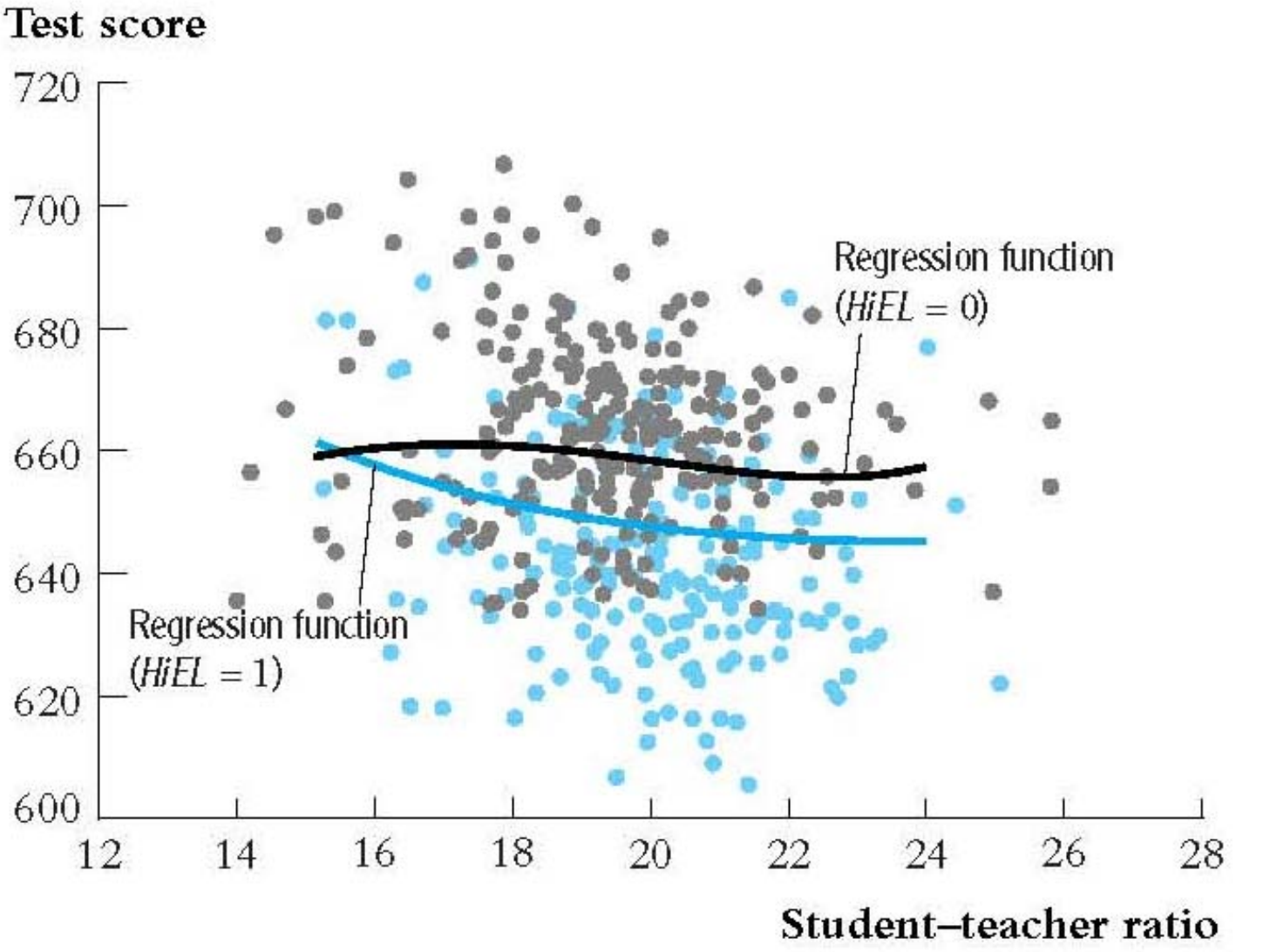
About question #2?

Interpreting the regression functions via plots:

First, compare the linear and nonlinear specifications:



Next, compare the regressions with interactions:



Summary: Nonlinear Regression Functions

- Using functions of the independent variables such as ln(X) or $X_1 \times X_2$, allows recasting a large family of nonlinear regression functions as multiple regression.
- Estimation and inference proceed in the same way as in the linear multiple regression model.
- Interpretation of the coefficients is model-specific, but the general rule is to compute effects by comparing different cases (different value of the original X's)
- Many nonlinear specifications are possible, so you must use judgment:
  - What nonlinear effect you want to analyze?
  - What makes sense in your application?