Week 1: Overview and Review of Probability and Statistics

Slava Mikhaylov

PUBLG088 Advanced Quantitative Methods

Week 1 Outline

Statistics

Course outline and logistics

Review of probability and statistics

- Policy problem
- Review of statistical theory
  - The probability framework
  - Estimation
  - Hypothesis testing
  - Confidence intervals
- Linear regression recap
  - The linear regression model
  - Multiple regression
  - Hypothesis testing

Statistics

"Its machine learning allows the computer to become smarter as it tries to answer questions and to learn as it gets them right or wrong." David Ferrucci (PI on Watson DeepQA at IBM Research).

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

W

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?" Hal Varian (Chief Economist at Google, 2009).

Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten post code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a satellite photo of Earth's surface by agricultural usage.

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Philosophy

- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working (simpler methods often perform as well as fancier ones!).
- This is an exciting research area, having important applications in science, industry and policy.
- Statistical learning is a fundamental ingredient in the training of a modern data scientist.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy – find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

The Netflix prize

- competition started in October 2006. Training data is ratings for 18,000 movies by 400,000 Netflix customers, each rating between 1 and 5.
- training data is very sparse – about 98% missing.
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data.
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins one million dollars.
- is this a supervised or unsupervised problem?

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

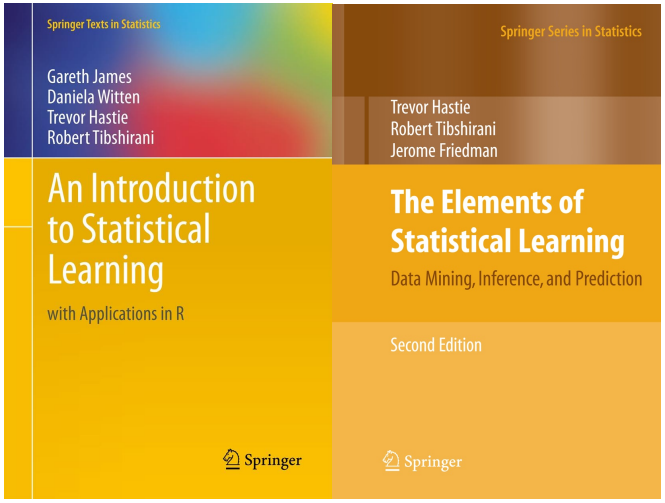
BellKor's Pragmatic Chaos wins, beating The Ensemble by a narrow margin.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap – both fields focus on supervised and unsupervised problems:
  - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".
- Machine learning has the upper hand in Marketing!

Course outline and logistics

Course Texts



- The course will cover most of the material in this Springer book (ISLR) published in 2013. Each chapter ends with an R lab, in which examples are developed. An electronic version of this book is available for free from the authors' websites.
- Some of the figures and lecture material are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.
- This Springer book (ESL) is more mathematically advanced than ISLR; the second edition was published in 2009. It covers a broader range of topics. A free electronic version of the book is available from the authors' websites.

Review of probability and statistics

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:

- 5th 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!)

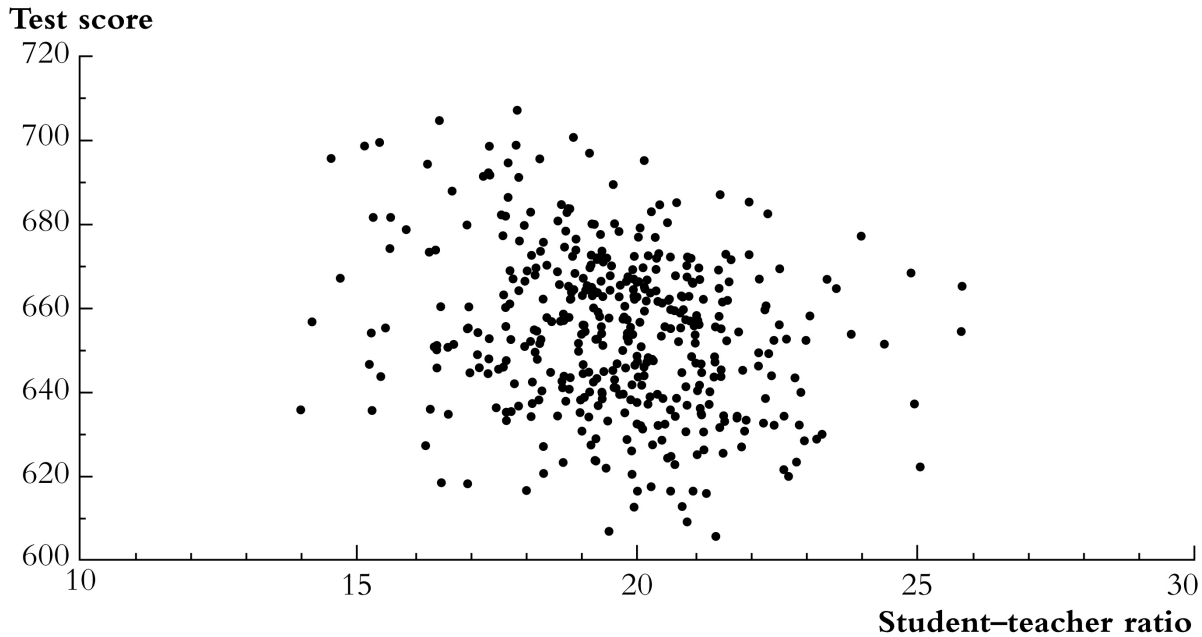
TABLE 4.1 Summary of the Distribution of Student–Teacher Ratios and Fifth-Grade Test Scores for 420 K–8 Districts in California in 1998

			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" ($STR \geq 20$) class sizes:

Class size	**Average score** $\bar{Y}$	**Standard deviation** (sB_{YB})	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

Estimation

$$\bar{Y}_{small} - \bar{Y}_{large} = 657.4 - 650.0 = 7.4$$

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

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

Hypothesis testing

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

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

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:

Class size	**Average score** $\bar{Y}$	**Standard deviation** (sB_{YB})	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.

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.

The probability framework for statistical inference

- Population, random variable, and distribution
- Moments of a distribution (mean, variance, standard deviation, covariance, correlation)
- Conditional distributions and conditional means
- Distribution of a sample of data drawn randomly from a population

Population, random variable, and distribution

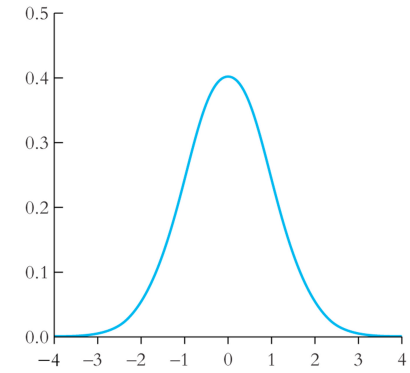
- Population
  - The group or collection of all possible entities of interest (school districts)
  - We will think of populations as infinitely large
- Random variable Y
  - Numerical summary of a random outcome (district average test score, district STR)

Population distribution of Y

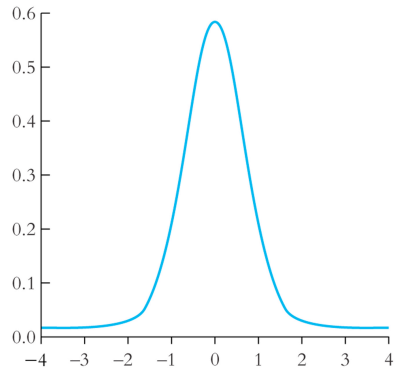
- The probabilities of different values of Y that occur in the population, e.g. $Pr[Y = 650]$ (when Y is discrete)
- or: The probabilities of sets of these values, e.g. $Pr[640 \leq Y \leq 660]$ (when Y is continuous).

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

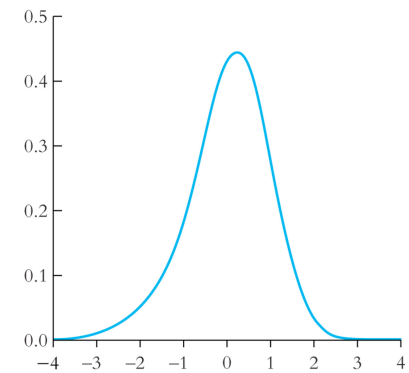
- mean = expected value (expectation) of Y= E(Y) = long-run average value of Y over repeated realizations of Y
- variance = E(Y - E[Y]) = measure of the squared spread of the distribution
- standard deviation = $\sqrt{variance}$
- skewness = measure of asymmetry of a distribution: skewness = 0 – distribution is symmetric; otherwise the distribution has long right or left tail.
- kurtosis=measure of mass in tails=measure of probability of large values: kurtosis = 3 – normal distribution; greater than 3 – heavy tails ("leptokurtotic")



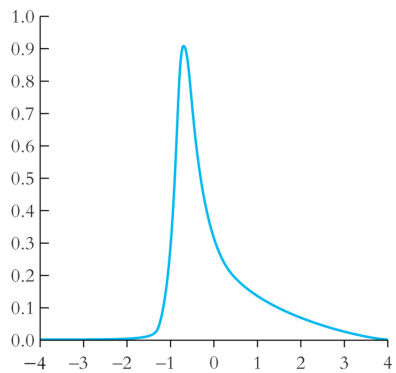
(a) Skewness = 0, kurtosis = 3



(b) Skewness = 0, kurtosis = 20



(c) Skewness = –0.1, kurtosis = 5



(d) Skewness = 0.6, kurtosis = 5

Two random variables: joint distributions and covariance

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

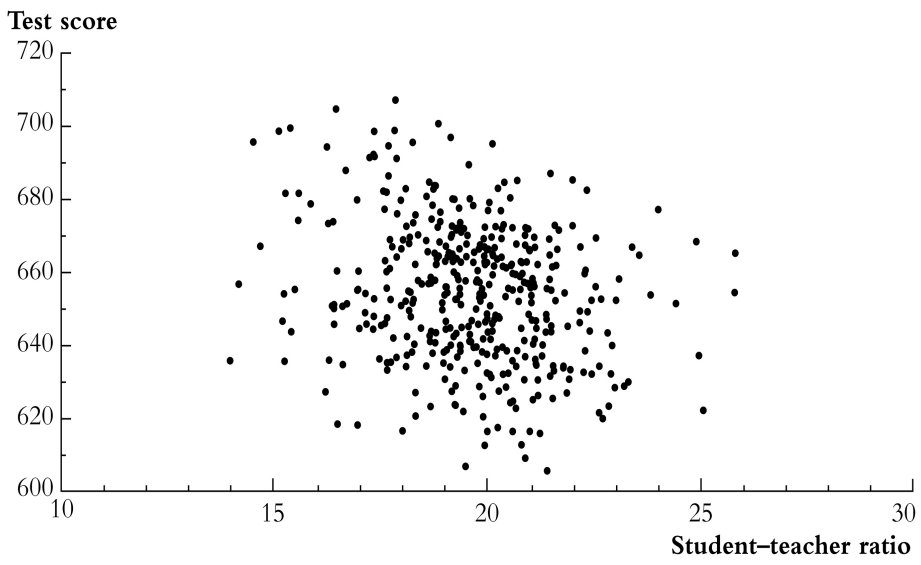
$$cov(X, Z) = E[(X - \mu_X)(Z - \mu_Z)]$$

- The covariance is a measure of the linear association between X and Z
- $cov(X, Z) > 0$ means a positive relation between X and Z
- If X and Z are independently distributed, then $cov(X, Z) = 0$ (but not vice versa!!)

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.

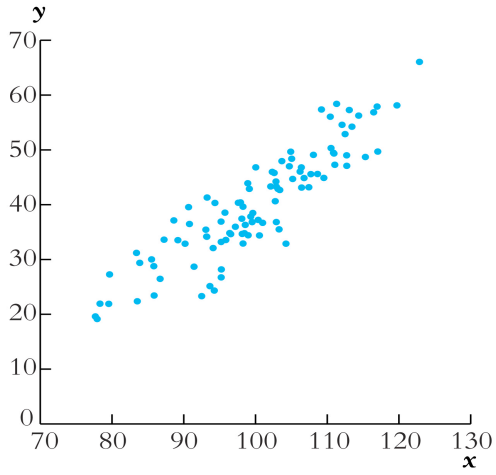


So is the correlation

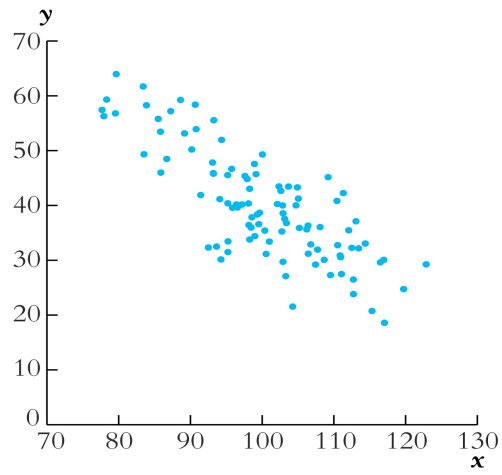
The correlation coefficient is defined in terms of the covariance:

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

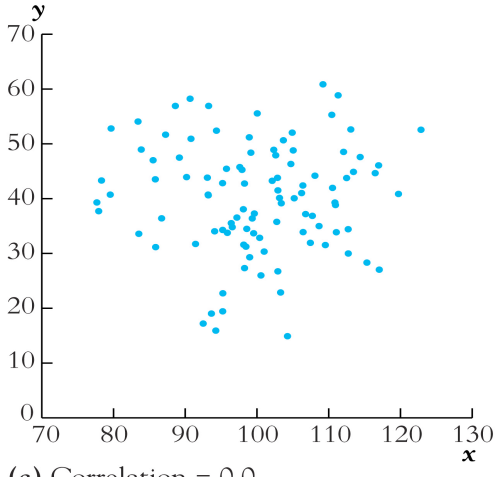
- $-1 \leq corr(X, Z) \leq 1$
- corr(X,Z) = 1 mean perfect positive linear association
- corr(X,Z) = -1 means perfect negative linear association
- corr(X,Z) = 0 means no linear association



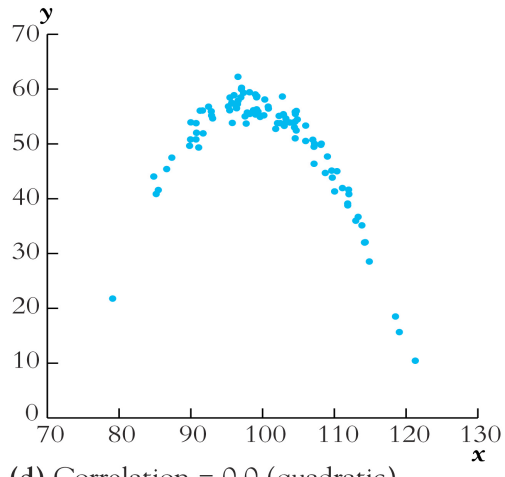
(a) Correlation = +0.9



(b) Correlation = –0.8



(c) Correlation = 0.0



(d) Correlation = 0.0 (quadratic)

Conditional distributions and conditional means

- Conditional distributions: The distribution of Y, given value(s) of some other random variable, X. E.g. 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)
- The difference in means is the difference between the means of two conditional distributions.

Distribution of a sample of data drawn randomly from a population

- 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.
- Because individuals no. 1 and no. 2 are selected at random, the value Y for the first has no information content for the second. Thus, Y_1 and Y_2 are independently distributed, and if they come from the same distribution then they are also identically distributed.
- That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d.).

Estimation

- $\bar{Y}$ is the natural estimator of the mean.
- $\bar{Y}$ is a random variable, and its properties are determined by the sampling distribution of $\bar{Y}$.
- 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 $\bar{Y}$ 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)

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.
- 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.

Confidence intervals

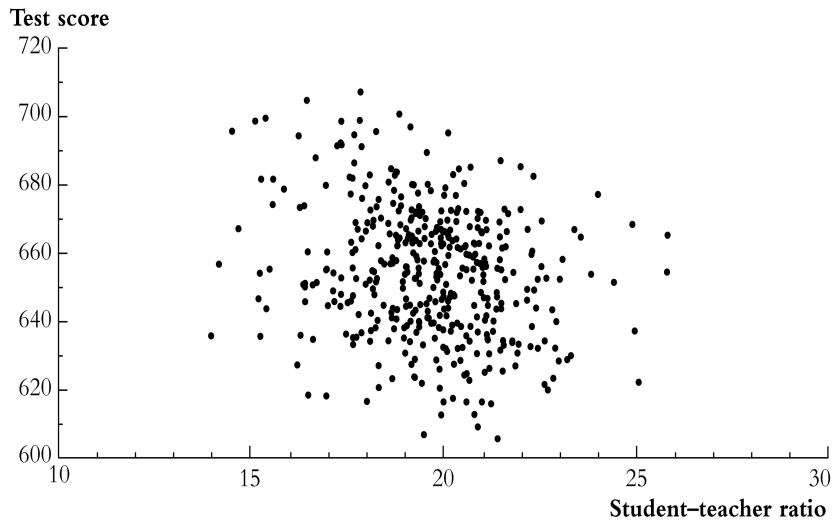
- A 95% confidence interval forμ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.
- The values of $Y_1, \ldots, Y_n$ and thus any functions of them – including the confidence interval are random. The confidence interval will differ from one sample to the next. The population parameter, μ_Y, is not random; we just don't know it.
- 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.

Back to the policy question

- What is the effect on test scores of reducing STR by one student/class?
- Have we answered this question?

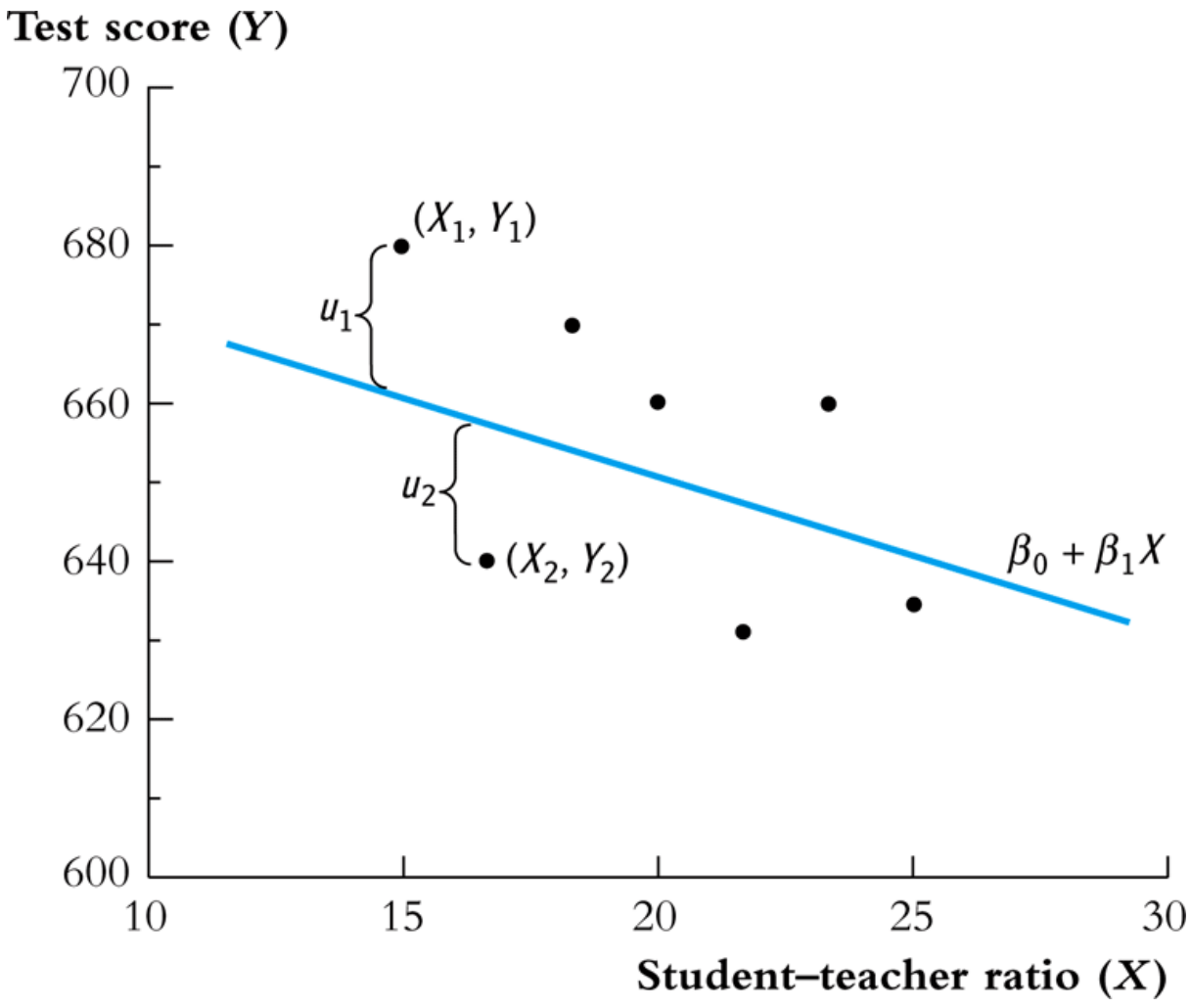
FIGURE 4.2 Scatterplot of Test Score vs. Student–Teacher Ratio (California School District Data)

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.



Running example: Class size and student performance

- Consider the problem faced by a school authority:
  - It is considering hiring additional teachers to reduce class; sizes
  - To evaluate this policy the authority would like to know how much student performance will increase as a result of this intervention;
- To help evaluate this policy, you have collected data on test scores and class sizes in 420 school districts in California in 1999. (See S & W, page 143 for a detailed description of the data)


Test score (Y)
700
680
660
640
620
600
10
15
20
25
30
Student–teacher ratio (X)
(X_1, Y_1)
(X_2, Y_2)
u_1
u_2
$\beta_0 + \beta_1 X$

The linear regression model

- The simplest way to summarize the relationship between two variables is to assume that they are linearly related
- We can express this with the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where:

- y_i is the dependent variable, or left-hand variable
- x_i is the independent variable, regressor, or right-hand variable
- u_i is the error term
- β_0 and β_1 are parameters to be estimated

The linear regression model (cntd.)

- The Population Regression Function is $\beta_0 + \beta_1$
- The subscript runs over observations $i = 1, \ldots, n$
- In our class size example
  - y_i is the average test score in the school district
  - x_i is the average class size in the school district
  - u_i contains all factors influencing test scores other than class size
  - β_1 is the effect of a one unit change in class size on test scores
  - What does β_0 represent?

Estimating the coefficients of the linear regression model

- If the parameter β_1 were known, it would be very easy to predict the effect of changes in class size.
- How can we estimate the size of β_1 from our data from school districts in California?
- The most widely used approach to estimating the parameters of the linear regression model is the ordinary least squares (OLS) method.

Ordinary Least Squares

- The OLS estimator chooses the regression coefficients so that the estimated regression line is "as close as possible" to the data.
- In particular it minimizes the sum of the squared deviations of the data from the regression line.
- Formally, from all possible β_0 and β_1, it chooses the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following expression:

$$\sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 \tag{2}$$

- The predicted value of y_i, denoted $\hat{y}_i$,is equal to $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Ordinary least squares (cntd.)

- With some algebra one can show that the solution to this minimization problem is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \tag{3}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

- Returning to our example, the estimated relationship between class size and test scores is

$$\hat{y}_i = 698.9 - 2.28 \times x_i$$

Ordinary least squares (cntd.)

- The interpretation of this result is that a decrease of the student teacher ratio by one student will increase test scores by 2.28 points.
- Is this a large or a small effect?
  - Relative to the sample mean of 654 this is an improvement of about 0.35 percent
  - Relative to the distribution of test scores it is equivalent to an improvement from the median to about the 55th percentile.

Why use OLS?

- The OLS estimator is the most popular estimator in applications
- The reason for this is that it has desirable statistical properties under certain assumptions:
  - It is unbiased and consistent
  - Under some additional assumptions it is also the most efficient estimator.
- We examine these conditions now in detail.

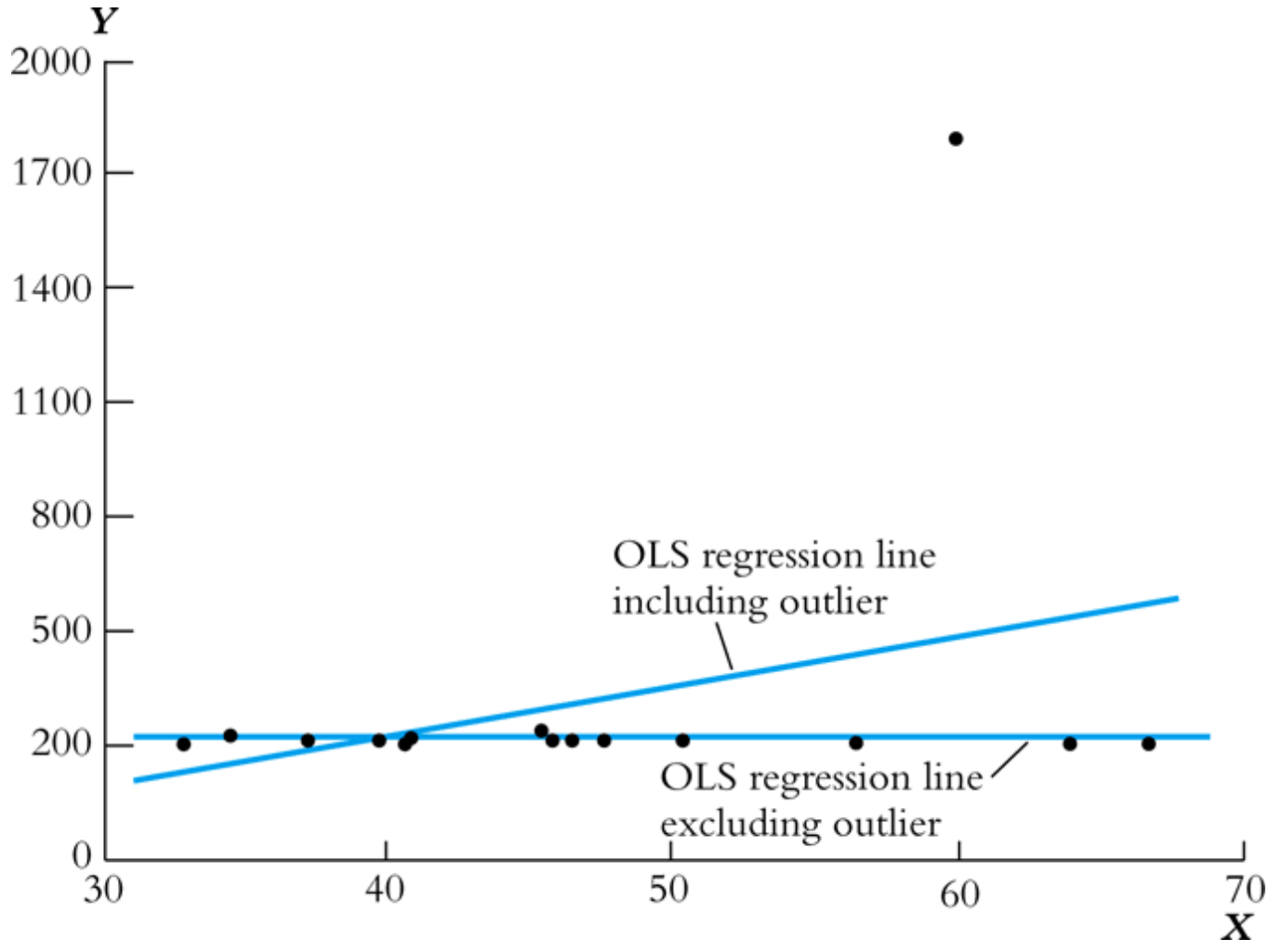
Assumptions of the OLS estimator

For the OLS estimator of the parameters β_0 and β_1 in (1) to be appropriate three key assumptions have to be satisfied:

1. Conditional Mean Independence Assumption: $E(u_i|X_i) = 0$
2. (X_i, Y_i) are i.i.d.: $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d.
3. Large outliers are unlikely

Assumptions of the OLS estimator (cntd.)

- Among these three assumptions Conditional Mean Independence is the most critical. Particularly if we want to use the language of causality, as we'll see next week.
- We will see later how to deal with violations of i.i.d. (e.g. time series analysis)
- Assumption 3 can be assessed from the data, but violation can lead to misleading estimation results.


Y
2000
1700
1400
1100
800
500
200
0
30
40
50
60
70
X
OLS regression line
including outlier
OLS regression line
excluding outlier

The sampling distribution of the OLS estimator

- The OLS estimator is consistent under conditions listed above.
- We now turn to the efficiency property of the distribution of the OLS estimator.
- It is possible to show that in the absence of heteroskedasticity the variance of the OLS estimator of β_1 in (1) is

$$var(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_x^2} \tag{5}$$

The sampling distribution of the OLS estimator (cntd.)

- What is the intuition behind this formula:
  - The larger the variance of the error term the less precise is the estimator of β_1
  - The estimator is more precise as the number of observations n increases
  - Finally, a larger variance of x (for a given σ_u^2) increases the precision of the estimator

The sampling distribution of the OLS estimator (cntd.)

- Even if we know that the OLS estimator of β_1 is consistent and also what its variance is, we still do not know its full distribution.
- It is possible to show that the estimator has a normal distribution if the error term is normally distributed.
- However, fortunately a version of the Central Limit Theorem implies that the estimator will be approximately normally distributed in large samples even if the error term is not normally distributed.
- We will therefore rarely have to rely on the assumption that the error term has a normal distribution to justify that the OLS estimator follows the normal distribution.

Omitted Variable Bias

- So far we have explained the variation in test scores only with the student teacher ratio.
- All other determinants of test scores are therefore included in the error term u.
- Which other determinants of test scores in the school districts of California can you think of?
- Is it problematic to leave these factors included in the error term u?

Omitted Variable Bias (cntd.)

- Omitting a variable from a regression will result in omitted variable bias if two conditions are met:
  - The omitted variable is correlated with the explanatory variable x
  - The omitted variable is a determinant of the dependent variable y
- Which variables may or may not meet these conditions in the test scores example?
- How would we solve this problem?

Multiple regression

- A fairly obvious response to the problem of omitted variable bias is to include further explanatory variables in the regression (1).
- We could, for example, use an alternative model with two explanatory variables x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{6}$$

- Note that we have now for simplicity dropped the subscript i.
- Is this a general solution to the problem of omitted variable bias?

Multiple regression (cntd.)

To use the OLS estimator to estimate β_1 and β_2 we need one additional assumption:

- There is no perfect multicollinearity between the explanatory variables.

Four Assumptions of the OLS estimator

1. Conditional Mean Independence Assumption: $E(u_i|X_i) = 0$
2. (X_i, Y_i) are i.i.d.: $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d.
3. Large outliers are unlikely
4. There is no perfect multicollinearity between the explanatory variables.

Examples of perfect multicollinearity

- Perfect linear combination of each other (percentages and fractions)
- Intercept is a column of 1s, so any perfect combination with it will also result in perfect multicollinearity.
- Dummy variable trap.

Imperfect multicollinearity

- Imperfect multicollinearity is defined as a situation in which there is a "high" degree of correlation between two explanatory variables, but they are not a perfect linear combination of each other.
- Such a correlation does not cause any problems for the OLS estimator. However:
  - it will be difficult to separately identify the influence of the two variables that are highly correlated
  - As a result standard errors will be large and t-statistics small
- There is no general solution to this "problem" (other than getting more data), or dropping some variables.

Interpreting the coefficients in a multiple regression

- In the multiple regression model the interpretation of the regression coefficient is somewhat different from the simple regression.
- In the simple regression (1) β_1 represents the effect of a one unit change in the explanatory variable on the dependent variable.
- In the multiple regression (6) β_1 represents the effect of a one unit change in x_1 on y holding x_2 constant.
- Formally, β_1 is the partial derivative of (6) with respect to x_1.

Measures of Fit

- How does our model perform? Are we any better than a random guess?
- What proportion of the variation in the dependent variable can be explained by the explanatory variables?
- How tightly the observations are clustered around the regression line?

R-squared

- The derivation of the R^2 starts from the identity

$$y_i = \hat{y}_i + \hat{u}_i \tag{7}$$

where

  - y_i is the actual value of the dependent variable for observation i
  - $\hat{y}_i$ is the value of the dependent variable predicted by the regression for observation i
  - $\hat{u}_i$ is the residual and is defined as the deviation of observation i from the regression line, i.e. $\hat{u}_i \equiv y_i - \hat{y}_i$

- Note that the residual $\hat{u}_i$ is not at all the same thing as the error term u_i of the regression model in (1) and (6).

R-squared (cntd.)

- It is possible to show that the total variation in the dependent variable can be decomposed into:

$$TSS = SSR + ESS \tag{8}$$

where

  - TSS (Total sum of squares) equals $\sum_{i=1}^{n}(y_i - \bar{y})^2$
  - ESS (Explained sum of squares) equals $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - SSR (Sum squared residuals) equals $\sum_{i=1}^{n}(y_i - \hat{y})^2$ or simply $\sum_{i=1}^{n}(\hat{u}_i)^2$

- Note that other textbooks use ESS to denote the concept that we (and Stock and Watson) denote with SSR.

R-squared (cntd.)

- The R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{9}$$

  and therefore varies between zero and one.
- How useful is the R^2?
  - A large R^2 most certainly does not imply that a regression represents a causal relationship
  - Similarly, a low R^2 does not by itself mean that a regression is hopeless
- We will learn to use the R^2 as a useful piece information even if it is not a "hard fact".

The Adjusted R-squared

- R-squared increases when you add a new variable, without corresponding increase in the fit of the model.
- This inflation is corrected through the "adjustment" to the number of independent variables in the model.
- Hence the Adjusted R-squared.

Standard Error of the Regression

- A different measure of fit: estimates the standard deviation of the error term u_i.
- SER is a measure of the spread of the distribution of Y around the regression line.

$$SER = \sqrt{\frac{SSR}{n-k-1}}$$

- In R output RMSE is almost the same as SER (tiny difference in dividing by n rather than $n-k-1$). Both are measures of the spread of Ys around the regression line.

Background example

- Consider the following problem: An "angry taxpayer" is of the opinion that reducing class sizes is a waste of (his) money as it has no impact on educational outcomes.
- How can we evaluate this statement using our data for the 420 Californian school districts?
- We know that a reduction in class size by one student improves test results by 2.28 points in this dataset.
- Is it possible to determine how incompatible this estimate is with the opinion of the angry taxpayer?

Null and alternative hypothesis

- One way to address the claim of the angry taxpayer is to specify the following null and alternative hypothesis:
  - H_0: the student teacher ratio has no effect on test scores
  - H_1: the student teacher ratio has an effect on test scores
- Note that one could also entertain the one-sided alternative hypothesis that the student teacher ratio has a positive effect on test scores.
- We want to test these hypotheses in the context of the simple regression:

$$testscr = \beta_0 + \beta_1 str + u \tag{10}$$

Estimating the standard error of $\hat{\beta}_1$

- To discriminate between H_0 and H_1 we need to know how the OLS estimator $\hat{\beta}_1$ of the parameter β is distributed under H_0.
- Stock&Watson Chapter 2 shows that the OLS estimator of β_1 in the absence of heteroskedasticity and autocorrelation has variance:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_u}{n\sigma^2_x} \tag{11}$$

- Unfortunately, this expression depends on σ^2_u and σ^2_x, which are unknown and need to be estimated.

Estimating the standard error of $\hat{\beta}_1$

- It is possible to show that

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{s^2_{\hat{u}}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{12}$$

  is an unbiased and consistent estimator of (11), where

$$s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i)^2 \tag{13}$$

- The square root of $s^2_{\hat{u}}$ is the standard error of the regression (SER) we discussed above.
- What $s^2_{\hat{u}}$ basically does is to use the residuals to estimate the variance of the underlying error term.

The t-test

- With the estimator (12) at hand we can compute the t-statistic

$$t = \frac{\hat{\beta}_1 - H_0}{\hat{\sigma}_{\hat{\beta}_1}} \tag{14}$$

  where $\hat{\sigma}_{\hat{\beta}_1}$ is usually referred to as the standard error of $\hat{\beta}_1$.
- Note that in the very common case where the null hypothesis is $\beta_1 = 0$ the t-statistic simplifies to $\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}$
- In large samples the t-statistic will approximately follow the standard normal distribution.

The t-test

- In small samples the t-statistic will follow the t-distribution if the error term u is normally distributed.
- If the error term is not normally distributed, the distribution of the t-statistic does not follow any simple distribution in small samples.
- Most regression packages (including R) nevertheless use the t-distribution to compute critical values for the t-statistic:
  - However, even for moderately large samples it is practically irrelevant whether we use the *t* or *normal* distribution
  - For small samples this means that the program implicitly assumes that the error term is (hopefully) normally distributed

Application to test scores

- For the regression of test scores on the student teacher ratio we obtain:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR \qquad (15)$$

  where the numbers in brackets are the standard errors of the coefficients.
- These are heteroskedasticity robust standard errors, which we will return to later.

Application to test scores

- To test the hypothesis of the angry taxpayer we have to compute:

$$\frac{\hat{\beta}_1 - H_0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-2.28 - 0}{0.52} = -4.38 \tag{16}$$

- The probability of observing a value of the t-statistic outside the interval $[1.96, 1.96]$ is less than five percent under the standard normal distribution.
- As the t-statistic is clearly outside this interval, the probability that H_0 is correct is less than five percent.
- We can therefore reject the claim of the angry taxpayer at the five percent significance level.

Statistical significance

- In the vast majority of t-tests the null hypothesis is that the coefficient is equal to zero.
- In this case the null hypothesis is often not even stated and you will encounter statements such as:
  - The coefficient is statistically significant at the x percent level
  - The coefficient is significant at conventional levels
  - The coefficient is highly significant
- In all of these statements the implicit null hypothesis (or simply "null") is that the coefficient of interest is equal to zero.
- We should not forget that t-test can nevertheless be used to test also other null hypotheses.
- For example, can we reject the null hypothesis that the true effect of the student teacher ratio on test results is -3.0?

Computing p-values

- Can we determine more precisely how unlikely the angry taxpayer's hypothesis is given our estimate for the Californian school districts?
- The p-value associated with the t-statistic does exactly that and can be defined in three equivalent ways:
  - It is the probability that the null hypothesis is true
  - It is the lowest (most stringent) significance level at which the null hypothesis can just be rejected.
  - Finally, it is the probability of getting a value of the t-statistic that is as large or larger (in absolute terms) as the observed t-statistic, when the null is true.

Computing p-values

- Formally, the p-value associated with a value of the t-statistic t_{act} in a two-sided test is the probability mass of the standard normal distribution outside the interval $[-t_{act}, t_{act}]$.
- How does the p-value change if we consider a one-sided alternative hypothesis?
- Modern regression packages will always report the p-value for any t-statistic (and also for other regression statistics).
- For the estimate of β_1 R reports a p-value of 0.000.
- What does this mean?

Confidence intervals for regression coefficients

- As in the case of estimating the mean, a simple extension of the logic of hypothesis testing are confidence intervals.
- The 95 percent confidence interval, for example, can be interpreted in two equivalent ways:
  - The true parameter will be within the confidence interval with a 95 percent probability.
  - The confidence interval contains all values of the parameter that cannot be rejected at the five percent significance level given our estimate.
- In the case of regression (15) the 95 percent confidence interval for the coefficient β_1 is the interval

$$[-2.28 - 1.96 \times 0.52, -2.28 + 1.96 \times 0.52] \qquad (17)$$

- How would you compute the 99 percent confidence interval?

The t-test in multiple regression

- We can also use the t-test to test hypothesis about one regression coefficient in a multiple regression.
- The formula used to compute the standard error is more complicated, but the same intuitions as in the case of the simple regression hold.
- Consider, for example, the following regression:

$$\begin{aligned} \widehat{TestScore} &= 686 - 1.10 \times STR - 0.650 \times PctEL \quad (18) \\ &\quad (8.7) \quad (0.43) \qquad\quad (0.031) \end{aligned}$$

- Which coefficients of this regression are statistically significant at the 5 percent level?

Tests of joint hypotheses

- We will often be interested in testing hypotheses that involve more than one coefficient.
- Possible examples would be:
  - The student teacher ratio has the same effect on test scores as the percentage of students on a subsidized lunch
  - Both the student teacher ratio and the percentage of students on a subsidized lunch do not affect test scores
- To evaluate these hypotheses we need something more powerful than the t-test and will use an F-test.

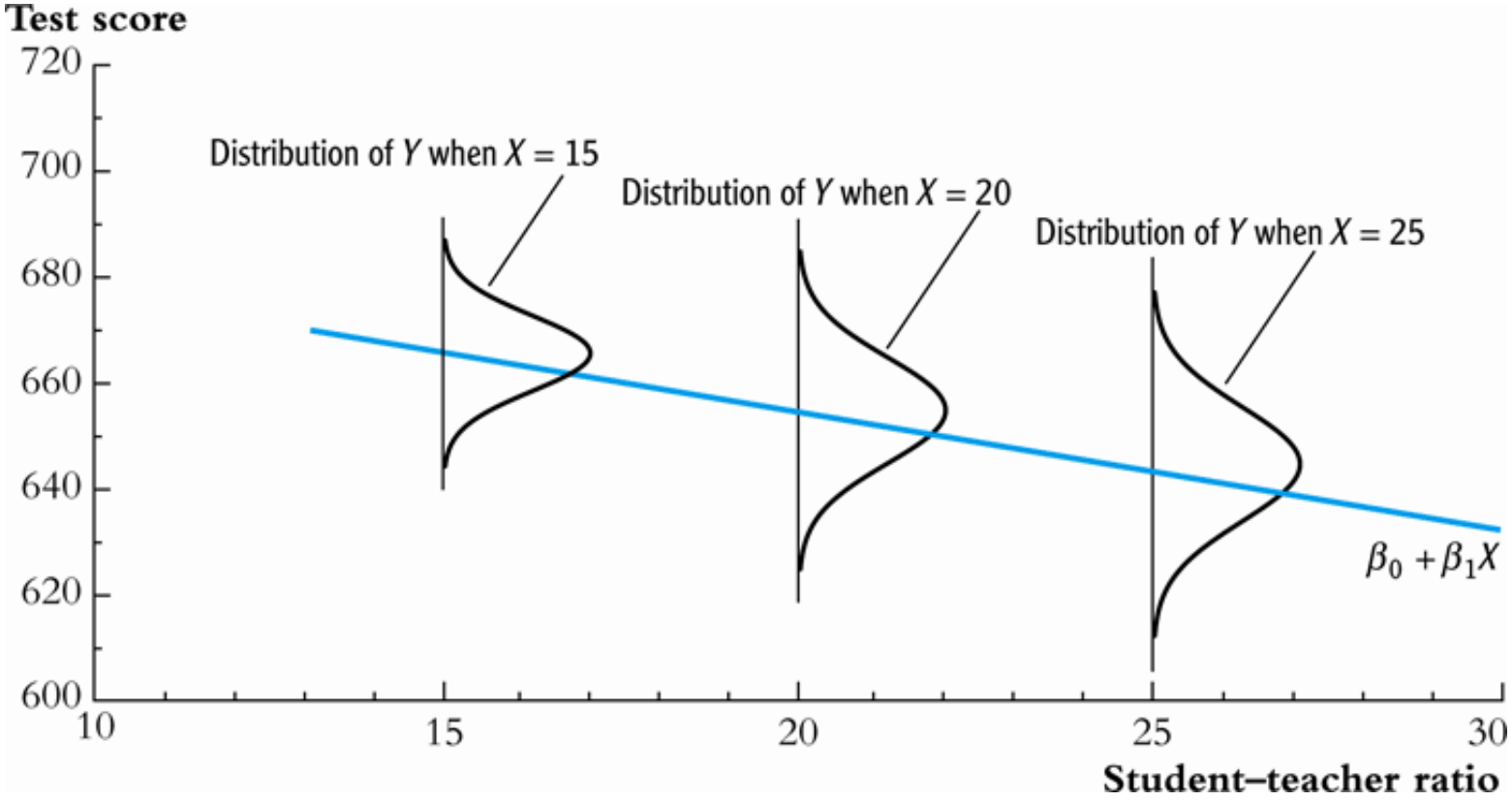
The F-test

- When the joint null has the two restrictions that $\beta_1 = 0$ and $\beta_2 = 0$, the F-statistic combines the two t-statistics t_1 and t_2:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \tag{19}$$

- The F-statistic compares the ability of the regression to fit the data in the presence and in the absence of the hypothesis that is being tested.
- Note that the a F-test of a hypothesis about one coefficient is equivalent to a t-test.

Homoskedasticity vs heteroskedasticity



Model specification in theory and practice

- Base specification and Alternative specification
- Rsq and Adjusted Rsq
- "Star gazing"

Presenting results

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

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.