Applied Econometrics 122B
Linear model review

Giuseppe Ragusa

gragusa@luiss.it
http://gragusa.org/

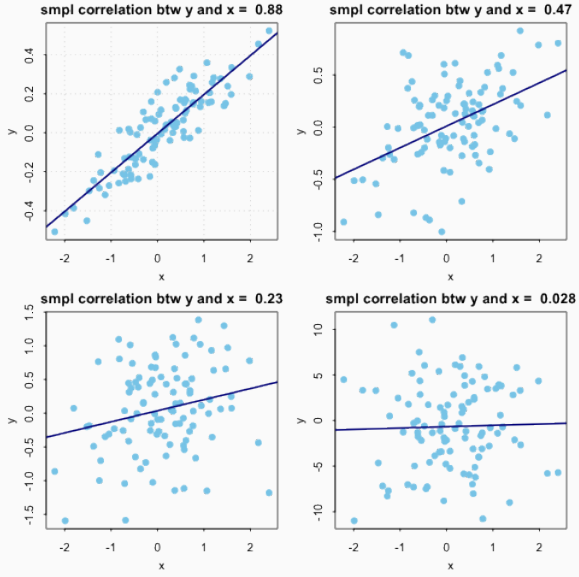
June 27, 2017

University of California, Irvine

Covariance and correlation

- Covariance and correlation are measure of linear dipendence between variables
- The best way of picturing linear association is to immagine to fit a line through the cloud of points in a scatterplot
- The better the could of points be summarized by a line, the stronger is the linear association

Covariance and correlation



The best “line”

In the previous graphics I draw a line through the cloud of points.

Question:

How did I draw this line?

- Put differently: which is the line that better approximate the cloud of points?

The best “line”

In the previous graphics I draw a line through the cloud of points.

Question:

How did I draw this line?

- Put differently: which is the line that better approximate the cloud of points?

Answer:

We use the so called method of least squares.

Method of least squares

Hystory

Carl Friedrich Gauss is credited with developing the fundamentals of the basis for least-squares analysis in 1795 at the age of eighteen. But Adrien Marie Legendre, a french mathematician, was the first to publish the method, however.

- Gauss applied this method to predict the location of Ceres, a dwarf planet discovered by Italian Giuseppe Piazzi in 1801.

Carl Gauss

The ordinary least squares (intuition)

Sample mean

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 taking the first derivative of the objective function and equating to 0, it is easy to see that the solution of $\bar{Y}$

$$-2\sum_{i=1}^{n} Y_i + nm = 0 \implies m = \bar{Y} = \sum_{i=1}^{n} Y_i/n$$

The ordinary least squares (intuition)

Now, we do not have to calculate the mean, but the coefficients of a line:

$$Y = b_0 + b_1 X,$$

where

- b_0 is the intercept
- b_1 is the slope

The ordinary least squares (intuition)

Now, we do not have to calculate the mean, but the coefficients of a line:

$$Y = b_0 + b_1 X,$$

where

- b_0 is the intercept
- b_1 is the slope

Least squares idea

Among all the possible values of (b_0, b_1), we pick the ones that give "predictions" that are closer (in the least squares sense) to the observed data:

$$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, \ldots, n$$

The ordinary least squares (intuition)

Assume: $b_0 = -0.6$ and $b_1 = 0.14$

Observed data

y	x
-0.032	-0.626
-0.885	0.184
-7.024	-0.836
-2.677	1.595
1.539	0.330
1.785	-0.820
-5.943	0.487

Predictions

$\hat{Y}_i$
$-0.6 + 0.14 \times (-0.626) = -0.68$
$-0.6 + 0.14 \times (0.184) = -0.57$
$-0.6 + 0.14 \times (-0.836) = -0.71$
$-0.6 + 0.14 \times (1.595) = -0.37$
$-0.6 + 0.14 \times (0.330) = -0.55$
$-0.6 + 0.14 \times (-0.820) = -0.71$
$-0.6 + 0.14 \times (-0.487) = -0.63$

Least squares (intuition)

Predictions and squares

$\hat{Y}_i$	$(\hat{Y}_i - Y_i)^2$
-0.68	0.431
-0.57	0.096
-0.71	39.776
-0.37	5.291
-0.55	4.379
-0.71	6.249
-0.63	29.28

Sum of squares

The sum of the squares for the line with $b_0 = -0.6$ and $b_1 = 0.14$ is

$$\begin{aligned} SSR =& 0.431 + 0.096 + 39.776 \\ &+ 5.291 + 4.379 + 6.249 + 29.28 \\ =& 85.50 \end{aligned}$$

Least squares

OLS

The least squares ("ordinary least squares" or "OLS") estimators of b_0 and b_1 solve

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

The OLS estimator solves:

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

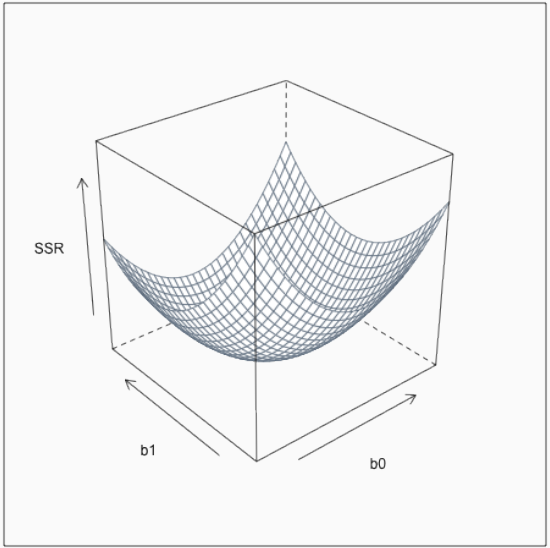
- The OLS estimator minimizes the average squared difference between the actual values of Y_i and the prediction ("predicted value") based on the estimated line $(b_0 + b_1 X_i)$
- We have a multivariate function in (b_0, b_1):

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

- *SSR* stands for sum of squared residuals
- The residuals are

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i, \quad i = 1, \ldots, n$$

Sum of squares: graphical representation



Refresh your math

- The minimum value of a function $f(x)$ is denoted

$$\min_x f(x).$$

- The argument for which the function achieves its minimum is denoted

$$\arg\min_x f(x).$$

- If the function is strictly convex, the necessary and sufficient condition for x_0 to be a minimizer of $f(x)$ is

$$\frac{d}{dx} f(x_0) = 0$$

Refresh your math

Example: the math

Let $f(x) = 2 + x^2$. Then,

$$\min_x f(x) = 2$$

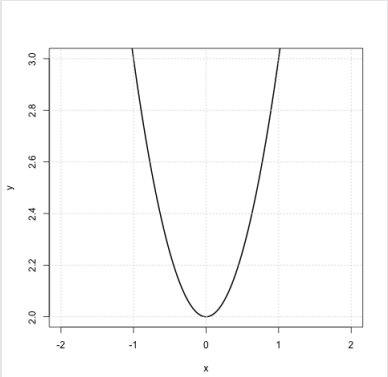
and

$$\arg\min_x f(x) = 0.$$

Also,

$$\frac{d}{dx} f(x) = 2x \implies \frac{d}{dx} f(0) = 0$$

Graph of $2 + x^2$



Least squares first order conditions

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

First order conditions

$$\frac{\partial \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{\partial b_0} = 0 \implies -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_1)^2}{\partial b_1} = 0 \implies -2 \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0$$

Solving for the first order conditions

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \tag{1}$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0 \tag{2}$$

- Two equations in two unknowns (b_0 and b_1)

Solving for the first order conditions

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \tag{1}$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0 \tag{2}$$

- Two equations in two unknowns (b_0 and b_1)
- Solve for b_0 in (1), to obtain

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} X_i$$

Solving for the first order conditions

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \tag{1}$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0 \tag{2}$$

- Two equations in two unknowns (b_0 and b_1)
- Solve for b_0 in (1), to obtain

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Substitute b_0 into (2) and solve for b_1

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Lest squares solution for b_0 and b_1

The convention is to denote the solution of the first order conditions as $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

Remarks

- The underlying assumption for the existence of a solution is that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0$$

(*When does this happen?*)

Remarks

- The underlying assumption for the existence of a solution is that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \neq 0$$

  (*When does this happen?*)
- If the FOCs can be solved, the FOCs are sufficient for the minimum (why?)

Remarks

- The least squares method that we just used is usually referred to as "ordinary least squares" (OLS)
- Although b_0 and b_1 are the solutions to a mathematical problem, their are function of sample statistics: sample mean of Y, sample mean X, variance X, and covariance of Y and X
- In due time, we will also see how to solve a more complicated OLS problem, one in which we fit planes or hyper-planes instead of lines.

Conditional expectations as object of interest

We are often interested in the expectation of a r.v. conditional to the value of another.

Example (Wage and education)

- You may (and probably should) what would be your wage after you graduate from college
- If educ denotes years of education and wage the wage rate, you are interested in this quantity

$$E[wage|educ = 16],$$

that is, the expected value of wages for individuals with 16 years of education, exactly the same number of years of education you will have

Conditional expectations as object of interest

Example (Wage and education, ctd.)

- Suppose now you want to know whether to continue your education with a master degree
- To see whether completing a master is worth your money you are concerned about this quantity

$$E[wage|educ = 17] - E[wage|educ = 16],$$

that is, the difference in wage due to the extra year of education

Conditional expectations as object of interest

Policy problem: testscore and str

A school superintend must decide whether to hire new 10 new teachers

- she faces a trade-off because hiring 40 new teachers:
  - will reduce the student-per-teacher (STR) ratio by 2, from 22 to 20
  - will increase expenditures by $1,800,000.

  If she reduces the student-teacher ratio by 2, what will the effect be on standardized test scores in her district?

Policy answer

$$E[testscore|str = 20] - E[testscore|str = 22]$$

A linear model for conditional expectations

- The conditional expectation is a population quantity and so we need to estimate its value
- Unfortunately, estimating conditional expectation is quite difficult is we are not willing to make some assumptions on its form
- A very powerful assumption is the linear one:

Linear assumption

Given a conditional expectation btw Y and X, we assume that this expectation is linear in X, that is,

$$E[Y|X] = \beta_0 + \beta_1 X,$$

where β_0 and β_1 are two parameters

Linear model and conditional expectations

$$E[testscore|str] = \beta_0 + \beta_1 str$$

Linear model and conditional expectations

$$E[testscore|str] = \beta_0 + \beta_1 str$$

- What is the expected value of testscore in a district with $str = 19$?

$$E[testscore|str = 19] = \beta_0 + \beta_1 \times 19$$

Linear model and conditional expectations

$$E[testscore|str] = \beta_0 + \beta_1 str$$

- What is the expected value of testscore in a district with $str = 19$?

$$E[testscore|str = 19] = \beta_0 + \beta_1 \times 19$$

- What is the expected value of testscore in a district with $str = 20$?

$$E[testscore|str = 20] = \beta_0 + \beta_1 \times 20$$

Linear model and conditional expectations

$$E[testscore|str] = \beta_0 + \beta_1 str$$

- What is the expected value of testscore in a district with $str = 19$?

$$E[testscore|str = 19] = \beta_0 + \beta_1 \times 19$$

- What is the expected value of testscore in a district with $str = 20$?

$$E[testscore|str = 20] = \beta_0 + \beta_1 \times 20$$

- What is the expected value of testscore in a district with $str = 23$?

$$E[testscore|str = 23] = \beta_0 + \beta_1 \times 23$$

Linear model and conditional expectations

Question: what is the effect of increasing str by 1?

$$\begin{aligned} E[testscore|str = 20] - E[testscore|str = 19] &= (\beta_0 + \beta_1 20) - (\beta_0 + \beta_1 19) \\ &= \beta_1 \end{aligned}$$

Interpretation of β_1

Under the linear assumption, β_1 is the change in the CE of `testscore` when `str` is increased by 1 unit

Linear model and conditional expectations

Question: what is the effect on the CE of increasing str by Δ?

$$E[testscore|str = 20 + \Delta] - E[testscore|str = 20] = \beta_1 \times \Delta$$

Interpretation of β_1

Under the linear assumption, $\beta_1 \times \Delta$ is the change in the CE of testscore when *str* is increased by Δ unit

Linear model and conditional expectations

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

Interpretation of β_1

$$\beta_1 = \frac{\partial E[Y_i|X_i]}{\partial X_i}$$

that is, β_1 is the effect on the CE of Y when X_i is increased by 1 unit

Linear model and conditional expectations

Interpretation of β_0

$$\beta_0 = E[Y_i|X_i = 0]$$

that is, the CE of Y when $X_i = 0$

The linear assumption

Linear assumption

If

$$E[Y|X] = \beta_0 + \beta_1 X$$

then we can write

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

where

$$E[u|X] = 0$$

.

- u is called the regression error
- it consists of omitted factors, or possibly errors in the measurement of Y.
  - In general, these omitted factors are other variables that influence Y, other than the X

The linear assumption

Implications of the linear assumption

- Since $E[u|X] = 0$ implies that $\text{cov}(u, X) = 0$, the linearity assumption can be interpreted as saying that we are ruling out linear relationship between u and X (that is, the correlation between u and X is zero).
- We will see that this assumption is usually questionable, but for now, we will stick to it

The population linear regression model - general notation

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

- X is the independent variable or regressor
- Y is the dependent variable
- $\beta_0 =$ intercept
- $\beta_1 =$ slope
- $u_i =$ the regression error

Terminology in a picture

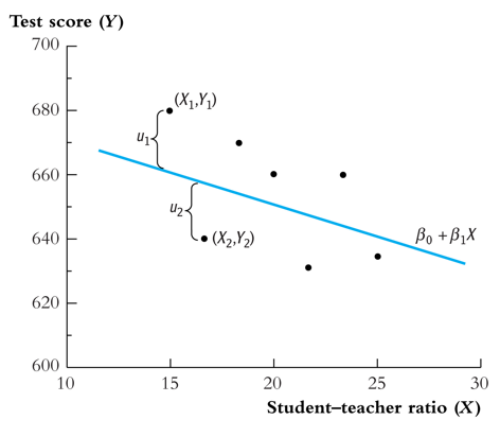


Figure 3: Observations on Y and X; the population regression line; and the regression error

Challenges ahead

The problems of statistical inference for linear regression are, 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:

Challenges ahead

The problems of statistical inference for linear regression are, 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 estimate β_0 and β_1? (answer: ordinary least squares). What are advantages and disadvantages of OLS?

Challenges ahead

The problems of statistical inference for linear regression are, 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 estimate β_0 and β_1? (answer: ordinary least squares). What are advantages and disadvantages of OLS?
- Hypothesis testing
  - How to test if β_1 (or β_0) is zero?

Challenges ahead

The problems of statistical inference for linear regression are, 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 estimate β_0 and β_1? (answer: ordinary least squares). What are advantages and disadvantages of OLS?
- Hypothesis testing
  - How to test if β_1 (or β_0) is zero?
- Confidence intervals
  - How to construct a confidence interval for β_1?

Estimation of β_0 and β_1

We already have a way of estimating the intercept and slope using the least squares method:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

Implementing OLS

Often the output of the regression is written as:

$$\widehat{testscore} = 698.933 - 2.280 \times str$$

- Estimated slope: $\hat{\beta}_1 = -2.2798$

Implementing OLS

Often the output of the regression is written as:

$$\widehat{testscore} = 698.933 - 2.280 \times str$$

- Estimated slope: $\hat{\beta}_1 = -2.2798$
- Estimated intercept: $\hat{\beta}_0 = 698.9329$

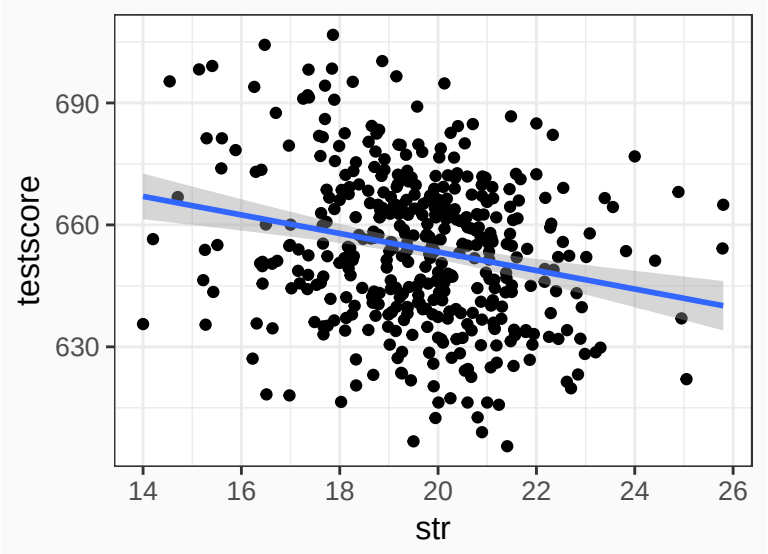
Implementing OLS

Often the output of the regression is written as:

$$\widehat{testscore} = 698.933 - 2.280 \times str$$

- Estimated slope: $\hat{\beta}_1 = -2.2798$
- Estimated intercept: $\hat{\beta}_0 = 698.9329$
- $\widehat{TestScore}$ denotes the estimated regression line

Graphic representation



Interpretation of estimated intercept and slope

- Districts with one more student per teacher on average have test scores that are 2.28 points lower;
- That is,

$$\frac{\Delta TestScore}{\Delta STR} = -2.28$$

- 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
- 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

Linear Regression Model: General

Although we have been discussing about the testscore and str, the above model is more general, so we will find useful to introduce more general notation:

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

is the **linear regression model with a single regressor.**

Linear Regression Model: General

Although we have been discussing about the testscore and str, the above model is more general, so we will find useful to introduce more general notation:

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

is the **linear regression model with a single regressor.**

- Y_i is the **dependent variable**
- X_i is the **independent variable** or **regressor**

Linear Regression Model: General

Although we have been discussing about the testscore and str, the above model is more general, so we will find useful to introduce more general notation:

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

is the **linear regression model with a single regressor.**

- Y_i is the **dependent variable**
- X_i is the **independent variable** or **regressor**
- $\beta_0 + \beta_1 X_i$ is the **population regression line**

Linear Regression Model: General

Although we have been discussing about the testscore and str, the above model is more general, so we will find useful to introduce more general notation:

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

is the **linear regression model with a single regressor.**

- Y_i is the **dependent variable**
- X_i is the **independent variable** or **regressor**
- $\beta_0 + \beta_1 X_i$ is the **population regression line**
- u_i is the **error term** incorporating all the factors responsible for the difference between Y_i and $X_i \beta_1$

Regression functions and conditional expectations

Conditional Expectation

If $E[u_i|X_i] = 0$ then

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

$\beta_0 + \beta_1 X_i$ is the conditional expectation of Y given X.

Regression functions and conditional expectations

Conditional Expectation

If $E[u_i|X_i] = 0$ then

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

$\beta_0 + \beta_1 X_i$ is the conditional expectation of Y given X.

Caveats

$E[u_i|X_i] = 0$ is a **big** assumption and we will question it in a bit. However, for now, we all assume it holds.

The Least Squares Assumptions (SW Section 4.4)

- What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it?

The Least Squares Assumptions (SW Section 4.4)

- What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it?
- Under what conditions is it an unbiased estimator of the true population parameters?

The Least Squares Assumptions (SW Section 4.4)

- What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it?
- Under what conditions is it an unbiased estimator of the true population parameters?
- 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)

The Least Squares Assumptions (SW Section 4.4)

- What, in a precise sense, are the properties of the OLS estimator? We would like it to be unbiased, and to have a small variance. Does it?
- Under what conditions is it an unbiased estimator of the true population parameters?
- 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

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

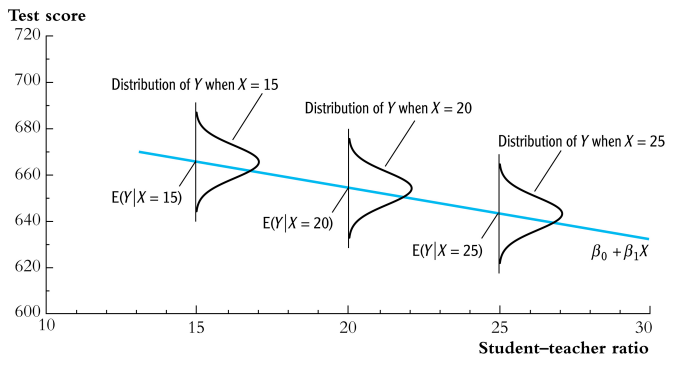
- The conditional distribution of u given X has mean zero, that is,

$$E(u_i | X_i = x) = 0$$

- (Y_i, X_i) are iid
- Large outliers in Y and X are rare

Least squares assumption 1: $E(u_i|X_i = x) = 0$.

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



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.

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 independently distributed of X

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 independently distributed of X
- Thus, in an ideal randomized controlled experiment, $E(u_i|X_i = x) = 0$ holds.

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 independently distributed of X
- Thus, in an ideal randomized controlled experiment, $E(u_i|X_i = x) = 0$ holds.
- In actual experiments, or with observational data, we will need to think hard about whether $E(u_i|X_i = x) = 0$ holds.

Note

Fact

If $E(u|X = x) = 0$, then $\text{cov}(u, X) = 0$. The converse is not true.

Thus, checking whether $E(u|X = x) = 0$ can be done by checking whether $\text{cov}(u, X) = 0$.

In particular, Assumption 1 will be violated if the other factors are correlated with X. (Again, $\text{cov}(u, X)$ maybe 0, but $E(u|X) \neq 0$).

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 entity is selected then, for that entity, X and Y are observed (recorded).

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 entity is selected then, for that entity, X and Y are observed (recorded).
- The main place we will encounter non-i.i.d. sampling is when data are recorded over time ("time series data") – this will introduce some extra complications.

Least squares assumption 3: Large outliers are rare

$$E(Y^4) < \infty, \quad E(X^4) < \infty$$

- A large outlier is an extreme value of X or Y

Least squares assumption 3: Large outliers are rare

$$E(Y^4) < \infty, \quad E(X^4) < \infty$$

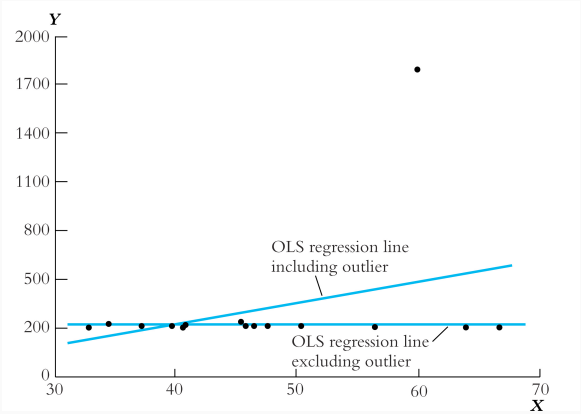
- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; str, family income, etc. satisfy this too).

Least squares assumption 3: Large outliers are rare

$$E(Y^4) < \infty, \quad E(X^4) < \infty$$

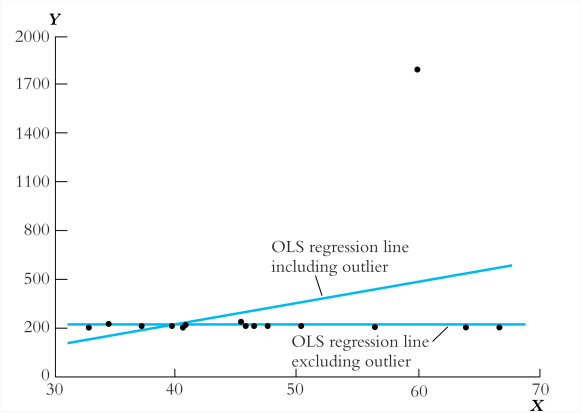
- A large outlier is an extreme value of X or Y
- On a technical level, if X and Y are bounded, then they have finite fourth moments. (Standardized test scores automatically satisfy this; str, family income, etc. satisfy this too).
- However, the substance of this assumption is that a large outlier can strongly influence the results

Outliers



- Is the lone point an outlier in X or Y?

Outliers



- Is the lone point an outlier in X or Y?
- In practice, outliers often are data glitches (coding/recording problems) – so check your data for outliers!

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

The OLS estimator is computed from a sample of data; a different sample gives 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 $\hat{\beta}_1$;

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

The OLS estimator is computed from a sample of data; a different sample gives 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 $\beta_1 = 0$;

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

The OLS estimator is computed from a sample of data; a different sample gives 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;

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

The OLS estimator is computed from a sample of data; a different sample gives 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...

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

The OLS estimator is computed from a sample of data; a different sample gives 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;

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

The OLS estimator is computed from a sample of data; a different sample gives 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

Population

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

Random variables

Y, X Ex: (Test Score, str)

Joint distribution of (Y, X)

The population regression function is linear $E(u_i|X_i) = 0$ (1st Least Squares Assumption)

X, Y have finite fourth moments (3rd L.S.A.)

Data Collection by simple random sampling

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

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

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

- What is $E(\hat{\beta}_1)$?
- What is $\mathrm{var}(\hat{\beta}_1)$?
- What is the distribution of $\hat{\beta}_1$ in small samples?
- What is the distribution of $\hat{\beta}_1$ in large samples?

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

Preliminary algebra

Given $Y_i = \beta_0 + \beta_1 X_i + u_i$ and taking means on both sides, noting that

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

we can express the model in mean deviation

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u}). \tag{3}$$

Substituting (1) into the expression for $\hat{\beta}_1$, we get

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

Mean and variance of the sampling distribution of $\hat{\beta}_1$, ctd.

We have that

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

Mean and variance of the sampling distribution of $\hat{\beta}_1$, ctd.

Expected value of $\hat{\beta}_1$

$$
\begin{aligned}
E\left[\hat{\beta}_1 - \beta_1\right] &= E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= E\left\{E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big| X_1, \ldots, X_n\right]\right\} \\
&= E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})E(u_i|X_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \\
&= 0
\end{aligned}
$$

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

- $E[\hat{\beta}_1 - \beta_1] = 0 \implies E[\hat{\beta}_1] = \beta_1$
- LSA #1 implies that $\hat{\beta}_1$ is unbiased;
- For details see App. 4.3

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

Next, calculate $\mathrm{var}(\hat{\beta}_1)$.

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

If $n \to \infty$, $\frac{(n-1)}{n}\hat{\sigma}_X^2 \xrightarrow{p} \sigma_X^2$, and $(X_i - \bar{X})u_i \xrightarrow{p} (X_i - \mu_X)u_i$. Thus,

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

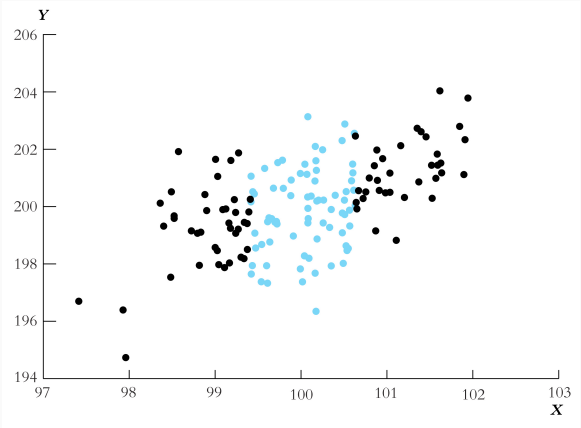
where $\nu_i = (X_i - \mu_X)u_i$.

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

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}(\nu_i)}{\sigma_X^4} = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4}$$

- The variance of $\hat{\beta}_1$ is inversely proportional to n— just like $\text{var}(\bar{Y})$.
- The larger the variance of X, the smaller the variance of $\hat{\beta}_1$
  - **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.

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



There are the same number of black and blue dots – using which would you get a more accurate regression line?

Large sample 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) approximation:

$$\hat{\beta}_1 \xrightarrow{d} N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right).$$

where

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma^4_X}.$$

Parallel btw the asymptotic distribution of β_1 and $\bar{Y}$

$\hat{\beta}_1$

- $E[\hat{\beta}_1] = \beta_1$
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$
- $\hat{\beta}_1 \xrightarrow{d} N(\beta_1, \sigma^2_{\hat{\beta}_1})$
- $\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma^4_X}$

$\bar{Y}$

- $E[\bar{Y}] = \mu_Y$
- $\bar{Y} \xrightarrow{p} \mu_Y$
- $\bar{Y} \xrightarrow{d} N(\mu_Y, \sigma^2_{\bar{Y}})$
- $\sigma^2_{\bar{Y}} = \frac{\sigma^2_Y}{n}$

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

If the three LS assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4} \propto \frac{1}{n}$$

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

If the three LS assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4} \propto \frac{1}{n}$$

- Other than its mean and variance, the exact distribution of is complicated and depends on the distribution of (X, u);

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

If the three LS assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4} \propto \frac{1}{n}$$

- Other than its mean and variance, the exact distribution of is complicated and depends on the distribution of (X, u);
- $\hat{\beta}_1 \xrightarrow{p} \beta_1$, (that is, $\hat{\beta}_1$ is consistent)

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

If the three LS assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4} \propto \frac{1}{n}$$

- Other than its mean and variance, the exact distribution of 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

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0, 1)$$

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

If the three LS assumptions hold, then

- The exact (finite sample) sampling distribution of $\hat{\beta}_1$ has:

$$E(\hat{\beta}_1) = \beta_1 \quad \text{(that is, } \hat{\beta}_1 \text{ is unbiased)}$$

$$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}((X_i - \mu_X)u_i)}{\sigma_X^4} \propto \frac{1}{n}$$

- Other than its mean and variance, the exact distribution of 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

$$\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim N(0, 1)$$

- This parallels the sampling distribution on $\bar{Y}$.

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;
- Also, we will cover some loose ends about regression:

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;
- Also, we will cover some loose ends about regression:
  - Regression when X is binary (0/1);

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;
- Also, we will cover some loose ends about regression:
  - Regression when X is binary (0/1);
  - Heteroskedasticity and homoskedasticity (this is new)

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;
- Also, we will cover some loose ends about regression:
  - Regression when X is binary (0/1);
  - Heteroskedasticity and homoskedasticity (this is new)
  - Efficiency of the OLS estimator (also new);

Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (SW Chapter 5)

- Now that we have the sampling distribution of OLS estimator, we are ready to perform hypothesis tests about β_1 and to construct confidence intervals about β_1;
- Also, we will cover some loose ends about regression:
  - Regression when X is binary (0/1);
  - Heteroskedasticity and homoskedasticity (this is new)
  - Efficiency of the OLS estimator (also new);
  - Use of the t-statistic in hypothesis testing (new but not surprising)

Hypothesis Testing and the Standard Error of $\hat{\beta}_1$

Objective

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:

$$H0 : \beta_1 = \beta_{1,0} \quad vs. \quad H_1 : \beta_1 \neq \beta_{1,0}$$

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

- Null hypothesis and one-sided alternative:

$$H0 : \beta_1 = \beta_{1,0} \quad vs. \quad H_1 : \beta_1 < (>)\beta_{1,0}$$

General approach to testing

Construct t-statistic, and compute p-value (or compare to N(0,1) critical value)

$$t = \frac{\text{Estimator - Hypothesized value}}{\text{Standard Error of the estimator}}$$

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

General approach to testing

Construct t-statistic, and compute p-value (or compare to N(0,1) critical value)

$$t = \frac{\text{Estimator - Hypothesized value}}{\text{Standard Error of the estimator}}$$

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

$$t = \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

General approach to testing

Construct t-statistic, and compute p-value (or compare to N(0,1) critical value)

$$t = \frac{\text{Estimator - Hypothesized value}}{\text{Standard Error of the estimator}}$$

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

$$t = \frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$$

- 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):

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}\left[(X_i - \mu_X)u_i\right]}{n\sigma_x^4} = \frac{\sigma_\nu^2}{n\sigma_X^4}, \text{ where } \nu_i = (X_i - \mu_X)u_i$$

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

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

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}\left[(X_i - \mu_X)u_i\right]}{n\sigma_x^4} = \frac{\sigma_\nu^2}{n\sigma_X^4}, \text{ where } \nu_i = (X_i - \mu_X)u_i$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_ν^2 and σ_X^4 by estimators constructed from the data:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{estimator of } \sigma_\nu^2}{(\text{estimator of } \sigma_X^4)}$$

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

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

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}\big[(X_i - \mu_X)u_i\big]}{n\sigma_x^4} = \frac{\sigma_\nu^2}{n\sigma_X^4}, \text{ where } \nu_i = (X_i - \mu_X)u_i$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_ν^2 and σ_X^4 by estimators constructed from the data:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{estimator of } \sigma_\nu^2}{(\text{estimator of } \sigma_X^4)} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\nu}_i}{\frac{1}{n}\sum_i^n (X_i - \bar{X})^4}$$

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

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

$$\text{var}(\hat{\beta}_1) = \frac{\text{var}\big[(X_i - \mu_X)u_i\big]}{n\sigma_x^4} = \frac{\sigma_\nu^2}{n\sigma_X^4}, \text{ where } \nu_i = (X_i - \mu_X)u_i$$

The estimator of the variance of $\hat{\beta}_1$ replaces the unknown population values of σ_ν^2 and σ_X^4 by estimators constructed from the data:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\text{estimator of } \sigma_\nu^2}{(\text{estimator of } \sigma_X^4)} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\nu}_i}{\frac{1}{n}\sum_i^n (X_i - \bar{X})^4}$$

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

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

Standard Error

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{\nu}_i}{\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^4}}$$

Remarks

OK, this is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates var(X).

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

Standard Error

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{\nu}_i}{\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^4}}$$

Remarks

OK, this is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates var(X).
- $SE(\hat{\beta}_1)$ is computed by regression software

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

Standard Error

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}_{\hat{\beta}_1}} = \sqrt{\frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat{\nu}_i}{\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^4}}$$

Remarks

OK, this is a bit nasty, but:

- It is less complicated than it seems. The numerator estimates var(v), the denominator estimates var(X).
- $SE(\hat{\beta}_1)$ is computed by regression software
- R has memorized this formula so you don't need to.

Summary: To test: $H0 : \beta_1 = \beta_{1,0}$ *vs.* $H_1 : \beta_1 \neq \beta_{1,0}$

- Construct the t-statistics
$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}}$$
- 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 ¡ 0.05;
  - in general, you reject at the $\alpha \times 100\%$ significance level if the p-value is ¡ α;
- This procedure relies on the large-n approximation; typically $n = 50$ is large enough for the approximation to be excellent.

Confidence Intervals for β_1

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:

$$(\hat{\beta}_1 - 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \times SE(\hat{\beta}_1))$$

Confidence Interval

$$\underset{}{\widehat{testscore}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\, str$$

$$\hat{\beta}_1 = -2.28, \quad SE(\hat{\beta}_1) = 0.4798255$$

- 95% Confidence interval

  $\left(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right)$

Confidence Interval

$$\underset{}{\widehat{testscore}} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\, str$$

$$\hat{\beta}_1 = -2.28, \quad SE(\hat{\beta}_1) = 0.4798255$$

- 95% Confidence interval

  $\left(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right)$

Confidence Interval

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\ str$$

$$\hat{\beta}_1 = -2.28, \quad SE(\hat{\beta}_1) = 0.4798255$$

- 95% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right) = \left(-2.2798081 \pm 1.96 \times 0.4798255\right) = (-3.2202662, -1.3393501)$$

- 90% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.64 \times SE(\hat{\beta}_1)\right)$$

Confidence Interval

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\ str$$

$$\hat{\beta}_1 = -2.28, \quad SE(\hat{\beta}_1) = 0.4798255$$

- 95% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right) = \left(-2.2798081 \pm 1.96 \times 0.4798255\right) = (-3.2202662, -1.3393501)$$

- 90% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.64 \times SE(\hat{\beta}_1)\right)$$

Confidence Interval

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\ str$$

$$\hat{\beta}_1 = -2.28, \quad SE(\hat{\beta}_1) = 0.4798255$$

- 95% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\right) = \left(-2.2798081 \pm 1.96 \times 0.4798255\right) = (-3.2202662, -1.3393501)$$

- 90% Confidence interval

$$\left(\hat{\beta}_1 \pm 1.64 \times SE(\hat{\beta}_1)\right) = (-2.2798081 \pm 1.64 \times 0.4798255) = (-3.2202662, -1.4928942)$$

Hypothesis Testing

We test whether str has any effect on testscore

$$H0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28} \ str$$

- The t-statistic is

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -4.7513271$$

Hypothesis Testing

We test whether str has any effect on testscore

$$H0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28} \ str$$

- The t-statistic is

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -4.7513271$$

- Since $|t| > 1.64$, we **reject** the null hypothesis at 10%;

Hypothesis Testing

We test whether str has any effect on testscore

$$H0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28} \ str$$

- The t-statistic is

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -4.7513271$$

- Since $|t| > 1.64$, we **reject** the null hypothesis at 10%;
- Since $|t| > 1.96$, we **reject** the null hypothesis at 5%;

Hypothesis Testing

We test whether str has any effect on testscore

$$H0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28}\ str$$

- The t-statistic is

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -4.7513271$$

- Since $|t| > 1.64$, we **reject** the null hypothesis at 10%;
- Since $|t| > 1.96$, we **reject** the null hypothesis at 5%;
- Can we reject at 1.35%? Check and see whether the p-value is greater than 0.0135 ....

Hypothesis Testing

We test whether str has any effect on testscore

$$H0 : \beta_1 = 0 \ vs. \ H_1 : \beta_1 \neq 0$$

$$\underset{}{testscore} = \underset{(9.47)}{698.93} - \underset{(0.48)}{2.28} \ str$$

- The t-statistic is
$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -4.7513271$$
- Since $|t| > 1.64$, we **reject** the null hypothesis at 10%;
- Since $|t| > 1.96$, we **reject** the null hypothesis at 5%;
- Can we reject at 1.35%? Check and see whether the p-value is greater than 0.0135 ....
- The p-value is $1.73e-05$, thus we **reject** at any significance level.

Hypothesis Testing

p-value

- The p-value reported by R is the p-value for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$;

Hypothesis Testing

p-value

- The p-value reported by R is the p-value for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$;
- If your test is different, you can't use this p-value;

Hypothesis Testing

p-value

- The p-value reported by R is the p-value for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$;
- If your test is different, you can't use this p-value;
- However, the test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ is the most common and has a special name: **significance test**

Hypothesis Testing

p-value

- The p-value reported by R is the p-value for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$;
- If your test is different, you can't use this p-value;
- However, the test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ is the most common and has a special name: **significance test**
- If you **can't reject** the null hypothesis we say that the coefficient β_1 is **not** statistically significant.

Omitted Variable Bias

- 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:

The two conditions for omitted variable bias

1. Z is a determinant of Y (i.e. Z is part of u)
2. Z is correlated with the regressor X (i.e. $\text{corr}(Z, X) \neq 0$)

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

Omitted variable bias, ctd.

Let's go back to the class size example:

$$testscore = \beta_0 + \beta_1 str + \underbrace{u}_{\beta_2 Z + \epsilon}$$

- English language ability (whether the student has English as a second language) plausibly affects standardized test scores: Z is a determinant of Y.
- 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, that is, $\hat{\beta}_1 \xrightarrow{p} \beta_1 + bias$

- What is the direction of this bias? (That is, what is the sign of *bias*?)
- What does common sense suggest? If common sense fails you, there is a formula...

California Schools Data

- The California School Dataset (`CASchool`) has data on the fraction of english learning in a district
- The variable is *english*

```
summary(CASchools$english)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       2       9      16      23      86
```

- Is this variable correlated, at least in the sample, with *str* and *testscore*?

Omitted variable bias, ctd.

A formula for omitted variable bias: recall the equation,

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

Under the assumption that $E[u_i|X_i] = 0$,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i\right] = E\left[(X_i - \bar{X})u_i\right] = \text{cov}(X_i, u_i) = 0$$

and, thus, there is not bias because:

$$E\left[\hat{\beta}_1\right] = \beta_1$$

But what if if $cov(X_i, u_i) \neq 0$?

Omitted variable bias, ctd.

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

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\frac{n}{n-1}s_X^2} \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2}$$

- If LSA #1 is correct, $\sigma_{Xu} = 0$ and
$$\hat{\beta}_1 \xrightarrow{p} \beta_1$$
- If LSA #1 is incorrect, $\sigma_{Xu} \neq 0$ and
$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_{Xu}}{\sigma_X^2}$$

Omitted variable bias, ctd.

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\frac{\sigma_{Xu}}{\sigma_X^2}}_{\text{bias}}$$

- *Since*

  $$u_i = \beta_2 Z_i + \epsilon_i$$

  the formula simplifies to

  $$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \underbrace{\beta_2 \frac{\sigma_{XZ}}{\sigma_X^2}}_{\text{bias}}$$

Omitted variable bias, ctd.

Consider the class size example. It is very likely that:

- $\beta_2 < 0$ - as it is reasonable to assume that districts with more english learners have lower testscore (the sample analysis also seems to suggest so)
- $\sigma_{XZ} > 0$ - the covariance between *str* and *english* is probably positive (the sample analysis also seems to suggest so)
- Thus, the bias is probably negative

$$\underset{(-)}{\beta_2} \frac{\underset{(+)}{\sigma_{XZ}}}{\sigma_X^2} < 0$$

in which case we say that $\hat{\beta}_1$ is *downward biased,* that is, it is smaller than the "true" β_1.

Omitted variable bias, ctd.

The two conditions for the omitted variable bias can be expressed in terms of β_2 and σ_{XZ}

1. Z is a determinant of Y
   - $\implies \beta_2 \neq 0$
2. Z is correlated with the regressor X
   - $\implies \sigma_{XZ} \neq 0$

Three ways to overcome omitted variable bias

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

The Population Multiple Regression Model

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

- Y is the dependent variable
  - $X_1, X_2, \ldots, X_n$ are the k independent variables (regressors)
- $(Y_i, X_{1i}, X_{2i}, \ldots, X_{ki})$ denote the ith observation on $Y_i, X_1, X_2, \ldots, X_k$
- β_0 = unknown population intercept
- β_1 = effect on Y of a change in X_1, holding $X_2 \ldots . X_k$ constant
- β_2 = effect on Y of a change in X_2, holding $X_1, X_3, \ldots, X_k$ constant
- $\vdots$
- β_k = effect on Y of a change in X_k, holding $X_1 \ldots . X_{k-1}$ constant
- u_i= the regression error (omitted factors)
  - Satisfies $E[u_i | X_1, \ldots, X_k] = 0$

Interpretation of coefficients in multiple regression

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

Consider changing X1 by DX1 while holding X2 constant:

- Population regression line before the change:

$$E[Y_i | X_{1i} = x_1, \ldots, X_{ki} = x_k] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- Population regression line, after the change: ?

$$E[Y_i | X_{1i} = x_1 + \Delta x, \ldots, X_{ki} = x_k] = \beta_0 + \beta_1 (x_1 + \Delta x) + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\underbrace{E[Y_i | X_{1i} = x_1 + \Delta x, \ldots, X_{ki} = x_k] - E[Y_i | X_{1i} = x_1, \ldots, X_{ki} = x_k]}_{\text{effect of increasing } X_1 \text{ by } \Delta x \text{ unit, holding } X_2, \ldots, X_k \text{ constant}} = \beta_1 \Delta x$$

The OLS Estimator in Multiple Regression

Assume for the moment that $k = 2$, that is, there are two regressors

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

- In this case the OLS estimator solves

$$\min_{\beta_0, \beta_1, \beta_2} \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) \right)^2$$

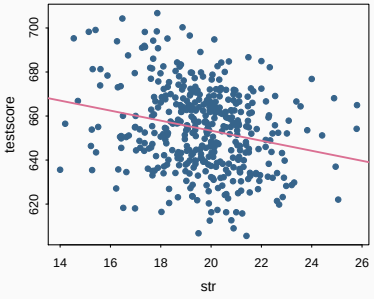
- The OLS estimator minimizes the average squared difference between the actual values of Y_i the prediction (predicted value) based on the estimated line.
- Generalization of the case with one regressors

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_{1i}) \right)^2$$

Graphic intuition

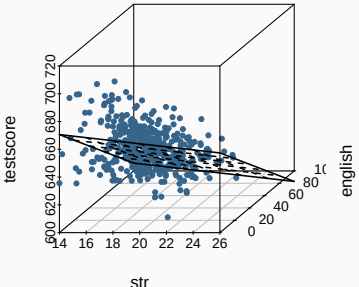
$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i})^2$$

Fits a line through points in $\mathbb{R}^2$



$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2$$

Fits a plane through points in $\mathbb{R}^3$



Matrix notation

The multivariate linear model

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

can also be written in matrix form

$$Y = X\beta + u$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

OLS in Matrix Form

- Using matrix notation, the minimization of the residuals sum of squares can be compactly rewritten as

$$\min_{\beta}(Y - X\beta)'(Y - X\beta)$$

- The first order conditions are

$$X'(Y - X\beta) = 0 \implies \underbrace{X'X}_{(k+1)\times(k+1)} \underbrace{\beta}_{(k+1)\times 1} = \underbrace{X'Y}_{(k+1)\times 1}$$

  which is a system of linear equations (recall $Ax = b$, where $A = X'X$, $x = \beta$, and $b = X'Y$)

- From which we obtain the OLS estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

  provided $(X'X)^{-1}$ is invertible (determinant $\neq 0$)

Example: the California test score data

```
lm(testscore ~ str, data = CASchools)


##
## Call:
## lm(formula = testscore ~ str, data = CASchools)
##
## Coefficients:
## (Intercept)          str
##      698.93        -2.28
```

```
lm(testscore ~ str + english, data = CASchools)


##
## Call:
## lm(formula = testscore ~ str + english, data = CASchools)
##
## Coefficients:
## (Intercept)          str      english
##      686.03        -1.10        -0.65
```

- What happens to the coefficient on *str*?

Measures of Fit for Multiple Regression

- 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 $X_1, \ldots, X_k$?
- $\bar{R}^2$= "adjusted R2" $= R^2$ with a degrees-of-freedom correction that adjusts for estimation uncertainty; $\bar{R}^2 < R^2$

SER and RMSE

As in regression with a single regressor, the SER and the RMSE are measures of the spread of the Ys 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

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}}) = \sum_{i=1}^{n} \hat{u}_i^2$$

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

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

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

The $\bar{R}^2$ (the "adjusted R2") corrects this problem by "penalizing" you for including another regressor – the does not necessarily increase when you add another regressor.

$$\begin{aligned}\bar{R}^2 &= 1 - \left(\frac{n-1}{n-k-1}\right)\frac{SSR}{TSS} \\ &= R^2 - \left(\frac{k}{n-k-1}\right)\frac{SSR}{TSS}\end{aligned}$$

?

Note that $\bar{R}^2 < R^2$, however if n is large

$$\left(\frac{k}{n-k-1}\right) \to 0,$$

and the two will be very close.

California Example

- Regression of *testscore* against *str*

  [1] "

  beginequation*=698.93- 2.28

  ,str,

  quad $R^2 = 0.0512$

  endequation*"

- Regression of *testscore* against *str* and *english*

  [1] "

  beginequation*=686.032- 1.101

  ,str- 0.65

  ,english,

  quad $R^2 = 0.426$

  endequation*"

What – precisely – does this tell you about the fit of univariate regression compared with the bivariate regression?

The Least Squares Assumptions for Multiple Regression

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

1. The conditional distribution of u given the X's has mean zero, that is,

   $$E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0,$$

   for all $(x_1, \ldots, x_k)$
2. $(Y_i, X_{1i}, \ldots, X_{ki})$, $i = 1, \ldots, n$, are i.i.d.
3. Large outliers are unlikely, $X_1, \ldots, X_k$ and Y have four moments

   $$E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y^4) < \infty$$

4. There is no perfect multicollinearity

Assumption #1: the conditional mean of u given the included Xs is zero.

$$E(u_i|X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$$

- 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 *(a)* belongs in the equation (so is in u) and *(b)* 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 (we will see this later)

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

- This is the same assumption as we had before for a single regressor.
- 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: Perfect multicollinearity

- **Perfect multicollinearity** is when one of the regressors is an exact linear function of the other regressors.
- If there is perfect multicollinearity, the OLS problem is not defined
- Modern computer software has ways to handle this problem. For instance, `R` drops one of the collinear variable

```
CASchools[["var0"]] <- CASchools[["str"]] * 2 + 5
## var0 is a linear transformation of str
lm(testscore ~ str + var0, data = CASchools)
```

```
##
## Call:
## lm(formula = testscore ~ str + var0, data = CASchools)
##
## Coefficients:
## (Intercept)          str         var0
##      698.93        -2.28           NA
```

Assumption #4: 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, ctd.

- 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.
- One of the consequences of imperfect multicollinearity is that the standard errors of the coefficients tend to be large. In that case, the test of the hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanatory. (*Why?*)

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 (Think of region of residence: Sicily, Lazio, Tuscany, etc.). 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. Lazio), **or**
  2. Omit the intercept
- What are the implications of (1) or (2) for the interpretation of the coefficients?

We will see this later on with an example.

Hypothesis Tests and Confidence Intervals for a Single Coefficient

- 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.
- Since
$$\frac{\hat{\beta}_1 - E[\hat{\beta}_1]}{SE(\hat{\beta}_1)}, \frac{\hat{\beta}_2 - E[\hat{\beta}_2]}{SE(\hat{\beta}_2)}, \cdots, \frac{\hat{\beta}_k - E[\hat{\beta}_k]}{SE(\hat{\beta}_k)}$$
are approximately distributed $N(0,1)$ (because, under the assumptions just given, the Central Limit Theorem applies)
- The hypothesis on β_1 (or any other coefficient) can be tested using the usual t-statistic
$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$
, and confidence intervals are constructed as
$$(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1))$$
- The standard errors of the estimated coefficients are calculated as in the single regression case (*mutatis mutandis).*

Example: The California class size data, ctd.

- The coefficient on *str* in (2) is the effect on *testscore* of a unit change in *str*, **holding constant** the percentage of English Learners in the district
- The coefficient on *str* falls by one-half (why?)
- 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 $H_0 : \beta_{str} = 0$ is

$$t = \frac{-1.10}{0.43} = -2.54,$$

  so we **reject** the hypothesis at the **5%** significance level
- We use heteroskedasticity-robust standard errors – for exactly the same reason as in the case of a single regressor.

Tests of joint hypotheses, ctd.

- $$H_0 : \beta_1 = 0, \text{ and } \beta_2 = 0$$

  $$vs.\ H_1 : \text{either } \beta_1 \neq 0 \text{ or } \beta 2 \neq 0 \text{ or both}$$

- 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 and are independently distributed (this isn't true in general – just in this example). Let t_1 and t_2 be the t-statistics:

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

The "one at time" test is:

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

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

$$
\begin{aligned}
&= \Pr_{H_0}[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] \\
&= 1 - \Pr_{H_0}[|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96] \\
&= 1 - \underbrace{\Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]}_{\text{(because } t_1 \text{ and } t_2 \text{ are independent by assumption)}} \\
&= 1 - (.95)^2 = 0.0975 = 9.75\%
\end{aligned}
$$

Which is **not** the desired 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-test (this is common practice)