Hypothesis Tests and Confident Intervals in Multiple Regression

Ming-Ching Luoh

2022.3.1.

Hypothesis Tests and Confidence Intervals for a Single Coefficient

- $\frac{\hat{\beta}_1 - \mathrm{E}(\hat{\beta}_1)}{\sqrt{\mathrm{Var}(\hat{\beta}_1)}}$ is approximately distributed $N(0,1)$ (CLT).
- Thus hypotheses on β_1 can be tested using the usual t-statistic, and confidence intervals are constructed as $(\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1))$.
- So as for $\beta_2, \cdots, \beta_k$.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are generally not independently distributed—neither are their t-statistics (more on this later).

Example: The California class size data

$$(1)\ \widehat{TestScore} = 698.9 - 2.28 \cdot STR$$
$$(10.4)\quad (0.52)$$
$$(2)\ \widehat{TestScore} = 696.0 - 1.10 \cdot STR - 0.650 \cdot PctEL$$
$$(8.7)\quad (0.43)\qquad (0.031)$$

- The coefficient on STR in (2) is the effect on *Test Score* of a unit change in STR, holding constant the percentage of English Learners in the district.
- Coefficient on STR falls by one-half.
- 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 for $\beta_{STR} = -1.10/0.43 = -2.54$, so we reject the hypothesis at the 5 % significance level.

Tests of Joint Hypotheses

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

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

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

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \text{ and } \beta_2 = 0 \\ vs.\ H_1 &: \quad \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both} \end{aligned}$$

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

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \text{ and } \beta_2 = 0 \\ vs.\ H_1 &: \quad \text{either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both} \end{aligned}$$

A **joint hypothesis** specifies a value for two or more coefficients, that is, it imposes restrictions on two or more coefficients.

- A "common sense" test is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this "common sense" approach doesn't work. The resulting test doesn't have the right significance level.

Here's why:

Calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual t-statistics.

- To simplify the calculation, suppose that $\hat{\beta}_1$ and $\hat{\beta}_2$ are independently distributed. 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 "common sense" test is

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

  What is the probability that this "common sense" test rejects H_0 when H_0 is actually true? (It should be 5%.)

Probability of incorrectly rejecting the null

$$
\begin{aligned}
&= \Pr_{H_0}(|t_1| > 1.96 \text{ and/or } |t_2| > 1.96) \\
&= \Pr_{H_0}(|t_1| > 1.96, |t_2| > 1.96) \\
&\quad + \Pr_{H_0}(|t_1| > 1.96, |t_2| \leq 1.96) \\
&\quad + \Pr_{H_0}(|t_1| \leq 1.96, |t_2| > 1.96)
\end{aligned}
$$

$$
\begin{aligned}
&= \Pr_{H_0}(|t_1| > 1.96) \times \Pr_{H_0}(|t_2| > 1.96) \\
&\quad + \Pr_{H_0}(|t_1| > 1.96) \times \Pr_{H_0}(|t_2| \leq 1.96) \\
&\quad + \Pr_{H_0}(|t_1| \leq 1.96) \times \Pr_{H_0}(|t_2| > 1.96) \\
&\quad (t_1, t_2 \text{ are independent by assumption}) \\
&= .05 \times .05 + .05 \times .95 + .95 \times .05 \\
&= .0975 = 9.75\%
\end{aligned}
$$

which is not the desired 5%.

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

- The size of the "common sense" test is not 5%.
- Its actual 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 App. 7.1). This is rarely used in practice.
- Use a different test statistic that test both β_1 and β_2 at once—the F-statistic.

The F-statistic

- The F-statistic tests all parts of a joint hypothesis at once.
- Formula for the special case of the joint hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ in a regression with two regressors.

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

  where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t_1 and t_2.
- Reject when F is "large."
- The F-statistic is large when t_1 and/or t_2 is large.
- The F-statistic corrects (in just the right way) for the correlation between t_1 and t_2.

Large-sample distribution of the F-statistic

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

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

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

- The *chi-squared* distribution with q degrees of freedom (χ^2_q) is defined to be the distribution of the sum of q independent squared standard normal random variables.
- In large samples, F-statistic is distributed as χ^2_q/q.
- **Selected large-sample critical values of χ^2_q/q**

q	5% critical value	
1	3.84	(why?)
2	3.00	
3	2.60	
4	2.37	
5	2.21	

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

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

- **Implementation in STATA**

  Use the "test" command after the regression.

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

F-test example, California class size data:

```
reg testscr str expn_stu pctel, r;

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

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

NOTE

```
test str expn_stu;
```

The test command follows the regression

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

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

There are q=2 restrictions being tested

The 5% critical value for q=2 is 3.00

Stata computes the p-value for you

The homoskedasticity-only F-statistic

To compute the homoskedasticity-only F-statistic if the error term u_i is homoskedastic:

- Use the previous formulas, but using homoskedasticity-only standard errors. Or
- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- The second method gives a simple formula.

The "restricted" and "unrestricted" regressions

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

Restricted population regression (that is, under H_0):

$$Test\ Score_i = \beta_0 + \beta_3 PctEL_i + u_i$$

Unrestricted population regression (under H_1):

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

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

- By how much must the R^2 increase for the coefficients on $Expn$ and $PctEL$ to be judged statistically significant?
- *Simple formula for the homoskedasticity-only F-statistic*

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

where

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

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

q = the number of restrictions under the null

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

Example:

Restricted regression:

$$\begin{aligned} \widehat{Test\ Score} &= 644.7 - 0.671 PctEL, R^2_{res} = 0.4149 \\ &\quad (1.0) \quad (0.032) \end{aligned}$$

Unrestricted regression:

$$\begin{aligned} \widehat{Test\ Score} &= 649.6 - 0.29 STR + 3.87 Expn \\ &\quad (15.5) \quad (0.48) \qquad (1.59) \\ &\quad - 0.656 PctEL \\ &\quad (0.032) \\ R^2_{unres} &= 0.4366, k_{unres} = 3, q = 2 \end{aligned}$$

Therefore,

$$\begin{aligned} F &= \frac{(R^2_{unres} - R^2_{res})/q}{(1 - R^2_{unres})/(n - k_{unres} - 1)} \\ &= \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = 8.01 \end{aligned}$$

The homoskedasticity-only F-statistic-summary

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

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

Summary: testing joint hypotheses

- The "common-sense" approach of rejecting if either of the t-statistics exceeds 1.96 rejects more than 5% of the time under the null (the size exceeds the desired significance level).
- The heteroskedasticity-robust F-statistic is built in to STATA ("test" command). This tests all q restrictions at once.
- For large n, F is distributed as $\chi^2_q/q (= F_{q,\infty})$.
- The homoskedasticity-only F-statistic is important historically (and thus in practice), and is intuitively appealing, but invalid when there is heteroskedasticity.

Testing Single Restrictions on Multiple Coefficients

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

Consider the null and alternative hypothesis,

$$H_0: \beta_1 = \beta_2 \; vs. \; H_1: \beta_1 \neq \beta_2$$

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

Two methods for testing single restrictions on multiple coefficients:

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

   Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression.

2. **Perform the test directly.**

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

Method 1: Rearrange the regression.

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

$$H_0: \beta_1 = \beta_2 \; vs. H_1: \beta_1 \neq \beta_2$$

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

$$\begin{aligned} Y_i &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i \\ Y_i &= \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i \end{aligned}$$

where

$$\begin{aligned} \gamma_1 &= \beta_1 - \beta_2 \\ W_i &= X_{1i} + X_{2i} \end{aligned}$$

(a) Original system:

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

$$H_0: \beta_1 = \beta_2 \; vs. H_1: \beta_1 \neq \beta_2$$

(b) Rearranged ("transformed") system:

$$\begin{aligned} Y_i &= \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i \\ \text{where } \gamma_1 &= \beta_1 - \beta_2, W_i = X_{1i} + X_{2i} \end{aligned}$$

so

$$H_0: \gamma_1 = 0 \; vs. H_1: \gamma_1 \neq 0$$

The testing problem is now a simple one: test whether $\gamma_1 = 0$ in specification (b).

Method 2: Perform the test directly

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

$$H_0: \beta_1 = \beta_2 \; vs. H_1: \beta_1 \neq \beta_2$$

Example:

$$\begin{aligned} TestScore_i &= \beta_0 + \beta_1 STR_i + \beta_2 Expn_i \\ &\quad + \beta_3 PctEL_i + u_i \end{aligned}$$

To test, using STATA, whether $\beta_1 = \beta_2$:

regress testscore str expn pctel, r

test str=expn

Model Specification for Multiple Regression

- The job of determining which variables to include in multiple regression— that is, the problem of choosing a regression specification— can be quite challenging, and no single rule applies in all situations.
- The starting point for choosing a regression specification is thinking through the possible sources of omitted variable bias.
- It is important to rely on your expert knowledge of the empirical problem and to focus on obtaining an unbiased estimate of the causal effect of interest.
- Do not rely solely on purely statistical measures of fit such as R^2 or $\bar{R}^2$.

Omitted Variable Bias in Multiple Regression

- The OLS estimators of the coefficients in multiple regression will have omitted variable bias if an omitted determinant of Y_i is correlated with at least one of the regressors.
- For example, students from affluent families often have more learning opportunities that do their less affluent peers, which could lead to better test scores.
- Moreover, if the district is a wealthy one, then the schools will tend to have larger budgets and lower student-teacher ratio.
- Omitting the students' economic background could lead to omitted variable bias in the regression of test scores on the student-teacher ratio and the percentage of English learners.

The Role of Control Variables in Multiple Regression

- A **control variable** is not the object of interest; rather it is a regreessor included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omotted variable bias.

For example:

$$\begin{aligned} \widehat{TestScore} &= 700.2 - 1.00 \cdot STR - 0.122 \cdot PctEL \\ &\quad (5.6) \quad (0.37) \qquad (0.0333) \\ &\quad -0.547 \cdot LchPct, \quad \bar{R}^2 = 0.773 \\ &\quad (0.024) \end{aligned}$$

$PctEL$ = % English Learners in the school district

$LchPct$ =% of students receiving a free/subsidized lunch

- Which variable is the variable of interest?
- Which variables are control variables? Do they have causal component? What do they control for?
- STR is the variable of interest.
- $PctEL$ probably has a direct causal effect (school is tougher if you are leaning English?) But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and $PctEL$ is correlated with those omitted causal variables.
- $PctEl$ is both a possible causal variable and a control variable.
- Same for $LchPct$.

Three interchangeable statements about what makes an effective control variable:

- An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.
- Holding constant the control variable(s), the variable of interest is “as if" randomly assigned.
- Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of Y.

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

$$\begin{aligned} \widehat{TestScore} &= 700.2 - 1.00 \cdot STR - 0.122 \cdot PctEL \\ &\quad (5.6) \quad (0.37) \qquad (0.0333) \\ &\quad -0.547 \cdot LchPct, \quad \bar{R}^2 = 0.773 \\ &\quad (0.024) \end{aligned}$$

- Does the coefficient on $LchPct$ have a causal interpretation? If so, then we should be able to boost test scores (by a lot! Do the math!) by simply eliminating the school lunch program, so that $LchPct = 0$!

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

Conditional mean independence

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

$$\mathrm{E}(u_i|X_i, W_i) = E(u_i|W_i)$$

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

Conditional mean independence, ctd.

Consider the regression model,

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

where X is the variable of interest and W is an effective control variable so that conditional mean independence $\mathrm{E}(u_i|X_i, W_i) = \mathrm{E}(u_i|W_i)$ holds. In addition, suppose that LSA #2, #3, and #4 hold. Then:

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

The math of conditional mean independence

Under conditional mean independence:

1. β_1 has a causal interpretation.

   The expected change in Y resulting from a change in X, holding W constant, is:

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

   Because under condtional mean independence,

$$\begin{aligned}
\mathrm{E}(u|X = x + \Delta x, W = w) &= \mathrm{E}(u|W = w) \\
\mathrm{E}(u|X = x, W = w) &= \mathrm{E}(u|W = w)
\end{aligned}$$

The math of conditional mean independence, ctd.

Under conditional mean independence:

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

Consider the regression model

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

where u satisfies the conditional mean indepence assumption. For convenience, suppose that $\mathrm{E}(u|W) = \gamma_0 + \gamma_2 W$, that is, $\mathrm{E}(u|W)$ is linear in W.

Thus, under conditional mean indepence,

$$\begin{aligned} E(u|X,W) &= \mathrm{E}(u|W) = \gamma_0 + \gamma_2 W \quad (1) \\ \text{Let } v &= u - \mathrm{E}(u|X,W) \quad (2) \end{aligned}$$

so that $\mathrm{E}(v|X,W) = 0$. Combining (1) and (2) yields,

$$\begin{aligned} u &= \mathrm{E}(u|X,W) + v \\ &= \gamma_0 + \gamma_2 W + v \quad (3) \end{aligned}$$

Now substitute (3) into the regression so that

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X + \beta_2 W + u && (4)\\
&= \beta_0 + \beta_1 X + \beta_2 W + \gamma_0 + \gamma_2 W + v \text{ from } (3)\\
&= (\beta_0 + \gamma_0) + \beta_1 X + (\beta_2 + \gamma_2) W + v\\
&\equiv \delta_0 + \beta_1 X + \delta_2 W + v && (5)
\end{aligned}
$$

- Because $E(v|X, W) = 0$ (from eq.(2)), equation (5) satisfies LSA#1 so that the OLS estimator of δ_0, β_1 and δ_2 in (5) are unbiased.
- Because the regressors in (4) and (5) are the same, the OLS coefficients in (4) satisfy, $E(\hat{\beta}_1) = \beta_1$ and $E(\hat{\beta}_2) = \delta_2 = \beta_2 + \gamma_2 \neq \beta_2$ in general.

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1) &= \beta_1 \\ \mathrm{E}(\hat{\beta}_2) &= \delta_2 = \beta_2 + \gamma_2 \neq \beta_2 \end{aligned}$$

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

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

Model Specification in Theory and in Pracitce

- In theory, when data are available on the omitted variable, the solution to omitted variable bias is to include the omitted variable in the regression.
- In practice, however, deciding whether to include a particular variable can be difficult and requires judgement.
- Our approach to the challenge of potential omitted variable bias is twofold.

- First, a base set of regressors should be chosen using a combination of expert judgment, economic theory, and knowledge of how the data were collected.
- The regression using this base set of regressors is referred to as a base specification, which contain the variables of primary interest and the control variables suggested by judgment and economic theory.
- Judgment and theory are rarely decisive, however, and often the variables suggested by theory are not the ones on which you have data.

- The next step is to develop a list of candidate alternative specifications, that is, alternative sets of regressors.
- If the estimates of the coefficients of interest are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification are reliable.
- If the estimates of the coefficients of interest change substantially across specifications, this often provide evidence that the original specification had omitted variable bias.

Interpreting R^2 and $\bar{R}^2$ in Practice

- An increase in the R^2 or $\bar{R}^2$ does not necessarily mean that an added variable is statistically significant.
- A high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable.
- A high R^2 or $\bar{R}^2$ does not mean there is no omitted variable bias.
- A high R^2 or $\bar{R}^2$ does not necessarily mean you have the most appropriate set of regressors, nor does a low R^2 or $\bar{R}^2$ necessarily mean you have an inappropriate set of regressors.

Analysis of the Test Score Data

Variables we would like to see in the California data set

School characteristics

- student-teacher ratio
- teacher quality
- computers (non-teaching resources) per student
- measures of curriculum design

Variables we would like to see in the California data set

Student characteristics

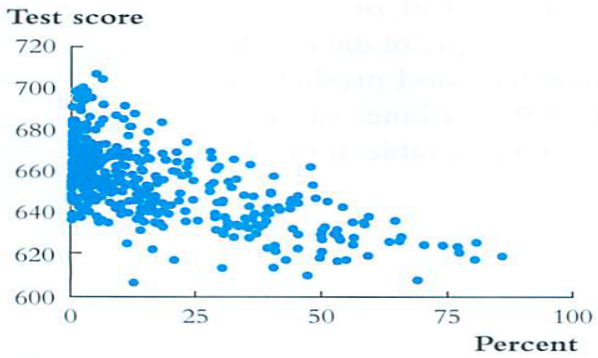
- English proficiency
- availability of extracurricular enrichment
- home learning environment
- parent's education level

Variables actually in the California class size data set

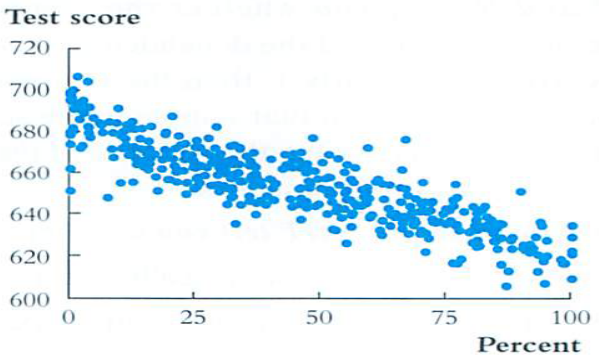
- student-teacher ratio (STR)
- percent English learners in the district (PctEL)
- percent eligible for subsidized/free lunch
- percent on public income assistance
- average district income

A look at more of the California data

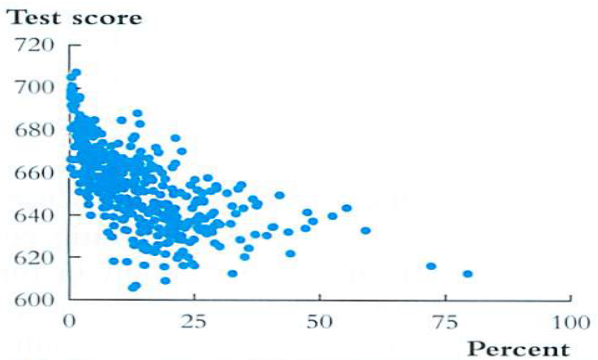
FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English learners



(b) Percentage eligible for subsidized lunch



(c) Percentage qualifying for income assistance

The correlation coefficients between the two variables are (a) -0.64 (b) -0.87 (c) -0.63. The correlation between subsidized lunch and income assistance is 0.74.

Presentation of regression results in a table

- Listing regressions in “equation" form can be cumbersome with many regressors and many regressions.
- Tables of regression results can present the key information compactly.
- Information to include:
  - variables in the regression (dependent and independent).
  - estimated coefficients.
  - standard errors.
  - results of F-tests of joint hypotheses.
  - some measure of fit (adjusted R^2).
  - number of observations.

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	−1.10	−1.00	−1.31	−1.01
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
	[−3.30, −1.26]	[−1.95, −0.25]	[−1.53, −0.47]	[−1.97, −0.64]	[−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650	−0.122	−0.488	−0.130
		(0.031)	(0.033)	(0.030)	(0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547		−0.529
			(0.024)		(0.038)
Percentage qualifying for income assistance (X_4)				−0.790	0.048
				(0.068)	(0.059)
Intercept	698.9	686.0	700.2	698.0	700.4
	(10.4)	(8.7)	(5.6)	(6.9)	(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. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

What scale should we use for the regressors?

- A practical question that arises in the regression analysis is what scale we should use for the regressors. For example, $PctEL$ or $FracEL = \frac{PctEL}{100}$ (fraction of english learners)?
- The general answer for choosing the scale of the variables is to make the regression results easy to read and to interpret.

- For example, the coefficient on $PctEL$ in Model (2) of Table 7.1 is -0.650.
- If instead the regressor had been $FracEL$, the regreesion would have had an identical R^2 and SER; however, the coefficient on $FracEL$ would have been -65.0 (i.e. -0.65×100).
- Another consideration is easy to read. For example, if a regressor is measured in dollars and has a coefficient of 0.00000356, it is easier to read if the regressor is converted to millions of dollars (10^6) and the coefficient 3.56 is reported.

Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on Y of a change in X_1, holding X_2 constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- If you can't measure the omitted variable, you still might be able to control for its effect by including a control variable.
- There is no simple recipe for deciding which variables belong in a regression - you must exercise judgment.
- One approach is to specify a base model - relying on a-priori reasoning - then explore the sensitivity of the key estimate(s) in alternative specifications.