

# Multiple Regression

## (SW Chapters 6 & 7)

**OLS estimate of the *Test Score/STR* relation:**

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

(10.4) (0.52)

Is this a credible estimate of the causal effect on test scores of a change in the student-teacher ratio?

*No*: there are omitted confounding factors (family income; whether the students are native English speakers) that bias the OLS estimator: *STR* could be “picking up” the effect of these confounding factors.

# Omitted Variable Bias

## (SW Section 6.1)

The bias in the OLS estimator that occurs as a result of an omitted factor is called *omitted variable* bias. For omitted variable bias to occur, the omitted factor “Z” must be:

1. a determinant of  $Y$ ; **and**
2. correlated with the regressor  $X$ .

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

In the test score example:

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

- Accordingly,  $\hat{\beta}_1$  is biased
- What is the direction of this bias?
  - What does common sense suggest?
  - If common sense fails you, there is a formula...

A formula for omitted variable bias: recall the equation,

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

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

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

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

Then

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

SO

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

where  $\approx$  holds with equality when  $n$  is large; specifically,

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu}, \text{ where } \rho_{Xu} = \text{corr}(X, u)$$

**Omitted variable bias formula:**  $\hat{\beta}_1 \xrightarrow{p} \beta_1 + \left( \frac{\sigma_u}{\sigma_X} \right) \rho_{Xu}$ .

If an omitted factor  $Z$  is *both*:

- (1) a determinant of  $Y$  (that is, it is contained in  $u$ ); *and*
- (2) correlated with  $X$ ,

then  $\rho_{Xu} \neq 0$  and the OLS estimator  $\hat{\beta}_1$  is biased.

The math makes precise the idea that districts with few ESL students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the ESL factor results in overstating the class size effect.

*Is this is actually going on in the CA data?*

**TABLE 5.1** Differences in Test Scores for California School Districts with Low and High Student Teacher Ratios, by the Percentage of English Learners in the District

	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	<i>n</i>	Average Test Score	<i>n</i>	Difference	<i>t</i> -statistic
All Districts	657.4	238	650.0	182	7.4	4.04
Percent of English Learners						
< 2.2%	664.1	78	665.4	27	-1.3	-0.44
2.2–8.8%	666.1	61	661.8	44	4.3	1.44
8.8–23.0%	654.6	55	649.7	50	4.9	1.64
> 23.0%	636.7	44	634.8	61	1.9	0.68

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

## Three ways to overcome omitted variable bias

1. Run a randomized controlled experiment in which treatment ( $STR$ ) is randomly assigned: then  $PctEL$  is still a determinant of  $TestScore$ , but  $PctEL$  is uncorrelated with  $STR$ . (*But this is unrealistic in practice.*)
2. Adopt the “cross tabulation” approach, with finer gradations of  $STR$  and  $PctEL$  (*But soon we will run out of data, and what about other determinants like family income and parental education?*)
3. Use a method in which the omitted variable ( $PctEL$ ) is no longer omitted: include  $PctEL$  as an additional regressor in a multiple regression.

# The Population Multiple Regression Model

## (SW Section 6.2)

Consider the case of two regressors:

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

- $X_1, X_2$  are the two *independent variables (regressors)*
- $(Y_i, X_{1i}, X_{2i})$  denote the  $i^{\text{th}}$  observation on  $Y, X_1$ , and  $X_2$ .
- $\beta_0$  = unknown population intercept
- $\beta_1$  = effect on  $Y$  of a change in  $X_1$ , holding  $X_2$  constant
- $\beta_2$  = effect on  $Y$  of a change in  $X_2$ , holding  $X_1$  constant
- $u_i$  = “error term” (omitted factors)

## *Interpretation of multiple regression coefficients*

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

Consider changing  $X_1$  by  $\Delta X_1$  while holding  $X_2$  constant:

Population regression line *before* the change:

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

Population regression line, *after* the change:

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

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

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

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

That is,

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

also,

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

and

$$\beta_0 = \text{predicted value of } Y \text{ when } X_1 = X_2 = 0.$$

# The OLS Estimator in Multiple Regression (SW Section 6.3)

With two regressors, the OLS estimator solves:

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

- The OLS estimator minimizes the average squared difference between the actual values of  $Y_i$  and the prediction (predicted value) based on the estimated line.
- This minimization problem is solved using calculus
- **The result is the OLS estimators of  $\beta_0$  and  $\beta_1$ .**

## Example: the California test score data

Regression of *TestScore* against *STR*:

$$\overline{\text{TestScore}} = 698.9 - 2.28 \times \text{STR}$$

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

$$\overline{\text{TestScore}} = 696.0 - 1.10 \times \text{STR} - 0.65 \text{PctEL}$$

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

# Multiple regression in STATA

```
reg testscr str pctel, robust;
```

Regression with robust standard errors

```
Number of obs =      420  
F( 2, 417) =    223.82  
Prob > F      =    0.0000  
R-squared     =    0.4264  
Root MSE     =    14.464
```

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

$$\widehat{TestScore} = 696.0 - 1.10 \times STR - 0.65 PctEL$$

What are the sampling distribution of  $\hat{\beta}_1$  and  $\hat{\beta}_2$ ?

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

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

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

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

- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in  $u$ ) and (2) is correlated with an included  $X$ , then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution – *if possible* – is to include the omitted variable in the regression.

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

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

**Assumption #3: finite fourth moments**

This is technical assumption is satisfied automatically by variables with a bounded domain (test scores, *PctEL*, etc.)

## Assumption #4: There is no perfect multicollinearity

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

*Example:* Suppose you accidentally include *STR* twice:

```
regress testscr str str, robust
```

```
Regression with robust standard errors
```

```
Number of obs =      420
```

```
F( 1, 418) =    19.26
```

```
Prob > F      =    0.0000
```

```
R-squared     =    0.0512
```

```
Root MSE     =    18.581
```

```
-----
```

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

```
-----
```

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

- In the previous regression,  $\beta_1$  is the effect on *TestScore* of a unit change in *STR*, holding *STR* constant (???)
- Second example: regress *TestScore* on a constant,  $D$ , and  $B$ , where:  $D_i = 1$  if  $STR \leq 20$ ,  $= 0$  otherwise;  $B_i = 1$  if  $STR > 20$ ,  $= 0$  otherwise, so  $B_i = 1 - D_i$  and there is perfect multicollinearity
- Would there be perfect multicollinearity if the intercept (constant) were somehow dropped (that is, omitted or suppressed) in the regression?
- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data

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

Under the four Least Squares Assumptions,

- The exact (finite sample) distribution of  $\hat{\beta}_1$  has mean  $\beta_1$ ,  $\text{var}(\hat{\beta}_1)$  is inversely proportional to  $n$ ; so too for  $\hat{\beta}_2$ .
- Other than its mean and variance, the exact distribution of  $\hat{\beta}_1$  is very complicated
- $\hat{\beta}_1$  is consistent:  $\hat{\beta}_1 \xrightarrow{p} \beta_1$  (law of large numbers)
- $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$  is approximately distributed  $N(0,1)$  (CLT)
- So too for  $\hat{\beta}_2, \dots, \hat{\beta}_k$

# Hypothesis Tests and Confidence Intervals for a Single Coefficient in Multiple Regression (SW Section 7.1)

- $\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$  is approximately distributed  $N(0,1)$  (CLT).
- Thus hypotheses on  $\beta_1$  can be tested using the usual  $t$ -statistic, and confidence intervals are constructed as  $\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1)\}$ .
- So too for  $\beta_2, \dots, \beta_k$ .
- $\hat{\beta}_1$  and  $\hat{\beta}_2$  are generally not independently distributed – so neither are their  $t$ -statistics (more on this later).

**Example:** The California class size data

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

(10.4) (0.52)

$$(2) \quad \overline{TestScore} = 696.0 - 1.10 \times STR - 0.650 PctEL$$

(8.7) (0.43) (0.031)

- The coefficient on  $STR$  in (2) is the effect on  $TestScores$  of a unit change in  $STR$ , holding constant the percentage of English Learners in the district
- Coefficient on  $STR$  falls by one-half
- 95% confidence interval for coefficient on  $STR$  in (2) is  $\{-1.10 \pm 1.96 \times 0.43\} = (-1.95, -0.26)$

# Tests of Joint Hypotheses

## (SW Section 7.2)

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

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

The null hypothesis that “school resources don’t matter,” and the alternative that they do, corresponds to:

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

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

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

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

vs.  $H_1$ : ***either***  $\beta_1 \neq 0$  ***or***  $\beta_2 \neq 0$  ***or both***

A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction 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:* Calculation of 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:

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

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

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

$$= \Pr_{H_0} [|t_1| > 1.96, |t_2| > 1.96]$$

$$+ \Pr_{H_0} [|t_1| > 1.96, |t_2| \leq 1.96]$$

$$+ \Pr_{H_0} [|t_1| \leq 1.96, |t_2| > 1.96] \quad (\text{disjoint events})$$

$$= \Pr_{H_0} [|t_1| > 1.96] \times \Pr_{H_0} [|t_2| > 1.96]$$

$$+ \Pr_{H_0} [|t_1| > 1.96] \times \Pr_{H_0} [|t_2| \leq 1.96]$$

$$+ \Pr_{H_0} [|t_1| \leq 1.96] \times \Pr_{H_0} [|t_2| > 1.96]$$

$(t_1, t_2 \text{ are independent by assumption})$

$$= .05 \times .05 + .05 \times .95 + .95 \times .05$$

$$= .0975 = 9.75\% - \text{which is } \mathbf{not} \text{ the desired } 5\%!!$$

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

- The size of the “common sense” test isn’t 5%!
- Its size actually 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. 5.3)
- Use a different test statistic that test both  $\beta_1$  and  $\beta_2$  at once: the  $F$ -statistic.

## The $F$ -statistic

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

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

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

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

Reject when  $F$  is “large”

The  $F$ -statistic testing  $\beta_1$  and  $\beta_2$  (special case):

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

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

## Large-sample distribution of the $F$ -statistic

Consider special case that  $t_1$  and  $t_2$  are independent, so

$\hat{\rho}_{t_1, t_2} \xrightarrow{p} 0$ ; in large samples the formula becomes

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

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

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

In large samples,  $F$  is distributed as  $\chi_q^2/q$ .

### **Selected large-sample critical values of $\chi_q^2/q$**

$q$	<u>5% critical value</u>	
1	3.84	( <i>why?</i> )
2	3.00	(the case $q=2$ above)
3	2.60	
4	2.37	
5	2.21	

*p-value using the F-statistic:*

*p-value* = tail probability of the  $\chi^2/q$  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
```

testscr	Coef.	Robust 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
```

*There are q=2 restrictions being tested*

```
F(   2,   416) =    5.43   The 5% critical value for q=2 is 3.00
Prob > F =    0.0047   Stata computes the p-value for you
```

*Two (related) loose ends:*

1. Homoskedasticity-only versions of the  $F$ -statistic
2. The “ $F$ ” distribution

## **The homoskedasticity-only (“rule-of-thumb”) $F$ -statistic**

To compute the homoskedasticity-only  $F$ -statistic:

- 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  $H_0$ ):

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

Unrestricted population regression (under  $H_1$ ):

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

- 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:*

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

where:

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

$R_{unrestricted}^2$  = 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:

$$\overline{\text{TestScore}} = 644.7 - 0.671 \text{PctEL}, \quad R^2_{\text{restricted}} = 0.4149$$

(1.0) (0.032)

Unrestricted regression:

$$\overline{\text{TestScore}} = 649.6 - 0.29 \text{STR} + 3.87 \text{Expn} - 0.656 \text{PctEL}$$

(15.5) (0.48)      (1.59)      (0.032)

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

so:

$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}}) / q}{(1 - R^2_{\text{unrestricted}}) / (n - k_{\text{unrestricted}} - 1)}$$
$$= \frac{(.4366 - .4149) / 2}{(1 - .4366) / (420 - 3 - 1)} = 8.01$$

## *The homoskedasticity-only $F$ -statistic*

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

- 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  $\chi^2_q/q$ .
- But if the errors are heteroskedastic, the large-sample distribution is a mess and is not  $\chi^2_q/q$

# The $F$ distribution

If:

1.  $u_1, \dots, u_n$  are normally distributed; and
2.  $X_i$  is distributed independently of  $u_i$  (so in particular  $u_i$  is homoskedastic)

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

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

- The  $F$  distribution is tabulated many places
- When  $n$  gets large the  $F_{q,n-k-1}$  distribution asymptotes to the  $\chi_q^2/q$  distribution:

**$F_{q,\infty}$  is another name for  $\chi_q^2/q$**

- For  $q$  not too big and  $n \geq 100$ , the  $F_{q,n-k-1}$  distribution and the  $\chi_q^2/q$  distribution are essentially identical.
- Many regression packages compute  $p$ -values of  $F$ -statistics using the  $F$  distribution (which is OK if the sample size is  $< 100$ )
- You will encounter the “ $F$ -distribution” in published empirical work.

*Digression: A little history of statistics...*

- The theory of the homoskedasticity-only  $F$ -statistic and the  $F_{q,n-k-1}$  distributions rests on implausibly strong assumptions (are earnings normally distributed?)
- These statistics dates to the early 20<sup>th</sup> century, when “computer” was a job description and observations numbered in the dozens.
- The  $F$ -statistic and  $F_{q,n-k-1}$  distribution were major breakthroughs: an easily computed formula; a single set of tables that could be published once, then applied in many settings; and a precise, mathematically elegant justification.

## *A little history of statistics, ctd...*

- The strong assumptions seemed a minor price for this breakthrough.
- But with modern computers and large samples we can use the heteroskedasticity-robust  $F$ -statistic and the  $F_{q,\infty}$  distribution, which only require the four least squares assumptions.
- This historical legacy persists in modern software, in which homoskedasticity-only standard errors (and  $F$ -statistics) are the default, and in which  $p$ -values are computed using the  $F_{q,n-k-1}$  distribution.

## Summary: the homoskedasticity-only (“rule of thumb”) $F$ -statistic and the $F$ distribution

- These are justified only under very strong conditions – stronger than are realistic in practice.
- Yet, they are widely used.
- *You* should use the heteroskedasticity-robust  $F$ -statistic, with  $\chi_q^2/q$  (that is,  $F_{q,\infty}$ ) critical values.
- For  $n \geq 100$ , the  $F$ -distribution essentially is the  $\chi_q^2/q$  distribution.
- For small  $n$ , the  $F$  distribution isn’t necessarily a “better” approximation to the sampling distribution of the  $F$ -statistic – only if the strong conditions are true.

## 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  $n$  large,  $F$  is distributed as  $\chi_q^2/q (= F_{q,\infty})$
- 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 (SW Section 7.3)

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

Consider the null and alternative hypothesis,

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

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

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 (“transform”) the regression*

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

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

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

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

or

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

where

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

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

(a) *Original system:*

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

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

(b) *Rearranged (“transformed”) system:*

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

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

so

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

The testing problem is now a simple one:

test whether  $\gamma_1 = 0$  in specification (b).

*Method 2: Perform the test directly*

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

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

*Example:*

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{STR}_i + \beta_2 \text{Expn}_i + \beta_3 \text{PctEL}_i + u_i$$

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

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

# Confidence Sets for Multiple Coefficients (SW Section 7.4)

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

What is a *joint* confidence set for  $\beta_1$  and  $\beta_2$ ?

A 95% *confidence set* is:

- A set-valued function of the data that contains the true parameter(s) in 95% of hypothetical repeated samples.
- The set of parameter values that cannot be rejected at the 5% significance level when taken as the null hypothesis.

The *coverage rate* of a confidence set is the probability that the confidence set contains the true parameter values

A “common sense” confidence set is the union of the 95% confidence intervals for  $\beta_1$  and  $\beta_2$ , that is, the rectangle:

$$\{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_2 \pm 1.96 \times SE(\hat{\beta}_2)\}$$

- What is the coverage rate of this confidence set?
- Is its coverage rate equal the desired confidence level of 95%?

Coverage rate of “common sense” confidence set:

$$\begin{aligned} & \Pr[(\beta_1, \beta_2) \in \{\hat{\beta}_1 \pm 1.96 \times SE(\hat{\beta}_1), \hat{\beta}_2 \pm 1.96 \times SE(\hat{\beta}_2)\}] \\ &= \Pr[\hat{\beta}_1 - 1.96SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1.96SE(\hat{\beta}_1), \\ & \quad \hat{\beta}_2 - 1.96SE(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + 1.96SE(\hat{\beta}_2)] \\ &= \Pr[-1.96 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 1.96, -1.96 < \frac{\hat{\beta}_2 - \beta_2}{SE(\hat{\beta}_2)} < 1.96] \\ &= \Pr[|t_1| < 1.96 \text{ and } |t_2| < 1.96] \\ &= 1 - \Pr[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] \leq 95\% ! \end{aligned}$$

Why?

*This confidence set “inverts” a test for which the size doesn’t equal the significance level!*

*Recall:* the probability of incorrectly rejecting the null

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

$$= \Pr_{H_0} [|t_1| > 1.96, |t_2| > 1.96]$$

$$+ \Pr_{H_0} [|t_1| > 1.96, |t_2| \leq 1.96]$$

$$+ \Pr_{H_0} [|t_1| \leq 1.96, |t_2| > 1.96] \quad (\text{disjoint events})$$

$$= \Pr_{H_0} [|t_1| > 1.96] \times \Pr_{H_0} [|t_2| > 1.96]$$

$$+ \Pr_{H_0} [|t_1| > 1.96] \times \Pr_{H_0} [|t_2| \leq 1.96]$$

$$+ \Pr_{H_0} [|t_1| \leq 1.96] \times \Pr_{H_0} [|t_2| > 1.96]$$

(if  $t_1, t_2$  are independent)

$$= .05 \times .05 + .05 \times .95 + .95 \times .05$$

$$= .0975 = 9.75\% - \text{which is } \mathbf{not} \text{ the desired } 5\%!!$$

Instead, use the acceptance region of a test that has size equal to its significance level (“invert” a valid test):

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

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

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

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

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

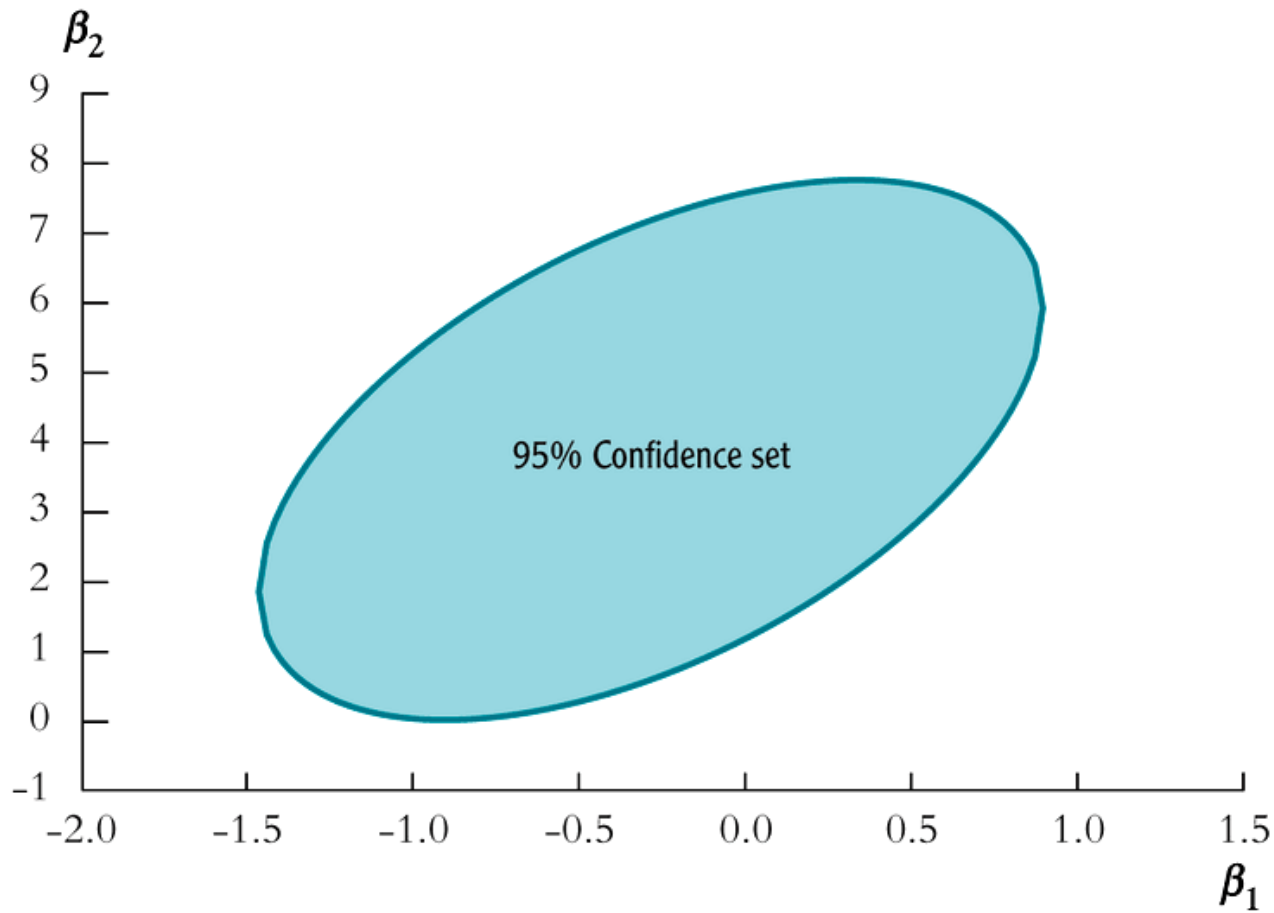
Now

$$\begin{aligned} F &= \frac{1}{2(1 - \hat{\rho}_{t_1, t_2}^2)} \times [t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2] \\ &= \frac{1}{2(1 - \hat{\rho}_{t_1, t_2}^2)} \times \\ &\quad \left[ \left( \frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)} \right)^2 + \left( \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right)^2 + 2\hat{\rho}_{t_1, t_2} \left( \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)} \right) \left( \frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)} \right) \right] \end{aligned}$$

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

# Confidence set based on inverting the $F$ -statistic

**FIGURE 5.1** 95% Confidence Set for  $\beta_1$  and  $\beta_2$



The 95% confidence set for  $\beta_1$  and  $\beta_2$  is an ellipse. The ellipse contains the pairs of values of  $\beta_1$  and  $\beta_2$  that cannot be rejected using the  $F$ -statistic at the 5% significance level.

## The $R^2$ , $SER$ , and $\bar{R}^2$ for Multiple Regression (SW Section 6.4)

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

As in regression with a single regressor, the  $SER$  (and the  $RMSE$ ) is a measure of the spread of the  $Y$ 's around the regression line:

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

The  $R^2$  is the fraction of the variance explained:

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

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

$\sum_{i=1}^n (Y_i - \bar{Y})^2$  – just as for regression with one regressor.

- The  $R^2$  always increases when you add another regressor – a bit of a problem for a measure of “fit”
- The  $\bar{R}^2$  corrects this problem by “penalizing” you for including another regressor:

$$\bar{R}^2 = 1 - \left( \frac{n-1}{n-k-1} \right) \frac{SSR}{TSS} \quad \text{so } \bar{R}^2 < R^2$$

## *How to interpret the $R^2$ and $\bar{R}^2$ ?*

- A high  $R^2$  (or  $\bar{R}^2$ ) means that the regressors explain the variation in  $Y$ .
- A high  $R^2$  (or  $\bar{R}^2$ ) does *not* mean that you have eliminated omitted variable bias.
- A high  $R^2$  (or  $\bar{R}^2$ ) does *not* mean that you have an unbiased estimator of a causal effect ( $\beta_1$ ).
- A high  $R^2$  (or  $\bar{R}^2$ ) does *not* mean that the included variables are statistically significant – this must be determined using hypotheses tests.

## ***Example: A Closer Look at the Test Score Data*** **(SW Section 7.6)**

*A general approach to variable selection and model specification:*

- Specify a “base” or “benchmark” model.
- Specify a range of plausible alternative models, which include additional candidate variables.
- Does a candidate variable change the coefficient of interest ( $\beta_1$ )?
- Is a candidate variable statistically significant?
- Use judgment, not a mechanical recipe...

*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...

**Student characteristics:**

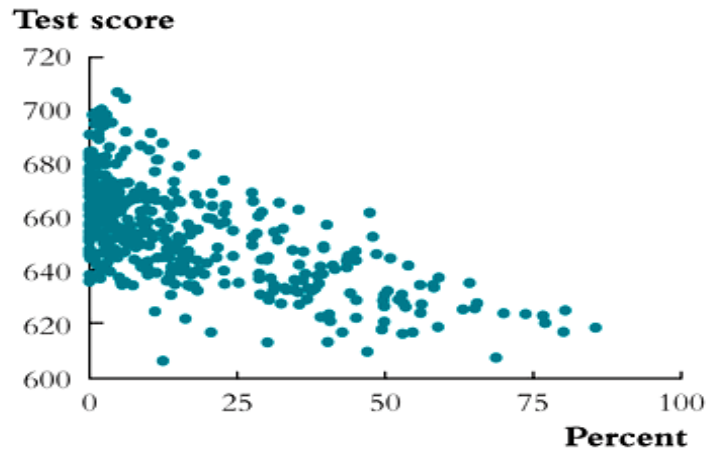
- English proficiency
- availability of extracurricular enrichment
- home learning environment
- parent's education level...

*Variables actually in the California class size data set:*

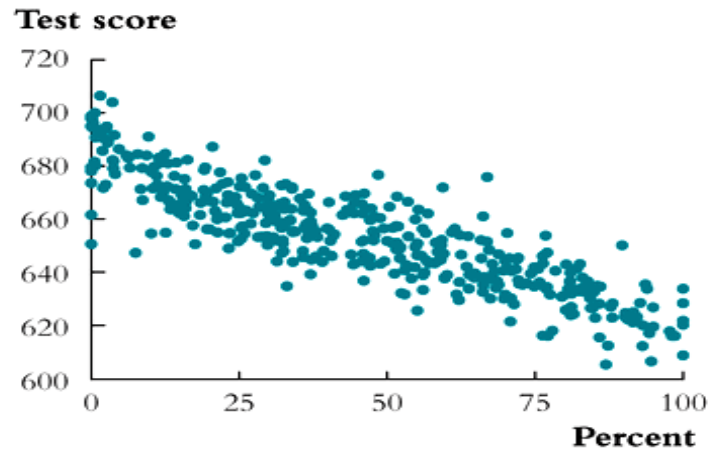
- student-teacher ratio (*STR*)
- percent English learners in the district (*PctEL*)
- percent eligible for subsidized/free lunch
- percent on public income assistance
- average district income

# A look at more of the California data

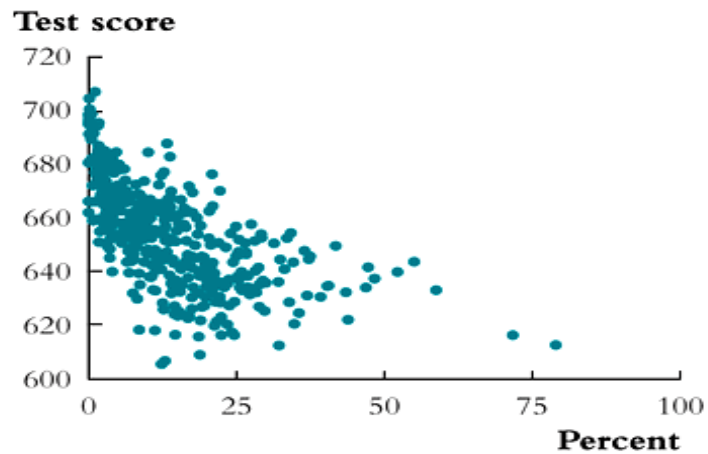
**FIGURE 5.2** Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percent of English language learners



(b) Percent qualifying for reduced price lunch



(c) Percent qualifying for income assistance

The scatterplots show a negative relationship between test scores and (a) the percentage of English learners (correlation =  $-0.64$ ), (b) the percentage of students qualifying for a subsidized lunch (correlation =  $-0.87$ ); and (c) the percentage qualifying for income assistance (correlation =  $-0.63$ ).

## *Digression: 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 pertinent joint hypotheses
  - some measure of fit
  - number of observations

**TABLE 5.2 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.**

<b>Regressor</b>	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>	<b>(4)</b>	<b>(5)</b>
Student-teacher ratio ( $X_1$ )	-2.28** (0.52)	-1.10* (0.43)	-1.00** (0.27)	-1.31** (0.34)	-1.01** (0.27)
Percent English learners ( $X_2$ )		-0.650** (0.031)	-0.122** (0.033)	-0.488** (0.030)	-0.130** (0.036)
Percent eligible for subsidized lunch ( $X_3$ )			-0.547** (0.024)		-0.529** (0.038)
Percent on public income assistance ( $X_4$ )				-0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
<b>Summary Statistics</b>					
<i>SER</i>	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
<i>n</i>	420.0	420.0	420.0	420.0	420.0

These regressions were estimated using the data on K-8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the \*5% level or \*\*1% significance level using a two-sided test.

## Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on  $Y$  of a change in  $X_1$ , holding  $X_2$  constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- 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.