Introductory Applied Econometrics Analysis using Stata

November 14 – 18, 2016

Dushanbe, Tajikistan

Allen Park and Jarilkasin Ilyasov

Linear Regression with Multiple Regressors

Outline

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

Based on Chapter 6 and 7. Stock and Watson. “Introduction to Econometrics” 3rd Edition.

Linear Regression with Multiple Regressors

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.

Linear Regression with Multiple Regressors

Omitted variable bias (cont.)

- 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*); **and**
  - (2) *Z* is correlated with the regressor *X* (*i.e.* corr(*Z*,*X*) ≠ 0)
- ***Both*** *conditions must hold for the omission of Z to result in omitted variable bias*.

Linear Regression with Multiple Regressors

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. 1st least square assumption (***E*(*u*|*X* = *x*) = 0)** is violated. What is the direction of this bias?
  - *What does common sense suggest?*

If common sense fails you, there is a formula…

Linear Regression with Multiple Regressors

The omitted variable bias formula:

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

- If an omitted variable *Z* is ***both***:
  - (1) a determinant of *Y* (that is, it is contained in *u*); ***and***
  - (2) correlated with *X*, then $\rho_{Xu} \neq 0$ and the OLS estimator $\hat{\beta}_1$ is biased and is not consistent.

ρ_{Xu} = corr(X_i, u_i) = correlation between the X_i and u_i

Linear Regression with Multiple Regressors

- For example, districts with few English second language (ESL) students (1) do better on standardized tests and (2) have smaller classes (bigger budgets), so ignoring the effect of having many ESL students factor would result in overstating the class size effect.

  Is this is actually going on in the data?

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

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

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

Linear Regression with Multiple Regressors

Causality and regression analysis

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

Linear Regression with Multiple Regressors

This raises a deeper question: how do we define β_1?

β_1 was defined as the slope of population regression line

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

Linear Regression with Multiple Regressors

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

- 1. We want to estimate the slope of a line through a scatterplot as a simple summary of the data to which we attach no substantive meaning.
  - *This can be useful at times, but isn't very interesting intellectually and isn't what this course is about.*
- 2. We want to make forecasts, or predictions, of the value of *Y* for an entity not in the data set, for which we know the value of *X*.
  - *Forecasting is an important job for economists, and excellent forecasts are possible using regression methods without needing to know causal effects. We will return to forecasting later in the course.*

Linear Regression with Multiple Regressors

- 3. We want to estimate the causal effect on *Y* of a change in *X*.
  - *This is why we are interested in the class size effect. Suppose the school board decided to cut class size by 2 students per class. What would be the effect on test scores? This is a causal question (what is the causal effect on test scores of STR?) so we need to estimate this causal effect.*

Linear Regression with Multiple Regressors

What, precisely, is a causal effect?

- "Causality" is a complex concept!
- Taking a practical approach to defining causality:
  - **A causal effect is defined to be the effect measured in an ideal randomized controlled experiment.**

Linear Regression with Multiple Regressors

Ideal Randomized Controlled Experiment

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

Linear Regression with Multiple Regressors

Three ways to overcome omitted variable bias

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

Linear Regression with Multiple Regressors

The Population Multiple Regression Model

Consider the case of two regressors:

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

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

Linear Regression with Multiple Regressors

Interpretation of coefficients in multiple regression

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

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

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

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

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

Linear Regression with Multiple Regressors

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

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

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

Linear Regression with Multiple Regressors

The OLS Estimator in Multiple Regression

With two regressors, the OLS estimator solves:

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

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

Linear Regression with Multiple Regressors

Multiple regression in STATA

```
reg testscr str pctel, robust;

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

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

Linear Regression with Multiple Regressors

Measures of Fit for Multiple Regression

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

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

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

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

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

Linear Regression with Multiple Regressors

SER and RMSE

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

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

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

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

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

Linear Regression with Multiple Regressors

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

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

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

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

Linear Regression with Multiple Regressors

The Least Squares Assumptions for Multiple Regression

But before we look at them, do we remember LSA for a single regression?

Linear Regression with Multiple Regressors

The Extended Least Squares Assumptions

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

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

Linear Regression with Multiple Regressors

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, $E(u_i | X_{1i} = x_1, \ldots, X_{ki} = x_k) = 0$.
2. $(X_{1i},\ldots,X_{ki}, Y_i)$, $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are unlikely: $X_1,\ldots, X_k$, and Y have four moments: $E(X_{1i}^4) < \infty, \ldots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$.
4. There is no perfect multicollinearity.

Linear Regression with Multiple Regressors

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

$$E(u|X1 = x1,\ldots, Xk = xk) = 0$$

- 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
- The best solution, if possible, is to include the omitted variable in the regression.
- A second, related solution is to include a variable that controls for the omitted variable (discussed in Ch. 7)

Linear Regression with Multiple Regressors

Assumption #2: ($X1i$,…,Xki,Yi), i =1,…,n, are i.i.d.

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

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

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

Linear Regression with Multiple Regressors

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
------------------------------------------------------------------------------
```

Linear Regression with Multiple Regressors

The Sampling Distribution of the OLS Estimator

Under the four Least Squares Assumptions,

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

Conceptually, there is nothing new here!

Linear Regression with Multiple Regressors

Multicollinearity, Perfect and Imperfect

- ***Perfect multicollinearity*** is when one of the regressors is an exact linear function of the other regressors.
  - If a variable is a fraction of another variable
  - **Dummy variable trap** → exclude one of the binary variables from the multiple regression

Linear Regression with Multiple Regressors

- ***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 or +1, that collinearity is imperfect.

Linear Regression with Multiple Regressors

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

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

Linear Regression with Multiple Regressors

Hypothesis Tests and Confidence Intervals in Multiple Regression

Outline

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

Linear Regression with Multiple Regressors

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.

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

Linear Regression with Multiple Regressors

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

```
reg testscr str pctel, robust;

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

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

$$TestScore = 686.0 - 1.10STR - 0.650PctEL$$
$$(8.7) \quad (0.43) \quad (0.031)$$

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

Linear Regression with Multiple Regressors

Tests of Joint Hypotheses

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

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

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

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

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

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

Linear Regression with Multiple Regressors

H_0: $\beta_1 = 0$ ***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.
- 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%)!

Linear Regression with Multiple Regressors

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

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

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

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

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

(because t_1 and t_2 are independent by assumption)

$$= 1 - (.95)^2$$

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

Linear Regression with Multiple Regressors

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

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

Two Solutions:

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

Linear Regression with Multiple Regressors

The *F*-statistic

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

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

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

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

Linear Regression with Multiple Regressors

Large-sample distribution of the *F*-statistic

- Consider the *special case* that t_1 and t_2 are independent, so

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

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

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

Linear Regression with Multiple Regressors

The chi-squared distribution

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

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

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

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

Linear Regression with Multiple Regressors

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

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

See Table 4 on page 807

Linear Regression with Multiple Regressors

```
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
The test command follows the regression

```
test str expn_stu;

 ( 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

Linear Regression with Multiple Regressors

Summary: testing joint hypotheses

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

Linear Regression with Multiple Regressors

Testing Single Restrictions on Multiple Coefficients

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

Consider the null and alternative hypothesis,

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

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

Linear Regression with Multiple Regressors

```
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
------------------------------------------------------------------------------
```

regress testscore str expn pctel, r

test str=expn

Linear Regression with Multiple Regressors

Regression Specification: variables of interest, control variables, and conditional mean independence

- We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant factors outside the school committee's control – such as outside learning opportunities (museums, etc), parental involvement in education (reading with mom at home?), etc.
- If we could run an experiment, we would randomly assign students (and teachers) to different sized classes.
- But with observational data, u_i depends on additional factors (museums, parental involvement, knowledge of English etc).

What if you cannot observe?

Linear Regression with Multiple Regressors

Control variables in multiple regression

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

Linear Regression with Multiple Regressors

- **Control variables: an example from the California test score data**

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

$PctEL$ = percent English Learners in the school district

$LchPct$ = percent of students receiving a free/subsidized lunch (only students from low-income families are eligible)

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

Linear Regression with Multiple Regressors

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

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

Linear Regression with Multiple Regressors

- **Three interchangeable statements about what makes an effective control variable:**

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

Linear Regression with Multiple Regressors

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

For example:

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

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

Linear Regression with Multiple Regressors

The math of control variables: conditional mean independence.

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

$E(ui|Xi, Wi) = E(ui|Wi)$ (conditional mean independence)

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

Linear Regression with Multiple Regressors

Consider the regression model,

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

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

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

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

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

Linear Regression with Multiple Regressors

Implications for variable selection and "*model specification*"

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

Linear Regression with Multiple Regressors

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

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

Linear Regression with Multiple Regressors

What about measures of fit?

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

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

Linear Regression with Multiple Regressors

Analysis of the Test Score Data Set

1. Identify the variable of interest: *STR*
2. Think of the omitted causal effects that could result in omitted variable bias
   - *Whether the students know English; outside learning opportunities; parental involvement; teacher quality (if teacher salary is correlated with district wealth) – there is a long list!*

Linear Regression with Multiple Regressors

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

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

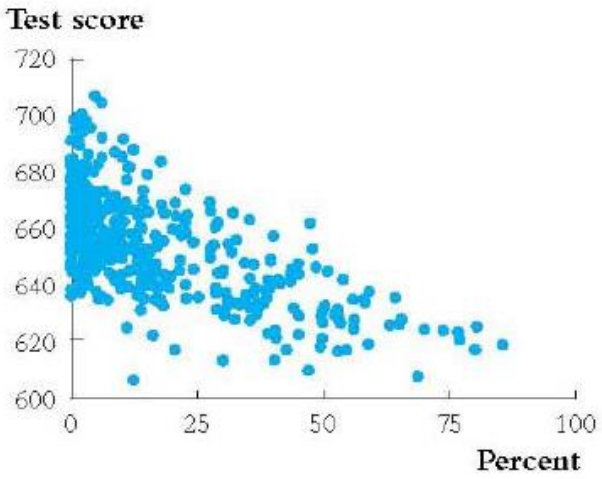
Linear Regression with Multiple Regressors

4. Also specify a range of plausible alternative models, which include additional candidate variables.
   - *It isn't clear which of the income-related variables will best control for the many omitted causal factors such as outside learning opportunities, so the alternative specifications include regressions with different income variables. The alternative specifications considered here are just a starting point, not the final word!*

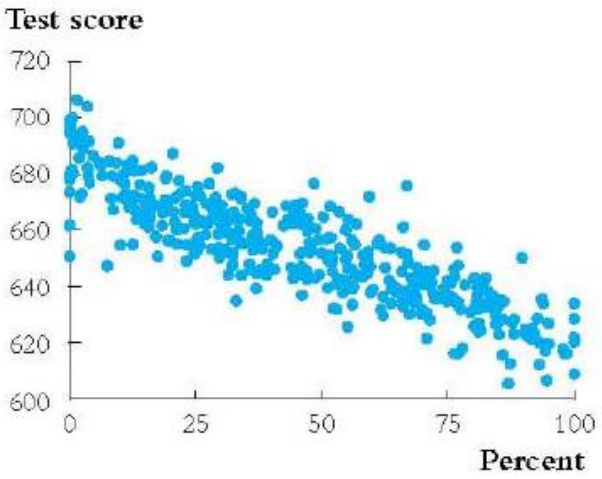
5. Estimate your base model and plausible alternative specifications ("sensitivity checks").

Linear Regression with Multiple Regressors

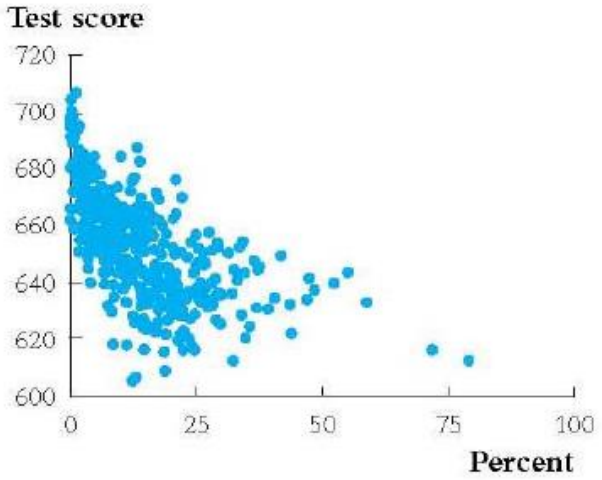
FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public income assistance (X_4)				−0.790** (0.068)	0.048 (0.059)
Intercept	698.9** (10.4)	686.0** (8.7)	700.2** (5.6)	698.0** (6.9)	700.4** (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\overline{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

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

Linear Regression with Multiple Regressors

Summary: Multiple Regression

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