ECON 7310 Elements of Econometrics
Week 5: Assessing Studies Based on Multiple Regression

David Du[1]

[1] University of Queensland

Draft

Outline

- Hypothesis tests and confidence intervals for one coefficient
- Joint hypothesis tests on multiple coefficients
- Other types of hypotheses involving multiple coefficients
- Variables of interest, control variables, and variable selection

Hypothesis Tests and Confidence Intervals for a Single Coefficient

(SW Section 7.1)

- 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.
- $\frac{\widehat{\beta}_1 - \beta_1}{SE(\widehat{\beta}_1)}$ is approximately distributed $\mathcal{N}(0, 1)$.
- Thus hypotheses on β_1 can be tested using the usual t-statistic, and confidence intervals are constructed as $\widehat{\beta}_1 \pm 1.96 \times SE(\widehat{\beta}_1)$.
- The same method applies to $\beta_2, \ldots, \beta_k$.

Example: The California class size data

1. Single Regressor:

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

2. Multiple Regressors:

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} STR - \underset{(0.031)}{0.650} PctEL$$

- The coefficient on STR in (2) is the effect on *TestScore* of a unit change in *STR*, holding constant the percentage of English Learners
- The coefficient on STR falls by one-half. The 95% confidence interval for coefficient on STR in (2) is $-1.10 \pm 1.96 \times 0.43 = (-1.95, -0.26)$
- The t-statistic testing $\beta_{STR} = 0$ is $t = -1.10/0.43 = -2.54$, so we reject the hypothesis at the 5% significance level

Standard errors in 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
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} PctEL$$

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

Tests of Joint Hypotheses (SW Section 7.2)

- Let *Expn* := expenditures per student, and consider

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

- The null hypothesis that "school resources do not matter," and the alternative that they do, corresponds to:

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both}$$

- A joint hypothesis specifies a value for two or more coefficients, i.e., it imposes a restriction on two or more coefficients simultaneously.
- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.

Why can't we just test the coefficients one at a time?

- 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 is not valid: the resulting test rejects too often under the null hypothesis (more than 5%)!
- The "one at time" test is to 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 "one at a time" test rejects H_0, when H_0 is actually true? (It should be 5%.) Suppose t_1 and t_2 are independent,

$$
\begin{aligned}
&\Pr(|t_1| > 1.96 \text{ and/or } |t_2| > 1.96 | H_0) \\
&\quad = 1 - \Pr(|t_1| < 1.96 | H_0) \times \Pr(|t_2| < 1.96 | H_0) \\
&\quad = 1 - (0.95)^2 \approx 9.75\% \neq 5\%
\end{aligned}
$$

- The size of the "common sense" test is not 5%! So, we will study F test.

The F-statistic

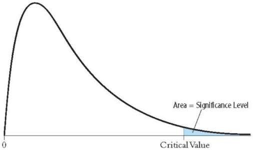
- The heteroskedasticity-robust F-statistic testing H_0 with q restrictions is approximately distributed as $\mathcal{F}_{q,\infty}$ when n is large.
- The critical values for the F-statistic can be found from the tables of $\mathcal{F}_{q,\infty}$. Note that the critical values depend on q.
- It is more convenient to conduct the hypothesis testing using p value

$$p\text{-value} = \Pr(\mathcal{F}_{q,\infty} > \widehat{F})$$

where $\widehat{F}$ is the value of the F statistic actually computed.
- Note that we do not use the homoskedasticity-only F-statistic for the same reason as why we do not use Student t distribution.
  - In economic data, errors are mostly heteroskedastic and normality assumption does not hold. But, n is typically large.

$\mathcal{F}_{q,\infty}$ distribution

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



Degrees of Freedom	Significance Level 10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

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    There are q=2 restrictions being tested
 ( 2)  expn_stu = 0.0
       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
```

Hence, we can reject $H_0 : \beta_1 = \beta_2 = 0$ at significance level of 1%.

Testing Single Restrictions on Multiple Coefficients (SW Section 7.3)

- Consider the regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i, \qquad i = 1, \ldots, n$$

- Consider the null and alternative hypothesis,

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

- This null imposes a single restriction ($q = 1$) on multiple coefficients – it is not a joint hypothesis with multiple restrictions, e.g., $H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$.
- There are two methods for testing single restrictions on multiple coefficients: (1) rearrange the regression (2) perform the test directly

Method 1: Rearrange ("transform") the regression

- We start from

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$$

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

- Add and subtract $\beta_2 X_{i1}$;

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} - \beta_2 X_{i1} + \beta_2 X_{i2} + \beta_2 X_{i1} + u_i \\ &= \beta_0 + (\beta_1 - \beta_2) X_{i1} + \beta_2 (X_{i1} + X_{i2}) + u_i \\ &= \beta_0 + \gamma X_{i1} + \beta_2 W_i + u_i \end{aligned}$$

  where $\gamma := \beta_1 - \beta_2$ and $W_i := X_{i1} + X_{i2}$.
- Test $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$.
- Then, this is equivalent to testing $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$

Method 2: Perform the test directly

- Again, we have

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i$$

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

- Example:

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

- In STATA, to test $H_0 : \beta_1 = \beta_2$ (two sided);

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

Confidence Sets for Multiple Coefficients (SW Section 7.4)

- Consider the regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + u_i, \qquad i = 1, \ldots, n$$

- What is a joint confidence set for β_1 and β_2?
- A 95% joint confidence set is:
  - A set-valued function of the data that contains the true coefficient(s) in 95% of hypothetical repeated samples.
  - Equivalently, the set of coefficient values that cannot be rejected at the 5% significance level.
- You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an F-test (why not just combine the two 95% confidence intervals?).

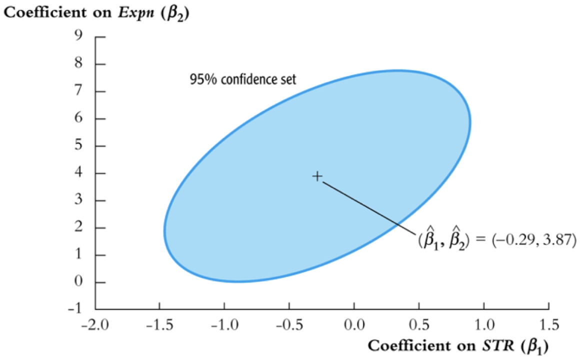
Joint confidence sets (continued)

- 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 on (the test it "inverts") has size of 5%
- 5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not;
- therefore the confidence set constructed as the non-rejected values contains the true value 95% of the time (in 95% of all samples).

Confidence set based on inverting the F-statistic

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



Regression Specification: variables of interest, control variables, and conditional mean independence (SW Section 7.5)

- 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.
  - Then STR_i would be independent of all the things in u_i, so $E(u_i|STR_i) = 0$.
  - Then, the OLS slope estimator in the regression of $TestScore_i$ on STR_i will be an unbiased estimator of the desired causal effect.

Regression Specification: control variables

- But with observational data, u_i depends on additional factors (museums, parental involvement, knowledge of English etc).
- If you can observe those factors (e.g., *PctEL*), then include them.
- But usually you can't observe all these omitted causal factors (e.g., parental involvement in homework).
- In this case, you can include **control variables**
- A **control variable** W is a variable that
  1. is correlated with (controls for) an omitted causal factor in the regression of Y on X,
  2. but does not necessarily have a causal effect on Y.

Control variables: an example from the California test score data

$$\widehat{Test\ Score} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00}STR - \underset{(0.033)}{0.122}PctEL - \underset{(0.024)}{0.547}LchPct$$

$$\overline{R}^2 = 0.773$$

PctEL = percentage of English Learners in the school district
LchPct = percentage 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?

Control variables example (continued)

$$\widehat{Test\ Score} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00}STR - \underset{(0.033)}{0.122}PctEL - \underset{(0.024)}{0.547}LchPct$$

$$\overline{R}^2 = 0.773$$

- *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
  - *PctEL* is correlated with those omitted causal variables.
  - So, *PctEL* is both a possible causal variable and a control variable.
- *LchPct* <u>might</u> have a causal effect (eating lunch helps learning)
  - It is also correlated with and controls for income-related outside learning opportunities.
  - So, *LchPct* is both a possible causal variable and a control variable.

Control variables (continued)

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

Control variables (continued)

Control variables need not be causal, and their coefficients generally DO NOT have a causal interpretation. For example,

$$\underset{}{\widehat{Test\ Score}} = \underset{(5.6)}{700.2} - \underset{(0.27)}{1.00 STR} - \underset{(0.033)}{0.122 PctEL} - \underset{(0.024)}{0.547 LchPct}$$

- 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 $LchPct = 0$!
- This is not reasonable!! In fact, studies show the opposite.

Control Variables: Conditional mean independence

- We need a mathematical statement for effective control variables. Formally, consider the regression model;

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i$$

  where X_i is the variable of interest and W_i is a control variable.
- W_i is an effective control variable if **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. $\widehat{\beta}_1$ is unbiased
  3. The coefficient on the control variable, $\widehat{\beta}_2$, is generally biased
  4. See Appendix 6.5 for the mathematics of 1-3.

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. This results in a **base** or **benchmark** model.
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.
   - Never ever just try to maximize R^2!

Digression 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, e.g., 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 any of the followings;
  - you have eliminated omitted variable bias.
  - you have an unbiased estimator of a causal effect (β_1).
  - the included variables are statistically significant – this must be determined using hypotheses tests.

Analysis of the Test Score Data Set (SW Section 7.6)

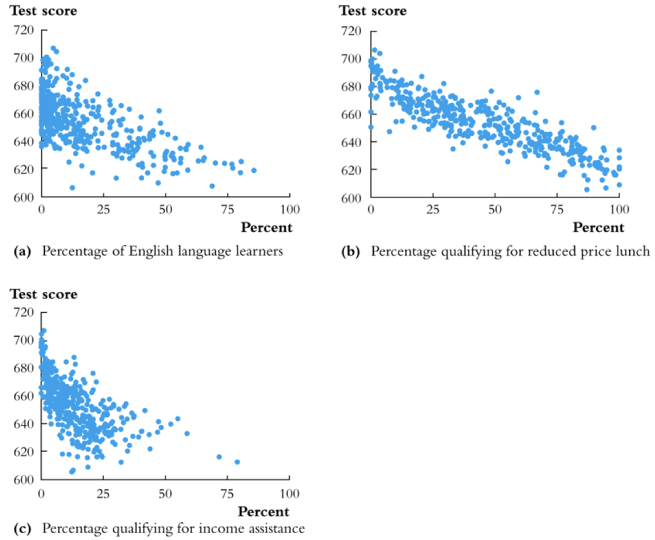
1. Identify the variable of interest: *STR*
2. Think of the omitted causal effects that could result in omitted variable bias;
   - whether the students know English;
   - outside learning opportunities;
   - parental involvement;
   - teacher quality (if teacher salary is correlated with district wealth)
   - there is a long list!
3. Include those omitted causal effects if you can or, if you can't, include control variables to construct a base 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.

Analysis of the Test Score Data Set, continued

4. Also specify a range of plausible alternative models, which include additional candidate variables.
   - It is not 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").

Test scores and California socioeconomic data ...

FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



Digression on presentation of regression results

- We have a number of regressions and we want to report them. It is awkward and difficult to read regressions written out in equation form.
- So it is conventional to report them in a table. The table should include:
  - estimated regression coefficients
  - standard errors
  - measures of fit
  - number of observations
  - relevant F-statistics, if any
  - Any other pertinent information.
- Find this information in the following table:

A Table to summarise estimation results

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

Dependent variable: average test score in the district.

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

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.