Lecture Notes #5 (Chapter 7) Economics 120B Econometrics

Prof. Dahl

UC San Diego

Outline

1. Hypothesis tests and confidence intervals for a single coefficie
2. Joint hypothesis tests on multiple coefficients
3. Other types of hypotheses involving multiple coefficients
4. How to decide what variables to include in a regression model

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

- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\operatorname{var}(\hat{\beta}_1)}}$ 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 $\beta_2, \ldots, \beta_k$.
- $\hat{\beta}_1$ 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) $\widehat{TestScore} = 698.9 - 2.28 \times STR$

$\qquad\qquad(10.4)\ \ (0.52)$

(2) $\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$

$\qquad\qquad(8.7)\ \ (0.43)\qquad\qquad(0.031)$

- 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
- 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 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:

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

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

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

Tests of joint hypotheses, ctd.

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

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

- A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this "one at a time" test isn't valid: the resulting test rejects too often under the null hypothesis (more than 5%)!

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

Because the rejection rate under the null isn't 5%. We'll calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual t-statistics. To simplify the calculation, suppose that $\hat{\beta}_1$ 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 "one at time" 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 "one at a time" test rejects H_0, when H_0 is actually true? (It *should* be 5%.)

Suppose t_1 and t_2 are independent (for this calculation).

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

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

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

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

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

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

(t_1, t_2 are independent by assumption)

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

= $.0975 = 9.75\%$ – which is ***not*** the desired 5%!!

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

- The size of the "common sense" test 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 that test both β_1 and β_2 at once: the F-statistic (this is common practice)

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?)

The F-statistic testing β_1 and β_2:

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

- 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 β's is 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

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

- 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.

The ***chi-squared*** distribution with q degrees of freedom (χ_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 χ_q^2/q.

Selected large-sample critical values of χ_q^2/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	

Computing the p-value using the F-statistic:

p-value = tail probability of the χ_q^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

------------------------------------------------------------------------------
             |               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
```

More on *F*-statistics: *a simple F-statistic formula that is easy to understand (it is only valid if the errors are homoskedastic, but it might help intuition).*

The homoskedasticity-only *F*-statistic

When the errors are homoskedastic, there is a simple formula for computing the "homoskedasticity-only" *F*-statistic:

- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- Compare the fits of the regressions – the R^2's – if the "unrestricted" model fits sufficiently better, reject the null

The "restricted" and "unrestricted" regressions

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

Unrestricted population regression (under H_1):

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

Restricted population regression (that is, under H_0):

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

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

By how much must the R^2 increase for the coefficients on *Expn* and *PctEL* to be judged statistically significant?

Simple formula for the homoskedasticity-only F-statistic:

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

where:

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

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

q = the number of restrictions under the null

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

- The bigger the difference between the restricted and unrestricted R^2's – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only F.

Example:

<u>Restricted regression</u>:

$$\widehat{TestScore} = \underset{(1.0)}{644.7} - \underset{(0.032)}{0.671}PctEL, \quad R^2_{restricted} = 0.4149$$

<u>Unrestricted regression</u>:

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}STR + \underset{(1.59)}{3.87}Expn - \underset{(0.032)}{0.656}PctEL$$

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

so

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

$$= \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = \mathbf{8.01}$$

Note: Heteroskedasticity-robust F = **5.43**…

The homoskedasticity-only F-statistic – summary

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

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

Digression: The *F* distribution

Your regression printouts might refer to the "*F*" distribution.

If the four multiple regression LS assumptions hold ***and***:

5. u_i is homoskedastic, that is, var(u|X_1,…,X_k) does not depend on X's
6. u_1,…,u_n are normally distributed

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* distribution is to the χ^2_q/q distribution what the t_{n-1} distribution is to the *N*(0,1) distribution**

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

- The F distribution is tabulated many places
- As $n \to \infty$, the $F_{q,n\text{-}k-1}$ distribution asymptotes to the χ_q^2/q distribution:

The $F_{q,\infty}$ and χ_q^2/q distributions are the same.

- For q not too big and $n \geq 100$, the $F_{q,n-k-1}$ distribution and the χ_q^2/q distribution are essentially identical.
- Many regression packages (including STATA) compute p-values of F-statistics using the F distribution
- You will encounter the F distribution in published empirical work.

Summary: the homoskedasticity-only *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 χ_q^2/q (that is, $F_{q,\infty}$) critical values.
- For $n \geq 100$, the *F*-distribution essentially is the χ_q^2/q distribution.
- For small *n*, sometimes researchers use the *F* distribution because it has larger critical values and in this sense is more conservative.

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_q^2/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

An example of a multiple regression analysis – and how to decide which variables to include in a regression…

A Closer Look at the Test Score Data

(SW Sections 7.5 and 7.6)

We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant student and school characteristics (but not necessarily holding constant the budget (*why*?)).

To do this we need to think about what variables to include and what regressions to run – and we should do this before we actually sit down at the computer. This entails thinking beforehand about your *model specification*.

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 (β_1)?
- Is a candidate variable statistically significant?
- Use judgment, not a mechanical recipe…
- Don't 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 $\bar{R}^2$ – but this loses sight of our real objective, an unbiased estimator of the class size effect.

- 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 (β_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.

Back to the test score application:

- *What variables would you want – ideally – to estimate the effect on test scores of STR using school district data?*
- *Variables actually in the California class size data set*:
  - student-teacher ratio (*STR*)
  - percent English learners in the district (*PctEL*)
  - school expenditures per pupil
  - name of the district (so we could look up average rainfall, for example)
  - percent eligible for subsidized/free lunch
  - percent on public income assistance
  - average district income
- *Which of these variables would you want to include?*

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 instead it is conventional to report them in a table.
- A table of regression results 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:

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. 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.