ECON3150/4150 Spring 2016

Lecture 9

Siv-Elisabeth Skjelbred

University of Oslo

February 15th

Last updated: February 12, 2016

- We remember that when we have a set of linear restrictions we use the F-test.
- The most common type of multiple restrictions is a set of exclusion restrictions.
- Exclusion restrictions involve a set of variables all hypothesized to have a zero effect on the dependent variable.
- Other restrictions may involve a certain value of one or more coefficients (possibly in addition to exclusion restrictions).
- To do a joint hypothesis test we needs two regressions:
  - The unrestricted model
  - The restricted model
- Based on these two models we compute the F-statistic

The F-statistic

$$F \equiv \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n-k-1)}$$

- SSR_R is the SSR for the restricted model.
- SSR_{UR} is the SSR for the unrestricted model.
- q is the number of restrictions
- $n - k - 1$ is the degrees of freedom in the unrestricted model.
- k is the number of independent variables in the unrestricted model.

The F-statistic

Intuition:

$$F = \frac{\text{Average loss in explanatory power under } H_0}{\text{Average unexplained variation under } H_A}$$

- The F-statistic measures the relative increase in the SSR when moving from the unrestricted to the restricted model.
- If F is high we lose a lot of explanatory power by our restrictions.
- The F-statistic is used for testing whether the increase in SSR from the unrestricted model to the restricted model is large enough to warrant the rejection of the null hypothesis.

The critical value of F-statistics

- Under the null hypothesis, and assuming that the OLS assumptions hold, the F-statistic is distributed as an F random variable with (q,n-k-1) degrees of freedom.
- This is written as $F \sim F_{q,n-k-1}$
- In large samples the F statistic is distributed $F_{q,\infty}$

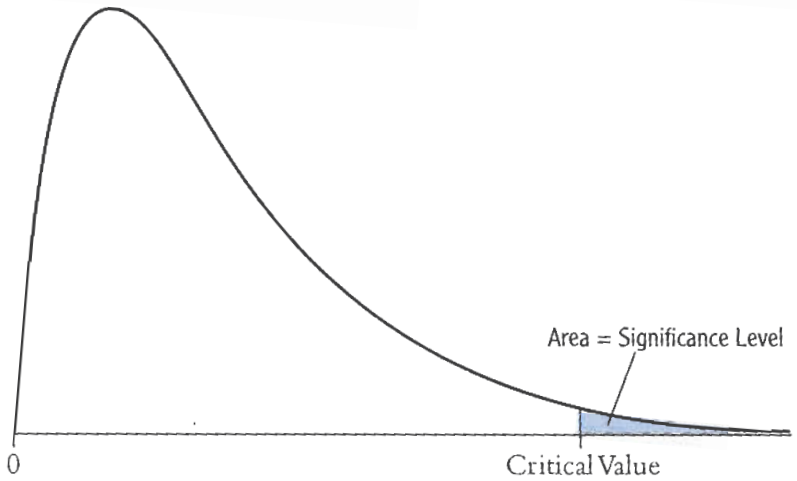
The F-test

- As with the t-statistic we will reject the null hypothesis when F is sufficiently large.
  - Choose a desired significance level
  - Find the critical value in the F-table associated with that significance level.
  - If $F > F^c$ reject the null.
- Note:
  - The variables can be jointly significant even if all the included variables are individually insignificant.
  - The variables can be jointly insignificant even when one (or more) of the variables are individually significant.

The F-distribution table

When the denominator degrees of freedom is large:

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



	Significance Level		
Degrees of Freedom	**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

The F-distribution table

For a 5% significance level:

TABLE 5B Critical Values for the F_{n_1, n_2} Distribution—5% Significance Level

Denominator Degrees of Freedom (n_2)	Numerator Degrees of Freedom (n_1)									
	1	2	3	4	5	6	7	8	9	10
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38

Example F-test with exclusion restriction

Unrestricted model:

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$$

with: n=88, SSR=156765. $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$
Restricted model:

$$price = \beta_0 + \beta_1 assess + u$$

with n=88, and $SSR = 165644$

$$F = \frac{\frac{(165644-156765)}{3}}{\frac{156765}{88-4-1}} = 1.567$$

Critical value at 5% for $F_{3,83} \approx F_{3,90} = 2.72$ Cannot reject the null.

F-test in stata

Most statistical softwares have built-in feature for testing multiple exclusion restrictions. The advantages are:

- Less likely to make a mistake.
- p-value are computed automatically.
- The problem of missing data is handled without any additional work on our part.

Stata

```
. reg price assess lotsize sqrft bdrms

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  4,    83) =  100.74
       Model |  761089.801     4  190272.45           Prob > F      =  0.0000
    Residual |  156764.704    83  1888.73138           R-squared     =  0.8292
-------------+------------------------------           Adj R-squared =  0.8210
       Total |  917854.506    87  10550.0518           Root MSE      =   43.46

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      assess |   .9082991   .1040386     8.73   0.000     .7013706    1.115228
     lotsize |   .0005867   .0004963     1.18   0.240    -.0004004    .0015738
       sqrft |  -.0005175   .0170849    -0.03   0.976    -.0344986    .0334636
       bdrms |   11.60249   6.549515     1.77   0.080    -1.424233    24.62921
       _cons |  -38.88702   21.49853    -1.81   0.074    -81.64673    3.872696
------------------------------------------------------------------------------

. test lotsize sqrft bdrms

 ( 1)  lotsize = 0
 ( 2)  sqrft = 0
 ( 3)  bdrms = 0

       F(  3,    83) =     1.57
            Prob > F =     0.2035
```

Example joint insignificance

The following example shows that a variable can be individually significant, but jointly insignificant with another variable.

```
1 . reg wage jc univ ne nc south black hispanic, robust

  Linear regression                                      Number of obs =       6763
                                                            F(  7,  6755) =  118.05
                                                            Prob > F      =  0.0000
                                                            R-squared     =  0.1114
                                                            Root MSE      =  5.0033

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          jc |   .5867815    .077754     7.55   0.000     .4343592    .7392038
        univ |   .7175047   .0286009    25.09   0.000     .6614379    .7735716
          ne |   .2013285   .2045864     0.98   0.325    -.1997254    .6023824
          nc |  -.4011147   .1922327    -2.09   0.037    -.7779514   -.0242779
       south |  -.6561036   .1870899    -3.51   0.000    -1.022859   -.2893485
       black |  -1.211106   .1966389    -6.16   0.000     -1.59658   -.8256322
    hispanic |  -.2937563   .2782542    -1.06   0.291    -.8392222    .2517096
       _cons |   9.474505   .1663662    56.95   0.000     9.148375    9.800635
------------------------------------------------------------------------------

2 . test nc=hispanic=0

   ( 1)  nc - hispanic = 0
   ( 2)  nc = 0

         F(  2,  6755) =     2.37
              Prob > F =     0.0932
```

Example joint significance

The following example shows that two variables can be individually insignificant, but jointly significant.

```
1 . reg lsalary years gamesyr hrunsyr rbisyr bavg, robust

  Linear regression                                      Number of obs =        353
                                                            F(  5,   347) =  136.52
                                                            Prob > F      =  0.0000
                                                            R-squared     =  0.6278
                                                            Root MSE      =  .72658

 ------------------------------------------------------------------------------
              |               Robust
      lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
 -------------+----------------------------------------------------------------
        years |   .0688626   .0156569     4.40   0.000     .0380682    .0996571
      gamesyr |   .0125521   .0026492     4.74   0.000     .0073416    .0177626
      hrunsyr |   .0144295   .0165958     0.87   0.385    -.0182115    .0470706
       rbisyr |   .0107657   .0071903     1.50   0.135    -.0033763    .0249078
         bavg |   .0009786   .0008162     1.20   0.231    -.0006267    .0025839
        _cons |   11.19242   .2366536    47.29   0.000     10.72696    11.65788
 ------------------------------------------------------------------------------

2 . test (hrunsyr=0) (rbisyr=0)

   ( 1)  hrunsyr = 0
   ( 2)  rbisyr = 0

         F(  2,   347) =    13.18
              Prob > F =     0.0000
```

The F-statistic and R^2

- As you saw in the example the SSR was large which may make computations based on SSR tedious.
- R-squared is always between 0 and 1 which makes it more convenient.
- Using the fact that $SSR = (1 - R^2)SST$ we can substitute for SSR_R and SST_{UR} into the F-statistic

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n - k - 1)}$$

- This is called the R-squared form of the F-statistic.

Example R-squared of F-statistic

In the house-price example the unrestricted model gives R^2 of 0.8292 while the restricted model gives R^2 of 0.8195.

$$F = \frac{0.8292 - 0.8195}{1 - 0.8292} * \frac{83}{3} = 1.57$$

Gives the same F-statistic as with using the SSR.

Example F-test general linear restrictions

Unrestricted model:

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$$

with: n=88, SSR=156765.
$H_0 : \beta_1 = 1$ & $\beta_2 = \beta_3 = \beta_4 = 0$ gives restricted model:

$$price - asess = \beta_0 + u$$

with n=88 and SSR=166116

$$F = \frac{\frac{166166-156765}{4}}{\frac{156765}{88-4-1}} = 1.24$$

Critical value at 5% for $F_{4,83} \approx F_{4,90} = 2.47$ Cannot reject the null.

The F-statistic of the regression

- The extreme version of the exclusion restriction is to test the hypothesis that **all** the slope coefficients are zero.
- Under this null hypothesis none of the regressors explain any of the variation in Y_i.
- The restricted model is then:

$$Y = \beta_0 + u$$

- Stata automatically reports the F-statistic for this hypothesis which is given by:

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

- This F-statistic determines the overall significance of the regression.

P-values for F-test

- Since the F-distribution depends on the numerator and denominator degrees of freedom there are numerous critical F-values.
- In the F-testing context the p-value is defined as:

$$p - value = P(\mathscr{F} > F)$$

- where $\mathscr{F}$ denotes an F random variable with (q,n-k-1) degrees of fredom and F is the actual value of the test statistic.
- The p-value is still the probability of observing a value of the test statistic at least as large as we did, given that the null hypothesis is true.
- As with t-testing once the p-value has been computed the F-test can be carried out at any significance level.

The relationship between F- and t-statistics

The relationship between t and F is apparent in the case with 1 or 2 restrictions.

- If q=1 the F-statistic tests a single restriction and the F-statistic is the square of the t-statistic
- Since t^2_{n-k-1} has an $F_{1,n-k-1}$ distribution the two approaches lead to exactly the same outcome for a two-sided test.
- If q=2 and we test exclusion restrictions then:

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

- t_1 and t_2 is the t-statistic associated with the two exclusion coefficients
- $\hat{\rho}_{t_1,t_2} t_1, t_2$ is an estimator of the correlation between the two t-statistics.

Why not use a repeated t-test?

- Will the F-test give the same result as a repeated t-test where we test each hypothesis separately?
- Assume that we want to test:

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

- against

$$H_0 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$

- We can regard this as a joint null hypothesis made up of:

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

- assume as a simplification that the t-tests for the two null hypothesis are stochastically independent with significance level ϵ_1 and ϵ_2.

Repeated t-test

The overall significance level is then:

$$P(\text{reject either } H_0' \text{ or } H_0''|H_0) =$$
$$1-P(\text{reject neither } H_0' \text{ nor } H_0''|H_0) =$$
$$1-(1-\alpha_1)(1-\alpha_2)$$
$$\alpha_1 + \alpha_2(1-$$
$$alpha_1)$$

If you set the two significance levels equal to each other then:

$$P(\text{reject either } H_0' \text{ or } H_0''|H_0) = \alpha + \alpha(1-\alpha) > \alpha$$

Thus the significance level of this joint test is larger than the level of each individual test.

The Bonferroni test of a joint hypothesis

- The Bonferroni test corrects the individual significance level so that the significance level of the test equals the desired significance level.
- In general the Bonferroni test can be conducted even when the t-statistics are correlated.
- The overall significance level of α is secured by choosing the significance level of each test so that:

$$\epsilon = \frac{\alpha}{m} \text{ (Bonferroni)}$$

- Where m is the number of individual tests.
- The F-test is the preferred method as it is a better test, but the Bonferroni method may be useful if you only have the regression results and not the data.

Heteroskedasticity and the F-statistic

- The formula for the F-statistic is computed under the homoskedasticity assumption.
- The robust command makes the standard errors heteroskedasticity robust but it does not alter the SSR (and the R^2) of the regression.
- The formula for heteroskedasticity robust F-statistic is complicated.
- If the the ZCM, large outliers are unlikely and no perfect collinearity assumptions hold, then under the null hypothesis:

$$F \xrightarrow{d} F_{q,\infty}$$

- Stata can compute heteroskedasticity robust F-statistic.

Example homoskedastic F-using formula

```
1 . reg testscr str expn_stu el_pct, robust

  Linear regression                                      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    -1.234002     .661203
     expn_stu  |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       el_pct  |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
         _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
  ------------------------------------------------------------------------------

2 . reg testscr el_pct, robust

  Linear regression                                      Number of obs =         420
                                                            F(  1,   418) =  436.58
                                                            Prob > F      =  0.0000
                                                            R-squared     =  0.4149
                                                            Root MSE      =  14.592

  ------------------------------------------------------------------------------
               |               Robust
     testscr   |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
       el_pct  |  -.6711562   .0321211   -20.89   0.000    -.7342952   -.6080172
         _cons |   664.7394   .9740374   682.46   0.000     662.8248    666.6541
  ------------------------------------------------------------------------------

3 . display (0.4366-0.4149)/2/((1-0.4366)/(420-3-1))
  8.0113596
```

Example homoskedastic F-using Stata

```
1 . reg testscr str expn_stu el_pct

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  3,   416) =  107.45
       Model |  66409.8837     3  22136.6279           Prob > F      =  0.0000
    Residual |  85699.7099   416  206.008918           R-squared     =  0.4366
-------------+------------------------------           Adj R-squared =  0.4325
       Total |  152109.594   419  363.030056           Root MSE      =  14.353

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863992   .4805232    -0.60   0.551    -1.230955     .658157
    expn_stu |   .0038679   .0014121     2.74   0.006     .0010921    .0066437
      el_pct |  -.6560227   .0391059   -16.78   0.000    -.7328924   -.5791529
       _cons |   649.5779   15.20572    42.72   0.000     619.6883    679.4676
------------------------------------------------------------------------------

2 . test str=expn_stu = 0

 ( 1)  str - expn_stu = 0
 ( 2)  str = 0

       F(  2,   416) =      8.01
            Prob > F =      0.0004
```

Example heteroskedasticity robust F

```
1 . reg testscr str expn_stu el_pct, robust

  Linear regression                                      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    -1.234002     .661203
     expn_stu  |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       el_pct  |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
         _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
  ------------------------------------------------------------------------------

2 . test str=expn_stu = 0

   ( 1)  str - expn_stu = 0
   ( 2)  str = 0

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

Hypothesis testing with non-nested models

- The restricted model is a nested version of the unrestricted model as the unrestricted model is a special case of the restricted model.
- The F-test can only be used to test nested models.
- To identify the effect of each you can run separate regressions with each of the variables.
- Use adjusted R-squared

Non-nested model

Model 1:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

Model 2:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 IQ + u$$

Which model is "best" if $\bar{R}^2 = 0.1527$ for model one and $\bar{R}^2 = 0.1595$ for model 2?

Non-nested models

$$salary = \beta_0 + \beta_1 years + \beta_2 bavg + \beta_3 gamesyr + \beta_4 hrunsyr + u$$

$$salary = \beta_0 + \beta_1 years + \beta_2 bavg + \beta_3 gamesyr + \beta_4 rbisyr + u$$

- Where salary is yearly salary of a baseball player, bavg is batting average, gamesyr is games per year, hrunsyr is homeruns per year and rbisyr is runs batted in per year.
- If both of hrunsyr and rbisyr are included in the same regression they are individually insignificant as they are so strongly correlated, while they are significant if they are included separately.
- The adjusted R-squared can serve as an indicator for which model is to prefer.
- Note: The dependent variable of the two models must be on the same functional form.

Model specification

The role of control variables

- We may have one (or more) independent variables for which we desire to estimate a causal effect.
- To avoid omitted variable bias we may include more variables in the regression.
- These variables are called control variables and are not the object of interest in the study.
- Thus we do only care about whether variable(s) of interest is unbiased.

Example

If estimating: Testscore $= \beta_0 + \beta_1 \text{STR} + \beta_2 \text{PctEL}$

- STR is likely to be correlated with "outside learning opportunities" (unmeasurable)
- Outside learning opportunities is correlated with the student's economic background (measurable).
- LchPct measure the fraction of economically disadvantaged children in the district.

Control variables

Control variable

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

$$TestScore = \beta_0 + \beta_1 STR + \beta_2 PctEL + \beta_3 LchPct$$

- STR is the student teacher ratio and the variable of interest
- PctEL is the percentage of english learners and probably has a direct causal effect, but it also serves as a control as it is correlated with outside learning opportunities.
- LctPct might have causal effect but it is correlated with and thus controls for income-related outside learning opportunities.

Control variables

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

Control variables

- Control variables are selected because they are correlated with omitted factors, that means that they are likely to be biased.
- This means that the zero conditional mean assumption will not hold.
- However, the control variable is effective if the mean of u does not depend on the variables of interest given the control variable. This is called the conditional mean independence.

Conditional mean independence

Conditional mean independence

$$E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$$

- Where X_1 represents the variable of interest and X_2 represents the control variable.
- The zero conditional mean assumption ensures unbiased estimates however it a strong assumption and can be replaced by the weaker conditional mean independence assumption.
- If the conditional mean independence assumption holds the variable of interest has a causal interpretation (but the control variables are potentially biased).

Specify population regression function

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 so serve as control variables.
4. Sensitivity check your model by alternative specifications.

Specify population regression function

Sensitivity check to OVB

- Specify base specification with variable of interest and control variables suggested by economic theory and expert judgment.
- Develop a list of candidate alternative specifications.
- If the estimates of the coefficients are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification are reliable.
- If they are not similar this often is evidence that the original specification had omitted variable bias.

Specify population regression function

How many variables should you include?

- Perform an hypothesis test to control whether variables belong to the model
- Reason whether there is likely or unlikely to be omitted variable bias.
- The question of what constitutes the right set of regressors is difficult as you must weigh issues of omitted variable bias, data availability, data quality and economic theory.

Reporting regression results

Reporting regression results

- Always report OLS coefficients, the standard errors and the number of observations used in estimation.
- For the key variables in the analysis you should interpret the estimated coefficients.
- Discuss both economic and statistical significance.
- The R-squared from the regression should always bee included.
- You should think about the scale of the variables so that it easy to read and interpret your regression results.
- If there is any relevant F-statistic then you should report this.

Tabular presentation

- If only a couple of models are being estimated the results can be summarized in equation form, but in many cases a table (or multiple tables) is to prefer.
- The dependent variable should be indicated clearly in the table, and the independent variables should be listed in the first column.
- It is common to indicate significance levels with stars.

California test score data set

- The book throughout the first chapters use a data set constituting the tests scores of Californian students.
- The primary interest is in establishing whether the student teacher ratio (STR) has a causal effect on the student tests scores.
- Factor such as outside learning opportunities are correlated with STR and provides potential of OVB.
- These factors are not directly measurable, but we can include control variables that are correlated with these omitted factors.
- If the control variables are adequate in the sense that the conditional mean independence assumption holds, then we can give the coefficient a causal interpretation.

California test score data set

- Potential background variables:
  - Percentage of students who are still learning English.
  - The percentage of students who are eligible for a subsidized or free lunch.
  - The percentage of students whose families qualify for a California income assistance program.

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
$\overline{R}^2$	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.

Dummy variable classification

If you believe two groups have different effect on Y you can:

- Run separate regressions for the two groups
- Allow the two groups to have different intercepts.

Men:

$$wage = \beta_1 + \beta_2 educ + u$$

Women:

$$wage = \beta_1 + \delta + \beta_2 educ + u$$

where δ is defined as the difference between men and women. Combined:

$$wage = \beta_1 + \delta F + \beta_2 educ + u$$

Dummy variable classification

If the qualitative explanatory variable has more than two categories:

- Example: type of schools: Technical schools, vocational schools, general schools
- The standard procedure is to chose one category as the reference category and define dummy variables for each of the others.
- In general it is good practice to select the most normal or basic category as the reference category
- Thus define a dummy for technical and one for vocational schools.
- Each of the dummy variables will have a coefficient which represents the extra Y of the school, relative to the reference category.

Dummy variable classification

- The specification with a particular reference group allows a test of the difference between other groups and the reference group.
- However, suppose that we were interested in the in the difference between vocational and technical schools?
- You can rerun the regression including a dummy for general schools and dropping the dummy for either vocational or technical schools.
- The function of Y for a given group will be exactly the same
- But the t-test now shows the null hypothesis, if vocational is reference group, the t-test for the coefficient of technical will now test whether the technical groups differ from vocational groups.

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.