Hypothesis Tests and Confidence Intervals in Multiple Regression

(SW Chapter 7)

Outline

1. Why multipple regression?
2. Simpson's paradox (omitted variables bias)
3. Hypothesis tests and confidence intervals for a single coefficient
4. Joint hypothesis tests on multiple coefficients
5. Other types of hypotheses involving multiple coefficients
6. How to decide what variables to include in a regression model?

Simpson’s papradox (omitted variables bias)

*Example (*PJ Bickel, EA Hammel, JW O'Connell - Science, 1975): Sex bias in admission to UC Berkeley in 1973? Over-all, applying men were more likely to be admitted than women, but within department, the picture is different: Women apply more often to departments with low admission!

	Applicants	**% admitted**
meen	8442	**44%**
Women	4321	35%

Department	**Men**		**Women**	
	Applicants	% admitted	Applicants	% admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

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

$\frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)}$ is approximately N(0,1) distributed, even when the standard error is robutly estimateed, when (6.4, p. 204):

1. $E\left[Y_i \mid X_{1i}, \cdots, X_{ki}\right] = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji}$.
2. $\left(Y_i, X_{1i}, \cdots, X_{ki}\right)$ are independently drawn from an infinite population.
3. Large outliers are unlikely.
4. There is no multicollinearity.

Regression estimators for different parameters are usually *correlated*, as are standard errors.

Tests and confidence intervals for β_j are straight forward from the normal distribution!

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

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

Heteroskedasticity-robust standard errors! No easy formula in multiple regression.

Tests of Joint Hypotheses

Question: do test results depend on economic school resources?
Kristin Clemet (former minister of education from Høyre): "It is not a matter of money"

Expn = expenditures per pupil does also measure economic resources

Model:

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

H_0: $\beta_1 = 0$ ***and*** $\beta_2 = 0$
vs. H_1: ***either*** $\beta_1 \neq 0$ ***or*** $\beta_2 \neq 0$ ***or both***

Three approaches for testing:

1. Test first $\beta_1 = 0$ then $\beta_2 = 0$, both at 5% level, and reject H_0 if at least one of the separate hypotheses is rejected. Simultaneous level more than 5%! Test statistics $t_j = |\hat{\beta}_j / SE(\hat{\beta})|$
2. Do the same, but apply Bonferroni: test the individual hypotheses at level 5/2=2.5%:
   $R_1 = \{H_{01} : \beta_1 \text{ is rejected at level } \alpha\}$, same for R_2,
   $P_0(R_1 \cup R_2) \leq P_0(R_1) + P_0(R_2) = 2\alpha$
3. Use the F-test which tests $H_0 : \beta_1^2 + \beta_2^2 = 0 \Leftrightarrow H_0 : \beta_1 = 0 \,\&\, \beta_2 = 0$

The *F*-statistic

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

Formula for the special case of the joint hypothesis $\beta_1 = 0$ and $\beta_2 = 0$ in a regression with two regressors:

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) = \frac{1}{2}\left(\frac{\left(t_1 - \hat{\rho}_{t_1,t_2} t_2\right)^2 + \left(1 - \hat{\rho}_{t_1,t_2}^2\right) t_2^2}{1 - \hat{\rho}_{t_1,t_2}^2}\right) \geq 0$$

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between the individual test statistics t_1 and t_2. F is a quadratic form in the two test statistics $t_j = \hat{\beta}_j / SE\left(\hat{\beta}\right)$ centered at the origin.

Reject when F is large.

The null distribution of F is approximately the F-distribution with q=2 degrees of freedom in the numerator and n-2-1 df in the denominator, see SW: Table 5A.

When the two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are uncorrelated, they are independent due to normality in large samples, and $\hat{\rho}_{t_1,t_2} \approx \rho_{t_1,t_2} = 0$ making

$$F \approx \frac{1}{2}(t_1^2 + t_2^2) \approx \frac{\chi^2_{df=2}}{df} \approx F_{2,\infty}$$

Implementation in STATA

Example, test scores: 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.

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

      test str expn_stu;                    Note: 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.Table!
```

The traditional 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 or the *SSR*'s
- if the "unrestricted" model fits sufficiently better, reject the null

Example, test scores: 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):

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i$$

- 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?

The homoskedasticity-only F-statistic (SW 7.13 or 7.14)

Example:

Restricted regression:

Fitted equation *TestScore* = 644.7 –0.671*PctEL*,

$R^2_{restricted}$ = 0.4149

Unrestricted regression:

Fitted equation *TestScore* = 649.6 – 0.29*STR* + 3.87*Expn* – 0.656*PctEL*

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

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

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

Summary: testing joint hypotheses

- 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 also in practice), and can help intuition, but isn't valid when there is heteroskedasticity

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

Two equivalent methods

1. ***Rearrange ("transform") the regression***

   Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression

   $$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + (\beta_1 - \beta_2) X_1 + \beta_2 (X_2 + X_1) = \beta_0 + \beta_1' X_1 + \beta_2 X_2'$$

   $$H_0 \Leftrightarrow \beta_1' = 0$$

2. ***Perform the test directly***

   Some software, including STATA, lets you test restrictions using multiple coefficients directly .

Example:

Is it total economic resources per student that matters?

(1) $TestScore_i = \beta_0 + \beta_1(1/STR_i) + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$

When Expn is measured per student in the unit of teacher sallery, the null model is

$TestScore_i = \beta_0 + \beta_1[(1/STR_i) + Expn_i] + \beta_3 PctEL_i + u_i$

In STATA, to test $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ in model (1):

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

Confidence Sets for Multiple Coefficients

A 95% ***joint confidence set*** for β_1 and β_2 is:

- A set-valued function of the data that contains the true parameter vector in 95% of hypothetical repeated samples.
- The set of parameter 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?*).

  - Let $F(\beta_{1,0},\beta_{2,0})$ be the (heteroskedasticity-robust) *F*-statistic testing the null hypothesis: $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$. The 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 (look it up!)

This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%

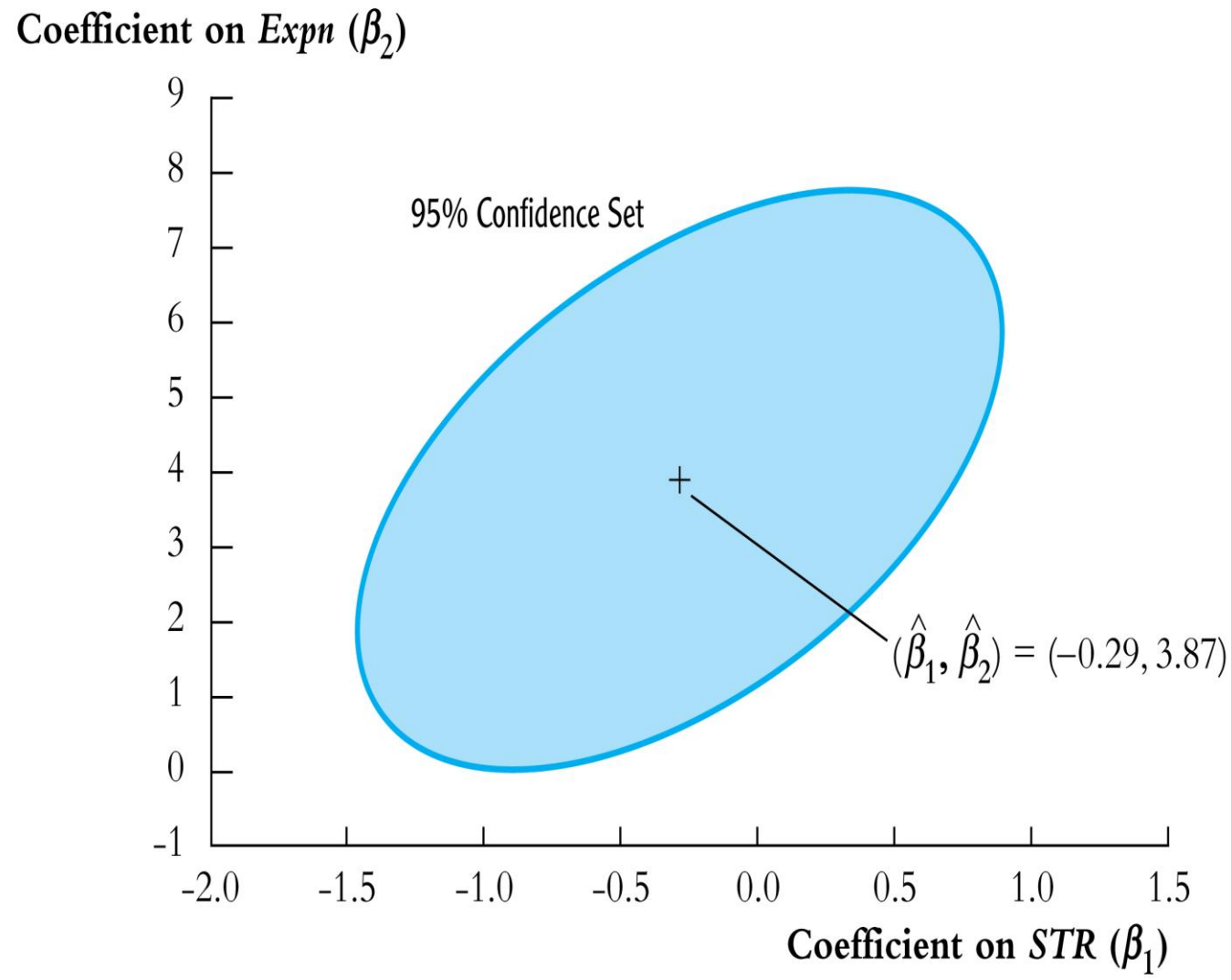
5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

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

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

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.



The test score example of a multiple regression analysis – how to decide which variables to include in a regression…

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*.

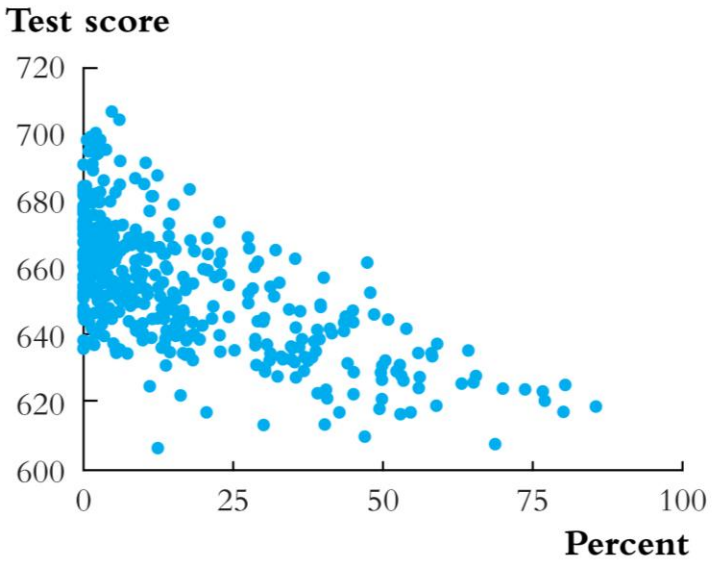
A general approach to variable selection and "*model specification*"

- 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!

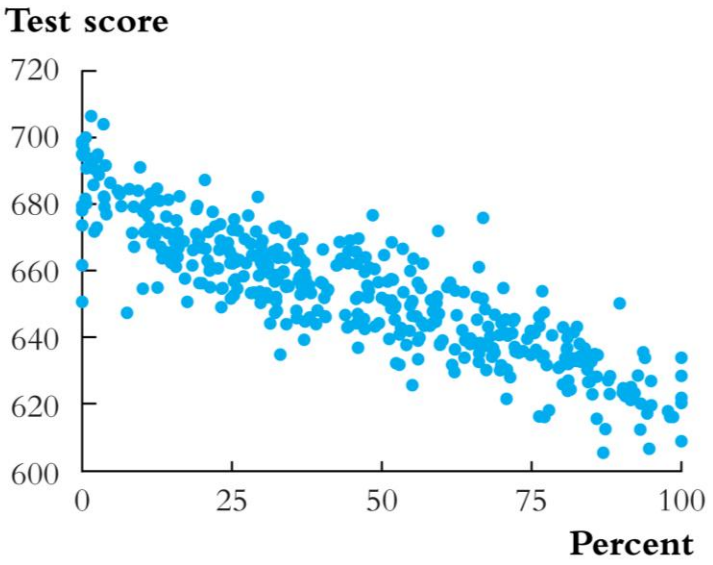
- *What variables would you want – ideally – to estimate the effect on test scores of STR using school district data?*
- *Variables actually in the California class size data set*:
  - student-teacher ratio (*STR*)

  - 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?*

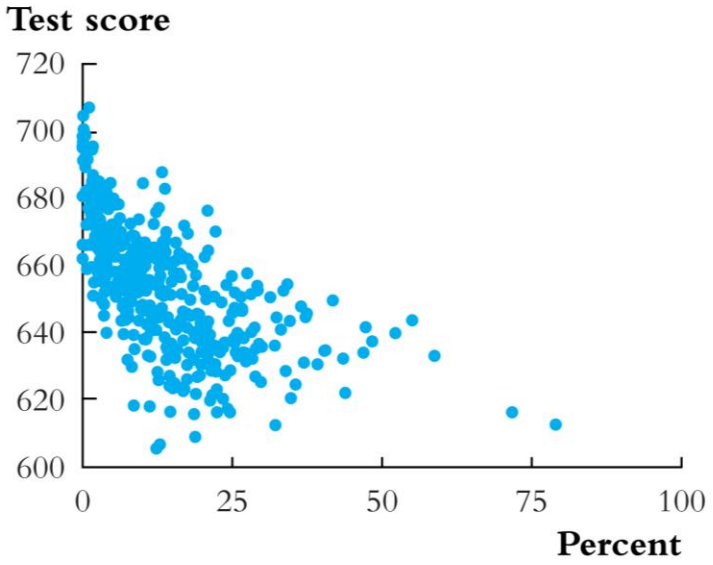
More California data…



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



(c) Percentage qualifying for income assistance

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
$\overline{R}^2$	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. 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.

Do SW: 7.8