

# Econ 312: Introduction to Econometrics

## Multiple Linear Regression

Sang-Yeob Lee

April 23, 2009

# The Least Squares Assumptions for Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \epsilon_i, \quad i = 1, \dots, n$$

- 1 The conditional distribution of  $\epsilon$  given the  $X$ 's has mean zero, that is,  $E(\epsilon | X_1 = x_1, \dots, X_k = x_k) = 0$
- 2  $(X_{1i}, X_{2i}, \dots, X_{ki}, Y_i)$ ,  $i = 1, \dots, n$  are i.i.d.
- 3 Large outliers are rare:  $X_1, \dots, X_k$ , and  $Y$  have fourth moments:  $E(X_{1i}^k) < \infty \dots, E(X_{ki}^k) < \infty \dots, E(Y_i^k) < \infty$
- 4 There is no perfect multicollinearity

## Assumption #1: the conditional mean of $\epsilon$ given the included $X$ s is zero.

$$E(\epsilon|X_1 = x_1, \dots, X_k = x_k) = 0$$

- This has the same interpretation as in regression with a single regressor.
- If an omitted variable (1) belongs in the equation (so is in  $\epsilon$ ) and (2) is correlated with an included  $X$ , then this condition fails
- Failure of this condition leads to omitted variable bias
- The solution - if possible - is to include the omitted variable in the regression

Ctd.

Assumption #2:  $(X_{1i}, \dots, X_{ki}, Y_i), i = 1, \dots, n$ , 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).

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
```

testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
str	-2.279808	.4798256	-4.75	0.000	-3.22298 -1.336637
str	(dropped)				
_cons	698.933	9.467491	73.82	0.000	680.3231 717.5428

Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.

- In the previous regression,  $\beta_1$  is the effect on TestScore of a unit change in STR, holding STR constant (???)
- We will return to perfect (and imperfect) multicollinearity shortly, with more examples

With these least squares assumptions in hand, we now can derive the sampling dist'n of  $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$

# The Sampling Distribution of the OLS Estimator

Under the four Least Squares Assumptions,

- The exact (finite sample) distribution of  $\hat{\beta}_1$  has mean  $\beta_1$ ,  $\text{Var}(\hat{\beta}_1)$  is inversely proportional to  $n$ ; so too for  $\hat{\beta}_2$  .
- 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:  $\hat{\beta}_1 \rightarrow^p \beta_1$
- $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}}$  is approximately distributed  $N(0, 1)$  (CLT)
- So too for  $\hat{\beta}_2, \dots, \hat{\beta}_k$

Conceptually, there is nothing new here!

## Multicollinearity, Perfect and Imperfect

Some more examples of perfect multicollinearity

- The example from earlier: you include STR twice
- Second example: regress *TestScore* on a constant,  $D$ , and  $B$ , where:  $D_i = 1$  if  $STR \leq 20$ ,  $= 0$  otherwise;  $B_i = 1$  if  $STR > 20$ ,  $= 0$  otherwise, so  $B_i = 1 - D_i$  and there is perfect multicollinearity
- Would there be perfect multicollinearity if the intercept (constant) were somehow dropped (that is, omitted or suppressed) in this regression?
- This example is a special case of

## The dummy variable trap

Suppose you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive - that is, there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other). If you include all these dummy variables and a constant, you will have perfect multicollinearity - this is sometimes called the dummy variable trap.

- Why is there perfect multicollinearity here?
- Solutions to the dummy variable trap:
  - 1 Omit one of the groups (e.g. Senior), or
  - 2 Omit the intercept
- What are the implications of (1) or (2) for the interpretation of the coefficients?

## Perfect multicollinearity, ctd.

- Perfect multicollinearity usually reflects a mistake in the definitions of the regressors, or an oddity in the data
- If you have perfect multicollinearity, your statistical software will let you know - either by crashing or giving an error message or by "dropping" one of the variables arbitrarily
- The solution to perfect multicollinearity is to modify your list of regressors so that you no longer have perfect multicollinearity.

## Imperfect multicollinearity

Imperfect and perfect multicollinearity are quite different despite the similarity of the names

- Imperfect multicollinearity occurs when two or more regressors are very highly correlated.
- Why this term? If two regressors are very highly correlated, then their scatterplot will pretty much look like a straight line - they are collinear - but unless the correlation is exactly 1, that collinearity is imperfect.

## Imperfect multicollinearity, ctd.

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

- Intuition: 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 are pretty much uninformative about what happens when  $X_1$  changes but  $X_2$  doesn't, 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.

Next topic: hypothesis tests and confidence intervals

# Hypothesis Tests and Confidence Intervals for a Single Coefficient in Multiple Regression

- $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{Var}(\hat{\beta}_1)}}$  is approximately distributed  $N(0, 1)$  (CLT)
- Thus hypotheses on  $\beta_1$  can be tested using the usual  $t$ -statistic, and confidence intervals are constructed as  $\{\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)\}$ .
- So too for  $\hat{\beta}_2, \dots, \hat{\beta}_k$
- $\hat{\beta}_1$  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

$$(1) \widehat{TestScore} = 698.9 - 2.28 \times STR$$

(9.47) (0.48)

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

(7.41) (0.38) (0.04)

- 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  $\{-1.101.96 \times 0.38\} = (-1.85, -0.35)$
- The t-statistic testing  $\beta_{STR} = 0$  is  $t = -1.10/0.38 = -2.90$ , so we reject the hypothesis at the 5% significance level

## Tests of Joint Hypotheses

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

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

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

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

- 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 cant 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:

$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$  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

$$\begin{aligned} &= Pr[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96] \\ &= Pr[|t_1| > 1.96, |t_2| > 1.96] + Pr[|t_1| > 1.96, |t_2| \leq 1.96] \\ &+ Pr[|t_1| \leq 1.96, |t_2| > 1.96] \quad (\text{disjoint events}) \\ &= Pr[|t_1| > 1.96] \times Pr[|t_2| > 1.96] + Pr[|t_1| > 1.96]Pr[|t_2| \leq 1.96] \\ &+ Pr[|t_1| \leq 1.96] \times Pr[|t_2| > 1.96] \quad (t_1, t_2 \text{ are independent by assumption}) \\ &= .05 \times .05 + .05 \times .95 + .95 \times .05 = .0975 = 9.75\% \\ &- \text{ which is not the desired } 5\%!! \end{aligned}$$

Use a different test statistic that test both 1 and 2 at once: the F-statistic (this is common practice)

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

$p$ -value = tail probability of the  $\chi^2/q$  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 ( $expn_s tu$ ) are both zero, against the alternative that at least one of the population coefficients is nonzero.

## F-test example, California class size data:

Source	SS	df	MS			
Model	66409.8837	3	22136.6279	Number of obs = 420		
Residual	85699.7099	416	206.008918	F( 3, 416) = 107.45		
				Prob > F = 0.0000		
				R-squared = 0.4366		
				Adj R-squared = 0.4325		
				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

```

. test str expn_stu
( 1) str = 0
( 2) expn_stu = 0
      F( 2, 416) = 8.01
      Prob > F = 0.0004
  
```

## The “restricted” and “unrestricted” regressions

Example: are the coefficients on STR and Expn zero?

- Unrestricted population regression (under  $H_1$ ):

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PCtEL_i + \epsilon_i$$

- Restricted population regression (that is, under  $H_0$ ):

$$TestScore_i = \beta_0 + \beta_3 PCtEL_i + \epsilon_i$$

- 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?

## Simple formula for the F-statistic:

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

where

- $R_{unrestricted}^2$  the  $R^2$  for the restricted regression
- $R_{restricted}^2$  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 F.

## Example:

- Restricted Regression:

$$\widehat{\text{Testscore}} = 644.7 - 0.671\text{PctEL}, R^2_{\text{restricted}} = 0.4149$$

- Unrestricted Regression:

$$\widehat{\text{Testscore}} = 649.6 - 0.29\text{STR} + 3.87\text{Expn} - 0.656\text{PctEL},$$

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

- so,

$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)} \sim F(q, n - k - 1)$$
$$\frac{(0.4366 - 0.4149)/2}{(1 - 0.4366)/(420 - 3 - 1)} = 8.01$$

## F-statistic summary

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

- The 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”
- These are justified only under very strong conditions - stronger than are realistic in practice.
- Yet, they are widely used.
- You should use the F-statistic, with  $\chi^2_q/q$  (that is,  $F_{q,\infty}$ ) critical values.
- For  $n \geq 100$ , the F-distribution essentially is the  $\chi^2_q/q$  distribution.
- For small  $n$ , sometimes researchers use the F distribution because it has larger critical values and in this sense is more conservative.