Making sense of Econometrics: Basics

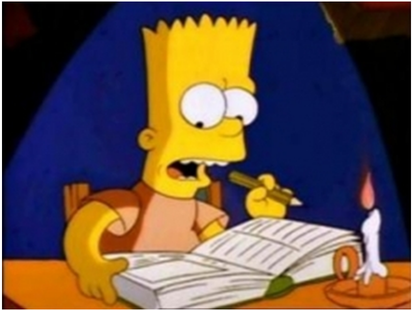
Lecture 7: Multicollinearity

Hany Abdel-Latif & Anita Staneva

Egypt Scholars Economic Society

November 22, 2014

Assignment & feedback



enter classroom at

http://b.socrative.com/login/student/

room name c28efb78

Outline

1. Multicollinearity
   - meaning
   - detection
   - example

nature of multicollinearity

- CLRM assumes no exact relationship among explanatory variables (**A6**)
- perfect multicollinearity
  - an exact relationship amongst the x's
  - is rarely encountered in practice, unless as a result of 'specification error' e.g., dummy variable trap
- imperfect multicollinearity
  - when explanatory variables are highly correlated
  - is a matter of degree
  - typically in macroeconomic time series data

لأن كـل عقل يفرق

perfect multicollinearity

- when there is a perfect linear relationship
- assume we have the following model

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

- where the sample values for X_2 and X_3 are

X_2	1	2	3	4	5	6
X_3	2	4	6	8	10	12

- we observe that $X_3 = 2X_2$
- although it seems we have two explanatory variables in fact it is only one
- X_2 is an exact linear function of X_3
- X_2 and X_3 are perfect collinear

consequences of perfect multicollinearity

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

where $X_{3t} = \lambda X_{2t}$

- every variation in X_{2t} will be paralleled by variation in X_{3t}
- no longer possible to separate the independent influences of the two on Y_t
- substituting for X_{3t} and collecting terms we get

$$\begin{aligned} Y_t &= \beta_1 + (\beta_2 + \lambda\beta_3) X_{2t} + u_t \\ &= \beta_1 + \beta_4 X_{2t} + u_t \end{aligned}$$

where $\beta_4 = \beta_2 + \lambda\beta_3$

consequences of perfect multicollinearity

$$Y_t = \beta_1 + \beta_4 X_{2t} + u_t$$

where $\beta_4 = \beta_2 + \lambda\beta_3$

- in which case β_4 can be estimated, but cannot decomposed to give separate estimates of β_2 and β_3
  - cannot obtain unique estimates of all the parameters
  - cannot conduct hypothesis testing
- OLS cannot be applied

consequences of imperfect multicollinearity

- OLS estimator are still BLUE, if other CLRM assumptions continue to hold
- however, the parameters will not be vary accurately estimated
  - estimated coefficient variances and standard errors will be large
  - t-ratios will be low and confidence interval wider
- if the multicollinearity is strong enough
  - bias towards failing to reject the null hypothesis $H_0 : \beta_j = 0$

detection of multicollinearity

- the classical symptom of strong multicollinearity is high R^2 with low t-ratios for individual coefficients
- no satisfactory formal statistical test exists
- informal tests
  - inspect the correlation coefficients for pair-wise combinations of the explanatory variables
  - run 'auxiliary regressions' of each of the explanatory variables k on $k-1$ other variables and inspect their R^2
  - drop one of the suspected multicollinear variables from the regression and see if the other variables become significant

remedies for imperfect multicollinearity

- drop one or more of the multicollinear variables
  - this solution can introduce specification bias
- transform the multicollinear variables
  - from a linear combination of the multicollinear variables
  - transform the equation into differences or logs
- increase the sample size since multicollinearity is ultimately a 'sample-specific' problem
- 'principal component analysis' or 'ridge regression', beyond the scope of this module

illustrative example

- consumption expenditure Y_i in relation to income X_{2i} and wealth X_{3i}

$$\hat{Y}_i = \underset{(6.7525)}{24.7747} + \underset{(0.8229)}{0.9415X_{2i}} - \underset{(0.0807)}{0.0424X_{3i}}$$

$$t = (3.6690) \qquad (1.1442) \qquad (-0.5261)$$

$$R^2 = 0.9635 \quad \bar{R}^2 = 0.9531 \quad F = 92.4019$$

- highly significant F-value while t-values are individually insignificant
- two variables are highly correlated and it is impossible to isolate the individual impact

illustrative example

- if we regress X_3 on X_2, we obtain

$$\hat{X}_{3i} = \underset{(29.4758)}{7.5454} + \underset{(0.1643)}{10.1909 X_{2i}}$$
$$t = (0.2560) \qquad (62.0405)$$
$$R^2 = 0.9979$$

- which shows there there is almost perfect collinearity between X_3 and X_2

illustrative example

- if we regress Y on X_2 only

$$\hat{Y}_i = \underset{(6.4138)}{24.4545} + \underset{(0.0357)}{0.5091}X_{2i}$$
$$t = (3.8128) \quad (14.2432)$$
$$R^2 = 0.9621$$

- in the first model (with both income and wealth), the income variable was statistically insignificant
- now the income variable is highly significant

examine residuals: informal

- if we regress Y on X_3 only

$$\hat{Y}_i = \underset{(6.874)}{24.411} + \underset{(0.0037)}{0.0498} X_{3i}$$
$$t = (3.551) \quad (13.29)$$
$$R^2 = 0.9567$$

- now wealth has a significant impact on consumption expenditure
- whereas earlier it has no effect on consumption expenditure