迴歸進階主題 1 與 2

1. Omitted Variable Bias (OVB) and 2. Multicollinearity

「缺失變數偏誤」 與 「多元共線性」 (取自 Wooldridge Chap 3)

迴歸進階主題 1: Omitted Variable Bias (OVB)

- 先閱讀以下課文片段

Deriving the bias caused by omitting an important variable is an example of **misspecification analysis**. We begin with the case where the true population model has two explanatory variables and an error term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \qquad [3.40]$$

and we assume that this model satisfies Assumptions MLR.1 through MLR.4.

Suppose that our primary interest is in β_1, the partial effect of x_1 on y. For example, y is hourly wage (or log of hourly wage), x_1 is education, and x_2 is a measure of innate ability. In order to get an unbiased estimator of β_1, we *should* run a regression of y on x_1 and x_2 (which gives unbiased estimators of β_0, β_1, and β_2). However, due to our ignorance or data unavailability, we estimate the model by *excluding* x_2. In other words, we perform a simple regression of y on x_1 only, obtaining the equation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \qquad [3.41]$$

We use the symbol "~" rather than "^" to emphasize that $\tilde{\beta}_1$ comes from an underspecified model.

When first learning about the omitted variable problem, it can be difficult to distinguish between the underlying true model, (3.40) in this case, and the model that we actually estimate, which is captured by the regression in (3.41). It may seem silly to omit the variable x_2 if it belongs in the model, but often we have no choice. For example, suppose that *wage* is determined by

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \qquad [3.42]$$

Since ability is not observed, we instead estimate the model

$$wage = \beta_0 + \beta_1 educ + v,$$

where $v = \beta_2 abil + u$. The estimator of β_1 from the simple regression of *wage* on *educ* is what we are calling $\tilde{\beta}_1$.

As it turns out, we have done almost all of the work to derive the bias in the simple regression estimator of $\tilde{\beta}_1$. From equation (3.23) we have the algebraic relationship $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the slope estimators (if we could have them) from the multiple regression

$$y_i \text{ on } x_{i1}, x_{i2}, i = 1, \ldots, n \qquad [3.43]$$

and $\tilde{\delta}_1$ is the slope from the simple regression

$$x_{i2} \text{ on } x_{i1}, i = 1, \ldots, n. \qquad [3.44]$$

- Slides p.20-22

- **Including irrelevant variables in a regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

No problem because $E(\hat{\beta}_3) = \beta_3 = 0$. ← = 0 in the population

However, including irrevelant variables may increase sampling variance.

- **Omitting relevant variables: the simple case**

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ ← True model (contains x_1 and x_2)

$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1$ ← Estimated model (x_2 is omitted)

- **Omitted variable bias**

$x_2 = \delta_0 + \delta_1 x_1 + v$ ← If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + (\beta_2 v + u)$$

If y is only regressed on x_1 this will be the estimated intercept

If y is only regressed on x_1, this will be the estimated slope on x_1

error term

- **Conclusion: All estimated coefficients will be biased**

- **Example: Omitting ability in a wage equation**

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

$$abil = \delta_0 + \delta_1 educ + v$$

Will both be positive

$$wage = (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)educ + (\beta_2 v + u)$$

The return to education β_1 will be overestimated because $\beta_2\delta_1 > 0$. It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.

- **When is there no omitted variable bias?**
  - If the omitted variable is irrelevant or uncorrelated

- 重要結果整理

$$\begin{aligned} E(\tilde{\beta}_1) &= E(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = E(\hat{\beta}_1) + E(\hat{\beta}_2)\tilde{\delta}_1 \\ &= \beta_1 + \beta_2\tilde{\delta}_1, \end{aligned} \qquad [3.45]$$

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- OVB 例題

EXAMPLE 3.6 Hourly Wage Equation

Suppose the model $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ satisfies Assumptions MLR.1 through MLR.4. The data set in WAGE1 does not contain data on ability, so we estimate β_1 from the simple regression

$$\widehat{\log(wage)} = .584 + .083\, educ$$
$$n = 526,\ R^2 = .186. \qquad [3.47]$$

This is the result from only a single sample, so we cannot say that .083 is greater than β_1; the true return to education could be lower or higher than 8.3% (and we will never know for sure). Nevertheless, we know that the average of the estimates across all random samples would be too large.

- OVB 例題

C6 Use the data set in WAGE2 for this problem. As usual, be sure all of the following regressions contain an intercept.

(i) Run a simple regression of *IQ* on *educ* to obtain the slope coefficient, say, $\tilde{\delta}_1$.

(ii) Run the simple regression of log(*wage*) on *educ*, and obtain the slope coefficient, $\tilde{\beta}_1$.

(iii) Run the multiple regression of log(*wage*) on *educ* and *IQ*, and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

(iv) Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

C3.6 (i) The slope coefficient from the regression *IQ* on *educ* is (rounded to five decimal places) $\tilde{\delta}_1 = 3.53383$.

(ii) The slope coefficient from log(*wage*) on *educ* is $\tilde{\beta}_1 = .05984$.

(iii) The slope coefficients from log(*wage*) on *educ* and *IQ* are $\hat{\beta}_1 = .03912$ and $\hat{\beta}_2 = .00586$, respectively.

(iv) We have $\hat{\beta}_1 + \tilde{\delta}_1\hat{\beta}_2 = .03912 + 3.53383(.00586) \approx .05983$, which is very close to .05984; the small difference is due to rounding error.

- 討論重點:

(1) OVB 會造成怎樣的偏誤? 其偏誤公式為何?
(2) 缺失的變數與既有變數的相關性對偏誤有何影響?

迴歸進階主題 2: Multicollinearity (和 VIF 的討論)

- Slides p.16

- **Assumption MLR.3 (No perfect collinearity)**

  "In the sample (and therefore in the population), none of the independent variables is constant and there are **no exact linear relationships** among the independent variables."

- **Remarks on MLR.3**
  - The assumption only rules out perfect collinearity/correlation between explanatory variables; imperfect correlation is allowed
  - If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated
  - Constant variables are also ruled out (collinear with intercept)

- Slides p.25:

Theorem 3.2 (Sampling variances of the OLS slope estimators)

Under assumptions MLR.1 – MLR.5:

Variance of the error term

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}, \quad j = 1, \ldots, k$$

Total sample variation in explanatory variable x_j:

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

R-squared from a regression of explanatory variable x_j on all other independent variables (including a constant)

- Slides p.27-30:

- The problem of almost linearly dependent explanatory variables is called **multicollinearity** (i.e. $R_j \to 1$ for some j)

- Note that multicollinearity is not a violation of MLR.3 in the strict sense
- Multicollinearity may be detected through "variance inflation factors"

$$VIF_j = 1/(1 - R_j^2)$$

As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10

- **An example for multicollinearity**

Average standardized test score of school

Expenditures for teachers

Expenditures for instructional materials

Other expenditures

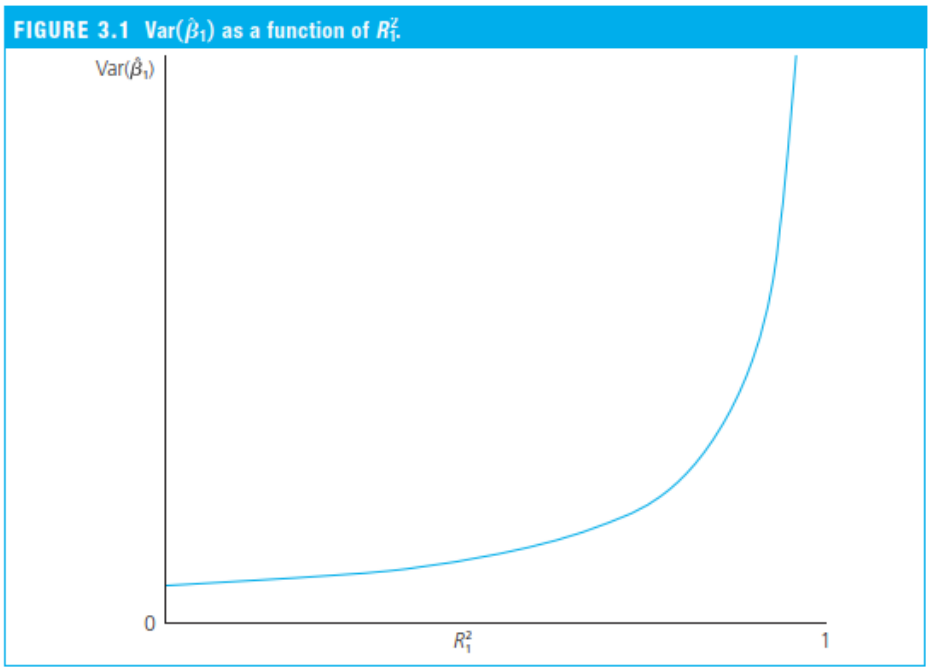
$$avgscore = \beta_0 + \beta_1 teachexp + \beta_2 matexp + \beta_3 othexp + \ldots$$

The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.

It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.

As a consequence, sampling variance of the estimated effects will be large.

- **Discussion of the multicollinearity problem**
  - In the above example, it would probably be better to lump all expen-diture categories together because effects cannot be disentangled
  - In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias)



- Slides p.31, 32, 34:

- The choice of whether to include a particular variable in a regression can be made by analyzing the **tradeoff** between bias and variance

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$ ← True population model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$ ← Estimated model 1

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$ ← Estimated model 2

- It might be the case that the likely omitted variable bias in the misspecified model 2 is overcompensated by a smaller variance

- **Variances in misspecified models (cont.)**

$$Var(\hat{\beta}_1) = \sigma^2 / \left[SST_1(1 - R_1^2)\right]$$

$$Var(\tilde{\beta}_1) = \sigma^2 / SST_1$$

Conditional on x_1 and x_2, the variance in model 2 is always smaller than that in model 1

- **Estimation of the sampling variances of the OLS estimators**

The true sampling variation of the estimated β_j

$$sd(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)} = \sqrt{\sigma^2 / \left[SST_j(1 - R_j^2)\right]}$$

Plug in $\hat{\sigma}^2$ for the unknown σ^2

The estimated sampling variation of the estimated β_j

$$se(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\hat{\sigma}^2 / \left[SST_j(1 - R_j^2)\right]}$$

- Multicollinearity 例題

C11 Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

(i) Run the simple regression of *math*4 on *pctsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?

(ii) Add the variables *lmedinc* and *free* to the equation. What happens to the coefficient on *pctsgle*? Explain what is happening.

(iii) Find the sample correlation between *lmedinc* and *free*. Does it have the sign you expect?

(vi) Does the substantial correlation between *lmedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.

(v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

C3.11 (i) The regression results are:

$$\widehat{math4} = 96.7704 - 0.8328pctsgle.$$

The percentage of children not in the married-couples families has a negative impact on percentage of satisfactory level of 4th grade math. The effect of single parenthood seem small. If, say, *pctsgle* increases by .10 (ten percentage points), the percentage of satisfactory level of 4th grade math is estimated to decrease by .08328 percentage, which is a small effect.

(ii) The estimated regression results are:

$$\widehat{math4} = 51.723 - 0.1996pctsgle - 0.3964free + 3.5601lmedinc.$$

The coefficient of *pctsgle* has negatively increased from -0.8328 to -0.1996. This means that, as the percentage of children not in married couples increases, the percentage of satisfactory level of 4th grade math decreases.

(iii) The sample correlation between *lmedinc* and *free* is -0.74. This is the expected relationship because as the median income increases, the eligibility of the free lunch decreases.

(iv) No, because high correlations among the variables *lmedinc* and *free* do not make it more difficult to determine the causal effect of single parenthood on student performance.

(v) $\text{VIF}_{pctsgle} = \frac{1}{1-R^2} = \frac{1}{1-0.3795} = 1.6116.$
$\text{VIF}_{free} = \frac{1}{1-R^2} = \frac{1}{1-0.4455} = 1.8034.$
$\text{VIF}_{lmedinc} = \frac{1}{1-R^2} = \frac{1}{1-0.3212} = 1.4732.$

By comparing the three variables, it is very clear that the variable *free* has the highest VIF. No, this knowledge does not affect the model to study the causal effect of single parenthood on math performance.

- 討論重點:

(1) 多元共線性會對迴歸模型帶來怎樣的影響?
(有一種說法是: 個別不顯著，但整體顯著，但何謂個別，何謂整體?)
(2) 所謂 VIF 是什麼意義? 怎麼計算它? 高到什麼地步會不可接受?
(3) Variance 是指誰的 variance? 它又與 se 有何關連?
(4) 在「要納進一個變數」與「不納進一個變數」的考量上，要怎樣取捨? (為什麼說 tradeoff between bias and variance?) 請整理自己對此問題的處理原則。