Class 13- Multiple Regression Model Estimation (Part III)

Pedram Jahangiry

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Including irrelevant variables in a regression model (overspecification)

x_3 is an irrelevant variable:

$\beta_3 = 0$ in the population, i.e. X_3 has no partial effect on y

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

Because we do not know that $\beta_3 = 0$, we are inclined to estimate the equation including x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Is there any problem?

including one or more irrelevant variables in a multiple regression model, or overspecifying the model, **does not** affect the **unbiasedness** of the OLS estimators.

However, it can have **undesirable effects** on the **variances** of the OLS estimators. (How? Use Venn Diagram)

Omitting relevant variables (Underspecification)

x_2 is a relevant variable:

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

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\, x_1 + \widehat{\beta_2}\, x_2$$ ← This estimated model should be used

---

$$\tilde{y} = \widetilde{\beta_0} + \widetilde{\beta_1} x_1$$ ← But due to our ignorance or data availability, this estimated model (x_2 is omitted) is used

This is an underspecified model.

Is there any problem? (use Venn Diagram)

Omitting relevant variables – Calculating the 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

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

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

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

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

Error term

All estimated coefficients will be biased (Why?)

Omitted Variable Bias

What is the bias in $\widetilde{\beta_1}$?

$$\mathrm{E}(\tilde{\beta}_1) = \mathrm{E}(\hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1) = \mathrm{E}(\hat{\beta}_1) + \mathrm{E}(\hat{\beta}_2)\tilde{\delta}_1$$
$$= \beta_1 + \beta_2\tilde{\delta}_1$$

which implies the bias in $\tilde{\beta}_1$ is

$$\mathrm{Bias}(\tilde{\beta}_1) = \mathrm{E}(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1$$

if x_1 and x_2 are uncorrelated in the sample, then $\tilde{\beta}_1$ is unbiased.

Direction of the bias:

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

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

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

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

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?

Example: Omitting ability in a wage equation

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.

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

What is the trade off here if we drop one of the explanatory variables (for example *othexp)*?

Discussion of the multicollinearity problem

- In the above example, it would probably be better to lump all expenditure 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)

- Only the sampling variance of the variables involved in multicollinearity will be inflated

- Note that multicollinearity is not a violation of MLR.3

Detecting multicollinearity

Multicollinearity may be detected through Variance Inflation Factors:

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

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)} \longrightarrow \text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j} \cdot \text{VIF}_j$$

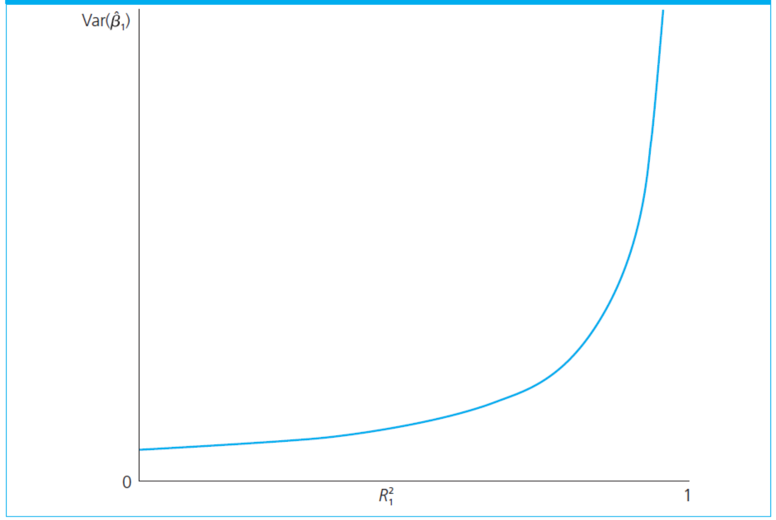

FIGURE 3.1 Var($\hat{\beta}_1$) as a function of R_1^2.

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