LECTURE 7

Introduction to Econometrics

Omitted & Irrelevant Variables

Hieu Nguyen

Fall semester, 2024

Specification of a regression

- We discussed the specification of a regression equation
- **Specification** consists of choosing:
  1. correct independent variables
  2. correct functional form
  3. correct form of the stochastic error term
- A **specification error** occurs if any of these choices is wrong
- In lecture 6, we discussed the correct functional form. Now we will learn how to deal with the other two in today's and the following two lectures.

On today's lecture

- We will talk about the problem of **not adding relevant** independent variables or **adding irrelevant** independent variables

- We will learn that
  - Omitting a relevant variable brings bias to our estimates of the other coefficients
  - Including an irrelevant variable increase the variance of our estimates of the other coefficients
  - Since in real estimation, it is often hard to judge whether or not to include a variable, we need economic theory and statistical tools to decide

OMITTING RELEVANT VARIABLES

- We omit a variable when we
  - forget to include it
  - do not have data for it
- This misspecification results in
  - not having the coefficient for this variable
  - biasing estimated coefficients of other variables in the equation −→ **omitted variable bias**

Omitted variables

- Where does the omitted variable bias come from?
- True model:

$$y_i = \beta x_i + \gamma z_i + u_i$$

- Model as it looks when we omit variable z:

$$y_i = \beta x_i + \tilde{u}_i$$

  implying

$$\tilde{u}_i = \gamma z_i + u_i$$

- We assume that $Cov(u_i, x_i) = 0$. But it does not guarantee that $Cov(\tilde{u}_i, x_i) = 0$ is also true. Because of correlation among independent variables, it can be

$$Cov(\tilde{u}_i, x_i) = Cov(\gamma z_i + u_i, x_i) = \gamma Cov(z_i, x_i) \neq 0$$

- The classical assumption is violated ⇒ biased (and inconsistent) estimate!!!

OMITTED VARIABLES

- For the model with omitted variable:

$$E\left(\beta^{omitted\ model}\right) = \beta + bias$$

- The bias is

$$bias = \gamma \times \alpha$$

  - Coefficients β and γ are from the true model:

$$y_t = \beta x_t + \gamma z_t + u_t$$

  - Coefficient α is from a regression of z on x:

$$z_t = \alpha x_t + e_t$$

- The bias disappears only if either $\gamma = 0$ or $\alpha = 0$.
  - If $\gamma = 0$, it is no more the problem of omitting relevant variable
  - Mostly $\alpha \neq 0$ because non-perfect collinearity among independent variables.

OMITTED VARIABLES

- **Intuitive explanation:**
  - if we leave out an important variable from the regression ($\gamma \neq 0$), coefficients of other variables are biased unless the omitted variable is uncorrelated with all included dependent variables ($\alpha \neq 0$).
  - the included variables pick up some of the effect of the omitted variable (if they are correlated), and the coefficients of included variables thus change causing the bias.
  - Example: what would happen if you estimated a production function with capital only and omitted labour?

OMITTED VARIABLES

- Example: estimating the price of chicken meat in the US

$$\hat{Y}_t = 31.5 - \underset{(0.08)}{0.73}\ PC_t + \underset{(0.05)}{0.11}\ PB_t + \underset{(0.02)}{0.23}\ YD_t$$

$$R^2 = 0.986 \quad , \quad n = 44$$

Y_t ... per capita chicken consumption
PC_t ... price of chicken
PB_t ... price of beef
YD_t ... per capita disposable income

OMITTED VARIABLES

- When we omit price of beef:

$$\hat{Y}_t = 32.9 - \underset{(0.08)}{0.70}\ PC_t + \underset{(0.01)}{0.27}\ YD_t$$

$$R^2 = 0.895 \quad , \quad n = 44$$

- Compare to the true model:

$$\hat{Y}_t = 31.5 - \underset{(0.08)}{0.73}\ PC_t + \underset{(0.05)}{0.11}\ PB_t + \underset{(0.02)}{0.23}\ YD_t$$

$$R^2 = 0.986 \quad , \quad n = 44$$

- We observe positive bias in the coefficient of *PC* (was it expected?)

Omitted variables

- **Determining the direction of bias: bias= $\gamma * \alpha$**
  - Where γ is a correlation between the omitted variable and the dependent variable (the price of beef and chicken consumption)
  - γ is likely to be positive

  - Where α is a correlation between the omitted variable and the included independent variable (the price of beef and the price of chicken)
  - α is likely to be positive

- **Conclusion: Bias in the coefficient of the price of chicken is likely to be positive if we omit the price of beef from the equation.**

OMITTED VARIABLES

- In reality, we usually do not have the true model to compare with

  Because we do not know what the true model is

  Because we do not have data for some important variable

- We can often recognize the bias if we obtain some unexpected results

- We can prevent omitting variables by relying on the theory

- If we cannot prevent omitting variables, we can at least determine in what way this biases our estimates

IRRELEVANT VARIABLES

- A second type of specification error is including a variable that does not belong to the model

- This misspecification
  - does not cause bias
  - but it increases the variances of the estimated coefficients of the included variables

Irrelevant variables

- True model:

$$y_i = \beta x_i + u_i \tag{1}$$

- **Model as it looks when we add irrelevant z:**

$$y_i = \beta x_i + \gamma z_i + \tilde{u}_i \tag{2}$$

- **We can represent the error term as $\tilde{u}_i = u_i - \gamma z_i$**

- **but since from the true model $\gamma = 0$, we have $\tilde{u}_i = u_i$ and there is no bias**
- **But the problem:**

$$Var(\beta^{(2)}) = \frac{\sigma^2}{(1-r_{XZ}^2)\sum_i x_i^2} > \frac{\sigma^2}{\sum_i x_i^2} = Var(\beta^{(1)})$$

Irrelevant variables

- **True model:**

$$\hat{Y}_t = 31.5 - \underset{(0.08)}{0.73 PC_t} + \underset{(0.05)}{0.11 PB_T} + \underset{(0.02)}{0.23 YD_t}$$

$$R^2 = 0.986, \quad n = 44$$

- **If we include irrelevant variable interest rate R_t**

$$\hat{Y}_t = 30.0 - \underset{(0.10)}{0.73 PC_t} + \underset{(0.06)}{0.12 PB_T} + \underset{(0.03)}{0.22 YD_t} + \underset{(0.21)}{0.17 R_t}$$

$$R^2 = 0.987, \quad n = 44$$

- **We observe that R_t is insignificant and standard errors of other variables increase**

Summary of the theory

- Bias – efficiency trade-off:

	Omitted variable	**Irrelevant variable**
Bias	Yes*	No
Variance	Decreases *	Increases*

* As long as we have correlation between x and z

Four important specification criteria

Does a variable belong to the equation?

1. *Theory:* Is the variable's place in the equation unambiguous and theoretically sound? Does intuition tells you it should be included?

2. *t-test:* Is the variable's estimated coefficient significant in the expected direction?

3. R^2*:* Does the overall fit of the equation improve (enough) when the variable is added to the equation?

4. *Bias:* Do other variables' coefficients change significantly when the variable is added to the equation?

FOUR IMPORTANT SPECIFICATION CRITERIA

- If all conditions hold, the variable belongs in the equation
- If none of them holds, the variable is irrelevant and can be safely excluded
- If the criteria give contradictory answers, most importance should be attributed to theoretical justification
- Therefore, if theory (intuition) says that variable belongs to the equation, we include it (even though its coefficients might be insignificant!).

EXAMPLE FOR SPECIFICATION CRITERIA

- **Examining the price elasticity of Brazilian coffee**

$$\widehat{COF} = 9.1 + 7.8\,P_{BC} + 2.4\,P_T + 0.0035\,Y$$
$$\qquad\qquad (15.6) \qquad (1.2) \qquad (0.0010)$$

$$R^2 = 0.60\,, \qquad n = 25$$

COF ... Brazilian coffee consumption
P_{BC} ... price of Brazilian coffee
P_T ... price of tea
Y ... disposable income

EXAMPLE FOR SPECIFICATION CRITERIA

- **Compare the following two regressions:**

$$\widehat{COF} = 9.3 \qquad\qquad + \underset{(1.0)}{2.4}\, P_T + \underset{(0.0009)}{0.0036}\, Y$$

$$R^2 = 0.58\,, \qquad n = 25$$

$$\widehat{COF} = 9.1 + \underset{(15.6)}{7.8}\, P_{BC} + \underset{(1.2)}{2.4}\, P_T + \underset{(0.0010)}{0.0035}\, Y$$

$$R^2 = 0.60\,, \qquad n = 25$$

- **It seems almost all four criteria in this case does not hold (except theory), P_{BC} is irrelevant variable, and we will conclude that Brazilian coffee is price inelastic.**

EXAMPLE FOR SPECIFICATION CRITERIA

- **But what if we add variable price of Colombian coffee (P_{CC})?**

$$\widehat{COF} = 10.0 + 8.0\,P_{CC} - 5.6\,P_{BC} + 2.6\,P_T + 0.0030\,Y$$
$$\qquad\qquad (4.0) \qquad (2.0) \qquad (1.3) \qquad (0.0009)$$

$$R^2 = 0.70\,, \qquad n = 25$$

$$\widehat{COF} = 9.1 + 7.8\,P_{BC} + 2.4\,P_T + 0.0035\,Y$$
$$\qquad\qquad (15.6) \qquad (1.2) \qquad (0.0010)$$

$$R^2 = 0.60\,, \qquad n = 25$$

- It seems almost all four criteria in this case hold, P_{CC} and P_{BC} **are relevant variables**, and we will **conclude** that Brazilian coffee is **price elastic**.

THE DANGER OVERSPECIFICATION

- ”If you just torture the data long enough, they will confess.”
- If too many specifications are tried:
  - The final result may have the desired properties only by chance
  - The statistical significance of the result is overestimated because the estimations of the previous regressions are ignored.
- How to solve this issue:
  - Keep the number of try of regressions low
  - Focus on theory (very important)
  - Save all regression you tried

Specification test

- **Ramsey's Regression Specification Test (RESET)**
  - Allows to detect possible misspecification
  - But cannot detect the source of misspecification
  - Two types of test based on the same intuition:
    - If the equation is correctly specified, nothing is missing in the equation and the residuals are **white noise.**

- **Assume we have:**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

RESET TYPE I

1. Run the regression: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
2. Save the predicted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$
3. Run the augmented regression:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \varepsilon_i$$

(more power can be included)

4. Test a standard F test with null hypothesis $\gamma_1 = \gamma_2 = 0$
    - If we can reject the null hypothesis, there is a misspecification problem in the model
    - Intuition: if the model is correct, y is well explained by x_i and z_i and addition of the predicted values raised to higher powers should not be significant.

RESET TYPE II

1. Run the regression: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$
2. Save the predicted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i$ and the residuals $e_i = y_i - \hat{y}_i$
3. Run the regression:

$$e_i = \alpha_0 + \alpha_1 \hat{y}_i + \alpha_2 \hat{y}_i^2 + \varepsilon_i$$

(more power can be included)

4. Test the null hypothesis $\alpha_1 = \alpha_2 = 0$ using F test
   - If we can reject the null hypothesis, there is a misspecification problem in the model
   - Intuition: if the model is correct, residuals should not display any pattern depending on the independent variables.

Summary

- Omitting a relevant variable brings bias to our estimates of the other coefficients
- Including an irrelevant variable increase the variance of our estimates of the other coefficients
- Since in real estimation, it is often hard to judge whether or not to include a variable, we need economic theory and statistical tools to decide