LECTURE 9

Introduction to Econometrics

Choosing explanatory variables

November 15, 2016

What we have learned so far

- We know what a linear regression model is and how its parameters are estimated by OLS
- We know what the properties of OLS estimator are
- We know how to test single and multiple hypotheses in linear regression models
- We know how to asses the goodness of fit using R^2
- We started to talk about the specification of a regression equation

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
- We discussed the choice of functional form on the previous lecture
- We will discuss the choice of independent variables today
- We will study the form of the error term on the next two lectures

On today's lecture

- We will learn that
  - omitting a relevant variable from an equation is likely to bias remaining coefficients
  - including an irrelevant variable in an equation leads to higher variance of estimated coefficients
  - our choice should be led by the economic theory and confirmed by a set of statistical tools

Omitted 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 $\longrightarrow$ **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:

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

- The classical assumption is violated $\Rightarrow$ biased (and inconsistent) estimate!!!

Omitted variables

- For the model with omitted variable:

$$E(\widehat{\beta}^{\text{omitted model}}) = \beta + \text{bias}$$

$$\text{bias} = \gamma * \alpha$$

  - Coefficients β and γ are from the true model

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

  - Coefficient α is from a regression of z on x, i.e.

$$z_i = \alpha x_i + e_i$$

- The bias is zero if $\gamma = 0$ or $\alpha = 0$ (not likely to happen)

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 labor?

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

Irrelevant variables

- True model:

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

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

- If we include interest rate R_t (irrelevant variable)

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

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

- 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!).

Estimating price elasticity of Brazilian coffee

- Should we include the price of Brazilian coffee into the equation?

$$
\begin{aligned}
\widehat{COF} &= 9.3 + \underset{(1.0)}{2.6}\ P_T + \underset{(0.0009)}{0.0036}\ Y \\
t &= \qquad 2.6 \qquad 4.0 \\
R^2 &= 0.58 \quad , \quad n = 25
\end{aligned}
$$

$$
\begin{aligned}
\widehat{COF} &= 9.1 + \underset{(15.6)}{7.8}\ P_{BC} + \underset{(1.2)}{2.4}\ P_T + \underset{(0.0010)}{0.0035}\ Y \\
t &= 0.5 \qquad 2.0 \qquad 3.5 \\
R^2 &= 0.60 \quad , \quad n = 25
\end{aligned}
$$

- The three criteria does not hold (theory is inconclusive) $\Rightarrow$ the price of Brazilian coffee does not belong to the equation (Brazilian coffee is price inelastic)

Estimating price elasticity of Brazilian coffee

- Really???
- What if we add price of Colombian coffee (P_{CC})?

$$
\begin{aligned}
\widehat{COF} &= \underset{}{10.0} + \underset{(4.0)}{8.0}\,P_{BC} - \underset{(2.0)}{5.6}\,P_{CC} + \underset{(1.3)}{2.6}\,P_T + \underset{(0.0010)}{0.0030}\,Y \\
t &= 2.0 \qquad -2.8 \qquad 2.0 \qquad 3.0 \\
R^2 &= 0.70 \quad , \quad n = 25
\end{aligned}
$$

$$
\begin{aligned}
\widehat{COF} &= 9.1 + \underset{(15.6)}{7.8}\,P_{CC} + \underset{(1.2)}{2.4}\,P_T + \underset{(0.0010)}{0.0035}\,Y \\
t &= \qquad 0.5 \qquad 2.0 \qquad 3.5 \\
R^2 &= 0.60 \quad , \quad n = 25
\end{aligned}
$$

- The three criteria hold $\Rightarrow$ the price of Brazilian coffee belongs to the equation!!! (Brazilian coffee is price elastic)

The danger of specification searches

- "If you just torture the data long enough, they will confess."
- If too many specifications are tried:
  - The final result has desired properties only by chance
  - The statistical significance of the results is overestimated because the estimations of the previous regressions are ignored
- How to proceed:
  - Keep the number of regressions estimated low
  - Focus on theoretical considerations: leave the insignificant variables in the equation if the theory predicts they should be included
  - Document all specifications investigated

Additional specification test

- Ramsey's Regression Specification Error Test (RESET)
  - allows to detect possible misspecification - tells you if all important variables are included or not
  - unfortunately does not allow to detect its source
- There are two forms of this test, both based on similar intuition:
  - If the equation is correctly specified, nothing is missing in the equation and the residuals are a white noise.
- We will derive the test for the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

RESET I

1. We run the regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
2. We save the predicted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2}$
3. We run an augmented regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 \widehat{y}_i^2 + \gamma_2 \widehat{y}_i^3 + \varepsilon_t$$

   (more powers of $\widehat{y}$ can be included)

4. We test $H_0: \gamma_1 = \gamma_2 = 0$ using a standard F-test.
5. If we reject H_0, there is a misspecification problem in our model.

- Intuition: If the model is correct, y is well explained by x_1 and x_2 and the predicted values of y (raised to higher powers) should not be significant.

RESET II

1. We run the regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
2. We save the predicted values $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2}$ and the residuals $e_i = y_i - \widehat{y}_i$
3. We run the regression

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

   (more powers of $\widehat{y}$ can be included)

4. We test $H_0 :\ \alpha_1 = \alpha_2 = 0$ using a standard F-test.
5. If we reject H_0, there is a misspecification problem in our model.

- Intuition: if the model is correct, residuals should not display any pattern depending on the explanatory variables.

Summary

- Omitted variable causes bias (and decreases variance)
  - sign of this bias can be predicted
- Included irrelevant variable increases variance (but does not cause bias)
  - such variable is insignificant in the regression
  - it does not contribute to the overall fit of the regression
- There is a set of criteria that help us to recognize correct specification
  - these criteria have to be applied with caution - theoretical justification has always priority over statistical properties
- Readings:
  - Studenmund Chapter 6, Wooldridge Chapter 9