Erasmus School of Economics

MOOC Econometrics

Lecture 3.2 on Model Specification: Specification

Michel van der Wel

Erasmus University Rotterdam

Bias-efficiency trade-off

Setting:

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

or

$$y = X\beta + \varepsilon$$

in matrix form.

Which variables should we include in X?

- Too few variables $\rightarrow$ Bias.
- Too many variables $\rightarrow$ Efficiency loss.
  (Even if all variables really matter!)

Consequences of omitting variables

DGP: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad \rightarrow \quad b_1$ and b_2

Model: $y = X_1\beta_1 + \tilde{\varepsilon} \quad \rightarrow \quad b_R$

Test

Express $E(b_R)$ as function of β_1 and β_2.

Answer:

$$\begin{aligned} E(b_R) &= E((X_1'X_1)^{-1}X_1'y), \\ &= E((X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)), \\ &= E((X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon)), \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + 0. \end{aligned}$$

Consequences of omitting variables

DGP: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad \rightarrow \quad b_1$ and b_2

Model: $y = X_1\beta_1 + \tilde{\varepsilon} \quad \rightarrow \quad b_R$

It holds:

- $E(b_R) = \beta_1 + \underbrace{(X_1'X_1)^{-1}X_1'X_2}_{P}\beta_2 = \beta_1 + P\beta_2$

  $\rightarrow$ Bias if $\beta_2 \neq 0$ (omitted variable bias).

- $Var(b_R) = Var(b_1) - PVar(b_2)P'$

  $\rightarrow$ Variance of b_R is smaller than that of b_1 (even if $\beta_2 = 0$!).

Decision metrics

Possible decision metrics:

- Information criteria
- Out-of-sample prediction

Information criteria

Commonly used information criteria:

- Akaike: $AIC = \log(s^2) + \frac{2k}{n}$
- Bayes: $BIC = \log(s^2) + \frac{k \log n}{n}$

with s the standard error of the regression and k the number of variables.

Test

Which information criterion imposes the strongest penalty on the number of variables?

Answer: Penalty is $2/n$ for AIC and $\log(n)/n$ for BIC; BIC imposes stronger penalty if $\log(n) > 2$, $n \geq 8$.

Out-of-sample prediction

Commonly used out-of-sample prediction metrics:

- $RMSE = \left(\frac{1}{n_f} \sum_{i=1}^{n_f} (y_i - \hat{y}_i)^2 \right)^{1/2}$
- $MAE = \frac{1}{n_f} \sum_{i=1}^{n_f} |y_i - \hat{y}_i|$

with n_f the number of observations "saved" for out-of-sample evaluation and $\hat{y}_i$ the i-th predicted value of the dependent variable.

Iterative selection methods

Commonly used methods to select explanatory variables:

- t-test and F-test
- Information criteria
- Out-of-sample predictions

Also iterative methods (based on tests) are commonly used:

- General-to-specific / backward elimination
- Specific-to-general / forward selection

TRAINING EXERCISE 3.2

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).