Erasmus
School of
Economics

MOOC Econometrics

Lecture 3.3 on Model Specification: Transformation

Michel van der Wel

Data transformation

Setting:

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

or

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

in matrix form.

What is the most appropriate form of the data (y and X)?

- All variables should be incorporated in a compatible manner.
- If this is not the case, data can be transformed.

Taking logarithms

Use for:

- Exponential growth.

Taking differences

Use for:

- Trending patterns

  $\rightarrow$ Statistical assumptions may not hold.

First difference:

$$\Delta y_i = y_i - y_{i-1}.$$

Test

What is the result if you take the difference of $y_i = i$?

Answer: For $y_i = i$ the difference is:

$$\Delta y_i = y_i - y_{i-1} = i - (i-1) = 1.$$

Non-linear effects

$$y_i = x_i'\beta + \varepsilon_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

has linear set-up with fixed marginal effects ($dy_i/dx_{ji} = \beta_j$).

Extension with interaction and quadratic terms:

$$y_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_{jj} x_{ji}^2 + \sum_{j=2}^{k} \sum_{h=j+1}^{k} \gamma_{jh} x_{ji} x_{hi} + \varepsilon_i.$$

Advantages:

- Get non-linear functional form.
- May provide economically meaningful specification.

Example (from lecture 2.2)

- $\log(\text{Wage})_i =$

  $\beta_1 + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i + \varepsilon_i.$

- $\log(\text{Wage})_i =$

  $\beta_1 + \beta_2 \text{Female}_i + \beta_3 \text{Age}_i + \beta_4 \text{Educ}_i + \beta_5 \text{Parttime}_i +$

  $\gamma_1 \text{Female}_i \text{Educ}_i + \gamma_2 \text{Age}_i^2 + \varepsilon_i.$

Now:

- Wage differential may depend on education ($\beta_2 + \gamma_1 \text{Educ}_i$).
- Age has non-linear effect for wage ($\beta_3 + 2\gamma_2 \text{Age}_i$).

Dummy variables

Quarterly data

$$y_i = \alpha_i + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where α_i is the quarter-specific mean level.

Use dummy variables:

- D_{hi} for $h = 1, 2, 3, 4$, with $D_{hi} = 1$ if observation i is in quarter h (and $D_{hi} = 0$ otherwise).
- Then

$$y_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i.$$

Dummy variables

$$y_i = \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i.$$

Test

Can we add a constant term to the above specification that has a dummy for each quarter?

Answer: Only if one of the dummy parameters is set to 0.

If we omit D_{1i}, so $\alpha_1 = 0$, we get

$$y_i = \alpha_1 + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \gamma_4 D_{4i} + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i,$$

where $\gamma_h = \alpha_h - \alpha_1$ for $h = 2, 3, 4$.

TRAINING EXERCISE 3.3

- 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).