Copenhagen Business School
HANDELSHØJSKOLEN

Applied Econometrics for Researchers

Dummy Variables and Moderation Effects

H.C. Kongsted
Copenhagen Business School, SÍ
hck.si@cbs.dk

Outline of lecture

- Qualitative information in the linear regression
  - Dummies and categorical variables
- Interaction effects:
  - Dummy variables and continuous variables
  - How to test interaction
  - How to interpret the estimated parameters
  - Marginal effects
  - Plotting interaction effects
- Empirical example: Innovation in British firms, rev.

Copenhagen Business School
HANDELSHØJSKOLEN

Including Qualitatitive Information in the Regression Model

Dummy variables

Copenhagen Business School
HANDELSHØJSKOLEN

- A dummy variable is a quantitative expression of a qualitative property of an observational unit - for example labour market status.
- Define STATUS. It has two values. Assume it's coded 1 for employed, 0 for others
- In a regression model, the parameter of the dummy expresses the impact of the dummy being one (employed) compared to the zero value (all other persons): non-employed persons are chosen as the reference category (baseline category)
- Could equally well have chosen employed as the reference: But then the interpretation of the dummy variable STATUS is changed.
- In fact, EMPLOYED would be a MUCH better name for this dummy as we have defined it here.
- Two categories: One dummy variable (and a base category).

Dummies and categorical variables

- A binary variable EMPLOYED may not be all that useful (what is included by the reference category?)
- Discrete variables with more than 2 categories e.g. High/Medium/Low, Industry (manufacturing/service/public),…
- General: *m* levels.
- Order of categories may be of essence (ordinal data) or not (nominal data)
- Categorical variables are split into as many dummy variables as levels (*m*). But one is left out of the equation to define the reference category: *m-1* dummies included
- Stata provides *factor variables:* an expansion which will expand terms containing categorical data into dummy variables:
  – reg y i.catvar x1 x2

Example

- We want to investigate if the price (P) of a flat is determined by a (continuous measure of) size in square metres (M) and location, where location is a categorical variable: inner city (I), suburbs (S), countryside (C)

$$P_i = b_0 + b_M M_i + b_I I_i + b_C C_i + u_i \qquad (1)$$

- We have included ***two*** dummy variables (0/1), one for each of I and C. Note that the suburb location is left out -- it would be a perfectly linear combination of the other two dummies and the intercept of the regression – exact multicollinearity
- Suburb flats are treated as the reference category
- For a countryside flat (*C=1, I=0)* we have the relationship:

$$P_i = (b_0 + b_C) + b_M M_i + u_i \qquad (2)$$

The dummy effect

Copenhagen Business School
HANDELSHØJSKOLEN

- For an inner city flat (*C=0, I =1)*

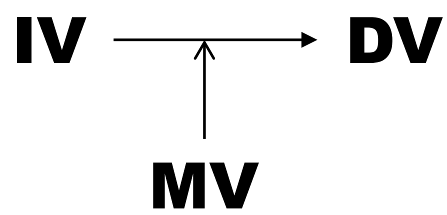
$$P_i = (b_0 + b_I) + b_M M_i + u_i \qquad (3)$$

- A dummy **shifts** the intercept of the function
- A dummy **does not** change the slope of the function
- Positive parameter estimate shifts the function upwards for observations within that category (”inner city premium”)
- Negative parameter estimate shifts the function downwards for observations in the category

Possibility of different slopes: Interactions

Why interactions?

- Theoretically, we may think that a moderating variable affects the magnitude of the effect of the IV on the DV



- A moderator affects the strength or even the sign of the relation between an IV and the DV
- The moderator interacts with the IV to predict the outcome → for *different* levels of a moderator we predict different impacts on the DV of a *given* change of the IV
- In terms of a linear regression, moderation exists if there is a significant <u>interaction</u> in addition to the "main" effects.

What kinds of variables are used?

- Moderator variable (MV) can be:
  - Continuous variables (e.g., size; R&D investments) or
  - Categorical variables (e.g., gender; country)
- We focus on:
  - An IV that is continuous
  - A MV that is either continuous or a dummy variable
  - A DV that is continuous (linear regression models)
- Categorical DV is discussed in the sessions on logits and probits

Moderation and dummy variables

- The slope of the regression function could be different between different categories of a moderating dummy variable
- Marginal change in y associated with a unit change in x depends on the observation's categorical placement
- In the example, the price of a flat may increase more steeply as the size of the flat increases in the inner city than in the countryside ("inner city premium on space").
- We investigate this with interactions

$$P_i = b_0 + b_M M_i + b_I I_i + b_C C_i + b_{MI}(M_i \times I_i) + b_{MC}(M_i \times C_i) + u_i \qquad (5)$$

- For the countryside flat (*C=1, I=0*) we now have:

$$P_i = (b_0 + b_C) + (b_M + b_{MC})M_i + u_i \qquad (6)$$

Moderation and dummy variables: interpretation

- The interaction effect "tilts" the regression line
- Depending on the magnitudes of the main estimate and the interaction estimate, we see different slope adjustments
  - Ex 1: If $b_M > 0$; $b_{MC} > 0 \rightarrow$ More steep positive slope
  - Ex 2: If $b_M < 0$; $b_{MC} < 0 \rightarrow$ More steep negative slope

- *Questions:*
- *What if one is positive and the other is negative?*
- *What is the role of the reference category here?*

Continuous moderator

- Y, X, and Z are continuous variables
- Base model: $Y=\alpha_0 + \alpha_1 X + \alpha_2 Z + \varepsilon$
- If the relationship of X and Y is moderated by Z, we need an extended model:

$$Y= \beta_0+ \beta_1\mathbf{X} + \beta_2\mathbf{Z} + \beta_3\mathbf{X*Z} + \varepsilon$$

- Model with **main** effects and **interaction** effect
- Note: This is still a linear regression model (linear in the β-parameters). But: How to interpret the parameters?

Continuous moderator: Example

- Innovation in British firms revisited
- In workshop 2, you are asked to consider *rdint* as a potential moderator and check whether it significantly moderates the relationship between *extsource* and *prodnew*.
- Base model (simplified): $prodnew = a_0 + a_1 extsource + a_2 rdint + u$
- If the relationship of *prodnew* and *extsource* is moderated by *rdint:*
- Extended: $prodnew = c_0 + c_1 extsource + c_2 rdint + \mathbf{c_3 extsource * rdint} + u$
- Model with **main** effects and **interaction** effect

Partial effects in base model

- Main effects in base model:

- $\partial Y/\partial X = \alpha_1$ The predicted change in Y of increasing one unit X, holding fixed all other factors affecting Y

- $\partial Y/\partial Z = \alpha_2$ The predicted change in Y of increasing one unit Z, holding fixed all other factors affecting Y

- What happens to the interpretation of the partial effect of X on Y when we add to the regression the interaction term X*Z, the product of the two variables?

Partial effects in interacted model

Copenhagen Business School
HANDELSHØJSKOLEN

- $\partial Y/\partial X = \beta_1 + \beta_3 Z$ Depends on value of Z: the effect of X on Y has to be evaluated at a "relevant" value of Z
- $\partial Y/\partial Z = \beta_2 + \beta_3 X$ Depends on value of X: the effect of Z on Y has to be evaluated at a "relevant" value of X
- Then: How can we interpret β_1, β_2, β_3?

Interpretation of estimated parameters

Copenhagen Business School
HANDELSHØJSKOLEN

- Main effects:

  β_1= the effect of X on Y **when Z=0**

  β_2= the effect of Z on Y **when X=0**

- Interaction effect:

  β_3 = the change in the slope of Y on X (Z) given a one unit change in Z (X)

Interpretation of main effects

- In many cases, the effect of X on Y when Z=0 (or Z on Y when X=0) is clearly not of interest: Z=0 could be “out of range”
- What does it mean to know the effect of the distance to city centre (in km) on house prices when the house size (in square meters) is zero?
- Product innovation: main effect of *extsources* is the partial effect for a firm that has no internal R&D. Relevant y/n?
- **Reparameterize** the model to obtain main effects with an interesting meaning -> Workshop 2
- Often we want to evaluate the effect of X on Y when Z is at its sample average

Copenhagen Business School
HANDELSHØJSKOLEN

Marginal effects

Marginal effects with interactions(1)

- The marginal effect of X on Y in model with interaction is $\partial Y/\partial X = \beta_1 + \beta_3 Z$
- We might be interested in evaluating the marginal effect of X on Y at many different values of Z, not only at its mean
- Manual calculation; but Stata has automated procedure
- The routine “**margins**” in Stata

Marginal effects with interactions(2)

- Check the help of “margins”
  - Keep track of interactions with #
  - Note: Difference between factor (i.) and continuous (c.) variables
- Innovation example (simplified):

  reg prodnew extsource rdint c.rdint#c.extsource

  margins, dydx(extsource)

  - Use the option “**at()**” to set the appropriate value(s) of the covariate(s)
  - Min, max, mean, mean+1sd, mean-1sd

Plotting the interaction effect

Copenhagen Business School
HANDELSHØJSKOLEN

- We are interested in understanding how the moderator affects the relationship of the IV and the DV
- Focus on the slope → the partial effect of X on Y moderated by Z
- In terms of original specification: $\partial Y/\partial X = \beta_1 + \beta_3 Z$
- Use the routine "**marginsplot**" in Stata
- Will plot the outcome of the most recent "**margins**" command

Plotting the interaction effect

Copenhagen Business School
HANDELSHØJSKOLEN