LECTURE #8

Econometrics I

PREDICTION & QUALITATIVE INFORMATION

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Summer semester 2024, April 16

In the previous lecture #7

- We focused on data scaling:

$$\hat{y} = \hat{\beta}_0 + \boxed{\frac{\hat{\beta}_1}{c}} c x_1 + \hat{\beta}_2 x_2 \qquad c\hat{y} = \boxed{c\hat{\beta}_0} + \boxed{c\hat{\beta}_1} x_1 + \boxed{c\hat{\beta}_2} x_2$$

  - variables in **logs** rescaled: **slopes not affected**, only the intercept changes.
- More on **log-level**, **level-log** functional forms, **quadratics**.
- We introduced models with **interaction terms**.
- We defined the adjusted $\bar{R}^2$ that controls for the number of explanatory variables: $\boxed{\bar{R}^2 = 1 - \frac{\frac{SSR}{n-k-1}}{\frac{SST}{n-1}}}$.
- We summarized **four important variable selection criteria**:
  1. theory, 2. OVB reduction, 3. $\bar{R}^2$, 4. t/F test.
- Readings for lecture #8:
  - Chapter 6: 6.4, Chapter 7: 7.1–3

Outline

Prediction and residual analysis

Multiple regression with qualitative information
- Single dummy independent variable
- Using dummy variables for multiple categories

Outline

Predictions in OLS cross-sectional data framework

- Up till now, we have been dealing with estimating the parameters, testing their significance, and judging the performance and quality of the model.
- What if we estimate our model on n observations and we want to estimate $\hat{y}_{n+1}$ (having $x_{n+1,1}, \ldots, x_{n+1,k}$)?
- In the cross-sectional data framework, this is referred to as **predicting** (not to be confused with **forecasting**).

Predicted/expected value

- We have the estimated model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k.$$

- **Predicted value** of y_{n+1} given a new set of explanatory variables is simply $\hat{y}_{n+1}$, i.e., we have the estimates $\hat{\beta}_j$ and we have the levels of the explanatory variables $x_{n+1,1}, \ldots, x_{n+1,k}$, which we simply input to the regression model above.
- However, this is only a point estimate that is subject to **sampling variation** as obtained using the OLS estimator.
- We thus need to construct confidence intervals around such estimated $\hat{y}_{n+1}$ to represent the uncertainty of the prediction.

Confidence intervals for predictions

- Let us have specific realizations of explanatory variables and label them $c_1, \ldots, c_k$ for $x_1 \ldots, x_k$, respectively.
- We are thus estimating a specific parameter, which can be written as

$$\theta \equiv \mathbb{E}(y|x_1 = c_1, \ldots, x_k = c_k) = \beta_0 + \beta_1 c_1 + \ldots + \beta_k c_k. \qquad (1)$$

- Let us use the same 'trick' as we did for testing hypotheses about linear combinations of parameters: rewrite the previous equation as

$$\beta_0 = \theta - \beta_1 c_1 - \ldots - \beta_k c_k,$$

plug it in $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$ and estimate the resulting model

$$y = \theta + \beta_1(x_1 - c_1) + \ldots + \beta_k(x_k - c_k) + u.$$

- The predicted value is then the intercept of the new model, i.e., not only do we get the estimate but also the corresponding **standard error** and, in turn, the **confidence intervals**.

Minimal variance for predictions

- Recall the variance of intercept in the simple regression model:

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- When is it minimal? What does it imply?
- When estimating θ, our explanatory variable becomes $x - c$.
- Rewrite the variance for θ as

$$\mathrm{Var}(\hat{\theta}) = \frac{\sigma^2 \sum (x_i - c)^2}{n \sum (x_i - c - \bar{x} + c)^2} = \frac{\sigma^2}{n} \frac{\sum (x_i - c)^2}{\sum (x_i - \bar{x})^2}.$$

- Now we find the first derivative of the variance with respect to c:

$$\frac{\partial \mathrm{Var}(\hat{\theta})}{\partial c} = -2 \frac{\sigma^2}{n} \frac{\sum (x_i - c)}{\sum (x_i - \bar{x})^2}.$$

- Let us finally find its minimum by laying it equal to 0:

$$\frac{\partial \mathrm{Var}(\hat{\theta})}{\partial c} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} (x_i - c) = 0 \quad \Leftrightarrow \quad c = \frac{\sum x_i}{n} = \bar{x}.$$

'Predicting' $\mathbb{E}(y|x_{n+1})$ vs. predicting y for a specific unit

- $\hat{\theta}$ (i.e., the predicted y) and more importantly its standard error and confidence intervals give the information about the **average** value of y for the subpopulation with a given set of realizations of explanatory variables.
- That is why there is no u in (1).
- However, to construct the confidence intervals for a prediction for a particular unit such as an individual or a firm, i.e., the **prediction interval**, we also need to take the uncertainty in u into consideration.

Predicting y for a specific unit

- y^0 is the value of the dependent variable for a specific unit given values $x_1^0, \ldots, x_k^0$, which can be standardly written as (cf. θ)

$$y^0 = \beta_0 + \beta_1 x_1^0 + \ldots \beta_k x_k^0 + u^0.$$

- **Prediction error** is defined as

$$\hat{e}^0 = y^0 - \hat{y}^0 = \beta_0 + \beta_1 x_1^0 + \ldots \beta_k x_k^0 + u^0 - \hat{y}^0.$$

- By the OLS assumptions, the prediction is unbiased, i.e., $\mathbb{E}(\hat{e}^0) = 0$.
- As only the u^0 and $\hat{y}^0$ are random variables, we have

$$\text{Var}(\hat{e}^0) = \text{Var}(u^0 - \hat{y}^0) = \text{Var}(u^0) + \text{Var}(\hat{y}^0) = \sigma^2 + \text{Var}(\hat{y}^0),$$

  which gives us

$$se(\hat{e}^0) = \sqrt{\hat{\sigma}^2 + \text{Var}(\hat{y}^0)}.$$

- We thus need to control both for the uncertainty in estimating y^0 and for the uncertainty of the unobserved error u^0.

A note on residual analysis

- Residuals are not only the 'estimates of the error term'.
- They can be further used for:
  - investment decisions (undervalued firm/stock/real estate),
  - assessment (below/above the OLS regression line),
  - discrimination (legal) cases.

Outline

Describing qualitative information

- Many qualitative factors can be represented as binary information: YES or NO (1 or 0)
- Such variables are usually labelled as **binary**, **zero-one**, or **dummy** variables.
- Classical examples: gender, employment, education, marital status, university degree, etc.

Outline

Single dummy independent variable

- Labelling the dummy variable as D, we can simply write a model as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k D + u.$$

- In practice, this gives us two models:

$$\begin{aligned} D_i = 0 \quad &: \quad y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} + u_i, \\ D_i = 1 \quad &: \quad y_i = (\beta_0 + \beta_k) + \beta_1 x_{i1} + \ldots + \beta_{k-1} x_{i,k-1} + u_i. \end{aligned}$$

- These models differ only in the intercepts (β_0 vs. $\beta_0 + \beta_k$), i.e., other effects are controlled for and assumed to be the same for both **groups**.
- This type is thus called an **intercept dummy**.

Dummy variable trap

- When constructing a model with dummy variables, we must not have a model which has both groups included in a model with the intercept.
- Let us have two dummy variables that are complementary, employed (E) and unemployed (U): we thus have $E + U = 1$.
- Remember how the X matrix in the matrix representation of the regression model looks:

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} & 1 & 0 \\ 1 & x_{21} & \dots & x_{2k} & 0 & 1 \\ . & x_{31} & \dots & x_{3k} & 1 & 0 \\ . & . & \dots & . & . & . \\ 1 & . & \dots & . & 1 & 0 \\ 1 & x_{n1} & \dots & x_{nk} & 0 & 1 \end{pmatrix}$$

- If we have the intercept in our model, we violate the MLR.3 assumption: this is called the **dummy variable trap**.

A base/benchmark group

- Dummy variable trap forces us to keep a specific group 'in the intercept'.
- Such group is referred to as the **base** or **benchmark group**.
- This gives us the interpretation of the coefficients of dummy variables as differences between the group included in the equation and the based group 'hidden' in the intercept.
- Alternatively, we can construct a model with all dummy variables but without an intercept. However, such specification is less useful for statistical testing and sometimes causes troubles with R^2 in econometric softwares.

Dummy variables and logarithms

- Nothing prevents us from constructing a model with a logarithmic dependent variable such as

$$\log(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k D + u.$$

- Interpretation here is the same as for a 'normal' independent variable, i.e., a semi-elasticity.
- Keeping other factors fixed, the dependent variable is approximately expected to be $100\beta_k\%$ higher for group $D = 1$ compared to group $D = 0$.

Outline

Using dummy variables for multiple categories

- Entire framework of a single dummy independent variable extends for multiple categories so that we can have a general model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-d} x_{k-d} + \\ + \beta_{k-d+1} D_1 + \ldots + \beta_k D_d + u,$$

  where d is the number of dummy variables.
- This gives us more information about the base group as well as more options for hypotheses testing.
- Beware the dummy variable trap here as well.
- Often, it is useful to create dummy variables as a combination of various criteria for a more straightforward interpretation.

Multiple categories: Example

- Let us have two characteristics of a person:
  - employment: E
  - citizenship: C
- Let us have a straightforward model $\quad wage = \beta_0 + \beta_1 E + \beta_2 C + u.$
- This, in practice, gives us a model of an expected wage for each group:

employed citizens	$wage = \beta_0 + \beta_1 + \beta_2 + u$
unemployed citizens	$wage = \beta_0 + \beta_2 + u$
employed foreigners	$wage = \beta_0 + \beta_1 + u$
unemployed foreigners	$wage = \beta_0 + u$

- Testing between groups can be complicated with similar specifications, so it can be easier to create one dummy for each specific group:

employed citizens	'EC'
unemployed citizens	'UC'
employed foreigners	'EF'
unemployed foreigners	'UF'

- Then, we can pick a group against which we want to test a statistical difference (using a usual t test) and specify the model accordingly:

$$wage = \gamma_0 + \gamma_1 UC + \gamma_2 EF + \gamma_3 UF + v.$$

Ordinal information

- Binary specification sometimes does not suffice.
- Assume a qualitative variable Q with possible values $0, 1, 2, 3$.
- We can either construct a model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \beta_k Q + u.$$

- Or, we can decompose Q into four new dummy categories:

$D_0 = 1$	if $Q = 0$,	$D_0 = 0$ otherwise
$D_1 = 1$	if $Q = 1$,	$D_1 = 0$ otherwise
$D_2 = 1$	if $Q = 2$,	$D_2 = 0$ otherwise
$D_3 = 1$	if $Q = 3$,	$D_3 = 0$ otherwise

  and construct a model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-3} x_{k-3} + \beta_{k-2} D_1 + \beta_{k-1} D_2 + \beta_k D_3 + u.$$

- What are the practical differences between these two approaches?

Seminars and the next lecture

- Seminars:
  - practicing interaction terms
  - model comparison
  - predictions & CIs
  - interpreting single dummy independent variable
- Next lecture #9:
  - interactions involving dummy variables (slope dummies)
  - binary dependent variable: LPM
  - more on interpreting discrete dependent variables
- Readings for lecture #9:
  - Chapter 7: 7.4–7