LECTURE #9

Econometrics I

QUALITATIVE ANALYSIS & LINEAR PROBABILITY MODEL

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

Summer semester 2024, April 23

In the previous lecture #8

- We summarized **four important variable selection criteria**:

  1. theory, 2. OVB reduction, 3. $\bar{R}^2$, 4. t/F test.

- For **predictions**, we derived the confidence/prediction intervals:
  - mean prediction $\mathbb{E}(y|x_{n+1})$ vs.
  - prediction for a specific unit: $se(\hat{e}^0) = \sqrt{\hat{\sigma}^2 + \mathrm{Var}(\hat{y}^0)}$.
- We introduced regression with **qualitative information**:
  - intercept dummy:

    $D_i = 1 \;:\; y_i = \boxed{(\beta_0 + \beta_k)} + \beta_1 x_{i,1} + \ldots + \beta_{k-1} x_{i,k-1} + u_i,$
  - dummy variable trap (base group),
  - multiple categories:

    $y = \beta_0 + \beta_1 x + \boxed{\beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3} + u,$
  - ordinal information: $Q = 0, 1, 2, 3.$
- Readings for lecture #9:
  - Chapter 7: 7.4–7

Home assignment #2

- Assigned on Thursday via SIS.
- Teams of two, one report.
- Matching spreadsheet.
- Delivered electronically in the .pdf format [5 MB max, R or other formats can be attached in .zip] via the **Study group roster** app (Lecture JEB109) in SIS.
- Deadline: Thursday, May 9, 2024, 23:59:59.
- 'Academic integrity'; solo $\Rightarrow$ 0.
- IES guide: AI tools when studying at IES.

Outline

More on binary independent variables
- Interactions involving dummy variables
- Testing for differences between groups: Chow test

A binary dependent variable: Linear probability model (LPM)

Outline

More on binary independent variables
- Interactions involving dummy variables
- Testing for differences between groups: Chow test

A binary dependent variable: Linear probability model (LPM)

Outline

More on binary independent variables

- Interactions involving dummy variables
- Testing for differences between groups: Chow test

A binary dependent variable: Linear probability model (LPM)

Interactions among dummy variables

- Similarly, as for quantitative explanatory variables, we can create interactions between/across dummy variables and between/across dummy and quantitative variables.
- The former is practically another way of forming multi-criterial dummy variables as these two approaches are effectively equivalent:
  - multi-criterial dummies are more suitable for testing differences between groups.
  - interacting dummies are more suitable for testing a general statement of significant interactions, e.g., whether gender wage differentials (discrimination) depend on ethnicity.
  - care must be taken with setting and understanding the base group correctly.
  - when including interaction among dummies, the separate ones are usually part of the model as well.

Interactions among dummy variables: Example

- Back to our example, we can have the following specification:

$$income = \beta_0 + \beta_1 E + \beta_2 C + \beta_3 E \cdot C + u.$$

- Figuring out the base group:
  - when dummy variables and the interaction products are zeros.
  - in this case, we need $E = 0$ and $C = 0$.

  $\Rightarrow$ our base group is an unemployed foreigner.
- To test against the other groups, we need to pick carefully:

employed citizens	$E = C = 1$	test $\theta = \beta_1 + \beta_2 + \beta_3 = 0$
unemployed citizens	$E = 0$ & $C = 1$	only test $\beta_2 = 0$
employed foreigners	$E = 1$ & $C = 0$	only test $\beta_1 = 0$

- Apparently, this can get a bit complicated.

Allowing for different slopes: Slope dummy

- We already know that a separate intercept dummy allows for different intercepts across multiple categories.
- Interaction term between a dummy and a quantitative explanatory variable functionally changes the model to allow for different slopes for specific groups.
- Consider the following model:

$$y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 Dx + u.$$

- Can this model be estimated? Is there a problem with the dummy variable trap?
- This, in practice, gives us two models for two subpopulations:

$$\begin{aligned} D_i = 0 \quad &: \quad y_i = \beta_0 + \beta_1 x_i + u_i, \\ D_i = 1 \quad &: \quad y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + u_i. \end{aligned}$$

- This specification allows for both the individual and joint significance testing (be aware of the high sample correlation between D and Dx).

Illustration: Different intercepts vs. different slopes

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u \\ &= (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) educ + u\end{aligned}$$

FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.

© Cengage Learning, 2013

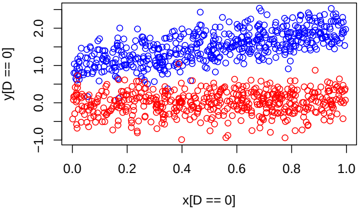
Source: Wooldridge (2012)

Illustration in R: Two subpopulations

$$
\begin{aligned}
y &= (\beta_0 + \delta_0 D) + (\beta_1 + \delta_1 D)x + u \\
\text{setup} &: \beta_0 = 1, \delta_0 = -1, \beta_1 = 1, \delta_1 = -1
\end{aligned}
$$

$$
\begin{aligned}
D_i = 0 &: y_i = \beta_0 + \beta_1 x_i + u_i \\
D_i = 1 &: y_i = u_i
\end{aligned}
$$

(a) $\sigma^2 = 0.1$



(b) $\sigma^2 = 0.5$

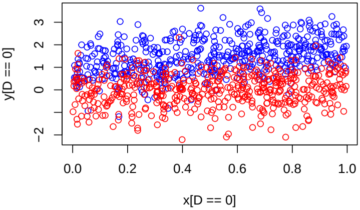
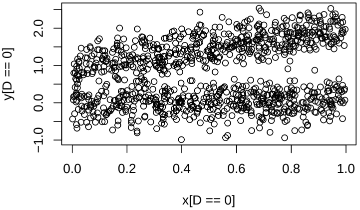


Illustration in R: Two subpopulations

$$\begin{aligned} y &= (\beta_0 + \delta_0 D) + (\beta_1 + \delta_1 D)x + u \\ \text{setup} &: \beta_0 = 1, \delta_0 = -1, \beta_1 = 1, \delta_1 = -1 \end{aligned}$$

$$\begin{aligned} D_i = 0 &: y_i = \beta_0 + \beta_1 x_i + u_i \\ D_i = 1 &: y_i = u_i \end{aligned}$$

(a) $\sigma^2 = 0.1$



(b) $\sigma^2 = 0.5$

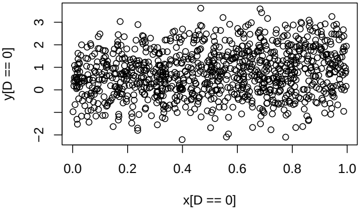


Illustration in R: Two subpopulations: (a) $\sigma^2 = 0.1$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.97345    0.02921   33.33   <2e-16 ***
D           -1.03492    0.04163  -24.86   <2e-16 ***
x            1.01965    0.04978   20.48   <2e-16 ***
I(D * x)    -0.90469    0.07086  -12.77   <2e-16 ***
---
Multiple R-squared:  0.857,   Adjusted R-squared:  0.8566

vs.

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47103    0.05393   8.734  < 2e-16 ***
x            0.57626    0.09181   6.276 5.16e-10 ***
---
Multiple R-squared:  0.03797,   Adjusted R-squared:  0.03701
```

Illustration in R: Two subpopulations: (a) $\sigma^2 = 0.5$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.94063    0.06532  14.401  < 2e-16 ***
D           -1.07809    0.09308 -11.583  < 2e-16 ***
x            1.04394    0.11131   9.379  < 2e-16 ***
I(D * x)    -0.78687    0.15845  -4.966 8.04e-07 ***
---
Multiple R-squared:  0.5416,   Adjusted R-squared:  0.5403

vs.

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.41680    0.06756   6.169 9.96e-10 ***
x            0.65866    0.11502   5.726 1.36e-08 ***
---
Multiple R-squared:  0.03181,   Adjusted R-squared:  0.03084
```

Outline

More on binary independent variables

- Interactions involving dummy variables
- Testing for differences between groups: Chow test

A binary dependent variable: Linear probability model (LPM)

Testing for differences between groups

- Testing among various dummy variable parameters allows us to uncover differences between groups.
- **Chow test** is frequently used to test the stability/equality of the parameters of the underlying population model for different groups.
- Assume the population model with $k = 1$: $\boxed{y = \beta_0 + \beta_1 x + u}$.
- Let us assume two subpopulations and define a dummy

$$D_i = \begin{cases} 0 & \text{if the } i\text{-th observation is from the first subpopulation,} \\ 1 & \text{if the } i\text{-th observation is from the second subpopulation.} \end{cases}$$

- Extend the population model

$$\begin{aligned} y_i &= \beta_0 + \delta_0 D_i + \beta_1 x_i + \delta_1 D_i x_i + u_i \\ &= (\beta_0 + \delta_0 D_i) + (\beta_1 + \delta_1 D_i) x_i + u_i. \end{aligned} \qquad (1)$$

- Using an F test, we can specify

$$H_0 : \delta_0 = 0 \text{ and } \delta_1 = 0 \qquad \text{vs.} \qquad H_1 : \delta_0 \neq 0 \text{ or } \delta_1 \neq 0.$$

Chow test

- For a standard F test, $D_i = 1$ defines the unrestricted model (1) while $D_i = 0$ defines the restricted model:

$$y_i = \beta_0 + \beta_1 x_i + u_i. \quad (2)$$

- If we estimate the restricted/**population** model (2) **separately** over the two subpopulations/**subsamples** and obtain the respective SSR_1 and SSR_2, it can be shown that $SSR_1 + SSR_2 = SSR_U$ from (1).
- SSR_R based on (1) is then simply the residual sum of squares from (2) estimated using the **pooled** dataset, which we label SSR_P.
- F statistic, or the **Chow statistic**, is then defined as:

$$F = \boxed{\frac{SSR_P - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \frac{n - 2(k+1)}{k+1}},$$

$$\text{compare to} \quad F = \frac{SSR_R - SSR_U}{SSR_U} \frac{n-k-1}{q}.$$

- Under the null hypothesis, i.e., that the population model does not differ for the subpopulations, $F \sim F_{k+1, n-2(k+1)}$.

Chow test* (alternative, less strict)

- As for a standard F test, we assume homoskedasticity and normality of the error term for an exact F distribution under the null, but the normality assumption can be dropped for large samples, and the statistic is then approximately F-distributed.
- Sometimes, the classical Chow test can be too strict as it does not allow any parameter to differ across subpopulations, including the intercept.
- Alternative specification of the test and its testing statistic can be written as:

$$F = \frac{SSR_P^* - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \frac{n - 2(k+1)}{k},$$

  where SSR_P^* is the residual sum of squares of the population model, which, however, includes a dummy variable for the **intercept shift**, i.e., we consider **one less restriction**.
- This also changes the limiting distribution to $F \sim F_{k,n-2(k+1)}$.

Outline

More on binary independent variables

- Interactions involving dummy variables
- Testing for differences between groups: Chow test

A binary dependent variable: Linear probability model (LPM)

A binary dependent variable

- Until now, we have only used qualitative variables as the independent ones in a model.
- However, also various dependent variables of interest have a **qualitative** or discrete nature (this semester, we only stick to the binary ones).
- Population model remains the same as for a quantitative dependent variable:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$$

but with y being binary, i.e., only 1 or 0.

Expected value of y

- As y is either 1 or 0, the interpretation of βs changes.
- Under MLR.1–4, the OLS estimator is still unbiased and consistent, and we thus still have

$$\mathbb{E}(y|X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- Importantly, y is a discrete random variable with a Bernoulli distribution

$$y = \begin{cases} 1 & \text{with probability } P, \\ 0 & \text{with probability } 1 - P. \end{cases}$$

- We can thus find the expected value

$$\boxed{\mathbb{E}(y|X)} = 1 \cdot \mathrm{P}(y = 1|X) + 0 \cdot \big(1 - \mathrm{P}(y = 1|X)\big) = \boxed{\mathrm{P}(y = 1|X)}.$$

- I.e., the expected value of the dependent variable y (given X) is the probability of y being 1 (given X).

Interpretation of βs in the LPM framework

- $\mathrm{P}(y = 1|X)$ is usually referred to as the **response probability**.
- Both $\mathrm{P}(y = 1|X)$ and $\mathrm{P}(y = 0|X) = 1 - \mathrm{P}(y = 1|X)$ are linear in β_j and also by definition linear functions of all x_j, so the model is called the **linear probability model (LPM)**

$$p(X) \equiv \mathrm{P}(y = 1|X) = \mathbb{E}(y|X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- Previous derivations allow for the interpretation of βs as:

$$\Delta p(X) = \Delta \mathrm{P}(y = 1|X) = \beta_j \Delta x_j,$$

  i.e., the change in the **probability of 'success'** (probability of y being 1) when x_j changes by one small unit.
- $\hat{y}$ is the predicted probability of 'success'.
- Dummies can be added as independent variables as well.

Shortcomings of the LPM

1. Estimated/predicted probability is **not bounded by 0 and 1**.
   - this can be partially solved either by (a) truncation to $\langle 0, 1\rangle$ or by (b) thresholding, i.e., setting a value of $\hat{y}$ separating 0 and 1 probabilities.
   - both potentially problematic: (a) too many 0 and 1, i.e., exact predicted probabilities 0% and 100%; (b) how to set the threshold?
2. **Constant marginal effect** Δx_j (often unrealistic).
3. Error term is inherently **heteroskedastic**:
   - MLR.5 violated, but OLS remains unbiased and consistent.
   - heteroskedasticity needs to be dealt with (lecture #10) as it is crucial for justifying the usual t and F statistics even in large samples.

$$\begin{aligned}\boxed{\mathrm{Var}(u|X)} &= \mathrm{Var}(y|X) = \mathbb{E}(y^2|X) - \big(\mathbb{E}(y|X)\big)^2 = \\ &= 1^2 \cdot p(X) + 0^2 \cdot \big(1 - p(X)\big) - \big(p(X)\big)^2 = \boxed{p(X)\big(1 - p(X)\big)}.\end{aligned}$$

4. Error term is **not normally distributed**.
   - because y takes on only two values, u also takes on only two possible values for given X.

Potential advantages of the LPM

Nonetheless, the LPM is still useful and often applied in economics, as it is:

1. Simple to estimate.
2. Intuitive and straightforward linear interpretation.
3. It usually works well for values of x_j near the sample averages.
4. However, be aware: R^2 and $\bar{R}^2$ no longer good measures of the goodness-of-fit for the LPM!

Seminars and the next lecture

- Seminars:
  - intercept and slope dummies
  - multiple categories and interactions with dummies
  - model construction
  - practicing Chow test in R
  - LPM in practice in R
- Next lecture #10:
  - heteroskedasticity: consequences for OLS
  - heteroskedasticity-robust inference
  - testing for heteroskedasticity
  - weighted least squares (WLS) estimator
  - LPM revisited
  - (F)GLS estimator
- Readings for lecture #10:
  - Chapter 8 **(8.4 'What If...' & 'Prediction and...' and 8.6 mandatory after lecture)**