Regression Analysis: Dummy Variables

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

Statistics and Econometrics

9/11/2020 : 1 - [illegible]

Jiahua Wu

382 Business School
j.wu@imperial.ac.uk

Imperial College Business School

Imperial means Intelligent Business

Roadmap

- Regression analysis with cross-sectional data
  - Basics: estimation, inference, analysis with dummy variables
  - More involved: model specification and data issues
- Advanced topics
  - Binary dependent variable models
  - Panel data analysis
  - Time series analysis

Outline (Wooldridge, Chap. 7.1 - 7.6)

- Qualitative information and dummy variables
- Interactions involving dummy variables
- The linear probability model

Outline

- Qualitative information and dummy variables
- Interactions involving dummy variables
- The linear probability model

Qualitative Information

- Many factors in empirical projects are qualitative (non-numerical) that take two values
  - Eg. gender, marriage, etc
- They can be modelled as binary valued variables $(0-1)$, known as dummy variables
  - Eg. female ($=1$ if are female, 0 otherwise), married ($=1$ if are married, 0 otherwise)
- The assignment of values $(0,1)$ is often determined by interpretation convenience

Dummy Independent Variables

- Eg. Wage model

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u,$$

where δ_0 characterise the gender difference in wage

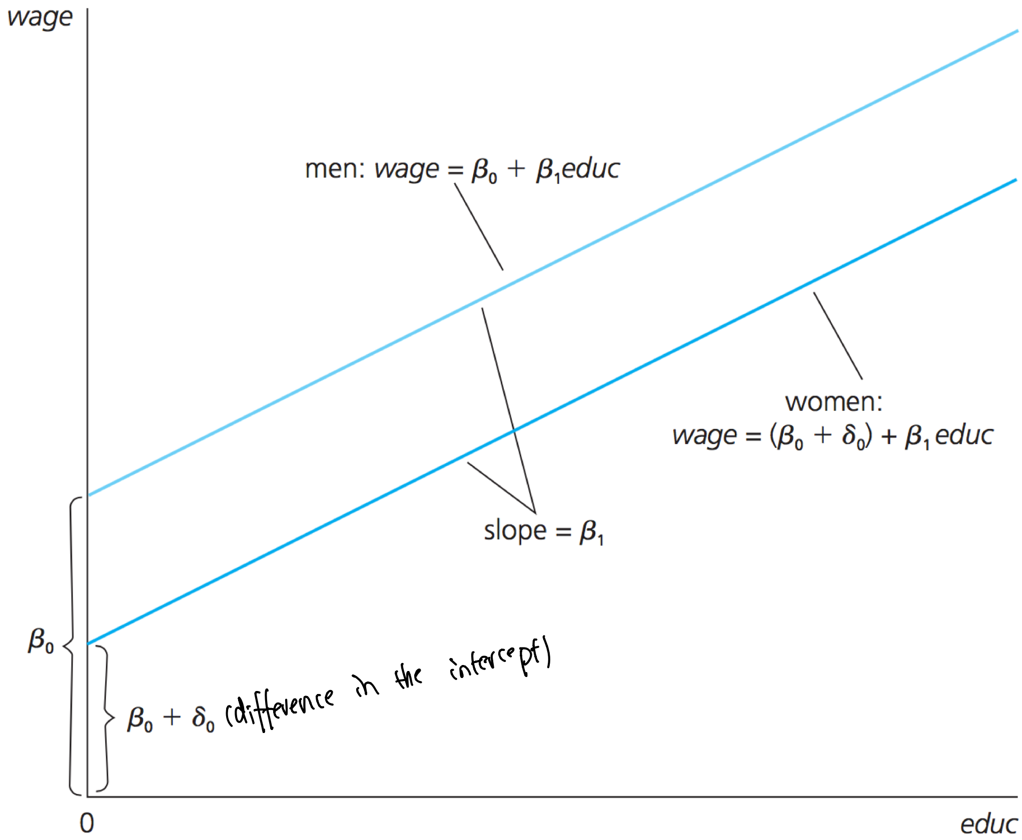
- The conditional expectation of *wage* is given by

$$\begin{aligned} E(wage|female = 1, educ) &= \beta_0 + \delta_0 + \beta_1 educ, \\ E(wage|female = 0, educ) &= \beta_0 + \beta_1 educ, \end{aligned}$$

where δ_0 represents an intercept shift

Dummy Independent Variables

$$wage = \beta_0 + \delta_0 female + \beta_1 educ \text{ for } \delta_0 < 0$$



Interpretation of Dummy

- Eg. Wage model (continued)

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

  - Would you add the *male* dummy in the model?

Interpretation of Dummy

- Eg. Wage model (continued)

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

  - Would you add the *male* dummy in the model?
  - Males are treated as the base group (against which comparisons are made)
  - We could regress *wage* on *male* and *educ*, where females would be base group (coefficient interpretation would be different)

Dummy Independent Variables

- Example 7.1 (wage1.RData) Use *educ*, *exper*, *tenure* and gender to explain hourly wage

- ```
  > load("wage1.RData")
  > wage.m1 <- lm(wage ~ female + educ + exper + tenure, data
      = data)
  ```
  ```

Dummy Independent Variables

expected wage for male with 0 of eductn. . . -
not meaningful

- Example 7.1. OLS SRF

$$\begin{aligned}\widehat{wage} &= \underset{(0.72)}{-1.57} - \underset{(0.26)}{1.81}\, female + \underset{(0.049)}{0.572}\, educ \\ &\quad + \underset{(0.012)}{0.025}\, exper + \underset{(0.021)}{0.141}\, tenure\end{aligned}$$

$n = 526, R^2 = 0.364$

  - Negative intercept is not meaningful here
  - Interpretation of the coefficient of *female*
    - A female worker is predicted to earn \$1.81 less than a male worker at the same level of *educ*, *exper* and *tenure*

- Compare the above with the simple regression

$$\widehat{wage} = \underset{(0.21)}{7.10} - \underset{(0.30)}{2.51}\, female, \quad n = 526, R^2 = 0.116$$

Natural Interpretation: average wage for men.

Dummy Independent Variables in a Log Model

- Eg. Wage model (continued): what if $y = \log(wage)$?

$$\widehat{\log(wage)} = \underset{(.102)}{.501} - \underset{(.037)}{.301}\,female + \underset{(.007)}{.087}\,educ + \underset{(.002)}{.005}\,exper + \underset{(.003)}{.017}\,tenure$$

$n = 526, R^2 = 0.392$

- What is the interpretation of the coefficient of *female*?

percentage interpretion.

Dummy Independent Variables in a Log Model

- Eg. Wage model (continued): what if $y = \log(wage)$?

$$\begin{aligned}\widehat{\log(wage)} &= \underset{(.102)}{.501} - \underset{(.037)}{.301}\,female + \underset{(.007)}{.087}\,educ \\ &\quad + \underset{(.002)}{.005}\,exper + \underset{(.003)}{.017}\,tenure\end{aligned}$$

$n = 526, R^2 = 0.392$

- What is the interpretation of the coefficient of *female*?
  - A female worker is predicted to earn 30.1% less than a male worker at the same level of *educ*, *exper* and *tenure*

Dummy Variables for Multiple Categories

- What if individuals are from more than two categories?
  - Eg. gender-marriage: single male, single female, married male, and married female base group
  - Eg. Wage model (again)

$$
\begin{aligned}
wage &= \beta_0 + \delta_1 SingleFemale + \delta_2 MarriedFemale \\
&\quad + \delta_3 MarriedMale + \beta_1 educ + \cdots + u
\end{aligned}
$$

- In general, for g groups, we need $g - 1$ dummy variables, with the intercept for the base group
- The coefficient on the dummy of a group is the difference in the intercepts between that group and the base group

Dummy Variables for Ordinal Information

- Consider a variable that takes multiple values, where the order matters but the scale is not meaningful
  - Eg. A government's credit rating is on the scale of $0-4$ with $0=$ very risky, $1=$ risky, $2=$ neutral, $3=$ safe, $4=$ very safe.
  - Can we just include an independent variable, say CR, and use the regression model

$$y=\beta_0+\beta_1 CR+\textit{other factors}?$$

Dummy Variables for Ordinal Information

- Consider a variable that takes multiple values, where the order matters but the scale is not meaningful
  - Eg. A government's credit rating is on the scale of $0-4$ with $0 =$ very risky, $1 =$ risky, $2 =$ neutral, $3 =$ safe, $4 =$ very safe.
  - Can we just include an independent variable, say *CR*, and use the regression model

$$y = \beta_0 + \beta_1 CR + \textit{other factors}?$$

  - Better to use separate dummies for multiple values: $CR_1 = 1$ if risky, $CR_2 = 1$ if neutral, $CR_3 = 1$ if safe, $CR_4 = 1$ if very safe
- If an ordinal variable takes too many values, then group them into a small number of categories
  - Eg. Business school rankings: not sensible to use a dummy for each value. Rather use 4 dummies to indicate if the rank is in top 10, 11-25, 26-40, 41-60, and the rest

Dummy Variables for Ordinal Information: An Example

- Example 7.7 (beauty.RData) Effects of attractiveness on wage.
  - The attractiveness of each person in the sample was ranked as "below average", "average", or "above average"
  - Use *educ*, *exper* and physical attractiveness of an individual to explain *wage*
  -
```
> data.male <- data %>% filter(female == 0)
> male.m1 <- lm(log(wage) ~ belavg + abvavg + educ +
     exper, data = data.male)
> data.female <- data %>% filter(female == 1)
> female.m1 <- lm(log(wage) ~ belavg + abvavg + educ +
     exper, data = data.female)
> stargazer(male.m1, female.m1, no.space = TRUE, align
     = TRUE)
```

Dummy Variables for Ordinal Information: An Example

	Dependent variable:	
	log(wage)	
	(1)	(2)
belavg	-0.173^{***}	-0.108
	(0.055)	(0.069)
abvavg	-0.038	0.038
	(0.039)	(0.051)
educ	0.066^{***}	0.083^{***}
	(0.007)	(0.009)
exper	0.015^{***}	0.011^{***}
	(0.001)	(0.002)
Constant	0.751^{***}	0.105
	(0.096)	(0.124)
Observations	824	436
R^2	0.181	0.199
Adjusted R^2	0.177	0.192
Residual Std. Error	0.490 (df = 819)	0.471 (df = 431)
F Statistic	45.175^{***} (df = 4; 819)	26.822^{***} (df = 4; 431)
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Outline

- Qualitative information and dummy variables
- Interactions involving dummy variables
- The linear probability model

Interactions among Dummy Variables

- Interacting dummy variables is like subdividing the group
- Eg. Wage model (controlling for gender and marital status)

$$
\begin{aligned}
wage &= \beta_0 + \delta_1 SingleFemale + \delta_2 MarriedFemale \\
&\quad + \delta_3 MarriedMale + \beta_1 educ + \cdots + u
\end{aligned}
$$

- Alternatively, we can use two dummy variables: *female* and *married*, and their interactions

$$
\begin{aligned}
wage &= \beta_0 + \delta_1 female + \delta_2 married \cdot female \\
&\quad + \delta_3 married + \beta_1 educ + \cdots + u
\end{aligned}
$$

Other Interactions with Dummies

- Interacting a dummy with a quantitative variable allows for different slope parameters
- Eg. Wage model

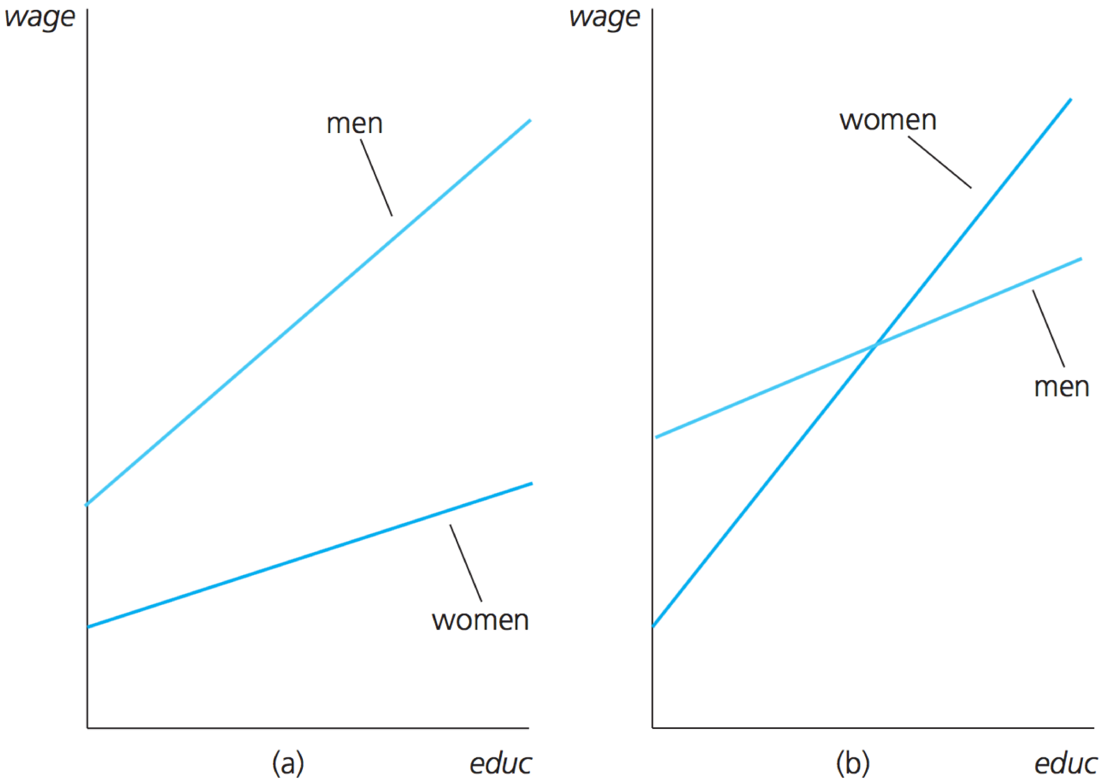
$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female) educ + u$$

  - $female = 0$: intercept and slope are β_0 and β_1
  - $female = 1$: intercept and slope are $(\beta_0 + \delta_0)$ and $(\beta_1 + \delta_1)$
  - Differences in intercept and slope are measured by δ_0 and δ_1, respectively

Other Interactions with Dummies

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u$$

$\delta_0 < 0, \delta_1 < 0$ $\delta_0 < 0, \delta_1 > 0$



Other Interactions with Dummies

- To estimate, we use OLS for

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + u,$$

  where δ_1 is the effect of the interaction of *female* and *educ*
- A number of hypotheses of interest can be tested in this model
  - The return to education is the same for men and women
  - Expected wages are the same for men and women who have the same level of education
- Testing hypotheses in `R`

Outline

- Qualitative information and dummy variables
- Interactions involving dummy variables
- The linear probability model

Linear Probability Model

- Consider the case where the dependent variable (response) is binary: $y = 0$ or 1
  - Eg. y represents whether or not: a person was employed last week; a household purchased a car last year.
- When the response (y) is influenced by a number of independent variables (x's), we may write

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- But how do we interpret the β coefficients?

Linear Probability Model

- Notice that for binary response

  $$P(y = 1|\mathbf{x}) = E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

  The PRF is the probability of "success" for given x's

- $P(y = 1|\mathbf{x})$ is known as the response probability, and the regression model with a binary dependent variable is called the linear probability model (LPM)

- The parameter β_j is interpreted as the change in the probability of success caused by a one-unit increase in x_j: $\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j$

- The interpretation of the predicted value is the predicted probability of success

Linear Probability Model: An Example

- Example (mroz.RData). Predict labour force participation by married women (7.29)
  - The dependent variable is *inlf*, whether the woman was in the labour force last year
  - The mechanics of OLS are the same as before

```
> load("mroz.RData")
> fitted.inlf <- lm(inlf ~ educ + kidslt6, data = data)
```

  - OLS SRF

$$\widehat{inlf} = \underset{(.095)}{.053} + \underset{(.008)}{.046}\, educ - \underset{(.033)}{.224}\, kidslt6,$$

  where

    - *educ*: years of education
    - *kidslt6*: number of children less than six years old

  - Holding everything else fixed, another year of education increases the probability of labour force participation by 0.046

Issues with Linear Probability Model

- The predicted probability can be outside $[0, 1]$
  - Linear function is not suitable for modelling probabilities
- For LPM, it can be shown that

$$Var(u|\mathbf{x}) = Var(y|\mathbf{x}) = P(y = 1|\mathbf{x})[1 - P(y = 1|\mathbf{x})]$$

That is, the conditional variance depends on x's (heteroskedasticity). It does not cause estimation bias but does invalidate the standard errors