Chapter 7: Dummy variable regression

Why include a qualitative independent variable? . . . 2

Simplest model **3**
- Simplest case . . . 4
- Example (continued) . . . 5
- Possible solution: separate regressions . . . 6
- Independent variable vs. regressor . . . 7
- Common slope model . . . 8
- Testing . . . 9

More general models **10**
- More than one quantitative independent variable . . . 11
- Polytomous independent variables . . . 12
- Example (continued) . . . 13
- Testing with polytomous independent variable . . . 14
- R commands . . . 15
- More than one qualitative independent variable . . . 16

Interaction **17**
- Definition . . . 18
- Interaction vs. correlation . . . 19
- Constructing regressors . . . 20
- Testing . . . 21
- Principle of marginality . . . 22
- Polytomous independent variables . . . 23
- Hypothesis tests . . . 24
- Standardized estimates . . . 25
- Interaction between categorical variables . . . 26

Why include a qualitative independent variable?

- We are interested in the effect of a qualitative independent variable (for example: do men earn more than women?)
- We want to better predict/describe the dependent variable. We can make the errors smaller by including variables like gender, race, etc.
- Qualitative variables may be confounding factors. Omitting them may cause biased estimates of other coefficients.

Simplest model

Simplest case

- Example:
  - Dependent variable: income
  - One quantitative independent variable: education
  - One dichotomous (can take two values) independent variable: gender
- Assume effect of either independent variable is the same, regardless of the value of the other variable (additivity, parallel regression lines) - See pictures from book.
- Usual assumptions on statistical errors: independent, zero means, constant variance, normally distributed, fixed X's or X independent of statistical errors.

Example (continued)

- Suppose that we are interested in the effect of education on income, and that gender has an effect on income.
- See pictures from book.
- Scenario 1: Gender and education are uncorrelated
  - Gender is not a confounding factor
  - Omitting gender gives correct slope estimate, but larger errors
- Scenario 2: Gender and education are correlated
  - Gender is a confounding factor
  - Omitting gender gives biased slope estimate, and larger errors

Possible solution: separate regressions

- Fit separate regression for men and women
- Disadvantages:
  - How to test for the effect of gender?
  - If it is reasonable to assume that regressions for men and women are parallel, then it is more efficient to use all data to estimate the common slope.

Independent variable vs. regressor

- Y=income, X=education, D=regressor for gender:

$$D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$$

- Independent variable = real variables of interest
- Regressor = variable put in the regression model
- In general, regressors are functions of the independent variables. Sometimes regressors are equal to the independent variables.

Common slope model

- $Y_i = \alpha + \beta X_i + \gamma D_i + \epsilon_i$
- For women ($D_i = 0$):
$$Y_i = \alpha + \beta X_i + \gamma \cdot 0 + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$
- For men ($D_i = 1$):
$$Y_i = \alpha + \beta X_i + \gamma \cdot 1 + \epsilon_i = (\alpha + \gamma) + \beta X_i + \epsilon_i$$
- See picture from book.
- What are the interpretations of α, β and γ?
- What happens if we code $D = 1$ for women and $D = 0$ for men?

Testing

- Test the partial effect of gender:
  - $H_0 : \gamma = 0$, $H_a : \gamma \neq 0$
  - Same as before:
    Compute t-statistic or incremental F-test
- Test the partial effect of education:
  - $H_0 : \beta = 0$, $H_a : \beta \neq 0$
  - Same as before:
    Compute t-statistic or incremental F-test
- Cystic fibrosis example.

9 / 26

More general models

10 / 26

More than one quantitative independent variable

- All methods go through, as long as we assume parallel regression surfaces.
- Model: $Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma D_i + \epsilon_i$.
- Women ($D_i = 0$):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma \cdot 0 + \epsilon_i \\ &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \end{aligned}$$

- Men ($D_i = 1$):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \gamma \cdot 1 + \epsilon_i \\ &= (\alpha + \gamma) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \end{aligned}$$

- Interpretation of α, $\beta_1, \ldots, \beta_k$, γ.

11 / 26

Polytomous independent variables

- Qualitative variable with more than two categories
- Example: Duncan data:
  - Dependent variable: Y=prestige
  - Quantitative independent variables:
    X_1=income and X_2=education
  - Qualitative independent variable: type (bc, prof, wc)
- D_1 and D_2 are regressors for type:

Type	D_1	D_2
Blue collar (bc)	0	0
Professsional (prof)	1	0
White collar (wc)	0	1

- If there are p categories, use $p-1$ dummy regressors.
  What happens if we use p regressors?

Example (continued)

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \epsilon$
- Blue collar ($D_{i1} = 0$ and $D_{i2} = 0$):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 + \epsilon_i \\ &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

- Professional ($D_{i1} = 1$ and $D_{i2} = 0$):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 + \epsilon_i \\ &= (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

- White collar ($D_{i1} = 0$ and $D_{i2} = 1$):

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 + \epsilon_i \\ &= (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \end{aligned}$$

Testing with polytomous independent variable

- Test partial effect of type, i.e., the effect of type controlling for income and education.
- $H_0 : \gamma_1 = \gamma_2 = 0$
- H_a: at least one $\gamma_j \neq 0$, $j = 1, 2$.
- Incremental F-test:
  - Null model:
    $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - Full model:
    $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \epsilon$
- What do the individual p-values in `summary(lm())` mean?
- First look at F-test, then at individual p-values

R commands

- Creating dummy variables by hand:
```
D1 <- (type=="prof")*1
D2 <- (type=="wc")*1
m1 <- lm(prestige~education+income+D1+D2)
```
- Letting R do things automatically:
```
m1 <- lm(prestige~education+income+type)
m1 <- lm(prestige~education+income+factor(type))
```
- The use of `factor()`:
  - `factor()` is not needed in this example, because the coding of the categories is in words: "bc", "prof", "wc".
  - It is essential to use `factor()` if the coding of the categories is numerical!
  - To be safe, you can always use `factor`.
- Example R-code

More than one qualitative independent variable

- Example: Y=prestige, X_1=income, X_2=education,

Type	D_1	D_2
Blue collar	0	0
Professional	1	0
White collar	0	1

and

Gender	D_3
Women	0
Men	1

- $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + \epsilon$
- What is the equation for men with professional jobs? And for women with white collar jobs?

16 / 26

Interaction

17 / 26

Definition

- Two variables are said to *interact* in determining a dependent variable if the partial effect of one depends on the value of the other.
- So far we only studied models without interaction.
- Interaction between a quantitative and a qualitative variable means that the regression surfaces are not parallel. See picture.
- Interaction between two qualitative variables means that the effect of one of the variables depends on the value of the other variable. Example: the effect of type of job on prestige is bigger for men than for women.
- Interaction between two quantitative variables is a bit harder to interpret, and we may consider that later.

18 / 26

Interaction vs. correlation

- First, note that in general, the *independent* variables are *not independent* of each other.
- Correlation:
  Independent variables are statistically related to each other.
- Interaction:
  Effect of one independent variable on the dependent variable depends on the value of the other independent variable.
- Two independent variables can interact whether or not they are correlated.

19 / 26

Constructing regressors

- Y=income, X=education, D=dummy for gender
- $Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \epsilon_i$
- Note $X \cdot D$ is a new regressor. It is a function of X and D, but not a linear function. Therefore we do not get perfect collinearity.
- Women ($D_i = 0$):
$$Y_i = \alpha + \beta X_i + \gamma \cdot 0 + \delta(X_i \cdot 0) + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$
- Men ($D_i = 1$)
$$\begin{aligned} Y_i &= \alpha + \beta X_i + \gamma \cdot 1 + \delta(X_i \cdot 1) + \epsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta) X_i + \epsilon_i \end{aligned}$$
- Interpretation of α, β, γ, δ.

20 / 26

Testing

- Testing for interaction is testing for a difference in slope between men and women. $H_0 : \delta = 0$ and $H_a : \delta \neq 0$.
- What is the difference between:
  - The model with interaction
  - Fitting two separate regression lines for men and women

21 / 26

Principle of marginality

- If interaction is significant, do not test or interpret main effects:
  - First test for interaction effect.
  - If no interaction, test and interpret main effects.
- If interaction is included in the model, main effects should also be included.
- See pictures of models that violate the principle of marginality.

22 / 26

Polytomous independent variables

- Create interaction regressors by taking the products of all dummy variable regressors and the quantitative variable.
- Example:
  - Y=prestige, X_1=education, X_2=income
  - D_1, D_2=dummies for type of job

$$
\begin{aligned}
Y = {} & \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 D_1 + \gamma_2 D_2 \\
& + \delta_{11} X_1 D_1 + \delta_{12} X_1 D_2 + \delta_{21} X_2 D_1 + \delta_{22} X_2 D_2 + \epsilon
\end{aligned}
$$

- Interpretation of parameters

23 / 26

Hypothesis tests

- When testing for main effects and interactions, follow principle of marginality
- Use incremental F-test
- Examples in R-code

24 / 26

Standardized estimates

- Do not standardize dummy-regressor coefficients.
- Dummy regressor coefficient has clear interpretation.
- By standardizing it, this interpretation gets lost. Therefore we don't standardize dummy regressor coefficients.
- Also, don't standardize interaction regressors. You can standardize the quantitative independent variable before taking its product with the dummy regressor.

25 / 26

Interaction between categorical variables

- Example: Does reproduction reduce lifespan of male fruitflies?
- Experiment:
  - male flies with 1 pregnant (not receptive) female per day
  - male flies with 8 pregnant females per day
  - male flies with 1 virgin (receptive) female per day
  - male flies with 8 virgin females per day
  - male flies without females
- Each group contains 25 fruitflies
- Available information:
  - Thorax length in mm
  - Percentage of time sleeping
  - Longevity in days
- See plots

26 / 26