Multiple Linear Regression (Dummy Variable Treatment)

CIVL 7012/8012
In Today’s Class

• Recap
• Single dummy variable
• Multiple dummy variables
• Ordinal dummy variables
• Dummy-dummy interaction
• Dummy-continuous/discrete interaction
• Binary dependent variables
Introducing Dummy Independent Variable

• **Qualitative Information**
  - Examples: gender, race, industry, region, rating grade, ...
  - A way to incorporate qualitative information is to use dummy variables
  - They may appear as the dependent or as independent variables

• **A single dummy independent variable**
  \[ wage = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + u \]

= the wage gain/loss if the person is a woman rather than a man (holding other things fixed)

Dummy variable:
- =1 if the person is a woman
- =0 if the person is man
Illustrative Example

• Graphical Illustration

Alternative interpretation of coefficient:

\[
\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)
\]

i.e. the difference in mean wage between men and women with the same level of education.

Intercept shift
Specification of Dummy Variables

- Dummy variable trap

This model cannot be estimated (perfect collinearity)

\[ \text{wage} = \beta_0 + \gamma_0 \text{male} + \delta_0 \text{female} + \beta_1 \text{educ} + u \]

When using dummy variables, one category always has to be omitted:

\[ \text{wage} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + u \]  

\[ \text{wage} = \beta_0 + \gamma_0 \text{male} + \beta_1 \text{educ} + u \]

Alternatively, one could omit the intercept:

\[ \text{wage} = \gamma_0 \text{male} + \delta_0 \text{female} + \beta_1 \text{educ} + u \]

Disadvantages:
1) More difficult to test for differences between the parameters
2) R-squared formula only valid if regression contains intercept

The base category are men

The base category are women
Interpretation of Dummy Variables

- Estimated wage equation with intercept shift

\[
\hat{\text{wage}} = -1.57 - 1.81 \text{female} + 0.572 \text{educ} \\
+ 0.025 \text{exper} + 0.141 \text{tenure}
\]

Holding education, experience, and tenure fixed, women earn 1.81\$ less per hour than men.

\(n = 526, R^2 = 0.364\)

- Does that mean that women are discriminated against?
  - Not necessarily. Being female may be correlated with other productivity characteristics that have not been controlled for.
Model with only dummy variables-
(Example-1)

- Comparing means of subpopulations described by dummies

\[
\hat{\text{wage}} = 7.10 - 2.51 \cdot \text{female} \\
(0.21) \quad (0.26)
\]

\[n = 526, R^2 = 0.116\]

- Discussion
  - It can easily be tested whether difference in means is significant
  - The wage difference between men and women is larger if no other things are controlled for; i.e. part of the difference is due to differences in education, experience and tenure between men and women.

Not holding other factors constant, women earn 2.51$ per hour less than men, i.e. the difference between the mean wage of men and that of women is 2.51$. 
Further example: Effects of training grants on hours of training

- Treatment group (= grant receivers) vs. control group (= no grant)
- Is the effect of treatment on the outcome of interest causal?

Model with only dummy variables-(Example-2)

\[ \text{hrsemp} = 46.67 + 26.25 \times \text{grant} - 0.98 \times \log(\text{sales}) \]
\[ (43.41) \quad (5.59) \quad (3.54) \]
\[ - 6.07 \times \log(\text{employ}), \quad n = 105, \quad R^2 = 0.237 \]
\[ (3.88) \]
Dependent log(y) and Dummy
Independent

- Using dummy explanatory variables in equations for log(y)

\[
\hat{\log}(price) = -1.35 + 0.168 \log(lotsize) + 0.707 \log(sqft) + 0.027 bdrms + 0.054 \text{colonial}
\]

(0.65) (0.038) (0.093) (0.029) (0.045)

\[n = 88, R^2 = 0.649\]

As the dummy for colonial style changes from 0 to 1, the house price increases by 5.4 percentage points.
Dummy variables for multiple categories

- Using dummy variables for multiple categories
  - 1) Define membership in each category by a dummy variable
  - 2) Leave out one category (which becomes the base category)

\[
\hat{\log}(wage) = 0.321 + 0.213 \text{ marrmale} - 0.198 \text{ marrfem} \\
- 0.110 \text{ singfem} + 0.079 \text{ educ} + 0.027 \text{ exper} - 0.00054 \text{ exper}^2 \\
+ 0.029 \text{ tenure} - 0.00053 \text{ tenure}^2 \\
\]

Holding other things fixed, married women earn 19.8% less than single men (= the base category)
Ordinal Dummy Variables

• **Incorporating ordinal information using dummy variables**

• **Example: City credit ratings and municipal bond interest rates**

Municipal bond rate

Credit rating from 0-4 (0=worst, 4=best)

\[ MBR = \beta_0 + \beta_1 CR + \text{other factors} \]

This specification would probably not be appropriate as the credit rating only contains ordinal information. A better way to incorporate this information is to define dummies:

\[ MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + \text{other factors} \]

Dummies indicating whether the particular rating applies, e.g. CR_1=1 if CR=1 and CR_1=0 otherwise. All effects are measured in comparison to the worst rating (= base category).
Interactions among dummy variables

- Interactions involving dummy variables
- Allowing for different slopes

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

- Interaction term

$$\beta_0 = \text{intercept men} \quad \beta_1 = \text{slope men}$$

$$\beta_0 + \delta_0 = \text{intercept women} \quad \beta_1 + \delta_1 = \text{slope women}$$

- Interesting hypotheses

$$H_0: \delta_1 = 0$$  
The return to education is the same for men and women

$$H_0: \delta_0 = 0, \delta_1 = 0$$  
The whole wage equation is the same for men and women
Graphical illustration

Interacting both the intercept and the slope with the female dummy enables one to model completely independent wage equations for men and women.
Dummy-Continuous / Discrete Interaction (2)

- Testing for differences in regression functions across groups
- Unrestricted model (contains full set of interactions)

\[ \text{cumgpa} = \beta_0 + \delta_0 \text{female} + \beta_1 \text{sat} + \delta_1 \text{female} \cdot \text{sat} + \beta_2 \text{hsperc} + \delta_2 \text{female} \cdot \text{hsperc} + \beta_3 \text{tothrs} + \delta_3 \text{female} \cdot \text{tothrs} + u \]

- Restricted model (same regression for both groups)

\[ \text{cumgpa} = \beta_0 + \beta_1 \text{sat} + \beta_2 \text{hsperc} + \beta_3 \text{tothrs} + u \]
Dummy-Continuous /Discrete Interaction (3)

- Null hypothesis

\[ H_0 : \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0 \]

- Estimation of the unrestricted model

\[
\hat{cumgpa} = 1.48 - 0.353 \text{ female} + 0.011 \text{ sat} + 0.0075 \text{ female} \cdot \text{sat} - 0.0085 \text{ hisperc} - 0.0055 \text{ female} \cdot \text{hisperc} + 0.023 \text{ tothrs} - 0.0012 \text{ female} \cdot \text{tothours}
\]

Tested individually, the hypothesis that the interaction effects are zero cannot be rejected.
Models (with Dummy Variables)

- **Joint test with F-statistic**

  \[ F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} \approx \frac{(85.515 - 78.355)/4}{78.355/(366 - 7 - 1)} = 8.18 \]

  Null hypothesis is rejected

- SSRr is the sum of squared residuals from the restricted regression, i.e., the regression where we impose the restriction.
- SSRur is the sum of squared residuals from the full model,
- q is the number of restrictions under the null and
- k is the number of regressors in the unrestricted regression.
Binary dependent variable

- A Binary dependent variable: the linear probability model
- Linear regression when the dependent variable is binary

\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \]

\[ E(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \]

\[ E(y|x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x) \]

\[ P(y = 1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \]

\[ \beta_j = \frac{\partial P(y = 1|x)}{\partial x_j} \]

If the dependent variable only takes on the values 1 and 0

Linear probability model (LPM)

In the linear probability model, the coefficients describe the effect of the explanatory variables on the probability that \( y = 1 \)
Binary dependent variable: Example-1

- Example: Labor force participation of married women

\[
\hat{inlf} = 0.586 - 0.0034 \text{nwifeinc} + 0.038 \text{educ} + 0.039 \text{exper} \\
- 0.00060 \text{exper}^2 - 0.016 \text{age} - 0.262 \text{kidslt6} \\
+ 0.130 \text{kidsge6}, \ n = 753, R^2 = 0.264
\]

Non-wife income (in thousand dollars per year)

If the number of kids under six years increases by one, the probability that the woman works falls by 26.2%.

Does not look significant (but see below)
Binary dependent variable: Example-2

- Example: Female labor participation of married women (cont.)

Graph for nwifeinc=50, exper=5, age=30, kindslt6=1, kidsge6=0

The maximum level of education in the sample is educ=17. For the given case, this leads to a predicted probability to be in the labor force of about 50%.

Negative predicted probability (but no problem because no woman in the sample has educ < 5).