BS2280 - Econometrics 1

Lecture 8 - Part 1: Dummy Variables

Dr. Yichen Zhu

Structure of today's lecture

Intended Learning Outcomes

- Introducing qualitative variables into a regression model
- Illustrating the impact of a dummy using a regression line diagram graphically
- Interpreting dummies

Background

- The standard multiple regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

- So far we have assumed that the Y and X are **quantitative variables.** For example,

$$EARNINGS_i = \beta_1 + \beta_2 S_i + \beta_3 EXP_i + u_i$$

- **Quantitative variables**: can be quantified, i.e. are recorded with actual numbers
- **Examples**: earnings, schooling years, population, income, sales, age, GDP, distance, etc.

Qualitative Variables

- **Question**: Is it feasible for one or more of the X variables to be qualitative variables in the regression model?

- **Qualitative variables**: also called categorical variables, capture events/status. The data on them is not recorded with numbers, but classified into categories (with textual descriptions).
- **Examples**:
  - gender (male or female)
  - race (black, white, Asian, etc.)
  - smoking status (smoker/non-smoker)
  - political stability (stable, non-stable)
  - ownership (state-owned, private-owned)

Implementing Qualitative Variables

- What is the effect of one or more qualitative X variables (alongside quantitative X variables) on Y?
- **Examples**:
  - Does gender-based wage discrimination exist in the labour market?
  - Does smoking have an impact on life expectancy?
  - Are private-owned firms more efficient/productive than state-owned firms?
  - Do M&As raise firm profitability?

Dummy Variables

- Assume, we want to estimate the following regression:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 \textit{Qualitative.Variable}_i + u_i$$

  where X_2 captures a quantitative variable
- **Problem**: Qualitative variables do not have numbers on them, how do we introduce them in regression analysis?

- We do this through **dummy variables**!!!
- Dummy variables convert categories into binary variables (can only take numbers 0 and 1)
- i.e. gender: assigning the value 1 for males or 0 for females

Gender
Female
Male
Male
Female

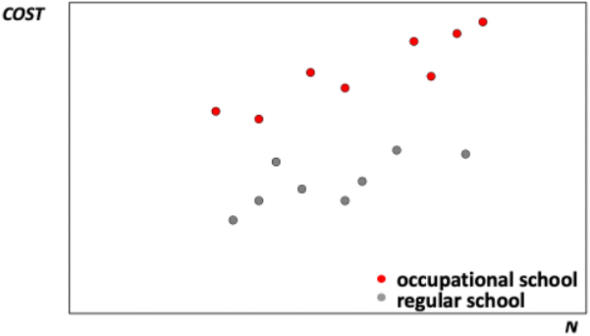
Gender
1
0
0
1

Example: Costs of Different School Types

- We have data on a sample of secondary schools
- This dataset has these variables:
  - *COST*: annual recurrent expenditure (¥)
  - *N*: number of students enrolled
  - *TYPE*: types of school: regular or occupational
- Note: Occupational schools aim to provide skills for specific occupations and they tend to be relatively expensive to run.

Example: Costs of Different School Types

- We present the data in a scatter plot
- Costs of occupational schools are higher than those of regular schools.



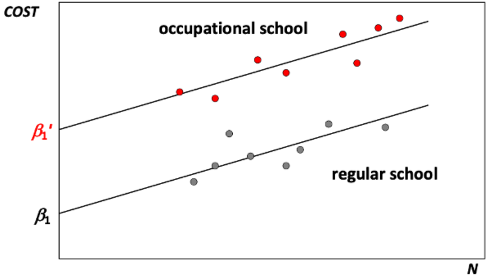
Example: Costs of Different School Types

- How can we estimate the difference in costs between regular and occupational schools?
- Two solutions

1. Run two separate regressions for the two types of school.
   Problem: Running regressions with two small samples instead of one large one will reduce the precision of the estimates of the coefficients.
2. Include a dummy variable in the regression
   Assumption: annual overhead cost is different for the two types of school, but the marginal cost is the same.

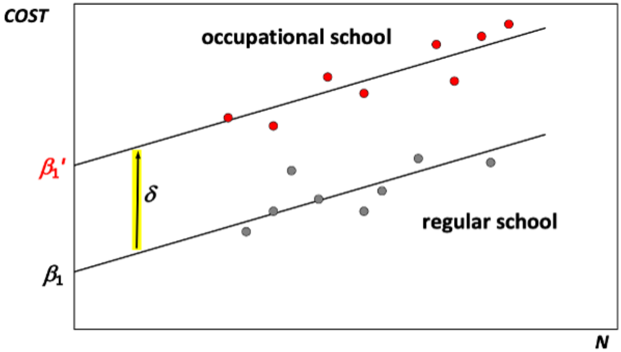
Example: Costs of Different School Types

- Occupational schools aim to provide skills for specific occupations and they tend to be relatively expensive to run.
- Therefore, we assume that the cost function for occupational schools has an intercept $\beta_1{}'$ that is greater than that for regular schools.



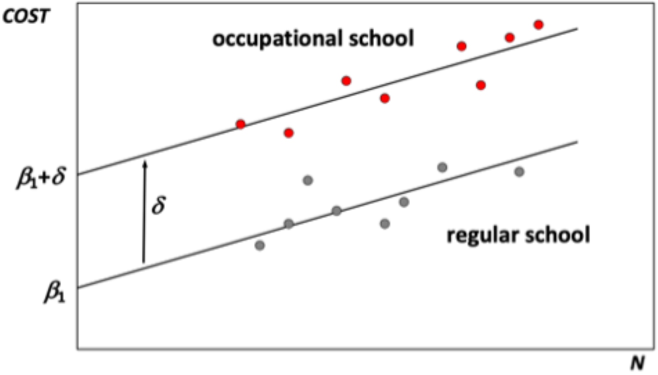
Regular school	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Occupational school	$COST_i = \beta_1{}' + \beta_2 N_i + u_i$

Example: Costs of Different School Types



Regular school	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Occupational school	$COST_i = \beta_1{}' + \beta_2 N_i + u_i$
Define difference in intercepts	$\delta = \beta_1{}' - \beta_1$

Example: Costs of Different School Types



Define difference in intercepts	$\delta = \beta_1{}' - \beta_1$ so rearrange get $\beta_1{}' = \beta_1 + \delta$
Regular school	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Occupational school	$COST_i = \beta_1{}' + \beta_2 N_i + u_i = \beta_1 + \delta + \beta_2 N_i + u_i$

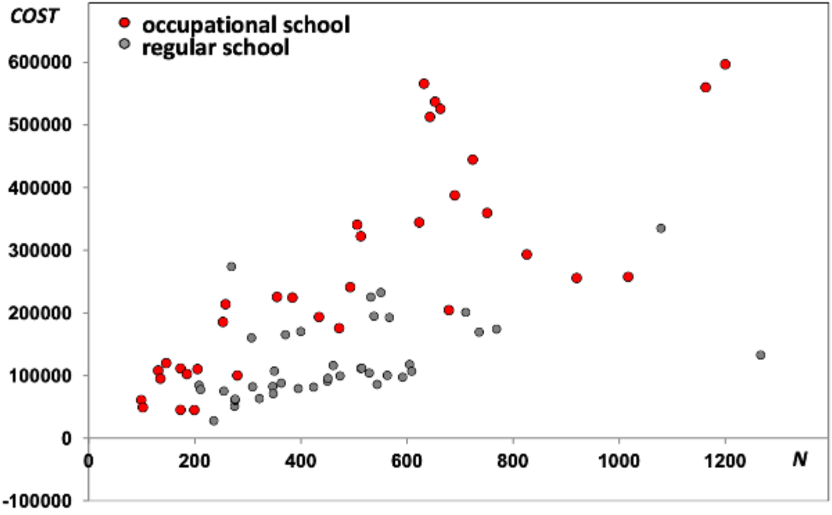
Example: Costs of Different School Types

- We can now combine above two functions by defining a dummy variable *TYPE*
- *TYPE*: A dummy variable taking a value of 0 if the school is a regular school, taking a value of 1 if the school is an occupational school

Regular school, $TYPE = 0$	$COST_i = \beta_1 + \beta_2 N_i + u_i$
Occupational school, $TYPE = 1$	$COST_i = \beta_1 + \delta + \beta_2 N_i + u_i$
Combined these two equations	$COST_i = \beta_1 + \delta TYPE_i + \beta_2 N_i + u_i$

- We will now use the actual data of 74 secondary schools in Shanghai.

Example: Costs of Different School Types



Example: Costs of Different School Types

- The table shows the data for the first 10 schools in the sample.
- *COST*: annual recurrent expenditure (¥)
- *N*: number of students enrolled
- *TYPE*: types of school: $Type = 0$ *Regular*; $Type = 1$ *Occupational*

School	Type	*COST*	*N*	*TYPE*
1	Occupational	345,000	623	1
2	Occupational	537,000	653	1
3	Regular	170,000	400	0
4	Occupational	526.000	663	1
5	Regular	100,000	563	0
6	Regular	28,000	236	0
7	Regular	160,000	307	0
8	Occupational	45,000	173	1
9	Occupational	120,000	146	1
10	Occupational	61,000	99	1

Example: Costs of Different School Types

- We estimate the following regression

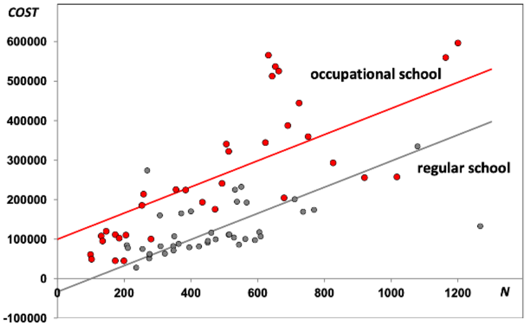
$$COST_i = \beta_1 + \beta_2 N_i + \delta TYPE_i + u_i$$

```
> costfit <- lm(COST~N+TYPE, data=schools)
> costfit

Call:
lm(formula = COST ~ N + TYPE, data = schools)

Coefficients:
(Intercept)            N         TYPE
   -33612.6        331.4     133259.1
```

Combined regression model	$\widehat{COST}_i = -33612.6 + 331.4 N_i + 133259.1\, TYPE_i$
Regular school, $TYPE = 0$	$\widehat{COST}_i = -33612.6 + 331.4 N_i + 133259.1 \times 0$ $= -33612.6 + 331.4 N_i$
Occupational school, $TYPE = 1$	$\widehat{COST}_i = -33612.6 + 331.4 N_i + 133259.1 \times 1$ $= 99646 + 331.4 N_i$

Example: Costs of Different School Types

Combined regression model	$\widehat{COST}_i = -33612.6 + 331.4N_i + 133259.1\,TYPE_i$
Regular school, $TYPE = 0$	$\widehat{COST}_i = -33612.6 + 331.4N_i$
Occupational school, $TYPE = 1$	$\widehat{COST}_i = 99646 + 331.4N_i$



Interpretation

- The dummy captures the difference in the average level of Y based on the dummy categories
- For example: On average, the costs of occupational schools ($TYPE = 1$) are 133,259 (¥) higher than for regular schools ($TYPE = 0$), ceteris paribus.

Combined regression model	$\widehat{COST}_i = -33612.6 + 331.4N_i + 133259.1\,TYPE_i$
Regular school, $TYPE = 0$	$\widehat{COST}_i = -33612.6 + 331.4N_i + 133259.1 \times 0$
Occupational school, $TYPE = 1$	$\widehat{COST}_i = -33612.6 + 331.4N_i + 133259.1 \times 1$

Hypothesis Testing

```
> summary(costfit)

Call:
lm(formula = COST ~ N + TYPE, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max
-253800  -49270   -8281   40403  257014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33612.55   23573.47  -1.426    0.158
N              331.45      39.76   8.337 3.97e-12 ***
TYPE        133259.08   20827.59   6.398 1.46e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89250 on 71 degrees of freedom
Multiple R-squared:  0.6156,	Adjusted R-squared:  0.6048
F-statistic: 56.86 on 2 and 71 DF,  p-value: 1.813e-15
```

$$\widehat{COST}_i = \underset{(23573.47)}{-33612.55} + \underset{(39.76)}{331.45}\ N_i + \underset{(20827.59)}{133259.08}\ TYPE_i \tag{1}$$

Hypothesis Testing, Use *t* test

- Hypothesis tests of dummies is similar to that of quantitative variables

$$COST_i = \beta_1 + \beta_2 N_i + \delta TYPE_i + u_i$$

$$\widehat{COST}_i = \underset{(23573.47)}{-33612.55} + \underset{(39.76)}{331.45}\ N_i + \underset{(20827.59)}{133259.08}\ TYPE_i \tag{2}$$

1. State the null and alternative hypotheses

Null Hypothesis	$H_0 : \delta = 0$
Alternative Hypothesis	$H_1 : \delta \neq 0$

2. Select the significance level. Significance level $\alpha = 5\%$
3. Select and calculate the test statistics

   Do not know the population variance σ^2, so use t statistic: $t = \frac{\hat{\delta} - \delta^0}{s.e.(\hat{\delta})} = \frac{\hat{\delta}}{s.e.(\hat{\delta})} = \frac{133259.08}{20827.59} = 6.398$
4. Set the decision rule. $n = 74$, degree of freedom = $n - k = 74 - 3 = 71$, $t_{crit,5\%} = 1.99$
5. Make statistical decisions. Make statistical decisions. $|t| = 6.398 > t_{crit,5\%} = 1.99$, reject the null $H_0 : \delta = 0$. *TYPE* is statistical significant, *TYPE* will affect *COST*.

Hypothesis Testing, Use p-value

$$\widehat{COST}_i = \underset{(23573.47)}{-33612.55} + \underset{(39.76)}{331.45} \; N_i + \underset{(20827.59)}{133259.08} \; TYPE_i \tag{3}$$

```
> summary(costfit)

Call:
lm(formula = COST ~ N + TYPE, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max
-253800  -49270   -8281   40403  257014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33612.55   23573.47  -1.426    0.158
N              331.45      39.76   8.337 3.97e-12 ***
TYPE        133259.08   20827.59   6.398 1.46e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89250 on 71 degrees of freedom
Multiple R-squared:  0.6156,     Adjusted R-squared:  0.6048
F-statistic: 56.86 on 2 and 71 DF,  p-value: 1.813e-15
```

- Use p-value: p-value = 1.46e-08 < 1%, variable is very significant (i.e. at the 1% level), reject null $H_0 : \delta = 0$.
- We conclude that *TYPE* is statistical significant, *TYPE* will affect *COST*.

Hypothesis Testing: Intercept

$$\widehat{COST}_i = \underset{(23573.47)}{-33612.55} + \underset{(39.76)}{331.45}\ N_i + \underset{(20827.59)}{133259.08}\ TYPE_i \tag{4}$$

```
> summary(costfit)

Call:
lm(formula = COST ~ N + TYPE, data = schools)

Residuals:
    Min      1Q  Median      3Q     Max
-253800  -49270   -8281   40403  257014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -33612.55   23573.47  -1.426    0.158
N              331.45      39.76   8.337 3.97e-12 ***
TYPE        133259.08   20827.59   6.398 1.46e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89250 on 71 degrees of freedom
Multiple R-squared:  0.6156,     Adjusted R-squared:  0.6048
F-statistic: 56.86 on 2 and 71 DF,  p-value: 1.813e-15
```

- For intercept, if we use *p*-value: *p*-value = 0.158 > 10%, variable is not significant (i.e. at the 10% level), cannot reject null $H_0 : \beta_1 = 0$.
- We conclude that the intercept is not statistical significant. Another possibility could be misspecification.

Student Task

- We regress hourly wages in USD on a gender dummy (1...female, 0...male) and an education variable (measured in years):

$$wage_i = \beta_1 + \beta_2 educ_i + \delta female_i + u_i$$

```
Call:
lm(formula = wage ~ educ + female, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9890 -1.8702 -0.6651  1.0447 15.4998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.62282    0.67253   0.926    0.355
educ         0.50645    0.05039  10.051  < 2e-16 ***
female      -2.27336    0.27904  -8.147 2.76e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.186 on 523 degrees of freedom
Multiple R-squared:  0.2588,	Adjusted R-squared:  0.256
F-statistic: 91.32 on 2 and 523 DF,  p-value: < 2.2e-16
```

- Using the R output, write down the wage equation for women and men separately and interpret the coefficients.
- Provide a regression line diagram to illustrate the differences between the wages of men and women based on education