ECO 372: Introduction to Econometrics

Spring 2025

Lecture 9: Dummy Variables in Econometric Modelling

Sakib Bin Amin, Ph.D.

Associate Professor in Economics

Director, Accreditation Project Team (APT)

Outline

Our objectives for this lecture will be to learn:

- What are Dummy Variables?
- Why Dummy Variables are Important?
- Measurement of Dummy Variables
- Modelling with Dummy Variables
- Visualisation of Dummy Variable Data
- Statistical Inferences of Dummy Variable Models
- Examples of Dummy Variable Uses From Literature
- Application of Dummy Variables in Models With STATA

What are Dummy Variables?

- So far the dependent and independent variables have been numerical or quantitative. For example: income, output, prices, costs, height, temperature.
- But this may not always be the case.
- There are occasions where the independent variables can be <u>qualitative in nature</u>.
- The qualitative variables are often known as dummy variables.
- Alternative names for dummy variables found in literature are : indicator variables, binary variables, categorical variables, and dichotomous variables.
- Some Examples: race, gender color, religion, nationality, geographical region, political upheavals, structural changes, policy reforms, natural calamity, party affiliation, etc.
- Used in all types of modelling framework: cross-section, time series, and panel.

Why Dummy Variables are Important?

- Dummy variables usually indicate the presence or absence of a "quality" or an attribute.
- For example, holding all other things constant, female workers are found to earn less than their male counterparts or nonwhite workers are found to earn less than whites.
- This pattern may result from gender or racial discrimination, but whatever the reason, qualitative variables such as gender and race seem to influence the independent variable(s).
- Therefore, they should be included among the explanatory variables, or the regressors for policy analysis or impact analysis if there is scope.
- Now, the question is, how to measure dummy variables?

Measurement of Dummy Variables

- Dummy are essentially nominal scale variables.
- In simple, a nominal scale variable does not have a natural order or ranking.
- One way we could "quantify" such attributes is by constructing artificial variables that take on values of 1 or 0.
- 1 indicating the presence (or possession) of that attribute and 0 indicating the absence of that attribute.
- This means a dummy variables can split the sample into two distinct groups.
- Mainly denoted by D.

$$D = 1; if\ gender\ is\ male$$

$$D = 0; if\ gender\ is\ female$$

- Other way around is also possible.

Visualisation of Dummy Variable Data

Respondent ID	Age	Education (Years of Education)	Gender (Male=1, Female=0)
1	18	12	1
2	19	10	0
3	20	8	0
4	21	7	0
5	23	12	0
6	25	12	1
7	26	12	0
8	28	12	0
9	31	12	0
10	33	8	0
11	35	12	1
12	37	5	1
13	19	5	1
14	20	12	1
15	21	12	1
16	23	12	1
17	25	16	1
18	26	13	1
19	28	13	0
20	30	16	0

Modelling with Dummy Variables

- Suppose we have the following single variable model:

$$Y_i = \beta_0 + \beta_1 D_i + u_i \ \ ; u_i \sim N\ (\mu, \sigma^2)$$

- If $D_i = 0$ then:

$$Y_i = \beta_0 + u_i$$

- If $D_i = 1$ then:

$$Y_i = \beta_0 + \beta_1 + u_i$$

- We can extend this modelling framework with other continuous variable(s):

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_{1i} + \cdots + \beta_k X_{ki} + u_i$$

- The use of dummy variables do not pose any new challenge for estimation, and we can use the traditional OLS to estimate parameters of the models.
- When all the explanatory variables are dummies, we refer such model as Analysis-of-Variance (ANOVA) models.
- Model with mix of dummy and continuous variables are called Analysis-of-Covariance (ANCOVA).

Statistical Inferences of Dummy Variable Models

- Let's take an simulated example.

- $D_i = 1$ is male and $D_i = 0$ is female.

$$Wage_i = \beta_0 + \beta_1 D_{gender,i} + \beta_2 Age_i + u_i$$

- We estimate the model with OLS and find the following:

$$\widehat{Wage}_i = 20 + 3.2 D_{gender,i} + 2.5 Age_i$$

- The observed data is split into 2 groups according to the gender dummy as shown earlier.

  - The group with $D_i = 0$ is called the baseline (i.e., female).

  - The group with $D_i = 1$ is called the target group/other group (i.e., male).

- β_1 of dummy quantifies the expected effect of the target group (i.e., male in our case).

Statistical Inferences of Dummy Variable Models

$$Wage_i = \beta_0 + \beta_1 D_{gender,i} + \beta_2 Age_i + u_i$$

$$\widehat{Wage}_i = 20 + 3.2 D_{gender,i} + 2.5 Age_i$$

- Expected value of wage if male (holding age constant):

$$E(Wage_i | D_{gender,i} = 1) = \beta_0 + \beta_1 = 20 + 3.2 = 23.2$$

- Expected value of wage if female(holding age constant):

$$E(Wage_i | D_{gender,i} = 0) = \beta_0 = 20$$

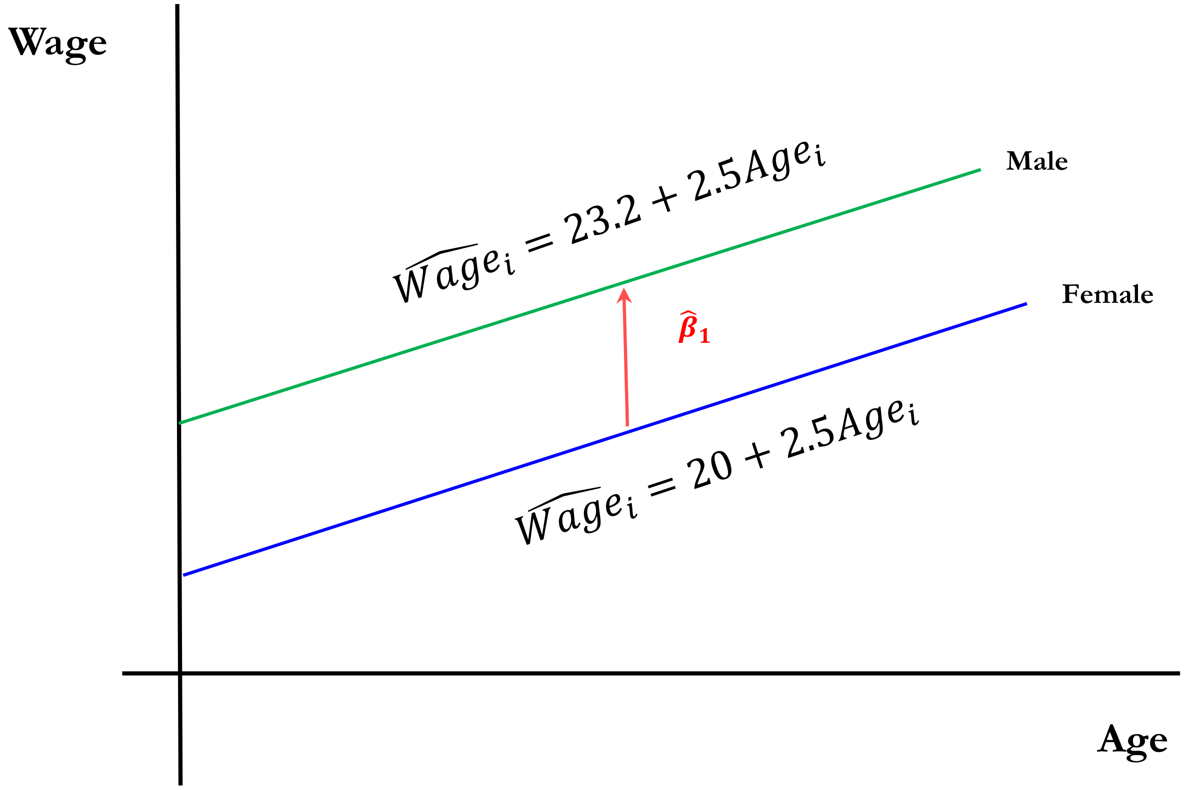
- Expected wage difference between male and female:

$$E(Wage_i | D_{gender,i} = 1) - E(Wage_i | D_{gender,i} = 0) = \beta_1 = 3.2$$

- Wage difference indicates males earn 3.2 units more compared to females, ceteris paribus.

Modelling and Statistical Inferences of Dummy Variable Models

- If we want to plot the models for $D_i = 0$ and $D_i = 1$, we will get something like the figure below:



Dummy Variable Trap

- The general principle of dummy variables can be extended to cases where there are several (but not infinite) discrete groups/categories.
- In some cases, this might lead to perfect multicollinearity in the model. This is known as the dummy variable trap.
- The good thing is all the current software algorithms can detect this issue automatically and remove one of the dummies from the model to avoid perfect multicollinearity.
- To illustrate, let's say we have three dummies that tell where the respondents live in a city:

Respondent ID	**C**	**D1 (north)**	**D2 (south)**	**D3 (east)**	**D4(west)**	**D1+D2+D3+D4**
1	1	1	0	0	0	1
2	1	0	1	0	0	1
3	1	0	0	1	0	1
4	1	0	0	0	1	1

- In this example, we can show, for example, $C = D1 + D2 + D3 + D4$, which is case of perfect multicollinearity
- Therefore as <u>a rule of thumb</u>, when closely related multiple dummies are present, <u>$N-1$ dummies</u> should be used, given N is the number of dummies.

Examples of Dummy Variable Uses From Literature

TABLE 4
Impact of Electricity Access on Women Empowerment in Urban Areas

Variables	Economic Freedom	Economic Decision	Household Decision	Mobility and Agency
Electricity (Hrs)	0.0336***	0.0572*	0.0550***	–0.0272
	(0.0108)	(0.0300)	(0.0199)	(0.0204)
Log (Income)	0.590	–5.159***	–0.440	–0.589
	(2.195)	(1.755)	(1.259)	(0.949)
Log (Income)^2	–0.0293	0.244***	0.00974	0.0306
	(0.107)	(0.0850)	(0.0628)	(0.0462)
Household Size	–0.0267	–0.0133	0.00913	–0.0264
	(0.0220)	(0.0497)	(0.0324)	(0.0280)
Household Head Sex (Female = 1)[a]	0.237***	0.567**	0.480***	0.696***
	(0.0890)	(0.223)	(0.160)	(0.152)
Woman Age	0.0517***	0.0627*	0.0844***	0.0113
	(0.0139)	(0.0320)	(0.0229)	(0.0228)
Woman Age^2	–0.000562***	–0.000662*	–0.000906***	–0.000164
	(0.000167)	(0.000384)	(0.000270)	(0.000268)
Ethnicity (Indigenous = 1)	–0.0138	–0.201	0.0290	–0.240
	(0.170)	(0.331)	(0.241)	(0.223)
Number of Children	0.122	0.188	–0.0128	–0.0776
	(0.0969)	(0.263)	(0.169)	(0.153)
Married (Yes = 1)	0.0239	–0.263	–0.0163	–0.135
	(0.134)	(0.292)	(0.203)	–0.0776
Controls				
District FE	Yes	Yes	Yes	Yes
Household Type FE	Yes	Yes	Yes	Yes
Instruments	76	76	76	76
Constant	–4.960	28.87***	0.539	1.799
	(11.23)	(10.47)	(6.810)	(5.611)
N	221	221	221	221

[a] Even though the female headed households are low (11%) in urban areas but 62.5% of those households are indigenous. We also believe some ethnicity aspects are suppressed due to urban dynamics. Perhaps, these together results high significance of the coefficients.
Robust standard errors in parentheses
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.15$

- Look at the Household Head variable. It's a dummy. It takes 1 for female and 0 otherwise.
- The estimation suggests that in urban areas of the Chittagong Hill Tracts of Bangladesh, women's economic freedom is higher by <u>0.237 units</u> in <u>female headed households</u> compared to <u>male headed households.</u>

Amin, S. B., Jamasb, T., Khan, F., & Nepal, R. (2024). Electricity access, gender disparity, and renewable energy adoption dynamics: The case of mountain areas of Bangladesh. *Economics of Energy & Environmental Policy*, 13 (1), DOI: 10.5547/2160-5890.13.1.sami

Application of Dummy Variables in Models with STATA

- We take the following models for estimations:

$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 Black_i + u_i$$
$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 SMSA_i + u_i$$
$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 Black_i + \beta_2 SMSA_i + u_i$$

Here:

$Black_i = 1 \; if \; the \; person \; is \; black$

$Black_i = 0 \; if \; the \; person \; is \; white$

$SMSA_i = 1 \; if \; the \; person \; is \; not \; from \; metropolitan \; area$

$SMSA_i = 0 \; if \; the \; person \; is \; from \; metropolitan \; area$

Application of Dummy Variables in Models with STATA

Table to show the summary of variable Black

1 if black	Freq.	Percent	Cum.
0	20,483	71.78	71.78
1	8,051	28.22	100.00
Total	28,534	100.00	

Table to show the summary of variable SMSA

1 if not SMSA	Freq.	Percent	Cum.
0	20,469	71.76	71.76
1	8,057	28.24	100.00
Total	28,526	100.00	

Table to show the distributions of black and white employees living in the metropolitan areas

1 if black	Freq.	Percent	Cum.
0	6,163	76.49	76.49
1	1,894	23.51	100.00
Total	8,057	100.00	

Application of Dummy Variables in Models with STATA

$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 Black_i + u_i$$

Linear regression

Number of obs	=	28,510
F(2, 28507)	=	1358.57
Prob > F	=	0.0000
R-squared	=	0.0944
Root MSE	=	.45499

ln_wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0196842	.0004233	46.51	0.000	.0188546	.0205138
black	-.1387483	.0059396	-23.36	0.000	-.1503903	-.1271064
_cons	1.142352	.0120125	95.10	0.000	1.118807	1.165897

Follow model selection handout for interpretation .

Estimate shows black employees tend to earn around 13.87% less than the white employees, considering all are constant.

Application of Dummy Variables in Models with STATA

$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 SMSA_i + u_i$$

```
Linear regression                               Number of obs     =     28,502
                                                F(2, 28499)       =    1919.00
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1272
                                                Root MSE          =     .44671
```

ln_wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0202263	.0004153	48.71	0.000	.0194124	.0210403
not_smsa	-.2370934	.0057822	-41.00	0.000	-.2484267	-.22576
_cons	1.154443	.0117055	98.62	0.000	1.1315	1.177386

Application of Dummy Variables in Models with STATA

$$lnWage_i = \beta_0 + \beta_1 Age_i + \beta_2 Black_i + \beta_2 SMSA_i + u_i$$

Linear regression

Number of obs	=	28,502
F(3, 28498)	=	1555.19
Prob > F	=	0.0000
R-squared	=	0.1484
Root MSE	=	.44126

ln_wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0200584	.0004116	48.73	0.000	.0192516	.0208653
black	-.1550226	.0056919	-27.24	0.000	-.166179	-.1438662
not_smsa	-.2471931	.0056668	-43.62	0.000	-.2583003	-.2360859
_cons	1.205903	.0117308	102.80	0.000	1.18291	1.228896

STATA Commands

```
webuse regsmpl
		tab black
		tab not_smsa
		tab black if not_smsa==1
*ols with black dummy
reg ln_wage age black, vce(r)
*ols with smsa dummy
reg ln_wage age not_smsa, vce(r)
*ols with both bakck and smsa dummies
reg ln_wage age black not_smsa, vce(r)

//general method of dummy variable creation with STATA
*Read: Speaking STATA: How best to generate indicator or dummy variables by Nicholas J. Cox and Clyde B. Schechter
sysuse auto
*suppose you need a dummy that shows 1=mpg higher than 30 and 0=otherwise
gen hi_mpg = 1 if mpg > 30
		ed hi_mpg mpg if mpg<=30.
		replace hi_mpg=0 if mpg<=30
		ed hi_mpg mpg
*the first line creates a variable hi_mpg that takes value 1 when mpg variable is greater than 30. Less than and equal to 30 are still missing.
**second line will show you a dot (.) in hi_mpg when mpg<=30. dot(.) indicates missing value in STATA.
** third line addresses this issue and makes all dots (.) as 0 when mpg is <=30 with "replace" command.
** fourth line shows the actual mpg and hi_mpg data in a spreadsheet with "ed".
```