ECO 391 Economics and Business Statistics

Lecture 9: Regression with Dummy Variables

Xiaozhou Ding

April 10, 2019

Dummy Variables

Dummy Variables

- Quantitative Variables: In previous chapters, most of the variables used in regression applications have been quantitative (Few of them are qualitative variables).
- Examples:
  - Income: Measured in dollars ranging from 0 up to $1,000,000 and higher.
  - Age: Measured in years ranging from 1 to 100 or perhaps higher.
  - Points scored, revenue, expenditures, height, miles per gallon, grade percentage earned...

Dummy Variables

- Qualitative Variables: Variables that represent qualities or characteristics and are not typically measured in numerical terms.
- Examples:
  - Male or female
  - Race (Caucasian, African-American, Native-American, Asian American, etc.)
  - Year at School: freshman, sophomore, junior, or senior
  - State residence, nationality, “yes or no” questions, responses of opinion ...
- Quantitative variables assume meaningful numerical values, whereas qualitative variables represent categories

Dummy Variables

- A qualitative variable with two categories can be associated with a dummy variable. A dummy variable is defined as a variable that assumes a value of 1 for one of the category and 0 for the other.
- A dummy variable is also referred to as an indicator variable.
- Various Uses of Dummy Variables:
  - war time vs. peace time
  - strike vs. nonstrike period
  - Urban v.s. rural
  - Treatment v.s. Non-treatment (control/comparison)
  - Male vs. female
  - Age over 55 vs. age below 55

Dummy Variable as Independent Variable

- Any observation for which the dummy variable is equal to zero is the base case, also referred as reference category.
- The coefficient of a dummy variable measures the difference between being in the base case and not being in the base case.
- The coefficient of a dummy variable capture the shift of intercept.

Qualitative Variables with Two Categories

- For example, suppose you are interested in the impact of gender on salary , gender wage gap between male and female professors. We might first define a dummy variable d that has the following structure:

$$\text{Let } d = 1 \text{ if gender} = \text{“male”}$$

$$\text{and } d = 0 \text{ if gender} = \text{“female”}$$

- This allows us to include a measure for gender in a regression model and quantify the impact of gender on salary.

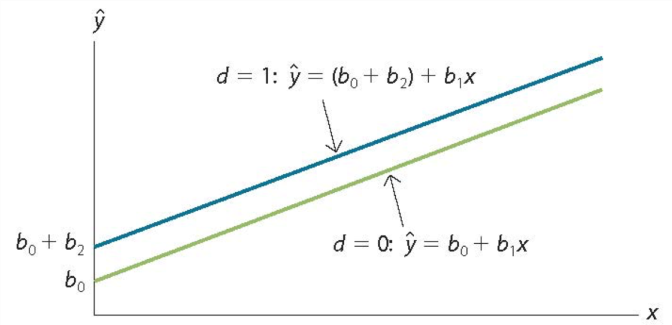
Qualitative Variables with Two Categories

- This allows us to include a measure for gender in a regression model and quantify the impact of gender on salary.
- y_i = annual salary in 1,000 dollars of professor$i$
- x_i = years of seniority (experience) in the workplace of person$i$
- $d = 1$ if professor i is male
- $d = 0$ if professor i is female
- Regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \epsilon$$

- Sample regression equation:
  - For a female professor $\hat{y} = b_0 + b_1 x_i$
  - For a male professor $\hat{y} = b_0 + b_1 x_i + b_2 \times 1 = b_0 + b_1 x_i + b_2 = (b_0 + b_2) + b_1 x_i$

Regression with a Dummy Variable



The slope is unchanged-only the intercept term differs, therefore these dummy variables are called intercept dummies.

- Suppose that we estimate this regression and the prediction equation is

$$\hat{y} = 29.43 + 1.23x + 13.88d$$

- The expected salary of a female professor with x_i years of seniority is:

$$E(y_i|x_i, d_i = 0) = 39.43 + 1.23x_i + 13.88 \times 0 = 39.43 + 1.23x_i$$

- The expected salary of a male professor with Xi years of seniority is:

$$E(y_i|x_i, d_i = 1) = 39.43 + 1.23x_i + 13.88 \times 1 = 53.31 + 1.23x_i$$

- What is starting salary for a male professor? $39,430
- What is starting salary for a female professor? $53,310
- Verbal interpretation: Female professors have a starting salary of $39,430 and for every year increase in seniority, the expected salary increases by $1,230. Male professors have a starting salary of $53,310 and for every year increase in seniority, the expected salary increases by $1,230.
- What if we put $d = 1$ if gender="female"

Test for Differences between the Categories of a Qualitative Variable

- The statistical tests discussed in Chapter 15 remain valid for dummy variables as well.
- We can perform a t-test for individual significance, form a confidence interval using the parameter estimate and its standard error, and conduct a F-test for joint significance.

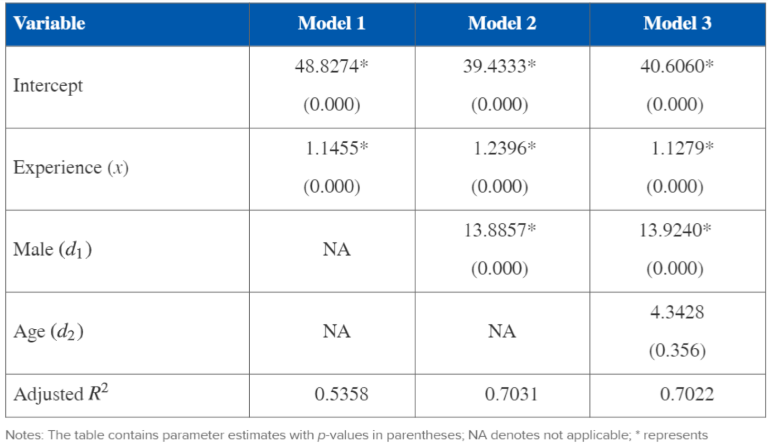
Example: Salaries, Gender, and Age

- Is there a gender effect in the salary study?

$$\begin{aligned} H_0 : &\beta_2 = 0 \quad \text{males and females are paid the same} \\ H_A : &\beta_2 \neq 0 \quad \text{there is a difference due to gender} \end{aligned}$$

- Given a value of the t_{df} test statstic of 4.86 and p-value of approximate 0.00, we reject the null hypothesis and conclude that the gender dummy variable is significant.
- For the age coefficient, t_{df} is 0.94, and the p-value is 0.36, so we fail to reject the null hypothesis. The evidence suggests that professors over 60 do not have significantly different salaries compared to those under 60.

Which Model You Should Use for Prediction



Variable	Model 1	Model 2	Model 3
Intercept	48.8274* (0.000)	39.4333* (0.000)	40.6060* (0.000)
Experience (x)	1.1455* (0.000)	1.2396* (0.000)	1.1279* (0.000)
Male (d_1)	NA	13.8857* (0.000)	13.9240* (0.000)
Age (d_2)	NA	NA	4.3428 (0.356)
Adjusted R^2	0.5358	0.7031	0.7022

Notes: The table contains parameter estimates with *p*-values in parentheses; NA denotes not applicable; * represents

Qualitative Variables with More than Two Categories

- Some qualitative variables have more than two classes.
- Example: Quarterly returns of a stock. Y is the return of a stock, X include some dummy variables indicating quarters.
  - Quarter 1: January – March
  - Quarter 2: April – June
  - Quarter 3: July – September
  - Quarter 4: October – December
- If there are four classes for the variable, we need to include 3 variables in the regression equation
- number of dummies = number of classes - 1

- Whichever one we do not include in the equation is the base case - let's say the 4th quarter
  - Q1 = 1 if the observation comes from the 1st quarter and 0 otherwise
  - Q2 = 1 if the observation comes from the 2nd quarter and 0 otherwise
  - Q3 = 1 if the observation comes from the 3rd quarter and 0 otherwise
- Practice
  - If an observation comes from the fourth quarter: Q1=0, Q2=0, Q3=0, Q4=1 (which is the base case)
  - If an observation comes from the first quarter: Q1=1, Q2=0, Q3=0, Q4=0
  - If an observation comes from the second quarter: Q1=0, Q2=1, Q3=0, Q4=0
  - If an observation comes from the third quarter: Q1=0, Q2=0, Q3=1, Q4=0

- The coefficient of Q1 measures the difference in being in the first quarter and the base case (4th quarter).
- Note you cannot have one Q variable and let it take on the values 0, 1, 2, 3.

Example: A Bond's Rating and Its Price

- Dependent variable: price of the bond
- Independent variables:
  - I_i = the interest rate of the ith bond in percentage points
  - T_i = the interest rate of the t-bills in percentage points on day bond was sold
  - $E_i = 1$ if income from the ith bond is tax exempt
  - Also each bond has a Moody's rating of AAA, AA, or A
- Questions: How many dummy variables do we need to account for this classification?

- The model is

$$P_i = \beta_0 + \beta_1 I_i + \beta_2 T_i + \beta_3 E_i + \beta_4 R_{i1} + \beta_5 R_{i2} + \epsilon_i$$

- Suppose that the estimated model turns out to be:

$$\widehat{P_i} = 93.615 + 0.764 I_i - 1.357 T_i + 6.166 E_i + 7.213 R_{i1} + 6.931 R_{i2}$$

- Interpret the coefficient of E_i:
  The price of a tax-exempt bond is \$6.17 higher than a nonexempt bond holding all else constant, on average.
- Interpret the coefficient of R_{i1}:
  The price of an AAA rated bond is \$7.21 higher than the price of an A-rated bond holding all else constant, on average.
- Interpret the coefficient of R_{i2}:
  The price of an AA rated bond is \$6.93 higher than the price of an A-rated bond holding all else constant, on average.

- How much more does a AAA bond sell for than a AA bond?
  - Premium of AAA – Premium of AA Bond (compared to a A bond)
  - 7.213 - 6.931 or 28 cents is the difference in price between an AAA rated bond and an AA rated bond.
- Predict the selling price of a tax-exempt bond with an AA rating where the interest rate is 6 percent, the T-bill rate is 8 percent.
  - So $I = 6$ and $T = 8$ and $E = 1$ and $R_{i1} = 0$ and $R_{i2} = 1$.
  - Plug those values into the estimate model:

$$\widehat{P} = 93.615 + 0.764 \times 6 - 1.35 \times 8 + 6.166 \times 1 + 7.213 \times 0 + 6.931 \times 1 = \$100.44$$

Avoiding the Dummy Variable Trap

- Given the intercept term, we exclude one of the dummy variables from the regression, where the excluded variable represents the reference category against which the others are assessed.
- If we included as many dummy variables as categories, this would create perfect multicollinearity in the data, and such a model cannot be estimated.
- So, we include one less dummy variable than the number of categories of the qualitative variable.

Interactions with Dummy Variables

- So far we have used dummy variables to allow shifts in the intercept.
- We can also use a dummy variable to create an interaction variable, which allows the estimated change in y to vary across x.
- The product xd captures the interaction between a quantitative dummy variable x and a dummy variable d. Together, the variables d and xd allow the intercept as well as the slope to vary between categories of a qualitative variable.

Modelling Interactions

- Consider the following regression model:

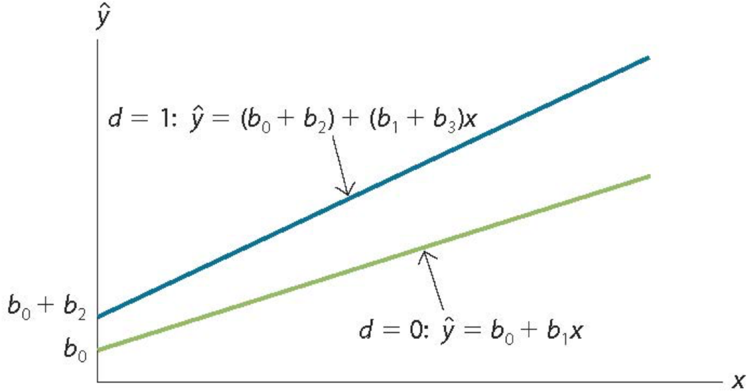
$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \epsilon$$

- When we use sample data to estimate this equation, we obtain:

$$\hat{y} = b_0 + b_1 x + b_2 d + b_3 xd$$

- When $d = 1$, $\hat{y} = (b_0 + b_2) + (b_1 + b_3)x$ and
- When $d = 0$, $\hat{y} = b_0 + b_1 x$

Shifts in the Intercept and the Slope


$\hat{y}$
$d = 1\colon\ \hat{y} = (b_0 + b_2) + (b_1 + b_3)x$
$b_0 + b_2$
b_0
$d = 0\colon\ \hat{y} = b_0 + b_1 x$
x

Testing for Significance

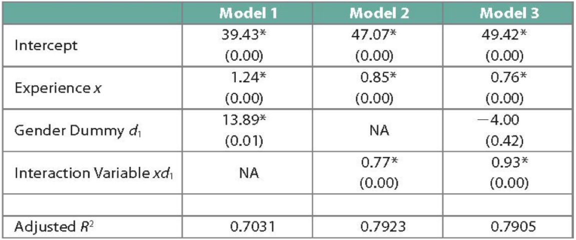
- Consider our regression model with both a dummy variable and an interaction variable:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \epsilon$$

- We are still able to perform the standard t test for individual significance and partial F test for joint significance. No adjustments need to be made to either procedure.

Example

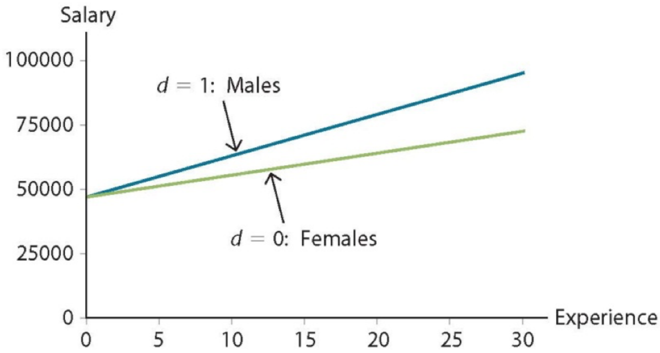
- We return to the introductory case and are interested in whether gender impacts salary differently at different levels of experience. Does additional experience get a higher reward for one gender over the other?
- Since age was not significant earlier, we consider three models, one with a dummy variable for gender, one with an interaction variable between gender and experience, and one with both a dummy variable and an interaction variable.
- As before, we keep experience as a quantitative explanatory variable.
  - Model 1: $y = \beta_0 + \beta_1 x + \beta_2 d + \epsilon$
  - Model 2: $y = \beta_0 + \beta_1 x + \beta_2 xd + \epsilon$
  - Model 3: $y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \epsilon$



	Model 1	Model 2	Model 3
Intercept	39.43* (0.00)	47.07* (0.00)	49.42* (0.00)
Experience x	1.24* (0.00)	0.85* (0.00)	0.76* (0.00)
Gender Dummy d_1	13.89* (0.01)	NA	−4.00 (0.42)
Interaction Variable xd_1	NA	0.77* (0.00)	0.93* (0.00)
Adjusted R^2	0.7031	0.7923	0.7905

- Model 1 use a Male dummy variable to allow salaries between males and females to differ by a fixed amount, irrespective of experience. The estimated model implied that, on average, males earn about 13.89×1000 more than females at all levels of experiences.
- Model 2 Uses an interaction variable $x \times d$ to allow the difference in salaries between males and females to vary with experience. Since the coefficient related with xd is significant. With every extra year of experience, the estimated difference in salaries between males and females increased by about $0.77 \times 1000 = 770$.
- Model 3 uses d and xd to allow a fixed as well as a varying difference in salaries between males and females. However, d1 now is no longer statistically significant. But xd is still significant, suggesting that with every extra year of experience, the estimated difference in salaries between males and females increases by $0.93 \times 100 = 930$.

Predicted Salaries



- The interaction term allows for male professors to have a different slope coefficient than female professors.
- Conceptually, experience impacts the salary of each gender differently.
  - For example, with 10 years of experience, what is the salary difference between males and females?

Binary Choice Model

- So far, we have been considering models where dummy variables are used as explanatory variables.
- There are, however, many applications where the variable of interest, the response variable, is binary.
- Consumer choice literature has many applications including whether to buy a house, join a health club, or go to graduate school.

Binary Choice Model

Definition

Regression models that use a dummy (binary) variable as the response variable are called binary choice models. They are also referred as discrete choice models or qualitative response models.

A linear regression models applied to a binary response variables is called a linear probability model (LPM)

Liner Probability Model

- When the response variable assumes values of only 0 and 1, we refer to the model as a binary choice model.
- A linear probability model (LPM) is formulated as

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

  where y assumes a 0 or 1 value and $P(y = 1)$ is the probability of sucess,
- Predictions with an LPM are interpreted as probabilities of y taking on the value of 1, and are made by

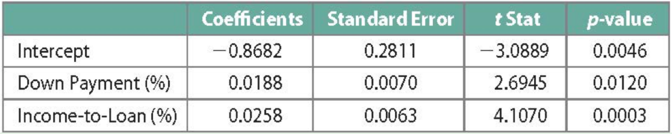
$$\widehat{P} = \hat{y} = b_0 + b_1 x$$

Example

- We are interested in whether the size of the down payment and the income-to-loan ratio influence whether or not an applicant is approved for a mortgage.
- Consider 30 loan applications at a bank (16 were approved). If x_1 is the down payment offered as a percentage of the loan, x_2 is the income-to-loan ratio of the applicant, and $y = 1$ if the applicant receives a loan, then we estimate the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Example



	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	−0.8682	0.2811	−3.0889	0.0046
Down Payment (%)	0.0188	0.0070	2.6945	0.0120
Income-to-Loan (%)	0.0258	0.0063	4.1070	0.0003

- Both variables are positive and significant, suggesting that a higher down payment and a higher income-to-loan ratio both make getting a loan more likely.
- To predict the probability that an applicant who wants a laon for \$200,000 with an income of \$60,000 and a down payment of 20%, we compute:

$$\hat{y} = -0.8682 + 0.0188 \times 20 + 0.0258 \left(\frac{60000}{200000} \times 100 \right) = 0.2818$$

Interpreting the Coefficients

- The coefficient of 0.0188 on Down Payment indicates that a 1 percent increase in the down payment will increase the probability of getting a loan by 0.0118 (1.88 percentage points)
- Similarly, a 1 percent increase in the income-to-loan ratio will increase the probability of getting a loan by 0.0258 (2.58 percentage points)
- One shortcoming of the LPM: if the down payment and the income-to-loan ratios are 60% and 30%, respectively, then the predicted probability is 1.0338, an impossible probability.

Weakness of LPM

- In the linear probability model, predicted values are interpreted as the probability that y equals 1.
- Suppose $\hat{y} = -0.2 + 0.4x$. If $x = 0$ is plausible, then we will predict that the probability that $y = 1$ is less than 0.
- Alternatively, the model may predict a probability greater than 1.
- By no means is this a fatal flaw of linear probability models, but certainly a shortcoming.

The Logit Model

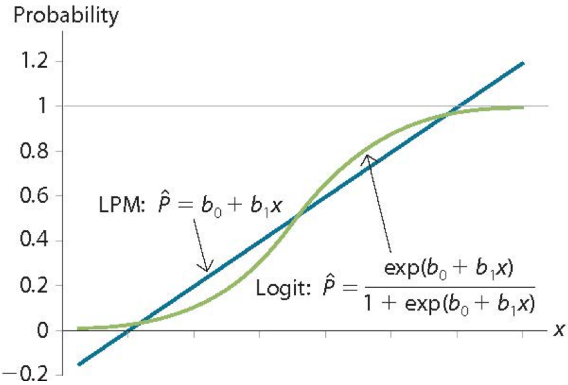
- In order to address the problem of the LPM that the predicted probabilities may be negative or greater than 1, we consider an alternative specification called the logistic model, more typically referred to as a logit model.
- A logit model uses a nonlinear regression function that insures that the result is always in the interval [0,1].
- However, this feature comes with a cost, as interpreting the coefficients becomes more complicated and estimation cannot be done by OLS.

Logistic Regression

- Consider the equation:
$$P = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$
- This nonlinear specification ensures that the probability is between 0 and 1 for all values of x.
- Unfortunately, the coefficients in this model cannot be estimated by ordinary least squares. A technique known as maximum likelihood estimation is used instead.
- Once the coefficients are estimated, the prediction function of this logistic regression is
$$\widehat{P} = \frac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$$

MLE is not offer in excel, but it is important to be able to interpret and make predictions with the estimated logit model

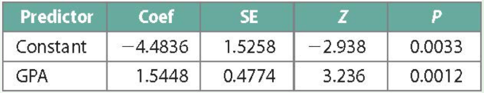
Logit versus Linear Probability Model



- The graph above shows the logit model and the LPM give similar $\widehat{P}$ values for most of the x range, given $b_1 > 0$.
- For small and large x values, the LPM may yield invalid probabilities.
- The logit model always return a value in [0,1].

Example

- There is a declining interest among teenagers in pursuing a career in science. In a survey, 50% high school students showed no interest in the sciences. An educator wants to determine if a student's interest in science is linked with the student's GPA.
- She collected 120 student's data and uses R to estimate a logit model where a student's choice of field (1=science, 0=other) is predicted by GPA.



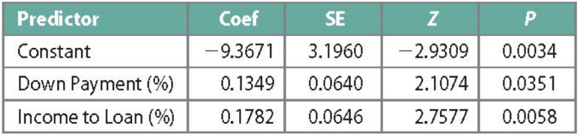
Predictor	Coef	SE	Z	P
Constant	−4.4836	1.5258	−2.938	0.0033
GPA	1.5448	0.4774	3.236	0.0012

- With a *p*-value of 0.0012, GPA is indeed a significant factor in predicting whether a student chooses science.
- Since the regression coefficient $b_1 = 1.5448$ is positive, we can infer that GPA exerts a positive influence on the predicted probability that a student chooses science.
- For a student with a GPA=3.0, we compute the predicted probability as

$$\widehat{P} = \frac{\exp(-4.4836 + 1.5448 \times 3.0)}{1 + \exp(-4.4836 + 1.5448 \times 3.0)} = 0.54$$

Compare LPM and Logit

- Let's revisit the mortgage example. If we instead utilize a logit model, then our estimated coefficients will be:

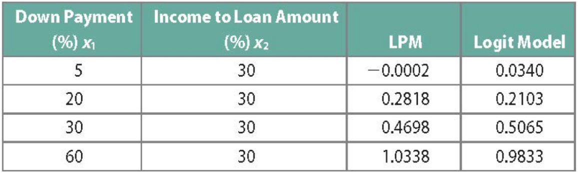


Predictor	Coef	SE	Z	P
Constant	−9.3671	3.1960	−2.9309	0.0034
Down Payment (%)	0.1349	0.0640	2.1074	0.0351
Income to Loan (%)	0.1782	0.0646	2.7577	0.0058

- Inserting the coefficients into the expression for $\widehat{P}$, we can find the predicted probability for loan approval:

$$\widehat{P} = \frac{\exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}{1 + \exp(-9.3671 + 0.1349x_1 + 0.1782x_2)}$$

Prediction Comparison



Down Payment (%) x_1	Income to Loan Amount (%) x_2	LPM	Logit Model
5	30	−0.0002	0.0340
20	30	0.2818	0.2103
30	30	0.4698	0.5065
60	30	1.0338	0.9833

- Compared to the linear probability model, the logit model does not predict probabilities less than zero or greater than one.
- Therefore, whenever possible, it is generally preferable to use the logit model rather than the linear probability model.