Topic 7

Dummy Variables and Multicollinearity

- In many applications in applied econometrics we wish to include as explanatory variables in our regression model **qualitative** or **categorical** variables.
- As the name suggests, categorical variables are variables which indicate what category a given observation fits in.
- Examples of categorical variables are:
  - An individual's gender.
  - An individual's race.
  - Whether or not an individual votes for a particular political party.
  - Whether or not an individual has private health insurance.
  - Whether or no an individual is employed or unemployed.
  - Whether or not an individual has ever been convicted of a crime.
- These variables provide qualitative rather than quantitative information.

- The simplest kind of categorical variable is one for which there are only two possible categories, such as employed/unemployed and married/unmarried.
- Categorical variables for which there are only two possible outcomes are called **binary variables**.
- The inclusion of categorical variables as regressors in the linear regression model enables us to estimate the effects of qualitative factors, such as race and gender, on the dependent variable.
- We can include qualitative information in a linear regression model by defining an appropriate **dummy variable** (or **indicator variable**).
- A binary dummy variable is a variable which takes on one of two values, typically 0 or 1, depending on whether or not a particular condition is satisfied.

- For example, we could define a gender dummy variable called *male* as follows:

$$male_i = \left\{ \begin{array}{l} 1 \text{ if individual i is male,} \\ 0 \text{ if individual i is female.} \end{array} \right| , i = 1, 2, ..., n.$$

- Alternatively, we could define a gender dummy variable called *female* as follows:

$$female_i = \left\{ \begin{array}{l} 1 \text{ if individual i is female,} \\ 0 \text{ if individual i is male.} \end{array} \right| , i = 1, 2, ..., n.$$

- By including a gender dummy variable in our linear regression model, we can estimate the effect of gender on the dependent variable.
- For example, suppose that we are interested in modelling the effect of gender and education on the average (hourly) wage.

- Assume that

$$E(wage_i|female_i, educ_i) = \beta_0 + \delta_0 female_i + \beta_1 educ_i. \quad (1)$$

Then, the average wage for females with a given level of education is

$$E(wage_i|female_i = 1, educ_i) = \beta_0 + \delta_0 + \beta_1 educ_i, \quad (2)$$

and the average wage for males with the same level of education is

$$E(wage_i|female_i = 0, educ_i) = \beta_0 + \beta_1 educ_i. \quad (3)$$

- Since

$$\begin{aligned} E(wage_i|female_i &= 1, educ_i) - E(wage_i|female_i = 0, educ_i) \\ &= \beta_0 + \delta_0 + \beta_1 educ_i - \beta_0 - \beta_1 educ_i \\ &= \delta_0, \end{aligned}$$

the coefficient δ_0 measures the difference between the average wage for females with a given level of education and the average wage of males with the same level of education.

- That is, δ_0 in

$$E(wage_i|female_i, educ_i) = \beta_0 + \delta_0 female_i + \beta_1 educ_i. \quad (1)$$

measures the marginal effect of gender on average wages, controlling for education.

- The regression equation associated with (1) is

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, i = 1, 2, ...n. \quad (4)$$

- Note that (4) implies that for a female the wage equation is

$$wage_i = \beta_0 + \delta_0 + \beta_1 educ_i + u_i,$$

and for a male the wage equation is

$$wage_i = \beta_0 + \beta_1 educ_i + u_i.$$

- Therefore, the regression model specified in (4) allows for a **different intercept** in the wage equation for males and females.
- For females the intercept is $\beta_0 + \delta_0$, and for males the intercept is β_0.

- For this reason, the dummy variable *female* is called an **intercept dummy**.
- We can test whether or not, controlling for education, averages wages vary by gender by estimating

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, i = 1, 2, ...n. \quad (4)$$

and testing

$$\begin{aligned} H_0 &: \delta_0 = 0 \\ H_1 &: \delta_0 < 0. \end{aligned}$$

- The form of the alternative hypothesis is based on the assumption that either gender has no effect on average wages, or the average wage for females is less than the average wage for males with the same level of education.

- Some possible justifications for this assumption are:
  - On average, females may be attracted to occupations which offer relatively low wages, because of the non-pecuniary characteristics of such occupations.
  - Females may lose seniority by taking time out of the labor market to raise children.
  - There may be gender based discrimination in the labor market.
- When data from a random sample of 526 observations is used to estimate

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, i = 1, 2, ...526, \quad (4)$$

by OLS we obtain the output reported in Figure 1 below.

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.6228	0.6725	0.9261	0.3548
FEMALE	-2.2734	0.2790	-8.1470	0.0000
EDUC	0.5065	0.0504	10.0505	0.0000
R-squared	0.2588	Mean dependent var		5.8961

Figure: 1

- Based on the output in Figure 1 we can conclude the following:

- Given that the p-value for testing the individual significance of the regressor female is zero, we would reject the

$$H_0 : \delta_0 = 0$$

in either a one-sided or two-sided test, at any significance level, and conclude that, controlling for education, gender does affect average wages.

- Since

$$\widehat{\delta}_0 = -2.27,$$

we estimate (or predict) that the average hourly wage of females is \$2.27 less than the average hourly wage of males with the same level of education.

- The introduction of an intercept dummy variable allows the mean of the dependent variable to vary by category.
- For example, as we have seen in the linear regression equation

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, i = 1, 2, ...526, \quad (4)$$

the intercept dummy variable *female* allows the average weekly wage of males and females to differ.

- However, intercept dummies don't allow the marginal effects of the explanatory variables to vary by category.

- For example, in the wage equation

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, \tag{4}$$

since

$$\frac{\partial wage_i}{\partial educ_i} = \beta_1$$

for both males and females, the marginal effect of education on wages is the same for males and females.

- We can allow for the possibility that the marginal effect of a regressor varies by category by including a **slope dummy variable** (also called an **interaction dummy**) in the regression equation.
- For example, suppose we wish to allow for the possibility that the marginal effect of education on wages varies by gender.

- We can do this by specifying the equation for the conditional mean of wages as

$$E(wage_i|female_i, educ_i) = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i. \quad (5)$$

- The regressor $female_i * educ_i$ is called a slope or interaction dummy variable.
- Equation (5) implies that

$$\begin{aligned} E(wage_i|female_i &= 1, educ_i) = \beta_0 + \delta_0 + \beta_1 educ_i + \delta_1 educ_i, \\ &= \beta_0 + \delta_0 + (\beta_1 + \delta_1) educ_i \end{aligned}$$

and

$$E(wage_i|female_i = 0, educ_i) = \beta_0 + \beta_1 educ_i.$$

- Therefore,

$$\begin{aligned} E(wage_i|female_i &= 1, educ_i) - E(wage_i|female_i = 0, educ_i,) \\ &= \delta_0 + \delta_1 educ_i. \end{aligned}$$

- For example,

$$\begin{aligned} E(wage_i|female_i &= 1,\ educ_i{=}10) - E(wage_i|female_i = 0, educ_i = 10 \\ &= \delta_0 + 10\delta_1. \end{aligned}$$

- The specification

$$E(wage_i|female_i, educ_i) = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i \quad (5)$$

also allows the marginal effect of education to vary by gender.

- To see this, note that, from (5),

$$\begin{aligned}\frac{\partial E(wage_i|female_i, educ_i)}{\partial educ_i} &= \frac{\partial(\beta_0 + \beta_1 educ_i + \delta_1 female_i * educ_i)}{\partial educ_i} \\ &= \frac{\beta_1 \partial(educ_i)}{\partial educ_i} + \frac{\delta_1 \partial female_i * educ_i}{\partial educ_i} \\ &= \beta_1 + \delta_1 female_i. \qquad (6)\end{aligned}$$

- Equation (6) implies that for females the marginal effect of education on the average wage is

$$\frac{\partial E(wage_i|female_i = 1, educ_i)}{\partial educ_i} = \beta_1 + \delta_1,$$

and for males the marginal effect of education on the average wage is

$$\frac{\partial E(wage_i|female = 0, educ_i)}{\partial educ_i} = \beta_1.$$

- That is, the coefficient δ_1 in

$$E(wage_i|female_i, educ_i) = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i \quad (5)$$

measures the difference between the marginal effect of education on the average wage of females and the marginal effect of education on the average wages of males.

- Since

$$\begin{aligned} E(wage_i|female_i &= 1, educ_i) - E(wage_i|female_i = 0, educ_i,) \\ &= \delta_0 + \delta_1 educ_i, \end{aligned}$$

we can test the null hypothesis that gender has no effect on average wages by estimating the unrestricted model

$$wage_i = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i + u_i \quad (7)$$

and performing an F test of

$$\begin{aligned} H_0 &: \delta_0 = \delta_1 = 0, \\ H_1 &: \delta_0 \text{ and/or } \delta_1 \neq 0. \end{aligned}$$

- The restricted model for the test is

$$wage_i = \beta_0 + \beta_1 educ_i + u_i. \tag{8}$$

- When we estimate (7) and (8) we obtain the output reported in Figure 2 and Figure 3 below.

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.2005	0.8436	0.2377	0.8122
FEMALE	-1.1985	1.3250	-0.9045	0.3661
EDUC	0.5395	0.0642	8.4001	0.0000
FEMALE*EDUC	-0.0860	0.1036	-0.8298	0.4070

R-squared	0.2598	Mean dependent var	5.8961
Sum squared resid	5300.1699	Schwarz criterion	5.1957

Figure: 2

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.9049	0.6850	-1.3210	0.1871
EDUC	0.5414	0.0532	10.1667	0.0000

R-squared	0.1648	Mean dependent var	5.8961
Sum squared resid	5980.6823	Schwarz criterion	5.2927

Figure: 3

- We test the null hypothesis that gender has no effect wages once we control for education by performing the following F test:

$$
\begin{aligned}
H_0 &: \delta_0 = \delta_1 = 0 \\
H_1 &: \delta_0 \text{ and/or } \delta_1 \neq 0 \\
\text{Significance level} &: \alpha = 0.05 \\
\text{Test stat and null distribution} &: \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \frac{(n-k-1)}{q} \\
&= \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \frac{(526-4)}{2} \sim F(2, 522) \\
F_{calc} &= \frac{(5980.6823 - 5300.1699)}{5300.1699} \frac{522}{2} \\
&= 33.51 \\
F_{crit} &= 3.07 \\
\text{Decision rule} &: \textit{reject } H_0 \text{ if } F_{calc} > Ft_{crit} \\
\text{Decision} &: \textit{Since } 33.51 > 3.07, \text{ we reject } H_0.
\end{aligned}
$$

- That is, we reject the null hypothesis that gender has no effect on wages once we control for education, in favor of the alternative hypothesis that it does have an effect.
- Since

$$\begin{aligned} \frac{\partial E(wage_i|female_i = 1, educ_i)}{\partial educ_i} &= \beta_1 + \delta_1, \\ \frac{\partial E(wage_i|female = 0, educ_i)}{\partial educ_i} &= \beta_1, \end{aligned}$$

we can test the null hypothesis that the marginal effect of education on the average wage is the same for males and females by estimating

$$wage_i = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i + u_i \qquad (7)$$

the and testing

$$\begin{aligned} H_0 &: \delta_1 = 0 \\ H_1 &: \delta_1 \neq 0. \end{aligned}$$

- The form of the alternative hypothesis reflects the fact that we have no strong prior belief that an additional year of education is more beneficial for one gender than it is for the other.
- From

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.2005	0.8436	0.2377	0.8122
FEMALE	-1.1985	1.3250	-0.9045	0.3661
EDUC	0.5395	0.0642	8.4001	0.0000
FEMALE*EDUC	-0.0860	0.1036	-0.8298	0.4070

R-squared	0.2598	Mean dependent var	5.8961
Sum squared resid	5300.1699	Schwarz criterion	5.1957

Figure: 2

we note that the p-value associated with the regressor $female_i * educ_i$ is

$$0.4070 > 0.05,$$

we fail to reject the null hypothesis that the marginal effect of education on wages is the same for males and females.

2 Perfect Multicollinearity I

- **Perfect multicollinearity** arises when there is an **exact** linear relationship between some or all of the regressors in the multiple linear regression model.
- When this occurs in the multiple linear regression model

$$y = X\beta + u,$$

the OLS estimator of β,

$$\widehat{\beta} = (X'X)^{-1}X'y,$$

cannot be computed because the matrix $X'X$ is singular and therefore the matrix $(X'X)^{-1}$ is not defined.

- If a subset of the regressors in a linear regression are perfectly collinear, Eviews will return an error message if you attempt to estimate the regression equation.

2 Perfect Multicollinearity II

- Perfect multicollinearity is a regression equation is highly unusual and it normally occurs because the researcher has made an elementary mistake in the way they have attempted to incorporate dummy variables into the regression model.
- To illustrate the problem, let's return to the wage equation

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, \tag{9}$$

  which we specified earlier.
- Suppose that we decide to extend (9) to include the dummy variable *male*, which we defined above.
- Our extended regression equation is

$$wage_i = \beta_0 + \delta_0 female_i + \gamma_0 male_i + \beta_1 educ_i + u_i. \tag{10}$$

2 Perfect Multicollinearity III

- For simplicity, assume that the sample size is 5 and that the first 3 individuals in the sample are female, each with 12 years of education, and the final 2 individuals are male, each with 10 years of education.
- In this simple example the X matrix is given by

$$\underset{(5\times 4)}{X} = \begin{bmatrix} 1 & 1 & 0 & 12 \\ 1 & 1 & 0 & 12 \\ 1 & 1 & 0 & 12 \\ 1 & 0 & 1 & 10 \\ 1 & 0 & 1 & 10 \end{bmatrix}.$$

2 Perfect Multicollinearity IV

- Let c_j denote the jth column of X. Note that

$$c_2 + c_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = c_1 .$$

Since

$$c_1 = c_2 + c_3$$

we have an exact linear relationship between the columns of X and the OLS estimator cannot be computed.

2 Perfect Multicollinearity V

- Note that, even if we exclude the male dummy variable, perfect multicollinearity would also arise in this example if, by chance, all five individuals in the sample had 12 years of education, since then the X matrix would be given by

$$\underset{(5x3)}{X} = \begin{bmatrix} 1 & 1 & 12 \\ 1 & 1 & 12 \\ 1 & 1 & 12 \\ 1 & 0 & 12 \\ 1 & 0 & 12 \end{bmatrix}$$

and

$$c_3 = 12c_1 .$$

That is, there is an exact linear relationship between columns 3 and 1 of the X matrix.

2 Perfect Multicollinearity VI

- Perfect multicollinearity is present in the regression equation

$$wage_i = \beta_0 + \delta_0 female_i + \gamma_0 male_i + \beta_1 educ_i + u_i \tag{10}$$

because of the way in which it has been specified.

- There are two categories for the regressor gender, male and female, and we have included a dummy variable for each category, together with an intercept.
- Including an intercept in the regression equation together with a dummy variable for every category always induces perfect multicollinearity.
- This fundamental error is called **the dummy variable trap**.
- There are two ways to avoid the dummy variable trap when specifying the wage equation.

1. Omit **one** of the gender dummies from the regression equation.

2 Perfect Multicollinearity VII

(a) If we omit the dummy variable *male* from the regression equation and specify the model as

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, \tag{11}$$

then

$$E(wage_i|educ_i, female_i = 1) - E(wage_i|educ_i, female_i = 0) = \delta_0.$$

In this case the coefficient δ_0 in (11) measures the difference between the average wage of females and males, with the same level of education.

2 Perfect Multicollinearity VIII

(b) If we omit the dummy variable *female* from the regression equation and specify the model as

$$wage_i = \beta_0 + \gamma_0 male_i + \beta_1 educ_i + u_i, \tag{12}$$

then

$$E(wage_i | educ_i, male_i = 1) - E(wage_i | educ_i, male_i = 0) = \gamma_0.$$

In this case the coefficient γ_0 in (12) measures the difference between the average wage of females and males, with the same level of education.

2. Retain both gender dummies and omit the intercept.
   - In this case we specify the regression equation as

$$wage_i = \alpha_0 female_i + \lambda_0 male_i + \beta_1 educ_i + u_i. \tag{13}$$

2 Perfect Multicollinearity IX

- Equation (13) implies that

$$\begin{aligned} E(wage_i|female_i &= 1, male_i = 0, educ_i) = \alpha_0 + \beta_1 educ_i, \\ E(wage_i|female_i &= 0, male_i = 1, educ_i) = \lambda_0 + \beta_1 educ_i. \end{aligned}$$

- Therefore,

$$\begin{aligned} E(wage_i|female_i &= 1, male_i = 0, educ_i) \\ -E(wage_i|female_i &= 0, male_i = 1, educ_i) \\ &= (\alpha_0 + \beta_1 educ_i) - (\lambda_0 + \beta_1 educ_i) \\ &= \alpha_0 - \lambda_0. \end{aligned} \tag{14}$$

2 Perfect Multicollinearity X

- It follows from (14) that when we omit the intercept and specify the regression equation as

$$wage_i = \alpha_0 female_i + \lambda_0 male_i + \beta_1 educ_i + u_i, \tag{13}$$

the difference between the average wage of females and males, with the same level of education, is

$$\alpha_0 - \lambda_0.$$

The estimated difference between the average wage of females and males, with the same level of education, is

$$\widehat{\alpha}_0 - \widehat{\lambda}_0,$$

where $\widehat{\alpha}_0$ and $\widehat{\lambda}_0$ are the estimates of α_0 and λ_0 we obtain when we estimate (13) by OLS.

- When we choose option 1 and omit the dummy variable for one of the categories, the category we omit is called the **base category**.

2 Perfect Multicollinearity XI

- For example, when we choose option 1 (a) and specify the model as

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, \tag{11}$$

the omitted category male is the base category.

- In general, the coefficient attached to the dummy variable for the included category always measures the difference between the conditional mean of the dependent variable for the included category and the conditional mean of the dependent variable for the base category.
- For example, in the regression equation

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i, \tag{11}$$

the coefficient δ_0 measures the difference between the average wages of females and males (the base category) with the same level of education.

3 Categorical Variables with many Categories I

- We next generalize the discussion of categorical variables to the case in which we have a categorical variable with more than two categories.
- For example, for most populations the categorical variable, race, will have more than two categories.
- Suppose that we wish to determine whether or not, controlling for education, the average wage varies by race in the United States of America (USA).
- For simplicity, assume that everyone in the USA belongs to one and only one of the racial categories white, black, Asian and Hispanic.

3 Categorical Variables with many Categories II

- We define the following dummy variables:

$$white_i = \left\{ \begin{array}{c} 1 \text{ if individual i is white} \\ 0 \quad \text{otherwise} \end{array} \right| , i = 1, 2, ..., n,$$

$$black_i = \left\{ \begin{array}{c} 1 \text{ if individual i is black} \\ 0 \quad \text{otherwise} \end{array} \right| , i = 1, 2, ..., n,$$

$$asian_i = \left\{ \begin{array}{c} 1 \text{ if individual i is Asian} \\ 0 \quad \text{otherwise} \end{array} \right| , i = 1, 2, ..., n,$$

$$hiss_i = \left\{ \begin{array}{c} 1 \text{ if individual i is Hispanic} \\ 0 \quad \text{otherwise} \end{array} \right| , i = 1, 2, ..., n,$$

- Note that the racial categories defined above are **exhaustive and mutually exclusive**. That is, every member of the population belongs to one and only one category.

3 Categorical Variables with many Categories III

- In order to avoid the dummy variable trap we must either omit the intercept or one of the race dummies when we specify the regression equation.
- If we omit the dummy variable *white* and specify the regression model as

$$wage_i = \beta_0 + \beta_1 black_i + \beta_2 asian_i + \beta_3 hiss_i + \beta_4 educ_i + u_i, \quad (15)$$

then *white* is the base category. In this case:

  - β_1 measures the difference between the average wage of blacks and the average wage of whites, with the same level of education.
  - β_2 measures the difference between the average wage of Asians and the average wage of whites, with the same level of education.
  - β_3 measures the difference between the average wage of Hispanics and the average wage of whites, with the same level of education.

3 Categorical Variables with many Categories IV

- Since

$$E(wage_i|black_i = 1, asian_i = hiss_i = 0, educ_i) = \beta_0 + \beta_1 + \beta_4 educ_i,$$

and

$$E(wage_i|asian_i = 1, black_i = hiss_i = 0, educ_i) = \beta_0 + \beta_2 + \beta_4 educ_i,$$

it follows that

$$\begin{aligned} E(wage_i|black_i &= 1, asian_i = hiss_i = 0, educ_i) \\ -E(wage_i|asian_i &= 1, black_i = hiss_i = 0, educ_i) \\ &= \beta_0 + \beta_1 + \beta_4 educ_i - (\beta_0 + \beta_2 + \beta_4 educ_i,) \\ &= \beta_1 - \beta_2. \end{aligned}$$

That is,

$$\beta_1 - \beta_2$$

measures the difference between the average wage of blacks and the average wage of Asians, with the same level of education.

3 Categorical Variables with many Categories V

- In summary:
  - The coefficient attached to a particular race dummy in

$$wage_i = \beta_0 + \beta_1 black_i + \beta_2 asian_i + \beta_3 hiss_i + \beta_4 educ_i + u_i, \quad (15)$$

    measures the difference between the average wage of the members of that racial group and the average wage of whites (the base group), with the same level of education.
  - The difference between the coefficients attached to any two racial dummies measures the difference between the average wage of those two racial groups, controlling for education.
  - When we estimate (15), $\widehat{\beta}_1$ is the **estimated** difference between the average wage of blacks and the average wage of whites (the base category), with the same level of education, and

$$\widehat{\beta}_1 - \widehat{\beta}_2$$

    is the **estimated** difference between the average wage of blacks and the average wage of Asians, with the same level of education.

4 Near Multicollinearity I

- **Near multicollinearity** (also called **imperfect multicollinearity**) arises when there is not an exact linear relationship between any subset of the regressors, but two or more of the regressors are highly correlated **in the sample at hand**.
- The presence of near multicollinearity (NM) does not violate any of the assumptions of the classical linear regression model, so the OLS estimator can still be computed and has the usual properties under assumptions A1 to A5.
- However, NM does pose serious problems for both estimation and hypothesis testing in the linear regression model.
- Proving the problems posed by NM is beyond the scope of this unit. We restrict ourselves to stating them without proof.
- **Problems for Estimation**: NM causes the standard errors of the $\widehat{\beta}s$ to be quite large, implying:

4 Near Multicollinearity II

- Our estimates of the true βs are not very precise.
- Our confidence intervals, which depend on the standard errors of the $\widehat{\beta} s$, will be wide.

- **Problems for hypothesis testing**: NM increases the probability that the sample value of our t statistic for testing the individual significance of a regressor will fall in the non-rejection region, even **when the null hypothesis is false**, thereby increasing the probability that we will fail to reject a false null (that is, increasing the probability of committing a type-2 error).
- Recall that the t statistic for testing

$$H_0 : \beta_j = 0$$

is

$$\frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \sim t(n-k-1).$$

4 Near Multicollinearity III

We reject the null if

$$t_{calc} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

falls in the rejection region. (See Figure 4 below).

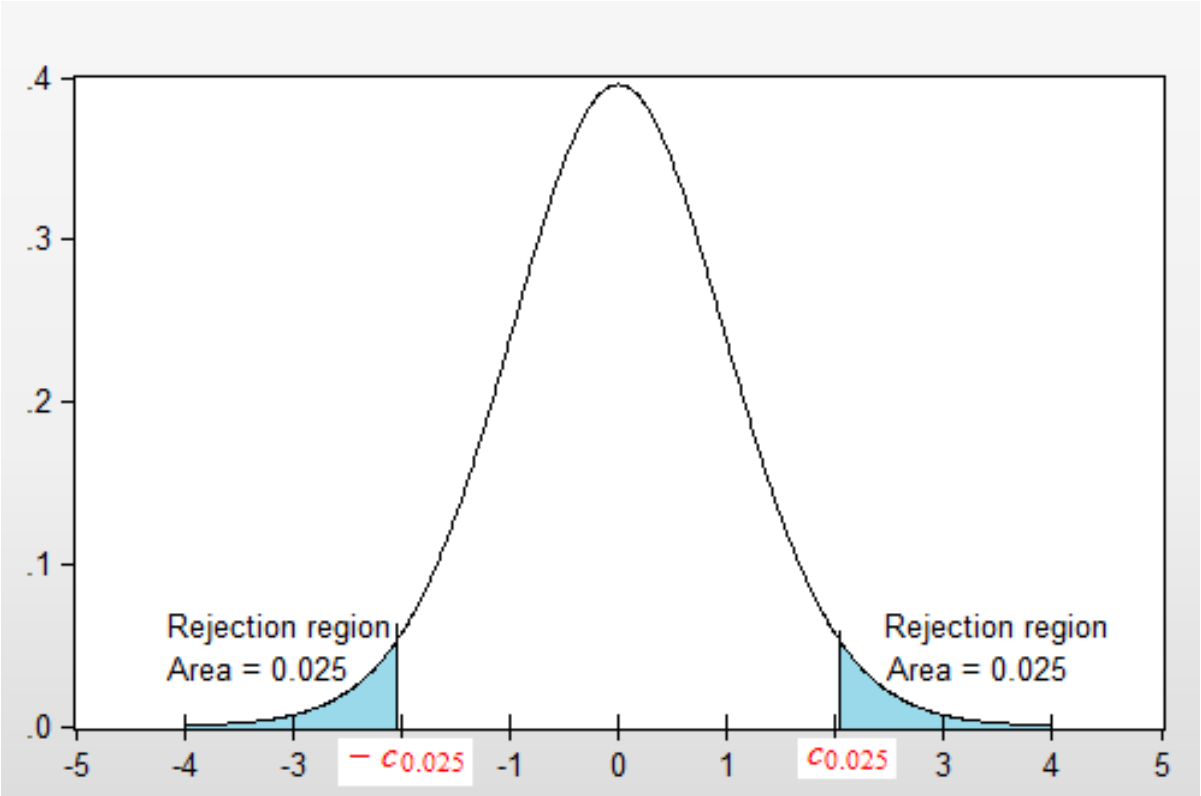

Figure 4

4 Near Multicollinearity V

- In the presence of NM, the $se(\widehat{\beta}_j)$ tends to be large, with the result that

$$t_{calc} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

  tends to be small in absolute value and is therefore less likely to fall in the rejection region (see Figure 4).
- Fortunately, although our t tests of individual significance are unreliable in the presence of NM, the F test for testing the joint significance of a subset of the regressors is not affected by NM.
- Recall that when we estimated

$$wage_i = \beta_0 + \delta_0 female + \beta_1 educ_i + \delta_1 female_i * educ_i + u_i \qquad (7)$$

  by OLS we obtained the output reported in Figure 5 below.

4 Near Multicollinearity VI

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.2005	0.8436	0.2377	0.8122
FEMALE	-1.1985	1.3250	-0.9045	0.3661
EDUC	0.5395	0.0642	8.4001	0.0000
FEMALE*EDUC	-0.0860	0.1036	-0.8298	0.4070

R-squared	0.2598	Mean dependent var	5.8961
Sum squared resid	5300.1699	Schwarz criterion	5.1957

5

4 Near Multicollinearity VII

- Notice that, since the relevant p-values are greater than 0.1, the regressors *female* and *female* $*$ *educ* are both individually insignificant at the 5% significance level in either a one-sided or a two-sided test.
- However, when we performed an F test of the null hypothesis

$$H_0 : \delta_0 = \delta_1 = 0,$$

  we rejected the null hypothesis and concluded that the regressors *female* and *female* $*$ *educ* are jointly significant.
- When a regression equation contains a subset of regressors which appear to be individually insignificant when we perform t tests of their individual significance, but appear to be jointly significant when we perform an F test of their joint significance, this may be an indication of NM.
- Unfortunately, there is very little that we can do to combat the effect of NM on estimation and hypothesis testing.

5 Models with a Binary Dependent Variable I

- In the preceding sections we have discussed the case in which one of more of the regressors is a binary variable.
- Below we briefly discuss the case in which the **dependent variable** is a binary variable.
- In some empirical studies we may be interested in modelling the probability of a particular outcome occurring. For example, we may interested in modelling:
  - the probability of an individual having private health insurance;
  - the probability of an individual voting for a particular political party;
  - the probability of an individual having their bank loan application approved;
  - the probability of an individual having an affair;

5 Models with a Binary Dependent Variable II

- Suppose that y_i is a binary random variable with the property that.

$$y_i = \left\{ \begin{array}{l} 1 \text{ with probability } p_i \\ 0 \text{ with probability } 1\text{-}p_i \end{array} \right| , i = 1, 2, ..., n. \tag{16}$$

- The outcome

$$y_i = 1$$

denotes some outcome of interest, such as individual i having private health insurance, voting for a particular political party, or having an affair.

- For concreteness, assume that

$$y_i = 1$$

denotes the outcome that individual i has private health insurance.

5 Models with a Binary Dependent Variable III

- It follows from

$$y_i = \left\{ \begin{array}{l} 1 \text{ with probability } p_i \\ 0 \text{ with probability 1-}p_i \end{array} \right| , i = 1, 2, ..., n. \quad (16)$$

that

$$E(y_i) = 1xp_i + 0x(1 - p_i) = p_i. \quad (17)$$

- That is, the mean of y_i is the probability that y_i takes on the value 1, which in our example, is the probability that individual i has private health insurance.
- The random variable p_i is called a **response probability**.
- In order to study the effect of various explanatory variables on p_i, we assume we need to specify a mathematical model for p_i.
- The simplest assumption to make is that

$$p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, i = 1, 2, ..., n. \quad (18)$$

5 Models with a Binary Dependent Variable IV

- Equation (18) specifies that p_i is a linear function of both the random variables $(x_{i1}, x_{i2}, ..., x_{ik})$ and the parameters $(\beta_0, \beta_1, ..., \beta_k)$.
- For example, in the private health insurance example, we might specify the following model for the response probability

$$p_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 age_i + \beta_3 income_i.$$

- Combining the fact that

$$E(y_i) = p_i \tag{17}$$

and (18) we obtain

$$\begin{aligned} E(y_i|x_{i1}, x_{i2}, ..., x_{ik}) &= p_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, i = 1, 2, ..., n \end{aligned} \tag{19}$$

- The associated linear regression model is

$$\begin{aligned} y_i &= E(y_i|x_{i1}, x_{i2}, ..., x_{ik}) + u_i \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, + u_i, i = 1, 2, ..., n. \quad (20) \end{aligned}$$

- It follows from (19) that the estimated or predicted response probability is given by

$$\begin{aligned} \widehat{p}_i &= \widehat{E}(y_i|x_{i1}, x_{i2}, ..., x_{ik}) \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + ... + \widehat{\beta}_k x_{ik} \\ &= \widehat{y}_i. \end{aligned} \quad (21)$$

5 Models with a Binary Dependent Variable VI

- Equation (21) states that the **estimated probability** of individual i having private health insurance is the predicted value of y_i we get when we estimate the linear regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, + u_i, i = 1, 2, ..., n,$$

where

$$y_i = \left\{ \begin{array}{l} 1 \text{ with probability p}_i \\ 0 \text{ with probability 1-p}_i \end{array} \right| , i = 1, 2, ..., n.$$

- For example, in the private health insurance example,

$$y_i = \left\{ \begin{array}{l} 1 \text{ if individual i has private health insurance} \\ 0 \text{ otherwise} \end{array} \right| , i = 1, 2, ..., n,$$

and we assume that

$$p_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 age_i + \beta_3 income_i, i = 1, 2, ..., n, \quad (22)$$

in which case the associated linear regression equation is

$$y_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 age_i + \beta_3 income_i + u_i. \quad (23)$$

- In (22), β_1 measures the marginal effect of education **on the probability** of an individual having private health insurance, controlling for gender, age and income.
- When we estimate (23) by OLS, $\widehat{\beta}_1$ is the estimated marginal effect of education on the probability of an individual having private health insurance, controlling for gender, age and income.

5 Models with a Binary Dependent Variable VIII

- It follows from

$$p_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 age_i + \beta_3 income_i, i = 1, 2, ..., n, \tag{22}$$

that the probability of a 40 year old female with 10 years of education and an income of \$50,000 having private health insurance is

$$p_i|(female_i = 1, educ_i = 10, age_i = 40, income_i = 50) = \beta_0 + \delta_0 + 10\beta_1 \tag{24}$$

and

$$\widehat{p}_i|(female_i = 1, educ_i = 10, age_i = 40, income_i = 50) = \widehat{\beta}_0 + \widehat{\delta}_0 + 10\widehat{\beta}_1 \tag{25}$$

- We obtain $(\widehat{\beta}_0, \widehat{\delta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)$ when we estimate

$$y_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 age_i + \beta_3 income_i + u_i. \tag{23}$$

5 Models with a Binary Dependent Variable IX

- The linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, + u_i, i = 1, 2, ..., n,$$

where

$$y_i = \left\{ \begin{array}{l} 1 \text{ with probability } p_i \\ 0 \text{ with probability 1-}p_i \end{array} \right| , i = 1, 2, ..., n,$$

is called the **linear probability model.**

- The name comes from the fact that the model assumes that

$$p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, i = 1, 2, ..., n. \qquad (18)$$

- It turns out that a linear probability model is not the best model to use in applications where

$$y_i = \left\{ \begin{array}{l} 1 \text{ with probability } p_i \\ 0 \text{ with probability 1-}p_i \end{array} \right| , i = 1, 2, ..., n.$$

- The reason is that the specification in (18) does not guarantee that

$$0 \leq p_i \leq 1, i = 1, 2, ..., n.$$

- There are superior alternative models available, called **logit** and **probit** models, for modeling binary dependent variables.
- The logit model replaces the assumption that

$$p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, i = 1, 2, ..., n, \qquad (18)$$

with the assumption that

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})}}.$$

The associated regression equation is

$$y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})}} + u_i, i = 1, 2, ..., n \qquad (26)$$

5 Models with a Binary Dependent Variable XI

- The regression model specified in (26) is highly nonlinear in the parameters and cannot be estimated by OLS.
- The regression model in (26) is usually estimated by an alternative estimation procedure called **maximum likelihood estimation.**
- For more information, the interested student is directed to ETC3410.