

Economics 326
Methods of Empirical Research in Economics

Lecture 15: Dummy variables

Vadim Marmer
University of British Columbia

March 24, 2011

Interval, Ordinal, and Categorical Variables

- ▶ Interval variable: is one where the difference between two values is *meaningful*. Example: “Education” when measured in years. There is a meaning to the difference between 12 and 10 years of education.
- ▶ In some data sets, education is reported as an ordinal variable: only the order between its values matters, but the difference has no meaning. Example: The following two variables are equivalent.

$$Education_i = \begin{cases} 1 & \text{if high-school graduate,} \\ 2 & \text{if college graduate,} \\ 3 & \text{if advanced degree.} \end{cases}$$
$$Education_i = \begin{cases} 1 & \text{if high-school graduate,} \\ 10 & \text{if college graduate,} \\ 234 & \text{if advanced degree.} \end{cases}$$

Interval, Ordinal, and Categorical Variables

- ▶ Categorical variable is one that has one or more categories, but there is no natural ordering to the categories
Examples: Gender, race, marital status, geographic location.

- ▶ The following two variables are equivalent:

$$Gender_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a } \textit{woman}, \\ 2 & \text{if observation } i \text{ corresponds to a } \textit{man}. \end{cases}$$

$$Gender_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a } \textit{man}, \\ 2 & \text{if observation } i \text{ corresponds to a } \textit{woman}. \end{cases}$$

- ▶ Categorical and ordinal variables are also called qualitative.
- ▶ Qualitative variables cannot be simply included in regression, because the regression technique assumes that all variables are interval.

Dummy variables

- ▶ A dummy variable is a binary zero-one variable which takes on the value one if some condition is satisfied and zero if that condition fails:

- ▶ $Female_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a woman,} \\ 0 & \text{if observation } i \text{ corresponds to a man.} \end{cases}$

- ▶ $Male_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a man,} \\ 0 & \text{if observation } i \text{ corresponds to a woman.} \end{cases}$

- ▶ Note that $Female_i + Male_i = 1$ for all observations i .

- ▶ $Married_i = \begin{cases} 1 & \text{if married,} \\ 0 & \text{otherwise.} \end{cases}$

Example

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

A single dummy independent variable

- ▶ Consider the following regression:

$$Wage_i = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i,$$

and assume that conditionally on all independent variables,
 $E(U_i) = 0$.

- ▶ If observation i corresponds to a woman, $Female_i = 1$, and

$$\begin{aligned} E(Wage_i | Female_i = 1, Educ_i, Exper_i, Tenure_i) &= \\ &= \beta_0 + \delta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i. \end{aligned}$$

- ▶ If observation i corresponds to a man, $Female_i = 0$, and

$$\begin{aligned} E(Wage_i | Female_i = 0, Educ_i, Exper_i, Tenure_i) &= \\ &= \beta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i. \end{aligned}$$

- ▶ Thus,

$$\begin{aligned} \delta_0 &= E(Wage_i | Female_i = 1, Educ_i, Exper_i, Tenure_i) - \\ &\quad - E(Wage_i | Female_i = 0, Educ_i, Exper_i, Tenure_i). \end{aligned}$$

An intercept shift

- ▶ The model:

$$Wage_i = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$$

- ▶ For men ($Female_i = 0$):, we can write the model as

$$Wage_i^M = \beta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ For women ($Female_i = 1$):, we can write the model as

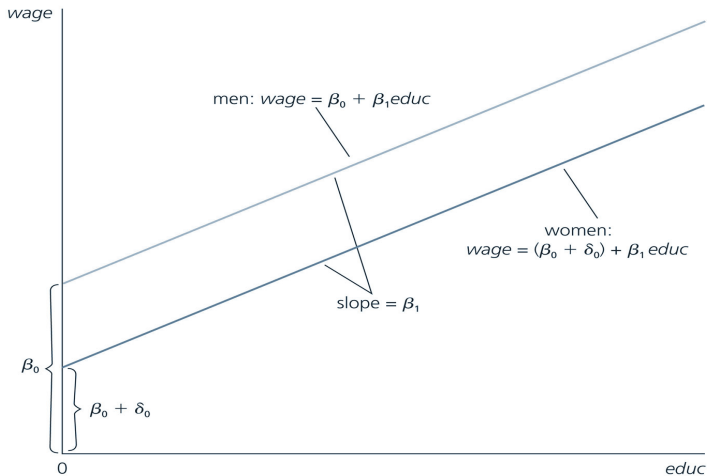
$$Wage_i^F = (\beta_0 + \delta_0) + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ In this case, men play the role of the base group.
- ▶ δ_0 measures the difference relatively to the base group.

An intercept shift

FIGURE 7.1

Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



Example

- ▶ Estimated equation:

$$\widehat{Wage}_i = - \begin{matrix} 1.57 \\ (0.72) \end{matrix} - \begin{matrix} 1.81 \\ (0.26) \end{matrix} Female_i + \begin{matrix} 0.572 \\ (0.049) \end{matrix} Educ_i + \\ + \begin{matrix} 0.025 \\ (0.012) \end{matrix} Exper_i + \begin{matrix} 0.141 \\ (0.021) \end{matrix} Tenure_i.$$

- ▶ The dependent variable is the wage per hour.
- ▶ $\hat{\delta}_0 = -1.81$ implies that a women earns \$1.81 less per hour than a man with the same level of education, experience, and tenure. (These are 1976 wages.)
- ▶ The difference is also statistically significant.

When the dependent variable is in the logarithmic form

- ▶ The model:

$$\ln(\text{Wage}) = \beta_0 + \delta_0 \text{Female} + \beta_1 \text{Educ} + \beta_3 \text{Exper} + \beta_4 \text{Tenure} + U.$$

- ▶ In this case,

$$\begin{aligned}\delta_0 &= \ln(\text{Wage}^F) - \ln(\text{Wage}^M) \\ &= \ln\left(\frac{\text{Wage}^F}{\text{Wage}^M}\right) \\ &= \ln\left(\frac{\text{Wage}^M + (\text{Wage}^F - \text{Wage}^M)}{\text{Wage}^M}\right) \\ &= \ln\left(1 + \frac{\text{Wage}^F - \text{Wage}^M}{\text{Wage}^M}\right) \\ &\approx \frac{\text{Wage}^F - \text{Wage}^M}{\text{Wage}^M}.\end{aligned}$$

- ▶ When the dependent variable is in the log form, δ_0 has a percentage interpretation.

Example

- ▶ Estimated equation:

$$\begin{aligned} \ln(\widehat{Wage}_i) = & 0.417 & - & 0.297 & Female_i & + & 0.080 & Educ_i & + \\ & (0.099) & & (0.036) & & & (0.007) & & \\ & + & 0.029 & Exper_i & - & 0.00058 & Exper_i^2 & + \\ & & (0.005) & & & (0.00010) & & \\ & + & 0.032 & Tenure_i & - & 0.00059 & Tenure_i^2. \\ & & (0.007) & & & (0.00023) & & \end{aligned}$$

- ▶ $\hat{\delta}_0 = -0.297$ implies that a woman earns 29.7% less than a man with the same level of education, experience and tenure.

Changing the base group

- ▶ Instead of

$$\ln(Wage_i) = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$$

consider:

$$\ln(Wage_i) = \theta_0 + \gamma_0 Male_i + \theta_1 Educ_i + \theta_3 Exper_i + \theta_4 Tenure_i + U_i.$$

- ▶ Since $Male_i = 1 - Female_i$,

$$\begin{aligned}\ln(Wage_i) &= \theta_0 + \gamma_0 Male_i + \theta_1 Educ_i + \theta_3 Exper_i + \theta_4 Tenure_i + U_i \\ &= \theta_0 + \gamma_0 (1 - Female_i) + \theta_1 Educ_i + \theta_3 Exper_i + \theta_4 Tenure_i + U_i \\ &= (\theta_0 + \gamma_0) - \gamma_0 Female_i + \theta_1 Educ_i + \theta_3 Exper_i + \theta_4 Tenure_i + U_i.\end{aligned}$$

- ▶ We conclude that $\delta_0 = -\gamma_0$, $\beta_0 = \theta_0 - \delta_0$, $\beta_1 = \theta_1$, and etc.:

$$\ln(Wage_i) = (\beta_0 + \delta_0) - \delta_0 Male_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ Thus, changing the base group has no effect on the conclusions.

The dummy variable trap

- ▶ Consider the equation:

$$\ln(Wage_i) = \beta_0 + \delta_0 Female_i + \gamma_0 Male_i + \\ + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ Recall that the intercept is a regressor that takes the value one for all observations.
- ▶ Since $Female_i + Male_i - 1 = 0$ for all observations i , we have the case of perfect multicollinearity, and such an equation cannot be estimated.
- ▶ One cannot include an intercept and dummies for all the groups!

The dummy variable trap

- ▶ One of the dummies has to be omitted and the corresponding group becomes the base group:
 - ▶ Men are the base group: $\ln(Wage_i) = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.
 - ▶ Women are the base group: $\ln(Wage_i) = \theta_0 + \gamma_0 Male_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.
- ▶ Alternatively, one can include both dummies without the intercept: $\ln(Wage_i) = \pi_0 Female_i + \pi_1 Male_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.
 - ▶ In Stata regression with no intercept can be estimated by using the option "no constant":
`regress Y X, noconstant`
 - ▶ The coefficients on the dummy variables lose the difference interpretation.

A slope shift and interactions

- ▶ We can also allow the returns to education to be different for men and women:

$$\ln(Wage_i) = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \delta_1 (Female_i \cdot Educ_i) + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ The variable $(Female_i \cdot Educ_i)$ is called an interaction.
- ▶ The equation for men ($Female_i = 0$):

$$\ln(Wage_i^M) = \beta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ The equation for women ($Female_i = 1$):

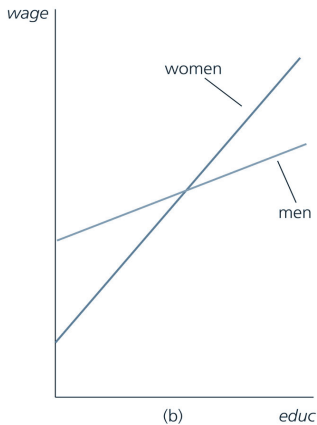
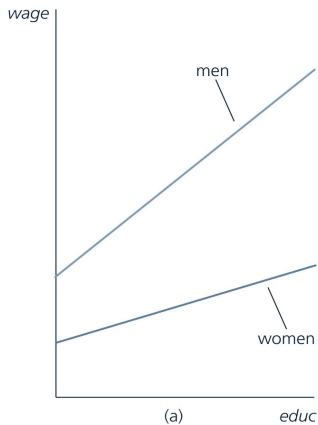
$$\ln(Wage_i^F) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- ▶ δ_1 can be interpreted as the difference in return to education between the women and men (the base group) after controlling for experience and tenure.

A slope shift

FIGURE 7.2

Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



Example

- ▶ Estimated equation:

$$\begin{aligned} \ln(\widehat{Wage}_i) = & 0.389 - 0.227 \text{ Female}_i + \\ & (0.119) \quad (0.168) \\ & + 0.082 \text{ Educ}_i - 0.0056 \text{ Female}_i \cdot \text{Educ}_i \\ & (0.008) \quad (0.0131) \\ & + 0.029 \text{ Exper}_i - 0.00058 \text{ Exper}_i^2 + \\ & (0.005) \quad (0.00011) \\ & + 0.032 \text{ Tenure}_i - 0.00059 \text{ Tenure}_i^2. \\ & (0.007) \quad (0.00024) \end{aligned}$$

- ▶ $\hat{\delta}_1 = -0.0056$ suggesting that the return to education for women is 0.56% less than for men, however it is not statistically significant. Thus, we can conclude that the return to education is the same for men and women.

Multiple categories

- ▶ In the previous examples, *Educ* was a quantitative variable: years of education.
- ▶ Suppose now that instead the education variable is ordinal:

$$Education = \begin{cases} 1 & \text{if high-school dropout,} \\ 2 & \text{if high-school graduate,} \\ 3 & \text{if some college,} \\ 4 & \text{if college graduate,} \\ 5 & \text{if advanced degree.} \end{cases}$$

- ▶ Only the order is important, and there is no meaning to the distance between the values.
- ▶ Adding such a variable to the regression will give a meaningless result.

Multiple categories

$$Education_i = \begin{cases} 1 & \text{if high-school dropout,} \\ 2 & \text{if high-school graduate,} \\ 3 & \text{if some college,} \\ 4 & \text{if college graduate,} \\ 5 & \text{if advanced degree.} \end{cases}$$

- ▶ Define 5 new dummy variables:

$$E_{1,i} = \begin{cases} 1 & \text{if high-school dropout,} \\ 0 & \text{otherwise.} \end{cases}, \quad E_{2,i} = \begin{cases} 1 & \text{if high-school graduate,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{3,i} = \begin{cases} 1 & \text{if some college,} \\ 0 & \text{otherwise.} \end{cases}, \quad E_{4,i} = \begin{cases} 1 & \text{if college graduate,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{5,i} = \begin{cases} 1 & \text{if advanced degree,} \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ To avoid the dummy variable trap, one of the dummies has to be omitted:

$$Wage_i = \beta_0 + \delta_0 Female_i + \delta_2 E_{2,i} + \delta_3 E_{3,i} + \delta_4 E_{4,i} + \delta_5 E_{5,i} + \text{Other Factors}$$

- ▶ Group 1 (high-school dropout) becomes the base group.
- ▶ δ_2 measures the wage difference between high-school graduates and high-school dropouts.
- ▶ δ_3 measures the wage difference between individuals with some college education and high-school dropouts.

Testing for structural breaks or differences in regression functions across groups

- ▶ Suppose for simplicity we have two groups. For example,
 - ▶ Male and female workers.
 - ▶ Observations before and after a certain date.
- ▶ We want to test if the intercept and all slopes are the same across the two groups.
- ▶ The model:

$$Y_i = \beta_{1,0} + \beta_{1,1}X_{1,i} + \dots + \beta_{1,k}X_{k,i} + U_i \text{ if } i \text{ belongs to Group 1}$$

$$Y_i = \beta_{2,0} + \beta_{2,1}X_{1,i} + \dots + \beta_{2,k}X_{k,i} + U_i \text{ if } i \text{ belongs to Group 2}$$

- ▶ The hypotheses:

$$H_0 : \beta_{1,0} = \beta_{2,0}, \beta_{1,1} = \beta_{2,1}, \dots, \beta_{1,k} = \beta_{2,k}.$$

$$H_1 : \beta_{1,j} \neq \beta_{2,j} \text{ at least for one } j \in \{0, 1, \dots, k\}.$$

Testing for structural breaks or differences in regression functions across groups

$$Y_i = \beta_{1,0} + \beta_{1,1}X_{1,i} + \dots + \beta_{1,k}X_{k,i} + U_i \text{ if } i \text{ belongs to Group 1}$$

$$Y_i = \beta_{2,0} + \beta_{2,1}X_{1,i} + \dots + \beta_{2,k}X_{k,i} + U_i \text{ if } i \text{ belongs to Group 2}$$

- ▶ The Chow F statistic:

$$F^{Chow} = \frac{(SSR_r - SSR_{ur}) / (k + 1)}{SSR_{ur} / (n - 2(k + 1))} = \frac{(SSR_r - (SSR_1 + SSR_2)) / (k + 1)}{(SSR_1 + SSR_2) / (n - 2(k + 1))},$$

where

- ▶ SSR_1 is the SSR obtained by estimating the model using only the observations from Group 1.
- ▶ SSR_2 is the SSR obtained by estimating the model using only the observations from Group 2.
- ▶ SSR_r is the SSR obtained by pooling the groups and estimating a single equation:

$$Y_i = \gamma_0 + \gamma_1 X_{1,i} + \dots + \gamma_k X_{k,i} + U_i \text{ for all } i\text{'s (Groups 1 and 2)}.$$

- ▶ H_0 of constancy or no structural break is rejected when

$$F^{Chow} > F_{k+1, n-2(k+1), 1-\alpha}.$$

Testing for structural breaks or differences in regression functions across groups

- ▶ The Chow test can also be performed using the dummy variables, and the two approaches are numerically equivalent.
- ▶ Define

$$D_i = \begin{cases} 1 & \text{observation } i \text{ belongs to Group 1,} \\ 0 & \text{otherwise.} \end{cases}$$

- ▶ Estimate the following single equation using all observations (Groups 1 and 2):

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_k X_{k,i} + \delta_0 D_i + \delta_1 (D_i \cdot X_{1,i}) + \dots + \delta_k (D_i \cdot X_{k,i}) + U_i.$$

- ▶ Test:

$$H_0 : \delta_0 = \delta_1 = \dots = \delta_k = 0.$$

$$H_1 : \delta_j \neq 0 \text{ for at least one } j \in \{0, 1, \dots, k\}.$$