Introductory Econometrics

Lecture 14: Dummy variables

Instructor: Ma, Jun

Renmin University of China

October 28, 2021

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.

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

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

- 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

person	*wage*	*educ*	*exper*	*female*	*married*
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, $\mathrm{E}[U_i] = 0$.

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

$$\mathrm{E}[Wage_i \mid Female_i = 1, Educ_i, Exper_i, Tenure_i] = \beta_0 + \delta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i.$$

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

$$\mathrm{E}[Wage_i \mid Female_i = 0, Educ_i, Exper_i, Tenure_i] = \beta_0 + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i.$$

- Thus,

$$\delta_0 = \mathrm{E}[Wage_i \mid Female_i = 1, Educ_i, Exper_i, Tenure_i] - \\ - \mathrm{E}[Wage_i \mid Female_i = 0, Educ_i, Exper_i, Tenure_i].$$

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

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

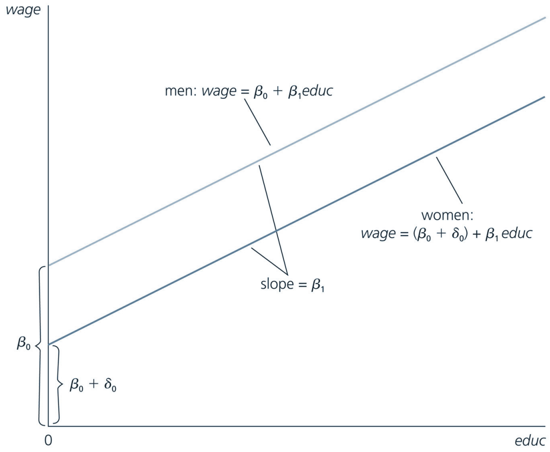
- For women ($Female_i = 1$):, we can write the model as

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

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

FIGURE 7.1

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



Example

- Estimated equation:

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

- 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:

$$\log(Wage) = \beta_0 + \delta_0 Female + \beta_1 Educ + \beta_3 Exper + \beta_4 Tenure + U.$$

- In this case,

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

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

Example

- Estimated equation:

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

- $\hat{\delta}_0 = -0.297$ 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

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

consider:

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

- Since $Male_i = 1 - Female_i$,

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

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

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

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

The dummy variable trap

- Consider the equation:

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

- 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!

- One of the dummies has to be omitted and the corresponding group becomes the base group:
  - Men are the base group: $\log(Wage_i) = \beta_0 + \delta_0 Female_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.
  - Women are the base group: $\log(Wage_i) = \theta_0 + \gamma_0 Male_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.
- Alternatively, one can include both dummies without the intercept: $\log(Wage_i) = \pi_0 Female_i + \pi_1 Male_i + \beta_1 Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i$.

  - 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:

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

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

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

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

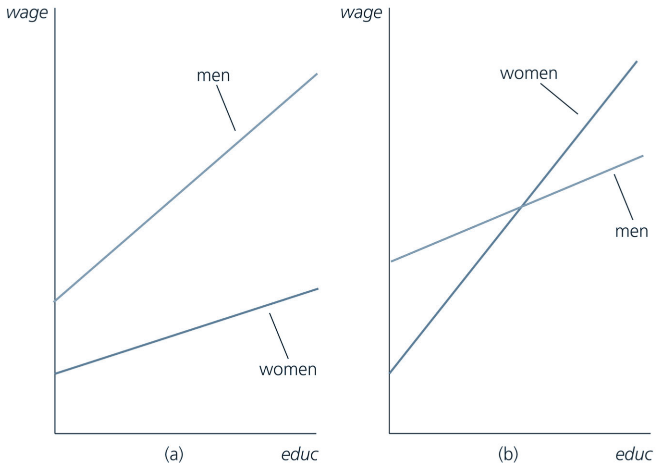
$$\log\left(Wage_i^F\right) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) Educ_i + \beta_3 Exper_i + \beta_4 Tenure_i + U_i.$$

- 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:

$$\begin{aligned}
\widehat{\log(Wage_i)} = {} & \underset{(0.119)}{0.389} - \underset{(0.168)}{0.227}\ Female_i \\
& + \underset{(0.008)}{0.082}\ Educ_i - \underset{(0.0131)}{0.0056}\ Female_i \cdot Educ_i \\
& + \underset{(0.005)}{0.029}\ Exper_i - \underset{(0.00011)}{0.00058}\ Exper_i^2 \\
& + \underset{(0.007)}{0.032}\ Tenure_i - \underset{(0.00024)}{0.00059}\ Tenure_i^2.
\end{aligned}$$

- $\hat{\delta}_1 = -0.0056$ 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.

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

- 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:

$$
\begin{aligned}
Y_i &= \beta_{1,0} + \beta_{1,1} X_{1,i} + \ldots + \beta_{1,k} X_{k,i} + U_i \text{ if } i \text{ belongs to Group 1} \\
Y_i &= \beta_{2,0} + \beta_{2,1} X_{1,i} + \ldots + \beta_{2,k} X_{k,i} + U_i \text{ if } i \text{ belongs to Group 2}
\end{aligned}
$$

- The hypotheses:

$$
\begin{aligned}
H_0 &: \beta_{1,0} = \beta_{2,0}, \beta_{1,1} = \beta_{2,1}, \ldots, \beta_{1,k} = \beta_{2,k}. \\
H_1 &: \beta_{1,j} \neq \beta_{2,j} \text{ at least for one } j \in \{0, 1, \ldots, k\}.
\end{aligned}
$$

$$
\begin{aligned}
Y_i &= \beta_{1,0} + \beta_{1,1} X_{1,i} + \ldots + \beta_{1,k} X_{k,i} + U_i \text{ if } i \text{ belongs to Group 1} \\
Y_i &= \beta_{2,0} + \beta_{2,1} X_{1,i} + \ldots + \beta_{2,k} X_{k,i} + U_i \text{ if } i \text{ belongs to Group 2}
\end{aligned}
$$

- 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:

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

- H_0 of constancy or no structural break is rejected when

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

- 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 (Groups1 and 2):

$$\begin{aligned} Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_k X_{k,i} \\ + \delta_0 D_i + \delta_1 \left(D_i \cdot X_{1,i} \right) + \ldots + \delta_k \left(D_i \cdot X_{k,i} \right) + U_i . \end{aligned}$$

- Test:

$$\begin{aligned} H_0 &: \quad \delta_0 = \delta_1 = \ldots = \delta_k = 0. \\ H_1 &: \quad \delta_j \neq 0 \text{ for at least one } j \in \{0, 1, \ldots, k\} . \end{aligned}$$