ECON 351* -- Introduction to NOTE 21

Introduction to Using Dummy Variable Regressors in Regression Models

- Consider the regression model for the average weekly earnings of individual workers given by the following **population regression equation (PRE)**:

$$\text{earn}_i = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{fem}_i + \beta_3 \text{fem}_i \text{ed}_i + u_i \qquad (1)$$

where u_i is an iid (independently and identically distributed) random error term, and the observable variables are defined as follows:

***earn**$_i$* = person i's **average weekly earnings** during a calendar year;

***ed**$_i$* = person i's years of **completed formal education**;

***fem**$_i$* = a **female *indicator* variable**, or ***dummy* variable**, defined such that
$\text{fem}_i = 1$ if person i is female, $= 0$ if person i is male.

- The earnings regression equation (1) contains two explanatory variables: the continuous variable ***ed**$_i$* and the indicator (dummy) variable regressor ***fem**$_i$*.

- The dummy variable regressor ***fem**$_i$* enters earnings regression equation (1) in two distinct ways:

1. it enters ***additively*** on its own as ***fem**$_i$*;

2. it enters ***multiplicatively*** interacted with ***ed**$_i$* in the regressor ***fem**$_i$**ed**$_i$*.

1. Interpreting the Simple Earnings Regression with a Female Dummy Variable Regressor

$$\text{earn}_i = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{fem}_i + \beta_3 \text{fem}_i \text{ed}_i + u_i \quad (1)$$

- ***Question 1:*** **What is the slope coefficient β_2 of the female indicator variable *fem_i*?**

  Answer: The slope coefficient β_2 of the dummy variable regressor *fem_i* is the **female-male *difference* in intercept coefficients**, i.e.,

  β_2 = the ***female*** **intercept coefficient – the *male* intercept coefficient**

- ***Question 2:*** **What is the slope coefficient β_3 of the female indicator interaction term *$fem_i ed_i$*?**

  Answer: The slope coefficient β_3 of the dummy variable interaction regressor *$fem_i ed_i$* is the **female-male *difference* in the slope coefficients of the regressor *ed_i*.**

  β_3 = the ***female*** **slope coefficient of *ed_i* – the *male* slope coefficient of *ed_i***

- ***Question 3:*** **What is the equation intercept coefficient β_0** in equation (1)**?**

  Answer: The equation intercept coefficient β_0 is the intercept of the earnings equation for males, for whom the female dummy variable *fem_i* equals 0, i.e., for whom *fem_i* = 0.

- ***Question 4:*** **What is the slope coefficient β_1 of the regressor *ed_i*** in equation (1)**?**

  Answer: The slope coefficient β_1 of the regressor *ed_i* in equation (1) is the slope coefficient of *ed_i* **for *males*** for whom *fem_i* = 0 by definition.

2. Demonstrating the Interpretation of the Regression Coefficients in Equation (1)

$$\text{earn}_i = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{fem}_i + \beta_3 \text{fem}_i \text{ed}_i + u_i \quad (1)$$

- The **population regression function (PRF)** corresponding to regression equation (1) gives the **conditional mean weekly earnings of female and male workers** with different levels of formal education (i.e., different values of the continuous explanatory variable ***ed***$_i$).

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{fem}_i + \beta_3 \text{fem}_i \text{ed}_i \quad (2)$$

The population regression function (2) contains two separate regression functions: (1) a female regression function; and (2) a male regression function.

- The ***female*** **population regression function** is obtained by setting the **female indicator variable** $\boldsymbol{fem_i}$ equal to 1 everywhere it appears in regression function (2). It gives the conditional mean earnings of female workers as a function of $\boldsymbol{ed_i}$, years of formal education.

Setting $\boldsymbol{fem_i = 1}$ in regression function (2) gives:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 + \beta_3 \text{ed}_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3)\text{ed}_i \quad (2f)$$

- The ***male*** **population regression function** is obtained by setting the **female indicator variable** $\boldsymbol{fem_i}$ equal to 0 everywhere it appears in regression function (2). It gives the conditional mean earnings of male workers as a function of $\boldsymbol{ed_i}$, years of formal education.

Setting $\boldsymbol{fem_i = 0}$ in regression function (2) gives:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_0 + \beta_1 \text{ed}_i \quad (2m)$$

- The ***female-male* difference in mean weekly earnings** is obtained by subtracting the male regression function (2m) from the female regression function (2f).

- The ***female* mean earnings function** is the female regression function (2f):

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) = \beta_0 + \beta_1\text{ed}_i + \beta_2 + \beta_3\text{ed}_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3)\text{ed}_i \quad (2f)$$

Note: $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_2}$ is the ***intercept*** of the mean earnings function **for *females***; $\boldsymbol{\beta_1}$ + $\boldsymbol{\beta_3}$ is the **slope coefficient** of the regressor $\boldsymbol{ed_i}$ **for *females*.**

- The ***male* mean earnings function** is the male regression function (2m):

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_0 + \beta_1\text{ed}_i \quad (2m)$$

Note: $\boldsymbol{\beta_0}$ is the ***intercept*** of the mean earnings function **for *males***; $\boldsymbol{\beta_1}$ is the **slope coefficient** of the regressor $\boldsymbol{ed_i}$ **for *males*.**

- Subtracting the male regression function (2m) from the female regression function (2f) gives the expression for the ***female-male difference* in mean weekly earnings**:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) - E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_0 + \beta_1\text{ed}_i + \beta_2 + \beta_3\text{ed}_i - \beta_0 - \beta_1\text{ed}_i = \beta_2 + \beta_3\text{ed}_i$$

i.e.,

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) - E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_2 + \beta_3\text{ed}_i \quad (3)$$

Result: Regression equation (2) implies that the ***female-male difference* in mean weekly earnings** is the linear function (3) of the explanatory variable $\boldsymbol{ed_i}$:

Interpretation of the Regression Coefficients in Regression Equation (1)

$$earn_i = \beta_0 + \beta_1 ed_i + \beta_2 fem_i + \beta_3 fem_i ed_i + u_i \qquad (1)$$

- The **population regression function (PRF)** corresponding to regression equation (1) is:

$$E(earn_i \mid ed_i, fem_i) = \beta_0 + \beta_1 ed_i + \beta_2 fem_i + \beta_3 fem_i ed_i \qquad (2)$$

- The ***female* mean earnings function** is the female regression function (2f):

$$E(earn_i \mid ed_i, fem_i = 1) = \beta_0 + \beta_1 ed_i + \beta_2 + \beta_3 ed_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) ed_i \qquad (2f)$$

- The ***male* mean earnings function** is the male regression function (2m):

$$E(earn_i \mid ed_i, fem_i = 0) = \beta_0 + \beta_1 ed_i \qquad (2m)$$

The ***female* intercept** coefficient = $\boldsymbol{\beta_0 + \beta_2}$

The ***male* intercept** coefficient = $\boldsymbol{\beta_0}$

Therefore $\boldsymbol{\beta_2}$ **= the *female* intercept coefficient ($\boldsymbol{\beta_0 + \beta_2}$) minus the *male* intercept coefficient ($\boldsymbol{\beta_0}$)**

The ***female* slope coefficient of *ed*$_i$** = $\boldsymbol{\beta_1 + \beta_3}$

The ***male* slope coefficient of *ed*$_i$** = $\boldsymbol{\beta_1}$

Therefore $\boldsymbol{\beta_3}$ **= the *female* slope coefficient of *ed*$_i$ ($\boldsymbol{\beta_1 + \beta_3}$) minus the *male* slope coefficient of *ed*$_i$ ($\boldsymbol{\beta_1}$)**

4. A More General Earnings Regression Model with a Female Dummy Variable Regressor

Consider now an expanded earnings regression model for female and male workers that allows for non-constant marginal earnings effects of $\boldsymbol{ed_i}$. The population regression equation for this model is:

$$\text{earn}_i = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 + \beta_3 \text{fem}_i + \beta_4 \text{fem}_i \text{ed}_i + \beta_5 \text{fem}_i \text{ed}_i^2 + u_i \tag{3}$$

Note that regression equation (3) allows for the possibly that the **marginal earnings effect of *ed*** may be ***increasing* or *decreasing*** in $\boldsymbol{ed_i}$.

- The **population regression function (PRF)** corresponding to regression equation (3) gives the **conditional mean weekly earnings of female and male workers** with different levels of formal education (i.e., different values of the continuous explanatory variable $\boldsymbol{ed_i}$).

  $$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 + \beta_3 \text{fem}_i + \beta_4 \text{fem}_i \text{ed}_i + \beta_5 \text{fem}_i \text{ed}_i^2 \tag{3p}$$

  The population regression function (3p) contains two separate regression functions: (1) a female regression function; and (2) a male regression function.

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 + \beta_3 \text{fem}_i + \beta_4 \text{fem}_i \text{ed}_i + \beta_5 \text{fem}_i \text{ed}_i^2 \quad (3p)$$

- The ***female*** **population regression function** is obtained by setting the **female indicator variable** $\boldsymbol{fem_i}$ equal to 1 everywhere it appears in regression function (3p). It gives the conditional mean earnings of female workers as a function of $\boldsymbol{ed_i}$, years of formal education.

  Setting $\boldsymbol{fem_i}$ **= 1** in regression function (3p) gives the ***female*** **mean earnings function**:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 + \beta_3 + \beta_4 \text{ed}_i + \beta_5 \text{ed}_i^2 = \beta_0 + \beta_3 + (\beta_1 + \beta_4)\text{ed}_i + (\beta_2 + \beta_5)\text{ed}_i^2 \quad (3f)$$

- The ***male*** **population regression function** is obtained by setting the **female indicator variable** $\boldsymbol{fem_i}$ equal to 0 everywhere it appears in regression function (3p). It gives the conditional mean earnings of male workers as a function of $\boldsymbol{ed_i}$, years of formal education.

  Setting $\boldsymbol{fem_i}$ **= 0** in regression function (3p) gives the ***male*** **mean earnings function**:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 \quad (3m)$$

- The ***female-male*** **difference in mean weekly earnings** is obtained by subtracting the male regression function (3m) from the female regression function (3f):

$$\begin{aligned} E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) - E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) \\ &= \beta_0 + \beta_1 \text{ed}_i + \beta_2 \text{ed}_i^2 + \beta_3 + \beta_4 \text{ed}_i + \beta_5 \text{ed}_i^2 - \beta_0 - \beta_1 \text{ed}_i - \beta_2 \text{ed}_i^2 \\ &= \beta_3 + \beta_4 \text{ed}_i + \beta_5 \text{ed}_i^2 \quad (3d) \end{aligned}$$

  Result: Regression equation (3) implies that the ***female-male difference*** **in mean weekly earnings** is the quadratic function (3d) of the continuous explanatory variable $\boldsymbol{ed_i}$.

Summary Interpretation of the Regression Coefficients in Regression Equation (3)

- The **population regression function (PRF)** corresponding to regression equation (3) is:

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i) = \beta_0 + \beta_1\text{ed}_i + \beta_2\text{ed}_i^2 + \beta_3\text{fem}_i + \beta_4\text{fem}_i\text{ed}_i + \beta_5\text{fem}_i\text{ed}_i^2 \qquad (3p)$$

- The ***female*** **mean earnings function** is the female regression function (3f):

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 1) = \beta_0 + \beta_1\text{ed}_i + \beta_2\text{ed}_i^2 + \beta_3 + \beta_4\text{ed}_i + \beta_5\text{ed}_i^2 = \beta_0 + \beta_3 + (\beta_1 + \beta_4)\text{ed}_i + (\beta_2 + \beta_5)\text{ed}_i^2 \qquad (3f)$$

- The ***male*** **mean earnings function** is the male regression function (3m):

$$E(\text{earn}_i \mid \text{ed}_i, \text{fem}_i = 0) = \beta_0 + \beta_1\text{ed}_i + \beta_2\text{ed}_i^2 \qquad (3m)$$

The ***female*** **intercept** coefficient = $\boldsymbol{\beta_0} + \boldsymbol{\beta_3}$
The ***male*** **intercept** coefficient = $\boldsymbol{\beta_0}$

Therefore $\boldsymbol{\beta_3}$ **= the *female* intercept coefficient ($\beta_0 + \beta_3$) <u>minus</u> the *male* intercept coefficient (β_0)**

The ***female*** **slope coefficient of** ed_i = $\boldsymbol{\beta_1} + \boldsymbol{\beta_4}$
The ***male*** **slope coefficient of** ed_i = $\boldsymbol{\beta_1}$

Therefore $\boldsymbol{\beta_4}$ **= the *female* slope coefficient of ed_i ($\beta_1 + \beta_4$) <u>minus</u> the *male* slope coefficient of ed_i (β_1)**

The ***female*** **slope coefficient of** ed_i^2 = $\boldsymbol{\beta_2} + \boldsymbol{\beta_5}$
The ***male*** **slope coefficient of** ed_i^2 = $\boldsymbol{\beta_2}$

Therefore $\boldsymbol{\beta_5}$ **= the *female* slope coefficient of ed_i^2 ($\beta_2 + \beta_5$) <u>minus</u> the *male* slope coefficient of ed_i^2 (β_2)**

The Marginal Earnings Effect of ed in Regression Equation (3)

- The **marginal earnings effect of *ed_i*** is obtained by partially differentiating the **population regression function (PRF)** for regression equation (3) with respect to ***ed_i***:

$$E(earn_i \mid ed_i, fem_i) = \beta_0 + \beta_1 ed_i + \beta_2 ed_i^2 + \beta_3 fem_i + \beta_4 fem_i ed_i + \beta_5 fem_i ed_i^2 \quad (3p)$$

$$\frac{\partial E(earn_i \mid ed_i, fem_i)}{\partial ed_i} = \beta_1 + 2\beta_2 ed_i + \beta_4 fem_i + 2\beta_5 fem_i ed_i \quad (4)$$

- The **marginal earnings effect of *ed_i* for *females*** is obtained by setting the female indicator variable ***fem_i* = 1** in the above expression (4) for the marginal earnings effect of ***ed_i***:

$$\frac{\partial E(earn_i \mid ed_i, fem_i = 1)}{\partial ed_i} = \beta_1 + 2\beta_2 ed_i + \beta_4 + 2\beta_5 ed_i = (\beta_1 + \beta_4) + 2(\beta_2 + \beta_5) ed_i \quad (4f)$$

- The **marginal earnings effect of *ed_i* for *males*** is obtained by setting the female indicator variable ***fem_i* = 0** in expression (4) for the marginal earnings effect of ***ed_i***:

$$\frac{\partial E(earn_i \mid ed_i, fem_i = 0)}{\partial ed_i} = \beta_1 + 2\beta_2 ed_i \quad (4m)$$

- The ***female-male difference* in the marginal earnings effect of *ed_i*** is obtained by subtracting (4m) from (4f):

$$\frac{\partial E(earn_i \mid ed_i, fem_i = 1)}{\partial ed_i} - \frac{\partial E(earn_i \mid ed_i, fem_i = 0)}{\partial ed_i} = \beta_4 + 2\beta_5 ed_i \quad (4d)$$