

Notes on Dummy RHS Variables, Interactive Terms

Copyright - Jonathan Nagler; Nov 10, 2006

Types of Discrete Variables: Ordinal versus Cardinal versus Categorical Variables:

Cardinal variable: the difference between the values j and the value $j + 1$ is the same as the difference between the values k and $k + 1$. Some common cardinal variables: wages, population.

Ordinal variable: $j + 1$ is bigger than j ; but the difference between $j + 1$ and j is not necessarily the difference between k and $k + 1$.

Some Common Ordinal Variables:

Responses to survey questions: “What is your view on legalization of marijuana?” 1 = strongly opposed; 2 = somewhat opposed; 3 = somewhat in favor; 4 = strongly in favor.

Income: 1 = \$0 - \$10,000; 2 = \$10,000 - \$25,000; 3 = \$25,001 - \$75,000; 4 = \$75,001 - \$100,000; 5 = \$100,001 and above.

Notice that income in this case is an example of transforming a cardinal variable into an ordinal variable.

Education: 1 = less than high-school; 2 = high-school grad; 3 = some college; 4 = college grad or beyond.

Categorical Variable: a discrete variable that categorizes a set.

Common Example: **Region:** 1 = east, 2 = south; 3 = west; 4 = north.

[Note: obviously (duh!) one would never put a categorical variable on the RHS of a regression model; **unless** it were either ordinal or cardinal.]

Dummy Variable: Generally used to refer to a discrete variable that can take the values 0 or 1.

Example: **Woman.**

We can generally replace any discrete variable with 1 or more binary variables ($k-1$). So we could have **education** on the RHS of a model (a four-category variable). Or, we could have three binary variables: **LT_highschool**, **HS_grad**, **some_coll**.

If we just have **education** on the right-hand side, then we are claiming that the increase from **LT_highschool** to **HS_grad** is the same as the increase from **HS_grad** to **some_coll**. That might be very wrong.

Simple Cases of RHS Variables.

Example 1: You have a binary (0/1 categorical) RHS variable (**men**), indicating that you think the intercept for respondents with this characteristic is different than the intercept for the other respondents.

$$Y_i = \beta_0 + \beta_1 * \mathbf{male}_i + \beta_2 * \mathbf{educ}_i + \beta_3 * \mathbf{Z}_i + \epsilon$$

The intercept for men is different than for women.

Example 2: You have $k-1$ binary RHS variables (**east**, **west**, **south**) indicating that you think the intercept varies across regions.

$$Y_i = \beta_0 + \beta_1 * \mathbf{N}_i + \beta_2 * \mathbf{S}_i + \beta_3 * \mathbf{W}_i \\ + \beta_4 * \mathbf{educ}_i + \beta_5 * \mathbf{Z}_i + \epsilon$$

[Where is \mathbf{E}_i ???

In the constant!

Multiple RHS Categorical Variables:

Consider a case with two categorical variables: 1) Gender (M/F), and 2) (N, S, E, W).

There are 8 possible categories ('cells') people can be in: {M,N}, {M,S}, {M,E}, {M,W}, {F,N}, {F,S}, {F,E}, {F,W}.

So you can create a new variable with 8 categories. To include it, include (K - 1) of those variables.

$$\begin{aligned} Y_i = & \beta_0 + \beta_1 * (\mathbf{M} * \mathbf{N}) + \beta_2 * (\mathbf{M} * \mathbf{S}) + \beta_3(\mathbf{M} * \mathbf{E}) \\ & + \beta_4(\mathbf{M} * \mathbf{W}) + \beta_5(\mathbf{F} * \mathbf{N}) + \beta_6(\mathbf{F} * \mathbf{S}) \\ & + \beta_7(\mathbf{F} * \mathbf{E}) + \epsilon \end{aligned}$$

What About Three Categorical Variables:

Now consider a case with three categorical variables: 1) Gender (M/F); 2) (N, S, E, W); and 3) race (C/B). [C for caucasian.]

There are 16 possible categories ('cells') people can be in: {M,N,C}, {M,S,C}, {M,E,C}, {M,W,C}, {F,N,C}, {F,S,C}, {F,E,C}, {F,W,C}, {M,N,B}, {M,S,B}, {M,E,B}, {M,W,B}, {F,N,B}, {F,S,B}, {F,E,B}, {F,W,B}.

With 16 possible categories ('cells'), you are now including 15 RHS binary variables.

Slightly More Complicated Case - Add a Continuous RHS Variable:

We want to interact a dummy variable (say region) with a ‘continuous’ variable such as education.¹

Say we want to allow for K possible slopes for education. And we want to allow for K possible intercepts. Then we would have:

$$\begin{aligned} Y_i = & \beta_0 + \beta_1 * educ_i * N_i + \beta_2 * educ_i * S_i + \\ & \beta_3 * educ_i * W_i + \beta_4 * educ_i * E_i + \\ & \beta_5 * N_i + \beta_6 * S_i + \beta_7 * W_i + \\ & \beta_8 * Z_i + \epsilon \end{aligned}$$

$$\frac{\delta Y}{\delta educ} = \beta_1 * N_i + \beta_2 * S_i + \beta_3 * W_i + \beta_4 * E_i$$

So the derivative of Y with respect to $educ$ for any given respondent is **EITHER**: $\beta_1, \beta_2, \beta_3$, **or** β_4 , *conditional* on which region the respondent lives in.

¹Note that we switch back and forth between treating education as discrete (ordinal), and continuous. We hope the bogus assumption does not hurt us too much.

Go back to our simplest example (**Example 1:**) The intercept for men is different than for women.:

$$Y_i = \beta_0 + \beta_1 * \mathbf{educ}_i + \beta_2 * \mathbf{male}_i + \epsilon$$

But now let the effect of education vary based on gender. There are two ways to do this:

$$Y_i = \beta_0 + \beta_1 \mathbf{educ}_i + \beta_2 \mathbf{male}_i + \beta_3 (\mathbf{educ}_i * \mathbf{male}_i) + \epsilon$$

$$Y_i = \gamma_0 + \gamma_1 \mathbf{male}_i + \gamma_2 (\mathbf{educ}_i * \mathbf{male}_i) + \gamma_3 (\mathbf{educ}_i * \mathbf{female}_i) + \epsilon$$

So, the first equation gives the impact of changes in education on Y as:

$$\frac{\delta Y}{\delta \mathbf{educ}} = \beta_1 + \beta_3 * \mathbf{male}_i$$

And the second equation gives the impact of changes in education on Y as:

$$\frac{\delta Y}{\delta \mathbf{educ}} = \gamma_2 * \mathbf{male}_i + \gamma_3 * \mathbf{female}_i$$

Hypothesis: the rate of return to education is higher for men than women.

Equation 1:

$$H1: (\beta_1 + \beta_3) > \beta_1$$

$$H0: (\beta_1 + \beta_3) = \beta_1$$

OR:

$$H1: \beta_3 > 0$$

$$H0: \beta_3 = 0$$

Equation 2:

$$H1: \gamma_2 > \gamma_3$$

$$H0: \gamma_2 = \gamma_3$$

OR:

$$H1: \gamma_2 - \gamma_3 > 0$$

$$H0: \gamma_2 - \gamma_3 = 0$$

Hypothesis Tests:

1) t-test on $\beta_3 = 0$

2) t-test on $\gamma_2 - \gamma_3 = 0$

3) F-test on the linear restriction that $\gamma_1 = \gamma_2$.

These should all produce exactly the same results (i.e., level of confidence).

One More Thing:

If you include an interactive term ($\mathbf{X} * \mathbf{Y}$) on the RHS, then you generally need to include \mathbf{X} and \mathbf{Y} as separate RHS variables.

If not, you will likely have omitted variable bias.