AEDECON 7140: Application and Interpretation of Dummy Variables

Explanatory variables are often qualitative or discrete in nature (consider age versus sex), so that some indicator must be constructed to represent them in a regression [Why?]. A dummy variable is an artificial variable constructed such that it takes the value of one whenever the qualitative phenomenon it represents occurs, and zero otherwise. [A 0-1 coding is not always the case though.] They can be used in classical linear regression (CLR) models just like any other continuous variables.

For example, you want to analyze the impact of job, sex, and experience on wage rate and have the following data (N = 50):

id	wage rate	job	sex	years of experience
1	10.6464	1	1	21
2	4.238418	2	2	8
3	6.534406	2	2	23
4	10.42548	3	1	27
5	9.845904	3	1	26
…	…	…	…	…

Suppose the raw data you received were coded such that “wage rate” and “years of experience” are actual levels of wage rate and years of experience; “job” = 1, 2, and 3 corresponds to doctor, professor, and lawyer respectively; and “sex” = 1, and 2 corresponds to male and female respectively. You created some dummy variables and obtained the following:

id	wage rate	job	doc	prof	lawyer	sex	male	female	experience
1	10.6464	1	1	0	0	1	1	0	21
2	4.238418	2	0	1	0	2	0	1	8
3	6.534406	2	0	1	0	2	0	1	23
4	10.42548	3	0	0	1	1	1	0	27
5	9.845904	3	0	0	1	1	1	0	26
…	…	…	…	…	…	…	…	…	…

* There is another variable called “random” in the data. We will not talk about this variable in this handout.

Let us look at several models.

1. We want to see how different jobs may affect wage rate and we run the following model: $Y = \alpha_D D_D + \alpha_P D_P + \alpha_L D_L + \varepsilon$ (M11a), where $D_D, D_P,$ and D_L are three dummy variables representing doctor, professor, and lawyer. Notice that this model does not have a constant. What if you add a constant but remove D_L (label this as M11b)? We include a constant and run: $Y = \beta_0 + \beta_D D_D + \beta_P D_P + \beta_L D_L + \varepsilon$ (M12). How would we interpret the results?

2. We also want to consider sex in the analysis. Let's see the following two models and compare results:
   2.1 $Y = \alpha_D D_D + \alpha_P D_P + \alpha_L D_L + \alpha_F D_F + \varepsilon$ (M21a)
   [Q: For the above model, why do we not do the following:
   $Y = \alpha_D D_D + \alpha_P D_P + \alpha_L D_L + \alpha_F D_F + \alpha_M D_M + \varepsilon$ (M21b)?]
   2.2 $Y = \beta_0 + \beta_D D_D + \beta_P D_P + \beta_F D_F + \varepsilon$ (M22)

3. What do we must assume in the above models in terms of the impact of sex to wage rate? Let us see the following two models:
   3.1 $Y = \alpha_{FD} D_{FD} + \alpha_{MD} D_{MD} + \alpha_{FP} D_{FP} + \alpha_{MP} D_{MP} + \alpha_{FL} D_{FL} + \alpha_{ML} D_{ML} + \varepsilon$ (M31)
   3.2 $Y = \beta_0 + \beta_{FD} D_{FD} + \beta_{MD} D_{MD} + \beta_{FP} D_{FP} + \beta_{MP} D_{MP} + \beta_{FL} D_{FL} + \varepsilon$ (M32)

4. Now we want to see how experience may affect the wage rate. We can do the following model: $Y = \gamma_0 + \gamma_D D_D + \gamma_P D_P + \gamma_E Exper + \varepsilon$ (M4)

5. We can also do: $Y = \gamma_0 + \gamma_D D_D + \gamma_P D_P + \gamma_E Exper + \gamma_{ED}(D_D Exper) + \gamma_{EP}(D_P Exper) + \varepsilon$ (M5)
   [Q: what is the difference between M4 and M5?]
   [Q: what do we gain by running M5 compared to running three separate regressions, each using just the data for a particular profession (i.e., each model has only a constant and the variable *Exper*)? If you compare the SSEs of all four equations, what would you find?]

6. Under M5, here is how we can calculate the marginal effects and their associated variances:
   Marginal effect of a continuous variable, say $Exper$ is: $\frac{\partial E[Y|Exper, D_D, \ldots]}{\partial Exper} = \gamma_E + \gamma_{ED} D_D$. Obviously, this effect is different depending on whether $D_D = 1$ or 0. For the variance of this marginal effect, we have (expressed using coefficient estimates from the model):
   $Var\left(\frac{\partial E[Y|Exper, D_D, \ldots]}{\partial Exper}\right) = Var(\hat{\gamma}_E) + D_D^2 Var(\hat{\gamma}_{ED}) + 2D_D Cov(\hat{\gamma}_E, \hat{\gamma}_{ED})$ (note that for a dummy variable, $D_D^2 = D_D$). This variance term is also different depending on whether $D_D = 1$ or 0.

   Marginal effect of a dummy variable, say D_D is: $\Delta E[Y|Exper, D_D, \ldots] = E[Y|Exper, D_D = 1, \ldots] - E[Y|Exper, D_D = 0, \ldots] = \gamma_D + \gamma_{ED} Exper$
   [Q: try if you can interpret these effects.] The variance of this marginal effect is:
   $Var(\Delta E[Y|Exper, D_D, \ldots]) = Var(\hat{\gamma}_D) + Exper^2 Var(\hat{\gamma}_{ED}) + 2Exper Cov(\hat{\gamma}_E, \hat{\gamma}_{ED})$
   [Q: Interpret these.]

7. Construct a model yourself where $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_6 X_3^2 + \beta_7 X_1 X_2 + \beta_8 X_2 X_3 + \beta_9 X_1 X_3 + \varepsilon$ and see if you can interpret the results when X variables are dummy or continuous variables, respectively.

Be careful deriving the marginal effects of dummy variables. It is fairly simple in a linear model as we have seen above but becomes more complicated in a model involving logs. We will talk about this in detail.

Invariance Property

To summarize, given x_1, x_2, and x_3 are three dummy coded dummy variables created from a 3-level categorical variable where $x_1 + x_2 + x_3 = 1$, a regression $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is equivalent to $y = \beta_1(1 - x_2 - x_3) + \beta_2 x_2 + \beta_3 x_3$. Rearranging this can obtain $y = \beta_1 + (\beta_2 - \beta_1)x_2 + (\beta_3 - \beta_1)x_3$.

Now suppose we run a regression $y = Intercept + B_2 x_2 + B_3 x_3$. In this regression, we assume the effect of x_1 is zero, that is, $B_1 = 0$. Compare this model with the model in the previous paragraph, it is clear that $Intercept = \beta_1$, $B_2 = \beta_2 - \beta_1$, and $B_3 = \beta_3 - \beta_1$. Furthermore, since we assume $B_1 = 0$, we have the convenient property that $B_2 - B_1 = \beta_2 - \beta_1$ and $B_3 - B_1 = \beta_3 - \beta_1$. Finally, $B_2 - B_3 = \beta_2 - \beta_3$. All these are saying is that the difference between marginal utilities of different levels should not change between the two methods. This is a case of the invariance property.

Effects Coding

Discrete variables do not have to be coded as dummy variables. For instance, you may also encounter a method called “effects coding” or equivalently “sum coding”. Using the above example and treating “lawyer” as the omitted category, effects coding suggests the following variables for “doc” and “prof”. The key reason to perform effects coding is to allow coefficients of all levels (including the omitted level) sum to zero. This has benefit when considering interaction effects between different levels of a discrete variable.

id	wage rate	job	doc	prof
1	10.6464	1	1	0
2	4.238418	2	0	1
3	6.534406	2	0	1
4	10.42548	3	-1	-1
5	9.845904	3	-1	-1
…	…	…	…	…

The following table explains difference between dummy and effects coding. One should carefully interpret these two coding schemes to realize they reveal identical information.

	Dummy coding			Effects coding		
	Variable 1	Variable 2	Variable 3	Variable 1	Variable 2	Variable 3
Two-level						
Level 1 (base)	1	0			-1	
Level 2	0	1			1	
Impact on y	β_1	β_2		$\gamma_1 = -\gamma_2$	γ_2	
Regression coefficient	$B_1 = 0$	$B_2 = \beta_2 - \beta_1$		$G_1 = 0$	$G_2 = \gamma_2 - \gamma_1$	
Equivalence	$Int_B = Int_G - G_2$	$B_2 = 2G_2$		$Int_G = Int_B + \frac{1}{2}B_2$	$G_2 = \frac{1}{2}B_2$	
Three-level						
Level 1 (base)	1	0	0		-1	-1
Level 2	0	1	0		1	0
Level 3	0	0	1		0	1
Impact on y	β_1	β_2	β_3	$\gamma_1 = -(\gamma_2 + \gamma_3)$	γ_2	γ_3
Regression coefficient	$B_1 = 0$	$B_2 = \beta_2 - \beta_1$	$B_3 = \beta_3 - \beta_1$	$G_1 = 0$	$G_2 = \gamma_2 - \gamma_1$	$G_3 = \gamma_3 - \gamma_1$
Equivalence	$Int_B = Int_G - (G_2 + G_3)$	$B_2 = G_2 + (G_2 + G_3)$	$B_3 = G_3 + (G_2 + G_3)$	$Int_G = Int_B + \frac{1}{3}(B_2 + B_3)$	$G_2 = B_2 - \frac{1}{3}(B_2 + B_3)$	$G_3 = B_3 - \frac{1}{3}(B_2 + B_3)$

Several more points:

- Look at the 3-level case, the two "equivalence" conditions for B_2 and B_3 are:
  $$B_2 = G_2 + (G_2 + G_3)$$
  $$B_3 = G_3 + (G_2 + G_3)$$
  Solving these for G_2 and G_3 gives
  $$G_2 = B_2 - \frac{1}{3}(B_2 + B_3)$$
  $$G_3 = B_3 - \frac{1}{3}(B_2 + B_3)$$
- From the above, notice that $B_2 - B_3 = G_2 - G_3$. This is an important conclusion that no matter which coding scheme is used, the differences between coefficients of the *non-omitted* categories are equivalent.
- Extension to more sets of discrete variables is straightforward:
  $$Int_B = Int_G - \sum_{k_1=2}^{K_1} G_{k_1} - \sum_{k_2=2}^{K_2} G_{k_2} - \cdots$$
  $$Int_G = Int_B + \frac{\sum_{k_1=2}^{K_1} B_{k_1}}{K_1} + \frac{\sum_{k_2=2}^{K_2} B_{k_2}}{K_2} + \cdots$$
  where the number of discrete levels of each set is given by K_i.

Mean-Centering

Another approach often taken by researchers in marketing, health science, and sociology is referred to as mean-centering. The method and implications of mean-centering is best explained in an example. Given the regression on employee income, $Income = \beta_0 + \beta_1 Female + \beta_2 Age$, we know the interpretation of β_0 is the income of an employee who is male and is of 0 years of age. We know that "0 years of age" does not make a lot of practical sense. Holding age at sample or population mean will be a better idea. Let us hold age at sample mean. We can say that the income of a male employee measured as the sample average age is $\beta_0 + \beta_2 \overline{Age}$. Suppose we mean-center variable Age, and estimate $Income = \gamma_0 + \gamma_1 Female + \gamma_2 (Age - \overline{Age})$, then γ_0 can be interpreted as the income of a male employee who is at the sample average age. Mean-centering has another benefit that is to reduce the chance that different variables are measured in different orders of magnitude, which may lead to convergence difficulty in complex models.

Concluding Remarks

Dummy variables are common in economic analysis and beyond such as:

- To indicate structural shifts or seasonal factors
- Being used in analysis of variance (ANOVA) and analysis of covariance (ANCOVA)
- Being used as dependent variables
    - Discrete analysis
- Being used in experimental design (involving a tremendously wide range of disciplines and applications, often wider than we think)
    - Economic treatment effects (such as difference in differences and regression discontinuity design).