Dummy Variables

Dummy variables are used extensively in Econometrics and applied research. They are used to model the impact (treatment effect) of public policies, randomized control trials, and 'natural' experiments. They are also used to model categorical variables and fixed-effects (used to explain unobserved heterogeneity in cross-section and panel models). Given their proclivity, it is important that we understand their mechanics.

You might be surprised by how much 'real-world' data is not continuous. For example, if you downloaded a survey dataset you will likely find that most variables are in fact categorical. This is the result of a number of factors, including accuracy (avoiding reporting bias) and just the nature of real-world variables. For example, sureys often ask people to report their income on a binned scale (e.g., \$0-\$10,000; \$10,000-\$20,000; …) as a lot of people do not know their exact annual or monthly salary. Variables about a households structure (e.g. "2 parents, with children"; "Single parent, without children"; …) are naturally categorical.

Even if a variable may have (close to) continuous measure (e.g. age in months/days; years of education), you might choose to model the variable as categorical since it may not be cardinal. For example, a bachelors degree may correspond to a 15 years of education, compared to 12 years of education for a high school diploma. However, is this 3-year difference equivalent to the difference in education of individuals with 5 and 2 years of education? You could say this about income (measured in \$'s), height/weight/BMI, and employment tenure, as these measures are cardinal.

Notation

Any dummy variable is a discrete random variable (often denoted D) that takes on two values: $D_i \in \{0, 1\}$. It corresponds to a true-false statement:

$$D_i = \mathbf{1}\{\text{"true"}\}$$

This simple function returns the value 1 if the statement inside is true (for unit i), and 0 otherwise. The For example, suppose there was a categorical variable that took on m distinct values, each with their own label,

$$X_i \in \{x_1 \text{ "label 1"}, x_2 \text{ "label 2"}, ..., x_m \text{ "label m"}\}$$

For example, $favouritecolour_i \in \{1 \text{ "red"}, 2 \text{ "blue"}, 3 \text{ "yellow"}, 4 \text{ "green"}\}$. Then for each value of X_i, you can define a dummy variable:

$$D_{im} = \mathbf{1}\{X_i = x_m\}$$

A common usage of dummy variables is in the evaluation of randomized and 'natural' experiments. Researchers will typically define a single dummy variable to denote treatment status: $D_i = \mathbf{1}\{\text{“treated”}\}$. The dummy variable splits the sample into two groups: treated ($D_i = 1$) and control ($D_i = 0$).

Dummy regressors

In a basic setting the use a of dummy variable might be relatively straight forward. For example, consider the above exmple of a single treatment-status dummy variable. We can include this in a univariate regression model (including a constant term),

$$Y_i = \beta_1 + \beta_2 D_i + \varepsilon_i$$

We can show that β_2 can be interpretted as the difference between the mean of Y for the treated and control group (see material on interpretting linear models),

$$\beta_2 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Unfortunately, this notation is not going to help us move forward, so let us be a little more formal/detailed. Suppose, a categorical regressor takes on two values $X \in \{1, 2\}$. We can then define two dummy variables,

$$\begin{aligned} D_{i1} =& \mathbf{1}\{X_i = 1\} \\ D_{i2} =& \mathbf{1}\{X_i = 2\} \end{aligned}$$

Since the values of X are exhaustive and mutually exclusive, it must be that,

$$D_{i1} + D_{i2} = 1$$

For every unit, their value of X is either 1 or 2.

Now, let us consider the linear model that includes these dummy variables as regressors:

$$Y_i = \beta_1 + \beta_2 D_{i1} + \beta_3 D_{i2} + \varepsilon_i$$

We immediately have a problem. This model is not identified. This is due to a rank violation (perfect colinearity between regressors).

Suppose the data was sorted on X. Then consider the matrix of regressors in this model:

$$X = [\ell, D_1, D_2] = \begin{bmatrix} 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \end{bmatrix}$$

This matrix has 3 columns, but its rank is 2. Column 3 (D_2) is given by column 1 (ℓ) minus column 2 (D_1).

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

If X is not full rank, then the $k \times k$ matrix $X'X$ is not invertible. And neither is $E[X_i X_i']$, implying a failure of identification. We cannot seperately identify all three β-parameters $[\beta_1, \beta_2, \beta_3]$.

We can demonstrate the result in another way. Given that $D_{i1} + D_{i2} = 1$ (where 1 represents the constant regressor), we can substitute D_{i2} with $1 - D_{i1}$.

$$\begin{aligned} Y_i =& \beta_1 + \beta_2 D_{i1} + \beta_3(1 - D_{i1}) + \varepsilon_i \\ =& (\beta_1 + \beta_3) + (\beta_2 - \beta_3) D_{i1} + \varepsilon_i \end{aligned}$$

which we can rewrite as,

$$Y_i = \beta_{1(2)} + \beta_{2(2)} D_{i1} + \varepsilon_i$$

I'm using the notation of an additional subscript-(2) to denote the parameters corresponding to the model where we exclude the dummy variable D_{i2}.

Now, the X_2 matrix (denoted as such because it excludes D_2) has two columns and is full rank. Thus, $[\beta_{1(2)}, \beta_{2(2)}]$ are both identified.

$$X_{(2)} = [\ell, D_1] = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

Thus, the model is given by,

$$Y = X_{(2)}\beta_{(2)} + \varepsilon$$

We can have also substituted D_{i1} with $1 - D_{i2}$. This would give us the model,

$$Y_i = \beta_{1(1)} + \beta_{2(1)}D_{i2} + \varepsilon_i$$

where $\beta_{1(1)} = \beta_1 + \beta_2$ and $\beta_{2(1)} = \beta_3 - \beta_2$.

or,

$$Y = X_{(1)}\beta_{(1)} + \varepsilon$$

Note, the vectors $\beta_{(1)}$ and $\beta_{(2)}$ are not the same. However, you can recover the one from the other. For example,

$$\begin{aligned}\beta_{1(1)} =& \beta_1 + \beta_2 = \beta_1 + \beta_3 + (\beta_2 - \beta_3) = \beta_{1(2)} - \beta_{2(2)} \\ \beta_{2(1)} =& \beta_3 - \beta_2 = -\beta_{2(2)}\end{aligned}$$

Finally, we can substitute in the constant with $D_{i1} + D_{i2}$. This gives us the model,

$$Y_i = \beta_{1(0)}D_{i1} + \beta_{2(0)}D_{i2} + \varepsilon_i$$

where $\beta_{1(0)} = \beta_1 + \beta_2$ and $\beta_{2(0)} = \beta_1 + \beta_3$.

or,

$$Y = X_{(0)}\beta_{(0)} + \varepsilon$$

While it might seem strange to exclude the constant term, it is implicitly included. This is because the regressors add up to one.

Dummy projections

We have considered three different models, each with a different excluded variable:

1. $Y = X_{(2)}\beta_{(2)} + \varepsilon$
2. $Y = X_{(1)}\beta_{(1)} + \varepsilon$
3. $Y = X_{(0)}\beta_{(0)} + \varepsilon$

Each model corresponds to a different projection matrix: $P_X = X(X'X)^{-1}X'$. I will denote these $P_{(2)}$, $P_{(1)}$, and $P_{(0)}$. It turns out that,

$$P_{(2)} = P_{(1)} = P_{(0)}$$

Implying the same result for the orthogonal projections: $M_{(2)}$, $M_{(1)}$, and $M_{(0)}$. It implies that the predicted values and residuals are the same for all three models.

Dummy covariates

The above result has important implications for discrete/categorical covariates. Suppose you had a regression with a single a continuous regressors X_1 and then a single categorical variable $X_2 \in \{x_1, ..., x_m\}$. There will be $k = m + 1$ parameters in this model. You would define a set of dummy variables for each category and include $m - 1$ in the model, assuming a constant is also included.[1]

$$Y = \beta_1 X_1 + \beta_{21(1)} + \beta_{22(1)} D_2 + ... + \beta_{2m(1)} D_m + \upsilon$$

This can be written as,

$$Y = X_1\beta_1 + X_{2(1)}\beta_{2(1)} + \upsilon$$

where,

$$X_{2(1)} = [1, D_2, D_3, ..., D_m]$$

The partitioned regression result tells us that the OLS-estimator for β_1 is given by,

$$(M_{2(1)}Y) = (M_{2(1)}X_1)\beta_1 + \xi$$

where $M_{2(1)}$ is the orthogonal projection matrix of all other regressors in the model. In this case, that is the orthogonal projection matrix of the matrix,

In this matrix, D_1 has been excluded as the base category. Above, we showed that,

$$M_{2(0)} = M_{2(1)} = M_{2(2)} = ... = M_{2(m)}$$

Thus, the choice of base category for the covariate regressors has no impact on our estimate of β_1, the parameter of interest. We could have written the same model, excluding D_2

$$Y = \beta_1 X_1 + \beta_{21(2)} + \beta_{22(2)} D_1 + \beta_{23(2)} D_3 + ... + \beta_{2m(2)} D_m + \upsilon$$

[1] I have deliberately moved the order of the variables so that the constant and dummy variables can be grouped together as a single set of regressors X_2. In this way, the model follows the notation from Handout 1, in which $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$.

Or even excluding the constant,

$$Y = \beta_1 X_1 + \beta_{21(0)} D_1 + \beta_{22(0)} D_2 + ... + \beta_{2m(0)} D_m + \upsilon$$

Regardless, of the base category, the estimator for β_1 remains unchanged.

Fixed Effects

The above result helps rationalize a fairly common notation in applied econometrics. Suppose you were studying the relationship between (log of) wages and education. We know that there is a lot of spatial heteroegeneity in wages: some occupations in certain cities pay a lot more than the same occupation in a different city. This could be for a whole range of reasons - some individual (differences in education/training, etc.) and some environmental (local ammenities, labour competition, number of employers etc.). Some of these location factors may be obserable, but some not. If you are willing to assume that these characteristics relate to the local area and not the specific individual, then you could model them as,

$$A_l = [A_{l1}, ..., A_{ls}]'$$

Where A_l is a $s \times 1$ vector of l-location's characteristics. You can use the notation $A_{l(i)}$ to link the individual to a specific location. Recall, these are common to all individuals in this location. We can think of this linear combination of location specific variables as a single location-specific parameter α_l.

$$\alpha_l = A'_{l(i)} \gamma$$

By doing this, we remove the need to observe all the variables in A_l. However, we must observe multiple people from the same location.

Suppose the model you have in mind is given by,

$$\ln(w_i) = \alpha + \beta edu_i + A'_{l(i)} \gamma + \varepsilon_i$$

We can't estimate this model because we can't observe all variables in A. However, we can estimate the model,

$$\ln(w_i) = \alpha_1 + \beta edu_i + \sum_{j=2}^{m} \alpha_{2j} \mathbf{1}\{location_i = j\} + \varepsilon_i$$

where we exclude the first (arbitrary) location. Or, we can drop the constant and estimate,

$$\ln(w_i) = \beta edu_i + \sum_{j=1}^{m} \alpha_j \mathbf{1}\{location_i = j\} + \varepsilon_i$$

Since, an individual will have only one location (in a given period of time), unit i in location j will just return,

$$\sum_{j=1}^{m} \alpha_j \mathbf{1}\{location_i = j\} = \alpha_j$$

So, the model can be written as,

$$\ln(w_{ij}) = \beta edu_{ij} + \alpha_j + \varepsilon_{ij}$$

You can think of α_j as a location-specific constant. This is also why it makes sense to drop the constant whe using fixed-effects notation. This constant explains all the unobserved spatial heterogeneity in the earnings of individuals.

From a modelling perspective, it is important to remember that this notation essentially hides a whole set of dummy variables that are implicitly included in the model. The number of regressors in this model will be $k = 1 + m$ where m is the number of locations.