Ch 7: **Dummy** (**binary**, **indicator**) **variables**

:Examples

Dummy variable are used to indicate the presence or absence of a characteristic. For example, define

$$female_i = \begin{cases} 1 \text{ if obs } i \text{ is female} \\ 0 \text{ otherwise} \end{cases}$$

$$male_i = \begin{cases} 1 \text{ if obs } i \text{ is male} \\ 0 \text{ otherwise} \end{cases}$$

or

$$married_i = \begin{cases} 1 \text{ if obs } i \text{ is married} \\ 0 \text{ otherwise} \end{cases}$$

$$ON_i = \begin{cases} 1 \text{ if obs } i \text{ lives in Ont.} \\ 0 \text{ otherwise} \end{cases}$$

$$G_i = \begin{cases} 1 \text{ if } x_{ij} \leq x_j^* \\ 0 \text{ otherwise} \end{cases}$$

We can use dummy variables to allow responses (regression coefficients) to vary across groups.

- For example, suppose we posit

$$wage_i = \beta_0(i) + \beta_1 educ_i + u_i \text{ with } \beta_0(i) = \beta_0 + \delta_0 female_i$$

$$\therefore\ wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i$$

- Notice

$$\begin{aligned}\delta_0 &= E(wage|female, educ) - E(wage|not\ female, educ) \\ &= E(wage|female, educ) - E(wage|male, educ)\end{aligned}$$

- Because $female_i + male_i = 1$, we could substitute out for the female dummy or the intercept and get alternate and equivalent parameterizations.

Alternative parameterizations

- Suppose that instead of

$$wage_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + u_i$$

we ran

$$wage_i = \alpha_0 + \gamma_0 male_i + \gamma_1 educ_i + u_i$$

Notice $Sp(X)$ is the same for both models.

- How are the coefficients in the two models above related? Substitute into the second specification

$$\begin{aligned} wage_i &= \alpha_0 + \gamma_0(1 - female_i) + \gamma_1 educ_i + u_i \\ &= (\alpha_0 + \gamma_0) - \gamma_0 female_i + \gamma_1 educ_i + u_i \end{aligned}$$

Therefore

$$\beta_0 = (\alpha_0 + \gamma_0) \quad \delta_0 = -\gamma_0 \quad \beta_1 = \gamma_1$$

- We know that the OLS estimates of the two models satisfy exactly the same restrictions as the population parameters.
- We know that a *t*-test of the null $\delta_0 = 0$ would yield exactly the same value as a *t*-test of the null $\gamma_0 = 0$
- A third and equivalent model is

$$wage_i = \beta_0 male_i + \alpha_0 female_i + \beta_1 educ_i + u_i$$

Notice that although this specification doesn't have an explicit constant as an intercept, we still have $\iota \in Sp(X)$. Unfortunately, most regression programs won't catch this and won't know how to calculate R^2 for this specification. For this reason, in practice we prefer a specification that explicitly contains a constant as a regressor.

- Notice that we can't include both dummy variables and the intercept as regressors (dummy variable trap). This would violate X has full column rank. The OLS estimator would no longer be unique (but $\hat{y}$ would be)

- No reason to focus only on the intercept. We could posit

$$wage_i = \beta_0(i) + \beta_1(i)educ_i + u_i$$

with

$$\beta_0(i) = \beta_0 + \delta_0 female_i$$

$$\beta_1(i) = \beta_1 + \delta_1 female_i$$

Exercise: What's the relation between the OLS estimates of this model and the OLS estimates obtained from separate regression of wages on education for men and women?

Ex. 1 Test if men and women are paid the same wage vs women are paid less.

- Estimate the model

$$wage_i = \beta_0 + \delta_0 female_i + u_i$$

- Test $H_0 : \delta_0 = 0$ vs. $H_1 : \delta_0 < 0$ using a *t*-test

Rks:

- This is the test for equality of two-sample means (assuming constant equal variances) from 2^{nd} yr stats
- Seems dumb as a test for *discrimination* but it gets interpreted that way all the time!
- OLS always estimates something: In this case, it's the BLP of wages given gender. Is that what we want?

Ex 2. Test if women's wages are lower because of discrimination.

- This is a good example of how hard it is to translate something we care about into a restriction on some parameters!
- The conceptual experiment involves changing gender, keeping everything else constant. To an excellent approximation, this experiment isn't feasible. (We do have some studies of the effect of "gender blind" applications on successfully landing the job)

- We need a model of what should determine wages, in the absence of discrimination. Let's say it is productivity and try to capture it in the specification

$$y_i = \beta_0 + \delta_0 female_i + \beta_1 educ_i + \beta_2 esper_i + \beta_3 tenure_i + u_i$$

- Test $H_0 : \delta_0 = 0$ vs. $H_1 : \delta_0 < 0$ using a *t*-test

Rks:

- This seems better as a way of controlling for differences that should matter, but it's not perfect.
- We don't have to worry about simultaneity bias–shocks that lead to higher wages will not cause people to change their sex. But we still can have correlation between regressors and the disturbance.

- What’s not measured that could be correlated with the female dummy?
  - Type of job (or industry). Would putting in industry dummies control for taste preferences (women prefer jobs with certain characteristics such as flexibility or not involving heavy physical labour)? Or would it amount to "over controlling" (discrimination manifests itself in creating some industries that are female ghettos)?
  - Commitment to the labour force. Upon graduation, salaries of new hires are indistinguishable by gender; then a gap opens up. Why? Is it a difference in investment on the job? Women are more likely to interrupt their careers for family responsibilities. Is this taste (biology) or response to market opportunities?

- Nonrandom sample. We only see women's wages if they work. Employment rates of women have been lower, historically, than those of men. Does this matter? Women who remain unmarried have higher wages. Is this a reflection of higher commitment to work? Do they get more on-the-job investment by displaying this commitment? (Men who remain unmarried earn less!)
- Measurement error. Women miss more days at work. Moretti (2005) estimates that days lost due to menstrual pain alone can account for almost half of the wage gap in Italian banks!

- Why focus on δ_0? Is the return to education, experience, or tenure different for men and women?

:Multiple categories

- A dummy variable allows us to divide the sample into two different groups. For example, $female_i$ allows us to divide the sample into females and "not" females.
- With two dummy variables, we can divide the sample into *four* groups. For example, $\{female_i, married_i\}$ yields
  1. female and married
  2. female and not married
  3. not female and married
  4. not female and not married
- With m dummy variables, we have 2^m different groups. If we gave each group their own regression coefficients, we would have $2^m \cdot K$ regression coefficients. This gets large very fast. We often posit that some of the responses are constant across groups as a way of imposing parsimony on the model.

- For example, we may specify

$$wage_i = \beta_0(i) + \beta_1 educ_i + \beta_2(i) esper_i + u_i$$

with

$$\beta_0(i) = \beta_0 + \delta_0 female_i * married_i +$$
$$\delta_1 female_i * simgle_i + \delta_2 male_i * simgle_i$$

$$\beta_2(i) = \beta_2 + \beta_3 * esper_i$$

This gives each of the four groups defined by $\{female_i, married_i\}$ their own intercept, but a common linear response to $educ_i$ and a common quadratic response to $esper_i$

- Q: What measures the difference in the expected wage of married and unmarried women?
- Q: What measures the difference in the expected wage of married and unmarried men?

Ex 7.6

$$\widehat{\ln wage} = \underset{(.100)}{.321} + \underset{(.055)}{.213}\,marrmale - \underset{(.058)}{.198}\,marrfem$$

$$- \underset{(.056)}{.110}\,simgfem + \underset{(.007)}{.079}\,educ + \underset{(.005)}{.027}\,esper$$

$$- \underset{(.00011)}{.00054}\,esper^2 + \underset{(.007)}{.029}\,tenure - \underset{(.00023)}{.00053}\,tenure^2$$

$$n = 526 \qquad R^2 = .461$$

Rk: $marrmale = married * male$, etc *educ*=yrs schooling, *tenure*=yrs on the job

- What's the coefficient on married male mean?
- What happens to your wage if you spend one more year on the job?
  $(.027 - .00054(2 * esper + 1) + .029 - .00053(2 * tenur + 1))$

:Ordinal information

Suppose we have three credit ratings:
Poor < *OK* < *Excellent*

- We could construct a variable

$$CR = \begin{cases} 0 \text{ if } Poor \\ 1 \text{ if } OK \\ 2 \text{ if } Excellent \end{cases}$$

and regress

$$y = \beta_0 + \beta_1 CR + \text{other factors}$$

But this says that response of going from *Poor* to *OK* exactly equals the response of going from *OK* to *Excellent*

- A better approach is to construct a family of dummy variables for each category

$$CR_1 = 1 \text{ if } CR = 1, \ 0 \text{ otherwise}$$

$$CR_2 = 1 \text{ if } CR = 2, \ 0 \text{ otherwise}$$

and regress

$$y = \beta_0 + \beta_1 CR_1 + \beta_2 CR_2 + \text{other factors}$$

We can always test if $2\beta_1 = \beta_2$

- To save space, data sets often give what look like ordinal values to data that aren't ordinal. For example,

$$\begin{aligned} PROV &= 1 \text{ if NFLD} \\ &= 2 \text{ if PEI} \\ &\vdots \\ &= 10 \text{ if BC} \end{aligned}$$

It doesn't makes ANY sense to run the regression

$$y = \beta_0 + \beta_1 PROV + \text{ other factors}$$

You should construct a family of dummy variables in the same way as described above for ordinal variables.

:Pooling many groups

- Suppose we have G groups in our sample, and for each of them we posit the regression model

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \cdots + \beta_{g,k}x_k + u \quad g = 1..G$$

- We can create $G - 1$ dummy variables D_g to indicate membership in group $g = 2..G$ and then combine all the regression models into a single one using

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \sum_{g=2}^{G} \left(\tilde{\beta}_{g,0} D_g + \tilde{\beta}_{g,1}(x_1 D_g) + \cdots + \tilde{\beta}_{g,k}(x_k D_g) \right) + u$$

where $x_1 D_g$ is a regressor formed by multiplying x_1 by D_g, etc

- The OLS estimates from the regression with all the data allow us to recover the OLS estimates for each of the group regressions.
- As long as the disturbances in each group have the same variance, we can use our usual test statistics for the general linear hypothesis to test

$$H_0 : \beta_{g,0} = \beta_0, \beta_{g,1} = \beta_1, \ldots, \beta_{g,k} = \beta_k \quad g = 1..G$$

- Often, we will choose to impose some of the restrictions as maintained hypotheses and test the remaining. For example, maintaining $\beta_{g,2} = \beta_2, \ldots, \beta_{g,k} = \beta_k \quad g = 1..G$, test

$$H_0 : \beta_{g,0} = \beta_0, \beta_{g,1} = \beta_1 \qquad g = 1..G$$

(Again, we already know how to impose the restrictions and to test a linear hypothesis)

:Using dummy variables for functional form

We can construct piece-wise linear regression functions using dummy variables.

- Suppose we believe the regression function is piecewise linear

$$y = \begin{cases} \beta_0 + \beta_1 x + u & x \leq x^* \\ \gamma_0 + \gamma_1 x + u & x > x^* \end{cases}$$

- Create a dummy variable

$$D = \begin{cases} 0 & x \leq x^* \\ 1 & \text{otherwise} \end{cases}$$

- We can write the regression function compactly as

$$y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 (Dx) + u$$

where $\beta_2 = \gamma_0 - \beta_0$ and $\beta_3 = \gamma_1 - \beta_1$. Notice that the function is *linear in parameters*, so we can use OLS.

- We can impose continuity at the "knot" x^* by restricting the parameters so that

$$\beta_0 + \beta_1 x^* = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x^*$$

- For more flexibility, we can generalize to having many knots, $\{x_1^*, x_2^*, \cdots x_m^*\}$
- For more flexibility, we can replace piecewise linear by piecewise cubic functions (cubic splines). This allows us to have functions in C^2 (continuous first and second derivatives).

- We can generalize to the case of many regressors, and write
$$y = X_1\beta_1 + f(X_2, \beta_2) + u$$
where $f(X_2, \beta_2)$ is a piecewise linear (or cubic) function.
- Suppose we an extreme version of the piecewise linear representation is the true model
$$y_i = \beta(x_i)x_i + u_i$$
where $\beta(x_i) \in \mathbb{R}$, and $E(u_i|X) = 0$. But you estimate
$$y_i = \gamma x_i + u_i$$
What's $E(\hat{\gamma}|X)$?
- Answer: $E(\hat{\gamma}|X) = \sum_i \omega_i \beta(x_i)$ where $\omega_i = x_i^2 / \sum x_i^2$.
- If $x_i \in \{x_1^*, x_2^*, \cdots x_M^*\}$ and p_m denotes $P(x_i = x_m^*)$, then we can rewrite the expression above as
$E(\hat{\gamma}|X) = \sum_m \omega_m \beta(x_m) p_m$

:Applications of the Frisch-Waugh Theorem with Dummy variables

- Recall that if the model is

$$y = X_1\beta_1 + X_2\beta_2 + u$$

then the FW theorem says

$$\widehat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2y$$

where $M_2 = I - X_2(X_2'X_2)^{-1}X_2$ is the matrix that gives the orthogonal projection onto the space orthogonal to $Sp(X_2)$.

- This means that we can compute $\widehat{\beta}_1$ from the regression of M_2y (or just y) on M_2X_1. (Using M_2y as the dependent variable yields exactly the same residual vector as the regression of y on X_1 and X_2).

- A leading example occurs if $X_2 = \iota$ (a vector of ones, i.e. the intercept). In this case, we see that the slope estimates, $\widehat{\beta}_1$, are obtained by regressing the dependent variable, expressed as a deviation from its mean, on the explanatory variables, also expressed as a deviation from mean, i.e. running the regression

$$\widetilde{y} = \widetilde{X}_1 \beta_1 + \widetilde{u}$$

where $\widetilde{y} = y - \overline{y}\iota$, and the columns of $\widetilde{X}_1$ are defined accordingly.

- Many other examples can be constructed using dumming variables. For example, suppose we write $X_2 = D$, where D denotes a matrix of dummy variables. For convenience, let's parameterize so that $\iota \in Sp(D)$, and write M_D rather than M_2.

- Suppose D includes two dummy variables to indicate membership in one of two mutually exclusive and exhaustive groups, say, sex ($D_{1i} = 1$ for females and $D_{2i} = 1$ for males).

$$(D'D) = \begin{pmatrix} D_1' \\ D_2' \end{pmatrix} \quad (D_1 D_2) = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}$$

where n_1 = number of females in the sample, and n_2 = number of males in the sample. With a bit of work, we see that

$$(D'D)^{-1}D'y = \begin{pmatrix} \overline{y}_F \\ \overline{y}_M \end{pmatrix}$$

where $\overline{y}_F$ is the mean value of y for females, and $\overline{y}_M$ is the mean value for men.

- We conclude that $\tilde{y} = M_D y$ is the dependent variable measured as a deviation from the gender mean, that is, the component

$$\tilde{y}_i = y_i - D_{1i}\overline{y}_F - D_{2i}\overline{y}_M$$

  The columns of the matrix $M_D X_1$ will have a similar interpretation. So only the deviations from gender specific means of the explanatory variables are used to estimate β_1, the response of y to X_1.

- Exercise: Show that if group 1 has only one observation, then running the regression with the two dummies gives the same coefficient on X_1 as just dropping that observation, and the coefficient on D_2 is the intercept from this "leave one out" regression.

- In time series applications, we often want to take account of "seasonal" effects (hour of the day, day of the week, month of the year, quarter of the year). Suppose we have S "seasons" and construct dummy variables $D_{si} = 1$ if obs i occurs in season s (0 otherwise). Proceeding as above, we see that

$$\tilde{y}_i = y_i - D_{1i}\overline{y}_1 - D_{2i}\overline{y}_2 - \cdots - D_{Si}\overline{y}_S$$

where $\overline{y}_s$ denotes the mean of y in season s. The columns of the matrix $M_D X_1$ will have a similar interpretation. So only the deviations from season specific means of the explanatory variables are used to estimate β_1

- (Panel Data Fixed Effects) Suppose we have T observations ($T \geq 2$) on the same person at different points in time. We could construct individual specific dummies, i.e. $D_{mi} = 1$ if obs i pertains to person $m = 1..M$. (M can be very large).
- Proceeding as above, we see that we can write

$$\tilde{y}_i = y_i - \sum_m D_{mi} \overline{y}_m$$

  so each person's observations on the dependent variable are expressed as deviations from that particular person's mean value.
- Similarly, the explanatory variables are written as deviations from that particular person's mean values. So it is only variation within individuals explanatory variables over time that can be used to estimate β_1; variation across individuals is ignored.

- So for example, in estimating the effect of marital status on wages, only those people that change marital status affect the estimated coefficient. Data on individuals who stay single or stay married within the sample is ignored.
- Notice that with individual specific dummies we cannot estimate the response to explanatory variables that DON'T vary over time for individuals (sex, ethnicity, father's education, etc.).

:Dummy Dependent variable

- LOTS of interesting economic outcomes are discrete (working/not working etc.). Suppose y is a dummy variable.

$$E(y|x) = 1 \cdot P(y = 1|x) + 0 \cdot P(y = 0|x)$$
$$= P(y = 1|x)$$

- In general, this must be a nonlinear regression function because

$$0 \leq P(y = 1|x) \leq 1$$

- Nonlinear models such as probit or tobit deal give the correct support for $E(y|x)$, but they are difficult to modify for other problems, such as correlation between regressors and the disturbance, or serial correlation in the disturbance.

- An approximation is to use the BLP rather than the conditional expectation, which in this case is called the Linear Probability Model (LPM)

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_k x_k + u$$

Notice that it makes no sense to assume $V(u_i|X) = \sigma^2$ since $V(y_i|X) = P(y = 1|X) - P^2(y = 1|X)$, so we'll need to develop new estimators for the covariance matrix of $\widehat{\beta}$ and new test statistics for the general linear hypothesis. We do that in the next lecture.

:Sample Selection

- Often our data are not random draws, but we only get to see the data if some condition is satisfied.
- For example, we may observe the explanatory variables for the i^{th} observation only if $x_i \in X^*$ (eg. very high income individuals may be missing from our sample of consumption decisions). Fortunately, changes in the marginal distribution of the regressors have no effect on our estimators and test statistics under the CLM (but see the example above where the regression coefficients vary with the regressor).
- If the sample we see depends on the properties of the dependent variable, then things are very different. For example, suppose we posit a wage regression for women of the form

$$\ln wage = \beta_0 + \beta_1 educ + \beta_2 esper + u$$

But we only observe the wages of working women. Let y be a dummy variable that equals 1 if the women works. Then the regression function we estimate from our sample is

$$E(\ln wage | X, y = 1) = X\beta + E(u | X, y = 1)$$

So there will be a bias.

- Examples of *self-selection* bias like this are numerous. See Wooldridge for examples related to *program evaluation*.