PSC 508

Jim Battista

Univ. at Buffalo, SUNY

Dummies

- Sometimes we want to include *categorical* variables in our models
- Numerical variables that don't necessarily have any inherent order and that just describe different categories
- Easy example: respondent sex in individual models

A simple dummy

- A simple dummy variable is just a variable that takes only one of two possible values – zero or one
- We can code even a simple dummy variable in more than one way
- Respondent sex for example
    - "Male" variable – 1 if man, 0 if woman
    - "Female" variable – if if woman, 0 if man
    - These will say the same thing and fulfill the same role in the regression

A simple example

```
. reg bushft male

      Source |       SS       df       MS              Number of obs =    1181
-------------+------------------------------           F(  1,  1179) =    2.24
       Model |  2505.56377     1  2505.56377           Prob > F      =  0.1350
    Residual |  1320313.98  1179  1119.85918           R-squared     =  0.0019
-------------+------------------------------           Adj R-squared =  0.0010
       Total |  1322819.54  1180  1121.03351           Root MSE      =  33.464

------------------------------------------------------------------------------
      bushft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   2.919655   1.951914     1.50   0.135    -.9099563    6.749267
       _cons |   53.92063    1.33325    40.44   0.000     51.30483    56.53644
------------------------------------------------------------------------------
```

- The coefficient on "male" means what a coefficient always does
- But because it can only go from zero to one, it says that men like Bush 2.9 points more than women

A more complex example

```
. reg bushft male lib_con partyid age education

      Source |       SS       df       MS              Number of obs =     896
-------------+------------------------------           F(  5,   890) =  229.93
       Model |  587060.669     5  117412.134           Prob > F      =  0.0000
    Residual |   454479.72   890  510.651371           R-squared     =  0.5636
-------------+------------------------------           Adj R-squared =  0.5612
       Total |  1041540.39   895  1163.73228           Root MSE      =  22.598

------------------------------------------------------------------------------
      bushft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |  -.6810617    1.51743    -0.45   0.654    -3.65922    2.297097
    lib_cons |   4.380368   .6516295     6.72   0.000    3.101458    5.659278
     partyid |   9.791417   .4415604    22.17   0.000    8.924796    10.65804
         age |   .0996235    .045958     2.17   0.030    .0094248    .1898223
   education |  -1.937508   .4721056    -4.10   0.000   -2.864078   -1.010938
       _cons |   11.97742   3.772359     3.18   0.002    4.573665    19.38118
------------------------------------------------------------------------------
```

- All else equal, men like Bush 0.68 points less than women do.

A more complex example

```
. reg bushft female lib_con partyid age education

      Source |       SS       df       MS              Number of obs =     896
-------------+------------------------------           F(  5,   890) =  229.93
       Model |  587060.669     5  117412.134           Prob > F      =  0.0000
    Residual |   454479.72   890  510.651371           R-squared     =  0.5636
-------------+------------------------------           Adj R-squared =  0.5612
       Total |  1041540.39   895  1163.73228           Root MSE      =  22.598

------------------------------------------------------------------------------
      bushft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   .6810617    1.51743     0.45   0.654    -2.297097     3.65922
    lib_cons |   4.380368   .6516295     6.72   0.000     3.101458    5.659278
     partyid |   9.791417   .4415604    22.17   0.000     8.924796    10.65804
         age |   .0996235    .045958     2.17   0.030     .0094248    .1898223
   education |  -1.937508   .4721056    -4.10   0.000    -2.864078   -1.010938
       _cons |   11.29636   3.824392     2.95   0.003     3.790482    18.80224
------------------------------------------------------------------------------
```

- All else equal, women like Bush 0.68 points more than men do.

Other uses for dummies

- Sometimes a variable might be coded in ways that don't make sense for your use
- Say education in the NES, when you have a theory about college graduates
- Can turn that variable into a dummy variable taking 1 if the respondent finished college and 0 otherwise
- LET'S DO THAT!

Creating a simple dummy variable

- In Stata:
  `generate dummy=variable==value if variable!=0`
- In R: `dummy<-1*(variable==value)`
- Both of these forms *should* preserve missing data as missing
- In either, you can substitute other expressions for `variable==value`
  - Stata:
    `generate college=education>5 if education !=.`
  - R: `college<-1*(education>5)`
  - Stata:
    `gen dummy=variable>4 & variable<8 if variable!=.`
  - R: `dummy<-1*(variable>4 & variable<8)`

Multiple categories

- Some variables have multiple categories in them
- Race and ethnicity for example – respondent can be any of several races
- Region – respondent or state can be from any of several regions
- Usual tactic:
  - Convert categorical variable with N categories into N-1 dummies
  - Why N-1? OLS will explode if one IV is a perfect linear combination of other IVs
  - . . . and including all the categories would make that happen

Except. . .

- We usually create all the dummies – we just exclude one from the regression
- That way we can easily change the reference category later
- Let's generate a set of "race" dummies in the NES

Multiple categories

- N-1 categories is the same thing that we did for single dummies
  - We didn't include one dummy for men and another for women
- Omitted category is the reference category
- Other categories are relative to it
- So if we omit the southeast region, the coefficient on the dummy variable for the Pacific northwest tells us the difference between the Pacific Northwest *and the southeast*
- If we omitted New England instead, the coefficient on PacNW would be the difference between the Pacific northwest and New England instead
- No one right way to organize these or choose a reference category
- Choose one that helps you tell your analytical story

Except. . .

- Changing the reference category is easy
    - Just add the reference category in, and remove another category
- Let's try this with race in the NES

Coding multiple categories in Stata

- Remember that the goal is to create a set of dummies that capture whatever we're interested in from the source categorical variable
- We need to preserve "missing-ness" in all the dummies that represent our source variable
- Say we want to code race as in the NES
  - `gen black=race==10 if race!=.`
  - `gen asian=race==20 if race!=.`
  - `gen nativeamerican=race==30 if race!=.`
  - `gen latino=race==40 if race!=.`
  - `gen anglo=race==50 if race!=.`

Coding multiple categories in Stata

- Another example: coding education into dummies for
  1. Didn't finish high school
  2. Finished high school, doesn't have BA
  3. Has BA or more
- Stata code:
  - `gen nohsdiploma=education<3 if education!=.`
  - `gen diploma_no_ba=education>2 & education<6 if education!=.`
  - `gen ba_or_more=education>5 if education!=.`

Coding multiple categories in R

- First example
  - `black<-1*(race==10)`
  - `asian<-1*(race==20)`
  - `nativeamerican<-1*(race==30)`
  - `latino<-1*(race==40)`
  - `anglo<-1*(race==50)`
- Second example
  - `no.hs.diploma<-1*(education<3)`
  - `diploma.no.ba<-1*(education>2 & education<6)`
  - `ba.or.more<-1*(education>5)`

Factors in R

- Another way to code multiple-category variables in R is as a "factor"
  - If coding race in the NES, try:
  - `racefactor<-factor(race)`
  - Automatically sets first category as reference/omitted category
- To choose whites as the reference category, change "contrasts"
  - `contrasts(racefactor)<-contr.treatment(5,base=5)`
  - More generally:
    - `contrasts(variable)<-contr.treatment(NumberOfCategories,base=DesiredCategory)`
    - Note that it wants the category number from 1 to N, not the value in the variable (5, not 50)

A complication

- Say we have a set of dummies for race and ethnicity and none are statistically significant
- Does that mean that race doesn't matter? That race isn't statistically significant?
    - Not necessarily – remember that we have one theoretical variable spanning multiple dummies in the regression
    - Possible that we may have chosen a reference category that masks real differences

Joint F tests

- To perform hypothesis tests on a single theoretical variable with multiple dummies, use a joint F test item Say we want to test whether the three dummies `dummy1, dummy2, dummy3` that code the source variable `sourcevariable` are jointly statistically significant
  - In Stata: run regression, then `test dummy1 dummy2 dummy3`
  - In R, it's more complex – have to do with anova
    1. Run model without the categorical variable
    2. `model1<-lm(dv ~ iv1+iv2,subset=!is.na(sourcevariable))`
    3. The rigamarole at the end ensures that we run the model for only those observations where our dummies aren't missing
    4. Run again with the dummy variables –
    5. `model2<-lm(dv1 ~ iv1+iv2+dummy1+dummy2+dummy3)`
    6. `anova(model1,model2)`
  - Let's try this!