Vanderbilt University
Leadership, Policy and Organizations
Class Number 9952
Spring 2018

Binary and Categorical Variables

Binary and categorical variables can be a headache to work with. It's worth taking some time to think about each step with these kinds of variables in order to make sure that they are being reported effectively.

Coding

First, it's worth thinking pretty carefully about how these variables will be coded. Are you sure that they are mutually exclusive and exhaustive? How about the numbers of categories? Are these appropriate for the task at hand? Are they really categorical or can they be thought of as ordered? How would you figure this out?

In general, it's better to favor fewer categories, but you need to make sure that your decisions reflect the important questions in your theoretical framework.

Below, I recode the race variables as they're constructed by NCES to be more useful in our analysis.

```
. recode byrace (4/5=4) (6=5) (7=6) (.=.), gen(byrace2)
(10633 differences between byrace and byrace2)
.
. label define byrace2 1 "Am.Ind." 2 "Asian/PI" 3 "Black" 4 "Hispanic" 5 "Multiraci
> al" 6 "White"
.
. label values byrace2 byrace
```

Binary Variables

Binary variables must always be constructed to be directional. Never have a binary variable for "sex," always construct this kind of binary variable as either "male" or "female." Binary variables in a regression represent an intercept shift– for the group in question, they increase or decrease the intercept by that amount.

```
. gen female=bysex==2
. replace female=. if bysex==.
(819 real changes made, 819 to missing)
.
. lab var female "Female"
```

Categorical Variables

When running a model with categorical variables, Stata won't always know what you're talking about. If the underlying variable is numeric, it will simply include that variable as numeric. This is not good. Instead, we need to use the `i.` formulation, which specifies not only that a given variable is to be understood as a factor variable, but also allows the user some fine-grained control over how this will be constructed.

Remember that categorical variables must always be interpreted relative to their reference category. We cover how to think about that next.

```
. // NOPE!
. eststo order1: svy: reg `y´ order_plan
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       361                  Number of obs      =     15129
Number of PSUs     =       751                  Population size    = 3055917.9
                                                Design df          =       390
                                                F(   1,    390)    =   1025.42
                                                Prob > F           =    0.0000
                                                R-squared          =    0.1261

------------------------------------------------------------------------------
             |             Linearized
    bynels2m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  order_plan |     .07355   .0022968    32.02   0.000     .0690342    .0780657
       _cons |   .2704247   .0059146    45.72   0.000     .2587962    .2820531
------------------------------------------------------------------------------

.
.
. //Proper factor notation
. eststo order1: svy: reg `y´ i.order_plan byses1 female
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       361                  Number of obs      =     14561
Number of PSUs     =       751                  Population size    =   2908622
                                                Design df          =       390
                                                F(   4,    387)    =    647.17
                                                Prob > F           =    0.0000
                                                R-squared          =    0.2507

------------------------------------------------------------------------------
             |             Linearized
    bynels2m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  order_plan |
  Votech/CC  |   .0174877   .0053254     3.28   0.001     .0070176    .0279577
  Four Year  |   .0899849   .0054533    16.50   0.000     .0792633    .1007065
             |
      byses1 |   .0629814   .0020981    30.02   0.000     .0588565    .0671064
      female |  -.0208619   .0026488    -7.88   0.000    -.0260696   -.0156542
       _cons |   .4048229   .0051525    78.57   0.000     .3946927     .414953
------------------------------------------------------------------------------

.
. esttab order1 using order1.rtf,  varwidth(50) label  ///
>                nodepvars             ///
>                    b(3)                  ///
>                 se(3)                    ///
>                r2 (2)                   ///
```

Table 1: Results of OLS, Dependent Variable= Math Scores

	(1)
Plans, Reference= No Plans/ Don’t Know	
—Votech/CC	1.749**
	(0.533)
—Four Year	8.998***
	(0.545)
SES	6.298***
	(0.210)
Female	-2.086***
	(0.265)
Constant	40.482***
	(0.515)
Observations	14561
R^2	0.25
Adjusted R^2	
F	647.17
DF model	4
DF residual	390

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
>               ar2 (2)                   ///
>               scalar(F  "df_m DF model"  "df_r DF residual" N)   ///
>               sfmt (2 0 0 0)               ///
>               replace
(output written to order1.rtf)

.
.
. esttab order1 using order1.rtf,  varwidth(50) label  ///
>     refcat(2.order_plan "Plans, Reference= No Plans/ Don´t Know",nolabel) ///
>         nobaselevels ///
>               nomtitles ///
>               nodepvars              ///
>                b(3)                   ///
>                se(3)                     ///
>               r2 (2)                    ///
>               ar2 (2)                   ///
>               scalar(F  "df_m DF model"  "df_r DF residual" N)   ///
>               sfmt (2 0 0 0)               ///
>               replace
(output written to order1.rtf)

.
```

This gives us a properly formatted table, like so

Quick Exercise

Run the above regression, but use parental education as a predictor. Create a properly formatted table with parental education as a categorical variable.

Reference Categories for Categorical Variables

It's important to put some thought into reference categories for category variables. If you have no other preference, then use the largest group. You can accomplish this via the `ib(freq).` command. You should put some careful thought into the contrasts you'd like to draw–which groups do you want to compare and why?

```
.
. //Proper factor notation: setting base levels
. eststo order2: svy: reg `y´ ib(freq).order_plan byses1 female
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       361                  Number of obs      =     14561
Number of PSUs     =       751                  Population size    =   2908622
                                                Design df          =       390
                                                F(   4,    387)    =    647.17
                                                Prob > F           =    0.0000
                                                R-squared          =    0.2507

------------------------------------------------------------------------------
             |             Linearized
    bynels2m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  order_plan |
No Plans/DK  |  -.0899849   .0054533   -16.50   0.000    -.1007065   -.0792633
  Votech/CC  |  -.0724972   .0028019   -25.87   0.000    -.0780059   -.0669885
             |
      byses1 |   .0629814   .0020981    30.02   0.000     .0588565    .0671064
      female |  -.0208619   .0026488    -7.88   0.000    -.0260696   -.0156542
       _cons |   .4948077   .0028725   172.25   0.000     .4891601    .5004553
------------------------------------------------------------------------------

.
. esttab order2 using order2.rtf,  varwidth(50) label  ///
>               nodepvars              ///
>                   b(3)                   ///
>                se(3)                     ///
>               r2 (2)                    ///
>               ar2 (2)                   ///
>               scalar(F  "df_m DF model"  "df_r DF residual" N)   ///
>               sfmt (2 0 0 0)               ///
>               replace
(output written to order2.rtf)

.
.
. esttab order2 using order2.rtf,  varwidth(50)   ///
>     refcat(1.order_plan "College Plans, Reference=Plans to go to College",nolabel
> ) ///
>         label ///
>                    nomtitles ///
>                        nobaselevels ///
>               nodepvars              ///
>                b(3)                  ///
```

Table 2: Results of OLS, Dependent Variable= Math Scores

	(1)
College Plans, Reference=Plans to go to College	
—No Plans/DK	-8.998***
	(0.545)
—Votech/CC	-7.250***
	(0.280)
SES	6.298***
	(0.210)
Female	-2.086***
	(0.265)
Constant	49.481***
	(0.287)
Observations	14561
R^2	0.25
Adjusted R^2	
F	647.17
DF model	4
DF residual	390

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
>                se(3)                    ///
>               r2 (2)                   ///
>               ar2 (2)                  ///
>               scalar(F  "df_m DF model"  "df_r DF residual" N)   ///
>               sfmt (2 0 0 0)               ///
>               replace
(output written to order2.rtf)

.
```

Quick Exercise

Run the regression above, but include parental education. This time, output the results with some college as the reference category for parental education.

Interactions

When interacting a binary variable with a categorical variable, you must do the FULL interaction–you can't just interact with one level. Same thing applies to continuous variables.

```
. // Factor notation, interaction
.
. //Proper factor notation: setting base levels
. eststo order3: svy: reg `y´ b3.order_plan##i.female byses1
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       361                  Number of obs     =     14,561
Number of PSUs     =       751                  Population size   =  2,908,622
                                                Design df         =        390
                                                F(   6,    385)   =     436.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2508

-----------------------------------------------------------------------------------
                  |             Linearized
         bynels2m |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
       order_plan |
  ---No Plans/DK  |  -8.525456   .6892792   -12.37   0.000    -9.880624   -7.170288
    ---Votech/CC  |  -7.123964   .3838901   -18.56   0.000    -7.878717   -6.369211
                  |
         1.female |  -1.908111   .3136299    -6.08   0.000    -2.524728   -1.291494
                  |
order_plan#female |
---No Plans/DK#1  |  -1.408776   1.010942    -1.39   0.164    -3.396355    .5788024
  ---Votech/CC#1  |  -.2453994   .5164566    -0.48   0.635    -1.260787    .7699881
                  |
           byses1 |   6.292627   .2098255    29.99   0.000     5.880097    6.705158
            _cons |   49.38647   .2962783   166.69   0.000     48.80397    49.96898
-----------------------------------------------------------------------------------

.
. esttab order3 using order3.`ttype´, varwidth(50) ///
>     refcat(1.order_plan "College Plans, Reference=Plans to go to College:" 1.order_plan#1.female "In
> teraction of Plans with Female:", nolabel) ///
>  interaction(" X ") ///
>    label ///
>                    nomtitles ///
>                        nobaselevels ///
>                nodepvars              ///
>                 b(3)                   ///
>                 se(3)                     ///
>                r2 (2)                    ///
>                ar2 (2)                   ///
>                scalar(F  "df_m DF model"  "df_r DF residual" N)   ///
>                sfmt (2 0 0 0)               ///
>                replace
(output written to order3.tex)
```

	(1)
College Plans, Reference=Plans to go to College:	
—No Plans/DK	-8.525***
	(0.689)
—Votech/CC	-7.124***
	(0.384)
Female=1	-1.908***
	(0.314)
Interaction of Plans with Female:	
—No Plans/DK X Female=1	-1.409
	(1.011)
—Votech/CC X Female=1	-0.245
	(0.516)
SES	6.293***
	(0.210)
Constant	49.386***
	(0.296)
Observations	14561
R^2	0.25
Adjusted R^2	
F	436.30
DF model	6
DF residual	390

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Using Margins

Once you're undertaking interactions with categorical variables, it's generally a good idea to interpret them using the margins command. In the below code I use margins to interpret the interaction between a categorical and a binary variable and to make a table with confidence intervals from the output.

```
. // Margins to figure out what´s going on
. margins, predict(xb) at((mean) byses1 order_plan=(1 2 3) female=(0 1)) post

Adjusted predictions                              Number of obs      =     13055
Model VCE    : Linearized

Expression   : Linear prediction, predict(xb)

1._at        : order_plan      =           1
               female          =           0
               byses1          =    .0400221 (mean)

2._at        : order_plan      =           1
               female          =           1
               byses1          =    .0400221 (mean)

3._at        : order_plan      =           2
               female          =           0
               byses1          =    .0400221 (mean)

4._at        : order_plan      =           2
               female          =           1
               byses1          =    .0400221 (mean)

5._at        : order_plan      =           3
               female          =           0
               byses1          =    .0400221 (mean)

6._at        : order_plan      =           3
               female          =           1
               byses1          =    .0400221 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .4111286   .0063766    64.47   0.000     .3985917    .4236655
          2  |   .3779598    .007331    51.56   0.000     .3635465     .392373
          3  |   .4251435   .0034677   122.60   0.000     .4183258    .4319613
          4  |   .4036084   .0034357   117.47   0.000     .3968536    .4103633
          5  |   .4963832   .0029395   168.86   0.000     .4906039    .5021625
          6  |   .4773021   .0025748   185.38   0.000     .4722399    .4823643
------------------------------------------------------------------------------

.
. esttab . using margins.rtf , margin label nostar ci ///
>     varlabels(1._at "No College Plans, Male" ///
>                   2._at "No College Plans, Female" ///
>                       3._at "Vo-Tech/Community College, Male" ///
>                           4._at "Vo-Tech/Community College, Female" ///
>                               5._at "Four-Year College Plans, Male" ///
>                                   6._at "Four-Year College Plans, Female" ) ///
>         replace
(output written to margins.rtf)
```

Table 3: Predicted Math Scores by College Plans and Sex

	(1)
No College Plans, Male	41.11 [39.86,42.37]
No College Plans, Female	37.80 [36.35,39.24]
Vo-Tech/Community College, Male	42.51 [41.83,43.20]
Vo-Tech/Community College, Female	40.36 [39.69,41.04]
Four-Year College Plans, Male	49.64 [49.06,50.22]
Four-Year College Plans, Female	47.73 [47.22,48.24]
Observations	13055

Marginal effects; 95% confidence intervals in brackets

(d) for discrete change of dummy variable from 0 to 1

Quick Exercise

Again include parental education, and generate predicted probabilities using the margins command. Then go back and choose a different reference category. Does a different reference category result in different predicted probablities?