Using Factor Variables in Stata

Thomas Elliott

January 16, 2014

Running regressions on categorical independent variables requires the creation of dummy variables. Recent versions of Stata have included a new way of defining and using categorical variables in regression analysis. For the help page on how to use this method, type `help fvvarlist` in the command window.

There are four basic factor operators in Stata:

Operator	Description
`i.`	specify a categorical variable (create indicators for categories)
`c.`	specify a variable as continuous (when specifying interaction terms)
`#`	specify an interaction between variables
`##`	specify factorial interaction between variables (all interactions and main effects)

1 Categorical Variables

I'll be using some data from the General Social Survey (GSS). To begin, let's say we want to regress `EDUC`, the highest year of education completed, on `POLVIEWS`, a seven category measure of political views. In GSS, `POLVIEWS` is a numeric variable, with each number representing a possible categorical answer. By default, Stata will treat `POLVIEWS` as a continuous variable, which would be a mistake.

```
. regress EDUC POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    POLVIEWS |  -.1248473   .0183561    -6.80   0.000    -.1608279   -.0888668
       _cons |   13.97524     .07997   174.76   0.000     13.81849    14.13199
------------------------------------------------------------------------------
```

Instead, we can use a factor operator to tell Stata that `POLVIEWS` is a categorical variable and to create indicators. We do this by prepending the factor variable with `i.`

```
. regress EDUC i.POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    POLVIEWS |
          2  |   .2452808   .1560552     1.57   0.116    -.0606091    .5511707
          3  |   .0027932   .1563809     0.02   0.986    -.3037352    .3093215
          4  |  -.8235205   .1428313    -5.77   0.000     -1.10349   -.5435512
          5  |  -.1274898   .1525671    -0.84   0.403    -.4265427    .1715632
          6  |  -.3355454   .1511146    -2.22   0.026    -.6317512   -.0393396
          7  |  -1.036269   .1918869    -5.40   0.000    -1.412394    -.660144
             |
       _cons |   13.86157   .1366483   101.44   0.000     13.59372    14.12942
------------------------------------------------------------------------------
```

By prepending `i.`, we told Stata to create indicators for each category of `POLVIEWS`. Notice that the the first category (`POLVIEWS == 1`) is left out. Stata by default will use the first value as the reference category. What if we wanted to use 4 (Moderate) instead? There are two ways to do this. The first is to specify it with the `i.` prefix, by adding `b#` where # is the value you want to use for the base. In our example, we wanted to use 4 for moderate, so we would prepend `POLVIEWS` with `ib4`.

```
. regress EDUC ib4.POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    POLVIEWS |
          1  |   .8235205   .1428313     5.77   0.000     .5435512     1.10349
          2  |   1.068801   .0860726    12.42   0.000      .900087    1.237516
          3  |   .8263137   .0866617     9.53   0.000     .6564446    .9961828
          5  |   .6960308   .0795739     8.75   0.000     .5400547    .8520068
          6  |   .4879752   .0767523     6.36   0.000       .33753    .6384204
          7  |  -.2127485   .1409817    -1.51   0.131    -.4890924    .0635954
             |
       _cons |   13.03805   .0415696   313.64   0.000     12.95657    13.11953
------------------------------------------------------------------------------
```

Notice that 4 is the new reference category. Using this method, we have to specify the reference category we want each time. But let's say we are doing a whole series of analyses in which moderate will be our reference category? Instead of specifying the base category each time, we can set it permanently with the `fvset base` command. The syntax for the command is:

`fvset base #` *`var`*

Where # is the value for the reference category and *var* is the factor variable.

```
. fvset base 4 POLVIEWS

. regress EDUC i.POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    POLVIEWS |
          1  |   .8235205   .1428313     5.77   0.000     .5435512     1.10349
          2  |   1.068801   .0860726    12.42   0.000      .900087    1.237516
          3  |   .8263137   .0866617     9.53   0.000     .6564446    .9961828
          5  |   .6960308   .0795739     8.75   0.000     .5400547    .8520068
          6  |   .4879752   .0767523     6.36   0.000       .33753    .6384204
          7  |  -.2127485   .1409817    -1.51   0.131    -.4890924    .0635954
             |
       _cons |   13.03805   .0415696   313.64   0.000     12.95657    13.11953
------------------------------------------------------------------------------
```

Since we've set the reference category with `fvset`, every time we run anything in which we declare `POLVIEWS` a factor variable with `i.`, Stata will now use 4 as the reference category each time.

2 Interaction

Interaction of variables in a regression analysis is specified by multiplying the two variables together. Previously, you need to do this by hand, which can get pretty hairy if you have a lot of categories:

```
gen male = SEX == 1
gen exlib = POLVIEWS == 1
gen lib = POLVIEWS == 2
gen slilib = POLVIEWS == 3
gen mod = POLVIEWS == 4
gen slicon = POLVIEWS == 5
gen con = POLVIEWS == 6
gen excon = POLVIEWS == 7
gen maleXexlib = male*exlib
gen maleXlib = male*lib
gen maleXslilib = male*slilib
gen maleXmod = male*mod
gen maleXslicon = male*slicon
gen maleXcon = male*con
gen maleXexcon = male*excon
```

And then we would need to include all of the above in our regression command. Instead, Stata gives us a shortcut in the form of `#`:

```
. regress EDUC i.SEX i.POLVIEWS SEX#POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.SEX |     .47561   .2759397     1.72   0.085    -.0652705     1.01649
             |
    POLVIEWS |
          2  |   .3613743   .2388327     1.51   0.130    -.1067713    .8295199
          3  |   .2344999   .2381342     0.98   0.325    -.2322765    .7012764
          4  |  -.5513693    .218091    -2.53   0.011    -.9788581   -.1238805
          5  |   .3614134   .2302676     1.57   0.117    -.0899433    .8127701
          6  |   .1740083   .2281394     0.76   0.446    -.2731767    .6211934
          7  |  -.5067274   .2843459    -1.78   0.075    -1.064085    .0506304
             |
SEX#POLVIEWS |
        2 2  |  -.2059436   .3152785    -0.65   0.514    -.8239337    .4120464
        2 3  |  -.4033632    .315548    -1.28   0.201    -1.021881    .2151551
        2 4  |   -.475527   .2883906    -1.65   0.099    -1.040813    .0897589
        2 5  |  -.8976255   .3074692    -2.92   0.004    -1.500308   -.2949428
        2 6  |  -.9415219   .3045796    -3.09   0.002    -1.538541   -.3445032
        2 7  |  -.9762233   .3855571    -2.53   0.011    -1.731969   -.2204774
             |
       _cons |   13.58937   .2087524    65.10   0.000     13.18019    13.99856
------------------------------------------------------------------------------
```

The # tells Stata to interact SEX and POLVIEWS. Note that when you use #, Stata assumes that the variables you are interacting are categorical variables. To interact with a continuous variable, you need to include the c. prefix:

```
. regress EDUC AGE i.POLVIEWS c.AGE#POLVIEWS, nohead
--------------------------------------------------------------------------------
          EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
           AGE |   .0042395   .0080479     0.53   0.598    -.0115355    .0200145
               |
      POLVIEWS |
            2  |   .6628868    .438429     1.51   0.131    -.1964956    1.522269
            3  |   .3190494   .4413477     0.72   0.470    -.5460541    1.184153
            4  |   .1462598   .4014724     0.36   0.716    -.6406825    .9332021
            5  |   1.026635   .4344239     2.36   0.018     .1751026    1.878167
            6  |   .7496973   .4314904     1.74   0.082    -.0960845    1.595479
            7  |   .6474008   .5718167     1.13   0.258    -.4734403    1.768242
               |
POLVIEWS#c.AGE |
            2  |  -.0091917   .0091819    -1.00   0.317    -.0271895    .0088061
            3  |  -.0068349   .0092625    -0.74   0.461    -.0249906    .0113209
            4  |  -.0208108   .0083921    -2.48   0.013    -.0372606    -.004361
            5  |  -.0244743   .0090242    -2.71   0.007    -.0421629   -.0067857
            6  |  -.0220221   .0088753    -2.48   0.013    -.0394189   -.0046252
            7  |   -.033427   .0112887    -2.96   0.003    -.0555543   -.0112996
               |
         _cons |   13.66125   .3835279    35.62   0.000     12.90948    14.41301
------------------------------------------------------------------------------
```

Notice that when we use #, we need to include the main effects separately. For Stata, # stands for just the interaction term. However, if we wanted to include all the interaction effects and main effects of the variables, we can use the operator ## - two pound signs tells Stata to include the interaction term and all main effects:

```
. regress EDUC SEX##POLVIEWS, nohead
------------------------------------------------------------------------------
        EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       2.SEX |     .47561   .2759397     1.72   0.085    -.0652705     1.01649
             |
    POLVIEWS |
          2  |   .3613743   .2388327     1.51   0.130    -.1067713    .8295199
          3  |   .2344999   .2381342     0.98   0.325    -.2322765    .7012764
          4  |  -.5513693    .218091    -2.53   0.011    -.9788581   -.1238805
          5  |   .3614134   .2302676     1.57   0.117    -.0899433    .8127701
          6  |   .1740083   .2281394     0.76   0.446    -.2731767    .6211934
          7  |  -.5067274   .2843459    -1.78   0.075    -1.064085    .0506304
             |
SEX#POLVIEWS |
        2 2  |  -.2059436   .3152785    -0.65   0.514    -.8239337    .4120464
        2 3  |  -.4033632    .315548    -1.28   0.201    -1.021881    .2151551
        2 4  |   -.475527   .2883906    -1.65   0.099    -1.040813    .0897589
        2 5  |  -.8976255   .3074692    -2.92   0.004    -1.500308   -.2949428
        2 6  |  -.9415219   .3045796    -3.09   0.002    -1.538541   -.3445032
        2 7  |  -.9762233   .3855571    -2.53   0.011    -1.731969   -.2204774
             |
       _cons |   13.58937   .2087524    65.10   0.000     13.18019    13.99856
------------------------------------------------------------------------------
```

Like before, Stata assumes variables used with the ## operator are categorical variables. This shortcut becomes really useful when we interact more than two variables - it will include all lower level interactions and main effects automatically. The next page shows the results of a three way interaction. It would have been incredibly tedious to create the many indicators and interactions by hand, so the ## operator is very useful in this case.

```
. regress EDUC SEX##POLVIEWS##RACE, nohead
-----------------------------------------------------------------------------------
             EDUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+----------------------------------------------------------------
            2.SEX |   .4441819   .3258858     1.36   0.173    -.1946001    1.082964
                  |
         POLVIEWS |
               2  |   .1188739   .2803632     0.42   0.672    -.4306772    .6684251
               3  |  -.0356943   .2777314    -0.13   0.898    -.5800867    .5086981
               4  |  -1.087544   .2549151    -4.27   0.000    -1.587213   -.5878742
               5  |    .006421   .2661787     0.02   0.981    -.5153265    .5281685
               6  |  -.1875431   .2631917    -0.71   0.476    -.7034357    .3283495
               7  |  -.6005161   .3213019    -1.87   0.062    -1.230313    .0292808
                  |
     SEX#POLVIEWS |
             2 2  |  -.1187812   .3712865    -0.32   0.749    -.8465548    .6089924
             2 3  |  -.3399038   .3698596    -0.92   0.358     -1.06488    .3850728
             2 4  |  -.4184207   .3395514    -1.23   0.218    -1.083989    .2471477
             2 5  |  -.8837868   .3581756    -2.47   0.014    -1.585861   -.1817123
             2 6  |  -.8971659   .3539616    -2.53   0.011     -1.59098   -.2033516
             2 7  |  -1.131034   .4451843    -2.54   0.011    -2.003657   -.2584102
                  |
             RACE |
               2  |  -1.891156   .5190791    -3.64   0.000    -2.908624   -.8736885
               3  |  -2.343537   .7408547    -3.16   0.002    -3.795717   -.8913583
                  |
         SEX#RACE |
             2 2  |   -.133122   .6765598    -0.20   0.844    -1.459274     1.19303
             2 3  |   1.202485   .9732746     1.24   0.217    -.7052699    3.110239
                  |
    POLVIEWS#RACE |
             2 2  |   .7228628   .6119335     1.18   0.238    -.4766128    1.922338
             2 3  |   1.200854   .8108066     1.48   0.139    -.3884407    2.790149
             3 2  |   .5775662   .6227936     0.93   0.354    -.6431966    1.798329
             3 3  |    1.19125   .8148141     1.46   0.144    -.4059001      2.7884
             4 2  |    1.68555   .5511719     3.06   0.002     .6051759    2.765925
             4 3  |   2.115555    .768735     2.75   0.006     .6087263    3.622383
             5 2  |   .4362598   .6237646     0.70   0.484    -.7864064    1.658926
             5 3  |   .8963568   .8258861     1.09   0.278    -.7224959    2.515209
             6 2  |   .3684955   .6523294     0.56   0.572    -.9101616    1.647153
             6 3  |   .1483274   .8292754     0.18   0.858    -1.477169    1.773824
             7 2  |  -.5472672   .7850557    -0.70   0.486    -2.086087    .9915522
             7 3  |  -2.532817   1.213427    -2.09   0.037    -4.911305   -.1543299
                  |
SEX#POLVIEWS#RACE |
           2 2 2  |   -.281755   .7945975    -0.35   0.723    -1.839278    1.275767
           2 2 3  |  -1.510215   1.084467    -1.39   0.164    -3.635923    .6154928
           2 3 2  |   .0788087   .8058376     0.10   0.922    -1.500746    1.658363
           2 3 3  |  -1.571014   1.096913    -1.43   0.152    -3.721118    .5790894
           2 4 2  |   .2295453   .7165159     0.32   0.749    -1.174926    1.634017
           2 4 3  |  -1.661019   1.014035    -1.64   0.101     -3.64867     .326632
           2 5 2  |    .565262   .8103806     0.70   0.485    -1.023198    2.153722
           2 5 3  |  -.7288761    1.09214    -0.67   0.505    -2.869623    1.411871
           2 6 2  |   .3871559   .8323598     0.47   0.642    -1.244386    2.018698
           2 6 3  |  -.4591114   1.104427    -0.42   0.678    -2.623942     1.70572
           2 7 2  |   1.587703   1.009605     1.57   0.116    -.3912651    3.566672
           2 7 3  |   2.466185   1.522995     1.62   0.105    -.5190987    5.451469
                  |
            _cons |   14.17687   .2446963    57.94   0.000     13.69723    14.65651
-----------------------------------------------------------------------------------
```