Sociology 7704: Regression Models for Categorical Data
Instructor: Natasha Sarkisian

Introduction to OLS Regression in Stata

To run an OLS regression:

```
. reg agekdbrn educ born sex mapres80

      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  4,  1086) =   51.24
       Model |  4954.03533     4  1238.50883            Prob > F      =  0.0000
    Residual |  26251.1232  1086   24.172305            R-squared     =  0.1588
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  31205.1586  1090  28.6285858            Root MSE      =  4.9165

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6122718   .0569422    10.75   0.000     .5005426     .724001
        born |   1.360161   .5816506     2.34   0.020     .2188749    2.501447
         sex |   -2.37973   .3075642    -7.74   0.000    -2.983218   -1.776243
    mapres80 |   .0243138   .0119552     2.03   0.042     .0008558    .0477718
       _cons |   16.95808   1.101139    15.40   0.000     14.79748    19.11868
------------------------------------------------------------------------------
```

Note that regression coefficients are partial slope coefficients; they indicate the change in the expected value of the dependent variable associated with one unit increase in the independent variable, when all other independent variables are held constant. These coefficients can potentially have two types of interpretation: cross-sectional and over time. Strictly speaking, all analyses we will do in this course are based on cross-sectional data.

To interpret the results, let's see how born and sex are coded:

```
. codebook born sex
-------------------------------------------------------------------------------
born                                                  was r born in this country
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  born

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  6/2765

            tabulation:  Freq.   Numeric  Label
                         2503         1  yes
                          256         2  no
                            6         .

-------------------------------------------------------------------------------
sex                                                             respondents sex
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  sex

                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  0/2765

            tabulation:  Freq.   Numeric  Label
                         1228         1  male
                         1537         2  female
```

To get standardized regression coefficients, we can use beta option:

```
. reg agekdbrn educ born sex mapres80, beta
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  4,  1086) =   51.24
       Model |  4954.03533     4  1238.50883            Prob > F      =  0.0000
    Residual |  26251.1232  1086   24.172305            R-squared     =  0.1588
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  31205.1586  1090  28.6285858            Root MSE      =  4.9165

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .6122718   .0569422    10.75   0.000                 .3108984
        born |   1.360161   .5816506     2.34   0.020                 .0651372
         sex |   -2.37973   .3075642    -7.74   0.000                -.2154051
    mapres80 |   .0243138   .0119552     2.03   0.042                 .0588174
       _cons |   16.95808   1.101139    15.40   0.000                       .
------------------------------------------------------------------------------
```

These coefficients indicate the number of standard deviations that agekdbrn increases per each one standard deviation increase in an independent variable.

Note on missing data: Stata estimation commands (e.g. regress, logit etc) automatically drop from the analysis all cases that miss data points on at least one of the variables used in the analyses (this is called listwise deletion). This can be very problematic when there is a lot of missing data and when the patterns of missing data are systematic (which is often the case); we will discuss the topic of dealing with missing data towards the end of the course.

In order to get your regression output to look nice, you can use estimates table. For example, for our regression model, we can run:

```
. est table,star b(%8.3f) varlabel stats(N) varwidth(40)
-------------------------------------------------------
                                Variable |   active
-----------------------------------------+-------------
        highest year of school completed |    0.612***
              was r born in this country |    1.360*
                          respondents sex |   -2.380***
mothers occupational prestige score (198 |    0.024*
                                Constant |   16.958***
-----------------------------------------+-------------
                                       N |     1091
-------------------------------------------------------
               legend: * p<0.05; ** p<0.01; *** p<0.001
```

This way you don't need to retype anything – it's closer to the journal format table. You can also have multiple models in the same table by storing the estimates of each and then using est table:

```
. quietly reg  agekdbrn educ born sex mapres80

. est store model1

. qui reg  agekdbrn educ born sex mapres80 sibs maeduc

. est store model2

. est table model1 model2, star b(%8.3f) varlabel stats(N) varwidth(40)
```

```
------------------------------------------------------------------------
                            Variable |   model1        model2
-------------------------------------+----------------------------------
        highest year of school completed |    0.612***      0.579***
              was r born in this country |    1.360*        1.423*
                        respondents sex |   -2.380***     -2.381***
mothers occupational prestige score (198 |    0.024*        0.016
          number of brothers and sisters |                 -0.128*
   highest year school completed, mother |                  0.029
                            Constant |   16.958***     17.820***
-------------------------------------+----------------------------------
                                   N |     1091          1035
------------------------------------------------------------------------
                          legend: * p<0.05; ** p<0.01; *** p<0.001
```

To find out more details and options, see help est_table as well as help estout (that is another command that allows to format the output in more complex ways).

If you are using nominal variables with more than just 2 categories or ordinal independent variables, you should not enter these variables in the model the same way you would use a continuous variable. For a nominal variable, that will result in nonsensical coefficients, because the categories are not really placed in any order so one unit increase is meaningless. For an ordinal variable, it's a stretch to use it in that fashion, because we assume equal distances among all categories. Before assuming that, we should test that assumption by introducing categories as separate variables. Here's how that's done in Stata.

```
. codebook  marital
-------------------------------------------------------------------------------
marital                                                          marital status
-------------------------------------------------------------------------------
type:  numeric (byte)
                label:  marital

                range:  [1,5]                        units:  1
        unique values:  5                    missing .:  0/2765

           tabulation:  Freq.   Numeric  Label
                        1269         1  married
                         247         2  widowed
                         445         3  divorced
                          96         4  separated
                         708         5  never married

. reg  agekdbrn educ born sex mapres80 i.marital
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  8,  1082) =   32.14
       Model |  5991.99195     8  748.998994           Prob > F      =  0.0000
    Residual |  25213.1666  1082  23.3023721           R-squared     =  0.1920
-------------+------------------------------           Adj R-squared =  0.1860
       Total |  31205.1586  1090  28.6285858           Root MSE      =  4.8273

------------------------------------------------------------------------------
      agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
          educ |   .5662673   .0570585     9.92   0.000     .4543094    .6782251
          born |   1.317066   .5740325     2.29   0.022     .1907232    2.443409
           sex |  -2.187909    .306421    -7.14   0.000    -2.789156   -1.586662
      mapres80 |   .0232956   .0117729     1.98   0.048     .0001953    .0463958
               |
       marital |
```

```
      widowed  |    .331999   .5584542     0.59   0.552    -.7637768    1.427775
     divorced  |  -.8996868   .3914891    -2.30   0.022    -1.667851   -.1315229
    separated  |  -2.101723   .7018116    -2.99   0.003    -3.478789   -.7246572
never married  |   -2.76481   .4698441    -5.88   0.000    -3.686719   -1.842901
               |
         _cons |   17.93003   1.111328    16.13   0.000     15.74943    20.11063
------------------------------------------------------------------------------
```

To change the reference group (omitted category):

```
. reg  agekdbrn educ born sex mapres80 ib5.marital

      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  8,  1082) =   32.14
       Model |  5991.99195     8  748.998994           Prob > F      =  0.0000
    Residual |  25213.1666  1082  23.3023721           R-squared     =  0.1920
-------------+------------------------------           Adj R-squared =  0.1860
       Total |  31205.1586  1090  28.6285858           Root MSE      =  4.8273

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5662673   .0570585     9.92   0.000     .4543094    .6782251
        born |   1.317066   .5740325     2.29   0.022     .1907232    2.443409
         sex |  -2.187909    .306421    -7.14   0.000    -2.789156   -1.586662
    mapres80 |   .0232956   .0117729     1.98   0.048     .0001953    .0463958
             |
     marital |
    married  |    2.76481   .4698441     5.88   0.000     1.842901    3.686719
    widowed  |   3.096809   .6574738     4.71   0.000     1.806741    4.386877
   divorced  |   1.865123   .5388517     3.46   0.001     .8078108    2.922436
  separated  |    .663087   .7909822     0.84   0.402    -.8889456     2.21512
             |
       _cons |   15.16522   1.133172    13.38   0.000     12.94176    17.38868
------------------------------------------------------------------------------
```

Alternatively, you can create permanent dichotomies for all categories:

```
. tab marital, gen(marital)
      marital |
       status |      Freq.     Percent        Cum.
--------------+-----------------------------------
      married |      1,269       45.90       45.90
      widowed |        247        8.93       54.83
     divorced |        445       16.09       70.92
    separated |         96        3.47       74.39
never married |        708       25.61      100.00
--------------+-----------------------------------
        Total |      2,765      100.00

. des marital*
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
marital         byte   %8.0g       marital    marital status
marital1        byte   %8.0g                  marital==married
marital2        byte   %8.0g                  marital==widowed
marital3        byte   %8.0g                  marital==divorced
marital4        byte   %8.0g                  marital==separated
marital5        byte   %8.0g                  marital==never married
```

```
. reg  agekdbrn educ born sex mapres80 marital2 marital3 marital4 marital5
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  8,  1082) =   32.14
       Model |  5991.99195     8  748.998994            Prob > F      =  0.0000
    Residual |  25213.1666  1082  23.3023721            R-squared     =  0.1920
-------------+------------------------------           Adj R-squared =  0.1860
       Total |  31205.1586  1090  28.6285858            Root MSE      =  4.8273
------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5662673   .0570585     9.92   0.000     .4543094    .6782251
        born |   1.317066   .5740325     2.29   0.022     .1907232    2.443409
         sex |  -2.187909    .306421    -7.14   0.000    -2.789156   -1.586662
    mapres80 |   .0232956   .0117729     1.98   0.048     .0001953    .0463958
    marital2 |    .331999   .5584542     0.59   0.552    -.7637768    1.427775
    marital3 |  -.8996868   .3914891    -2.30   0.022    -1.667851   -.1315229
    marital4 |  -2.101723   .7018116    -2.99   0.003    -3.478789   -.7246572
    marital5 |   -2.76481   .4698441    -5.88   0.000    -3.686719   -1.842901
       _cons |   17.93003   1.111328    16.13   0.000     15.74943    20.11063
------------------------------------------------------------------------------
```

Hypothesis testing after regression models

Oftentimes, we need to test hypotheses to find out whether coefficients are jointly significant, or to assess whether certain coefficients are equal to each other.

```
. reg  agekdbrn educ born sex mapres80 i.marital
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  8,  1082) =   32.14
       Model |  5991.99195     8  748.998994            Prob > F      =  0.0000
    Residual |  25213.1666  1082  23.3023721            R-squared     =  0.1920
-------------+------------------------------           Adj R-squared =  0.1860
       Total |  31205.1586  1090  28.6285858            Root MSE      =  4.8273

------------------------------------------------------------------------------
      agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
          educ |   .5662673   .0570585     9.92   0.000     .4543094    .6782251
          born |   1.317066   .5740325     2.29   0.022     .1907232    2.443409
           sex |  -2.187909    .306421    -7.14   0.000    -2.789156   -1.586662
      mapres80 |   .0232956   .0117729     1.98   0.048     .0001953    .0463958
               |
       marital |
      widowed  |    .331999   .5584542     0.59   0.552    -.7637768    1.427775
     divorced  |  -.8996868   .3914891    -2.30   0.022    -1.667851   -.1315229
    separated  |  -2.101723   .7018116    -2.99   0.003    -3.478789   -.7246572
never married  |   -2.76481   .4698441    -5.88   0.000    -3.686719   -1.842901
               |
         _cons |   17.93003   1.111328    16.13   0.000     15.74943    20.11063
------------------------------------------------------------------------------

. ereturn list

scalars:
                  e(N) =  1091
               e(df_m) =  8
               e(df_r) =  1082
                  e(F) =  32.14260719876214
                 e(r2) =  .1920192757413473
               e(rmse) =  4.827253060400016
                e(mss) =  5991.991948028161
```

```
               e(rss) =  25213.16662209091
              e(r2_a) =  .1860452962643886
                e(ll) =  -3261.080117285556
              e(ll_0) =  -3377.390032713303
              e(rank) =  9

macros:
            e(cmdline) : "regress agekdbrn educ born sex mapres80 i.marital"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
            e(depvar) : "agekdbrn"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "regress_estat"

matrices:
                  e(b) :  1 x 10
                  e(V) :  10 x 10

functions:
             e(sample)

. mat list e(b)

e(b)[1,10]
                                                      1b.         2.          3.          4.          5.
        educ       born         sex    mapres80     marital     marital     marital     marital     marital       _cons
y1   .56626729   1.3170661  -2.1879089   .02329556          0   .33199897  -.89968677  -2.1017231  -2.7648101   17.930029
```

A joint significance test:

```
. test 2.marital 3.marital 4.marital 5.marital

 ( 1)  2.marital = 0
 ( 2)  3.marital = 0
 ( 3)  4.marital = 0
 ( 4)  5.marital = 0

       F(  4,  1082) =   11.14
            Prob > F =    0.0000
```

Testing whether categories 2, 3, and 4 can be combined:

```
. test 2.marital=3.marital

 ( 1)  2.marital - 3.marital = 0

       F(  1,  1082) =    4.00
            Prob > F =    0.0457

. test 2.marital=4.marital, acc

 ( 1)  2.marital - 3.marital = 0
 ( 2)  2.marital - 4.marital = 0

       F(  2,  1082) =    4.34
            Prob > F =    0.0133
```

For an ordinal variable, hypothesis testing allows us to evaluate whether each one unit increase produces the same change in the dependent variable:

```
. reg  agekdbrn educ born sex mapres80 i.degree
      Source |       SS       df       MS              Number of obs =    1091
-------------+------------------------------           F(  8,  1082) =   32.94
       Model |  6111.91384     8   763.98923           Prob > F      =  0.0000
    Residual |  25093.2447  1082  23.1915386           R-squared     =  0.1959
-------------+------------------------------           Adj R-squared =  0.1899
       Total |  31205.1586  1090  28.6285858           Root MSE      =  4.8158

------------------------------------------------------------------------------
       agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
           educ |   .0506574   .1089486     0.46   0.642     -.163117    .2644317
           born |   1.267439    .570358     2.22   0.026     .1483064    2.386572
            sex |  -2.192157   .3025278    -7.25   0.000    -2.785764   -1.598549
       mapres80 |   .0225168   .0118318     1.90   0.057    -.0006991    .0457326
                |
         degree |
   high school  |   1.934153   .6048514     3.20   0.001     .7473387    3.120968
junior college  |   2.201938   .8713455     2.53   0.012     .4922196    3.911656
      bachelor  |   4.446438   .9701565     4.58   0.000     2.542837    6.350039
      graduate  |   7.624749   1.215111     6.27   0.000     5.240509    10.00899
                |
          _cons |   21.78773   1.329524    16.39   0.000     19.17899    24.39647
------------------------------------------------------------------------------
```

The increases are 1.93, 0.27, 2.24, 3.18, so they look unequal (especially because the junior college category is not really different from high school category), but let's test it:

```
. test 1.degree=(2.degree-1.degree)

 ( 1)  2*1.degree - 2.degree = 0

       F(  1,  1082) =    4.53
            Prob > F =    0.0335

. test 1.degree=(3.degree-2.degree), acc

 ( 1)  2*1.degree - 2.degree = 0
 ( 2)  1.degree + 2.degree - 3.degree = 0

       F(  2,  1082) =    2.41
            Prob > F =    0.0905

. test 1.degree=(4.degree-3.degree), acc

 ( 1)  2*1.degree - 2.degree = 0
 ( 2)  1.degree + 2.degree - 3.degree = 0
 ( 3)  1.degree + 3.degree - 4.degree = 0

       F(  3,  1082) =    3.98
            Prob > F =    0.0078
```