Stata Lecture Notes
Class 4

The following show the example of propensity scores shown in the Class 4 Lecture Notes. These notes make use of the dataset found in the Stata file `nmes_pro2.dta` After opening this dataset, run the following command to make sure that variables are stored as numeric values rather than string values:

```
. destring, replace
```

1. Example similar to Section 2 of the Lecture Notes: Estimating the log(OR) of MSCD, comparing smokers to non-smokers, while controlling for income. (We are controlling for income here, rather than controlling for poverty status as done in the lectures notes.)

There are 5 different values for the income variable:

```
. tab income

1-poor,2-ne |
         ar |
 poor,3-low |
income,4-mi |
       ddle |
income,5-hi |
  gh income |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,547       11.38       11.38
          2 |        727        5.35       16.73
          3 |      2,027       14.91       31.64
          4 |      4,198       30.89       62.53
          5 |      5,093       37.47      100.00
------------+-----------------------------------
      Total |     13,592      100.00
```

We can estimate the log(OR) of MSCD, comparing smokers to non-smokers separately in each of these income groups (strata). We also get the standard error of the estimate from this logistic regression:

```
. logit mscd eversmk if income==1

Logistic regression                             Number of obs     =      1,256
                                                LR chi2(1)        =      18.60
                                                Prob > chi2       =     0.0000
Log likelihood = -456.08275                     Pseudo R2         =     0.0200

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .7770725   .1863209     4.17   0.000     .4118902    1.142255
       _cons |   -2.44988   .1536727   -15.94   0.000    -2.751073   -2.148687
------------------------------------------------------------------------------

. logit mscd eversmk if income==2

Logistic regression                             Number of obs     =        595
                                                LR chi2(1)        =       5.83
```

```
                                                Prob > chi2       =     0.0157
Log likelihood = -255.00946                     Pseudo R2         =     0.0113

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .5529128    .231917     2.38   0.017     .0983638    1.007462
       _cons |  -1.993728   .1801735   -11.07   0.000    -2.346861   -1.640594
------------------------------------------------------------------------------

. logit mscd eversmk if income==3

Logistic regression                             Number of obs     =      1,664
                                                LR chi2(1)        =      31.10
                                                Prob > chi2       =     0.0000
Log likelihood = -672.74659                     Pseudo R2         =     0.0226

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .8198172   .1526641     5.37   0.000      .520601    1.119033
       _cons |  -2.289262   .1272643   -17.99   0.000    -2.538695   -2.039828
------------------------------------------------------------------------------

. logit mscd eversmk if income==4

Logistic regression                             Number of obs     =      3,631
                                                LR chi2(1)        =      11.96
                                                Prob > chi2       =     0.0005
Log likelihood = -1292.6444                     Pseudo R2         =     0.0046

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .3704381   .1085268     3.41   0.001     .1577294    .5831468
       _cons |  -2.262048   .0869631   -26.01   0.000    -2.432493   -2.091604
------------------------------------------------------------------------------

. logit mscd eversmk if income==5

Logistic regression                             Number of obs     =      4,499
                                                LR chi2(1)        =      13.63
                                                Prob > chi2       =     0.0002
Log likelihood = -1368.3422                     Pseudo R2         =     0.0050

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .3965334   .1091236     3.63   0.000     .1826552    .6104117
       _cons |    -2.5424    .088411   -28.76   0.000    -2.715683   -2.369118
------------------------------------------------------------------------------
```

We can use these values to fill in the table below and calculate the pooled estimate using Woolf's Method for pooling odds ratio estimates:

Income level (stratum)	logOR	SE	1/var = 1/SE^2	Weight	Weight*logOR
1	.777	.186	28.91	.112	.087
2	.553	.232	18.58	.072	.040
3	.820	.153	42.72	.166	.136

4	.370	.109	84.17	.327	.121
5	.397	.110	82.64	.322	.128
Pooled			257.02	~ 1.00	.512

So our estimate of this log(OR), pooling across the income strata, is 0.512. The standard error for this estimate is sqrt(1/257.02) = .062. We can see that this matches the results from a logistic regression for predicting MSCD from eversmk while controlling for the indicators of the income strata:

```
. logit mscd eversmk i.income

Logistic regression                             Number of obs     =     11,645
                                                LR chi2(5)        =     121.21
                                                Prob > chi2       =     0.0000
Log likelihood = -4049.3711                     Pseudo R2         =     0.0147

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .5166411   .0621103     8.32   0.000     .3949072    .6383751
             |
      income |
          2  |   .3067016   .1426604     2.15   0.032     .0270924    .5863109
          3  |   .1925588   .1112764     1.73   0.084     -.025539    .4106566
          4  |  -.0791885   .1010522    -0.78   0.433    -.2772472    .1188703
          5  |  -.3439344   .1009508    -3.41   0.001    -.5417943   -.1460745
             |
       _cons |  -2.278649   .0951062   -23.96   0.000    -2.465053   -2.092244
------------------------------------------------------------------------------
```

2. Example from to Section 3 of the Lecture Notes: Estimating the log(OR) of MSCD, comparing smokers to non-smokers, while controlling for the propensity score for smoking:

First, we need to calculate the propensity scores. We do this by fitting a logistic regression mode to predict the probability of (propensity for) being a smoker (eversmk) from the other covariates of interest. In this case we control for the covariates of male, age (using splines), income, and education level, while including an interaction term between age and male.

```
. mkspline newage 65 newage_sp65= age, marginal

. gen male_newage=male*newage

. gen male_newage_sp65=male*newage_sp65

. logit eversmk male newage newage_sp65 male_newage male_newage_sp65 i.income i.educate

Logistic regression                             Number of obs     =     11,645
                                                LR chi2(12)       =    1280.62
                                                Prob > chi2       =     0.0000
Log likelihood = -7328.5935                     Pseudo R2         =     0.0804

----------------------------------------------------------------------------------
         eversmk |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
            male |  -.4718718   .2926374    -1.61   0.107    -1.045431    .1016871
          newage |  -.0027621   .0034017    -0.81   0.417    -.0094292     .003905
     newage_sp65 |  -.0568172   .0078003    -7.28   0.000    -.0721055   -.0415289
```

```
     male_newage |   .0296449   .0054042     5.49   0.000     .0190529    .0402368
male_newage_sp65 |  -.0190136    .012415    -1.53   0.126    -.0433466    .0053193
                 |
          income |
              2  |  -.1286167   .1065875    -1.21   0.228    -.3375243    .0802909
              3  |  -.0486653   .0802896    -0.61   0.544      -.20603    .1086994
              4  |  -.1630927   .0717029    -2.27   0.023    -.3036277   -.0225577
              5  |  -.1950945   .0728702    -2.68   0.007    -.3379174   -.0522715
                 |
         educate |
              2  |   .4716228   .0746741     6.32   0.000     .3252642    .6179814
              3  |   .4566217     .06203     7.36   0.000     .3350451    .5781984
              4  |   .1599544   .0741713     2.16   0.031     .0145812    .3053275
                 |
           _cons |  -.0686404   .2019025    -0.34   0.734     -.464362    .3270812
------------------------------------------------------------------------------
```

Next, we use this logistic regression model to predict the probability of being a smoker for each person in the dataset. This probability is the propensity score:

```
. predict PrC
```

We then create groups (strata) based on these propensity scores. In this case, we use the quintiles of the propensity scores to create 5 groups where individuals within each group have similar risk of being a smoker based on the other covariates:

```
. centile PrC, centile(20(20)100)

                                                       -- Binom. Interp. --
    Variable |       Obs  Percentile    Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
         PrC |    13,592         20    .4184538         .4132604    .4225965
             |                   40    .5153694         .5143821    .5161662
             |                   60    .6099819         .5971178    .6192712
             |                   80    .7361238         .7333673    .7393156
             |                  100    .8407061         .8407061    .8407061*

* Lower (upper) confidence limit held at minimum (maximum) of sample
```

We call this new grouping variable group and can look at the breakdown of MSCD within each group for both the smoking and non-smoking group:

```
. gen group=1 if PrC <0.418
(10,932 missing values generated)

. replace group=2 if PrC >= 0.418 & PrC < .515
(2,668 real changes made)

. replace group=3 if PrC >= 0.515 & PrC <.610
(2,799 real changes made)

. replace group=4 if PrC >= 0.610 & PrC < 0.736
(2,688 real changes made)

. replace group=5 if PrC >=0.736 & PrC <0.841
(2,721 real changes made)

. tab group mscd if eversmk==0, row

+----------------+
| Key            |
|----------------|
|   frequency    |
```

```
| row percentage |
+----------------+

           |         mscd
     group |         0          1 |     Total
-----------+----------------------+----------
         1 |     1,251        203 |     1,454
           |     86.04      13.96 |    100.00
-----------+----------------------+----------
         2 |     1,150         97 |     1,247
           |     92.22       7.78 |    100.00
-----------+----------------------+----------
         3 |     1,118         35 |     1,153
           |     96.96       3.04 |    100.00
-----------+----------------------+----------
         4 |       606         48 |       654
           |     92.66       7.34 |    100.00
-----------+----------------------+----------
         5 |       492         50 |       542
           |     90.77       9.23 |    100.00
-----------+----------------------+----------
     Total |     4,617        433 |     5,050
           |     91.43       8.57 |    100.00

. tab group mscd if eversmk==1, row

+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+

           |         mscd
     group |         0          1 |     Total
-----------+----------------------+----------
         1 |       633        150 |       783
           |     80.84      19.16 |    100.00
-----------+----------------------+----------
         2 |       928        129 |     1,057
           |     87.80      12.20 |    100.00
-----------+----------------------+----------
         3 |     1,196         89 |     1,285
           |     93.07       6.93 |    100.00
-----------+----------------------+----------
         4 |     1,467        183 |     1,650
           |     88.91      11.09 |    100.00
-----------+----------------------+----------
         5 |     1,487        333 |     1,820
           |     81.70      18.30 |    100.00
-----------+----------------------+----------
     Total |     5,711        884 |     6,595
           |     86.60      13.40 |    100.00
```

We could estimate the log(OR) of MSCD, comparing smokers to non-smokers separately in each of these propensity score groups (strata). We could also get the standard error of the estimate from this logistic regression and then combine these strata-specific estimates using Woolf's Method for pooling odds ratio estimates. Or, we can simply include the indicators for these propensity score groups in our logistic regression model:

```
. logit mscd eversmk i.group

Logistic regression                             Number of obs     =     11,645
                                                LR chi2(5)        =     286.96
                                                Prob > chi2       =     0.0000
Log likelihood = -3966.4945                     Pseudo R2         =     0.0349

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .5496893   .0664281     8.27   0.000     .4194926     .679886
             |
       group |
          2  |  -.6101314   .0917195    -6.65   0.000    -.7898982   -.4303645
          3  |   -1.35853   .1100753   -12.34   0.000    -1.574274   -1.142787
          4  |  -.7203958    .093844    -7.68   0.000    -.9043266    -.536465
          5  |  -.1923965   .0849841    -2.26   0.024    -.3589623   -.0258307
             |
       _cons |  -1.890916   .0651202   -29.04   0.000     -2.01855   -1.763283
------------------------------------------------------------------------------
```

Our estimate of this log(OR), pooling across the income strata, is 0.550. We can compare this estimated log(OR) to what we would get if we individually controlled for all of the same covariates in a multiple logistic regression model:

```
. logit mscd eversmk male newage newage_sp65 male_newage male_newage_sp65 i.income
i.educate

Iteration 0:   log likelihood =  -4109.977
Iteration 1:   log likelihood = -3723.0194
Iteration 2:   log likelihood = -3646.5678
Iteration 3:   log likelihood = -3645.3432
Iteration 4:   log likelihood =  -3645.339
Iteration 5:   log likelihood =  -3645.339

Logistic regression                             Number of obs     =     11,645
                                                LR chi2(13)       =     929.28
                                                Prob > chi2       =     0.0000
Log likelihood =  -3645.339                     Pseudo R2         =     0.1131

---------------------------------------------------------------------------------
            mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+---------------------------------------------------------------
         eversmk |   .6543806   .0693947     9.43   0.000     .5183696    .7903916
            male |  -.6277045    .715148    -0.88   0.380    -2.029369    .7739597
          newage |   .0911537   .0085118    10.71   0.000     .0744708    .1078365
     newage_sp65 |  -.0378389   .0130328    -2.90   0.004    -.0633827    -.012295
     male_newage |   .0165958   .0119243     1.39   0.164    -.0067755     .039967
male_newage_sp65 |  -.0388879    .018824    -2.07   0.039    -.0757823   -.0019936
                 |
          income |
              2  |   .1177465   .1480643     0.80   0.426    -.1724542    .4079471
              3  |   .0224403   .1161032     0.19   0.847    -.2051178    .2499984
              4  |   -.113896   .1078534    -1.06   0.291    -.3252848    .0974928
              5  |  -.2286938   .1117192    -2.05   0.041    -.4476595   -.0097281
                 |
         educate |
              2  |   .2836474   .1249633     2.27   0.023     .0387238     .528571
              3  |   .1354226   .1066018     1.27   0.204     -.073513    .3443582
              4  |   -.148143   .1194866    -1.24   0.215    -.3823324    .0860463
                 |
           _cons |  -8.253345   .5284758   -15.62   0.000    -9.289139   -7.217551
---------------------------------------------------------------------------------
```