Class 4 Outline

1. Goal of controlling for potential confounding and set-up
2. Stratification to account for potential confounding
3. Propensity score strategy and detailed example
4. Comparison of propensity score results to that obtained by multivariable logistic regression
5. Pros and cons of constructing propensity scores
6. References

1

0. Learning Objectives

- Identify and review possible methods to control for potential confounding.
- Define and construct a propensity score for a major covariate of interest based on the possible confounders of the association between it and the outcome of interest.
- Review a detailed example to show the construction and use of propensity scores to control for potential confounding.

2

1. Goal of Controlling for Potential Confounding

To estimate the effect of a "treatment" or "risk factor" (e.g., ever smoking) on an outcome (e.g. major smoking caused disease) by *comparing otherwise similar persons with and without the risk factor*.

3

1.1 Set-Up

- Health response: Y = 1,0
  - Major smoking caused disease (MSCD)
- Binary treatment or risk factor: Z = 1,0
  - Ever smoker
- Potential confounders: X
  - Age
  - Gender
  - SES: Poverty, education; marital status, seat belt use

4

2. Stratification to Account for Potential Confounding

- Stratify by the covariate
- Woolf's method for pooling (combining) odds ratio estimates
- Multivariable logistic regression

5

2.1 Controlling for One Covariate

- Stratify by the covariate
- Estimate the difference in mean outcome or log odds ratio within each covariate stratum
- Pool the stratum-specific estimates of effects absent any evidence of qualitative effect modification

6

2.1a Example: MSCD, Ever Smoker, Poverty

Poverty Level	Probability of MSCD (n)		Log OR	Std Error
	Ever smokers	Never smokers		
1 (Poverty)	.076 (181)	.042 (213)	.630	.439
2	.081 (86)	.089 (101)	-.099	.526
3	.122 (285)	.043 (296)	1.11	.336
4	.092 (682)	.052 (651)	.613	.220
5 (No Poverty)	.076 (758)	.042 (823)	.623	.220

7

2.2 Woolf's Method for Pooling (Combining) Odds Ratio Estimates

- Weight each odds ratio estimate inversely proportional to the variance of the estimate
- Give more weight to less variable estimates
- Combine or pool $\log_e$ OR estimates

8

2.2a Pool the Evidence Using Weighted Mean of Log OR

Stratum	log OR	se	1/var= $1/se^2$	weight= (1/var)/total	weight·logOR
1	.63	.439	5.19	.088	.0554
2	- .099	.526	3.61	.061	-.0060
3	1.11	.336	8.86	.150	.1665
4	.613	.220	20.66	.350	.2146
5	.623	.220	20.66	.350	.2181
Pooled			58.98	1.00	0.65

$$se_{\log OR} = \sqrt{\frac{1}{58.98}} = 0.130$$

9

2.3 Multivariable Logistic Regression

```
.logit mscd eversmk i.POVSTALB

Logistic regression                               Number of obs   =       4078
                                                  LR chi2(5)      =      32.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -995.96268                       Pseudo R2       =     0.0162
------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .6558862    .129402     5.07   0.000     .4022629    .9095095
_IPOVSTALB_2 |   .4213909   .3396267     1.24   0.215    -.2442652    1.087047
_IPOVSTALB_3 |    .362854   .2632888     1.38   0.168    -.1531826    .8788906
_IPOVSTALB_4 |   .2106937   .2401061     0.88   0.380    -.2599055     .681293
_IPOVSTALB_5 |   .0022913   .2406694     0.01   0.992     -.469412    .4739946
       _cons |  -3.136465   .2292715   -13.68   0.000    -3.585829   -2.687102
------------------------------------------------------------------------------
```

- A faster method of pooling evidence!
- Regressing Y on X and indicators of the strata is identical to weighting the log ORs inversely related to their variances

10

3. What to Do with Many Confounders?

- Stratify on all confounder combinations
  - Large number of strata, hard to make tables
- Match each smoker to a few "similar" non-smokers; not bad, but does not use all the data
- Stratify on a single derived variable chosen so that the distribution of all the covariates is similar for the two treatment groups within each stratum of the variable.
  - One such variable is the **propensity score**

11

3.1 Propensity Score Definition

- Definition: $p(X) = Pr(Z=1|X)$
  - The propensity score is the probability of being "treated" (smoking) as a function of the potential confounders
- Fact: the distribution of X given p(X) is the same whether Z=1 or Z=0
  - The treated (smokers) and untreated (non-smokers) within a propensity score stratum are alike with respect to the covariates (age, gender, SES variables)

12

3.2 Propensity Score Strategy

- Estimate the propensity score using logistic regression or other classification method
  - Similar to that performed in the example in Class 15, Biostat 622
- Stratify into quintiles of the estimated propensity score
- Estimate the treatment effect within each stratum
- Pool the estimates across strata

13

3.2a Estimating the Propensity Score

- Question: does the rate of ever smoking differ for men and woman *who are otherwise similar*?
- Major variables in a larger data set:
  - eversmk = ever smoker : 1-yes; 0-no
  - age =age at survey
  - male (0-female; 1-male)
  - educate (1-college grad; 2-some college;3-high school grad; 4 - other)
  - poor (poverty level) (1-at or below poverty line; 2 - up to twice line;…;5 - 5 or more times)

14

3.2b Estimating the Propensity Score

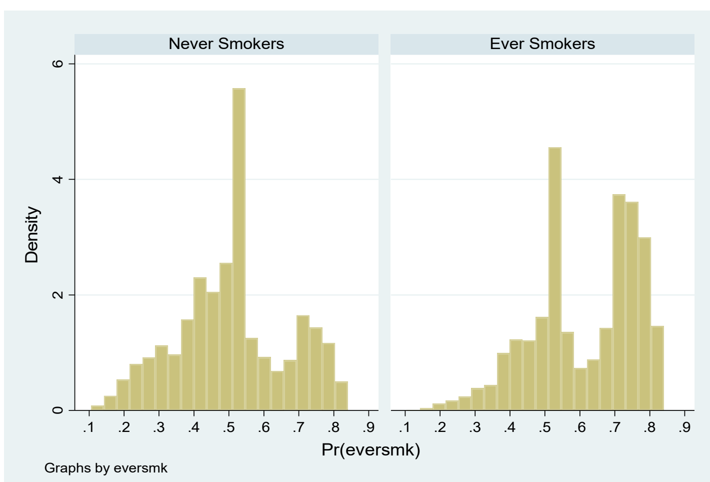
```
. mkspline newage 65 newage_sp65= age, marginal
. gen male_newage=male*newage
. gen male_newage_sp65=male*newage_sp65

. logit eversmk male newage newage_sp65 male_newage male_newage_sp65 i.poor 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
                 |
            poor |
              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
----------------------------------------------------------------------------------
. predict PrC
```

15

3.3 Distribution of Propensity Scores for Never and Ever Smokers

Propensity Score=Pr(Smoker|age, gender, SES)



16

3.4 Creating Quintiles for Predicted Probabilities of Ever Smoking

```
. 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*


. 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)
```

17

3.5 Probability of MSCD by Ever Smoking within Quintiles (Groups)

Never Smokers

```
. tab group mscd if eversmk==0, row

           |         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
```

Ever Smokers

```
. tab group mscd if eversmk==1, row

           |         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
```

18

3.6 Log OR of MSCD by Ever Smoking within Propensity Score Quintiles

Propensity Score Quintile	Probability of MSCD (n) Ever	Never	Log_e OR	Std error
1	.1916 (783)	. 1396 (1454)	0.379	
2	. 1220 (1057)	.0778 (12471)	0.499	
3	.0693 (1285)	.0304 (1153)	0.866	
4	.1109 (1650)	.0734 (654)	0.454	
5	.1830 (1820)	.0923 (542)	0.791	

19

3.7a Another Way: Logistic Regression of Ever Smoking within Quintiles

```
. logit mscd eversmk if group==1
Logistic regression                             Number of obs     =      2,237
------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .3786574   .1182029     3.20   0.001     .1469841    .6103307
       _cons |  -1.818493   .0756668   -24.03   0.000    -1.966797   -1.670188
------------------------------------------------------------------------------

. logit mscd eversmk if group==2

Logistic regression                             Number of obs     =      2,304

------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .4995869   .1414509     3.53   0.000     .2223482    .7768256
       _cons |  -2.472806     .10573   -23.39   0.000    -2.680033   -2.265579
------------------------------------------------------------------------------
```

20

3.7b Another Way: Logistic Regression of Ever Smoking within Quintiles

```
. logit mscd eversmk if group==3
Logistic regression                               Number of obs     =      2,438
------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |    .865847   .2038086     4.25   0.000     .4663894    1.265305
       _cons |  -3.463949   .1716563   -20.18   0.000    -3.800389   -3.127508
------------------------------------------------------------------------------

. logit mscd eversmk if group==4
Logistic regression                               Number of obs     =      2,304
------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .4541904    .169203     2.68   0.007     .1225586    .7858221
       _cons |  -2.535679    .149945   -16.91   0.000    -2.829566   -2.241792
------------------------------------------------------------------------------

. logit mscd eversmk if group==5
Logistic regression                               Number of obs     =      2,362
------------------------------------------------------------------------------
        mscd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     eversmk |   .7900811   .1603371     4.93   0.000     .4758261    1.104336
       _cons |  -2.286455   .1484335   -15.40   0.000    -2.577379    -1.99553
------------------------------------------------------------------------------
```

21

3.8 Pooling the Evidence in a Single Logistic Regression

```
. 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
------------------------------------------------------------------------------

.
```

22

3.9 Findings from Propensity Score Analysis

- We estimate that the odds of having a major smoking caused disease is exp(.55)=1.73 times as high among ever smokers versus never smokers *who have similar demographic and SES characteristics*
- 95% CI: (exp(.42)= *1.52*, exp(.68)=*1.97*) times higher

23

4. Compare Propensity Score Results to Multiple Logistic Regression

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


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
                 |
            poor |
              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
```

24

5.1a Pros and Cons of Propensity Scores

- Organizes the analysis into 2 steps
  - Probability of treatment given the covariates: there is sometimes prior knowledge about this probability, for example in randomized trials ($p(X)=.5$)
  - Comparison of treatment groups within strata of assignment probability
- Easy to picture the evidence for the binary treatment effect

25

5.1b Pros and Cons of Propensity Scores

- Most natural with binary treatment
  - Extensions possible, but they are awkward
- Not as simple to study effect modifications (interactions)
- No method controls for unmeasured confounders, regardless of what is claimed

26

6.0 References

- Propensity scores can also be used to perform weighted analyses

- References
  - Rosenbaum and Rubin, 1983. *Biometrika*, 70: 41-55.
  - Rubin. 1997. *Annals of Internal Medicine*, 127: 757-763.

27