Section 9c

Propensity scores

Controlling for bias & confounding in observational studies

Logistic regression and propensity scores

Consider comparing an outcome in two “treatment” groups: A vs B. In a randomized clnicial trial, the randomization process provides that, on average, values of the potential covariates/confounders are similar between the two groups, thus eliminating bias.

But in an observational study with **no** randomization, we may need to control for many measured covariates that are both related to the outcome and are different in the two treatment groups.

As we know, this can create very messy ANOVA/regression models. In the case of continuous covariates, one might doubt that assumptions of linearity/parallelism are true. In ANOVA models, presence of complicated significant multiway interactions may be difficult to explain.

What to do? If there were only a few covariates, we could make strata from each covariate pattern. Within each stratum, there would be no relationship between treatment group vs. covariates since the covariates would all have the same value in the stratum. That is, the association between treatment and covariates would disappear in each stratum.

But this is impractical if there are many covariates with many levels. There are too many potential strata.

However, if we had a model where we **knew** the probability of each person being **assigned** to treatment A (= 1- prob of assignment to B), statisticians have shown that one can then **stratify** on this probability. Within each stratum, it turns out that the value of the covariates are roughly the **same** between the two treatments**!** That is, it is not necessary to make strata with identical

covariate patterns, only identical probabilities. It is sufficient to stratify only on the probability of being assigned to treatment A (vs B). Forming such strata will "automatically" create comparability! That is, within any one stratum, the X values **will be similar** between treatments A and B if everyone in the stratum has about the same probability of being assigned to A! (Even though, in fact, some we assigned to B).
Of course, within a stratum, the number of cases in group A will not be the same as the number of cases in group B.

While we don't in fact know the probability of assignment exactly, we can model it (using logistic regression, for example).

The propensity score therefore is the (estimated) probability (or any monotonically related score, such as the logit) of being assigned to treatment A (vs B). We stratify on this score/probability to obtain comparability and eliminate the association between treatment and covariates.

We do this when we are not really interested in the relation between the covariates and the outcome. We also don't really care if the propensity (i.e. logistic) model is "correct" or has any actual meaning as long as it lets us create strata where there is comparabilty between the two treatments within each stratum.

Is a new treatment for "whiter teeth" better than the standard treatment? Sample of n=350 people.

t test - comparing mean gray scale scores

Unadjusted scores - observational study

This is not a randomized trial

group	n	mean	sd	sem
STD	208	39.45	24.1	1.67
NEW	142	42.51	20.8	1.75
difference		3.06		2.49

t= -1.23 p=0.219

Covariate comparison

	STD, n=208			**NEW, n=142**			p value
age	**mean**	SD	sem	**mean**	SD	sem	
	22.36	6.47	0.45	**24.4**	6.33	0.53	**0.004**
sugar use	**6.10**	3.08	0.21	**5.84**	3.06	0.26	0.435
		PCT	**SE**		**PCT**	**SE**	
male		28.4%	3.1%		47.2%	4.2%	**0.0003**
floss		28.9%	3.1%		35.9%	4.0%	0.1629
yearly cleaning		31.7%	3.2%		32.4%	3.9%	0.896
drink coffee		42.3%	3.4%		74.7%	3.7%	**<0.0001**
drink tea		30.8%	3.2%		62.7%	4.1%	**<0.0001**
use mouthwash		22.1%	2.9%		25.4%	3.7%	0.4827

Covariates not the same.

Logistic regression with treatment(tx)as the outcome to estimate propensity (probability) of being assigned to the new treatment (= 1 – prob of assignment to standard treatment).

```
The LOGISTIC Procedure

 Ordered                Total
   Value     tx     Frequency
       1     new          142
       2     std          208   n= 350

          Model Fit Statistics

                                 Intercept
                 Intercept          and
Criterion          Only         Covariates
AIC               474.682          433.440
SC                478.540          468.162
-2 Log L          472.682          415.440

R-Square    0.1509    Max-rescaled R-Square    0.2036

        Testing Global Null Hypothesis: BETA=0

Test                  Chi-Square       DF     Pr > ChiSq
Likelihood Ratio         57.2422        8         <.0001
Score                    54.3801        8         <.0001
Wald                     48.7461        8         <.0001

                                   Standard
Parameter    DF   Estimate         Error    Chi-Square    Pr > ChiSq
Intercept     1    -1.7982        0.5417       11.0185        0.0009
age           1     0.0214        0.0196        1.1945        0.2744
male          1     0.3898        0.2559        2.3201        0.1277
floss         1     0.3280        0.2601        1.5905        0.2073
clean         1    -0.0543        0.2556        0.0450        0.8319
sugar         1    -0.0401        0.0393        1.0400        0.3078
coffee        1     0.9042        0.2767       10.6771        0.0011
tea           1     0.8681        0.2570       11.4094        0.0007
mwash         1    -0.1009        0.2844        0.1258        0.7228

Association of Predicted Probabilities and Observed Responses
 Percent Concordant     72.4    Somers' D    0.451
 Percent Discordant     27.4    Gamma        0.452
 Percent Tied            0.2    Tau-a        0.218
 Pairs                 29536    c            0.725
```

Estimated propensity=1/(1+exp(-logit))

gray scale means by propensity stratum

	STD	**STD**	NEW	**NEW**	n	**mean**	p value	propen score
stratum	n	**mean**	n	**mean**		**difference**		
1	83	**21.3**	4	**27.5**	87	**6.2**	0.5304	0-.2
2	49	**43.9**	39	**36.9**	88	**-7.0**	0.0915	0.2-0.4
3	38	**53.9**	50	**40.6**	88	**-13.3**	0.0014	0.4-0.6
4	38	**58.9**	49	**50.2**	87	**-8.7**	0.0358	0.6+
total n	208		142		350			
adjusted mean		**44.5**		**38.8**		**-5.7**	0.06	
unadj mean		39.4		42.5		**3.1**	0.21	
adj mean stratum 2,3,4		52.2		42.5		**-9.7**		

ANOVA for gray score by tx group & stratum

```
         Type 3 Tests of Fixed Effects
                 Num       Den
Effect            DF        DF      F Value      Pr > F
tx                 1       342         3.53      0.0610
stratum            3       342        14.71      <.0001
tx*stratum         3       342         1.24      0.2963
```

```
         Least Squares Means     df=342
                   (Mean)     Standard
Effect     tx      Estimate    Error
tx         new     38.7852     2.6938
tx         std     44.4894     1.3977
```

```
Effect       tx   stratum   mean      std error
tx*stratum   new    1       27.5000    9.5834
tx*stratum   std    1       21.3373    2.1038

tx*stratum   new    2       36.8974    3.0692
tx*stratum   std    2       43.8571    2.7381

tx*stratum   new    3       40.5800    2.7106
tx*stratum   std    3       53.8684    3.1093

tx*stratum   new    4       50.1633    2.7381
tx*stratum   std    4       58.8947    3.1093
```

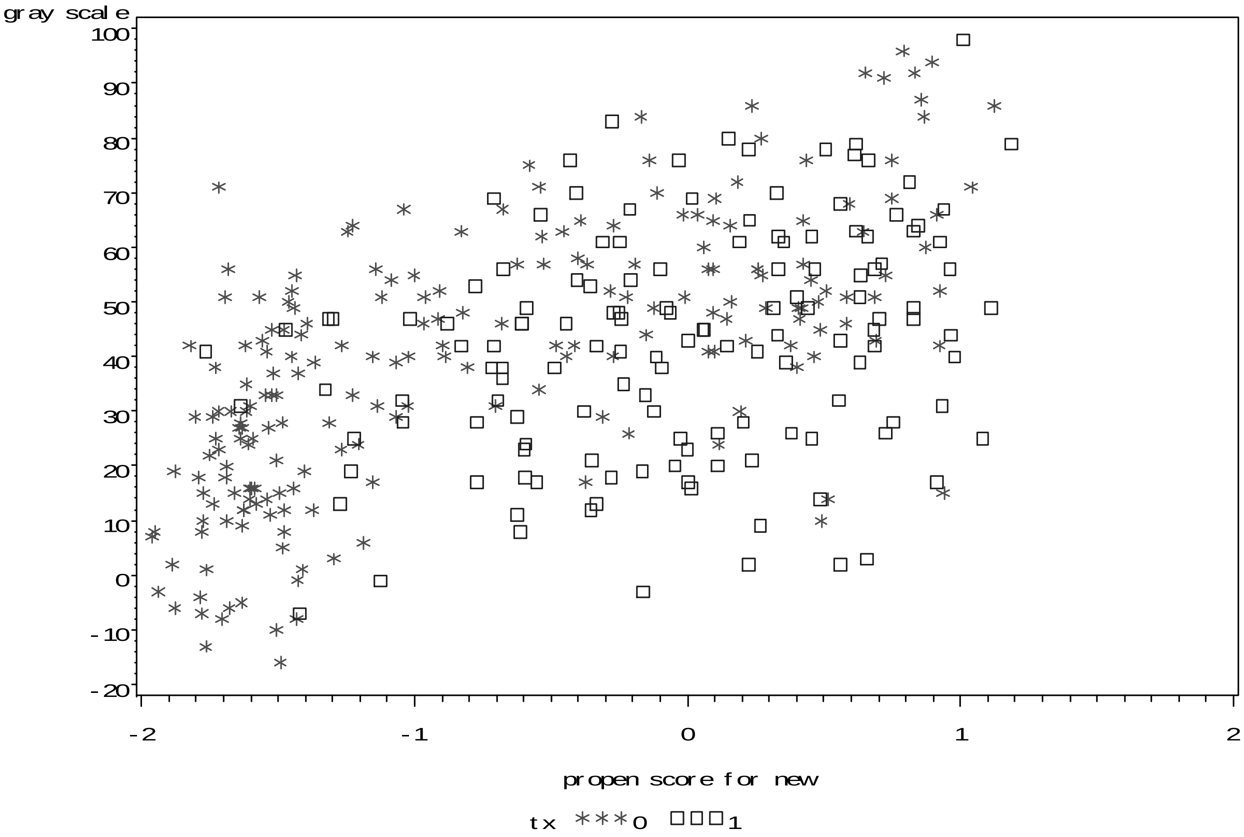
```
     Differences of Least Squares Means

                                Standard
Effect   tx   vs   tx   Estimate   Error    DF    t Value    Pr > |t|
tx       std       new   5.7042   3.0349   342    -1.88       0.0610
```

mean score

Tx	**Unadjusted**	**Adjusted**
Std	39.447	44.4894
New	42.507	38.7852
Diff (new-std)	3.06	-5.704

p value	0.21	0.06

Gray scale versus propensity score by group
* = STD whitener   □ = NEW whitener



The propensity score for choosing the NEW whitener is a function of eight covariates (age, sugar use, gender, flossing, tooth cleaning, drink coffee, drink tea, use mouthwash). It is the logit from the logistic regression. The higher the score, the more likely one is assigned (or chose) the NEW treatment.

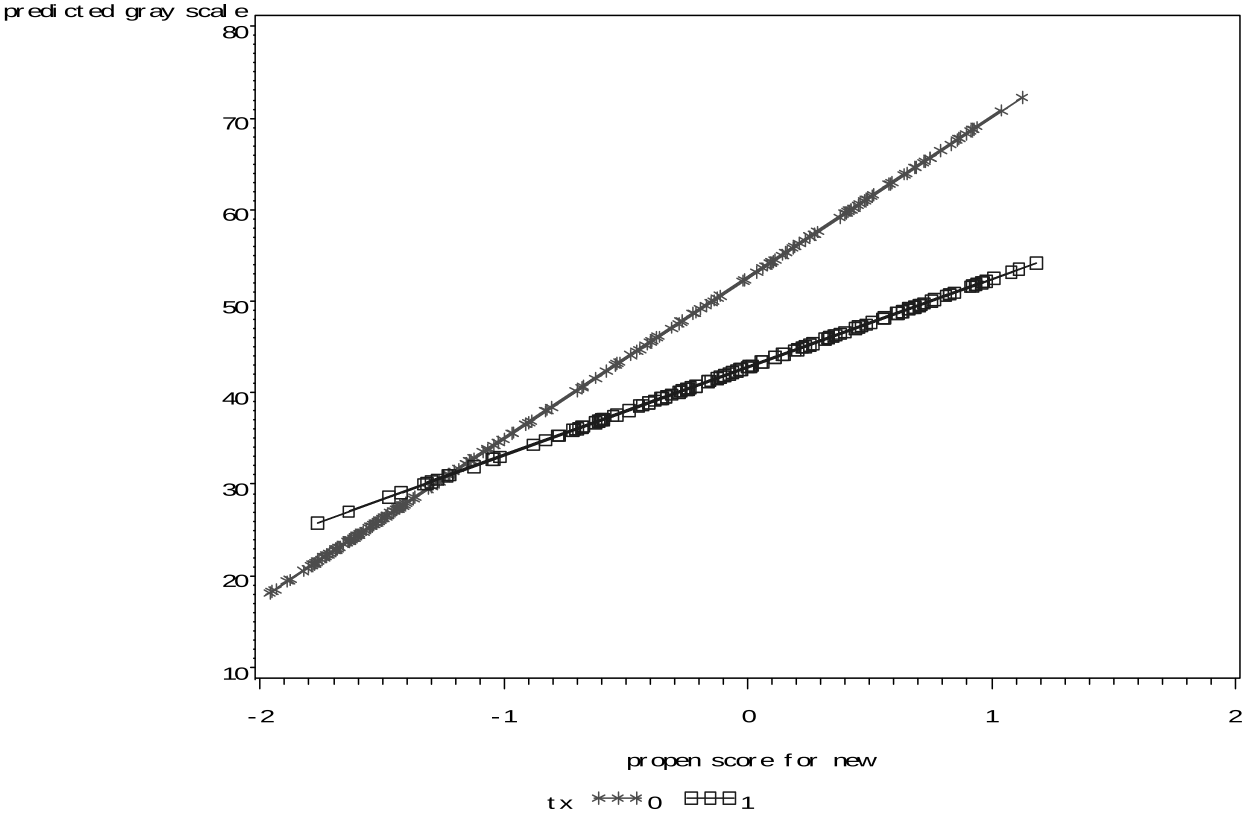
```
The REG Procedure
Dependent Variable: score gray scale

Number of Observations Used        350

                        Analysis of Variance

                             Sum of       Mean
Source               DF     Squares     Square    F Value  Pr> F
Model                 3      59728      19909.0    56.31   <.0001
Error               346     122337       353.6
Corrected Total     349     182182065

Root MSE           18.80360    R-Square      0.3281
Dependent Mean     40.68857    Adj R-Sq      0.3222
Coeff Var          46.21346

                          Parameter Estimates

                           Parameter  Std
Variable                DF Estimate   Error t Value   Pr>|t|
Intercept                1    52.574   1.688  31.15 <.0001
New tx                   1    -9.768   2.312  -4.23 <.0001
Propen score (logit)     1    17.558   1.433  12.25 <.0001
New tx * propen score 1       -7.942   2.759  -2.88 0.0042
```

“New tx” is coded 1 for new, 0 for old

Q – If the propensity score is a good proxy for the covariates, what should happen if any or all of the 8 covariates are added to the above model? Are they needed?

```
   Regression model based mean gray scale (Ŷ) as a function
of (logit) propensity score to choose new whitner
```

```
  * = STD whitener    □ = NEW whitener
```



As the propensity to choose the NEW treatment increases, the mean difference between the two treatments increases.

Comparing covariates by strata

mean age

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	18.0	24.8	25.5	25.6
NEW	25.2	23.5	23.7	25.8
p value	0.0668	0.2648	0.1696	0.8743

mean sugar use

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	6.55	5.63	6.05	5.76
NEW	7.62	6.66	5.55	5.33
p value	0.4616	0.1587	0.3865	0.5455

pct male

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	3.6%	24.5%	44.7%	71.1%
NEW	0.0%	30.8%	46.0%	65.3%
p value	0.078	0.514	0.906	0.566

pct who floss

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	20.5%	34.7%	26.3%	42.1%
NEW	25.0%	23.1%	30.0%	53.1%
p value	0.838	0.225	0.702	0.307

pct who get yearly tooth cleaning

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	26.5%	40.8%	34.2%	28.9%
NEW	75.0%	25.6%	32.0%	34.7%
p value	0.070	0.126	0.827	0.566

pct drink coffee

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	0.0%	34.7%	86.8%	100.0%
NEW	0.0%	46.2%	78.0%	100.0%
p value	1.000	0.274	0.271	1.000

pct drink tea

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	0.0%	8.2%	57.9%	100.0%
NEW	0.0%	25.6%	60.0%	100.0%
p value	1.000	**0.040**	0.842	1.000

pct use mouthwash

tx	stratum 1	stratum 2	stratum 3	stratum 4
STD	19.3%	14.3%	28.9%	31.6%
NEW	50.0%	25.6%	16.0%	32.7%
p value	0.226	0.186	0.150	0.915

Algorithm Summary for Estimating the Propensity Score (Dehejia)

1. Start with a parsimonious logit model to estimate the propensity score.

2. Sort the data according to estimated propensity score (from lowest to highest).

3. Stratify all observations such that estimated propensity scores within a stratum for treated and comparison units are close (no significant difference); e.g., start by dividing observations into strata of equal score range (0-0.2,...,0.8-1). Rubin suggests 4-10 strata.

4a Check that covariates are similar (balanced) within each stratum. For all covariates, differences in means (or proportions) between treated versus comparison units within each stratum should not be significantly different from zero.
4b. If covariates are balanced between treated and comparison observations for all strata, stop- this is successful.
4c. If covariates are not balanced for some stratum, divide the stratum into finer strata and re-evaluate.
4d. If a particular covariate is not balanced for many strata, modify the logit by adding interaction terms and/or higher-order terms of the covariate and re-evaluate.

Advantages of Propensity score analysis

1. Reduces all the covariates to one dimension

2. Easy to check if the two groups being compared overlap on the score (ie on the covariates)

3. Does not extrapolate beyond the range of the data (unlike linear regression)

4. Robust – Does not matter if model for propensity score is miss specified as long as covariates are the same in the strata made by the score

Drawbacks to Propensity analysis

Can only be used when there are two treatment groups of interest.

If the mean treatment (A-B) difference varies from one (propensity) stratum to the next, this is a crude indication that the mean difference varies by covariate pattern. That is, there may be a treatment x covariate interaction.

If this is of concern, one may have to run the usual multivariate model to identify and report on the interactions.

Getting the same mean difference across strata imply that the mean difference is the same for all covariate patterns.

Limitations of Propensity score methods

"good" example - unique propensity for each covariate pattern

In this example, we are comparing mean SBP in drug A to drug B and we assume there are only two other covariates, gender and smoking.
Below, every unique combination of these two covariates has a different probability (propensity) of getting drug A.

mean SBP - all possible "true" strata

gender	smoking	mean Drug A	n for a	mean Drug B	n for b	mean difference
male	smoker	140	600	120	75	20
female	smoker	125	90	137	35	-12
male	non smoker	133	80	120	40	13
female	non smoker	120	10	140	20	-20
overall - ignore covariates (incorrect)		137.3	780	125.9	170	11.4
overall- stratum adjusted (correct)		129.5		129.25		**0.25**

Since each covariate combination has a different propensity, the propensity analysis will make a different stratum for each covariate pattern as above.

gender	smoking	Proportion on Drug A = propensity
male	smoker	89%
female	smoker	72%
male	non smoker	67%
female	non smoker	33%

The predicted probabilities from the logistic model exactly match the observed probabilities

logit(Prob of getting drug A) =
 - 0.6931 + 1.386 male + 1.638 smoker - 0.25 male* smoker

(male=1 for male, 0 for female, smoker=1 for smoker, 0 for non smoker)

Propensity score analysis limitations (continued)

"bad example" - very different covariate patterns have same propensity

In this example, two very different covariate patterns (male smokers & female smokers) have the same propensity (same probability getting drug A)

mean SBP - all possible "true" strata

gender	smoking	mean Drug A	n for a	mean Drug B	n for b	mean difference
male	smoker	140	600	120	75	20
female	smoker	125	280	137	35	-12
male	non smoker	133	80	120	40	13
female	non smoker	120	10	140	20	-20
overall - ignore covariates (incorrect)		135	970	126	170	9.0
overall- stratum adjusted (correct)		129.5		129.25		**0.25**

gender	smoking	Proportion on Drug A = propensity
male	smoker	89%
female	smoker	89%
male	non smoker	67%
female	non smoker	33%

Since the male smokers and female smokers have the same propensity, the propensity analysis puts both these covariate patterns into the same stratum as below. All smokers are in the same stratum ignoring gender, even though gender influences SBP.

mean SBP- strata based on propensity

stratum	propensity	mean Drug A	n for a	mean Drug B	n for b	mean difference
1	89%	135	880	125	110	10
2	67%	133	80	120	40	13
3	33%	120	10	140	20	-20
overall-ignore covariates		135	970	126	170	9.0
overall-propensity stratum adjusted		126.5		130		**-3.5**

Here, the "stratum adjusted" mean difference is not quite correct!
It is -3.5 instead of 0.25 even though stratum 1 has 68% male smokers for both drugs.