

## Lecture 19

1. Comparing groups with an observational study
2. Unbalanced predictors
3. AER example
4. Propensity scores
5. Matching
6. Treatment effect estimated from sample matched on propensity scores

1

### Comparing treatments with an observational study

Comparison of treatments aims to compare *effects* of treatments on something.

**Experiment:** researcher *assigns* treatment to subject

Researcher makes a change and observes the effect. If subjects were alike except for treatment (by randomization), difference in effect was caused by treatments.

**Observational study:** subjects choose their own treatment

Subjects may be different, and difference relates to both choice of treatment and outcome.

Subject differences may cause part or all of treatment difference.

2

Challenge for observational studies:

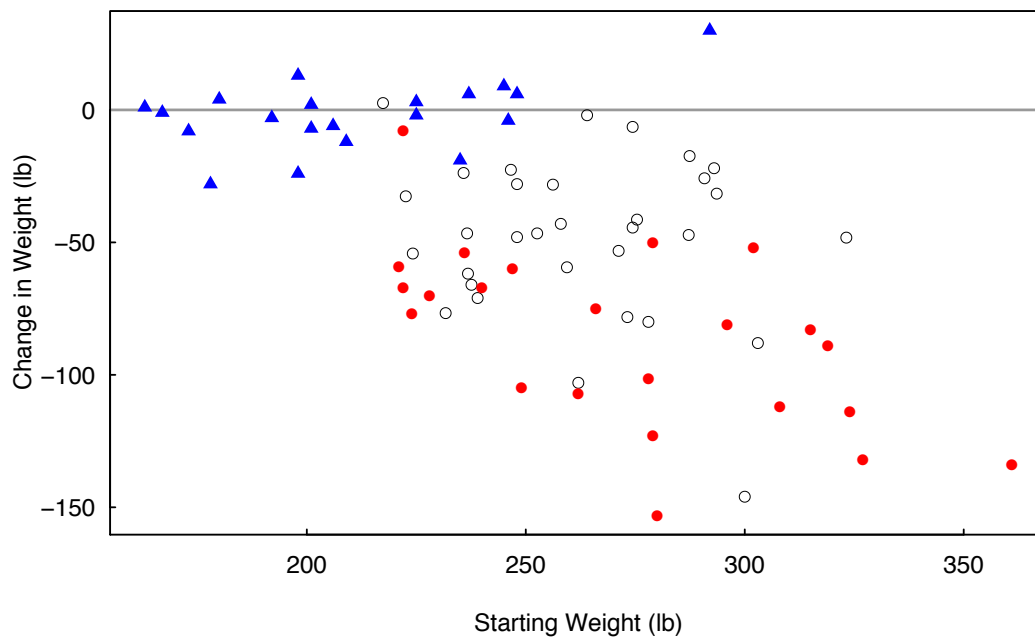
**show subjects in treatment groups alike**

Example 1. Observational data (1992): retrospectively collected weight losses in very overweight diabetic patients who received one of three treatments:

- gastric-bypass surgery
- very-low-calorie liquid diet
- ▲ standard medical care

3

Treatments: standard medical care (▲), very-low-calorie liquid diet (○), or gastric-bypass surgery (•).



4

Must adjust comparison for baseline weight.

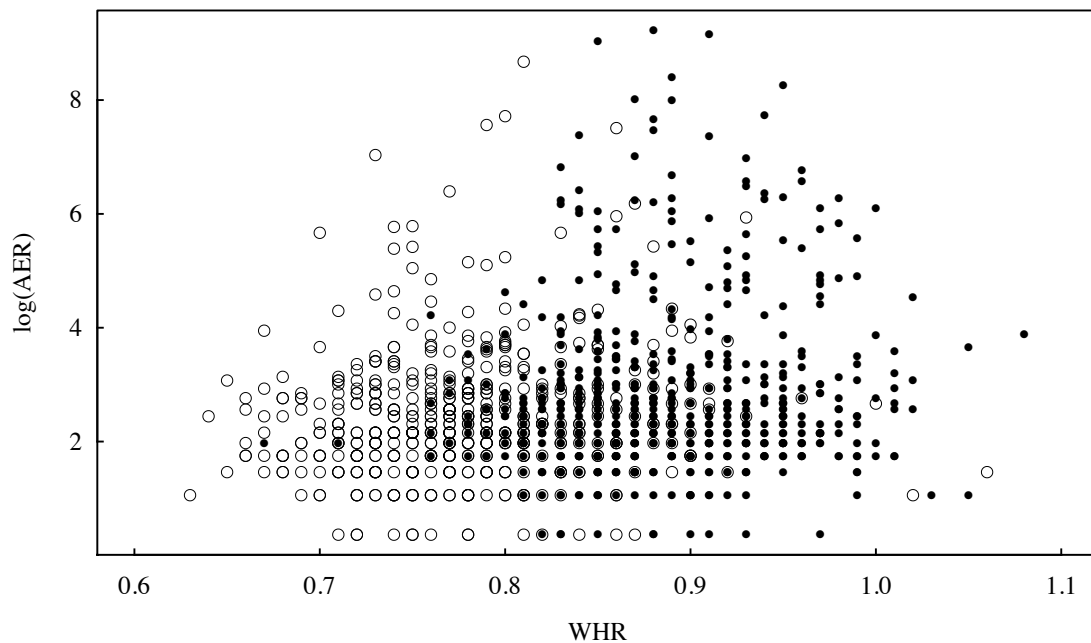
1. Use regression (weight change on baseline weight) to adjust for baseline weight.
2. Stratify on baseline weight, use strata that contain large enough samples from at least 2 treatment groups.

Stratification is difficult with small samples and minimal overlap.

What if there were 20 other baseline variables that were also different?

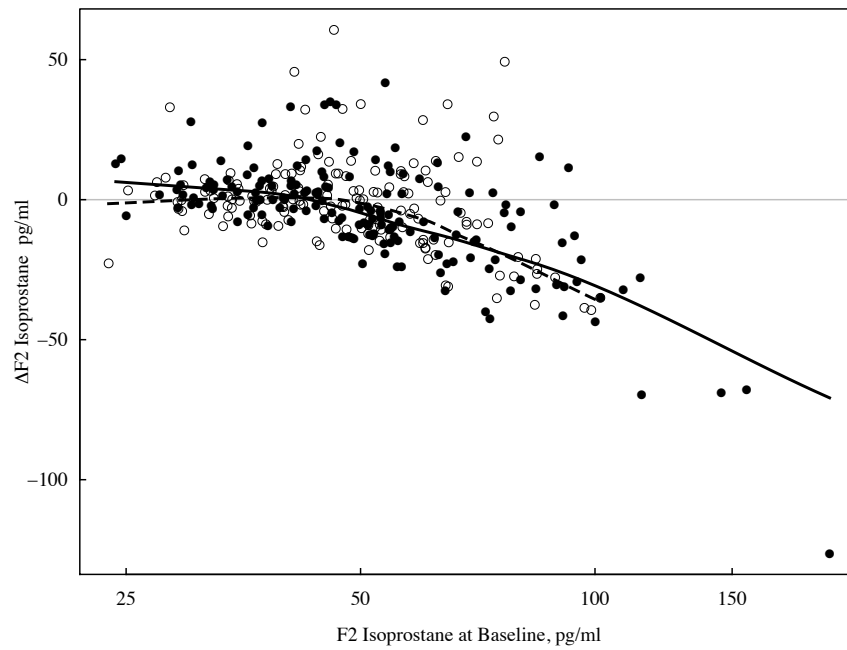
5

Example 2. Compare albumin excretion rate (AER) between genders, adjusting for waist-hip ratio (WHR).



6

Example 3. Randomization does not guarantee balance. Clinical trial with two treatments:

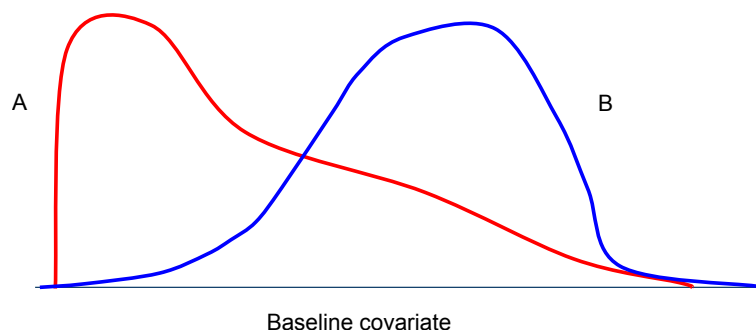


7

### Differences in baseline characteristics

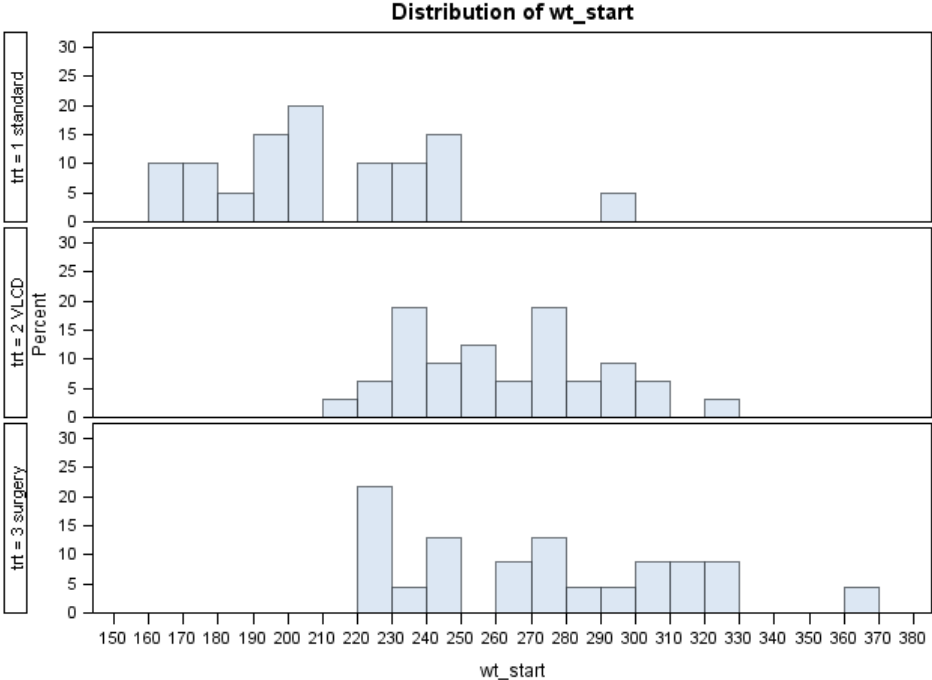
Baseline characteristics may differ between observational treatment groups

- **limited overlap** of range
- **imbalance**: similar range, but different distributions

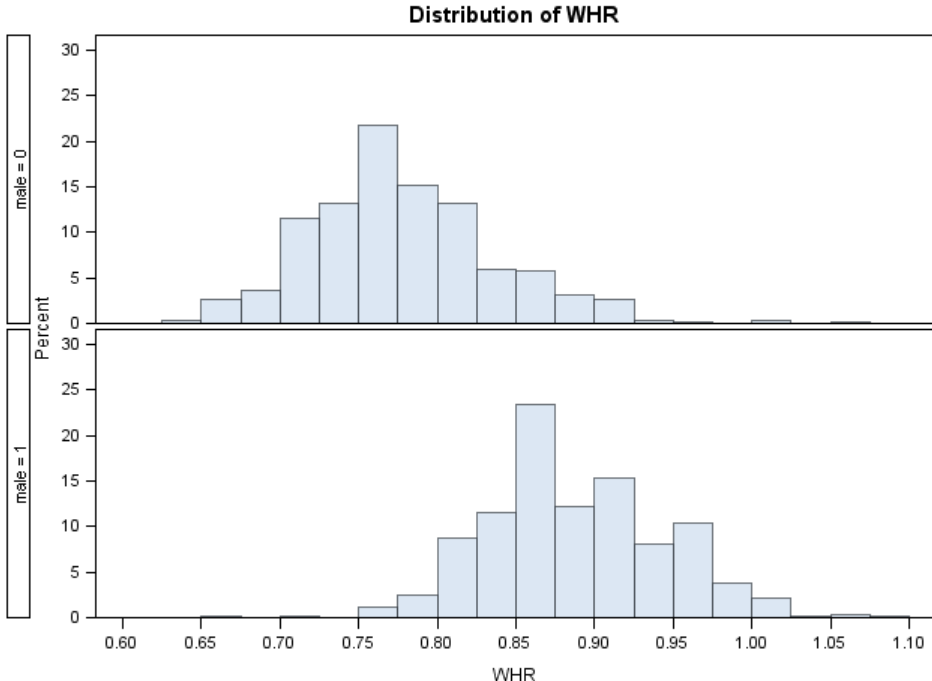


8

Limited overlap of range: gastric-bypass baseline weights



Limited overlap, different distributions:



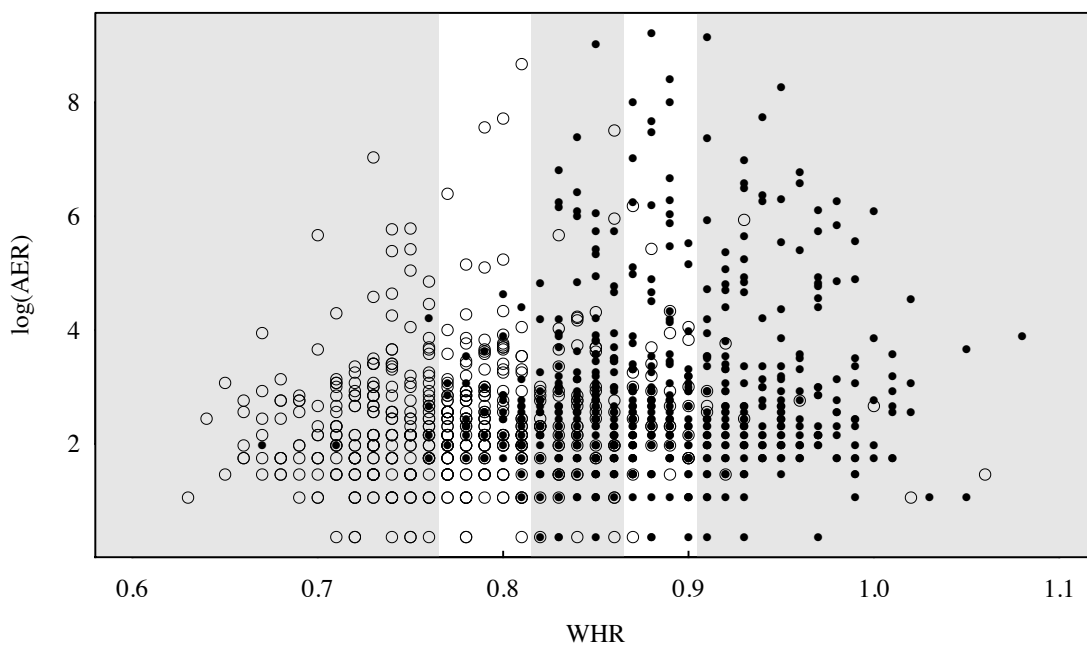
Code for stacked panels of histograms:

```
data rename; change variable levels to sort correctly
  set pubh.med_vs_surgery;
  trt="1 standard";
  if (code=2) then trt="2 VLCD";
  if (code=3) then trt="3 surgery";

ODS graphics on;
Proc Univariate noprint data= rename;
  var wt_start;
  class trt;
  histogram wt_start / nrows=3 endpoints = 150 to 380 by 10;
run;
ODS graphics off;
```

11

AER example: tried restricting comparison to “balanced” middle quintiles:



12

SEX	QWHR (Rank for Variable WHR)					Total
Frequency	0	1	2	3	4	
F	223	173	64	30	11	501
M	6	49	204	167	230	656
Total	229	222	268	197	241	1157

“Balance” is not great. Also ignores question of balance in other predictors.

13

**Table 1. Descriptive Statistics by Sex**

	Men (n = 675)	Women (n = 510)	P
Intensive therapy during DCCT	327 (48)	261 (51)	0.352
Age (y)	38 ± 7	38 ± 7	0.341
Duration of diabetes (y)	15.9 ± 5	16.5 ± 5	0.046
HbA <sub>1c</sub> (%)	8.2 ± 1.3	8.3 ± 1.5	0.537
AER (mg/24 h)	16.2 (14.6-18.1)	11.9 (10.8-13.2)	<0.001
Subjects with elevated AER	136 (20)	81 (16)	0.054
BMI (kg/m <sup>2</sup> )	26.7 ± 4	26.1 ± 4	0.008
WHR	0.89 ± 0.06	0.78 ± 0.06	<0.001
Waist circumference (cm)	90.4 ± 10	80.0 ± 10	<0.001
SBP (mm Hg)	122 ± 13	117 ± 13	<0.001
DBP (mm Hg)	78 ± 9	75 ± 8	<0.001
Serum creatinine (mg/dL)	1.0 ± 0.1	0.8 ± 0.1	<0.001
Total cholesterol (mg/dL)	188 ± 37	190 ± 37	0.449
Triglycerides (mg/dL)	96 ± 67	84 ± 66	0.003
HDL cholesterol (mg/dL)	50 ± 12	61 ± 14	<0.001
LDL cholesterol (mg/dL)	119 ± 31	112 ± 30	<0.001
LDL/HDL cholesterol ratio	2.5 ± 0.9	2.0 ± 0.8	<0.001
Smoking status (yes)	131 (19)	99 (19)	0.999

Source: *American Journal of Kidney Diseases*, 2006; 47: 223–32.

14

*Question:* Is there a gender difference in AER in this population of type I diabetics, adjusting for WHR, age, LDL/HDL ratio, HBA1c level, duration of diabetes, smoking status, and SBP?

From Table 1, significant baseline differences in WHR, LDL/HDL ratio, duration of diabetes, SBP.

- Stratification on 4 variables:
  - 5 strata each gives  $4^5 = 1024$  subgroups, but total sample is  $n = 1185$
  - 4 strata each gives  $4^4 = 256$  subgroups, roughly 4 per subgroup
  - Stratification doesn't work here even though we have a fairly large sample size.
- Matching on 4 variables: no automatic procedure. Proc Sort, then pair by hand—difficult.

15

### **Propensity Score**

A different way to balance these characteristics between genders:

1. Compute each subject's **propensity score** = probability subject is male, given their WHR, age, LDL/HDL ratio, HBA1c level, duration of diabetes, smoking status, and SBP.  
Use logistic regression, output fitted probabilities.
2. Form matched pairs based on propensity score. (SAS macro for this.)
3. Perform planned analysis on matched data, but ignore the pairing in the analysis.

16

Calculate propensity scores in logistic regression:

```
Proc Logistic descending data=ph6470.whrdata;  
    model male = WHR age ldl_hdl HBA1C duration smoking SBP;  
    output out=AER_P pred=pscore ; fitted p-hat is propensity score
```

ODS graphics on;

```
Proc Univariate noprint data= AER_P ;
```

```
var pscore;
```

```
class male;
```

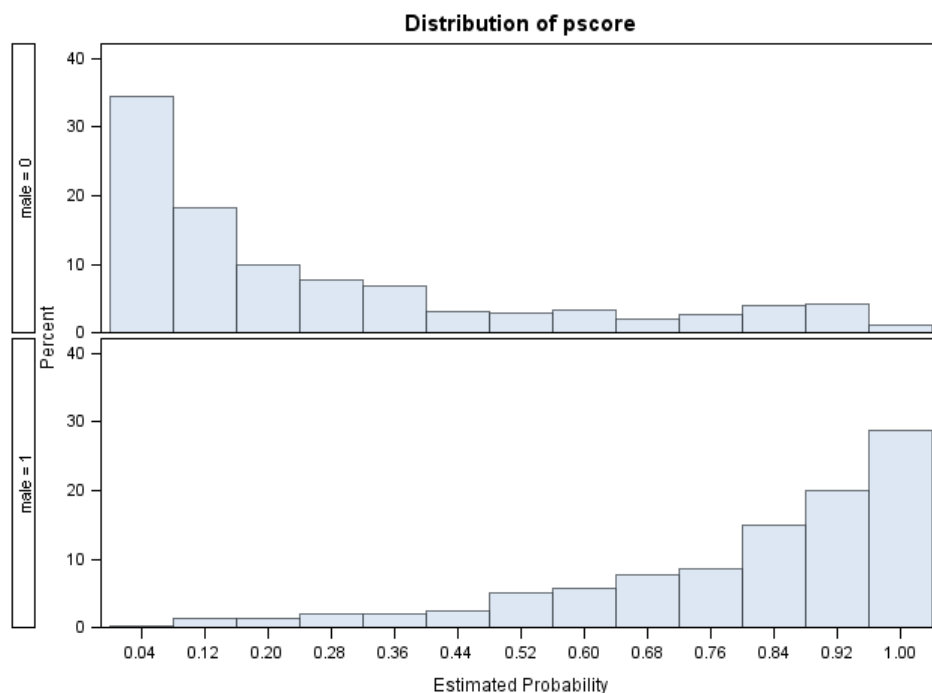
```
histogram pscore / nrows=2 ;
```

```
run;
```

ODS graphics off;

17

Propensity scores have very different distributions by gender



18

## 2. Forming matched pairs

Use macro **%PSmatching** adapted from M. Coca-Perraillon (1987).

Make separate Treatment and Control datasets such that:

- Control data includes: idC = subject\_id, pscoreC = propensity score  
Treatment data includes: idT, pscoreT  
Other variables are discarded.
- *method of matching*: caliper or NN (nearest neighbor)

19

Make two separate datasets:

```
data T C;
  set AER_P;
  if male=0 then do;
    idC = id; pscoreC = pscore; output C;
  end;
  if male=1 then do;
    idT = id; pscoreT = pscore; output T;
  end;
```

20

Call the macro:

```
%include "X:\ PC SAS examples\SAS Macros\PSmatching.sas";  
  
%PSMatching(datatreatment= T, datacontrol= C,  
method= caliper, caliper=.1,  
numberofcontrols= 1, replacement=no, out=PS_match_cal);
```

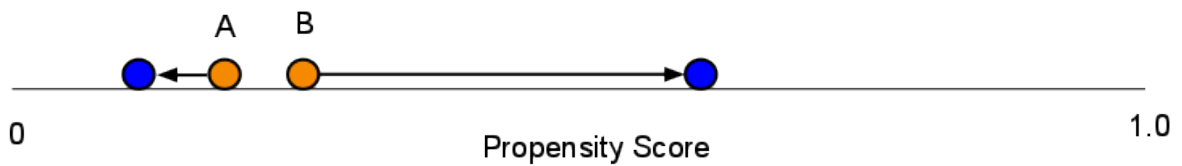
**caliper value** = max radius for matching

**replacement** = yes/no whether controls can be matched to more than one case

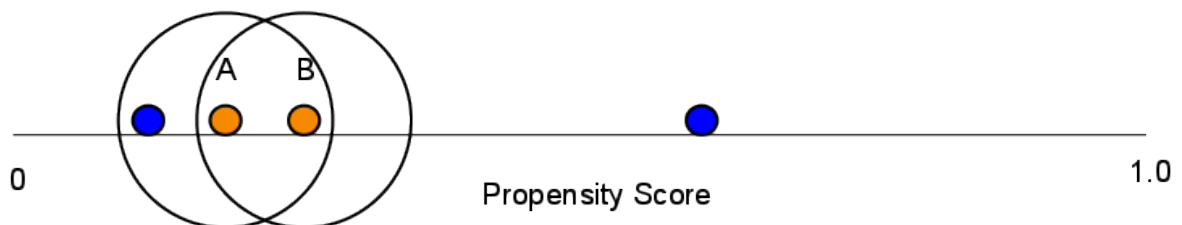
**out** = output data set name

21

Nearest-neighbor : matches each treatment observation to nearest control, pairs can be any distance apart

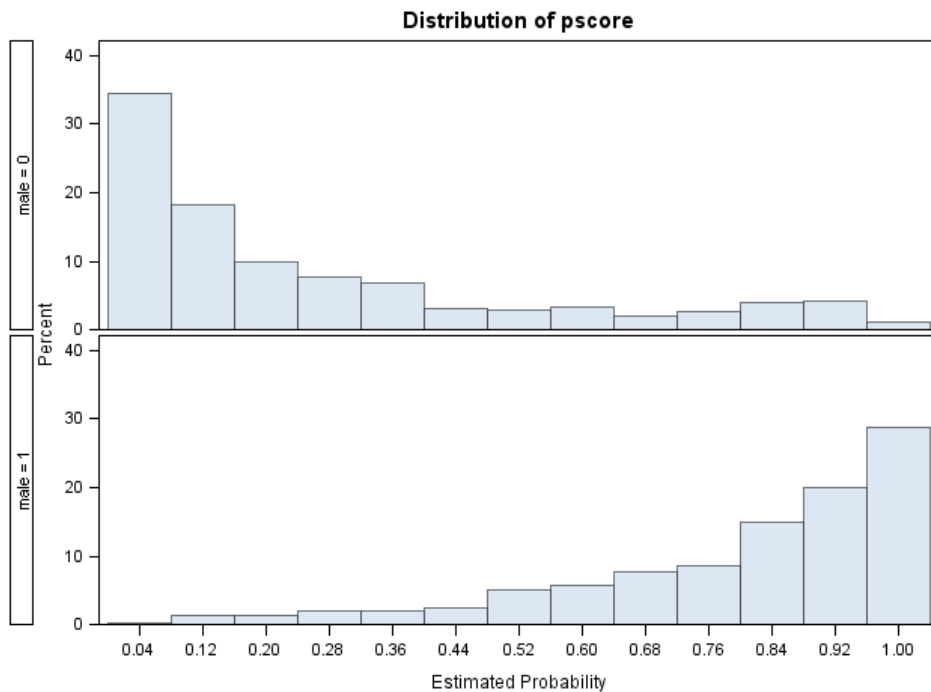


Caliper matching: matches each treatment observation to nearest control within caliper = maximum distance (no match for observation B)



22

How large should the caliper (maximum distance) be? Used 0.1



23

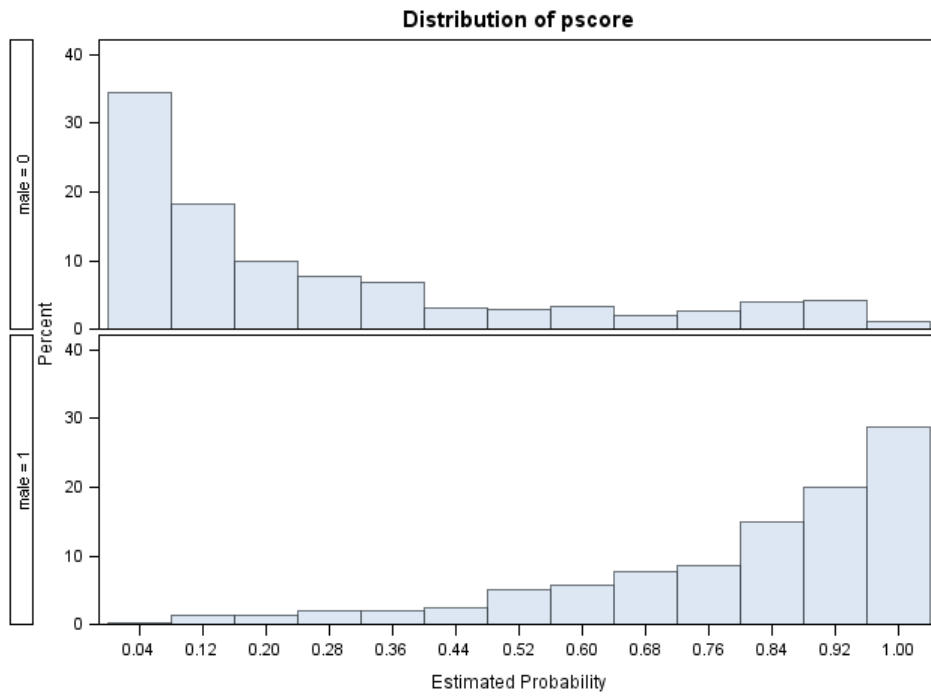
Obs	Id		Matched	
	Selected	PScore	To	PScore
	Control	Control	TreatID	Treat
1	5303	0.72641	27198	0.72443
2	5219	0.98164	13175	0.98473
3	17086	0.88649	12015	0.88515
4	4232	0.97758	6233	0.97410
5	22208	0.93011	14166	0.92881
6	5380	0.95024	9145	0.95109
7	17322	0.91933	12152	0.91260
8	8005	0.78349	7152	0.78517
9	20166	0.47323	19165	0.46772
10	26001	0.62917	21199	0.62902

All other variables gone; need to separate observations for merging.

24

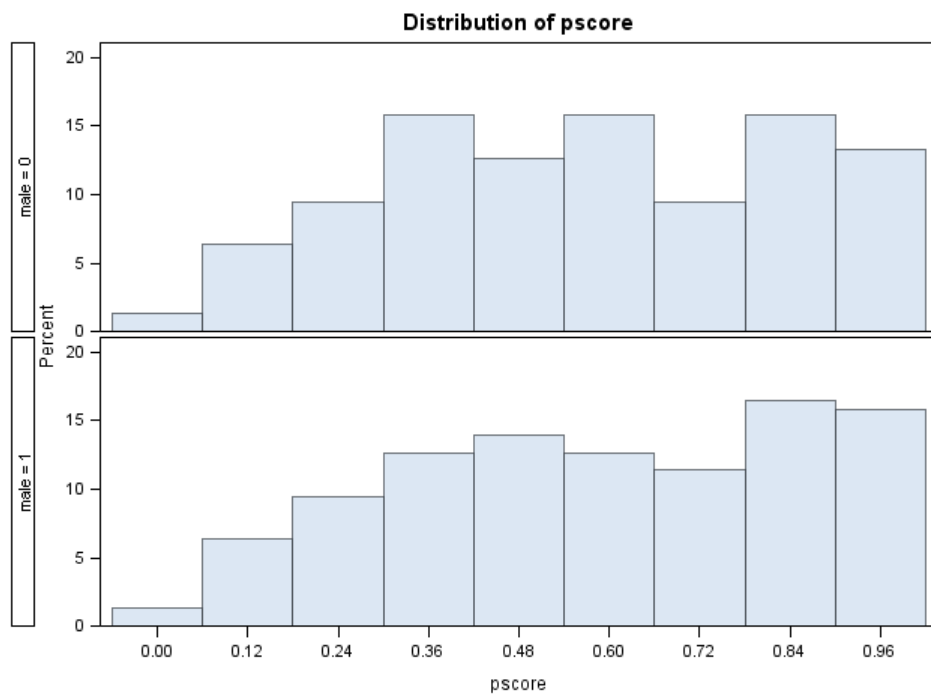
Make histogram of propensity scores by treatment group to check balance.

Full data ( $n = 1118$ ), no matching:



25

Matched data ( $n = 366$ )



26

After separating into T and C datasets, now we must separate T-C pairs into separate observations. (Make 2 observations from 1; see LSB §6.10.)

```
data pairs_cal;
  set PS_match_cal;
  id = IdSelectedControl; pscore = PScoreControl;
  male=0; pair = _N_; pair is observation number
  output;
  id = MatchedToTreatID; pscore = PScoreTreat;
  male=1; pair = _N_;
  output;
  keep id pair male;
```

27

Then merge pairing back with full data.

```
proc sort data=pairs_cal;
  by id;
proc sort data=ph6470.whrdata;
  by id;

data matched_pairs;
  merge pairs_cal ph6470.whrdata;
  by id;
  if (pair NE .); restrict to matched pairs
```

28

**Check that propensity-score matching has worked:** no significant differences between groups. If necessary, repeat with additional variables in propensity score, smaller caliper for matching.

**Table 1. Descriptive Statistics by Sex**

	Men (n = 675)	Women (n = 510)	P
Intensive therapy during DCCT	327 (48)	261 (51)	0.352
Age (y)	38 ± 7	38 ± 7	0.341
Duration of diabetes (y)	15.9 ± 5	16.5 ± 5	0.046
HbA <sub>1c</sub> (%)	8.2 ± 1.3	8.3 ± 1.5	0.537
AER (mg/24 h)	16.2 (14.6-18.1)	11.9 (10.8-13.2)	<0.001
Subjects with elevated AER	136 (20)	81 (16)	0.054
BMI (kg/m <sup>2</sup> )	26.7 ± 4	26.1 ± 4	0.008
WHR	0.89 ± 0.06	0.78 ± 0.06	<0.001
Waist circumference (cm)	90.4 ± 10	80.0 ± 10	<0.001
SBP (mm Hg)	122 ± 13	117 ± 13	<0.001
DBP (mm Hg)	78 ± 9	75 ± 8	<0.001
Serum creatinine (mg/dL)	1.0 ± 0.1	0.8 ± 0.1	<0.001
Total cholesterol (mg/dL)	188 ± 37	190 ± 37	0.449
Triglycerides (mg/dL)	96 ± 67	84 ± 66	0.003
HDL cholesterol (mg/dL)	50 ± 12	61 ± 14	<0.001
LDL cholesterol (mg/dL)	119 ± 31	112 ± 30	<0.001
LDL/HDL cholesterol ratio	2.5 ± 0.9	2.0 ± 0.8	<0.001
Smoking status (yes)	131 (19)	99 (19)	0.999

29

	Men (n = 183)	Women (n = 183)	P-value
Age (y)	37.4 ± .5	36.9 ± .5	.528
Duration of diabetes (y)	15.9 ± .3	16.2 ± .3	.665
HbA1c (%)	8.3 ± .1	8.4 ± .1	.861
BMI	25.7 ± .3	27.6 ± .3	<.001
WHR	0.84 ± .004	0.83 ± .004	.043
SBP (mm Hg)	120 ± 1	119 ± 1	.476
DBP (mm Hg)	77 ± .6	76 ± .6	.116
Total Cholesterol (mg/dL)	182 ± 3	197 ± 3	.003
Triglycerides (mg/dL)	85 ± 6	101 ± 6	.062
HDL (mg/dL)	52 ± 1	56 ± 1	.012
LDL (mg/dL)	113 ± 2	120 ± 2	.039
LDL/HDL ratio	2.3 ± .06	2.3 ± .06	.911
Smoking status (yes)	20%	17%	.504

Mean ± SE; highlighted variables used in propensity score

30

### 3. Planned analysis, restricted to paired data

*Question:* Is there a gender difference in AER in this population of type I diabetics, adjusting for WHR, age, LDL/HDL ratio, HBA1c level, duration of diabetes, smoking status, and SBP?

```
Proc GLM data=matched_pairs ;  
  title3 "Matched pairs propensity scores";  
  class male;  
  model log_AER =  
    male WHR age ldl_hdl HBA1C duration smoking SBP;  
  lsmeans male / CL pdiff;
```

*Do not include pair in the model.*

31

Analysis restricted to about 1/3 of the data.

#### The GLM Procedure

```
Number of Observations Read    366  
Number of Observations Used    313
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
male	1	0.12405750	0.12405750	0.09	0.7699
WHR	1	4.20048430	4.20048430	2.90	0.0894
AGE	1	10.81919991	10.81919991	7.48	0.0066
ldl_hdl	1	11.33429034	11.33429034	7.83	0.0055
HBA1C	1	23.30375953	23.30375953	16.11	<.0001
DURATION	1	6.01581291	6.01581291	4.16	0.0423
SMOKING	1	0.01380717	0.01380717	0.01	0.9222
SBP	1	46.33839016	46.33839016	32.03	<.0001

Most of the adjustors are still significant.

32

*Question:* Is there a gender difference in AER in this population of type I diabetics, adjusting for WHR, age, LDL/HDL ratio, HBA1c level, duration of diabetes, smoking status, and SBP?

Matched pairs from propensity scores				H0:LSMean1=
Least Squares Means				LSMean2
male	log_aer	95% Confidence Limits		Pr >  t
	LSMEAN			
0	2.642726	2.453707	2.831746	0.7699
1	2.602861	2.413236	2.792486	

These are gender means for log(AER), adjusted for the whole list, from data balanced using propensity scores.

33

Alternative matching procedure, which sometimes gives larger matched sample:

1. Compute propensity scores
2. *Divide propensity scores into deciles (0, 1, 2, ..., 9)*
3. *Match on the deciles, instead of the propensity scores*
4. Perform planned analysis on matched data, but ignore the pairing in the analysis.

34

```

Proc Rank data=AER_P out=AER_P groups=10;
    var pscore;
    ranks pscore_decile;
data T1 C1;
    set AER_P;
    if male=0 then do;
        idC = id; pscoreC = pscore_decile; output C1;
    end;
    if male=1 then do;
        idT = id; pscoreT = pscore_decile; output T1;
    end;
%PSMatching(datatreatment= T1, datacontrol= C1, method= caliper,
numberofcontrols= 1, caliper=.25 , match within decile
replacement=no, out=PS_match_decile);

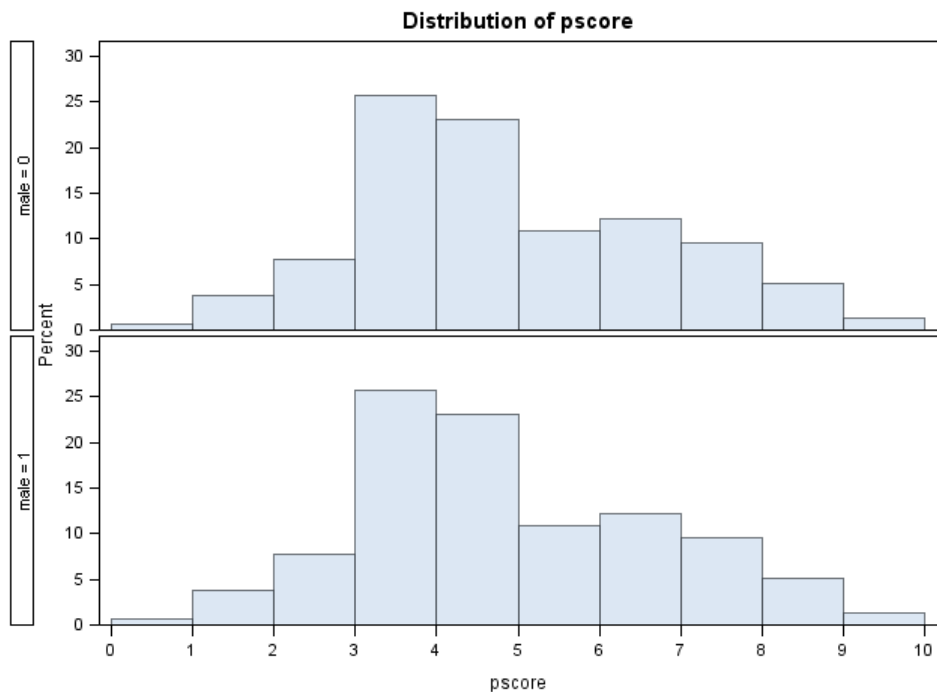
```

35

Obs	Id		Matched	
	Selected Control	PScore Control	To TreatID	PScore Treat
1	5010	5	27198	5
2	26101	9	13175	9
3	41145	7	12015	7
4	17240	8	6233	8
5	22208	7	14166	7
6	26096	8	9145	8
7	20006	7	12152	7
8	24018	5	7152	5
9	16141	3	19165	3
10	1384	4	21199	4

36

Matched sample using deciles ( $n = 362$ )



37

Gives very similar answer:

Least Squares Means

	log_aer	LSMean2	H0:LSMean1=
male	LSMEAN	Pr >  t	
0	2.64242411	0.7452	
1	2.59933335		

male	log_aer	95% Confidence Limits	
	LSMEAN		
0	2.642424	2.458529	2.826319
1	2.599333	2.414841	2.783825

38

## References

Gelman and Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, §10.1–10.3

RB D’Agostino: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statist. Med.* 1998; 17, 2265–2281.

PR Rosenbaum and DB Rubin: The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983; 70: 41–55.

Description of matching macro:

M. Coca-Perraillon (1987) “Local and global optimal propensity score matching.”

Large literature on causal inference from observational studies:  
see §10.8 in Gelman and Hill.