

Lecture 23

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
7. Matching deciles of 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:

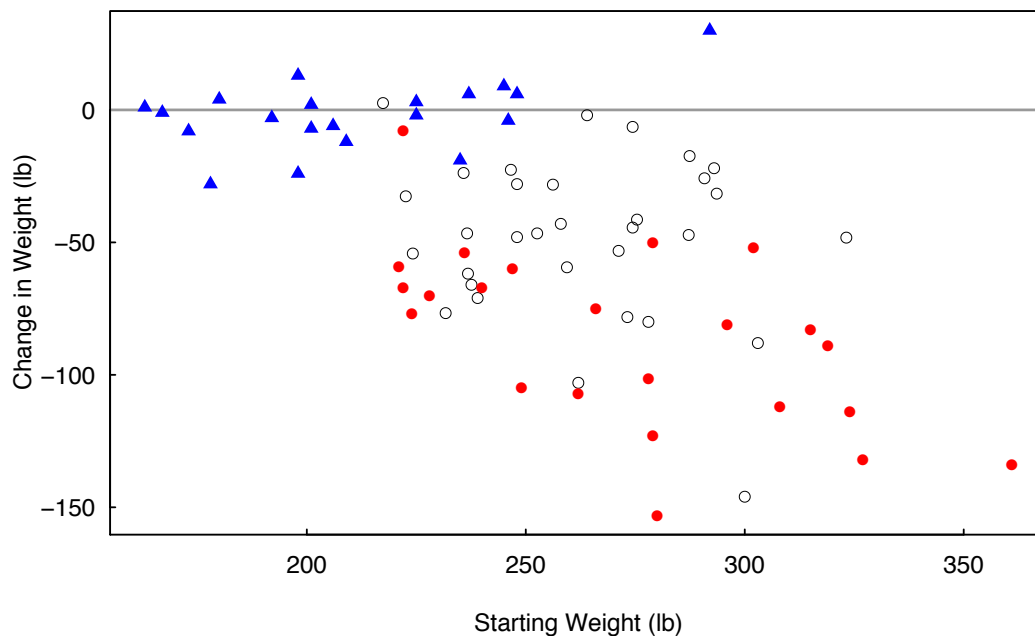
show subjects in treatment groups alike

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.

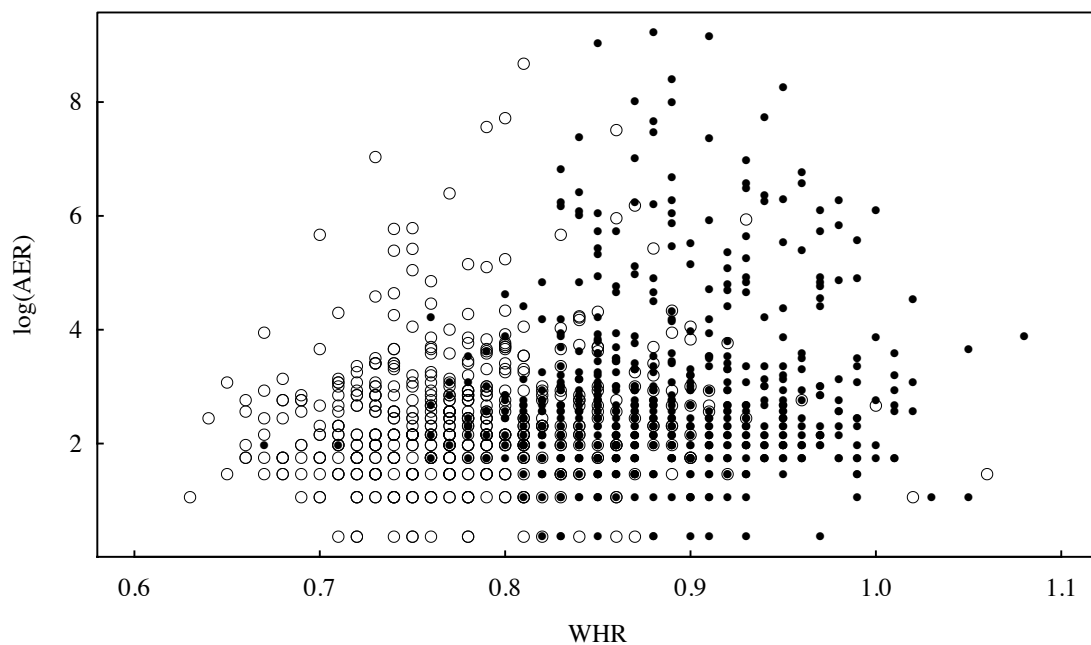
1. Use regression (weight change on baseline weight) to adjust for baseline weight.
2. Stratify on baseline weight, use strata that contain large enough samples from at least 2 treatment groups.

Stratification is difficult with small samples and minimal overlap.

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

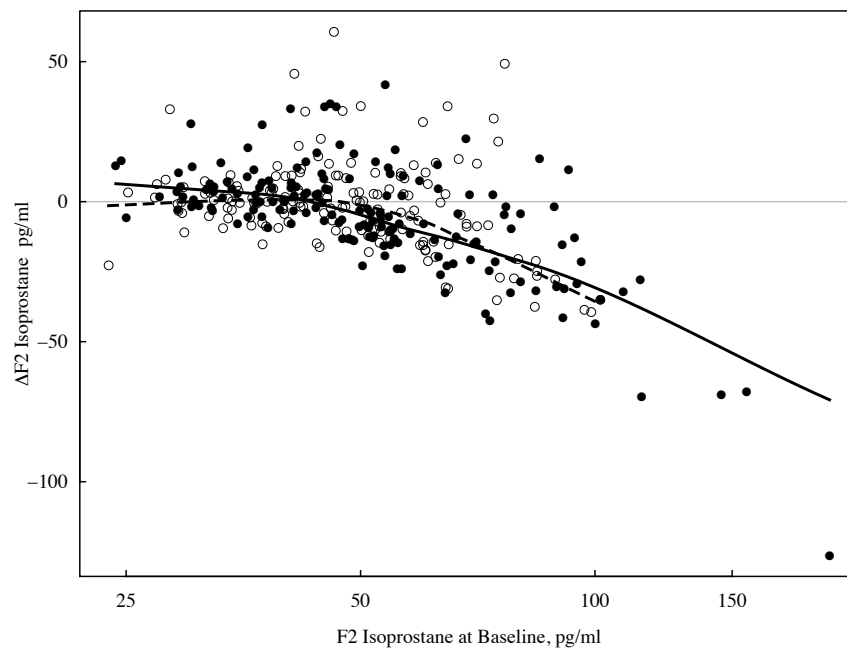
5

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:

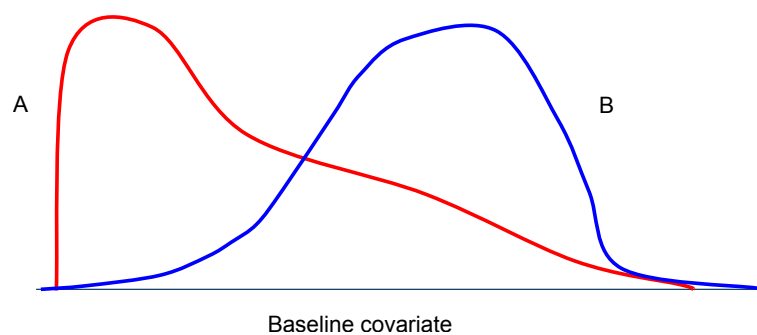


7

Differences in baseline characteristics

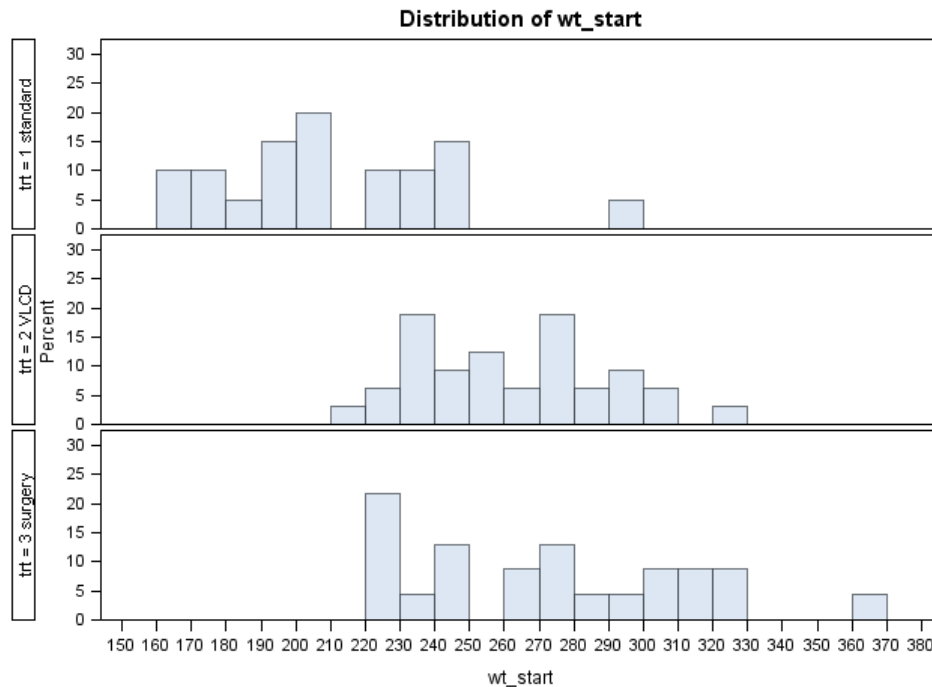
Baseline characteristics may differ between observational treatment groups

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



8

Limited overlap of range: gastric-bypass baseline weights



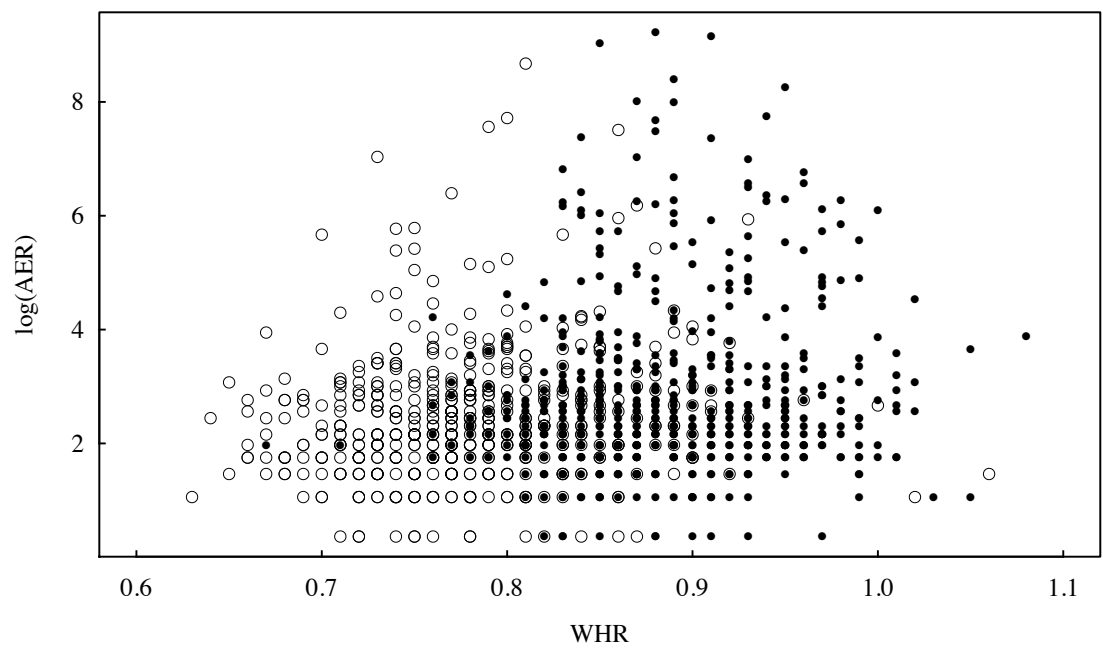
9

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";

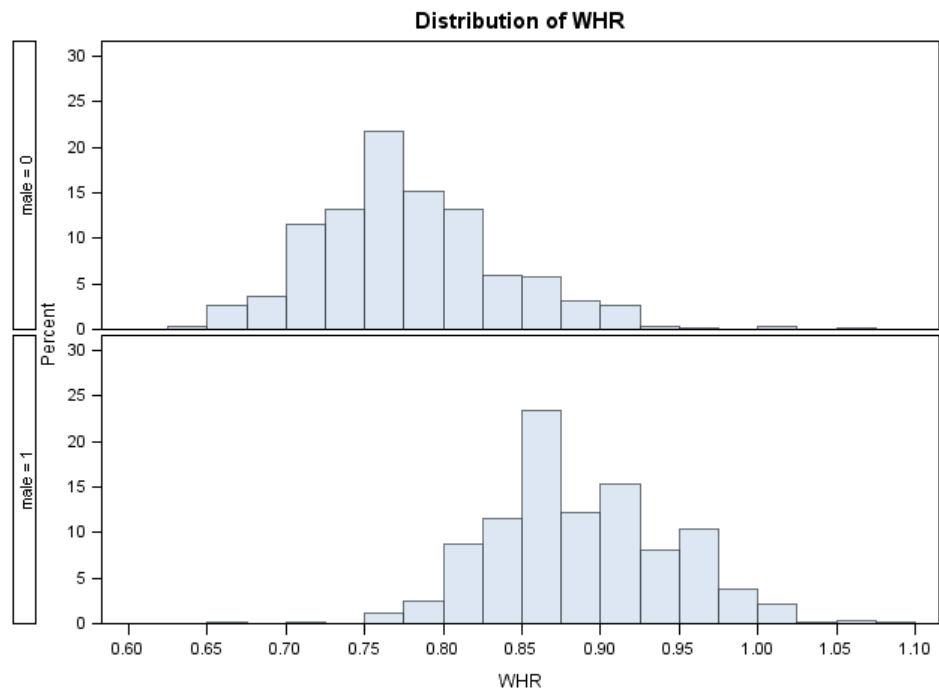
ODS graphics on;
Proc Univariate noprint data= rename;
  var wt_start;
  class trt;
  histogram wt_start / nrows=3 endpoints = 150 to 380 by 10;
run;
ODS graphics off;
```

Question: Is there a gender difference in AER in this population of type I diabetics, adjusting for WHR?



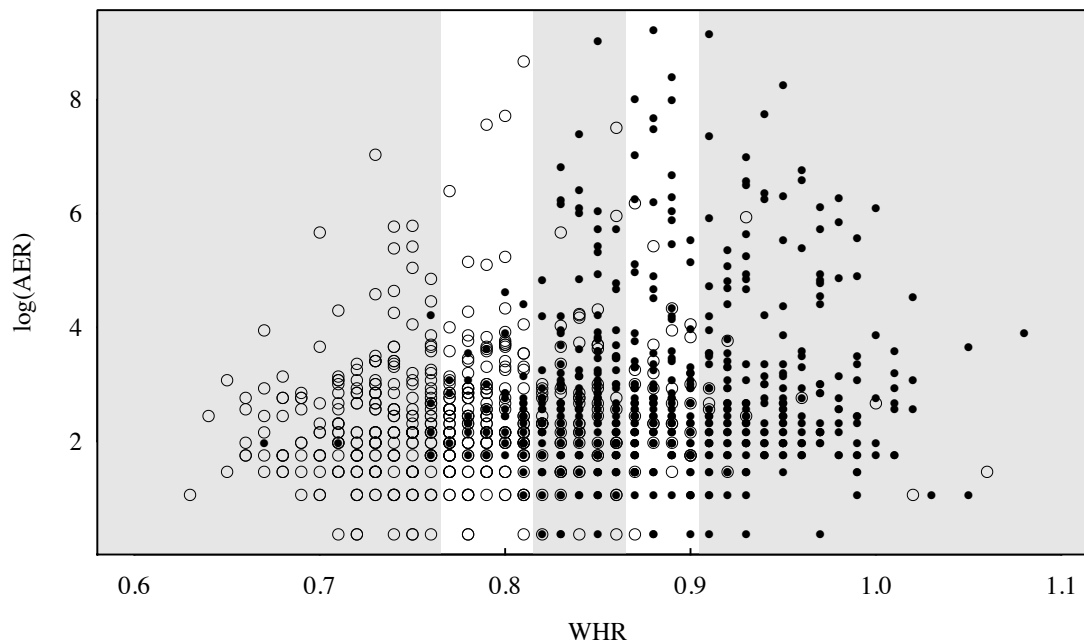
11

Limited overlap, different distributions:



12

Divide into quintiles of WHR, restrict comparison to “balanced” middle quintiles:



13

```
Proc Rank data=AER out=AER_1 groups=5 ;
    var WHR;          variable to rank--do only 1 at a time
    ranks QWHR;       name of new variable holding ranks
```

Adds rank variable to data set.

Missing values in any variables in VAR list not ranked.

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 very close. Also ignores question of balance in other predictors.

14

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 _{1c} (%)	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 ²)	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

NOTE. N = 1,185. Values expressed as mean ± SD or number (percent); for AER, geometric mean and 95% confidence interval are given. To convert serum creatinine in mg/dL to μ mol/L, multiply by 88.4; cholesterol (total, LDL, or HDL) in mg/dL to mmol/L, multiply by 0.02586; triglycerides in mg/dL to mmol/L, multiply by 0.01129.

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

15

From Table 1, significant baseline differences between genders in duration of diabetes, BMI, WHR, SBP, DBP, creatinine, triglycerides, HDL, LDL.

Investigator's question: Adjusting for WHR, age, LDL/HDL ratio, HBA1c level, duration of diabetes, smoking status, and SBP, is there a gender difference in AER?

- 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.

16

Propensity Score

1. Compute each subject's **propensity score** = probability (logit scale) that subject is male, given duration of diabetes, BMI, WHR, SBP, DBP, creatinine, triglycerides, HDL, LDL (significant baseline differences between genders).
Also common: use all baseline predictors + interactions.
Use logistic regression, output fitted probabilities on logit scale (approx normally distributed).
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.

17

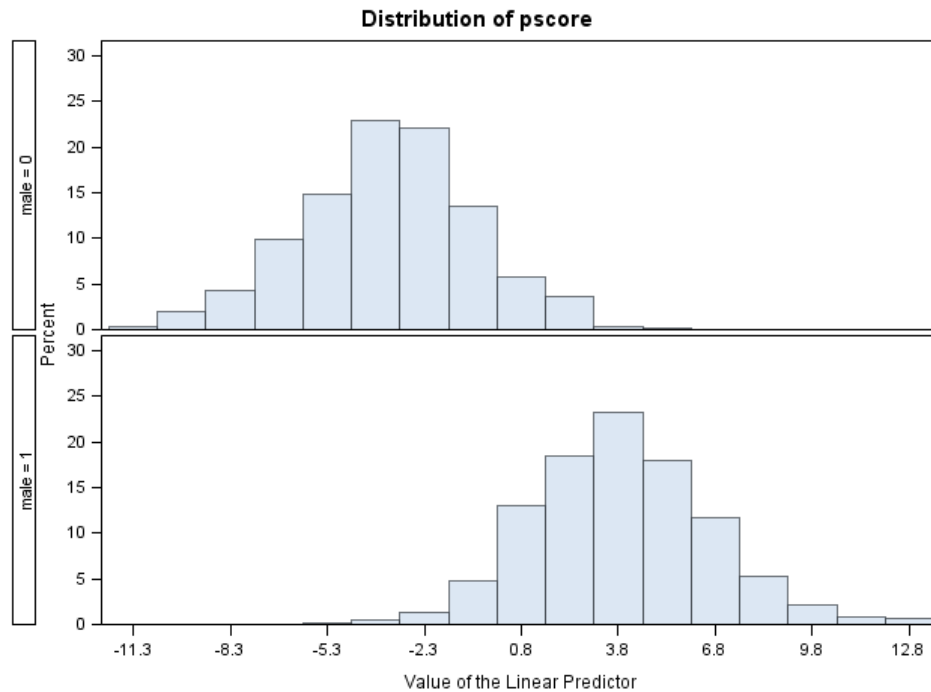
Calculate propensity scores in logistic regression:

```
Proc Logistic descending data=ph6470.whrdata;  
  model male = duration BMI WHR SBP DBP REN_SCR TRIG HDL LDL;  
  output out=AER_P xbeta =pscore ;  
  xbeta gives prediction on logit scale = propensity score
```

```
ODS graphics on;  
Proc Univariates noprint data= AER_P ;  
  var pscore;  
  class male;  
  histogram pscore / nrows=2 ;  
  run;  
ODS graphics off;
```

18

Propensity scores have very different distributions by gender



19

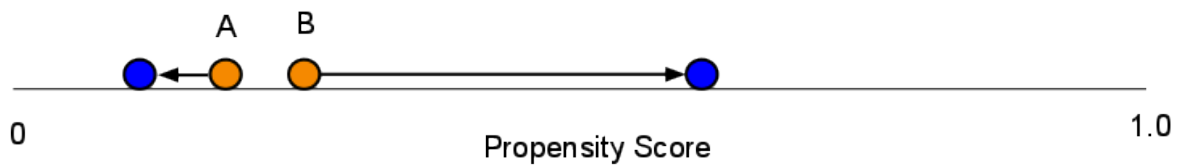
2. Form matched pairs

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

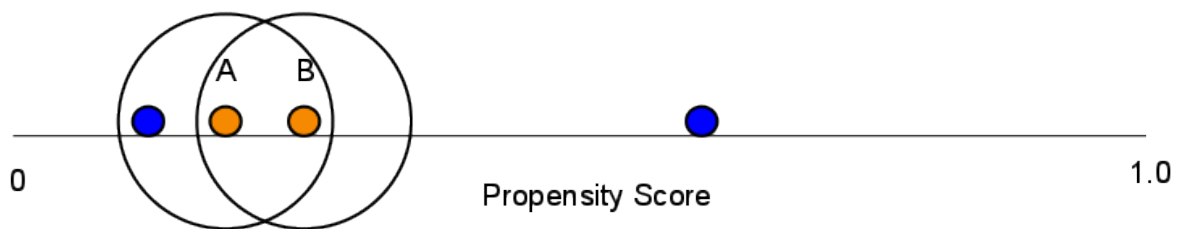
Make separate Treatment and Control datasets such that:

- Control dataset includes: idC = subject_id, pscoreC = propensity score
Treatment dataset includes: idT, pscoreT
Other variables are discarded by the macro.
- Macro offers two methods of matching: caliper or nearest neighbor (NN)

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

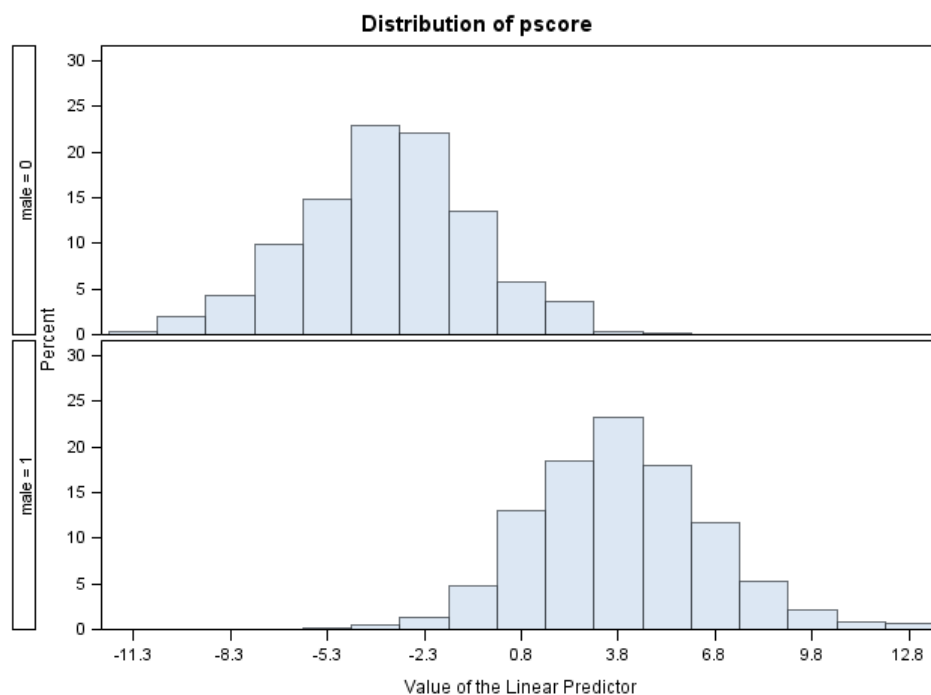


Better: Caliper matching: matches nearest within caliper = maximum distance. In diagram, there's no match for observation B.



21

Use caliper matching for propensity scores. How large should the caliper be?



22

Make two separate datasets:

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

23

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

24

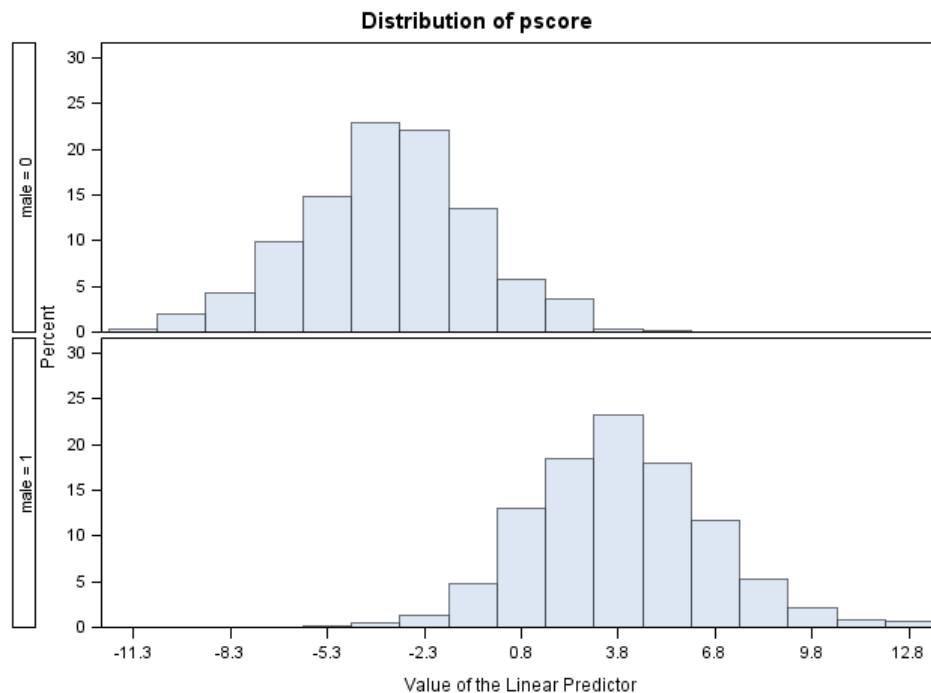
Obs	Id		Matched	
	Selected	PScore	To	PScore
	Control	Control	TreatID	Treat
1	23014	0.67726	27198	0.69198
2	26096	2.20864	12015	2.30824
3	27213	1.76857	14166	1.71937
4	1246	0.74520	9145	0.74557
5	26101	4.46764	12152	4.39405
6	22128	0.48823	7152	0.46361
7	1024	0.77259	18006	0.74018
8	10026	0.49336	20096	0.45877
9	17240	4.08784	2124	4.18447
10	19012	5.47371	15161	5.55885

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

25

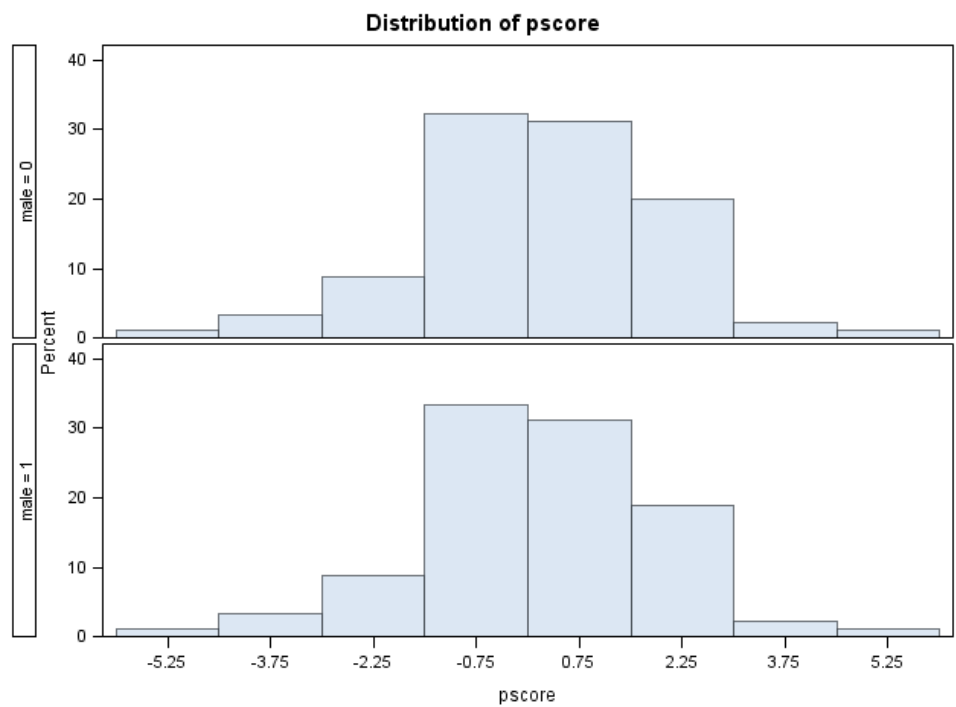
Make histogram of propensity scores by treatment group to check balance.

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



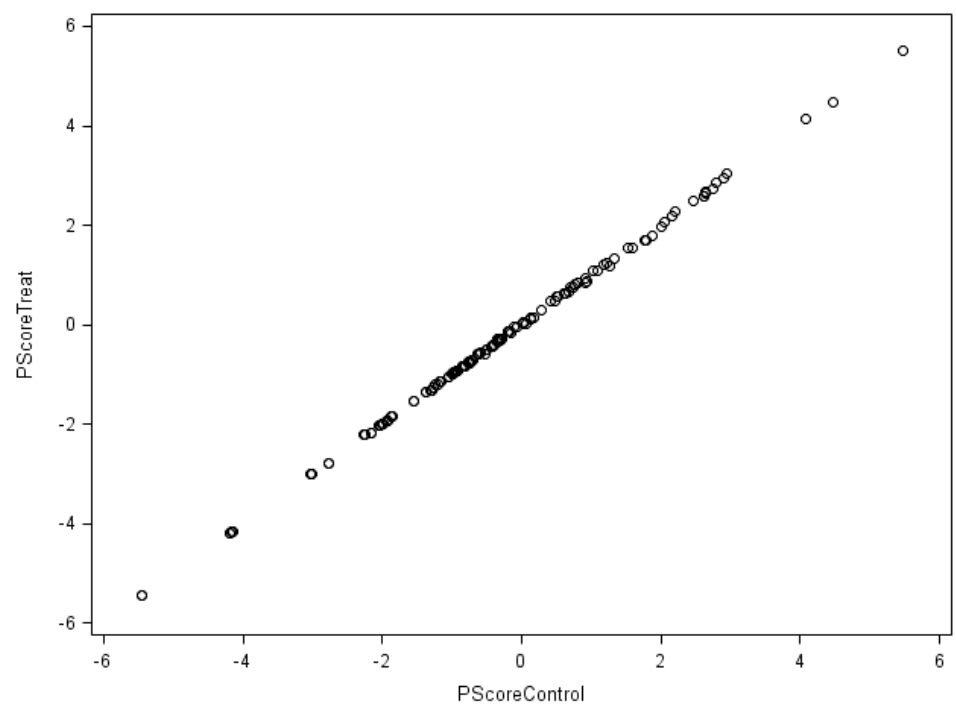
26

Matched data ($n = 228$, 21% of original data)



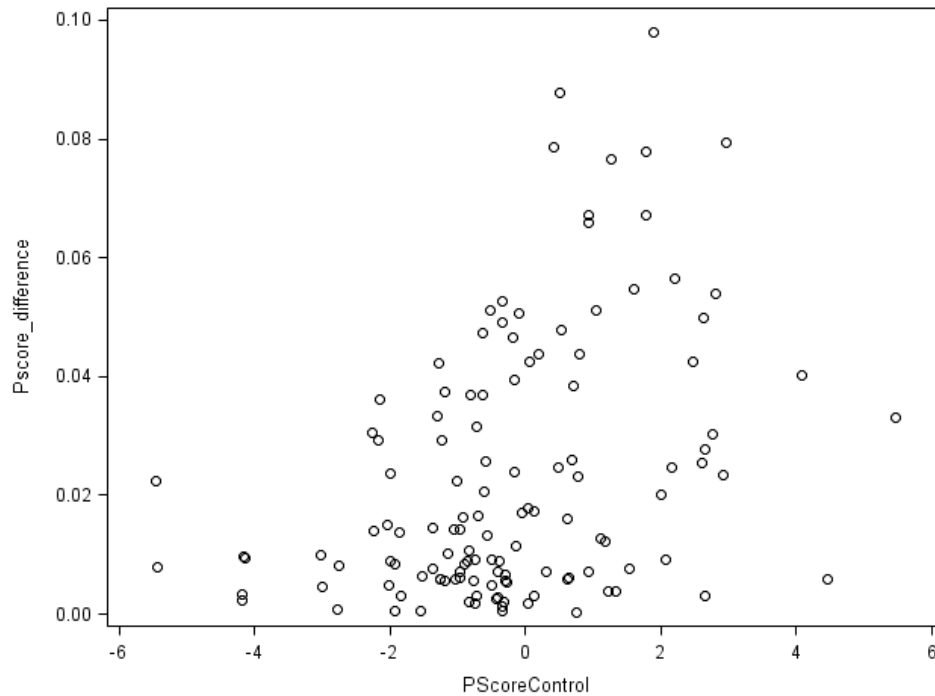
27

Plot pairs of matched propensity scores:



28

Detrended plot of absolute differences between matched propensity scores:



29

Obs	Id		Matched	
	Selected	PScore	To	PScore
	Control	Control	TreatID	Treat
1	23014	0.67726	27198	0.69198
2	26096	2.20864	12015	2.30824
3	27213	1.76857	14166	1.71937

Make two separate observations from each T-C pairs. (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;
```

30

Then merge paired data 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
```

31

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 _{1c} (%)	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 ²)	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

NOTE. N = 1,185. Values expressed as mean ± SD or number (percent); for AER, geometric mean and 95% confidence interval are given. To convert serum creatinine in mg/dL to μ mol/L, multiply by 88.4; cholesterol (total, LDL, or HDL) in mg/dL to mmol/L, multiply by 0.02586; triglycerides in mg/dL to mmol/L, multiply by 0.01129.

32

	Men (n = 114)	Women (n = 114)	P-value
Age (y)	36.3 ± 7	38.0 ± 7	.072
Duration of diabetes (y)	15.7 ± 5	16.0 ± 5	.662
HbA1c (%)	8.3 ± 1.5	8.3 ± 1.4	.815
BMI	26.1 ± 4	26.5 ± 4	.501
WHR	0.84 ± .05	0.83 ± .07	.077
SBP (mm Hg)	120 ± 11	119 ± 13	.609
DBP (mm Hg)	77 ± 9	75 ± 8	.298
Serum creatinine (mg/dL)	.88 ± .12	.86 ± .11	.085
Total Cholesterol (mg/dL)	187 ± 32	193 ± 46	.281
Triglycerides (mg/dL)	89 ± 79	105 ± 115	.229
HDL (mg/dL)	56 ± 16	55 ± 14	.453
LDL (mg/dL)	113 ± 27	115 ± 30	.684
LDL/HDL ratio	2.18 ± .9	2.22 ± .9	.767
Smoking status (yes)	23%	21%	.745

Mean ± SD; highlighted variables used in propensity score

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.

Analysis restricted to about 1/5 of the original data.

The GLM Procedure

Number of Observations Read	228
Number of Observations Used	204

Source	DF	Type III SS	Mean Square	F Value	Pr > F
male	1	0.20300451	0.20300451	0.18	0.6690
WHR	1	3.14422209	3.14422209	2.84	0.0935
AGE	1	7.40515567	7.40515567	6.69	0.0104
HBA1C	1	29.69799901	29.69799901	26.83	<.0001
DURATION	1	4.33733789	4.33733789	3.92	0.0492
SMOKING	1	6.78536623	6.78536623	6.13	0.0141
SBP	1	30.48229714	30.48229714	27.54	<.0001

Most of the adjustors are still significant.

35

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?

		H0:LSMean1=LSMean2	
		Pr > t	
male	log_aer LSMEAN		
0	2.58776245	0.6690	
1	2.65242871		

		95% Confidence Limits	
male	log_aer LSMEAN		
0	2.587762	2.381809	2.793716
1	2.652429	2.442295	2.862562

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

36

Alternative matching procedure, which sometimes gives larger matched sample:

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

37

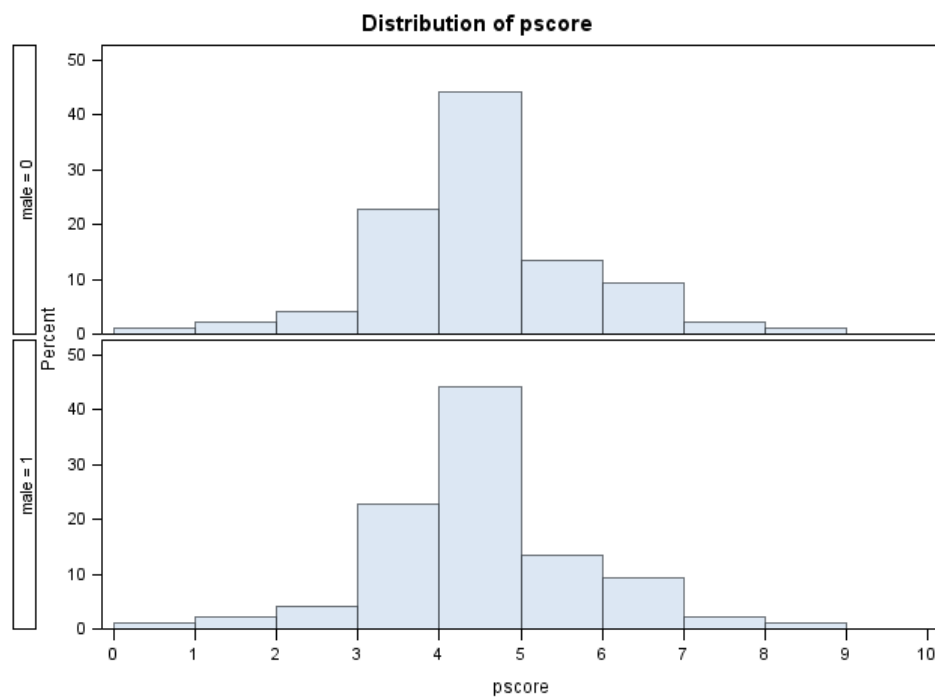
```
Proc Rank data=AER_P out=AER_P groups=10; start with propensity scores
    var pscore;
    ranks pscore_decile;
data T1 C1;
    set AER_P;
    if male=0 then do;
        idC = id; pscoreC = pscore_decile; output C1;
    end;
    if male=1 then do;
        idT = id; pscoreT = pscore_decile; output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

38

Obs	Id		Matched	
	Selected	PScore	To	PScore
	Control	Control	TreatID	Treat
1	14306	4	27198	4
2	4232	6	13175	6
3	2008	5	12015	5
4	17240	7	6233	7
5	26096	5	14166	5
6	41145	4	9145	4
7	26101	7	12152	7
8	19015	4	7152	4
9	16244	6	21199	6
10	10026	4	18006	4

39

Matched sample using deciles ($n = 117$ from each gender)



40

Gives very similar answer: no gender difference in AER, after adjustment.

		H0:LSMean1=
		LSMean2
male	log_aer LSMEAN	Pr > t
0	2.78541696	0.4589
1	2.67425579	

		95% Confidence Limits	
male	log_aer LSMEAN		
0	2.785417	2.581162	2.989672
1	2.674256	2.466120	2.882391

41

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.

42