Causal inference foundations and applications in environmental health sciences

Jay S. Kaufman, PhD
Dept of Epidemiology, Biostatistics & Occupational Health
McGill University, Montreal, Quebec, Canada

Sunday August 26th
Shaw Centre, Ottawa
Room 4
13:00 – 16:00

Outline for Sunday August 26th, 2018

Jay Kaufman, McGill University	**13:00 – 14:25 (85 min)**
Break	**14:25 – 14:35 (10 min)**
Alexander Keil, UNC-CH	**14:35 – 16:00 (85 min)**

1st Half Topic:	**85 minutes => 114 slides**

Causation vs. Association, DAGs and Confounding	(3 - 21)
Propensity Scores	(23 - 48)
IPTW/Marginal Structural Models/G-Formula	(49 - 116)

Hernán & Robins. *Causal Inference,* in press.
https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Causal inference is necessary for medical and public policy decision-making because we hope to optimize some outcome.

Causal inference is about inherently unobservable things (i.e. the future under different scenarios)

Because we can't directly observe what we want to know, we model it.

Good models
Bad models

From 1999 to 2009, the number of Americans who fell into a swimming pool and drowned each year is correlated with the number of films in which Nicholas Cage appeared that year.

Shall we reduce the number of pool drownings by keeping Cage off the screen?

Three main inferential targets of these models:

1) Real world in the present

surveillance, descriptive study

2) Real world in the future

clinical prediction model

3) Hypothetical world in the future

causal inference, etiologic study

The inferential target determines the adjustment strategy.

Most people here are interested in 3)

If you are trying to estimate the causal effect of a treatment, your job is to PREDICT what would happen in the FUTURE if you did thing A compared to what would happen if you did thing B.

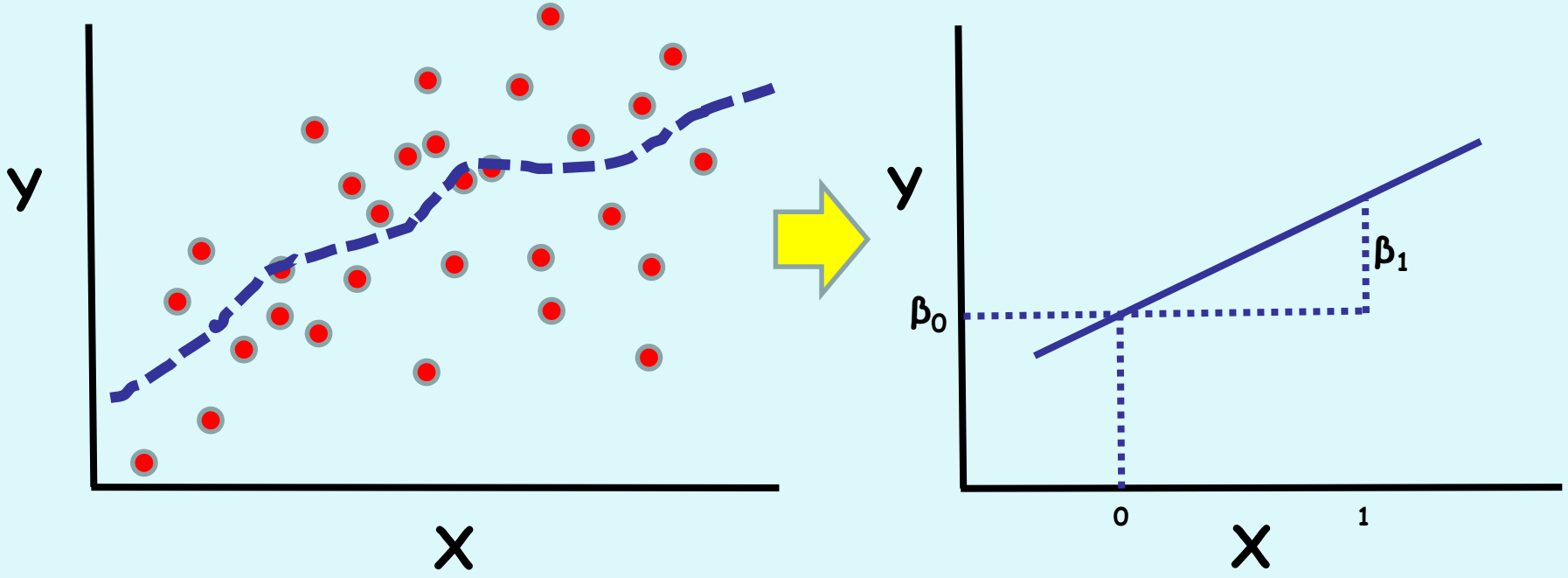
To do this from observational data, you must often adjust statistically for factors that are associated with the treatment and the outcome.

You observe: $Pr(Y|X=x)$

You want to know: $Pr(Y|SET[X=x])$

This is the intervention you want to know about, but unfortunately you don't really get to "SET" anything.

Statistical models are used to estimate relationships between variables in observational data sets.



But it is mechanistic knowledge or structural assumptions that allow us to infer causal effects from these relationships (not statistical considerations)

Read: $Pr(Y|SET[X=x])$

as: $Pr(Y|SET[X=x_1])$ versus $Pr(Y|SET[X=x_2])$

x_1 and x_2 are the levels at which you intervene to set the treatment; contrast is usually a difference or ratio.

Causal inference from passively observed data requires not only structural identification, but also:

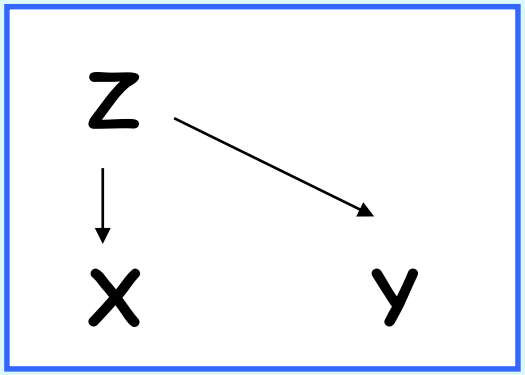
positivity (there are sufficient data available on the treatment and outcome in the range of interest)

consistency (the way that people came to be treated in the data set is comparable to the way that you plan to treat them in your intervention)

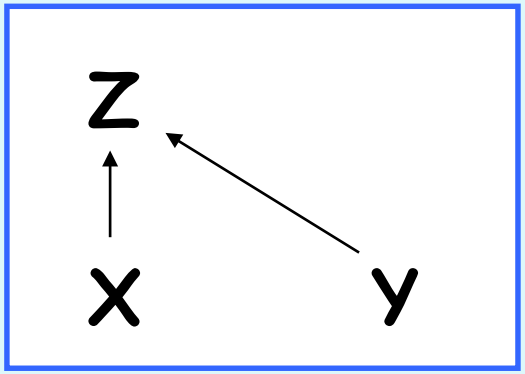
correct specification of statistical models

Three main structural threats to validity:

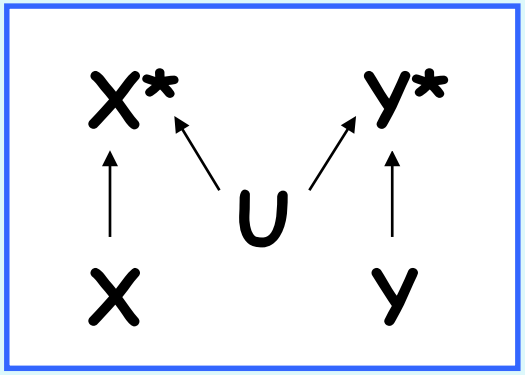
Confounding Bias



Selection Bias



Information Bias



Identification versus estimation

If you can't get be guaranteed to get the right answer as $n \to \infty$, then you have an identification problem.

DAGs are all about expected values, and therefore are focused entirely on identification.

This has some drawbacks for application to real-world studies with $n \lll \infty$.

For example, RCT DAG shows no confounding, but an RCT can, by bad luck, have an imbalanced covariate. This is confounding, even though there is no confounding in the correctly specified DAG.

Statistical Adjustment:

"The philosophers have only interpreted the world, the point, however, is to change it." --- Karl Marx

There are two kinds of analysis:

descriptive and etiologic (causal)

If you are doing descriptive analysis, you show a picture of the world as it really is. No "adjustments". Why not?

Because the real world is unadjusted.

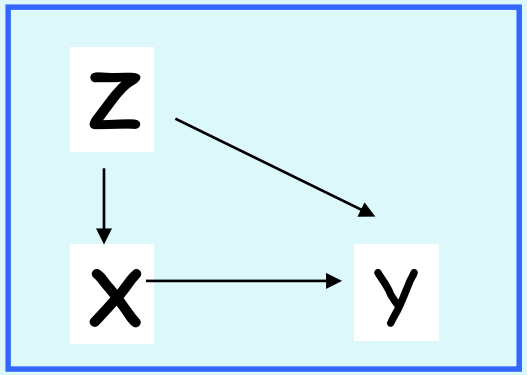
If you are doing an etiologic (causal) analysis, your job is to identify what would happen if you intervened on the world in some specific way.

In order to do this from observational data, you must often adjust statistically for factors that are associated with the exposure and outcome under study.

You observe: $Pr(Y|X=x)$

You want to know: $Pr(Y|SET[X=x])$

The adjustment tradition in epidemiology and the social sciences exists to link these two quantities:



$$Pr(Y|X=x) \neq Pr(Y|SET[X=x])$$

BUT!

$$\Sigma Pr(Y|X=x, Z=z)Pr(Z=z) = Pr(Y|SET[X=x])$$

Read: $Pr(Y|SET[X=x])$

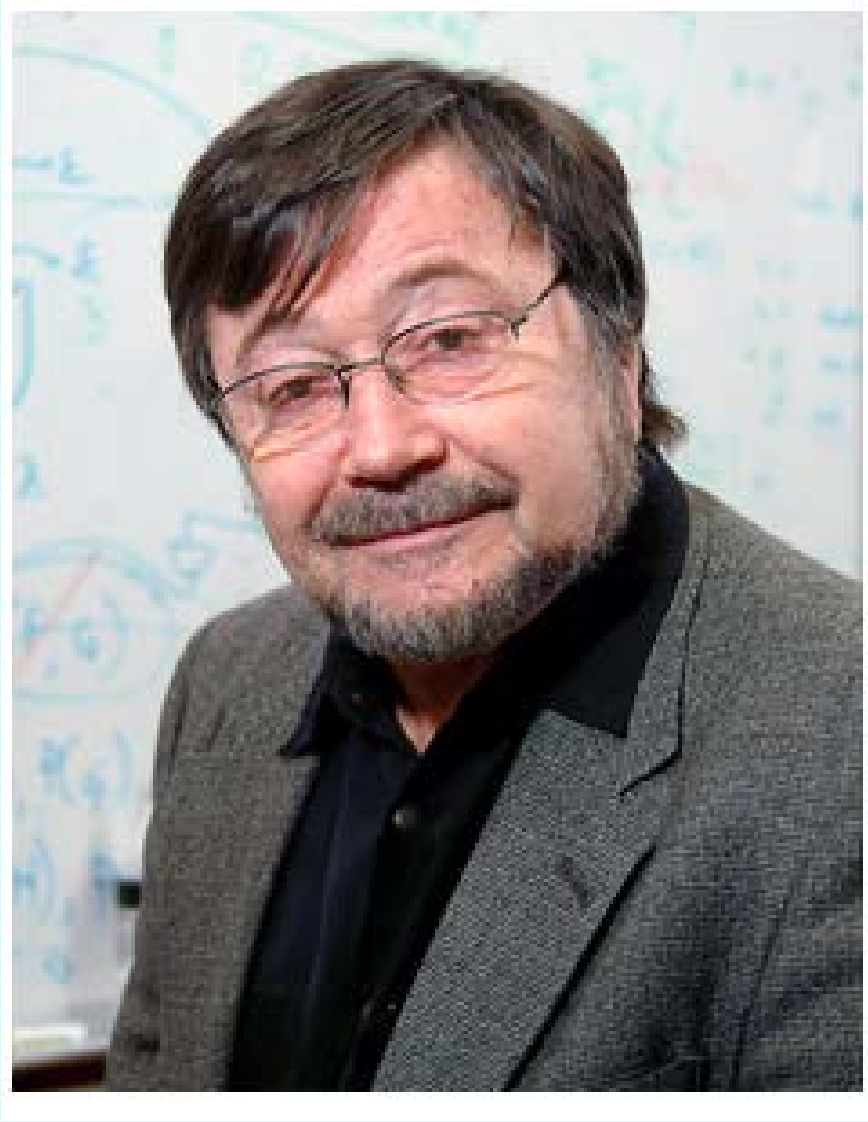
as: $Pr(Y|SET[X=x_1])$ **versus** $Pr(Y|SET[X=x_2])$

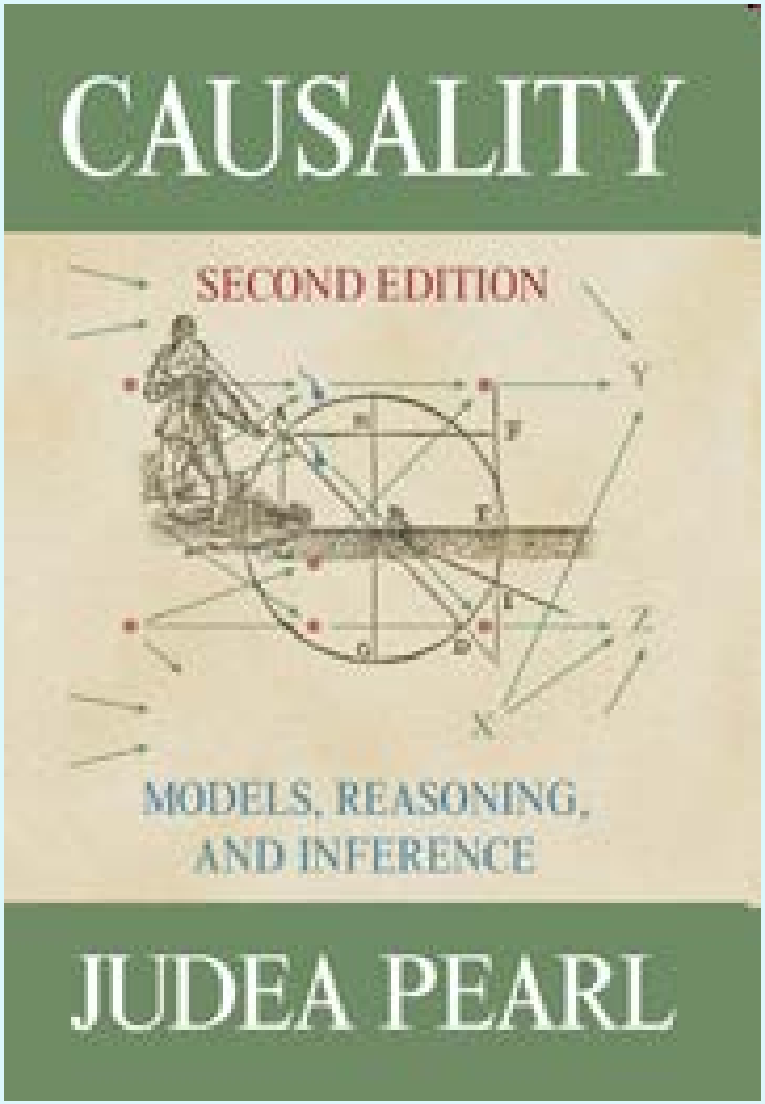
where x_1 and x_2 are two levels at which you can intervene to set the exposure, and the contrast is usually a difference or ratio.

Clearly, quantities intermediate between exposure and outcome are not "confounders", they are just the way that the exposure has the effect that it has.

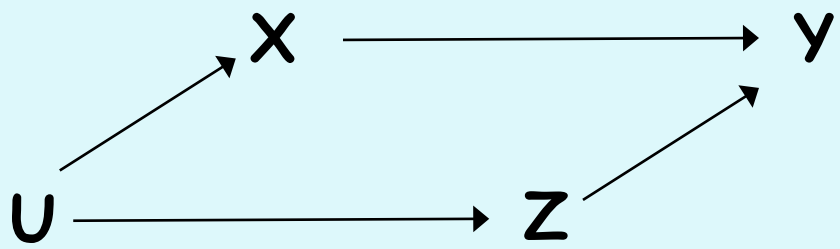
(See: Kaufman *American J of Law & Medicine* 2017
Schisterman et al *Epidemiology* 2009, etc)

Graphical language for encoding subject matter knowledge about causal structure





Compare a graphical model with a typical parametric epidemiologic model, such as logistic regression:



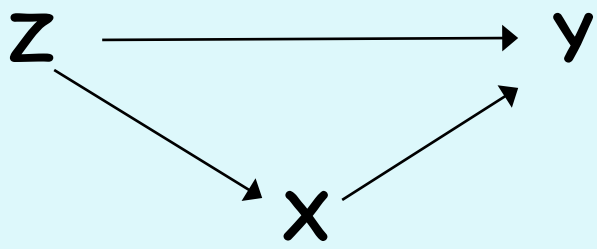
The graphical model asserts only that: $Y = f(X, Z, \varepsilon_y)$
and that $X = f(U, \varepsilon_X)$ and $Z = f(U, \varepsilon_Z)$

The logistic regression model:

$$E(Y|X,Z) = \left[\frac{e^{(\alpha+\beta_1 X+\beta_2 Z)}}{1+e^{(\alpha+\beta_1 X+\beta_2 Z)}} \right]$$

makes *MANY* assertions, including the multiplicative interaction of X and Z, and the linearity of the ln(odds) of Y across all values of X and Z.

Furthermore, a graphical model can represent many assumptions that cannot be encoded in a typical statistical model:



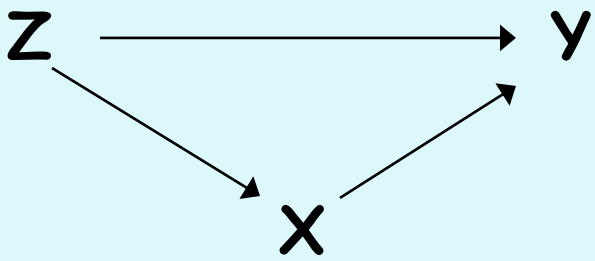
The graphical model asserts that: $X = f(Z, \varepsilon_X)$

The logistic regression model:

$$E(Y|X,Z) = \left[\frac{e^{(\alpha+\beta_1 X+\beta_2 Z)}}{1+e^{(\alpha+\beta_1 X+\beta_2 Z)}} \right]$$

cannot easily represent this constraint, even if it is know by the investigators to be true on subject matter grounds (e.g., Z = SEX, X = SMOKING)

Confounding

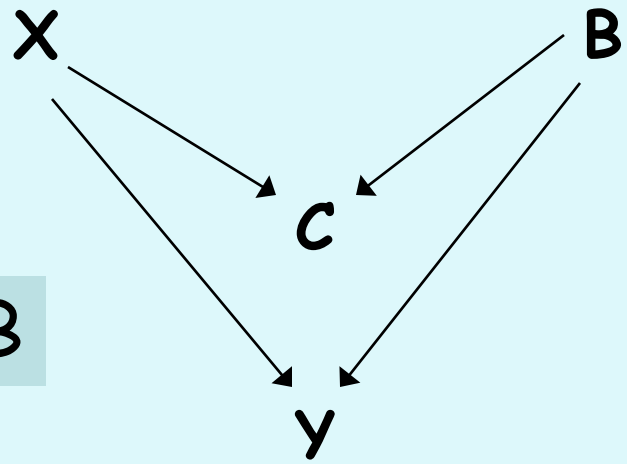


Confounding is a divergence between two kinds of conditional probability distributions of *Y*:

the distribution given that we find *X* at the value *x* (estimable from the data), and the distribution given that we intervene to force *X* to take the value *x*.

With Pearl's SET notation, express confounding as:

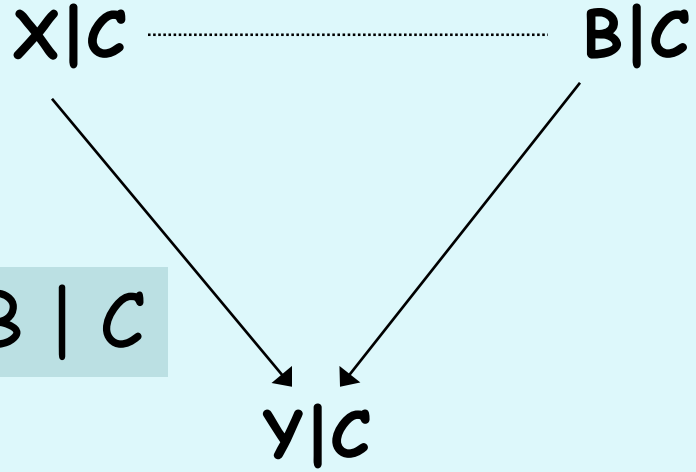
$$\Pr(Y = y \mid SET[X = x]) \quad \neq \quad \Pr(Y = y \mid X = x)$$



$X \perp\!\!\!\perp B$

NO CONFOUNDING FOR THE EFFECT OF X ON Y

No confounding in the source population when there is no backdoor path, or when a backdoor path is blocked by a collider.



$X \not\perp\!\!\!\perp B \mid C$

~~NO~~ CONFOUNDING FOR THE EFFECT OF X ON Y, GIVEN C

Most causal inference methods assume that you have no unmeasured confounders:

- Regression
- Propensity Scores
- Marginal Structural Models
- G-methods (SNMs, G-Formula, etc)

"Quasi-Experimental" Methods use structural assumptions to achieve identification even in the presence of unmeasured confounding:

- Instrumental variables
- Regression Discontinuity
- Fixed Effects Differences in Differences

Some causal inference methods achieve identification based on extrapolation of a parametric model.

Semi-parametric methods (e.g. propensity scores, IPTW, TMLE, etc) rely less on model form. Letting a computer pick the model reduces "wish bias".

Non-parametric methods (e.g. matching) require no model at all.

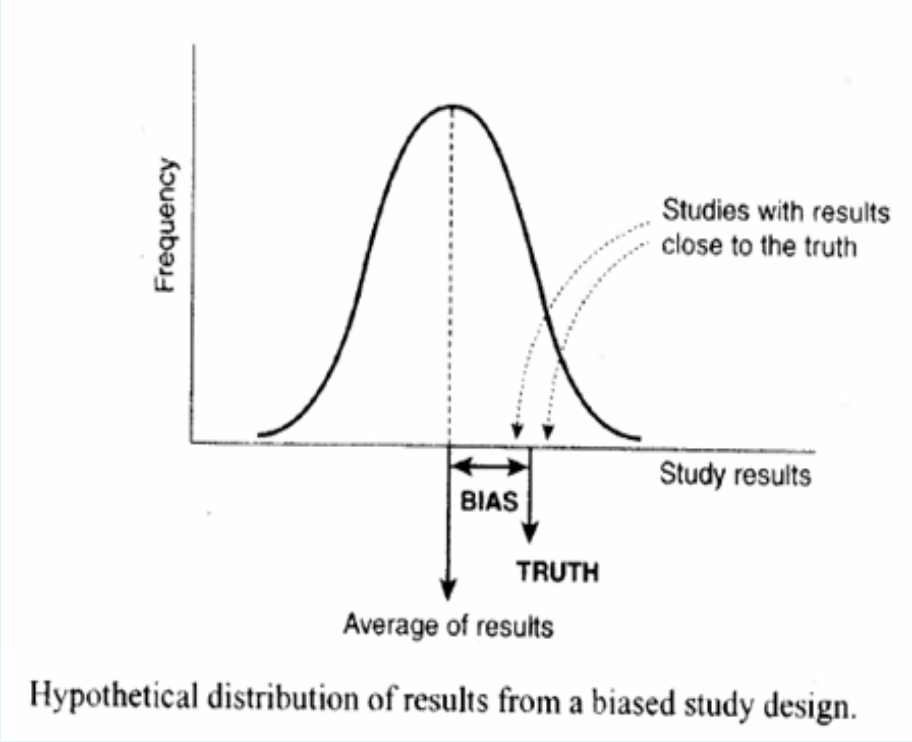
Doubly robust methods require that at least one model be right, but not both.

Computer intensive methods (e.g. bootstrapping) reduce reliance on distributional assumptions.

Confounding is a bias:

Validity and Bias: The epidemiologist's goal is the most VALID and PRECISE estimate possible of the causal effect of exposure on disease.

Error comes from sampling variability (lack of precision) and bias (lack of validity).



Hypothetical distribution of results from a biased study design.

Szklo & Nieto,
2nd edition, 2007

Big data context:

All usual threats to validity still apply:

confounding bias
selection bias
information bias

} None are reduced as n → large

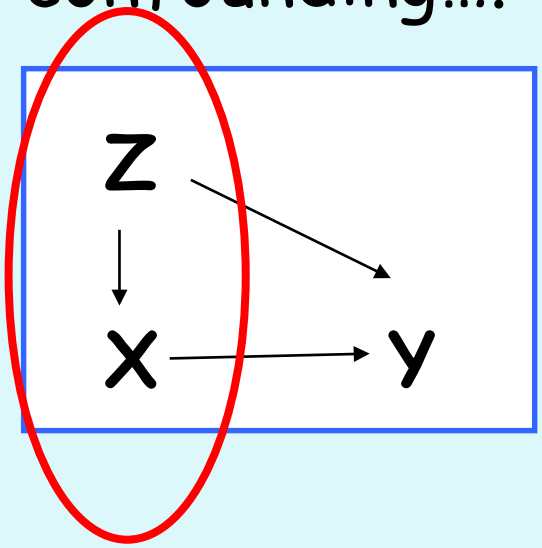
Big data means that random error should ↓

So for precision: BIGGER REALLY IS BETTER

Precision gain may not be "worth the price" if data quantity is negatively correlated with data quality (as often happens).

Exposure Modeling

OK, now back to confounding….



All confounding arrived via arrows into X.

So modeling receipt of treatment is a way to control for confounding. Pr(X=1|Z) is "the propensity score"

Rosenbaum, P.R. & Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70(1), 41-55.

- over 20,000 citations
- causal inference is a missing data problem
- defines a balancing score b(z), is a function of the observed covariates z such that the conditional distribution of x given b(x) is the same for treated (x = 1) and control (x = 0) units, that is: $z \perp\!\!\!\perp x \mid b(z)$.

Identifies the propensity score as the coursest possible balancing score.

"Weakly Ignorable Treatment Assignment" within strata of Z

Conditioning on Z sufficient adjustment to control for all confounding bias if, within each stratum of Z, observed exposure X is statistically independent of the potential response $(Y|SET[X=x])$, for each imposed value x.

Write: $Y_X \perp\!\!\!\perp X \mid Z$

where Y_x is the potential response of Y to treatment x
i.e., $(Y|SET[X=x])$

Target Populations:

Each subject i has a pair of potential outcomes: $Y_i(0)$ and $Y_i(1)$ for binary treatment that takes values 0 and 1.

Each subject i receives only one treatment, so 50% of values are missing.

For subject i, individual effect of treatment is $Y_i(1) - Y_i(0)$

Average treatment effect (ATE) = $E[Y_i(1) - Y_i(0)]$

An alternative target population is to consider the effect only in the exposed.

This is called the ATT (average treatment effect for the treated) or ETT (effect of treatment on the treated)

ATT is defined as $E[Y(1) - Y(0)|X = 1]$.

In an RCT, ACE = ATT because, due to randomization, the treated population will not, on average, differ systematically from the overall population.

Applied researchers must decide whether ATE or ATT is of greater utility or interest in their particular context.

e.g. for estimating effectiveness of an intensive, structured smoking cessation program, ATT might make more sense.

In contrast, for effect on smoking cessation of an information brochure given by family physicians to patients who are current smokers, ATE may be of greater interest.

Also possible to define ATU (but not commonly used).

Propensity Scores

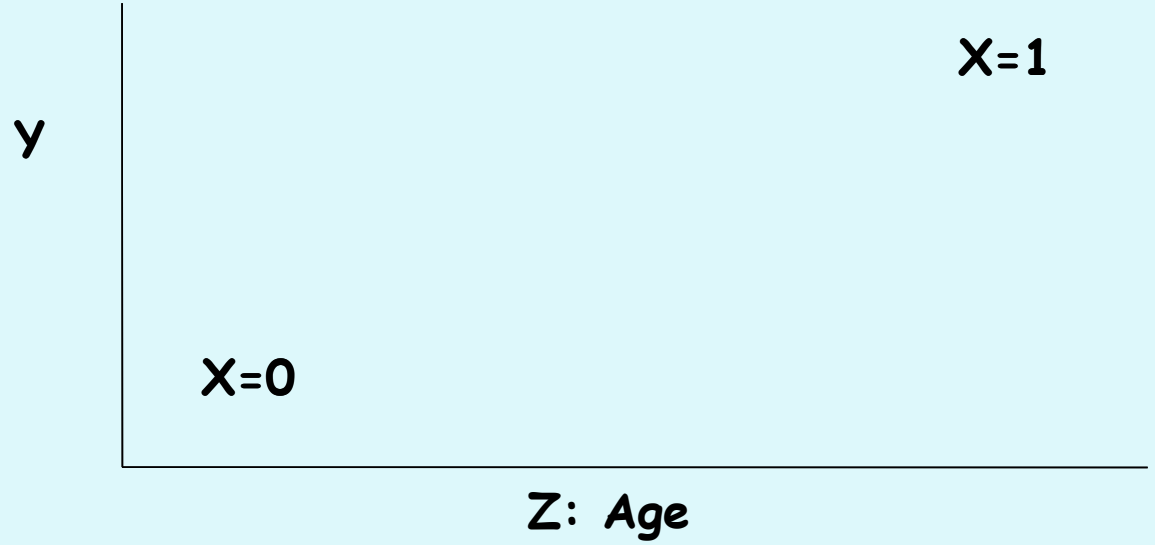
Typically, many background characteristics need to be controlled in a study of an observed exposure. Propensity score technology reduces the collection of background characteristics to a single "composite" characteristic that appropriately summarizes the collection.

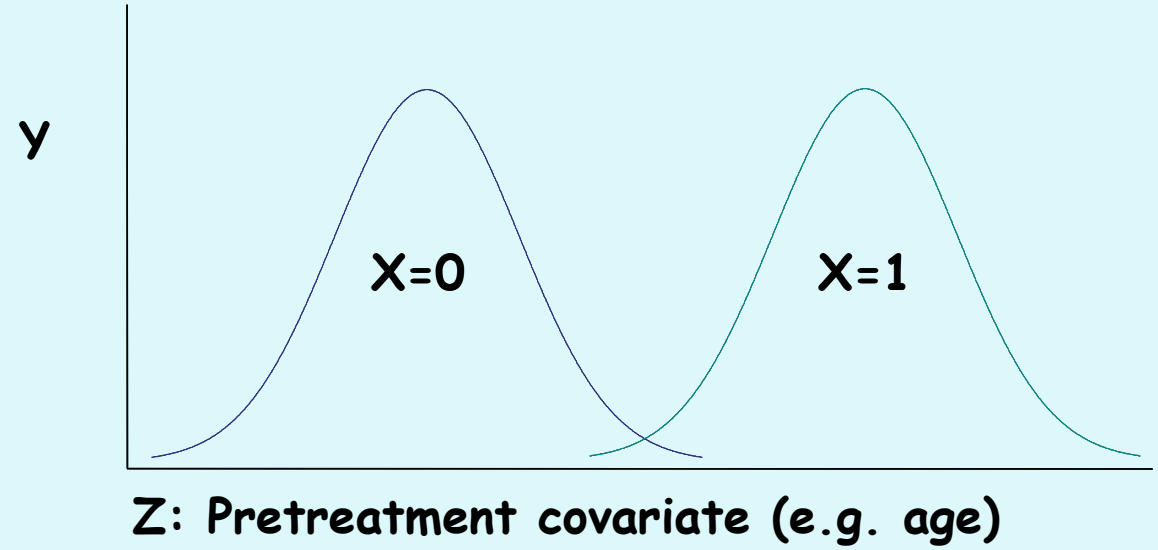
This reduction:

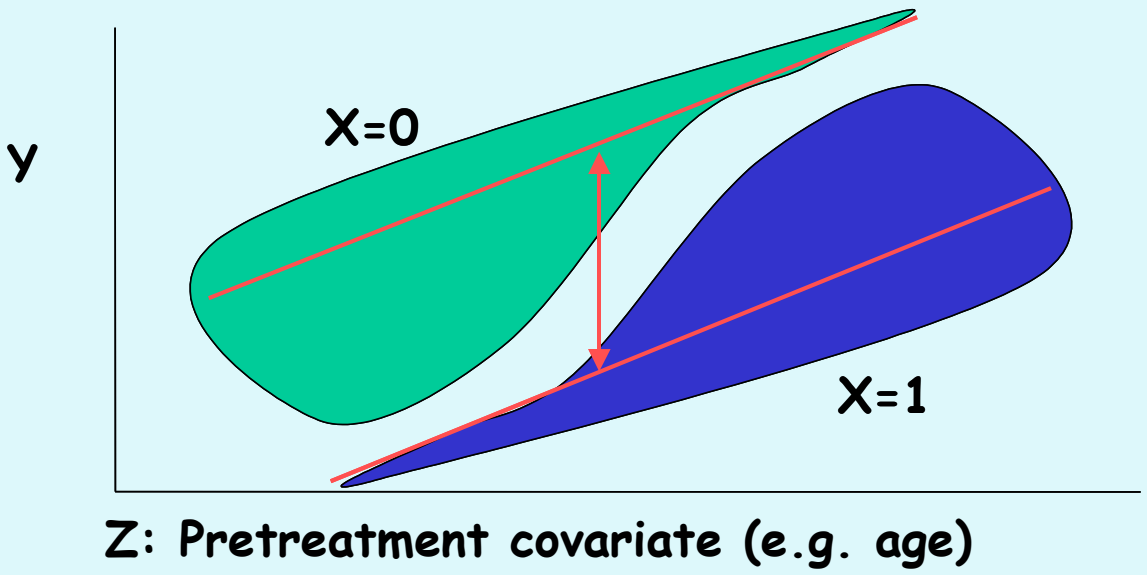
many characteristics → one composite characteristic

allows the *straightforward* assessment of whether the exposed and unexposed groups overlap enough on background characteristics to allow sensible estimation of causal effects from the data set.

Use to CHECK BALANCE and assess covariate sufficiency.


X=1
Y
X=0
Z: Age


y
X=0
X=1
Z: Pretreatment covariate (e.g. age)


y
X=0
X=1
Z: Pretreatment covariate (e.g. age)

The propensity score is just the estimated Pr(X=1|Z).

How to use this information?

At least 4 common methods have been defined:

1) stratification (or subclassification) on the propensity score (R & R's original idea)

2) propensity score matching

3) covariate adjustment using the propensity score

4) inverse probability of treatment weighting (IPTW)

Which Variable to Include

- Best to include: (↓ Bias)

  factors that affect outcome and are
  correlated with receipt of treatment

- May be beneficial to include (↑ precision):

  factors that affect outcome and are
  not correlated with receipt of treatment

- May be harmful to include (↓ precision):

  factors that do not affect outcome but
  are correlated with receipt of treatment

- Worst to include: (↑ Bias)

  factors that are affected by the exposure
  or by the outcome

Brookhart MA, et al. Variable selection for propensity score models. AJE 2006; 163(12): 1149-56

Comparing propensity score distributions

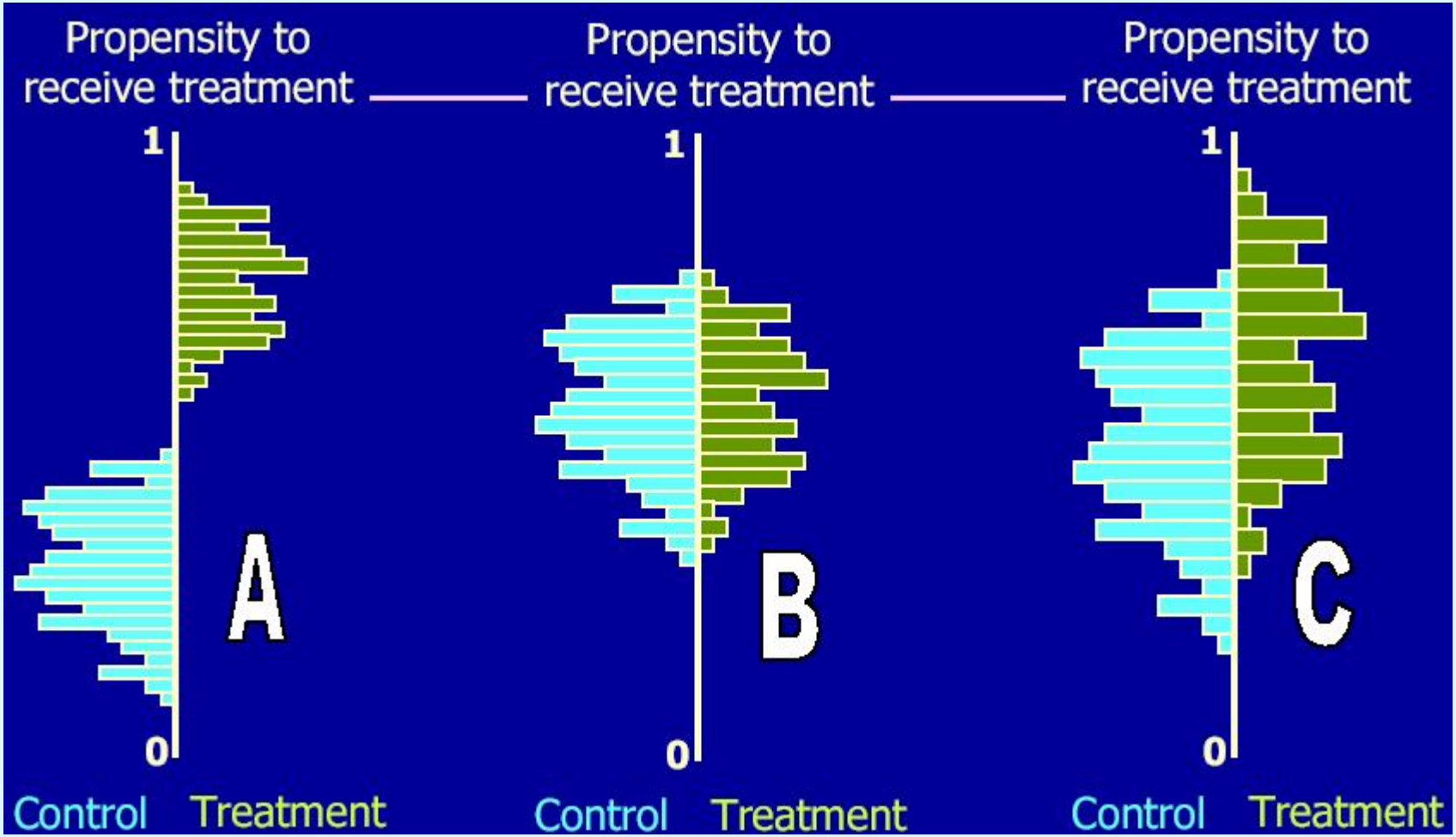
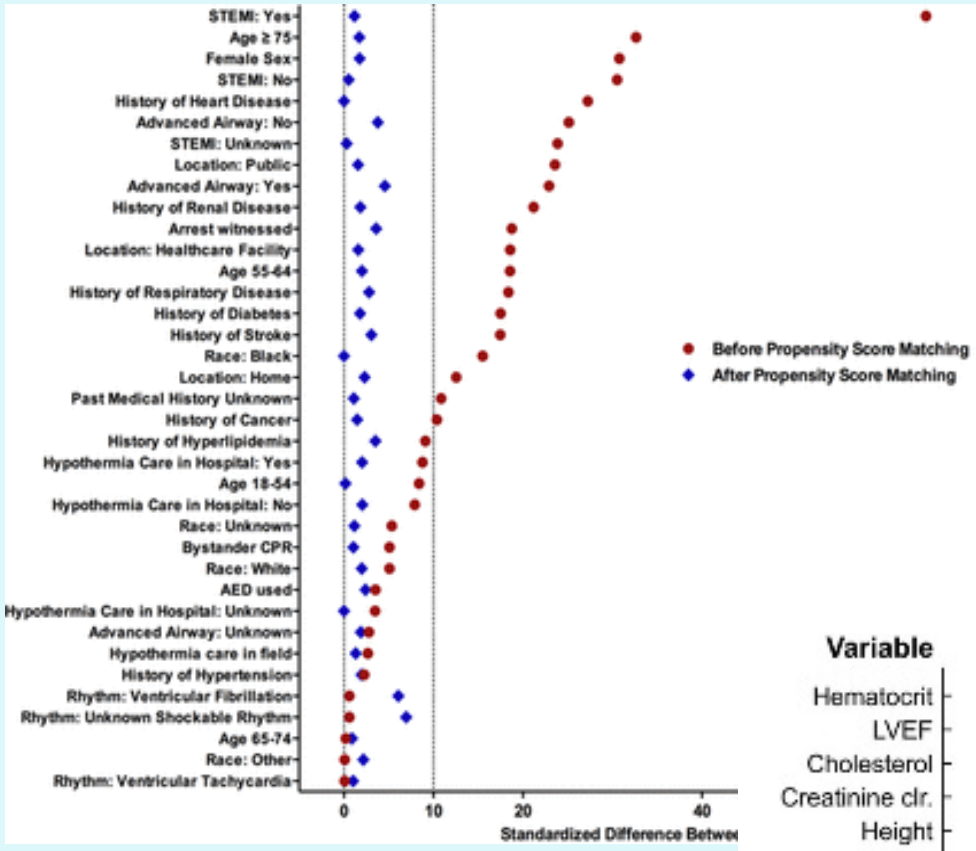
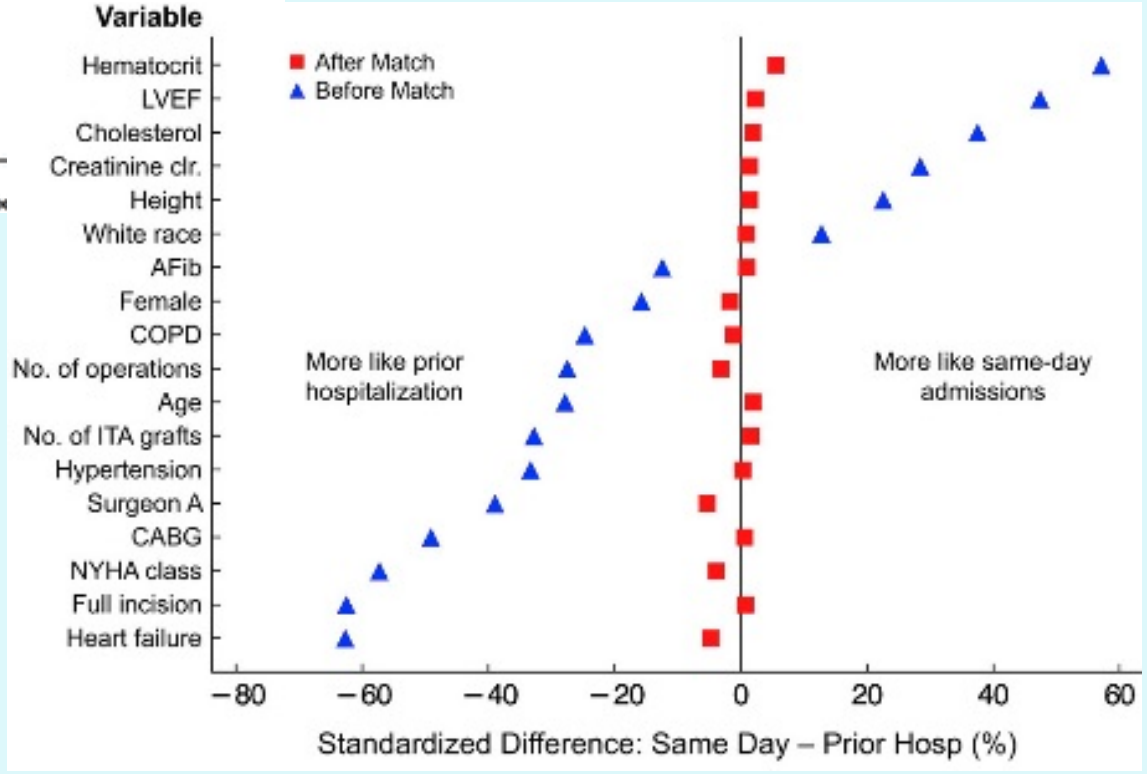


Figure from: Thomas E Love, PhD (https://cwru.pure.elsevier.com/en/persons/thomas-e-love-2)



Vyas et al. Early Coronary Angiography and Survival After Out-of-Hospital Cardiac Arrest. *Circulation* 2015.



Winger et al. Propensity-score analysis in thoracic surgery: When, why, and an introduction to how. *Journal of Thoracic and Cardiovascular Surgery* 2016

PS Advantages:

- Non-parametric, good face validity
- Captures interactions among covariates
- Examination of PS distributions exposes non-comparability (Rubin 1997)
- Single propensity score can be used in analysis of effect of exposure on >1 outcome
- Matching & stratification are robust to misspecification of PS
- Multidimensional matching with less loss of potential matches

PS Disadvantages:

- Balance may not be achieved (missing confounders, deterministic predictors, small sample size)
- Unmeasured confounders – only controlled to the extent that they are correlated with measured confounders
- Residual confounding within PS strata
- Direct adjustment relies on model form (Rubin *PDS* 2004)
- Potential adjusting for variables affected by exposure or other non-confounders
- Modeling Pr(X=1|Z) may be hard for rare exposure with many Z
- No effect estimates obtained for covariates
- Generalizability (if overlap is limited)
- Inconvenient for non-binary exposures

Consider the famous example of the impact of the NSW job-training program, using the "nswre74" dataset

```
. des

Contains data from C:\Wednesday\nswre74.dta
  obs:           445
 vars:            11                          20 May 2013 12:01
 size:        10,235
-------------------------------------------------------------------------------
storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
treat           byte    %9.0g      noyes      treatment group?
age             byte    %9.0g                 age (years)
ed              byte    %9.0g                 years of education
black           byte    %9.0g      noyes      black race?
hisp            byte    %9.0g      noyes      hispanic ethnicity?
married         byte    %9.0g      noyes      married?
nodeg           byte    %9.0g      noyes      no high school degree?
re74            float   %9.0g                 1974 earnings
re75            float   %9.0g                 1975 earnings
re78            float   %9.0g                 1978 earnings
age2            float   %9.0g                 age squared
-------------------------------------------------------------------------------
Sorted by:
```

BALANCE DIAGNOSTICS

Propensity score is a balancing score: conditional on true propensity score, distribution of measured baseline covariates is independent of treatment assignment.

But we don't know the true propensity score. Must estimate from data, which requires checking balance.

This is a major advantage of PS over regression models.

In strata of subjects with similar PS, distribution of measured baseline covariates should be similar between treated and untreated subjects.

We have a way to find out of our model is adequate.

BALANCE DIAGNOSTICS

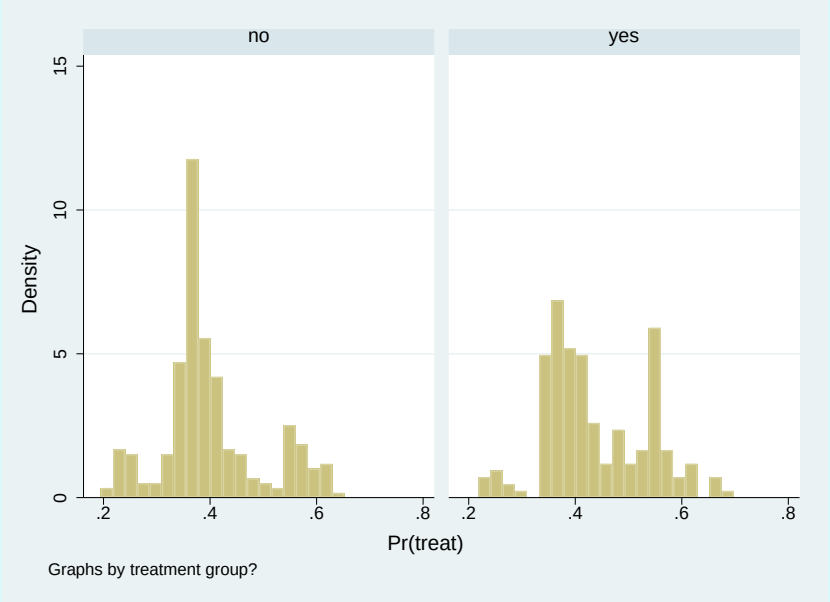
For matched data, compare treated and untreated subjects within the propensity score matched sample.

For IPTW, compare treated and untreated subjects in the inverse weighted sample

For data subclassified on the PS, compare treated and untreated subjects within the same strata of PS.

For 1-to-1 matching, the **standardized difference** is used to compare the means of continuous and binary variables between treatment groups. Multilevel categorical variables are represented with a set of binary indicator variables.

```
* build propensity score models for nsw ("by hand")

* predict treatment status
logit treat age age2 ed black hisp nodeg married re75 re74

* output predicted probability of treatment
predict ps

* summary statistics for propensity score, by treatment
bysort treat: sum ps, det
histogram ps, by(treat)
```



Good overlap

```
* nearest neighbor matching (1:1) without replacement

* randomly order data in case match ties
set seed 123456 // set seed to reproduce results

gen ranorder = runiform()
sort ranorder

* create propensity score
psmatch2 treat age age2 ed black hisp nodeg married re75 re74, logit neighbor(1) noreplacement
```

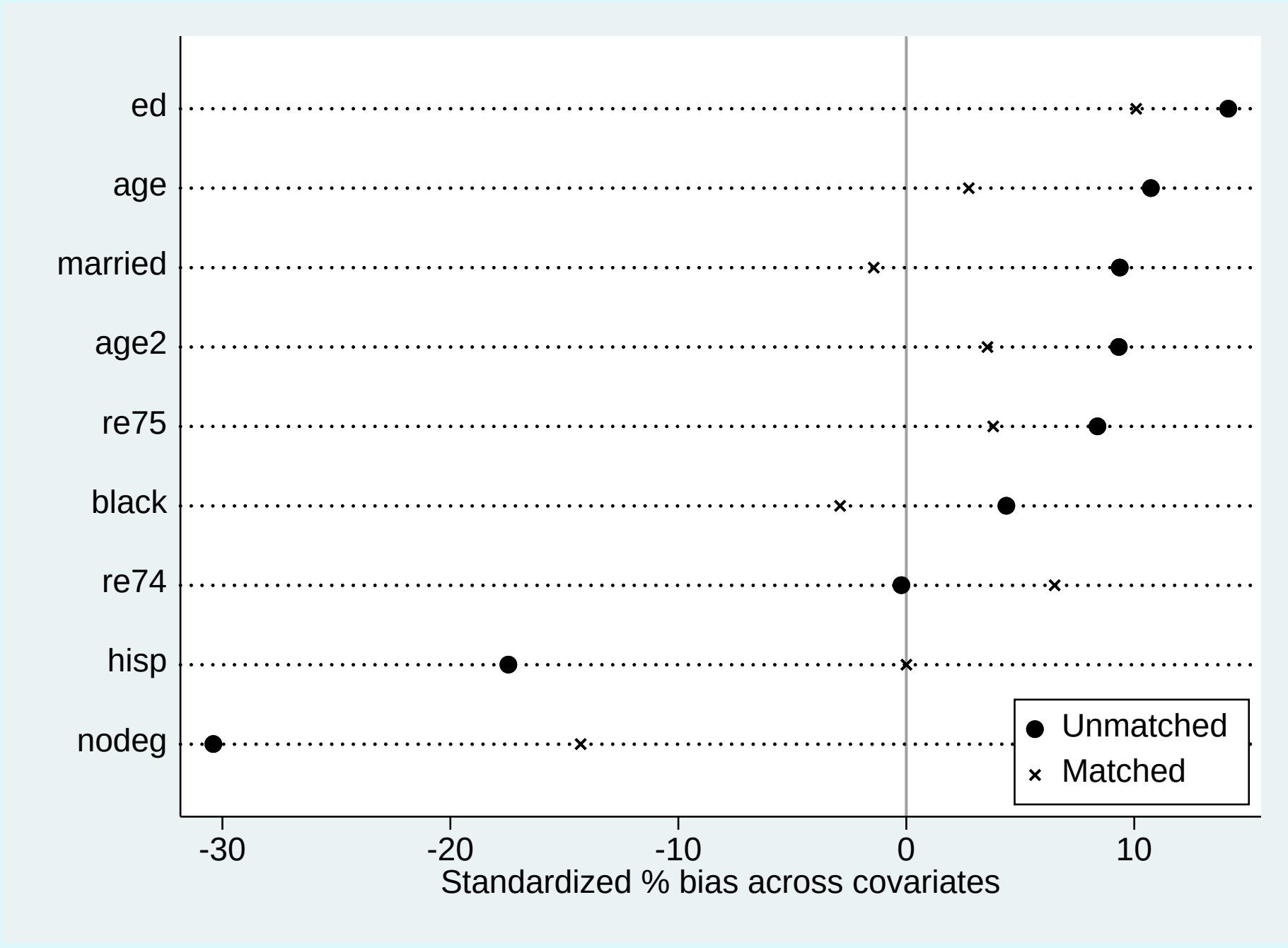
```
* evaluate balance (Rosenbaum and Rubin formula)
pstest age age2 ed black hisp nodeg married re75 re74, both graph
```

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/V(C)
age	U	25.816	25.054	10.7		1.12	0.265	1.03
	M	25.816	25.622	2.7	74.5	0.27	0.788	1.12
age2	U	717.39	677.32	9.3		0.97	0.333	1.01
	M	717.39	702.11	3.6	61.9	0.35	0.726	1.13
ed	U	10.346	10.088	14.1		1.50	0.135	1.55*
	M	10.346	10.162	10.1	28.6	0.94	0.346	1.36*
black	U	.84324	.82692	4.4		0.45	0.649	.
	M	.84324	.85405	-2.9	33.8	-0.29	0.772	.
hisp	U	.05946	.10769	-17.5		-1.78	0.076	.
	M	.05946	.05946	0.0	100.0	-0.00	1.000	.
nodeg	U	.70811	.83462	-30.4		-3.22	0.001	.
	M	.70811	.76757	-14.3	53.0	-1.30	0.195	.
married	U	.18919	.15385	9.4		0.98	0.327	.
	M	.18919	.19459	-1.4	84.7	-0.13	0.895	.
re75	U	1532.1	1266.9	8.4		0.87	0.382	1.08
	M	1532.1	1411.7	3.8	54.6	0.35	0.725	0.92
re74	U	2095.6	2107	-0.2		-0.02	0.982	0.74*
	M	2095.6	1750.8	6.5	-2910.6	0.70	0.486	1.12

```
* if variance ratio outside [0.75; 1.34] for U and [0.75; 1.34] for M
```

Sample	Ps R2	LR chi2	p>chi2	MeanBias	MedBias	B	R	%Var
Unmatched	0.028	17.04	0.048	11.6	9.4	40.1*	1.00	40
Matched	0.006	2.92	0.967	5.0	3.6	17.7	1.26	20

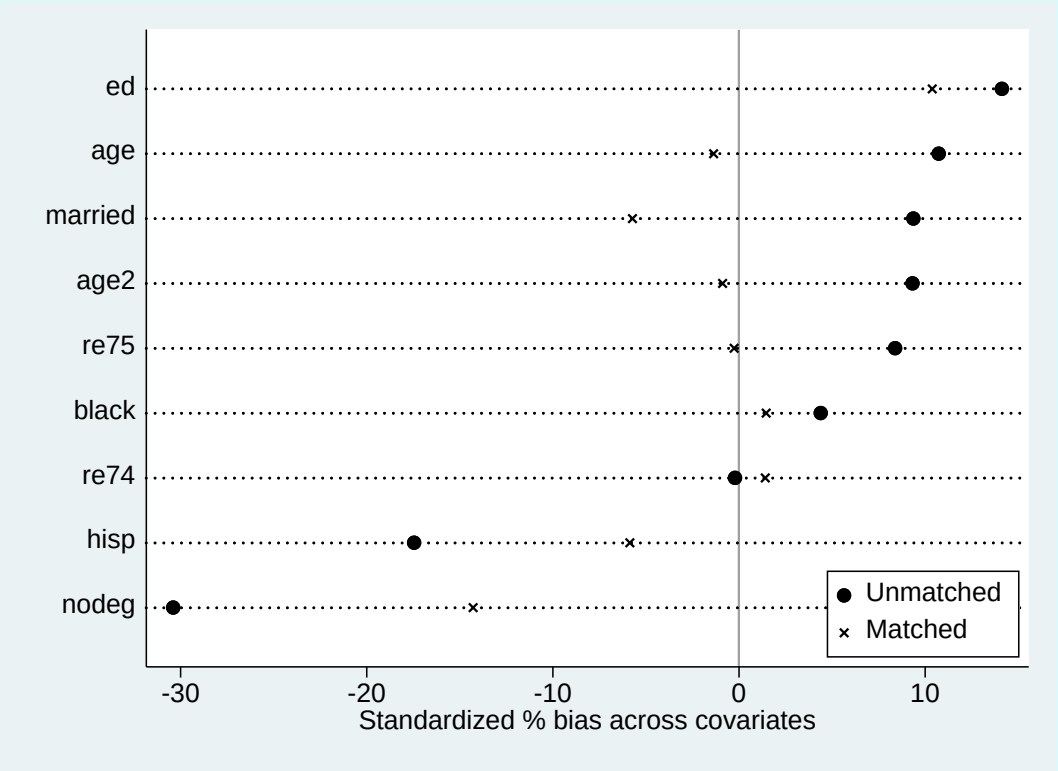
```
* if B>25%, R outside [0.5; 2]
```


ed
age
married
age2
re75
black
re74
hisp
nodeg
Unmatched
Matched
-30
-20
-10
0
10
Standardized % bias across covariates

Can we improve the model a bit?

```
* generate squared term for 1974 earnings
gen re74_2 = re74*re74
label var re74_2 "1974 earnings squared"

* randomly order data in case match ties
replace ranorder = runiform()
sort ranorder

* create propensity score
psmatch2 treat age age2 ed black hisp nodeg married re75 re74 re74_2, logit neighbor(1) noreplacement

* evaluate balance
pstest age age2 ed black hisp nodeg married re75 re74, both graph

* checking balance by hand...

Standardized difference for age = 1.4
Standardized difference for age2 = 0.9
Standardized difference for ed = 9.4
Standardized difference for black = 1.5
Standardized difference for hisp = 6.8
Standardized difference for nodeg = 13.0
Standardized difference for married = 5.5
Standardized difference for re75 = 0.3
Standardized difference for re74 = 1.5
```

Better, but "ed" and "nodeg" still don't look so good.



Treatment effect in final matched sample:

```
regress re78 treat if _weight==1
Number of obs   =       370
------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1762.777   719.3954     2.45   0.015     348.1358    3177.419
       _cons |   4586.366   508.6894     9.02   0.000     3586.063    5586.669
------------------------------------------------------------------------------
```

Tx effect = $1763 (95% CI: 348, 3177)

"Doubly robust":

```
regress re78 treat age age2 ed black hisp nodeg married re75 re74 if _weight==1
Number of obs   =       370
------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1712.479   714.1617     2.40   0.017     308.0126    3116.945
         age |    386.866   328.6266     1.18   0.240     -259.409    1033.141
        age2 |  -5.408275   5.392483    -1.00   0.317     -16.0131    5.196549
          ed |   383.7983   260.5215     1.47   0.142    -128.5417    896.1384
       black |   -2242.44   1252.403    -1.79   0.074    -4705.408    220.5279
        hisp |   190.7841    1829.93     0.10   0.917    -3407.946    3789.514
       nodeg |    101.479   1113.154     0.09   0.927    -2087.644    2290.601
     married |   -341.127    960.039    -0.36   0.723    -2229.134     1546.88
        re75 |   .0345891    .154948     0.22   0.823    -.2701306    .3393089
        re74 |   .1187688   .1048129     1.13   0.258    -.0873557    .3248932
       _cons |  -3871.263   5476.255    -0.71   0.480    -14640.83    6898.306
------------------------------------------------------------------------------
```

Tx effect = $1712 (95% CI: 308, 3117)

Treatment effects from -psmatch2- command:

```
psmatch2 treat age age2 ed black hisp nodeg married re75 re74 re74_2, logit neighbor(1)
noreplacement outcome(re78)
```

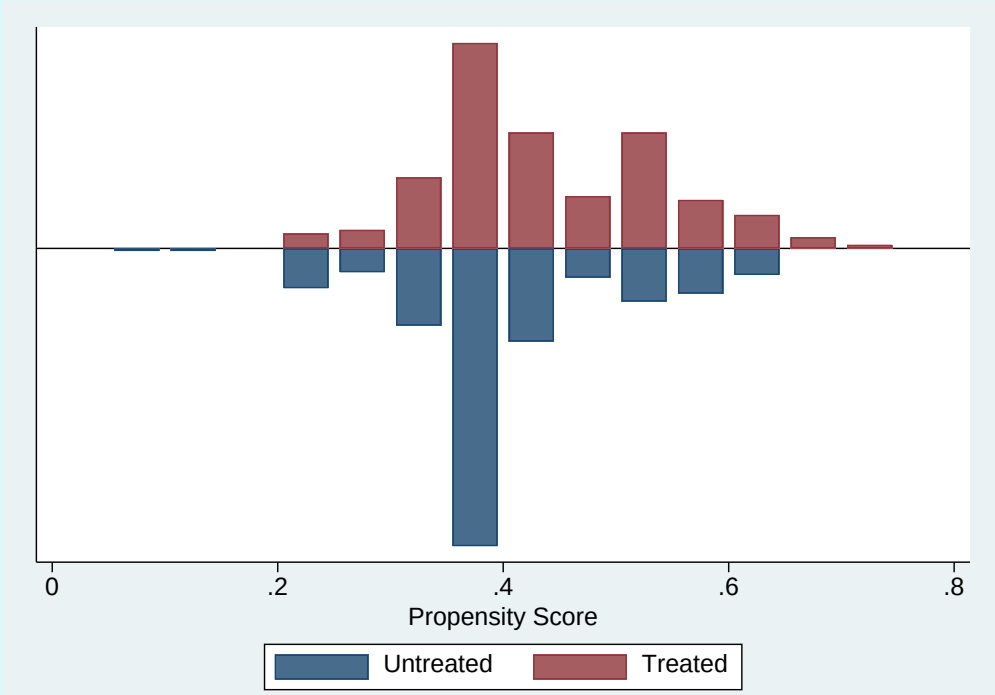
Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6349.1435	4554.80112	1794.34238	632.853392	2.84
	ATT	6349.1435	4586.36616	1762.77735	719.395391	2.45

Note: S.E. does not take into account that the propensity score is estimated.

```
tab _treat _support
```

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	260	260
Treated	185	185
Total	445	445

```
psgraph, name(pscore1, replace)
```




```
* bootstrapped standard error for ATT:

. bootstrap r(att), reps(500): psmatch2 treat
logit neighbor(1) noreplacement outcome(re78)
```

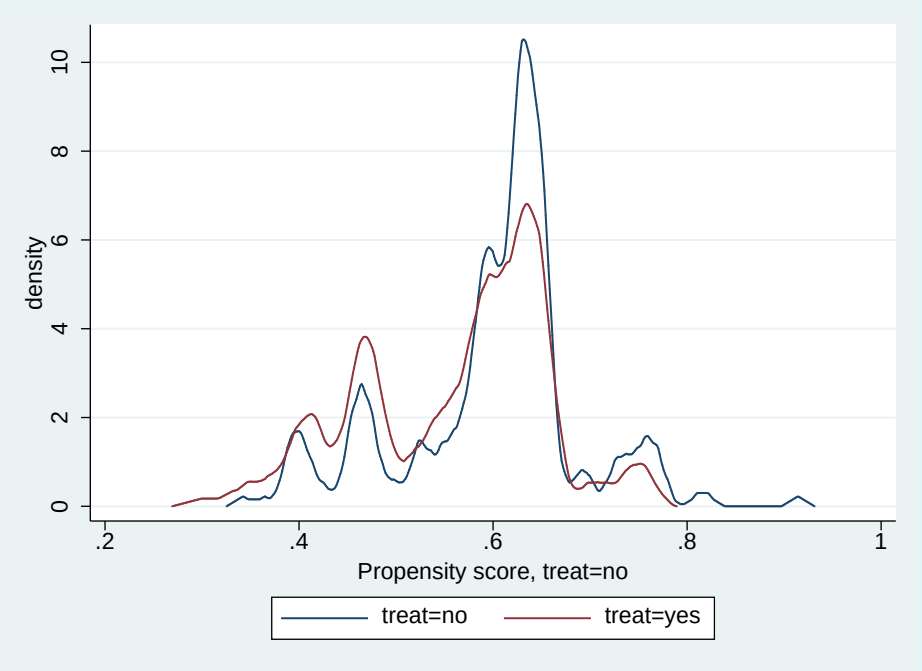
	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
tx effect	1762.777	726.0403	2.43	0.015	339.7644	3185.79

Using Stata's new "treatment effects" suite of commands:

```
teffects psmatch (re78) (treat age ed black hisp married nodeg re74 age2 re74_2), atet

Treatment-effects estimation                   Number of obs      =        445
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: logit                                     max =          8
------------------------------------------------------------------------------
             |              AI Robust
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ATT |  1721.515   739.8562     2.33   0.020     271.4232    3171.606
------------------------------------------------------------------------------

teffects psmatch (re78) (treat age ed black hisp married nodeg re74 age2 re74_2)
------------------------------------------------------------------------------
             |              AI Robust
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ATE |  1381.962   608.0156     2.27   0.023     190.2732    2573.651
------------------------------------------------------------------------------

teffects psmatch (re78) (treat age ed black hisp married nodeg re74 age2 re74_2), nneighbor(2) nopvalues
cformat(%9.2f) pformat(%5.2f) sformat(%8.2f) caliper(.2)

------------------------------------------------------------------------------
             |              AI Robust
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ATE |   1539.94     560.53                        441.33     2638.55
------------------------------------------------------------------------------
```

Post-estimation commands:

```
teffects overlap
```



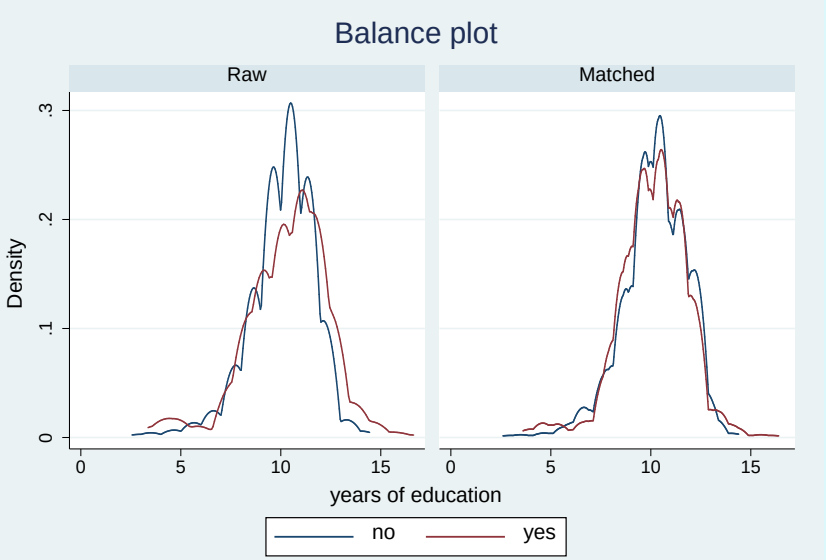

```
tebalance summarize, baseline

  Covariate balance summary
                                                 Raw      Matched
                          ----------------------------------------
                          Number of obs =          445          890
                          Treated obs   =          185          445
                          Control obs   =          260          445
                          ----------------------------------------
  ------------------------------------------------------------------
                  |           Means                  Variances
                  |    Control     Treated       Control    Treated
  ----------------+-------------------------------------------------
              age |   25.05385    25.81622      49.81176    51.1943
               ed |   10.08846    10.34595      2.606044   4.042714
            black |   .8269231    .8432432      .1436739   .1329025

etc. . .
```

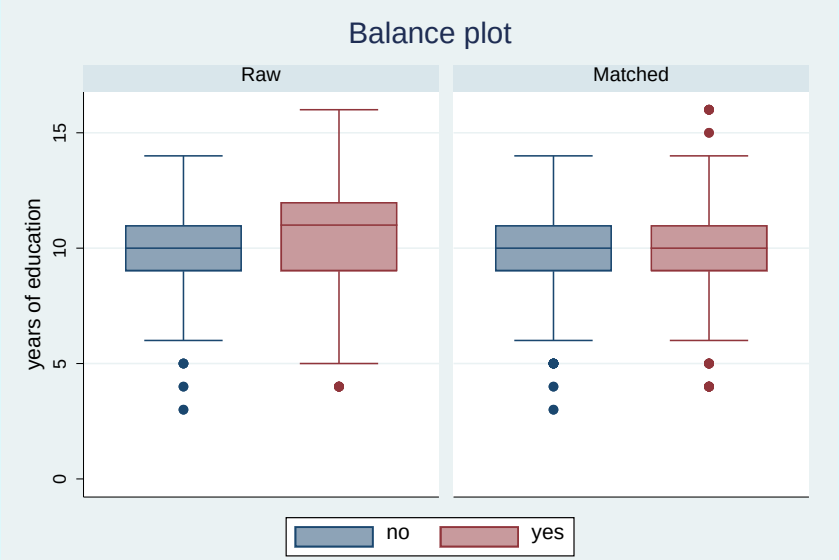
Post-estimation commands:

```
tebalance summarize
```

	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
age	.1072771	.0217026	1.027755	.8960536
ed	.1412198	-.0409838	1.551284	1.315371
black	.0438866	.005008	.9250286	.9911161
hisp	-.1745611	-.0240977	.5828804	.930887
married	.0936407	.0181852	1.180212	1.033589
nodeg	-.3039864	.0273399	1.499755	.9626536
re74	-.0021599	-.0554885	.7380953	.8065356
age2	.0932032	.0054639	1.011541	.805725
re74_2	-.0611646	-.0505421	.5038162	.7137261

```
tebalance density ed
tebalance box ed
```





Easier to specify more flexible models and check balance:

```
teffects psmatch (re78)(treat age ed black hisp married nodeg c.age#(c.age c.ed i.nodeg) c.re74#(c.re74 i.black))

Treatment-effects estimation                   Number of obs      =        445
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Treatment model: logit                                     max =          8
------------------------------------------------------------------------------
             |               AI Robust
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ATE |   1479.321   711.7633     2.08   0.038     84.29022    2874.351
------------------------------------------------------------------------------

tebalance summarize
        Covariate balance summary
                                                 Raw      Matched
                        ----------------------------------------
                        Number of obs =          445          890
                        Treated obs   =          185          445
                        Control obs   =          260          445
                        ----------------------------------------
  ------------------------------------------------------------------
                  |Standardized differences          Variance ratio
                  |        Raw     Matched           Raw    Matched
  ----------------+-------------------------------------------------
              age |   .1072771    .0496652      1.027755   .9333797
               ed |   .1412198   -.0285369      1.551284   1.135913
            black |   .0438866   -.0416538      .9250286   1.075269
             hisp |  -.1745611    .0156963      .5828804   1.046103
          married |   .0936407   -.0234989      1.180212   .9610636
            nodeg |  -.3039864    .0054267      1.499755   .9926484
          age#age |   .0932032     .033652      1.011541   .8480046
           age#ed |   .1554148    .0223087      1.214733   1.016357
        nodeg#age |  -.2152312    .0118524      1.330088    .951223
        re74#re74 |  -.0611646    -.023542      .5038162   .6547103
       black#re74 |
             no   |  -.0497845    .1128004      .3545242   1.309552
            yes   |   .0183761   -.0373243      .8288651   .8377786
  ------------------------------------------------------------------
```

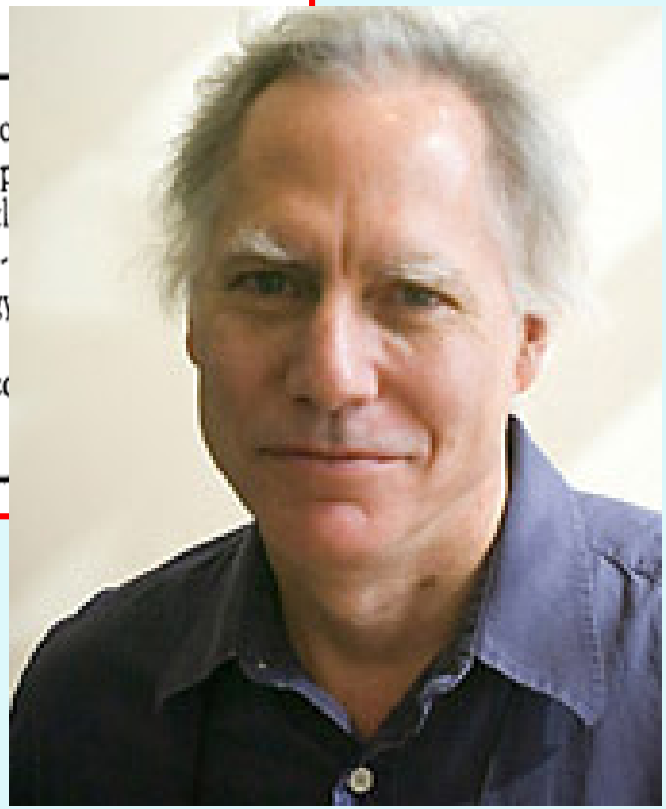
Further developments in exposure modeling for confounder control and causal inference:

Marginal Structural Models and Causal Inference in Epidemiology

James M. Robins,[1,2] *Miguel Ángel Hernán,*[1] *and Babette Brumback*[2]

In observational studies with exposures or treatments that vary over time, standard approaches for adjustment of confounding are biased when there exist time-dependent confounders that are also affected by previous treatment. This paper introduces marginal structural models, a new class of causal models that allow for impro
founding in those situations. The p
structural model can be consistentl
class of estimators, the inverse-
weighted estimators. (Epidemiology

**Keywords: causality, counterfactuals, epidemiologic methods, longitudinal data, structural models, c
variables**



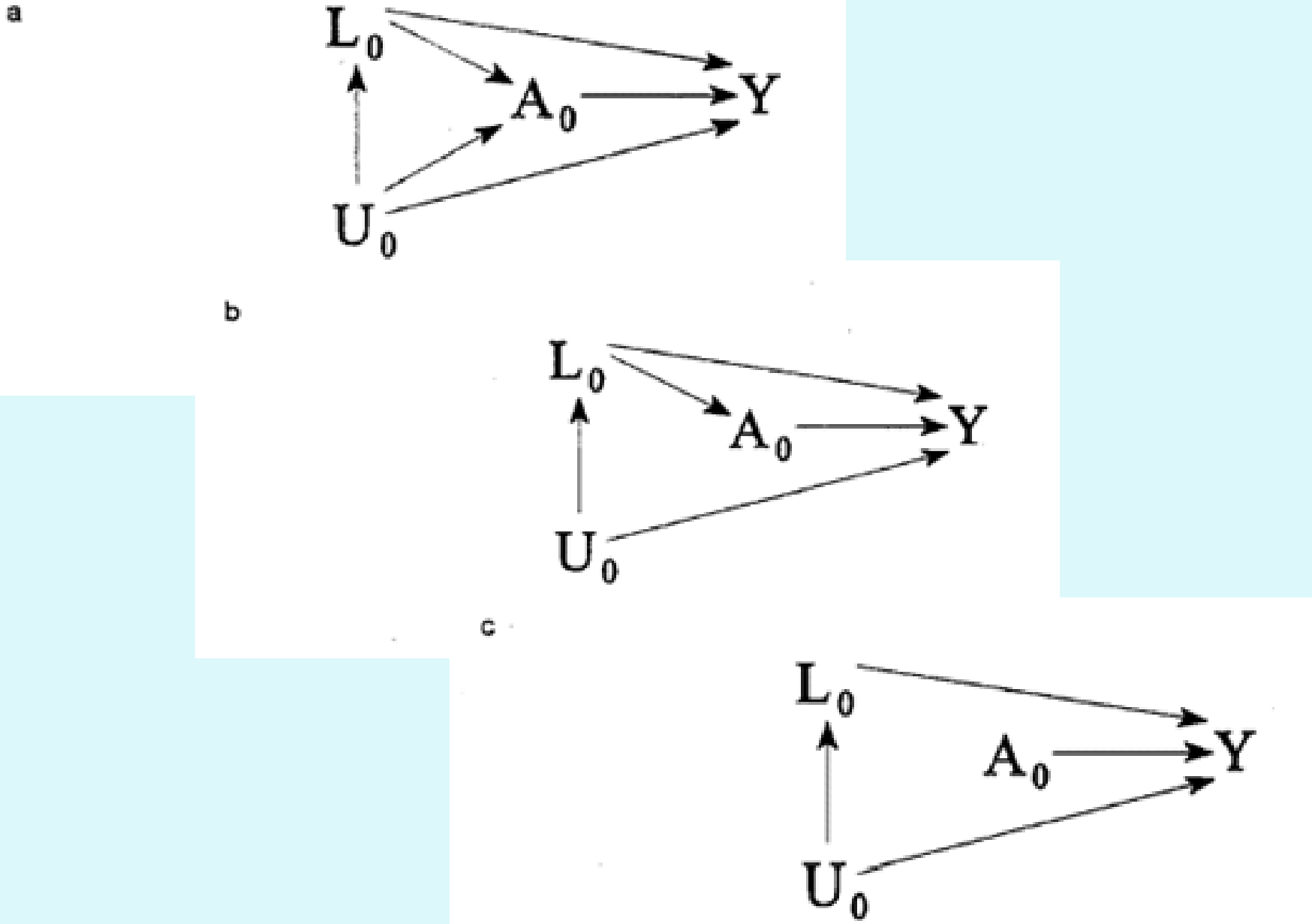
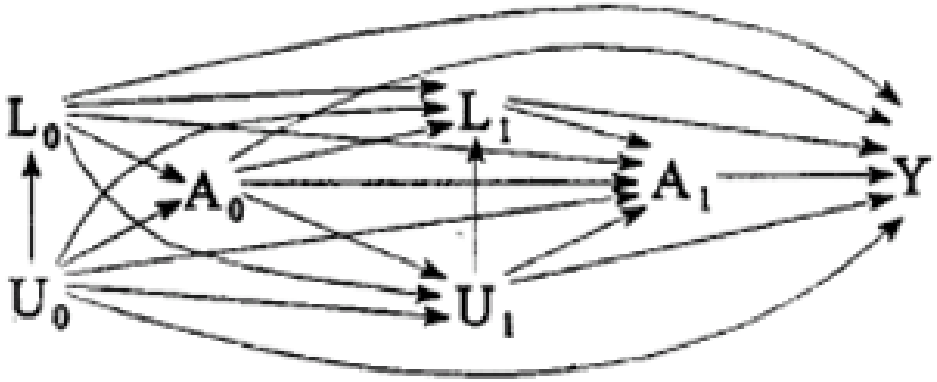
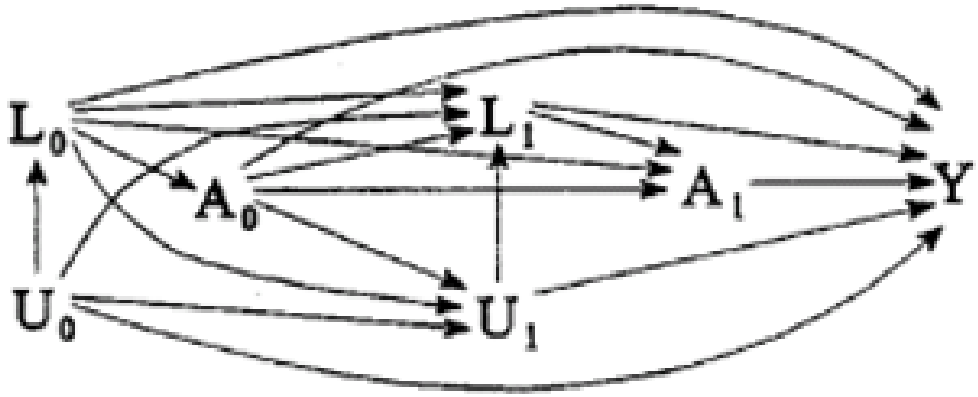


FIGURE 2. Causal graphs for a point exposure A_0.

a



b



c

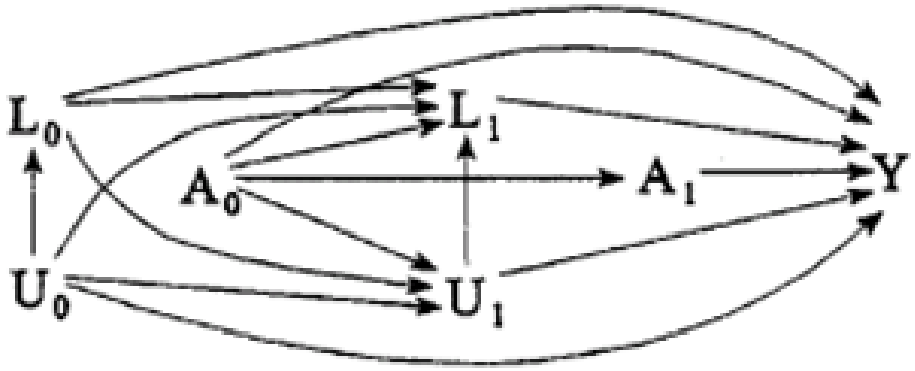


FIGURE 1. Causal graphs for a time-dependent exposure.

The average causal effect is the average over individual causal effects, where the individual effect is the contrast between counterfactual exposure states.

$Y_{a_0=1}$ = the outcome a subject would have if $a_0 = 1$

$Y_{a_0=0}$ = the outcome a subject would have if $a_0 = 0$

$Y_{a_0=1} - Y_{a_0=0}$ = individual casual effect on the difference scale

If no confounding, then:

$$Pr[Y_{a_0=1} = 1] - Pr[Y_{a_0=0} = 1] =$$
$$cRD = Pr[Y=1|A_0=1] - Pr[Y=1|A_0=0]$$

The causal RD, RR, and OR can also be expressed in terms of the parameters of the following linear, log linear, and linear logistic models for the two counterfactual probabilities $pr(Y_{a0=1} = 1)$ and $pr(Y_{a0=0} = 1)$:

$$pr[Y_{a0} = 1] = \psi_0 + \psi_1 a_0 \qquad (1)$$

$$\log pr[Y_{a0} = 1] = \theta_0 + \theta_1 a_0 \qquad (2)$$

$$\text{logit } pr[Y_{a0} = 1] = \beta_0 + \beta_1 a_0 \qquad (3)$$

where Y_{a0} is $Y_{a0=1}$ if $a_0 = 1$ and Y_{a0} is $Y_{a0=0}$ if $a_0 = 0$.

causal RD = ψ_1, causal RR = $e^{\theta 1}$, and causal OR = $e^{\beta 1}$.

Models 1-3 are saturated MSMs.

Marginal because they model the marginal distribution of the counterfactual random variables $Y_{a0=1}$ and $Y_{a0=0}$ rather than the joint distribution.

Saturated, because each has two unknown parameters so each model places no restriction on the possible values of the two unknown probabilities $pr(Y_{a0=1}=1)$ and $pr(Y_{a0=0}=1)$.

These models do not include covariates because they are, by definition, models for causal effects on the entire source population, not models for observed associations.

Crude RD, RR, and OR can also be expressed in terms of the parameters of the following saturated linear, log linear, and linear logistic models for the observed outcome Y.

$$\Pr[Y=1|A_0=a_0] = \psi'_0 + \psi'_1 a_0 \qquad (4)$$

$$\log \Pr[Y=1|A_0=a_0] = \theta'_0 + \theta'_1 a_0 \qquad (5)$$

$$\text{logit } \Pr[Y=1|A_0=a_0] = \beta'_0 + \beta'_1 a_0 \qquad (6)$$

These are models for associations observed when comparing subpopulations (defined by levels of treatment) of the source population (i.e. not causal effects).

Crude RD = ψ'_1, crude RR = $e^{\theta'1}$, and crude OR = $e^{\beta'1}$. The parameters of the associational models 4-6 will differ from the parameters of the MSMs 1-3, except when treatment is unconfounded.

Because models 4-6 are models for the observed data, unbiased estimates of the model parameters can be obtained (assuming no selection bias or measurement error).

When treatment is unconfounded, these same estimates will also be unbiased for the corresponding causal parameters of models 1-3.

<u>Under No Confounding:</u>

causal RD = ψ_1 is equivalent to crude RD = ψ'_1

causal RR = $e^{\theta 1}$ is equivalent to crude RR = $e^{\theta' 1}$

causal OR = $e^{\beta 1}$ is equivalent to crude OR = $e^{\beta' 1}$

No Unmeasured Confounders

In the real world of observational epidemiologic data, exposure will be confounded, so crude (unadjusted) model of the data will not have a causal interpretation.

i.e, the crude association parameter will not equal the corresponding causal parameter.

Assume that you have no remaining unmeasured confounders given data on measured confounders L_0.

Unbiased estimates of the causal parameters ψ_1, θ_1, and β_1 obtained by performing a weighted analysis.

This is the innovation of this article (weighting vs adjustment)

Specifically, using a weighted regression model each subject i is assigned a weight w_i equal to the inverse of the conditional probability of receiving the treatment that was actually received.

$$w_i = 1 \,/\, pr[A_0 = a_{0i} | L_0 = l_{0i}]$$

where, l_{0i} is the observed value of L_0 for subject i.

The true weights w_i are unknown but can be estimated from the data in a preliminary logistic regression of A_0 on L_0. For example, the logistic regression model:

$$\text{logit } pr[A_0 = 1 | L_0 = l_0] = \alpha_0 + \alpha_1 l_0$$

$$\text{logit } pr[A_0 = 1 | L_0 = l_0] = \alpha_0 + \alpha_1 l_0$$

Then if A_0 is tx, L_0 is the column vector of covariates and α_1 is a row vector of unknown parameters to be estimated, one obtains fitted estimates for α_0 and α_1 via standard logistic regression software. For a subject i with $A_0 = 0$ and $L_0 = l_{0i}$:

$$w_i = \frac{1}{pr(A_0 = 0 \mid L_0 = l_{0i})} =$$

$$\frac{1}{\left(\frac{e^{\alpha_0 + \alpha_1 l_{0i}}}{1 + e^{\alpha_0 + \alpha_1 l_{0i}}} \right)} = (1 + e^{\alpha_0 + \alpha_1 l_{0i}})$$

$$\text{logit } pr[A_0 = 1 | L_0 = l_0] = \alpha_0 + \alpha_1 l_0$$

On the other hand, for a subject i with $A_0 = 1$ and $L_0 = l_{0i}$:

$$w_i = \frac{1}{pr(A_0 = 1 | L_0 = l_{0i})} =$$

$$\frac{1 + e^{\alpha_0 + \alpha_1 l_{0i}}}{e^{\alpha_0 + \alpha_1 l_{0i}}} = (1 + e^{-\alpha_0 - \alpha_1 l_{0i}})$$

If there are no unmeasured confounders given data on L_0 (L_0 is a sufficient adjustment set) then one can control confounding by modifying the crude analysis by weighting each subject i by w_i.

Denominator of w_i is the probability that subject i had his/her own observed treatment. Hence "IPTW".

The effect of weighting is to create a pseudopopulation consisting of w_i copies of each subject i.

For example, if $Pr[A_0=1|L_0=l_0] = 0.25$ for a given treated subject, then $w_i = 1/0.25 = 4$, and so the subject contributes four copies of him/herself to the pseudopopulation.

This new pseudopopulation has the following two important properties:

1) in the pseudopopulation, unlike the actual population, A_0 is unconfounded by the measured covariates L_0.

2) $pr(Y_{a0=1} = 1)$ and $pr(Y_{a0=0} = 1)$ in the pseudopopulation are the same as in the true study population so that the causal RD, RR, and OR are the same in both populations.

It follows that one can unbiasedly estimate the causal RD, RR, and OR by a standard crude analysis in the pseudopopulation.

Appendix

Analyze data in Table A1 under the assumption of no unmeasured confounders given L_0.

TABLE A1. Observed Data from a Point-Treatment Study with Dichotomous Treatment A_0, Stratified by the Confounder L_0

	$L_0 = 1$		$L_0 = 0$	
	$A_0 = 1$	$A_0 = 0$	$A_0 = 1$	$A_0 = 0$
$Y = 1$	108	24	20	40
$Y = 0$	252	16	30	10
Total	360	40	50	50

Ignore sampling variability and thus the distinction between parameters of the source population and their empirical estimates.

Under the assumption of no unmeasured confounder, $pr(Y_{a0=1} = 1)$ is a weighted average of the L_0-stratum-specific risks among the treated with weights proportional to the distribution of L_0 in the entire study population (i.e. target).

That is, $pr(Y_{a0=1} = 1)$ is given by:

$$\sum_{l_0} pr[Y = 1 | A_0 = 1, L_0 = l_0] pr[L_0 = l_0]$$

where the sum is over the possible values of L_0.

The L_0-standardized risk of outcome in the exposed.

$$\sum_{l_0} pr[Y = 1 | A_0 = 1, L_0 = l_0] pr[L_0 = l_0]$$

Calculating from Table A1, this L_0-standardized risk in the treated group is estimated as $pr(Y_{a0=1} = 1)$ =

(108/360)(0.8)+(20/50)(0.2) = (0.3*0.8)+(0.4*0.2) = 0.32

Similarly, $pr(Y_{a0=0} = 1)$ is the L_0-standardized risk in the untreated:

$$\sum_{l_0} pr[Y = 1 | A_0 = 0, L_0 = l_0] pr[L_0 = l_0]$$

which, from Table A1, is

(24/40)(0.8)+(40/50)(0.2) = (0.6*0.8)+(0.8*0.2) = 0.64.

If $pr(Y_{a0=1} = 1) = 0.32$ and $pr(Y_{a0=0} = 1) = 0.64$

It follows that:

causal RD = 0.32 – 0.64 =-0.32

Causal RR = 0.32 / 0.64 = 0.50

Causal OR = (0.32/0.68) / (0.64/0.36) = 0.26

We estimated these causal effects by standardizing the risks, not by standardizing the effect parameters.

Note that these differ from the crude parameters computed from Table A2 (ignoring data on the confounder L_0):

TABLE A2. Crude Data from the Point-Treatment Study of Table A1

	$A_0 = 1$	$A_0 = 0$
$Y = 1$	128	64
$Y = 0$	282	26
Total	410	90

Thus, $\psi_1 = -0.32$, $\theta_1 = \log 0.50$, and $\beta_1 = \log 0.26$

in models 1-3 differ from the parameters

$$\psi'_1 = -0.40,\ \theta'_1 = \log 0.044,\ \text{and}\ \beta'_1 = \log 0.18$$

of models 4-6.

It is well known that the causal RD and causal RR (but not the causal OR) are also equal to weighted averages of the stratum-specific RDs and RRs.

The causal RD equals the standardized RD (*sRD*) where:

$$sRD = \sum_{l_0} RD_{l_0} pr[L_0 = l_0]$$

and $RD_{l0} = pr[Y = 1|A_0 = 1, L_0 = l_0] - pr[Y = 1|A_0 = 0, L_0 = l_0]$ is the risk difference in stratum l_0.

$RD_{l0=0} = (0.4 - 0.8) = -0.4$

$RD_{l0=1} = (0.3 - 0.6) = -0.3$

SRD = -0.4(0.2) + -0.3(0.8) = -0.08 + -0.24 = -0.32

$$sRD = \sum_{l_0} RD_{l_0} pr[L_0 = l_0]$$

The traditional approach to estimating the causal RD is to calculate this *sRD*.

Likewise this works for the sRR:

$RR_{l0=0}$ = (0.4 / 0.8) = 0.5

$RR_{l0=1}$ = (0.3 – 0.6) = 0.5

sRR = 0.5(0.2) + 0.5(0.8) = 0.1 + 0.4 = 0.50

The IPTW method is an alternative approach to estimation of the causal RD and RR that, in contrast to the approach based on calculating the *sRD*, allows generalization to unsaturated MSMs in longitudinal studies with time-varying treatments.

Table A3 displays the data from the study in a different format.

It gives the number of subjects with each of the possible combinations of l_0, a_0, and y, as well as the weight

$w = 1/pr[A_0 = a_0|L_0 = l_0]$ associated with each.

Table A3: Inverse Probability of Treatment Weights w and Composition of the Pseudopopulation in a Point-Treatment Study (Robins et al 2000)

L_0	A_0	Y	N Observed Population	$Pr(A_0\|L_0)$	w	N Pseudo Population
1	1	1	108	0.9	1.11	120
1	1	0	252	0.9	1.11	280
1	0	1	24	0.1	10	240
1	0	0	16	0.1	10	160
0	1	1	20	0.5	2	40
0	1	0	30	0.5	2	60
0	0	1	40	0.5	2	80
0	0	0	10	0.5	2	20

The final column of the table represents the number of subjects in the weighted pseudopopulation for each combination of (l_0, a_0, y). Note that the w_i need not be whole numbers or sum to 1, and so the number of subjects in the pseudopopulation can be greater than the number in the actual population.

TABLE A4. Pseudopopulation Created by Inverse Probability of Treatment Weighting from a Point-Treatment Study with Dichotomous Treatment A_0, Stratified by the Confounder L_0

	$L_0 = 1$		$L_0 = 0$	
	$A_0 = 1$	$A_0 = 0$	$A_0 = 1$	$A_0 = 0$
Y = 1	120	240	40	80
Y = 0	280	160	60	20
Total	400	400	100	100

Note that $A_0 \perp\!\!\!\perp L_0$ in this Table!

It can be seen that L_0 and A_0 are unassociated in the pseudopopulation, which implies that A_0 is unconfounded.

The lack of association between L_0 and A_0 implies that in the pseudopopulation, the L_0-standardized risk in the treated
equals the crude risk $pr(Y = 1|A_0 = 1) = 0.32$ and the L_0-standardized risk in the untreated equals the crude risk $pr(Y = 1|A_0 = 0) = 0.64$.

This means that we can calculate the crude RD or RR in Table A4 and it will have a causal interpretation:

TABLE A5. Crude Data from the Pseudopopulation of Table A4

	$A_0 = 1$	$A_0 = 0$
$Y = 1$	160	320
$Y = 0$	340	180
Total	500	500

```
. csi 160 320 340 180

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |       160          320  |        480
        Noncases |       340          180  |        520
-----------------+------------------------+------------
           Total |       500          500  |       1000
                 |                         |
            Risk |       0.32         0.64  |       0.48
                 |                         |
                 |      Point estimate     |
-----------------|-------------------------|
 Risk difference |             -0.32       |
      Risk ratio |              0.50       |
-----------------+-------------------------|
```

It follows under assumption of no unmeasured confounding given L_0, that crude RD, RR, and OR in the pseudopopulation equal causal RD, RR, and OR in the actual population.

IPTW analysis estimates crude parameters of the pseudo-Population, thus causal parameters of the actual population.

Stabilized Weights

The probabilities $\mathbf{pr}[A_0 = a_{0i} | L_0 = l_{0i}]$ may vary greatly between subjects when components of L_0 are strongly associated with A_0.

This variability can result in extremely large values of the weight w_i for a few subjects, and these few subjects will contribute a very large number of copies of themselves to the pseudopopulation and thus will dominate the weighted analysis, with the result that the IPTW estimator will have a large variance and a potentially skewed distribution.

For unsaturated MSMs, this variability can be somewhat mitigated by replacing the weight w_i by the stabilized weight: $\mathbf{sw_i = pr}[A_0 = a_{0i}]/\mathbf{pr}[A_0 = a_{0i} | L_0 = L_{0i}]$.

For example, suppose A_0 was unconfounded so that A_0 and L_0 are unassociated and $pr[A_0 = a_{0i}] = pr[A_0 = a_{0i}|L_0 = l_{0i}]$.

Then $sw_i = 1$, and each subject contributes the same weight.

When A_0 is confounded, sw_i will not be constant but will vary around 1, depending on a subject's value of L_0.

sw_i will still tend to be much less variable than w_i.

When using the weight sw_i rather than the weight w_i the estimates of the parameters of an MSM remain unbiased and will generally be less variable.

$pr[A_0 = a_{0i}]$ and $pr[A_0 = a_{0i}|L_0 = l_{0i}]$ are unknown and must be estimated.

$pr[A_0 = a_{0i}]$ can be estimated as the proportion of subjects in the study sample with A_0 equal to a_{0i}.

Table A3 displayed the appendix data for number of subjects with each of the possible combinations of l_0, a_0, and y, as well as the weight $1/pr[A_0 = a_0|L_0 = l_0]$ associated with each.

$pr[A_0 = 1] = (108+252+20+30)/500 = 410/500 = 0.82$

$pr[A_0 = 0] = (24+16+40+10)/500 = 90/500 = 0.18$

L_0	A_0	Y	N	$pr(A_0\|L_0)$	w	sw	weighted Pseudo N	stabilized Pseudo N
1	1	1	108	0.9	1.11	0.9111	120	98.4
1	1	0	252	0.9	1.11	0.9111	280	229.6
1	0	1	24	0.1	10	1.8	240	43.2
1	0	0	16	0.1	10	1.8	160	28.8
0	1	1	20	0.5	2	1.64	40	32.8
0	1	0	30	0.5	2	1.64	60	49.2
0	0	1	40	0.5	2	0.36	80	14.4
0	0	0	10	0.5	2	0.36	20	3.6

New Stabilized Table A5: Crude Data from the Stabilized Pseudopopulation of Table A4

	$A_0=1$	$A_0=0$
Y=1	131.2	57.6
Y=0	278.8	32.4
Total	410	90

```
. csi 1312 576 2788 324

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |      1312          576 |       1888
        Noncases |      2788          324 |       3112
-----------------+------------------------+------------
           Total |      4100          900 |       5000
                 |                        |
            Risk |      0.32         0.64 |     0.3776
                 |                        |
                 |      Point estimate    |
-----------------|------------------------|
 Risk difference |             -0.32      |
      Risk ratio |              0.50      |
```

Limitations of Marginal Structural Models

IPTW estimators will be biased and thus MSMs should not be used in studies in which at each time k there is a covariate level l_k such that all subjects with that level of the covariate are certain to receive the identical treatment a_k.

For example, suppose that in some stratum, treatment is impossible given the covariates.

In this instance $\mathbf{pr[A_0 = a_k | L_0 = l_k] = 0}$

Which implies that $\mathbf{1/pr[A_0 = a_k | L_0 = l_k] = 1/0 =}$ **undefined**

As probability of tx →0, weight →infinity.

Limitations of Marginal Structural Models

This implies that MSMs should not be used in occupational cohort studies.

For example, take an occupational cohort study in which A_k is level of exposure to an industrial chemical at time k and $L_k = 1$ if subject is off work at time k and $L_k = 0$ otherwise. Then all subjects with $L_k = 1$ have $A_k = 0$, because all subjects off work are necessarily unexposed.

So weights go to infinity.

Similarly, in a study of the effect of screening on mortality from cervical cancer, women who have had their cervix operatively removed by time k so that $L_k = 0$ cannot receive exposure (screening) at that time, so MSMs should not be used.

"Doubly Robust" estimator (Robins *JASA* 1994):

Combines IPT weighting with regression adjustment within treatment categories:

Remains a consistent causal estimator if either:

1) the exposure weighting model is correctly specified but the outcome regression models is not,

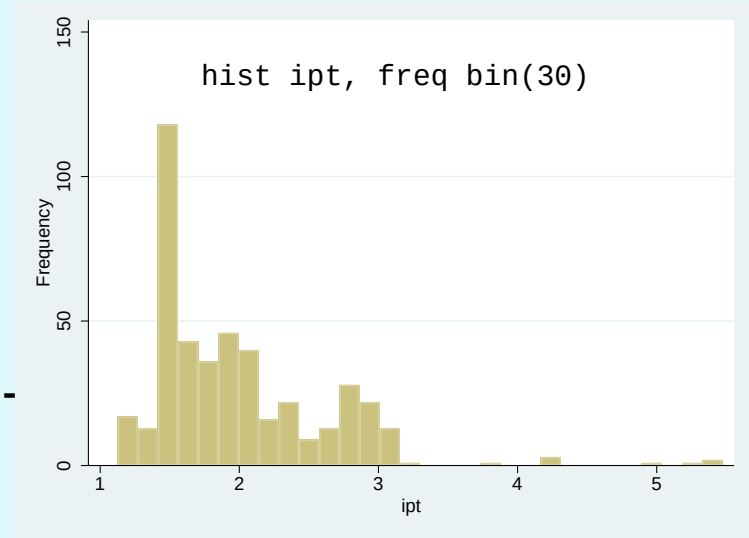
or...

2) the outcome regression model is correctly specified but the exposure weighting model is not.

Lunceford & Davidian *Stat Med* 2004

Return to the example NSW data set used earlier:

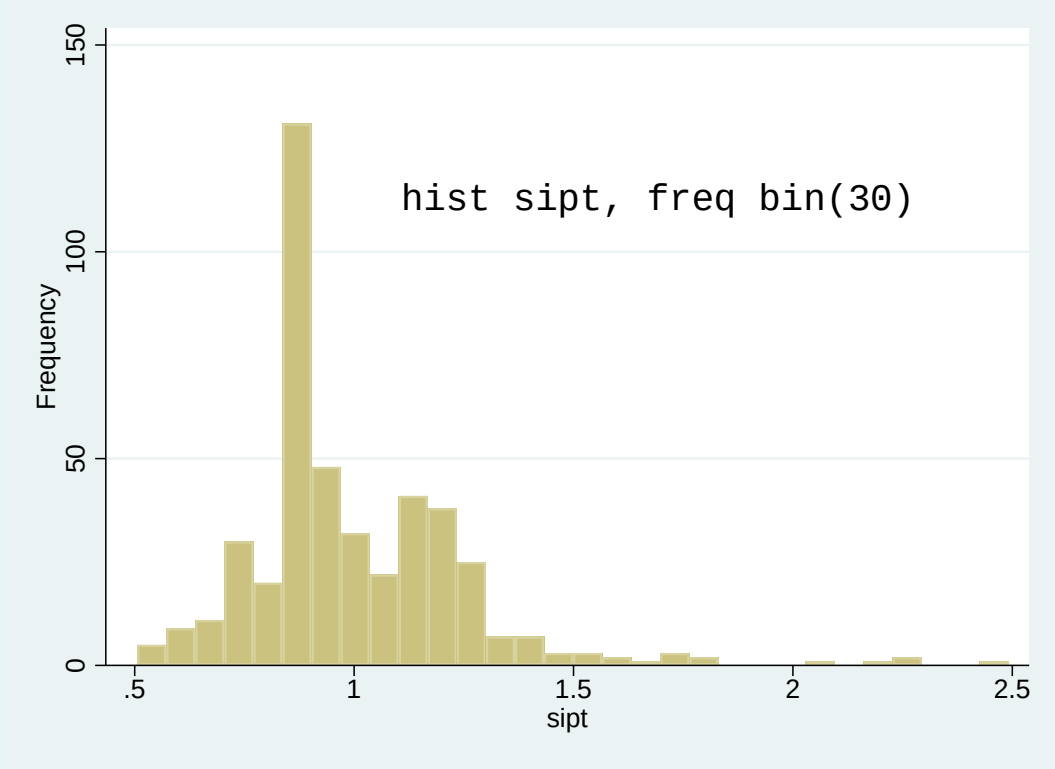
```
import delimited "C:\Users\jkaufm5\propensity.csv"
quietly logit treat age educ black##married hisp nodegr re74
                re75 u74 u75 age_2 educ_2 re74_2 re75_2

predict ps
replace ps = (ps*treat) + ((1-ps)*(1-treat))
(260 real changes made)

gen ipt = 1/ps
sum ipt, detail

                            ipt
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.212611       1.125102
 5%     1.330638       1.127132
10%     1.480171       1.202106       Obs                 445
25%     1.522238       1.210468       Sum of Wgt.         445

50%     1.816627                      Mean           1.998664
                        Largest       Std. Dev.      .6399586
75%      2.30197       4.985171
90%     2.857443       5.323654       Variance       .4095471
95%     3.001669       5.432951       Skewness       1.861643
99%      4.28544       5.471707       Kurtosis        8.71183
```



```
*test standardized difference without weighting*

mean re75_std, over(treat)
lincom [re75_std]1 - [re75_std]0
------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0841478   .0968099     0.87   0.385    -.1061148    .2744104
------------------------------------------------------------------------------

*test standardized difference with weighting*

mean re75_std [pw=ipt], over(treat)
lincom [re75_std]1 - [re75_std]0
------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0028258   .0971702    -0.03   0.977    -.1937964    .1881448
------------------------------------------------------------------------------
```

```
reg re78 treat

Linear regression                                      Number of obs =     445
------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1794.343   632.8536     2.84   0.005     550.5749    3038.111
       _cons |   4554.802    408.046    11.16   0.000     3752.856    5356.749
------------------------------------------------------------------------------

reg re78 treat [pw=ipt], cluster(id)

Linear regression                                      Number of obs =     445
                                  (Std. Err. adjusted for 445 clusters in id)
------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1626.067   687.8781     2.36   0.019     274.1651    2977.968
       _cons |   4540.558   344.7552    13.17   0.000     3863.003    5218.112
------------------------------------------------------------------------------
```

```
sum treat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       treat |       445    .4157303    .4934022          0          1

local p = r(mean)
gen sipt = (`p'*ipt)*treat + (((1-`p')*ipt)*(1-treat))
sum sipt

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        sipt |       445    .9991493    .2533877   .5056734   2.492409
```



```
reg re78 treat [pw=sipt], cluster(id)
(sum of wgt is   444.62

Linear regression                                       Number of obs =     445
                                        (Std. Err. adjusted for 445 clusters in id)
------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   1626.067   687.8781     2.36   0.019     274.1651    2977.968
       _cons |   4540.558   344.7552    13.17   0.000     3863.003    5218.112
------------------------------------------------------------------------------

teffects ipw (re78) (treat age educ black##married hisp nodegr re74
                         re75 u74 u75 age2 educ2 re742 re752), nopvalues

Treatment-effects estimation                   Number of obs     =        445
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ATE |   1626.067   655.6775      340.9623    2911.171
-------------+----------------------------------------------------------------
      POmean |   4540.558   337.0657      3879.921    5201.194
------------------------------------------------------------------------------
Warning: convergence not achieved
```

NSW example has point treatment with binary exposure where MSMs have no particular advantage over PS.

The real benefit of MSMs is for generalizations such as longitudinal data or continuous exposure.

Turn next to example in Hernán & Robins Chapter 12

Causal question "what is the average causal effect of smoking cessation on body weight gain?"

Use data from the NHEFS (National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study)

Dataset:

https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2012/10/nhefs_stata.zip

Goal is to estimate the average causal effect of smoking cessation (treatment) A on weight gain (outcome) Y .

Use 1566 cigarette smokers aged 25-74 years who had NEHFS baseline visit and follow-up about 10 years later.

Treated $A = 1$ if they reported having quit smoking before follow-up visit, and untreated $A = 0$ otherwise.

Weight gain Y measured (in kg) as the body weight at follow-up visit minus body weight at baseline.

Most people gained weight, but quitters gained more weight on average.

Average weight gain was 4.5 kg in quitters, and 2.0 kg in non-quitters.

$E[Y^{a=1}]$ is mean weight gain that would have been observed if all individuals had quit smoking before follow-up visit.

$E[Y^{a=0}]$ is mean weight gain that would have been observed if all individuals in the population had not quit smoking.

ACE on the difference scale is $E[Y^{a=1}] - E[Y^{a=0}]$

The associational difference $E[Y|A=1] - E[Y|A=0]$ = 4.5 - 2.0 = 2.3 kg is generally different from the causal difference $E[Y^{a=1}] - E[Y^{a=0}]$

Hernán & Robins select 9 potential baseline confounders: sex (0: M, 1: F), age (years), race (0: white, 1: other), education (5 categories), intensity and duration of smoking (cigarettes per day and years of smoking), physical activity in daily life (3 categories), recreational exercise (3 categories), and weight (in kg).

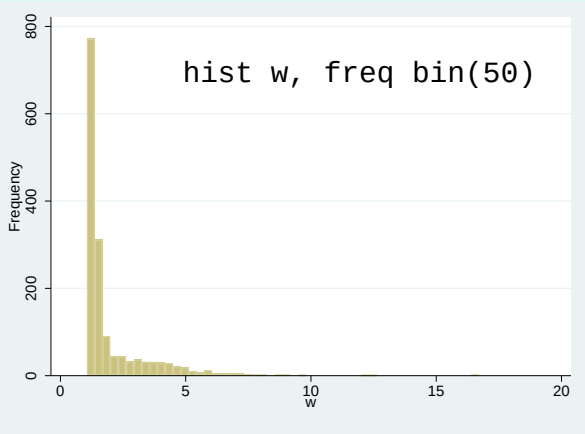
IP weighting creates a pseudo-population in which the arrow from the confounders L to the treatment A is removed.

If the confounders L are sufficient to block all backdoor paths from A to Y , then all confounding is eliminated in the pseudo-population.

Then association between A and Y in the pseudo-population consistently estimates the causal effect of A on Y .

Use logistic regression:

```
logit qsmk sex race c.age##c.age ib(last).education c.wt71##c.wt71
c.smokeintensity##c.smokeintensity ib(last).exercise
ib(last).active c.smokeyrs##c.smokeyrs
predict p_qsmk, pr

gen w=.
replace w=1/p_qsmk if qsmk==1
replace w=1/(1-p_qsmk) if qsmk==0
summarize w, det
```



The estimated IP weights ranged from 1.05 to 16.7, with mean = 2.00.

Since average weight was 2, the effective sample size doubled.

Association is causation in the pseudopopulation, so fit an associational model in the inverse weighted data:

```
regress wt82_71 qsmk [pweight=w], cluster(seqn)

Number of obs     =      1,566
------------------------------------------------------------------------------
     wt82_71 |      Coef.   Robust SE      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |   3.440535   .5258294     6.54   0.000     2.409131     4.47194
       _cons |   1.779978   .2248742     7.92   0.000     1.338892    2.221065
------------------------------------------------------------------------------
```

Causal estimate = 3.4 kg (95CI 2.4, 4.5)

```
tab sex qsmk [aw=w]

           | quit smoking between
           |   baseline and 1982
       sex | No smokin  Smoking c |     Total
-----------+----------------------+----------
         0 | 382.51458  382.52246 | 765.03704
         1 | 401.62064  399.34231 | 800.96296
-----------+----------------------+----------
     Total | 784.13522  781.86478 |     1,566
```

A ⫫ L in pseudopopulation

Why not create a pseudo-population in which A and Y are independent but in which the average weight is 1 instead of 2.

Stabilized weight!

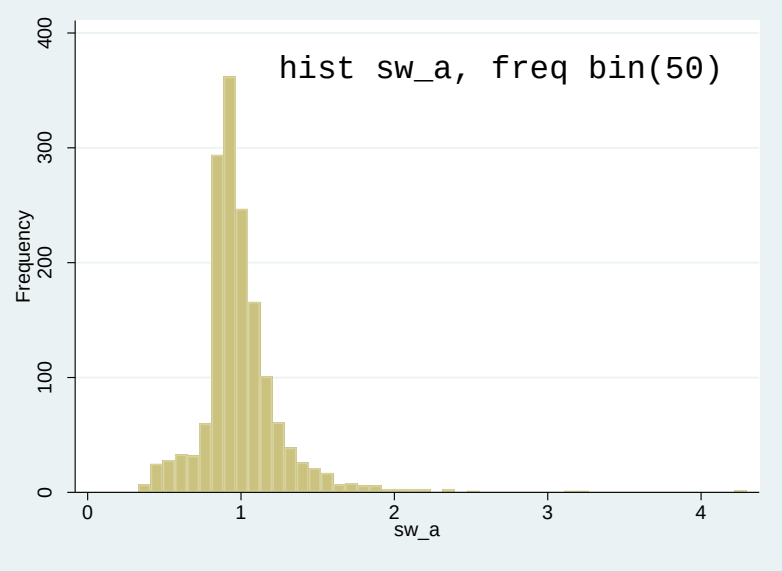
This pseudo-population would be the same size as the study population.

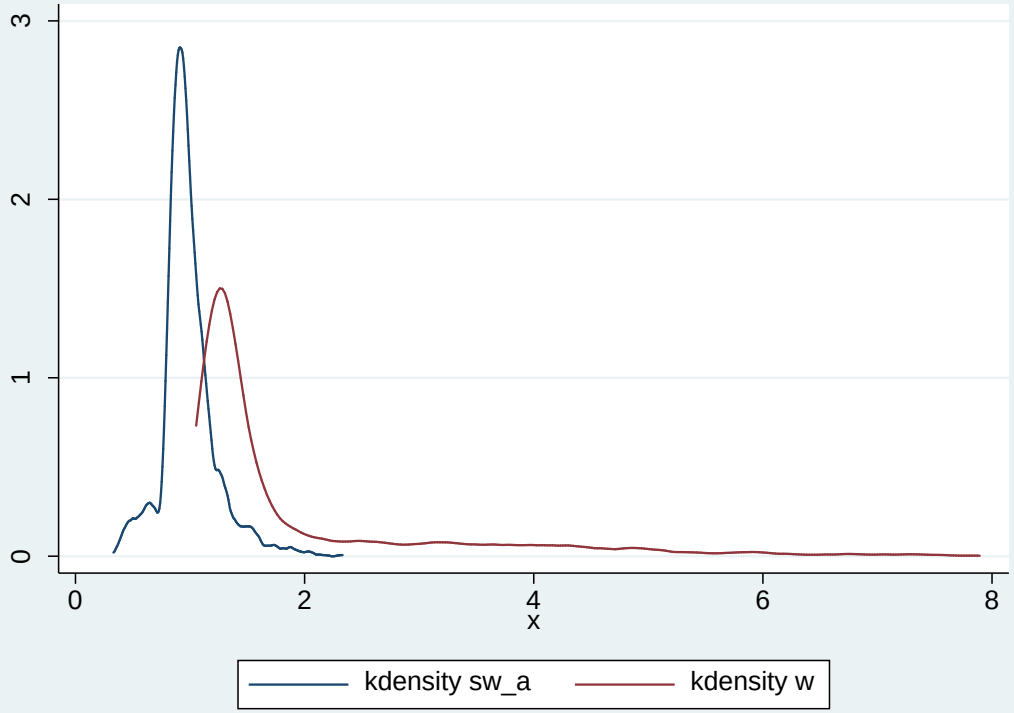
Assign to the treated subjects a probability Pr[A=1] of receiving treatment.

Assign to untreated subjects a probability Pr[A=0] of not receiving treatment.

```
/* estimation of denominator of ip weights*/
logit qsmk sex race c.age##c.age ib(last).education ib(last).active
c.wt71##c.wt71 c.smokeintensity##c.smokeintensity c.smokeyrs##c.smokeyrs
ib(last).exercise
predict pd_qsmk, pr

/* estimation of numerator of ip weights*/
logit qsmk
predict pn_qsmk, pr

gen sw_a=.
replace sw_a=pn_qsmk/pd_qsmk if qsmk==1
replace sw_a=(1-pn_qsmk)/(1-pd_qsmk) if qsmk==0
summarize sw_a, det
```

Mean weight = 1.00






```
twoway (kdensity sw_a if w < 8) (kdensity w if w < 8)
```

```
regress wt82_71 qsmk [pweight=sw_a], cluster(seqn)
(sum of wgt is 1,564.19025221467)

Number of obs     =      1,566
------------------------------------------------------------------------------
             |               Robust
     wt82_71 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |   3.440535   .5258294     6.54   0.000     2.409131     4.47194
       _cons |   1.779978   .2248742     7.92   0.000     1.338892    2.221065
------------------------------------------------------------------------------
```

In fact, not much change in CI width in this example.

The advantages of working with an unconditional estimate:

Estimate the causal effect of quitting smoking A on the risk of death D by 1992.

A=1 defined as a smoker who quit
D=1 defined as a participant who died by 1992

Model is: $ln\left(\frac{Pr(D^a=1)}{Pr(D^a=0)}\right) = \beta_0 + \beta_1 A$

Therefore, the effect parameter is $e^{\beta 1}$ = the causal odds ratio of death contrasting quitting versus not quitting.

Estimate this causal model by fitting an associational model in the pseudo-population:

$$ln\left(\frac{Pr(D=1|A)}{Pr(D=0|A)}\right) = b_0 + b_1 A$$

First, estimate the stabilized weights sw_a

```
logit qsmk sex race c.age##c.age ib(last).education
     c.smokeintensity##c.smokeintensity c.smokeyrs##c.smokeyrs ib(last).exercise
     ib(last).active c.wt71##c.wt71
```

Then obtain propensity scores:

```
predict pd_qsmk, pr
```

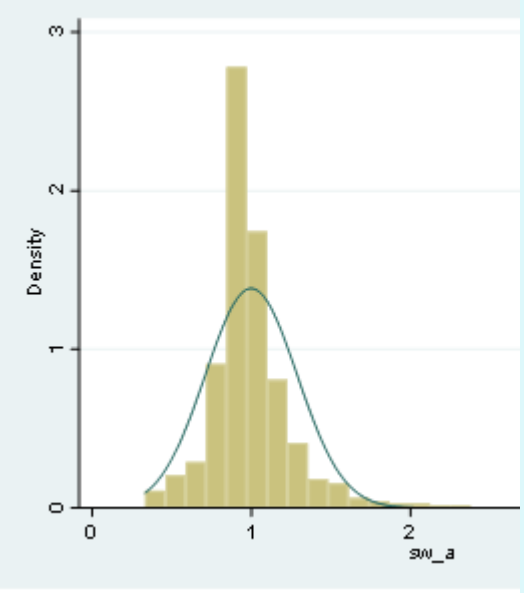
Then get the numerator of the stabilized weight:

```
logit qsmk
predict pn_qsmk, pr
```

Then construct the stabilized weight as Pr(A) / Pr(A|L):

```
gen sw_a=.
replace sw_a=pn_qsmk/pd_qsmk if qsmk==1
replace sw_a=(1-pn_qsmk)/(1-pd_qsmk) if qsmk==0
sum sw_a, det
hist sw_a, norm
```

We confirm that weights are symmetrically distributed around a mean of 1.00



Now fit the weighted logistic regression model, which estimates the associational parameters in the pseudopopulation:

```
logistic death qsmk [pweight=sw_a], cluster(seqn) cformat(%6.2f)

Logistic regression                     Number of obs     =      1,566
------------------------------------------------------------------------------
             |               Robust
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |       1.03       0.16     0.19   0.848        0.76    1.40
       _cons |       0.23       0.02   -18.88   0.000        0.19    0.26
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Therefore we conclude that on the odds ratio scale, the effect of quitting smoking on short-term mortality is null: OR = 1.0 (95% CI: 0.8, 1.4).

This is very different from the crude association:

```
logistic death qsmk, cformat(%6.2f)
------------------------------------------------------------------------------
             |      Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |       1.40       0.20     2.39   0.017        1.06    1.86
------------------------------------------------------------------------------
```

Observed OR = 1.4 (95% CI: 1.1, 1.9).

This causal estimate is not very exciting (because it's null), but it is still good to note that any GLM will work here, not just logistic.

This allows us to avoid the OR (which has no causal interpretation due to non-collapsibility concerns explained earlier).

For example, use binomial regression to get causal RR:

```
binreg death qsmk [pweight=sw_a], rr cluster(seqn) cformat(%6.2f)

Generalized linear models                  No. of obs      =      1,566
Variance function: V(u) = u*(1-u)          [Bernoulli]
Link function    : g(u) = ln(u)            [Log]

------------------------------------------------------------------------------
             |               Semirobust
       death | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |       1.02       0.13     0.19   0.848         0.80        1.32
       _cons |       0.18       0.01   -26.28   0.000         0.16        0.21
------------------------------------------------------------------------------
Note: _cons estimates baseline risk.
```

For a large effect magnitude, the OR and RR would be very different.

Or risk difference:

```
binreg death qsmk [pweight=sw_a], rd cluster(seqn) cformat(%6.2f)

Generalized linear models                  No. of obs      =      1,566
Variance function: V(u) = u*(1-u)          [Bernoulli]
Link function    : g(u) = u                [Identity]

------------------------------------------------------------------------------
             |             Semirobust
       death | Risk Diff.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        qsmk |       0.00       0.02     0.19   0.849        -0.04        0.05
       _cons |       0.18       0.01    15.52   0.000         0.16        0.21
------------------------------------------------------------------------------
```

Or the standardized absolute risks in the exposed and unexposed:

```
binreg death i.qsmk [pweight=sw_a], rd cluster(seqn) cformat(%6.4f)
margins i.qsmk, cformat(%6.2f)

Adjusted predictions                       Number of obs     =      1,566
Expression   : Predicted mean death, predict()
------------------------------------------------------------------------------
             |            Delta-method
             |    Margin  Std. Err.      z   P>|z|     [95% CI]
-------------+----------------------------------------------------------------
      qsmk=0 |    0.1839   0.0118    15.52   0.000   0.1606   0.2071
      qsmk=1 |    0.1884   0.0208     9.05   0.000   0.1476   0.2292
------------------------------------------------------------------------------
```

International Journal of Epidemiology, 2017, 756–762
doi: 10.1093/ije/dyw323
Advance Access Publication Date: 30 December 2016
Education Corner

Education Corner

An introduction to g methods

Ashley I Naimi[1]*, Stephen R Cole[2] and Edward H Kennedy[3]

[1]Department of Epidemiology, University of Pittsburgh, [2]Department of Epidemiology, University of North Carolina at Chapel Hill and [3]Department of Statistics, Carnegie Mellon University

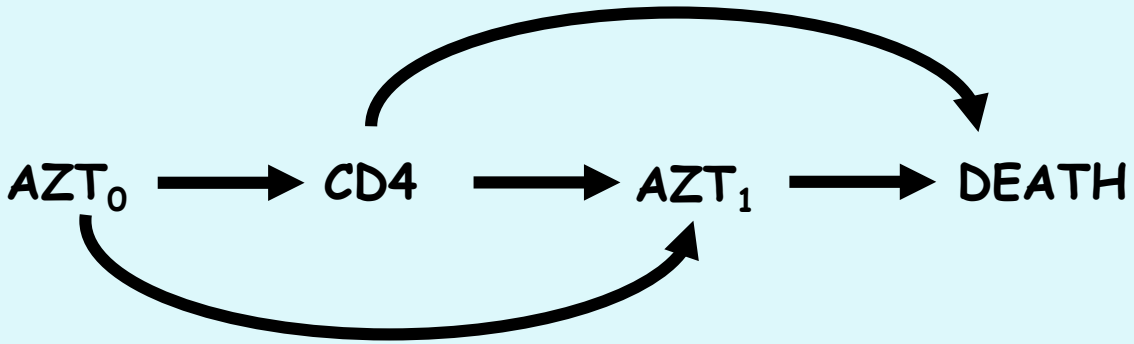
*Corresponding author. Department of Epidemiology University of Pittsburgh 130 DeSoto Street Pittsburgh, PA 15261 ashley.naimi@pitt.edu

Robins' named his models "g methods" for "general", to enable the identification and estimation of the effects of all kinds of exposures or intervention plans.

A family of methods that include the g formula, marginal structural models, and structural nested models.

Traditional regression models require no feedback between time-varying treatments and time-varying confounders.

Before 1986, no solution to this general problem.



Naimi example concerns the effect of treatment for HIV on (continuous) CD4 count.

Table 1 presents data from a hypothetical cohort study (A=1 for treated, A=0 otherwise).

A_0	Z_1	A_1	Y	N
0	0	0	87.29	209,271
0	0	1	112.11	93,779
0	1	0	119.65	60,654
0	1	1	144.84	136,293
1	0	0	105.28	134,781
1	0	1	130.18	60,789
1	1	0	137.72	93,903
1	1	1	162.83	210,527

Treatment measured at baseline (A_0) and at follow up (A_1).

One covariate is elevated HIV viral load (Z=1 means >200 copies/ml, Z=0 otherwise)

Outcome Y is CD4 count at the end of follow up in cells/mm^3.

N is large, so no sampling variability issues or measures.

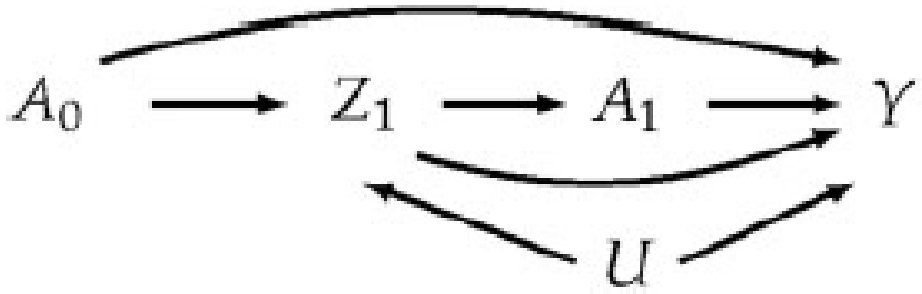


Figure 1. Causal diagram representing the relation between anti-retroviral treatment at time 0 (A_0), HIV viral load just prior to the second round of treatment (Z_1), anti-retroviral treatment status at time 1 (A_1), the CD4 count measured at the end of follow-up (Y), and an unmeasured common cause (U) of HIV viral load and CD4.

Based on the DAG, $E(Y|A_0, A_1, Z)$ may be composed of several parts:

effects of A_0, Z, and A_1;
3 two-way interactions between A_0, Z, and A_1;
1 three-way interactions between A_0, Z, and A_1;

Naimi focuses on the ACE of always taking treatment (a_0=a_1=1) compared to never taking treatment (a_0=a_1=0).

$$\varphi = E(Y^{a_0=1,a_1=1}) - E(Y^{a_0=0,a_1=0})$$

$$= E(Y^{a_0=1,a_1=1} - Y^{a_0=0,a_1=0})$$

This ACE is a marginal effect because it does not compute contrasts within covariate strata and then combine these (like regression or Mantel-Haenszel)

Write the effect as

$$E(Y^{a0,a1}) - E(Y^{0,0}) = \varphi_0 a_0 + \varphi_1 a_1 + \varphi_2 a_0 a_1$$

i.e. that the ACE φ may be composed of two exposure main effects φ_0 and φ_1 their two-way interaction φ_2.

This marginal effect differs from a conditional effect that is calculated within covariate strata.

For example, a conditional effect would make sense if you were interested in effect measure modification by Z.

Conditioning on Z=1 would answer the question: what is the effect of A_0 and A_1 in those with high viral load?

But for now, effect of interest is just $\varphi = \varphi_0 + \varphi_1 + \varphi_2$.

Standard Methods

```
input a0 z a1 y n
0 0 0  87.29 209271
0 0 1 112.11  93779
0 1 0 119.65  60654
0 1 1 144.84 136293
1 0 0 105.28 134781
1 0 1 130.18  60789
1 1 0 137.72  93903
1 1 1 162.83 210527
end

expand n
gen avga = (a0+a1)/2
reg y avga, nohead cformat(%6.2f)
reg y avga z, nohead cformat(%6.2f)
reg y a0, nohead cformat(%6.2f)
reg y a0 z, nohead cformat(%6.2f)
reg y a1, nohead cformat(%6.2f)
reg y a1 z, nohead cformat(%6.2f)
```

Standard Methods

Model	Estimate of β_1
$\beta_0 + \beta_1(A_0 + A_1)/2$	60.9
$\beta_0 + \beta_1(A_0 + A_1)/2 + \beta_2 Z$	42.6
$\beta_0 + \beta_1 A_0$	27.1
$\beta_0 + \beta_1 A_0 + \beta_2 Z$	18.0
$\beta_0 + \beta_1 A_1$	38.9
$\beta_0 + \beta_1 A_1 + \beta_2 Z$	25.0

In the first model, $\beta_1 = 60.9$ cells/mm^3 is the crude difference in mean CD4 count for the always treated compared to the never treated.

Second model gives 42.6 cells/mm^3 for the Z-adjusted difference in mean CD4 count for the same contrast.

Estimate the MSM using IPTW:

MSM is: $E(Y^{a0,a1}) = \beta_0 + \varphi_0 a_0 + \varphi_1 a_1 + \varphi_2 a_0 a_1$

where $\beta_0 = E(Y^{0,0})$ is an intercept parameter and the ACE $\varphi = E(Y^{1,1} - Y^{0,0}) = \varphi_0 + \varphi_1 + \varphi_2$

First obtain predicted probabilities of the observed tx.

There are 2 A_1 values (1 vs 0) for each of 4 levels of Z x A_0.

Additionally, there are 2 possible A_0 values (1 vs 0).

Four possible exposure regimes (A_0, A_1): 00, 10, 01, and 11

For each Z value, obtain from the table the predicted probability of the exposure that was actually received.

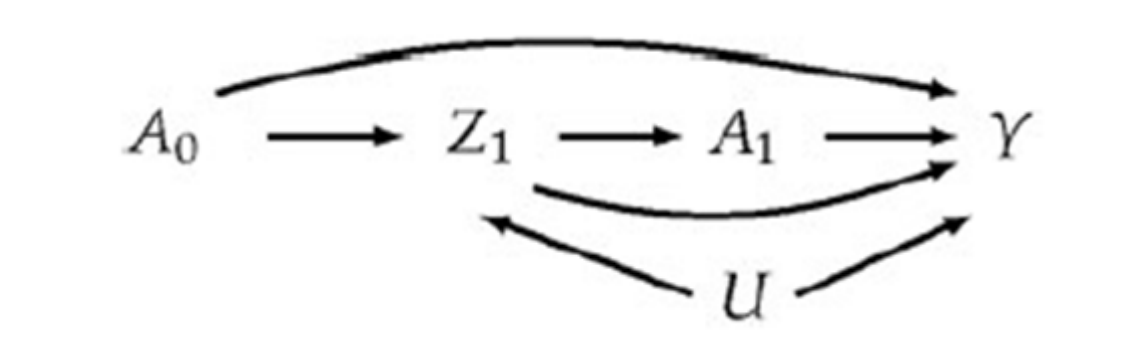
Estimate the MSM using IPTW (cont):

No variables affect A_0, so this probability is 0.5 for all individuals in the sample.

According to the DAG, A_1 is not directly affected by A_0

Therefore the Z-specific probabilities of A_1 are constant across levels of A_0 (because $A_1 \perp\!\!\!\perp A_0 \mid Z$)

In a DAG where A_0 affects A_1, the Z-specific probabilities of A_1 would vary across levels of A_0.



Estimate the MSM using IPTW (cont):

```
* IPTW
quietly sum a1
scalar pa1 = r(mean)
quietly sum a1 if z==0
scalar pa1z0 = r(mean)
quietly sum a1 if z==1
scalar pa1z1 = r(mean)

scalar w000 = (0.5 * (1-pa1z0))
scalar w001 = (0.5 * (pa1z0))
scalar w010 = (0.5 * (1-pa1z1))
scalar w011 = (0.5 * (pa1z1))
scalar w100 = (0.5 * (1-pa1z0))
scalar w101 = (0.5 * (pa1z0))
scalar w110 = (0.5 * (1-pa1z1))
scalar w111 = (0.5 * (pa1z1))

scalar sw000 = (0.5*(1-pa1))/(0.5 * (1-pa1z0))
scalar sw001 = (0.5*pa1)/(0.5 * (pa1z0))
scalar sw010 = (0.5*(1-pa1))/(0.5 * (1-pa1z1))
scalar sw011 = (0.5*pa1)/(0.5 * (pa1z1))
scalar sw100 = (0.5*(1-pa1))/(0.5 * (1-pa1z0))
scalar sw101 = (0.5*pa1)/(0.5 * (pa1z0))
scalar sw110 = (0.5*(1-pa1))/(0.5 * (1-pa1z1))
scalar sw111 = (0.5*pa1)/(0.5 * (pa1z1))
```

Table 4. Stabilized inverse probability weights and Pseudo-population obtained by using inverse probability weights

A_0	Z_1	A_1	Y	sw	Pseudo N
0	0	0	87.23	0.72	151222.84
0	0	1	112.23	1.62	151680.46
0	1	0	119.79	1.62	98110.06
0	1	1	144.78	0.72	98789.40
1	0	0	105.25	0.72	97395.08
1	0	1	130.25	1.62	98321.62
1	1	0	137.80	1.62	151884.02
1	1	1	162.80	0.72	152596.51

```
gen sw = .
replace sw = (0.5*(1-pa1))/(0.5 * (1-pa1z0)) if a0==0 & z==0 & a1==0
replace sw = (0.5*pa1)/(0.5 * (pa1z0)) if a0==0 & z==0 & a1==1
replace sw = (0.5*(1-pa1))/(0.5 * (1-pa1z1)) if a0==0 & z==1 & a1==0
replace sw = (0.5*pa1)/(0.5 * (pa1z1)) if a0==0 & z==1 & a1==1
replace sw = (0.5*(1-pa1))/(0.5 * (1-pa1z0)) if a0==1 & z==0 & a1==0
replace sw = (0.5*pa1)/(0.5 * (pa1z0)) if a0==1 & z==0 & a1==1
replace sw = (0.5*(1-pa1))/(0.5 * (1-pa1z1)) if a0==1 & z==1 & a1==0
replace sw = (0.5*pa1)/(0.5 * (pa1z1)) if a0==1 & z==1 & a1==1

reg y i.a0##i.a1 [pw=sw], nohead cformat(%6.2f)

------------------------------
           y |      Coef.
-------------+----------------
        1.a0 |      25.02
        1.a1 |      25.00
       a0#a1 |
        1 1  |      -0.01
       _cons |     100.02
------------------------------
```

thus ACE φ = $E(Y^{1,1} - Y^{0,0})$= φ_0 + φ_1 + φ_2 = 25+25+0=50

Weighting the observed data by the inverse of the probability of observed exposure yields a "pseudo-population" in which treatment at the second time point (A_1) is no longer related to and thus no longer confounded by previous viral load (Z).

Weighting a regression model for the outcome by the inverse probability of treatment enables us to account for the fact that Z both confounds A_1 and is affected by A_0.

```
table a1 a0 z [pweight = sw], contents(freq )

----------------------------------------------------
           |                  z and a0
           | -------- 0 -------    -------- 1 -------
        a1 |         0          1           0          1
-----------+----------------------------------------
         0 |  151,222  97,394.8    98,106.2   151,886
         1 |  151,681  98,321.9    98,789.1   152,596
----------------------------------------------------
```

Now Replicate this Estimate Using the G-Formula

Start with a mathematical representation of the data generating mechanism (the joint density of the observed data).

Factor the joint density in a way that respects the temporal ordering of the data by conditioning each variable on its history.

$$f(y, a_1, z, a_0) = f(y|a_1, z, a_0)Pr(a_1|z,a_0)Pr(z|a_0)Pr(a_0)$$

Marginalize over the distribution of A_1, Z and A_0 to get the marginal mean of Y:

$$E(Y) = \sum_{a_1,z,a_0} E(Y|a_1, z, a_0) \Pr(a_1|z, a_0) \Pr(z|a_0) \Pr(a_0)$$

Under intervention on A_1 and A_0, however, these are certain to take values of 1 or 0, and so these probabilities $\Pr(a_1|z, a_0)$ and $\Pr(a_0)$ are struck from the factorization:

This leaves only:

$$E(Y^{a1,a0}) = \sum_{z} E(Y|a_1, z, a_0) \Pr(z|a_0)$$

This equation is the so-called "g-formula"

In the Naimi example therefore, the expected CD4 count under exposure vector 0,0 can be calculated as:

$$E(Y^{0,0}) = E(Y|A_1 = 0, Z = 1, A_0 = 0)\Pr(Z = 1|A_0 = 0) + E(Y|A_1 = 0, Z = 0, A_0 = 0)\Pr(Z = 0|A_0 = 0)$$

Weighting the observed outcome's conditional expectation by the conditional probability that Z=z accounts for the fact that Z is itself affected by A_0, but also confounds the effect of A_1 on Y.

Likewise for exposure vector 1,1:

$$E(Y^{1,1}) = E(Y|A_1 = 1, Z = 1, A_0 = 1)\Pr(Z = 1|A_0 = 1) + E(Y|A_1 = 1, Z = 0, A_0 = 1)\Pr(Z = 0|A_0 = 1)$$

And similarly for any other combination of A_0 and A_1

```
* g-formula
quietly sum y if a1==0 & z==0 & a0==0
scalar y000 = r(mean)
quietly sum y if a1==0 & z==0 & a0==1
scalar y001 = r(mean)
quietly sum y if a1==0 & z==1 & a0==0
scalar y010 = r(mean)
quietly sum y if a1==1 & z==0 & a0==0
scalar y100 = r(mean)
quietly sum y if a1==0 & z==1 & a0==1
scalar y011 = r(mean)
quietly sum y if a1==1 & z==1 & a0==0
scalar y110 = r(mean)
quietly sum y if a1==1 & z==0 & a0==1
scalar y101 = r(mean)
quietly sum y if a1==1 & z==1 & a0==1
scalar y111 = r(mean)
quietly sum z if a0==1
scalar z1 = r(mean)
quietly sum z if a0==0
scalar z0 = r(mean)

scalar Y00 = (y010*z0)+(y000*(1-z0))
scalar Y11 = (y111*z1)+(y101*(1-z1))
scalar Y01 = (y011*z1)+(y001*(1-z1))
scalar Y10 = (y110*z0)+(y100*(1-z0))

disp "Y00 = ", Y00
disp "Y01 = ", Y01
disp "Y10 = ", Y10
disp "Y11 = ", Y11
```

Using the data in the table yields:

$E(Y^{0,0}) = 100.0$

$E(Y^{1,0}) = 125.0$

$E(Y^{0,1}) = 125.0$

$E(Y^{1,1}) = 150.0$

One can model each probability from the data in this way, potentially with many covariate adjustments.

Therefore φ = 150.0-100.0 = 50.0

This is the average causal effect of treatment (always on versus always off) on CD4 count.

Same as the estimate obtained via IPTW.

The Naimi paper also uses a 3rd method (g-estimation).

COFFEEBREAK