

Counterfactuals and Potential Outcomes Estimation

Dave Glidden
Biostatistics 210

1

Potential Outcomes

- $Y_i(1)$: Glucose ith person who exercises
- $Y_i(0)$: Glucose ith person no exercise
- $Y_i(1) - Y_i(0)$: causal effect in ith person
would be rigorous to observe this!
- Only get to see one realization
even in a randomized trial
possible exception: crossover trials

Average Causal Effect

- Effect of exposure in population
- A.C.E. $E[Y(1)] - E[Y(0)]$
 $= 1/n \sum Y_i(1) - Y_i(0)$
- Average causal effect
- Very elegant “target”
- Why is it causal?
*compares exposure in all $i=1, \dots, n$
the identical population*

3

Table 9.1. Potential Outcomes Framework for the Effect of Exercise

Sub- Population	Person	$Y(1)$	$Y(0)$	$Y(1)-Y(0)$
$C = 0$	1	94	99	-5
	2	98	99	-1
	3	99	98	1
	4	103	105	-2
	5	96	99	-3
$C = 1$	6	95	94	1
	7	92	95	-3
	8	94	95	-1
	9	90	94	-5
	10	99	101	-2
Overall	Mean	96	98	-2

4

Fundamental Problem

- Only get to see actual exposure
not potential outcome
- E_i : exposure in the i th person
- E_i : associated w/ factor(s) C related to outcome
- C_i : -- confounder. Associated with exposure and outcome
- Distorts the observed difference in exposure groups

5

Table 9.2. Effect of Exercise Using Observed Data Only

Sub-Population	Person	$Y(1)$	$Y(0)$	Estimated Difference
$C = 0$	1		99	
	2		99	
	3		98	
	4		105	
	5	96		
$C = 1$	6	95		
	7		95	
	8	94		
	9	90		
	10	99		
Overall	Mean	94.8	99.2	-4.4

$C=1$: family history of diabetes, $C=0$: no family history of diabetes

6

Missing Data Problem

- What is mean of $Y(1)$?
- Missing in some people
they were exposed to $E_i=0$
- Missingness associated with C
- C is associated with latent $Y(1)$
- Classic missing data problem

“Fundamental problem” is a missing data problem

7

Randomization Assumption

- E_i is independent of $Y(1)$ and $Y(0)$
- If true can identify $E[Y(1)]$ and $E[Y(0)]$
estimate them rigorously
- Observed difference identifies causal effect
- Causal inference: without having counterfactuals

Why we love the RCT

Potential outcome is “missing completely at random”

8

Conditional Independence

- Relaxes this assumption
- If C is a confounder...
 - suppose within levels of C , E and $Y(1), Y(0)$ are not associated
 - exposure random within levels of the confounder
- Adjustment identifies effect within C
- Assumption behind “adjustment” for confounders

9

Conditional Independence

- Synonymous with “no unmeasured confounders” beyond C
- $E(Y | E=e, C) = E(Y(e) | C)$
- Estimate effect of exposure with C
and combine in some sense
- Basis of several methods
regression, inverse weighting, propensity scores, potential outcomes estimation (g-computation)

This is an MAR assumption

10

Causal Estimation

11

Glucose and Exercise

- $Y_i(1)$: Glucose ith person who exercises
- $Y_i(0)$: Glucose ith person no exercise
- $Y_i(1) - Y_i(0)$: causal effect in ith person
- Can't observe. Can estimate it's average $E\{Y(1) - Y(0)\}$
- Need to estimate: $E[Y(1)], E[Y(0)]$

12

Table 9.2. Effect of Exercise Using Observed Data Only

Sub-Population	Person	$Y(1)$	$Y(0)$	Estimated Difference
C = 0	1		99	
	2		99	
	3		98	
	4		105	
	5	96		
C = 1	6	95		
	7		95	
	8	94		
	9	90		
	10	99		
Overall	Mean	94.8	99.2	-4.4

13

Potential Outcomes

- Estimate: $E[Y(1)]$, $E[Y(0)]$
- Mean glucose in population
 $E(Y(1))$ if entire population exercises
 $E(Y(0))$ if entire population doesn't exercise
- Estimating is a missing data problem
- MAR with respect to variable C
which is the two-level confounder

14

Major Methods

Model Outcome

Potential Outcomes/G-computation

Model Exposure

Inverse Weighting, Propensity Scores

all rely on no unmeasured confounders

15

The Means

	E=1	E=0	Diff
C=0	96	100.25	-4.25
C=1	94.5	95	-0.5

16

Weighting Population

- $E(Y(x)) = P(C=1) \cdot E(Y(x)|C=1) + P(C=0) \cdot E(Y(x)|C=0)$
- $E(Y(1)) = 5/10 \cdot 96 + 5/10 \cdot 94.5 = 95.25$
- $E(Y(0)) = 5/10 \cdot 95 + 5/10 \cdot 100.25 = 97.625$
- $E(Y(1)) - E(Y(0)) = -2.4$

17

Table 9.2. Effect of Exercise Using Observed Data Only

Sub-Population	Person	$Y(1)$	$Y(0)$	Estimated Difference
C = 0	1	96	99	
	2	96	99	
	3	96	98	
	4	96	105	
	5	96	100.25	
C = 1	6	95	95	
	7	94.5	95	
	8	94	95	
	9	90	95	
	10	99	95	
Overall	Mean	95.3	97.6	-2.4

18

Potential Outcome

- Step 1: Identify confounders: C
- Step 2: Regress Y given C and E
- Step 3: Set $E_i=e$, estimate $E(Y_i|E_i=e, C_i)$
mean, proportion, predicted survival
- Step 4: $E\{Y(e)\} = n^{-1} \sum_{i=1}^n E(Y_i|E_i = e, C_i)$

19

Fracture Data

Fracture Intervention Trial

- 2669 placebo participants in FIT
- Followed for mean of 3.9 years
- 523 new fractures of any kind
- 5.1/100 person year 0.95 CI (4.7 to 5.5)
- Baseline bone mineral density (BMD)
- estimate $Y(e)$ where BMD = e
- confounders; age, body mass index

20

Relative Risk Regression

```
. glm fracture blbmd bmi age, link(log) family(binomial) nolog eform
```

Generalized linear models		No. of obs	=	2669
Optimization	: ML	Residual df	=	2665
		Scale parameter	=	1
Deviance	= 2548.539609	(1/df) Deviance	=	.9563
Pearson	= 2653.92697	(1/df) Pearson	=	.995845
Variance function: V(u) = u*(1-u)		[Bernoulli]		
Link function : g(u) = ln(u)		[Log]		
Log likelihood = -1274.269805		AIC	=	.9578642
		BIC	=	-18476.87

fracture	Risk Ratio	OIM Std. Err.	z	P> z	[95% Conf. Interval]	
blbmd	.6229577	.0338921	-8.70	0.000	.5599494	.6930559
bmi	1.015919	.0096703	1.66	0.097	.9971412	1.03505
age	1.017513	.0067336	2.62	0.009	1.004401	1.030797
_cons	.7162968	.4579807	-0.52	0.602	.2045775	2.508003

21

Potential Outcomes

$$E(Y(x_1)) = n^{-1} \sum_{i=1}^n \exp\{\beta_0 + \beta_1 x_1 + \beta_2 \text{bmi}_i + \beta_3 \text{age}_i\}$$

$$E(Y(x_0)) = n^{-1} \sum_{i=1}^n \exp\{\beta_0 + \beta_1 x_0 + \beta_2 \text{bmi}_i + \beta_3 \text{age}_i\}$$

Can calculate marginal summarizes based on comparison of x1 to x0

22

Margins Command

```
. margins, at(blbmd=(0.636 0.542)) post
```

```
1._at      : blbmd      =      .636
```

```
2._at      : blbmd      =      .542
```

		Delta-method		z	P> z	[95% Conf. Interval]	
		Margin	Std. Err.				
E(Y(x1)) ₁ ^{-at}		.1431981	.0084859	16.87	0.000	.1265662	.1598301
E(Y(x0)) ₂		.2298682	.0091738	25.06	0.000	.2118878	.2478485

$$RR = E(Y(x1)) / E(Y(x0)) = 0.143/0.230 = 0.62$$

23

Marginal Relative Risk

$$\log RR = \log (E(Y(x1))/E(Y(x0)))$$

```
. nlcom (logRR: log(_b[1._at]) - log(_b[2._at]))
```

```
logRR: log(_b[1._at]) - log(_b[2._at])
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
logRR	-.4732766	.0544051	-8.70	0.000	-.5799087	-.3666445

```
. dis exp( -.4732766)
.62295774
```

```
. dis exp(-.5799087)
.55994949
```

```
. dis exp( -.3666445)
.69305598
```

*Identical to the relative risk from
the conditional model*

24

Logistic Regression

```
. logistic fracture c.bmdsp*##c.bmi##c.age
```

Logistic regression

Number of obs = 2669
 LR chi2(11) = 99.27
 Prob > chi2 = 0.0000
 Pseudo R2 = 0.0376

Log likelihood = -1270.831

fracture	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
bmdsp1	.0060128	.0640367	-0.48	0.631	5.17e-12	6990156
bmdsp2	.4289443	5.758955	-0.06	0.950	1.60e-12	1.15e+11
bmi	.6878283	1.707098	-0.15	0.880	.0053078	89.13511
c.bmdsp1#c.bmi	1.091443	.4809919	0.20	0.843	.4601316	2.588929
c.bmdsp2#c.bmi	1.117461	.5956207	0.21	0.835	.3931258	3.176388
age	.7869287	.665148	-0.28	0.777	.1501278	4.124865
c.bmdsp1#c.age	1.054416	.1593224	0.35	0.726	.7841428	1.417845
c.bmdsp2#c.age	1.024268	.1990111	0.12	0.902	.6998882	1.49899
c.bmi#c.age	1.003063	.035319	0.09	0.931	.9361739	1.074732
c.bmdsp1#c.bmi#c.age	.9992013	.0062572	-0.13	0.898	.9870123	1.011541
c.bmdsp2#c.bmi#c.age	.997932	.0077008	-0.27	0.788	.9829522	1.01314
_cons	1.17e+10	6.97e+11	0.39	0.698	1.87e-41	7.32e+60

25

Margins After Logistic

```
. margins, at(blbmd=(0.636 0.542)) post
```

1._at : blbmd = .636

2._at : blbmd = .542

	Margin	Delta-method Std. Err.	z	P> z	[95% Conf. Interval]	
E(Y(x1)) ₁ ^{-at}	.1414725	.0094845	14.92	0.000	.1228833	.1600618
E(Y(x0)) ₂	.2359769	.0120128	19.64	0.000	.2124323	.2595215

$$RR = E(Y(x1)) / E(Y(x0)) = 0.141/0.236 = 0.60$$

Marginal Relative Risk

$$\log RR = \log (E(Y(x1))/E(Y(x0)))$$

```
. nlcom (logRR: log(_b[1._at]) - log(_b[2._at]))
```

```
logRR: log(_b[1._at]) - log(_b[2._at])
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
logRR	-.5116282	.0791684	-6.46	0.000	-.6667954 -.3564609

```
. dis exp( -.5116282)  
.59951865
```

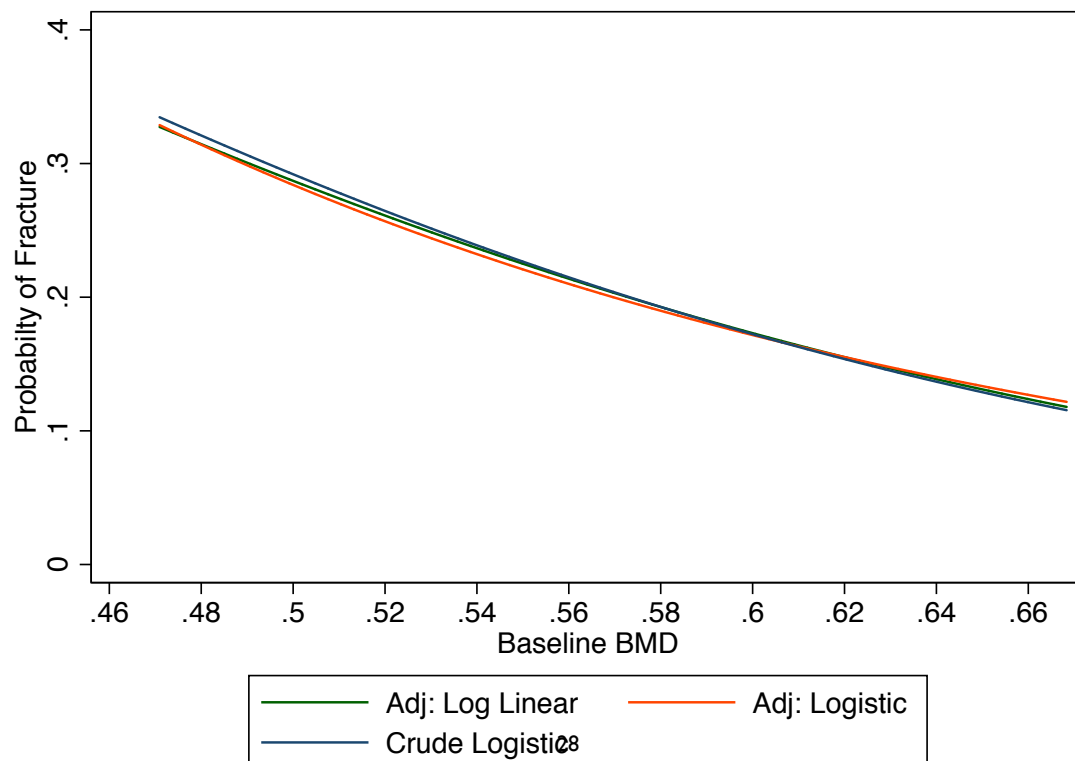
```
. dis exp(-.6667954)  
.51335103
```

Rather similar results

```
. dis exp(-.3564609)  
.70014985
```

27

Fracture Risk by BMD



The Big Idea

- If we have adjustment set C
- Model $E(Y | E, C)$ allows g-computation
- Could be very complex (machine learning)
- Just have to set values to $E=1, E=0$
- Plug into algorithm
- Extract predictions (CIs are harder)

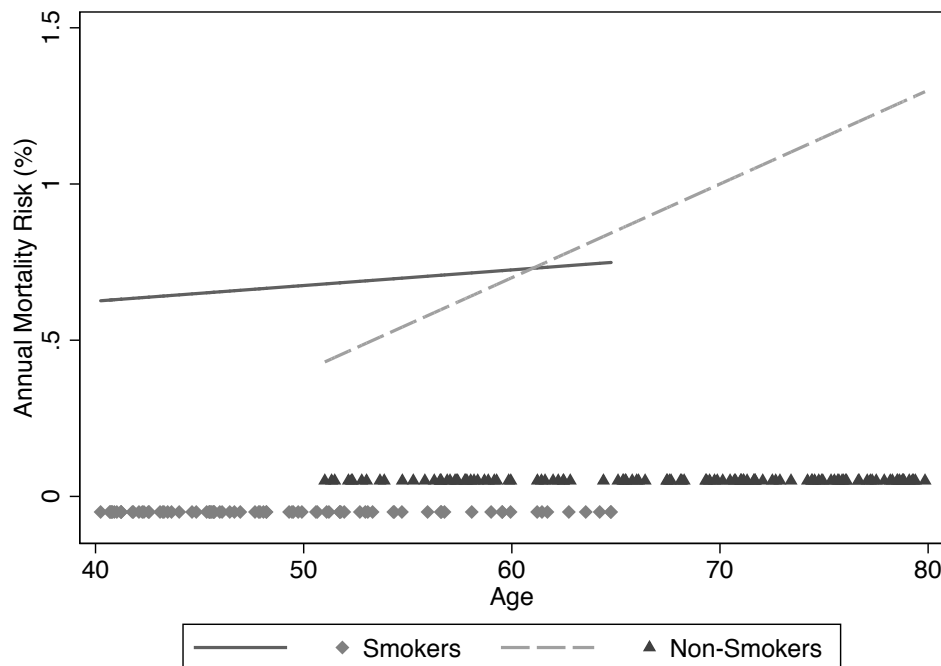
29

Smoking and Mortality

- Compare mortality in a group of younger smokers v. older non-smokers?
- Suppose age is only confounder?
- How to estimate the causal OR or HR
- How to estimate the ACE?

30

What is Causal Effect?



31

Overlap and Positivity

- Positivity: for every covariate combination there is a change of exposure and non-exposure. *Property of the population.*
- Overlap: there are obs with exposure and non exposure in all covariate combination. *Property of the data.*

32

Dilemma of Positivity

- Easy to assess with a single confounder
- In a $2 \times 2 \times 2$ table, immediately evident
- Clear from smoking/age graph
- More complex with multiple variables
- Overlap not clear from regression:
propensity offers some strategies

33

What if I lack overlap?

- Acknowledge it as a limitation
- Finesse the question: effect of smoking in middle aged men
- Could rely on extrapolation from regression
heavily dependent on modeling choices

34

The modeling dilemma

- Handling non-linearities
- Interactions
- These are particular concerns with
 - small samples
 - poor overlap
- Extrapolation is fundamentally different use of regression

35

Consider Interactions

- Quantitative v. qualitative interaction
- Absence can aid credibility:
effect doesn't vary with study site
- Presence can aid understanding:
effect more pronounced in young people
- Limit number of interactions:
false positive
- Consider most plausible ones first

36

No Unmeasured Confounders

- Essential
particularly important for weak effects
- Unverifiable
fundamental dilemma of causal inference
- A surrogate can stand in for unmeasured confounder
- Value for an adjusted effect even if it is not completely free of confounding

37

Potential Outcomes

- Pursues rigorous “causal” targets
- So-called “marginal” approach
- Comparing factors in standardized population
- Answer depends on confounder mix
- Doesn’t assume homogeneous effect
- Model $f(Y|C,E)$: familiar in that sense
- No clear generalization of hazard ratios

38