Causal Analysis Using SAS® Software

Clay Thompson
SAS Institute Inc.

BASUG Webinar
25 January 2023

Many research questions are causal in nature

What is the effect of **T** (treatment) on **Y** (outcome)?

How does smoking cessation affect body weight?

Can a program for at-risk youth reduce the juvenile crime rate?

Does music training enhance academic performance?

A causal effect is a contrast between potential outcomes

Neyman (1923), Rubin (1974)

Obs	Y_1
1	37.4
2	36.6
3	35.5
4	36.7
5	32.7
6	33.6
7	33.5
8	31.1

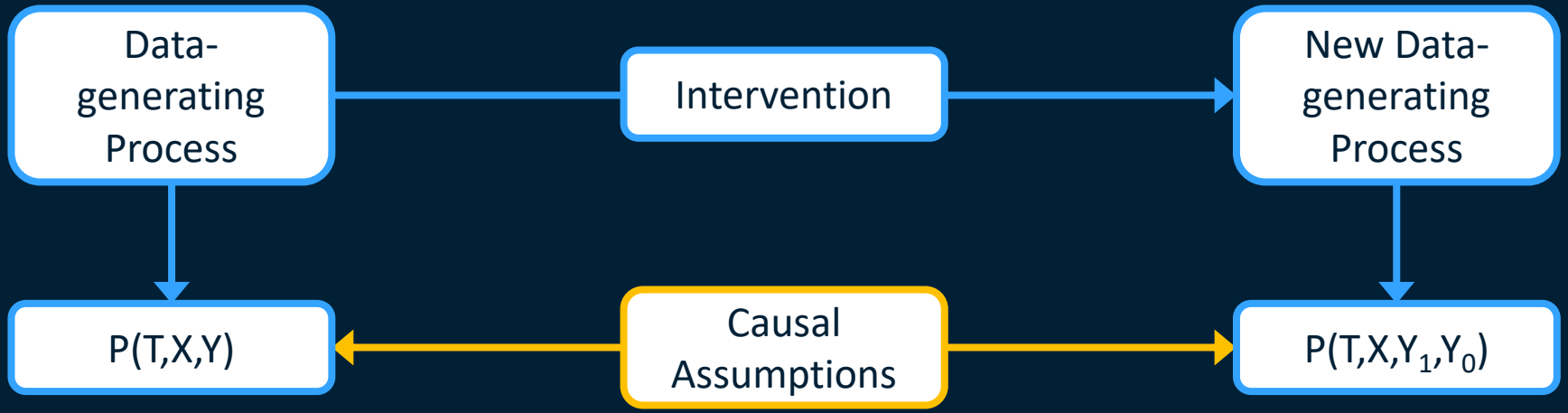
Obs	Y_0
1	39.5
2	38.0
3	37.3
4	38.0
5	34.3
6	35.0
7	35.4
8	32.4

Obs	T	Y_1	Y_0	Y
1	0	?	39.5	39.5
2	0	?	38.0	38.0
3	0	?	37.3	37.3
4	1	36.7	?	36.7
5	0	?	34.3	34.3
6	1	33.6	?	33.6
7	1	33.5	?	33.5
8	1	31.1	?	31.1

$$\text{ATE} = E[Y_1 - Y_0]$$

$$\text{ATT} = E[Y_1 - Y_0 \mid T=1]$$

A causal analysis is a statistical analysis plus causal assumptions



- Stable Unit Treatment Value Assumption (SUTVA)
- Causal Consistency
- Positivity
- No Unmeasured Confounders

There are three major approaches you can use to estimate a total treatment effect

		Treatment Model	
		No	Yes
Outcome Model	No		**PS weighting and matching methods**
	Yes	**Regression adjustment methods**	**Doubly robust methods, “causal ML”**

Outline

- Example: smoking cessation and body weight change
  - Estimating the ATT by matching on the propensity score
  - Estimating the ATT by inverse probability weighting
  - Estimating the ATE by regression adjustment
  - Estimating the ATE with doubly robust methods

- Example: preK enrollment and subsequent academic performance
  - Using a directed acyclic graphic to choose model covariates
  - Exploring causal mechanisms through mediation analysis

What is the effect of quitting smoking on body weight change?

Adapted from Hernán & Robins (2023)

- Data: A subset of NHANES 1 Epidemiologic Follow-Up Study (NHEFS)
- Collect medical and behavioral information in an initial physical examination
- Follow-up interviews completed approximately 10 years later
- Treatment (Quit): indicator of smoking cessation during the 10-year period
- Outcome (Change): change in body weight (kg)
- Assume all missingness is ignorable

The data include a subject's level of physical activity, smoking habits, and demographic information

- Activity: Level of daily activity (0, 1, 2)
- Age: Age in 1971 (yrs)
- BaseWeight: Weight in 1971 (kg)
- Education: Level of education (0,1,2,3,4)
- Exercise: Level of regular recreational exercise (0,1,2)
- PerDay: Number of cigarettes smoked per day
- Race: 0 for white; 1 otherwise
- Sex: 0 for male; 1 for female
- Weight: Weight at the follow-up interview (kg)
- YearsSmoke: Number of years a subject has smoked

Estimating the ATT by matching on the propensity score

PROC PSMATCH

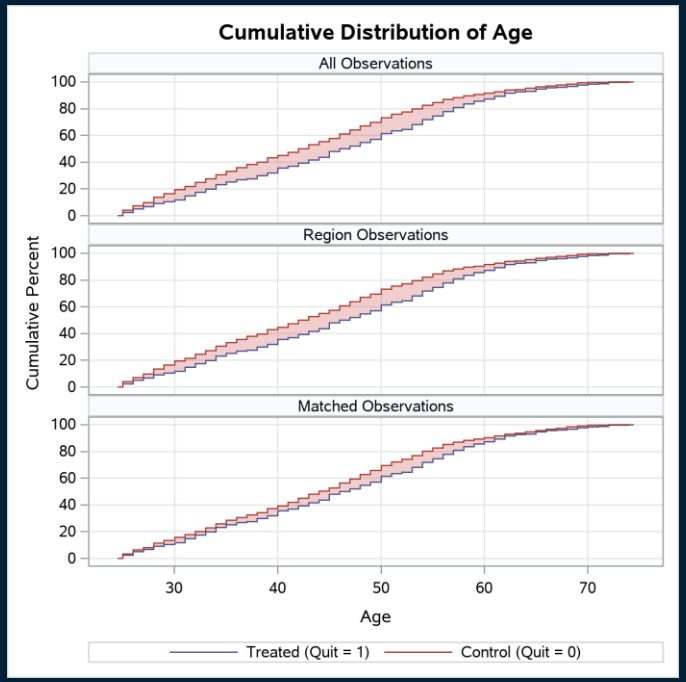
A propensity score–based matching analysis involves three important steps

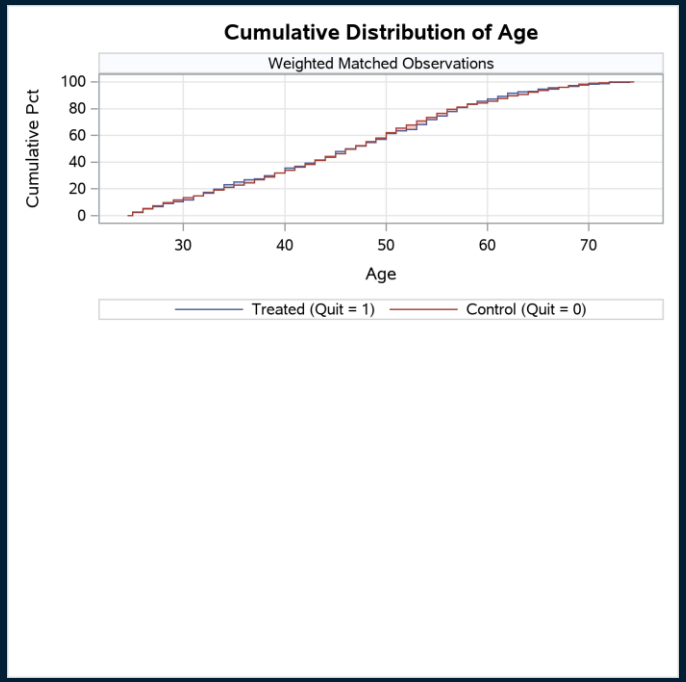

```
data SmokingNoResp;
    set SmokingWeight;
    drop change weight;
run;
```

PROC PSMATCH can fit a PS model, perform matching, assess balance, and create an output data set

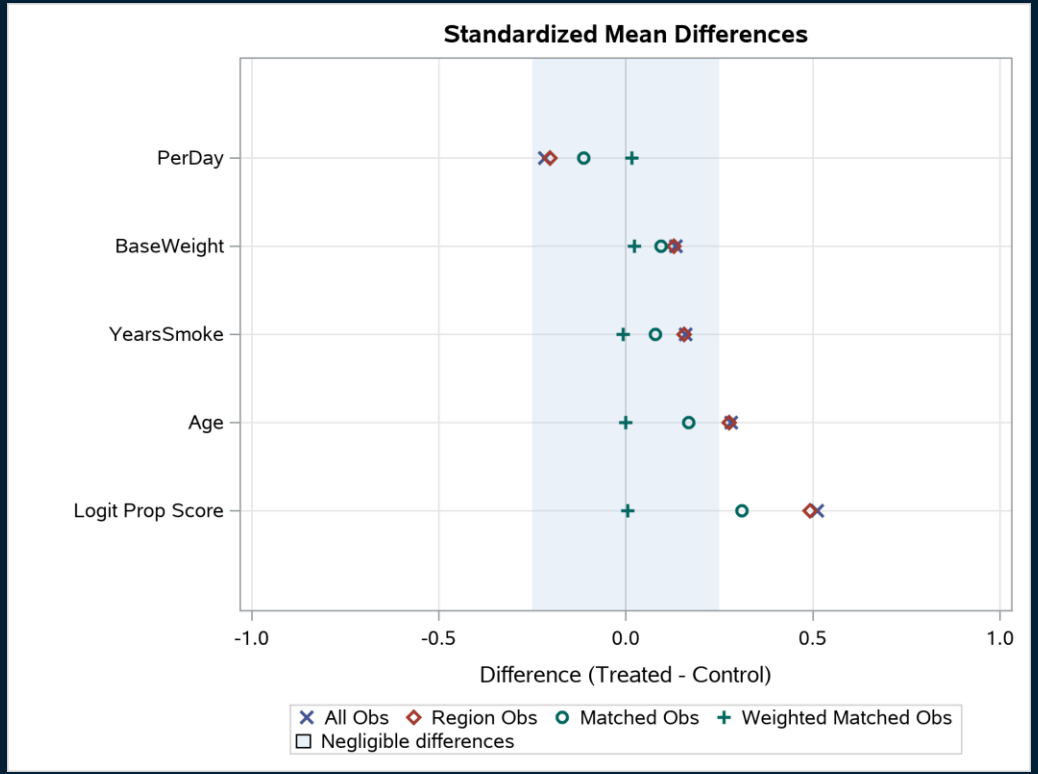
```
proc psmatch data=SmokingNoResp;
   class Activity Education Exercise Quit Sex;
   psmodel Quit(Treated='1') = Sex Age Education Exercise Activity
                               YearsSmoke PerDay;
   match distance=lps
         method=varratio(kmin=1 kmax=4)
         caliper=.5;
   assess lps var=(Age YearsSmoke BaseWeight PerDay) /
              plots=(CDFPlot BoxPlot StdDiff);
   output out(obs=all)= SmokeMatched1 weight=matchattwgt
                                      matchid=MatchID;
run;
```

In the input data set, subjects who quit smoking tended to be older than those who did not quit





A standardized mean differences plot provides a concise graphical assessment of balance for multiple covariates



For the outcome analysis, use PROC TTEST with the weights from PSMATCH

```
data SmokeAnalyze1;
   merge SmokeMatched1
         SmokingWeight;
   by ID;
run;

proc ttest data=SmokeAnalyze1;
   class Quit;
   var Change;
   weight matchattwgt;
run;
```

Quit	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		1.2458	0.7992	1.6925	4.5692	4.3781	4.7779
1		4.5251	3.6684	5.3818	8.7483	8.1831	9.3979
Diff (1-2)	Pooled	-3.2792	-4.1171	-2.4414	6.0627	5.8469	6.2951
Diff (1-2)	Satterthwaite	-3.2792	-4.2447	-2.3138			

Estimating the ATT by inverse probability weighting

PROC PSMATCH

PROC CAUSALTRT

PROC PSMATCH uses IPW when no MATCH or STRATA statement is specified

```
proc psmatch data=SmokingWeight;
   class Sex Education Exercise Activity Quit;
   psmodel Quit(Treated='1') = Sex Education Age Exercise Activity
                               YearsSmoke PerDay;
   assess lps var=(Age YearsSmoke BaseWeight PerDay) /
              plots=(CDFPlot BoxPlot WgtCloud(ref=6));
   psweight nlargest=5 weight=att;
   output out=SmokeIPW1 weight=attwgt;
   id ID;
run;
```

```
proc genmod data=SmokeIPW1;
   class Quit(desc) ID;
   model Change = Quit;
   repeated subject=ID;
   weight attwgt;
run;
```

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.2495	0.2448	0.7697	1.7292	5.10	<.0001
Quit	1	3.2756	0.4993	2.2969	4.2543	6.56	<.0001
Quit	0	0.0000	0.0000	0.0000	0.0000	.	.

PROC CAUSALTRT can directly estimate the ATT by IPW

```
proc causaltrt data=SmokingWeight method=ipwr att;
   class Sex Education Activity Quit;
   psmodel Quit(Event='1') = Sex Education Age Exercise Activity
                             YearsSmoke PerDay;
   model Change;
   bootstrap seed=1776;
run;
```

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > |Z|
POM	1	4.5251	0.4352	0.4282	3.6720	5.3781	3.7187	5.3879	10.40	<.0001
POM	0	1.2495	0.2565	0.2595	0.7467	1.7522	0.7345	1.7439	4.87	<.0001
ATT		3.2756	0.4815	0.4855	2.3319	4.2193	2.3671	4.2215	6.80	<.0001

Estimating the ATE by regression adjustment

PROC CAUSALTRT

PROC GLIMMIX

bart Action Set, PROC BART

PROC CAUSALTRT fits a generalized linear model separately within each treatment condition

```
proc causaltrt data=SmokingWeight;
   class Activity Education Exercise Quit Sex / descending;
   psmodel Quit(Event='1');
   model Change = Activity Age Education Exercise
                  PerDay Sex YearsSmoke;
run;
```

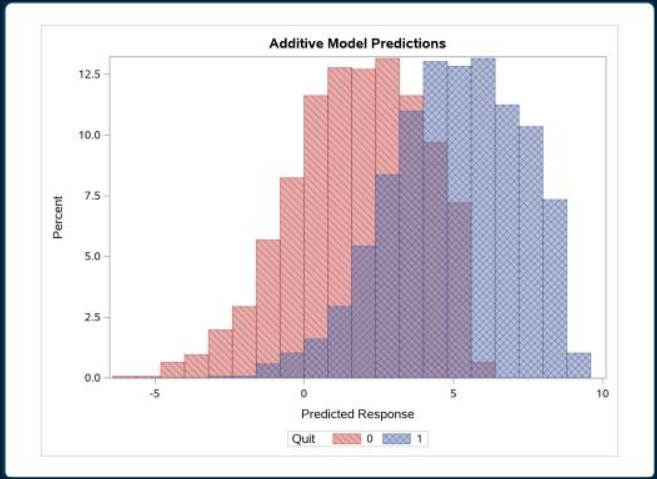
Analysis of Causal Effect

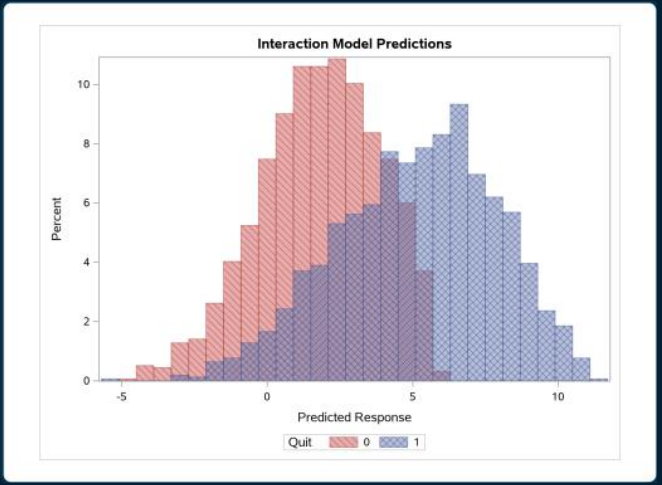
Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	1	5.1407	0.4638	4.2317	6.0496	11.08	<.0001
POM	0	1.8160	0.2163	1.3921	2.2399	8.40	<.0001
ATE		3.3247	0.5072	2.3306	4.3188	6.55	<.0001

A model with all treatment-confounder interactions is comparable to models fit separately

```
proc glimmix data=smokingweight;
   class Sex Race Education Exercise Activity Quit ;
   model Change =  Quit Sex Age Education Exercise
                   Activity YearsSmoke PerDay ;
   margins Quit / diff;
   lsmeans Quit / diff om;
run;
```

```
proc glimmix data=SmokingWeight;
   class Sex Race Education Exercise Activity Quit ;
   model Change =  Quit Sex*Quit Age*Quit Education*Quit
                   Exercise*Quit Activity*Quit
                   YearsSmoke*Quit PerDay*Quit ;
   margins quit / diff;
   lsmeans quit / diff om;
run;
```





A model with all treatment-confounder interactions is comparable to models fit separately

```
proc glimmix data=smokingweight;
   class Sex Race Education Exercise Activity Quit ;
   model Change =  Quit Sex Age Education Exercise
                   Activity YearsSmoke PerDay ;
   margins Quit / diff;
   lsmeans Quit / diff om;
run;
```

```
proc glimmix data=SmokingWeight;
   class Sex Race Education Exercise Activity Quit ;
   model Change =  Quit Sex*Quit Age*Quit Education*Quit
                   Exercise*Quit Activity*Quit
                   YearsSmoke*Quit PerDay*Quit ;
   margins quit / diff;
   lsmeans quit / diff om;
run;
```

Quit Margins

Quit	Estimate	Standard Error	DF	t Value	Pr > |t|
0	1.7994	0.2223	1552	8.09	<.0001
1	5.0591	0.3821	1552	13.24	<.0001

Differences of Quit Margins

Quit	_Quit	Estimate	Standard Error	DF	t Value	Pr > |t|
0	1	-3.2597	0.4461	1552	-7.31	<.0001

Quit Margins

Quit	Estimate	Standard Error	DF	t Value	Pr > |t|
0	1.8160	0.2225	1540	8.16	<.0001
1	5.1407	0.4023	1540	12.78	<.0001

Differences of Quit Margins

Quit	_Quit	Estimate	Standard Error	DF	t Value	Pr > |t|
0	1	-3.3247	0.4597	1540	-7.23	<.0001

Bayesian additive regression trees (BART) are a popular model type for effect estimation

```
proc bart data=mycas.SmokingWeight
     seed=1972 trainInMem;
   class Sex Race Education
         Exercise Activity Quit;
   model Change = Quit Sex Age
         Education Exercise
         Activity YearsSmoke
         PerDay;
   store mycas.swModel;
run;
```

```
proc cas;
   action bart.bartScoreMargin /
   table = {name="smokingWeight"}
   restore = {name="swModel"}
   margins= {
      { name="Cessation",
        at={{var="Quit" value="1"}}
      }
      { name="No Cessation",
        at={{var="Quit" value="0"}}
      }}
   differences = {
      { label="Cessation Difference"
        refMargin="No Cessation"
        evtMargin="Cessation"} };
   run;
quit;
```

Predictive Margins

Description	Estimate	95% Equal-Tail Interval	
Cessation	5.21910	4.49393	5.95288
No Cessation	1.77900	1.39608	2.21010

Predictive Margin Differences

Description	Estimate	95% Equal-Tail Interval	
Cessation Difference	3.4401	2.5413	4.2345

Estimating the ATE with doubly-robust methods

PROC CAUSALTRT

causalAnalysis.caEffect

deepEcon.deepCausal

A doubly robust method requires that you specify models for both the treatment and the outcome

```
proc causaltrt data=SmokingWeight covdiffps plots=outcomebyweight;
   class Activity Education Exercise Quit Sex / descending;
   psmodel Quit(Event='1') = Activity Age Education Exercise
                             PerDay Sex YearsSmoke;
   model Change = Activity Age BaseWeight Exercise Sex;
   bootstrap seed=1682 plots(unpack)=hist;
run;
```

Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > |Z|
POM	1	5.0832	0.4495	0.4637	4.2021	5.9643	4.1806	5.9854	11.31	<.0001
POM	0	1.7783	0.2156	0.2171	1.3557	2.2009	1.3152	2.1702	8.25	<.0001
ATE		3.3049	0.4911	0.4943	2.3423	4.2675	2.2922	4.2748	6.73	<.0001

Use TMLE to incorporate machine learning methods into semiparametric efficient estimators

- Machine learning methods excel at predicting outcomes
  - Corresponding confidence intervals are typically absent or insufficient
  - For causal problems, you need to predict Y_t, not Y
- TMLE is
  - Non-/semiparametric
  - Doubly robust
  - Maximally efficient
  - Substitution estimator

TMLE Part I: Create a propensity score model

```
regression.logistic /
    class={"Sex", "Race", "Education",
           "Exercise", "Activity"},
    model={depvar={{name="Quit", options={event="1"}}},
           effects={"Sex", "Race", "Education", "Exercise",
                    "Activity", "Age",
                    "YearsSmoke", "PerDay"}},
    output={casout={name="swDREstData", replace="True"},
            copyvars="All",
            pred="pTrt"},
    table="SmokingWeight";
run;
```

TMLE Part II: Create an outcome model

```
bart.bartGauss /
    inputs={"Sex", "Race", "Education", "Exercise", "Quit",
            "Activity", "Age", "YearsSmoke", "PerDay"},
    nMC="200",
    nTree="100",
    nominals={"Sex", "Race", "Education", "Exercise",
              "Activity", "Quit"},
    seed="2156",
    store={name="bartOutMod", replace="True"},
    table="swDREstData",
    target="Change";
run;
```

TMLE Part III: Estimate the causal effect

```
causalAnalysis.caEffect /
    difference={{evtLev="1"}},
    method="aipw",
    outcomeModel={predName="P_Change",
                  store="bartOutMod"},
                  outcomeVariable="Change",
    pom={{trtLev="1", trtProb="pTrt"},
         {trtLev="0", trtProb="pCnt"}},
    table="swDREstData",
    treatVar="Quit";
run;
```

POM Differences		
Treatment Levels		
Event	Reference	Difference
1	0	3.3375

POM Estimates	
Treatment Level	Estimate
1	5.12564
0	1.78811

You can use the deepEcon.deepCausal action to implement doubly/debiased machine learning (DML) methods based on dNNs!

What is the effect of PreK enrollment on subsequent student performance?

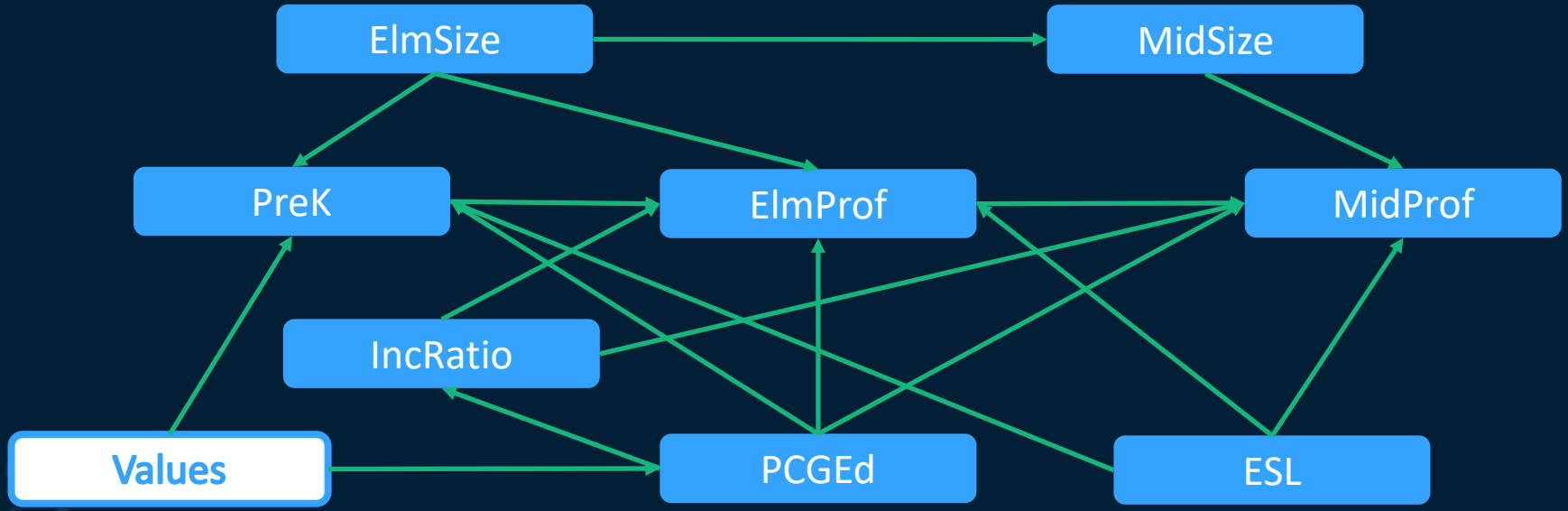
A simulated example

- PreK: indicator for enrollment in a PreK program
- MidProf: indicator for reading proficiency at the end of 8th grade
- ElmProf: indicator for reading proficiency at the end of 4th grade
- ElmSize: average 4th grade class size
- ESL: indicator for English as a second langange
- IncRatio: ratio off household income to the federal poverty line
- MidSize: average 8th grade class size
- PCGEd: classification variable for primary caregiver's education

Using a directed acyclic graph to choose model covariates

PROC CAUSALGRAPH

A causal diagram is a directed acyclic graph that encodes causal assumptions about the data generating process



Determine which covariates form a valid statistical adjustment to estimate a causal effect

```
proc causalgraph;
   model "ReadingProf"
      ElmProf => MidProf,
      ElmSize => ElmProf MidSize PreK,
      ESL IncRatio => ElmProf MidProf PreK,
      MidSize => MidProf,
      PCGEd => ElmProf IncRatio MidProf PreK,
      PreK => ElmProf,
      Values => PCGEd PreK;
   latent Values;
   identify PreK => MidProf;
run;
```

Covariate Adjustment Sets for ReadingProf

Causal Effect of PreK on MidProf

	Size	Minimal	Covariates					
			ElmProf	ElmSize	ESL	IncRatio	MidSize	PCGEd
1	4	Yes		*	*	*		*
2	5	No		*	*	*	*	*

Use the covariates from PROC CAUSALGRAPH with your preferred method of estimation

```
proc causaltrt data=ReadingObs;
   class ESL MidProf PCGEd PreK / desc;
   psmodel PreK = ElmSize ESL IncRatio PCGEd;
   model MidProf = ElmSize ESL IncRatio PCGEd;
   bootstrap seed=1976;
run;
```

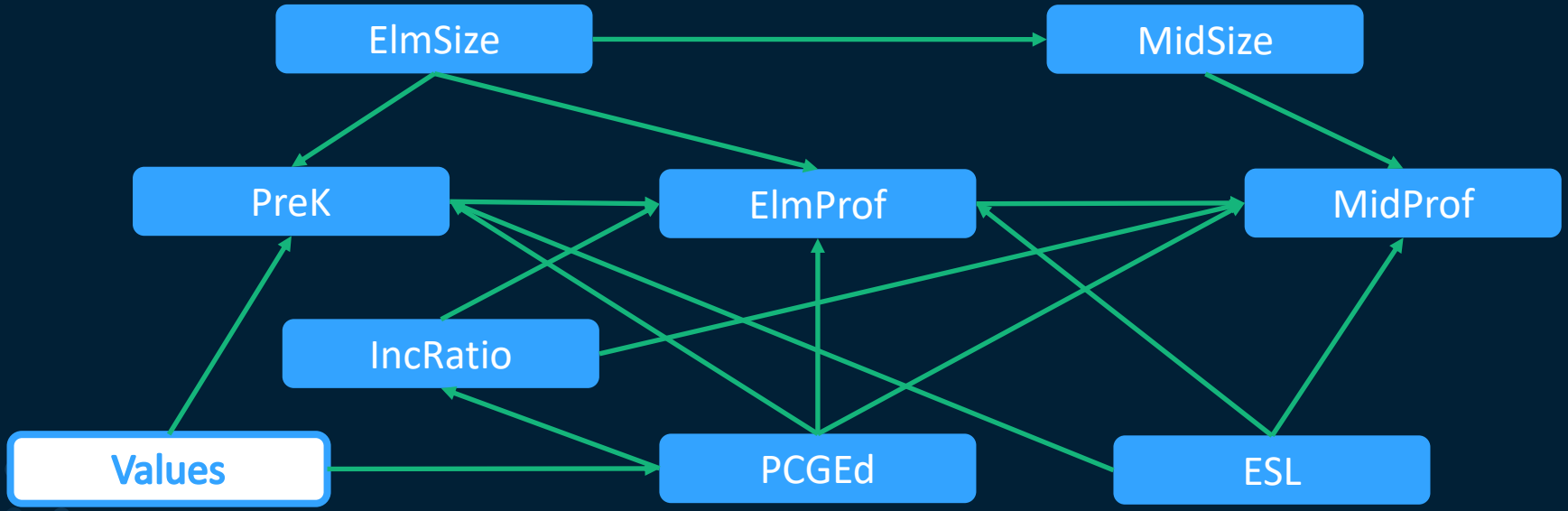
Analysis of Causal Effect

Parameter	Treatment Level	Estimate	Robust Std Err	Bootstrap Std Err	Wald 95% Confidence Limits		Bootstrap Bias Corrected 95% Confidence Limits		Z	Pr > |Z|
POM	1	0.7855	0.00670	0.00646	0.7723	0.7986	0.7718	0.7979	117.18	<.0001
POM	0	0.7528	0.00562	0.00574	0.7417	0.7638	0.7419	0.7644	133.90	<.0001
ATE		0.03270	0.00872	0.00873	0.01561	0.04980	0.01487	0.04813	3.75	0.0002

Exploring causal mechanisms through mediation analysis

PROC CAUSALMED

To what extent is the effect of interest mediated by improved proficiency in elementary school?



A mediation analysis decomposes the total effect into direct and indirect components

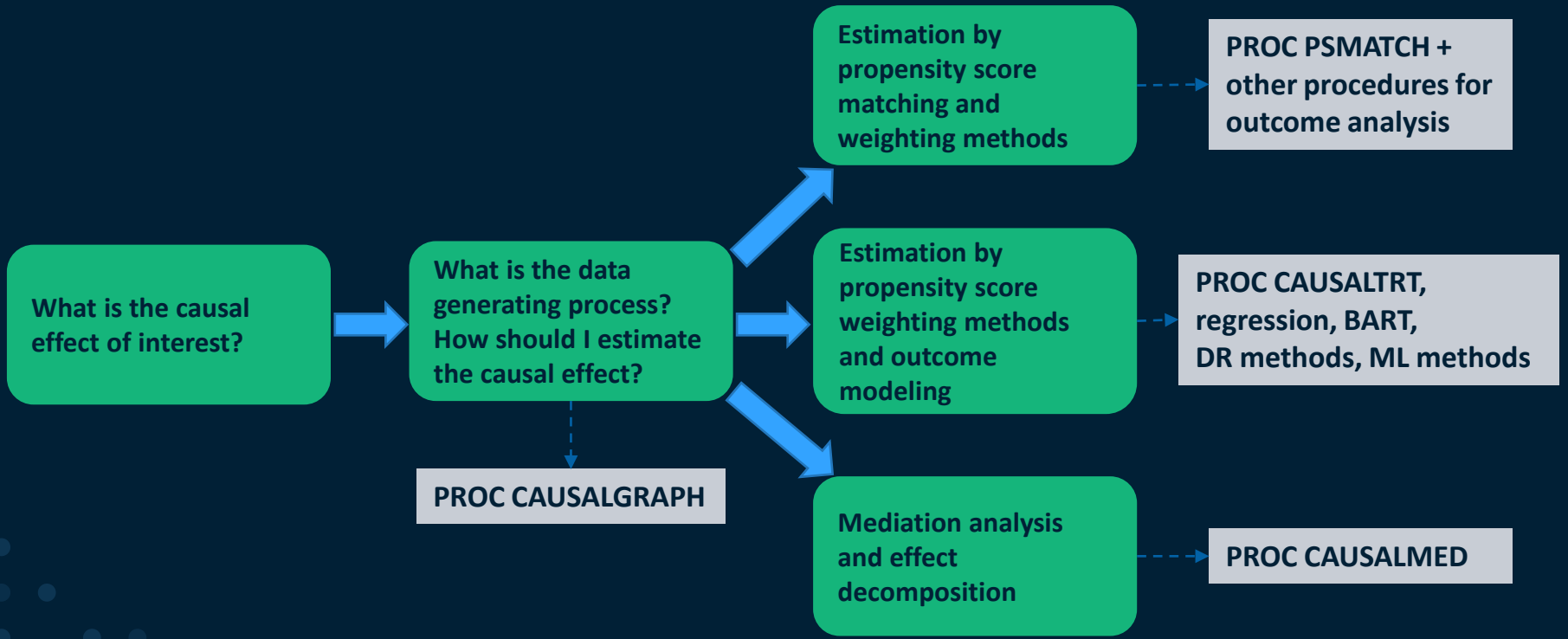
```
proc causalmed data=ReadingObs;
   class ESL PCGEd PreK / desc;
   model MidProf = Prek | ElmProf;
   mediator ElmProf = PreK;
   covar ElmSize ESL IncRatio PCGEd;
run;
```

Summary of Effects

	Estimate	Standard Error	Wald 95% Confidence Limits		Z	Pr > |Z|
Total Effect	0.03396	0.008735	0.01684	0.05108	3.89	0.0001
Controlled Direct Effect (CDE)	0.02537	0.008546	0.008624	0.04212	2.97	0.0030
Natural Direct Effect (NDE)	0.02555	0.008559	0.008778	0.04233	2.99	0.0028
Natural Indirect Effect (NIE)	0.008407	0.001927	0.004631	0.01218	4.36	<.0001
Percentage Mediated	24.7556	7.6304	9.8003	39.7109	3.24	0.0012
Percentage Due to Interaction	-0.7598	1.4251	-3.5530	2.0333	-0.53	0.5939
Percentage Eliminated	25.2832	7.6608	10.2683	40.2981	3.30	0.0010

Summary

You can build a causal analysis workflow with SAS procedures and actions



Causal analysis procedures in SAS 9

Procedure	Primary Use	Release (Year)
PROC PSMATCH	Assessment of covariate balance; creation of matched data sets for causal effect estimation	SAS 9.4M4 (2016) SAS/STAT 14.2
PROC CAUSALTRT	Direct estimation of a causal effect	SAS 9.4M4 (2016) SAS/STAT 14.2
PROC CAUSALMED	Decomposition of a (total) causal effect into direct and indirect effects	SAS 9.4M5 (2017) SAS/STAT 14.3
PROC CAUSALGRAPH	Analysis of graphical causal models; construction of sound statistical strategies for causal effect estimation	SAS 9.4M6 (2018) SAS/STAT 15.1

Causal analysis procedures in SAS Viya 4

Procedure	Primary Use	Release (MM/YY)
bart Action Set, PROC BART (SAS Visual Statistics)	Bayesian additive regression trees, including predictive margins	2022.1.1 (05/22) 2022.09 LTS (09/22)
causalanalysis Action Set (SAS Visual Statistics)	Estimation of potential outcome means and causal effects	2022.11 (11/22)
deepecon Action Set, PROC DEEPCAUSAL (SAS Econometrics)	Doubly/debiased machine learning of causal effects and policies via dNNs	2021.1.4 (08/21) 2021.2 LTS (10/21)

The SAS Programming Runtime Environment (SPRE) in Viya 4 provides access to licensed SAS 9 PROCs

Thank you

sas.com