SAS® GLOBAL FORUM 2021

Paper 1151-2021

Getting Started with Structural Causal Modeling in SAS/STAT

John Amrhein, McDougall Scientific Ltd.

INTRODUCTION

In a broad sense, correlation is any relationship between two variables. Causation is a stronger relationship because the value of one variable induces the second variable to take its value. Causation can be quantified when the two variables are measured close in time, but with the cause always preceding the outcome. To establish a causal relationship, we must isolate the cause and outcome from other possible causal factors by intervening in the world we are measuring. To say this another way, causation is all about understanding the process that generated our data. It is not enough to discover relationships in the data, we must understand how those relationships were created.

This paper is a gentle introduction to causal inference using the method of structural causal models. The first section defines causality and describes the conditions necesary to declare causality from an analysis. This is followed by an introduction to directed acyclic graphs, or DAGs, which are instrumental to understanding and fitting structural models. Using the SASHELP.HEART data set, a DAG is constructed to support several analyses. For several causal factors of interest, we use the CAUSALGRAPH Procedure to identify adjustment sets containing factors that confound the causal factor of interest. Having identified the confounders, we use the CAUSALMED and CAUSALTRT Procedures to fit our structural causal models to estimate causal effects.

CONCEPTS AND PROPERTIES OF CAUSALITY

Causality Defined

An outcome O causally depends on a prior event C if and only if * The occurrence of C implies that O will occur, and * The absence of C implies that O will not occur.

However, the world is usually not so simple because outcomes require a series of prior causes to occur. For example, eating too much cheesecake will result in weight gain, but only if you do not exercise enough to counter the caloric intake. Therefore, at least two causes are necessary, eating too much cheesecake and not exercising. A causal mechanism or chain of causal dependence is one in which C causes O if and only if there is a sequence of events C, D_1, D_2, $\ldots$, D_k, O such that each event in the sequence causally depends on the previous.

Note that the definition concerns the manner in which the data are generated.

Counterfactual Thinking

If you observe both events C and O, then how can you know whether O will be absent in the absence of C? You cannot reverse time and refrain from eating cheesecake to observe whether you gain weight. Counterfactual thinking is imagining a world that was not observed but might have been.

The fundamental problem of causality is that, at the level of an individual (or experimental unit), only one world can exist, and the alternative cannot be observed; only one outcome can be observed on each unit. For example, it is impossible to verify if abstaining from eating cheesecake results in no weight gain. The good news is that, although unit-level causality cannot be quantified, under specific conditions that you will read about shortly, population-level causal effects can be.

Potential Outcomes

Potential outcomes is a concept that organizes and establishes a formal framework for counterfactual thinking. Potential outcomes are events that are possible under alternative paths. For example, an intervention of interest might be eating cheesecake and the outcome of interest is weight gain. We have two possible paths; eat the cheesecake and refrain from eating the cheesecake, and two potential outcomes, weight gain or no weight gain.

Unit	Treatment Outcome	No Treatment Outcome	Causal Effect
1	Y_1(1)	Y_1(0)	Y_1(1) - Y_1(0)
2	Y_2(1)	Y_2(0)	Y_2(1) - Y_2(0)
3	Y_3(1)	Y_3(0)	Y_3(1) - Y_3(0)
...	...	...	...
N	Y_N(1)	Y_N(0)	Y_N(1) - Y_N(0)
Mean	Y(1)	Y(0)	Y(1) - Y(0)

Table 1. Potential Outcomes

Table 1 shows N individuals in an experiment. Y_i(t) is the potential outcome for the i^{th} individual; t=1 indicates treatment and t=0 indicates control (no treatment). For each individual, the observed outcome is shown in black whereas the alternative outcome is in red. Only one is possible because each individual can either receive the treatment or not. The difference between the treatment and no-treatment outcomes is the causal effect, which cannot be estimated at the individual level. To estimate a causal effect, we calculate the mean response for treatment, calculate the mean response under no treatment, and, after controlling for confounders (which we will discuss at length), compute the difference. This is known as the average causal effect (ACE) or average treatment effect (ATE).

There are a few assumptions that are needed to properly estimate ACE or ATE.

Assumptions for Causal Modeling

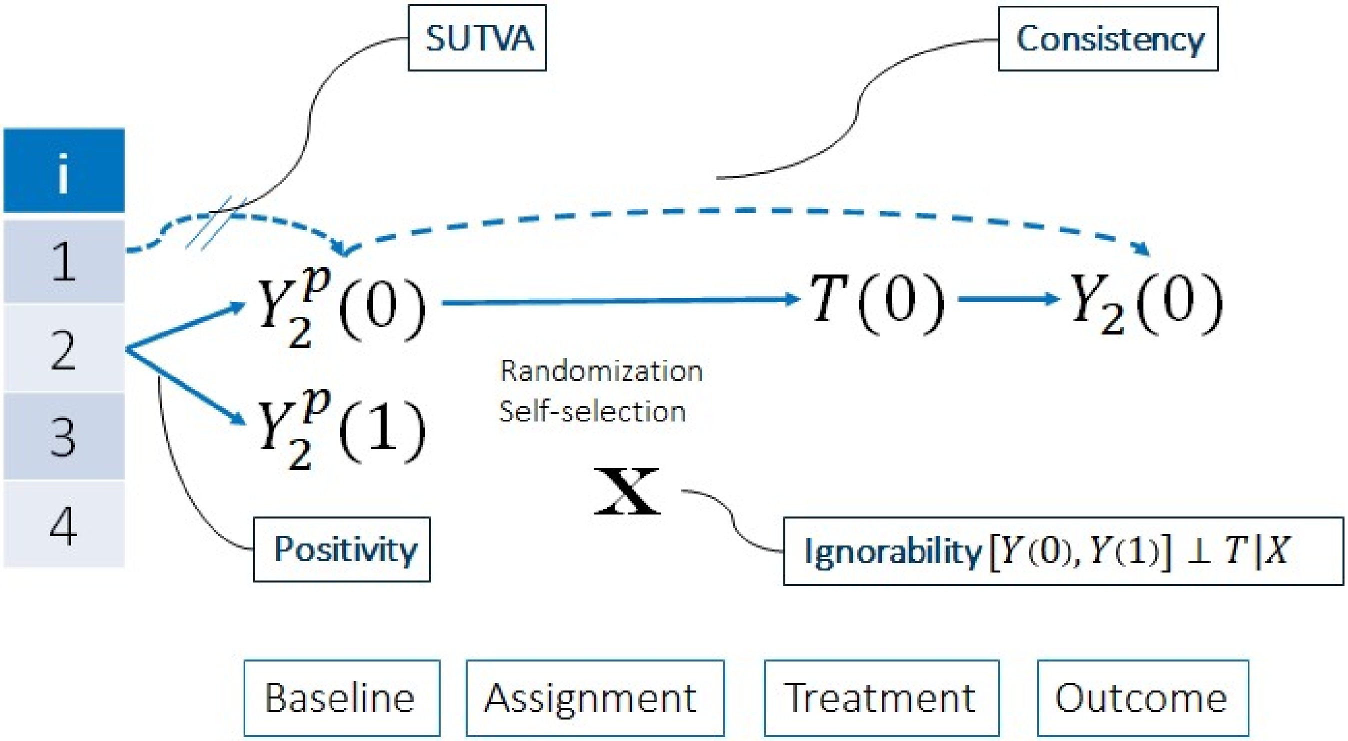


Figure 1. Assumptions for Causal Models

Figure 1 illustrates the conditions that must be met to declare cause and effect from your analysis. * Positivity: Every study subject (unit) has a positive probability of being assigned to each of the treatments. * Stable Unit Treatment Value Assignment (SUTVA): A unit's potential outcome is not affected by another unit's treatment assignment. * Consistency: A unit's potential outcome under a treatment is its realized outcome. * Ignorability: Given an appropriate set of confounders (Xs), a unit's mechanism for being assigned to treatment can be ignored.

Without positivity, it does not make sense to consider potential outcomes; if an individual has no chance of receiving one of the treatments, then there is no causal effect of treatment for that individual. SUTVA is necessary for causal effect to be defined, at the individual level, using only that individual. Sometimes it might be necessary to increase the number of potential outcomes to satisfy SUTVA, but that is a complication beyond the scope of this paper. Consistency allows us to use observed outcomes to estimate causal effect.

Ignorability is also known as unconfoundedness, 'no missing confounders', or exchangeability. Satisfying this assumption is at the heart of isolating a cause of interest from other possible causes. Formally, we assume

$[Y_i(0), Y_i(1)]$ |, where X is the set of confounding covariates. The key to a valid estimate of ACE is to account for confounding factors; that is, factors related to both the treatment assignment and the outcome. For example, in our cheesecake example, suppose we also measure each individual's rating about the importance of good health. One can imagine that the more important good health is to someone, the more probable that person will avoid cheesecake and the more likely they are to exercise during the study. An individual's view about health is related both to the treatment (eating cheesecake) and the outcome (weight gain). Therefore, we must control for opinions about health to understand the cause-and-effect relationship between eating cheesecake and weight gain. Later we will use graphs to visualize such relationships.

There are several methods to control for confounders; via randomization, propensity scores, and stratification. We will learn how to choose an appropriate set of confounders, called an adjustment set, when we introduce the CAUSALGRAPH Procedure.

METHODS OF CAUSAL ANALYSES

Randomized Controlled Trials

Randomized Controlled Trials (RCT) have been the gold standard to estimate causal effects. RCTs are a broad class of experimental designs, usually in which humans are the experimental unit. RCTs are common in medical and drug development studies. The 'C' in RCT indicates a control group in the study against which a treatment or intervention is compared. The control group does not receive the treatment or intervention under study. The 'R' in RCT indicates randomization; i.e. the study participants are randomly assigned to the treatment or control group. Randomization eliminates the ability of study participants to choose whether they receive the treatment, thereby removing potential confounding between causal factors. For example, suppose an intervention is being studied to determine if it helps job seekers find work. If one of the sexes is more inclined to participate in the intervention and is more likely to find work, then sex confounds the effect of the intervention. Randomization balances the distributions of sex between the intervention and control groups, nullifying the potential confounding by sex by not allowing individuals to choose their group. To declare causality of the intervention, this must hold true for all potential confounding factors associated with individuals (e.g. age, income, etc.). Randomization isolates the intervention from other causal factors by balancing the distribution of confounders between intervention and control.

A confounding factor is one that is related to both the treatment (the cause) and the outcome (the effect). If individuals with a specific profile are more likely to choose a treatment compared to those in other profiles, then the distributions of the covariates making up the profile might differ across the treatment groups, inducing relationships between cause and covariates. Randomization prevents these relationships. Formally, we can accept the assumption that potential outcomes and treatment assignment are independent; Y(t) T, where T is treatment assignment. This is known as exchangeability in the RCT and epidemiological literature (see Greenland and Robins); we called this "ignorability" in the section about assumptions. Exchangeability means that the mean outcome of the treated would remain the same even if all subjects were treated, or the mean outcome of the control would be the same even if all subjects were not treated. Said another way, if the two groups were swapped, the resulting outcome means would not change.

Modeling the Treatment: Propensity Scores

Some studies cannot use randomization; for example, studies that are not prospective but instead use data that have already been collected for other reasons (e.g. medical records, customer transactions, utility consumption). Studies analyzing retrospective data have no opportunity to randomly assign individuals to treatment. However, observational data often contain the variables needed to conduct an analysis of interest. Propensity scores offer a way to isolate the cause of interest from confounding factors when randomization was not used.

Propensity scores are probabilities of being assigned to the treatment group, t = 1; Pr(T=1|X=x). They are estimated, usually by logistic regression, by specifying the binary treatment indicator as the response and covariates as predictors. A new data set is created in which the propensity scores are used to: * Match treatment and control observations by similar propensity scores for matched (pairs) analysis * Compute weights for a weighted analysis * Stratify observations by similar propensity scores to conduct a stratified analysis

The idea is to create conditional independence between the potential outcomes and the treatment assignment. Recall that under complete randomization we can assume [Y_i(0), Y_i(1)] T. Using propensity scores, we assume [Y_i(0), Y_i(1)] T|X, where X is the set of confounding covariates used to estimate the propensity scores. The difficulty is finding the appropriate set X, noting that all confounders must be in X; we cannot have any unobserved confounding.

The PSMATCH Procedure provides methods of weighting, stratification, and matching. The objective is to achieve a balance of confounder distributions across treatment groups. Therefore, PROC PSMATCH assesses covariate balance by comparing distributions between the adjusted treated and control groups and creates a new data set for analysis of the cause of interest. You must specify the set X that PROC PSMATCH uses to estimate the propensity scores. PROC PSMATCH will not help you identify an appropriate set X, nor will it tell you if you have used an appropriate set.

See Lamm et al. (2019) for more information about potential outcomes analysis using PROC PSMATCH.

You can also use the CAUSALTRT Procedure to model propensity scores, which we introduce at the end of the next section.

Modeling the Treatment: Structural Causal Models

We now turn our attention to structural causal modeling (SCM). “Structural” refers to a pre-specified data generating process which we will represent graphically. “Causal” implies that the graphical model will adhere to conditions that permit causal interpretation of the relationships.

In this section we describe visual representation of SCMs using directed acyclic graphs, define three graph constructs useful for understanding adjustment sets, link graphical models to model analyses, and demonstrate the use of the procedures in the CAUSAL family of procedures: CAUSALGRAPH, CAUSALMED, and CAUSALTRT.

Directed Acyclic Graphs

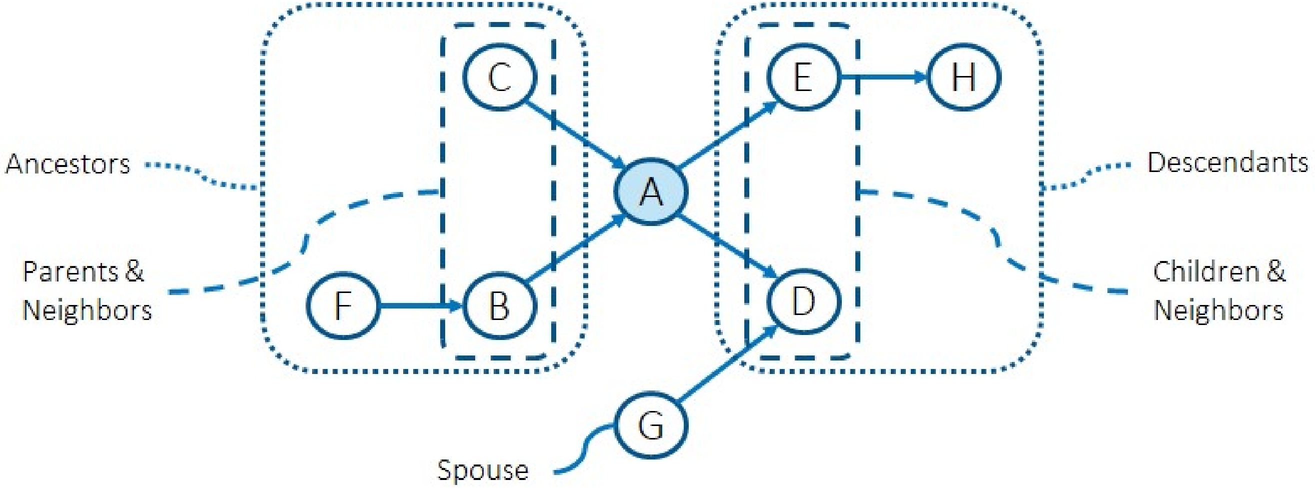
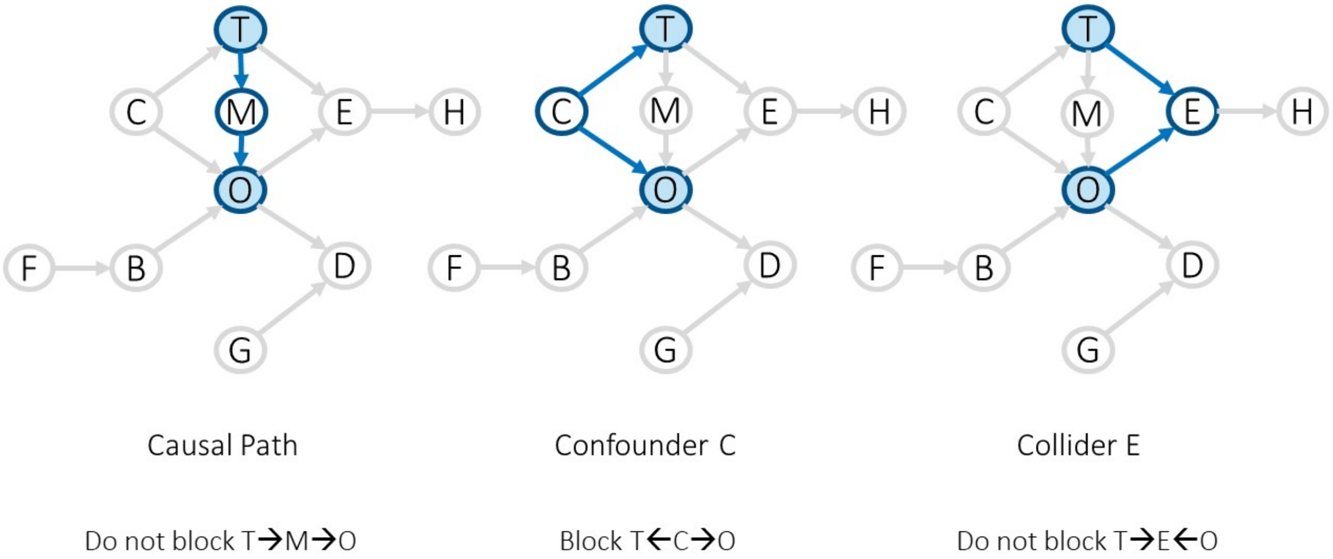


Figure 2. A Directed Acyclic Graph (DAG)

A graphical model is a visualization of relationships between variables. Variables are represented by nodes and relationships by connecting edges or arcs. Depending on the type of model, connectors can be single- or double-headed arrows. Connectors can begin and end at the same node. The absence of a connector between two nodes means that the variables are independent of each other. Figure 2 shows a directed acyclic graph (DAG), a special kind of graphical model. Letters A – G represent nodes. The graph is directed because each arc between two nodes is uniquely directed and is acyclic because no cycles or loops exist (e.g. A→B→C→A). A node from which a directed edge starts is called the parent of the node to which the edge is directed; a node on which a directed edge ends is called the child of the node from which it originates. For example, nodes B and C are parents of node A, and node A is the common child of nodes B and C. With respect to node A, nodes B, C, and F are ancestors, and nodes D, E, and H are descendants. Parents and children of a node are the neighbors of the node. Spouses are nodes sharing the same child, such as A and G.



Causal Path

Confounder C

Collider E

Do not block T→M→O

Block T←C→O

Do not block T→E←O

Figure 3. Graph Constructs and Blocking

In Figure 3, two nodes, T and M, have been added to our DAG example. We are interested in estimating the causal effect of treatment T on outcome O.

There are three basic constructs of nodes (or sets of nodes) in a DAG. 1. A causal path is a directed path from cause to outcome. The path might have intermediate nodes called mediators, like node M in Figure 3. 2. A confounding node, like node C, is a parent of both the cause of interest and the outcome. 3. A collider node, like node E, is a child of both the cause of interest and the outcome. This construct is called “endogenous selection”.

The cause of interest, the treatment, and the outcome are associated in the causal and confounding constructs. They are not associated via a collider. To isolate the causal effect of interest, we need to disassociate the treatment and outcome in all confounding paths. At the same time, we must not disassociate the treatment and outcome in the causal path nor introduce a confounding association via a collider.

To disassociate two nodes that have a natural association via a confounding path, we must “block” the path.
1. We do not want to block the T->M->O path because this is the cause of interest. If we mistakenly do block this path, the resulting bias in our analysis is called “overcontrol bias”. 2. We do want to block the T->C->O path to disassociate O from T via C. Doing so will isolate the causal path of interest. If we mistakenly do not block a confounding path, then our analysis will suffer “confounding bias”. 3. We do not want to block the T->E->O path because there is no association between T and O in this path. If we block this path, then we will create a confounding association and introduce “endogenous selection bias”.

Figure 4. Colliders and Endogenous Selection Bias

A simple example might help explain endogenous selection bias. Consider an outcome of risk status for Coronary Heart Disease (CHD). One of the measured variables is quality of life (QOL), quantified using a questionnaire. The outcome of interest, CHD, causes below normal QOL. A CHD treatment under study is expected to lower the risk of CHD and increase QOL, perhaps by returning patients to normal levels of daily activities. Figure 4 is the graphical model, where T is the treatment node, D is the disease node and Q is the QOL node. The objective of the study is to estimate the direct effect of treatment on CHD risk. Focusing on QOL, if you do not know the QOL level, then you have no information regarding the treatment status or CHD risk. However, if you know that QOL is very low, then you know that it is more likely that the patient is not receiving the treatment and that CHD risk is high. By setting or controlling the level of QOL, i.e. by blocking the Q node, you will induce a relationship between treatment and disease that did not naturally exist.

“Blocking” is a graphical term, and “adjusting, controlling, and conditioning” are statistical terms. In the next section we link graphs to analyses and state that blocking in graphs is accomplished through adjustment in modeling.

D-Separation

Graphical models help us visualize data generating processes and variable dependencies. But the graphical models are not analyses that estimate causal effects; we need statistical analysis methods. They are connected through the concept of d-separation. To make the transition from graphical models to inferential statistical methods, it is helpful to be aware of d-separation.

The “D” in d-separation stands for “directional”. Two nodes are d-separated when there is no connecting path between them. If a connecting path does exist, then they are d-connected.

In graphical terms: * Two nodes are d-separated when every path between them is blocked * Two nodes are d-connected if at least one path between them is unblocked

The transition to statistical methods is (Pearl et al., 2016, p46): * If two nodes are d-separated, then the variables they represent are definitely independent * If two nodes are d-connected, then the variables they represent are possibly or likely dependent

We block paths via conditioning on or adjusting for variables whose nodes are in confounding paths, or by the presence of colliders.

PROC CAUSALGRAPH tells us which nodes to condition on by identifying valid adjustment sets. It also indicates which nodes to never include in an adjustment set.

$CAUTION$ Causal graphs are comprised of nodes that represent variables. The graphs do not consider the shape of relationships between variables or distributions of those variables. When transitioning to causal analyses, it is important to keep this in mind if you use parametric methods, e.g. regression. Ideally, we would use non-parametric methods. But this is not always possible.

SASHELP.HEART

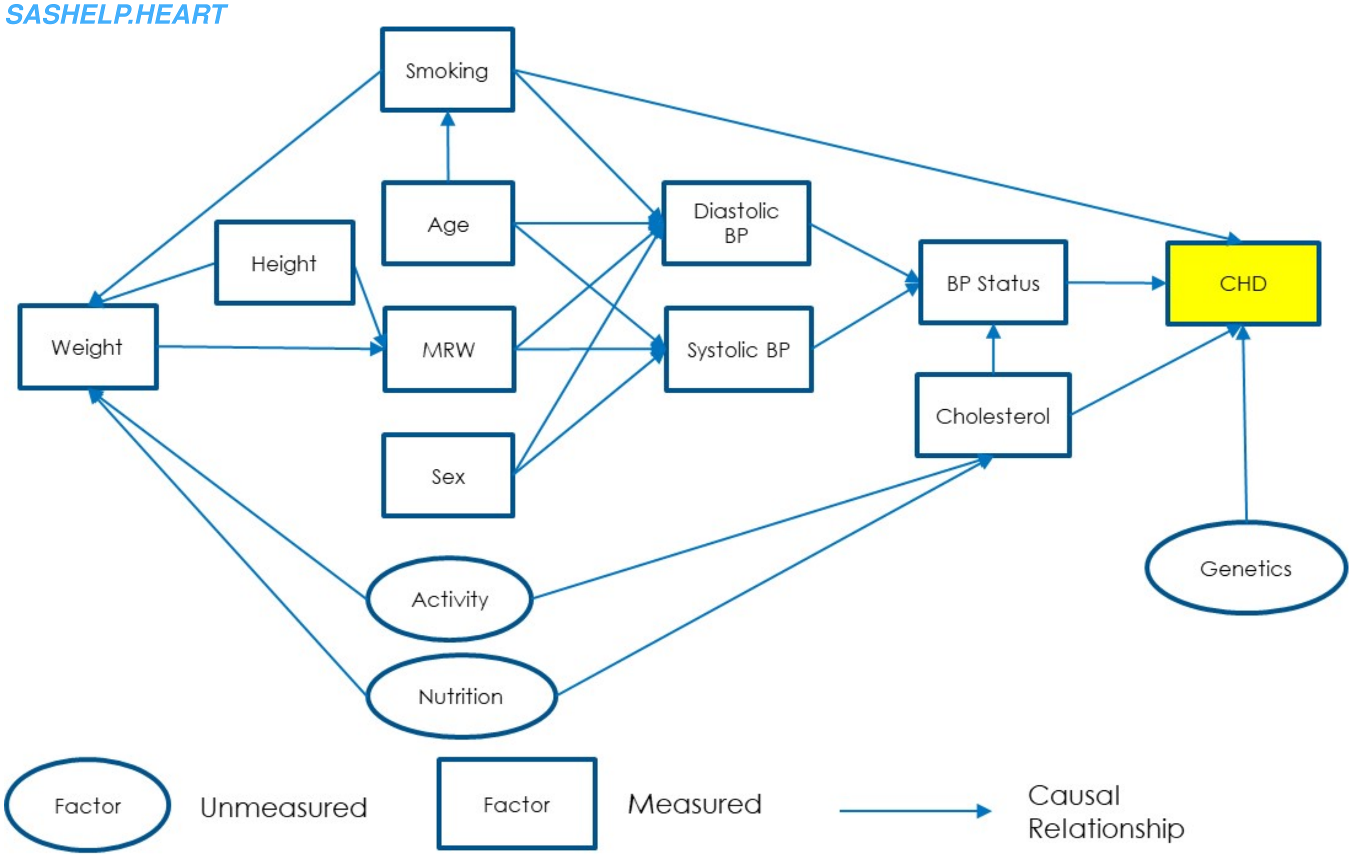


Figure 5. DAG of Coronary Heart Disease

To demonstrate causal analyses, we use the SASHELP.HEART data set. This data set contains 5209 adults who participated in the Framingham heart study (https://www.framinghamheartstudy.org/fhs-about/). SASHELP.HEART was used by Fechtner in a 2018 PhUse EU paper to discuss propensity score matching for causal inferences using the PSMATCH Procedure. You can compare the propensity score method described in her paper to the analyses we do here.

The Framingham heart study was initiated in 1948 in Framingham, Massachusetts to study risk factors associated with coronary heart disease (CHD). It is ongoing. The HEART data set in your SASHELP library is a subset of the full data set and includes 17 variables.

Suppose that the graph in this slide represents our belief about the relationships between the variables. We might be interested in estimating a causal relationship between assumed risk factors and Coronary Heart Disease (CHD), such as: * Weight * Metropolitan Relative Weight (MRW) which is an alternative to Body Mass Index (BMI) * Cholesterol * Smoking * Age

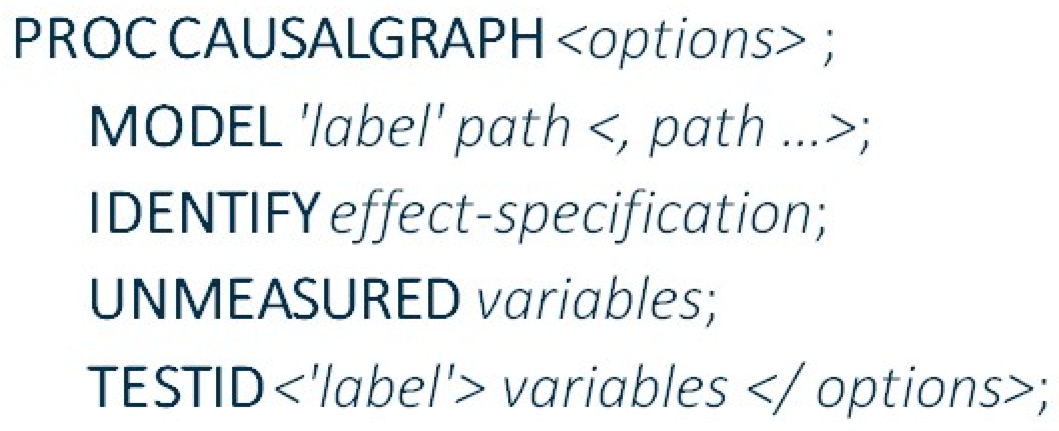
For each of these, we can use PROC CAUSALGRAPH to tell us which covariates to include in an adjustment set.

Variable	Label	Description
AgeAtDeath	Age at Death	Not needed for this analyses
AgeAtStart	Age at Start	Integer age in years when participant added to the study

Variable	Label	Description
AgeCHDdiag	Age CHD Diagnosed	Integer age in years when CHD first diagnosed
BP_Status	Blood Pressure Status	Normal, High, or Optimal
Chol_Status	Cholesterol Status	Borderline, Desirable, High
Cholesterol		Integer milligrams of cholesterol per deciliter of blood
DeathCause	Cause of Death	Not needed for this analysis
Diastolic		Amount of pressure in arteries between heart beats
Height		Height in inches to two decimal places
MRW	Metropolitan Relative Weight	Integer percent of weight compared to reference weight for height
Sex		Female or Male
Smoking		Integer 0, 1, or 5 to 60 by 5
Smoking_Status	Smoking Status	Non-smoker, Light (1-5), Moderate (6-15), Heavy (16-25), or Very Heavy (>25)
Status		Not needed for this analysis
Systolic		Maximum pressure heart exerts while beating
Weight		Integer weight in pounds
Weight_Status	Weight Status	Normal, Overweight, Underweight

Table 2. Variables in SASHELP.HEART

The CAUSALGRAPH Procedure



```
PROC CAUSALGRAPH <options> ;
    MODEL 'label' path <, path ...>;
    IDENTIFY effect-specification;
    UNMEASURED variables;
    TESTID <'label'> variables </ options>;
```

Figure 6. PROC CAUSALGRAPH Syntax

The CAUSALGRAPH Procedure in SAS/STAT® Software identifies nodes that must be, or can be, blocked to isolate a causal path of interest. This can be for planning an analysis of data already collected, or for planning a data collection effort for a designed experiment or study. If the data have not yet been collected, you can use PROC CAUSALGRAPH to learn which variables must be measured. When a causal path can be isolated, i.e. all confounding paths can be blocked, we say that the causal path is identified. An adjustment set includes the variables that you need to condition on, or adjust for, during your analysis to facilitate valid causal inferences.

Unlike many SAS Procedures, PROC CAUSALGRAPH does not have a DATA= option on the PROC CAUSALGRAPH statement; no data are needed. You specify your DAG on the MODEL statement and the causal path of interest on the IDENTIFY statement. If any of the nodes represent variables that are (or will be) unmeasured, you specify them on the UNMEASURED statement. Unmeasured nodes cannot be blocked, so it is important to specify such nodes.

See Thompson (2019) for an introduction to the CAUSALGRAPH Procedure.

IDentifying Adjustment Sets for the HEART data

Before we run PROC CAUSALGRAPH for our model of coronary heart disease, we modify SASHELP.HEART using the following DATA step to create a copy of the data set in the WORK library and to create the variables in Table 3. The frequency table indicates 197 participants were diagnosed with CHD within 5 years of registering for the study.

Variable	Label	Description
BP_status	Blood Pressure Status	Assigned Optimal group to Normal group
Chd5yr		= 1 if CHD diagnosis within age at start + 5 years, = 0 otherwise
Cholesterol		Divided by 10 for less granular scale

```
data heart;
  set sashelp.heart;
  cholesterol=cholesterol/10;
  chd5yr = (ageatstart le agechddiag le (ageatstart+5));
  if bp_status="Optimal" then bp_status="Normal";
  if smoking_status ne "Non-smoker" then smoking_status="Smoker";
run;

title "Number of Participants with CHD at 5 Years";
proc freq data=heart;
  tables chd5yr;
run;
```

We encode the CHD DAG in Figure 5 using the CAUSALGRAPH step below. Note that the node labels are not data set variables; there is no input data set. The causal path of interest is cholesterol –> CHD, which we specify on the IDENTIFY statement.

Review the “Variables in Model” table and the “Graphical Model Summary” table to ensure that you coded your DAG as you intended.

PROC CAUSALGRAPH will list all possible adjustment sets unless you override this default. This model has 92 possible adjustment sets that are listed in increasing size. You should take note of variables that do not appear in any adjustment set; e.g. the unmeasured variables are omitted.

Minimal adjustment sets are those for which no subset is also a valid adjustment set. Minimal does not mean the smallest adjustment set. The smallest adjustment set will be minimal, but a minimal set might not be the smallest.

```
title "Adjustment sets for HEART data";
proc causalgraph;
   model "Cholesterol->CHD"
      height ==> weight mrw,
      weight ==> mrw,
      smoking ==> weight diastolic systolic chd,
      age ==> smoking diastolic systolic,
      mrw sex ==> diastolic systolic,
      diastolic systolic ==> bp_status,
      nutrition activity ==> cholesterol weight,
      cholesterol ==> bp_status chd,
      bp_status genetics ==> chd;
   unmeasured nutrition activity genetics;
   identify cholesterol ==> chd;
run;
```

If you want to display only minimal adjustment sets, specify the MINIMAL option on the PROC CAUSALGRAPH statement.

If you want to display only adjustment sets below a specific size, use the MAXSIZE=n option on the PROC CAUSALGRAPH statement. Using the keyword MIN for n will cause PROC CAUSALGRAPH to display only the smallest adjustment set.

By default, PROC CAUSALGRAPH lists a maximum of 100 adjustment sets. You can override this default to raise or lower the limit by using the MAXLIST=n option on the PROC CAUSALGRAPH statement.

Use the TESTID statement to check if an adjustment set of interest is valid for causal analysis of the path specified on the IDENTIFY statement. Below we specify the minimal set plus bp_status on the TESTID statement to see if the set of three is a valid adjustment set.

The usual table of adjustments will be omitted when you specify the TESTID statement. Instead, you will produce a table only for the path(s) on the TESTID statement.

BP Status cannot be added to the adjustment set. Refer to the CHD DAG in Figure 5. Because BP Status is a mediator between cholesterol and CHD in the causal path of interest, we should not adjust for it. If we do, then our analysis will suffer from over control bias.

```
proc causalgraph;
   model "Cholesterol->CHD"
      height ==> weight mrw,
      weight ==> mrw,
      smoking ==> weight diastolic systolic chd,
      age ==> smoking diastolic systolic,
      mrw sex ==> diastolic systolic,
      diastolic systolic ==> bp_status,
      nutrition activity ==> cholesterol weight,
      cholesterol ==> bp_status chd,
      bp_status genetics ==> chd;
   unmeasured nutrition activity genetics;
   identify cholesterol ==> chd;
   testid mrw smoking bp_status;
run;
```

We put our PROC CAUSALGRAPH step inside a small macro loop so we could easily identify adjustment sets for our five causal factors of interest. * Weight * Metropolitan Relative Weight (MRW) * Cholesterol * Smoking * Age

Note that we specified the MINIMAL option on the PROC statement.

```
%macro ID(cause);
title "&CAUSE";
proc causalgraph minimal;
   model "&CAUSE"
```

```
      height ==> weight mrw,
      weight ==> mrw,
      smoking ==> weight diastolic systolic chd,
      age ==> smoking diastolic systolic,
      mrw sex ==> diastolic systolic,
      diastolic systolic ==> bp_status,
      nutrition activity ==> cholesterol weight,
      cholesterol ==> bp_status chd,
      bp_status genetics ==> chd;
   unmeasured nutrition activity genetics;
   identify &CAUSE ==> chd;
run;
%mend;
%ID(MRW)
%ID(WEIGHT)
%ID(CHOLESTEROL)
%ID(SMOKING)
%ID(AGE)
```

The results of the multiple executions of PROC CAUSALGRAPH are summarized in table 3. Note that no adjustment is need for the causal path AGE –> CHD.

Model	Size	Min	age	bp_status	cholesterol	MRW	diastolic	height	sex	smoking	systolic	weight
MRW	2	Yes			*					*		
MRW	2	Yes						*				*
Weight	3	Yes			*			*		*		
Cholesterol	2	Yes				*				*		
Cholesterol	3	Yes					*			*	*	
Cholesterol	3	Yes						*		*		*
Smoking	1	Yes	*									
Age	0	Yes										

Table 3. Minimal adjustment sets for 5 causal paths in the CHD study

The CAUSALMED Procedure

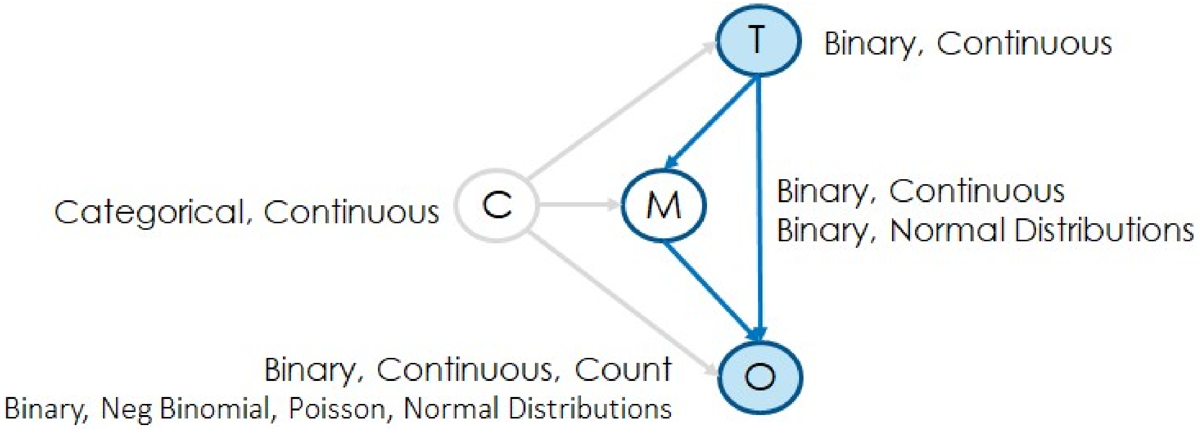
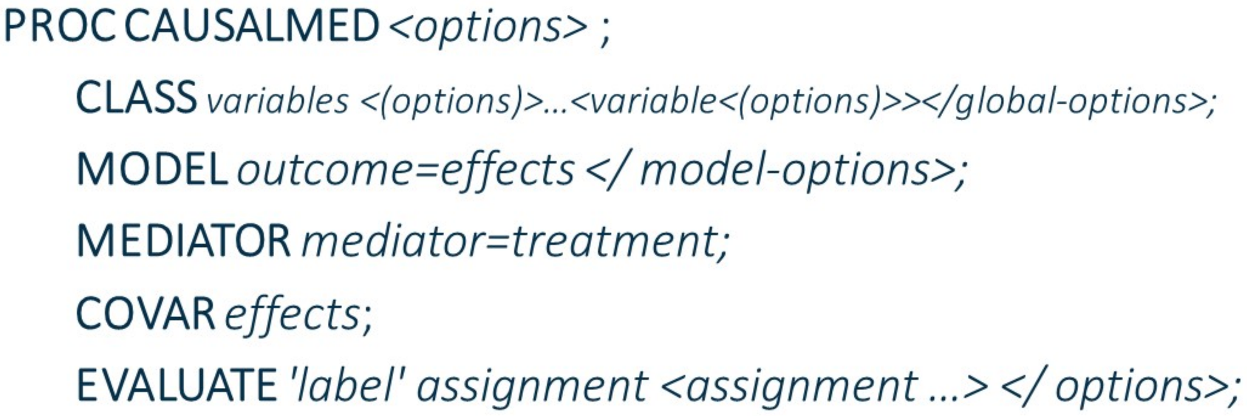

T
Binary, Continuous
Categorical, Continuous
C
M
Binary, Continuous
Binary, Normal Distributions
Binary, Continuous, Count
Binary, Neg Binomial, Poisson, Normal Distributions
O

Figure 7. PROC CAUSALMED Distributions and Measurement Scales

Now that we have identified which nodes to block, we can conduct an analysis that adjusts for the variables represented by those nodes. The CAUSALMED Procedure is in our toolbox of choices. PROC CAUSALMED conducts mediation analyses for causal interpretation. Figure 7 displays our simple graph model with treatment, mediator, outcome, and confounder nodes. An edge connecting treatment to outcome has been added to facicilitate separate estimates for direct and indirect (via the mediator) effects. CAUSALMED fits generalized linear models for the measurement scales listed in Figure 7.

The mediator and outcome nodes, which will be response variables in the fitted models, have distributional assumptions. CAUSALMED currently supports the distributions listed next to those nodes in Figure 7.



```
PROC CAUSALMED <options> ;
    CLASS variables <(options)>...<variable<(options)>></global-options>;
    MODEL outcome=effects </ model-options>;
    MEDIATOR mediator=treatment;
    COVAR effects;
    EVALUATE 'label' assignment <assignment ...> </ options>;
```

Figure 8. PROC CAUSALMED Syntax

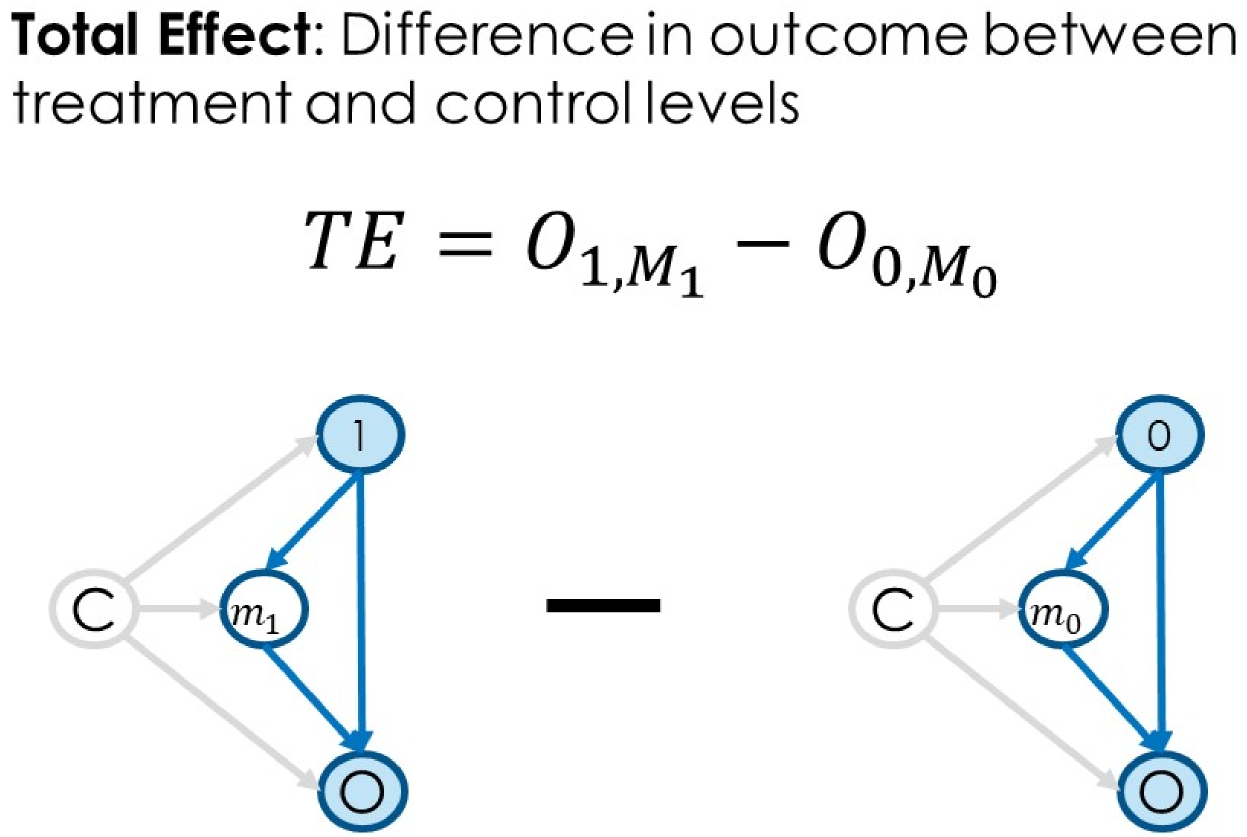
Figure 8 displays selected syntax for the CAUSALMED Procedure. You must specify one, and only one, for each of the treatment, mediator, and outcome variables. You may specify multiple confounders and their interactions. You specify categorical variables on the CLASS statement as you do in other procedures. On the MODEL statement, the outcome variable, CHD5YR, is specified to the left of the = sign and the treatment and mediator variables to the right. The MEDIATOR statement is, in effect, another modeling statement. Specify the mediator variable to the left of the = sign and the treatment variable to the right. The confounders that you list on the COVAR statement are used in both the treatment and mediator models. You can specify interactions using the vertical line and star notation as in other modeling procedures. For example, we could use MRW|SMOKING to include the main effects and two-way interaction in the models. Use the EVALUATE statement to specify levels of a covariate at which you want causal estimates. In the PROC CAUSALMED step below, we are estimating the causal effect of cholesterol on cardiac heart disease. From our DAG, we presume that blood pressure status has a mediating effect. From our indentification analysis using PROC CAUSALGRAPH, we know that we need to control for MRW and SMOKING. On the EVALUATE statement, we are requesting cholesterol effect estimates for someone who smokes 60 cigarettes per day. By default, PROC CAUSALMED will use the mean of the covariates. This is like the “AT variable=value” syntax on an LSMEANS statement in a modeling procedure.

See Yung et al. (2018) for an introduction to PROC CAUSALMED.

```
data heart;
  set sashelp.heart;
  cholesterol=cholesterol/10;
  chd5yr = (ageatstart le agechddiag le (ageatstart+5));
  if bp_status="Optimal" then bp_status="Normal";
  if smoking_status ne "Non-smoker" then smoking_status="Smoker";
run;
proc causalmed data=heart;
  title "Mediation Analysis: Cholesterol";
  class bp_status(ref='Normal') chd5yr(ref='0');
  model chd5yr=cholesterol bp_status;
  mediator bp_status=cholesterol;
  covar mrw smoking;
  evaluate "Smoking 3 packs" smoking=60;
run;
```

Review the Model Information table to ensure that you specified the model as you intended. Then review the Response Profile and Mediator Profile to ensure that counts for the different levels of the two binary variables are what you expect. For binary variables, we also get notes indicating the levels that the logistic regressions are modeling. We controlled the reference levels using the REF= option on the CLASS statement.

The Summary of Effects table presents a lot of information. To interpret these results, we must understand the quantities that are being estimated.

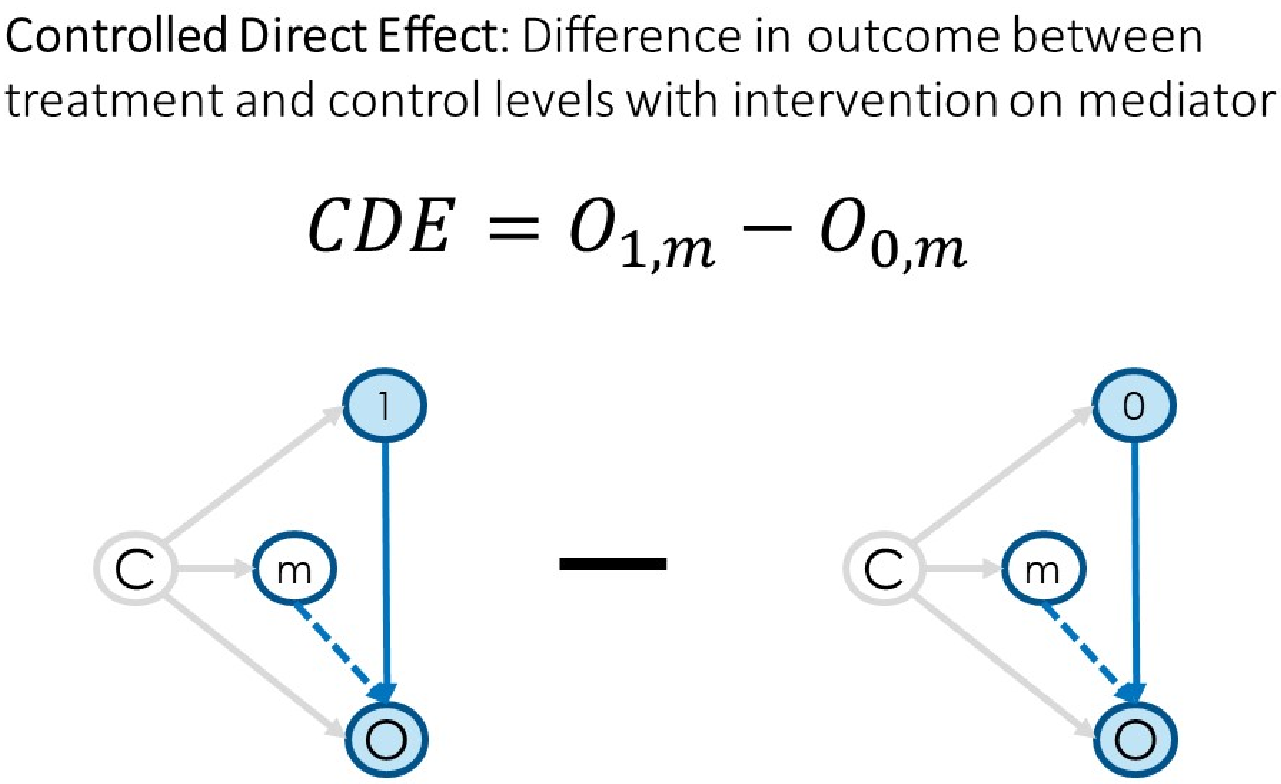


Total Effect: Difference in outcome between treatment and control levels

$$TE = O_{1,M_1} - O_{0,M_0}$$

Figure 9. Total Effect

$O_{t,m}$ is the value of the outcome at treatment t and mediator m. For simplicity, Figure 9 assumes that treatment and mediator are both binary; level 0 stands for a control group. Because treatment has a causal relationship with the mediator, we distinguish the value of the mediator when treatment is 0 versus 1. M_t is the value of the mediator when treatment is at level t. Total effect is the sum of direct (treatment to outcome) and indirect (mediator to outcome) effects, which we denote using solid lines in Figure 9. Note that we are not intervening on the mediator, rather it takes the value caused by the treatment; when under treatment, the mediator assumes the value M_1 and when under control, the mediator assumes value M_0.

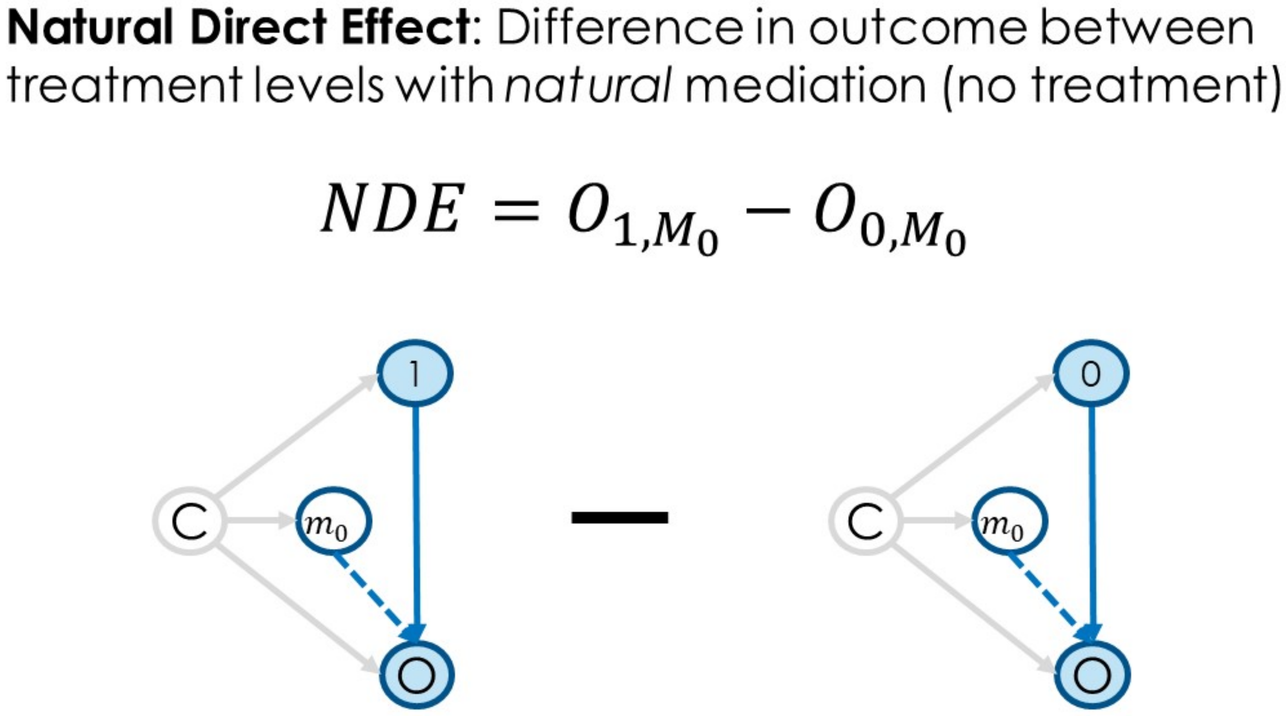


Controlled Direct Effect: Difference in outcome between treatment and control levels with intervention on mediator

$$CDE = O_{1,m} - O_{0,m}$$

Figure 10. Controlled Direct Effect

Direct effects can be controlled by intervening on the mediator; i.e. setting it to a value. By doing so, the indirect effect (mediator to outcome) is the same in both summands, and the difference is only the direct effect (treatment to outcome). In Figure 10, we denote this situation by removing the edges from treatment to mediator (setting the value of the mediator removes any effect of treatment on mediator), and by using dashed edges for the indirect effect (setting the value of the mediator forces the indirect effect to be equal under treatment and control.



Natural Direct Effect: Difference in outcome between treatment levels with *natural* mediation (no treatment)

$$NDE = O_{1,M_0} - O_{0,M_0}$$

Figure 11. Natural Direct Effect

The natural direct effect is the controlled direct effect when we set the mediator at the value it obtains under control. In Figure 11, we denote this situation by indicating the mediator value as M_0.

Natural Indirect Effect: Difference in outcome between treated with treatment and control mediation

$$NIE = O_{1,M_1} - O_{1,M_0}$$

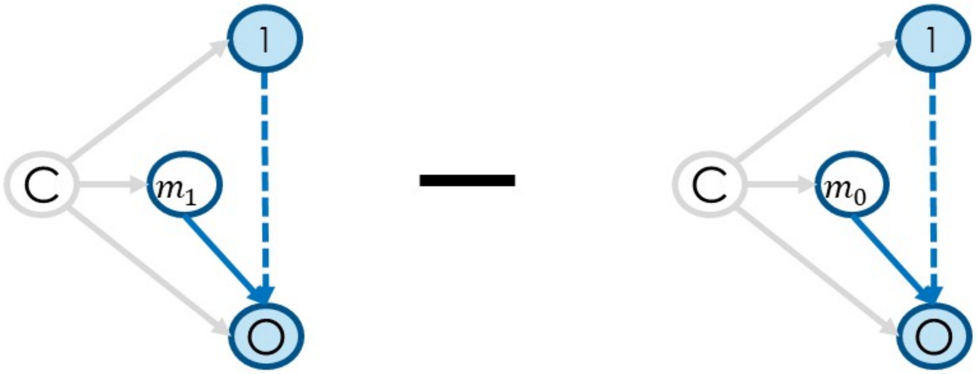


Figure 12. Natural Indirect Effect

The natural indirect effect is estimated by setting the treatment to the treated value and the mediator to different values; the values it obtains under treatment and under control such that the control is subtracted from the treated. In Figure 12 we denote this sitation by omitting edges from treatment to mediator (because we are specifying the mediator values), using dashed edges from treatment to outcome (because we set treatment to 1 in both summands), and solid edges from mediator to outcome (because we allow the mediator to assume its estimated effect under treatment and control. The direct effect is the same in both summands, so the difference is equal to the indirect effect only.

The total effect is the sum of the natural direct and natural indirect effects.

Returning to the output from our initial run of PROC CAUSALMED, the Summary of Effects table displays estimates for the total, controlled direct, and natural direct and indirect effects. Because CHD5YR is a binary outcome, PROC CAUSALMED fit a logistic regression model and presents both odds ratios and relative risks. Because these data were generated prospectively and have not been over-sampled, we can use the relative risk estimates.

In the CHD example, our treatment, cholesterol, is continuous. Recall that we divided the raw values by 10. Therefore, our effect estimates correspond to a 10 unit increase in cholesterol. A 10 unit increase in cholesterol directly causes an increase of 6.87% in the risk of coronary heart disease within 5 years. It indirectly causes an additional increase of 1.39% via blood pressure status. The total effect is 8.26% increase in the risk of CHD within 5 years. The percent of the total effect due to mediation is 16.88% and is also provided in the output. These estimates were produced at the mean values of the covariates.

Recall that, using the EVALUATE statement, we requested additional estimates for the subpopulation of those smoking 3 packs per day (at mean MRW). The relative risk estimates did not change, so smoking more cigarettes does not impact the effect of cholesterol on CHD. This does not mean that smoking does not affect CHD!

To estimate the affect of smoking on coronary heart disease we can run the PROC CAUSALMED step below. We again refer to our PROC CAUSALGRAPH output to ensure we use the correct adjustment set to analyze smoking. Age is the only covariate that we must control. Smoking appears to have a small but significant direct effect, but no indirect effect via blood pressure status.

```
data heart;
  set sashelp.heart;
  cholesterol=cholesterol/10;
  chd5yr = (ageatstart le agechddiag le (ageatstart+5));
  if bp_status="Optimal" then bp_status="Normal";
  if smoking_status ne "Non-smoker" then smoking_status="Smoker";
run;
proc causalmed data=heart;
  title "Mediation Analysis: Smoking";
  class bp_status(ref='Normal') chd5yr(ref='0');
  model chd5yr=smoking bp_status;
  mediator bp_status=smoking;
  covar ageatstart;
run;
```

The CAUSALTRT Procedure

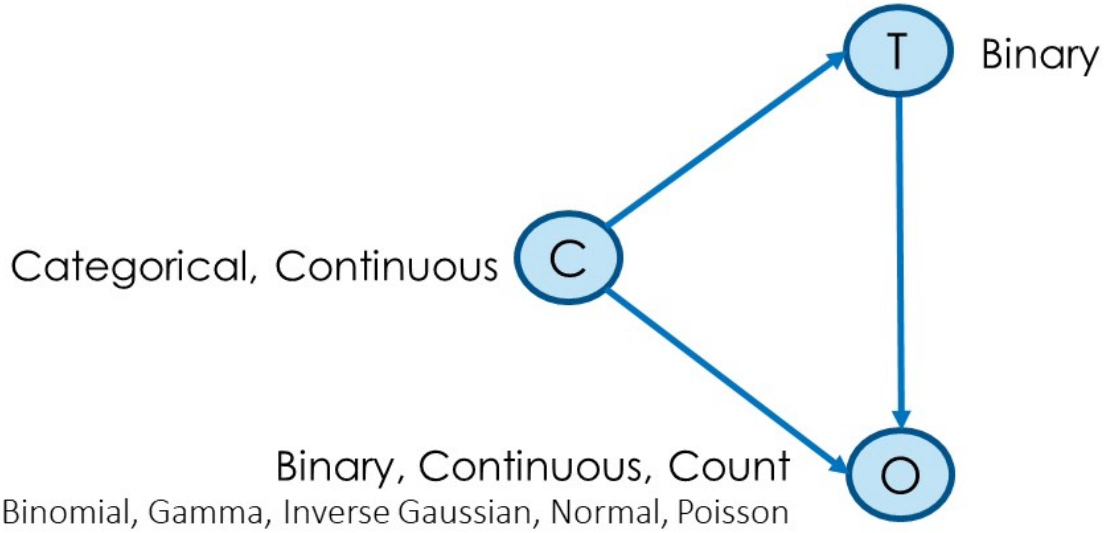

T
Binary
Categorical, Continuous
C
Binary, Continuous, Count
Binomial, Gamma, Inverse Gaussian, Normal, Poisson
O

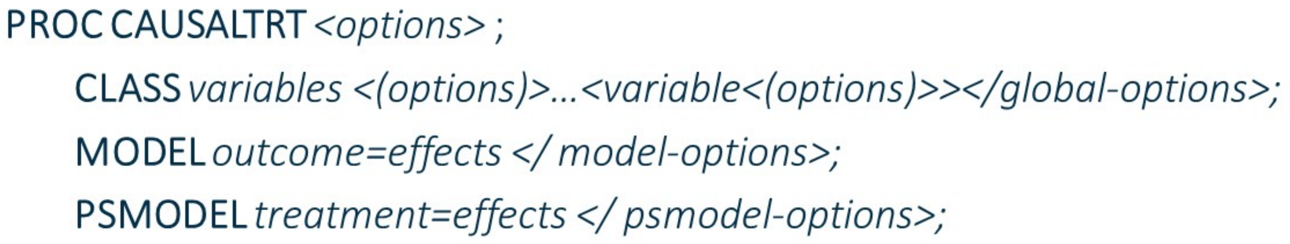
Figure 13. PROC CAUSALTRT Distributions and Measurement Scales

Another tool in our causal toolbox is the CAUSALTRT Procedure. PROC CAUSALTRT estimates the Average Treatment Effect (ATE) or Average Treatment Effect of the Treated (ATT) for analyses in which the treatment is binary.

PROC CAUSALTRT fits one or both of two models: * A model of the treatment assignment * A model of the outcome

PROC CAUSALTRT can fit both models, either each alone or together for robust estimation of causal effects.

The model of the treatment assignment estimates propensity scores. Recall that propensity scores are probabilities of being assigned to the treatment group, t = 1; Pr(T=1|X=x). They are estimated, usually by logistic regression, by specifying the binary treatment indicator as the response and covariates as predictors. We also learned that we can use the propensity scores to compute weights for a weighted analysis.



```
PROC CAUSALTRT <options> ;
    CLASS variables <(options)>...<variable<(options)>></global-options>;
    MODEL outcome=effects </ model-options>;
    PSMODEL treatment=effects </ psmodel-options>;
```

Figure 14. PROC CAUSALTRT Syntax

Figure 14 shows the syntax for the statements used below. A comprehensive discussion of all the capabilities of PROC CAUSALTRT is beyond the scope of this paper. The reader is encouraged to review the CAUSALTRT documentation and examples for an awareness of all features. See Lamm and Yung (2017) for an introduction to PROC CAUSALTRT.

You specify categorical MODEL and PSMODEL inputs on the CLASS statement. Options, including REF=, are available as in other procedures.

Specify the model for the outcome using the MODEL statement. If the outcome is binary, you can use the REF= option with the outcome variable. Model options include DIST= and LINK= to control the generalized linear model that you fit.

Specify the model for the treatment on the PSMODEL statement. "PS" stands for propensity score. You are specifying the model to estimate the probability of someone receiving the treatment. This should include all pre-treatment (baseline) factors that you believe affect an individual's selection or exposure of the treatment of interest.

How do you know which covariates to include in the models? You need to satisfy the assumption that treatment assignment and outcome are independent. Conditional independence is satisfied if you adjust for all confounders; factors that impact both the treatment and outcome. The adjustment set is identified by PROC CAUSALGRAPH. If the graphical model that we specified in PROC CAUSALGRAPH is an accurate representation of the data generating process, and we specify the adjustment set on both the MODEL and PSMODEL statements, then we satisfy the independence assumption.

You do not need to specify both models. However, you do need to identify both the treatment and outcome variables, which you can do without fitting the model by specifying either statement without input variables, i.e. no variables to the right of the equal sign (=).

Method	Description
AIPW	Doubly robust augmented IPW. Requires both MODEL and PSMODEL.
IPW	Inverse Probability Weighted. Requires PSMODEL.
IPWR	IPW with ratio adjustment. Requires PSMODEL.
IPWS	IPW with ratio and scale adjustment. Requires PSMODEL.
IPWREG	Doubly robust IPW regression adjustment. Requires MODEL and PSMODEL.
REGADJ	Regression adjustment. Requires MODEL.

Table 4. Estimation Methods available in PROC CAUSALTRT

Table 4 lists available estimation methods which you control using the METHOD= option on the PROC CAUSALTRT statement. If both models are specified, then AIPW is used by default. If only the outcome model is specified, then REGADJ is used by default. And if only the treatment model is specified, then IPWR is used. If neither model is specified, then REGADJ is used. AIPW and IPWREG are known as doubly robust because the final ATE combines estimates from both the propensity score and regression methods. These methods are called doubly robust because you obtain unbiased estimates of the ATE even if one of the models, treatment our outcome model, is mis-specified.

We have written the PROC CAUSALTRT step below to estimate the ATE for blood pressure status. Recall from our PROC CAUSALGRAPH results that cholesterol and smoking comprise an adjustment set for the causal path from blood pressure status to coronary heart disease. Therefore, we specify cholesterol and smoking as inputs in both the model for treatment assignment (PSMODEL statement) and for the outcome (MODEL statement). There is no procedural requirement that the models be the same. Because the procedure supports multiple distributions for the outcome, we specify the binomial distribution on the MODEL statement. The PALL option on the PROC CAUSALTRT statement requests all output tables, some of which are not displayed by default.

Review the model information to be sure your model was specified as intended. The Analysis of Causal Effect table is the main output table and is produced by default (even if PALL is not specified). The table shows the population mean outcome for each level of the binary treatment variable. The ATE is the difference.

It is important to keep in mind that the ATE is not derived from model parameter estimates; rather it is a difference of mean predictions.

```
data heart;
  set sashelp.heart;
  cholesterol=cholesterol/10;
  chd5yr = (ageatstart le agechddiag le (ageatstart+5));
  if bp_status="Optimal" then bp_status="Normal";
  if smoking_status ne "Non-smoker" then smoking_status="Smoker";
run;

proc causaltrt data=heart pall;
  title "ATE Analysis: BP Status";
  psmodel bp_status(ref='Normal') = cholesterol smoking;
  model chd5yr(ref='0') = cholesterol smoking / dist=bin;
run;
```

CONCLUSION

This paper is a gentle introduction to causality and structural causal models. The first step in a causal analysis using structural causal models is to draw your system of variables in a directed acyclic graph. Then code your DAG in PROC CAUSALGRAPH to identify adjustment sets of confounders that you must control in your causal analysis. With an adjustment set identified, if you can satisfy the four assumptions and if you practice good modeling principles, then you can safely make causal inferences.

We introduced to modeling procedures, PROC CAUSALMED, for mediation analyses, and PROC CAUSALTRT, for estimates of average treatment effect. PROC CAUSALTRT provides two modeling approaches; modeling the treatment by fitting a propensity score model, and modeling the outcome by fitting a structural causal model.

REFERENCES

Amrhein, J. and Wang F. (2018). “Bayesian Concepts: An Introduction.” Paper 1863-2018. In Proceedings of the SAS Global Forum 2018 Conference. Cary, NC: SAS Institute Inc.

Fechtner, S. (2018). “The Propensity Score Matching.” Paper RW03. In Proceedings of the PhUSE EU Connect 2018 Conference.

Greenland, S. and Robins, J. (2009). Identifiability, Exchangeability, and Confounding Revisited. Epidemiologic Perspectives & Innovations, 6:4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2745408/

Lamm, M. and Yung, Y-F. (2017). “Estimating Causal Effects from Observational Data with the CAUSALTRT Procedure” Paper SAS374-2017. In Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc.

Lamm, M., Thompson, C., and Yung, Y-F. (2019). “Building a Propensity Score Model with SAS/STAT® Software: Planning and Practice.” Paper 3056-2019. In Proceedings of the SAS Global Forum 2019 Conference. Cary, NC: SAS Institute Inc.

Madhanagopal, B. and Amrhein, J. (2019). “Analyzing Structural Causal Models Using the CALIS Procedure.” Paper 3765-2019. In Proceedings of the SAS Global Forum 2019 Conference. Cary, NC: SAS Institute Inc.

Pearl, J., Glymour, M., and Jewell, N. (2016). Causal Inference in Statistics: A Primer. John Wiley & Sons Ltd.

Schafer, J. and Kang, J. (2008). Average Causal Effects from Nonrandomized Studies: A Practical Guide and Simulated Example. Psychological Methods, Vol. 13, No. 4, 279-313.

Thompson, C. (2019). “Causal Graph Analysis with the CAUSALGRAPH Procedure.” Paper SAS2998-2019. In Proceedings of the SAS Global Forum 2019 Conference. Cary, NC: SAS Institute Inc.

Yung, Y-F., Lamm, M., and Zhang, W. (2018). “Causal Mediation Analysis with the CAUSALMED Procedures.” Paper SAS1991-2018. In Proceedings of the SAS Global Forum 2018 Conference. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

The author sincerely thanks Clay THompson of SAS Institute Inc. for his suggestions to improve an earlier version of this paper.

RECOMMENDED READING

Pearl, J. and Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books. New York, NY.

CONTACT INFORMATION

Your comment and questions are valued and encouraged. Contact the author at:

John Amrhein

Vice President, Managing Director

McDougall Scientific Ltd.

jamrhein@mcdougallscientific.com

www.mcdougallscientific.com