SAS/STAT® 15.1 User's Guide

The CAUSALGRAPH Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

Chapter 34
The CAUSALGRAPH Procedure

Contents

Overview: CAUSALGRAPH Procedure

The CAUSALGRAPH procedure examines the structure of graphical causal models and suggests statistical strategies or steps that enable researchers to estimate causal effects that have valid causal interpretations. Causal models are encoded in the form of directed acyclic graphs (Pearl 2009a, b), which are the primary input for the procedure. Henceforth, the input causal graph or diagram for the procedure is assumed to be a directed acyclic graph.

A causal graph depicts the causal relationships among variables in the context of a specific data generation process, which includes observational and experimental situations. Either the data have been collected previously or they will be collected in the future. The focus of a causal graph is usually on the relationship between the treatment variables and the outcome variables. Specifically, one of the primary goals of causal graph analysis is to determine how to estimate the causal effects of the treatment variables on the outcome variables.

Although the causal effect that is defined in PROC CAUSALGRAPH is called a treatment effect, it is not confined to effects that result from controlled treatments (such as effects in an experiment). Rather, the treatment might represent an intervention (such as smoking cessation versus control), an exposure to a condition (such as an infectious agent), or an existing characteristic of subjects (such as high versus low socioeconomic status). For example, the analysis in the section "Getting Started: CAUSALGRAPH Procedure" on page 2246 considers the effect of maternal exposure to persistent perfluoroalkyl substances on the duration of breastfeeding, and the analysis in Example 34.3 considers the link between an individual's measured serum urate and risk of cardiovascular disease.

Given the input causal graph and the treatment and outcome variables of interest, you can use the CAUSALGRAPH procedure to explore some formal properties of causal graphs, including the following:

- the causal and noncausal (associative) paths between the treatment and outcome variables
- the adjustment sets of variables that can be used to remove or block the spurious or confounding associations between the treatment and outcome variables
- the set of variables that can be used as instruments to estimate the causal treatment effect of interest

Essentially, because of the possible presence of spurious associations and unmeasured confounding, you cannot determine the causal effect of the treatment directly. Identifying and removing spurious association are especially important in observational studies where the treatment conditions are not randomly assigned to individuals. PROC CAUSALGRAPH uses the formal properties of a causal graph to suggest statistical strategies that can lead to unbiased estimation of causal treatment effects. For example, the list of adjustment covariates that the procedure produces can be used as input for an appropriate statistical procedure, such as PROC CAUSALTRT or PROC PSMATCH, to estimate the magnitude of a causal effect. For more information, see Chapter 36, "The CAUSALTRT Procedure," or Chapter 98, "The PSMATCH Procedure." The estimation process is also illustrated by the analysis in Example 34.7.

The CAUSALGRAPH procedure suggests statistical strategies that are based on the assumptions of a causal model that specifies the relationships between variables of interest (Elwert 2013). This causal model supplements the available data and cannot be discerned from those data (Pearl 2009b). Therefore, specifying a causal graph that accurately reflects the data generating process is essential in causal graph analysis. Domain-specific knowledge and familiarity with the data collection process can help you produce reasonably accurate causal graphs.

Using PROC CAUSALGRAPH to examine the identifiability of causal effects in causal graphs requires an understanding of the terminology, concepts, and assumptions of graphical causal models. For more information about these technical aspects, see the section "Details: CAUSALGRAPH Procedure" on page 2260.

Features of the CAUSALGRAPH Procedure

PROC CAUSALGRAPH provides several criteria for identifying causal treatment effects. In particular, you can use the METHOD= option in the PROC CAUSALGRAPH statement to specify any one of the following identification criteria:

- constructive backdoor criterion (Van der Zander, Liśkiewicz, and Textor 2014)
- backdoor criterion (Pearl 2009b)
- instrumental variables (Van der Zander, Textor, and Liśkiewicz 2015)

The constructive backdoor criterion (METHOD=ADJUSTMENT), also called the adjustment criterion, finds all valid adjustment sets that consist of observed variables only. The backdoor criterion (METHOD=BACKDOOR) similarly finds adjustment sets that consist of observed variables, but with a slightly stronger criterion. The backdoor criterion is computationally more efficient than the adjustment criterion, but it might not find every possible valid adjustment set. The appeal of the backdoor criterion is that it has an intuitive interpretation and provides a fast method of constructing valid adjustment sets (Elwert 2013). The instrumental variable method (METHOD=IV) finds instrumental variables to deal with the presence of unmeasured confounding between the treatment and outcome variables. This is useful because unmeasured confounding is a situation in which the adjustment and backdoor criteria might fail.

To identify the sets of adjustment covariates or instrumental variables, PROC CAUSALGRAPH has two primary modes of operation:

- The LIST option in the PROC CAUSALGRAPH statement enables you to enumerate the criteria under which it is possible to estimate a causal effect.
- The TESTID statement specification enables you to test whether a user-specified criterion is valid for estimating a causal effect.

You can use both of these modes in a single run of the procedure. Various options are available to fine-tune the output listing of the requested criteria. You can use these options to limit the number of criteria that are listed, sort the listed criteria, improve the searching and listing efficiency, and so on.

In the CAUSALGRAPH procedure, every causal model must be a directed acyclic graph (DAG). You can input causal graphs or models by using the MODEL statement. The MODEL statement supports a pathlike syntax to input causal relationships among variables. For example, to specify the causal path $X \rightarrow Y$, you can use either the `X ==> Y` or `Y <== X` syntax in the MODEL statement. You can also specify multiple causal relationships as a chain of causal paths: for example, `X ==> Y ==> Z`, `Z <== X ==> Y <== W`, and so on. Each edge in a causal path represents a direct causal effect of one variable on another variable. For more information about the causal interpretation of directed graphs, see the section “Causal Graph Theory” on page 2261.

PROC CAUSALGRAPH performs the following semantic validation checks for every model that you specify:

- The model should be weakly connected. That is, there should be a path between any pair of variables when all edges are treated as undirected.
- The model cannot contain any directed cycles.

The procedure also supports the specification of bidirected edges (or paths). A bidirected edge syntax, such as `X <==> Y` (for $X \leftrightarrow Y$), is interpreted as unmeasured confounding between the two variables, so that the graph is still a DAG. That is, `X <==> Y` is equivalent to `X <== L ==> Y` (for $X \leftarrow L \rightarrow Y$), where the node L represents some unmeasured variable, which you specify in the UNMEASURED statement.

It is important to distinguish between measured and unmeasured variables in a graphical model analysis. Variables that you list in the UNMEASURED statement are treated as unmeasured or unobserved. All other variables are treated as measured or observed. In order to make causal effect estimation meaningful, your treatment and outcome variables must always be measured. All other variables can either be measured or unmeasured. Unmeasured variables in a causal model cannot be included in a statistical analysis, and thus you cannot use them in any identification criterion for causal treatment effects.

In practice, there are several reasons why you might want to specify a variable to be unmeasured:

- The variable corresponds to a latent construct that cannot be measured.
- The variable was not measured or will not be measured in the data set for the causal effect estimation.
- The data that are collected for the variable are not considered reliable enough to include in the causal effect estimation.

Although PROC CAUSALGRAPH distinguishes between measured and unmeasured variables, it does not distinguish between classification and continuous variables in an analysis.

You can specify multiple causal models in a single run of the procedure by using multiple MODEL statements. This enables you to compare the identification criteria that are produced for alternative models. Or, as in Example 34.4, you can search for identification criteria that are simultaneously valid for all specified models.

The CAUSALGRAPH procedure also includes the following important features:

- identification of joint causal effects from multiple treatments on multiple outcomes. For more information about the interpretation of joint effects, see the section "Identifying Joint Treatment Effects" on page 2267.
- enumeration of all the observationally testable assumptions that are encoded by a causal model
- listing of causal and noncausal treatment-to-outcome paths that are blocked or nonblocked under specified adjustment sets of variables

Getting Started: CAUSALGRAPH Procedure

This example demonstrates how you can use the CAUSALGRAPH procedure to determine which covariates in a causal model you must control in order to estimate a treatment effect that has a valid causal interpretation.

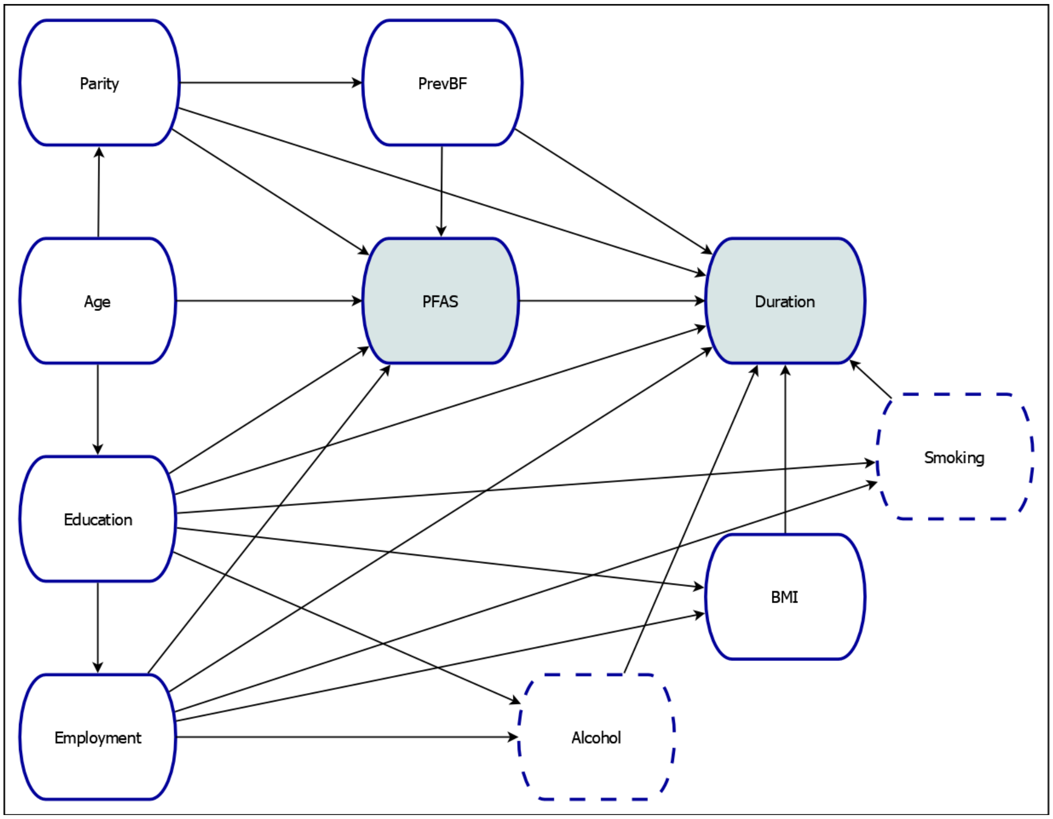
The causal model shown in Figure 34.1 has been adapted from Timmermann et al. (2017) and examines the relationship between maternal exposure to persistent perfluoroalkyl substances (PFAS) and breastfeeding duration (Duration) among residents of the Faroe Islands. The model includes the following variables:

- PFAS: the treatment variable
- Duration: the outcome variable
- Age: age of the mother at the child's birth

- Education: indicator of whether the mother had any postprimary education
- Employment: a categorical variable that describes the employment condition of the mother (employed, unemployed, homemaker, and so on)
- Parity: indicator of whether this was the mother's first childbirth
- Alcohol: indicator of whether the mother consumed alcohol during the pregnancy
- Smoking: indicator of whether the mother smoked cigarettes during the pregnancy
- BMI: prepregnancy body mass index of the mother
- PrevBF: indicator of prior breastfeeding experience

The treatment (PFAS) and outcome (Duration) variables are shaded in Figure 34.1. For this example, it is assumed that the variables Alcohol and Smoking are not observed (for example, because the data are considered to be unreliable).

Figure 34.1 Causal Model of the Effect of Persistent Perfluoroalkyl Substances on Breastfeeding Duration



The statistical association between the variables PFAS and Duration that would be measured in an observational study reflects a combination of true causal association and additional spurious or noncausal association. In order to isolate the true causal association between PFAS and Duration, you must devise a strategy to eliminate the noncausal association. One way to do this is to find an adjustment set. You can use the CAUSALGRAPH procedure to construct all possible adjustment sets that can be used to identify the causal effect of PFAS on Duration, subject to the assumptions that are encoded in the causal model in Figure 34.1. If at least one such adjustment set exists, then it is possible to estimate the causal effect by using observational data. For more information about adjustment sets and identifying causal effects, see the section "Identification and Adjustment" on page 2265.

The following statements invoke PROC CAUSALGRAPH to define and analyze the causal model and construct the adjustment sets:

```
proc causalgraph;
   model "Timm17TwoLatent"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration Employment PFAS BMI Alcohol Smoking,
      Employment ==> Duration PFAS BMI Alcohol Smoking,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
   unmeasured Alcohol Smoking;
run;
```

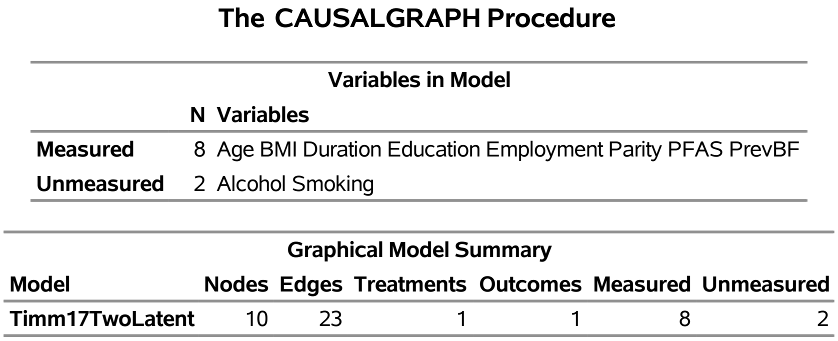
In an analysis that uses PROC CAUSALGRAPH, you must specify at least one causal model in a MODEL statement. You can also specify multiple models. Each MODEL statement must begin with a quoted string that provides a unique name for the model. This example labels the model as **Timm17TwoLatent**, a reference to its original publication. The remainder of the MODEL statement specifies the variables and their causal relationships (as indicated by directed edges). In this example, the MODEL statement encodes the model shown in Figure 34.1.

In the IDENTIFY statement, you specify the causal effect of interest. You can use this statement to specify one or more treatment variables and one or more outcome variables. The treatment and outcome variables are separated by a single right arrow, **==>**. This example studies the causal effect of the variable PFAS on the variable Duration.

The UNMEASURED statement specifies variables that are not observed and thus cannot be included in any adjustment set. In this example, the variables Alcohol and Smoking are treated as unmeasured.

The output in Figure 34.2 summarizes the variables and edges in the causal model that is specified in the MODEL statement. You can use this information as a qualitative check of the model specification.

Figure 34.2 Input Summary Tables for the Causal Model in Figure 34.1



```
                    The CAUSALGRAPH Procedure

                         Variables in Model

                 N  Variables
    Measured     8  Age BMI Duration Education Employment Parity PFAS PrevBF
    Unmeasured   2  Alcohol Smoking

                          Graphical Model Summary
 Model            Nodes  Edges  Treatments  Outcomes  Measured  Unmeasured
 Timm17TwoLatent     10     23           1         1         8           2
```

In this example, the CAUSALGRAPH procedure uses the constructive backdoor criterion (METHOD=ADJUSTMENT; see Van der Zander, Liśkiewicz, and Textor 2014) to construct all valid adjustments. You can change the default criterion by specifying the METHOD= option in the PROC CAUSALGRAPH statement.

The adjustment sets are displayed in Figure 34.3. For the model in Figure 34.1, there are four valid adjustment sets. Each row of Figure 34.3 contains an adjustment set, and the variables in each set are indicated in the table by an asterisk. Assuming that the causal model is accurate, you can estimate the causal effect of PFAS on Duration by using any one of these adjustment sets.

Figure 34.3 Adjustment Sets for the Causal Model in Figure 34.1

Covariate Adjustment Sets for Timm17TwoLatent

Causal Effect of PFAS on Duration

			Covariates					
	Size	**Minimal**	**Age**	**BMI**	**Education**	**Employment**	**Parity**	**PrevBF**
1	4	Yes			*	*	*	*
2	5	No	*		*	*	*	*
3	5	No		*	*	*	*	*
4	6	No	*	*	*	*	*	*

The table also indicates the size of each set and whether or not the set is minimal. An adjustment set is minimal if no proper subset of the set is also a valid adjustment set. In this example, there is one minimal set that contains four covariates that you must adjust for in order to estimate the specified causal effect. You can use one of these covariate adjustment sets as input for an appropriate statistical procedure, such as PROC PSMATCH or PROC CAUSALTRT, to estimate the magnitude of the specified causal effect. For an illustration of how you can use an adjustment set to estimate a causal effect, see Example 34.7.

By default, PROC CAUSALGRAPH constructs every possible adjustment set for the specified causal effect. You can use the MAXLIST=, MAXSIZE=, and MINIMAL options in the PROC CAUSALGRAPH statement to refine the adjustment sets that are computed. You can modify the displayed output by using the LIST, NOLIST, NOPRINT, or PSUMMARY option in the PROC CAUSALGRAPH statement.

Syntax: CAUSALGRAPH Procedure

The following statements are available in the CAUSALGRAPH procedure:

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

You can specify only one UNMEASURED statement and only one IDENTIFY statement for each PROC CAUSALGRAPH statement. The following sections describe the PROC CAUSALGRAPH statement and then describe the other statements in alphabetical order.

PROC CAUSALGRAPH Statement

PROC CAUSALGRAPH < *options* > ;

The PROC CAUSALGRAPH statement invokes the procedure. The *options* listed in Table 34.1 are available in the PROC CAUSALGRAPH statement.

Table 34.1 Options Available in the PROC CAUSALGRAPH Statement

Option	Description
Analysis Options	
COMMON	Requests adjustment sets common to all models
IMAP	Requests conditional independence assumptions
LIST	Requests possible identification criteria
METHOD=	Specifies the method to use for identification
MINIMAL	Requests only minimal adjustment sets
NOLIST	Excludes the construction of possible identification criteria
Output Options	
COMPACT	Suppresses the display of unused variables
CYCLES	Displays directed cycles
DISCONNECTED	Suppresses warnings for models that contain disjoint DAGs
MAXLIST=	Specifies the maximum number of identification criteria to print
MAXSIZE=	Specifies the maximum size of constructed adjustment sets
NOPRINT	Suppresses display of all output
NOSORT	Requests an unsorted list of adjustment sets
NTHREADS=	Specifies the maximum number of processing threads
ORDERMODELS	Orders the model output displays according to the model labels
PSUMMARY	Displays only the identification analysis summary

COMMON < (ONLY) >

requests adjustment sets common to all models. If you specify the COMMON option, the adjustment sets common to all models are computed in addition to the adjustment sets specific to each model. You can specify the COMMON(ONLY) option to compute only the adjustment sets common to all models. The COMMON option is ignored if you specify METHOD=IV.

COMPACT

suppresses the display of unused variables in a table. By default, the procedure prints one column for every variable when adjustment sets or conditional instrumental variables are displayed. Specify this option if you want to print only the columns corresponding to variables that appear in at least one adjustment set or conditioning set.

CYCLES < =*n* | =ALL >

displays directed cycles in each model that you specify by using a MODEL statement. By default, the procedure issues an error if any model that you specify in a MODEL statement contains a directed cycle. If you specify this option, the procedure does not issue an error. This is useful if you need to investigate and revise a large causal model that contains directed cycles.

By default, the procedure displays only one directed cycle (if such a cycle exists) for each model. You can change the number of cycles that are printed for each model by specifying the optional argument. For a positive integer *n*, CYCLES=*n* prints at most *n* directed cycles. To print all cycles, specify CYCLES=ALL.

This option has no effect for a model that contains no directed cycles.

DISCONNECTED

DISCONNECT

suppresses warnings for models that contain disjoint directed acyclic graphs (DAGs). By default, the procedure issues a warning if a model that you specify in a MODEL statement is not weakly connected (that is, if the model consists of two or more disjoint DAGs). Although the procedure can still continue to analyze disconnected models, this warning provides a safeguard against accidental misspecification. You can use this option to turn off the warning. This option does not have any effect for a model that is connected.

IMAP < =GLOBAL | LOCAL >

requests a list of the conditional independence properties (an independence map) that are encoded by each causal model. For more information about conditional independence properties in causal models, see the section "Statistical Properties of Causal Models" on page 2264.

GLOBAL produces a list of global Markov properties. For a causal model that is encoded in a DAG, the global Markov property corresponds to d-separation (Koller and Friedman 2009).

LOCAL produces a list of local Markov properties. For a causal model that is encoded in a DAG, the local Markov property is the set of nodes that are jointly independent of a given node, conditional on the parents of that node (Koller and Friedman 2009).

By default, IMAP=LOCAL.

When a conditional independence property includes an unmeasured variable, the implications of that property (for example, zero partial correlation) cannot be tested in a data set. Such properties are still included in the independence map as an implication of a causal model.

LIST

requests the possible identification criteria for each model. The type of the identification criteria that are constructed depends on the value that you specify in the METHOD= option. If METHOD=ADJUSTMENT or BACKDOOR, a list of adjustment sets is produced for each model. If METHOD=IV, a list of instrumental variables is produced for each model.

The LIST option is used by default if you specify an IDENTIFY statement but no TESTID statement. Otherwise, the NOLIST option is used.

You cannot specify both the LIST and NOLIST options.

MAXLIST=*n* | ALL

specifies the maximum number of identification criteria to be printed when you specify the LIST option. By default, the procedure prints up to 100 entries. To change the maximum number of entries to be printed, you can specify a positive integer *n*. To print all entries, specify MAXLIST=ALL.

MAXSIZE=*n* | MIN

specifies the maximum size of the adjustment sets to be constructed when you specify the LIST option. By default, the procedure prints all valid adjustment sets without regard to the size of the set. You can specify the MAXSIZE= option with a positive integer *n* to change the maximum size of the sets to be printed. To see only those adjustment sets with the smallest size, specify MAXSIZE=MIN.

The MAXSIZE=MIN option is not the same as the MINIMAL option. Every adjustment set that has the smallest size is minimal, but not every minimal adjustment set necessarily has the smallest size.

The MAXSIZE= option is ignored if you specify METHOD=IV.

METHOD=ADJUSTMENT | BACKDOOR | IV

specifies the method to use for identifying a causal effect. For more information about identifying causal effects, see the section "Identification and Adjustment" on page 2265.

ADJUSTMENT specifies the constructive backdoor criterion (Van der Zander, Liśkiewicz, and Textor 2014). The constructive backdoor criterion is equivalent to the adjustment criterion (Shpitser, VanderWeele, and Robins 2010). You can use this criterion to find all valid adjustment sets that consist of observed variables only. This method contains the backdoor criterion (METHOD=BACKDOOR) as a special case.

BACKDOOR specifies the backdoor adjustment criterion (Pearl 2009b). The backdoor adjustment criterion is similar to the constructive backdoor criterion (METHOD=ADJUSTMENT) in that it finds valid adjustment sets that consist of observed variables only. The backdoor criterion is stronger than the constructive backdoor criterion, so it is computationally more efficient than the constructive backdoor criterion, although it might not find every valid adjustment set. The appeal of the backdoor criterion is that it has an intuitive interpretation and is more widely known (Pearl 2009b).

This method is a special case of the constructive backdoor criterion. This means that every adjustment set that satisfies the backdoor adjustment criterion also satisfies the constructive backdoor criterion.

IV specifies the ancestral instrument criterion (Van der Zander, Textor, and Liśkiewicz 2015). This method finds all traditional and ancestral instruments for a causal model. If you use this criterion with the LIST option, the procedure first checks to see whether each observed variable can be used as a classical instrument (that is, the variable is a valid instrument without conditioning on any other variables). If not, then the procedure searches for a conditioning set that instrumentalizes the observed variable.

In order to use the ancestral instrument criterion, every model that you specify by using a MODEL statement must contain a single treatment variable and a single outcome variable that are directly connected by an edge.

By default, METHOD=ADJUSTMENT.

MINIMAL

MIN

requests a list of only minimal adjustment sets when you specify the LIST option. A minimal adjustment set is a set for which no proper subset is also a valid adjustment set. If you specify prescribed adjustment variables in the IDENTIFY statement, then the minimal condition is checked only among those sets that contain the prescribed variables.

The MINIMAL option is not the same as the MAXSIZE=MIN option. Every adjustment set that has the smallest size is minimal, but not every minimal adjustment set necessarily has the smallest size.

The MINIMAL option is ignored if you specify METHOD=IV.

NOLIST

excludes the construction of possible identification criteria. If you specify this option, the procedure does not construct any identification criteria for any model. This option is used by default if you specify a TESTID statement or if you do not specify an IDENTIFY statement.

You cannot specify both the LIST and NOLIST options.

NOPRINT

suppresses all displayed output.

NOSORT

requests an unsorted list of adjustment sets when METHOD=ADJUSTMENT or BACKDOOR. By default, the procedure prints adjustment sets in order of increasing set size. The NOSORT option enables more efficient output because the sets do not need to be sorted. This can be especially useful for large models when you have a larger number of adjustment sets. For more information about options that improve the performance of PROC CAUSALGRAPH, see the section "Time Requirements" on page 2269.

The NOSORT option is ignored if you specify METHOD=IV.

NTHREADS=*n*

specifies the maximum number of simultaneous computational threads available to the procedure. By default, the procedure uses the values of the THREADS and CPUCOUNT system variables to determine the number of computational threads. Multithreading is available when you specify more than one MODEL statement. There is no performance benefit in using a value of *n* greater than the number of models. To disable multithreading within the procedure, specify NTHREADS=1.

ORDERMODELS

prints the model results in alphabetical order by model label. However, the default behavior is to print the model results in the order in which the models appear in the input specifications.

PSUMMARY

displays only the summary of the causal effect identification analysis.

IDENTIFY Statement

IDENTIFY *effect-specification* **;**

The IDENTIFY statement specifies the causal effects to be analyzed. The *effect-specification* has the following form:

treatment-variables right-arrow outcome-variables < | *adjustment-variables* >

Each of the *treatment-variables*, *outcome-variables*, and *adjustment-variables* contains a list of variables in any of the following forms:

- *variables*
- {*variables*}
- [*variables*]
- (*variables*)

The use of braces, brackets, or parentheses for grouping variables is optional but highly recommended because it clearly distinguishes the roles of different types of variables within the *effect-specification*.

The *right-arrow* that indicates the causal direction can be of different forms. For information about specifying arrows in the CAUSALGRAPH procedure, see the section "Arrow or Edge Specification" on page 2267.

The following is an example for analyzing the effect of a single treatment variable on a single outcome variable without any prescribed adjustment variables:

```
identify x ==> y;
```

The following examples are equivalent for specifying a joint causal effect analysis that includes a prescribed set of adjustment variables:

```
identify {x1-x3} ==> {y1 y5} | {c1-c4 z};
identify {x1 x2 x3} ==> y1 y5 | {c1-c4 z};
identify x1 x2 x3 ==> y1 y5 | c1 c2 c3 c4 z;
```

You must use an IDENTIFY statement if you use a TESTID statement or if you use the LIST option in the PROC CAUSALGRAPH statement. You cannot specify more than one IDENTIFY statement.

Specifying Causal Effects

A causal effect consists of one or more treatment variables and one or more outcome variables. The treatment variables and the outcome variables are each specified as a list of one or more valid SAS variable names. The two variable lists are separated by a single right arrow. Every treatment variable and every outcome variable must be measured or observed. This means that you cannot include any treatment or outcome variable in the UNMEASURED statement.

If you specify more than one treatment variable or more than one outcome variable, PROC CAUSALGRAPH attempts to identify the joint effect. For more information about the interpretation of joint treatment effects, see the section "Identifying Joint Treatment Effects" on page 2267. However, you cannot specify multiple treatment or multiple outcome variables if you specify the METHOD=IV option in the PROC CAUSALGRAPH statement.

You cannot use the same variable as a treatment variable and an outcome variable. Every treatment variable and every outcome variable must appear in at least one MODEL statement, or else the effect is ill-defined. In that case, the missing variable is ignored. In addition, if you use an IDENTIFY statement, then every model must contain at least one treatment variable and at least one outcome variable.

Prescribing Adjustment Variables

If you want to prescribe a set of adjustment variables in the *effect-specification*, use a vertical bar (|) after the list of outcome variables and then specify the list of adjustment variables. This set of prescribed adjustment variables is included in every adjustment set that is tested or constructed by the CAUSALGRAPH procedure. The set of prescribed adjustment variables is ignored if you specify the METHOD=IV in the PROC CAUSALGRAPH statement.

An adjustment variable cannot also be a treatment variable or an outcome variable. Every adjustment variable must appear in at least one MODEL statement, or else it is not defined and is ignored. Every adjustment variable must be measured or observed. This means that you cannot include any prescribed adjustment variable in the UNMEASURED statement.

MODEL Statement

MODEL '*label*' *path* <, *path* . . . > **;**

where *label* represents a name that you assign to the model and *path* represents either of the following specifications:

- a *directed-path*
- a *covariance-path*

Details about the syntax and interpretation of these different types of *paths* are described later in this section. Here are some examples:

```
model 'Example1'
   X1 ==> X2,
   X3 <== X2,        /* same as: X2 ==> X3 */
   X3 ==> X4 X5,     /* same as: X3 ==> X4, X3 ==> X5 */
   <==> {X2 X5 X6}; /* latent confounding between X2, X5, and X6 */

model 'Example2'
   X1 ==> X2 ==> X3 <== X4 <==> X5;
```

You must specify at least one MODEL statement in an analysis.

The *label* is enclosed within quotation marks and can be any string of characters. Every model that you specify using a MODEL statement must have a unique *label*. The *labels* for models are not case-sensitive.

A MODEL statement specifies a causal model in the form of a directed acyclic graph (DAG). A DAG consists of nodes that represent variables in the model and edges that represent causal relationships between pairs of variables. For more information about how to use a DAG to represent a causal model, see the section "Causal Graph Theory" on page 2261. You specify the causal relationships in a DAG in accordance with the *path* syntax in the MODEL statement.

The following subsections explain the *path* syntax. Each *path* is either a *directed-path* or a *covariance-path*.

Directed-paths

A *directed-path* has the following form:

variables arrow variables < *arrow variables* . . . >

The *directed-path* continues alternating between *arrows* and *variables* and terminates with either a comma (which ends the path) or a semicolon (which ends the MODEL statement). Each *variables* argument contains a list of variable names. Optionally, you can enclose this list of names in curly braces, square brackets, or parentheses. This means that all the following forms are equivalent:

- *variables*
- {*variables*}
- [*variables*]
- (*variables*)

The use of braces, brackets, or parentheses for grouping variables is optional but highly recommended because it clearly identifies the edges that are associated with each variable.

Each *arrow* in a *directed-path* defines a set of edges in a DAG. Each variable in the list preceding the *arrow* is linked by an edge to each variable in the list following the *arrow*. The direction of the edge is given by the direction of the *arrow*. An *arrow* in a *directed-path* can be a right arrow (**==>**), a left arrow (**<==**), or a bidirected arrow (**<==>**). For more information about representing arrows in the CAUSALGRAPH procedure, see the section "Arrow or Edge Specification" on page 2267.

The procedure does not allow multiple edges of the same type between two variables. If the same edge is specified more than once in a MODEL statement, the repeated specifications are ignored. Variable names are not case-sensitive in the procedure.

Here are some examples of specifying *directed-paths*:

```
model 'M1'
   Y ==> Z,
   U <== W,
   X U V ==> Y,
   W ==> Z <== M N ==> Y;
```

The following MODEL statement specifications of "M2" are equivalent:

```
model 'M2' V1 V2 ==> X1-X3 A <== B C <==> D;
model 'M2' {V1 V2} ==> {X1-X3 A} <== {B C} <==> D;
model 'M2' V1 V2 ==> X1-X3 A,
           D <==> {B C} ==> X1-X3 A;
model 'M2' V1 ==> X1, V1 ==> X2, V1 ==> X3, V1 ==> A,
   V2 ==> X1, V2 ==> X2, V2 ==> X3, V2 ==> A,
   X1 <== B, X2 <== B, X3 <== B, A <== B,
   X1 <== C, X2 <== C, X3 <== C, A <== C,
   B <==> D, C <==> D;
```

Covariance-paths

A *covariance-path* has the following form:

<==> *variables*

The **<==>** syntax represents a bidirected arrow. For more information about representing arrows in the CAUSALGRAPH procedure, see the section "Arrow or Edge Specification" on page 2267. The *variables* argument contains a list of variable names. Optionally, you can enclose this list of names in curly braces, square brackets, or parentheses. This means that all the following forms are equivalent:

- *variables*
- {*variables*}
- [*variables*]
- (*variables*)

The use of braces, brackets, or parentheses for grouping variables is optional but highly recommended because it clearly identifies the edges that are associated with each variable.

You use a *covariance-path* to specify covariances between pairs of variables that are not explained by the causal paths in the model. Essentially, these covariances are equivalent to assuming latent confounding between each pair of variables that you specify in the *covariance-path*. That is, one bidirected edge is added to the model for each pair of unique variables in the *covariance-path*. Thus the following specifications are equivalent:

```
model 'MyModel' <==> {X1 X2};
model 'MyModel' X1 <==> X2;
```

Furthermore, because a bidirected edge represents latent confounding, this is also equivalent to the following specification:

```
model 'MyModel' X1 <== L ==> X2;
unmeasured L;
```

For more information about the interpretation of bidirected edges in the CAUSALGRAPH procedure, see the section "Causal Graph Theory" on page 2261.

Here are some examples of specifying *covariance-paths*:

```
model 'M3'
   <==> {X1 X2},
   <==> {X3-X5 X8};
```

The following MODEL statement specifications of "M4" are equivalent:

```
model 'M4' <==> {X1-X3 Y Z};
model 'M4' <==> {X1 X2 X3 Y Z};
model 'M4'
   <==> {X1 X2},
   <==> {X1 X3},
   <==> {X1 Y},
   <==> {X1 Z},
   <==> {X2 X3},
   <==> {X2 Y},
   <==> {X2 Z},
   <==> {X3 Y},
   <==> {X3 Z},
   <==> {Y Z},
model 'M4'
   X1 <==> X2 X3 Y Z,
   X2 <==> X3 Y Z,
   X3 <==> Y Z,
   Y <==> Z;
```

TESTID Statement

TESTID < '*label*' > *variable-list* < / *options* > **;**

The TESTID statement tests whether a criterion that you specify is valid for identifying a causal effect. The form of the *variable-list* depends on which method you specify in the METHOD= option in the PROC CAUSALGRAPH statement and is described later in this section. The causal effect to be identified is specified by the IDENTIFY statement. You must specify the IDENTIFY statement in order to use a TESTID statement.

For example, the following code tests two specified adjustment sets to determine whether each set is sufficient for the unbiased estimation of the causal effect of X on Y:

```
proc causalgraph method=adjustment;
   model 'TESTID Demo'
      Z ==> X ==> Y,
      U => X Y,
      W => Z U;
   identify X => Y;
   testid U;
   testid U W;
run;
```

The following example tests whether you can use the variable Z as an instrument for the causal effect of X on Y when U is unmeasured:

```
proc causalgraph method=iv;
   model 'TESTID Demo'
      Z ==> X ==> Y,
      U => X Y,
      W => Z U;
   identify X => Y;
   testid Z;
   unmeasured U;
run;
```

For more information about when an estimation criterion is valid, see the section "Identification and Adjustment" on page 2265.

When METHOD=ADJUSTMENT or BACKDOOR, or when you omit the METHOD= option, *variable-list* is a list of adjustment variables. This list of variables is tested to see whether it forms a valid adjustment set according to the criterion that you specify in the METHOD= option. For these adjustment methods, the list of variables can be empty. If you specify a set of prescribed adjustment variables in the IDENTIFY statement, then these prescribed adjustment variables are added to the list of variables that you specify in each TESTID statement. If *variable-list* is empty and there are no prescribed adjustment variables in the IDENTIFY statement, then the TESTID statement tests whether the causal effect is identified without any adjustment.

When METHOD=IV, *variable-list* consists of a single variable name. This variable name is required and cannot be empty. The single variable is tested to see whether it is an instrumental variable that can be used to estimate the causal effect that you specify in the IDENTIFY statement. If you specify a set of prescribed adjustment variables in the IDENTIFY statement, these variables are ignored.

Regardless of the value of the METHOD= option, every variable in a TESTID statement must be observed. This means that any variable name in *variables* cannot also be specified in the UNMEASURED statement.

You can specify multiple TESTID statements in each run of the CAUSALGRAPH procedure. Each proposed criterion that you specify in a TESTID statement is tested using each model that you specify in a MODEL statement.

The *label* is optional and can be any string of characters enclosed in quotation marks. If you specify a *label*, it must be the first item in the TESTID statement. If you do not specify a *label*, then a test name is automatically generated. The autogenerated name has the form “Test*i*”, where *i* is the smallest positive integer such that each test name is unique. It is recommended that you use your own labels so that you can identify the tests easily in the output results.

You can specify the following *options* in the TESTID statement:

CONDITIONAL={*variables*}

CONDITIONAL=[*variables*]

CONDITIONAL=(*variables*)

CONDITIONAL=*variable*

performs a test for a conditional instrumental variable, where the *variables* or *variable* is a list of conditioning variables. You can specify this option only if you specify METHOD=IV in the PROC CAUSALGRAPH statement. If you use this option, the procedure determines whether the instrumental variable that you specify in the TESTID statement is a conditional instrument associated with the specified conditioning set of *variables*.

PATHS < =(*path-types*) >

PATHS < =*path-type* >

creates an enumeration of the proper paths from the treatment variables to the outcome variables in each model. You cannot use this option if you specify METHOD=IV in the PROC CAUSALGRAPH statement.

A proper path is a path that begins with a treatment variable and does not contain any other treatment variables. The procedure determines whether each path is causal or noncausal. It also applies the adjustment set that is indicated by the TESTID statement and then determines whether each path is blocked or unblocked.

You can use the *path-types* argument to change the paths that are displayed. You can specify the following *path-type* values:

ALL	displays all paths.
BLOCKED	displays blocked paths.
CAUSAL	displays causal paths.
NONBLOCKED	displays nonblocked paths.
NONCAUSAL	displays noncausal paths.

If you do not specify *path-types*, then PATHS=(ALL) is used by default. You can specify multiple values for *path-types* at the same time. For example, you can specify PATHS=(CAUSAL BLOCKED) to display all paths that are both causal and blocked.

UNMEASURED Statement

UNMEASURED *variables* **;**

The UNMEASURED statement specifies a list of variables that are treated as unmeasured. You might want to specify a variable as unmeasured in the following situations:

- The variable corresponds to a latent construct that cannot be measured.
- The variable was not measured or will not be measured in the data set for the causal effect estimation.
- The data that are collected for the variable are not considered reliable enough to include in the causal effect estimation.

You can specify only one UNMEASURED statement.

If you specify multiple models by using multiple MODEL statements, each variable that you specify in the UNMEASURED statement is treated as unmeasured in every model. In addition, any variable that you specify in the IDENTIFY statement or in a TESTID statement cannot also be specified in the UNMEASURED statement. That is, the treatment and outcome variables, the adjustment variables, and the instrumental variables are all assumed to be measured or observed.

Details: CAUSALGRAPH Procedure

In a simple randomized controlled study, experimental units are randomly assigned to either a treatment group or a control (that is, untreated or unexposed) group. As a result of this randomization step, the two groups have the same distribution of covariates. Thus you can estimate the causal effect of the treatment by a direct comparison of the outcome variables in the two groups.

In a nonrandomized study, such as an observational study, the units of interest are not randomized to the treatment and control groups. As a result, some covariates can be correlated with both the group assignment and the outcome variables. In this case, the values of the outcome variables are determined both by the causal effect of treatment and by spurious association with covariates. Because the covariates confound the causal effect, you cannot determine the causal effect of the treatment without some form of adjustment (for example, matching or stratification) to remove the spurious association.

In fact, spurious association that is induced by confounding covariates can also be present in imperfect randomized controlled studies. For example, because of idiosyncratic background factors, subjects might not fully comply with the treatment protocol.

You can use the CAUSALGRAPH procedure to determine when, and under what circumstances, it is possible to estimate a causal effect that has a valid causal interpretation. To do so, you define causal models in the form of directed acyclic graphs (DAGs). The procedure accepts the input DAG and outputs implications of the models that are useful for causal analysis and effect estimation. The sections that follow describe the link between causal models and DAGs and discuss important properties of DAGs that you can use to identify causal effects.

Statistical and Causal Concepts

A statistical model is a mathematical rule for computing quantities of interest from the joint distribution function for a set of observed variables. For example, you can use a statistical model to answer queries such as the probability of a specific event and how that probability would change as different variables are observed (Pearl 1993). However, such models and data are purely associative. Because associations typically contain a mix of causal and noncausal components, statistical approaches alone are not sufficient to answer causal queries (Elwert 2013). This is because a causal query seeks to predict the effect of an action that intervenes in and changes the data generating process (Pearl 2009b). Such queries are essential for estimating the effects of treatments or assessing the impact of public policies (Pearl 1993). In order to answer a causal query by using data from a nonrandomized experiment, you must supplement the joint distribution function with a set of causal assumptions that, together, form a causal model.

Causal Graph Theory

A causal model represents beliefs about the data generating process that is being studied. That is, it defines the causal relationships that determine how the value of each variable is determined. These beliefs reflect an existing state of knowledge, including expert opinion and past experience. Constructing a causal model requires not only expertise in the subject matter being studied but also knowledge of the measurement process so that the factors affecting each variable can be accurately reflected in the model.

DAGs provide a formal semantics for defining and manipulating a causal model. The theoretical developments that link DAGs with causal analysis are most closely associated with the work of Pearl and his colleagues. See Pearl (2009b) for a detailed overview or Pearl (2010) for an abbreviated review. Lauritzen (1996) provides a technical treatment of the probabilistic properties of graphical models (including DAGs). Koller and Friedman (2009) and Spirtes, Glymour, and Scheines (2001) provide extensive treatments of the computational tools available for analyzing data by using graphical models. For an accessible summary of using DAGs to identify causal effects, see Elwert (2013) and Elwert and Winship (2014). After you decide on a valid identification strategy, Schafer and Kang (2008) provide a useful and accessible summary of computational tools for estimating the average causal effect.

Components of a Causal Graph

A DAG consists of three components (Elwert 2013):

- nodes
- edges
- missing edges

Each node in the DAG represents a variable that is assumed to play a causal role in the process being studied. Each variable can have any distribution. It is not necessary for every node in the DAG to correspond to a variable that has been (or could be) measured. For example, some variables in the DAG might correspond to latent constructs that are assumed to play a causal role in the process being modeled but that cannot possibly be observed directly. By convention, error random variables (independent error terms) are not represented in a DAG (Elwert 2013) unless the error random variable is a common cause of two or more variables that the model already includes (Spirtes, Glymour, and Scheines 2001).

All edges in a DAG are directed. That is, an edge consists of an arrow that points from one node into another node. An edge is a graphical representation of a causal assumption. Specifically, an edge in a DAG represents an assumed *possible* direct causal effect of one variable on another (Elwert 2013). These causal effects are assumed to be deterministic (Pearl 1993), but they are fully nonparametric in the sense that each edge can have any functional form (Elwert and Winship 2014). There can be at most one edge between a pair of nodes in a DAG.

Because each edge in a DAG is given a causal interpretation, each edge is associated with a temporal ordering of a pair of nodes. For this reason, the DAG cannot contain a directed cycle. The CAUSALGRAPH procedure performs a semantic validation of every model to verify that it does not contain a directed cycle.

PROC CAUSALGRAPH allows bidirected edges to be specified. A bidirected edge is interpreted as unmeasured confounding between two variables, and thus the graph is still a DAG. For example, the edge

$$X \leftrightarrow Y$$

is interpreted as the pair of edges

$$X \leftarrow L \rightarrow Y$$

where the node L represents some unmeasured variable.

Missing edges in a DAG (that is, where two nodes are *not* directly connected by an edge) indicate an assumption of exactly zero direct causal effect. Thus a missing edge in a DAG represents a much stronger assumption than an edge. This is the *strong null hypothesis* for graphical models; it is known in the econometrics literature as an *exclusion restriction* (Elwert 2013). Missing edges have implications for the statistical properties that are implied by a causal model. For more information about these properties, see the section “Statistical Properties of Causal Models” on page 2264.

Together the nodes, edges, and missing edges in a DAG form a causal model that encodes researchers’ assumptions about a data generating process. Starting with these data generating assumptions, you can use a set of graphical rules that operate on the DAG to derive statements of statistical association. For more information about sources of statistical association in DAGs, see the section “Sources of Association and Bias” on page 2263.

Terminology

Two variables in a DAG are *adjacent* if they are directly connected by a single edge.

A *path* is an ordered list of variables in which no variable appears more than once and consecutive variables in the list are adjacent in the graph. The edges that connect consecutive nodes in a path can point in either direction.

A path is *causal* if, for every consecutive pair of variables on the path, the arrow that connects the two variables points toward the latter variable. A path that is not causal is called *noncausal*. A path is *proper* if it begins with a treatment variable and does not contain any other treatment variables. A path can also be *blocked* or *nonblocked*. For information about blocked and nonblocked paths, see the section “Statistical Properties of Causal Models” on page 2264.

A DAG encodes specific relationships between the variables in a causal model. It is standard practice to describe these relationships by using familial adjectives; for example, see Koller and Friedman (2009); Pearl (2009b); Elwert and Winship (2014); Van der Zander, Liśkiewicz, and Textor (2014). For an edge

$$P \rightarrow Q$$

P is the *parent* of Q and Q is the *child* of P. If there is a causal path from a variable S to a variable T, then T is a *descendant* of S and S is an *ancestor* of T. Thus the set of descendants of a variable S is the set of all variables that are caused (either directly or indirectly) by S. Similarly, the set of ancestors of a variable T is the set of all direct or indirect causes of T.

For a variable V on a path, V is a *collider* on the path if it has two arrows (one on each side) that point to it. A variable that is not a collider is a *noncollider*. The definition of a collider is path-specific. A variable can be a collider on one path but a noncollider on another path.

Statistical association between sets of variables can be divided into two components: a causal component and a noncausal or spurious component. If all spurious association can be removed, the causal effect is said to be *identified*. *Identification analysis* is the process of determining whether a causal effect can be identified and, if so, how to identify that effect. For more information about identification analysis, see the section "Identification and Adjustment" on page 2265.

Spurious association is typically removed through some form of statistical *adjustment* or *conditioning*. For example, you could compute an adjustment by including a variable as a regressor in a regression model or by stratifying the analysis by levels of the variable. You can use the causal model to determine which variables must be included in an adjustment set.

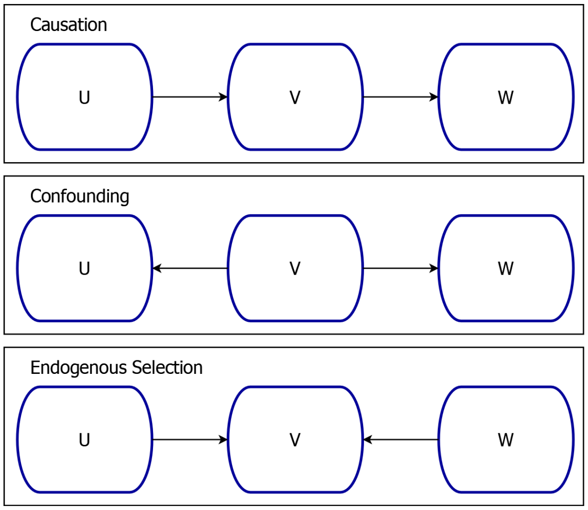
Sources of Association and Bias

A causal model that is represented by a DAG has unambiguous implications for the manner in which information can flow in the underlying data generating process. This flow of information is encapsulated by three graphical constructs that can be used to assemble every path in a DAG (Elwert 2013). The three constructs, which correspond to the three fundamental sources of association in a causal model, are as follows:

- causation
- confounding
- endogenous selection

These constructs are summarized graphically in Figure 34.4.

Figure 34.4 Three Fundamental Sources of Association



In the causal structure

$$U \rightarrow V \rightarrow W$$

the variables U and W are associated, and this association is the result of the causal chain. If you were to condition on the mediating variable V, then this would block the flow of information such that U and W would no longer be associated.

In the confounding structure

$$U \leftarrow V \rightarrow W$$

there is no causal path that relates U and W. However, U and W are still associated. This association is induced by the confounding variable V, the common parent of U and W. If you were to condition on the common cause V, then this would block the flow of information such that U and W would no longer be associated.

In the endogenous selection structure

$$U \rightarrow V \leftarrow W$$

the variables U and W jointly determine the value of their common child V, but U and W are not associated. However, if you were to condition on the common outcome V, then this would create a flow of information such that U and W would then be associated. For examples of endogenous selection, see Elwert and Winship (2014).

Loosely speaking, if you have a treatment variable (such as U) and an outcome variable (such as W) in a causal analysis, the goal is to eliminate noncausal association between U and W and leave the causal association unchanged. Thus the three fundamental graphical constructs correspond not only to the three fundamental sources of association but also to the three fundamental sources of bias. Generally, when you have a set of treatment and outcome variables, if you control for a variable on a causal path, you block the flow of information along that causal path. This is called overcontrol bias. Similarly, if you fail to control for a confounding common cause, some of the association between the treatment and outcome variables is the result of this confounding. This is called confounding bias. Finally, if you control for a common outcome, you create association between the treatment and outcome variables that is not causal. This is called endogenous selection bias. For an accessible discussion of the three fundamental sources of association and bias, see Elwert (2013) and Elwert and Winship (2014).

Statistical Properties of Causal Models

There are two ways to interpret the assumptions that are encoded within a DAG:

- A DAG is a "formal language for organizing claims about external interventions and their interactions" (Pearl 1993). For more information about this interpretation, see the section "Components of a Causal Graph" on page 2261.
- A DAG is a set of structures that define the flow of information between a set of variables. For more information about this interpretation, see the section "Sources of Association and Bias" on page 2263.

These two interpretations are equivalent under two additional assumptions (Elwert 2013):

- The variables in a DAG satisfy the *local Markov property*.
- The DAG satisfies the *weak faithfulness* property.

The local Markov property states that every variable in the DAG is statistically independent, conditional on its parents, of its set of nondescendants. In other words, the joint distribution function that is defined by the data generating process *factorizes* over the DAG (Koller and Friedman 2009).

The weak faithfulness property is discussed after the definition of d-separation.

A path in a DAG is said to be *d-separated* by a set of variables $\mathbf{Z}$ if either of the following conditions holds:

- The path contains a chain $U \rightarrow V \rightarrow W$ or a fork $U \leftarrow V \rightarrow W$ such that $V \in \mathbf{Z}$.
- The path contains a collider $U \rightarrow V \leftarrow W$ such that $V \notin \mathbf{Z}$ and such that no descendant of V is in $\mathbf{Z}$.

A path that is d-separated is said to be *blocked*; otherwise it is *nonblocked*. A set of variables $\boldsymbol{X}$ is d-separated from a set of variables $\boldsymbol{Y}$ by a set of variables $\boldsymbol{Z}$ if every path between a node in $\boldsymbol{X}$ and a node in $\boldsymbol{Y}$ is blocked.

The blocked/nonblocked terminology reflects the flow of information in a causal model. If a path is blocked, then information does not flow through that path. If the path is nonblocked, then information might flow through that path. The link between d-separation and information flow is embodied in the assumption of *weak faithfulness*. Weak faithfulness states that if two variables, X and Y, are not d-separated in a DAG, then the two variables are dependent in at least one distribution that factorizes over the DAG. The practical importance of faithfulness is that it does not permit the exact cancellation of the effects in a path (Elwert 2013). The use of weak faithfulness (rather than faithfulness) is consistent with the interpretation of edges as possible, rather than certain, effects (Spirtes, Glymour, and Scheines 2001; Pearl 2009b; Elwert 2013).

By interpreting a causal model as a DAG that represents the flow of association between variables, you can transform the causal assumptions that underlie a DAG into conditional independence statements. Specifically, if two variables are d-separated in a DAG by a set $\mathbf{Z}$, then those two variables must be statistically independent conditional on $\mathbf{Z}$. In other words, d-separation is a *global Markov property*. If a conditional independence statement contains only observed variables, then you can perform a statistical test by using the observed data to see whether the independence statement holds. Thus, the d-separation criterion determines the set of observationally testable implications of a causal model (Elwert 2013).

In fact, the global Markov property for DAGs (d-separation) and the local Markov property for DAGs are logically equivalent (Koller and Friedman 2009). If you have a complete list of either the local or global Markov properties, you can derive the other list by using the semigraphoid axioms (Pearl and Verma 1987; Geiger and Pearl 1988). In the CAUSALGRAPH procedure, you can use the IMAP option in the PROC CAUSALGRAPH statement to request a list of these properties.

Identification and Adjustment

The statistical association between a pair of variables can be divided into two components: a causal component and a noncausal or spurious component. If all spurious association can be removed, the causal effect is identified. Thus one possible approach to identification is *identification by adjustment*, which is the basis of causal effect identification in regression and matching (Elwert and Winship 2014).

When you use identification by adjustment, you seek an *adjustment set*, a set of variables that, when controlled for in an analysis, blocks all the noncausal paths in a DAG without blocking any causal paths in that same DAG. The causal property of a path is inherited from the direction of the edges in a model. That is, the causal property is a property of the causal model and does not change during an analysis. However, whether a path is blocked depends not only on the structure of the DAG that represents the causal model but also on the set of variables that are included in the adjustment set. Thus, you must carefully choose an adjustment set so as to remove all confounding bias without introducing any overcontrol or endogenous selection bias.

The following criterion is necessary and sufficient for a valid adjustment set (Shpitser, VanderWeele, and Robins 2010; Perković et al. 2018). For a set of treatment variables $\boldsymbol{X}$ and a set of outcome variables $\boldsymbol{Y}$, a set of observed variables $\boldsymbol{Z}$ is a valid adjustment set if all the following conditions are present:

- $\boldsymbol{Z}$ blocks all noncausal paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- No variable in $\boldsymbol{Z}$ lies on a causal path or descends from a causal path from $\boldsymbol{X}$ to $\boldsymbol{Y}$.

- No variable in $\boldsymbol{Z}$ is a descendant of any variable on a causal path (except possibly the variables in $\boldsymbol{X}$).

This criterion is identical to the constructive backdoor criterion (Van der Zander, Liśkiewicz, and Textor 2014). In the CAUSALGRAPH procedure, you can specify the METHOD=ADJUSTMENT option in the PROC CAUSALGRAPH statement to list all adjustment sets that satisfy the constructive backdoor criterion.

Pearl's backdoor criterion (Pearl 2009b) is a stronger criterion and thus can be used to produce a smaller list of adjustment sets. In PROC CAUSALGRAPH, you can specify the METHOD=BACKDOOR option to list all adjustment sets that satisfy the backdoor criterion.

When a valid adjustment set $\boldsymbol{Z}$ has been identified, you can estimate the total causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ by using the stratification estimator (Shpitser, VanderWeele, and Robins 2010; Elwert 2013). For discrete data, this estimator has the form

$$P(\mathbf{Y} = \mathbf{y} | \mathrm{do}(\mathbf{X} = \mathbf{x})) = \sum_{z} P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z})$$

where the do-operator is intended to emphasize the interpretation of a causal effect as the result of an action or intervention (Pearl 2009b). In the language of potential outcomes, the preceding expression describes the distribution of potential outcomes under the assumption that the potential outcome and the treatment are independent after you condition on $\boldsymbol{Z}$. For more information about the relationship to potential outcomes, see the section "Causal Graphs and Potential Outcomes" on page 2266.

Although the stratification estimator is fully nonparametric, it is rarely used in practice. Instead, the distribution functions in the estimator are replaced by parametric functions. This task requires considerable care and expertise. Causal effect identification is a nonparametric concept. A poorly specified parametric model can produce biased or incorrect results. For a useful summary of computational tools that you can use to estimate an average causal effect, see Schafer and Kang (2008).

The existence of an adjustment set is sufficient, but not necessary, to determine that a causal effect is identified. When an adjustment set does not exist, it might still be possible to estimate a causal effect by using a different method. For example, it is sometimes possible to estimate a causal effect by using an instrumental variable approach even when there is unobserved confounding between a treatment and an outcome. You can use the METHOD=IV option in the PROC CAUSALGRAPH statement to see whether a causal effect can be identified using an instrumental variable.

Causal Graphs and Potential Outcomes

The CAUSALGRAPH procedure emphasizes the graphical or structural model framework of causality as developed by Pearl, among others (Pearl 2009b). This is in contrast with the Neyman-Rubin potential outcomes framework (Neyman, Dabrowska, and Speed 1990; Rubin 1980, 1990). Although the notation differs, these two frameworks are equivalent in the sense that any theorem that can be proved in one framework can also be proved in the other framework (Galles and Pearl 1998; Elwert 2013). For an extended discussion, see Pearl (2009b, chap. 7) and Pearl (2012).

For a single outcome variable *Y* and a single treatment variable *X*, the potential outcome *Y(x)* is a random variable that describes the possible values of the outcome variable for an experimental unit that is associated with the treatment *X*=*x*. The identity

$$P(Y(x) = y) = P(Y = y | \mathrm{do}(X = x))$$

establishes a map between the potential outcomes framework and the structural model framework (Pearl 2009b).

The do-operator is meant to emphasize the interpretation of a causal effect as the effect of an action or intervention (Pearl 2009b). That is, the quantity $P(Y = y|\mathrm{do}(X = x))$ reflects the distribution of the outcome variable that would result from the hypothetical act of intervening and imposing the condition X=x (Elwert 2013).

The correspondence between potential outcomes and the do-operator, together with the stratification estimator, indicates another important link between the potential outcomes framework and the structural model framework (Elwert 2013). In particular, the conditional ignorability of treatment assignment (in the potential outcomes framework) is exactly equivalent to the criterion for the existence of an adjustment set. That is, $Y(x) \perp X | \mathbf{Z}$ if and only if $\mathbf{Z}$ satisfies the adjustment criterion (Shpitser, VanderWeele, and Robins 2010). For more information about the adjustment criterion, see the section "Identification and Adjustment" on page 2265. Significantly, the adjustment criterion involves only observed quantities, whereas conditional ignorability requires reasoning about counterfactuals that might be unobserved (Elwert 2013).

Arrow or Edge Specification

In the CAUSALGRAPH procedure, edges in a causal model are specified in the MODEL statement as either right, left, or bidirected arrows. In addition, the treatment and outcome variables are separated by a right arrow in the IDENTIFY statement.

The following table summarizes the edges and the syntax that you can specify in PROC CAUSALGRAPH:

Edge Type	Syntax	Statements
Right arrow	>, ->, ->, -->, =>, ==>, ===>	MODEL, IDENTIFY
Left arrow	<, <-, <-, <--, <=, <==, <===	MODEL
Bidirected arrow	<>, <->, <->, <-->, <=>, <==>, <===>	MODEL

Limitations of the CAUSALGRAPH Procedure

Causal Models with Directed Cycles

PROC CAUSALGRAPH analyzes DAGs that represent a causal model. These DAGs cannot contain a directed cycle. This can lead to difficulties in situations where two variables seem to cause (directly or indirectly) each other. In such situations, a common approach is to introduce additional variables so as to describe the data generating process at a more refined temporal scale (Greenland, Pearl, and Robins 1999; Elwert 2013).

Identifying Joint Treatment Effects

The CAUSALGRAPH procedure enables you to specify multiple treatment variables and multiple outcome variables in an identification analysis.

When you specify multiple treatment variables, the causal effect is interpreted as a *joint causal effect*. That is, the causal effect is interpreted as the hypothetical result of imposing specific values on all treatment variables simultaneously (Elwert 2013). In the language of the do-operator, the joint causal effect is treated as a conjunction so that $\mathrm{do}(X_1 = x_1, X_2 = x_2)$ is interpreted as $\mathrm{do}(X_1 = x_1)$ and $\mathrm{do}(X_2 = x_2)$.

You can also interpret multiple treatment variables as sequential treatment actions, provided that the treatment sequence is predetermined (Elwert 2013). However, you cannot use PROC CAUSALGRAPH to assess the identifiability of a dynamic treatment regime.

When you specify multiple outcome variables, each outcome is interpreted separately as a unique causal effect. Although the interpretation is separate, PROC CAUSALGRAPH constructs only those adjustment sets that are valid for every outcome variable. In some situations, there might not be any such adjustment sets, even though it is possible to identify the causal effect on each outcome separately. For example, if the causal effect of X on Y_1 can be identified only with an adjustment set $\mathbf{Z}_1$ and the causal effect of X on Y_2 can be identified only with an adjustment set $\mathbf{Z}_2$ for disjoint sets $\mathbf{Z}_1$ and $\mathbf{Z}_2$, then there is no adjustment set that is valid for both outcome variables simultaneously.

Causal Effect Identification Is a Population Concept

A causal effect that you estimate from observational data cannot have a valid causal interpretation unless those data are supplemented by a set of causal assumptions in the form of a causal model (Pearl 2009b). However, the causal model represents assumed relationships between variables at the population level and not at the level of an individual subject. Therefore, the theory that describes causal effect identification by using DAGs does not consider sampling variability. The conditions for identification are valid in the asymptotic limit (as the number of observations increases) (Elwert 2013). For this reason, a successful identification strategy (using either an adjustment set or a conditional instrumental variable) is a necessary first step to estimate a causal effect by using data from a nonrandomized experiment (Elwert and Winship 2014). You should carefully consider the role of sampling variability when estimating a causal effect and when examining the testable implications of a model.

Causal Effect Identification Is a Nonparametric Concept

The identifiability of a causal effect is a fully nonparametric concept in the sense that it does not depend on distributional or functional forms for the variables and edges in a causal model. However, an identification strategy as well as any estimate that is computed by that strategy should be understood to be conditional on the validity of the assumed causal model (Elwert 2013). In addition, when a causal effect is shown to be identified (for example, using an adjustment set), this does not mean that you can freely choose a parametric estimator in order to quantify the effect. The suitability of a parametric estimator is contingent on parametric assumptions. These assumptions are separate from the assumptions of the causal model and must be justified for each specific situation (Elwert 2013).

Dealing with Nonidentified Causal Effects

When a causal effect cannot be identified in a particular causal model, there are a couple of actions that you can take. First, you can revise the assumptions of the causal model to see whether the data generating process might be equally well described by an alternative model. In some cases, this might involve testing the observable implications of a model against existing data. For more information about the testable implications of a model, see the section “Statistical Properties of Causal Models” on page 2264. Second, you can consider observing additional variables. This might take the form of adding observations for a previously unmeasured variable or adding new variables and edges to an existing model (Pearl 2009b). However, adding edges to an existing set of variables never helps, and might harm, identification (Pearl 2009b; Elwert and Winship 2014).

Time Requirements

The following table summarizes the approximate computational cost to perform various tasks by using the CAUSALGRAPH procedure. The computational costs are computed for a DAG that contains N nodes (variables) and M edges.

For certain tasks, such as listing all adjustment sets, the output might grow exponentially large with the size of the graph. For these tasks, the computational cost is given in terms of delay complexity. The delay complexity is the computational cost that is associated with producing each element of the output set (Takata 2010).

The time required to...	**Is roughly proportional to**
Test an adjustment	N^2
Find one adjustment	N^2
List all adjustments	$N(N+M)$ delay
List all minimal adjustments	N^3 delay
Test an instrument	$N+M$
List all instruments	NM delay
List local Markov properties	$N(N+M)$
List global Markov properties	$N(N+M)$ delay

In general, the algorithms to test or find a single adjustment set have a computational cost proportional to $(N+M)$ (Van der Zander, Liśkiewicz, and Textor 2014). However, PROC CAUSALGRAPH always checks whether an adjustment set is minimal, and this operation is associated with a computational cost proportional to N^2.

By default, PROC CAUSALGRAPH sorts adjustment sets so that the smallest sets are printed first. You can use the MAXLIST= option in the PROC CAUSALGRAPH statement to limit the printed output, but this does not change the computational cost unless you also specify the NOSORT option in the same statement.

In certain situations, you might simply want to ascertain whether any adjustment set exists, regardless of which variables that adjustment set includes. To do this efficiently, you can request a single adjustment set by combining the NOSORT option with the MAXLIST=1 option. This enables the procedure to quickly find and display a single adjustment set. In the special case where you specify METHOD=ADJUSTMENT together with the NOSORT and MAXLIST=1 options in the PROC CAUSALGRAPH statement, the procedure switches to an even more efficient algorithm (Van der Zander, Liśkiewicz, and Textor 2014). For an illustration, see Example 34.2.

ODS Table Names

PROC CAUSALGRAPH assigns a name to each table that it creates. You can use these names to refer to the table when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 34.2. The specific statements and options that produce these output tables are shown in the last two columns. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 34.2 ODS Tables Produced by the CAUSALGRAPH Procedure

ODS Table Name	Description	Statement	Option
AdjustList	List of adjustment sets	PROC CAUSALGRAPH	LIST
AdjustTest	Test results of adjustment sets	TESTID	Default
ConnectedComponents	List of connected components	PROC CAUSALGRAPH	Default
DirectedCycles	List of directed cycles	PROC CAUSALGRAPH	Default
GraphicalModelSummary	Summary of the causal models		Default
Imap	List of derived conditional independence assumptions	PROC CAUSALGRAPH	IMAP
InstrumentList	List of instrumental variables	PROC CAUSALGRAPH	LIST
InstrumentTest	Test results of instrumental variables	TESTID	Default
Paths	Summary of proper paths in a model	TESTID	PATHS
VariableInfo	Summary of variables in the causal models		Default

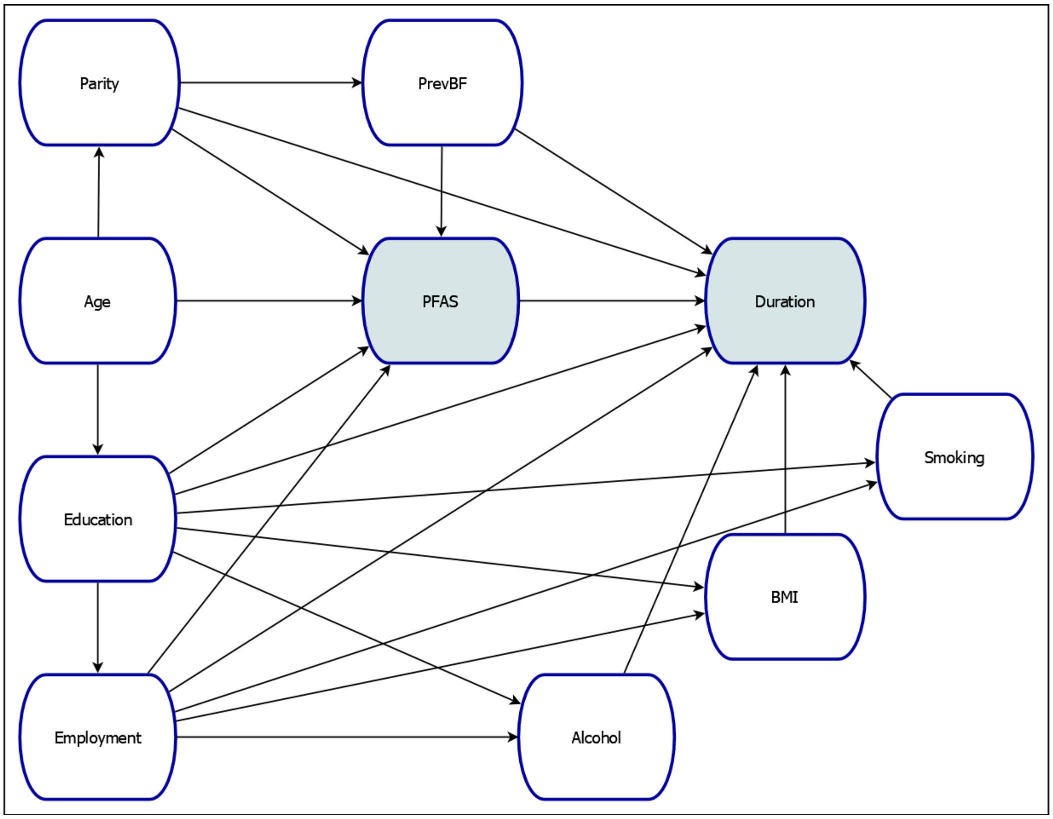
Examples: CAUSALGRAPH Procedure

Example 34.1: Constructing Adjustment Sets

This example illustrates how you can use the CAUSALGRAPH procedure to construct adjustment sets from a causal model. An adjustment set is a set of variables that can be used to remove noncausal association between the treatment and outcome variables in a causal model. If an adjustment set exists, then the causal effect of the treatment variables on the outcome variables is identified. The identification implies that it is possible to estimate the specified causal effects from data. For more information about identifying causal effects in a causal model, see the section "Identification and Adjustment" on page 2265.

The causal model shown in Figure 34.5, which has been adapted from Timmermann et al. (2017), examines the relationship between maternal exposure to persistent perfluoroalkyl substances (PFAS) and breastfeeding duration among residents of the Faroe Islands. For a summary of the variables in the model, see the example in the section "Getting Started: CAUSALGRAPH Procedure" on page 2246. For this example, it is assumed that all variables are observed.

Figure 34.5 Causal Model of the Effect of Persistent Perfluoroalkyl Substances on Breastfeeding Duration



In a naive approach to identify the causal effect of the variable PFAS on the variable Duration, you might consider an adjustment set that includes all observed covariates. The following statements invoke the CAUSALGRAPH procedure to test whether such an adjustment set is valid:

```
proc causalgraph;
   model "Timm17AllObs"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration Employment PFAS BMI Alcohol Smoking,
      Employment ==> Duration PFAS BMI Alcohol Smoking,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
   testid "All Covariates" Age Education Employment Parity Alcohol
      Smoking BMI PrevBF;
run;
```

In the MODEL statement, you specify the causal model to be analyzed. The quoted string in the statement labels the model. The remainder of the MODEL statement specifies all the variables and edges in the model. These variables and edges reflect the hypothesized data generating process shown in Figure 34.5.

In the IDENTIFY statement, you specify the causal effect of interest. You can use this statement to specify one or more treatment variables and one or more outcome variables. The treatment and outcome variables are separated by a single right arrow, ==>. In this example, you are interested in testing the identification of the causal effect of the variable PFAS on the variable Duration.

Because the METHOD= option is not specified in the PROC CAUSALGRAPH statement, the procedure uses the constructive backdoor criterion (METHOD=ADJUSTMENT) by default to test the identification of the causal effect by the adjustment set that you specify in the TESTID statement.

The "Variables in Model" and "Graphical Model Summary" tables (Output 34.1.1) summarize the variables and edges in the causal model. You can use this information as a qualitative check of the model specification.

Output 34.1.1 Input Summary Tables for the Causal Model in Figure 34.5

The CAUSALGRAPH Procedure

Variables in Model

	N	Variables
Measured	10	Age Alcohol BMI Duration Education Employment Parity PFAS PrevBF Smoking
Unmeasured	0	

Graphical Model Summary

Model	Nodes	Edges	Treatments	Outcomes	Measured	Unmeasured
Timm17AllObs	10	23	1	1	10	0

Output 34.1.2 displays the results of the test for adjustment. Although the proposed adjustment is sufficient to identify the causal effect, it is not a minimal adjustment. If you use this adjustment set, the estimation of the causal effect might be computationally inefficient. In addition, you have to collect data for all these variables in order to estimate the causal effect.

Output 34.1.2 Adjustment Set Test Output for the Causal Model in Figure 34.5

Covariate Adjustment Test: All Covariates

Causal Effect of PFAS on Duration

				Covariates							
Model	**Size**	**Valid**	**Minimal**	**Age**	**Alcohol**	**BMI**	**Education**	**Employment**	**Parity**	**PrevBF**	**Smoking**
Timm17AllObs	8	Yes	No	*	*	*	*	*	*	*	*

You can use the CAUSALGRAPH procedure to see whether there are any smaller adjustment sets that can be used to identify the causal effect shown in Figure 34.5. The following statements produce a list of all possible adjustment sets that can be used to estimate the causal effect of PFAS on Duration in the model:

```
proc causalgraph;
   model "Timm17AllObs"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration Employment PFAS BMI Alcohol Smoking,
      Employment ==> Duration PFAS BMI Alcohol Smoking,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
run;
```

As before, the METHOD= option is not specified here, so the constructive backdoor criterion (METHOD=ADJUSTMENT) is used by default. In this case, because no TESTID statement is specified, all adjustment sets are listed. Alternatively, when you specify a TESTID statement, you can use the LIST option in the PROC CAUSALGRAPH statement to list all adjustment sets.

As shown in Output 34.1.3, the causal model contains 16 valid adjustment sets that can be used to identify the causal effect of PFAS on Duration. These 16 sets differ in size. The last set contains eight covariates that you have tested previously by using the TESTID statement. If data collection cost is a concern, then you can consider other adjustment sets with smaller sizes. As shown in Output 34.1.3, the first 15 sets all have smaller sizes, ranging from four to seven. The first set is marked as a minimal because the identification of the specified causal effect would not hold if any one of these four variables were removed from the adjustment set.

Output 34.1.3 Adjustment Set List Output for the Causal Model in Figure 34.5

Covariate Adjustment Sets for Timm17AllObs

Causal Effect of PFAS on Duration

			Covariates							
	Size	Minimal	Age	Alcohol	BMI	Education	Employment	Parity	PrevBF	Smoking
1	4	Yes				*	*	*	*	
2	5	No	*			*	*	*	*	
3	5	No		*		*	*	*	*	
4	5	No			*	*	*	*	*	
5	5	No				*	*	*	*	*
6	6	No	*	*		*	*	*	*	
7	6	No	*		*	*	*	*	*	
8	6	No	*			*	*	*	*	*
9	6	No		*	*	*	*	*	*	
10	6	No		*		*	*	*	*	*
11	6	No			*	*	*	*	*	*
12	7	No	*	*	*	*	*	*	*	
13	7	No	*	*		*	*	*	*	*
14	7	No	*		*	*	*	*	*	*
15	7	No		*	*	*	*	*	*	*
16	8	No	*	*	*	*	*	*	*	*

You can use the MAXLIST=, MAXSIZE=, or MINIMAL option in the PROC CAUSALGRAPH statement to reduce the number of adjustment sets that are displayed. For example, the following statements use the MINIMAL option to display only minimal adjustment sets:

```
proc causalgraph minimal;
   model "Timm17AllObs"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration Employment PFAS BMI Alcohol Smoking,
      Employment ==> Duration PFAS BMI Alcohol Smoking,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
run;
```

The table in Output 34.1.4 shows all the minimal adjustment sets for the causal model. In this example, there is only a single minimal adjustment set.

Output 34.1.4 Minimal Adjustment Sets for the Causal Model in Figure 34.5

Covariate Adjustment Sets for Timm17AllObs

Causal Effect of PFAS on Duration

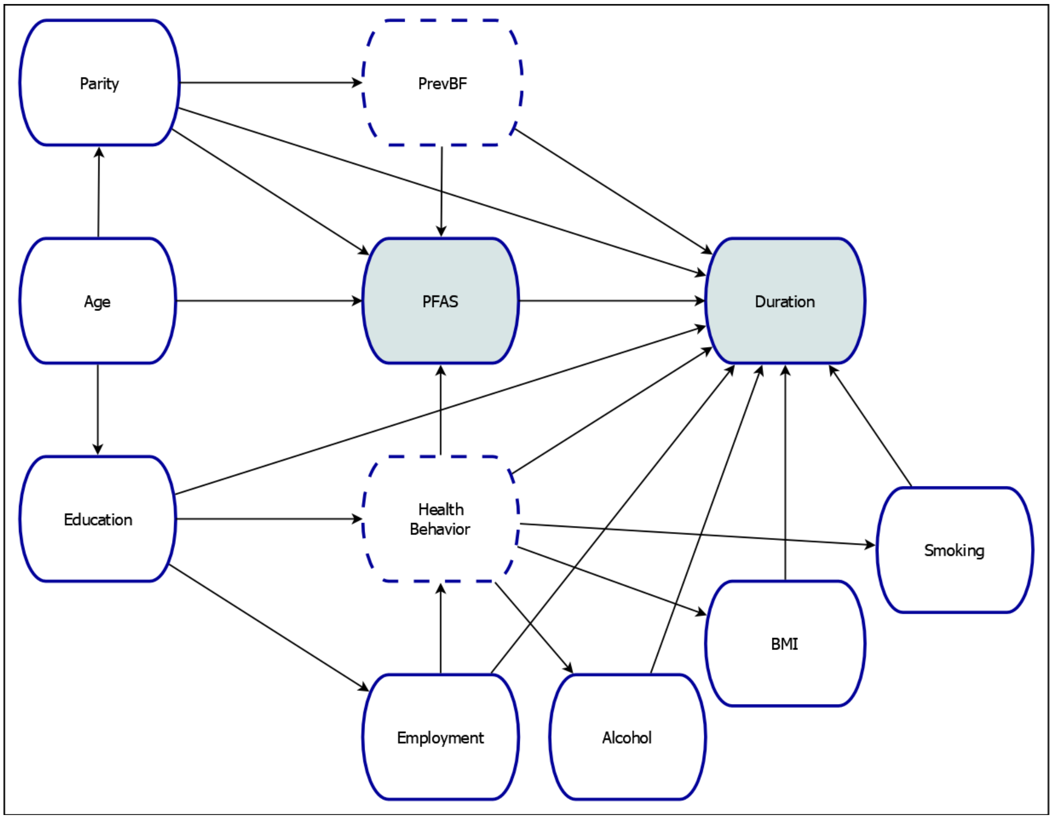
			Covariates							
	Size	Minimal	Age	Alcohol	BMI	Education	Employment	Parity	PrevBF	Smoking
1	4	Yes				*	*	*	*	

Example 34.2: Searching Efficiently for an Adjustment Set

This example illustrates how you can use the CAUSALGRAPH procedure to quickly determine whether an adjustment set exists in a particular causal model. If an adjustment set exists, then the causal effect of the treatment variables on the outcome variables is identified. The identification implies that it is possible to estimate the specified causal effects from the data. For this reason, you might want to determine whether such a set exists, regardless of what that set might be. For more information about identifying causal effects in a causal model, see the section "Identification and Adjustment" on page 2265.

The causal model shown in Figure 34.6, which has been adapted from Timmermann et al. (2017), examines the relationship between maternal exposure to persistent perfluoroalkyl substances (PFAS) and breastfeeding duration among residents of the Faroe Islands. For a summary of the variables in the model, see the section "Getting Started: CAUSALGRAPH Procedure" on page 2246. For this example, the causal model includes an additional variable, HealthBehavior, which is assumed to be a latent construct that represents the degree to which an individual's behavior is considered to be healthful. The variables HealthBehavior and PrevBF are assumed to be unobserved, as in Timmermann et al. (2017).

Figure 34.6 Causal Model of the Effect of Persistent Perfluoroalkyl Substances on Breastfeeding Duration



The following statements invoke the CAUSALGRAPH procedure to determine whether it is possible to find an adjustment set that can be used to estimate the causal effect:

```
proc causalgraph method=adjustment maxlist=1 nosort;
   model "Timm17HealthBehavior"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration HealthBehavior Employment,
      HealthBehavior ==> PFAS Duration BMI Alcohol Smoking,
      Employment ==> HealthBehavior Duration,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
   unmeasured PrevBF HealthBehavior;
run;
```

In the MODEL statement, you specify the causal model to be analyzed. The quoted string in the statement labels the model. The remainder of the MODEL statement specifies all the variables and edges in the model. These variables and edges reflect the hypothesized data generating process shown in Figure 34.6.

In the IDENTIFY statement, you specify the causal effect of interest. You can use this statement to specify one or more treatment variables and one or more outcome variables. The treatment and outcome variables are separated by a single right arrow, `==>`. In this example, you are interested in testing the identification of the causal effect of the variable PFAS on the variable Duration.

In the UNMEASURED statement, you specify variables that are not observed and thus cannot be included in an adjustment set. In this example, the variables PrevBF and HealthBehavior are unmeasured.

The METHOD=ADJUSTMENT uses the constructive backdoor criterion for identifying causal effects. The MAXLIST=1 prints only a single adjustment set. The NOSORT option specifies that it is not necessary to print the smallest adjustment sets first. These options enable the procedure to use an efficient algorithm to quickly find an adjustment set, if at least one adjustment set exists. For more information about the computational complexity of the algorithms that PROC CAUSALGRAPH uses, see the section "Time Requirements" on page 2269.

As shown in Output 34.2.1, for the causal model in Figure 34.6, it is not possible to use an adjustment set to identify the effect of PFAS on Duration.

Output 34.2.1 Adjustment Set List Summary Note

```
NOTE: There are no adjustment sets satisfying the specified criteria for Timm17HealthBehavior.
```

Although you cannot use an adjustment set to estimate the causal effect in Figure 34.6, it is still possible to estimate the causal effect if you are willing to make additional parametric assumptions in the model. See Example 34.5 for an example.

Example 34.3: Testing Adjustments and Enumerating Paths

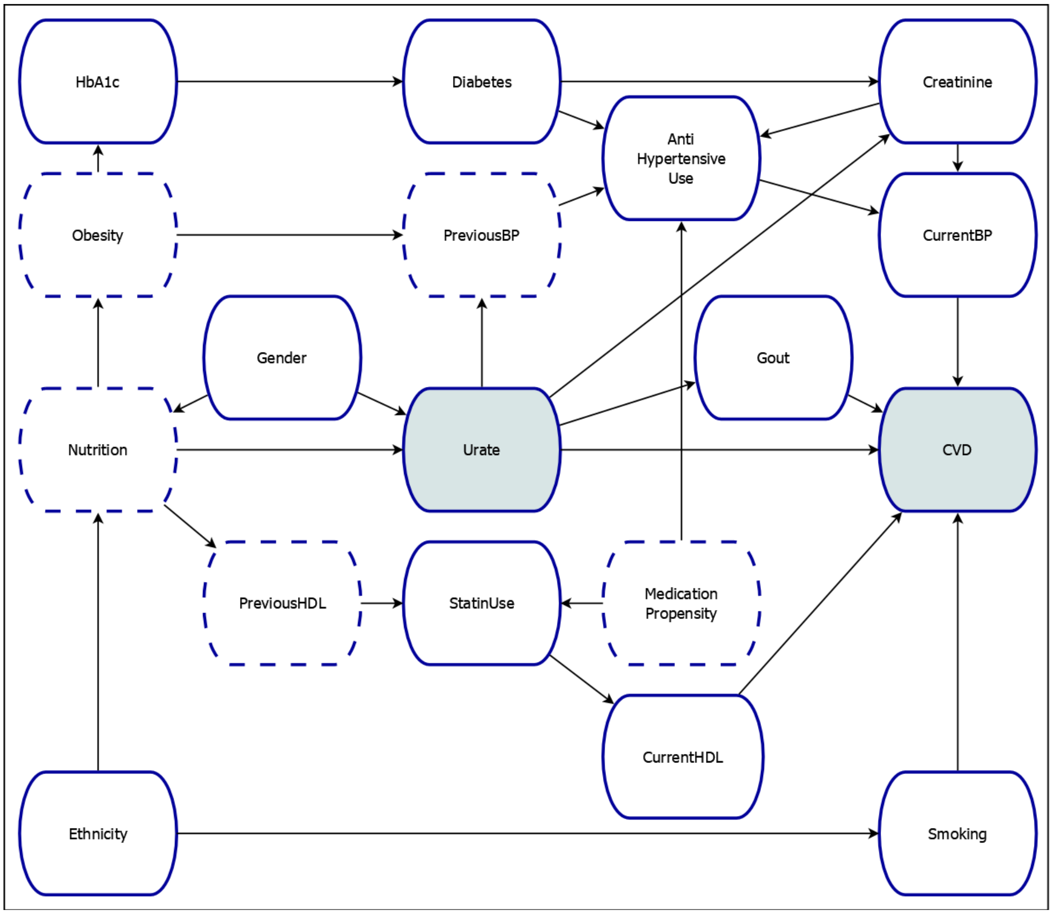
This example illustrates how you can use the CAUSALGRAPH procedure to determine whether a proposed adjustment set is valid for estimating a causal effect in a particular causal model. It also illustrates how you can enumerate the paths that connect the treatment and outcome variables in the model. If a causal effect cannot be identified in a causal model, the enumeration of paths can provide information that might help you modify the study design in order to enable identification of the causal effect.

The causal model shown in Figure 34.7 is used by Thornley et al. (2013) to examine the relationship between an individual's serum urate and risk of cardiovascular disease. The model includes the following variables:

- Urate: the treatment variable
- CVD: the outcome variable
- AntiHypertensiveUse: indicator of antihypertensive drug use
- Creatinine: measured serum creatinine level
- Diabetes: indicator of diabetes diagnosis
- Ethnicity: classification variable for ethnicity
- Gender: indicator for biological male
- Gout: indicator of gout diagnosis
- HbA1c: measured glycated hemoglobin
- MedicationPropensity: latent construct that reflects an individual's propensity to take prescribed medication
- Nutrition: latent construct that reflects diet or nutrition
- Obesity: indicator of body mass index ≥ 30
- CurrentBP: measured blood pressure
- CurrentHDL: measured HDL cholesterol
- PreviousBP: previous (prior to study) blood pressure
- PreviousHDL: previous (prior to study) HDL cholesterol
- Smoking: indicator of current smoking status
- StatinUse: indicator of statin drug use

The variables MedicationPropensity and Nutrition correspond to latent constructs and thus cannot be observed. It is also assumed that the variables PreviousBP, PreviousHDL, and Obesity are not observed.

Figure 34.7 Causal Model of the Effect of Serum Urate on Risk of Cardiovascular Disease



According to this causal model, the statistical association between the variables Urate and CVD reflects a combination of true causal association and additional spurious or noncausal association. In order to isolate the true causal association between these two variables, Thornley et al. (2013) consider adjustment for CurrentHDL, Ethnicity, Gender, HbA1c, and Smoking. The following code invokes the CAUSALGRAPH procedure to test whether this adjustment set can be used to estimate the causal effect of Urate on CVD according to the causal model:

```
proc causalgraph compact;
   model "Thor12"
      AntiHypertensiveUse ==> CurrentBP,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      CurrentBP ==> CVD,
      CurrentHDL ==> CVD,
      Diabetes ==> AntiHypertensiveUse Creatinine,
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      HbA1c ==> Diabetes,
      MedicationPropensity ==> AntiHypertensiveUse StatinUse,
      Nutrition ==> PreviousHDL Urate Obesity,
      Obesity ==> PreviousBP HbA1c,
      PreviousBP ==> AntiHypertensiveUse,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL,
      Urate ==> PreviousBP Creatinine CVD Gout;
   identify Urate ==> CVD;
   unmeasured Nutrition Obesity PreviousBP MedicationPropensity PreviousHDL;
   testid CurrentHDL Ethnicity Gender HbA1c Smoking;
run;
```

In the MODEL statement, you specify the causal model to be analyzed. The quoted string in the statement labels the model. The remainder of the MODEL statement specifies all the variables and edges in the model. These variables and edges reflect the hypothesized data generating process shown in Figure 34.7.

In the IDENTIFY statement, you specify the causal effect of interest. You can use this statement to specify one or more treatment variables and one or more outcome variables. The treatment and outcome variables are separated by a single right arrow, **==>**. In this example, you are interested in testing the identification of the causal effect of Urate on CVD.

In the UNMEASURED statement, you specify variables that are not observed and thus cannot be included in an adjustment set. In this example, five variables are specified as unmeasured.

Because the METHOD= option is not specified in the PROC CAUSALGRAPH statement, the procedure uses the constructive backdoor criterion (METHOD=ADJUSTMENT) by default to test the identification of the causal effect by the adjustment set that you specify in the TESTID statement.

The COMPACT option in the PROC CAUSALGRAPH statement displays the output table in a compact manner. For this example, this means that a column is added for a covariate in the adjustment test output table only if that variable actually appears in the test.

As shown in Output 34.3.1, the proposed adjustment set is marked "No" in the Valid column, so it is not sufficient to estimate the causal effect of Urate on CVD.

Output 34.3.1 Adjustment Set Test for the Model in Figure 34.7

Covariate Adjustment Test: Test1

Causal Effect of Urate on CVD

				Covariates				
Model	**Size**	**Valid**	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**HbA1c**	**Smoking**
Thor12	5	No	No	*	*	*	*	*

To see why the proposed adjustment set is not valid, you can request an enumeration of the proper paths that connect the treatment to the outcome in the model. You can also use the procedure to search for a valid adjustment set. The following code invokes the CAUSALGRAPH procedure to perform these two tasks:

```
   proc causalgraph compact list;
      model "Thor12"
         AntiHypertensiveUse ==> CurrentBP,
         Creatinine ==> AntiHypertensiveUse CurrentBP,
         CurrentBP ==> CVD,
         CurrentHDL ==> CVD,
         Diabetes ==> AntiHypertensiveUse Creatinine,
         Ethnicity ==> Nutrition Smoking,
         Gender ==> Nutrition Urate,
         Gout ==> CVD,
         HbA1c ==> Diabetes,
         MedicationPropensity ==> AntiHypertensiveUse StatinUse,
         Nutrition ==> PreviousHDL Urate Obesity,
         Obesity ==> PreviousBP HbA1c,
         PreviousBP ==> AntiHypertensiveUse,
         PreviousHDL ==> StatinUse,
         Smoking ==> CVD,
         StatinUse ==> CurrentHDL,
         Urate ==> PreviousBP Creatinine CVD Gout;
      identify Urate ==> CVD;
      unmeasured Nutrition Obesity PreviousBP MedicationPropensity PreviousHDL;
      testid Gender HbA1c Ethnicity Smoking
         CurrentHDL / paths=(noncausal nonblocked);
   run;
```

The PATHS option in the TESTID statement requests an analysis of the proper paths in the model when the variables in the TESTID statement are being adjusted for identifying the causal effect. By default, the procedure prints all proper paths when you specify the PATHS option. You can change this behavior by specifying additional suboptions. In this example, the NONCAUSAL and NONBLOCKED suboptions display only noncausal paths that are not blocked. For more information about the terminology of paths in DAGs, see the section "Terminology" on page 2262.

Output 34.3.2 shows the proper paths in the model that are not causal and are not blocked. There are two such paths. If you were to use the proposed adjustment set, some of the association between Urate and CVD would be attributable to these two noncausal paths, and the causal effect would not be estimated correctly. The first path is not blocked because the variable StatinUse is a collider on the path and one of its descendants, the variable CurrentHDL, appears in the adjustment set. The second path does not contain any colliders but is not blocked because it does not contain any element of the proposed adjustment set. For more details about the flow of information in DAGs, see the section "Statistical Properties of Causal Models" on page 2264.

Output 34.3.2 Analysis of Proper Paths

Treatment-to-Outcome Paths

Test Test1 for Model Thor12

Causal Effect of Urate on CVD Adjusted for {CurrentHDL Ethnicity Gender HbA1c Smoking}

	Causal	Blocked	Paths
1*	No	No	Urate <== Nutrition ==> PreviousHDL ==> StatinUse <== MedicationPropensity ==> AntiHypertensiveUse ==> CurrentBP ==> CVD
2*	No	No	Urate <== Nutrition ==> Obesity ==> PreviousBP ==> AntiHypertensiveUse ==> CurrentBP ==> CVD

*** indicates a biasing path**

Because the adjustment set in the TESTID statement leads to two nonblocked, noncausal paths, it is not valid for identifying the specified causal effect. The next question to ask is whether there are any valid adjustment sets. You can use the LIST option in the PROC CAUSALGRAPH statement to address this question. The LIST option lists all valid adjustment sets.

The note in Output 34.3.3 summarizes the results of the search for any adjustment set that can be used to identify the effect of Urate on CVD in the model in Figure 34.7. For this causal model, it is not possible to use an adjustment set to estimate the causal effect.

Output 34.3.3 Adjustment Set List Summary Note

NOTE: There are no adjustment sets satisfying the specified criteria for Thor12.

In order to obtain identifiability, you might consider collecting additional data. For example, if you were to collect data for the variables Obesity and PreviousHDL so that these two variables are no longer unmeasured, then you could block the two noncausal paths in Output 34.3.2. The following code demonstrates this:

```
proc causalgraph compact;
   model "Thor12"
      AntiHypertensiveUse ==> CurrentBP,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      CurrentBP ==> CVD,
      CurrentHDL ==> CVD,
      Diabetes ==> AntiHypertensiveUse Creatinine,
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      HbA1c ==> Diabetes,
      MedicationPropensity ==> AntiHypertensiveUse StatinUse,
      Nutrition ==> PreviousHDL Urate Obesity,
      Obesity ==> PreviousBP HbA1c,
      PreviousBP ==> AntiHypertensiveUse,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL,
      Urate ==> PreviousBP Creatinine CVD Gout;
   identify Urate ==> CVD;
   unmeasured Nutrition PreviousBP MedicationPropensity;
   testid Gender HbA1c Ethnicity Smoking
      CurrentHDL PreviousHDL Obesity;
run;
```

As shown in Output 34.3.4, the proposed adjustment set is marked "Yes" in the Valid column, so it is sufficient to estimate the causal effect of Urate on CVD. For more information about possible approaches to obtain identifiability, see the section "Dealing with Nonidentified Causal Effects" on page 2268.

Output 34.3.4 Adjustment Set Test

Covariate Adjustment Test: Test1

Causal Effect of Urate on CVD

				Covariates						
Model	**Size**	**Valid**	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**HbA1c**	**Obesity**	**PreviousHDL**	**Smoking**
Thor12	7	Yes	No	*	*	*	*	*	*	*

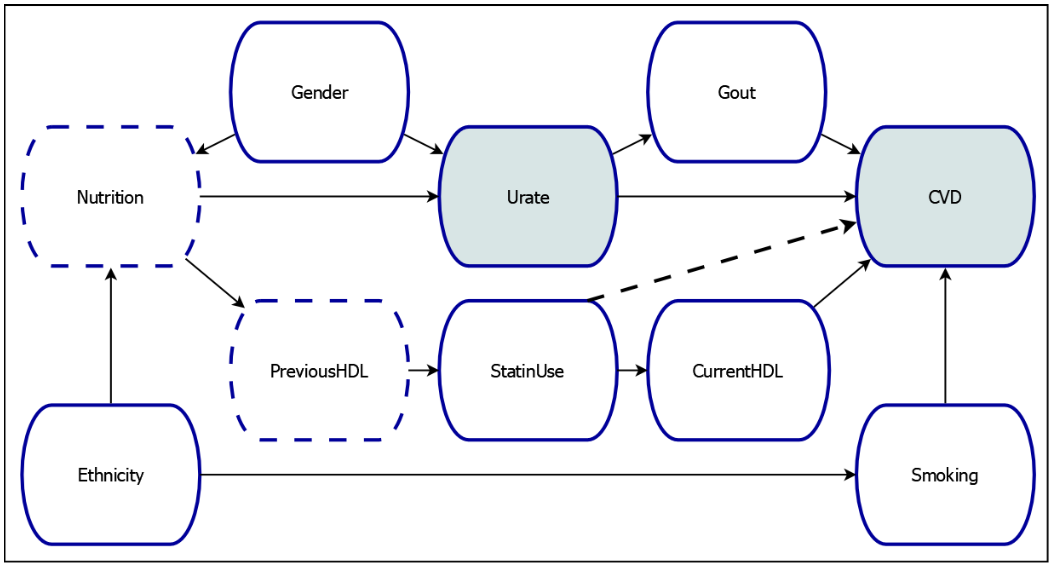
Example 34.4: Finding Adjustment Sets Common to Multiple Models

This example illustrates how you can use the CAUSALGRAPH procedure to find adjustment sets that you can use to estimate a causal effect when you are uncertain of the exact structure of the causal model or when alternative causal models might be possible for the data. For instance, you might not be sure whether to include an edge in a causal model, what direction an edge should have, or what covariates to include in the model. You can use multiple MODEL statements in the procedure to specify every plausible causal model and then use the COMMON option in the PROC CAUSALGRAPH statement to search for adjustment sets that are valid for all the models.

For a single causal model, if an adjustment set exists, then you can use this adjustment set to estimate the causal effect. If multiple causal models share a common adjustment set, then you can use this set to estimate the causal effect, no matter which causal model reflects the true data generating process. For more information about identifying causal effects in a causal model, see the section "Identification and Adjustment" on page 2265.

The models in this example are derived from a larger model that was developed by Thornley et al. (2013). These models examine the relationship between an individual's serum urate and risk of cardiovascular disease. See Example 34.3 for a summary of the variables in the models. Figure 34.8 depicts two plausible causal assumptions for the effect of the variable StatinUse on the variable CVD. One model assumes that all the causal influence of StatinUse on CVD is strictly mediated by the variable CurrentHDL. The other model includes a direct effect of StatinUse on CVD, in addition to the mediated effect. These two models differ by a single edge, which is indicated by the broken arrow in Figure 34.8. In both models in this example, the variables Nutrition and PreviousHDL are assumed to be unmeasured and thus are represented by nodes with broken outlines.

Figure 34.8 Two Possible Causal Models of the Effect of Serum Urate on Risk of Cardiovascular Disease



The following statements invoke PROC CAUSALGRAPH to construct adjustment sets for the two models in Figure 34.8:

```
proc causalgraph common;
   model "Thor12SimpleHDL"
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      Nutrition ==> PreviousHDL Urate,
      CurrentHDL ==> CVD,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL,
      Urate ==> CVD Gout;
   model "Thor12AltHDL"
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      Nutrition ==> PreviousHDL Urate,
      CurrentHDL ==> CVD,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL CVD,
      Urate ==> CVD Gout;
   identify Urate ==> CVD;
   unmeasured Nutrition PreviousHDL;
run;
```

When you analyze multiple causal models, you must use a separate MODEL statement for each causal model. Each MODEL statement must begin with a quoted string that provides a unique name for the model. The remainder of the MODEL statement specifies all the variables and edges in the model. The model **Thor12SimpleHDL** assumes that the causal association between StatinUse and CVD is strictly mediated by the variable CurrentHDL, whereas the model **Thor12AltHDL** includes a direct effect between StatinUse and CVD.

In the IDENTIFY statement, you specify the causal effect of interest. You can use this statement to specify one or more treatment variables and one or more outcome variables. The treatment and outcome variables are separated by a single right arrow, **==>**. In this example, you are interested in testing the identification of the causal effect of the variable Urate on the variable CVD.

In the UNMEASURED statement, you specify variables that are not observed and thus cannot be included in an adjustment set. In this example, the variables Nutrition and PreviousHDL are specified as unmeasured.

The COMMON option in the PROC CAUSALGRAPH statement requests a list of all adjustment sets that are common to all the models that you specify in the MODEL statements. In this example, the procedure produces three tables of adjustment sets: one for each model, and one that displays common adjustments. You can use the COMMON(ONLY) option to produce only the table of common adjustments without also producing separate tables of adjustment sets for each model.

The adjustment sets for the models **Thor12SimpleHDL** and **Thor12AltHDL** are shown in Output 34.4.1 and Output 34.4.2, respectively. Although both models contain the same number of variables, **Thor12SimpleHDL** has one less edge and a larger number of valid adjustment sets. This is usually, but not always, expected. Adding edges to a model does not generally improve the identifiability of a causal effect, and it is often harmful to identification. For more information about obtaining identifiability in a model, see the section "Dealing with Nonidentified Causal Effects" on page 2268.

Output 34.4.1 Adjustment Sets for `Thor12SimpleHDL`

The CAUSALGRAPH Procedure

Covariate Adjustment Sets for Thor12SimpleHDL

Causal Effect of Urate on CVD

			Covariates					
	Size	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**Gout**	**Smoking**	**StatinUse**
1	2	Yes	*	*				
2	2	Yes	*				*	
3	2	Yes		*				*
4	2	Yes					*	*
5	3	No	*	*	*			
6	3	No	*	*			*	
7	3	No	*	*				*
8	3	No	*		*		*	
9	3	No	*				*	*
10	3	No		*	*			*
11	3	No		*			*	*
12	3	No			*		*	*
13	4	No	*	*	*		*	
14	4	No	*	*	*			*
15	4	No	*	*			*	*
16	4	No	*		*		*	*
17	4	No		*	*		*	*
18	5	No	*	*	*		*	*

Output 34.4.2 Adjustment Sets for `Thor12AltHDL`

The CAUSALGRAPH Procedure

Covariate Adjustment Sets for Thor12AltHDL

Causal Effect of Urate on CVD

			Covariates					
	Size	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**Gout**	**Smoking**	**StatinUse**
1	2	Yes		*				*
2	2	Yes					*	*
3	3	No	*	*				*
4	3	No	*				*	*
5	3	No		*	*			*
6	3	No		*			*	*
7	3	No			*		*	*
8	4	No	*	*	*			*
9	4	No	*	*			*	*
10	4	No	*		*		*	*
11	4	No		*	*		*	*
12	5	No	*	*	*		*	*

Output 34.4.3 shows the adjustment sets that are common to both models. Coincidentally, the adjustment sets that are valid in both models are the same as the adjustment sets that are valid in the model that has the extra edge. Because at least one such common adjustment set exists, this analysis suggests that it is possible to estimate the causal effect of Urate on CVD by using any adjustment set from Output 34.4.3, regardless of whether there is a direct effect of StatinUse on CVD.

Output 34.4.3 Common Adjustment Sets

The CAUSALGRAPH Procedure

Covariate Adjustment Sets Common to All Models

Causal Effect of Urate on CVD

			Covariates					
	Size	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**Gout**	**Smoking**	**StatinUse**
1	2	Yes		*				*
2	2	Yes					*	*
3	3	No	*	*				*
4	3	No	*				*	*
5	3	No		*	*			*
6	3	No		*			*	*
7	3	No			*		*	*
8	4	No	*	*	*			*
9	4	No	*	*			*	*
10	4	No	*		*		*	*
11	4	No		*	*		*	*
12	5	No	*	*	*		*	*

Example 34.5: Identifying a Causal Effect by Using Instrumental Variables

This example illustrates how you can use the CAUSALGRAPH procedure to find an instrumental variable in a causal model. By using an instrumental variable, you can identify a causal effect even when there is unobserved confounding between a treatment variable and an outcome variable, which is a situation where the adjustment criterion might fail.

In Example 34.2, a causal model, adapted from Timmermann et al. (2017), is used to examine the relationship between maternal exposure to persistent perfluoroalkyl substances (PFAS) and breastfeeding duration among residents of the Faroe Islands. That example shows that you cannot construct an adjustment set to estimate the causal effect of the treatment variable PFAS on the outcome variable Duration. This is because there is confounding bias between the treatment and outcome variables that results from the unobserved variables HealthBehavior and PrevBF.

In many cases where there is unmeasured confounding, you can still estimate the causal effect if you are willing to assume that certain edges in the causal model have a particular parametric form. Then it might be possible to use an instrumental variable (Angrist, Imbens, and Rubin 1996; Imbens 2014). The following statements invoke PROC CAUSALGRAPH to list possible instrumental variables that you can use to estimate the causal effect:

```
proc causalgraph method=iv;
   model "Timm17HealthBehavior"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration HealthBehavior Employment,
      HealthBehavior ==> PFAS Duration BMI Alcohol Smoking,
      Employment ==> HealthBehavior Duration,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
   unmeasured PrevBF HealthBehavior;
run;
```

In the MODEL statement, you specify the same causal model as in Example 34.2. In the IDENTIFY statement, you specify that the causal effect of the variable PFAS on the variable Duration is of interest. By specifying the METHOD=IV option in the PROC CAUSALGRAPH statement, you use an instrumental variable (or a conditional instrumental variable) to identify the causal effect. The UNMEASURED statement specifies that the variables PrevBF and HealthBehavior are not observed and thus cannot be included as an instrument or in any conditional set.

As shown in the table in Output 34.5.1, there is one variable that you can use as an instrument to identify the causal effect of PFAS on Duration. The variable Age is a conditional instrument. That is, Age becomes an instrument after its association with the outcome variable Duration is blocked by conditioning on the five variables shown in Output 34.5.1. If an instrumental variable is a classical instrument (that is, you can use the variable as an instrument without conditioning on any other variables), then the procedure prints an empty conditioning set for that variable.

Output 34.5.1 Instrumental Variables for the Causal Model

Instrumental Variables for Timm17HealthBehavior

Causal Effect of PFAS on Duration

		Conditionals						
	Instrument	**Age**	**Alcohol**	**BMI**	**Education**	**Employment**	**Parity**	**Smoking**
1	Age		*	*	*	*	*	

The conditional sets that are produced in constructing an instrument might not be minimal. For instance, the following test shows that you can use Age as a conditional instrument if you adjustment for only the variables Education and Parity:

```
proc causalgraph method=iv;
   model "Timm17HealthBehavior"
      Age ==> Parity PFAS Education,
      Parity ==> PrevBF Duration PFAS,
      PrevBF ==> PFAS Duration,
      PFAS ==> Duration,
      Education ==> Duration HealthBehavior Employment,
      HealthBehavior ==> PFAS Duration BMI Alcohol Smoking,
      Employment ==> HealthBehavior Duration,
      BMI Alcohol Smoking ==> Duration;
   identify PFAS ==> Duration;
   unmeasured PrevBF HealthBehavior;
   testid "Minimal CIV" Age / conditional = (Education Parity);
run;
```

The model specification, the causal effect of interest, and the identification criterion are the same as those in the previous CAUSALGRAPH analysis in this example. What is new here is the inclusion of the TESTID statement. You want to investigate whether Age can serve as a conditional instrumental variable for the causal effect, where the conditioning variables are Education and Parity. This set of conditional variables is a proper subset of the proposed conditional variables in the previous analysis. As shown in Figure 34.5.2, the value for the Valid column is “Yes,” providing an affirmative answer to this investigation.

Output 34.5.2 Instrumental Variable Test Output

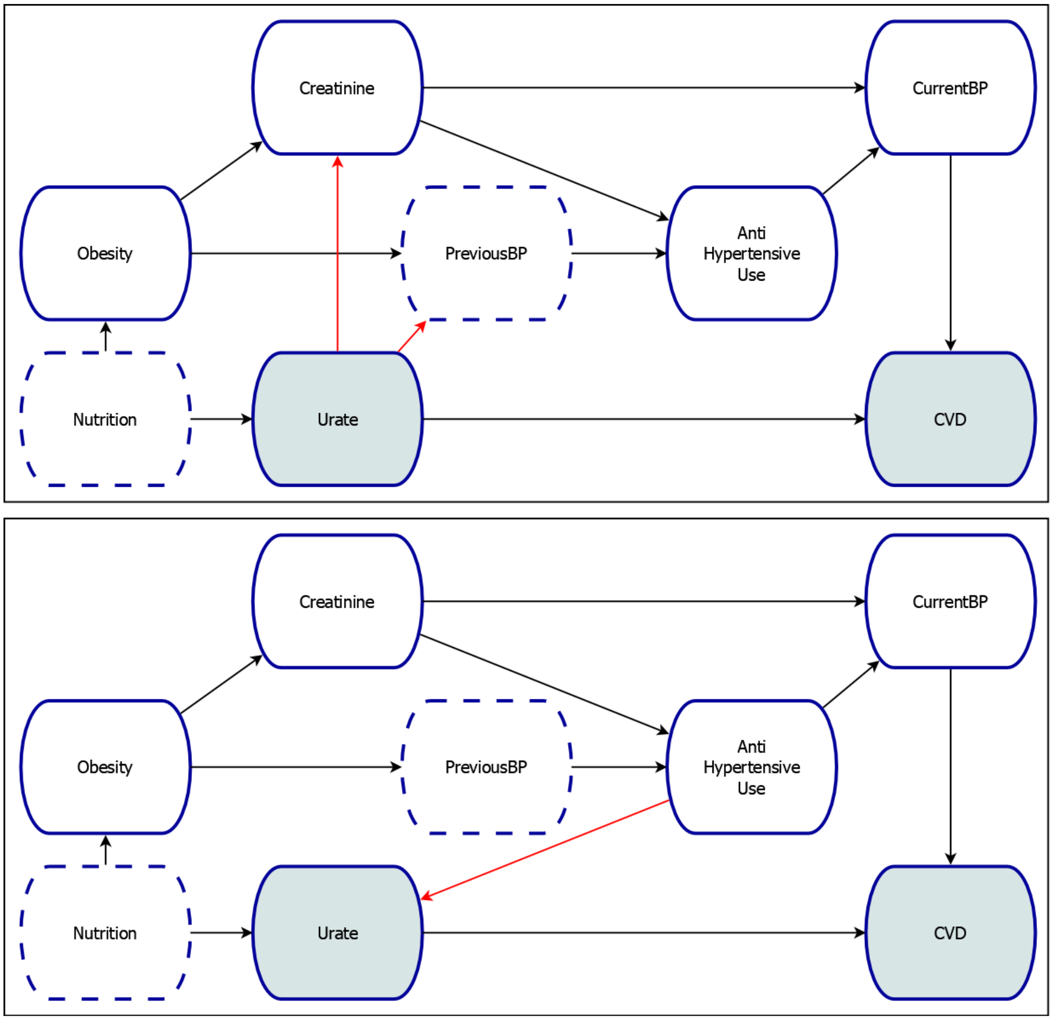
```
                          Instrumental Variable Test: Minimal CIV
                              Causal Effect of PFAS on Duration
                                                      Conditionals
Model                Instrument  Valid  Age  Alcohol  BMI  Education  Employment  Parity  Smoking
Timm17HealthBehavior Age         Yes                           *                    *
```

Example 34.6: Distinguishing Models with Data

This example illustrates how you can use the CAUSALGRAPH procedure to examine the statistical implications of the assumptions that are encoded in a causal model. When you have multiple plausible models that might describe the same data set, you can use these statistical implications to decide which model best represents the data generating process.

Two plausible causal models for describing an individual’s serum urate and risk of cardiovascular disease are shown in Figure 34.9. The model are derived from a larger model that was developed by Thornley et al. (2013). See Example 34.3 for a summary of the variables in the models. In one of the two models, blood pressure and antihypertensive medication use are assumed to mediate the effect of the variable Urate on the variable CVD. In the other model, the causal direction is reversed, and the variable AntiHypertensiveUse is assumed to exert a causal effect on Urate. Differences between the two models are highlighted in red. The treatment and outcome variables are shaded in both models. For this example, the variable Nutrition corresponds to a latent construct and thus is not measured or observed. It is also assumed that the variable PreviousBP is not measured.

Figure 34.9 Two Possible Causal Models of the Effect of Serum Urate on Risk of Cardiovascular Disease



When you have multiple possible causal models, sometimes there is a common adjustment set that is valid for all models. In this case, you can use the adjustment technique to estimate the causal effect from the data, no matter which model best represents the true data generating process. See Example 34.4 for an illustration. The following statements invoke PROC CAUSALGRAPH to construct a common adjustment set:

```
proc causalgraph common(only);
   model "Thor12SimpleBP"
      AntiHypertensiveUse ==> CurrentBP,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      Nutrition ==> Urate Obesity,
      Obesity ==> PreviousBP Creatinine,
      CurrentBP ==> CVD,
      PreviousBP ==> AntiHypertensiveUse,
      Urate ==> PreviousBP Creatinine CVD;
   model "Thor12AltBP"
      AntiHypertensiveUse ==> CurrentBP Urate,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      Nutrition ==> Urate Obesity,
      Obesity ==> PreviousBP Creatinine,
      CurrentBP ==> CVD,
      PreviousBP ==> AntiHypertensiveUse,
      Urate ==> CVD;
   identify Urate ==> CVD;
   unmeasured Nutrition PreviousBP;
run;
```

When you analyze multiple causal models, you must use a separate MODEL statement for each causal model. Each MODEL statement must begin with a quoted string that provides a unique name for the model. In this example, you specify the two models shown in Figure 34.9 in the two model statements, respectively, by using the labels **Thor12SimpleBP** and **Thor12AltBP**.

In the IDENTIFY statement, you specify the causal effect of interest. In this example, you are investigating the causal effect of Urate on CVD.

The COMMON(ONLY) option in the PROC CAUSALGRAPH statement lists all adjustment sets that are common to all the models that you specify using MODEL statements. In this example, the procedure produces output only for the common adjustments. You can use the COMMON option without the (ONLY) suboption to produce separate adjustment sets for each model, in addition to the common adjustment sets.

The note in Output 34.6.1 shows that, in this example, it is not possible to find a common adjustment set. Thus, either you must find separate adjustment sets, then estimate the causal effect separately by using each model, or you must determine which model best represents the data generating process. The IMAP option provides analyses of model properties that might be useful for the latter action.

Output 34.6.1 Common Adjustment Note

```
NOTE: There are no adjustment sets common to all models that satisfy the specified criteria.
```

The causal assumptions that are encoded in a graphical model have implications for the statistical properties of the data generating process. You can test the implied statistical properties against the available data (subject to the usual limitations of sampling error and hypothesis testing). If the property does not hold in the data, then you should consider revising or discarding the model. For more information about the statistical implications of a graphical model, see the section "Statistical Properties of Causal Models" on page 2264. If you have multiple models, then you can compare the statistical implications of the models to find a property that is implied in one model but not the others. You can then test this property in the data and use the corresponding test results to determine which model best represents the true data generating process.

The following statements invoke the procedure to enumerate these statistical properties for both models in this example. For each model, the table of conditional independence properties that the procedure produces is saved to a data set by the ODS OUTPUT statement.

```
proc causalgraph imap=global;
   model "Thor12SimpleBP"
      AntiHypertensiveUse ==> CurrentBP,
      Creatinine ==> AntiHypertensiveUse CurrentBP,
      Nutrition ==> Urate Obesity,
      Obesity ==> PreviousBP Creatinine,
      CurrentBP ==> CVD,
      PreviousBP ==> AntiHypertensiveUse,
      Urate ==> PreviousBP Creatinine CVD;
   identify Urate ==> CVD;
   unmeasured Nutrition PreviousBP;
```

```
      ods output Imap = SimpleBPIndep;
   run;

   proc causalgraph imap=global;
      model "Thor12AltBP"
         AntiHypertensiveUse ==> CurrentBP Urate,
         Creatinine ==> AntiHypertensiveUse CurrentBP,
         Nutrition ==> Urate Obesity,
         Obesity ==> PreviousBP Creatinine,
         CurrentBP ==> CVD,
         PreviousBP ==> AntiHypertensiveUse,
         Urate ==> CVD;
      identify Urate ==> CVD;
      unmeasured Nutrition PreviousBP;
      ods output Imap = AltBPIndep;
   run;
```

The IMAP=GLOBAL option in each invocation of the procedure produces a table of global Markov properties for every model that you specify using a MODEL statement. Each global Markov property consists of two variables that are statistically independent conditional on another (possibly empty; in this case, the independence is unconditional) set of variables. For more information about the global Markov properties that are associated with a model, see the section "Statistical Properties of Causal Models" on page 2264. If every variable in a Markov property is observed, then you can perform statistical tests (for example, you can test for zero partial correlation) by using the available data to see whether the property can be falsified. Independence properties that involve one or more unmeasured variables cannot be tested.

The following code prints the first 10 observed conditional independence properties for each model. This example focuses on the first 10 statements for the sake of brevity. In practice, you perform the following analysis on the complete sets of independence properties.

```
   proc print data=SimpleBPIndep(obs=10);
      var Set1 Set2 CondSet;
      where Observable = 1;
   run;

   proc print data=AltBPIndep(obs=10);
      var Set1 Set2 CondSet;
      where Observable = 1;
   run;
```

The first 10 observed conditional independence properties for the two models are displayed in Output 34.6.2 and Output 34.6.3. You can now compare the two models. For both models, there are four conditioning sets such that the variable AntiHypertensiveUse is conditionally independent of the variable CVD, and these four conditioning sets are the same for both models. Thus, you cannot distinguish between the two models by testing for conditional independence between AntiHypertensiveUse and CVD. Next, you compare the conditional independence properties for the variables Creatinine and CVD. There are four such properties for the model Thor12SimpleBP, but there are five such properties for the model **Thor12AltBP**. The model **Thor12AltBP** encodes the statistical implication that Creatinine and CVD are independent conditional on the set (AntiHypertensiveUse, CurrentBP, Obesity), but this implication is not encoded in **Thor12SimpleBP**.

This means that if you were to find a nonzero partial correlation between Creatinine and CVD (after partialing out AntiHypertensiveUse, CurrentBP, and Obesity), you would have evidence to reject **Thor12AltBP**. You can continue this analysis for every independence property that is unique to one of the two models. The best model is the one whose conditional independence properties most closely match the zero partial correlations in the available data.

Output 34.6.2 Conditional Independence Properties of `Thor12SimpleBP`

Obs	Set1	Set2	CondSet
1	AntiHypertensiveUse	CVD	CurrentBP Urate
3	AntiHypertensiveUse	CVD	CurrentBP Obesity Urate
9	AntiHypertensiveUse	CVD	Creatinine CurrentBP Urate
12	AntiHypertensiveUse	CVD	Creatinine CurrentBP Obesity Urate
73	Creatinine	CVD	CurrentBP Urate
75	Creatinine	CVD	CurrentBP Obesity Urate
81	Creatinine	CVD	AntiHypertensiveUse CurrentBP Urate
83	Creatinine	CVD	AntiHypertensiveUse CurrentBP Obesity Urate
141	CurrentBP	Obesity	AntiHypertensiveUse Creatinine
142	CurrentBP	Obesity	AntiHypertensiveUse Creatinine Urate

Output 34.6.3 Conditional Independence Properties of `Thor12AltBP`

Obs	Set1	Set2	CondSet
1	AntiHypertensiveUse	CVD	CurrentBP Urate
3	AntiHypertensiveUse	CVD	CurrentBP Obesity Urate
9	AntiHypertensiveUse	CVD	Creatinine CurrentBP Urate
11	AntiHypertensiveUse	CVD	Creatinine CurrentBP Obesity Urate
37	Creatinine	CVD	CurrentBP Urate
39	Creatinine	CVD	CurrentBP Obesity Urate
45	Creatinine	CVD	AntiHypertensiveUse CurrentBP Urate
47	Creatinine	CVD	AntiHypertensiveUse CurrentBP Obesity
48	Creatinine	CVD	AntiHypertensiveUse CurrentBP Obesity Urate
79	Creatinine	Urate	AntiHypertensiveUse Obesity

Example 34.7: Applying an Adjustment Set to Estimate a Causal Effect from Data

This example illustrates how you can use the CAUSALGRAPH procedure to estimate the magnitude of a causal effect that has a valid causal interpretation. To compute such an estimate from a data set, you can use the following general approach:

1. Carefully consider the data generating process, and create a list of causal assumptions that accurately represents that process. Encode these assumptions in a graphical causal model. For more information about causal assumptions and graphical models, see the section "Causal Graph Theory" on page 2261.

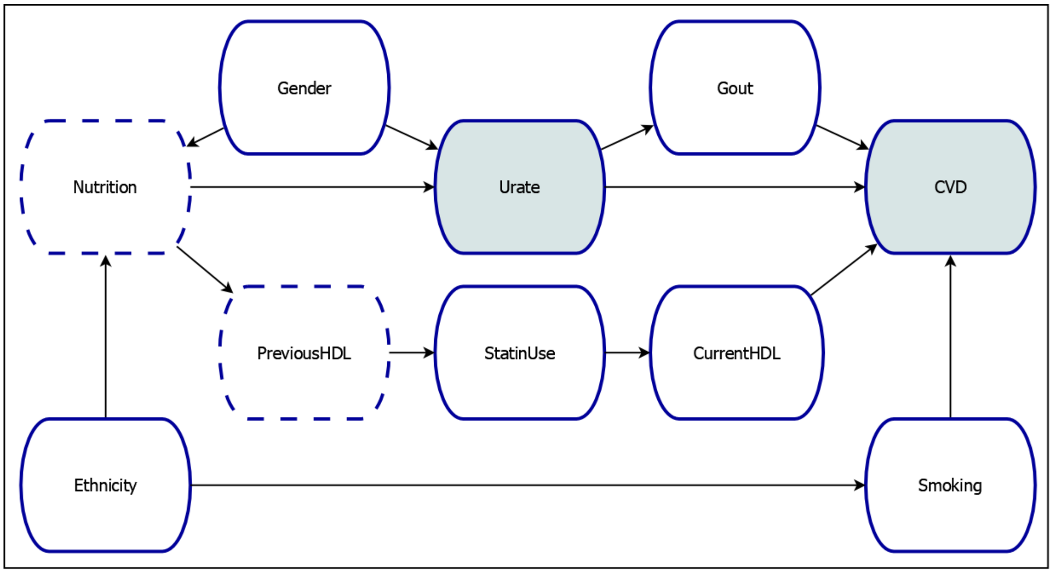
2. Use this graphical model to find a valid identification strategy, such as an adjustment set.

3. Use the identification results to construct an estimator, such as the stratification estimator.

In most practical situations, the true data generating process is not known exactly. In such situations, you must define a causal model that is assumed to represent the data generating process. To construct this causal model, you might rely on expert opinion, established scientific theory, prior experience, or any other source of substantive knowledge. This example uses simulated data, so that the data generating process is treated as known and the impact of identification and adjustment on causal effect estimation can be illustrated. For more information about identification strategies and adjustment sets, see the section “Identification and Adjustment” on page 2265.

A causal model that relates an individual’s serum urate to the risk of cardiovascular disease is shown in Figure 34.10. This model is derived from a larger model that was developed by Thornley et al. (2013). See Example 34.3 for a summary of the variables in the figure. The treatment and outcome variables are shaded in the model. For this example, the variable Nutrition corresponds to a latent construct and thus is not measured or observed. It is also assumed that the variable PreviousHDL is not measured.

Figure 34.10 Causal Model of the Effect of Serum Urate on Risk of Cardiovascular Disease



The following DATA step creates a simulated data set that is consistent with the model in Figure 34.10. Thus, it also defines the true data generating process.

```
data CVDdata;
   drop ii Nutrition PreviousHDL;
   call streaminit(1000);
   array EthProb[6] _temporary_ (0.60, 0.18, 0.13, 0.05, 0.01, 0.03);
   array SmokeRates[6] _temporary_ (0.17, 0.10, 0.17, 0.07, 0.22, 0.15);
   array EthNut[6] _temporary_ (0.20, 0.18, 0.08, 0.03, 0.11, 0.04);
   do ii = 1 to 79000;
      Gender = rand("Bernoulli" , .5);
      Ethnicity = rand("Table" , of EthProb[*]);
      Smoking = rand("Bernoulli", SmokeRates[Ethnicity]);
      Nutrition = 0.5 - Gender + 10.0*rand("Normal", 0, EthNut[Ethnicity]);
      PreviousHDL = 55 + 4.0*Nutrition;
      if      PreviousHDL<40 then StatinUse = rand("Bernoulli", 0.90);
      else if PreviousHDL<60 then StatinUse = rand("Bernoulli", 0.65);
      else                        StatinUse = rand("Bernoulli", 0.05);
      CurrentHDL = 55 + rand("Normal",0.0,7) + StatinUse*rand("Normal",4.0,0.5);
      Urate = 6.0 + 0.4*Nutrition + 1.5*Gender;
      Gout = rand("Bernoulli", logistic(-8.0 + 0.90*Urate));
      CVD = rand("Bernoulli", logistic(-1.2 - 0.04*CurrentHDL + 0.2*Gout +
         0.65*Smoking + 0.1*Urate));
      output;
   end;
run;
```

The first 10 lines of the simulated data set are shown in Output 34.7.1.

Output 34.7.1 First 10 Lines of the Simulated Data Set

Obs	Gender	Ethnicity	Smoking	StatinUse	CurrentHDL	Urate	Gout	CVD
1	Female	Hispanic	No	0	45.9241	6.41547	1	0
2	Male	Hispanic	No	1	66.9528	7.43840	0	0
3	Female	NativeAmer	No	1	54.5367	6.42775	0	0
4	Female	Hispanic	No	1	67.7733	5.32636	0	0
5	Female	WhiteNonHisp	No	0	64.2067	6.67543	0	0
6	Male	WhiteNonHisp	No	1	53.0462	7.22069	1	0
7	Female	AfricanAmer	Yes	1	49.4850	5.15609	0	0
8	Male	Hispanic	Yes	0	57.8065	6.47476	0	0
9	Female	WhiteNonHisp	No	0	57.7600	6.52005	0	1
10	Male	AfricanAmer	No	1	71.0460	7.30762	0	0

The following code uses the simulated data to create a table of summary statistics for the variable Urate:

```
proc means data=CVDdata;
   var Urate;
   ods output Summary=SampleMeansOutput;
run;
```

The summary statistics are shown in Output 34.7.2. You use the ODS OUTPUT statement to store the summary statistics for Urate in an output data set. You use this information later in the analysis to define the treatment and control levels for the causal effect of interest.

Output 34.7.2 Summary Statistics for Urate

Analysis Variable : Urate				
N	**Mean**	**Std Dev**	**Minimum**	**Maximum**
79000	6.7460150	0.8936312	3.0665877	10.3763707

In this example, the treatment or exposure variable Urate is continuous. Moreover, the effect of this variable on the mediator variable Gout and the outcome variable (CVD) is nonlinear. Because there are no natural treatment and control levels for Urate, you must somehow define the causal effect of interest. A common causal effect measure is the average treatment effect (ATE) or the expected risk difference, which is the difference in the expected potential outcome values between well-defined treatment and control conditions or levels. For more information about the definitions of causal effect measures in the potential outcome framework, see the section "Causal Effects: Definitions, Assumptions, and Identification" on page 2394 in Chapter 36, "The CAUSALTRT Procedure."

In this example, you consider the causal effect of interest to be the expected risk difference in CVD that is associated with a change in Urate from a control condition to a treatment condition. Two possibilities for defining the control and treatment conditions are considered here. In this way, you can explore how the magnitude of the causal effect depends on the values of the treatment variable that are being considered.

First, you consider the causal effect of a unit change in Urate, centered around the population mean. Then, in the potential outcomes notation, the causal effect of interest is the expected risk difference

$$\text{UnitEff} = E[\text{CVD}(\text{Urate} = \mu + 0.5)] - E[\text{CVD}(\text{Urate} = \mu - 0.5)]$$

where μ is the population mean of Urate. In this causal effect definition, the control condition is defined as a half unit below the population mean of Urate, and the treatment condition is defined as a half unit above the population mean of Urate.

Second, you consider the causal effect of a change of one standard deviation in Urate, also centered around the population mean. The causal effect is now defined as the expected risk difference

$$\text{StdEff} = E[\text{CVD}(\text{Urate} = \mu + 0.5\sigma)] - E[\text{CVD}(\text{Urate} = \mu - 0.5\sigma)]$$

where σ is the population standard deviation of Urate.

For demonstration purposes, the two population causal effects are computed by generating a large number of potential outcomes (1,000,000,000 replications) according to the true data generating process mentioned earlier. By this method, the population effect, UnitEff, is found to be 0.0076, and the standardized population effect, StdEff, is found to be 0.0068. These values are the target causal effects that you are estimating from the random sample. Now the most interesting questions are whether and how the CAUSALGRAPH procedure can help find a statistical strategy that provides accurate estimates of these target causal effects.

Given the causal model for the data, you can use the procedure to analyze the identifiability of the causal effect of Urate on CVD. The following code uses the procedure to list valid adjustment sets that can be used to identify this causal effect. For brevity, you can use the MAXSIZE=2 option to construct only those adjustment sets that have no more than two elements.

```
proc causalgraph maxsize=2;
   model "Thor12SimpleHDL"
      Ethnicity ==> Nutrition Smoking,
      Gender ==> Nutrition Urate,
      Gout ==> CVD,
      Nutrition ==> PreviousHDL Urate,
      CurrentHDL ==> CVD,
      PreviousHDL ==> StatinUse,
      Smoking ==> CVD,
      StatinUse ==> CurrentHDL,
      Urate ==> CVD Gout;
   identify Urate ==> CVD;
   unmeasured Nutrition PreviousHDL;
run;
```

In the MODEL statement, you specify the causal model to be analyzed. The quoted string in the statement labels the model. The remainder of the MODEL statement specifies all the variables and edges in the model. These variables and edges reflect the data generating process shown in Figure 34.10.

In the IDENTIFY statement, you specify that the causal effect of interest is the effect of Urate on CVD. In the UNMEASURED statement, you specify that the variables Nutrition and PreviousHDL are not measured or observed.

The list of adjustment sets that the procedure produces is shown in Output 34.7.3. Notice that the empty set does not appear in this list. This means that the marginal association between Urate and CVD cannot be used to estimate a causal effect that has a valid causal interpretation. Rather, you must use an alternative estimation strategy, such as estimation by adjustment that uses one of the adjustment sets in Output 34.7.3. As is shown later in this example, the failure to perform such an adjustment results in biased estimation of the causal effects.

Output 34.7.3 Possible Adjustment Sets for the Model in Figure 34.10

Covariate Adjustment Sets for Thor12SimpleHDL

Causal Effect of Urate on CVD

			Covariates					
	Size	**Minimal**	**CurrentHDL**	**Ethnicity**	**Gender**	**Gout**	**Smoking**	**StatinUse**
1	2	Yes	*	*				
2	2	Yes	*				*	
3	2	Yes		*				*
4	2	Yes					*	*

You can use any of the adjustments sets in Output 34.7.3 to obtain an estimate for the effect of Urate on CVD that has a valid causal interpretation. In particular, the set {Smoking, StatinUse} is a valid adjustment set. This set also has the useful property that both variables in the set are binary classification variables. Thus one possible way to estimate the causal effect is to stratify the analysis by the levels of these variables.

As mentioned earlier, two causal effects are being estimated. One is the unstandardized unit effect of Urate on CVD, denoted as UnitEff. The other is the standardized unit effect of Urate on CVD, denoted as StdEff. Both of these causal effects are defined in terms of the differences in expected CVD potential outcome values. These potential outcomes are evaluated at some Urate treatment and control levels that are defined in terms of population parameters. Because these population parameters and hence the treatment and control levels are not known, you need to estimate them from the sample. The following code computes two sets of sample values for the treatment and control levels of Urate from the table of summary statistics that you created earlier in this example. These computed values are stored in the data set ScoreData that you will use to estimate the two causal effects of interest.

```
data _null_;
   set SampleMeansOutput;
   call symputx("UrateMean",Urate_Mean);
   call symputx("UrateStd", Urate_StdDev);
   call symputx("UrateUnit1", Urate_Mean + 0.5);
   call symputx("UrateUnit0", Urate_Mean - 0.5);
   call symputx("UrateStd1", Urate_Mean + 0.5*Urate_StdDev);
   call symputx("UrateStd0", Urate_Mean - 0.5*Urate_StdDev);
run;

data ScoreData;
   set SampleMeansOutput;
   keep Urate Test;

   Test  = "UnitTreat  ";
   Urate = &UrateUnit1;
   output;
```

```
      Test  = "UnitControl";
      Urate = &UrateUnit0;
      output;

      Test  = "StdTreat   ";
      Urate = &UrateStd1;
      output;

      Test  = "StdControl ";
      Urate = &UrateStd0;
      output;
   run;
```

Now, the following code performs logistic regression analyses that are stratified by levels of the two adjustment variables that are suggested by the results of PROC CAUSALGRAPH:

```
   proc sort data=CVDdata;
      by Smoking StatinUse;
   run;
   proc logistic data=CVDdata noprint;
      by Smoking StatinUse;
      model CVD(event='1') = Urate;
      score data=ScoreData out=ProbStrat;
   run;
```

The MODEL statement specifies the outcome variable CVD and the treatment variable Urate. The effect of Urate on CVD is estimated within each of the four strata separately. However, to estimate the aforementioned population causal effects of interest, you need to apply the logistic regression results to estimate the expected CVD values under the specific Urate treatment and control levels of interest. You can accomplish this by using the SCORE statement.

The SCORE statement estimates the probability of a CVD event (that is, the probability that CVD=1), which is the expected value of CVD, for each value of Urate that is specified in the data set ScoreData. In this example, these values of Urate correspond to the treatment and control levels for defining the causal effects of interest. Hence, the SCORE statement estimates the expected CVD values under the required treatment and control levels in each stratum. The OUT= option in the SCORE statement saves the expected CVD values in the data set ProbStrat. These expected CVD values are shown in the column P_1 in Output 34.7.4.

Output 34.7.4 Posterior Probabilities for Each Strata

Obs	Smoking	StatinUse	Test	Urate	P_1
1	0	0	UnitTreat	7.24602	0.06929
2	0	0	UnitControl	6.24602	0.05994
3	0	0	StdTreat	7.19283	0.06876
4	0	0	StdControl	6.29920	0.06041
5	0	1	UnitTreat	7.24602	0.05829
6	0	1	UnitControl	6.24602	0.05489
7	0	1	StdTreat	7.19283	0.05811
8	0	1	StdControl	6.29920	0.05506
9	1	0	UnitTreat	7.24602	0.12451
10	1	0	UnitControl	6.24602	0.10476
11	1	0	StdTreat	7.19283	0.12338
12	1	0	StdControl	6.29920	0.10574
13	1	1	UnitTreat	7.24602	0.11226
14	1	1	UnitControl	6.24602	0.09918
15	1	1	StdTreat	7.19283	0.11153
16	1	1	StdControl	6.29920	0.09984

The two binary adjustment variables result in four strata for analysis. Within each stratum, you can compute the unstandardized unit effect by the difference in P_1 between UnitTreat and UnitControl, and you can compute the standardized effect by the difference in P_1 between StdTreat and StdControl. However, none of these effects within strata are the causal effect estimates themselves. The estimate of the causal effect UnitEff must be computed by using the weighted average of the difference in P_1 between UnitTreat and UnitControl in the strata, where the weights are the sample sizes of the strata. Similarly, the estimate of the causal effect StdEff must be computed by using the weighted average of the difference in P_1 between StdTreat and StdControl in the strata. These estimates of the causal effects are shown in the column Stratified Estimation in Output 34.7.6.

As discussed previously, if you attempt to estimate the effect of Urate on CVD by using the marginal association between the two variables (that is, no adjustment), confounding covariates bias the estimation results. Strictly for the purpose of demonstrating such a biased result, the following PROC LOGISTIC code performs a logistic regression that is not stratified by any covariate:

```
proc logistic data=CVDdata noprint;
   model CVD(event='1') = Urate;
   score data=ScoreData out=ProbNaive;
run;
```

Just as before, the SCORE statement in PROC LOGISTIC estimates the expected CVD values under the required Urate treatment and control levels. The corresponding estimates of the expected CVD values are shown in Output 34.7.5. As in the stratified estimation, the target causal effects are estimated by computing the related differences in P_1, except that in this case no weighted average is computed for the differences.

Output 34.7.5 Unadjusted Posterior Probabilities

Obs	Test	Urate	P_1
1	UnitTreat	7.24602	0.073217
2	UnitControl	6.24602	0.064062
3	StdTreat	7.19283	0.072702
4	StdControl	6.29920	0.064521

Now, you have two sets of estimation results. One set of results is computed by using a stratified estimator that is based on an adjustment strategy. The other set of results is computed by using the naive marginal association between the treatment and outcome variables. The estimates of the causal effects UnitEff and StdEff that are computed using these two estimators are shown in Output 34.7.6. For comparison, the target true values of the causal effects are also shown.

Output 34.7.6 Causal Effect Estimation Summary

		True	Stratified	Unadjusted
Obs	Effect	Effect	Estimation	Estimation
1	UnitEff	0.007620	0.007766	0.009155
2	StdEff	0.006789	0.006940	0.008181

The estimates that are computed by using the stratified estimator very closely approximate the true values. This is expected, because the set {Smoking,StatinUse} is a valid adjustment set for the data generating process shown in Figure 34.10. However, the estimates that are computed by using the naive unadjusted logistic regression results do not agree with the true values. This is also expected, because the PROC CAUSALGRAPH analysis in Output 34.7.3 shows that the empty set is not a valid adjustment. Thus, this example demonstrates the usefulness of causal graph theory to identify causal effects in confounding situations. By devising a stratified estimator of the causal effects, this example also demonstrates how to implement a good statistical estimation strategy based on the identification results from PROC CAUSALGRAPH.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). “Identification of Causal Effects Using Instrumental Variables.” *Journal of the American Statistical Association* 91:444–455.

Elwert, F. (2013). “Graphical Causal Models.” In *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan, 245–273. Dordrecht: Springer.

Elwert, F., and Winship, C. (2014). “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.” *Annual Review of Sociology* 40:31–53.

Galles, D., and Pearl, J. (1998). “An Axiomatic Characterization of Causal Counterfactuals.” *Foundations of Science* 3:151–182.

Geiger, D., and Pearl, J. (1988). “On the Logic of Causal Models.” In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, edited by R. D. Shacter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, 136–147. Amsterdam: North-Holland.

Greenland, S., Pearl, J., and Robins, J. M. (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology* 10:37–48.

Imbens, G. W. (2014). "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29:323–358.

Koller, D., and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9." *Statistical Science* 5:465–472. Translated and edited by Dabrowska and Speed from the Polish original by Neyman (1923).

Pearl, J. (1993). "Comment: Graphical Models, Causality and Intervention." *Statistical Science* 8:266–269.

Pearl, J. (2009a). "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3:96–146.

Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Pearl, J. (2010). "An Introduction to Causal Inference." *International Journal of Biostatistics* 6:1–62.

Pearl, J. (2012). "The Causal Foundations of Structural Equation Modeling." In *Handbook of Structural Equation Modeling*, edited by R. H. Hoyle, 68–91. New York: Guilford Press.

Pearl, J., and Verma, T. (1987). "The Logic of Representing Dependencies by Directed Graphs." In *Proceedings of the Sixth National Conference on Artificial Intelligence*, 374–379. AAAI Press.

Perković, E., Textor, J., Kalisch, M., and Maathuis, M. (2018). "Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs." *Journal of Machine Learning Research* 18:1–62.

Rubin, D. B. (1980). "Comment on D. Basu, 'Randomization Analysis of Experimental Data: The Fisher Randomization Test'." *Journal of the American Statistical Association* 75:591–593.

Rubin, D. B. (1990). "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5:472–480.

Schafer, J. L., and Kang, J. (2008). "Average Causal Effects from Nonrandomized Studies: A Practical Guide and Simulated Example." *Psychological Methods* 13:279–313.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2010). "On the Validity of Covariate Adjustment for Estimating Causal Effects." In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, edited by P. Grünwald and P. Spirtes, 527–536. Corvallis, OR: AUAI Press.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. 2nd ed. Cambridge, MA: MIT Press.

Takata, K. (2010). "Space-Optimal, Backtracking Algorithms to List the Minimal Vertex Separators of a Graph." *Discrete Applied Mathematics* 158:1660–1667.

Thornley, S., Marshall, R. J., Jackson, R., Gentles, D., Dalbeth, N., Crengle, S., Kerr, A., and Wells, S. (2013). "Is Serum Urate Causally Associated with Incident Cardiovascular Disease?" *Rheumatology* 52:135–142.

Timmermann, C. A. G., Budtz-Jørgensen, E., Petersen, M. S., Weihe, P., Steuerwald, U., Nielsen, F., Jensen, T. K., and Grandjean, P. (2017). "Shorter Duration of Breastfeeding at Elevated Exposures to Perfluoroalkyl Substances." *Reproductive Toxicology* 68:164–170.

Van der Zander, B., Liśkiewicz, M., and Textor, J. (2014). "Constructing Separators and Adjustment Sets in Ancestral Graphs." In *Proceedings of the Thirtieth Conference on Causal Inference: Learning and Prediction*, edited by N. L. Zhang and J. Tian, 11–24. Corvallis, OR: AUAI Press.

Van der Zander, B., Textor, J., and Liśkiewicz, M. (2015). "Efficiently Finding Conditional Instruments for Causal Inference." In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, edited by Q. Yang and M. Wooldridge, 3242–3249. Palo Alto, CA: AAAI Press.

Subject Index

Syntax Index