My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Your R. Name

May 200x

Approved for the Division
(Mathematics)

Advisor F. Name

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class.

List of Abbreviations

Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

Background

1.1 Graphs, D-separation, Causality

The idea of cause and effect has been studied and discussed in philosophy for centuries, but the formalization of causality in mathematics, statistics, and computer science is much more recent. One framework in particular, Judea Pearl's Structural Causal Models (SCMs), (Pearl, 2009) is flexible, widely used, and mathematically elegant. However, before we can give SCMs a serious treatment, it is helpful to introduce definitions for the graphical machinery that most analysis of SCMs rely on.

1.1.1 Directed Acylic Graphs

Definition 1.1.1. *A* **graph** G *is a pair* $G = (V, E)$ *where* V *denotes a set of nodes (sometimes called vertices) and* $E \subseteq \{(i, j)|i, j \in V\}$ *denotes a set of edges between nodes.*

Conventions differ, but for our purposes, an ordered pair $(i, j) \in E$ is considered to be a **edge**, read as "an edge from node i to node j". For easy reading, we write $i \to j$ whenever $(i, j) \in E$. In the figures that frequently accompany graphs, an edge (i, j) is depicted as an arrow from node i to node j.

Often times we are interested in how different nodes in a particular graph are or are not connected to one another. In this spirit we define paths.

Definition 1.1.2. *A* **path** p *between nodes* $v_1, v_n \in V$ *is a sequence* $v_1, v_2, ..., v_n \in V$ *of distinct nodes such that either* $(v_i, v_{i+1}) \in E$ *or* $(v_i, v_{i+1}) \in E$ *edge always exists between* v_i *and* v_{i+1}*. When* $v_i \to v_{i+1}$ *(*$(i, i-1) \in E$*) for all* $i \in \{1, 2, ..., n\}$*, we say* p *is a* **directed path**.

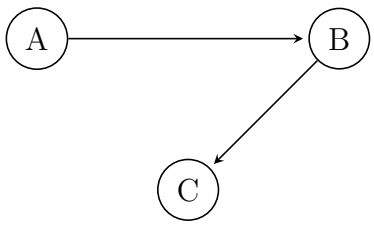


Figure 1.1: Graphs corresponding to $V = \{A, B, C\}$, $E = \{(A, B), (B, C)\}$

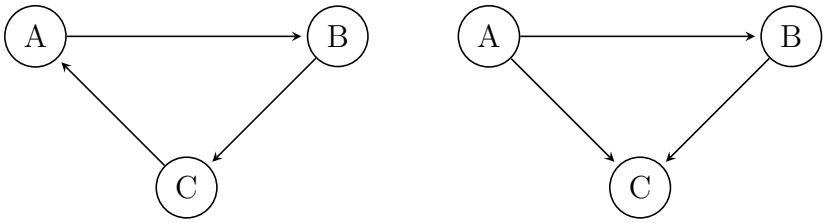


Figure 1.2: Cyclic (left) and acyclic (right) graphs on the nodes A, B, C

For a directed graph G and a particular node $v \in V$, we define several sets of related nodes. The **parents** of v, denoted PA_v, are the nodes with a directed edge ending at v, $\{j \in V | j \to v\}$ and similarly the **children**, denoted CH_v, of v are those nodes which v has a directed edge to, $\{j \in V | v \to j\}$.

Extending these definitions, we define the **ancestors** of v, AN_v, as all the nodes from which a directed path to v exists, $\{j \in V | j \to ... \to v\}$. Likewise we define **descendants** DE_v of v as the nodes for which a directed path from v exists, $\{j \in V | v \to ... \to j\}$ [ref Peters chp 6.1]. With these in place, we progress to the next definition.

Definition 1.1.3. *A directed graph $G = (V, E)$ is said to be an* **directed acyclic graph** *(DAG) if for all nodes $v \in V$, $DE_v \cap AN_v = \emptyset$. That is, a directed graph is a DAG when there is never a directed path from a node to itself.*

Now that we have defined DAGs, we can advance to a relevant application, Bayesian networks.

1.1.2 Conditional Independence and Bayesian Networks

The idea of conditional independence is at the basis of graphical approaches to causality. Intuitively, conditional independence statements tell us what we can *ignore*. When $X \perp\!\!\!\perp Y | Z$, the conditional probability $P(x|y, z)$ can be replaced with $P(x|z)$, i.e. we can ignore the value of y as long as we know z. More formally:

Definition 1.1.4 (Conditional Independence)**.** *Two sets of random variables $\mathbf{X}$ and $\mathbf{Y}$ are said to be* **conditionally independent** *given a third set $\mathbf{Z}$ when $P(\mathbf{x}, \mathbf{y}|\mathbf{z}) = P(\mathbf{x}|\mathbf{z})P(\mathbf{y}|\mathbf{z})$. Alternatively, this condition can be formulated as $P(\mathbf{x}|\mathbf{y}, \mathbf{z}) = P(\mathbf{x}|\mathbf{z})$. We symbolize this statement as $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$.*

This is useful because it allows us to dramatically simplify many probabilistic expressions, making them easier to evaluate. Bayesian networks are a popular way of encoding a particular "factorization" of a joint distribution. This requires a quick review of the chain rule for probability, an extension of the definition of conditional probability that allows joint distribution to be factored into the product of conditionals.

Definition 1.1.5 (Chain Rule)**.** *A joint distribution $P(x_1, x_2, ..., x_n)$ may be factored as $P(x_1, x_2, ..., x_n) = P(x_1|x_2, ..., x_n)P(x_2|x_3, ...x_n) \cdots P(x_{n-1}|x_n)P(x_n)$.*

Since the variables may be ordered in any way, this allows for many distinct factorizations. So, for any particular factorization, conditional independences allow it to be simplified through the removal of extraneous conditioning variables.

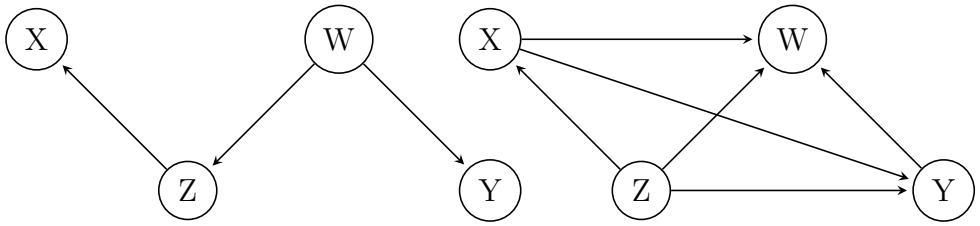


Figure 1.3: Bayesian networks corresponding to $P(x|z)P(y|w)P(z|w)P(w)$ (left) and $P(w|x,y,z)P(y|x,z)P(x|z)P(z)$ (right)

Example 1. *For a joint distribution $P(x,y,z,w)$ with the independences $X \perp\!\!\!\perp (Y,W)|Z$ and $Y \perp\!\!\!\perp W|Z$, the chain rule and conditional independence allows us to expression the joint as follows:*

$$P(x,y,z,w) = P(x|y,z,w)P(y|z,w)P(z|w)P(w) = P(x|z)P(y|w)P(z|w)P(w)$$

Notice that if we had started with a different factorization, such as $P(x,y,z,w) = P(w|x,y,z)P(y|x,z)P(x|z)P(z)$ the conditional independences do not allow us to further simplify the expression.

For a particular (simplified) factorization, the associated Bayesian network is the DAG formed by drawing an edge from each variable being conditioned on to the variable on the left of the conditioning bar. For our two factorizations, the associated Bayesian networks are the DAGs displayed in figure 1.4.

This process is reversible, and given a particular Bayesian network we may "read off" the associated joint distribution as $P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i|PA_{x_i})$ (Pearl, 2009). However, the real usefulness of Bayesian networks comes from way that further conditional independence relationships are encoded graphically.

1.1.3 D-Separation

Graphical structure can allow us to reason about causal and probabilistic relationships. One of the essential tools is d-separation. Graphically, d-separation statements formalize the idea that two sets of nodes can be "separated" from each other by a third set. To develop the intuition, consider the very basic DAG $A \to B \to C$. There is exactly one directed path from A to C, and the path passes through B. So to get to C from A you need to pass through B. In this way, we say that B blocks the directed path from A to C. This notion of blocking is extended to form the definition of d-separation (**?**):.

Definition 1.1.6. *Given a directed graph G, a path p from node v_1 to node v_n is said to be* **blocked** *by a set S (with $S \cap \{v_1, v_n\} = \emptyset$) if:*

1. *$v_j \in S$ and p contains a* **chain***: $v_{j-1} \to v_j \to v_{j+1}$ or $v_{j-1} \leftarrow v_j \leftarrow v_{j+1}$, or*

2. *$v_j \in S$ and p contains a* **fork***: $v_{j-1} \leftarrow v_j \to v_{j+1}$, or*

3. *$v_j \notin S$ and $DE_{v_j} \cap S = \emptyset$ and p contains a* **collider***: $v_{j-1} \to v_j \leftarrow v_{j+1}$. As we will see, colliders play an especially important role in the analysis of causal DAGs.*

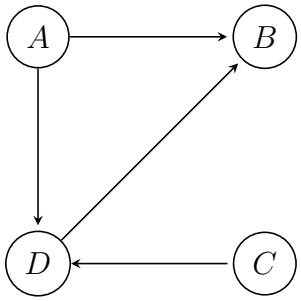
Example 2. *Consider a path $A \to B \to C \leftarrow D \to E$ between nodes A and E. There are many sets that block this path. For one, there is a collider on the path. This means that any set which does not include C or any elements of DE_C would successfully block the path. However, we could also pick $S = \{B\}$, $S = \{D\}$, or $S = \{B, D\}$ as blocking sets since the path contains a chain at B and a fork at D. Importantly, if we have either of these nodes in our blocking set, we could include C as well because the first blocking condition would be satisfied, even though the second would not. In fact, the only $S \subseteq \{B, C, D\}$ that would not block the path is $S = \{C\}$.*

With blocking defined, we address d-separation.

Definition 1.1.7. *For a DAG G, two sets of nodes $A, B \in G$ are said to be* **d-separated** *by a set S if all paths between A and B are blocked by S. We denote d-separation by the symbol $\perp\!\!\!\perp_G$, with the statement "In DAG G, A is d-separated from B by S" expressed as $A \perp\!\!\!\perp_G B|S$.*

Remark 1.1.1. *If $A \perp\!\!\!\perp_G B|S$ then $B \perp\!\!\!\perp_G A|S$*

Example 3. *Consider the DAG G below.*



Which nodes can be d-separated, and by which blocking sets? Nodes which are connected by an edge cannot be d-separated. To see why, take A and B. Although some paths between A and B can be blocked, the path $A \to B$ contains no nodes other than A and B and therefore cannot be blocked. Since there is no set that can block the path, A and B are not able to be d-separated. So consider A and C, two nodes which are not connected by an edge. There are two paths between A and C, $A \to B \leftarrow D \leftarrow C$ and $A \to D \leftarrow C$. Both of these paths include a collider. Of the four possible blocking sets, $\emptyset, \{B\}, \{D\}, \{B, D\}$, all but $\emptyset$ include a collider node for paths that are not blocked by other means. Then $A \perp\!\!\!\perp_G C|\emptyset$ and no other sets will work. What about B and C? Again there are two paths: $B \leftarrow D \leftarrow C$ and $B \leftarrow A \to D \leftarrow C$. The second path contains a collider, and therefore is blocked by $\emptyset$, but the first path cannot be. To block $B \leftarrow D \leftarrow C$ we must have D in our blocking set. So, $B \perp\!\!\!\perp_G C|D$, but we can also include A since doing so will not unblock a path with a collider. Then $B \perp\!\!\!\perp_G C|D, A$.

The reason that the d-separation symbol $\perp\!\!\!\perp_G$ is so similar to the symbol for conditional independence $\perp\!\!\!\perp$ is that in any Bayesian network, the set of d-separation statement correspond to a set of conditional independences in the joint distribution. This allows us to define

a notion of compatibility between particular Bayesian networks and some joint distributions over the same variables.

Definition 1.1.8. *For a Bayesian network G over variables $X_1, ..., X_n$ and a joint distribution $P^*(x_1, ..., x_n)$ we say that P^* is* **compatible** *with G if for every conditional d-separation statement $\mathbf{Y} \perp\!\!\!\perp_G \mathbf{Z}|\mathbf{W}$ for $\mathbf{Y}, \mathbf{Z}, \mathbf{W} \subset \{X_1, ..., X_n\}$ there is a corresponding conditional independence $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{W}$ in the distribution P^*.*

Importantly, this notion does not require that every conditional independence in P^* has a corresponding d-separation in the graph. For instance, a probability distribution in which every variable in unconditionally independent from every other variable is compatible with any Bayesian network over those variables. However, the only probability distribution compatible with the Bayesian network without any edges is that fully independent distribution. This notion of compatibility is useful for two reasons: it defines a class of distributions which are possible under a particular Bayesian network and it allows comparison between a suggested Bayesian network and available data which is alleged to conform to this network. This becomes particularly important when Bayesian networks represent a particular *causal* order, as described in the next subsection.

1.1.4 Structural Causal Models

Many questions in the natural and social sciences involve understanding the way that one set of factors (e.g. genes, medicines, social policies) influence others (e.g. disease risk, mortality, childhood poverty rates). In these cases, the language and of classical statistics is often inadequate. The tools of hypothesis testing, (linear) regression, interval estimation, etc. can tell us much about relationships in the data, but except in the highly constrained setting of randomized controlled trials, very little about cause and effect. These limitations are not just theoretical. Many longstanding controversies and debates within the sciences come down to disagreements over which factors associated with a particular outcome are causing the outcome.

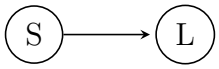
A paradigmatic example of such a controversy is over the relationship between cigarette smoking and lung cancer. Today, it is uncontroversial that smoking cigarettes (especially habitually over a long period of time) increases a person's risk of contracting lung cancer. This is a causal relationship, and we might express our belief in this relationship as a ***causal model*** in which a person's risk of contracting lung cancer is a *function* of a person's smoking habits (for simplicity, we pretend that smoking is the only such causal factor).

So, let L be a random variable denoting whether or not a person contracts lung cancer and S be a random variable denoting smoking. Under our causal model in which smoking is the only structural cause of lung cancer, we express L as a function of S:

$$L := f(S, N_L),$$
$$S := N_S.$$

Here, the use of $:=$ denotes an assignment rather that an algebraic relationship (**?**) and N_L is a noise term capturing random effects on lung cancer risk, and N_S a noise variable for cigarette smoking. Together, these two assignments are a structural causal model. Now that

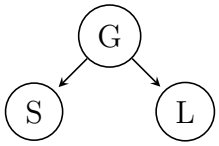
we have an SCM, we can express the structure as a DAG. Since L is assigned as a function of smoking but S is not caused by any factors in our model, we would express our model as a directed edge going from S to L:



However, historically, this causal structure was not the only one suggested to explain the association between lung cancer and smoking. One particularly illustrious detractor of this theory was British statistician R. A. Fisher, a foundational figure in modern statistics (and a heavy smoker) [ref Fisher]. In Fisher's view, there was no causal effect of smoking on lung cancer, rather, an underlying genetic factor was responsible for both a predisposition to smoking and a higher risk of developing lung cancer. So Fisher's causal model would have looked like:

$$L := f(G, N_L),$$
$$S := f(G, N_S),$$
$$G := N_G.$$

Where G is the genetic factor. As before, the structure of this model can be expressed as a graph:



In this way, every SCM entails a directed graph. Although it is entirely possible for an SCM to entail a graph that has cycles, we focus our attention on the case where the graph entailed is a DAG. Importantly, these two (graphical) models imply different observational joint distributions. Without having specified the functions, little can be said about the joint distribution entailed by the model, however, one important property is built in to every SCM.

Definition 1.1.9 (Markov property)**.** *In any SCM C with entailed graph G, if $A \perp\!\!\!\perp_G B|S \implies A \perp\!\!\!\perp B|S$ for all disjoint sets of nodes A, B, S. That is, every d-separation in the graph corresponds to a conditional independence in the joint distribution. This fact is implied by the definition of SCMs and is called the* **Markov Property**.

In other words, structural assignments (without cycles) induce a Bayesian network over the involved variables that the observational distribution must be compatible with. This result is a powerful tool that allows us to test how well differing causal models comport with the observational data. Returning to Fisher's suggestion that an underlying genetic cause was responsible for the association of smoking with lung cancer, consider that by the Markov property $S \perp\!\!\!\perp L|G$. Let's assume we have a population of identical twins, some of whom were smoking discordant. Since we know that identical twins are genetically identical we could assess whether or not smoking and lung cancer are independent within twin pairs. If not, then something must be wrong with the model.

1.2 Selection Bias and Missing Data

Most statistical inference techniques are designed for the case in which samples are 'representative' of the population. That is, for a finite population of size N, each unit in the population $X_1, ..., X_N$ are equally likely to appear in the sample. When the domain is continuous, we can extend this idea as saying that the distribution of a sampled value is the same as the "true" underlying distribution. However, this is often unattainable and therefore standard methods are inappropriate (see Chapter 4 for an overview of more appropriate methods). This problem is captured by two related phenomena: selection bias and missing data.

1.2.1 Selection Bias

Selection bias occurs whenever samples are non-representative. In the finite population case, this corresponds to some units being more likely to be sampled than others. The problem with this is serious. As a basic example, consider a sample $\tilde{\mathbf{X}}$ taken from $X_1, ..., X_N$ where each $X_i \sim \text{Bern}(p)$. If we want to estimate p from a representative sample, all we need is to take the mean of our sample to get an unbiased estimator. However, if values of $X_1, ..., X_N$ are sampled such that the sampling probability of X_i is larger when $X_i = 1$, then any sample, no matter how larger, will give biased estimated of p.

Selection bias occurs often in practice but can be difficult to identify. Unlike the case of missing data (described in section 1.2.2), selection bias is often invisible - we don't know anything about the units that weren't sampled. The problem presented by selection bias has been a thorn in the side of empirical research for decades.

One classic example of selection bias is a 1936 presidential election poll published in an American magazine called The Literary Digest. The election was between Franklin Delano Roosevelt, the democratic incumbent, and Alfred Landon, a republican. The Literary Digest conducted a poll drawing a sample from their readership and by telephone and using records of automobile ownership. The sample size was extremely large, with over 2 million survey respondents. Their prediction, that Landon would comfortably win the election, proved embarrassing. Not only did Landon lose, but he lost in one of the largest electoral landslides in American history: 523 electoral votes went to Roosevelt and 8 went to the Digest's predicted winner. Shortly after, the magazine ceased publication.

The precise details of what caused this failure are not fully known. One view that has become the popular explanation was that although their survey had a plenty large sample size, it was non-representative of the population as a whole. Particularly, both being subscribed to The Literary Digest and owning a telephone or car were more common among more affluent people, and so their survey skewed wealthier than the population as a whole, and therefore underestimated the proportion of voters who would vote for Roosevelt, a candidate with largely working class support. In fact, although this explanation almost certainly accounts for a part of the failure, it is unlikely to be complete. Although it is true that the survey over-sampled more affluent voters who were more likely to support Landon, a separate poll conducted by George Gallup in 1937 found that in even voters who owned both a telephone and a car were more likely to favor Roosevelt than Landon. This means that there must have been other factors at play, and in particular, that Landon voters were more

likely to return the survey than Roosevelt voters (Squire, 1988), Lusinchi (2012). This is called (non) response bias, and we will see that it too falls under the selection bias framework proposed here.

Selection bias can be even more subtle than this. Medical studies into the effects of a certain health conditions rely on samples of people with that condition. This can introduce a phenomenon called "survivorship bias" wherein the most severe cases are less likely to be included in the analysis because they are more likely to result in death (Delgado-Rodríguez & Llorca, 2004). Of course, similar effects exist outside of medicine.

One such case which has made its way into statistical legend concerns the work done by Abraham Wald and the Statistical Research Group (SRG) of which he was a member. The SRG was a working group of statisticians and mathematicians based at Columbia who applied statistical methods to problems faced by the American military during World War II (**?**). Asked to reduced the number of planes being taken down by enemy fire through discovering the areas of the plane most vulnerable to damage and reinforcing them, Wald pushed back on the military's suggestion to armor the areas which (within the returning aircraft) were most affected by gunfire. In fact, Wald believed that the opposite - reinforcing the areas *not* damaged in the returning aircraft - would provide the best protection. His reasoning was simple: all regions of the aircraft were similarly likely to be hit, but only those planes hit in the regions which were not essential were able to return. In other words, the full set of aircraft would have damage uniformly throughout the body of the plane, but selection into the set of returning aircraft depended on the planes not being so badly damaged that they went down (**?**). So, the areas not damaged in the returning set must have been the areas which, if damaged, would lead to the plane going down. In this way, Wald was used additional pieces of information (uniformity of hits to the plane and the fact that downed planes had been hit in vital regions) to effectively make use of selection-biased data which otherwise would have led to dangerously wrong conclusions.

The importance of establishing the mechanisms by which selection occurs has led to the development of graphical representations for the selection process. In the 2000s, a handful of papers were published using graphs to represent selection bias in the context of medical research (Geneletti et al., 2008) (Hernán et al., 2004) (Haneuse et al., 2009). More recently, a group of computer scientists has generalized this work in terms of causal graphs to make substantial headway on the problem (Pearl & Bareinboim, 2011) (Bareinboim & Pearl, 2012) (Bareinboim et al., 2014). This latter body of work is essential to our study of the problem and is reviewed in the next chapter.

1.2.2 Missing Data

Like selection bias, missing data occurs when some variables are not observed for each unit (row). Unlike selection bias, missing data can always be detected - missing values are not known, but the fact that they are missing is. In the standard missing data notation, $R_{Y_i} = 1$ corresponds to the i^{th} observation of a random variable Y being missing. In some cases, missingness follows a very simple structure where a particular variable is not always observed (often the response variable), but covariates are available for each unit. Table 3.1 gives an example of what is meant by this.

One example of this effect is found in studies that observe units over a long period of

Y	X_1	X_2	X_3	X_4	R_Y
y_1	x_{11}	x_{12}	x_{13}	x_{14}	0
-	x_{21}	x_{22}	x_{23}	x_{24}	1
-	x_{31}	x_{32}	x_{33}	x_{34}	1
y_2	x_{41}	x_{42}	x_{43}	x_{44}	0

Table 1.1: Hypothetical table representing missing data with one partially observed variable.

X_1	X_2	X_3	R_{X_1}	R_{X_2}	R_{X_3}
-	-	x_{13}	1	1	0
-	x_{22}	x_{23}	1	0	0
x_{31}	x_{32}	x_{33}	0	0	0
-	x_{42}	-	1	0	1

Table 1.2: Hypothetical table representing missing data with all variables partially observed.

time, such as longitudinal studies. In these cases, researchers have to deal with drop-out: some participants stop participating in the study before it finishes, and therefore the final outcome is not recorded for these units. Since the pre-treatment covariates are recorded for all participants before dropout occurs, only the outcome variable is missing for the units affected by dropout. This situation is a prime candidate for a technique called inverse probability weighting, which we explore in the next chapter (Hernán et al., 2004).

Sometimes the missingness structure is more complex in that different variables are missing from different units. So while one survey respondent might answer a question on income but decline to answer a question about marital status, another might do just the opposite. A range of approaches have been developed: sometimes, rows with missing variables may be safely deleted, other times, imputation based methods are more appropriate (Schafer & Graham, 2002). For instance, survey respondents may choose not to answer some questions, but unlike in the example above, the unanswered questions can differ between units (table 3.2). In cases like this, missing data are often imputed: that is, the missing values are predicted on the basis of non-missing values, but weight based approaches like inverse probability weighting can also be applied (Little & Rubin, 1986) ,(Seaman & White, 2011).

The degree of the problem posed by missing data will depend on the form that the missingness takes. In the worst cases, a variable causes its own missingness, an effect called "self-censoring". When this happens, the problem becomes very hard for the same reason selection bias is so vexing: the sampling distribution for the missing variable will never align with the population distribution and without external data or added assumptions, there is no way to say anything about what this underlying distribution is likely to be. In more favorable cases, whether or not an entry is missing for a particular unit will only depend on the set of observed values for that unit.

Since such a wide range of missingness structures are possible, a large literature has developed to address the problem in various contexts (Schafer & Graham, 2002) (Heckman, 1979) (Little & Rubin, 1986), much more so than with selection bias. Indeed, missing data has several qualities that often make it easier to handle than selection bias. When some (but not all) features for a particular unit are not measured, those features that are measured

can often inform on the features that do not. In many cases, a model can be specified to replace or "impute" the missing values (Donders et al., 2006). In the best case, the rows with missing data can be excluded from the analysis, but unless strict assumptions are met, this method will produce biased results and potentially greatly reduce sample size. Still, even with the wealth of methods developed for handling missing data, it can be difficult or even impossible to determine which are appropriate in a particular situation (Mohan & Pearl, 2019). Just as with selection bias, this fact motivates the recent formulation of the missing data problem in terms of causal graphs as discussed in chapter 3.

Chapter 2

Graphical Representations of Selection Bias

2.1 Introduction

In chapter 4, we give an overview of relevant methods for selection bias and missing data, including several older papers that describe sample selection graphically. However, these papers mostly made use of the graphs as conceptual models of selection for particular problems without digging into the implications of such structure in the context of particular study designs (Hernán et al., 2004), (Geneletti et al., 2008). Bareinboim, Tian, and Pearl's 2014 paper "Recovering from Selection Bias in Causal and Statistical Inference" provides a formulation of selection bias explicitly in terms of causal DAGs and d-separation, making it an especially important resource for developing our study of the topic. In particular, the paper provides graphical conditions (in terms of d-separations) under which a *conditional* distribution $P(y|x)$ can be obtained despite selection bias. When this can be done, we say that the $P(y|x)$ can be "recovered", from a causal graph containing a selection mechanism. To illustrate the results in a practical context, we will apply the graphical approach to the Literary Digest election polling problem described in section 1.2.1.

Bareinboim et. al is broken into three sections, the first two of which are the most substantial and the focus of our attention. The first gives conditions for when a distribution can be recovered if only the biased distribution is accessible, the second gives conditions for the case where population level data is available for some variables. As we will see, the presence of external data turns out to be quite useful.

2.2 Representing Selection with DAGs

Although the topic of this paper is selection *bias*, the graphical framework is really a model of selection itself. That is, it can model cases where selection does not induce bias as well as cases in which it does. Representing selection in this way requires augmenting a structural causal model over a set of variables $\mathbf{V}$ with a node S that represents inclusion in the sample. The "manifest" (or observed) distribution is then $P(\mathbf{v}|S = 1)$, that is, the joint over our variables $\mathbf{V}$ conditioned on inclusion within the sample.

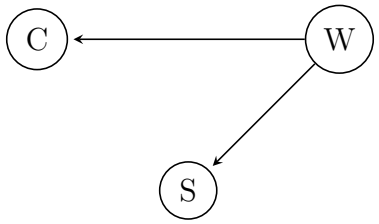


Figure 2.1: Graph corresponding to selection bias based on income

In the finite population setting, we can think about this labeling every unit in the population with $S = 1$ if they are included in our sample and $S = 0$ otherwise. The advantage of this is that S is can now be represented as another variable within the causal model, and therefore the d-separations within the associated causal graph give access conditional independences which allow selection to be ignored. This framework captures the non-biasing selection case as a selection graph in which there are no edges to S. In such a graph, $\mathbf{V} \perp\!\!\!\perp_G S|\emptyset$ and so $P(\mathbf{v}|S = 1) = P(\mathbf{v})$.

Finally, it should be emphasized that this method completely ignores sample size. In the real world, any sample is going to be finite and therefore any estimate of the population distribution, regardless of selection mechanism, will be subject to sampling variation. Mitigating this kind of uncertainty is not a concern of this work. Instead, it focuses purely on the effect of selection on the distribution from which the sample is being drawn. In other words, the focus is on the effect of selection on the sample as the sample becomes infinitely large.

2.3 Recovering Conditionals Without External Data

In the basic case, we only have access to a distribution corresponding to a causal graph G with a set of nodes $\mathbf{V}$ equipped with a selection node S. The distribution $P(\mathbf{v}|S = 1)$ corresponds to the distribution of values that are selected for inclusion in the sample. So, returning to our election example, in the simplest model in which only income affects selection probability, our graph would look like figure 2.1. In the figure, C is a random variable for the candidate a person intends to vote for, W is a measure of income and S, as usual, is the binary variable denoting inclusion in the sample. Although in practice we would probably like to know $P(c)$, the population level distribution of intended votes, for now we restrict ourselves to $P(c|w)$, the conditional distribution of intended vote given SES.

In this vein, Bareinboim et al. define **s-recoverability** as the criterion which must be satisfied for a conditional distribution such as $P(c|w)$ to be obtainable despite the selection mechanism. We reproduce their definition exactly and give an explanation of why it makes sense:

Definition 2.3.1. *Given a causal graph G_s augmented with a node S encoding the selection mechanism, the distribution $Q = P(y|\mathbf{x})$ is said to be **s-recoverable** from selection biased data in G_s if the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{v}|S = 1)$. Formally, for every two probability*

distributions P_1 and P_2 compatible with G_s, $P_1(\mathbf{v}|S=1) = P_2(\mathbf{v}|S=1) > 0$ implies $P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$.

In the example, if we are to recover $P(c|w)$ we need to express $P(c|w)$ as values derivable from $P(c, w|S=1)$ and the conditional independences implied by the d-separations within the graph. This is what is meant by "if the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{v}|S=1)$", and it is worth exploring why this fairly intuitive formulation is equivalent to the formal definition given in the last sentence.

The agreement of P_1 and P_2 on the joint distribution $P(\mathbf{v}|S=1)$ means that P_1 and P_2 produce the same observational distribution under selection bias. Therefore, when $P_1(\mathbf{v}|S=1) = P_2(\mathbf{v}|S=1) \implies P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$ we know that no matter the differences between P_1 and P_2, they must produce the same conditional $P(y|\mathbf{x})$. So, if this is true for every P_1, P_2, then all distributions consistent with the observational $P(\mathbf{v}|S=1)$ produce the same $P(y|\mathbf{x})$.

This leads to the first important result of the paper in the following theorem:

Theorem 1. *The distribution $P(y|\mathbf{x})$ is s-recoverable from G_s if and only if $S \perp\!\!\!\perp_{G_s} Y|\mathbf{X}$.*

The proof of $\impliedby$ is quite long, but $\implies$ is easy.

Proof. Assume that $S \perp\!\!\!\perp_{G_s} Y|\mathbf{X}$. Then by the Markov property, $S \perp\!\!\!\perp Y|X$ and $P(y|\mathbf{x}, S=1) = P(y|\mathbf{x})$. So the 'recovered' distribution is simply the biased distribution. □

Remarkably, this theorem means the *only* time that $P(y|\mathbf{x})$ is recoverable is when $\mathbf{X}$ d-separates S and Y in the graph. In our application, we can see that $C \perp\!\!\!\perp_G S|W$, and so $P(c|w) = P(c|w, S=1)$. So our study gives us direct access to the conditional distribution for candidate preference by income. This is potentially useful but restricted. What if we wish to recover a conditional distribution but do not want to condition on the full separating set? To do so, we will need external measurements.

2.4 Recovery With External Data

Sometimes a researcher will conduct a survey and collect variables on survey participants for which we know the population level distribution. For instance, the US Census collects and publishes data on occupation, race, income, and many other variables from the entire US population. If some of these variables are measured in our selection biased study, this information can be used to recover distributions that would otherwise be unavailable.

To elaborate on this topic we will introduce a more complex model of the selection mechanism in our voter poll. As mentioned in the introduction, more recent research suggests that the assumption that income was the only variable affecting the probability of being included in the poll is not sufficient. A better explanation includes the fact that Roosevelt supporters who received the survey were much less likely to respond than Landon supporters Lusinchi (2012). This effect is called non-response bias, and is really another form of selection bias since it arises from the differing probabilities of being included in the final sample. The exact mechanism of the non-response effect in this poll is not and likely will never be known.

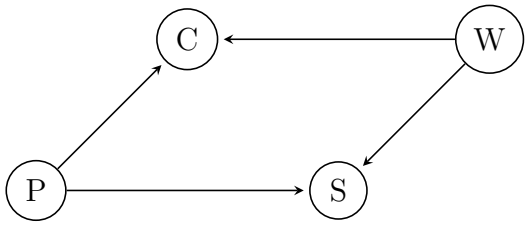


Figure 2.2: Graph corresponding to selection bias based on income and party registration

In fact, if candidate preference directly influences the likelihood of survey response, (i.e. $C \to S$), the two nodes cannot be d-separated and we will be out of luck. Being optimistic, however, let's say we have reason to believe that another factor is the culprit. Particularly, the political party to which a survey respondent is registered could be mediating the relationship between voter preference and response probability. This model is represented in the figure below.

Where P is a binary variable denoting the party registration status of the voter. Under this model, we have a path $S \leftarrow U \to V$ which is not blocked by W. So $S \not\perp\!\!\!\perp_G Y|X$ and by the previous theorem, we cannot recover $P(c|w)$. Of course, we could recover $P(c|w, p)$ since $S \perp\!\!\!\perp_G V|\{W, P\}$, but assuming we are still interested in $P(c|w)$, what can be done?

Well, we could consult the US census as well as state voter records and gather the joint distribution for income and party registration $P(u, w)$. This is what we would call *external data* since it is not affected by the selection mechanism. It turns out that in this case, we can do something. To address this situation, Bareinboim et al. introduce a revised definition of s-recoverability which we again reproduce and comment on.

Definition 2.4.1. *Given a causal graph G_s augmented with a node S encoding the selection mechanism, the distribution $Q = P(y|\mathbf{x})$ is said to be* ***s-recoverable*** *from selection bias in G_s with external data over $\mathbf{T} \subseteq \mathbf{V}$ and selection biased data over $\mathbf{M} \subseteq \mathbf{V}$ if the the assumptions embedded in the causal model render Q expressible in terms of the distribution under selection bias $P(\mathbf{m}|S = 1)$ and $P(\mathbf{t})$, both positive. Formally, for every two probability distributions P_1 and P_2 compatible with G_s, if they agree on the available distributions $P_1(\mathbf{v}|S = 1) = P_2(\mathbf{v}|S = 1) > 0$, $P_1(\mathbf{t}) = P_2(\mathbf{t})$ they must agree on the query distribution $P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$.*

We can see that this definition follows the same structure as the original s-recoverability definition, but is expanded to allow for the use of population level distributions. The paper's second theorem follows directly from this definition.

Theorem 2. *If there is a set $C \subseteq V$ such that $P(\mathbf{c}, \mathbf{x})$ is measured in the population and $Y \perp\!\!\!\perp_{G_s} S|\{C, X\}$ then $P(y|\mathbf{x})$ is s-recoverable as*

$$P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c}|\mathbf{x})$$

The theorem is a straightforward application of the law of total probability.

Proof. By assumption, we have the external distribution $P(\mathbf{c}, \mathbf{x})$ and therefore $P(\mathbf{c}|\mathbf{x})$, and as usual we have $P(\mathbf{v}|S = 1)$. So can apply the law of total probability and the conditional independence $Y \perp\!\!\!\perp S|\{C, X\}$ to write:

$$P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c})P(\mathbf{c}|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1)P(\mathbf{c}|\mathbf{x})$$

□

Therefore, if we wish to recover $P(c|w)$ party registration affects response probability, we can do so using the external data $P(p, w)$.

2.5 A Useful Extension

As we have seen, Bareinboim et al. is focused on the recovery of conditional distributions, i.e. $P(c|w)$. However, we would often like to have the unconditional distribution $P(v)$. Although it is only mentioned obliquely at the end of the second section, the results they prove give a simple condition for when $P(c)$ is recoverable using external data. In both sections we have seen that $P(c|w)$ was recoverable. Then, assuming that we have the external data for $P(w)$, we can use the law of total probability to write:

$$P(c) = \sum_{w} P(c|w)P(w).$$

More generally, we can formulate this result as a corollary, although it is not listed as such in the paper.

Corollary 2.5.1. If there exists a set $\mathbf{X} \subseteq V$ such that $P(y|\mathbf{x})$ is s-recoverable and $P(\mathbf{x})$ is available externally, then $P(y)$ is recoverable as $P(y) = \sum_{\mathbf{x}} P(y|\mathbf{x})P(\mathbf{x})$.

Being able to recover marginal distributions (rather than conditionals) will come in handy later when we compare recovery from selection bias to recovery from missing data.

Chapter 3

Graphical Approaches to Missing Data

The advantages of graphical representations of selection are also found in graphical representations of missing data. As with selection, the dependence structure of missing data determines the whether a particular quantity is accessible despite missingness. This fact has long been recognized (Mohan & Pearl, 2019), (Little & Rubin, 1986) and led to the widespread adoption of a classification system based on these relationships.

3.1 Rubin's Taxonomy of Missing Data

Donald Rubin's work on the missing data problem is likely the most influential out of any researcher on the topic. Rubin developed several notable techniques to address missing data including multiple imputation, the expectation-maximization algorithm, and the theory of propensity scores (Little & Rubin, 1986). We will discuss these methods in more details later in the next chapter, but Rubin's taxonomy (Rubin, 1976) of the structure of missingness is important enough to introduce now. Consider a case with a set of variables V_m potentially subject to missingness, a set V_o of fully observed variables, and a set U of unmeasured variables (Mohan et al., 2013). Define R as the set containing the variables that dictates missingness among the elements of V_m. So if $V_m = \{Y_1, Y_2\}$ then $R = \{R_{Y_1}, R_{Y_2}\}$ where $R_{Y_i} = 1$ when Y_i is missing (Mohan et al., 2013). We consider three cases:

- The best case is when missingness is unrelated to all the variables, i.e. $R \perp\!\!\!\perp V_m \cup V_o \cup U$. In this case, the data are called *missing completely at random* (MCAR). When this condition holds, analysis can safely be conducted on only the rows which are fully recorded.

- If missingness depends only on the fully observed variables, i.e. $R \perp\!\!\!\perp V_m \cup U | V_o$ we say that the data are *missing at random* (MAR). Many techniques have been developed for data satisfying this condition. Overall, these techniques are more complex that those appropriate for MCAR data but are highly effective (Schafer & Graham, 2002). This type of missingness is sometimes called "ignorable" not because it can actually

be ignored, but because it does not pose a threat with the right techniques (**?**). Note that all MCAR data is also MAR.

- When the previous condition does not apply, the data are said to be *missing not at random* (MNAR). This is the worst case, and is sometimes called "non-ignorable" since the techniques used to effectively adjust for MAR data cannot generally be applied. However, as we will see, the graphical approach shows that some MNAR situations are in fact recoverable.

These definitions are actually slightly adapted from Rubin's original formulation because they refer to conditional independencies among the variables rather than independencies at the unit level. That is, Rubin's original definitions allowed for data to be considered MAR when in each *row*, the missingness of whichever variables were observed for the particular unit depended only on the observed values for that unit. This means that our definition is slightly more restrictive than Rubin's but this alteration makes it much more conducive to graphical representation (Mohan et al., 2013) (Schafer & Graham, 2002).

To further develop the intuition, we describe variations on survey non-response satisfying each of these conditions. Consider a very simple survey that asks each respondent for their education E and their annual income I. Assume that age is fully observed but that income is not.

- (MCAR) A bug in the software used to conduct the survey means that a uniformly random selection of respondents do not receive the income question. Everyone who receives the income question answers it.

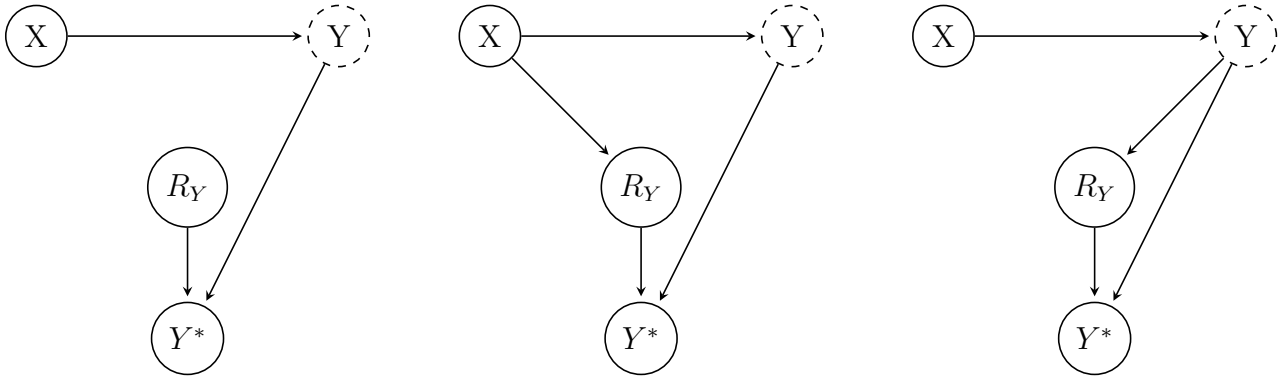
- (MAR) Respondents with higher levels of education are more likely to respond to the income question that those with lower education. However, within strata of education, the question is answered with equal probability.

- (MNAR) High income respondents are less likely to respond to the income question than those with lower income. It does not matter if education is associated with non-response within the strata of income.

However, although we can describe or display particular situations under which each condition would hold, the task becomes more difficult in practice. For one, it can be shown that it is generally not possible to test for whether or not data satisfy the missing at random condition (Schafer & Graham, 2002). This limitation means that given a particular dataset containing missing values, it (usually) is not possible to determine whether the missing values are missing at random or not without further assumptions.

3.2 Graphical Innovations

The difficulties in establishing the precise conditions under which the definitions apply as well as the conditions which allow for distributions of interest to be recovered led to the recent formulation of the missing data problem in graphical terms. As with the analogous literature we have explored on selection bias, this work also was headed by Judea Pearl



(a) (b) (c)

Figure 3.1: From left to right, bivariate cases of MCAR, MAR, and MNAR missing data patterns

and one of his graduate students, Karthika Mohan (Mohan et al., 2013) (Mohan & Pearl, 2019). Fortunately, familiarity with Bareinboim's framework makes understanding Mohan and Pearl's straightforward.

As above, we consider four sets of variables: fully observed variables V_o, partially observed variables V_m, unobserved variables U, and the missingness variables R. To make the missingness process more explicit, we add to this a set of "proxy" variables $V_m^* = \{Y_i^* | Y_i \in V_m\}$ which reflect what we actually observe, such that:

$$y_i^* = \begin{cases} \text{"missing" if } r_{y_i} = 1 \\ y_i \text{ if } r_{y_i} = 0 \end{cases}$$

Notice that Y_i^* is a function of R_{Y_i} and Y_i. Since we are working with causal models, this fact implies that $PA_{Y_i^*} = \{R_{Y_i}, Y_i\}$. In this sense, R_{Y_i} is not just an indicator of missingness, but the variable which enforces missingness of Y^* in the causal model. With these sets constructed we define the missingness graph:

Definition 3.2.1. *A missingness graph, denoted* ***m-graph****, is a causal graph* $(\mathbf{V}, E)$ *where* $\mathbf{V} = V_o \cup V_m \cup V^* \cup U \cup R$ *and* E *is the set of edges in the graph.*

Looking at figure 3.1 (a) and (b) we have the following d-separation relationships: (a) $R_Y \perp\!\!\!\perp Y$ and (b) $R_Y \perp\!\!\!\perp Y|X$. This corresponds to the missing completely at random and the missing at random conditions, respectively. However, in the final scenario, (c), there is no set that d-separates R_Y from Y. In each case, we have access to the joint $P(X, Y^*, R_Y)$. By definition of Y^*, when we condition on $R_Y = 0$ we get $P(x, y|R_Y = 0)$, the distribution of "complete cases".

Suppose we wish to know the population level joint $P(x, y)$. In cases (a) we have that $P(x, y|R_Y = 0) = P(x, y)$ and we are done. In case (b), the d-separation in the graph imply that $P(Y|x) = P(y|x, R_Y = 0)$ and so we apply the chain rule to get $P(x, y) = P(y|x)P(x) = P(y|x, R_Y = 0)$ and again we succeed. Indeed, our definitions of MCAR and MAR mean this

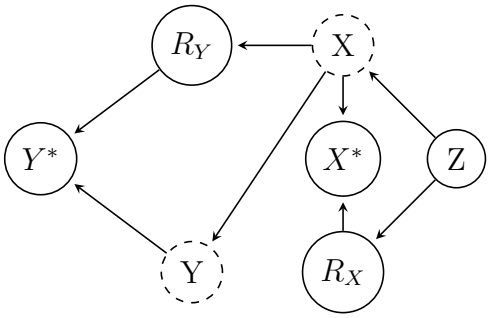


Figure 3.2: A three variable MNAR m-graph

method works for any such missingness structure. However, the lack of independence in the final case means that we cannot separate Y from its missingness mechanism, and therefore we cannot factor out the conditional. In fact recovery is not possible in this case because Y and R_Y are neighbors (Mohan et al., 2013).

3.3 Recovering Distributions Under MNAR

With missing data, as with selection bias, what we really care about is not the distribution we have - the one subject to missingness or selection - but the population level distribution. We have seen that under the MCAR and MAR missingness conditions, the population distribution is straightforward to obtain. However, when missingness can be affected by the variables that are missing, the problem becomes much more difficult. The major task of Mohan and Pearl's papers is providing graphical characterizations for those situations in which recovery of a particular joint or condition distribution are possible or impossible. Conveniently, definition of recovery given in Mohan et. al 2013 is equivalent to the defintion used in Bareinboim 2014, making moving between the two easy.

The first theorem provides a general condition under which a "query" Q - meaning a probabilistic quantity such as a joint or conditional distribution - can be recovered.

Theorem 3. *Given a m-graph G and an observed distribution $P(v^*, v_o, R)$ a query Q is recoverable if Q can be decomposed into an ordered factorization or a sum of such factorizations, suc h that every factor $Q_i = P(y_i|x_i)$ satisfies $Y_i \perp\!\!\!\perp (R_{X_i}, R_{Y_i})|X_i$. Then, each Q_i can be recovered as $P(y_i^*|x_i^*, R_{Y_i} = 0, R_{X_i} = 0) = P(y_i|x_i)$ (Mohan et al., 2013).*

What this theorem is saying is fairly simple: Q can be recovered when it can be expressed as a sum or product of quantities that can be recovered individually.

Example 4. *Consider the m-graph in figure 3.2. We notice that the one fully observed variable Z does not d-separate (R_X, R_Y) from (X, Y) and therefore conclude that this m-graph corresponds to an MNAR missingness structure. However, we do have some relevant d-separations:*

1. $X \perp\!\!\!\perp R_X|Z$
2. $Y \perp\!\!\!\perp (R_Y, R_X, Z)|X$

Now, by the law of total probability and the chain rule for probability, we get that:

$$\begin{aligned} P(x,y) &= \sum_z P(x,y,z) \\ &= \sum_z P(x,y|z)P(z) \\ &= \sum_z P(y|x,z)P(x|z)P(z) \end{aligned}$$

When we apply our d-separations, we get that $P(y|x,z,R_Y=0,R_X=0) = P(y^*|x^*)$ *and* $P(x^*|z,R_X=0) = P(x|z)$ *and therefore:*

$$\sum_z P(y|x,z)P(x|z)P(z) = \sum_z P(y^*|x^*,z,R_Y=0,R_X=0)P(x^*|z,R_X=0)P(z)$$

So, we have succeed: the marginal distribution $P(x,y)$ *is expressible as a sum and product of distributions which are recoverable simply using the d-separations, and so* $P(x,y)$ *is recoverable by theorem 4.*

The second theorem aims a little higher: necessary *and* sufficient conditions for recovering the full joint distribution $P(v_o, v_m)$.

Theorem 4. *Given a m-graph G with no edges between R nodes the joint distribution* $P(v_o, v_m)$ *can be recovered if and only if there is no variable* $X \in V_m$ *such that:*

1. *X and* R_X *are neighbors*
2. *X and* R_X *are connected by a path in which all intermediate nodes are colliders and elements of* $V_m \cup V_o$ *(i.e. at least partially observed, not an R node, and not a proxy).*

When neither of these conditions apply, the joint $P(v_m, v_o)$ *is recovered as:*

$$\frac{P(v, R=0)}{\prod_i P(R_i=0|Mb(R_i), R_{Mb(R_i)}=0)}$$

Where $Mb(R_i)$ *is any* ***Markov blanket*** *for* R_i*: a set of nodes* $Mb(R_i)$ *such that* $R_i \perp\!\!\!\perp \mathbf{V}\backslash\{R_i, Mb(R_i)\}|Mb(R_i)$*. In many cases, this will simply be* PA_{R_i}*. Similarly,* $R_{Mb(R_i)}$ *is the set of R variables associated with the elements of* $Mb(R_i)$ *(Mohan & Pearl, 2019).*

This final expression is a little difficult to parse, so we clarify what each component means in the following example.

Example 5. *Lets return to the m-graph described in the previous example (figure 3.4). We check the two conditions laid out in the theorem:*

1. *The two variables subject to missingness, Y and X both lack an edge to their corresponding missingness node, and so the first condition is satisfied.*

2. *The only colliders in the graph are the two proxy nodes: X^* and Y^*. Therefore, no path contains a collider which is an element of $V_o \cup V_m$.*

So, the population level joint $P(x, y, z)$ may be recovered as given in the theorem. This gives:

$$P(x, y, z) = \frac{P(x, y, z, R_X = 0, R_Y = 0)}{P(r_X|z)P(r_Y|x, R_x = 0)}$$

It is worth noting here how much stronger these results are than anything that was shown for selection bias. There, we saw that (without external data) the only distributions which can be recovered are conditionals $P(y|\mathbf{x})$ for which $Y \perp\!\!\!\perp_G S|\mathbf{X}$. When data is missing, a relatively weak condition, no self-censoring and no collider paths, is all that is needed to recover the full population distribution, as long as no edges exist between R nodes. Intuitively, this gets back to the point that selection bias is much worse than missing data. With missingness, we know something about every unit, and we know what we don't know (that is, we can tell when a value is missing). Under selection bias, all we have is our biased data - we know nothing about the units that were not sampled.

3.3.1 Conditions for Non-Recoverability

As important as it is to know when a particular quantity can be recovered, it is sometimes just as useful to know when a quantity cannot be. To this end, some attention has been paid to establishes those graphical conditions that prevent recoverability. One of these we have already seen: theorem 5 gives that whenever X and R_X are neighbors or X and R_X are connected by a path in which all intermediate nodes are colliders and elements of $V_m \cup V_o$, the joint $P(V_m, V_o)$ is not recoverable. In fact, the necessity direction of this theorem is proven using two lemmas which are useful on their own.

We reproduce these lemmas below:

Lemma 3.3.1. In the graph $X \rightarrow R_X$, $P(X)$ is not recoverable.

Lemma 3.3.2. If a quantity Q is not m-recoverable in graph G, it will not be recoverable after the addition of any edge to the graph.

The second is especially useful because it says that if a quantity is not recoverable in a particular subgraph of our missingness graph, we can use lemma 2 and a simple induction to show that the quantity is not recoverable in the full graph, either. This is exactly what is done to prove that a single edge from a node X to R_X implies that $P(V_m, V_o)$ is not recoverable. Indeed, this makes clear that when an edge $X \rightarrow R_X$ exists, the marginal $P(\mathbf{S})$ cannot be recovered if $X \in \mathbf{S}$.

By similar reasoning, **?** give a similar corollary for the non-recoverability of conditionals, which we also reproduce.

Corollary 3.3.1. For disjoint subsets $\mathbf{X}$ and $\mathbf{Y}$, $P(\mathbf{Y}|\mathbf{X})$ is not recoverable if one of the following conditions is true:

Y and R_Y are neighbors

There is a collider path connecting R_Y and Y such that all intermediate nodes are in $\mathbf{X}$.

This condition is obviously similar to the condition for recovering the joint, and in chapter 5, we continue in this direction to derive more conditions for non-recovery.

Chapter 4

Earlier Approaches to Selection Bias and the Missing Data Problem

In statistics and many related fields, forms of the effect we are calling "selection bias" and "missing data" have gone by many names. In epidemiology and biostatistics, Berkson's bias (sometimes Berkson's paradox) describes a phenomenon in which two conditions that are independent at the population level become dependent within a sample as a result of both conditions affecting the likelihood of sample inclusion. This too is a kind of selection bias, as is Neyman's bias which results from survivors of a disease being selected for study but not those who died from the disease. In fact there are far more names for particular selection biases than we can describe here (Delgado-Rodríguez & Llorca, 2004). One explanation for this extensive and overlapping nomenclature is that prior to the widespread use of graphical models, it was difficult to give any concise definition of selection bias that was specific enough to be useful. Selection bias is essentially about a sample with a different distribution than the population, but when that non-representivity can come from so many places this characterization is hardly useful.

Unfortunately, the different threads within the literature has caused certain terms to be overloaded. As we have seen, the tendency for types of survey recipients to respond more or less than others - (non) response bias - falls neatly into the framework of selection bias, but the same phrase is used to describe the selective answering of particular questions, which is a problem of missing data. Even the terms "missing data" and "selection bias" have this problem. For instance, James Heckman's famous "Heckman Correction" is widely described as a technique for dealing with selection bias. However, under our definitions, the method actually addresses missing data.

In this section we will review a handful of notable earlier approaches to defining, correcting, and recognizing selection bias and missing data that have appeared within the literature, being careful to note which problem is being addressed. The purpose here is two-fold. For one, the historical background is generally useful to anyone looking for a thorough treatment of the subject. More pointedly, we will use the examples of previous methods to argue for the utility of the causal/graphical formulation used in this paper, as well as interpreting the techniques in that light.

4.1 Case Analysis

The easiest and most obvious way of "addressing" missing data is to ignore units which have missing values. There are a couple of techniques that take this approach. These techniques are widely used in practice despite significant undesirable theoretical properties (Little & Rubin, 1986) which we discuss. In some cases, the adverse effects might be small - such as when missingness is very uncommon - but in other cases such techniques dramatically bias the results.

4.1.1 Complete-case Analysis

In "complete-case" analysis only rows which are complete - not missing *any* values - are included in the analysis. This can be thought of as deleting any row which has a missing value. There are two main problems with this approach. The first has to do with the structure of the missingness. In any case other than MCAR, the most restrictive of our taxonomy, complete case analysis produces distributions different from the population distribution (Little & Rubin, 1986). This is because complete case analysis is essentially a process by which data subject to missingness is made into data affected by selection bias. Since we remove any rows with missing values, we can think about complete case analysis as the process of constructing a complete data set subject to selection. Particularly, in the new data set, $S = \begin{cases} 1 \text{ if } R = 0 \\ 0 \text{ if } R = 1 \end{cases}$ and so in the corresponding graph, and edge would exist between each parent of an R node and the selection node. Therefore, only under MCAR do we have a graph with no edges between the selection node and the rest of the graph.

The second problem with complete-case analysis is that in many contexts, it is highly inefficient (Little & Rubin, 1986). Consider a survey with a large number of questions and suppose we wish to know the distribution of the respondents' answers to a particular question. To conduct a complete-case analysis in such a situation would mean throwing out every row in which the respondent had failed to answer any of the many questions, even if they had answered the question of interest. In such situations, complete-case analysis results in sample sizes which are far smaller than the sample size of the original data set. This, of course, has consequences for the power of statistical tests, the size of confidence intervals, etc.

Despite these drawbacks, complete-case analysis is very common. As above, this is mathematically appropriate when data are MCAR, but this is often not the case. Nonetheless, complete case analysis may be "good enough" for some purposes, especially when the number of excluded rows is small enough that any reasonable values for the missing data in the column of interest are unlikely to dramatically affect the analysis (Schafer & Graham, 2002).

4.1.2 Available-case Analysis

The inefficiency of complete-case analysis has an obvious (partial) remedy. Instead of deleting every row which is subject to missingness in any column, we first select the variables of interest and then only delete rows which are missing values in the corresponding columns. This method is called "available-case" analysis or sometimes "pairwise deletion" (Schafer & Graham, 2002). Of course, when the number of variables of interest is close to

the overall number of variables this doesn't help much, but when only a couple are needed much more data can be preserved. However, this method does nothing to address the main theoretical problem of bias: once again, data which is not MCAR is generally not appropriate for available case analysis.

4.2 Imputation based Techniques

Imputation is the process of "guessing" missing values such that an analysis can be performed on the imputed data set. Imputation methods vary widely in their simplicity and practicality (Schafer & Graham, 2002). Broadly, imputation techniques fall into one of two categories: "single" imputation and "multiple" imputation. As the names suggest, the difference between the methods is that single imputation replaces missing values with a single imputed value whereas multiple imputation replaces missing values with multiple plausible values, effectively creating multiple datasets on which the analysis can be performed. We give an overview of both approaches.

4.2.1 Single Imputation

In single imputation, the task is to replace missing values with "reasonable" guesses. There are several techniques used to generate such values all of which rely on the missingness being MCAR or MAR. We summarize them below from simplest to most complex and then discuss their advantages and disadvantages.

- Mean Imputation: For a variable Y subject to missingness, the missing values are replaced with the mean of the observed values of Y (Little & Rubin, 1986).
- Hot Deck Imputation: For a variable Y subject to missingness, the missing values are replaced with the value from another row. Sometimes, this row is chosen randomly, whereas other times the row is selected based on similarity of observed covariates (Little & Rubin, 1986), (Schafer & Graham, 2002).
- Regression Imputation: For a variable Y subject to missingness, a model f is formulated based on the fully observed variables $\mathbf{X}$ is formulated and trained on the rows in which Y is observed. Then, the model prediction $\hat{y}_i = f(\mathbf{x}_i)$ is used to imputed each missing value of Y (Schafer & Graham, 2002).
- Random Regression Imputation: Same as regression imputation except rather than replacing missing y_i with $\hat{y}_i$ we sample it from the conditional distribution implied by the model.

In general, the methods at the end of this list are preferable to the methods at the beginning. Starting with mean imputation, there are two main problems. The first is that unless the data is MCAR, the conditional (on fully observed $\mathbf{X}$) expected value of the replacement $E[y|\mathbf{x}]$ will in general be different from the unconditional mean $E[y]$ which is used to impute. This means that when the data is not MCAR, the estimates based on the imputed data will be biased. The other problem is that because the imputed values are all

the same (the mean) the variance of the imputed column will be deflated (**?**). This occurs regardless of three structure of the missingness.

Hot-deck imputation attempts to solve this problem by adding noise in the form of a random value taken from the observed values of Y. So, if $Y = [1, 0, 1, 0, 0, m]$ (where m indicates missingness) than the missing value would be imputed as 1 with probability 2/5 and 0 with probability 3/5. This is generally preferable to mean imputation because while both methods preserve the (unconditional, observed) expected value of Y, but hot-deck imputation also preserves the (unconditional, observed) variance of Y. However, the major problem is unaffected: as long and Y and $\mathbf{X}$ are not independent, the information contained in the observed variables is wasted and estimates become biased. This means that the required assumption is very strong: for both mean imputation and hot-deck imputation are appropriate only when $Y \perp\!\!\!\perp (\mathbf{X}, R_Y)$. This means that not even MCAR missingness is enough: there cannot be any dependence between Y and $\mathbf{X}$ (by the same token, MAR data could be appropriate as long as $Y \perp\!\!\!\perp \mathbf{X}$).

The solution to this problem is fairly clear: a model of the form $Y \sim X$ should be constructed to predict the missing values of Y and then used to impute those values. Accordingly, this approach is called regression imputation. In the standard case, this is done using an estimate for $E[y|x]$, meaning that two rows (with Y missing) which agree on X will always be imputed with the same value. Analogous to the problem with mean imputation, the lack of randomness has the effect of producing exaggerating the covariance between X and Y (Schafer & Graham, 2002).

Fortunately, most models provide a way of adding their uncertainty back in. In the case of linear regression, the error term is assumed to be normally distributed and center at 0 with variance that can be estimated from the residuals. In Bayesian modeling, a posterior distribution is estimated, and can be sampled from directly (**?**). Either way, missing values can be drawn according to these distributions, fixing the covariance problems that arise in non-random regression imputation. Both of these methods require that the data is MAR rather than MCAR, meaning that with a correctly specified model, they can be justified in situations in which hot-deck and mean imputation will be inappropriate. Nonetheless, the estimation of quantities such as correlation, and especially confidence interval coverage, can still suffer even if MAR holds (Schafer & Graham, 2002).

4.2.2 Multiple Imputation

The problems with single imputation methods led researchers to search for more sophisticated techniques. Donald Rubin introduced first introduced the concept of "multiple" imputation (MI) in 1977 (**?**) but its classic treatment is Rubin's 1987 book on the topic. The method has become widespread is and regarded as the best technique for imputation and among the best technique for the analysis of missing data (**?**), (Schafer & Graham, 2002). Because of the large literature on the subject, attempting to fully describe the theory behind MI would go beyond the scope of this work.

Nonetheless, we hope to give a reasonable summary of how MI works and why it is considered generally superior to single imputation. The reason that MI is referred to as "multiple" imputation is that rather than replacing each missing value with a single guess, each is replaced with a vector of length $M > 1$. In effect, this means M datasets are created

and analyzed separately before being combined.

Some subtlety is involved in the way that these M imputations are made. It would be possible to draw each from the posterior distribution same model (or with noise term in linear regression), however, this method would not account for the uncertainty in the terms associated with the model. To do this properly is easiest in the Bayesian context. In that case, each parameter associated with the model has its own distribution. To construct the M imputed values, we first take M draws from the distribution of each parameter, creating M models from which we can impute the missing values (**?**).

The most common way of combining the M datasets is simply to take the arithmetic mean of the estimates constructed in each analysis. So, in the basic setting where Y is subject to missingness and $\mathbf{X}$ is fully observed such that we can model $P(R_Y = 0|\mathbf{x}, \boldsymbol{\delta})$ with some parameters $\boldsymbol{\delta}$, the standard proceedure for constructing an estimator $\hat{\theta}$ with a model using multiple imputation is as follows :

1. Draw $\hat{\boldsymbol{\delta}}_i$ for $i \in \{1, ..., M\}$ from the distribution associated with $\boldsymbol{\delta}$.
2. For each missing value of y, draw $\hat{y}_i$ from $P(y|\mathbf{x}, \hat{\boldsymbol{\delta}}_i)$ for $i \in \{1, ..., M\}$.
3. In each of the M datasets, compute $\hat{\theta}_i = T(y_i, \mathbf{x}_i)$ for the estimator function T.
4. Average the estimates as: $\hat{\theta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i$.

In theory, many of the desirable properties of MI rely on taking $M \to \infty$ (**?**). However, in practice, fairly small values of M (such as 20) can work well (Schafer & Graham, 2002).

4.3 The Heckman Correction

James Heckman's well-cited 1979 paper "Sample Selection Bias as a Specification Error" proposes a method for overcoming "sample selection bias" that is among the most prominent selection adjustment techniques (Heckman, 1979). Heckman was an economist, and his correction technique comes in the context of economic modeling, particularly linear regression models. Because it is situated within this framework, the correction is parametric - it requires the assumption of normally distributed noise. We consider the situation in the response variable is missing for some observations. Since we are interested modeling the population values and coefficients, this poses a potential problem. The title of Heckman's paper gives a hint to his strategy: specification error refers to the omission of a relevant variable from a model.

The method proceeds in two parts: first, a model is constructed for sample inclusion and second, an expected error term is calculated such that it can be included in the regression model to remove the bias associated with its exclusion. Understanding the specifics of Heckman's method requires substantial elaboration that is mostly irrelevant to our discussion, so in this section, we outline the approach without delving into the derivations. However, interested readers can find a detailed discussion of the technique (slightly adapted from the original to ignore finite sample concerns) in the appendix, or read Heckman's paper (Heckman, 1979).

4.3.1 Premise

Although presented as a method for addressing selection bias, Heckman's technique fits better into the missing data literature. In the traditional formulation, Heckman is attempting to estimate model coefficients in the context of a finite sample. So far, we have been considering distributions rather than samples, and for the sake of continuity (as well as avoiding excessive indexing) we will present the method in the context of distributions. Fortunately, the ideas are essentially the same.

The problem is as follows: a response variable Y is partially observed and assumed to fit the linear regression assumptions:

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

Where as usual $\epsilon \sim N(0, \sigma^2)$.

for an observed vector of variables X. The missingness of Y is governed by a variable Z which is not observed but also assumed to follow the linear regression assumptions for another set of fully observed covariates $\mathbf{W}$ (often assumed to be a superset of $\mathbf{X}$) such that:

$$Y = \mathbf{W}\boldsymbol{\delta} + \tau$$

With $(\epsilon, \tau) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ for some covariance matrix $\boldsymbol{\Sigma}$. Particularly, missingness occurs when Z exceeds some particular value c, i.e. $R_Y = 1$ when $Z > c$. Without losing generality, it is often assumed that $c = 0$.

Heckman's insight was that under this set-up, the biased estimate of β_0 produced by running a regression on only the fully observed rows can be treated as a "specification error" - the failure to include a relevant variable. This value can be computed analytically as $\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ where σ_τ is the standard deviation of τ, ϕ is the standard normal pdf and Φ is the standard normal cdf. The details of this calculation are somewhat involved and therefore are located in the appendix alongside some examples.

However, with this out of the way, the method is simple to implement. There are three steps:

1. First, a probit model is employed to estimate $\frac{\boldsymbol{\delta}}{\sigma_\tau}$, i.e. $P(S = 1|\mathbf{W}) = \Phi(\frac{\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})$.
2. This estimated quantity is used to get an estimate of $\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$.
3. The estimate of $\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ is included in the corrected model $Y = \mathbf{X}\boldsymbol{\beta} + \frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})} + \epsilon$. The coefficient on $\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ is then an estimate of $\gamma\sigma_\tau = \rho\sqrt{\sigma_\epsilon}$. In this model, the estimate of the intercept,$\hat{\beta}_0$, is now consistent, meaning that as the sample size increases $\hat{\beta}_0 \to \beta_0$ in probability.

4.3.2 Graphing Heckman

Although Heckman's method makes assumptions which are not captured by a causal graph (such as normality of ϵ, τ), it is still instructive to supply the graph that Heckman's setup

induces. One thing to note: when Y is not observed, its residual cannot be calculated. Therefore, we include R_ϵ as a node in the graph as a child of Z with the idea that ϵ is not observed when $Z > c$.

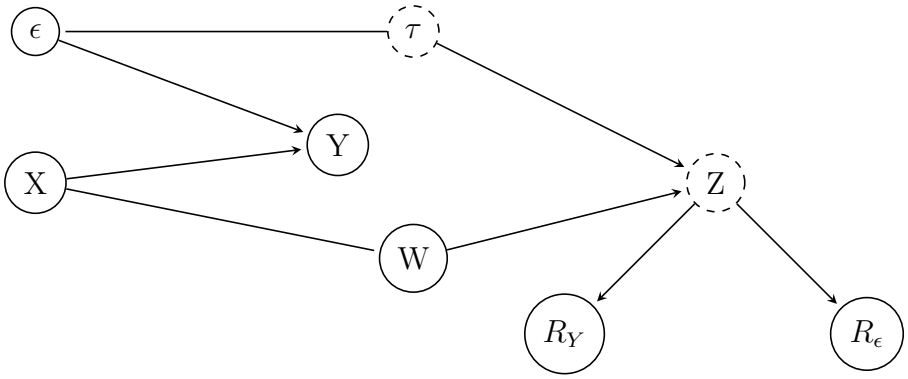


Figure 4.1: PDAG corresponding to the structural equations in Heckman's set up. We leave arrows off the edges between ϵ and τ and between X and W since the directions are not given by the structural equations. Dashed circles indicate unobserved variables.

Several insights can be gained from viewing this graph in the light of more recent methods. First, since $Y \perp\!\!\!\perp_G R_Y|Z$, if Z were observed it would be possible to recover $P(y|x)$ (and therefore $E[y|x]$) from $P(x,y) = \sum_z P(x,y|z)P(z) = \sum_z P(x,y|z,R_Y=0)P(z)$. However, when Z is not observed, the chain of unobserved variables ϵ, τ, Z between Y and R_Y mean that $P(y|x)$ is not recoverable.

One fact that comes up during the derivation is and is also clear in the graph is that when the edge between ϵ and τ is inactive, the sampling procedure will not affect the estimates since $Y \perp\!\!\!\perp_G S|X$ and so $P(y|x) = P(y|x,R_Y=1)$. Since (ϵ,τ) is bivariate normal, this occurs exactly when ϵ and τ are uncorrelated.

4.4 Weight Based Approaches

Another family of techniques worth mentioning is the broad array of weight based estimators. Versions of the technique are old and have uses well outside of adjusting for selection bias a (Horvitz & Thompson, 1952), but we give an interpretation particularly suited to our definition of selection bias. Given a particular (non-representative) sample, the goal is to construct a 'new' sample which follows the population distribution by giving the rows different 'weights' depending on their probability of being included in the sample.

Probability weighting techniques are associated with adjusting for confounding bias for causal inference, but methods have been developed that address selection bias as well (**?**). They are widely applied in survey statistics, where the goal is to generalize the results of the survey to some broad group such as likely voters, American adults, or members a a particular religious group. If costs and logistical concerns are ignored, the survey would be conducted using a "simple random sample" of the target population - that is - a sample in which every member of the target population is equally likely to be included in the sample.

Unfortunately, this is only rarely possible. Often times, some members of the population are much hard to reach than others, which was (part of) the problem we saw earlier with the Literary Digest survey before the 1936 election. When this happens, some units in the population are more likely to be sampled than others, and it is helpful to notate this by assigning the i^{th} member of the population of size N has sampling probability p_i (**?**).

For instance, in a simple random sample of size n, we have that $p_i = \frac{n}{N}$ for all $i \in \{1, ..., N\}$. Notice that $\sum_i p_i = n$ rather than 1. Then, regardless of the sampling pattern, each unit is assigned a weight $w_i = \frac{1}{p_i}$ called, appropriately, the inverse probability of selection weight.

However, the question remains about how such weights might be calculated. Sometimes, the specifics of the survey design allow for weights to be calculated, and in others, models (such as logistic regression) are specified (**?**). In either case, formulating selection graphically makes clear exactly what quantity needs to be estimated to construct the weights. In general, p_i could depend on almost any factor associated with the unit: income, age, name, location, primary language, etc. could all conceivably impact the likelihood of a person being included in the sample. When a graphical model is specified, we simplify the problem to estimating $p_i = P(S = 1|PA_S)$.

External data is often used to help calculate this value. One such approach is outlined by Pew Research, one of the most well known American polling organizations. Pew uses a combination of census data and representative survey data to construct a "synthetic population" distribution over a carefully selection set of measures that mirrors the real demographic profile of the United States. Then, the sampled units are compared with the synthetic population and a random forest model is fit to estimate the probability of sample inclusion for each of the units. Regardless of how weights these calculated, many quantities can be estimated without bias as long as the weights are calculated correctly (often a challenge). If a variable Y is measured in each of the n units, we can estimate $E[Y]$ using the weights as:

$$\hat{y} = \frac{\sum_j^n}{\sum_j^n w_j}$$

Much more complex estimators can also be constructed such as for linear regression or generalized linear models (Haneuse et al., 2009). Additionally, similar techniques can be applied for missing data. There, the weights are derived from probability of missingness, which can best estimated directly from the data (under the MAR assumption) (Seaman & White, 2011).

4.4.1 Probability Weighting for Distributions

For most of this paper, our primary concern is recovering entire distributions affected by selection bias rather than making a particular estimator consistent. Less work exists on this question, but, a similar technique which also relies on $P(S = 1|\mathbf{v})$ can be applied.

When dealing with distributions affected by selection bias, our target distribution is $P(\mathbf{v})$ and we have access to $P(\mathbf{v}|S = 1)$. Then, Bayes rule gives that (Cortes et al., 2008):

$$P(\mathbf{v}) = \frac{P(\mathbf{v}|S = 1)P(S = 1)}{P(S = 1|\mathbf{v})} = P(\mathbf{v}|S = 1)\frac{P(S = 1)}{P(S = 1|PA_S)}$$

So, with knowledge of $P(S = 1|PA_S)$ and $P(S = 1)$ we can fully recover $P(\mathbf{v})$ by again multiplying by the target distribution.

So clearly knowing $P(S = 1)$ and $P(S = 1|x)$ is quite powerful. As with Heckman, this method requires information about the selection mechanism, but doesn't assume population level information about the other variables. This is a real difference between inverse probability weighting or the Heckman correction and the graphical approaches we examine next. However, unlike Heckman, inverse probability weighting does not require assumptions about the parametric structure of the data and instead works with non-parametric values.

4.5 Getting Graphical

By the first decade of the 2000's, Pearl's causality framework had become well-known not just in computer science but in applied fields as well. Accordingly, this period marks the first attempts to formulate selection bias in graphical terms. Specifically, we discuss two influential papers that drawn on Pearl's work.

4.5.1 Hernán et. al

Miguel Hernán's 2004 paper "A Structural Approach to Selection Bias" sets out to distinguish between selection bias and other problematic features of studies using the logic of causality. Although some other work had used DAGs to represent selection into a study (Robins, 2001), (Pearl, 1995), Hernán's paper went further in that it gave an explicitly graphical interpretation of presence bias caused by selection. Hernán, an epidemiologist, is particularly concerned with presence of selection bias within case-control and cohort studies, and his paper proceeds primarily by drawing on examples of such studies in which selection bias is a problem. As is often the case in medical studies, the random variables considered are mostly binary, meaning that instead of looking for conditional distributions in general, related quantities such as risk ratios or odds ratios are considered. However, the concepts are very similar.

The core of Hernán's argument is that selection bias occurs when we *condition on common effects* of the exposure (treatment) and the outcome through selection. That is, the effect of the treatment X on the outcome Y is affected by selection bias when both selection and the treatment depend on X or an ancestor (cause) of X. In Barenboim's 2014 paper, we saw a version of this statement in theorem 1: $P(y|\mathbf{x})$ is recoverable as $P(y|\mathbf{x}, S = 1)$ when $Y \perp\!\!\!\perp_G S|X$. By the definition of d-separation, This definition can be contrasted with confounding, which has its roots in common *causes*. Hernán gives an simple example illustrating what is meant by this which we reproduce.

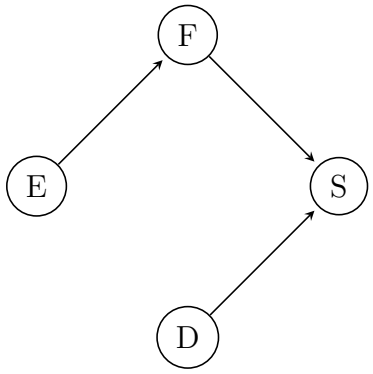


Figure 4.2: Study design for the Hernń's example

Here, E is the treatment, estrogen supplements, F denotes women who fractured their hips, and D represents having a heart attack. We want to know if estrogen use increases the risk of having a heart attack. The study participants were gathered in such that the controls, the women who did not have a heart attack, were disproportionately women who had fractured their hips. This could be because the sample was taken from a particular hospital. Since estrogen use decreases a woman's risk of breaking her hip, this means that women taking estrogen are underrepresented within the controls. So, if we look at $P(d|e, S = 1)$ then we will find a positive relationship between F and E since selection into the study is a caused by both estrogen use (indirectly through decreased hip factors) and having a heart attack. So by selecting our sample in the way that we have, we have conditioned on a common effect of both the treatment and the outcome, biasing our inferences. In the reader's digest survey,

Although Hernán's work was important in categorizing selection bias through graph structure, the paper was not primarily concerned with correcting for selection bias. There is a short section on how to overcome such bias, but it was limited to inverse probability weighting, a technique we have already discussed. Fortunately, the next paper we examine does.

4.5.2 Geneletti et. al

Following up on Hernán's work characterizing selection bias, another group of epidemiologists, Geneletti et. al give a method for correcting selection bias in retrospective case-control studies using DAGs and conditional independence (Geneletti et al., 2008). In spirit this is very similar to Bareinboim's work but limited to a specific study design such that further assumptions can be leveraged. We now briefly review the structure of a retrospective case-control study.

Unlike experiments where treatments are assigned randomly and outcomes measured after treatment, or cohort studies where individuals are tracked over a long period to see if they develop a disease, retrospective case-control studies (often just called case-control studies) select individuals for sample inclusion *after* they have contracted a disease (Woodward, 1999). Two groups are selected: one which has a particular disease/condition and one which does not. Then, covariates are gathered for both groups, with the hope of discovering some

factors which are more common in one group that the other. So for instance, we could gather a sample of lung cancer patients and another of adults without lung cancer. We would likely discover that after adjusting for relevant demographic factors (age, sex, SES, etc.) smoking was more common in the lung cancer group and then construct an estimate of the risk ratio for smoking and lung cancer. However, this kind of study is particularly susceptible to selection bias (Woodward, 1999) (Geneletti et al., 2008). This can happen for different reasons, but one of most prominent is that the treatment (smoking, in our example) might cause patients to be admitted to the hospital for a reason unrelated to lung cancer, but upon their admission be discovered to have lung cancer. In this way, we might see a dependence between lung cancer and smoking that is caused by smoker's greater likelihood to be diagnosed. Similarly, when both cases and controls are selected from the hospital Berkson's bias could create a false negative association (Hernán et al., 2004).

Much like the first theorem of Bareinboim et al. (2014), Geneletti goes about recovering a conditional distribution by constructing a "bias breaking" set that d-separates (though they do not use the term) the selection node from treatment. The paper uses a similar Hernán's, concerning the effect of estrogen use on an unspecified outcome Y where vaginal bleeding B is a symptom of estrogen use and a cause of selection.

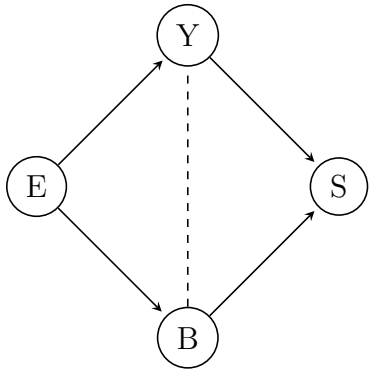


Figure 4.3: Study design for the Geneletti's example. The undirected dashed line indicates that the edge may or may not be present and could be oriented in either direction.

As is common in medical studies, the goal is to estimate the "odds ratio", $\frac{P(E=1|Y=1)P(E=0|Y=0)}{P(E=1|Y=0)P(E=0|Y=1)}$, which can be derived directly from the conditional $P(e|y)$. The authors write that when E and S are associated (i.e. $E \not\perp\!\!\!\perp_G Y$) bias can be introduced to the this odds ratio estimated using $p(y|e, S = 1)$. However, as we can see from the graph, $E \perp\!\!\!\perp_G S|(Y, B)$, or in the authors' words, (Y, B) is a "bias breaking set". They further assume that $P(b|y)$ is known as additional data, and therefore,

$$P(e|y) = \sum_b P(e|y, b)P(b|y) = \sum_b P(e|y, b, S = 1)P(b|y)$$

Although they do not attempt to generalize their result to more general graphical structures, it is worth noting that this result is closely related to the external data recovery theorem in (Bareinboim et al., 2014).

Chapter 5

Simultaneous Selection Bias and Missing Data

In Chapters 2 and 3 we discussed the contributions of two groups of researchers on the related topics of graphical characterizations of selection bias and graphical characterizations of missing data. Although these literatures share terminology, approach, and even authors, there has been no effort to date to apply the machinery of graphical models to cases in which both selection bias and missing data are present. We feel that this is a gap in the research that should be filled.

As we have argued in the previous chapters, missing data and selection bias are fundamentally related. In both cases we have some available data which doesn't quite match with what the true population data would look like. Indeed, many of the situations in which selection bias is a major concern are also canonical cases for missing data. For instance, almost any survey must contend with selection bias: both the people who are approached for the survey and the people who when approached actually respond are likely to differ from those who are not. In many observational medical studies, the study group is taken from a hospital. This can present a couple problems: Berkson's bias means that such a sampling procedure can create spurious negative correlations between conditions, and in the United States, hospitalized people are likely to be higher income than those with similar conditions who are not in the hospital. Often, these studies observe units over a long period of time, leading to one of the other classic forms of missing data: study drop-out (censoring). Clearly, it would be desirable to have a framework which describes these related situations in unified terms.

Fortunately, the existing graphical approaches to selection bias and missing data are highly compatible with each other.

5.1 Basics

In fact, we can define a selection-missingness graph (sm-graph) exactly as we would want. Just as in the selection or missingness graphs, we define recoverability, sticking close to the definitions in Mohan et al. (2013) and (Bareinboim et al., 2014).

Definition 5.1.1 (sm-recoverability)**.** *A relation Q (such as a conditional or marginal prob-*

ability) is recoverable in G if and only if Q can be expressed in terms of the observed distribution $P(V_m^*, V_o, \mathbb{R}|S=1)$. Formally, for every two probability distributions P_1 and P_2 compatible with G, $P_1(V_m^*, V_o, \mathbf{R}|S=1) = P_2(V_m^*, V_o, \mathbf{R}|S=1)$ implies that $P_1(Q) = P_2(Q)$.

5.2 Easy Facts

This set-up lends itself to a handful of obvious but nonetheless potentially useful facts about recovery from an sm-graph. First of all, as we discussed in the section on complete-case analysis, "deleting" every row which has missing values can be thought of as imposing an additional selection condition on the data generating process. This gives us a simple sufficient (but far from necessary) condition for recovering a distribution. The graphical interpretation of this fact is encoded in the following definition.

Theorem 5. *In an sm-graph G with no edges between the R variables and S, no edges between a variable and its associated missingness node, and no collider paths, the necessary and sufficient condition for recovering $P(y|\mathbf{x})$ is that $Y \perp\!\!\!\perp_G S|\mathbf{X}$.*

Proof $\implies$ We first follow the proof of theorem 3 from Mohan & Pearl (2019) to show that $P(v|S=1)$ is recoverable. We have that:

$$\begin{aligned} P(\mathbf{v}, \mathbf{R}=0|S=1) &= P(\mathbf{v}|S=1)P(\mathbf{R}=0|\mathbf{v}, S=1) \\ &= P(\mathbf{v}|S=1)P(\mathbf{R}=0|\mathbf{v}) \qquad\qquad = P(\mathbf{v}|S=1)\prod_i P(R_i=0|pa_{R_i}) \end{aligned}$$

Giving $P(\mathbf{v}|S=1) = \frac{P(\mathbf{v},\mathbf{R}=0|S=1)}{\prod_i P(R_i=0|pa_{R_i})} = \frac{P(\mathbf{v},\mathbf{R}=0|S=1)}{\prod_i P(R_i=0|pa_{R_i}, R_{PA_{R_i}})}$, much as in the original. The only difference comes in the second equality. This is immediate from $\mathbf{R} \perp\!\!\!\perp S|\mathbf{v}$, which is true by the assumption that there are no edges between S and and R node. So, since $P(\mathbf{v}|S=1)$ is recoverable, we can ignore the missingness and directly apply the result from (Bareinboim et al., 2014).

A more useful idea comes from considering that we can essentially treat the missingness and selection mechanisms as separate. That is, if we can recover the joint from missing data when selection is not present, i.e. $P(v_m, v_o)$, then we can recover $P(v_m, v_o|S=1)$ in the sm-graph. This means that in such cases we then look to the conditions for recovery in the selection graph while ignoring the missingness mechanisms.

Theorem 6. *In an sm-graph G with no edges between the R variables and S and such that $PA_S \subset V_o$, a quantity Q derivable from $P(v_o, v_m)$ is recoverable if and only if Q is s-recoverable in $G\backslash(\mathbf{R} \cup V_m^*)$ and Q is m-recoverable in $G\backslash S$.*

Proof (If) Assume that Q is s-recoverable in $G\backslash(\mathbf{R} \cup V_m^*)$ and Q is m-recoverable in $G\backslash S$. First we decompose

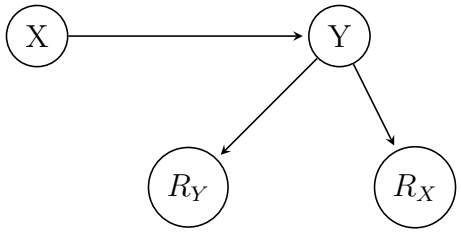


Figure 5.1: Case in which $P(X)$ is not recoverable

$$
\begin{aligned}
P(v_m^*, v_o, \mathbf{r}|S=1) &= P(v_m^*, v_o, |\mathbf{r}, S=1)P(\mathbf{r}|S=1) \\
&= \frac{P(\mathbf{r}, S=1|v_m^*, v_o)P(v_m^*, v_m^*)P(\mathbf{r}|S=1)}{P(\mathbf{r}, S=1)} \\
&= \frac{P(S=1|v_m^*, v_o)P(\mathbf{r}|v_m^*, v_o)P(v_m^*, v_m^*)P(\mathbf{r}|S=1)}{P(\mathbf{r}, S=1)} \\
&= \frac{P(S=1|v_o)P(\mathbf{r}, v_m^*, v_o)P(\mathbf{r}|S=1)}{P(\mathbf{r}, S=1)} \\
&= \frac{P(v_o|S=1)P(S=1)P(\mathbf{r}, v_m^*, v_o)P(\mathbf{r}|S=1)}{P(v_o)P(\mathbf{r}|S=1)P(S=1)} \\
&= \frac{P(v_o|S=1)P(\mathbf{r}, v_m^*, v_o)}{P(v_o)}
\end{aligned}
$$

Consider two probability distributions P_1, P_2 with $P_1(v_m^*, v_o, \mathbf{r}|S=1) = P_2(v_m^*, v_o, \mathbf{r}|S=1)$. By the above,

(Only if) Without loss of generality, assume that Q is not s-recoverable in $G\backslash(\mathbf{R} \cup V_m^*)$. Then there exist two probability distributions P_1 and P_2 such that $P_1(v_m, v_o|S=1) = P_2(v_m, v_o|S=1)$ but $P_1(Q) \neq P_2(Q)$. Then, define P_1', P_2' as $P_1'(v_m, v_o, \mathbf{r}|S=1) = P_1(v_m, v_o|S=1)P(\mathbf{r})$ and $P_2'(v_m, v_o, \mathbf{r}|S=1)P_2(\mathbf{r})$ are compatible with G (when all edges to and from $\mathbf{R} \cup V_m^*$ are inactive), we have $P_1()$

Theorem 7. *A distribution $P(X)$ is not m-recoverable if there exists $Y \in PA_{R_X}$ such that Y is self-censoring ($Y \in PA_{R_Y}$) and Y is a neighbor of X.*

Proof (Similar to Mohan et al. (2013) section 5.5)The proof requires a lemma proven in Mohan et al. (2013), copied below.

With this result in mind, consider figure 5.1.

The lemma gives that if $P(X)$ is not recoverable in the graph (G) displayed in figure 5.1, $P(X)$ is non-recoverable in *any* missingness graph which contains G as a subgraph. Then, it is sufficient to provide a direct counter-example. We construct two probability distributions, P_1, P_2 compatible with G such that P_1, P_2 agree on the observed distributions but $P_1(x) \neq P_2(x)$. Since there is only one d-separation in G ($X \perp\!\!\!\perp (R_X, R_Y)|Y$) the only condition that must hold for compatibility is that $P_i(x, y, r_x, r_y) = P(x|y)P(y, r_x, r_y)$.

Claim: A marginal distribution $P(\mathbf{X})$ can be recovered in m-graph G if and only if $P(\mathbf{X})$ can be recovered in the m-graph G' formed by deleting any self-censoring node Y and its

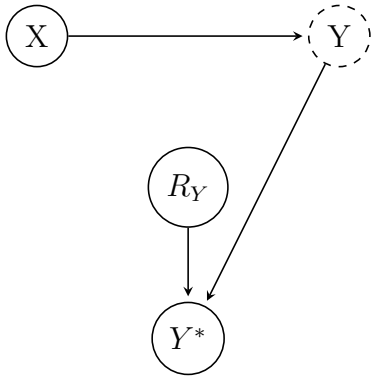


Figure 5.2: "Hard" MNAR Case

associated R_Y and redrawing replacing every edge from Y to a missingness node with an edge from each parent and child of Y in $V_m \cup V_o$ to that missingness node.

$\implies$ We prove the condition in three cases.

Case 1: Y has no parents and no children in $V_m \cup V_o$. In this case, $Y \perp\!\!\!\perp_G (V_m \cup V_o \backslash Y) | \emptyset$

5.3 External Data Recovery

In overcoming selection bias, access to population distributions for a subset of the variables has been shown to be useful. In the graphical methods outlined in chapter 2, separate graphical conditions were developed for recovery. Older methods for selection bias adjustment such as survey weighting also depend on knowing population distributions. However, none of the recent graphical work on missing data has addressed the utility of external data in that context. Intuitively, we can imagine it might: if data for one variable is missing, it makes sense that knowing what that data *would* look like if it were not subject to missingness. In fact, this is true. Consider the following example:

Example 6. *In figure 5.1, we display what is sometimes called a "hard" case of MNAR missingness: Y contributes to its own missingness (self-censoring). Without external data, we could neither recover $P(x, y)$ or $P(y|x)$. We could recover $P(x|y)$ as $P(x|y) = P(x|y^*, R_Y = 0)$. However, if we assume that we know $P(y)$ then we simply apply Bayes' rule to get:*

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

We can also recover the joint $P(x, y)$ (and therefore the conditional) directly as $P(x|y)P(y)$.

Given that recovery under missingness is sometimes possible with external data, we want to establish conditions which describe all cases in which recovery must be possible. We will describe such conditions for recovering both joint distributions and conditionals with or without external data.

Theorem 8. *A conditional distribution $P(y|x)$ with $X, Y \in V_o \cup V_m$ is recoverable if and only if at least one on of the following statements holds:*

1. $Y \in Vo$ *and*

Bibliography

Bareinboim, E., & Pearl, J. (2012). Controlling selection bias in causal inference. In N. D. Lawrence, & M. Girolami (Eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, vol. 22 of *Proceedings of Machine Learning Research*, (pp. 100–108). La Palma, Canary Islands: PMLR. http://proceedings.mlr.press/v22/bareinboim12.html

Bareinboim, E., Tian, J., & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. *Proceedings of the National Conference on Artificial Intelligence*, *4*, 2410–2416.

Bishop, C., Leite, W., & Snyder, P. (2018). Using propensity score weighting to reduce selection bias in large-scale data sets. *Journal of Early Intervention*, *40*, 105381511879343.

Bushway, S., Johnson, B. D., & Slocum, L. A. (2007). Is the magic still there? the use of the heckman two-step correction for selection bias in criminology. *Journal of Quantitative Criminology*, *23*(2), 151–178. https://doi.org/10.1007/s10940-007-9024-4

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. In Y. Freund, L. Gyorfi, G. Turán, & T. Zeugmann (Eds.), *Algorithmic Learning Theory*, (pp. 38–53). Berlin, Heidelberg: Springer Berlin Heidelberg.

Delgado-Rodríguez, M., & Llorca, J. (2004). Bias. *Journal of Epidemiology & Community Health*, *58*(8), 635–641. https://jech.bmj.com/content/58/8/635

Donders, R., van der Heijden, G., Stijnen, T., & Moons, K. (2006). Review: A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*, 1087–91.

Fienberg, S. E., & Tanur, J. M. (1996). Reconsidering the fundamental contributions of fisher and neyman on experimentation and sampling. *International Statistical Review / Revue Internationale de Statistique*, *64*(3), 237–253. http://www.jstor.org/stable/1403784

Geneletti, S., Richardson, S., & Best, N. (2008). Adjusting for selection bias in retrospective, case–control studies. *Biostatistics*, *10*(1), 17–31. https://doi.org/10.1093/biostatistics/kxn010

Guo, S., & Fraser, M. (2015). *Propensity Score Analysis*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications. https://books.google.com/books?id=V3c2ngEACAAJ

Haneuse, S., Schildcrout, J., Crane, P., Sonnen, J., Breitner, J., & Larson, E. (2009). Adjustment for selection bias in observational studies with application to the analysis of autopsy data. *Neuroepidemiology*, *32*, 229–39.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161. http://www.jstor.org/stable/1912352

Hernán, M., Hernández-Díaz, S., & Robins, J. (2004). A structural approach to selection bias. *Epidemiology (Cambridge, Mass.)*, *15*, 615–25.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446

Little, R. J. A., & Rubin, D. B. (1986). *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc.

Lusinchi, D. (2012). "president" landon and the 1936 literary digest poll: Were automobile and telephone owners to blame? *Social Science History*, *36*(1), 23–54.

Mohan, K., & Pearl, J. (2019). Graphical models for processing missing data.

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, (p. 1277–1285). Red Hook, NY, USA: Curran Associates Inc.

Pearl, J. (1985). *Bayesian Networks: A Model of Self-activated Memory for Evidential Reasoning*. Report. UCLA, Computer Science Department. https://books.google.com/books?id=1sfMOgAACAAJ

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. http://www.jstor.org/stable/2337329

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

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. vol. 1.

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, *12*(3).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. http://www.jstor.org/stable/2335739

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Seaman, S., & White, I. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, *22*.

Squire, P. (1988). Why the 1936 literary digest poll failed. *The Public Opinion Quarterly*, *52*(1), 125–133. http://www.jstor.org/stable/2749114

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, *18*(1), 327–350. https://doi.org/10.1146/annurev.so.18.080192.001551

Woodward, M. (1999). *Epidemiology: Study Design and Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. https://books.google.com/books?id=0qjB1WBrjjEC

Appendix

.1 Heckman

.2 Deriving the Bias

Heckman is viewing selection bias as being the result of an omitted variable which we can obtain. Here we give the derivation of what this variable actually is, closely following the original paper but with some clarifications and changes to notation as in (**?**). We know that:

$$E[Y|\mathbf{X}, S = 1] = E[\mathbf{X}\boldsymbol{\beta} + \epsilon|\mathbf{X}, R_Y = 0] = \mathbf{X}\boldsymbol{\beta} + E[\epsilon|\mathbf{X}, R_Y = 0].$$

Our independence assumption gives that $E[\epsilon|\mathbf{X}, S = 1] = E[\epsilon|R_Y = 0]$ and so this is the quantity we wish to estimate. Now, $S = 1 \implies \mathbf{W}\boldsymbol{\delta} + \tau > 0$ and we write:

$$E[\epsilon|\mathbf{X}, R_Y = 0] = E[\epsilon|\mathbf{W}\boldsymbol{\delta} + \tau > 0] = E[\epsilon|\tau > -\mathbf{W}\boldsymbol{\delta}].$$

Since (ϵ, τ) follows a bivariate normal distribution and each have a marginal mean of 0, the conditional expectation $E[\epsilon|\tau] = \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau$ where ρ is the correlation between ϵ and τ and σ_ϵ^2, σ_τ^2 are the respective variances. Let $\gamma = \rho\frac{\sigma_\epsilon}{\sigma_\tau}$ giving $E[\epsilon|\tau] = \gamma\tau$. Notice that when $\rho = 0$ (meaning ϵ and τ are independent), the γ term is equal to 0 and therefore the selection does not bias the estimate. We may then apply the law of total expectation (sometimes call Adam's law):

$$E[\epsilon|\tau] = E[E[\epsilon|\tau]|\tau] = E[\gamma\tau|\tau] = \gamma E[\tau|\tau]$$

So, when $\tau > -\mathbf{W}\boldsymbol{\delta}$, $E[\tau|\tau] = E[\tau|\tau > -\mathbf{W}\boldsymbol{\delta}]$ and we have that:

$$E[\epsilon|\tau > -\mathbf{W}\boldsymbol{\delta}] = \gamma E[\tau|\tau > -\mathbf{W}\boldsymbol{\delta}]$$

We recognize this as the expectation of a truncated normal, which is given by the inverse Mills ratio $E[X|X > x] = \sigma\frac{\phi(\frac{x-\mu}{\sigma})}{1-\Phi(\frac{x-\mu}{\sigma})}$ for $X \sim N(\mu, \sigma^2)$. As usual, ϕ is the density function for the standard normal and Φ is the distribution function. By assumption, $\mu = 0$. Therefore,

$$E[\epsilon|\tau > -\mathbf{W}\boldsymbol{\delta}] = \gamma\sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1 - \Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}.$$

So the bias term we want to estimate is $\gamma\sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$. This is done in several parts Heckman (1979).

Something to note: unlike the approach we see in the more contemporary selection bias literature, Heckman's method asks for external information on the selection mechanism itself, not the variables being measured. In addition to this, Heckman makes very strong parametric assumptions. Sometimes when such assumptions are made, the method is still relatively robust to (mild) violations. However, Heckman's correction's does not generally have this property, and violations of the set-up can cause seriously biased estimates (Little & Rubin, 1986). This lack of robustness has led to Heckman's method becoming less popular in recent years (Bushway et al., 2007).

.2.1 Examples

Heckman developed his technique in the context of economic research, and its implementation typically relies on further assumptions justified by economic theory. Particularly, Z is often understood as an *unmeasured* continuous random variable following the linear regression assumptions (Winship & Mare, 1992). Sometimes, this variable is not directly measurable - such as an individual propensity or ability - that is nonetheless assumed to exist. We quickly review a number of fairly standard examples used to introduce Heckman's method (Heckman, 1979), (Guo & Fraser, 2015).

1. Estimating the wages women not in the workforce would make if they entered the workforce using data for women currently in the workforce. Selection occurs because the women in the workforce differ from those who are not.

2. Estimating the effect of unionization of worker wages - similar to the last example, unionized workers might have joined the union because they were unsatisfied with their wages pre-unionization.

3. Estimating effects of schooling/training programs of worker productivity. Once again, the problem is that the people who choose to participate in such programs cannot be used to represent the workers who did not make that

In the spirit of some of these examples, we describe a relevant variation in more detail. Suppose that we are working with Reed College academic services to measure the effect of tutoring on student homework grades in the introductory chemistry sequence. We have collected a collected a set of relevant covariates ($\mathbf{X}$) on all students in the course such as major, high school grades, basic demographics, etc. The response variable (Grade Change) is the difference between the student's average homework scores before their first tutoring appointment and the average score after. Since some students do not seek tutoring, their value of Y is missing. We assume that the response variable follows the linear regression assumptions:

$$\text{Grade Change} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

For normally distributed ϵ with mean 0. As students who participate in tutoring might differ from students who do not, it would be incorrect to use the model created using data for students that did seek tutoring to estimate the impact that tutoring would have on students

who did not chose to be tutored. So, we must create a model for a student's probability of going to tutoring. To do this though Heckman's technique, we must further assume that a student seeks tutoring a when some unmeasured value crosses a threshold. There are multiple options here, and in practice this a stage that requires existing economic theory to be justified. In our case, these details are not important we assume that Tutoring Value is, for each student, the perceived value of going to tutoring minus the students value of an equal amount of time for a different activity, such that a student goes to tutoring if Tutoring Value is positive. We have further collected from each student covariates capturing their motivation to do well in their courses, their perception of the usefulness of tutoring, and the amount of free time they have. Together with the covariates we measured for X, these variables constitute $\mathbf{W}$, and again we require linear regression assumptions are satisfied:

$$\text{Tutoring Desire} = \mathbf{W}\boldsymbol{\delta} + \tau$$

With τ being normally distributed with mean 0 such that ϵ and τ are jointly bivariate normal with correlation ρ. We can then implement the correction as follows:

1. Define the selection variable S as 1 for student with data for Grade Change and 0 for those who do not.
2. Create a probit model $P(S = 1|W)$ which yields as estimate of $\frac{\boldsymbol{\delta}}{\sigma_\tau}$ as the coefficients on $\mathbf{W}$.
3. We then estimate the inverse Mill's ratio $\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ and include the term in the regression equation, giving consistent estimates of $E[\text{Grade Change}|\mathbf{X}]$.

.2.2 Conditional Expectation of Bivariate Normal Distribution

We want to show that for $(\epsilon, \tau) \sim N((0,0), \mathbf{\Sigma})$ (where $\mathbf{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & \rho\sigma_\epsilon\sigma_\tau \\ \rho\sigma_\epsilon\sigma_\tau & \sigma_\tau^2 \end{pmatrix}$ and ρ is the correlation of ϵ and τ) the conditional $E[\epsilon|\tau] = \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau$. The normal distribution has the property that the covariance (therefore correlation) of determines the dependence structure. Consider the random variable $\epsilon - \frac{\rho}{\sigma_\tau}\tau$. We want to show that this random variable is independent of τ. This is true when their covariance is 0, i.e. when $E[\tau(\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau)] - E[\tau]E[\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau] = 0$. In fact, since $E[\tau] = 0$ by assumption, we want to show that $E[\tau(\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau)] = 0$

We have:

$$\begin{aligned} E[\tau(\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau)] &= E[\tau\epsilon] - \rho\frac{\sigma_\epsilon}{\sigma_\tau}E[\tau^2] \\ &= \rho\sigma_\tau\sigma_\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}E[\tau^2] \\ &= \rho\sigma_\tau\sigma_\epsilon - \rho\sigma_\tau \\ &= 0 \end{aligned}$$

As desired. Then, applying independence, we have that:

$$\begin{aligned}
E[\epsilon|\tau] &= E[\epsilon - \frac{\rho}{\sigma_\tau}\tau + \frac{\rho}{\sigma_\tau}\tau|\tau] \\
&= E[\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau|\tau] + E[\rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau|\tau] \\
&= E[\epsilon - \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau] + \rho\frac{\sigma_\epsilon}{\sigma_\tau}E[\tau|\tau] \\
&= E[\epsilon] + \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau \\
&= \rho\frac{\sigma_\epsilon}{\sigma_\tau}\tau.
\end{aligned}$$

.2.3 Truncated Normal and the Inverse Mills Ratio

We show that $E[\tau|\tau \geq t] = \sigma_\tau\frac{\phi(t)}{1-\Phi(t)}$. By the definition of truncated density, the truncated PDF $f_{\tau|\tau>t}$ is:

$$f_{\tau|\tau>t}(x) = \frac{f_\tau(x)}{1 - F_\tau(t)} \text{ for } x > t$$

Where f is the density function for τ and F is the distribution function for τ. In particular, these are $f(x) = \frac{1}{\sigma_\tau}\phi(\frac{x}{\sigma_\tau})$ and $F(x) = \Phi(\frac{x}{\sigma_\tau})$ giving $f_{\tau|\tau>t}(x) = \frac{\phi(\frac{x}{\sigma_\tau})}{\sigma_\tau(1-\Phi(\frac{t}{\sigma_\tau}))}$ for $x > t$. To find the expected value, we integrate.

$$\begin{aligned}
E[\tau|\tau > -t] &= \int_t^\infty x f_{\tau|\tau>t}(x) \\
&= \int_t^\infty \frac{\phi(\frac{x}{\sigma_\tau})}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))} \\
&= \frac{1}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))}\int_t^\infty x\phi(\frac{x}{\sigma_\tau}) \\
&= \frac{1}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))}\int_t^\infty \frac{x}{\sigma_\tau\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma_\tau^2}} \\
&= \frac{1}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))}\left(-\frac{\sigma_\tau e^{-\frac{x^2}{2\sigma_\tau{}^2}}}{\sqrt{2\pi}}\Big|_t^\infty\right) \\
&= \frac{1}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))}\frac{\sigma_\tau e^{-\frac{t^2}{2\sigma_\tau{}^2}}}{\sqrt{2\pi}} \\
&= \frac{\sigma_\tau^2\phi(\frac{t}{\sigma_\tau})}{\sigma_\tau(1 - \Phi(\frac{t}{\sigma_\tau}))} \\
&= \sigma_\tau\frac{\phi(\frac{t}{\sigma_\tau})}{(1 - \Phi(\frac{t}{\sigma_\tau}))}
\end{aligned}$$

Which is the desired result.