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

Chapter 1: The First **3**
1.1 Background: Graphs, D-separation, Causality 3
1.1.1 Directed Acylic Graphs 3
1.1.2 Bayesian Networks 4
1.1.3 D-Separation 4
1.1.4 Structural Causal Models 6

Chapter 2: Bareinboim et al 2014 . . . . . **9**
2.1 Introduction 9
2.2 Representing Selection with DAGs 9
2.3 A Famous Case of Selection Bias 10
2.4 Recovering $P(y|x)$ Without External Data 10
2.5 Recovery With External Data 12
2.6 A Useful Extension 13

Chapter 3: Earlier Approaches to Selection Bias **15**
3.1 Old Stuff 15
3.2 The Heckman Correction 15
3.2.1 Premise 16
3.2.2 Implementing the Correction 16
3.3 Probability Weighting 18
3.4 Getting Graphical 18
3.4.1 Hernán et. al 19
3.4.2 Geneletti et. al 20

Bibliography **23**

Appendix **25**
.1 Heckman 25
.1.1 Conditional Expectation of Bivariate Normal Distribution 25
.1.2 Truncated Normal and the Inverse Mills Ratio 25

List of Tables

List of Figures

1.1 Graphs corresponding to $V = \{A, B, C\}$, $E = \{(A, B), (B, C)\}$ 3
1.2 Cyclic (left) and acyclic (right) graphs on the nodes A, B, C 4

2.1 Graph corresponding to selection bias based on income 11
2.2 Graph corresponding to selection bias based on income and party registration 12

3.1 Basic DAG for inverse probability weighting example 18
3.2 Study design for the Hernń's example 19
3.3 Study design for the Geneletti's example 21

Abstract

The preface pretty much says it all.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

The First

1.1 Background: 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), [ref Pearl 1996] 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 edge $(i, j) \in E$ 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**.

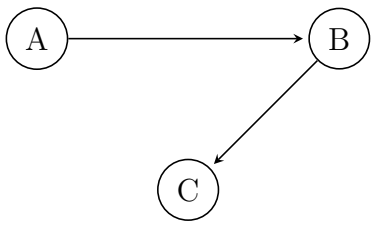


Figure 1.1: Graphs corresponding to $V = \{A, B, C\}$, $E = \{(A, B), (B, C)\}$

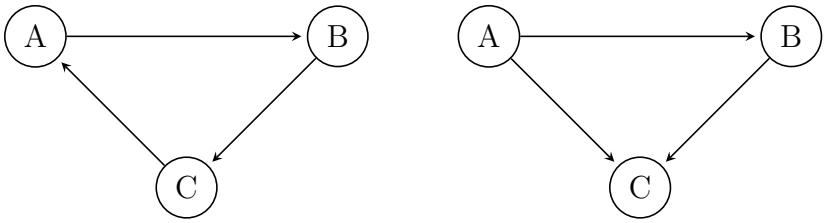


Figure 1.2: Cyclic (left) and acyclic (right) graphs on the nodes A, B, C

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 Bayesian Networks

Before we do so, we review some basic definitions from probability.

Definition 1.1.4. *For random variables $X_1, X_2, ..., X_n$, the* ***joint distribution*** *with associated joint probability function $P(x_1, x_2, ..., x_n)$ gives the probability of every possible combination of values for $X_1, X_2, ..., X_n$. For any subset $\{Y_1, ..., Y_k\} \subseteq \{X_1, X_2, ..., X_n\}$, the probability distribution for $Y_1, Y_2, ..., Y_k$, is called the* ***marginal distribution****. Finally, for a subset*

1.1.3 D-Separation

Graphical structure can allow us to reason about causal and statistical relationships. One of the essential tools is d-separation. 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. We extend this notion [ref Pearl 2008], [ref Peters 2017].

Definition 1.1.5. *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***:* $v_{j-1} \leftarrow v_j \rightarrow v_{j+1}$*, or*

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

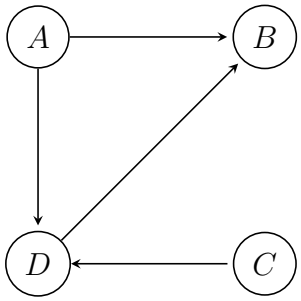
Example 1. *Consider a path* $A \rightarrow B \rightarrow C \leftarrow D \rightarrow E$ *between nodes* A *and* E*. There are many sets that block this path. For one, there is a collider on the path. This means that any set which does not include* C *or any elements of* DE_C *would successfully block 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 1.1.6. *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 2. *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 $A \rightarrow B$ 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, $A \rightarrow B \leftarrow D \leftarrow C$ and $A \rightarrow D \leftarrow C$. 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 $B \leftarrow A \rightarrow D \leftarrow C$. 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 $B \perp\!\!\!\perp_G C|D, A$.

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.

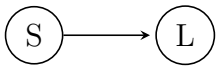
A paradigmatic example of such a controversy is over the relationship between cigarette smoking and lung cancer. Today, it is uncontroversial that smoking cigarettes (especially habitually over a long period of time) increases a person's risk of contracting lung cancer. This is a causal relationship, 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 [ref Peters] 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:



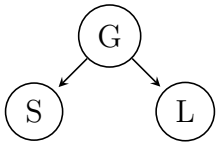
However, historically, this causal structure was not the only one suggested to explain the association between lung cancer and smoking. One particularly illustrious detractor of this theory was British statistician R. A. Fisher, a foundational figure in modern statistics (and a heavy smoker) [ref Fisher]. 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 1.1.7 (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**.

[This seems important enough to reproduce a proof but I will have to spend a little more time than I have tonight]

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.

Chapter 2

Bareinboim et al 2014

2.1 Introduction

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. However, the content of the paper is highly abstracted and does not attempt to provide real-life cases in which their techniques are useful. For this reason, we will summarize the main results of the paper and interpret them in the context of some realistic examples.

The paper 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, and the last gives a brief discussion of recovering causal effects.

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 biased distribution is then $P(\mathbf{v}|S = 1)$, that is, the joint for our variables $\mathbf{V}$ conditioned on inclusion within the sample.

2.3 A Famous Case of Selection Bias

One classic example of selection bias is the 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 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, this explanation 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 falls under the selection bias framework proposed here.

For now, though, let's assume that income was the only variable that was responsible for the non-representative sample, and that this information was collected from respondents as part of the survey. However, we will return to this case later to add some depth.

One of the important features of selection bias is that unlike sampling variation, it cannot be reduced with sufficiently large sample sizes. For this reason, in the rest of the paper we will discuss a selection biased joint distribution for this survey, and thereby ignore sample size.

2.4 Recovering $P(y|x)$ Without External Data

In the simplest 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. This means that we do not have access to the full joint distribution $P(\mathbf{v})$ but rather the conditional joint distribution $P(\mathbf{v}|S = 1)$. This corresponds to the joint distribution of values that are selected to be sampled. So, returning to our election example, our graph would look like: Where C is a random variable for the candidate a person intends to vote for, W is measures income and S is a binary variable denoting inclusion in the study. 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.

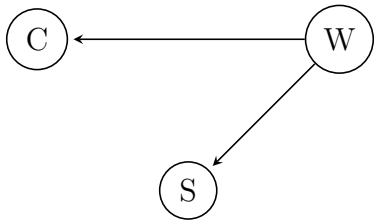


Figure 2.1: Graph corresponding to selection bias based on income

In the 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 under selection bias. We reproduce their definition exactly and give an explanation of why it makes sense.

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 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 $P_1(y|\mathbf{x}) = P_2(y|\mathbf{x})$.*

If we have a selection biased distribution, we have access to a particular joint distribution, $P(\mathbf{v}|S = 1)$ where $\mathbf{V}$ is the set of all nodes other than S. So if we are to recover $P(y|\mathbf{x})$ ($\mathbf{X} \subset \mathbf{V}$, $Y \in \mathbf{V}$) we need to express $P(y|\mathbf{x})$ as values derivable from $P(\mathbf{v}|S = 1)$ 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.

Consider that 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 $P(y|\mathbf{x})$. The first important result of the paper is 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. □

Clearly this is a useful result since it is easy to test whether two nodes are d-separated by a set. In our application, we can see that $V \perp\!\!\!\perp_G S|W$, and so $P(v|w) = P(v|w, S = 1)$. 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.

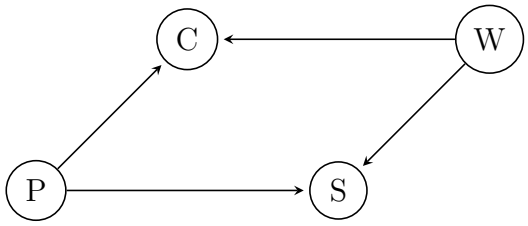


Figure 2.2: Graph corresponding to selection bias based on income and party registration

2.5 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. In fact, if candidate preference directly influences the likelihood of survey response, (i.e. $C \rightarrow S$), 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 $S \leftarrow U \rightarrow V$ 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 2.5.1. *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 $P(p|w)$.

2.6 A Useful Extension

As we have seen, Bareinboim et al. is focused on the recovery of conditional distributions, i.e. $P(v|w)$. 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 $P(v)$ is recoverable using external data. In both sections we have seen that $P(v|w)$ was recoverable. Then, assuming that we have the external data for $P(w)$, we can use the law of total probability to write:

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

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

Theorem 3. *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* $P(y) = \sum_{\mathbf{x}} P(y|\mathbf{x})P(\mathbf{x})$.

Chapter 3

Earlier Approaches to Selection Bias

In statistics and many related fields, forms of the effect we are calling "selection bias" have gone by many names. 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. 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.

In this section we will review a handful of notable earlier approaches to defining, correcting, and recognizing selection bias that have appeared within the literature. 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 (haven't really done this part yet).

3.1 Old Stuff

I want this section to be review of the really old content on selection bias (like the origin of the term and so on), but so far I have not found much.

3.2 The Heckman Correction

James Heckman's well-cited 1979 paper "Sample Selection Bias as a Specification Error" proposes a method for overcoming 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 familiar situation in which a non-representative sample is drawn from a population. 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. We now describe Heckman's method in detail.

3.2.1 Premise

Heckman's original paper is concerned with estimating 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 excess indexing) we will present the method in the context of distributions. Fortunately, the math is essentially the same.

Assume that we have access to the distributions of a dependent variable Y, a vector of regressor variables $\mathbf{X}$, and a normally distributed noise term ϵ which is independent of $\mathbf{X}$. Further assume that the linear regression assumption holds, i.e.

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

Now, assuming we have the joint distribution $P(y, \mathbf{x}, \epsilon)$ we can get β using standard techniques. However, if a selection mechanism is present, this may not be possible. Heckman's correction can be applied when the selection mechanism follows a specific form. Particularly, assume that we have access to another set of random variables that also follow the linear model pattern:

$$Z = \mathbf{W}\boldsymbol{\delta} + \tau.$$

Again, $\mathbf{W}$ is a vector of regressors, $\boldsymbol{\delta}$ is a coefficient vector and τ is normally distributed and independent of $\mathbf{W}$. Oftentimes, it is assumed that $X \subset W$, but this is not needed. We also require that the joint $P(\tau, \epsilon)$ follows the bivariate normal distribution. Heckman's correction can be applied when the selection mechanism is of the form:

$$S = I(Z > c) = I(\mathbf{W}\boldsymbol{\delta} + \tau > c).$$

For some constant c. Without losing generality, we assume that $c = 0$. So selection occurs when a second random variable, Z, passes a particular threshold.

3.2.2 Implementing the Correction

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:

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

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

$$E[\epsilon|\mathbf{X}, S=1] = E[\epsilon|\mathbf{W}\boldsymbol{\delta} + \tau > 0] = E[\epsilon|\tau > -\mathbf{W}\boldsymbol{\delta}].$$

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 $E[\epsilon|\tau] = \gamma\tau$. 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 two parts:

1. First, a linear model for $E[Z|W]$ is constructed and $\hat{\sigma}_\tau$ is estimated from the squared sum of the residuals. Then, probit model is employed to estimate $\boldsymbol{\delta}$, i.e. $P(S=1|\mathbf{W}) = \Phi(\frac{\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})$. The estimate $\hat{\boldsymbol{\delta}}$ is calculated as $\hat{\boldsymbol{\delta}} = \tilde{\boldsymbol{\delta}}\hat{\sigma}_\tau$ where $\tilde{\boldsymbol{\delta}}$ are the model coefficients of the probit model.

2. This estimated quantity is used to get an estimate of $\sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$.

3. The estimate of $\sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ is included in the corrected model $Y = \mathbf{X}\boldsymbol{\beta} + \sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})} + \epsilon$. The coefficient on $\sigma_\tau\frac{\phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}{1-\Phi(\frac{-\mathbf{W}\boldsymbol{\delta}}{\sigma_\tau})}$ is then an estimate of γ.

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), (Bushway et al., 2007).

3.3 Probability Weighting

Another technique worth mentioning is a very general approach called inverse probability weighting. Versions of the technique are old and have uses well outside of adjusting for selection bias (Horvitz & Thompson, 1952), but we give an interpretation particularly suited for this task. Given a particular (potentially non-representative) sample, the goal is to construct a 'new' sample which follows the population distribution. Doing this requires knowing something about the selection rule, particularly knowing the probability that any particular set of values is sampled Cortes et al. (2008). Inverse probability weighting admits a particularly clear interpretation using our graphical notation. Consider a simple two variable case following the DAG below.

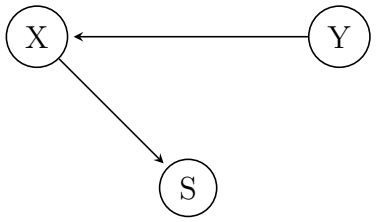


Figure 3.1: Basic DAG for inverse probability weighting example

Now, instead of assuming that we have access to the population distribution $P(x)$ as in the external data section, but rather that we know or can estimate $P(S = 1)$. One context in which this is plausible is in drop-out over the course of a longitudinal study (Hernán et al., 2004). Particularly, selection into the final analysis is the result of *not* dropping out at some point over the course of the study. If drop out does not occur at random, selection bias is likely. However, it would be possible to estimate drop-out probability by comparing the body of participants at the outset of the study to the group that remains. Since we have $P(x, y|S = 1)$, knowing $P(S = 1)$ gives $P(x, y, S = 1) = P(x, y|S = 1)P(S = 1)$ and thereby use Baye's rule to get:

$$P(x, y) = \frac{P(x, y|S = 1)P(S = 1)}{P(S = 1|x, y)} = \frac{P(S = 1)}{P(S = 1|x)}P(x, y|S = 1).$$

So clearly knowing $P(S = 1)$ 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.

3.4 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.

3.4.1 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. This 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.

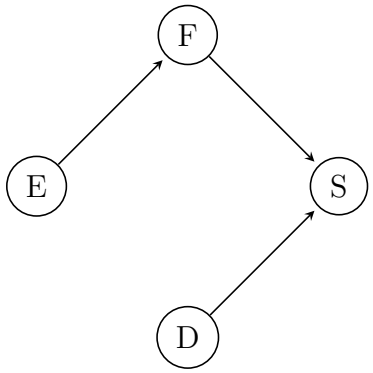


Figure 3.2: Study design for the Hernń's example

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.

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 not based in the conditional independences implied by the causal structure. Fortunately, the next paper we examine does.

3.4.2 Geneletti et. al

(This section is unfinished, it's late)

Following up on Hernán's work characterizing selection bias, another epidemiologist, Sara Geneletti and her co-authors 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 she does not use the term) the selection node from treatment. The paper uses nearly the same example as Hernán's but here

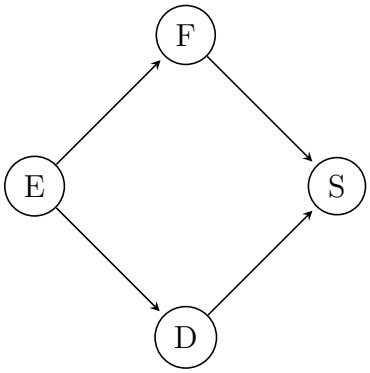


Figure 3.3: Study design for the Geneletti's example

Bibliography

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.

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

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

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.

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

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

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

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

.1.1 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:

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

.1.2 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.