When Observed Data Just Isn't Enough: Omitted Variable Based Sensitivity Analysis for Sample Selection as a Threat to Internal Validity [DRAFT]

Adam Rohde[*], Chad Hazlett[†]

December 2022

Abstract

Sample selection is a common threat to the internal validity of causal effect estimates. While Rohde and Hazlett (20XX) discusses these threats at length and provides guidance on how covariate adjustment can be use to address them, what should researchers do when observed covariates are insufficient to eliminate these threats? We show how researchers can use omitted variable based sensitivity analyses to do this. In particular, we discuss the omitted variable based sensitivity analyses of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) and how these can be leveraged to evaluate threats from sample selection. However, since sample selection as a threat to internal validity is typically the result of collider stratification, the parameters in such sensitivity analyses can be difficult to interpret. We show how more interpretable expressions for the sensitivity parameters in these frameworks can be derived in some simple, parametric settings. Using these as a guide, we also propose bounds on the parameters for general, non-parametric settings using notions from information theory. A worked example and discussion are provided.

1 Introduction

Sample selection bias is a common threat to the internal validity of causal effect estimates. Rohde and Hazlett (20XX) discuss this problem in depth and provide a comprehensive framework for evaluating these threats and whether covariate adjustment can be used to eliminate them. But what should be done when covariate adjustment using observed data is insufficient to remove the threats to internal validity from sample selection and to identify causal effects? In this paper, we show how researchers can use sensitivity analysis. Rohde and Hazlett (20XX) show that sample selection as a threat to the internal validity of causal effect estimates can be viewed as an omitted variable problem. Therefore, we can conduct omitted variable bias based sensitivity analyses when considering threats to internal validity from sample selection, when observed covariates are insufficient to identify internally valid causal effects.

Suppose an investigator is interested in estimating the effect of a treatment, D, on an outcome, Y, for the selected sample alone or for the subpopulation for which the selected sample is a representative sample. Suppose further that the investogator know that $Y_d \not\perp\!\!\!\perp D|X, S=1$ but $Y_d \perp\!\!\!\perp D|W, X, S=1$ based on the tools discussed in Rohde and Hazlett (20XX). $Y_d \perp\!\!\!\perp D|W, X, S=1$ is a conditional ignorability statement that can be used to identify internally valid casual effects. $Y_d[i]$ is a potential outcome; that is, value the variable Y would have taken for unit i, if the variable D for unit i had been set, possibly counterfactually, to the value d. If W is not observed, we can consider it as an omitted variable and use an omitted variable based sensitivity analysis to understand the threats to internal validity posed by sample selection. In this case, W would be a variable that blocks non-causal paths or spurious associations created by sample selection.[1] Figure 1 presents a couple simple examples. The graphs in Figure 1 are internal selection graphs, which explicitly show how sample selection alters the relationships between variables in the selected sample. See Rohde and Hazlett (20XX) for how such graphs can be constructed from directed acyclic graphs and for further discussion of all the concepts just discussed.

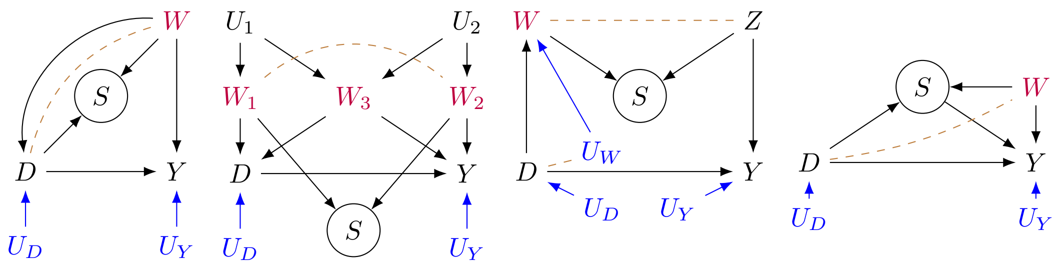
In this paper we discuss the omitted variable based sensitivity analyses of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) and how these can be leveraged to evaluate the threats that sample selection poses for internal validity. We discuss how the sensitivity parameters in such frameworks, in the sample selection setting, can be difficult to interpret. We show how alternative expressions for these sensitivity parameters can be derived in terms of more easily interpreted quantities in some simple, parametric settings. We then propose a bound on the difficult to interpret sensitivity parameters for general, non-parametric settings again in terms of more easily interpreted quantities. Finally, we provide a worked example and discussion.

[*]Department of Statistics, UCLA. adamrohde@ucla.edu

[†]Associate Professor, Departments of Statistics & Political Science, UCLA. chazlett@ucla.edu

[1]W could also be a simple common cause confounder of the Y, D relationship. This would put you in the settings already discussed in Cinelli and Hazlett (2020) and Chernozhukov et al. (2022).

Figure 1: Internal selection graph examples



2 Omitted variable based sensitivity analyses

2.1 Expressions for omitted variable bias

We consider three settings in which an investigator may consider the threats to internal validity from sample selection as a omitted variable bias problem. Each of these settings allow us to use expressions or bounds on such bias as the basis for a sensitivity analysis.

1. We may suppose that the investigator is interested in estimating a fully linear regression model, using the selected sample: $\left[Y = \beta_{Y\sim D|X,S=1}D + X\beta_{Y\sim X|D,S=1} + \epsilon_s\right]|S=1$. However, we know that $Y_d \not\perp\!\!\!\perp D|X,S=1$ and $Y_d \perp\!\!\!\perp D|W,X,S=1$; so $\beta_{Y\sim D|X,S=1}$ contains some bias, relative to what we would estimate if we were to include W in the regression: $\left[Y = \beta_{Y\sim D|W,X,S=1}D + X\beta_{Y\sim X|D,W,S=1} + \beta_{Y\sim W|D,X,S=1}W + \epsilon_l\right]|S=1$. We will refer to $\beta_{Y\sim D|X,S=1}$ as θ_s and $\beta_{Y\sim D|W,X,S=1}$ as θ_l, for the parameter of interest for the short and long regression, respectively.
2. We may also suppose that the investigator is interested in estimating a partially linear model: $[Y = \theta_l D + f_l(X,W) + \epsilon_l]|S = 1$. But, as in the fully linear case, they are only able to estimate $[Y = \theta_s D + f_s(X) + \epsilon_s]|S=1$. In these first two cases, we assume that the user is responsibly considering how inclusion of a covariate in these regressions differs from fully non-parametric conditioning and adjustment for the covariates, before considering that they want to know how inclusion of W in the regression changes θ_s.
3. We may also suppose that the investigator is interested in estimating a linear functional of the conditional expectation function of the outcome in a fully non-parametric setting, like $\theta_l = \mathbb{E}[Y_1 - Y_0|S=1] = \mathbb{E}[f_Y(1,X,W) - f_Y(0,X,W)|S=1]$ for a binary treatment D, where $Y_d = f_Y(d,X,W,U_Y)$ is the equation for Y in the structural causal model under intervention to set $D=d$. Again, the investigator is only able to estimate $\theta_s = \mathbb{E}[f_Y^*(1,X) - f_Y^*(0,X)|S=1]$, where $f_Y^*(D,X) \triangleq \mathbb{E}[Y|D,X] = \mathbb{E}[f_Y(D,X,W)|D,X]$.

In all three settings, there will be some bias resulting from not adjusting for W in our estimate. The bias is $\theta_s - \theta_l$. While the cause of the bias may be sample selection or common cause confounding or both, we can leverage an omitted variable bias frameworks to conduct sensitivity analysis to see how θ_s would change if we had included W in our estimation. For each of these settings, Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide expressions or bounds on the bias that can be expressed in terms of simple sensitivity parameters that *capture the relationships between the variables in the selected sample.*

1. Cinelli and Hazlett (2020) show that omitted variable bias for internally valid OLS regression can be expressed as

$$|\widehat{\text{bias}}| = \widehat{\text{se}}(\hat{\beta}_{Y\sim D|X,S=1})\sqrt{\text{df}_{S=1}\frac{R^2_{YW|D,X,S=1}R^2_{WD|X,S=1}}{1-R^2_{WD|X,S=1}}}$$

   where $\widehat{\text{se}}(\hat{\beta}_{Y\sim D|X,S=1}) = \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\sqrt{\text{df}_{S=1}}\widehat{\text{SD}}(D^{\perp X}|S=1)}$ is the standard error from the regression using the selected sample and $\text{df}_{S=1}$ are that regression's degrees of freedom. See Appendix A Section A.1 for the full derivation.
2. Chernozhukov et al. (2022) show that omitted variable bias in partially linear setting can be bounded by an expression in terms of $\eta^2_{YW|D,X,S=1}$, $\eta^2_{DW|X,S=1}$, and terms estimable from the data, where $\eta^2_{YW|D,X,S=1}$ and $\eta^2_{DW|X,S=1}$ are Pearson's correlation ratios (or the non-parametric R^2s).[2] $\eta^2_{YW|D,X,S=1}$ is the proportion of residual variation in Y explained by W. $\eta^2_{DW|X,S=1}$ is the proportion of residual variation in D explained by W.

[2] $\eta^2_{DW|X,S=1} = \frac{\text{Var}(\mathbb{E}[D|W,X,S=1]|S=1) - \text{Var}(\mathbb{E}[D|X,S=1]|S=1)}{\text{Var}(D|S=1) - \text{Var}(\mathbb{E}[D|X,S=1]|S=1)} = \frac{\eta^2_{D\sim WX|S=1} - \eta^2_{D\sim X|S=1}}{1-\eta^2_{D\sim X|S=1}}$. $\eta^2_{YW|D,X,S=1}$ can be similarly interpreted.

3. Chernozhukov et al. (2022) also show that omitted variable bias in fully non-parametric setting can be bounded by an expression in terms of $\eta^2_{YW|D,X,S=1}$ and a second term that, in the case of targeting $\theta_l = \mathbb{E}[Y_1 - Y_0|S=1]$ with a binary treatment D, is the "average gain in the conditional precision with which we predict D by using W in addition to X," which is somewhat similar to $\eta^2_{DW|X,S=1}$.

Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide thorough discussions of sensitivity analysis using these bias expressions, in addition to tools and examples for conducting such analysis. However, this discussion hinges on the ability for practitioners to interpret the sensitivity parameters on which this sort of sensitivity analysis relies. As we discuss next, this is complicated by sample selection.

2.2 Difficulty interpreting sensitivity parameters

Since the bias we are worried about might be a result of sample selection and the effect we are interested in is for the selected sample in hand alone, we allow for either of the sensitivity parameters $R^2_{YW|D,X,S=1}$ or $R^2_{WD|X,S=1}$ (or $\eta^2_{YW|D,X,S=1}$ or $\eta^2_{DW|X,S=1}$) to contain a purely statistical relationship that results from stratifying to $S=1$ where S is a collider. Sensitivity analysis should leverage external knowledge about the relationships captured by these sensitivity parameters to inform the range of plausible values that the sensitivity parameters may take. This can then be used to determine how θ_s may change if W were included in the estimation. However, such external knowledge will be difficult to obtain when one of these sensitivity parameters contains a purely statistical (i.e., non-causal) relationship created by sample selection. This is because the association captured by the sensitivity parameter does not result from structural relationships in which one variable causes another. Instead, the association results from the often counterintuitive phenomenon of conditioning on a collider. In what follows, we consider how we might deal with this problem by appealing to relationships between the variables in the full population, as opposed to the selected sample, and focus on $R^2_{WD|X,S=1}$ and $\eta^2_{DW|X,S=1}$. Similar discussion could apply to $R^2_{YW|D,X,S=1}$ or $\eta^2_{YW|D,X,S=1}$. We do not fully address the non-parametric case, since not all of the sensitivity parameters can be expressed as R^2s or η^2s, but if the sample selection collider alters the association between Y and W, then our discussion will still apply. We start by building a sense how these sensitivity parameters can be expressed in terms of structural relationships between the variables in the full population in very simple settings.

Binary random variables To provide some intuition about how we will try to understand $R^2_{DW|X,S=1}$ and $\eta^2_{DW|X,S=1}$, we consider the case where W, D are binary. We assume that data are generated according to the the simple collider graph: $D \rightarrow S \leftarrow W$. Here $X = \{\emptyset\}$. $R^2_{DW|S=1}$ can be written in terms of six probabilities as shown in Equation 1.[3] See Appendix A Section A.2 for the complete derivation.

$$R^2_{DW|S=1} = [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2 \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{pmatrix}} \tag{1}$$

The relationship between W and D in the selected sample ($S=1$) can be expressed in terms of the relationships between S and W, D, in the full population, where we also need $P(D=1), P(W=1)$. These quantities should be easy for researchers to reason about, since they capture structural (i.e., causal) relationships between the variables. The key here is that $R^2_{DW|X,S=1}$ can be understood in terms of structural relationships in the full population.

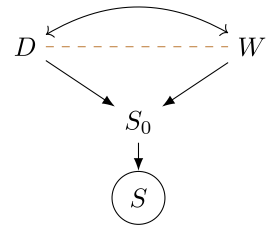
Truncated multivariate normal random variables To provide additional intuition into how we might try to think about $R^2_{DW|X,S=1}$ and $\eta^2_{DW|X,S=1}$, we consider another simple case where W, D, S have the causal structure shown in Figure 2 and W, D, S_0 have a multivariate normal joint distribution and $S = \mathbf{1}[S_0 \geq \mathrm{C}]$ for some $\mathrm{C} \in \mathbb{R}$. Here $X = \{\emptyset\}$. S_0 is a hypothesized latent variable that captures how W and D relate to S. The bidirected edge captures that W, D could have some relationship other than that created by conditioning on S. Within the stratum $S=1$, we have a truncated multivariate normal joint distribution. Using the properties of truncated normals, we get the expression for $R^2_{WD|S=1}$ in Equation 2.[4] See Appendix A Section A.3 for the complete derivation. Heckman (1979) relies on truncated normal variables but connections to that work are not explored here.

$$R^2_{DW|S=1} = \left(\frac{\rho_{DW} - \rho_{SD}\rho_{SW}\theta}{\sqrt{1-\rho^2_{SD}\theta}\sqrt{1-\rho^2_{SW}\theta}}\right)^2, \text{ where } \theta \text{ is a function of } P(S=1) \text{ or C} \tag{2}$$

[3] $P_{W=w} = P(W=w)$, $P_{D=d} = P(D=d)$, $P_{S=1} = P(S=1)$, $P_{S=1|wd} = P(S=1|W=w, D=d)$, $P_{W=0} = 1 - P_{W=1}$ and $P_{D=0} = 1 - P_{D=1}$.

[4] In Equation 2, if $\rho_{WD} = 0$, then $R^2_{DW|S=1} = \frac{R^2_{SD}R^2_{SW}\theta^2}{\sqrt{1-R^2_{SD}\theta}\sqrt{1-R^2_{SW}\theta}}$.

Figure 2: Internal selection graph for truncated multivariate normal example



The relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S, W and S, D as well as between W and D, in the full population, where we also need $P(S=1)$, the probability of selection. All of these quantities should be easy for researchers to reason about, since they capture structural (i.e., causal) relationships between the variables in the population.

Partial correlation and "constant selection effects" It is important to note that truncation on S or stratification to $S=1$ (i.e., sample selection) is not the same as conditioning on S. Conditioning on S (linearly) would give Equation 3, based on the partial correlation formula.[5] This does not equal $R^2_{WD|S=1}$ in general. Equations 2 and 3 are remarkably similar, however, with their only differences being the need to account for where truncation happens (or the probability of selection). Equation 3 holds for linear conditioning on S, without any restrictions on the distribution or relationships between W, D, and S. Equation 2 only holds for truncated normals.

$$R^2_{DW|S} = \left(\frac{\rho_{DW} - \rho_{SD}\rho_{SW}}{\sqrt{1-\rho^2_{SD}}\sqrt{1-\rho^2_{SW}}} \right)^2 \tag{3}$$

We might wonder under what circumstances we would be able to use Equation 3 to inform our discussion of $R^2_{WD|S=1}$. We explore this in Appendix A Section A.4. The idea is to assume something like "constant selection effects" (akin to constant treatment effects) between the $S=1$ and $S=0$ strata. Such an approach also requires some assumptions about the strata specific variances for W and D. While this could be used as a first pass analysis, the assumptions are typically unrealistic and using this approach could underestimate $R^2_{WD|S=1}$. In Appendix A Section A.5, we discuss a simple bound on $R^2_{WD|S=1}$ that relies on the partial correlation formula. However, this bound is typically uninformative (not less than 1) and so we do not discuss it here.

In the next section, we propose bounds on $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ (as well as $R^2_{WD|X,S=1}$ and $\eta^2_{DW|X,S=1}$) for the non-parametric case in which we make no restrictions on the distribution or relationships between W, D, and S. In spirit, these bounds are similar to Equations 1 and 2 in that they ask us to reason about the relationships between W, D, and S (i.e., the sample selection mechanism) in the population, as well as the probability of selection. These population relationships are structural causal relationships. As such, researchers should be able to appeal to external knowledge, expertise, and data in order to understand the range of plausible strengths of these relationships.

3 Proposal

Since one of our sensitivity parameters might contain some spurious association created by sample selection (and is therefore difficult to interpret), we aim to bound this sensitivity parameter with structural relationships from the full population, about which investigators should be able to reason easily. We start by noting that $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1}$. This is easy to see since $\eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \sup_f \left[\text{Cor}^2(D, f(W)|S=1)\right]$. (Doksum and Samarov, 1995; Chernozhukov et al., 2022) $\eta^2_{D,W|S=1}$ measures portion of the variation in D that can be explained by $\mathbb{E}[D|W,S=1]$, the conditional expectation function (CEF).[6] We will bound $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ by appealing to mutual information.

[5] In Equation 3, if $\rho_{WD} = 0$, then $R^2_{DW|S} = \frac{R^2_{SD}R^2_{SW}}{\sqrt{1-R^2_{SD}}\sqrt{1-R^2_{SW}}}$. Without restrictions on the distribution or relationships between W, D, and S, recall that $\rho_{WD} = 0$ does not mean that W and D are marginally independent.

[6] The law of total variance tells us that $\text{Var}(D|S=1) = \text{Var}(\mathbb{E}[D|W,S=1]|S=1) + \mathbb{E}[\text{Var}(D|W,S=1)|S=1]$. (Aronow and Miller, 2019)

What is mutual information? Before we try to work with mutual information, what is it? Mutual information is a measure of how similar the joint distribution is to the product of the marginal distributions. Therefore, it is a measurement of the total dependence between A and B, whether this dependence is linear or non-linear. It makes no assumptions about the distribution of A and B or the form their dependence takes. Mutual information between A and B, $\mathrm{MI}(A,B)$, can be thought of as the reduction in uncertainty about A that results from learning the value of B. (Smith, 2015) Mutual information is defined in the following ways, where D_{KL} is KL divergence and H is entropy.

$$\begin{aligned}\mathrm{MI}(A,B) &= D_{\mathrm{KL}}\left(P_{(A,B)} \| P_A \otimes P_B\right)\\ &= \sum_a \sum_b P_{(A,B)}(a,b)\log\left(\frac{P_{(A,B)}(a,b)}{P_A(a)P_B(b)}\right)\\ &= H(A)+H(B)-H(A,B)\end{aligned}$$

There are also notions of conditional mutual information and joint mutual information and entropy. See Ihara (1993); MacKay (2003); Cover and Thomas (2006) for details.[7] Mutual information measures the amount of Shannon information revealed about A as a result of knowing B. Shannon information of an event is defined as $I_A(a) = \log(1/P_A(a))$. Events that occur with certainty are perfectly unsurprising and hence have no information. As the probability of an event decreases, the surprise that the event occurred increases, and so does the information content. The entropy of a random variable is the average information of the outcomes of the variable, $H(A) = \sum_a P_A(a)\log(1/P_A(a))$, and can be thought of as the uncertainty in the variable's outcomes. (MacKay, 2003) In practice, interpreting mutual information can be difficult, so we appeal to a normalized version that has nice properties discussed below.

Mutual information for Gaussians In order to connect $R^2_{DW|X,S=1}$ and $\eta^2_{DW|X,S=1}$ with mutual information, we draw inspiration from the relationship between R^2 and mutual information for random variables with Gaussian distributions. For random variables, W and D, with a bivariate Gaussian joint distribution, there is an exact relationship between R^2 (i.e., squared correlation coefficient) and mutual information (MI). (Ihara, 1993; Cover and Thomas, 2006)

$$\mathrm{MI}(W;D) = -\frac{1}{2}\log(1-R^2_{DW}) \iff R^2_{DW} = 1-\exp(-2\times \mathrm{MI}(W;D))$$

Many authors have considered this type of transformation of mutual information as a way to obtain something like a non-parametric correlation based on mutual information. See Linfoot (1957); Kent (1983); Joe (1989); Kojadinovic (2005); Speed (2011); Kinney and Atwal (2014); Asoodeh et al. (2015); Smith (2015); Shevlyakov and Vasilevskiy (2017); Laarne et al. (2021). Drawing on Lu (2011)'s L-measure,[8] we create a useful normalized version of mutual information for arbitrary random variables. See Appendix A Section A.6.

Bounds This normalization of mutual information and Theorem 1 allow us to build interpretable bounds on $R^2_{D,W|S=1}$, $\eta^2_{D,W|S=1}$, $R^2_{D,W|X,S=1}$, and $\eta^2_{D,W|X,S=1}$. Theorem 1 can be applied to the case for which S is a collider between D and W (e.g., $D \to S \leftarrow W$), providing a guide to how conditioning on a collider alters the relationship between the parents of the collider. We leverage our normalized version of mutual information, which we call normalized scaled mutual information (NSMI), to give interpretable bounds on $R^2_{D,W|S=1}$, $\eta^2_{D,W|S=1}$, $R^2_{D,W|X,S=1}$, and $\eta^2_{D,W|X,S=1}$ that rely on Theorem 1. These bounds can be found in Theorems 2 and 3. These results are proved and NSMI is defined in Appendix A Section A.6. While framed in the context of conditioning on a collider, S, these results hold for stratification to $S=1$ in general.

Theorem 1. *For random variables D, W, S, conditioning on S alters the relationship between D and W according to the expression $MI(D;W|S) = MI(D;W) + MI(S;[D,W]) - MI(S;D) - MI(S;W)$. Therefore, the change in dependence due to conditioning on S can be characterized using mutual information according to $MI(D;W|S) - MI(D;W) = MI(S;[D,W]) - MI(S;D) - MI(S;W)$. The dependence is not changed when $MI(S;[D,W]) = MI(S;D) + MI(S;W)$.*

When S is binary, it is also possible to write $MI(D;W|S) = p(S=1)MI(D;W|S=1) + p(S=0)MI(D;W|S=0)$, meaning that $MI(D;W|S=1) \leq \frac{MI(D;W|S)}{p(S=1)} = \frac{MI(D;W)+MI(S;[D,W])-MI(D;S)-MI(W;S)}{p(S=1)}$.

Theorem 2. *For random variables D, W, S, for which S is a collider on a path from D to W in G^+_S that, if conditioned on, could alter the relationship between D and W (e.g., $D \to S \leftarrow W$), the $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ resulting after stratification to $S=1$ can be bounded in the following ways:*

[7] We've shown the definition of mutual information for discrete random variables but there are analogous definitions for all random variables.

[8] The L-measure takes the form $1-\exp(-2\times \mathrm{IF}\times \mathrm{MI})$, where IF is an "inflation factor" that ensures that the L-measure takes appropriate values for arbitrary variables, not just continuous variables.

1. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(S;[D,W])]}{[1-NSMI(S;D)][1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
2. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(D;S|W)]}{[1-NSMI(S;D)]}\right)^{\frac{1}{p(S=1)}}$
3. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(W;S|D)]}{[1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
4. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(D;S|W)])^{\frac{1}{p(S=1)}}$
5. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(W;S|D)])^{\frac{1}{p(S=1)}}$
6. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(S;[D,W])])^{\frac{1}{p(S=1)}}$

Theorem 3. *For random variables D, W, S, X, for which S is a collider on a path from D to W in G^+_S that, if conditioned on, could alter the relationship between D and W (e.g., $D \to S \leftarrow W$), the $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

1. $R^2_{D,W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
2. $R^2_{D,W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - [1-NSMI(D;X|S=1)]\left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
3. $\eta^2_{D,W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$
4. $\eta^2_{D,W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - [1-NSMI(D;X|S=1)]\left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$

where $R^2_{D\sim X|S=1}$ or $\eta^2_{D\sim X|S=1}$ is estimated from the data. We can approximate or inform the choice of $NSMI(D; X|S = 1)$ using the estimated $R^2_{D\sim X|S=1}$ or $\eta^2_{D\sim X|S=1}$.[9] *These bounds are all analogous to bound 1 in Theorem 2. Analogs to bounds 2 - 6 in Theorem 2 could also be formed.*

Normalized scaled mutual information (NSMI) NSMI is a mutual information based measure of dependence between random variables. It measures the full dependence relationship of two random variables, not just the linear dependence or dependence related through the conditional expectation function. NSMI is a useful measure of dependence between random variables in that it satisfies the properties discussed in Rényi (1959), Smith (2015), Lu (2011), and others as the properties possessed by "an appropriate measure of dependence."[10][11]

1. NSMI is defined for arbitrary pairs of random variables.
2. NSMI is symmetric.
3. NSMI takes values between 0 and 1.
4. NSMI equals 0 if and only if the variables are independent.
5. NSMI equals 1 if and only if the variables a strict dependence (functional relationship).
6. NSMI is invariant to marginal, one-to-one transformations of the variables.
7. If the variables are Gaussian distributed, then NSMI equals their R^2.
8. $\text{NSMI}(D; \mathbb{E}[D|W, S = 1]|S = 1) = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D,W|S=1}$.

All but the last of these are discussed in Rényi (1959),Smith (2015), and Lu (2011). The last property results from how we've defined NSMI. "Furthermore, MI is invariant under monotonic transformations of variables. This means that the MI correlation coefficient of a non-linear model (X, Y) matches the Pearson correlation of the linearized model $(f(X), g(Y))$. General conditions for f and g are described in" Ihara (1993). (Laarne et al., 2021) The "MI correlation coefficient" discussed in Laarne et al. (2021) is defined similarly to NSMI for continuous variables. Thus, NSMI can be thought of as the R^2 of the linearized model $(f(X), g(Y))$.[12]

[9] We cannot directly estimate $\text{NSMI}(D; X|S = 1)$, since we cannot estimate $\mathbf{\Omega}$ or IF which are based on $\eta^2_{D\sim W,X|S=1}$ and $\text{MI}(D; [W, X]|S = 1)$.

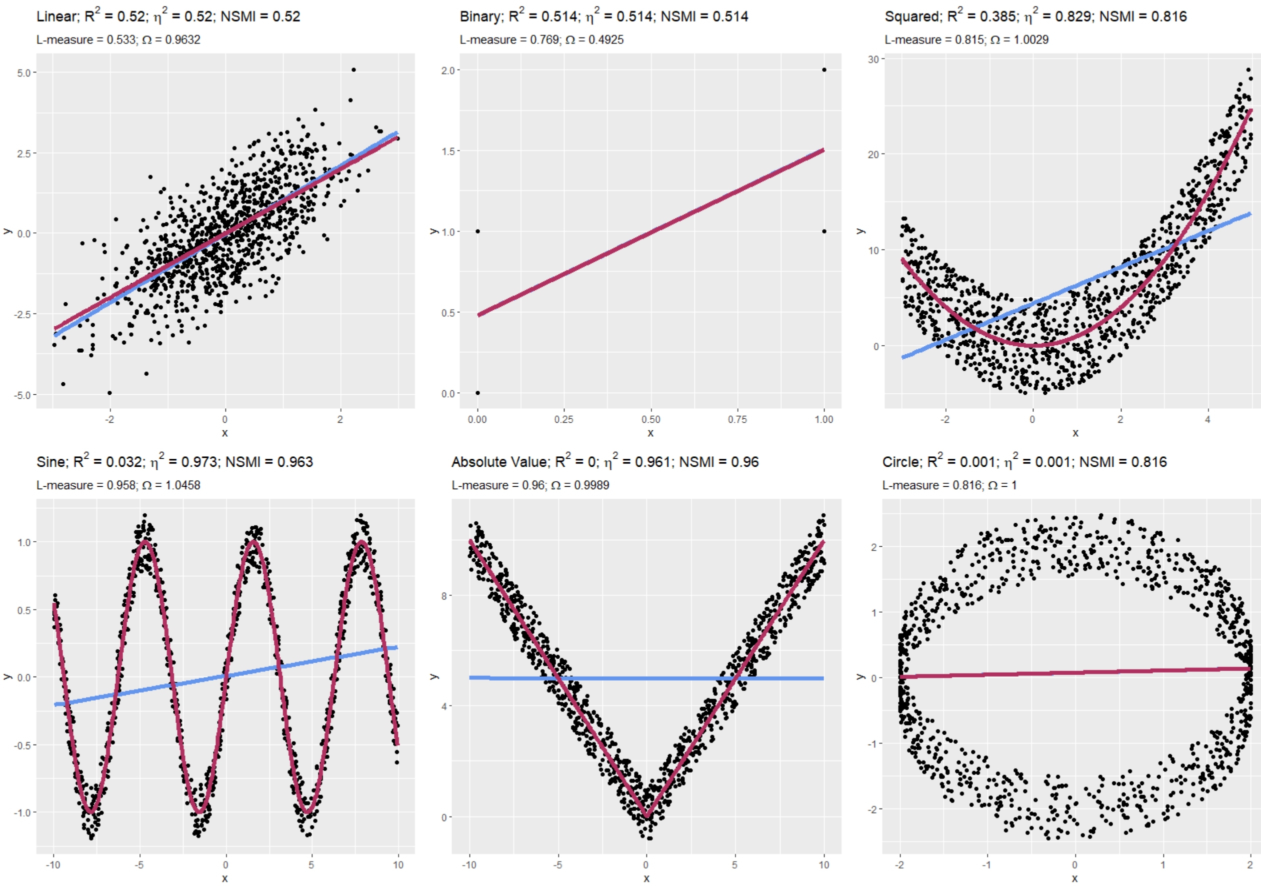
[10] Mutual information satisfies properties 1, 2, 4, and 6. Squared Pearson correlation (i.e., R^2) satisfies properties 1, 2, 3, 5, and 7. η^2 also does not satisfy all of these properties. See Rényi (1959) for further discussion.

[11] The transformation $\ell^2(\text{MI}(X; Y)) = 1 - \exp(-2 \times \text{MI}(X; Y))$ ensures that properties 2, 3, 6, and 7 are satisfied; it is the transformation that turns mutual information into an R^2 for Gaussian distributed variables. The transformation $L^2(\text{MI}(X; Y)) = 1 - \exp(-2 \times \text{IF} \times \text{MI}(X; Y))$ is the square of Lu (2011)'s L-measure, where IF is chosen to ensure that properties 1 and 5 are satisfied, while also maintaining properties 2, 3, 6, and 7. The transformation $\text{NSMI}(X; Y) \triangleq L^2_{\mathbf{\Omega}}(\text{MI}(X; Y)) = 1 - \exp(-2 \times \mathbf{\Omega} \times \text{IF} \times \text{MI}(X; Y))$ is our normalized and scaled measure of mutual information, where $\mathbf{\Omega} \geq 0$ is also chosen to ensure that property 8 is satisfied, while also maintaining properties 1 through 7. Lu (2011) demonstrates that properties 1 through 7 hold for the L-measure. Given this, it is trivial to see that they also hold for NSMI.

[12] It is worth noting that, although we might be more comfortable thinking about correlations and R^2's, they are not necessarily capturing what we expect. First, correlation and R^2 capture only the strength of linear association; these do not necessarily capture an intuitive sense of dependence but one restricted to linear relationships. Also, "Mutual Information is a nonlinear function of ρ which in fact makes it additive. Intuitively, in the

We also provide some examples to help readers gain some familiarity with NSMI. In Figures 3 and 4, we show 12 different types of bivariate relationships with the corresponding R^2, η^2, and NSMI. In these examples, we estimate NSMI using the "rmi" and "infotheo" R packages and η^2 with the "KRLS" R package using samples of 1000 data points. (Michaud, 2018; Meyer, 2014; Hainmueller and Hazlett, 2014; Ferwerda et al., 2017) There is estimation error in these, since mutual information can be difficult to estimate in practice, but the Figures should still be informative.[13] We see that NSMI is larger that η^2 but is often very comparable. When η^2 does a poor job of capturing the full relationship between the variables, NSMI can be much larger than η^2. Lu (2011)'s L-measure is close to or larger than NSMI. So it is possible to reason about the L-measure as an approximation or as a bound on NSMI. See Appendix A Section A.6 for more detail on NSMI and the L-measure. See Figure 8 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. In the Gaussian case, NSMI equals R^2; and so interpretation of NSMI should be familiar.

Figure 3: NSMI Examples. These are generated with various linear and non-linear relationships between x and y. The blue line is an linear fit. The red line is a flexible fit or the true non-linear relationship.

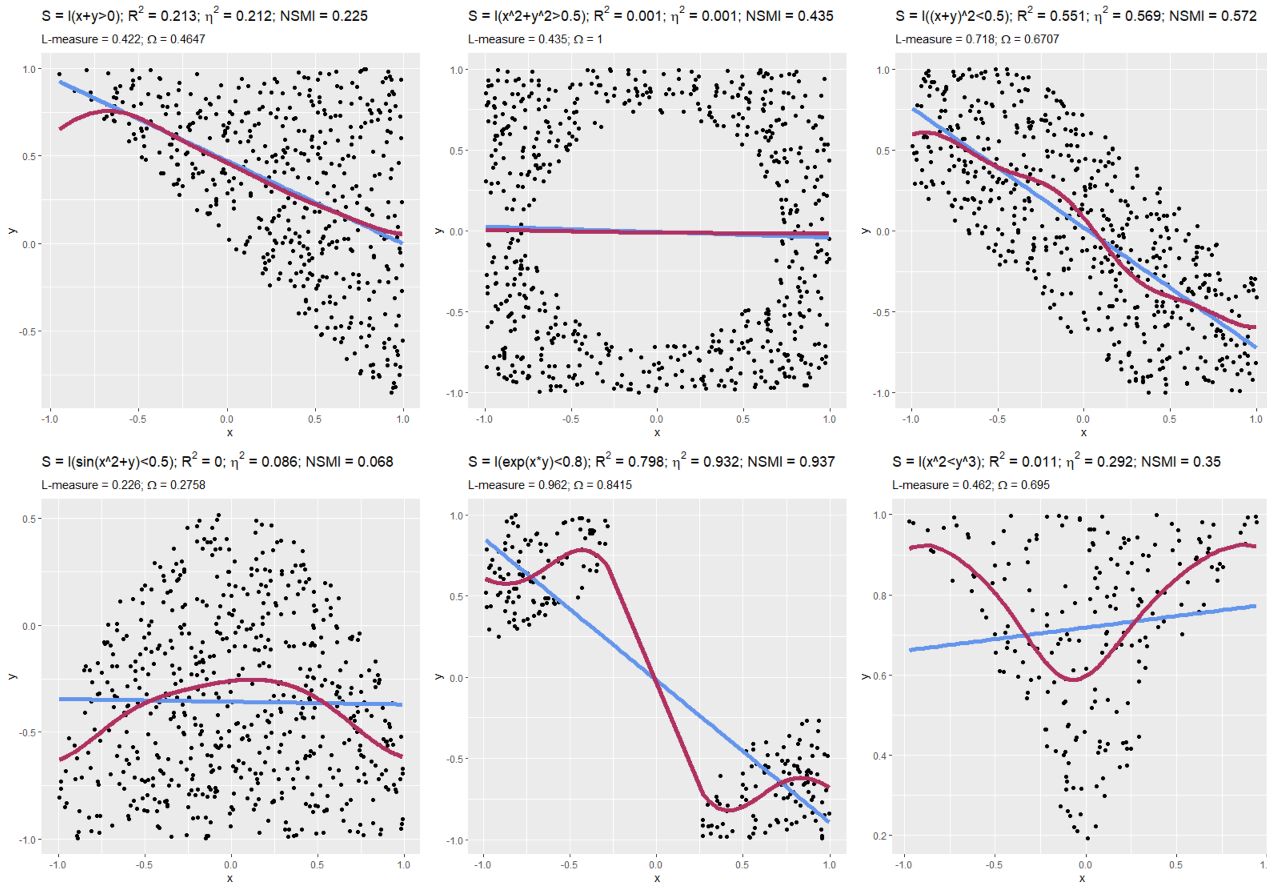


Discussion of bounds Bounds 1 through 3 in Theorem 2 are tighter than bounds 4 through 6, but require additional sensitivity parameters as well as some knowledge about how mutual information works. That is, since some of the NSMI quantities are related in the bounds in Theorem 2, users need to take care to reason about coherent combinations of the NSMI quantities. In particular, the bounds all take the form $1-(\tau)^{\frac{1}{p(S=1)}}$ but with different τ; τ must take a value between 0 and 1. This reflects the fact that $1-(1-\text{NSMI}(W;D|S))^{\frac{1}{p(S=1)}}$ equals bounds 1 through 3 and $\text{NSMI}(W;D|S)$ takes values between 0 and 1. This, in turn, reflects that $\text{MI}(D;W|S) = \text{MI}(D;W) + \text{MI}(S;[D,W]) - \text{MI}(S;D) - \text{MI}(S;W) \geq 0$. For this

Gaussian case, ρ should never be interpreted linearly: a ρ of $\frac{1}{2}$ carries ≈ 4.5 times the information of a $\rho = \frac{1}{4}$, and a ρ of $\frac{3}{4}$ 12.8 times!" (Taleb, 2019) "One needs to translate ρ into information. See how $\rho = 0.5$ is much closer to $[\rho =]0$ than to a $\rho = 1$. There are considerable differences between .9 and .99." (Taleb, 2019) See Figure 8 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. See Figure 9 for a plot of the relationship between mutual information and R^2 for Gaussian variables, this is also the normalization curve we use. Mutual information can capture our intuitive sense of dependence better than correlation and R^2 even in the simple Gaussian case.

[13] In addition, we present the L-measure and $\mathbf{\Omega}$. See the discussion in Appendix A Section A.6 for more detail on NSMI and $\mathbf{\Omega}$.

Figure 4: More NSMI Examples. These are generated by selecting a non-random sample from two uniform random variables. S is the sample selection variable. The blue line is an linear fit. The red line is a flexible fit.

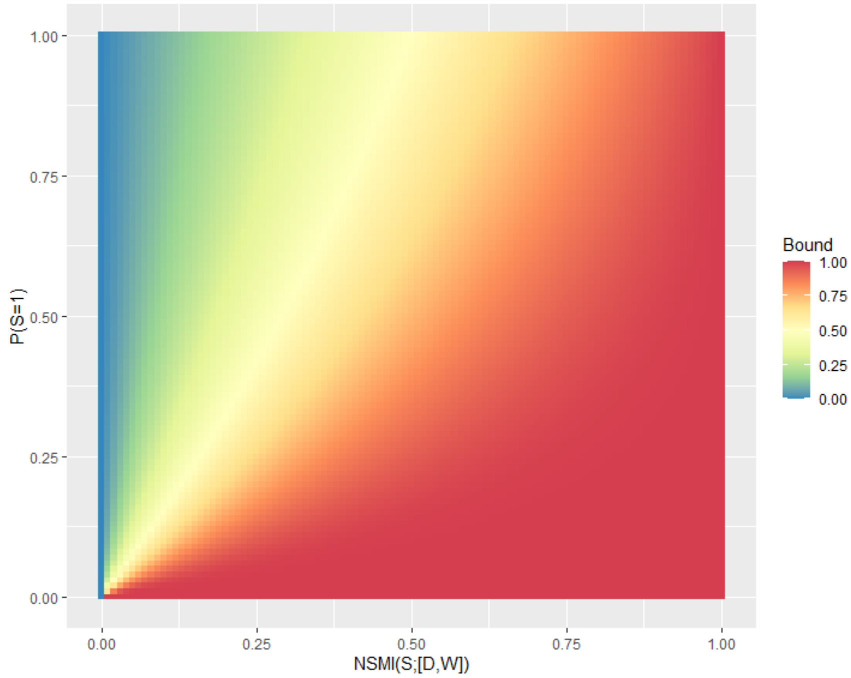


reason, we encourage users unfamiliar with mutual information to use bounds 4 through 6, where the condition that $\tau \in [0,1]$ will always be satisfied given NSMI values between 0 and 1. If W and D are assumed to be marginally independent, then $\text{NSMI}(D;W) = 0$ and this term can be removed from the bounds. Which bound is most useful depends on the relationships that practitioners feel comfortable reasoning about in terms of NSMI's.

We consider in detail bound 6 from Theorem 2. This bound is an expression of normalized scaled mutual information for the marginal mutual information between D and W, for the mutual information between S and $[D,W]$ together, and the probability of selection, $P(S=1)$.[14] As we saw in the case of binary random variables and truncated normal random variables, we have an expression in terms of structural (i.e., causal) relationships between the variables in the full population. In Figure 5, we show bound 6 from Theorem 2 changes for different values of $\text{NSMI}(S;[D,W])$ and $p(S=1)$. For this, we assume that that W, D are marginally independent and so $\text{NSMI}(D;W) = 0$ and the bound becomes $B \triangleq 1 - (1 - \text{NSMI}(S;[D,W]))^{\frac{1}{p(S=1)}}$. As $p(S=1) \to 1$, $B \to \text{NSMI}(S;[D,W])$. As $p(S=1) \to 0$, $B \to 1$. As $\text{NSMI}(S;[D,W]) \to 1$, $B \to 1$. As $\text{NSMI}(S;[D,W]) \to 0$, $B \to 0$. These dynamics are easy to see in the expression for the bound itself. They reflect the bounds on $\text{MI}(W;D|S=1)$ that we then scale and normalize. It is worth noting that this bound is not always informative (i.e., smaller than 1); small probabilities of selection can lead to high bounds, regardless of the value for $\text{NSMI}(S;[D,W])$. This reflects that, when the selection probability is small, $\text{NSMI}(S;[D,W])$ carries much less information about the stratum $S=1$ than it does about the stratum $S=0$.

[14] $P(S=1)$ can be thought of as the proportion of the population that the sub population for which our selected sample is a representative sample represents. It is not the size of our data sample relative to the size of the population. Note that it is important to have a clear sense of the population from which the sample has been selected here, but this is already required to be able to think about the sample selection mechanism in the first place and hence know whether conditioning on W will yield conditional ignorability or not. Alternatively, we might think about the $P(S=1)$ that would bring the estimated result to zero.

Figure 5: Bound 6 from Theorem 2 on $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ given values for NSMI$(S;[D,W])$ and $p(S=1)$ and assuming NSMI$(D;W)=0$



4 Discussion

4.1 Worked example

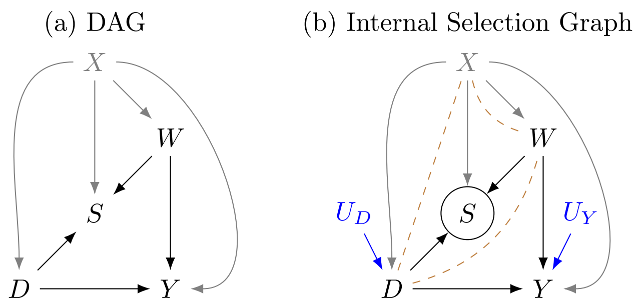
We start our discussion with a worked example demonstrating how a sensitivity analysis for sample selection may proceed based on the discussion above. Hazlett (2020) considers the effect of being directly harmed in the conflict in Darfur in early 2000s on attitudes about peace using a survey of individuals in refugee camps. The paper controls for things like village, gender, and other important covariates. While adjustment for these covariates likely reduces non-causal association between the treatment and outcome, being harmed may effect whether someone re-entered the conflict (and hence was not captured in the survey). An individual's pro-peace predisposition (before the conflict) may be a common cause of both whether they re-entered the conflict and their peace attitudes at the time of the survey. This may cause there to be a generalized non-causal path running from harm to pro-peace predisposition to attitude about peace that could threaten the internal validity of estimated effects. In Figure 6, D is direct harm, Y is attitudes about peace, W is pro-peace predisposition (before the conflict), S is being in the survey from the refugee camp (i.e., did not re-enter the conflict), and X is observed covariates like village and gender. Hazlett (2020) is able to adjust for the observed covariates, but the path $D - -Z \rightarrow Y$ cannot be blocked since pro-peace predisposition is not observed. That is, we are able to estimate an effect controlling for age, gender, village, and other covariates. But this effect could be biased by the spurious relationship created by sample selection as a collider between direct harm (D) and pro-peace predisposition (W). Hazlett (2020) estimate that peace index is .09 to .10 units higher among those directly harmed.

Outcome: *peace factor*

Treatment:	Est.	S.E.	t-value	RV
directly harmed	0.097	0.023	4.184	13.9%
df = 783				

The process that would drive individuals back into the conflict would "act more powerfully for men of fighting age because in this context, few women or elderly participate directly in the armed opposition groups. If such a process drove the results, we would see the apparent effect most strongly among young men but should see little or no apparent effect among women or the elderly who are far less likely to join the opposition. This is not the case." (Hazlett, 2020) We might then claim that the effect of direct harm on peace attitudes among women and the elderly is perhaps not biased by this sample selection

Figure 6: Possible sample selection bias in Hazlett (2020)



mechanism. But this sample selection mechanism might not allow us to obtain an internally valid effect estimate for fighting age men.

We can directly use our bounds from above in the the sensitivity analysis from Cinelli and Hazlett (2020) using the software from Cinelli et al. (2020). This omitted variable bias based sensitivity analysis requires that we consider hypothetical values for $R^2_{WD|X,S=1}$ and $R^2_{YW|D,X,S=1}$. Inspecting Figure 6, we see that $R^2_{YW|D,X,S=1}$ captures just the causal path $W \to Y$, that is, the strength of the relationship between pro-peace predisposition (W) and attitudes about peace after the conflict (Y) after controlling for observed covariates (X) like village and gender. This is a structural relationship that we will be able to build some intuition about. On the other hand, $R^2_{WD|X,S=1}$ is more difficult. This captures the path $W - -D$ and relates to the strength of the relationship between pro-peace predisposition (W) and direct harm (D) within the selected sample of refugees that did not reenter the fight after controlling for observed covariates (X) like village and gender. These variables do not have a direct relationship in the population that we can build intuition about that would allow us to directly reason about $R^2_{WD|X,S=1}$. In fact, the assumption in Figure 6 is that pro-peace predisposition (W) is independent of direct-harm (D) in the population, conditional on the observed covariates like village and gender. So their entire relationship is created as a result of conditioning on a collider due to sample selection.

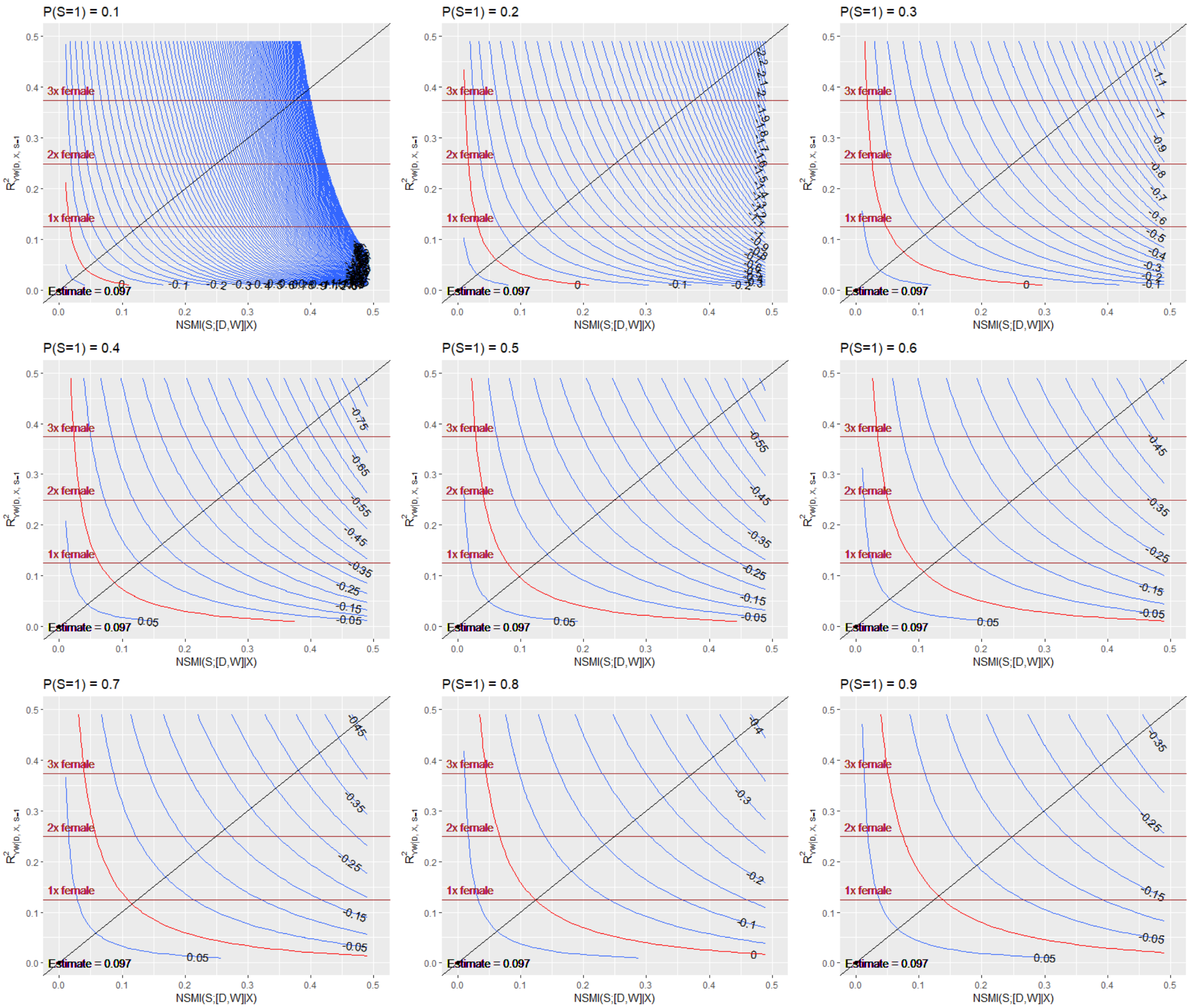
Given the difficulty in interpreting and building intuition for $R^2_{DW|X,S=1}$, we use bound 2 from Theorem 3 to bound $R^2_{DW|X,S=1}$, where we assume that D and W are independent conditional on X in the population. We also choose a bound analogous to bound 6 from Theorem 2. The bound we use is therefore

$$R^2_{WD|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - [1-\text{NSMI}(D;X|S=1)]\,[1-\text{NSMI}(S;[D,W]|X)]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right),$$

where we estimate $R^2_{D\sim X|S=1}$ from the data and approximate $\text{NSMI}(D;X|S=1)$ based on that estimate. We simply assume that we have the worst case where $R^2_{WD|X,S=1}$ equals this bound and substitute this into the bias expression provided in Cinelli and Hazlett (2020) and use this to calculate revised estimates for hypothesized values of $R^2_{YW|D,X,S=1}$, $\text{NSMI}(S;[D,W]|X)$, and $p(S=1)$. Contour plots that show the surface of revised estimates for the breadth of values these three sensitivity parameters can take are displayed in Figure 7. We can take $1-p(S=1)$ to represent the portion of refugees that reentered the fighting and were, therefore, not eligible to be captured by the survey. Focusing on the individuals who survived their injuries, we might suppose that no more than say 20% of individuals reentered the fighting. If we believe this is plausible, we might then consider the contour plot at the bottom middle of Figure 7. We may believe that no unobserved variable will explain more of the outcome than the female variable (or a multiple of the female variable). We show 1x, 2x, and 3x how much the female variable explains of the outcome in the contour plots. In the $p(S=1)=0.8$ panel of Figure 7, we see that assuming that $R^2_{YW|D,X,S=1}$ equals the partial R^2's between the female variable and the outcome, peace index, an $\text{NSMI}(S,[D,W]|X)$ of about 0.13 or so would bring our effect estimate to zero.

Do we think that an $\text{NSMI}(S,[D,W]|X)$ of 0.13 or more is plausible? If S and $[D,W]$ shared a joint Gaussian distribution, then this would correspond to an R^2 of 0.13; and recall that NSMI can be thought of as the R^2 of the linearized model. We might suspect that most of the decision to reenter the fight would be explained by gender, age, and village, all of which we control for. The question becomes to what extent does reentering the fight depend on pro-peace predisposition and direct harm, after controlling for the observed covariates. The relationship between direct harm and reentering the fight is likely dramatically diminished by controlling for gender, which likely captures much of the relationship between these two. Similarly, gender and age likely dramatically reduce the dependence between pro-peace predisposition and reentering the fight. Other determinants of reentering the fight might matter much more than direct harm and pro-peace predisposition like having family

Figure 7: Sensitivity analysis contour plots for Darfur example. Contours represent revised effect estimates.



members still in conflict zone. Moreover, a complex decision like this likely is the result of numerous social and personal factors, as is any social phenomenon. If having family members remaining in the conflict zone and other factors (that we might assume are not related to direct harm or peace attitudes) are the predominant determinants of reentering the fight, after controlling for observed covariates, then we might have a case that there is a small $\text{NSMI}(S, [D, W]|X)$, perhaps less that 0.13, which is a fairly week dependence relationship in general. Again, all the discussion in this paragraph assumes that pro-peace predisposition explains just as much of peace attitudes as female. If say, pro-peace predisposition explains less than half as much of peace attitudes as female, then we could tolerate an $\text{NSMI}(S, [D, W]|X)$ of up to perhaps 0.25. For Gaussians, this would be an R^2 of 0.25 or a correlation of 0.5. Perhaps reentering the fight would not have such a strong dependence on pro-peace predisposition and direct harm if almost all of the decision to reenter the fight, controlling for gender and age, is determined by having family members still in conflict zone and other determinants of selection, not attitudes about peace and direct harm. In this setting, we might be able to conclude that our effect estimate would not change sign of our estimate, despite this level of sample selection bias. On the other hand, if it is reasonable to believe more than 20% of refugees reentered the fight, perhaps it is more likely that the sign of the estimate could change due to the amount of sample selection bias.

4.2 Overview of Sensitivity Analysis

4.3 Discussion of alternative approaches

References

Aronow, P. M. and Miller, B. T. (2019). *Foundations of Agnostic Statistics.* Cambridge University Press.

Asoodeh, S., Alajaji, F., and Linder, T. (2015). On maximal correlation, mutual information and data privacy. In *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*, pages 27–31.

Chernozhukov, V., Cinelli, C., Newey, W., Sharma, A., and Syrgkanis, V. (2022). Long story short: Omitted variable bias in causal machine learning. Working Paper 30302, National Bureau of Economic Research.

Cinelli, C., Ferwerda, J., and Hazlett, C. (2020). sensemakr: Sensitivity analysis tools for ols in r and stata.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience.

Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57.

Doksum, K. and Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5):1443–1473.

Ferwerda, J., Hainmueller, J., and Hazlett, C. J. (2017). Kernel-based regularized least squares in r (krls) and stata (krls). *Journal of Statistical Software*, 79(3):1–26.

Hainmueller, J. and Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168.

Hazlett, C. (2020). Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, 64(5):844–870.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Ihara, S. (1993). *Information Theory for Continuous Systems.* WORLD SCIENTIFIC.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):157–164.

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173.

Kinney, J. B. and Atwal, G. S. (2014). Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359.

Kojadinovic, I. (2005). On the use of mutual information in data analysis : an overview.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions: Models and Applications.* Woley.

Laarne, P., Zaidan, M. A., and Nieminen, T. (2021). ennemi: Non-linear correlation detection with mutual information. *SoftwareX*, 14:100686.

Linfoot, E. (1957). An informational measure of correlation. *Information and Control*, 1(1):85–89.

Lu, S. (2011). Measuring dependence via mutual information. *Master's thesis, Queen's University, Canada.*

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

Manjunath, B. G. and Wilhelm, S. (2021). Moments calculation for the doubly truncated multivariate normal density. *Journal of Behavioral Data Science*, 1(1):17–33.

Meyer, P. E. (2014). *infotheo: Information-Theoretic Measures.* R package version 1.2.0.

Michaud, I. (2018). *rmi: Mutual Information Estimators.* R package version 0.1.1.

Nguyen, T. Q., Dafoe, A., and Ogburn, E. L. (2019). The magnitude and direction of collider bias for binary variables. *Epidemiologic Methods*, 8(1):20170013.

Rohde, A. and Hazlett, C. (20XX). Revisiting sample selection as a threat to internal validity: New lessons, examples, and tools. *XXXX*, XX(XX):XXX–XXX.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10:441–451.

Shevlyakov, G. and Vasilevskiy, N. (2017). A modification of linfoot's informational correlation coefficient. *Austrian Journal of Statistics*, 46(3-4):99–105.

Smith, R. (2015). A mutual information approach to calculating nonlinearity. *Stat*, 4(1):291–303.

Speed, T. (2011). A correlation for the 21st century. *Science*, 334(6062):1502–1503.

Taleb, N. N. (2019). Fooled by correlation: Common misinterpretations in social "science".

A Appendix

A.1 Traditional OVB and its reparameterization

Cinelli and Hazlett (2020) reparameterize omitted variable bias in terms of partial R^2's in the hopes of making sensitivity analysis more straight forward and the sensitivity parameters more interpretable. Traditional OVB analysis uses the Frisch–Waugh–Lovell theorem as follows.

$$\begin{aligned}
\hat{\beta}_{Y\sim D|X,S=1} &= \frac{\widehat{\text{Cov}}(D^{\perp X}, Y^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \frac{\widehat{\text{Cov}}(D^{\perp X}, \hat{\beta}_{Y\sim D|W,X,S=1}D^{\perp X} + \hat{\beta}_{Y\sim W|D,X,S=1}W^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \hat{\beta}_{Y\sim D|W,X,S=1} + \hat{\beta}_{Y\sim W|D,X,S=1}\frac{\widehat{\text{Cov}}(D^{\perp X}, W^{\perp X}|S=1)}{\widehat{\text{Var}}(D^{\perp X}|S=1)} \\
&= \hat{\beta}_{Y\sim D|W,X,S=1} + \hat{\beta}_{Y\sim W|D,X,S=1}\hat{\beta}_{W\sim D|X,S=1}
\end{aligned}$$

Cinelli and Hazlett (2020) then take the following additional steps to rewrite bias.

$$\begin{aligned}
\implies \widehat{\text{bias}} &= \hat{\beta}_{Y\sim D|X,S=1} - \hat{\beta}_{Y\sim D|W,X,S=1} = \hat{\beta}_{Y\sim W|D,X,S=1}\hat{\beta}_{W\sim D|X,S=1} \\
&= \widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(W^{\perp D,X}|S=1)}\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)\frac{\widehat{\text{SD}}(W^{\perp X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)} \\
&= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\frac{\widehat{\text{SD}}(W^{\perp X}|S=1)}{\widehat{\text{SD}}(W^{\perp D,X}|S=1)}\widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1) \\
&= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\frac{\widehat{\text{Cor}}(Y^{\perp D,X}, W^{\perp D,X}|S=1)\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)}{\sqrt{1-\widehat{\text{Cor}}(W^{\perp X}, D^{\perp X}|S=1)^2}}
\end{aligned}$$

We can then see that the magnitude of bias can be written in terms of partial R^2's and summary information that is typical in standard OLS output.

$$\begin{aligned}
\implies |\widehat{\text{bias}}| &= \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\widehat{\text{SD}}(D^{\perp X}|S=1)}\sqrt{\frac{R^2_{YW|D,X,S=1}R^2_{WD|X,S=1}}{1-R^2_{WD|X,S=1}}} \\
&= \text{se}(\hat{\beta}_{Y\sim D|X,S=1})\sqrt{\text{df}_{S=1}\frac{R^2_{YW|D,X,S=1}R^2_{WD|X,S=1}}{1-R^2_{WD|X,S=1}}}
\end{aligned}$$

where $\text{se}(\hat{\beta}_{Y\sim D|X,S=1}) = \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\sqrt{\text{df}_{S=1}}\widehat{\text{SD}}(D^{\perp X}|S=1)}$ is the standard error from the regression using the selected sample and $\text{df}_{S=1}$ are that regression's degrees of freedom.

A.2 $R^2_{DW|X,S=1}$ for binary random variables

Nguyen et al. (2019) provide the expression in Equation 4 for $\text{Cov}(W,D|S=1)$ for binary variables W, D, S in their Lemma 1.

$$\text{Cov}(W,D|S=1) = \frac{1}{P(S=1)^2}\begin{bmatrix} P(W=1,D=1,S=1)P(W=0,D=0,S=1)- \\ P(W=1,D=0,S=1)P(W=0,D=1,S=1) \end{bmatrix} \tag{4}$$

To simplify things, we assume that data are generated according to the the simple collider graph: $D \to S \leftarrow W$. Nguyen et al. (2019) show that in this setting, we can write $\text{Cov}(W,D|S=1)$ as in Equation 5, where $P_{W=w} = P(W=w)$, $P_{D=d} = P(D=d)$, $P_{S=1} = P(S=1)$, and $P_{S=1|wd} = P(S=1|W=w,D=d)$.

$$\text{Cov}(W,D|S=1) = \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P^2_{S=1}}[P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \tag{5}$$

We can then express $\mathrm{Cor}(W, D|S = 1)$ as follows.

$$
\begin{aligned}
\mathrm{Cor}(W, D|S = 1) &= \frac{\mathrm{Cov}(W, D|S = 1)}{\sqrt{\mathrm{Var}(W|S = 1)\mathrm{Var}(D|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P_{S=1}^2\sqrt{\mathrm{Var}(W|S = 1)\mathrm{Var}(D|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \sqrt{\frac{P_{W=1}^2P_{D=1}^2P_{W=0}^2P_{D=0}^2}{P_{S=1}^4P(W = 1|S = 1)P(W = 0|S = 1)P(D = 1|S = 1)P(D = 0|S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \sqrt{\frac{P_{W=1}^2P_{D=1}^2P_{W=0}^2P_{D=0}^2}{P(W = 1, S = 1)P(W = 0, S = 1)P(D = 1, S = 1)P(D = 0, S = 1)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \sqrt{\frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P(S = 1|W = 1)P(S = 1|W = 0)P(S = 1|D = 1)P(S = 1|D = 0)}} \\
&= [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \sqrt{\frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{pmatrix}}}
\end{aligned}
$$

So we see that $R^2_{WD|S=1}$ can be written in terms of six probabilities $(P_{S=1|11}, P_{S=1|00}, P_{S=1|10}, P_{S=1|01}, P_{W=1}, P_{D=1})$ as shown in Equation 6, since $P_{W=0} = 1 - P_{W=1}$ and $P_{D=0} = 1 - P_{D=1}$.

$$
R^2_{WD|S=1} = [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2 \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\begin{pmatrix}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{pmatrix}} \tag{6}
$$

The relationship between W and D in the selected sample ($S = 1$) can be expressed in terms of the relationships between S and W, D, in the full population, where we also need $P(D = 1), P(W = 1)$. In this setting, $P(S = 1)$ is actually not needed directly, since it cancelled out. All of these quantities should be easy for researchers to reason about, since they capture structural (i.e., causal) relationships between the variables.

A.3 $R^2_{DW|X,S=1}$ for truncated multivariate normal random variables

To provide some intuition into how we might try to think about $R^2_{DW|X,S=1}$, we consider the simple case where W, D, S have the causal structure shown in Figure 2 and W, D, S_0 have a multivariate normal joint distribution and $S = \mathbf{1}[S_0 \geq \mathrm{C}]$ for some $\mathrm{C} \in \mathbb{R}$. Here $X = \{\emptyset\}$. S_0 is a hypothesized latent variable that captures how W and D relate to S. The bidirected edge captures that W, D could have some relationship other than that created by conditioning on S. Within the stratum $S = 1$, we have a truncated multivariate normal joint distribution.

The post-selection covariance matrix The pre-selection covariance matrix for S_0, D, W can be written as $\Sigma = \left[\begin{array}{c|cc} \sigma^2_{S_0} & \sigma_{S_0D} & \sigma_{S_0W} \\ \hline \sigma_{S_0D} & \sigma^2_D & \sigma_{WD} \\ \sigma_{S_0W} & \sigma_{WD} & \sigma^2_W \end{array}\right] = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$, where $\Sigma_{11} = \sigma^2_{S_0}$, $\Sigma_{12} = \Sigma_{21}^\top = \begin{bmatrix}\sigma_{S_0D} & \sigma_{S_0W}\end{bmatrix}$, and $\Sigma_{22} = \begin{bmatrix}\sigma^2_D & \sigma_{WD} \\ \sigma_{WD} & \sigma^2_W\end{bmatrix}$. Since we're interested in how the relationships between the variables change due to selection, we're interested in the covariance matrix after truncation, which can be written in terms of the pre-selection covariances: $\Sigma^* = \begin{bmatrix} K_{11} & K_{11}\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{21}\Sigma_{11}^{-1}K_{11} & \Sigma_{22} - \Sigma_{21}(\Sigma_{11}^{-1} - \Sigma_{11}^{-1}K_{11}\Sigma_{11}^{-1})\Sigma_{12}\end{bmatrix} = \begin{bmatrix}\Sigma^*_{11} & \Sigma^*_{12} \\ \Sigma^*_{21} & \Sigma^*_{22}\end{bmatrix}$, where $K_{11} = \sigma^2_{S_0}\left[1 + \frac{\mathrm{C}\phi(\mathrm{C})}{1-\Phi(\mathrm{C})} - \left(\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}\right)^2\right] = \sigma^2_{S_0}\left[1 + \mathrm{C}\gamma - \gamma^2\right]$, letting $\gamma = \frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}$, which is the inverse Mills ratio. (Kotz et al., 2000; Manjunath and Wilhelm, 2021) $S = \mathbf{1}[S_0 \geq \mathrm{C}] \iff P(S = 1) = P(S_0 \geq \mathrm{C}) = P(S_0 \leq -\mathrm{C}) = \Phi(-\mathrm{C}) \iff \mathrm{C} = -\Phi^{-1}(P(S = 1))$ (here we assume $S_0 \sim \mathcal{N}(0, 1)$, which can be done without loss of generality; see below). $\phi(\cdot)$, $\Phi(\cdot)$, and $\Phi^{-1}(\cdot)$ are the pdf, cdf, and quantile function of the standard normal distribution. Now, we're interested in Σ^*_{22}, which contains σ^*_{DW}, the covariance between D and W after truncation.

$$\begin{aligned}
\Sigma_{22}^* &= \Sigma_{22} - \Sigma_{21}(\Sigma_{11}^{-1} - \Sigma_{11}^{-1}K_{11}\Sigma_{11}^{-1})\Sigma_{12} \\
&= \Sigma_{22} - \Sigma_{21}\left(\frac{1}{\sigma_{S_0}^2} - \frac{\sigma_{S_0}^2\left[1 + \mathrm{C}\gamma - \gamma^2\right]}{\sigma_{S_0}^4}\right)\Sigma_{12} \\
&= \Sigma_{22} - \frac{\delta}{\sigma_{S_0}^2}\Sigma_{21}\Sigma_{12}, \text{ where } \delta = \left[1 + \mathrm{C}\gamma - \gamma^2\right] \\
&= \begin{bmatrix} \sigma_D^2 & \sigma_{WD} \\ \sigma_{WD} & \sigma_W^2 \end{bmatrix} - \frac{\delta}{\sigma_{S_0}^2}\begin{bmatrix} \sigma_{S_0D} \\ \sigma_{S_0W} \end{bmatrix}\begin{bmatrix} \sigma_{S_0D} & \sigma_{S_0W} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_D^2 & \sigma_{WD} \\ \sigma_{WD} & \sigma_W^2 \end{bmatrix} - \frac{\delta}{\sigma_{S_0}^2}\begin{bmatrix} \sigma_{S_0D}^2 & \sigma_{S_0D}\sigma_{S_0W} \\ \sigma_{S_0D}\sigma_{S_0W} & \sigma_{S_0W}^2 \end{bmatrix} \\
\implies \sigma_{ab}^* &= \sigma_{ab} - \frac{\sigma_{S_0a}\sigma_{S_0b}}{\sigma_{S_0}^2}\delta, \ \forall a, b \in \{D, W\} \\
\implies \sigma_{DW}^* &= \sigma_{DW} - \frac{\sigma_{S_0D}\sigma_{S_0W}}{\sigma_{S_0}^2}\delta
\end{aligned}$$

We can assume $S_0 \sim \mathcal{N}(0,1)$ WOLOG Suppose that $S_0 = aD + bW + U_{S_0} = X^\top\xi + U_{S_0}$, where $U_{S_0} \sim \mathcal{N}(\mu, \sigma)$, $X = \begin{bmatrix} D & W \end{bmatrix}$, and $\xi = \begin{bmatrix} a & b \end{bmatrix}$. Since D, W, U_{S_0} are all normal random variables, so is S_0. Let $S_0' = \frac{S_0 - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]} = \frac{a}{\mathrm{SD}[S_0]}D + \frac{b}{\mathrm{SD}[S_0]}W + \frac{1}{\mathrm{SD}[S_0]}U_{S_0} - \frac{\mathbb{E}[S_0]}{\mathrm{SD}[S_0]} = X^\top\xi' + \frac{1}{\mathrm{SD}[S_0]}U_{S_0} - \frac{\mathbb{E}[S_0]}{\mathrm{SD}[S_0]}$, where $\xi' = \begin{bmatrix} \frac{a}{\mathrm{SD}[S_0]} & \frac{b}{\mathrm{SD}[S_0]} \end{bmatrix}$. Since we standardized S_0 to get S_0', we know that $S_0' \sim \mathcal{N}(0,1)$. We also know that S_0' is still a linear function of D, W, and U_{S_0}. It's also easy to see how this can be extended so that X and ξ include other variables and path coefficients. Finally, we can see that

$$\begin{aligned}
S = \mathbf{1}[S_0 \geq \mathrm{C}] &= \mathbf{1}\left[\frac{S_0 - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]} \geq \frac{\mathrm{C} - \mathbb{E}[S_0]}{\mathrm{SD}[S_0]}\right] = \mathbf{1}[S_0' \geq \mathrm{C}'] \\
\iff P(S = 1) &= P(S_0' \geq \mathrm{C}') = \Phi(-\mathrm{C}') \iff \mathrm{C}' = -\Phi^{-1}(P(S = 1))
\end{aligned}$$

So we can adjust the path coefficients we're considering and use S_0' rather than S_0 and just think of $S_0 \sim \mathcal{N}(0,1)$. As we saw above, we can then just consider the entire relationship between S and other variables, rather than the relationships with S_0, since we can assume $S_0 \sim \mathcal{N}(0,1)$.

Expression for $R_{WD|S=1}^2$ We can derive an expression similar to the partial correlation formula for truncated correlation and hence R^2. We can see that this is almost identical to the partial correlation formula, but for the δ's. This clarifies the difference between conditioning and truncation for normal random variables.

$$\begin{aligned}
\rho_{WD|S=1} = \rho_{WD|S_0 \geq \mathrm{C}} = \rho_{WD}^* = \frac{\sigma_{WD}^*}{\sigma_D^*\sigma_W^*} &= \frac{\sigma_{WD} - \frac{\sigma_{S_0D}\sigma_{S_0W}}{\sigma_{S_0}^2}\delta}{\sqrt{\sigma_D^2 - \frac{\sigma_{S_0D}^2}{\sigma_{S_0}^2}\delta}\sqrt{\sigma_W^2 - \frac{\sigma_{S_0W}^2}{\sigma_{S_0}^2}\delta}} \\
&= \frac{\rho_{WD}\sigma_D\sigma_W - \frac{\rho_{S_0D}\sigma_{S_0}\sigma_D\rho_{S_0W}\sigma_{S_0}\sigma_W}{\sigma_{S_0}^2}\delta}{\sqrt{\sigma_D^2 - \frac{(\rho_{S_0D}\sigma_{S_0}\sigma_D)^2}{\sigma_{S_0}^2}\delta}\sqrt{\sigma_W^2 - \frac{(\rho_{S_0W}\sigma_{S_0}\sigma_W)^2}{\sigma_{S_0}^2}\delta}} = \frac{\sigma_D\sigma_W\left(\rho_{WD} - \rho_{S_0D}\rho_{S_0W}\delta\right)}{\sigma_D\sigma_W\sqrt{1 - \rho_{S_0D}^2\delta}\sqrt{1 - \rho_{S_0W}^2\delta}} \\
&= \frac{\rho_{WD} - \rho_{S_0D}\rho_{S_0W}\delta}{\sqrt{1 - \rho_{S_0D}^2\delta}\sqrt{1 - \rho_{S_0W}^2\delta}} \\
\rho_{WD|S_0} &= \frac{\rho_{WD} - \rho_{S_0D}\rho_{S_0W}}{\sqrt{1 - \rho_{S_0D}^2}\sqrt{1 - \rho_{S_0W}^2}}
\end{aligned}$$

We see that the relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S_0, W and S_0, D as well as between W and D, in the full population. We also need $P(S = 1)$, the probability of selection or the cut point C, since $\delta = f(P(S = 1))$. If $\rho_{WD} = 0$, then $R_{WD|S=1}^2 = \frac{R_{S_0D}^2 R_{S_0W}^2\delta}{\sqrt{1 - R_{S_0D}^2\delta}\sqrt{1 - R_{S_0W}^2\delta}}$.

Expression in terms of relationships with S, not S_0 We now explore how we can express $R_{WD|S=1}^2$ in terms of the relationships between S, W and S, D, rather than between S_0, W and S_0, D. This is useful, since here S_0 is a hypothesized latent variable, not a substantive variable. We can express ρ_{S_0W} and ρ_{S_0D} in terms of ρ_{SW} and ρ_{SD}. To see this, we borrow two results from Ding and Miratrix (2015). Assume (X_1, X_2) follows a bivariate normal with mean (μ_1, μ_2) and variance $\begin{pmatrix} \sigma_1^2 & \sigma_{12} = \rho_{12}\sigma_1\sigma_2 \\ \sigma_{12} = \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Then for $Z_1 \sim \mathcal{N}(0,1)$, $Z_2 \sim \mathcal{N}(0,1)$, and independent from X_1, X_2 we can write

$$
\begin{aligned}
X_1 &= \mu_1 + \sigma_1 Z_1 \implies Z_1 = \frac{X_1 - \mu_1}{\sigma_1} \\
X_2 &= \mu_2 + \sigma_2 \left[\rho_{12} Z_1 + \sqrt{1-\rho_{12}^2} Z_2\right] \\
&= \mu_2 + \sigma_2 \rho_{12} Z_1 + \sigma_2 \sqrt{1-\rho_{12}^2} Z_2 \\
&= \mu_2 + \sigma_2 \rho_{12} \left[\frac{X_1 - \mu_1}{\sigma_1}\right] + \sigma_2 \sqrt{1-\rho_{12}^2} Z_2 \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1} X_1 + \sigma_2 \sqrt{1-\rho_{12}^2} Z_2 \\
\implies \mathbb{E}[X_2|X_1 \geq \alpha] &= \mathbb{E}[\mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1} X_1 + \sigma_2 \sqrt{1-\rho_{12}^2} Z_2 | X_1 \geq \alpha] \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1} \mathbb{E}[X_1|X_1 \geq \alpha] \\
\mathbb{E}[X_2|X_1 < \alpha] &= \mathbb{E}[\mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1} X_1 + \sigma_2 \sqrt{1-\rho_{12}^2} Z_2 | X_1 < \alpha] \\
&= \mu_2 - \rho_{12}\frac{\sigma_2}{\sigma_1}\mu_1 + \rho_{12}\frac{\sigma_2}{\sigma_1} \mathbb{E}[X_1|X_1 < \alpha] \\
\implies \mathbb{E}[X_2|X_1 \geq \alpha] - \mathbb{E}[X_2|X_1 < \alpha] &= \rho_{12}\frac{\sigma_2}{\sigma_1}\left(\mathbb{E}[X_1|X_1 \geq \alpha] - \mathbb{E}[X_1|X_1 < \alpha]\right) = \rho_{12}\frac{\sigma_2}{\sigma_1}\left(\frac{f_1(\alpha)}{F_1(-\alpha)} - \frac{-f_1(\alpha)}{F_1(\alpha)}\right) \\
&= \rho_{12}\frac{\sigma_2}{\sigma_1} f_1(\alpha)\left(\frac{1}{F_1(-\alpha)} + \frac{1}{F_1(\alpha)}\right) = \rho_{12}\frac{\sigma_2}{\sigma_1} f_1(\alpha)\left(\frac{F_1(\alpha) + F_1(-\alpha)}{F_1(\alpha)F_1(-\alpha)}\right) \\
&= \rho_{12}\frac{\sigma_2}{\sigma_1}\frac{f_1(\alpha)}{F_1(\alpha)F_1(-\alpha)} = \frac{\sigma_{12}}{\sigma_1^2}\frac{f_1(\alpha)}{F_1(\alpha)F_1(-\alpha)}
\end{aligned}
$$

If the marginal distribution of X_1 is $\mathcal{N}(0,1)$ then this becomes $\mathbb{E}[X_2|X_1 \geq \alpha] - \mathbb{E}[X_2|X_1 < \alpha] = \sigma_{12}\frac{\phi(\alpha)}{\Phi(\alpha)\Phi(-\alpha)}$. (Ding and Miratrix, 2015) Therefore, we have

$$
\begin{aligned}
\mathbb{E}[W|S=1] - \mathbb{E}[W|S=0] = \mathbb{E}[W|S_0 \geq \mathrm{C}] - \mathbb{E}[W|S_0 < \mathrm{C}] = \sigma_{S_0W}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})} \\
\mathbb{E}[D|S=1] - \mathbb{E}[D|S=0] = \mathbb{E}[D|S_0 \geq \mathrm{C}] - \mathbb{E}[D|S_0 < \mathrm{C}] = \sigma_{S_0D}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})}
\end{aligned}
$$

For random variables X, B where $B \sim \text{Bernoulli}(p)$, $\text{Cov}(X, B) = \sigma_{xb} = p(1-p)\left[\mathbb{E}(X|B=1) - \mathbb{E}(X|B=0)\right]$. (Ding and Miratrix, 2015) So we see that

$$
\begin{aligned}
\sigma_{SD} &= P(S=1)(1-P(S=1))\left[\mathbb{E}(D|S=1) - \mathbb{E}(D|S=0)\right] \\
&= \Phi(\mathrm{C})\Phi(-\mathrm{C})\left[\mathbb{E}(D|S=1) - \mathbb{E}(D|S=0)\right] \\
&= \Phi(\mathrm{C})\Phi(-\mathrm{C})\sigma_{DL}\frac{\phi(\mathrm{C})}{\Phi(\mathrm{C})\Phi(-\mathrm{C})} = \sigma_{S_0D}\phi(\mathrm{C}) \\
\iff \sigma_{S_0D} &= \frac{\sigma_{SD}}{\phi(\mathrm{C})} \\
\iff \rho_{S_0D} &= \frac{\sigma_{SD}}{\sigma_D\sigma_{S_0}\phi(\mathrm{C})}\frac{\sigma_S}{\sigma_S} = \rho_{DS}\frac{\sigma_S}{\sigma_{S_0}\phi(\mathrm{C})} = \rho_{DS}\frac{\sqrt{P(S=1)(1-P(S=1))}}{\sigma_{S_0}\phi(\mathrm{C})} = \rho_{DS}\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\sigma_{S_0}\phi(\mathrm{C})} = \rho_{DS}\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}.
\end{aligned}
$$

The last equality uses $S_0 \sim \mathcal{N}(0,1)$ We can do the same thing for ρ_{WL}. So we have that $\rho_{S_0D} = \rho_{SD}\xi$ and $\rho_{S_0W} = \rho_{SW}\xi$, where $\xi = \frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}$ can be written as a function of $P(S=1)$. We can then write

$$
\rho_{WD|S=1} = \frac{\rho_{WD} - \rho_{SD}\rho_{SW}\theta}{\sqrt{1-\rho_{SD}^2\theta}\sqrt{1-\rho_{SW}^2\theta}},
$$

where $\theta = \xi^2\delta$ can be written as functions of $P(S=1)$ or C. First, recall that $\xi = \frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}$, $\delta = \left[1 + \mathrm{C}\gamma - \gamma^2\right]$, and $\gamma = \frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}$. So we can write θ in terms of C as follows or in terms of $P(S=1)$ by plugging in $\mathrm{C} = -\Phi^{-1}(P(S=1))$.

$$\begin{aligned}\theta = \xi^2\delta &= \left(\frac{\sqrt{\Phi(\mathrm{C})\Phi(-\mathrm{C})}}{\phi(\mathrm{C})}\right)^2 \left[1 + \mathrm{C}\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})} - \left(\frac{\phi(\mathrm{C})}{1-\Phi(\mathrm{C})}\right)^2\right] \\ &= \frac{\Phi(\mathrm{C})(1-\Phi(\mathrm{C}))}{\phi(\mathrm{C})^2} + \frac{\mathrm{C}\Phi(\mathrm{C})}{\phi(\mathrm{C})} - \frac{\Phi(\mathrm{C})}{1-\Phi(\mathrm{C})}\end{aligned}$$

If $\rho_{WD} = 0$, then $R^2_{WD|S=1} = \frac{R^2_{SD}R^2_{SW}\xi^2\delta}{\sqrt{1-R^2_{SD}\xi^2\delta}\sqrt{1-R^2_{SW}\xi^2\delta}}$. We now see that the relationship between W and D in the selected (truncated) sample can be expressed in terms of the relationships between S, W and S, D as well as between W and D, in the full population, where we also need $P(S=1)$, the probability of selection. All of these quantities should be easy for researchers to have knowledge about and to reason about, since they capture structural (i.e., causal) relationships between the variables.

A.4 "Constant selection effects"

Suppose we would like to assume something like constant treatment effects but for R^2 between D and W after sample selection (e.g., something like $R^2_{WD|S=1}$ equals $R^2_{WD|S=0}$) as a way of simplifying our analysis of $R^2_{WD|S=1}$. What assumptions might make sense? What expression would this provide for $R^2_{WD|S=1}$? First, we expand $\mathrm{Cor}(W,D|S)$ into an expression of $\mathrm{Cor}(W,D|S=1)$ and $\mathrm{Cor}(W,D|S=0)$. Note that this is not a convex combination. That is the coefficients on $\mathrm{Cor}(W,D|S=1)$ and $\mathrm{Cor}(W,D|S=0)$ do not sum to 1.

$$\begin{aligned}\mathrm{Cor}(W,D|S) &= \frac{\mathrm{Cov}(W,D|S)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\ &= \frac{p(S=1)\mathrm{Cov}(W,D|S=1) + p(S=0)\mathrm{Cov}(W,D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\ &= \frac{p(S=1)\mathrm{Cov}(W,D|S=1)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} + \frac{p(S=0)\mathrm{Cov}(W,D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)} \\ &= \frac{p(S=1)\mathrm{SD}(W|S=1)\mathrm{SD}(D|S=1)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)}\mathrm{Cor}(W,D|S=1) + \frac{p(S=0)\mathrm{SD}(W|S=0)\mathrm{SD}(D|S=0)}{\mathrm{SD}(W|S)\mathrm{SD}(D|S)}\mathrm{Cor}(W,D|S=0) \\ &= \sqrt{(A)(B)}\mathrm{Cor}(W,D|S=1) + \sqrt{(1-A)(1-B)}\mathrm{Cor}(W,D|S=0) \\ \text{where } A &= \frac{p(S=1)\mathrm{Var}(W|S=1)}{\mathrm{Var}(W|S)} = \frac{p(S=1)\mathrm{Var}(W|S=1)}{p(S=1)\mathrm{Var}(W|S=1) + p(S=0)\mathrm{Var}(W|S=0)} \in [0,1] \\ B &= \frac{p(S=1)\mathrm{Var}(D|S=1)}{\mathrm{Var}(D|S)} = \frac{p(S=1)\mathrm{Var}(D|S=1)}{p(S=1)\mathrm{Var}(D|S=1) + p(S=0)\mathrm{Var}(D|S=0)} \in [0,1]\end{aligned}$$

If we assume that

- $\mathrm{Cor}(W,D|S=1) = \mathrm{Cor}(W,D|S=0)$; this makes $\mathrm{Cor}(W,D|S) = \left[\sqrt{(A)(B)} + \sqrt{(1-A)(1-B)}\right]\mathrm{Cor}(W,D|S=1)$
- $\mathrm{Var}(W|S=1) = \mathrm{Var}(W|S=0)$; this makes $A = p(S=1)$
- $\mathrm{Var}(D|S=1) = \mathrm{Var}(D|S=0)$; this makes $B = p(S=1)$

These three together make $\mathrm{Cor}(W,D|S) = [p(S=1) + (1-p(S=1))]\,\mathrm{Cor}(W,D|S=1) = \mathrm{Cor}(W,D|S=1) \implies R^2_{WD|S=1} = R^2_{WD|S}$. We can then leverage the partial correlation formula to arrive at

$$R^2_{WD|S=1} = R^2_{WD|S} = \left(\frac{R_{WD} - R_{SW}R_{SD}}{\sqrt{1-R^2_{SW}}\sqrt{1-R^2_{SD}}}\right)^2$$

A.5 An often uninformative bound

In this section, we consider an bound on $R^2_{WD|S=1}$ that follows an approach similar to the last section but where we do not make the assumptions from that section. From above, we have that

$$\mathrm{Cor}(W,D|S) = \sqrt{(A)(B)}\mathrm{Cor}(W,D|S=1) + \sqrt{(1-A)(1-B)}\mathrm{Cor}(W,D|S=0)$$

So we see that

$$
\begin{aligned}
R^2_{WD|S} = \mathrm{Cor}^2(W,D|S) &= \left[\sqrt{(A)(B)}\mathrm{Cor}(W,D|S=1) + \sqrt{(1-A)(1-B)}\mathrm{Cor}(W,D|S=0)\right]^2 \\
&= \underbrace{(A)(B)R^2_{WD|S=1}}_{\geq 0} + \underbrace{(1-A)(1-B)R^2_{WD|S=0}}_{\geq 0} + 2\sqrt{A(1-A)B(1-B)}R_{WD|S=1}R_{WD|S=0}
\end{aligned}
$$

$$
\begin{aligned}
\implies R^2_{WD|S=1} &\leq \min\left(\frac{1}{AB}\left[R^2_{WD|S} - 2\sqrt{A(1-A)B(1-B)}R_{WD|S=1}R_{WD|S=0}\right], 1\right) \\
&\text{We see that } R_{WD|S=1}R_{WD|S=0} \text{ is minimized when } R_{WD|S=1}R_{WD|S=0} = -1. \\
&\leq \min\left(\frac{1}{AB}\left[R^2_{WD|S} + 2\sqrt{A(1-A)B(1-B)}\right], 1\right) \\
&\text{Note that } 2\sqrt{A(1-A)B(1-B)} \text{ is maximized at } \frac{1}{2} \text{ when } A = B = \frac{1}{2}. \\
&= \min\left(\frac{1}{AB}\left[\left(\frac{R_{WD} - R_{SW}R_{SD}}{\sqrt{1-R^2_{SW}}\sqrt{1-R^2_{SD}}}\right)^2 + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)
\end{aligned}
$$

We can show that

$$
\begin{aligned}
\mathrm{Var}(W|S) &= \mathrm{Var}(W) - \frac{\mathrm{Cov}^2(W,S)}{\mathrm{Var}(S)} = \mathrm{Var}(W)\left(1 - \frac{\mathrm{Cov}^2(W,S)}{\mathrm{Var}(W)\mathrm{Var}(S)}\right) = \mathrm{Var}(W)\left(1 - \mathrm{Cor}^2(W,S)\right) = \mathrm{Var}(W)\left(1 - R^2_{SW}\right) \\
\mathrm{Var}(D|S) &= \mathrm{Var}(D) - \frac{\mathrm{Cov}^2(D,S)}{\mathrm{Var}(S)} = \mathrm{Var}(D)\left(1 - \frac{\mathrm{Cov}^2(D,S)}{\mathrm{Var}(D)\mathrm{Var}(S)}\right) = \mathrm{Var}(D)\left(1 - \mathrm{Cor}^2(D,S)\right) = \mathrm{Var}(D)\left(1 - R^2_{SD}\right)
\end{aligned}
$$

This means that

$$
\begin{aligned}
A &= \frac{p(S=1)\mathrm{Var}(W|S=1)}{\mathrm{Var}(W|S)} = \frac{p(S=1)}{1-R^2_{SW}}\frac{\mathrm{Var}(W|S=1)}{\mathrm{Var}(W)} = \frac{p(S=1)}{1-R^2_{SW}}\Theta_W \\
B &= \frac{p(S=1)\mathrm{Var}(D|S=1)}{\mathrm{Var}(D|S)} = \frac{p(S=1)}{1-R^2_{SD}}\frac{\mathrm{Var}(D|S=1)}{\mathrm{Var}(D)} = \frac{p(S=1)}{1-R^2_{SD}}\Theta_D
\end{aligned}
$$

So we get a bound on $R^2_{WD|S=1}$:

$$
\begin{aligned}
R^2_{WD|S=1} &\leq \min\left(\frac{1}{AB}\left[\frac{(R_{WD} - R_{SW}R_{SD})^2}{(1-R^2_{SW})(1-R^2_{SD})} + 2\sqrt{A(1-A)B(1-B)}\right], 1\right) \\
&\text{where } A = \frac{p(S=1)}{1-R^2_{SW}}\Theta_W,\ B = \frac{p(S=1)}{1-R^2_{SD}}\Theta_D,\ \Theta_W = \frac{\mathrm{Var}(W|S=1)}{\mathrm{Var}(W)}, \text{ and } \Theta_D = \frac{\mathrm{Var}(D|S=1)}{\mathrm{Var}(D)}.
\end{aligned}
$$

The relationship between W and D in the selected sample can be expressed in terms of the relationships between S, W and S, D as well as between W and D, in the full population, where we also need $P(S=1)$, Θ_W, and Θ_D. There are at least two problems with this bound. First, Θ_W and Θ_D may not be easy to reason about or to have prior knowledge about. Second, the bound is very often equal to 1. In fact, the bound very often equals 1 when $P(S=1)$ is at all far from 1. So this is not a very useful bound.

A.6 Normalized Scaled Mutual Information Bound

Connecting $R^2_{DW|S=1}$ and $\eta^2_{DW|S=1}$ We start by noting that $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1}$. This is easy to see since $\eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \sup_f \left[\mathrm{Cor}^2(D, f(W)|S=1)\right]$. (Doksum and Samarov, 1995; Chernozhukov et al., 2022) $\eta^2_{D,W|S=1}$ measures portion of the variation in D that can be explained by $\mathbb{E}[D|W, S=1]$, the conditional expectation function (CEF).[15]

[15] The law of total variance tells us that $\mathrm{Var}(D|S=1) = \mathrm{Var}(\mathbb{E}[D|W,S=1]|S=1) + \mathbb{E}[\mathrm{Var}(D|W,S=1)|S=1]$. (Aronow and Miller, 2019)

Correlation and mutual information for Gaussians In order to connect $R^2_{DW|X,S=1}$ and $\eta^2_{DW|X,S=1}$ with mutual information, we draw inspiration from the relationship between R^2 and mutual information for random variables with Gaussian distributions. For random variables, W and D, with a bivariate Gaussian joint distribution, there is an exact relationship between R^2 (i.e., squared correlation coefficient) and mutual information (MI); see Equation 7. (Ihara, 1993; Cover and Thomas, 2006) Can we use something like this transformation to create a useful normalized version of mutual information for non-Gaussian random variables?

$$\mathrm{MI}(W;D) = -\frac{1}{2}\log(1-R^2_{WD}) \iff R^2_{WD} = 1-\exp(-2\times\mathrm{MI}(W;D)) \tag{7}$$

A variation on the L-measure We follow the approach to normalizing mutual information laid out in Lu (2011) in transforming mutual information onto the range [0,1]. This is a variation on the transformation that holds for random variables with Gaussian joint distributions we saw in Equation 7. Many authors have considered this type of transformation of mutual information as a way to obtain something like a non-parametric correlation based on mutual information. See Linfoot (1957); Kent (1983); Joe (1989); Kojadinovic (2005); Speed (2011); Kinney and Atwal (2014); Asoodeh et al. (2015); Smith (2015); Shevlyakov and Vasilevskiy (2017); Laarne et al. (2021). Lu (2011) introduces the L-measure. We define the squared L-measure in Equation 8.

$$L^2(X,Y) \triangleq 1-\exp\left(-2\times\mathrm{IF}\times\mathrm{MI}(X;Y)\right),\ \text{where IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/A)}\right) \text{ and } A = \sup_{U,V\in\mathcal{A}_{X,Y}} \mathrm{MI}(U;V)^{16} \tag{8}$$

IF is a mutual information "inflation factor." We need to increase mutual information so that it goes to infinity when X, Y have a strict dependence for all types of variables, not just continuous variables. (Lu, 2011) shows that

- $A = \min[H(X), H(Y)]$, when X, Y are both discrete. This implies that $\mathrm{IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/\min[H(X),H(Y)])}\right) \geq 1$ since $\frac{\mathrm{MI}(X;Y)}{\min[H(X),H(Y)]} \in [0,1]$. $\mathrm{MI}(X;Y) \leq \min[H(X), H(Y)]$ since $H(X), H(Y)$ are the information content of X, Y. The idea is to inflate mutual information so that $\mathrm{IF}\times\mathrm{MI}(X;Y) \to +\infty$ as X, Y become more dependent. The relationship is a strict dependence when $\mathrm{MI}(X;Y) = \min[H(X), H(Y)]$. So A gives us the right level of inflation.
- $A = H(Y)$, when Y is discrete and X is continuous. This implies that $\mathrm{IF} = \left(\frac{1}{1-(\mathrm{MI}(X;Y)/H(Y))}\right) \geq 1$. Similar ideas apply here as in the last bullet.
- $A = 1$, when X, Y are both continuous which implies that $\mathrm{IF} = 1$, since $\mathrm{MI}(X;Y) = +\infty$ for continuous variables with a strict dependence. Here no inflation is necessary.

This makes the squared L-measure is a good normalization of mutual information in that it ensures that "it is defined for any pair of random variables, it is symmetric, its value lies between 0 and 1, it equals 0 if and only if the random variables are independent, it equals 1 if there is a strict dependence between the random variables, it is invariant under marginal one-to-one transformations of the random variables, and if the random variables are Gaussian distributed, it equals" their R^2. (Lu, 2011)

For our purposes, a first question is: "does something like Equation 9 hold?" That is, when does $R^2_{W,\mathbb{E}[D|W,S=1]|S=1}$ equal $1-\exp\left(-2\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right)$? We know that this would hold when D and $\mathbb{E}[D|W,S=1]$ have a Gaussian joint distribution within $S=1$. For arbitrarily distributed variables, the relationship between D and $\mathbb{E}[D|W,S=1]$ is linear. So we would expect $R^2_{D,\mathbb{E}[D|W,S=1]|S=1}$ and $1-\exp\left(-2\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right)$ to provide similar portraits of the dependency between D and $\mathbb{E}[D|W,S=1]$.

$$\eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} \overset{?}{\approx} 1-\exp\left(-2\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \tag{9}$$

$$\eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = 1-\exp\left(-2\times\boldsymbol{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \tag{10}$$

The question becomes whether we can alter the squared L-measure for $\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$ to exactly recover $\eta^2_{D,W|S=1}$. We do this by introducing an additional mutual information scaling factor $\boldsymbol{\Omega} \triangleq \frac{-\frac{1}{2}\log(1-R^2_{D,\mathbb{E}[D|W,S=1]|S=1})}{\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)} \geq 0$. See Equation 10. This additional scaling factor, $\boldsymbol{\Omega}$, removes any discrepancy between the way that $R^2_{D,\mathbb{E}[D|W,S=1]|S=1}$ and the squared L-measure measure dependence between $D, \mathbb{E}[D|W,S=1]$ on the scale [0,1]. Next, the data processing inequality tells us that $\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1) \leq \mathrm{MI}(W;D|S=1)$, since $\mathbb{E}[D|W,S=1]$ is a function of W.[17] (Cover and Thomas, 2006)

[16] Lu (2011) defines $\mathcal{A}_{X,Y}$, U, and V in the following way: For two arbitrary random variables X and Y, with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, let $\mathcal{A}_{X,Y}$ be the set of all bivariate random vectors (U,V) on $\mathcal{X}\times\mathcal{Y}$ with the same marginal distributions as X and Y. Let $\mathrm{MI}(U;V)$ represent the mutual information between the random variables U and V.

[17] When the relationship between W and D is highly non-linear, $\mathrm{MI}(W;D|S=1)$ may be much larger than $\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$.

It is also easy to see that $L^2_{\mathbf{\Omega}}(a) \triangleq 1 - \exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times a) \in [0,1]$ is a monotonic increasing function of $a \in [0, +\infty)$,[18] which means that $L^2_{\mathbf{\Omega}}(\mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\mathrm{MI}(W; D|S=1))$. Thus, we have the relationship in Equation 11.

$$R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = L^2_{\mathbf{\Omega}}(\mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\mathrm{MI}(W; D|S=1)) \tag{11}$$

What can we say about $\mathbf{\Omega}$? Including $\mathbf{\Omega}$ in $L^2_{\mathbf{\Omega}}(\mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1))$ essentially cancels out $\mathrm{IF} \times \mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)$ and undoes the L-measure transformation to simply return $R^2_{D,\mathbb{E}[D|W,S=1]|S=1}$. This is not a problem, since our goal is simply to find a normalization of mutual information quantities that allows us to write the bound $\eta^2_{D,W|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} \leq L^2_{\mathbf{\Omega}}(\mathrm{MI}(W; D|S=1))$. As we discuss in the next paragraph, we will reason about quantities like $L^2_{\mathbf{\Omega}}(\mathrm{MI})$ directly. We do not need to directly reason about or interpret either the raw mutual information quantities, IF, or $\mathbf{\Omega}$. Moreover, due to the construction of $\mathbf{\Omega}$, it should take values less than or equal to 1; meaning we could instead use the L-measure as a bound. This is because the transformation of $R^2_{D,\mathbb{E}[D|W,S=1]|S=1}$ in the numerator of $\mathbf{\Omega}$ is the transform that turns R^2's into mutual information for Gaussian variables. So it is an approximation to the mutual information between D and $\mathbb{E}[D|W, S=1]$, but limited to their linear relationship. If the relationship between D and $\mathbb{E}[D|W, S=1]$ is fully captured by $R^2_{D,\mathbb{E}[D|W,S=1]|S=1}$, then $\mathbf{\Omega}$ should be very close to 1. If there is some other way that D and $\mathbb{E}[D|W, S=1]$ relate, then $\mathbf{\Omega}$ will be less than 1, since $\mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)$ captures the full relationship and IF appropriately scales mutual information for arbitrary random variables. Therefore, we might choose to consider the L-measure without scaling by $\mathbf{\Omega}$ either as an approximation or as a bound. Simulated examples support this discussion. See Figures 3 and 4.

Normalized scaled mutual information Our approach is to scale and normalize the mutual information using $L^2_{\mathbf{\Omega}}(\cdot)$. Scaling mutual information plays an important role in relating $\eta^2_{D,W|S=1}$ and $\mathrm{MI}(D; W|S=1)$. We will refer to any mutual information quantity scaled by $\mathbf{\Omega} \times \mathrm{IF}$ as scaled mutual information (SMI). Any mutual information quantity that is both scaled and then normalized using $L^2_{\mathbf{\Omega}}(\cdot)$ will be referred to as normalized scaled mutual information (NSMI). NSMI values are much easier to interpret than raw mutual information values. NSMI is a useful measure of dependence between random variables in that it satisfies the properties discussed in Rényi (1959), Smith (2015), Lu (2011), and others as the properties possessed by "an appropriate measure of dependence."[19][20]

1. NSMI is defined for arbitrary pairs of random variables.
2. NSMI is symmetric.
3. NSMI takes values between 0 and 1.
4. NSMI equals 0 if and only if the variables are independent.
5. NSMI equals 1 if and only if the variables a strict dependence (functional relationship).
6. NSMI is invariant to marginal, one-to-one transformations of the variables.
7. If the variables are Gaussian distributed, then NSMI equals their R^2.[21]
8. $\mathrm{NSMI}(D; \mathbb{E}[D|W, S=1]|S=1) = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D,W|S=1}$.

All but the last of these are discussed in Rényi (1959), Smith (2015), and Lu (2011). The last property results from how we've defined NSMI. "Furthermore, MI is invariant under monotonic transformations of variables. This means that the MI correlation coefficient of a non-linear model (X, Y) matches the Pearson correlation of the linearized model $(f(X), g(Y))$. General conditions for f and g are described in" Ihara (1993). (Laarne et al., 2021) This statement focuses on continuous variables and the setting where the linearized model is created using monotonic transformations. $\mathbf{\Omega}$ will equal 1 for a linearized model. So NSMI can be interpreted as the squared Pearson correlation (i.e., R^2) of the linearized model. Figure 9 shows the normalization curve; the normalization of SMI is precisely the normalization that turns mutual information into R^2 for Gaussian variables. Using this terminology, we see that Equation 11 implies Equation 12.

[18] IF in $L^2_{\mathbf{\Omega}}(a)$ is based on $\mathrm{MI}(D; \mathbb{E}[D|W, S=1]|S=1)$.

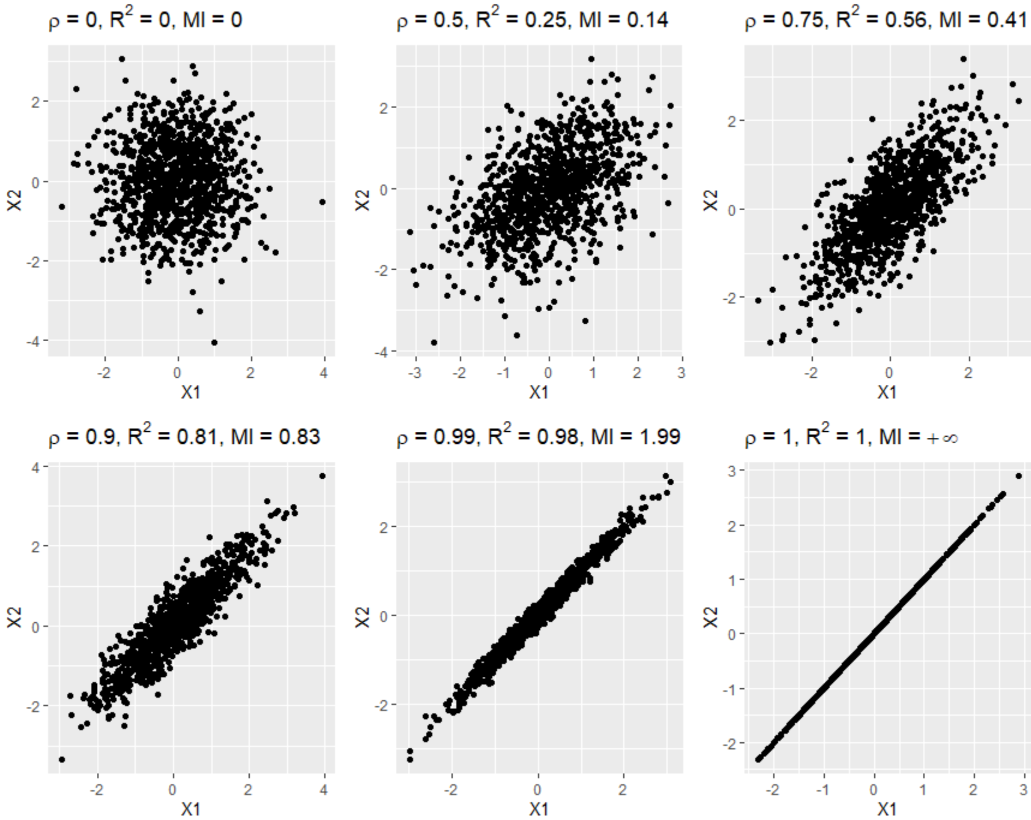
[19] Mutual information satisfies properties 1, 2, 4, and 6. Squared Pearson correlation (i.e., R^2) satisfies properties 1, 2, 3, 5, and 7.

[20] The transformation $\ell^2(\mathrm{MI}(X;Y)) = 1 - \exp(-2 \times \mathrm{MI}(X;Y))$ ensures that properties 2, 3, 6, and 7 are satisfied; it is the transformation that turns mutual information into an R^2 for Gaussian distributed variables. The transformation $L^2(\mathrm{MI}(X;Y)) = 1 - \exp(-2 \times \mathrm{IF} \times \mathrm{MI}(X;Y))$ is the square of Lu (2011)'s L-measure, where IF is chosen to ensure that properties 1 and 5 are satisfied, while also maintaining properties 2, 3, 6, and 7. The transformation $\mathrm{NSMI}(X;Y) \triangleq L^2_{\mathbf{\Omega}}(\mathrm{MI}(X;Y)) = 1 - \exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(X;Y))$ is our normalized and scaled measure of mutual information, where $\mathbf{\Omega} \geq 0$ is also chosen to ensure that property 8 is satisfied, while also maintaining properties 1 through 7. Lu (2011) demonstrates that properties 1 through 7 hold for the L-measure. Given this, it is trivial to see that they also hold for NSMI.

[21] It is worth noting that, although we might be more comfortable thinking about correlations and R^2's, they are not necessarily capturing what we expect. "Mutual Information is a nonlinear function of ρ which in fact makes it additive. Intuitively, in the Gaussian case, ρ should never be interpreted linearly: a ρ of $\frac{1}{2}$ carries ≈ 4.5 times the information of a $\rho = \frac{1}{4}$, and a ρ of $\frac{3}{4}$ 12.8 times!" (Taleb, 2019) "One needs to translate ρ into information. See how $\rho = 0.5$ is much closer to $[\rho =]0$ than to a $\rho = 1$. There are considerable differences between .9 and .99." (Taleb, 2019) See Figure 8 for a series of plots that illustrate how changes in correlation and R^2 compare to changes in mutual information for standard Gaussian random variables. See Figure 9 for a plot of the relationship between mutual information and R^2 for Gaussian variables, this is also the normalization curve we use.

$$R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq \text{NSMI}(W; D|S=1) \tag{12}$$

Figure 8: Correlation is non-linear. Scatter plots of standard Gaussian random variables with different correlations. Correlation of 0.5 is much more similar to correlation of 0 than to correlation of 1.

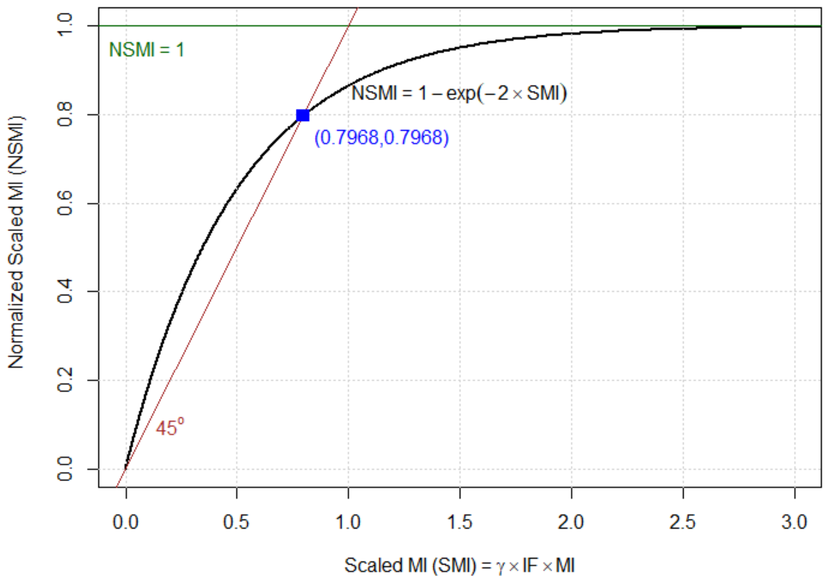


Mutual information bounds Equation 12 seems nice. But have we solved our original problem of finding a bound on $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ in terms of structural descriptions of the relationships between the variables in the population? No we haven't. $\text{MI}(W; D|S=1)$ and $\text{NSMI}(W; D|S=1)$ both contain the spurious association between W and D created by sample selection. We now aim to find structural descriptions of the relationships between the variables in the population that can bound $\text{MI}(W; D|S=1)$. These can then be normalized to provide bounds on $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$. We start by considering $\text{MI}(D; W|S)$. Using properties of mutual information (Cover and Thomas, 2006), we can show Equation 13.

$$\begin{aligned}\text{MI}(D; W|S) &= \text{MI}(D; W) + \text{MI}(S; D|W) - \text{MI}(S; D) \\ &= \text{MI}(D; W) + [\text{MI}(S; [D, W]) - \text{MI}(S; W)] - \text{MI}(S; D) \\ &= \text{MI}(D; W) + \text{MI}(S; [D, W]) - \text{MI}(S; D) - \text{MI}(S; W)\end{aligned} \tag{13}$$

$\text{MI}(S; [D, W]) = \text{MI}(S; W) + \text{MI}(S; D|W)$ is the mutual information between S and $[D, W]$ jointly. We now consider bounds on $\text{MI}(D; W|S=1)$. When S is binary, two positive terms (one for $S=1$ and one for $S=0$) are being summed to create $\text{MI}(D; W|S)$. See Equation 14.

Figure 9: Normalization of Scaled Mutual Information



$$
\begin{aligned}
\mathrm{MI}(D;W|S) &= \int_S D_{\mathrm{KL}}\left(P_{(D,W)|S} \| P_{D|S} \otimes P_{W|S}\right) dP_S \\
&= \sum_{s\in\{0,1\}} p(S=s) \sum_d \int_w p(d,w|S=s) \log\left[\frac{p(d,w|S=s)}{p(d|S=s)p(w|S=s)}\right] dd\; dw \\
&= \sum_{s\in\{0,1\}} p(S=s) D_{\mathrm{KL}}\left(P_{(D,W)|S=s} \| P_{D|S=s} \otimes P_{W|S=s}\right) \\
&= p(S=1) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=1} \| P_{D|S=1} \otimes P_{W|S=1}\right) + p(S=0) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=0} \| P_{D|S=0} \otimes P_{W|S=0}\right) \\
&= p(S=1)\mathrm{MI}(D;W|S=1) + p(S=0)\mathrm{MI}(D;W|S=0)
\end{aligned}
\tag{14}
$$

From Equations 13 and 14, we have that

$$
\begin{aligned}
\mathrm{MI}(D;W|S=1) &\le \frac{\mathrm{MI}(D;W|S)}{p(S=1)} \\
&= \frac{\mathrm{MI}(D;W) + \mathrm{MI}(S;[D,W]) - \mathrm{MI}(D;S) - \mathrm{MI}(W;S)}{p(S=1)} \\
&= \frac{\mathrm{MI}(D;W) + \mathrm{MI}(S;D|W) - \mathrm{MI}(D;S)}{p(S=1)} \\
&\le \frac{\mathrm{MI}(D;W) + \mathrm{MI}(S;[D,W])}{p(S=1)}
\end{aligned}
\tag{15}
$$

This gives us the simple results in Theorem 1.

Theorem 1. *For random variables D, W, S, conditioning on S alters the relationship between D and W according to the expression $MI(D;W|S) = MI(D;W) + MI(S;[D,W]) - MI(S;D) - MI(S;W)$. Therefore, the change in dependence due to conditioning on S can be characterized using mutual information according to $MI(D;W|S) - MI(D;W) = MI(S;[D,W]) - MI(S;D) - MI(S;W)$. The dependence is not changed when $MI(S;[D,W]) = MI(S;D) + MI(S;W)$.*

When S is binary, it is also possible to write $MI(D;W|S) = p(S=1)MI(D;W|S=1) + p(S=0)MI(D;W|S=0)$, meaning that $MI(D;W|S=1) \le \frac{MI(D;W|S)}{p(S=1)} = \frac{MI(D;W)+MI(S;[D,W])-MI(D;S)-MI(W;S)}{p(S=1)}$.

So we see that we have a bound on $\mathrm{MI}(D;W|S=1)$. Every component of these bounds is something that we might have external knowledge or intuition on. From Theorem 1, we have a few relationships we can consider as bounds on $\mathrm{MI}(D;W|S=1)$. Others are also likely possible.

1. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D;W])-\mathrm{MI}(D;S)-\mathrm{MI}(W;S)}{p(S=1)}$
2. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(D;S|W)-\mathrm{MI}(D;S)}{p(S=1)}$
3. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(W;S|D)-\mathrm{MI}(W;S)}{p(S=1)}$
4. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(D;S|W)}{p(S=1)}$
5. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(W;S|D)}{p(S=1)}$
6. $\mathrm{MI}(D;W|S=1) \leq \frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D;W])}{p(S=1)}$

It is important to note that these vary in the tightness of the bound. The first three bounds are all equivalent. But the last three are not as tight, since these involve the exclusion of at least one term that is subtracted from the numerator of the first three bounds. If W and D are marginally independent, then the term $\mathrm{MI}(D;W)$ will be zero in all the bounds.

Interpretable bounds on $R^2_{DW|S=1}$ and $\eta^2_{DW|S=1}$ We now combine Equation 12 with Equation 15 to get interpretable bounds on $R^2_{DW|S=1}$ and $\eta^2_{DW|S=1}$. We start by considering only one such bound. But others are possible.

$$
\begin{aligned}
R^2_{DW|S=1} &\leq \eta^2_{DW|S=1} = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} \\
&= 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \text{ since } \mathbf{\Omega} = \frac{-\frac{1}{2}\log(1-R^2_{W,\mathbb{E}[W|D,S=1]|S=1})}{\mathrm{IF}\times\mathrm{MI}(W;\mathbb{E}[W|D,S=1]|S=1)} \\
&\leq 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(W;D|S=1)\right) = \mathrm{NSMI}(W;D|S=1) \text{ by the data processing inequality} \\
&\leq 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\frac{\mathrm{MI}(D;W)+\mathrm{MI}(S;[D,W])}{p(S=1)}\right) \text{ by Eqn. 15} \\
&= 1-\exp\left(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W)-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W])\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left(\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W))\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W]))\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left([1-1+\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W))][1-1+\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;D,W]))]\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left([1-(1-\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(D;W)))][1-(1-\exp(-2\times\mathbf{\Omega}\times\mathrm{IF}\times\mathrm{MI}(S;[D,W])))]\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left(\ [1-\mathrm{NSMI}(D;W)][1-\mathrm{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}}
\end{aligned}
\tag{16}
$$

Therefore, our first interpretable bound is captured by Equation 17.

$$
R^2_{DW|S=1} \leq \eta^2_{DW|S=1} \leq 1-\left(\ [1-\mathrm{NSMI}(D;W)][1-\mathrm{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}} \tag{17}
$$

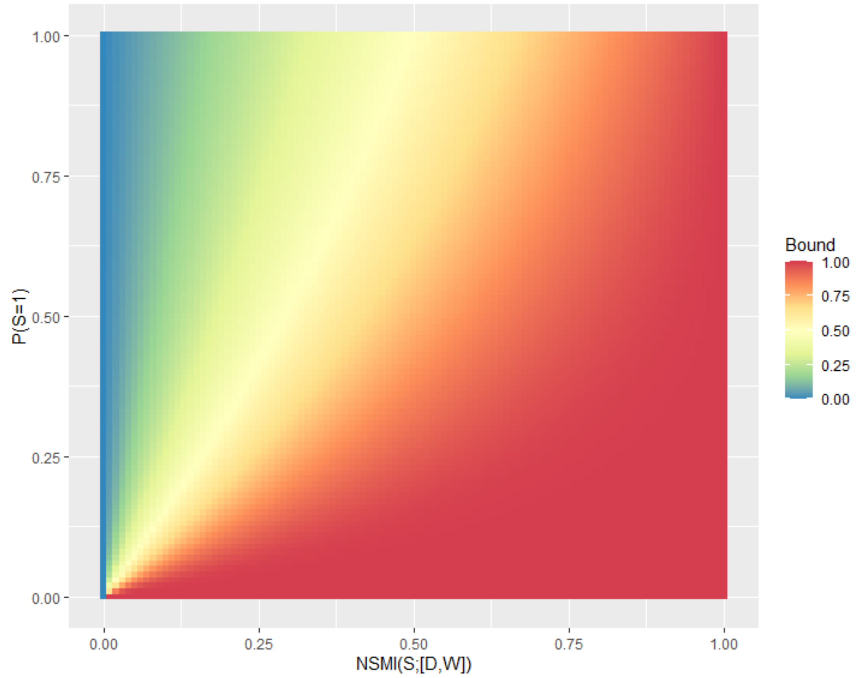
This bound is an expression of normalized scaled mutual information for the marginal mutual information between D and W, for the mutual information between S and $[D,W]$ together, and the probability of selection, $P(S=1)$. As we saw in the case of binary random variables and truncated normal random variables, we have an expression in terms of structural (i.e., causal) relationships between the variables in the full population. In Figure 10, we show how the bound in Equation 17 changes for different values of $\mathrm{NSMI}(S;[D,W])$ and $p(S=1)$. For this, we assume that that W,D are marginally independent and so $\mathrm{NSMI}(D;W)=0$ and the bound becomes $B \triangleq 1-(1-\mathrm{NSMI}(S;[D,W]))^{\frac{1}{p(S=1)}}$. As $p(S=1)\to 1$, $B\to \mathrm{NSMI}(S;[D,W])$. As $p(S=1)\to 0$, $B\to 1$. As $\mathrm{NSMI}(S;[D,W])\to 1$, $B\to 1$. As $\mathrm{NSMI}(S;[D,W])\to 0$, $B\to 0$. These dynamics are easy to see in the expression for the bound itself. They reflect the bounds on $\mathrm{MI}(W;D|S=1)$ that we then scale and normalize. It is worth noting that this bound is not always informative (i.e., smaller than 1); small probabilities of selection can lead to high bounds, regardless of the value for $\mathrm{NSMI}(S;[D,W])$. This reflects that, when the selection probability is small, $\mathrm{NSMI}(S;[D,W])$ carries much less information about the stratum $S=1$ than it does the stratum $S=0$.

Following a similar approach as we did in obtaining the bound in Equation 17, we arrive at Theorem 2.

Theorem 2. *For random variables D,W,S, for which S is a collider on a path from D to W in G^{+}_{S} that, if conditioned on, could alter the relationship between D and W (e.g., $D\to S\leftarrow W$), the $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ resulting after stratification to $S=1$ can be bounded in the following ways:*

1. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1-\left(\frac{[1-NSMI(D;W)][1-NSMI(S;[D,W])]}{[1-NSMI(S;D)][1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
2. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1-\left(\frac{[1-NSMI(D;W)][1-NSMI(D;S|W)]}{[1-NSMI(S;D)]}\right)^{\frac{1}{p(S=1)}}$

Figure 10: Bounds (from Equation 17) on $R^2_{D,W|S=1}$ and $\eta^2_{D,W|S=1}$ given values for NSMI$(S;[D,W])$ and $p(S=1)$ and assuming NSMI$(D;W)=0$



3. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - \left(\frac{[1-NSMI(D;W)][1-NSMI(W;S|D)]}{[1-NSMI(S;W)]}\right)^{\frac{1}{p(S=1)}}$
4. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(D;S|W)])^{\frac{1}{p(S=1)}}$
5. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(W;S|D)])^{\frac{1}{p(S=1)}}$
6. $R^2_{D,W|S=1} \leq \eta^2_{D,W|S=1} \leq 1 - ([1-NSMI(D;W)][1-NSMI(S;[D,W])])^{\frac{1}{p(S=1)}}$

Bounds 1 through 3 in Theorem 2 are tighter than bounds 4 through 6, but require additional sensitivity parameters as well as some knowledge about how mutual information works. That is, since some of the NSMI quantities are related in the bounds in Theorem 2, users need to take care to reason about coherent combinations of the NSMI quantities. In particular, the bounds all take the form $1-(\tau)^{\frac{1}{p(S=1)}}$ but with different τ; τ must take a value between 0 and 1. This reflects the fact that $1-(1-\text{NSMI}(W;D|S))^{\frac{1}{p(S=1)}}$ equals bounds 1 through 3 and NSMI$(W;D|S)$ takes values between 0 and 1. This, in turn, reflects that $\text{MI}(D;W|S) = \text{MI}(D;W) + \text{MI}(S;[D,W]) - \text{MI}(S;D) - \text{MI}(S;W) \geq 0$. For this reason, we encourage users unfamiliar with mutual information to use bounds 4 through 6, where the condition that $\tau \in [0,1]$ will always be satisfied given NSMI values between 0 and 1. If W and D are assumed to be marginally independent, then NSMI$(D;W)=0$ and this term can be removed from the bounds. Which bound is most useful depends on the relationships that practitioners feel comfortable reasoning about in terms of NSMI's.

Incorporating Covariates It is also fairly straightforward to incorporate covariates, X. We now turn to bounding $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$. The approach is very similar to the above. Equation 18 follows from the usual expressions of $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$ and the fact that $R^2_{D\sim W,X|S=1} \leq \eta^2_{D\sim W,X|S=1}$.[22]

$$\begin{aligned} R^2_{D,W|X,S=1} &= \frac{R^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \leq \frac{\eta^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \\ \eta^2_{D,W|X,S=1} &= \frac{\eta^2_{D\sim W,X|S=1} - \eta^2_{D\sim X|S=1}}{1 - \eta^2_{D\sim X|S=1}} \end{aligned} \tag{18}$$

[22]We are not able to directly link $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$ as we did the versions that did not include X. If X has a very non-linear relationship with D and/or W, then it is not clear how $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$ relate. In this discussion, we simply bound them separately.

We can estimate $R^2_{D\sim X|S=1}$ and $\eta^2_{D\sim X|S=1}$ in Equation 18 from the selected sample, since neither involves W. Since both portions of Equation 18 are expressions of things we can estimate from the data and $\eta^2_{D\sim W,X|S=1}$, we now turn to bounding $\eta^2_{D\sim W,X|S=1}$ in Equation 19. Note that, as in the above discussion, $\mathbf{\Omega}$ should take values less than or equal to 1. So we could chose to omit it and simply reason about the L-measure as a bound. See the above discussion.

$$\eta^2_{D\sim W,X|S=1} = 1 - \exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;[W,X]|S=1)) \text{ where } \mathbf{\Omega} = \frac{-\frac{1}{2}\log(1-\eta^2_{D\sim W,X|S=1})}{\mathrm{IF} \times \mathrm{MI}(D;[W,X]|S=1)} \tag{19}$$

From Equation 19, we have two options for how to proceed. First, we could use Theorem 1 with W replaced with $[W,X]$ to arrive at Equation 20.

$$\mathrm{MI}(D;[W,X]|S=1) \leq \frac{\mathrm{MI}(D;[W,X]|S)}{p(S=1)} = \frac{\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)}{p(S=1)} \tag{20}$$

Using Equations 19 and 20 we arrive at Equation 21.

$$\begin{aligned}
\eta^2_{D\sim W,X|S=1} &= 1 - \exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;[W,X]|S=1)) \\
&\leq 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \left[\frac{\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)}{p(S=1)}\right]\right) \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times [\mathrm{MI}(D;[W,X]) + \mathrm{MI}(S;[D,W,X]) - \mathrm{MI}(D;S) - \mathrm{MI}([W,X];S)]\right)^{\frac{1}{p(S=1)}} \\
&= 1 - \left[\frac{\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;[W,X]))\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(S;[D,W,X]))}{\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;S))\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}([W,X];S))}\right]^{\frac{1}{p(S=1)}} \\
&= 1 - \left[\frac{[1-\mathrm{NSMI}(D;[W,X])][1-\mathrm{NSMI}(S;[D,W,X])]}{[1-\mathrm{NSMI}(D;S)][1-\mathrm{NSMI}([W,X];S)]}\right]^{\frac{1}{p(S=1)}}
\end{aligned} \tag{21}$$

Second, we could use Theorem 1 with everything conditioned on X and the fact that $\mathrm{MI}(D;W|X,S) = p(S=1)\mathrm{MI}(D;W|X,S=1) + p(S=0)\mathrm{MI}(D;W|X,S=0)$ to arrive at the second equation in Equation 22. The first equation in Equation 22 just comes from the definition of $\mathrm{MI}(D;[W,X]|S=1)$.

$$\begin{aligned}
\mathrm{MI}(D;[W,X]|S=1) &= \mathrm{MI}(D;X|S=1) + \mathrm{MI}(D;W|X,S=1) \text{ and} \\
\mathrm{MI}(D;W|X,S=1) &\leq \frac{\mathrm{MI}(D;W|X,S)}{p(S=1)} = \frac{\mathrm{MI}(D;W|X) + \mathrm{MI}(S;[D,W]|X) - \mathrm{MI}(D;S|X) - \mathrm{MI}(W;S|X)}{p(S=1)}
\end{aligned} \tag{22}$$

Using Equations 19 and 22 we arrive at Equation 23.

$$\begin{aligned}
\eta^2_{D\sim W,X|S=1} &= 1 - \exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;[W,X]|S=1)) \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times [\mathrm{MI}(D;X|S=1) + \mathrm{MI}(D;W|X,S=1)]\right) \\
&\leq 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \left[\mathrm{MI}(D;X|S=1) + \frac{\mathrm{MI}(D;W|X) + \mathrm{MI}(S;[D,W]|X) - \mathrm{MI}(D;S|X) - \mathrm{MI}(W;S|X)}{p(S=1)}\right]\right) \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;X|S=1)\right) \\
&\times \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \frac{\mathrm{MI}(D;W|X) + \mathrm{MI}(S;[D,W]|X) - \mathrm{MI}(D;S|X) - \mathrm{MI}(W;S|X)}{p(S=1)}\right) \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;X|S=1)\right) \\
&\times \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times [\mathrm{MI}(D;W|X) + \mathrm{MI}(S;[D,W]|X) - \mathrm{MI}(D;S|X) - \mathrm{MI}(W;S|X)]\right)^{\frac{1}{p(S=1)}} \\
&= 1 - \exp\left(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;X|S=1)\right) \\
&\times \left[\frac{\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;W|X))\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(S;[D,W]|X))}{\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(D;S|X))\exp(-2 \times \mathbf{\Omega} \times \mathrm{IF} \times \mathrm{MI}(W;S|X))}\right]^{\frac{1}{p(S=1)}} \\
&= 1 - [1-\mathrm{NSMI}(D;X|S=1)]\left[\frac{[1-\mathrm{NSMI}(D;W|X)][1-\mathrm{NSMI}(S;[D,W]|X)]}{[1-\mathrm{NSMI}(D;S|X)][1-\mathrm{NSMI}(W;S|X)]}\right]^{\frac{1}{p(S=1)}}
\end{aligned} \tag{23}$$

Equations 18, 21, and 23 combine to provide the following bounds on $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$. As before, which bound is most useful depends on what the researcher is most comfortable reasoning about.

Theorem 3. *For random variables D, W, S, X, for which S is a collider on a path from D to W in G_S^+ that, if conditioned on, could alter the relationship between D and W (e.g., $D \rightarrow S \leftarrow W$), the $R^2_{D,W|X,S=1}$ and $\eta^2_{D,W|X,S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

1. $R^2_{D,W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
2. $R^2_{D,W|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right)$
3. $\eta^2_{D,W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - \left[\frac{[1-NSMI(D;[W,X])][1-NSMI(S;[D,W,X])]}{[1-NSMI(D;S)][1-NSMI([W,X];S)]}\right]^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$
4. $\eta^2_{D,W|X,S=1} \leq \frac{1}{1-\eta^2_{D\sim X|S=1}} \times \left(1 - [1 - NSMI(D;X|S=1)]\left[\frac{[1-NSMI(D;W|X)][1-NSMI(S;[D,W]|X)]}{[1-NSMI(D;S|X)][1-NSMI(W;S|X)]}\right]^{\frac{1}{p(S=1)}} - \eta^2_{D\sim X|S=1}\right)$

where $R^2_{D\sim X|S=1}$ or $\eta^2_{D\sim X|S=1}$ is estimated from the data. We can approximate or inform the choice of $NSMI(D;X|S=1)$ using the estimated $R^2_{D\sim X|S=1}$ or $\eta^2_{D\sim X|S=1}$.[23] *These bounds are all analogous to bound 1 in Theorem 2. Analogs to bounds 2 - 6 in Theorem 2 could also be formed.*

[23] We cannot directly estimate NSMI$(D;X|S=1)$, since we cannot estimate $\mathbf{\Omega}$ or IF which are based on $\eta^2_{D\sim W,X|S=1}$ and MI$(D;[W,X]|S=1)$.