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 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. 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 the general, non-parametric settings by drawing on 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? 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 often 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.[1] Suppose further that the investigator knows or is willing to assume 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.[2] Figure 1 presents 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 (DAGs) 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. Using these parametric settings as inspiration, we then propose bounds 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 brief discussion.

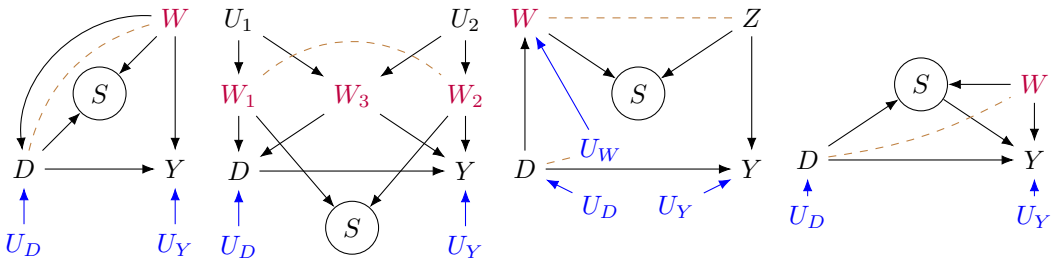
[*]Department of Statistics, UCLA. adamrohde@ucla.edu

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

[1]See Rohde and Hazlett (20XX) for further discussion of internal validity, causal effects for the selected sample, and causal effects for the subpopulation for which the selected sample is representative.

[2]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

We review 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 for or bounds on such bias as the basis for a sensitivity analysis. These settings are those considered in Cinelli and Hazlett (2020) and Chernozhukov et al. (2022).

Linear Model We may be interested in estimating a linear regression model, using the selected sample, like in Equation 1a. 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, like in Equation 1b. 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" regressions, respectively.

$$\left[Y = \beta_{Y\sim D|X,S=1} D + X\beta_{Y\sim X|D,S=1} + \epsilon_s\right] | S=1 \tag{1a}$$

$$\left[Y = \beta_{Y\sim D|W,X,S=1} D + X\beta_{Y\sim X|D,W,S=1} + \beta_{Y\sim W|D,X,S=1} W + \epsilon_l\right] | S=1 \tag{1b}$$

Partially Linear Model Alternatively, we may be interested in estimating a partially linear model, as in Equation 2b. But, as in the fully linear case, they are only able to estimate Equation 2a, which omits W. In both the linear and partially linear cases, we assume that the user is responsibly considering how a linear or partially linear model (and the inclusion of a covariate in these models) differs from a fully non-parametric setting, before considering that they want to know how inclusion of W in the model changes θ_s.

$$[Y = \theta_s D + f_s(X) + \epsilon_s] | S=1 \tag{2a}$$

$$[Y = \theta_l D + f_l(X, W) + \epsilon_l] | S=1 \tag{2b}$$

Non-parametric Model 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 Equation 3b 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[3] under intervention to set $D=d$. Again, the investigator is only able to estimate Equation 3a, where $f^*_Y(D, X) \triangleq \mathbb{E}[Y|D, X] = \mathbb{E}[f_Y(D, X, W)|D, X]$.

$$\theta_s = \mathbb{E}[f^*_Y(1, X) - f^*_Y(0, X)|S=1] \tag{3a}$$

$$\theta_l = \mathbb{E}[Y_1 - Y_0|S=1] = \mathbb{E}[f_Y(1, X, W) - f_Y(0, X, W)|S=1] \tag{3b}$$

In all three settings, there will be some bias resulting from not adjusting for W in our estimate. The bias is $\theta_s - \theta_l$. The cause of the bias may be sample selection or common cause confounding or both. Again, see Rohde and Hazlett (20XX) for more on how sample selection can be thought of as an omitted variable problem. 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.

2.1 Expressions for omitted variable bias

For each of the settings above, Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide expressions for or bounds on the omitted variable bias that can be expressed in terms of simple sensitivity parameters that *capture the relationships between the variables in the selected sample.*

[3] See Rohde and Hazlett (20XX) for more discussion of structural causal models under sample selection.

Linear Model Following Cinelli and Hazlett (2020), we can show that omitted variable bias for internally valid OLS regression can be expressed as in Equation 4.

$$\widehat{|\text{bias}|} = \widehat{\text{se}}(\hat{\beta}_{Y\sim D|X,S=1}) \sqrt{\text{df}_{S=1} \frac{R^2_{Y\sim W|D,X,S=1} R^2_{W\sim D|X,S=1}}{1 - R^2_{W\sim D|X,S=1}}} \tag{4}$$

$\widehat{\text{se}}(\hat{\beta}_{Y\sim D|X,S=1}) = \frac{\widehat{\text{SD}}(Y^{\perp D,X}|S=1)}{\sqrt{\text{df}_{S=1}}\widehat{\text{SD}}(D^{\perp X}|S=1)}$ is equal to the standard error running a regression using the selected sample using typical statistical software and $\text{df}_{S=1}$ are that regression's degrees of freedom. See Appendix A Section A.1 for the full derivation. $R^2_{Y\sim W|D,X,S=1}$ is a partial R^2 that equals the fraction of residual variation in Y explained by W after partialling out both D and X, in the selected sample. $R^2_{W\sim D|X,S=1}$ is a partial R^2 that equals the fraction of the residual variation in D explained by W after partialling out X, in the selected sample. See Cinelli and Hazlett (2020) for further discussion of how to interpret partial R^2s.

Partially Linear Model Chernozhukov et al. (2022) show that omitted variable bias in partially linear setting can be bounded by an expression in terms of $\eta^2_{Y\sim W|D,X,S=1}$, $\eta^2_{D\sim W|X,S=1}$, and terms estimable from the data. $\eta^2_{Y\sim W|D,X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ are Pearson's correlation ratios (or non-parametric R^2s).[4] $\eta^2_{Y\sim W|D,X,S=1}$ is the proportion of residual variation in Y explained by W. $\eta^2_{D\sim W|X,S=1}$ is the proportion of residual variation in D explained by W. See Chernozhukov et al. (2022) for further discussion of how to interpret partial η^2s.

Non-parametric Model Chernozhukov et al. (2022) also show that omitted variable bias in fully non-parametric setting can be bounded by an expression in terms of $\eta^2_{Y\sim W|D,X,S=1}$ 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 $\eta^2_{D\sim W|X,S=1}$.

The sensitivity parameters above are either R^2s or η^2s, quantities with which most researchers will have some familiarity. While these may be familiar measures of dependence, they do not have all properties of dependence measures that are desirable. For example, an R^2 of zero can exist for dependent variables and an R^2 reflects only the linear relationship between variables. But they do have several useful properties and certain appeal as measures of dependence. See Rényi (1959) for a discussion of these measures of dependence and the properties that make good measures of dependence. Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) provide thorough discussions of sensitivity analysis using these bias expressions and reasoning about these types of sensitivity parameters, 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. That is, ignoring any limitations as measures of dependence, researchers intending to conduct a sensitivity analysis using these approaches must be able to build an understanding of the relationships between the variables based on first principles, existing literature, intuition, and subject matter expertise. Such understanding will, necessarily, reflect causal relations between the variables.[5] As we discuss next, obtaining such knowledge is complicated by sample selection.

2.2 Difficulty interpreting sensitivity parameters in selected samples

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 alone, we allow for either of the sensitivity parameters $R^2_{Y\sim W|D,X,S=1}$ or $R^2_{W\sim D|X,S=1}$ (or $\eta^2_{Y\sim W|D,X,S=1}$ or $\eta^2_{D\sim W|X,S=1}$) to contain a purely statistical relationship that results from stratifying to $S=1$ where S is a collider, rather than just causal relationships that operate in the full population. 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 conditioning on a collider due to sample selection.[6] This is because the association captured by the sensitivity parameter does not result from structural relationships in which one variable causes another.

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

[5] It is important to recognize that the R^2s or η^2s in the expressions for or bounds on omitted variable bias from Cinelli and Hazlett (2020) and Chernozhukov et al. (2022) are statistical measures of dependence between the variables. These could measure direct causal realtionships between variables. Often, however, the variables might not have a direct causal relationship. In such cases, the R^2s or η^2s may be capturing a chain of causal relationships. Researchers should be clear about such causal relationships when building their understanding of the relationship between W and Y. There could be more information available about the constituent relationships or maybe the true relationship becomes murkier.

[6] By "purely statistical relationship," we mean one that arises due to conditioning, as opposed to a relationship that exists causally in the population from which the sample was selected.

Instead, the association results from or is changed by the often counterintuitive phenomenon of conditioning on a common effect (a collider). Such associations do not exist in the population from which the sample was selected and will be difficult to understand from first principles, previous studies (unless those studies suffered from similar sample selection), intuition, or subject matter expertise concerning mechanisms. See the worked example below for an example. 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.[7] In our discussion, we focus on $R^2_{W\sim D|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. Similar discussion could apply to $R^2_{Y\sim W|D,X,S=1}$ or $\eta^2_{Y\sim W|D,X,S=1}$. 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 for how these sensitivity parameters can be expressed in terms of structural relationships between the variables in the full population in simple parametric settings.

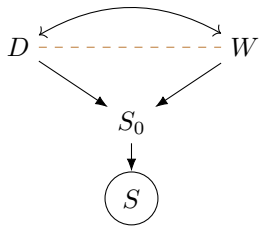
Binary random variables To provide some intuition about how we will try to understand $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$, we consider the case where W, D are binary. We assume that data are generated according to a simple collider graph: $D \rightarrow S \leftarrow W$. Here $X = \{\emptyset\}$. $R^2_{D\sim W|S=1}$ can be written in terms of six probabilities as shown in Equation 5.[8] See Appendix A Section A.2 for the complete derivation.

$$R^2_{D\sim W|S=1} = [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2 \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\left(\begin{array}{l}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{array}\right)} \tag{5}$$

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 capture structural (i.e., causal) relationships between the variables. The key here is that $R^2_{D\sim W|X,S=1}$ can be understood in terms of structural relationships in the full population.

Truncated multivariate normal random variables Let's consider another parametric setting to provide additional intuition into how we might try to think about $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. Take the 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 $\mathrm{C} \in \mathbb{R}$. Again $X = \{\emptyset\}$. S_0 can be thought of as 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 $R^2_{W\sim D|S=1}$ in Equation 6.[9] See Appendix A Section A.3 for the complete derivation.[10]

Figure 2: Internal selection graph for truncated multivariate normal example



$$R^2_{D\sim W|S=1} = \left(\frac{\rho_{D\sim W} - \rho_{S\sim D}\rho_{S\sim W}\theta}{\sqrt{1-\rho^2_{S\sim D}\theta}\sqrt{1-\rho^2_{S\sim W}\theta}}\right)^2, \text{ where } \theta \text{ can be written as a function of } P(S=1) \text{ or C} \tag{6}$$

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. Again, these quantities capture structural (i.e., causal) relationships between the variables in the population.

[7] If one or both of the sensitivity parameters does not contain a relationship altered by collider stratification, then the parameters for the selected sample will be the same as those for the population. The exception is when sample selection blocks a causal path that operates in the population but not in the selected sample. See Rohde and Hazlett (20XX) for more discussion.

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

[9] In Equation 6, if $\rho_{W\sim D} = 0$, then $R^2_{D\sim W|S=1} = \frac{R^2_{S\sim D}R^2_{S\sim W}\theta^2}{\sqrt{1-R^2_{S\sim D}\theta}\sqrt{1-R^2_{S\sim W}\theta}}$.

[10] Heckman (1979) relies on truncated normal variables and having some data on the full population; connections to that work are not explored here.

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 7, based on the partial correlation formula.[11] This does not equal $R^2_{W\sim D|S=1}$ in general. Equations 6 and 7 are remarkably similar, however, with their only differences being the need to account for where truncation happens (or the probability of selection). Equation 7 holds for linear conditioning on S, without any restrictions on the distribution or relationships between W, D, and S. Equation 6 only holds for truncated normals.

$$R^2_{D\sim W|S} = \left(\frac{\rho_{D\sim W} - \rho_{S\sim D}\rho_{S\sim W}}{\sqrt{1-\rho^2_{S\sim D}}\sqrt{1-\rho^2_{S\sim W}}} \right)^2 \tag{7}$$

We might wonder under what circumstances we would be able to use Equation 7 to inform our discussion of $R^2_{W\sim D|S=1}$. 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 $R^2_{W\sim D|S=1}$. In Appendix A Section A.5, we discuss a simple bound on $R^2_{W\sim D|S=1}$ 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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ (as well as $R^2_{W\sim D|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$) 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 5 and 6 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 a combination of first principles, previous studies and existing literature, intuition, and subject matter expertise to understand the range of plausible strengths of these relationships.

3 Proposal

Since one of the sensitivity parameters in the omitted variable bias expression 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 more easily. We will bound $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ by appealing to mutual information.

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 of two random variables, A and B, is to the product of their 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. As we will discuss further below, it turns out that mutual information has a number of useful properties for measuring dependence that are not present in R^2s and η^2s. Mutual information between A and B, $\mathrm{MI}(A;B)$, can be thought of as the information obtained (or reduction in uncertainty) about variable A that results from learning the value of variable 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 useful notions of conditional mutual information and joint mutual information and entropy. See Ihara (1993); MacKay (2003); Cover and Thomas (2006) for details.[12] Mutual information measures the amount of Shannon information revealed about A as a result of knowing B. Shannon information (or surprisal) 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

[11] In Equation 7, if $\rho_{W\sim D} = 0$, then $R^2_{D\sim W|S} = \frac{R^2_{S\sim D}R^2_{S\sim W}}{\sqrt{1-R^2_{S\sim D}}\sqrt{1-R^2_{S\sim W}}}$. 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.

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

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) While mutual information can be an improvement as a measure of dependence over R^2 or η^2, in practice, interpreting mutual information can be difficult. Therefore, we appeal to a normalized version that has nice properties discussed below.

Mutual information for Gaussians In order to connect $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ 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)

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

This relationship do not hold for arbitrary random variables, but 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); Lu (2011); Speed (2011); Kinney and Atwal (2014); Asoodeh et al. (2015); Smith (2015); Shevlyakov and Vasilevskiy (2017); Laarne et al. (2021), among others. Lu (2011) presents such a measure of dependence that is defined for arbitrary variables and that has many nice properties. We employ a slight variation on Lu (2011)'s L-measure, to create a useful normalized version of mutual information for our purposes. The L-measure takes the form $L(\text{MI}) = 1 - \exp(-2 \times \text{IF} \times \text{MI})$, where IF is an "inflation factor" that ensures that the L-measure takes appropriate values for arbitrary variables, not just continuous variables. See Appendix A Section A.6 for details.

Bounds This normalization of mutual information and Theorem 1 allow us to build interpretable bounds on $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ without any assumptions on the distrubutions or relationships between the variables. 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 $R^2_{D\sim W|S=1}$, $\eta^2_{D\sim W|S=1}$, $R^2_{D\sim W|X,S=1}$, and $\eta^2_{D\sim W|X,S=1}$ that rely on Theorem 1. These bounds can be found in Theorems 2 and 3. These results are proved and NSMI is defined in detail 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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

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

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 $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ resulting after stratification to $S = 1$ can be bounded in the following ways:*

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

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

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

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

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 $\eta^2_{D\sim X|S=1}$.[13] *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. We show in Appendix A Section A.6 that, for two random variables (X, Y), *$NSMI(X;Y)$ can be thought of as a measure of the **proportion** of the **certainty** in the outcomes of X, after we learn the value of Y, that is **gained** as a result of learning the value of Y.* (As opposed to the proportion of the certainty in the outcomes of X, after we learn the value of Y, that existed before we learned the value of Y.) Of course, this connects to the typical interpretation of mutual information as the "amount of information" obtained about X as a result of learning the value of Y. NSMI can indeed be interpreted just as a normalized and scaled version of mutual information; but it also has this additional interpretation that is somewhat similar to thinking of R^2 as the proportion of variance in one variable explained by another variable.

NSMI and the L-measure it is based on are useful measures of dependence between random variables in that they satisfy the properties discussed in Rényi (1959), Smith (2015), Lu (2011), and others as the properties possessed by "an appropriate measure of dependence."[14][15]

1. NSMI is defined for arbitrary pairs of random variables.[16]
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 have 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.[17]

(Laarne et al., 2021) also discusses a very similar transformation of mutual information and notes: "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 $(f(X), g(Y))$.[18]

We also provide 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;

[13] 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 $\text{MI}(D;[W,X]|S=1)$. See Appendix for discussion of $\mathbf{\Omega}$ and IF.

[14] 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.

[15] 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.

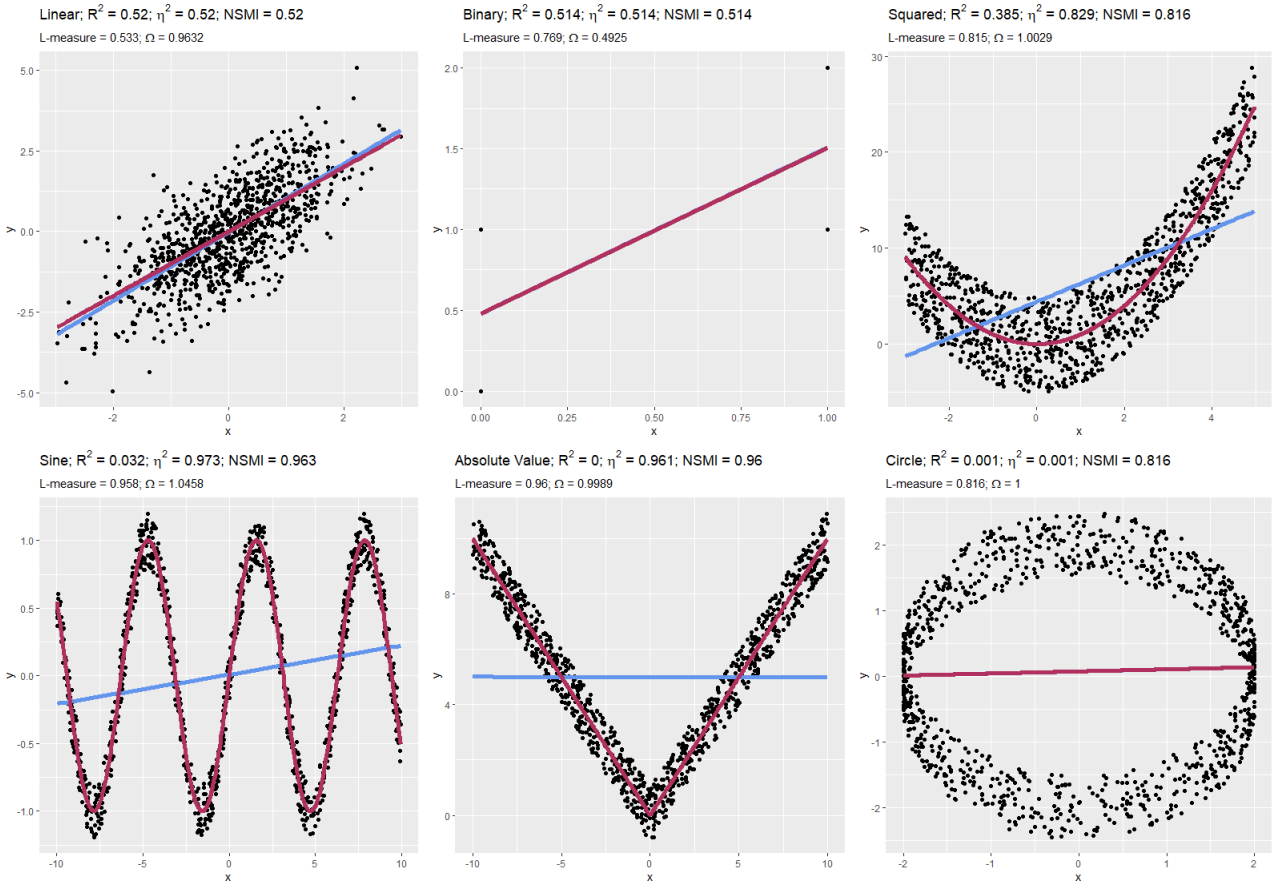
[16] When there are multiple unobserved variables contained in W, we can consider them in combination and consider their NSMI. That is, let $W = \{W_1, W_2, \ldots, W_k\}$ and consider NSMI like $\text{NSMI}(S;[D,W]) = \text{NSMI}(S;[D,W_1,W_2,\ldots,W_k])$ or $\text{NSMI}(D;W) = \text{NSMI}(D;[W_1,W_2,\ldots,W_k])$.

[17] Based on how we've defined NSMI, we also have the property that $\text{NSMI}(D;\mathbb{E}[D|W,S=1]|S=1) = R^2_{D,\mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D\sim W|S=1}$.

[18] 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 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.

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.[19] 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.

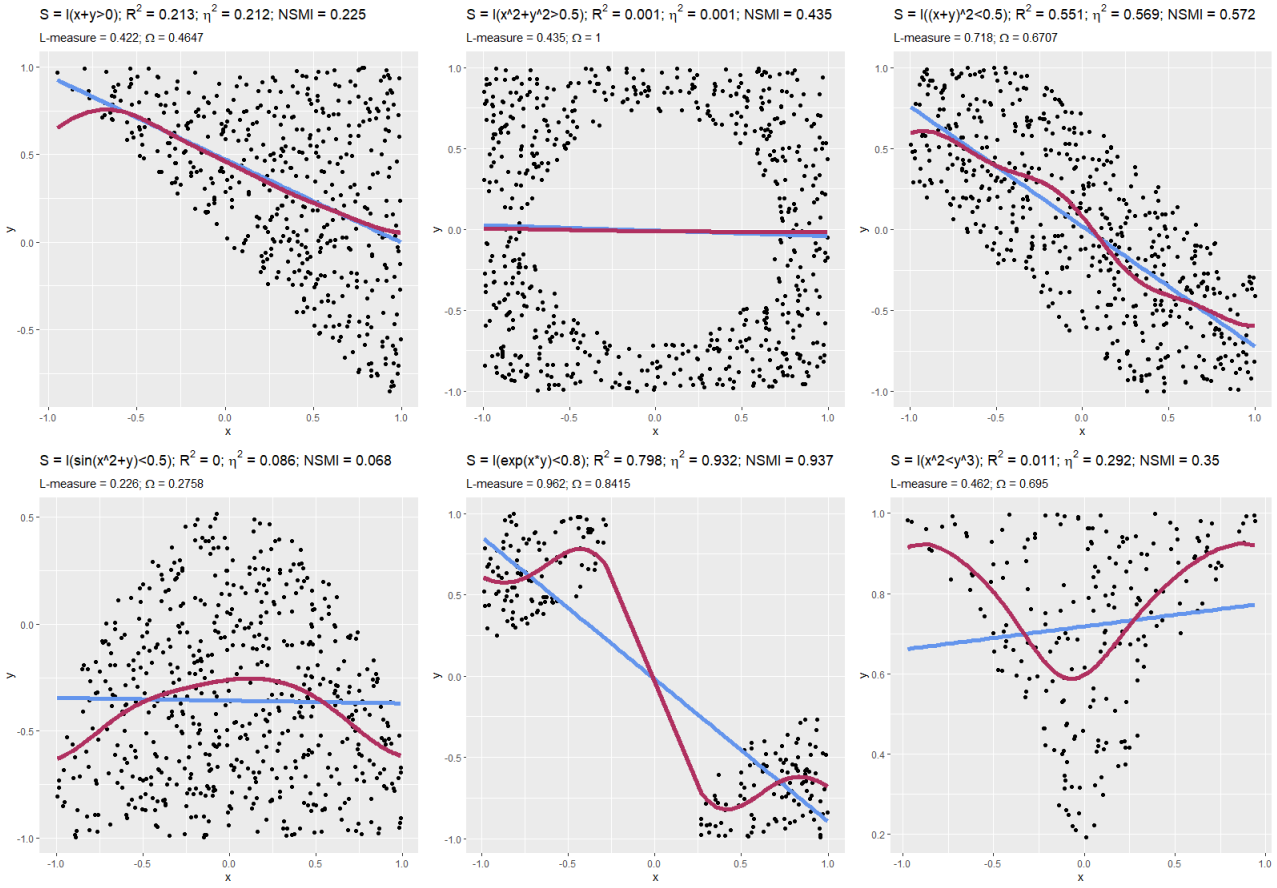


Discussion of bounds Theorems 2 and 3 contain several bounds. All the bounds presented in Theorem 3 correspond to bound 1 from Theorem 2. Analogs to bounds 2-6 from Theorem 2 can also be created for the case where there are covariates X. We expect that the simplest bound to use will often be bound 6 from Theorem 2. See the worked example below for an example of how bound 6 from Theorem 2 can be adapted to include covariates.

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 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.

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

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.



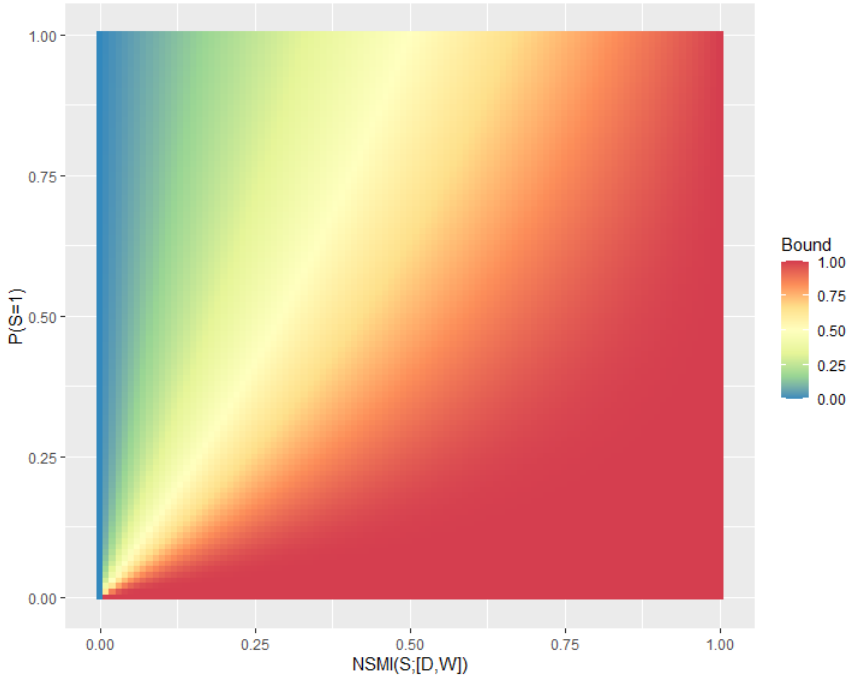
We consider in detail bound 6 from Theorem 2. This bound is an expression of normalized scaled mutual information for the marginal mutual information between D and W, for the mutual information between S and $[D, W]$ together, and the probability of selection, $P(S = 1)$.[20] As we saw in the case of 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 how 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 $S = 0$.

4 Worked example

We now turn our attention to a real application and demonstrate 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 article argues that "violence was targeted

[20] $P(S = 1)$ 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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ given values for NSMI$(S;[D,W])$ and $p(S=1)$ and assuming NSMI$(D;W)=0$



by village and gender but was indiscriminate beyond this" and that the "evidence is consistent not with the 'angry' response but rather with claims of a 'pro-peace' or 'weary' effect of exposure to violence." The author describes that "Most refugees or internally displaced persons left their homes during 2003 to 2004. A large number of those in the Western regions of West Darfur made the decision to cross the border into eastern Chad. Very few of these refugees had returned home by the time of this survey in mid-2009, when approximately 250,000 Darfurians were registered in refugee camps in eastern Chad." The study relies on data "drawn from a survey conducted between April and June of 2009 by the 'Darfurian Voices' team with support of the US Department of State... The full survey was thus representative of adult refugees (eighteen years or older) from Darfur, living in the twelve Darfurian refugee camps in eastern Chad at the time of sampling."

The study controls for things like village, gender, and other important covariates in estimating the causal effect of being directly harmed on attitudes about peace. 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 create a 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). Using OLS regression (among other estimation strategies), Hazlett (2020) estimates that peace index is .09 to .10 units higher among those directly harmed. See Table 1.

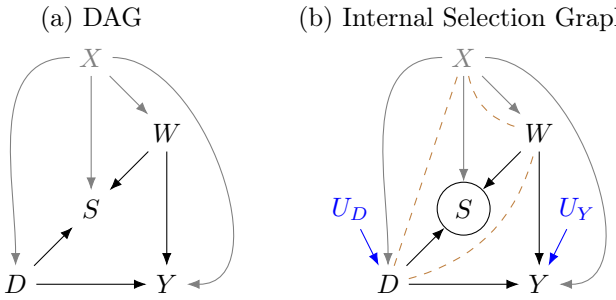
Table 1: Causal effect estimate from Hazlett (2020)

Outcome: *peace factor*

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

The process that would drive individuals back into the conflict would "act more powerfully for men of fighting age because

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



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

We can use the bounds from Theorems 2 and 3 in a sensitivity analysis for linear regression following 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 $R^2_{W\sim D|X,S=1}$ and $R^2_{Y\sim W|D,X,S=1}$. Inspecting Figure 6, we see that $R^2_{Y\sim W|D,X,S=1}$ 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, reasoning about $R^2_{W\sim D|X,S=1}$ 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 $R^2_{W\sim D|X,S=1}$. 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 $R^2_{D\sim W|X,S=1}$, we use bound 2 from Theorem 3 to bound $R^2_{D\sim W|X,S=1}$, 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

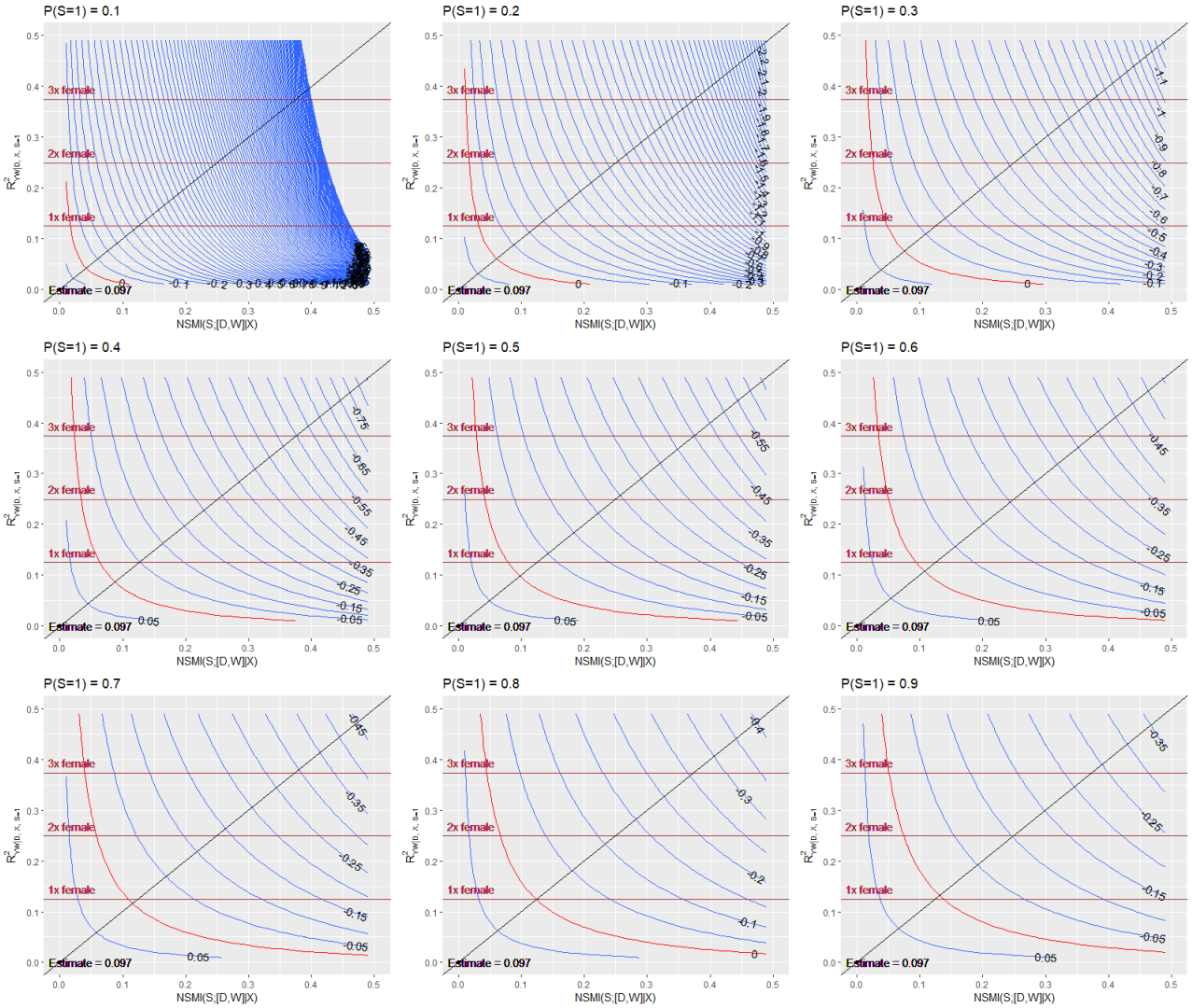
$$R^2_{W\sim D|X,S=1} \leq \frac{1}{1-R^2_{D\sim X|S=1}} \times \left(1 - \left[1 - \text{NSMI}(D;X|S=1)\right]\left[1 - \text{NSMI}(S;[D,W]|X)\right]^{\frac{1}{p(S=1)}} - R^2_{D\sim X|S=1}\right),$$

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 $R^2_{W\sim D|X,S=1}$ equals this bound and substitute the bound 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. In reality it is likely far less than 20%. 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, including pro-peace predisposition, will explain more of the outcome than 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 $R^2_{Y\sim W|D,X,S=1}$ equals the partial R^2 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? $\text{NSMI}(S,[D,W]|X)$ can be thought of as the proportion of the certainty inherent to whether someone re-enters the fight, given that we know whether they were directly harmed and their pro-peace pre-disposition (as well as their gender, village, and other observed covariates), that is gained as

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



a result of learning whether they were directly harmed and their pro-peace pre-disposition. 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. Overall, the relationship (i.e., shared or mutual information) between direct harm and reentering the fight largely stems from the common cause of gender. Conditional on gender, direct harm and reentering the fight likely share dramatically less information. However, there may still be some weak direct effect of direct harm on reentering the fight. This could result from some harmed individuals being more apprehensive than unharmed individuals to return. It could also result from from some harmed individuals being more vengeful. The latter is likely not a strong effect and the former is also likely weak, considering relatively few refugees returned overall. Similarly, the relationship (i.e., shared or mutual information) between pro-peace predisposition and reentering the fight largely stems from the common causes of gender and age. Conditional on gender and age, there still could be a direct effect of pro-peace predisposition on reentering the fight. This might result from individuals with less peaceful pre-conflict attitudes being more likely to reenter the fight. However, pre-conflict peace attitudes and direct harm, together, are likely much less important determinants of who reentered the fighting than are familial ties (e.g., having family members in conflict zone or in the refugee camps), concerns about saftey, gender, and age. That is, the sample selection mechanism is likely primarily driven by factors other than pre-conflict attitudes and direct harm. A complex decision like this likely is the result of numerous social and personal factors, as is any social phenomenon. So we should expect that there is a relatively weak dependency between these three as captured by 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 NSMI$(S, [D, W]|X)$ of up to perhaps 0.25. Perhaps reentering the fight would not have such a strong dependence on pro-peace predisposition and direct harm in essentially any plausible setting. In this case, we might be able to conclude that our effect estimate would not change sign of our estimate, despite this level of sample selection bias. Additionally, the true portion of refugees that re-entered the fight is likely much less that 20%. So even larger NSMI$(S, [D, W]|X)$'s are likely tolerable.

This type of analysis shifts criticism away from whether or not there exists a threat of sample selection bias towards substantive discussions like the previous paragraph that attempt to discern whether some substantive structural relationships meet a threshold level of strength that would change the conclusions of the estimated effect. We hope that this example provides some guidance to how such sensitivity analysis can be conducted in practice.

5 Discussion

Other approaches to sensitivity analysis for sample selection have been proposed. Smith and VanderWeele (2019) discuss approaches with sensitivity parameters that capture the relationships between unobserved variables and the observed variables, as we do here. However, building intuition for their parameters may not be as familiar as using R^2s or η^2s. Moreover, in their discussion of "the selected population as the target population," they only provide heuristic guidance on how researchers might deal with the counterintuitive nature of sensitivity parameters capturing associations between marginally independent variables that are made dependent due to sample selection. Thompson and Arah (2014) also present an approach for sensitivity analysis for sample selection. This approach requires specifying sensitivity parameters that filter into a model of the selection mechanism. Greenland (2003); Hernán et al. (2004); Elwert and Winship (2014); Infante-Rivard and Cusson (2018); Arah (2019), and many others also provide insightful discussions into sample selection bias and potential remedies. The benefit of our approach is the ease of interpretation of the sensitivity parameters and connections to the very useful, existing omitted variable bias based sensitivity analysis frameworks of Cinelli and Hazlett (2020) and Chernozhukov et al. (2022).

References

Arah, O. A. (2019). Analyzing selection bias for credible causal inference. *Epidemiology*, 30(4):517–520.

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.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40(1):31–53. PMID: 30111904.

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.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

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.

Hernán, M., Hernández-Díaz, S., and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, 15(5):615–625.

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

Infante-Rivard, C. and Cusson, A. (2018). Reflection on modern methods: selection bias—a review of recent developments. *International Journal of Epidemiology*, 47(5):1714–1722.

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, L. H. and VanderWeele, T. J. (2019). Bounding bias due to selection. *Epidemiology*, 30(4):509–516.

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".

Thompson, C. A. and Arah, O. A. (2014). Selection bias modeling using observed data augmented with imputed record-level probabilities. *Annals of Epidemiology*, 24(10):747–753.

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.

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

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 $R^2_{D\sim W|X,S=1}$ for binary random variables

Nguyen et al. (2019) provide the expression in Equation 8 for $\text{Cov}(W, D|S=1)$ for binary variables W, D, S in their Lemma 1.

$$\text{Cov}(W, D|S=1) = \frac{1}{P(S=1)^2}\begin{bmatrix} P(W=1,D=1,S=1)P(W=0,D=0,S=1)- \\ P(W=1,D=0,S=1)P(W=0,D=1,S=1) \end{bmatrix} \tag{8}$$

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 9, where $P_{W=w} = P(W=w)$, $P_{D=d} = P(D=d)$, $P_{S=1} = P(S=1)$, and $P_{S=1|wd} = P(S=1|W=w, D=d)$.

$$\text{Cov}(W, D|S=1) = \frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{P^2_{S=1}}[P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}] \tag{9}$$

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

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

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 10, since $P_{W=0} = 1 - P_{W=1}$ and $P_{D=0} = 1 - P_{D=1}$.

$$R^2_{W\sim D|S=1} = [P_{S=1|11}P_{S=1|00} - P_{S=1|10}P_{S=1|01}]^2\frac{P_{W=1}P_{D=1}P_{W=0}P_{D=0}}{\left(\begin{array}{c}[P_{S=1|11}P_{D=1} + P_{S=1|10}P_{D=0}][P_{S=1|01}P_{D=1} + P_{S=1|00}P_{D=0}]\times \\ [P_{S=1|11}P_{W=1} + P_{S=1|01}P_{W=0}][P_{S=1|10}P_{W=1} + P_{S=1|00}P_{W=0}]\end{array}\right)} \tag{10}$$

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 $R^2_{D\sim W|X,S=1}$ for truncated multivariate normal random variables

To provide some intuition into how we might try to think about $R^2_{D\sim W|X,S=1}$, 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 $S = \mathbf{1}[S_0 \geq \text{C}]$ for some $\text{C} \in \mathbb{R}$. 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 $K_{11} = \sigma^2_{S_0}\left[1 + \frac{\text{C}\phi(\text{C})}{1-\Phi(\text{C})} - \left(\frac{\phi(\text{C})}{1-\Phi(\text{C})}\right)^2\right] = \sigma^2_{S_0}\left[1 + \text{C}\gamma - \gamma^2\right]$, letting $\gamma = \frac{\phi(\text{C})}{1-\Phi(\text{C})}$, which is the inverse Mills ratio. (Kotz et al., 2000; Manjunath and Wilhelm, 2021) $S = \mathbf{1}[S_0 \geq \text{C}] \iff P(S=1) = P(S_0 \geq \text{C}) = P(S_0 \leq -\text{C}) = \Phi(-\text{C}) \iff \text{C} = -\Phi^{-1}(P(S=1))$ (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.

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

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 $R^2_{WD|S=1}$ 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.

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

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 $R^2_{W\sim D|S=1} = \frac{R^2_{S_0\sim D}R^2_{S_0\sim W}\delta}{\sqrt{1 - R^2_{S_0\sim D}\delta}\sqrt{1 - R^2_{S_0\sim W}\delta}}$.

Expression in terms of relationships with S, not S_0 We now explore how we can express $R^2_{WD|S=1}$ 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

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

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

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

The last equality uses $S_0 \sim \mathcal{N}(0,1)$ We can do the same thing for ρ_{WL}. So we have that $\rho_{S_0 D} = \rho_{SD}\xi$ and $\rho_{S_0 W} = \rho_{SW}\xi$, 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 $R^2_{W\sim D|S=1} = \frac{R^2_{S\sim D}R^2_{S\sim W}\xi^2\delta}{\sqrt{1-R^2_{S\sim D}\xi^2\delta}\sqrt{1-R^2_{S\sim W}\xi^2\delta}}$. 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 $R^2_{W\sim D|S=1}$ equals $R^2_{W\sim D|S=0}$) as a way of simplifying our analysis of $R^2_{W\sim D|S=1}$. What assumptions might make sense? What expression would this provide for $R^2_{W\sim D|S=1}$? 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 $\mathrm{Cor}(W, D|S) = [p(S=1) + (1-p(S=1))]\,\mathrm{Cor}(W, D|S=1) = \mathrm{Cor}(W, D|S=1) \implies R^2_{W\sim D|S=1} = R^2_{W\sim D|S}$. We can then leverage the partial correlation formula to arrive at

$$R^2_{W\sim D|S=1} = R^2_{W\sim D|S} = \left(\frac{R_{W\sim D} - R_{S\sim W}R_{S\sim D}}{\sqrt{1-R^2_{S\sim W}}\sqrt{1-R^2_{S\sim D}}}\right)^2$$

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

$$R^2_{W\sim D|S} = \text{Cor}^2(W,D|S) = \left[\sqrt{(A)(B)}\text{Cor}(W,D|S=1) + \sqrt{(1-A)(1-B)}\text{Cor}(W,D|S=0)\right]^2$$
$$= \underbrace{(A)(B)R^2_{W\sim D|S=1}}_{\geq 0} + \underbrace{(1-A)(1-B)R^2_{W\sim D|S=0}}_{\geq 0} + 2\sqrt{A(1-A)B(1-B)}R_{W\sim D|S=1}R_{W\sim D|S=0}$$

$$\implies R^2_{W\sim D|S=1} \leq \min\left(\frac{1}{AB}\left[R^2_{W\sim D|S} - 2\sqrt{A(1-A)B(1-B)}R_{W\sim D|S=1}R_{W\sim D|S=0}\right], 1\right)$$

We see that $R_{W\sim D|S=1}R_{W\sim D|S=0}$ is minimized when $R_{W\sim D|S=1}R_{W\sim D|S=0} = -1$.

$$\leq \min\left(\frac{1}{AB}\left[R^2_{W\sim D|S} + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)$$

Note that $2\sqrt{A(1-A)B(1-B)}$ is maximized at $\frac{1}{2}$ when $A = B = \frac{1}{2}$.

$$= \min\left(\frac{1}{AB}\left[\left(\frac{R_{\sim WD} - R_{S\sim W}R_{S\sim D}}{\sqrt{1-R^2_{S\sim W}}\sqrt{1-R^2_{S\sim D}}}\right)^2 + 2\sqrt{A(1-A)B(1-B)}\right], 1\right)$$

We can show that

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

This means that

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

So we get a bound on $R^2_{W\sim D|S=1}$:

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

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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ We start by noting that $R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1}$. This is easy to see since $\eta^2_{D\sim W|S=1} = R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} = \sup_f \left[\text{Cor}^2(D, f(W)|S=1)\right]$. (Doksum and Samarov, 1995; Chernozhukov et al., 2022) $\eta^2_{D\sim W|S=1}$ measures portion of the variation in D that can be explained by $\mathbb{E}[D|W,S=1]$, the conditional expectation function (CEF).[21]

[21] 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)

Correlation and mutual information for Gaussians In order to connect $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ 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 11. (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?

$$\text{MI}(W;D) = -\frac{1}{2}\log(1-R^2_{W\sim D}) \iff R^2_{W\sim D} = 1-\exp(-2\times \text{MI}(W;D)) \tag{11}$$

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 11. 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 12.

$$L^2(X,Y) \triangleq 1-\exp\left(-2\times \text{IF}\times \text{MI}(X;Y)\right), \text{ where IF} = \left(\frac{1}{1-(\text{MI}(X;Y)/A)}\right) \text{ and } A = \sup_{U,V\in\mathcal{A}_{X,Y}} \text{MI}(U;V)^{22} \tag{12}$$

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 IF $= \left(\frac{1}{1-(\text{MI}(X;Y)/\min[H(X),H(Y)])}\right) \geq 1$ since $\frac{\text{MI}(X;Y)}{\min[H(X),H(Y)]} \in [0,1]$. $\text{MI}(X;Y) \leq \min[H(X), H(Y)]$ since $H(X), H(Y)$ are the information content of X, Y. The idea is to inflate mutual information so that IF $\times$ MI$(X;Y) \to +\infty$ as X, Y become more dependent. The relationship is a strict dependence when $\text{MI}(X;Y) = \min[H(X), H(Y)]$. So A gives us the right level of inflation.
- $A = H(Y)$, when Y is discrete and X is continuous. This implies that IF $= \left(\frac{1}{1-(\text{MI}(X;Y)/H(Y))}\right) \geq 1$. Similar ideas apply here as in the last bullet.
- $A = 1$, when X, Y are both continuous which implies that IF $= 1$, since $\text{MI}(X;Y) = +\infty$ 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 13 hold?" That is, when does $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ equal $1-\exp\left(-2\times \text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right)$? 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 $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ and $1-\exp\left(-2\times \text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right)$ to provide similar portraits of the dependency between D and $\mathbb{E}[D|W,S=1]$.

$$\eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} \stackrel{?}{\approx} 1-\exp\left(-2\times \text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \tag{13}$$

$$\eta^2_{D\sim W|S=1} = R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1} = 1-\exp\left(-2\times \boldsymbol{\Omega}\times \text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \tag{14}$$

The question becomes whether we can alter the squared L-measure for $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)$ to exactly recover $\eta^2_{D\sim W|S=1}$. We do this by introducing an additional mutual information scaling factor $\boldsymbol{\Omega} \triangleq \frac{-\frac{1}{2}\log(1-R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1})}{\text{IF}\times\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)} \geq 0$. See Equation 14. This additional scaling factor, $\boldsymbol{\Omega}$, removes any discrepancy between the way that $R^2_{D\sim\mathbb{E}[D|W,S=1]|S=1}$ 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 $\text{MI}(D;\mathbb{E}[D|W,S=1]|S=1) \leq \text{MI}(W;D|S=1)$, since $\mathbb{E}[D|W,S=1]$ is a function of W.[23] (Cover and Thomas, 2006)

[22] 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 $\text{MI}(U;V)$ represent the mutual information between the random variables U and V.

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

It is also easy to see that $L^2_{\mathbf{\Omega}}(a) \triangleq 1 - \exp(-2 \times \mathbf{\Omega} \times \text{IF} \times a) \in [0,1]$ is a monotonic increasing function of $a \in [0, +\infty)$,[24] which means that $L^2_{\mathbf{\Omega}}(\text{MI}(D; \mathbb{E}[D|W, S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\text{MI}(W; D|S=1))$. Thus, we have the relationship in Equation 15.

$$R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} = R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} = L^2_{\mathbf{\Omega}}(\text{MI}(D; \mathbb{E}[D|W, S=1]|S=1)) \leq L^2_{\mathbf{\Omega}}(\text{MI}(W; D|S=1)) \tag{15}$$

What can we say about $\mathbf{\Omega}$? Including $\mathbf{\Omega}$ in $L^2_{\mathbf{\Omega}}(\text{MI}(D; \mathbb{E}[D|W, S=1]|S=1))$ essentially cancels out $\text{IF} \times \text{MI}(D; \mathbb{E}[D|W, S=1]|S=1)$ and undoes the L-measure transformation to simply return $R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1}$. 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 $\eta^2_{D\sim W|S=1} = R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} \leq L^2_{\mathbf{\Omega}}(\text{MI}(W; D|S=1))$. As we discuss in the next paragraph, we will reason about quantities like $L^2_{\mathbf{\Omega}}(\text{MI})$ 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 $R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1}$ 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 $R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1}$, 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 $\text{MI}(D; \mathbb{E}[D|W, S=1]|S=1)$ 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 $\eta^2_{D\sim W|S=1}$ and $\text{MI}(D; W|S=1)$. We will refer to any mutual information quantity scaled by $\mathbf{\Omega} \times \text{IF}$ 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."[25][26]

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.[27]
8. $\text{NSMI}(D; \mathbb{E}[D|W, S=1]|S=1) = R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} = \eta^2_{D\sim W|S=1}$.

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 15 implies Equation 16.

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

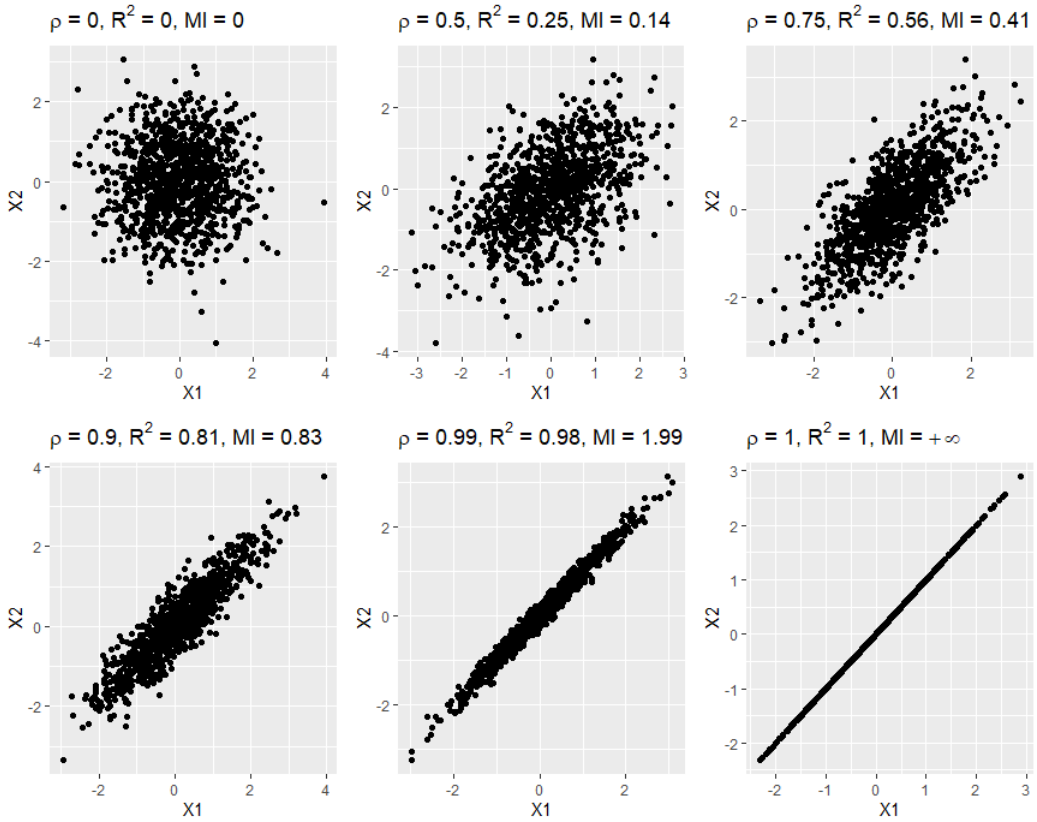
[25] Mutual information satisfies properties 1, 2, 4, and 6. Squared Pearson correlation (i.e., R^2) satisfies properties 1, 2, 3, 5, and 7.

[26] 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.

[27] 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.

$$R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq \text{NSMI}(W; D|S=1) \tag{16}$$

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.

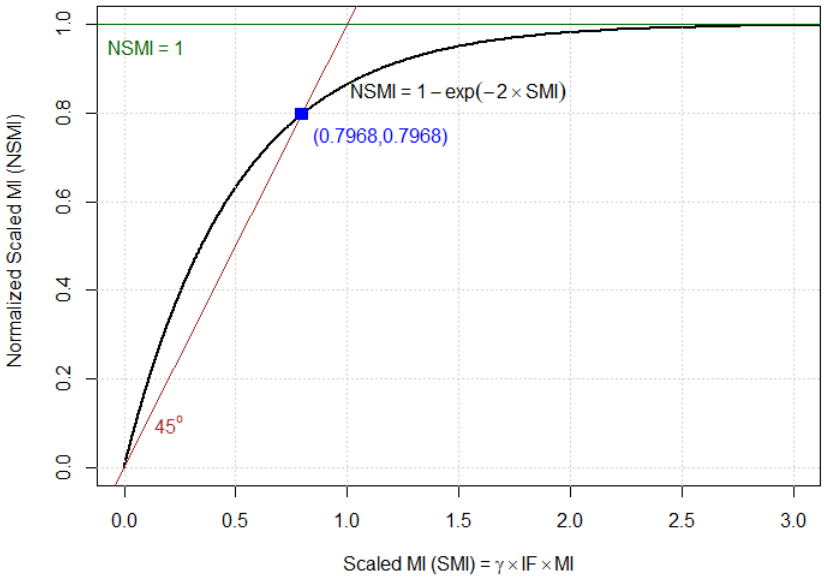


Interpreting NSMI For discrete random variables, X and Y, entropy and conditional entropy, $H(X)$ and $H(X|Y)$, are both positive. Recall that entropy can be thought of as a measure of the uncertainty or surprise in a random variable's outcomes. Further, $\text{MI}(X;Y) = H(X) - H(X|Y)$. It is easy to see that Equation 17 holds.

$$\begin{aligned}
\text{NSMI}(X;Y) &= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times \text{MI}(X;Y)\right) \\
&= 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times [H(X) - H(X|Y)]\right) \\
&= 1 - \exp\left([-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X)] - [-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X|Y)]\right) \\
&= 1 - \frac{\exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X)\right)}{\exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X|Y)\right)} \\
&= 1 - \frac{1 - [1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X)\right)]}{1 - [1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X|Y)\right)]} \\
&= 1 - \frac{1 - \text{NSH}(X)}{1 - \text{NSH}(X|Y)} = 1 - \frac{\text{C}(X)}{\text{C}(X|Y)} = \frac{\text{C}(X|Y) - \text{C}(X)}{\text{C}(X|Y)} \\
&\text{where } \text{NSH}(X) \triangleq 1 - \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X)\right) \\
&\text{and } \text{C}(X) \triangleq 1 - \text{NSH}(X) = \exp\left(-2 \times \boldsymbol{\Omega} \times \text{IF} \times H(X)\right)
\end{aligned} \tag{17}$$

$\text{NSH}(X)$ is a normalized and scaled version of entropy that uses the same normalization and scaling as NSMI. This means that $\text{NSH}(X)$ takes values between zero and one and is a measure of the uncertainty in the outcomes of X, since it is a monotonic transformation of entropy. How might we think about the $1 - \text{NSH}(X)$ and $1 - \text{NSH}(X|Y)$ terms that appear in the expression for $\text{NSMI}(X;Y)$ in Equation 17? $1 - \text{NSH}(X)$ close to zero means that there is a large amount of uncertainty

Figure 9: Normalization of Scaled Mutual Information



in the outcomes of X. $1 - \text{NSH}(X)$ close to one means that there is very little uncertainty in the outcomes of X. As expected, $1 - \text{NSH}(X)$ captures something very similar to $\text{NSH}(X)$, but with the meaning of large and small values reversed. We might, therefore, call $\text{C}(X) \triangleq 1 - \text{NSH}(X)$ a measure of *lack of* surprise or *certainty*.

Thus, $\text{NSMI}(X;Y) = \frac{\text{C}(X|Y) - \text{C}(X)}{\text{C}(X|Y)}$, can be thought of as a measure of the **proportion** of the **certainty** in the outcomes of X, after we learn the value of Y, that is **gained** as a result of learning the value of Y. (As opposed to the proportion of the certainty in the outcomes of X, after we learn the value of Y, that existed before we learned the value of Y, which equals $\frac{\text{C}(X)}{\text{C}(X|Y)}$. Note that $\text{NSMI}(X;Y) + \frac{\text{C}(X)}{\text{C}(X|Y)} = 1$.)

Note that $\text{NSMI}(X;Y)$ is not a measure of the proportion of the *uncertainty* in X that is reduced by learning Y, which would be captured by $\frac{\text{NSH}(X) - \text{NSH}(X|Y)}{\text{NSH}(X)}$. But these two are closely related. Indeed, we can write $\text{NSMI}(X;Y) = \frac{\text{C}(X|Y) - \text{C}(X)}{\text{C}(X|Y)} = \frac{\text{NSH}(X) - \text{NSH}(X|Y)}{1 - \text{NSH}(X|Y)}$. We see that the two share a numerator. It is only the denominator that differs. Both measure the change in information we have about X but take this as a proportion of different quantities. Note that this intuition applies to both NSMI and the L-measure.

Mutual information bounds Equation 16 seems nice. But have we solved our original problem of finding a bound on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ 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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$. We start by considering $\text{MI}(D;W|S)$. Using properties of mutual information (Cover and Thomas, 2006), we can show Equation 18.

$$\begin{aligned}\text{MI}(D;W|S) &= \text{MI}(D;W) + \text{MI}(S;D|W) - \text{MI}(S;D) \\ &= \text{MI}(D;W) + [\text{MI}(S;[D,W]) - \text{MI}(S;W)] - \text{MI}(S;D) \\ &= \text{MI}(D;W) + \text{MI}(S;[D,W]) - \text{MI}(S;D) - \text{MI}(S;W)\end{aligned} \tag{18}$$

$\text{MI}(S;[D,W]) = \text{MI}(S;W) + \text{MI}(S;D|W)$ 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 19.

$$
\begin{aligned}
\mathrm{MI}(D;W|S) &= \int_{\mathcal{S}} D_{\mathrm{KL}}\left(P_{(D,W)|S} \| P_{D|S} \otimes P_{W|S}\right) dP_S \\
&= \sum_{s\in\{0,1\}} p(S=s) \sum_{d} \int_{w} p(d,w|S=s) \log\left[\frac{p(d,w|S=s)}{p(d|S=s)p(w|S=s)}\right] dd\ dw \\
&= \sum_{s\in\{0,1\}} p(S=s) D_{\mathrm{KL}}\left(P_{(D,W)|S=s} \| P_{D|S=s} \otimes P_{W|S=s}\right) \\
&= p(S=1) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=1} \| P_{D|S=1} \otimes P_{W|S=1}\right) + p(S=0) \times D_{\mathrm{KL}}\left(P_{(D,W)|S=0} \| P_{D|S=0} \otimes P_{W|S=0}\right) \\
&= p(S=1)\mathrm{MI}(D;W|S=1) + p(S=0)\mathrm{MI}(D;W|S=0)
\end{aligned}
\tag{19}
$$

From Equations 18 and 19, we have that

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

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 $MI(D;W|S=1) \leq \frac{MI(D;W|S)}{p(S=1)} = \frac{MI(D;W)+MI(S;[D,W])-MI(D;S)-MI(W;S)}{p(S=1)}$.*

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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ We now combine Equation 16 with Equation 20 to get interpretable bounds on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$. We start by considering only one such bound. But others are possible.

$$
\begin{aligned}
R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} &= R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1} \\
&= 1-\exp\left(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)\right) \text{ since } \mathbf{\Omega}=\frac{-\frac{1}{2}\log(1-R^2_{D\sim \mathbb{E}[D|W,S=1]|S=1})}{\text{IF}\times \text{MI}(D;\mathbb{E}[D|W,S=1]|S=1)} \\
&\leq 1-\exp\left(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(W;D|S=1)\right) = \text{NSMI}(W;D|S=1) \text{ by the data processing inequality} \\
&\leq 1-\exp\left(-2\times \mathbf{\Omega}\times \text{IF}\times \frac{\text{MI}(D;W)+\text{MI}(S;[D,W])}{p(S=1)}\right) \text{ by Eqn. 20} \\
&= 1-\exp\left(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(D;W)-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(S;[D,W])\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left(\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(D;W))\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(S;[D,W]))\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left([1-1+\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(D;W))][1-1+\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(S;D,W]))]\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left([1-(1-\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(D;W)))][1-(1-\exp(-2\times \mathbf{\Omega}\times \text{IF}\times \text{MI}(S;[D,W])))]\right)^{\frac{1}{p(S=1)}} \\
&= 1-\left(\ [1-\text{NSMI}(D;W)][1-\text{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}}
\end{aligned}
\tag{21}
$$

Therefore, our first interpretable bound is captured by Equation 22.

$$
R^2_{D\sim W|S=1} \leq \eta^2_{D\sim W|S=1} \leq 1-\left(\ [1-\text{NSMI}(D;W)][1-\text{NSMI}(S;[D,W])]\ \right)^{\frac{1}{p(S=1)}} \tag{22}
$$

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 22 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 the stratum $S=0$.

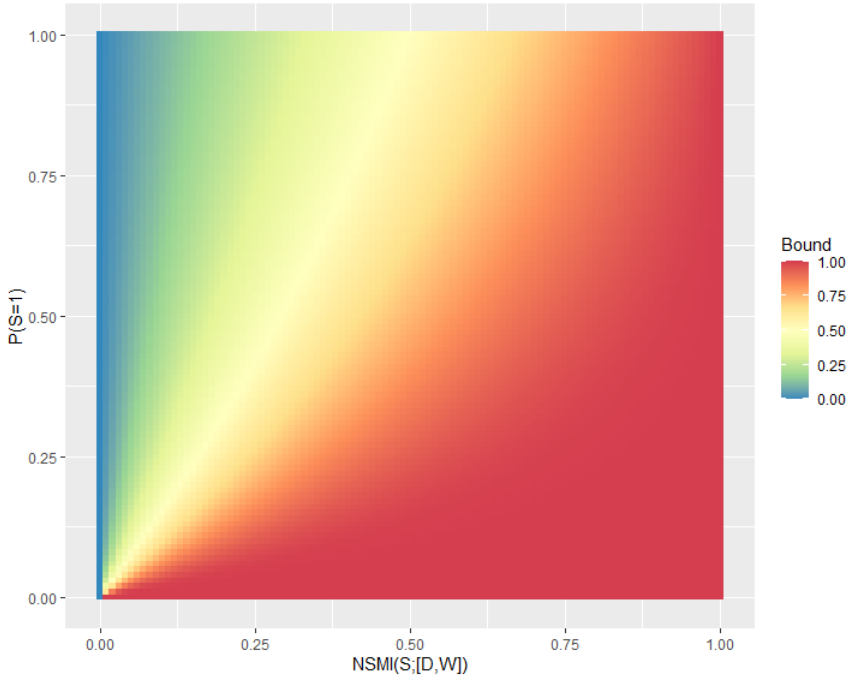
Following a similar approach as we did in obtaining the bound in Equation 22, we arrive at Theorem 2.

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 $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ resulting after stratification to $S=1$ can be bounded in the following ways:*

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

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 $\text{NSMI}(W;D|S)$ 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 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.

Figure 10: Bounds (from Equation 22) on $R^2_{D\sim W|S=1}$ and $\eta^2_{D\sim W|S=1}$ given values for NSMI$(S;[D,W])$ and $p(S=1)$ and assuming NSMI$(D;W)=0$



Incorporating Covariates It is also fairly straightforward to incorporate covariates, X. We now turn to bounding $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. The approach is very similar to the above. Equation 23 follows from the usual expressions of $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ and the fact that $R^2_{D\sim W,X|S=1} \leq \eta^2_{D\sim W,X|S=1}$.[28]

$$\begin{aligned} R^2_{D\sim W|X,S=1} &= \frac{R^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \leq \frac{\eta^2_{D\sim W,X|S=1} - R^2_{D\sim X|S=1}}{1 - R^2_{D\sim X|S=1}} \\ \eta^2_{D\sim W|X,S=1} &= \frac{\eta^2_{D\sim W,X|S=1} - \eta^2_{D\sim X|S=1}}{1 - \eta^2_{D\sim X|S=1}} \end{aligned} \tag{23}$$

We can estimate $R^2_{D\sim X|S=1}$ and $\eta^2_{D\sim X|S=1}$ in Equation 23 from the selected sample, since neither involves W. Since both portions of Equation 23 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 24. 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.

$$\eta^2_{D\sim W,X|S=1} = 1 - \exp(-2 \times \mathbf{\Omega} \times \text{IF} \times \text{MI}(D;[W,X]|S=1)) \text{ where } \mathbf{\Omega} = \frac{-\frac{1}{2}\log(1-\eta^2_{D\sim W,X|S=1})}{\text{IF} \times \text{MI}(D;[W,X]|S=1)} \tag{24}$$

From Equation 24, we have two options for how to proceed. First, we could use Theorem 1 with W replaced with $[W,X]$ to arrive at Equation 25.

$$\text{MI}(D;[W,X]|S=1) \leq \frac{\text{MI}(D;[W,X]|S)}{p(S=1)} = \frac{\text{MI}(D;[W,X]) + \text{MI}(S;[D,W,X]) - \text{MI}(D;S) - \text{MI}([W,X];S)}{p(S=1)} \tag{25}$$

Using Equations 24 and 25 we arrive at Equation 26.

[28] We are not able to directly link $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ 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 $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ relate. In this discussion, we simply bound them separately.

$$
\begin{aligned}
\eta^2_{D\sim W,X|S=1} &= 1-\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times \text{MI}(D;[W,X]|S=1))\\
&\leq 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\left[\frac{\text{MI}(D;[W,X])+\text{MI}(S;[D,W,X])-\text{MI}(D;S)-\text{MI}([W,X];S)}{p(S=1)}\right]\right)\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times[\text{MI}(D;[W,X])+\text{MI}(S;[D,W,X])-\text{MI}(D;S)-\text{MI}([W,X];S)]\right)^{\frac{1}{p(S=1)}}\\
&= 1-\left[\frac{\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;[W,X]))\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(S;[D,W,X]))}{\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;S))\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}([W,X];S))}\right]^{\frac{1}{p(S=1)}}\\
&= 1-\left[\frac{[1-\text{NSMI}(D;[W,X])][1-\text{NSMI}(S;[D,W,X])]}{[1-\text{NSMI}(D;S)][1-\text{NSMI}([W,X];S)]}\right]^{\frac{1}{p(S=1)}}
\end{aligned}
\tag{26}
$$

Second, we could use Theorem 1 with everything conditioned on X and the fact that $\text{MI}(D;W|X,S) = p(S=1)\text{MI}(D;W|X,S=1)+p(S=0)\text{MI}(D;W|X,S=0)$ to arrive at the second equation in Equation 27. The first equation in Equation 27 just comes from the definition of $\text{MI}(D;[W,X]|S=1)$.

$$
\begin{aligned}
&\text{MI}(D;[W,X]|S=1) = \text{MI}(D;X|S=1)+\text{MI}(D;W|X,S=1) \text{ and}\\
&\text{MI}(D;W|X,S=1) \leq \frac{\text{MI}(D;W|X,S)}{p(S=1)} = \frac{\text{MI}(D;W|X)+\text{MI}(S;[D,W]|X)-\text{MI}(D;S|X)-\text{MI}(W;S|X)}{p(S=1)}
\end{aligned}
\tag{27}
$$

Using Equations 24 and 27 we arrive at Equation 28.

$$
\begin{aligned}
\eta^2_{D\sim W,X|S=1} &= 1-\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times \text{MI}(D;[W,X]|S=1))\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times[\text{MI}(D;X|S=1)+\text{MI}(D;W|X,S=1)]\right)\\
&\leq 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\left[\text{MI}(D;X|S=1)+\frac{\text{MI}(D;W|X)+\text{MI}(S;[D,W]|X)-\text{MI}(D;S|X)-\text{MI}(W;S|X)}{p(S=1)}\right]\right)\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;X|S=1)\right)\\
&\times\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\frac{\text{MI}(D;W|X)+\text{MI}(S;[D,W]|X)-\text{MI}(D;S|X)-\text{MI}(W;S|X)}{p(S=1)}\right)\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;X|S=1)\right)\\
&\times\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times[\text{MI}(D;W|X)+\text{MI}(S;[D,W]|X)-\text{MI}(D;S|X)-\text{MI}(W;S|X)]\right)^{\frac{1}{p(S=1)}}\\
&= 1-\exp\left(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;X|S=1)\right)\\
&\times\left[\frac{\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;W|X))\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(S;[D,W]|X))}{\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(D;S|X))\exp(-2\times\boldsymbol{\Omega}\times \text{IF}\times\text{MI}(W;S|X))}\right]^{\frac{1}{p(S=1)}}\\
&= 1-[1-\text{NSMI}(D;X|S=1)]\left[\frac{[1-\text{NSMI}(D;W|X)][1-\text{NSMI}(S;[D,W]|X)]}{[1-\text{NSMI}(D;S|X)][1-\text{NSMI}(W;S|X)]}\right]^{\frac{1}{p(S=1)}}
\end{aligned}
\tag{28}
$$

Equations 23, 26, and 28 combine to provide the following bounds on $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$. 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 $R^2_{D\sim W|X,S=1}$ and $\eta^2_{D\sim W|X,S=1}$ resulting after stratification to $S=1$ can be bounded in the following ways:*

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

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* $\eta^2_{D\sim X|S=1}$.[29] *These bounds are all analogous to bound 1 in Theorem 2. Analogs to bounds 2 - 6 in Theorem 2 could also be formed.*

[29] 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$(D;[W,X]|S=1)$. See Appendix for discussion of $\mathbf{\Omega}$ and IF.