Contents

Confounders, Mediators, and Colliders: An Overview

To conduct and interpret the results of regression analysis in a causal manner, it is usually necessary[1] to have a theoretical model of the causal relationships between the variables. A purely data-driven analysis will not provide insights into causal relationships, no matter how sophisticated the analysis methods may be. However, even basic statistical analysis[2] can yield valuable information about causal relationships among variables, if we have an accompanying causal theory.

Suppose we have three variables: X, Y, and Z. Based on our theory, we believe that Y, the outcome variable, is potentially influenced by X, which we can express as X → Y. In our regression analysis, we will consider Y as the dependent variable and X as the critical independent variable. However, let's say we discover that the third variable, Z, is correlated with both X and Y. In this situation, our course of action depends on how our theory explains the role of Z. [3]

- Confounder: If Z is a confounder, it means that it has a causal impact on both X (Z → X) and Y (Z → Y). In this case, we need to include Z as a control variable in our regression analysis to account for its influence.
- Mediator: If Z is a mediator, it means that X has a causal impact on Z (X → Z) and Z has a causal impact on Y (Z → Y). Depending on our specific objectives, we can choose to include or exclude the mediator from our regression analysis. However, if we decide to include Z, we must be cautious when interpreting the causal impact of X, as part of the effect may be mediated through Z.
- Collider: If Z is a collider, it means that X has a causal impact on Z (X → Z) and Y also has a causal impact on Z (Y → Z). In this case, collider variables should always be excluded from our analysis as they can introduce biased results.

To summarize, when we encounter a situation where a third variable, Z, is correlated with both X and Y, we need to consider its role as a confounder, mediator, or collider. [4] The appropriate course of action depends on our theoretical

[1] When data comes from controlled experiments or situations that resemble such experiments (this occurs typically in lab sciences but also occasionally in social sciences), causal analysis might be possible without the need for a detailed theory.

[2] For example, the variance-covariance matrix for X, Y and Z can provide us the slope coefficients for all possible linear regressions involving those three variables (as long as we don't have interaction terms).

[3] In the following discussion, we will exclude the possibility of simultaneous reverse causality. This means that if X has a causal impact on Y, we assume that Y cannot have a causal impact on X, either directly or indirectly. To help conceptualize this, we can consider X, Y, and Z as implicitly measured at different points in time. As the future cannot influence the past or the present, any causal relationship must be one-directional, following the arrow of time. However, it's important to note a few minor caveats:

i. Forecasts of the future: Forecasts or predictions about the future can indeed influence the present. If relevant, these forecasts need to be included in the model to account for their impact on the variables.
ii. Time periods of measurement: In practice, many variables are measured over periods of time rather than at different points in time and therefore reverse causality can be a common feature of real-world data (e.g., GDP and Consumption are measured over a quarter and can cause each other).
iii. Quantum mechanics: In the field of quantum mechanics, there are intriguing phenomena where the future appears to affect the present. Recent advancements in this area were recognized with the 2022 Physics Nobel Prize. However, in our day-to-day world, causality typically runs from the past through the present to the future.

[4] There are three other possibilities.

i. Z may be causally related *only to* X (in either direction) but not causally related to Y. In this case, Z can be useful as an instrumental variable. An instrumental variable should be correlated with the independent variable (X) but not directly associated with the dependent variable (Y). Including Z as an instrumental variable can help address endogeneity issues.

understanding. We may need to include Z as a control variable if it is a confounder, be cautious about interpreting the causal impact of X if Z is a mediator, or exclude Z entirely if it acts as a collider.

I. Confounders

A confounder is a variable that influences both the outcome (dependent variable) and the independent variable of interest (critical independent variable). Failing to account for a confounder can lead to omitted variable bias (OVB) and incorrect causal interpretations of the results. We have extensively studied OVB in Eco 105 and prior lectures of Eco 205, including the derivation of relevant formulas. Thus, this discussion serves as a brief review emphasizing the significance of addressing confounders to avoid biased causal interpretations.

Constructed example for a confounder

Suppose we aim to examine the impact of following public health guidelines (P) on the severity of illness due to Covid (S) using the regression model:

$S = \beta_0 + \beta_1 P + \varepsilon$

However, the coefficient β_1 may be biased because of the presence of confounding variables. In this case, a confounder could be the existence of pre-existing health conditions or comorbidities (C). Individuals with comorbidities might be more likely to adhere to public health guidelines due to their increased vulnerability to Covid infections.

The bias can arise either on the supply side, where individuals with comorbidities receive preferential access to vaccines, or on the demand side, where those with comorbidities recognize their higher susceptibility and exhibit greater adherence to public health guidelines.

Given the specified regression equation, the coefficient β1 will not solely capture the effect of P on S but also the effect of C due to the correlation between comorbidities and the extent to which individuals follow public health guidelines. This situation can lead to Simpson's paradox, where the direction of the relationship observed in the entire sample gets reversed when examining subgroups.

To accurately estimate the effect of following public health guidelines (P) on the severity of illness (S), it is crucial to address the confounding effect of comorbidities (C) by incorporating appropriate control variables or employing advanced statistical techniques to account for the potential bias introduced by confounders.

In terms of a DAG (directed acyclic graph) the confounder can be depicted as follows:

ii. Z may be causally related *only to* Y (in either direction) but not causally related to X. If Z causes Y, it might be useful to include Z in the analysis to provide a better overall explanation of Y. However, if Y causes Z, then Z should be omitted from the analysis, similar to other collider variables, to prevent introducing biased results.

iii. In addition to having their own independent influences on Y, Z and X may also each influence *how* the other effects Y. In this case, an interaction term would be required to capture the non-zero cross-derivative.

We will construct our hypothetical example[5] using Excel (see spreadsheet, "Confounders") and work through possible regressions to examine how they match up to our constructed theoretical model:

Suppose you have a sample of 100 people where 30 people have a comorbidity.

- Create a column labeled 'C' where the first 30 values are 1 and the next 70 values are 0. Note, this assumption implies that the mean[6] of C (μ_C) = **0.3** and the variance[7] of C (σ_C^2) = **0.21**.

Suppose the relationship between following public health guidelines and comorbidities is given by:

P = **0.3** + **0.4** C + ε where ε is normally distributed with mean **0** and standard deviation of **0.3**

- Create a column of 100 values. Label it as 'epsilon raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.3**)
- Copy this column over and use "paste special as values" into a new column and label it as 'epsilon'. This step is for convenience because, by default, Excel recomputes all values each time you enter a new formula which causes all the random numbers to keep changing each time.
- Create a column labeled 'P' and fill it using the formula P = **0.3** + **0.4**C + ε
- The mean value of P[8] should approximately equal = $\mathbf{0.3} + \mathbf{0.4} \times \mu_C = \mathbf{0.3} + \mathbf{0.4} \times \mathbf{0.3}$ = 0.42
- The variance of P[9] should approximately equal $\mathbf{0.4^2} \times \sigma_C^2 + \sigma_\varepsilon^2 = \mathbf{0.4^2} \times \mathbf{0.21} + \mathbf{0.3^2}$ = 0.12
- The covariance of P and C[10] should approximately equal $\mathbf{0.4} \times \sigma_C^2 = \mathbf{0.4} \times \mathbf{0.21}$ = 0.084

Suppose the relationship between severity of illness, vaccination status, and comorbidities is given by:
S = **4** - **0.3** P + **1.5** C + u where u is normally distributed with mean **0** and variance **0.2**.

- Create a column of 100 values. Label it as 'u raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.2**)
- Again, copy this column over and use paste special as values into a new column and label it as 'u'
- Create a column labeled 'S' and fill it using the formula S = **4** - **0.3** P + **1.5** C + u

Now we will estimate the regression, first without controlling for the confounder and then in a version where we control for the confounder.

- To do this we need to first create three new contiguous columns copying over S, P, and C. Contiguous columns are required if you wish to use Excel for multiple regression.
- Estimate the regression S = β_0 + β_1 P + ε using the entire sample. What is the sign and statistical significance for the coefficient of P? What would be (an incorrect) naïve interpretation of this coefficient?

[5] C is assumed to be dummy variable (i.e., a person is classified as either having one or more comorbidities or not having any comorbidity). P is assumed to be a continuous variable that typically ranges from 0 (i.e., ignoring public health guidelines) to 1 (i.e., following guidelines perfectly) although values could be negative (e.g., if a person actively seeks to do the opposite of the guidelines) or greater than 1 (e.g., if a person goes even beyond the recommended guidelines in terms of avoiding exposure to the virus).
[6] For dummy variables, the mean equals the fraction of observations with a value of 1.
[7] For dummy variables, variance equals the product of the fraction of observations with a value of 1 and the fraction of variables with a value of 0.
[8] Formula for average for linear expressions.
[9] Formula for variance, assuming covariance of epsilon and C = 0.
[10] This comes from rearranging the formula for the slope coefficient in a linear regression with one variable, i.e.., 0.4 = covariance(P,C) / variance(C)

In the simulation, the value obtained is positive 0.9 suggesting that severity of infection is positively correlated with the extent to which public health guidelines are followed! Based on the p-value, this coefficient is statistically significant even at 0.01% level. A naïve interpretation would suggest that following public health guidelines is associated with (some naïve folks might even say 'causes') greater severity of illness.

- Now, estimate the regression $S = \beta_0 + \beta_1 P + \beta_2 C + \varepsilon$ using the entire sample. What is the new sign and value for β_1?

  Once the confounder is included, the coefficient changes to -0.3 and is statistically significant even at the 0.01% level. Based on the constructed model, this coefficient can be interpreted causally.
- Recall that the theoretical value for the omitted variable bias[11] is y **1.5** x 0.084/(0.12) = 1.05.
- In this example, Simpson's paradox is easiest to see graphically. Graph the full sample and plot the trendline and include the equation on the graph. On the same graph plot each subsample and plot their respective trendlines and include their respective equations.[12]

In our example, we know the true relationship among our variables. In our data generating process, **C was built in as a confounder**. If the goal of the study is to answer the question, "Does following public health guidelines help to reduce severity of illness?" then, from our regressions we can see how omitting the confounder leads to an incorrect conclusion, just as is predicted by statistical theory.

II. Mediators

A mediator is a variable that affects the outcome (dependent variable) and is also influenced by the independent variable of interest (critical independent variable). It acts as an intermediate factor through which the critical independent variable influences the dependent variable. Whether to include or exclude a mediator in a regression analysis depends on the research objective. However, it is important to carefully interpret the coefficient of the independent variable in both cases. Ignoring the role of a mediator while interpreting causality leads to overcontrol bias.

When the goal is to determine the overall causal impact of the critical independent variable, typically for policy purposes, two approaches can be taken:

1. Conduct a regression where the dependent variable is the outcome variable and include both the critical independent variable and the mediator as independent variables. The coefficient of the independent variable represents the partial derivative and should be interpreted accordingly. To obtain the total derivative, which reflects the overall causal impact, a second regression is required where the dependent variable is the mediator and the independent variable is the critical independent variable. By combining the coefficients from these two regressions using a formula, the total derivative, representing the overall causal impact, can be computed. This approach requires the estimation of two separate regressions to provide insights into the overall causal impact of the critical independent variable and the extent to which it operates through the mediator.

[11] Recall from class lectures that Bias = β_2 x Covariance(P,C)/Variance (P)

[12] For the model here, estimating each subsample separately (i.e., stratifying by the confounding dummy variable) is equivalent to including C as well as the interaction term P x C in the regression. Including such an interaction term (or stratifying, i.e., conducting the analysis for each subsample differently) would be a particularly good idea if the effect of following public health guidelines varies based on whether a person has comorbidities.

2. Alternatively, a simpler approach involves estimating a single regression without including the mediator. Although this method is convenient as it requires only one regression, it does not provide information about the extent to which the independent variable acts through the mediator.

Both approaches have their advantages and limitations. The first approach allows for a comprehensive assessment of the overall causal impact and the mediating effect, while the second approach is more straightforward but lacks insight into the mediation pathway. The choice between these approaches depends on the specific research objectives and the level of understanding required regarding the mediating mechanism.

Constructed example for a mediator

Suppose we want to investigate the impact of receiving prenatal healthcare (H) on birthweight (W), an important indicator of infant health at birth. Along with these variables, we also have data on the extent to which dietary and lifestyle guidelines (G) were followed during pregnancy. Initially, we include G in the regression equation as we suspect it could be a confounding variable:

$W = \beta_0 + \beta_1 H + \beta_2 G + \varepsilon$

However, it is worth considering that parents might be more inclined to learn and adhere to dietary and lifestyle guidelines if they receive prenatal healthcare. Drawing from my personal experiences as a parent, this is a plausible scenario. In such a case, G functions as a mediator rather than a confounder. Consequently, the coefficient β_1 would not accurately measure the true effect of receiving prenatal healthcare on infant health. Instead, β_1 represents only the partial derivative of W with respect to H. i.e., $\beta_1 = \frac{\partial W}{\partial H}$, while we may be more interested in the total derivative, $\frac{dW}{dH}$.

The limited interpretation of β_1 as a partial derivative can also be understood through the formal interpretation of the coefficient. We interpret β_1 as the association between W and H ***holding constant*** *all other included variables*. In other words, β_1 tells us how receiving prenatal health care changes birthweight under the assumption that G is unchanged.

However, for policy purposes, the total derivative for W with respect to H (i.e., $\frac{dW}{dH}$) is likely more relevant. Mathematically, this value equals $\frac{dW}{dH} = \beta_1 + \beta_2 \frac{\partial G}{\partial H}$

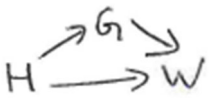
In other words, to find the impact of H on W we need to consider its direct impact ($= \beta_1$) as well as its indirect impact through G ($= \beta_2 \frac{\partial G}{\partial H}$).

To find $\frac{\partial G}{\partial H}$, we need to estimate a regression where G is the dependent variable and H is the key independent variable. While estimating this regression, however, we need to be sure that we are not omitting any confounders as that would give us a biased coefficient as discussed in the prior section.[13]

On the other hand, if we only care about $\frac{dW}{dH}$, it is adequate to directly estimate the regression $W = \alpha_0 + \alpha_1 H + \varepsilon$ and treat α_1 as the coefficient of interest for policy purposes.

In terms of a DAG, the mediator can be depicted as follows:

[13] In the hypothetical example I construct, I do not build in any confounders. However, in practice, education and income levels of parents would be likely confounders since they are likely to affect both G and H.



Once again, we construct our hypothetical example[14] using Excel (see spreadsheet "Mediators") and work through possible regressions to examine how they match up to our constructed theoretical model:

Suppose you have a sample of 100 people where 50 people receive prenatal healthcare.

- Create a column labeled 'H' where the first 50 values are 1 and the next 50 values are 0. Note, this assumption implies that the mean of H (μ_H) = **0.5** and the variance of H (σ_H^2) = **0.25**.

Suppose the relationship between following pregnancy-related lifestyle and dietary guidelines and receiving prenatal healthcare is given by:

G = **0.2** + **0.5 H** + ε where ε is normally distributed with mean **0** and standard deviation of **0.1**

- Create a column of 100 values. Label it as 'epsilon raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.1**)
- Copy this column over and use paste special as values into a new column and label it as 'epsilon'.
- Create a column labeled 'G' and fill it using the formula G = **0.2** + **0.5** H + ε
- The mean value of G should approximately equal = $\mathbf{0.2} + \mathbf{0.5} \times \mu_H = \mathbf{0.2} + \mathbf{0.5} \times \mathbf{0.5}$ = 0.45
- The variance of G should approximately equal $\mathbf{0.5^2} \times \sigma_H^2 + \sigma_\varepsilon^2 = \mathbf{0.5^2} \times \mathbf{0.25} + \mathbf{0.1^2}$ = 0.07
- The covariance of G and H should approximately equal $\mathbf{0.5} \times \sigma_H^2 = \mathbf{0.5} \times \mathbf{0.25}$ = 0.13

Suppose the relationship between birthweight, prenatal healthcare, and the extent to which lifestyle and dietary guidelines were followed during pregnancy is given by:
W = **6** + **0.3** H + **1.5** G + u where u is normally distributed with mean **0** and variance **0.2**.

- Create a column of 100 values. Labeled it as 'u raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.2**)
- Copy this column over and use paste special as values into a new column and label it as 'u'
- Create a column labeled 'W' and fill it using the formula W = **6** + **0.3** H + **1.5** G + u

Now we will estimate the causal impact of H on W using two different approaches. In the first approach, we will use two regressions: the first regression will feature 'W' as the dependent variable and will include 'G' and 'H' as independent variables while the second regression will feature 'G' as the dependent variable and 'H' as the independent variable. In the second approach we will estimate just one regression that features 'W' as the dependent variable and 'H' as the only independent variable.

- To do this we need to first create three new contiguous columns copying over W, H, G. Contiguous columns are required if you wish to use Excel for multiple regression.

[14] H is a dummy variable (i.e., it equals 1 for parents who received prenatal care). G is assumed to be a continuous variable that typically ranges from 0 (i.e., ignoring dietary and lifestyle guidelines) to 1 (i.e., following guidelines perfectly) although values could be negative (e.g., if a person actively seeks to do the opposite of the guidelines) or greater than 1 in theory (e.g., if a person goes even beyond the recommended guidelines).

- Estimate the regression $W = \beta_0 + \beta_1 H + \beta_2 G + \varepsilon$. What is the value for the coefficient of H? How would you interpret this coefficient formally? Is this coefficient useful for directly assessing the value of prenatal healthcare?

  In the simulation, the coefficient for H is 0.32 = $\frac{\partial W}{\partial H}$. To interpret this formally, one could say, "Receiving prenatal healthcare is associated with a 0.32 pounds increase in birthweight, holding constant the extent to which dietary and lifestyle guidelines were followed." This coefficient is not particularly useful for assessing the value of prenatal healthcare because it ignores how prenatal healthcare influences the extent to which dietary and lifestyle guidelines were followed.

- Now, estimate the regression $G = \gamma_0 + \gamma_1 H + \varepsilon$. How would you use these results in conjunction with the results in the previous regression to obtain a better measure for assessing the value of prenatal healthcare? Are there any concerns one should be aware of while estimating these regressions?

  To assess the overall value of prenatal healthcare on birthweight, we should consider $\frac{dW}{dH}$ rather than $\frac{\partial W}{\partial H}$. Our estimate for this value in the simulation is: $\frac{dW}{dH} = 0.32 + 1.49 \times 0.51 = 1.08$. This value is significantly larger than the partial derivative calculated previously. Therefore, we can see that prenatal healthcare has a direct impact on birthweight (for example, perhaps through appropriate medical interventions) as well as an indirect effect through changing parents' lifestyle and dietary choices. Though this is not a factor in the example created in the simulation, one important real-world concern is that our estimate for γ_1 could itself be biased if there are confounders like income or education which could influence both H and G. Omitting these confounders could lead to an overestimation of the impact of H on G.

- Finally, estimate the regression $W = \alpha_0 + \alpha_1 H + \varepsilon$ to directly obtain the overall value of prenatal healthcare. How close are the results obtained from the different approaches?

  The value for dW/dH in the regression with H as the only independent variable = 1.08 which matches the previous result exactly.

A more complicated constructed example for a mediator

In the more complex example, let's introduce a new variable called socioeconomic status (SES). This variable influences both the likelihood of receiving prenatal healthcare (H) and the degree to which dietary and lifestyle guidelines are followed (G). Moreover, the extent to which guidelines are followed is not only influenced by socioeconomic status but also by whether parents have received prenatal healthcare. Similar to the previous example, birthweight (W) is ultimately influenced by the adherence to guidelines and the availability of prenatal care during the pregnancy. This example can be depicted through the following DAG:

In this example, if our goal is to find the causal impact of H on W, SES is a confounder, and G is a mediator.

Once again, we can adopt two different approaches.

- Approach 1 uses two regressions. The first regression includes the mediator and the confounder as independent variables with the outcome variable as the dependent variable. The second regression includes the mediator as the dependent variable and has the confounder and the critical independent variable as the independent variables:

$W = \beta_0 + \beta_1 H + \beta_2 G + \beta_3 SES + \varepsilon$
$G = \gamma_0 + \gamma_1 H + \gamma_2 SES + \varepsilon$
The value of interest is the overall derivative $\frac{dW}{dH} = \beta_1 + \beta_2 \frac{\partial G}{\partial H} = \beta_1 + \beta_2 \gamma_1$
Note that it is important to include the confounder in both regressions. This approach where we estimate two regressions helps us identify the overall impact of H $\left(\frac{dW}{dH} = \beta_1 + \beta_2 \gamma_1\right)$, the impact of H while holding G constant $\left(\frac{\partial W}{\partial H} = \beta_1\right)$, and the extent to which H affects W through G $\left(\beta_2 \frac{\partial G}{\partial H} = \beta_2 \gamma_1\right)$

- Approach 2 is simpler as it uses only one equation where the mediator is excluded but the confounder is included:
  $W = \alpha_0 + \alpha_1 H + \alpha_2 SES + \varepsilon$
  In this equation, $\frac{dW}{dH} = \alpha_1 + \alpha_2 \frac{\partial SES}{\partial H}$. However, we know from our causal theory that H does not cause SES (rather, SES has a causal impact on H), and therefore from a causal perspective, $\frac{\partial SES}{\partial H} = 0$. Since $\frac{\partial SES}{\partial H} = 0$ from a causal perspective, $\frac{dW}{dH} = \alpha_1$. As can be seen from the Excel example, this approach yields the same answer as Approach 1.

In brief, the Excel file (see spreadsheet "Mediators2") is constructed as follows:

- A column 'SES' is constructed with 50 values of 0 and 50 values of 1. SES is a binary variable which equals one for parents of high socioeconomic status.
- A column 'upsilon' is created as a random error distributed normally with mean 0 and standard deviation 0.2.
- A column 'H' is constructed using the equation H = 0.3 + 0.4*SES + random error upsilon which is then rounded to have 0 decimal places. All values of H are also either 0 or 1 and indicate whether parents availed of prenatal healthcare.
- A column 'epsilon' is created as a random error distributed normally with mean 0 and standard deviation 0.05.
- A column 'G' is constructed using the equation G = 0.22+0.25*SES + 0.35*H + random error epsilon. G measures the extent to which parents followed dietary and lifestyle guidelines during the pregnancy.
- A column 'u' is created as a random error distributed normally with mean 0 and variance 0.05.
- A column 'W' is constructed using the equation W = 6 + 0.3*H + 1.5*G + random error u. W measures the birthweight of the infant.

With these equations, theoretically[15], from a causal perspective, $\frac{dW}{dH} = 0.3 + 1.5 * 0.35 = 0.825$
In the simulation, approach 1 and 2, both yield a value of $\frac{dW}{dH} = 0.776$

The Excel file also presents several incorrect regressions that omit the confounder so you can observe how that affects results.

III. Colliders

A collider is a variable that is influenced by both the outcome (dependent variable) and the independent variable of interest (critical independent variable). When a collider is included in the analysis, it can result in collider bias (CB), which

[15] Technically, the fact that H is rounded complicates how the confounding occurs and therefore resulting underlying theoretical derivatives can only be considered approximations.

can lead to incorrect interpretations of causality in the results. Mathematically, collider bias is equal in magnitude to omitted variable bias but acts in the opposite direction.

Constructed example for a collider

Suppose we are studying the gender wage gap and examining whether a person's income is influenced by their gender within a firm. We have collected data on gender [16] (M), income (I), and the number of neckties owned (N) by each individual. We find that N is correlated with both I and M in our statistical correlation matrix. Based on the correlations if we mistakenly include N in the regression model $I = \beta_0 + \beta_1 M + \beta_2 N + \varepsilon$, even though N is theoretically caused by I (i.e., individuals with higher incomes tend to own more neckties), we will introduce collider bias.

In terms of a DAG the collider can be depicted as follows:

Once again, we construct our hypothetical example using Excel (see spreadsheet "Colliders") and work through possible regressions to examine how they match up to our constructed theoretical model:

Suppose you have a sample of 100 people where 50 people identify as cis male.

- Create a column labeled 'M' where the first 50 values are 0 and the next 50 values are 1. Note, this assumption implies that the mean of M (μ_M) = **0.5** and the variance of M (σ_M^2) = **0.25**.

Suppose *the employer actively discriminates in favor of cis males* according to the following relationship:

I = 40000 + 10000 M + ε where ε is normally distributed with mean **0** and standard deviation of **3000**

- Create a column of 100 values. Label it as 'epsilon raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **3000**)
- Copy this column over and use paste special as values into a new column and label it as 'epsilon'.
- Create a column labeled 'I' and fill it using the formula I = **40000** + **100004** M + ε

Suppose the relationship determining the number of neckties a person owns is given by:

N = **- 6** + **0.0002** I + **5** M + u where u is normally distributed with mean **0** and standard deviation of **0.5**

- Create a column of 100 values. Label it as 'u raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.5**)
- Copy this column over and use paste special as values into a new column and label it as 'u'.
- Create a column labeled 'N' and fill it using the formula[17] N = round(**- 6** + **0.0002** I + **5** M + u, 0)

Now we will estimate the regressions, first while controlling for the collider and then without the collider.

[16] In this example, for simplicity, we are assuming M is a binary variable equaling 1 for someone who identifies as a cis-male and equaling 0 for everyone else. Our hypothetical research question therefore involves examining whether there is bias in favor of cis males relative to everyone else.

[17] For the values of the neckties to be sensible, you need to round it. However, this will cause any related derivatives to be approximations.

- Estimate the regression $I = \beta_0 + \beta_1 M + \beta_2 N + \varepsilon$. What is the sign and value for the coefficient of M?
  In the simulation, the value obtained when the collider is included is -8479.8. The coefficient is negative, large in magnitude, and statistically significantly significant even at the 0.001 % level of significance. This ***erroneous*** analysis suggests that ***the employer discriminates against cis men***.

- Now, estimate the regression $I = \beta_0 + \beta_1 M + \varepsilon$. What is the new sign and value for the coefficient of M?
  Once the collider is excluded, the coefficient changes to 10398.7 and is statistically significant even at the 0.001% level. Based on the constructed model, this coefficient can be interpreted causally. The extent of discrimination in favor of cis men matches what was built into the model (i.e., 10,000).

The above example might seem silly but there are several instances even among highly reputed published journal articles where one could argue that the included 'control' variables are colliders rather than confounders or mediators. For a good discussion of this issue, I would recommend reading Chapter 3 of Scott Cunningham's book, "Causal Inference: The Mixtape." An alternative book entirely devoted to finding causality through models represented as DAGs is Judea Pearl and Dana Mackenzie's "Book of Why: The New Science of Cause and Effect."

Constructed example of an implicit collider

A collider might sometimes be implicit (i.e., a potential control variable acts like a collider due to causal relationships involving unobservable variables). To illustrate this issue, I will construct a simple example[18] and present six different theories that could be used to model the example. Before I get into the example, let me state an underlying motivating fact: according to The Pew Research Center, the median woman earned about 82 cents on the dollar relative to the median man in 2022.

In this complex example, we have several variables of interest. Let's clarify their meanings:

- M: Male, represents gender. It is used to distinguish between cis males and others in the analysis.
- D: Discrimination, an unobservable variable whose causal impact on wage we are trying to assess.
- O: Occupation level, represents the extent to which the occupation is high paying. Depending on its role in the analysis, it could act as a confounder, mediator, or implicit collider.
- A: Ability, an unobservable variable denoting some measure of workplace productivity. It is observed by the firm but not included in the available data.
- Wage: The outcome variable, indicating the individual's wage.

With multiple variables and complicated relationships, it can be hard to identify variables as confounders, mediators, and colliders. While a full discussion of DAGs is beyond the scope of this handout, in general, when there are more than three variables involved, it helps to list all the pathways through which the key independent variable is connected to the dependent variable. A pathway is a continuous connection (with arrows going in either direction) of variables from the key independent variable to the dependent variable without any cycling through the same variables (recall the 'A' in DAG stands for acyclic). Once this is done, every pathway can be classified as an open pathway or a closed pathway. An open pathway is one where none of the variables is a collider (recall that a collider variable has arrows entering it from both sides). Controlling for a variable on an open pathway shuts that pathway down which can then help us isolate the

[18] This example is an application of one of the DAGs discussed in Chapter 3 of Scott Cunningham's book, "Causal Inference: The Mixtape." Sometime in maybe early 2023, I saw a discussion of this DAG on Twitter and my example is based on that discussion. Unfortunately, I am unable to find links to the tweets on which I am basing my example.

effect of other pathways. Controlling for a collider on a closed pathway, on the other hand, opens it up and causes collider bias.

Let us consider a few alternative theories of how M, D, O, A, and W could be causally related.

1. Theory 1: Gender-based discrimination causes all the differences in wages between genders.

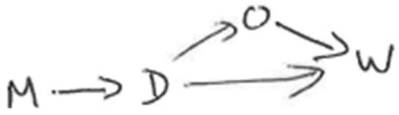
   There is only one pathway here M → D → W. It is the pathway of interest and can be estimated with a simple regression that has W as the dependent variable and M as the key independent variable.

2. Theory 2: Wages differ between genders due to occupational choice rather than due to discrimination, i.e., men *choose* higher-paying occupations while others *choos*e lower-paying occupations. According to this theory, since occupations are chosen freely, people of different genders choose occupations differently for benign reasons such as different innate preferences or comparative advantage considerations between home and labor market activity. If Theory 2 is correct, analysis based on Theory 1 will be flawed due to confounder or omitted variable bias.

   In terms of the DAG, notice there are two pathways here, both of which are open.
   (i) M → D → W
   (ii) M → O →W

   Since we are interested in the causal impact of D rather than the causal impact of M, it is necessary to include O as a control variable in order to shut down the open pathway (ii). On the other hand, if we were interested in the *overall* causal impact of ***M*** rather than the causal impact of ***D***, we would leave both paths open, i.e., we would not include O as a control variable.

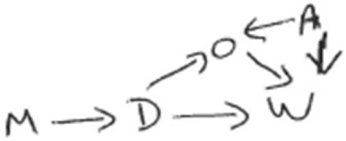
3. Theory 3: Another perspective could be that gender-based discrimination forces most people who are not males into lower paying occupations. This could be at the societal level (e.g., men are less likely to be subjected to social opprobrium for missing their children's school events) or at the employer or industry level (e.g., bosses promote or select men into higher paying occupations at higher rates). Notice that with this as our theory, the causal impact of D is mediated through O (there is an arrow that goes from D to O). Occupation therefore acts as a mediator, i.e., as one of the channels through which discrimination affects wages. With this theory, the simplistic analysis based on Theory 1 is justifiable. However, a more refined analysis with the two regressions approach as discussed in the Mediators section of this document might allow us to distinguish the relative strength of the two channels through which D affects W (the direct channel, D → W, and the indirect channel, D → O → W).



   In terms of the DAG, notice there are two pathways here, both of which are open.
   (i) M → D → W
   (ii) M → D → O →W

Since we are interested in the *overall* causal impact of D on W rather than just its direct causal impact on W, our regression model should keep both paths open. In this case, the convenient option would be to not control for O and estimate the same regression that we estimated for Theory 1. If we were to control for O, we would have to estimate a second regression with O as the dependent variable and M as the independent variable to find the relative importance of the two channels.

4. Theory 4: Yet another perspective could be that selection into higher paying occupations is driven partly by discrimination in favor of men and partly by ability. Wages are determined partly by discrimination, partly by occupation, and partly by ability. In such a situation, occupation acts as a collider variable, not directly (wages don't cause occupations), but due to the omitted unobservable variable "ability." Including occupation will therefore result in biased coefficient estimates.

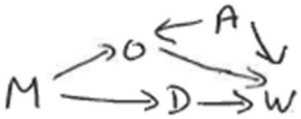


In terms of the DAG, there are three pathways here. The first two pathways are open pathways while the third pathway is closed due to the collider variable O.

(i) M → D → W
(ii) M → D → O →W
(iii) M → D → O ← A →W

If our goal is to find the overall impact of D on W, we would like to keep the first two paths open and keep the third path closed. Since the first two pathways are open pathways, they will remain open as long as we do not control for O. Since the third pathway is already closed due to collider O, we should not control for the collider if we wish to keep it closed. On the other hand, if we erroneously decide to control for O, this will shut down pathway (ii) and open pathway (iii) and will therefore bias our results in two ways.

5. Theory 5: For our fifth perspective, let us assume that that selection into higher paying occupations is not driven by discrimination. Rather, it is partly by differences in freely chosen preferences across genders and partly by ability. Wages are determined partly by discrimination, partly by occupation, and partly by ability.



In terms of the DAG, there are three pathways here. The first two pathways are open pathways while the third pathway is closed due to the collider variable O.

(i) M → D → W
(ii) M → O →W
(iii) M → O ← A →W

If our goal is to find the causal effect of D on W, we need to isolate the effect based on pathway (i). This requires closing the open pathway (ii) which requires controlling for O. However, controlling for O opens the closed pathway (iii). **Unfortunately, there is no empirical resolution to this issue**.

To illustrate this example, I am going to create a dataset based on Theory 4 and analyze it using different regressions to demonstrate the occurrence of collider bias. Since it is based on Theory 4, the dataset will be built using the following assumptions:

a. The distribution of ability is the same across genders
b. A combination of ability and discrimination will determine occupation level
c. A combination of ability, discrimination, and occupation level will determine wages

Once again, we will construct our simplistic hypothetical example[19] using Excel (see spreadsheet "Colliders2") and work through possible regressions to examine how they match up to our constructed theoretical models:

- Create a column labeled 'Male' and fill it with 1's for the first 50 rows and with 0's for the next 50 rows.
- Create a column labeled 'Ability'. Construct ability to be such that the first five males have ability = 10, the next five have an ability = 9 and so on until the last five males have an ability = 1. Use the exact same approach to construct ability values for those who are not males. Notice, by construction, ability does not vary across genders in the example.

Suppose that occupation level is determined partly through gender (due to discrimination) and partly due to ability:
Occupation Level = **2** + **4** Male + **3** Ability + υ where υ is normally distributed with mean **0** and standard deviation of **0.5**

- Create a column of 100 values. Label it as 'upsilon raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.5**)
- Copy this column over and use paste special as values into a new column and label it as 'upsilon'.
- Create a column labeled 'Occupation Level' with the formula Occupation Level = **2** + **4** Male + **3** Ability + υ. **Notice that in terms of who gets into higher occupation Levels, the example builds in discrimination against those who are not males.**

Suppose Wages are determined partly through gender (due to discrimination), partly through ability, and partly through occupation level:
Wage = **1** + **2** Male + **3** Ability + **1.5** Occupation Level + ε where ε is normally distributed with mean **0** and standard deviation of **0.8**

- Create a column of 100 values. Label it as 'epsilon raw' and fill it out using the following formula: =NORMINV(RAND(), **0**, **0.8**)
- Copy this column over and use paste special as values into a new column and label it as 'epsilon'.
  Create a column labeled 'Wage' with the formula Wage = **1** + **2** Male + **3** Ability + **1.5** Occupation Level + ε. **Notice that in terms of wages, the example builds in *further* discrimination against those who are not males.**

Now we will estimate three regressions:

The first regression is based on Theory 1, Theory 3, and Theory 4 (all these three theories suggest the same regression model). Recall, that the model itself was built based on Theory 4. The model based on all of these theories requires only the dependent variable (W) and the key independent variable (M) without controlling for O.

[19] Once again, Male is again assumed to be a binary variable (i.e., a person is classified as either a cis male or not a cis male) for enhancing the simplicity and clarity of the example. This pedagogical choice is not meant to discount the significance of considering non-binary identities whenever these are available in real world data. Ability is assumed to vary uniformly between 10 and 1 across all genders. Occupation level is a continuous variable assumed to measure the extent to which an occupation is high-paying.

- Estimate the regression Wage = $\beta_0 + \beta_1$ Male + ε. What is the sign and value for the coefficient of Male?

  In the simulation, the value obtained when the collider is excluded is 7.92. The coefficient is positive, and large in magnitude, but is statistically significant only at the 10% level of significance. This value matches the effect of discrimination built into the model (2 + 1.5 *4) =8.

- Now, estimate the regression Wage = $\beta_0 + \beta_1$ Male + β_2 Occupation Level + ε. What is the new sign and value for the coefficient of Male?

  When the collider variable, Occupation Level, is included, the coefficient changes to -2.16 and is statistically significant even at the 0.001% level. This ***erroneous analysis*** suggests that ***employers discriminate against males***.

- Finally, estimate the regression Wage = $\beta_0 + \beta_1$ Male + β_2 Occupation Level + β_3 Male x Occupation Level + ε. What is the new sign and value for the coefficient of Male? Does adding the interaction term help resolve the collider bias?

  Including the interaction term along with the collider variable does not help resolve the collider bias. The coefficient for male is statistically significant and equals -2.16. However, due to the interaction term, this coefficient is not meaningful by itself. Computing the wage premium for Males at the minimum, mean, and maximum values of occupation levels of yield the values -2.54, -2.16, and -1.79 suggesting that that ***employers discriminate against males*** at all occupation levels.

Notice that the same dataset could be said to reflect Theory 5 as well (in this case though, the difference between genders in selection into occupations is a result of free choice rather than discrimination). If this is the case, the analysis with the simple regression above includes the causal effect of the open pathway M → O →W even though this pathway is not the result of discrimination. Unfortunately, if Theory 5 is the true theory, there is no way to isolate the desired pathway M → D →W, and we must turn to other forms of analysis or obtain data on the unobserved ability variable for a more sophisticated model.

Illustrative, simple, non-regression-based version of previous example

While the above example is precise and shows how collider bias can affect the results of regressions, the underlying mechanism may be getting obscured with all the equations. The following example is simple and hopefully useful for illustrative purposes (see Excel spreadsheet "Colliders2Simple").

Once again, assume the same DAG as Theory 4. This time assume there are just two occupation levels – supervisor, and worker. Also assume there are 100 individuals of whom 50 are men. The distribution of ability is exactly the same across genders and ranges from 10 to 1 (with 5 individuals at each value).

Employers discriminate in favor of men and promote anyone with ability over 7 to be a supervisor. However, among those who are not men, only those with an ability level of 10 get promoted to the supervisor level.

Employers further discriminate in favor of males when it comes to wages. For male supervisors, they offer a wage equal to two times their ability. Among non-male supervisors, the offered wage is 1.8 times their ability. Similarly, for male workers they offer a wage equal to their ability. For non-male workers, they offer a wage of 0.9 times their ability.

A few notable observations about this constructed dataset:

- Among supervisors, males have a lower average ability than non-males (8.5 versus 10), because the selection process was less restrictive for males.

- Among workers too, males have a lower average ability (3.5 versus 5), because many able non-males were discriminated against and have remained as workers despite higher ability.
- Since wages are a function of ability, among supervisors, males have a lower average wage than non-males (17 versus 18), despite the discrimination in favor of males for wages among supervisors.
- Similarly, since wages are a function of ability, among workers too, males have a lower average than non-males (3.5 versus 4.5), despite the discrimination in favor of males for wages among workers.
- Overall though males earn higher than non-males (8.9 versus 5.85)

While I have not conducted any regressions on the basis of this simple example, it is straightforward to see that with an unobservable collider variable involved in the data creation, a statistical analysis that controls for whether an individual is a supervisor would conclude that employers discriminate against males (they receive a lower wage at each level). However, if we exclude the collider variable (occupation level of supervisor or worker) from consideration, we would indeed find that discrimination occurs in favor of men, as was built into the model.

References

Cunningham, S., 2021. Causal Inference: The Mixtape. Yale University Press, New Haven ; London.

Kochhar, S., 2023. The Enduring Grip of the Gender Pay Gap. Pew Research Center's Social & Demographic Trends Project. URL https://www.pewresearch.org/social-trends/2023/03/01/the-enduring-grip-of-the-gender-pay-gap/ (accessed 6.26.23).

Pearl, J., Mackenzie, D., 2018. The Book of Why: The New Science of Cause and Effect, 1st edition. ed. Basic Books.