POLS 602-PS4-Joseph Kim

Joseph Kim

2025-12-13

#Part 1-Q1 A confounder is a variable that has a causal relationship with both the independent variable and the dependent variable, meaning it causally affects both. A collider, by contrast, is also related to both the independent variable and the dependent variable, but in the opposite direction: it is causally affected by them rather than causing them. Confounders generate bias when they are omitted from a regression model and should therefore be controlled for to block backdoor paths. Colliders, however, do not bias estimates unless researchers condition on them; controlling for a collider opens a spurious association between the independent variable and the outcome and introduces bias. Thus, proper model specification requires controlling for confounders while avoiding conditioning on colliders, based on the underlying causal structure rather than statistical associations alone.

#Part 1-Q2 Controlling for a collider can create bias because, although the collider is not part of the causal pathway between the independent variable and the dependent variable, it is causally affected by both. Conditioning on the collider induces a spurious statistical association between the independent variable and the dependent variable, even if no such association exists causally. In DAG terms, controlling for a collider opens a previously blocked path between the independent variable and the outcome, thereby generating bias rather than removing it.

#Part 1-Q3 While statistical summaries and correlations can provide descriptive information about how variables are associated, they do not allow us to infer causal relationships. In particular, correlations cannot identify the underlying causal structure or DAG that generates the observed data. As a result, they cannot tell us whether a variable is a confounder, mediator, or collider, nor whether controlling for it would reduce or introduce bias. Decisions about which variables to include in a model therefore must be based on causal reasoning rather than statistical associations alone.

#Part 1-Q4 A kitchen sink regression refers to an approach in which researchers include a large set of variables in a regression model under the assumption that controlling for more variables will improve causal inference. This approach is problematic because not all variables are causally related to the dependent variable in a way that reduces bias. Some variables may be colliders or mediators, and conditioning on them can either introduce bias or block part of the causal effect of interest. As a result, kitchen sink regression can lead to incorrect causal conclusions by ignoring the underlying causal structure.

#Part 1-Q5 A backdoor path is a non-causal path between an independent variable and a dependent variable that creates a spurious association due to common causes, rather than a direct causal effect. These paths typically arise through confounders that influence both the treatment and the outcome. Multiple regression helps block backdoor paths by conditioning on appropriate confounders, thereby isolating the causal effect of the independent variable on the outcome. When the correct set of confounders is controlled for, the remaining association reflects the causal effect rather than bias from omitted variables.

#Part 2-Q1 T = private school Y = test score C = parents' income (confounder) M = class size (mediator)

Confounder (C) to block backdoor bias Mediator (M) to block the indirect path T→M→Y

Regression Code on R lm(test_score ~ private_school + parent_income + class_size, data = dat)

Which variables are necessary?

parent_income (C) — necessary to remove confounding (backdoor path T←C→Y) class_size (M) — necessary to isolate the direct effect (otherwise you estimate total effect)

Which variables should not be included for Q1

Collider (K) (e.g., parent involvement): do not control (induces bias) Instrument (Z): not needed for unbiased OLS given observed confounder; can add noise / instability Exogenous Y-only variable (U_Y): not necessary (may improve precision, but not required for identification)

#Part 2-Q2 Because the outcome variable (test scores) is continuous and the data-generating process is linear by construction, I estimate all models using ordinary least squares (OLS). To recover the total effect of private school attendance on test scores, I exclude the mediator (class size) while continuing to control for the confounder (parents' income).

#Part 2-Q3 The results change substantially depending on which additional variable is controlled for. When I control for the collider (parent involvement), the estimated effect of private school attendance on test scores becomes biased, as conditioning on the collider opens a spurious path between the treatment and the outcome. In contrast, controlling for the exogenous variable that affects test scores only (student ability) does not meaningfully change the estimated treatment effect, since this variable is not part of any backdoor path; it may improve precision but is not required for identification. Finally, controlling for the instrument (distance to the nearest private school) does not reduce bias and can increase estimation noise, as instruments are relevant for identifying causal effects only when confounding cannot be addressed through direct control. Overall, these results illustrate that adding variables mechanically can worsen causal inference depending on their role in the underlying causal structure.

#Part 2-Q4 Taken together, the readings and the simulation results suggest that variables should be included in a regression model based on their role in the underlying causal structure rather than their statistical association with the outcome. Confounders should be controlled for because they block backdoor paths and reduce bias, while colliders should be excluded because conditioning on them opens spurious paths and introduces bias. Mediators should be included or excluded depending on whether the goal is to estimate a direct or total effect, as controlling for mediators changes the estimand rather than improving identification. More generally, the simulation demonstrates that indiscriminately adding variables—such as instruments or outcome-only predictors—does not necessarily improve causal inference and may even be harmful, underscoring the importance of causal reasoning over mechanical model selection.

#R codes for drawing a DAG library(dagitty) library(ggdag) library(ggplot2)

dag_private <- dagitty(" dag { # Confounder income -> private income -> score

Causal effect + mediator private -> class_size class_size -> score private -> score

Exogenous Y-only variable ability -> score

Instrument (exogenous T-only cause) distance -> private

Collider (caused by T and Y) private -> involve score -> involve } ")

Plot (automatic layout)

ggdag(dag_private, layout = "stress") + theme_dag() + labs(title = "DAG: Private School → Test Scores") + theme(plot.title = element_text(hjust = 0.5, face = "bold"))