More Than Just Associations: An Introduction to Causal Inference for Sport Science

Master thesis

From

Simon Nolte

German Sport University Cologne

Cologne 2024

Thesis supervisor:

Dr. Oliver Jan Quittmann

Institute of Movement and Neurosciences

Affirmation in lieu of an oath

Herewith I affirm in lieu of an oath that I have authored this Bachelor thesis independently and did not use any other sources and tools than indicated. All citations, either direct quotations or passages which were reproduced verbatim or nearby-verbatim from publications, are indicated and the respective references are named. The same is true for tables and figures. I did not submit this piece of work in the same or similar way or in extracts in another assignment.

_____

Personally signed

Abstract

Zusammenfassung (German Abstract)

Table of Contents

Abstract

Zusammenfassung (German Abstract)

Table of Contents i

List of Figures iii

List of Tables iii

1 Introduction 1
- 1.1 Relevance . . . 1
- 1.2 Previous Research . . . 2
- 1.3 Aim . . . 2

2 Theoretical Background 3
- 2.1 Causility, Associations, and (In)dependence . . . 3
- 2.2 Graphical Causal Models . . . 3
- 2.3 Error Terms in Causal Modeling . . . 5
- 2.4 Modeling Causal Systems . . . 7
- 2.5 Confounders and Colliders . . . 7
- 2.6 Conditioning Rules: The Backdoor Criterion . . . 9

3 Methods 10
- 3.1 Data Set . . . 10
- 3.2 Causal Models Development . . . 10
- 3.3 Statistical Modeling and Evaluation . . . 10

4 Results 11
- 4.1 Confounding . . . 11
- 4.2 Collider Bias . . . 11
- 4.3 Application of the Backdoor-Criterion . . . 11
- 4.4 Development of a Causal Model for Endurance Performance . . . 11

5 Discussion 12
- 5.1 Applications in Sport Science . . . 12
  - 5.1.1 Causality in Observational Data . . . 12
  - 5.1.2 Identification of Confounders . . . 12
  - 5.1.3 Understanding Big Data . . . 12
  - 5.1.4 Study Design . . . 12
  - 5.1.5 Predicting Hypothetical Outcomes with Counterfactuals. . . . 12
- 5.2 Challenges and Limitations . . . 12
  - 5.2.1 Need for Theoretical Models . . . 12
  - 5.2.2 Data Quality . . . 12
  - 5.2.3 Complex Systems . . . 12
  - 5.2.4 Communicating Causality . . . 12

5.3 Perspectives and Further Possibilities 12
5.3.1 Modeling Unobserved Variables, Missing Data, and Measurement Error 12
5.3.2 Sampling and Survivorship Bias 12
5.3.3 Longitudinal Data 12
5.3.4 Causal Modeling Workflows in Sport Science Practice 12

6 Conclusion **13**

References **14**

A Appendix **15**
A.1 Mathematical Background 15
A.2 Simulations 16
A.3 Technical Details 17
A.3.1 Session Info 17
A.3.2 Packages 18

List of Figures

1 A simple graphical causal model with two variables. 4
2 A more complex graphical causal model with four variables. 4
3 A simple causal path, with random error. 6
4 Random errors in a causal path. 7
5 A causal path blocked by conditioning. 8
6 A graphical example of confounding. 9
7 A graphical example of collider bias. 10

List of Tables

1 Introduction

1.1 Relevance

Empirical research is acquiring knowledge through systematic observations by analyzing data. Data analysis typically encompasses three primary tasks: description, prediction, and causal inference (1, 2). Description means characterizing features in a subset of a population. Prediction means forecasting outcomes based on available data. Causal inference means making claims about causality — what would have happened under different circumstances.

Most research in sport science is of causal nature. We want to understand how sports works with the ultimate goal to intervene: If we understand why certain people or teams are winning a competition, we can use that knowledge to adjust training and tactics. Likewise, in health contexts, we seek for sport intervention that change an individual's fitness to ultimately increase well-being compared to if not intervention were undertaken. All of this is causal thinking.

Research has devised a framework for conducting studies that can infer causality without knowledge of the exact underlying causal mechanisms: the randomized controlled trial (RCT). But in sport science, RCTs are often not feasible, because of the difficulty or undesirability of implementing randomized interventions, particularly in the context of elite sports (3). Consequently, causality must often be inferred through alternative designs, such as observational studies. The field of causal inference offers tools for this particular task.

An association on its own does not inherently indicate causality, echoing the famous adage: "correlation does not imply causality." Associations observed in data may indeed stem from causality, but they can also arise from different types of bias, resulting in spurious associations. Conversely, causality does not necessarily imply correlation. Genuine causal relationships might remain obscured within the data. Distinguishing between associations and causal relationships necessitates looking beyond the data itself: it requires considering the causal model.

Causal data analysis requires something that is not relevant to most description and prediction tasks: A scientific model informed by expert-domain knowledge, that depicts the causal nature of the phenomena under investigation. This causal model serves as the foundation for all causal inference. By adhering to the rules implied by the causal model, we can analyze our data in a manner that allows for the estimation of causal effects.

I will start by establishing a working definition of causality and providing an overview of the historical development of causal inference as a research field. Following this, I will outline recent advancements in applied causal inference in across various disciplines, and offer an overview over the (mostly non-existing) literature of causal inference in sport science.

1.2 Previous Research

1.3 Aim

The aim of this thesis is to bring the methods of causal inference to sport science. The overachieving goal is to demonstrate the utility and necessity of causal methods for data analysis in sport science. I start with introducing key concepts of causal models using directed acyclic graphs. Using real-world and simulated example data, I will demonstrate concepts of collider bias, confounding, and conditioning in sport science. I will discuss opportunities that causal inference brings to sport science as well as challenges and limitations of adopting such approaches.

I aim to make the thesis as accessible as possible to readers who are new to causal inference. Detailed methodologies of modeling and mathematical formulations will be included in the appendices. My objective is to ensure that the thesis is understandable for any sport scientist with some basic statistical education. Instead of critiquing current statistical practices in sport science, the objective of this work is to showcase the effectiveness of methods that extend beyond these practices.

2 Theoretical Background

2.1 Causility, Associations, and (In)dependence

2.2 Graphical Causal Models

Graphical models are the easiest way to conceptualize causal systems. Pioneered by XXXX and YYYY, they allow to visualize causal relationships, which eases development and understanding of causal models.

A graphical causal model visualizes the exposure, outcome, covariates, and their (assumed) causal relationship. In the following, we will usually note the exposure (or independent variable) with X, the outcome (or dependent variable) with Y, and covariates (all other variables) with other letters. Note, that depending on the research question, the assignment of exposure, outcome, and covariates may change within the same model, so there can be instances, where the conventional naming cannot be followed.

Variables in graphical causal model are connected by arrows. An arrow between X and Y means, that a direct causal relationship between the two is assumed (see Figure 1) . The direction of the arrow tells the direction of causality. As in Figure 1, $X \rightarrow Y$ means that X causes Y (and not the other way around). For our definition of causality this means, if we intervene on X we expect to see a change in Y.

The direction of causality has to be determined by theoretical knowledge. It cannot be found in the data itself. Suppose that in our first example in Figure 1, X is biological sex and Y is endurance performance. It appears obvious, that a causal relationship between both exist (though it is certainly much more complicated than that seen in the simple model). However, the fact that it is sex that causes performance — and not the other way around — is based purely on theoretical knowledge and understanding of the world. There are neither randomized trials for proof (because you cannot randomly assign sex), nor controlled interventions (because you cannot easily intervene on sex) possible. Ultimately, the direction of causality is an assumption by the researcher.

Causal systems in the world are usually more complex than consisting of only exposure and outcome, and so are the graphical causal models depicting them. A more complex graph is displayed in Figure 2. X and Y are not directly connected anymore, but indirectly via B. This is called a *causal path*. We will later see, that some models also have non-causal paths.

The graph in Figure 2 is called an directed acyclic graph (DAG). It is directed, because all paths have arrows (the direction of causality is set). It is acyclic, because there are no circular paths in it. Finally, it is a graph. All graphs in this thesis will be DAGs, because the presented tools work only for these, and most research problems can be adequately formulated using them.

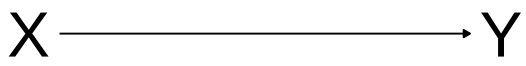

Figure 1: A simple graphical causal model with two variables. The variable X (exposure) is assumed to cause the variable Y (outcome). No other variables are believed to influence this process.

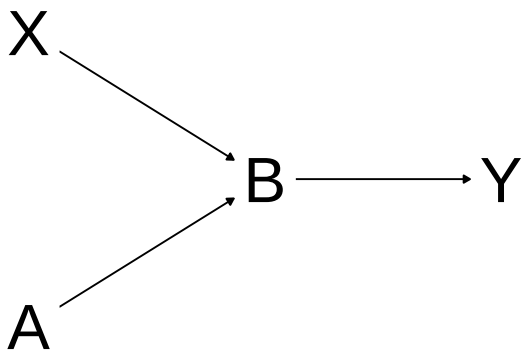

Figure 2: A more complex graphical causal model with four variables. X and A both cause B, which in turn causes Y.

What is inside a DAG is as important as what is not inside it. A DAG should depict all causal relations important to the research question. If two variables are not connected, we assume that they do not causally relate to each other. For example, in Figure 2, there is no direct connection between X and A, or between X and Y.

DAGs tell a story. For example, we can assign the variables in Figure 2 to a very simple model of endurance performance. Let X be the biological sex, A the nutrition status, B the physiological capacity to perform endurance tasks, and Y the endurance performance in a competition. Our model assumes that sex and nutrition both directly affect the physiological capacity, and this in turn affects the performance. On the other hand, it assumes that sex and nutrition are not causally related, and that both sex and nutrition have no direct effect on performance, but only an indirect effect via physiological capacity.

2.3 Error Terms in Causal Modeling

If we knew the true causal model and could measure all variables perfectly, we could exactly determine all causal effects. In reality, this is impossible. One of the main reasons are unobserved factors, that influence our relevant variables in the model. This could be things like random measurement error or biological variability. Taken together with the fact, that we can always only investigate causal effects in a sample of the population, our research will only result in an estimate of the causal effect.

As with any statistical analysis, we aim for unbiased and precise estimates. Unbiased means, that on average our estimate will correspond to the true value of the variable. Precise means, that the estimate will have a small variance, or in other words, that a single estimate is sufficiently near to the average estimate. Random error terms add imprecision, but not bias to our model. We will later learn scenarios, that introduce bias.

To demonstrate the concepts of precision and bias of causal effect estimated, I will use toy data simulations in the following. These simulations create samples (n = 100) of data corresponding to the simulated causal model including random error terms. Each simulated sample is modeled to yield a single causal effect estimate. I then visualize the distribution of simulated estimated. More details of the simulation procedure can be found in Section A.2.

Figure 3 demonstrates how unobserved factors (random error terms) add uncertainty to a causal effect estimate. Without the random error term, each sample would give the exact true causal effect. With the random error terms, some samples will give estimates that differ from the true causal effect. One main goal of causal inference is to create models, whose estimates do not differ systematically and do not differ too much from the true effect.

Certainty in causal effect estimates is higher in simpler models. The main reason for this is that simpler models have less random error terms. This can be demonstrated by comparing a simple causal relation with a causal path (a chain).

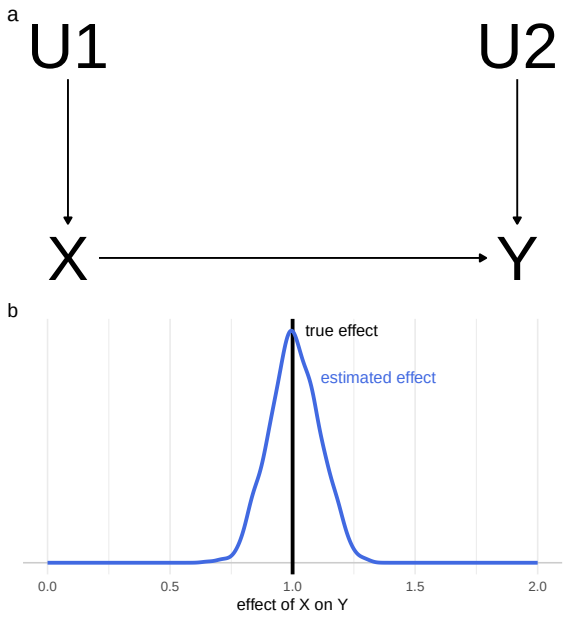

Figure 3: A simple causal path, with random error. (a) X causes Y, but both variables are influenced by other unobserved variables (random error). (b) A simulation of the model. The density plot shows the distribution of k = 1000 simulations of the model with random error terms. The random error adds uncertainty to the estimate of the causal effect, but no bias (i.e., on average, the true causal effect can be correctly estimated).

Along a causal path, information is generally lost, even if the causal effects are unaltered. The culprits are the additional error terms of intermediate variables (see Figure 4). Chains therefore introduce uncertainty, but no bias, to a causal effect estimate.

For an example from sport science think of two different causal effects. First, the effect of a running intervention on mitochondrial density. Second, the effect of a running intervention on endurance performance. Even if we assume in the second case, that the effect is directly chained trough mitochondrial density (i.e., $intervention \rightarrow density \rightarrow performance$), the effect on endurance performance is harder to estimate. The main reason is, that endurance performance will be influenced by additional unobserved factors, that will not influence mitochondrial density, for example motivation, pacing, or day-to-day variability.

Viewing the causal model in Figure 4, we have to reconsider that the arrows drawn in a DAG are as interesting as the arrows not drawn. In the example, all unobserved error terms are parent nodes, meaning that they are not influenced by any relevant variable, so also not by each other. This is a general assumption regarding unobserved error terms: We assume random errors to be uncorrelated. As soon as errors influence each other (directly or via other variables), we should model them explicitly to yield the best estimates.

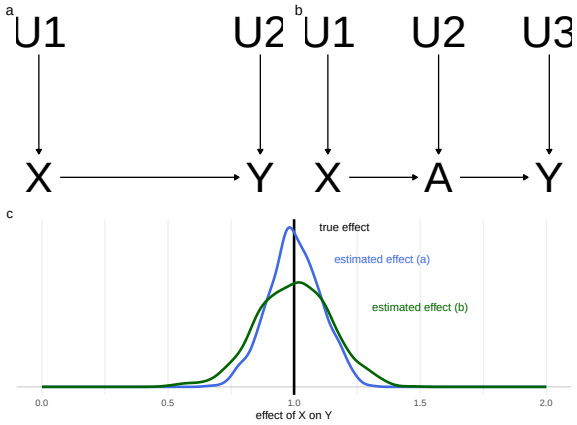

Figure 4: Random errors in a causal path. (a) X causes Y directly. Both variables are influenced by random errors. (b) X causes Y via A. All three variables are influenced by random errors. (b) A simulation of the effect of X on Y in both models. The chain introduces additional uncertainty in the effect estimate, but no bias.

2.4 Modeling Causal Systems

DAGs can be understood as linear models.

DAGs are an abstract concept to describe research problems. This level of abstraction allows to plan a study and its data analysis on a conceptual level. For the actual data analysis, a DAG has to be filled with data and functions.

Arrows in a DAG represent a causal relationship, but they do not specify the type of this relationship. $X \rightarrow Y$ can mean a strong linear affect of X on Y, or a weak negative curvelinear relation. While the researcher is free to choose what exact type of relation a DAG assumes for a set of two variables, a linear relationship is the common choice.

2.5 Confounders and Colliders

Confounders are variables that influence both the exposure and the outcome causally (see Figure 6 a). The confounder creates a spurious (non-causal) association between exposure and outcome. Conceptually, a confounder gives a set of similar information (knowledge) to both exposure and outcome. This leads to both sharing a set of information, regardless of their actual causal relationship. The actual causal relationship is biased.

Confounders can be controlled for by conditioning on them in a model. This removes the entire bias and preserves the actual causal relationship.

Let's take an example by looking at Figure 6. We are interested in the relationship between the (average) 5000-m time trial speed and the (average) 100-m sprint speed. We assume, that being fast in an endurance task reduces the ability to sprint fast, and thus reduces the 100-m speed. Therefore, we are interested in the causal relationship between X (endurance

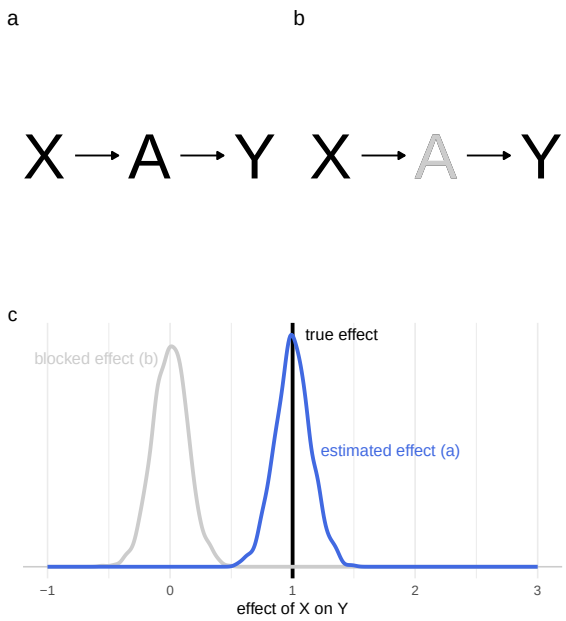

Figure 5: A causal path blocked by conditioning. (a) X causes Y via A. (b) The causal path is blocked, because the analysis conditions on A. As all affects of X on Y trail through A, no causal effect remains. (c) A simulation of the effect of X on Y in both models. Blocking removes the true causal effect entirely.

speed) and Y sprinting speed. Note that this is a very simplistic causal model, as we could also model the unobserved ability to sprint and ability to perform endurance tasks, as well as their potential causes.

Our model has a collider A, the biological sex. From expert knowledge, we know that sex causally influences both sprinting and endurance performance, mainly via anthropometry and physiology. Sex thus biases the causal relationship between sprinting and endurance performance. To remove this bias, the analysis has to control for sex. For a discrete variable such as sex is typically documented as, controlling for means in practice stratifying the analysis by it. Assuming our causal model is correct — which holds of course not true in our simple example here — controlling for sex gives us the true (unbiased) causal relationship between endurance and sprinting performance.

Colliders pose a more subtle form of bias. A collider is a variable that is causally influenced by the exposure and the outcome (see Figure 7 a). Per se, colliders do not cause harm. But when they are conditioning on the introduce bias into a model. This collider bias can be understood by the following: A collider combines knowledge from both its source, the exposure and the outcome, and thus also of their causal relationship. If this combined knowledge is being removed from a model by conditioning on the collider, then some of the actual causal relationship between exposure and outcome is also removed.

Consider the causal relationship between X post-lactate in a ramp test and Y maximum oxygen uptake in the same ramp test. Basically, our question is if more lactate causes a

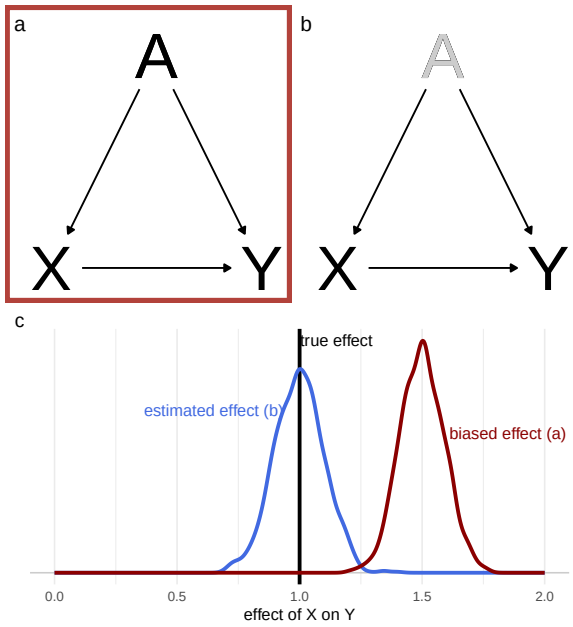

Figure 6: A graphical example of confounding. Both X and Y share a common cause A. (a) This confounder biases determining the causal effect of X on Y. (b) Conditioning on A removes the bias in the analysis. (c) A simulation of the effect of X on Y in both models.

higher or lower maximum oxygen uptake. In our model, both lactate and VO2max influence the maximum velocity in the ramp test. This appears reasonable, as individuals with a more capable glycolytic or oxidatative energy metabolism are likely to outperform their counterparts that have neither in term of the maximum velocity. The maximum velocity is thus the collider B. Conditioning on it will bias our model.

2.6 Conditioning Rules: The Backdoor Criterion

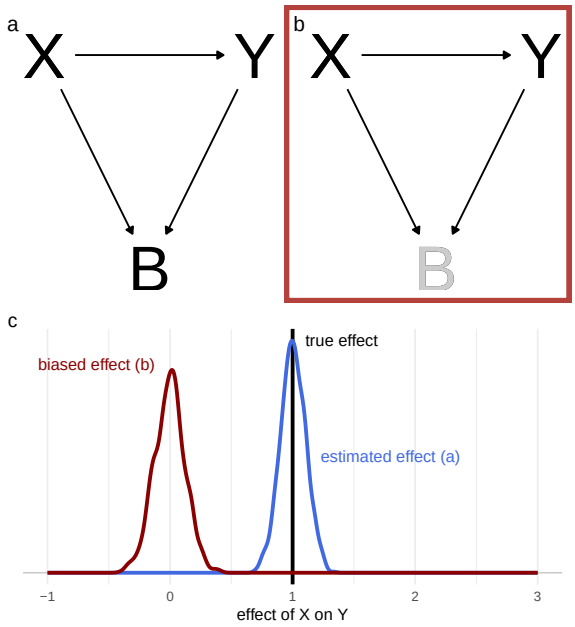

Figure 7: A graphical example of collider bias. Both X and Y directly affect the collider B. (a) As long as B is not conditioned on, the causal effect of X on Y is unbiased. (b) Conditioning on B will introduce bias in the model. (c) A simulation of the effect of X on Y in both models.

3 Methods

I conducted all analyses in this thesis using R version 4.3.1 (4) in the RStudio IDE version 2023.09.1.494 (5). The thesis was written in Quarto version 1.3.450 (6). The default settings and attached packages are documented in Appendix Section A.3. The DAGs in this thesis were drawn using the ggdag R package (7), which is based on the software daggity (8). All source code of this project is available at GitHub.

3.1 Data Set

3.2 Causal Models Development

3.3 Statistical Modeling and Evaluation

4 Results

4.1 Confounding

4.2 Collider Bias

4.3 Application of the Backdoor-Criterion

4.4 Development of a Causal Model for Endurance Performance

5 Discussion

5.1 Applications in Sport Science

5.1.1 Causality in Observational Data

5.1.2 Identification of Confounders

5.1.3 Understanding Big Data

5.1.4 Study Design

5.1.5 Predicting Hypothetical Outcomes with Counterfactuals.

5.2 Challenges and Limitations

5.2.1 Need for Theoretical Models

5.2.2 Data Quality

5.2.3 Complex Systems

5.2.4 Communicating Causality

5.3 Perspectives and Further Possibilities

5.3.1 Modeling Unobserved Variables, Missing Data, and Measurement Error

5.3.2 Sampling and Survivorship Bias

5.3.3 Longitudinal Data

5.3.4 Causal Modeling Workflows in Sport Science Practice

6 Conclusion

References

1. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: A classification of data science tasks. *CHANCE*. 2019;32(1):4249.
2. Carlin JB, Moreno-Betancur M. On the uses and abuses of regression models: A call for reform of statistical practice and teaching. 2023; Available from: http://arxiv.org/abs/2309.06668.
3. Bullock GS, Ward P, Hughes T, Thigpen CA, Cook CE, Shanley E. Using Randomized Controlled Trials in the Sports Medicine and Performance Environment: Is It Time to Reconsider and Think Outside the Methodological Box? *Journal of Orthopaedic & Sports Physical Therapy* [Internet]. 2023; Available from: https://www.jospt.org/doi/10.2519/jospt.2023.11824. doi:10.2519/jospt.2023.11824.
4. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: 2023. Available from: https://www.R-project.org/.
5. Posit team. *RStudio: Integrated development environment for r*. Boston, MA: Posit Software, PBC; 2023. Available from: http://www.posit.co/.
6. Allaire JJ, Teague C, Scheidegger C, Xie Y, Dervieux C. *Quarto*. 2023. Available from: https://github.com/quarto-dev/quarto-cli.
7. Barrett M. *Ggdag: Analyze and create elegant directed acyclic graphs*. 2024. Available from: https://github.com/r-causal/ggdag.
8. Textor J, Zander B van der, Gilthorpe MS, Liśkiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: The r package 'dagitty'. *International Journal of Epidemiology*. 2016;45(6):1887–94.

A Appendix

A.1 Mathematical Background

A.2 Simulations

For demonstrating the basic concepts of causal inference I use simulations of simple linear models. The exposure is normally distributed as $X \sim N(0, 1)$. For the simplest causal inference path of $X \rightarrow Y$, Y is a linear combination of X and an (in reality unobserved) error term $U_1 \sim N(0, 1)$. Therefore, the true causal effect of X on Y equals 1. More complex simulation models work in the same way, with each variable given by a linear combination of its ancestor variables and a random error term.

The causal effect in each simulation is estimated by a linear regression model. For the simplest model of $X \rightarrow Y$, this means estimating the regression coefficient b_1 of $Y = b_1 * X + b_0 + \epsilon$ via ordinary least square estimation with the R Code `lm(Y ~ X, data)`. For each simulation, the estimated regression coefficient is assumed to be the best unbiased estimate of the causal effect, creating a distribution of estimated causal effects.

A.3 Technical Details

A.3.1 Session Info

```
sessionInfo()
```

```
R version 4.3.1 (2023-06-16 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 11 x64 (build 22631)

Matrix products: default


locale:
[1] LC_COLLATE=German_Germany.utf8  LC_CTYPE=German_Germany.utf8
[3] LC_MONETARY=German_Germany.utf8 LC_NUMERIC=C
[5] LC_TIME=German_Germany.utf8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] patchwork_1.2.0 ggplot2_3.5.0   ggdag_0.2.12    dagitty_0.3-4

loaded via a namespace (and not attached):
 [1] viridis_0.6.5      utf8_1.2.4         generics_0.1.3     tidyr_1.3.1
 [5] stringi_1.8.3      digest_0.6.35      magrittr_2.0.3     evaluate_0.23
 [9] grid_4.3.1         fastmap_1.1.1      rprojroot_2.0.4    jsonlite_1.8.8
[13] ggrepel_0.9.5      gridExtra_2.3      purrr_1.0.2        fansi_1.0.6
[17] viridisLite_0.4.2  scales_1.3.0       tweenr_2.0.3       cli_3.6.2
[21] rlang_1.1.3        graphlayouts_1.1.1 polyclip_1.10-6    tidygraph_1.3.1
[25] munsell_0.5.0      cachem_1.0.8       withr_3.0.0        yaml_2.3.8
[29] tools_4.3.1        memoise_2.0.1      dplyr_1.1.4        colorspace_2.1-0
[33] here_1.0.1         boot_1.3-28.1      curl_5.2.1         vctrs_0.6.5
[37] R6_2.5.1           lifecycle_1.0.4    stringr_1.5.1      V8_4.4.2
[41] MASS_7.3-60        ggraph_2.2.1       pkgconfig_2.0.3    pillar_1.9.0
[45] gtable_0.3.4       glue_1.7.0         Rcpp_1.0.12        ggforce_0.4.2
[49] xfun_0.43          tibble_3.2.1       tidyselect_1.2.1   rstudioapi_0.16.0
```

```
[53] knitr_1.45          farver_2.1.1        htmltools_0.5.8     igraph_2.0.3
[57] labeling_0.4.3      rmarkdown_2.26      compiler_4.3.1
```

A.3.2 Packages

```
# p_used <- suppressMessages(unique(renv::dependencies(path = "../")$Package))
# p_inst <- as.data.frame(installed.packages())
# out <- p_inst[p_inst$Package %in% p_used, c("Package", "Version")]
# rownames(out) <- NULL
# out
```