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 Background . . . 1
- 1.2 Previous Research . . . 1
- 1.3 Aim . . . 1

2 Theoretical Background 2
- 2.1 Graphical Causal Models . . . 2
- 2.2 Modelling causal systems . . . 2
- 2.3 Colliders and Confounders . . . 2
- 2.4 Conditoning Rules: The backdoor criterion . . . 2
- 2.5 Counterfactuals . . . 2

3 Methods 3
- 3.1 Data set . . . 3
- 3.2 Creating causal models . . . 3
- 3.3 Statistical modelling and evaluation . . . 3
- 3.4 Simulations . . . 3

4 Results 4
- 4.1 Confounding . . . 4
- 4.2 Collider bias . . . 4
- 4.3 Application of the backdoor-criterion . . . 4
- 4.4 Development of a causal model for endurance performance . . . 4

5 Discussion 5
- 5.1 Applications in Sport Science . . . 5
  - 5.1.1 Causality in observational data . . . 5
  - 5.1.2 Identification of confounders . . . 5
  - 5.1.3 Understanding big data . . . 5
  - 5.1.4 Study design . . . 5
  - 5.1.5 Outcome projection . . . 5
- 5.2 Challenges and Limitations . . . 5
  - 5.2.1 Need for theoretical models . . . 5
  - 5.2.2 Data Quality . . . 5
  - 5.2.3 Complex Systems . . . 5

5.3 Perspectives and Further Possibilities 5
5.3.1 Unobserved variables and measurement error 5
5.3.2 Sampling and survivorship bias 5
5.3.3 Longitudinal data 5
5.3.4 Causal modelling workflows in sport science practice 5

6 Conclusion **6**

References **7**

A Appendix **8**
A.1 Mathematical Background 8
A.2 Technical Details 8
A.2.1 Session Info 8
A.2.2 Packages 9

List of Figures

1 A simple dag 2

List of Tables

1 Introduction

1.1 Background

1.2 Previous Research

1.3 Aim

2 Theoretical Background

2.1 Graphical Causal Models

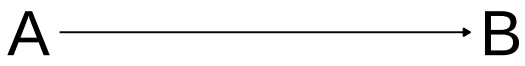

Figure 1: A simple dag

2.2 Modelling causal systems

2.3 Colliders and Confounders

2.4 Conditoning Rules: The backdoor criterion

2.5 Counterfactuals

3 Methods

I conducted all analyses in this thesis using R version 4.3.1 (1) in the RStudio IDE version 2023.09.1.494 (2). The thesis was written in Quarto version 1.3.450 (3). The default settings and attached packages are documented in Appendix Section A.2. The DAGs in this thesis were drawn using the ggdag R package (4), which is based on the software daggity (5). All source code of this project is available at GitHub.

3.1 Data set

3.2 Creating causal models

3.3 Statistical modelling and evaluation

3.4 Simulations

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 Outcome projection

5.2 Challenges and Limitations

5.2.1 Need for theoretical models

5.2.2 Data Quality

5.2.3 Complex Systems

5.3 Perspectives and Further Possibilities

5.3.1 Unobserved variables and measurement error

5.3.2 Sampling and survivorship bias

5.3.3 Longitudinal data

5.3.4 Causal modelling workflows in sport science practice

6 Conclusion

References

1. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: 2023. Available from: https://www.R-project.org/.
2. Posit team. *RStudio: Integrated development environment for r*. Boston, MA: Posit Software, PBC; 2023. Available from: http://www.posit.co/.
3. Allaire JJ, Teague C, Scheidegger C, Xie Y, Dervieux C. *Quarto*. 2023. Available from: https://github.com/quarto-dev/quarto-cli.
4. Barrett M. *Ggdag: Analyze and create elegant directed acyclic graphs*. 2024. Available from: https://github.com/r-causal/ggdag.
5. 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 Technical Details

A.2.1 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] 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      withr_3.0.0        cachem_1.0.8       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.2.2 Packages

```
p_used <- unique(renv::dependencies(path = "../")$Package)
```

```
Finding R package dependencies ... Done!
```

```
p_inst <- as.data.frame(installed.packages())
out <- p_inst[p_inst$Package %in% p_used, c("Package", "Version")]
rownames(out) <- NULL
out
```

```
    Package Version
1   dagitty   0.3-4
2     ggdag  0.2.12
3   ggplot2   3.5.0
4      here   1.0.1
5      renv   1.0.5
6 rmarkdown    2.26
```