Causal Parameter Estimation and Application On Sports Data

Stylianos Grammatikakis, supervised by Sofia Triantafillou
University of Crete, Department of Mathematics and Applied Mathematics

Introduction

The thesis is divided into two parts:

- The first part provides a detailed explanation of the fundamental concepts of causality and causal inference, with a focus on G-Estimation. G-Estimation is a method used to estimate the effect of an intervention adjusting for potential confounders.

- The second part applies G-Estimation to sports data, specifically analyzing NBA star Stephen Curry's performance and its impact on the Golden State Warriors' results. The study focuses on whether playing consecutive games affects Curry's performance and if his form influences team outcomes. Using nine seasons of box score data from 2009 to 2018, covering 878 games, the analysis provides insights into the relationship between Curry's performance and team success.

Methodology

We start by defining a treatment variable A (binary) and an outcome of interest Y (continuous or binary).

$Y^{A=a}$: The outcome variable that would have been observed under $A = a$.

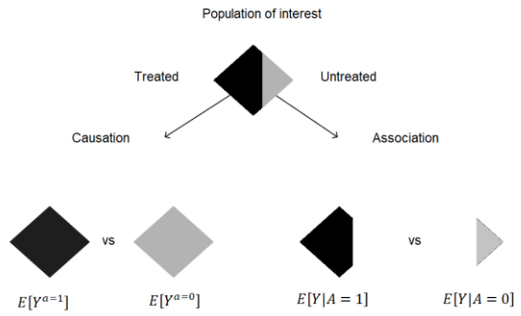
Average causal effect of A on Y:

$$E[Y^{A=1} - Y^{A=0}]$$

Estimating average causal effects with observational data requires **conditional exchangeability** of the treated and untreated (for some variables L), that is

$$Y^A \perp A|L$$

Association vs Causation:



Confounding is a form of lack of exchangeability between the treated and untreated when treatment and outcome share a common cause.

A sufficient set for confounding adjustment is a set of potential confounder variables L, for which conditional exchangeability holds.

G-Estimation estimates the average causal effect of A on Y adjusting for potential confounders L.

$$E[Y^{A=1} - Y^{A=0}|L] = \beta_1$$

Working with the rank preserving model

$$Y^{a=0} = Y - \beta_1 a$$

and considering $Y^{a=0}(\beta) = Y - \beta a$, we fit the logistic model

$$logit\left(P\left[A = 1|Y^{a=0}(\beta), \mathbf{L}\right]\right) = a_0 + a_1\left(Y - \beta a\right) + \mathbf{a_2 L}$$

Season	Treatment P-value	L_1 P-value	L_2 P-value	A.C.E	W-L	C.R	O.R
2009-2010	0.519	0.414	0.839	**1.335**	26 – 56	13	26
2010-2011	0.036	0.599	0.656	**-3.029**	36 – 46	12	20
2012-2013	0.09	0.72	0.012	**1.093**	47 – 35	6	10
2013-2014	0.694	0.134	0.713	**3.351**	51 – 31	6	8
2014-2015	0.434	0.58	0.787	**4.641**	67 – 15	1	1
2015-2016	0.142	0.577	0.061	**0.661**	73 – 9	1	1
2016-2017	0.096	0.489	0.393	**-0.238**	67 – 15	1	1
2017-2018	0.002	0.3	0.368	**7.148**	58 – 26	2	3

G-algorithm produces a consistent estimator $\hat{\beta}$ of β_1 in closed form, which results to a logistic model where a_1 is as close to 0 as possible. The idea is based on conditional exchangeability which means that $Y^{a=0}$ does not predict A given L.

Application

We define the following **outcomes** of interest:

- Y_{GS} =Game score metric for Steph Curry.
- $Y_R = \begin{cases} 1, & team\ won\ the\ game \\ 0, & team\ lost\ the\ game \end{cases}$

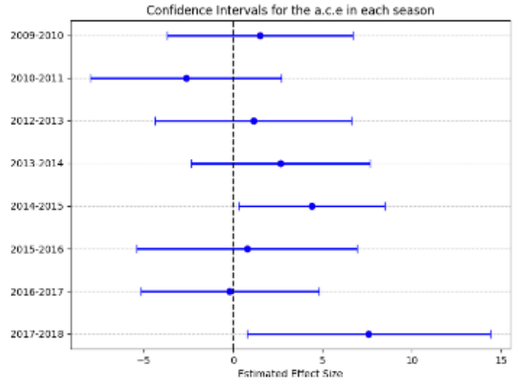
Treatment variables:

- $T_C = \begin{cases} 1, & the\ game\ he\ plays\ is\ a\ day\ away \\ & \quad from\ the\ previous\ one \\ 0, & otherwise \end{cases}$
- $T_F = \begin{cases} 1, & The\ mean\ of Y_{GS}\ in\ 4\ previous\ games \geq 20 \\ 0, & The\ mean\ of Y_{GS}\ in\ 4\ previous\ games < 20 \end{cases}$

Possible **confounders**:

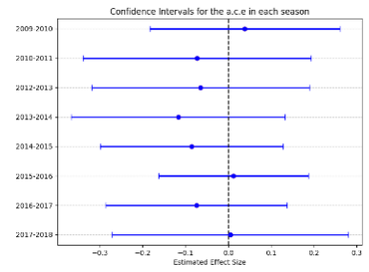
- L_1 = Curry's number of minutes played in the previous game
- L_2 = Points opponent score – Points GSW scored in the previous game

Results from 8 structural nested mean models (for each season) to estimate the average causal effect of T_C on Y_{GS} adjusting for L_1, L_2 and confidence intervals for the estimated effect.



Same for the average causal effect of T_F on Y_R adjusting for L_2.

Season	Treatment P-value	L_2 P-value	A.C.E	W-L	C.R	O.R
2009-2010	0.000	0.809	**0.037**	26 – 56	13	26
2010-2011	0.000	0.607	**-0.07**	36 – 46	12	20
2012-2013	0.000	0.012	**-0.057**	47 – 35	6	10
2013-2014	0.000	0.686	**-0.115**	51 – 31	6	8
2014-2015	0.000	0.762	**-0.074**	67 – 15	1	1
2015-2016	0.002	0.067	**0.033**	73 – 9	1	1
2016-2017	0.000	0.427	**-0.065**	67 – 15	1	1
2017-2018	0.000	0.514	**0.015**	58 – 26	2	3



Conclusion

- **G-estimation** could yield more accurate results if we had more enriched data (e.g. play by play data for sports) and adjusting for other potential confounder variables like the strength of the opponent in each game.
- **Causality** is a powerful tool for understanding relationships between events, but its accuracy depends on assumptions and conditions that we may never know if they are satisfied. In sports, causal inference can help uncover the true impact of various factors on athletes' performance, guiding coaches and organizations in making informed decisions and developing new strategies.

References

- Miguel A. Hermán & James M. Robins, Causal Inference: What If, October 1, 2019.
- Stijn Vansteelandt & Marshall Joffe, Structural Nested Models and G-estimation: The Partially Realized Promise, Institute of Mathematical Statistics, 2014.