Causal Analysis in Social Research

Walter R Davis

National Institute of Applied Statistics Research Australia
University of Wollongong

Frontiers in Social Statistics Methodology
8 February 2017

Causal analysis is back

- Social and health policy research, evaluation and prediction is an inherently causal framework.
- In the absence of experimental data, researchers have tried to draw causal inferences based on observational data.
- There are legitimate questions whether this is even possible and it became increasingly unfashionable in the social sciences to discuss models in a causal framework, relying instead on a vocabulary of “association.”
- However, over the last two decades, the causal framework has come back into fashion, especially in economics and increasingly in the other social sciences, but in a somewhat different form.

Some key sources

- Donald Rubin: *Matched Sampling for Causal Effects* (Cambridge UP, 2006).
- Judea Pearl: *Causality: Models, Reasoning and Inference* (Cambridge UP, 2009).
- Morgan & Winship: *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (Cambridge UP, 2nd ed 2015).
- Peter Austin: "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, 2011, 46:399-424.
- Bollen & Pearl: "Eight myths about causality and structural equation models" in Morgan (ed.) *Handbook of Causal Analysis for Social Research* (Springer, 2013).

Some software

- twang tutorial (R): https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf
- twang-mnps (R): https://cran.r-project.org/web/packages/twang/vignettes/mnps.pdf
- twang/mnps for SAS: http://www.rand.org/statistics/twang/mnps-sas-tutorial.html
- Matching (R): Sekhon, JS 2011. “Multivariate and propensity score matching software with automated balance optimization: The Matching package for R,” *Journal of Statistical Software*, 42:1-52.
- Causal effects in Stata: http://www.stata.com/features/treatment-effects/

Why causality matters
. . . and is unavoidable?

Simpson's paradox

Pearl (2009:174-182) presents a nice example based on Simpson's paradox. Simpson's paradox is the situation where a relationship between two variables is positive at an aggregate level but is much weaker or even negative at a sub-group level. In probability terms, for a cause $C = 0, 1$ (e.g. a drug) and an effect $E = 0, 1$ (e.g. recovery) and a third, non-experimental factor F:

$$P(E = 1|C = 1) > P(E = 1|C = 0)$$
$$P(E = 1|C = 1, F = 1) < P(E = 1|C = 0, F = 1)$$
$$P(E = 1|C = 1, F = 0) < P(E = 1|C = 1, F = 0)$$

Which table to consult?

Marginal	E=1	E=0	Tot	Rate
C=1	20	20	40	50%
C=0	16	24	40	40%

F=0	E=1	E=0	Tot	Rate
C=1	18	12	30	60%
C=0	7	3	10	70%

F=1	E=1	E=0	Tot	Rate
C=1	2	8	10	20%
C=0	9	21	30	30%

What is F?

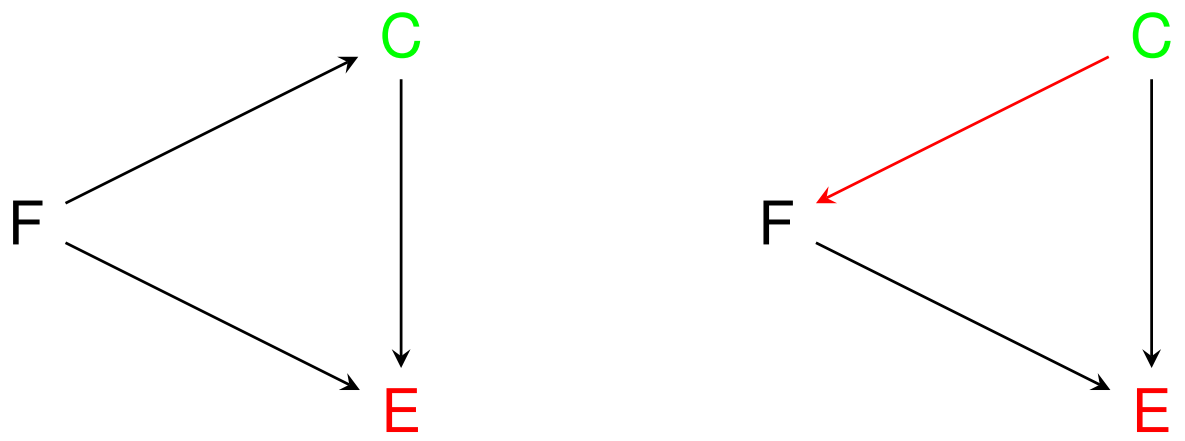


Figure: Simpson's Paradox – What is F?

Key characteristics I

Assume that we are interested in the "effect" of a treatment T (usually 0/1) on an outcome Y.

- Observational vs. experimental: There is a *selection* mechanism rather than an assignment mechanism. Those that select the treatment may differ from those that do not in characteristics that also affect Y. Moreover, those who would benefit most from the treatment may be more likely to select the treatment.
- Cause vs. association: Action is not the same as observation; the current interest in causal modeling is less about understanding what has been going on and is more to assess the effect of a change in one part of a system.
- Atomic vs. systemic: The focus therefore is generally on a small portion of a complex causal system, often a single causal effect from one T to one Y, not estimation for the entire system.

Key characteristics II

- Counterfactuals: What would have happened (what is likely to happen) if the treatment had/had not been applied (to a particular sub-group).
- Stronger foundations: Much greater attention is now paid to theoretical justification, incorporation of past research, clarity of assumptions, model identification, determining the conditions under which the causal effect may be testable and, when possible, testing whether those conditions exist.
- Nonlinear, non-parametric: The literature makes a big deal about how the theoretical approach makes no assumptions about functional form or distribution of error terms, etc. In practice however, estimation will generally require assumptions and most research continues to rely on generalized linear (mixed) models.

Conceptual similarities

There is overlap with:

- the traditional structural equations framework;
- missing data and imputation;
- matching/linking;
- self-selection and non-compliance in experimental settings;
- standardization (e.g. age-sex) of survey sub-group estimates;
- and even data fusion.

A graphical model for clarity

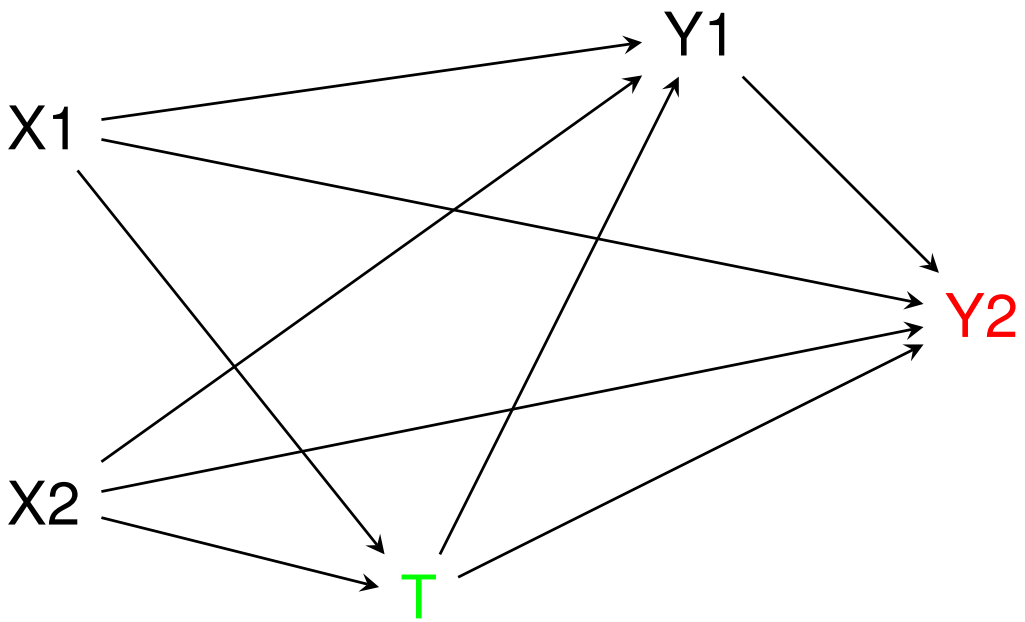


Figure: Causal Model – Directed Acyclic Graph (DAG)

The well-known problems remain

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

- At the end of the day, observational data is what's available and the causal model can only be consistent or inconsistent with those data.
- Many different causal models would lead to the same observational data and they cannot be distinguished.
- There is always a set of untestable causal assumptions underlying a causal model.
- Identification and estimation of the causal effect relies on an assumption that applying the treatment does not change the other causal mechanisms affecting Y.

Two main streams

Potential Outcomes (aka Rubin causal models): From a multitude of articles by Donald Rubin, Paul Rosenbaum and many other colleagues from about 1975-1995 and again in the 2000s. This is the main stream in economics and it has developed its own terminology. The main criticism within economics has come from James Heckman but more in the sense that this ties in with traditional structural equations.

Graphical Models: From a multitude of articles by Judea Pearl and many other colleagues from about 1985 to today. Algebraically there is little/no difference between Pearl and Rubin (as any number of Pearl articles demonstrate) but certainly the basics of causal modeling, especially model identification, tend to be easier to grasp from a graphical perspective. Pearl has also established the connection between Rubin's work and traditional structural equations.

The main notion is that, if we could, we would run an experiment where half received the treatment . . . then turn back time and give the other half the treatment. Then for every person, we would know their value of Y without the treatment (Y_i^0) and with the treatment (Y_i^1) and we could easily determine the effect of the treatment.

We would like to estimate the average treatment effect (ATE) $E(Y^1 - Y^0)$ observed across all observations. Of course all we observe are Y_1^1 and Y_0^0 where the subscript denotes whether the unit was in the treatment or non-treatment group. The notion then is to consider the potential outcomes, or "counterfactuals", Y_1^0 and Y_0^1.

Average treatment effect (ATE): $E(Y^1 - Y^0)$ which is $E(Y^1) - E(Y^0)$.

Average treatment effect for the treated (ATT): $E(Y_1^1 - Y_1^0)$.

Average treatment effect for the control (ATC): $E(Y_0^1 - Y_0^0)$.

There will be times when the ATT (or ATC) is identified but the ATE is not. There will also be times when the ATE may be identified in the population but, within a given sample, only the ATT can be reliably estimated.

The naive estimator

The naive estimator of the ATE is $E(Y_1^1) - E(Y_0^0)$. This can be rewritten in several ways in terms of the potential outcomes. Define $\delta_t = E(Y_t^1 - Y_t^0)$ (for $t = 0$ to 1) and π as the proportion of the population in $T = 1$. These can also be considered after conditioning on a set of confounding variables X:

$$\begin{aligned} naive &= ATE + E(Y_1^0 - Y_0^0) + (1-\pi)E(\delta_1 - \delta_0) \\ &= ATT + E(Y_1^0 - Y_0^0) \\ &= ATC + E(Y_1^1 - Y_0^1) \end{aligned}$$

Assumptions of the naive estimator I

The naive estimator only results in the true ATE when those differences in expectation resolve to zero (or magically cancel one another). The assumption that $E(Y_1^0) = E(Y_0^0)$ is an assumption that those who selected the treatment would have had the same outcome compared with those who did not select the treatment. Its violation would result in **baseline bias** (Morgan & Winship 2015).

The assumption that $E(\delta_1) = E(\delta_0)$ is an assumption that the treatment effect for those who chose the treatment is the same as for those who did not. Morgan & Winship refer to its violation as **differential treatment effect bias**. In the real world, one factor that often plays a role in a decision such as getting a post-graduate degree or buying a house or exercising vigorously is the (often correct) opinion about whether that treatment will be of benefit.

Assumptions of the naive estimator II

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

When both assumptions hold, the ATE, ATT and ATC are all the same and the (conditional) naive estimator is unbiased. When there is no baseline bias, the naive estimator is unbiased for the ATT. There is rarely interest in the ATC and $E(Y_1^1 - Y_0^1)$ being zero is unlikely unless both assumptions for the ATE are met anyway.

Identification and estimation by conditioning

For a DAG, if we can determine and measure the complete set of variables X that enter into both the function for T and the function for Y, then the causal path from T to Y is identified and estimable from observed data. This is the *back door criterion* (Pearl 2009). More generally, condition on the "parents" of T.

There are three main approaches to estimation by conditioning:

- Inverse probability of treatment weighting (IPTW) or related matching/balancing weighting methods
- Weighted Regression using IPTW ("doubly robust estimators")
- Instrumental variable estimation (for "selection on the unobservables," i.e. the back door criterion does not apply in full)

Back door criterion

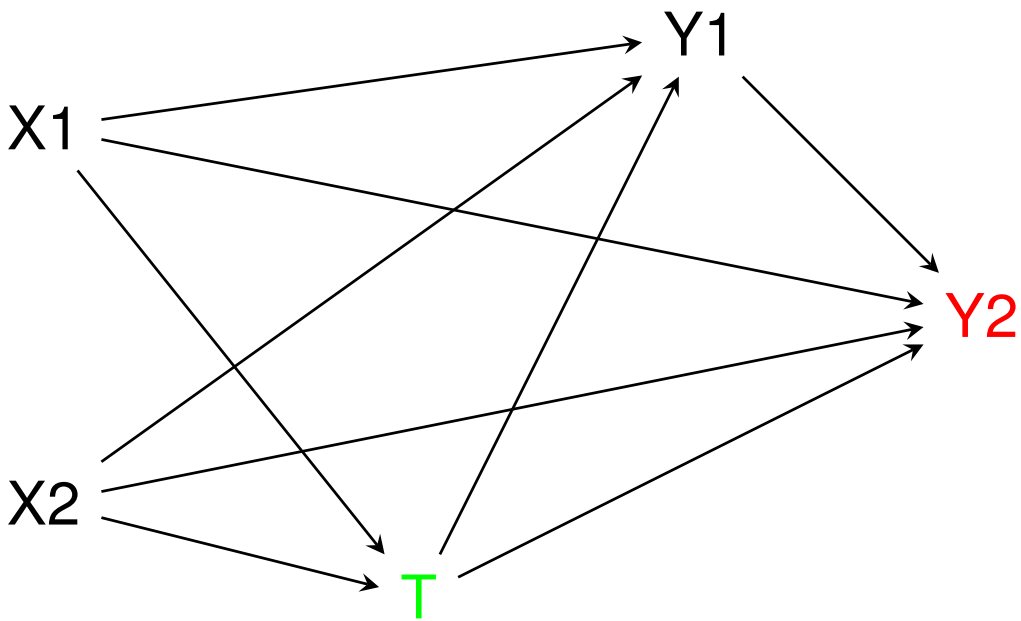


Figure: Causal Model – Directed Acyclic Graph (DAG)

Propensity scoring/ matching/balancing approaches I

If the propensity score was known (and had complete overlap) then all that would be needed is to weight by its inverse and calculate the difference in weighted means of Y by T. More realistically, build a model for T conditional on X:

- This model can be used to generate propensity scores which can then be used to weight the data for analysis (IPTW). The model and subsequent weights can incorporate existing survey weights.
- The linear predictor $X\beta$ (usually preferred to the propensity score) can be used to match treatment and control cases. There are several approaches:

Propensity scoring/ matching/balancing approaches II

- Create strata: 5 of equal size reduces about 90% of the bias (Cochran 1968; Rosenbaum & Rubin 1984). Or you can optimize the number of strata and allocation to minimize within-stratum variances.
- Nearest neighbor: Many variations on this – strict nearest neighbor; a weighted group of nearest neighbors; a weighted group of neighbors within a range; kernel matching. See Morgan & Winship for a summary.

- Optimal balancing: The Matching package in R uses a genetic matching algorithm to achieve this. The twang package uses Generalized Boosted Models. See also Morgan & Winship.

Regression approaches

Not recommended on its own generally. Even if all of the relevant X are measured, the model breaks down if there is heterogeneity in the causal effect across individuals or propensity strata not incorporated into the model. That is, if the functional form is not correct.

Doubly robust regression

Perform a weighted regressions of Y on X, using the IPTW/matching/balancing weights. In a sense, this gives you two chances to get the model right or at least (hopefully) a second chance to eliminate some of the remaining confounding.

Selection on the observed

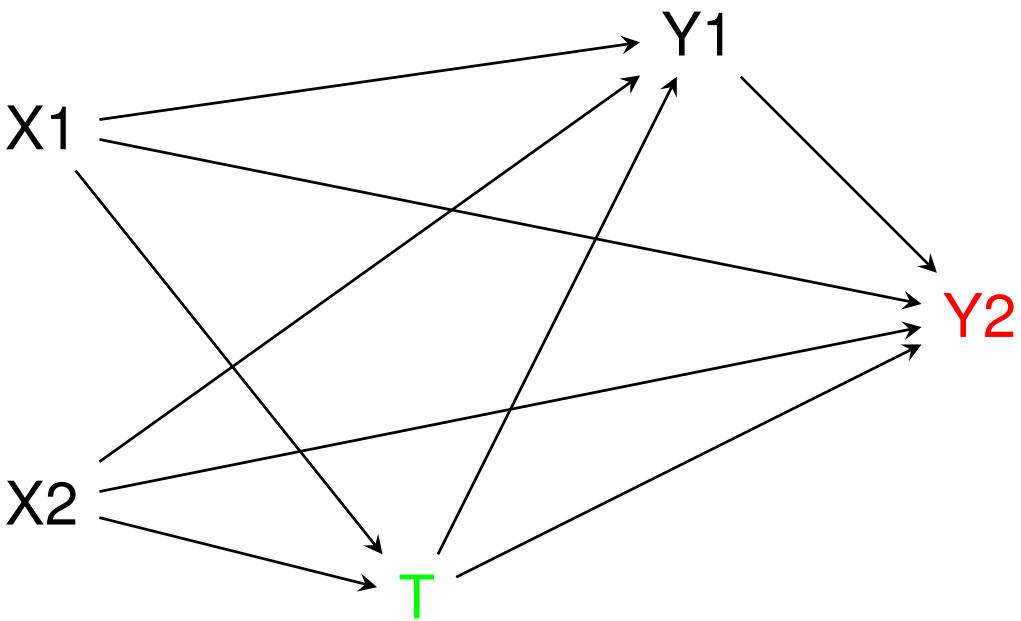


Figure: Causal Model – Directed Acyclic Graph (DAG)

Selection on the unobserved

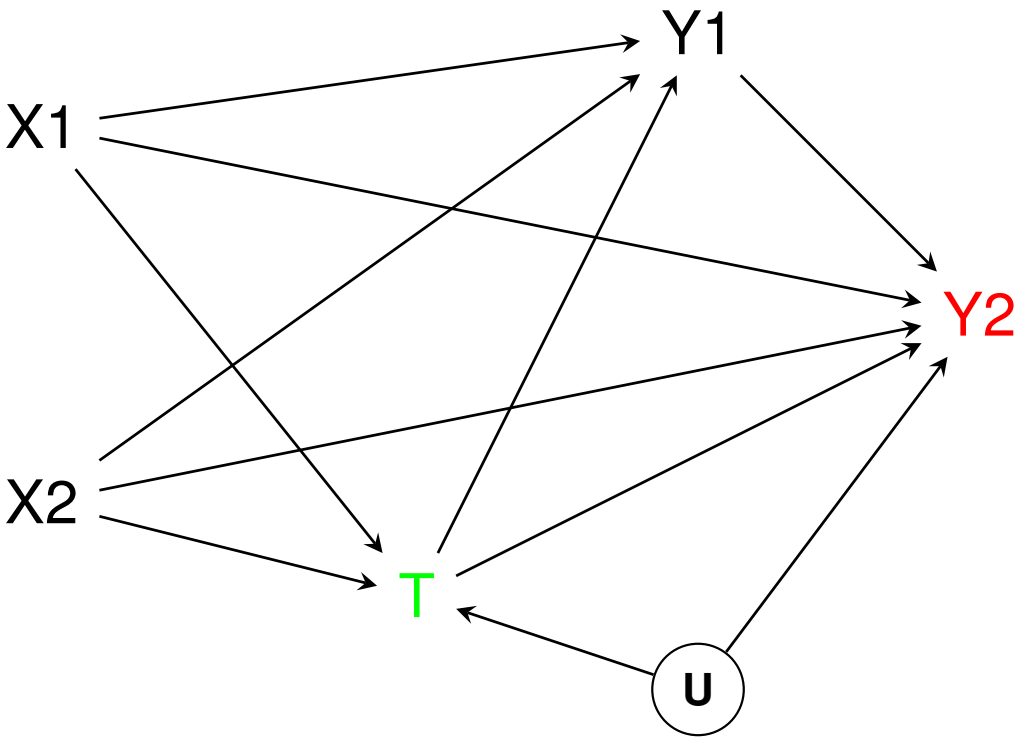


Figure: Causal Model with Latent/Unobserved

Instrumental variable Z

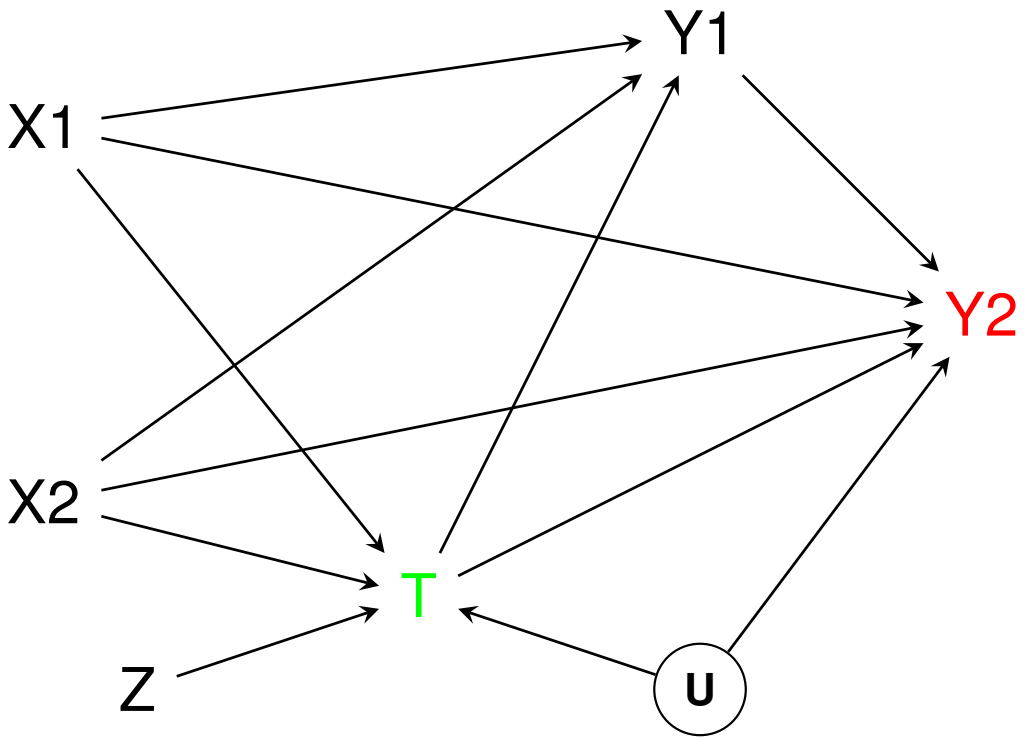


Figure: Instrumental Variable: Z associated with $T|X$ but not $Y|X, T$

Matching/weighting and overlap, common support, causal effect heterogeneity

For the ATE, the weighted regression approaches rely, in at least a practical sense, on the assumption that either we observe all outcomes of T for all values of X (overlap) or that the effect of T on Y is constant across all values of X.

If in the population, $T = 0$ for all cases where $X < 5$, then it would be at best extrapolation to provide any estimate of Y^1 in that sub-group. We would have no data to identify or estimate the value of δ in that sub-group so any such estimate would require the assumption that δ in that group has the same value as δ in some other sub-group (or other similar identifying assumption). That is, that the causal effect is homogenous. In a sample, this lack of overlap may arise from sampling variability as well.

Matching/weighting and overlap, common support, causal effect heterogeneity, pt 2

For example, Morgan & Winship (2015) investigating the effect of (US) Catholic schools on achievement, the distribution of the propensity scores was such that about 25% of the public school sample had estimated propensity scores below the minimum of the Catholic school students.

One approach is to estimate only for the “region of common support” (also omitting <1% of Catholic students with very high propensities). See Heckman et al 1997, 1998. This effect is neither the ATE nor the ATT, sometimes it is called the “common-support ATT.” This raises questions of what the population of inference is.

A worked multinomial treatment example I

This is a teaching example and should not be cited for substantive or policy reasons. Borrowing from Morgan & Winship's analysis (itself borrowing from James Coleman's classic research), I have taken publicly available Australian data from the 2012 Programme for International Student Assessment (PISA). PISA data is a stratified, two-stage sample of approximately 14,500 year-10 Australian students from nearly 800 schools. I have drawn a $1/3$ systematic sample across larger schools (retaining all small schools) and approximately re-weighted.

The research interest is the causal effect of Catholic and Independent school enrolment on student scores in the PISA maths test, conditioned on whether the school is in a Metropolitan area, an index of socioeconomic and cultural status of the parents, the child's immigration status, whether the student comes from a single parent

A worked multinomial treatment example II

family, whether a language other than English is spoken at home, whether the student is of aboriginal or Torres Strait Islander status and the student's gender.

The models

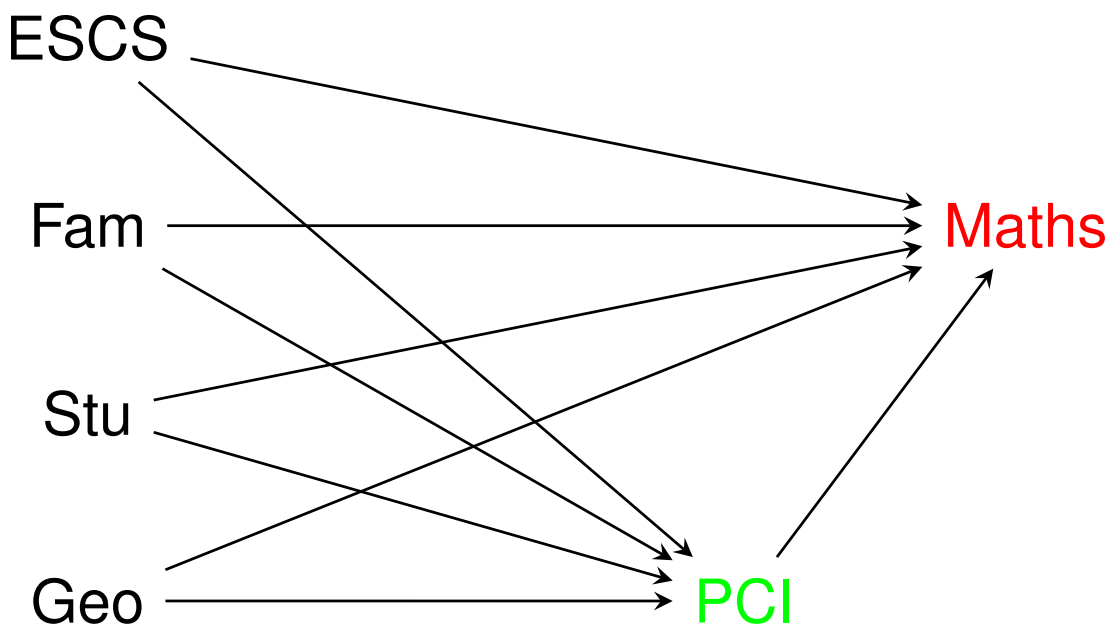


Figure: Model for PISA 2012 example

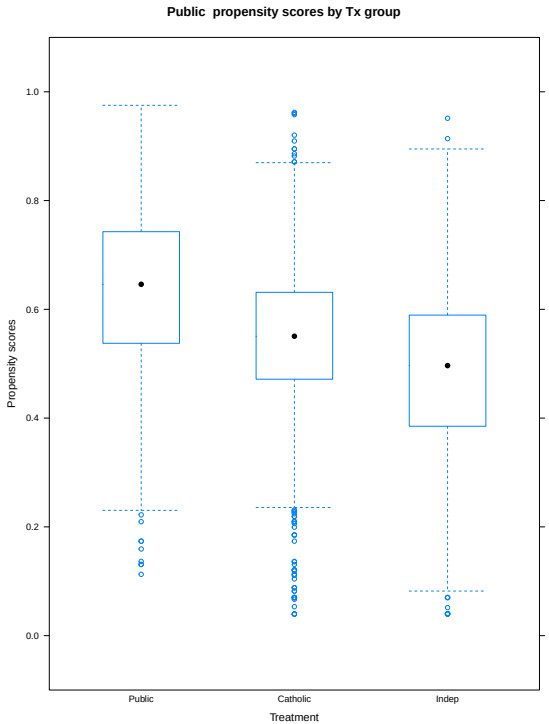
Multiple values of treatment

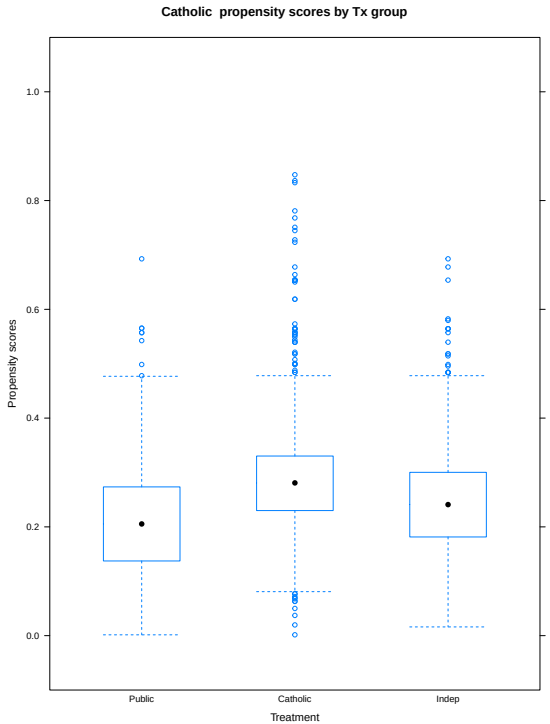
Analysis follows that of McCaffrey et al (2013), implemented via the R package *twang-mnps*.

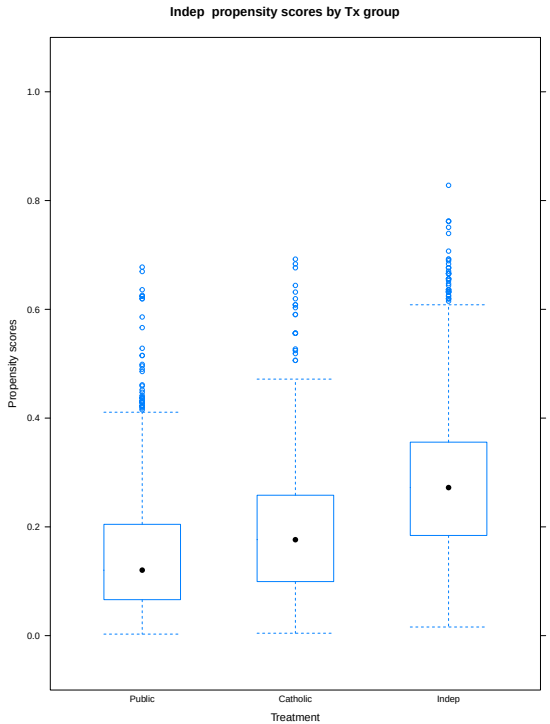
Most causal analysis with a multinomial treatment variable uses multinomial logistic regression for the propensity scoring but McCaffrey et al recommend a generalized boosting model.

As I understand it, these models iteratively fit localized tree regressions then smooth them, to achieve a more optimal balance of the covariates. They cite literature showing that GBM generally performs better than multinomial logistic. GBMs are implemented in twang.

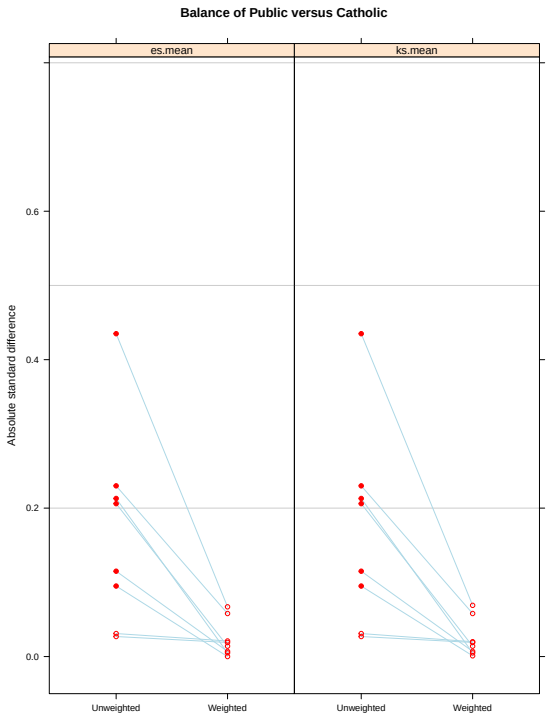
Diagnostics – Propensity Scores

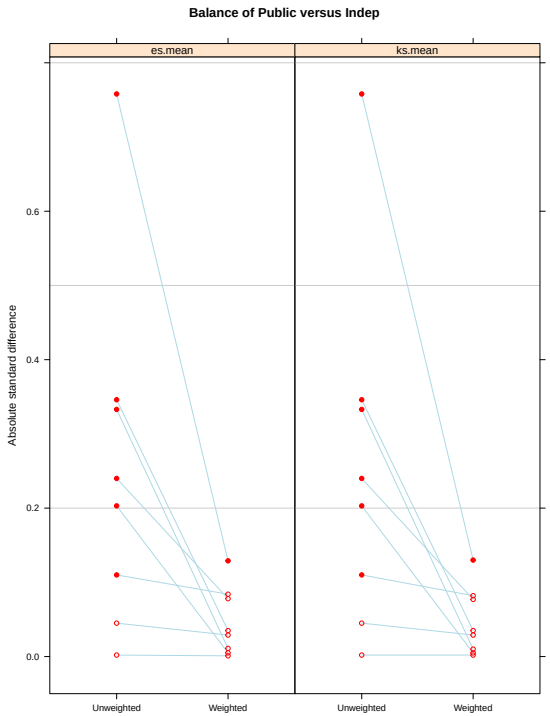


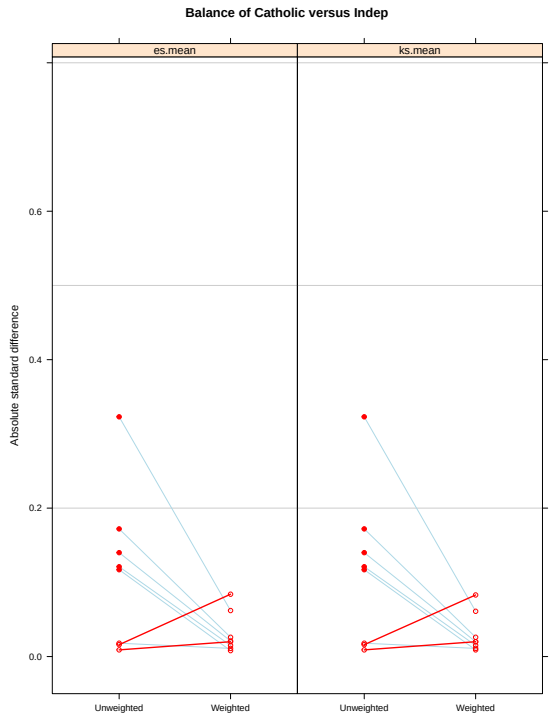




Diagnostics – Balance of Covariates I







Some key results

Variable	Svy Mean	IPTW mean
Public	499	506
Cath	521	515
Indep	541	531

Some key results

Variable	Svy Reg	DR reg	DR reg2
Inter	501	503	504
Cath	8.1	6.7	7.7
Indep	18.9	21.1	21.3
ESCS	38.4	37.4	36.0
Prov_Rem	-8.9	-9.2	-10.6
ATSI	-43.8	-42.2	-41.9
Female	-11.6	-10.1	-10.5

Other references

McCaffrey, DF et al (2013). “A tutorial on propensity score estimation for multiple treatments using generalized boosted models” *Statistics in Medicine*, 32:3388-3414.

Cochran, WG (1968). “The effectiveness of adjustment by subclassification in removing bias in observational studies,” *Biometrics*, 24:295-313.

Rosenbaum, PR and DB Rubin (1984). “Reducing bias in observational studies using subclassification on the propensity score,” *JASA*, 79:516-524.