CAUSAL INFERENCE FOR OBSERVATIONAL DATA OBTAINED FROM A COMPLEX SURVEY

David A. Binder[1]

ABSTRACT

Mary Thompson's 2004 Gold Medal address included a discussion of causation and causality in observational studies. One of Mary Thompson's passions has been on making analytic inferences from surveys with a complex design. Since population-based surveys are observational studies, it seems natural to ask what the impact of the survey design might be for making causal inferences. First, we give a brief review of the literature on causal inference from observational studies. Using Cox and Wermuth's (2004) delineation of various levels of causality, we investigate which assumptions are suitable for the ignorability of the survey design.

KEY WORDS: Levels of causality, Stable unit treatment value assumption, Survey design ignorability

RÉSUMÉ

En 2004, l'allocution de Mary Thompson, la récipiendaire de la médaille d'or incluait une discussion sur la causalité pour des études observationnelles. L'une des passions de Mary Thompson fut de mener des analyses inférentielles d'enquêtes à plan de sondage complexe. Puisque les enquêtes sur la population sont des études observationnelles, il semble naturel de s'interroger sur l'impact du plan de sondage dans les inférences causales. Nous débuterons par une courte revue de littérature de l'inférence causale d'études par observation. En utilisant la délimitation de Cox et Wermuth (2004) pour différents niveaux de causalité, nous étudierons les hypothèses valables permettant d'ignorer le plan de sondage.

MOTS CLÉS : Niveaux de causalité; Hypothèse de la stabilité de valeur de traitement de l'unité; Ignorabilité du plan de sondage.

"Any claim from an observational study is most likely to be wrong." - (Young and Karr – 2011)

1. BACKGROUND

Mary Thompson's illustrious career spans several important topics, including survey methods, estimation for partially observed Markov or semi-Markov models, and inferential and design issues for complex longitudinal surveys. Her Statistical Society of Canada Gold Medal address in 2004 was on "Understanding associations: Implications for the design and analysis of longitudinal surveys". In choosing a topic for this paper in honour of Mary Thompson's retirement, I decided to study the issue of causal inference for data from a complex survey, since this seems to overlap two of the topics about which Mary Thompson is passionate.

We introduce the topic with a short review of the distinction between analytic and descriptive inferences made from complex survey data. We note some peculiarities of causal inferences made from observational data, and discuss the Fundamental Problem of Causal Inference described in Holland (1986). Cox and Wermuth (2004) provide a comprehensive framework for describing various levels of causality. We discuss the impact of survey design on causal inference within this framework.

[1] David A. Binder, 49 Bertona St,. Nepean, ON, K2G 4G7, dbinder49 at hotmail dot com

1.1 Inference from Population-based Surveys

A distinguishing feature of population-based surveys is that the target population is finite. The main objective is to estimate finite population quantities, such as means, totals, ratios, etc. In some instances the sample is also used to estimate analytic quantities, such as the parameters of a model. In this case, it is convenient to assume that the finite population from which the sample was drawn is actually the result of a random realization from a statistical model. However for making causal inferences, since the sample was not obtained via a controlled experiment, the causal effects can be assessed only from the point of view of having observational data, rather than experimental data.

For estimating model-parameters based on data obtained from a survey with a complex design, it is well-known that a design-based approach can lead to valid inferences, for large sample sizes, provided that the first moments of the model are correctly specified, and the sampling fraction is negligible - see Binder and Roberts (2003). Important exceptions do exist, such as estimating the variance components of random effects in a random or fixed effects model. On the other hand, if the higher order moments of the model are not correctly specified, or if there are missing variables in the assumed model, model-based inferences may lead to misleading conclusions.

1.2 Problems with Causal Inferences

Pearl (2009), for example, distinguishes associational assumptions from causal assumptions. Associational assumptions are testable in principle, but causal assumptions cannot be verified (even in principle) unless one resorts to experimental control. Some authors – for example, Rogosa (1997) – would go so far as to say that causal models do not support scientific conclusions. However, in spite of this dilemma, Thompson (2004) pointed out that "careful interpretation of associations can lead to understanding glimpses of causality". Freedman (1999) gives some historical examples where associations have lead to correct and incorrect causality conclusions. He mentions, however, that for some situations, there may be good reasons why controlled experiments are not possible and observational studies are the only choice. These may include ethical considerations.

The *Fundamental Problem of Causal Inference* was described by Holland (1986). This problem is the fact that it is impossible to observe the value of the outcome of interest for all the treatment values.

Here is a simple example of this quandary, given by Rubin (2005). Suppose that in very large randomized experiments the concomitant variable X is the number of plants established in each plot, the primary outcome R is the yield in each plot, the treatment, C, is a new fertilizer ($C=1$), and the control is the standard fertilizer ($C=0$). In each experiment, half of the units are randomly assigned to the active treatment and half of the units are assigned to the control treatment. For each of four plots, we observe X_{obs} and $R_{obs.}$ In Table 1, we give the observed data.

Table 1 – Agricultural Fertilizer Trial - Observed Data

Fraction	Observed Data		
of Sample	C	X_{obs}	R_{obs}
1/ 4	0	2	10
1/ 4	1	3	10
1/ 4	0	3	12
1/ 4	1	4	12

What conclusion might be reached? It turns out that the following linear relationship holds exactly.

$$R_{obs} = 6 - 2C_{obs} + 2X_{obs}. \qquad (1.1)$$

The conclusion looks obvious. For a given number of plants established in each plot, the average effect of the active treatment would be to decrease the yield by 2 units. If only real life were always so simple!

However, the observed data do not tell the whole story. We denote by $X(C)$ and $R(C)$ the potential outcomes for $C=0$ and $C=1$. In Table 2 we give all the outcomes.

Table 2 – Agricultural Fertilizer Trial - Potential Outcomes and Observed Data

Fraction	Potential Outcomes				Observed Data		
of Sample	*X(1)*	*X(0)*	*R(1)*	*R(0)*	*C*	X_{obs}	R_{obs}
1/ 4	3	2	10	10	0	2	10
1/ 4	3	2	10	10	1	3	10
1/ 4	4	3	12	12	0	3	12
1/ 4	4	3	12	12	1	4	12

In Table 2, we see that for *R(0)* and *R(1)*, the yields under the control and the treatment, there is no treatment effect in any of the plots! This directly contradicts to the so-called "obvious" conclusion reached from the observed data in Table 1! The Fundamental Problem of Causal Inference has reared its ugly head.

1.3 Common Assumptions for the Validity of Causal Inferences

In order to make causal inferences in light of the Fundamental Problem of Causal Inference, it is necessary to make some assumptions, which, in general, are unverifiable.

1.3.1 Propensity score matching

Propensity score matching is used to try to create a scenario that can be analyzed using methods similar to methods used for designed experiments, by correcting for the imbalance on the covariates. It is assumed that, conditional on the covariates, the treatment assignment is independent of the outcome of interest. This is sometimes referred to as "no unmeasured confounding".

A common procedure to implement propensity score matching is, first, to run a logistic regression with the response variable as the dependent variable and appropriate conditioning variables as explanatory variables. The propensity score is the predicted probability that the response variable is 1, for each unit. Each participant is matched to one or more nonparticipants on propensity score, using one of a variety of matching techniques, such as nearest neighbour matching. The analysis is then performed using methods appropriate for non-independent matched samples. Some issues about weighting when using propensity score methods are given in Austin (2009).

1.3.2 Rubin Causal Model

Central to Rubin's Causal Model is the *Stable Unit Treatment Value Assumption* (SUTVA). Here, it is assumed that one unit's outcomes are unaffected by another unit's treatment assignment. The conditional probabilities of being assigned to each treatment level are not chosen by the investigators but can be consistently estimated from the data. The treatment levels are not assigned by the investigator but correspond to well-defined interventions.

In my opinion, one needs to consider the validity of this very carefully. For example, would the introduction of a mandatory screening policy for a disease in one province affect the likelihood of being screened for that disease in another province? If so, is it still possible to estimate the effect of implementing the mandatory screening policy?

2. IMPACT OF THE SURVEY DESIGN

2.1 Some Basic Sampling Theory Results

There is very little in the literature where the impact of the survey design on making causal inferences is discussed. Godambe and Thompson (1997) do look at estimation problems for this case. Wang, Scharfstein, Tan, and MacKenzie (2009) consider estimation of the causal effect of a treatment on an outcome from observational data collected in two phases.

To discuss this more formally, we start with a simple case of a Poisson model with loglinear parameters. The population values are N_r^R $(r = 0,1)$, with means μ_r^R, where

$$\log \mu_r^R = \lambda + \lambda_r^R \text{, and } \lambda_0^R + \lambda_1^R = 0 . \qquad (2.1)$$

We denote by D , a 0-1 design variable, and we assume that

$$\log \mu_{rd}^{RD} = \lambda + \lambda_r^R + \lambda_d^D + \lambda_{rd}^{RD} \text{, (with side conditions).} \qquad (2.2)$$

The parameters of interest here correspond to the treatment means; namely, μ_r^R $(r = 0,1)$. Note that this model is not a causal model, but it will be useful in the causal modeling context in Section 2.2.1.

For count data given by ($n, n_r^R, n_d^D, n_{rd}^{RD}$), the loglikelihood function is given by

$$\ell(\boldsymbol{\mu}) = \sum_{r,d} \left[-\mu_{rd}^{RD} \left(\frac{n_d^D}{N_d^D} \right) + n_{rd}^{RD} (\lambda + \lambda_r^R + \lambda_d^D + \lambda_{rd}^{RD}) \right] . \qquad (2.3)$$

The maximum likelihood estimators for μ_r^R are given by

$$\hat{\mu}_r^R = \sum_d \left[\left(\frac{N_d^D}{n_d^D} \right) n_{rd}^{RD} \right] . \qquad (2.4)$$

We note that the factors N_d^D / n_d^D in (2.4) are the usual sampling weights.
If, however, it can be assumed that $\lambda_{rd}^{RD} = 0$ (independence of the design variable and the variable of interest), then the maximum likelihood estimators for μ_r^R are given by

$$\hat{\mu}_r^R = \frac{N}{n} n_r^R , \qquad (2.5)$$

so that the design information can be ignored.

This is a simple case of a more general result that when the sampling design is informative, an approach that ignores the design variables can lead to inappropriate conclusions. In general, the sample design is informative when the distribution of the observed outcomes is different from the distribution for outcomes generated from the model with no additional effect of the design variables (other than those already in the model). If the sampling design is not informative, then the sampling design can be ignored. For more complex situations, see Binder and Roberts (2009).

2.2 Cox and Wermuth's Framework for Causality

Cox and Wermuth (2004) provide a comprehensive framework for inferring causality. In this framework, there are four main types of variables:

- Primary responses (*R*),
- Intermediate variables (*I*),
- Potential causes (*C*), and
- Background variables (*B*).

Ignoring for the moment the intermediate variables, notionally we can regard the variables as ordered *B*, then *C*, then *R*, so that it is natural to represent the distribution for the random variable as

$$f_{RCB} = f_{R|CB} f_{C|B} f_B . \qquad (2.6)$$

When the intervening (causal) variable *C* has no backward effect on *B*, Lauritzen (2000) uses the notation "||" to replace the conditioning sign,

$$f_{R\|C} = \int f_{R|CB} f_B db. \tag{2.7}$$

This is Pearl's (2000) definition of a causal effect. We see that the focus is on how the response variable *R* changes as the causal variable *C* changes, having marginalized over *B*.

Cox and Wermuth (2004) discuss what they refer to as three levels of causality (zero-level, first-level, second-level). For *zero-level causality*, there is a statistical association with clearly established ordering from cause to response, which cannot be removed by conditioning on **allowable** alternative features. An allowable feature must have the attribute of not being affected by the causal variable. For stochastic outcomes on the response variable, this is rare.

We consider the loglinear model given by

$$\log \mu_{rc}^{RC} = \lambda + \lambda_r^R + \lambda_c^C + \lambda_{rc}^{RC}. \tag{2.8}$$

Note that even though this model may appear symmetric in *R* and *C*, we are really assuming that conditional on *C*, the outcomes for *R* are independent and identically distributed, with probability depending on the value of *C*. The marginal model for *C* is assumed to be a Poisson model.

Zero-level causality implies that for **any** allowable intermediate variable *I*, the model that includes *I* must be of the form

$$\log \mu_{rci}^{RCI} = \lambda + \lambda_r^R + \lambda_c^C + \lambda_i^I + \lambda_{rc}^{RC}. \tag{2.9}$$

Because the focus is on the effect that *C* has on *R*, the parameters of interest in this model are λ_{rc}^{RC}. Now, since intermediate variables could include design variables, we see that ZERO-LEVEL CAUSALITY MUST HAVE AN IGNORABLE SAMPLING DESIGN. If, however, the strong assumption of zero-level causality does not hold, then ignoring the design may have negative implications.

The case of *first-level causality* is the one most immediately relevant in many applications. Faced with two or more possible interventions in a system, the aim is to compare the outcomes that would arise under the different interventions. For example, if C_0 and C_1 are two treatments, only one of which can be observed, we would like to compare the outcome observed when C_1 is used to the outcome that would have been observed had C_0 been used, and vice versa. This definition of causality is explicitly comparative. To formalize first-level causality in a simple case, we consider a Poisson model with loglinear parameters. For each individual, there are two possible outcomes, depending on the value of the causal variable. We denote these by R_0 and R_1. When C_0 is applied to a unit, we observe R_0, and when C_1 is applied to a unit, we observe R_1. The model here is given by

$$\log \mu_{r_0,r_1}^{R_0,R_1} = \lambda + \lambda_{r_0}^{R_0} + \lambda_{r_1}^{R_1} + \lambda_{r_0,r_1}^{R_0,R_1}. \tag{2.10}$$

It can be shown that the two-factor interactions cannot be estimated from the available data without further assumptions. A common assumption made is the stable unit treatment value assumption (SUTVA), so that all two-factor interactions are zero. However, other assumptions could also lead to an identifiable model. Shaffer and Chinchilli (2002), for example, assume that $n_{10} = 0$. This could correspond to an assumption such as a placebo leading to a successful outcome implies that the treatment would lead to the same outcome.

2.2.1 Impact of the Survey Design

We now ask what can be surmised if we also consider a model that includes a survey design variable such as *D* used in Section 2.1. First, we note that if the population values can be assumed to have been generated by model (2.10), then under the assumption of no unmeasured confounding, the likelihood for the population counts is equivalent to independent realizations from two univariate distributions, similar to the form given by (2.1). Since we are now in the same framework as the simple non-causal model described in Section (2.1), we see that if the sampling is informative, the design information should be incorporated in the analysis.

Also Godambe and Thompson (1997) showed through an optimal estimating equation approach, that the design weights should be used.

Godambe and Thompson (1997) suggest estimating propensities using sampling weights. When propensity scores are used, they suggest using the doubly-robust weighting method. This is supported by Wang et al. (2009) in their simulations for a case of a two-phase sample design.

We turn now to *second-level causality* as described in Cox and Wermuth (2004). As already stated, to find convincing evidence about the generating process in general requires assembly of evidence of various kinds. Nevertheless an important first step towards level-two causality may often be analysis involving the intermediate variable or variables *I*. These may indicate possible pathways between potential causal variables *C* and the response *R*. Detailed interpretation will have the limitations of observational studies. Even in the simpler discussion of potential causes, it may sometimes be dangerous to disregard *I* totally, for this may indicate some unexpected and in a sense unwanted consequence of the intervention for which some account needs to be taken.

If we suppose that careful design and analysis have established a pattern of dependencies or associations or have provided reasonable evidence of first- or zero-level causality, then the question of explaining how these dependencies or associations arose is often posed. What underlying generating process was involved, i.e. what is underlying the structure observed?

A general concern is the notion of averaging an effect over the distribution of *B*. Cox and Wermuth (2004) state that while this is sometimes convenient, in general the marginalization is a bad idea, notably because it discourages the study of interactions between the causal variables and additional features included in the background variables. Such interactions may be crucial for interpretation. In this case, the appropriate distribution for causal interpretation is $f_{R|CB}$, not $f_{R||CB} = \int f_{R|CB} f_B db$.

However, marginalizing over *B* really deals with the following question: Given a probability distribution over a set of variables (estimated from appropriate data) and given only *C*= *c*, what can be inferred about *R*? When marginalizing over *B* in a complex survey, informative sampling can now play a role, especially if B is not directly observed, or is not included in the model. Marginalizing over *B* would answer the question of what would be the impact on the mean of the response variable over the whole population be if the whole population had been exposed to *C*= *c*.

When a second-level causality model is used to study the effect of the causal variables on the response variables, it would be appropriate to incorporate the design information as in the case of first-level causality, since this offers some protection against informative sample designs.

3. CONCLUSIONS

Making analytic inferences from complex survey data involves assuming a model for the finite population. The Fundamental Problem of Causal Inference implies that unverifiable assumptions are needed when dealing with observational data. These assumptions should not be taken lightly. Cox and Wermuth (2004) provide a useful framework for discussing the impact of survey design on inference. For most typical cases, ignoring the survey design information may lead to misleading conclusions.

ACKNOWLEDGMENTS

I would like to thank Mary Thompson for several helpful remarks.

REFERENCES

Austin, Peter (2009). "The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies". *Statistics in Medicine*, **29**, 2137-2148.

Binder, D.A. and Roberts, G. (2003). "Design-based and model-based methods for estimating model parameters". In: R.L. Chambers and C.J Skinner eds., *Analysis of Survey Data*, , 29-48. Chichester: Wiley.

Binder, D.A. and Roberts, G. (2009). "Design- and model-based inference for model parameters". In: D. Pfeffermann and C. R. Rao, eds., *Handbook of Statistics – Sample Surveys: Inference and Analysis, Volume 29B*., 33–54. The Netherlands: North-Holland.

Cox, D.R. and Wermuth, Nanny (2004). "Causality: a statistical view". *Int. Statist. Rev.*, **72**, 285-305.

Freedman, David (1999). "From association to causation: Some remarks on the history of statistics". *Statistical Science*, **14**, 243-258.

Godambe, V.P. & Thompson, M.E. (1996/97). "Optimal estimation in a causal framework". *J. Ind. Soc. Ag.Statist.*, **49**, 21-46.

Holland, Paul W. (1986). "Statistics and Causal Inference (with discussion)". *J. Am. Statist. Assoc.*, **81**, 945-970.

Lauritzen, S.L. (2000). "Causal inference from graphical models". In: O.E. Barndorff-Nielsen et al., eds., *Complex Stochastic Systems*, 63–107. London: Chapman and Hall.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Pearl, Judea (2009). "Causal inference in statistics: An overview". *Statistics Surveys*, **3**, 96–146.

Rogosa, David (1987). "Casual models do not support scientific conclusions: A comment in support of Freedman." *Journal of Educational Statistics*, **12**, 185-195.

Rubin, Donald B. (2005). "Causal inference using potential outcomes: Design, modeling, decisions". *J. Am. Statist. Assoc.*, **100**, 322-331.

Shaffer , Michele L. and Chinchilli, Vernon M. (2002). "Using Counterfactuals to Account for Treatment Failures in Clinical Trials". *Joint Statistical Meetings - Biometrics Section*, 3173-3178.

Thompson, Mary E. (2004). SSC Gold Medal address. 2004 Annual Meeting of the Statistical Society of Canada, Montréal.

Wang, W, Scharfstein, D., Tan, Z, and MacKenzie, E.J. (2009). "Causal inference in outcome-dependent two-phase sampling designs". *J. R. Statist. Soc. B*, **71**, 947–969.

Young, S. Stanley and Karr, Alan (2011). "Deming, data and observational studies". *Significance*, **8**, 116-120.