A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Tadeusz Bednarski[*]

CAUSALITY ANALYSIS OF SURVEY NONRESPONSE – A COUNTERFACTUAL PERSPECTIVE

Abstract. Declining participation rates in social surveys stimulate research to better understand nonresponse mechanisms and their impact on related statistical inference. In this paper we focus on potential causal relationship between nonresponse and job finding in survey unemployment studies. Selected approaches are discussed from the perspective of counterfactual causality concept.

Key words: survey non-response, counterfactual analysis, causality testing.

I. INTRODUCTION

Declining participation rates in surveys and censuses become more and more a worry of government statistical agencies. The descriptive power of their large data sets, meant to inform about important economic and social processes, declines as well. According to US 2012 Government Wide Managerial Report on Federal Employee Viewpoint Survey the response rate was 46 percent. Of the 82 agencies participating in the survey, only 67 agencies had a response rate of 50 percent or higher. Similar situation concerns other important labor market studies, like the European Union Labor Force Survey or the US Current Population Surveys. Even though there is a vast literature on practices for improving survey participation – a good example could be a book by I. Stoop et al. (2010) where in addition a comprehensive overview of non-response literature is given – the problem persists.

In unemployment duration studies, households or individuals, once included in the sample, may be interviewed periodically over time. They may refuse to cooperate or they may simply be unable to cooperate. So, in addition to initial non-response there can be a high fraction of individuals who drop from the sample during the study period. This raises additional concern about possible bias resulting from non-response. To remedy the situation organizations have implemented respondent incentives, enhanced interviewer training, changes to field

[*] Professor, Chair of Statistics, University of Wroclaw.

procedures, and experimentation with alternate modes of response. On the scientific side we observe an increasing interest in analysis of inferential consequences of potential sampling bias and in understanding the causal mechanism of non-response.

The main body of statistical literature is oriented towards associative inference – relationship between random variables is usually described by convenient features of their joint distribution. Even regression models, particularly aimed at causal relationship between phenomena, cannot in fact go beyond distributional statements, unless some exogenous information or specially designed experiments are given. The causality detection, in a way natural between events ordered in time under properly designed experiments, becomes a much more controversial and speculative task for non experimental data. The other difficulty with causality stems from the fact that its notion is neither unique nor formally strict. We shall adopt here the notion of causality based on counterfactuals which is very intuitive as it mostly relates to unavoidable succession of events, processes or phenomena.

Statisticians try to understand better the real impact of high nonresponse rate on results of data analysis and in particular to single out situation where the missing observations may be of smaller harm. These kinds of studies concentrate in particular on better comprehension and description of causality mechanisms in non-response reactions. In some cases like the unemployment duration studies it is possible to model the causal or missing data mechanism and then analyze its impact on correct inference (Van den Berg et al. (2006), Pyy-Martikainen and Rendtel (2008)).

The aim of this note is to discuss analysis of nonresponse mechanisms from the perspective of counterfactual causality notion. A special attention is devoted to approaches that seem specially promising for modeling of causality nonresponse in the context of observational studies.

The following chapter discusses the definition of causality used in the paper and it gives a brief history of statistical methodology designed for causal studies. Chapter 3 summarizes research dealing with scope and effect of non-response on unemployment time studies. Section 4 discusses a special statistical model aimed to detect the so called non-response causality in the context of counterfactual causality.

II. COUNTERFACTUAL ANALYSIS OF CAUSATION AND STATISTICAL INFERENCE

Counterfactuals refer to conditional statements of the form "If *A* were the case, *B* would be the case". Counterfactual notion of causation in addition require a comparison between what actually happened and what would have happened in the absence of the intervention *A*.

A is a cause of B if and only if a) *event A precedes B and b*) *if A does not occur B does not happen*.

This definition will be our basic reference point in further discussions. Though it would not pass contemporary requirements in a philosophical debate on causality it is a core of any causality notion. Debates on most adequate notions of causality are still going on among philosophers and they probably will never end. Even a brief description of those philosophical studies is beyond the scope of this paper. We also leave it to reader's intuition imagining what is meant by an event or a process, the meaning of causality in population studies and finally refer to omnipresent dilemma between feedback and causality for events in many medical, economic and social studies. Tight connections between causal and counterfactual relations in the probabilistic setup are extensively discussed in J.Pearl (2000, 2009).

Statistical assessment of causality requires however at least a comment on the probabilistic version of the above definition. In experimental reality the causal effect usually varies and is not deterministic. Explaining the merits of causal inference in statistics Holland (1986) refers to Suppes (1970) who expresses causality between temporally ordered events via conditional probabilities of $P(A|B)$ and $P(B)$. He actually relates to other possible causes of B then just A, eliminating so called spurious cases. In further description of statistical methodology related to causality we shall refer to notations and concepts of Neyman, Rubin and Holland, since their description is invariably concerned with practice of statistical inference. The simplest causality assessment with statistical methodology may be described as follows.

Suppose U denotes a population of experimental units. For a randomly selected unit u let $Y(u)$ be the random variable with distribution depending on whether it is in a treatment or control group. Therefore, to each unit u we can in fact assign two random variables $Y_t(u)$ and $Y_c(u)$, of which only one is observed and the other is named a potential outcome. The causal effect for a single unit u is measured by the difference $Y_t(u) - Y_c(u)$ and cannot usually be observed – units assigned to treatment cannot be at the same time in control group. The population causal effect denoted by the expectation $E(Y_t(u) - Y_c(u))$ is not observed as well but it can be estimated. We can sample units from the population, assign them to treatment and control groups by randomization and then use the difference of sample means . To understand how the estimation process is related to proving counterfactual causality we can take for event *A "treatment"* , for *~A "control"* and for *B "change in EY"*. The counterfactual causality condition between events A and B is then satisfied if $\mathrm{E}(Y_t(u) - Y_c(u)) \neq 0$ (*event A precedes B and if A does not occur B does not happen*). Notice that causality can be deduced only if we can consistently estimate $E(Y_t(u) - Y_c(u))$ and the consistency can always be achieved by random assignment of experimental units.

The modern history of statistical methodology based on counterfactual causality starts in 1923 with Spława-Neyman paper published in "Roczniki Nauk Rolniczych i Leśnych" – the paper was translated into the English language for Statistical Science in 1990 by Dąbrowska and Speed. Neyman gives there a probabilistic description of completely randomized design of field experiments. He considers $m = v \times n$ plots on which *v* varieties might be applied and potential yields Y_{ij} corresponding to *ith* variety and *jth* plot. The best estimate of the yield from the *ith* variety would be

$$a_i = \frac{1}{m}\sum\nolimits_{j=1}^{m} Y_{ij}$$

while causal effects resulting from different varieties could be given by the differences $a_i - a_k$. In his reasoning probability appears when we realize that one variety can be assigned to a given number *n* of plots and the assignment should be random. He notices that

$$E\left(\frac{1}{n}\sum\nolimits_{l=1}^{n} X_{i,i_l} - \frac{1}{n}\sum\nolimits_{l=1}^{n} X_{k,k_l}\right) = a_i - a_k$$

if $X_{i,i_1},...,X_{i,i_n}$ and $X_{k,k_1},...,X_{k,k_n}$ are random samples from the population of potential yields $\{Y_{ij}, i = 1,...,v,\ j = 1,...,m\}$. The accuracy of estimation depends on the variance of the above difference of arithmetic means. The higher the number of observations the simpler analytically is the situation. However, a smaller number of observations makes the observations "more" dependent. Neyman thus relates the calculus of probability with statistical modeling of field experiments.

Rubin (1974, 1975, 1978, 1990, 2004) and Imbens G. W. & Rubin D. B. (1997) developed the above setup to complex experimental and observational studies using matching, missing data approach and Bayesian analysis. One should bear in mind that assessment of causality on the basis of non-experimental data becomes much more difficult and always requires some exogenous knowledge or special model assumptions. In experimental studies there is a clear temporal order of events in question and moreover it is the experimenter who evokes the cause *A*. In observational studies the situation is usually not so clear. Take for instance the potential impact of economic university studies on future annual income. We will never be sure to what extent the higher salaries associated with such studies result from acquired knowledge and to what extent they stem from inner managerial talents of candidates interested in economics. Another example of an observational study could be between change in unemployment rate and inflation. Not only it would be hard to imagine experi-

mental study to settle any doubts in understanding relations between the two macroeconomic features but in any case it might be virtually impossible to exclude some sort of feedback between them. Difficulties in statistical analysis of causality, based on observational data in social science, result also from complexity of events of interest and of their recursive character.

III. NON-RESPONSE MECHANISM IN UNEMPLOYMENT STUDIES

Labor force surveys, carried out regularly in many countries constitute a good source of studies of influence of such individual and social characteristics as age, gender, education, work experience or family status on unemployment duration. However, due to high non-response rate, the value of such studies is frequently questionable. Some remedy could be a better understanding of the nonresponse mechanism and a better knowledge of its influence on resulting bias in statistical inference. Identification of the causal non-response mechanisms becomes sometimes possible when the survey information can be combined with individual administrative records held by employment agencies.

An example of advanced study of this kind is given by Van den Berg et al. (2006), where combined survey information with administrative records was used to assess the effect and magnitude of non-response in an unemployment duration study. More precisely the authors try to answer the following questions:

– Are there unobserved personal characteristics affecting both the duration outcome and the attitude towards survey participation?

– Is there a direct causal effect from accepting a job on the probability that a survey interview can take place?

They propose methods to distinguish between two explanations for non-response in survey practice, related to the above questions: selectivity – due to observed and related unobserved determinants of durations of unemployment and a causal effect of job exit. To explain the selectivity they propose, using the framework of the Cox model, to compare baseline hazards for response and non-response parts of the entire sample. To detect the causality effect they examine the hazard rates of exit out of unemployment $\lambda(t|Z,X)$ around $t = c$, where c is the survey time, t is the unemployment duration, Z is the binary non-response indicator and X is a vector of explanatory variables. They argue that under the causal effect the time dependent conditional probability $P(Z = 1|T = t,X)$ has to jump downwards at time $t = c$, while $P(Z = 0|T = t,X)$ has to jump upwards at the same time. It is suggested that statistical verification of causality can be done on the basis of estimation of the hazard rate in a piecewise constant hazard rate model. Their method is however conditioned on a fixed time distance between unemployment entrance and the survey moment.

Another detailed study of survey non-response mechanism for unemployment duration data, where intensive use of register data was made, is given in Pyy-Martikainen and Rendtel (2008). They use the first five waves of the Finnish European Community Household Panel survey data combined at person-level with longitudinal register data. The register data were used as a source of information on unemployment spells and covariates. They study the determinants and impact of initial nonresponse and attrition on the distribution of unemployment spells. Their approach is based on analysis of data missingness mechanism, as it was described in Rubin (1976). The simplest case when we can ignore the process that causes missing data is when the missing data are missing completely at random and the observed data are observed at random. Usually one has to take into account the influence of other factors that affect variables of interest as well as the missingness mechanism. Rubin states general conditions under which ignoring the process that causes missing data always leads to correct inferences. He calls it "missing at random" (MAR). According to Pyy-Martikainen and Rendtel (2008) presence of MAR can be deduced for both the initial non-response and the attrition in the following way:

– If the initial non-response mechanism is MAR, none of the spell covariates should explain the probability of non-response.

– In the attrition model, a MAR non-response mechanism would imply that the spell covariates measured after the last obtained interview should not affect the probability of non-response.

Statistical study of relationships between spell covariates and nonresponse lead to conclusion that the two types of missingness are not MAR but also that initial non-response and attrition are different processes driven by different background variables.

It is important to realize differences of the two approaches in understanding the non-response mechanism. Pyy-Martikainen and Rendtel (2008) concentrate on those aspects of non-response which can be described in terms of ignorable or non-ignorable missingness mechanism. Since this is not directly verifiable they find logical consequences of MAR in their statistical model that are feasible for statistical verification. The influence of explanatory variables on initial non-reponse and attrition is then measured by "associational" reasoning.

Van den Berg et al. (2006) explicitly state possible sources of non-responce and try to model them. They in particular visualize causality non-response as a temporal process in which causal effect of job obtaining results in immediate change of non-response probability. This kind of description is closely related to the more universal definition of causality based on counterfactuals. This point of view will be further elaborated.

IV. MODELING NONRESPONSE CAUSALITY IN THE CONTEXT OF COX MODEL

The Cox model is frequently applied in unemployment duration studies in order to assess influence of explanatory variables on unemployment spell distribution. Statistical relationship between unemployment duration T and the vector of explanatory variables X is described by the Cox proportional hazard model (Cox (1972)) via conditional hazard

$$\lambda(t \mid x) = \lambda_0(t)\exp(\beta' x)$$

where λ_0 is called baseline hazard while β is a vector of regression parameters. The partial likelihood estimator of β solves the following Cox score function equation

$$\int \left[y - \frac{\int x I_{t \geq w} \exp(\beta' x) dF_n(t,x)}{\int I_{t \geq w} \exp(\beta' x) dF_n(t,x)} \right] dF_n(w,y) = 0$$

where F_n is the empirical distribution function of time and covariates. The time censoring variable is suppressed for simplicity of considerations. The other parameter of interest – the cumulated baseline hazard

$$\Lambda(t) = \int_0^t \lambda_0(u) du$$

can be assessed by Breslow (1975) estimator.

As in Van den Berg et al. (2006) we add to our time and covariate data the non-response indicator variable Z giving it a very precise probabilistic meaning. Since causal non-response in particular reflects individual reluctance to survey participation we can define it formally as

$$Z = b_1 I_{C \leq T} + b_2 I_{C > T}$$

where C is a random survey time and T unemployment spell. The variables b_1 and b_2 are independent Bernoulli variables with success probabilities $p_1=p$ and $p_2=p+\varepsilon$ respectively, describing nonresponse chance with p depending possibly on C and explanatory variables X. Notice that if survey time is given and we look at Z as a function of T then the chance of Z being 1 changes (if ε is not equal to 0) at the moment of unemployment termination. This is a natural model

of causal relationship between two events: “finding job” and “changing attitude towards participation in survey”, where “changing attitude” means change in chance of participation decision. This is consistent with counterfactual definition of causality.

Bednarski and Borowicz (2010) suggested verifying significance of Z within the standard inference process for the Cox model to test causality of job finding and survey non-response. Bednarski (2013) gave a precise formal justification and meaning of the testing procedure. It was expressed by the following fact:

Suppose β is the true value of the regression parameter in the Cox regression model. Then the following expression, corresponding to the Cox score function,

$$\int \left[\bar{z} - \frac{\int z I_{t \geq w} \exp(\beta_0 x + \beta' x) dF(t,x,x)}{\int I_{t \geq w} \exp(\beta_0 x + \beta' x) dF(t,z,x)} \right] dF(w,\bar{z},y)$$

where $F(t,z,x)$ denotes the joint distribution of time to exit from unemployment, z the non-response variable and x the covariates, is equal to zero at $\beta_0 = 0$ if and only if $\varepsilon = 0$.

It F is replaced by sample cumulative distribution F_n in the above equation, then depending on whether non-response causality is absent or not, the estimator of β_0 will tend to 0 or to a quantity different from 0. The statistical variability of the estimator described by the Gaussian law leads then to the test. The resulting method, though designed for very specific application constitutes a promising way to verify causal relationship between events for observational data.

REFERENCES

Bednarski T., Borowicz F. (2010) Analysis of non-response causality in labor market surveys.Acta Universitatis Lodziensis. Folia Oeconomica 253, s. 217-224, Lodz 2010.

Bednarski T. (2013). On robust causality nonresponse testing in duration studies under the Cox Model. Statistical Papers. DOI 10.1007/s00362-013-0523-0.

Breslow, N. E. (1975) Analysis of Survival Data under the Proportional Hazards Model". International Statistical Review 43, 45-58.

Cox, R.D. (1972) Regression model and life tables. J. Roy. Statist. Soc. Ser. B 34, 187-220.

Imbens G. W. and Rubin D. B. (1997). Bayesian Inference for Causal Effects in Randomized Experiments. The Annals of Statistics. Vol. 25, No. 1, 305-327.

Holland P.W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association. Theory and Methods. Vol. 81, No 396, 945-960.

Neyman J. 1923 (1990). “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” *Statistical Science* 5 (4): 465-472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Pearl J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl J. (2009). Causal inference in statistics: An overview. Statistics Survey. Vol 3. 96-146.

Pyy-Martikainen, M. and Rendtel, U. (2008) Assessing the impact of initial nonresponseand attrition in the analysis of unemployment duration with panel surveys. Advances In Statistical Analysis, Vol. 92, 297-318.

Spława-Neyman J. (1923). Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych. „Roczniki Nauk Rolniczych i Leśnych" t. 10, s. 1-51.

Stoop I., Billiet J., Koch A. and Fitzgerald R. (2010) Improvind Survey Response, Lessons Learned from European Social Survey. Wiley.

Rubin D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. Vol. 66, No. 5 688-701.

Rubin D.B. (1975). Bayesian inference for causality: the importance of randomization. Proc. Social Statistics Section. Am. Statist. Assoc. p. 233-239.

Rubin D. (1976). Inference and missing data. Biometrika 63, 581-592.

Rubin, D.B. (1978). Bayesian inference for causal effects. The role of randomization. Ann. Statist. 7, 34-58.

Rubin, D.B. (1990). Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. Statistical Science. Vol.5 No. 4, 472-480.

Rubin, D.B. (2004). Direct and Indirect Causal Effects via Potential Outcomes. Scandinavian Journal of Statistics 31, 161-170.

Suppes P.C. (1970). A probabilistic Theory of Causality. Amsterdam, North Holland.

Van den Berg, G. J., Lindeboom M. M., Dolton P. (2006). Survey nonresponse and the duration of unemployment, Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 169, Number 3, 585-604, (2006).

Tadeusz Bednarski

ANALIZA PRZYCZYNOWOŚCI ODMOWY W BADANIACH SONDAŻOWYCH W PERSPEKTYWIE KONTRFAKTYCZNEJ

Warunek nieobciążoności próby w statystycznych badaniach społecznych praktycznie nigdy nie jest spełniony, a w sondażowych badaniach rynku pracy poziom odmowy udziału niejednokrotnie przekracza 40%. Dokładność wniosków statystycznych w takich sytuacjach może być poprawiona lepszym zrozumieniem mechanizmu odmowy. Szczególnie niekorzystną sytuacją w statystycznej analizie danych bezrobocia jest zależność pomiędzy czasem poszukiwania pracy i odmową udziału. W pracy omawia się metodę weryfikacji takiego mechanizmu odmowy w perspektywie kontrfaktycznej analizy przyczynowości.