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 3(302), 2014

Tadeusz Bednarski[*], *Filip Borowicz*[**]

ON NONRESPONSE CAUSALITY TESTING IN ROTATING PANEL DESIGNS UNDER THE COX MODEL

Abstract. High survey nonresponse in unemployment duration studies may have a strong effect on inference if exit from unemployment affects the chance of nonresponse (nonresponse causality). In rotational studies large part of the nonresponse results from panel attritions. A method to test the presence of the causality mechanism for rotating panel designs is proposed and its asymptotic consistency is proved under the Cox regression model. An application to real labor data and a simulation study are shown.

Keywords: biased sample, testing non-response causality, unemployment data analysis

I. INTRODUCTION

The inference for the Cox regression model, commonly applied to duration data in social studies, may be sensitive to sample bias. The nonresponse may be specially destructive if event defining the time variable affects the nonresponse probability. The mechanism is then called causal.

The high nonresponse rate is in particular present in labor force surveys (LFS), carried out regularly in the EU countries. The inferential value of these surveys could be probably improved if the mechanisms of nonresponse were better known. Unfortunately, such detailed studies are seldom possible, in fact, only when a combination of survey and administrative records is available.

Extensive longitudinal studies of survey nonresponse and attrition are given e.g. in Romeo (1997), O'Muircheartaigh and Campanelli (1999), Little and Rubin (2002), van den Berg (1994) and Groves (2006). Pyy-Martikainen and Rendtel (2008) show how register data combined at person-level with survey data can be used to conduct an extended type of nonresponse analysis in panel surveys. They

[*] Full Professor, Institute of Economic Sciences, Wrocław University, Poland.
[**] M.Sc., Institute of Economic Sciences, Wrocław University, Poland.

demonstrate that initial nonresponse and attrition mechanisms are nonignorable with respect to analysis of unemployment spells. An important review of longitudinal methods in economics for labor market data can also be found in Hackman and Singer (1985).

A causality testing method proposed in Bednarski and Borowicz (2010) and statistically elaborated in Bednarski (2013) was intended for studying the initial nonresponse mechanism in unemployment data. Here, the method is specialized to attrition in rotating panel designs. Data for such studies are "easier" to obtain, since unemployment status of individuals missing in one panel can be frequently retrieved in later panels while studies of causality in initial nonresponse require combination of survey and registered data information.

To explain the causality effect (van den Berg et al., (2006)) examine the hazard rates of exit out of unemployment $\lambda(t \mid Z, X)$ around $t = c$, where c is the survey time, t is the unemployment duration, Z is the binary nonresponse indicator, X is a vector of explanatory variables. They argue that under the causal effect the time dependent conditional probability $P(Z = 1 \mid T = t, X)$ has to jump downwards at time $t = c$, while $P(Z = 0 \mid T = t, X)$ has to jump upwards at the same time. The suggested empirical application of this heuristic method was based on a piecewise constant hazard rate model. The method is then limited to large sample sizes and it requires a fixed time distance between unemployment entrance and the survey moment. Moreover, as shown in Bednarski (2013), it is several times less efficient compared to a proposal in Bednarski and Borowicz (2010).

As in Bednarski and Borowicz (2010) the method proposed here is derived under the supposition that population distribution follows approximately the Cox regression model. It consists in adding to the set of explanatory variables the indicator variable Z

$$Z = \begin{cases} 1 & \text{nonresponse at time } D \\ 0 & \text{response at time } D \end{cases}$$

and studying its significance using the partial likelihood estimation. The variable D, assumed independent of T, represents a random instant of time between the moment of inflow into unemployment and the moment of the survey date (for the attrition studies one of randomly chosen moments in the rotating panel design). It was shown that the regression coefficient corresponding to Z is zero if, and only if, there is no causality effect. Therefore, standard statistical packages can be used to test the effect. It is further assumed that we have complete information on sample variables and for each individual we know the value of Z. Censoring determined by the time of study termination will be allowed.

II. THE MODEL AND TESTING METHOD

The formal relationship between unemployment duration T and the vector of explanatory variables X is further described by the Cox proportional hazards model (Cox (1972)) with conditional hazard $\lambda(t \mid x) = \lambda_o(t)exp(x'\beta)$, where λ_o is the baseline hazard and β is a vector of regression parameters. The partial likelihood estimator of β is then the solution of the score function equation $L_{F_n}(\beta) = 0$, where

$$L_{F_n}(\beta) = \int \left[y - \frac{\int x I_{t \wedge c \geq w} \exp(\beta' x) dF_n(t,c,x)}{\int I_{t \wedge c \geq w} \exp(\beta' x) dF_n(t,c,x)} \right] I_{(w \leq \bar{c})} dF_n(w, \bar{c}, y),$$

F_n is the empirical distribution of time, censoring and covariate variables. The censoring variable in the inner integrals is denoted by c and it is denoted by $\bar{c}$ in the outer integral. The cumulated baseline hazard $\Lambda(t) = \int_0^t \lambda_0(u)du$ is usually estimated by the Breslow estimator (Breslow (1975))

$$\hat{\Lambda}(t) = \sum_{i:T_i \leq t} \frac{1}{\sum_{j \in R(T_i)} \exp(\hat{\beta}' X_j)},$$

where T_i, X_i are sample observations, $\hat{\beta}$ is the partial likelihood estimator and $R(T_i)$ is the risk set at time T_i. The risk set denotes all individuals who are at risk at time T_i – the individuals unemployed at time T_i. Censoring time variable is assumed independent of T given the values of covariates X.

The causal nonresponse means here dependence of Z on event $(T < D)$ only if covariate X and survey time D independent of T are given. It is formally described by the formula

$$Z = b_1 I_{D \leq T} + b_2 I_{D > T}, \tag{1}$$

where b_1 and b_2 are Bernoulli variables, independent of T and C when X is given, with success probabilities $p_1 = p$ and $p_2 = p + \gamma$ respectively, with p and γ depending possibly on X. Lack of causality is equivalent to identical

success probabilities for the two Bernoulli random variables $(\gamma = 0)$ at any fixed values of D and X.

As mentioned earlier, the proposed testing method is very simple in use. It consists in including Z into the list of explanatory variables and performing a standard inference using the Cox regression model. The following theorem justifies the consistency of the testing method. Its analogue for uncensored time variable in connection with initial nonresponse analysis was proved in Bednarski (2013). The proof is given in the Appendix.

Theorem 1 *Suppose that the support of survey distribution is contained in the support of unemployment time distribution. Assume also that the binomial probabilities have the form* $p_1 = p(x)$, $p_2(x) = p(x) + \varepsilon\gamma(x)$ *where* $\gamma(x)$ *is strictly positive with* $\varepsilon \geq 0$, β *is the true regression parameter in the Cox regression model, the time variable* T *and censoring* C *given* X *are independent, while nonresponse* Z *is given in (1). Then the following expression*

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

corresponding to the Cox score function, where $F(t, c, z, x)$ *denotes the joint distribution of time to exit from unemployment, censoring time, the nonresponse variable* z *and covariates* x *is equal to* 0 *at* $\beta_0 = 0$ *if and only if* $\varepsilon = 0$.

The nonzero value of $L_F(\beta_0, \beta)$ for $\beta_0 = 0$ and $\varepsilon \neq 0$ implies that the equation consistently detects causality. When F is replaced by the empirical distribution function then a test based on the partial likelihood estimator of β_0 lets us verify the hypotheses H_0: non-causality versus H_1: causality. A standard argumentation can be used to show that the distribution of $\sqrt{n}(\hat{\beta}_0 - \beta_0)$ is approximately Gaussian with mean zero and estimable standard deviation under the null hypothesis. The important feature of the testing method, which follows from the theorem's assumption on the supports of survey and time distributions, is that quite arbitrary survey time designs are allowed. In particular D can be randomly chosen from periodically scheduled survey times as it will be done for the real unemployment data in the following chapter.

III. DATA DESCRIPTION AND ANALYSIS

The unemployment data used in this study cover the period from February 2006 to December 2009 for the Lower Silesia Province in Poland. They constitute a part of a large - scale LFS study carried out by the Polish Central Statistical Office since 1992, gradually modified according to Eurostat recommendations. The LFS methodology is based on the definitions of the economically active population. People aged 15 or more are perceived either as economically active or inactive (outside the labour force).

The population of economically active is observed through households. Each sample amounts to about 54 thousand dwellings for the entire country. Samples are selected quarterly according to a rotation system (rotating panel design) – each sample is employed two quarters in the survey, two quarters break and again two quarters in the survey and then out. Each individual's survey questionnaire contains over 100 questions with lists of possible answers (altogether over 450 !). Though the number and order of questions depend on the individual's status, questions on demographic, social and educational issues apply to all individuals in the sample.

One of important objectives of the large scale LFS, apart from provision of basic labor statistics, is to determine the effect of explanatory variables on the distribution of job search time. It is well known that LFS have high initial nonresponse. In rotating panel designs, where we can longer trace individual's economic status, there might be additional loss of information and inferential bias due to attrition.

The presented method of causality (in attrition) testing required information on nonresponse status Z and unemployment time T for each individual. Due to required answering pattern the unemployment time was determined in months. The idea then was to use the information on nonresponse provided by the four interviews for all individuals. To determine the values of Z the interviews were selected randomly and independently of T in the existing sample. A special care had to be taken in censored cases – under the causality hypothesis the unemployed individuals absent in a series of consecutive interviews were not legally censored. To reduce the number of such cases the last interview was excluded for determination of Z.

In our data analysis the samples with complete 6 quarters observation period were included, giving the initial set of 8325 individuals from 3524 households. Only 780 individuals out of the initial set declared unemployment in at least one of the 4 interviews. Due to missing or inconsistent information the final study sample was composed of 573 individuals.

Table 1 shows results of estimation under the Cox regression model. The number of explanatory variables was reduced by the Akaike method from 25 to 12.

The last row of the Table shows result of testing causality in attrition. With *p*-value 0.81 we are quite confident that the attrition nonresponse is unrelated or very weakly related to exit moment from unemployment.

Table 1. Results of estimation for LFS Silesian data using the Cox regression model

Variable	coef	exp(coef)	*p*-value
Age	–0.0339	0.967	0.00018
general secondary education	–0.4675	0.627	0.07900
primary education	–0.3184	0.727	0.14000
relationship to householder- partnership	–0.3139	0.731	0.07500
relationship to householder- grand(parents/children)	–1.8422	0.158	0.06900
Married	–0.5012	0.6060	0.00920
professional practice more than 5 years	0.3913	1.479	0.02500
registered in labor office	–0.3606	0.697	0.02700
compensation for unemployed	0.4312	1.539	0.02200
number of people in household > 3	–0.2634	0.768	0.14000
household's main livelihood – retirement, pensions	–0.6891	0.502	0.00031
household's main livelihood – other sources	–1.1400	0.320	0.00084
nonresponse indicator *Z*	0.0378	1.039	0.81000

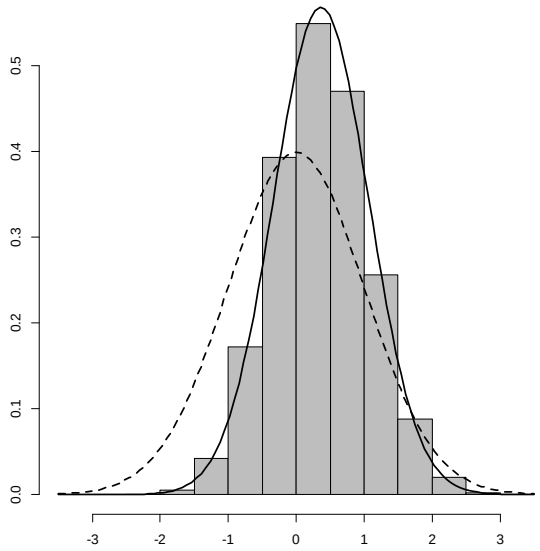

Figure 1. Histogram of standardized values of test statistics for causality verification (1000 runs). Dashed line – standard normal density

The causality test in Table 1 was determined in a single post factum random interview choice, generated for each individual by the Monte Carlo method. To have an idea what is the distributional effect of many post factum randomizations, the simulations were repeated 1000 times. It was found that the frequency of noncausality acceptance was 0.98. Figure 1 shows the empirical distribution of the standardized test statistic for this simulation and compares it with the standard normal distribution. We can see a smaller variability of the test statistic compared to the standard normal distribution. The difference can be explained by the fact that the simulations were done conditionally on the observed unemployment times and covariates.

To understand better the discrepancy between the empirical distribution of standardized test statistics and the standard normal distribution a small Monte Carlo study was designed. In the first step a sample of size 500 was taken from the Cox regression model with $\lambda(t,x) = exp(-x_1 - x_2 - 0.5x_3)$, where x_i were independent Bernoulli 0.5 success probability random variables. Then, for each "individual", the first interview time, say w_1, was independently generated from the uniform $[0,1]$ distribution. To imitate the rotating panel design the survey times w_2, w_3, w_4 were respectively $w_i = w_{i-1} + 1$ for $i = 2,3,4$. The variable w_4 imitating termination of the study was the censoring time. The probabilities p_1, p_2 defining the nonresponse indicator Z were equal 0.5 for each selected survey moment. Given the sample from the Cox regression model the variables w_i and Z_i were generated 1000 times and in each case the standardized test statistic for causality was computed.

In the second step the sampling process from the Cox model was repeated 1000 times and for every sample random variables w_i and Z were generated once for each individual. The testing procedure was applied to determine the value of test statistic in each sample case. Figure 2 depicts density histograms of the test statistics for the first and the second simulation case. It explains smaller variability of the empirical distributions when the only variability source comes from random survey time choice.

To check effects of time discretization the simulations were repeated for unemployment times conveniently rounded to imitate job search given in months. The simulation results were basically unchanged. Also different values were given to the nonresponse probabilities. The results in terms of differences in variability were roughly the same.

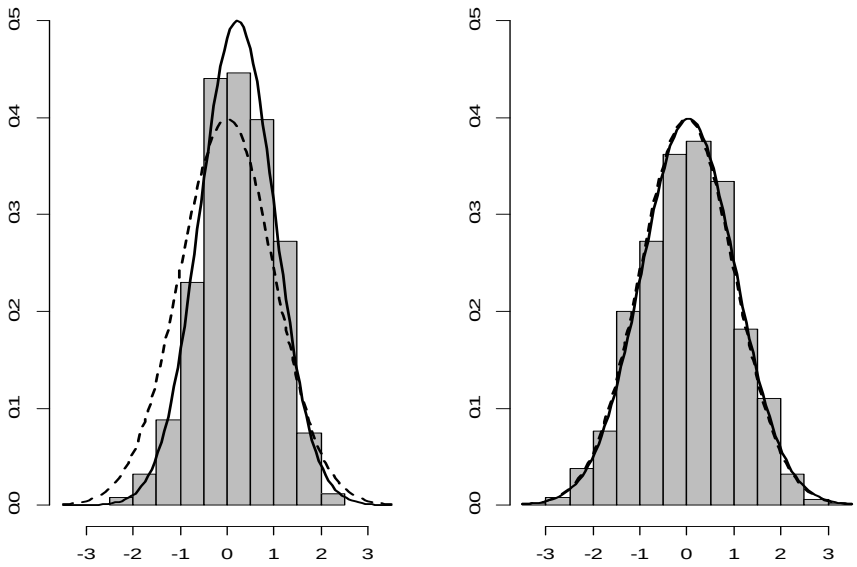

Figure 2. Density histograms of standardized values of test statistics for causality verification from simulation experiment. Dashed line – standard normal density

IV. CONCLUSIONS

The dependence of job finding on nonresponse in unemployment duration studies may result in highly biased inference when the Cox regression model is applied. The aim of the paper was to verify a causal relationship between exit from unemployment and attrition for the LFS in Lower Silesia Province in Poland. The data were a part of a large- scale rotating panel study carried out by the Polish Central Statistical Office since 1992. For most of the uncensored individuals in the sample, including the cases of attrition, it was possible to determine their unemployment time by combining the history in rotating panel questionnaires. A testing method of Bednarski and Borowicz (2010) and of Bednarski (2013) based on the Cox model was then adapted to verify the causality mechanism in attrition nonresponse. The noncausality hypothesis was not rejected and we concluded that the attrition has a minor influence (if any) on statistical inference based on the Cox model for this particular LFS study. Supplementary Monte Carlo experiments were given to better justify the conclusions.

V. APPENDIX – PROOF OF THEOREM 1

Let $F(t,c,z,x)$ denote the distribution of unemployment time T, censoring C, nonresponse variable Z and covariates X. By independence of T and C given the covariates X and independence of Z and the censoring variable C given X we can write

$$dF(t,c,z,x) = dF_z(z \mid t,x)dF(t \mid x)dF_c(c \mid x)dG(x)$$

where $F_z(z \mid t,x)$ is the conditional distribution of Z given T and X, $F(t \mid x)$ denotes the distribution of T given X, $F_c(c \mid x)$ is the distribution of censoring time C given X while G is the marginal of X. If S denotes the distribution of survey time D, then at fixed value of X, $P(Z=1,T=t \mid X=x) = $ $= [p(x)+\varepsilon\gamma(x)(1-S(t))]f(t \mid x)$, where $f(t \mid x)$ is the conditional density for T given X, and consequently

$$P(Z=1,T=t) = \int [p(x)+\varepsilon\gamma(x)(1-S(t))]f(t \mid x)dG(x).$$

To show that $L_F(\beta_0,\beta)=0$ when $\varepsilon=0$, β is the true parameter value and $\beta_0 = 0$ we can argue as in Bednarski (2013).

For the other part of the proof it is shown that $L_F(0,\beta)$ is strictly increasing with respect to ε, which implies that $\beta_0 = 0$ cannot be a solution of $L_F(\beta_0,\beta)=0$ if $\varepsilon > 0$. To verify the monotonnicity we compute the derivative of $L_F(0,\beta)$ with respect to ε.

Since

$$L_F(0,\beta) = \int I_{w\le\bar{c}}[p(y)+\varepsilon\gamma(y)(1-S(w))]f(w \mid y)dF_c(\bar{c} \mid y))dG(y)dw$$

$$-\int \frac{\int I_{w\le t\wedge c}[p(x)+\varepsilon\gamma(x)(1-S(t))]e^{\beta' x}f(t \mid x)dF_c(c \mid x)dG(x)dt}{\int I_{w\le t\wedge c}e^{\beta' x}f(t \mid x)dF_c(c \mid x)dG(x)dt} \cdot$$

$$\cdot f(w \mid y)dF_c(\bar{c} \mid y)dG(y)dw$$

is linear in ε, integration with respect to censoring variable followed by differentiation gives

$$\frac{\partial L_F(0,\beta)}{\partial \varepsilon} = \int (1-F_c(w\,|\,y))[\gamma(y)(1-S(w))]\,f(w\,|\,y)dG(y)dw$$

$$-\int (1-F_c(w\,|\,y))\frac{\int I_{w\le t}(1-F_c(w\,|\,x))[\gamma(x)(1-S(t))]e^{\beta' x} f(t\,|\,x)dG(x)dt}{\int I_{w\le t}(1-F_c(w\,|\,x))e^{\beta' x} f(t\,|\,x)dG(x)dt}\cdot$$

$$\cdot\, f(w\,|\,y)dG(y)dw$$

Therefore, the replacement of $S(t)$ by $S(w)$ in the second summand leads to

$$\frac{\partial L_F(0,\beta)}{\partial \varepsilon} > \int (1-F_c(w\,|\,y))[\gamma(y)(1-S(w))]\,f(w\,|\,y)dG(y)dw$$

$$-\int (1-F_c(w\,|\,y))\frac{\int I_{w\le t}(1-F_c(w\,|\,x))[\gamma(x)(1-S(w))]e^{\beta' x} f(t\,|\,x)dG(x)dt}{\int I_{w\le t}(1-F_c(w\,|\,x))e^{\beta' x} f(t\,|\,x)dG(x)dt}\cdot$$

$$\cdot\, f(w\,|\,y)dG(y)dw$$

with the right-hand side of the inequality equal to

$$\int (1-F_c(w\,|\,y))[\gamma(y)(1-S(w))]\,f(w\,|\,y)dG(y)dw$$

$$-\int (1-F_c(w\,|\,y))\frac{\int (1-F(w\,|\,x))(1-F_c(w\,|\,x))[\gamma(x)(1-S(w))]e^{\beta' x} dG(x)}{\int (1-F(w\,|\,x))(1-F_c(w\,|\,x))e^{\beta' x} dG(x)}\cdot$$

$$\cdot\, f(w\,|\,y)dG(y)dw$$

$$= \int (1-F_c(w\,|\,y))[\gamma(y)(1-S(w))]\,f(w\,|\,y)dG(y)dw$$

$$-\int (1-F_c(w\,|\,y))\frac{\int (1-F_c(w\,|\,x))[\gamma(x)(1-S(w))]\,f(w\,|\,x)dG(x)}{\int (1-F_c(w\,|\,x))\,f(w\,|\,x)dG(x)}\cdot$$

$$\cdot\, f(w\,|\,y)dG(y)dw$$

$$= \int (1-F_c(w\,|\,y))[\gamma(y)(1-S(w))]\,f(w\,|\,y)dG(y)dw$$

$$-\int\int (1-F_c(w\,|\,x))[\gamma(x)(1-S(w))]\,f(w\,|\,x)dG(x)dw = 0$$

REFERENCES

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

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

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.

Groves R. (2006), *Nonresponse Rates and Nonresponse Bias in Household*, Surveys Public Opinion Quarterly, Special Issue 70, 646–67.

Heckman J. J., Singer B. S. (1985), *Longitudinal Analysis of Labor Market Data*, Econometric Society Monographs 10, Cambridge University Press.

Little R. J. A., Rubin D. B. (2002), *Statistical Analysis with Missing Data*, 2nd ed, New York, Wiley.

O'Muircheartaigh C., Campanelli P.A. (1999), *Multilevel exploration of the role of interviewers in survey nonresponse,* Journal of the Royal Statistical Society: Series A (Statistics in Society), 162, 437–446.

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

Romeo C. J. (1997), *Measuring information loss due to inconsistencies in duration data from longitudinal surveys,* Journal of Econometrics 78, 159–177.

Van den Berg G. J., Lindeboom M., Ridder G. (1994). *Attrition in longitudinal panel data and the empirical analysis of dynamic labour market behavior,* J. Appl. Econometr. 9: 421–435.

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

Tadeusz Bednarski, *Filip Borowicz*

TESTOWANIE PRZYCZYNOWOŚCI UBYTKU RESPONDENTÓW DLA PANELI ROTACYJNYCH W PRZYPADKU MODELU COXA

Absencja respondentów w ankietowych badaniach rynku pracy może znacząco wpływać na obciążenie estymatorów rozkładu czasu poszukiwania pracy przez osoby bezrobotne, w przypadku gdy znalezienie pracy przez ankietowanego wpływa na szansę odmowy udziału w badaniu (efekt przyczynowy absencji). Rotacyjne panelowe badania ankietowe, takie jak na przykład BAEL, są dodatkowo narażone na „wyczerpywanie się" danych, ponieważ część wylosowanych jednostek rezygnuje w trakcie trwania cyklu badawczego. W pracy proponuje się metodę testowania efektu przyczynowego związanego z „wyczerpywaniem się" danych. Przedstawia się także zastosowanie zaproponowanej metody dla danych BAEL oraz uzupełnia się wyniki empiryczne symulacjami. Istotnym wnioskiem badań jest stwierdzenie braku przyczynowości związanej z wyczerpywaniem się danych w Badaniach Aktywności Ekonomicznej Ludności prowadzonych przez GUS.