Stat Papers (2014) 55:221–231
DOI 10.1007/s00362-013-0523-0

REGULAR ARTICLE

On robust causality nonresponse testing in duration studies under the Cox model

Tadeusz Bednarski

Received: 26 March 2012 / Revised: 2 April 2013 / Published online: 4 May 2013

Abstract High survey nonresponse in unemployment duration studies may have a strong effect on inference if the so called causal mechanism is present. A robust method of testing the causal nonresponse is proposed for data sets where survey information can be combined with complete administrative records. It is assumed that population distribution follows approximately the Cox regression model. Formal justification of the method and a comparative simulation study are included.

Keywords Robust testing and estimation · Cox model · Biased sample · Testing non-response causality · Unemployment data

Mathematics Subject Classification (2000) 62F03 · 62P25 · 62D05 · 62N01

1 Introduction

The Cox model is commonly applied in duration studies where a high nonresponse rate may result in excessive estimation and testing bias. In social time to event studies the nonresponse reflects human reluctance which if it is related to order of appearance of moments of survey and the event in question, like finding a job, is called causal. 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 et al. (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 a panel survey. They demonstrate

T. Bednarski (✉)
Institute of Economic Sciences, Wroclaw University,
ul. Uniwersytecka 22/26, 50-145 Wroclaw, Poland
e-mail: t.bednarski@prawo.uni.wroc.pl

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 Heckman and Singer (1985).

Labor force surveys, carried out regularly in the EU countries, provide information on unemployment rate and constitute a good source for studies of influence of individual characteristics on unemployment duration. However, due to high nonresponse rate, the value of such studies may be questionable, unless the relation between mechanisms of nonresponse and resulting estimation bias is known sufficiently well. Verification of nonresponse mechanisms becomes possible when the survey information can be combined with individual administrative records held by employment agencies. It is important to realize that conclusions from such studies can be useful for future surveys, to assess the impact of nonresponse on resulting estimation bias.

Results proposed here integrate into a formal (mathematical) framework a method of testing causality nonresponse, outlined heuristically in Bednarski and Borowicz (2010). The subject studied is very much related to Van den Berg et al. (2006), where combined survey information with administrative records was used to assess the mechanism, effect and magnitude of nonresponse in an unemployment duration study. Van den Berg et al. propose a method to distinguish between two explanations for nonresponse in survey practice: *selectivity* - due to observed and related unobserved determinants of durations of unemployment on nonresponse, and *causality nonresponse*, a causal effect of a job exit on nonresponse. A simuation study is made to compare efficiency of the two causality testing methods.

To explain the causality effect van den Berg et al. 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 nonresponse indicator, X is a vector of explanatory variables, and 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, where T is the random variable denoting the unemployment duration. The verification of causality is based on application of a piecewise constant hazard rate model. Their method is however conditioned on a fixed time distance between unemployment entrance and the survey moment.

The method proposed here requires the population distribution to follow the Cox regression model. It consists in incorporating into the set of explanatory variables the indicator variable Z

$$Z = \begin{cases} 1 \text{ nonresponseattime} C \\ 0 \text{ responseattime} C \end{cases}$$

and studies its significance using the partial likelihood estimation. The variable C, assumed independent of T, represents a random instant of time between the moment of inflow into unemployment and the moment of the survey date. We shall argue that in a reasonable range of situations the regression coefficient next to Z is zero if and only if there is no causality effect. Standard statistical packages can then be used to test the causality effect.

From now on it is assumed that we have complete information on sample variables and for each individual we know the value of Z.

2 The model and the methods

The statistical relationship between unemployment duration T and the vector of explanatory variables X will be further described by the Cox proportional hazard model (Cox 1972) with conditional hazard

$$\lambda(t|x) = \lambda_o(t)exp(x'\beta)$$

where λ_o is the baseline hazard and β is a vector of regression parameters.

To simplify the formulas below, a single integral sign will be used for multivariate functions as it is frequently used in Lebesque integration. The partial likelihood estimator of β can then be written as a solution of $L_{F_n}(\beta) = 0$, where

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

and F_n is the empirical distribution of time and covariate variables (the censoring variable is suppressed since we assume complete information on the time variable in the sample). The cumulated baseline hazard

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

is usually estimated by the Breslow estimator (Breslow 1975).

Bednarski (1993) introduced a robust version of the partial likelihood estimator while Grzegorek (1993) (see also Bednarski 2007) proposed a robustified version of Breslow cumulated baseline hazard estimator. The robust estimator of regression parameters is based on smooth weighting incorporated into the partial likelihood score function

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

while the robust estimator of the baseline hazard (without censoring) is given by

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

where T_i, X_i are sample observations, $\hat{\beta}$ is the robust estimator and $R(T_i)$ is the risk set at time T_i. The risk set denotes all individuals which are at risk at time T_i - the individuals unemployed at time T_i.

If the weights $A(w, x)$ are sufficiently regular, then $L_F(\beta)$ becomes Fréchet differentiable with respect to F at a given Cox distribution. The differentiability in turn yields good asymptotic properties of estimates and ensures basic qualitative robustness properties, like continuity of the statistical functional and nonzero breakdown point (Bednarski 1993).

The causal nonresponse means here dependence of Z on event $(T < C)$ only. To be more precise we define the random mechanism associated with nonresponse. It will be assumed that

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

where C is a random survey time, independent of T, while b_1 and b_2 are Bernoulli variables with success probabilities $p_1 = p$ and $p_2 = p + \epsilon$ respectively, with p depending possibly on C and X and independent of T when X is given. Lack of causality is equivalent to identical success probabilities for the two Bernoulli random variables ($\epsilon = 0$) at any fixed values of C and X. We naturally assume further that the variables T and C are independent and moreover it is supposed that their distributions do not have disjoint supports.

As mentioned earlier the proposed testing method is very simple. It consists in including Z into the list of explanatory variables and performing a standard inference using the Cox regression model. The following theorem is stated for the nonrobust - partial likelihood estimation. Its proof given below can be altered to the robust version of the estimator similarly as it is done in Bednarski (1993).

Theorem 1 *Suppose β is the true regression parameter in the Cox regression model. Then the following expression, corresponding to the Cox score function*

$$\int \left[\bar{z} - \frac{\int z I_{t \ge w} \exp(\beta_0 z + \beta' x) dF(t, z, x)}{\int I_{t \ge w} \exp(\beta_0 z + \beta' x) dF(t, z, x)} \right] dF(w, \bar{z}, y) \tag{1}$$

where $F(t, z, x)$ *denotes the joint distribution of time to exit from unemployment, the nonresponse variable Z and covariates X, is equal to zero at $\beta_0 = 0$ if and only if $\epsilon = 0$.*

The nonzero value of the expression (1) for $\beta_0 = 0$ when $\epsilon \neq 0$ implies that the method consistently detects causality. Therefore we can apply the partial likelihood to estimate the regression parameters for the covariates (Z, X) and then use the estimator $\hat{\beta}_0$ corresponding to variable Z to verify the hypotheses H_0: non-causality versus H_1: causality. The distribution of $\sqrt{n}(\hat{\beta}_0 - \beta_0)$ is approximately Gaussian with mean zero and estimable standard deviation under the null hypothesis. Verification of the hypotheses has very practical consequences—under the null hypothesis the estimation for the Cox regression model remains asymptotically unbiased no matter how high the nonresponse rate is!

Proof Let $F(t, z, x)$ denote the distribution of time T, nonresponse variable Z and covariates X and put $dF(t, z, x) = dF(z|t, x)dF(t|x)dG(x)$, where $F(t|x)$ denotes

the marginal distribution of T given X while $G(x)$ stands for the distribution of X. If S denotes the distribution of survey time C then at fixed value of X

$$P(Z=1, T=t|X=x) = [p(x) + \epsilon(1 - S(t))]f(t|x)$$

where $f(t|x)$ is the marginal density for T given X and consequently

$$P(Z=1, T=t) = \int [p(x) + \epsilon(1 - S(t))]f(t|x)dG(x)$$

Therefore, when $\epsilon = 0$, β is the true parameter value and $\beta_0 = 0$ the following equations hold

$$\begin{aligned} L_F(\beta_0, \beta) &= \int p(y)f(w|y)dG(y)dw \\ &\quad - \int \frac{\int (1 - F(w|x))p(x)e^{\beta' x}dG(x)}{\int (1 - F(w|x)))e^{\beta' x}dG(x)} f(w|y)dG(y)dw \\ &= \int p(y)f(w|y)dG(y)dw - \int \frac{\int p(x)f(w|x)dG(x)}{\int f(w|x)dG(x)} f(w|y)dG(y)dw \\ &= \int p(y)f(w|y)dG(y)dw - \int p(x)f(w|x)dG(x)dw = 0. \end{aligned}$$

Since the score function corresponds here to a strictly concave objective function (the log of partial likelihood) which is maximized by (β_0, β), the solution $\beta_0 = 0$ must be unique.

For the other part of the proof we shall study the sign of the derivative of $L_F(\beta_0, \beta)$ with respect to ϵ at $\beta_0 = 0$ and show that it is positive. By continuity of the derivative we will be able to conclude that in an open neighbourhood of 0 the expression $L_F(0, \beta)$ can not be equal to zero for $\epsilon \neq 0$.

Since

$$\begin{aligned} L_F(0, \beta) &= \int [p(y) + \epsilon(1 - S(w))]f(w|y)dG(y)dw \\ &\quad - \int \frac{\int [p(x) + \epsilon(1 - S(t))]f(t|x)dG(x)}{\int f(w|x)dG(x)} f(w|y)dG(y)dw \end{aligned}$$

is linear in ϵ the derivative can be expressed as

$$\int (1 - S(w))f(w|y)dG(y)dw - \int \frac{\int_{t \geq w} (1 - S(t))f(t|x)dG(x)dt}{1 - F(w)} dF(w)$$

and it is equal to

$$- \int S(w)dF(w) + \int \frac{\int_{t \geq w} S(t)dF(t)}{1 - F(w)} dF(w)$$

where $F(.)$ stands here for the time distribution. Now, since

$$\frac{\int_{t\geq w} S(t)dF(t)}{1-F(w)} \geq \frac{\int_{t\geq w} S(w)dF(t)}{1-F(w)} = S(w)$$

it follows that the derivative is positive if for all w in an open interval of positive probability F it holds

$$\frac{\int_{t\geq w} S(t)dF(t)}{1-F(w)} > S(w),$$

which is implied by nonorthogonality of F and S. □

Corollary *Special care is needed when we suspect existence of statistically significant unobserved time determinants. The method needs not to work if Z depends on unobserved explanatory variables that influence time distribution. In general there seem to be serious formal and conceptual difficulties in assessing the influence of unobserved determinants in duration models and distinguishing it from the nonresponse causality effect in practical situations. The following chapter discusses this issue in the context of some Monte Carlo experiments.*

3 Simulations

Several Monte Carlo experiments were carried out to evaluate efficiency of the proposed method, its robustness and sensitivity to existence of unobserved explanatory variable. The method was also compared with Van den Berg et al. (2006) methodology based on the piecewise constant intensity model.

Experiment 1. Comparison with a method based on a piecewise constant intensity model. The time T representing the unemployment spell is generated from the Cox model, given by the intensity

$$\lambda(t|edu) = 0.1\exp(edu),$$

where edu is a binary explanatory variable intended to imitate "level of education" (1 – higher education, 0 – lack of higher education), generated from the binomial distribution with a fixed success probability equal 0.3. The expected value of the time variable is $ET \approx 7.9$. The survey time is set equal to 12 yielding $P(T > 12) \approx 0.22$ and the nonresponse mechanism is given by

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

for independent Bernoulli variables b_1 and b_2 with success probability p equal to 0.1, 0.2, 0.3 and ϵ =0.0, 0.1, 0.2, 0.3.

In the Van den Berg et al. (2006) proposal the time axis is divided into equal time segments (0, 2], (2, 4], ..., (18, 20], (20, ∞) and the piecewise constant intensity model with intensity parameters $\lambda_1, \lambda_2, ..., \lambda_{11}$ is used. Since by these assumptions

Table 1 Frequency of nonresponse causality detection for the Van den Berg et al. and the new method.

p_1	p_2	$\epsilon = 0$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
0.2	0.1	41 54	69 675	160 993	348 1000
0.2	0.2	59 48	80 547	180 986	374 1000
0.2	0.3	39 49	90 501	205 969	420 1000
0.3	0.1	45 44	82 625	167 989	325 1000
0.3	0.2	60 48	57 547	157 975	366 1000
0.3	0.3	57 53	85 508	205 970	411 1000

the intensities $\lambda_1, ..., \lambda_{11}$ are equal, the zero difference between intensities λ_6 and λ_7 is tested for the subsample of respondents to asses the efficiency of the method.

The *R* language is used for the simulations: "coxph" function for the new method and a general-purpose optimization routine "optim" to estimate the piecewise constant intensity model.

Table 1 summarizes results of the experiment repeated 1,000 times for sample size equal 1,000 (Table 1). For each fixed value of ϵ two columns of numbers indicate detection frequency of causal nonresponse at significance level 0.05 (causality is present when $\epsilon > 0$). The first one corresponds to the van den Berg et al. method, the second one to the new method. The results show high superiority of the new method under the above model conditions.

Experiment 2. Effect of complex model conditions on nonresponse causality testing. When the Cox model is applied to real data we have to take into account the sensitivity of inference on existence of unobserved explanatory variables related to time distributions and possible data contamination. In this Monte Carlo experiment we cover these situations. The time variable is generated from the Cox model with intensity either

$$\lambda(t|x) = (1/12)\exp(x)$$

where x is Bernoulli with $p = 0.5$ or

$$\lambda(t|x, v) = (1/12)\exp(x + v)$$

where v is Bernoulli with success probability 0.2, independent of x and not observed by the statistician. The nonresponse is generated by

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

with b_1 and b_2 independent Bernoulli with success probabilities either $p_1 = 0.2+0.3x$ or $p_1 = 0.2+0.3x+0.2v$ and $p_2 = p_1+\epsilon$. The nonresponse depends then on observed and possibly unobserved covariates. The contamination consists in replacing 5 % of randomly selected time observations by 50. Notice that the value 50 is a mild and rather difficult to identify outlier - the chance of observing the value larger or equal 50 for an exponential distribution with rate 1/12 is 0.0155.

The sample size was 1,000 and the experiment was repeated 1,000 times in each of the following experimental situations:

Table 2 Comparison of ple and robust causality testing and regression estimation for the Cox model

Causality	Cont	H_0	Regression	Se of reg
No unobserved variable				
$\epsilon = 0$	0	0.984	1.007	0.073
		0.984	**1.008**	**0.095**
$\epsilon = 0.1$	0	0.344	0.955	0.074
		0.260	**0.937**	**0.095**
$\epsilon = 0$	1	0.966	0.668	0.068
		0.968	**0.942**	**0.090**
$\epsilon = 0.1$	1	0.618	0.637	0.069
		0.332	**0.888**	**0.090**
Unobserved variable present				
$\epsilon = 0$	0	0.974	0.923	0.072
		0.974	**0.934**	**0.092**
$\epsilon = 0.1$	0	0.362	0.873	0.073
		0.298	**0.867**	**0.093**
$\epsilon = 0$	1	0.972	0.616	0.068
		0.978	**0.879**	**0.087**
$\epsilon = 0.1$	1	0.594	0.578	0.069
		0.390	**0.811**	**0.088**
Nonresponse depends on unobserved variable				
$\epsilon = 0$	0	0.668	0.896	0.072
		0.620	**0.899**	**0.092**
$\epsilon = 0.1$	0	0.038	0.844	0.073
		0.021	**0.835**	**0.092**
$\epsilon = 0$	1	0.800	0.602	0.068
		0.662	**0.846**	**0.087**
$\epsilon = 0.1$	1	0.218	0.557	0.069
		0.026	**0.780**	**0.087**

Robust results are given in bold

- Cox model without v
- Cox model with unobserved v
- nonresponse independent or dependent on unobserved variable v
- data contaminated or not

The data generation and estimation process are accomplished with R and the coxrobust package. Table 2 gives acceptance frequencies of $H_0 : \epsilon = 0$ (column 3) and it indicates the mean value of estimated regression coefficient of x variable (column 4) along with the mean value of its standard error (column 5). Causality column indicates the presence of causality mechanism in the simulations ($\epsilon = 0.1$). The second column indicates the presence of contaminated observations.

The first part of the table corresponds to the data generating mechanism given by the Cox model without unobserved variable v. In the second one the unobserved variable

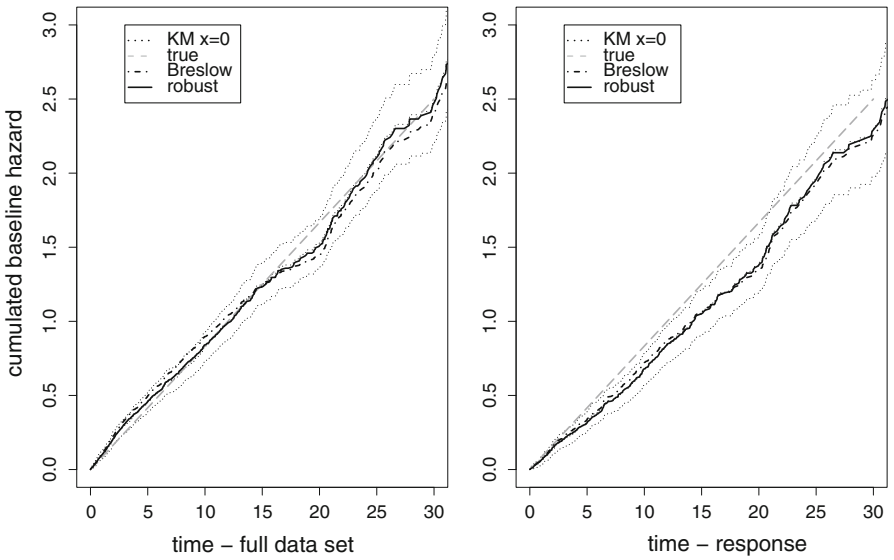

Fig. 1 Cumulated hazard estimation for the full data set (*left*) and response (*right*) under causality nonresponse, unobserved explanatory variable and contamination.

v is present however it does not affect the nonresponse mechanism. The third part takes into account all the effects.

The robust method shows in general superiority over ple estimation both in testing power and in estimation of the regression parameter. The other noticeable feature is that the presence of unobserved variable in the nonresponse mechanism lowers efficiency in testing and estimation. Moreover under contamination and the unobserved covariate present in nonresponse mechanism the robust method more frequently rejects the null hypothesis when it is "true" even though results of estimation of the regression parameter are still superior there compared to ple.

Experiment 3. Effect of complex model conditions on estimation of cumulated baseline hazard. The experiment shows exemplary results of cumulated baseline hazard estimation in a single run for the entire sample and for its response part. The sample generation conditions ($n = 1{,}000$) correspond to part 3 of Table 2 where we have all the unfavorable effects present: unobserved explanatory variable which influences the nonresponse pattern, causality nonresponse and the contamination. The distribution of variables T, X, V is exactly as in the experiment 2 except for the contamination where, to avoid abrupt changes in cumulated hazard, 5 % of randomly selected time observations were given value 20 incremented by exponential variable with expectation 5. The nonresponse probabilities are chosen as $p_1 = 0.2 + 0.3x + 0.2v$; $p_2 = p_1 + 0.2$, respectively. Figure 1 gives the following curves:

- theoretical baseline hazard
- Breslow estimate
- robust modification of Breslow estimate
- $-$log survival for $x = 0$ estimated by Kaplan Meier with 95 % confidence bands

We can see that there is only a minor difference between the Breslow estimate and its robust counterpart. In fact when the explanatory variables are bounded (here we use only indicator variables) this is a rule—there are no influential components among $exp(\hat{\beta}X_i)$. The weights A do not affect very much the cumulated hazard estimator though they improve estimation of the regression parameters as seen in Table 2. The exponential components may, however, become influential for large values X. The robust modification of Breslow estimator (Bednarski 2007) can then be of help. Another noticeable feature is that theoretical cumulated hazard may differ significantly from the estimates. Repeated analysis of cumulated hazard estimation for the complete and response part of the sample, on many runs of the program, was not sufficient to give a clear distinction between effects due to presence of unobserved variable, causality effect and contamination.

4 Conclusions

High survey nonresponse in unemployment duration studies may have strong effect on inference if so called causal mechanism is present (exit from unemployment changes the chance of nonresponse). A formal description of the mechanism is given and a robust method of testing its presence is justified for data sets where survey information can be combined with complete administrative records when population distribution follows approximately the Cox regression model. Comparison of the method with the one proposed in Van den Berg et al. (2006) shows that it is essentially more efficient and very simple in application. A Monte Carlo study shows the estimation and testing results for the Cox regression model under dependence of nonresponse probability on explanatory variables, under data contamination and under dependence of unemployment time on unobserved explanatory variables. The method shows high consistency under the influence of observed explanatory variables and resistance to outliers. It is, however, less resistant to the influence of unobserved explanatory variables on the nonresponse probabilities. A simulated estimation of cumulated hazards indicates that the nonresponse effect can not be consistently detected there under the influence of unobserved explanatory variables.

References

Bednarski T (1993) Robust estimation in the Cox regression model. Scand J Stat 20:213–225
Bednarski T (2007) On a robust modification of Breslows cumulated hazard estimator. Comput Stat Data Anal 52:234–238
Bednarski T., Borowicz F. (2010) Analysis of non-response causality in labor market surveys. Acta Universitatis Lodziensis. Folia Oeconomica 253, 217–224, Lodz 2010.
Breslow NE (1975) Analysis of survival data under the proportional hazards model. Int Stat Rev 43:45–58
Cox RD (1972) Regression model and life tables. J R Stat Soc Ser B 34:187–220

Groves R (2006) Nonresponse rates and nonresponse bias in household. Surv Public Opin Q 70(Special Issue):646–667
Grzegorek, K. (1993) On robust estimation of baseline hazard under the Cox model and via Frechet differentiability. Preprint of the Institute of Mathematics of the Polish Academy of Sciences 518.
Heckman JJ, Singer BS (1985) Longitudinal analysis of labor market data, econometric society monographs 10. Cambridge University Press, Cambridge
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
O'Muircheartaigh C, Campanelli PA (1999) Multilevel exploration of the role of interviewers in survey nonresponse. J R Stat Soc Ser A 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. Adv Stat Anal 92:297–318
Romeo CJ (1997) Measuring information loss due to inconsistencies in duration data from longitudinal surveys. J Econ 78:159–177
Van den Berg GJ, Lindeboom M, Ridder G (1994) Attrition in longitudinal panel data and the empirical analysis of dynamic labour market behaviour. J Appl Econ 9:421–435
Van den Berg GJ, Lindeboom M, Dolton P (2006) Survey nonresponse and the duration of unemployment. J R Stat Soc Ser A 169:585–604