Controling the Healthy Worker Effect in Occupational Epidemiology

Jinheum Kim[1] and Chung Mo Nam[2]

Summary

The healthy worker effect is an important issue in occupational epidemiology. We proposed a new statistical method to test the relationship between exposure and time to death in the presence of the healthy worker effect. In this study, we considered the healthy worker hire effect to operate as a confounder and the healthy worker survival effect to operate as a confounder and an intermediate variable. The basic idea of the proposed method reflects the length bias-sampling caused by changing one's employment status. Simulation studies were also carried out to compare the proposed method with the Cox proportional hazards models. According to our simulation studies, both the proposed test and the test based on the Cox model having the change of the employment status as a time-dependent covariate seem to be satisfactory at an upper 5% significance level. The Cox models, however, are inadequate with the change, if any, of the employment status as time-independent covariate. The proposed test is superior in power to the test based on the Cox model including the time-dependent employment status.

Key words : Healthy worker effect, Length bias-sampling, Score statistic

1. Introduction

In occupational epidemiology, much attention has been paid in explaining the phenomenon of the healthy worker effect [1,2]. The healthy worker effect can be defined as a phenomenon where the mortality of individuals exposed to a specific risk is lower than that of a general population. It can be divided into two important parts, a healthy worker hire effect and a healthy worker survival effect [3-5]. The former arises from health workers being employed more likely than those who are relatively less healthy on an initial selection process. The latter refers to a continuing selection process such that the possibility of an individual still remaining employed in workplace is larger in healthy workers than that in unhealthy workers. We can consider the healthy worker hire effect as a confounding effect caused by the difference in health status among individuals at the time of

[1] (445-743) 경기도 화성시 봉담읍 와우리 산 2-2호 수원대학교 통계정보학과 조교수

[2] (120-749) 서울시 서대문구 신촌동 134 연세대학교 의과대학 예방의학교실 조교수

employment. It can be easily controlled including the health status of individuals in a model. There are two major approaches to control the healthy worker survival effect according on how to deal with the employment status. One approach contains the following: a method to restrict an analysis to those who have large survival since their initial hire and then stratify them on the employment status [6]; a method of lagging the exposure up to a predetermined time [7]; a method of involving current employment status as an indicator variable in a regression model [8,9]. On the other hand, Robins [5] explained the healthy worker survival effect as a phenomenon caused by a time-dependent employment status operating simultaneously as a confounder and an intermediate variable. He asserted that the standard methods, including Cox model with the time-dependent employment status, will be biased [10]. Instead, he proposed the so-called G-null set, G-algorithm and structural nested failure time model (SNFTM), to remove the healthy worker survival effect [5,11]. Although his algorithms were theoretically well derived, they still have some drawbacks such as complication in computational process, dependency on given data set, and restriction to checking the adequacy of proposed methods through simulations under various configurations. On the contrary, Nam and Zelen [12] have shown that both a logrank test and a stratified logrank test have very high significance levels in the comparison of two survival distributions with time-dependent intermediate variable. They propose both a conceptual model and a statistical model to remove the bias due to the length bias sampling came from a time-dependent intermediate variable.

In this study, we first construct a conceptual model which is able to control the healthy worker effect and propose a test procedure based on the model. We perform simulations to compare the proposed test with the tests based on Cox model in terms of significance level and power.

2. Score method

The healthy worker effect originates from both the health status of an individual at the time of employment which operates as a confounder and a time-dependent employment status at time t which operates as a confounder and an intermediate variable. The healthy worker hire effect can be easily removed by treating the health status as a confounder in a model. The healthy worker survival effect can be explained with the following model: the exposure amount at t may affect an employment status at $t+1$ and reversely the employment status at $t+1$ may affect an exposure amount at $t+2$, and so on. To control the healthy worker survival effect in our theoretical model, we introduce a methodology on the basis of Nam and Zelen [12]. Let define three conceptual times; survival time (T_0) conditional on that the employment status is not changed, survival time (T_1) conditional on that the employment status has been changed, and waiting time W to the change of employment status. Since T_0 and T_1 are conceptual variables and observed survival times may be truncated according to employment status, the methods using only observed survival times lead to the biased results. It is reasonable to assume that the change of employment status can be observed through competing relation between T_0 and W.

Let survival functions of T_0, T_1 and W be $Q_0(t) = Pr(T_0 > t)$, $Q_1(t) = Pr(T_1 > t)$, and $G(t) = Pr(W > t)$, respectively. Let $Y(t)$ be the exposure amount of an individual at t and $S(t)$ be a $p \times 1$ vector of values of time-dependent and/or time-independent covariates at t. Consider a model, for $i = 0, 1$,

$$Q_{i(t)} = Q_{i0}(t)^{\exp\{\beta_i Y(t) + \gamma_i' S(t)\}}, \qquad (1)$$

where Q_{i0} is the survival function corresponding to the unknown baseline hazard function at the exposure amount of 0 and β_i and γ_i are the unknown regression parameters. The subscript $i = 0, 1$ indicates with or without the change of the employment status. The $\gamma_i (i = 0, 1)$ are nuisance parameters and our hypothesis of interest corresponds to

$$H_0 : \beta_0 = \beta_1 = 0 \ vs. \ H_1 : \text{atleast}\, \beta_i \neq 0\, (\mathrm{i} = 0, 1)$$

i.e., there is no effect of exposure on survival regardless of the employment status. Define $T = (1 - Z) T_0 + Z T_1$, where $Z = I(W \leq T_0)$. That is, $T = T_1$ if the employment status of an individual has been changed $(Z = 1)$ or $T = T_0$ otherwise $(Z = 0)$. Assume that T is subject to censoring and that conditional on covariates the survival and censoring time are independent. Define a censoring indicator as $\delta = 1$ if a observation is uncensored or $\delta = 0$ otherwise. The vector (t, δ, w, z, y, s) corresponds to the information for a single observation. Following the arguments of Nam and Zelen [12], we have the log-likelihood based on a single observation as

$$\begin{aligned} l(\theta \mid t, \delta, z, y, s) = {} & (1 - z)[\delta\{\beta_0 y(t) + \gamma_0' s(t)\} + \exp\{\beta_0 y(t) + \gamma_0' s(t)\} \log Q_{00}(t)] \\ & + z[\delta\{\beta_1 y(t) + \gamma_1' s(t)\} + \exp\{\beta_1 y(t) + \gamma_1' s(t)\} \log\{Q_{10}(t)/Q_{10}(w)\} \\ & + \exp\{\beta_0 y(t) + \gamma_0' s(t)\} \log Q_{00}(t)] + Re, \end{aligned}$$

where $\theta = (\beta_0, \beta_1, \gamma_0', \gamma_1')'$ and "Re" denotes the terms not involving $\beta_0, \beta_1, \gamma_0$ and γ_1. Therefore, the log-likelihood for n observations is

$$l_{n=} \sum_{k=1}^{n} l(\theta \mid t_k, \delta_k, w_k, z_k, y_k, s_k).$$

For $k = 1, \ldots, n$, define $N_k(t) = I(T_k \leq t, \delta_k = 1)$, $R_k(t) = I(T_k \geq t)$, and $Z_k(t) = I(W_k \leq t)$. Under model (1), the natural estimates of $Q_{i0}(t)(i = 0, 1)$, according to the arguments of Nam and Zelen [12], are given as

$$\log\hat{Q}_{00}(t) = -\int_0^t \frac{\sum_{k=1}^{n}\{1-Z_k(u)\}dN_k(u)}{\sum_{k=1}^{n}\{1-Z_k(u)\}R_k(u)\exp\{\beta_0 y_k(u)+\gamma_0' s_k(u)\}},$$

$$\log\hat{Q}_{10}(t) = -\int_0^t \frac{\sum_{k=1}^{n} Z_k(u)dN_k(u)}{\sum_{k=1}^{n} Z_k(u)R_k(u)\exp\{\beta_1 y_k(u)+\gamma_1' s_k(u)\}}.$$

Define $\hat{\gamma}_i(i=0,1)$ to be the restricted maximum likelihood estimate of γ_i as a solution to $\frac{\partial l_n}{\partial\gamma_0}=0$ or $\frac{\partial l_n}{\partial\gamma_1}=0$ subject to $\beta_0=\beta_1=0$. Since we can not derive explicit form of $\hat{\gamma}_i$ from (8) or (9), we may rely on a Newton-Raphson method to do this. Also, define score statistics $\hat{U}_1$ and $\hat{U}_2$ as follows:

$$\hat{U}_1 = \frac{\partial l_n}{\partial\beta_0}\,\Big|_{\beta_0=\beta_1=0,\hat{\gamma}_0,\hat{\gamma}_1,\hat{Q}_{00},\hat{Q}_{10}}$$

$$= \sum_{k=1}^{n}\int_0^{\infty}\left[y_k(u)-\frac{\sum_{j=1}^{n}\{1-Z_j(u)\}R_j(u)\exp\{\gamma_0' s_j(u)\}y_j(u)}{\sum_{j=1}^{n}\{1-Z_j(u)\}R_j(u)\exp\{\gamma_0' s_j(u)\}}\right]1-Z_k(u)dN_k(u),$$

$$\hat{U}_2 = \frac{\partial l_n}{\partial\beta_1}\,\Big|_{\beta_0=\beta_1=0,\hat{\gamma}_0,\hat{\gamma}_1,\hat{Q}_{00},\hat{Q}_{10}}$$

$$= \sum_{k=1}^{n}\int_0^{\infty}\left[y_k(u)-\frac{\sum_{j=1}^{n}Z_j(u)R_j(u)\exp\{\gamma_1' s_j(u)\}y_j(u)}{\sum_{j=1}^{n}Z_j(u)R_j(u)\exp\{\gamma_1' s_j(u)\}}\right]Z_k(u)dN_k(u).$$

By the standard multivariate theory, the null distribution of score vector, $U=(\hat{U}_1,\hat{U}_2)'$, asymptotically follows a bivariate normal with mean 0 and variance-covariance matrix

$$\hat{\Sigma}=\begin{pmatrix}\hat{\sigma}_{11} & 0\\ 0 & \hat{\sigma}_{22}\end{pmatrix},$$

where the exact formula of $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ in $\hat{\Sigma}$ are given in Appendix. Based on this result, we propose a score test statistic for testing $H_0:\beta_0=\beta_1=0$ against H_1 : at least $\beta_i\neq 0\,(i=0,1)$ as

$$X^2 = U'\hat{\Sigma}^{-1}U = \hat{U}_1^{\,2}/\hat{\sigma}_{11}+\hat{U}_2^{\,2}/\hat{\sigma}_{22}, \tag{12}$$

which, under H_0, asymptotically follows a χ^2 distribution with two degrees of freedom. We reject H_0 in favor of H_1 at the significance level α when $X^2\geq\chi^2_\alpha(2)$, where $\chi^2_\alpha(2)$ is the $100 \times (1-\alpha)$ percentile point of the χ^2 distribution with two degrees of freedom

References

[1]Choi BCK. Definition, sources, magnitude, effect modifiers, and strategies of reduction of the healthy worker effect. *Journal of Occupational Medicine* 1992; **34**:979-988.

[2]Checkoway H, Pearce N, Crawford-Brown DJ. *Research Methods in Occupational Epidemiology : Monographs in Epidemiology and Biostatistics.* Oxford University Press: New York, 1989.

[3]Arrighi HM, Hertz-Picciotto I. The evolving concept of the healthy worker survival effect. *Epidemiology* 1994; **5**:189-196.

[4]Arrighi HM, Hertz-Picciotto I. Controlling the healthy worker survival effect: an example of arsenic exposure and respiratory cancer. *Occupational and Environmental Medicine* 1996; **53**:455-462.

[5]Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survival effect. *Mathematical Modelling* 1986; **7**:1393-1512.

[6]Fox AJ, Collier PF. Low mortality rates in industrial cohort studies due to selection for work and survival in the industry. *British Journal of Preventive and Social Medicine* 1976; **30**:225-230.

[7]Gilbert E. Some confounding factors in the study of mortality and occupational exposures. *American Journal of Epidemiology* 1982; **116**:177-188.

[8]Gilbert E, Marks S. An analysis of the mortality of workers in a nuclear facility. *Radiation Research* 1979; **79**:122-148.

[9]Steenland K, Stayner L. The importance of employment status in occupational cohort mortality studies. *Epidemiology* 1991; **2**:418-423

[10]Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the of AIDS patients. *Epidemiology* 1992a; **3**:319-336.

[11]Robins JM. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992b; **79**:321-334.

[12]Nam CM, Zelen M. Comparing the survival of two groups with an intermediate clinical event. *Lifetime Data Analysis* 2001; **7**:5-19.