Controlling the healthy worker effect in the presence of an intermediate variable

Jinheum Kim¹, Chung Mo Nam²*,† and Dae Ryong Kang³

¹Department of Applied Statistics, University of Suwon, Suwon P.O. Box 77, Suwon, 445-743, South Korea
²Department of Preventive Medicine and Public Health, Yonsei University College of Medicine, C.P.O. Box 8044, Seoul, 120-749, South Korea
³Graduate School of Health Science and Management, Yonsei University, Seoul, 120-749, South Korea

SUMMARY

The healthy worker effect is an important issue in occupational epidemiology. We propose a new statistical method to test the relationship between exposure and time to death in the presence of the healthy worker effect. The basic idea of the proposed method reflects length bias sampling caused by job changes. According to simulation studies, both the proposed test and a test based on the Cox model including job changes as a time-dependent covariate seem to be satisfactory at the upper 5% significance level. However, the Cox models involving job changes as a time-independent covariate are unsatisfactory. The proposed test is superior in power to the test based on the Cox model with time-dependent employment status. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: Cox model; healthy worker effect; intermediate variable; length bias sampling; score test

1. INTRODUCTION

In occupational epidemiology, much attention has been paid in explaining the healthy worker effect (Choi, 1992; Checkoway et al., 1989). 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, as a healthy worker hire effect (HWHE) and a healthy worker survival effect (HWSE) (Arrighi and Hertz-Picciotto, 1994, 1996; Robins, 1986). The former arises from health workers being more likely to be employed than those who are relatively less healthy on an initial selection process. The latter refers to a continuing selection process such that the probability of an individual still remaining employed in a workplace is greater in healthy workers than in unhealthy workers.

We can consider the HWHE as a confounding effect caused by the difference in health status among individuals at hire. It can be easily controlled by including the health status of individuals in a model. There are two major approaches to control the HWSE depending on how to deal with employment
status. One approach contains the following: a method to restrict an analysis to those who have enough survival since their initial hire and then stratify them on employment status (Fox and Collier, 1976); a method of lagging the exposure up to a predetermined time (Gilbert, 1982); and a method of involving current employment status as an indicator variable in a regression model (Gilbert and Marks, 1979). On the other hand, Robins (1986) explained the HWSE as a phenomenon caused by time-dependent employment status operating simultaneously as an intermediate and a confounding variable. He asserted that the standard methods, including a Cox model with time-dependent employment status, will be biased (Robins et al., 1992). Instead he proposed a so-called G-null set, G-algorithm and structural nested failure time model (SNFTM) to remove the HWSE (Robins, 1986, 1992b). Although his algorithms were theoretically well derived, they still have some drawbacks such as complications in the computational process, dependency on a given data set, and restriction to checking the adequacy of proposed methods through simulations under various configurations. On the contrary, Nam and Zelen (2001) 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 a time-dependent intermediate variable. They proposed both a conceptual model and a statistical model to remove the bias due to length bias sampling.

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 in Section 2. We perform simulations to compare the proposed test with the test based on the Cox model in terms of significance level and power in Section 3. A discussion is provided in Section 4.

2. SCORE METHOD

The healthy worker effect originates from the health status of an individual at hire which operates as a confounder and from time-dependent employment status at $t$ which operates as an intermediate and a confounding variable. The HWHE can be easily removed by treating health status as a confounder in a model. The HWSE can be explained with the following model: the exposure amount at $t$ may affect employment status at $t+1$, and conversely employment status at $t+1$ may affect an exposure amount at $t+2$, and so on. To control the HWSE in our theoretical model, we introduce a methodology on the basis of Nam and Zelen (2001). Let us define three conceptual times: survival time ($T_0$) conditional on employment status not changing; survival time ($T_1$) conditional on employment status changing; and waiting time ($W$) to job changes. 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 job changes can be observed through competing relations between $T_0$ and $W$.

Let $Y_i, i = 0, 1$, be the exposure amount of an individual and $S$ be a $p \times 1$ vector of time-independent covariates. The subscript $i = 0, 1$ indicates without or with job changes. Conditional on $Y_i = y_i, S = s$, let survival functions of $T_0, T_1$ and $W$ be $Q_0(t), Q_1(t)$ and $G(w)$, respectively. Also, the probability density functions of $T_0, T_1$ and $W$ will be denoted by $q_0(t), q_1(t)$ and $g(w)$, respectively. Define $T = (1 - Z)T_0 + ZT_1$, where $Z = I(W \leq T_0)$. That is, $T = T_1$ if employment status of an individual has been changed ($Z = 1$) or $T = T_0$ otherwise ($Z = 0$).

As in Nam and Zelen (2001), an observation on survival time without experiencing job changes implies that the waiting time to job changes has been right censored, i.e.

$$f(t, z = 0) = q_0(t)G(t)$$ (1)
The event of \( Z = 1 \) implies that the waiting time to job changes is observed and the survival time without experiencing job changes must exceed the observed waiting time, i.e. \( f(w, z = 1) = Q_0(w)g(w) \). Also, note that when \( Z = 1, T = T_1 \), the waiting time is observed with \( w < T_0 \) followed by the distribution for survival time with experiencing job changes past \( w \), i.e.

\[
f(t, w, z = 1) = f(w, z = 1) f(t | w, z = 1) = Q_0(w)g(w) q_1(t) / Q_1(w), \quad \text{for } 0 < w < t < \infty \quad (2)
\]

Consider a model, for \( i = 0, 1 \),

\[
Q_i(t) = Q_{0i}(t) \exp\{\beta_i Y_i + \gamma_i S\}
\]

where \( Q_{0i} \) is the baseline survival function and \( \beta_i \) and \( \gamma_i \) are the unknown regression parameters. The \( \gamma_i, i = 0, 1 \), are nuisance parameters and our hypothesis of interest corresponds to

\[
H_0 : \beta_0 = \beta_1 = 0 \quad \text{vs.} \quad H_1 : \text{at least } \beta_i \neq 0, \quad i = 0, 1
\]

i.e. there is no effect of exposure on survival regardless of employment status. 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 an observation is uncensored or \( \delta = 0 \) otherwise. The vector \((x, \delta, w, z, y, s)\) corresponds to the information for a single observation. Here \( x \) denotes the minimum of survival time and censoring time. Under our assumed model (3), using (1) and (2), we have the likelihood based on a single observation as

\[
L(\theta | x, \delta, w, z, y, s) = f(x, w, z = 1)^{\delta} \left\{ \int_{s}^{\infty} f(u, w, z = 1) du \right\}^{(1-\delta)z} \\
\times f(x, z = 0)^{\delta(1-z)} \{ P(T_0 > x, W > x) \}^{(1-\delta)(1-z)} \\
= \{ Q_0(w) g(w) q_1(x) / Q_1(w) \}^{\delta} \{ Q_0(w) g(w) q_1(x) / Q_1(w) \}^{(1-\delta)z} \\
\times \{ q_0(x) G(x) \}^{\delta(1-z)} \{ Q_0(x) G(x) \}^{(1-\delta)(1-z)}
\]

where \( \theta = (\beta_0, \beta_1, \gamma_0, \gamma_1) \). Also, the log-likelihood is given by

\[
l(\theta | x, \delta, w, z, y, s) = (1 - z) [\delta(\beta_0 y_0 + \gamma_0 s) + \exp(\beta_0 y_0 + \gamma_0 s) \log Q_{00}(x)] \\
+ z [\delta(\beta_1 y_1 + \gamma_1 s) + \exp(\beta_1 y_1 + \gamma_1 s) \log \{ Q_{10}(x) / Q_{01}(w) \} \\
+ \exp(\beta_0 y_0 + \gamma_0 s) \log Q_{00}(w)] + \Re
\]

where ‘\( \Re \)’ denotes the terms not involving \( \beta_0, \beta_1, \gamma_0 \) and \( \gamma_1 \). Therefore, the log-likelihood for \( n \) observations is

\[
l_n = \sum_{k=1}^{n} l(\theta | t_k, \delta_k, w_k, z_k, y_k, s_k)
\]

and the score functions associated with this log-likelihood are as follows:
Let the Newton–Raphson method to do this. Also, define score statistics

\[ U_1 = \frac{\partial l_n}{\partial \beta_0} = \sum_{k=1}^{n} \left[ (1 - z_k) \{ \delta_k y_{0k} + y_{0k} \exp(\beta_0 y_{0k} + \gamma'_0 s_k) \log Q_{10}(x_k) \} + z_k \{ y_{0k} \exp(\beta_0 y_{0k} + \gamma'_0 s_k) \log Q_{10}(w_k) \} \right] \]

\[ U_2 = \frac{\partial l_n}{\partial \beta_1} = \sum_{k=1}^{n} z_k \{ \delta_k y_{1k} + y_{1k} \exp(\beta_1 y_{1k} + \gamma'_1 s_k) \log \{ Q_{10}(x_k)/Q_{10}(w_k) \} \} \]

\[ U_3 = \frac{\partial l_n}{\partial \gamma_0} = \sum_{k=1}^{n} \left[ (1 - z_k) \{ \delta_k s_k + s_k \exp(\beta_0 y_{0k} + \gamma'_0 s_k) \log Q_{10}(x_k) \} \right. \]

\[ \left. + z_k \{ s_k \exp(\beta_0 y_{0k} + \gamma'_0 s_k) \log Q_{10}(w_k) \} \right] \]

\[ U_4 = \frac{\partial l_n}{\partial \gamma_1} = \sum_{k=1}^{n} z_k \{ \delta_k s_k + s_k \exp(\beta_1 y_{1k} + \gamma'_1 s_k) \log \{ Q_{10}(x_k)/Q_{10}(w_k) \} \} \]

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 (3), the natural estimates of \( Q_{i0}(t), i = 0, 1 \), according to the arguments of Nam and Zelen (2001), are given as

\[ \log \hat{Q}_{10}(t) = - \int_{0}^{t} \frac{\sum_{k=1}^{n} \{ 1 - Z_k(u) \} dN_k(u)}{\sum_{k=1}^{n} \{ 1 - Z_k(u) \} R_k(u) \exp(\beta_0 y_{0k} + \gamma'_0 s_k)} \]

\[ \log \hat{Q}_{10}(t) = - \int_{0}^{t} \frac{\sum_{k=1}^{n} Z_k(u) dN_k(u)}{\sum_{k=1}^{n} Z_k(u) R_k(u) \exp(\beta_1 y_{1k} + \gamma'_1 s_k)} \]

Let \( \hat{U}_3 \) and \( \hat{U}_4 \) be \( U_3 \) and \( U_4 \) evaluated at \( \beta_0 = \beta_1 = 0 \) after substituting \( Q_{i0}, i = 0, 1 \), in (6) and (7) by estimates (8) and (9). In particular, we have:

\[ \hat{U}_3 = U_3 \bigg|_{\beta_0 = \beta_1 = 0} = \sum_{k=1}^{n} \int_{0}^{\infty} \left[ s_k - \frac{\sum_{j=1}^{n} \{ 1 - Z_j(u) \} R_j(u) \exp(\gamma'_0 s_j)}{\sum_{j=1}^{n} \{ 1 - Z_j(u) \} R_j(u) \exp(\gamma'_0 s_j)} \right] \left\{ 1 - Z_k(u) \right\} dN_k(u) \]

\[ \hat{U}_4 = U_4 \bigg|_{\beta_0 = \beta_1 = 0} = \sum_{k=1}^{n} \int_{0}^{\infty} \left[ s_k - \frac{\sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma'_1 s_j)}{\sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma'_1 s_j)} \right] Z_k(u) dN_k(u) \]

Define \( \hat{\gamma}_i, i = 0, 1 \), to be the restricted maximum likelihood estimate of \( \gamma_i \) as a solution to \( \hat{U}_3 = 0 \) or \( \hat{U}_4 = 0 \). Since we cannot derive the explicit form of \( \hat{\gamma}_i \) from (10) or (11), we may rely on a Newton–Raphson method to do this. Also, define score statistics \( \hat{U}_1 \) and \( \hat{U}_2 \) as follows:
CONTROLLING THE HEALTHY WORKER EFFECT

\[ \hat{U}_1 = U_1 \bigg|_{\beta_0 = \hat{\beta}_1 = 0, \gamma_0, \gamma_1, \Psi_0, \Psi_1} \]

\[ = \sum_{k=1}^{n} \int_0^\infty \left[ y_{1k} - \sum_{j=1}^{n} \{1 - Z_j(u)\} R_j(u) \exp(\gamma'_j y_j) \right] \{1 - Z_k(u)\} dN_k(u) \]  

\[ \hat{U}_2 = U_2 \bigg|_{\beta_0 = \hat{\beta}_1 = 0, \gamma_0, \gamma_1, \Psi_0, \Psi_1} \]

\[ = \sum_{k=1}^{n} \int_0^\infty \left[ y_{1k} - \sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma'_j y_j) \right] Z_k(u) dN_k(u) \]  

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 \( \mathbf{0} \) and estimated variance–covariance matrix

\[ \hat{\Sigma} = \begin{pmatrix} \hat{\sigma}_{11} & 0 \\ 0 & \hat{\sigma}_{22} \end{pmatrix} \]

where the exact formulas of \( \hat{\sigma}_{11} \) and \( \hat{\sigma}_{22} \) in \( \hat{\Sigma} \) are given in the 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

\[ X^2 = U' \hat{\Sigma}^{-1} U = \hat{U}_1^2 / \hat{\sigma}_{11} + \hat{U}_2^2 / \hat{\sigma}_{22} \]  

which, under \( H_0 \), asymptotically follows a \( \chi^2 \) distribution with two degrees of freedom. We reject \( H_0 \) in favor of \( H_1 \) at the significance level \( \alpha \in (0, 1) \) when \( \chi^2 \geq \chi^2_{\alpha}(2) \), where \( \chi^2_{\alpha}(2) \) is the \( 100 \times (1 - \alpha) \) percentile point of the \( \chi^2 \) distribution with two degrees of freedom.

3. SIMULATION STUDIES

We perform simulations to investigate the finite-sample performance of the proposed score test \( (14) \) in terms of empirical significance level and power. Also, the proposed test is compared with the Wald test based on various Cox proportional hazard models. We considered three covariates, i.e. exposure amount \( Y(t) \), time-independent health status \( S \) having binary values, and job changes, \( Z \) or \( \tilde{Z}(t) \). The Cox models included in simulations are as follows:

(I) model with time-dependent exposure amount and health status, i.e.

\[ \lambda^I(t) = \lambda_0(t) \exp\{\alpha_{11} Y(t) + \alpha_{12} S\} \]

(II) model with time-dependent exposure amount, health status, and employment status at the end of study, i.e.

\[ \lambda^{II}(t) = \lambda_0(t) \exp\{\alpha_{21} Y(t) + \alpha_{22} S + \alpha_{23} Z\} \]
(III) model with time-dependent exposure amount, health status, and time-dependent employment status, i.e.

\[ \lambda_{III}(t) = \lambda_0(t)\exp\{\alpha_{31}Y(t) + \alpha_{32}S + \alpha_{33}Z(t)\} \]

In addition, denote the proposed model in (3) by (IV). Here, \( \lambda_0(\cdot) \) denotes the baseline hazard function and \( \lambda^m(\cdot)(m = I, II, III) \) the hazard function for the survival time of an individual with covariates corresponding to the respective model. Note that the hypotheses corresponding to \( H_0 : \beta_i = 0, i = 0, 1, \) are based on models (I)–(III),

\[ H_0 : \sigma_{it} = 0 \quad \text{vs.} \quad H'_1 : \alpha_{i1} \neq 0 \quad i = 1, 2, 3 \]

According to health status and employment status, we parameterized average survival times at an exposure amount of 0 as \((m_{0h}, m_{0u}, m_{1h}, m_{1u})\), where \( m_{0h}(or \ m_{0u}) \) is the average survival time of healthy (or unhealthy) individuals without experiencing job changes and \( m_{1h}(or \ m_{1u}) \) is that of healthy (or unhealthy) individuals who did experience job changes. Simulations include four configurations: \((m_{0h}, m_{0u}, m_{1h}, m_{1u}) = (2.0, 2.0, 2.0, 2.0)\), say \([A]\), or \((2.0, 1.5, 2.0, 1.5)\), say \([B]\), or \((2.0, 2.0, 1.5, 1.5)\), say \([C]\), or \((2.0, 1.5, 1.5, 1.0)\), \([D]\). The \([A]\) corresponds to the case of having no HWHE and HWSE, and \([B]\), \([C]\) and \([D]\) to those of having only HWHE, having only HWSE, and having both HWHE and HWSE, respectively. We assume the baseline survival distribution to be a Weibull distribution with shape parameter of 2. The scale parameters of this distribution can be determined from fixed values of \((m_{0h}, m_{0u}, m_{1h}, m_{1u})\). The exposure amounts of individuals in the healthy group are generated from uniform \((\frac{y_0}{2}, 1)\) and those in unhealthy group from uniform \((\frac{y_1}{2}, 1)\). Basically, we assume that healthy individuals may be more exposed to risk more than unhealthy ones. The waiting times to job changes are generated from

\[ Q_w(t) = Q_{w0}(t)\exp(\eta Y(t)) \]

where \( Q_{w0} \) is the baseline survival function and \( \eta \) is the unknown regression parameter. Set \( \eta = 0.5 \) and \( Q_{w0} \) as a Weibull distribution of shape parameter of 2. Likewise, the scale parameter of this distribution is directly determined from a fixed value, say 2, of average waiting time at an exposure amount of 0. In this way, we assume that the more exposed to risk the sooner job changes happen. Moreover, we assume that if the exposure amount of an individual is \( y_0 \) at the beginning stage and then job changes happened at \( w \), the exposure amounts after that time are generated from uniform \((0, y_0)\), i.e. the change on employment status makes the individual less exposed to risk. To be specific, let \((\alpha_0, \lambda_0)\) and \((\alpha_1, \lambda_1)\) be pairs of shape and scale parameters of Weibull distribution related to survival times depending on employment status and \((\alpha_w, \lambda_w)\) be those of Weibull distribution related to waiting time to job changes. Fix \( \beta_0 \) and \( \beta_1 \) in model (3). Assume \( y_0 \) is the exposure amount up to job changes and \( y_1 \) is after that event. Then, we have \((t_0, t_1, w)\) as follows:

\[ t_0 = -\log(1 - u_0)/\{\exp(\beta_0y_0)c_0\}^{1/\alpha_0} \]

\[ w = -\log(1 - u_w)/\{\exp(\beta_0y_0)c_w\}^{1/\alpha_w} \]

\[ t_1 = \{[-\log(1 - u_1) - \exp(\beta_0y_0)c_0w^{\alpha_0} + \exp(\beta_1y_1)c_1w^{\alpha_1}] / \{\exp(\beta_1y_1)c_1\}\}^{1/\alpha_1} \]

where \( c_0 = \lambda_0^{\alpha_0}, c_1 = \lambda_1^{\alpha_1}, c_w = \lambda_w^{\alpha_w} \) and \( u_0 \) and \( u_w \) are uniform \([0, 1]\) variates and also \( u_1 \) is a uniform \([1 - \exp\{-\exp(\beta_0y_0)c_0w^{\alpha_0}\}, 1]\) variate. Censoring times are generated from an exponential distribution with mean 2. All simulations were carried out with sample size 200 in each health status group and
were replicated 1000 times. Tables 1 and 2 provide empirical levels and powers of the proposed test and Wald test based on the Cox model.

Table 1 presents empirical upper 5% significance levels over four different configurations under no censoring and heavy censoring when the null hypothesis is true. First, focus on the entries of the first four rows in Table 1, i.e. no censoring case. Each entry in Table 1 was calculated from 1000 replications, so the observed Type I error probabilities have the standard error of about 0.007. The values in column ‘% Resp.’ denote the percentage of individuals experiencing job changes. When no

<table>
<thead>
<tr>
<th>(m_{0h}, m_{0u}, m_{1h}, m_{1u})</th>
<th>% Resp.</th>
<th>% Cens.</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2.0,2.0,2.0,2.0)</td>
<td>59.2</td>
<td>0</td>
<td>0.047</td>
</tr>
<tr>
<td>(2.0,1.5,2.0,1.5)</td>
<td>52.1</td>
<td>0</td>
<td>0.046</td>
</tr>
<tr>
<td>(2.0,2.0,1.5,1.5)</td>
<td>59.2</td>
<td>0</td>
<td>0.874</td>
</tr>
<tr>
<td>(2.0,1.5,1.5,1.0)</td>
<td>52.1</td>
<td>0</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Table 2. Empirical powers of Wald test based on model (III) and proposed score test (model (IV)) based on model (1) with 1000 simulations for two configurations; (m_{0h}, m_{0u}, m_{1h}, m_{1u}) = (2.0, 2.0, 1.5, 1.5) or (2.0, 1.5, 1.5, 1.0)

<table>
<thead>
<tr>
<th>(\beta_0, \beta_1)</th>
<th>% Resp.</th>
<th>% Cens.</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>(m_{0h}, m_{0u}, m_{1h}, m_{1u}) = (2.0, 2.0, 1.5, 1.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.4,0.0)</td>
<td>51.9</td>
<td>0.070</td>
<td>0.124</td>
</tr>
<tr>
<td>(0.8,0.0)</td>
<td>44.4</td>
<td>0.212</td>
<td>0.323</td>
</tr>
<tr>
<td>(1.2,0.0)</td>
<td>37.1</td>
<td>0.495</td>
<td>0.707</td>
</tr>
<tr>
<td>(1.5,0.0)</td>
<td>32.2</td>
<td>0.735</td>
<td>0.898</td>
</tr>
<tr>
<td>(0.0,0.4)</td>
<td>59.3</td>
<td>0.231</td>
<td>0.233</td>
</tr>
<tr>
<td>(0.0,0.8)</td>
<td>59.3</td>
<td>0.657</td>
<td>0.714</td>
</tr>
<tr>
<td>(0.0,1.2)</td>
<td>59.3</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>(0.0,1.5)</td>
<td>59.3</td>
<td>0.999</td>
<td>1.000</td>
</tr>
<tr>
<td>(m_{0h}, m_{0u}, m_{1h}, m_{1u}) = (2.0, 1.5, 1.5, 1.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.4,0.0)</td>
<td>44.8</td>
<td>0.108</td>
<td>0.141</td>
</tr>
<tr>
<td>(0.8,0.0)</td>
<td>37.6</td>
<td>0.332</td>
<td>0.402</td>
</tr>
<tr>
<td>(1.2,0.0)</td>
<td>30.9</td>
<td>0.634</td>
<td>0.777</td>
</tr>
<tr>
<td>(1.5,0.0)</td>
<td>26.4</td>
<td>0.844</td>
<td>0.932</td>
</tr>
<tr>
<td>(0.0,0.4)</td>
<td>52.1</td>
<td>0.223</td>
<td>0.210</td>
</tr>
<tr>
<td>(0.0,0.8)</td>
<td>52.1</td>
<td>0.602</td>
<td>0.646</td>
</tr>
<tr>
<td>(0.0,1.2)</td>
<td>52.1</td>
<td>0.915</td>
<td>0.958</td>
</tr>
<tr>
<td>(0.0,1.5)</td>
<td>52.1</td>
<td>0.986</td>
<td>0.999</td>
</tr>
</tbody>
</table>
HWHE and HWSE exist ([A]), we found that model (II) did not control the level due to length bias sampling, which is caused by dividing it into changed and unchanged groups of employment status. When the HWHE exists but HWSE does not ([B]), all the tests except for model (II) control the significance level well. On the contrary, when the HWSE exists but HWHE does not ([C]), models (I) and (II) not considering HWSE have very large Type I error probabilities, yet the levels of model (III) and (IV) are well controlled. When both HWHE and HWSE exist ([D]), we have the same conclusion as in [C]. Also, we found that the results under heavy censoring are very similar to those under no censoring.

As we mentioned before, models (I) and (II) do not control HWSE at all and so we considered only model (III) in power comparison with model (IV). Also, among four configurations considered in Table 1, we only focused on \((m_{0h}, m_{0u}, m_{1h}, m_{1u}) = (2.0, 2.0, 1.5, 1.5) \) or \((2.0, 1.5, 1.5, 1.0)\) having HWSE regardless of HWHE. Table 2 presents the empirical powers of the Wald test based on model (III) and proposed score test (model (IV)) based on model (1) with an upper 5% Type I error by varying the \((\beta_0, \beta_1)\) pair. We found that the proposed test is more powerful than the Wald test regardless of two different configurations of \((m_{0h}, m_{0u}, m_{1h}, m_{1u})\). Also, similar results were given in the case of heavy censoring.

4. DISCUSSION

We proposed a method to test the relation between exposure and survival in the presence of the healthy worker effect. In order to remove the bias caused by length bias sampling, which comes from the stratification of a changed and unchanged group of employment status, we introduced three conceptual times, such as the waiting time for job changes to occur and survival times depending on whether or not an individual experiences job changes. Then, experiencing job changes was modelled as a phenomenon arising from competing relations between waiting time to job changes and survival time without experiencing job changes. We found from simulations that the upper 5% Type I error was well controlled for both our proposed test and the Wald test based on the Cox model with covariates such as time-dependent exposure amount, health status at hire, and time-dependent employment status. In power comparison, however, when \(\beta_1 = 0\), i.e. there was no relation between exposure and survival after employment status has been changed, the proposed test was more powerful than the Wald test based on the Cox model by varying \(\beta_0\). This came from a difference in the definitions of risk set between model (III) and model (IV). Also, it implied that our risk set is more likely to effectively control the bias owing to length bias sampling. Furthermore, when \(\beta_0 = 0\), i.e. there was no relation between the exposure and survival time without experiencing job changes, the proposed test was also superior to the Wald test.

As mentioned before, the methods for controlling the HWSE that have been developed by researchers (Fox and Collier, 1976; Gilbert, 1982; Gilbert and Marks, 1979; and Steenland and Stayner, 1991) may be considered as data-oriented models rather than theoretical ones. As shown in simulations, however, treating employment status as a time-independent covariate like model (II) or stratification according to employment status resulted in violating the Type I error. This implies that their methods need more detailed theoretical investigation. The proposed method can be easily obtained by adjusting the risk set in the Cox model and requires relatively little computation work as compared to Robins’ methods so that our test is very practical and applicable to real data. Moreover, our method is applicable to data having no follow-up for the exposure after job changes. This can be done only focusing on the inference of \(\beta_0\) with \(\beta_1 = 0\). In fact, in occupational epidemiology studies, real data of this kind are more frequent than those having complete follow-up exposures. In these situations, information loss due to ignoring incomplete data is so unavoidable that we need a further
study on this problem. Finally, our method can be extended to cases having more than two intermediate covariates.

ACKNOWLEDGEMENTS

The authors wish to thank the associate editor and the referees for helpful comments which have led to important improvement of the earlier versions. This work was partially supported by a Korea Research Foundation Grant (KRF-2002-015-CP0055).

REFERENCES


APPENDIX

For simplicity, we first introduce some notations. For $l = 1, 2$

$$e_a^l = \sum_{j=1}^{n} \left\{1 - Z_j(u)\right\} R_j(u) \exp(\gamma_j^l s_j) y_{0j}^l, \quad e_c^l = \sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma_j s_j) y_{1j}^l$$

$$h_u^{\otimes l} = \sum_{j=1}^{n} \left\{1 - Z_j(u)\right\} R_j(u) \exp(\gamma_j \delta_j s_j) s_j^{\otimes l}, \quad h_c^{\otimes l} = \sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma_j s_j) s_j^{\otimes l}$$

$$g_{u11}^l = \sum_{j=1}^{n} \left\{1 - Z_j(u)\right\} R_j(u) \exp(\gamma_j \delta_j s_j) y_{0j} s_j, \quad g_{c11}^l = \sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma_j s_j) y_{1j} s_j$$

$$d_a = \sum_{j=1}^{n} \left\{1 - Z_j(u)\right\} R_j(u) \exp(\gamma_j s_j), \quad d_c = \sum_{j=1}^{n} Z_j(u) R_j(u) \exp(\gamma_j s_j)$$

$$E_u^c = e_a^l / d_a, \quad E_u^h = h_u^{\otimes 1} / d_a, \quad E_c^c = e_c^l / d_c, \quad E_c^h = h_c^{\otimes 1} / d_c$$

$$V_u^a = e_a^l / d_a, \quad V_u^h = h_u^{\otimes 2} / d_a, \quad V_c^c = e_c^l / d_c, \quad V_c^h = h_c^{\otimes 2} / d_c$$

where for a column vector \( \mathbf{a}, \mathbf{a} \odot 1 = \mathbf{a} \) and \( \mathbf{a} \odot 2 = \mathbf{a}^t \). Using the notations, we have the following:

\[
\begin{align*}
\mathbf{f}_1 &= \frac{\partial U_1}{\partial \beta_0} \big|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} = -\sum_{j=1}^{n} \int_{0}^{\infty} \left\{ V_1^e - (E_1^u)^2 \right\} \left\{ 1 - Z_j(u) \right\} \text{d}N_j(u) \\
\mathbf{f}_2 &= \frac{\partial U_2}{\partial \beta_1} \big|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} = -\sum_{j=1}^{n} \int_{0}^{\infty} \left\{ V_2^e - (E_2^v)^2 \right\} Z_j(u) \text{d}N_j(u) \\
\mathbf{f}_3 &= \left\{ \frac{\partial U_3}{\partial \gamma} \bigg|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} \right\}' \\
\mathbf{f}_4 &= \left\{ \frac{\partial U_4}{\partial \gamma} \bigg|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} \right\}' \\
\mathbf{f}_5 &= \frac{\partial U_3}{\partial \gamma} \bigg|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} = -\sum_{j=1}^{n} \int_{0}^{\infty} \left\{ (E_3^c)^2 - (E_3^c)^2 \right\} Z_j(u) \text{d}N_j(u) \\
\mathbf{f}_6 &= \frac{\partial U_4}{\partial \gamma} \bigg|_{\beta_0 = \beta_1 = 0, \gamma, \dot{\gamma}, \mathbf{Q}_0, \mathbf{Q}_0} = -\sum_{j=1}^{n} \int_{0}^{\infty} \left\{ V_3^c - (E_3^h)^2 \right\} Z_j(u) \text{d}N_j(u)
\end{align*}
\]

Hence, \( \sigma_{11} = f_1 - f_3 f_5^{-1} f_3 \) and \( \sigma_{22} = f_2 - f_4 f_6^{-1} f_4 \).