ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 162, 2002

Witold Kupść[*], *Ewa Kawalec*[**], *Bogdan Jasiński*[*]

SOME PRACTICAL PROBLEMS OF MULTIVARIATE SURVIVAL ANALYSIS OF EPIDEMIOLOGICAL STUDIES

ABSTRACT. In the epidemiological analysis of chronic diseases (most often cardiovascular or cancer) the main problem of interest is the estimation of the risk of death (or getting ill) related to set of characteristics called risk factors. For epidemiological studies typical features are:

- large sample size (at least 1000 persons),
- long follow up period for survival analysis (5 or more years),
- large percentage of censored observations (patients who survive the whole time of study, more than 90%),
- large number of registered risk factors.

Some practical problems that concern the statistical analysis of the epidemiological data are following:

- selection of the survival function model,
- selection of the variables included into the model,
- inclusion of interaction and/or higher order effect into the model.

Some solutions of presented problems were applied to the Polish Part of Cardiovascular Diseases Prevention Program (Euro 8202). The program was conducted in 1976–1982 years with long follow up period concerning mortality till 1994 year. The program covered 8603 working men aged 40–59 years in two regions – Warsaw and South-Eastern Poland. Most of statistical analyses were performed on the basis of standard Statistical Analysis System (SAS) package.

Key words: epidemiological study, risk factors, proportional hazards regression, logistic regression.

I. INTRODUCTION

In the epidemiological analysis of chronic diseases (most often cardiovascular or cancer) the main problem of interest is the estimation of the risk of death (or getting ill) related to set of characteristics called risk factors.

[*]Dr, National Institute of Cardiology, Warszawa, Department of Epidemiology and Prevention of CVD.

[**]Dr, Jagiellonian University of Kraków, Institute of Public Health, Collegium Medicum.

For epidemiological studies typical features are:

- large sample size (at least 1000 persons),
- long follow up period for survival analysis, 5 or more years,
- large percentage of patients who survive the whole time of study, (censored observations, more than 90%),
- large number of registered risk factors.

Some practical problems that concern the statistical analysis of the epidemiological data are following:

- selection of the survival function model,
- selection of the variables included into the model,
- inclusion of interaction and/or higher order effect into the model.

II. THE WORKING EXAMPLE

Some solutions of presented problems were applied to the Polish Part of Cardiovascular Diseases Prevention Program (Euro 8202). Details of the Program are described elsewhere – WHO Collaborative Group (1986), R y w i k et al. (1975).

The Program was conducted in 1976–1982 years and after it the long follow up period concerning mortality was performed till 1994.

The Program covered 8603 working men aged 40–59 years in two regions – Warsaw and South-Eastern Poland.

The analysed primary outcome was total mortality divided into three groups of causes of death: cardiovascular diseases, cancer and death from all other causes.

The following 11 individual risk factors were included into analyses:

- age (AGE) years,
- smoking habit (SMOKE) 0 – no, 1 – yes,
- systolic blood pressure (SBP) per 10 mmHg,
- diastolic blood pressure (DBP) per 10 mmHg,
- total cholesterol level (CHOL) mg/dl,
- body mass index (BMI) (weight kg)/ (height m)^2,
- physical activity at work (more than 50% of work time PHAC) 0 – no, 1 – yes,
- marital status (MARIT) 0 – married, 1 – other,
- diabetes (DIAB) 0 – no, 1 – yes,
- family history of CVD (FHIST) 0 – no, 1 – yes,
- coughing up of the phlegm (COUGH) 0 – no, 1 – yes,
- region (REG) 0 – South-East Poland, 1 – Warsaw.

Most of statistical analyses were performed on the basis of standard Statistical Analysis System (SAS) package (SAS Technical Report P–217 (1991)).

The main results concerning the baseline values of analysed risk factors and distribution of deaths registered within period of 18 years are presented in Table 1 and Table 2 respectively.

Table 1

Mean (fraction), SD of risk factors at baseline

Risk factor	Warsaw (N = 5562)		South-East Poland (N = 3041)	
	mean	SD	mean	SD
TIME	15.85	3.27	18.40	3.84
AGE (years)	47.40	5.11	46.35	4.69
SMOKE (0 – no, 1 – yes)	0.61	0.49	0.68	0.47
SBP (mmHg)	135.48	18.05	125.98	16.58
DBP (mmHg)	87.03	10.99	80.14	10.55
CHOL (mg/dl)	207.33	38.09	187.31	36.81
BMI (kg/m^2)	26.31	3.40	24.84	3.25
PHAC (0 – no, 1 – yes)	0.92	0.28	0.66	0.48
MARIT (0 – married, 1 – other)	0.08	0.28	0.04	0.19
DIAB (0 – no, 1 – yes)	0.02	0.14	0.01	0.08
FHIST (0 – no, 1 – yes)	0.25	0.43	0.18	0.38
COUGH (0 – no, 1 – yes)	0.43	0.49	0.59	0.49

Table 2

Distribution of registered deaths according to cause and region

Region		Cause of death (VIII and IX ICD Rev.)			
		CVD 390–459	Cancer 140–209	All others	Total
Warsaw	N_c	445	284	306	1035
	%	43.0	27.4	29.6	100.0
South-East	N_c	343	194	124	661
	%	51.9	29.4	18.7	100.0

III. SELECTION OF THE RISK FUNCTION MODEL

Formulation of two most popular survival models analysed in epidemiological studies may be following: there are N individuals in a cohort, each observed with a set of k potential risk factors:

$$\mathbf{z}' = (z_1, z_2, \ldots, z_k) \tag{1}$$

measured at the beginning of the follow-up time period of length T. The response variable Y is defined by:

$$Y = \begin{cases} 1 & \text{if the disease or death occured during time } T \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

1) The logistic model of risk factors z with coefficients β_0 and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$ is given as follows:

$$Pr(Y = 1 \mid \mathbf{z}, T) = \{1 + \exp(\beta_0 - \boldsymbol{\beta}'\mathbf{z})\}^{-1} \tag{3}$$

and corresponding probability of survival of the follow-up period is

$$Pr(Y = 0 \mid \mathbf{z}, T) = S(T \mid \mathbf{z}, \boldsymbol{\beta}) = \exp(\beta_0 - \boldsymbol{\beta}'\mathbf{z})\{1 + \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{z})\}^{-1} \tag{4}$$

2) The proportional hazard Cox's model is based on a hazard rate having the form

$$\lambda(t, \mathbf{z}) = \lambda_0(t)\exp(\boldsymbol{\alpha}'\mathbf{z}) \tag{5}$$

with the survival function given by

$$S(T \mid \mathbf{z}, \boldsymbol{\alpha}) = \exp[-\int_0^T \lambda_0(t)\exp(\boldsymbol{\alpha}'\mathbf{z})dt] \tag{6}$$

In epidemiology the association between outcome Y and the given risk factor z is measured in logistic regression by odds ratio OR, defined for two different levels z_1 and z_0:

$$OR = P_1/(1-P_1) : P_0/(1-P_0) \tag{7}$$

which in logistic regression is expressed by $OR = \exp(\beta)$.

In the Cox's model values of $\exp(\alpha)$ are interpreted as hazard rate (HR):

$$HR = h_1(t)/h_0(t) \tag{8}$$

Green, Symons (1983) investigated the problem of relationship between a vector of logistic regression coefficients β and proportional hazard coefficients α and the main results are following:
for logistic model if we denote

$$u = \exp(-\beta_0 - \boldsymbol{\beta}'\mathbf{z}) \tag{9}$$

then

$$-\ln(S(\mathrm{T} \mid \mathbf{z}, \boldsymbol{\beta}) = \ln(1+\mathrm{u})/\mathrm{u} \tag{10}$$

and by linearized Taylor series expansion

$$\ln(1+u)/u \cong 1/u \tag{11}$$

if u^{-1} is small. The approximation

$$-\ln(S(T \mid \mathbf{z}, \boldsymbol{\beta})) \cong \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{z}) \tag{12}$$

holds for $\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{z})$ small. From the other hand if for Cox model the underlying hazard

$$\lambda_0(t) = \mathrm{const} = \lambda \tag{13}$$

then

$$\int_0^{\mathrm{T}} \lambda_0(t)\exp(\boldsymbol{\alpha}'\mathbf{z})dt = \lambda \mathrm{T}\exp(\boldsymbol{\alpha}'\mathbf{z}) \tag{14}$$

and the logistic and Cox model survival functions are approximately equivalent when

$$\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{z}) \cong \lambda \mathrm{T}\exp(\boldsymbol{\alpha}'\mathbf{z}) \tag{15}$$

and the natural correspondence is suggested: $\beta_0 = \ln(\lambda T)$ and $\beta_i \cong \alpha_i$.

The relation holds for small λ, T and β_i i *e.* for rather rare disease intensity, short period of follow-up and not too large effects of risk factors. Estimated values of hazard rates for Cox's model are presented in Table 3 and odds ratios for logistic regression in Table 4.

Table 3

Cox's hazard rates

RISK FACTOR	SOUTH-EAST	WARSAW	COMBINED
AGE	1.07[b]	1.09[b]	1.08[b]
SMOKE	1.66[b]	2.07[b]	1.93[b]
SBP	1.13[b]	1.10[b]	1.11[b]
DBP	1.05	1.12[b]	1.10[b]
CHOL	1.01	1.00	1.01
BMI	0.99	0.98[a]	0.98[a]
PHAC	1.05	0.84	0.96
MARIT	0.80	1.47[b]	1.26[b]
DIAB	2.59[b]	2.27[b]	2.26[b]
FHIST	1.13	1.00	1.03
COUGH	1.15	1.15[a]	1.16[b]
REG	–	–	1.18[b]

[a] $p < 0.05$, [b] $p < 0.01$

Table 4.

Logistic regression – odds ratios

RISK FACTOR	SOUTH-EAST	WARSAW	COMBINED
AGE	1.10[b]	1.11[b]	1.10[b]
SMOKE	1.98[b]	2.36[b]	2.22[b]
SBP	1.13[b]	1.12[b]	1.12[b]
DBP	1.10	1.14[a]	1.12[b]
CHOL	1.01	1.00	1.01
BMI	0.99	0.98	0.98[a]
PHAC	1.11	0.78[a]	0.98
MARIT	0.97	1.51[b]	1.37[b]
DIAB	2.53[b]	2.52[b]	2.51[b]
FHIST	1.13	0.95	1.00
COUGH	1.15	1.15	1.16[a]
REG	–	–	0.63[b]

[a] $p < 0.05$, [b] $p < 0.01$

IV. SELECTION OF VARIABLES

In each statistical package for regression analysis, linear or non-linear, there are several procedures concerning the selection of the „best subset" $q \leq k$ variables into the model. Selection procedures cover two typical classes: forward and backward selection K l e i n b a u m (1996).

The forward selection starts with one variable (or few preselected) and adds sequentially at each step one with the smallest value of p for testing of appropriate hypothesis, continuing procedure till all variables with $p < \alpha$ will be included into the model.

The backward selection starts with the whole set of k variables and excludes step by step one with largest p value till only the variables with $p < \alpha$ will stay in the model.

Criterion of including (or excluding) a variable in (or from) the model in Cox's or logistic regression packages is most often a p value for testing the hypothesis of investigated α_i or β_i coefficient for variable i: H_0: $\alpha_i = 0$ or $\beta_i = 0$. The stopping rule presumes that all variables selected into model showed $p < \alpha$ with typical $\alpha = 0.05$.

In epidemiological investigation the recommended procedure is a backward one. It allows to estimate the impact of all analysed explanatory variables on the outcome and to compare it with the „best selected subset".

T a b l e 5.

Results of PHREG ANALYSIS for full model and after backward selection procedure (total mortality combined region)

RISK FACTOR	FULL MODEL	AFTER SELECTION
AGE	1.09[b]	1.08[b]
SMOKE	2.07[b]	1.98[b]
SBP	1.10[b]	1.11[b]
DBP	1.12[b]	1.10[b]
CHOL	1.00	–
BMI	0.98[a]	–
PHAC	0.84	–
MARIT	1.47[b]	1.29[b]
DIAB	2.27[b]	2.22[b]
FHIST	1.00	–
COUGH	1.15[a]	1.14[b]

[a] $p < 0.05$, [b] $p < 0.01$

In epidemiology some of explanatory variables should be included into the model regardless of the result of the statistical test. However generally the variables significant in the full model are also included in the „best subset". Results of reduction of initial full model including 11 variables by SAS procedure PHREG with backward selection are presented in Table 5.

V. EFFECT OF INTERACTION AND HIGH ORDER COMPONENT

Interaction in logistic and Cox's regression models is defined as multiplicative one. For hazard ratio we have:

$$HR(t, z_1) = \exp(\alpha_1),\ HP(t, z_2) = \exp(\alpha_2)$$

$$HR(t, z_1 z_2) = \exp(\alpha_1 + \alpha_2 + \delta)$$

and

$$HR(t, z_1 z_2)/[HR(t, z_1) \cdot HR(t, z_2)] = \exp(\delta)$$

The multiplicative interaction does not hold when:

$$HR(t, z_1 z_2) = HR(t, z_1) \cdot HR(t, z_2) \text{ or } \delta = 0$$

When interactions are included into model then the formulae for HR contain coefficients δ_i and corresponding values of variables z_i.

The standard error $S_E(l)$ for the estimator l of L is calculated as error of linear combination of errors of estimators b_i and d_i of parameters β_i and δ_i

$$S_E^2(l) = \text{var}(l) = \text{var}(b) + \Sigma(z_i)^2 \text{var}(d_i) + 2\Sigma z_i \text{cov}(b, d_i)$$

what allows to calculate the confidence limits (H_L, H_U) for HR:

$$HR_L = \exp\{b + d_1 z_1 + \ldots + d_r z_r - u_\alpha S_E(l)\}$$

$$HR_U = \exp\{b + d_1 z_1 + \ldots + d_r z_r + u_\alpha S_E(l)\}$$

where u_α – quantile of order $1 - \alpha/2$ of normal distribution N(0, 1).

Analysis of interaction REGION*SMOKE is presented in Table 6.

Table 6

An example of analysis of interaction of risk factors with region term [a]

Variable	Parameter Estimate	Standard Error	p	HR	HR_L	HR_U
REG	0.4558	0.5468	0.4044	1.58	0.54	4.61
...	...	...	...	...	...	...
SMOKE	0.5201	0.0973	0.0001	1.68	1.39	2.04
REG*SMOKE	0.2166	0.1210	0.0734	1.24	0.98	1.57

[a] Adjusted for remaining 9 risk factors

The covariance matrix of estimates which elements were applied for calculation of confidence limits is presented in Table 7.

Table 7.

Covariance matrix of estimates

	REG	SMOKE	REG*SMOKE
REG	0.2989	0.0146	– 0.0221
SMOKE		0.0095	– 0.0093
REG*SMOKE			0.0146

In presence of interaction REG*SMOKE effect of SMOKE is evaluated as:

$HR_{SMK} = \exp(0.5201 + 0.2166\ \mathrm{REG})$

For South-East Poland (REG = 0) the result is

$HR_{SMK.S\text{-}E} = 1.68,\ HR_L = 1.39,\ HR_U = 2.04$

For Warsaw (REG = 1) we have

$S_E^2\ (\mathrm{SMK.War}) = 0.0095 + 0.0146 + 2\cdot(-0.0093) = 0.0055$

$S_E = 0.074,\ u_{0.975}\cdot S_E = 0.145$

and $HR_{SMK.War} = \exp(0.7367) = 2.09,\ HR_L = 1.81,\ HR_U = 2.42$

In similar manner the curvilinear effect of variable BMI was additionally analyzed. The quadratic effect was introduced into the model:

	Parameter Estimate	Standard Error	p	HR	HR_L	HR_U
BMI	– 0.3482	0.0592	0.0001	0.71	0.63	0.79
BMI^2	0.0062	0.0011	0.0001	1.01	1.00	1.02

The hypothesis H_0: $\beta_{BMI\wedge 2} = 0$ was tested and rejected and hazard ratio for $\exp(\beta_{BMI} + \beta_{BMI\wedge 2})$ and its confidence interval was calculated:

$HR = 0.71$, $HR_L = 0.63$, $HR_U = 0.80$

VI. CONCLUSIONS

1. The Cox's proportional hazards model is most popular and most informative for survival analyzing. It allows estimating the whole survival curve as well as the influence of risk factors on selected outcome end points. However, if the main problem of interest is the estimation of the risk factors impact, then logistic regression model may be used quite appropriate. The logistic model is not influenced by large number of censored observations at the same time (time at the end of the study) and obtained estimates of odds ratios are easy to interpret in medical applications.
2. The backward selection of the variables included in the regression model allows to estimate the changes of the goodness of fit measures and relation among analyzed variables in comparison between full and reduced model. Possibility of fixing some variables in the model enables to analyze their importance regardless of the results of the tests.
3. The effect of multiplicative interaction is easy to estimate in standard statistical packages. However the interpretation of the interaction is difficult, especially for continuous variables, where selection of points to calculate the hazard rates for interaction is quite arbitrally.

REFERENCES

World Health Organization European Collaborative Group: European Collaborative Trial of Multifactorial Prevention of Coronary Heart Disease. Final report on the 6-year results, „The Lancet", 1986, 869–872.

R y w i k S., K o r e w i c k i J., M i k o ł a j c z y k W. et al. (1975), *Methodology of Polish Prevention Trial on Cardiovascular Disease Epidemiology. (in Polish)*, „Przegl. Lek", **32**, 510.

SAS Technical Report P-217, SAS/STAT Software: The PHREG Procedure, Version 6 (1991). SAS Institute Inc., 6, Cary, NC.

G r e e n M. S., S y m o n s M. J. (1983), *A comparison of the Logistic Risk Function and the Proportional Hazards Model in Prospective Epidemiological Studies*, „J. Chron. Dis.", **36**, 715–724.

K l e i n b a u m D. G. (1996), *Survival Analysis*, Springer.

Witold Kupść, Ewa Kawalec, Bogdan Jasiński

NIEKTÓRE PRAKTYCZNE PROBLEMY WIELOZMIENNEJ ANALIZY PRZEŻYĆ W BADANIACH EPIDEMIOLOGICZNYCH

Jednym z głównych celów epidemiologicznych badań nad chorobami przewlekłymi (najczęściej układu krążenia lub nowotworowymi) jest oszacowanie ryzyka zachorowania lub zgonu w zależności od zespołu cech – czynników ryzyka.

Badania epidemiologiczne charakteryzują się najczęściej następującymi własnościami:

- duża liczebność próby, powyżej 1000 badanych;
- długi okres obserwacji badanych osób, ponad kilka lat;
- wysoka frakcja (ok. 90%) osób, które przeżyły cały okres badania bez incydentu chorobowego, tzw. cenzorowanie administracyjne;
- duża liczba czynników ryzyka rejestrowanych w badaniu.

Analiza statystyczna badania epidemiologicznego wymaga, między innymi, rozwiązania następujących problemów:

- wybór modelu funkcji oceniającej ryzyko,
- selekcja badanych w modelu czynników ryzyka,
- ocena wzajemnego oddziaływania (interakcji) badanych czynników i ocena nieliniowych efektów ich oddziaływania.

Rozwiązanie przedstawionych zadań przeprowadzono na przykładzie analizy wyników Polskiego Programu Prewencji Chorób Układu Krążenia przeprowadzonego w latach 1976–1982, obejmującego 8603 mężczyzn zatrudnionych w zakładach pracy w dwóch regionach Polski – Warszawy i Polski Południowo-Wschodniej i rozszerzonego o obserwację postępującą w zakresie zgonu do roku 1994.