ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

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

ON ROBUST INFERENCE FOR THE COX MODEL – THE COXROBUST PACKAGE

ABSTRACT. A methodology supporting the COXROBUST package, designed for the robust inference in the Cox regression model, is presented. Basic functions of the package and its data analytic capabilities are described.

Key words: robust estimation, survival data analysis.

I. INTRODUCTION

The statistical relationship between survival time T and the vector of covariates Z, in the Cox proportional hazards model, is described by the *hazard function*, which is defined for fixed $Z = z$ as

$$\lambda(t, z) = \lambda_0(t)\exp(z\beta),$$

where $\lambda_0(t)$ is the baseline hazard function and $\beta \in R^k$ is a vector of regression parameters. A standard statistical inference for the above model was built up in several important steps.

Cox's (1972, 1975) proposed a method of estimation of the regression parameters β which consists in maximization of the partial likelihood function and it is equivalent to solving the following score function equation:

$$\sum_{i=1}^{n}\left[Z - \frac{\sum_{T_j \geq T_i} Z_j \exp(\beta' Z_j)}{\sum_{T_j \geq T_i} \exp(\beta' Z_j)}\right] I_{T_i \leq C_i} = 0,$$

[*] Professor, Institute of Economic Sciences, Wrocław University.
[**] Msc, Institute of Economic Sciences, Wrocław University.

where T_i, C_i, Z_i denote respectively the time, censoring and covariate variables in the sample. Breslow (1974) gave an estimator of the cumulated hazard function $\Lambda(t) = \int_0^t \lambda_0(u)du$ based on the Cox partial likelihood estimation. Finally a number of goodness of fit testing methods of the model were proposed, including the standard significance testing based on the partial likelihood.

The following sections indicate basic steps of the robust methodology used for the package COXROBUST and describe its basic functions in the process of statistical inference for survival data.

II. ROBUST METHODOLOGY

It was shown in Samuels (1978) and in Bednarski (1989) that inference for the Cox model based on the partial likelihood, can be very sensitive to even small departures from the model (see also Ried and Crépeau (1985)).

Bednarski (1993), using the notion of Fréchet differentiability of statistical functionals, proposed a robust method of estimation of regression coefficients based on a modification of the partial likelihood estimator. The method, leading to a functional-estimator $\beta(F_n)$ consists in solving the equation:

$$L(F_n, \beta, A) = \int A(w,y)\left[y - \frac{\int A(w,z) z I_{a \wedge t \geq w} \exp(\beta' z) dF_n(t,a,z)}{\int A(w,z) I_{a \wedge t \geq w} \exp(\beta' z) dF_n(t,a,z)} \right] I_{w \leq c} dF_n(w,c,y) = 0,$$

where F_n denotes the empirical distribution function of the sample $(T_1, C_1, Z_1), \ldots, (T_n, C_n, Z_n)$, t is the time failure variable, c is the censoring variable and z is the covariate k-dimensional vector. If weights are equal 1 then the method reduces to the partial likelihood estimation. The differentiability property ensures that the following expansion:

$$\sqrt{n}(\beta(F_n) - \beta(F)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_{\beta_0, A}(X_i) + o_P(1),$$

where $\psi_{\beta_0, A}(x)$ is the influence function of the robust estimator, holds for the Cox model distribution F with true parameter value β_0 and for its infinitesimal neighborhood.

Grzegorek (1993), using a similar functional oriented methodology, proposed a robust estimator of the cumulated hazard. His estimator was a suitably weighted variant of the Breslow estimator:

$$\Lambda_{A,F_n}(t) = \int \frac{A(w,y) I_{w \le \min(t,c)}}{\int A(w,z) I_{w \le \min(u,a)} \exp\left(\beta(F_n)' z\right) dF_n(u,a,z)} dF_n(w,c,y).$$

Minder and Bednarski (1996) proposed a heuristic method of goodness of fit testing for the Cox regression model. It is based on a standardized difference of the Kaplan Meier estimator and an estimator of survival function following from the Cox regression model. Finally Bednarski and Borowicz (2007) studied the behavior of a Wald significance test based on the robust estimator $\beta(F_n)$.

III. THE PACKAGE COXROBUST

The package was built up as an extension of the standard R language SURVIVAL package designed for the survival data analysis. The type of functions of the package and mode of their use are fully compatible with the R standards. The package consists of the following three basic functions for the data analysis:

- The function coxr(formula, data, subset, na.action, trunc = 0.95, f.weight = c("linear", "quadratic", "exponential"), singular.ok = TRUE, model = FALSE), estimating the regression parameters of the model, where the arguments have the following meaning:

o formula – a formula object, with the response on the left of a ~ operator, and the terms on the right. The response must be a survival object as returned by the Surv function.

o data – a data frame in which to interpret the variables named in the formula, or in the subset.

o subset – expression saying that only a subset of the rows of the data should be used in the fit.

o na.action – a missing-data filter function, applied to the model.frame, after any subset argument has been used.

o trunc – roughtly, quantile of the sample T_i exp(β'Z_i), it determines the trimming level for the robust estimator

o f.weight – type of weighting function, default is "quadratic"

o singular.ok – logical value indicating how to handle collinearity in the model matrix. If TRUE, the program will automatically skip over columns of the X matrix that are linear combinations of earlier columns. In this case the coeffi-

cients for such columns will be NA, and the variance matrix will contain zeros. For ancillary calculations, such as the linear predictor, the missing coefficients are treated as zeros.

o model – a logical value indicating whether model frame should be included as a component of the returned value.

- The plot function plot(x, caption = c("Full data set", "First quartile", "Second quartile", "Third quartile", "Fourth quartile"), main = NULL, xlab = "log time", ylab = "standardized survival differences", ..., color = TRUE), where

o x – coxr object, typically result of coxr,
o caption – captions to appear above the plots,
o xlab – title for the x axis,
o ylab – title for the y axis,
o main – overall title for the plot,
o ... other parameters to be passed through to plotting functions,
o color – if FALSE grayscale mode is used.

- A function for generation of non-contaminated and contaminated samples from the Cox regression model: gen_data(n, beta, cont = 0, p.censor = 0) with arguments

o n – number of observations.
o beta – vector of regression coefficients.
o cont – fraction of contaminated observations.
o p.censor – probability of censoring.

A detailed description of the numerous values and components returned by the above functions can be found in package documentation (http://cran.r-project.org/doc/packages/coxrobust.pdf). The coxr function uses results from Bednarski (1993), Grzegorek (1993) and from Bednarski and Mocarska (2006) for the estimation of the regression parameter and from Bednarski and Borowicz (2007) for significance testing. The plot function is based on Minder and Bednarski (1996) proposition of goodness of fit assessment for the Cox model.

IV. EXEMPLARY USE OF THE PACKAGE

The data in the following example were generated using the gen_data function. A sample of size 200 for $\beta = (1,0.1,2)$ with censoring frequency 0.3 and 5% contamination was generated into data frame "a".

```
> a <- gen_data(200, c(1, 0.1, 2), cont = 0.05, p.censor = 0.30)
> result <- coxr(Surv(time, status) ~ X1 + X2 + X3, data = a, trunc = 0.9)
> print(result)

Call:
coxr(formula = Surv(time, status) ~ X1 + X2 + X3, data = a, trunc = 0.9)

Partial likelihood estimator
       coef exp(coef) se(coef)        p
X1 0.340953      1.41   0.0809 2.52e-05
X2 0.000602      1.00   0.0761 9.94e-01
X3 0.791971      2.21   0.0844 0.00e+00

Wald test=172 on 3 df, p=0

Robust estimator
    coef exp(coef) se(coef)        p
X1 0.836      2.31    0.233 3.36e-04
X2 0.244      1.28    0.102 1.74e-02
X3 1.642      5.16    0.335 9.45e-07

Extended Wald test=43.6 on 3 df, p=1.79e-09
```

Then the generated data in the data frame were used for the inference using the coxr function. The standard output consists of estimation results of the regression coefficients for the partial likelihood estimator and for the robust estimator. Both estimations are supplemented by the p-values of the Wald test. One can see that estimation results, though strikingly different for the two methods, under slightly contaminated data indicate significance of the regression variables in both cases. Let us notice, however, that the partial likelihood method gives unacceptable discrepancy with the true parameter values, which is not the case for the robust method. To see better the fit of the model under the partial likelihood and robust estimation we recommend the use of the plot function. The function shows four strata of "possibly" equal size, resulting from the sample ordered by the value of the linear predictor and defined by corresponding quartiles.

For each stratum a standardized difference between the nonparametric Kaplan Meier estimator – a natural reference survival function – and an estimator of the survival function resulting from the Cox model is computed and plotted for the non-robust (black color) and robust (gray color). The X axis, where the log times of the observed values are marked, correspond to the zero difference of the two estimates of the survival function. Each plot contains the 95% confidence intervals. We can easily see that the robust method leads to better estimation both in the domain of the regression parameters and in the domain of the survival functions under the contamination.

```
> plot(result)
```

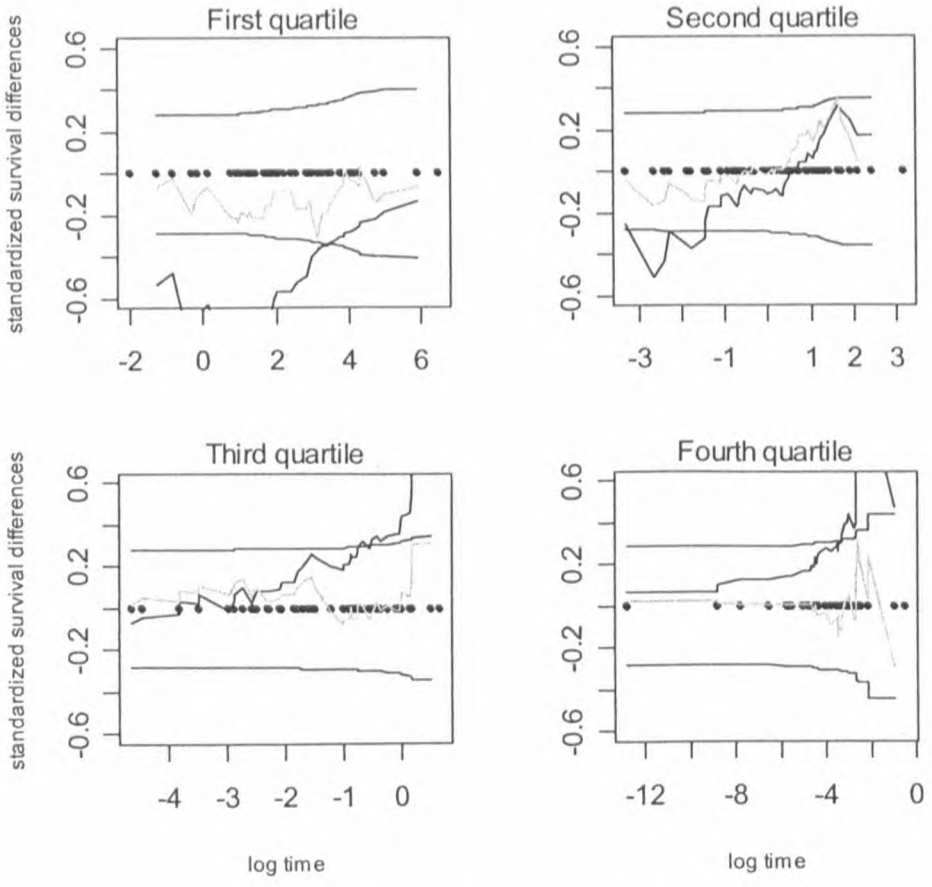

Fig. 1. Standardized differences between the nonparametric Kaplan Meier estimator and the estimator of survival function resulting from the Cox model

Analysis of real data requires first reading the data from a file into the data frame. Then all the inference steps are the same as above.

REFERENCES

Bednarski T. (1989), On sensitivity of Cox's estimator, *Statistics and Decisions*. 7, 215–228.

Bednarski T. (1993), Robust estimation in Cox's regression model, *Scandinavian Journal of Statistics* 20, 213–225.

Bednarski T., Mocarska E. (2006), On robust model selection within the Cox model, *Econometrics Journal*, Vol. 9, Issue 2, 279–290.

Breslow N.E. (1974), Covariance analysis of censored survival data, *Biometrics* 30, 579–594.

Cox D. (1975), Partial likelihood, *Biometrika* 62(2), 269–276.

Cox, D.R. (1972), Regression models and life tables, *Journal of the Royal Statistical Society*, Series B, 34, 187–220.

Grzegorek K. (1993), On robust estimation of baseline hazard under the Cox model and via Fréchet differentiability, Preprint of the Institute of Mathematics of the Polish Academy of Sciences, 518.

Krug B. (1998), Robust Estimation in Selected Additive Models, Institute of Mathematics of the Polish Academy of Sciences, PhD thesis.

Lin D.Y. (1991), Goodness-of-fit analysis for the Cox regression model based on a class of parameter estimators, *J. Amer. Statist. Assoc* 86, 725–729.

Minder Ch. and Bednarski T. (1996), A robust method for proportional hazards regression, *Statistics in Medicine* 15, 1033–1047.

Reid N. and Crépeau H. (1985), Influence functions for proportional hazards regression, *Biometrika* 72, 1–9.

Samuels S. (1978), Robustness for survival estimators, Unpublished Ph.D. thesis. Dept. of Biostatistics, Univ. of Washington.

Tadeusz Bednarski, Filip Borowicz

ODPORNY TEST ISTOTNOŚCI DLA MODELU COXA – TEORIA I ZASTOSOWANIA

Pierwsza części wykładu będzie prezentacją programu statystycznego służącego odpornej estymacji w modelu Cox'a – aplikacja opracowana na podstawie metody T. Bednarskiego (1993). Program ten została dołączony do zestawu pakietów statystycznych języka R, budowanego w ramach otwartego projektu.

Teoretyczna część wykładu poświęcona będzie zagadnieniu odpornej weryfikacji istotności modelu Cox'a. Przedstawione będą wyniki analityczne związane z granicznym rozkładem stosownie zmodyfikowanej statystyki Walda – bazującej na odpornej estymacji parametrów regresji w modelu Cox'a. Ponadto zaprezentowane będą wyniki Monte Carlo umożliwiające ocenę stopnia bliskości rozkładu asymptotycznego z wynikami empirycznymi dla skończonych prób. Zaproponowana statystyka testowa została dołączona do wspomnianego wcześniej pakietu statystycznego.