COMPUTER APPROACH IN CONSTRUCTING SURVIVAL REGRESSION MODELS AND TESTING ADEQUACY

E.V. Chimitova, M.M. Remneva, M.A. Vedernikova
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: chim@mail.ru

Abstract

The paper is devoted to the problems of constructing proportional hazard Cox model and testing goodness-of-fit. The semiparametric and various parametric proportional hazard models have been considered. The goodness-of-fit test statistic distributions by samples of residuals have been investigated by the Monte-Carlo method. The problem of testing goodness-of-fit by randomly censored samples has been discussed. The Kolmogorov-Smirnov, Cramer-von Mises-Smirnov and Anderson-Darling goodness-of-fit tests have been considered.

1 The proportional hazard Cox model

The main objective in many studies is to understand and exploit the relationship between lifetime and covariates. Data often include covariates that might be related to lifetime: for example, in a survival study for lung cancer patients (data are given in [2]) include such factors as the age, general condition of the patient, the type of tumors, the number of months from diagnosis of lung cancer, the type of chemotherapy treatment and the prior therapy.

Suppose that each individual in a population has a lifetime T_x under a vector of covariates $x = (x_1, x_2, ..., x_m)^T$. Let denote by $S_x(t)$ the survival function which is defined as $S_x(t) = P(T_x \geq t) = 1 - F_x(t)$, the hazard rate function $\lambda_x(t) = \lim\limits_{\Delta t \to 0} \frac{P(t \leq T_x \leq t + \Delta t | T_x \geq t)}{\Delta t} = \frac{f_x(t)}{S_x(t)}$, and we denote by $\Lambda_x(t) = \int\limits_0^t \lambda_x(u)du = -\ln(S_x(t))$ the cumulative hazard rate function of T_x.

In medical and epidemiological studies, lifetimes are typically right censored. The observed data are usually presented as $(t_1, \delta_1), ..., (t_n, \delta_n)$, where $\delta_i = 1$ if t_i is an observed lifetime, and $\delta_i = 0$ if t_i is censoring time which means that lifetime of i-th individual is greater than t_i.

The best known lifetime regression model is the proportional hazard (PH) model introduced by Cox [1], which can be written as following

$$\Lambda_x(t, \beta) = r(x, \beta)\Lambda_0(t), \tag{1}$$

where $\Lambda_0(t)$ is the baseline hazard rate function, $r(x, \beta)$ is a nonnegative function and β is a vector of regression parameters. Various parameterizations of the function

$r(x,\beta)$ can be used in practice such as the linear form $r(x,\beta) = 1 + \beta' x$, logistic form $r(x,\beta) = \ln(1 + e^{\beta' x})$ and the most commonly used logarithmically linear form is

$$r(x,\beta) = e^{\beta' x} \tag{2}$$

If the baseline hazard function $\Lambda_0(t)$ is unknown we have the semiparametric Cox model. Unknown parameters β in case of parameterization (2) can be estimated maximizing the partial log-likelihood function [1]

$$\ln\left(\tilde{L}(t,\beta)\right) = \sum_{i=1}^{n} \delta_i \left[\beta' x^i - \ln\left(\sum_{j:\, t_j \geq t_i} \exp\left(\beta' x^j\right)\right)\right], \tag{3}$$

where x^i is the covariate vector associated with the i-individual. Nonparametric estimate of $\Lambda_0(t)$ corresponding to the Cox model (1) can be written as

$$\hat{\Lambda}_0(t) = \sum_{i:\, t_i \leq t} \left[\delta_i \Big/ \sum_{j:\, t_j \geq t_i} \exp\left(\beta' x^j\right)\right].$$

Fully parametric PH models specify in (1) both $r(x,\beta)$ and $\Lambda_0(t,\theta)$ parametrically Various parametric families of models are used in survival analysis. Among them the exponential, Weibull, log-normal, inverse Gaussian, gamma distributions occupy the central position because of their demonstrated usefulness in a wide range of situations For example, the Weibull PH model can be written as

$$\Lambda_x(t,\beta,\theta) = r(x,\beta) \cdot \left(\frac{t}{\theta_1}\right)^{\theta_0}. \tag{4}$$

Unknown parameters β and θ can be estimated maximizing the log-likelihood function

$$\ln\left(L(t,\beta,\theta)\right) = \sum_{i=1}^{n} \left[\delta_i \ln\left(r(x^i,\beta) \cdot \lambda_0(t_i,\theta)\right) + r(x^i,\beta) \ln S_0(t_i,\theta)\right], \tag{5}$$

which for Weibull PH model with parameterization (2) becomes

$$\ln\left(L(t,\beta,\theta)\right) = \sum_{i=1}^{n} \left[\delta_i \left(\beta' x^i + \ln\left(\frac{\theta_0 t_i^{\theta_0 - 1}}{\theta_1^{\theta_0}}\right)\right) - \left(\frac{t_i}{\theta_1}\right)^{\theta_0} \cdot e^{\beta' x^i}\right].$$

2 Testing goodness-of-fit with parametric PH models

After estimating unknown model parameters the goodness-of-fit hypothesis for the model should be tested. It is an essential part of statistical analysis, because if the model is not appropriate than conclusions made on the basis of obtained model would be incorrect. There are various methods for testing goodness-of-fit of data to a probability model. One approach to testing goodness-of-fit with parametric PH model is based on residuals such as the cumulative hazard (exponential) residuals

$$\hat{R}_i = r(x^i,\hat{\beta}) \cdot \Lambda_0\left(t_i,\hat{\theta}\right). \tag{6}$$

If the model is appropriate the $\hat{R}_1, \ldots, \hat{R}_n$ is a censored sample of the standard exponential distribution. The hypothesis about goodness-of-fit of the sample of residuals (6) to the standard exponential distribution can be tested with the classical Kolmogorov-Smirnov, Cramer-von Mises-Smirnov, Anderson-Darling tests. It should be noted that we have a composite hypothesis, for which test statistic distributions $G(S|H_0)$ are affected by a number of factors: the form of assuming lifetime distribution $F_0(t,\theta)$; the type and the number of estimated parameters, the method of parameter estimation and other factors [4, 5]. So, approximate p-values can be obtained by simulation.

The purpose of this paper is to investigate test statistic distributions $G(S|H_0)$ by computer simulation methods for various parametric families. Let us consider at first an uncensored sample case.

In [4, 5] the approximations of statistic distribution models and the tables of percentage points were obtained for testing composite hypotheses by the Kolmogorov-Smirnov, Cramer-von Mises-Smirnov, Anderson-Darling tests using the maximum likelihood estimates of unknown parameters. In this paper we have investigated statistic distributions in testing goodness-of-fit of samples of residuals (6) to the standard exponential distribution. It has been shown that test statistic distributions turn out to be strongly dependent on both the kind of $r(x,\beta)$ and chosen parameterization $\Lambda_0(t,\theta)$. For example, in figure 1 there are simulated Kolmogorov statistic distributions $G(S_k|H_0)$ when testing goodness-of-fit with the Weibull PH model (4). There are three distributions of the Kolmogorov statistic for various functions $r(x,\beta)$. It should be noted that the Kolmogorov statistic distribution for the logarithmically linear form (2) coincides with the approximation of statistic distribution obtained in [4, 5] for samples without covariates.

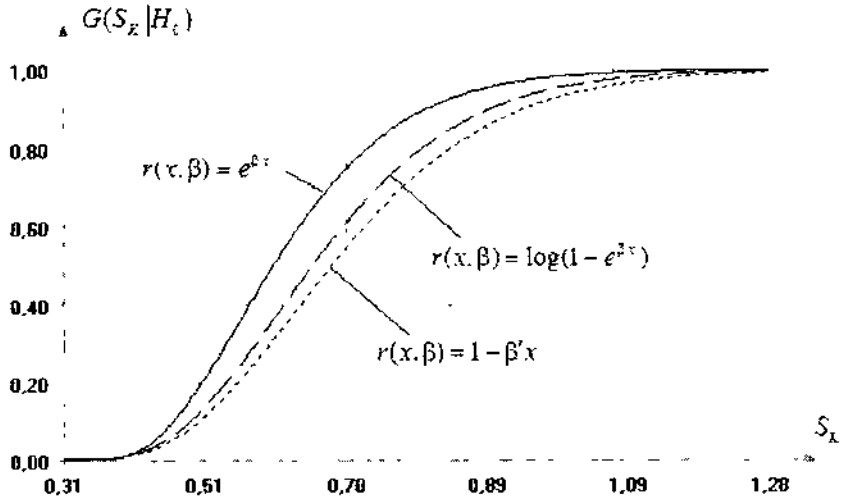

Figure 1: The Kolmogorov test statistic distributions

A similar result has been obtained for Cramer-von Mizes-Smirnov and Anderson-Darling tests.

In case of censored samples approximate p-values in testing goodness-of-fit can be obtained by simulation only if there is sufficient knowledge of the censoring process. It is quite possible if we have type I or type II censoring, but in case of random censoring process which often occurs in survival analysis there is a problem of ambiguity in simulating censored observations because the distribution of censoring times is unknown. It has been shown that test statistic distributions may significantly differ for various distributions of censoring times. The algorithm for simulation of statistic distribution for testing goodness-of-fit with the parametric PH model by type I or type II censored data can be written as follows.

1. Generate response values $t_1, ..., t_n$ from $F_x(t; \hat{\beta}, \hat{\theta})$ according to the tested model. where $\hat{\beta}$ and $\hat{\theta}$ are MLEs obtained by the source data.

2. Transform complete sample $t_1, ..., t_n$ into censored sample according to the plan of experiment.

3. Calculate the MLEs of β and θ by obtained censored sample for given values of covariates.

4. Obtain residuals $\hat{R}_1, ... \hat{R}_n$ by (6).

5. Calculate the test statistic.

By repeating the process N times a random sample from the test statistic distribution is generated.

3 Acknowledgements

The research has been partially supported by the Russian Foundation for Basic Research (Grant No. 09-01-00056-a) and the Federal targeted program of the Ministry of Education and Science of the Russian Federation "Scientific and scientific-educational personnel of innovative Russia"

References

[1] Cox D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society. Series B.* Vol. **34**, pp. 187-220.

[2] Kalbfleisch J.D., Prentice R.L. (1980). *The statistical analysis of failure time data.* John Wiley and Sons, Inc., New York.

[3] Lawless J.F. (2003). *Statistical models and methods for lifetime data.* John Wiley and Sons. Inc., Hoboken, New Jersey.

[4] Lemeshko B.Yu., Lemeshko S.B. (2009). Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. Part 1. *Measurement Techniques.* Vol. **52**, No. 6. pp. 555-565.

[5] Lemeshko B.Yu., Lemeshko S.B. (2009). Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. Part II. *Measurement Techniques.* Vol. **52**, No. 8. pp. 799-812.