ON ANALYSIS OF MARTINGALE RESIDUALS IN LIFETIME MODELS

Petr Volf*

ÚTIA AV ČR, Praha

Abstract: Residual analysis is used commonly in statistical tests of model fit, i.e. of good correspondence between data and model. We shall recall the notion of martingale residuals defined for the case of lifetime models based on failure intensities. Analysis of these residual processes has already been studied by many authors. Nevertheless, Bayes approach to this problem is just developing. We shall present a Bayes procedure of estimation in models containing nonparametric components. Further, Bayes construction of residual processes and model goodness-of-fit assessing will be proposed and demonstrated on an example with Cox's regression model.

Key words: Bayes statistics, Cox's model, goodness-of-fit, martingale residuals.

1 Martingale residuals

Let us first recall a basic scheme of survival analysis, for the beginning without considering any dependence on covariates. We shall switch to regression models later. Let us imagine that we observe survival times of n objects of the same type, so that we describe their survivals by i.i.d. (independent, identically distributed) set of random variables T_i. Alternatively, we may consider their **counting processes** $N_i(t)$, each having maximally 1 count (at the time of failure, T_i), or it can be censored without failure. For this case, we have indicator processes (of being in risk) $Y_i(t)$, which $=0$ after failure or censoring, $Y_i(t) = 1$ otherwise. As lifetimes are i.i.d., corresponding counting processes have the same common **hazard rate** $h(t) \geq 0$. Cumulated hazard rate is then $H(t) = \int_0^t h(s)ds$. It follows that the intensity of $N_i(t)$ is $a_i(t) = h(t) \cdot Y_i(t)$. Notice a difference between those two notions: The hazard rate is a characteristics of distribution, namely here $h(t) = -d\bar{F}(t)/dt$ where $\bar{F}(t) = 1 - F(t)$ is a survival function, complement to distribution function, while intensity depends on realization of process $Y_i(t)$. It is assumed that data are observed on a finite time interval $t \in [0, T]$, $N_i(0) = 0$.

We can as well define sums of individual characteristics, namely counting process $N(t) = \sum_{i=1}^n N_i(t)$ counting number of failures, further $Y(t) = \sum_{i=1}^n Y_i(t)$, cumulated intensities $A_i(t) = \int_0^t a_i(s)ds$ and $A(t) = \sum_{i=1}^n A_i(t)$, so that here $A(t) = h(t) * Y(t)$.

In theoretical studies on lifetime models, many results are based on martingale – compensator decomposition of counting process, namely that $N_i(t) = A_i(t) + M_i(t)$, so that also $N(t) = A(t) + M(t)$, where $M_i(t)$, $M(t)$ are martingales with $EM(t) = 0$, conditional variance processes (conditioned by corresponding filtration, a nondecreasing set of σ-algebras $\mathcal{F}(t^-)$) are $\langle M_i \rangle(t) = A_i(t)$, $\langle M \rangle(t) = A(t)$. From practical point of view, these limits from the left of σ-algebras $\mathcal{F}(t^-)$ represent actually all observations available at (just before) time t. Naturally, martingales have non-correlated increments, and here are also non-correlated mutually (for different i).

Then it is quite reasonable to consider a **RESIDUAL PROCESS** (martingale residuals)

$$R(t) = N(t) - \hat{A}(t) = M(t) + A(t) - \hat{A}(t)$$

as a tool for testing model fit. Here $\hat{A}(t)$ is estimated cumulated intensity. Hence, residual process is constructed from observed data, and its properties depend mainly on properties of

*email: volf@utia.cas.cz

estimator of cumulated hazard rate, because $\hat{A}(t) = \int_0^t Y(s)d\hat{H}(s)$. Tests are then performed either graphically or numerically, critical borders for assessing the goodness-of-fit are then based on asymptotic properties of estimates.

1.1 Properties of residuals

The most common estimator of cumulated hazard rate $H(t)$ is so called Nelson-Aalen estimator, which has the form:

$$\hat{H}(t) = \int_0^t \sum_{i=1}^{n} \frac{dN_i(s)}{\sum_{j=1}^{n} Y_j(s)} = \int_0^t \frac{dN(s)}{Y(s)},$$

so that it is a piecewise constant function with jumps $d\hat{H}(s) = dN(s)/Y(s)$ at times where failure occurred. Its asymptotic properties, namely uniform on $[0,T]$ consistency in probability and asymptotic normality when $n \to \infty$, are well known (for review of survival analysis, see for instance Kalbfleisch and Prentice, 2002). More precisely, the following convergence in distribution on $[0,T]$ to Brown process $\mathcal{B}$ holds:

$$\sqrt{n}(\hat{H}(t) - H(t)) \to \mathcal{B}(V(t)), \quad V(t) = \int_0^t \frac{h(s)ds}{c_0(s)},$$

where we assume the existence of $c_0(s) = P - lim\frac{Y(s)}{n}$, uniform in $[0,T]$, $c_0(s) \geq \varepsilon > 0$.

Hence, it is possible to construct Kolmogorov-Smirnov type confidence bands for $H(t)$, also point-wise confidence intervals. Consistent (again uniform in $0,T$) estimator of $V(t)$ is available, too: $\hat{V}(t) = \int_0^t \frac{n\,dN(s)}{Y(s)^2}$.

However, in present contribution we are more interested in properties of residual process $R(t) = N(t) - \hat{A}(t)$. Notice that $\hat{A}(t) = N(t)$ directly, so that it is preferred to construct residuals in data subgroups (strata), $S \subset \{1,..,n\}$. Thus, let us define

$$R_S(t) = N_S(t) - \hat{A}_S(t) = M_S(t) + A_S(t) - \hat{A}_S(t),$$

where we denoted again $N(t) = \sum_{i=1}^{n} N_i(t)$, $N_S(t) = \sum_{i\in S} N_i(t)$, similarly for $Y(t)$, $M(t)$, $A(t)$, $\hat{A}(t)$. As

$$\hat{A}_S(t) = \int_0^t \sum_{i\in S} d\hat{H}(r)Y_i(r) = \int_0^t \frac{dN(r)}{Y(r)} \cdot Y_S(r) =$$

$$= \int_0^t \frac{dH(r)Y(r) + dM(r)}{Y(r)} \cdot Y_S(r) = A_S(t) + \int_0^t \frac{dM(r)}{Y(r)} \cdot Y_S(r),$$

we obtain that (with notation $\bar{S}$ - complement of S)

$$R_S(t) = M_S(t) - \int_0^t \frac{dM(r)}{Y(r)} \cdot Y_S(r) = \int_0^t \frac{dM_S(r)Y_{\bar{S}}(r) - dM_{\bar{S}}(r)Y_S(r)}{Y(r)}.$$

From its structure it follows that process $R_S(t)$ has non-correlated increments, conditional variance (conditioned by σ-algebras $\mathcal{F}(t^-)$) of $\frac{1}{\sqrt{n}}dR_S(t)$ is

$$\frac{dH(t)}{nY(t)^2}(Y_{\bar{S}}(t)Y_S(t)^2 + Y_{\bar{S}}(t)^2Y_S(t)) \sim dH(t)\frac{c_S(t)c_{\bar{S}}(t)}{c_0(t)},$$

where we again assume that there exist P-limits $Y_S(t)/n \to c_S(t), Y_{\bar{S}}(t)/n \to c_{\bar{S}}(t), Y(t)/n \to c_0(t)$, uniform in $t \in [0,T]$, positive. Then asymptotic normality, namely the convergence to Brown process $\frac{1}{\sqrt{n}}R_S(t) \to \mathcal{B}(V_R(t))$ holds, and asymptotic variance function $V_R(t)$ is consistently estimable by:

$$\hat{V}_R(t) = \int_0^t \frac{d\hat{H}(r)Y_S(r)Y_{\bar{S}}(r)}{nY(r)} = \int_0^t \frac{dN(r)Y_S(r)Y_{\bar{S}}(r)}{nY(r)^2}.$$

Hence, the process (provided assumptions of our model hold)

$$\frac{1}{\sqrt{n}} \frac{R_S(t)}{(1+\hat{V}_R(t))}$$

behaves asymptotically as Brown bridge process, so that Kolmogorov-Smirnov (or other similar, as Cramer-von Mises) criterion can be used directly for it. In the present case, we assume a simple model of survival times (without any non-heterogeneity), so that the method can be used for assessing the homogeneity of different subsamples S.

For direct comparison of two subsamples C and D the following variant can be considered: Let $H(t)$ be estimated from a set of data D and residuals are constructed in another set C. Under hypothesis that $H(t)$ is the same in both sets, we have

$$R_{C,D}(t) = N_C(t) - \hat{A}_C(t) = M_C(t) - \int_0^t \frac{dM_D(r)}{Y_D(r)} \cdot Y_C(r).$$

Hence, conditional variance

$$var \frac{1}{\sqrt{n}} dR_{C,D}(t) \sim dH(t) \frac{c_C(t)}{c_D(t)} c_0(t),$$

again with assumed existence of P-limits $\frac{Y_C}{n} \to c_C$, $\frac{Y_D}{n} \to c_D$, $\frac{Y}{n} \to c_0$ as above.

The case considered in the present part was rather simple, in such a case the tests of model fit can be performed directly with the aid of estimated cumulative hazard rates or distribution functions (recall well known Product limit estimate of Kaplan and Maier). However, the test are not so straightforward in cases of regression models. That is why we defined residual process in a simple setting first, and we shall continue by description of Bayes variant of residual analysis.

2 Bayes version of residual process

Bayes approach to statistical models considers all models components (i.e. the parameters as well as non-parametrized components) as random quantities, initially with a prior probability distribution. The result of statistical analysis is then a posterior distribution of those model components, i.e. estimate is a distribution. Actually it is the likelihood function 'modulated' by prior distribution.

From another point of view, it is possible to say that while the "standard statistics" studies the variation of data and its consequence when inserted to given functions (estimators), in Bayes statistics the main concern is variation of 'parameters', data are taken as fixed.

Today, Bayes analysis is often connected with (supported by) MCMC (Markov Chain Monte Carlo) methods. They are based on certain algorithms of random sampling (Gibbs sampler, Metropolis-Hastings procedure) and are used for obtaining approximate representation of posterior distribution. More about MCMC can be found elsewhere, for instance in Gamerman (1997).

In the case considered here we deal with nonparametric hazard rate. For Bayes solution, its representation can be made from piecewise-constant functions (or from splines or from other functional basis), as in Arjas and Gasbarra (1994). Parameters are then points of changes of hazard rate, also their number in $[0, T]$, and levels of hazard rate in intervals between these points. Arjas and Gasbarra (1994) show how MCMC generation can follow Gibbs sampler combined with an "accept-reject" sampling method.

Once we have posterior sample (i.e. last M representatives of aposteriori distribution obtained by MCMC procedure) of 'hazard rates', $h^{(m)}(t)$, we can construct from them a sample of cumulated intensities in subgroup S and corresponding residuals:

$$A_S^{(m)}(t) = \int_0^t h^{(m)}(r) Y_S(r) dr, \quad R_S^{(m)}(t) = N_S(t) - A_S^{(m)}(t).$$

Further, point-wise (at each t) sample quantiles of $R_S^{(m)}(t)$ are obtained immediately, showing so called credibility intervals (Bayesian version of confidence intervals) for $R_S(t)$. Hence, if 0 is inside, hypothesis of good fit is not rejected.

Methods for construction of confidence bands (of Bayes type) on whole interval $[0, T]$ are studied intensively. They can utilize ideas developed for construction of multivariate quantiles, or even non-parametric regression quantiles, also just developing area of 'depth of data' analysis can contribute to this task solution.

3 Residuals in regression models

In the follow-up, we shall assume that distribution of time-to failure, and consequently also intensity, depends on covariates. It means that we have to select a regression model for hazard rate and after its evaluation it is necessary to test the model fit. It is actually the proper situation where the residual analysis should be used. We shall discuss here just two types of regression models, skipping another frequently used, namely the accelerated failure time model. More details about regression models in survival analysis can be found in many monographs, let us again mention Kalbfleish and Prentice (2002).

3.1 Additive (Aalen's) regression model

In this model, hazard function is specified as $h(t, z) = z \cdot \beta(t)$, where z stands instead values of covariates, $\beta(t)$ are functions of time, both z and β are p-dimensional. Their domains should ensure that $h(t, z) \geq 0$. As a rule, the first covariate component is taken fixed to 1, so that $\beta_1(t)$ has the meaning of a 'baseline' hazard function. In the sequel, by index i, $i = 1, .., n$ we shall denote individual objects, while by k, $k = 1, .., p$ components of vectors β, z.

The covariates themselves, $Z_i(t)$, can be different for each object and can change in time. Individual intensity of $N_i(t)$ is then

$$a_i(t) = Z_i(t) \cdot \beta(t) \cdot Y_i(t), \;\; i = 1, \ldots, n.$$

Cumulated functions $B_k(t) = \int_0^t \beta_k(s)\, ds$ are estimated by weighted least squares method. As $dN_i(t) = X_i(t) dB(t) + dM_i(t)$, where $X_i(t) = Z_i(t) \cdot Y_i(t)$, then

$$\hat{B}(t) = \int_0^t (X(r)' W(r) X(r))^{-1} X(r)' W(r) dN(r),$$

where $W(r)$ is a matrix of weights; the simplest choice $W(r) = I_n$, identity matrix, optimal weights are $W(r) = diag\{1/a_i(r)\}$, in praxis $\hat{a}_i(r)$ are used, computation is iterated.

Consistency and asymptotic normality of $\hat{B}(t)$ are straightforward, it holds that the term

$$\sqrt{n}(\hat{B}(t) - B(t)) = \sqrt{n} \int_0^t \overline{X}(r) dM(r),$$

where $\overline{X}(r) = (X(r)' W(r) X(r))^{-1} X(r)' W(r)$, is asymptotically distributed as a Gaussian process with independent increments (i.e. Brown process) and its covariance function is estimable consistently (uniformly on $[0, T]$) by

$$n \int_0^t \overline{X}(s)\, D(s, B(s))\, \overline{X}'\, ds$$

provided such a limit exists (here $D(s, B(s))$ is a diagonal matrix with components $a_i(s)$).

It follows that the case is similar to the case of nonparametrized hazard rate treated in 1-st part. It means that it is possible develop the asymptotic distribution of residuals, again coinciding with Brown process distribution. It is described in detail in Volf (1996). Even Bayes residual analysis can follow the same scheme as in the preceding part, each function $\beta_k(t)$ has to be modelled separately.

3.2 Cox's regression model

As we shall se, the case differs in certain aspects from preceding one, due more complicated asymptotic properties. Now, the hazard rate is specified as $h(t,z) = h_0(t)exp(z \cdot \beta)$, with processes of covariates $Z_i(t)$ and parameter β (both p-dimensional), $h_0(t)$ is a baseline hazard rate, a nonnegative function.

Then intensity of i-th process $N_i(t)$ is

$$a_i(t) = h(t, Z_i(t)) \cdot Y_i(t).$$

Parameter β is estimated from partial log-likelihood

$$L_p = \sum_{i=1}^{n} \int_0^T log\{\frac{exp(Z_i(t)\beta)}{\sum_{k=1}^n exp(Z_k(t)\beta) \cdot Y_k(t)}\} dN_i(t),$$

by an iterative procedure (of Newton-Raphson as a rule), cumulated baseline hazard $H_0(t) = \int_0^t h_0(r)dr$ is then estimated as

$$\hat{H}_0(t) = \int_0^t \frac{dN(r)}{\sum_{k=1}^n exp(Z_k(t)\hat{\beta}) \cdot Y_k(t)}.$$

Theory on properties of estimates is collected elsewhere, first time the results has been established by Andersen and Gill (1982). Estimates are consistent, asymptotically normal, however, neither $\sqrt{n}(\hat{H}_0(t) - H_0(t))$ nor residual process are martingales.

3.3 Residuals in Cox's model

Residuals are sometimes formulated more generally, as

$$dR(t) = \sum_{i=1}^{n} K_i(t) \cdot (dN_i(t) - d\hat{A}_i(t)),$$

with some (convenient) 'weight' processes $K_i(t)$, for instance if $K_i(t) = Z_i(t)$ (p-dimensional), $R(t)$ is then estimated score process (the first derivative) of L_p, while $K_i(t) = 1[i \in S]$ yields stratified residuals. Stratified residuals (the simplest case) are then expressed as

$$dR_S(t) = dM_S(t) + dH_0(t)C_S(\beta_0, t) - d\hat{H}_0(t) \cdot C_S(\hat{\beta}, t),$$

where $d\hat{H}_0(t) = \frac{dN(t)}{C(\hat{\beta},t)}$ and $C_S(\beta,t) = \sum_{i\in S} exp(Z_i(t)\beta) \cdot Y_i(t)$, $C(\beta,t) = \sum_{i=1}^n exp(Z_i(t)\beta) \cdot Y_i(t)$.

If we take approximately $\hat{\beta} \sim \beta_0$, we obtain expression similar to previous 'non-regression' case. Exact approach (for good overview see Kraus, WDS 2004) uses Taylor expansion of the last term at β_0. Then $R_S(t)/\sqrt{n}$ is expressed with the aid of a martingale and a nonrandom function, with asymptotic distribution as a Gaussian process, however with rather complicated covariance structure.

Hence, random generation of would-be residual processes with 'ideal' distribution under hypothesis of model fit is possible (but not easy). It is actually a bootstrapping, by which we obtain a sample of 'ideal' residual processes. Then, for instance their absolute maxima are compared with maximal residual computed from our data. Or other characteristics can be compared. That is why, the practical tests of Cox's model fit is mostly performed just graphically, comparing visually how far are residuals in group S from zero line, or, equivalently, $\hat{A}_S(t)$ from $N_S(t)$, as proposed in Arjas (1988). Thus, it seems that in the case of Cox's model the Bayes analysis could offer an easiest tool for model fit assessing.

3.4 A Bayes procedure in Cox's model

Let us briefly summarize starting points and procedure of Bayes analysis in the Cox's model setting, namely the variant applied in the following numerical example. Metropolis-Hastings steps of MCMC procedure were used to obtain samples representing posterior distributions of β. Values of β-s were proposed from a prior (a sufficiently wide uniform, in our case), accepted or rejected with the use of partial likelihoods proportion. Then, to each β, a representation of $H_0(t)$ was generated, similarly as in the 1-st part. i.e. from a piecewise constant prior. Finally, we obtained a sample of both, $\beta^{(m)}, h_0^{(m)}(t)$, $m = 1, .., M$, from them the intensities and residuals (in a group S, say) were derived:

$$A_S^{(m)}(t) = \int_0^t h_0^{(m)}(r) \sum_{i \in S} exp\left(Z_i(r)\beta^{(m)}) \cdot Y_i(r)\right) dr, \;\; R_S^{(m)}(t) = A_S^{(m)}(t) - N_S(t).$$

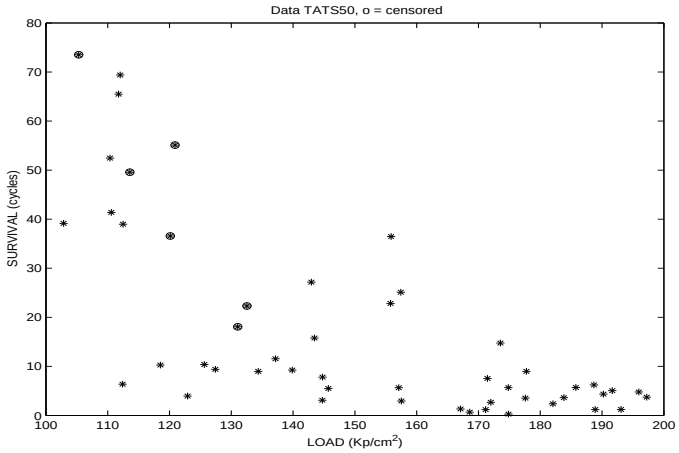

Figure 1: Data: Load (Kp/cm^2) on x axis, survival on y axis, censored items denoted by 'o'

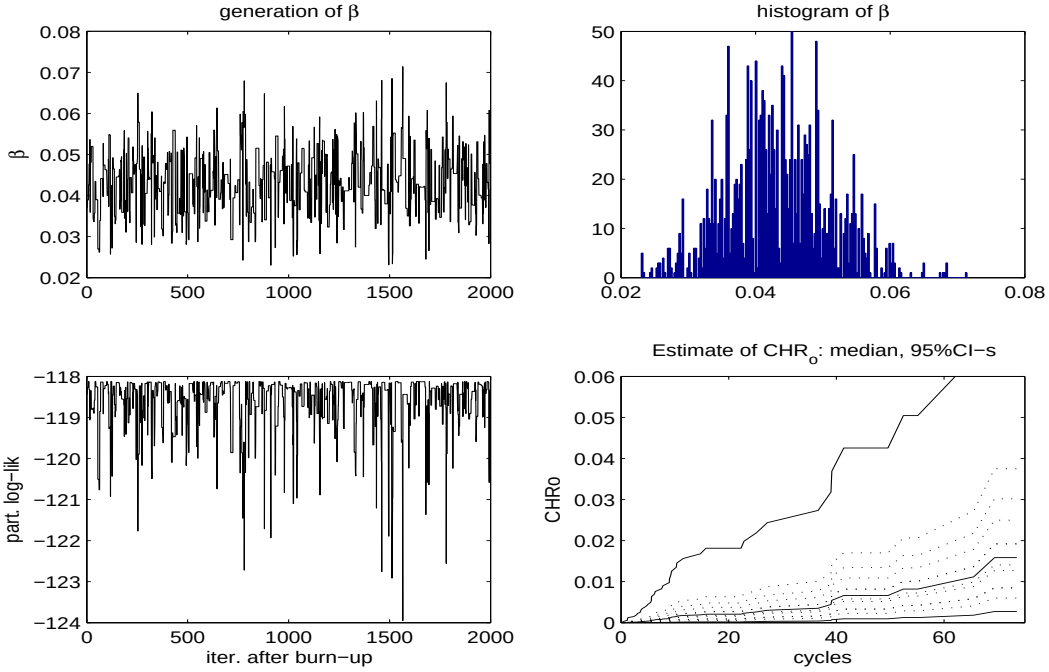

Figure 2: Results of Bayes - MCMC estimation in Cox's model

4 Example

We analyzed the data – survival of 50 metal parts tested by cycles of vibrations under different stress level. The data are on Figure 1, some of them are censored from above (survival is randomly censored from the right). It is seen that the survival is significantly smaller under larger load. To express this dependence, we selected the Cox's model. Standard analysis in Cox's model yielded estimated $\beta = 0.0429$, with asymptotic 95% confidence interval (0.0279, 0.0578).

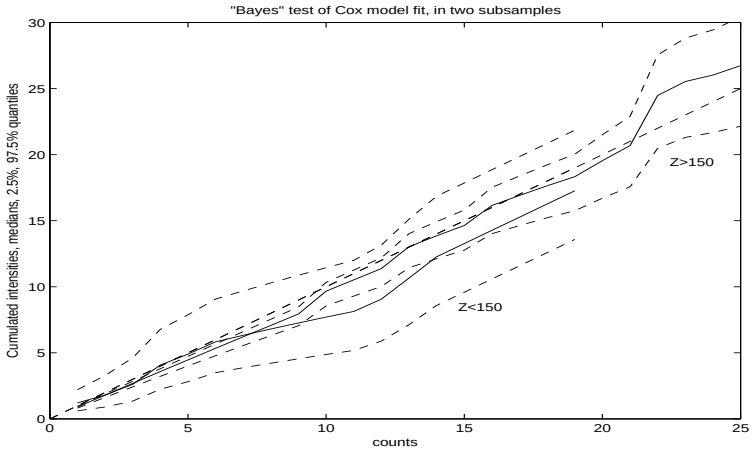

Figure 3: Characterization of residual processes in 2 groups, estimated intensities are plotted against counting processes.

Then, Bayes analysis, following the procedure described above, has been performed. We generated 5000 β-s, take last 2000 as a representation of posterior distribution of β. To each $\beta^{(m)}$ 200 instances of $H_0(t)$ were generated, we always took just the last of them. In such a way, a representation with M=2000 members was obtained. Together, it was used to model fit analysis. Posterior of β had the mean 0.0432 and standard deviation 0.77.

Figure 2. shows (above left) the course of MCMC generation of β, last $M = 2000$ iterations from 5000. Above right there is the histogram of final sample of $\beta^{(m)}$, $m = 1, ..., M$. Below left is development of log of partial likelihood of β, and finally (below right) the characteristics of generated representation $H_0(t)^{(m)}$ of baseline cumulative hazard function are shown, namely point-wise sample medians and 2.5%. 97.5% quantiles (several functions from set $H_0(t)^{(m)}$ are displayed, too, by dots).

Figure 3 then displays estimated cumulated intensities $\hat{A}_S(t)$, more precisely, from obtained sample we display their point-wise medians (full) and 95% credibility intervals (dashed) plotted against counts $N_S(t)$, in 2 groups: $Z < 150$, $Z > 150$). It is seen that graphs are concentrated around diagonal line (dashed, too), showing good Cox's model fit. As it has been said, it is a way proposed for instance by Arjas (1988), quite equivalent to plotting the residuals around zero line.

5 Conclusion

Models of lifetime often have to incorporate a dependence on covariates. The sense of the present contribution was to show several variants of goodness-of-fit tests as procedures supporting a proper regression model selection. It was also shown that sometimes, as in the case of Cox's regression model, Bayes approach could be a reasonable alternative to standard analysis.

Acknowledgement: The research has been supported by the project of MŠMT ČR No. 1M06047, The Center of Quality and Reliability.

References

[1] Andersen P.K., Gill R. (1982) *Cox's regression model for counting processes: a large sample study.* Ann. Statist. 10, 1100-1120.

[2] Arjas E (1988) *A graphical method for assessing goodness of fit in Cox's proportional hazard model.* J. Amer. Statist. Assoc. 83, 204-212.

[3] Arjas E., Gasbarra D. (1994) *Nonparametric bayesian inference from right censored survival data, using Gibbs sampler.* Statist. Sinica 4, 505-524.

[4] Gamerman D., (1997) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall/CRC Press, Boca Raton.

[5] Kraus D. (2004) *Testing goodness of fit of hazard regression models.* In Proceedings of WDS, Part I, MFF UK Praha 2004, 6-12.

[6] Kalbfleisch J.D., R.L. Prentice R.L. (2002) *The Statistical Analysis of Failure Time Data.* Wiley, New York.

[7] Volf P. (1996) *Analysis of generalized residuals in hazard regression models.* Kybernetika 32, 501-510.