New Residuals for Cox Regression and Their Application to Outlier Screening

Alessandra Nardi
Department of Systems Theory, University of Teramo,
Viale Crucoli 122, I-64100 Teramo, Italy

and

Michael Schemper
Department of Medical Computer Sciences, Vienna University, Spitalgasse 23, A-1090 Vienna, Austria
e-mail: michael.schemper@akh-wien.ac.at

SUMMARY. The identification of individuals who 'died far too early' or 'lived far too long' as compared to their survival probabilities from a Cox regression can lead to the detection of new prognostic factors. Methods to identify outliers are generally based on residuals. For Cox regression, only deviance residuals have been considered for this purpose, but we show that these residuals are not very suitable. Instead, we develop and propose two new types of residuals: the suggested log-odds and normal deviate residuals are simple and intuitively appealing and their theoretical properties and empirical performance make them very suitable for outlier identification. Finally, various practical aspects of screening for individuals with outlying survival times are discussed by means of a cancer study example.

KEY WORDS: Outlier screening; Residuals; Survival analysis.

1. Introduction
The detection of single individuals whose survival is poorly predicted by a fitted Cox (1972) regression, i.e., of individuals who 'died far too early' or 'lived far too long' as compared to the estimated survival probabilities for their covariate pattern, can be of even greater impact for medical research than standard analysis results. Further medical investigation of the characteristics of those individuals who died much earlier or much later, if at all, than predicted by a model may ultimately result in the detection of hitherto unknown favourable or unfavourable conditions for the development of a disease.

In Section 4, we analyse a prostate cancer study consisting of 310 patients for whom survival after treatment and seven prognostic factors were recorded. After fitting a Cox model to this data set, we identify those patients who were most poorly predicted by the model. A search of unknown prognostic factors can then originate from a thorough investigation of the medical records of these patients.

The detection of outliers within the framework of the proportional hazards model (Cox, 1972) has received almost no attention despite the frequent use of this model in medical research and the continued interest in residuals for the model (Grambsch and Therneau, 1994).

The topic of outliers is relatively well developed for the general linear model (cf., Beckmann and Cook, 1983; Barnett and Lewis, 1984; Hawkins, 1980). Some procedures exist to accommodate outliers, while others aim at identifying them. For the latter case, simple tests exist to check if the residuals of certain individuals, selected prior to an inspection of the data, are too large to have occurred by chance. Furthermore, the largest observed residual can be tested for being too large to have occurred by chance. While the test for certain selected individuals, based on the distribution of studentized residuals, is of little practical interest, the test for the largest observed residual, being based on the distribution of the maximum studentized residual, is of more interest. The situation of most practical interest, testing for multiple outliers without *a priori* information on their number, cannot be addressed by automatic testing: A repeated single outlier test may fail in the presence of multiple outliers (masking), while a block test for multiple outliers may declare certain observations as outlying that are not (swamping). We therefore agree with Beckmann and Cook (1983) that one should not routinely seek to reject outliers on the basis of formal tests.

Given this rather limited state of formal outlier testing for the general linear model and also judging the usefulness of such tests within Cox models, we consider outlier concepts mainly relevant to screen for outlying survival times as described in Section 4. This screening is comparable to the screening of individuals by reference ranges for clinical laboratory data. In either situation, the percentage of individuals expected to be outlying in an adequately modelled reference population has to be selected.
Statistics suitable for outlier detection should increase in absolute value as the related observations move away from their predicted values. This is the property of residuals. They should also be easy to interpret and should be known to follow a standard theoretical distribution under a properly specified model, assuming no outliers present. Finally, in order to improve the quality of any graphical procedure, symmetry and unimodality of their sampling distribution are crucial. Such residuals are suggested in Section 2, and their empirical properties are dealt with in Section 3.

Though residuals (cf., Cox and Snell, 1968; Crowley and Hu, 1977) have been historically the first type of residuals considered with Cox regression, the following residuals are currently proposed (Schoenfeld, 1982; Barlow and Prentice, 1988; Therneau, Grambsch, and Fleming, 1990; Collett, 1994) for different applications: (1) martingale residuals to assess the functional form of the influence of a covariate in a model, already accounting for other covariates, (2) Schoenfeld residuals for checking the proportional hazards assumption for a covariate, (3) score residuals for demonstrating an individual’s leverage on parameter estimates, and (4) deviance residuals in the context of outlier identification (cf., SAS/STAT, 1997, p. 593).

However, deviance residuals have no reference sampling distribution and the approximation by a standard normal distribution (Fleming and Harrington, 1991) appears unsatisfactory even without censoring (see Section 3). In the following section, we introduce two new residuals that are better suited for outlier identification, as will be shown in Section 3.

2. Definition of Log-Odds and Normal Deviate Residuals

2.1 General Aspects

We consider prediction of survival by Cox’s model as perfect for individual i if \( \hat{S}_i(t_i) = 0.5 \), where \( \hat{S}_i(t_i) \) denotes the estimated value of the survival function for individual i at his observed death time \( t_i \). Alternatively, we regard individual i as well predicted if his observed survival time and his estimated median survival time, according to a fitted Cox regression model, agree.

Instead of directly measuring discrepancies between observed and estimated median failure time, which may not always be possible, we measure this discrepancy indirectly by comparing the estimated survival probability at failure time with 0.5.

Since attention focuses on the median survival time, failure versus nonfailure by estimated median time can be regarded as a binary variable, implicitly assuming a binomial model. Therefore, two types of residuals appear particularly appropriate, using the logit and the probit transformation of the survival function, i.e., (1) log-odds residuals: \( L_i = \log[\hat{S}_i(t_i)/(1 - \hat{S}_i(t_i))] \) and (2) normal deviate residuals: \( N_i = \Phi^{-1}(\hat{S}_i(t_i)) \), where \( \Phi \) denotes the normal cumulative distribution function.

Note that increasing departures from a perfect prediction are reflected by increasing absolute values of both the suggested residuals. Their sampling distributions, assuming \( \hat{S}_i() \) known, are the logistic for \( L_i \), with \( E(L_i) = 0 \) and \( \text{var}(L_i) = \pi^2/3 \), and the standard normal for \( N_i \), respectively.

Both results follow from the uniform distribution of \( F(T) = 1 - S(T) \), where \( F(T) \) is the distribution function of \( T \). If the unknown survival function is replaced by its estimator and assuming a correctly specified model, \( \hat{L}_i \) and \( \hat{N}_i \) converge in probability to \( L_i \) and \( N_i \) (Nardi, 1996).

2.2 The Censored Case

Observed survival times \( t_i \) (\( 1 \leq i \leq n \)) are symbolized by \( t_i^c \) if they are censored. At survival time \( t_i \), \( l_i \) and \( n_i \) denote the observed values of \( \hat{L}_i \) and \( \hat{N}_i \) and are

\[ l_i = \log[\hat{S}_i(t_i)/(1 - \hat{S}_i(t_i))] \]

and

\[ n_i = \Phi^{-1}(\hat{S}_i(t_i)) \],

while log-odds and normal deviate residuals evaluated at \( t_i^c \) are denoted by \( l_i^c \) and \( n_i^c \).

There are various options for accommodating both types of residuals to censoring (Crowley and Hu, 1977). These options use the fact that the true survival time is greater than the observed censored one and therefore the distribution of the unknown true residual is related to the uniform distribution of \( S_i(t_i^c) \) in \([0, S_i(t_i^c)]\).

Thus, replacing the unknown value \( \hat{S}_i(t_i) \) by its conditional median value, \( \hat{S}_i(t_i^c)/2 \), we get

\[ l_i^m = \log[\hat{S}_i(t_i^c)/(2 - \hat{S}_i(t_i^c))] \]

and

\[ n_i^m = \Phi^{-1}(\hat{S}_i(t_i^c)/2). \]

The derivation of expected mean values from the conditional distributions of \( L_i \) and \( N_i \), \( l_i^c \) and \( n_i^c \), is slightly more involved (see Appendix), with

\[ l_i^c = l_i - \frac{1 + \exp(l_i^c)}{\exp(l_i^c)} \log(1 + \exp(l_i^c)) \]

and

\[ n_i^c = -\frac{\exp(-0.5(n_i^c)^2)}{\sqrt{2\pi}\hat{S}_i(t_i^c)}. \]

Due to the averaging process, \( l_i^c \) and \( n_i^c \) (and similarly \( l_i^m \) and \( n_i^m \)) tend to be less extreme than the corresponding unobservable \( l_i \) and \( n_i \) and the empirical distributions of both the proposed residuals are more concentrated than the theoretical ones.

When obtaining expected mean or median values of residuals from censored survival times, it should be noted that the information from a short censored survival time is too limited to identify it as outlying. In fact, for an extremely short censored time, \( l_i^c \) and \( n_i^c \) (or \( l_i^m \) and \( n_i^m \)) will be close to zero as an expected \( \hat{S}_i(t_i^c) \approx 0.5 \) will be assumed.

In order to avoid the concentration effect of averaging, an alternative option in the presence of censoring might be to randomly sample from the conditional distributions and to proceed in the spirit of Rubin’s (1987) multiple imputation. Since our goal is to identify outlying survival times, this option is of little interest. Censored survival times are candidates for outliers. Therefore, it is particularly appealing for the purpose of outlier identification to compute for each censored survival time a probability \( P_i \) that, if uncensored, the corresponding residual would belong to the \( \alpha \)-percent residuals (usually \( \alpha = 0.05 \) or \( \alpha = 0.025 \)) with largest negative values (related
Cox Regression Residuals and Outlier Screening

3. Empirical Studies of the Suggested Residuals

3.1 Empirical Distributions

In this section, distributions of deviance, log-odds, and normal deviate residuals are described and compared. For this purpose, 10,000 trials were generated, each of sample sizes $n$ of 20, 40, 80, 200, and 400. Values of a single covariate $X$ were assumed binomially distributed with $p(X = 1) = p(X = 2) = 0.5$, and corresponding survival times $T$ were assumed exponentially distributed with hazards $h(t | X = 1) = 1$ and $h(t | X = 2) = 2$. To explore the properties of residuals under censoring, the generated samples were censored using the procedure by Gehan and Thomas (1969) to model a clinical trial. Subjects were assumed to enter a study in a constant rate in an interval $(0, Z)$ and then to fail according to the prescribed survival time distribution. For each experimental condition, a value of $Z$, the time of analysis, was determined as in Lininger et al. (1979) to achieve an expected 40% censoring of survival times.

A Cox model was fitted to the generated data of each trial, and deviance, log-odds, and normal deviate residuals of the last subject of each trial were calculated. Their empirical distributions in 10,000 trials are given in Figure 1 for $n = 200$. Distributional results for other sample sizes are very similar.

![Figure 1](image-url)
Figure 2. Residual plots for prostate cancer example. The abscissae give the predictor $x\hat{\beta}$. Residuals for uncensored (*) and censored (○) data are distinguished.

(Nardi, 1996). The empirical studies were carried out within SAS (1990) using its random number generators. We learn that, even under no censoring, the empirical distribution of deviance residuals departs substantially from the standard normal distribution. In contrast, log-odds and normal deviate residuals’ empirical distributions agree quite well with their theoretical reference distributions, the logistic and the standard normal, respectively. Results for normal deviate and log-odds residuals practically agree, as both are transforms of $\hat{S}_i(t_i)$ with appropriate limiting distributions and each type of residual can be transformed into the other. In case of censoring, the approximations to the empirical distributions are far from perfect for all three types of residuals but are much worse for deviance residuals. The bimodal shape of the deviance residuals’ empirical distribution can be attributed to the mixture of censored and uncensored data residuals, for which the contributions to the ordinary likelihood differ. Figure 2d shows the local concentration of the empirical distribution of log-odds residuals due to replacing censored residuals by their conditional expected values. As expected, only the left tail of the distribution is involved since censoring mainly affects long survival times. For screening purposes, however, this unfavourable effect under censoring can be avoided by using probabilities $P_i$ (see Section 2) that the residual belongs to a critical region had the observation not been censored. The performance of this procedure is explored in the following section.

3.2 Screening by Log-Odds or Normal Deviate Residuals
We continue to use the simulation study set-up of Section 3.1 but now count the number of residuals of the second subjects of the generated trials that exceeded critical limits and also add the probabilities $P_i$ (see Section 2) that censored times of the second subjects could be outlying. To be useful for clinical research, outlying subjects who appear to have ‘lived too long’ have to be distinguished from those who appear to have ‘died too early,’ and therefore separate critical regions are defined as follows:

$$R_{TED,L} = \{l_2 : l_2 > w_{1-\alpha}\}$$

$$R_{TLL,L} = \{l_2 : l_2 < w_{\alpha}\}$$
Cox Regression Residuals and Outlier Screening

**Table 1**
Coverage of critical regions in outlier screening by log-odds residuals
at one-sided significance levels $\alpha$ (10,000 simulated samples of size $n$)

<table>
<thead>
<tr>
<th></th>
<th>No censoring</th>
<th></th>
<th>40% Censoring</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 0.05$</td>
<td>$\alpha = 0.01$</td>
<td>$\alpha = 0.05$</td>
<td>$\alpha = 0.01$</td>
</tr>
<tr>
<td>‘Too long lived’</td>
<td>20</td>
<td>0.042</td>
<td>0.0008</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.049</td>
<td>0.010</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.049</td>
<td>0.010</td>
<td>0.050</td>
</tr>
<tr>
<td>‘Too early died’</td>
<td>20</td>
<td>0.036</td>
<td>0.0008</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>0.046</td>
<td>0.008</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.051</td>
<td>0.009</td>
<td>0.051</td>
</tr>
</tbody>
</table>

Note: The entries in this table are average $P_2$, where $P_2$ denotes the probability that the second observation in a simulated sample belongs to the critical region. For uncensored times, $P_2$ is one if the residual belongs to the critical region and is zero otherwise.

and

$$R_{TED,N} = \{ n_2 : n_2 > z_{1-\alpha} \}$$

$$R_{TLL,N} = \{ n_2 : n_2 < z_\alpha \},$$

where $R_{A,B}$ defines the critical region for direction $A$ of the alternative hypothesis ($TLL$ = ‘too long lived,’ $TED$ = ‘too early died’), using residuals of type $B$; $z_\alpha$ and $w_\alpha$ denote the $\alpha$ percentiles of standard normal and logistic distributions, respectively.

The estimated coverage of the critical regions (= empirical size) for uncensored and for 40% censored samples is given in Table 1 by significance level ($\alpha = 0.05$ and $\alpha = 0.01$), direction of alternative hypothesis, and sample size ($n = 20$, 80, and 200). The agreement of results by log-odds and normal deviate residuals permits giving results only for the former. The good empirical performance indicates the suitability of the proposed residuals for screening of outlying survival times.

4. Example and Aspects of Application
We illustrate an application of the proposed residuals in the screening for outlying survival times in a Veteran’s Administration randomized treatment study of prostate cancer. The data set has been published (Andrews and Herzberg, 1985) and provides a typical example for an analysis of prognostic factors.

Though screening for outlying observations is meaningful for data sets of almost any size, we have selected the subset of $n = 310$ patients below age 75 years because, for this smaller sample, graphical presentation of results can better be demonstrated. Censoring occurred in 34% of survival times.

For the analysis using Cox’s model, we have taken into account the treatment ($\leq 0.2$ mg diethylstilbestrol versus $>0.2$ mg) and six further dichotomous factors: weight index ($<100$, $\geq 100$), performance rating (normal, limitation of activity), serum haemoglobin (HG; $<12$ g/ml, $\geq 12$ g/ml), size of primary lesion ($<30$ cm$^2$, $\geq 30$ cm$^2$), Gleason stage/grade category (SG; $\leq 10$, $>10$), history of cardiovascular disease (HX; no, yes).

If we were interested in a search of unknown prognostic factors, we would suggest reexamining the medical records of those patients whose log-odds or normal deviate residuals lie in the lower or upper critical regions and of those patients with censored survival times that have a high probability of belonging to the ‘too long lived’ critical region if their times had been uncensored. Critical values—for suggested percentiles of distribution functions of 0.025, 0.975, 0.05, 0.95—are $\pm 1.96$ and $\pm 1.64$ for normal deviate residuals and $\pm 3.66$ and $\pm 2.94$ for log-odds residuals, respectively. These critical limits are used in a way analogous to reference ranges for clinical laboratory data used for medical check-ups. In the former case, new prognostic factors can be detected; in the latter case, hidden diseases in apparently healthy individuals.

For our example, we give a plot of normal deviate residuals versus the predictor $x \beta$ in Figure 2. We do not detect an association in this plot—taking into account that censored data residuals are represented by expected values for conditional distributions; but we identify four patients with residuals exceeding critical limits as defined by 0.025 and 0.975 percentiles.

From Table 2, we learn that patients 50 and 293, whose prognostic factors are very unfavourable, are still alive when their estimated survival probabilities are 0.019 and 0.060, respectively. The reverse holds true for patients 437 and 451, who had died unexpectedly early with respect to their favourable prognoses. Conclusions from log-odds and normal deviate residuals are identical. If expected medians are used instead of means, results are similar except for patient 293, who just misses the critical limits for the 0.025 percentile. This patient has a 42% chance ($= 0.025/0.060$) of belonging to the expected 2.5% of patients who most strikingly exceeded their expected survival time, while the same probability equals one for patient 50. For no other patient with censored survival did this chance exceed 30%.

While the plot for log-odds residuals looks very similar to that for normal deviate residuals (and therefore is not shown), the plots for Cox–Snell, martingale, and deviance residuals are given for comparison in Figure 2 as well. We recognize that the patterns for uncensored data and for censored data residuals agree for normal deviate and deviance residuals; however, deviance residuals’ values for uncensored and censored data are further apart, with a positive and negative shift for uncensored and censored values, respectively. The deviance residuals’ distribution in the prostate cancer example is in agreement with the more general results of Figure 1b, show-
ing an excess of extreme positive residuals (15 individuals are suggested as ‘too early died’). As discussed in previous sections, we do not recommend deviance residuals for outlier screening. The distributions of Cox–Snell and martingale residuals are clearly unsuitable for graphical presentation.

Multiple outliers may not only be indicative of some unknown favourable or unfavourable condition but may also indicate a weakness of a fitted model (e.g., that no satisfactory fit is provided for early deaths), in the data (e.g., that some regions in the factor space have inadequate coverage), or in both.

If, in the prostate cancer example, two interactions (HX × SG, HX × HG) that are significant at α = 0.05 are added to the original main effects model, four patients appear to have ‘died far too early’ but none to have ‘lived far too long’ (again for upper and lower tail probabilities of 0.025 each). It is often difficult to decide on a single best model, in particular if we have selected interaction or nonlinear effects terms from a great number of candidates. A particular choice of such terms may amount to a better fit of outlying observations but may not offer the chance to explore the possible presence of further prognostic factors. We emphasise that screening for outlying survival times is always exploratory and conditional on one or more finally accepted models.

5. Conclusions
In this paper, we have focused on screening for outlying survival times, which we think to be the most relevant usage of our residuals in practice. By deriving the distribution of the maximum observed log-odds or normal deviate residual, formal outlier tests can be obtained, as are available for the general linear model (see Section 1). While the practical relevance of such procedures is questioned in principle (cf., Beckman and Cook, 1983), censoring of survival times and the identification of outliers becomes ‘too early died’ as are suggested by such procedures.

By way of conclusion, we recommend routinely screening for outlying survival times by means of log-odds or normal deviate residuals whenever medical studies of survival are statistically analysed. For this purpose, a SAS macro ‘SURRES’ (with files surres.sas and surres.fortran) is offered by ftp to VM.AKH-Wien.ac.at, specifying the userid ‘Biometry.’

Acknowledgements
We thank Pierluigi Conti, University of Rome ‘La Sapienza,’ and Alexandra Kaider, Vienna University, for their valuable cooperation on theoretical and programming aspects of our project. We also acknowledge helpful comments by Terry M. Therneau, Mayo Clinic, USA, on an early version of this paper.

résumé
L’identification des individus “morts beaucoup trop tôt” ou “vivant beaucoup trop longtemps” par comparaison à leur probabilité de survie peut mener à la détection de nouveaux facteurs pronostiques. Les méthodes d’identification de données “aberrantes” sont généralement basées sur les résidus. Dans ce cadre, pour le modèle de Cox seuls les résidus de déviance ont été considérés, mais nous montrons que ces résidus ne conviennent pas vraiment. Nous proposons plutôt deux nouveaux types de résidus: les log-odds suggérés et les résidus normaux deviates sont simples, intuitivement attractifs, leurs propriétés théoriques et leurs performances empiriques les rendent très appropriés pour l’identification de données “aberrantes.” En dernier lieu, différents aspects pratiques de détection d’individus avec des temps de survie “aberrants” sont discutés à partir d’un exemple d’étude en cancérologie.

References
Cox Regression Residuals and Outlier Screening


APPENDIX

Derivation of Conditionally Expected Value of Suggested Residuals under Censoring

A1. Normal Deviate Residuals

The distribution of $N_i$ conditioned to the event \{ $N_i \leq n_i^c$ \} is given by

$$\Pr(N_i \leq x \mid N_i \leq n_i^c) = \left( \int_{-\infty}^{x} I_{(-\infty,n_i^c)}(y) \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \right) / \Phi(n_i^c),$$

where $I_W(y)$ is the indicator function of the set $W$. Therefore, the conditional density is

$$f_{N_i \mid N_i \leq n_i^c}(x) = \left( e^{-x^2/2} I_{(-\infty,n_i^c)}(x) \right) / \sqrt{2\pi} \Phi(n_i^c),$$

and the expected value of the conditional distribution is

$$n_i^c = \mathbb{E}(N_i \mid N_i \leq n_i^c) = \int_{-\infty}^{n_i^c} x \frac{1}{\sqrt{2\pi} \Phi(n_i^c)} e^{-x^2/2} dx = -\frac{1}{\sqrt{2\pi} \Phi(n_i^c)} e^{-n_i^c^2/2}.$$

Finally, replacing $n_i^c$ with $\Phi^{-1}\{S(t_i^c)\}$, we get

$$n_i^c = -\frac{1}{\sqrt{2\pi} S(t_i^c)} e^{-n_i^c^2/2}.$$

A2. Log-Odds Residuals

The derivation of log-odds residuals' conditional expected value follows the same steps as the previous one. In detail,

$$\Pr(L_i \leq x \mid L_i \leq l_i^c) = \left( \int_{-\infty}^{x} I_{(-\infty,l_i^c)}(y) \frac{1}{1 + e^{-y^2}} dy \right) / F(l_i^c),$$

where $F(\cdot)$ is the c.d.f. of the standardized logistic distribution. Then

$$l_i^c = \mathbb{E}(L_i \mid L_i \leq l_i^c) = \frac{1}{F(l_i^c)} \int_{-\infty}^{l_i^c} x \frac{e^{-x}}{(1 + e^{-x})^2} dx = \frac{1}{F(l_i^c)} \left\{ \frac{e^{l_i^c} - 1}{1 + e^{l_i^c}} \log \left( 1 + e^{l_i^c} \right) \right\} = l_i^c - \frac{1 + e^{l_i^c}}{e^{l_i^c}} \log \left( 1 + e^{l_i^c} \right).$$