1. Introduction

Cox proportional hazard model (Cox, 1972) is the most popular method for assessing covariates effects on time to event in the presence of censoring. Its contribution to the analysis of real data spread firstly in medical studies and from then it stemmed to become a relevant tool in various fields of application, such as economics, industrial reliability, agriculture, biological and physical sciences.

Cox regression and its extensions are today the standard for time to event data analysis. The wide spread of the methodology implies an increasing interest in diagnostic measures to assess the correct specification of the model, to evaluate goodness of fit and to detect outliers.

The nature of the data in survival analysis and the semiparametric structure of Cox model do not allow for a direct extension of the usual diagnostic tools from linear regression and generalized linear models. Its peculiarities make it hard to define a general diagnostic tool, thus the plethora of definition of residuals related to the Cox model.

In this work we suggest a diagnostic methodology based on the conditional contribution to the partial likelihood that is developed in analogy to standard tools in linear regression model.

The paper is structured as follows: in Section 2 we describe briefly the context in which our proposal is embedded, in Section 3 we define the conditional predicted residual in Cox model, Section 4 will present an exploratory simulation study to test the proposal while in Section 5 it will be tested on the famous kidney dataset (Collet, 2003). Finally in Section 6 some conclusions will be drawn.

2. Background

To make the paper self-contained the proportional hazard regression model proposed by Cox will be briefly recalled in this section.
In Cox model (Cox, 1972) and in survival analysis in general, (Kalbfleisch and Prentice, 2002), time is the dependent variable and ascertaining its dependence on covariates is the main goal of the analysis. We assume we have \( n \) cases and for each of them we consider two random variables: \( Y_i, i = 1, ..., n \), is the real time to event for this subject and \( C_i, i = 1, ..., n \) is a random variable representing the censoring process. In survival analysis \( C_i \) is assumed to be independent of \( Y_i \), or less restrictively non informative, with respect to \( Y_i \). On each unit we observe the vector \((t_i, \delta_i, z_i)\), \( t_i \) is the time until failure if the event has been observed (\( \delta_i = 1 \)), or the time until censoring (\( \delta_i = 0 \)), if the unit has been removed from the study for reasons not related to the event of interest.

The proportional hazard model represents the hazard function as the product of two components: a non-parametric part representing the hazard of an hypothetical subject with all covariates identically equal to zero (this component contains all the information of the effect of time on the hazard function), and a term, constant with respect to time, that accounts for the effects of the covariates.

The hazard function can then be written as:

\[
\lambda(t) = \lambda(t, z) = \lambda_0(t) \exp(\beta^T z)
\]

Where \( \beta \) is a \( p \)-vector of unknown parameters, \( z \) is a vector of known covariates, \( \lambda_0(\cdot) \) is usually referred to as baseline hazard. This is a proportional hazard model since the covariates act multiplicatively on the hazard; its semiparametric nature is due to the complete non parametric specification of the baseline hazard function and the complete parametric form of the term in which covariates are involved.

Estimation in Cox model is based on the partial likelihood function which is the part of the full likelihood independent of the underlying baseline hazard. Estimates obtained via the maximization of the partial likelihood have been shown to have similar features of those obtained using full likelihood although some information is lost, not negligible for small data set and informative censoring.

The partial likelihood \( L(\beta) \) is defined as follows: let \( t_{(1)}, t_{(2)} ... t_{(d)} \) be ordered observed failure times, \( (d \leq n) \) in increasing order and \( \phi_k = i \) if the \( i-th \) subject fails at \( t_k \) then

\[
L(\beta) = \Pi_k \Pr(\phi_k = i | H_k) = \Pi_k \frac{\exp(\beta^T z_k)}{\sum_{l \in R_k} \exp(\beta^T z_l)}
\]

Where \( H_k \) contains all the information on the history of both events and censoring up to the \( k-th \) event.
Several goodness of fit tests have been proposed to evaluate global and local fitting, correct specification of the model, in its general hypothesis of proportionality and in the specification of its regressive components. Among these, martingale residuals (Barlow and Prentice, 1988) are worth mentioning, since they will be used as reference during the study. They have a skewed distribution and are defined as the difference between the indicator variable $\delta_i$ and the cumulative hazard assigned by the model to an individual with failure time $t_i$ (Therneau, Grambsch, Fleming 1990)

$$M_i = \delta_i - \Lambda_i$$

where $\Lambda_i = \int_0^{t_i} \lambda_i(u)du$.

High values for $\Lambda_i(t_i)$, indicate high cumulative risk of death, therefore a highly negative martingale residual.

Martingale residual properly transformed in order to gain symmetricity and nearly normality are generally called deviance residuals

$$\text{dev}(M_i) = \text{sgn}(M_i)\sqrt{-2(M_i + \delta_i \ln(\delta_i - M_i))}$$

Despite their name, $\text{dev}(M_i)$, as defined above, are not real deviance residuals. Standard deviance residuals, as introduced by Pregibon in generalized linear model, assume the meaning of a discrepancy measure between estimated and observed values. This discrepancy is evaluated by comparing the contribution to the likelihood for each single observation in the current model and in the full model, while deviance residual for the Cox model as have been defined above, do not refers directly to any likelihood, neither the complete one, nor the partial one on which model fitting and estimates are based upon. The reason why they are called deviance residuals is that their definition resembles deviance residual in Poisson regression, but, as a nuisance parameter, the unspecified baseline hazard is still involved.

While martingale residuals and therefore deviance residuals, are evaluated at time conditional on covariates, Shoenfield (Schoenfeld, 1982) residuals compare the observed values of covariates with the corresponding expected value returned by the model.

$$r_{sch_i} = X_i(t_i) - E(X_i(t_i)|\beta)$$

The result will be a matrix with as many rows as events and a column for each covariate. Shoenfield residuals are evaluated with respect to partial likelihood on a covariate scale and are not defined for censored units.
3. Conditional predicted residual in Cox model

Our proposal, starting from the standard deviance residual definition, suggests a residual based on the contribution of each single unit to the partial likelihood.

A first step is therefore to determine the contribution to the likelihood when dealing with partial likelihood.

When independence holds, the standard likelihood is a product of terms, each associated to a single unit, therefore the contribution to the likelihood is easily identified. Instead partial likelihood is factorized over time or, equivalently with respect to the risk set. Then the contributes of each unit to the partial likelihood is present in more than one term in the factorization; nonetheless they are still well defined and not independent because they are conditional to the process $H_k$, the history of the process till the $k-th$ event.

Namely, a subject that experienced the event at time $t(k)$ contributes to the likelihood with the product of two terms, the first is its contributes at time $t(k)$

$$\frac{\exp(\beta^T z_k)}{\Sigma_{i \in R_k} \exp(\beta^T z_i)}$$

But it also contributes via its presence in the risk set of the previous events. This contributes are summarized in a product of $k-1$ factors, as follows

$$\Pi_{s=1}^{k-1} \frac{\Sigma_{i \in R_{s+1} \neq k} \exp(\beta^T z_i)}{\Sigma_{i \in R_s} \exp(\beta^T z_i)}$$

The last term is defined for all observations even for censored ones. According to the structure of the partial likelihood the contribution of each unit depends on the ranking it occupies with respect the sequence of events. Time between events is not involved directly or indirectly.

Due to the semi-parametrical nature of Cox proportional hazard model, it is not possible apply the notion of saturated model in this framework. In fact if we define a saturated model as one that reproduces exactly the observed data, this requirement is satisfied by a null model with only the non parametric component. On the contrary, if we define it as the one that has one parameter for each observation, the structure of the model itself does not allow to reproduce the whole information. In both cases we will have to refer to the baseline function that is actually not involved in partial likelihood.

We suggest to compare the contribute to the partial likelihood,
\[
\frac{\exp(\beta^T z_k)}{\sum_{l \in R_k} \exp(\beta^T z_l)} \exp(\beta^T z_k) \prod_{s=1 \atop s \neq k}^{k-1} \frac{1}{\sum_{l \in R_s} \exp(\beta^T z_l)} \exp(\beta^T z_l)
\]

With a conditional maximum contribute

\[
\frac{\exp(\beta^T z_{k^*})}{\sum_{l \in R_{k^*}} \exp(\beta^T z_l)} \exp(\beta^T z_{k^*}) \prod_{s=1 \atop s \neq k^*}^{k^*-1} \frac{1}{\sum_{l \in R_s} \exp(\beta^T z_l)} \exp(\beta^T z_l)
\]

Where \( k^* \) is the predicted rank for the \( k \)-th unit, estimated on the base of the other \( n - 1 \) units, i.e. the rank that maximized the contribution to the partial likelihood for that unit conditional to the other \( n - 1 \); risk set \( R \) above are adjusted to the corresponding ranking.

This proposal gets further support from an analogy with a similar quantities in normal linear model. In fact it can be looked at as a generalization of predicted residuals that compare observed value with estimated value to which the considered unit has not contributed to.

4. An Exploratory Simulation Study

A brief simulation study to test the ability of the proposed method to detect outliers will be presented. Datasets of \( n=40 \) randomly generate units will be considered. For each unit two covariates, one quantitative \( (X_{i1}) \) and one qualitative \( (X_{i2}) \), will be randomly drawn according to the following scheme:

\[
X_{i1} \sim N(\mu = 0, \sigma = 1) \quad X_{i2} \sim Bin(1, \theta = 0.5) \quad i = 1, ..., n
\]

Times to event have been generated according to:

\[
T_i = \text{abs} \left[ \frac{1}{x_i^T \beta} + N(0, \sigma) \right], \quad \sigma = \{0.005, 0.008\}, x_i = (X_{i1}; X_{i2}), \beta = (1; 2) \quad (1)
\]

Generating times to event according to equation (1) allows the same Gaussian noise to have different effect according to the value of the linear predictor, namely the same noise will have a stronger effect on units with large values of the linear predictor, therefore simulating the presence of possible outliers in the same dataset. In Figure 1 the allocation matrices for the optimal ranking for both levels of noise have been displayed. The elements in the dataset that already have a ranking optimal conditional to the remaining \( n-1 \) have been displayed in green; red has
been used to depict units that are not allocated in an optimal position according to the estimated model. The green dots are mainly associated with small values of the linear predictor, for whom the noise is less effective. As it was to be expected, a smaller level of noise implies a smaller displacement needed to optimize the ranking of the data; with increasing noise the ranking that the units have in the dataset are less and less optimal.

**Figure 1 – Allocation matrices for the simulated data with increasing level of noise**

![Allocation matrices](image)

In Figure 2 the standardized conditional residuals (in red) and the standardized deviance residuals (in black) have been displayed for the simulated sample for both levels of noise. On the right hand side part of the picture the values of $\log(t)$ and of the linear predictor for each element in the dataset have been displayed. Bars have been colored according to whether they were considered outliers (i.e. $abs(residual) \geq 2$). With the smaller level of noise both deviance residuals and conditional residuals select the same points in the dataset as outliers (blue bars in the top right quadrant of Figure 2). With increasing noise, one point has been selected by the proposed method as being an outlier (the red bar in the bottom right quadrant of Figure 2) and one point has been selected by both deviance and conditional residuals as being an outlier (blue bar).

It is worth noticing that the element that has been selected as having too high a residual by the proposed diagnostic tool is actually a unit that, considering the value of the linear predictor, should have died much later then when it actually did, conditionally to the other $n-1$ elements in the dataset.
5. Kidney Infection Data

In this section the proposed residuals will be computed on a vastly used dataset in survival data analysis, namely the catheter infection data (McGilchrist and Aisbett, 1991). For our purposes we will be using only a subset of those patients, as in Collett (2003).

In the treatment of some kidney related diseases, dialysis could help remove waste materials in the blood. The use of such a technique could result in the development of an infection at the site where the catheter is inserted. To cure the infection, the catheter must then be removed. McGilchrist and Aisbett (1991) recorded the time from insertion until infection (in days) for a group of kidney patients. It is possible that the catheter may be removed for reason other than infection; such a case results in right censored data. Following Collet’s approach
only 13 out of 39 patients, namely those suffering from glomerulo neptiritis, acute neptiritis and polycytic kidney disease have been removed from the analysis and only those with diseases coded as type 3 (“other”) in the original paper from McGichrist and Aisbett, have been considered.

In Table 1 for each of these 13 patient, the time, the status (1=infection, 0=other reasons), the age and gender have been displayed. Only one right censored observation is present in the dataset, all the others experienced the event due to infection during follow-up.

### Table 1 – Kidney infection data (Collet, 2003)

<table>
<thead>
<tr>
<th>Patient</th>
<th>Time</th>
<th>Status</th>
<th>Age</th>
<th>Gender</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>1</td>
<td>28</td>
<td>M</td>
<td>Other</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>1</td>
<td>44</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>1</td>
<td>32</td>
<td>M</td>
<td>Other</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>1</td>
<td>16</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>1</td>
<td>10</td>
<td>M</td>
<td>Other</td>
</tr>
<tr>
<td>6</td>
<td>54</td>
<td>0</td>
<td>42</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>7</td>
<td>119</td>
<td>1</td>
<td>22</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>8</td>
<td>141</td>
<td>1</td>
<td>34</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>9</td>
<td>185</td>
<td>1</td>
<td>60</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>10</td>
<td>292</td>
<td>1</td>
<td>43</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>11</td>
<td>402</td>
<td>1</td>
<td>30</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>12</td>
<td>447</td>
<td>1</td>
<td>31</td>
<td>F</td>
<td>Other</td>
</tr>
<tr>
<td>13</td>
<td>536</td>
<td>1</td>
<td>17</td>
<td>F</td>
<td>Other</td>
</tr>
</tbody>
</table>

Only 3 male subjects are present in the dataset and they all experienced the event within a month from the catheter insertion.

A Cox model, considering Age and Gender as covariates, has been fitted to the data, resulting in the following estimates (Table 2):

### Table 2 – Estimates of the parameter of cox proportional hazard model for kidney infection data

<table>
<thead>
<tr>
<th>covariate</th>
<th>coef</th>
<th>exp(coef)</th>
<th>Std. err</th>
<th>Z</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.0304</td>
<td>1.0308</td>
<td>0.0262</td>
<td>1.16</td>
<td>0.250</td>
</tr>
<tr>
<td>Gender(F)</td>
<td>-2.7108</td>
<td>0.0665</td>
<td>1.0959</td>
<td>-2.47</td>
<td>0.013</td>
</tr>
</tbody>
</table>

As it is also quite clear from the data in Table 1, patients in the dataset tend to show an increasing risk with age and a significant effect of gender, females being more resilient to developing the infection.

In Figure 3 the time to event for each patient and the time vs linear predictor have been displayed. Patients that showed a standardized deviance residual (resdev) or a standardized conditional residual (rescond) greater than 2 have been
highlighted in the plots. According to the deviance residuals there is no patient showing a behavior that is not coherent with the fitted model (resdev≥2). Conditional residuals, on the other hand, singled out the 4th patient in the dataset (red bar in Figure 3), i.e. a 16 years old female that developed an infection after 24 days and that shows a standardized conditional residual greater than 2.

**Figure 3 – Time to event for each unit in the kidney data and linear predictor vs time.**

In Table 3 the original ranking and the optimal ranking assigned according to the best conditional contribution to likelihood have been displayed.

Patient number 4, being a Female aged 16, should develop an infection much later in time, since young age is a protective factor and females have a lower risk of developing the infection according to the fitted model. This is also confirmed by the plot of the linear predictor vs time in Figure 3. The time to event tends to decrease with the increase of the value of the linear predictor. Patient 4 shows a very small value of the linear predictor, associated with a too small value of the time to event.

The optimal ranking method would, conditionally to the other 12 patients, expect her to develop the infection after all the other patients. She is clearly an anomalous patient since she develops the infection sooner than a male with almost the same age (patient 5).
Table 3 – Observed and Optimal ranking for kidney infection data obtained via the proposed method.

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Gender</th>
<th>Time</th>
<th>offset</th>
<th>Original Ranking</th>
<th>Assigned Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28</td>
<td>M</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>F</td>
<td>15</td>
<td>7</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>M</td>
<td>22</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>F</td>
<td>24</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>M</td>
<td>30</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>F</td>
<td>54</td>
<td>5</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>F</td>
<td>119</td>
<td>5</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>F</td>
<td>141</td>
<td>2</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>60</td>
<td>F</td>
<td>185</td>
<td>5</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>43</td>
<td>F</td>
<td>292</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>F</td>
<td>402</td>
<td>1</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>31</td>
<td>F</td>
<td>447</td>
<td>1</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>F</td>
<td>536</td>
<td>0</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

6. Conclusions

In this paper, a new approach to determining the residuals for the Cox model has been suggested. The proposal is in line with the idea of deviance residuals in generalized linear models, i.e. the difference between the log likelihood for the full model and the log likelihood for the fitted model. Since there is no unique definition of full model in Cox regression, the most common alternatives available in the literature are not based on likelihood. Our proposal is based on determining the optimal ranking of each element in the dataset conditionally on the ranking of the other \( n-1 \) elements. This is achieved by selecting, for each unit, the rank that yields the best conditional contribution for that subject, to the log likelihood of the model, given the ordering of the other \( n-1 \) subjects. The proposal is still under testing and has been tried on a small simulation study and on a real benchmark dataset, both times yielding very promising results. As future agenda we plan on testing the proposal on a large simulations study.

Bibliography


**SUMMARY**

Diagnostic tools based on optimal ranking in the Cox Model

Parameter estimates for Cox proportional hazard model are achieved via the maximization of the partial likelihood. Nonetheless, diagnostic tools and local fitting measures (residuals) are based on the complete likelihood.

Partial likelihood in Cox model entails a factorization of the contributes of each unit based on the ranking of each unit according to the time they have experienced the event. This contribute is conditional on the units that have experienced the event before the unit that is being considered. Such a structure suggests the possibility to use diagnostic tools based on the conditional contributes of each unit to the partial likelihood. In this paper we propose a diagnostic approach based on the optimal ranking of each unit conditional to the others.

________________________________________

 Luciano NIEDDU, UNINT – Università degli Studi Internazionali di Roma, l.nieddu@unint.eu
Cecilia VITIELLO, Sapienza, Università di Roma, Dipartimento di Statistica, cecilia.vitiello@uniroma1.it