Visualising Survival Data Regression Models
Using Pseudo-Observations

Maja Pohar Perme
Department of Biomedical Informatics, University of Ljubljana
Vruzoj trg 2, 1000 Ljubljana, Slovenia
E-mail: maja.pohar@mf.uni-lj.si

Per Kragh Andersen
Department of Biostatistics, University of Copenhagen
O. Farimagsgade 5, PB 2099, DK 1014 Copenhagen K, Denmark
E-mail: pka@biostat.ku.dk

Abstract. Methods for visualising data are an essential part of model fitting procedures and are commonly used within all fields of statistics. Various graphical checks can be performed, either using scatter plots of the data itself or some kind of informative residuals. However, in survival data, with the existence of censored observations as one of its defining properties, these elementary plots are not meaningful as the censored observations cannot be sensibly plotted. In this paper, we review a recently introduced general solution of this problem that is based on pseudo-observations. These are defined for each individual at any point of the follow-up time and therefore offer a way around the censoring problems. Using pseudo-observations, we can apply methods analogous to those in regression with binary outcomes and plot various kinds of scatter plots or residuals that give an important insight into the quality of the data fit. An important property of this approach is that it applies to any hazard regression model, with the Cox and the additive model being the focus of this paper. We describe methods for single as well as multiple covariate cases and illustrate them using simulated data sets.

Keywords. graphical goodness of fit methods, proportional hazards, additive hazards, pseudo-observations, regression models, survival data

1. Introduction

For data without censoring, graphical checks are often performed using either various kinds of informative residuals or scatter plots of data itself. Often, a smoothed curve is superimposed on these plots, this is especially essential in the case of binary data. In the case of hazard regression models for survival data, these graphical methods do not apply because of the presence of censored observations. In this paper, we shall focus on two most commonly used regression models, that is the Cox model [3], where the hazard function \( \alpha(t) \) is specified as

\[
\alpha(t|Z) = \alpha_0(t)e^{\beta Z}
\]  

(1)

and the additive model of Lin and Ying [4] with the hazard function defined as

\[
\alpha(t|Z) = \alpha_0(t) + \beta Z.
\]  

(2)

Here, \( \alpha_0(t) \) denotes the baseline hazard, vector \( Z \) stands for the covariates in the model and vector \( \beta \) for its coefficients. Each of these models has two assumptions:

- coefficient in \( \beta \) is constant in time. This corresponds to the proportional hazards (PH) in the case of the Cox model and time-constant hazard differences in the additive model;
- linearity of the covariate effect (as opposed to a more general function \( g(Z) \) in equations (1) and (2)).
2. Pseudo-observations

Pseudo observations are defined \[1\] for each individual \(i\) and for any time \(t\) within the follow-up period as

\[ S_i(t) := n \hat{S}(t) - (n - 1) \hat{S}_{-i}(t) \quad (3) \]

Here, \( \hat{S}(t) \) denotes the Kaplan-Meier estimate of the survival curve at time \(t\) based on the whole sample (of size \(n\)) and \( \hat{S}_{-i}(t) \) denotes the Kaplan-Meier estimate based on the sample from which the individual \(i\) has been excluded. When no censoring is present in the sample, the pseudo-observations equal to the survival indicator for each individual \(I(T_i > t)\). The property that makes pseudo-observations truly useful is the fact that they are still defined in the case of censoring; a value is defined for any censored individual even after his time of censoring, that is at times when the indicator \(I(T_i > t)\) can no longer be observed.

The definition (3) implies that the pseudo-observation \(S_i(t)\) can be intuitively understood as the individual contribution to the overall Kaplan-Meier estimate. And while these individual values may be hard to interpret by themselves, their average equals to the Kaplan-Meier estimate \((1/n \sum S_i(t) = \hat{S}(t)\) in all but the last follow-up time) and the average of pseudo-observations over a subgroup of individuals is very close to the Kaplan-Meier estimate for that subgroup, provided that the censoring distribution does not differ between subgroups. The important fact to note here is that pseudo-observations are calculated using only the survival times and censoring status, and can be later averaged within subgroups with respect to the covariates of interest.

3. Plotting on survival scale

The effect of a categorical variable on survival can be easily visualised by comparing the Kaplan-Meier curves within each group defined by the variable. However, the Kaplan-Meier estimate is always representing a group experience and therefore cannot be used in the case of a continuous variable. On the contrary, the pseudo-observations are given for each individual. Therefore, to get an idea about the effect of a continuous variable, we order the individuals with respect to that covariate and smooth them (the individual values are too varied to give a clear picture) - this gives us a nonparametric estimate of \(S(t|Z)\). A plot of the pseudo-observations smoothed over variable and time is given on Figure 1. As three-dimensional plots may be quite hard to read, Figure 1 can be projected to a two-dimensional plot with profile curves corresponding to chosen time-points, see Figure 2.

While these plots give a good idea of the effect of the covariate, they provide little information on the sensibility of a model we might want to fit to the data. To get an idea about that, we turn to comparing the observed data to the fitted model. We define pseudo-residuals as the difference between the observed values, represented by the pseudo observations \(S_i(t)\), and the fitted values \(\hat{S}(t|Z_i)\) that can be calculated for each individual using the estimated coefficients. We standardize these differences to get what
we define as standardized pseudo-residuals:

\[ \hat{\epsilon}_i(t) := \frac{S_i(t) - \hat{S}(t|Z_i)}{\sqrt{\hat{S}(t|Z_i)[1 - \hat{S}(t|Z_i)]}}. \]  

(4)

In graphical checking of models (1) and (2) we propose plotting the residuals \( \hat{\epsilon}_i(t) \) with respect to the covariate in question at several chosen time-points and superimposing a smoothed average. In this way, each individual plot can give information about the functional form of the covariate, while the comparison of plots at different time-points can give an idea of the behavior of the covariate effect in time. In this way, the residuals present a very general approach that can provide a good initial illustration of the model fit. However, it turns out that the effects may need to be rather large to be observed. And as particular deviations from a given model may be difficult to read from graphs in the survival probability scale, it proves to be useful to consider an appropriate transformation of the scale to get a more specific information. We deal with this idea in the next section.

4. Checking model specific assumptions on transformed scales

Considering again Figure 2, we can see clearly that survival improves with increasing values of the \( Z \) covariate, however, the curvature of the profile curves does not provide any specific additional information. The Cox model (1) survival function given the covariates is calculated as

\[ S(t|Z) = e^{-A_0(t)\exp(\beta Z)}. \]  

(5)

where \( A_0 \) is the cumulative baseline hazard function. Therefore, to get an idea of the linearity of the covariate effect, we should rather turn to the cloglog scale in which the covariate effects are linear, provided that the data really follows the specified model:

\[ \log(-\log S(t|Z)) = \log A_0(t) + \beta Z. \]  

(6)

As for the effect of time - we can see that only the constant \( \log A_0(t) \) changes with time in (6), therefore, profile curves corresponding to the situation with \( \beta \) constant in time (PH assumption) should be parallel. As an illustration, we present Figures 3 and 4, with the former one corresponding to a situation with the data following a Cox model with a negative coefficient \( \beta \) (the curves are close to being linear and parallel) and the latter to a model in which the effect of the \( Z \) covariate is clearly quadratic and therefore the linearity assumption is violated (the curves are parallel but far from linear). Turning to the additive model...
pose for checking the additive model, is to smooth the pseudo-observations, transform them with the logarithm and plot them at chosen time points divided by those chosen time values. If the data follow the additive model this should, as in the Cox model case, result in parallel lines with the slope \( \beta \). Figure 5 presents a case of the effect changing in time (starting as a positive effect and then turning into being negative) while Figure 6 presents a situation in which both assumptions are violated simultaneously - the quadratic effect changes its sign with time.

(2), we can use the same idea, the only difference is in choosing the proper transformation function. In this case, the survival function is given by

\[
S(t|Z) = e^{-[A_0(t)+\beta Z t]} \tag{7}
\]

and an expression that is linear in the covariate can be obtained by the logarithmic transformation:

\[
\frac{-\log S(t|Z)}{t} = \frac{A_0(t)}{t} + \beta Z. \tag{8}
\]

Therefore, the plotting procedure we pro-

5. The multivariate case

All the methods described in the previous section apply to the case of one covariate. However, when we have, for example, two covariates in the Cox model, the data will no longer follow the Cox model if we only focus on one of them and ignore the other. In other words, there is no reason the curves on the cloglog scale with respect to a certain covariate are linear and parallel if more than one covariate was included in the Cox model that generated the data. The same observation holds for the additive model, though in that case only mutually

Figure 4: A violation of the linearity assumption in the Cox model.

Figure 5: Additive model with the coefficient \( \beta \) changing in time.

Figure 6: Additive model with both the coefficient \( \beta \) changing in time and the effect of \( Z \) being non-linear.
dependent covariates are problematic.

The additional step we therefore need to perform in the multivariate case is to compare the actual curves on the appropriate scale to the curves we expect if the data follows the model. We illustrate the procedure on a data set, simulated to follow the Cox model case with two covariates \( Z_1 \) and \( Z_2 \). Figure 7 presents the curves with respect to covariate \( Z_1 \). These curves need to be compared to Figure 8, that presents the values of the fitted model (with both covariates), ordered with respect to the \( Z_1 \) and smoothed. We can conclude that the observed curves on Figure 7 correspond well to the fitted ones (Figure 8) and therefore there is no indication of the goodness of fit problems. This can be also seen from Figure 9 that presents the difference of the two graphs. The term \( \beta_1 Z_1 \) could be added to this last graph to get a plot that can be interpreted in the same way as those in the previous section.

![Figure 7: The cloglog transformed smoothed pseudo-observations from a data set generated with a Cox model with two covariates, plotted with respect to \( Z_1 \).](image1)

![Figure 8: The cloglog transformed smoothed predicted survival (based on the Cox model fit using both covariates).](image2)

![Figure 9: The difference of the graphs presented on Figure 7 and 8.](image3)

6. Conclusion

Pseudo-observations present a general approach to visualising survival data and checking the assumptions of any hazard regression model. Though alternative methods already exist for most of the presented graphs and checks \([2,8,9,5]\), they are all tied to the model specific ways of tackling the censoring problem. Furthermore, an important property of the graphs presented in Sections 4 and 5 is that both violations can be inspected simultaneously, a property that is only matched by the martingale difference residuals plots introduced by Sasieni and Winnet \([7]\) for the Cox model case. Another favourable property of the pseudo-observations approach is that all the
graphs presented in Sections 4 and 5 can be supplemented by goodness of fit tests [1]. To further facilitate the practical use of the described methods, an R function is available from the internet address: www.mf.uni-lj.si/ibmi-english/biostat-center/programje/pseu.r, the many details skipped in this review can be found in [6].

References


