Survival Analysis for Worcester Heart Attack Study

Ryan Cooper

May 18, 2017

1 Introduction

The Worcester Heart Attack Study data set consists of right-censored survival data from 500 patients who experienced an acute miocardial infarction. In this report we model the survival rate using a proportional hazards model. The first part of the analysis consists of analyzing important covariates in predicting survival. These covariates are then used in a backward selection procedure to fit a proportional hazards model. Next we look at the explained variation in some of the models created in the selection process. Lastly, we use the residuals from the model to check the proportional hazards assumption as well as the functional form of the covariates.

2 Analysis

In this data set the time till death for each patient is measured in days from the hospital admission for acute myocardial infarction. Among the 500 patients in the study 215 patients died. To model the time till death, we assume the data follows the Cox Proportional Hazards (PH) model,

$$\lambda(t|\boldsymbol{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\boldsymbol{Z}}$$

where $\boldsymbol{Z}$ is the vector of covariates listed in Table 1. For each covariate we fit a univariate PH model. Table 1 shows the p-value resulting from the partial likelihood ratio test. From the table we see that all variables except for av3 and cvd are significant at the 0.05 level.

Next we use a backward stepwise selection procedure to obtain a parsimonious model. In our model selection procedure we leave out the covariate Cardiogenic Shock due to the lack of patients with this symptom (only 17 out of the 215 non-censored patients had cardiogenic shock). Starting with all covariates significant at the 0.2 level, we use backward stepwise selection removing non-significant variables at level 0.10. Table 2 shows the resulting main effects model with covariates age, gender, hr, diasbp, bmi and chf. At this stage we add interactions to the model and again use backward stepwise selection removing non-significant variables at level 0.05. The only interaction that survived the selection procedure was age by gender. Table 3 shows fit of the pruned main effects interaction model. Note that the partial likelihood ratio test comparing the main effects model to the final model is 0.008. Hence the interaction in the final model provides a significant improvement in the fit.

Label	Description	Value	p-value
age	Age at hospital admission	Years	< 0.001
gender	Gender	Male/Female	0.006
hr	Initial Heart Rate	Beats per minute	< 0.001
sysbp	Initial Systolic Blood Pressure	$mmHg$	0.041
diasbp	Initial Diastolic Blood Pressure	$mmHg$	< 0.001
bmi	Body Mass Index	kg/m^2	< 0.001
cvd	History of Cardiovascular Disease	Yes/No	0.084
afb	Atrial Fibrillation	Yes/No	0.002
sho	Cardiogenic Shock	Yes/No	< 0.001
chf	Congestive Heart Complications	Yes/No	< 0.001
av3	Complete Heart Block	Yes/No	0.252
miord	MI Order	First/Recurrent	0.002
mitype	MI type	Q-wave/non Q-wave	< 0.001

Table 1: Regression variables with partial likelihood ratio test p-value for determining variable significance in predicting survival

	coef	exp(coef)	se(coef)	z	p
age	0.050	1.051	0.007	7.558	0.000
genderFemale	-0.272	0.762	0.144	-1.891	0.059
hr	0.011	1.011	0.003	3.839	0.000
diasbp	-0.011	0.989	0.004	-3.022	0.003
bmi	-0.045	0.956	0.016	-2.781	0.005
chfYes	0.779	2.180	0.147	5.314	0.000

Table 2: Estimated coefficients in the main effects model with standard error and two-sided Wald test p-value.

To get a sense of how much the model explains the variation in survival time, we look at a measure of explained variation. More specifically, we calculate Generalized R^2 for some of our PH models. This is defined in terms of the Kullback-Liebler information. For the PH model the sample R^2 is given by

$$\hat{\rho}^2 = 1 - e^{-\Gamma}$$

where

$$\Gamma = 2(\log \boldsymbol{L}(\hat{\boldsymbol{\beta}}) - \log \boldsymbol{L}(\mathbf{0}))/k$$

and k is the number of events. The Generalized R^2 for the main effects model and the final model are 0.619 and 0.632, respectively. Hence the explained variation between the two models is roughly the same.

Lastly, we asses the proportional hazards assumption and the functional form of the covariates in the final model using test statistics proposed by Lin et al. (1993) and Wei (1984) based on cumulative residuals. Table 4 shows the p-values for the score process test for proportionality and the test for cumulative martingale residuals. The p-value for the score process test shows that the proportional hazards assumption is appropriate for all covariates in the final model at the 0.05 significance level. The p-value for the cumulative martingale

	coef	exp(coef)	se(coef)	z	p
age	0.062	1.064	0.008	7.455	0.000
genderFemale	2.257	9.554	0.960	2.350	0.019
diasbp	-0.011	0.989	0.003	-3.119	0.002
hr	0.011	1.011	0.003	3.869	0.000
bmi	-0.044	0.957	0.016	-2.740	0.006
chfYes	0.778	2.177	0.146	5.341	0.000
age:genderFemale	-0.032	0.968	0.012	-2.651	0.008

Table 3: Estimated oefficients in the final model with standard error and two-sided Wald test p-value.

	Score p-value	MG p-value
age	0.648	0.626
gender	0.552	NA
hr	0.342	0.58
diasbp	0.538	0.694
bmi	0.518	0.082
chf	0.638	NA
age:gender	0.394	0.562

Table 4: p values for Score Process test and Cumulative Martingale residuals.

residuals test shows that the scale for all the covariates is acceptable with the possible exception of bmi. Similarly, Figure 1 shows the observed cumulative residuals over random realizations of the cumulative residuals under the model assumption. Again we conclude that the functional form of the continuous covariates in the final model is appropriate with the possible exception of bmi.

Figure 1: Observed cumulative residuals versus continuous covariates with 50 random realizations under the model.