Summary Report of a Survival Analysis
Using Proportional Hazards Models

Purpose

1) Run a complete data analysis using an assumed Proportional Hazards (PH) model;
2) Select the appropriate parameters for the PH model;
3) Assess the fit and perform inference.

Data description

The data consists of 863 kidney transplant patients.

We consider the data set from a study designed to assess the effect of some factors on the survival time of patients who took kidney transplant. The *TIME* variable contains survival time or on-study time in days after a kidney transplant. The variable *STATUS* has a value of 1 (dead) for those events at time, and has a value of 0 (alive) for those right censored.

The covariates included in the analyses are:

(i) *gender*: 1 = male, 2 = female;

(ii) *race*: 1 = white, 2 = black;

(iii) *age*: age in years.

We assume a PH model beforehand and let *gamma = 1269* for the baseline hazard rate $\lambda_0(t)$.

Software for analyzing

In this case, we cover the basics of modeling using the R software package. R is open source and can be downloaded from http://www.r-project.org/.

The following topics are addressed in this case:

- Import data into R;

- Fit Proportional Hazards (PH) models and find the appropriate reduced model, including: variable selection using the Akaike Information Criterion[1] (AIC) and the F-test, obtaining regression coefficients;
- Influence analysis for the PH model fit;
- Obtain inferences for parameters of interest, including: the hazard ratio for any possible covariate combination of the reduced model, and correspondent survivor curves;
- Model diagnostics.

PH model

The PH model is the most general of the regression models since it is not based on any assumptions concerning the nature or shape of the underlying survival distribution[2]. Survival analysis typically examines the relationship of the survival distribution to covariates. Most commonly, this examination entails the specification of a linear-like model for the *log* hazard[3].

Let t_i be a random variable denoting the failure time for the *ith* subject, and let x_{i1}, x_{i2}, …, x_{ip} be the values of p covariates for that same subject. In this case, we assume that $\underline{X_i}$ is independent with time. A PH model is given as

$$\lambda_i(t;\alpha,\beta,\underline{x}) = \lambda_0(t,\alpha)g(\underline{x},\beta)$$

Since the hazard function of $\lambda_i(t;\alpha,\beta,\underline{x})$ is strictly positive, consider the *log* link,

$$\log(g(\underline{x},\beta)) = \underline{x}^T\beta,$$

this means that: $g(\underline{x},\beta) = \exp\{\underline{x}^T\beta\}$.

Then, the PH model can be usually expressed as Cox Proportional Hazards Model:

$$\lambda_i(t;\alpha,\beta,\underline{x}) = \lambda_0(t,\alpha)\exp\{\underline{x}^T\beta\}.$$

Assume beforehand that this PH model has a baseline hazard rate as:

$$\lambda_0(t,\alpha) = \exp\{\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3\}$$

[1] Model selection, http://www.webpages.uidaho.edu/~brian/stat401ch13_01.pdf.
[2] Survival/Failure Time Analysis. http://www.statsoft.com/TEXTBOOK/stsurvan.html.
[3] Fox, J. 2002. Cox Proportional-Hazards Regression for Survival Data. http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

Where $(t-\gamma)_+ = \max\{0, t-\gamma\}$ and γ = 1269.

Thus, the hazard ratio for subjects with covariate vectors $\underline{x}_a$ and $\underline{x}_b$ is

$$\frac{\lambda_a(t;\alpha,\beta,\underline{x}_a)}{\lambda_b(t;\alpha,\beta,\underline{x}_b)} = \exp\{(\underline{x}_a^T - \underline{x}_b^T)\beta\}$$

Under a discrete time framework in this case, the cumulative hazard function

$$H(t_k) = \sum_{j=1}^{k} \lambda(t_k),$$

and the survival function is given as:

$$S(t_k) = \exp\{-H(t_k)\}.$$

PH model analysis

We mainly use the *coxph()* function in R to perform the analysis.

- Obtain parameter estimates

  The Cox PH model used in *coxph()* leaves the baseline hazard rate λ_0 unspecified. However, in this case, the baseline hazard rate $\lambda_0(t,\alpha) = \exp\{\alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 (t-\gamma)_+^3\}$ was assumed beforehand, so we take two major steps to obtain all parameter estimates: (i) use the *coxph()* in R to get the estimates of β; (ii) use the *lm()* function in R to get the estimates of α given the estimates of β.

- Select variables

  Covariate data are available for several comparisons (e.g., gender, race, and age), and in such cases, it is necessary to determine which variables should be included in the fitted PH model.

  Variable selection for β can be performed using **a forward and backward stepwise procedure** that searches all possible models to determine which model minimizes the AIC (R package: *MASS*; R function: *stepAIC*). This approach adds covariates to the model when this improves goodness of fit, but does not generate an overfit model with unnecessary covariates, since the AIC includes a penalty term for each explanatory variable added to the model.

Variable selection for α can be performed using the F-test generated by the *lm()* function. Or we can use R function *stepwise()* to obtain the same results more readily.

Residual analysis or influence analysis can be applied to verify the variable selection.

- Perform diagnostic analyses to evaluate the adequacy of model fit

  Residual analysis is used to evaluate whether or not observations are characterized by the model. In this case, the assumption of a PH model appears to be supported.

  Model fit can be evaluated based upon a graphical comparison between the empirical Kaplan–Meier survival curves and the fitted survival curves generated from the final reduced PH model. The results show that the final reduced PH fits the data well.

Conclusion

We obtain the final reduced PH model as follows.

	Parameter	Estimate	Std. Error	z or t	P	
Reduced model	β_3 age	0.051068	0.007136	7.156	8.3e-13	***
	α_0 (Intercept)	-8.926e+00	2.397e-02	-372.458	< 2e-16	***
	α_2 haz.time2 (t^2)	7.362e-07	8.529e-08	8.631	2.36e-14	***
	α_3 haz.time3 (t^3)	-2.514e-10	4.658e-11	-5.397	3.28e-07	***
	α_4 new.time3 $(t-\gamma)_+^3$	5.679e-10	1.105e-10	5.140	1.03e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1 Parameter estimates of the reduced PH model

The final reduced PH model in this case can be expressed as:

$$\lambda_i(t;\hat{\alpha},\hat{\beta},\underline{x}) = \lambda_0(t,\hat{\alpha})\exp\{\underline{x}^T\hat{\beta}\}$$

$$= \exp\{\hat{\alpha}_0+\hat{\alpha}_2 t^2+\hat{\alpha}_3 t^3+\hat{\alpha}_4(t-\gamma)_+^3\}\exp\{x_{age}\hat{\beta}_3\},$$

Where $\hat{\beta}_3$*=0.051068*, $\hat{\alpha}_0$*=-8.926e+00*, $\hat{\alpha}_2$*=7.362e-07*, $\hat{\alpha}_3$*=-2.514e-10*, $\hat{\alpha}_4$ *=5.679e-10*, $(t-\gamma)_+ = \max\{0,t-\gamma\}$ and γ = 1269.

We can conclude that:

- The variables *gender* and *race* play **little** effect on the kidney transplant patients. It means that: no matter males or females and no matter blacks or whites, these factors of *gender* and *race* may be insignificant in this case.
- From Table 1, $\hat{\beta}_3 = 0.051068$, the standard error for the relative risk is *0.007136*. Since $exp(\hat{\beta}_3) = 1.052394$, this result indicates that the relative risk of dying increases about 5% for each one year increase in age, and the *age* effect is nearly consistent for *gender* (males and females) and for *race* (blacks and whites).
- We also observe that $\lambda_0(t, \hat{\alpha}) = \exp\{\hat{\alpha}_0 + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 t^3 + \hat{\alpha}_4 (t-\gamma)_+^3\}$ can be depicted as Figure 1. From Figure 1, we know that: holding the covariate *age* constant, as time goes by, the risk of dying would like to increase faster due to the increase in $\lambda_0(t, \hat{\alpha})$ after kidney transplant.

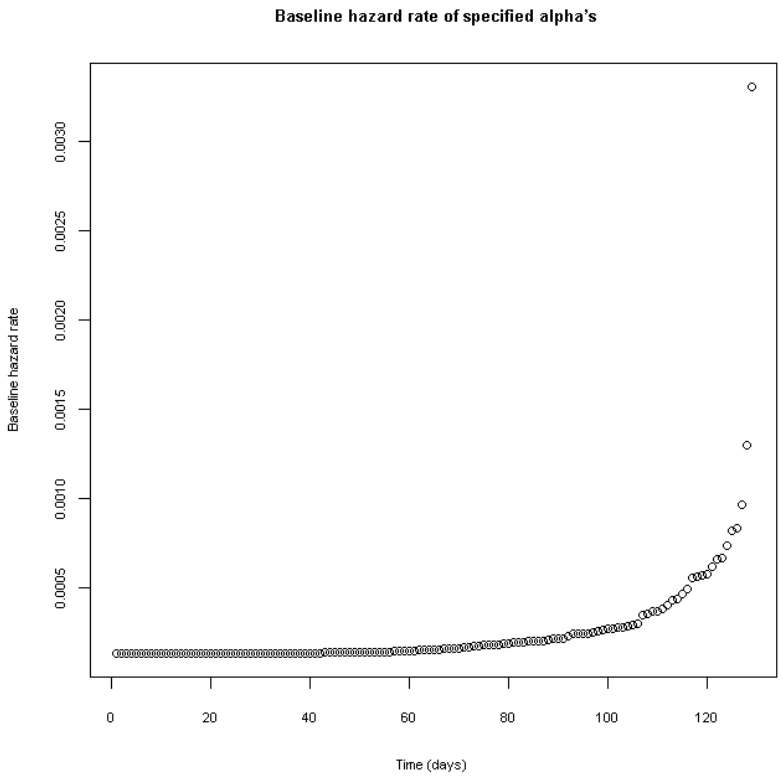

Figure 1 Trend of the baseline hazard rate generated from the final reduced PH model

In terms of Figure 2 as follows, the general K-M estimate roughly falls into the interval of the survival curves of *age=40* and *age=50* of the reduced PH model.

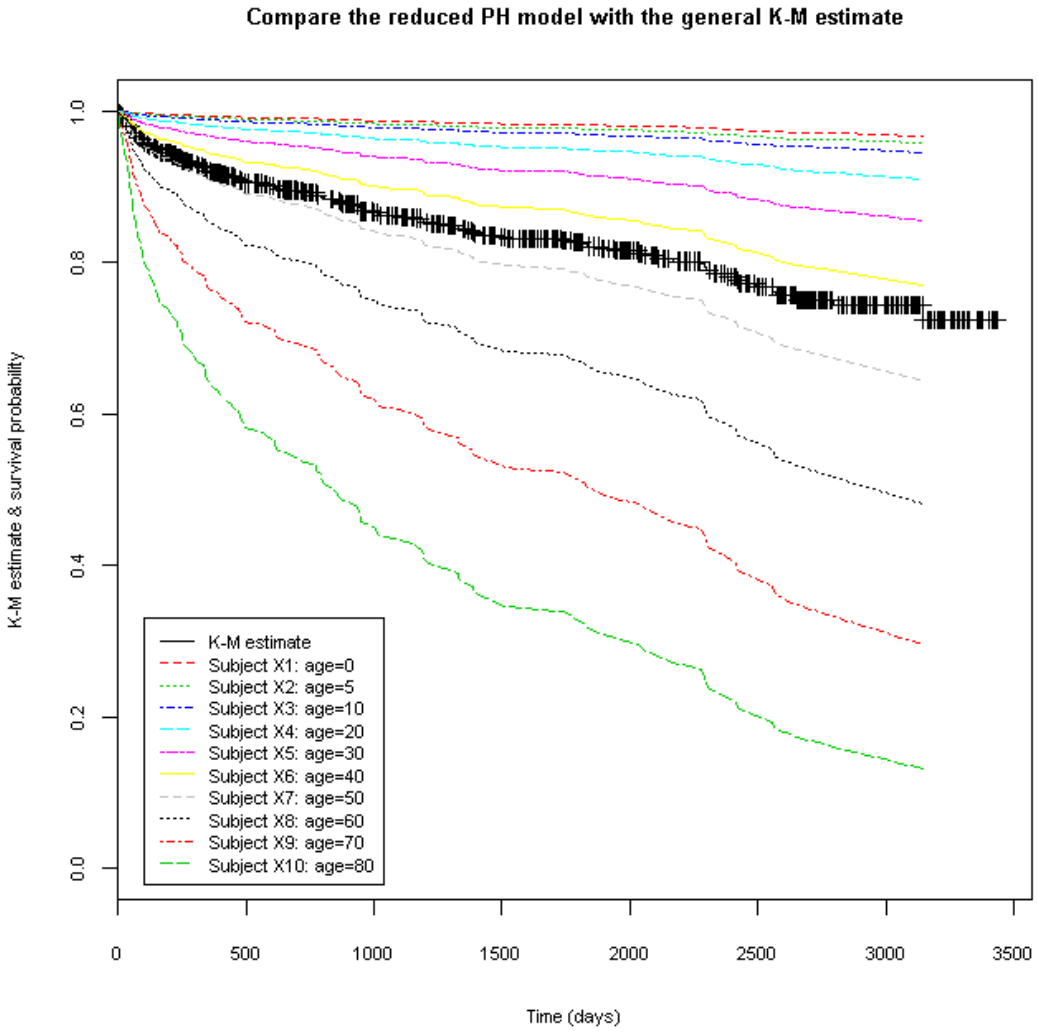

Figure 2 Graphical comparison between the general K-M estimate and the final fitted curves

We can conclude that:

- The increase in *age* is associated with decreased survival. This means that the young patients may have higher probabilities to live longer than the old patients.
- The survival probability for an individual who took kidney transplant decreases more quickly as age increases.

- Generally speaking, the survival time of the patients around 45 years old are more likely to be predicted by this reduce PH model.
- Generally speaking, a large proportion of the patients who need kidney transplant may be the ones 40 to 50 years old.

The median of *age* is *43* for the patients in this case. In terms of Figure 3 as follows, we see that the K-M estimate for *age < 43* **match**es the survival curves of *age=30* of the reduced PH model, and the K-M estimate for *age >= 43* falls into the interval of the survival curves of *age=50* and *age=60* of the reduced PH model.

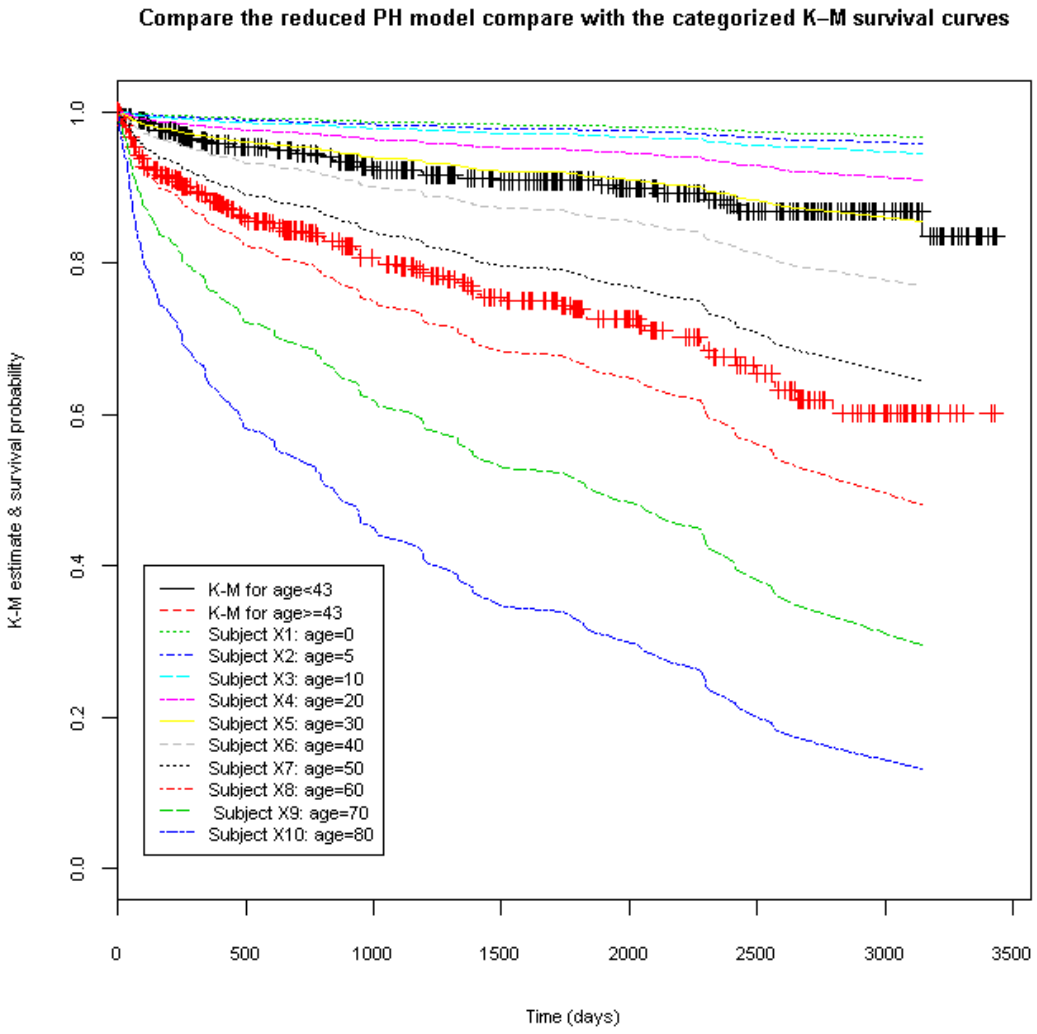

Figure 3 Graphical comparison between the categorized K–M curves and the final fitted curves

We can conclude that:

- For the patients less than 43 years old, the survival time of the patients around 30 years old are more likely to be predicted by this reduce PH model.
- For the patients more than 43 years old, the survival time of the patients around 55 years old are more likely to be predicted with this reduce PH model.
- Most of the patients less than 43 years old who need kidney transplant may be the ones 20 to 40 years old.
- Most of the patients more than 43 years old who need kidney transplant may be the ones 50 to 60 years old.

Figure 4 on the next page shows that we can use the distinguishably categorized K-M curves to verify the final reduced PH model. We classify the patients into eight groups by age as:

Group 1	age between 0 to 10
Group 2	age between 10 to 20
Group 3	age between 20 to 30
Group 4	age between 30 to 40
Group 5	age between 40 to 50
Group 6	age between 50 to 60
Group 7	age between 60 to 70
Group 8	age between 70 to 80

Table 2 Classify the patients into eight groups by age

Comparing the distinguishably categorized K-M curves in Figure 4 with the survival curves generated by the final reduced PH model in Figure 5, we see that: there are big differences between two figures in the patients with age 40 to 50, the patients with age 50 to 60, and the patients with age 70 to 80.

We can conclude that:

- The final reduced PH model can not well estimate the group with age 40 to 50 years old, and the group with age 50 to 60 years old, since their empirical K-M curves intersect with each other.
- We are not sure of the situation of the patients who are older than 70 years old because of insufficient information.

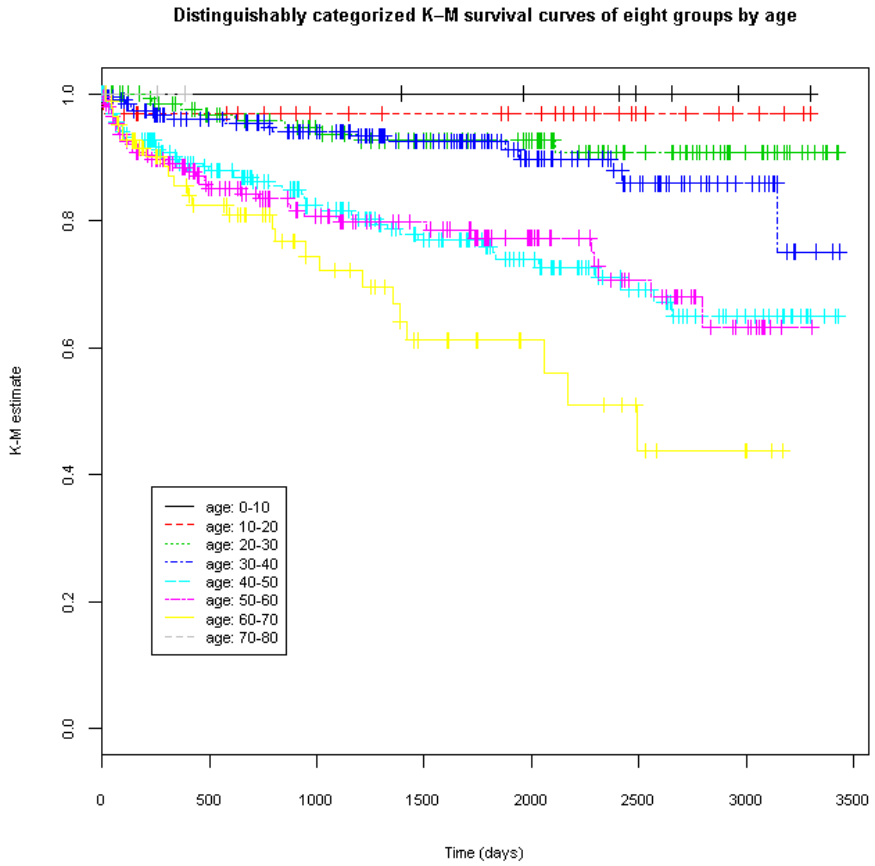

Figure 4 Distinguishably categorized K–M survival curves of eight groups by age

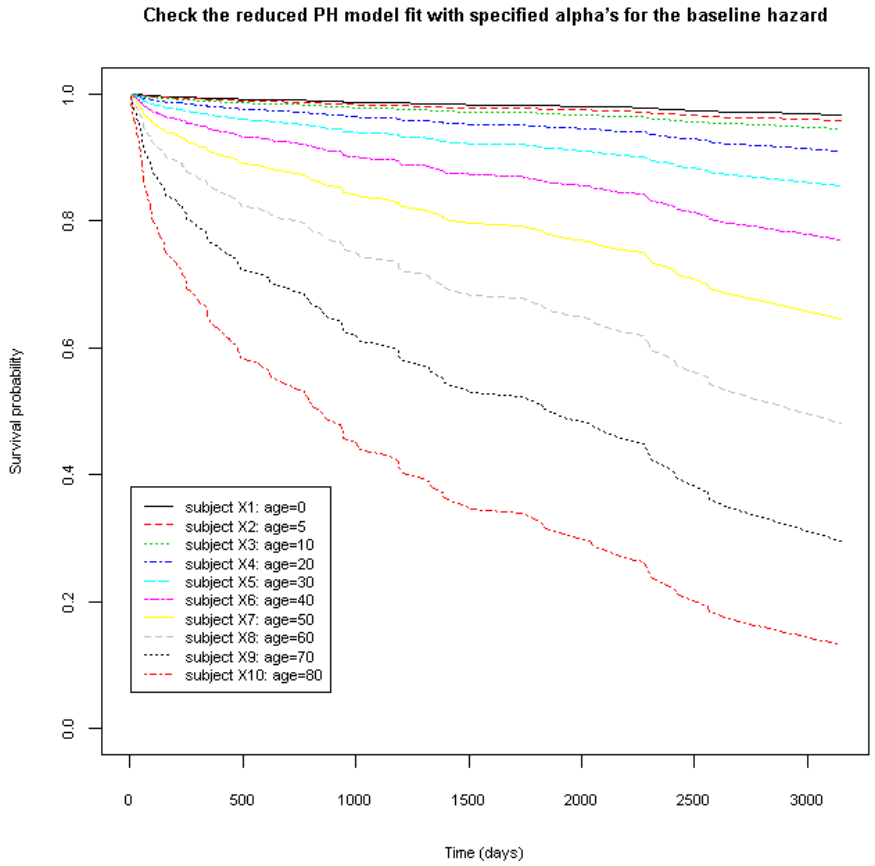

Figure 5 Check the reduced PH model fit with specified alpha's for the baseline hazard

Suggestion

The final selected reduced PH model may be not the best model to fit the data set in this case, since there are still relationships can not be explored. For example, the final reduced PH model can not well estimate the group with age 40 to 50 years old and the group with age 50 to 60 years old.

This also means that we can use residual analysis to evaluate whether certain observations were poorly characterized by the model.

Another way to improve this analysis is: maybe we can use time-varying covariates or a mixed PH model.