Summary Report of a Survival Analysis Using a Parametric Accelerated Failure Time Model

Purpose:

1) Run a complete data analysis using a parametric Accelerated Failure Time (AFT) model;
2) Select the appropriate log-location-scale family for the survival time T_i;
3) Assess the fit and perform inference.

Data description:

The data consists of 137 patients who participated in a clinical trial for two treatment regimens for lung cancer.

We consider the data set from a study designed to assess the effect of a new treatment on the survival time of patients with lung cancer. The *TIME* variable contains survival time in days of after a treatment. The variable *STATUS* has a value of 1 for those events at time, and has a value of 0 for those right censored.

The covariates included in the analyses are:

(i) *trt*: 1=standard treatment, 2=new treatment;

(ii) *ctype*: cell type, 1=squamous, 2=smallcell, 3=adeno, 4=large;

(iii) *dtime*: days from diagnosis to randomization;

(iv) *age*: in years at the time of a treatment;

(v) *prior*: prior therapy, 0=no, 1=yes.

Software for analyzing:

In this case, we cover the basics of modeling time-to-event data 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;
- Model selection using graphic methods and step Akaike's Information Criterion (AIC);
- Fit Accelerated Failure Time (AFT) models, find the appropriate reduced model, and obtain inferences for parameters of interest, including: regression coefficients, median time-to-events (TTE's) for any possible covariate combination of the reduced model, and correspondent survivor curves;
- Model diagnostics.

AFT model

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. An AFT model is then

$$\log T_i = b_0 + b_1 x_{i1} + ... + b_p x_{ip} + s e_i$$

where e_i is a random disturbance term, and $b_0, ..., b_p$ and s are parameters to be estimated. As for the natural log transformation of T_i, many popular survival distributions *T's* have the property $Y = logT$ where Y is from a location-scale family, and this transformation also ensures that predicted values of T are positive.

In other words, the AFT model can be specified as

$$\log T_i = X_i^T \underline{b} + s e_i$$

The AFT model treats the logarithm of survival time $logT_i$ as the response variable and includes an error term ε_i that is assumed to follow a particular distribution (see Table 1 in Technical Report).

AFT model analysis[1]

Usually, X_1 represents the type of treatment, in the absence of covariates β_1, a single variable X_1 is often defined as a 0|1 indicator variable distinguishing between control and experimental treatments. When covariate data are available, additional terms $\beta_2 x_{i2}, \ldots, \beta_p x_{ip}$ are included in the model, where the added variables represent factors such as gender, date of birth and age.

- Select model

  An initial step in fitting an AFT model is determining which distribution should be specified for the survival time T_i. Under the AFT model parameterization, the distribution chosen for T_i dictates the distribution of the error term ε_i. For instance, if survival times T_i's are modeled as a Weibull distribution, the error term is assumed to follow an extreme-value distribution (see Table 2 in Technical Report).

  For each comparison, preliminary models are fit in where the T_i's are modeled using the exponential, Weibull, Gamma, log-logistic and log-normal distributions. Graphic methods provide ways to check assumptions concerning the form of a lifetime distribution and its relationship to covariates, and the appropriate distribution can be selected as the one which minimizes the Akaike's Information Criterion (AIC).

  In almost every case, the Weibull distribution and the log-normal distribution are the most appropriate based upon the AIC criterion.

- Select variables

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

[1] Swindell, W. 2009. Accelerated Failure Time Models Provide a Useful Statistical Framework for Aging Research, Experimental Gerontology 44 (2009): 190–200.

Variable selection 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. Therefore, residual analysis or influence analysis can be applied to verify the variable selection.

- Perform diagnostic analyses to evaluate the adequacy of model fit

  Model fit is evaluated based upon a graphical comparison between empirical Kaplan–Meier survival curves and fitted or "predicted" survival curves generated from the final AFT model.

  Residual analysis is used to evaluate whether certain observations are poorly characterized by the model, and case deletion influence measures are analyzed to determine whether some observations exerted strong influence on parameter estimates. (I skip this part.)

Conclusion

We obtain the estimated parameters of the reduced log normal AFT model in the following table.

Table 1 K-M survival curves by treatment

	Parameter	Value	Std.Error	z	p
Reduced model	b_0 Intercept	4.674	0.123	37.90	0.00e+00
	b_4 prior	-0.496	0.160	-3.10	1.96e-03
	b_5 smallcellTRUE	-0.739	0.163	-4.55	5.49e-06
	b_6 adenoTRUE	-0.842	0.195	-4.32	1.55e-05
	s^* Log(scale)	-0.190	0.063	-3.01	2.60e-03

This means that: we have the reduced model as

$$\log T_i = b_0 + b_4 x_{i4} + b_5 x_{i5} + b_6 x_{i6} + se_i .$$

Here the value of *scale* refers to the estimate of s and $s^* = log(\sigma)$.

We can conclude that:

- According to the analysis, the variables *trt*, *dtime*, *age*, and *ctype=large* play little effect on these patients with lung cancer. It means that: (i) new *treatment* can not improve the survival time of these patients, and this point of view can also be known from Figure 1 that shows there is almost no difference after new treatment applied; (ii) *age* seems not the factor to effect the survival time; (iii) *Large* cell type may be insignificant in this trial.
- We obtain the significant variables *prior*, *ctype=squamous*, *ctype=smallcell*, *ctype=adeno*. But according to their negative coefficients, it means that: (i) having *prior* therapy may reduce the survival time; (ii) *ctype=squamous*, *ctype=smallcell*, *ctype=adeno* may have negative effect to the patients; (iii) the variable *prior* may weigh less than others, which means: compared with three selected cell types, it plays less effect on the survival time.

Figure 1 K-M survival curves by treatment

Table 2 shows median TTE values for subject X_1 to X_6.

Table 2 Median TTE values for subject X_1 to X_6

	Squamous	*Small Cell*	*Adeno*
Prior = 0	107.10262	51.14441	46.13765
Prior = 1	65.21646	31.14263	28.09394

We can conclude that: (i) the patients not taking *prior* therapy may live longer than others; (ii) the patients with *squamous* cell type may live longer than those patients with other two cell types; (iii) the patients with *adeno* cell type may have shorter survival time than those patients with other two cell types; (iv) there are some overlaps of 95% confidence intervals between *smallcell* and *adeno,* and this also means: *squamous* cell type tends to perform more significant in this trial.

According to Figure 2, the K-M curve of *squamous* roughly falls into the interval between X_1 and X_2, and X_3 and X_5 respectively fit the K-M curves of *smallcell* and *adeno*.

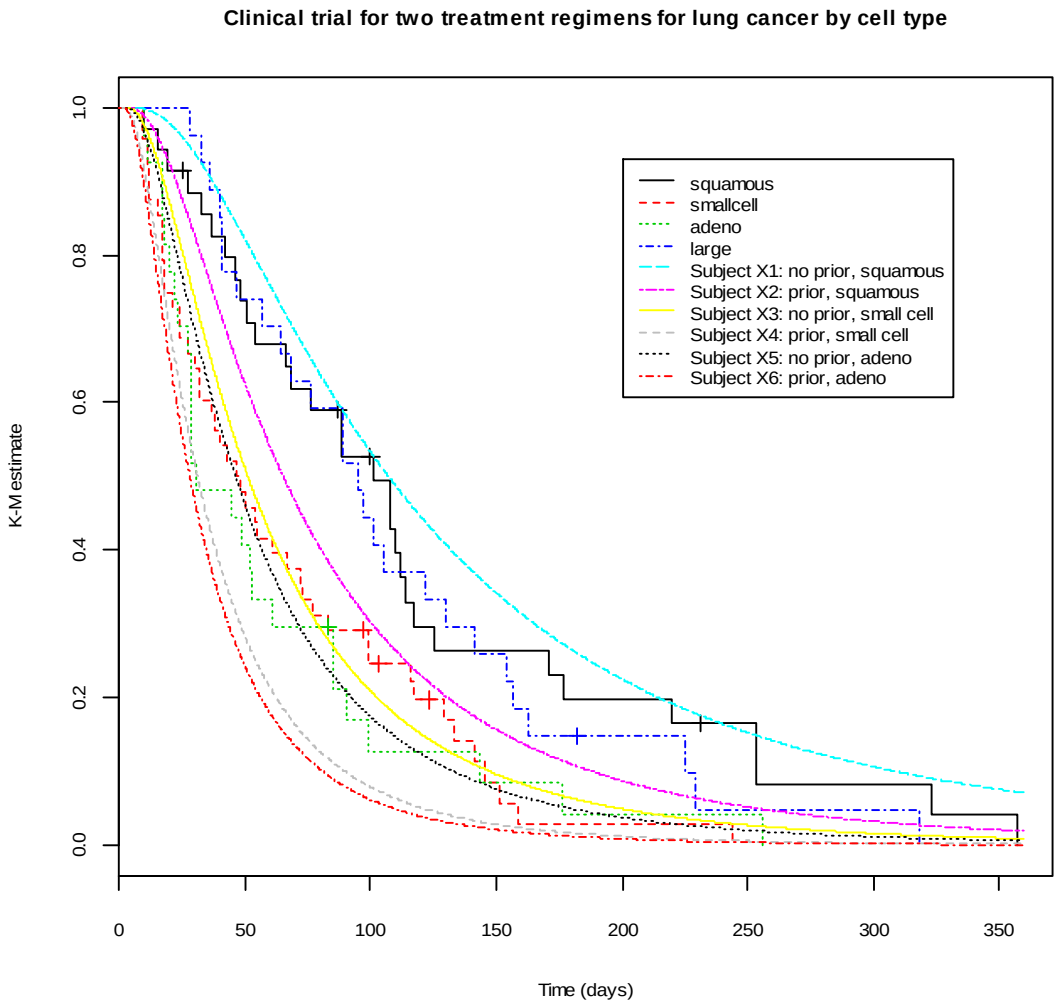

Figure 2 Graphical comparison between K-M curves and final fitted curves by cell type

We can conclude that: (i) the survival time of the patients with *squamous* are likely easier to be predicted, and having *prior* therapy may cause negative effect; (ii) the survival time of the patients with *smallcell* and *prior* therapy and the patients with *adeno* and *no prior* therapy may be predicted by this model; (iii) as discussed in the part of median TTE, the patients with *squamous* may have higher probabilities to live longer than those patients with other two cell types.

According to Figure 3, only X_2 seems to fit the K-M curve of *prior*.

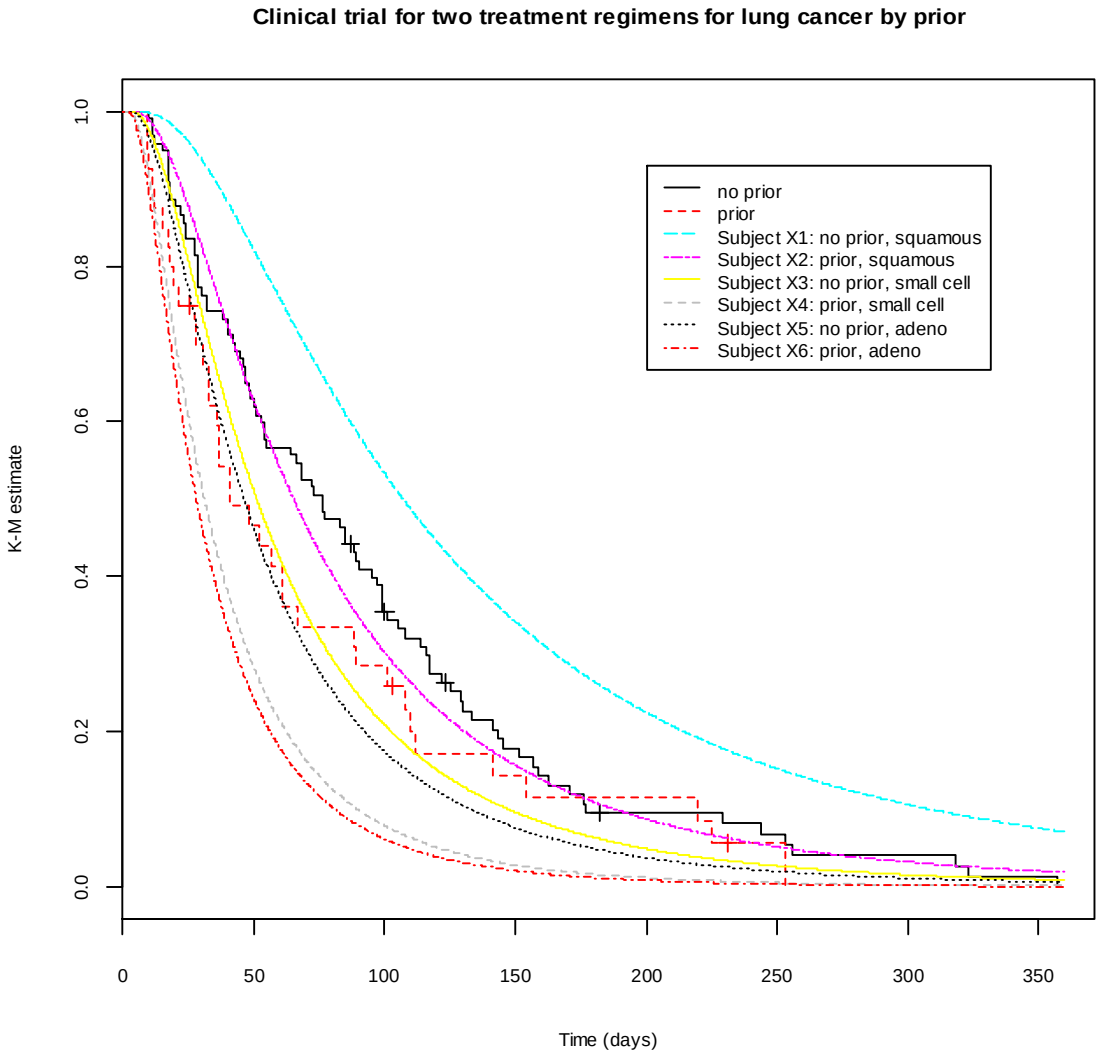

Figure 3 Graphical comparison between K-M curves and final fitted curves by *prior*

We can conclude that: (i) having *prior* therapy or not is not so important as the variables of cell type; (ii) the survival time of the patients with *squamous* are likely to have some unidentified relationship with the variable - having *prior* therapy or not.

Suggestion

The final selected reduced log normal AFT model may not be the best model to fit the data set in this case, since there are still relationships can not be explored.

As we can see from the Figure 5, 6 and 7 in Technical Report, the log normal probability plots do not fit well with the time at two ends. This means that: we can use residual analysis to evaluate whether certain observations are poorly characterized by the AFT model, and employ case deletion influence measures to determine whether some observations exert strong influence on parameter estimates.

Another way to improve this analysis is: maybe we can use a mixed AFT model.