ISyE 6404 CP.1: Proportional Hazards Regression

Yuan Gao, Kevin Lee, Akshay Govindaraj,
Yijun (Emma) Wan, Peter Williams, Ruixuan Zhang

Date: 2018-10-22

Contents

Workload Distribution

Below is a description of tasks completed by each team member for this project:

Team Member	Task Description
Yuan Gao	Data Description, PH-regression, Reference and Literature Review
Kevin Lee	Summary of References, Description of AFT model and R Implementation
Akshay Govindaraj	Bootstrap for Computing Confidence Interval
Yijun (Emma) Wan	Comparison between PH-regression and AFT
Peter Williams	Code Compilation, R Debugging, Latex Formatting, Visualization
Ruixuan Zhang	Co-work in Description of AFT procedure and Comparison to PH

Proportional Hazards Regression Tasks

1. Data Description & PH-regression

Task: Locate a data set in the field of your interest, e.g., eCommerce, medical study, drug development, supply-chain/logistics operations, to practice the PH-regression technique. Note that we need to predict both hazard-rate and the survival function at an input x0.

For this exercise, we relied on a dataset in the 'survival' package in R, that describes the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored. Each patient has exactly 2 observations, and the dataset consists of the following variables:

1) patient: id assigned to patient
2) time: recurrence time to infection
3) status: event status
4) age: in years
5) sex: 1=male, 2=female
6) disease: disease type (0=GN, 1=AN, 2=PKD, 3=Other)
7) frail: frailty estimate from original paper

A preview of the data is shown here:

Table 1: Preview: Kidney Dataset

id	time	status	age	sex	disease	frail
1	8	1	28	1	Other	2.3
1	16	1	28	1	Other	2.3
2	23	1	48	2	GN	1.9
2	13	0	48	2	GN	1.9
3	22	1	32	1	Other	1.2
3	28	1	32	1	Other	1.2
4	447	1	31	2	Other	0.5
4	318	1	32	2	Other	0.5

First, we compute univariate Cox PH effect estimates for four variables outlined below, and then we fit multi-variate Cox PH effects using two variables to describe how the factors jointly impact survival.

The table output below shows the regression beta coefficients, the effect sizes (given as hazard ratios) and statistical significance for each of the variables in relation to overall survival. Note that a separate model was fit including each of the following covariates, individually (effects not estimate jointly):

Table 2: Cox PH Regression Results

Covariate	Effect Estimate (beta)	Hazard Ratio (rho)	p-value
age	0.005	0.088	0.609
sex	-0.838	0.435	0.005
diseaseGN	0.351	-0.052	1.463
diseaseAN	0.380	0.091	0.993
diseasePKD	-0.260	-0.008	0.608
frail	1.008	0.083	0.000

Here are some of the key findings from the analysis here:

1) Statistical significance, reported as a p-value from a Wald statistic, evaluates whether the beta coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variables sex and frail have statistically significant coefficients, but age and disease do nott. From here on, we focus on the variables, sex (gender) and frailty (frail).

2) The regression coefficients are positive signs meaning that the hazard (risk of death) is higher, and thus the estimated prognosis is worse for subjects with higher values of that variable.

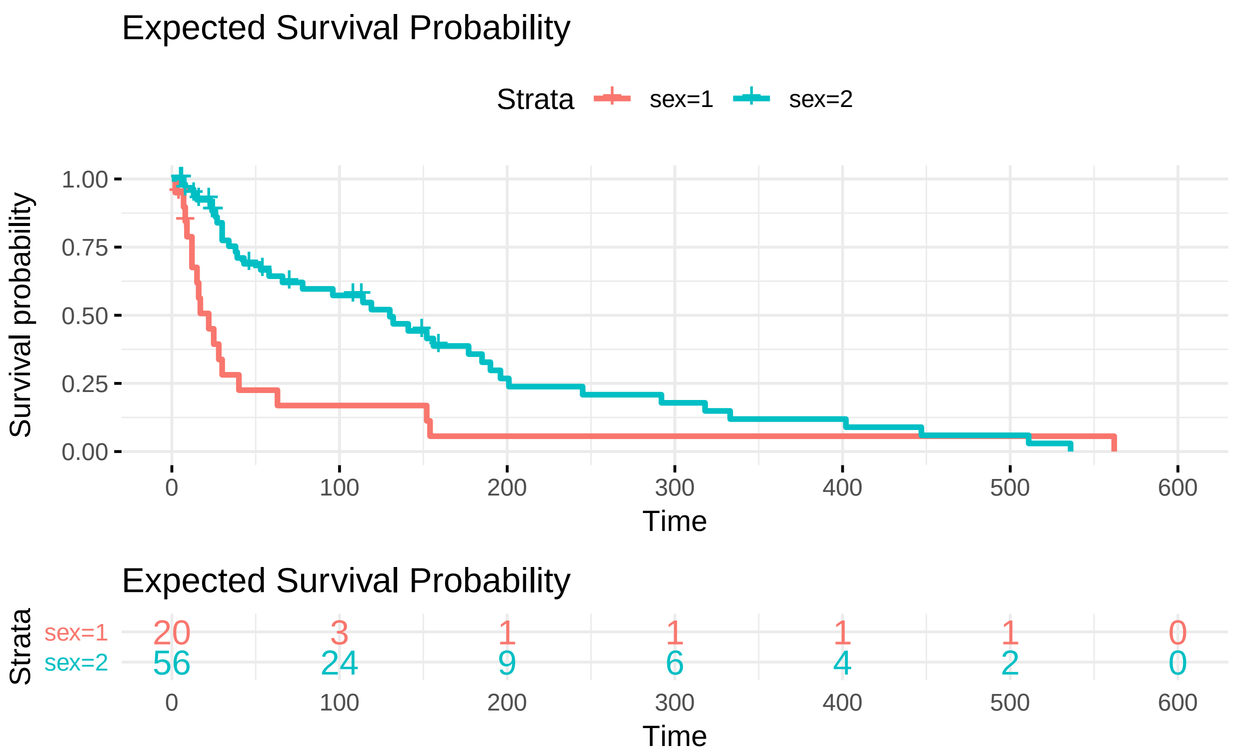
3) Hazard ratios are the exponentiated coefficients exp(coefficient) which gives the effect size of the variable age. So, it gives us the predict about the hazard ratio for any given variables. The variable sex is encoded as a numeric vector. 1: male, 2: female. The beta coefficient for sex $= -0.838$ indicates that females have lower risk of kidney infection (lower survival rates) than males, based on model estimates.

4) Confidence intervals of the hazard ratios is shown by the upper and lower 95% confidence intervals for the hazard ratio exp(coefficient), which displayed below.

5) Global statistical significance p-values are also reportable, for three alternative tests: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent.

6) The R package 'survival' provides a function, 'cox.zph()' which shows the data are sufficiently consistent with the assumption of proportional hazards with respect to each of the variables separately as well as globally.

To demonstrate some of these functionalities, we visualize the expected survival proportion at any given point for a particular risk group. In this case, we visualize the survival function, layered with a model that estimates the impact of the variable 'sex', group (strata), here:



Further, utilizing survival model functions in R, and given new data from a patient who is Male (Sex = 1), with a fraility measure of (Frail = 2), we can generate expected survival probabilities with upper and lower confidence intervals. A snapshot is displayed below, using the 'coxph' function in the 'survival' package in R:

Table 3: Prediction Table: Sex = 1, Frail = 2

n	time	survival_prob	lower	upper
76	2	0.957	0.877	1.000
76	4	0.957	0.877	1.000
76	5	0.957	0.877	1.000
76	6	0.957	0.877	1.000
76	7	0.867	0.736	1.000
76	8	0.774	0.612	0.978
76	9	0.726	0.554	0.952
76	12	0.618	0.425	0.897

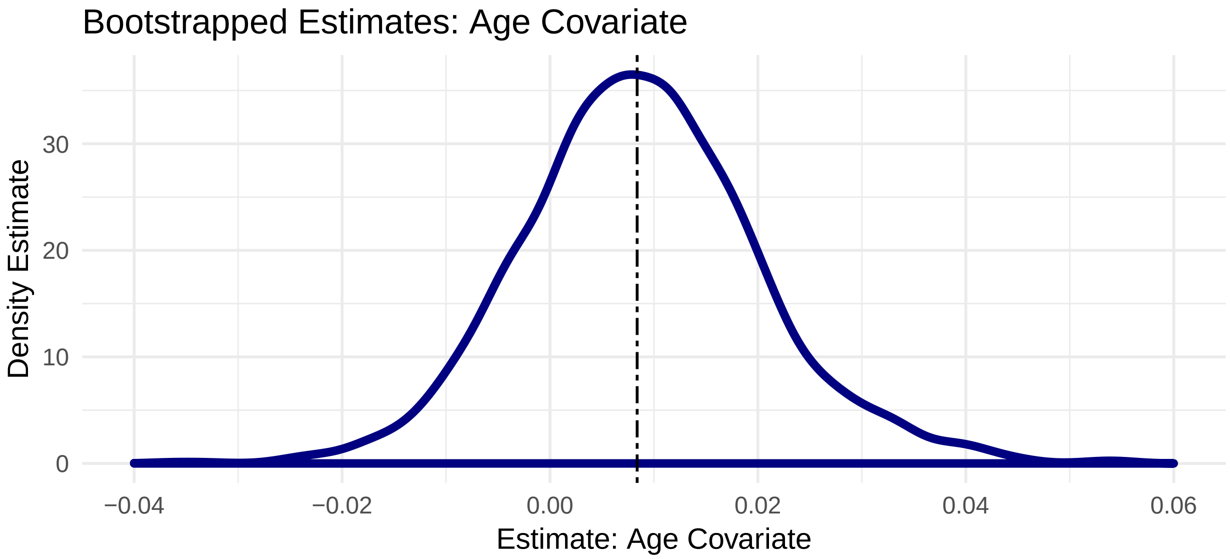
2. Confidence Interval - Bootstrap Method

Task: Follow #4 in EP-2 to construct a 90% confidence interval based on the bootstrap method.

To get a sense of the stability of the effects estimated above, we rely on the bootstrap technique, and more specifically the percentile method utilizing the kidney infection dataset described above. To generate bootstrap estimates, we rely on the 'boot' library in R.

To illustrate the functionality of the bootstrap library in R, we visualize the spread, using a density plot in R, of the effect estimate for the 'age' variable, in context of the survival function, below:

```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```



The visualization above, highlights a symmetry of effect estimates for the variable age. This symmetry allows for utilization of the 'percentile' method in calculating effect estimate intervals, that is utilized below. Also note that where range of bootstrapped estimates does overlap with 0, it can be indicative of a lack of statistical significance. The results for the rest of the variables displayed below:

Table 4: Bootstrap Effect Estimates (Percentile Range)

Covariate	Percentile: Lower 5th	Percentile: Upper 95th
sex	-3.296	-1.567
age	-0.009	0.028
DiseaseGN	-0.530	0.891
DiseaseAN	0.038	1.439
DiseasePKD	-3.953	-0.821
frail	1.453	2.672

Note: Estimates Based on R=2000 Bootstrap Replications

Based on the results from the bootstrapping exercise, the hazard variables sex and frail have more significant effect estimates, since the middle 90% of effect estimates did not overlap 0, as shown above in the percentile range table above. Other findings and comments from this analysis are highlighted below:

1) With PH regression models, the best specified hazard variables are generally the ones which do not change with time. Since the variable 'age' is in this dataset, and it certainly changes with time it should be used with caution.

2) For the variable 'sex', or gender, we get the boostrap estimate of the value of the hazard

coefficient consistently between (-3.253 and -1.622). Since the numeric representation for men is 1, we can conclude that men are at a higher risk than women for kidney infection.

3) The variable, 'frail', or a measure of frailty, is the most significant hazard variable, given that meaning of the variable, it makes sense that the coefficient is positive, meaning more frail patients are at a higher risk of infection.

4) It is unclear as to which disease is more hazardous. The values seems to suggest that DiseaseAN is the most dangerous but it is hard to conclude when the lower bound is very close to zero (0.0379).

3. References & Literature Review

Task: Skim through the CP-1 references given in Files> Projects> CP-1> Reference> directory to write a two-page report for summarizing the work there.

Ref-1.0 and Ref-1.1 introduces the semiparametric regression, which means the combination of parametric(may be mis specified and inconsistent) and nonparametric models. So, there are many assumptions should be satisfied or verified. Although some assumptions are difficult to verify, they are generally satisfied for well-behaved estimators. But, they make Andrews' MINPIN, the semiparametric estimator has the same asymptotic distribution as the idealized estimator obtained by replacing the nonparametric estimate with the true function, not easy to implement. Semiparametric regression is typically used in cases where a nonparametric model may not perform well enough or where a parametric model error distribution is unknown. Semiparametric regression owns three popular methods: 1) partially linear, 2) index and 3) varying coefficient models.

Reference 2.0 and 2.1 focus on introducing the process and characters of PH model. Firstly it describes the definition of survivor function S(t) , hazard function h(t) and cumulative hazard function H(t) which bring interpretability, analytic simplifications and modeling simplifications. Secondly, it introduces the situation with censored data which is classified into three categories: 1) clearly informative: Type I or II Censoring. 2) noninformative and 3) Less clear situation. Thirdly, after the introduction, they move to the process of PH model, which shows X are time independent. And PH model is popular due to robustness, non-negative hazards, easily compute the hazard ratio and can estimate h(t,X) and S(t,X). Fourthly, reference talks two types of likelihood in ML Estimation of the Cox PH Model. 1) full likelihood 2) partial likelihood which considers probabilities for subject who fail and does not consider probabilities for censored subjects explicitly. Full likelihood allows us to estimate the baseline hazard function, but the partial likelihood allows us to make estimates of parameter B while accounting for censored data. Once partial likelihood is maximized with respect to the regression parameter, it can then be used to eventually estimate the baseline hazard function. A similar partial likelihood process (penalized partial likelihood) can be utilized in addition to iterative sure independence screening to allow high-dimensional variable selection, which apparently has very small false selection while maintaining small mean squared error. This type of selection through penalization is an extension of classical selection techniques such a stepwise and bootstrap procedures to be used for PH regression, which means simulations can be used to demonstrate the viability of said selection methods Fifthly, reference also introduces the estimation of survival curves called adjusted survival curves. The last but not the least, as the content in reference 1, there are several assumptions in semiparametric regression, so does PH model.

Reference 2.2 gives us a real world case in population-based cancer survival analysis using PH model. Patient survival rates provide such a measure to effectively diagnose and treat the cancers that arise and require a means of measuring progress in this specific area. There are several measures including cause-specific survival, which can estimate net survival, and relative survival which is calculated by observed survival proportion divided by expected survival proportion. And, it uses flexible parametric models with restricted cubic splines to fit relative survival curve other than cox model that cannot be applied to model a difference in two rates.

Reference 2.3 talks about variable selection for Cox's proportional hazards model in high dimensional space by extending the sure screening procedure to an iterative version. It extends the key idea of SIS and ISIS to handle Cox's proportional hazards model: 1) ranking by marginal utility 2) Conditional feature ranking and iterative feature selection3) new variants of SIS and ISIS for reducing FSR. Finally, numerical simulation studies have shown encouraging performance of the proposed method in comparison with other techniques such as LASSO.

Reference 3.1 introduces the Bayesian variable selection for proportional hazards regression models with right censored data where a nonparametric prior is specified for the baseline hazard rate with the use of discrete gamma process and a semi-automatic parametric informative prior specification that focuses on the observables for the regression coefficients and the model space. In addition, it proposes a Markov chain Monte Carlo method to compute the posterior model probabilities.

Reference 4.1 proposes a piecewise exponential representation of the original survival data to link hazard regression with estimation schemes, which is based on the Poisson likelihood. And it makes recent advances for model building in exponential family regression accessible and in the nonproportional hazard regression. The reason why coming the above new method is that recent statistical methods typically introduce additional difficulties, such as immediate appeal in terms of flexibility, when a subset of covariates and the corresponding modeling alternatives have to be chosen. The article implement a two-stage stepwise selection approach, an approach based on doubly penalized likelihood, and a component wise functional gradient descent approach will be adapted to the piecewise exponential regression problem.

Reference 5.1 shows using machine learning to do survival analysis. Because there are always some censored data which can be effectively handled using survival analysis techniques. Although above references developed traditional statistical approaches to overcome this censoring issue. In addition, many machine learning algorithms are adapted to effectively handle survival data and tackle other challenging problems that arise in real world data. It provides a comprehensive and structured review of the representative statistical methods along with the machine learning techniques used in survival analysis and provide a detailed taxonomy of the existing methods. One can perform survival trees, neural networks, bayesian methods (as discussed prior), support vector machines, boosting, and other such advanced machine learning techniques.

4. Implementation: Reference Procedure

Task: Outline steps for implementing one of the studied procedure addressed in the reference. You DO NOT need to implement them, but describe how to do it.

A method of survival analysis provided in the references was the accelerated failure time (AFT) model. This is a parametric alternative to the Cox Proportional Hazards model, which is originally semi-parametric. Several steps below show how to implement the regression based on AFT model.

Step-1: As the AFT model is parametric, certain assumptions must first be met. The big difference (assumption) for an AFT model is that it assumes an underlying distribution for its survival times, while PH regression does not. The general equation for an accelerated failure time model is

$$log(T) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon\sigma$$

where T is the survival time, the β values are coefficients of each predictor's effect on the log(T) (since survival times are always positive). The sigma term in the model equation is a scale parameter, which depends upon the shape of the underlying distribution, and the error term is assumed to be independent and identically distributed (as well as independent of the X-values). Notice that in an AFT model, the survival time will "contract" or "stretch" as a function of the model's predictor variables (slightly different interpretation compared to Cox PH). This stems from the idea that in AFT models the predictor variables have a multiplicative event on the log of survival time. As it acts as a parametric alternative to the Cox PH regression model, the AFT model also commonly employs the Weibull and Exponential distributions, though the typical distribution used is the log-logistic distribution.

Step-2: As we can see from the simplified AFT model equation, AFT models (parametric as they are) are analogous to generalized linear regression models in the sense that the results of the model can be interpreted similarly to how a typical regression can be. In interpreting the coefficients, one must note that the equation is based on a logarithm of time, so once we get the values of the B coefficients, we must rearrange the equation.

$$\hat{\gamma} = exp(\hat{\alpha})$$

Where $\hat{\alpha}$ can be estimated from historical data. This allows us to calculate our acceleration factor. For example, if we have a positive acceleration factor, for example 2, then we would be able to say that the variable increases our survival time by a factor of 2. If we have a negative value of 2 as our acceleration factor, the survival time would be "accelerated" or shortened by a factor of 2. Keep in mind that the predictor variables' effects are indeed multiplicative. If use the MLE method to estimate the parameters, based on the preset distribution and acceleration factor, it is feasible to compute the full likelihood:

$$L(\beta, \sigma) = \prod_{i=1}^{n} [\frac{1}{\sigma} f_0(\frac{logt_i - x^{\mathrm{T}}\beta}{\sigma})]^{\delta_i} S_0(\frac{logt_i - x^{\mathrm{T}}\beta}{\sigma})^{1-\delta_i}$$

where δ_i is the indicator to show that whether data is censored or not, if an observation is right-censored, the value of δ_i should be 0, otherwise it should be 1. Specifically, the likelihood is full rather than partial (different from Cox PH regression). Let $\epsilon_i = \frac{logt_i - x^{\mathrm{T}}\beta}{\sigma}$,thus, the log-likelihood become:

$$l(\beta, \sigma) = \sum_{i=1}^{n} [-\delta_i log\sigma + \delta_i log f_0(\epsilon_i) + (1 - \delta_i) log S_0(\epsilon_i)]$$

where $S_0(\epsilon_i)$ is called "baseline survival function" and $f_0(\epsilon_i)$ is the corresponding probability density function. Usually, they are the survival function and PDF of log-logistic. What's more $S_1(t) = S_0(\hat{\gamma}t) = S_0(e^{\hat{\alpha}t})$ By differentiating the log-likelihood function, we can maximize the log-likelihood and estimate the coefficients β and the scale parameter $log\sigma$.

Step-3: Once the regression is performed, one can even compare the effectiveness of models with differing underlying distributions through AIC comparisons. One must simply compare the AIC values of each different model to see which has the lower AIC value, though this is not the only method of comparison. As we would have a model with multiple predictor variables, one can also perform variable selection to try to cut variables that are not as explanatory in the context of survival time or that are simply insignificant based on given p-value (or simply removed through stepwise AIC).

Implementation in R: In R, it is simple to create an AFT model for survival analysis. Similarly to how we plot and interpret data as above in PH regression, we use the surv() function to establish that our model will be used to interpret survival data. The difference with AFT models lies in the usage of the survreg() function, which in R allows us to establish parameters (most notably distribution shape) for our model. This allows us to specify whether we want to use a Weibull, exponential, gamma, or log-logistic distribution. There are certain advantages to each type of distribution depending on circumstance, but log-logistic has the advantage of allowing non-monotonic hazard functions, which the Weibull distribution cannot replicate.

Note: As with most comparisons between parametric and non(or semi-)parametric methods, parametric models have the advantage in terms of ease of interpretability (the goal of a parametric model is to find a certain parameter that helps build a model representative of the data. This also leads to a disadvantage: parametric models require a foundational distribution of the data with which to base the model off of, which may not always be feasible. A violation of assumptions in a parametric model would then lead us to potentially sub-optimal results.

5. Result Review

Task: Discuss whether the results getting in (1) and (4) might be different (in what way).

First of all, accelerated failure time (AFT) model assumes an underlying distribution for its survival times, however, pH regression does not require any assumptions. As a parametric model, AFT needs to set up a baseline survival function, which means that the hazard function comes from a certain distribution.

Secondly, the relationship between either the survival time or the logarithm of the survival time and the features is linear. We can also get the full likelihood instead of partial likelihood by using MLE. Full likelihood or full model indicates that AFT can do the prediction.

Last but not least, AFT assumes that all the predictors of observations will either accelerate or decelerate the survival time. All of the impact on survival time by predictors can be seen and analyzed in AFT model.

In conclusion, parametric methodologies support a stronger evidence than nonparametric ones. However, nonparametric methodologies can be used in a more general way, because they do not require any assumptions in advance. If a given dataset does not meet the required assumptions, parametric model may not be the optimal choice.

Code Appendix

PH Regression Model Fitting Code

The following code demonstrates our approach to fit each potential co-variate independently to estimate its impact in context of the survival function, utilized in task 1:

```
#extract name of var, estimate, and p-value
#helper to extract model summaries in tabular format
get_cox_ph_results <- function(coxph_res){
  zph <- cox.zph(coxph_res)
  res_table <- data.frame(covariate = names(coxph_res$coefficients),
            beta = as.numeric(coxph_res$coefficients),
            pvalue =
             as.numeric(summary(coxph_res)$coefficients[colnames(summary(coxph_res)$coefficients
            rho = as.numeric(zph$table[,1])[!is.na(zph$table[,1])] )
  return(res_table)
}

variables <- c('age','sex','disease','frail')

results <- do.call('rbind',
          lapply(variables, function(x){
                  model_formula <- as.formula(paste0("Surv(time, status) ~ ",x))
          model_res <- coxph(formula = model_formula, data = kidney)
          return(get_cox_ph_results(model_res))
          }))
```

Bootstrap Procedure

The following outlines our bootstrap procedure for section 1.3, utilizing functionality from the *boot* package in R:

```
# must pass indices argument so that bootstrap can randomly choose - in this case row index
# in this case only using age as a covariate below
suppressPackageStartupMessages( library(boot))
#bootstrap function - age coefficient only here
bs_fun <- function(data, indices){
    bs_dat <- data[indices,]
    res.cox <- coxph(Surv(time, status) ~ age , data = bs_dat)
    return(as.numeric(res.cox$coefficients["age"]))
}

bs_res <- boot(kidney, bs_fun, R=2000) # R = number of replications

#plot(bs_res) #if symmetric - ok to use percentile approach
# type options include: "norm", "basic", "perc", "stud"

bs_ci <- boot.ci(bs_res, conf = 0.95, var.t0 = NULL, type = 'perc')
```

Questions?

Contact: ygao390, kylee20, ywan40, agovindaraj6, pwilliams60, rzhang438 | @gatech.edu