CP-1 Part 1: ISyE 6404

Yuan Gao, Kevin Lee, Akshay Govindaraj
Yijun (Emma) Wan, Peter Williams, Ruixuan Zhang
ygao390, kylee20, ywan40, agovindaraj6, pwilliams60, rzhang438 | @gatech.edu

2018-10-20

Contents

Workload Distribution

Below is a description of tasks and the distribution of work (%) by team member for this project:

Team Member	Task Description
Yuan Gao	TBD
Kevin Lee	TBD
Akshay Govindaraj	TBD
Yijun (Emma) Wan	TBD
Peter Williams	TBD
Ruixuan Zhang	TBD

PH-Regression Data-Fit

Dataset intro

Tasks

1) 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.

```
suppressPackageStartupMessages( library(survival))
suppressPackageStartupMessages( library(survminer))
suppressPackageStartupMessages( library(boot))
data("kidney")
knitr::kable(head(kidney))
```

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

For this exercise, I located a dataset in the survival package, 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. It has seven variables:

1)patient: id

2)time: time

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

First, I compute the univariate Cox analyses for the four variables; then we'll fit multivariate cox analyses using two variables to describe how the factors jointly impact on survival.

```
res.cox <- coxph(Surv(time, status) ~ age, data = kidney)
res.cox
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age, data = kidney)
##
##        coef exp(coef) se(coef)    z    p
## age 0.00458   1.00459  0.00896 0.51 0.61
##
## Likelihood ratio test=0.26  on 1 df, p=0.6
## n= 76, number of events= 58
```

```
res.cox.zph <- cox.zph(res.cox)
res.cox.zph
```

```
##        rho chisq     p
## age 0.0876 0.472 0.492
```

```
res.cox <- coxph(Surv(time, status) ~ sex, data = kidney)
res.cox
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = kidney)
##
##       coef exp(coef) se(coef)     z      p
## sex -0.838     0.433    0.297 -2.82 0.0047
##
## Likelihood ratio test=7.07  on 1 df, p=0.008
## n= 76, number of events= 58
```

```
res.cox.zph <- cox.zph(res.cox)
res.cox.zph
```

```
##       rho chisq       p
## sex 0.435  10.8 0.00102
```

```
res.cox <- coxph(Surv(time, status) ~ disease, data = kidney)
res.cox
```

```
## Call:
## coxph(formula = Surv(time, status) ~ disease, data = kidney)
##
##              coef exp(coef) se(coef)     z    p
## diseaseGN   0.351     1.421    0.354  0.99 0.32
## diseaseAN   0.380     1.463    0.336  1.13 0.26
## diseasePKD -0.260     0.771    0.507 -0.51 0.61
##
## Likelihood ratio test=2.66  on 3 df, p=0.4
## n= 76, number of events= 58
```

```
res.cox.zph <- cox.zph(res.cox)
res.cox.zph
```

```
##                 rho   chisq     p
## diseaseGN  -0.05159 0.14339 0.705
## diseaseAN   0.09068 0.45068 0.502
## diseasePKD -0.00838 0.00425 0.948
## GLOBAL           NA 1.04211 0.791
```

```
res.cox <- coxph(Surv(time, status) ~ frail, data = kidney)
res.cox
```

```
## Call:
## coxph(formula = Surv(time, status) ~ frail, data = kidney)
##
##        coef exp(coef) se(coef)    z       p
## frail 1.008     2.741    0.179 5.63 1.8e-08
##
## Likelihood ratio test=26.81  on 1 df, p=2e-07
## n= 76, number of events= 58
```

```
res.cox.zph <- cox.zph(res.cox)
res.cox.zph
```

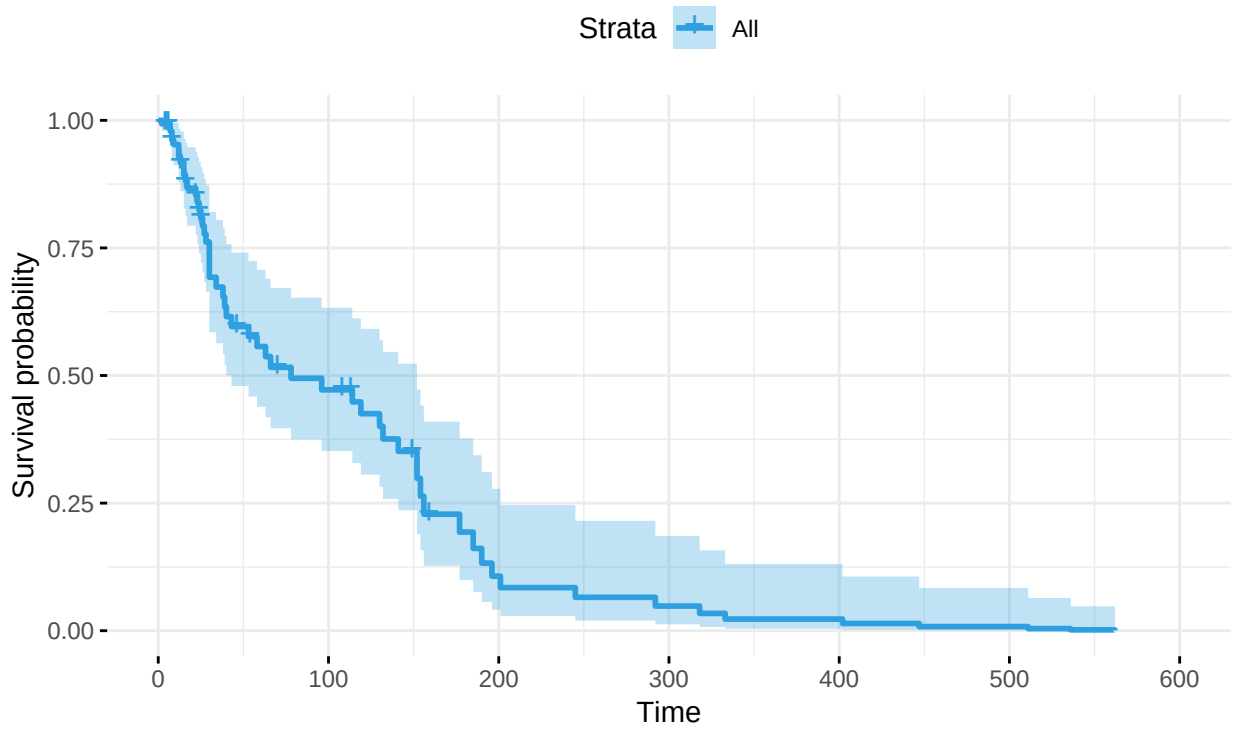
```
##          rho chisq     p
## frail 0.0828 0.247 0.619
```

The syntax above gives us the method to get the survival function of any given variables.The output above 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.

Following is the Cox regression results Analysis:

1) Statistical significance, z, gives the Wald statistic value,evaluates whether the beta coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variable sex and frail have highly statistically significant coefficients, but age and disease don't. So, following we focus on the sex and frail.
2) The regression coefficients are positive signs meaning that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable.
3) Hazard ratios are the exponentiated coefficients (exp(coef)) which give 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 (lower survival rates) than males, in these data.
4) Confidence intervals of the hazard ratios is shown by the upper and lower 95% confidence intervals for the hazard ratio (exp(coef)).
5) Global statistical significance p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent.
6) cox.zph() function shows the data are sufficiently consistent with the assumption of proportional hazards with respect to each of the variables separately as well as globally.

```
res.cox <- coxph(Surv(time, status) ~ sex+frail, data =  kidney)
ggsurvplot(survfit(res.cox), data = kidney, palette = "#2E9FDF",ggtheme = theme_minimal())
```




```
new_dat = data.frame(sex = c(1), frail = c(2))
survfit(res.cox, newdata = new_dat)
```

```
## Call: survfit(formula = res.cox, newdata = new_dat)
##
##       n  events  median 0.95LCL 0.95UCL
##      76      58      15      12      30
```

The picture is to visualize the predicted survival proportion at any given point for a particular risk group. Besides, the syntax above is to make predict about the survival interval of any given new data.

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

3) 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. Semi-parametric 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) 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. As the AFT model is parametric, certain assumptions must first be met.

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) Discuss whether the results getting in (2) and (4) might be different (in what way).