Final Project for STAT 7122/8122, Fall 2018.
Due December 13, 2018

Aanalyzing TRACE Data. Please email me your R code along with your submission.

Data Source: The TRACE study group. Jensen, G.V., Torp-Pedersen, C., Hildebrandt, P., Kober, L., F. E. Nielsen, Melchior, T., Joen, T. and P. K. Andersen (1997), Does in-hospital ventricular fibrillation affect prognosis after myocardial infarction? European Heart Journal 18, 919-924.

The TRACE study group (Jensen et al. (1997)) studied the prognostic importance of various risk factors on mortality for approximately 6600 patients after myocardial infarction. The TRACE data included in the timereg package is a random sample of 1877 of these patients. See the data using: data(TRACE); names(TRACE). See more about the data description using help(TRACE).

Format: This data frame contains the following columns:

ID a numeric vector. Patient code.

STATUS a numeric vector code. Survival status. 9: dead from myocardial infarction, 0: alive, 7: dead from other causes.

TIME a numeric vector. Survival time in years.

CHF a numeric vector code. Clinical heart pump failure, 1: present, 0: absent.

DIABETES a numeric vector code. Diabetes, 1: present, 0: absent.

VF a numeric vector code. Ventricular fibrillation, 1: present, 0: absent.

WMI a numeric vector. Measure of heart pumping effect based on ultrasound measurements where 2 is normal and 0 is worst.

SEX a numeric vector code. 1: female, 0: male.

AGE a numeric vector code. Age of patient.

The risk factors for myocardial infarction related mortality include age, sex (female=1, male=0), clinical heart failure (CHF) (present=1), ventricular fibrillation (VF) (present=1), and diabetes (present=1). The failure time of interest is the time to death since myocardial infarction (time in years). The myocardial infarction related mortality is indicated with status=9; the censoring is indicated with status not equal to 9. The end of follow-up time is $\tau = 8$ years. Some risk factors were expected to have strongly time-varying effects, in particular, ventricular fibrillation. The effect of diabetes was also expected to decay with time.

The project includes answering the following problems.

(1) Find out how many deaths took place in a 3-year period after the myocardial infarction, and how many took place within the first month. Find out the percentage of censoring for this data set.

(2) Fit the nonparametric additive hazards regression model with sex, clinical heart failure (CHF), ventricular fibrillation (VF), diabetes, and age as regression covariates.

(3) Plot estimated cumulative coefficients along with 95% confidence intervals.

(4) Which risk factors have strong time-varying effects? Give your conclusions based on some hypothesis testing procedures. Please specify your hypotheses, test statistics and report your p-values.

(5) Based on your hypothesis testing results from (4), build a semiparametric additive model, indicating which covariates are modeled with time-varying effects and which are modeled with constant effects. Present your estimation for constant effects and the estimation for the cumulative time-varying effects along with their 95% confidence intervals.

(6) Conduct hypothesis testing to check if each of these covariates has significant effect on the myocardial infarction related mortality. Please specify your hypotheses, test statistics and report your p-values.

(7) Suppose that the analysis in (5) and (6) are based on correct models, please write a paragraph to your boss (the doctor who knows little about survival analysis) stating the findings of your data analysis.

(8) Plot the cumulative residuals versus age for fitting the nonparametric additive model in (2). Do you observe evidence of lack of fit of the model to the TRACE data? If lack of fit is observed, what are your suggestions to improve model fit? Does the fit improves if you fit the model with categorized AGE into 3 groups (lower, middle and upper) ?