Group Projects on Monte Carlo Simulation Design.

P8160 Advanced Statistical Computing

Project 1 Design a simulation study to compare three survival models;

Project 2 Design a simulation study to assess the three hypothesis testings]

Project 4 Design a simulation study to compare variable selection methods;

Project 4 Design a simulation study to compare two bootstrapping methods for propensity-score matching.

Project 5 Design a simulation study to compare two clustering methods with non-normal data

Project 1: Design a simulation study to compare three survival models

Background: Suppose $T \in [0, \infty)$ is the time to a event of interest, such as death, disease onset, device failure, etc. To analyze such data, we define a **survival function** S as

$$S(t) = \Pr(T > t) = \int_t^\infty f(s)ds$$

It measures the probability of "survive" beyond time t. If T is the time to death, then $S(t)$ is the probability of living longer than t. A closely-related concept, **hazard function** h, is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \in (t, t + \Delta t)|T > t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

where $f(t)$ is the density function of T. The hazard function measures the instantaneous risk of failure at time t giving that a patient has survived until time t.

Proportional hazards model is one of the primary approaches to investigate the efficacy of a treatment (X) on a survival time T. It assumes that the hazard ratio for the i-th patient at a time t is

$$h_i(t) = h_0(t)exp(x_i\theta),$$

where

- $h_0(t)$ is the baseline hazard function,
- x_i is a binary treatment indicator variable coded 0 for control and 1 for the treatment
- θ is the parameter of interest, which is the log hazard ratio for the treatment effect. θ measures the relative hazard reduction due to treatment in comparison to the control.

This approach is referred to as **proportional hazards** because the hazard rate

$$\frac{h(t|x_1)}{h(t|x_2)} = \exp[\beta^T(x_1 - x_2)]$$

does not depend on t.

There are different ways to formulate the baseline hazard function $h_0(t)$, which lead to different models and estimations.

An exponential proportional-hazards model assumes the baseline hazard function is a constant

$$h_0(t) = \lambda$$

A Weibull proportional-hazards model assumes the baseline hazard function follows Weibull distribution, where

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

for $\gamma > 0$

A Cox proportional-hazards model leaves $h_0(t)$ unspecified.

Note that exponential distribution is a special case of Weibull distribution where $\lambda = 1$. Hence, among the three models, the exponential proportional-hazards model is the most restrictive model, while the Cox model is the most general one.

Your tasks: Design a simulation study to compare the accuracy and efficiency of the estimated treatment effects (β) from the three models under various baseline hazard functions and evaluate their robustness against misspecified baseline hazard functions. Based on your numerical investigations, write a practical recommendation for general users to choose a suitable model.

R codes for Hazard Ratio Estimation

```r
# Exponential
fit.exponential <- survreg(Surv(y) ~ x[, 1] + x[, 2], dist = "exponential")
summary(fit.exponential)
- fit.exponential$coefficients[-1]

# Weibull
fit.weibull <- survreg(Surv(y) ~ x[, 1] + x[, 2], dist = "weibull")
summary(fit.weibull)
- fit.weibull$coefficients[-1] / fit.weibull$scale

# Cox
fit.cox <- coxph(Surv(y) ~ x[, 1] + x[, 2])
summary(fit.cox)
```

Note that in `survreg`, the output is parameterized differently (see Chapter 2.2 and 2.3 of Kalbfleisch and Prentice). Therefore we need some extra transform to obtain β.

Reference:

Cox, David R (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B. 34 (2): 187–220.

Project 2: Design a simulation study to assess the three hypothsis testings

Background: Continue from the background introduction in Project 1. In many randomized clinical trials, the primary outcome is a time-to-event. The treatment effect is summarized by the hazard ratio (HR) between the control and treatment arms, under the proportional hazards assumption. The logrank test is widely used to determine the treatment effect since it is the optimal (most powerful) rank test under the proportional hazards assumption.

Let $t_1 < t_2 < \cdots < t_J$ are J distinct failure times; The log-rank test statistics calculate the difference between "observed" and "expected" number of faitures (under H_0) at each observed failure time, and aggregate them overtime.

$$S_{Logrank} = \frac{\sum_{j=1}^{J}(O_j - E_j)}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0,1) \text{ under } H_0,$$

where O_j and E_j are "observed" and "expected" numbers of failures at the jth failure time, and V_j is the variance of the observed number of failures.

During the last decade, more and more researchers reported that non-proportional hazards (non-PH) occur fairly often in clinical trials. For example, a therapy could shown effective at an early stage, but then 'wearing off' over time; or a therapy could show a "delayed" effect which only manifest after a period of time.

To adjust for potential non-proportional hazard functions, Fleming and Harrington considered "weighted log rank tests"

$$S^w_{Logrank} = \frac{\sum_{j=1}^J W_j(O_j - E_j)}{\sqrt{\sum_{j=1}^J W_j V_j}},$$

where the weight function $w(t) = \widehat{S}(t)^\rho(1-\widehat{S}(t))^\gamma$ include two parameters ρ and γ; Choosing $(\rho = 0, \gamma = 1)$ puts more weight on late events, $(\rho = 1, \gamma = 0)$ puts more weight on early events. Researchers also considered combining different weighting schemes with appropriate multiple comparison controls. The following R functions

R codes for log-rank and weighted log-rank tests

```
library(nph)
# log-rank
logrank.test(dat$y, dat$event, dat$group, rho   =  0,  gamma =  0 )

# weighted log-rank for a late effect
logrank.test(dat$y, dat$event, dat$group, rho   =  0,  gamma =  1 )

# weighted log-rank for an early effect
logrank.test(dat$y, dat$event, dat$group, rho   =  1,  gamma =  0 )

# take maximum  over the previous three log-rank with multiple comparison control
logrank.maxtest(dat$y, dat$event, dat$group)
```

Your tasks: Design a set of distributions/models under both proportional-hazard and non-proportional-hazard assumptions, and carry out a simulation study to compare performance of those hypothesis tests in those models. Based on your numerical investigations, write a practical recommendation for general users to choose a suitable testing tool.

Reference Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. Biometrika 69, 553-566.

Project 3: Design a simulation study to compare variable selection methods

Background: When the number of candidate predictors in a linear model is large, variable selection is a common practice to find an optimal model that balances between model fitness and model complexity.

Step-wise forward method: Starting with the empty model, and iteratively adds the variables that best improves the model fit. That is often done by sequentially adding predictors with the largest reduction in AIC. For linear models,

$$AIC = n\ln(\sum_{i=1}^n (y_i - \widehat{y}_i)^2/n) + 2p,$$

where $\widehat{y}_i$ is the fitted values from a model, and p is the dimension of the model (i.e.,number of predictors plus 1).

Automated LASSO regression LASSO is another popular method for variable selection. It estimates the model parameters by optimizing a penalized loss function:

$$\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \| \sum_{k=1}^{p} |\beta_k|$$

where λ is a tunning parameter. Cross-validation (CV) is the most common selection criteria for LASSO.

Your tasks: In modern applications with high-dimensional covariates, traditional variable selection methods often struggle with the presence of "weak" predictors. Design a simulation study to investigate and illustrate (1) how well each of the two methods in identifying weak and strong predictors, and (2) how missing "weak" predictors impacts the parameter estimations.

To do so, you need to simulate data with a combination of `strong'',`weak-but-correlated" and "weak-and-independent" predictors. Their definition can be found in the following.

Definition of strong signals —

$$S_1 = \{j : |\beta_j| > c\sqrt{log(p)/n}, \text{ some } c > 0, 1 \leq j \leq p\}$$

Definition of weak-but-correlated signals —

$$S_2 = \{j : 0 < |\beta_j| \leq c\sqrt{log(p)/n}, \text{ some } c > 0, \text{corr}(X_j, X_j') \neq 0, \text{for some } j' \in S_1, 1 \leq j \leq p\}$$

Definition of weak-and-independent signals —

$$S_3 = \{j : 0 < |\beta_j| \leq c\sqrt{log(p)/n}, \text{ some } c > 0, \text{corr}(X_j, X_j') = 0, \text{for all } j' \in S_1, 1 \leq j \leq p\}$$

```r
# A sample codes for generating a data with a combinaton of true predictors and null predictors.
n <- 1000
p <- 50
X <- matrix(rnorm(n * p), n, p)
b.true <- rnorm(p) * (runif(p) < 0.2)
cat("True non-zero effects:", which(b.true != 0), "\n")
Y <- 1 + X %*% b.true + rnorm(n)
df <- data.frame(cbind(X, Y))
names(df)[p + 1] <- "y"

# Forward Selection
fit.forward <- step(object = lm(y ~ 1, data = df),
                    scope = formula(lm(y ~ ., data = df)), direction = "forward", k = 2n, trace = 0) # 
summary(fit.forward)

# LASSO
fit.lasso <- cv.glmnet(X, Y, nfolds = 10, type.measure = "mse") # 5-fold CV using mean squared error
param.best <- fit.lasso$glmnet.fit$beta[, fit.lasso$lambda == fit.lasso$lambda.1se] # one standard-erro
param.best[param.best != 0]
```

R codes for forwarding selection and LASSO

Project 4: Compare bootstrap methods for propensiy score matching

Background: Propensity-score matching has been widely used to estimate the effects of treatments, exposures and interventions from observational data. It is an effective way to reduce or minimize the confounding effects.

An important issue, however, is how to estimate the standard error of the estimated treatment effect when a propensity-score matching is used. Direct variance estimation is often hard. Some researchers consider the use of bootstrapping to estimate the sampling variability of treatment effects to ensure accurate inferences.

Design a simulation study to examine the performance of two bootstrap methods to estimate the sampling variability of treatment effect estimates obtained from a nearest-neighbor propensity-score matching. Using this matching approach, a treated subject is selected at random from an observational data. This treated subject is then matched to the untreated subject whose propensity score is closest to that of the treated subject. Matching without replacement was used, so that once an untreated subject was selected for matching to a given treated subject, that untreated subject was no longer eligible for matching subsequent treated subjects.

Estimating treatment effects. For continuous outcomes, the effect of treatment can be estimated as the difference between the mean outcome in the treated subjects in the matched sample and the mean outcome in the untreated subjects in the matched sample. For binary outcomes, the effect of treatment can be estimated as the difference between the proportions in the treated subjects and their matched untreated subjects.

We consider two bootstrap approaches to estimate the variances of the treatment effect.

1. the simple bootstrap: obtain a bootstrap sample by bootstrapping matched pairs, and estimate the treatment effect from bootstrap sample. The standard deviation of the estimated treatment effects across the B bootstrap samples is used as an estimate of the standard error of the estimated treatment effect in the original propensity-score-matched sample.

2. the complex bootstrap: the complex bootstrap attempts to incorporate two additional sources of variability compared with that addressed by the simple bootstrap: variability in estimating the propensity-score model and variability in the formation of the propensity-score-matched sample. Using this approach, a bootstrap sample are drawn from the original (unmatched) observational data. From that bootstrap sample, the propensity-score model is re-estimated, and a propensity-score-matched sample is re-formed using the nearest-neighbor propensity-score matching. The treatment effect is then estimated from the marched sample.

Your tasks: Design a simulation study to compare the performance of the simple bootstrap and the complex bootstrap in estimating the sample variablities of the estimated treatment effects using propensity-score matching. You can consider a linear logistic regression as the propensity model. Report your findings, and make recommendations on whether bootstrap is a suitable approach, and if so, which bootstrap method should be used.

Project 5: Design a simulation study to compare two clustering methods with non-normal data

Clustering is a powerful unsupervised learning approach to discover intrinsic groups in an unstructured data set. Many clustering methods have been proposed in the literature. Among them, k-Means and Latent Class Analysis (LCA) are the two best-known methods applied widely in medical applications.

K-means is a simple nonparametric approach grouping observations based on their similarities and spatial locations. LCA, on the contrary, is a model-based method that assumes Gaussian Mixture distributions. You can find simple introductions to these two clustering methods and the R packages/functions in what follows.

https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

https://cran.r-project.org/web/packages/mclust/vignettes/mclust.html#clustering

Your tasks: Both methods are proven successful and well-tested when the normally or elliptically distributed data. In real world applications, data often exhibit non-normal features, including asymmetry/skewed, multimodality, heavy-tails, and the presence of outliers. How well and how robust they perform for non-elliptical distributions are less understood and investigated. Design a simulation study to comprehensively assess and compare the performance of K-means and LCA when the data are non-normal. (You may stay with bivariate cases.)