STA242 Proposal for final project

Chi Po Choi (UCD ID:912494157), Amy Kim (UCD ID:912492829)

April 24, 2015

1 What we are trying to do

We would like to write a R package for the *Hypercube estimators* [1] introduced by Rudolf Beran. We have noticed there is no package available to utilize the hypercube esimating method, so it would be nicer if there is such a package, then people can get estimators by more convenient way.

2 Description of Hypercube estimators

Given a data set (y, X), we fit the model:

$$y = X\beta + \epsilon.$$

Let $\eta = \mathrm{E}(y) = X\beta$. We would like to find an estimator $\hat{\eta}$ so that the risk $\mathrm{E}|\hat{\eta} - \eta|^2$ is minimized. One should notice that the least square estimator of η does not minimized the risk.

Define the hypercube estimator of η to be

$$\hat{\eta}_{\mathrm{H}}(V) = XV(VX'XV + I_p - V^2)^{-1}VX'y$$

where V is a symmetric matrix with all eigenvalues $\in [0, 1]$. We can compute the V so that the risk $\mathrm{E}|\hat{\eta}_{\mathrm{H}}(V) - \eta|^2$ is minimized. Thus, we obtain an estimator which is better than the least square estimator.

One advantage of hypercube estimator $\hat{\eta}_{\mathrm{H}}(V)$ is that computation of the inverse of $VX'XV + I_p - V^2$ is numerically stable. It is proved in Beran's article [1].

Another advantage is that hypercube estimator $\hat{\eta}_{\mathrm{H}}(V)$ is a generalization of penalized least squares estimator and submodel least square estimator.

1. Penalized least squares estimator

   $$\hat{\eta}_{\mathrm{pls}} = X\hat{\beta}_{\mathrm{pls}}$$

   where

   $$\hat{\beta}_{\mathrm{pls}} = \arg\min_{\beta}[|y - X\beta|^2 + \beta' W\beta] = (X'X + W)^{-1}X'y$$

2. Submodel least squares estimators:

   $$\hat{\eta}_{\mathrm{sub}} = X_0 X_0^{+} y \qquad \text{where } R(X_0) \subset R(X) \text{ and } {}^{+} \text{ denotes Moore-Penrose pseudoinverse}$$

3 Description of the R package

We would like to write a `R` package for the hypercube estimators. Following the guideline in Leisch's tutorial [2] .The package should contain functions which facilitate data analysis using hypercube estimators. We will also define corresponding classes and methods. We want classes and methods to be compatible with the class `lm` in `R`.

For example, we will implement the function `helm()` (stands for HyperCube Estimator Linear Model), `sublm()` (Submodel least square estimation) and `plm()` (Penalized least square estimation). Those functions return variables of the class `helm` with methods `summary`, `predict`, `coef`, `residuals`, etc.

4 Our plan on the R package

- We implement functions for hypercube estimators.
- We define classes and methods for hypercube estimators.
- We include example data sets.
- We write full documentation on the classes and functions in the package.
- We maintain a git repository for the package.
- The package satisfies the CRAN repository policy and will be eventually available in the CRAN repository .

5 Expected difficulties

- We want our package to be accessible to those people are not familiar with the details of the hypercube estimators. We need to design an "easy interface" which make the functions `sublm` and `plm` just work. It may be difficult to design a robust way to provide suitable V for different user-input data.
- Besides the "easy interface", we want more feasible interface which allows advanced users do some small tweaks. We want to make those advanced users feel that using our package is easier than writing their own codes. To find the balance between "easy" and "hackable" may be difficult.
- As mentioned in the description of Hypercube estimators, hypercube estimators are numerical stable [1]. We want our package do have this advantage. We need to make sure that the functions we implement can handle large data set.
- Our team does not have previous experience in building softwares. We will need some time to learn.

6 Software

- We use `RStudio` because it has convenient features for building package.
- If possible, we will try to develop the package purely in R.

References

[1] Rudolf Beran. Hypercube estimators: Penalized least squares, submodel selection, and numerical stability. *Computational Statistics & Data Analysis*, 71:654–666, 2014.

[2] Friedrich Leisch. Creating r packages: A tutorial. 2008.