Bio 754: Homework 3

May 3, 2011

See the class site for due date, guidance on how to answer, and notice of any corrections or clarifications. For problems which require computation, please append neat and commented R code as an appendix to your homework.

1. [**Bootstrap**] In your own words, write a description of how use of the (standard, non-parametric, observation-resampling) bootstrap approximates simple random sampling from superpopulation F, and in particular how this approximation is used in the construction of approximate confidence intervals. You description should cover;

   (a) Bootstrap intervals constructed from quantiles of the bootstrapped point estimates

   (b) Bootstrap intervals constructed from the mean and variance of the bootstrapped point estimates

   Notes: You are **not** asked to describe the computations involved. Instead, in words not equations, describe every approximation involved, stating what is being approximated. Write for an audience of statisticians who understand superpopulation-based parameter definitions, and frequentist inference, but who are unfamiliar with the bootstrap.

2. [**Bootstrap Hypothesis Test**] For the stamp problem described in Lecture 10, write code to calculate $\hat{h}_1$, perform the smoothed bootstrap, and calculate a bootstrap p-value for testing H_0 : number of modes = 1. The stamp data are available in the `bootstrap` R package and can be accessed like so:

```
> library(bootstrap)
> data(stamp)
> dat = stamp[[1]]
```

3. [**Choice of working correlation matrix in GEE**] This question examines the effect of using different correlation structures, designs and sample sizes in GEE using a simulation study. It is also an exercise in writing code systematically; please take care to break the required programming into small tasks, and write individual functions to

do each of these tasks. You may use an R package to get the `gee` estimates described below.

For the marginal model

$$\mathbb{E}[\,Y_{ij}|X_{ij} = x_{ij}\,] = \beta_0 + \beta_1 x_{ij},$$

we consider estimating functions of the form

$$\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \mathbf{0}_p,$$

where $\boldsymbol{W}_i$ is an $n_i \times n_i$ working covariance matrix, and the other notation follows the lecture notes. We shall assume throughout that;

- $\beta_0 = 0$, $\beta_1 = 0.6$
- $\boldsymbol{Y}_i \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \phi\boldsymbol{V}_i)$ with $\phi = 1$. You will find the `mvtnorm` package helpful.

The factors that will vary are;

- The number of clusters; $n = 8, 20, 60$
- The design;

  Design I is balanced. It has $n_i = 4$, for all clusters. In each cluster, we see $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}\} = \{8, 10, 12, 14\}$

  Design II is unbalanced. While we have $n_i = 3$ for all clusters, we see equal numbers of clusters with $\boldsymbol{x}_i = \{8, 10, 12\}, \{8, 10, 14\}, \{8, 12, 14\}, \{10, 12, 14\}$
- The true covariance and the working covariance matrix. The form of these is $\phi\boldsymbol{R}_i$. For the true covariance, we consider Exchangeable and AR-1 correlation structures, with $\alpha = 0.5$ or $\alpha = 0.9$, and $\phi = 1$. For the working covariance, we consider these structures, with unknown ϕ and unknown α, and additionally the Independence working correlation matrix, with unknown ϕ.

Present your results in the following table, and write a paragraph or two summarizing your findings. (Bonus: give the output as figures, not tables)

			Coverage Working Cov			Var($\hat{\beta}_1$) Working Cov			Relative Efficiency Working Cov		
n	Design	True Corr	Ind	Exch	AR-1	Ind	Exch	AR-1	Ind	Exch	AR-1
8	I	Exchangeable $\alpha = 0.5$									
8	I	Exchangeable $\alpha = 0.9$									
8	I	AR-1 $\alpha = 0.5$									
8	I	AR-1 $\alpha = 0.9$									
8	II	Exchangeable $\alpha = 0.5$									
8	II	Exchangeable $\alpha = 0.9$									
8	II	AR-1 $\alpha = 0.5$									
8	II	AR-1 $\alpha = 0.9$									
20	I	Exchangeable $\alpha = 0.5$									
20	I	Exchangeable $\alpha = 0.9$									
20	I	AR-1 $\alpha = 0.5$									
20	I	AR-1 $\alpha = 0.9$									
20	II	Exchangeable $\alpha = 0.5$									
20	II	Exchangeable $\alpha = 0.9$									
20	II	AR-1 $\alpha = 0.5$									
20	II	AR-1 $\alpha = 0.9$									
60	I	Exchangeable $\alpha = 0.5$									
60	I	Exchangeable $\alpha = 0.9$									
60	I	AR-1 $\alpha = 0.5$									
60	I	AR-1 $\alpha = 0.9$									
60	II	Exchangeable $\alpha = 0.5$									
60	II	Exchangeable $\alpha = 0.9$									
60	II	AR-1 $\alpha = 0.5$									
60	II	AR-1 $\alpha = 0.9$									

Table of true coverage (percentage), at the nominal 95% level, variance of the estimated slopes, and efficiency of the estimates of the slope. Efficiencies are given relative to the most efficient estimator, in each row

4. [**Perform a Service For Future Students**] Report up to 5 typos in the notes, for 1 bonus point each. Corrections of mathematical formulae will be particularly appreciated.