Stat 8053 (Geyer) Fall 2014
Homework Assignment 1
Due Friday, September 26, 2014

Instructions: do not hand in a mess. You can have an appendix or "supplementary" material that is a mess, but we want to see a concise description of what you did and *more important* what you conclude from what you did and your argument as to why what you did implies what you conclude.

1-1. The dataset `scor` in the `bootstrap` package, which is a CRAN package that goes with the book *An Introduction to the Bootstrap* by Efron and Tibshirani (Chapman & Hall, 1993), gives scores on 5 tests (mechanics, vectors, algebra, analysis, and statistics) taken by 88 students. Since these are all somewhat related mathy subjects, Efron and Tibshirani suggest using principal components for dimension reduction

```
library(bootstrap)
data(scor)
foo <- eigen(var(scor), symmetric = TRUE)
```

gives the eigenvalues and eigenvectors of the sample variance matrix of the data. They look at the first two eigenvalues and the ratio of the sum of these eigenvalues to the sum of all five eigenvalues (the "fraction of variance explained by the first two principle components). They also bootstrap these two eigenvalues, this ratio, and the corresponding two eigenvectors. They do a really kludgy analysis of the bootstrap distribution of the eigenvectors using boxplots. You should be able to think of something better.

One issue with bootstrapping this analysis is that the signs of eigenvectors are arbitrary. You need to fix up the bootstrap eigenvectors so they point in more or less the same direction as the corresponding eigenvectors for the original data.

(a) Do a nonparametric bootstrap of these data. Make 95% confidence intervals (not adjusted for simultaneous coverage) for all five eigenvalues and for the ratio of the sum of the first k eigenvalues to the sum of all eigenvalues (for each k). Use some reasonable method for your bootstrap confidence intervals (your choice).

(b) For the first eigenvector devise a (bootstrap) method that gives some sort of statistical inference about the direction of this eigenvector. Your method should say something about the variability as a five-dimensional vector. What does your method say about this vector?

1-2. The dataset Duncan in the car package, which is a CRAN package that goes with the book *An R Companion to Applied Regression* by Fox and Weisberg (SAGE Publications, 2010), contains four variables: type (of occupation), income (percent of males in occupation earning $3500 or more in 1950), education (percent of males in occupation in 1950 who were high-school graduates), prestige (percent of raters in NORC study rating occupation as excellent or good in prestige). Despite the oldness of the data (published in 1961) and despite the rather odd way it has been reduced to these four variables, we will analyze it. There are several "outliers" so we will use robust regression.

(a) Compare the results of using lm, ltsreg, and rlm, the latter with both methods ("M" and "MM"), on the formula

```
prestige ~ income + education
```

with these data.

(b) The ltsreg function does not come with standard errors. Compute bootstrap standard errors for it.

(c) Use the nonparametric bootstrap (bootstrapping residuals rather than cases) to check the standard errors for rlm, method "MM". Make 95% confidence intervals for the regression coefficients (not adjusted for simultaneous coverage) based on the standard errors given by the summary function and also make nonparametric bootstrap t confidence intervals.

1-3. The dataset LakeHuron in the R core (what you get with every install) is a time series. We are going to fit some time series models to it. An AR(k) time series with mean zero has the form

$$X_n = \rho_1 X_{n-1} + \rho_2 X_{n-2} + \cdots + \rho_k X_{n-k} + Y_n$$

where the Y_i are independent and identically distributed mean zero normal random variables (also Y_n is independent of X_1, ..., X_{n-1}). The R code

```
ar.mle(foo, order.max = k, aic = FALSE)
```

fits this model to the time series data foo. If instead the aic = FALSE is omitted

```
ar.mle(foo)
```

then it selects the order k by AIC. The predict function does prediction of *future data*, for example,

```
out <- ar.mle(foo)
predict(out, n.ahead = 5)
```

predicts the next five (future) data points. It also gives standard errors if the argument `se.fit = TRUE` is given (which is the default).

(a) Fit an AR(k) model to the `LakeHuron` data. If `aic = TRUE`, what order is selected?

(b) Do a subsampling bootstrap fitting AR models to the `LakeHuron` data, again allowing order selection by AIC. Ideally we want the subsample size b to be large compared to one and small compared to n, but $\sqrt{n}$ is too small to use the `ar.mle` function with its defaults. Hence the need to either use a larger batch length or a more stable estimation method. Your choice. (The help page for ar.mle also describes two other estimation methods.) The question we want this subsampling bootstrap to answer is how stable the order selection is. What does the bootstrap say about the order selection probabilities?

(c) Predict five steps ahead based on your fit to the observed data.

(d) Do another subsampling bootstrap fitting AR models (with order selection by AIC) and predicting five steps ahead for the `LakeHuron` data. The question we want this subsampling bootstrap to answer is how good the prediction is and how good the standard errors given by the predict function are. What does this bootstrap say about that?

Be careful about relating the subsample predictions to the full sample predictions. Note that if this were ordinary linear regression, the variance for a "prediction interval" is $\sigma^2 + c/n$, where c is a constant (a complicated function of the model matrices for the original data and the new data being predicted). Hence prediction errors do not obey the "square root law" even if the estimates do (*part* of the prediction error obeys the square root law and part is constant, not a function of n). Since the help page for `predict.ar` does not tell us exactly how it calculates the prediction errors, doing this really well will not be easy. Do something sensible.