STAT-S 470/670 Homework 6

Due: Wednesday 11/18/2015

1. Show that the jackknife estimate and the jackknife SE are exactly the same as sample mean $\bar{x}$ and standard error $se(\bar{x}) = s/\sqrt{n}$ when $\hat{\theta} = \bar{x}$ (sample mean).

2. Show that the standard error of the pseudo-values (jackknife SE) is the same as the standard deviation of the "leave-out-one" values multiplied by $(n-1)/\sqrt{n}$; i.e., that the standard error of the pseudo-values (jackknife SE) is the same as the standard error of the "leave-out-one" values multiplied by $(n-1)$.

3. Let x = LSAT, y = GPA (below, I multiplied GPA by 100):

x	576	635	558	578	666	580	555	661	651	605	653	575	545	572	594
y	339	330	281	303	344	307	300	343	336	313	312	274	276	288	296

   Let $\theta = 0.5\log_e((1+\rho)/(1-\rho))$, where ρ is the correlation coefficient between x and y, to be estimated by Pearson's correlation coefficient ($r = 0.77637$).

   (a) Fisher showed that the distribution of $\hat{\theta} = 0.5\log_e((1+r)/(1-r))$ is roughly Gaussian with mean $0.5\log_e((1+\rho)/(1-\rho))$ and variance $1/(n-3)$. Use Fisher's result to calculate an approximate 95% CI for θ.

   (b) Use the jackknife method to estimate θ, and derive an approximate 95% CIs for θ.

   (c) Stem-and-leaf the pseudo-values.Do you see anything odd? Examine the scatterplot. Remove the first observation from the sample and repeat (b).

   (d) Use the bootstrap method to estimate θ.

   (e) Derive approximate 95% CIs for θ via the bootstrap. And compare them with the ones you obtained in (a), (b) and (d).

4. In simple linear regression we are interested in modeling the response y as a simple linear function of some predictor variable x under the model

$$y = \alpha + \beta x + \epsilon$$

   where α is the y-intercept, β is the slope and ϵ is a vector of errors or departures from the line. In resistant regression we are interested in estimating the y-intercept and slope parameters in a resistant fashion. For example, the program run.rrline() is a program which will calculate estimators for the parameters $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ in a resistant fashion, however I have no closed form estimators for the standard errors $\text{SE}(\hat{\alpha})$ and $\text{SE}(\hat{\beta})$. Write a program in R, which will perform non-parametric bootstrap re-sampling of a dataset and return the bootstrap estimators for the population parameters $(\hat{\alpha}_B, \hat{\beta}_B)$, the bootstrap estimators for the standard errors $\{\hat{\text{SE}}(\hat{\alpha}), \hat{\text{SE}}(\hat{\beta})\}$, the bootstrap estimators for the bias of $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ and an estimator the out of sample predictive error (or out-of-bag error) $\hat{\text{Err}}$. Your function should be constructed so that it has two arguments: a data frame and the number of bootstrap replicates, R. The input data frame should contain a column for the y variable and a column for the x variable, and the column names of the data frame can be labeled as such. Once you have constructed this function practice using this function on the "faithful" dataset in R. For this particular faithful data set, treat x-variable as waiting and the y-variable as eruptions.

5. In this problem, I will ask you to do something similar as the previous problem, except this time use k-fold cross-validation rather than bootstrapping. That is to say, for the program run.rrline() construct a function which will calculate the k-fold cross-validation estimator for the out-of-sample error $\hat{\text{Err}}$. Your function should be constructed so that it has two arguments: a data frame and the number of desired CV folds k. Like the previous problem, the input data frame should contain a column for the y variable and a column for the x variable, and the column names of the data frame can be labeled as such. Once you have constructed this function practice using this function on the "faithful" dataset in R, treating the x-variable as waiting and the y-variable as eruptions.