Assignment 6 – Bootstrapping

Math 154, Computational Statistics
Fall 2015, Maria Martinez

Due: Tuesday, October 13, 2015, noon

NOT RELATED TO HW SCORE:

Total hours spent on assignment: 4:20- _____

Number of different "sittings" to finish assignment: _____

Assignment

1. Problem 9 in section 5.4 of *An Introduction to Statistical Learning*, `http://www-bcf.usc.edu/~gareth/ISL/`.

```
require(MASS)
data(Boston)
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

   (a) Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$.

```
  mu_hat = mean(Boston$medv)
  mu_hat
```

```
## [1] 22.53281
```

   (b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret the result.

```
  SE = sd(Boston$medv)
  SE
```

```
## [1] 9.197104
```

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does the estimate compare to your answer from (b)?

```
  set.seed(1)
  require(boot)

  #sampletmean <- function(x,d,trimperc){
  #  return(mean(x[d], trim=trimperc))
  #}

  boot.mean <- function(data=Boston, index=1:506){
    mean(data$medv[index])
  }
  bootMean = boot(Boston, statistic = boot.mean, R=1000)
  SEMean = sd(bootMean)
```

```
## Error in is.data.frame(x):  (list) object cannot be coerced to type 'double'
```

```
  bootMean = as.numeric(boot(Boston, statistic = boot.mean, R=1000)[1])
  bootMean
```

```
## [1] 22.53281
```

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of `medv`. Compare it to the results obtained using `t.test(Boston$medv)`.

```
  confint1 = mu_hat - 2*SE
  confint2 = mu_hat + 2*SE
  unlist(list(confidence = confint1, interval = confint2))
```

```
## confidence   interval
##   4.138598  40.927014
```

```
  t.test(Boston$medv)
```

```
##
##  One Sample t-test
##
## data:  Boston$medv
## t = 55.1111, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

```
  mu_hat + qt(c(.025,.975),nrow(Boston)-1)*SE
```

```
## [1]  4.463508 40.602105
```

(e) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of medv in the population.

```
medianB = median(Boston$medv)
medianB
```

```
## [1] 21.2
```

(f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

```
boot.median<-function(data=Boston, index=1:506){
  median(data$medv[index])
}
as.numeric(boot(Boston,boot.median, R=1000)[1])
```

```
## [1] 21.2
```

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$. (Use the `quantile()` function.)

```
boot(Boston, boot.quantile, R=1000)
```

```
## Error in boot(Boston, boot.quantile, R = 1000):  object 'boot.quantile'
not found
```

(h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

2. Building on problem 9 of section 5.4 for each parameter (true population mean & median `medv` (median value of owner-occupied homes in $1000s in Boston in the 1970s)) find the following CI:

   (a) Bootstrap percentile interval.

   (b) BS-t interval

   (c) t-CI with the BS SE

   The percentile and BS-t intervals are given by the `boot.ci` function. Note that in order to get the t-cutoffs (for the BS-t interval), the second layer of bootstrapping was needed. To get the t interval with BS SE, use the SE from the `boot` output.

3. Why are there two levels of bootstrapping needed to calculate the BS-t interval?

4. Explain why the percentile interval is range respecting but the bootstrap-t interval is not. An example might help the explanation.

5. **The effect of outliers.** We know that outliers can strongly influence some statistics. Consider an observational study on whether the full moon is associated with agressive behavior in nursing home patients with dementia. The number of incidents of agressive behavoir was

recorded each day for 12 weeks. A "moon day" is one when there was a full moon that day, the day before, or the day afterward. The data below give the average number of aggressive incidents for moon days and other days for each of 15 subjects. A few things to note about the example:

- An independent two sample analysis is not appropriate here because the data are paired within each subject (that is, when bootstrapping, consider the paired nature of the data).
- There are a few observations with unusual observations (difficult to see in the raw data).
- The unusual observations may affect the intervals.

```
patient = 1:15
moonIncidents = c(3.33, 3.67, 2.67, 3.33, 3.33, 3.67, 4.67,
                  2.67, 6, 4.33, 3.33, 0.67, 1.33, .33, 2)
otherIncidents = c(.27, .59, .32, .19, 1.26, .11, .3, .4, 1.59,
                   .6, .65, .69, 1.26, .23, .38)
moondata = data.frame(patient, moonIncidents, otherIncidents)
```

(a) Bootstrap the mean of the differences with and without the three low values. How do these values influence the shape, variance, and bias of the bootstrap distribution?

```
meanDiff <- function(data=moondata, index=1:15){
  mean(data$moonIncidents[index] - data$otherIncidents[index])
}
moonMean = boot(moondata, statistic = meanDiff, R=1000)
```

(b) Give the percentile and t-CI (with BS SE) intervals from the bootstrap distributions with and without the outliers. Discuss the differences.