The University of British Columbia

Data Science 581 Modelling and Simulation II

Lab Assignment 1 Solutions

Exercises

Do the following exercises, and hand in #1, #2, #3, #6 and #7 for credit.

1. Enter the function `SEcalculator()` in R. Use it to obtain an estimate of the standard error of the following:

   (a) the mean annual change of level in Lake Huron. Compare with the usual estimate $s/\sqrt{n}$.

   ```
   SEcalculator(changes, mean)
   ## [1] 0.07567375
   sd(changes)/sqrt(length(changes))
   ## [1] 0.07605443
   ```

   The bootstrap estimate is similar to the usual estimate.

   (b) the variance of the annual change of level in Lake Huron.

   ```
   SEcalculator(changes, var)
   ## [1] 0.07837701
   ```

   In this case, calculating the conventional standard error estimate is fairly complicated.

2. Write a function called `q90()` which takes a single argument `x` and returns the 90th percentile of `x`, specifically,

   ```
   quantile(x, prob = 0.9)
   ```

   Then estimate the standard error of the 90th percentile estimate for the annual changes in the Lake Huron water levels.

   ```
   q90 <- function(x) {
       quantile(x, prob = 0.9)
   }
   SEcalculator(changes, q90)

   ## [1] 0.1445608
   ```

3. Suppose X_1 and X_2 are independent Bernoulli random variables with common parameter p.

   (a) Write down the likelihood function for p.
   hint: PDF for Bernoulli distribution is:

   $$f(x,p) = p^x(1-p)^1 - x, x \in \{0,1\}$$

   solution:

   $$L(p) = p^{X_1+X_2}(1-p)^{2-X_1-X_2}.$$

(b) Suppose the observed data are $x_1 = 1$ and $x_2 = 0$. If the possible parameter values lie in the set $\{.2, .7, .9\}$, find the maximum likelihood estimate of p.
solution:

$$L(.2) = .2(.8) = .16.$$
$$L(.7) = .7(.3) = .23.$$
$$L(.9) = .9(.1) = .09.$$

Therefore, the maximum likelihood estimate of p is 0.7.

(c) Refer to the preceding part, and suppose the possible parameter values lie in the set $\{p : 0 < p < 1\}$. Find the maximum likelihood estimate of p.
solution:

$$\log(L) = \log(p) + \log(1 - p)$$

Differentiating with respect to p gives

$$\frac{1}{p} - \frac{1}{1-p}$$

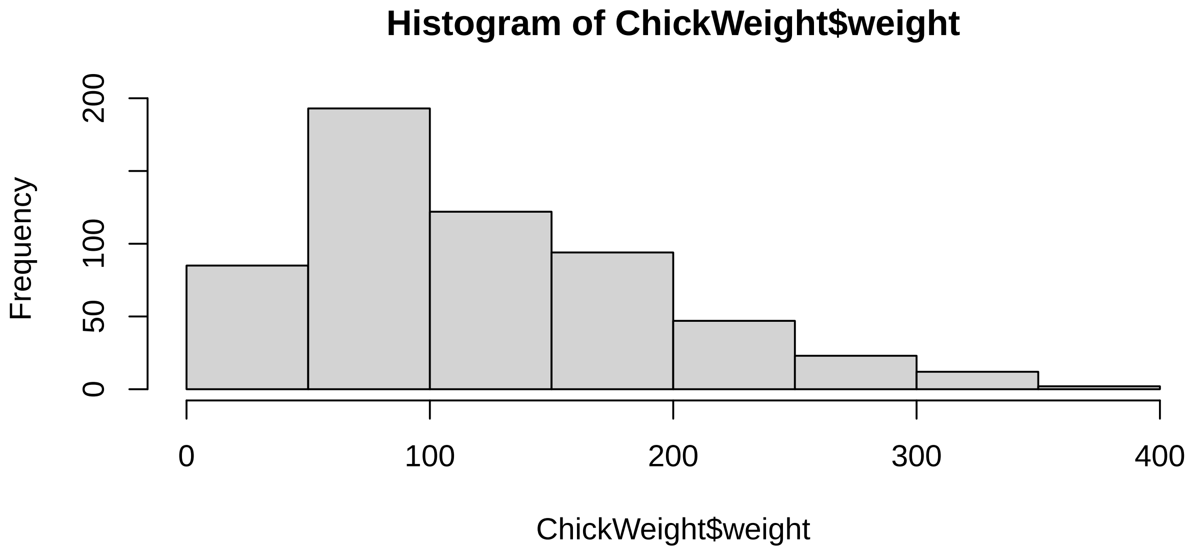
Setting this to 0, and solving for p gives $\widehat{p} = 0.5$.

4. Identify the names of the `ChickWeight` data frame, and use the `$` operator and `hist()` to plot a histogram of the weights.

```
names(ChickWeight)

## [1] "weight" "Time"   "Chick"  "Diet"

hist(ChickWeight$weight)
```



5. Enter the following table of data into a file called `demo.txt`:

```
name age height
Mary  22   171
Bob    23   175
Sue     21   166
Kim     19   167
```

Read the data into a data frame called `demo`, and use the `$` and `mean()` function to calculate the average of the heights.

```r
demo <- read.table("demo.txt", header=TRUE)
mean(demo$height)
```

6. Consider the `p13.2` data frame in the *MPV* package. Use the following code to obtain it:

```r
install.packages("MPV") # if MPV is not installed
library(MPV) # loads MPV

install.packages("boot") # if boot is not installed
library(boot) # loads boot
```

(a) Read the help file on `p13.2`, i.e. `help(p13.2)` to obtain information on the data. Why is this an example of data for which binary logistic regression makes sense?

The home ownership variable is binary, so this could be used as the response variable in a logistic regression model.

(b) Fit the binary logistic model to the data, using the `glm()` function, assign the output to an object called `p13.glm`.

```r
p13.glm <- glm(y ~ x, data = p13.2, family=binomial)
```

(c) By using the command `is.list(p13.glm)`, find out whether the output from `glm()` is a list.

```r
is.list(p13.glm)

## [1] TRUE
```

The output is a list.

(d) Write out the fitted logit model.

```r
coef(p13.glm)

##   (Intercept)             x
## -8.7395139021  0.0002009056
```

The fitted model is

$$\text{logit}(p) = -8.74 + 2.009 \times 10^{-4}x$$

where p is the probability of home ownership and x is family income.

7. Install the package `boot()`. Read the help file for `boot()` which executes the resampling of your dataset and calculation of your statistic(s) of interest on these samples. Before calling boot, you need to define a function that will return the statistic(s) that you would like to bootstrap. The first argument passed to the function should be your dataset. The second argument can be an index vector of the observations in your dataset to use or a frequency or weight vector that informs the sampling probabilities.

   consider the following sample:

```
set.seed(123) # use this seed replicate same result
x <- c(12, 14, 14, 15, 18, 21, 25, 29, 32, 35)

#define function to return mean, we include all observations
myMean <- function(x,i){mean(x[i])}
```

   We can pass `myMean` to `boot()` to calculate SE for the mean of the above sample.

   Create a function for calculating median and report the SE for statistic using 200 bootstrapping samples by using `boot()`.

```
library(boot)

##
## Attaching package:  'boot'
## The following object is masked from 'package:MPV':
##
##     motor
## The following object is masked from 'package:lattice':
##
##     melanoma

 myMedian <- function(x, i) median(x[i])
 boot(data = x, statistic = myMedian, R =200)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = x, statistic = myMedian, R = 200)
##
##
## Bootstrap Statistics :
##     original  bias    std. error
## t1*     19.5    1.24        4.56
```

   reported SE is guite large meaning the variance of the median is large. We might choose other statistic to better summarize the population.

8. Consider the `p13.7` data frame in *MPV* package.

(a) Read the help file on `p13.7`. Why would modelling the number fractures in coal seams as a function of the other variables be an example of Poisson regression?

The number of fractures is a non-negative count. A Poisson distribution is a possible model to try.

(b) Fit a Poisson regression model relating the numbers of fractures to `x2`. Write out the fitted model.

```r
coal.glm <- glm(y ~ x2, data = p13.7, family=poisson)
coef(coal.glm)
```

```
## (Intercept)          x2
##     -3.3286      0.0523
```

$$\widehat{\log \lambda} = -3.329 + 0.052 x_2$$