STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 9

Example

The data set Bangladesh has measurements on water quality from 271 wells in Bangladesh. There are two missing values in the chlorine variable. Use the following R code to remove these two observations

```
library(resampledata)
data("Bangladesh")
head(Bangladesh)
```

```
##   Arsenic Chlorine Cobalt
## 1    2400      6.2   0.42
## 2       6    116.0   0.45
## 3     904     14.8   0.63
## 4     321     35.9   0.68
## 5    1280     18.9   0.58
## 6     151      7.8   0.35
```

```
df <- with(Bangladesh, Chlorine[!is.na(Chlorine)])
```

- Find a 95% CI for the mean μ of chlorine levels in Bangladesh wells.
- Find the 95% bootstrap percentile, bootstrap t, and Bca confidence intervals for the mean chlorine level, and compare results. Which confidence interval will you report?

Chapter 9: Regression Models

Basic of Linear Regression

- Quantifying the relationship between dependent and independent variable(s).
- We are concerned with linear regression in which the mean of the response Y observed at value of $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ the explanatory variable vector is $E(Y|X = \mathbf{x}) = \mu(\mathbf{x}) \equiv \mathbf{x}'\beta$.
- Use of scatter plot/matrix to investigate relationship.
- Diagnostics plots to check the assumptions.
- Look at regression coefficients and confidence intervals.
- Can be used for predictions.

Basic of Linear Regression

- If the least squares estimation procedure is used to estimate the regression parameters and the model is reasonable and the noise term can be considered to be independent and identically distributed random variables with mean 0, and finite variance σ^2, bootstrapping will not add anything.
- That is because of the Gauss – Markov theorem that asserts that the least squares estimates of the regression parameters are unbiased and have minimum variance among all unbiased estimators.
- Moreover, if the residuals can be assumed to have a Gaussian distribution, the least squares estimates have the nice additional property of being maximum likelihood estimates and are therefore the most efficient (accurate) estimates.
- If the reader is interested in a thorough treatment of linear regression, Draper and Smith (1981, 1998) are very good sources.

Linear Regression Model

Why bootsrap?

- Looking at Bootstrap distributions helps you understand what is going in the regression.
- We can use methods other than least squares method that are less affected by outliers.
- We can study the uncertainty (e.g., variance, MSE, or CI) of the fitted estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Simple Linear Regression Model

The simple linear regression models is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n,\ \text{with } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

More info:

- $E(\widehat{\beta}_1) = \beta_1,\ \mathsf{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{SXX}.$
- $E(\widehat{\beta}_0) = \beta_0,\ \mathsf{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right).$
- Residuals (e_i) with $E(e_i) = 0$ and $\mathsf{Var}(e_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the leverage of i^{th} data point.
- Standardized residuals – sometimes studentized residuals are used for diagnostics.
- Leverage and Influence points

Two types of Bootstrap regression

We can treat the predictors as random (potentially changing from sample to sample) or we can treat them as fixed.

- Random x resampling
  - it is also called **observation resampling** or **case resampling**.
  - Resample observations as with correlation example.
- Fixed x resampling
  - it is also called is called **model based resampling** or **residual resampling**.
  - Resample residuals as follows
    - Fit a model and compute residuals
    - Generate the bootstrap data by $Y* = (\text{Fit}) +$ Bootstrap sample of OLS residuals.

Random x resampling

Assume that we want to fit a regression model with response variable y and predictors $x_1, x_2, \ldots, x_p$. We have a sample of n observations $z_i' = (y_i, x_{i1}, \ldots, x_{ik}),\ i = 1, \ldots, n.$

In random-x resampling, we simply select R bootstrap samples of the z_i', fitting the model and saving the coefficients from each bootstrap sample.

We can then construct confidence intervals for the regression coefficients.

Algorithm for Random x resampling (observation resampling)

For $b = 1, \ldots, B$,

- sample $i_1^*, i_2^*, \ldots, i_n^*$ randomly with replacement from $\{1, 2, \ldots, n\}$
- for $j = 1, 2, \ldots, n$, set $\mathbf{x}_j^* = \mathbf{x}_{i_j^*}, y_j^* = y_{i_j^*}$
- fit least squares regression to $(\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_n^*, y_n^*)$, giving estimates $\hat{\beta}_0^*(b)$, $\hat{\beta}_1^*(b)$, $\hat{\sigma}^2(b)$.

Example

Data set: geyser

A version of the eruptions data from the ‘Old Faithful’ geyser in Yellowstone National Park, Wyoming. This version comes from Azzalini and Bowman (1990) and is of continuous measurement from August 1 to August 15, 1985.
A data frame with 299 observations on 2 variables

- **duration** numeric Eruption time in mins
- **waiting** numeric Waiting time for this eruption

Example

Let's to the regression of **waiting** on **duration**

```
library(MASS)
data(geyser)
geyser.lm <- lm(waiting ~ duration, data = geyser)
summary(geyser.lm)
```

```
##
## Call:
## lm(formula = waiting ~ duration, data = geyser)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.5084  -8.1683  -0.4892   7.5365  29.1416
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99.3099     1.9569   50.75   <2e-16 ***
## duration     -7.8003     0.5368  -14.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.64 on 297 degrees of freedom
## Multiple R-squared:  0.4155, Adjusted R-squared:  0.4136
## F-statistic: 211.2 on 1 and 297 DF,  p-value: < 2.2e-16
```

Example

Using the boot library it is much easier to perform the calculations in R.

```
library(boot)

boot.geyser <- function(data, indices){
data <- data[indices,] # select obs. in bootstrap sample
mod <- lm(waiting ~ duration, data = data)
coefficients(mod) # return coefficient vector
}

geyser.boot <- boot(geyser, boot.geyser, 5000)
geyser.boot
```

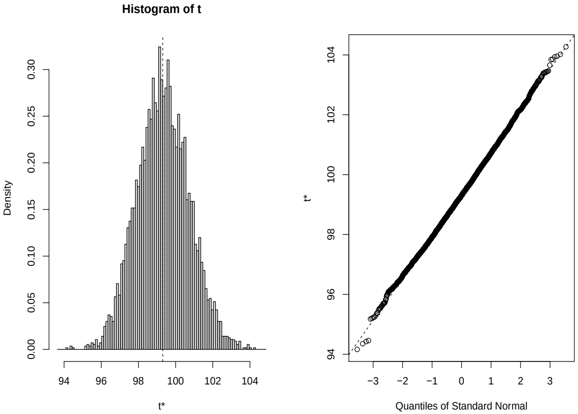
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = geyser, statistic = boot.geyser, R = 5000)
##
##
## Bootstrap Statistics :
##      original        bias    std. error
## t1* 99.309856  0.016130792    1.3969402
## t2* -7.800326 -0.001113983    0.4542609
```

```
names(geyser.boot)
```

```
##  [1] "t0"        "t"         "R"         "data"      "seed"      "statistic"
##  [7] "sim"       "call"      "stype"     "strata"    "weights"
```

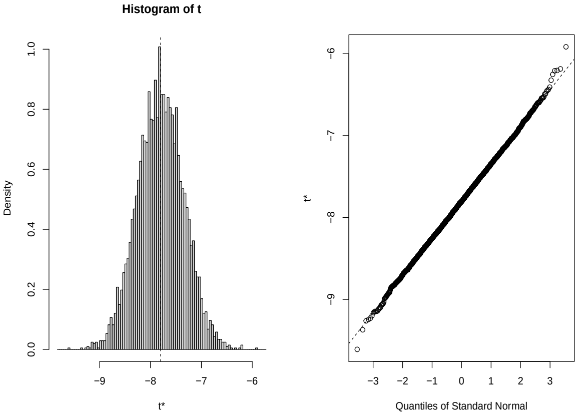
Example

```
plot(geyser.boot, index = 1)
```



Example

```
plot(geyser.boot, index = 2)
```



```r
confint(geyser.lm)
```

```
##                 2.5 %     97.5 %
## (Intercept) 95.458632 103.161080
## duration    -8.856725  -6.743927
```

```r
boot.ci(geyser.boot, index = 1, type = c("norm", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
## 
## CALL : 
## boot.ci(boot.out = geyser.boot, type = c("norm", "perc", "bca"), 
##     index = 1)
## 
## Intervals : 
## Level      Normal             Percentile            BCa          
## 95%   ( 96.56, 102.03 )   ( 96.67, 102.14 )   ( 96.69, 102.14 )  
## Calculations and Intervals on Original Scale
```

```r
boot.ci(geyser.boot, index = 2, type = c("norm", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
## 
## CALL : 
## boot.ci(boot.out = geyser.boot, type = c("norm", "perc", "bca"), 
##     index = 2)
## 
## Intervals : 
## Level      Normal             Percentile            BCa          
## 95%   (-8.690, -6.909 )   (-8.678, -6.909 )   (-8.670, -6.901 )  
## Calculations and Intervals on Original Scale
```

Fixed x resampling (Residual resampling)

If the observations in the Geyser data set are meant to represent a larger ‘population’ of eruptions. If it makes sense to think that the x values used in the study are fixed with replication of the study. And the response are random because of the error component in the model.

If the values of the predictor are set by the experimenter, then Fixed x sampling is the way to go.

Fixed x resampling

How can we generate bootstrap replications when the model-matrix $\boldsymbol{X}$ is fixed?

- One way to proceed is to treat the fitted values $\hat{y}_i$ from the model as giving the expectation of the response for the bootstrap samples.
- Attaching a random error to each $\hat{y}_i$ produces a fixed x bootstrap sample.
- The errors could be generated parametrically from a normal distribution with mean 0 and variance $\hat{\sigma}^2$ (the estimated error variance in the regression), if we are willing to assume that the errors are normally distributed, or nonparametrically, by resampling residuals from the original regression.
- We would then regress the bootstrapped values $y^*(b)$ on the fixed X matrix to obtain bootstrap replications of the regression coefficients.

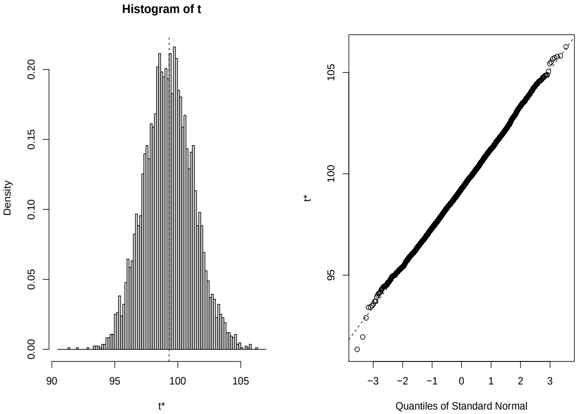
Example

```
fit <- fitted(geyser.lm)
e <- residuals(geyser.lm)
X <- model.matrix(geyser.lm)

boot.geyser.fixed = function(data, indices) {
y_b <-  fit + e[indices]
mod <- lm(y_b ~ X - 1)
coefficients(mod)
}

geyser.fixed.boot <- boot(geyser, boot.geyser.fixed, 5000)

geyser.fixed.boot
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = geyser, statistic = boot.geyser.fixed, R = 5000)
##
##
## Bootstrap Statistics :
##      original        bias    std. error
## t1* 99.309856 -0.030264487    1.9507000
## t2* -7.800326  0.006958766    0.5321568
```

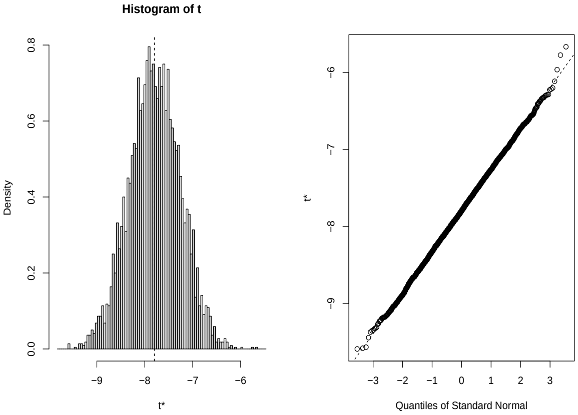
Example

```
plot(geyser.fixed.boot, index = 1)
```



Example

```
plot(geyser.fixed.boot, index = 2)
```



Calculating Confidence Intervals

```
boot.ci(geyser.fixed.boot, index = 1, type = c("norm", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = geyser.fixed.boot, type = c("norm", "perc",
##     "bca"), index = 1)
##
## Intervals :
## Level      Normal              Percentile            BCa
## 95%   ( 95.52, 103.16 )   ( 95.47, 103.28 )   ( 95.59, 103.37 )
## Calculations and Intervals on Original Scale
```

```
boot.ci(geyser.fixed.boot, index = 2, type = c("norm", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 5000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = geyser.fixed.boot, type = c("norm", "perc",
##     "bca"), index = 2)
##
## Intervals :
## Level      Normal              Percentile            BCa
## 95%   (-8.850, -6.764 )   (-8.864, -6.745 )   (-8.864, -6.745 )
## Calculations and Intervals on Original Scale
```

Comparison

Two results should be close, but can be rather different in cases of outliers.

1. Resampling observations

- Allows predictors to vary over samples
- Robust to model specification

2. Resampling residuals

- Fixes the design, as might be needed for certain problems
- closely mimics classical OLS results, but requires model to hold