Homework 9

Xupeng Chen

2017/12/05

1

(i) paired bootsrap

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

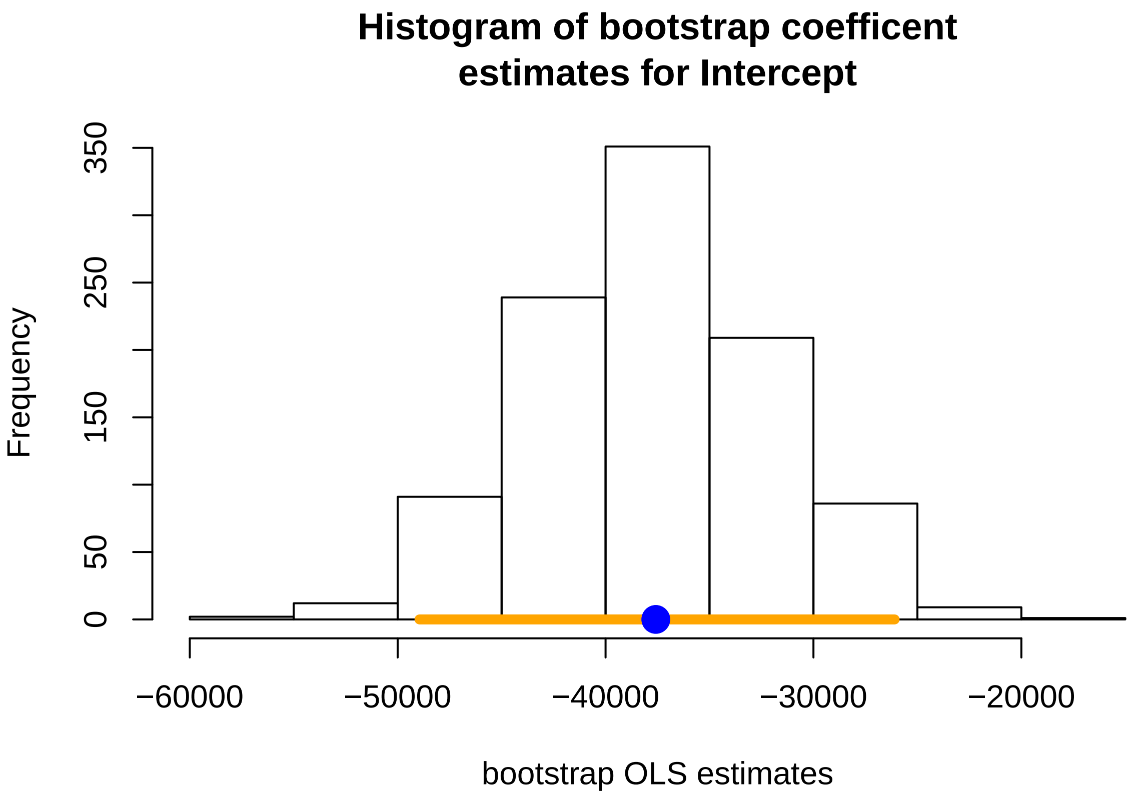
```r
library(boot)
library(bootstrap)
data <-read.csv('data.csv')

target<-select(data,Gr_Liv_Area,Central_Air,SalePrice)
Central_Air <- target$Central_Air == "Y"
target$Central_Air <- as.numeric(Central_Air)
fit <- lm(SalePrice ~ . , data = target)
beta_hat <- fit$coefficients
n <- dim(target)[1]
p <- dim(target)[2] - 1

set.seed(1111)
boot_house <- target[sample(1:n, replace = TRUE), ]
boot_fit <- lm(SalePrice ~ ., data = boot_house)
boot_beta <- boot_fit$coefficients
B <- 1000
boot_beta <- matrix(0, B, p + 1) # there are 8 predictors and 1 for intercept term
variable.names <- c("Intercept", colnames(target)[1:2])
colnames(boot_beta) <- variable.names
for(b in 1:B){
boot_house <- target[sample(1:n, replace = TRUE), ]
boot_fit <- lm(SalePrice ~ ., data = boot_house)
boot_beta[b, ] <- boot_fit$coefficients
}
par(mfrow = c(1, 1))
k <- 1
hist(boot_beta[, k], main = paste("Histogram of bootstrap coefficent
estimates for ", variable.names[k], sep = ""),
xlab = "bootstrap OLS estimates")
```

```r
CI <- quantile(boot_beta[, k], probs = c(0.025, 0.975))

segments(CI[1], 0, CI[2], 0, lwd = 5, col = "orange")
points(beta_hat[k], 0, col = "blue", cex = 2, pch = 16)
```




```r
# CI for each regression coefficients
apply(boot_beta, 2, quantile, probs = c(0.025, 0.975))
```

```
##       Intercept Gr_Liv_Area Central_Air
## 2.5%  -48948.30    105.0432    44572.31
## 97.5% -26097.19    122.1410    58703.12
```

residual bootstrap

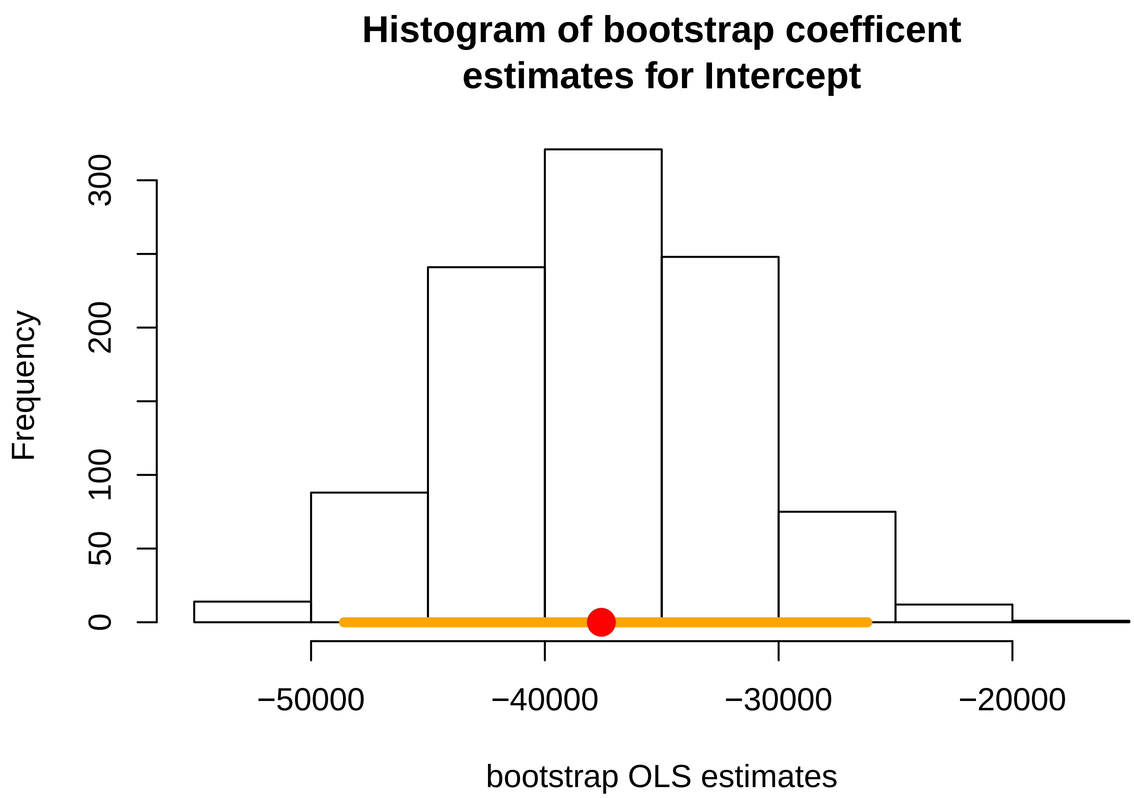
```r
set.seed(1111)
fit <- lm(SalePrice ~ . , data = target)
residuals <- fit$residuals
mean(residuals)
```

```
## [1] 2.121291e-12
```

```r
central_res <- residuals - mean(residuals)
boot_res <- sample(central_res, replace = TRUE)
boot_response <- fit$fitted.values + boot_res
boot_sample <- target
boot_sample[, 1] <- boot_response
boot_fit <- lm(SalePrice ~ ., data = boot_sample)
boot_beta_res <- boot_fit$coefficients

B <- 1000
boot_beta_res <- matrix(0, B, p + 1) # there are 8 predictors and 1 for intercept term
```

```r
variable.names <- c("Intercept", colnames(target)[1:2])
colnames(boot_beta_res) <- variable.names
for(b in 1:B){
boot_res <- sample(central_res, replace = TRUE)
boot_response <- fit$fitted.values + boot_res
boot_sample <- target
boot_sample[, 3] <- boot_response
boot_fit <- lm(SalePrice ~ ., data = boot_sample)
boot_beta_res[b, ] <- boot_fit$coefficients
}
par(mfrow = c(1, 1))
k <- 1
hist(boot_beta_res[, k], main = paste("Histogram of bootstrap coefficent
estimates for ", variable.names[k], sep = ""),
xlab = "bootstrap OLS estimates")
CI <- quantile(boot_beta_res[, k], probs = c(0.025, 0.975))
segments(CI[1], 0, CI[2], 0, lwd = 5, col = "orange")
points(beta_hat[k], 0, col = "red", cex = 2, pch = 16)
```




```r
# CI for each regression coefficients
apply(boot_beta_res, 2, quantile, probs = c(0.025, 0.975))
```

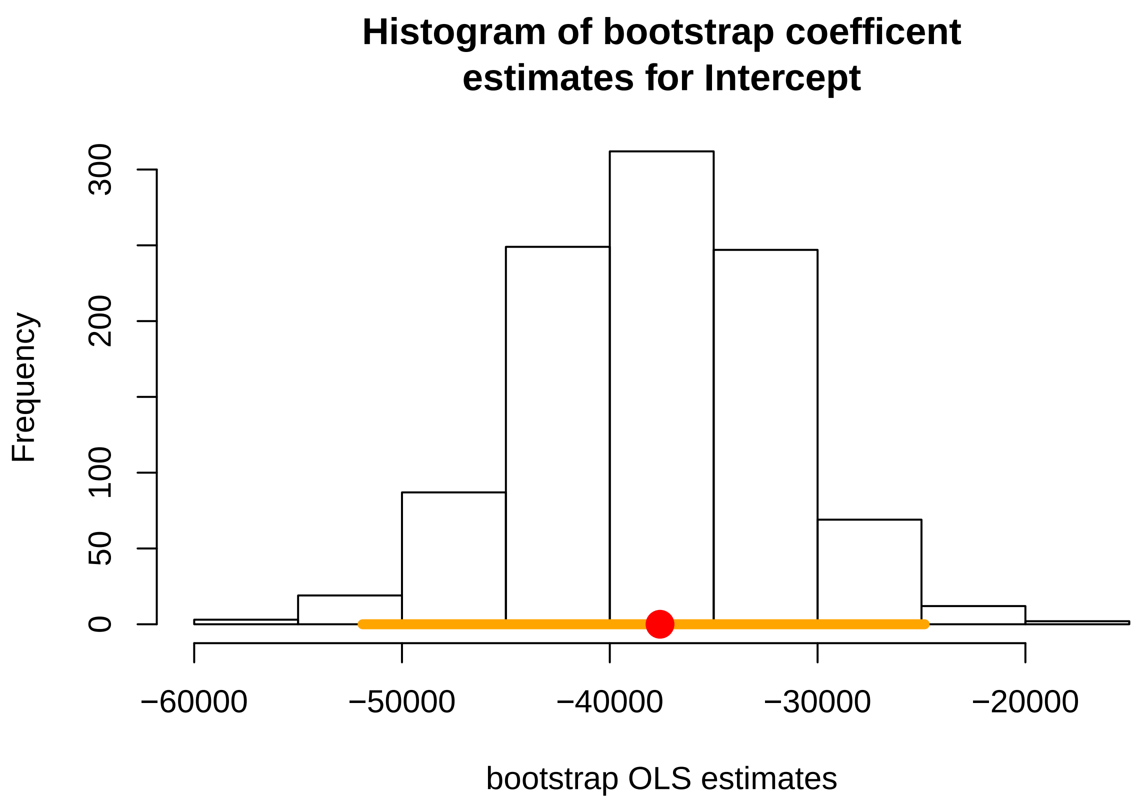
```
##       Intercept Gr_Liv_Area Central_Air
## 2.5%  -48596.48    108.3915    41376.45
## 97.5% -26205.72    119.4049    61420.32
```

(ii)

$$P(A \cap B) = 1 - P(A^c \cup B^c) \geq 1 - P(A^c) - P(B^c) A = \{\beta_1 \in [m, n]\}, B = \{\beta_2 \in [s, t]\} \text{we can have: } P(A^c) + P(B^c) \leq 0.05 \text{to}$$

use paired bootstrap

```
data<-read.csv("data.csv")
library(dplyr)
set.seed(1111)
target<-select(data,Gr_Liv_Area,Central_Air,SalePrice)
Central_Air <- target$Central_Air == "Y"
target$Central_Air <- as.numeric(Central_Air)
fit <- lm(SalePrice ~ . , data = target)
beta_hat <- fit$coefficients
n <- dim(target)[1]
p <- dim(target)[2] - 1
set.seed(0)
boot_house <- target[sample(1:n, replace = TRUE), ]
boot_fit <- lm(SalePrice ~ ., data = boot_house)
boot_beta <- boot_fit$coefficients
B <- 1000
boot_beta <- matrix(0, B, p + 1) # there are 8 predictors and 1 for intercept term
variable.names <- c("Intercept", colnames(target)[1:2])
colnames(boot_beta) <- variable.names
for(b in 1:B){
boot_house <- target[sample(1:n, replace = TRUE), ]
boot_fit <- lm(SalePrice ~ ., data = boot_house)
boot_beta[b, ] <- boot_fit$coefficients
}
par(mfrow = c(1, 1))
k <- 1
hist(boot_beta[, k], main = paste("Histogram of bootstrap coefficent
estimates for ", variable.names[k], sep = ""),
xlab = "bootstrap OLS estimates")
CI <- quantile(boot_beta[, k], probs = c(0.0125, 0.9875))
segments(CI[1], 0, CI[2], 0, lwd = 5, col = "orange")
points(beta_hat[k], 0, col = "red", cex = 2, pch = 16)
```




```
# CI for each regression coefficients
apply(boot_beta, 2, quantile, probs = c(0.0125, 0.9875))
```

```
##         Intercept Gr_Liv_Area Central_Air
## 1.25%   -51896.13    105.0305    43782.71
## 98.75%  -24847.29    122.6561    59651.90
```

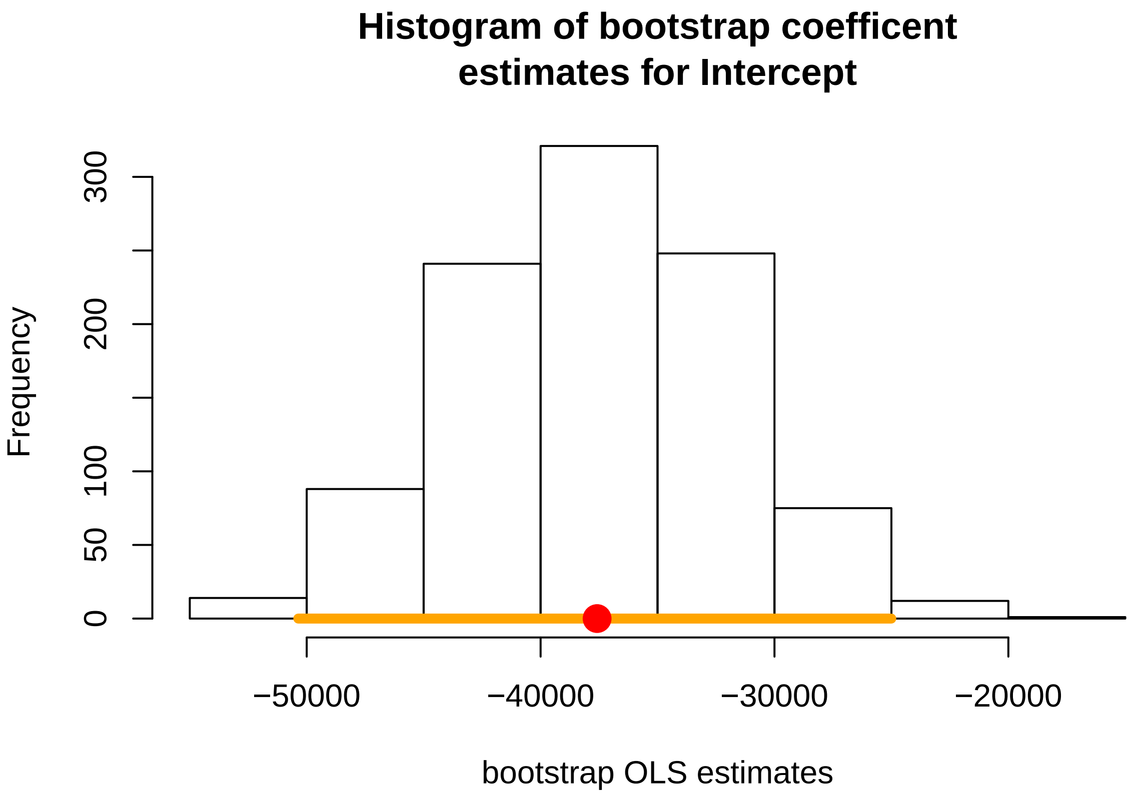
use residual bootstrap:

```
set.seed(1111)
library(dplyr)
fit <- lm(SalePrice ~ . , data = target)
residuals <- fit$residuals
mean(residuals)
```

```
## [1] 2.121291e-12
```

```
set.seed(1111)
central_res <- residuals - mean(residuals)
boot_res <- sample(central_res, replace = TRUE)
boot_response <- fit$fitted.values + boot_res
boot_sample <- target
boot_sample[, 1] <- boot_response
boot_fit <- lm(SalePrice ~ ., data = boot_sample)
boot_beta_res <- boot_fit$coefficients
B <- 1000
boot_beta_res <- matrix(0, B, p + 1) # there are 8 predictors and 1 for intercept term
variable.names <- c("Intercept", colnames(target)[1:2])
colnames(boot_beta_res) <- variable.names
for(b in 1:B){
boot_res <- sample(central_res, replace = TRUE)
```

```r
boot_response <- fit$fitted.values + boot_res
boot_sample <- target
boot_sample[, 3] <- boot_response
boot_fit <- lm(SalePrice ~ ., data = boot_sample)
boot_beta_res[b, ] <- boot_fit$coefficients
}
par(mfrow = c(1, 1))
k <- 1
hist(boot_beta_res[, k], main = paste("Histogram of bootstrap coefficent
estimates for ", variable.names[k], sep = ""),
xlab = "bootstrap OLS estimates")
CI <- quantile(boot_beta_res[, k], probs = c(0.0125, 0.9875))
segments(CI[1], 0, CI[2], 0, lwd = 5, col = "orange")
points(beta_hat[k], 0, col = "red", cex = 2, pch = 16)
```




```r
# CI for each regression coefficients
apply(boot_beta_res, 2, quantile, probs = c(0.0125, 0.9875))
```

```
##        Intercept Gr_Liv_Area Central_Air
## 1.25%  -50359.84    107.7793    40217.04
## 98.75% -25006.42    120.1388    62444.63
```

so the confidence region is:
$[105.0305, 122.6561] \times [40217.04, 62444.63]$

(iii) from the normal distribution assumption $\epsilon \sim N(0, \sigma^2)$
we can calculate from the above that the CI of ϵ is $[2.6035 \times 10^9, 3.0002 \times 10^9]$

2

(i)It is a composite hypothesis.
reason: the sample is in the distribution of B(n, p), H1 is p does not equal to 0.5. So it has many potential value(and many potential expectation).

(ii)It depends. if the parameter θ can be solely determined by the expectation of the population distribution, it is a simple hypothesis. Otherwise it is a composite distribution.