# DVM Bootstrap - Work 4

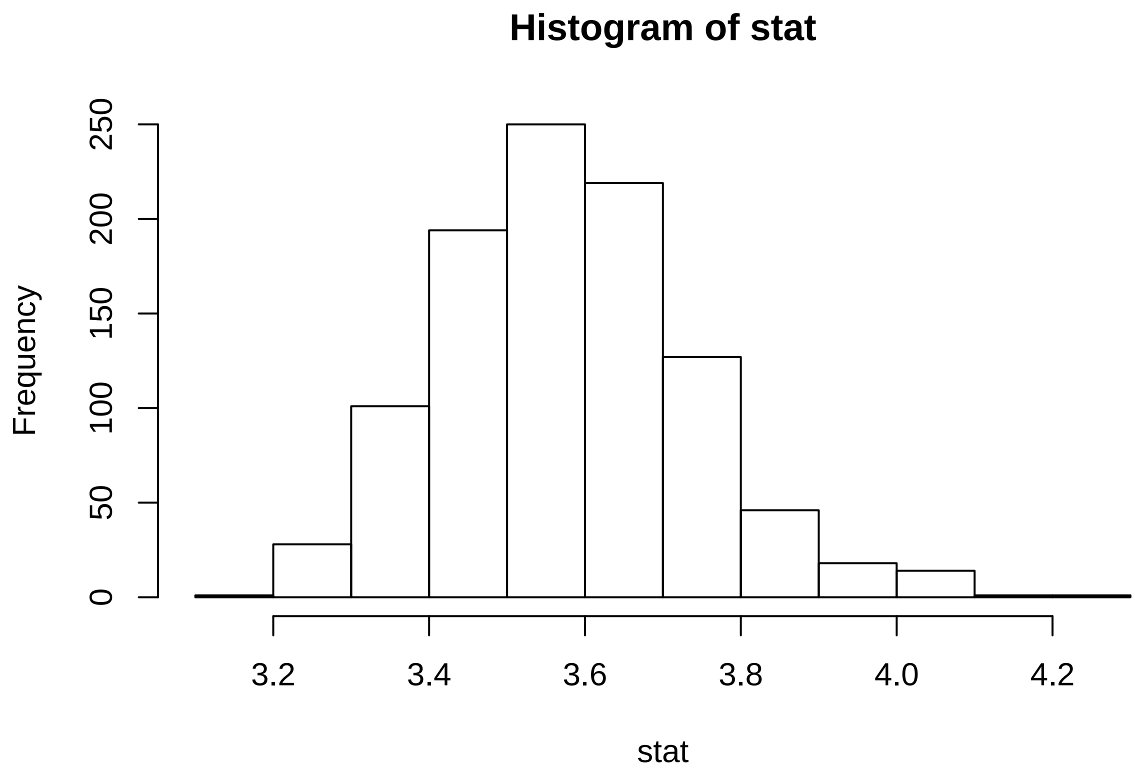
*Jeremy Williams and Spyridoula Chrysikopoulou-Soldatou*

*November 15, 2017*

## Question 1 - CI based on nonparametric boot-strap

### Calculate 95% CI based on nonparametric boot-strap for the population trimmed mean (25%), standard deviation and coefficient of variation.

```
index=c(3.65, 4.03, 4.58, 4.61, 4.70, 4.85, 3.21, 3.93, 3.15, 3.00, 2.93,
        3.56, 4.13, 3.68, 3.88, 3.25, 3.92, 3.99, 3.04, 3.10, 3.20, 3.35,
        3.19, 3.10)

n = length(index)

#Set number of bootstrap samples
nsim=1000

#create a vector in which to store the results
stat = numeric(nsim)

#Set up a loop to generate a series of bootstrap samples
for (i in 1:nsim){
  index1= sample(index, n, replace=T)
  stat[i] = mean(index1, trim = 0.25)}

par(mfrow=c(1,1))
hist(stat)
```




```r
#95% Confidence Interval:
quantile(stat,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 3.296396 3.930083
```

```r
#Bootstrap estimate:
mean(stat)
```

```
## [1] 3.578645
```

```r
#Estimated standard error:
sd(stat)
```

```
## [1] 0.1606031
```

```r
#Coefficients of Variation
sd(stat)/mean(stat)
```

```
## [1] 0.0448782
```

### Question 2 - Bootstrap CI (three methods)

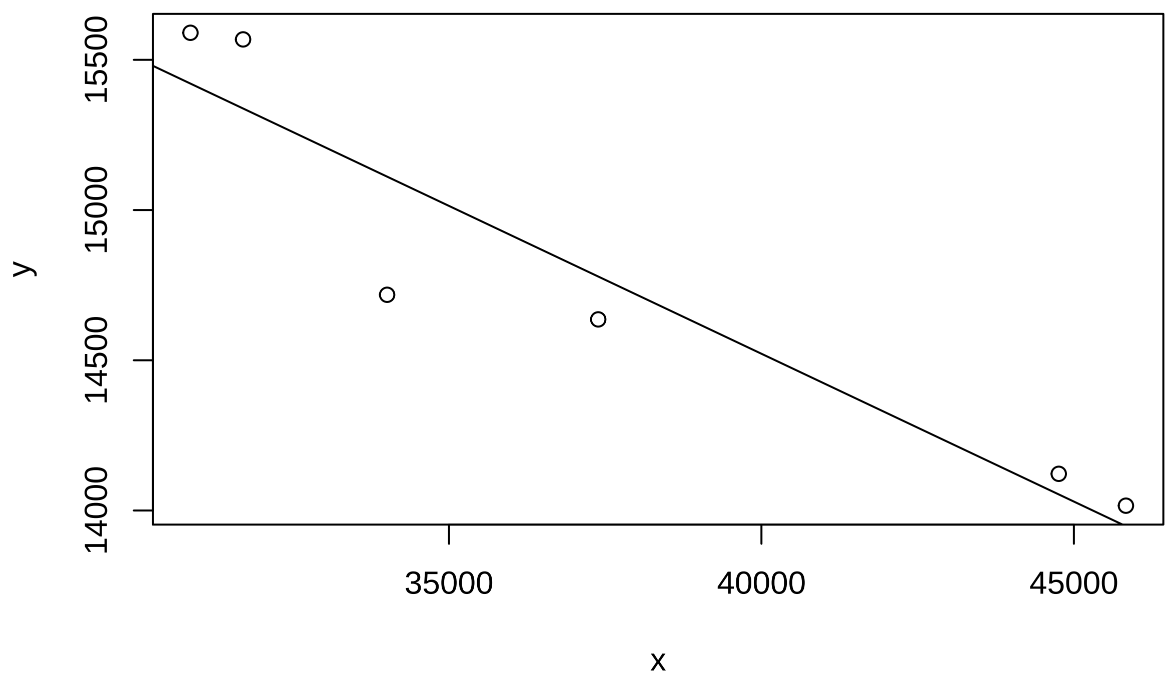
**Calculate 95% bootstrap CI (three methods) of the second-hand price for a car (same model and year) with 50000 km.**

```r
y<-c(14636, 14122, 14016, 15590, 15568, 14718)
x<-c(37388,44758,45833,30862,31705,34010)
```

```r
data2 <- data.frame(x,y)
data2
```

```
##       x     y
## 1 37388 14636
## 2 44758 14122
## 3 45833 14016
## 4 30862 15590
## 5 31705 15568
## 6 34010 14718
```

```r
#Visually estimate some starting parameter values
par(mfrow=c(1,1))
plot(x,y)
abline(lm(y ~ x,data2))
```




```r
#Model
cars<-lm(y ~ x, data2)

#Summary of the Model
summary(cars)
```

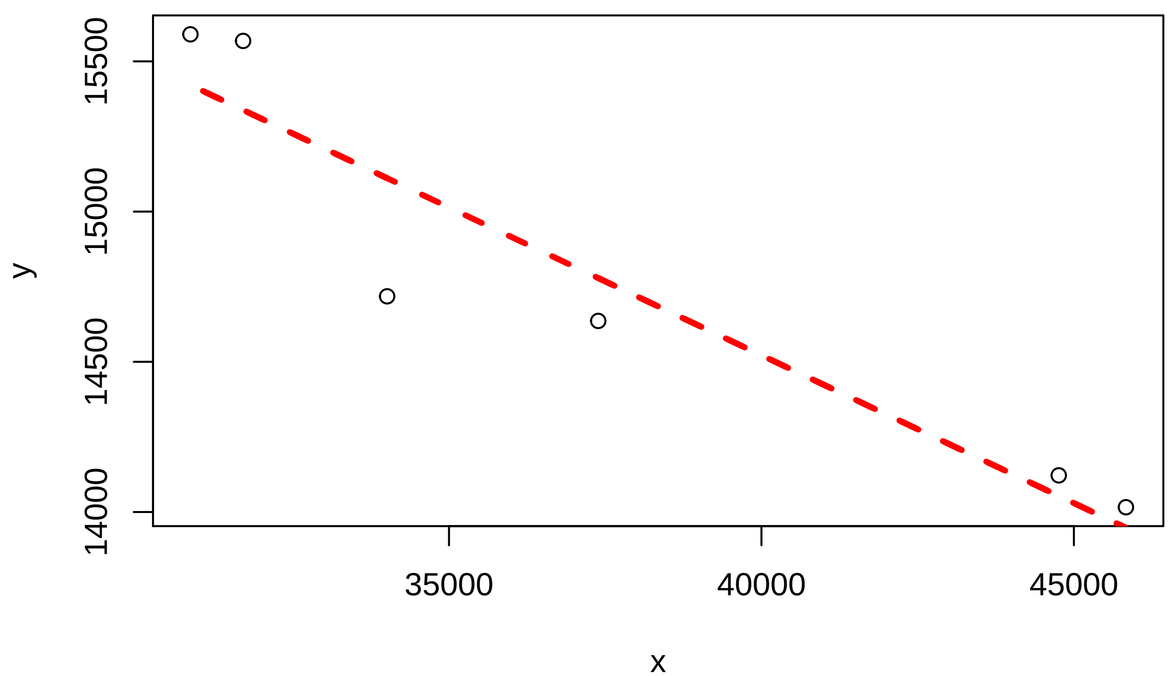
```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
```

```
##       1       2       3       4       5       6
## -142.74   68.54   68.33  169.04  230.00 -393.17
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.846e+04  6.710e+02  27.507 1.04e-05 ***
## x           -9.841e-02  1.771e-02  -5.558  0.00513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.8 on 4 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8567
## F-statistic: 30.89 on 1 and 4 DF,  p-value: 0.005132
```

```r
#Estimation of goodness of fit
cor(y,predict(cars))
```

```
## [1] 0.9409255
```

```r
#Plot the fit
par(mfrow=c(1,1))
plot(x,y)
lines(x,predict(cars),lty=2,col="red",lwd=3)
```



### Predicated Value of Model

```r
#install.packages("tidyverse")
suppressMessages(suppressWarnings(library(tidyverse)))

cars <- data.frame(cbind(km = c(37388, 44758, 45833, 30862, 31705, 34010),
        price = c(14636, 14122, 14016, 15590, 15568, 14718)))
cars
```

```
##      km price
## 1 37388 14636
## 2 44758 14122
## 3 45833 14016
## 4 30862 15590
## 5 31705 15568
## 6 34010 14718
```

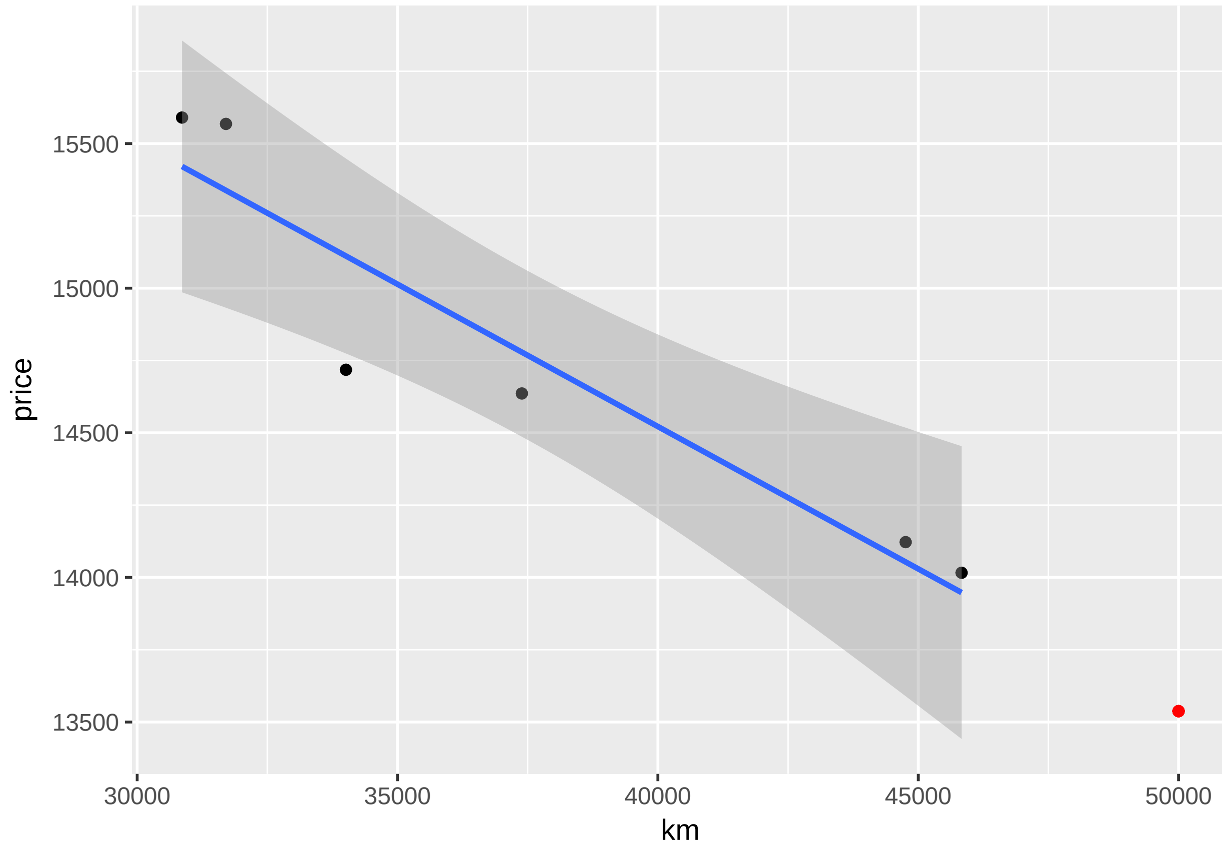
```r
lm.model <- lm(price ~ km, cars)
summary(lm.model)
```

```
##
## Call:
## lm(formula = price ~ km, data = cars)
##
## Residuals:
##       1       2       3       4       5       6
## -142.74   68.54   68.33  169.04  230.00 -393.17
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.846e+04  6.710e+02  27.507 1.04e-05 ***
## km          -9.841e-02  1.771e-02  -5.558  0.00513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.8 on 4 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8567
## F-statistic: 30.89 on 1 and 4 DF,  p-value: 0.005132
```

```r
new.car <- 50000
predicted.price <- coef(lm.model)[[1]] + new.car * coef(lm.model)[[2]]
predicted.price
```

```
## [1] 13537.6
```

```r
#Visualation of the Predicated Value
ggplot(cars) +
  geom_point(aes(x = km, y = price)) +
  geom_point(aes(x = new.car, y = predicted.price), colour = "red") +
  geom_smooth(aes(x = km, y = price), method = lm)
```




**Method I: bootstraping the pairs (x,y)**

```
y<-c(14636, 14122, 14016, 15590, 15568, 14718)
x<-c(37388,44758,45833,30862,31705,34010)

data2 <- data.frame(x,y)
data2
```

```
##       x     y
## 1 37388 14636
## 2 44758 14122
## 3 45833 14016
## 4 30862 15590
## 5 31705 15568
## 6 34010 14718
```

```
#Model
cars<-lm(y ~ x, data2)

#Summary of the Model
summary(cars)
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
```

```
## Residuals:
##       1       2       3       4       5       6
## -142.74   68.54   68.33  169.04  230.00 -393.17
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.846e+04  6.710e+02  27.507 1.04e-05 ***
## x           -9.841e-02  1.771e-02  -5.558  0.00513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.8 on 4 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8567
## F-statistic: 30.89 on 1 and 4 DF,  p-value: 0.005132
```

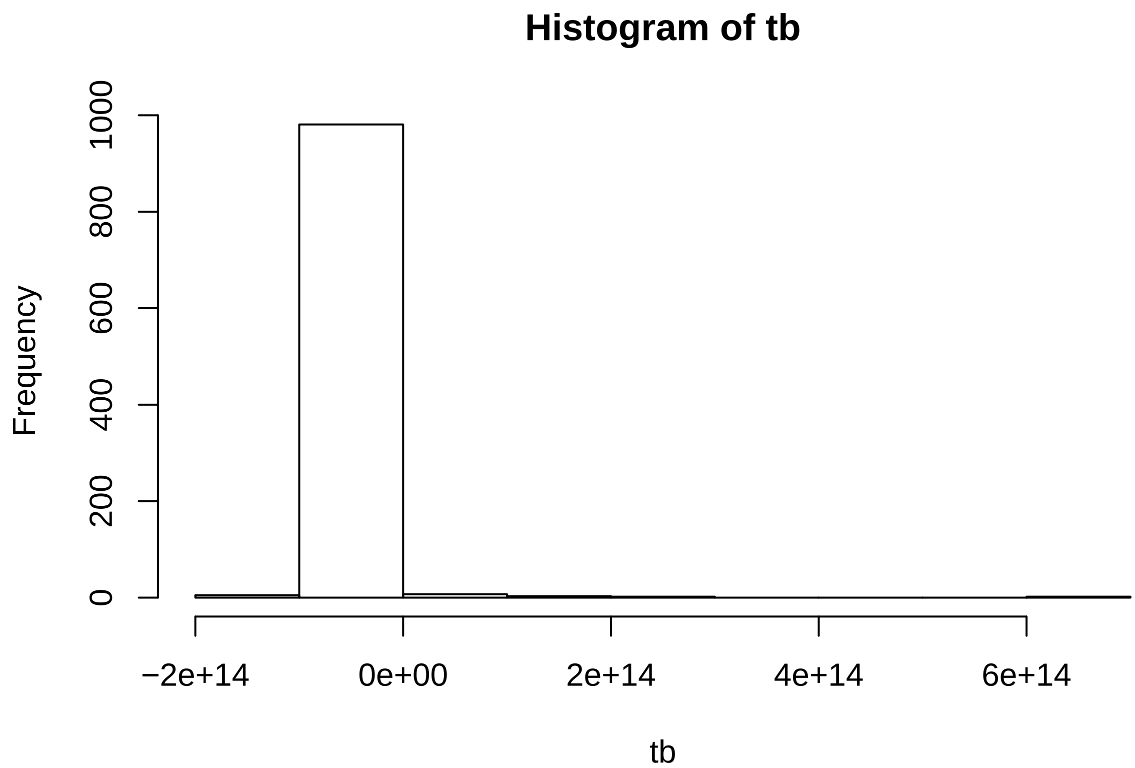
```r
#Bootstrap sample process
n<-length(x); data<-data2
theta<-summary(cars)$coefficients[1]
sdtheta<-summary(cars)$coefficients[3]
nb<-1000; z<-seq(1,n);tb<-numeric(nb);predb<-numeric(nb)
thetab<-numeric(nb)
suppressWarnings(for(i in 1:nb){
  zb<-sample(z,n,replace=T)
  carsb<-lm(data[zb,2] ~ data[zb,1])
  data[zb,2]
  data[zb,1]
  thetab[i]<-summary(carsb)$coefficients[1]
  sdthetab<-summary(carsb)$coefficients[3]
  tb[i]<-(thetab[i]-theta)/sdthetab
  predb[i]<-summary(carsb)$coefficients[1] + summary(carsb)$coefficients[2]*50000})

#The Bootstrap-t 95% CI of price :
theta+sdtheta*quantile(tb,0.025)
```

```
##     2.5%
## 5857.963
```

```r
theta+sdtheta*quantile(tb,0.975)
```

```
##    97.5%
## 51740.98
```

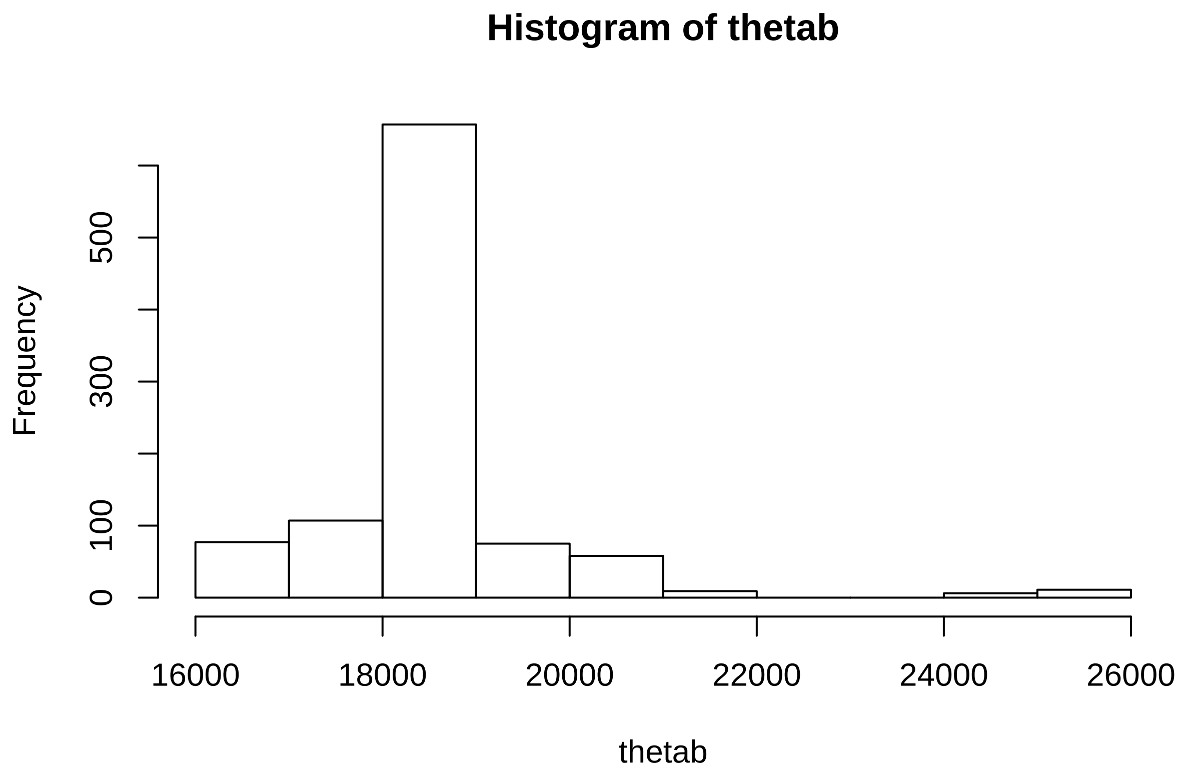
```r
par(mfrow=c(1,1))
hist(tb)
```




```r
#The CI based in the quantile method:
quantile(thetab,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 16724.81 21063.15
```

```r
par(mfrow=c(1,1))
hist(thetab)
```



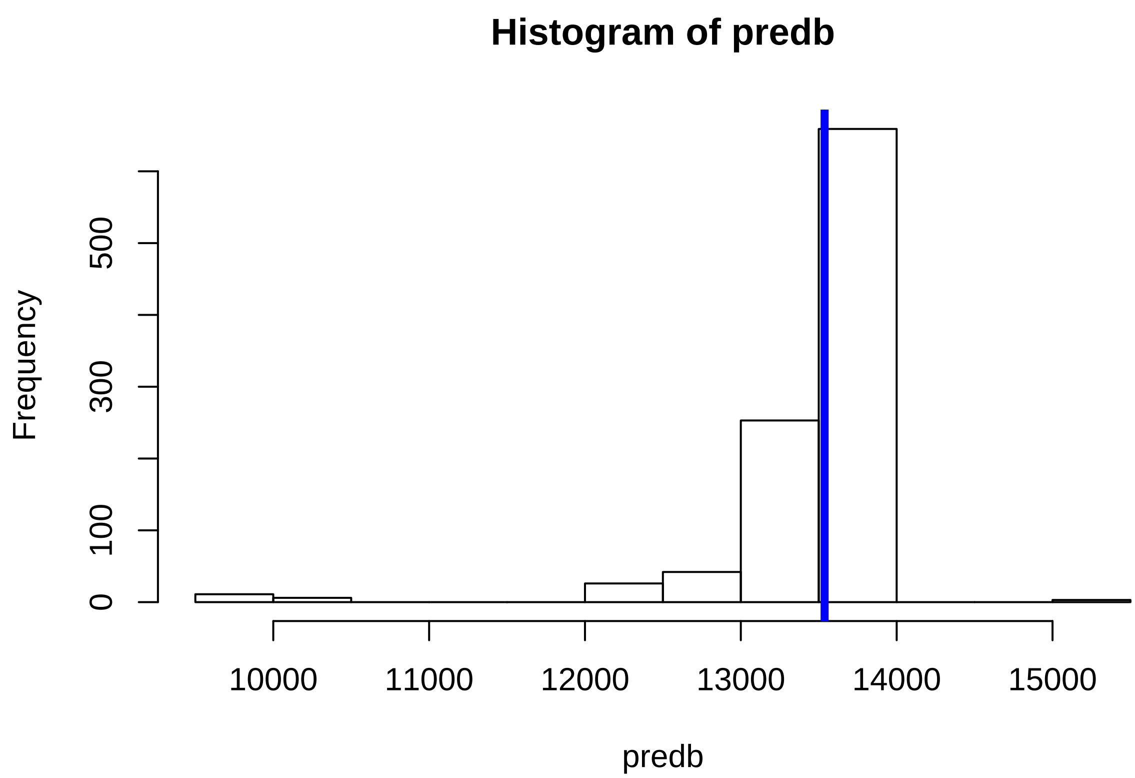

```r
#Predicted Price from the model
pred1<-summary(cars)$coefficients[1] + summary(cars)$coefficients[2]*50000
pred1
```

```
## [1] 13537.6
```

```r
#The quantile 95% CI of the prediction at x=50000:
quantile(predb,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 12198.84 13807.35
```

```r
par(mfrow=c(1,1))
hist(predb)
abline(v=pred1, lwd=4, col="blue")
```




**Method II: bootstraping the residuals**

```
y<-c(14636, 14122, 14016, 15590, 15568, 14718)
x<-c(37388,44758,45833,30862,31705,34010)

data2 <- data.frame(x,y)
data2
```

```
##       x     y
## 1 37388 14636
## 2 44758 14122
## 3 45833 14016
## 4 30862 15590
## 5 31705 15568
## 6 34010 14718
```

```
#Model
cars<-lm(y ~ x, data2)

#Residuals of the model
res<-residuals(cars)
res
```

```
##          1          2          3          4          5          6
## -142.73955   68.53705   68.32706  169.04130  230.00035 -393.16620
```

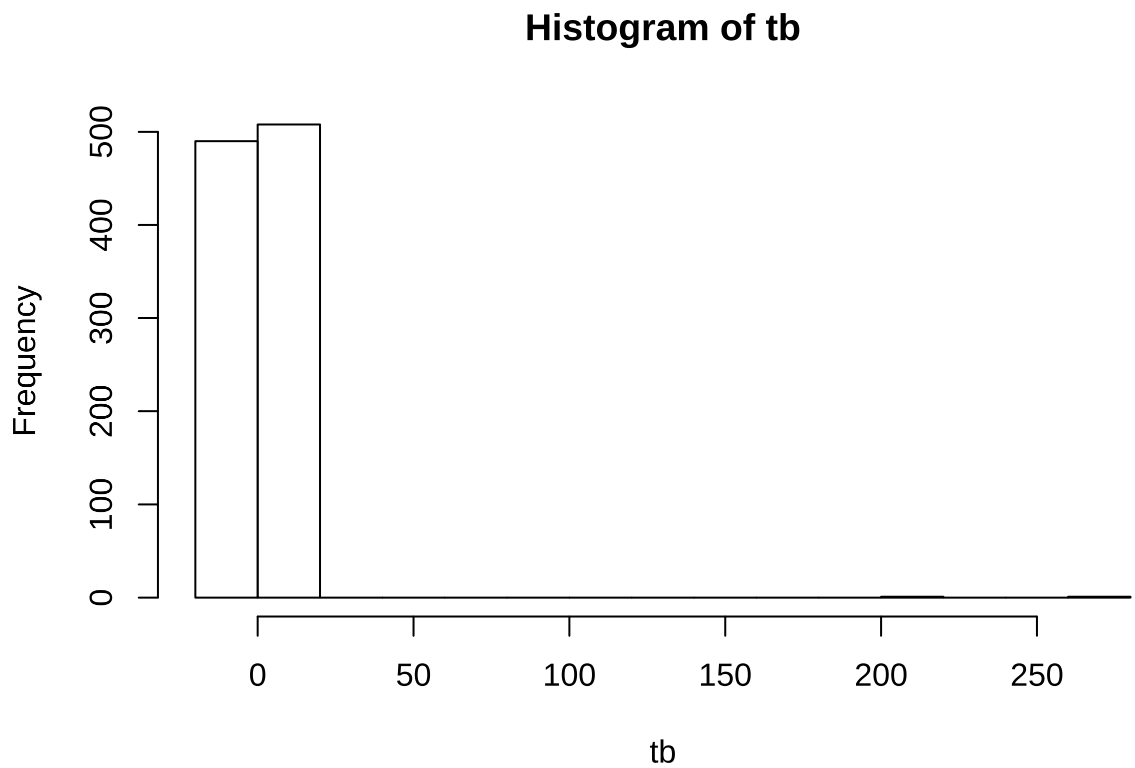
```r
#Bootstrap sample process
n<-length(x); c1<-summary(cars)$coefficients[1]
c2<-summary(cars)$coefficients[2]
sdtheta<-summary(cars)$coefficients[3]
nb<-1000; tb<-numeric(nb); predb<-numeric(nb)
thetab<-numeric(nb)
suppressWarnings(for(i in 1:nb){
  rb<-sample(res,n,replace=T)
  yb<-c1 + (c2*x) + rb
  carsb<-lm(yb ~ x, data2)
  thetab[i]<-summary(carsb)$coefficients[1]
  sdthetab<-summary(carsb)$coefficients[3]
  tb[i]<-(thetab[i]-c1)/sdthetab
  predb[i]<-summary(carsb)$coefficients[1] + summary(carsb)$coefficients[2]*50000})

#The bootstrap-t 95% CI of b1 :
c1+sdtheta*quantile(tb,0.025)
```

```
##     2.5%
## 16751.47
```

```r
c1+sdtheta*quantile(tb,0.975)
```

```
##    97.5%
## 20527.76
```

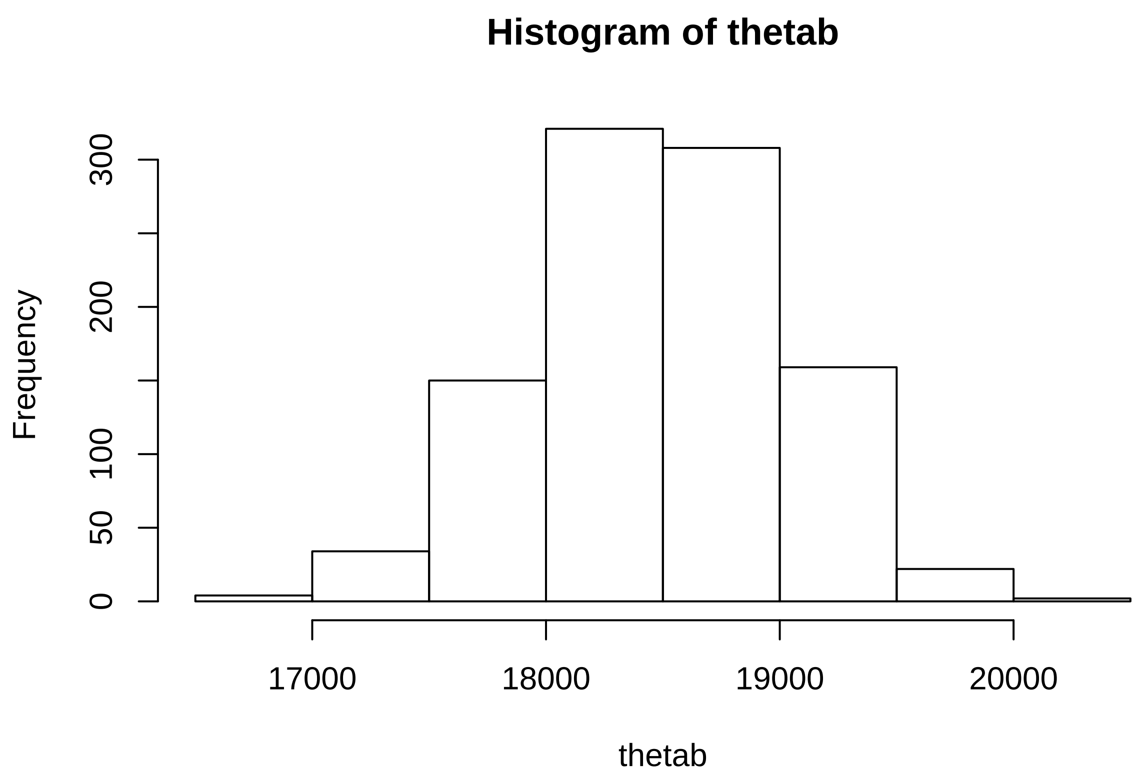
```r
par(mfrow=c(1,1))
hist(tb)
```




```
#The CI based in the quantile method,
quantile(thetab,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 17365.61 19495.23
```

```
par(mfrow=c(1,1))
hist(thetab)
```



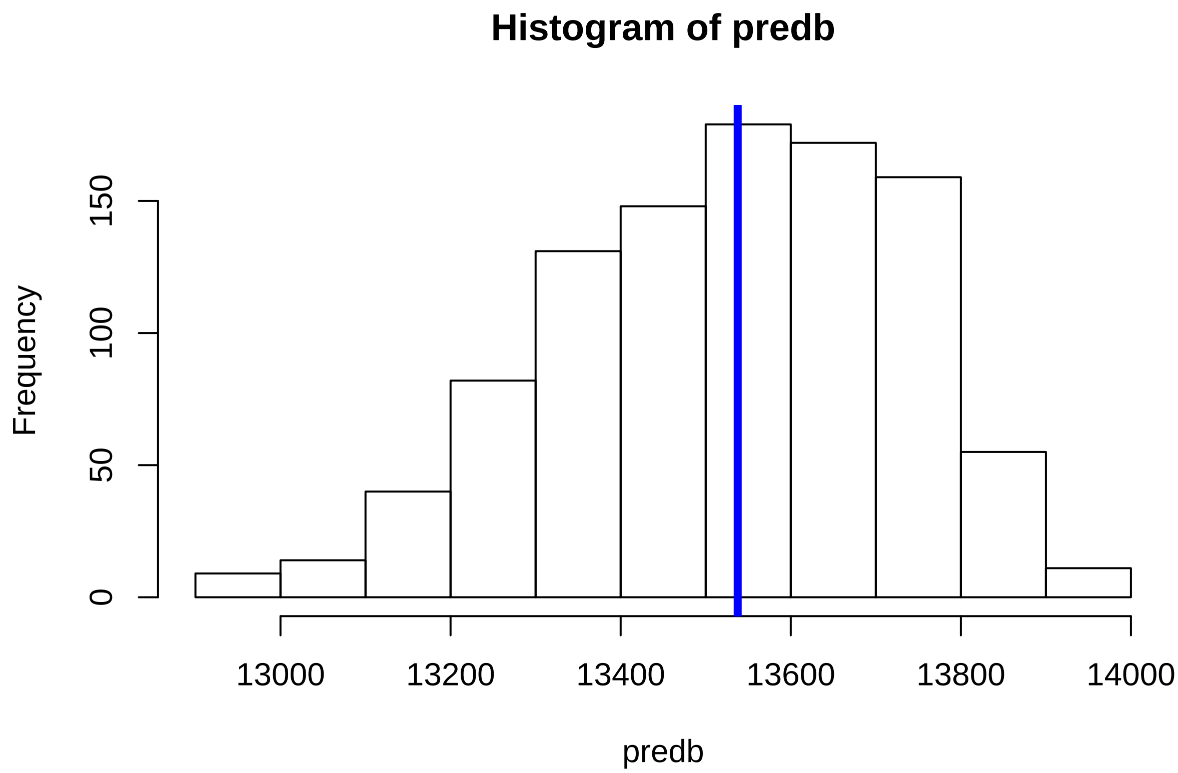

```
#Predicted Price from the model
pred1<-summary(cars)$coefficients[1] + summary(cars)$coefficients[2]*50000
pred1
```

```
## [1] 13537.6
```

```
#The quantile 95% CI of the prediction at x=50000:
quantile(predb,c(0.025,0.975))
```

```
##     2.5%    97.5%
## 13108.00 13857.05
```

```
par(mfrow=c(1,1))
hist(predb)
abline(v=pred1, lwd=4, col="blue")
```




**Looking at all predications, methods and bootstrapping processes, it can be said that "Method II: bootstraping the residuals" provides that best results with respect to Bootstrap CI methods.**