Computer Intensive Methods using R

Part 6: Cross validation, Jackknife and Bias estimation

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

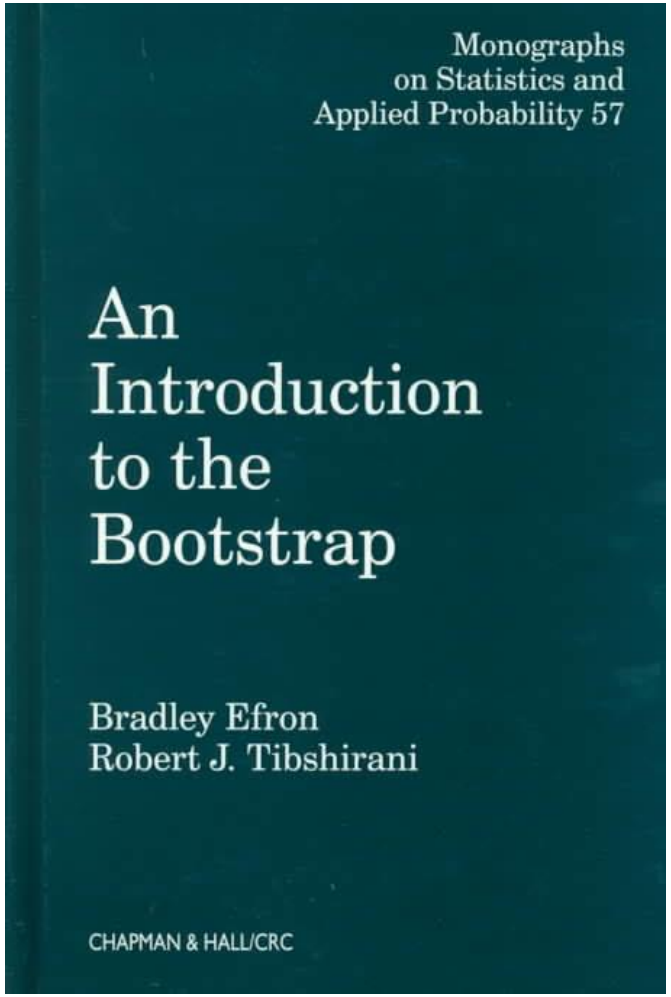
Overview of the course

- Selected topics:
  - Estimation of bias.
  - Cross-validation of prediction error.
  - The Jackknife.

Reference

- Bradley Efron and Robert J. Tibshirani (1994): An introduction to bootstrap.

- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.



Course materials

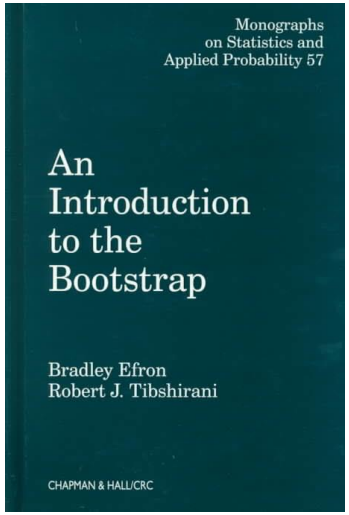
- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

YouTube tutorials

- YouTube tutorials about bootstrap using R:
  1. One-sample bootstrap CI for the mean (host: LawrenceStats): https://www.youtube.com/watch?v=ZkCDYAC2iFg.
  2. Using the non-parametric bootstrap for regression models in R (host:Ian Dworkin):https://www.youtube.com/watch?v=ydtOTctg5So.
  3. Performing the Non-parametric Bootstrap for statistical inference using R (host: Ian Dworkin): https://www.youtube.com/watch?v=TP6r5CTd9yM
  4. Using the sample function in R for resampling of data - absolute basics (host: Ian Dworkin):https://www.youtube.com/watch?v=xE3KGVT6VLE
  5. Permutation tests in R - the basics (host:Ian Dworkin):https://www.youtube.com/watch?v=ZiQdzwB12Pk.
  6. Bootstrap Sample Technique in R software (host: Sarveshwar Inani):https://www.youtube.com/watch?v=tb6wb9ZdPH0
  7. Bootstrap confidence intervals for a single proportion (host: LawrenceStats):https://www.youtube.com/watch?v=ubX4QEPqx5o
  8. Bootstrapped prediction intervals (host:James Scott):https://www.youtube.com/watch?v=c3gD_PwsCGM.

- https://www.youtube.com/watch?v=gcPIyeqymOU

Estimates for bias



Chapter 10

Topics

- Bias.
- Estimation of bias using bootstrap.
- Example:
  - The patch data.
  - Ratio statistic.

The probability distribution

Let X be a random variable such that

$$\boxed{X \sim F(\theta)}$$

F is the probability distribution of X.

θ is an unknown parameter to be estimated.

We assume that

$$\boxed{\theta = t(X_1, X_2, \ldots, X_N)}$$

The empirical distribution

The empirical distribution function is defined to be the discrete distribution that puts probability of 1/n on each value of x_i

$$F \rightarrow \left(x_1, x_2, \ldots, x_n\right)$$

$$P(A) = \hat{F} = \frac{\#(x_i \in A)}{n}$$

The plug-in principle

The plug-in estimate of the parameter

$$\theta = t(F)$$

is defined as

$$\hat{\theta} = t(\hat{F})$$

We use the same function from F, $t(F)$ on the empirical distribution

Bias

Bias: on average, how far is the statistic from the parameter ?

$$bias_F = bias_F(\hat{\theta}, \theta) = \boxed{E_F(\hat{\theta}) - \theta}$$

observed

unknown parameter

The bootstrap estimate for the bias

We would like to apply the bootstrap method in order to estimate the bias of a statistics.

This could be very useful if the distribution of the statistics in unknown.

$$bias_{\hat{F}} = bias_{\hat{F}}(\hat{\theta}, \hat{\theta}^*) = E_{\hat{F}}(\hat{\theta}^*) - \hat{\theta}$$

Observed statistics

The bootstrap algorithm

The observed data

$$x_1, x_2, \ldots\ldots\ldots, x_n$$

B bootstrap samples

$$x_1^*, x_2^*, \ldots\ldots\ldots, x_n^* \qquad x_1^*, x_2^*, \ldots\ldots\ldots, x_n^* \qquad x_1^*, x_2^*, \ldots\ldots\ldots, x_n^*$$

The bootstrap replicates

$$\theta_1^* \qquad \theta_b^* \qquad \theta_B^*$$

The bootstrap estimate for the bias

We first approximate the distribution of the statistics using bootstrap.

We estimate the expected value of the bootstrap replicates by

$$\boxed{\bar{\hat{\theta}}^* = \hat{\theta}^* = \frac{\sum_{b=1}^{B} \hat{\theta}_b^*}{B} = \hat{E}_{\hat{F}}(\hat{\theta}^*)}$$

The bootstrap estimate for the bias

The estimate for the bias

$$\boxed{\hat{b}_{\hat{F}} = \hat{\theta}^{*} - \hat{\theta}}$$

The mean of the bootstrap replicates

The observed statistics

Example: the patch data

- Eight subjects used medical patches design to decrease the level of a certain hormone in the blood.
- Each subject was measured three times, at baseline (using a placebo patch), using old patch and using new patch.
- In R:

```
> help(patch)
```

Example: the patch data

Bioequivalence study.

The FDA criterion for bioequivalence is that the expected value of the new patch match the expected value of the new patches so that

$$\frac{\left|E(new \quad patch) - E(old \quad patch)\right|}{E(old \quad patch) - E(placebo \quad patch)} \leq 0.2$$

The test statistic

We define 2 variables

$$z = oldpatch - placebo$$
$$y = newpatch - oldpatch$$

Ratio statistic

$$\theta = \frac{E_F(y)}{E_F(z)}$$

F is the joint distribution of y and z

What is the distribution of θ ?

The plug in estimate

parameter

$$\theta = \frac{E_F(y)}{E_F(z)}$$

parameter estimate

$$\hat{\theta} = \frac{\bar{y}}{\bar{z}} = \frac{1/8\sum_{i=1}^{8} y_i}{1/8\sum_{i=1}^{8} z_i}$$

The bootstrap algorithm

The observed data

$$\boxed{x_i = (z_i, y_i)}$$

$$x_1, x_2, \ldots\ldots\ldots, x_n$$

Since the distribution of the statistic is unknown we use bootstrap to approximate the distribution.

We resample pairs.

B bootstrap samples

$$x_1^*, x_2^*, \ldots\ldots\ldots, x_n^* \qquad \theta_1^*$$

$$x_1^*, x_2^*, \ldots\ldots\ldots, x_n^* \qquad \theta_b^* \longrightarrow \hat{\theta}_b^* = \frac{\bar{y}^*}{\bar{z}^*} = \frac{1/8\sum_{i=1}^{8} y_i^*}{1/8\sum_{i=1}^{8} z_i^*}$$

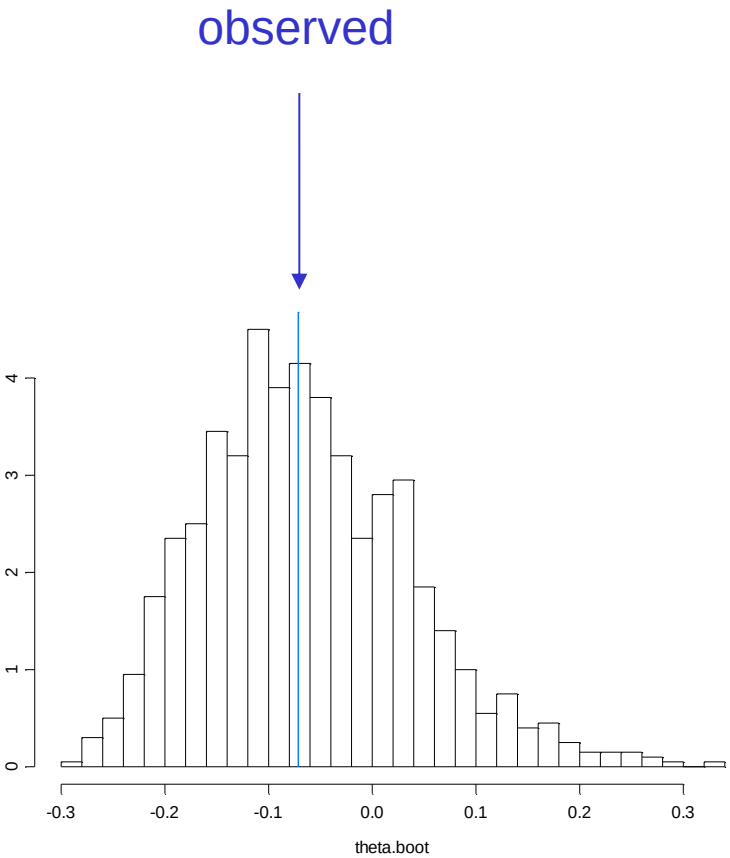
$$x_1^*, x_2^*, \ldots\ldots\ldots, x_n^* \qquad \theta_B^*$$

Data and observed statistic

```
   P     OLD  NEW
1  9243 17649 16449
2  9671 12013 14614
3 11792 19979 17274
4 13357 21816 23798
5  9055 13850 12560
6  6290  9806 10157
7 12412 17208 16570
8 18806 29044 26325

> mean(y)
[1] -452.25
> mean(z)
[1] 6342.375
> theta.obs <- mean(y)/mean(z)
> theta.obs
[1] -0.0713061
```

The observed ratio is -0.0713

The bootstrap replicates

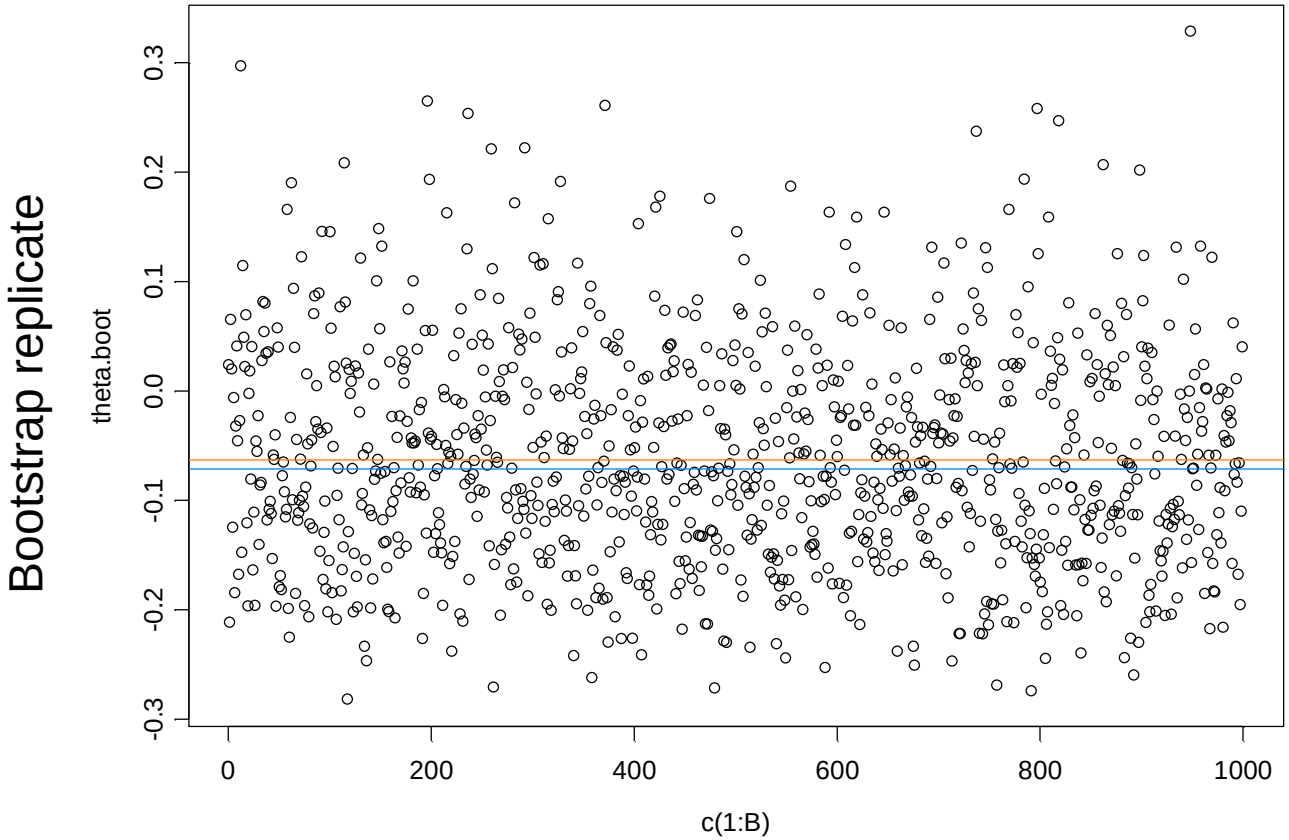


1000 bootstrap replicates.

Asymmetric distribution for the ratio.

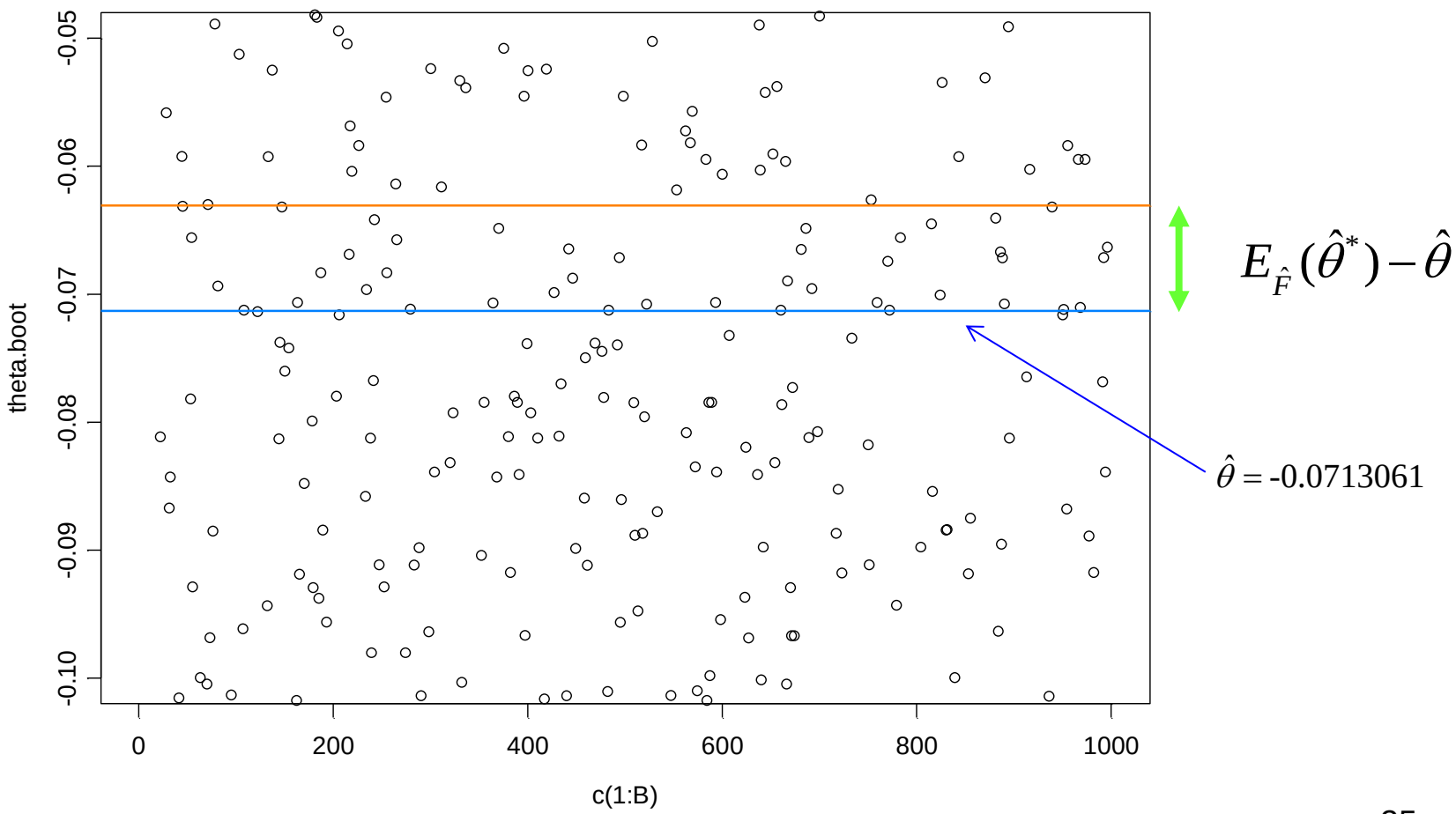
```
n<-length(z)
B<-1000
index<-c(1:n)
theta.boot<-c(1:B)
for(i in 1:B)
{
cat(i)
i.boot<-sample(index, size=n, replace=T)
y.boot<-y[i.boot]
z.boot<-z[i.boot]
theta.boot[i]<-mean(y.boot)/mean(z.boot)
}

hist(theta.boot,col=0,nclass=30,probability=T)
lines(c(theta.obs,theta.obs),c(0,5),lwd=2,col=6)
```

The bootstrap replicates and the bias

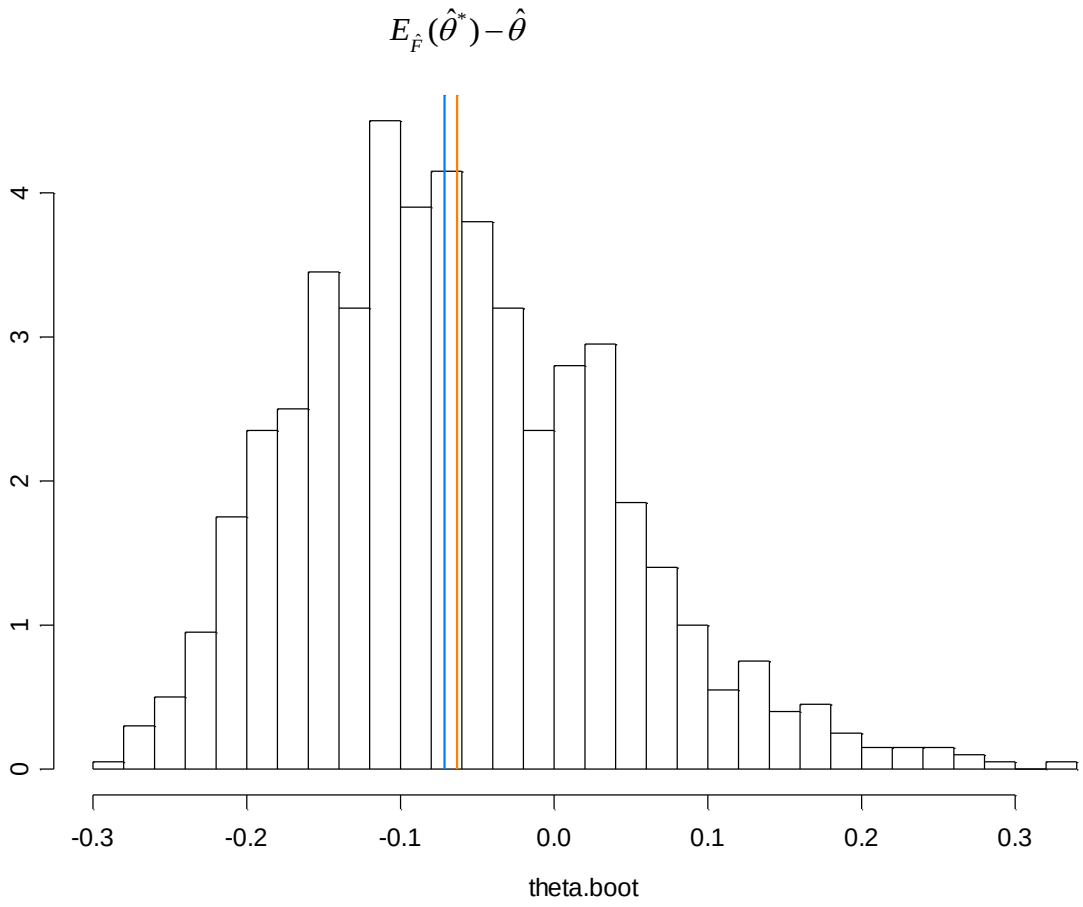


Bootstrap iteraion

The bootstrap replicates and bias



The bootstrap replicates and bias



Estimate for the bias

$$\hat{b}_{\hat{F}} = \hat{\theta}^* - \hat{\theta}$$

```
> m.boot <- mean(theta.boot)
> b.boot <- m.boot - theta.obs
> b.boot
[1] 0.008231291
```

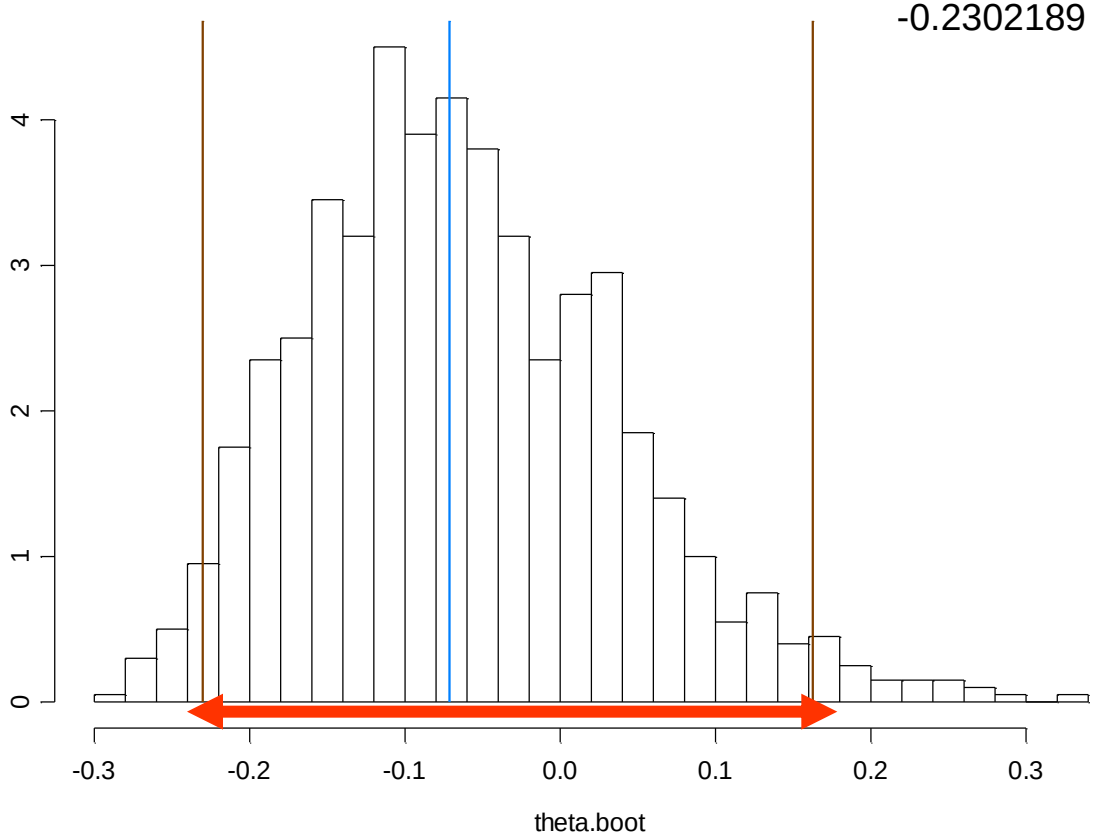
```
b.boot <- m.boot - theta.obs
```

The mean of the bootstrap replicates

The observed statistic

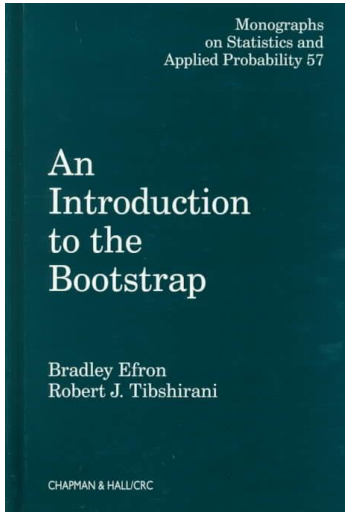
Bioequivalence ?

```
> quantile(theta.boot, probs = c(0.025, 0.975))
      2.5%     97.5%
-0.2302189 0.1626573
```



$$|\theta| \leq 0.2$$

Cross-validation of prediction error



Chapter 17

Topics

- Cross validation for regression problems.
- K-fold CV.
- Leave one out cross validation.
- Examples:
  - The hormone data.

The prediction problem

Suppose that we have an outcome y and we would like to predict the outcome (maybe as a function of other covariates).

Linear regression models etc.

Prediction error

The predictive model:

$$M(y, x, \theta)$$

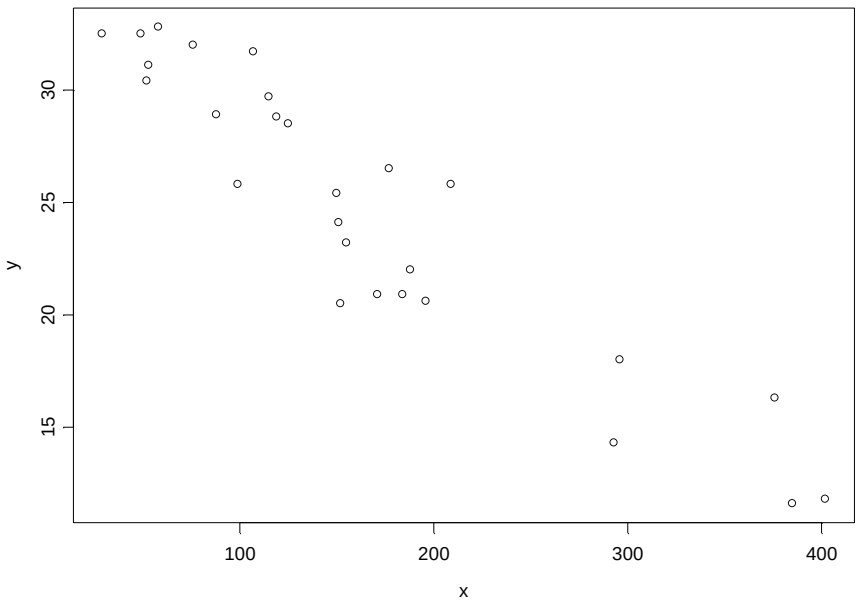
The error that we make when we predict the outcome using the model:

$$PE = E\left(y - \hat{y}\right)^2$$

Example: residuals sum of squares in regression.

Example: The hormone data

Amount in milligrams of anti-inflammatory hormone remaining in 27 devices, after a certain number of hours (hrs) of wear.



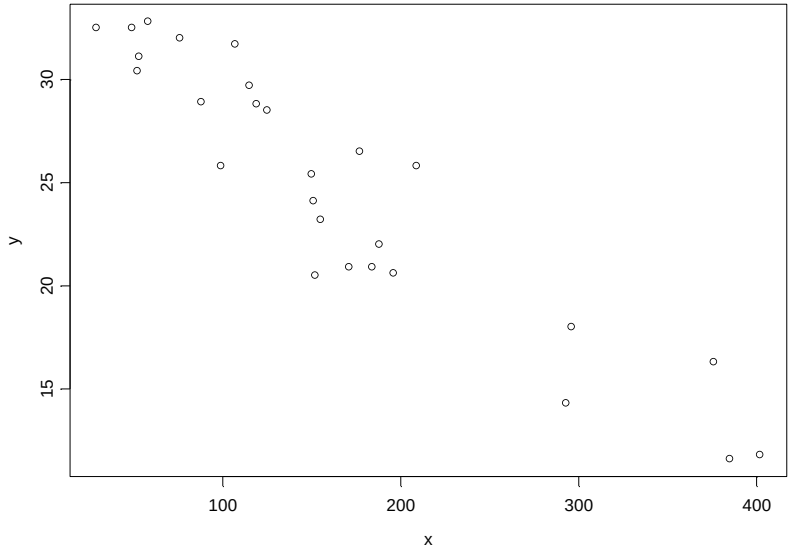
In R:

```
> help(hormone)
```

Data structure

$$\left(x_1, y_1\right)$$
$$\left(x_2, y_2\right)$$
$$\left(x_i, y_i\right)$$
$$\cdot$$
$$\left(x_n, y_n\right)$$

27 pairs of hormone levels and hours.

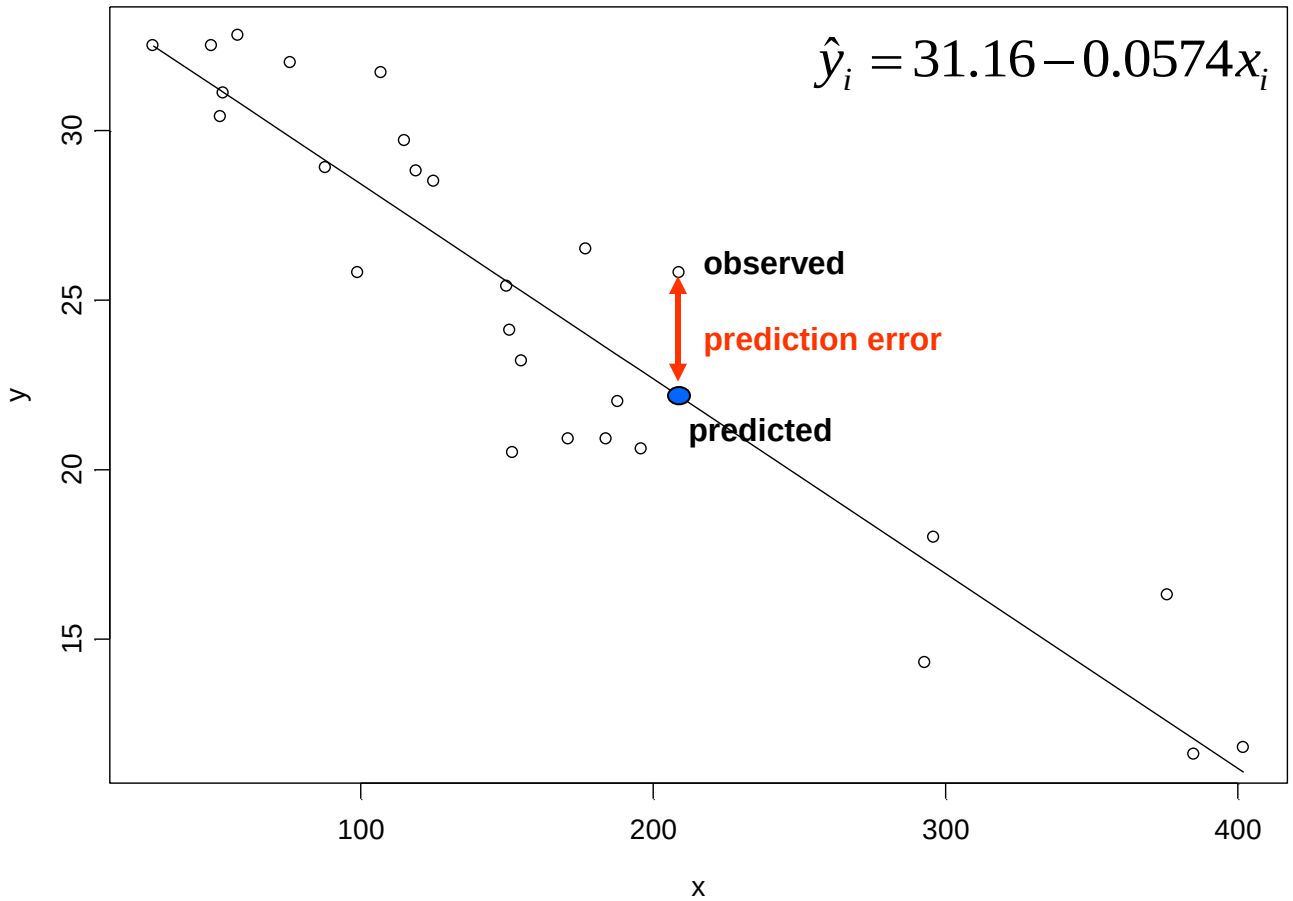


Model formulation

We assume that the amount of the hormone is (y) is a function of the hours (x):

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Data and predicted values



Estimated model and prediction error

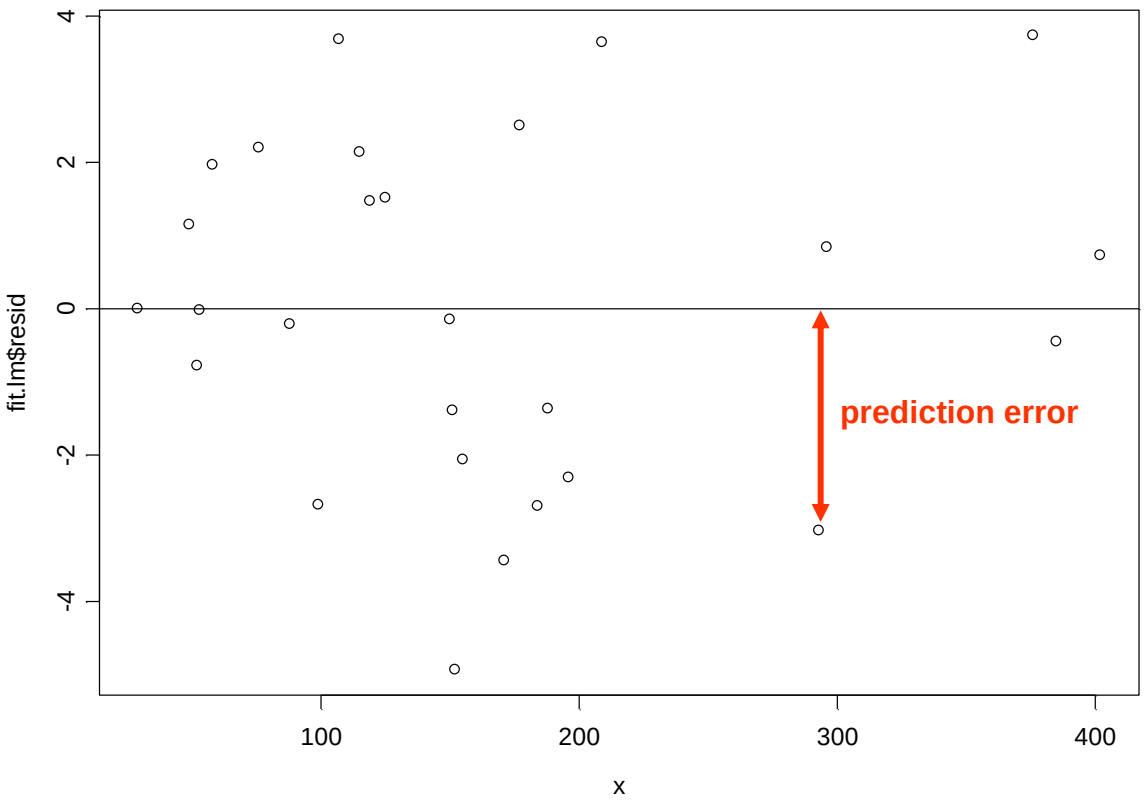
$$\hat{y}_i = 31.16 - 0.0574 x_i$$

$$\boxed{\frac{RSE}{n} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n}}$$

Mean of residuals sum of squares

```
> summary(fit.lm)

Call: lm(formula = y ~ x)
Residuals:
    Min      1Q   Median     3Q   Max
 -4.936 -1.728 -0.02287 1.739 3.732

Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)  34.1675   0.8672    39.3999   0.0000
          x  -0.0574   0.0045   -12.8683   0.0000

Residual standard error: 2.378 on 25 degrees of
     freedom
Multiple R-Squared: 0.8688
F-statistic: 165.6 on 1 and 25 degrees of freedom,
     the p-value is 1.584e-012

Correlation of Coefficients:
  (Intercept)
x -0.8494
```

Residuals



The prediction error

We used the data twice:

1. To estimate the model.
2. To calculate the prediction error.

As a results, our estimate for the prediction error tends to be optimistic.

How can we validate the model ?

Suppose that we use one dataset to estimate the model and another dataset to calculate the prediction error.

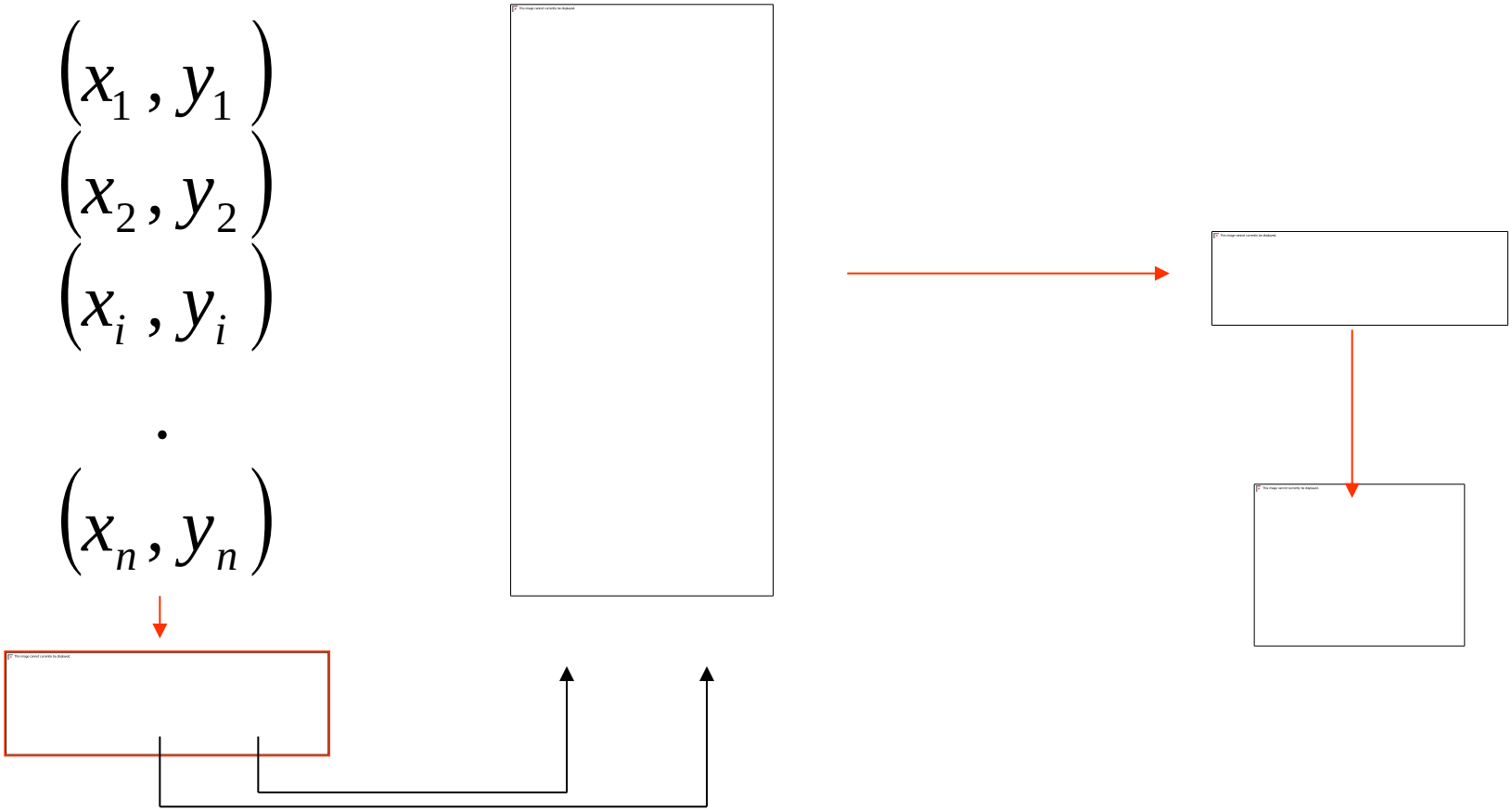
This means that we do not use the observation from the second dataset to estimate the parameters but only to calculate the prediction error.

Prediction error from new data

Estimate the model

Use the parameter estimates form the first data to calculate the prediction error

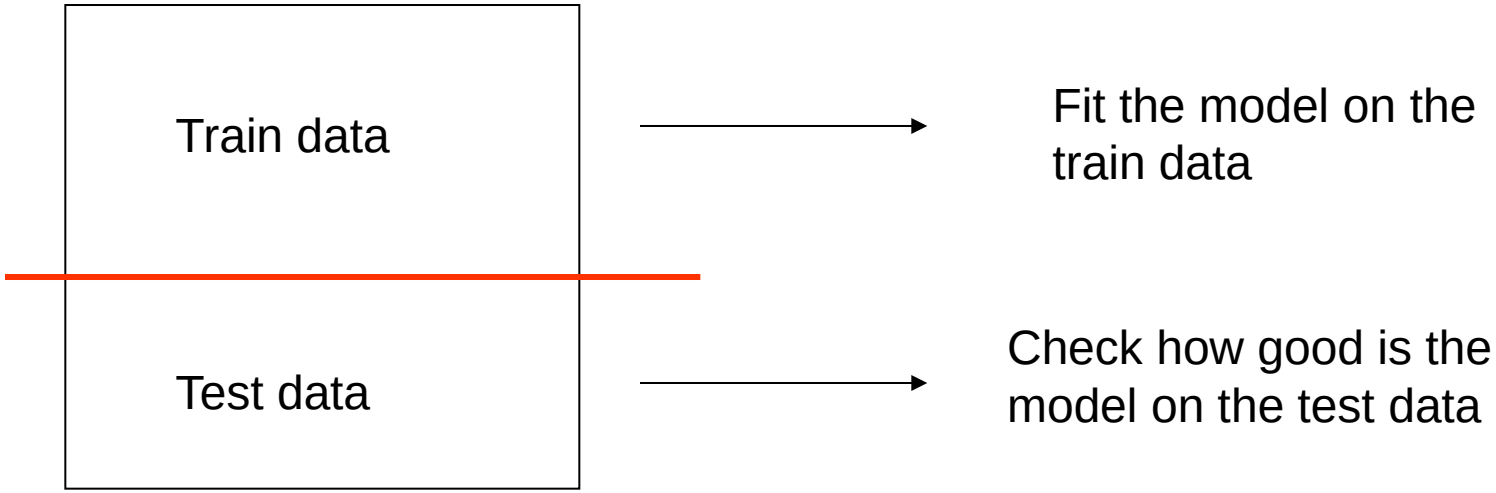
$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$(x_i, y_i)$$
$$\cdot$$
$$(x_n, y_n)$$



Cross validation

The idea is to split the sample into two parts

1. The train sample: the sample from which we estimate the parameters
2. The test sample: the sample which we use to calculate the prediction error (but not to estimate the parameters)



K-fold cross-validation

1. Split the sample into K equal parts.
2. For the kth part, fit the model to the other K-1 parts and calculate the prediction error of the fitted model when prediction the kth part.
3. Repeat for k=1,2,..,K.

k=n: leave one out cross validation

Step 1

Leave one observation out

$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\boxed{(x_i, y_i)}$$
$$\cdot$$
$$(x_n, y_n)$$

Step 2

Estimate the model using n-1 observations

$$y_j = \alpha + \beta x_j + \varepsilon_j$$

$$\hat{\alpha}, \hat{\beta}$$

Step 3

Calculate the prediction value for the observation which left out

$$\hat{y}_i^{-i} = \hat{\alpha} + \hat{\beta} x_i$$

prediction error

$$y_i - \hat{y}_i^{-i}$$

Cross-validation score

The prediction error is calculated for each observation.

The cross validation score versus the residuals sum of squares.

$$CV = \frac{1}{n}\sum_{i=1}^{m}\left(y_i - \hat{y}_i^{-i}\right)^2$$

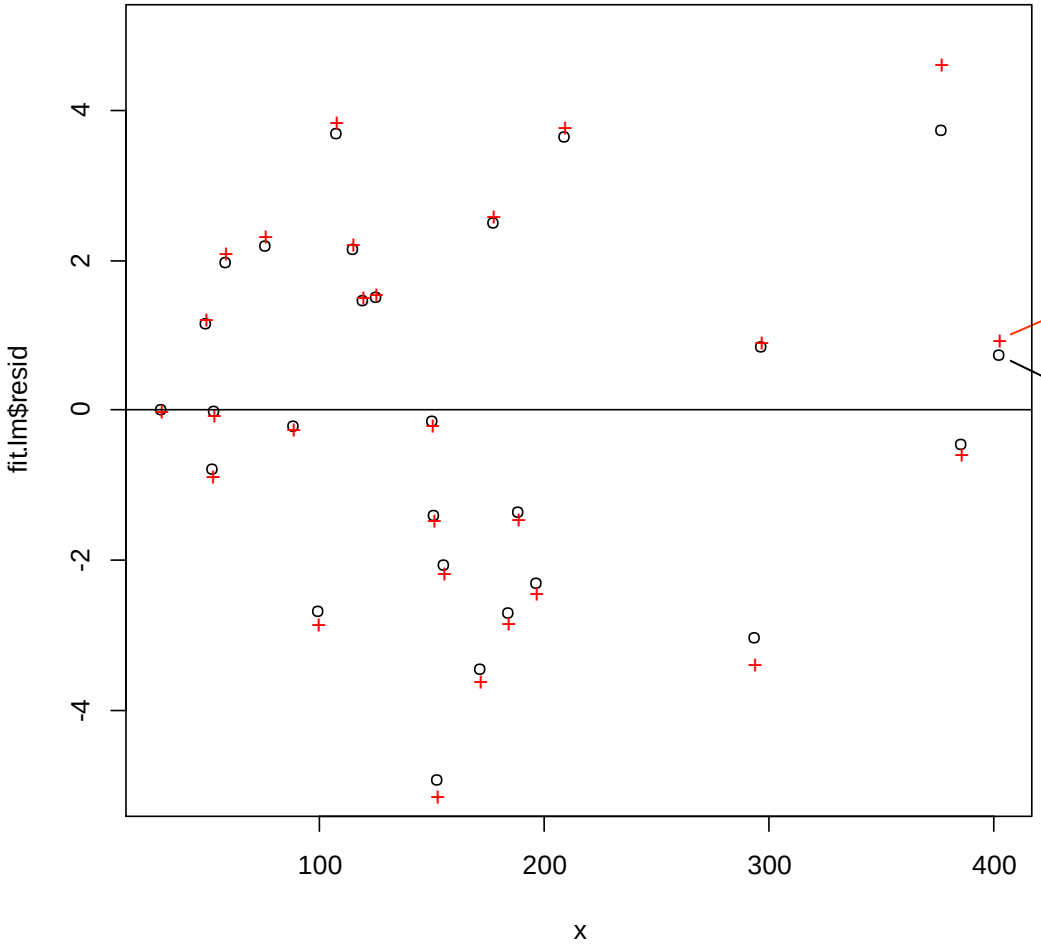
$$\frac{RSE}{n} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}$$

CV score and RSE

In our example the CV score is about 7.3% higher than the residual sum of squares.

```
> cv.score
[1] 2.455137
> cv.score/(sqrt(sum((y - fit.lm$fit)^2)/27))
[1] 1.072869
```

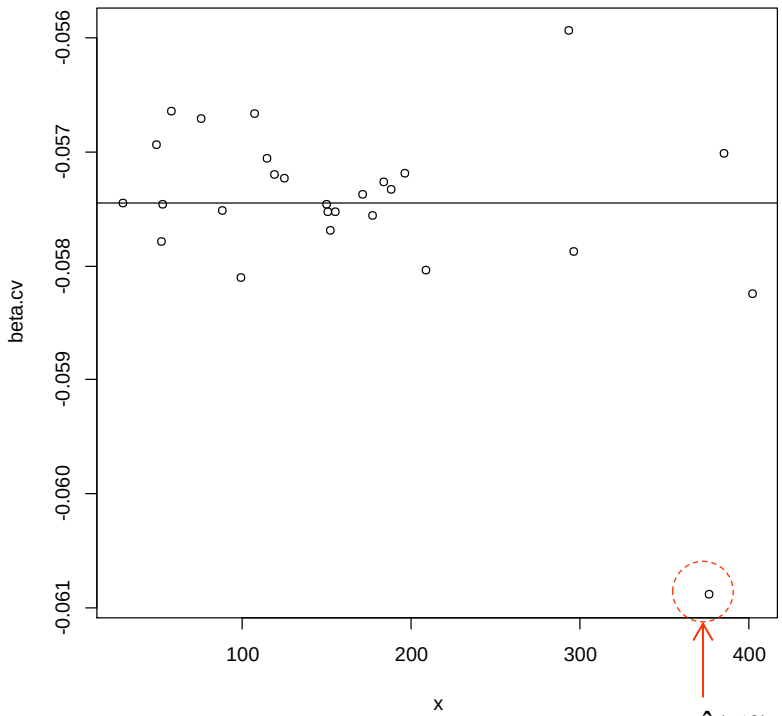
Cross validated residuals



Residual from the cross validation

Residual from the linear regression model with all data

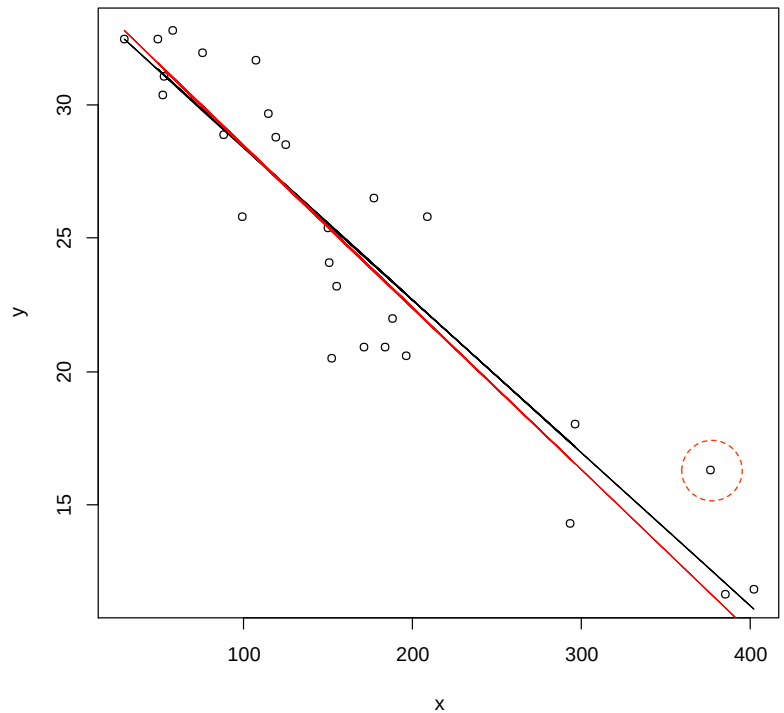
Parameter estimate for the slope



$\hat{\beta} = -0.05745$

$\hat{\beta}^{(-10)} = -0.0608$

Parameter estimate for the slope when the i'th observation is not included.



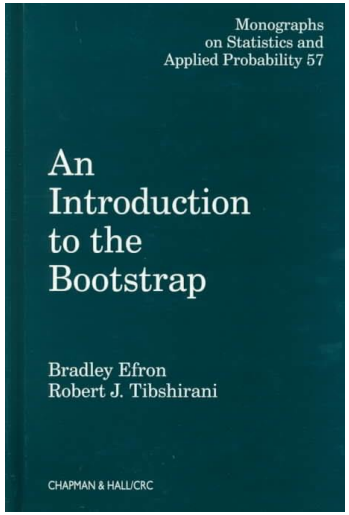
Fitted model with and without the observation

```
> x[10]
[1] 376
> y[10]
[1] 16.3
```

R code for LOOCV

```
> n <- length(x)
> beta.cv <- fit.cv <- c(1:n)
> for(i in 1:n) {
   cat(i)
   x.cv <- x[ - c(i)]
   y.cv <- y[ - c(i)]
   fit.lm.cv <- lm(y.cv ~ x.cv)
   fit.cv[i] <- fit.lm.cv$coeff[1] + fit.lm.cv$coeff[2] * x[i]
   beta.cv[i] <- fit.lm.cv$coeff[2]
}
1234567891011121314151617181920212223242526 27
> res.cv <- y - fit.cv
> cv.score <- sqrt(sum((res.cv^2))/n)
> cv.score
[1] 2.455137
> cv.score/(sqrt(sum((y - fit.lm$fit)^2)/27))
[1] 1.072869
```

The jackknife



Chapter 11

Topics

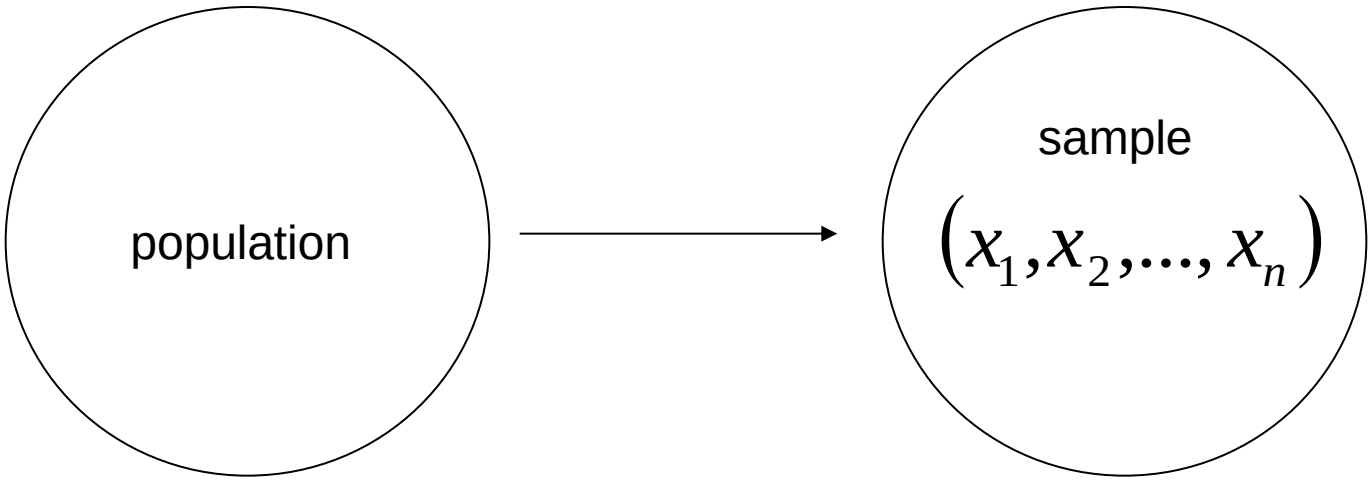
- The jackknife method.
- Standard error of the mean.
- Estimation of bias.
- Examples:
  - The airquality data.
  - The patch data.

Both R objects.

A random Sample from F

We observed a random sample from the probability distribution F

$$F \rightarrow (x_1, x_2, \ldots, x_n)$$



The jackknife sample

The observed sample

$$x = \left(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n\right)$$

The jackknife sample

$$x_{(i)} = \left(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right)$$

The jackknife replicate

We calculate the plug-in estimate for each jackknife sample. The jackknife replicate:

$$\hat{\theta}_{(i)} = t(x_{(i)}) = t\left(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right)$$

The jackknife estimator for the unknown parameter:

$$\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$$

The jackknife standard error

The jackknife estimate for the standard error is given by

$$\boxed{se_{jack} = \left[\frac{n-1}{n} \sum \left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)} \right)^2 \right]^{\frac{1}{2}}}$$

Standard error: bootstrap and jackknife

Bootstrap standard error

$$se(\hat{\theta}) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}^* \right)^2 \right\}^{0.5}$$

Jackknife standard error

$$se_{jack} = \left[\frac{n-1}{n} \sum \left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)} \right)^2 \right]^{\frac{1}{2}}$$

n-1: inflation factor

The inflation factor

The inflation factor *(n-1)/n* is needed since the jackknife deviation is smaller than the bootstrap deviation:

$$\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2 < \left(\hat{\theta}_b^* - \hat{\theta}_{(.)}^*\right)^2$$

Jackknife sample is more similar to the observed sample than a bootstrap sample.

$$x_{(i)} = \left(x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\right)$$

Jackknife sample

$$x^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right)$$

bootstrap sample

$$x^{obs} = \left(x_1, x_2, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n\right)$$

The inflation factor

Why *n-1* ?

Jackknife estimate for the standard error for the sample mean:

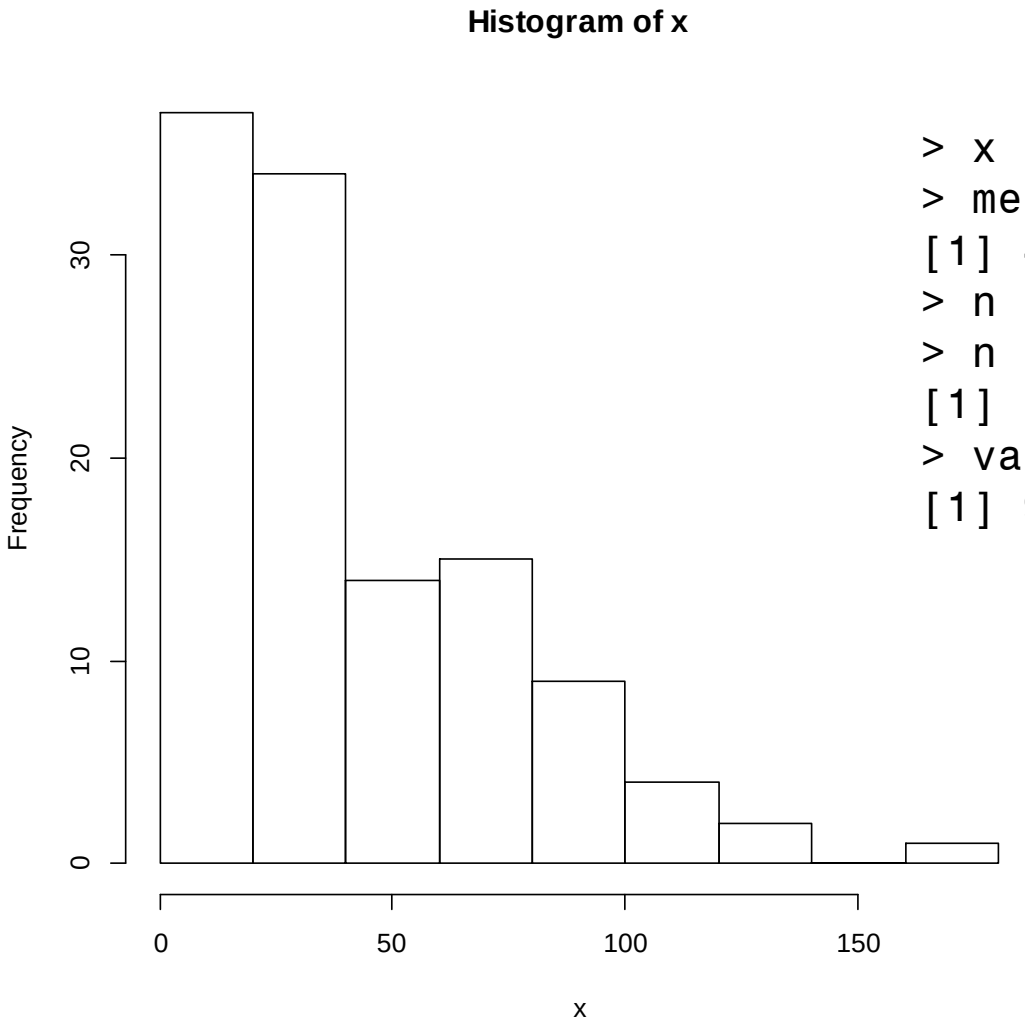
$$se_{jack} = \left[\frac{n-1}{n} \sum \left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)} \right)^2 \right]^{\frac{1}{2}} = \left[\frac{1}{n(n-1)} \sum \left(x_i - \bar{x} \right)^2 \right]^{\frac{1}{2}}$$

$$\left[\frac{(n-1)}{n(n-1)} \sum \left(x_i - \bar{x} \right)^2 \right]^{\frac{1}{2}} = \left[\frac{1}{n} \sum \left(x_i - \bar{x} \right)^2 \right]^{\frac{1}{2}}$$

This is only true when the parameter of interest is the sample mean.

In practice, n-1 is an arbitrary choice.

Example: the airquality data



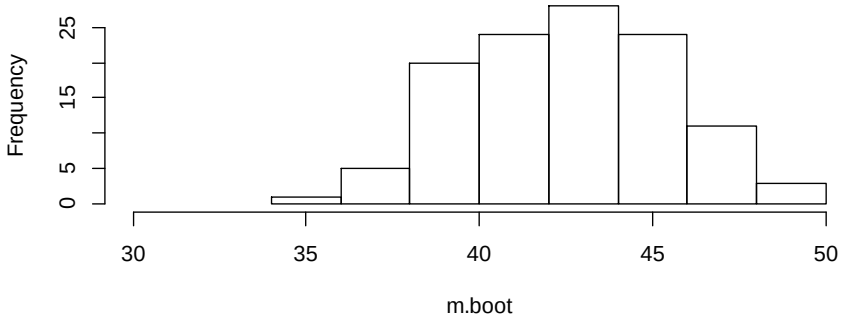
Histogram of ozone levels

```
> x <- na.omit(airquality$Ozone)
> mean(x)
[1] 42.12931
> n <- length(x)
> n
[1] 116
> var(x)/n
[1] 9.381039
```

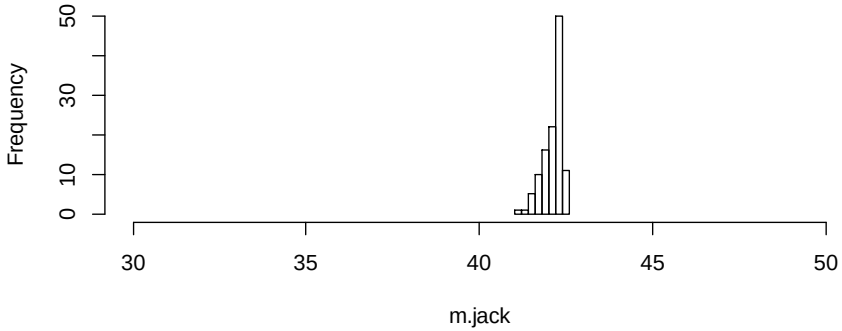
$$\bar{x} = 42.12931$$

Why do we need the inflation factor ?

bootstrap



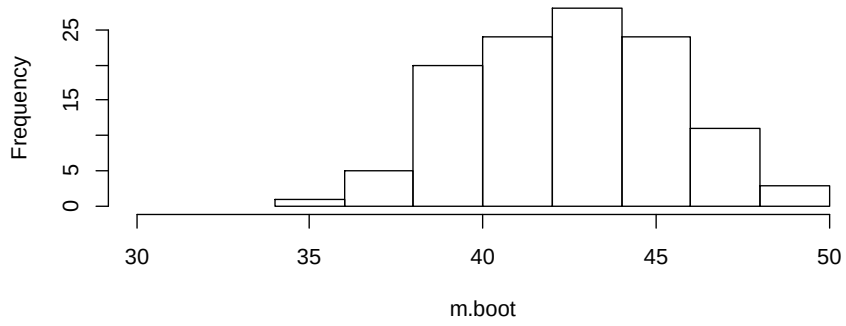
jackknife

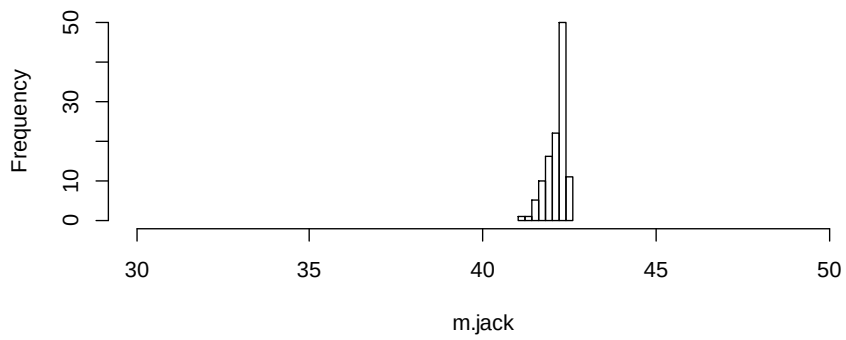



```
> x <- air$ozone
> hist(x, col = 0)
> mean(x)
[1] 3.247784
> n <- length(x)
> var(x)/n
[1] 0.007142405
> m.boot <- m.jack <- c(1:n)
> for(i in 1:n) {
    cat(i)
    x.jack <- x[ - c(i)]
    m.jack[i] <- mean(x.jack)
    x.boot <- sample(x, size = n,
    replace = T)
    m.boot[i] <- mean(x.boot)
}
```

Bootstrap and jackknife approximation for the distribution of the sample mean

Why do we need the inflation factor ?



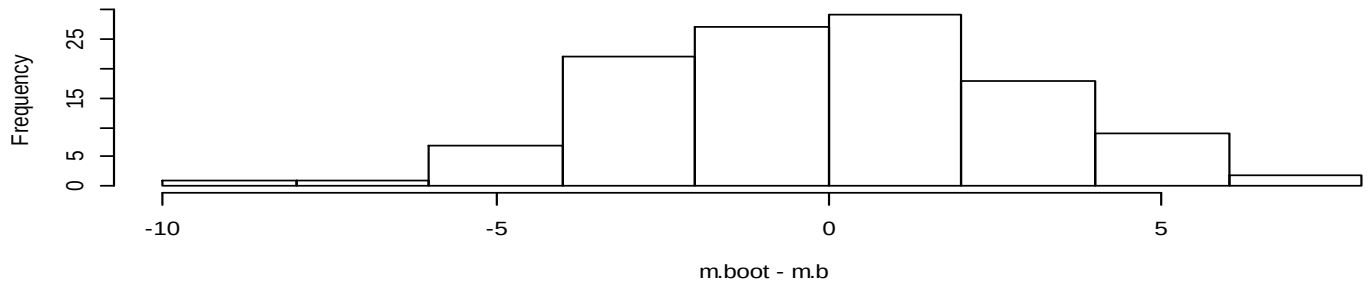


The jackknife datasets are more similar (on average) to the original dataset than the bootstrap datasets.

This is a problem if we would like to estimate the standard error.

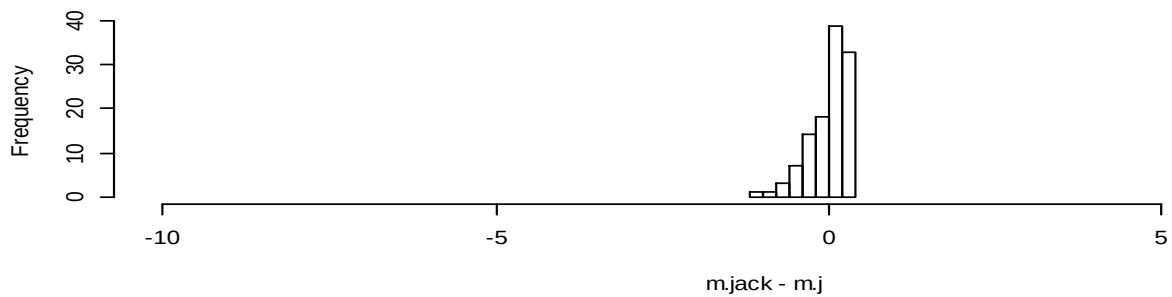
Solution: use an inflation factor.

Why do we need the inflation factor ?



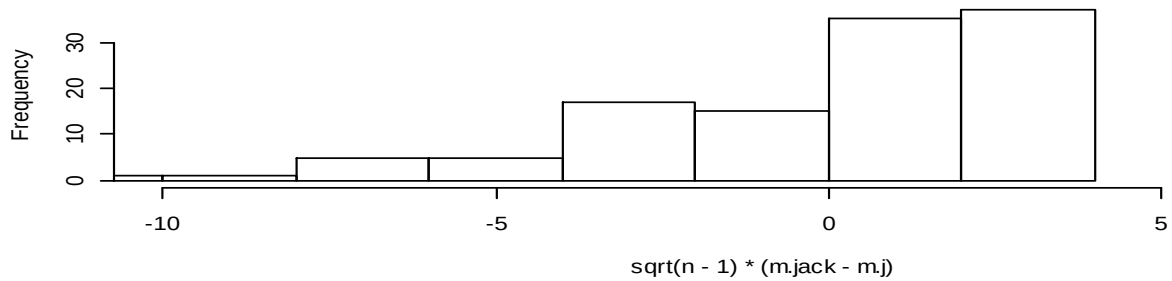
bootstrap

$$\hat{\theta}_b^* - \overline{\hat{\theta}}_b^*$$



jackknife

$$\hat{\theta}_{(i)} - \hat{\theta}_{(\ )}$$



Infelated jackknife

$$\sqrt{n-1} \times \left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)$$

The air data: standard error for the sample mean - bootstrap and jackknife

There is a mistake in these results (this is not the sir dataset)…

```
> (sum((m.jack - mean(m.jack))^2) * ((n - 1)/n))^(0.5)
[1] 0.08451275
> var(x)/sqrt(111)
[1] 0.0752499
```

```
> B <- 500
> m.boot <- c(1:B)
> for(i in 1:B) {
   cat(i)
   x.boot <- sample(x, size = n, replace = T)
   m.boot[i] <- mean(x.boot)
}
> (sum((m.boot - mean(m.boot))^2)/(B - 1))^0.5
[1] 0.08862128
```

The jackknife estimate for the bias

Similar to the bootstrap, one can use the jackknife in order to estimate the bias (or standard error)
of a particular statistic.

It becomes very useful for “non standard” statistic for with the distribution is unknown.

The bootstrap and jackknife estimate for the bias

Bootstrap estimate for the bias

$$\hat{b}_{\hat{F}} = \hat{\theta}^* - \hat{\theta}$$

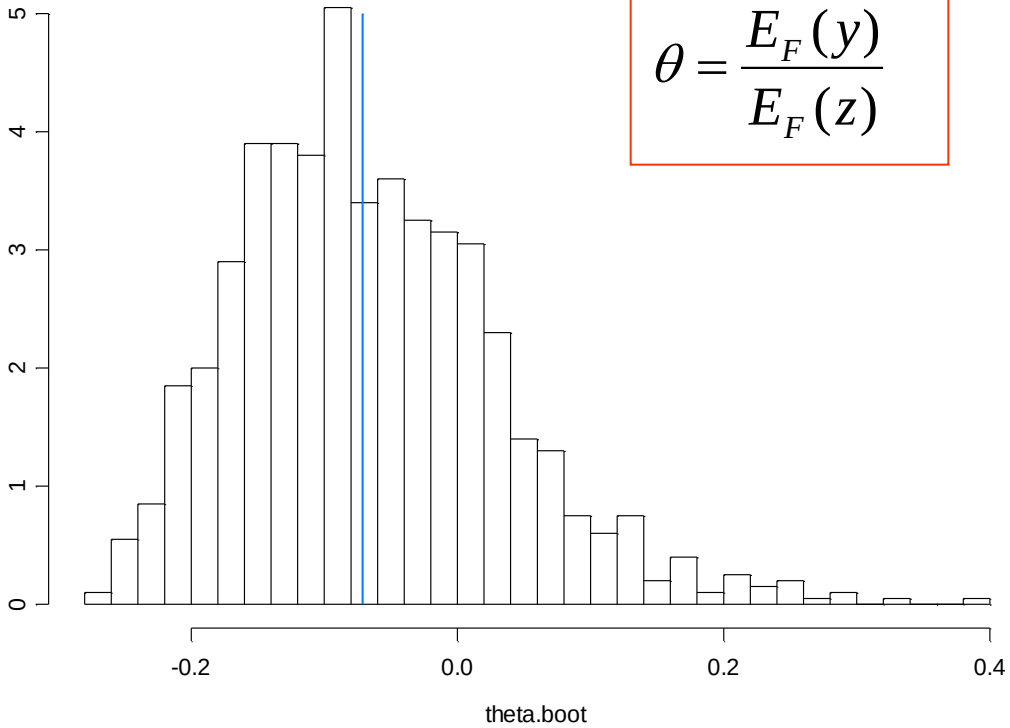
Jackknife estimate for the bias

$$\hat{b}_{\hat{F},jack} = (n-1)\left(\hat{\theta}_{(.)} - \hat{\theta}\right)$$

$$\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$$

The patch data

Bootstrap estimate for the bias



$$\theta = \frac{E_F(y)}{E_F(z)}$$

```
> z <- p$V2 - p$V1
> y <- p$V3 - p$V2
> p
     V1    V2    V3
1  9243 17649 16449
2  9671 12013 14614
3 11792 19979 17274
4 13357 21816 23798
5  9055 13850 12560
6  6290  9806 10157
7 12412 17208 16570
8 18806 29044 26325

> mean(y)
[1] -452.25
> mean(z)
[1] 6342.375
> theta.obs <- mean(y)/mean(z)
  theta.obs
```

```
[1] -0.0713061
```

```
> m.boot <- mean(theta.boot)
> b.boot <- m.boot - theta.obs
> b.boot
[1] 0.0050834
```

The patch data: bias

$$\hat{\theta}_{(i)} = \frac{\bar{y}_{(i)}}{\bar{z}_{(i)}} = \frac{1/7\sum_{j=1}^{7} y_{j,(i)}}{1/7\sum_{j=1}^{7} z_{j,(i)}}$$

```
> theta.obs <- mean(y)/mean(z)
> theta.obs
[1] -0.0713061
> n <- length(z)
> m.jack <- c(1:n)
> for(i in 1:n) {
   cat(i)
   z.jack <- z[ - c(i)]
   y.jack <- y[ - c(i)]
   m.jack[i] <-
   mean(y.jack)/mean(z.jack)
}
12345678
> (n - 1) * (mean(m.jack) - theta.obs)
[1] 0.008002488
```

Jackknife estimate for the bias

$$\hat{b}_{\hat{F},jack} = (n-1)\left(\hat{\theta}_{(.)} - \hat{\theta}\right)$$