Contents

Chapter 4

Resampling Methods

4.1 Non-parametric computational estimation

Let $x_1, \ldots, x_n$ be a realization of the i.i.d. r.vs $X_1, \ldots, X_n$ with a c.d.f. F.

We are interested in the precision of estimation of a population parameter θ_F. One possibility is to estimate θ_F by $\theta_{\hat{F}}$, where $\hat{F}$ is the empirical distribution funtion. We will denote an estimator of a parameter θ by $\hat{\theta}$.

Examples

1. $$\theta_F = E(X) = \int_{-\infty}^{\infty} x f(x) dx,$$

   where $f(x) = \frac{dF(x)}{dx}$. Then

   $$\theta_{\hat{F}} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

   which is the sample mean. Here we assign equal probability, $\frac{1}{n}$, to each realization of X.

2. $$\theta_F = var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx.$$

   Then

   $$\theta_{\hat{F}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

   which is the variance of the sample.

3.

$$\theta_F = F(c) = P(X \leqslant c)$$

Then

$$\theta_{\hat{F}} = \frac{1}{n}\#\{i : x_i \leqslant c\}$$

Question
How good is $\hat{\theta} = \theta_{\hat{F}}$ as an estimator of θ_F?

Three common measures of goodness are:

$$Bias_\theta(\hat{\theta}) = E_F(\hat{\theta}) - \theta \tag{4.1}$$

$$se_\theta(\hat{\theta}) = \sqrt{var(\hat{\theta})} \tag{4.2}$$

$$MSE_\theta(\hat{\theta}) = E_F\left[(\hat{\theta} - \theta)^2\right] \tag{4.3}$$

It is easy to see that

$$MSE_\theta(\hat{\theta}) = var(\hat{\theta}) + \left(Bias_\theta(\hat{\theta})\right)^2 \tag{4.4}$$

Namely:

$$E\left[(\hat{\theta} - \theta)^2\right] = E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) =$$
$$E\left[\hat{\theta}^2 - 2\hat{\theta}(E(\hat{\theta}) - Bias_\theta(\hat{\theta})) + \left(E(\hat{\theta}) - Bias_\theta(\hat{\theta})\right)^2\right] =$$
$$E\left[\hat{\theta}^2 - 2\hat{\theta}E(\hat{\theta}) - 2\hat{\theta}Bias_\theta(\hat{\theta}) + \left(E(\hat{\theta})\right)^2 - 2E(\hat{\theta})Bias_\theta(\hat{\theta}) + \left(Bias_\theta(\hat{\theta})\right)^2\right] =$$
$$E\left[\hat{\theta}^2 - 2\hat{\theta}E(\hat{\theta}) + \left(E(\hat{\theta})\right)^2\right] + \left(Bias_\theta(\hat{\theta})\right)^2 =$$
$$var(\hat{\theta}) + \left(Bias_\theta(\hat{\theta})\right)^2$$

Also note that

$$\sqrt{MSE_\theta(\hat{\theta})} = \sqrt{var(\hat{\theta}) + \left(Bias_\theta(\hat{\theta})\right)^2} = se_\theta(\hat{\theta}) \times \sqrt{1 + \left(\frac{Bias_\theta(\hat{\theta})}{se_\theta(\hat{\theta})}\right)^2} \doteq$$
$$se_\theta(\hat{\theta}) \times \left[1 + \tfrac{1}{2}\left(\frac{Bias_\theta(\hat{\theta})}{se_\theta(\hat{\theta})}\right)^2\right]$$

Problem
How to calculate $Bias_\theta(\hat{\theta})$, $se_\theta(\hat{\theta})$ and $MSE_\theta(\hat{\theta})$?

If we knew the distribution F then we could calculate expected value and variance of the estimator $\hat{\theta}$ directly from definitions. It may be difficult if $f(x) = \frac{dF}{dx}$ is complicated. Then a practical alternative is simulation:

- Generate a large number of random samples from the population with the c.d.f. F and calculate a value of $\hat{\theta}$ for each sample.
- The mean and variance of the set of generated values of $\hat{\theta}$ will give a good approximation to $E_F(\hat{\theta})$ and $var_F(\hat{\theta})$.

What if F is unknown? Then the simulation from F is impossible. In such situations a further approximation is to replace F by $\hat{F}$. Let θ^* be the value of θ calculated at the random sample from $\hat{F}$. The idea is that

$$Bias_\theta(\hat{\theta}) \approx Bias_{\hat{\theta}}(\theta^*)$$

and

$$var_F(\hat{\theta}) \approx var_{\hat{F}}(\theta^*).$$

The heuristic reasoning is that $\hat{F}$ is close to F and so the relationship of $\theta_{\hat{F}}$ to θ_F should be close to the relationship of θ_{F^*} to $\theta_{\hat{F}}$, as shown in the diagram below.

$$\begin{array}{ccccc} \text{True unknown} & & \text{Empirical} & & \text{Resampled} \\ F & \xrightarrow{data} & \hat{F} & \xrightarrow{resampling} & F^* \\ \downarrow & & \downarrow & & \downarrow \\ \theta_F & & \theta_{\hat{F}} & & \theta_{F^*} \end{array}$$

4.2 Bootstrap estimates of bias, standard error and MSE

Assume we do not know F. The bootstrap estimates of $Bias_\theta(\hat{\theta})$, $se_\theta(\hat{\theta})$ and $MSE_\theta(\hat{\theta})$ are obtained by substituting $\hat{F}$ for F in (4.1), (4.2) and (4.3), respectively. $\hat{F}$ is the distribution which assigns probability $\frac{1}{n}$ to each observation x_i. So, a random sample from $\hat{F}$ is just a random sample from the set $\{x_1, \ldots, x_n\}$ with replacement. The procedure to calculate the estimates is the following:

- construct N samples of size n from $\{x_1, \ldots, x_n\}$ with replacement;
- denote the bootstrap samples by $\{x_1, \ldots, x_n\}_i$, $i = 1, \ldots, N$;
- denote by θ_i^* the value of the estimator calculated for the i-th bootstrap sample;
- calculate sample mean and variance of bootstrap estimates θ_i^*, $i = 1, \ldots, n$, that is $\bar{\theta}^* = \frac{1}{N}\sum_{i=1}^N \theta_i^*$, $\frac{1}{N-1}\sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2$;

$Bias_\theta(\hat{\theta})$, $var_F(\hat{\theta})$, $se_F(\hat{\theta})$ and $MSE_\theta(\hat{\theta})$ are approximated, respectively, by $Bias_{\hat{\theta}}(\theta^*)$, $var_{\hat{F}}(\theta^*)$, $se_{\hat{\theta}}(\theta^*)$ and $MSE_{\hat{\theta}}(\theta^*)$ which are further approximated by

$$\widehat{Bias}_{\hat{\theta}}(\theta^*) = \bar{\theta}^* - \hat{\theta},$$

$$\widehat{var}_{\hat{F}}(\theta^*) = \frac{1}{N-1}\sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2,$$

$$\widehat{se}_{\hat{F}}(\theta^*) = \sqrt{\frac{1}{N-1}\sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2},$$

$$\widehat{MSE}_{\hat{\theta}}(\theta^*) = \frac{1}{N-1}\sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2 + (\bar{\theta}^* - \hat{\theta})^2.$$

The following diagram represents the bootstrap resampling method:

Empirical distribution	Bootstrap samples of size n		Bootstrap replications of $\hat{\theta}$	Bootstrap estimates
	$\{x_1, \ldots, x_n\}_1$	$\rightarrow$	θ_1^*	
$\hat{F}$	$\{x_1, \ldots, x_n\}_2$	$\rightarrow$	θ_2^*	
$\{x_1, \ldots, x_n\}$	$\vdots$	$\vdots$	$\vdots$	
	$\{x_1, \ldots, x_n\}_N$	$\rightarrow$	θ_N^*	$\Downarrow$
				bias: $\bar{\theta}^* - \hat{\theta}$ variance: $\frac{1}{N-1}\sum_{i=1}^N (\theta_i^* - \bar{\theta}^*)^2$

Example: sample mean and sample median
Let $x_1, \ldots, x_n$ be a realization of the i.i.d. r.vs $X_1, \ldots, X_n$ with a c.d.f. F. Consider $\theta_F = E_F(X_i)$ and let $\theta_{\hat{F}} = \bar{X}$. We know that the sample mean $\bar{X}$ is an unbiased estimator of $E(X_i)$. What is the bootstrap bias of the mean?

$$Bias_{\hat{\theta}}(\theta^*) = E_{\hat{F}}(\theta^*) - \hat{\theta} = E_{\hat{F}}(\bar{X}) - \bar{X} = 0$$

Is the estimate of the bootstrap bias of the sample mean, $\widehat{Bias}_{\hat{\theta}}(\bar{X})$, equal to zero as well?

Let {2.3 3.4 2.5 3.2 2.7 2.6 3.1 3.5 2.9 2.5}

be a sample from a population with a c.d.f. F.

Here the sample mean is 2.87 and the sample median is 2.80.

15 bootstrap samples										replicates of sample mean	median
{3.2	2.5	3.2	3.2	3.4	3.2	2.5	2.7	2.5	2.5}	2.89	2.95
{2.3	3.4	3.5	2.9	2.6	3.5	2.5	2.9	2.9	3.1}	2.96	2.90
{2.5	2.3	3.1	2.5	3.4	3.1	2.3	3.1	3.5	3.5}	2.93	3.10
{2.3	2.3	2.6	3.4	2.5	2.6	2.3	3.1	2.6	2.5}	2.62	2.55
{2.5	3.5	2.9	3.4	2.5	3.4	2.6	2.3	2.3	3.2}	2.86	2.75
{2.5	2.5	2.5	3.4	2.9	3.5	3.5	2.7	3.5	3.2}	3.02	3.05
{3.5	3.4	2.6	2.5	2.9	3.4	3.2	2.3	3.1	2.9}	2.98	3.00
{3.1	2.5	3.1	2.3	2.6	2.7	3.2	3.4	3.4	2.5}	2.88	2.90
{3.1	3.4	3.1	3.1	2.5	2.6	3.2	2.7	3.5	2.3}	2.95	3.10
{2.6	2.5	2.3	2.5	2.3	2.6	3.2	2.5	2.7	2.5}	2.57	2.50
{2.9	3.5	2.9	2.3	2.7	2.7	2.6	2.5	2.9	3.1}	2.81	2.80
{2.9	3.1	3.5	2.3	2.7	2.3	3.5	3.2	2.5	2.9}	2.89	2.90
{3.4	2.9	3.2	3.2	3.1	2.9	3.4	2.7	3.5	3.2}	3.15	3.20
{2.7	3.1	3.4	3.2	3.5	2.5	3.2	2.9	2.5	3.4}	3.04	3.15
{2.5	3.2	3.5	2.5	2.7	3.1	3.2	2.9	2.5	3.4}	2.95	3.00

The bootstrap estimate of the sample mean is the average of the replicates θ_i^*, that is

$$\bar{\theta}^* = \frac{1}{15}(2.89 + 2.96 + 2.93 + \ldots + 2.95) = 2.898$$

Hence, the estimate of bootstrap bias of the sample mean is

$$\widehat{Bias}_{\hat{\theta}}(\theta^*) = \bar{\theta}^* - \hat{\theta} = 2.898 - 2.87 = 0.098 \neq 0.$$

So, the answer to the question if the estimate of the bootstrap bias of the sample mean, $\widehat{Bias}_{\hat{\theta}}(\bar{X})$, is equal to zero, is GENERALLY NOT.

Similar calculations for the above data and the bootstrap samples show that the estimate of the bootstrap bias for the sample median is

$$\widehat{Bias}_{\hat{\theta}}(\theta^*) = \bar{\theta}^* - \hat{\theta} = 2.916 - 2.80 = 0.116$$

However, we do not know what is the true bias of the sample median, so we do not know how good this estimate is.

The following question arises: How big should the bootstrap sample be to get a high probability that the estimate of the bootstrap bias of an estimator is close to the true value of the bias (known or unknown)?

The Central Limit Theorem says that the distribution of an average is approximately normal if the sample size is large and the variance is finite. Applying it to the bootstrap replicates θ_i^* we get

$$P\left(|\bar{\theta}^* - E_{\hat{F}}(\theta^*)| < 2\frac{\widehat{se}_{\hat{F}}(\theta^*)}{\sqrt{N}}\right) \cong 0.95. \tag{4.5}$$

This means

$$P\left(|\widehat{Bias}_{\hat{\theta}}(\theta^*) - Bias_{\hat{\theta}}(\theta^*)| < 2\frac{\widehat{se}_{\hat{F}}(\theta^*)}{\sqrt{N}}\right) \cong 0.95.$$

Hence, we need N such that $2\frac{\widehat{se}_{\hat{F}}(\theta^*)}{\sqrt{N}}$ is very small, for example 0.001.

Example: The patch data (Efron, Tibshirani, 1993, page 127)
Eight subjects wore medical patches designed to increase the blood levels of a certain natural hormone. Each subject had his blood levels of the hormone measured after wearing three different patches: a placebo patch, which had no medicine in it, an old patch which was from a lot manufactured at an old plant, and a new patch, which was from a lot manufactured at a newly opened plant. The purpose of the experiment was to show that the new plant was producing patches equivalent to those from the old plant. The observations are in the table below.

subject	placebo p_i	old patch old_i	new patch new_i	oldpatch - placebo $z_i = old_i - p_i$	newpatch - oldpatch $y_i = new_i - old_i$
1	9243	17649	16499	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19979	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

The Food and Drug Administration (FDA) criterion for the *bioequivalence* is that the expected value of the new patches match that of the old patches in the sense that

$$\frac{|E(new) - E(old)|}{E(old) - E(placebo)} \leqslant 0.2$$

This means that the new patch should match the old one within 20% of the amount of hormone the old drug adds to placebo blood levels. Denote the parameter of interest by θ, i.e.,

$$\theta = \frac{E(new) - E(old)}{E(old) - E(placebo)}$$

and let us assume that the pairs $x_i = (z_i, y_i)$ are realization of i.i.d. bivariate r.vs $X_i = (Z_i, Y_i)$ with unknown c.d.f. F. Then the parameter θ is

$$\theta = \frac{E_F(Y)}{E_F(Z)}.$$

The natural estimator of θ is

$$\hat{\theta} = \frac{\bar{Y}}{\bar{Z}}$$

and its value for the given observations is

$$\hat{\theta} = \frac{-452.3}{6342} = -0.0713$$

whose absolute value is less then 0.2.

Below is GenStat program doing the calculations.

```
scalar [8] ndata
scalar [2000] nboots
variate [nvalues=ndata] z ,y
read z
8406 2342 8187 8459 4795 3516 4796 10238:
read y
-1200 2601 -2705 1982 -1290 351 -638 -2719:
calc thetahat = mean(y)/mean(z)

variate [nvalues = nboots] thetastar
variate [nvalues = ndata] bootsample1, bootsample2, bootset

for i=1...nboots
    calc seed=43*i+13
    calc bootset = urand(seed)
    calc bootset=int(ndata*bootset+1)
    calc bootsample1$[1...ndata]= y$[bootset$[1...ndata]]
    calc bootsample2$[1...ndata]= z$[bootset$[1...ndata]]
    calc thetastar$[i]=mean(bootsample1)/mean(bootsample2)
endfor

hist thetastar
calc bootbias = mean(thetastar) - thetahat
calc bootvar = var(thetastar)
calc bootse = sqrt(bootvar)

print bootbias, bootvar, bootse, thetahat

calc a = 2*bootse/sqrt(nboots)
print a
```

Here are the results:

```
Histogram of thetastar

        -0.24   28 **
-0.24 - -0.16  292 ************************
-0.16 - -0.08  617 ***************************************************
-0.08 -  0.00  551 *********************************************
 0.00 -  0.08  347 *****************************
 0.08 -  0.16  108 *********
 0.16 -  0.24   42 ****
 0.24 -  0.32   14 *
 0.32 -  0.40    1
 0.40 -          0

    bootbias     bootvar      bootse    thetahat       a
    0.009905     0.01013      0.1007    -0.07131    0.004501
```

The value a is very small which indicates that the estimate of the bias is good. Also, comparing the bootstrap estimate of the bias of an estimator to the bootstrap estimate of its standard error gives some information on the quality of the estimator. Here we have:

$$\frac{\widehat{Bias}_{\hat{\theta}}(\theta^*)}{\widehat{se}_{\hat{F}}(\theta^*)} = \frac{0.009905}{0.1007} = 0.0983615,$$

which is rather small, again indicating a good estimate of the bias.

Now, we may correct our estimate of θ. By definition $Bias_\theta(\hat{\theta}) = E(\hat{\theta}) - \theta$, so a better estimate of the parameter might be

$$\hat{\theta} - \widehat{Bias}_{\hat{\theta}}(\theta^*) = -0.07131 - 0.009905 = -0.081215$$

whose absolute value is still less than the maximum of 0.2 allowed for the new patch to be considered equivalent to the old one.

4.3 The Jackknife

The Jackknife is one of the oldest resampling methods. Here we get replications of an estimator $\hat{\theta}$ by constructing new samples simply omitting one observation at a time. So, we get n samples of size $n-1$. Here is the procedure:

Empirical distribution	Jackknife samples of size $n-1$		Jackknife replications of $\hat{\theta}$	Jackknife estimates of
	$\{x_2, x_3, \ldots, x_n\}^*$	$\rightarrow$	$\theta^*_{(1)}$	
$\hat{F}$	$\{x_1, x_3 \ldots, x_n\}^*$	$\rightarrow$	$\theta^*_{(2)}$	
$\{x_1, \ldots, x_n\}$	$\vdots$	$\vdots$	$\vdots$	
	$\{x_1, \ldots, x_{n-1}\}^*$	$\rightarrow$	$\theta^*_{(n)}$	$\Downarrow$
				bias:
				$(n-1)(\bar{\theta}^* - \hat{\theta})$
				where $\bar{\theta}^* = \frac{1}{n}\sum_{i=1}^n \theta^*_{(i)}$
				variance:
				$\frac{n-1}{n}\sum_{i=1}^n (\theta^*_{(i)} - \bar{\theta}^*)^2$

Here, $\theta^*_{(i)}$ is calculated in the same way as $\hat{\theta}$ except that the $i-th$ observation is omitted. The Jackknife estimator of bias and variance of $\hat{\theta}$ are defined to be:

$$Bias_{\hat{\theta}}(\theta^*_{Jack}) = (n-1)(\bar{\theta}^* - \hat{\theta}) \tag{4.6}$$

$$var_{\hat{\theta}}(\theta^*_{Jack}) = \frac{n-1}{n}\sum_{i=1}^{n}(\theta^*_{(i)} - \bar{\theta}^*)^2 \tag{4.7}$$

For simple types of θ the Jackknife estimator can be calculated explicitly.

Example: Jackknife estimator of the mean and of its variance

Mean
Let θ be the expected value of a r.v. X with a c.d.f. F and let the estimator of θ be the average of a random sample of size n, i.e., $\hat{\theta} = \bar{X}$. The Jackknife replications of $\hat{\theta}$ are calculated as:

$$\theta^*_{(i)} = \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j.$$

Is the Jackknife estimate of bias of the mean equal to zero?

$$\bar{\theta}^* = \frac{1}{n}\sum_{i=1}^{n}\theta^*_{(i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} X_j =$$

$$\frac{1}{n}\frac{1}{n-1}(n-1)\sum_{i=1}^{n} X_i = \bar{X} = \hat{\theta}.$$

The Jackknife estimate of bias is

$$Bias_{\hat{\theta}}(\theta^*_{Jack}) = (n-1)(\bar{\theta}^* - \hat{\theta}) = (n-1)(\bar{X} - \bar{X}) = 0.$$

Variance of the mean
Here we calculate the Jackknife variance of the mean.

$$var_{\hat{\theta}}(\theta^*_{Jack}) = \frac{n-1}{n}\sum_{i=1}^{n}(\theta^*_{(i)} - \bar{\theta}^*)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j=1, j\neq i}^{n} X_j - \bar{X}\right)^2 =$$

$$\frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}(n\bar{X} - X_i) - \bar{X}\right)^2 = \frac{n-1}{n}\sum_{i=1}^{n}\frac{(\bar{X} - X_i)^2}{(n-1)^2} =$$

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}S^2.$$

So, the Jackknife estimator of the variance of the mean is the familiar one.

Example: Opinion survey
An opinion survey asked a random sample of 200 people a *yes/no* question of whom 75 answered *yes*. The estimate of the population proportion p of those who would answer *yes* is estimated as $\hat{p} = \frac{75}{200} = \frac{3}{8}$. A social science researcher is interested in a parameter

$$\theta = p(1-p) = pq.$$

A natural estimate of the parameter is

$$\hat{\theta} = \hat{p}\hat{q} = \frac{3}{8}\frac{5}{8} \cong 0.2344.$$

Is the estimator $\hat{\theta} = \hat{p}\hat{q}$ biased?

If we knew that the probability of answering *yes* is the same for each person, then we could use the Bernoulli(p) distribution and calculate the bias. However, it may be rather unlikely that all people have the same attitude to the questioned problem and so this assumption may not be feasible. Then a non-parametric method can help. Here, we calculate the Jackknife estimate of bias of $\hat{\theta} = \hat{p}\hat{q}$.

Let a random variable X_i have two values: 1 if the answer is *yes* and 0 if the answer is *no*. Then the sum

$$\sum_{i=1}^{n} X_i$$

is the number of *yes* answers in the survey. Also, note that

$$n\hat{p} = \sum_{i=1}^{n} X_i.$$

Then the i-th Jackknife replication of $\hat{\theta}$, which is based on the sample of size $n-1$, can be written as

$$\theta^*_{(i)} = \begin{cases} \frac{n\hat{p}-1}{n-1}\left(1-\frac{n\hat{p}-1}{n-1}\right) & \text{if } X_i = 1 \\ \frac{n\hat{p}}{n-1}\left(1-\frac{n\hat{p}}{n-1}\right) & \text{if } X_i = 0. \end{cases}$$

So, we have

$$\bar{\theta}^* = \frac{1}{n}\sum_{i=1}^{n}\theta^*_{(i)} = \frac{1}{n}\left[n\hat{p}\frac{n\hat{p}-1}{n-1}\left(1-\frac{n\hat{p}-1}{n-1}\right) + (n-n\hat{p})\frac{n\hat{p}}{n-1}\left(1-\frac{n\hat{p}}{n-1}\right)\right]$$

Note that in the above formula $n\hat{p}$ is the number of ones and $n-n\hat{p}$ is the number of zeros. Simplifying the above formula we get

$$\bar{\theta}^* = \frac{n(n-2)}{(n-1)^2}\hat{p}(1-\hat{p}) = \frac{n(n-2)}{(n-1)^2}\hat{p}\hat{q}.$$

Hence, the Jackknife estimator of bias of parameter $\hat{\theta}$ is

$$Bias_{\hat{\theta}}(\theta^*_{Jack}) = (n-1)(\bar{\theta}^* - \hat{\theta}) = (n-1)\left(\frac{n(n-2)}{(n-1)^2}\hat{p}\hat{q} - \hat{p}\hat{q}\right) = -\frac{\hat{p}\hat{q}}{n-1} = -\frac{\hat{\theta}}{n-1}.$$

So, in our example we get

$$\widehat{Bias}_{\hat{\theta}}(\theta^*_{Jack}) = -\frac{1}{199} \times 0.2344 \cong -0.0012.$$

Now we may correct the initial estimate of θ by the estimate of the bias to obtain

$$\hat{\theta}_{new} = 0.2344 - (-0.0012) = 0.2356.$$

4.4 The parametric bootstrap

The parametric bootstrap is a useful method for bias and variance estimation when we know that the sample comes from a population with a c.d.f. F, which is a member of a parametric family of distributions indexed by an unknown parameter, or set of parameters, ϕ:

$$F \in \{F_\phi : \phi \in \Phi\}.$$

For example

- F belongs to the family of normal distributions, i.e.,

  $$F \in \{F_{(\mu,\sigma^2)} : \mu \in R, \sigma^2 \in R_+\},$$

  where $F_{(\mu,\sigma^2)}$ is a c.d.f. of a normal r.v. with expected value μ and variance σ^2. Here $\phi = (\mu, \sigma^2)$.

- F belongs to the family of exponential distributions, i.e.,

  $$F \in \{F_\lambda : \lambda > 0\},$$

  where F_λ is a c.d.f. of an exponential r.v. Here $\phi = \lambda$.

In such situation, the empirical distribution $\hat{F}$ is a member of the family with parameter $\hat{\phi}$ - an estimate of ϕ. Hence, to calculate bias and variance of some parameter θ we simulate bootstrap samples from the empirical distribution $\hat{F} \equiv F_{\hat{\phi}}$ and calculate bootstrap replications of $\hat{\theta}$. Note that θ may be the same as ϕ, but is may also be a function of ϕ. The idea of the procedure of the parametric bootstrap is given below:

True distribution		Empirical distribution		Simulated distribution
F_ϕ	$\overrightarrow{data, est. of \phi}$	$F_{\hat{\phi}}$	$\overrightarrow{sampling from F_{\hat{\phi}}}$	$F^*_{\hat{\phi}}$
$\downarrow$		$\downarrow$		$\downarrow$
θ_{F_ϕ}		$\theta_{F_{\hat{\phi}}}$		$\theta_{F^*{}_{\hat{\phi}}}$

Example: Empirical exponential distribution
Let$X_i \sim Exp(\lambda)$, for all $i = 1, \ldots, n$, where λ is unknown. The c.d.f. F and the density function f of an exponential r.v. are following:

$$F(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

$$f(x) = \begin{cases} 0 & \text{if } x \leqslant 0 \\ \lambda e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

We know that $E(X_i) = \frac{1}{\lambda}$ and $\bar{X}$ is an unbiased estimator of $\frac{1}{\lambda}$. Hence, $\hat{\lambda} = \frac{1}{\bar{x}}$ may be used to get the empirical distribution, i.e., $F_{\hat{\lambda}} = F_{\frac{1}{\bar{x}}}$. So, we may generate samples from $Exp(\frac{1}{\bar{x}})$. Next section explains how to do it.

4.4.1 Simulation of random variables

Continuous random variables

Lemma 4.1 (See Lemma 3.1)
Let $\tilde{F}$ be a c.d.f. of a continuous r.v. and let $U \sim Uniform[0, 1]$. Random variable $Z = \tilde{F}^{-1}(U)$ has the c.d.f. $\tilde{F}$.

Proof
We want to show that $P(Z \leqslant z) = \tilde{F}(z)$.

$$P(Z \leqslant z) = P(\tilde{F}^{-1}(U) < z) = P\left(\tilde{F}(\tilde{F}^{-1}(U)) < \tilde{F}(z)\right) = P(U < \tilde{F}(z)).$$

Now, $0 \leqslant \tilde{F}(z) \leqslant 1$ and $P(U \leqslant u) = u$ if $u \in [0, 1]$. So,

$$P(U \leqslant \tilde{F}(z)) = \tilde{F}(z)$$

and

$$P\left(\tilde{F}^{-1}(U) \leqslant \tilde{F}^{-1}(\tilde{F}(z))\right) = \tilde{F}(z)$$

That is

$$P(Z \leqslant z) = \tilde{F}(z).$$

□

GenStat, like many statistical packages, simulates independent observations of the r.v. $U \sim Uniform[0, 1]$. The directive is URAND(seed).

Suppose we need simulations of a continuous r.v. Z with a c.d.f. F. We obtain simulations of uniform r.v. U and transform it to $Z = F^{-1}(U)$.

Another GenStat command GRANDOM may be used for simulation of samples from populations with some distributions.

Example: Simulating exponential distribution
Let $Z \sim Exp(\lambda)$. Put

$$u = 1 - e^{-\lambda z}.$$

Then

$$z = -\frac{1}{\lambda} ln(1-u).$$

So,

$$Z = F_{\lambda}^{-1}(u) = -\frac{1}{\lambda} ln(1-u).$$

Hence, we simulate u from the uniform distribution on interval $[0, 1]$ and calculate $z = -\frac{1}{\lambda} ln(1-u)$.

Assume that we do not know λ, but we have an estimate $\hat{\lambda} = 2$. Check that for the following independent values of the uniform r.v. U we get the respective values of the $Exp(2)$ r.v. Z:

u	z
0.6317	0.5068
0.1897	0.1052
0.1824	0.1007
0.9322	1.3455
0.4338	0.2844
0.5839	0.4374

Discrete random variables

1. $Z \sim$ Bernoulli(p)

z	1	0
$P(Z = z)$	p	$1-p$

Generate u from $Uniform[0, 1]$ and put $z = h(u) = \begin{cases} 1 & \text{if } u \leqslant p \\ 0 & \text{if } u > p \end{cases}$

2. $Z \sim Binomial(n, p)$

   Z is a sum of n independent Bernoulli(p) r.vs so, we generate $u_1, \ldots, u_n$, all from $Uniform[0, 1]$ and put $Z = \sum_{i=1}^{n} h(u_i)$.

3. $Z \sim Discrete\ Uniform$

z	0	1	...	k
$P(Z = z)$	$\frac{1}{k}$	$\frac{1}{k}$	...	$\frac{1}{k}$

   Generate u from $Uniform[0, 1]$ and put $z = h(u) = int(ku)$.

4.4.2 Example: parametric bootstrap

An electronic component is known to have a useful life represented by an exponential density with rate λ failures per hour. The mean time to failure is thus $E(X) = \frac{1}{\lambda}$. A series of 100 observations gave an estimate of the mean $\bar{x} = 50$ hours. There were 35 components which had a shorter life than a day. An engineer is interested in the fraction of the components that would fail in less than 24 hours.

Examine the GenStat output given below and answer the following questions:

1. What is the parameter of interest to the engineer?
2. What method is used to improve the initial estimate of the parameter? Explain why this method is used.
3. Briefly describe the purpose of the commands in lines
   - 1 - 5
   - 6
   - 7 - 10
   - 11 - 12
   - 13 - 19
4. Comment on the results of the output.

GenStat output:

```
1  scalar [100] ndata
2  scalar [500] nboots
3  variate [nvalues = nboots] thetastar
4  pointer [nvalues=nboots] bsamplep
```

```
  5  variate [nvalues=ndata]  bsamplep[]
  6  calc  thetahat = 0.35
  7  for i=1...nboots
  8    calc s = 2*i+5
  9    GRANDOM [DISTRIBUTION=exponential; NVALUES=100; SEED=s;\
              MEAN=50; VARIANCE=2500] bsamplep[i]
 10  endfor
 11  calc thetastar$[1...nboots] = 1-exp((-1/mean(bsamplep[]))*24)
 12  histogram thetastar

Histogram of thetastar

      - 0.275    0
0.275 - 0.300    1
0.300 - 0.325    6 **
0.325 - 0.350   49 ************
0.350 - 0.375  128 ********************************
0.375 - 0.400  163 ****************************************
0.400 - 0.425   96 ************************
0.425 - 0.450   45 ***********
0.450 - 0.475   10 ***
0.475 -          2 *

 13  calc bootheta = mean(thetastar)
 14  calculate bootbias = bootheta - thetahat
 15  calculate bootvar = var(thetastar)
 16  calc bootse = sqrt(bootvar)
 17  calc a = bootbias/bootse
 18  calc newtheta = thetahat - bootbias
 19  print bootheta, thetahat, bootbias, bootse, a, newtheta

  bootheta    thetahat    bootbias    bootse        a    newtheta
    0.3866      0.3500     0.03661     0.03020    1.212     0.3134
```

Here are the answers to the above questions:

1. The parameter of interest is $\theta = P(X < 24)$, where $X \sim Exp(\lambda)$.

2. The method is 'parametric bootstrap'. It is used because we know what parametric family of distributions is involved, although we do not know the value of the parameter which indexes the family, i.e., we

do not know λ. However, we have an estimate of λ and we may simulate values of the parameter of interest θ from the empirical distribution.

3. Lines 1 - 5 declare scalars and pointers.

4. Line 6 assigns value 0.35 to the initial estimate of θ.

5. Lines 7 -10 generate random samples of size 100 from the exponential distribution with mean 50 and save them in variables bsamplep[i]. This is done 500 times in the loop FOR.

6. Lines 11 - 12 calculate replications of the estimate of the parameter θ, i.e., $\widehat{P(X < 24)} = \theta_i^* = 1 - e^{-\frac{1}{\bar{x}_i^*}24}$ for the samples from $Exp(\frac{1}{\bar{x}})$ and draw a histogram of the estimates.

7. Lines 13 - 19:

   - line 13 calculates the average of the bootstrap replications of $\hat{\theta}$, i.e., $\bar{\theta}^*$;
   - lines 14, 15 and 16 calculate bootstrap estimates of bias; variance and standard error of $\hat{\theta}$, respectively;
   - line 17 calculates the ratio of the bootstrap estimates of bias and standard error;
   - line 18 calculates the corrected value of the estimate;
   - line 19 prints out the calculated values.

8. The histogram of the replications θ_i^* and the calculated bootstrap bias show that the initial estimate may not be accurate, or the empirical distribution is not quite right.

4.5 The smoothed bootstrap

In the simple nonparametric bootstrap we have assumed that the empirical distribution assigning equal mass to each observation, $\hat{F}$, is a suitable estimate of F. However, $\hat{F}$ is discrete and it is natural to ask if a smooth estimate of F might be better, particularly when we expect F to be continuous.

$\hat{F}$ is the c.d.f. which places an atom of probability with mass $\frac{1}{n}$ to each data point. Smoothing consists of replacing each data point with a continuous distribution of total mass $\frac{1}{n}$ centered at the point. The most common smoothing

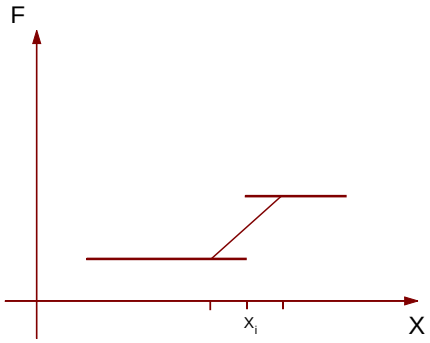

Figure 4.1: Smoothing the empirical distribution function.

distribution is uniform on interval $[-h, h]$. The uniformly smoothed empirical c.d.f. $\hat{F}^U$ is similar to $\hat{F}$ except that the jumps of size $\frac{1}{n}$ at each data point are replaced by straight lines with slope $\frac{1}{2nh}$ which pass through the midpoint of the jump. See Figure (4.1).

Another common smoothing distribution is $N(x_i, h^2)$.

In any case, there is the question of the choice of h. If it is too small, then the resulting distribution will not be very smooth, if it is too large, then the smoothed portions overlap and we loose information given in the original data.

Simulation from $\hat{F}^U$ proceeds in two stages:

1. generate a bootstrap sample in the usual way,
2. add a simulated r.v. from $Uniform[-h, h]$ to each member of the bootstrap sample.

Then calculate bias and variance of and estimator θ as in the simple bootstrap.

4.6 The balanced bootstrap

Since bootstrap samples are chosen randomly and independently, 'unrepresentative' collections of samples may occur, that is some values may occur many more times than other. In a balanced bootstrap, each of the n observations is constrained to occur exactly N times in the N samples. Hence, each bootstrap sample is a random sample from $\hat{F}$ but the samples are no longer

independent (for example, knowing the first $N-1$ samples tell us what the Nth sample must be).

To implement a balanced bootstrap we need a random permutation of the vector

$$(1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots, n, n, \ldots, n)'$$

In the randomized vector we use the first n entries to index the first bootstrap sample, the next n entries to index the second bootstrap sample, and so on. For example, let $n = 10$ and $N = 2$ and let the sample from a population be (see Practical 6) :

9.6 10.4 13.0 15.0 16.6 17.2 17.3 21.8 24.0 33.8

The randomized vector might be

$$(2, 2, 3, 8, 4, 6, 1, 9, 7, 4, 5, 6, 1, 5, 8, 9, 3, 10, 10, 7)'$$

which gives the following two bootstrap samples:

10.4 10.4 13.0 21.8 16.6 17.2 9.6 24.0 17.3 15.0

16.6 17.2 9.6 16.6 21.8 24.0 13.0 33.8 33.8 17.3

Using the frequencies of occurrence we may put the bootstrap samples in the following table:

data	9.6	10.4	13.0	15.0	16.6	17.2	17.3	21.8	24.0	33.8
frequencies	1	2	1	2	0	1	1	1	1	0
frequencies	1	0	1	0	2	1	1	1	1	2

Now, consider $\hat{\theta} = \bar{X}$. The value of the estimator obtained from the original sample is $\bar{x} = 17.87$, the bootstrap replications of the estimate are: $\bar{x}_1^* = 15.37$ and $\bar{x}_2^* = 20.37$ which give the mean of $\bar{x}^* = 17.87$. The balanced bootstrap forces the average value of $\theta_i^* = \bar{x}_i^*$ to be the same as the value of $\hat{\theta} = \bar{X}$.

4.7 Bootstrapping bivariate data

Suppose we have bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, a sample of bivariate i.i.d. random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ each with the same c.d.f. $F_{X,Y}$. The way of bootstrapping this kind of data depends on what we know about the relationship between X and Y.

4.7.1 Non-parametric bootstrap

This method is appropriate when the pairs (x_i, y_i) are the random sample, we have no prior control over the values of r.vs X and Y and the model of Y in terms of X is either unknown or theoretically untractable.

Bootstrapping:

- For each bootstrap sample randomly choose n numbers $j_i, j_2, \ldots, j_n$ from $\{1, 2, \ldots, n\}$ with replacement.
- Then, the bootstrap sample consists on the pairs $(x_{j_1}, y_{j_1}), (x_{j_2}, y_{j_2}), \ldots, (x_{j_n}, y_{j_n})$ (x and y get bootstrapped together).
- Calculate the replications of the parameter estimate for each bootstrap sample and average the replications.

4.7.2 Fully parametric bootstrap

Suppose we know that

$$Y_i = f(X_i, \psi) + \varepsilon_i,$$

where the r.v. ε follows a distribution which belongs to a known parametric family of distributions. For example

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

However, we neither know the values of the parameters ψ nor the error distribution parameters (α, β, σ^2 in the example). If we can control the values of r.v. X then we obtain a bootstrap sample by fixing each value x_i and simulating y_i from the fitted model $\hat{Y}_i$. For example from

$$N(\hat{\alpha} + \hat{\beta} x_i, \hat{\sigma}^2),$$

where $\hat{\alpha}$, $\hat{\beta}$and $\hat{\sigma}^2$ are preliminary estimates of the parameters.

Here we fit the model to the observed data, then simulate random samples from the fitted model, and refit the model with the simulated samples.

4.7.3 Semi-parametric bootstrap

Suppose that we know the expected model, i.e., $E(Y_i) = f(X_i, \psi)$ up to the model parameters ψ but we do not know the distribution of the errors $Y_i - E(Y_i)$. If we can control the values of X then we can bootstrap in the following way:

- Estimate the parameters ψ and obtain the residuals $r_i = y_i - \widehat{E(Y_i)}$.
- Fix the x_i values and bootstrap the residuals, i.e., randomly choose n numbers $j_i, j_2, \ldots, j_n$ from $\{1, 2, \ldots, n\}$ with replacement and put $(x_i, \widehat{E(Y_i)} + r_{j_i})$ for $i = 1, \ldots, n$ as a bootstrap bivariate sample.

Here we fit the expectation part of the model, then resample with replacement from the residuals, add the resampled residuals to the fitted expectation to get bootstrap samples, to which we refit the model.

Example of semi-parametric bootstrap
In recent years, physicians used the *dividing reflex* to reduce abnormally rapid heartbeats in humans by briefly submerging the patient's face in cold water. The reflex, triggered by cold water temperatures, is an involuntary neural response that shuts off circulation to the skin, muscles, and internal organs and diverts extra oxygen-carrying blood to the heart, lungs and brain. A research physician conducted an experiment to investigate the effects of various cold water temperatures on the pulse rate of small children. From his earlier experience, the physician knew that the expected pulse rate may be modeled as a linear function of water temperature, however he had no information about the measurement error distribution.

The relationship between X (water temperature) and Y (pulse rate) is assumed to be linear, so

$$E(Y) = \alpha + \beta X.$$

However, there is no information about the error distribution. We will use semi-parametric bootstrap.

The following GenStat program calculates bootstrap approximation of bias and of variance of the estimators of the slope β and the intersection α. Initial estimates are calculated (given in the output) and used to obtain residuals. The residuals are bootstrapped and the fitted values are corrected by the residuals. Then, the bootstrap replications of the estimates of the slope and the intersection are calculated. Finally the bias and the variance are obtained.

```
 1  scalar [7] ndata
 2  scalar [1000] nboots
 3  variate [nvalues=ndata] pulse, temp
 4  read pulse

      Identifier   Minimum      Mean   Maximum    Values   Missing
          pulse     55.00     63.43     72.00         7         0

 5  read temp

      Identifier   Minimum      Mean   Maximum    Values   Missing
          temp     1.000     7.000     13.00         7         0

 6 graph x=temp; y=pulse

         I
         I
         I                                                 *
    70.0 I
         I                                  *       *
         I
         I                           *
         I
         I
         I
    60.0 I
         I                    *
         I            *
         I    *
         I
         I
    50.0 I
        -+---------+---------+---------+---------+---------+---------+---
         0.0       2.5       5.0       7.5      10.0      12.5      15.0
                             pulse  v. temp using symbol *

 7  model pulse
 8  fit  temp

***** Regression Analysis *****

 Response variate: pulse
     Fitted terms: Constant, temp
```

```
*** Summary of analysis ***
             d.f.      s.s.       m.s.      v.r.
Regression    1       258.04    258.036    110.47
Residual      5        11.68      2.336
Total         6       269.71     44.952

Percentage variance accounted for 94.8
Standard error of observations is estimated to be 1.53

*** Estimates of parameters ***
             estimate        s.e.      t(5)
Constant       52.80         1.16     45.35
temp            1.518        0.144    10.51

 9  scalar slophat, inthat
 10  rkeep fittedvalues = f
 11  rkeep estimates = e
 12  calc resid = pulse - f
 13  calc inthat = e$[1]
 14  calc slophat = e$[2]
 15  print inthat, slophat

      inthat     slophat
       52.80       1.518

 16  variate [nvalues = nboots] slopstar, intstar
 17  matrix [rows=nboots; columns = ndata] bootmat
 18  pointer [nvalues=nboots] bootsubp, bootpulp
 19  variate [nvalues=ndata] bootsubp[], bootpulp[]
 20  calculate bootmat = urand(131)
 21  calculate bootmat = int(ndata*bootmat +1)
 22  equate bootmat; bootsubp
 23  calculate bootpulp[] = f + resid$[bootsubp[]]
 24  for i=1...nboots
 25      model bootpulp[i]
 26      fit [print = *]  temp
 27      rkeep estimates = e
 28      calc intstar$[i] = e$[1]
 29      calc slopstar$[i] = e$[2]
 30  endfor
```

```
 31  hist intstar

Histogram of intstar

      - 49.6    0
 49.6 - 50.4    8 *
 50.4 - 51.2   43 *******
 51.2 - 52.0  157 *************************
 52.0 - 52.8  258 ****************************************
 52.8 - 53.6  290 **********************************************
 53.6 - 54.4  188 ******************************
 54.4 - 55.2   54 *********
 55.2 - 56.0    2
 56.0 -         0
Scale:  1 asterisk represents 6 units.

 32  hist slopstar

Histogram of slopstar

     - 1.1    0
 1.1 - 1.2    2
 1.2 - 1.3   40 *******
 1.3 - 1.4  144 ************************
 1.4 - 1.5  264 *********************************************
 1.5 - 1.6  300 ***************************************************
 1.6 - 1.7  190 *********************************
 1.7 - 1.8   50 ********
 1.8 - 1.9    8 *
 1.9 -        2
Scale:  1 asterisk represents 6 units.

 33  calc slopbias = mean(slopstar) - slophat
 34  calc slopvar = var(slopstar)
 35  calc slopse = sqrt(slopvar)
 36  print slopbias, slopvar, slopse

    slopbias     slopvar      slopse
   -0.004125     0.01512      0.1229
```

```
37  calc intbias = mean(intstar) - inthat
38  calc intvar = var(intstar)
39  calc intse = sqrt(intvar)
40  print intbias, intvar, intse

    intbias      intvar       intse
    0.05197      0.9856      0.9928
```

4.7.4 Summary of the bivariate bootstrap

Relationship between X and Y	X controlled	X not controlled
known apart from the parameters	regression or GLM or parametric bootstrap	bootstrap x_i, then do parametric bootstrap
$E(Y_i)$ known apart from the parameters	fix the x_i and bootstrap the residuals	bootstrap x_i, then bootstrap the residuals
unknown	?	bootstrap the pairs (x_i, y_i)

So far we have used the bootstrap method to assess the properties of estimators based on a random sample. However, a major advantage of the bootstrap is that it can be used in an enormous range of statistical problems, including very complicated ones.

For a complicated statistical model involving many random variables, it is important to distinguish between:

1. non-parametric bootstrap in which we sample with replacement from all the data and estimate unknown parameters for each bootstrap sample,

2. fully parametric boostrap in which we fit the model to the observed data, then simulate random sample from the fitted model and finally refit the model to the simulated samples,

3. semi-parametric bootstrap in which we fit the expectation part of the model, then resample with replacement from the residuals, adding the resampled residuals to the fitted expectation to get new samples, to which we refit the model.

The distinction between the three cases is what we can assume. In the fully parametric bootstrap we must be prepared to assume that the statistical

model is valid for some values of the parameters: the only problem is to find these values. In the semi-parametric bootstrap we must assume that the expectation part of the model is valid, but need not assume anything about the distribution of errors. In the non-parametric bootstrap we need not assume that any part of the model is valid.

If the whole model is valid then the parametric bootstrap will give better estimates.

There is no invariant formula for deciding which method to apply. It is a question of the statistician's judgement based on understanding the model and the underlying phenomenon.

4.8 Cross-validation

When a statistical model has been fitted to data, a good way to test the goodness-of-fit is to assess how well the model predicts any future data. However, often there is no additional data available. If future data become available, it is desirable to refit the model to all data.

Cross-validation provides methods of making use of all the data both to fit the model and to assess the goodness-of-fit. These are resampling methods, computer intensive, requiring refitting the model many times.

Method 1: leave-one-out samples
For each i in turn, omit the $i-th$ datum, fit the model to the $n-1$ remaining data and use the fitted model to predict the outcome at the $i-th$ point.

The cross-validation residual is

$r_i = \textit{predicted value at the } i-th \textit{ point } - \textit{ actual outcome at the the } i-th \textit{ point.}$

If the model fits well, then

$$\sum_{i=1}^{n} r_i^2$$

will be small. To choose between two models, choose the one with the smaller value of $\sum_{i=1}^{n} r_i^2$.

Method 2: construction and test samples
Randomly divide the observed data into two sets: the construction sample

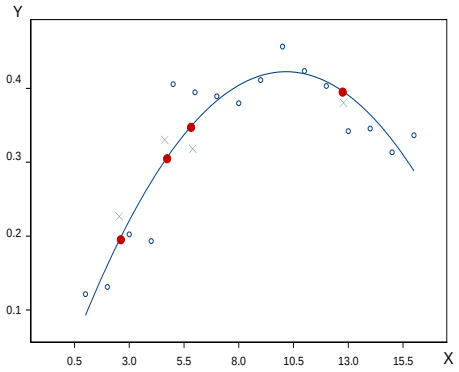

Figure 4.2: Method 2 of cross-validation: open blue circles - construction sample, cross - test sample, solid red points - prediction of the test sample

and the test sample. Fit the model with the construction sample, then test the goodness-of-fit with the test sample. Use the explanatory variables of the test sample to predict the outcomes from the fitted model, then compare them with the observed outcomes. Repeat the procedure, with new random division into construction and test samples, so that every observed data point will occur several times in both construction and test samples.