CENTRE DE RECHERCHE PUBLIC
Microarray Center

the Luxembourg Bioinformatics Network

Introduction to R and Statistical Data Analysis

PART II

Petr Nazarov

petr.nazarov@crp-sante.lu

17-10-2011

OUTLINE PART II

- **Descriptive statistics in R (8)**
  - sum, mean, median, sd, var, cor, *etc*.
- **Principle component analysis and clustering (9)**
  - PCA, k-means clustering, hierarchical clustering
- **Random numbers (10)**
  - random number generators, distributions
- **Statistical tests (11)**
  - t-test, Wilcoxon test, multiple test correction.
- **ANOVA and Linear regression (12)**
  - ANOVA, linear regression

Look for corresponding scripts at
`http://edu.sablab.net/r2010/scripts`

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the geographic variation of Iris flowers in the Gaspé Peninsula.

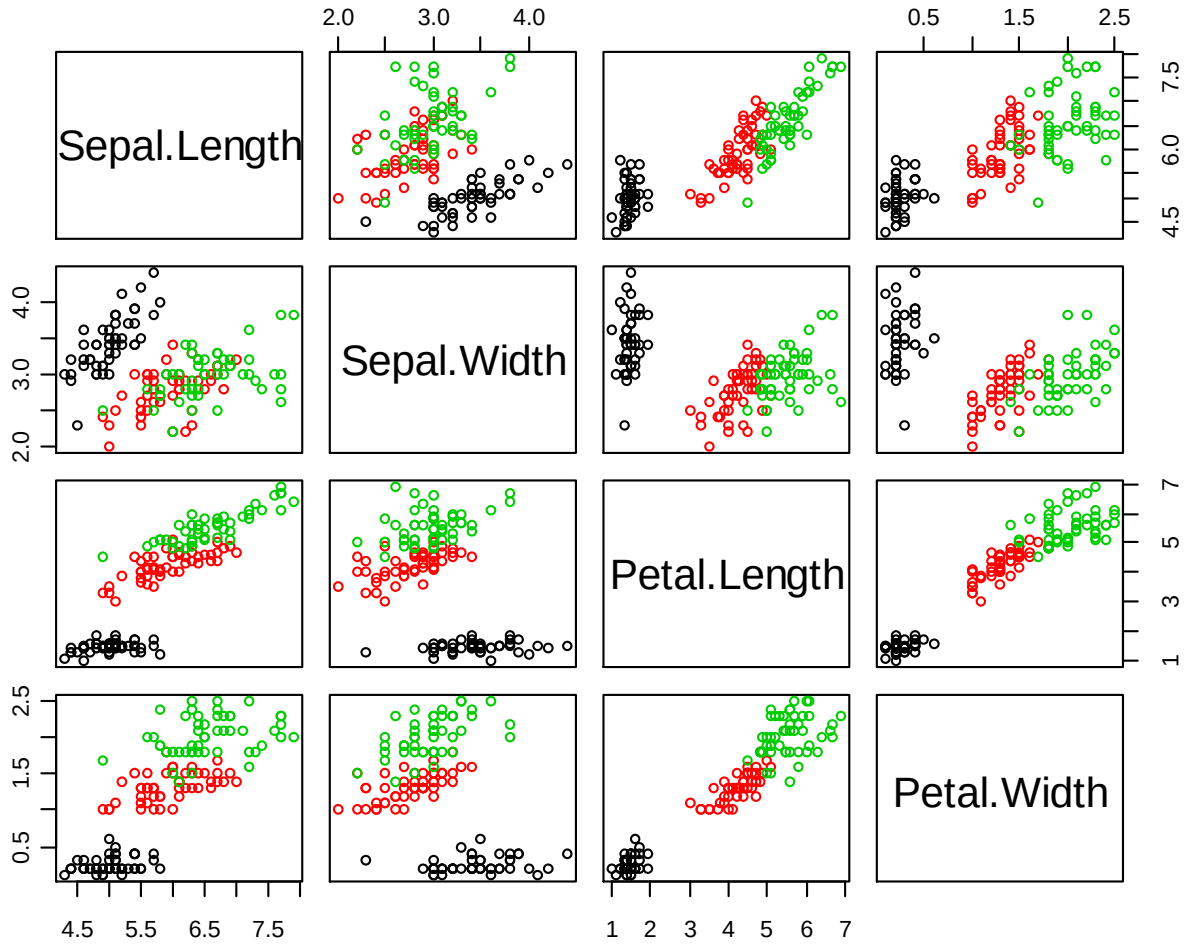
The dataset consists of 50 samples from each of three species of Iris flowers (Iris **setosa**, Iris **virginica** and Iris **versicolor**). Four features were measured from each sample, they are the length and the width of sepal and petal, in centimeters. Based on the combination of the four features, Fisher developed a linear discriminant model to distinguish the species from each other.

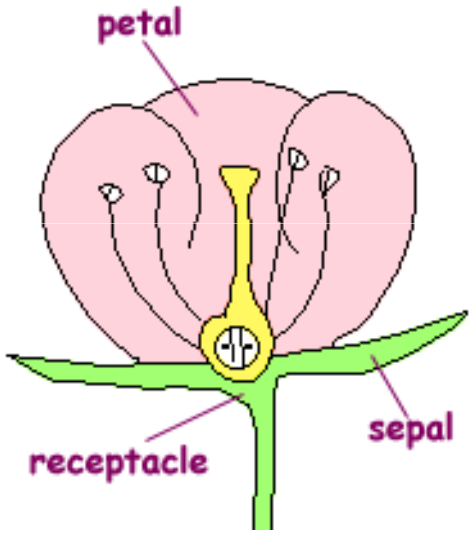
Iris setosa

Iris versicolor

Iris virginica

```
iris
str(iris)
## plot iris data
x11()
plot(iris[,-5])
plot(iris[,-5],
     col = iris[,5])
```



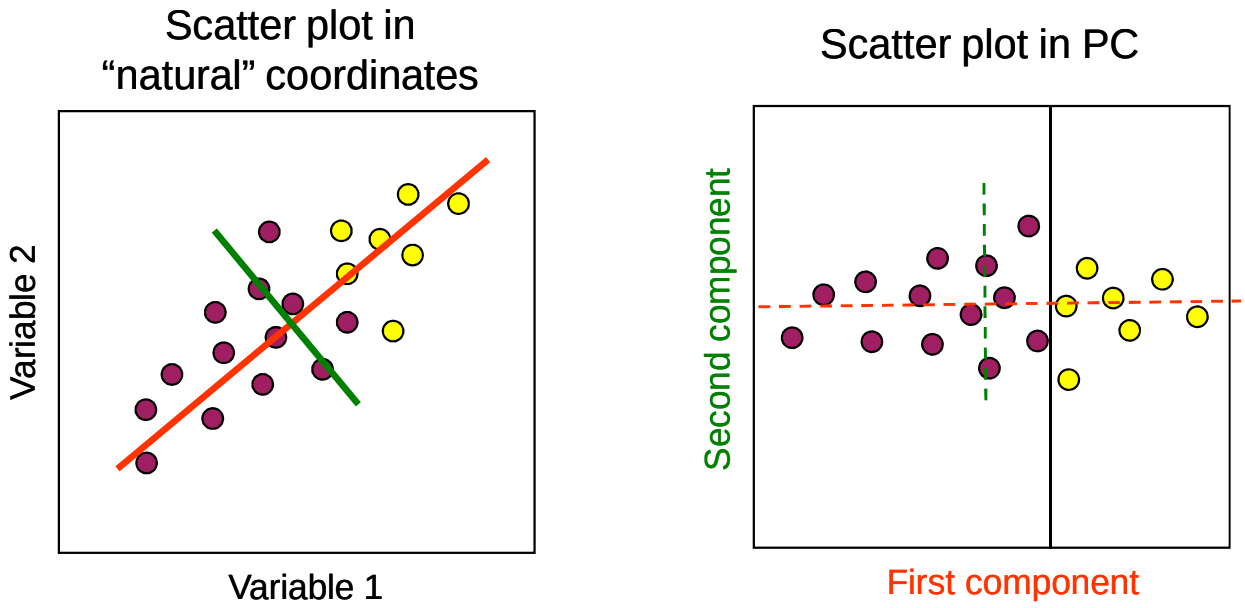


http://urbanext.illinois.edu/gpe/case4/c4facts1a.html

Principal component analysis (PCA)
is a vector space transform used to reduce multidimensional data sets to lower dimensions for analysis. It selects the coordinates along which the variation of the data is bigger.

20000 genes → 2 dimensions

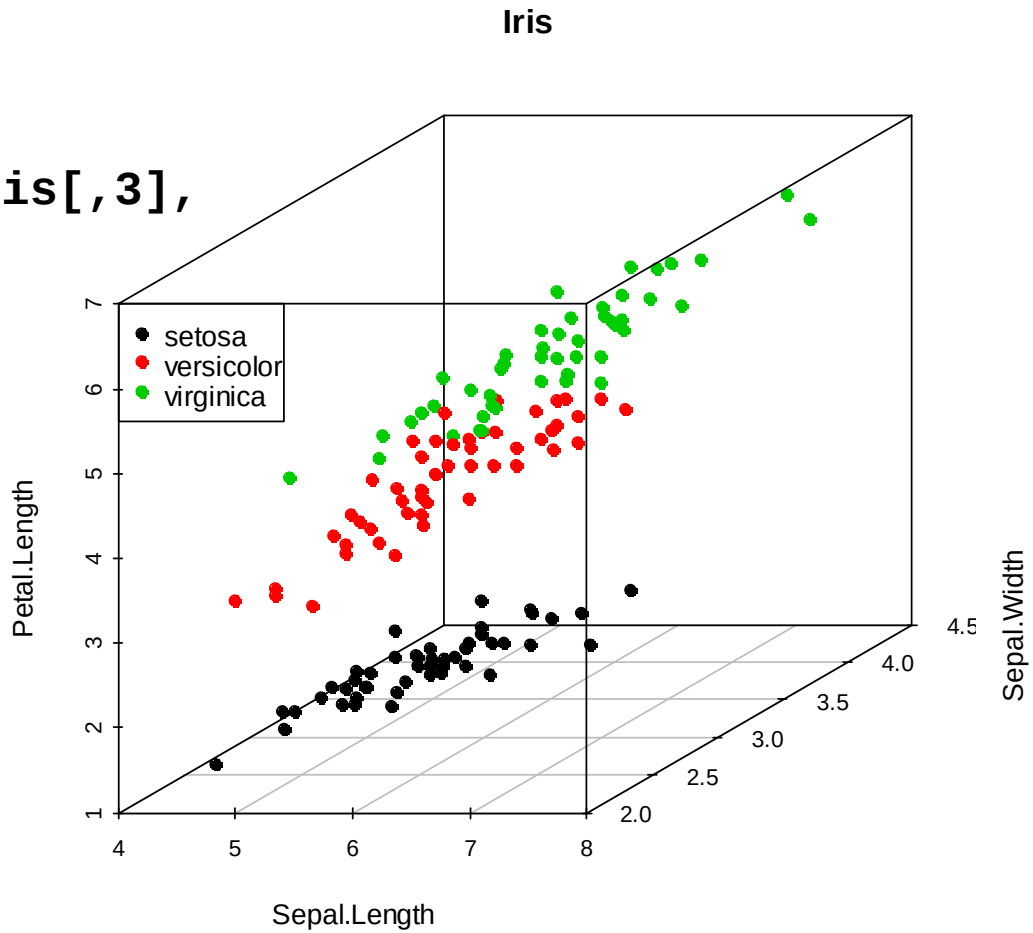
For the simplicity let us consider 2 parametric situation both in terms of data and resulting PCA.



Instead of using 2 “natural” parameters for the classification, we can use the first component!

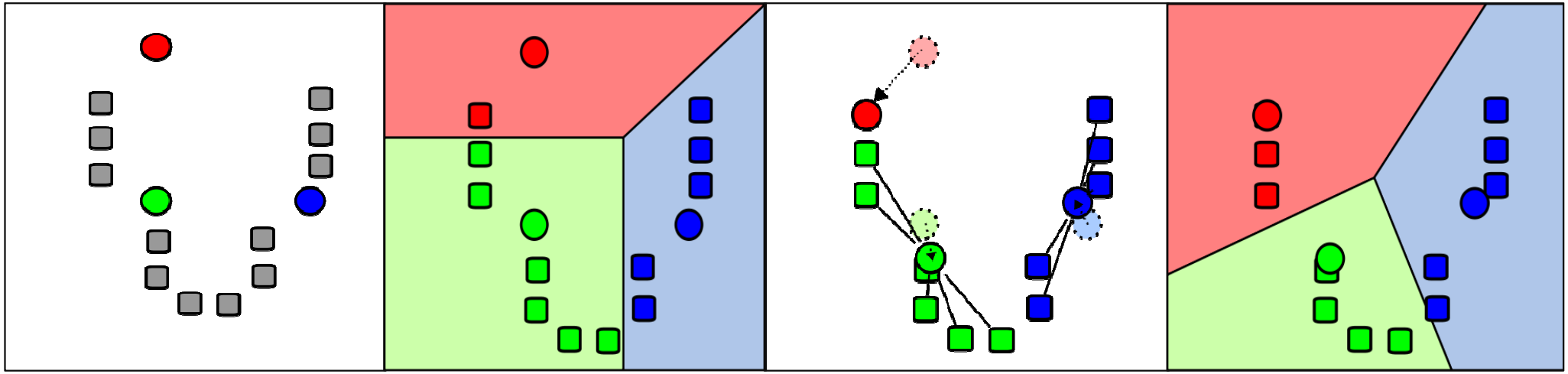
```
Data = as.matrix(iris[,-5])
row.names(Data) = as.character(iris[,5])
classes = as.integer(iris[,5])


## plot data in 3d
library(scatterplot3d)
x11()
scatterplot3d(iris[,1],iris[,2],iris[,3],
     pch=19,color=classes,
     main = "Iris",
     xlab = names(iris)[1],
     ylab = names(iris)[2],
     zlab = names(iris)[3])
legend(4,7,levels(iris$Species),
       col=c(1,2,3),pch=19)
```



k-Means Clustering

k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.



1) *k* initial "means" (in this case k=3) are randomly selected from the data set (shown in color).

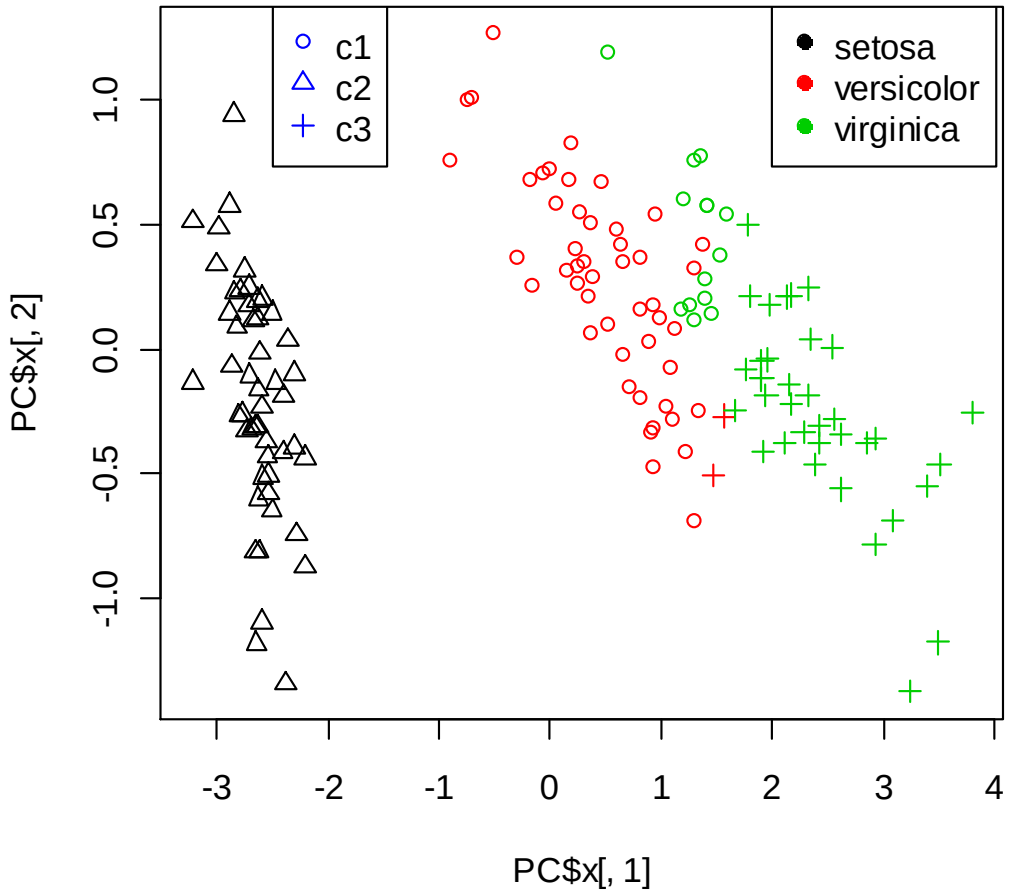
2) *k* clusters are created by associating every observation with the nearest mean.

3) *The centroid of each of the k clusters becomes the new means.*

4) Steps 2 and 3 are repeated until convergence has been reached.

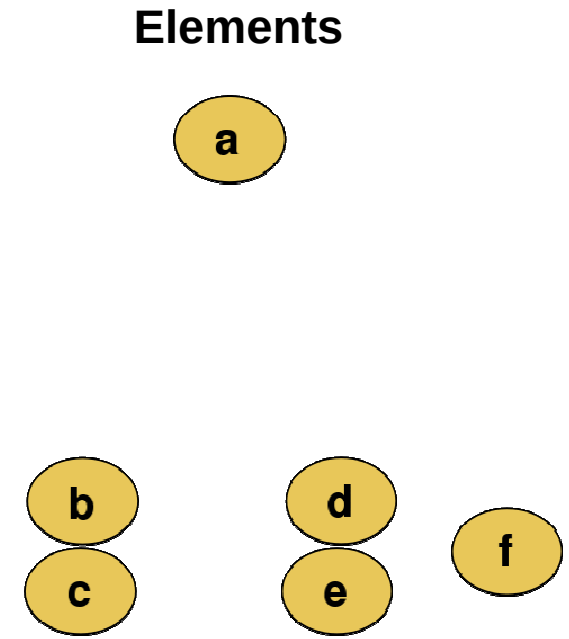
http://wikipedia.org

```
clusters = kmeans(x=Data,centers=3,nstart=10)$cluster
x11()
plot(PC$x[,1],PC$x[,2],col = classes,pch=clusters)
legend(2,1.4,levels(iris$Species),col=c(1,2,3),pch=19)
legend(-2.5,1.4,c("c1","c2","c3"),col=4,pch=c(1,2,3))
```

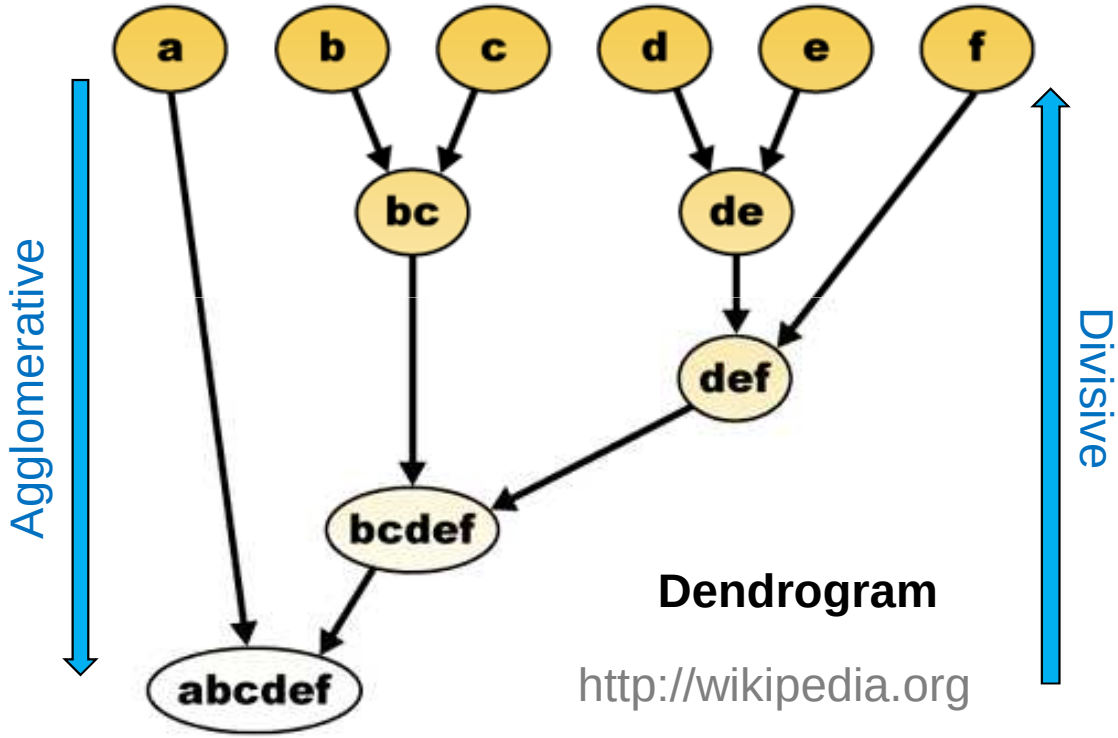


Hierarchical Clustering

Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a **dendrogram.** The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.
Algorithms for hierarchical clustering are generally either **agglomerative**, in which one starts at the leaves and successively merges clusters together; or **divisive**, in which one starts at the root and recursively splits the clusters.

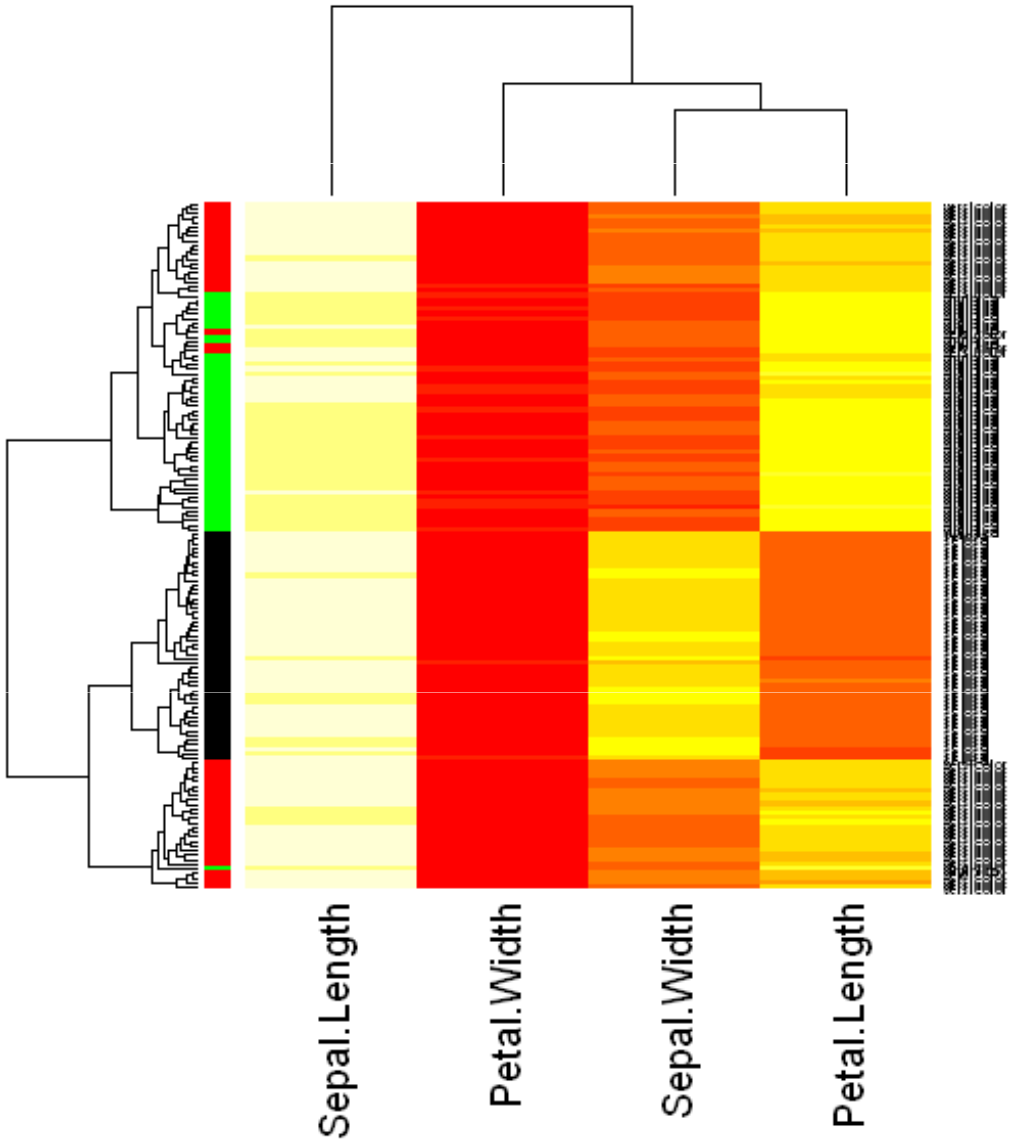


Distance: *Euclidean*



```
## use heatmap
heatmap(data)

## use heatmap with colors
color = character(length(classes))
color[classes == 1] = "black"
color[classes == 2] = "red"
color[classes == 3] = "green"
heatmap(data,RowSideColors=color)
```



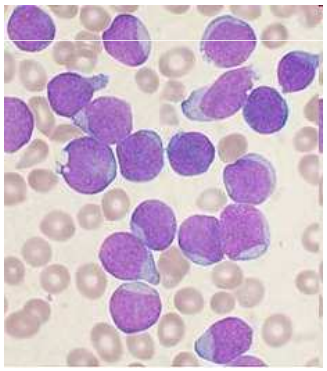
Iris setosa

Iris versicolor

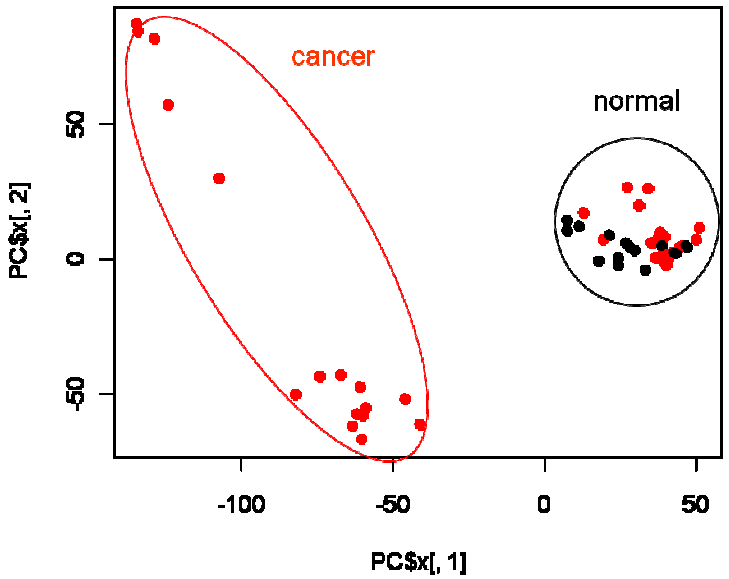
Iris virginica

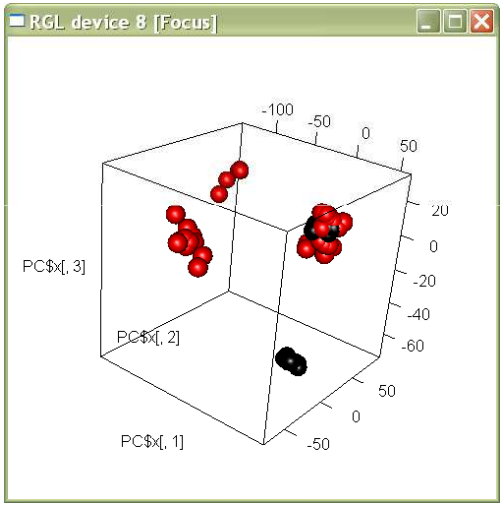
http://edu.sablab.net/data/txt/all_data.txt

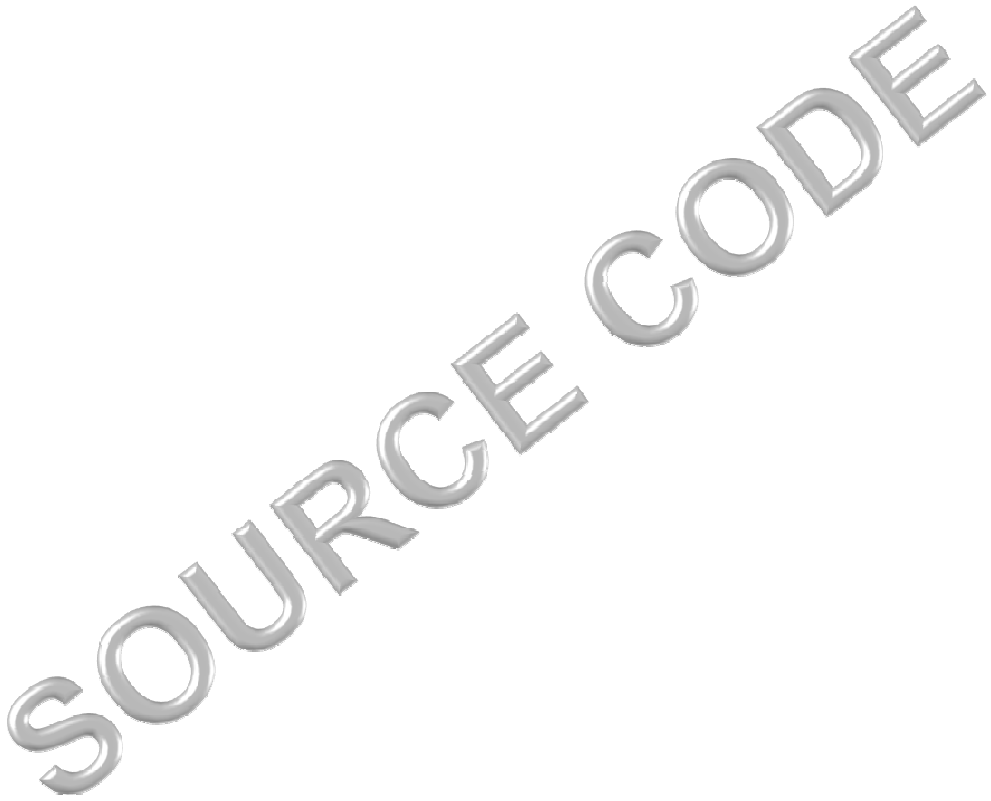
Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts.

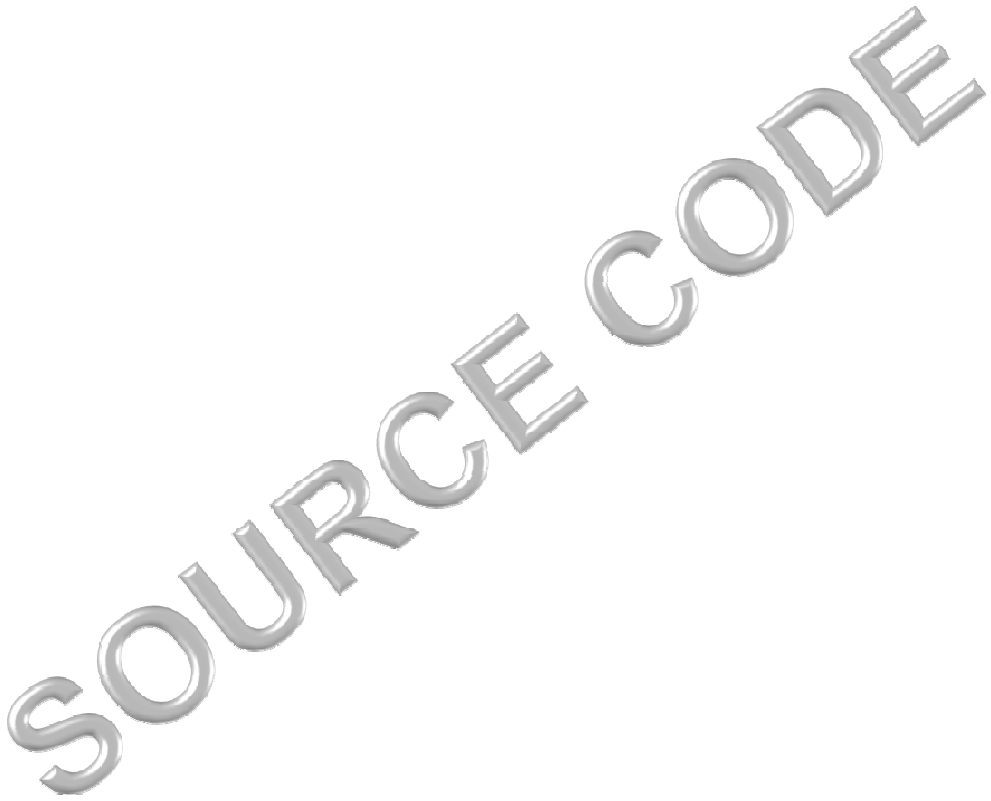


all_data.txt contains the results of full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in $\log_2$ scale.






SOURCE CODE


SOURCE CODE

Means for more than 2 populations

We have measurements for 5 conditions. Are the means for these conditions equal?

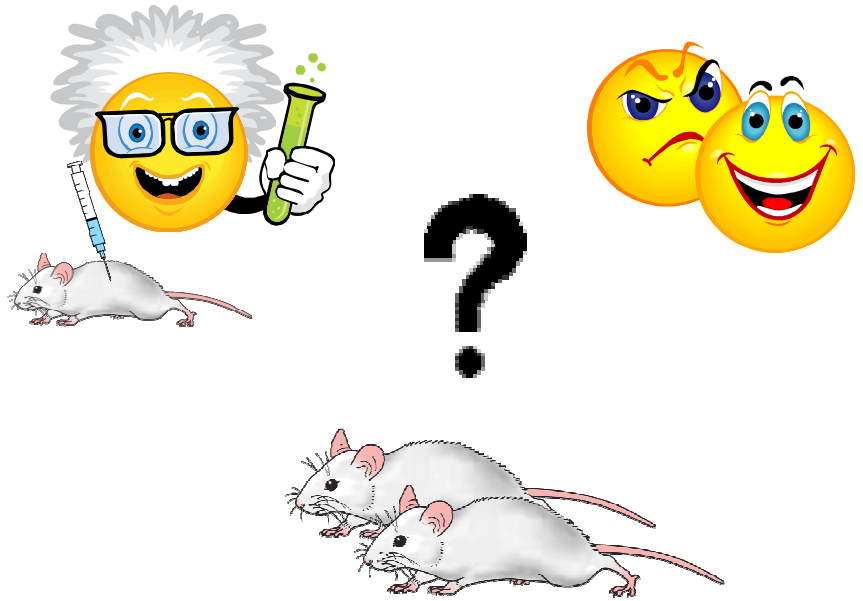
If we would use pairwise comparisons, what will be the probability of getting error?

Number of comparisons: $C_2^5 = \frac{5!}{2!3!} = 10$

Probability of an error: $1-(0.95)^{10} = 0.4$

Validation of the effects

We assume that we have several factors affecting our data. Which factors are most significant? Which can be neglected?

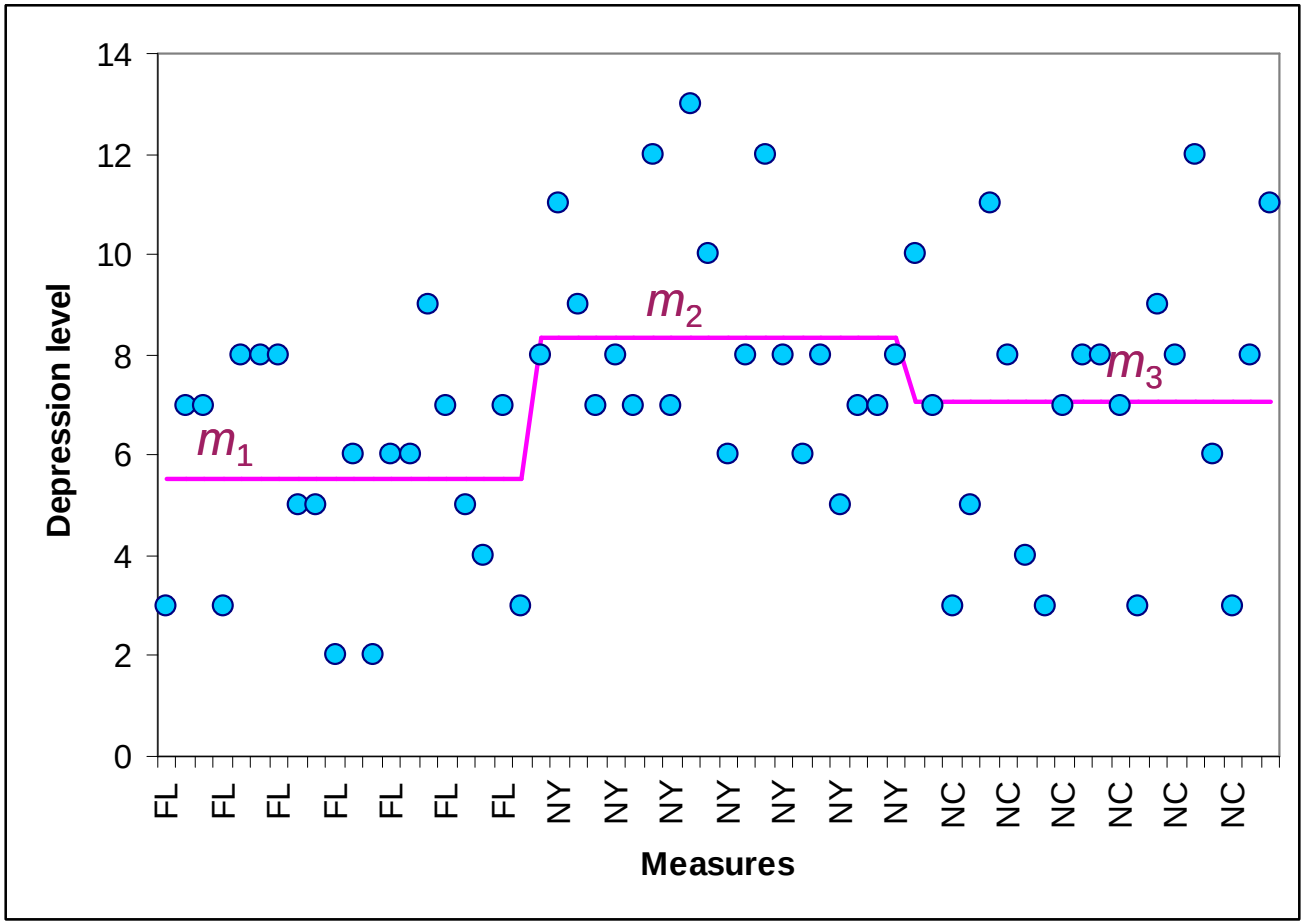


ANOVA example from Partek™

http://easylink.playstream.com/affymetrix/ambsymposium/partek_08.wvx

H_0: $\mu_1 = \mu_2 = \mu_3$

H_a: not all 3 means are equal



salaries.txt

Data Editor

File Edit Help

	Salary.week	Occupation	Gender
1	872	Financial Manager	Male
2	859	Financial Manager	Male
3	1028	Financial Manager	Male
4	1117	Financial Manager	Male
5	1019	Financial Manager	Male
6	519	Financial Manager	Female
7	702	Financial Manager	Female
8	805	Financial Manager	Female
9	558	Financial Manager	Female
10	591	Financial Manager	Female
11	747	Computer Programmer	Male
12	766	Computer Programmer	Male
13	901	Computer Programmer	Male
14	690	Computer Programmer	Male
15	881	Computer Programmer	Male
16	884	Computer Programmer	Female
17	765	Computer Programmer	Female
18	685	Computer Programmer	Female
19	700	Computer Programmer	Female

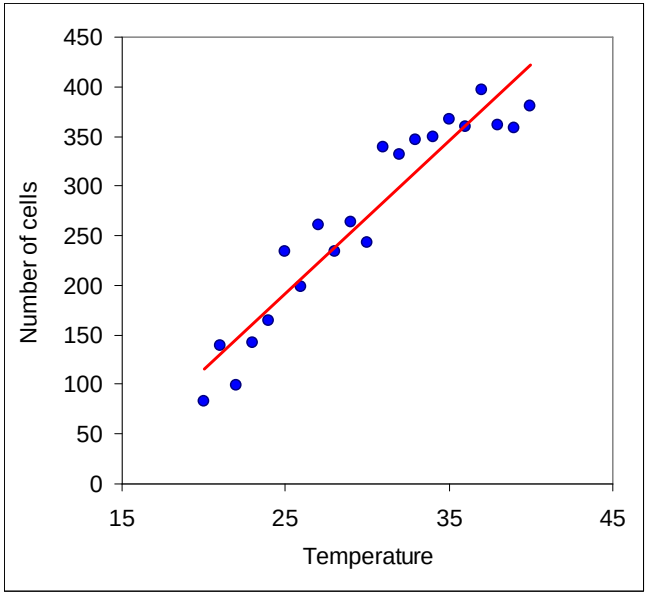
Regression model

The equation describing how y is related to x and an error term; in simple linear regression, the regression model is $y = \beta_0 + \beta_1 x + \varepsilon$

Regression equation

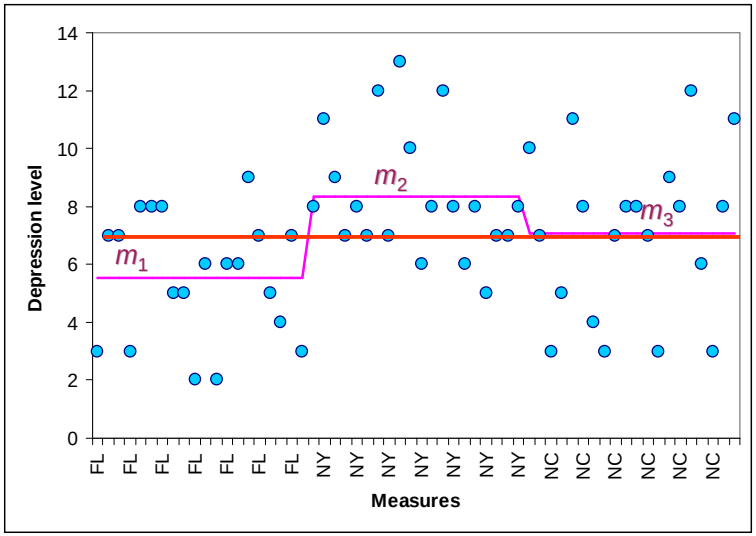
The equation that describes how the mean or expected value of the dependent variable is related to the independent variable; in simple linear regression,

$E(y) = \beta_0 + \beta_1 x$

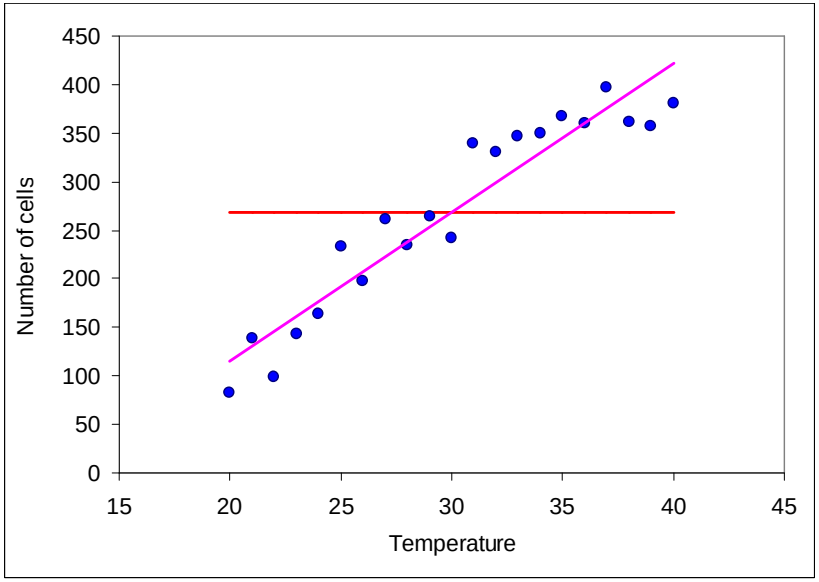


- Model for a simple linear regression:

$$y(x) = \beta_1 x + \beta_0 + \varepsilon$$



$$SST = SSTR + SSE$$



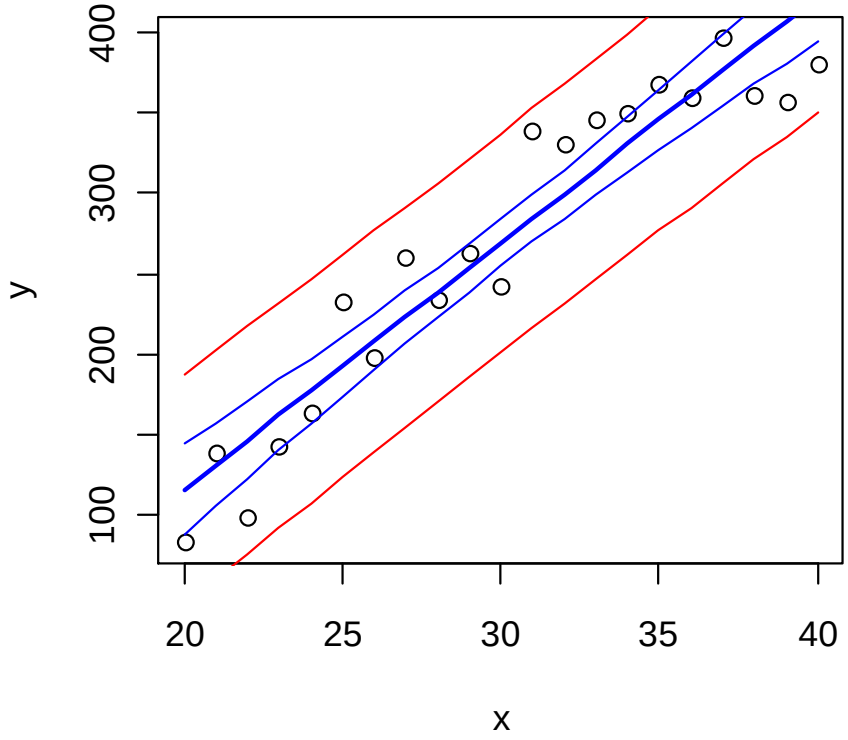
$$SST = SSR + SSE$$

cells.txt

Data Editor

File Edit Help

	Temperature	Cell.Number
2	21	139
3	22	99
4	23	143
5	24	164
6	25	233
7	26	198
8	27	261
9	28	235
10	29	264
11	30	243
12	31	339
13	32	331
14	33	346
15	34	350
16	35	368
17	36	360
18	37	397
19	38	361
20	39	358
21	40	381



SOURCE CODE

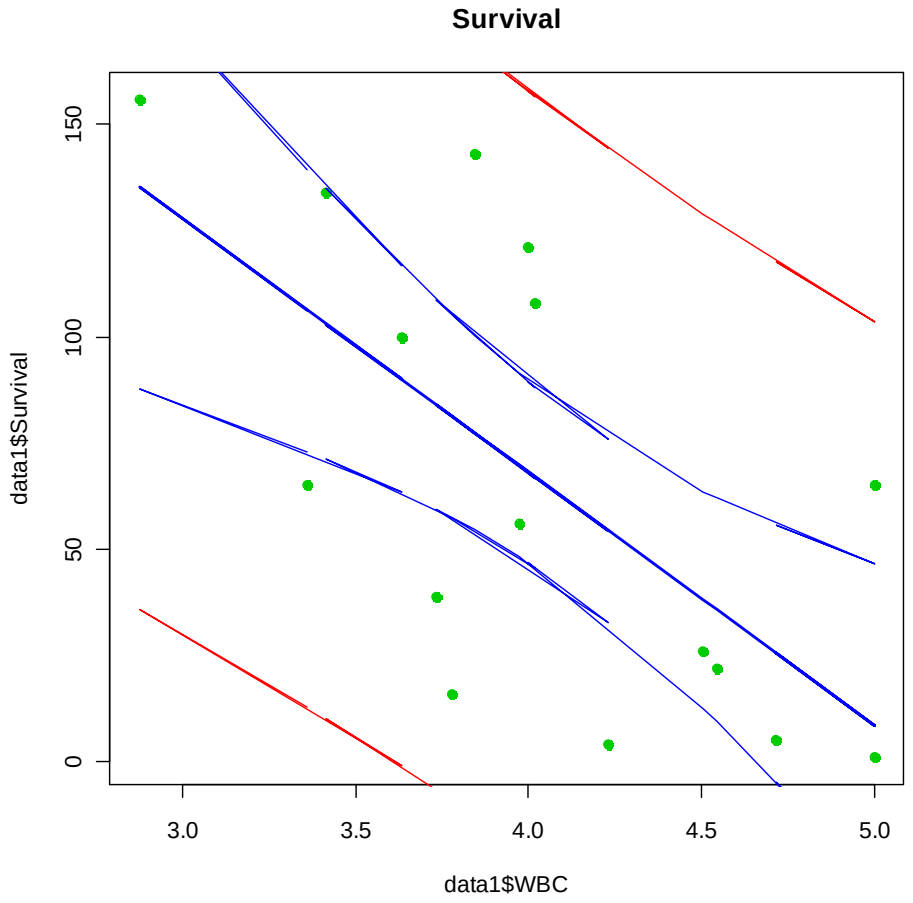
12.3. Solution of Task 12

leukemia.txt

Data Editor

File Edit Help

	WBC	Survival	AG
1	2300	65	Positive
2	750	156	Positive
3	4300	100	Positive
4	2600	134	Positive
5	6000	16	Positive
6	10500	108	Positive
7	10000	121	Positive
8	17000	4	Positive
9	5400	39	Positive
10	7000	143	Positive
11	9400	56	Positive
12	32000	26	Positive
13	35000	22	Positive
14	1e+05	1	Positive
15	1e+05	1	Positive
16	52000	5	Positive
17	1e+05	65	Positive
18	4400	56	Negative
19	3000	65	Negative
20	4000	17	Negative
21	1500	7	Negative
22	9000	16	Negative
23	5300	22	Negative



SOURCE CODE

Thank you for your attention

Questions ?