This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2020.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

An introduction to R: Short Version (2017)

Part 3: statistical modeling

Developed by

Dan Lin (Hasselt University) and Ziv Shkedy (Hasselt University)

LAST UPDATE: 15/10/2017

Overview

1. Statistical modeling in R: simple linear regression.
2. Statistical modeling in R: one-way ANOVA.
3. Statistical modeling in R: logistic regression.

Statistical modeling 1: Simple linear regression

Reading the cars data

The data is available in R, use, help(cars)

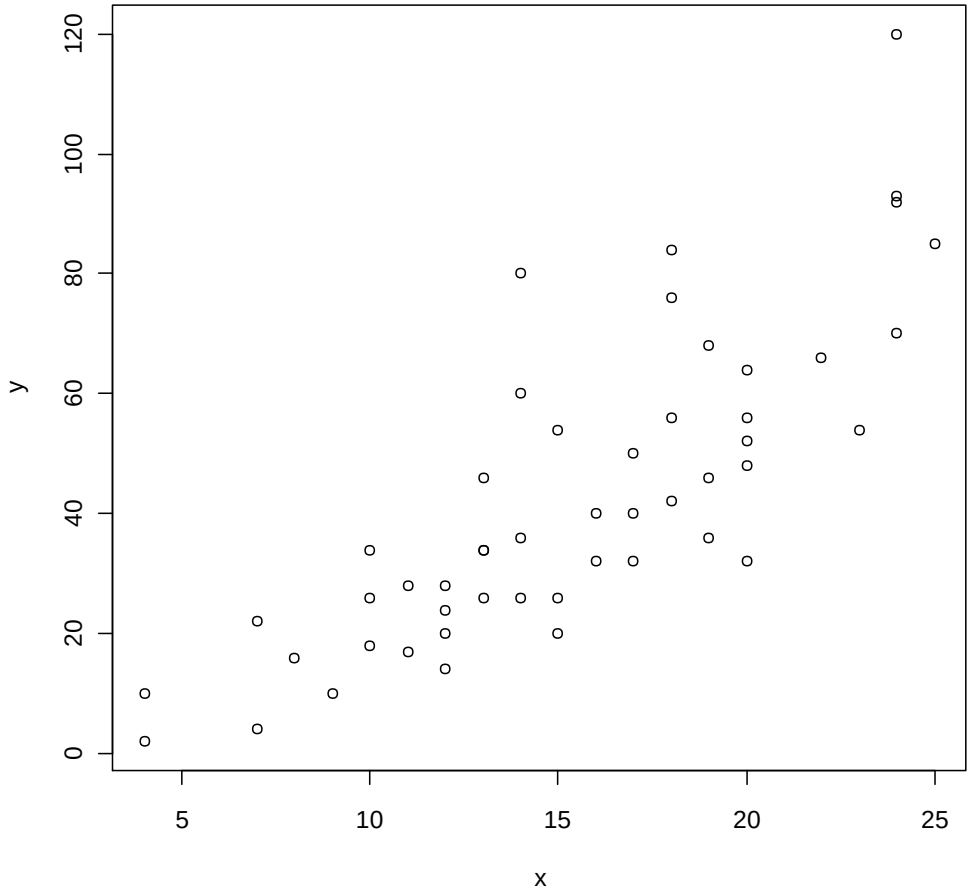
The cars data

```
> help(cars)
```

```
cars                    package:datasets                    R Documentation

Speed and Stopping Distances of Cars

Description:

     The data give the speed of cars and the distances taken to stop.
     Note that the data were recorded in the 1920s.

Usage:

     cars

Format:

     A data frame with 50 observations on 2 variables.

       [,1]  speed  numeric  Speed (mph)
       [,2]  dist   numeric  Stopping distance (ft)

Source:

     Ezekiel, M. (1930) _Methods of Correlation Analysis_.  Wiley.
```

The cars data

```
> x<-carsdat[,2]
> y<-carsdat[,3]
> plot(x,y)
```



The lm() function

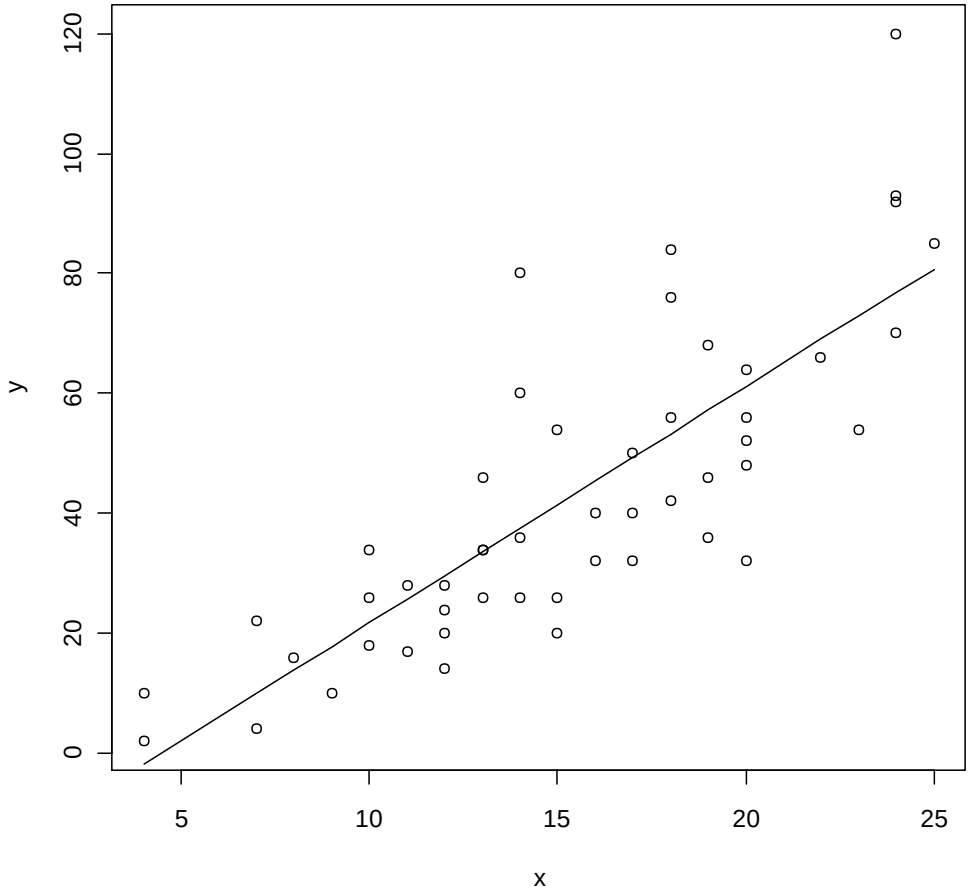
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

General call of the `lm()` function

```
lm(response~predictor)
```

Data and predicted model

```
> fit.1<-lm(y~x)
> plot(x,y)
> lines(x,fit.1$fit)
```



The “output”

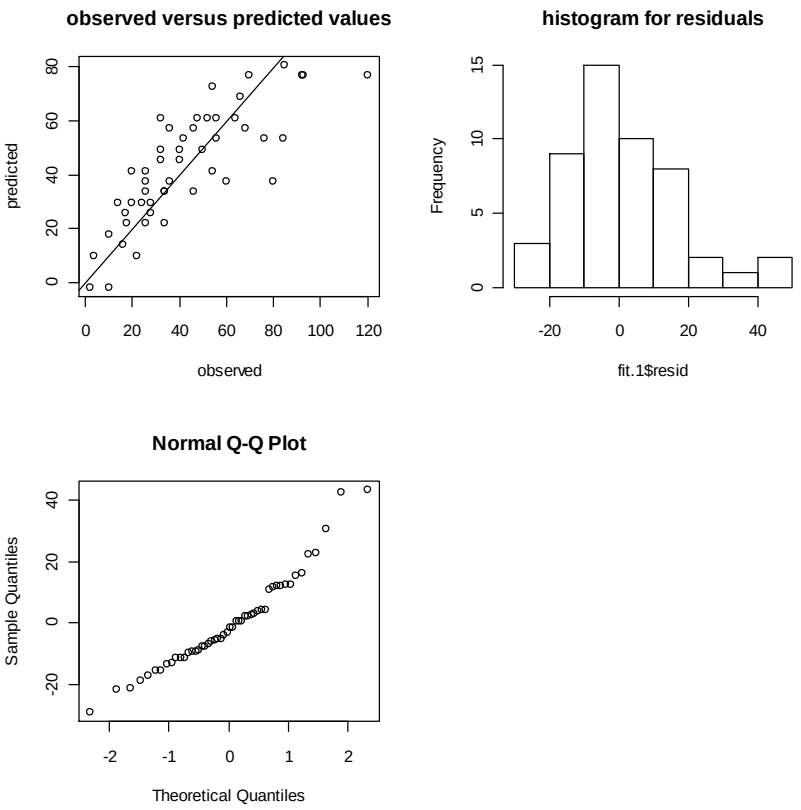
ANOVA table for the model

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  21186 21185.5  89.567 1.490e-12
    ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’
    0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(fit.1)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
x             3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
    0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of
    freedom
Multiple R-squared: 0.6511,     Adjusted R-squared:
    0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.490e-12
```

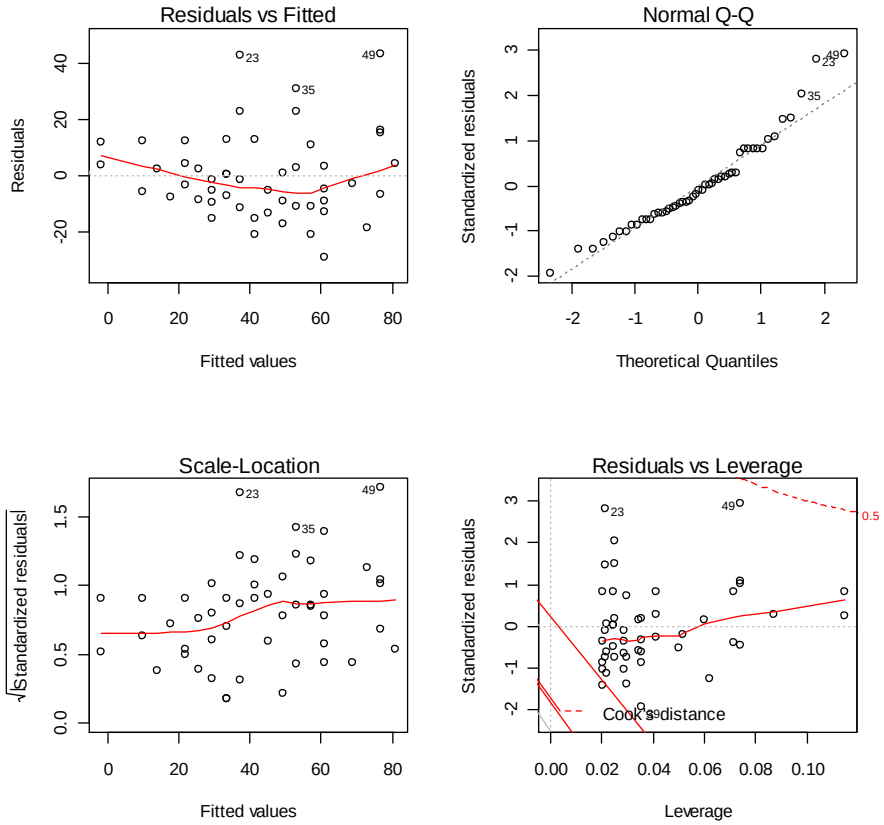
Graphical output

```
> par(mfrow=c(2,2))
> plot(y,fit.1$fit,xlab="observed",
  ylab="predicted")
> abline(0,1)
> title("observed versus predicted
   values")
> hist(fit.1$resid,col=0,main=" ")
> title("histogram for residuals")
> qqnorm(fit.1$resid)
```



Default plots

```
> plot(fit.1)
```



Practical session

- The airquality is a dataset available in R.
- Fit a simple linear regression model in which the ozone level is the response and the wind speed is the predictor.
- Test the hypothesis that the slope is zero.
- Use the default plots of an lm() object to produce the diagnostic plot.

Statistical modeling 2: One way ANOVA

Examples:

The chick data

The cash data

Example 1: The chick dataset in R

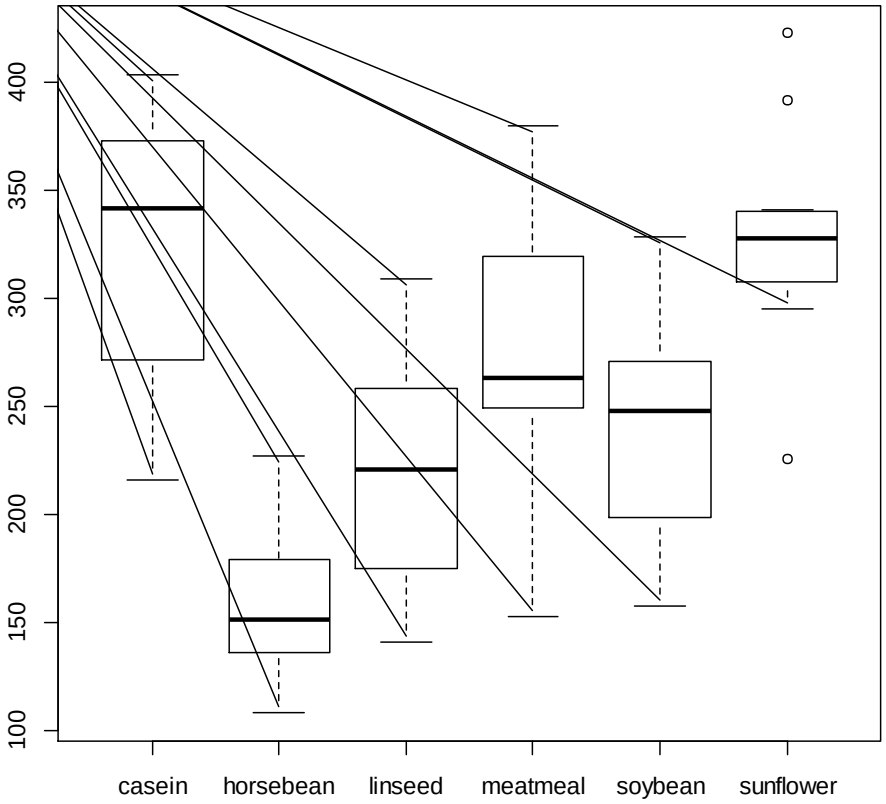
```
> chickwts
   weight      feed
1     179 horsebean
2     160 horsebean
3     136 horsebean
4     227 horsebean

16    203   linseed
17    148   linseed
18    169   linseed

23    243   soybean
24    230   soybean
25    248   soybean
```

> help(chickwts)

An experiment was conducted to measure the effectiveness of various feed supplements on the growth rate of chickens.

Boxplot by group

```
> w<-chickwts[,1]
> feed<-chickwts[,2]
> boxplot(split(w,feed))
```



Mean by group

```
> tapply(w,feed,mean)
   casein horsebean   linseed  meatmeal   soybean sunflower
 323.5833  160.2000  218.7500  276.9091  246.4286  328.9167
```

One-Way ANOVA model: model formulation

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Parameters: fixed but unknown and needed to be estimated

Random error, assumed to follow normal distribution with constant varaince.

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Model assumptions are:

1. The random error is nornal distributed.
2. The varaince is constant across the factor levels.

The Null Hypothesis: No diet effect

- For a model in which the factor has 5 (the diet group) levels we wish to test the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$$

- This means that we want to test if the means across all factor levels are equal.

- Mind that: we test if the parameters (μ_j) are equal, not is the sample means ($\bar{Y}_j$).

Test Statistic

Within group sum of squares

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.} \right)^2$$

Between group sum of squares

$$SSB = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i.} - \bar{Y}.. \right)^2$$

$$F = \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{MSB}{MSW}$$

The test statistic, F, is the ratio between the mean of the between sum of squares (SSB) and the mean of the within sum of squares.

The aov() function

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

```
aov(response ~ factor)
```

```
> a.model=aov(w~feed)
> summary(a.model)
```

Test Statistic

Between group sum of squares/dgree of fredom

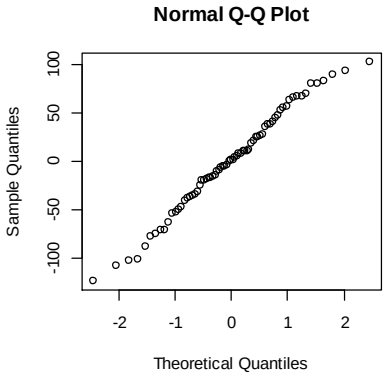
Within group sum of squares/dgree of fredom

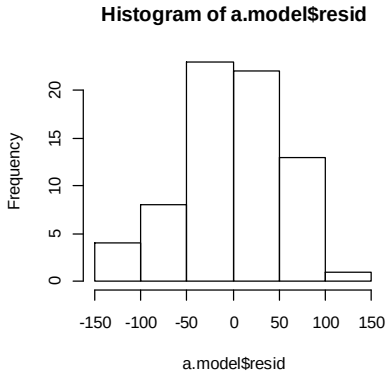
$$= \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{MSB}{MSW} = F$$

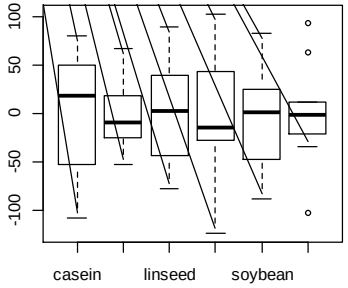
```
> a.model=aov(w~feed)
> summary(a.model)
            Df Sum Sq Mean Sq F value   Pr(>F)
feed         5 231129   46226   15.37 5.94e-10 ***
Residuals   65 195556    3009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Diagnostic plot

```
> par(mfrow=c(2,2))
> qqnorm(a.model$resid)
> hist(a.model$resid,col=0)
> boxplot(split(a.model$resid,feed))
```







One-Way ANOVA model: alternative model formulation

$$Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij}$$

Mean of the reference group

Diet effect

$$\sum_{i=1}^{I} \alpha_i = 0$$

Random error, assumed to follow normal distribution with constant varaince.

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Model assumptions are:

1. The random error is nornal distributed.
2. The varaince is constant across the factor levels.

Estimation of the model in R

$$Y_{ij} = \mu_0 + \alpha_i + \varepsilon_{ij}$$

```
lm(response~predictor)
```

```
> lm.fit<-lm(w~feed)
```

Estimation of the model in R

```
> summary(lm.fit)

Call:
lm(formula = w ~ feed)

Residuals:
     Min       1Q   Median       3Q      Max
-123.909  -34.413    1.571   38.170  103.091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   323.583     15.834  20.436  < 2e-16 ***
feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
feedsoybean    -77.155     21.578  -3.576 0.000665 ***
feedsunflower    5.333     22.393   0.238 0.812495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.5064
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

323.583: the mean of the casein

323.583-163.383= 160.2000, the mean of the horsebeen group

The AVOVA table

```
Residual standard error: 54.85 on 65 degrees of freedom
Multiple R-squared:  0.5417,    Adjusted R-squared:  0.5064
F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

$$54.85 = \sqrt{3009}$$

```
> anova(lm.fit)
Analysis of Variance Table

Response: w
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       5 231129   46226  15.365 5.936e-10 ***
Residuals 65 195556    3009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example 2: Reading the cash data

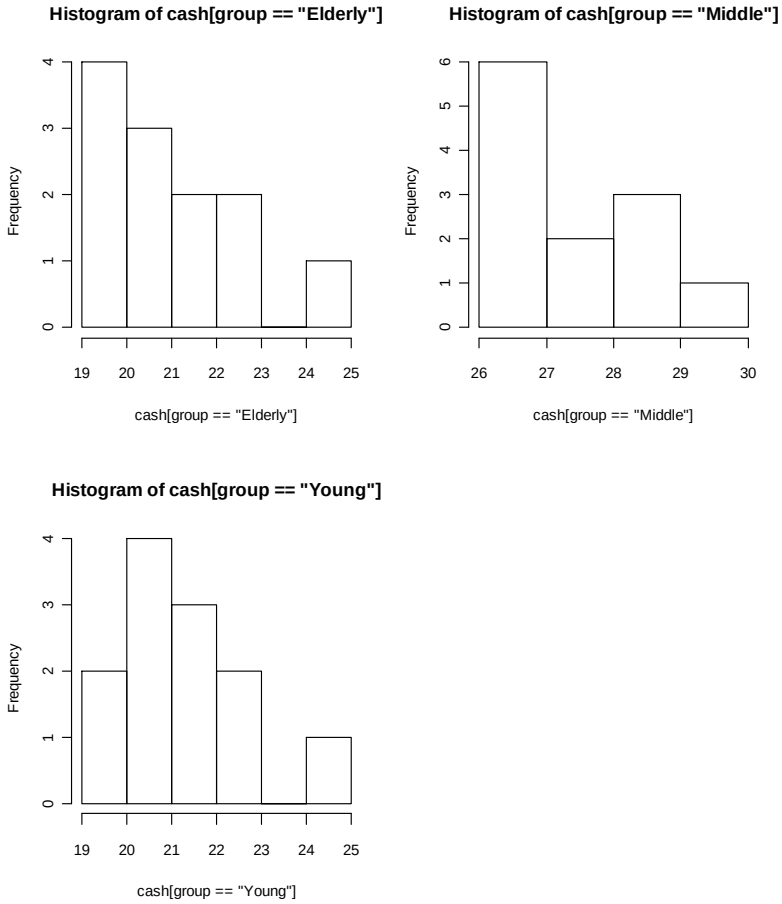
```
> cashdat<-
  read.table('c:\\projects\\wseda\\Rintro\\cashdat.txt',
  header=FALSE,na.strings="NA", dec=".")
> dim(cashdat)
[1] 36  2
> names(cashdat)<-c("cash","group")
> attach(cashdat)
```

The data

```
> print(cashdat)
   cash   group
1    23   Young
2    25   Young
.     .      .
.     .      .
11   21   Young
12   21   Young
13   28  Middle
.     .      .
.     .      .
24   29  Middle
25   23 Elderly
26   20 Elderly
35   22 Elderly
36   21 Elderly
```

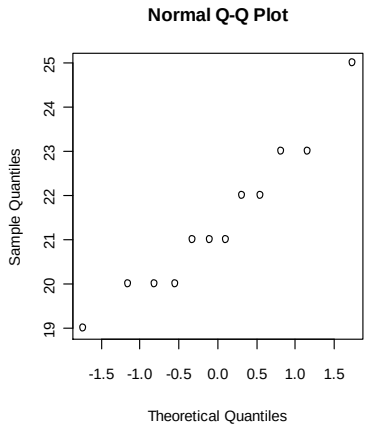
Histograms by group

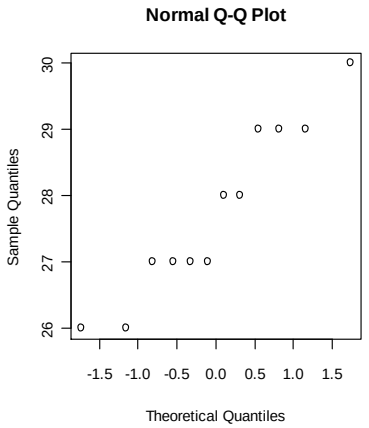
```
> par(mfrow=c(2,2))
> hist(cash[group=="Elderly"],col=0)
> hist(cash[group=="Middle"],col=0)
> hist(cash[group=="Young"],col=0)
```

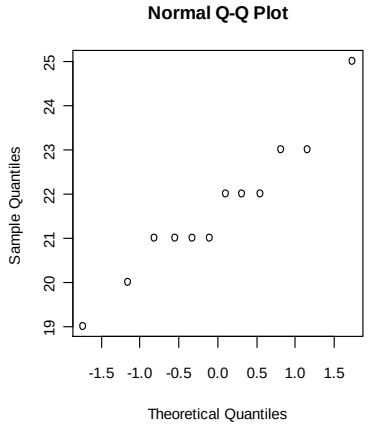


qq normal plots by group

```
> par(mfrow=c(2,2))
> qqnorm(cash[group=="Elderly"])
> qqnorm(cash[group=="Middle"])
> qqnorm(cash[group=="Young"])
```





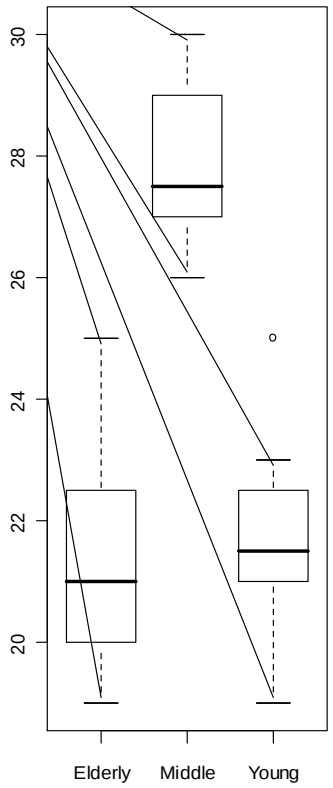


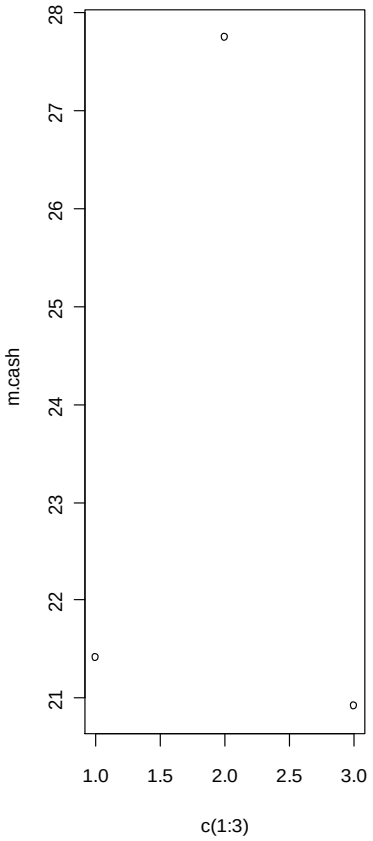
Boxplot and dotplot

```
> par(mfrow=c(1,2))
> boxplot(split(cash,group))
```

```
> tapply(cash,group,mean)
 Elderly   Middle    Young
21.41667 27.75000 21.66667
```

```
> m.cash<-c(21.41667,27.75,20.91667)
> names1<-c("Elderly","Middle","Young")
> plot(c(1:3),m.cash)
```





One-Way ANOVA model: model formulation

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Parameters: fixed but unknown and needed to be estimated

Random error, assumed to follow normal distribution with constant varaince.

$$\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

Model assumptions are:

1. The random error is nornal distributed.
2. The varaince is constant across the factor levels.

The Null Hypothesis: No treatment effect

- For a model in which the factor has three levels we wish to test the null hypothesis:

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

- This means that we want to test if the means across all factor levels are equal.

- Mind that: we test if the parameters (μ_j) are equal, not is the sample means ($\bar{Y}_j$).

Test Statistic

Within group sum of squares

$$SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.} \right)^2$$

Between group sum of squares

$$SSB = \sum_{i=1}^{I} n_i \left(\bar{Y}_{i.} - \bar{Y}.. \right)^2$$

$$F = \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{MSB}{MSW}$$

The test statistic, F, is the ratio between the mean of the between sum of squares (SSB) and the mean of the within sum of squares.

The aov() function

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

```
aov(response ~ factor)
```

```
>Fit.aov<-aov(cash~group)
>summary(Fit.aov)
```

Test Statistic

Between group sum of squares/dgree of fredom

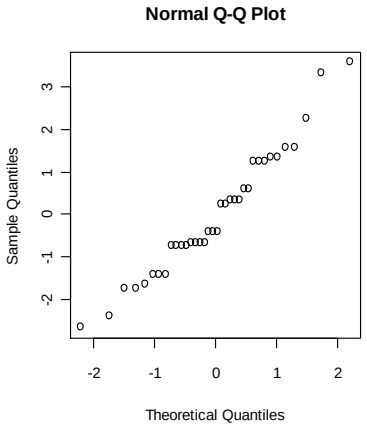
Within group sum of squares/dgree of fredom

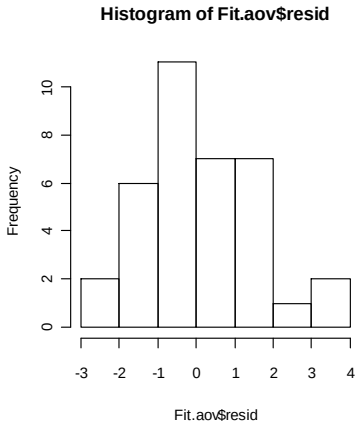
$$= \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{MSB}{MSW} = F$$

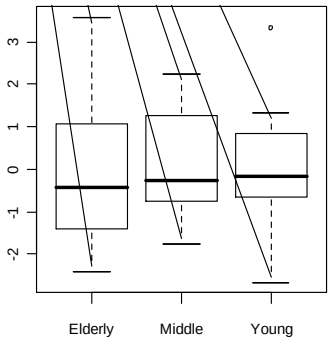
```
> summary(Fit.aov)
             Df  Sum Sq Mean Sq F value    Pr(>F)
group         2 308.722 154.361  67.172 2.322e-12 ***
Residuals    33  75.833   2.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Diagnostic plot

```
> par(mfrow=c(2,2))
> qqnorm(Fit.aov$resid)
> hist(Fit.aov$resid,col=0)
> boxplot(split(Fit.aov$resid,group))
```







Practical session (a)

- Create the following data frame in R

	y	treatment
1	10	A
2	12	A
3	13	A
4	15	A
5	10	B
6	9	B
7	9	B
8	11	B
9	10	C
10	15	C
11	13	C
12	8	C

- Use one-way ANOVA model to test the null hypothesis of no treatment effect

Practical session (b)

- Create the following data frame in R

	y	treatment
1	10	A
2	12	A
3	13	A
4	15	A
5	10	B
6	9	B
7	9	B
8	11	B
9	10	C
10	15	C
11	13	C
12	8	C

- Use one-way ANOVA model to test the null hypothesis of no treatment effect

Statistical modeling 3: Logistic regression

Examples:

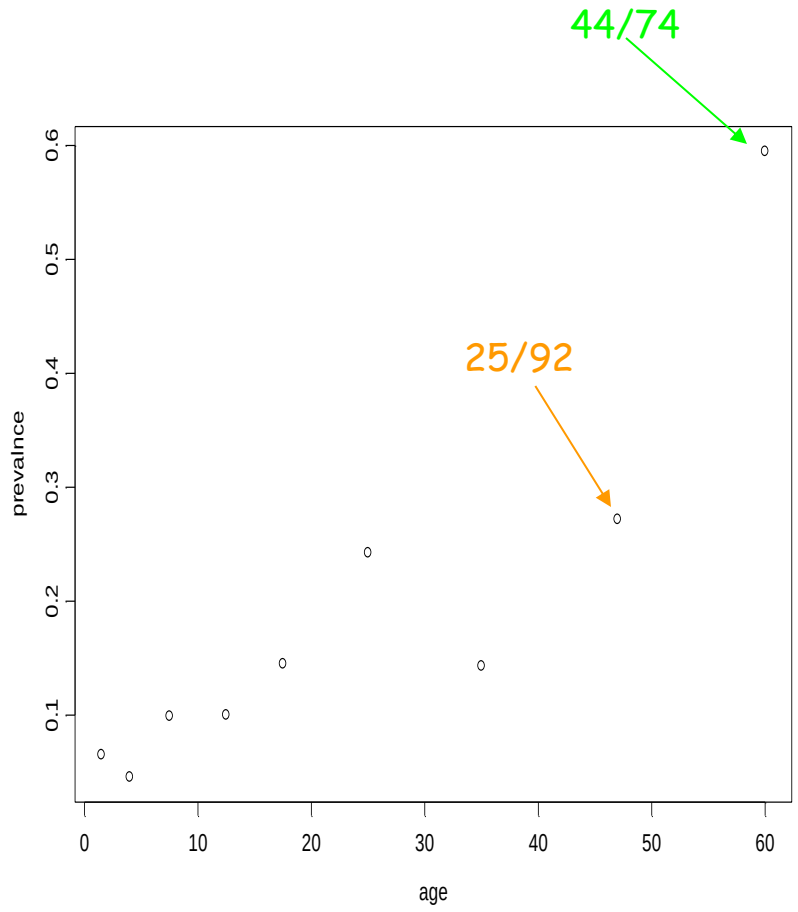
Serological data

Example : Serological data of malaria

- In this example the information about each subject in the experiment is the disease status (infected or not by malaria) and the age group of the subject.
- The variables are: the sample size, the number of sero-positive at each sample size (=the number of infected subjects) and the age.

Example : serological data

Age group	Mid age	Sero positive	Sample size
	1.5	8	123
	4.0	6	132
	7.5	18	182
	12.5	14	140
	17.5	20	138
	25.0	39	161
	35.0	19	133
	47.0	25	92
	60.0	44	74



Reading the data

```
> sero<-read.table('c:\\projects\\wseda\\Rintro\\sero1.txt',
header=FALSE,na.strings="NA", dec=".")

> print(sero)
  V1   V2  V3 V4
1  1  1.5 123  8
2  2  4.0 132  6
3  3  7.5 182 18
4  4 12.5 140 14
5  5 17.5 138 20
6  6 25.0 161 39
7  7 35.0 133 19
8  8 47.0  92 25
9  9 60.0  74 44
```

Example : serological data

Mid age	Sero positive	Sample size
1.5	8	123
4.0	6	132
7.5	18	182
12.5	14	140
17.5	20	138
25.0	39	161
35.0	19	133
47.0	25	92
60.0	44	74

$$Z_i = \begin{cases} 1 & sero \quad pos. \\ 0 & sero \quad neg. \end{cases}$$

$$Y_i = \sum Z_i$$

Number of sero-positive at each age group

$$Y_i \sim B(n_i, \quad P_i)$$

n_i: sample size at each age group

P_i is the probability to be infected (the prevalence). We use logistic regression in order to model the prevalence as a function of age

$$\log it(P_i) = \alpha + \beta \times age$$

The probability of infection

$$P = \frac{e^{\alpha+\beta\ age}}{1+e^{\alpha+\beta\ age}}$$

If β>0 then there is a positive association between the probability and age. This means that the probability of infection increase with age.

If β<0 then there is a negative association between the probability and age. This means that the probability of infection decrease with age.

The glm() function

$$Y_i \sim B(n_i, \quad P_i) \qquad \log it(P_i) = \alpha + \beta \times age$$

glm(pos/ntot ~ age, family=binomial(link = "logit"))

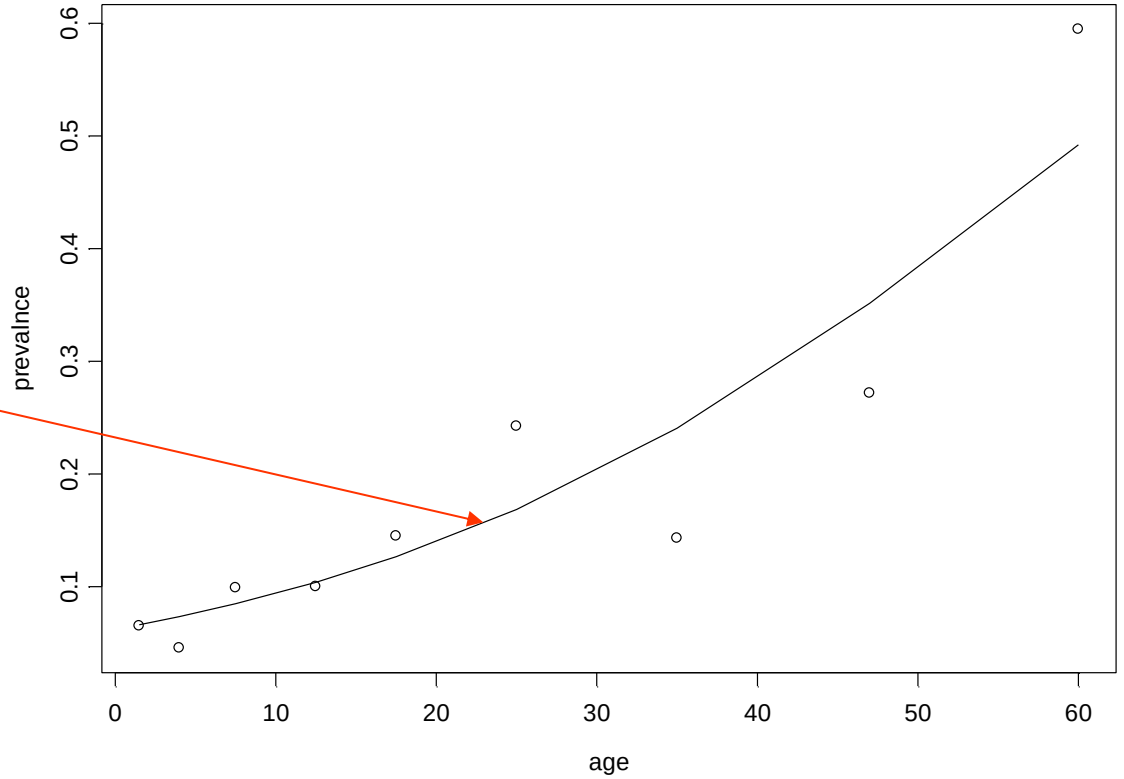
The glm() function

```
> fit.glm<- glm(pos/ntot ~ age, family=binomial(link = "logit"))
> summary(fit.glm)

Call:
glm(formula = pos/ntot ~ age, family = binomial(link = "logit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.24364  -0.09726   0.01479   0.06756   0.19568

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.79677    1.79832  -1.555    0.120
age          0.04718    0.04668   1.011    0.312

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.31775  on 8  degrees of freedom
Residual deviance: 0.18094  on 7  degrees of freedom
AIC: 8.062

Number of Fisher Scoring iterations: 5
```

Data and predicted values

$$\log it\left(\hat{P}_i\right) = -2.71 + 0.044 \times age$$

$$\hat{P}_i = \frac{e^{-2.71+0.044\times age}}{1+e^{-2.71+0.044\times age}}$$



Discussion

- Data: input for the analysis
- R Objects: output of the analysis.
- R functions: `lm(),glm(), aov().`
- $.