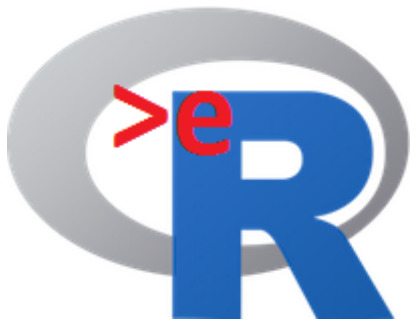




This course was developed as a part of the VLIR-UOS Cross-Cutting projects:

- 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

Applied Generalized Linear Models (GLM) using R (PART 1)

Developed by

Tadesse Awoke (Gondar University), Said Mussa (Mekelle University) and Ziv
Shkedy (Hasselt University), Fetene Tekle (J & J)

LAST UPDATE: 24/02/2024



ER-BioStat

Email: erbiostat@gmail.com



<https://github.com/eR-Biostat>



@erbiostat

Reference list

- Main reference:
 - Dobson (2002): An introduction to generalized linear models.
- Other references:
 - McCullagh and Nelder (1983): Generalized linear models (first edition).
 - Collet D(1994): Modeling Binary data.
 - Lindsey (1997): Applying generalized linear models.



Software

- Two main R functions:
 - Linear models in R: the `lm()` function.
 - Generalized linear models in R: the `glm()` function in R.
- All R programs for the examples presented in the slides are available online.



Datasets

- Data are given as a part of R programs for the course.
- External datasets (which are not given as a part of the R code) and used for illustration are available online.

Topics (part 1)

1. Analysis of Variance
2. Linear regression models with normal error
3. Generalized linear models
4. Exponential Family
5. Generalized linear model function in R
6. Models for Binary data
7. Estimation and confidence intervals
8. Inference
9. Model Selection
10. Model diagnostic

Topics (part 2)

- 11. Poisson Regression
- 12. Beyond Poisson and binomial distributions:
models with different link functions and/or distributions
- 13. Poisson regression and log linear models
- 14. Over dispersion

Chapter 1: Analysis of Variance (ANOVA)

Donson: chapter 2

Lindsey: chapter 9

McCullagh & Nelder: chapter 3

Example 1: A Biopharmaceutical Problem

- A group of 24 rats were randomized into two treatment groups: active drug and placebo
- After the administration of the drug, the rat was placed on a surface, and the distanced traveled by the rat (in meters) was measured.

The data

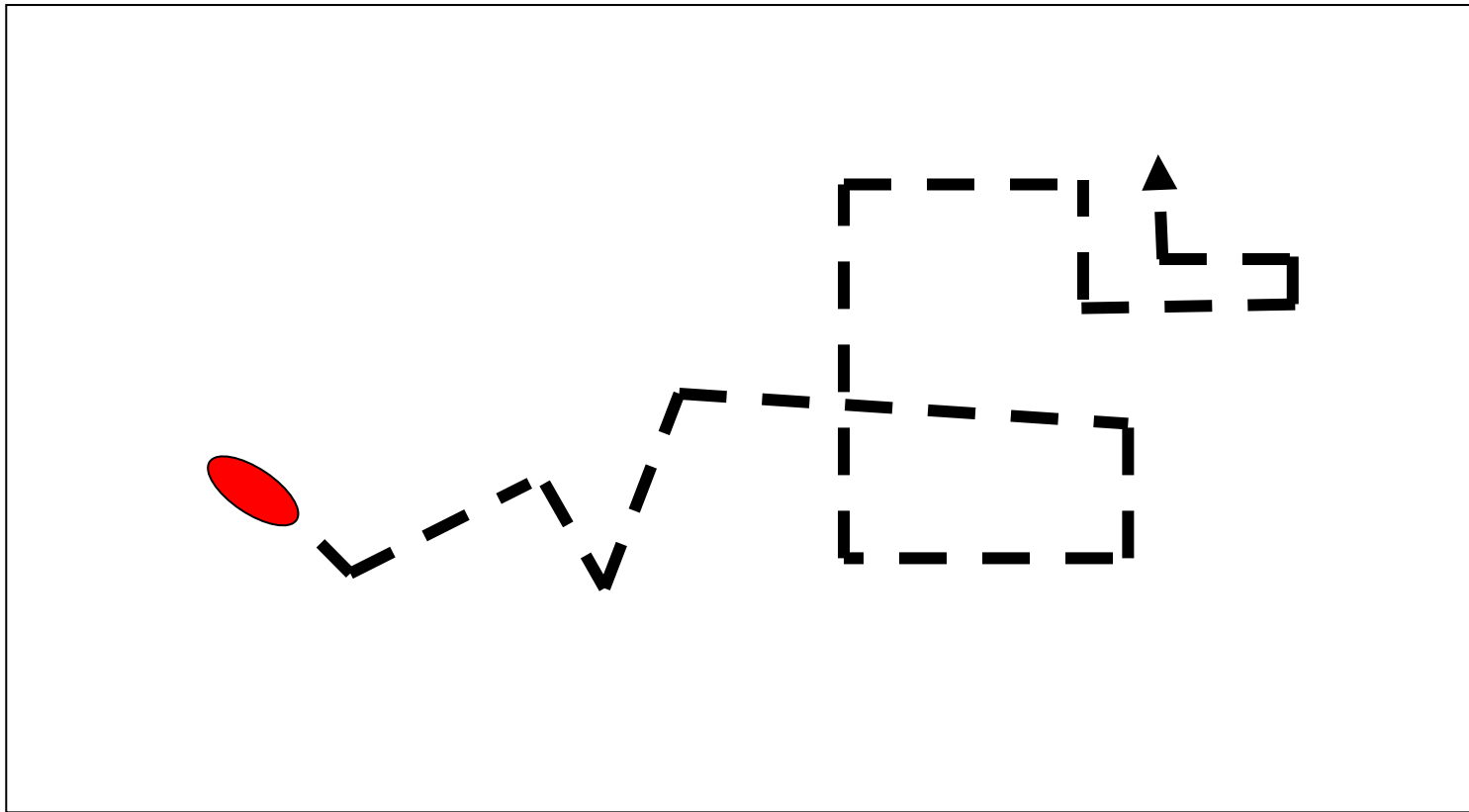
22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Response

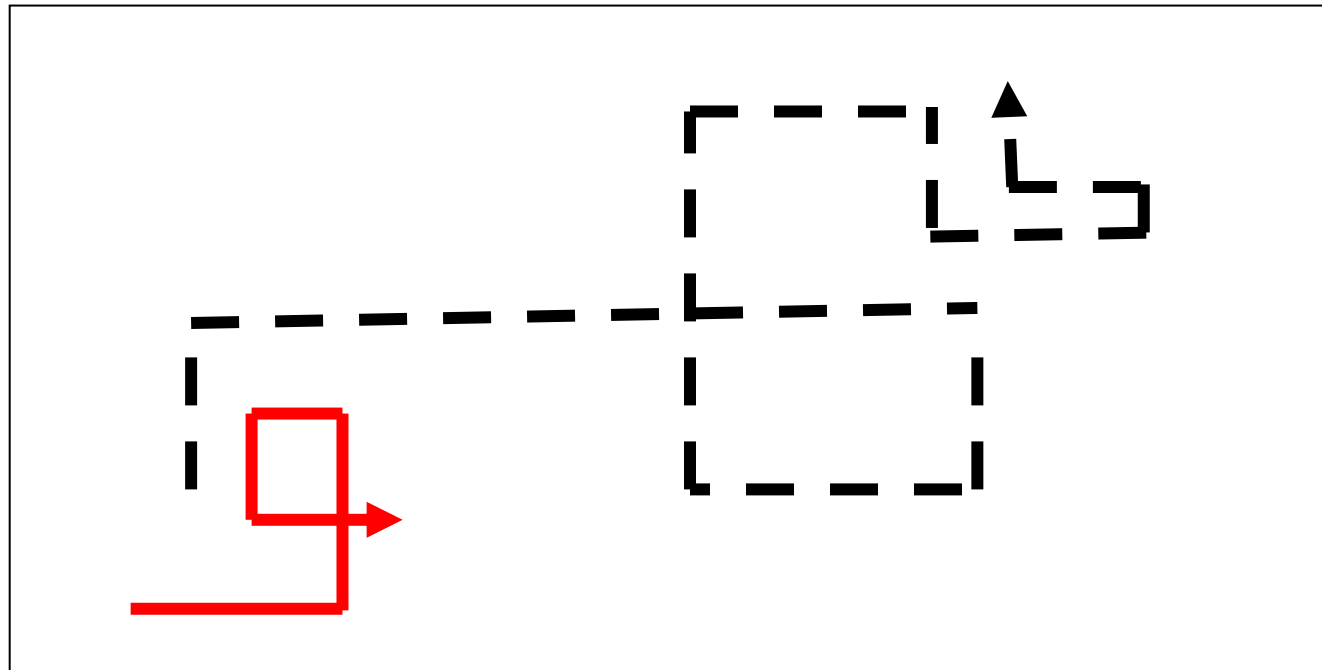
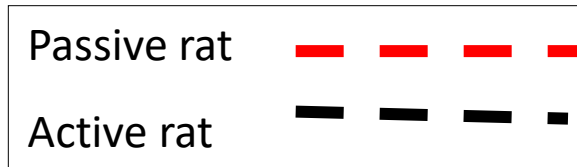
Treatment group

The Evaluation of the Rat performance in distance

Y_i is the distance traveled by the rat during the experiment.



Description of the Experiment



It is assumed that a successful drug increase the distance traveled by the rat.

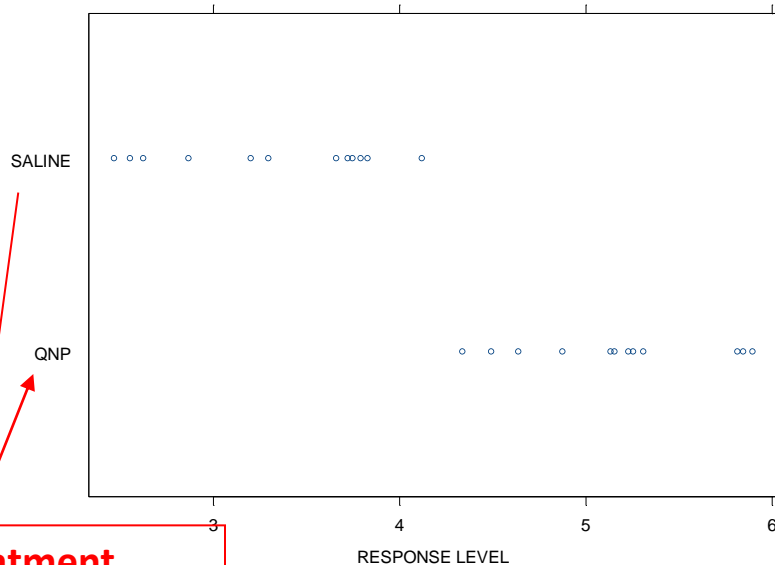
The Scientific Question

- Does the drug increase the distance traveled by the rat ?

**A good drug is expected to improve the rats' performance,
i.e. to increase the distance travel by the rat**

Graphical display of the data (1)

A strip plot



The data

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Response

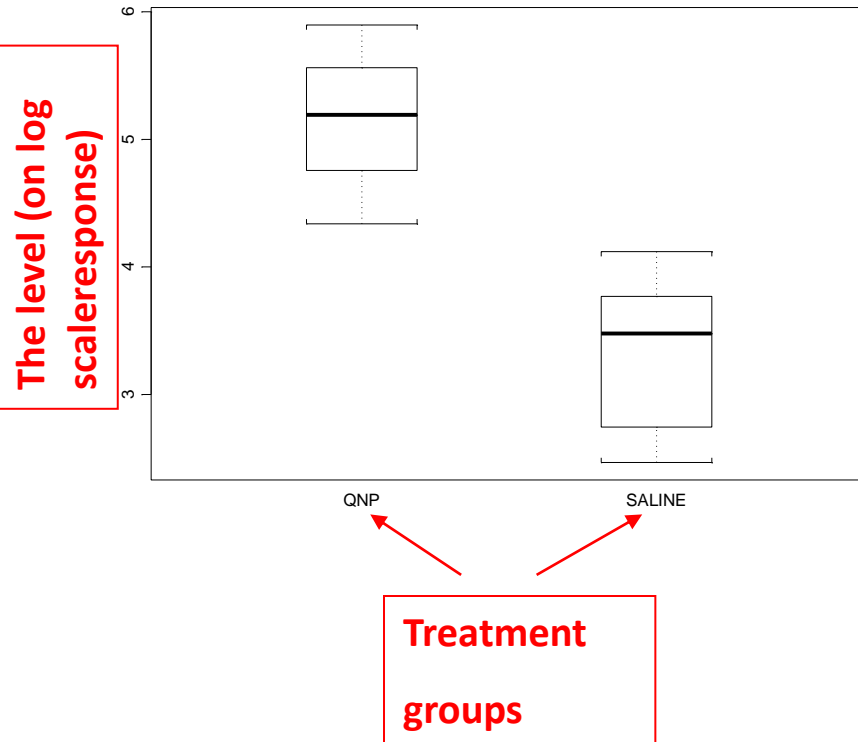
Treatment groups

The response level (on log scale)

Treatment group

Graphical display of the data (2)

A boxplot plot



22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

The data

Response

Treatment group

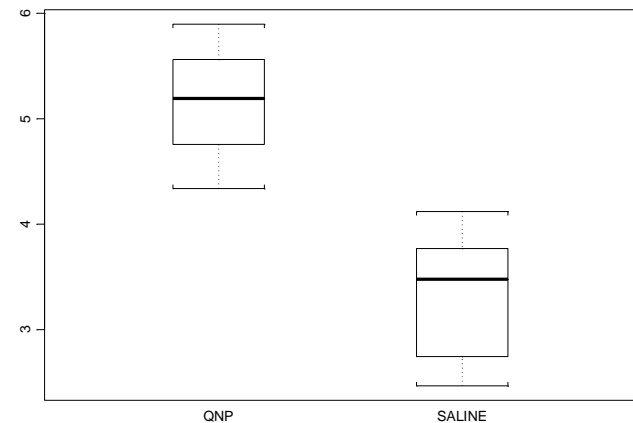
Boxplot by treatment group

The data in R:

```
> dist<-c(186.6145,103.3529,191.3850,334.9845,89.2831,  
          345.5070,169.5161,173.1491,130.9634,363.4392,  
          76.5340,202.1145,12.8458,44.3092,41.3581,24.5560,  
          61.5525,38.8464,27.0107,45.9960,13.7927,42.4009,17.5861,  
          11.7937)  
> gr<-c(rep(0,12),rep(1,12))
```

The boxplot:

```
> boxplot(split(dist,gr))
```



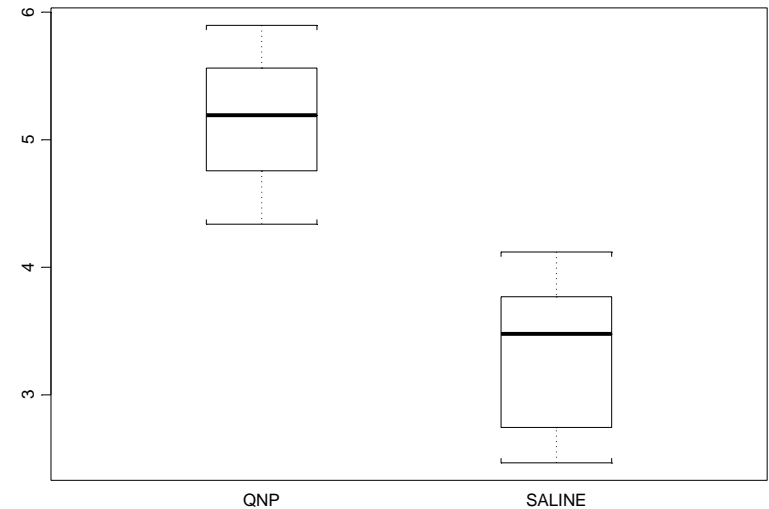
Groups' means

```
> tapply(dist, as.factor(gr), mean)
```

0	1
197.23694	31.83734

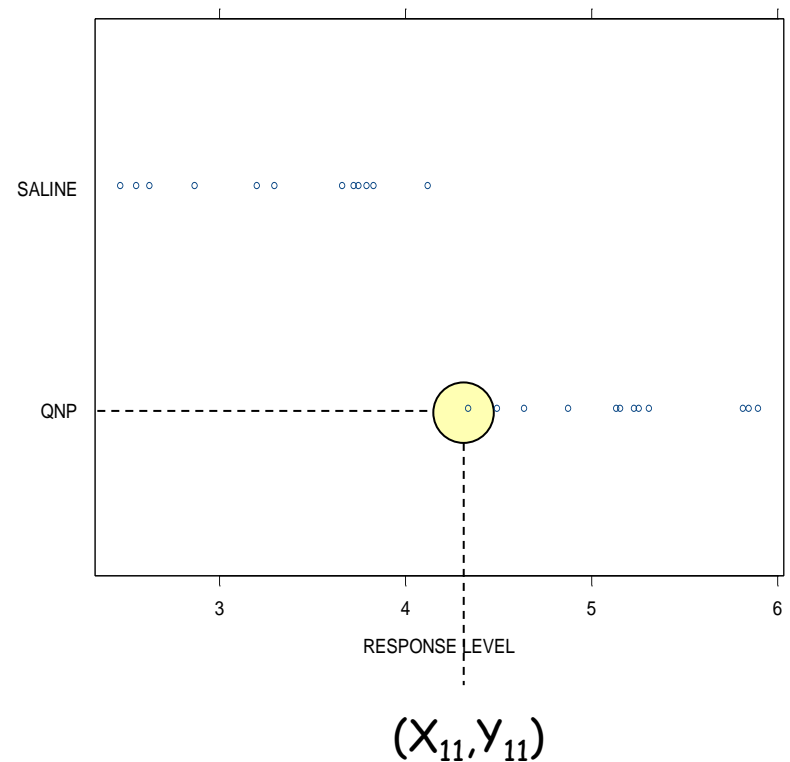
```
> tapply(dist, as.factor(gr), median)
```

0	1
179.88180	32.92855



ANOVA Terminology

- The distance traveled is the dependent variable. This is the response.
- The treatment group is the independent variable and it called the factor. In this example the factor has two levels.



Data Structure

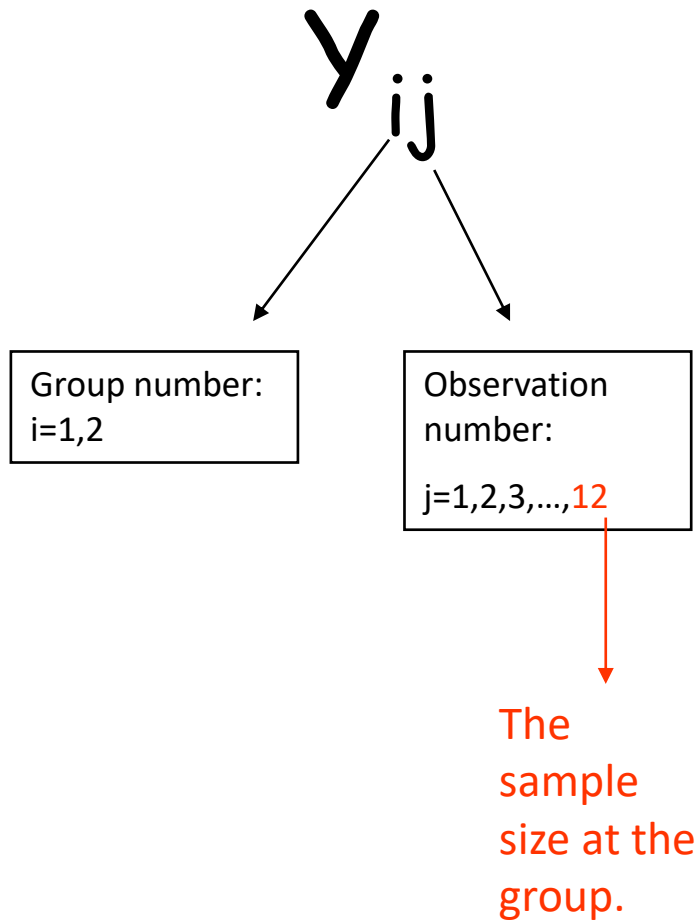
- We have two variables, the factor (x) and the response (Y).
- The value of X is equal for all subjects from the same treatment group. This value is the factor level.

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

The factor: the
treatment
group

The response:
the distance
traveled (y_i)

Data Structure: notation (1)



22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145
1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Group 1:

$i=1$

$n_1=12$

Group 2:

$i=2$

$n_2=12$

y_{212} :
Observation
number 12 in
group 2

Data Structure: notation (2)

Number of Group: l

Sample size: n

$n=n_1+n_2+\dots,n_k$

Overall mean: $\bar{Y}_{..}$

Mean of group i: $\bar{Y}_i.$

Sample size in group i: n_i

22	QNP	186.6145
11	QNP	103.3529
4	QNP	191.3850
16	QNP	334.9845
7	QNP	89.2831
13	QNP	345.5070
2	QNP	169.5161
20	QNP	173.1491
19	QNP	130.9634
8	QNP	363.4392
10	QNP	76.5340
24	QNP	202.1145

1	SALINE	12.8458
17	SALINE	44.3092
15	SALINE	41.3581
6	SALINE	24.5560
23	SALINE	61.5525
18	SALINE	38.8464
5	SALINE	27.0107
12	SALINE	45.9960
21	SALINE	13.7927
14	SALINE	42.4009
3	SALINE	17.5861
9	SALINE	11.7937

Group 1: The group mean

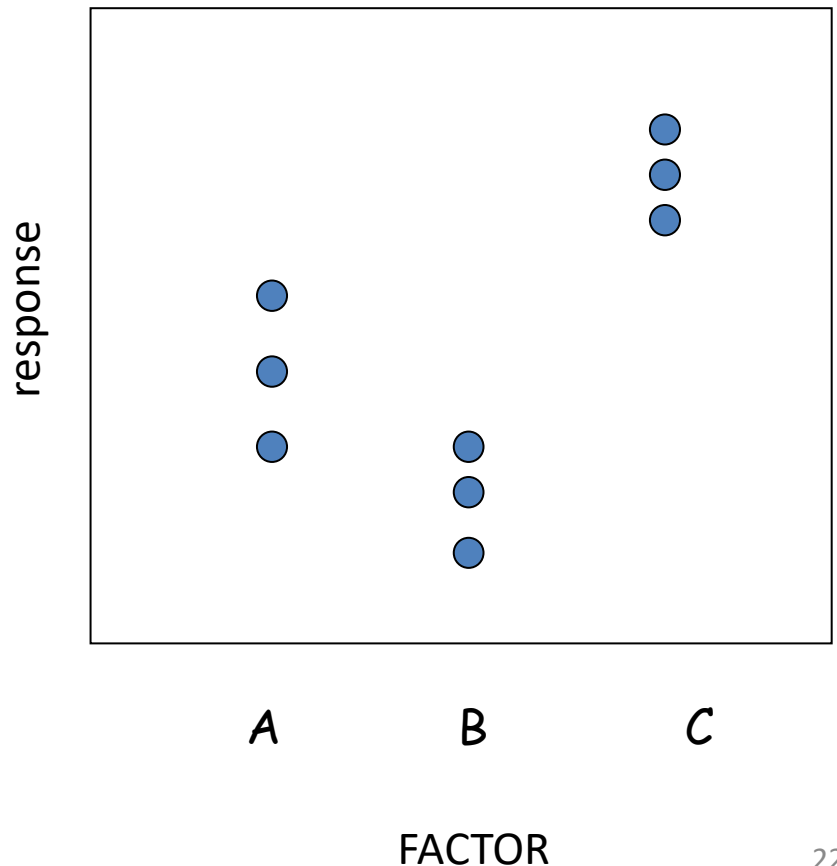
$\bar{Y}_1.$

Group 2: The group mean

$\bar{Y}_2.$

What is a One-Way ANOVA Model ?

- A One-Way ANOVA model is a statistical model which aims to explain the **variability of the response variable**.
- The question of primary interest is **IF THE MEAN RESPONSE IS DIFFERENT** across the factor levels.



one-way ANOVA: testing of hypotheses

$$H_0 \text{ \& } H_1$$

Testing of hypotheses

- The sample per treatment group (i. e, each level of the treatment factor) is a sample of a population.
- We want to test whether the means of the populations across the factor levels are equal or not.
- The averages of the populations are parameters (but unknown parameters). We want to estimate these parameters.

Populations and the factor levels (and assumptions)

Populations	Distribution of the populations
1) (QNP)	$N(\mu_1, \sigma^2)$
2) (SLINE)	$N(\mu_2, \sigma^2)$

population mean

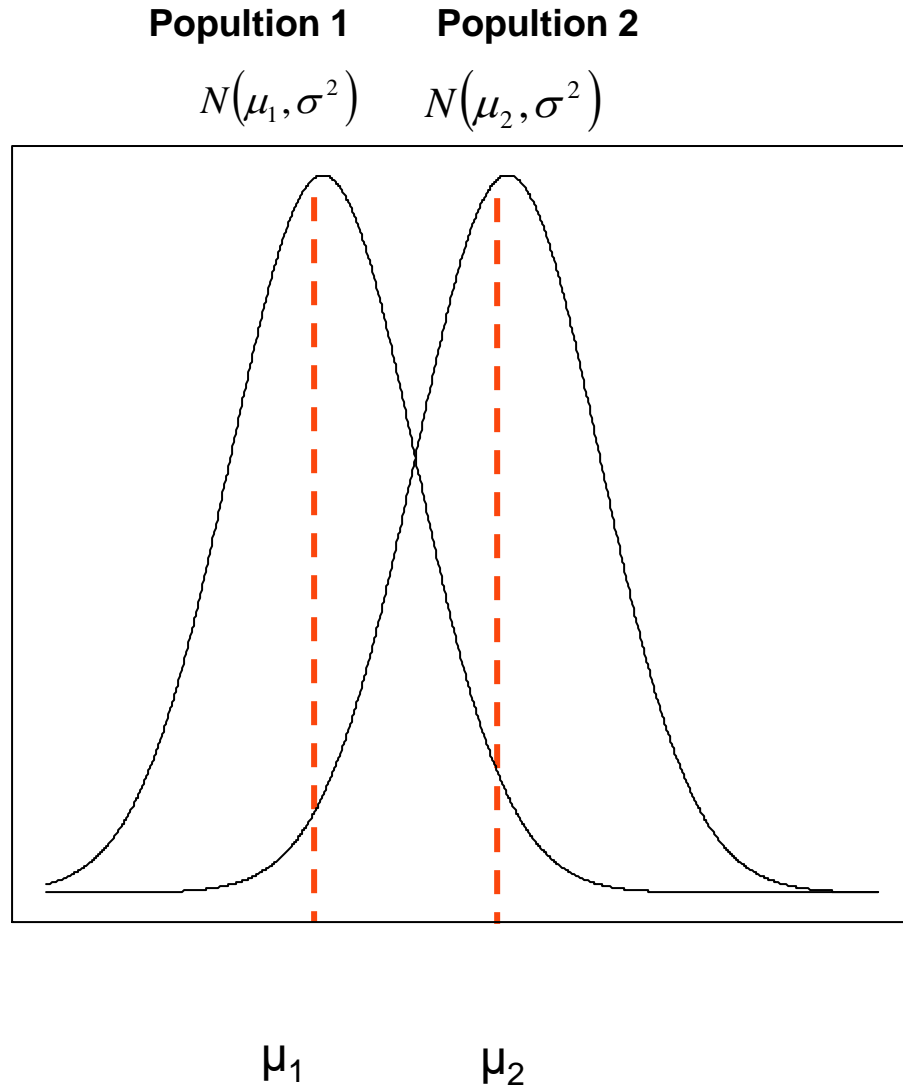
Population
variance

- There are two factor levels (groups): active drug and placebo.
- Each subject within a treatment group is a random sample from a population.
- We assume that

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

Two populations

- We assume that the variance is constant (σ^2).
- The null hypothesis is not rejected if the means are equal



Formulation of the null hypothesis

The null hypothesis states that (for K populations) the average of the K populations is the same.

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$$

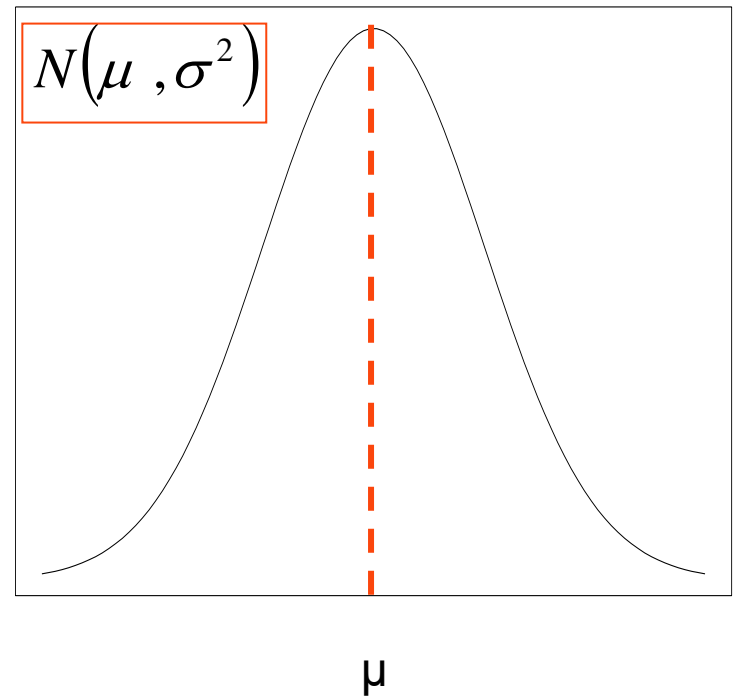
In other words, there is no effect of the treatment.

The null hypothesis

- Under the null hypothesis the means in the populations are equal.
- This means that:

$$Y_{ij} \sim N(\mu, \sigma^2)$$

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K$$



What is the alternative?

$$H_1 : \mu_i \neq \mu_l$$

For at least one pair of i and l
($i, l = 1, 2, \dots, K$)

$$\Rightarrow Y_{ij} \sim N(\mu_i, \sigma^2) \text{ and } Y_{lj} \sim N(\mu_l, \sigma^2) \\ \text{for } i \neq l, i, l = 1, 2, \dots, K$$

one-way ANOVA: inference

$$H_0 \text{ \& } H_1$$

Two Sources of Variability

- The main concept in ANOVA models, and in particular One-way ANOVA is to decompose the total variability of the response into two parts.

total variability = variability **within** the groups + variability **between** the groups

- An ANOVA model is a model in which we explain the total variability with these two sources.

A very simple example

- One factor experiment.
- The factor has three levels (1,2,3).
- Three observations at each level.

```
> resp<-c(2,3,4,5,6,7,1,2,3)
> gr<-c(1,1,1,2,2,2,3,3,3)
> data.frame(resp,gr)
  resp gr
1     2  1
2     3  1
3     4  1
4     5  2
5     6  2
6     7  2
7     1  3
8     2  3
9     3  3
```

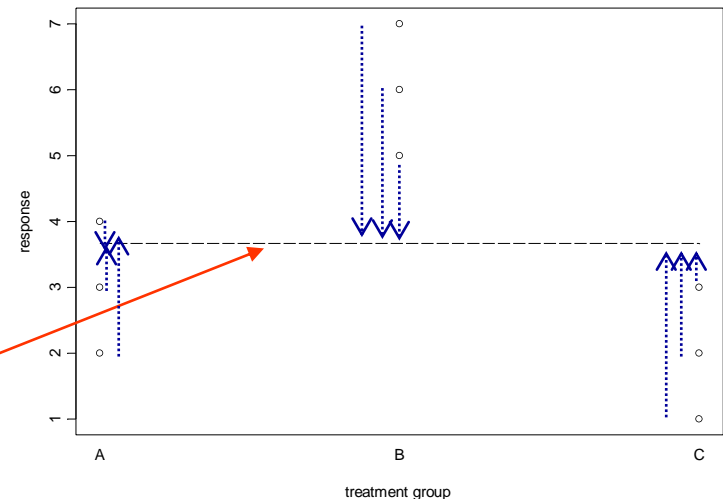
group	Y_{ij}	Group mean
1 1 1	2 3 4	3
2 2 2	5 6 7	6
3 3 3	1 2 3	2

Overall mean: 3.6666

Two Sources of Variability: the total variability

The total sum of squares (SST) is the sum of squared distance between the observations from the overall mean.

The overall
mean=3.6667



$$(2 - 3.666)^2 + (3 - 3.666)^2 + (4 - 3.666)^2 + \dots + (2 - 3.666)^2 + (3 - 3.666)^2 = 32$$

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

Two Sources of Variability: the variability within the groups

The sum of squares within the groups is the sum of squared difference between the observations at each group to the group mean.

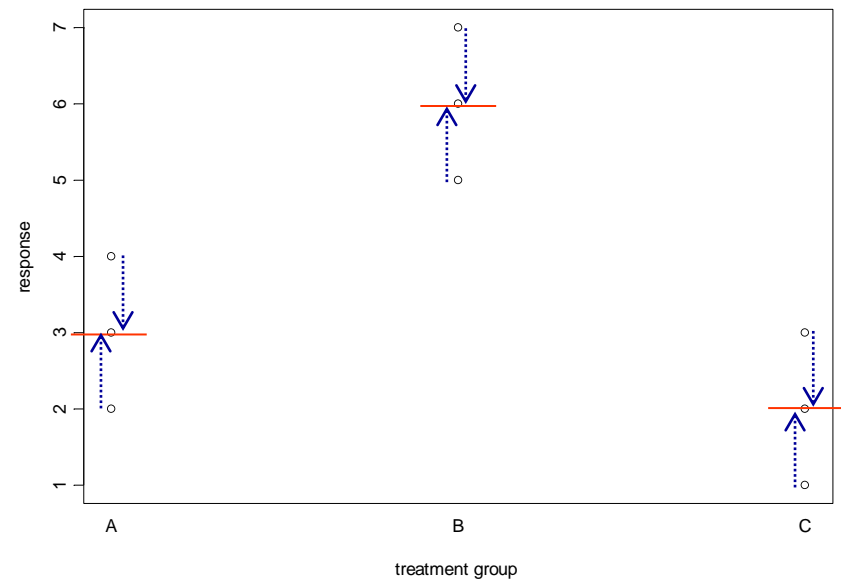
$$A \quad (2-3)^2 + (3-3)^2 + (4-3)^2 = 2$$

$$B \quad (5-6)^2 + (6-6)^2 + (7-6)^2 = 2$$

$$C \quad (1-2)^2 + (2-2)^2 + (3-2)^2 = 2$$

6

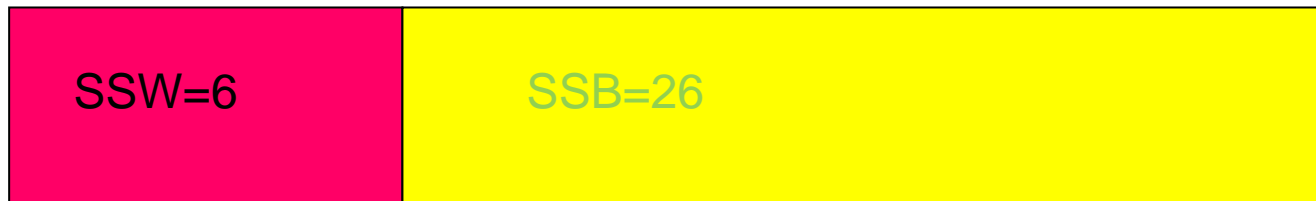
$$SSW = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$



Groups means: 3 (group A), 6 (group B) and 2 (group C)

Two Sources of Variability

Total variability



Variability within the
groups

Variability between
the groups

$$SST = SSW + SSB$$


In the slides for the class we use the notations:

$$SST = SSE + SSTR$$

The function `aov()` in R

Analysis Of Variance:

`aov(response~predictor(s))`



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

Two Sources of variability in R

```
> fit.1<-aov(resp~as.factor(gr))
```

```
> summary(fit.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(gr)	2	26	13	13	0.006592 **
Residuals	6	6	1		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 variance.

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

Model assumptions are:

1. The random error is normal distributed.
2. The variance 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 about the sample means (\bar{Y}_j).

Test Statistic **F**

Within group sum of squares

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

Between group sum of squares

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

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

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.

Test Statistic in R

Within group sum of squares/dgree of fredom

Between group sum of squares/dgree of fredom

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

```
> fit.1<-aov(resp~as.factor(gr))
> summary(fit.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(gr)	2	26	13	13	0.006592 **
Residuals	6	6	1		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of example 1: distance in rat experiment

```
> fit.2<-aov(dist~as.factor(gr))
> summary(fit.2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
as.factor(gr)	1	164142	164142	32.131	1.062e-05	***
Residuals	22	112389	5109			

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A typical example of one-way ANOVA (1)

- For an experiment with I treatments we have I groups
- Each group receives different treatment
- Sample size at each group: n_i
- We assume that each group is a sample from a population

Example 2: phosphate concentration in plasma (Table 6.18 in Dobson (2002))

- The response variable is the concentration of phosphate in the plasma.
- 3 treatment groups:
Hyperinsulinemic obese (HZ), Non Hyperinsulinemic obese (NHZ), and Controls (R).

HZ	NHZ	R
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.5
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

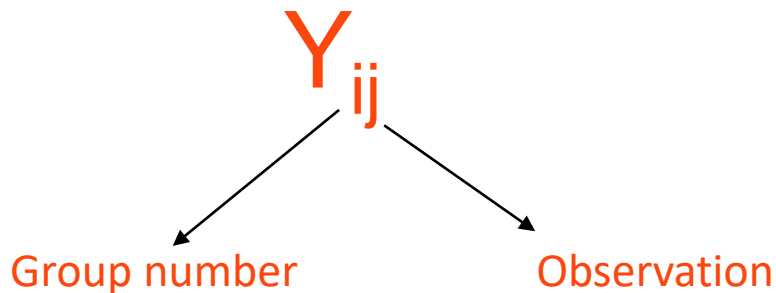
Data structure and notation

Example of
one-way
ANOVA

Response: Y_{ij} .

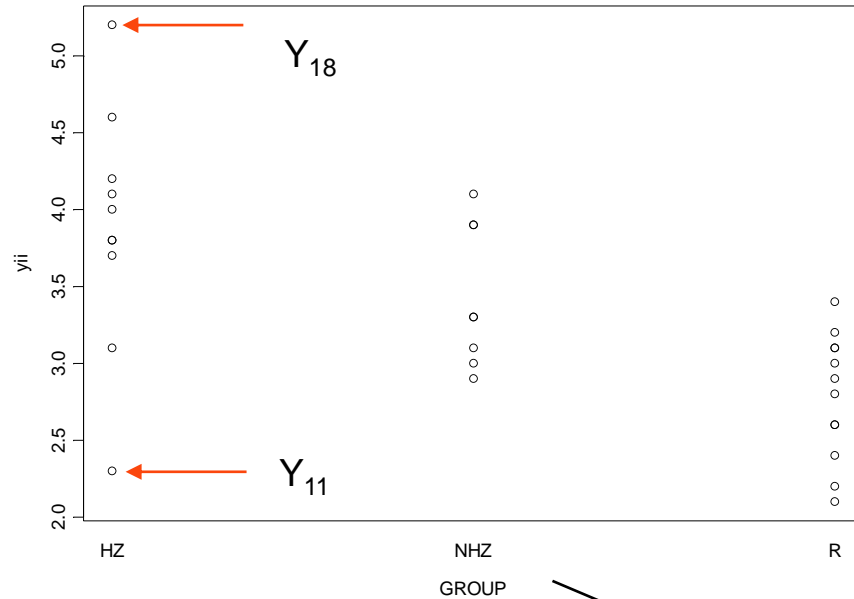
HZ	NHZ	R
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.5
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

Y_{ij} = observation j in group i



HZ $i=1$	NHZ $i=2$	R $i=3$
Y_{11}	Y_{21}	Y_{31}
Y_{12}	Y_{22}	Y_{32}
Y_{13}	Y_{23}	Y_{33}
Y_{14}	Y_{24}	Y_{34}
Y_{15}	Y_{25}	Y_{35}
Y_{16}	Y_{26}	Y_{36}
Y_{17}	Y_{27}	Y_{37}
Y_{18}	Y_{28}	Y_{38}
Y_{19}		Y_{39}
Y_{110}		Y_{310}
Y_{111}		Y_{311}
		Y_{312}

Scatterplot of the data

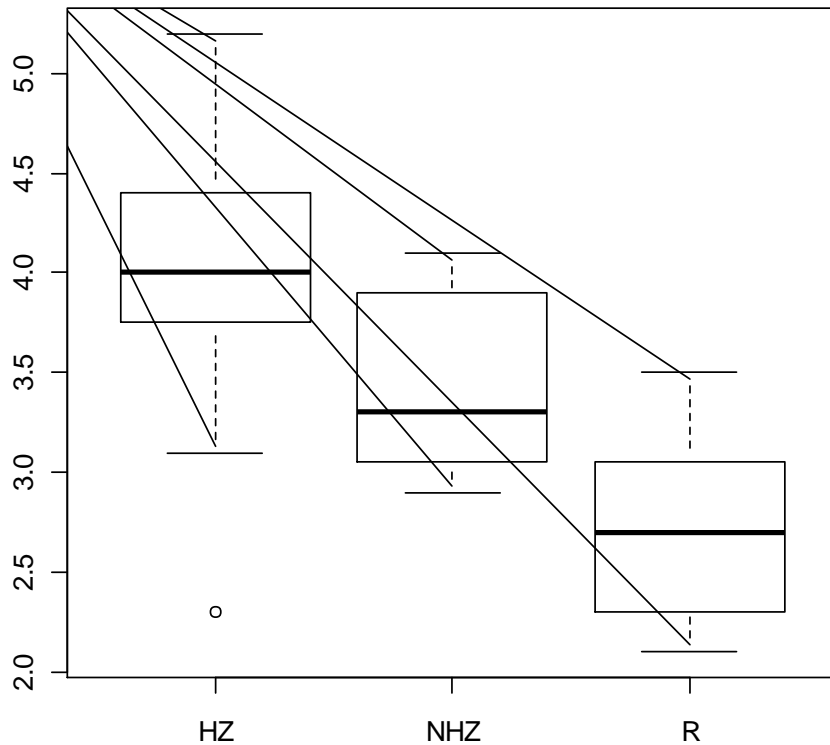


	HZ	NHZ	R
Y_{11}	2.3	3.0	3.0
	4.1	4.1	2.6
	4.2	3.9	3.1
	4.0	3.1	2.2
	4.6	3.3	2.1
	4.6	2.9	2.4
Y_{18}	3.8	3.3	2.8
	5.2	3.9	3.5
	3.1		2.9
	3.7		2.6
	3.8		3.1
			3.2

Response values.

Treatment group

Boxplot of the data



Patterns in the medians

Variability.

Populations

population	distribution
1) (HZ)	$N(\mu_1, \sigma^2)$
2) (NHZ)	$N(\mu_2, \sigma^2)$
3) (R)	$N(\mu_3, \sigma^2)$

- Mean and variance in the population.
- Normal distribution.

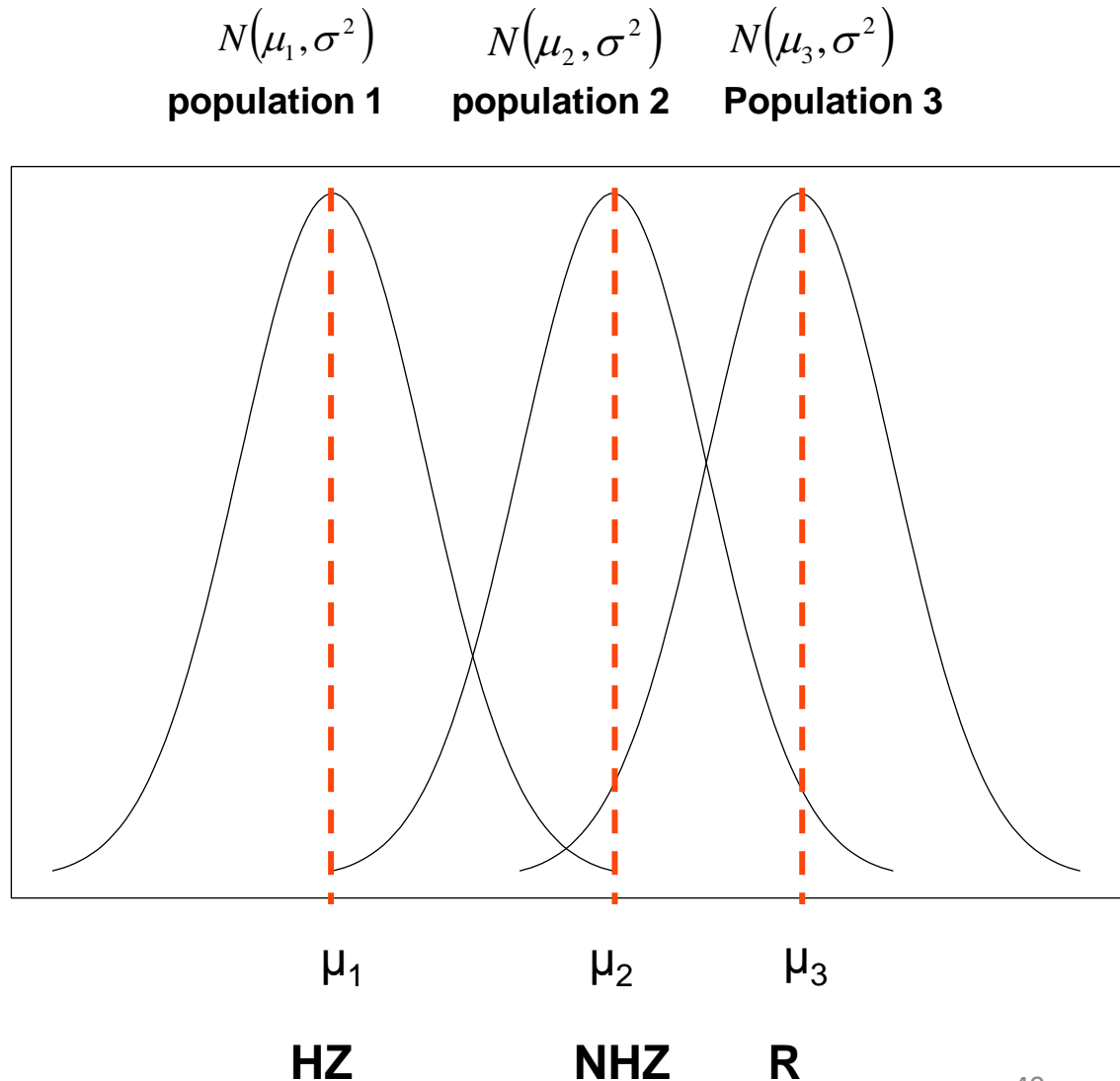
$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

Mean in the
population

Variance in the
population.

Three populations

- Constant variance: σ^2 .
- The null hypothesis



Formulation of the null hypothesis

Under the null hypothesis:

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

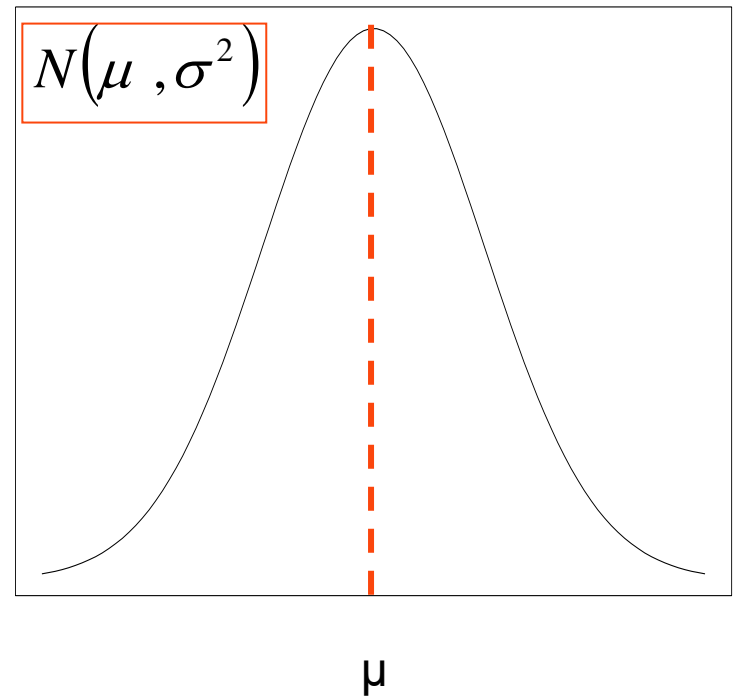
No effect of the treatment.

The null hypothesis

- Under the null hypothesis the means are equal

$$Y_{ij} \sim N(\mu, \sigma^2)$$

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



What is the alternative hypothesis ?

Under the alternative hypothesis

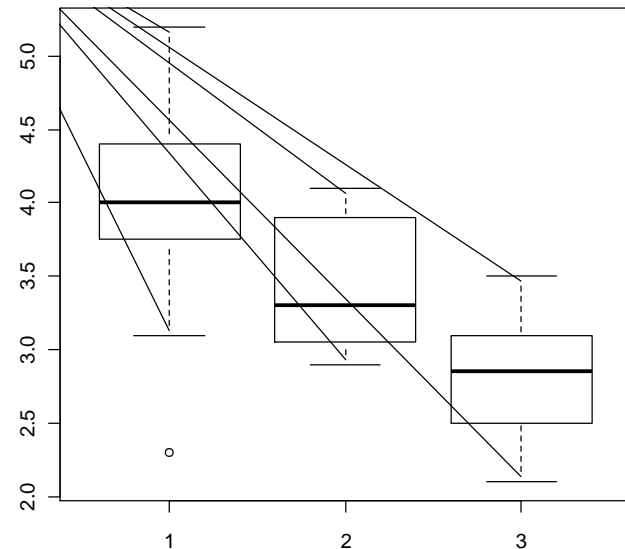
$$H_1 : \mu_i \neq \mu_l \quad \text{For at least one pair } i \text{ and } l \\ (i, l = 1, 2, \dots, k)$$

$$\mu_1 \neq \mu_2 \quad \text{and/or} \quad \mu_1 \neq \mu_3 \quad \text{and/or} \quad \mu_2 \neq \mu_3$$

Example 2: data in R

```
> con<-c(2.3,4.1,4.2,4.0,4.6,4.6,3.8,5.2,3.1,3.7,3.8,  
+        3.0,4.1,3.9,3.1,3.3,2.9,3.3,3.9,  
+        3.0,2.6,3.1,2.2,2.1,2.4,2.8,3.5,2.9,2.6,3.1,3.2)  
> gr<-c(rep(1,11),rep(2,8),rep(3,12))
```

```
> boxplot(split(con,gr))
```



Analysis of variance in R

```
> fit.1 <- aov(con ~ as.factor(gr))
```

```
> summary(fit.1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(gr)	2	7.6926	3.8463	11.318	0.0002499 ***
Residuals	28	9.5152	0.3398		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

One-Way ANOVA model

- The one way ANOVA model is a statistical model which we use in order to test the null hypothesis that the mean response across the factor level equal.
- It does not tell us which one is different.

One-Way ANOVA model

- Which groups are different ?
 - Post-hoc Pairwise comparisons.
 - Multiplicity issues.

Chapter 2:

Linear regression models with normal error

Donson: chapter 2.

Lindsey: chapter 9.

McCullagh & Nelder: chapter 3.

Simple linear regression model

$$Y_i = \beta_0 + \beta_1 \times x_i + \varepsilon_i$$

Y_i is the response variable.

x_i is the predictor (independent variable).

The observation is the pair (Y_i, x_i) ..

Sample of size n : $(Y_1, x_1), (Y_2, x_2), \dots, (Y_n, x_n)$

β_0 and β_1 are the unknown parameters of the model.

ε_i is a stochastic random variable (unobserved).

The error terms $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

- We assume for $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$

1 $\varepsilon_i \sim N(0, \sigma^2)$

$$E(\varepsilon_i) = 0$$

2

$$Var(\varepsilon_i) = \sigma^2$$

3

The distribution of the response

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

$$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_i}_{\mu_i}, \sigma^2) \quad i = 1, \dots, n$$

$$E[Y_i] = \beta_0 + \beta_1 x_i = \mu_i$$

$$Y_i \sim N(\mu_i, \sigma^2) \quad i = 1, \dots, n$$

The parameters to be estimated

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

- β_0 en β_1 are unknown parameters.
- β_1 – the slope.
- β_0 - the intercept.

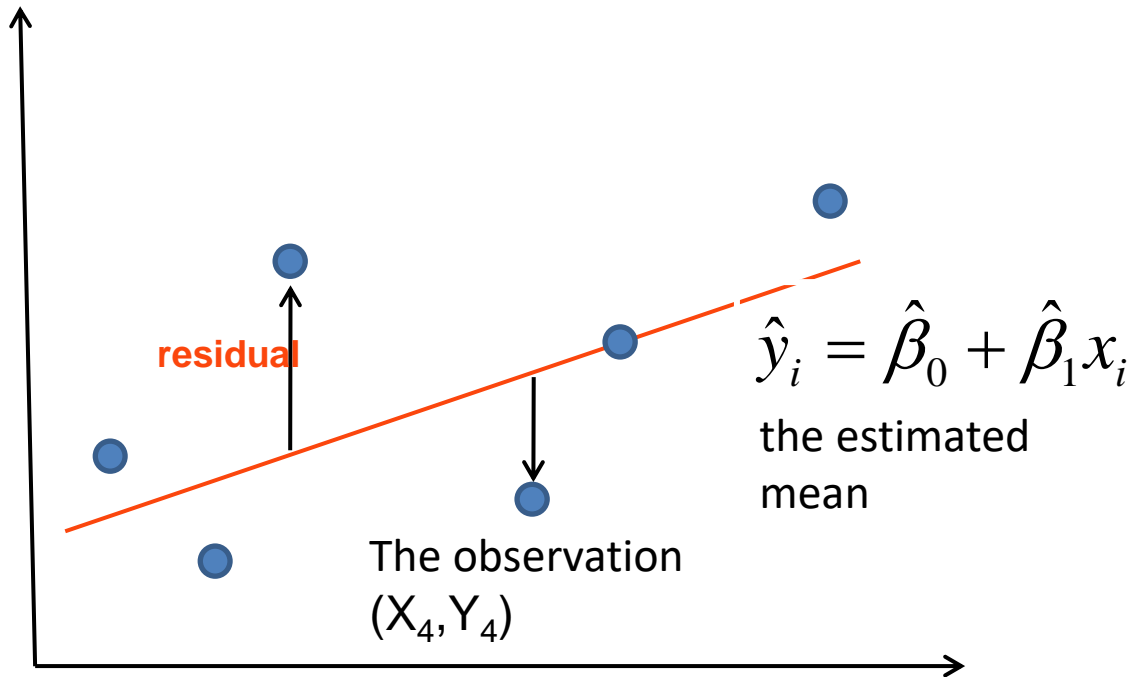
$$\varepsilon_i \sim N(0, \sigma^2)$$

The variance of the random error

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = \sigma^2$$

The residual



$$r_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i}$$

The estimated mean

Matrix notations

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y \\ \vdots \\ Y_n \end{bmatrix}$$

$$X = \begin{bmatrix} X_{11} & & X_{p1} \\ & & \\ & & \\ X_{1n} & & X_{pn} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Error structure (I)

$$E(Y) = E(X\beta) + E(\varepsilon) = X\beta = \mu$$

$$COV(Y) = \sigma^2 I$$

The response variables Y have equal variance they are uncorrelated (Identically Independently Distributed-iid).

Distribution of the response

Density function of the response

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y - \mu)^2}{2\sigma^2}\right)$$

Likelihood function

$$L(y_1, \dots, y_n, \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right)$$

Example 1: Plant weight data (Dobson)

Genetically similar seeds are randomly assigned to be raised either in

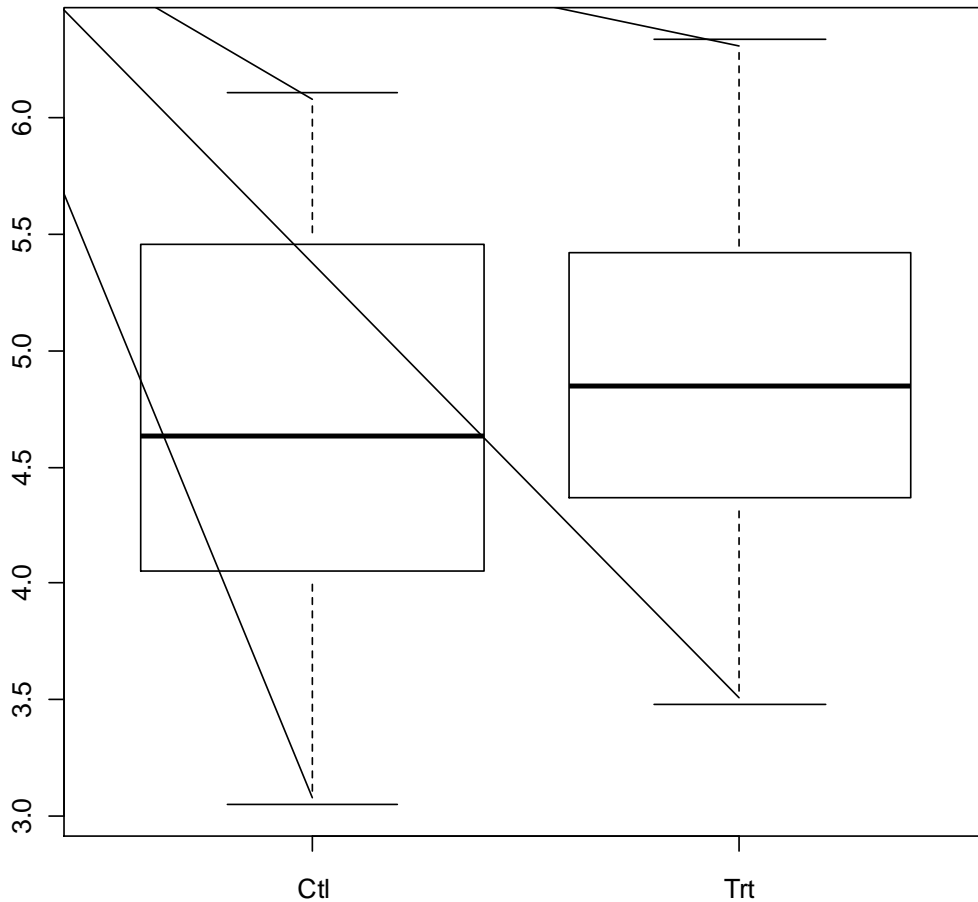
1. a nutritionally enriched environment (treatment)-treatment A in Dobson Table 6.6
2. standard conditions (control)

Response: dried weight in grams of the seeds.

Table 1.1: Plant weight from two different growing conditions

Control (1)	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment (2)	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

The data



The main question:
Are the mean in the
two treatment groups
equal ?

Model formulation (1): oneway ANOVA model

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

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

Assumptions:

1. Normality
2. Constant variance.

See slide 34

Model formulation (2): linear regression model

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

One-Way ANOVA as a linear regression model.

$$x_i = \begin{cases} 1 & T \\ 0 & C \end{cases}$$

Categorical predictor.

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 x_i & T \\ \beta_0 & C \end{cases}$$

Mean (per group).

R-Code and Output

```
> ctl <- c(4.17,5.18,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89, 4.32,4.69)
> group <- gl(2,20,labels=c("Ctl","Trt"))
> weight <- c(ctl,trt)
> cbind(weight,group)
```

 **gl()** function generates factor levels

Table 1.1: Plant weight from two different growing conditions

Control (1)	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
Treatment (2)	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69

The `aov()` and `lm()` functions in R

```
>lm(response ~ predictor(s))
```

```
>aov(response ~ factor(s))
```

One-way ANOVA as linear regression model

```
> fit.D9 <- lm(weight ~ group)
```

```
> summary(fit.D9)
```

Call:

```
lm(formula = weight ~ group)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.0710	-0.4692	0.0885	0.1983	1.3690

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.9920	0.2165	23.061	8.14e-15 ***
groupTrt	-0.3310	0.3061	-1.081	0.294

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6845 on 18 degrees of freedom

Multiple R-squared: 0.06098, Adjusted R-squared: 0.008817

F-statistic: 1.169 on 1 and 18 DF, p-value: 0.2939

ANOVA model

```
> fit.aov<-aov(weight ~ group)
> summary(fit.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	0.548	0.5478	1.169	0.294
Residuals	18	8.435	0.4686		

```
> anova(fit.aov)
```

Analysis of Variance Table

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	0.5478	0.5478	1.169	0.2939
Residuals	18	8.4349	0.4686		

$$\sqrt{0.4686} = 0.6845$$

Example 2: Body weight and gestational age (section 2.2.2 in Dobson)

Birth weights (g) and estimated gestational age (weeks) of 12 male and female babies born in a certain hospital.

Two predictors: age and gender.

Birth weight and gestational age for male and female babies

	Male		Female	
	Age (weeks)	Birth weight (g)	Age (weeks)	Birth weight (g)
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Means	38.33	3024.00	38.75	2911.33

Data in R

```
> bage<-c(40,38,40,35,36,37,41,40,37,38,40,38)
> gage<-c(40,36,40,38,42,39,40,37,36,38,39,40)
> bwei<-
  c(2968,2795,3163,2925,2625,2847,3292,3473,2628,3176,3421,2975)
> gwei<-
  c(3317,2729,2935,2754,3210,2817,3126,2539,2412,2991,2875,3231)
> age<-c(bage,gage)
> weight<-c(bwei,gwei)
> gender <- gl(2,12,24,labels=c("M","F"))
> dat2<-data.frame(weight,age,gender)
>Dat2
  weight age gender
1    2968  40      M
2    2795  38      M
3    3163  40      M
4    2925  35      M
5    2625  36      M
6    2847  37      M
```

The data

```
> plot(age,weight,pch=" ", main="Scatter Plot of Age and Weight")  
> points (age[gender=="F"],weight[gender=="F"],pch="o")  
> points (age[gender=="M"],weight[gender=="M"],pch="+")
```

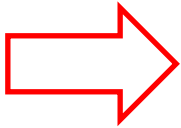


Does the growth rate equal for male and female ?

Model formulation

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2) \quad i = 1, \dots, n, j = 1(M), 2(F).$$

$$E(Y_{ij}) = \mu_{ij}$$



- Mean structure:
 - 4 possible models.

Model 0: model formulation

Gender and age do not have influence on the response.

$$Y_{ij} \sim N(\mu, \sigma^2) \quad i = 1, \dots, n, j = 1(M), 2(F).$$

$$E(Y_{ij}) = \mu$$

```
fit.lm.0 <- lm(weight ~ 1, data=dat2)
```

Model 0: R output

```
> fit.lm.0 <- lm(weight ~ 1, data=dat2)
```

```
> summary(fit.lm.0)
```

Call:

```
lm(formula = weight ~ 1, data = dat2)
```

Residuals:

Min	1Q	Median	3Q	Max
-555.67	-182.92	-16.17	216.83	505.33

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2967.67	57.58	51.54	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 282.1 on 23 degrees of freedom

Model 1: model formulation

Response is a function of age

$$Y_{ij} \sim N(\mu_i, \sigma^2) \quad i = 1, \dots, n, j = 1(M), 2(F).$$

$$\begin{aligned} E(Y_{ij}) = \mu_i &= \beta_0 + \beta_1 x_i \\ &= \beta_0 + \beta_1 \text{Age}_i \end{aligned}$$

```
fit.lm.1 <- lm(weight ~ age, data=dat2)
```

Model 1: R output

```
> summary(fit.lm.1)
```

Call:

```
lm(formula = weight ~ age, data = dat2)
```

Residuals:

Min	1Q	Median	3Q	Max
-262.032	-158.292	8.355	88.147	366.496

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1485.0	852.6	-1.742	0.0955 .
age	115.5	22.1	5.228	3.04e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 192.6 on 22 degrees of freedom

Multiple R-squared: 0.554, Adjusted R-squared: 0.5338

F-statistic: 27.33 on 1 and 22 DF, p-value: 3.04e-05

Model 2: model formulation

Response is a function of age and gender

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2) \quad i = 1, \dots, n, j = 1(M), 2(F).$$

$$\begin{aligned} E(Y_{ij}) = \mu_{ij} &= \beta_0 + \beta_1 x_i + \beta_2 G_i \\ &= \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{gender}F \end{aligned}$$

$$\text{where } \text{gender}F = \begin{cases} 1 & , \text{ girl} \\ 0 & , \text{ boy} \end{cases}$$

```
fit.lm.2 <- lm(formula = weight ~ age + gender, data = dat2)
```

Model 2: R output

```
summary(fit.lm.2)
```

```
Call:
```

```
lm(formula = weight ~ age + gender, data = dat2)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-257.49	-125.28	-58.44	169.00	303.98

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1610.28	786.08	-2.049	0.0532	.
age	120.89	20.46	5.908	7.28e-06	***
genderF	-163.04	72.81	-2.239	0.0361	*

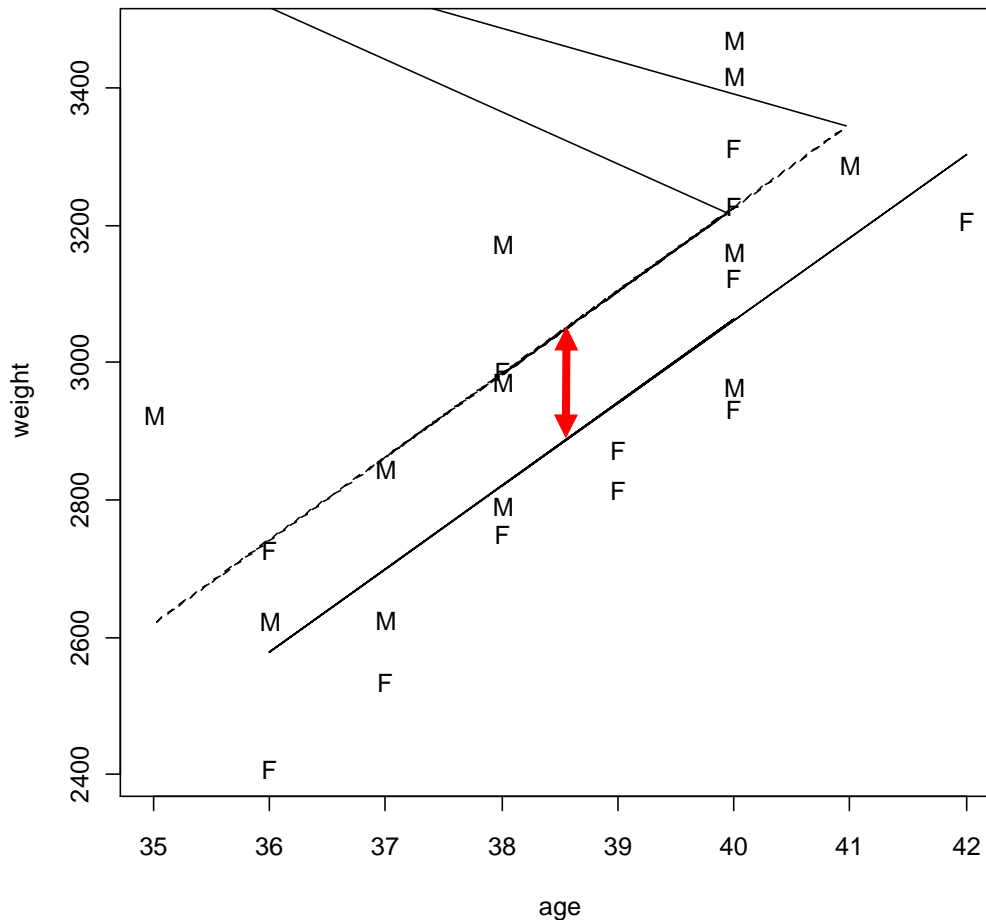
Signif. codes:	0	'***'	0.001	'**'	0.01
			'*'	0.05	'.'
				0.1	' '
					1

```
Residual standard error: 177.1 on 21 degrees of freedom
```

```
Multiple R-squared: 0.64, Adjusted R-squared: 0.6057
```

```
F-statistic: 18.67 on 2 and 21 DF, p-value: 2.194e-05
```

Data and predicted model (model 2)



Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1610.28	786.08	-2.049	0.0532	.
age	120.89	20.46	5.908	7.28e-06	***
genderF	-163.04	72.81	-2.239	0.0361	*

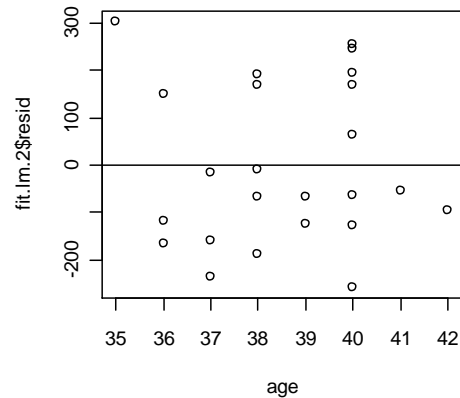
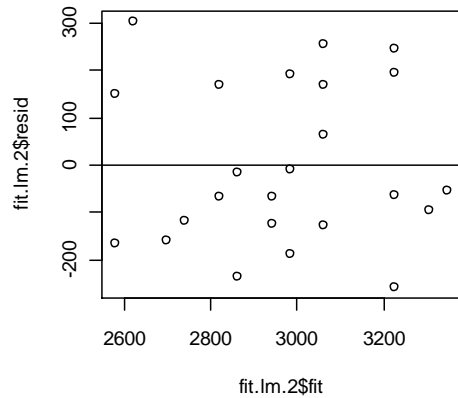
For boys

$$Y_{ij} = -1610.28 + 120.89 * \text{Age}$$

For girls

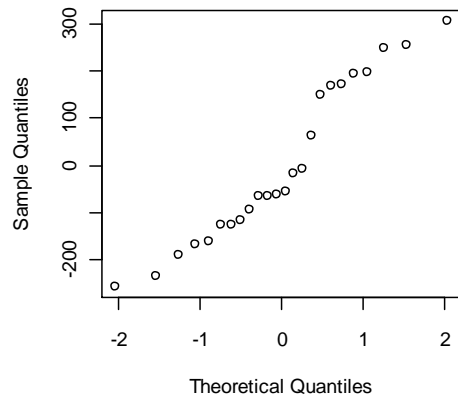
$$\begin{aligned}
 Y_{ij} &= -1610.28 + 120.89 * \text{Age} - 163.04 * (1) \\
 &= -1610.28 - 163.04 + 120.89 * \text{Age} \\
 &= -1773.32 + 120.89 * \text{Age}
 \end{aligned}$$

Diagnostic plots



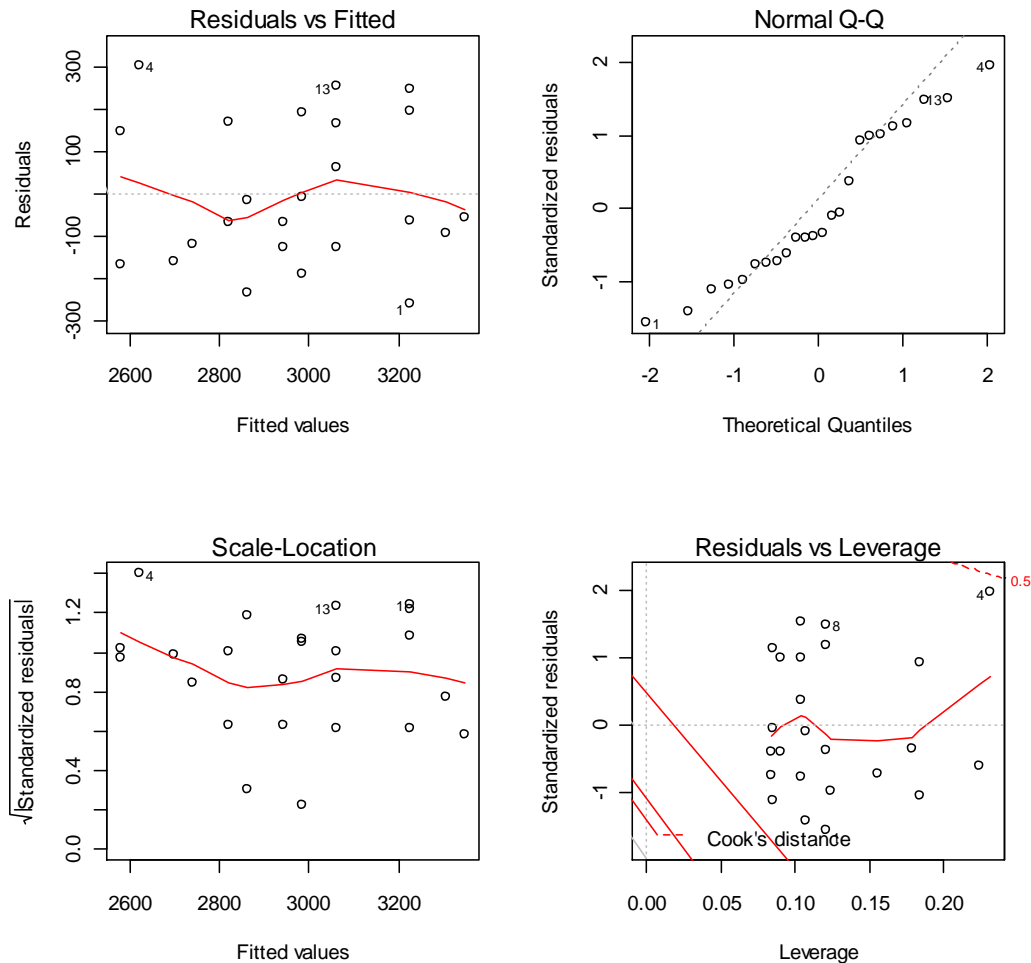
```
par(mfrow=c(2,2))  
plot(fit.lm.2$fit, fit.lm.2$resid)  
abline(0,0)  
plot(age, fit.lm.2$resid)  
abline(0,0)  
qqnorm(fit.lm.2$resid)
```

Normal Q-Q Plot



Diagnostic plots

```
> plot(fit.lm.4)
```



The likelihood function

Likelihood function

$$L(y_1, \dots, y_n, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \prod_{i=1}^n \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2} \right)$$

$-2\log(L)$

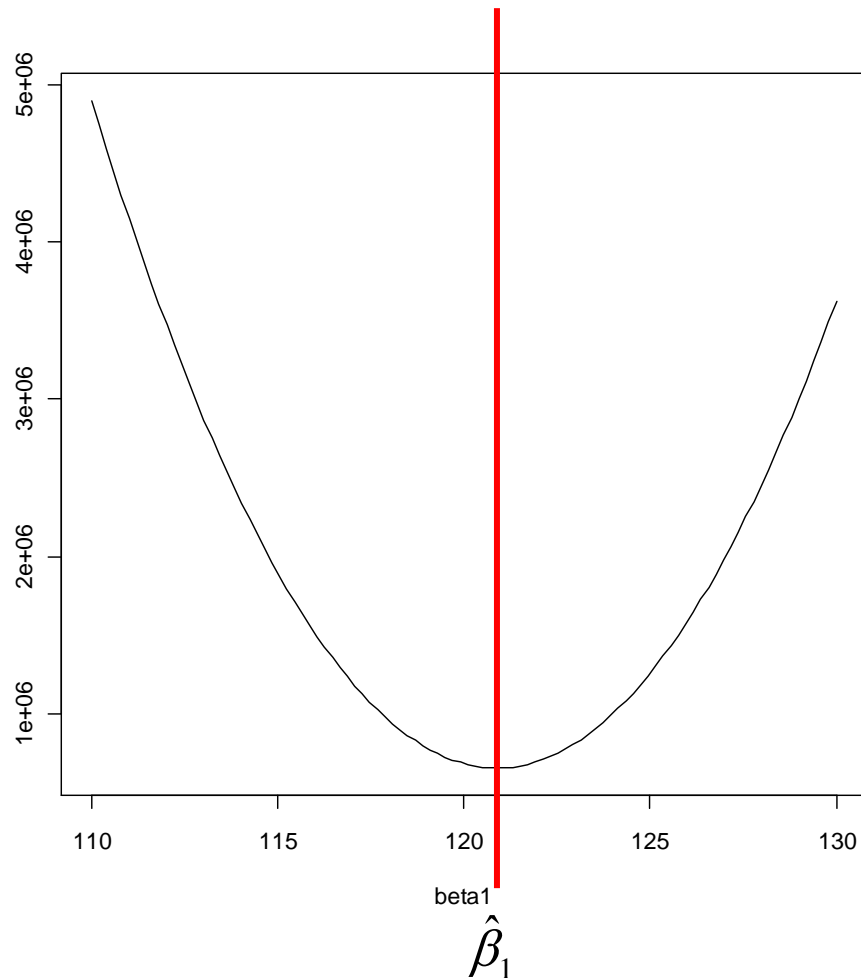
$$-2\ell = n \log(2\pi\sigma^2) + \sum_{i=1}^n \frac{-(y_i - \mu_i)^2}{\sigma^2}$$

The likelihood function

Maximizing the likelihood is equivalent to minimize the sum of squares

$$\sum_j \sum_{i=1}^n (y_{ij} - \mu_{ij})^2$$

The likelihood function



$$RSS(\mu) = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \beta_1 x_i + \hat{\beta}_2 G_i))^2$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1610.28	786.08	-2.049	0.0532 .
age	120.89	20.46	5.908	7.28e-06 ***
genderF	-163.04	72.81	-2.239	0.0361 *

$$RSS(\mu) = \sum_{i=1}^n (y_i - (1610.2 + \beta_1 x_i + 163.04 G_i))^2$$

The value that
maximizes the
likelihood: 120.89

Model 2/3: model formulation

Let us consider two models

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_i$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_i$$

What is the difference between the models ?

Model 2 & 3: sum of squares

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_1 x_{ij}$$

$$RSS_2(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_1 x_{ij}))^2$$

3 parameters

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j} x_{ij}$$

$$RSS_3(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j} x_{ij}))^2$$

4 parameters

Model 2 & 3 in R

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_1 x_{ij}$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j} x_{ij}$$

```
> fit.lm.2 <- lm(weight ~ age + gender, data=dat2)
> fit.lm.3 <- lm(weight ~ age + gender+age:gender, data=dat2)
```

Model 2 & 3 in R

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_1 x_{ij}$$

Call:

```
lm(formula = weight ~ age + gender, data = dat2)
```

Residuals:

Min	1Q	Median	3Q	Max
-257.49	-125.28	-58.44	169.00	303.98

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1610.28	786.08	-2.049	0.0532 .
age	120.89	20.46	5.908	7.28e-06 ***
genderF	-163.04	72.81	-2.239	0.0361 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 177.1 on 21 degrees of freedom

Multiple R-squared: 0.64, Adjusted R-squared: 0.6057

F-statistic: 18.67 on 2 and 21 DF, p-value: 2.194e-05

Model 2 & 3 in R

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

The slope of age depends on gender.

```
> summary(fit.lm.3)
```

Call:

```
lm(formula = weight ~ age + gender + age:gender, data = dat2)
```

Residuals:

Min	1Q	Median	3Q	Max
-246.69	-138.11	-39.13	176.57	274.28

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1268.67	1114.64	-1.138	0.268492	
age	111.98	29.05	3.855	0.000986	***
genderF	-872.99	1611.33	-0.542	0.593952	
age:genderF	18.42	41.76	0.441	0.663893	

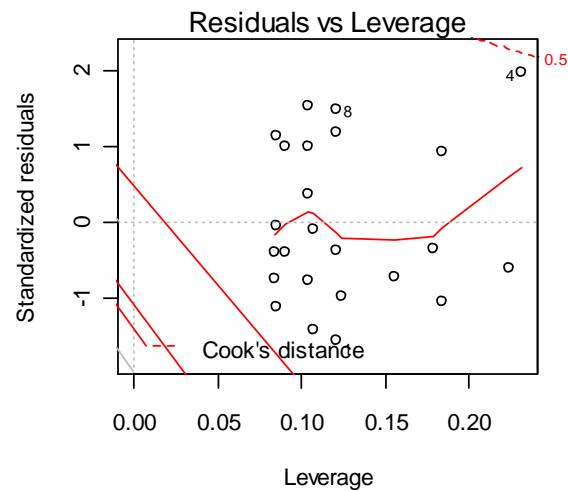
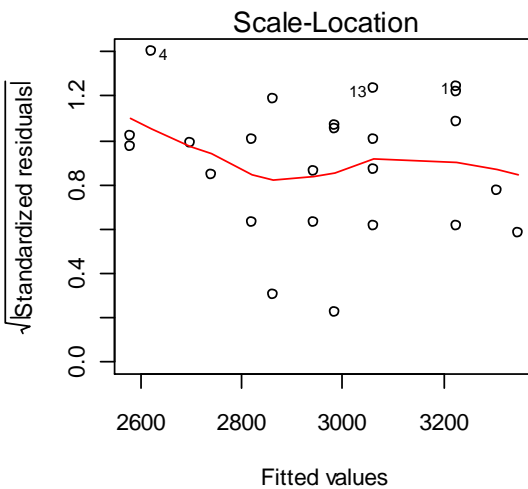
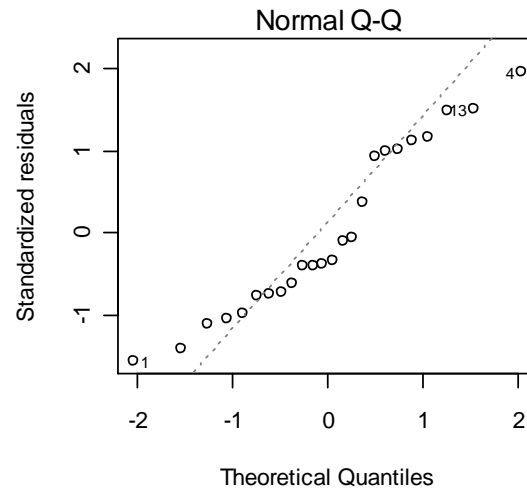
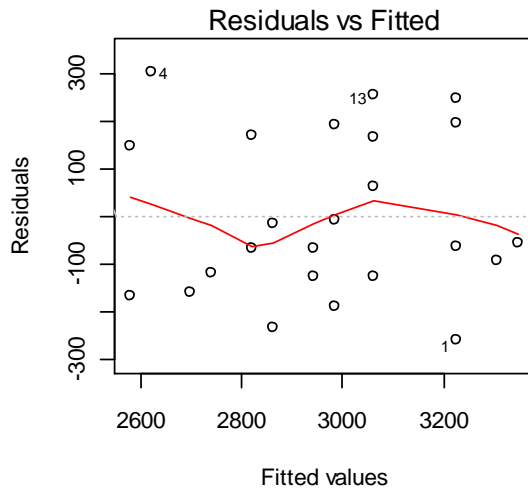
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 180.6 on 20 degrees of freedom

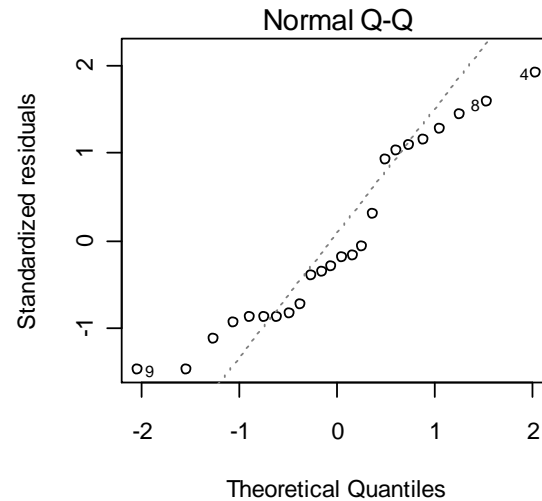
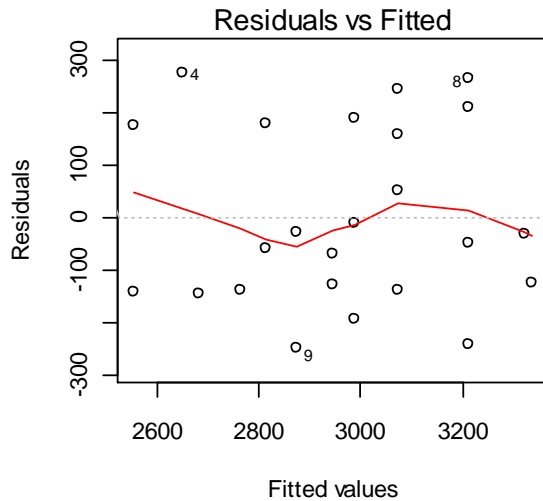
Multiple R-squared: 0.6435, Adjusted R-squared: 0.59

F-statistic: 12.03 on 3 and 20 DF, p-value: 0.0001010

Diagnostic plot model 2



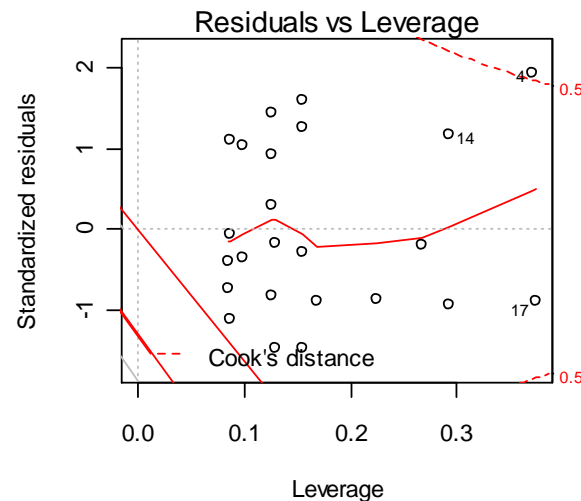
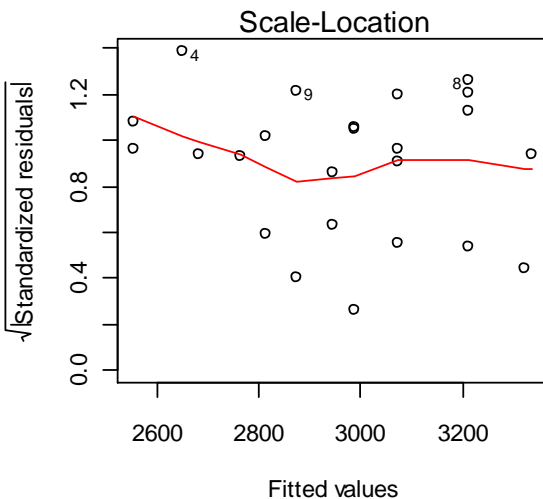
Diagnostic plot model 3



For both models: no systematic patterns in relation to fitted values.

Points in the QQ normal plot close to the line

Very little difference between the models.



Model 2 & 3: F test

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

The only difference between the models is that model 3 has two different slopes.

We can formulate the following hypotheses

$$H_0 : E(Y_{ij}) = \beta_{0j} + \beta_{1j}x_{ij}$$

$$H_1 : E(Y_{ij}) = \beta_{0j} + \beta_{1j}x_{ij}$$

Model 2 & 3: F test

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

$$RSS_2(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

$$RSS_3(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

3 parameters

4 parameters

$RSS_2(\mu_{ij})$ with $(N - 3)$ df

$RSS_3(\mu_{ij})$ with $(N - 4)$ df

$$F = \frac{(RSS_2(\mu_{ij}) - RSS_3(\mu_{ij})) / (4 - 3)}{RSS_3(\mu_{ij}) / (24 - 4)}$$

Under the null hypothesis

$$F \sim f(1, 20)$$

Model 2 & 3: F test in R

$$RSS_2(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

$$RSS_3(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

$$F = \frac{(RSS_2(\mu_{ij}) - RSS_3(\mu_{ij})) / (4 - 3)}{RSS_3(\mu_{ij}) / (24 - 4)}$$

```
> anova(fit.lm.2, fit.lm.3)
```

Analysis of Variance Table

Model 1: weight ~ age + gender

Model 2: weight ~ age + gender + age:gender

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	658771				
2	20	652425	1	6346.2	0.1945	0.6639

General F test

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

$$E(Y_{ij}) = \mu_{ij} = \beta_{0j} + \beta_{1j}x_{ij}$$

$$RSS_R(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

$$RSS_F(\mu_{ij}) = \sum_{i=1}^n (y_i - (\hat{\beta}_{0j} + \beta_{1j}x_{ij}))^2$$

$$RSS_R(\mu_{ij}) \text{ with } df_R = N - m$$

$$RSS_F(\mu_{ij}) \text{ with } df_F = N - (m + p)$$

$$F = \frac{(RSS_R(\mu_{ij}) - RSS_F(\mu_{ij})) / (df_R - df_F)}{RSS_F(\mu_{ij}) / df_F}$$

Under the null hypothesis

$$F \sim f((df_R - df_F), df_F)$$

Chapter 3:

Generalized linear models

Donson: chapter 3.

Lindsey: chapter 1.

McCullagh & Nelder: chapter 2.

Generalized linear models (GLM)

A framework for model fitting.

Examples:

- when an outcome (a response) is measured as a success or failure.
- when we count the number of events over a fixed period.

Generalized linear models (GLM) are used to fit fixed effect models to certain types of data that are not normally distributed.

Generalized – not limited to normally distributed data.

Linear – models use a linear combination of variables to ‘predict’ the response.

Components of a GLM

1. **Random component**- the **probability distribution** of the response.
2. **Systematic component (linear predictor)**: the predictor variables are (e.g., X_1 , X_2 , etc). These variable enter to the model in a linear manner.

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

3. **Link function**-Specify the relationship between the mean random component (i.e., $E(Y)$) and the systematic component.

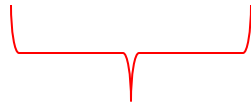
Example 1: linear regression models

Random component: the distribution of the response

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_\varepsilon^2)$$

The systematic component: the linear predictor

$$E(Y_i) = \beta_0 + \beta_1 x_i$$



Linear
predictor

The link function

$$\eta = \beta_0 + \beta_1 X_i$$

$$g(E(Y_i)) = \eta$$


$g = 1$, identity function



Link function

Components of a GLM: linear regression models

For the case with p predictors (and p unknown parameters)

$$E(Y_i) = \mu_i = \sum_{j=1}^p \beta_j x_j$$


$$\eta = \sum_{j=1}^p \beta_j x_j$$

The link function (=the link between the random and the systematic part)

$$Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$$

$$g(\mu) = g(E(Y_i)) = \eta$$

$$g = 1$$

Example 2: binary data

Dichotomous (binary) with a fixed numbers of trials
(Binomial distribution) Success/failure.

Dose response experiment (Table 7.2 in Dobson):

Dose	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Beetles	59	60	62	56	63	59	62	60
Killed	6	13	18	28	52	53	61	60

Random component: example of binary data

Dose	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Beetles	59	60	62	56	63	59	62	60
Killed	6	13	18	28	52	53	61	60

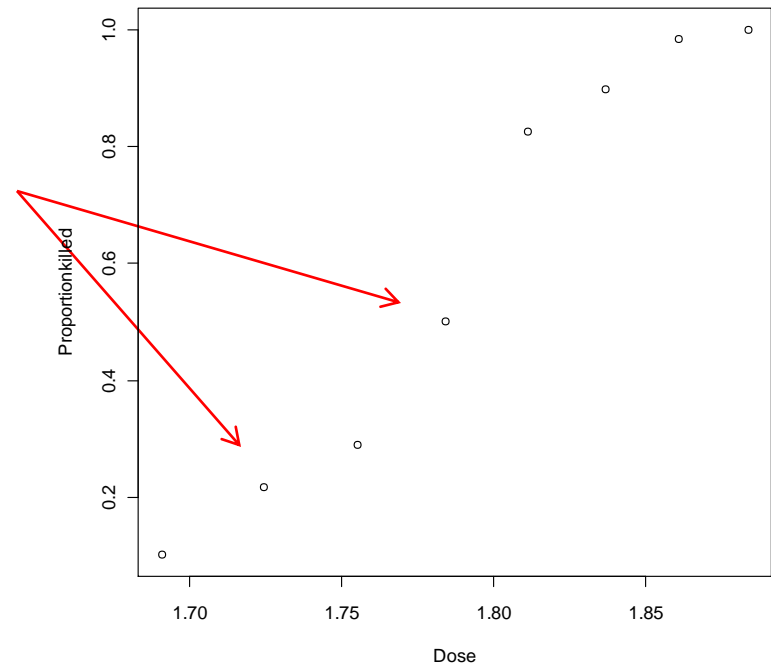
Proportion of the killed beetles

$$Y_{ij} = \begin{cases} 1 & \text{alive} \\ 0 & \text{killed} \end{cases}$$

$$\frac{\sum Y_{ij}}{n_j}$$

$$Y_{ij} \sim B(1, \pi_{ij})$$

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_{ij}$$

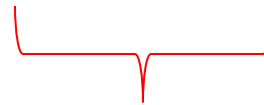


Systematic component: dependency of the predictor – the linear predictor

The systematic component of the model consists of a set of explanatory variables and some linear function of them.

$$\pi_j = f(dose_i) = f(d_i)$$

$$\pi_j = f(d_i) = f(\beta_0 + \beta_1 d_j)$$



The linear predictor

The Link function

The expected values of
the response variable

$$E(Y_{ij}) = \pi_j$$

The systematic part

$$\pi_j = f(\beta_0 + \beta_1 d_j) = f(\eta)$$

$$\pi_j = \frac{e^{\beta_0 + \beta_1 d_j}}{1 + e^{\beta_0 + \beta_1 d_j}}$$

The logistic
function to
describe the
mean, $E(Y_{ij})$, as
a function of the
linear predictor

$$g(E(Y_{ij})) = g(\pi_j) = \eta$$

Values between 0
and 1

The Link function (**logit link function** for binary data)

The link between the expected values of the response variable and the linear predictor

$$g(\pi_j) = \log\left(\frac{\pi_j}{1 - \pi_j}\right)$$

$$\log\left(\frac{\pi_j}{1 - \pi_j}\right) = \log\left(e^{\beta_0 + \beta_1 d_j}\right)$$

$$\Rightarrow g(\pi_j) = \log\left(e^{\beta_0 + \beta_1 d_j}\right) = \beta_0 + \beta_1 d_j = \eta$$

Example 3: count data

- In a list of 41 events, respondents were asked to note which had occurred within the last 18 months.
- The result is given as:

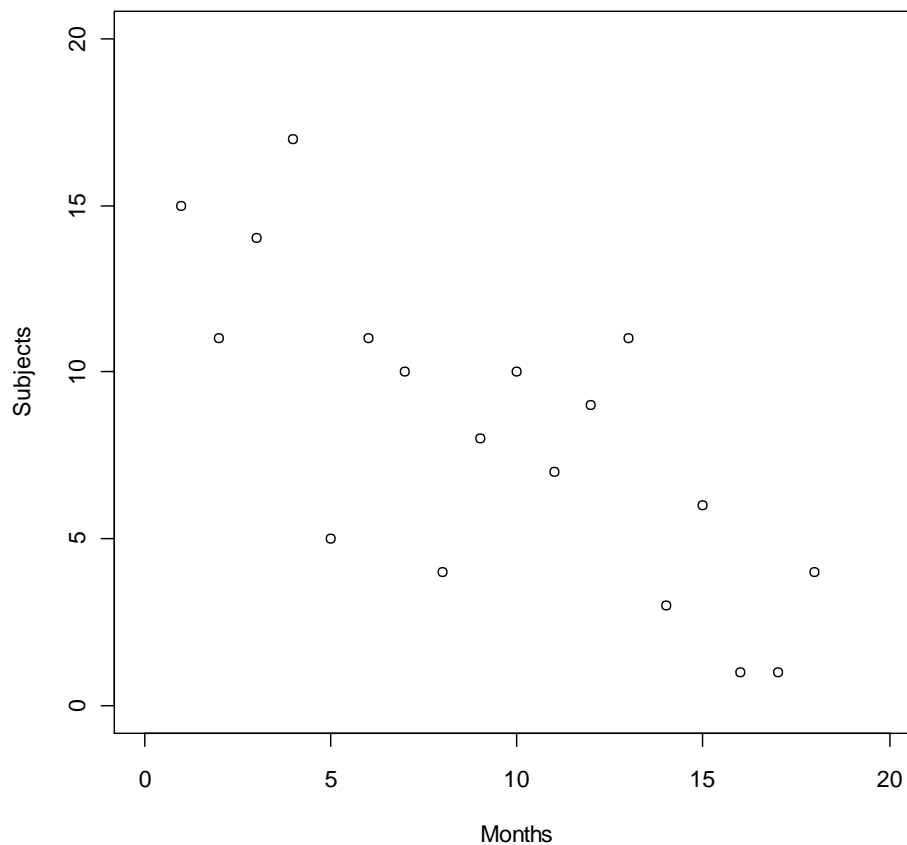
Month	1	2	3	4	5	6	7	8	9
Respondents	15	11	14	17	5	11	10	4	8
Month	10	11	12	13	14	15	16	17	18
Respondents	10	7	9	11	3	6	1	1	14

$$Y_t \sim \text{Poisson}(\mu(t))$$

Random component: example of count data

$$Y_t \sim \text{Poisson}(\mu_t)$$

$$E(Y_t) = \mu_t$$



Systematic component: dependency of the predictor – the linear predictor

$$\mu_t = f(\text{time}) = f(t) = f(\underbrace{\beta_0 + \beta_1 t}_{\text{The linear predictor}}) = f(\eta)$$

The linear predictor

$$\mu_t = f(\beta_0 + \beta_1 t) = e^{\beta_0 + \beta_1 t}$$

The Link function: count data (**log link**)

The expected values of
the response variable

$$E(Y_t) = \mu_t$$

The systematic part

$$\mu_t = e^{\beta_0 + \beta_1 t}$$

$$g(E(Y_t)) = g(\mu_t) = \eta$$

$$g(\mu_t) = \log(\mu_t) = \log(e^{\beta_0 + \beta_1 t}) = \beta_0 + \beta_1 t = \eta$$

Example 4: mortality rate (Table 3.2, Dobson)

Number of deaths from coronary heart diseases and population size per 5 years age group in new south Wales, Australia 1991.

Data in R:

```
> age<-c(32,37,42,47,52,57,62,67)
> deaths<-c(1,5,5,12,25,38,54,65)
> pop<-
c(17742,16554,16059,13083,10784,9645,10706,9933)
> data.frame(age,deaths,pop,(deaths/pop)*100000)
  age deaths   pop      rate per year
1  32      1 17742      5.636343
2  37      5 16554     30.204180
3  42      5 16059     31.135189
4  47     12 13083     91.722082
5  52     25 10784    231.824926
6  57     38  9645    393.986522
7  62     54 10706    504.390062
8  67     65  9933    654.384375
```

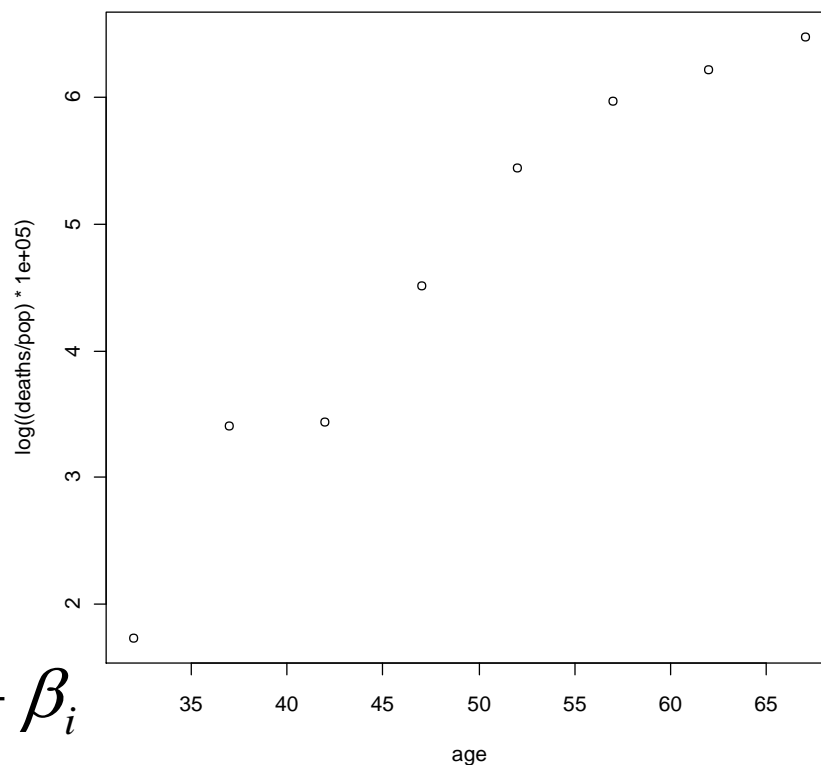
Random component: example of count data

$$Y_i \sim \text{Poisson}(\mu_i)$$

$$E(Y_i) = \mu_i$$

$$\mu_i = n_i e^{\beta_i}$$

$$g(\mu_i) = \log(\mu_i) = \log(n_i) + \beta_i$$



Chapter 4:

The Exponential family

Donson: chapter 3.

Lindsey: chapter 1.

McCullagh & Nelder: chapter 2.

The exponential family

Most of the commonly used statistical distributions, e.g. Normal, Binomial and Poisson, are members of the exponential family of distributions.

$$f(y) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}$$

Where ϕ is the dispersion parameter and θ is the canonical parameter and

$a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions

The Exponential family

- The parameters θ_i and ϕ are essentially location and scale parameters.
- It can be shown that if Y_i has a distribution in the exponential family then it has mean and variance

$$E(Y_i) = \mu_i = b(\theta_i)$$

And

$$Var(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi_i)$$

Example: normal distribution

$$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \mu_i)^2}{2\sigma^2}}$$
$$= \exp \left\{ \left[y_i \mu_i - \frac{\mu_i^2}{2} \right] \frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right\}$$



$$\theta_i = \mu_i,$$

$$b(\theta_i) = \theta_i^2 / 2$$

$$a_i(\phi) = \sigma^2$$

$$c(y_i, \phi) = -[y_i^2 / \phi + \log(2\pi\phi)] / 2.$$

Example: Bernoulli distribution

$$Y = \begin{cases} 1 & \text{if even of interest has occurred} \\ 0 & \text{Otherwise} \end{cases}$$

$$p(y | \theta) = \theta^y (1 - \theta)^{1-y} = \exp\left\{y \log \frac{\theta}{1 - \theta} + \log(1 - \theta)\right\}$$



$$a = 1$$

$$b(\theta) = \log(1 + \exp(\theta))$$

$$c(y) = 1$$

$$E(y) = \mu = b'(\theta) = e^{\theta} (1 + \exp(\theta))^{-1}$$

$$\text{var}(y) = \mu(1 - \mu)$$

Example: Binomial distribution

$$Z_i = \begin{cases} 1 \\ 0 \end{cases} \quad \longrightarrow \quad Y_i = \sum_{i=1}^n Z_i \quad \longrightarrow \quad Y_i \sim B(n, \pi_i)$$

$$p(y_i | \theta) = \binom{n_i}{y_i} \theta^{y_i} (1 - \theta)^{n - y} =$$

$$\exp \left\{ y_i \log \left[\frac{\theta_i}{1 - \theta_i} \right] + n_i \log(1 - \theta_i) + \log \binom{n_i}{y_i} \right\}$$

$$a_i(\phi) = 1, \quad b(\theta_i) = \log(1 + \exp(\theta_i))$$

$$c(y) = \log \binom{n_i}{y_i}$$

$$E(y) = \mu = b'(\theta_i) = e^{\theta} (1 + \exp(\theta_i))^{-1}$$

$$\text{var}(y) = \mu(1 - \mu) / n$$

Poisson distribution

$$Y_i \sim \text{Poisson}(\mu)$$

$$f(y_i, \theta_i) = e^{-\theta_i} \frac{\theta_i^{y_i}}{y_i!} \exp\{y_i \log \theta_i - \theta_i - \log(y_i!)\}$$

$$a_i(\phi) = 1$$

$$b(\theta) = \exp(\theta)$$

$$c(y) = -\log(y!)$$

$$E(y) = \mu = b'(\theta) = \exp(\theta)$$

$$\text{var}(y) = \mu$$

Gamma distribution

$$f(y_i; \mu_i, \nu) = \left(\frac{\nu}{\mu_i} \right) \frac{y_i^{\nu-1} e^{-\frac{\nu y_i}{\mu_i}}}{\Gamma(\nu)}$$
$$= \exp \left\{ \begin{aligned} &[-y_i / \mu_i - \log(\mu_i)]\nu + (\nu - 1) \log(y_i) \\ &+ \nu \log(\nu) - \log[\Gamma(\nu)] \end{aligned} \right\}$$

where

$$\theta_i = -1 / \mu_i,$$

$$b(\theta_i) = -\log(-\theta_i)$$

$$a_i(\phi) = 1 / \nu, \text{ and}$$

$$c(y_i, \phi) = (\nu - 1) \log(y_i) + \nu \log(\nu) - \log[\Gamma(\nu)].$$

The Canonical link function: Poisson distribution

The canonical link function is given by

$$g(b') = X\beta = \theta$$

Where **b** is obtained from the general exponential density form .

The link function

$$Y_i \sim \text{Poisson}(\mu)$$

$$g(\mu) = \log(\mu) = \theta$$

$$a_i(\phi) = 1$$

$$b(\theta) = \exp(\theta)$$

$$c(y) = -\log(y!)$$

$$E(y) = \mu = b'(\theta) = \exp(\theta)$$

$$\text{var}(y) = \mu$$

The canonical link function: Binomial distribution

$$Z_i = \begin{cases} 1 \\ 0 \end{cases} \quad \longrightarrow \quad Y_i = \sum_{i=1}^n Z_i \quad \longrightarrow \quad Y_i \sim B(n, \pi_i)$$

$$p(y_i | \theta) = \exp \left\{ y_i \log \left[\frac{\theta_i}{1 - \theta_i} \right] + n_i \log(1 - \theta_i) + \log \binom{n_i}{y_i} \right\}$$

The link function

$$g(\mu) = \log \left(\frac{\mu}{1 - \mu} \right)$$

$$\log \left(\frac{\mu}{1 - \mu} \right) = \log \left(\frac{\frac{e^\theta}{1 + e^\theta}}{\frac{1}{1 + e^\theta}} \right) = \log(e^\theta)$$

$$a_i(\phi) = 1, \quad b(\theta_i) = \log(1 + \exp(\theta_i))$$

$$c(y) = \log \binom{n_i}{y_i}$$

$$E(y) = \mu = b'(\theta_i) = e^{\theta_i} (1 + \exp(\theta_i))^{-1}$$

$$\text{var}(y) = \mu(1 - \mu) / n$$

The canonical link function: Normal distribution

$$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu_i)^2}{2\sigma^2}}$$
$$= \exp \left\{ \left[y_i \mu_i - \frac{\mu_i^2}{2} \right] \frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right\}$$



$$Y_i \sim N(\mu, \sigma^2)$$

$$g(\mu) = 1 \times \mu$$

$$\theta_i = \mu_i,$$

$$b(\theta_i) = \theta_i^2 / 2$$

$$\mu = b'(\theta_i) = \frac{2\theta}{2}$$

Canonical link function

Table showing the distribution with their link function and its name

Distribution	Link function ($g(\mu)$)	Name
Bernoulli	$\log(\mu/(1-\mu))$	Logit
Binomial	$\log(\mu/(k-\mu))$	Logit
Negative Binomial	$\log(\mu/(k+\mu))$	Logit
Poisson	$\log(\mu)$	Log
Gamma/ Exponential	$1/\mu$	Inverse
Normal	μ	Identity

Chapter 5

Generalized linear model function in R

The `glm()` Function

Generalized linear models can be fitted in R using the `glm()` function, which is similar to the `lm` function for fitting linear models.

The arguments to a `glm()` call are as follows:

```
glm(formula,family,link,data,...)
```

The first argument of the function is a model formula, which defines the response and linear predictor.

From `glm ()` function we have the following family

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

Link Function

- The link function links the response mean μ to the linear predictor η .

- Identity: $g(\mu) = \mu$

- Log: $g(\mu) = \log(\mu)$

- Logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$

- Probit : $g(\mu) = \phi^{-1}(\mu)$

- Comp. Log-log: $g(\mu) = \log(-\log(1-\mu))$

- Power: $g(\mu) = \mu^{\lambda}$, Where λ is the value in the power entry field.

Mainly for binary data (we will speak about this in a later stage in the course)


Link Function and distribution

- For each response distribution in the exponential family, there exists a special link function, the canonical link, for which $\theta = \eta$. The canonical links expressed in terms of the mean parameter μ are
- Normal: $g(\mu) = \mu$
- Inverse Gaussian $g(\mu) = \mu^{-2}$
- Gamma $g(\mu) = \mu^{-1}$
- Poisson $g(\mu) = \log(\mu)$
- Binomial $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- **Note** Some links are not appropriate for all distributions; logit, probit, and complementary log-log links are only appropriate for the binomial distribution.

Example: binary data with logit link

- Model with Binomial family and logit link function
- Fitting the model with the glm() function:

```
> model1 <- glm(Y ~ X*d, family=binomial(link=logit))
```

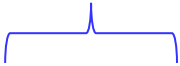


$$\eta = \beta_0 + \beta_1 x_i + \beta_2 d_{ij} + \beta_3 x_i * d_{ij}$$

A model with two predictors and interaction.

Alternative code

```
> model1 <- glm(y ~ X+d+X:d, family=binomial(link=logit))
```



$$\eta = \beta_0 + \beta_1 x_i + \beta_2 d_{ij} + \beta_3 x_i * d_{ij}$$

Extractor functions in R

- The `glm` function returns an object of class `c("glm", "lm")`.
- There are several `glm` or `lm` methods available for accessing/displaying components of the `glm` object, including:
 - `residuals()`
 - `fitted()`
 - `predict()`
 - `coef()`
 - `deviance()`
 - `formula()`
 - `summary()`

Extractor functions in R

- `Summary` to obtain more detailed information about the model :
- `residuals` or `resid`, for the deviance residuals
- `fitted` or `fitted.values`, for the fitted values (estimated probabilities)
- `predict`, for the linear predictor (estimated logits)
- `coef` or `coefficients`, for the coefficients, and
- `deviance`, for the deviance.

The `predict()` function in R

- The `predict()` function obtains predictions and optionally estimates standard errors of those prediction from a fitted glm objects.
- The general call is:

```
predict(object, newdata = NULL, type = c("link",  
"response", "terms"), se.fit = FALSE, dispersion = NULL,  
terms = NULL, na.action = na.pass, ...)
```

The `update()` function in R

- The `update()` function in R can be used to modify a fitted model by dropping some of the terms.
- The general call of the function is given as:

```
Update(old model, ~, . - or + the term we want to drop/ad)
```


Chapter 6:

Models for Binary data

Donson: chapter 7.

Lindsey: chapter 2.

McCullagh & Nelder: chapter 4.

Binary data

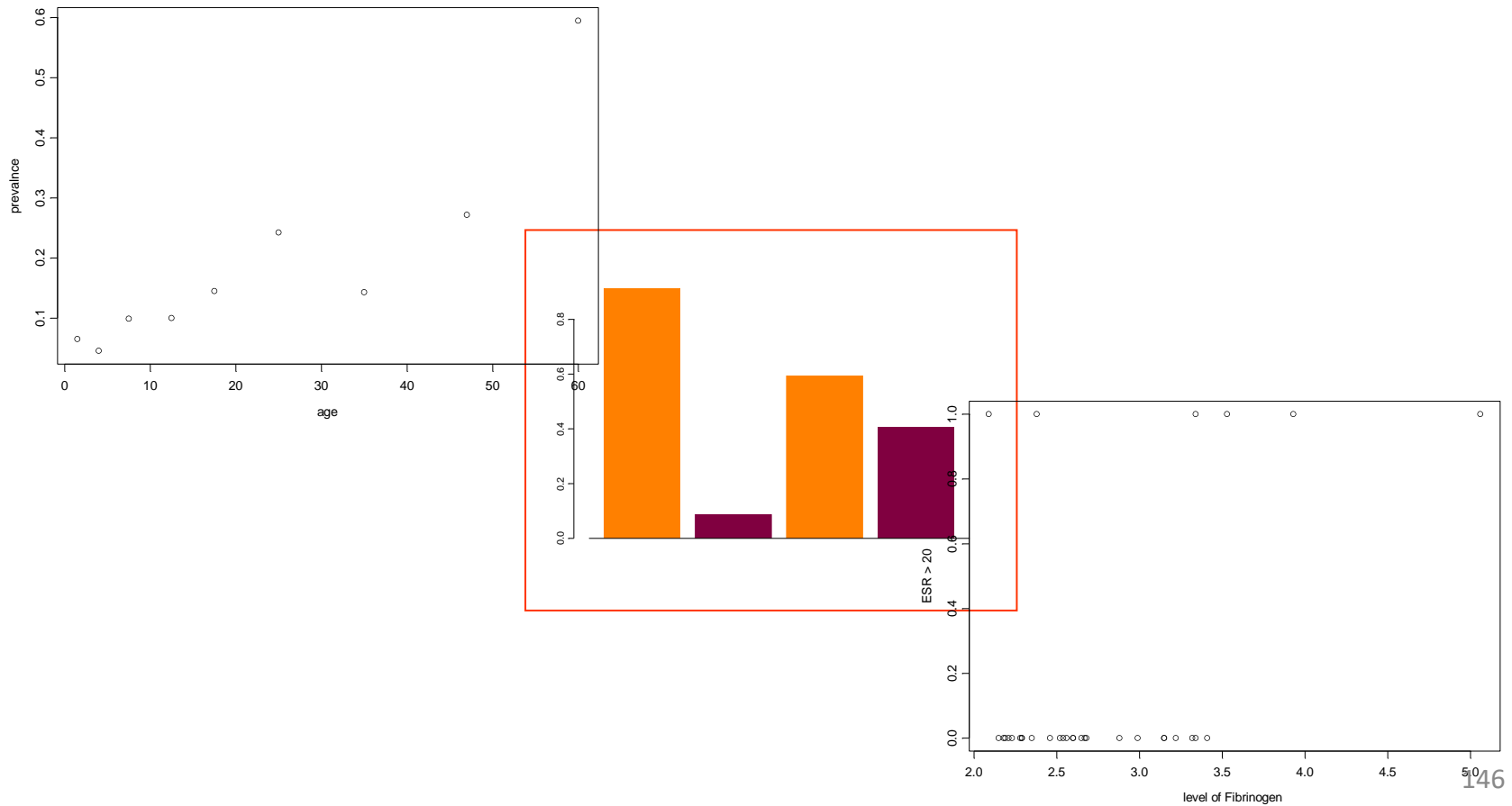
- Binary data may occur in two forms
 - ungrouped in which the variable can take one of two values, say success/failure
 - grouped in which the variable is the number of successes in a given number of trials
- The natural distribution for such data is the *Binomial* (n, p) *distribution*; where in the first case $n = 1$

Exploring Binary Data

If our aim is to model a binary response, we would first like to explore the relationship between that response and potential explanatory variables.

- When the explanatory variables are categorical, a simple approach is to calculate proportions within subgroups of the data.
- When some of the explanatory variables are continuous, plots can be more helpful.

Example tour



Example 1: The Aspirin and Myocardial Infarction Data

- Relationship between aspirin use and heart attacks
- 5-year randomized study
- does regular aspirin intake reduces mortality from cardiovascular disease?

Group	Myocardial Infarction		Total
	Yes	No	
Placebo	189	10845	11034
Asprin	104	10933	11037

Example 1: The Aspirin and Myocardial Infarction Data

The question of primary interest is:

Does regular aspirin intake reduces mortality from cardiovascular disease?

$$Y_i = \begin{cases} 1 & \text{Myocardial Infarction} \quad \textit{Yes} \\ 0 & \text{Myocardial Infarction} \quad \textit{No} \end{cases}$$

The response variable

Example 2: smoked mice

In order to investigate the influence of smoking on lung cancer a group of 55 mice were randomized into two treatment groups.

In the first group (the treated group), each animal was enclosed in a chamber that was filled with the smoke of one cigarette every hour in 12 hours day.

The second group (the control group) were kept in their chambers for 12 hours with out smoke. After one year an autopsy was carried out.

The response is the present and absent of a tumor.

The second variable in the data is the treatment group.

Example 2: smoked mice

The question of primary interest is:

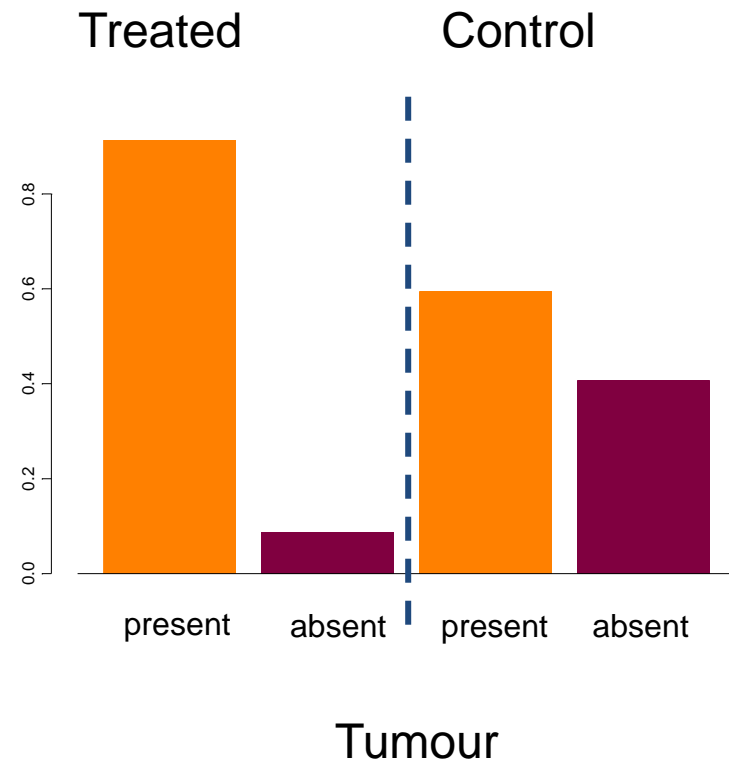
DOSE SMOKE INCREASE THE RISK FOR CANCER ?

$$Y_i = \begin{cases} 1 & \text{tumour present} \\ 0 & \text{tumour absent} \end{cases}$$

↙
The response variable

Example 2: smoked mice

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55



Example 2: smoked mice

	Tumour present	Tumour absent	Total
Treated	21	2	23
Control	19	13	32
Total	20	15	55

We want to model the probability to develop a tumour given the treatment group.

This is an example of grouped data.

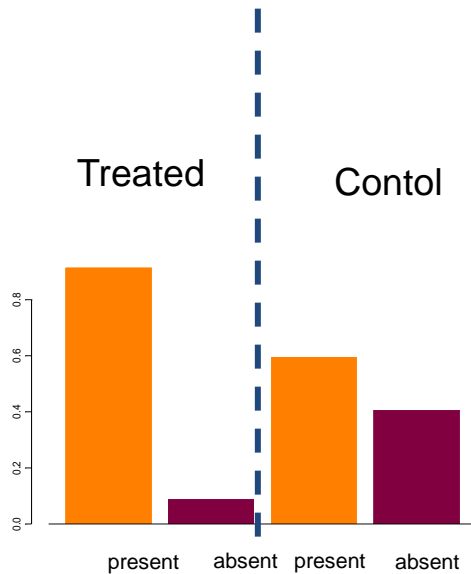
We do not have information about individuals in the sample, but only about the counts in different combinations of the experiment.

Individual data can be extracted from the table.

In terms of statistical modeling, the response is binary (tumor absent/tumor present).

The predictor, the treatment group, is also binary.

Example 2: smoked mice



In the treated group, 21/23 (91%) of the mice develop tumour. In the control group only 19/32 (59%).

The aim of the analysis is to determine if this difference is only due to chance or if the smoke increase the risk for tumour.

Example 3: Serological data

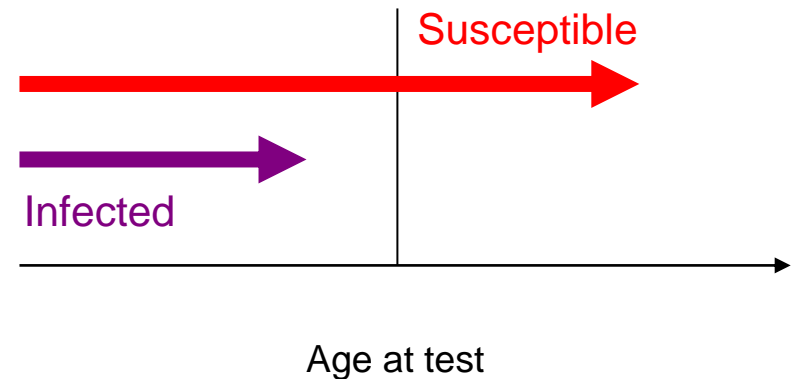
Antibodies produced in response to an infectious disease like malaria remain in the body after the individual has recovered from the disease. A serological test detects the presence or absence of such antibodies. An individual with such antibodies is termed seropositive.

Example 3: Serological data

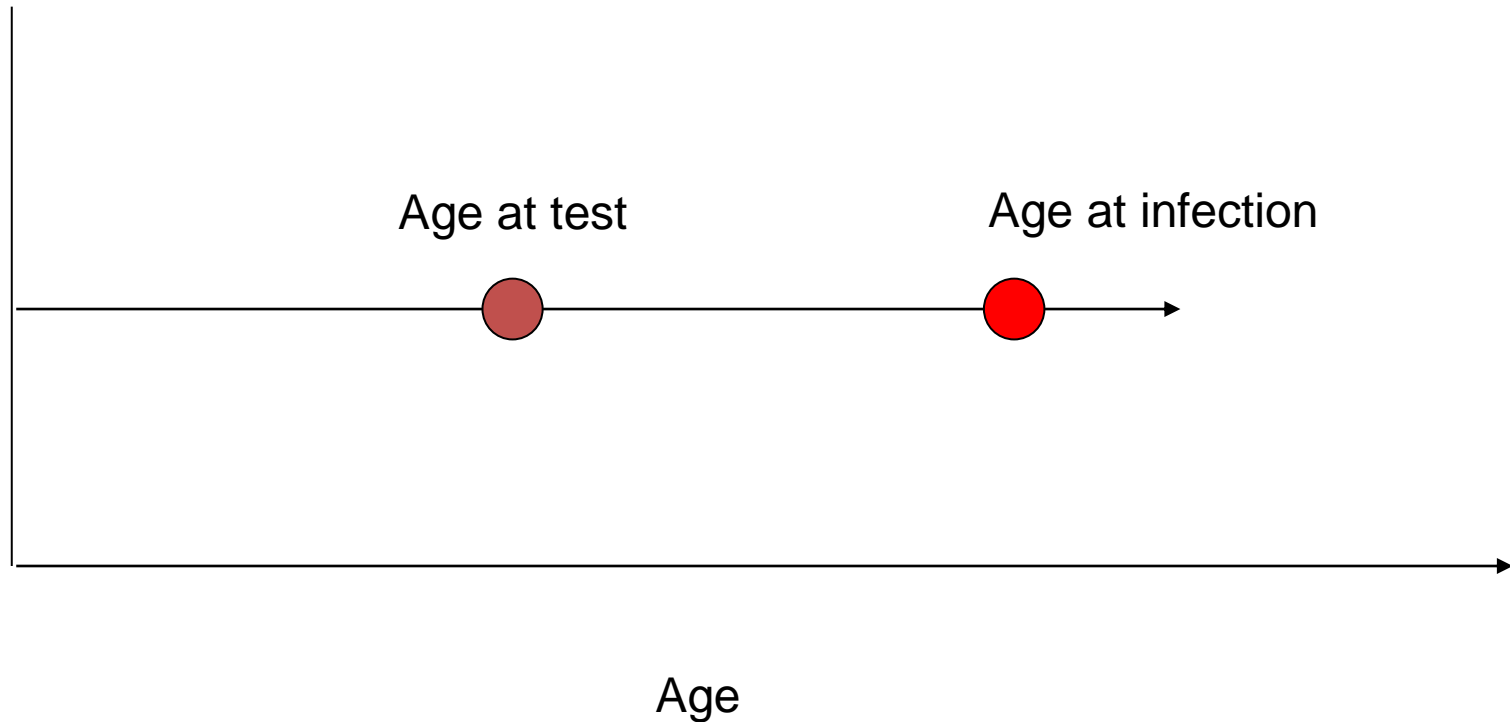
- A sample which taken at a certain time point.
- The information for each individual:
 1. Age at test.
 2. Infected or not.
- Prevalence of sero-positivity In the sample:
$$P(a)$$

This is the probability to become infected before the age at test.

- Sero-prevalence data

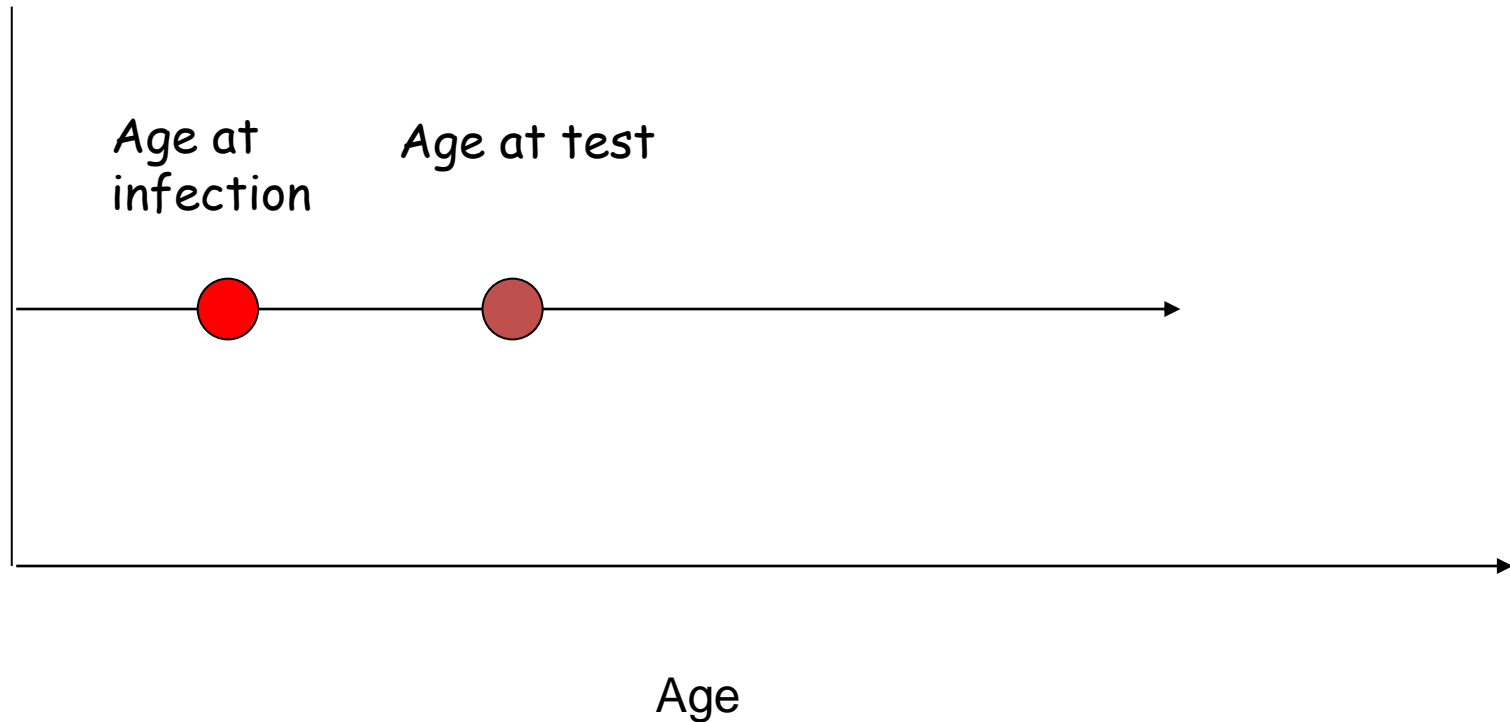


Example 3: serological data



- Sero-Negative: infected after the test.

Example 3: serological data



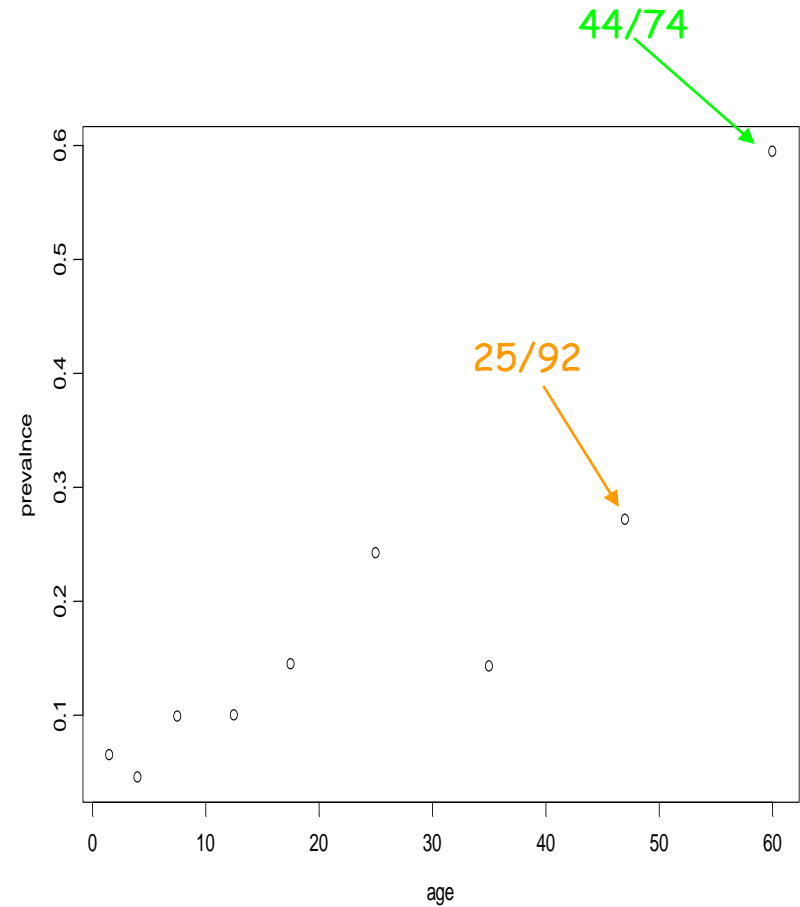
- Sero-Positive: infected before the test.

Example 3: 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.

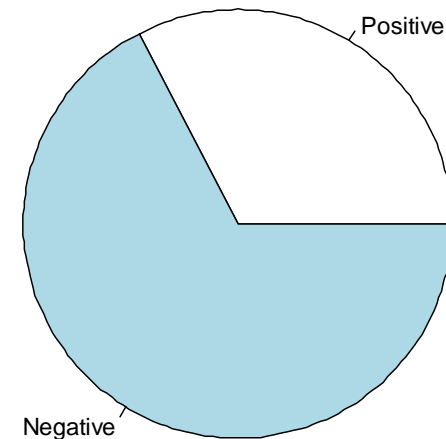
Example 3: serological data

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



Example 4: HIV data

- Consider the HIV data set and the model for HIV (the outcome variable, yes/no or 1/0).
- **Covariates:**
-, age group (also coded 1/0).
- Age group was coded 1 for people younger than 40.7 years
- Age
- Response: HIV status (32.6% are positive).

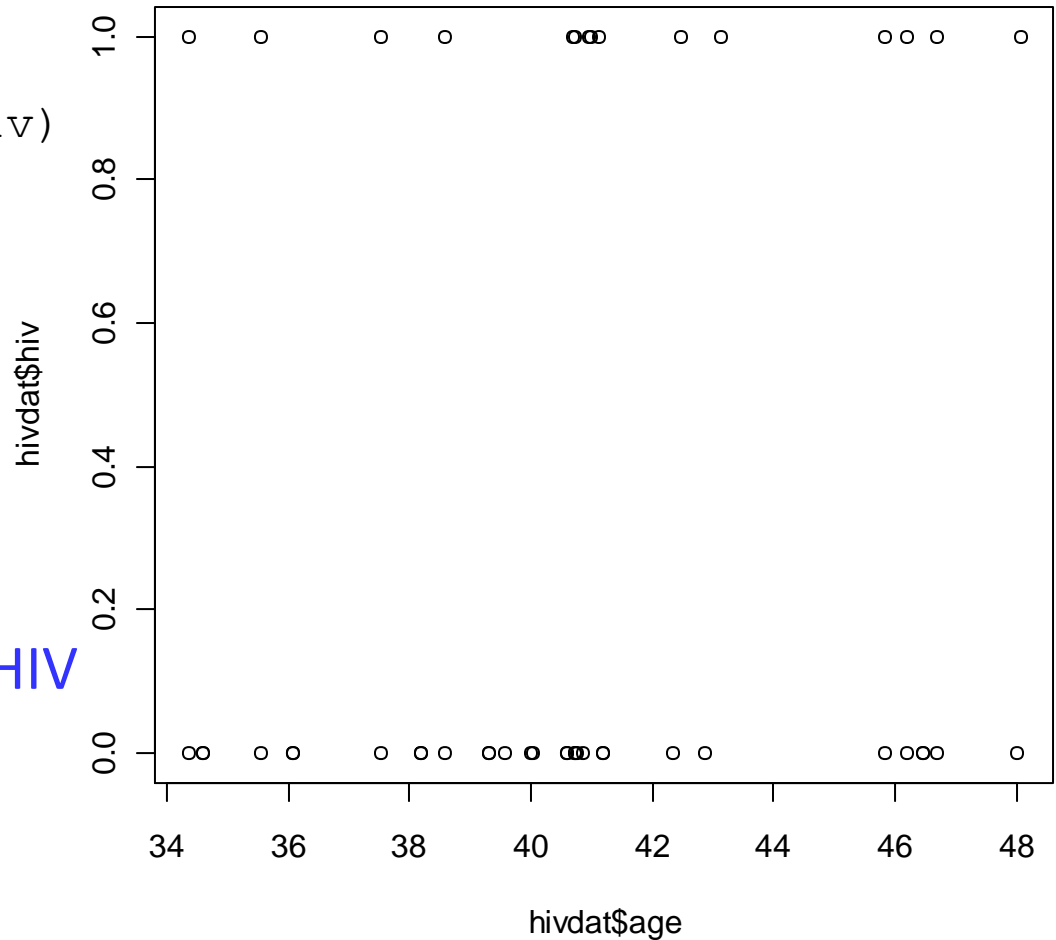


Example 4: HIV data

```
> par(mfrow=c(1,1))  
> plot(hivdat$age,hivdat$hiv)
```

- Continuous predictor.
- Age as predictor variable.
- Response: HIV STATUS

Does the probability to be HIV positive depends on age ?



Example 5: toxicity example (Budworm)

Collett (1991) describes an experiment on the toxicity of the pyrethoid trans - cypermethrin to the tobacco budworm.

Batches of 20 moths of each sex were exposed to varying doses of the pyrethoid for three days and the **number of dead or knocked down** in each batch was recorded:

Sex	Dose (μ g)					
	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Predictor: $\log(\text{dose})$

Example 6: Heart Disease (Dipankar Bandyopadhyay, Ph.D.)

Our outcome is heart disease, and in order to use the ordinal levels of snoring, we need to select scores.

A set (0, 2, 4, 5) seems to capture the relative magnitude of the differences among the categories.

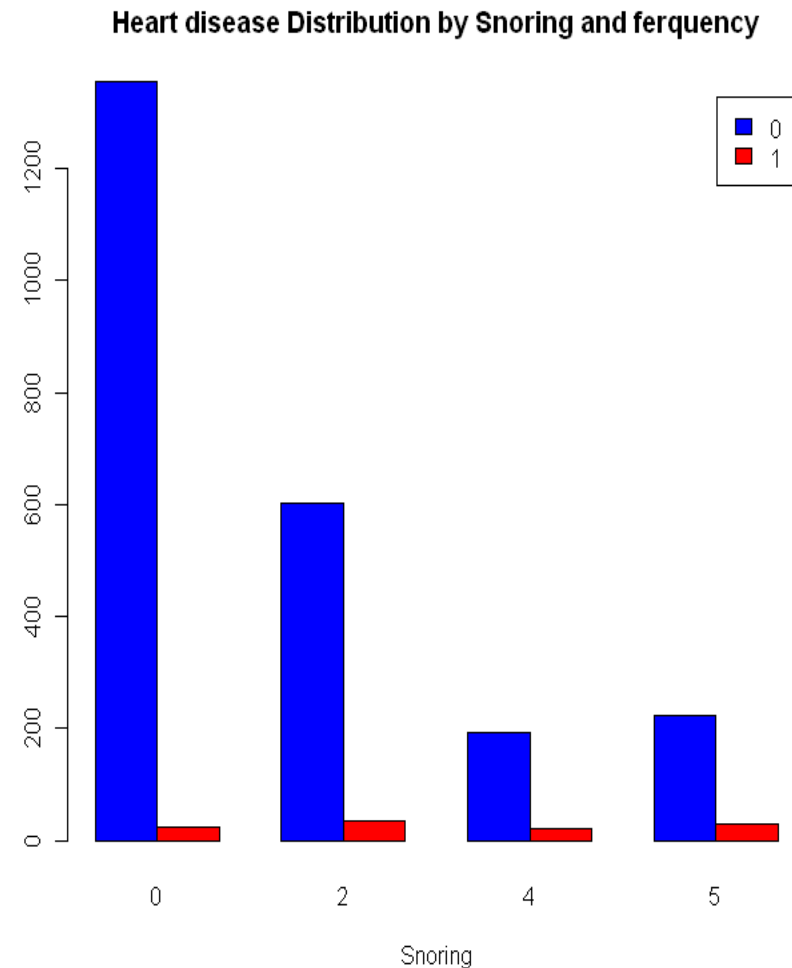
Snoring	Heart Disease		Proportion
	Yes	No	Yes
Never	24	1355	0.017
Occasionally	35	603	0.055
Nearly every night	21	192	0.099
Every Night	30	224	0.118

Example 6: Heart Disease data

```
> par(mfrow=c(1,1))  
> plot(snoring, dhyes)
```

- Categorical predictor.
- Snoring as predictor variable.
- Response: Heart disease (yes|No)

Does the probability to be heart disease depends on snoring ?



Modeling Binary data

Binary data

$$Z_i = \begin{cases} 1 & P \\ 0 & 1 - P \end{cases}$$

The observation is a binary variable with takes the value of 1 with probability P.

$$Z_1, Z_2, Z_3 \dots Z_{n_i}$$

P is the success probability, i.e. $P(Z=1)$.

$$E(Z_i) = P_i$$

The expected value of Z is equal to P.

The sum of binary random variables

$$Z_i = \begin{cases} 1 & P \\ 0 & 1-P \end{cases}$$

$$Z_1, Z_2, Z_3 \dots Z_{n_i}$$

$$E(Z_i) = P_i$$

$$Y_i = \sum_{i=1}^{n_i} Z_i$$

$$Y_i \sim \text{Bin}(n_i, P_i)$$

Often we want to model the sum of the binary variables Y .

If $Z \sim B(1, P)$ then $Y \sim B(n, P)$.

$E(Z) = P$ and $E(Y) = nP$.

Example 1: The Aspirin and Myocardial Infarction Data

The question of primary interest is:

does regular aspirin intake reduces mortality
from cardiovascular disease?

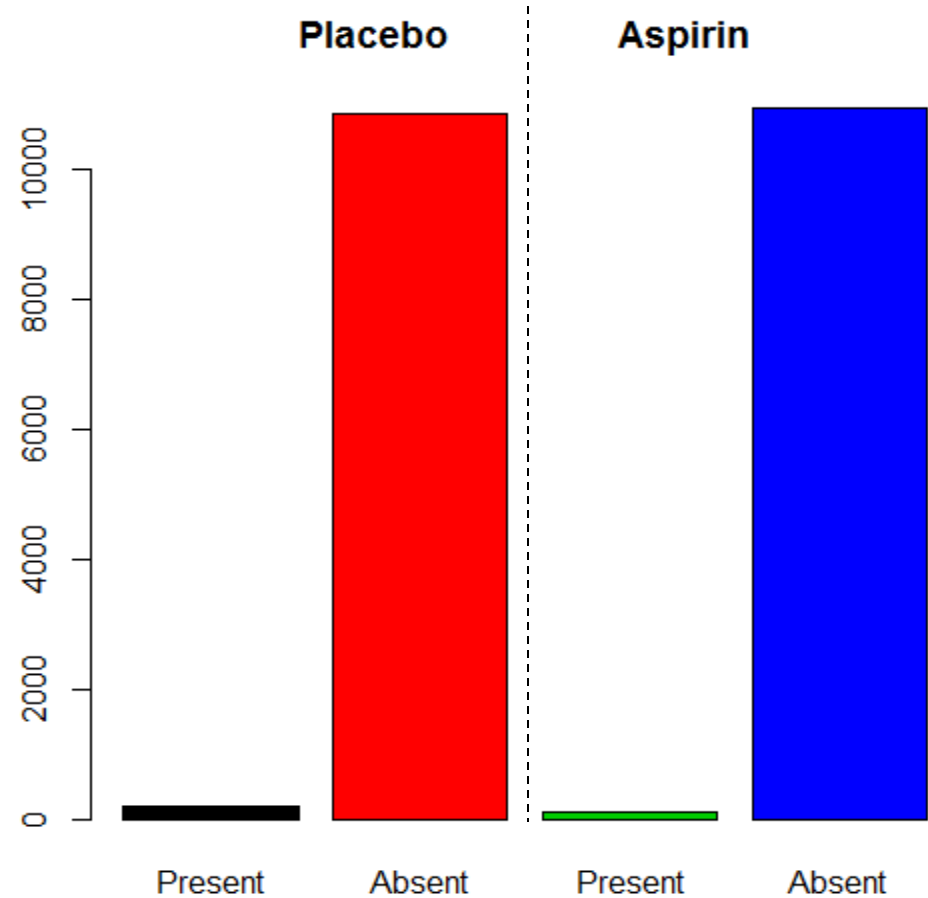
$$Z_i = \begin{cases} 1 & \text{cardiovascular present} \\ 0 & \text{cardiovascular absent} \end{cases}$$

The probability of succsses

- The probability of success $P(Z=1)$. This is the probability to have cardiovascular disease. We want to see if Aspirin intake has an effect on the probability to have Myocardial infarction.

The Data

Myocardial Infarction			
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037



Data in R

```
> resp<-as.factor(c(rep(1,189),rep(0,10845),rep(1,104),rep(0,10933)))
> trt<-as.factor(c(rep(1,189),rep(1,10845),rep(2,104),rep(2,10933)))
```

Myocardial Infarction			
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$$trt_i = \begin{cases} 1 & \text{Aspirin} \\ 2 & \text{Placebo} \end{cases}$$

$$resp_i = \begin{cases} 1 & \text{Yes} \\ 0 & \text{No} \end{cases}$$

Data structure:

```
> cbind(resp,trt)
      resp trt
[1,]      2  1
[2,]      2  1
[3,]      2  1
[4,]      2  1
[5,]      2  1
[6,]      2  1
      .  .
      .  .
[22066,]  1  2
[22067,]  1  2
[22068,]  1  2
[22069,]  1  2
[22070,]  1  2
[22071,]  1  2
```

Sample size 

Data structure in R

- Data are given in table format.
- The variable count is the number of cases in each category.

```
> table(trt, resp)
      resp
trt      0      1
  1 10845   189
  2 10933   104
>
```

Model formulation

We want to model the probability to have Myocardial infarction given the aspirin intake.

The model for P- logit transformation

$$\text{logit}(P) = \beta_0 + \beta_j$$

```
< fit.myoc<-glm(resp~trt,family=binomial(link = "logit"))
```

The estimated model in R

```
> summary(fit.myoc)
```

Call:

```
glm(formula = resp ~ as.factor(trt), family = binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.1859	-0.1859	-0.1376	-0.1376	3.0544

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-4.04971	0.07337	-55.195	< 2e-16 ***
as.factor(trt)2	-0.60544	0.12284	-4.929	8.28e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3114.7 on 22070 degrees of freedom
Residual deviance: 3089.3 on 22069 degrees of freedom
AIC: 3093.3

Number of Fisher Scoring iterations: 7

$$\log it(\hat{P}_i) = \hat{\beta}_0 + \hat{\beta} \times Aspirin$$

How do we interpret the parameters from the output above ?

The parameter estimate for the effect of the placebo group is -4.04971. The parameter estimate for the effect of the Aspirin intake is -0.60544.

The odds ratio, θ , is equal to 0.5458342. If $\theta < 1$ than the odds for a Myocardial infarction in the Aspirin intake group is smaller than the odds for Myocardial infarction in the placebo group. This means that the aspirin reduces the risk of myocardial infarction.

Example 2: smoked mice

The question of primary interest is:

DOSE THE SMOKE INCREASE THE RISK FOR CANCER ?

$$Z_i = \begin{cases} 1 & \text{tumour present} \\ 0 & \text{tumour absent} \end{cases}$$

The response variable

The probability of succsses

- The probability of success $P(Z=1)$. This is the probability to have tumour. We want to see if treatment (smoke) has an effect on the probability to develop a tumour.

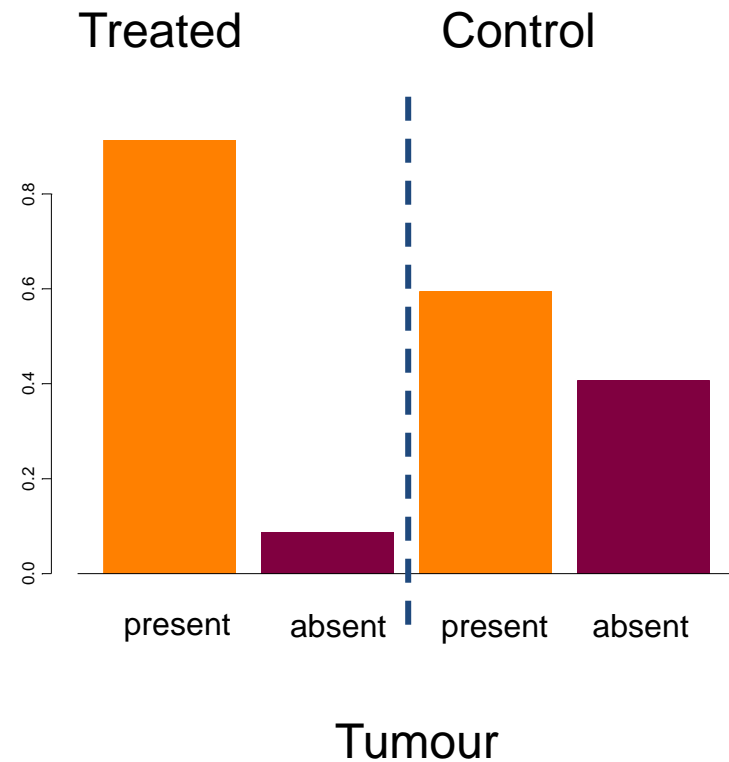
Data structure in R

- Data are given in table format.
- The variable count is the number of cases in each category.

```
> table(trti, resp)
      resp
trti    0    1
    1 21    2
    2 19   13
```

The Data

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55



Model formulation

The individual data

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55

$$Z_i = \begin{cases} 1 & \text{tumour present} \\ 0 & \text{tumour absent} \end{cases}$$

Number of subjects with
tumour

$$Y_i = \sum Z_i$$

Distribution of Y

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

The model for P- logit transformation

$$\log it(P) = \beta_0 + \beta_j$$

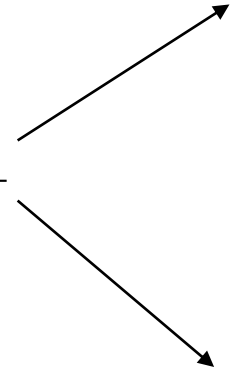
The probability

$$P = \frac{e^{\beta_0 + \beta_j}}{1 + e^{\beta_0 + \beta_j}}$$

The parameter β_j is the treatment effect.

Note that we have two treatment groups and it is dummy coding for treatment effect, the $\beta_{\text{control}} = \beta_0$

The probability

$$P = \frac{e^{\beta_0 + \beta_j}}{1 + e^{\beta_0 + \beta_j}}$$


The diagram shows a general formula on the left with two arrows pointing to specific formulas on the right. The top arrow points from the β_j term in the general formula to the $\beta_{treatment}$ term in the top specific formula. The bottom arrow points from the β_j term in the general formula to the β_0 term in the bottom specific formula.

$$P = \frac{e^{\beta_0 + \beta_{treatment}}}{1 + e^{\beta_0 + \beta_{treatment}}}$$
$$P = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

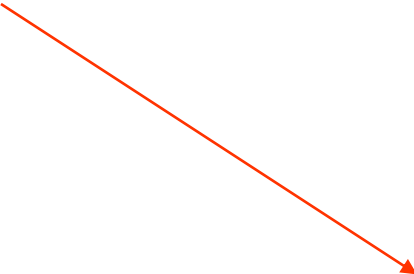
The probability to have tumor for the treatment group.

The probability to have tumor for the control group.

Logistic regression in R

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55

```
fit.mice<-glm(resp~trti,family=binomial(link = "logit"))
```


$$\log it(P_i) = \beta_0 + \beta \times treatment$$



model status= treat

The estimated model in R

```
> summary(fit.mice)
```

Call:

```
glm(formula = resp ~ trti, family = binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0211	-1.0211	-0.4265	1.3422	2.2101

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
$\hat{\beta}_0$ → (Intercept)	-2.3514	0.7400	-3.177	0.00149	**
$\hat{\beta}$ → trti2	1.9719	0.8229	2.396	0.01656	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

$$\log it(\hat{P}_i) = \hat{\beta}_0 + \hat{\beta} \times treatment$$

How do we interpret the parameters ?

Coefficients:

```
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3514    0.7400  -3.177 0.00149 **
trti2        1.9719    0.8229   2.396 0.01656 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The parameter estimate for the effect of the control group is -2.3514. The parameter estimate for the effect of the treatment group (the smoked group) is equal to **1.9719**.

- ➔ The probability of tumour in control and treatment groups are 0.086955 and 0.406247, respectively.
- ➔ The odds of tumour in control and treatment groups are 0.095236 and 0.684203, respectively.

How do we interpret the parameters ?


- ➔ The probability of tumour in control and treatment groups are 0.086955 and 0.406247, respectively.
- ➔ The odds of tumour in control and treatment groups are 0.095236 and 0.684203, respectively.
- ➔ The odds ratio θ is **7.184314**.
- ➔ If $\theta > 1$ than the odds for a tumour in the treatment group is larger than the odds for a tumour in the control group. This means that the probability for tumour in the treatment group is **LARGER** than the probability for tumour in the control group.

The odds ratio: estimation

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.3514	0.7400	-3.177	0.00149	**
trti2	1.9719	0.8229	2.396	0.01656	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



For a factor predictor variable,

$$\theta = \exp(\beta).$$

In our example: $\theta = \exp(1.9719) = 7.184314$.

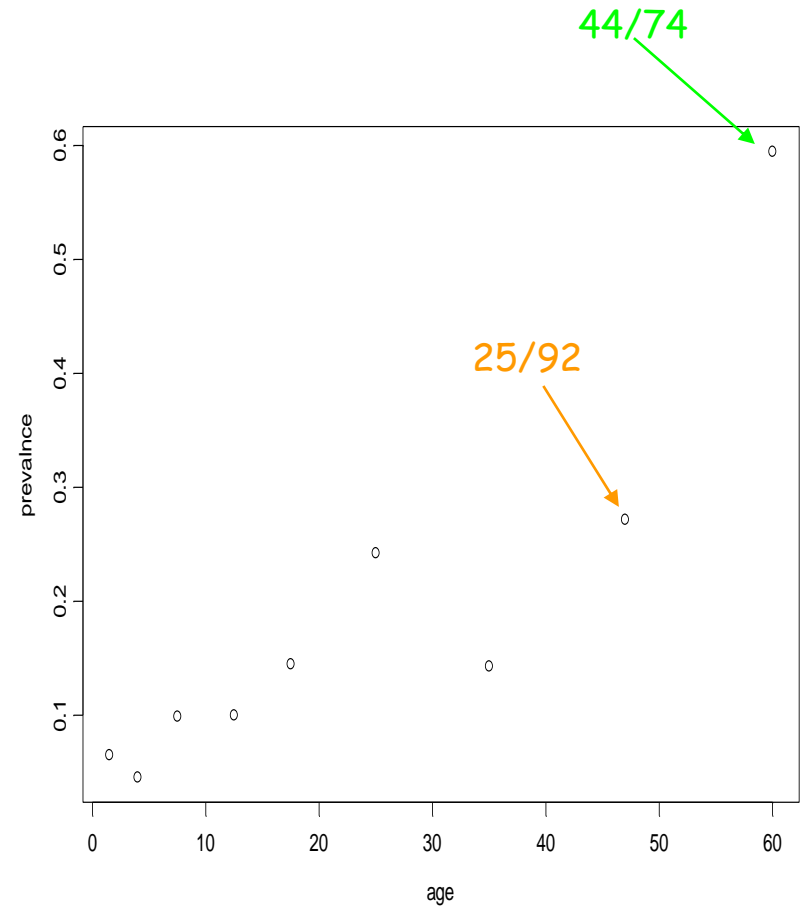
The odds ratio: point estimator

The inverse of the odds ratio, θ , is equal to 0.139.

➔ The odds for a tumour in the control group is smaller than the odds for a tumour in the treatment group. This means that the probability for tumour in the control group is SMALLER than the probability for tumour in the treatment group.

Example 3: 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



Example 3: Data structure in R

- This is an example in which the predictor (age) is continuous.
- We want to model the probability of infection as a function of age.

```
cbind(agei, posi, negi)
```

```
agei posi negi
```

```
[1,] 1.5  8 115
```

```
[2,] 4.0  6 126
```

```
[3,] 7.5 18 164
```

```
[4,] 12.5 14 126
```

```
[5,] 17.5 20 118
```

```
[6,] 25.0 39 122
```

```
[7,] 35.0 19 114
```

```
[8,] 47.0 25  67
```

```
[9,] 60.0 44  30
```

Example 3: 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

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

$$Y_i = \sum Z_i$$

Number of sero-positive at each age group

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

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

$$\log it(P_i) = \beta_0 + \beta \times \text{age}$$


The probability of infection

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

$$P = \frac{e^{\beta_0 + \beta \text{ age}}}{1 + e^{\beta_0 + \beta \text{ age}}}$$

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

Logistic regression in R

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


pos/N

```
fit.malaria<-glm(cbind(posi,negi)~agei,  
family=binomial(link="logit"))
```

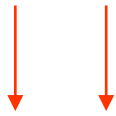
$$\text{logit}(P_i) = \beta_0 + \beta \times \text{age}$$

model pos/N=age



Parameters estimate

$$\log it(\hat{P}_i) = a + b \times age$$


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

```
> summary(fit.malaria)
```

Call:

```
glm(formula = cbind(posi, negi) ~ agei, family =  
binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.78685	-1.31863	-0.05053	0.66752	2.38275

Coefficients:

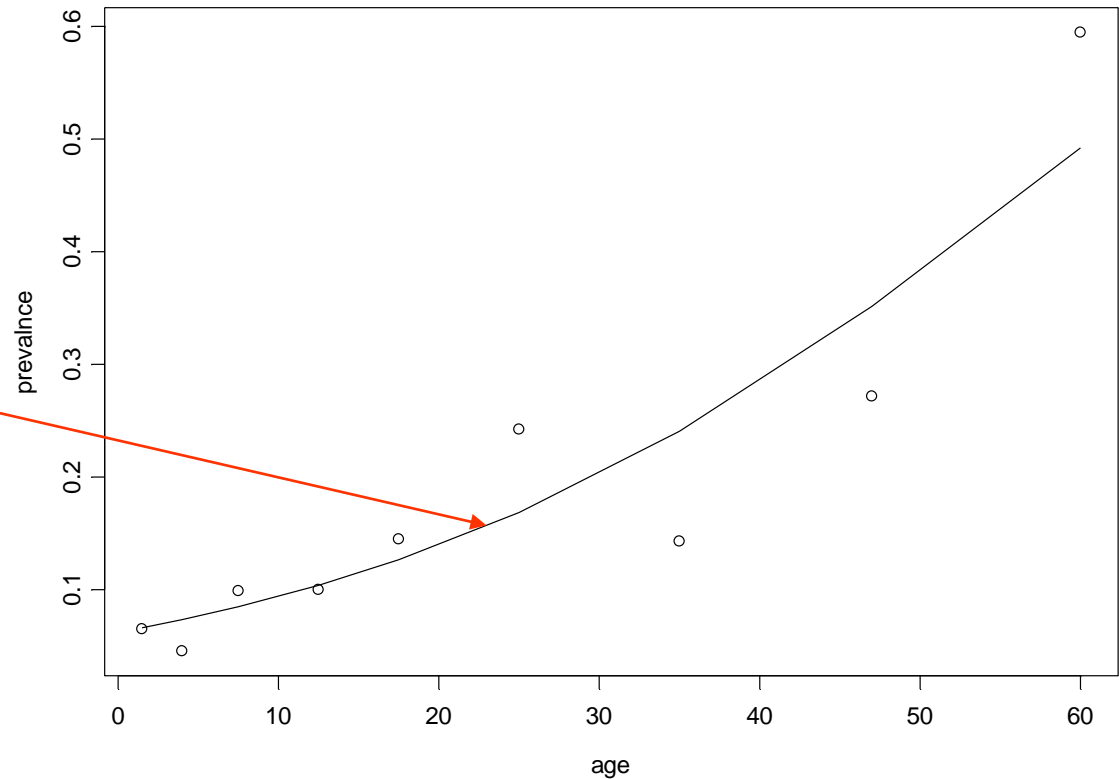
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.714074	0.151740	-17.886	<2e-16 ***
agei	0.044672	0.004511	9.904	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Data and predicted values

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

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



The odds ratio: point estimator

```
> exp(0.044672)
[1] 1.045685
```

How to calculate the odds ratio ?

The odds ratio is given by

$$\theta = \exp(\beta).$$

In our example $\theta = \exp(0.0447) = 1.046$.

Implies **per unit increase** of age the odds to be infected by malaria increase by 4.6%

Example 4: HIV data

- Dependency of the probability to be HIV positive on different covariates.

$$Y_i = \begin{cases} 1 & \text{HIV +} \\ 0 & \text{HIV -} \end{cases}$$

$$Y_i \sim B(1, \pi)$$

$$X_i = \text{age}_i$$

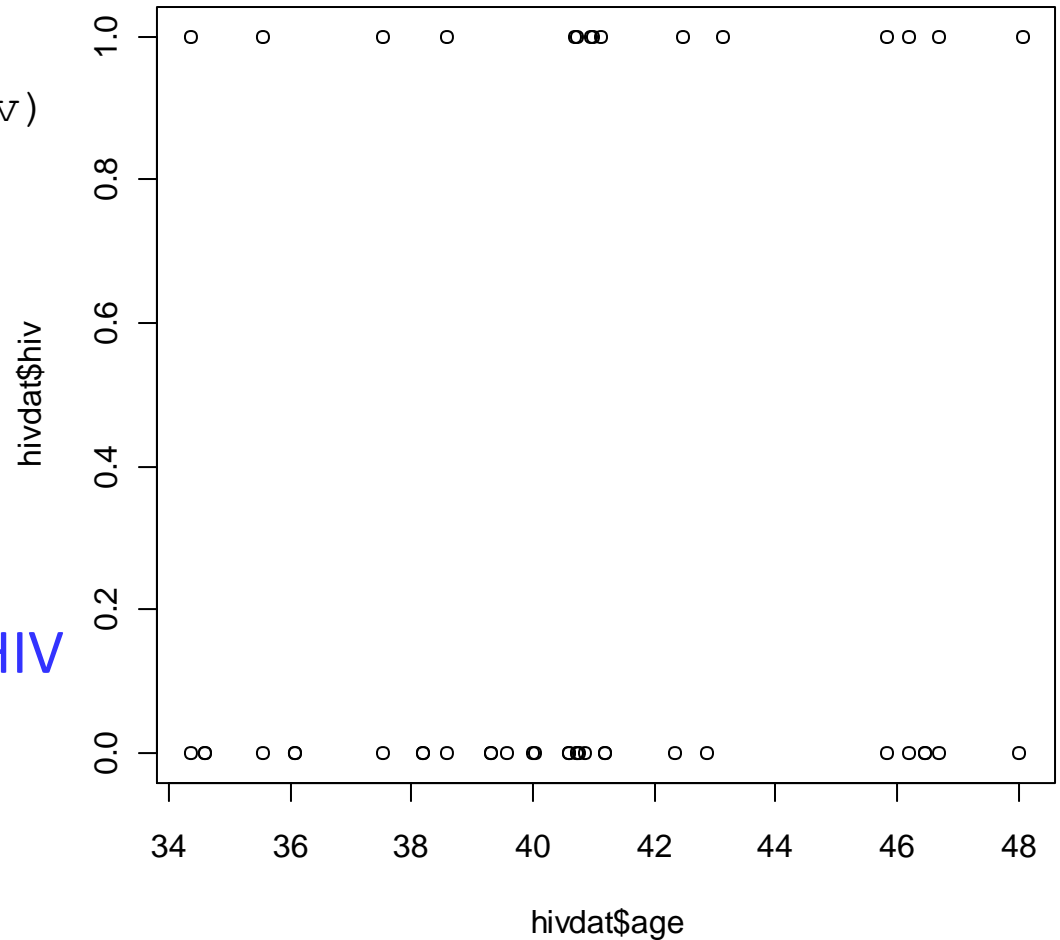
Does the probability to be HIV positive depends on age

Example 4: HIV data

```
> par(mfrow=c(1,1))  
> plot(hivdat$age,hivdat$hiv)
```

- Continuous predictor.
- Age as predictor variable.
- Response: HIV STATUS

Does the probability to be HIV positive depends on age ?



Model formulation

$$Y_i \sim B(1, \pi)$$

$$E(Y_i) = \pi$$

$$\pi = f(X_i) = f(\text{age}_i)$$

The GLM

$$Y_i \sim B(1, \pi)$$

$$E(Y_i) = \pi$$

$$\pi = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

$$g(E(Y_i)) = g(\pi_i) = \beta_0 + \beta_1 X_i$$

The GLM in R

```
> hiv.fit1 <- glm(hiv ~ age, family=binomial(link = "logit"),  
                  data= hivdat)  
  
> summary(hiv.fit1)
```

Call:

```
glm(formula = hiv ~ age, family = binomial(link = "logit"), data =  
hivdat)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.79597	3.43622	-1.105	0.269
age	0.07492	0.08314	0.901	0.367

(Dispersion parameter for binomial family taken to be 1)

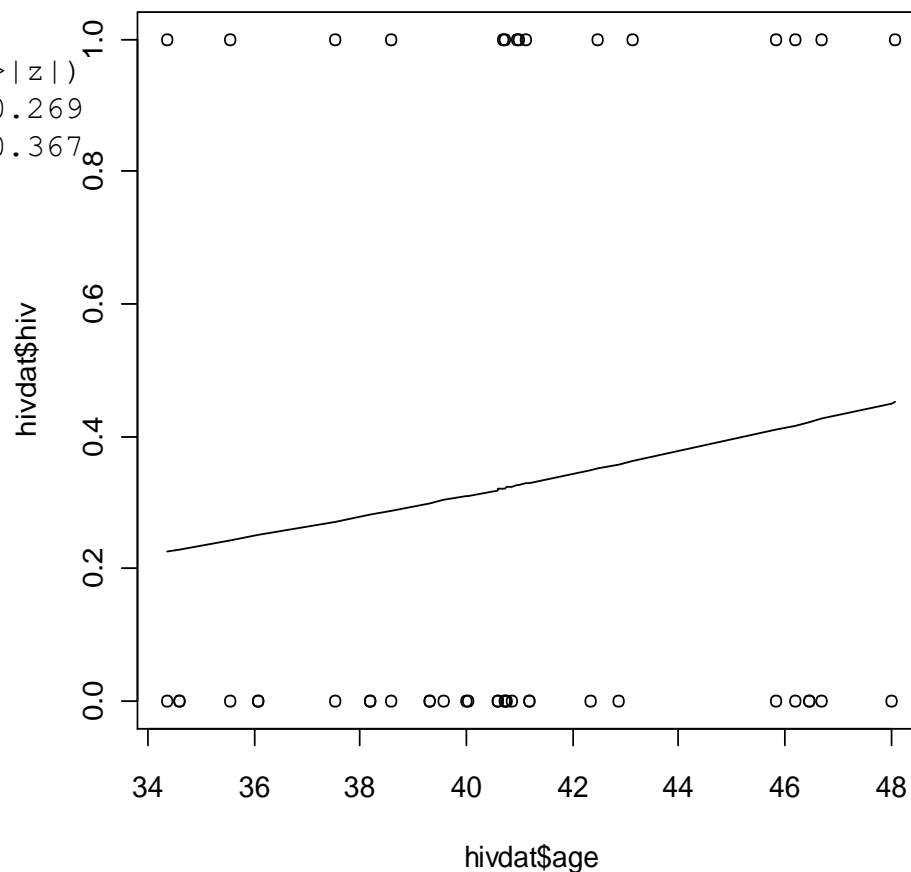
The data and fitted model plot

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.79597	3.43622	-1.105	0.269
age	0.07492	0.08314	0.901	0.367

$$g(\pi_i) = -3.79 + 0.0749 \times age_i$$

$$\pi_i = \frac{e^{-3.79+0.0749 \times age_i}}{1 + e^{-3.79+0.0749 \times age_i}}$$



The odds ratio: point estimator

```
> exp(0.07492)
[1] 1.077798
```

How to calculate the odds ratio ?

$\theta = \exp(\beta)$.

In our example $\theta = \exp(0.07492) = 1.07798$.

As age increases by one unit the odds to be HIV positive increase by 7.8%

Example 5: toxicity example (Budworm)

Collett (1991) describes an experiment on the toxicity of the pyrethoid trans - cypermethrin to the tobacco budworm.

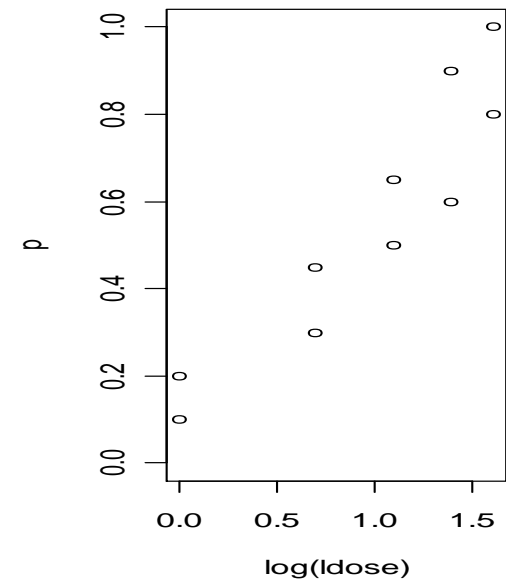
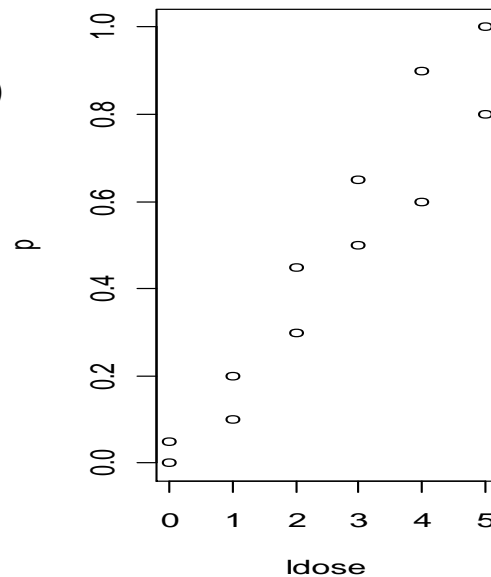
Batches of 20 moths of each sex were exposed to varying doses of the pyrethoid for three days and the **number knocked out** in each batch was recorded:

Sex	Dose (μ g)					
	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

Predictor: $\log(\text{dose})$

Data and Plot in R

```
> ldose <- rep(0:5, 2)
> numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
> sex <- factor(rep(c("M", "F"), c(6, 6)))
> SF <- cbind(numdead, numalive=20-numdead)
> p<-numdead/20
> par(mfrow=c(1,2))
> plot(p ~ ldose)
> plot(p ~ log(ldose))
```



Model formulation

the expected values of
The response variable

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_j$$

$$P(Y_{ij} = 1) = P(\text{knocked out})$$

The systematic part

$$\pi_j = f(\text{dose} \quad \text{gender})$$

$$\eta = \text{dose} + \text{gender} + \text{dose} * \text{gender}$$

$$g(E(Y_{ij})) = g(\pi_j) = \eta$$

Model formulation

Distribution of the
response

$$Y_{ij} \sim \text{Bin}(n(d_j), \pi_j)$$

$$P(Y_{ij} = 1) = P(\text{ko}) = \pi_j$$

The linear predictor

$$\eta = \beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}$$

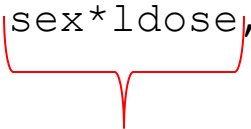
$$E(Y_{ij}) = \pi_j = \frac{e^{\beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}}}{1 + e^{\beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}}} = \frac{e^\eta}{1 + e^\eta}$$

$$g(E(Y_{ij})) = g(\pi_j) = \eta$$

Model with Binomial family and logit link function

Fitting the model with the glm() function:

```
> budworm.lg <- glm(SF ~ sex*ldose, family=binomial)
```



$$\eta = \beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}$$

Alternative code

```
> budworm.lg <- glm(SF ~ sex+ldose+sex:ldose, family=binomial)
```



R output

Call:

```
glm(formula = SF ~ sex * ldose, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08	***
sexM	0.1750	0.7783	0.225	0.822	
ldose	0.9060	0.1671	5.422	5.89e-08	***
sexM:ldose	0.3529	0.2700	1.307	0.191	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

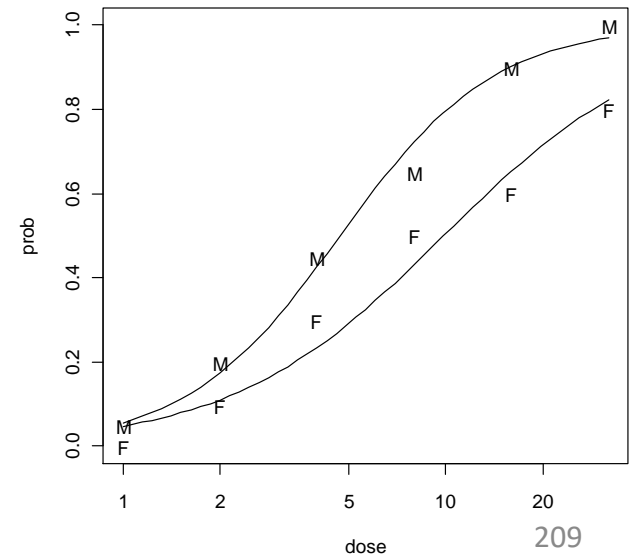
Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

Exp(0.906)=2.47 implies unit increase of dose increase the number of
knocked out 2.47 times

Plot of observed and predictive probability of death for male and female budworms

```
> plot(c(1,32), c(0,1), type = "n", xlab = "dose",  
+      ylab = "prob", log = "x")  
> text(2^ldose, numdead/20, as.character(sex))  
> ld <- seq(0, 5, 0.1)  
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,  
+      sex=factor(rep("M", length(ld)), levels=levels(sex)))),  
+      type = "response")  
> lines(2^ld, predict(budworm.lg, data.frame(ldose=ld,  
+      sex=factor(rep("F", length(ld)), levels=levels(sex)))),  
+      type = "response"))
```



Example 6:Heart Disease(Dipankar Bandyopadhyay, Ph.D.)

	Heart Disease		Proportion
	Yes	No	Yes
Snoring			
Never	24	1355	0.017
Occasionally	35	603	0.055
Nearly every night	21	192	0.099
Every Night	30	224	0.118

Our outcome is heart disease, and in order to use the ordinal levels of snoring, we need to select scores.

A set (0, 2 , 4, 5) seems to capture the relative magnitude of the differences among the categories.

Data structure in R

- Data are given in table format.
- The variable count is the number of cases in each category.

```
> table(snoring, dhyes)
      dhyes
snoring  0    1
      0 1355  24
      2  603  35
      4  192  21
      5  224  30
```

```
> fit.snoring<-  
  glm(dhyes~as.factor(snoring), family=binomial(link="logit"))
```

The estimated model in R

```
> summary(fit.snoring)
```

Call:

```
glm(formula = dhyes ~ as.factor(snoring), family = binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.5014	-0.3359	-0.1874	-0.1874	2.8464

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.0335	0.2059	-19.590	< 2e-16	***
as.factor(snoring)2	1.1869	0.2695	4.404	1.06e-05	***
as.factor(snoring)4	1.8205	0.3086	5.900	3.64e-09	***
as.factor(snoring)5	2.0231	0.2832	7.144	9.06e-13	***

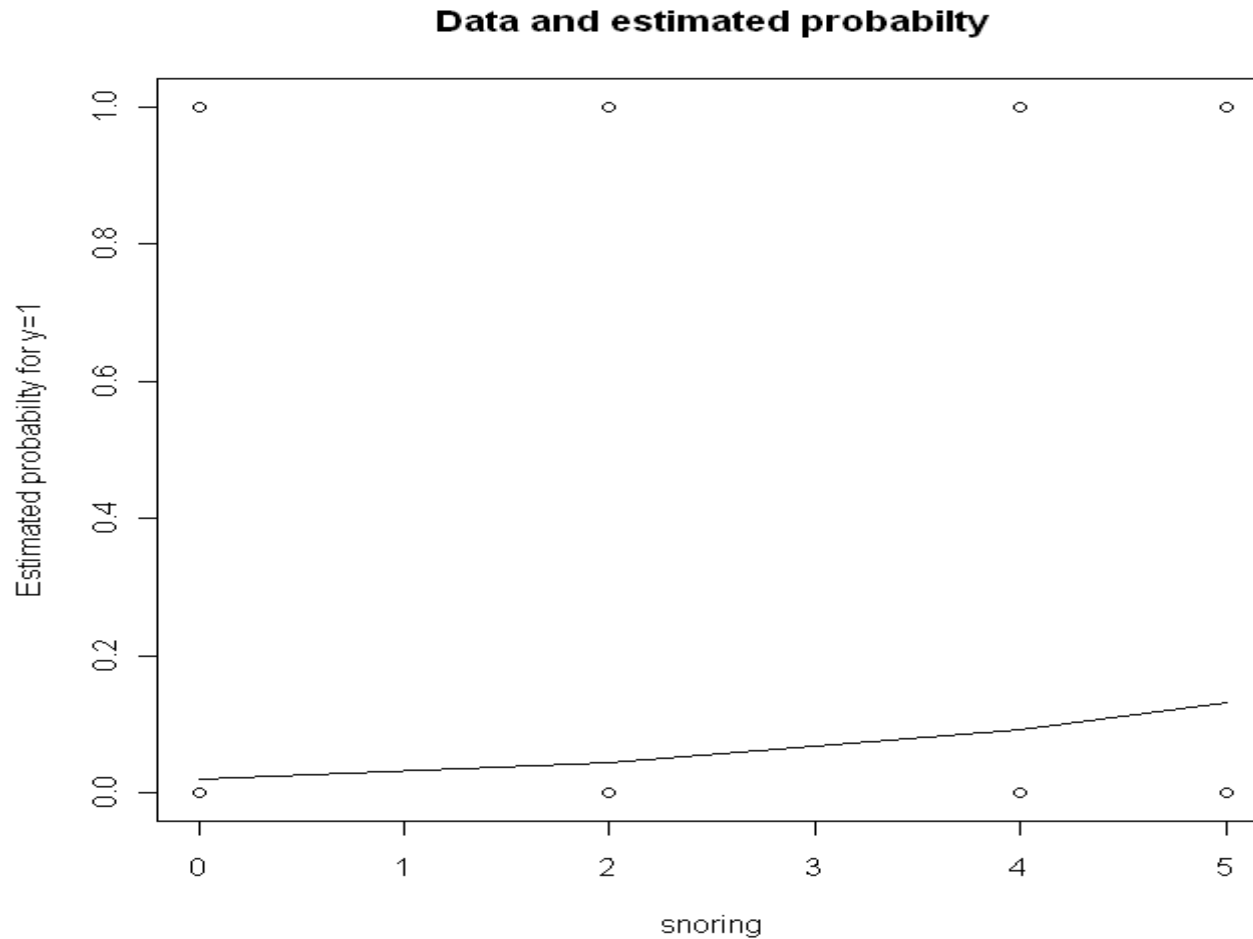
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 900.83 on 2483 degrees of freedom
Residual deviance: 834.92 on 2480 degrees of freedom
AIC: 842.92

Number of Fisher Scoring iterations: 6

Data and predicted probability



The estimated model in R

```
➤fit.snoringCont<-glm(dhyes~snoring,family=binomial(link="logit"))
➤summary(fit.snoringCont)
```

➤Call:

```
glm(formula = dhyes ~ snoring, family = binomial(link = "logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.5331	-0.3010	-0.2036	-0.2036	2.7882

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.86625	0.16621	-23.261	< 2e-16 ***
snoring	0.39734	0.05001	7.945	1.94e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 900.83 on 2483 degrees of freedom
Residual deviance: 837.73 on 2482 degrees of freedom
AIC: 841.73

Number of Fisher Scoring iterations: 6

$$\log it(\hat{P}_i) = -3.87 + 0.397 \times Snoring$$

Chapter 7:

Estimation and confidence Interval

Donson: chapter 4.

Lindsey: chapter 2.

McCullagh & Nelder: chapter 4.

Estimation of model parameters

A single algorithm can be used to estimate the parameters of an exponential family using maximum likelihood.

The log-likelihood for the samples y_1, y_2, \dots, y_n is

$$l = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + c(y_i, \theta_i)$$

The maximum likelihood estimates are obtained by solving the score equation

$$U(\beta_j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0$$

For parameters β_j .

The score function

We assume that

$$\phi_i = \frac{\phi}{a_i}$$

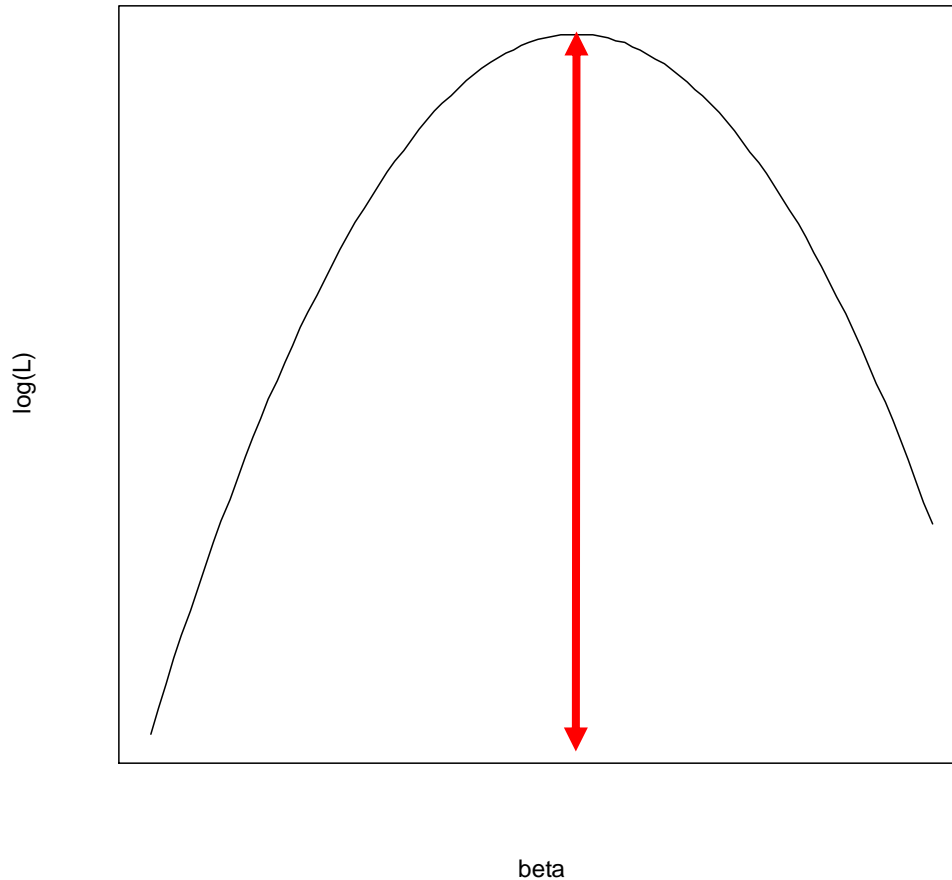
Where ϕ is a single dispersion parameter and a_i are known **prior weights**; for example binomial proportions with known index n_i have $\phi=1$ and $a_i=n_i$

The estimating equations are then

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{a_i (y_i - \mu_i)}{V(\mu_i)} \times \frac{x_{ij}}{g'(\mu_i)} = 0$$

Which does not depend on ϕ (which may be unknown)

The score function



At the maximum:

$$U(\beta_j) = \frac{\partial l}{\partial \beta_j} = 0$$

Example : toxicity example (Budworm)

Predictor: log(dose)

Sex	Dose (μ g)					
	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

See example 5 in Chapter 6

$$\eta = \beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}$$

```
> budworm.lg <- glm(SF ~ sex+ldose+sex:ldose, family=binomial)
```

The R output

Parameter estimates:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08	***
sexM	0.1750	0.7783	0.225	0.822	
ldose	0.9060	0.1671	5.422	5.89e-08	***
sexM:ldose	0.3529	0.2700	1.307	0.191	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Partial t-tests test the significance of each coefficient in the presence of others. Thus, only intercept and ldose were found to be significant at 5% level of significance.

Fisher scoring

The score function

$$U = \frac{d\ell}{d\beta}$$

The first derivative of the score

$$U' = \frac{dU}{d\beta}$$

A general method of solving score equations is the iterative algorithm **Fisher's Method of scoring** (derived from a Taylor's expansion of $U(\beta)$)

$$\beta^{(r+1)} = \beta^{(r)} + \frac{U^{(r)}(\beta)}{U'^{(r)}(\beta)}$$

Fisher scoring

The score function

$$U = \frac{d\ell}{d\beta}$$

The expected value of the score

$$E(U) = 0$$

The variance of the score

$$\text{Var}(U) = E(U^2) - [E(U)]^2 = E(U^2)$$

$$E(U^2) = -E\left(\frac{\partial U}{\partial \beta}\right) = I(\beta)$$

Update in the r th iteration

$$\beta^{(r+1)} = \beta^{(r)} + \frac{U^{(r)}(\beta)}{I^{(r)}(\beta)}$$

Fisher scoring

With some mathematics it can be shown that

$$\beta^{(r+1)} = \left(X^T W^{(r)} X \right)^{-1} X^T W^{(r)} z^{(r)}$$

That is the score equations for a weighted least squares regression of $\mathbf{z}^{(r)}$ on \mathbf{X} with weights $W^{(r)} = \text{diag}(w_i)$, where

$$z_i^{(r)} = \eta_i^{(r)} + \left(y_i - \mu_i^{(r)} \right) g' \left(\mu_i^{(r)} \right)$$

$$w_i^{(r)} = \frac{a_i}{V \left(\mu_i^{(r)} \right) \left(g' \left(\mu_i^{(t)} \right) \right)^2}$$

Standard errors

The estimates $\hat{\beta}$ have the usual properties of maximum likelihood estimators. In particular, $\hat{\beta}$ is asymptotically

$$N(\beta, i^{-1})$$

Where

$$i(\beta) = \phi^{-1} X^T W X$$

Standard errors for β_j may therefore be calculated as the square roots of the diagonal elements of

$$\text{cov}(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}$$

In which $\phi(X^T \hat{W} X)^{-1}$ is a by-product of the final **IWLS** iteration. If ϕ is unknown, an estimate is required.

Standard error

There are practical difficulties in estimating the dispersion ϕ by Maximum likelihood.

Therefore it is usually estimated by **method of moments**. If β was known an unbiased estimate of $\phi = \{a_i \text{ var}(Y)\} / v(\mu_i)$

Would be

$$\frac{1}{n} \sum_{i=1}^n \frac{a_i (y_i - \mu_i)^2}{V(\mu_i)}$$

Allowing for the fact that β must be estimated we obtain

$$\frac{1}{n-p} \sum_{i=1}^n \frac{a_i (y_i - \mu_i)^2}{V(\mu_i)}$$

R output for the toxicity example

Call:

```
glm(formula = SF ~ sex * ldose, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08	***
sexM	0.1750	0.7783	0.225	0.822	
ldose	0.9060	0.1671	5.422	5.89e-08	***
sexM:ldose	0.3529	0.2700	1.307	0.191	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756 on 11 degrees of freedom
Residual deviance: 4.9937 on 8 degrees of freedom
AIC: 43.104

Number of Fisher Scoring iterations: 4

Example: the beetle data

Dose	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Beetles	59	60	62	56	63	59	62	60
Killed	6	13	18	28	52	53	61	60

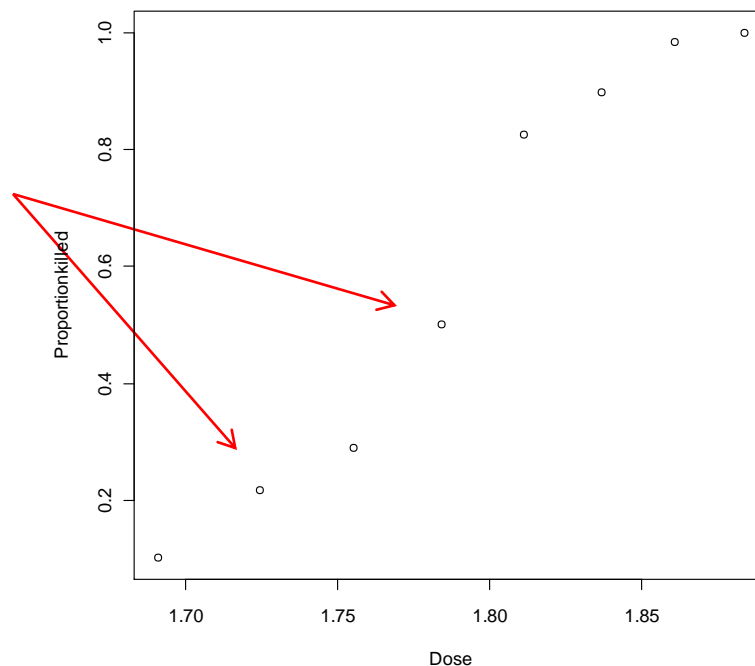
Proportion of the killed beetles

$$Y_{ij} = \begin{cases} 1 & \text{alive} \\ 0 & \text{killed} \end{cases}$$

$$\frac{\sum Y_{ij}}{n_j}$$

$$Y_{ij} \sim B(1, \pi_{ij})$$

$$E(Y_{ij}) = P(Y_{ij} = 1) = \pi_{ij}$$



The Link function and linear predictor

The expected values of
the response variable

$$E(Y_{ij}) = \pi_j$$

The systematic part

$$\pi_j = f(\beta_0 + \beta_1 d_j) = f(\eta)$$

$$\pi_j = \frac{e^{\beta_0 + \beta_1 d_j}}{1 + e^{\beta_0 + \beta_1 d_j}}$$

The logistic
function to
describe the
mean, $E(Y_{ij})$, as
a function of the
linear predictor

$$g(E(Y_{ij})) = g(\pi_j) = \eta$$

Values between 0
and 1

The model in R

```
> model.conf <-glm(cbind(killed,unkilled)~Dose, family=binomial("cloglog"),
                    data=beetle)
```

```
> summary(model.conf)
```

Call:

```
glm(formula = cbind(killed, unkilled) ~ Dose, family = binomial("cloglog"),
    data = beetle)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.80329	-0.55135	0.03089	0.38315	1.28883

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.572	3.240	-12.21	<2e-16 ***
Dose	22.041	1.799	12.25	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom
Residual deviance: 3.4464 on 6 degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4

Confidence interval

- A $(1-\alpha)100\%$ confidence interval for the parameter of the model can be defined as:

$$(\beta_i \pm Z_{\alpha/2} \times se(\beta_i))$$

```
> C.I<-c(model.conf$coeff[2]-1.96*1.799,model.conf$coeff[2]+1.96*1.799)
```

```
> C.I
```

```
      Dose      Dose  
18.51513 25.56721
```

95% C.I for the slope.

$$\exp\{\beta_i \pm Z_{\alpha/2} \times se(\beta_i)\}$$

Confidence interval in R

Confidence interval in R can be defined using the formula:

```
> confint(object, parm, level = 0.95, ...)
```

If the `parm` option is missed, then R will compute confidence interval for all parameters in the model.

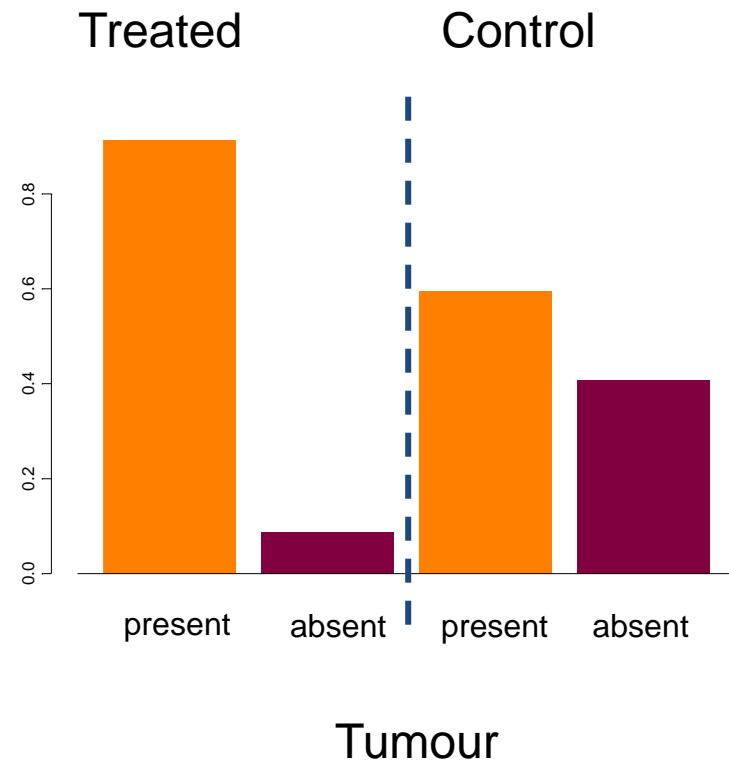
```
> library(MASS)
```

```
> model.conf <- glm(cbind(killed, unkilled) ~ Dose, family = binomial("cloglog"),  
data = beetle)
```

```
> confint(model.conf, level = 0.95)  
Waiting for profiling to be done...  
                2.5 %      97.5 %  
(Intercept) -46.2037 -33.53869  
Dose          18.6903  25.72251
```


Example : mice data

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55



Model formulation

	Tumour present	Tumour absent	Total
Treated	21	2	23
Contol	19	13	32
Total	20	15	55

We want to model the probability to develop a tumour given the treatment group.

The individual data

$$Z_i = \begin{cases} 1 & \text{tumour present} \\ 0 & \text{tumour absent} \end{cases}$$

Number of subjects with tumour

$$Y_i = \sum Z_i$$

Distribution of Y

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

The model for P- logit transformation

$$\log it(P) = \beta_0 + \beta_j$$

How do we interpret the parameters ?

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.3514	0.7400	-3.177	0.00149	**
trti2	1.9719	0.8229	2.396	0.01656	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The parameter estimate for the effect of the control group is -2.3514. The parameter estimate for the effect of the treatment group (the smoked group) is equal to **1.9719**.

Example : mice data

```
> library(MASS)
> fit.mice <- glm(resp ~ trti, family = binomial(link = "logit"))

> confint(fit.mice, level = 0.95)
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) -4.1843896 -1.125530
trti2        0.5295788  3.909658
```

Interpretation

The odds ratio, θ , is equal to $\exp(1.9719)=7.18$. If $\theta > 1$ than the odds for a tumour in the control group is smaller than the odds for a tumour in the treatment group.

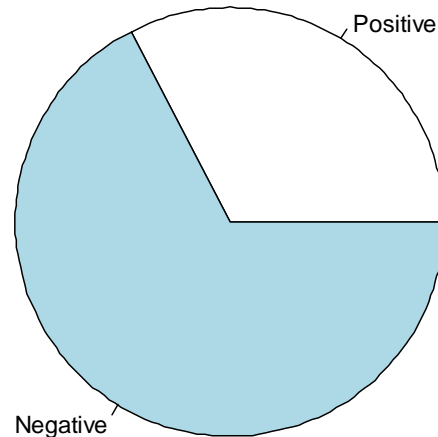
```
> exp(1.9719)
[1] 7.184314
```

95% C.I for the odds ratio:

```
> exp(confint(fit.mice, level=0.95))
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) 0.0152315 0.3244804
trti2       1.6982169 49.8818870
```

Example 4: HIV data

- Consider the HIV data set and the model for HIV (the outcome variable, yes/no or 1/0).
- **Covariates:**
 - age group (also coded 1/0).
 - Age group was coded 1 for people younger than 40.7 years
 - Age
- Response: HIV status (32.6% are positive).



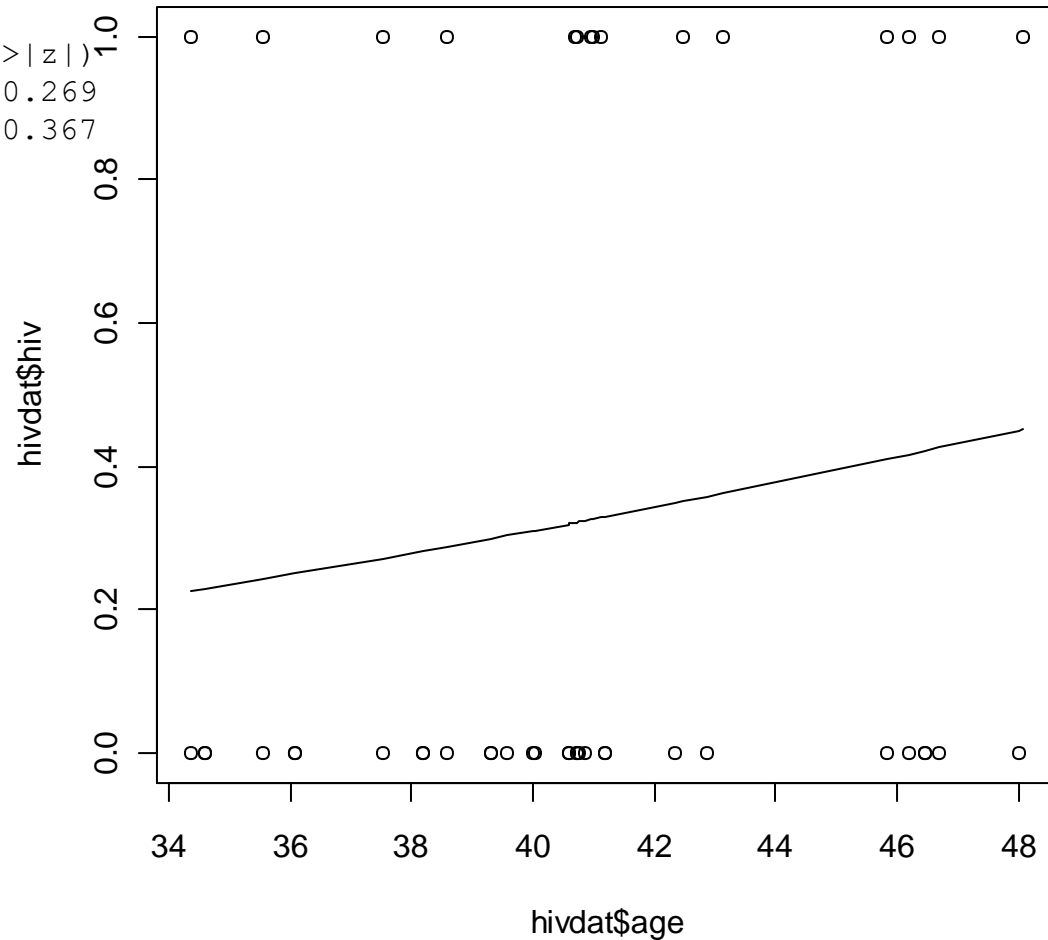
Example 3: HIV data

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.79597	3.43622	-1.105	0.269
age	0.07492	0.08314	0.901	0.367

$$g(\pi_i) = -3.79 + 0.0749 \times age_i$$

$$\pi_i = \frac{e^{-3.79+0.0749 \times age_i}}{1 + e^{-3.79+0.0749 \times age_i}}$$



95% C.I for the parameter estimates

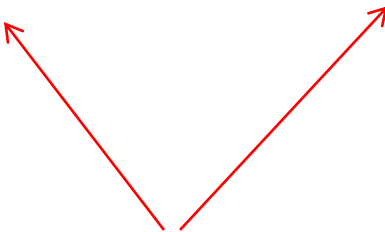
```
> confint(hiv.fit1 , level=0.95)
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) -10.86100806  2.8239564
age          -0.08678935  0.2445192
```


95% C.I for the odds ratio

```
> exp(confint(hiv.fit1, level=0.95))
```

```
Waiting for profiling to be done...
```

	2.5 %	97.5 %
(Intercept)	1.919217e-05	16.843358
age	9.168702e-01	1.277007


$$\exp\{\beta_i \pm Z_{\alpha/2} \times se(\beta_i)\}$$

Chapter 8

Inference

Donson: chapter 5.

Lindsey: chapter 9.

McCullagh & Nelder: chapter 3.

Inference

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

$$H_0 : \beta_j = 0$$

$$H_0 : \beta_j \neq 0$$

We can test the above hypothesis using:

1. Wald test
2. Likelihood ratio test

Wald Test

Asymptotic distribution of the ML estimator

$$\hat{\beta} \sim N(\beta, \phi(X'WX)^{-1})$$

We wish to test the null hypothesis

$$H_o : \beta_j = 0 \quad \text{versus} \quad H_1 : \beta_j \neq 0$$

Test statistic

$$Z_j = \frac{\hat{\beta}_j}{\sqrt{\phi(X'\hat{W}X)^{-1}_{jj}}} \quad , \quad Z_j \sim N(0,1)$$

Which is asymptotically $N(0,1)$ under H_o

Example : toxicity example (Budworm)

Predictor: log(dose)

Sex	Dose (μ g)					
	1	2	4	8	16	32
Male	1	4	9	13	18	20
Female	0	2	6	10	12	16

See example 5 in Chapter 6

$$\eta = \beta_0 + \beta_1 G_i + \beta_2 d_{ij} + \beta_3 G_i \times d_{ij}$$

```
> budworm.lg <- glm(SF ~ sex+ldose+sex:ldose, family=binomial)
```

Variance/covariance matrix of the estimates in R (for the toxicity example)

Variance covariance matrix for the parameter estimates

$$V = \phi(X'WX)^{-1}$$

```
> summary(budworm.lg)$cov.unscaled
```

	(Intercept)	sexM	ldose	sexM:ldose
(Intercept)	0.3054769	-0.3054769	-0.08394089	0.08394089
sexM	-0.3054769	0.6057665	0.08394089	-0.18661802
ldose	-0.0839409	0.0839409	0.02792296	-0.02792296
sexM:ldose	0.0839409	-0.1866180	-0.02792296	0.07289473

Variance/covariance matrix

- The variance can be written in terms of μ and the canonical link function g as:

$$\text{var}(y) = ag'^{-1}(\mu)$$

- The variance matrix

$$\text{var}(y) = V$$

- Fixed effect models assumes that the observations are uncorrelated, therefore the variance matrix is diagonal.
- Diagonal terms=variances of each parameter

Wald test in R (toxicity example)

Call:

```
glm(formula = SF ~ sex * ldose, family = binomial)
```

Coefficients:

	Estimate	Std. Error	z	value	Pr(> z)	
(Intercept)	-2.9935	0.5527	-5.416	6.09e-08	***	
sexM	0.1750	0.7783	0.225	0.822		
ldose	0.9060	0.1671	5.422	5.89e-08	***	
sexM:ldose	0.3529	0.2700	1.307	0.191		

The likelihood ratio statistic

Consider two models:

The model with the maximum number of parameters that can be estimated: the saturated model.

The model of interest with k parameters.

The likelihood ratio:

$$\lambda = \frac{L(\hat{\beta}_{\max})}{L(\hat{\beta})}$$

The likelihood ratio statistic

The likelihood ratio provides a goodness to fit of the model of interest.

Log likelihood ratio

$$\log(\lambda) = \ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}; y)$$

Large value of $\log(\lambda)$ indicates a poor fit.

The deviance

2Log likelihood ratio

$$D = 2\log(\lambda) = 2\left[\ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}; y)\right]$$

Large value of deviance indicates a poor fit.

The deviance

Let us assume that we have two models: M1 and M2.

Deviance of M1:

$$D_{M1} = 2\log(\lambda) = 2\left[\ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}_{M1}; y)\right]$$

Deviance of M2:

$$D_{M2} = 2\log(\lambda) = 2\left[\ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}_{M2}; y)\right]$$

The deviance

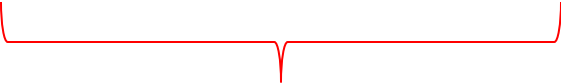
The difference between the deviance of M1 and M2:

$$\Delta D = D_{M1} - D_{M2} = 2 \left[\ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}_{M1}; y) \right] - 2 \left[\ell(\hat{\beta}_{\max}; y) - \ell(\hat{\beta}_{M2}; y) \right]$$


$$\Delta D = D_{M1} - D_{M2} = 2 \left[\ell(\hat{\beta}_{M1}; y) - \ell(\hat{\beta}_{M2}; y) \right]$$

Likelihood ratio test

Consider two model with the following linear predictors:

$$\eta = \beta_0 + \beta_1 G_i + \beta_2 d_{ij}$$


Full model


$$\eta = \beta_0 + \beta_2 d_{ij}$$



Redcued model

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_1 \neq 0$$

Model formulation

Model 1  $\eta = \beta_0 + \beta_1 \times \log(dose)$

Model 2  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(dose)$

Model 3  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(dose) + \beta_3 \times sex \times \log(dose)$

Model 1 in R

```
> budworm.lg1 <- glm(SF ~ ldose, family=binomial)
> summary(budworm.lg1)
Call:
glm(formula = SF ~ ldose, family = binomial)
Deviance Residuals:
      Min       1Q   Median       3Q      Max
-1.7989  -0.8267  -0.1871   0.8950   1.9850
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.7661     0.3701  -7.473 7.82e-14 ***
ldose         1.0068     0.1236   8.147 3.74e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.876  on 11  degrees of freedom
Residual deviance: 16.984  on 10  degrees of freedom
AIC: 51.094
```

Model 2 in R

```
> budworm.lg2 <- glm(SF ~ sex + ldose, family=binomial)
> summary(budworm.lg2)
Call:
glm(formula = SF ~ sex + ldose, family = binomial)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.10540  -0.65343  -0.02225   0.48471   1.42944
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4732     0.4685  -7.413 1.23e-13 ***
sexM           1.1007     0.3558   3.093  0.00198 **
ldose          1.0642     0.1311   8.119 4.70e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 124.876  on 11  degrees of freedom
Residual deviance:  6.757  on  9  degrees of freedom
AIC: 42.867

Number of Fisher Scoring iterations: 4
```

The likelihood ratio test in R

```
> anova(budworm.lg1,budworm.lg2)
```

Analysis of Deviance Table

Model 1: SF ~ ldose

Model 2: SF ~ sex + ldose

	Resid.	Df	Resid. Dev	Df	Deviance
1		10	16.9840		
2		9	6.7571	1	10.227

The likelihood ratio test

The difference between the deviance of M1 and M2:

$$\Delta D = D_{M1} - D_{M2} = 2 \left[\ell(\hat{\beta}_{M1}; y) - \ell(\hat{\beta}_{M2}; y) \right]$$

Under the null hypothesis:

$$\Delta D = D_{M1} - D_{M2} \sim \chi^2_{(p-q)}$$

The likelihood ratio test in R

```
> anova.glm(budworm.lg1,budworm.lg2,test="Chisq")  
Analysis of Deviance Table
```

```
Model 1: SF ~ ldose
```


```
Model 2: SF ~ sex + ldose
```


	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	10	16.9840			
2	9	6.7571	1	10.227	0.001384 **

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model formulation

Model 1  $\eta = \beta_0 + \beta_1 \times \log(dose)$

Model 2  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(dose)$

Model 3  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(dose) + \beta_3 \times sex \times \log(dose)$

Model 3 in R

```
> budworm.lg3<- glm(SF ~ sex*ldose, family=binomial)
> summary(budworm.lg3)
Call:
glm(formula = SF ~ sex * ldose, family = binomial)
Deviance Residuals:
      Min       1Q   Median       3Q      Max
-1.39849  -0.32094  -0.07592   0.38220   1.10375
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.9935     0.5527  -5.416 6.09e-08 ***
sexM           0.1750     0.7783   0.225  0.822
ldose          0.9060     0.1671   5.422 5.89e-08 ***
sexM:ldose     0.3529     0.2700   1.307  0.191
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 124.8756  on 11  degrees of freedom
Residual deviance:  4.9937  on  8  degrees of freedom
AIC: 43.104
Number of Fisher Scoring iterations: 4
```

Likelihood ratio test

$$\eta = \beta_0 + \beta_1 \text{sex} + \beta_2 \log(d)$$

$$\eta = \beta_0 + \beta_1 \text{sex} + \beta_2 \log(d) + \beta_3 \text{sex} \log(d)$$

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

Model 2 versus model 3

```
> anova.glm(budworm.lg2,budworm.lg3,test="Chisq")
Analysis of Deviance Table
```

```
Model 1: SF ~ sex + ldose
```

```
Model 2: SF ~ sex * ldose
```

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	9	6.7571			
2	8	4.9937	1	1.7633	0.1842

We cannot reject the null hypothesis

ANOVA () in R

```
> anova.glm(budworm.lg1,budworm.lg2,budworm.lg3,test="Chisq")
```

Analysis of Deviance Table

Model 1: SF ~ ldose

Model 2: SF ~ sex + ldose

Model 3: SF ~ sex * ldose

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	10	16.9840			
2	9	6.7571	1	10.2270	0.001384 **
3	8	4.9937	1	1.7633	0.184209

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The `update()` function in R

- The `update()` function in R can be used to modify a fitted model by dropping some of the terms.
- The general formulation of the function is given as:

```
Update(old model, ~, . - or + the term we want to drop/ad)
```

The `update()` function in R: example

We would like to drop the interaction term of model 3:

$$\eta = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \log(d) + \beta_3 \times \text{sex} \times \log(d)$$


The diagram illustrates the process of dropping an interaction term from a model. At the top, a linear model equation is shown: $\eta = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \log(d) + \beta_3 \times \text{sex} \times \log(d)$. The interaction term $\beta_3 \times \text{sex} \times \log(d)$ is circled in blue. A blue double-headed arrow points from this circled term to the `-sex:ldose` part of the R code below. A red arrow points from the `budworm.lg3` part of the R code to the first three terms of the equation.

```
updatefit<-update(budworm.lg3, ~. -sex:ldose)
```

The original model

The update () function in R: example

$$\eta = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \log(d) + \beta_3 \times \text{sex} \times \log(d) \Rightarrow \eta = \beta_0 + \beta_1 \times \text{sex} + \beta_2 \times \log(d)$$

```
> summary(updatefit)
```

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.4732	0.4685	-7.413	1.23e-13	***
sexM	1.1007	0.3558	3.093	0.00198	**
ldose	1.0642	0.1311	8.119	4.70e-16	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 124.876 on 11 degrees of freedom
```

```
Residual deviance: 6.757 on 9 degrees of freedom
```

```
AIC: 42.867
```

```
Number of Fisher Scoring iterations: 4
```


Chapter 9: Model Selection

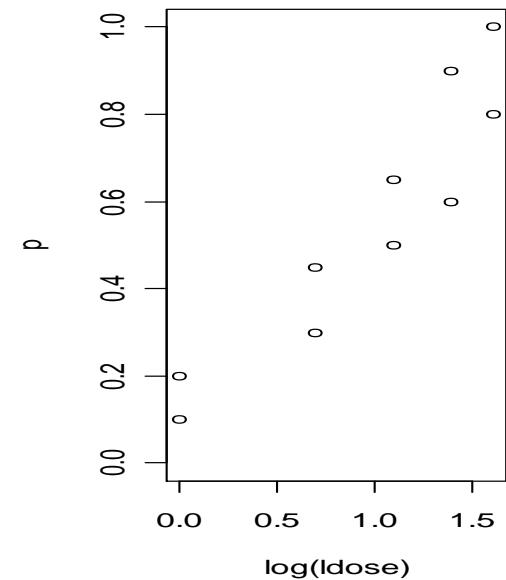
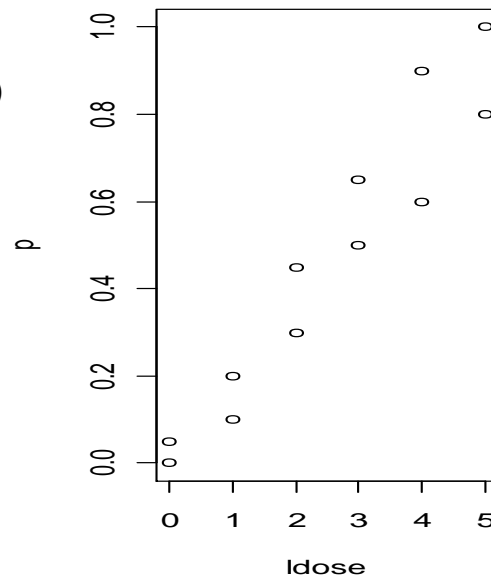
Donson: chapter 4.

Lindsey: chapter 3 (3.3.2 + A.1.4).


McCullagh & Nelder: chapter 2.

Example 1: Budworm Data and Plot in R


```
> ldose <- rep(0:5, 2)
> numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
> sex <- factor(rep(c("M", "F"), c(6, 6)))
> SF <- cbind(numdead, numalive=20-numdead)
> p<-numdead/20
> par(mfrow=c(1,2))
> plot(p ~ ldose)
> plot(p ~ log(ldose))
```



Example 1: model formulation

Model 1  $\eta = \beta_0 + \beta_1 \times \log(d)$

2 parameters

Model 2  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(d)$

3 parameters

Model 3  $\eta = \beta_0 + \beta_1 \times sex + \beta_2 \times \log(d) + \beta_3 \times sex \times \log(d)$

4 parameters

Example of three nested models.

Deviance

The deviance of a model is defined as

$$D = 2\phi(l_{sat} - l_{mod})$$

Where l_{mod} is the log-likelihood of the fitted model and l_{sat} is the log-likelihood of the **saturated model**.

In the saturated model, the number of parameters is equal to the number of observations, so

$$\hat{y} = y$$

For linear regression with Normal data, the deviance is equal to the residual sum of squares

Likelihood and the number of parameters

```
> budworm.lg1 <- glm(SF ~ ldose, family=binomial)
> budworm.lg2 <- glm(SF ~ sex + ldose, family=binomial)
> budworm.lg3 <- glm(SF ~ sex*ldose, family=binomial)
>
> >
> logLik(budworm.lg1)
'log Lik.' -23.54722 (df=2)
> logLik(budworm.lg2)
'log Lik.' -18.43373 (df=3)
> logLik(budworm.lg3)
'log Lik.' -17.55206 (df=4)
```

$-\log(L)$ increases as the number of parameters increases.

Deviance and the number of parameters

```
> budworm.lg1$null.deviance
[1] 124.8756
> budworm.lg1$deviance
[1] 16.98403
> budworm.lg2$null.deviance
[1] 124.8756
> budworm.lg2$deviance
[1] 6.757064
> budworm.lg3$null.deviance
[1] 124.8756
> budworm.lg3$deviance
[1] 4.993727
```

Deviance decreases as the number of parameters increases.

Akaike Information Criterion (AIC)

- The Akaike information criterion (AIC) defines as:

$$AIC = -2\log(\textit{likelihood}) + 2.p$$

- The model with minimal AIC tries to find an optimal compromise between model fit and model complexity.
- The R function `stepAIC()` of the package MASS provides such a functionality.
- The direction option specifies the strategy.

Goodness-of-fit and model complexity

$$AIC = \underbrace{-2\log(\textit{likelihood})}_{\text{Goodness-of-fit}} + \underbrace{2.p}_{\text{Complexity}}$$

```
> extractAIC(budworm.lg1, k=2)
[1] 2.00000 51.09443
> extractAIC(budworm.lg2, k=2)
[1] 3.00000 42.86747
> extractAIC(budworm.lg3, k=2)
[1] 4.00000 43.10413
```

Goodness-of-fit and model complexity

```
> summary(budworm.lg3)
```

Call:

```
glm(formula = SF ~ sex * ldose, family = binomial)
```

AIC: 43.104

```
> library(MASS)
```

```
> stepAIC(budworm.lg3, direction = "backward")
```

Start: AIC=43.1

SF ~ sex * ldose

Starting point

	Df	Deviance	AIC
- sex:ldose	1	6.7571	42.867
<none>		4.9937	43.104

Step: AIC=42.87

SF ~ sex + ldose

In the first step the interaction is dropped

	Df	Deviance	AIC
<none>		6.757	42.867
- sex	1	16.984	51.094
- ldose	1	118.799	152.909

```
Call: glm(formula = SF ~ sex + ldose, family = binomial)
```

Coefficients:

(Intercept)	sexM	ldose
-3.473	1.101	1.064

Final model

Degrees of Freedom: 11 Total (i.e. Null); 9 Residual

Null Deviance: 124.9

Residual Deviance: 6.757 AIC: 42.87

Model selection

- The basic idea of the procedure is to start from a given model (null model) and take a series of steps by either deleting or adding a term in the model from a list of candidates for inclusion, called the *scope* of the search and defined by a model formula.

The criteria seen before will be used in model selection which involves

- choice of distribution and link function
- covariate(s) to include in the model

Example 2: data and model formulation

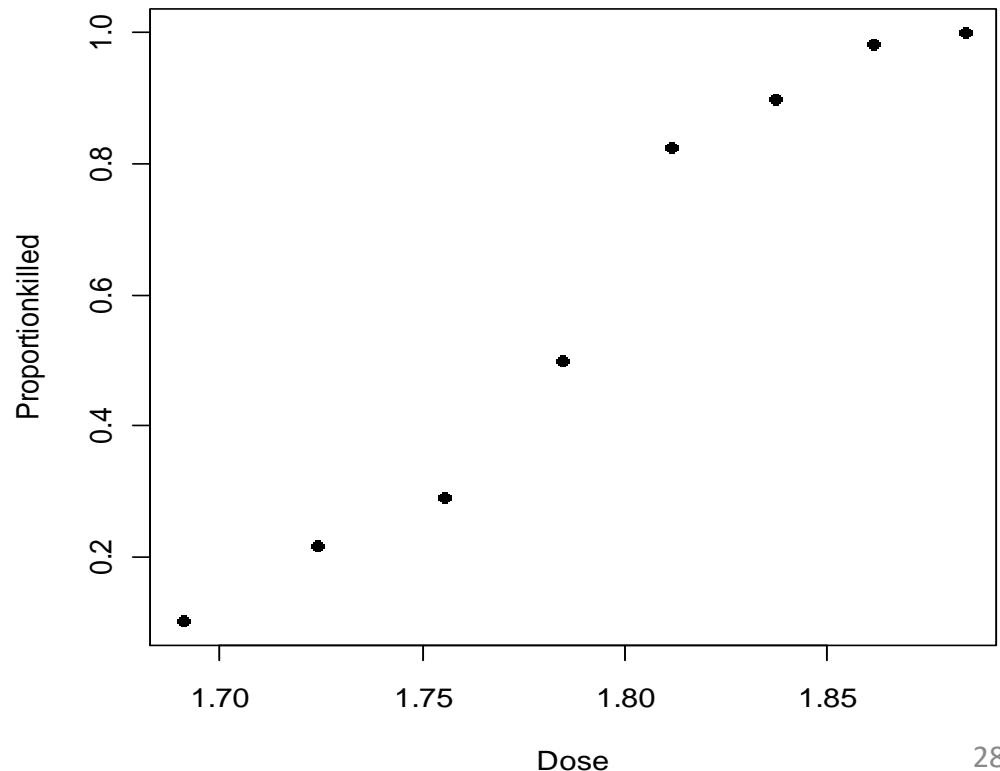
```
beetle<-read.table("C:...../beetle.txt", header = TRUE)
attach(beetle)
Proportionkilled<-killed/beetles
plot(Proportionkilled~Dose, main="Proportion of the killed beetles")
```

$$Y_i \sim \text{Bin}(\pi(d_i), n_i)$$

$$g(\pi(d_i)) = \beta_0 + \beta_1 d_i$$

g is the link function:

- logit.
- probit.
- cloglog.



Model with logit link

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 d_i$$

Where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 d_i)}{1 + \exp(\beta_0 + \beta_1 d_i)}$$

Model 1: binomial with logit link

```
> t1 <-glm(cbind(killed,unkilled)~Dose, family=binomial("logit"))
```

```
➤ summary(t1)
```

```
➤ Call:
```

```
glm(formula = cbind(killed, unkilled) ~ Dose, family =  
     binomial("logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5941	-0.3944	0.8329	1.2592	1.5940

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.717	5.181	-11.72	<2e-16 ***
Dose	34.270	2.912	11.77	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202 on 7 degrees of freedom

Residual deviance: 11.232 on 6 degrees of freedom

AIC: 41.43

Number of Fisher Scoring iterations: 4

Model with probit link

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 d_i$$

Where

$$\Phi = \int_{-\infty}^{\beta_0 + \beta d_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right) dz$$

Model2: with probit link

```
> t2 <-glm(cbind(killed,unkilled)~Dose, family=binomial("probit"))
> summary(t2)
```

Call:

```
glm(formula = cbind(killed, unkilled) ~ Dose, family = binomial("probit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5714	-0.4703	0.7501	1.0632	1.3449

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-34.935	2.648	-13.19	<2e-16 ***
Dose	19.728	1.487	13.27	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202 on 7 degrees of freedom

Residual deviance: 10.120 on 6 degrees of freedom

AIC: 40.318

Number of Fisher Scoring iterations: 4

Model with c-log-log link

$$g(\pi_i) = \log(-\log(1 - \pi_i)) = \beta_0 + \beta_1 d_i$$

Where

$$\pi_i = 1 - e^{-(\beta_0 + \beta d_i)}$$

$$1 - \pi_i = e^{-e^{(\beta_0 + \beta d_i)}}$$

$$\log(1 - \pi_i) = -e^{(\beta_0 + \beta d_i)}$$

$$\log(-\log(1 - \pi_i)) = \log(e^{(\beta_0 + \beta d_i)}) = \beta_0 + \beta d_i$$

Model 3: with cloglog link

```
> t3 <-glm(cbind(killed,unkilled)~Dose, family=binomial("cloglog"))
> summary(t3)
```

Call:

```
glm(formula = cbind(killed, unkilled) ~ Dose, family = binomial("cloglog"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.80329	-0.55135	0.03089	0.38315	1.28883

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.572	3.240	-12.21	<2e-16 ***
Dose	22.041	1.799	12.25	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom
Residual deviance: 3.4464 on 6 degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4

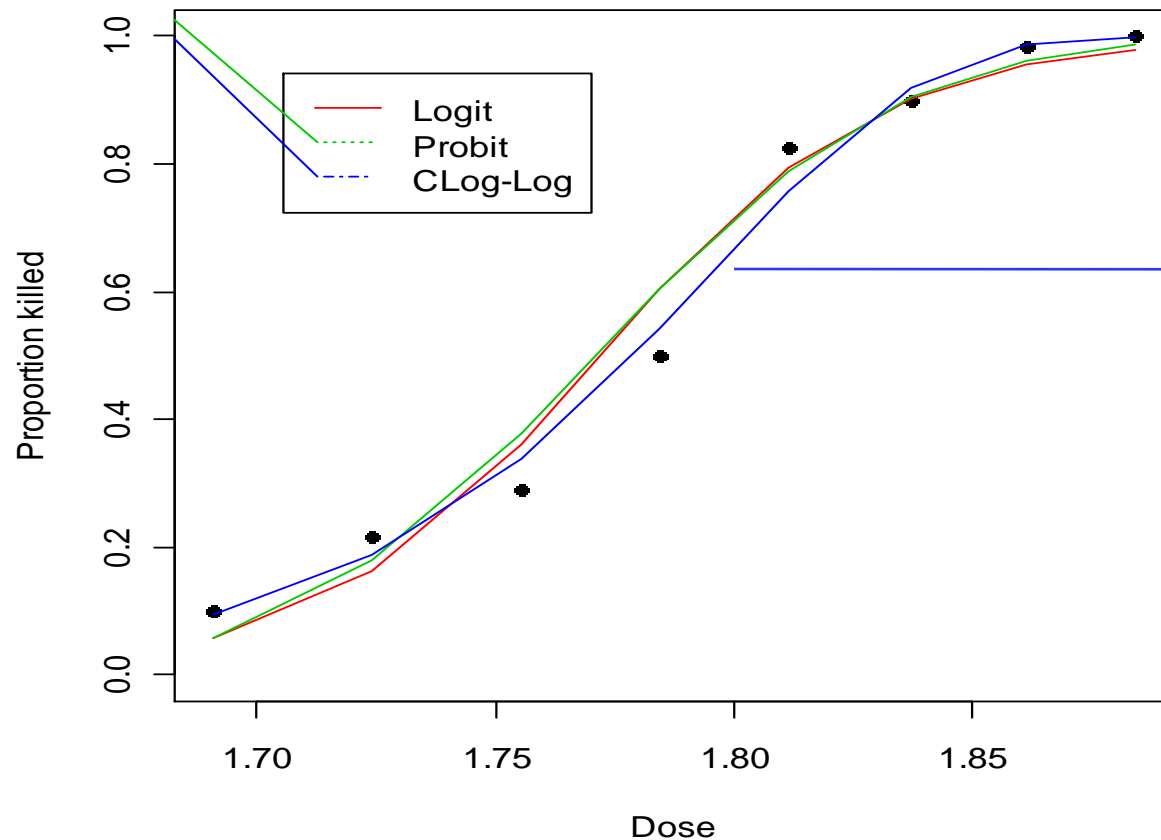
Model selection based on AIC

- Selection of terms for deletion or inclusion is based on Akaike's information criterion (AIC).
- In R, the function “`extractAIC(model)`” will give AIC .

Model	Likelihood	No parameters	AIC
Logit	-18.71513	2	41.43
Probit	-18.15890	2	40.318
Cloglog	-14.82224	2	33.44

- According to the AIC criteria, the model with cloglog link function will be chosen as a good model.

Plot of the estimated models



Best model

```
> extractAIC(t1, k=2)
[1] 2.00000 41.43027
> extractAIC(t2, k=2)
[1] 2.0000 40.3178
> extractAIC(t3, k=2)
[1] 2.00000 33.64448
```


Chapter 10: Model diagnostic

```
> library(boot)
```

```
> library(graphics)
```

Donson: chapter 7.

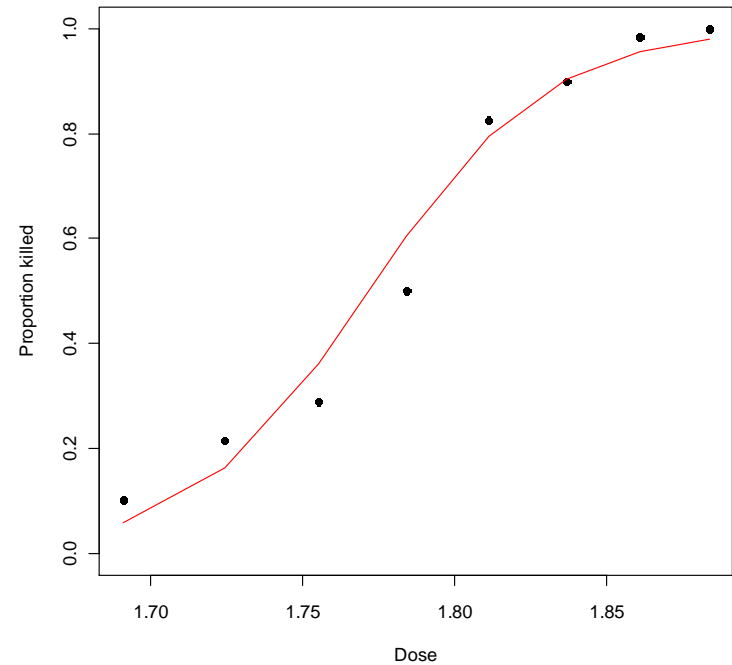
Lindsey: Appendix B.

McCullagh & Nelder: chapter 2.

Example 1: the beetle example

Consider beetle example
with the logit model

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 d_i$$



Example 1: the beetle example in R

```
> t1 <-glm(cbind(killed,unkilled)~Dose, family=binomial("logit"))
```

```
> summary(t1)
```

Call:

```
glm(formula = cbind(killed, unkilled) ~ Dose, family = binomial("logit"))
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5941	-0.3944	0.8329	1.2592	1.5940

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.717	5.181	-11.72	<2e-16 ***
Dose	34.270	2.912	11.77	<2e-16 ***

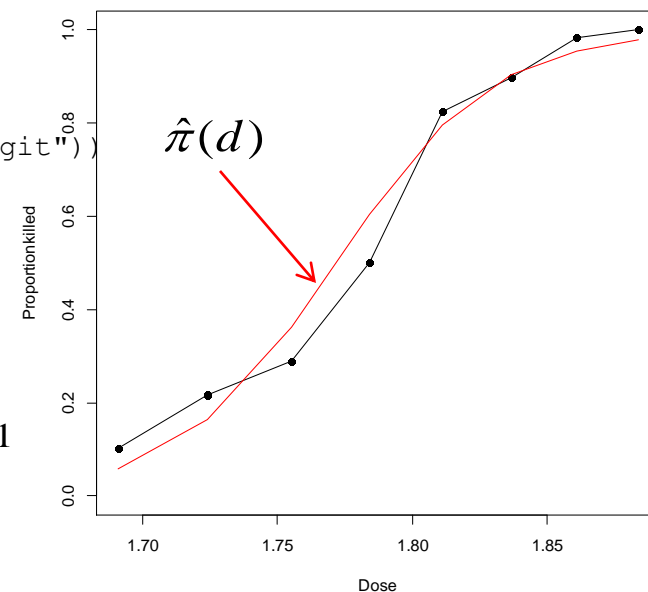
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202 on 7 degrees of freedom
 Residual deviance: 11.232 on 6 degrees of freedom
 AIC: 41.43

$$\hat{\beta}_0 = -60.71$$

$$\hat{\beta}_1 = 34.27$$



Residual Analysis

Several kinds of residuals can be defined for GLMs:

- **Raw response:** $R_i = y_i - \hat{\mu}_i$
- **working:** from the working response in the IWLS algorithm
- **Pearson**

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- Such that $\sum_i (r_i^P)^2$ equals the generalized Pearson statistic
- **deviance:** r_i^D such that $\sum_i (r_i^D)^2$ equals the deviance.

These definitions are all equivalent for Normal models

Raw residuals in R

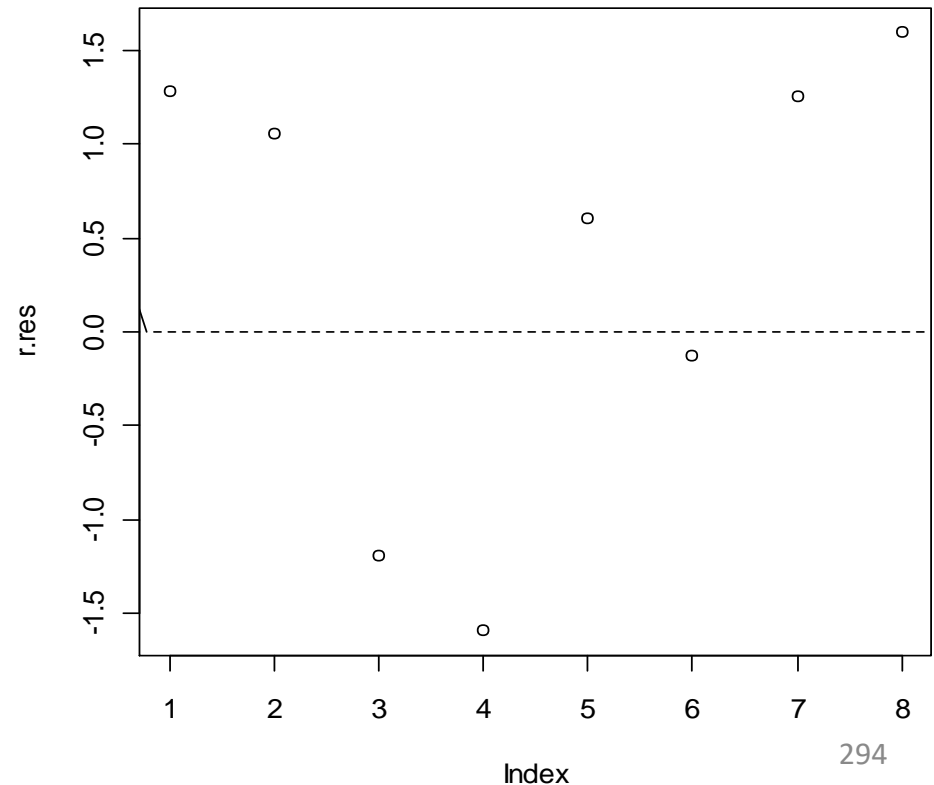
The raw residual is defined as:

$$r_i = y_i - \hat{\mu}_i$$

```
library(boot)
library(graphics)
r.res<-resid(t1)
par(mfrow=c(2,2))
plot(r.res)
abline(h=0, lty=2)
```

For binary data, these residuals
are not really informative

why not ?



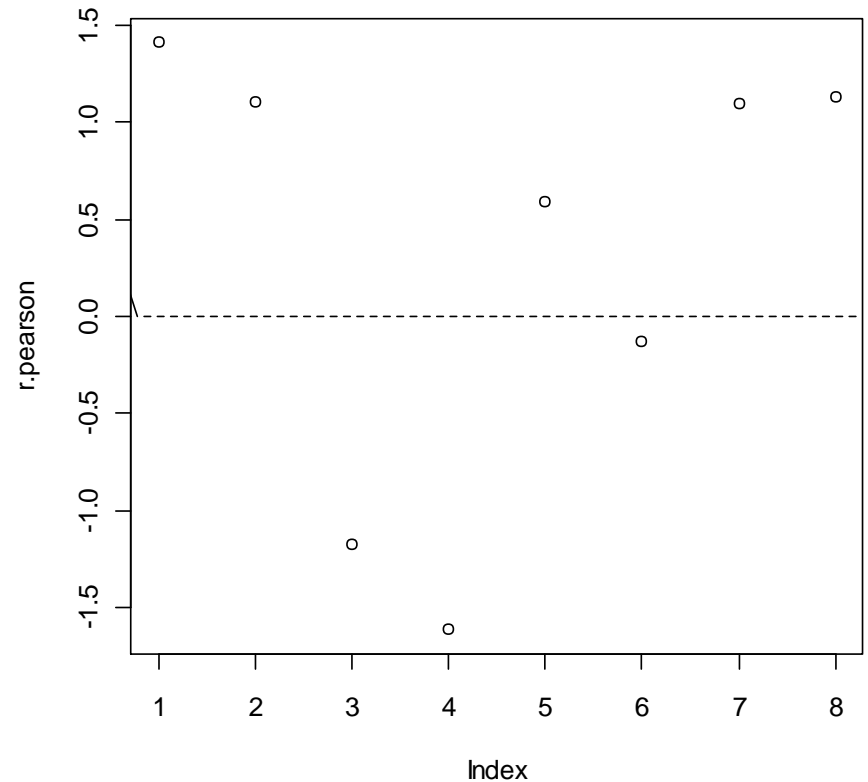
Pearson residual in R

Pearson residual :

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

$$\sum_i (r_i^P)^2$$

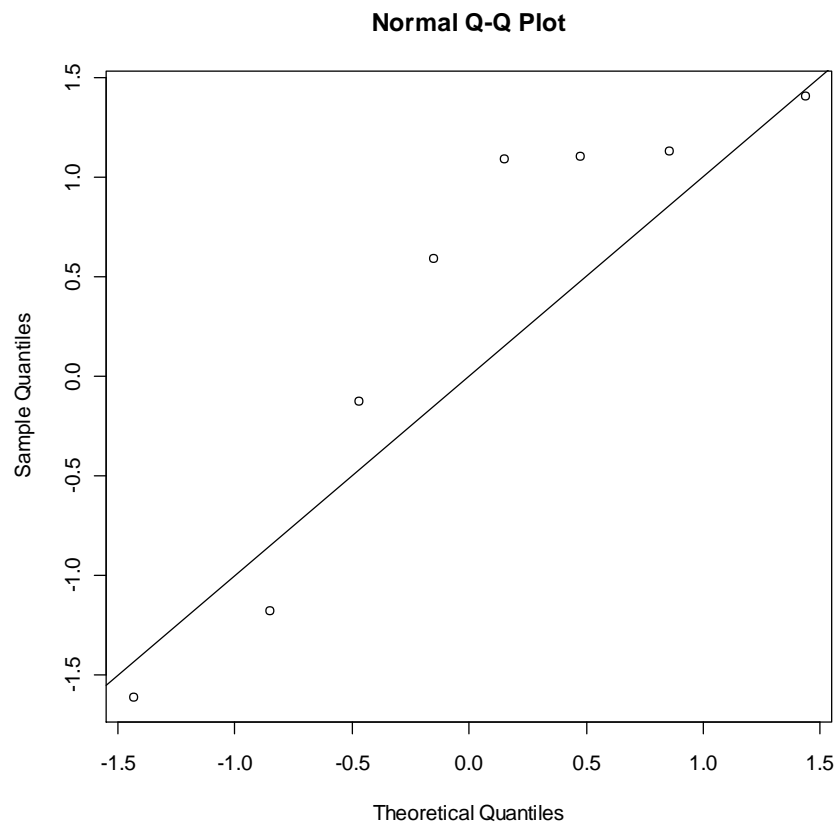
```
> r.pearson <- resid(t1, type="pearson")  
> plot(r.pearson)  
> abline(h=0, lty=2)
```



Pearson residuals

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \sim N(0,1)$$

```
> par(mfrow=c(1,1))  
> qqnorm(r.pearson)  
> abline(0,1)
```



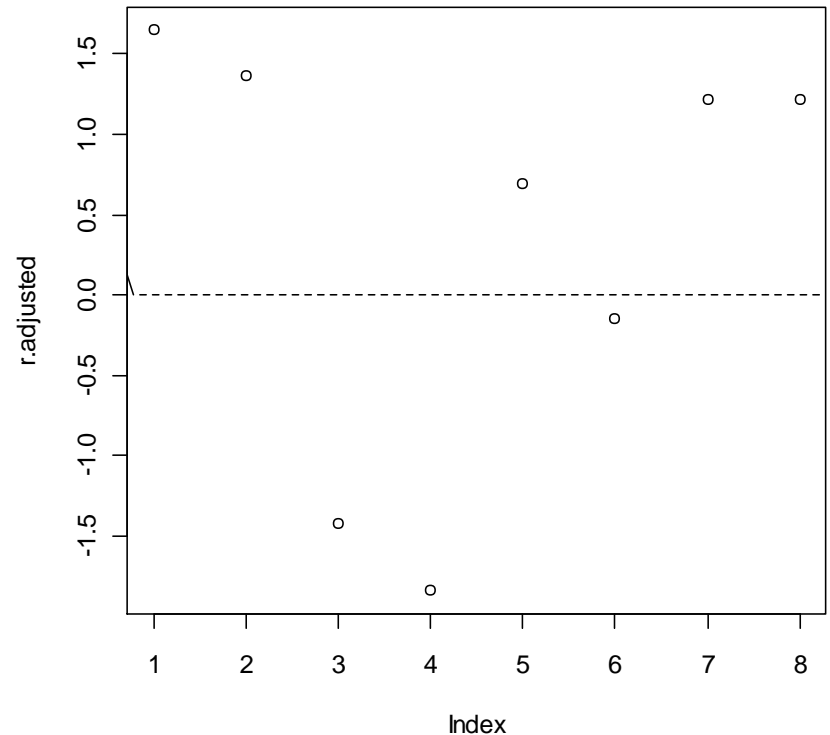
Adjusted residual in R

The adjusted residual

$$e_i^a = \frac{e_i^p}{(1 - H_{ii})^{1/2}}$$

see slide
299

```
> hii <- hatvalues(t1)
> r.adjusted <- r.pearson/sqrt(1 - hii)
> plot(r.adjusted)
> plot(r.adjusted)
> abline(h = 0, lty = 2)
```



Deviance residual

$$r_i^d = \text{sign}(Y_i - n_i \hat{p}_i) \left(2y_i \ln \left(\frac{Y_i}{n_i \hat{p}_i} \right) + 2(n_i - y_i) \ln \left(\frac{n_i - Y_i}{n_i (1 - \hat{p}_i)} \right) \right)$$

$$\Rightarrow D = \sum_{i=1}^n (r_i^D)^2$$

High leverage and influential points in logistic regression

Linear models :

$$Y = X\beta + \varepsilon \quad \hat{\beta} = (X^t X)^{-1} X^t Y \quad \hat{Y} = X\hat{\beta} = HY, \\ H = X(X^t X)^{-1} X^t \quad H^2 = H$$

$$\Rightarrow Y - \hat{Y} = (I - H)Y \\ = (I - H)(Y - \hat{Y}) \quad \text{since } H\hat{Y} = H(HY) = H^2Y = HY = \hat{Y} \\ = (I - H)(\hat{e})$$

\Rightarrow raw residuals satisfy $\hat{e} = (I - H)\hat{e}$

logistic regression

$$e^P = (I - H)e^P, \text{ where } e_i^P = \frac{Y_i - n_i \hat{p}_i}{(n_i \hat{p}_i (1 - \hat{p}_i))^{1/2}}$$

(Reference: Pregibon (1981))

High leverage points in logistic regression

$$\mathbf{e}^P = \underbrace{(\mathbf{I} - \mathbf{H})}_{\mathbf{M}} \mathbf{e}^P \quad \Rightarrow \quad e_i^a = \frac{e_i^p}{(1 - H_{ii})^{1/2}}$$

M spans residual space \mathbf{e}^P .

This suggests that small m_{ii} (or large h_{ii}) should be useful in detecting extreme points in the design space \mathbf{X} .

We have

$$\sum_{i=1}^n h_{ii} = p$$

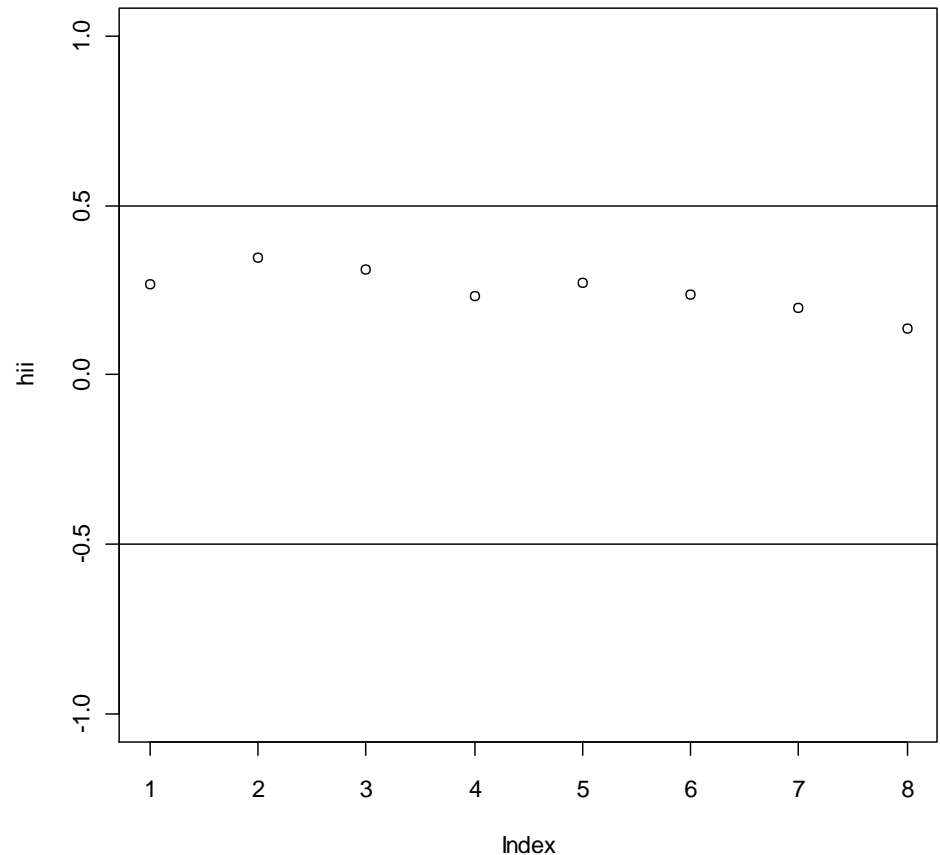
Therefore we consider

$$h_{ii} > \frac{2p}{n}$$

as “high leverage points”.

High leverage points in logistic regression

```
> hii <- hatvalues(t1)
> sum(hii)
[1] 2
> plot(hii,ylim=c(-1,1))
> 2*2/8
[1] 0.5
> abline(0.5,0)
> abline(-0.5,0)
```



Cook's distance in logistic regression

Using LRT it can be shown that

$$\left\{ \beta : -2 \ln \left\{ \frac{L(\beta)}{L(\hat{\beta})} \right\} \leq \chi^2_{1-\alpha, p} \right\} \text{ is an approx. } 100 (1-\alpha)\% \text{ CI for } \beta$$

$$\Rightarrow D_i = -2 \left\{ \ln \frac{L(\beta)}{L(\hat{\beta})} \right\}$$

measures change in the parameter when i^{th} observation removed;
difficult to calculate.

Cook's distance...

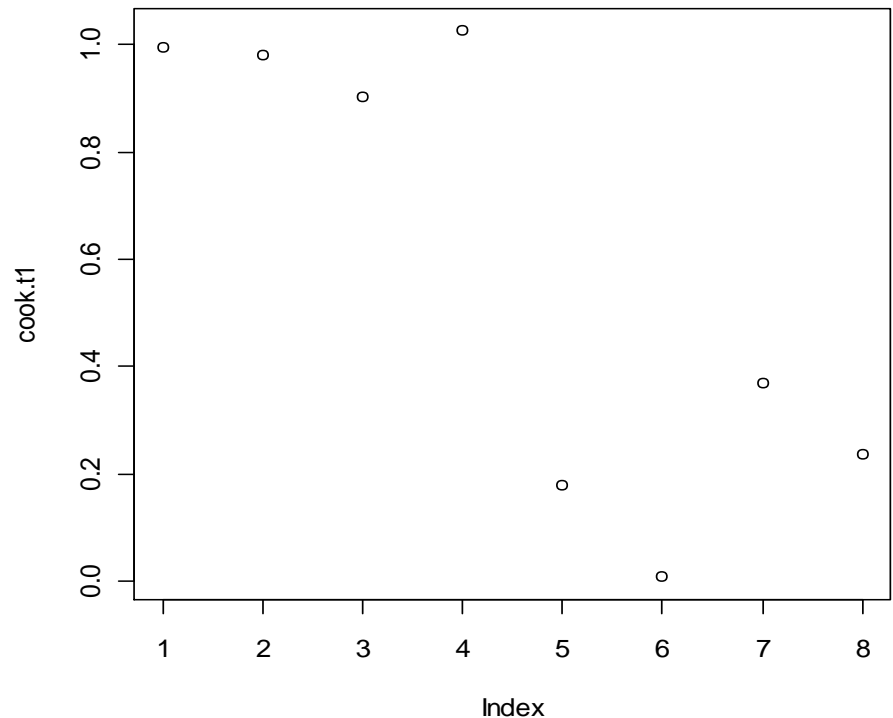
- Using Taylor expansion we have:

$$\left\{ \beta : -2 \ln \left\{ \frac{L(\beta)}{L(\hat{\beta})} \right\} \leq \chi^2_{1-\alpha, p} \right\} \approx \left\{ \beta : (\beta - \hat{\beta})^t \mathbf{X}^t \hat{\mathbf{D}} \mathbf{X} (\beta - \hat{\beta}) \leq \chi^2_{1-\alpha, p} \right\}$$

$$\Rightarrow D_i \approx (\hat{\beta}_{-i} - \hat{\beta})^t \mathbf{X}^t \hat{\mathbf{D}} \mathbf{X} (\hat{\beta}_{-i} - \hat{\beta})$$

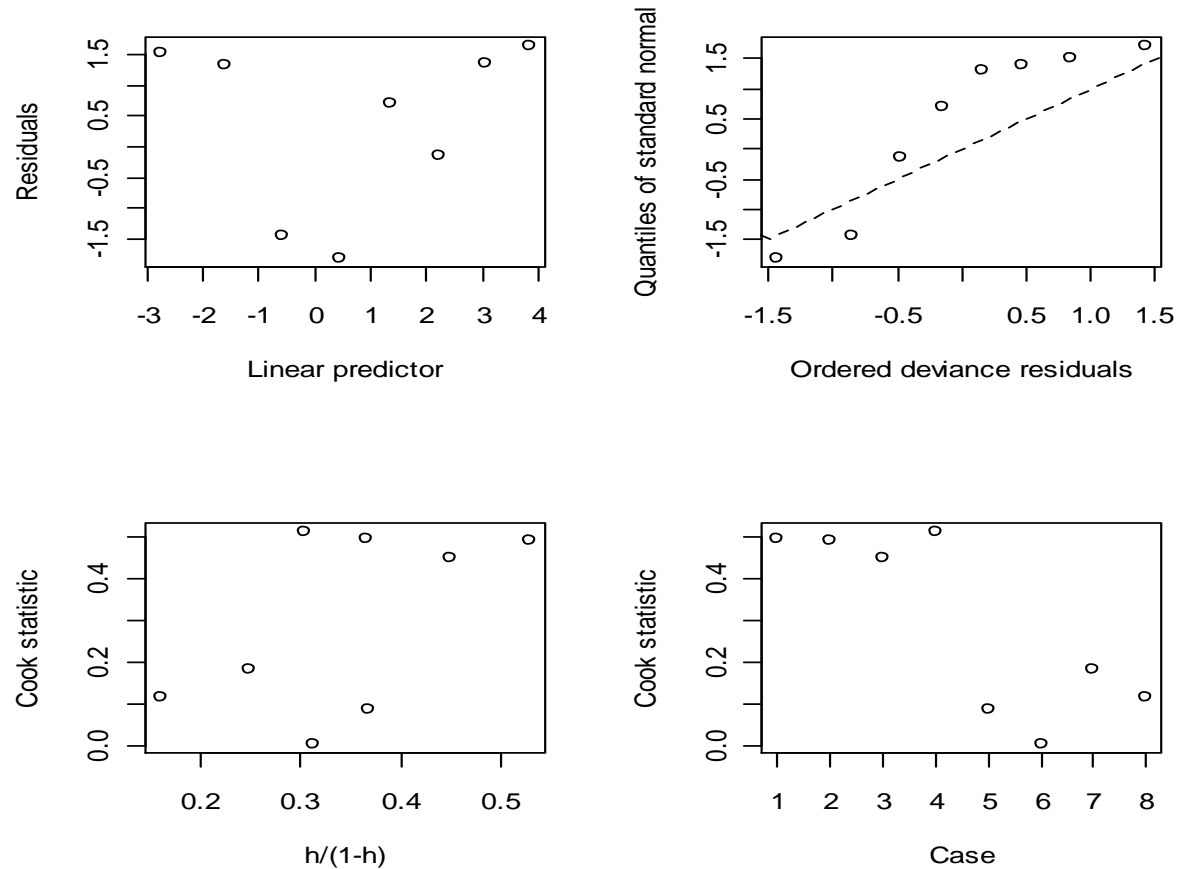
Cook's distance in R

```
> p.t1 <- length(coef(t1))  
> cook.t1 <- ((r.pearson^2) * hii)/((1 - hii)^2)  
> cook.t11 <- cooks.distance(t1) * p.t1  
> plot(cook.t1)
```



Diagnostic with R: the beetle data with logit link function

```
> library(boot)
> glm.diag.plots(beatlefit)
```



summary

- Model formulation: distribution, linear predictor and link functions.
- Estimation and inference.
- Model selection.
- Model diagnostic.

Extra Example

Effect of drug on cardiac death
(McCullagh & Nelder 1983)

Effect of drug on cardiac deaths

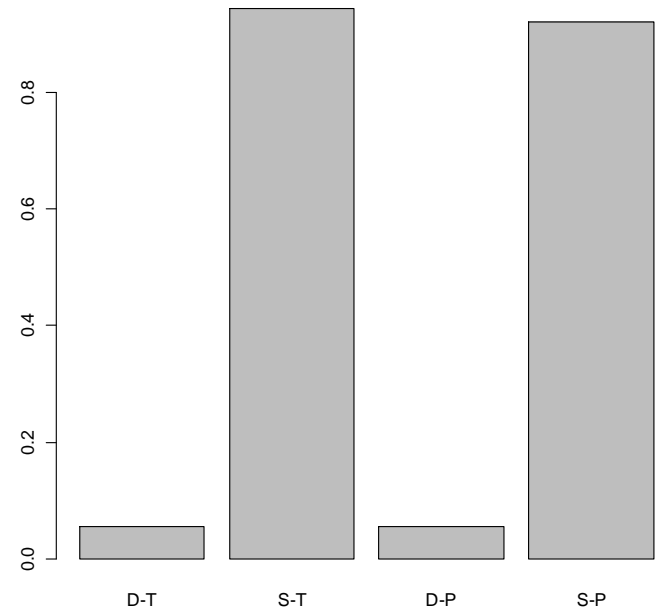
Study of the effect of a drug on cardiac death.

Patients treated with:

Drug: sulphapyrazone.

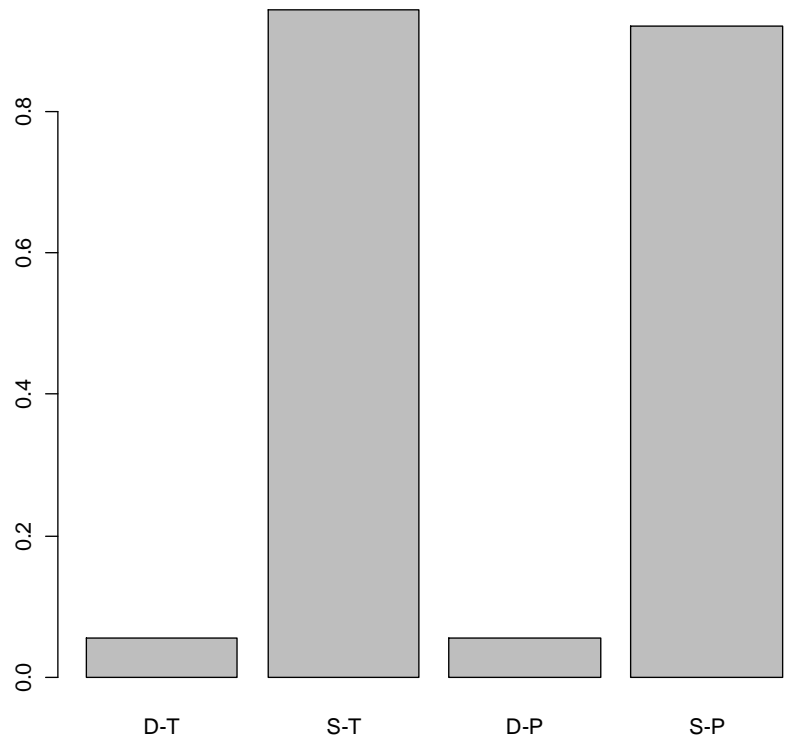
Placebo.

Number of deaths and survivors (from the total) were recorded.



Effect of drug on cardiac deaths

```
>d<-c(41,60)
>s<-c(692,682)
> n<-c(733,742)
> gr<-c("T","P")
> cbind(d,s,n)
  d s n
[1,] 41 692 733
[2,] 60 682 742
```



2 X 2 table

y_{00}	y_{01}	n_1	π_{00}	π_{01}	$\pi_{0.}$
y_{10}	y_{11}	n_2	π_{10}	π_{11}	$\pi_{1.}$
m	$n - m$	n	$\pi_{.0}$	$\pi_{.1}$	1

$$OR = \frac{\frac{\pi_{00}}{1 - \pi_{00}}}{\frac{\pi_{10}}{1 - \pi_{10}}}$$

2 X 2 table

$$\begin{array}{ccc}
 \pi_{00} & \pi_{01} & \pi_{0.} \\
 \pi_{10} & \pi_{11} & \pi_{1.} \\
 \pi_{.0} & \pi_{.1} & 1
 \end{array}$$

$$OR = \frac{\frac{\pi_{00}}{1 - \pi_{00}}}{\frac{\pi_{10}}{1 - \pi_{10}}}$$

$$\hat{\pi}_{00} = \frac{y_{00}}{n_1}, 1 - \hat{\pi}_{00} = \frac{y_{01}}{n_1}$$

$$\hat{\pi}_{10} = \frac{y_{10}}{n_2}, 1 - \hat{\pi}_{10} = \frac{y_{11}}{n_2}$$

$$OR = \frac{\frac{y_{00}}{y_{01}}}{\frac{y_{10}}{y_{11}}} = \frac{y_{00} \times y_{11}}{y_{01} \times y_{10}}$$

The odds ratio

y_{00}	y_{01}	n_1	> cbind(d,s,n)			
			d	s	n	
y_{10}	y_{11}	n_2	[1,]	41 692 733		T
			[2,]	60 682 742		C
m	$n - m$	n		deaths		
						totals

$$OR = \varphi = \frac{y_{00} \times y_{11}}{y_{01} \times y_{10}}$$

$$\varphi = \frac{41 \times 682}{60 \times 692} = 0.6735$$

$$\log(\varphi) = -0.3953$$

What does an OR=0.6735 mean ?

Conditional likelihood for 2 X 2 table

$$y_{00} \quad y_{01} \quad n_1$$

$$y_{10} \quad y_{11} \quad n_2$$

$$m \quad n - m \quad n$$

$$y_{00} \sim B(n_1, \pi_{00})$$

$$y_{10} \sim B(n_2, \pi_{10})$$

$$y_{00} + y_{10} = m$$

Conditional likelihood for 2 X 2 table

$$y_{00} \quad y_{01} \quad n_1$$

$$y_{10} \quad y_{11} \quad n_2$$

$$m \quad n - m \quad n$$

$$y_{00} \sim B(n_1, \pi_{00})$$

$$y_{10} \sim B(n_2, \pi_{10})$$

$$y_{00} + y_{10} = m$$

$$x_i = \begin{cases} 1 & T \\ 0 & P \end{cases}$$

Conditional likelihood for 2 X 2 table

$$y_{00} \sim B(n_1, \pi_{00})$$

$$y_{10} \sim B(n_2, \pi_{10})$$

$$x_i = \begin{cases} 1 & T \\ 0 & P \end{cases}$$

$$y_i \sim B(n_i, \pi_i)$$

$$g(\pi_i) = \beta_0 + \beta_1 x_i$$

$$\pi_i = \begin{cases} \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} & T \\ \frac{e^{\beta_0}}{1 + e^{\beta_0}} & P \end{cases}$$

The odds ratio

$$\varphi = OR = \frac{\frac{\pi_1}{1 - \pi_1}}{\frac{\pi_2}{1 - \pi_2}} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

$$\log(\varphi) = \beta_1$$

Inference

$$y_i \sim B(n_i, \pi_i)$$

$$g(\pi_i) = \beta_0 + \beta_1 x_i$$

$$H_0 : \pi_T = \pi_P$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \pi_T \neq \pi_P$$

$$H_1 : \beta_1 \neq 0$$

$$H_0 : \varphi = e^{\beta_1} = 1$$

$$H_1 : \varphi = e^{\beta_1} \neq 1$$

R output

Call:

```
glm(formula = d/n ~ gr, family = "binomial")
```

Deviance Residuals:

```
[1] 0 0
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4307	3.6680	-0.663	0.508
grT	-0.3953	5.6912	-0.069	0.945

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4.9015e-03 on 1 degrees of freedom

Residual deviance: 2.4213e-16 on 0 degrees of freedom

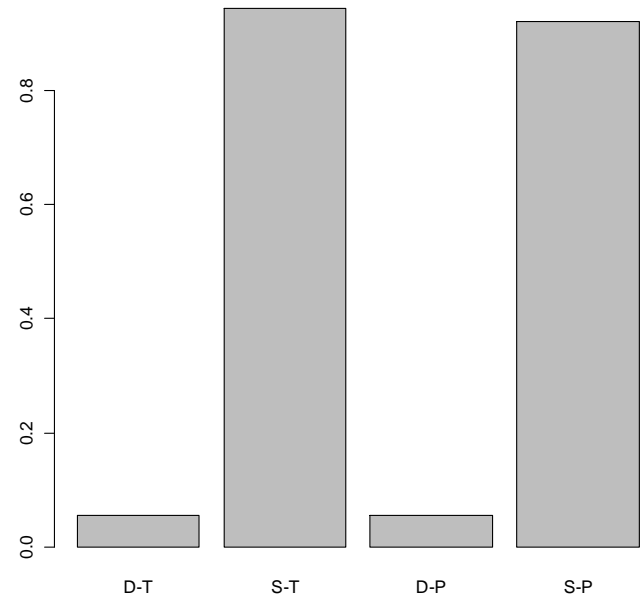
AIC: 4.2838

Effect of drug on cardiac deaths

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.4307	3.6680	-0.663	0.508
grT	-0.3953	5.6912	-0.069	0.945

We cannot reject the null hypothesis.



Example

Habitat preferences of lizards
(McCullagh & Nelder 1983)

Section 4.6, page 128 (first edition)

Habitat preferences of lizards

- A study consists of two lizards type: Grohami and Opalinus.
- Response: number of sites (from the total) occupied by Grahami lizards.
- Covariates:
 1. Height of the site (H).
 2. Diameter (D).
 3. Sun condition of the site (S, sun/ shade).
 4. Time of the day (T).

Habitat preferences of lizards

> habitat

	G	Total	S	D	H	T
1	20	22	S1	D1	H1	Early
2	8	9	S1	D1	H1	Mid
3	4	8	S1	D1	H1	Late
4	13	13	S1	D1	H2	Early
5	8	8	S1	D1	H2	Mid
6	12	12	S1	D1	H2	Late
7	8	11	S1	D2	H1	Early
8	4	5	S1	D2	H1	Mid
9	5	8	S1	D2	H1	Late
10	6	6	S1	D2	H2	Early
11	0	0	S1	D2	H2	Mid
12	1	2	S1	D2	H2	Late
13	34	45	S2	D1	H1	Early
14	69	89	S2	D1	H1	Mid
15	18	28	S2	D1	H1	Late
16	31	36	S2	D1	H2	Early
17	55	59	S2	D1	H2	Mid
18	13	16	S2	D1	H2	Late
19	17	32	S2	D2	H1	Early
20	60	92	S2	D2	H1	Mid
21	8	16	S2	D2	H1	Late
22	12	13	S2	D2	H2	Early
23	21	26	S2	D2	H2	Mid
24	4	8	S2	D2	H2	Late

S: sun conditions sun / shade).

D: diameter (<2 / > 2).

H: height (< 5 / > 5).

T: time of day (early/ mid day/late).

Habitat preferences of lizards: model formulation

$$y_{ijkl} \sim B(n_{ijkl}, \pi_{ijkl})$$

Total sample size.



Number of sites occupied by Grahmi lizards.

π_{ijkl} = The probability that a site is occupied by Grahmi lizards.

$$\begin{aligned} g(\pi_{ijkl}) &= \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \dots \\ &= \beta_0 + \beta_1 S + \beta_2 D + \beta_3 H + \beta_4 T + \dots \end{aligned}$$

Habitat preferences of lizards: model formulation in R

Main effects model in R

$$g(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l$$

```
> f1<-glm((G/Total)~H+D+S+T,family="binomial",data=habitat)
```

R output

```
> summary(f1)
```

Call:

```
glm(formula = (G/Total) ~ H + D + S + T, family = "binomial",  
     data = habitat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.50878	-0.11019	0.02009	0.26466	0.52322

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.0618	1.4060	1.466	0.143
HH2	1.0631	1.1222	0.947	0.343
DD2	-0.8798	1.0841	-0.812	0.417
SS2	-0.6415	1.0884	-0.589	0.556
TLate	-1.2054	1.2761	-0.945	0.345
TMid	0.0587	1.4590	0.040	0.968

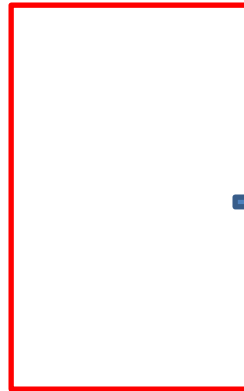
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4.6730 on 22 degrees of freedom
Residual deviance: 1.5417 on 17 degrees of freedom
(1 observation deleted due to missingness)
AIC: 28.658

Interpretation

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.0618	1.4060	1.466	0.143
HH2	1.0631	1.1222	0.947	0.343
DD2	-0.8798	1.0841	-0.812	0.417
SS2	-0.6415	1.0884	-0.589	0.556
TLate	-1.2054	1.2761	-0.945	0.345
TMid	0.0587	1.4590	0.040	0.968



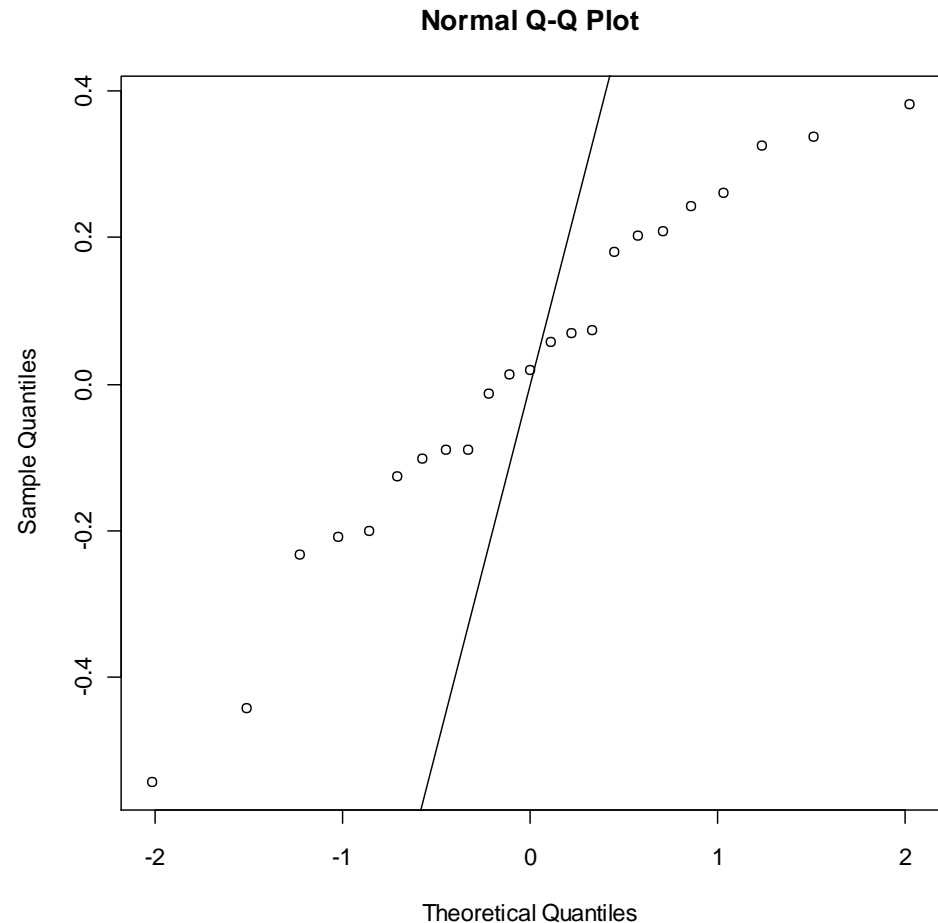
All the parameters estimates are not significant.

We will look at this problem again when we will speak about over/under dispersion of binomial data.

diagnostic

```
>r.pearson<-resid(f1, type="pearson")  
> par(mfrow=c(1,1))  
> qqnorm(r.pearson)  
> abline(0,1)
```

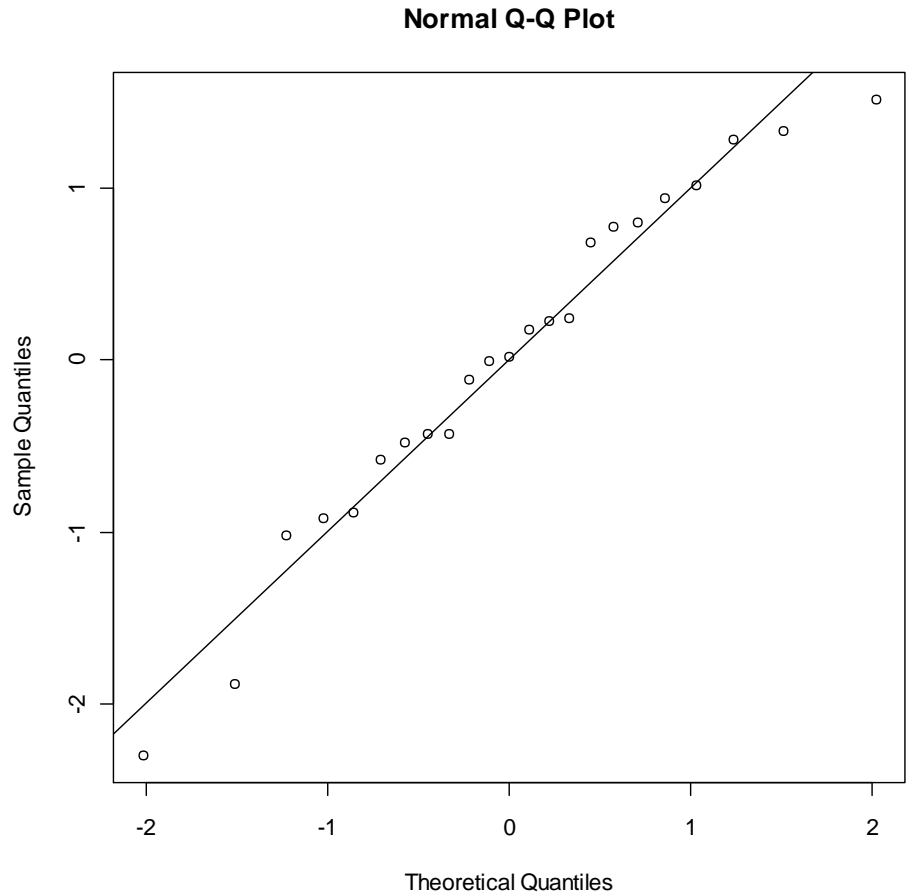
$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \sim N(0,1)$$



diagnostic

```
> mean(r.pearson)
[1] 0.01454735
> var(r.pearson)
[1] 0.05871372
> r.s<-(r.pearson-mean(r.pearson))/
  sqrt((var(r.pearson)))
> qqnorm(r.s)
> abline(0,1)
```

The variance of
pearson residual is
much smaller than 1



Models with two-way interactions

$$g(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \alpha\beta_{ij}$$

```
f2<-glm((G/Total)~H+D+S+T+T*S,family="binomial",data=habitat)
f3<-glm((G/Total)~H+D+S+T+T*H,family="binomial",data=habitat)
f4<-glm((G/Total)~H+D+S+T+T*D,family="binomial",data=habitat)
f5<-glm((G/Total)~H+D+S+T+S*H,family="binomial",data=habitat)
f6<-glm((G/Total)~H+D+S+T+S*D,family="binomial",data=habitat)
f7<-glm((G/Total)~H+D+S+T+H*D,family="binomial",data=habitat)
```

Model selection: the deviance

```
> deviance(f1)
```

```
[1] 1.541658
```

```
> deviance(f2)
```

```
[1] 1.379657
```

```
> deviance(f3)
```

```
[1] 1.327497
```

```
> deviance(f4)
```

```
[1] 1.526889
```

```
> deviance(f5)
```

```
[1] 1.518356
```

```
> deviance(f6)
```

```
[1] 1.538425
```

```
> deviance(f7)
```

```
[1] 1.364903
```

```
f3<-glm((G/Total)~H+D+S+T+T*H,  
        family="binomial"  
        ,data=habitat)
```

Model selection: AIC

```
extractAIC(f1)
[1] 6.00000 28.65782
> extractAIC(f2)
[1] 8.00000 32.57349
> extractAIC(f3)
[1] 8.00000 32.40805
> extractAIC(f4)
[1] 8.00000 32.52206
> extractAIC(f5)
[1] 7.00000 30.74255
> extractAIC(f6)
[1] 7.00000 30.77231
> extractAIC(f7)
[1] 7.00000 29.69527
```

$$g(\pi_{ijkl}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l$$

The main effect model is the model with the smallest AIC.