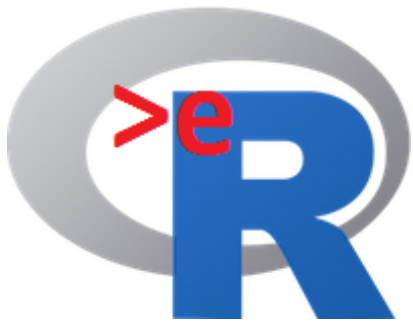This course was developed as a part of the VLIR-UOS Cross-Cutting projects:

- 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

Modelling Binary Data using R

Developed by

Nasima Akhter, Adetayo Kasim (Durham University,UK) and Ziv Shkedy (Hasselt University, belgium)

LAST UPDATE: 15/10/2017

Visit us on Facebook   ER-BioStat

GitHub   https://github.com/eR-Biostat
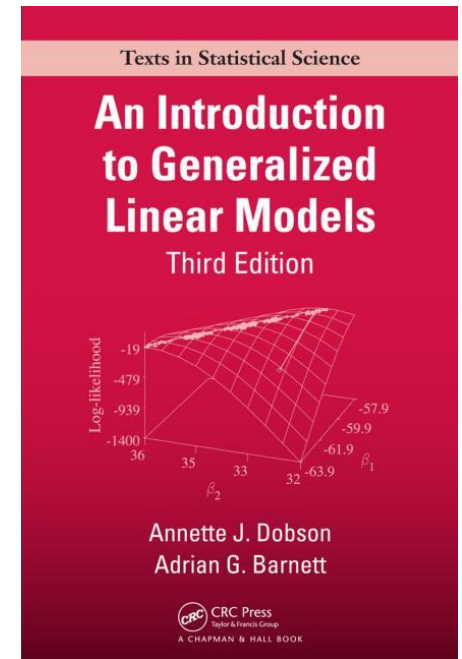
Email: erbiostat@gmail.com

twitter   @erbiostat

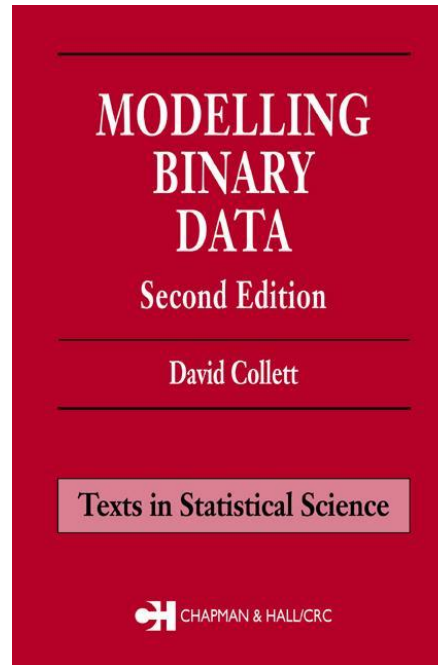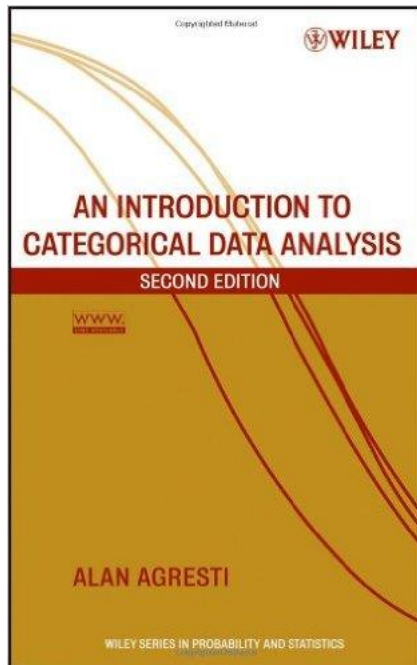# Outline

- Introduction

- Analysis of *2x2* contingency tables

- Analysis of I *x J* contingency tables

- GLMs

- Basic introduction to logistic regression

- Modelling binary data

- Multiple logistic regression

References

An Introduction to Categorical Data Analysis, Second Edition. Alan Agresti. WILEY. Wiley Series in Probability and Statistics.

Modelling Binary Data, Second Edition. David Collett. Texts in Statistical Science. CHAPMAN & HALL/CRC.

An Introduction to Generalized Linear Models, Third Edition. Annette J. Dobson, Adrian G. Barnett. Texts in Statistical Science. CRC Press, Taylor & Francis Group. A CHAPMAN & HALL BOOK.

Software

- R packages and functions:
  - `glm()`
  - `Prop.test()`
  - `bstat`

Course materials

- Slides.
- R programs.

Course materials: YouTube turorials

- Relative Risk, Odds Ratio and Risk Difference (aka Attributable Risk) (hosted by mark Marin): https://www.youtube.com/watch?v=V_YNPQoAyCc
- How to perform a Chi-Square test for Independence in R (hosted by thatRnerd): https://www.youtube.com/watch?v=i-pRb7dNakE
- Chi-Square Test, Fishers Exact Test, and Cross Tabulations in R (hosted by mark Marin): https://www.youtube.com/watch?v=POiHEJqmiC0
- Running a Chi-Squared Test of Independence in RStudio (hosted by UTSSC): https://www.youtube.com/watch?v=LnaeG0MzQVw
- Logistic Regression using R | Data Science | Machine Learning (hosted by Analytics University): https://www.youtube.com/watch?v=nubin7hq4-s
- Logistic Regression in RStudio (hosted by Tom Sherratt): https://www.youtube.com/watch?v=iyU2CkHrfQk
- Fitting a Logistic Regression Model in R (hosted by Jeff Hamrick): https://www.youtube.com/watch?v=awTyuiE4jC8
- Understanding the Summary Output for a Logistic Regression in R (hosted by Jeff Hamrick):https://www.youtube.com/watch?v=xl5dZo_BSJk

E-learning system using R

Biostatistics

Part 1

Introduction

Introduction

- In health, education, medical and social sciences, we frequently deal with dichotomous or binary outcomes.

- For example, we may have data on presence (Yes) or absence (No) of an event. For example; presence or absence of :

  ➢ Anaemia

  ➢ Ebola

  ➢ Diabetes

Introduction

- Binary data are often described by the occurrence of an event relative to the total number of trials.

- For example, suppose 20 males and 40 females students registered for a stats workshop.

Introduction

- The total sample is 60 which is the sum of the two possible outcomes (Male or Female)

- The proportion of male students is 20 out of 60 i.e 0.333

- The proportion of female students is 40 out of 60 i.e 0.667

- The sum of the proportion for two mutually exclusive outcomes should sum to 1.

Introduction

- Let Y represents the two possible outcome from an event

$$Y = \begin{cases} 1 \text{ if the outcome is postive/success} \\ \\ 0 \text{ if the outcome is negative/failure} \end{cases}$$

- Let $p = P(Y = 1)$ be the probability of success

- Let $(1 - p) = P(Y = 0)$ be the probability of failure

Introduction

- **Distribution**

$$Y \sim Bernoulli(p) \qquad OR \qquad Y \sim Bern(p)$$

- **Probability function**

$$P(Y = y) = p^y(1-p)^{1-y}$$

➢ $Mean = p$

➢ $Variance = p(1-p)$

Introduction

Bernoulli distribution represents a single trial of an event. For example, tossing a coin once. However, real life events rarely occurred in singleton. They often occur as consecutive Bernoulli processes. The probability of success from a consecutive Bernoulli processes can be represented as a Binomial distribution with probability (p) and number of trials (n)

Introduction

- Let $Y_1, Y_2, \ldots, Y_N$ represent a consecutive Bernoulli process from N trials.

$$Y_i = \begin{cases} 1 \text{ if the outcome is postive/success} \\ \\ 0 \text{ if the outcome is negative/failure} \end{cases}$$

- Let $p = P(Y_i = 1)$ be the probability of success

- Let $(1 - p) = P(Y_i = 0)$ be the probability of failure

Introduction: Bernoulli distribution in R

- Let $Y_1, Y_2, \ldots, Y_5$ represent a consecutive Bernoulli process from N trials.

$$Y_i = \begin{cases} 1 \text{ if the outcome is postive/success} \\ \\ 0 \text{ if the outcome is negative/failure} \end{cases}$$

- Let $p = P(Y_i = 1) = 0.7$, be the probability of success

```
> rbinom(5,1,0.7)
[1] 1 0 1 1 1
```

Introduction: Bernoulli distribution in R

- Let $Y_1, Y_2, \ldots, Y_5$ represent a consecutive Bernoulli process from N trials.

  - Let $p = P(Y_i = 1) = 0.7$, be the probability of success

```
> rbinom(5,1,0.7)
[1] 1 1 1 1 1
> rbinom(5,1,0.7)
[1] 1 1 1 1 1
> rbinom(5,1,0.7)
[1] 1 0 1 1 1
> rbinom(5,1,0.7)
[1] 1 1 1 0 1
> rbinom(5,1,0.7)
[1] 1 1 1 1 1
> rbinom(5,1,0.7)
[1] 1 1 1 1 0
> rbinom(5,1,0.7)
[1] 0 0 1 0 1
```
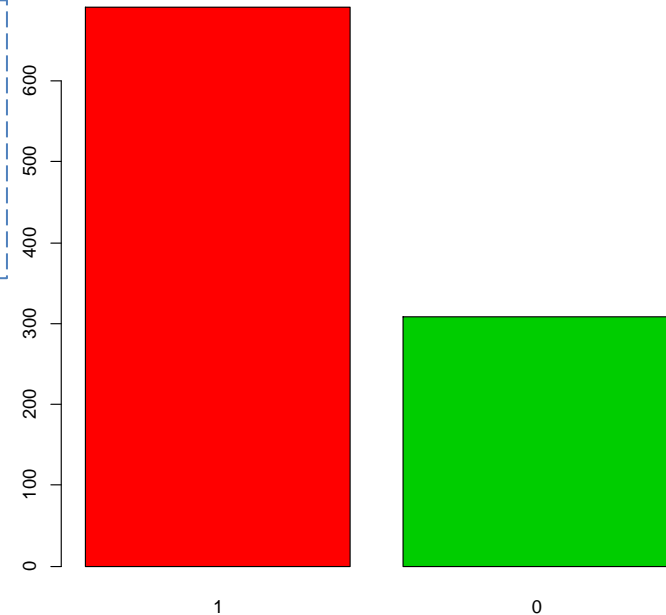
7 samples of size 5 from Bernoulli distribution with P=0.7

Introduction: Bernoulli distribution in R

- Let $Y_1, Y_2, \ldots, Y_{1000}$ represent a consecutive Bernoulli process from N trials.

  - Let $p = P(Y_i = 1) = 0.7$, be the probability of success

```
> y<-rbinom(1000,1,0.7)
> n1<-sum(y)
> n1
[1] 691
> barplot(c(n1,1000-n1),col=c(2,3))
```

Introduction

- Number of successes (Y)

$$Y = \sum_{i=1}^{N} Y_i$$

- Number of failures

$$= N - Y$$

- Note that success and failure are mutually exclusive. Both can not occur simultaneously in a single trial.

Introduction

- **Distribution of success**

$$Y \sim Binomial(N, p) \qquad \text{OR} \qquad Y \sim B(N, p)$$

- **Probability function**

$$P(Y = x) = \binom{N}{x} p^x (1-p)^{N-x}$$

➢ $Mean = Np$

➢ $Var(Y) = Np(1-p)$

Introduction

- **Mean**

$$E(Y) = Np$$

- **Variance**

$$Var(Y) = Np(1 - p)$$

➢ How is this different from Normal distribution?

➢ What is the potential problem with the parametrisation of binomial distribution?

$Y \sim B(3, 0.5)$

$E(Y) = 1.5$

$Var(Y) = 3 \times 0.5 \times 0.5$

Probability function

Y	0	1	2	3
P(Y=x)	0.125	0.375	0.375	0.125
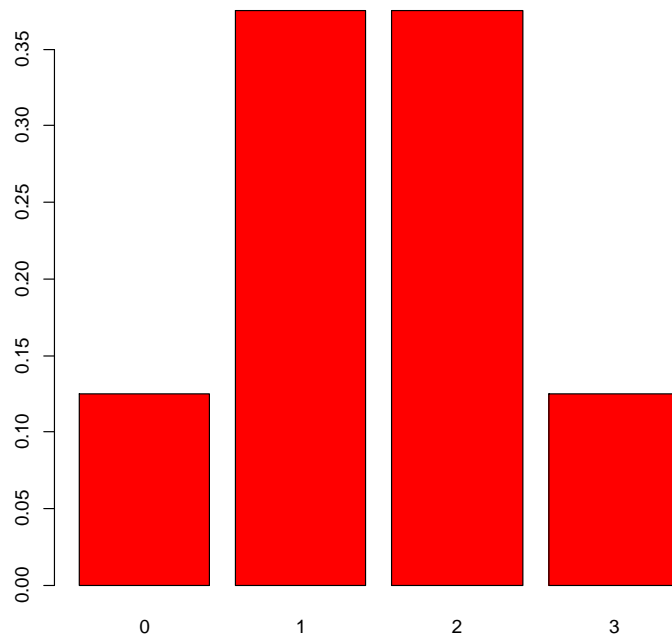
$$P(Y = 2) = \binom{3}{2} 0.5^2 (1 - 0.5)^{3-2}$$

$$P(Y = 2) = 3 \; X \; 0.5^3 = 0.375$$

$$Y \sim B(3,0.5)$$
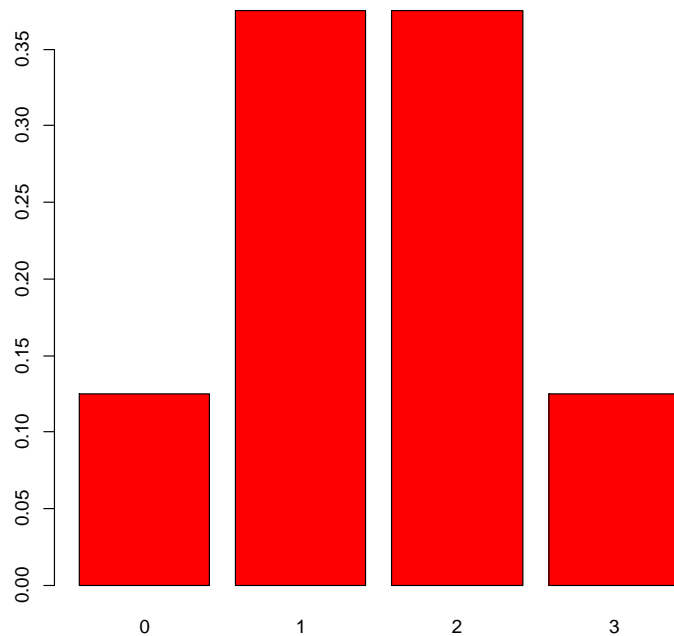
$$E(Y) = 1.5$$

$$Var(Y) = 3 \times 0.5 \times 0.5$$

1 sample from B(3,0.5)

```
> y<-rbinom(1,3,0.5)
> y
[1] 1
```

10 samples from B(3,0.5)

```
> y<-rbinom(10,3,0.5)
> y
 [1] 0 1 1 3 3 0 1 1 1 1
> mean(y)
[1] 1.2
> var(y)
[1] 1.066667
```

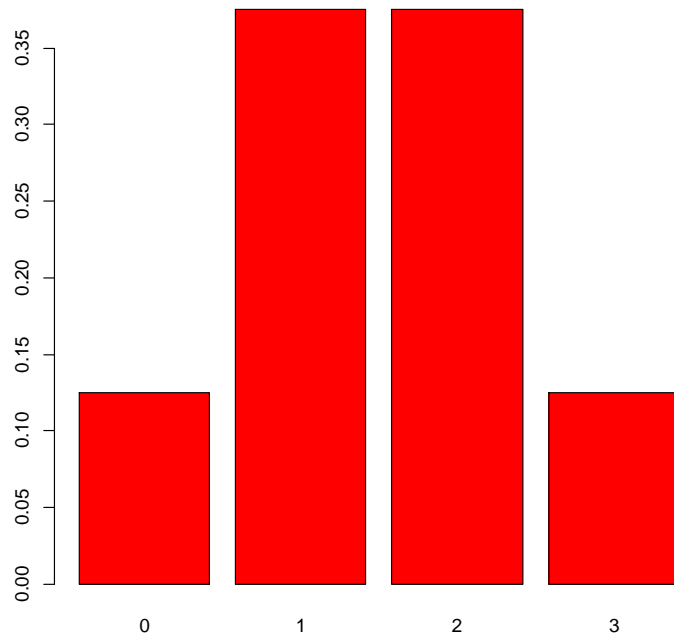$$Y \sim B(3, 0.5)$$

$$E(Y) = 1.5$$

$$Var(Y) = 3 \times 0.5 \times 0.5$$

10000 samples from B(3,0.5)

```
> y<-rbinom(10000,3,0.5)
> table(y)
y
    0    1    2    3
1289 3827 3703 1181
> mean(y)
[1] 1.4776
> var(y)
[1] 0.7435726
```


Histogram of y

24

$$Y \sim B(3, 0.5)$$
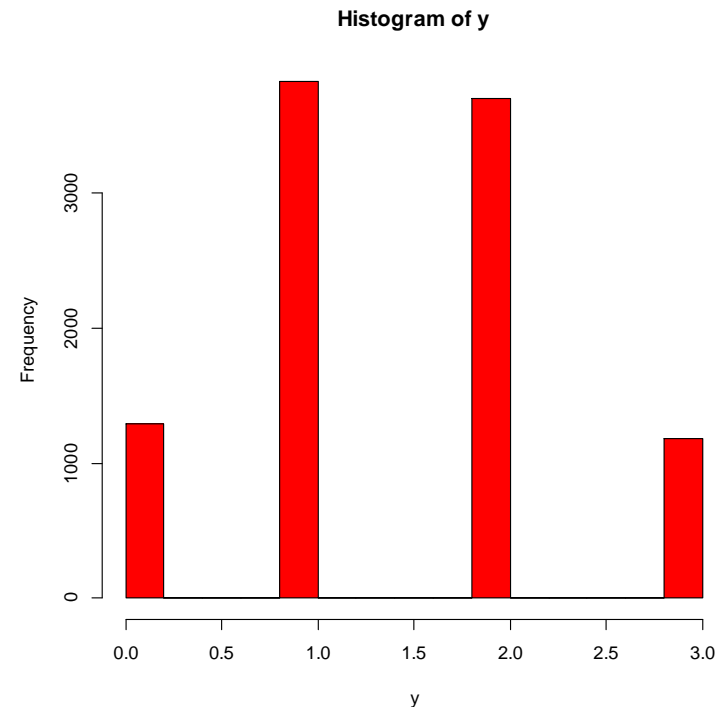
$$E(Y) = 1.5$$

$$Var(Y) = 3 \times 0.5 \times 0.5$$

1000000 samples from B(3,0.5)

```
> y<-rbinom(1000000,3,0.5)
> table(y)
y
      0       1       2       3
125339  375225  374107  125329
> mean(y)
[1] 1.499426
> var(y)
[1] 0.7513364
```
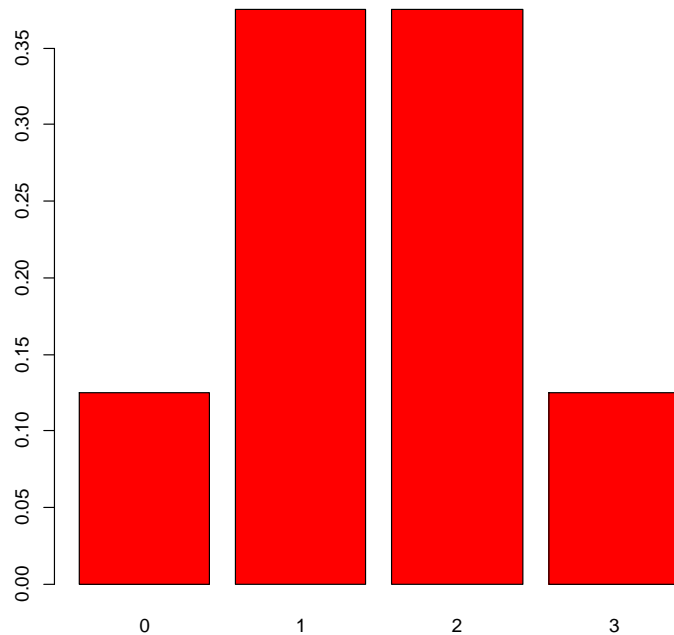

Histogram of y
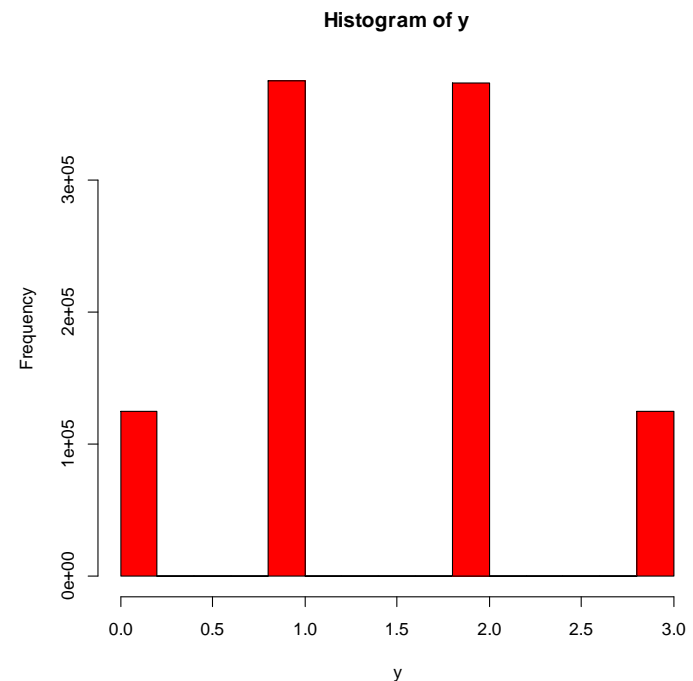
Introduction

Suppose you take a sample of 10 independent biologist to determine how many of them used valid statistical methods if the probability of using valid statistical methods is 0.8.

1. What is the probability that 10 out of 10 biologist used valid statistical methods?

2. What is the probability that none of the biologist used valid statistical methods?

3. What is the probability that 5 out of the 10 biologist used valid statistical methods?
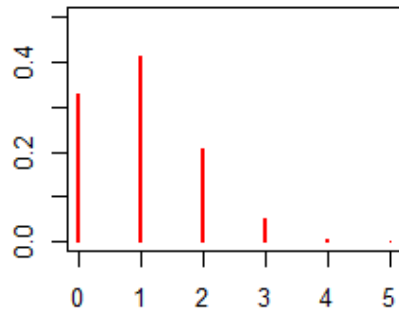
(i) N =5
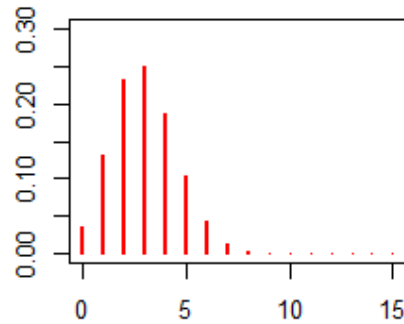
(ii) N =15
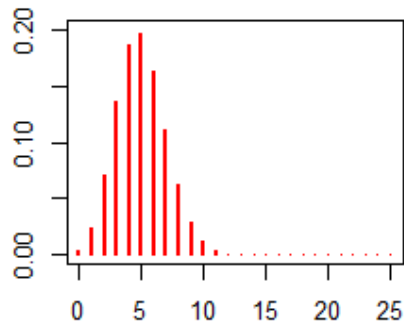
(iii) N =25
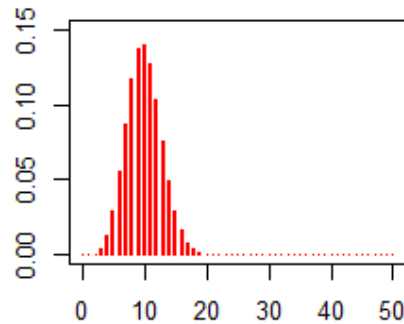
(iv) N =50

Normal approximation of Binomial distribution ($p = 0.2$)

Introduction

There other distributions for categorical data that are not covered in this course. These distributions include:

- ➢ Hypergeometric distribution

- ➢ Multinomial distribution

- ➢ Dirichlet distribution

- ➢ Negative binomial distribution

Part 2

Analysis of 2 x 2 contingency tables

2x2 Contingency table

- A contingency table / cross tab is explores the frequency distribution of an outcome(Y) .

- The table displays the frequency of an outcome variable (Y) at each level of explanatory variable X.

Gender	Anemic		Total
	Yes	No	
Male	n_{11}	n_{12}	n_{1+}
Female	n_{21}	n_{22}	n_{2+}

- The main question is whether the columns (Y) and the rows (X) are independent.

2x2 Contingency table

- General notation

Gender	Anemic		Total
	Yes	No	
Male	n_{ij}		n_{i+}
Female			
total	n_{+j}		n_{++}

- The main question is whether the columns (Y) and the rows (X) are independent.

2x2 Contingency table

- Independence in a 2X2 contingency table can be defined in three ways.

  ➢ Risk Difference (RD): Independent if the difference between the probabilities is zero.

  ➢ Relative Risk (RR): independent if the ratio of the probabilities equals one.

  ➢ Odds Ratio (OR): independent if the ratio of the odds equals one.

Risk Difference

Risk Difference

Gender	Child Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

Why is this a 2 x 2 contingency table ???

Risk Difference

Gender	Child Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

$$Risk = P\left(\text{Child Anemic}\right)$$

counts
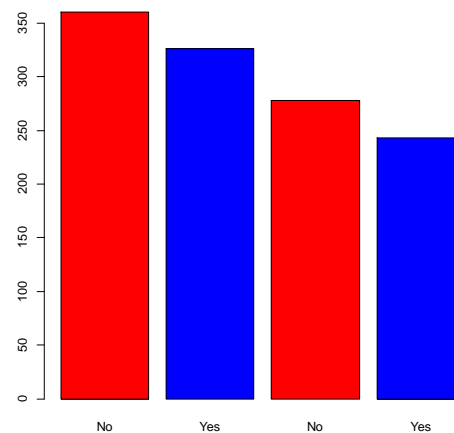
male   female

proportions
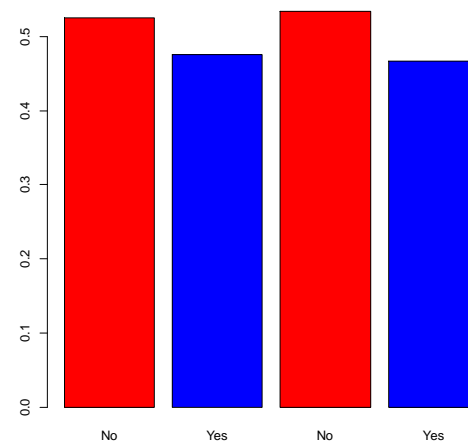
Risk Difference: inference

Gender	Child Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

$$H_0 : P_M = P_F$$
$$H_1 : P_M \neq P_F$$

$$H_0 : P_M - P_F = 0$$
$$H_1 : P_M - P_F \neq 0$$

$$\hat{P}_M \qquad \hat{P}_F$$

$$P_M = P_M \left(\text{Child Anemic} \right)$$
$$P_F = P_F \left(\text{Child Anemic} \right)$$
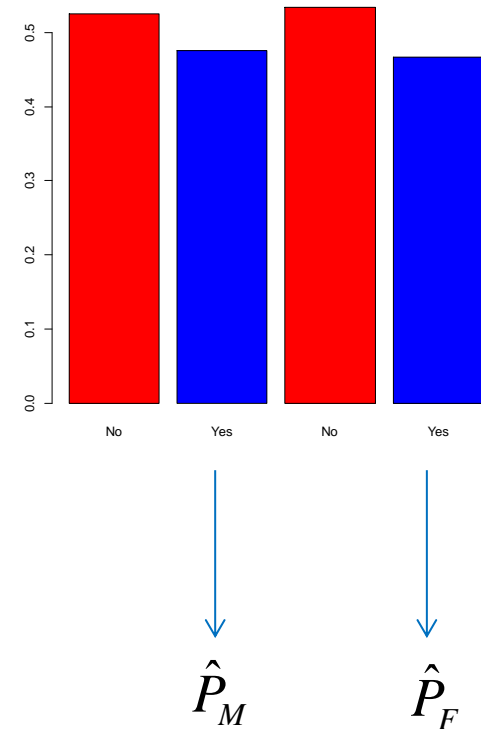
Risk Difference: estimation

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2}$

$$\widehat{p_1} = \frac{n_{11}}{n_{1+}} = \frac{326}{686} = 0.475$$

$$\widehat{p_2} = \frac{n_{21}}{n_{2+}} = \frac{243}{521} = 0.466$$

Gender	Child Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2} = 0.475 - 0.466$

$$RD = 0.009$$

Risk Difference

- $var(RD) = var(\widehat{p_1} - \widehat{p_2}) = var(\widehat{p_1}) + var(\widehat{p_2})$

$$var(\widehat{p_1}) = \frac{p_1(1 - p_1)}{n_{1+}} = \frac{0.475 * 0.525}{686} = 0.0004$$

$$var(\widehat{p_2}) = \frac{p_2(1 - p_2)}{n_{2+}} = \frac{0.466 * 0.525}{521} = 0.0005$$

- $var(RD) = var(\widehat{p_1} - \widehat{p_2}) = 0.0004 + 0.0005 = 0.0009$

- $se(RD) = \sqrt{var(RD)} = \sqrt{0.0009} = 0.03$

Risk Difference

- **Test for independence**

$$H_0: RD = 0 \qquad \text{VS} \qquad H_1: RD < 0$$

OR

$$H_0: RD = 0 \qquad \text{VS} \qquad H_1: RD > 0$$

OR

$$H_0: RD = 0 \qquad \text{VS} \qquad H_1: RD \neq 0$$

Risk Difference

- Test for independence

$$Z = \frac{p_{1-}\, p_2}{\sqrt{\dfrac{p_1(1-p_1)}{n_{1+}} + \dfrac{p_2(1-p_2)}{n_{2+}}}} = \frac{0.009}{0.03} = 0.3$$

- Two sided test, $\alpha = 0.05$, p-value= 0.7580

- Note that Z is approximated with standard Normal distribution $N(0,1)$

Risk Difference

- Confidence interval

$$P\left[-Z_{\frac{\alpha}{2}} \leq \frac{\widehat{p_1} - \widehat{p_2} - (p_1 - p_2)}{se(\widehat{p_1} - \widehat{p_2})} \leq Z_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

- The 95% confidence interval is

$$RD \pm 1.96 * se(RD)$$

$$= 0.009 \pm 1.96 * 0.03$$

$$= (-0.0498; \; 0.0678)$$

Risk Difference: Example 1

- **Read Data into R**

anemic <- read.csv("H:/D/WASP 1/2015/Data/Data for case study/Child

anaemia.csv", header=TRUE)

- **Check the Data in R using:**

  ➢ fix(anemic)

  ➢ head(anemic)

Risk Difference

Risk Difference

- **Construct 2 x2 Contingency table**

genderAnemic <- table(anaemic$Child_Gender, anaemic$Child_Anemic)

genderAnemic

- **Check the output:**

	No	Yes
Boy	360	326
Girl	278	243

Risk Difference

- **Calculate RD and test for significant results**

RDanemic <- prop.test(x=genderAnemic[,2], n=rowSums(genderAnemic),

      correct = FALSE)

RD <- round(- diff(RDanemic$estimate), 3)

RDCI <- round(RDanemic$"conf.int", 3)

RDpvalue <- round(RDanemic$"p.value", 4)

Risk Difference

- **Calculate RD and test for significant results**

RDanemic <- prop.test(x=genderAnemic[,2], n=rowSums(genderAnemic),

      correct = FALSE)

```
> RDanemic

        2-sample test for equality of proportions without continuity
        correction

data:  genderAnemic[, 2] out of rowSums(genderAnemic)
X-squared = 0.0922, df = 1, p-value = 0.7614
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.04803838  0.06565421
sample estimates:
   prop 1    prop 2
0.4752187 0.4664107
```

Risk Difference

- **Results**

> RD = 0.009
>
> 95% Confidence interval = $(-0.0498; \ 0.0678)$
>
> P value = 0.7614 **???**

- **Interpretation**

There is no significant association between child gender and anaemia.

There was a 0.9% difference in the probability of anemia between the gender.

Risk Difference: example 2

- Suppose we are interested in investigating whether younger children were more prone to anaemia than the older children.

  ➢ We need to create a contingency or a cross tabulation table with the outcome variable (Anaemia) on the columns and the explanatory variable (Age category of children) on the rows.

Risk Difference: example 2

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

Risk Difference

- **Construct 2 x2 Contingency table**

ageAnemic <- table(anemic$Child_Agecat, anaemic$Child_Anemic)

ageAnemic

- **Check the output:**

	No	Yes
24-59 months	469	310
6-23 months	169	259

Risk Difference: formulation of the hypotheses

- **A 2 x2 Contingency table**

	No	Yes
24-59 months	469	310
6-23 months	169	259

$$H_0 : P_{6-23} - P_{24-59} = 0$$
$$H_1 : P_{6-23} - P_{24-59} \neq 0$$

$$H_0 : Risk_{6-23} - Risk_{24-59} = 0$$
$$H_1 : Risk_{6-23} - Risk_{24-59} \neq 0$$

$$H_0 : RD = 0$$
$$H_1 : RD \neq 0$$

Risk Difference

- **Calculate RD and test for significant results**

RDanemic <- prop.test(x=ageAnemic[,2], n= rowSums(ageAnemic) , correct

= FALSE)

```
> RDanemic

        2-sample test for equality of proportions without continuity
        correction

data:  ageAnemic[, 2] out of rowSums(ageAnemic)
X-squared = 47.5894, df = 1, p-value = 5.255e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2648663 -0.1495219
sample estimates:
   prop 1    prop 2
0.3979461 0.6051402
```

Risk Difference

- **Calculate RD and test for significant results**

RDanemic <- prop.test(x=ageAnemic[,2], n= rowSums(ageAnemic) , correct

= FALSE)

RD <- round(- diff(RDanemic$estimate),3)

RDCI <- round(RDanemic$"conf.int",3)

RDpvalue <- round(RDanemic$"p.value",4)

Risk Difference

- **Results**

> RD = -0.207
>
> 95% Confidence interval = (-0.265; -0.150)
>
> P value = < 0.0001

- **Interpretation**

There is a significant association between child age and anaemia. Younger children have 20.7% more risk of anaemia than younger children.

Relative Risk (RR)

Relative Risk (RR)

Suppose we want to estimate Relative Risk (RR) for occurrence of anaemia among boys and girls.

Gender	Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

$$RR = \frac{Risk_M}{Risk_F}$$

Relative Risk (RR)

- $Relative\ Risk\ (RR) = \dfrac{\hat{p}_1}{\hat{p}_2}$

$$\hat{p}_1 = \frac{n_{11}}{n_{1+}} = \frac{326}{686} = 0.475$$

$$\hat{p}_2 = \frac{n_{21}}{n_{2+}} = \frac{243}{521} = 0.466$$

- $Relative\ Risk(RR) = \dfrac{\hat{p}_1}{\hat{p}_2} = \dfrac{0.475}{0.466}$

$$RR = 1.02$$

Relative Risk (RR)

$$RR = \frac{Risk_M}{Risk_F} = 1 \Rightarrow Risk_M = Risk_F$$

$$\log(RR) = \log\left(\frac{Risk_M}{Risk_F}\right) = \log(Risk_M) - \log(Risk_F)$$

$$RR = 1 \Rightarrow \log(RR) = 0 \Leftrightarrow Risk_M = Risk_F$$

Relative Risk (RR)

- Log transform RR to convert it to a linear scale

$$log(RR) = \log(\hat{p}_1) - \log(\hat{p}_2)$$

$$\log(\hat{p}_1) = \log(0.475) = -0.7444$$

$$\log(\hat{p}_2) = \log(0.466) = -0.7636$$

- $log(RR) = \log(\hat{p}_1) - \log(\hat{p}_2)$

$$\log(RR) = 0.0192$$

Relative Risk (RR)

- $var(\log(RR)) = var\left(\log\left(\frac{\hat{p}_1}{\hat{p}_2}\right)\right)$

$$= \frac{(1 - \hat{p}_1)}{\hat{p}_1 n_{1+}} + \frac{(1 - \hat{p}_2)}{\hat{p}_2 n_{2+}}$$

$$= \frac{(1 - 0.475)}{0.475 * 686} + \frac{(1 - 0.466)}{0.466 * 521}$$

- $var(\log(RR)) = 0.0038$

  ➢ $se(\log(RR)) = \sqrt{0.0038} = 0.06$

Relative Risk (RR)

- Test for independence

$$H_0: \log(RR) = 0 \qquad \text{vs} \qquad H_1: \log(RR) < 0$$

OR

$$H_0: \log(RR) = 0 \quad \text{vs} \qquad H_1: \log(RR) > 0$$

OR

$$H_0: \log(RR) = 0 \quad \text{vs} \qquad H_1: \log(RR) \neq 0$$

Relative Risk (RR)

- Test for independence

$$Z = \frac{\log(\hat{p}_1) - \log(\hat{p}_2)}{\sqrt{\dfrac{(1 - \hat{p}_1)}{\hat{p}_1 n_{1+}} + \dfrac{(1 - \hat{p}_2)}{\hat{p}_2 n_{2+}}}} = \frac{0.0192}{0.06} = 0.32$$

- Two sided test, $\alpha = 0.05$, p-value= 0.7517

- Note that Z is approximated with standard Normal distribution $N(0,1)$

Relative Risk (RR)

- Test for independence

$$P\left[-Z_{\frac{\alpha}{2}} \leq \frac{\log(\widehat{p_1}) - \log(\widehat{p_2}) - (\log(p_1) - \log(p_2))}{se(\log(\widehat{p_1}) - \log(\widehat{p_2}))}\right] = 1 - \alpha$$

- The 95% confidence interval for log(RR) is

$$\log(RR) \pm 1.96 * se(\log(RR))$$

$$= 0.0192 \pm 1.96 * 0.06$$

$$= (-0.0984; 0.1368)$$

- The 95% confidence interval for RR is

$$= (0.91; 1.15) \; ???$$

Relative Risk (RR) in R

- **Construct 2 x2 Contingency table**

genderAnemic <- table(anaemic$Child_Gender, anaemic$Child_Anemic)

genderAnemic

- **Check the output:**

	No	Yes
Boy	360	326
Girl	278	243

Relative Risk (RR)

- **Calculate RD and test for significant results**

```
##install.packages("bstats")

library(bstats)

RRanemic <- oddsratio(x=genderAnemic[,2], n=rowSums(genderAnemic))
```

```
Data:
         Event Size
Sample 1   326  686
Sample 2   243  521

Relative risk:   1.018884
 95 % confidence intervals
                  LL        UL
transform  2.1440305 2.144030
Asymptotic 0.9106626 1.139967
Score      0.9036104 1.151010
```

Relative Risk (RR)

- **Calculate RD and test for significant results**

```
##install.packages("bstats")

library(bstats)

RRanemic <- oddsratio(x=genderAnemic[,2], n=rowSums(genderAnemic))

RR <- round(RRanemic$RR,3)

RRCI <- round(RRanemic$RRCI,3)
```

Relative Risk (RR)

- **Results**

> ➢ RR = 1.019
>
> ➢ 95% Confidence interval = (0.911; 1.140)

- **Interpretation**

There is no significant association between child gender and anaemia.

A male child has 1.9% more risk of anaemia than a female child.

Relative Risk (RR): example 2

- Suppose we are interested in investigating whether younger children were more prone to anaemia than the older children.

  - ➢ We need to create a contingency or a cross tabulation table with the outcome variable (Anaemia) on the columns and the explanatory variable (Age category of children) on the rows.

Relative Risk (RR)

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

$$RR = \frac{Risk_{6-23}}{Risk_{24-59}} = \frac{P_{6-23}(Amenic)}{P_{24-59}(Amenic)}$$

Relative Risk (RR)

- **Construct 2 x2 Contingency table**

ageAnemic <- table(anemic$Child_Agecat, anaemic$Child_Anemic)

ageAnemic

- **Check the output:**

```
> ageAnemic

                 No  Yes
  24-59 months  469  310
  6-23 months   169  259
```

Relative Risk (RR)

- **Calculate RD and test for significant results**

RRanemic <- oddsratio(x=ageAnemic[,2], n=rowSums(ageAnemic))

```
> RRanemic

Data:
        Event Size
Sample 1   310  779
Sample 2   259  428


Relative risk:   0.6576097
 95 % confidence intervals
                    LL          UL
transform  1.6525086 1.6525086
Asymptotic 0.5950998 0.7266858
Score      0.5862009 0.7385974
```

Relative Risk (RR)

- **Calculate RD and test for significant results**

RRanemic <- oddsratio(x=ageAnemic[,2], n=rowSums(ageAnemic))

RR <- round(RRanemic$RR,3)

RRCI <- round(RRanemic$RRCI,3)

Relative Risk (RR)

- **Results**

> - RR = 0.658
>
> - 95% Confidence interval = (0.595; 0.727)

- **Interpretation**

There is a significant association between child age and anaemia. The risk of anaemia for younger children is 34.2% less than the risk of anaemia for younger children.

Relative Risk (RR)

- **Results**

> - RR = 0.658
>
> - 95% Confidence interval = (0.595; 0.727)
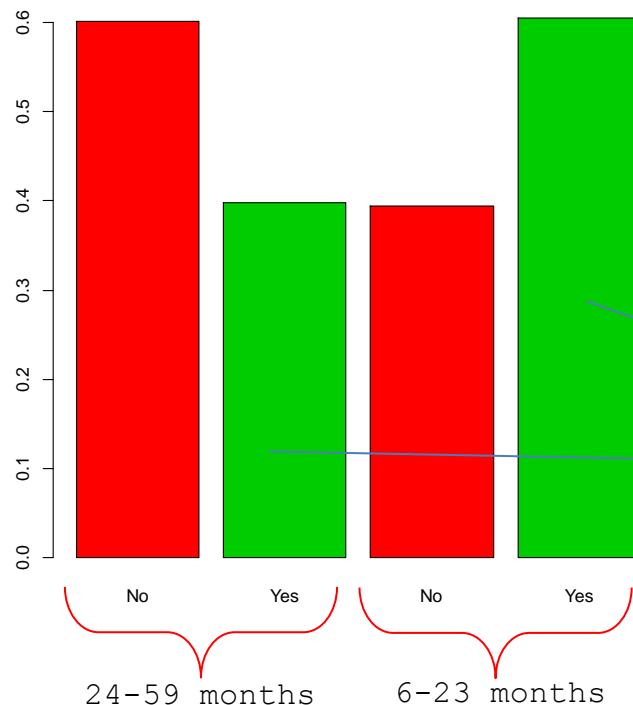

```
> ageAnemic

              No Yes
24-59 months  469 310
6-23 months   169 259
```

RR

Odds Ratio

Odds Ratio (OR)

- Both Risk Difference (RD) and Relative Risk (RR) are based on the proportion of "success" (Yes). They do not directly account for the proportion of "failures"

- Odds explicitly account for both success and failures. It is a ratio of successes to failures. We can calculate the Odds of anaemic separately for male and female children.

Odds Ratio (OR)

Gender	Anemic		Total
	Yes	No	
Male	n_{11}	n_{12}	n_{1+}
Female	n_{21}	n_{22}	n_{2+}

- The main question is whether the columns (Y) and the rows (X) are independent.

Odds Ratio (OR)

- We can Odds Ratio in terms of cell counts

➢ Odds(Male) $= \dfrac{n_{11}}{n_{12}}$

➢ Odds (Female) $= \dfrac{n_{21}}{n_{22}}$

➢ OR $= \dfrac{\text{Odds (Male)}}{\text{Odds (Female)}} = \left(\dfrac{n_{11}}{n_{12}}\right) / \left(\dfrac{n_{21}}{n_{22}}\right)$

$$OR = \dfrac{n_{11} * n_{22}}{n_{12} * n_{21}}$$

Odds Ratio (OR)

Gender	Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

$$\text{OR} = \frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{326 * 278}{360 * 243} = 1.04$$

Odds Ratio (OR)

➢ Odds(Male) $= \dfrac{326}{360} = 0.905$

➢ Odds (Female) $= \dfrac{243}{278} = 0.874$

$$OR = \frac{\text{Odds (Male)}}{\text{Odds (Female)}}$$

$$= \frac{0.905}{0.874} = 1.04$$

Odds Ratio (OR)

- Odds and Odd ratio can also be calculated from probabilities.

  $\triangleright$ Odds(Male) $= \dfrac{\hat{p}_1}{1-\hat{p}_1}$

  $\triangleright$ Odds (Female) $= \dfrac{\hat{p}_2}{1-\hat{p}_2}$

- OR $= \dfrac{\text{Odds (Male)}}{\text{Odds (Female)}} = \left(\dfrac{\hat{p}_1}{1-\hat{p}_1}\right) / \left(\dfrac{\hat{p}_2}{1-\hat{p}_2}\right)$

$$OR = \dfrac{\hat{p}_1 * (1-\hat{p}_2)}{\hat{p}_2 * (1-\hat{p}_1)}$$

Odds Ratio (OR)

$$OR = \frac{\hat{p}_1 * (1 - \hat{p}_2)}{\hat{p}_2 * (1 - \hat{p}_1)}$$

$$OR\ (Anemic) = \frac{0.475(1 - 0.466)}{0.466(1 - 0.475)}$$

$$OR\ (Anemic) = 1.04$$

Odds Ratio (OR)

- Hypothesis Testing

  ➢ Similar to relative risk (RR), hypothesis testing for Odds Ratio (OR) is difficult on its original scale.

  ➢ Instead the hypothesis testing is often performed on log scale.

Odds Ratio (OR)

- Hypothesis Testing

  - $log(OR) = \log(p1/1 - p1) - \log(p2/1 - p2)$

  - $log(OR) = logit(p1) - logit(p2)$

- Logit is used to transform odds of an event from a bounded range of $0\ to\ 1$ to a continuous scale that can take negative and positive values

Odds Ratio (OR)

$$Var\big(log(OR)\big) = \left[\frac{1}{n_{11}} + \frac{1}{n_{12}}\right] + \left[\frac{1}{n_{21}} + \frac{1}{n_{22}}\right]$$

$$se\big(log(OR)\big) = \sqrt{\left[\frac{1}{n_{11}} + \frac{1}{n_{12}}\right] + \left[\frac{1}{n_{21}} + \frac{1}{n_{22}}\right]}$$

Odds Ratio (OR)

- Test for independence

$$H_0: \log(OR) = 0 \quad \text{VS} \quad H_1: \log(OR) < 0$$

OR

$$H_0: \log(OR) = 0 \quad \text{VS} \quad H_1: \log(OR) > 0$$

OR

$$H_0: \log(OR) = 0 \quad \text{VS} \quad H_1: \log(OR) \neq 0$$

Odds Ratio (OR)

- Test for independence

$$Z = \frac{\text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_2)}{\sqrt{\left[\frac{1}{n_{11}} + \frac{1}{n_{12}}\right] + \left[\frac{1}{n_{21}} + \frac{1}{n_{22}}\right]}} = \frac{0.0361}{0.12} = 0.30$$

- Two sided test, $\alpha = 0.05$, p-value= 0.7580

- Note that Z is approximated with standard Normal distribution $N(0,1)$

Odds Ratio (OR)

- Test for independence

$$P\left[-Z_{\frac{\alpha}{2}} \leq \frac{\text{logit}(\widehat{p_1}) - \text{logit}(\widehat{p_2})}{\sqrt{\left[\frac{1}{n_{11}} + \frac{1}{n_{12}}\right] + \left[\frac{1}{n_{21}} + \frac{1}{n_{22}}\right]}}\right] = 1 - \alpha$$

- The 95% confidence interval for log(RR) is

$$\log(OR) \pm 1.96 * se(\log(OR))$$

$$= 0.0361 \pm 1.96 * 0.12$$

$$= (-0.1991; 0.2713)$$

- The 95% confidence interval for RR is

$$= (0.82; 1.31) \; ???$$

Odds Ratio (OR)

- To convert the estimated Log(OR) and its confidence intervals to the same scale as OR, used the following:

    ➢ $OR = exp^{(\log(OR))}$

    ➢ $95\% \; CI \; for \; OR = exp^{(95\% \; CI \; for \; \log(OR))}$

- Note that $95\% \; CI$ for OR is not always symmetric.

Odds Ratio (OR) in R : example 1

- **Construct 2 x2 Contingency table**

genderAnemic <- table(anaemic$Child_Gender,anaemic$Child_Anemic)

genderAnemic

- **Check the output:**

	No	Yes
Boy	360	326
Girl	278	243

Odds Ratio (OR)

- **Calculate RD and test for significant results**

library(bstats)

ORanemic <- oddsratio(x=genderAnemic[,2], =rowSums(genderAnemic))

```
Data:
        Event Size
Sample 1   326   686
Sample 2   243   521

Odds ratio:       1.035985
 95 % confidence intervals
                      LL              UL
Asymptotic 8.245977e-01 1.301563e+00
Exact      1.000000e+06 1.000000e+06
Score      8.245721e-01 1.301603e+00
```

Odds Ratio (OR)

- **Calculate RD and test for significant results**

library(bstats)

ORanemic <- oddsratio(x=genderAnemic[,2], =rowSums(genderAnemic))

OR <- round(ORanemic $OR,3)

ORCI <- round(ORanemic $ORCI,3)

Odds Ratio (OR)

- **Results**

  - RR = 1.036

  - 95% Confidence interval = (0.84; 1.267)

- **Interpretation**

  There is no significant association between child gender and anaemia. The odd of anaemia for a male child is 3.6% higher than the odd of anaemia for a female child.

Odds Ratio (OR): example 2

- Suppose we are interested in investigating whether younger children were more prone to anaemia than the older children.

  ➢ We need to create a contingency or a cross tabulation table with the outcome variable (Anaemia) on the columns and the explanatory variable (Age category of children) on the rows.

Odds Ratio (OR)

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

$$OR = \frac{259 \times 469}{169 \times 310} = 0.431$$
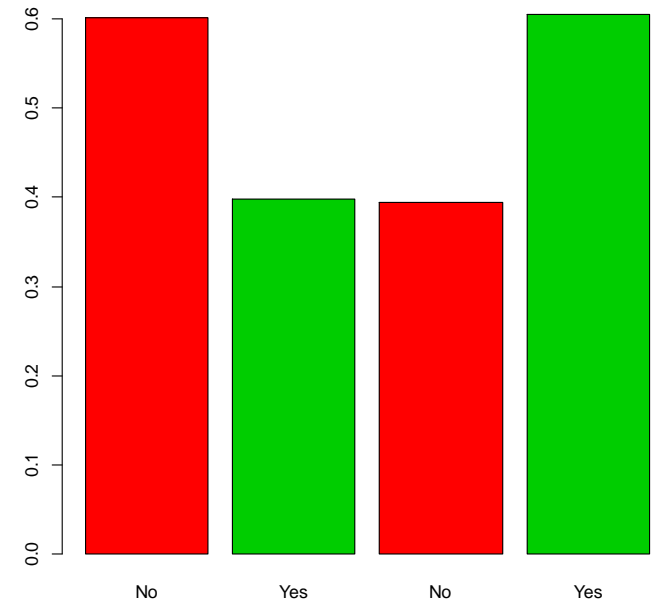
Odds Ratio (OR)

- **Construct 2 x2 Contingency table**

ageAnemic <- table(anemic$Child_Agecat, anaemic$Child_Anemic)

ageAnemic

- **Check the output:**

	No	Yes
24-59 months	469	310
6-23 months	169	259

Odds Ratio (OR) in R

- **Calculate RD and test for significant results**

```
ORanemic <- oddsratio(x=ageAnemic[,2], n=rowSums(ageAnemic))
```

```
Data:
          Event Size
Sample 1   310  779
Sample 2   259  428

Odds ratio:      0.4312964
 95 % confidence intervals
                    LL            UL
Asymptotic 3.388857e-01 5.489065e-01
Exact      1.000000e+06 1.000000e+06
Score      3.388502e-01 5.489639e-01
```

Odds Ratio (OR)

- **Calculate RD and test for significant results**

ORanemic <- oddsratio(x=ageAnemic[,2], n=rowSums(ageAnemic))

OR <- round(ORanemic$OR,3)

ORCI <- round(ORanemic$ORCI,3)

```
> OR <- round(ORanemic$OR,3)
> OR
[1] 0.431
> ORCI <- round(ORanemic$ORCI,3)
> ORCI
                   LL        UL
Asymptotic 3.39e-01 5.49e-01
Exact      1.00e+06 1.00e+06
Score      3.39e-01 5.49e-01
```

Odds Ratio (OR)

- **Results**

> OR = 0.431
>
> 95% Confidence interval = (0.352; 0.528)

- **Interpretation**

There is a significant association between child age and anaemia. The odds of anaemia for older children is 43.1% of the odds of anaemia for younger children

Summary

- Risk difference is the measure of absolute difference in observed disease/ event/ exposure between two groups

- Same difference may mean different thing for different event. Clinical importance of the risk difference has to be judged based on the context.

- RD= 0 mean that the estimated effects are independent

- $-\infty \leq RD \leq +\infty$

Summary

- $RR = 1$ means that the estimated effects are independent

- $RR > 1$ = success probabilities are higher in the intervention group than the comparison group.

- $RR < 1$ = success probability of an event is less in the intervention group than the comparison group.

- $\mathbf{0 \leq RR \leq +\infty; -\infty \leq log(RR) \leq +\infty}$

Summary

- OR = 1 corresponds to independence.

- OR >1 means the odd of an event is higher in group 1 than in group 2.

- OR <1 means that the odd of event is smaller in group 1 than in group 2.

- $0 \leq OR \leq +\infty; -\infty \leq log(OR) \leq +\infty$

Summary

- **Relationship between RR and OR.**

$$OR = \frac{P_1(1-P_2)}{P_2(1-P_1)} \quad ; \quad RR = \frac{P_1}{P_2}$$

$$OR = \frac{P_1}{P_2} * \frac{(1-P_2)}{(1-P_1)} \quad ; \quad OR = RR * \frac{(1-P_2)}{(1-P_1)}$$

$$OR = RR \; iff \; \frac{(1-P_2)}{(1-P_1)} \cong 1 \quad \text{i.e for a rare event}$$

Practical I

- **Using the same dataset (Child anaemia.csv)**

  ➢ Investigate whether there is association between the likelihood of child anaemia and mother anaemia using risk difference (RD), relative risk (RR) and odds ratio (OR)

  ➢ Interpret your results.

Chi-squared test for independence

Analysis of I x J contingency tables

Analysis of IxJ contingency tables

- The main goal of analysing a contingency table is to test independence between rows and columns.

- In our case study, the null hypothesis is that there is no association between anaemia prevalence and socio-economic status. Therefore, the distribution of outcome categories should be independent of the explanatory variable

Analysis of IxJ contingency tables

- **2 x 2 contingency table**

Explanatory	Outcome		Total
	Yes	No	
A	n_{11}	n_{12}	n_{1+}
B	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

- **2 x 2 contingency table**

Explanatory	Outcome		Total
	Yes	No	
A	n_{ij}		n_{i+}
B			
Total	n_{+j}		n_{++}

Analysis of IxJ contingency tables

- **4 x 2 contingency table**

Explanatory	Outcome		Total
	Yes	No	
A	n_{11}	n_{12}	n_{+1}
B	n_{21}	n_{22}	n_{+2}
C	n_{31}	n_{32}	n_{+3}
D	n_{41}	n_{42}	n_{+4}
Total	n_{+1}	n_{+2}	n_{++}

- **4 x 3 contingency table**

Explanatory	Outcome			Total
	Large	Medium	Small	
A	n_{11}	n_{12}	n_{13}	n_{1+}
B	n_{21}	n_{22}	n_{23}	n_{2+}
C	n_{31}	n_{32}	n_{33}	n_{3+}
D	n_{41}	n_{42}	n_{43}	n_{4+}
Total	n_{+1}	n_{+2}	n_{+3}	n_{++}

Analysis of IxJ contingency tables

- Independence test in a generalised two-way contingency tables of **nominal** outcomes can be tested using;

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j}$$

$$\pi_{ij} = \frac{n_{ij}}{n_{++}} \qquad ; \qquad \pi_{i+} = \frac{n_{i+}}{n_{++}} \qquad ; \qquad \pi_{+j} = \frac{n_{j+}}{n_{++}}$$

- If the independent assumptions holds, then the distribution of the cell counts is independent of the rows and the columns.

Probability under independence

$$H_0: \pi_{ij} = \pi_{i+}\pi_{+j}$$

- For two independent events:

$$P(A \cap B) = P(A) \times P(B)$$

- In a I X J table :

$$P(X = i \cap Y = j) = P(X = i) \times P(Y = j)$$

$$\pi_{ij} = \pi_{i+} \times \pi_{+j}$$

Analysis of IxJ contingency tables

- Under the null model we can calculate the expected cell frequencies ($\hat{\mu}_{ij}$) as:

$$n_{++} \times \hat{\pi}_{ij} = n_{++} \times \left(\hat{\pi}_{j+} \hat{\pi}_{+j} \right) = n_{++} \times \frac{n_{i+}}{n_{++}} \times \frac{n_{+j}}{n_{++}} \quad \Rightarrow \quad \hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}$$

- We can use $Chi - square\ test$ to compare the expected frequencies under the null model with the observed frequencies:

Analysis of IxJ contingency tables

- **Pearson Chi-square statistics**

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}; \qquad X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

➤ O_{ij} = observed cell counts for row i and column j

➤ E_{ij} = Expected cell counts for row i and column j

➤ X^2 ~Chi-Square distribution with $(I-1)(J-1)$ degree of freedom (df)

Analysis of IxJ contingency tables

- Investigate whether there is association between child location and child anaemia.

Areas	Anemic		Total
	Yes	No	
A	101	99	200
B	83	117	200
C	112	89	201
D	74	126	200
Total	370	431	801

Analysis of IxJ contingency tables

- Matrix of the observed cell counts (O_{ij})

Areas	Anemic	
	Yes	No
A	O_{11}=101	O_{12}=99
B	O_{21}=83	O_{22}=117
C	O_{31}=112	O_{32}=89
D	O_{41}=74	O_{42}=126

Analysis of IxJ contingency tables

- Matrix of the expected values (E_{ij}).

Areas	Anemic	
	Yes	No
A	$E_{11} = \dfrac{200*370}{801}$ =92.4	$E_{12} = \dfrac{200*431}{801}$ =107.6
B	$E_{21} = \dfrac{200*370}{801}$ =92.4	$E_{22} = \dfrac{200*431}{801}$ =107.6
C	$E_{31} = \dfrac{201*370}{801}$ =92.8	$E_{32} = \dfrac{201*431}{801}$ =108.2
D	$E_{41} = \dfrac{200*370}{801}$ =92.4	$E_{42} = \dfrac{200*431}{801}$ =107.6

Chi-square test in R

```
areaAnemic <- table(nonMissingAnemic$Areas, nonMissingAnemic$Child_Anemic,

        exclude=FALSE)

nplus. <- rowSums(areaAnemic)

n.plus <- colSums(areaAnemic)

npluplus <- sum(areaAnemic)

Oij <- areaAnemic

Eij <- (nplus.%*%t(n.plus))/npluplus

tmp  <- ((Oij-Eij)^2)/Eij

X2 <- sum(tmp)

df <- (nrow(areaAnemic)-1)*(ncol(areaAnemic)-1)

pvalue <- pchisq(X2, df,lower.tail = FALSE))
```

Chi-square test in R

- **Results**

  - $X^2 = 17.4$

  - Pvalue = 0.0006

- **Interpretation**

  There is a significant association between child location and child anaemia.

Chi-square test in R

- Definition of the variables

```
> Anemic<- as.factor(c(rep("Yes",101),rep("No",99),rep("Yes",83),rep("No",117)

                  ,rep("Yes",112),rep("No",89),rep("Yes",74),rep("No",126)))

> Areas<-as.factor(c(rep("A",101),rep("A",99),rep("B",83),rep("B" ,117),rep("C",112),rep("C"

      ,89),rep("D",74),rep("D" ,126)))
```

Chi-square test in R

- ## Chi-square for independence

```
> areaAnemic<-table(Anemic,Areas)

> areaAnemic

      Areas

Anemic    A    B    C    D

   No    99  117   89  126

   Yes  101   83  112   74
```

```
> chiArea <- chisq.test(areaAnemic,correct = FALSE)
> chiArea


        Pearson's Chi-squared test

data:  areaAnemic
X-squared = 17.4074, df = 3, p-value = 0.0005827
```

Example: 2 X 2 table: example 1

- Suppose we are interested in investigating whether younger children were more prone to anaemia than the older children.

  - ➢ We need to create a contingency or a cross tabulation table with the outcome variable (Anaemia) on the columns and the explanatory variable (Age category of children) on the rows.

Example: a 2 X 2 table

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

$$H_0 : P_M = P_F$$
$$H_1 : P_M \neq P_F$$

$$P_M = P_M(\text{Child Anemic})$$
$$P_F = P_F(\text{Child Anemic})$$
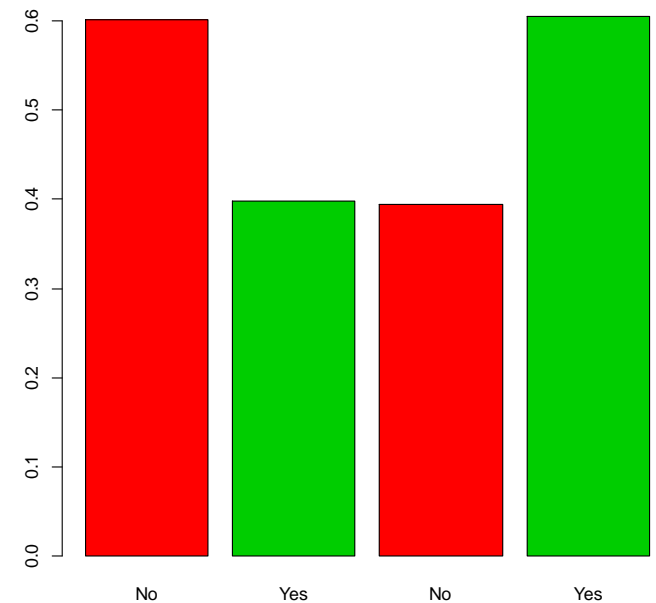
Risk Difference: estimation

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2}$

$$\widehat{p_1} = \frac{n_{11}}{n_{1+}} = \frac{259}{428} = 0.605$$

$$\widehat{p_2} = \frac{n_{21}}{n_{2+}} = \frac{310}{779} = 0.397$$

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779

- $Risk\ Difference\ (RD) = \widehat{p_1} - \widehat{p_2} = 0.605 - 0.379$

$$RD = 0.208$$

Risk Difference

- Test for independence

$$Z = \frac{p_{1-}\, p_2}{\sqrt{\dfrac{p_1(1-p_1)}{n_{1+}} + \dfrac{p_2(1-p_2)}{n_{2+}}}} = -7.041399$$

- Two sided test, $\alpha = 0.05$, p-value= <0.001

- Note that Z is approximated with standard Normal distribution $N(0,1)$

Risk Difference in R

```
> RDanemic <- prop.test(x=ageAnemic[,2],  n= rowSums(ageAnemic) ,
correct = FALSE)
>
> RDanemic


        2-sample test for equality of proportions without
continuity
        correction

data:  ageAnemic[, 2] out of rowSums(ageAnemic)
X-squared = 47.5894, df = 1, p-value = 5.255e-12
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2648663 -0.1495219
sample estimates:
   prop 1    prop 2
0.3979461 0.6051402
```
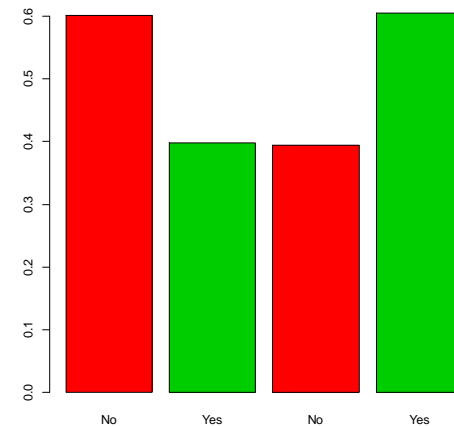
Example: a 2 X 2 table

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779


```
> Oij <- ageAnemic
> Oij

                No Yes
  24-59 months 469 310
  6-23 months  169 259
>
> nplus. <- rowSums(ageAnemic)
> nplus.
24-59 months  6-23 months
         779          428
> n.plus <- colSums(ageAnemic)
> n.plus
 No Yes
638 569
> npluplus <- sum(ageAnemic)
> npluplus
[1] 1207
```

```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
            No      Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Example: a 2 X 2 table

Age categories	Anemic		Total
	Yes	No	
6-23 months	259	169	428
24-59 months	310	469	779
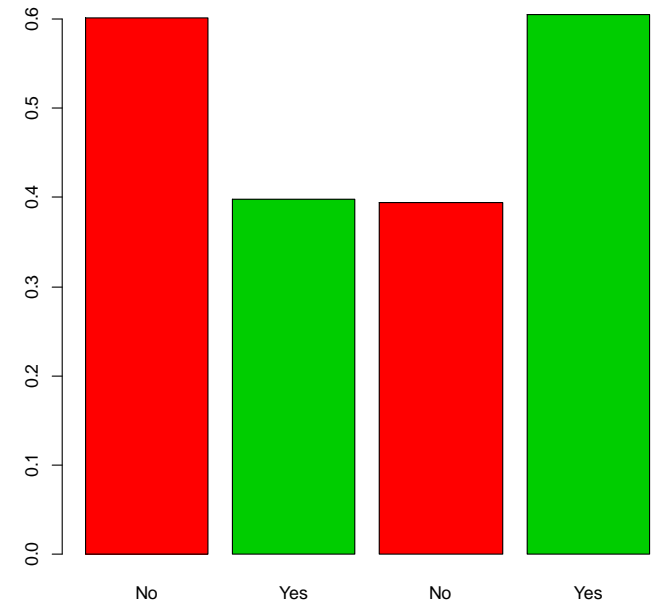

```
 chi.sq <- chisq.test(ageAnemic,correct = FALSE)
> chi.sq

        Pearson's Chi-squared test

data:  ageAnemic
X-squared = 47.5894, df = 1, p-value = 5.255e-12
```

Example: a 2 X 2 table – OR for rhe expected table

counts

proportions
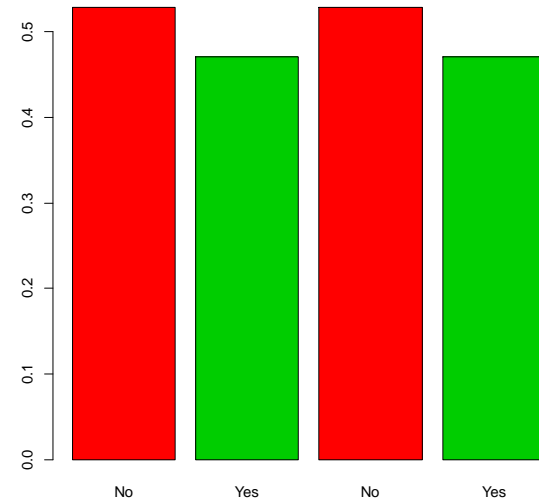

```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
          No      Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Expected value

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$
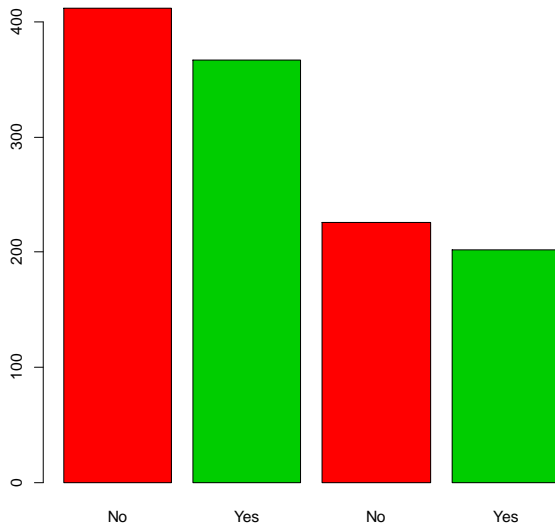
```
> Eij <- (nplus.%*%t(n.plus))/npluplus
> Eij
           No      Yes
[1,] 411.7664 367.2336
[2,] 226.2336 201.7664
>
> tmp  <- ((Oij-Eij)^2)/Eij
> X2 <- sum(tmp)
> X2
[1] 47.58941
```

Expected value

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

```
> ORanemic <- oddsratio(x=Eij[,2], n=rowSums(Eij))
> ORanemic

Data:
         Event  Size
Sample 1   367   779
Sample 2   201   428

Odds ratio:       1.006002
 95 % confidence intervals
                      LL              UL
Asymptotic 7.943026e-01 1.274123e+00
Exact       1.000000e+06 1.000000e+06
Score       7.943301e-01 1.274079e+00
```

OR for the expected table !!

Why OR=1 ?

Practical II

- Investigate whether there is association between Party Identification and Gender using Chi-square test (should be done manually in R).

Gender	Party Identification		
	Democrat	Independent	Republicans
Females	279	73	225
Male	165	47	191

- Verify your results using $chisq.test()$ function in R

Summary

- Chi-square is only valid when comparing nominal categorical variables. When either/both of the variables are ordinal, please discuss with a more experience colleague or your lecturer.

- There are other tests for two-way tables that are not covered in this course. This includes:
  - ➤ Trend test (M^2) test for ordinal outcomes
  - ➤ Fisher's exact test
  - ➤ Likelihood-ratio statistics

Part 3

Generalized linear models: a short introduction

Generalized linear models (GLM)

A framework for model fitting.

Examples:

- when an outcome 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.

$$\alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

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

$$Y_i \sim N(\alpha + \beta X_i, \sigma_\varepsilon^2)$$

The systematic component: the linear predictor

$$E(Y_i) = \alpha + \underbrace{\beta x_i}_{\text{Linear predictor}}$$
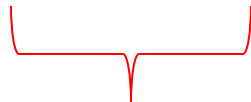
Linear predictor

The link function

$$\eta = \alpha + \beta X_i$$

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

$$\underbrace{g = 1}_{\text{Link function}}$$

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)

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

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

$$g = 1$$

Example 2: binary data

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

Dose response experiment:

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

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

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

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

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

Proportion of the killed beetles

140
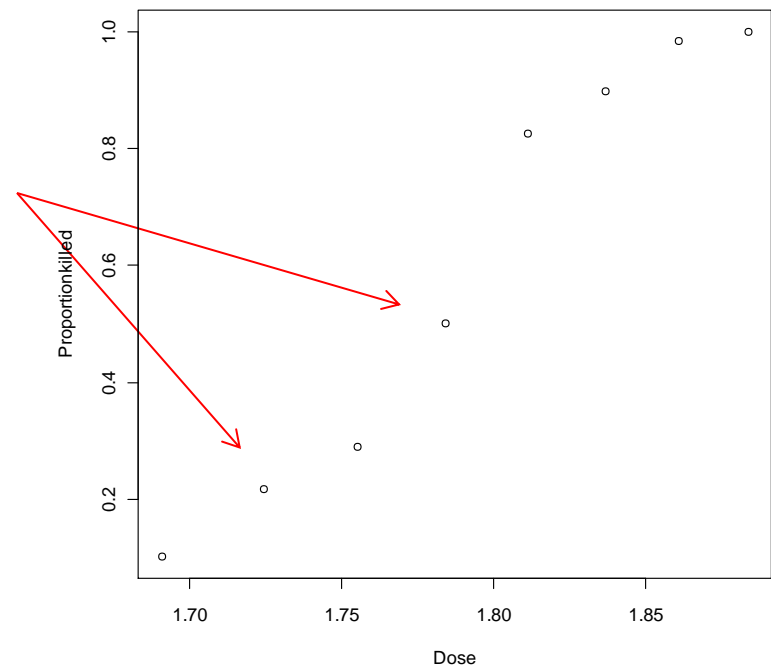
Systematic component: dependency of the predictor – the linear predictor

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

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

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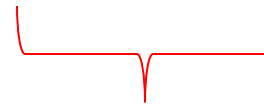
The linear predictor

The Link function

The expected values of the response variable

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

The systematic part

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

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

$$g(E(Y_{ij})) = g(\pi_j) = \eta$$
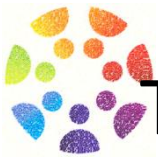
The Link function (logit link function for binary data)

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

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

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

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

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

$$Y_t \sim Poisson(\mu(t))$$

Random component: example of count data

$$Y_t \sim Poisson(\mu_t)$$

$$E(Y_t) = \mu_t$$
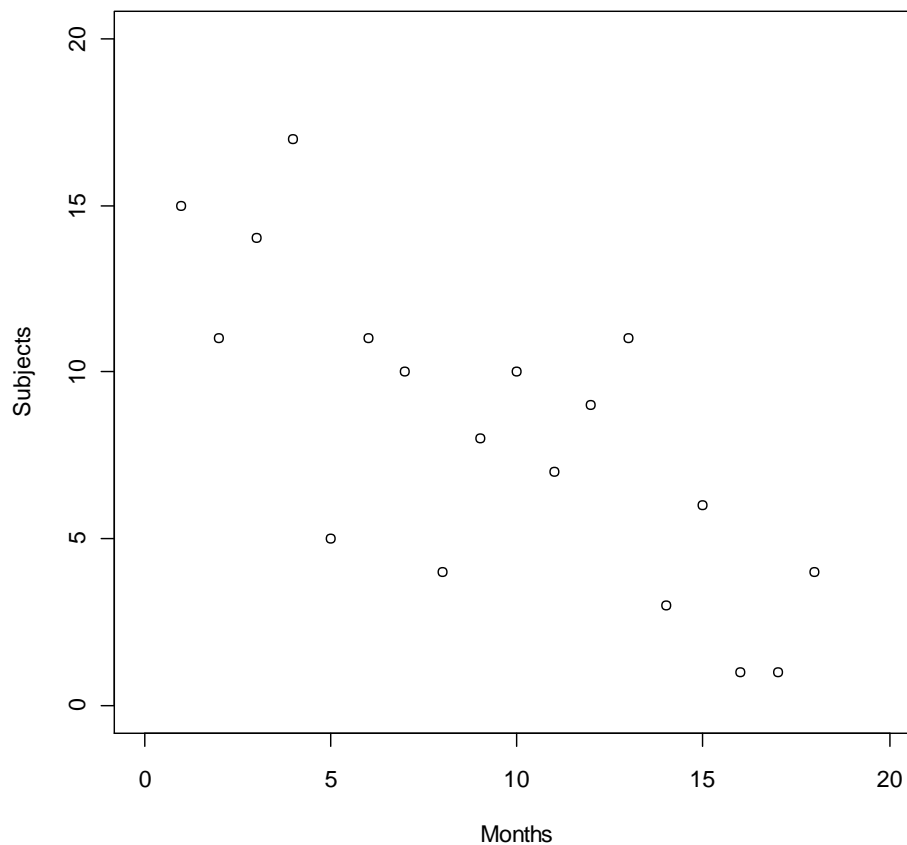
Systematic component: dependency of the predictor – the linear predictor

$$\mu_t = f(time) = f(t) = f(\beta_0 + \beta_1 t)$$

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

$$E(Y_t) = \mu_t$$

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

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

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

Example 4: mortality rate

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

```
> 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
```

$$\frac{d_i}{n_i} = r_i$$

$$d_i = r_i \times n_i$$

Random component: example of count data

$$Y_i \sim Poisson(\mu_i)$$

$$E(Y_i) = \mu_i$$

rate

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

$$g(\mu_i) = \log(\mu_i) = \log(n_i) + \beta_i$$
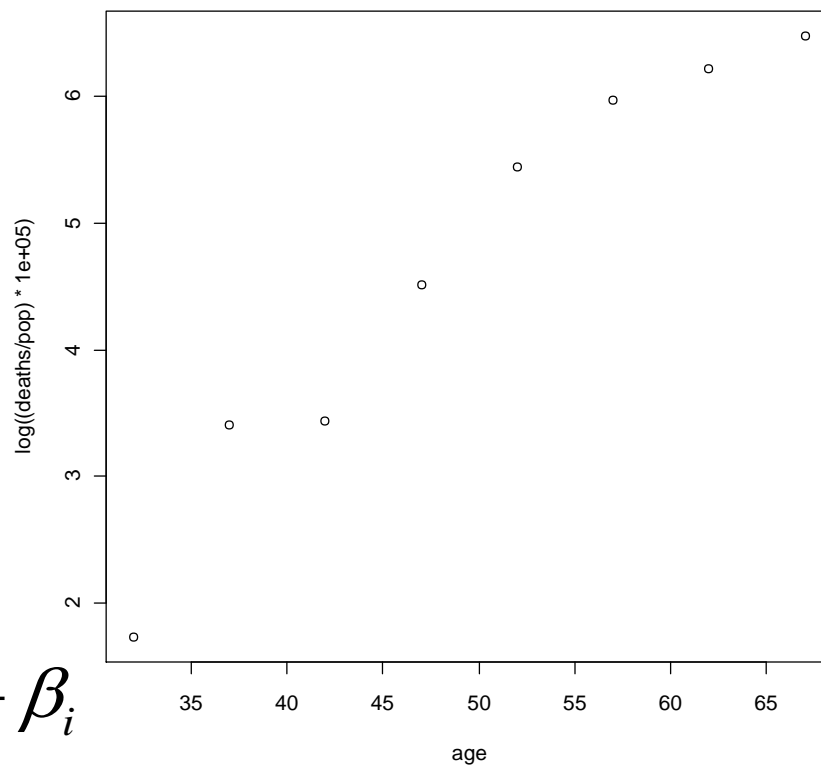
Inference about Model Parameters

Example: Poisson regression

$$Y_i \sim Poisson(\mu_i)$$

$$E(Y_i) = \mu_i$$

$$g(\mu_i) = \alpha + \beta \times X_i$$

The Wald test statistic

$$H_0 : \beta = 0$$

$$z = \frac{\hat{\beta}}{SE}$$

Under the null hypothesis:

$$z \sim \chi_1^2$$

Inference about Model Parameters

Example:

$$Y_i \sim H(\mu_i)$$

$$E(Y_i) = \mu_i$$

$$g(\mu_i) = \alpha + \beta \times X_i$$

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

The likelihood-ratio approach

$$H_0 : g(\mu) = \alpha$$

$$H_1 : g(\mu) = \alpha + \beta \times X$$

The LRT

$$-2\log\left(\frac{\ell_0}{\ell_1}\right) = 2\left(\log(\ell_0) - \log(\ell_1)\right)$$

Under the null hypothesis:

$$-2\log\left(\frac{\ell_0}{\ell_1}\right) \sim \chi_1^2$$

A Wald 95% confidence interval for a model parameter

Example: Poisson regression

The Wald 95% C.I

$$Y_i \sim Poisson(\mu_i)$$

$$E(Y_i) = \mu_i$$

$$g(\mu_i) = \alpha + \beta \times X_i$$

$$\hat{\beta} \pm 1.96 \times SE$$

The Deviance

The Saturated model:
A model that has a separate parameter for each observation, and it provides a perfect fit to the data.

Log likelihood:

$$L_S$$

A model with M parameters:

$$g(\mu_i) = \beta_0 + \beta_1 \times X_1 + ... + \beta_{m-1} \times X_{m-1}$$

Log likelihood:

$$L_M$$

$$L_S \leq L_M \quad \text{Why ?}$$

$$Deviance = -2(L_M - L_S)$$

$$-2\log\left(\frac{\ell_0}{\ell_1}\right) = 2(\log(\ell_0) - \log(\ell_1)) = Deviance_0 - Deviance_1$$

Part 4

introduction to logistic regression

Basic concept

We have mostly focused on investigating association between two categorical variables. The problem is that we cannot really investigate association between a binary outcome and a continuous explanatory variable without having to first categorise the continuous variable. Also, the simple analysis becomes cumbersome as we move to higher way table. For example three-way or four-way table. As a result, we are going to discuss a more formal modelling framework for binary data with flexibility to accommodate different data types and several explanatory variable.

Basic concept

In general we model observed data from an experiment assuming the underlying distribution of the data is known. This distribution dependent analysis is commonly refers to as parametric models. The most commonly used distributions are:

> ➤ Normal distribution for a continuous data

> ➤ Binomial distribution a dichotomous data

> ➤ Poison distribution for a count data

Basic concept

- Normal Distribution

$$\mathbf{Y|}\ \boldsymbol{x} \sim N(\boldsymbol{x}:\ \mu, \sigma^2)$$

$$\mathbf{g(}\mu) = \ \beta_0 + \beta_1 X$$

- **g(.)** is called link function that guarantee the linearity and additivity of the model. The default link is "identity link"
- σ^2 is the variance of the residuals in a general linear model or regression model

Basic concept

A bioassay experiment was designed to investigate the potency of a compound. The experiment consist of 5 groups of 40 mice. Each group was injected with combination of an infecting dose of a culture of pneumococci and one of five doses of the anti pneumococcus serum.

➢ **Outcome variable:** death from pneumonia within 7 days of inoculation
➢ **Explanatory variable:** Dose

Basic concept

Dose of serum	Number of deaths	Sample size
0.0028	35	40
0.0056	21	40
0.0112	9	40
0.0225	6	40
0.0450	1	40

Basic concept
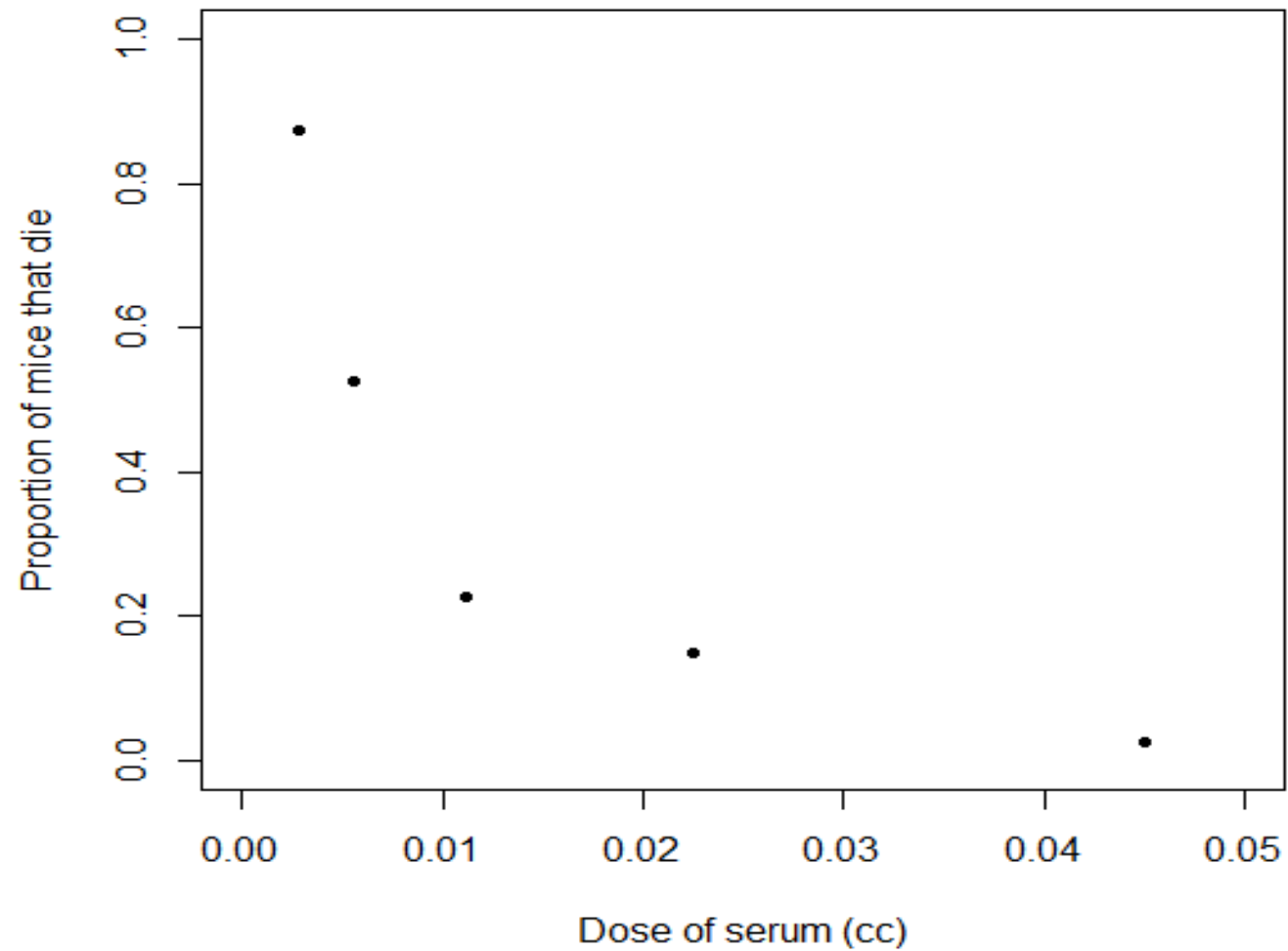
Basic concept

- **Generalized Linear Model in R**

object $<-$ glm(outcome variable $\sim$ explanatory, family(link), data)

- ➤ Outcome variable: the name of the outcome variable Explanatory variable: the name of the predictor(s)
- ➤ Family: the underlying distribution
- ➤ Link: the transformation function for the expectation/mean
- ➤ Data: name of the data

Basic concept

- **Normal Distribution with Identity link**

  glm(outcome variable ~ explanatory, family(link), data)

  - ➢ Outcome variable: **y/n**

  - ➢ Explanatory variable: Dose

  - ➢ Family: gaussian

  - ➢ Link: identity

  - ➢ Data: Anti_pneumococcuserum

Basic concept: normal regression

- **Codes**

```
> fit.1 <- glm(y/n~dose, family= gaussian(link=identity))
```

$$Y_i \sim N(\alpha + \beta \times dose_i, \sigma_\varepsilon^2)$$

$$E(Y_i) = \alpha + \beta \times dose_i$$

Basic concept: normal regression

- **Codes**

> fit.1 <- glm(y/n~dose, family= gaussian(link=identity))

```
> summary(fit.1)

Call:
glm(formula = y/n ~ dose, family = gaussian(link = identity))

Deviance Residuals:
      1         2         3         4         5
 0.2800   -0.0250   -0.2350   -0.1283    0.1083

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6400     0.1574   4.067   0.0268 *
dose        -16.0749     6.7781  -2.372   0.0984 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.05414835)

    Null deviance: 0.46700  on 4  degrees of freedom
Residual deviance: 0.16245  on 3  degrees of freedom
AIC: 3.0551

Number of Fisher Scoring iterations: 2
```

- **Implication**

$$\widehat{p}_i = 0.64 - 16 Dose_i$$
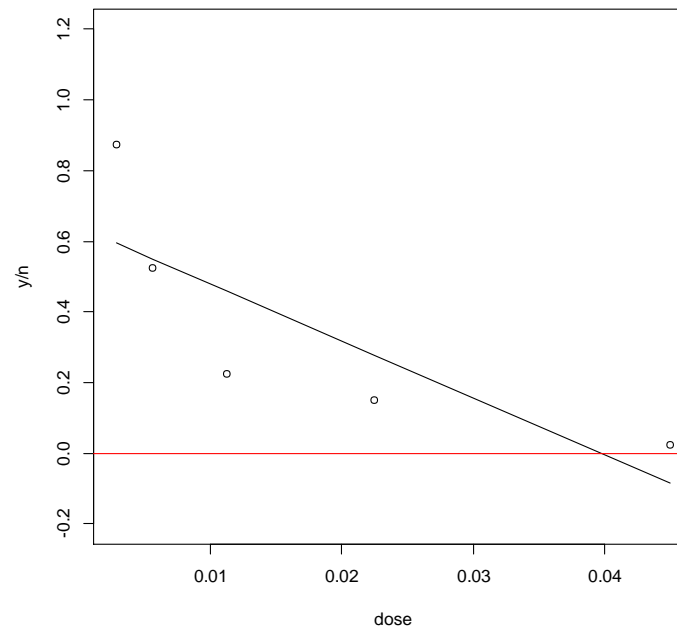
➢ If $Dose_i = 0.0028$, $p_i = 0.5952$

➢ If $Dose_i = 0.0450$, $p_i = -0.08$

object $<-$ glm(outcome variable $\sim$ explanatory, family(link), data)

➢ Outcome variable: **y/n** Explanatory variable: Dose

➢ Family: gaussian

➢ Link: **Log**

➢ Data: Anti_pneumococcuserum

Basic concept: normal regression with log link

- **Codes**

```
> fit.2 <- glm(y/n~dose, family= gaussian(link=log))
```

$$Y_i \sim N(e^{\alpha + \beta \times dose_i}, \sigma_\varepsilon^2)$$

$$E(Y_i) = e^{\alpha + \beta \times dose_i}$$

$$g\left(E(Y_i)\right) = g\left(E(Y_i)\right) = \alpha + \beta \times dose_i$$

- **Codes**

> fit.2 <- glm(y/n~dose, family= gaussian(link=log))

```
> summary(fit.2)

Call:
glm(formula = y/n ~ dose, family = gaussian(link = log))

Deviance Residuals:
        1          2          3          4          5
 0.02354   -0.03612   -0.01868    0.10472    0.02341

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2562     0.1333   1.923   0.1502
dose        -148.9399    28.5161  -5.223   0.0137 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.00457447)

    Null deviance: 0.467000  on 4  degrees of freedom
Residual deviance: 0.013722  on 3  degrees of freedom
AIC: -9.3016

Number of Fisher Scoring iterations: 9
```

- **Implication**

$$\hat{p}_i = \exp(0.2562 - 148.9399 Dose_i)$$

- ➤ If $Dose_i = 0.0028$, $p_i = 0.8514$
- ➤ If $Dose_i = 0.0450$, $p_i = 0.0016$
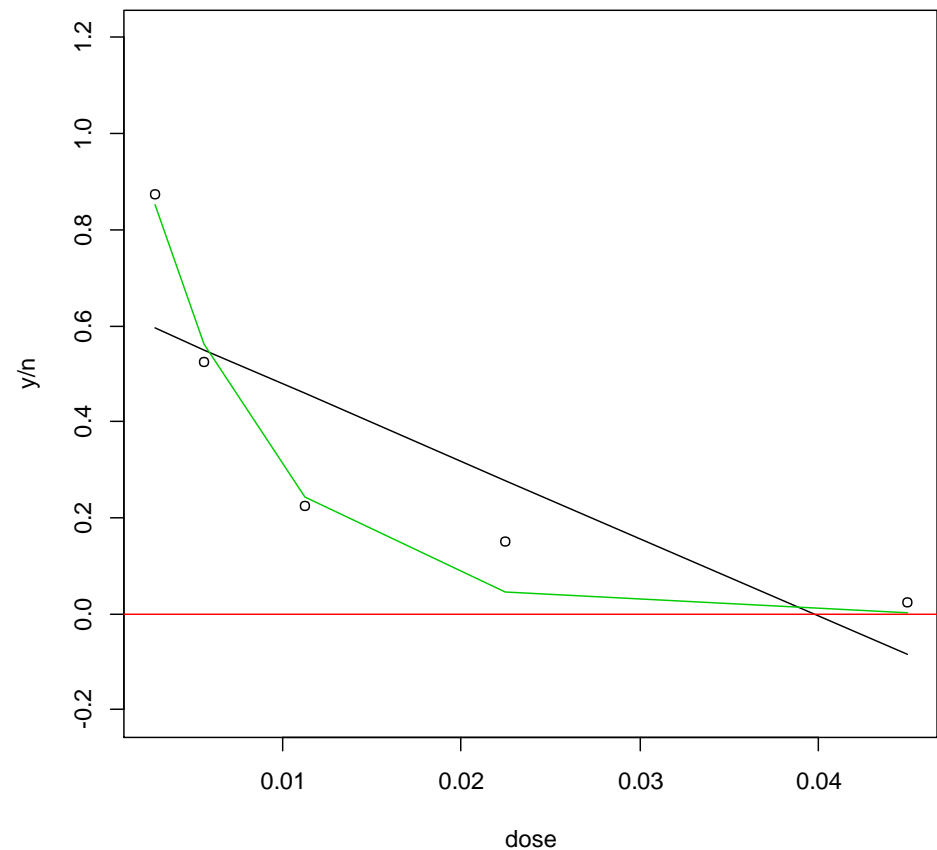
Logistic regression

- Simple Logistic Regression:

$$Y \sim B(N, \pi(x))$$

$$g(\pi(x)) = \beta_0 + \beta_1 X$$

$$\text{Log}\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X$$

$$\text{Logit}(\pi(x)) = \beta_0 + \beta_1 X$$

- **g(.) is a logit link, which is the default for a logistic regression.**

Logistic regression

- Back transformation from logit to probability

$$\text{Logit}(\pi(x)) = \beta_0 + \beta_1 X$$

$$\text{Log}\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 X$$

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

- Note that β_0 and β_1 are log of odds ratio. This will be clarified further with examples.

Logistic regression
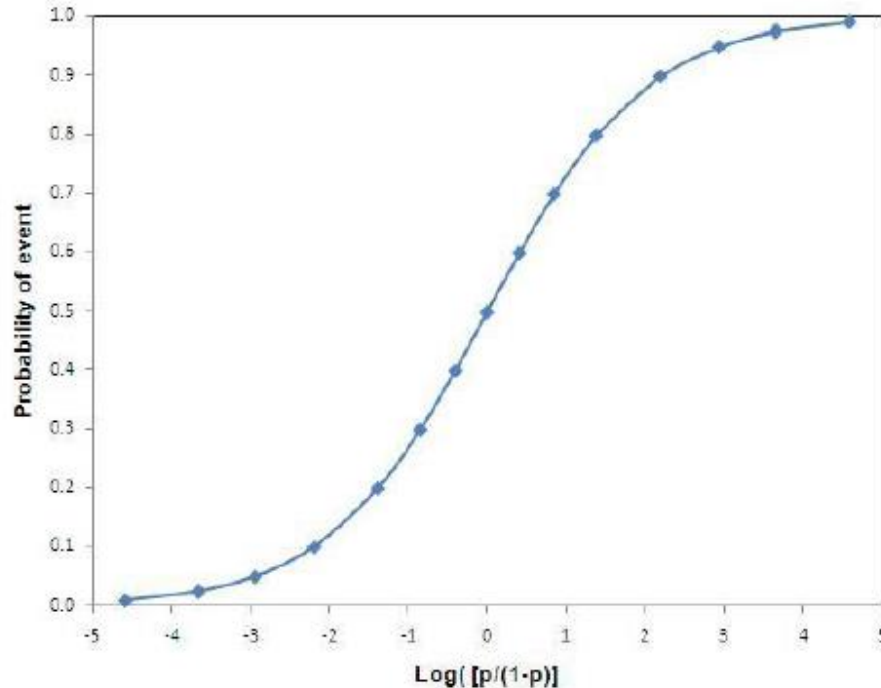
- The S-shaped curve is more appropriate than the straight line and is better at dealing with probabilities.

Logistic regression

- **Binomial Distribution with logit link**

$$object <- glm(outcome\ variable \sim explanatory,\ family(link),\ data)$$

- ➢ Outcome variable: **y/n**

- ➢ Explanatory variable: Dose

- ➢ Family: binomial

- ➢ Link: logit

- ➢ Data: Anti_pneumococcuserum

Logistic regression

- **Codes**

```
fit.3 <- glm(y/n~dose, family= binomial(link=logit))
```

$$Y_i \sim B\left(n_i, \pi_i\right)$$

$$E(Y_i) = \pi_i$$

$$\pi_i = \frac{e^{\alpha+\beta\times dose_i}}{1+e^{\alpha+\beta\times dose_i}}$$

$$g\left(\pi_i\right) = \log\left(\frac{\pi_i}{1+\pi_i}\right) = \alpha + \beta \times dose_i$$

Logistic regression

- **Codes**

fit.3 <- glm(y/n~dose, family= binomial(link=logit))

```
> summary(fit.3)

Call:
glm(formula = y/n ~ dose, family = binomial(link = logit))

Deviance Residuals:
      1        2        3        4        5
 0.4313  -0.1474  -0.3620   0.1193   0.2110

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.218      1.853   0.657    0.511
dose        -146.693    166.733  -0.880    0.379

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.2513  on 4  degrees of freedom
Residual deviance: 0.3975  on 3  degrees of freedom
AIC: 7.0168

Number of Fisher Scoring iterations: 6
```

- **Implication**

$$\widehat{p_i} = \frac{\exp(1.2179 - 146.6927 Dose_i)}{1 + \exp(1.2179 - 146.6927 Dose_i)}$$

$$\widehat{OR}(Dose) = \exp(-146.6927)$$

$$\cong 0.00$$

- ➤ If $Dose_i = 0.0028$, $\hat{p}_i = 0.6915$

- ➤ If $Dose_i = 0.0450$,, $\hat{p}_i = 0.0046$

- ➤ For a unit increase in dose, the odds of death decreases by 100%.

Logistic regression: data and predicted model

- **Implication**

$$\widehat{p_i} = \frac{\exp(1.2179 - 146.6927 Dose_i)}{1 + \exp(1.2179 - 146.6927 Dose_i)}$$

Which model is the best ?
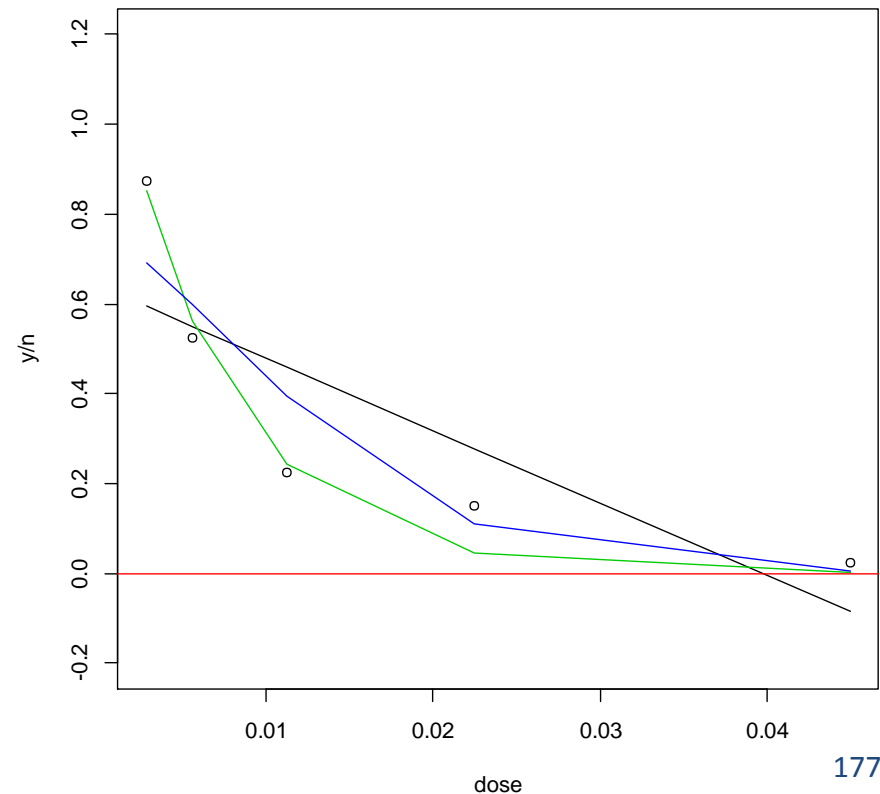
Logistic regression: example 2

- Binary data can also be analysed using probit or cloglog transformation. However, logistic regression model based on the logit link provides most intuitive and practically relevant interpretation. It is the most common form of binomial model in medical and health research.

Logistic regression in R: example 2

- Fit a logistic regression model to investigate the association between the likelihood of child anaemia and child gender.

Gender	Child Anemic		Total
	Yes	No	
Male	326	360	686
Female	243	278	521

Logistic regression in R

- **House Keeping**

```
anemic <-
read.csv("C:/projects/VLIR/CrossCutting/CoursesUpdated/BinaryKasim/dat
a/Child anaemia.csv", header=TRUE)
anemic$y <- ifelse(anemic$Child_Anemic=="Yes",1,0)
anemic$gender <- ifelse(anemic$Child_Gender=="Boy",1,0)
```

```
> head(data.frame(anemic$y,anemic$gender))
  anemic.y anemic.gender
1        1             0
2        1             1
3        0             1
4        1             0
5        1             0
6        0             0
> table(anemic$y,anemic$gender)

      0   1
  0 278 360
  1 243 326
```

Logistic regression in R

- **Logistic regression**

fit.1 <- glm(y~gender, family=binomial(link=logit),data=anemic)

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

$$E(Y_i) = \pi_i$$

$$\pi_i = \frac{e^{\alpha + \beta \times gender_i}}{1 + e^{\alpha + \beta \times gender_i}}$$

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 + \pi_i}\right) = \alpha + \beta \times gender_i$$

Logistic regression in R

- **Logistic regression**

fit.1 <- glm(y~gender, family=binomial(link=logit),data=anemic)

```
> summary(fit.1)

Call:
glm(formula = y ~ gender, family = binomial(link = logit), data = anemic)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.136   -1.136   -1.121    1.220    1.235

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.13456    0.08782  -1.532    0.125
gender       0.03535    0.11644   0.304    0.761

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1669.3  on 1206  degrees of freedom
Residual deviance: 1669.2  on 1205  degrees of freedom
AIC: 1673.2

Number of Fisher Scoring iterations: 3
```

Logistic regression in R

$$\text{Logit}(\hat{p}(x)) = -0.1346 + 0.0354X$$

$$X = \begin{cases} 1 \; if \; gender \; is \; male \\ \\ 0 \; if \; gender \; is \; female \end{cases}$$

$$\text{Logit}(\hat{p}(x)) = -0.1346 + 0.0354X$$

$$\text{Logit}(\hat{p}(x = female)) = -0.1346$$

$$\text{Logit}(\hat{p}(x = male)) = -0.1346 + 0.0354$$

$$\frac{Odds(Male)}{Odds(Female)} = \exp(0.0354)$$

Check !!!

Logistic regression in R

$$\pi_i = \frac{e^{\alpha + \beta \times gender_i}}{1 + e^{\alpha + \beta \times gender_i}}$$

$$\beta = \log(OR)$$

```
> table(anemic$y,anemic$gender)

     0    1
0  278  360
1  243  326

> (278*326)/(243*360)
[1] 1.035985
> exp(0.03535)
[1] 1.035982
```

- **Confidence Intervals**

```
> confint(fit.1)
Waiting for profiling to be done...
            2.5 %     97.5 %
(Intercept) -0.3071231 0.03733876
gender      -0.1928169 0.26374113
```

Logistic regression in R

$$\text{Logit}(\hat{p}(x)) = -0.1346 + 0.0354X$$

$$X = \begin{cases} 1 \; if \; gender \; is \; male \\ \\ 0 \; if \; gender \; is \; female \end{cases}$$

$$\text{Logit}(\hat{p}(x)) = -0.1346 + 0.0354X$$

$$\text{Logit}(\hat{p}(x = female)) = -0.1346$$

$$\text{Logit}(\hat{p}(x = male)) = -0.1346 + 0.0354$$

$$\frac{Odds(Male)}{Odds(Female)} = \exp(0.0354)$$

Check !!!

Logistic regression in R

- What is the odds ratio between boys and girls?

- What is the odds of anaemia for a male child?

- What is the odds of anaemia for a female child?

- What is the probability that a male child is anaemic?

- What is the probability that a female child is anaemic?

- Interpret your results

Logistic regression: example 3

- Investigate whether there is association between child anaemia and child location

Area	Anemic		Total
	Yes	No	
A	101	99	200
B	83	117	200
C	112	89	201
D	74	126	200
Total	370	431	801

Logistic regression in R

- **House Keeping**

anemic<- anemic[anemic$Areas!="Missing value",]

anemic$Areas <- as.factor(as.character(anemic$Areas))

anemic$y <- ifelse(anemic$Child_Anemic=="Yes",1,0)

```
> head(data.frame(anemic$y,anemic$Areas))
  anemic.y anemic.Areas
1        1       Area C
2        1       Area C
3        0       Area C
4        1       Area C
5        1       Area C
6        0       Area C
> table(anemic$y,anemic$Areas)

    Area A Area B Area C Area D
  0     99    117     89    126
  1    101     83    112     74
```

Logistic regression in R

- **Fitting the model**

Fit.1 <- glm(y~Areas, family=binomial(link=logit),data=anemic)

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

$$E(Y_i) = \pi_i$$

$$\pi_i = \frac{e^{\alpha + \beta \times Area_i}}{1 + e^{\alpha + \beta \times Area_i}}$$

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 + \pi_i}\right) = \alpha + \beta \times Area_i$$

Logistic regression in R

- **Fitting the model**

Fit.1 <- glm(y~Areas, family=binomial(link=logit),data=anemic)

```
> summary(fit.1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.0200     0.1414   0.141  0.88754
AreasArea B  -0.3633     0.2015  -1.803  0.07135 .
AreasArea C   0.2099     0.2004   1.047  0.29504
AreasArea D  -0.5522     0.2036  -2.712  0.00668 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1105.8  on 800  degrees of freedom
Residual deviance: 1088.3  on 797  degrees of freedom
AIC: 1096.3

Number of Fisher Scoring iterations: 4
```

Logistic regression in R

- **Confidence Intervals**

```
> confint(fit.1)
Waiting for profiling to be done...
                 2.5 %        97.5 %
(Intercept) -0.2575115  0.29776988
AreasArea B -0.7597756  0.03070055
AreasArea C -0.1825877  0.60367063
AreasArea D -0.9535298 -0.15473561
```

Logistic regression in R

- What is the odds ratio area A and B?

- Calculate the odd of anaemia for each area.

- Calculated probability of child anaemia for each area?

- Interpret your results

Logistic regression in R

Montgomery and peck(1982) describe a study on the compressive strength of an alloy fastener used in the construction of aircraft. Ten pressure loads, increasing in units of 200 psi from 2500 psi to 4300 psi, were used with different number of fasteners being tested at each of these loads. Is there an association between fastener failure and load?

Logistic regression in R

Unlike previous examples, we would like to investigate association between a binary outcome and a continuous explanatory variable. In other words, does a unit increase in loading increase the probability of fastener failure? Suppose the probability for fastener failure is p. The logistic regression model is

$$logit(p_i) = \beta_0 + \beta_1 load_i$$

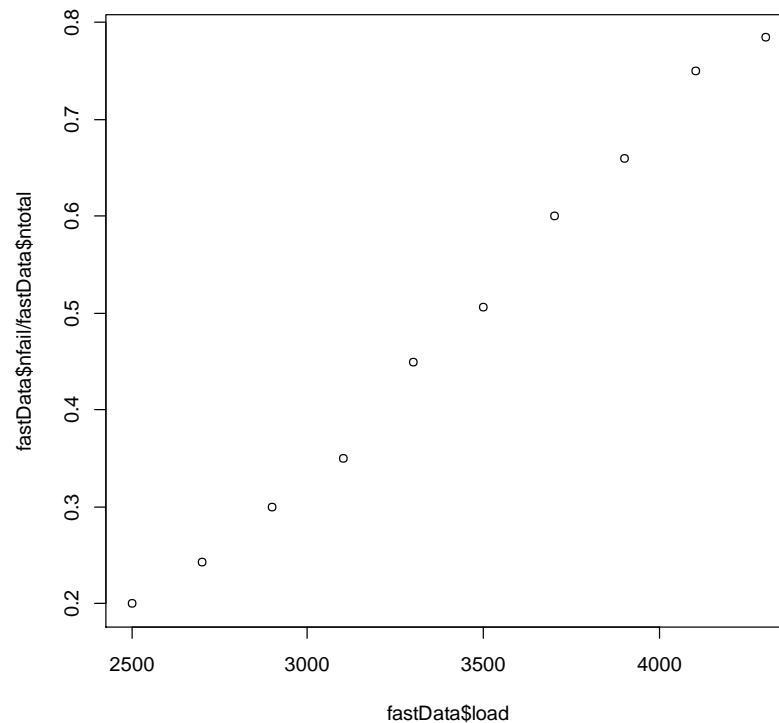# Logistic regression in R

- **Code**

```
fastData <- read.cvs("I:/Ethiopia/data/fast.cvs",header=TRUE)
```

```
> fastData
    load ntotal nfail
1   2500     50    10
2   2700     70    17
3   2900    100    30
4   3100     60    21
5   3300     40    18
6   3500     85    43
7   3700     90    54
8   3900     50    33
9   4100     80    60
10  4300     65    51
```

- **Model formulation**

fit.1 <- glm(cbind(nfail,ntotal-nfail)~load,family=binomial(link="logit"),data=fastData)

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

$$E(Y_i) = \pi_i$$

$$\pi_i = \frac{e^{\alpha+\beta \times load_i}}{1+e^{\alpha+\beta \times load_i}}$$

$$g(\pi_i) = \log\left(\frac{\pi_i}{1+\pi_i}\right) = \alpha + \beta \times load_i$$

Logistic regression in R

- **Code**

```
fit.1 <- glm(cbind(nfail,ntotal-nfail)~load,family=binomial(link="logit"),data=fastData)
```

```
> summary(fit.1)

Call:
glm(formula = cbind(nfail, ntotal - nfail) ~ load, family = binomial(link = "logit"),
    data = fastData)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.3397115  0.5456932  -9.785   <2e-16 ***
load         0.0015484  0.0001575   9.829   <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: 112.83207  on 9  degrees of freedom
Residual deviance:   0.37192  on 8  degrees of freedom
AIC: 49.088

Number of Fisher Scoring iterations: 3
```

Logistic regression in R: data and predicted model
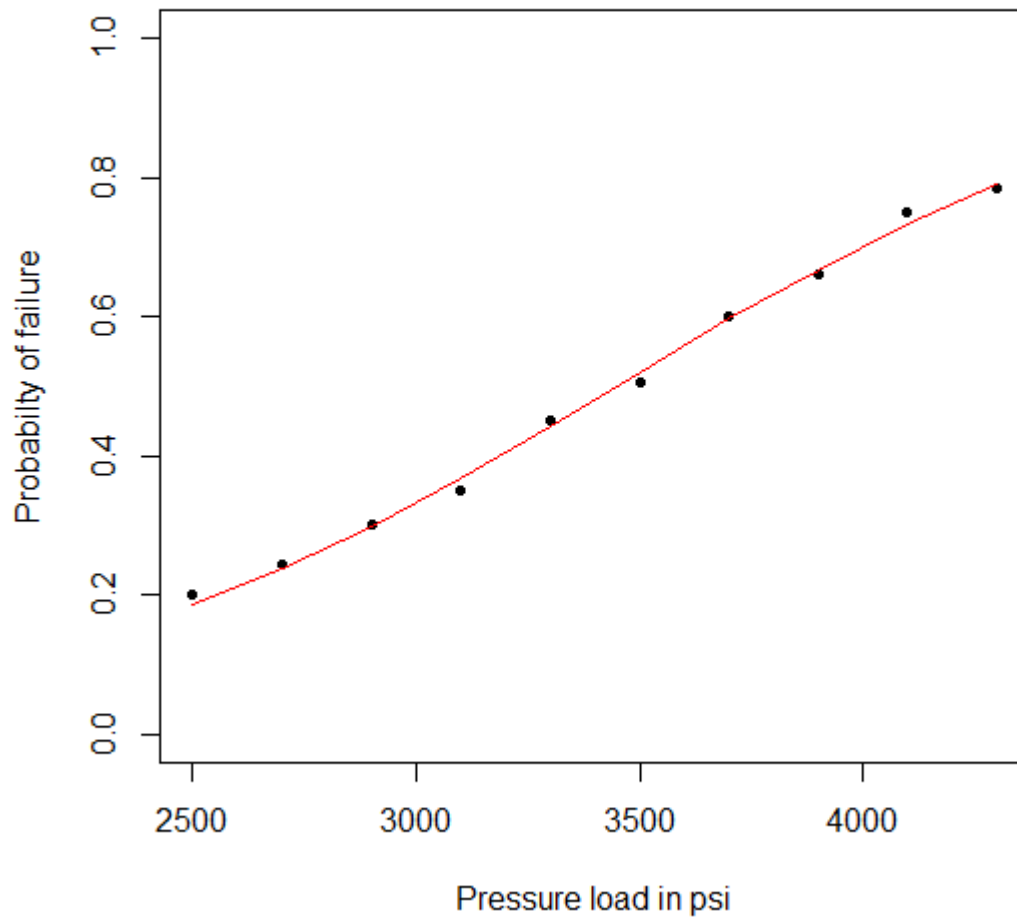
Logistic regression in R

$$\text{Logit}(\hat{p}(X = load)) = -5.3397 + 0.0015 * \text{Load}$$

$$\text{Logit}(\hat{p}(X = load + 1)) = -5.3397 + 0.0015 * (\text{Load} + 1)$$

$$\frac{Odds(X=load+1)}{Odds(X=load)} = \exp(0.0015)$$

Check !!!

Logistic regression in R

- What is the odds ratio corresponding to 200psi change in load?

- What is the odds ratio corresponding to 800 psi change in load?

- What is the probability of fastener failure for a loading of 3100psi?

- Interpret your results

Practical IIIa: Child Anaemia data

- Fit logistic regression models to investigate associations between mother anaemia status and

    i. Mother's age

    ii. Mother's location (Areas)

    iii. Socioeconomic status (ses)

- From each of the model,

    i. calculate the odds ratio and its associated confidence interval between the respective categories

    ii. Calculate the probability of mother's anaemia for the different categories of the explanatory variables

Practical IIIb: esrData

A reseacher was to Examine the extent to which the disease state of an individual refelcted in his/her ESR reading is related to levels of two plasma proteins, fibrinogen and γ-globulin.

Practical IIIb: esrData

- Fit logistic regression models to investigate association between disease state (y =1 if ESR level > 20; 0 otherwise) and

  i.   Fibrinogen

  ii.   Globulin

- Calculate the odds ratio and probabiities for

  i.   1.5 unit changes in fibrinogen level

  ii.   5 units changes in globulin level

Part 5

Modeling Binary data

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

Example tour
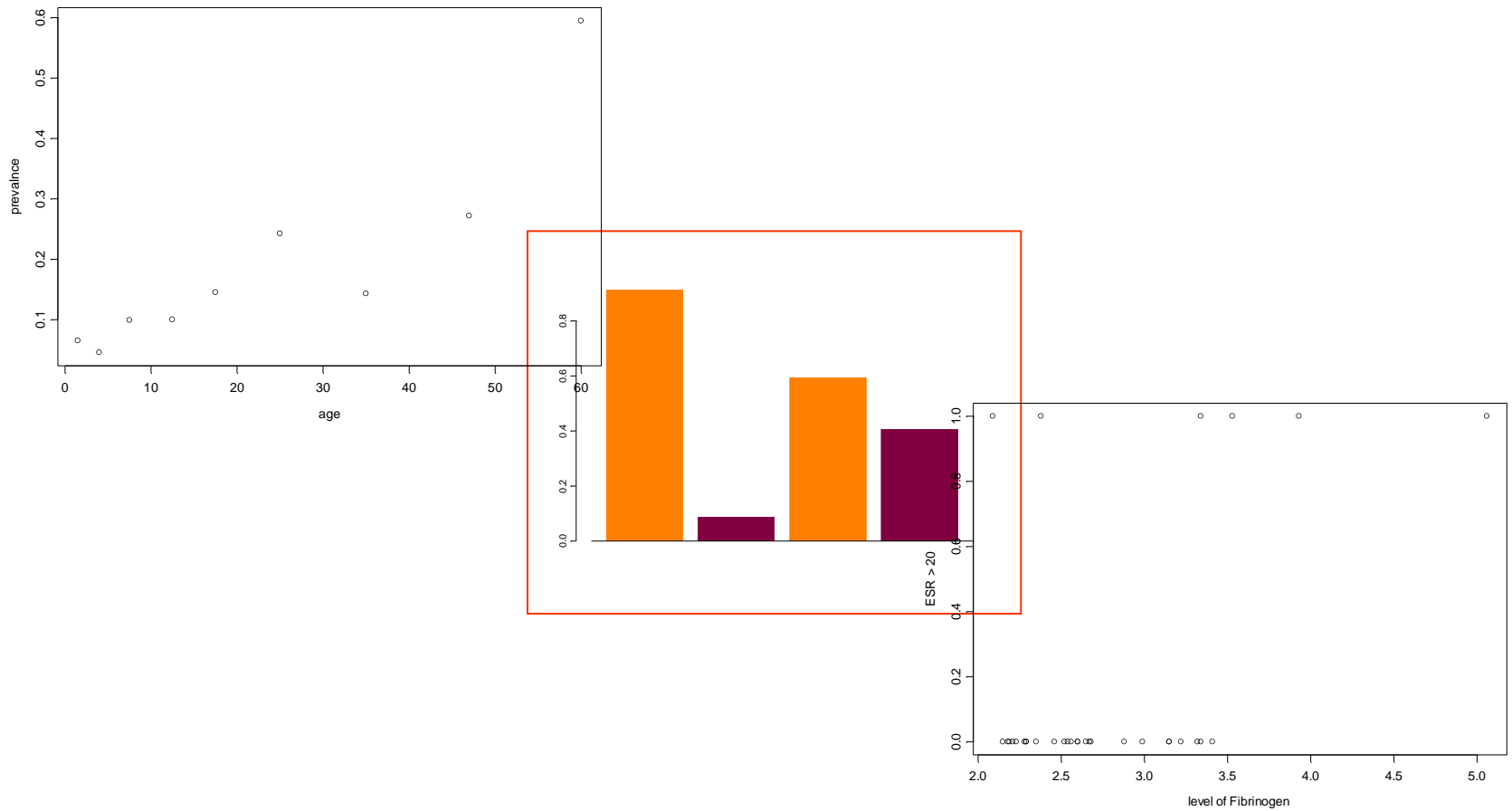
Example 1: The Aspirin and Myocardial Infarction Data

- 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?

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

The response variable

Example 2: smoked mice

In order to investigate the influence of smoking on lung censer 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 cambers 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.

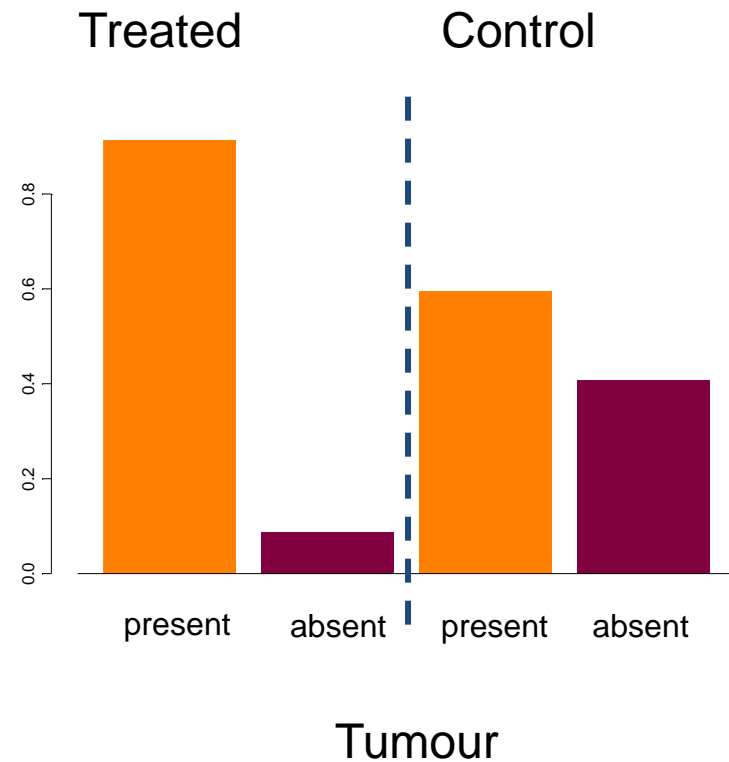# Example 2: smoked mice

The question of primary interest is:

DOSE THE SMOKE INCREAE THE RISK FOR CANSER ?

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

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
Contol	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

Treated

Contol

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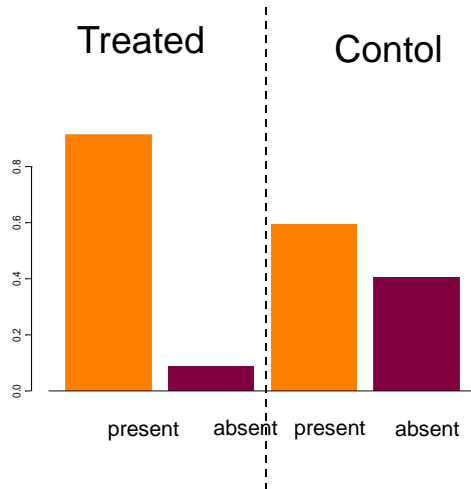
In the treated group, 21/23 (91%) of the mice develop tumour. In the control group only 19/32 (59%).

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-prevalnce data

Age at test

Example 3: serological data

Age at test   Age at infection

Age
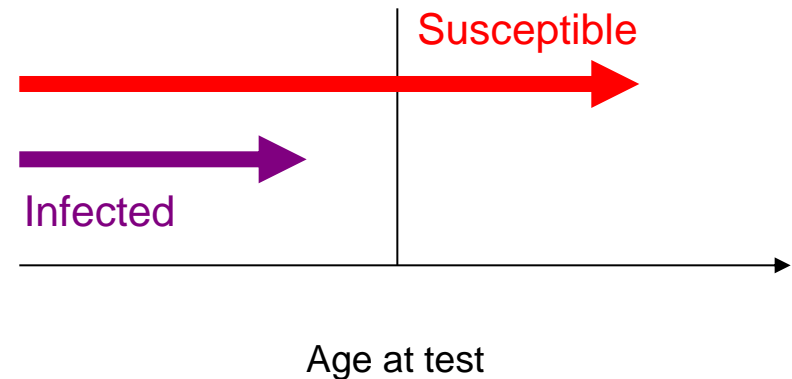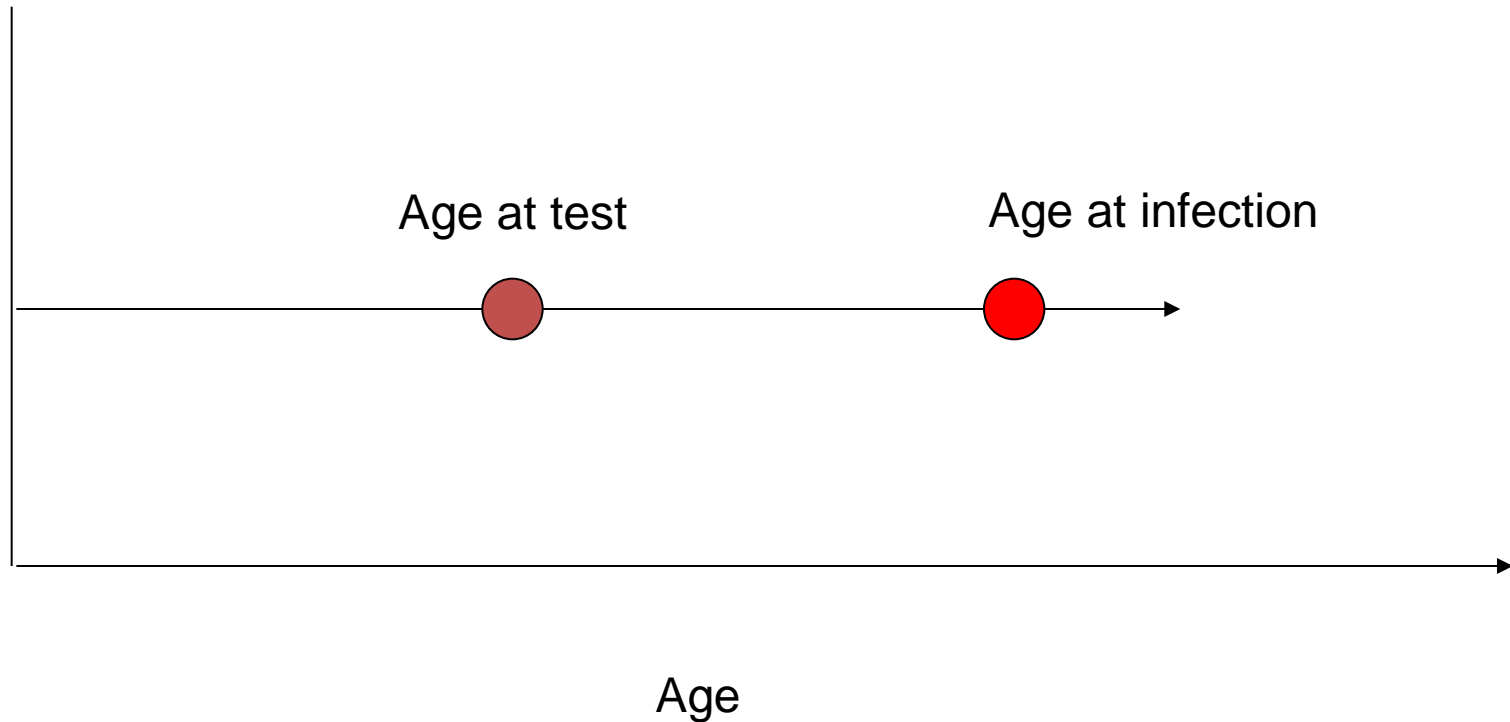
• Sero-Negative: infected after the test.

Example 3: serological data

Age at infection   Age at test

Age
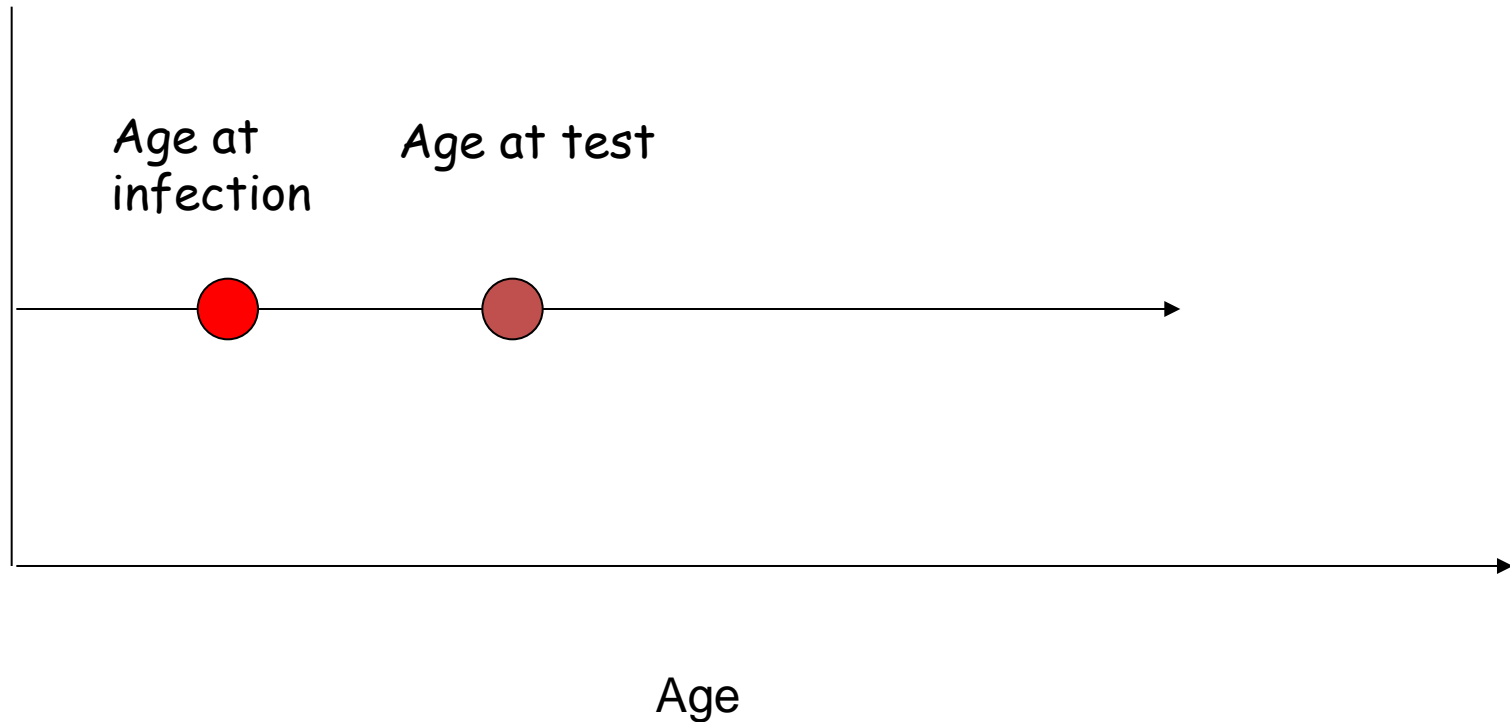
•Sero-Positive: infected before the test.

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

44/74

25/92
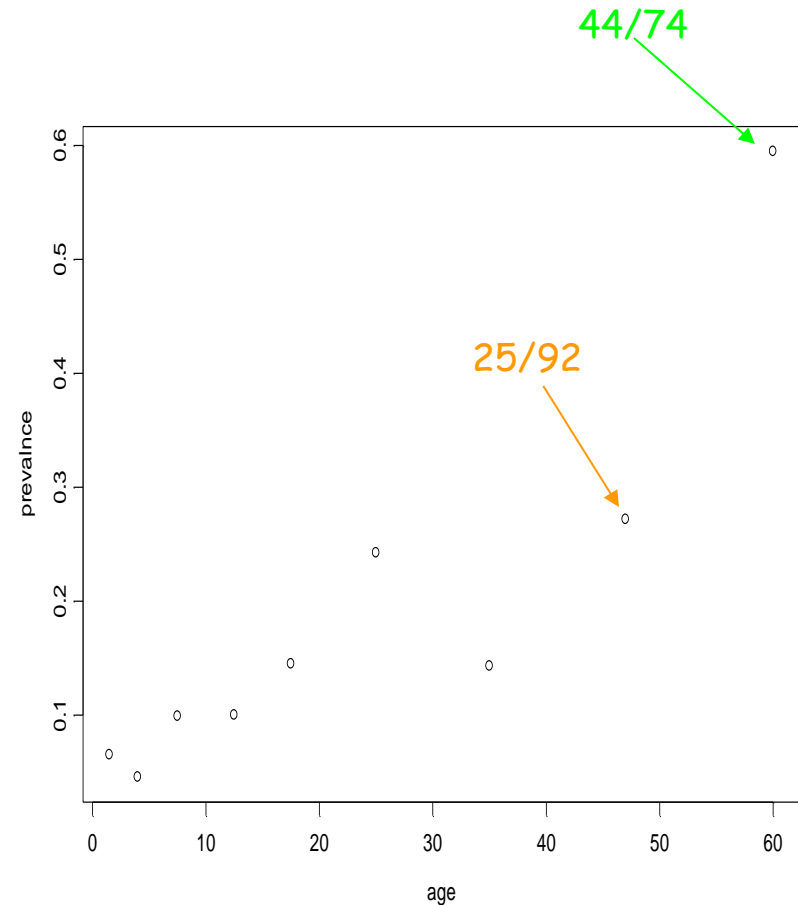
Example 4: HIV data

- Consider the HIV data set and the model for HIV (the outcome variable, yes/no or 1/0).

- Covariates:

- Silicosis and 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).
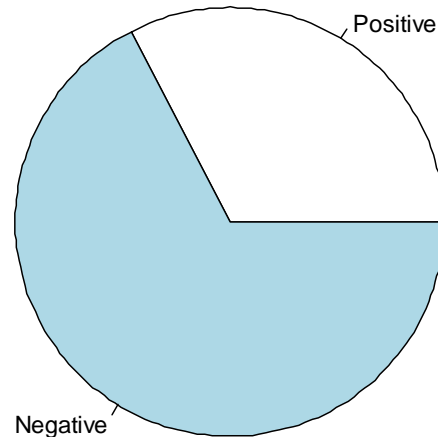
Example 4: HIV data

```
> 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 ?
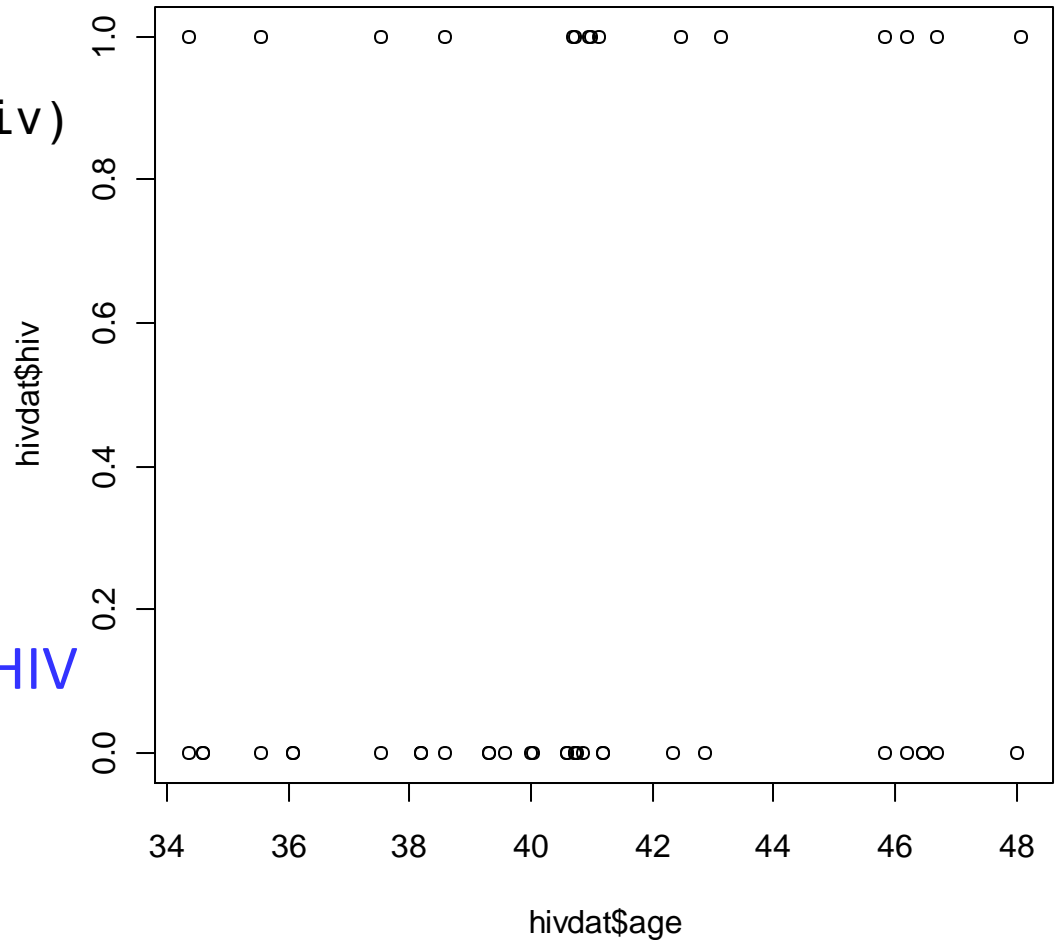
Example 5: toxicity example (Budworm)

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 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(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.

	Heart Disease		Proportion
Snoring	Yes	No	Yes
Never	24	1355	0.017
Occationally	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)
```

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

Does the probability to be heart disease depends on snoring ?

Heart disease Distribution by Snoring and ferquency
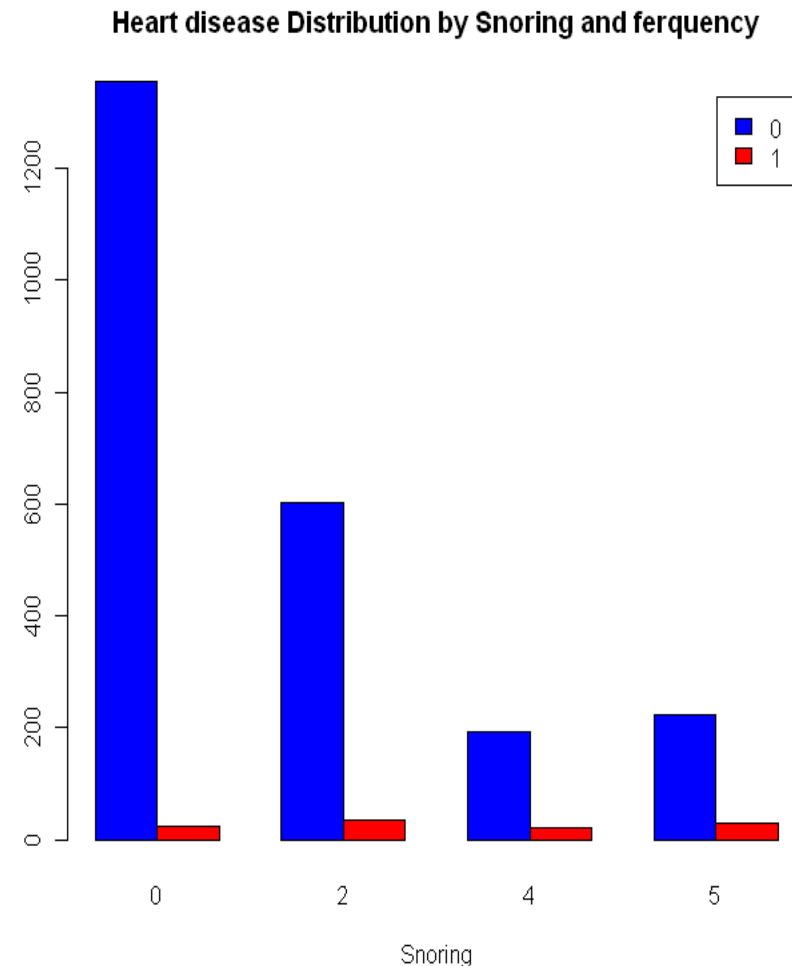
Modeling Binary data

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.

$$Z_1, Z_2, Z_3 ... Z_{n_i}$$

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

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

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

$$E(Z_i) = P_i$$

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

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

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

If Z~B(1,P) then Y~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?

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

The probability of sucsses

- 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
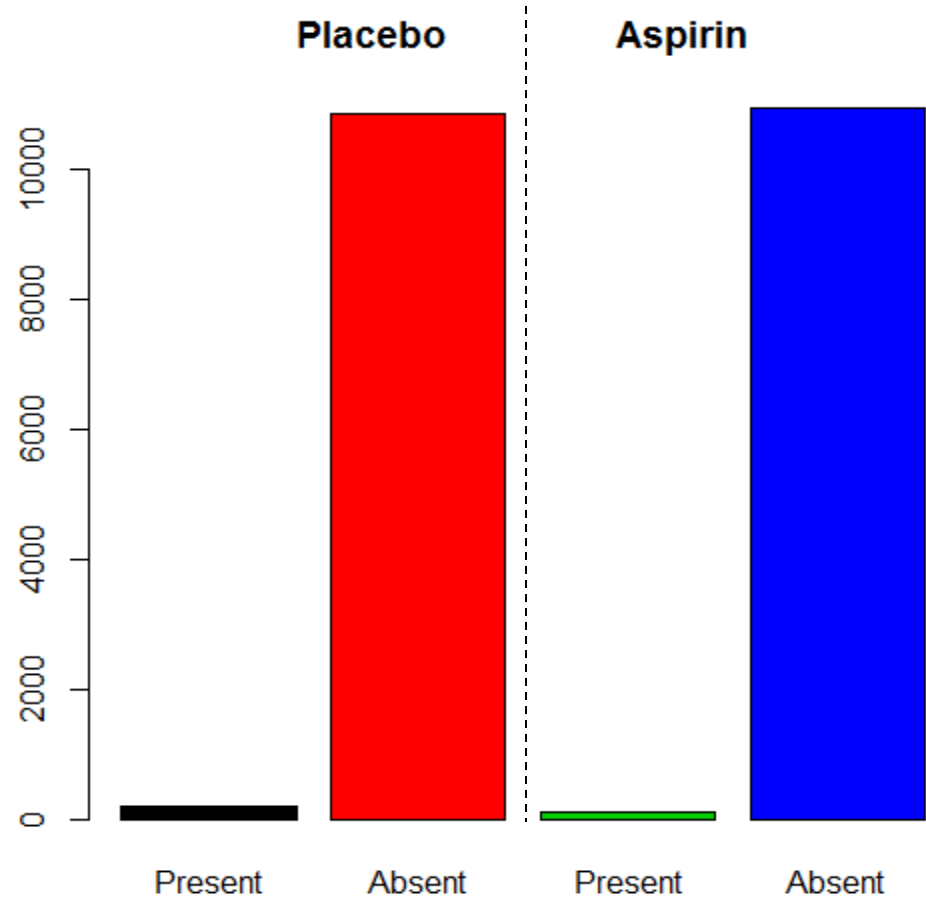
Data structure in R

- 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

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

```
< 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
```

$\hat{\alpha}$

$\hat{\beta}$

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

How do we interpreat the parameters from the output above ?

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

The odds ratio, θ, is equal to 0.5458342. If θ < 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 1: Aspirin

The problem as a GLM

Results from glm()

```
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 ***
---
```

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

$$\exp(\hat{\beta}) = \exp(-0.60544)$$

$$\frac{1}{\exp(\hat{\beta})}$$

```
> 1/exp(-0.60544)
[1] 1.832058
```

The problem as a 2 X 2 table

Results from glm()

```
Coefficients:
                 Estimate Std. Error
(Intercept)      -4.04971    0.07337
as.factor(trt)2  -0.60544    0.12284
---
```

$$\frac{1}{\exp(\hat{\beta})}$$

```
> 1/exp(-0.60544)
[1] 1.832058
```

Analysis of a 2 X 2 table

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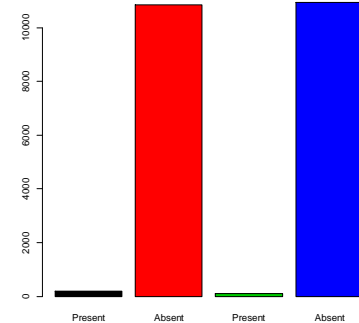


```
> RRAspirin1 <- oddsratio(x=Aspirin1[,1],
> RRAspirin1

Data:
             Event   Size
Sample 1       189  11034
Sample 2       104  11037

Odds ratio:        1.832054
```

Example 2: smoked mice

The question of primary interest is:

DOSE THE SMOKE INCREAE THE RISK FOR CANSER ?

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

The response variable

The probability of sucsses

- 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

Treated   Control
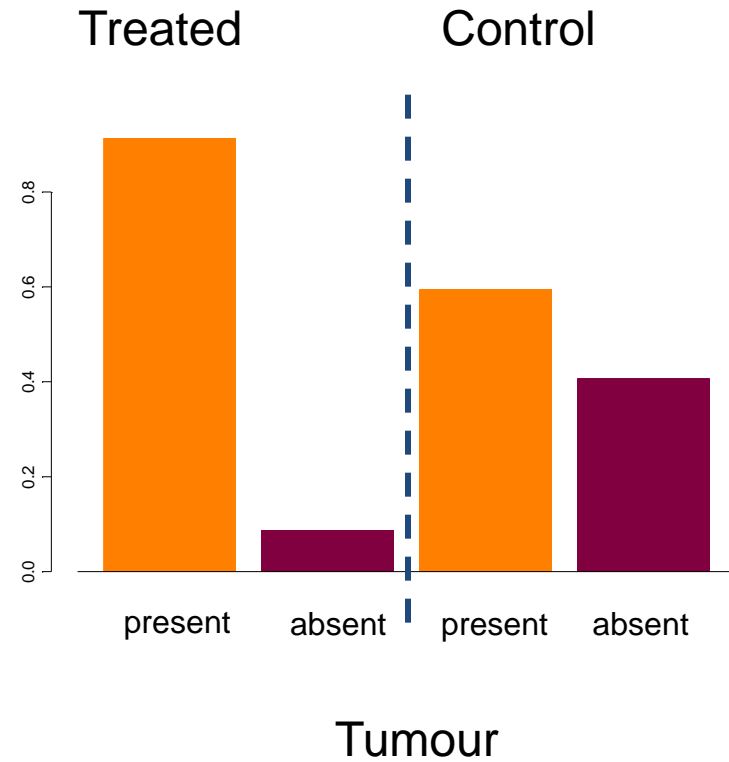
Tumour

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

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

Number of subjects with tumour

$$Y_i = \sum Z_i$$

Distribution of Y

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

The model for P- logit transformation

$$\log it(P) = \mu + \beta_j$$
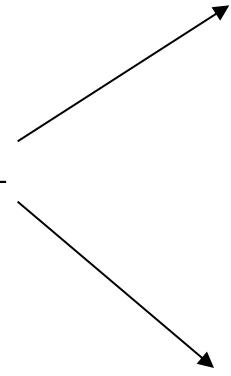
The probability

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

The parameter β_j is the treatment effect.

Note that we have two treatment groups and since the sum of the effects is zero it follows that $\beta_{treatment} = -\beta_{control}$.

The probability

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

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

The probability to have tumor for the treatment group.

$$P = \frac{e^{\mu+\beta_{control}}}{1+e^{\mu+\beta_{control}}}$$

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"))
```

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

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|)
(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
```

$\hat{\alpha}$

$\hat{\beta}$

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

How do we interpreat 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 -1.9719. The parameter estimate for the effect of the treatment group (the smoked group) is equal to 1.9719.

The odds ratio: point estimator

The odds ratio, θ, is equal to 0.139. If θ < 1 than 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.

The odds ratio: how do we calculate the value of θ

```
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: $\theta = \exp(1.9719) = 7.184314$.

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

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
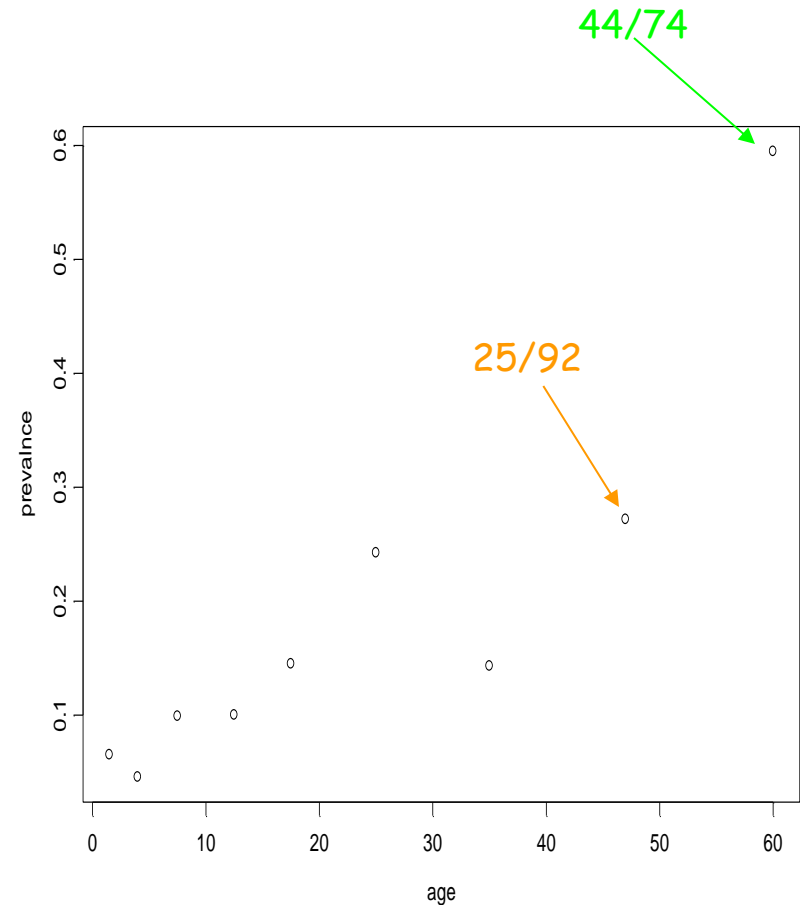
44/74

25/92

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

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

$$Y_i = \sum Z_i$$

Number of sero-positive at each age group

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

n_i: sample size at each age group

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

$$\log it(P_i) = \alpha + \beta \times age$$
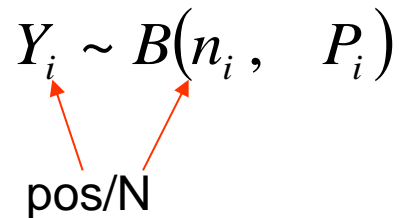
The probability of infection

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

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

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

logistic in R

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

pos/N

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

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

model pos/N=age

Parameters estimate

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

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

```
> 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

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

$$\hat{P}_i = \frac{e^{-2.71+0.044\times age}}{1+e^{-2.71+0.044\times age}}$$
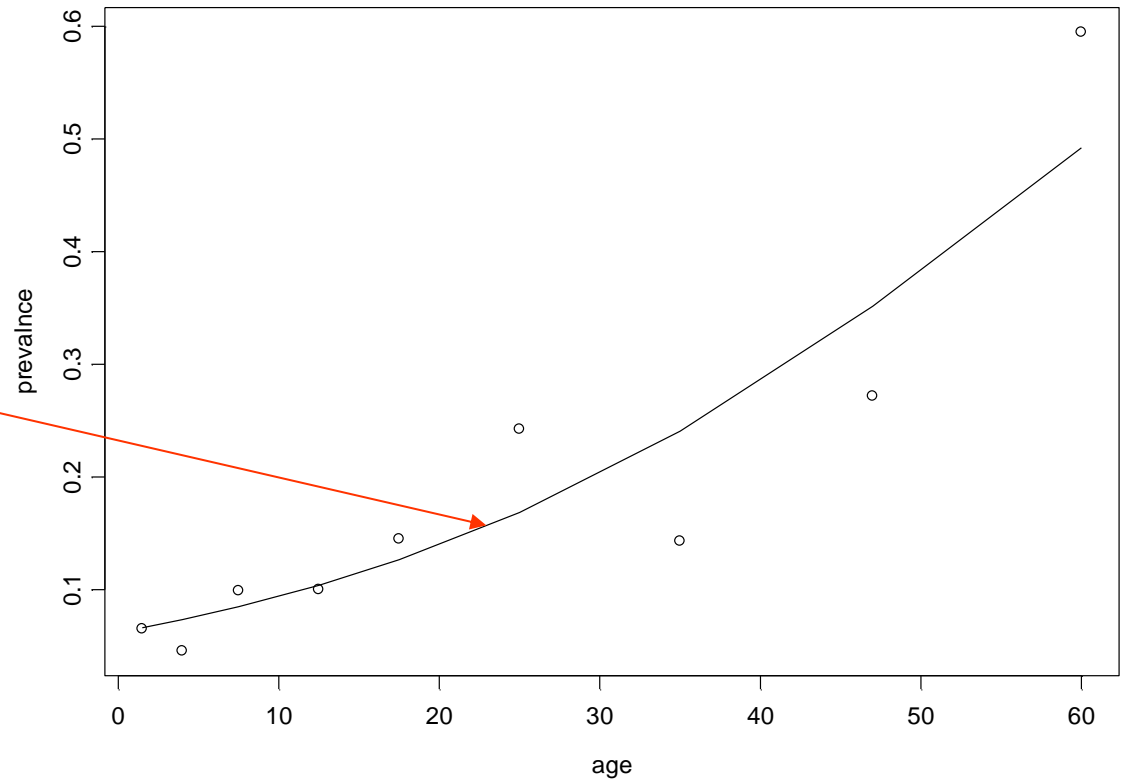
The odds ratio: point estimator

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

How to calculate the odds ratio ? For continuous predictor the odds ratio is given by

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

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

Implies per unit increase of 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.

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

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

$$X_i = age_i$$

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 ?
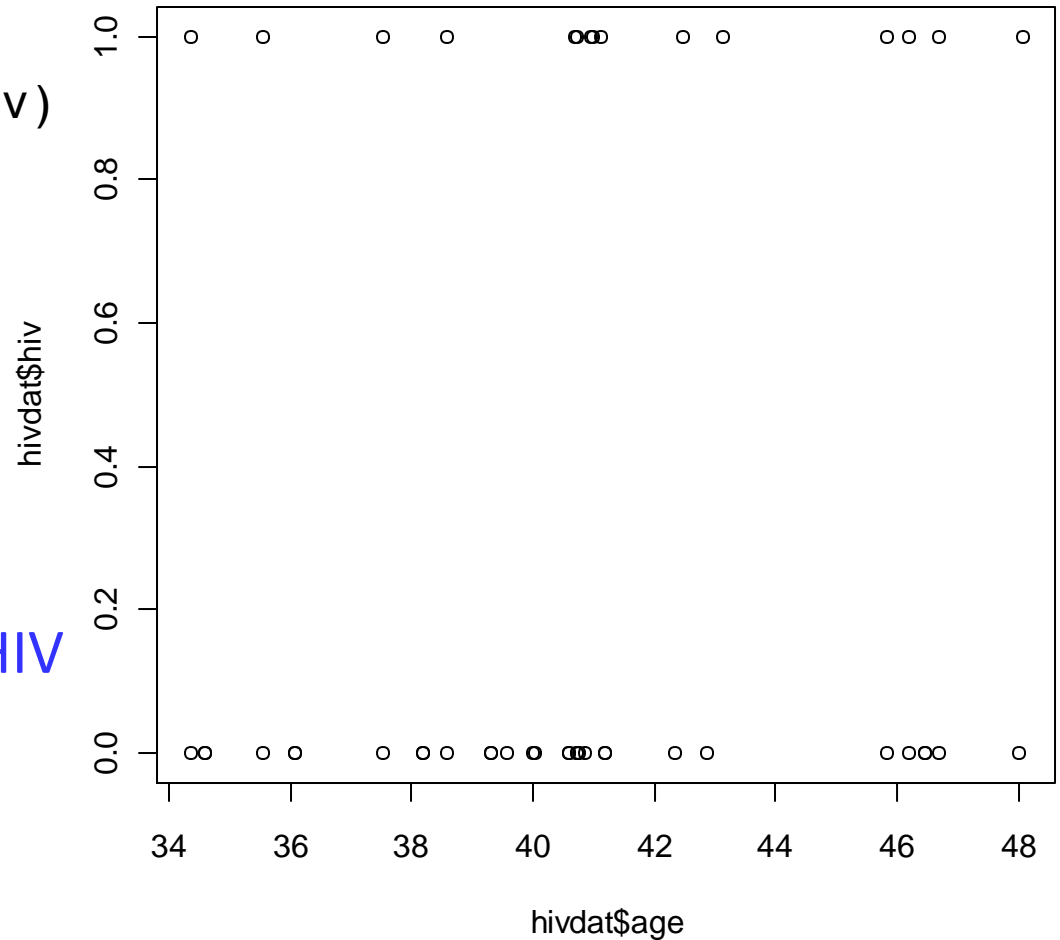
Model formulation

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

$$E(Y_i) = \pi$$

$$\pi = f(X_i) = f(age_i)$$

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
```

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

$$\pi_i = \frac{e^{-3.79 + 0.0749 \times age_i}}{1 + e^{-3.79 + 0.0749 \times age_i}}$$
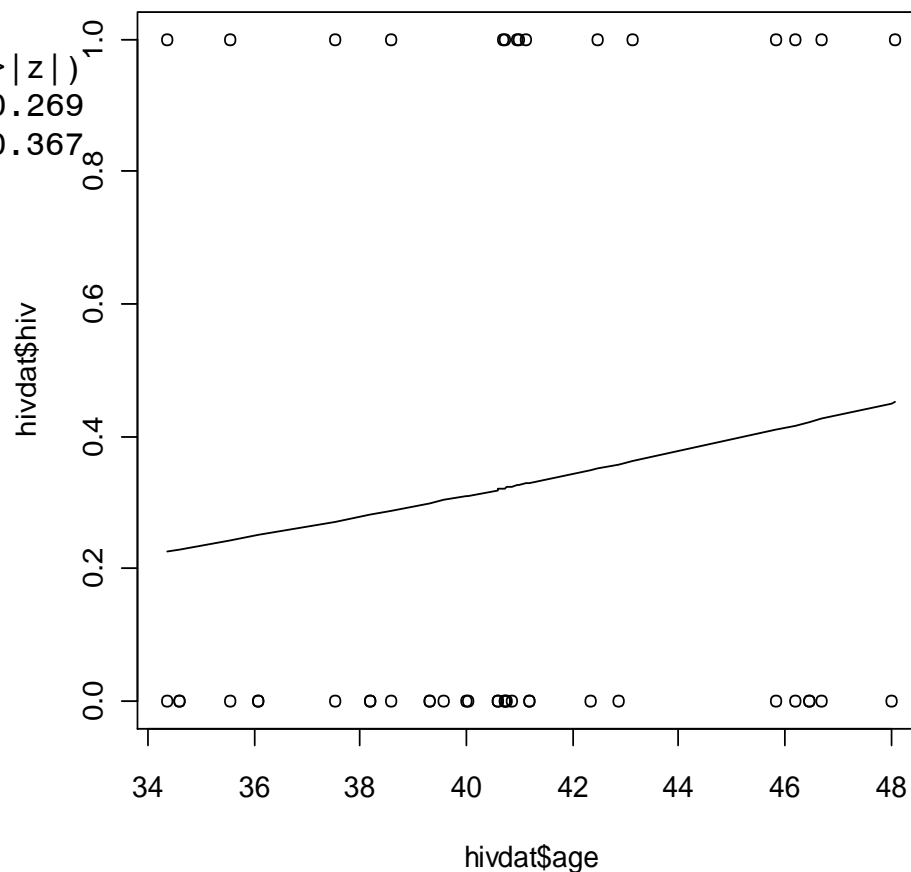
The odds ratio: point estimator

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

How to calculate the odds ratio ? For continuous predictor the odds ratio is given by

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

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

As age increases the probability to have HIV posetive increase by 7.8%

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(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))
```
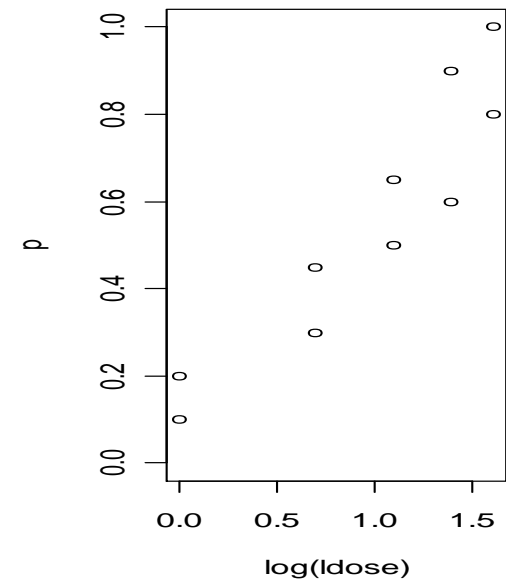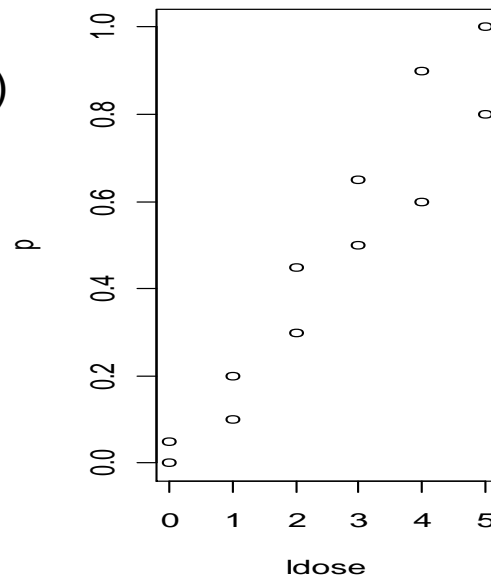
Model formulation

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

$$\pi_j = f(dose \quad gender)$$

$$\eta = dose + gender + dose*gender$$

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

Model formulation

Distribution of the response

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

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

The linear predictor

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

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

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

Data 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)
```

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,...)
```

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

Fitting the model with the glm() function:

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

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

Alternative code

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

Summary of fit using glm for Binomial

```
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

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

| Snoring | Heart Disease | | Proportion |
	Yes	No	Yes
Never	24	1355	0.017
Occationally	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
```

The estimated model in R

➢summary(fit.snoring)
➢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

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

Example using Figure

Data and estimated probabilty
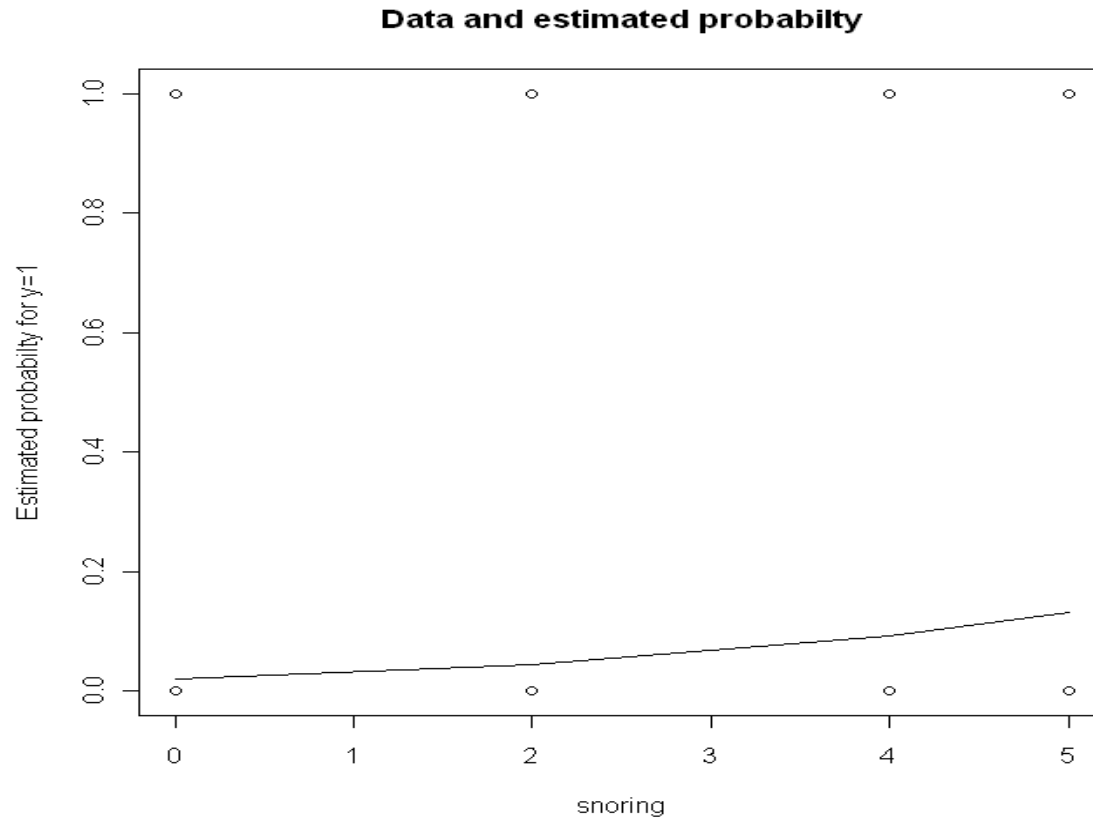
Extractor functions in R

- 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()`

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 form is;

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

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"))
```

Data and predicted model
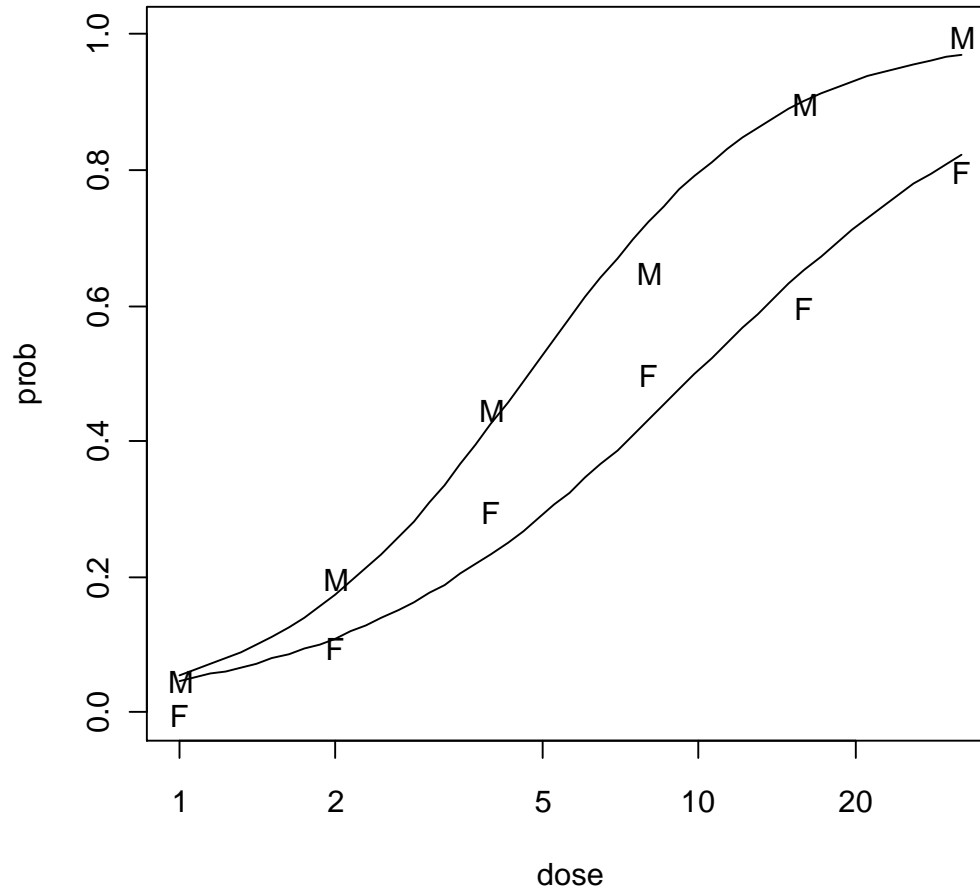
Part 6

Multiple Logistic Regression

Multivariable logistic regression

Simple logistic regression as previously discussed relies on a very strong assumption that the association between an outcome and explanatory variable does not depend on any other factor. In most real life scenario there is always one or more variables that are also significantly associated with the outcome.

Multivariable logistic regression

- Simple logistic regression

$$\text{Logit}(\pi(x)) = \beta_0 + \beta_1 X$$

- Multivariable logistic regression

$$\text{Logit}(\pi(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Multivariable logistic regression

Fit a logistic regression model to investigate associations between child anaemia and child age accounting for mother's anaemic status

$$\text{Logit}(\pi(Child\ Anaemia)) = \beta_0 + \beta_1 ChildAge + \beta_2 MothersAge$$

Multivariable logistic regression

- **House Keeping**

```
anemic <- read.csv("I:/Ethiopia/data/Child anaemia.csv",header=TRUE)

anemic$y <- ifelse(anemic$Child_Anemic=="Yes",1,0)

anemic$motherAnemic <- ifelse(anemic$Mother_Anaemic=="Anemic",1,0)

anemic$childAge <- ifelse(anemic$"Child_Agecat"=="24-59 months",1,0)
```

- **model**

```
fit1 <- glm(y~childAge+motherAnemic , family=binomial(link=logit) ,data=anemic)

fit1Coef <- round(summary(fit1)$coef,4)

fit2 <- glm(y~childAge+motherAnemic+childAge*motherAnemic,

        family=binomial(link=logit),data=anemic)

fit2Coef <- round(summary(fit2)$coef,4)
```

Multivariable logistic regression

- **Simple logistic regression**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.4269	0.0989	4.3175	0
childAge	-0.8410	0.1230	-6.8355	0

- **Multivariable logistic regression**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2608	0.1066	2.4458	0.0145
childAge	-0.8680	0.1243	-6.9850	0.0000
motherAnemic	0.5143	0.1240	4.1486	0.0000

Multivariable logistic regression

Fit a logistic regression model to investigate associations between child anaemia and child age accounting for mother's anaemic status and its **interaction** with child age

$$\text{Logit}(\pi(Child\ Anaemia)) =$$
$$\beta_0 + \beta_1 ChildAge + \beta_2 MotherAnemic$$
$$+ \beta_3 ChildAge * MotherAnemic$$

Multivariable logistic regression

- **Model**

fit3 <- glm(y~childAge+motherAnemic+childAgemotherAnemic ,

　　　　family=binomial(link=logit) ,data=anemic)

Fit3Coef <- round(summary(fit1)$coef,4)

- **Multivariable logistic regression**

	Estimate	Std. Error		z value	Pr(>|z|)
(Intercept)	0.3704	0.1207	3.0681	0.0022	
childAge	-1.0454	0.1537	-6.8035	0.0000	
motherAnemic		0.1702	0.2108	0.8074	0.4194
childAge:motherAnemic		0.5189	0.2599	1.9965	0.04590

Multivariable logistic regression

- **Model1: No Interaction**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.2608	0.1066	2.4458	0.0145
childAge	-0.8680	0.1243	-6.9850	0.0000
motherAnemic	0.5143	0.1240	4.1486	0.0000

- **Model2: With Interaction – BEST MODEL???**

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3704	0.1207	3.0681	0.0022
childAge	-1.0454	0.1537	-6.8035	0.0000
motherAnemic	0.1702	0.2108	0.8074	0.4194
childAge:motherAnemic	0.5189	0.2599	1.9965	0.0459

Multivariable logistic regression

A study investigates perception of people on the role of women in the society. The question is whether people agree or disagree that "women should take care of the homes and leave running the country up to men". Is there an association between year of education of participant and the probability to agree with the statement?

Multivariable logistic regression

Model0: $\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.5033	0.1784	14.0298	0
Years	-0.2707	0.0154	-17.5614	0

Multivariable logistic regression

$$\text{Model1: Logit}(\pi(agree)) = \beta_0 + \beta_1 Year + \beta_1 Year^2 + \beta_3 Sex +$$

$$\beta_4 Year * Sex + \beta_5 Year^2 * Sex$$

	Estimate	Std. Error	z value		Pr(>|z|)
(Intercept)	0.6822	1.1963	0.5702	0.5685	
Years	-0.0483	0.2201	-0.2194		0.8263
Sex	0.9405	0.8288	1.1348	0.2565	
Year2	-0.0048	0.0100	-0.4755		0.6344
Years:Sex	-0.0904	0.1526	-0.5923		0.5536
Sex:Year2	0.0004	0.0070	0.0544		0.9566

No association between probability to agree and the explanatory variables?

Model selection

Sometimes lack of significance is not the same thing as lack of importance. To objectively choose between models, we need an objective criterion that takes into account a measure of:

- ➤ Goodness of fit
- ➤ Model complexity

Goodness of fit

Goodness of fit of a model implies how well does the specified model explains the data. This can be quantified in terms of the log-likelihood of the model.

$$\log \hat{L}_c = \sum_i \left\{ \log \binom{n_i}{y_i} + y_i \log \hat{p}_i + (n_i - y_i)\log(1 - \hat{p}_i) \right\}$$

Goodness of fit

Model 0: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$

$$Log(\hat{L}_0) = \sum_{i=1} log \binom{n_i}{y_i} + y_i log\left(expit(\hat{\beta}_0 + \hat{\beta}_1 X_{i1})\right) + (n_i - y_i)log\left(1 - expit(\hat{\beta}_0 + \hat{\beta}_1 X_{i1})\right)$$

Model 1: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{2i}$

$$Log(\hat{L}_1) = \sum_{i=1} log \binom{n_i}{y_i} + y_i log\left(expit(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{21})\right) + (n_i - y_i)log\left(1 - expit(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{21})\right)$$

Goodness of fit

- **Deviance (D)**

The difference in goodness of fit between two models can be quantified by **Deviance (D)**:

$$D = -2Log(\hat{L}_0) - \left(-Log(\hat{L}_1)\right)$$

$$D = -2Log(\hat{L}_0) + Log(\hat{L}_1)$$

- The model with the bigger likelihood is the better model**???**

Complexity

The complexity of a model can be quantified as the number of parameters (k) to be estimated in the model. Alternative this can also be defined as the degree of freedom.

Model 0: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$

$$k_0 = 2$$

Model 1: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{2i}$

$$k_1 = 3$$

Model selection

- Model chosen based only goodness of fit will be biased towards the model with more parameters. **Why???**

- Model chosen based only complexity will over penalised the model with more parameters. **Why???**

Model selection

- Akaike Information Criterion (AIC)

$$\text{AIC} = -2Log(L) + \alpha p$$

Model 0: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$

$$AIC_0 = -2Log(\hat{L}_0) + 2\alpha$$

Model 1: $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{2i}$

$$AIC_1 = -2Log(\hat{L}_1) + 3\alpha$$

- Smaller AIC means better model. **Why ???**

Model selection

- **Likelihood Ratio Test**

The main challenge with using AIC and Deviance for model selection is how to answer the question of "How much small is small?" or "How much large is large". A better approach for nested models is to formally test the importance of the difference factors to be excluded from the a model.

Model selection

- **Likelihood Ratio Test**

Full Model (F): $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{2i}$

Reduced Model (R): $\text{Logit}(\hat{\pi}_i(x)) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$

F

| Intercept | X_{i1} | X_{2i} |

R

| Intercept | X_{i1} | ??? |

$$D = F - R \sim X^2(df_F - df_R)$$

Model selection

Fit logistic a regression model to investigate associations between child anaemia and child age accounting for mother's anaemic status and its **interaction** with child age.

- **Do we need interaction in this model?**

Model selection

- **Full Model**

$$\text{Logit}(\pi(Child\ Anaemia)) =$$

$$\beta_0 + \beta_1 ChildAge + \beta_2 MotherAnemic$$

$$+ \beta_3 ChildAge * MotherAnemic$$

- **Reduced Model**

$$\text{Logit}(\pi(Child\ Anaemia)) = \beta_0 + \beta_1 ChildAge + \beta_2 MothersAge$$

Model selection

- **Code**

```
Dev <- anova(fit1,fit2)[2,4]

Dev

pvalue <- round(pchisq(Dev[2,4],Dev[2,3],lower.tail=FALSE),4)

pvalue
```

- **Deviance**

	Resid. Df	Resid. Dev	Df	Deviance	**Pvalue**
1	1204	1604.2			
2	1203	1600.2	1	3.9552	**0.0467**

Model selection

A study investigates perception of people on the role of women in the society. The question is whether people agree or disagree that "women should take care of the homes and leave running the country up to men". Model the association between year of education of participant and the probability to agree with the statement, account for other potential predictors.

Model selection

MODEL	Parameterization
0	$\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year$
1	$\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year + \beta_3 Sex$
2	$\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year + \beta_3 Sex + \beta_4 Year * Sex$
3	$\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year + \beta_1 Year^2 + \beta_3 Sex + \beta_4 Year * Sex + \beta_5 Year^2 * Sex$

Model selection

- Deviance, degree of freedom and AIC for the different models

MODEL	df	Deviance	AIC
0	39	64.03	206.1
1	38	64.01	208.1
2	37	57.1	203.2
3	35	55.49	205.5

Model selection

- Likelihood ratio test model 2 vs 3

Model	Resid. Df	Resid. Dev	Df	Deviance	**Pvalue**
2	37	57.103			
3	35	55.487	2	1.615	**0.4459**

- Likelihood ratio test model 1 vs 2

Model	Resid. Df	Resid. Dev	Df	Deviance	**Pvalue**
1	38	64.007			
2	37	57.103	1	6.9039	0.0317

Model selection

- The most parsimonious model for the data is:

$$\text{Logit}(\pi(agree)) = \beta_0 + \beta_1 Year + \beta_3 Sex + \beta_4 Year * Sex$$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1935	0.5441	2.1935	0.0283
Years	-0.1526	0.0468	-3.2622	0.0011
Sex	0.9047	0.3601	2.5127	0.0120
Years:Sex	-0.0814	0.0311	-2.6175	0.0089

- **Interpret the results???**

Practical IVa: Child Anaemia data

- Fit a logistic regression model to investigate associations between mother anaemia status and mother's age accounting for other risk factors as well as potential interactions. Perform likelihood ratio test to justify your most parsimonious model.

- **Interpret your results.**

Practical IVb: esrData

- Fit a logistic regression model to investigate the association between disease state (y =1 if ESR level > 20; 0 otherwise) and the two protein. Perfom likelihood ratio test to just you rmost parsimonious model.

- **Interprete your results**