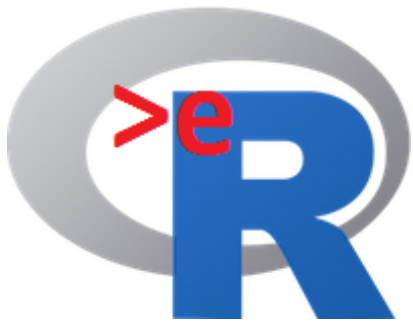




This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

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



The >eR-Biostat initiative
Making R based education materials in
statistics accessible for all

Basic concepts in statistical modeling using R: Simple Logistic Regression

Developed by

Legesse Kassa Debusho (UNISA, South Africa), Ziv Shkedy (Hasselt University, Belgium)
and

Tadele Worku Mengesha (Gondar University), Abdisa Gurmessa (Jmma University)

LAST UPDATE: 15/10/2017



ER-BioStat



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

Email: erbiostat@gmail.com



@erbiostat

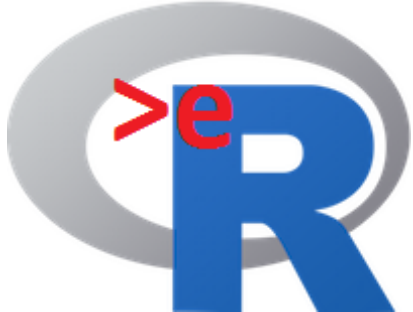


contents

- Logistic regression:
 - Examples.
 - The `glm()` function in R.
 - Fitting logistic regression models using the `glm()` function in R: 5 examples.

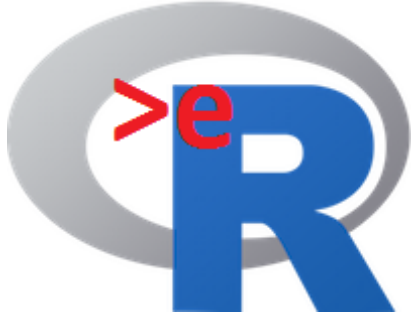
YouTube tutorials

- YouTube tutorials are available for:
 - Logistic Regression using R | Data Science | Machine Learning (host by Analytics University) :
<https://www.youtube.com/watch?v=nubin7hq4-s>
 - Logistic Regression Analysis in R (host by Dr. Bo Han) :
<https://www.youtube.com/watch?v=eScK5w5JcHI>
 - Statistics with R: Example of logistic regression (host by Phil Chan)
<https://www.youtube.com/watch?v=xElScuasns>



R program and Datasets

- Simple linear regression:
 - Introduction and model formulation.
 - Fitting a simple linear regression model using the `lm()` function in R.
 - Model diagnostic.
 - Model diagnostic in R.



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

The response variable

A binary variable:

$$Y_i = \begin{cases} 1 & \text{presence} \\ 0 & \text{absence} \end{cases}$$

A example:

$$Y_i = \begin{cases} 1 & \text{Diabetes} \\ 0 & \text{Healthy} \end{cases}$$

Bernoulli random variables

- Let Y_1, Y_2, \dots, Y_N represent a sample of Bernoulli random variables from N trials.

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

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

The predictor(s)

Our aim is to model the dependence of the probability of success upon known predictors.

$$Y_i = \begin{cases} 1 & \text{presence} \\ 0 & \text{absence} \end{cases} \quad \Rightarrow \quad P(Y_i = 1) = P(Y_i = \text{presence}) = P(\text{success})$$

$$P(Y_i = 1) = f(\text{predictors}) = f(X_1, X_2, \dots)$$

Logistic regression model

Our aim is to model the dependence of the probability of success on known predictors.

Example:

$$Y_i = \begin{cases} 1 & \text{Diabetes} \\ 0 & \text{Healthy} \end{cases}$$

$$P(Y_i = \text{Diabetes}) = f(\text{predictors}) = f(\text{diet}, \text{age}, \dots)$$

The model that we use to model the dependence between diabetes and the predictors is **logistic regression model**.



Examples

Example 1: Smoked mice

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

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

The second group (the control group) were kept in their chambers for 12 hours without smoke.

After One year an autopsy was carried out.

The response is the presence and absence of a tumour.

The second variable in the data is the treatment group.

Smoked mice: the response variable

The question of primary interest is:

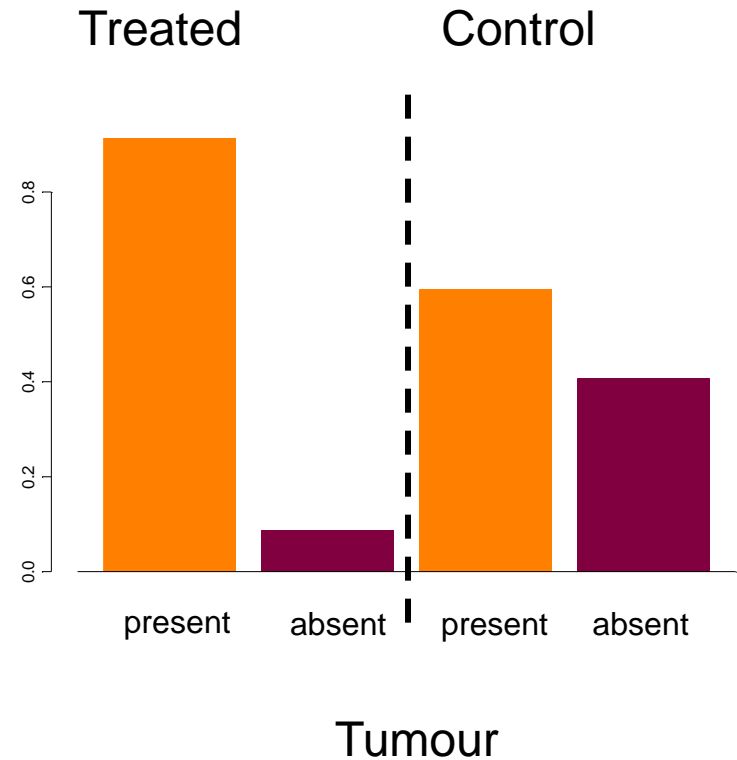
DOSE THE SMOKE INCREASE THE RISK FOR
CANCER ?

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

The response variable

Smoked mice: the data

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



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.

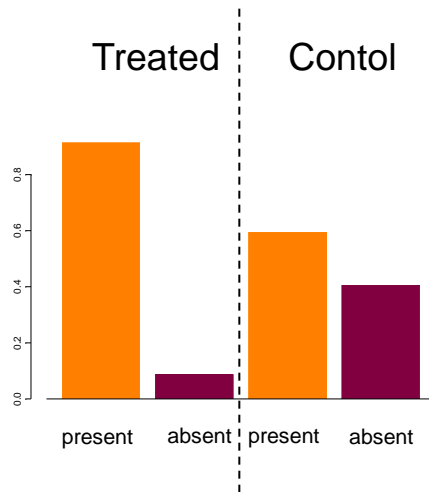
We do not have information about individuals in the sample, but only about the counts in different combinations of the experimant.

Individual data can be extracted from the table.

In terms of statistical modeling, the response is binary (tumour absent/tumour presnt).

The predictor, the treatment group, is also binary.

Response and predictor



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 increases the risk for tumour.

Response:

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

Predictor:

$Treatment_i$ (treated / control)

$$P(Y_i = 1) = P(\text{tumour}) = f(\text{treatment})$$

Example 2: 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.

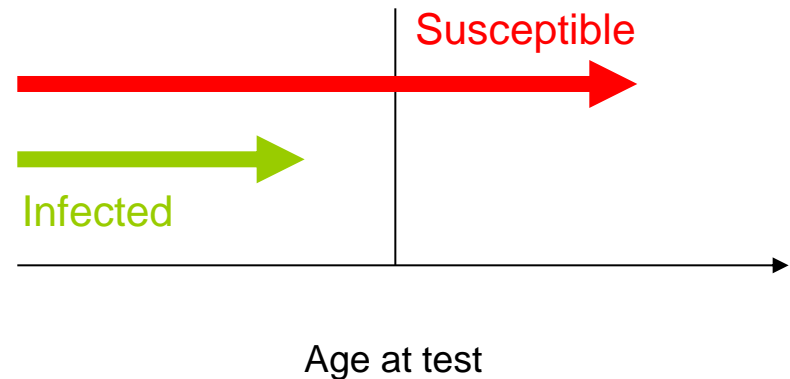
Example 2: Serological data

- 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:

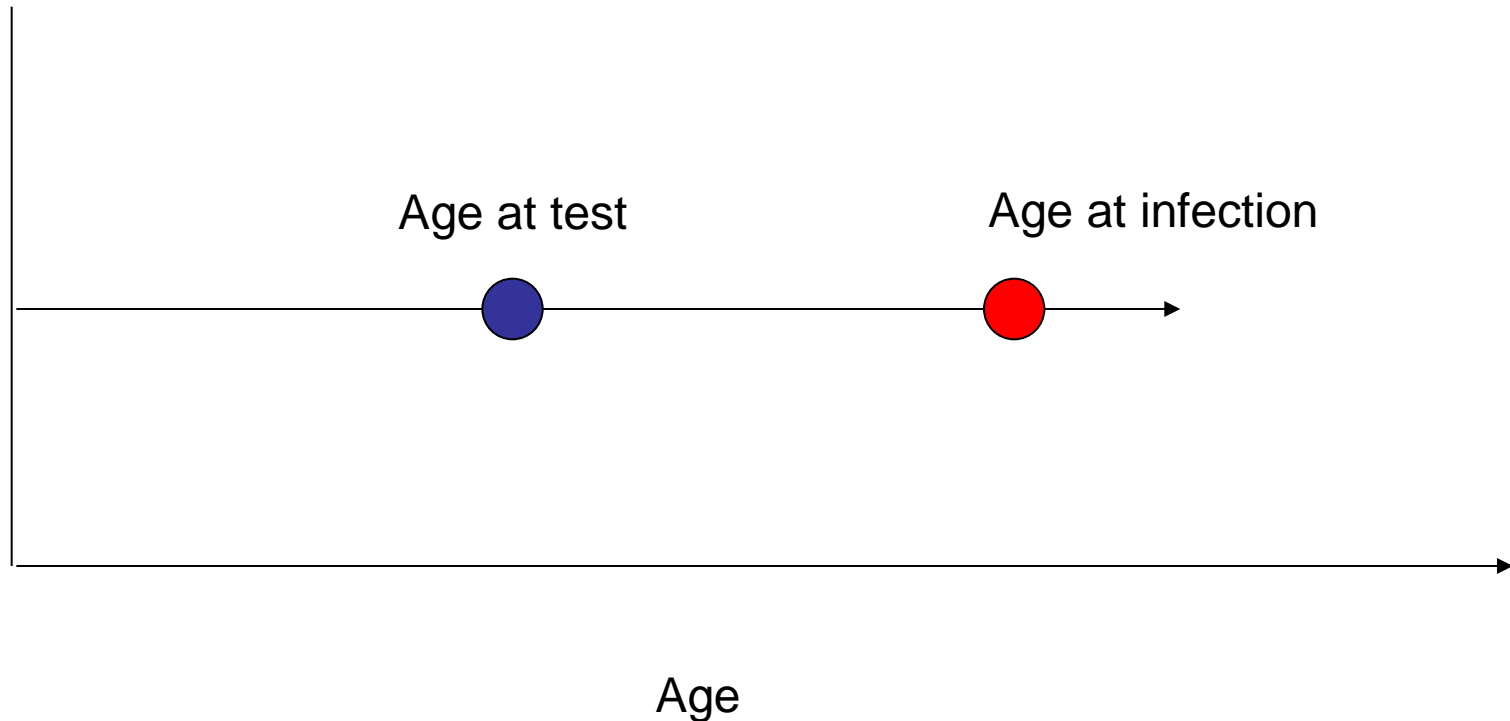
$$\pi(a)$$

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

- Sero-prevalence data

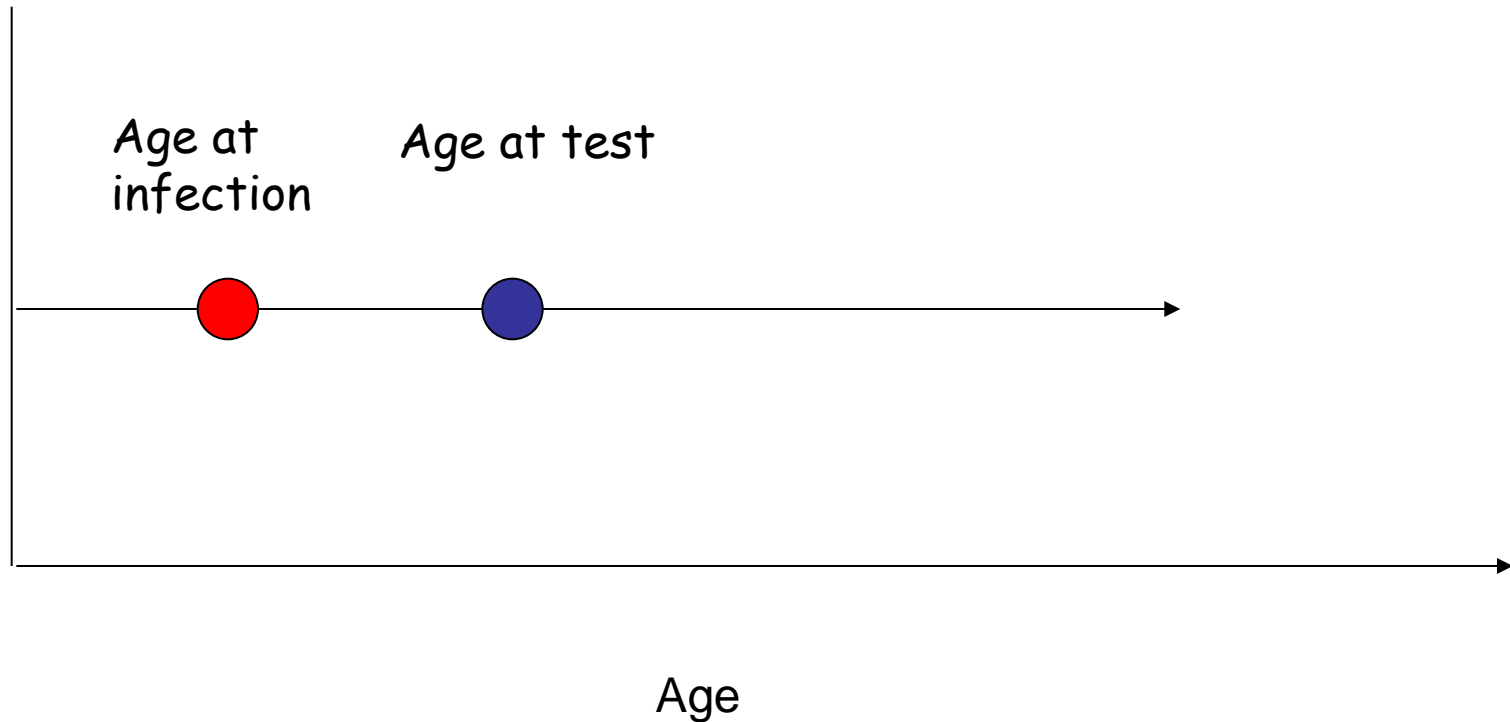


Current status data: sero-negative



- Sero-Negative: infected after the test.

Current status data: sero-positive



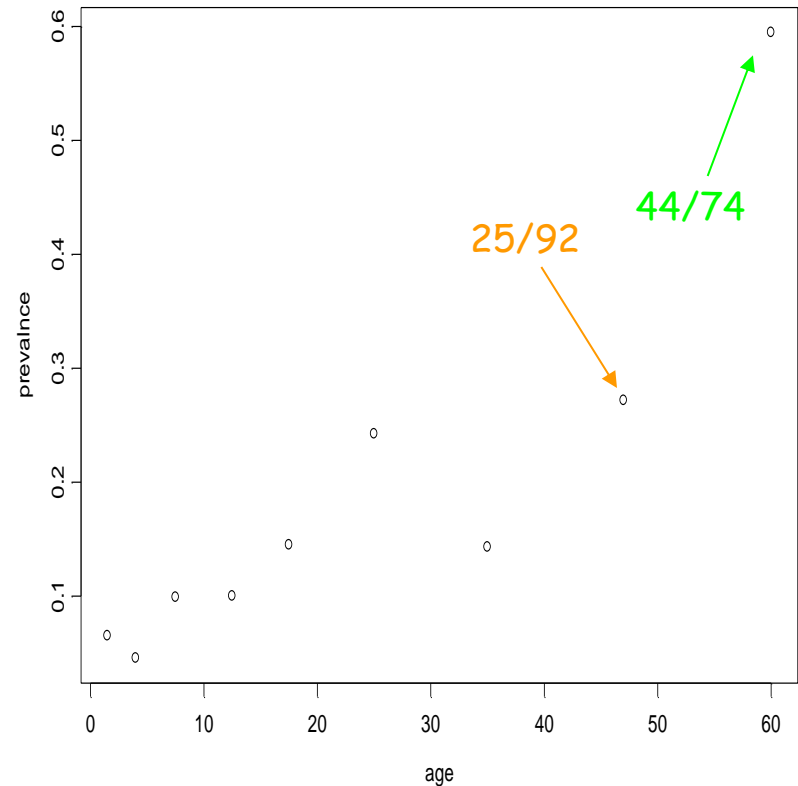
- Sero-Positive: infected after the test.

Example 2: Serological data

Malaria in Brasil

| 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 |

What is the relationship between infection and age ?



Example 2: 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 |

Response:

$$Y_{ij} = \begin{cases} 1 & \text{Sero +} \\ 0 & \text{Sero -} \end{cases}$$

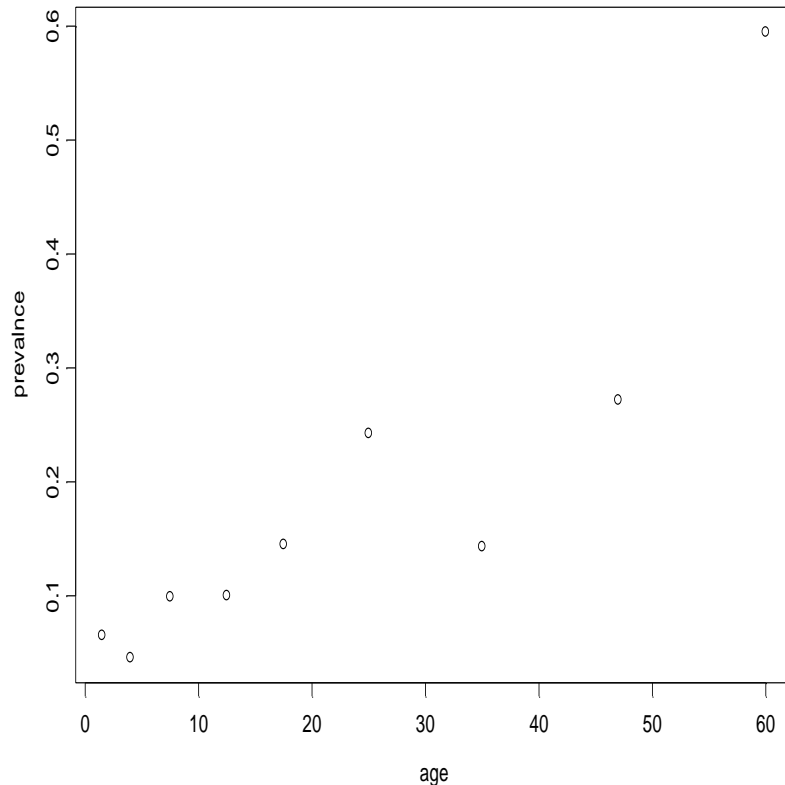
Number of Sero+ in age group i:

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Sample size at age group i:

$$n_i$$

Example 2: Serological data



Response: number of infected (sero+):

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Predictor: age

$$P(Y_i = 1) = P(\text{sero+}) = f(\text{age})$$

Example 3: Bioassay

A bioassay experiment is an experiment designed to assess the potency of a compound by means of the response produced when it is administered to a living organism.

In this example the protective effect of a particular serum (serum 32) on the bacterium associated with the occurrence of pneumonia is under investigation.

Study design:

The experiment consists 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.

Bioassay data: response and predictor

The response of the number of deaths within 7 days from injection.

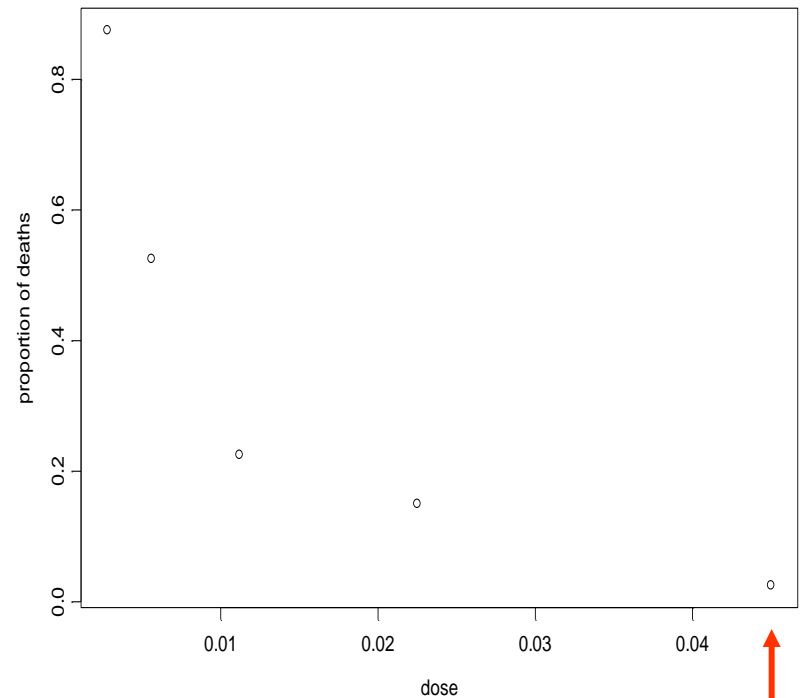
The dose level is the predictor.

The question of primary interest:

What is the relationship between the injected dose
and the number of deaths ?

Example 3: the data

| 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 |



Example 3: the data

| 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 |

Response:

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

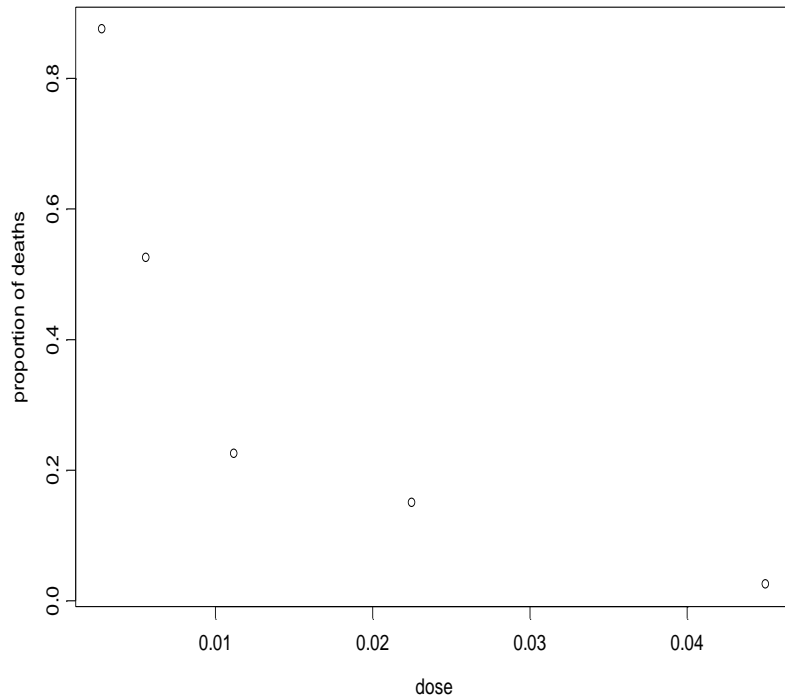
Number of deaths in dose level i:

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Sample size at dose level i:

$$n_i$$

Example 3: response and predictor



Response: number of deaths at each dose level:

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Predictor: dose

$$P(Y_i = 1) = P(\text{death}) = f(\text{dose})$$

Example 4: Determination of ESR

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells settle out of suspension in blood plasma when measured under standard condition.

The ESR increase if the levels of certain proteins in the blood increase.

Rheumatic diseases, chronic diseases and infections increase these proteins level.

From that reason the determination of the ESR is one of the most commonly used screening tests performed on samples bloods.

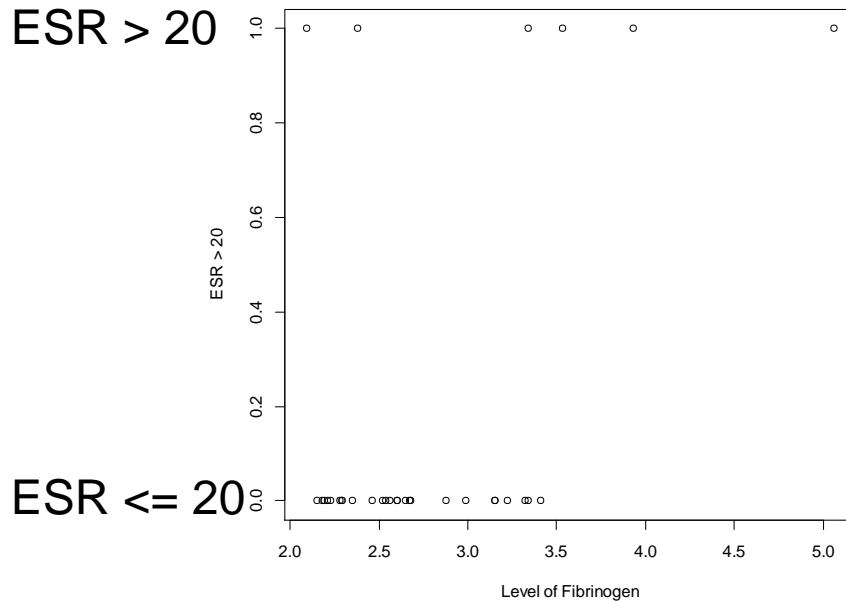
Determination of ESR: The data

| Individual | Fib | Glob | Y |
|------------|---------|------|---|
| 1 | 1 2.52 | 38 | 0 |
| 2 | 2 2.56 | 31 | 0 |
| 3 | 3 2.19 | 33 | 0 |
| 4 | 4 2.18 | 31 | 0 |
| 5 | 5 3.41 | 37 | 0 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| 19 | 19 2.60 | 38 | 0 |
| 20 | 20 2.23 | 37 | 0 |
| 21 | 21 2.88 | 30 | 0 |
| 22 | 22 2.65 | 46 | 0 |
| 23 | 23 2.09 | 44 | 1 |
| 24 | 24 2.28 | 36 | 0 |
| 25 | 25 2.67 | 39 | 0 |
| 26 | 26 2.29 | 31 | 0 |
| 27 | 27 2.15 | 31 | 0 |
| 28 | 28 2.54 | 28 | 0 |
| 29 | 29 3.93 | 32 | 1 |
| 30 | 30 3.34 | 30 | 0 |
| 31 | 31 2.99 | 36 | 0 |
| 32 | 32 3.32 | 35 | 0 |

An example of individual data. For each subject we have the response and the proteins level.

Does the Fibrinogen level (proteins in the blood) influence the ESR rate ?

Example 4: determination of ESR



Response:

$$Y_i = \begin{cases} 1 & ESR > 20 \\ 0 & ESR \leq 20 \end{cases}$$

Predictor: Fibrinogen level.

$$P(Y_i = 1) = P(ESR > 20) = f(\text{Fibrinogen level})$$

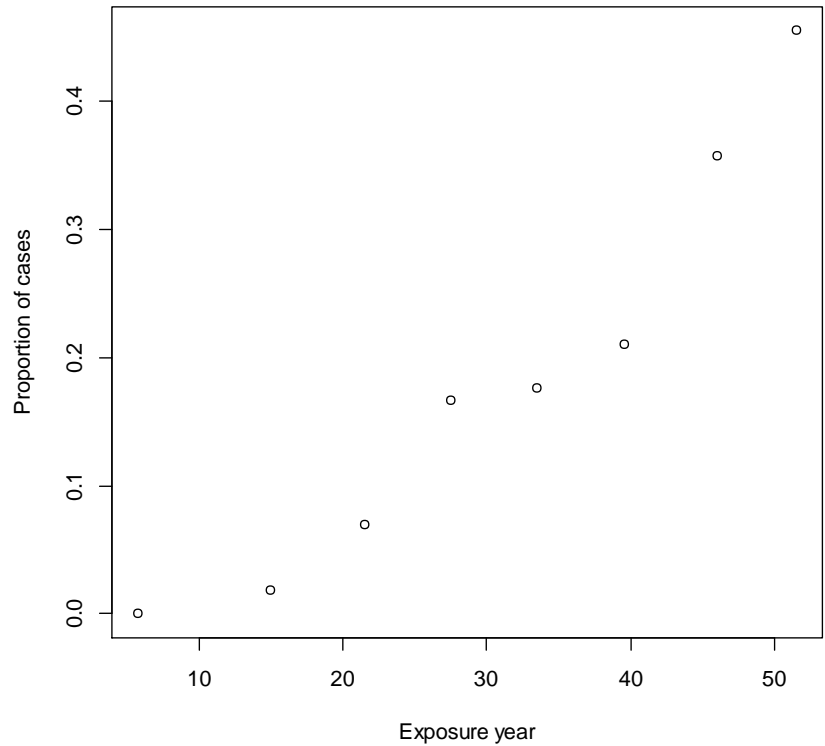
Example 5: Pneumoconiosis amongst coal miners

Pneumoconiosis amongst groups of coal miners with varying exposure to coal dust.

Does exposure time increase the probability to have the disease ?

The data

| | Years | Cases | Miners |
|---|-------|-------|--------|
| 1 | 5.8 | 0 | 98 |
| 2 | 15.0 | 1 | 54 |
| 3 | 21.5 | 3 | 43 |
| 4 | 27.5 | 8 | 48 |
| 5 | 33.5 | 9 | 51 |
| 6 | 39.5 | 8 | 38 |
| 7 | 46.0 | 10 | 28 |
| 8 | 51.5 | 5 | 11 |



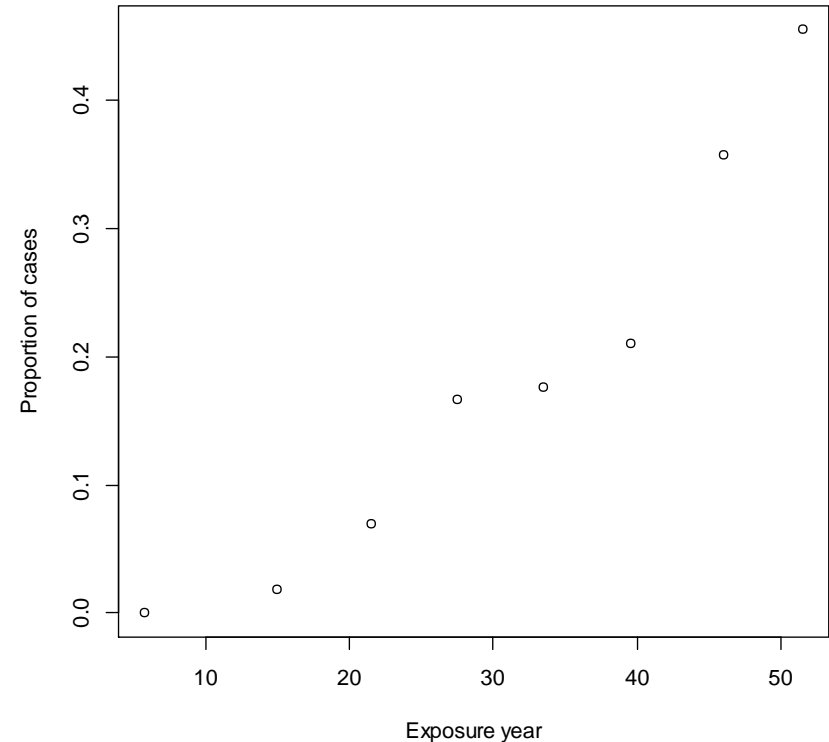
Example 5: response and predictor

Response:

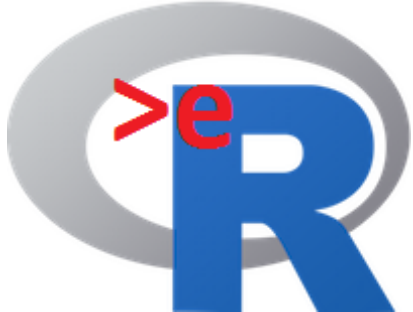
$$Y_{ij} = \begin{cases} 1 & \text{Pneumoconiosis} \\ 0 & \text{healthy} \end{cases}$$

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Predictor: years of exposure to coal dust.



$$P(Y_i = 1) = P(\text{Pneumoconiosis}) = f(\text{time})$$



Fitting logistic regression models using the
`glm()` function in R

The glm() Function in R

- Generalized linear models can be fitted in R using the `glm()` function, which is similar to the `lm()` function for fitting linear models.
- Arguments in the `glm()` call are as follows:

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

The glm() Function in R

- For binary data, the general call of the glm() function has the form

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




this defines a logistic regression model, i.e. a model for binary data with logit link function.

The glm() Function: zero/one data.

- For a zero/one data (for example the ESR data):

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



response~predictor 1 + predictor 2+ \top .

The glm() Function: grouped data

- For grouped data (for example, the serological data)

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

`positive/sample size ~ predictor 1 + predictor 2 +`

Number of successes

Sample size in the
category



Fitting logistic regression models using `glm()`
function in R: 4 examples

Example 1: Smoked mice

The question of primary interest is:

DOSE THE SMOKE INCREASE THE RISK FOR
CANCER ?

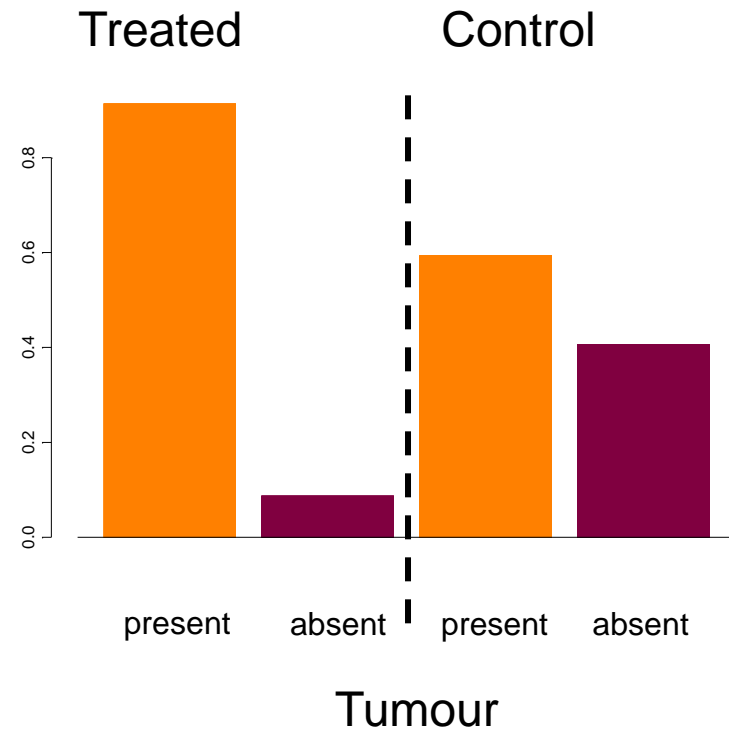
$$Y_i = \begin{cases} 1 & \text{tumour present} \\ 0 & \text{tumour absent} \end{cases}$$

↙
The response variable

Data structure in R

```
> mice <- data.frame(Treatm=c("Treated", "Control"),  
+                     Tumour = c(21,19), Total = c(23,32))  
> attach(mice)  
> mice
```

| | Treatm | Tumour | Total |
|---|---------|--------|-------|
| 1 | Treated | 21 | 23 |
| 2 | Control | 19 | 32 |



Model formulation

The individual data

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

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

Number of subjects
with tunour

$$Y_i = \sum X_{ij}$$

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

Distribution of Y

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

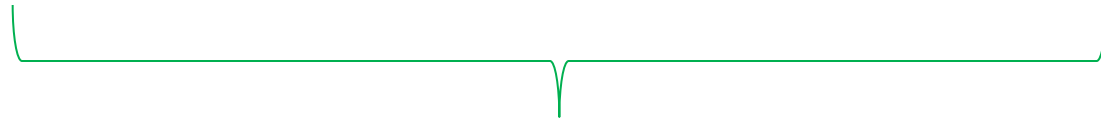
The model for P

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

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

Fitting the model with the glm() function:

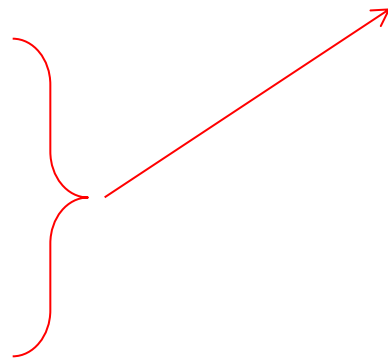
```
> fit2.mice <- glm(cbind(Tumour , Total-Tumour)~factor(Treatm),  
                  data = mice, family = binomial("logit"))
```



$$\text{logit}(P_i) = \alpha + \beta \times \text{treatment}$$

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

$$P_i = \frac{e^{\alpha + \beta \times \text{treatment}_i}}{1 + e^{\alpha + \beta \times \text{treatment}_i}}$$



R output

```
> summary(fit2.mice)
```

Call:

```
glm(formula = cbind(Tumour, Total - Tumour) ~ factor(Treatm),  
     family = binomial("logit"), data = mice)
```

Deviance Residuals:

```
[1] 0 0
```

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) |
|-----------------------|----------|------------|---------|----------|
| (Intercept) | 0.3795 | 0.3599 | 1.054 | 0.2917 |
| factor(Treatm)Treated | 1.9719 | 0.8229 | 2.396 | 0.0166 * |

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7.6349 on 1 degrees of freedom
Residual deviance: 0.0000 on 0 degrees of freedom
AIC: 10.421

Number of Fisher Scoring iterations: 4

The odds ratio

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

$$OR = \frac{21 \times 13}{19 \times 2}$$

```
> OR1 <- (21*13) / (19*2)
```

```
> OR1
```

```
[1] 7.184211
```

```
> log(OR1)
```

```
[1] 1.971886
```

```
> summary(fit2.mice)$coeff
```

| | Estimate | Std. Error | z value | Pr(> z) |
|-----------------------|-----------|------------|----------|-----------|
| (Intercept) | 0.3794896 | 0.3599370 | 1.054322 | 0.2917354 |
| factor(Treatm)Treated | 1.9718856 | 0.8229056 | 2.396248 | 0.0165639 |

$$\hat{\beta} = \log(OR)$$

$$OR = \exp(1.971886) = 7.184.$$

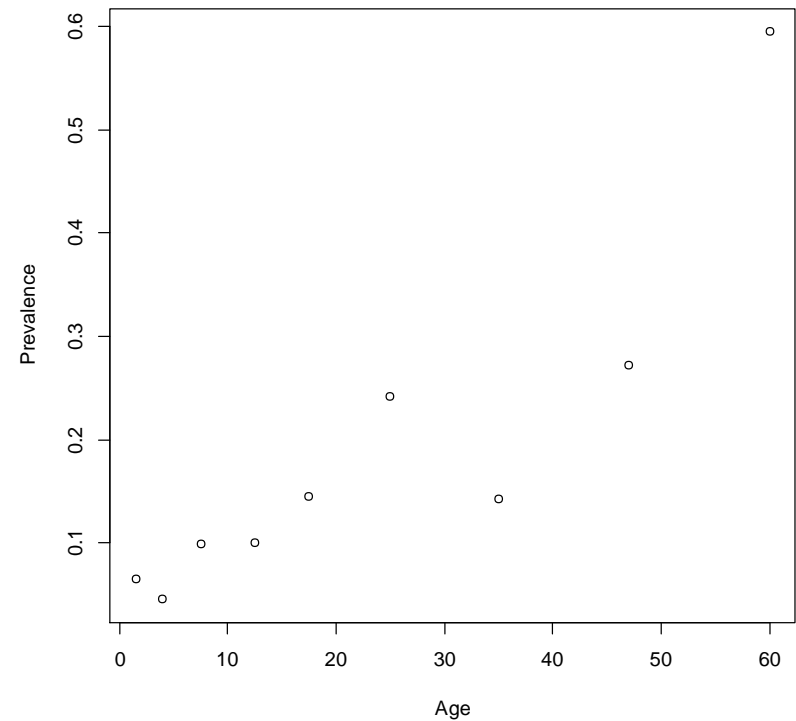
Example 2 (Serological data): Data structure in R

```
Serolog <- read.table('c:/... /Serological.txt',  
+                     header = TRUE, na.strings = "NA", dec = ".")  
> attach(Serolog)  
> print(Serolog)
```

| | Age | N | pos |
|---|------|-----|-----|
| 1 | 1.5 | 123 | 8 |
| 2 | 4.0 | 132 | 6 |
| 3 | 7.5 | 182 | 18 |
| 4 | 12.5 | 140 | 14 |
| 5 | 17.5 | 138 | 20 |
| 6 | 25.0 | 161 | 39 |
| 7 | 35.0 | 133 | 19 |
| 8 | 47.0 | 92 | 25 |
| 9 | 60.0 | 74 | 44 |

Example 2: Serological data

```
p <- pos/N  
plot(p ~ Age, xlab = "Age",  
      ylab = "Prevalence")
```



Model formulation

| 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 |

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

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Number of sero-positive at each age group

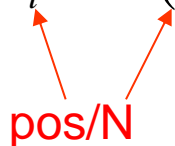
$$Y_i \sim B(n_i, P_i)$$

n_i : sample size at each age group

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

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

glm() function in R

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


pos/N

```
> fit.Sero <- glm(pos/N ~ Age, data = Serolog, family = binomial)
```

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

model pos/N=age



Parameters estimate

```
> summary(fit.Sero)
```

Call:

```
glm(formula = pos/N ~ Age, family = binomial, data = Serolog)
```

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|---------|---------|---------|
| -0.24363 | -0.09726 | 0.01479 | 0.06756 | 0.19568 |

Coefficients:

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.31775 on 8 degrees of freedom

Residual deviance: 0.18094 on 7 degrees of freedom

AIC: 8.0619

Number of Fisher Scoring iterations: 5

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



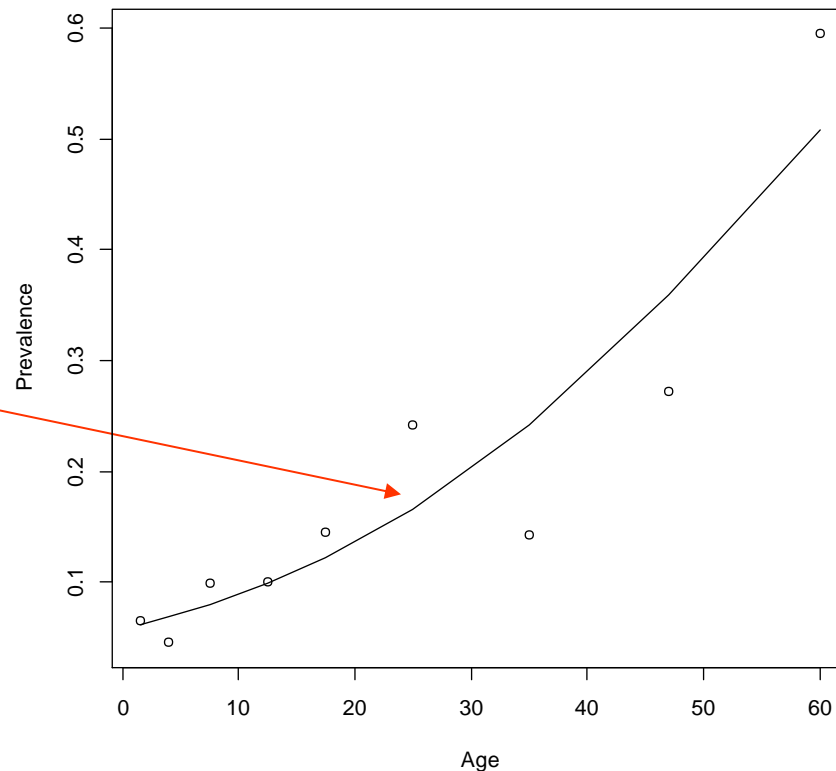
$$\log it(\hat{P}_i) = 2.71 + 0.044 \times age$$

Data and predicted values

```
> p <- pos/N  
> plot(p ~ Age, xlab = "Age", ylab = "Prevalence")  
> lines(Age, fit.Sero$fit)
```

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

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



Example 3: Bioassay

The response of the number of deaths within 7 days from injection.

The dose level is the predictor.

The question of primary interest:

What is the relationship between the injected dose and the number of deaths ?

Data structure in R

```
> serum <- read.table('c:/T.../Serum.txt',  
+   header = TRUE, na.strings = "NA", dec = ".")  
> print(serum)
```

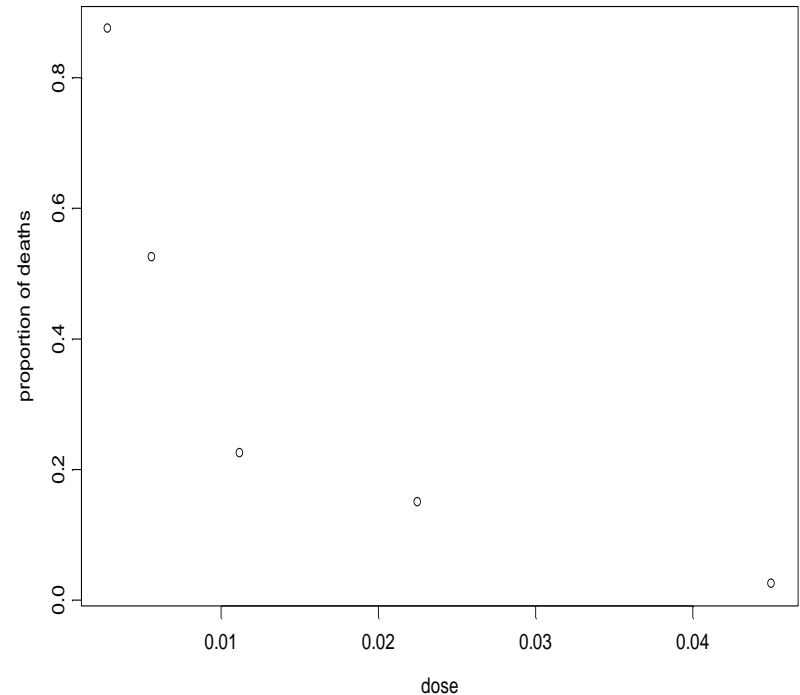
```
      dose death  N  
1 0.0028    35 40  
2 0.0056    21 40  
3 0.0112     9 40  
4 0.0225     6 40  
5 0.0450     1 40
```

| 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 |

The data

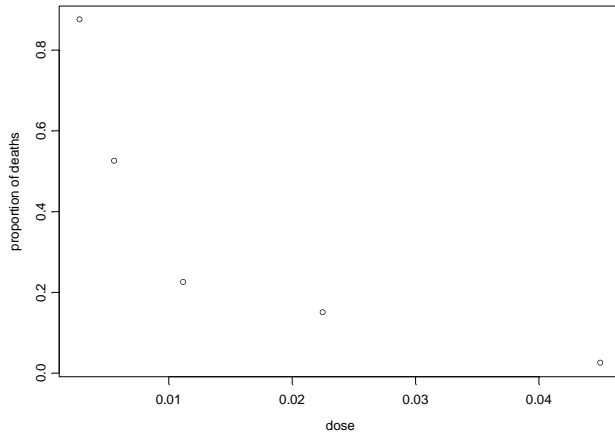
```
> print(serum)
      dose death  N
1 0.0028    35  40
2 0.0056    21  40
3 0.0112     9  40
4 0.0225     6  40
5 0.0450     1  40
```

```
> plot(death/N ~ ldose,
      data = serum, xlab = "Dose",
      ylab = "Proportion of deaths")
```

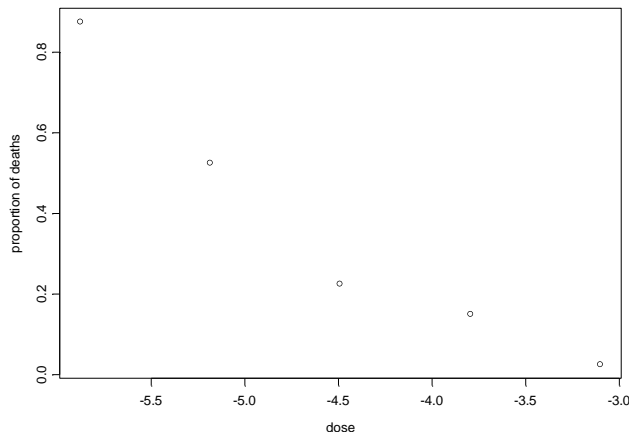


Using log(dose) as predictor

Original scale



Log scale



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

Y: Number of deaths

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

The model is fitted with dose on log scale

$$P_i = \frac{e^{\alpha + \beta \times \log(dose)}}{1 + e^{\alpha + \beta \times \log(dose)}}$$

R script for the model

```
> fit.serum <- glm(death/N ~ ldose, data = serum,  
+ family = binomial)
```

Response:
number of
deaths

Sample size at each
dose level

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

```
print(serum)
  dose death  N
1 0.0028   35 40
2 0.0056   21 40
3 0.0112    9 40
4 0.0225    6 40
5 0.0450    1 40
```

Outout

```
> summary(fit.serum)
```

Call:

```
glm(formula = death/N ~ ldose, family = binomial, data = serum)
```

Deviance Residuals:

| 1 | 2 | 3 | 4 | 5 |
|---------|----------|----------|---------|----------|
| 0.13193 | -0.09818 | -0.11361 | 0.17236 | -0.02366 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|----------|------------|---------|----------|
| (Intercept) | -9.189 | 7.938 | -1.158 | 0.247 |
| ldose | -1.830 | 1.610 | -1.136 | 0.256 |

(Dispersion parameter for binomial family taken to be 1)

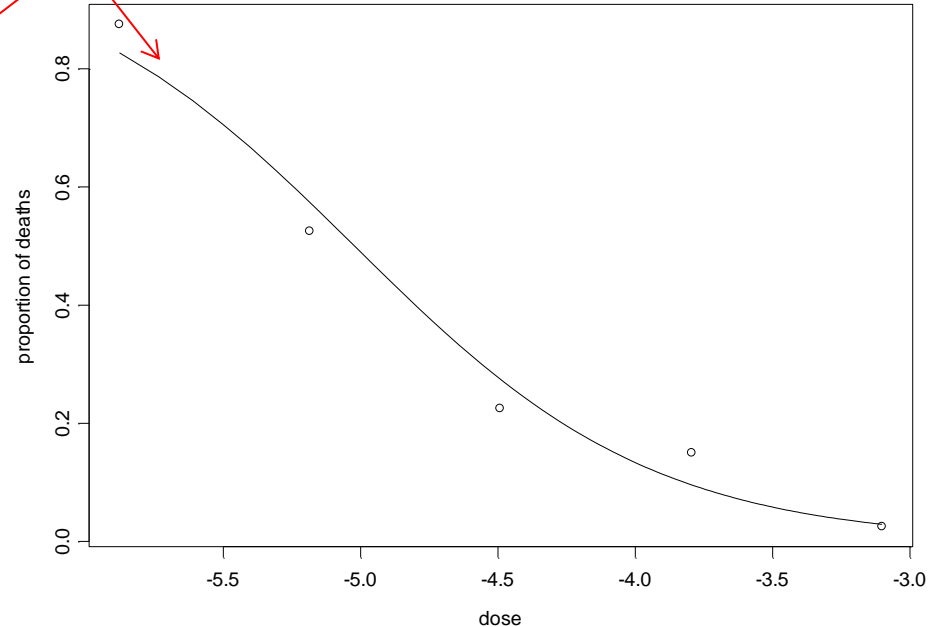
Null deviance: 2.251289 on 4 degrees of freedom
Residual deviance: 0.070222 on 3 degrees of freedom

Data and fitted model

```
> plot(death/N ~ ldose, data = serum, xlab = "Dose",  
      ylab = "Proportion of deaths")  
> lines(serum$ldose, fit.serum$fit)
```

Fitted values:

$$\hat{P}_i = \frac{e^{-9.189 - 1.830 \times \log(dose)}}{1 + e^{-9.189 - 1.830 \times \log(dose)}}$$



ED50

Consider the following logistic regression model:

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

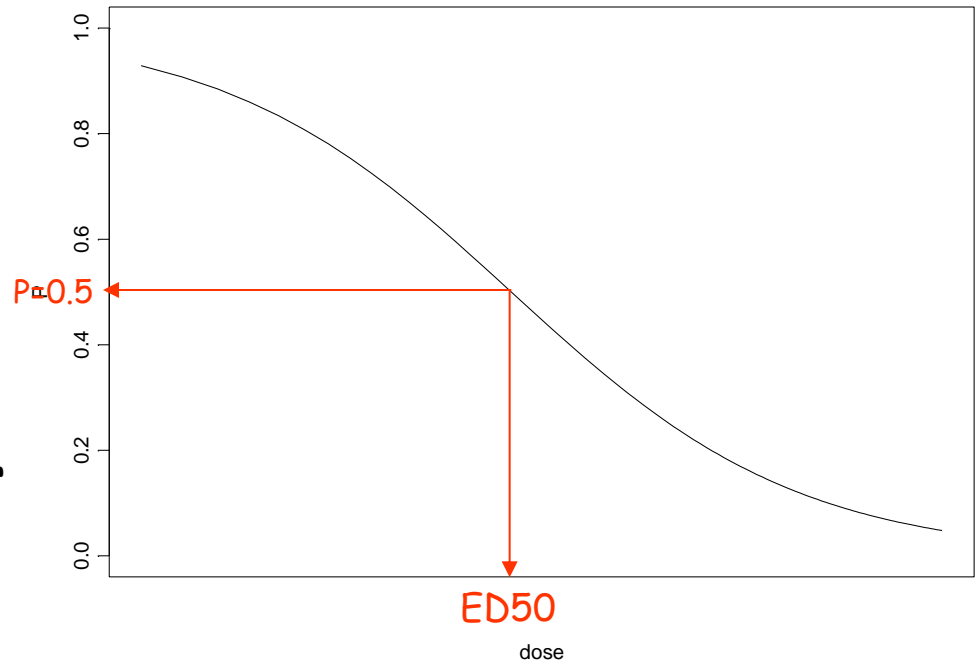
With

$$P_i = \frac{e^{\alpha + \beta \times dose}}{1 + e^{\alpha + \beta \times dose}}$$

The **ED50** is the dose level for which the **probability** for a response is equal to **0.5**, this means that

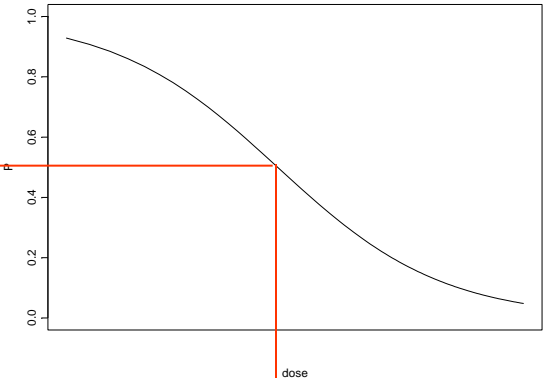
$$0.5 = \frac{e^{\alpha + \beta \times \log(dose)}}{1 + e^{\alpha + \beta \times \log(dose)}}$$

This dose level is the ED50 (on log scale)



How to calculate the ED50 ?

$$0.5 = \frac{e^{\alpha + \beta \times ED50}}{1 + e^{\alpha + \beta \times ED50}} \longleftarrow 0.5 = \frac{e^{\alpha + \beta \times dose}}{1 + e^{\alpha + \beta \times dose}}$$



Logit of 0.5:

$$\log it(0.5) = \log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0$$

Logit of P:

$$\log it(P) = \log\left(\frac{P}{1-P}\right) = \alpha + \beta \times dose$$

For $P=0.5$, $dose=ED50$, this means that

$$\alpha + \beta \times ED50 = 0 \quad \text{.....} \rightarrow \quad ED50 = -\frac{\alpha}{\beta}$$

ED50

Example 4: Determination of ESR

- The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells settle out of suspension in blood plasma when measured under standard condition.
- Response: binary (zero/one).

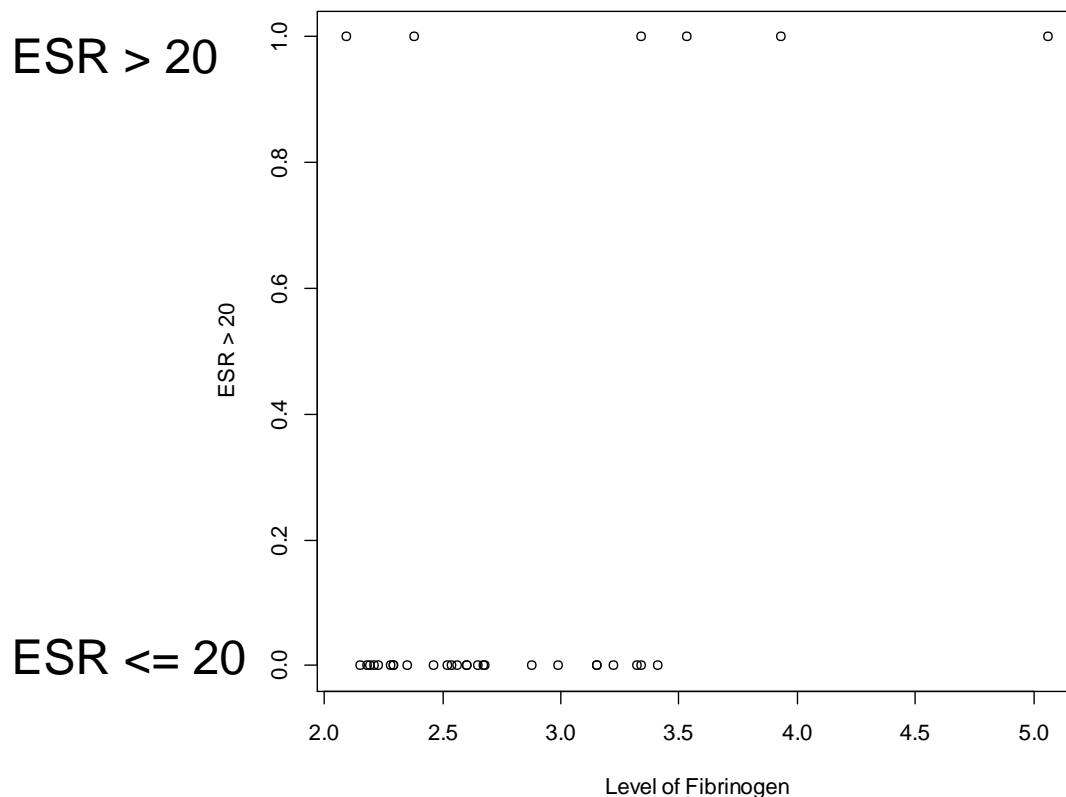
Data structure in R

```
> serum <- read.table('c:/....Serum.txt',  
+   header = TRUE, na.strings = "NA", dec = ".")  
> print(serum)
```

| | dose | death | N |
|---|--------|-------|----|
| 1 | 0.0028 | 35 | 40 |
| 2 | 0.0056 | 21 | 40 |
| 3 | 0.0112 | 9 | 40 |
| 4 | 0.0225 | 6 | 40 |
| 5 | 0.0450 | 1 | 40 |

The data: zero/one data

```
> plot(Y ~ Fib, data = esr, xlab = "Level of Fibrinogen",  
      ylab = "ESR > 20")
```



```
> print(esr)
```

| Individual | Fib | Glob | Y |
|------------|------|------|---|
| 1 | 2.52 | 38 | 0 |
| 2 | 2.56 | 31 | 0 |
| 3 | 2.19 | 33 | 0 |
| . | | | |
| 13 | 5.06 | 37 | 1 |
| 14 | 3.34 | 32 | 1 |
| 15 | 2.38 | 37 | 1 |
| 16 | 3.15 | 36 | 0 |
| 17 | 3.53 | 46 | 1 |
| 18 | 2.68 | 34 | 0 |
| 19 | 2.60 | 38 | 0 |

R script for the model

```
> fit.esr <- glm(Y ~ Fib, data = esr, family = binomial)
```

$$Y_i = \begin{cases} 1 & ESR > 20 \\ 0 & ESR \leq 20 \end{cases}$$

predictor

$Y \sim \text{Fib}$



$$\log \text{it}(P_i) = \alpha + \beta \times \text{Fib}_i$$

R output

```
Call:
glm(formula = Y ~ Fib, family = binomial, data = esr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9298  -0.5399  -0.4382  -0.3356   2.4794

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8451     2.7703  -2.471   0.0135 *
Fib           1.8271     0.9009   2.028   0.0425 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

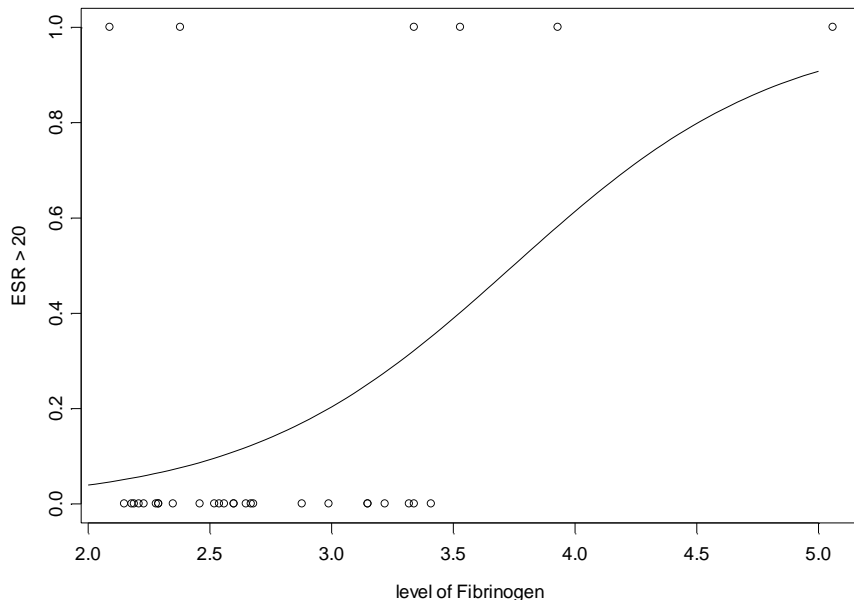
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 24.840  on 30  degrees of freedom
AIC: 28.84

Number of Fisher Scoring iterations: 5
```

Data and fitted model

```
> plot(Y ~ Fib, data = esr, xlab = "Level of Fibrinogen",  
      ylab = "ESR > 20")  
> lines(Fib, fit.esr$fit)
```



$$\hat{P}_i = \frac{e^{\hat{\alpha} + \hat{\beta} \times \text{Fib}_i}}{1 + e^{\hat{\alpha} + \hat{\beta} \times \text{Fib}_i}}$$

```
> summary(fit.esr)$coeff
```

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|-----------|------------|-----------|------------|
| (Intercept) | -6.845075 | 2.7702849 | -2.470892 | 0.01347765 |
| Fib | 1.827081 | 0.9008553 | 2.028162 | 0.04254367 |

$$\hat{\alpha} = -6.845075$$

$$\hat{\beta} = 1.827081$$

Example 5: Pneumoconiosis amongst coal miners

Pneumoconiosis amongst groups of coal miners with varying exposure to coal dust.

Does exposure time increase the probability to have the disease ?

A YouTube tutorial:

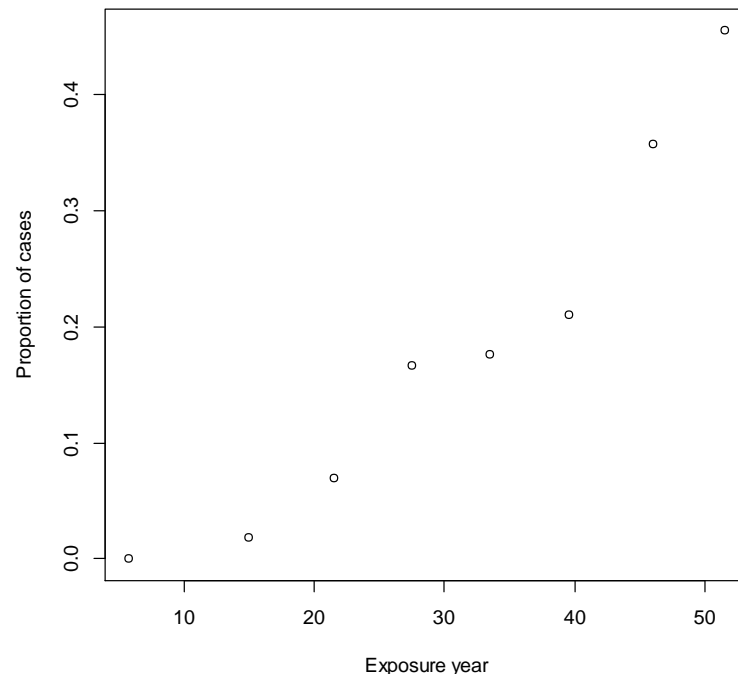
Statistics with R: Example of logistic regression (host by Phil Chan):

<https://www.youtube.com/watch?v=xElScuasns>

Data structure in R

```
> Years<-c(5.8,15.0,21.5,27.5,33.5,39.5,46.0,51.5)
> Cases<-c(0,1,3,8,9,8,10,5)
> Miners<-c(98,54,43,48,51,38,28,11)
> CW<-cbind(Cases,Miners-Cases)
> CW
```

| | Cases | |
|------|-------|----|
| [1,] | 0 | 98 |
| [2,] | 1 | 53 |
| [3,] | 3 | 40 |
| [4,] | 8 | 40 |
| [5,] | 9 | 42 |
| [6,] | 8 | 30 |
| [7,] | 10 | 18 |
| [8,] | 5 | 6 |



```
> plot(Years,Cases/Miners, xlab = "Exposure year", ylab = "Proportion of cases")
```

Variables and model formulation

```
> data.frame(Years,Cases,Miners)
```

| | Years | Cases | Miners |
|---|-------|-------|--------|
| 1 | 5.8 | 0 | 98 |
| 2 | 15.0 | 1 | 54 |
| 3 | 21.5 | 3 | 43 |
| 4 | 27.5 | 8 | 48 |
| 5 | 33.5 | 9 | 51 |
| 6 | 39.5 | 8 | 38 |
| 7 | 46.0 | 10 | 28 |
| 8 | 51.5 | 5 | 11 |

Y_i

n_i

$$Y_{ij} = \begin{cases} 1 & \text{Pneumoconiosis} \\ 0 & \text{healthy} \end{cases}$$

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

Number of infected at each exposure group

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

n_i : sample size at each exposure group

We use logistic regression to model the probability of infection a function of exposure time in years:

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

R script for the model

```
> fit.miners2 <- glm(CW~ Years, family = binomial)
```

```
> CW
Cases
[1,] 0 98
[2,] 1 53
[3,] 3 40
[4,] 8 40
[5,] 9 42
[6,] 8 30
[7,] 10 18
[8,] 5 6
```

CW ~ Years



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

Predictor: exposure time in years

R output

```
> summary(fit.miners2)

Call:
glm(formula = CW ~ Years, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6625  -0.5746  -0.2802   0.3237   1.4852


Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.79648     0.56859  -8.436  < 2e-16 ***
Years         0.09346     0.01543   6.059 1.37e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 56.9028  on 7  degrees of freedom
Residual deviance:  6.0508  on 6  degrees of freedom
AIC: 32.877

Number of Fisher Scoring iterations: 4
```

$$\log it(\hat{P}_i) = \hat{\alpha} + \hat{\beta} \times \text{exp osure}$$


$$\log it(\hat{P}_i) = -4.79648 + 0.09346 \times \text{exp osure}$$

Data and predicted model

```
> plot(Years,Cases/Miners, xlab = "Exposure year",  
       ylab = "Proportion of cases",ylim=c(0,0.6))  
> lines(Years,fit.miners2$fit)
```

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | -4.79648 | 0.56859 | -8.436 | < 2e-16 | *** |
| Years | 0.09346 | 0.01543 | 6.059 | 1.37e-09 | *** |

$$\hat{\alpha} = -4.79648$$

$$\hat{\beta} = 0.09346$$

$$\hat{P}_i = \frac{e^{\hat{\alpha} + \hat{\beta} \times \text{Exposure}_i}}{1 + e^{\hat{\alpha} + \hat{\beta} \times \text{Exposure}_i}}$$

