Computer Intensive Methods using R

Part 5: modeling

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

Overview of the course

- Modeling:
  - Linear regression.
  - Generalized Linear Models.

Overview of the course (part 1)

The Bootstrap algorithm

Introduction:
- Sampling from a population.
- The empirical distribution.
- Plug in principle.

Estimation:
- Accuracy of statistics.
- Confidence intervals.

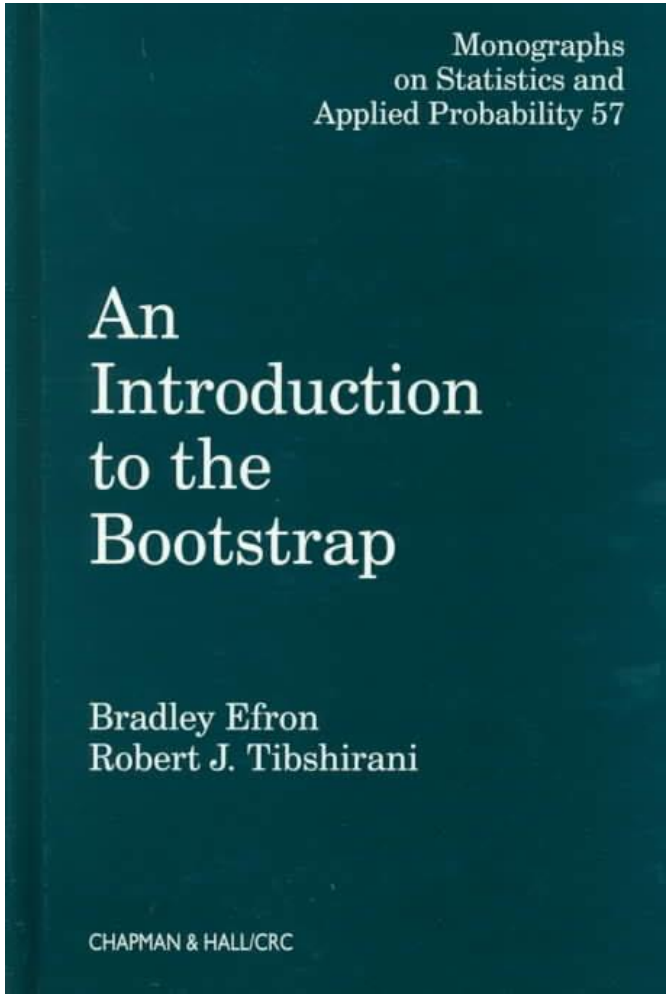
Inference:
- One sample tests.
- Two-samples tests.
- Bootstrap and permutation tests.

Modeling:
- Linear regression models.
- GLMs.

Reference

- Bradley Efron and Robert J. Tibshirani (1994): An introduction to bootstrap.

- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.



Course materials

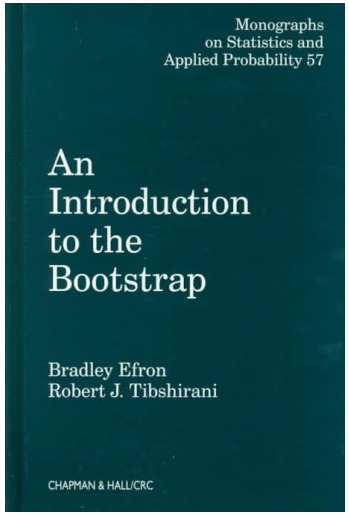
- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

YouTube tutorials

- YouTube tutorials about bootstrap using R:
  1. One-sample bootstrap CI for the mean (host: LawrenceStats): https://www.youtube.com/watch?v=ZkCDYAC2iFg.
  2. Using the non-parametric bootstrap for regression models in R (host:Ian Dworkin):https://www.youtube.com/watch?v=ydtOTctg5So.
  3. Performing the Non-parametric Bootstrap for statistical inference using R (host: Ian Dworkin): https://www.youtube.com/watch?v=TP6r5CTd9yM
  4. Using the sample function in R for resampling of data - absolute basics (host: Ian Dworkin):https://www.youtube.com/watch?v=xE3KGVT6VLE
  5. Permutation tests in R - the basics (host:Ian Dworkin):https://www.youtube.com/watch?v=ZiQdzwB12Pk.
  6. Bootstrap Sample Technique in R software (host: Sarveshwar Inani):https://www.youtube.com/watch?v=tb6wb9ZdPH0
  7. Bootstrap confidence intervals for a single proportion (host: LawrenceStats):https://www.youtube.com/watch?v=ubX4QEPqx5o
  8. Bootstrapped prediction intervals (host:James Scott):https://www.youtube.com/watch?v=c3gD_PwsCGM.

- https://www.youtube.com/watch?v=gcPIyeqymOU

Linear regression



Chapter 9

Topics

- Simple linear regression models:
- Bootstrap algorithms:
  - Non parametric bootstrap.
  - Semi parametric bootstrap.
  - Parametric bootstrap.
- Estimation & C.I.
- Inference.
- Robust regression (the cell datasets)
- Multiple regression models (the tooth strength data).

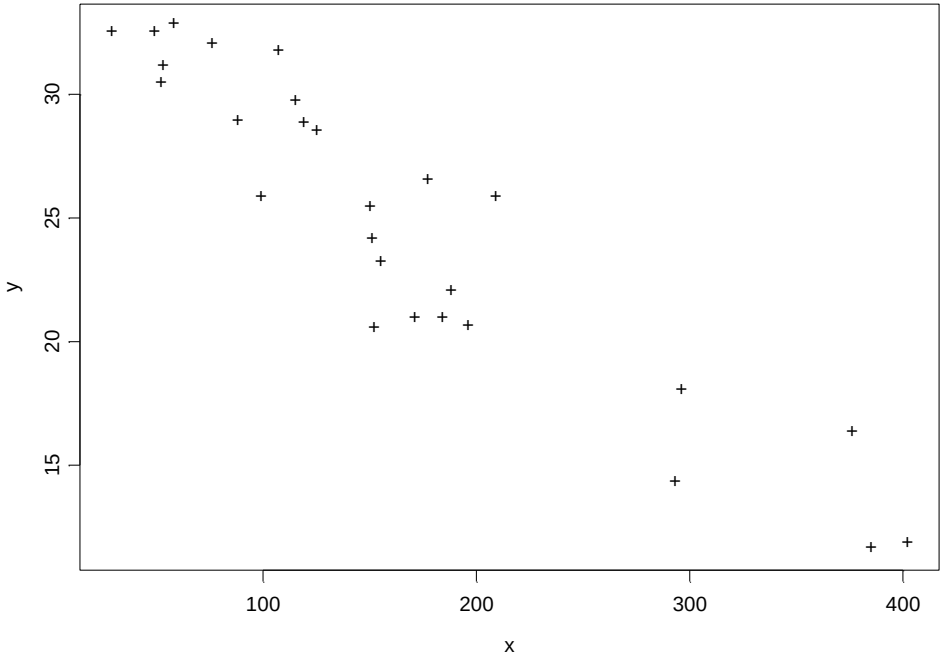
The hormone dataset.

The hormone dataset

- Amount in milligrams of anti-inflammatory hormone remaining in 27 devices after a certain number of hours (hrs) of wear.
- Variables:
  - Hormone level
  - Hours

- In R

```
> help(hormone)
```



Model formulation

We assume that the hormone level (y) is a function of the hours.

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

The hormone mean:

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

Model formulation

Observed data

$$x_1 \quad y_1$$
$$x_2 \quad y_2$$
$$\cdot \quad \cdot$$
$$\cdot \quad \cdot$$
$$x_n \quad y_n$$

The regression model

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

Assumptions ???

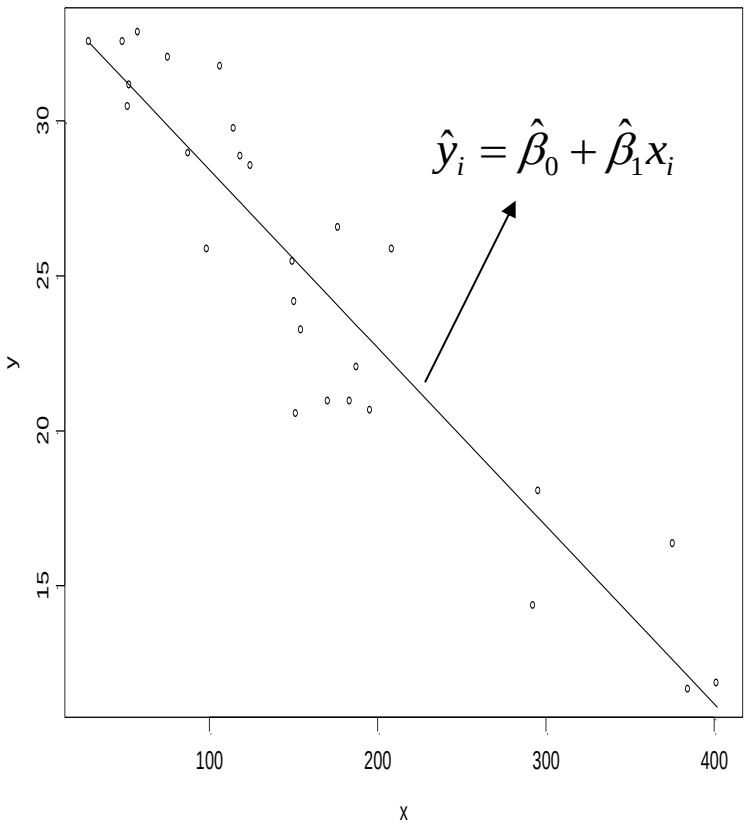
Estimated model and residuals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{e}_i = y_i - \hat{y}_i$$

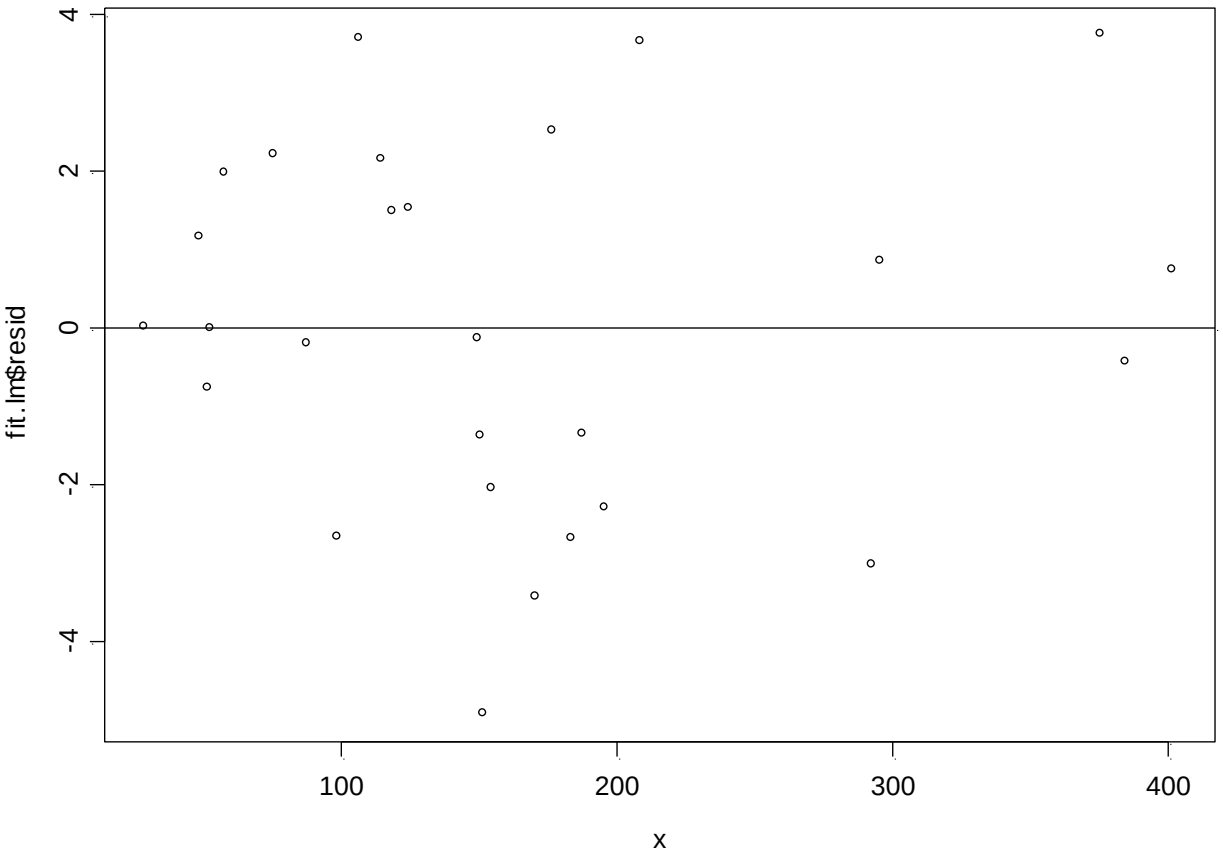
Data and estimated model

```
> fit.lm <- lm(y ~ x)
> summary(fit.lm)

Call: lm(formula = y ~ x)
Residuals:
    Min      1Q   Median     3Q   Max
 -4.936 -1.728 -0.02287 1.739 3.732

Coefficients:
                Value Std. Error  t value Pr(>|t|)
(Intercept)  34.1675   0.8672    39.3999   0.0000
          x  -0.0574   0.0045   -12.8683   0.0000

Residual standard error: 2.378 on 25 degrees of
     freedom
Multiple R-Squared: 0.8688
F-statistic: 165.6 on 1 and 25 degrees of freedom,
     the p-value is 1.584e-012

Correlation of Coefficients:
  (Intercept)
x -0.8494
```



The residuals

$$\hat{e}_i = y_i - \hat{y}_i$$



bootstrapping pairs versus bootstrapping residuals

Data

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ x_n & y_n \end{array}$$

$\downarrow$

Bootstrapping pairs

Estimated model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residuals

$$\hat{e}_i = y_i - \hat{y}_i$$

$\downarrow$

Bootstrapping residuals

The probability model for linear regression

We assume that the hormone level (y) is a function of the hours.

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

The error distribution

$$F \rightarrow (\varepsilon_1, \ldots \varepsilon_n) \qquad E_F(\varepsilon) = 0$$

The hormone mean:

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

The probability model for linear regression (1)

Two components: the unknown parameters and the unknown distribution of the error term.

$$P = (\beta, F)$$

Unknown parameters

Unknown distribution of the error

Assumption: The error between Y and the mean does not depend on X.

The probability model for linear regression (2)

$$P = (\beta, F)$$

Unknown parameters

$$F = F(X, Y)$$

Unknown joint distribution of X and Y

Assumption: No assumption about the error distribution.

bootstrapping pairs versus bootstrapping residuals

$$P = (\beta, F)$$

$$F = F(X, Y)$$

$$F \rightarrow (\varepsilon_1, \ldots \varepsilon_n)$$

Bootstrapping pairs

Bootstrapping residuals

What is the difference ?

Re sampling pairs

Data

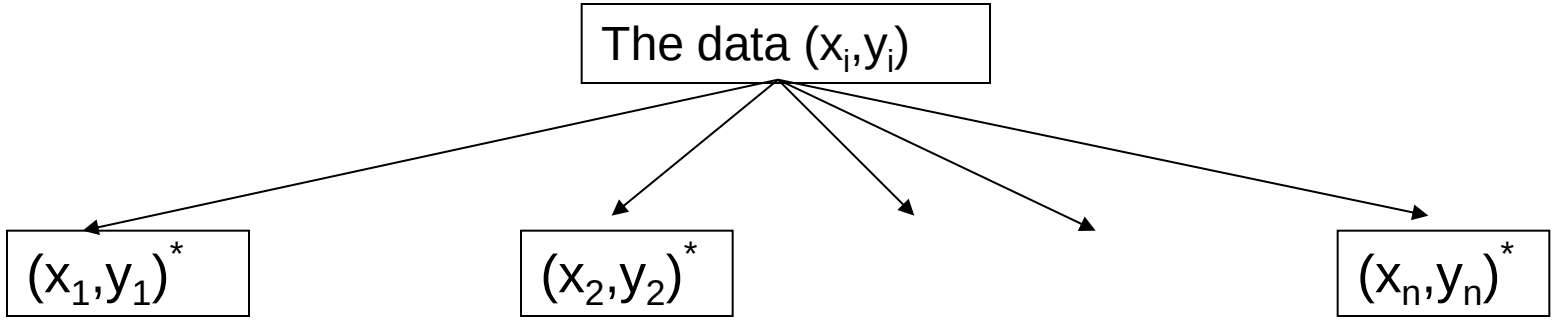
$$\begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ . & . \\ . & . \\ x_n & y_n \end{matrix}$$

Resample pairs with replacement

Bootstrap data

$$\begin{matrix} (x_1 & y_1)^* \\ (x_2 & y_2)^* \\ . & . \\ . & . \\ (x_n & y_n)^* \end{matrix}$$

Bootstrap estimate for $E(y|x)$

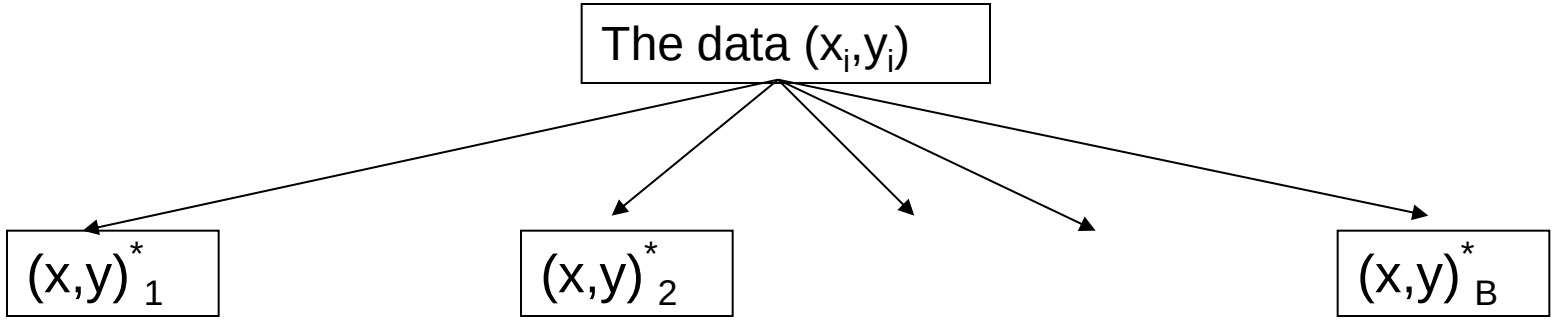


A bootstrap sample

$$\hat{E}(y \mid x)^*$$

An estimates for $E(y|x)$

Bootstrap estimate for $E(y|x)$



B bootstrap samples

$$\hat{E}(y\,|\,x)_1^* \quad \hat{E}(y\,|\,x)_2^* \quad \hat{E}(y\,|\,x)_3^* \qquad\qquad \hat{E}(y\,|\,x)_B^*$$

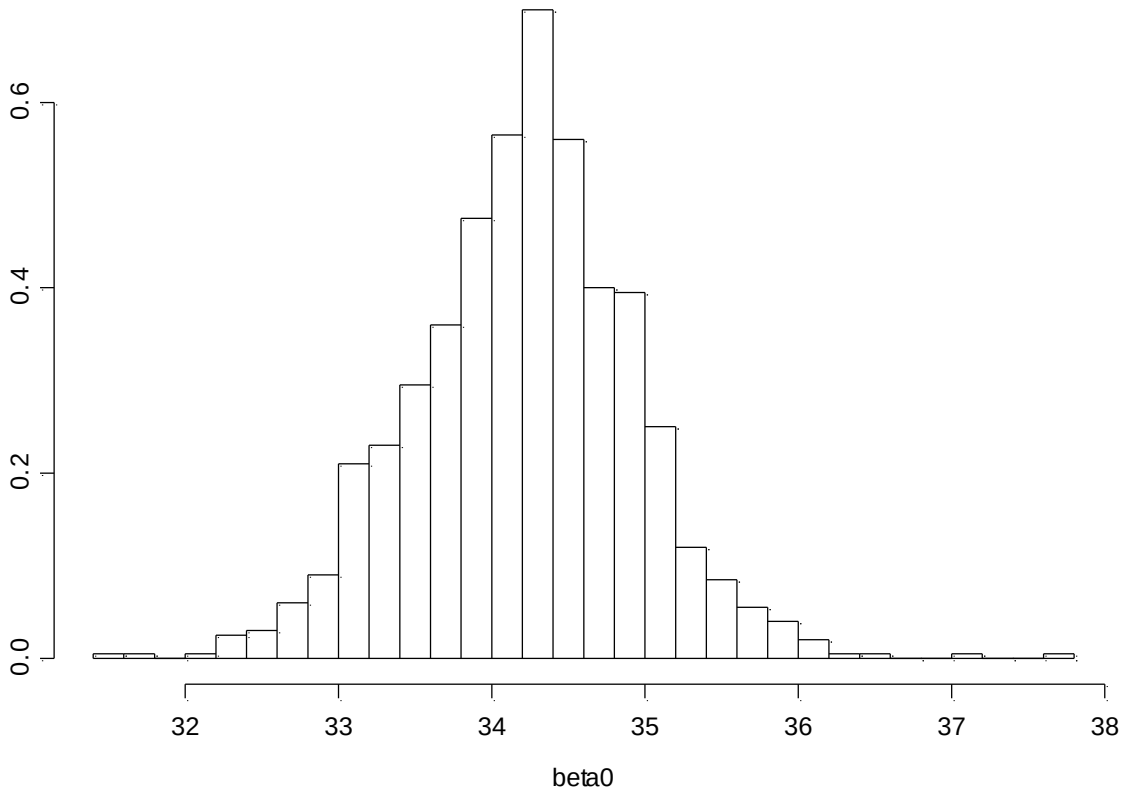
B bootstrap estimates for $E(y|x)$

R code the bootstrap (non parametric)

```
> n <- length(x)
> B <- 1000
> index <- c(1:n)
> beta0 <- beta1 <- c(1:B)
> for(i in 1:B) {
   cat(i)
   i.boot <- sample(index, size = n, replace = T)
   y.boot <- y[i.boot]
   x.boot <- x[i.boot]
   fit.boot <- lm(y.boot ~ x.boot)
   beta0[i] <- fit.boot$coeff[1]
   beta1[i] <- fit.boot$coeff[2]
}
```

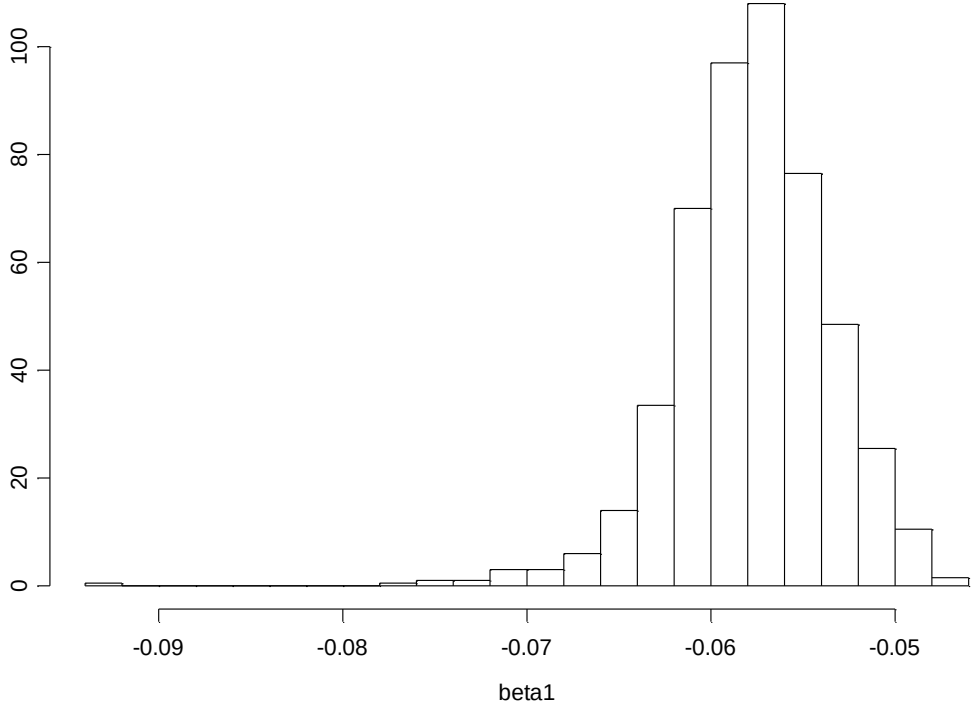
$$\begin{matrix} (x_1 & y_1)^* \\ (x_2 & y_2)^* \\ . & . \\ . & . \\ (x_n & y_n)^* \end{matrix}$$

Distribution of the bootstrap replicates for the intercept




```
> quantile(beta0, probs = c(0.025, 0.975))
     2.5%    97.5%
 32.79732 35.60828
```

Distribution of the bootstrap replicates for the slope




```
> quantile(beta1, probs = c(0.025, 0.975))
        2.5%       97.5%
 -0.06671698 -0.05016021
```

Re sampling residuals

$$\begin{matrix} x_1 & y_1 \\ x_2 & y_2 \\ . & . \\ . & . \\ x_n & y_n \end{matrix}$$

Estimate the parameters from the data

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimate the residuals

$$\hat{e}_i = y_i - \hat{y}_i$$

Re sampling residuals

Step 1: calculate the residuals

$$\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n$$

Step 2: bootstrap the residuals

The bootstrap algorithm

Residuals from the fitted model

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n$$

For b in 1:B

A bootstrap samples for e

$$\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^*$$

Bootstrap replicates for Y

$$y_1^* = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{e}_1^* \qquad y_2^* = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{e}_2^* \qquad y_n^* = \hat{\beta}_0 + \hat{\beta}_1 x_n + \hat{e}_n^*$$

The bootstrap algorithm

Residuals from the null model (the “observed data”)

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n$$

B bootstrap samples

$$\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^* \qquad \hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^* \qquad \hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^*$$

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i^*$$

A bootstrap data with n pairs !!!!

The bootstrap algorithm

Bootstrap data

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n \xrightarrow{\text{Bootstrap sample of e}} \begin{array}{c} e_1^* \\ e_2^* \\ \\ \\ e_n^* \end{array} \begin{array}{c} \rightarrow \\ \rightarrow \\ \\ \\ \rightarrow \end{array} \begin{array}{c} (x_1, \hat{\beta}_0 + \hat{\beta}_1 x_1 + e_1^*) \\ (x_2, \hat{\beta}_0 + \hat{\beta}_1 x_2 + e_2^*) \\ \\ \\ (x_n, \hat{\beta}_0 + \hat{\beta}_1 x_n + e_n^*) \end{array}$$

The bootstrap algorithm

Bootstrap data

$$(x_1, \hat{\beta}_0 + \hat{\beta}_1 x_1 + e_1^*) \longrightarrow (x_1, y_1^*)$$

$$(x_2, \hat{\beta}_0 + \hat{\beta}_1 x_2 + e_2^*) \longrightarrow (x_2, y_2^*)$$

$$(x_n, \hat{\beta}_0 + \hat{\beta}_1 x_n + e_n^*) \longrightarrow (x_n, y_n^*)$$

$$\downarrow$$

$$y_i^* = \beta_0 + \beta_1 x_i + \delta_i$$

R code for the bootstrap (residuals)

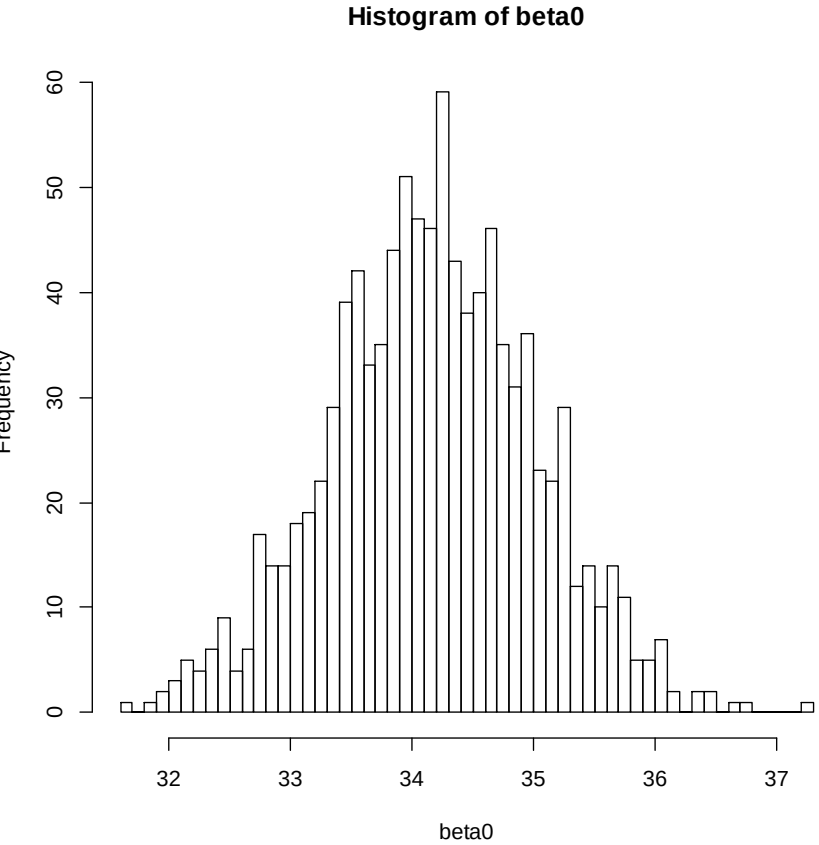
```
> fit.lm <- lm(y ~ x)
> ei <- fit.lm$resid
> n <- length(x)
> B <- 1000
> beta0 <- beta1 <- c(1:B)
> for(i in 1:B) {
   cat(i)
   e.boot <- sample(ei, size = n, replace = T)
   y.boot <- fit.lm$coeff[1] + fit.lm$coeff[2]*x + e.boot
   x.boot <- x
   fit.boot <- lm(y.boot ~ x.boot)
   beta0[i] <- fit.boot$coeff[1]
   beta1[i] <- fit.boot$coeff[2]
}
```

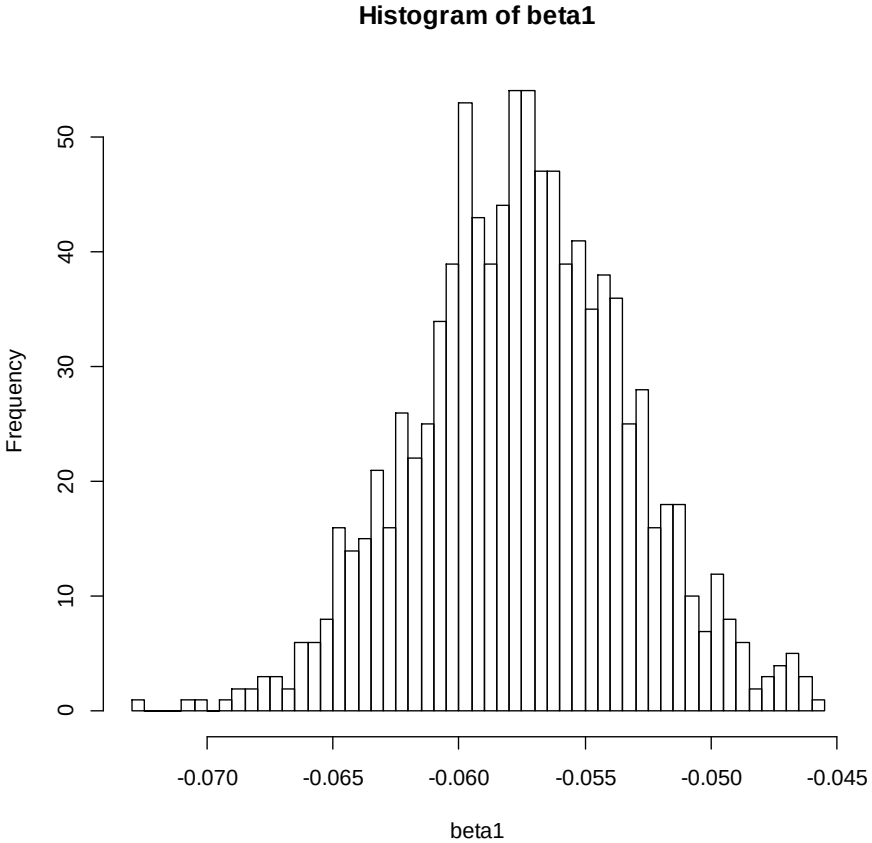
$$\hat{e}_1, \hat{e}_2, \ldots\ldots, \hat{e}_n$$

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i^*$$

$$y_i^* = \beta_0 + \beta_1 x_i + \delta_i$$

Distribution of the bootstrap replicates for the intercept and slope






```
> quantile(beta0,probs=c(0.025,0.975))
    2.5%    97.5%
32.43178 35.80418
```

```
> quantile(beta1,probs=c(0.025,0.975))
       2.5%       97.5%
-0.06559971 -0.04911397
```

bootstrapping pairs versus bootstrapping residuals

$$P = (\beta, F)$$

$$F = F(X, Y) \qquad\qquad F \rightarrow (\varepsilon_1, \ldots \varepsilon_n)$$

What is the difference ?

Bootstrapping pairs:

We do not make any assumption about the distribution of the error .

We do not make an assumption about constant variance.

Bootstrap estimate for the S.E

The Bootstrap estimate for the S.E can be calculated as before, i.e.,

$$\boxed{S.E.(\hat{\beta}_1) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_b^* - \hat{\beta}^* \right)^2 \right\}^{0.5}}$$

With

$$\hat{\beta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b^*$$

Test of hypotheses

$$H_0 : E(y_i) = \beta_0$$

$$H_1 : E(y_i) = \beta_0 + \beta_1 x_i$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Calculate the observed statistics for the parameter of primary interest.

Re sample B bootstrap samples UNDER the null hypothesis and calculate the bootstrap replicates for statistics.

Calculate Monte Carlo p values

$$P = \frac{\#\left\{\hat{\beta}_1^* \geq \hat{\beta}_1\right\} + 1}{B + 1}$$

Test of hypotheses

Re sampling under the null hypothesis

Bootstrapping pairs

Very easy: fix X and bootstrap Y

Bootstrapping residuals

Obtain the residuals under the null model.

Bootstrap the residuals and calculate the bootstrap replicates under the null.

Parametric bootstrap

We assume that the hormone level (y) is a function of the hours.

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

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

Estimate the unknown parameter and generate a bootstrap sample:

$$y_i^* \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2\right)$$

Parametric bootstrap

- For each bootstrap sample:
  - For i=1,..,n

$$\boxed{y_i^* \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2\right)}$$

$$\left(x_1, y_1^*\right) \quad \left(x_1, y_1^*\right) \quad \left(x_1, y_1^*\right)$$

$$\left(x_2, y_2^*\right) \quad \left(x_2, y_2^*\right) \quad \left(x_2, y_2^*\right)$$

$$\underbrace{\left(x_n, y_n^*\right)}_{\text{b=1}} \quad \underbrace{\left(x_n, y_n^*\right)}_{} \quad \underbrace{\left(x_n, y_n^*\right)}_{\text{b=B}}$$

$$\boxed{\hat{\beta}_0^*, \hat{\beta}_1^*} \quad \boxed{\hat{\beta}_0^*, \hat{\beta}_1^*} \quad \boxed{\hat{\beta}_0^*, \hat{\beta}_1^*}$$

Fitted model for the hormone data

```
> fit.lm<-lm(hormone$amount~hormone$hrs)
> summary(fit.lm)

Call:
lm(formula = hormone$amount ~ hormone$hrs)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9357 -1.7282 -0.0229  1.7388  3.7323 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.167528   0.867197   39.40  < 2e-16 ***
hormone$hrs -0.057446   0.004464  -12.87 1.58e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.378 on 25 degrees of freedom
Multiple R-squared:  0.8688,    Adjusted R-squared:  0.8636 
F-statistic: 165.6 on 1 and 25 DF,  p-value: 1.584e-12
```

Parameter estimates:

$$\hat{\beta}_0 = 34.167528$$

$$\hat{\beta}_1 = -0.057446$$

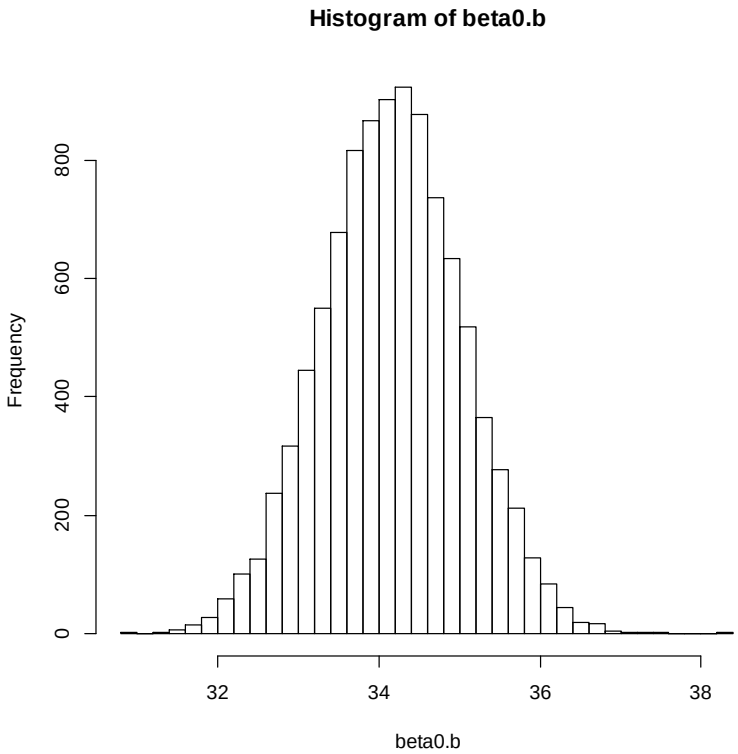
$$\hat{\sigma} = 2.378$$

Parametric bootstrap:

$$y_i^* \sim N\left(\hat{\beta}_0 + \hat{\beta}_1 x_i, \hat{\sigma}^2\right)$$

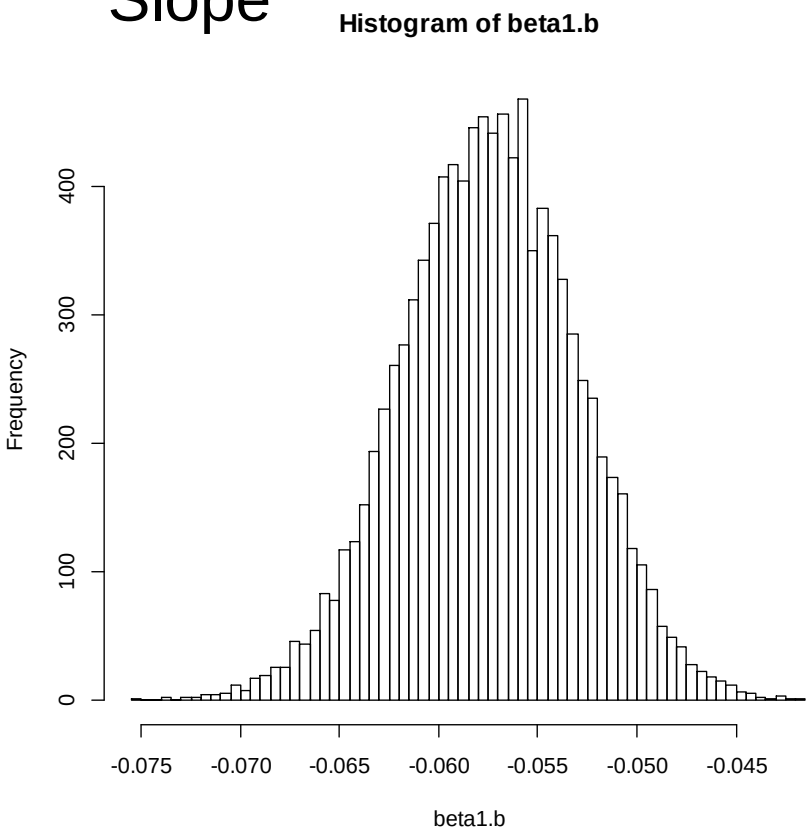
Distribution of the bootstrap replicates for the intercept and slope

Intercept




```
> quantile(beta0.b,probs=c(0.025,0.975))
    2.5%    97.5%
32.48882 35.86673
```

Slope




```
> quantile(beta1.b,probs=c(0.025,0.975))
       2.5%       97.5%
-0.06613581 -0.04891995
```

R code for the parametric bootstrap

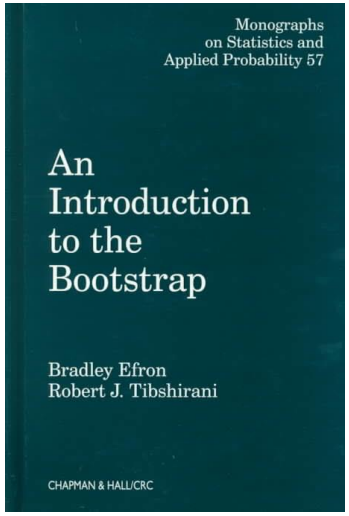
Non parametric bootstrap

```
n<-length(hormone$amount)
fit.lm<-lm(hormone$amount~hormone$hrs)
summary(fit.lm)

B<-10000
beta0.b<-beta1.b<-c(1:B)
index<-c(1:n)
for (i in 1:B)
{
index.b<-sample(index,n,replace=TRUE)
hormone.b<-hormone[index.b,]
fit.lm.b<-
lm(hormone.b$amount~hormone.b$hrs)
beta0.b[i]<-summary(fit.lm.b)$coeff[1,1]
beta1.b[i]<-summary(fit.lm.b)$coeff[2,1]
}
```

Parametric bootstrap

```
n<-length(hormone$amount)
fit.lm<-lm(hormone$amount~hormone$hrs)
summary(fit.lm)
beta0<-summary(fit.lm)$coeff[1,1]
beta1<-summary(fit.lm)$coeff[2,1]
sigma<-2.378

B<-10000
beta0.b<-beta1.b<-c(1:B)
amount.b<-c(1:n)
for (i in 1:B)
{
for(j in 1:n)
{
amount.b[j]<-
rnorm(1,beta0+beta1*hormone$hrs[j],sigma)
}
fit.lm.b<-lm(amount.b~hormone$hrs)
beta0.b[i]<-summary(fit.lm.b)$coeff[1,1]
beta1.b[i]<-summary(fit.lm.b)$coeff[2,1]
}
```

Example:
The cell data

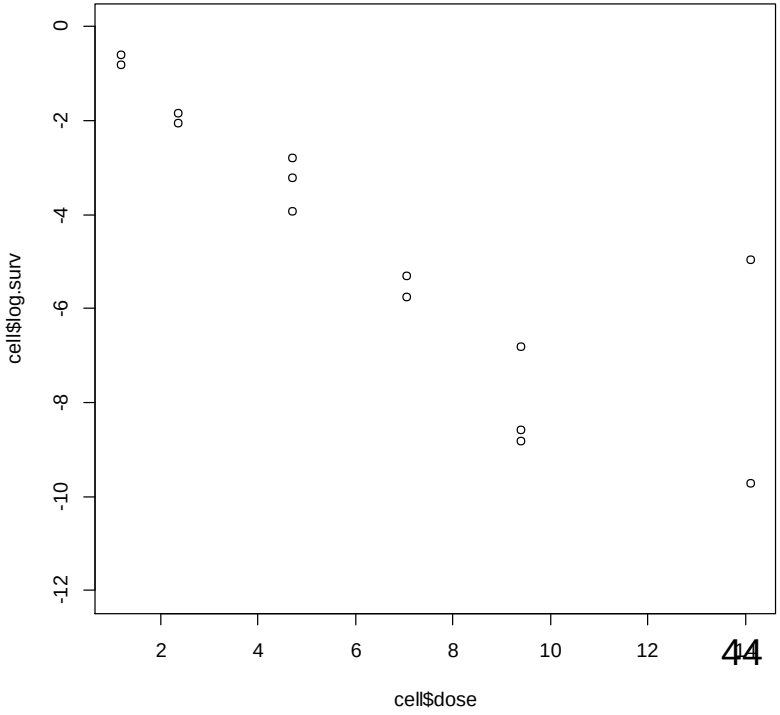


Chapter 9, Section 9.6 & 9.7

The cell data

- Data on cell survival under different radiation doses.
- Variables:
  - Dose
  - logarithm of proportion of survival.
- In R:

```
> help(cell)
```



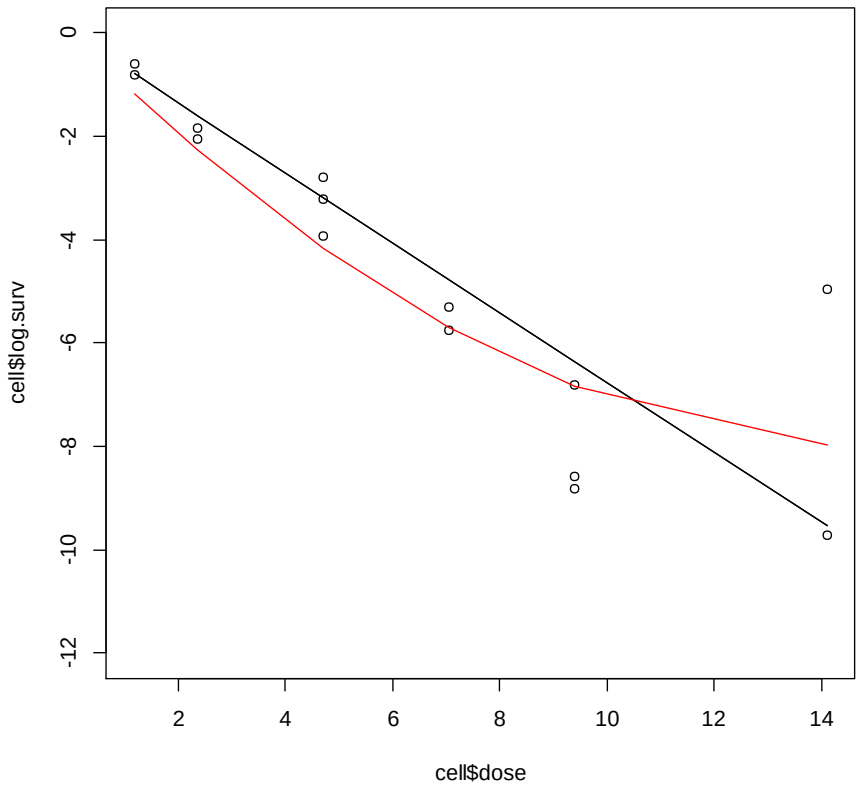
Models for log(survival proportion)

$$M_1 : y_i = \beta_1 z_i + \varepsilon_i$$

$$M_2 : y_i = \beta_1 z_i + \beta_2 z_1^2 + \varepsilon_i$$

```
> summary(fit.lm2)

Call:
lm(formula = y ~ -1 + x + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97720 -0.03754  0.30231  0.55116  3.00441

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x  -1.04910    0.15871   -6.61  2.5e-05 ***
x2  0.03433    0.01395    2.46     0.03 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 1.378 on 12 degrees of freedom
Multiple R-squared:  0.9457,    Adjusted R-squared:
0.9366
F-statistic: 104.5 on 2 and 12 DF,  p-value: 2.566e-08
```

Dara and estimated models (all data, n=14).

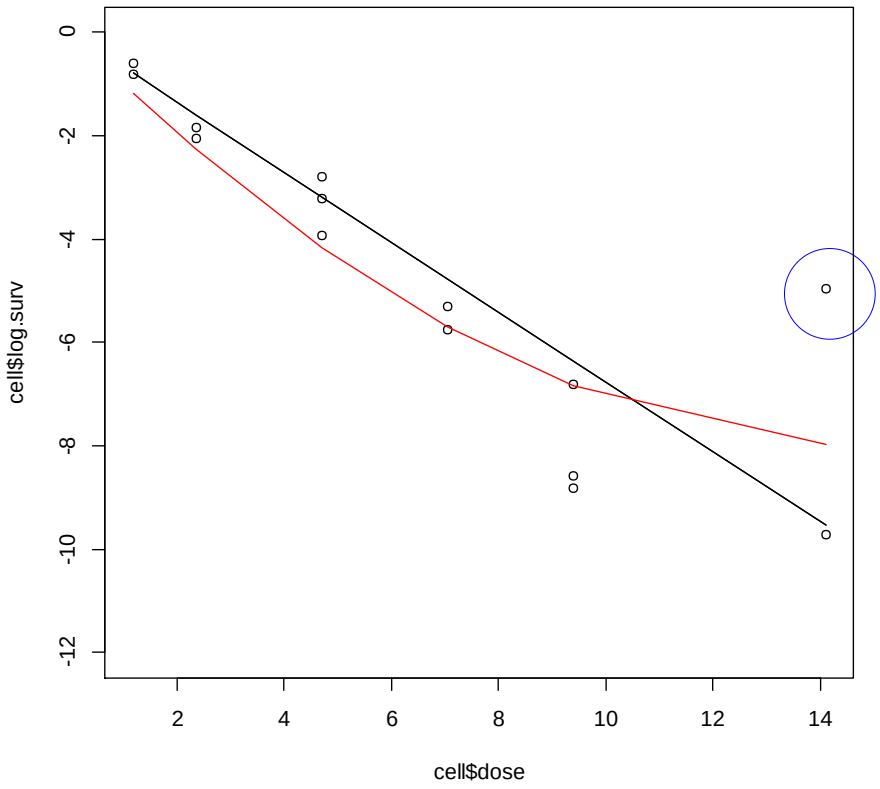


The influence of observation 13

The main question: what is the effect of observation 13 on the fitted models ?

An unusual value for y:

```
> x<-cell$dose
> x[13]
[1] 14.1
> y[13]
[1] -4.962
```



Models for log(survival proportion)

The quadratic model with and without observation 13.

$$M_2 : y_i = \beta_1 z_i + \beta_2 z_1^2 + \varepsilon_i$$

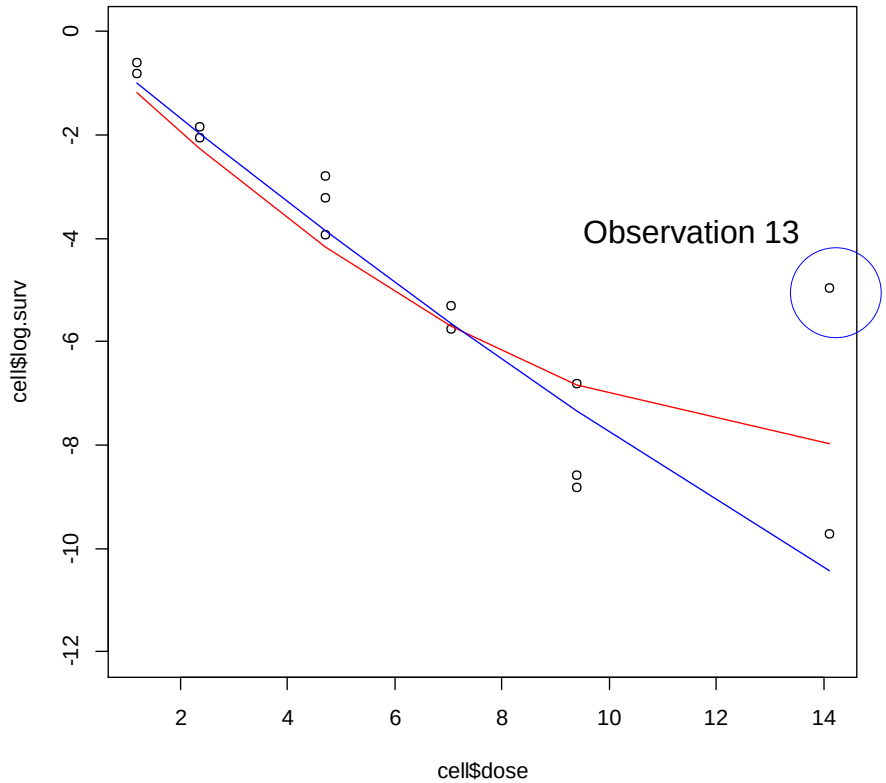
Parameter estimates for n=14

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x  -1.04910    0.15871   -6.61  2.5e-05 ***
x2  0.03433    0.01395    2.46     0.03 *
---
```

Parameter estimates for n=13

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
x13  -0.861950   0.094280  -9.142  1.8e-06 ***
x132  0.008646   0.009076   0.953    0.361 ---
```

Dara and estimated models
(without observation 13,n=13).



Least median squares

We fit the model:

$$y_i = \beta_1 z_i + \varepsilon_i$$

Estimate the unknown parameter by minimizing the least median sum of squares:

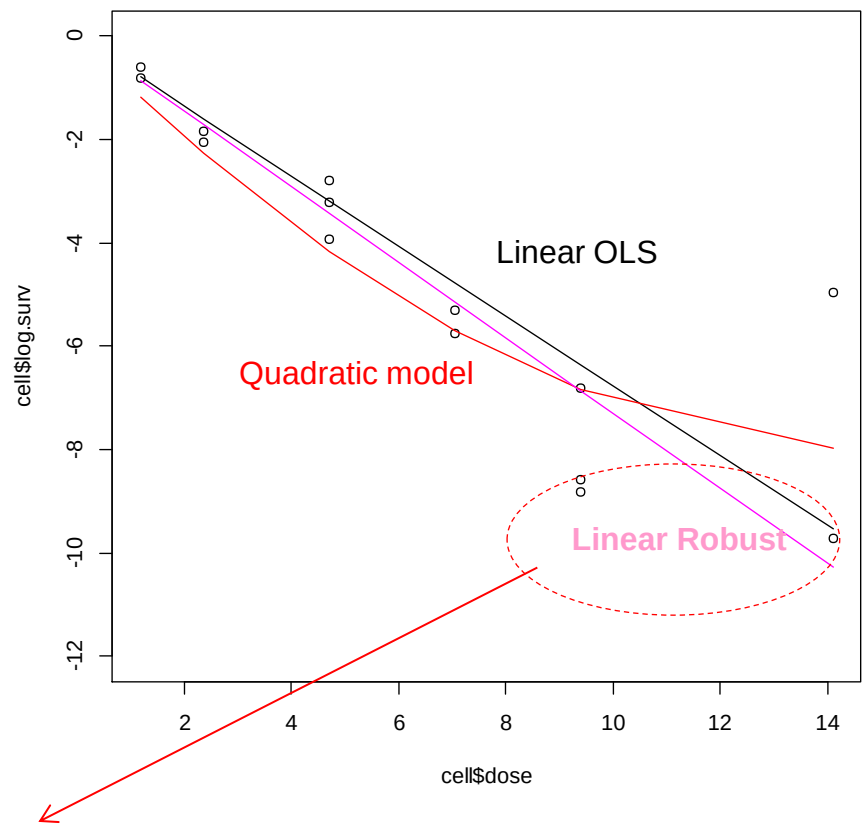
$$MSR(\hat{\beta}) = median\left(y_i - \beta_1 z_i\right)^2$$

OLS estimator minimizes:

$$RSS = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_1 z_i\right)^2$$

Robust regression (linear model)

```
> summary(fit.rob)

Call: rlm(formula = y ~ -1 + x)
Residuals:
     Min       1Q   Median       3Q      Max
-1.94911 -0.45955 -0.04152  0.24648  5.32183

Coefficients:
  Value    Std. Error t value
x  -0.7293   0.0237   -30.7394

Residual standard error: 0.6154 on 13 degrees
of freedom
```

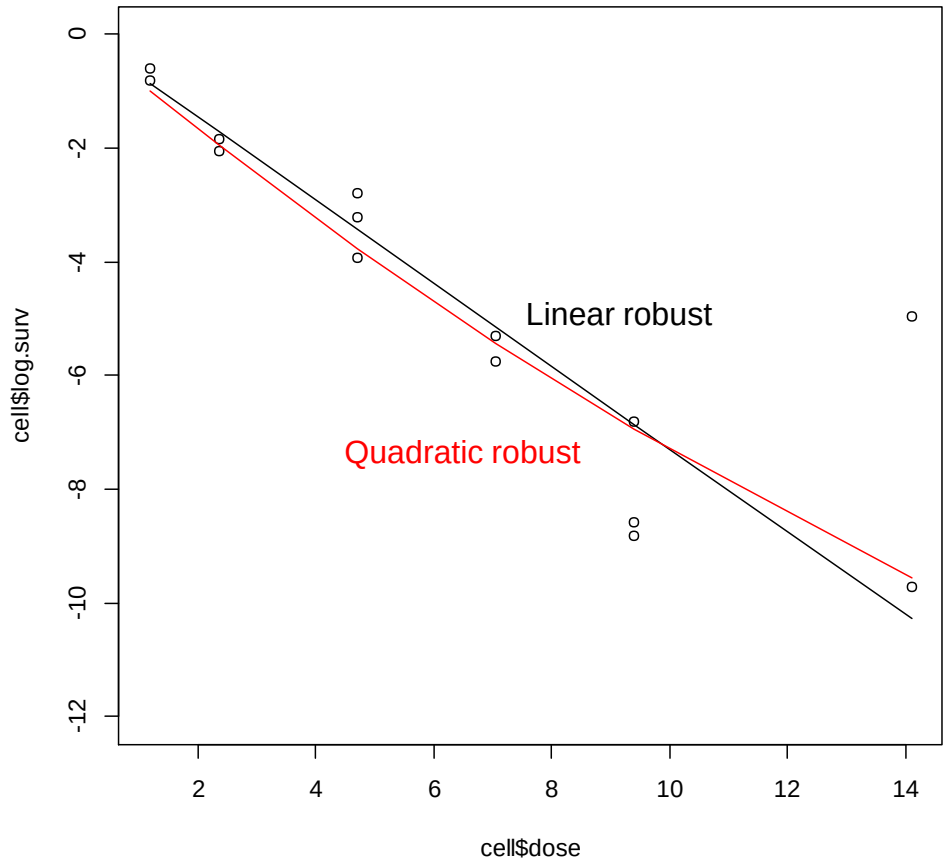


- The model was fitted using the R function `rlm()`.
- Use `help(rlm)` to see which method is used to fit the model.

Robust regression (quadratic model)

$$y_i = \beta_1 z_i + \beta_2 z_i^2 + \varepsilon_i$$

```
> fit.rob2<-rlm(y~-1+x+x2)
> summary(fit.rob2)

Call: rlm(formula = y ~ -1 + x + x2)
Residuals:
    Min      1Q  Median      3Q     Max
-1.8525 -0.1664  0.1294  0.3429  4.5919

Coefficients:
   Value    Std. Error t value
x   -0.8637   0.0706   -12.2285
x2   0.0132   0.0062     2.1257
```

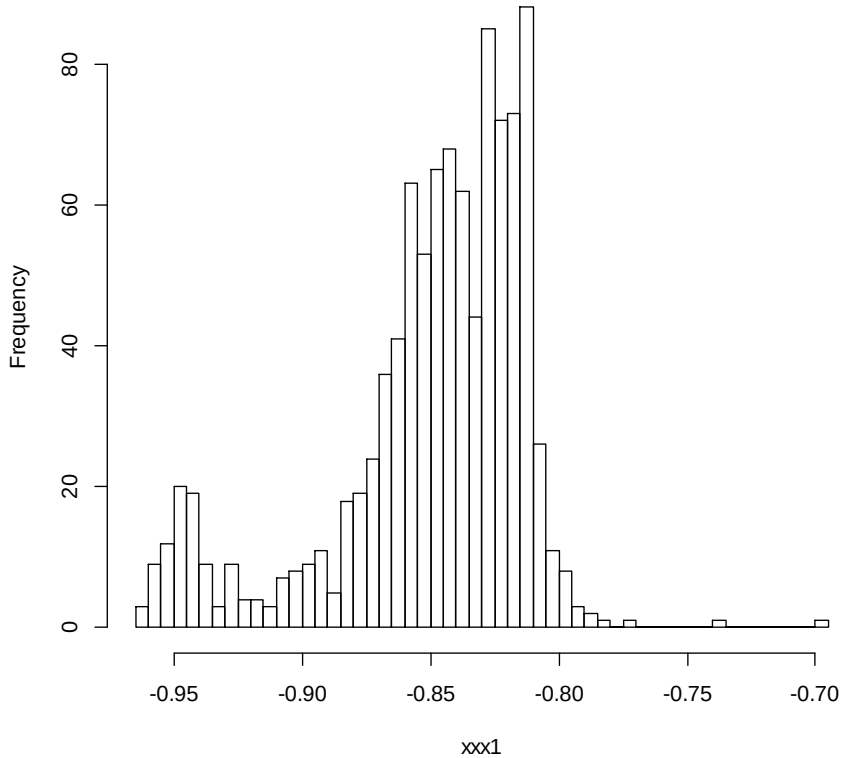


95% percentile bootstrap intervals for the prediction model

Non parametric bootstrap.

Bootstrap replicates for the predicted value at the first dose level.

$$\hat{y}^*_{1,d_1}, \hat{y}^*_{2,d_1}, \ldots, \hat{y}^*_{B,d_1}$$



$$\left(\hat{y}_{i,0.025}, \hat{y}_{i,0.975}\right) \Rightarrow$$

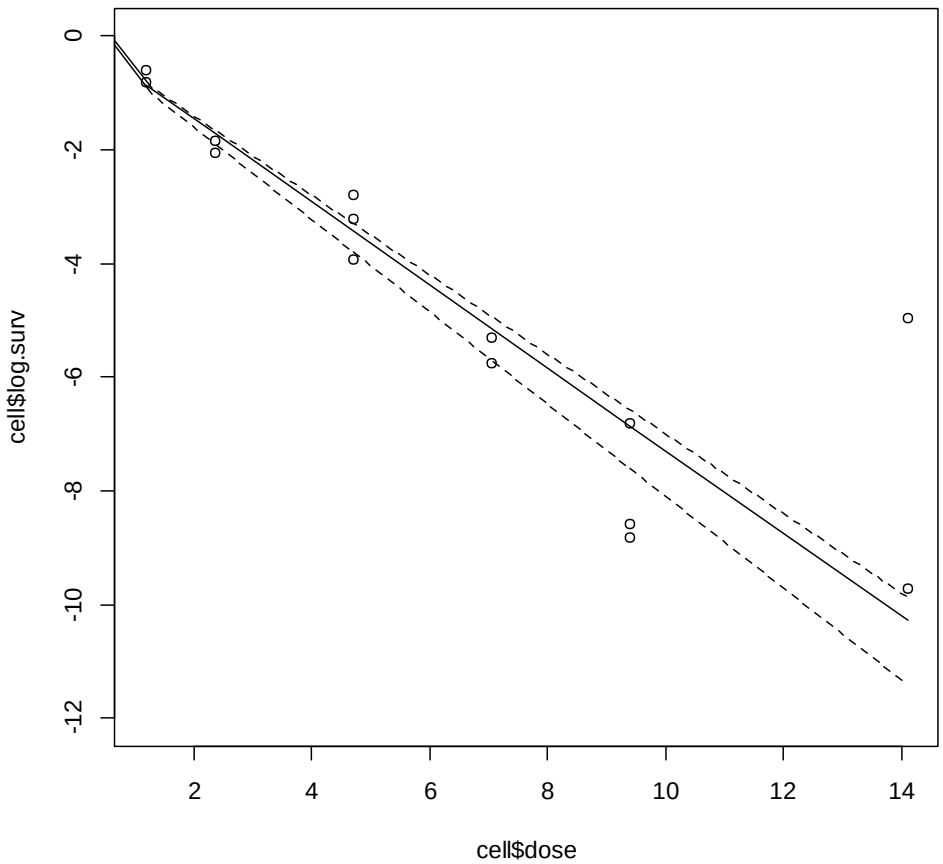
```
> ci[i,]
[1] -0.9499772 -0.8227408
```

95% percentile bootstrap intervals for the prediction model

Bootstrap C.I for the predicted value for all dose levels:

$$\left(\hat{y}_{i,0.025}, \hat{y}_i, \hat{y}_{i,0.975}\right)$$

Predicted value form the regression model.

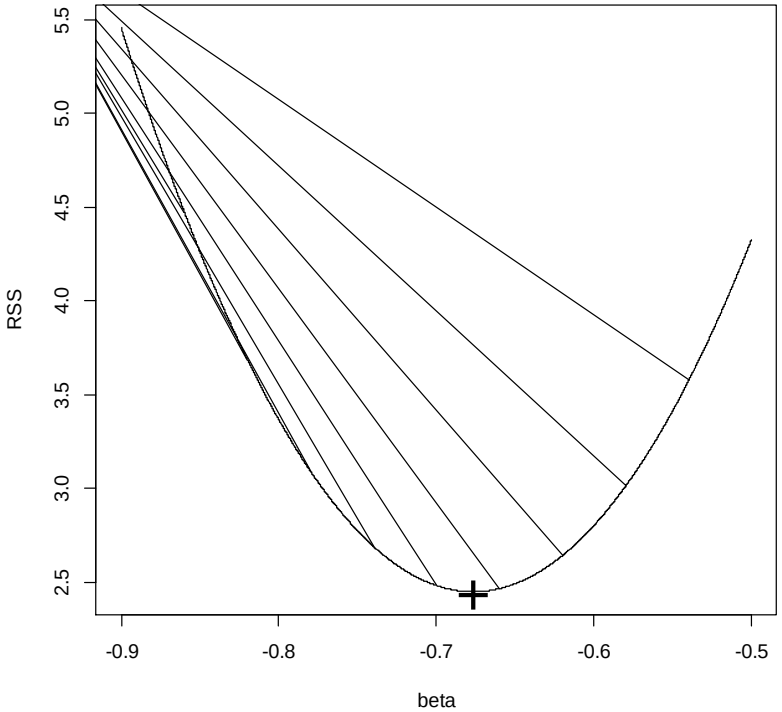


Models for log(survival proportion)

$$M_1 : y_i = \beta_1 z_i + \varepsilon_i$$

Estimate the unknown parameter
by the residuals sum of squares:

$$RSS = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \beta_1 z_i\right)^2$$



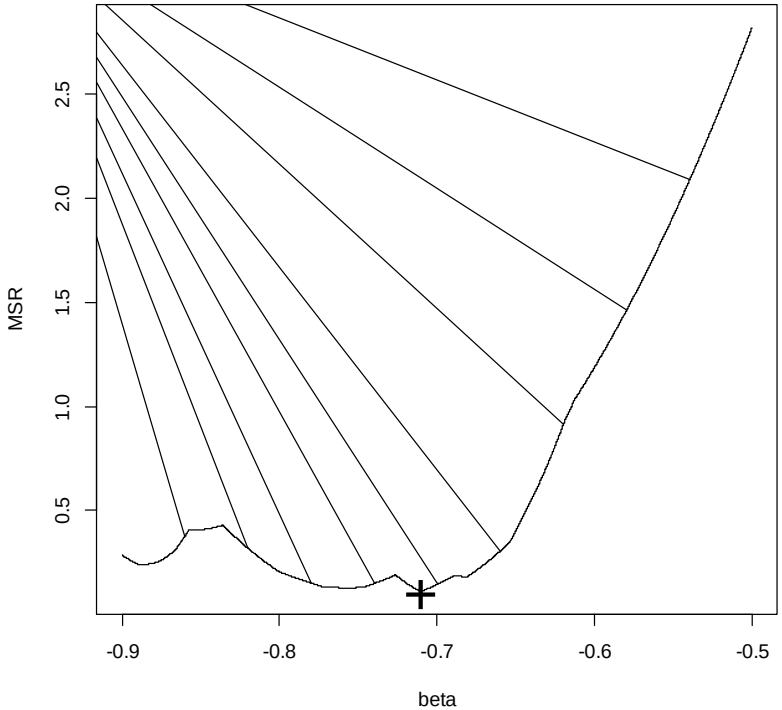

```
[1] 2.449635
> beta.rss
[1] -0.6764976
```

Models for log(survival proportion)

$$M_1 : y_i = \beta_1 z_i + \varepsilon_i$$

Estimate the unknown parameter
by the median squared residual:

$$MSR(\hat{\beta}) = median\left(y_i - \beta_1 z_i\right)^2$$



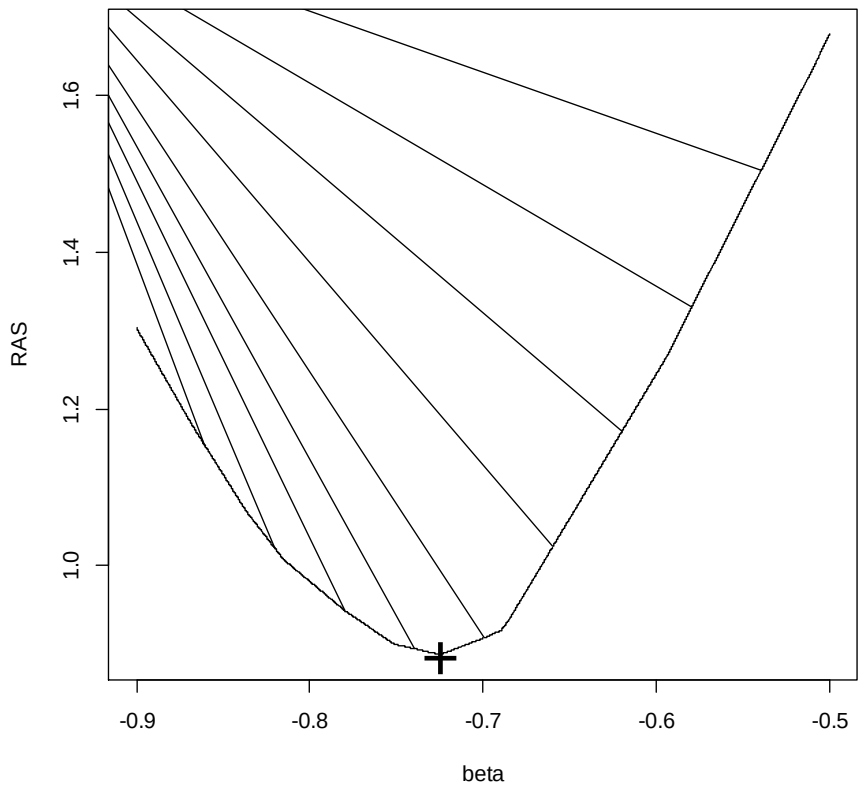

```
> m.MSR
[1] 0.1114881
> beta.msr
[1] -0.710101
```

Models for log(survival proportion)

$$M_1 : y_i = \beta_1 z_i + \varepsilon_i$$

Estimate the unknown parameter
by the sum of absolute residuals:

$$RAS = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \beta_1 z_i\right|$$



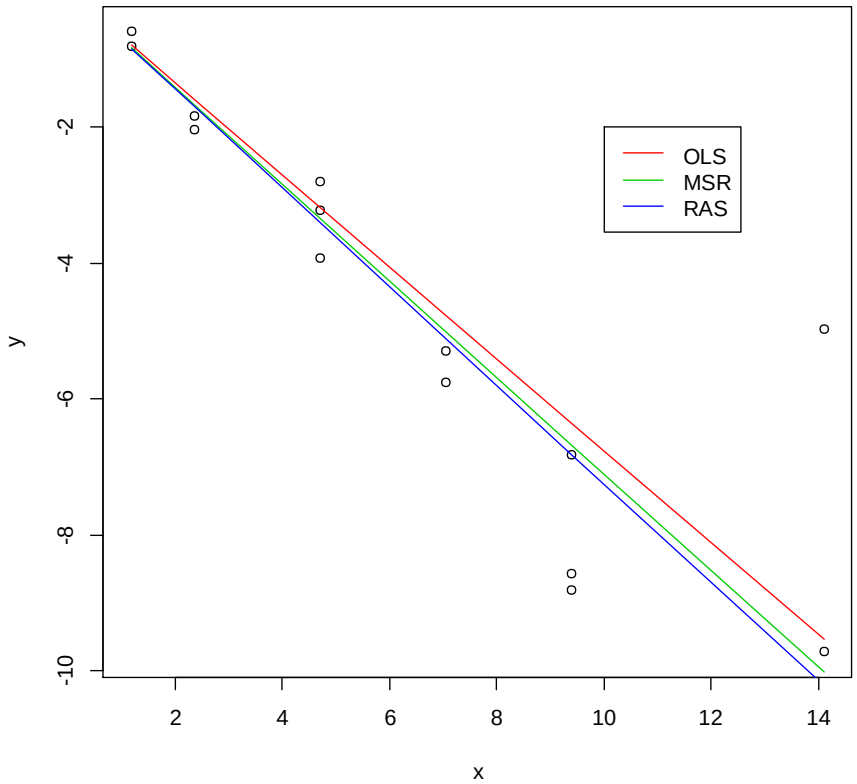

```
> m.RAS
[1] 0.8859395
> beta.ras
[1] -0.7247025
```

The three models

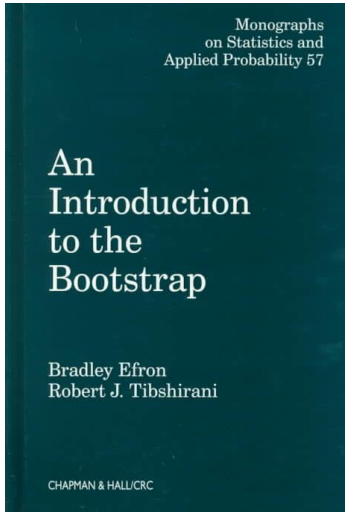
For a linear model,

$$M_1 : y_i = \beta_1 z_i + \varepsilon_i$$

The effect of observation 13 is minimal (for the linear model).



Example:
The tooth strength data



Chapter 14, Section 14.5

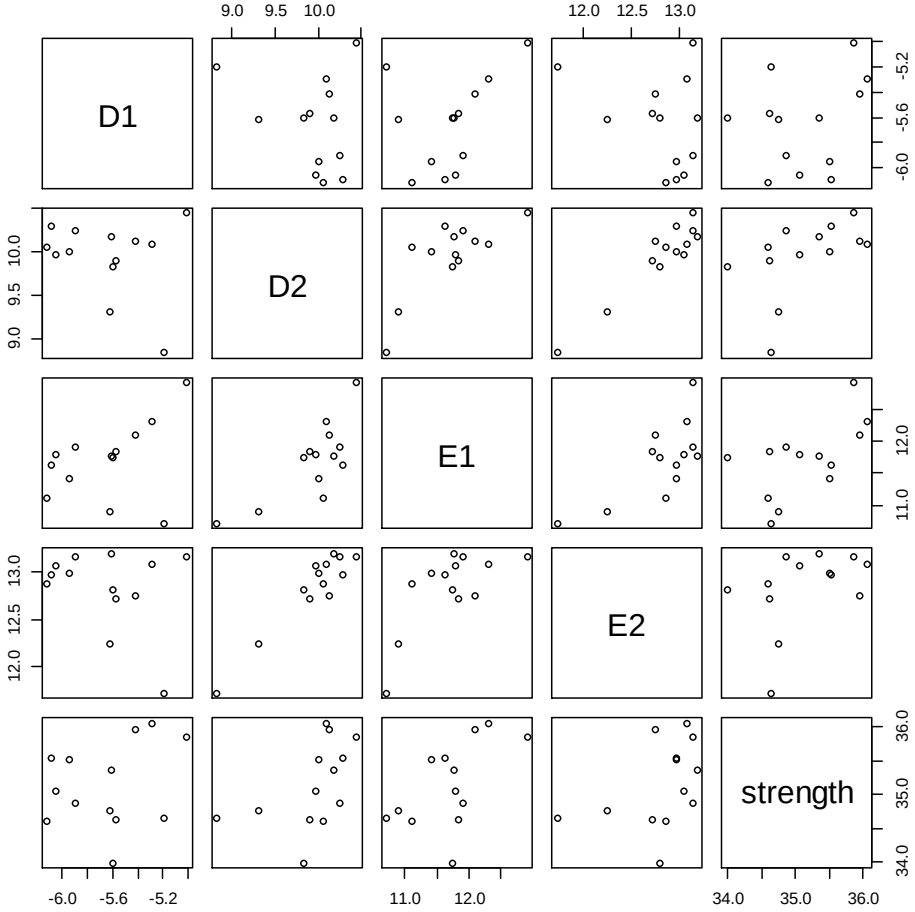
The tooth strength data

- Thirteen accident victims have had the strength of their teeth measured.
- We want to predict teeth strength from measurements not requiring destructive testing.
- Four variables are observed obtained for each subject:
  - D1,D2 are difficult to obtain.
  - E1,E2 are easy to obtain.
- In R

```
> help(tooth)
```

The Tooth Strength Data

```
> tooth
   patient     D1     D2    E1    E2 strength
1        1 -5.288 10.091 12.30 13.08    36.05
2        2 -5.944 10.001 11.41 12.98    35.51
3        3 -5.607 10.184 11.76 13.19    35.35
4        4 -5.413 10.131 12.09 12.75    35.95
5        5 -5.198  8.835 10.72 11.73    34.64
6        6 -5.598  9.837 11.74 12.80    33.99
7        7 -6.120 10.052 11.10 12.87    34.60
8        8 -5.572  9.900 11.85 12.72    34.62
9        9 -6.056  9.966 11.78 13.06    35.05
10      10 -5.010 10.449 12.91 13.15    35.85
11      11 -6.090 10.294 11.63 12.97    35.53
12      12 -5.900 10.252 11.91 13.15    34.86
13      13 -5.620  9.316 10.89 12.25    34.75
```



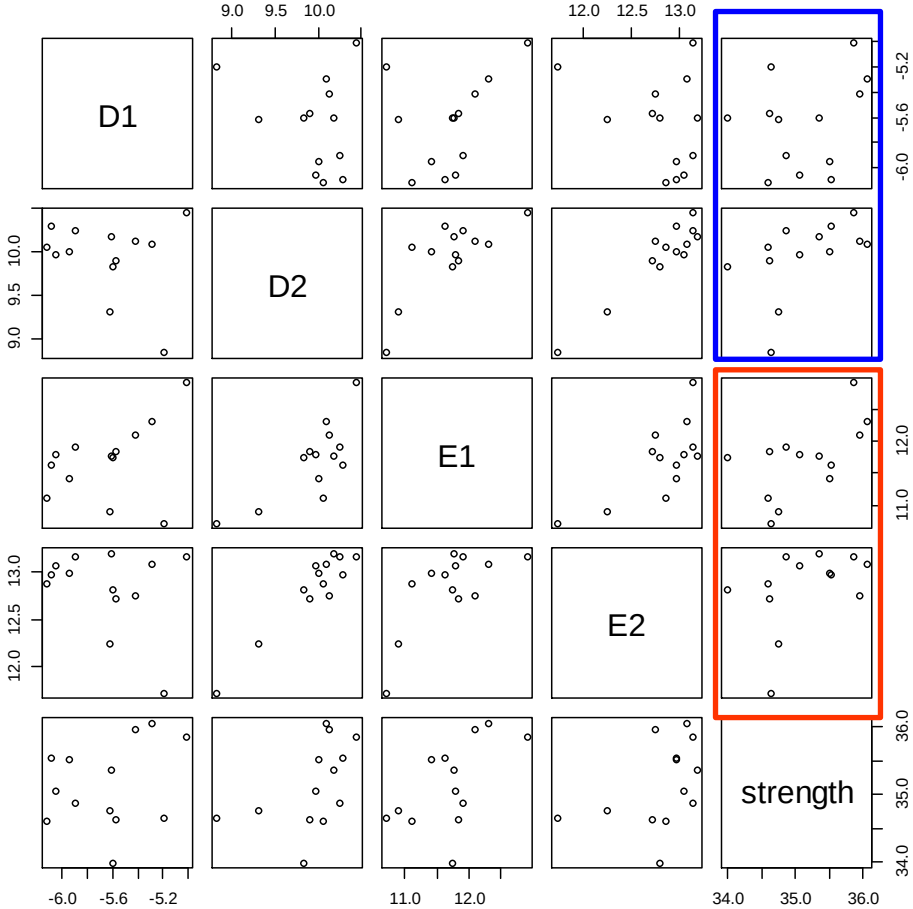
The tooth strength data

Two competing models:

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_{Di}$$

$$y_i = \alpha_0 + \alpha_1 E_{1i} + \alpha_2 E_{2i} + \varepsilon_{Ei}$$

The main question, which model lead to lower residuals ?



Fitted models: the tooth strength data

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \varepsilon_{Di}$$

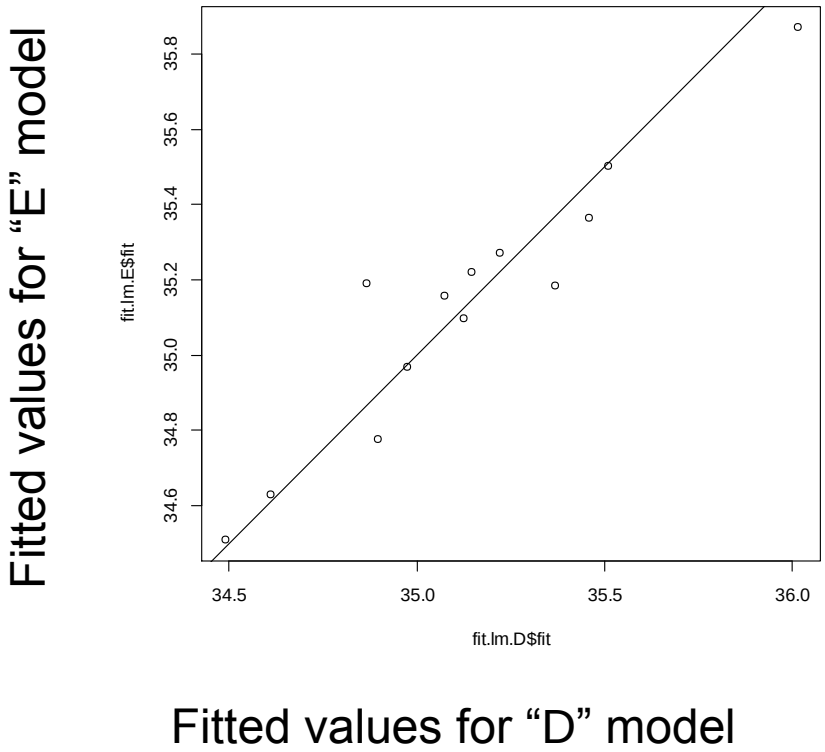
```
> fit.lm.D<-lm(strength~D1+D2,data=tooth)
> summary(fit.lm.D)

Call:
lm(formula = strength ~ D1 + D2, data = tooth)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0843 -0.2951  0.1411  0.4049  0.5395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.4915     3.8726   7.874 1.35e-05 ***
D1            0.6991     0.4337   1.612   0.1380
D2            0.8637     0.3597   2.401   0.0373 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.5254 on 10 degrees of freedom
Multiple R-squared:  0.412,     Adjusted R-squared:
0.2944
F-statistic: 3.504 on 2 and 10 DF,  p-value: 0.07028
```

Higher R^2

$$y_i = \alpha_0 + \alpha_1 E_{1i} + \alpha_2 E_{2i} + \varepsilon_{Ei}$$

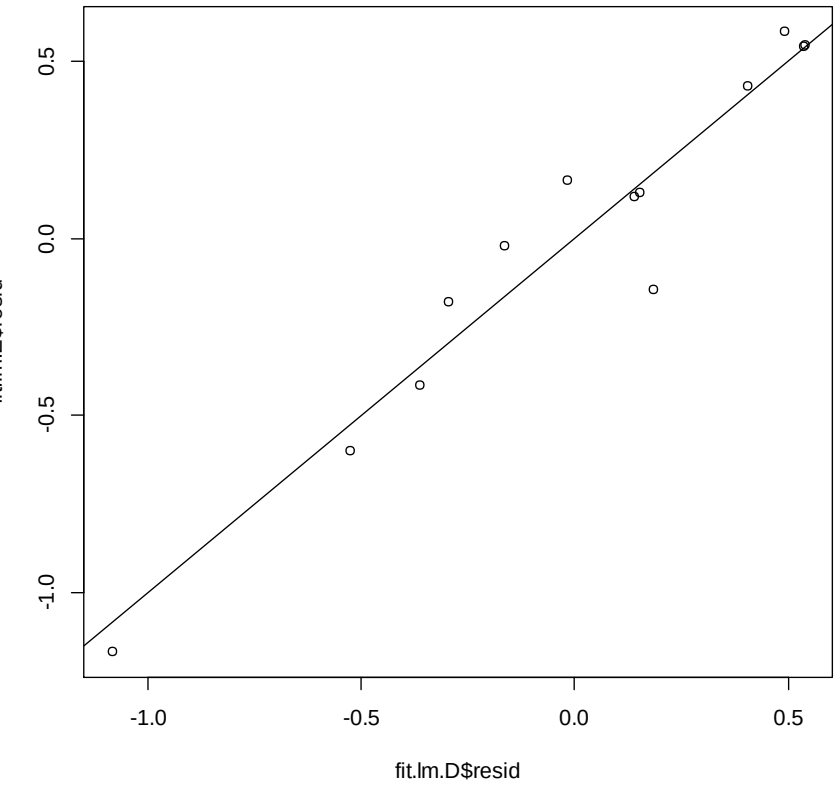
```
> fit.lm.E<-lm(strength~E1+E2,data=tooth)
> summary(fit.lm.E)

Call:
lm(formula = strength ~ E1 + E2, data = tooth)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1681 -0.1782  0.1203  0.4313  0.5846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.66657    5.02179   5.509 0.000258 ***
E1           0.59761    0.39095   1.529 0.157355
E2           0.03716    0.55306   0.067 0.947754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.5593 on 10 degrees of freedom
Multiple R-squared:  0.3336,    Adjusted R-squared:
0.2003
F-statistic: 2.503 on 2 and 10 DF,  p-value: 0.1314
```

Predicted values by the two models



- Fitted values of the two models:
  - Correlated.
  - Which one is a better prediction ?

Residuals



$$RSE(E) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i(E)\right)^2$$

$$RSE(D) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i(D)\right)^2$$

```
> RSS.E<-sum((tooth$strength-fit.lm.E$fit)^2)
> RSS.E
[1] 3.12817
> RSS.D<-sum((tooth$strength-fit.lm.D$fit)^2)
> RSS.D
[1] 2.760094
```

$$RSE(E) > RSE(D)$$

Residuals

$$\hat{\theta} = \frac{1}{n}\left[RSE(E) - RSE(D)\right]$$

```
> theta<-(RSS.E-RSS.D)/n
> theta
[1] 0.02831355
```

$$RSE(E) > RSE(D)$$

$$\hat{\theta} = 0.0283 > 0$$

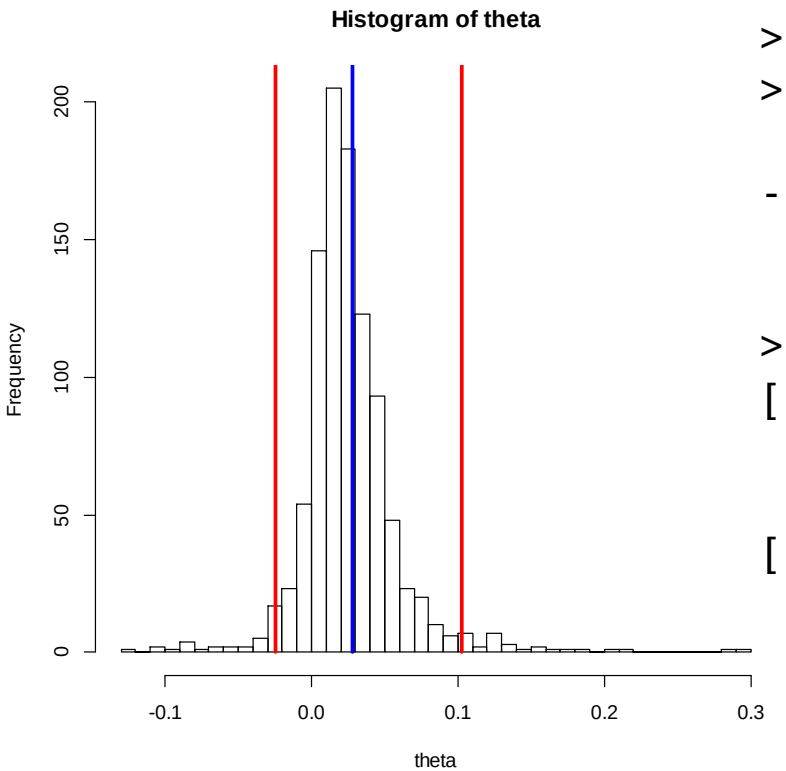
- What is the standard error of θ ?
- What is the distribution of θ ?

Non parametric bootstrap

- Draw B bootstrap samples (re sample pairs).
- For each sample, fit the two models and obtain the bootstrap replicates:

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^* \qquad RSE(E)_1^*, RSE(E)_2^*, \ldots, RSE(E)_B^* \qquad RSE(D)_1^*, RSE(D)_2^*, \ldots, RSE(D)_B^*$$

Bootstrap distribution of $\hat{\theta}$

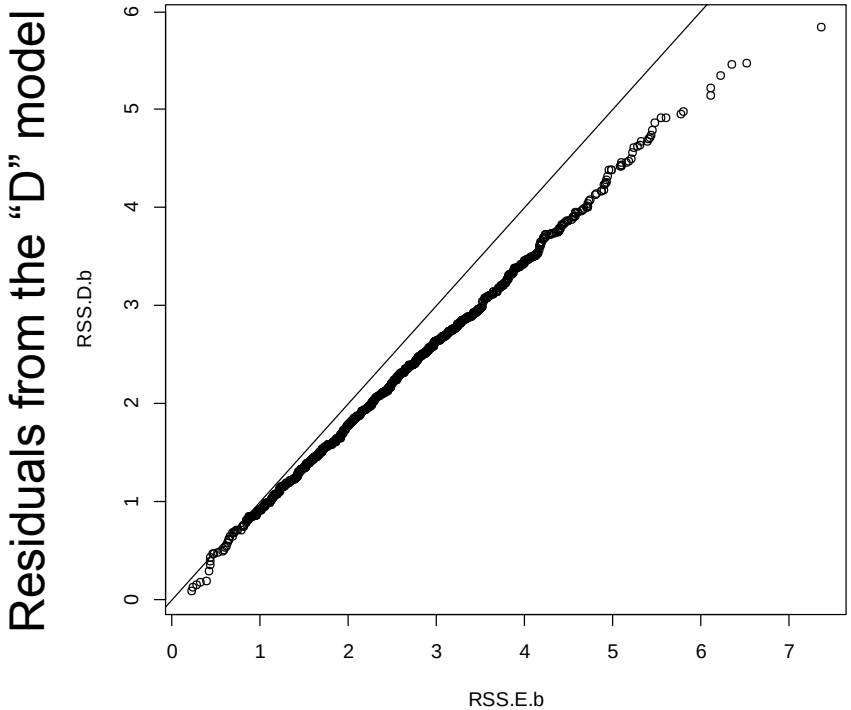



```
> ci<-quantile(theta,probs=c(0.025,0.975))
> ci
       2.5%       97.5%
-0.02450361  0.10272732


> sqrt(var(theta))
[1] 0.03357638

 sum(theta<theta.o)
[1] 622
```

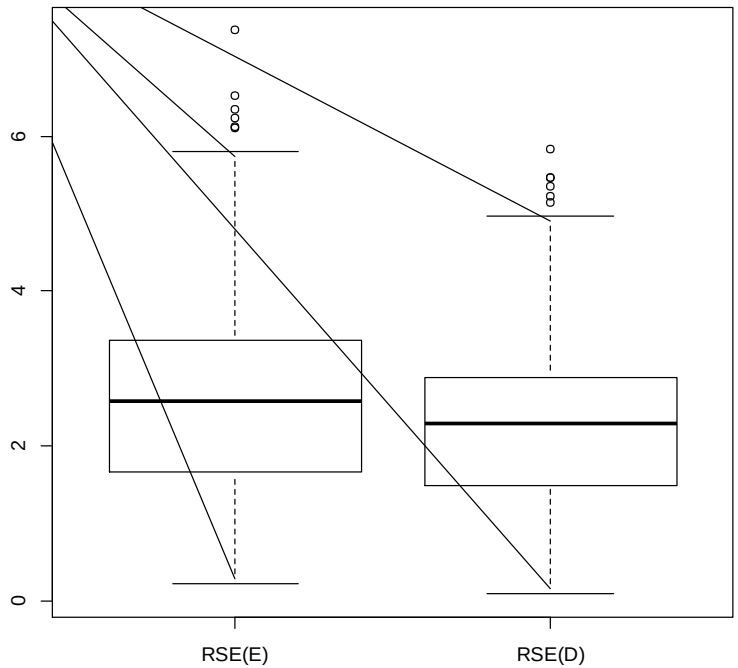
62.2% of $\hat{\theta}_b^* < \hat{\theta}$

Residuals sum of squares obtained for two models

qqplot:



Residuals from the "E" model



Higher residuals sum of squares obtained for the "E" model.

R code for the tooth strength data

```
pairs(tooth[,-c(1)])
fit.lm.D<-lm(strength~D1+D2,data=tooth)
summary(fit.lm.D)
fit.lm.E<-lm(strength~E1+E2,data=tooth)
summary(fit.lm.E)


par(mfrow=c(1,1))
plot(fit.lm.D$fit,fit.lm.E$fit)
abline(0,1)
plot(fit.lm.D$resid,fit.lm.E$resid)
abline(0,1)


n<-length(tooth$strength)
RSS.E<-sum((tooth$strength-
fit.lm.E$fit)^2)
RSS.E
RSS.D<-sum((tooth$strength-
fit.lm.D$fit)^2)
RSS.D
theta.o<-(RSS.E-RSS.D)/n
theta.o
```

R code for the tooth strength data

```
B<-1000
RSS.E.b<-RSS.D.b<-theta<-c(1:B)
index<-c(1:13)
for(i in 1:B)
{
index.b<-sample(index,n,replace=TRUE)
tooth.b<-tooth[index.b,]
fit.lm.D.b<-lm(strength~D1+D2,data=tooth.b)
fit.lm.E.b<-lm(strength~E1+E2,data=tooth.b)
RSS.E.b[i]<-sum((tooth.b$strength-fit.lm.E.b$fit)^2)
RSS.D.b[i]<-sum((tooth.b$strength-fit.lm.D.b$fit)^2)
theta[i]<-(RSS.E.b[i]-RSS.D.b[i])/n
}

hist(theta,nclass=50)
ci<-quantile(theta,probs=c(0.025,0.975))
ci
lines(c(ci[1],ci[1]),c(0,500),col=2,lwd=3)
lines(c(ci[2],ci[2]),c(0,500),col=2,lwd=3)
lines(c(theta.o,theta.o),c(0,500),col=4,lwd=3)

sqrt(var(theta))
sum(theta<theta.o)
qqplot(RSS.E.b,RSS.D.b)
abline(0,1)
boxplot(RSS.E.b,RSS.D.b,names=c("RSE(E)","RSE(D)"))
```

Bootstratp for generalized Linear Models (GLM)

Topics

- GLMs.
- Models for binary data:
  - Non parametric bootstrap.
  - Parametric bootstrap.
- Estimation & C.I.
- Inference.
- Examples:
  - Birth weights (g) and estimated gestational age (without bootstrap).
  - Dose-response data (the beetle data).
  - Serological data (malaria).

Both datasets are external dataset.

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.

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.

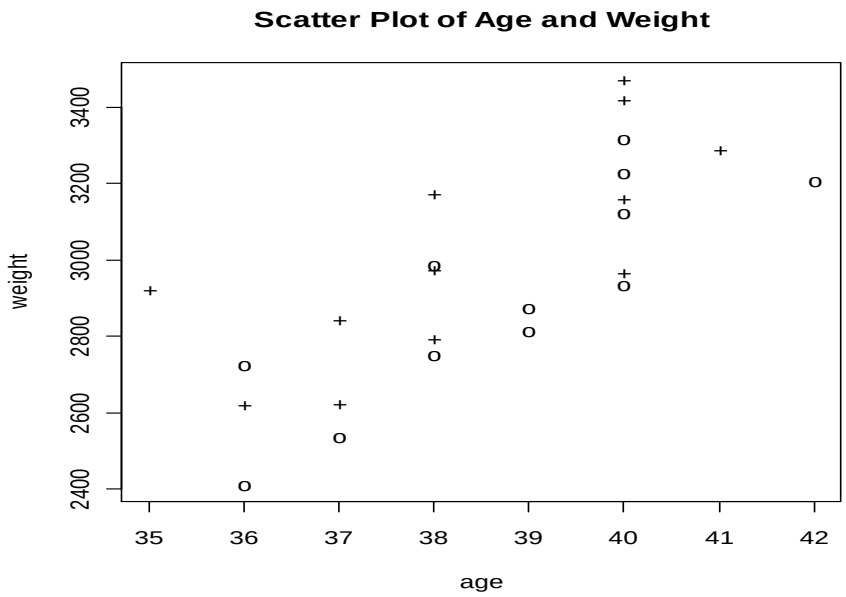
$$\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

3. **Link function-**Specify the relationship between the mean E(Y) and the systematic component.

Example 1: Body weight and gestational age

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

Two predictors: age and gender.



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 componnet: the linear predictor

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

The link function

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

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

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

Components of a GLM: linear regression models

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

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

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

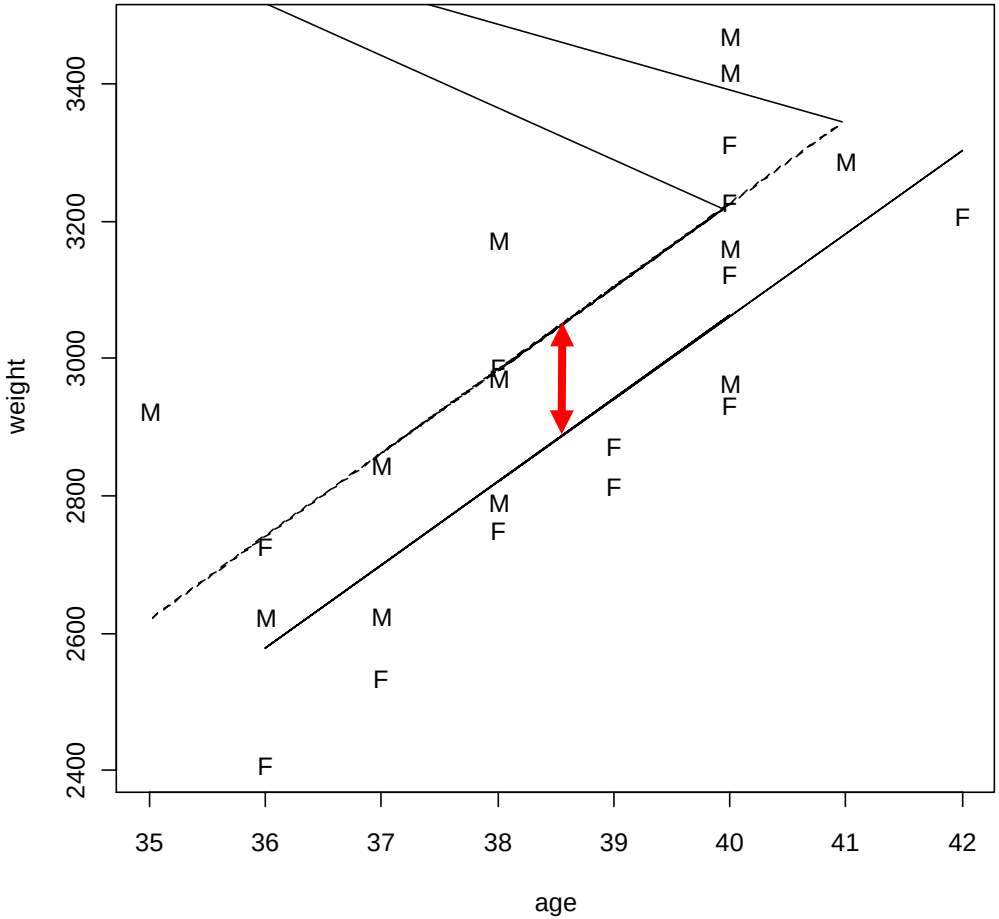
The link function (=the link between the random and the systematic part)

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

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

$$g = 1$$

Example 1 : linear regression model



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

$$\underbrace{\beta_0 + \beta_1 \times age_i + \beta_2 \times gender_i}$$

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

Example 2:
dose-response experiment

Dose-response data

Binary outcome 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

In R (external file):

```
beetle<-read.table("C:/projects/GLM/data4glm/beetle.txt", header = TRUE)
```

Random component: example of binary data

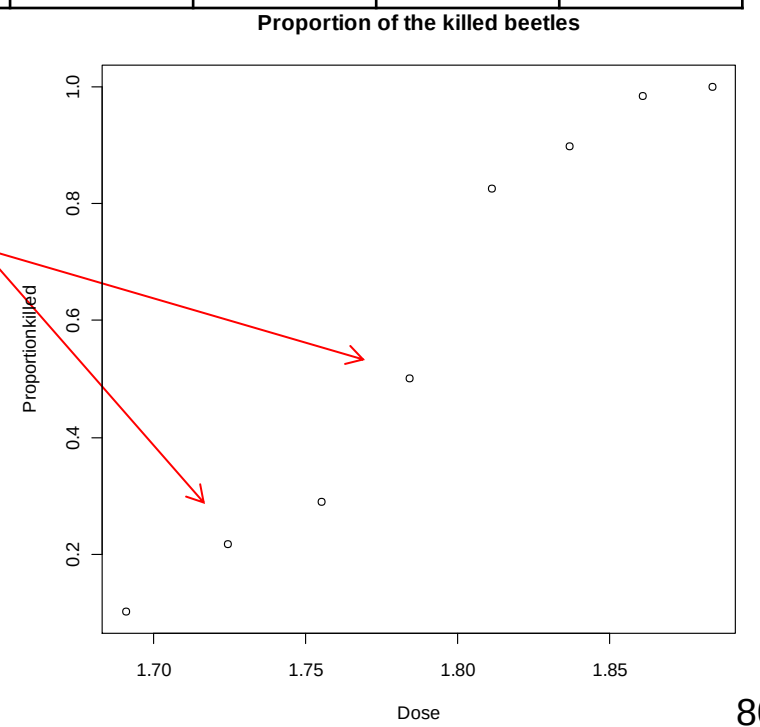
Dose	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
Beetles	59	60	62	56	63	59	62	60
Killed	6	13	18	28	52	53	61	60

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

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

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

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



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.

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

$$\pi_j = f(d_j) = f(\underbrace{\beta_0 + \beta_1 d_j}_{\text{The linear predictor}})$$

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

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

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

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

Parametric bootstrap

We assume that the number of deaths (y) is a function of the dose.

Population

$$y_j = \sum Y_{ij} \Rightarrow y_j \sim B(n_j, \pi_j)$$

Bootstrap sample

$$y_j^* \sim B(n_j, \hat{\pi}_j)$$

$$\hat{\pi}_j = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 d_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 d_j)}$$

Parametric bootstrap

One bootstrap sample
For i=1,..,n

GLM

$$y_j^* \sim B\left(n_j, \hat{\pi}_j\right) \Longleftrightarrow \pi_j = f(x_j)$$

$$\left(x_1, y_1^*\right)$$
$$\left(x_2, y_2^*\right)$$
$$\underbrace{\left(x_n, y_n^*\right)}_{b=1}$$

$$\left(x_1, y_1^*\right)$$
$$\left(x_2, y_2^*\right)$$
$$\underbrace{\left(x_n, y_n^*\right)}$$

$$\left(x_1, y_1^*\right)$$
$$\left(x_2, y_2^*\right)$$
$$\underbrace{\left(x_n, y_n^*\right)}_{b=B}$$

Non parametric bootstrap

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

The observation:

$$(Y_{ij}, d_j)$$

"zero/one" data

Data structure

$$\begin{array}{cl} \begin{bmatrix} Y_{11} & d_1 \\ Y_{11} & d_1 \\ \vdots & \vdots \\ Y_{n_1 1} & d_1 \end{bmatrix} & \text{Dose level 1} \\ \begin{bmatrix} Y_{12} & d_2 \\ \vdots & \vdots \\ Y_{n_2 2} & d_2 \end{bmatrix} & \text{Dose level 2} \\ \begin{bmatrix} \vdots & \vdots \\ Y_{1K} & d_K \end{bmatrix} & \text{Dose level K} \\ \begin{bmatrix} \quad & \quad \end{bmatrix} & \end{array}$$

Non parametric bootstrap

Estimation of π_j

Bootstrap with replacement pairs

$$\begin{bmatrix} 1 & d_1 \\ 1 & d_1 \\ \vdots & \vdots \\ 0 & d_1 \\ 0 & d_2 \\ \vdots & \vdots \\ 1 & d_2 \\ \vdots & \vdots \\ 0 & d_K \\ 1 & d_K \end{bmatrix}$$

Dose level 1

Dose level 2

Dose level K

Data structure

$$\begin{bmatrix} Y_{11} & d_1 \\ Y_{11} & d_1 \\ \vdots & \vdots \\ Y_{n_1 1} & d_1 \\ Y_{12} & d_2 \\ \vdots & \vdots \\ Y_{n_2 2} & d_2 \\ \vdots & \vdots \\ Y_{1K} & d_K \\ & \end{bmatrix}$$

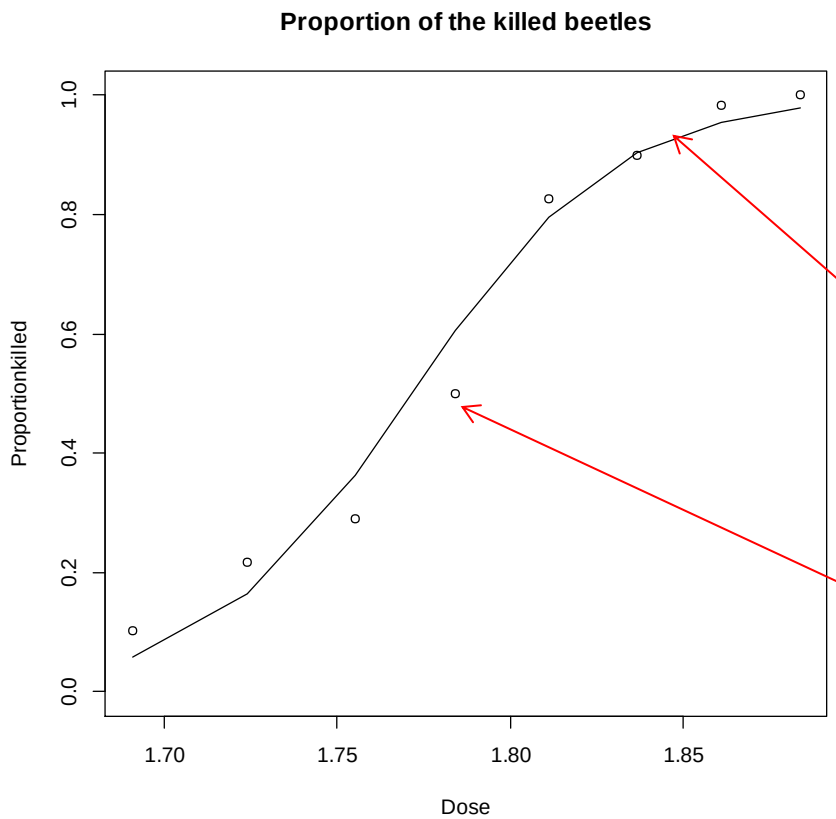
Dose level 1

Dose level 2

Dose level K

Resampling by dose level ?

Data and fitted model



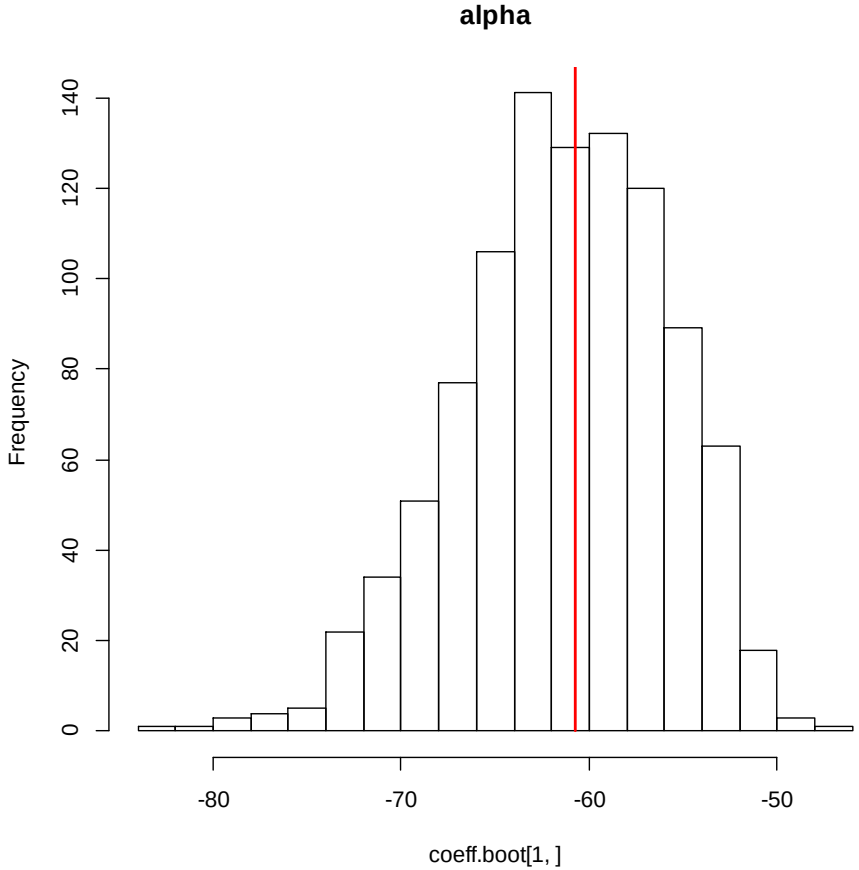
Two options for the estimates of the probability:

$$\hat{\pi}_j = \frac{\exp(-60.71 + 34.27 \times d_j)}{1 + \exp(-60.71 + 34.27 \times d_j)}$$

$$\hat{\pi}_j = \frac{y_j}{n_j}$$

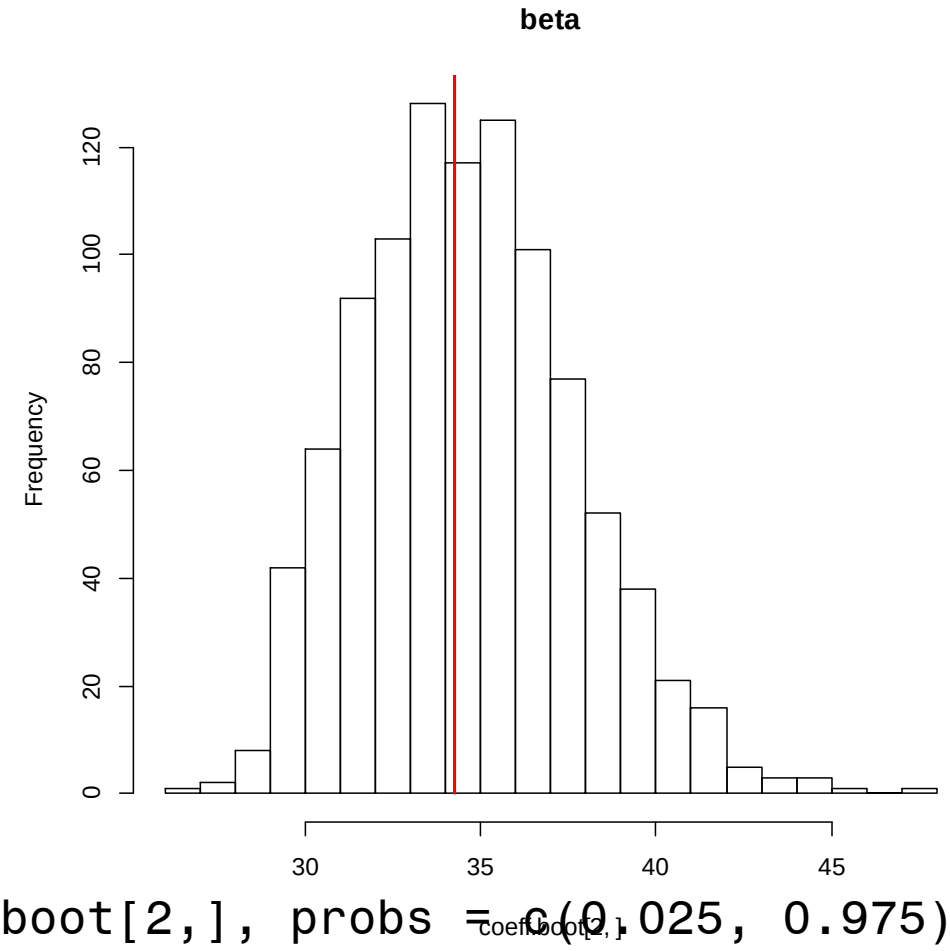
```
> fit.beetles<-glm(cbind(killed,unkilled)~Dose,family=binomial(link = "logit"))
> summary(fit.beetles)$coefficients
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -60.71745   5.180701 -11.71993 1.007549e-31
Dose         34.27033   2.912134  11.76811 5.698445e-32
```

Distribution of the bootstrap replicates for the intercept



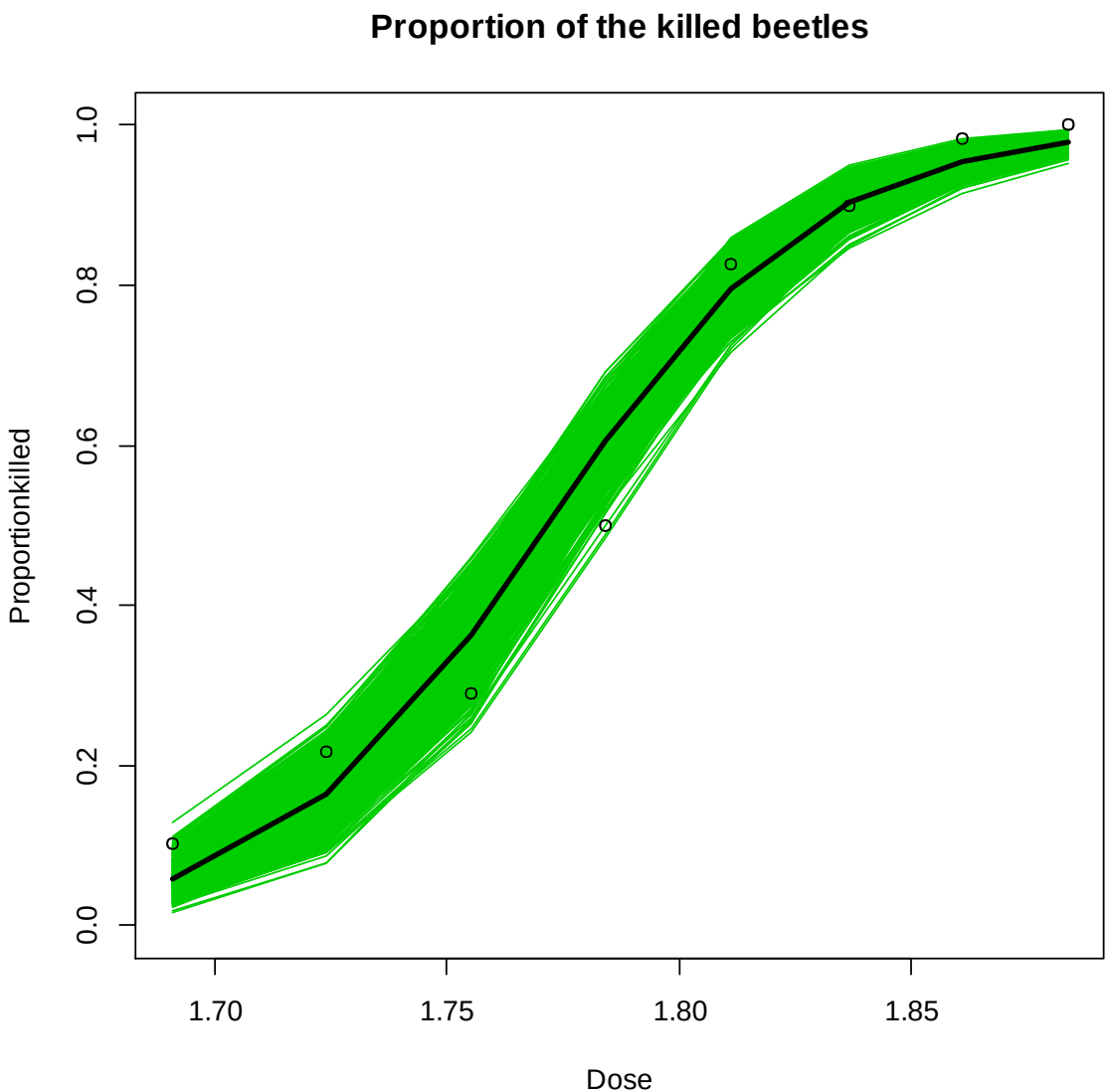

```
> quantile(coeff.boot[1,], probs = c(0.025, 0.975))
     2.5%     97.5%
-73.11089 -52.12720
```

Distribution of the bootstrap replicates for the slope




```
quantile(coeff.boot[2,], probs = c(0.025, 0.975))
    2.5%    97.5%
29.48047 41.16337
```

Data and predicted models



R code

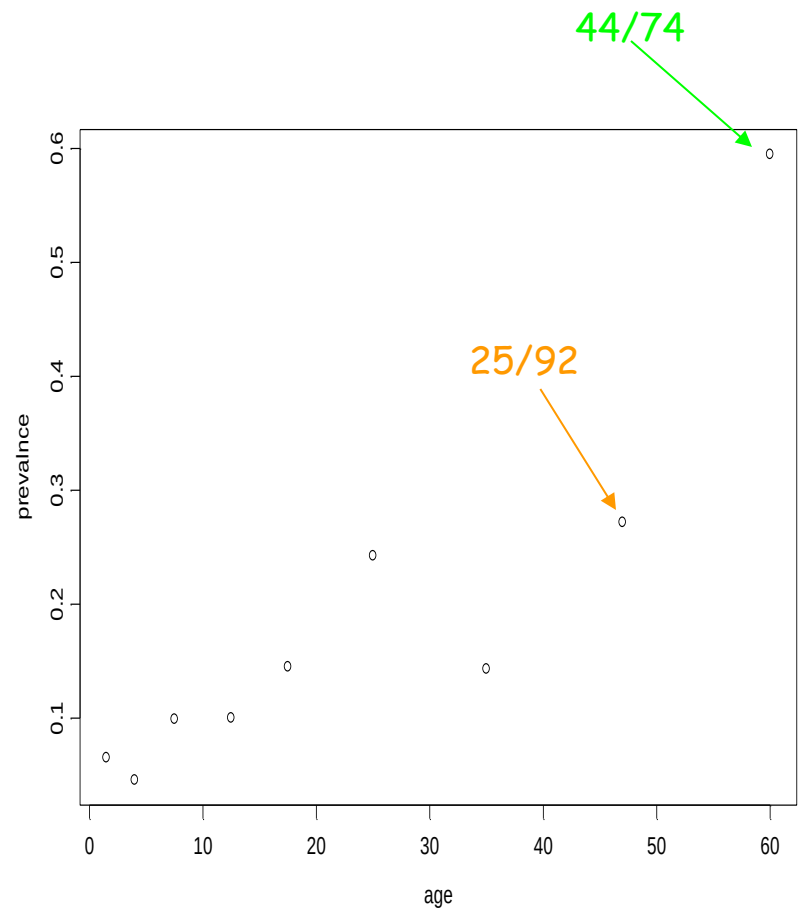
```
attach(beetle)
p.b<-killed/beetles
B<-1000
prob.boot<-matrix(0,8,B)
test.stat<-coeff.boot<-matrix(0,2,B)
pos.boot<-c(1:8)
for(b in 1:B)
{
for(i in 1:8)
{
pos.boot[i]<-sum(rbinom(beetles[i],1,p.b[i]))
}
neg.boot<-beetles-pos.boot
fit.boot<-glm(cbind(pos.boot,neg.boot)~Dose,family=binomial(link = "logit"))
prob.boot[,b]<-fit.boot$fit
coeff.boot[,b]<-fit.boot$coefficients
test.stat[,b]<-summary(fit.boot)$coefficients[,3]
}
```

Example 3: serological data

Estimation & C.I

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



Serological data

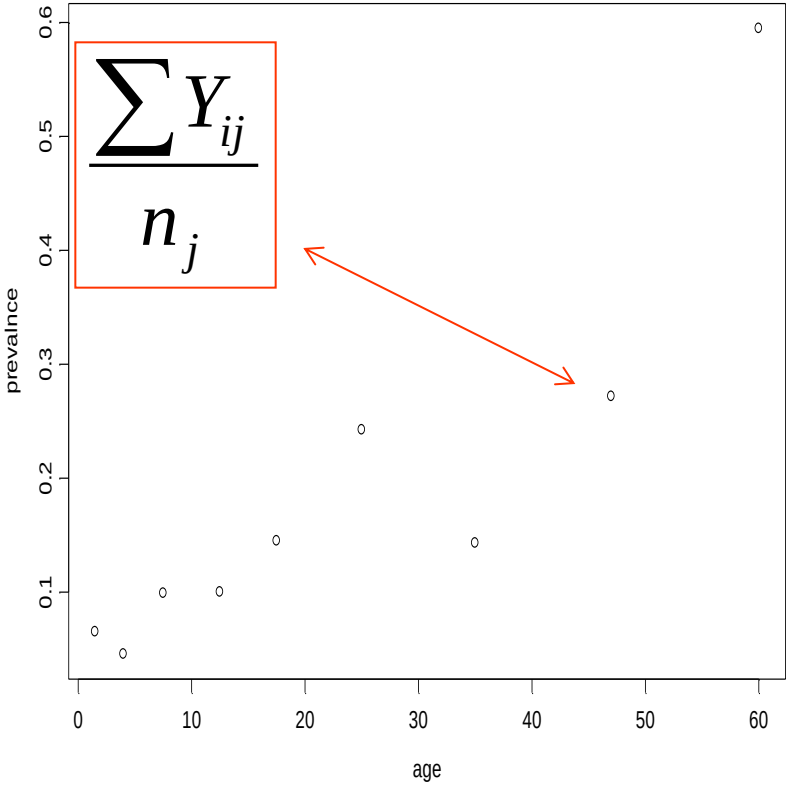
$$Y_{ij} = \begin{cases} 1 \\ 0 \end{cases}$$

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

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

$$Y_j = \sum Y_{ij}$$

$$Y_j \sim B(n_j, \pi_j)$$



The GLM

The distribution of the response

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

The expected values of the response variable

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

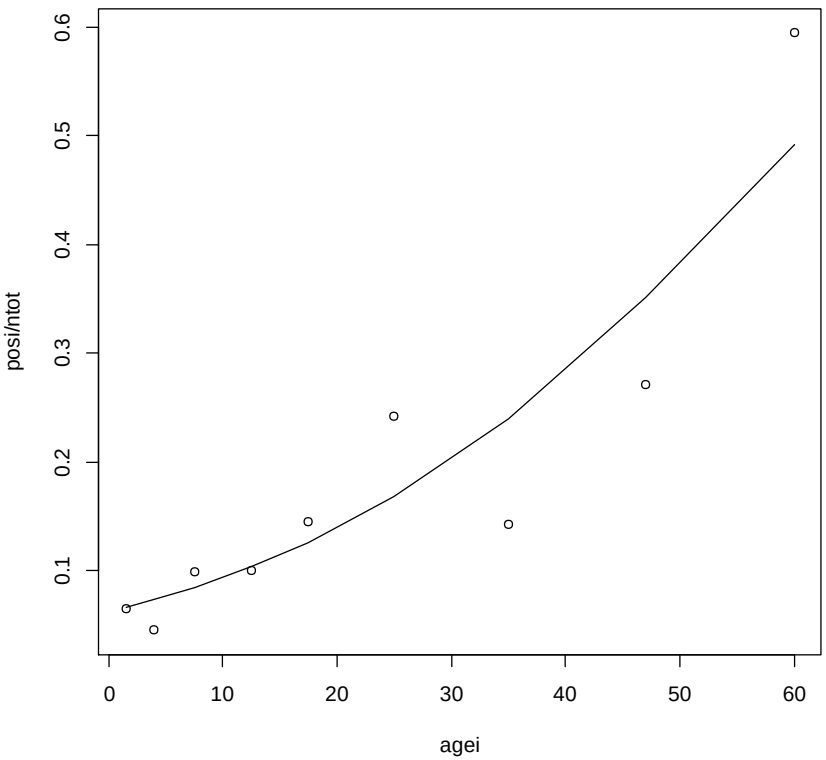
The systemetic part

$$\pi_j = f(\beta_0 + \beta_1 \times age_j) = f(\eta)$$

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

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

Data and fitted model



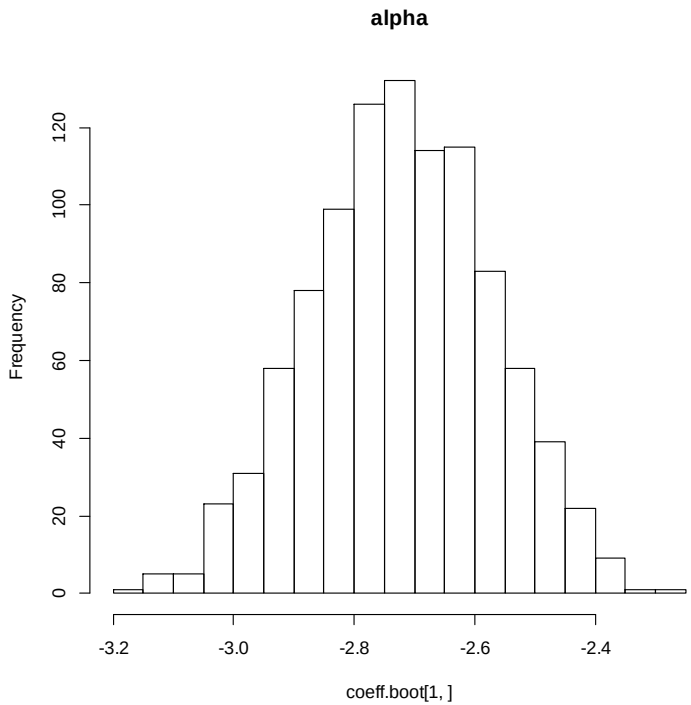
$$\hat{\beta}_0 = -2.7140$$

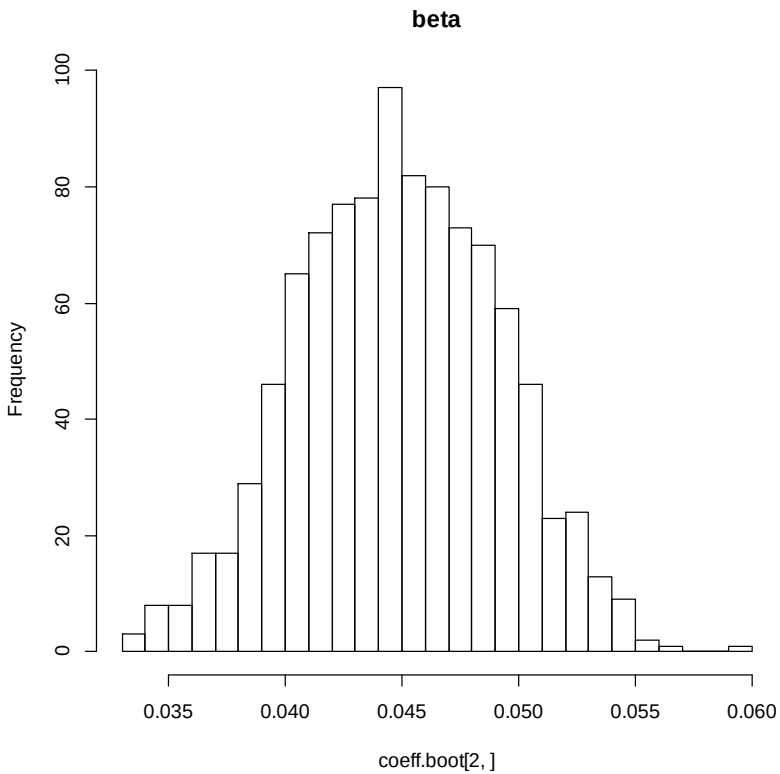
$$\hat{\beta}_1 = 0.04467$$

$$\pi(age) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1)}$$

```
> fit.malaria1<-glm(cbind(posi,negi)~agei,family=binomial(link = "logit"))
> summary(fit.malaria1)$coefficients
               Estimate  Std. Error    z value     Pr(>|z|)
(Intercept) -2.71407363 0.151740046 -17.886337 1.506942e-71
agei         0.04467246 0.004510723   9.903614 4.014980e-23
```

Distribution of bootstrap replicates for the intercept an slope

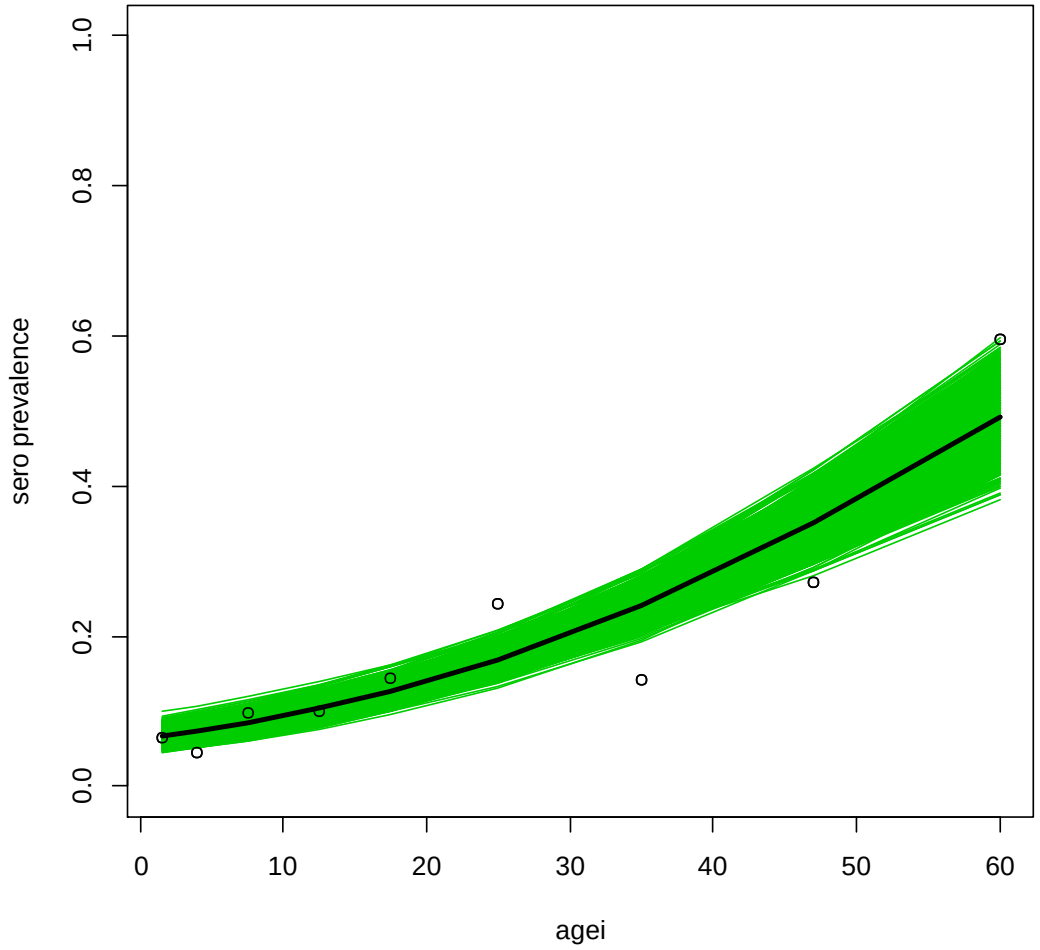





```
> quantile(coeff.boot[1,],prob c(0.025,0.975))
     2.5%     97.5%
-3.019566 -2.433120
```

```
> quantile(coeff.boot[2,],probs=c(0.025, 0.975))
      2.5%      97.5%
0.03646014 0.05301919
```

Data and bootstrap prediction band



Bootstrap test

Hypothesis testing

$$y_j \sim B(n_j, \pi_j)$$

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

$$\eta = \beta_0 + \beta_1 \times X_j$$

A GLM

We would like to test the hypothesis that the covariate has no effect on the probability

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Test of hypotheses

$$H_0 : g(\pi_j) = \beta_0$$
$$H_1 : g(\pi_j) = \beta_0 + \beta_1 \times age_j$$

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

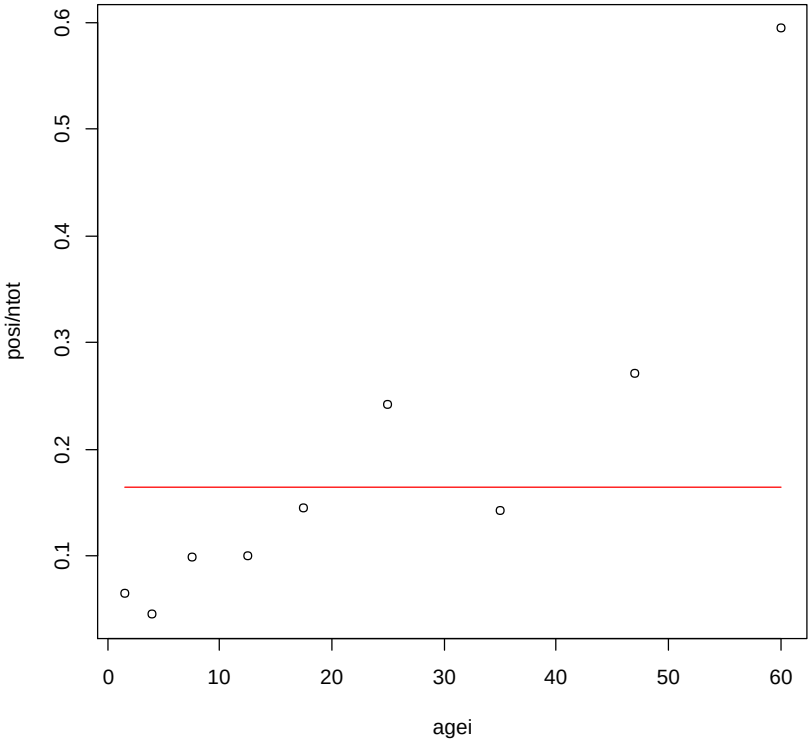
Calculate the observed statistics for the parameter of primary interest.

Re sample B bootstrap samples UNDER the null hypothesis and calculate the bootstrap replicates for statistics.

Calculate Monte Carlo p values

$$P = \frac{\#\left\{\hat{\beta}_1^* \geq \hat{\beta}_1\right\} + 1}{B + 1}$$

The null model



Under the null hypothesis

$$y_j \sim B(n_j, \pi)$$

$$\hat{\beta}_0 = -1.626$$

$$\hat{\pi} = 0.16425$$

```
> summary(fit.malaria0)$coefficients
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -1.626901 0.07873747 -20.66235 7.559447e-95
> eta<- -1.62690
> prob.i<-exp(eta)/(1+exp(eta))
> prob.i<-rep(prob.i,9)
> prob.i
[1] 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555
```

Test of hypotheses: parametric bootstrap

Re sampling under the null hypothesis

$$\pi(age) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

$$y_j \sim B\left(n_j, 0.16425\right)$$

The null hypothesis and the distribution of the response under the null hypothesis

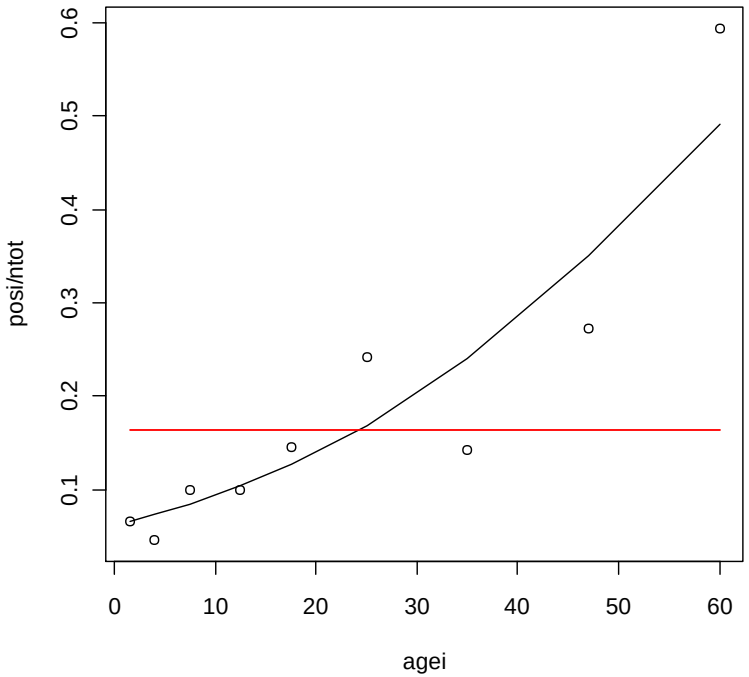
$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

$$H_0 : \log it(P_i) = \alpha$$

$$H_1 : \log it(P_i) = \alpha + \beta \times age_i$$

$$y_i \sim B(n_i, P_i) \qquad P_i = \frac{e^{\alpha}}{1+e^{\alpha}}$$



Parametric bootstrap

The data

$$((y_1, n_1, age_1), (y_2, n_2, age_2), \ldots, (y_n, n_n, age_n))$$

Bootstrap sample (y_i^*, n_i, age_i)

Bootstrap sample (y_i^*, n_i, age_i)

Bootstrap sample (y_i^*, n_i, age_i)

For each age group:

$$y_i^* \sim B(n_i, P_{H_0})$$

Parametric bootstrap

For each bootstrap sample:

$$y_i^* \sim B(n_i, P_{H_0})$$

$$g(P_i) = \alpha + \beta \times age_i$$

$$\Downarrow$$

$$\hat{\alpha}_b^*, \hat{\beta}_b^*$$

The distribution under the null hypothesis

$$\hat{\beta} \overset{H_0}{\sim} G$$

The bootstrap replicates

$$\hat{\beta}_1^*, \hat{\beta}_2^*, \ldots \hat{\beta}_B^*$$

$$\hat{\beta} \sim \hat{G}_{H_0}$$

Parametric bootstrap

One bootstrap sample
For i=1,..,n

$$y_j^* \sim B(n_j, 0.16425)$$

$$\underbrace{\begin{matrix}(x_1, y_1^*)\\(x_2, y_2^*)\\(x_n, y_n^*)\end{matrix}}_{b=1} \qquad \underbrace{\begin{matrix}(x_1, y_1^*)\\(x_2, y_2^*)\\(x_n, y_n^*)\end{matrix}} \qquad \underbrace{\begin{matrix}(x_1, y_1^*)\\(x_2, y_2^*)\\(x_n, y_n^*)\end{matrix}}_{b=B}$$

Parametric bootstrap

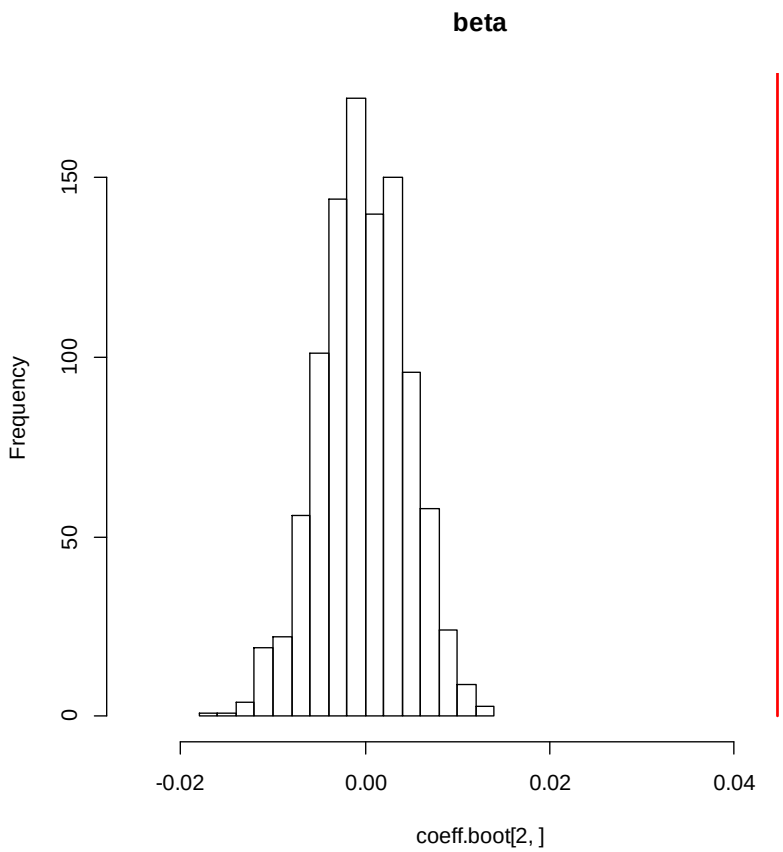
For each bootstrap sample

$$\left(x_1, y_1^*\right)$$
$$\left(x_2, y_2^*\right)$$
$$\left(x_n, y_n^*\right)$$

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

$$\hat{\beta}_{1,1}^*, \hat{\beta}_{1,2}^*, \ldots\ldots\ldots, \hat{\beta}_{1,B}^*$$

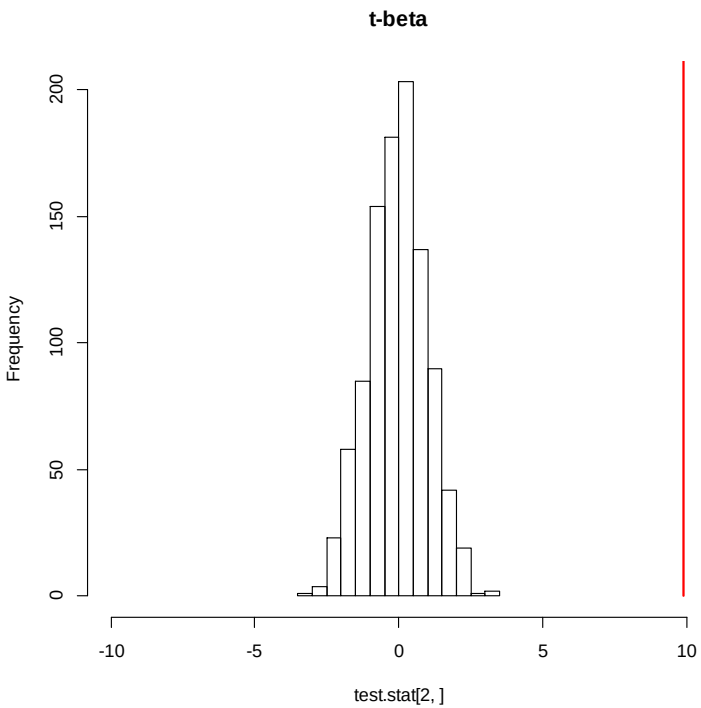
The distribution of the parameter estimate under the null hypothesis



$$\hat{\beta}^*_{1,1}, \hat{\beta}^*_{1,2}, \ldots\ldots\ldots, \hat{\beta}^*_{1,B}$$

$$P = \frac{\#\left\{\hat{\beta}^*_1 \geq \hat{\beta}_1\right\} + 1}{B+1}$$

The distribution of the test statistiec under the null hypothesis



$$t_b^* = \frac{\hat{\beta}_{1,b}^*}{S.E(\hat{\beta}_{1,b}^*)}$$

```
> fit.malaria1<-glm(cbind(posi,negi)~agei,family=binomial(link = "logit"))
> summary(fit.malaria1)$coefficients
               Estimate  Std. Error    z value     Pr(>|z|)
(Intercept) -2.71407363 0.151740046 -17.886337 1.506942e-71
agei         0.04467246 0.004510723   9.903614 4.014980e-23
```

Test of hypotheses: non parametric bootstrap

Parametric bootstrap: re sampling under the null hypothesis

$$\pi(age) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

Under the null hypothesis age does not influence the prevalence.

Test of hypotheses:non parametric bootstrap

Parametric bootstrap:

Re sampling under the null hypothesis

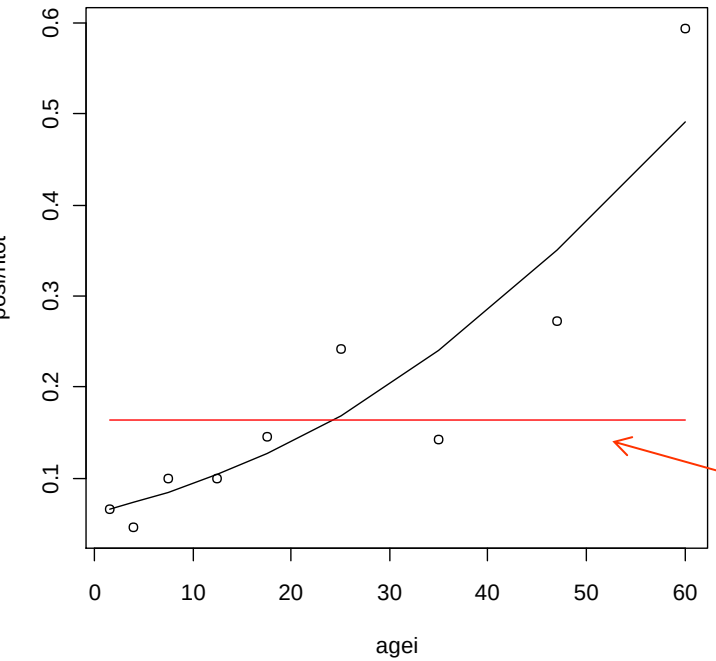
$$\pi(age) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

Non parametric bootstrap:
Bootstrap with replacement:
fix x and resample Y
(zero/one data)

Data structure

$$\begin{array}{cl} \begin{bmatrix} Y_{11} & d_1 \\ Y_{11} & d_1 \\ \vdots & \vdots \\ Y_{n_1 1} & d_1 \\ Y_{12} & d_2 \\ \vdots & \vdots \\ Y_{n_2 2} & d_2 \\ \vdots & \vdots \\ Y_{1K} & d_K \end{bmatrix} & \begin{array}{l} \text{Dose level 1} \\ \\ \text{Dose level 2} \\ \\ \text{Dose level K} \end{array} \\ \text{fixed} & \end{array}$$

Parametric bootstrap in R




```
> eta<- -1.62690
> prob.i<-exp(eta)/(1+exp(eta))
> prob.i<-rep(prob.i,9)
> prob.i
[1] 0.1642555 0.1642555 0.1642555 0.1642555 0.1642555
0.1642555 0.1642555
[8] 0.1642555 0.1642555
```

Probability to be infected under the null hypothesis

```
for(b in 1:B)
{
for(i in 1:9)
{
pos.boot[i]<-sum(rbinom(ntot[i],1,prob.i[i]))
}
neg.boot<-ntot-pos.boot
fit.boot<-glm(cbind(pos.boot,neg.boot)~agei,family=binomial(link = "logit"))
prob.boot[,b]<-fit.boot$fit
coeff.boot[,b]<-fit.boot$coefficients
test.stat[,b]<-summary(fit.malaria1)$coefficients[,3]
}
```