Linear regression and correlation

Ariel Alonso Abad

Catholic University of Leuven

Association and correlation, their scientific relevance

- Discovering associations is fundamental in science
- Many scientific hypotheses are stated in terms of correlation or lack of correlation
- Although correlation does not imply causation, causation does imply correlation. That is, although a correlational study cannot definitely prove a causal hypothesis, it may rule one out
- Some variables simply cannot be manipulated for ethical reasons. Other variables, such as birth order, sex, and age are inherently correlational because they cannot be manipulated and, therefore, the scientific knowledge concerning them must be based on correlation evidence

Association and correlation, their scientific relevance

- Once correlation is known it can be used to make predictions
- When we know a score on one measure we can make a more accurate prediction of another measure that is highly related to it. The stronger the relationship between/among variables the more accurate the prediction
- Practical evidence from correlation studies can lead to testing that evidence under controlled experimental conditions
- Complex correlational statistics like multiple regression and partial correlation allow the correlation between two variables to be recalculated after the influence of other variables is removed

Kalama study

Kalama study

As part of an investigation into the physical development of children a health scientist measured the age (in months) and the height (in cm) of 12 children in the Kalama province in Egypt.

Research question: Is there a relationship between length and age?

Data

age	18	19	20	21	22	23	24	25	26	27	28	29
height	76.1	77.0	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

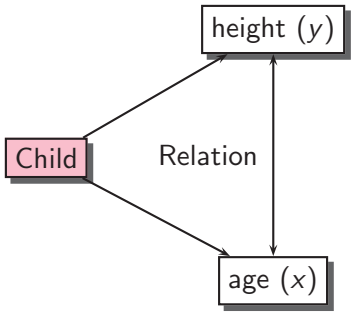
Kalama study

Kalama study

As part of an investigation into the physical development of children a health scientist measured the age (months) and the height (cm) of 12 children in the Kalama province in Egypt.

Research question: Is there a relationship between length and age?

Two variables measured for every child in the sample



- Strength of the association
- Is the relationship linear, quadratic, exponential?

Pearson correlation coefficient

Correlation coefficient

$$r_{xy}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

- Range: $-1 \leq r_{xy} \leq 1$
- Perfect positive correlation between x and y: $r_{xy} = 1$
- No correlation between x en y: $r_{xy} = 0$
- Perfect negative correlation between x and y: $r_{xy} = -1$

Reading the data in R

```
> # Defining working directory
> setwd("C:\R-code-data")
>
> ## Reading the data
>
> kalama=read.table("kalama.txt", header=T)
> kalama
>
   age height
1   18   76.1
2   19   77.0
3   20   78.1
4   21   78.2
5   22   78.8
6   23   79.7
7   24   79.9
8   25   81.1
9   26   81.2
10  27   81.8
11  28   82.8
12  29   83.5
>
```

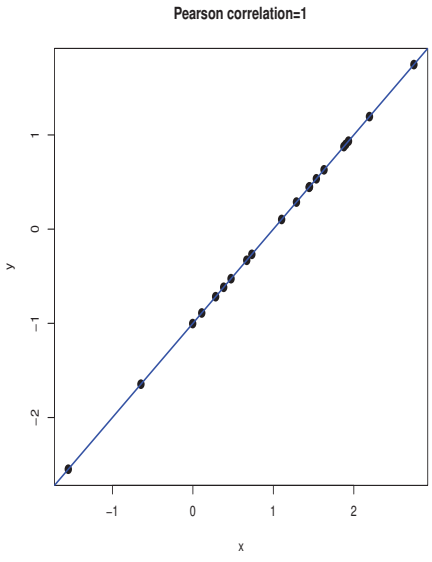

Descriptive statistics in R

```
> ## Descriptive Statistics
>
> options(digits=2)
> descrip.kalama<-stat.desc(kalama[,c("age","height")],basic=TRUE, desc=TRUE)
> descrip.kalama
>
                 age  height
nbr.val       12.00  12.000
min           18.00  76.100
max           29.00  83.500
range         11.00   7.400
sum          282.00 958.200
median        23.50  79.800
mean          23.50  79.850
SE.mean        1.04   0.665
CI.mean.0.95   2.29   1.463
var           13.00   5.301
std.dev        3.61   2.302
coef.var       0.15   0.029
>
```

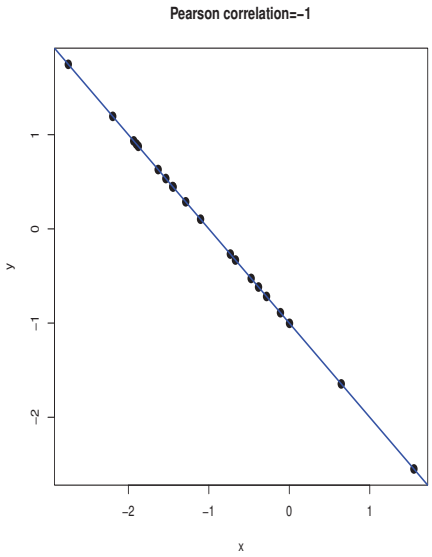
Estimating correlations in R

```
> ## Calculating the covariance and correlatio
> cov.age.height<-cov(kalama$age,kalama$height)
> corr.age.height<-cor(kalama$age,kalama$height)
> cov.age.height
[1] 8.3
> corr.age.height
[1] 0.99
> ## Testing if the population correlation is zero
> corr.age.height.test= cor.test(kalama$age, kalama$height,
+                        alternative="two.sided", method = "pearson")
> corr.age.height.test

        Pearson's product-moment correlation

data:  kalama$age and kalama$height
t = 30, df = 10, p-value = 4.428e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.98 1.00
sample estimates:
 cor
0.99
>
```

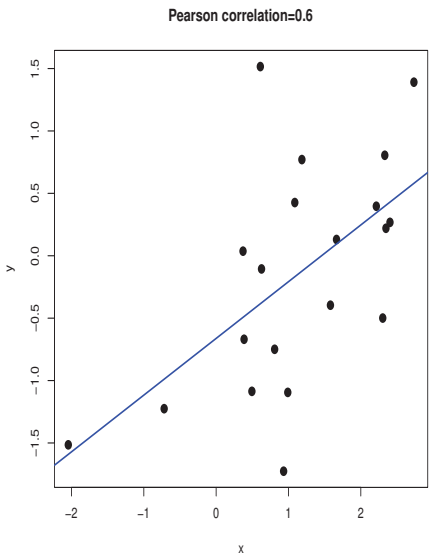

Pearson correlation coefficient



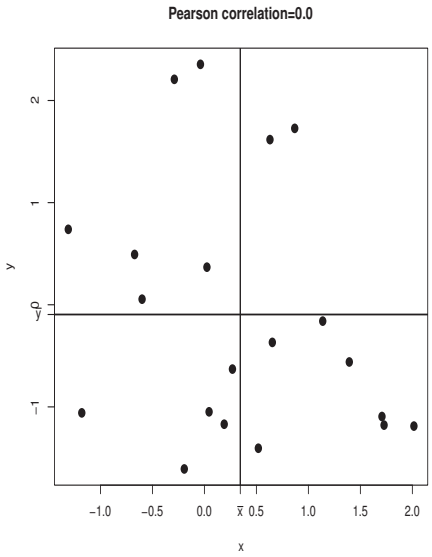
Pearson correlation coefficient



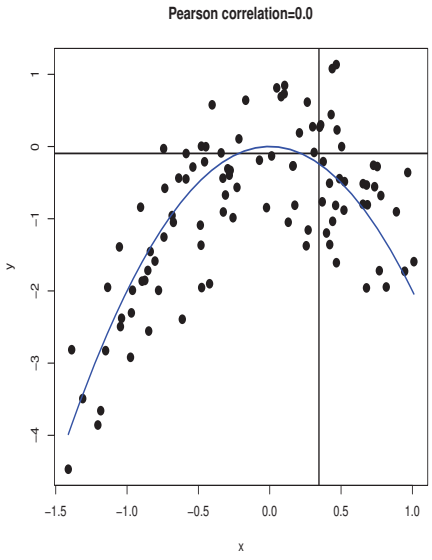
Pearson correlation coefficient



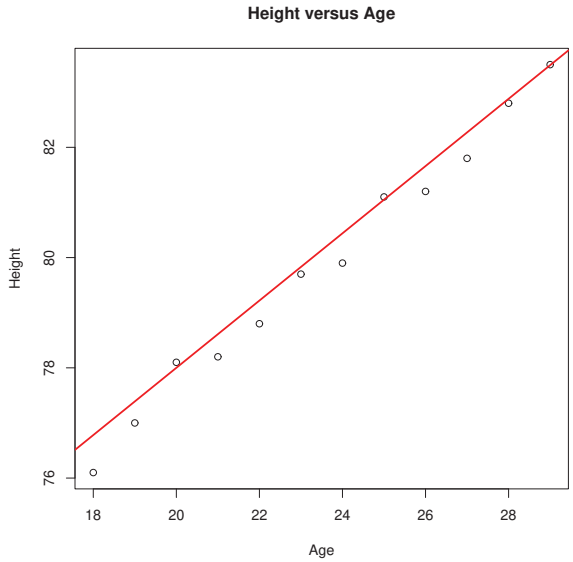
Pearson correlation coefficient



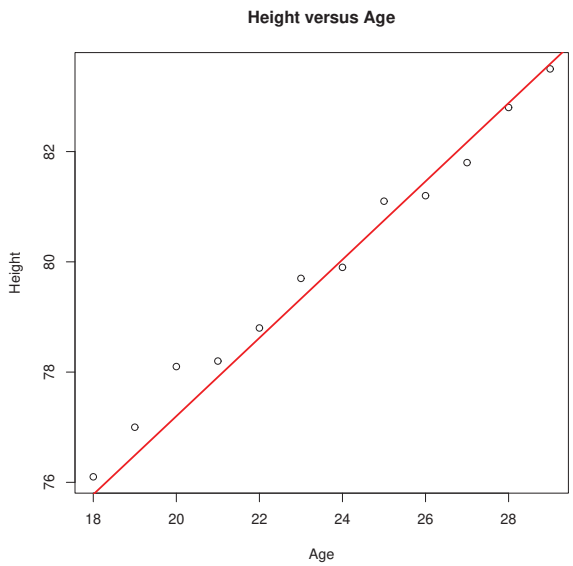
Pearson correlation coefficient



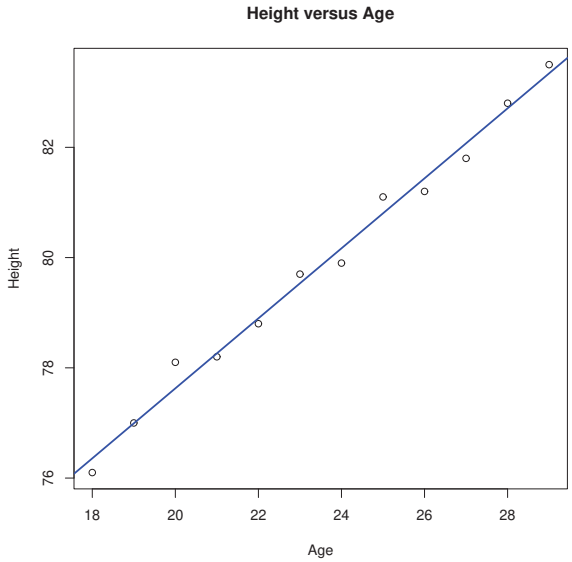
Kalama study ($r_K = 0.994$): Best line



Kalama study ($r_K = 0.994$): Best line



Kalama study ($r_K = 0.994$): Best line



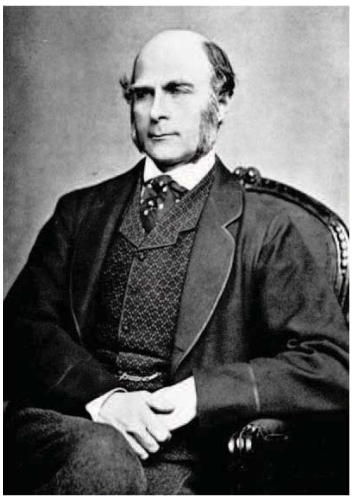
When to use Regression Analysis

- Regression analysis is used for explaining or modeling the relationship between a single variable Y, called the response output or dependent variable, and one or more predictor or explanatory variables, $\boldsymbol{X}^{'} = (X_1, \ldots, X_p)$
- When $p = 1$ it is called **simple** regression but when $p > 1$ it is called **multiple** regression
- When there is more than one Y, then it is called multivariate multiple regression which we won't be covering here
- The response must be a continuous variable but the explanatory variables can be continuous, discrete or categorical

Regression Analysis: Possible objectives

- Prediction of future observations
- Assessment of the effect of, or relationship between, explanatory variables on the response
- A general description of data structure
- Extensions exist to handle multivariate responses, binary responses (logistic regression analysis) and count responses (Poisson regression)

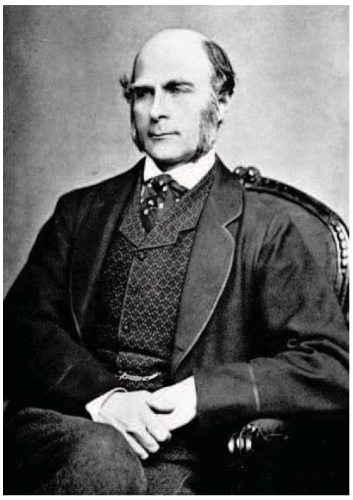
Francis Galton



- Cousin of Charles Darwin
- Regression and correlation
- The phenomenon of regression towards the mean

"Regression towards mediocrity in hereditary stature". Journal of the Anthropological Institute 15 (1886), 246-263.

Francis Galton

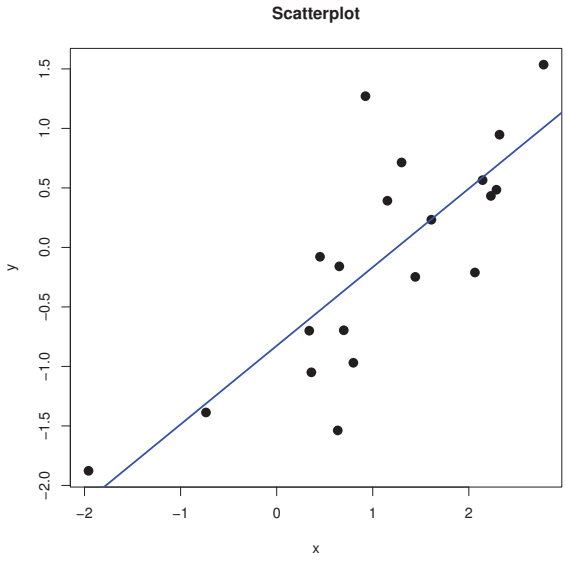


Francis Galton coined the term regression to mediocrity in 1875 in reference to the simple regression equation in the form

$$\frac{y-\bar{y}}{s_y} = r\left(\frac{x-\bar{x}}{s_x}\right).$$

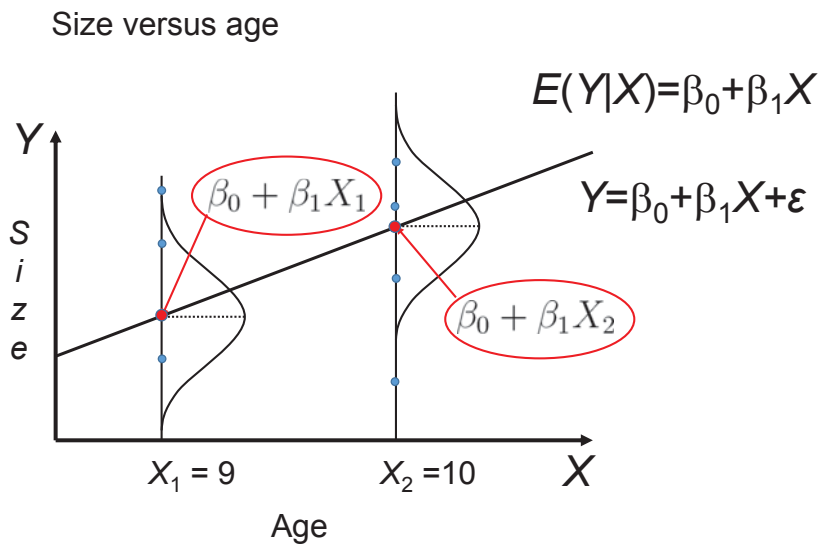
Sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers. The **regression** effect.

Linear regression



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Linear regression



Formal Statement of the Model

For each unit $i = 1, \ldots, n$, the value of explanatory variable X_i and the response Y_i are recorded. *Simple Linear Regression* model.

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

Assumptions

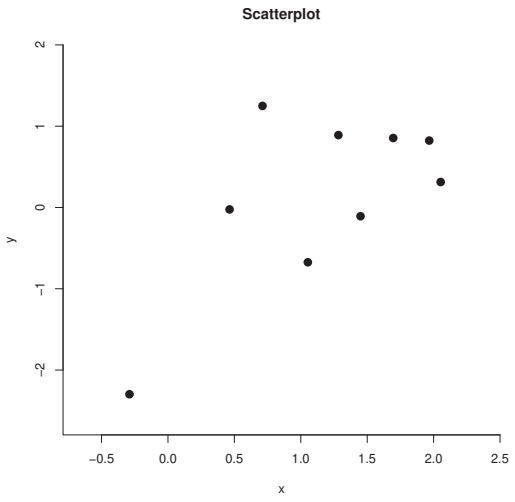
1. The value of X_i is precisely known.
2. Y_i is a continuous random variable.
3. β_0 and β_1 are parameters. That is, they are: unknown, constant and do not depend on the research unit.
4. ε_i is a random error term. It is not observable.

Formal Statement of the Model

Additional assumptions

5. For two different units, i and j, ε_i and ε_j are independent.
6. X_i and ε_i are independent.
7. $\varepsilon_i \sim N(0, \sigma^2)$ for all i, i.e., ε_i is normally distributed with $E(\varepsilon_i) = 0$, and $Var(\varepsilon_i) = \sigma^2$ for all i

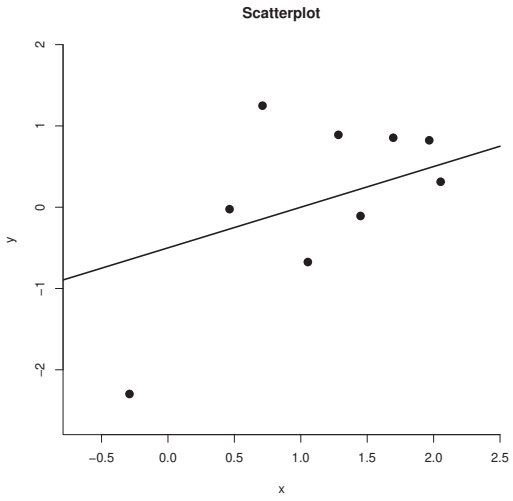
Least squares method



Which line fits the data best?

- $Y = \beta_0 + \beta_1 X + \varepsilon$

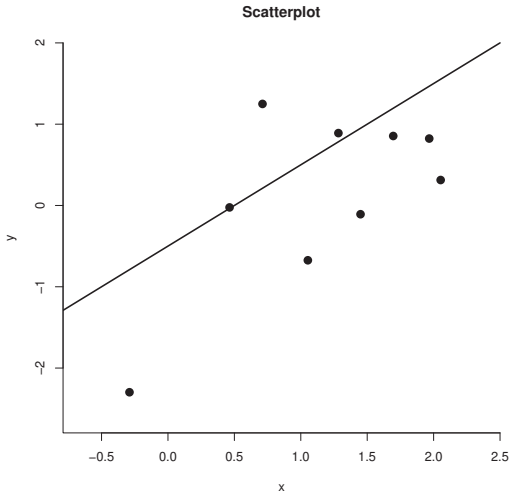
Least squares method



Which line fits the data best?

- $\hat{Y} = -0.5 + 0.5X$

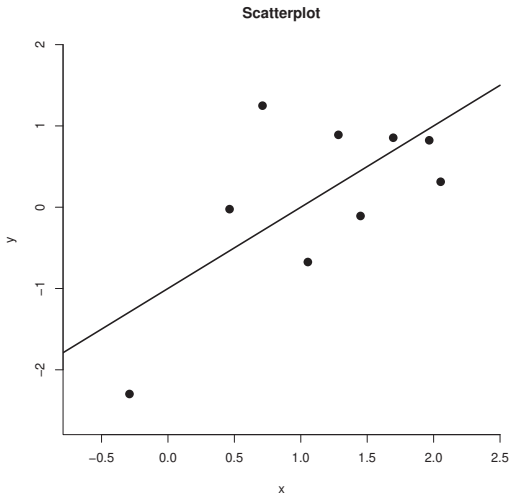
Least squares method



Which line fits the data best?

- $\hat{Y} = -0.5 + X$

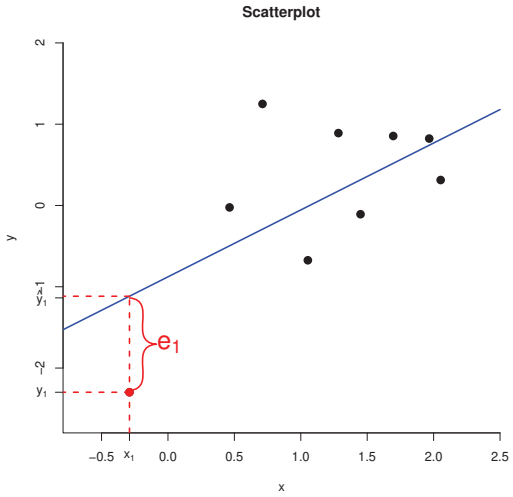
Least squares method



Which line fits the data best?

- $\hat{Y} = -0.26 + 0.45X$

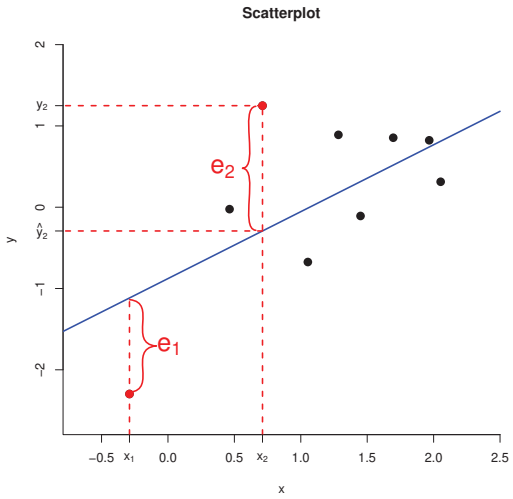
Least squares method



Which line fits the data best?

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$
$$y_1 = \hat{y}_1 + e_1$$
$$e_1 = y_1 - \hat{y}_1$$

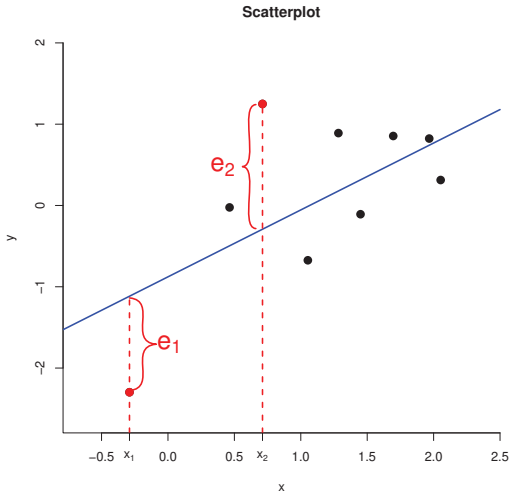
Least squares method



Which line fits the data best?

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2$$
$$y_2 = \hat{y}_2 + e_2$$
$$e_2 = y_2 - \hat{y}_2$$

Least squares method



Find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_i e_i^2,$$

where

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Estimation of the Regression Parameters

One needs to minimize

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=n}^{n} e_i^2 = \sum_{i=n}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Taking partial derivates of $Q(\hat{\beta}_0, \hat{\beta}_1)$ w.r.t. $\hat{\beta}_0$ and $\hat{\beta}_1$ and setting the resulting expressions equal to zero leads to the so-called *Normal Equations*

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \\ \sum X_i Y_i &= \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \end{aligned}$$

Estimation of the Regression Parameters

A little algebra yields the ordinary least squares estimators for the parameters (OLS)

$$\hat{\beta}_1 = b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{SS_{XY}}{SS_{XX}}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

where

$$SS_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \text{ and } SS_{XX} = \sum (X_i - \bar{X})^2$$

Estimation of the Regression Parameters

A little algebra yields the ordinary least squares estimators for the parameters (OLS)

Estimated model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = b_1 = r_{xy} \frac{S_y}{S_x}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

where

$$S_x^2 = \frac{1}{n-1} \sum \left(X_i - \bar{X} \right)^2 \text{ and } S_y^2 = \frac{1}{n-1} \sum \left(Y_i - \bar{Y} \right)^2$$

Fitted Regression Line

With $\hat{\beta}_0$ and $\hat{\beta}_1$ one can compute the fitted model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The estimated model should be *close* to the true linear regression model. Thus, one can think of $\hat{Y}_i$ as the estimated mean response at $X = X_i$

What about σ^2?

Estimation of σ^2

The minimum value of $Q(\hat{\beta}_0, \hat{\beta}_1)$ is denoted as *SSE*

- It is the sum of squares deviations between the observations and the fitted line.
- It is a measure of how well the fitted line fits the data.

$$\begin{aligned} SSE =& Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \end{aligned}$$

where e_i is called the *residual* for observation i.

Estimation of σ^2

Note

1. The residual $e_i = Y_i - \hat{Y}_i$ is the difference between observed and predicted values at X_i.
2. We can think of e_i as an estimator of the error ε_i.
3. The residuals are a fundamental tool to check the adequacy of the model.

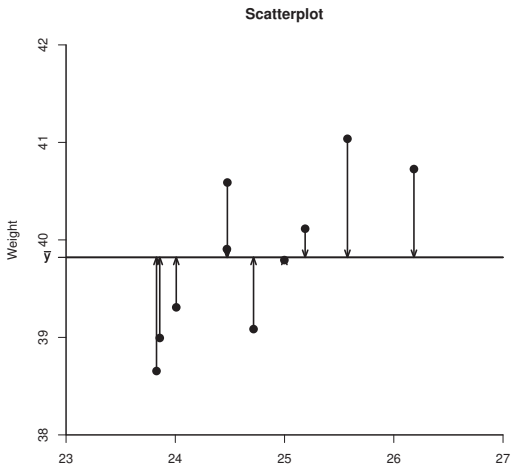
Estimation of σ^2

- Recall that σ^2 is the common variance for $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.
- Because $e_1, e_2, \ldots, e_n$ estimate the ε's, *SSE* should provide some information about the true variance σ^2.
- In fact,

$$s^2 = MSE = \frac{SSE}{n-2}$$

is an *unbiased* estimator of σ^2.

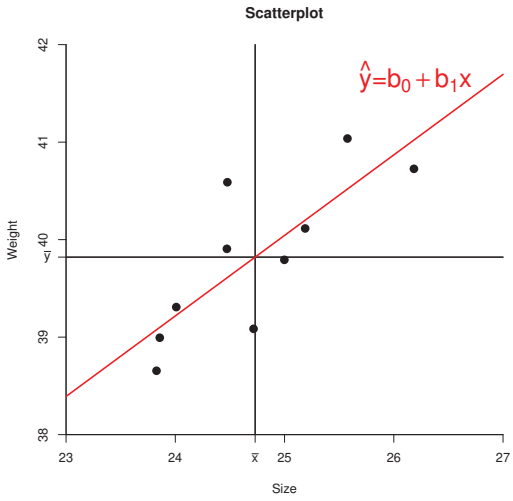
Sources of variation



Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

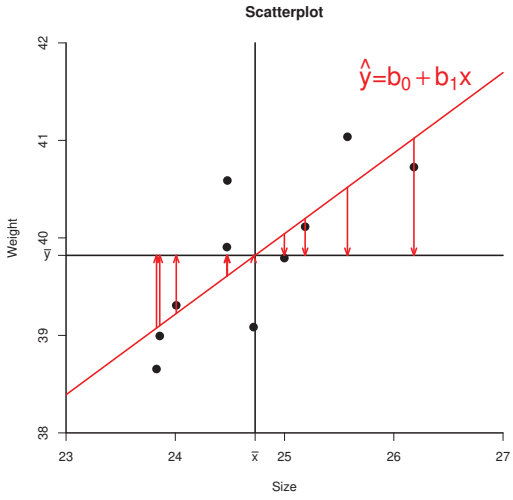
Sources of variation



Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

Sources of variation

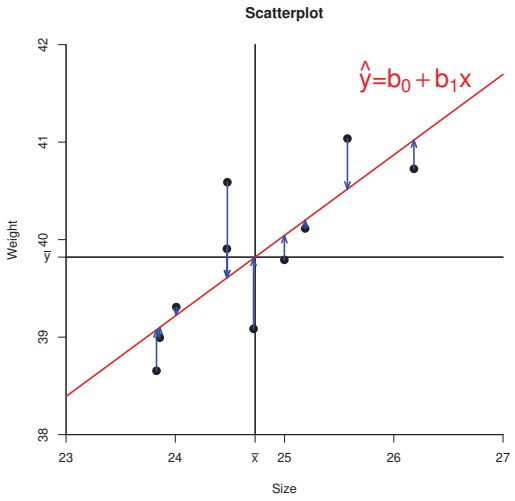


Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

Sources of variation



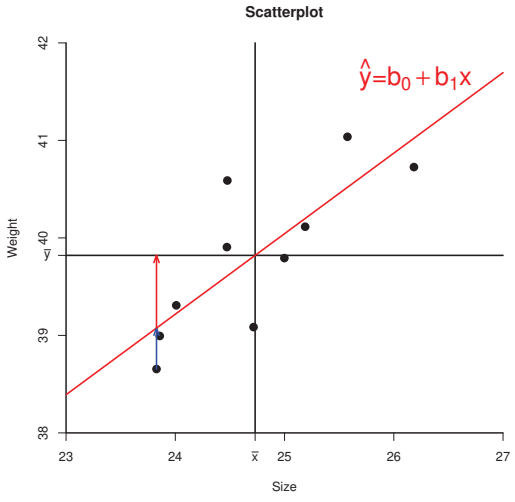
Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

Sources of variation



Variation in Y

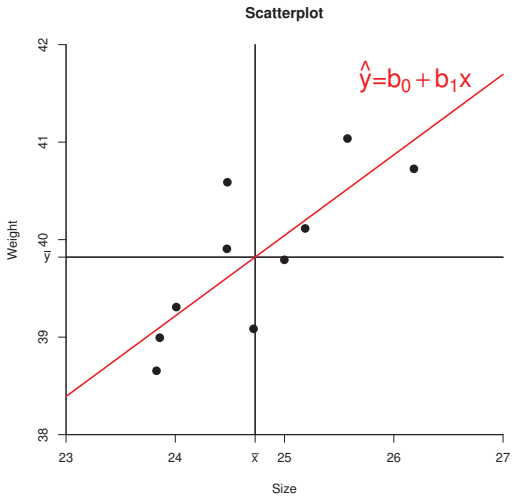
$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSTO = SSR + SSE$$

Sources of variation



Variation in Y

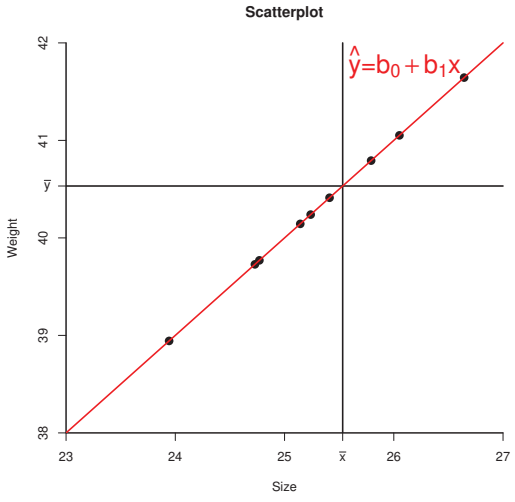
$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSTO = SSR + SSE$$

Sources of variation



Variation in Y

$$SSTO = \sum_i (y_i - \bar{y})^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSE = 0$$

$$SSTO = SSR$$

The sum of the squares

$$SSTO = SSR + SSE$$

$SSTO$: Total variation in the response Y

SSE: The variation in Y not explained by the model

SSR: The variation in Y explained by the model

The sum of the squares

We can decompose the total sum of squares in two different sums of squares: the residual and regression sum of squares.

Coefficient of determination

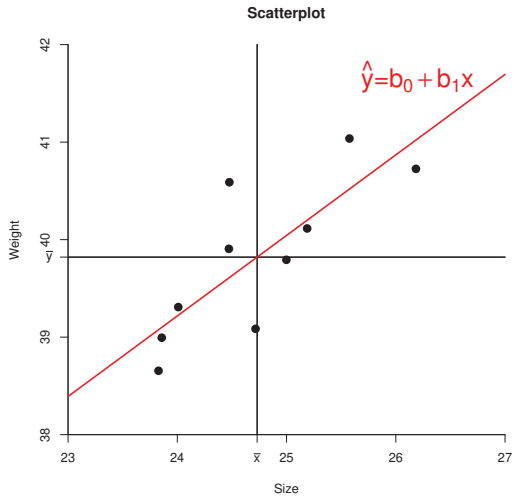
$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

Coefficient of determination

The coefficient of determination is a measure of the proportion of the total variation in the response that can be explained by the linear regression model.

- The coefficient of determination is always between 0 en 1

Sources of variation

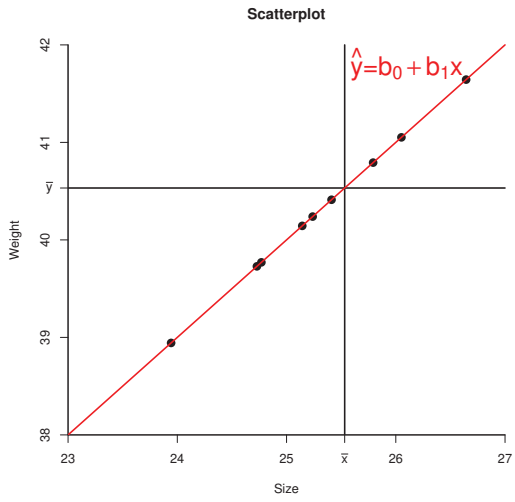


Variation van Y

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$
- The larger the better

Sources of variation



Variation van Y

$$R^2 = \frac{SSR}{SSTO} = 1$$

Linear regression: R code and output

```
> ## Fitting the model
>
> res<-lm(height~age, data=kalama)
> kalama.anova<-anova(res)
> kalama.summary<-summary(res)
> kalama.anova
>
Analysis of Variance Table

Response: height
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 57.655  57.655  879.99 4.428e-11 ***
Residuals 10  0.655   0.066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```


Linear regression: R code and output

```
> kalama.summary
>
Call:
lm(formula = height ~ age, data = kalama)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27238 -0.24248 -0.02762  0.16014  0.47238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  64.9283     0.5084  127.71  < 2e-16 ***
age           0.6350     0.0214   29.66 4.43e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.256 on 10 degrees of freedom
Multiple R-squared:  0.9888,    Adjusted R-squared:  0.9876
F-statistic:   880 on 1 and 10 DF,  p-value: 4.428e-11
>
```

Kalama Study: R Output

Anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

Kalama Study: R Output

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

- $y = \beta_0 + \beta_1 \cdot x + \epsilon$
- $b_0 = \bar{y} - b_1 \bar{x} = 64.928$
- $b_1 = r_{xy} \frac{s_y}{s_x} = 0.635$
- What does this p-value give?

Inference

$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.9283	0.5084	127.71	0.0000
age	0.6350	0.0214	29.66	0.0000

Kalama Study

Anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

- $r_{xy}(x, y) = \dfrac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}} = 0.994$
- $R^2 = \dfrac{SSR}{SST} = \dfrac{57.65}{58.31} = 0.989$

$$r_{xy} = \sqrt{R^2}$$

Kalama Study

Anova

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	1	57.65	57.65	879.99	0.0000
Residuals	10	0.66	0.07		
Total	11	58.31			

- $\hat{\sigma}^2 = MSE = \dfrac{SS_{Error}}{12-2} = \dfrac{0.66}{10} = 0.07$
- $R^2 = \dfrac{SSR}{SST} = \dfrac{57.65}{58.31} = 0.989$
- A substancial proportion of the variation in the outcome, 98.9%, is explained by the linear regression model.

Multiple linear regression

A multiple regression model is used to explain a dependent variable Y in terms of one or more independent variables $\boldsymbol{X}^{'} = \left(1, X_1, \ldots, X_{p-1}\right)$.

If Y is a quantitative random variable and the elements in $\boldsymbol{X}$ can take both quantitative and qualitative values, then one can consider a *regression model*

$$Y = f(\boldsymbol{X}) + \epsilon,$$

with $\boldsymbol{X}$ and ϵ independent and $E(\epsilon) = 0$, $\text{Var}(\epsilon) = \sigma^2$. Often, it is also assumed that ϵ is normally distributed.

The previous model essentially describes the average behavior of Y as a function $f(\cdot)$ of $\boldsymbol{X}$, i.e., $E(Y) = f(\boldsymbol{X})$.

The regression model

Taylor's theorem states that if f is differentiable at certain point $\boldsymbol{a} \in \mathbb{R}^{p-1}$ then

$$f(\boldsymbol{X}) = f(\boldsymbol{a}) + (\boldsymbol{X} - \boldsymbol{a})'\boldsymbol{\beta}_* + |\mathrm{X} - \mathrm{a}|h(\boldsymbol{X}), \qquad \lim_{\boldsymbol{X} \to \boldsymbol{a}} h(\boldsymbol{X}) = 0.$$

Therefore, at least locally (close to $\boldsymbol{a}$), $f(\cdot)$ can often be approximated by a *linear* model, i.e., $f(\boldsymbol{X}) = \boldsymbol{X}'\boldsymbol{\beta} = \beta_0 + \sum \beta_j X_j$.

$$\begin{aligned} Y &\approx \boldsymbol{X}'\boldsymbol{\beta} + \epsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon \end{aligned}$$

The previous regression model is linear in the parameters and, hence, it is called a linear regression model.

Non-linear Regression Model: $Y = \beta_0 + \beta_1 X_1^{\beta_2} + \varepsilon$

General linear regression model

Model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon$$

where

- $X_1, \ldots, X_{p-1}$ are known predictor variables.
- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ are unknown parameters.
- ϵ is an error term. It is often assumed that $\epsilon \sim N(0, \sigma^2)$.

Interpretation of the parameters

$$E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

- This response function is a hyperplane, which is a plane in more than two dimensions.
- The parameter β_k indicates the change in the mean response $E(Y|\boldsymbol{X})$ with a unit increase in the predictor variable X_k, when all other predictor variables in the regression model are held constant.
- $E(Y|\boldsymbol{X} = 0) = \beta_0$. The intercept gives the average response when all covariates are zero. It may not be interpretable unless the covariates are centered.

Categorical covariates: Dummy variables

Example

- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Main effects model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2$

Categorical covariates: Dummy variables

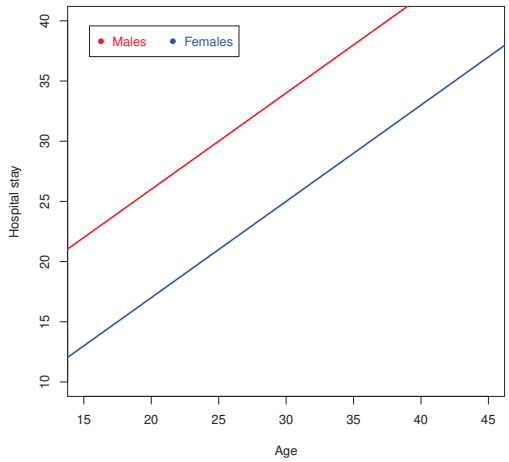
Example

- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Main effects model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$

Main effects model: Parallel lines

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = (\beta_0 + \beta_2) + \beta_1 X_1$



Categorical covariates: Dummy variables

Example

- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1$

Categorical covariates: Dummy variables

Example

- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$

Categorical covariates: Dummy variables

Example

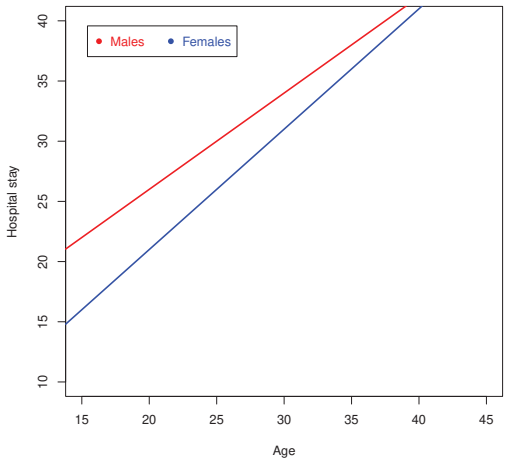
- Y length in hospital stay
- X_1: patient's age
- X_2: gender coded as female (1) - male (0)
- Interaction model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$

- It is still a linear model: $X_3 = X_1 X_2$

Interaction model: Non-parallel lines

Males	Females
$E(Y) = \beta_0 + \beta_1 X_1$	$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1$



Categorical covariates: Dummy variables

Example

- Y length in hospital stay
- X_1: patient's age
- X_2: female (1) - male (0)
- X disability status: 3 levels
  1. Not disabled
  2. Partially disabled
  3. Fully disabled

Categorical covariates: Dummy variables

For a factor with $r = 3$ levels, one needs to consider $(r - 1) = 2$ indicator (dummy) variables as predictors:

$$x_3 = \begin{cases} 1 & \text{Not disabled} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Partially disabled} \\ 0 & \text{otherwise} \end{cases}$$

Main effects model

$$Y = \beta_0 + \beta_1 X_1 + \overbrace{\beta_2 X_2}^{\text{gender}} + \underbrace{\beta_3 X_3 + \beta_4 X_4}_{\text{disability status}} + \varepsilon$$

Categorical covariates: Dummy variables

For a factor with $r = 3$ levels, one needs to consider $(r - 1) = 2$ indicator (dummy) variables as predictors:

$$x_3 = \begin{cases} 1 & \text{Not disabled} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{Partially disabled} \\ 0 & \text{otherwise} \end{cases}$$

Interaction model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_3 + \beta_4 X_4}_{\text{disability status}} + \underbrace{\beta_5 X_1 X_3 + \beta_6 X_1 X_4}_{\text{interaction: disability-age}} + \varepsilon$$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$, with $X_3 = X_1^2$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i}$$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Transformed variables:

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Great flexibility

- Polynomial regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$
- Transformed variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Transformed variables:

$$Y = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

Matrix formulation

Let us consider the following multiple regression model for the ith subject in the study, with $i = 1, 2, \ldots, n$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{p-1\,i} + \varepsilon_i$$

Collecting all the information on all subjects into vectors and matrices, the previous model can be written as

$$\underbrace{\begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{p-1\,1} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 1 & x_{1n} & \ldots & x_{p-1\,n} \end{pmatrix}}_{\boldsymbol{X}} \cdot \underbrace{\begin{pmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_{p-1} \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Matrix formulation

General linear regression model

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}} \cdot \underset{(p\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

- $\boldsymbol{Y}$ is the response vector.
- β is a vector of parameters.
- $\boldsymbol{X}$ is a matrix of known covariates with no measurement error.
- ε is a vector of errors with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2 \cdot \boldsymbol{I}$, where $\boldsymbol{I} = diag(1)$ is the so-called identity matrix.
- ε and $\boldsymbol{X}$ are assumed to be independent of each other.

Interpreting the the model: $\boldsymbol{Y}$

- $\boldsymbol{Y}$ is called the regressand, response variable, criterion variable, or dependent variable.
- The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables.
- Alternatively, there may be an operational reason, in which case there need be no presumption of causality.

Interpreting the the model: X

- X: Its elements are called regressors, explanatory variables, covariates, input variables, predictor variables, or independent variables.
- Usually a constant is included as one of the regressors. The corresponding element of β is called the intercept.
- Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.

Interpreting the the model: X

- Sometimes one of the regressors can be a non-linear function of another regressor, as in polynomial regression. The model remains linear as long as it is linear in the parameter vector β.
- The regressors may be viewed either as random variables, which one simply observes, or they can be considered as predetermined fixed values which one can choose.
- Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.

Interpreting the the model: β

- β is a p-dimensional parameter vector. Its elements are called effects, or regression coefficients.
- Statistical estimation and inference in linear regression focuses on β.
- The elements of this parameter vector are interpreted as the partial derivatives of the dependent variable with respect to the various independent variables.

Interpreting the the model: ε

- ε is called the error term, disturbance term, or noise.
- This variable captures all other factors which influence the dependent variable $\boldsymbol{Y}$ other than the regressors $\boldsymbol{X}$.
- The relationship between the error term and the regressors, for example whether they are correlated, is a crucial step in formulating a linear regression model, as it will determine the method to use for estimation.
- Typically, one assumes that $\boldsymbol{X}$ and ε are independent.

Estimating the model

Like before, the parameters can be estimated based on the ordinary least squares criterion (OLS)

$$Q = \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$$
$$= \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \cdots - \hat{\beta}_p X_{p-1\,i})^2$$

i.e., finding the values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}$ that minimize Q.

The solution to this optimization problem is given by the solution $\hat{\boldsymbol{\beta}}$ of the system of normal equations

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y} \qquad \Rightarrow \qquad \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Fitted values and residuals

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_{p-1} X_{p-1\,i}$
- Residuals $e_i = Y_i - \hat{Y}_i$
  - $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
  - $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
  - $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$
  - $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$ with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ (hat matrix)
  - $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$
  - $\text{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$

The sum of the squares

$$SST = \sum_i \left(Y_i - \bar{Y}\right)^2, \; SSR = \sum_i \left(\hat{Y}_i - \bar{Y}\right)^2, \; SSE = \sum_i \left(Y_i - \hat{Y}_i\right)^2$$

$$SST = SSR + SSE$$

SST: Total variation in the response

SSR: The variation explained by the model (covariates)

SSE: The variation not explained by the model (covariates)

- Coefficient of determination: $R^2 = \dfrac{SSR}{SST}$, interpretation idem
- $\hat{\sigma}^2 = MSE = \dfrac{SSE}{n-p}$

Inferences

$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \quad \mathrm{H}_A :$ not all β_k equal zero

Anova Table

Source of variation	SS	df	MS
Regression	SSR	$p-1$	$MSR = \dfrac{SSR}{p-1}$
Error	SSE	$n-p$	$MSE = \dfrac{SSE}{n-p}$
Total	$SSTO$	$n-1$	

- Under the null $F = \dfrac{MSR}{MSE} \sim F(p-1, n-p)$

Inferences

$$\mathrm{H}_0 : E(Y|\boldsymbol{X}) = \beta_0 \quad \mathrm{H}_A : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

Anova Table

Source of variation	SS	df	MS
Regression	SSR	$p-1$	$MSR = \dfrac{SSR}{p-1}$
Error	SSE	$n-p$	$MSE = \dfrac{SSE}{n-p}$
Total	$SSTO$	$n-1$	

- Under the null $F = \dfrac{MSR}{MSE} \sim F(p-1, n-p)$

Inferences: β_k

$$\mathrm{H}_0 : \beta_k = 0 \quad \mathrm{H}_A : \beta_k \neq 0$$

- Test statistics:

$$t = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \sim t(1-\alpha/2; n-p)$$

- Confidence interval:

$$\hat{\beta}_k \pm t(1-\alpha/2; n-p) s\{\hat{\beta}_k\}$$

Comparing nested models

Likelihood ratio tests

- Null hypothesis of interest equals $H_0 : \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta},0}$, for some subspace $\Theta_{\boldsymbol{\beta},0}$ of the parameter space $\Theta_{\boldsymbol{\beta}}$

Comparing nested models

Likelihood ratio tests

- Null hypothesis of interest equals $H_0 : \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta},0}$, for some subspace $\Theta_{\boldsymbol{\beta},0}$ of the parameter space $\Theta_{\boldsymbol{\beta}}$

- For instance,

  $\mathrm{H}_0 : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 \quad \mathrm{H}_A : E(Y|\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$

Comparing nested models

Likelihood ratio tests

- Null hypothesis of interest equals $H_0 : \boldsymbol{\beta} \in \Theta_{\boldsymbol{\beta},0}$, for some subspace $\Theta_{\boldsymbol{\beta},0}$ of the parameter space $\Theta_{\boldsymbol{\beta}}$
- For instance,

$$\mathrm{H}_0 : \beta_3 = \beta_4 = 0 \quad \mathrm{H}_A : \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0$$

- Notation:
  - L_{ML}: ML likelihood function
  - $\widehat{\boldsymbol{\beta}}_{ML,0}$: MLE under H_0
  - $\widehat{\boldsymbol{\beta}}_{ML}$: MLE under general model

Likelihood ratio tests

- Test statistic:

$$-2 \ln \lambda_N = -2 \ln \left[\frac{L_{ML}(\widehat{\boldsymbol{\beta}}_{ML,0})}{L_{ML}(\widehat{\boldsymbol{\beta}}_{ML})} \right]$$

- Asymptotic null distribution: χ^2 with d.f. equal to the difference in dimension of $\Theta_{\boldsymbol{\beta}}$ and $\Theta_{\boldsymbol{\beta},0}$.
- An equivalent F-test can also be used.

Patient satisfaction

Case study

A hospital administrator wanted to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index).

The administrator randomly selected 46 patients and collected data on the previous variables. Larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

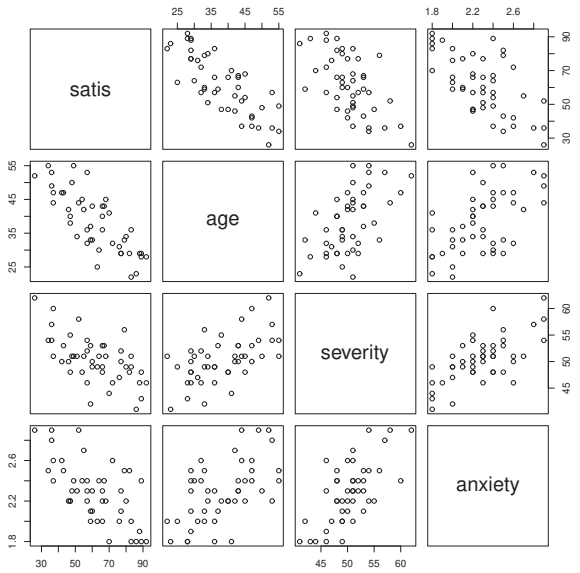
R code: Patient satisfaction

```
> ## Reading the data
>
> satisfaction=read.table("satisfaction.txt", header=T)
> head(satisfaction,10)
>
   satis age severity anxiety
1     48  50       51     2.3
2     57  36       46     2.3
3     66  40       48     2.2
4     70  41       44     1.8
5     89  28       43     1.8
6     36  49       54     2.9
7     46  42       50     2.2
8     54  45       48     2.4
9     26  52       62     2.9
10    77  29       50     2.1
>
```

R code: Patient satisfaction

```
> ## Exploring the data
>
> cor(satisfaction)
>
              satis        age   severity    anxiety
satis     1.0000000 -0.7867555 -0.6029417 -0.6445910
age      -0.7867555  1.0000000  0.5679505  0.5696775
severity -0.6029417  0.5679505  1.0000000  0.6705287
anxiety  -0.6445910  0.5696775  0.6705287  1.0000000
>
> options(digits=2)
> descrip.satisfaction<-stat.desc(satisfaction,basic=TRUE, desc=TRUE)
> descrip.satisfaction
>
               satis     age severity anxiety
nbr.val        46.00   46.00  4.6e+01  46.000
min            26.00   22.00  4.1e+01   1.800
max            92.00   55.00  6.2e+01   2.900
range          66.00   33.00  2.1e+01   1.100
median         60.00   37.50  5.0e+01   2.300
mean           61.57   38.39  5.0e+01   2.287
SE.mean         2.54    1.31  6.4e-01   0.044
var           297.10   79.53  1.9e+01   0.090
std.dev        17.24    8.92  4.3e+00   0.299
coef.var        0.28    0.23  8.6e-02   0.131
>
> plot(satisfaction)
>
```


R code: Patient satisfaction



R code: Patient satisfaction

```
> ## Fitting the model
>
>
> satisfaction.lm<-lm(satis~age+severity+anxiety, data=satisfaction)
> satisfaction.summary<-summary(satisfaction.lm)
> satisfaction.summary

Call:
lm(formula = satis ~ age + severity + anxiety, data = satisfaction)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491     18.126    8.74  5.3e-11 ***
age           -1.142      0.215   -5.31  3.8e-06 ***
severity      -0.442      0.492   -0.90    0.374
anxiety      -13.470      7.100   -1.90    0.065 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10 on 42 degrees of freedom
Multiple R-squared:  0.682,     Adjusted R-squared:  0.659
F-statistic: 30.1 on 3 and 42 DF,  p-value: 1.54e-10
>
```


R code: Patient satisfaction

```
> ## Likelihood ratio test null model versus full model
>
>
> satisfaction.lm.int<-lm(satis~1, data=satisfaction) # Null model
> anova(satisfaction.lm.int,satisfaction.lm)          # Null versus full
>
Analysis of Variance Table

Model 1: satis ~ 1
Model 2: satis ~ age + severity + anxiety
  Res.Df   RSS Df Sum of Sq    F  Pr(>F)
1     45 13369
2     42  4249  3      9120 30.1 1.5e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

R code: Patient satisfaction

```
> ## Likelihood ratio test null model versus full model
>
> satis = β0 + ε ⇔ satis = β0 + β1 age + β2 severity + β3 anxiety + ε
> satisfaction.lm.int<-lm(satis~1, data=satisfaction) # Null model
> anova(satisfaction.lm.int,satisfaction.lm)          # Null versus full
>
Analysis of Variance Table

Model 1: satis ~ 1
Model 2: satis ~ age + severity + anxiety
  Res.Df   RSS Df Sum of Sq    F  Pr(>F)
1     45 13369
2     42  4249  3      9120 30.1 1.5e-10 ***   H0 : β1 = β2 = β3 = 0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> ## Previous analysis with summary function

Multiple R-squared:  0.682,     Adjusted R-squared:  0.659
F-statistic: 30.1 on 3 and 42 DF,  p-value: 1.54e-10  H0 : β1 = β2 = β3 = 0
>
```


R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1   8275    8275   81.80 2.1e-11 ***
severity   1    481     481    4.75   0.035 *
anxiety    1    364     364    3.60   0.065 .
Residuals 42   4249     101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1   8275    8275   81.80 2.1e-11 ***
severity   1    481     481    4.75   0.035 *
anxiety    1    364     364    3.60   0.065 .
Residuals 42   4249     101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ## Previous analysis with summary function
>
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491     18.126    8.74  5.3e-11 ***
age           -1.142      0.215   -5.31  3.8e-06 ***
severity      -0.442      0.492   -0.90    0.374
anxiety      -13.470      7.100   -1.90    0.065 .
>
```


R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.lm2<-lm(satis~age+anxiety+severity, data=satisfaction)
> satisfaction.anova2<-anova(satisfaction.lm2)
> satisfaction.anova2
>
Analysis of Variance Table

Response: satis
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1   8275    8275   81.80 2.1e-11 ***
anxiety    1    763     763    7.55  0.0088 **
severity   1     82      82    0.81  0.3741
Residuals 42   4249     101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> ## Previous analysis with summary function
>
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  158.491     18.126    8.74  5.3e-11 ***
age           -1.142      0.215   -5.31  3.8e-06 ***
severity      -0.442      0.492   -0.90    0.374
anxiety      -13.470      7.100   -1.90    0.065 .
>
```

R code: Patient satisfaction

```
> ## Sequential building of the model
>
> satisfaction.anova<-anova(satisfaction.lm)
> satisfaction.anova
>
Analysis of Variance Table

Response: satis
          Df Sum Sq Mean Sq F value  Pr(>F)
age        1   8275    8275   81.80 2.1e-11 ***
severity   1    481     481    4.75   0.035 *
anxiety    1    364     364    3.60   0.065 .
Residuals 42   4249     101
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```


Final model

Final model:

$$Y_i = 145.941 - 1.2X_{1i} - 16.742X_{3i} + \varepsilon_i$$

```
> ## Final model
>
> satisfaction.lm.final<-lm(satis~age+anxiety, data=satisfaction)
> satisfaction.final.summary<-summary(satisfaction.lm.final)
> satisfaction.final.summary
>
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.941     11.525   12.66  4.2e-16 ***
age           -1.200      0.204   -5.88  5.4e-07 ***
anxiety      -16.742      6.081   -2.75   0.0086 **
---
>
```

Inference for mean response

- In many applications one wants to estimate and/or make inferences about the mean of the response Y for a given value of the predictors $\boldsymbol{X}_h$

$$Y_h = E(Y|\boldsymbol{X}_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$$

- Point estimate: $\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$.
- Confidence intervals are used for inferences.

Inference for mean response

- $E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$
- $\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$
- $s^2\left(\hat{Y}_h\right) = \boldsymbol{X}_h'\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_h$
- $s^2(\hat{Y}_h) = MSE(\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h)$
- Hence $1-\alpha$ confidence limits are

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)\cdot s(\hat{Y}_h)$$

Prediction of a new observation

- Let now assume that one wants to predict a new observation $Y_{h,new}$ corresponding to a given level of the predictor variable $\boldsymbol{X} = \boldsymbol{X}_h$.
- This is typically done by constructing a $1 - \alpha$ prediction interval, i.e., by finding values $Y_{h,low}$ and $Y_{h,up}$ so that

$$P(Y_{h,low} \leq Y_{h,new} \leq Y_{h,up} | \boldsymbol{X} = \boldsymbol{X}_h) = 1 - \alpha$$

- The prediction interval is given by

$$\hat{Y}_h \pm t(1 - \alpha/2; n - p) s\{pred\}$$

where $s^2\{pred\} = MSE + s^2(\hat{Y}_h)$

R code: Predicting a new observation

```
> ## Predicting a new observation
>
> newdata = data.frame(age=43, anxiety=2.7)
> pred.w.plim <- predict(satisfaction.lm.final, newdata, interval="predict")
> pred.w.clim <- predict(satisfaction.lm.final, newdata, interval = "confidence")
> pred.w.plim
>
  fit lwr upr
1  49  28  70
>
> pred.w.clim
>
  fit lwr upr
1  49  44  54
>
```

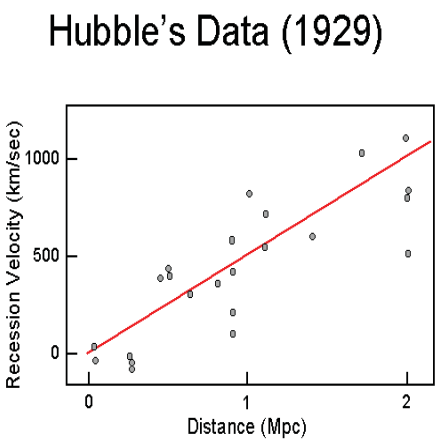
Logistic Regression

Ariel Alonso Abad

Catholic University of Leuven

Linear regression

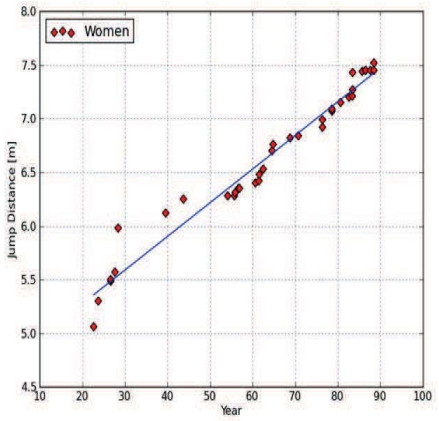
Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$



- Hubble's law

Linear regression

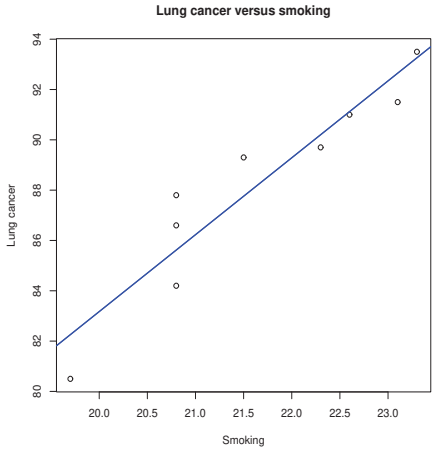
Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$



- Hubble's law
- Sport

Linear regression

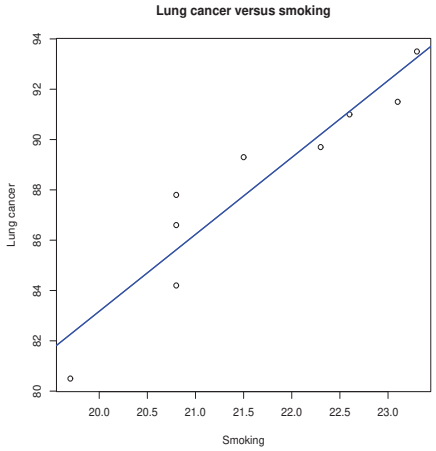
Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$



- Hubble's law
- Sport
- Smoking and lung cancer

Linear regression

Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

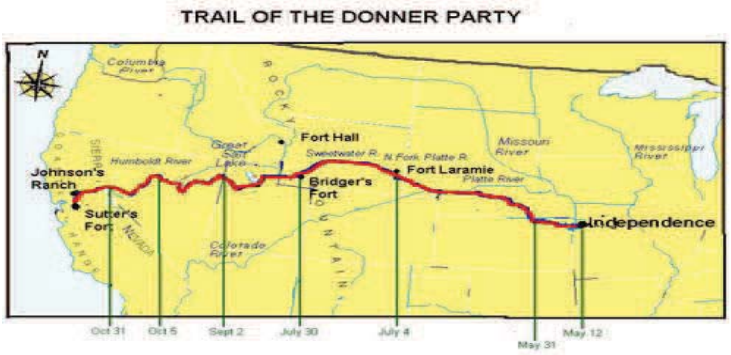


- Hubble's law
- Sport
- Smoking and lung cancer
- Linear regression versus Logistic regression

Legends of America: Donner party data

Donner party expedition

In 1846, the Donner and Reed families left Illinois for California, a 2500 mile journey that would become one of the greatest tragedies in USA history. Stranded in Sierra Nevada by a series of snowstorms, they were rescued in April of the following year. 40 members died, some (or perhaps all) of those that survived did so by resorting to cannibalism.



Donner party data

- 88 persons
- Variables: survival, gender and age
- Taking into account age, are the chances of survival larger for women than for men?

Donner party data

Name	Sex	Age	Survivg
Antoine	Male	23	No
Breen, Mary	Female	40	Yes
Breen, Patrick	Male	40	Yes
Burger, Charles	Male	30	No
Denton, John	Male	28	No
Dolan, Patrick	Male	40	No
Donner, Elizabeth	Female	45	No
Donner, George	Male	62	No
Donner, Jacob	Male	65	No
Donner, Tamsen	Female	45	No
Eddy, Eleanor	Female	25	No
Eddy, William	Male	28	Yes
Elliot, Milton	Male	28	No
Fosdick, Jay	Male	23	No
Fosdick, Sarah	Female	22	Yes
Foster, Sarah	Female	23	Yes
Foster, William	Male	28	Yes
Graves, Eleanor	Female	15	Yes
Graves, Elizabeth	Female	47	No
Graves, Franklin	Male	57	No
Graves, Mary	Female	20	Yes
Graves, William	Male	18	Yes
Halloran, Luke	Male	25	No
Hardkoop, Mr.	Male	60	No
Herron, William	Male	25	Yes
Noah, James	Male	20	Yes
Keseberg, Lewis	Male	32	Yes
Keseberg, Phillipine	Female	32	Yes
McCutcheon, Amanda	Female	24	Yes
McCutcheon, William	Male	30	Yes
Murphy, John	Male	15	No
Murphy, Lavina	Female	50	No
Pike, Harriet	Female	21	Yes
Pike, William	Male	25	No
Reed, James	Male	46	Yes
Reed, Margaret	Female	32	Yes
Reinhardt, Joseph	Male	30	No
Shoemaker, Samuel	Male	25	No
Smith, James	Male	25	No
Snyder, John	Male	25	No
Spitzer, Augustus	Male	30	No
Stanton, Charles	Male	35	No
Trubode, J.B.	Male	23	Yes
Williams, Baylis	Male	24	No
Williams, Eliza	Female	25	Yes

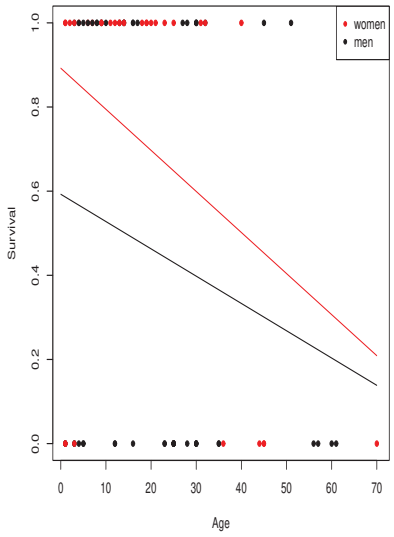
- Dependent variable is binary:

$$Y = \begin{cases} 1 \text{ Survived,} \\ 0 \text{ Died.} \end{cases}$$

- Independent predictors:

$$age, fem = \begin{cases} 1 \text{ for women,} \\ 0 \text{ for men.} \end{cases}$$

Exploring the data



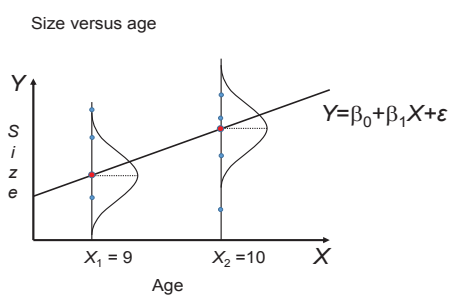
- Survived = 1, Died = 0
- Graph is not as informative as in linear regression

Linear regression

- Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- The expected value (average) of Y is modeled as a linear function of the predictors

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Simple linear regression



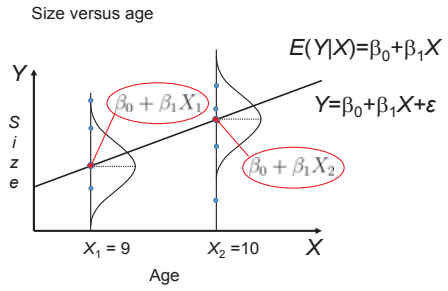
- $Y = \beta_0 + \beta_1 X + \varepsilon$

Linear regression

- Basic model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- The expected value (average) of Y is modeled as a linear function of the predictors

$$E(Y|X_1, X_2, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Simple linear regression



- $Y = \beta_0 + \beta_1 X + \varepsilon$
- $E(Y|X) = \beta_0 + \beta_1 X$

Binary outcome: Problem

The expected value (average) of a binary variable is a **probability**

$$E(Y|X_1, X_2, \ldots, X_p) = P(Y = 1|\boldsymbol{X})$$

where $P(Y = 1|\boldsymbol{X})$ gives the probability as a function of the covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$

$$\boxed{P(Y = 1|\boldsymbol{X})} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$0 \leq P(Y = 1|\boldsymbol{X}) \leq 1$$

but $\eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ is not always between 0 and 1.

Binary outcome: Problem

The expected value (average) of a binary variable is a **probability**

$$E(Y|X_1, X_2, \ldots, X_p) = P(Y = 1|\boldsymbol{X})$$

where $P(Y = 1|\boldsymbol{X})$ gives the probability as a function of the covariates $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$

$$P(Y = 1|age, fem) = \beta_0 + \beta_1 age + \beta_2 fem$$

$$0 \leq P(Y = 1|age, fem) \leq 1$$

but $\eta(age, fem) = \beta_0 + \beta_1 age + \beta_2 fem$ is not always between 0 and 1.

Binary outcome: Problem

Problem with the linear model

$$P(Y = 1|fem) = \beta_0 + \beta_1 fem$$

Assume that $\beta_0 = 0.5$ en $\beta_1 = -1$. What is the survival probability for a woman?

$$P(Y = 1|fem) = 0.5 - fem$$

For women: $fem = 1$ thus

$$\begin{aligned} P(Y = 1|fem = 1) &= 0.5 - fem \\ &= 0.5 - 1 = -0.5 \end{aligned}$$

A probability should always be between 0 and 1!

Binary outcome: Solution

- Transform $P(Y = 1|\boldsymbol{X})$:

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \ln\frac{P(Y = 1|\boldsymbol{X})}{P(Y = 0|\boldsymbol{X})} = \ln\left(\frac{P(Y = 1|\boldsymbol{X})}{1 - P(Y = 1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\left[P(Y = 1|\boldsymbol{X})\right] \leq \infty$
- Model:

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Transform back:

$$P(Y = 1|\boldsymbol{X}) = \frac{e^{\eta(\boldsymbol{X})}}{1 + e^{\eta(\boldsymbol{X})}}$$

Binary outcome: Solution

- Transform $P(Y = 1|\boldsymbol{X})$:

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \ln\frac{P(Y = 1|\boldsymbol{X})}{P(Y = 0|\boldsymbol{X})} = \ln\left(\frac{P(Y = 1|\boldsymbol{X})}{1 - P(Y = 1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\left[P(Y = 1|\boldsymbol{X})\right] \leq \infty$
- Model:

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Transform back:

$$P(Y = 1|\boldsymbol{X}) = \frac{e^{\eta(\boldsymbol{X})}}{1 + e^{\eta(\boldsymbol{X})}} = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Now is $P(Y = 1|\boldsymbol{X})$ always between 0 en 1

Binary outcome: Solution

- Transform $P(Y = 1|\boldsymbol{X})$:

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \ln \frac{P(Y = 1|\boldsymbol{X})}{P(Y = 0|\boldsymbol{X})} = \ln\left(\frac{P(Y = 1|\boldsymbol{X})}{1 - P(Y = 1|\boldsymbol{X})}\right)$$

- $-\infty \leq \text{logit}\left[P(Y = 1|\boldsymbol{X})\right] \leq \infty$
- Model:

$$\text{logit}\left[P(Y = 1|age, fem)\right] = \eta(age, fem) = \beta_0 + \beta_1 age + \beta_2 fem$$

- Transform back:

$$P(Y = 1|age, fem) = \frac{e^{\eta(age,fem)}}{1 + e^{\eta(age,fem)}} = \frac{e^{\beta_0+\beta_1 age+\beta_2 fem}}{1 + e^{\beta_0+\beta_1 age+\beta_2 fem}}$$

- Now is $P(Y = 1|age, fem)$ always between 0 en 1

Binary outcome: Solution

Logistic Model

$$P(Y = 1|fem) = \frac{e^{\beta_0+\beta_1 fem}}{1 + e^{\beta_0+\beta_1 fem}}$$

Assume that $\beta_0 = 0.5$, $\beta_1 = -1$. What is the survival probability for a woman?

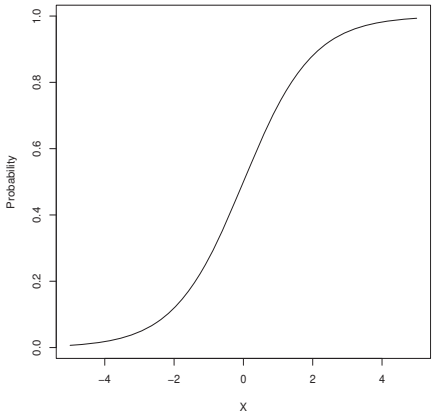
$$P(Y = 1|fem) = \frac{e^{0.5-fem}}{1 + e^{0.5-fem}}$$

For women: $fem = 1$ thus

$$P(Y = 1|fem = 1) = \frac{e^{0.5-1}}{1 + e^{0.5-1}} = \frac{e^{-0.5}}{1 + e^{-0.5}} = 0.378$$

Now the survival probability for a woman is between 0 en 1!

Model for the probability: X continuous



$$P(Y = 1|X) = \frac{e^X}{1 + e^X}$$

Logarithm (log)

Different notations: log

Estimating the parameters

$$\text{logit}\left[P(Y = 1|\boldsymbol{X})\right] = \eta(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Ordinary least squares (OLS) is not suitable
- Method of maximum likelihood (ML) is the adequate choice

Maximum Likelihood Estimation

Find the values of the parameters so that the likelihood is maximized. In other words, the ML estimates (MLE) are the values of the parameters that make the observed data most likely to have been observed. For linear models OLS and ML are equivalent.

Fitting the model in R

```
> ## Home
>
> setwd("C:\Users\Logistic-Regression"')
>
> ## Reading the data
>
> donner<-read.table("donner-class.txt", row.names = 1, header=TRUE)
> head(donner,10)
                       Age Outcome    Sex Family.name Status
Breen_Edward_           13       1   Male       Breen Family
Breen_Margaret_Isabella  1       1 Female       Breen Family
Breen_James_Frederick    5       1   Male       Breen Family
Breen_John              14       1   Male       Breen Family
Breen_Margaret_Bulger   40       1 Female       Breen Family
Breen_Patrick           51       1   Male       Breen Family
Breen_Patrick_Jr.        9       1   Male       Breen Family
Breen_Peter              3       1   Male       Breen Family
Breen_Simon_Preston      8       1   Male       Breen Family
Donner_Elitha_Cumi      13       1 Female    G_Donner Family
>
```


Fitting the model in R

```
> ## Keeping only the variables of interest
>
> donner.na<-na.omit(subset(donner,select=c('Age','Outcome','Sex')))
> donner.na$fem = as.numeric(donner.na$Sex=="Female")
> head(donner.na,10)
>
                       Age Outcome    Sex fem
Breen_Edward_           13       1   Male   0
Breen_Margaret_Isabella  1       1 Female   1
Breen_James_Frederick    5       1   Male   0
Breen_John              14       1   Male   0
Breen_Margaret_Bulger   40       1 Female   1
Breen_Patrick           51       1   Male   0
Breen_Patrick_Jr.        9       1   Male   0
Breen_Peter              3       1   Male   0
Breen_Simon_Preston      8       1   Male   0
Donner_Elitha_Cumi      13       1 Female   1
>
```

Fitting the model in R

```
> ## Fitting a logistic regression
>
> donner.log<-glm(Outcome ~ Age + fem,data=donner.na,family=binomial(link="logit"))
> summary(donner.log)

Call:
glm(formula = Outcome ~ Age + fem, family = binomial(link = "logit"), data = donner.na)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8828  -1.0383   0.6511   1.0261   1.7386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.55382    0.41788   1.325   0.1851
Age         -0.03561    0.01525  -2.336   0.0195 *
fem          1.06798    0.48229   2.214   0.0268 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 108.87  on 85  degrees of freedom
AIC: 114.87

Number of Fisher Scoring iterations: 4
```


Donner party data

Model

$$\text{logit}\left[P(Y=1|age,fem)\right]=\beta_0+\beta_1 age+\beta_2 fem$$

where $fem=1$ for women $fem=0$ for men. Equivalently

$$P(Y=1|age,fem)=\frac{e^{\beta_0+\beta_1 age+\beta_2 fem}}{1+e^{\beta_0+\beta_1 age+\beta_2 fem}}$$

Estimated model

$$\text{logit}\left[\hat{P}(Y=1|age,fem)\right]=0.553-0.035age+1.067fem$$

$$\hat{P}(Y=1|age,fem)=\frac{e^{0.553-0.035age+1.067fem}}{1+e^{0.553-0.035age+1.067fem}}$$

Interpretation of the coefficients

Linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Association between (y, x): $r_{xy} = \dfrac{\text{cov}(x, y)}{\sigma_y \sigma_x}$
  - Range: $-1 \leq r_{xy} \leq 1$
  - Positive correlation between x en y: $r_{xy} > 0$ ($\beta_1 > 0$)
  - No correlation between x en y: $r_{xy} = 0$ ($\beta_1 = 0$)
  - Negative correlation between x en y: $r_{xy} < 0$ ($\beta_1 < 0$)
- $\beta_1 = r_{xy} \dfrac{\sigma_y}{\sigma_x}$

Association in a 2×2 cross-table

Hypothetical example

		Predictor: Female $fem = 1$	Predictor: Male $fem = 0$
Criterium	Survived ($Y = 1$)	$P(Y = 1) = \frac{2}{3}$	$P(Y = 1) = \frac{1}{3}$
	Died ($Y = 0$)	$P(Y = 0) = \frac{1}{3}$	$P(Y = 0) = \frac{2}{3}$

Association in a 2×2 cross-table

Odds

- Odds of surviving for women:

$$\Theta_{survival|women} = \frac{P(Y=1|fem=1)}{P(Y=0|fem=1)} = \frac{2/3}{1/3} = 2$$

  $\Rightarrow$ for every 2 women that survive 1 dies

- Odds of surviving for men:

$$\Theta_{survival|men} = \frac{P(Y=1|fem=0)}{P(Y=0|fem=0)} = \frac{1/3}{2/3} = \frac{1}{2} = 0.5$$

  $\Rightarrow$ for every man that survives 2 die

(term comes from horse racing)

Association in a 2×2 cross-table

Odds Ratio

Ratio of odds or odds ratio = is a measure of association in 2×2 cross-tables

$$\text{Odds Ratio: } OR = \frac{\Theta_{survival|women}}{\Theta_{survival|men}} = \frac{2}{0.5} = 4$$

- Interpretation: the odds of survival for women are 4 times larger than the odds of survival for men

(if the odds for men are 0.5 to 1 then for women they are 2 to 1)

Association in a 2×2 cross-table

		Predictor: $X = 1$ Female	Predictor: $X = 0$ Male
Criterium	$Y = 1$	$P(Y = 1) = \frac{2}{3}$	$P(Y = 1) = \frac{2}{3}$
	$Y = 0$	$P(Y = 1) = \frac{1}{3}$	$P(Y = 1) = \frac{1}{3}$

$$\Theta_{survival|women} = \frac{P(Y = 1|fem = 1)}{P(Y = 0|fem = 1)} = \frac{2/3}{1/3} = 2$$

$$\Theta_{survival|men} = \frac{P(Y = 1|fem = 0)}{P(Y = 0|fem = 0)} = \frac{2/3}{1/3} = 2$$

$$OR = \frac{\Theta_{survival|women}}{\Theta_{survival|men}} = \frac{2}{2} = 1$$

Properties of odds ratios

Odds Ratio

- $0 < OR < \infty$
- $OR = 1 \Leftrightarrow$ independence
- $OR > 1 \Leftrightarrow$ positive association
- $OR < 1 \Leftrightarrow$ negative association

Interpretation of the coefficients

Dichotomous predictor ($X = 0$ of 1 like gender) probability model

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \begin{cases} \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} & X = 1, \text{female} \\ \frac{e^{\beta_0}}{1+e^{\beta_0}} & X = 0, \text{male} \end{cases}$$

		Predictor	
		$X = 1$ Female	$X = 0$ Male
Criterium	$Y = 1$ Survived	$P(Y=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(Y=1) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
	$Y = 0$ Died	$P(Y=0) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$P(Y=0) = \frac{1}{1+e^{\beta_0}}$

Computing the odds ratio

		Predictor	
		$X = 1$ Female	$X = 0$ Male
Criterium	$Y = 1$	$\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\frac{e^{\beta_0}}{1+e^{\beta_0}}$
	$Y = 0$	$\frac{1}{1+e^{\beta_0+\beta_1}}$	$\frac{1}{1+e^{\beta_0}}$

$$OR = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)}$$

Computing the odds ratio

Criterium	Predictor: $X = 1$ Female	Predictor: $X = 0$ Male
$Y = 1$	$\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\frac{e^{\beta_0}}{1+e^{\beta_0}}$
$Y = 0$	$\frac{1}{1+e^{\beta_0+\beta_1}}$	$\frac{1}{1+e^{\beta_0}}$

$$OR = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} = e^{\beta_1}$$

$$\begin{aligned} OR &= e^{\beta_1} \\ \ln(OR) &= \beta_1 \end{aligned}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Gender: The odds of survival for a woman are 3 times larger than the odds of survival for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$

Age=25

Age=25

$$\Theta_{survival|women} = 3 \cdot \Theta_{survival|men}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Gender: The odds of survival for a woman are 3 times larger than the odds of survival for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$

Age=50

Age=50

$$\Theta_{survival|women} = 3 \cdot \Theta_{survival|men}$$

Interpretation of the coefficients

Logistic regression model

$$\text{logit}\left[P(Y = 1|X)\right] = \beta_0 + \beta_1 X$$

- Association between (y, x): OR
  - Range: $0 < OR < \infty$
  - Positive association between x en y: $OR > 1$ $(\beta_1 > 0)$
  - No association between x en y: $OR = 1$ $(\beta_1 = 0)$
  - Negative association between x en y: $OR < 1$ $(\beta_1 < 0)$
- $\beta_1 = \ln(OR)$

Continuous predictor X

- The interpretation is analogous to the one given for dummy predictor
- For instance, consider two ages
  - A1: Age=X
  - A2: A year older, Age=$X+1$

$$\Theta_{survival|X+1} = \frac{P(Y=1|X+1)}{P(Y=0|X+1)} \qquad \Theta_{survival|X} = \frac{P(Y=1|X)}{P(Y=0|X)}$$

$$\Theta_{survival|X+1} = e^{\beta_1} \cdot \Theta_{survival|X}$$

Donner party data

- $\text{logit}(\hat{P}(Y=1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356} = 0.965 \approx 0.96$, an increase of one year in age is associated with a 4% decrease in the odds of survival

Age=26

Age=25

$$\Theta_{survival|26,men} = 0.96 \cdot \Theta_{survival|25,men}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356} = 0.965 \approx 0.96$, an increase of one year in age is associated with a 4% decrease in the odds of survival

Age=51

Age=50

$$\Theta_{survival|51,men} = 0.96 \cdot \Theta_{survival|50,men}$$

Continuous predictor X

- A one unit change in X is not always meaningful
- For instance, consider two ages
  - A1: Age=X
  - A2: c years older, Age=$X + c$

$$\Theta_{survival|X+c} = \frac{P(Y = 1|X + c)}{P(Y = 0|X + c)} \qquad \Theta_{survival|X} = \frac{P(Y = 1|X)}{P(Y = 0|X)}$$

$$\Theta_{survival|X+c} = e^{\beta_1 \cdot c} \cdot \Theta_{survival|X}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356 \cdot 10} = 0.70$, a 10 years increase in age is associated with a 30% decrease in the odds of survival

Age=30

Age=20

$$\Theta_{survival|30,women} = 0.70 \cdot \Theta_{survival|20,women}$$

Donner party data

- $\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$
- Age: $\hat{\beta}_1 = -0.0356 \Rightarrow \hat{OR} = e^{-0.0356 \cdot 10} = 0.70$, a 10 years increase in age is associated with a 30% decrease in the odds of survival

Age=60

Age=50

$$\Theta_{survival|60,women} = 0.70 \cdot \Theta_{survival|50,women}$$

More predictors

- The same approach as above: change in logit cause by increasing predictor X_j by 1 unit and keeping all the other predictors fixed is back transformed into a change in odds ratio
- Often called "adjusted odds ratio" ("adjusted" by other predictors)

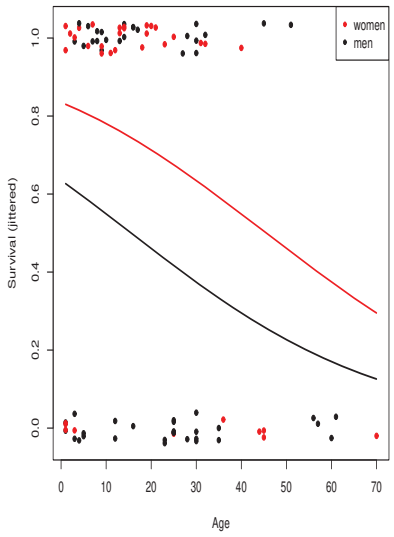
Donner party data

- Estimated model

$$\text{logit}(\hat{P}(Y = 1|age, fem)) = 0.553 - 0.035age + 1.067fem$$

- Gender: The survival odds for a woman are 3 times larger than the survival odds for a man of the same age: $\hat{OR} = e^{1.067} \approx 3$
- Age: A 10 year increase in age is associated with a 30% decrease in the survival odds for both men and women ($\hat{OR} = e^{-0.0356 \cdot 10} = 0.70$)

Donner party data



$$\hat{P}(Y = 1|age, fem) = \frac{e^{(0.55-0.04age+1.07fem)}}{1 + e^{(0.55-0.04age+1.07fem)}}$$

Donner party data: Model fit

- Model fit

Parameter	Estimate	Std. Error	z value	p-value
(Intercept)	0.553	0.417	1.325	0.1850
Age	-0.035	0.015	-2.336	0.0195
fem	1.067	0.482	2.214	0.0268

- 95% confidence interval for effect: Age, β_1

$$\left[\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right]$$

- 95% BI for β_1: $-0.035 \pm 1.96 \cdot 0.015 \Rightarrow (-0.067, -0.006)$

Donner party data: Model fit

- Model fit

Parameter	Estimate	Std. Error	z value	p-value
(Intercept)	0.553	0.417	1.325	0.1850
Age	-0.035	0.015	-2.336	0.0195
fem	1.067	0.482	2.214	0.0268

- 95% confidence interval for odds ratio: Age, $OR = e^{\beta_1}$

$$\left[\exp\left(\hat{\beta}_1 - 1.96 \cdot SE(\hat{\beta}_1)\right), \exp\left(\hat{\beta}_1 + 1.96 \cdot SE(\hat{\beta}_1)\right)\right]$$

- 95% BI for $OR = e^{\beta_1}$: $(e^{-0.067}, e^{-0.006}) = (0.934, 0.993)$

Estimating odds ratios in R

```
> ## Odds ratios
>
> exp(donner.log$coefficients)
>
(Intercept)         Age         fem
  1.7398953   0.9650211   2.9094868     e^β0, e^β1, e^β2

> exp(confint(donner.log))
>
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.7748972 4.0431170
Age         0.9348223 0.9930661
fem         1.1543365 7.7529827

> exp(cbind(OR =donner.log$coefficients, confint(donner.log)))
>
Waiting for profiling to be done...
                   OR     2.5 %    97.5 %
(Intercept) 1.7398953 0.7748972 4.0431170
Age         0.9650211 0.9348223 0.9930661
fem         2.9094868 1.1543365 7.7529827
>
```

Estimating odds ratios in R

```
> ## Odd ratio for Survival after 10 years increased
>
> exp(donner.log$coefficients*10)
 (Intercept)          Age          fem
2.542325e+02 7.004356e-01 4.346742e+04

> exp(c(OR =donner.log$coefficients[2]*10, confint(donner.log)[2,]*10))
>
Waiting for profiling to be done...
   OR.Age     2.5 %    97.5 %
0.7004356 0.5096720 0.9327850
>
```


Plotting the survival probabilities in R

```
## Plotting the logit curve
>
> logit<-function(x)log(x/(1-x))
> ilogit<-function(x,a,b)exp(a+b*x)/(1+exp(a+b*x))
>
> ## Plotting survival for men versus women
>
> cl=coef(donner.log)
> plot(donner.na$Age,jitter(donner.na$Outcome,.2),col=cols,pch=20,
+      cex=1.2,xlab="Age",ylab="Status (jittered)")
> curve(ilogit(cl[1]+cl[2]*x+cl[3]*0,0,1),add=T)
> curve(ilogit(cl[1]+cl[2]*x+cl[3]*1,0,1),add=T,col="red")
> legend("topright",pch=20,lty="solid",col=c("red","black"),c("women","men"))
>
```

Predicting the outcome

- One can use the model to predict the outcome of certain groups of interest
- For instance, in the Donner party study one may want to predict
  - The survival probability of a man with an average age (20.22 years)
  - The survival probability of a woman with an average age (20.22 years)

Predicting the outcome

Donner party example:

$$\text{logit}\left[\hat{P}(Y=1|age,fem)\right] = 0.553 - 0.035age + 1.067fem$$

with $fem = 1$ for women and $fem = 0$ for men

a) Survival probability for a man with the average age 20.22 ($fem = 0$)

$$\hat{P}(Y=1|20.22,man) = \frac{e^{(0.553-0.035\cdot 20.22+1.067\cdot 0)}}{1+e^{(0.553-0.035\cdot 20.22+1.067\cdot 0)}} = \frac{e^{-0.1547}}{1+e^{-0.1547}}$$

$$= \frac{0.8566}{1.8566} = 0.4614$$

Predicting the outcome

Donner party example:

$$\text{logit}\left[\hat{P}(Y=1|age,fem)\right] = 0.553 - 0.035age + 1.067fem$$

with $fem = 1$ for women and $fem = 0$ for men

b) Survival probability for a woman with the average age 20.22 ($fem = 1$)

$$\hat{P}(Y=1|20.22,woman) = \frac{e^{(0.553-0.035\cdot 20.22+1.067)}}{1+e^{(0.553-0.035\cdot 20.22+1.067)}}$$

$$= \frac{e^{0.9123}}{1+e^{0.9123}} = \frac{2.49}{3.49} = 0.7134$$

Predicting the outcome in R

```
> ## Predicted probabilities of survival
>
> newdata2<-data.frame(fem=1, Age=mean(donner.na$Age))
> newdata2$greP<-predict(donner.log,newdata=newdata2,type="response")
> newdata2
>
  fem      Age     greP
1   1 20.22727 0.711279
>
> newdata3<-data.frame(fem=0, Age=mean(donner.na$Age))
> newdata3$greP<-predict(donner.log,newdata=newdata3,type="response")
> newdata3
>
  fem      Age      greP
1   0 20.22727 0.4585025
>
> newdata4<-data.frame(fem=c(0,1),Age=mean(donner.na$Age))
> newdata4$greP<-predict(donner.log,newdata=newdata4,type="response")
> newdata4
>
  fem      Age      greP
1   0 20.22727 0.4585025
2   1 20.22727 0.7112790
>
```

Model building and model selection

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> George E. P. Box

- Until now we have pretended that the relevant covariates and the structure of the model are both known
- In reality the situation is often more complex: frequently neither the relevant covariates nor the structure of the model are known before hand

Model building and model selection

- Thus, in practice scientists often consider several models (theories) to describe and explain reality
- For instance, in the Donner party example one may wonder if the effect of gender on survival varies across age or not, i.e., one may want to consider the model

$$\begin{aligned}\text{logit}\left[\hat{P}(Y=1|age,fem)\right] =& \beta_0+\beta_1 age+\beta_2 fem+\beta_3 age\cdot fem\\ =& \beta_0+\beta_1 age+(\beta_2+\beta_3\cdot age)fem\end{aligned}$$

Interaction model in R

```
> ## Interaction model
>
> m4<-glm(Outcome ~ Age*fem,data=donner.na,family=binomial(link="logit"))
> summary(m4)

Call:
glm(formula = Outcome ~ Age * fem, family = binomial(link = "logit"),
    data = donner.na)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9888  -1.0532   0.5961   1.0727   1.6317

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.39779    0.48139   0.826    0.409
Age         -0.02789    0.01911  -1.460    0.144
fem          1.47859    0.82469   1.793    0.073
Age:fem     -0.01977    0.03166  -0.624    0.532
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 108.47  on 84  degrees of freedom
AIC: 116.47

Number of Fisher Scoring iterations: 4
>
```


Interaction model

$$\text{logit}\left[\hat{P}(Y=1|age,fem)\right] = 0.398 - 0.028age + 1.478fem - 0.020age \cdot fem$$

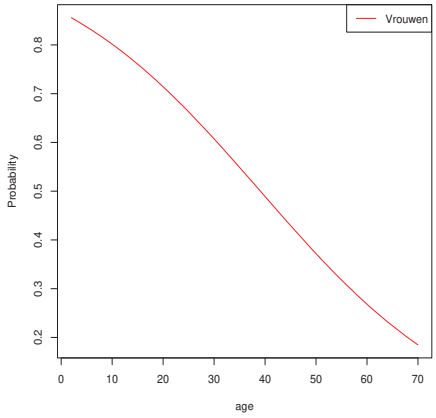
For women $fem = 1$

$$\begin{aligned}\text{logit}\left[\hat{P}(Y=1|age,women)\right] &= 0.398 - 0.028age + 1.478 - 0.020age \\ &= 1.876 - 0.048age\end{aligned}$$

$$\hat{P}(Y=1|age,women) = \frac{e^{1.876-0.048age}}{1+e^{1.876-0.048age}}$$

Interaction model

For women $fem = 1$



$$\hat{P}(Y = 1|age) = \frac{e^{1.876-0.048age}}{1 + e^{1.876-0.048age}}$$

Interaction model

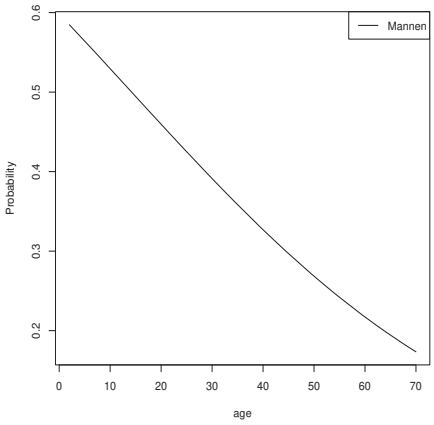
For men $fem = 0$

$$\text{logit}\left[\hat{P}(Y = 1|age, men)\right] = 0.398 - 0.028age$$

$$\hat{P}(Y = 1|age, men) = \frac{e^{0.398-0.028age}}{1 + e^{0.398-0.028age}}$$

Interaction model

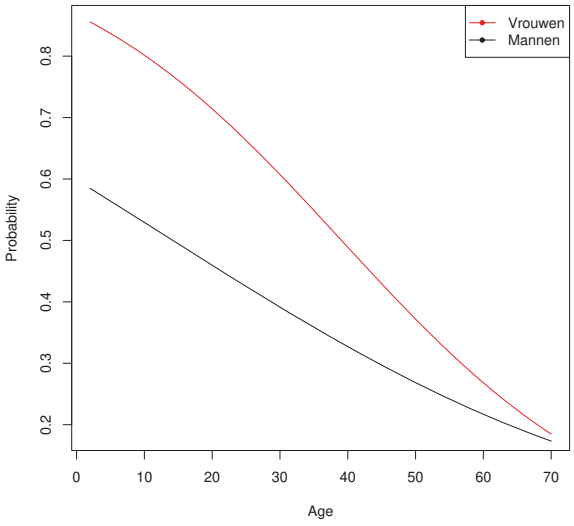
For men $fem = 0$



$$\hat{P}(Y = 1|age) = \frac{e^{0.398-0.028age}}{1 + e^{0.398-0.028age}}$$

Interaction model

Women versus men



Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: $\text{logit}\left[\hat{P}(Y = 1|age, women)\right] = 1.876 - 0.048age$
- $OR(age + 10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

Age=30

Age=20

$$\Theta_{survival|30,women} = 0.62 \cdot \Theta_{survival|20,women}$$

Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: $\text{logit}\left[\hat{P}(Y = 1|age, women)\right] = 1.876 - 0.048age$
- $OR(age + 10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

Age=60

Age=50

$$\Theta_{survival|60,women} = 0.62 \cdot \Theta_{survival|50,women}$$

Donner party data

- How do the odds of survival of a woman compare to those of a 10 years younger woman?
- For women: $\text{logit}\left[\hat{P}(Y = 1|age, women)\right] = 1.876 - 0.048age$
- $OR(age + 10, age) = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$

$$OR(age + 10, age) = \frac{Odds(survival, age + 10)}{Odds(survival, age)} = \text{Exp}(\hat{\beta}_1 \cdot 10) = e^{-0.48} = 0.62$$

Interpretation

A 10 years increase in age is associated with a 40% (30% in the model without interaction) decrease in the odds of survival

Where do the models come from?

- Sometimes a set of models is provided based on subject-matter theory, the so-called mechanistic models. One example is the PK/PD models used in pharmacokinetic/pharmacodynamics
- In practice good theory is very rare. Most often some simple restrictions are placed on the behavior one expects to find, for example, linear models, factorial models with limited interactions, etc. These models are sometimes called empirical models
- Nowadays model classes are available that can approximate many data generating mechanism. Furthermore, the computational resources to fit such models are rapidly increasing
- Model building and model selection

Model building: General principles

- **Goal**: To find a model that fits the data reasonably well without unnecessary complexity
- Model building is art and science: There are no clear, defined and fixed rules that you can automatically follow, but just general principles

P1 Use your **previous** scientific knowledge

- What are the research questions?
- What does the theory say?
- Are there results from previous studies?
- What does the *common sense* suggest?

Model building: General principles

P2 Interactions between predictors in the model should be included based on theory and plausibility

- Usually it is not necessary to evaluate all possible interactions
- Interactions between more than two predictors: very sound theoretical considerations necessary to include them

P3 There is a preference for so-called *hierarchical models* (also known as the principle of marginality)

- If the model includes an interaction, the corresponding main effects should be included as well
- If the model includes a quadratic term (x^2), a linear term should be included as well
- An intercept should be always included

Model building: General principles

P4 There is a distinction between observational and experimental studies

- Experimental research: Often limited set of factors is examined
- Model construction often less important ("true" model may be almost completely determined by the design)

P5 Importance of **replication**: A single study is **not** conclusive evidence of existence of an effect

P6 Groups or sets of predictors may belong together and, hence, move together in and out of the model

- For instance, personality can be represented using five predictors, the five factors of the Big Five
- For instance, a categorical predictor with more than two categories is included in the model using a set of dummy variables. In the final model, these dummy variables may or may not be included together

Model building: General principles

P7 Be aware of the issues associated with automatic selection procedures (stepwise, forward, backward, etc.)

- Each test is conditional on the results of the previous tests
- Distribution of these conditional statistics not fully understood
- Problems with the frequentist interpretation of α
- Multiple comparison problem
- In which sense is the final model best or optimal?
- No measure of model uncertainty

P8 Construction of a model is an iterative and creative process

Model building: General principles

P9 Inference after model construction and model selection: There is debate over which approach is correct. Active research area

P10 The objective of a study may also be the **prediction** of the criterion

- For instance, researchers may want to use a predictor(s) X to predict an outcome(s) Y
- *Understanding* is less important and, therefore, other principles can play a role

Explanation vs Prediction

- Explanation is like doing scientific research.
- Prediction is like doing engineering development. All that matters is that it works. And if the aim is prediction, model choice should be based on the quality of the predictions
- Why select a model at all?
  - It does seem a widespread misconception that model selection is about choosing **the** best model
  - For explanation one should be open to the possibility of there may be several (roughly) equally good explanatory models
  - For prediction one may want to do model averaging rather than model selection (expert opinion analogy)

Donner party data

- Reasonable models/theories

1. $\text{logit}\left[P(Y = 1|age)\right] = \beta_0 + \beta_1 age$
2. $\text{logit}\left[P(Y = 1|fem)\right] = \beta_0 + \beta_2 fem$
3. $\text{logit}\left[P(Y = 1|age, fem)\right] = \beta_0 + \beta_1 age + \beta_2 fem$
4. $\text{logit}\left[P(Y = 1|age, fem)\right] = \beta_0 + \beta_1 age + \beta_2 fem + \beta_3 fem \cdot age$

Which model should we use?

Model selection

What are you looking for?

- Model selection: One wants, given the sample, to choose a model that can describe the underlying distribution of the data
- But one only has limited information, namely the sample, and therefore one can not determine with complete certainty the underlying data generating mechanism
- Thus one looks for the most "likely" model, given your sample

- Competing models can be formally compared via
  - Nested models: Wald test, LRT
  - Nested and non-nested models: AIC (Akaike Information Criterion), BIC (Bayesian *Information* Criterion)
- Keep research question in mind!

Information criteria

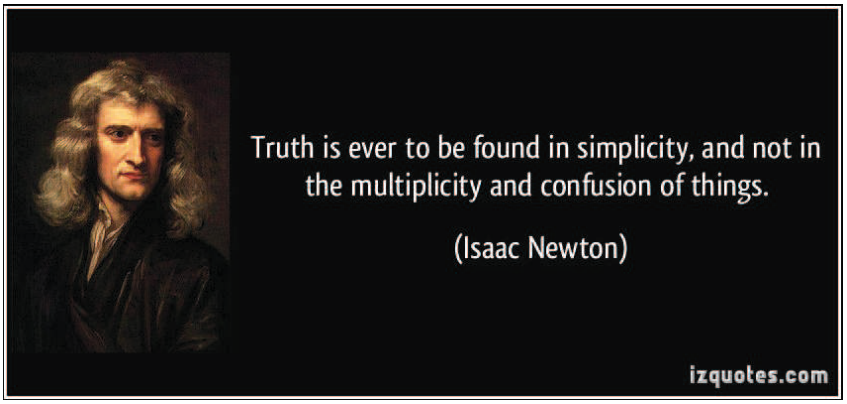
- AIC and BIC can be used to compare nested and non-nested models

$$\begin{aligned} \text{AIC} = & \quad -2\log(\text{LMAX}) + 2(\#\text{ of parameters}) \\ \text{BIC} = & \quad -2\log(\text{LMAX}) + \log(n)(\#\text{ of parameters}) \end{aligned}$$

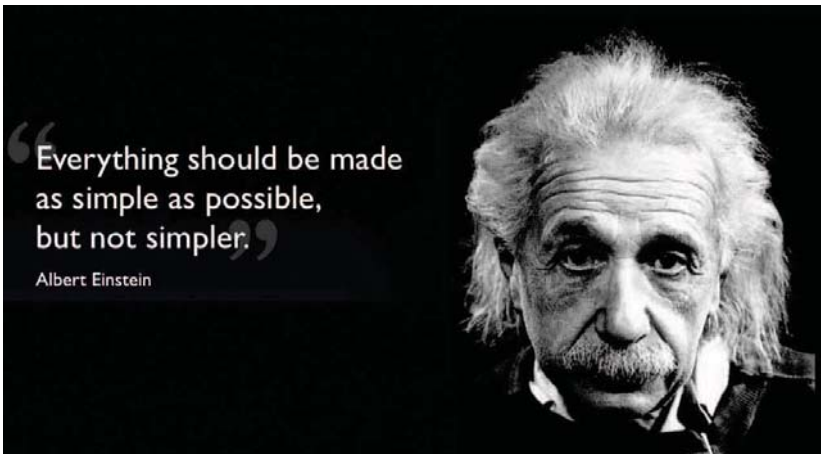
- *Penalty* for complexity, i.e., for the number of parameters used

- **Occam's razor**: Other things being equal, simpler explanations are generally better than more complex ones

Occam's razor



Occam's razor



Information criteria: AIC and BIC

- Smaller is better
- AIC and BIC are not equivalent. They have different characteristics and look for different models (different definitions of "best"). My personal choice: AIC
- AIC selects from a list of competing models the model that is "closest" to the underlying model
- "Closest" can be rigorously defined, namely, the model that minimizes the expected estimated Kullback-Leibler divergence cross-entropy

Akaike information criterion: AIC

- A single AIC value is meaningless. AIC values are meaningful only when they are compared with other AIC values
- The Akaike-weights are easier to interpret: Posterior probability that the model is the "best" model in the Kullback-Leibler sense
- Suppose one has a list of R competing models/theories then
  - Find the model with the smallest AIC_{min}
  - For every model i, compute $\Delta_i = \text{AIC}_i - \text{AIC}_{\text{min}}$
  - For every model i, compute the Akaike–weights

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{i=1}^{R}\exp(-\frac{1}{2}\Delta_i)}$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

$$w_T = \exp\left(-\tfrac{0.00}{2}\right) + \exp\left(-\tfrac{1.60}{2}\right) + \exp\left(-\tfrac{3.15}{2}\right) + \exp\left(-\tfrac{4.01}{2}\right) = 1.7909$$

$$w_1 = \frac{\exp\left(-\frac{0.00}{2}\right)}{w_T} = \frac{1}{1.7909} = 0.559$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

$$w_T = \exp\left(-\tfrac{0.00}{2}\right) + \exp\left(-\tfrac{1.60}{2}\right) + \exp\left(-\tfrac{3.15}{2}\right) + \exp\left(-\tfrac{4.01}{2}\right) = 1.7909$$

$$w_2 = \frac{\exp\left(-\frac{1.60}{2}\right)}{w_T} = \frac{0.4493}{1.7909} = 0.251$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

$$w_T = \exp\left(-\tfrac{0.00}{2}\right) + \exp\left(-\tfrac{1.60}{2}\right) + \exp\left(-\tfrac{3.15}{2}\right) + \exp\left(-\tfrac{4.01}{2}\right) = 1.7909$$

$$w_3 = \frac{\exp\left(-\frac{3.15}{2}\right)}{w_T} = \frac{0.2070}{1.7909} = 0.115$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

$$w_T = \exp\left(-\tfrac{0.00}{2}\right) + \exp\left(-\tfrac{1.60}{2}\right) + \exp\left(-\tfrac{3.15}{2}\right) + \exp\left(-\tfrac{4.01}{2}\right) = 1.7909$$

$$w_4 = \frac{\exp\left(-\frac{4.01}{2}\right)}{w_T} = \frac{0.1347}{1.7909} = 0.075$$

Donner party data: Model selection via AIC

- Model selection

Rank	Covariates			AIC	Δ_i	Akaike–weights
3	*age*			118.02	3.15	0.115
4	*fem*			118.88	4.01	0.075
1	*age*	*fem*		114.88	0.00	0.559
2	*age*	*fem*	*age* · *fem*	116.47	1.60	0.251

- Two models/theories, 1 and 2, seem to have some degree of support
- Some non-negligible level of model uncertainty
- A framework for scientific discussion: Which theory is more biologically plausible?

Akaike–weights in R

```
> ## Fitting the models
>
> donner.list=list()
>
> donner.list[[1]]=glm(Outcome ~ Age,data=donner.na,family=binomial(link="logit"))
> donner.list[[2]]=glm(Outcome ~ fem,data=donner.na,family=binomial(link="logit"))
> donner.list[[3]]=glm(Outcome ~ Age + fem,data=donner.na,family=binomial(link="logit"))
> donner.list[[4]]=glm(Outcome ~ Age*fem,data=donner.na,family=binomial(link="logit"))
>
> donner.modnames <- c("Age", "Sex", "Age+Sex", "Age+Sex+Age:Sex")
>
```


Akaike–weights in R

```
> ## Akaike weights with AICcmodavg
>
> donner.aictab=aictab(cand.set = donner.list, modnames = donner.modnames)
> donner.aictab
>
Model selection based on AICc:

                K   AICc Delta_AICc AICcWt Cum.Wt     LL
Age+Sex         3 115.15       0.00   0.56   0.56 -54.43
Age+Sex+Age:Sex 4 116.95       1.80   0.23   0.79 -54.23
Age             2 118.16       3.01   0.13   0.92 -57.01
Sex             2 119.02       3.87   0.08   1.00 -57.44

>
```

Model averaging

Model averaging is one of several methods for making formal inference from multiple models (Burnham and Anderson 2002). This approach is quite different from standard variable selection methods where inference is made only from the selected model. Model averaging admits from the beginning of the analysis that there is substantial uncertainty as to what model is best and what combination of variables is important. On the contrary, selection methods such as stepwise selection pick a single best model. Inference is then conditional on this model and variables not in the model are, therefore, deemed unimportant. These are two very different approaches.

P. M. Lukacs et al., Ann Inst Stat Math (2010) 62:117–125

Model average

- In presence of model uncertainty one may want to base inferences on several, similarly plausible, models instead of one single best model.
- One way of doing this is using a new type of model averaging estimator which averages $\hat{\beta}_i$ across several models.
- When calculating the model averaging estimator, one may consider only those models that contain the β_i parameter or, alternatively, all the models in the set of candidate models.
- The latter option is called the shrinkage estimator.

Model average

- The model averaging estimator takes the form

$$\tilde{\beta}_i = \sum_{i=1} w_j \hat{\beta}_{ij},$$

where w_j is the Akaike weight of model g_j and $\hat{\beta}_{ij}$ is the MLE of β_i calculated using model g_j.

- When using the shrinkage estimator, i.e., if all models in the candidate set $\{g_1, \ldots g_R\}$ are used to compute the average, then $\hat{\beta}_{ij} \equiv 0$ if variable i is not included in model g_j.

Model average

- The unconditional variance of $\tilde{\beta}_i$ is estimated as

$$\widehat{\text{Var}}(\tilde{\beta}_i) = \sum_{i=1} w_j \left[\widehat{\text{Var}}(\hat{\beta}_{ij}|g_j) + (\hat{\beta}_{ij} - \tilde{\beta}_i)^2 \right]$$

where w_j is the Akaike weight of model g_j and $\hat{\beta}_{ij}$ and $\widehat{\text{Var}}(\hat{\beta}_{ij}|g_j)$ are the MLE of β_i and its corresponding variance, calculated using model g_j.

- When using the shrinkage estimator, i.e., if all models in the candidate set $\{g_1, \ldots g_R\}$ are used to compute the average, then $\widehat{\text{Var}}(\hat{\beta}_{ij}|g_j) \equiv 0$ if variable i is not included in model g_j.

Model averaging in R

```
> ## Model average results
>
> modavg(cand.set= donner.list, parm="Age", second.ord=TRUE,
+ modnames = donner.modnames,  uncond.se="revised", exclude = list("Age:fem"),
+ conf.level=0.95, warn = TRUE)
>
Multimodel inference on "Age" based on AICc

AICc table used to obtain model-averaged estimate:

        K   AICc Delta_AICc AICcWt Estimate   SE
Age     2 118.16       3.01   0.18    -0.04 0.01
Age+Sex 3 115.15       0.00   0.82    -0.04 0.02

Model-averaged estimate: -0.04
Unconditional SE: 0.02
95% Unconditional confidence interval: -0.07, -0.01
>
```


Model averaging in R

```
> ## Model average results
>
> modavg(cand.set= donner.list, parm="fem", second.ord=TRUE,
+ modnames = donner.modnames, uncond.se="revised", exclude = list("Age:fem"),
+ conf.level=0.95, warn = TRUE)
>
Multimodel inference on "fem" based on AICc

AICc table used to obtain model-averaged estimate:

        K   AICc Delta_AICc AICcWt Estimate   SE
Sex     2 119.02       3.87   0.13     1.11 0.46
Age+Sex 3 115.15       0.00   0.87     1.07 0.48

Model-averaged estimate: 1.07
Unconditional SE: 0.48
95% Unconditional confidence interval: 0.13, 2.01
>
```

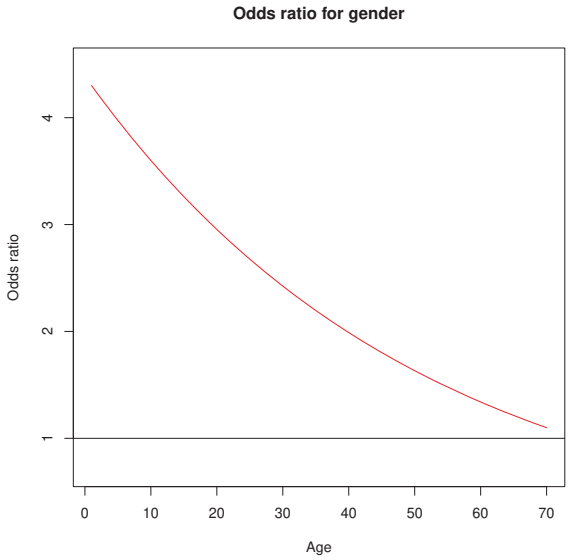
Donner party data: Conclusions

- **Main effect model**: Based on the data the estimated odds of survival for a woman is 3 ($\hat{OR} = e^{1067} \approx 3$) times larger than the corresponding odds of survival for a man of the same age, with a 95% CI for the odds ratio $(1.15, 7.75)$, approximately
- There is *model selection uncertainty*: **Interaction model** has a relatively large Akaike-weight
- However, both models indicate that the odds of survival are larger for women than for men

Odd ratio with the interaction model in R

```
> ## Gender odd ratio in the interaction model
>
> x=seq(1,70,0.01)
> y=exp(coef(m4)[3]+coef(m4)[4]*x)
>
> plot(x,y, type = "n", ylim=c(0.7, 4.5), xlab = "Age", ylab = "Odds ratio",
+ main="Odds ratio for gender")
> lines(x,y, lty = 1, col="red")
> abline(h=1)
>
```

Odd ratio in the interaction model



Donner party data: Conclusions

- It is important to note that this is an observational study (causal interpretations are therefore not justified)
- The sample was not drawn at random (inferences to a larger population are not strictly justified)

Models for longitudinal data

Ariel Alonso Abad

Catholic University of Leuven

Studying changes over time

- Longitudinal data: Studying changes over time.
- Exploratory data analysis.
- Introduction to multilevel models.
- Formulation and interpretation of the models.
- Implementation in R.
- Statistical inference
  - Fixed effects
  - Random effects

Studying changes over time

- Changes over time play pivotal role in science.
- Original ideas
  - ⇒ British astronomer George Biddel Airy 1861.
  - ⇒ Laird and Ware (1982): Life sciences.
  - ⇒ Goldstein (1979): Humanities.
- Computing power and software available in the 1990s.
- Synthesis: Intra and inter individual changes need to be modeled.

Different names similar models

- Individual growth models.
- Random coefficient models.
- Multilevel models.
- Mixed models.
- Hierarchical (linear) models.
- Growth curve models.

Why multilevel?

- Level 1: Changes within individuals.
  - ⇒ Can we describe the time evolution for each individual with a linear function?
- Level 2: Changes between individuals.
  - ⇒ Are the individuals different at the beginning of the study?
  - ⇒ Do they evolve differently over time?

Distinguishing quality

Longitudinal Studies: Repeated measurements over time (Waves)

- ⇒ Metric: Time, age, weeks since treated...
- ⇒ Spacing: Equal time intervals?
- ⇒ Time structure: All individual measured at the same time points?
- ⇒ Balanced: Same number of measurements for all individuals?

Cross-Sectional Studies: Only one measurement per subject. Nothing can be concluded about time changes.

Effect of early dietary intervention on children IQ

- 103 African American, low income families. Randomized to
  - 58 early intervention program.
  - 45 control group.
- Evaluated on ages 12, 18, and 24 months.
- Research question: Effect of the early intervention on the evolution of cognitive performance?

Effect of early dietary intervention on children IQ

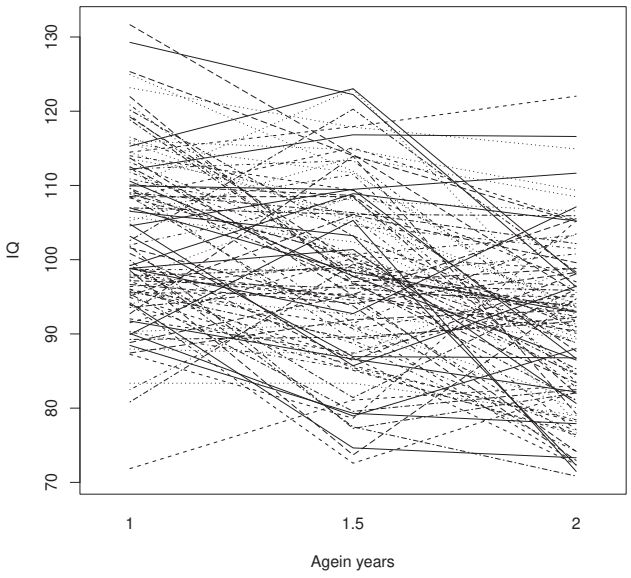
id	program	age	cog
1	1	1.00	106.98
1	1	1.50	98.31
1	1	2.00	92.91
2	1	1.00	108.86
2	1	1.50	100.29
2	1	2.00	85.30
3	1	1.00	112.52
3	1	1.50	96.77
3	1	2.00	83.43
4	1	1.00	90.24
4	1	1.50	85.27
4	1	2.00	76.41
5	1	1.00	105.71
5	1	1.50	102.40
5	1	2.00	88.79
6	1	1.00	93.89
6	1	1.50	85.10
6	1	2.00	76.66
7	1	1.00	109.94
⋮	⋮	⋮	⋮

- Fully balanced: Age=1.0, 1.5 and 2.0 years.
- PROGRAM: 1-intervention, 0-control.
- COG is a nationally normed scale.

Exploratory analysis

- Spaghettiplot: Individual profiles. Points are joined with lines.
- Descriptive tables.
- Box plots.
- Mean plots.
- Individual regressions.

Effect of early intervention: Spaghetti-plot



Spaghettiplot: R code

Let us get started with R:

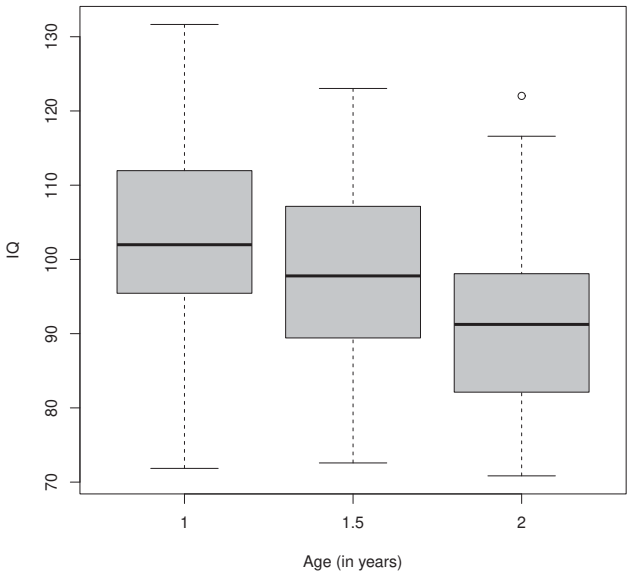
```
> ## Reading in the early.int data
>
> early.int1 <- read.table("earlyint.txt", header=T, sep=",")
>
> ## Attach data to the search path
>
> attach(early.int1)
>
> ## Spaghettiplot
>
> n=length(unique(id))
> interaction.plot(age,id,cog, xlab="Agein years", ylab="IQ",
+   legend=F)
>
```


Means per time point and group

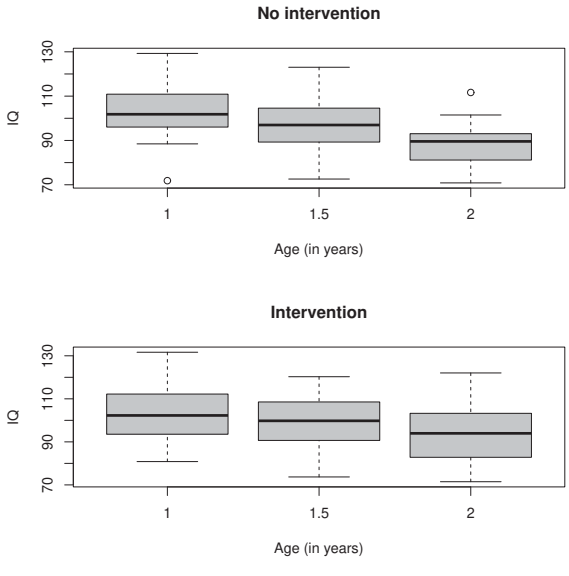
			Program	
			0	1
Age	Resp	Statistics		
1	IQ	n	45	58
		Mean	103.93	102.93
		Sd	11.01	11.78
1.5	IQ	n	45	58
		Mean	96.91	99.18
		Sd	11.93	12.02
2	IQ	n	45	58
		Mean	87.68	92.99
		Sd	9.05	12.13

```
> ## Descriptives
>
> ## Mean:
> early.mean=tapply(cog,list(age,program),mean)
>
> ## Standard deviation:
> early.sd=tapply(cog,list(age,program),sd)
>
> ## Variance:
> early.var=tapply(cog,list(age,program),var)
>
> ## Frequency:
> early.n=table(age,program)
>
```

Boxplot



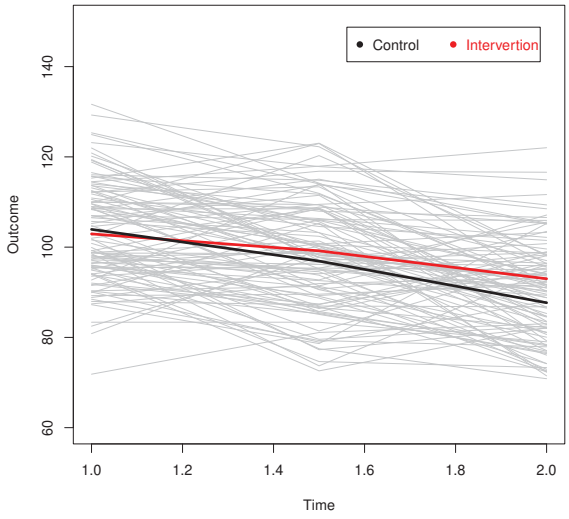
Boxplot per program



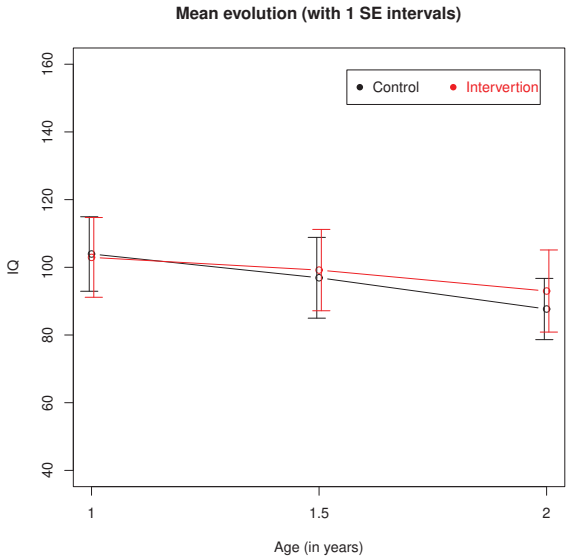
R code

```
> ## Boxplots:
>
> boxplot(cog~age,xlab="Age (in years)",ylab="IQ")
>
> ## Boxplots per program
>
> par(mfrow=c(2,1))
> boxplot(cog[program==0]~age[program==0],main="No intervention",
+ main="No intervention",xlab="Age (in years)",ylab="IQ")
>
> boxplot(cog[program==1]~age[program==1],main="Intervention",
+ main="No intervention",xlab="Age (in years)",ylab="IQ")
>
```


Mean evolution



Mean evolution



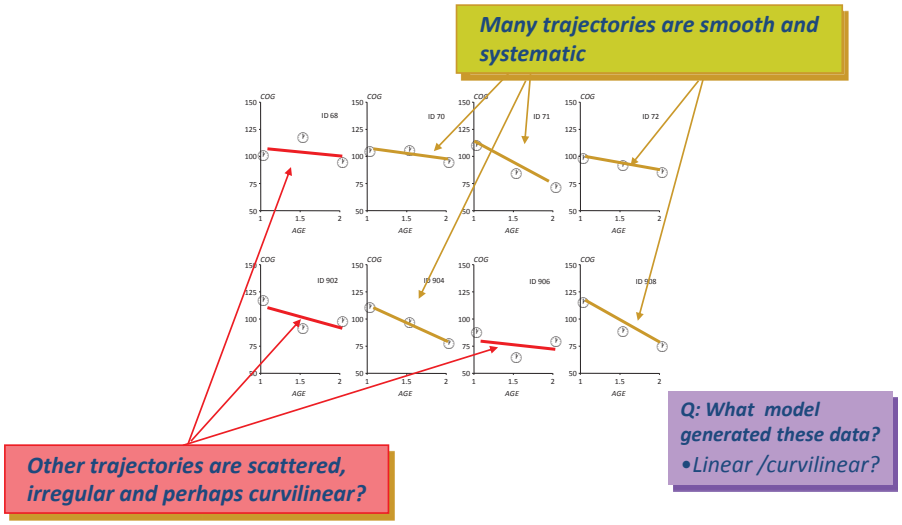
R code

```
####################################################################
#           General function to plot error bars                    #
####################################################################

errbar=function(x,y,height,width,lty=1,col="black")
arrows(x,y,x,y+height,angle=90,length=width,lty=lty,
col=col)
arrows(x,y,x,y-height,angle=90,length=width,lty=lty,
col=col)

> ## Plotting mean evolutions
>
> plot(age[id==1],early.mean[,1],type="b",xlim=c(1,2),
+ ylim=c(40,160),xlab="Age (in years)",ylab="IQ",axes=F,
+ main="Mean evolution (with 1 SE intervals)")
> axis(side=1,at=c(1,1.5,2),labels=c(1,1.5,2))
> axis(side=2,at=seq(40,160,20))
>
> box()
> points(age[id==1],early.mean[,2],type="b",col="red")
> errbar(age[id==1]-.005,early.mean[,1],early.sd[,1],.1)
> errbar(age[id==1]+.005,early.mean[,2],early.sd[,2],.1,col="red")
>
```

Correlations: R code

```
> ## Reshaping the data into a wide form
> early.int2 <- reshape(early.int1,
+  timevar = "age", idvar = c("id", "program"), direction = "wide")
> early.int2
>
     id program     cog.1   cog.1.5     cog.2
1     1       1 106.98289  98.31060  92.91342
2     2       1 108.86019 100.29307  85.29502
3     3       1 112.52438  96.76684  83.42649
4     4       1  90.24428  85.27380  76.41052
5     5       1 105.70738 102.39839  88.78872
6     6       1  93.88987  85.09601  76.66209
.........................................
>
> ## Correlation between the IQ scores  at  different ages
> cor(early.int2[,3:5])
>
            cog.1   cog.1.5     cog.2
cog.1   1.0000000 0.5816070 0.3263912
cog.1.5 0.5816070 1.0000000 0.4371109
cog.2   0.3263912 0.4371109 1.0000000
>
```


Linear regression per person



Linear regression per person

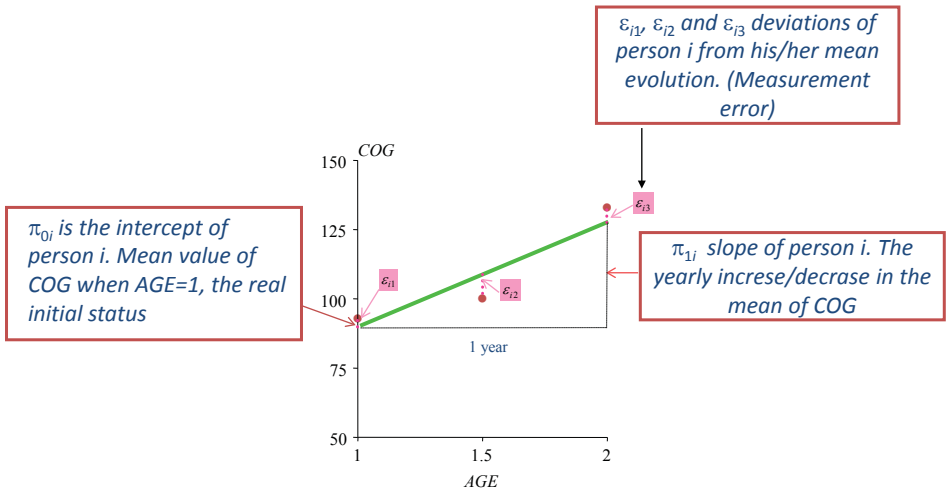
Model

Model for subject i

$$Y_{ij} = \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

- Y_{ij} denotes COG for subject i at Age_{ij}.
- π_{0i} intercept for subject i at $Age_{ij} = 1$.
- π_{1i} slope for subject i.
- ε_{ij} error term $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$.

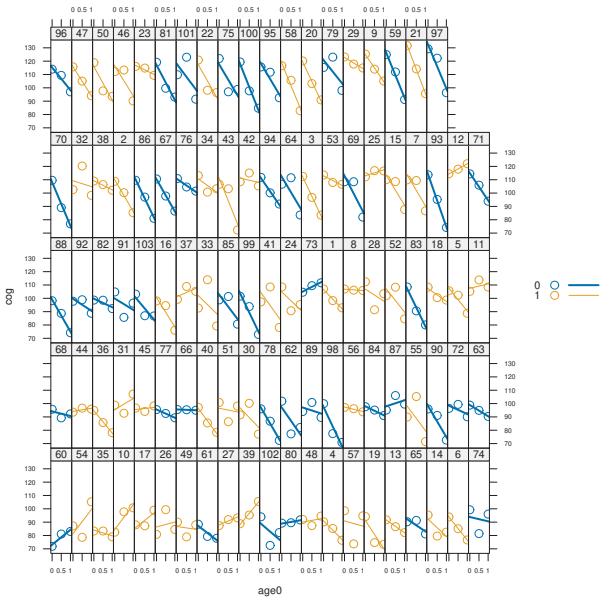
Interpretation of the model



Linear regression per person: Trellis graph

- The aspect ratio of the panels (ratio of the height to the width) chosen according to an algorithm described in Cleveland (1993) to facilitate comparison of slopes
- The effect is to have the slopes of the lines on the page distributed around ± 45, thereby making it easier to detect systematic changes in slopes
- The panels have been ordered (from left to right starting at the bottom row) by increasing intercept
- If there were a correlation between initial status (intercept) and rate of change (slope) then slopes would show an increasing trend (or a decreasing trend) in the left to right, bottom to top ordering.

Linear regression per person



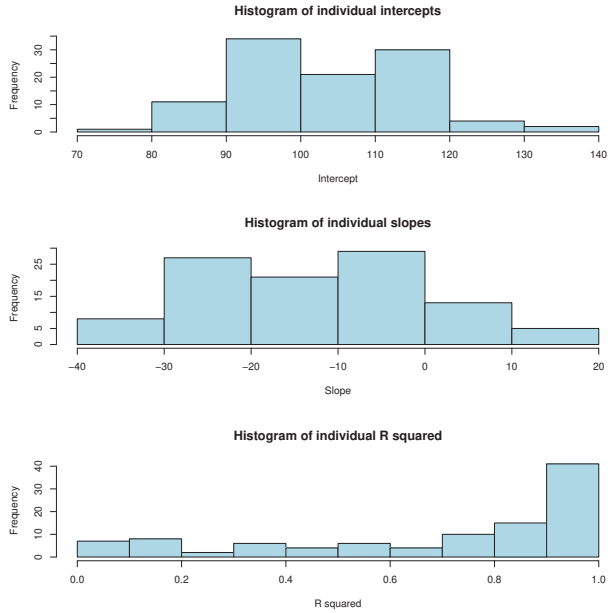
Linear regression per person: R code

```
>## Creating the time variable
>
early.int1$age0<-early.int1$age-1
>
> ## Displaying the linear regression per person
>
> cf<-sapply(early.int1$id, function(x)
+    coef(lm(cog~age0, data=subset(early.int1, id==x))))
>
> Sx<-reorder(early.int1$id, cf[1,])
>
> xyplot(cog ~ age0|Sx,groups=program,data=early.int1,
+ type=c('p','r'),auto.key=T,aspect="xy",
+ par.settings=list(axis.text=list(cex=0.6),
+ fontsize=list(text=8, points=10)),
+ scales=list(
+ x=list(
+ at=c(0,0.5,1),
+ labels=c("0","0.5","1")))
)
>
```


Linear regression per person: R code

```
> ## Linear regression per participant of cog on age
>
> ## Coefficients
> lin.reg.coef <- by(early.int1, early.int1$id,
+          function(data) coef(lm(cog ~ age0, data=data)))
> lin.reg.coef1 <- unlist(lin.reg.coef)
> names(lin.reg.coef1) <- NULL
> lin.reg.coef2=matrix(lin.reg.coef1,length(lin.reg.coef1)/2,2,byrow = TRUE)
>
> ## R squared
> lin.reg.r.squared <- by(early.int1, early.int1$id,
+          function(data) summary(lm(cog ~ age, data=data))$r.squared )
lin.reg.r.squared1<- as.vector(unlist(lin.reg.r.squared))
>
> ## Histograms
> par(mfrow=c(3,1))
> hist(lin.reg.coef2[,1],xlab="Intercept",col="lightblue",main="Histogram of individual intercepts")
> hist(lin.reg.coef2[,2],xlab="Slope",col="lightblue",main="Histogram of individual slopes")
> hist(lin.reg.r.squared1,xlab="R squared",col="lightblue",main="Histogram of individual R squared")
>
```

Between subject variability



Multilevel models

Level 1

$$Y_{ij} = \boxed{\pi_{0i} + \pi_{1i}(Age_{ij} - 1)} + \boxed{\varepsilon_{ij}}$$

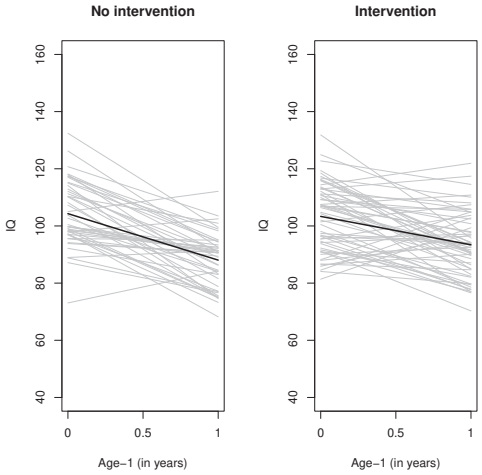
$\Rightarrow$ Structural part of the level 1. How individuals evolve.

$\Rightarrow$ Random part of Level 1. How individuals deviate from their own evolution.

$\Rightarrow$ Why do π_{0i} and π_{1i} vary?

$\Rightarrow$ Is due to the effect of the intervention program?

Multilevel models



⇒ Program is not the entire story.

⇒ How can we handle the unexplained variability?

Linear regression per person and group: R code

```
> ## Plotting individual regression lines per group
>
> reg.coef=cbind(lin.reg.coef2, early.int1[early.int1$age==1,]$program)
>
> mean.int<-tapply(reg.coef[,1],reg.coef[,3],mean)
> mean.slope<-tapply(reg.coef[,2],reg.coef[,3],mean)
>
> par(mfrow=c(1,2))
> plot(age,cog,type="n",xlim=c(1,2),ylim=c(40,160),main="No intervention",
+      xlab="Age-1 (in years)",ylab="IQ",axes=F)
> axis(side=1,at=c(1,1.5,2),labels=c(1,1.5,2))
> axis(side=2,at=seq(40,160,20))
> box()
> for (i in 1:103)
+ if (reg.coef[i,3]==0)
+ curve(cbind(1,x)%*%reg.coef[i,1:2],add=T,col="gray")
> curve(cbind(1,x)%*%c(mean.int[1],mean.slope[1]),add=T,lwd=2)
>
> plot(age,cog,type="n",xlim=c(1,2),ylim=c(40,160),main="Intervention",
+      xlab="Age-1 (in years)",ylab="IQ",axes=F)
> axis(side=1,at=c(1,1.5,2),labels=c(1,1.5,2))
> axis(side=2,at=seq(40,160,20))
> box()
> for (i in 1:103)
+ if (reg.coef[i,3]==1)
+ curve(cbind(1,x)%*%reg.coef[i,1:2],add=T,col="gray")
> curve(cbind(1,x)%*%c(mean.int[2],mean.slope[2]),add=T,lwd=2)
```

Multilevel models

Level 1

$$Y_{ij} = \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

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

Level 2

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} PROG_i + b_{0i} & \text{explaining the intercept} \\ \pi_{1i} = \gamma_{10} + \gamma_{11} PROG_i + b_{1i} & \text{explaining the slope} \end{cases}$$

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_0 & \sigma_{01} \\ \sigma_{01} & \sigma^2_1 \end{pmatrix} \right]$$

Parameters interpretation

Symbol	Definition
σ^2_0	Level 2 residual variance in true intercept π_{0i} across all individuals in the population, after controlling for program participation
σ^2_1	Level 2 residual variance in true slope π_{1i} across all individuals in the population, after controlling for program participation
σ_{01}	Level 2 residual covariance between true intercept π_{0i} and slope π_{1i} across all individuals in the population, after controlling for program participation

Explaining variation:

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} PROG_i + b_{0i} \\ \pi_{1i} = \gamma_{10} + \gamma_{11} PROG_i + b_{1i} \end{cases}$$

Control Group $PROG_i = 0$

$$\begin{cases} \pi_{0i} = \gamma_{00} + b_{0i} \\ \pi_{1i} = \gamma_{10} + b_{1i} \end{cases}$$

Intervention Group $PROG_i = 1$

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} + b_{0i} \\ \pi_{1i} = \gamma_{10} + \gamma_{11} + b_{1i} \end{cases}$$

Parameters interpretation

Symbol	Definition
σ_0^2	Level 2 residual variance in true intercept π_{0i} across all individuals in the population, after controlling for program participation
σ_1^2	Level 2 residual variance in true slope π_{1i} across all individuals in the population, after controlling for program participation
σ_{01}	Level 2 residual covariance between true intercept π_{0i} and slope π_{1i} across all individuals in the population, after controlling for program participation

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right]$$

Final model

Hierarchical model

$$\begin{cases} Y_{ij} = & \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij} \\ \pi_{0i} = & \gamma_{00} + \gamma_{01} PROG_i + b_{0i} \\ \pi_{1i} = & \gamma_{10} + \gamma_{11} PROG_i + b_{1i} \end{cases}$$

Distributional assumptions

$$\begin{cases} \varepsilon_{ij} \sim & N(0, \sigma_\varepsilon^2) \\ \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim & N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}\right] \end{cases}$$

One single model

Model

$$\begin{aligned} Y_{ij} = \quad & \gamma_{00} + \gamma_{01} PROG_i + \gamma_{10}(Age_{ij} - 1) + \gamma_{11} PROG_i(Age_{ij} - 1) + \\ & b_{0i} + b_{1i}(Age_{ij} - 1) + \\ & \varepsilon_{ij} \end{aligned}$$

Distributional Assumptions

$$\begin{aligned} \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \\ \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right] \end{aligned}$$

One single model

Model

$$\begin{aligned} Y_{ij} = \quad & \overbrace{\gamma_{00} + \gamma_{01} PROG_i + \gamma_{10}(Age_{ij} - 1) + \gamma_{11} PROG_i(Age_{ij} - 1)}^{\text{Fixed effects}} + \\ & b_{0i} + b_{1i}(Age_{ij} - 1) + \rightarrow \text{Random effects} \\ & \varepsilon_{ij} \rightarrow \text{Error} \end{aligned}$$

Distributional Assumptions

$$\begin{aligned} \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \\ \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right] \end{aligned}$$

Expected evolution: Control

Control group $PROG_i = 0$

$$Y_{ij} = \gamma_{00} + \gamma_{10}(Age_{ij} - 1) + b_{0i} + b_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

$$\mathrm{E}\,(Y_{ij}|PROG_i = 0) = \gamma_{00} + \gamma_{10}(Age_{ij} - 1)$$

Intervention group $PROG_i = 1$

$$Y_{ij} = (\gamma_{00} + \gamma_{01}) + (\gamma_{10} + \gamma_{11})(Age_{ij} - 1) + b_{0i} + b_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

$$\mathrm{E}\,(Y_{ij}|PROG_i = 1) = (\gamma_{00} + \gamma_{01}) + (\gamma_{10} + \gamma_{11})(Age_{ij} - 1)$$

Hypotheses of interest

Hierarchical model

$$\begin{cases} Y_{ij} = & \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij} \\ \pi_{0i} = & \gamma_{00} + \gamma_{01} PROG_i + b_{0i} \\ \pi_{1i} = & \gamma_{10} + \gamma_{11} PROG_i + b_{1i} \end{cases}$$

Hypotheses of interest

$$H_0 : \gamma_{01} = 0 \quad H_1 : \gamma_{01} \neq 0$$

$$H_0 : \gamma_{11} = 0 \quad H_1 : \gamma_{11} \neq 0$$

Fitting the model

Model

$$Y_{ij} = \gamma_{00} + \gamma_{01} PROG_i + \gamma_{10}(Age_{ij} - 1) + \gamma_{11} PROG_i (Age_{ij} - 1) + b_{0i} + b_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

$\Rightarrow$ Parameters are estimated via

- **Maximum likelihood (ML).**
- **Restricted maximum likelihood (REML).**
- **What is that?**

$\Rightarrow$ R: lmer (packages: nlme, lme4 or arm)

A 2-stage Model Formulation: A bit of theory

Stage 1

- Response Y_{ij} for ith subject, measured at time t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$
- Response vector $\boldsymbol{Y_i}$ for ith subject: $\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$
- **Stage 1 model**:

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon_i}$$

- $\boldsymbol{Z}_i$ is a $(n_i \times q)$ matrix of known covariates
- $\boldsymbol{\beta}_i$ is a q-dimensional vector of subject-specific regression coefficients
- $\boldsymbol{\varepsilon_i} \sim N(0, \Sigma_i)$, often $\Sigma_i = \sigma^2 \boldsymbol{I}_{n_i}$
- Note that the above model describes the observed variability within subjects

Dietary intervention example

The 1-stage model

$$Y_{ij} = \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij}$$

can be rewritten in matrix form as

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i$$

where

$$\overbrace{\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{pmatrix}}^{\boldsymbol{Y}_i} = \overbrace{\begin{pmatrix} 1 & Age_{i1} - 1 \\ 1 & Age_{i2} - 1 \\ 1 & Age_{i3} - 1 \end{pmatrix}}^{\boldsymbol{Z}_i} \overbrace{\begin{pmatrix} \pi_{0i} \\ \pi_{1i} \end{pmatrix}}^{\boldsymbol{\beta}_i} + \overbrace{\begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{pmatrix}}^{\boldsymbol{\varepsilon}_i}$$

A 2-stage Model Formulation: A bit of theory

Stage 2

- Between-subject variability can now be studied from relating the $\boldsymbol{\beta}_i$ to known covariates
- **Stage 2 model**:

$$\boldsymbol{\beta}_i = \boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i$$

- $\boldsymbol{K}_i$ is a $(q \times p)$ matrix of known covariates
- $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression parameters
- $\boldsymbol{b}_i \sim N(0, \boldsymbol{D})$

Dietary intervention example

The 2-stage model

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} PROG_i + b_{0i} \\ \pi_{1i} = \gamma_{10} + \gamma_{11} PROG_i + b_{1i} \end{cases}$$

can be rewritten in matrix form as

$$\boldsymbol{\beta}_i = \boldsymbol{K}_i \boldsymbol{\beta} + \boldsymbol{b}_i$$

where

$$\overbrace{\begin{pmatrix} \pi_{0i} \\ \pi_{1i} \end{pmatrix}}^{\boldsymbol{\beta}_i} = \overbrace{\begin{pmatrix} 1 & PROG_i & 0 & 0 \\ 0 & 0 & 1 & PROG_i \end{pmatrix}}^{\boldsymbol{K}_i} \overbrace{\begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{pmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}}^{\boldsymbol{b}_i}$$

The general linear mixed-effects model

- A 2-stage approach can be performed explicitly in the analysis
- Combining the two stages into one model leads to:

$$\begin{cases} \boldsymbol{Y}_i &= \boldsymbol{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon_i} \\ \boldsymbol{\beta}_i &= \boldsymbol{K}_i \boldsymbol{\beta} + \boldsymbol{b}_i \end{cases}$$

- and plugging $\boldsymbol{\beta}_i$ into the expression for $\boldsymbol{Y}_i$

$$\Rightarrow \boldsymbol{Y}_i = \underbrace{\boldsymbol{Z}_i \boldsymbol{K}_i}_{\boldsymbol{X}_i} \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon_i} = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon_i}$$

The general linear mixed-effects model

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \varepsilon_{\boldsymbol{i}} \\ \boldsymbol{b}_i \sim N(0, \boldsymbol{D}), \quad \varepsilon_{\boldsymbol{i}} \sim N(0, \Sigma_i), \\ \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \varepsilon_1, \ldots, \varepsilon_{\boldsymbol{N}} \text{ independent} \end{cases}$$

- Terminology:
  - Fixed effects: β
  - Random effects: $\boldsymbol{b}_i$
  - Variance components: elements in $\boldsymbol{D}$ and Σ_i

Hierarchical versus marginal model

- The general linear mixed model (LMM) is given by:

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \varepsilon_{\boldsymbol{i}} \\ \boldsymbol{b}_i \sim N(0, \boldsymbol{D}), \quad \varepsilon_{\boldsymbol{i}} \sim N(0, \Sigma_i), \\ \boldsymbol{b}_1, \ldots, \boldsymbol{b}_N, \varepsilon_1, \ldots, \varepsilon_{\boldsymbol{N}} \text{ independent} \end{cases}$$

- It can be rewritten as:

$$f(\boldsymbol{Y}_i|\boldsymbol{b}_i) = N(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \Sigma_i),$$

$$f(\boldsymbol{b}_i) = N(0, \boldsymbol{D})$$

Hierarchical versus marginal model

- It is therefore also called a hierarchical model:
  - A model for $\boldsymbol{Y}_i$ given $\boldsymbol{b}_i$: $f(\boldsymbol{Y}_i|\boldsymbol{b}_i)$
  - A model for $\boldsymbol{b}_i$: $f(\boldsymbol{b}_i)$
- Marginally, we have that $\boldsymbol{Y}_i$ is distributed as:

$$f(\boldsymbol{Y}_i) = \int f(\boldsymbol{Y}_i|\boldsymbol{b}_i) f(\boldsymbol{b}_i)\, d\boldsymbol{b}_i = N(\boldsymbol{X}_i\beta, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \Sigma_i)$$

- Hence, very specific assumptions are made about the dependence of the mean and covariance on the covariates $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$:
  - **Implied mean :** $\boldsymbol{X}_i\beta$
  - **Implied covariance :** $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \Sigma_i$
- The hierarchical model always implies a marginal one, **NOT** vice versa

Estimation of the Marginal Model

- Recall that the general linear mixed model equals

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon_i}$$

$$\left.\begin{array}{l} \boldsymbol{b}_i \sim N(0, \boldsymbol{D}) \\ \boldsymbol{\varepsilon_i} \sim N(0, \Sigma_i) \end{array}\right\} \text{ independent}$$

- The implied marginal model equals

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\beta, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \Sigma_i)$$

- Inferences based on the marginal model do not explicitly assume the presence of random effects representing the natural heterogeneity between subjects

Estimation of the Marginal Model

- Notation:
  - $\boldsymbol{\beta}$: vector of fixed effects (as before)
  - $\boldsymbol{\alpha}$: vector of all variance components in $\boldsymbol{D}$ and Σ_i
  - $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$: vector of all parameters in marginal model
- Marginal likelihood function:

$$L_{ML}(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (2\pi)^{-n_i/2} |\boldsymbol{V}_i(\boldsymbol{\alpha})|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha}) (\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right] \right\}$$

- If $\boldsymbol{\alpha}$ were known, MLE of $\boldsymbol{\beta}$ equals

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{y}_i,$$

  where $\boldsymbol{W}_i$ equals $\boldsymbol{V}_i^{-1}$.

Estimation of the Marginal Model

- In most cases, $\boldsymbol{\alpha}$ is not known, and needs to be replaced by an estimate $\widehat{\boldsymbol{\alpha}}$
- Two frequently used estimation methods for $\boldsymbol{\alpha}$:
  - Maximum likelihood
  - Restricted maximum likelihood

Maximum Likelihood Estimation (ML)

- $\widehat{\boldsymbol{\alpha}}_{ML}$ obtained from maximizing

$$L_{ML}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}))$$

with respect to $\boldsymbol{\alpha}$

- The resulting estimate $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\alpha}}_{ML})$ for $\boldsymbol{\beta}$ will be denoted by $\widehat{\boldsymbol{\beta}}_{ML}$
- $\widehat{\boldsymbol{\alpha}}_{ML}$ and $\widehat{\boldsymbol{\beta}}_{ML}$ can also be obtained from maximizing $L_{ML}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, i.e., with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously.

Restricted Maximum Likelihood Estimation (REML)

- We first combine all models

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$$

into one model

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$$

in which

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix}, \quad \boldsymbol{V}(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{V}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{V}_N \end{pmatrix}$$

- The data are transformed orthogonal to $\boldsymbol{X}$ ($\boldsymbol{A}'\boldsymbol{X} = 0$):

$$\boldsymbol{U} = \boldsymbol{A}'\boldsymbol{Y} \sim N(0, \boldsymbol{A}'\boldsymbol{V}(\boldsymbol{\alpha})\boldsymbol{A})$$

Restricted Maximum Likelihood Estimation (REML)

- The MLE of $\boldsymbol{\alpha}$, based on $\boldsymbol{U}$, is called the REML estimate and is denoted by $\widehat{\boldsymbol{\alpha}}_{REML}$
- The resulting estimate $\widehat{\boldsymbol{\beta}}(\widehat{\alpha}_{REML})$ for $\boldsymbol{\beta}$ will be denoted by $\widehat{\boldsymbol{\beta}}_{REML}$
- $\widehat{\boldsymbol{\alpha}}_{REML}$ and $\widehat{\boldsymbol{\beta}}_{REML}$ can also be obtained from maximizing

$$L_{REML}(\boldsymbol{\theta}) = \left| \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i(\boldsymbol{\alpha}) \boldsymbol{X}_i \right|^{-\frac{1}{2}} L_{ML}(\boldsymbol{\theta})$$

  with respect to $\boldsymbol{\theta}$, i.e., with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously.

- $L_{REML}\left(\boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})\right)$ is the likelihood of the error contrasts $\boldsymbol{U}$, and is often called the REML likelihood function. It is **NOT** the likelihood for the original data $\boldsymbol{Y}$

Restricted versus Maximum Likelihood Estimation

- The **MLE** is **negatively biased** for the estimation of **variance components**, but the bias gets smaller for larger sample sizes (asymptotically unbiased)
- **REML** is **unbiased** for the estimation of **variance components** and, therefore, it may be a better option for small sample sizes
- Likelihood ratio tests (LRT) based on REML require **exactly the same fixed effects specification** in both models (Why?). So, comparing models with different fixed effects (a common scenario) using an LRT, requires ML

Fitting the model: R code

```
> ## Installing the packages
>
> install.packages("lme4")
> install.packages("arm")
> install.packages("nlme")
>
> ## Loading the packages
>
> library(lme4)
> library(lattice)
> library(nlme)
> library(arm)
> library(car)
>
> ## Creating the time variable
>
> early.int1$age0<-early.int1$age-1
>
> ## Fitting the model with ML
>
> early.lmer1<-lmer(cog~1+age0*program+(1 + age0|id), REML = FALSE,
+                   data=early.int1)
>
```


R code: Remarks

- $(1 + age0|id)$ subject specific part: $b_{0i} + b_{1i}(Age_{ij} - 1)$
- Intercept is default: $(age0|id)$
- $age0 * program$: Fixed effects

$$\gamma_{00} + \gamma_{01} PROG_i + \gamma_{10}(Age_{ij} - 1) + \gamma_{11} PROG_i(Age_{ij} - 1)$$

- Default estimation procedure is REML.
- $REML = FALSE$ calculates MLE!

R Output

```
> summary(early.lmer1)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: cog ~ 1 + age0 * program + (1 + age0 | id)
   Data: early.int1

     AIC      BIC   logLik deviance df.resid
  2332.5   2362.4  -1158.3   2316.5      301

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.25361 -0.59088  0.02132  0.56849  2.29366

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 84.02    9.166
          age0        39.44    6.280    -0.55
 Residual             60.31    7.766
Number of obs: 309, groups:  id, 103

Fixed effects:
             Estimate Std. Error t value
(Intercept)  104.3007     1.7274   60.38
age0         -16.2555     1.8860   -8.62
program       -0.9646     2.3020   -0.42
age0:program   6.3187     2.5133    2.51

Correlation of Fixed Effects:
            (Intr) age0   progrm
age0        -0.629
program     -0.750  0.472
age0:progrm  0.472 -0.750 -0.629
>
```


R Output

```
> display(early.lmer1)
>
lmer(formula = cog ~ 1 + age0 * program + (1 + age0 | id), data = early.int1,
    REML = FALSE)
            coef.est coef.se
(Intercept)  104.30     1.73
age0         -16.26     1.89
program       -0.96     2.30
age0:program   6.32     2.51

Error terms:
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 9.17
          age0        6.28     -0.55
 Residual             7.77
---
number of obs: 309, groups: id, 103
AIC = 2332.5, DIC = 2316.5
deviance = 2316.5
>
> anova(early.lmer1)
>
Analysis of Variance Table
             Df Sum Sq Mean Sq  F value
age0          1 6256.8  6256.8 103.7473
program       1  134.8   134.8   2.2344
age0:program  1  381.2   381.2   6.3208
>
```

Inference for the Fixed Effects

- Estimate for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{y}_i,$$

  where $\boldsymbol{W}_i = \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha})$ and $\boldsymbol{\alpha}$ replaced by its ML or REML estimate

- Conditional on $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})$ is asymptotically multivariate normal with mean $\boldsymbol{\beta}$ and covariance

$$\text{Var}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i \right)^{-1}$$

- In practice one again replaces $\boldsymbol{\alpha}$ by its ML or REML estimate

Approximate Wald Test

- For any known matrix L, consider testing

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = 0, \quad \text{versus} \quad H_A : \boldsymbol{L}\boldsymbol{\beta} \neq 0$$

- Wald test statistic:

$$G = \widehat{\boldsymbol{\beta}}' \boldsymbol{L}' \left[\boldsymbol{L} \underbrace{\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{X}_i \right)^{-1}}_{\text{Var}(\widehat{\boldsymbol{\beta}})^{-1}} \boldsymbol{L}' \right]^{-1} \boldsymbol{L}\widehat{\boldsymbol{\beta}}$$

- Conditional on $\boldsymbol{\alpha}$ the asymptotic null distribution of G is χ^2 with rank(L) degrees of freedom

Approximate t-test and F-test

- Wald test based on

$$\text{Var}(\widehat{\boldsymbol{\beta}}(\boldsymbol{\alpha})) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1}$$

- In practice $\boldsymbol{\alpha}$ is replaced by an estimate but...
- The variability introduced from replacing $\boldsymbol{\alpha}$ by some estimate is not taken into account in Wald tests
- Therefore, Wald tests will only provide valid inferences in sufficiently large samples
- This is often solved by replacing the χ^2 distribution by an appropriate F-distribution (and the normal by a t).

Approximate t-test and F-test

- For any known matrix $\boldsymbol{L}$, consider testing

$$H_0 : \boldsymbol{L}\boldsymbol{\beta} = 0, \quad \text{versus} \quad H_A : \boldsymbol{L}\boldsymbol{\beta} \neq 0$$

- F test statistic:

$$F = \frac{\widehat{\boldsymbol{\beta}}' \boldsymbol{L}' \left[\boldsymbol{L}\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}_i^{-1}(\widehat{\boldsymbol{\alpha}}) \boldsymbol{X}_i\right)^{-1} \boldsymbol{L}'\right]^{-1} \boldsymbol{L}\widehat{\boldsymbol{\beta}}}{\text{rank}(\boldsymbol{L})}.$$

- Approximate null-distribution of F is F with numerator degrees of freedom equal to rank($\boldsymbol{L}$)

Approximate t-test and F-test

- Approximate null-distribution of F is F with numerator degrees of freedom equal to rank($\boldsymbol{L}$)
- Denominator degrees of freedom to be estimated from the data:
  - Satterthwaite approximation
  - Kenward and Roger approximation
  - ...
- In the context of longitudinal data, all methods typically lead to large numbers of degrees of freedom, and therefore also to very similar p-values.
- For univariate hypotheses (rank($\boldsymbol{L}$) $= 1$) the F-test reduces to a t-test

Testing fixed effects in LMM

> Perhaps I can try again to explain why I don't quote p-values or, more to the point, why I do not take the "obviously correct" approach of attempting to reproduce the results provided by SAS. Let me just say that, although there are those who feel that the purpose of the R Project - indeed the purpose of any statistical computing whatsoever - is to reproduce the p-values provided by SAS, I am not a member of that group.
>
> Douglas Bates at [R] lmer, p-values and all that
>
> https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Testing fixed effects in LMM

Most of the research on tests for the fixed-effects specification in a mixed model begin with the **assumption** that these statistics will have an F distribution with a known numerator degrees of freedom and the only purpose of the research is to decide how to obtain an approximate denominator degrees of freedom. I don't agree.

Douglas Bates at [R] lmer, p-values and all that

https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Testing fixed effects in LMM

- In general it is **not** clear that the null distribution of the computed ratio of sums of squares is really an F distribution, for any choice of denominator degrees of freedom.
- When the responses are normally distributed and the design is balanced, nested etc. (i.e. the classical LMM situation), the scaled deviances and differences in deviances are exactly F-distributed and looking at the experimental design (i.e., which treatments vary/are replicated at which levels) tells us what the relevant degrees of freedom are.

Testing fixed effects in LMM

- When the data are not classical (crossed, unbalanced), we might still **assume** that the deviances are approximately F-distributed but that we don't know the real degrees of freedom. This is what the Satterthwaite, Kenward-Roger, Fai-Cornelius, among other approximations are supposed to do
- Situation worsens when dealing with discrete responses (binary, Poisson, etc)

Testing the effects in R

```
> ## Calculating confidence intervals for the fixed effects via Wald, bootstrap and profile likelihood
> confint(early.lmer1,par=5:8,method="Wald",oldNames = FALSE) # Only for fixed effects vc will return NA
>
                              2.5 %     97.5 %
(Intercept)              100.915099 107.686389
age0                     -19.951908 -12.559005  ## Significant
program                   -5.476393   3.547128  ## Not significant
age0:program               1.392766  11.244657  ## Significant
>
> confint(early.lmer1,method="boot",boot.type ="perc",oldNames = FALSE,nsim=500)
>
                               2.5 %       97.5 %
sd_(Intercept)|id          6.9406669  11.17900578
cor_age0.(Intercept)|id   -1.0000000  -0.06533987
sd_age0|id                 0.7563938   9.45484576
sigma                      6.7327885   8.78499590
(Intercept)              100.5754354 108.09445268
age0                     -20.3249215 -12.16307504  ## Significant
program                   -5.4742982   4.03860091  ## Not significant
age0:program               1.8325498  11.24816569  ## Significant
>
> confint(early.lmer1, level = 0.95,method="profile",oldNames = FALSE)
>
                              2.5 %     97.5 %
sd_(Intercept)|id          7.009249  11.406182
cor_age0.(Intercept)|id          -1         1
sd_age0|id                 0.000000   9.975352
sigma                      6.814978   8.953279
(Intercept)              100.883287 107.718200
age0                     -19.986640 -12.524273  ## Significant
program                   -5.518786   3.589521  ## Not significant
age0:program               1.346481  11.290942  ## Significant
>
```

Getting p-values

```
> ## Get the KR-approximated degrees of freedom
>
> require(pbkrtest)
> early.lmer1.df.KR <- get_Lb_ddf(early.lmer1, fixef(early.lmer1))
>
> ## Get p-values from the t-distribution using the t-values and approximated
> ## degrees of freedom
>
> early.lmer1.coef=coef(summary(early.lmer1))
> early.lmer1.p.KR <- cbind(early.lmer1.coef,df=early.lmer1.df.KR, 2 * (1 - pt(abs(early.lmer1.coef[,3]),
                    early.lmer1.df.KR)))
> early.lmer1.p.KR
               Estimate Std. Error    t value       df
(Intercept)  104.3007437   1.727403 60.3800996 105.2575 0.000000e+00
age0         -16.2554565   1.885981 -8.6190984 105.2575 7.416290e-14
program       -0.9646326   2.301962 -0.4190479 105.2575 6.760350e-01
age0:program   6.3187112   2.513286  2.5141232 105.2575 1.344683e-02
>
```


Getting p-values with lmerTest

```
## Another way to get the p-values require(lmerTest) and refit the model
>
> require(lmerTest)
> early.lmer1<-lmer(cog~1+age0*program+(1 + age0|id), REML = FALSE, data=early.int1)
> summary(early.lmer1)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: cog ~ 1 + age0 * program + (1 + age0 | id)
   Data: early.int1

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 84.02    9.166
          age0        39.44    6.281    -0.55
 Residual             60.31    7.766
Number of obs: 309, groups:  id, 103

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  104.3007     1.7274 102.9999  60.380  < 2e-16 ***
age0         -16.2555     1.8860 103.0001  -8.619 8.55e-14 ***
program       -0.9646     2.3020 102.9999  -0.419   0.6761
age0:program   6.3187     2.5133 103.0001   2.514   0.0135 *
---
>
```

lmerTest: anova function

```
## Another way to get the p-values require(lmerTest) and refit the model
>
> require(lmerTest)
>
> ## Type III analysis the same as the one obtained with the summary function
>
> anova(early.lmer1)
>
Type III Analysis of Variance Table with Satterthwaite's method
            Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
age0        4480.2  4480.2     1   103 74.2889 8.547e-14 ***
program       10.6    10.6     1   103  0.1756   0.67605
age0:program 381.2   381.2     1   103  6.3208   0.01348 *
---
>
> ## Type I sequential model building
>
> anova(early.lmer1, type=1)
>
Type I Analysis of Variance Table with Satterthwaite's method
            Sum Sq Mean Sq NumDF DenDF  F value  Pr(>F)
age0        6256.8  6256.8     1   103 103.7471 < 2e-16 ***
program       89.7    89.7     1   103   1.4869 0.22549
age0:program 381.2   381.2     1   103   6.3208 0.01348 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```


Getting p-values with lme

```
> ## Fitting the model with lme
>
> require(nlme)
> early.lme1<-lme(cog~1+age0*program, random=~1+age0|id, method = "ML", data=early.int1)
> summary(early.lme1)

Linear mixed-effects model fit by maximum likelihood
  Data: early.int1
       AIC      BIC    logLik
  2332.532 2362.398 -1158.266

Random effects:
 Formula: ~1 + age0 | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 9.166182 (Intr)
age0        6.280528 -0.554
Residual    7.765843

Fixed effects:  cog ~ 1 + age0 * program
                Value Std.Error  DF  t-value p-value
(Intercept)  104.30074  1.738689 204 59.98815  0.0000
age0         -16.25546  1.898308 204 -8.56313  0.0000
program       -0.96463  2.317003 101 -0.41633  0.6781
age0:program   6.31871  2.529714 204  2.49780  0.0133

Number of Observations: 309
Number of Groups: 103
>
```

Testing the effects in R with lme

```
> ## CI lme
>
> intervals(early.lme1)

Approximate 95% confidence intervals

 Fixed effects:
                 lower        est.      upper
(Intercept)  100.894899 104.3007437 107.706589
age0         -19.973972 -16.2554565 -12.536941
program       -5.531097  -0.9646326   3.601831
age0:program   1.363362   6.3187112  11.274060

 Random Effects:
  Level: id
                           lower       est.     upper
sd((Intercept))        7.2611304  9.1661817 11.571048
sd(age0)               3.1375541  6.2805285 12.571907
cor((Intercept),age0) -0.7944973 -0.5540066 -0.163302

 Within-group standard error:
   lower     est.    upper
6.774495 7.765843 8.902259
>
```


Likelihood ratio test

$$H_0 : \boldsymbol{\beta} \in \Theta_0 \quad H_1 : \boldsymbol{\beta} \in \Theta_0^C$$

- Notation:
  - L_{ML}: Likelihood function
  - $\widehat{\boldsymbol{\beta}}_{ML0}$: Point in Θ_0 that maximizes L_{ML}
  - $\widehat{\boldsymbol{\beta}}_{ML}$: Point in $\Theta = \Theta_0 \cup \Theta_0^C$ that maximizes L_{ML}
- Test statistic:

$$-2 \ln \lambda_N = -2 \ln \left[\frac{L_{ML}(\widehat{\boldsymbol{\beta}}_{ML0})}{L_{ML}(\widehat{\boldsymbol{\beta}}_{ML})} \right] \overset{H_0}{\rightarrow} \chi^2(df)$$

- $df = \dim(\Theta) - \dim(\Theta_0)$.

Early dietary intervention study

Hierarchical model

$$\begin{cases} Y_{ij} = & \pi_{0i} + \pi_{1i}(Age_{ij} - 1) + \varepsilon_{ij} \\ \pi_{0i} = & \gamma_{00} + \gamma_{01} PROG_i + b_{0i} \\ \pi_{1i} = & \gamma_{10} + \gamma_{11} PROG_i + b_{1i} \end{cases}$$

Three models considered for the second level

- No effect of program $\gamma_{01} = \gamma_{11} = 0$ (early.lmer1.noprog)
- Program has an effect only on the intercept $\gamma_{11} = 0$ (early.lmer1.intprog)
- Program has an effect on both intercept and slope (early.lmer1)

Likelihood ratio tests in R

```
> ## Likelihood ratio tests
> early.lmer1.noprog<-lmer(cog~1+age0+(1 + age0|id), REML = FALSE, data=early.int1)
> early.lmer1.intprog<-lmer(cog~1+age0+program+(1 + age0|id), REML = FALSE, data=early.int1)
> anova(early.lmer1.noprog,early.lmer1.intprog,early.lmer1)
>
Data: early.int1
Models:
early.lmer1.noprog: cog ~ 1 + age0 + (1 + age0 | id)
early.lmer1.intprog: cog ~ 1 + age0 + program + (1 + age0 | id)
early.lmer1: cog ~ 1 + age0 * program + (1 + age0 | id)
                  Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
early.lmer1.noprog   6 2336.8 2359.2 -1162.4   2324.8
early.lmer1.intprog  7 2336.7 2362.8 -1161.3   2322.7 2.0840      1    0.14885
early.lmer1          8 2332.5 2362.4 -1158.3   2316.5 6.1345      1    0.01326 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Effect of early dietary intervention on children IQ

Conclusions level 2 model

$$\begin{cases} \hat{\pi}_{0i} = & 103.758 + b_{0i} \\ \hat{\pi}_{1i} = & -15.882 + 5.656 PROG_i + b_{1i} \end{cases}$$

$\Rightarrow$ Children in the intervention and control group have the same average initial scores. Expected?

$\Rightarrow$ The average cognitive performance decreased in both groups but less in the intervention group.

Assessing the random effects

- Empirical Bayes inference
- Best linear unbiased prediction
- Example: Early dietary intervention
- Shrinkage
- Example: Early dietary intervention
- A theoretical illustration

Assessing the random effects

- Recall that the general linear mixed model equals

$$\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(\boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i, \Sigma_i), \quad \boldsymbol{b}_i \sim N(0, \boldsymbol{D})$$

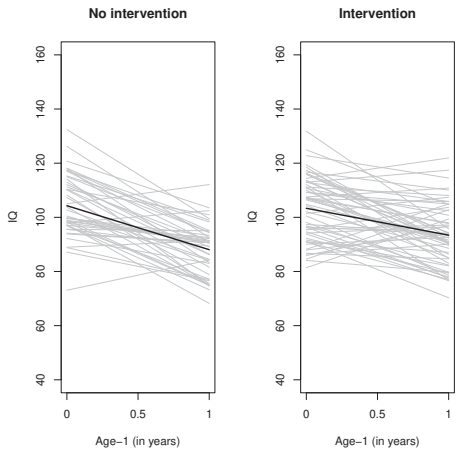
- Marginally,

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\beta, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \Sigma_i)$$

- Thus, random effects $\boldsymbol{b}_i$ reflect how the evolution of the ith subject deviates from the expected evolution $\boldsymbol{X}_i\beta$, i.e., how the evolution of the ith subject deviates from the average evolution in the population

- Estimation of $\boldsymbol{b}_i$ helpful for detecting outlying profiles or predicting individual trajectories

Assessing the random effects



$\Rightarrow$ $\boldsymbol{b}_i$ reflect how the evolution for the ith subject deviates from the average

$\Rightarrow$ Some subjects are above/below the average at the beginning of the study

$\Rightarrow$ The evolution of individual subjects differs from the average evolution

Assessing the random effects

- The term "estimates" of the random effects is some times used in the literature
- Random effects are not, strictly speaking, parameters but unobserved random variables
- One does not estimate the random effects in the same sense that one estimates parameters
- $f(\boldsymbol{b}_i) = N(0, \boldsymbol{D})$ can be interpreted as the prior distribution of $\boldsymbol{b}_i$, i.e., its distribution before the data are collected
- Hence, it is natural to base the prediction of $\boldsymbol{b}_i$ on the posterior distribution $f(\boldsymbol{b}_i|\boldsymbol{Y}_i)$ using Bayesian methods

Assessing the random effects

- Applying Bayes theorem the posterior density of $\boldsymbol{b}_i$ is

$$f(\boldsymbol{b}_i|\boldsymbol{Y}_i) = \frac{f(\boldsymbol{Y}_i|\boldsymbol{b}_i)\ f(\boldsymbol{b}_i)}{\int f(\boldsymbol{Y}_i|\boldsymbol{b}_i)\ f(\boldsymbol{b}_i)\, d\boldsymbol{b}_i} \propto f(\boldsymbol{Y}_i|\boldsymbol{b}_i)\ f(\boldsymbol{b}_i)$$

$$\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_i - \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{W}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)\right)' \Lambda_i^{-1} \left(\boldsymbol{b}_i - \boldsymbol{D}Z_i'\boldsymbol{W}_i(\boldsymbol{Y}_i - X_i\beta)\right)\right\}$$

for some positive definite matrix Λ_i

- Posterior distribution:

$$\boldsymbol{b}_i|\boldsymbol{Y}_i \sim N\left(\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{W}_i(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta), \Lambda_i\right)$$

Assessing the random effects

- Posterior mean $E[\boldsymbol{b}_i|\boldsymbol{Y}_i]$ used to predict $\boldsymbol{b}_i$

$$\widehat{\boldsymbol{b}}_i(\theta) = E[\boldsymbol{b}_i|\boldsymbol{Y}_i] = \int \boldsymbol{b}_i f(\boldsymbol{b}_i|\boldsymbol{Y}_i)\, d\boldsymbol{b}_i = \boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{W}_i(\alpha)(\boldsymbol{Y}_i - \boldsymbol{X}_i\beta)$$

- $\widehat{\boldsymbol{b}}_i(\theta)$ is normally distributed with covariance matrix

$$\text{var}(\widehat{\boldsymbol{b}}_i(\theta)) = \boldsymbol{D}\boldsymbol{Z}_i'\left\{\boldsymbol{W}_i - \boldsymbol{W}_i\boldsymbol{X}_i\left(\sum_{i=1}^{N}\boldsymbol{X}_i'\boldsymbol{W}_i\boldsymbol{X}_i\right)^{-1}\boldsymbol{X}_i'\boldsymbol{W}_i\right\}\boldsymbol{Z}_i\boldsymbol{D}$$

- Inferences for $\boldsymbol{b}_i$ should account for the variability in $\widehat{\boldsymbol{b}}_i$

Assessing the random effects

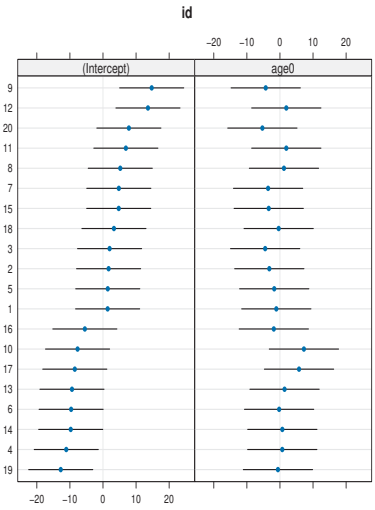
- Parameters in θ are replaced by their ML or REML estimates, obtained from fitting the marginal model

- $\widehat{\boldsymbol{b}}_i = \widehat{\boldsymbol{b}}_i(\widehat{\theta})$ is called the Empirical Bayes estimate/prediction of $\boldsymbol{b}_i$

R Output

```
> ## We are going to work with  the full model for illustration
> summary(early.lmer1)

Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: cog ~ 1 + age0 * program + (1 + age0 | id)
   Data: early.int1

     AIC      BIC   logLik deviance df.resid
  2332.5   2362.4  -1158.3   2316.5      301

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.25362 -0.59088  0.02131  0.56850  2.29366

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 84.02    9.166
          age0        39.44    6.281    -0.55
 Residual             60.31    7.766
Number of obs: 309, groups:  id, 103

Fixed effects:
             Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  104.3007     1.7274 102.9999  60.380  < 2e-16 ***
age0         -16.2555     1.8860 103.0001  -8.619 8.55e-14 ***
program       -0.9646     2.3020 102.9999  -0.419   0.6761
age0:program   6.3187     2.5133 103.0001   2.514   0.0135 *
---
```


Effect of early dietary intervention on children IQ




```
## Plotting the random intercept and slope for a subset of
## the data. In this case the first 20 individuals given
## in s

r.int=ranef(early.lmer1, condVar=TRUE)
s=1:20

r.int=lapply(r.int, function(x) {

s2=which(rownames(x) %in% s)
  x=x[s2, ]

attributes(x)$postVar=attributes(x)$postVar[, , s2]

return(x)
})

class(r.int)="ranef.mer"
dotplot(r.int)
```

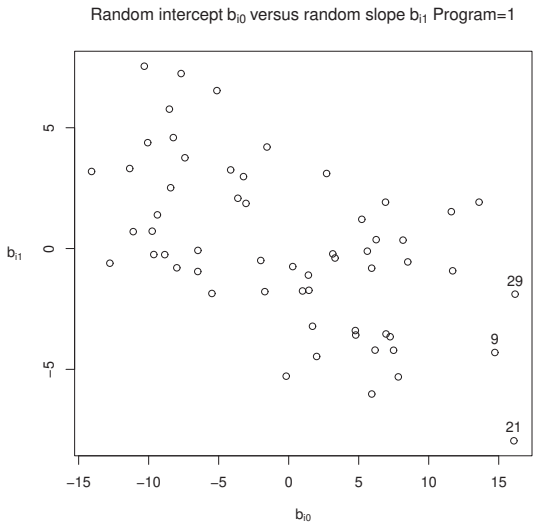
R Output

```
> ## Random effects covariance matrix
>
> D.early=unclass(VarCorr(early.lmer1))$id
> D.early
>
            (Intercept)      age0
(Intercept)    84.01915 -31.89344
age0          -31.89344  39.44432
attr(,"stddev")
(Intercept)        age0
   9.166196    6.280471
>
attr(,"correlation")
            (Intercept)       age0
(Intercept)   1.0000000 -0.5540131
age0         -0.5540131  1.0000000
>
```


Predicting the random effects in R

```
> ## Predicted random effects
> early.lmer1.re=ranef(early.lmer1)$id
>
> head(early.lmer1.re,10)
>
   (Intercept)       age0
1     1.406077 -1.0998359
2     1.700796 -3.2167090
3     1.996373 -4.4674210
4   -11.103346  0.6994383
5     1.444727 -1.7280502
6    -9.633042 -0.2478932
7     4.787545 -3.5797912
8     5.221731  1.2096573
9    14.723746 -4.3027959
10   -7.682804  7.2427588
>
> plot(early.lmer1.re[1:58,],
+ main="Random intercept (b0i) versus random slope (b1i).Program=1")
>
```

Random intercept ($\widehat{b}_{0i}$) versus random slope ($\widehat{b}_{1i}$)



Best Linear Unbiased Prediction (BLUP)

- Often, parameters of interest are linear combinations of fixed effects in $\boldsymbol{\beta}$ and random effects in $\boldsymbol{b}_i$
- For example, a subject-specific slope is the sum of the average slope and the subject-specific random slope. In the case study a child that did not receive the intervention has slope

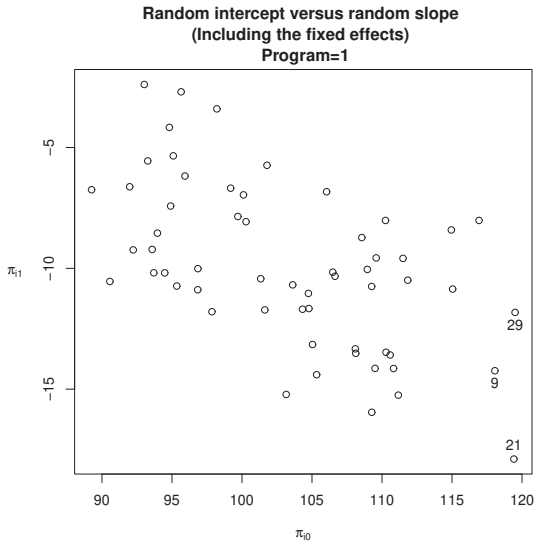
$$\pi_{1i} = \gamma_{10} + b_{1i}$$

- In general, suppose $\boldsymbol{u} = \boldsymbol{\lambda}'_{\beta}\boldsymbol{\beta} + \boldsymbol{\lambda}'_{b}\boldsymbol{b}_i$ is of interest
- Conditionally on $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{u}} = \boldsymbol{\lambda}'_{\beta}\widehat{\boldsymbol{\beta}} + \boldsymbol{\lambda}'_{b}\widehat{\boldsymbol{b}}_i$ is BLUP:
  - Linear in the observations $\boldsymbol{Y}_i$
  - Unbiased for $\boldsymbol{u}$
  - Minimum variance among all unbiased linear estimators

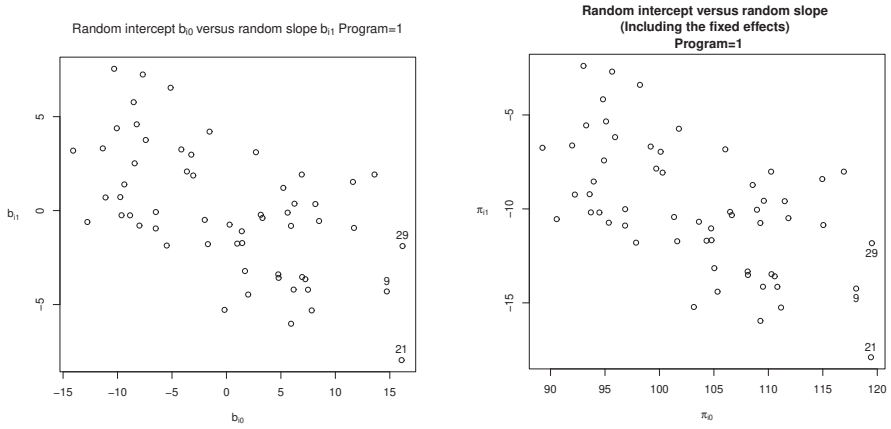
Intercept and slope: OLS versus LMM estimates

```
> ## Creating the subject specific intercepts and slopes
> ## Here we have to use the model fitted with lme
>
> ind.coef=coef(early.lme1)
> head(ind.coef)
  (Intercept)      age0    program age0:program
1   105.70683 -17.35531 -0.9646326     6.318711
2   106.00156 -19.47223 -0.9646326     6.318711
3   106.29715 -20.72297 -0.9646326     6.318711
4    93.19743 -15.55611 -0.9646326     6.318711
5   105.74548 -17.98354 -0.9646326     6.318711
6    94.66773 -16.50345 -0.9646326     6.318711
>
> prog=early.int1[early.int1$age0==0,]$program
> int.subject=ind.coef[,1]+ind.coef[,3]*prog
> slope.subject=ind.coef[,2]+ind.coef[,4]*prog
>
> plot(int.subject[1:58],slope.subject[1:58], xlab=expression(pi[i0]), ylab="",
+ main="Random intercept versus random slope (Including the fixed effects)
+ Program=1")
> mtext(expression(pi[i1]),  side = 2, line = 3, las = 1)
>
```

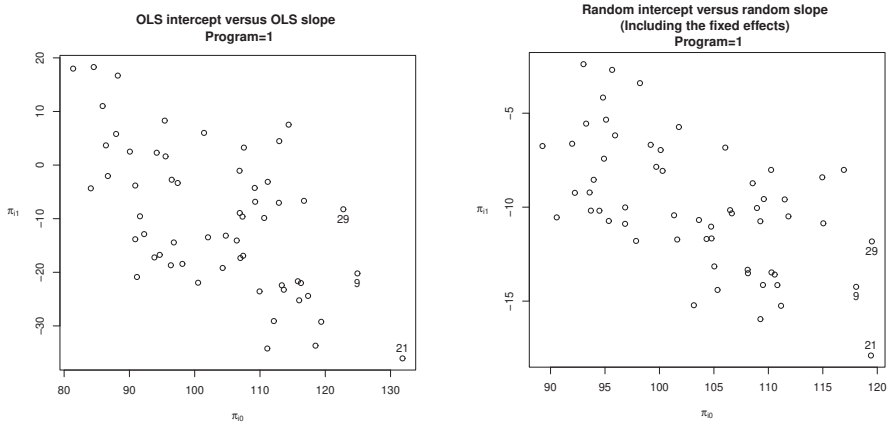

Random intercept ($\widehat{\pi}_{0i}$) versus random slope ($\widehat{\pi}_{1i}$)



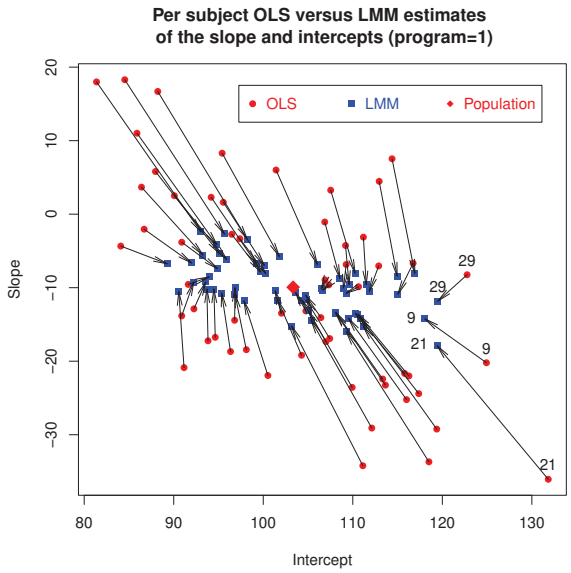
$(\widehat{b}_{0i}, \widehat{b}_{1i})$ versus $(\widehat{\pi}_{0i}, \widehat{\pi}_{1i})$



OLS versus LMM estimates



OLS versus LMM estimates



OLS versus LMM estimates

- In general, the per-subject slopes and intercepts from the mixed-effects model (LMM) are closer to the population estimates than are the within-subject OLS estimates
- This pattern is sometimes described as a shrinkage of coefficients toward the population values
- John Tukey chose to characterize this process in terms of the estimates for individual subjects "borrowing strength" from each other
- In a mixed-effects model we assume that the levels of a grouping factor are a selection from a population and, as a result, can be expected to share characteristics to some degree

Shrinkage Estimators $\widehat{b}_i$

- Consider the prediction of the evolution of the ith subject:

$$\begin{aligned}\widehat{\boldsymbol{Y}}_i &\equiv \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\widehat{\boldsymbol{b}}_i \\ &= \boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}) \\ &= \left(\boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\right)\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{V}_i^{-1}\boldsymbol{Y}_i \\ &= \Sigma_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\widehat{\boldsymbol{\beta}} + \left(\boldsymbol{I}_{n_i} - \Sigma_i\boldsymbol{V}_i^{-1}\right)\boldsymbol{Y}_i\end{aligned}$$

- Let us look more closely at this expression

Shrinkage Estimators $\widehat{b}_i$

$\widehat{\boldsymbol{Y}}_i$ is a weighted mean of two factors

$$\widehat{\boldsymbol{Y}}_i \equiv \boxed{\Sigma_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}} + \boxed{\left(\boldsymbol{I}_{n_i} - \Sigma_i\boldsymbol{V}_i^{-1}\right)\boldsymbol{Y}_i}$$

$\Rightarrow$ Factor 1: Population-averaged profile $\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}$ with weight $\Sigma_i\boldsymbol{V}_i^{-1}$

$\Rightarrow$ Factor 2: Individual data $\boldsymbol{Y}_i$ with weight $\left(\boldsymbol{I}_{n_i} - \Sigma_i\boldsymbol{V}_i^{-1}\right)$

Shrinkage Estimators $\widehat{b}_i$

- Note that the population average $\boldsymbol{X}_i\widehat{\beta}$ gets much weight if the residual variability Σ_i is 'large' in comparison to the total variability $\boldsymbol{V}_i$.
- This phenomenon is usually called shrinkage

> The observed data are shrunk towards the prior average profile $\boldsymbol{X}_i\beta$.

- This is also reflected in the fact that for any linear combination $\lambda'\boldsymbol{b}_i$ of random effects,

$$\text{var}(\lambda'\widehat{\boldsymbol{b}}_i) \leq \text{var}(\lambda'\boldsymbol{b}_i).$$

Models for cluster data

Ariel Alonso Abad

Catholic University of Leuven

Models for Clustered Data

The Rat Pup Example

The data come from a study in which 30 female rats were randomly assigned to receive one of three doses of an experimental compound (variable **treat** with levels: high, low or control). Although 10 female rats were initially assigned to receive each treatment dose, three of the female rats in the high-dose group died, so there are no data for their litters. In addition, litter sizes (variable **lts**) varied widely, ranging from 2 to 18 pups. The sex of the pups was also recorded (variable **sex** taking value zero for males)

Objective of the study: To compare the birth weights (variable **w**) of pups from litters born to female rats that received the high- and low-dose treatments to the birth weights of pups from litters that received the control treatment.

Jose Pinheiro and Doug Bates, (2000) Mixed-Effects Models in S and S-PLUS.

The Rat Pup Example

- Two-level clustered data from a cluster randomized trial
- Each litter (cluster) was randomly assigned to a specific level of treatment
- Rat pups (units of analysis) nested within litters
- Birth weights of rat pups within the same litter are likely to be correlated because the pups shared the same maternal environment

Exploring the data in R

```
> ## Reading the data
> ratpup <- read.table("rat_pup.dat", h = T)
> ratpup$sex1[ratpup$sex == "Female"] <- 1
> ratpup$sex1[ratpup$sex == "Male"] <- 0
> attach(ratpup)
>
> ## Table describing the  data
> g <- function(x)c(N=length(x),Mean=mean(x,na.rm=TRUE),
+ SD=sd(x,na.rm=TRUE), Min=min(x,na.rm=TRUE),Max=max(x,na.rm=TRUE))
> summarize(weight,by=llist(treatment,sex),g)
>
  treatment    sex weight     Mean        SD  Min  Max
1   Control Female     54 6.116111 0.6851179 3.68 7.57
2   Control   Male     77 6.471039 0.7537880 4.57 8.33
3       Low Female     65 5.837538 0.4504964 4.75 7.73
4       Low   Male     61 6.025082 0.3803403 5.25 7.13
5      High Female     32 5.851562 0.6001887 4.48 7.68
6      High   Male     33 5.918485 0.6909058 5.01 7.70
>
```


Exploring the data in R

Treatment	Sex	N obs	Mean	SD	Minimum	Maximum
Control	Female	54.00	6.12	0.69	3.68	7.57
Control	Male	77.00	6.47	0.75	4.57	8.33
Low	Female	65.00	5.84	0.45	4.75	7.73
Low	Male	61.00	6.03	0.38	5.25	7.13
High	Female	32.00	5.85	0.60	4.48	7.68
High	Male	33.00	5.92	0.69	5.01	7.70

- The experimental treatments appear to have a negative effect on mean birth weight for males and females
- Sample mean birth weight of males are consistently higher than those of females within all levels of treatment

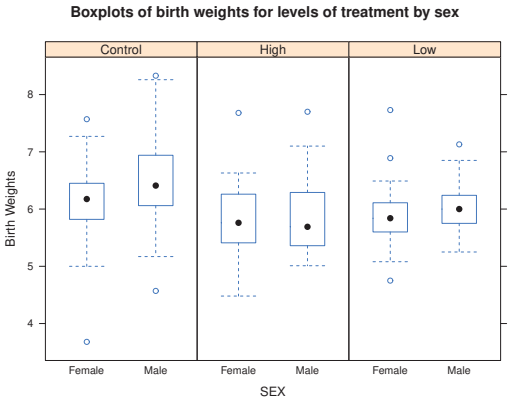
Exploring the data in R

```
> ## Comparing the distributions of birth weights
> ## for each treatment by sex combination
>
> library(lattice)  # trellis graphics
> library(grid)
>
> bwplot(weight ~ sex|treatment, data=ratpup,aspect = 2,
+  ylab="Birth Weights", xlab="SEX",
+  main = "Boxplots of birth weights for levels of treatment by sex")
>
```


Birth weights for levels of treatment by sex

- Males appear to have a higher median birth weight than females in the low and control groups, but not in the high group
- The distribution of birth weight appears to be roughly symmetric at each level of treatment and sex

Birth weights for levels of treatment by sex

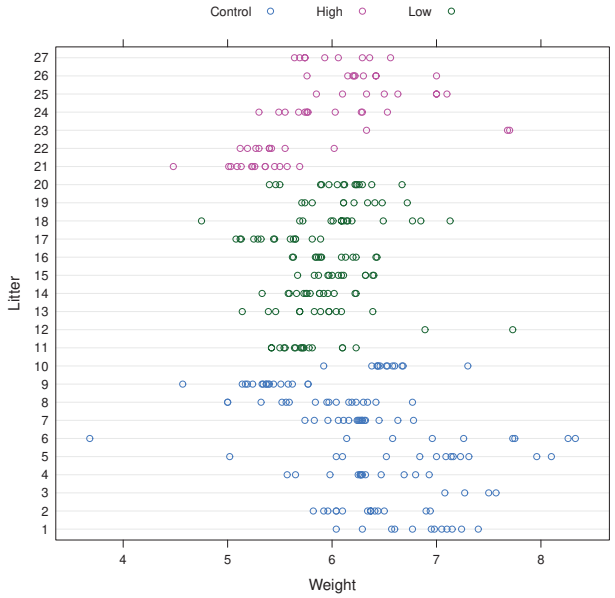


- Lower birth weight for the high- and low-dose treatments compared to the control group
- Variance of the birth weight is similar for males and females within each treatment but appears to differ across treatments

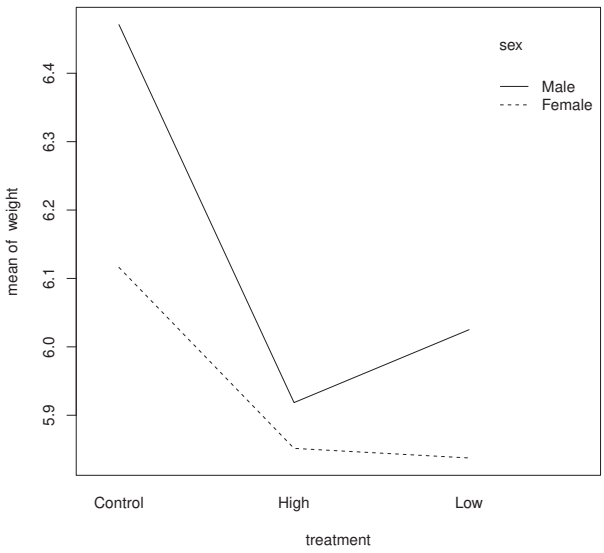
Exploring the data in R

```
> ## Comparing the distributions of birth weights for each treatment
>
> dotplot(litterid ~ weight,group=treatment, data =ratpup,
+ xlab="Weight", ylab="Litter",
+ auto.key=list(space="top", column=3, cex=.8, title="",
+                cex.title=1, lines=FALSE, points=TRUE) )
> with(ratpup, interaction.plot(treatment,sex,weight))
>
```

Exploring the data in R



Exploring the data in R



Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

Level 2:

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} treat_{1i} + \gamma_{02} treat_{2i} + \gamma_{03} ls_i + b_{0i} \\ \pi_{1i} = \gamma_{10} + \gamma_{20} treat_{1i} + \gamma_{30} treat_{2i} \end{cases}$$

where $treat_{1i}$ and $treat_{2i}$ are level 2 indicator variables for high and low treatment levels, ls_i is the litter size and $b_{0i} \sim N\left(0, \sigma^2_b\right)$

$\Rightarrow$ Birth weights of pups vary **within** litter due to differences in gender and in other unaccounted factors (ε_{ij})

Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$$

Level 2:

$$\begin{cases} \pi_{0i} = \gamma_{00} + \gamma_{01} treat_{1i} + \gamma_{02} treat_{2i} + \gamma_{03} ls_i + b_{0i} \\ \pi_{1i} = \gamma_{10} + \gamma_{20} treat_{1i} + \gamma_{30} treat_{2i} \end{cases}$$

where $treat_{1i}$ and $treat_{2i}$ are level 2 indicator variables for high and low treatment levels, ls_i is the litter size and $b_{0i} \sim N\left(0, \sigma^2_b\right)$

$\Rightarrow$ Birth weights vary **between** litters due to differences in treatment, litter size and other litter-specific characteristics unaccounted for by the model (b_{0i})

$\Rightarrow$ Notice that treatment may affect males and females pups differently

One single model

Model

$$\begin{aligned} w_{ij} = \quad & \gamma_{00} + \gamma_{01} treat_{1i} + \gamma_{02} treat_{2i} + \gamma_{03} ls_i + \\ & \gamma_{10} sex_{ij} + \gamma_{20} treat_{1i} sex_{ij} + \gamma_{30} treat_{2i} sex_{ij} + \\ & b_{0i} + \varepsilon_{ij} \end{aligned}$$

Distributional Assumptions

$$b_{0i} \sim N\left(0, \sigma_b^2\right) \text{ and } \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

Fitting the homocedastic model in R

Model 1

```
> ## Fitting the model
>
> library(nlme)
>
> meanfull.hom <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                     random = ~1 | litterid, ratpup, method = "REML")
>
```

- The factor() function is not necessary for treatment, because the original treatment variable has string values High, Low, and Control, and will therefore be considered as a factor automatically
- We also do not need to declare sex1 as a factor, because it is an indicator variable having only values of 0 and 1

Fitting the homocedastic model in R

Model 1

```
> ## Fitting the model
>
> library(nlme)
>
> meanfull.hom <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                     random = ~1 | litterid, ratpup, method = "REML")
>
```

- lme() treats the lowest level (alphabetically or numerically) of a factor as the reference category. This means that "Control" will be the reference category of treatment. The reference level can be changed using

  `treatment=relevel(treatment,ref="High")`

Fitting the homocedastic model in R

Model 1

```
> ## Fitting the model
>
> library(nlme)
>
> meanfull.hom <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                     random = ~1 | litterid, ratpup, method = "REML")
>
```

- `random = 1 | litterid`, includes a random effect (intercept) for each level of litter in the model
- `method = "REML"`, specifies that the default REML estimation method is to be used

Fitting the homocedastic model in R

```
> summary(meanfull.hom)
>
Linear mixed-effects model fit by REML
 Data: ratpup
       AIC      BIC    logLik
  419.1043 452.8775 -200.5522

Random effects:
 Formula: ~1 | litterid
        (Intercept) Residual
StdDev:   0.3106722 0.404337

Fixed effects: weight ~ treatment + sex1 + litsize + treatment:sex1
                      Value  Std.Error  DF   t-value p-value
(Intercept)         8.323340 0.27333009 292 30.451605  0.0000
treatmentHigh      -0.906057 0.19154238  23 -4.730320  0.0001
treatmentLow       -0.467040 0.15818328  23 -2.952521  0.0071
sex1               -0.411688 0.07315410 292 -5.627679  0.0000
litsize            -0.128382 0.01875336  23 -6.845819  0.0000
treatmentHigh:sex1  0.107023 0.13176318 292  0.812239  0.4173
treatmentLow:sex1   0.083866 0.10568189 292  0.793568  0.4281

.............................................................

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-7.47250744 -0.50014749  0.02911668  0.57348178  3.00962055

Number of Observations: 322
Number of Groups: 27
>
```


Fitting the homocedastic model in R

```
> anova(meanfull.hom)
>
                numDF denDF  F-value p-value
(Intercept)         1   292 9093.772  <.0001
treatment           2    23    5.082  0.0149
sex1                1   292   52.602  <.0001
litsize             1    23   47.374  <.0001
treatment:sex1      2   292    0.466  0.6282
>
```

- The anova() function performs a series of Type I (or sequential) F-tests for the fixed effects in the model, each of which are conditional on the preceding terms in the model specification
- For example, the F-test for sex1 is conditional on the treatment effects, but the F-test for treatment is not conditional on the sex1 effect

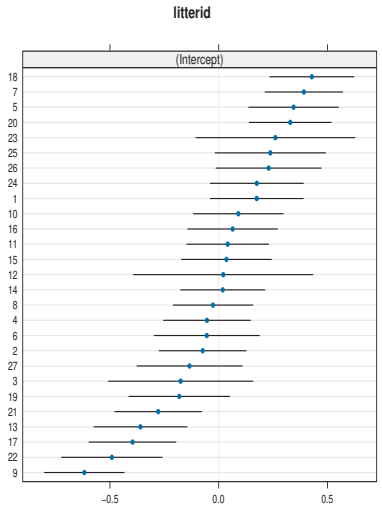
Fitting the homocedastic model in R

```
> anova(meanfull.hom)
>
                numDF denDF  F-value p-value
(Intercept)         1   292 9093.772  <.0001
treatment           2    23    5.082  0.0149
sex1                1   292   52.602  <.0001
litsize             1    23   47.374  <.0001
treatment:sex1      2   292    0.466  0.6282
>
```

Model fitted using REML

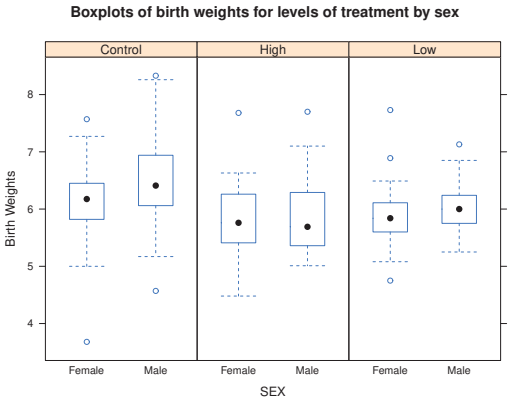
The model was fitted using REML and, therefore, different mean structures cannot be compare!

Effect of early dietary intervention on children IQ




```
> ## Display the random effects (EBLUPs) from the model.
>
> random.effects(meanfull.hom)
>
   (Intercept)       (Intercept)
1   0.17480024    17 -0.39636862
2  -0.07362296    18  0.42802095
3  -0.17490203    19 -0.18110865
4  -0.05376249    20  0.32903707
5   0.34446954    21 -0.27813901
6  -0.05480208    22 -0.49096620
7   0.39153638    23  0.26053476
8  -0.02616704    24  0.17537803
9  -0.61772106    25  0.23748827
10  0.09017150    26  0.22966911
11  0.04136696    27 -0.13396497
12  0.02072931
13 -0.35981737
14  0.01847368
15  0.03549783
16  0.06416884
>
```

Modeling the covariance structure



- Previous model assumes that the within litter variability σ^2_ε is constant across treatment
- The variances of the birth weights are similar for males and females within each treatment but appear to differ across treatments

Covariance structure: Testing homoscedasticity

Hence, one wants to test if the variance of the residuals (σ^2_ε) is the same (homogeneous) for the three treatment groups (high, low, and control)

$$H_0 : \sigma^2_{high} = \sigma^2_{low} = \sigma^2_{control} = \sigma^2_\varepsilon$$

- REML-based likelihood ratio test to compare two models (mean structure stays the same):

  Model 1: All three variances equal (meanfull.hom)

  Model 2: All three variances different (meanfull.het)

- The asymptotic null distribution of this test statistic is a χ^2 with 2 degrees of freedom

Covariance structure: Testing homoscedasticity

- At this moment the lmer() function does not allow users to fit models with heterogeneous error variance structures
- Therefore, we will work with the function lme() from the package nlme
- lme() and lmer() are similar but there are some differences in syntax and output that will be explained in the following

Fitting the heterocedastic model in R

Model 2

```
> ## Fitting a heterocedastic model
>
> meanfull.het <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                     random = ~1 | litterid, ratpup, method = "REML",
+                     weights = varIdent(form = ~1 | treatment))
>
```

- The arguments of the lme() function are the same as those used to fit Model 1, with the addition of the weights argument
- The argument

  `weights = varIdent(form = ~ 1 |treatment)`

  sets up a heterogeneous residual variance structure, with observations at different levels of treatment having different residual variance parameters

Fitting the heterocedastic model in R

```
> summary(meanfull.het)
>
Linear mixed-effects model fit by REML
 Data: ratpup
       AIC     BIC    logLik
  381.8847 423.163 -179.9423

Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.3134846 0.5147948

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | treatment
 Parameter estimates:
  Control       Low      High
1.0000000 0.5649830 0.6394383

Fixed effects: weight ~ treatment + sex1 + litsize + treatment:sex1
                      Value  Std.Error  DF   t-value p-value
(Intercept)         8.345294 0.27464753 292 30.385468  0.0000
treatmentHigh      -0.903277 0.19215903  23 -4.700672  0.0001
treatmentLow       -0.466292 0.15908908  23 -2.931013  0.0075
sex1               -0.408131 0.09303486 292 -4.386865  0.0000
litsize            -0.130007 0.01848708  23 -7.032332  0.0000
treatmentHigh:sex1  0.094666 0.12919527 292  0.732737  0.4643
treatmentLow:sex1   0.076013 0.10811858 292  0.703053  0.4826

.............................................................
```


Fitting the heterocedastic model in R

```
Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.3134846 0.5147948

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | treatment
 Parameter estimates:
  Control       Low      High
1.0000000 0.5649830 0.6394383
```

- Random effects portion of the output: Estimated residual standard deviation equal to 0.5147948
- Parameter estimates: Values by which the residual standard deviation should be multiplied to obtain the estimated standard deviation of the residuals in each treatment group
- This multiplier is 1.0 for the control group (the reference). Multipliers for the low and high treatment groups are very similar

Heterocedastic versus homocedastic model

The variance of the residuals (σ^2_ε) is the same (homogeneous) for the three treatment groups

$$H_0 : \sigma^2_{high} = \sigma^2_{low} = \sigma^2_{control} = \sigma^2_\varepsilon$$

```
> ## Heterocedastic versus homocedastic model
>
> anova(meanfull.hom,meanfull.het)
>
             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
meanfull.hom     1  9 419.1043 452.8775 -200.5522
meanfull.het     2 11 381.8847 423.1630 -179.9423 1 vs 2 41.21964  <.0001
>
```


Heterocedastic model

```
Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.3134846 0.5147948

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | treatment
 Parameter estimates:
  Control       Low      High
1.0000000 0.5649830 0.6394383
```

- $\sigma_{high} = 0.5147948 \cdot 0.6394383$, $\sigma_{low} = 0.5147948 \cdot 0.5649830$ and $\sigma_{control} = 0.5147948 \cdot 1$

Heterocedastic model

```
Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.3134846 0.5147948

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | treatment
 Parameter estimates:
  Control       Low      High
1.0000000 0.5649830 0.6394383
```

- $\sigma_{high} = 0.329179$, $\sigma_{low} = 0.290850$ and $\sigma_{control} = 0.5147948$
- Is $\sigma^2_{high} = \sigma^2_{low}$?

High-low dose: Equal residual variance

Hence, one wants to test if the variance of the residuals in the high and low dose groups are the same

$$H_0 : \sigma^2_{high} = \sigma^2_{low}$$

- REML-based likelihood ratio test to compare two models (mean structure stays the same):

  Model 2: All three variances different (meanfull.het)

  Model 3: $\sigma^2_{high} = \sigma^2_{low}$ (meanfull.hilo)

- The asymptotic null distribution of this test statistic is a χ^2 with 1 degrees of freedom

High-low dose: Equal residual variance

```
> ## High-low dose: Equal residual variance
>
> ratpup$trtgrp[treatment=="Control"] <- 1
> ratpup$trtgrp[treatment == "Low" | treatment == "High"] <- 2
>
> meanfull.hilo <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                      random = ~1 | litterid, ratpup, method = "REML",
+                      weights = varIdent(form = ~1 | trtgrp))
>
> summary(meanfull.hilo)
> anova(meanfull.hilo)
>
```


Fitting the heterocedastic model in R

```
> summary(meanfull.hilo)
>
Linear mixed-effects model fit by REML
 Data: ratpup
       AIC      BIC    logLik
  381.0807 418.6065 -180.5404

Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.3145679 0.5147878

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | trtgrp
 Parameter estimates:
         1         2
1.0000000 0.5905487

Fixed effects: weight ~ treatment + sex1 + litsize + treatment:sex1
                     Value  Std.Error  DF   t-value p-value
(Intercept)         8.350351 0.27567833 292 30.290196  0.0000
treatmentHigh      -0.901844 0.19140146  23 -4.711793  0.0001
treatmentLow       -0.466596 0.15999337  23 -2.916347  0.0078
sex1               -0.408195 0.09303540 292 -4.387529  0.0000
litsize            -0.130383 0.01856367  23 -7.023574  0.0000
treatmentHigh:sex1  0.092026 0.12461723 292  0.738473  0.4608
treatmentLow:sex1   0.076397 0.10939797 292  0.698337  0.4855
.............................................................
>
```

High-low dose: Equal residual variance

Hence, one wants to test if the variance of the residuals in the high and low dose groups are the same

$$H_0 : \sigma^2_{high} = \sigma^2_{low}$$

```
> ## High-low dose: Equal residual variance
>
> anova(meanfull.het,meanfull.hilo)
>
              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
meanfull.het      1 11 381.8847 423.1630 -179.9423
meanfull.hilo     2 10 381.0807 418.6065 -180.5404 1 vs 2 1.196053  0.2741
>
```


Is there a litter effect?

- Can the random effects (b_{0i}) associated with the litter–specific intercepts be omitted from Model 3?
- One do not directly test the significance of the random litter–specific intercepts, but rather tests a hypothesis related to the variance of the random litter effects.
- The null and alternative hypotheses can be written as follows:

$$H_0 : \sigma^2_b = 0 \text{ versus } H_1 : \sigma^2_b > 0$$

Is there a litter effect?

- Although hypothesis tests are often phrased in terms of parameter restrictions, they basically compare the quality of the fit obtained from two nested models
- Likelihood ratio tests (LRTs) are a valuable tool to compare nested models
- An approximate reference distribution for a LRT is the χ^2_γ where γ, the degrees of freedom, is determined by the difference in the number of parameters for the models H_1 and H_0
- Hence, the LRT for testing $H_0 : \sigma^2_b = 0$ versus $H_1 : \sigma^2_b > 0$ has an approximate reference distribution χ^2_1

Is there a litter effect?

- However, the argument for using a χ^2_1 distribution **does not apply** when the parameter value being tested is on the boundary of the parametric space
- The asymptotic null distribution of the test statistic is a mixture of χ^2 distributions, with 0 and 1 degrees of freedom, and equal weights of 0.5
- As shown in Pinheiro and Bates (2000) Section 2.5, the p-value from the χ^2_1 distribution will be "conservative"in the sense that it is larger than a simulation-based p-value would be
- In the worst-case scenario the χ^2_1-based p-value will be twice as large as it should be

Is there a litter effect?

```
> ## Is there a litter effect?
>
> meanfull.hilo.nolitter <- gls(weight ~ treatment + sex1 + litsize +
+         treatment:sex1, data = ratpup, weights = varIdent(form = ~1 | trtgrp))
>
> summary(meanfull.hilo.nolitter)
>
```


Is there a litter effect?

```
Generalized least squares fit by REML
  Model: weight ~ treatment + sex1 + litsize + treatment:sex1
  Data: ratpup
       AIC      BIC   logLik
  489.6521 523.4252 -235.826

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | trtgrp
 Parameter estimates:
       1         2
1.0000000 0.7060188

Coefficients:
                     Value  Std.Error   t-value p-value
(Intercept)          8.201712 0.15902776  51.57409  0.0000
treatmentHigh       -0.976414 0.10624042  -9.19060  0.0000
treatmentLow        -0.456018 0.08700180  -5.24147  0.0000
sex1                -0.339911 0.10616682  -3.20167  0.0015
litsize             -0.121478 0.01008518 -12.04524  0.0000
treatmentHigh:sex1   0.180960 0.14941228   1.21114  0.2267
treatmentLow:sex1    0.076386 0.13035758   0.58597  0.5583

.....................................................

Residual standard error: 0.5980885
Degrees of freedom: 322 total; 315 residual
```

Is there a litter effect?

Is there a litter effect?

$$H_0 : \sigma_b^2 = 0 \text{ versus } H_1 : \sigma_b^2 > 0$$

```
> ## Is there a litter effect?
>
> anova(meanfull.hilo.nolitter,meanfull.hilo)
>
                       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
meanfull.hilo.nolitter     1  9 489.6521 523.4252 -235.8260
meanfull.hilo              2 10 381.0807 418.6065 -180.5404 1 vs 2 110.5713  <.0001
>
```


Is there a litter effect?

```
## Simulation based (exact) restricted likelihood ratio test based on
## simulated values from the finite sample distribution for testing
## whether the variance of a random effect is 0 in a linear mixed model
## with  known correlation structure of the tested random
## effect and i.i.d. errors.
>
> require(RLRsim)
> exactRLRT(meanfull.hilo)

        simulated finite sample distribution of RLRT.

        (p-value based on 10000 simulated values)

data:
RLRT = 129.43, p-value < 2.2e-16

>
```

Modeling the mean structure

```
> ## Fitting the final model using ML
>
> meanfull.hilo.ml <- lme(weight ~ treatment + sex1 + litsize + treatment:sex1,
+                         random = ~1 | litterid, ratpup, method = "ML",
+                         weights = varIdent(form = ~1 | trtgrp))
>
> summary(meanfull.hilo.ml)
```


Modeling the mean structure

```
> summary(meanfull.hilo.ml)
>
Linear mixed-effects model fit by maximum likelihood
 Data: ratpup
       AIC      BIC    logLik
  357.1317 394.8773 -168.5659

Random effects:
 Formula: ~1 | litterid
        (Intercept)  Residual
StdDev:   0.2882595 0.5123784

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | trtgrp
 Parameter estimates:
        1         2
1.0000000 0.5897706
Fixed effects: weight ~ treatment + sex1 + litsize + treatment:sex1
                     Value  Std.Error  DF  t-value p-value
(Intercept)         8.350608 0.26150064 292 31.93341  0.0000
treatmentHigh      -0.904757 0.18092616  23 -5.00070  0.0000
treatmentLow       -0.466869 0.15105108  23 -3.09080  0.0052
sex1               -0.406590 0.09357754 292 -4.34495  0.0000
litsize            -0.130402 0.01755814  23 -7.42689  0.0000
treatmentHigh:sex1  0.093026 0.12521954 292  0.74290  0.4581
treatmentLow:sex1   0.075602 0.10998665 292  0.68737  0.4924
.............................................................

Number of Observations: 322
Number of Groups: 27
```

Modeling the mean structure

```
> anova(meanfull.hilo.ml)
>
               numDF denDF   F-value p-value
(Intercept)        1   292 10274.678  <.0001
treatment          2    23     4.810  0.0180
sex1               1   292    59.906  <.0001
litsize            1    23    55.438  <.0001
treatment:sex1     2   292     0.315  0.7303
>
```


Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim \begin{cases} N(0, 0.51^2), \text{Control} \\ N(0, 0.30^2), \text{Low/High dose} \end{cases}$$

Level 2:

$$\begin{cases} \pi_{0i} = 8.35 - 0.90 treat_{1i} - 0.47 treat_{2i} - 0.13 ls_i + b_{0i} \\ \pi_{1i} = -0.41 \\ b_{0i} \sim N(0, 0.29^2) \end{cases}$$

Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim \begin{cases} N(0, 0.51^2), \text{Control} \\ N(0, 0.30^2), \text{Low/High dose} \end{cases}$$

Level 2:

$$\begin{cases} \pi_{0i} = 8.35 - 0.90 treat_{1i} - 0.47 treat_{2i} - 0.13 ls_i + b_{0i} \\ \pi_{1i} = -0.41 \\ b_{0i} \sim N(0, 0.29^2) \end{cases}$$

$\Rightarrow$ Birth weights of pups vary **within** litter due to differences in gender and in other unaccounted factors (ε_{ij})

Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim \begin{cases} N(0, 0.51^2), \text{Control} \\ N(0, 0.30^2), \text{Low/High dose} \end{cases}$$

Level 2:

$$\begin{cases} \pi_{0i} = 8.35 - 0.90 treat_{1i} - 0.47 treat_{2i} - 0.13 ls_i + b_{0i} \\ \pi_{1i} = -0.41 \\ b_{0i} \sim N(0, 0.29^2) \end{cases}$$

$\Rightarrow$ Litters in the high/low dose have pups with smaller average birth weights. In addition, litter size has a negative impact on the average birth weight and there is extra variability from other unknown factors

Ariel Alonso Multilevel Models 39 / 39

Hierarchical model interpretation

Level 1: ANOVA type model

$$w_{ij} = \pi_{0i} + \pi_{1i} sex_{ij} + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim \begin{cases} N(0, 0.51^2), \text{Control} \\ N(0, 0.30^2), \text{Low/High dose} \end{cases}$$

Level 2:

$$\begin{cases} \pi_{0i} = 8.35 - 0.90 treat_{1i} - 0.47 treat_{2i} - 0.13 ls_i + b_{0i} \\ \pi_{1i} = -0.41 \\ b_{0i} \sim N(0, 0.29^2) \end{cases}$$

$\Rightarrow$ Litters in the high/low dose have pups with smaller average birth weights. In addition, litter size has a negative impact on the average birth weight and there is extra variability from other unknown factors

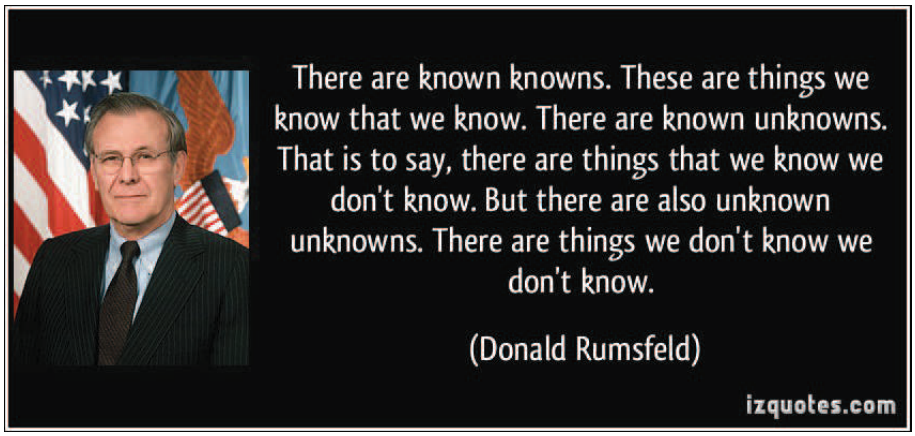
$\Rightarrow$ Treatment affects males and females pups equally

Missing Data: Problems, risks and solutions

Ariel Alonso Abad

Catholic University of Leuven

Missing data: The unknown unknowns



United States Secretary of Defense, Donald Rumsfeld talking about the missing WMD

Missing data problem

- Missing data: Ubiquitous presence in science
- Highly technical/mathematical field
- Only basic definitions, principles and methods
  - Less mathematically involved
  - Easy to implement in standard packages
- Advanced topics
  - Pattern mixture models
  - Selection models
  - Shared parameter models
  - Bodyguard theorem

Titanic

On the 10 April 1912 the largest passenger steamship in the world left Southampton England, to New York City. At 23:40 on 14 April, it struck an iceberg and sank at 2:20 the following morning, resulting in the deaths of 1,517 people in one of the deadliest peacetime maritime disasters in history.

Titanic

On the 10 April 1912 the largest passenger steamship in the world left Southampton England, to New York City. At 23:40 on 14 April, it struck an iceberg and sank at 2:20 the following morning, resulting in the deaths of 1,517 people in one of the deadliest peacetime maritime disasters in history.

Ariel Alonso Missing Data 4 / 73

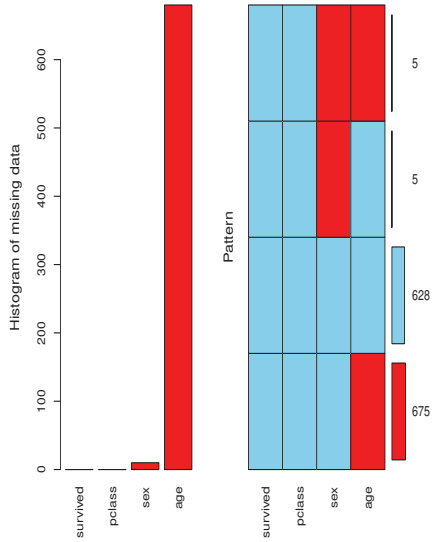
Titanic: Missing data

- Information on 1313 passengers
- Variables:
  - Survival: Y values 1/0.
  - age in years.
  - class: 1st, 2nd, 3rd.
  - sex: 1 male, 0 female.
- Adjusting by age, had class and gender an effect on survival?

Reading the data in R

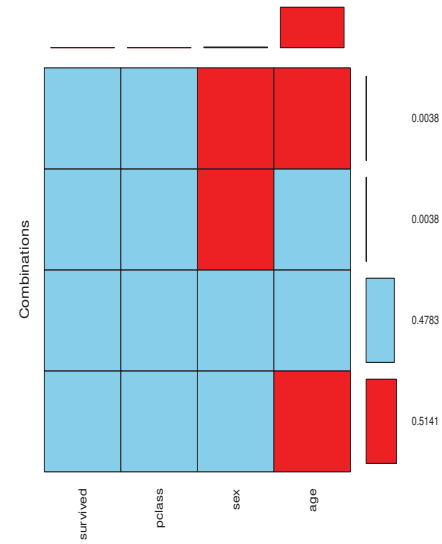
```
> ## Needed libraries
>
> library(mice)
> library(lattice)
> library(VIM)
> library(aod)
> library(BaM)
>
> ## Reading the data
>
> titanic.missing <- read.table("titanicmissing.txt", header=T, sep=",")
> head(titanic.missing,10)
>
   survived pclass sex     age
1         1    1st   0 29.0000
2         0    1st   0  2.0000
3         0    1st   1 30.0000
4         0    1st   0 25.0000
5         1    1st   1  0.9167
6         1    1st   1 47.0000
7         1    1st   0 63.0000
8         0    1st   1 39.0000
9         1    1st   0 58.0000
10        0    1st   1 71.0000
11        0    1st   1 47.0000
12        1    1st   0 19.0000
13        1    1st   0      NA
14        1    1st   1      NA
15        0    1st   1      NA
>
```

Titanic: Missing data



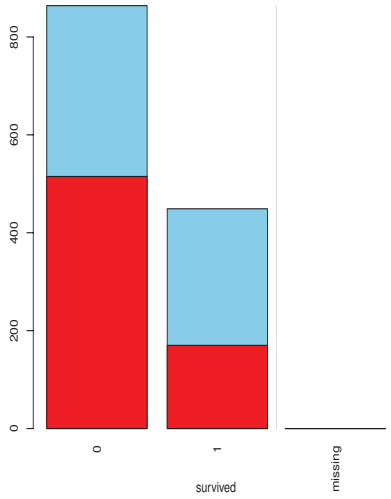

```
> ## Exploring the missingness (VIM library)
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)
>
> ## Amount of missigness in age for each survived group
> barMiss(titanic.missing[,c("survived","age")])
>
> ## Amount of missigness in age for each sex group
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data



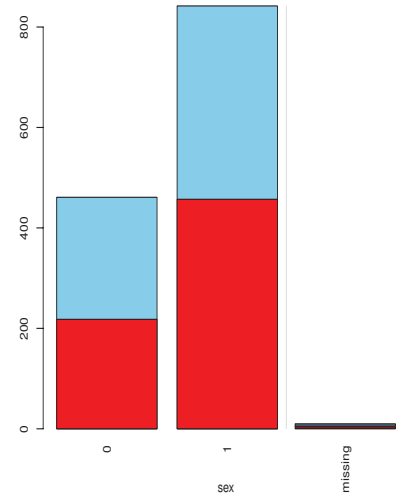

```
> ## Exploring the missingness (VIM library)
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)
>
> ## Amount of missigness in age for each survived group
> barMiss(titanic.missing[,c("survived","age")])
>
> ## Amount of missigness in age for each sex group
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```

Titanic: Missing data



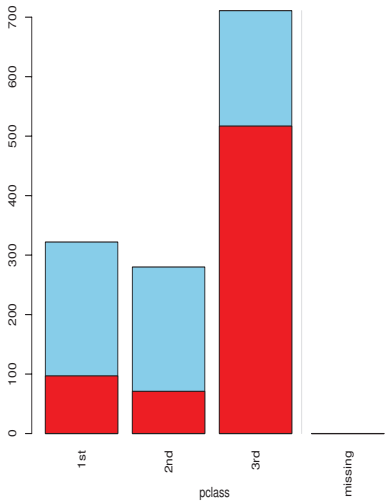

```
> ## Exploring the missingness (VIM library)
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)
>
> ## Amount of missigness in age for each survived group
> barMiss(titanic.missing[,c("survived","age")])
>
> ## Amount of missigness in age for each sex group
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data



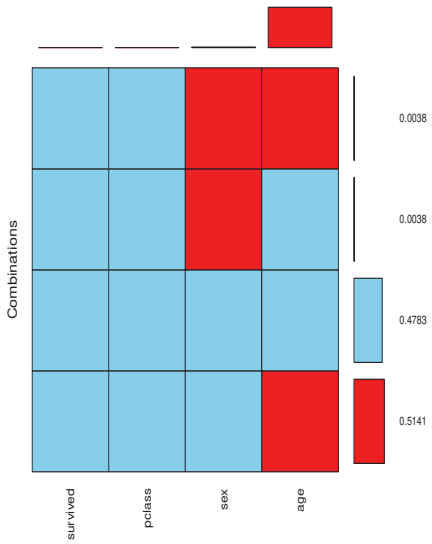

```
> ## Exploring the missingness (VIM library)
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)
>
> ## Amount of missigness in age for each survived group
> barMiss(titanic.missing[,c("survived","age")])
>
> ## Amount of missigness in age for each sex group
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```

Titanic: Missing data




```
> ## Exploring the missingness (VIM library)
>
> titanic.missing.aggr=aggr(titanic.missing,numbers=TRUE,
+ prop=FALSE, ylab=c("Histogram of missing data","Pattern"))

> titanic.missing.aggr
>
 Missings in variables:
 Variable Count
      sex    10
      age   680
>
> aggr(titanic.missing, combined=TRUE, numbers = TRUE,
+ prop = TRUE, cex.numbers=0.87, varheight = FALSE)
>
> ## Amount of missigness in age for each survived group
> barMiss(titanic.missing[,c("survived","age")])
>
> ## Amount of missigness in age for each sex group
> barMiss(titanic.missing[,c("sex","age")])
> histMiss(titanic.missing)
>
```


Titanic: Missing data



- Only 628 completers.
- Age: more than 50% missing.
- Complete case analysis.

Titanic: Model

Analysis Model

$$\text{logit}\left[P(Y=1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot class_2 + \beta_3 \cdot class_3 + \beta_4 \cdot age$$

Equivalently

$$P(Y=1|class, sex, age) = \frac{e^{\beta_0+\beta_1 sex+\beta_2 class_2+\beta_3 class_3+\beta_4 age}}{1+e^{\beta_0+\beta_1 sex+\beta_2 class_2+\beta_3 class_3+\beta_4 age}}$$

Analyzing the data in R

```
> ## Fitting a logistic regression model for the complete cases
>
> titanic.logistic.omit<-glm(survived ~ pclass + sex + age, family=binomial, data = titanic.missing)
> summary(titanic.logistic.omit)
>
Call:
glm(formula = survived ~ pclass + sex + age, family = binomial,
    data = titanic.missing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9519  -0.6500  -0.3172   0.5857   2.6875

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.439356   0.470042   9.445  < 2e-16 ***
pclass2nd   -1.466980   0.282904  -5.185 2.16e-07 ***
pclass3rd   -2.793512   0.338744  -8.247  < 2e-16 ***
sex         -3.085718   0.240780 -12.816  < 2e-16 ***
age         -0.047645   0.008747  -5.447 5.12e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 862.77  on 627  degrees of freedom
Residual deviance: 536.08  on 623  degrees of freedom
  (685 observations deleted due to missingness)
AIC: 546.08

Number of Fisher Scoring iterations: 5
```

Analyzing the data in R

```
> ## Global effect of class
>
> wald.test(b=coef(titanic.logistic.omit), Sigma=vcov(titanic.logistic.omit),
+ Terms=2:3)
>

Wald test:
----------

Chi-squared test:
X2 = 68.3, df = 2, P(> X2) = 1.4e-15

>
```


Titanic: Results

Coefficient	Explanation	Estimate	Std. Error	z value	p-value
β_0	Intercept	4.43	0.470	9.45	0.00
β_1	sex	-3.09	0.241	-12.82	0.00
β_2	2nd	-1.47	0.282	-5.19	0.00
β_3	3rd	-2.79	0.339	-8.25	0.00
β_4	age	-0.05	0.009	-5.45	0.00

χ^2 test for the effect of class

χ^2	df	p-value
68.3	2	0.00

Analyzing the data in R

```
> ## Odds ratios
> exp(cbind(OR =titanic.logistic.omit$coefficients,
+ confint(titanic.logistic.omit)))
>
Waiting for profiling to be done...
                     OR       2.5 %       97.5 %
(Intercept) 84.72034359 34.83140641 220.47300701
pclass2nd    0.23062102  0.13097332   0.39776333
pclass3rd    0.06120586  0.03083499   0.11664061
sex          0.04569723  0.02805441   0.07223138
age          0.95347203  0.93687359   0.96961325
>
```

- Odds of survival 77% smaller in 2nd class than in 1st class
- Odds of survival 93% smaller in 3rd class than in 1st class
- Odds of survival 95% smaller for men than for women
- An increase of one year in age is associated with a decrease of 5% in the odds of survival

Titanic: Results

Probability of surviving by gender and class after fixing age to its mean value $age = 31.27$

	Probability of Surviving		
Sex	1st	2nd	3rd
Male	0.47	0.18	0.05
Female	0.95	0.82	0.54

- Huge effect of class
  - Males: 1st class nine times more chance than 3rd class.
  - Females: 1st class two times more chance than 3rd class.
- Huge effect of gender: in 1st, 2nd and 3rd class women had 2, 5 and 11 times more chance to survive than men.

Missing data

- Common in many scientific investigations
  - A questionnaire got lost
  - Some subjects did not report their income
  - A machine got broken
- Determining the appropriate analytic approach is a major question
  - Throw them away?
  - Make a guess about their values?
  - Use the information available?
- Development of statistical methods has been an active area of research.

Missing data

Missing data:

Observations that are intended to be made but are not made.

Two possible, but distinct, goals

- Make inferences that would apply to the population targeted by the complete sample.
- Make inferences that would apply to those subject remaining in the study, or with complete data (on relevant variables).

We will focus on the first of these.

Missing data: Common strategies

Complete Cases: In complete cases (CC), sometimes also called listwise deletion (LD), all cases with missing values are deleted. Following deletion, conventional methods are used to derive estimates from the remaining, complete cases. **Default in many software.**

Available Cases: In available cases (AC), also called pairwise deletion (PD), each moment is estimated separately using cases with values for the pertinent variables.

Mean substitution: Special case of imputation. Substitution of missing values with the simple (grand) mean (MS).

Last Observation Carried Forward: Special case of imputation. Whenever a value is missing, the last observed value is substituted (LOCF).

Does it really matter?

Missing data: Simulation I

Generated 50 000 observations from the bivariate normal (X, Y) with correlation $\rho = 0.5$. So at the population level

$$\mu_X = E(X) = 0 \qquad \text{Var}(X) = 1$$
$$\mu_Y = E(Y) = 0 \qquad \text{Var}(Y) = 1$$
$$\rho = \text{Corr}(X, Y) = 0.5$$

Two settings where Y is always observed and X sometimes missing

- X is missing with probability 0.5.
- X is missing if $Y < 0$.

X is missing with probability 0.5

Obs	X	Y
1	-1.62	-0.05
2	0.49	0.13
3	-0.19	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

X is missing with probability 0.5

- CC means deleting cases where X is missing.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

Obs	X	Y
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮
⋮	⋮	⋮
⋮	⋮	⋮

X is missing with probability 0.5

- CC means deleting cases where X is missing.
- Following deletion, conventional methods are used to derive estimates from the remaining, complete cases.

	Population values			Complete Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.99	0.499
Y	0	1		0.001	1.00	

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.

Obs	X	Y
1	NA	-0.05
2	NA	0.13
3	NA	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

- To calculate $\overline{Y}$
- To calculate $\overline{X}$
- To calculate Corr(X, Y)

X is missing with probability 0.5

- AC: each moment is estimated separately using cases with values for the pertinent variables.
- $E(Y)$ and $Var(Y)$ would be estimated using all the cases.
- $E(X)$, $Var(X)$, and $cov(X, Y)$ would be estimated using only the cases with values for X.

	Population values			Available Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.99	0.499
Y	0	1		0.003	1.00	

X is missing with probability 0.5

- Mean imputation: All missing values in X are imputed using $\bar{X}_n$.
- Following imputation, conventional methods are used to derive estimates.

	Population values			Mean imputation		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.007	0.496	0.249
Y	0	1		0.003	1.00	

X is missing if $Y < 0$

Obs	X	Y
1	-1.62	-0.05
2	0.49	0.13
3	-0.19	-0.59
4	-0.29	0.79
5	-1.56	-2.25
6	0.94	0.07
7	-1.01	-0.82
8	1.90	-0.12
9	-1.05	-0.38
10	-0.56	-0.89
⋮	⋮	⋮

	X	Y
1	NA	-0.05
2	0.49	0.13
3	NA	-0.59
4	-0.29	0.79
5	NA	-2.25
6	0.94	0.07
7	NA	-0.82
8	NA	-0.12
9	NA	-0.38
10	NA	-0.89
⋮	⋮	⋮

X is missing if $Y < 0$

- CC means deleting cases where X is missing since $Y < 0$.

	Population values			Complete Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.397	0.841	0.185
Y	0	1		0.798	0.363	

- AC: Moments estimated using cases with values for the pertinent variables.

	Population values			Available Cases		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.397	0.841	0.185
Y	0	1		0.005	0.996	

X is missing if $Y < 0$

- Mean imputation.

	Population values			Mean imputation		
	Mean	Var	Corr	Mean	Var	Corr
X	0	1	0.5	0.399	0.421	0.009
Y	0	1		0.003	1.00	

Simulation I: Conclusion

Case I: X is missing with probability 0.5

- No major problems.
- CC and AC worked fine.
- MS failed.

Case II: X is missing if $Y < 0$. All methods seem to fail.

What is going on?

Missing Data: Formal definitions

Let $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)^T$ denote the data at hand. One is interested in the regression of Y_i versus a vector of covariates $\boldsymbol{x}_i$ and split $\boldsymbol{z}_i = (\boldsymbol{z}_i^{obs}, \boldsymbol{z}_i^{mis})^T$.

Notice that the latter partition is subject-specific. In addition, let $\boldsymbol{r}_i$ denote a vector of zeros and ones with $r_{ik} = 1$ if z_{ik} is observed and zero otherwise.

- Missing completely at random (MCAR): $P(\boldsymbol{r}_i \mid \boldsymbol{z}_i) = P(\boldsymbol{r}_i)$
- Missing at random (MAR): $P(\boldsymbol{r}_i \mid \boldsymbol{z}_i) = P(\boldsymbol{r}_i \mid \boldsymbol{z}_i^{obs})$
- Missing not at random (MNAR): $P(\boldsymbol{r}_i \mid \boldsymbol{z}_i) = P(\boldsymbol{r}_i \mid \boldsymbol{z}_i^{obs}, \boldsymbol{z}_i^{mis})$

Missing not at random: MNAR

The probability that an observation is missing depends on subject information that is not observed, like the value of the missing observation itself

- Studying mental health: People who have been diagnosed as depressed may report their mental status less often than others.
- Asking for income level: Missing data may be more likely to occur when the income level is relatively high/low.

MNAR: Highly problematic.

The only way to obtain unbiased estimates of the parameters is to model missingness. In other words, we would need to write a model that accounts for the missing data mechanism and **hope** this model is approximately correct.

Missing at random: MAR

The probability that an observation is missing depends on subject information that is present, i.e., missingness can be described using observed subject variables

- Depressed people may be less inclined to report their income and, hence, missingness in income will be related to depression. If mental status is always observed then missingness in income is MAR.
- Depressed people may also have a lower income. A high rate of missing data among depressed individuals $\Rightarrow$ observed mean income might be much larger than it would be without missing data.
- The probability of drop out may depend on the treatment received.

MAR: No simple methods.

Generally, under MAR, simple techniques like complete and available case analysis and overall mean imputation, give biased results.

Missing completely at random: MCAR

The probability that an observation is missing is not related to any other subject characteristics

- Equipment malfunctioned.
- The weather was terrible.
- Data were not entered correctly.

MCAR: Most methods work.

Although very inefficient, some simple techniques like complete and available case analysis will give unbiased results under MCAR. However, MS and LOCF **do not** work in this setting neither.

Simulation I: Remarks

Three settings where Y is always observed and X sometimes missing

- MCAR: X is missing with probability 0.5.
- MAR: X is missing if $Y < 0$.

Titanic: Simulations II

Simulations mimicking Titanic data set

- Age simulated mimicking the original data.
- Gender: $sex \sim$ Bernoulli(0.5). For men $sex = 1$.
- Only two classes considered $class = 1$ indicating first class.
- Survival (Y) like in case study and

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = 2.18 + 1.93 \cdot class - 3.04 \cdot sex - 0.04 \cdot age$$

Titanic: The incomplete data

Generating the missing data

- 2500 datasets were generated each with 1000 passengers.
- Missing data created for age.
- The probability of age being missing depending on:
  - Class: First class less chance of missing age
  - Survival: Survivors less chance of missing age
  - Missing generating mechanism.

$$\text{logit}\left[P(r = 0|class, Y)\right] = 2.11 - 1.5 \cdot class - 2.85 \cdot Y$$

$r = 0$ implies that age is missing.

Titanic: Simulations II

Analysis

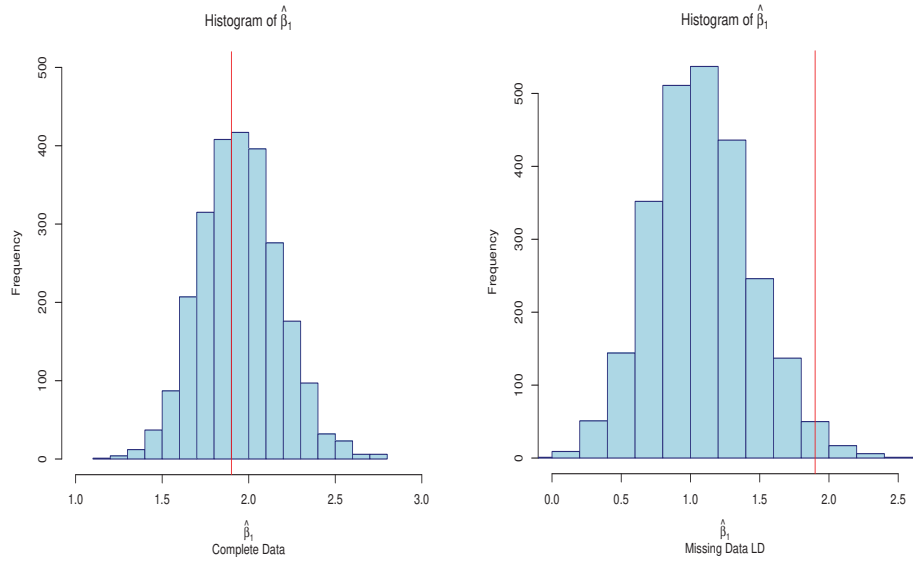
Model:

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot class + \beta_2 \cdot sex + \beta_3 \cdot age$$

Two types of analysis

- Data sets without missing values (Complete data).
- Data sets with missing values analyzed using complete cases (also called listwise deletion LD). Around 50% of the observations in each dataset had missing age.

Simulations II: Results



How to handle missing data?

Three methods to handle missing values:

- Complete Cases: Just analyse individuals with complete data.
- Multiple imputation (MI): Stochastically fill in missing values using observed data
  - Create multiple complete datasets
  - Apply complete-data estimators to each
  - Combine estimates (Rubin's Rules)
- Inverse probability weighting (IPW): Like Complete Cases but weight every individual by the inverse of P(no-missing)=(his probability of not having missing information).

Important

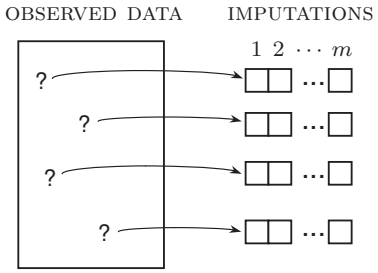
CC valid only under MCAR. IPW and MI valid under MAR!!!

Multiple imputations

Why multiple imputations?

- Single imputation techniques overestimate precision, since no correction is made for the uncertainty introduced from imputing the missing observations.
- This additional variability in the estimates is made explicit by generating multiple completed data sets.
- Each time replace missing values $\boldsymbol{z}_{mis}$ by draws from the conditional distribution $f(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \widehat{\psi})$, rather than by the average of that distribution.

A simulation-based approach to missing data



- Generate $m > 1$ plausible versions of $\boldsymbol{z}_{mis}$.
- Analyze each of the m datasets by standard complete-data methods.
- Combine the results.

Rubin (1987) calls this the repeated-imputation inference method

How to impute the missing values

Imputation model: For instance for the Titanic simulations one can impute age for subject i using the model $f(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\psi})$

$$age_i = \gamma_0 + \gamma_1 class_i + \gamma_2 sex_i + \gamma_3 Y_i + \epsilon_i$$

- First this model is fitted to the completers to estimate the parameters $\widehat{\boldsymbol{\psi}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3, \widehat{\sigma}^2_\epsilon)$.
- Note that there is also variability introduced from replacing $\boldsymbol{\psi}$ in $f(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\psi})$ by an estimate.
- However, we usually have an estimate for the variation in $\widehat{\boldsymbol{\psi}}$: $\widehat{\boldsymbol{\psi}} \sim N(\boldsymbol{\psi}, \widehat{\Sigma}_\psi)$.
- Drawing $\boldsymbol{\psi}$ from $N(\widehat{\boldsymbol{\psi}}, \widehat{\Sigma}_\psi)$ accounts for this additional variation.

The imputation algorithm

- Draw $\boldsymbol{\psi}^{(k)}$ from $N(\widehat{\boldsymbol{\psi}}, \widehat{\Sigma}_\psi)$
- Draw $\boldsymbol{z}^{(k)}_{mis}$ from $f(\boldsymbol{z}_{mis} \mid \boldsymbol{z}_{obs}, \boldsymbol{\psi}^{(k)})$
- Using the completed data $(\boldsymbol{z}_{obs}, \boldsymbol{z}^{(k)}_{mis})$, calculate an estimate $\widehat{\boldsymbol{\theta}}^{(k)}$ for the parameter $\boldsymbol{\theta}$ of interest, as well as its covariance matrix $\boldsymbol{U}^{(k)}$
- Repeat this m times
- Note that $\boldsymbol{U}^{(k)}$ reflects the sampling uncertainty, i.e., the uncertainty in the estimates of $\boldsymbol{\theta}$ due to the fact that only a finite sample is available.
- We can now obtain inferences for $\boldsymbol{\theta}$ from pooling the estimates

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m} \sum_{k=1}^{m} \widehat{\boldsymbol{\theta}}^{(k)}$$

The imputation algorithm

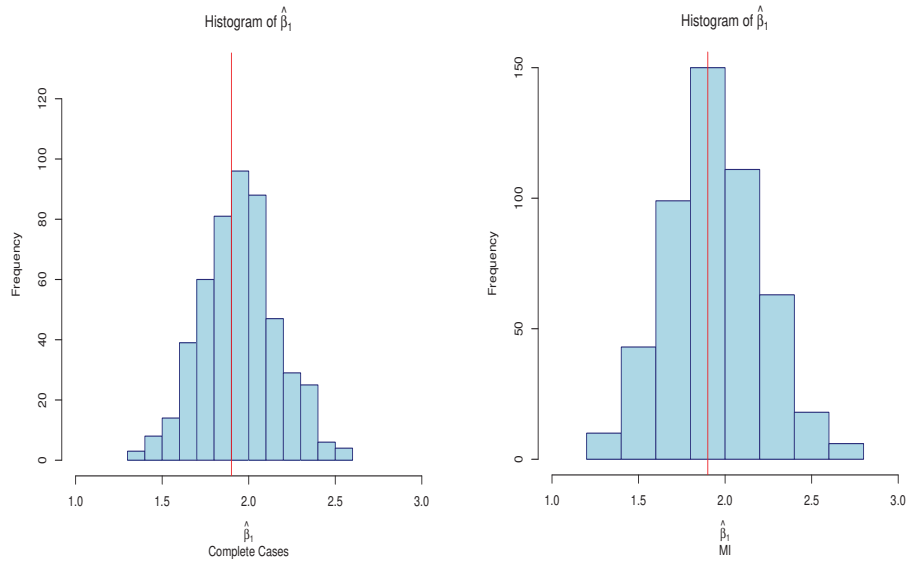
- The covariance matrix of $\widehat{\theta}$ equals

$$\text{var}(\widehat{\theta}) = \widehat{\boldsymbol{W}} + \left(\frac{m+1}{m}\right)\widehat{\boldsymbol{B}}$$

where $\widehat{\boldsymbol{W}} = \dfrac{\sum_{k=1}^{m} \boldsymbol{U}^{(k)}}{m}$ and $\widehat{\boldsymbol{B}} = \dfrac{\sum_{k=1}^{m}\left(\widehat{\theta}^{(k)} - \widehat{\theta}\right)\left(\widehat{\theta}^{(k)} - \widehat{\theta}\right)'}{m-1}$.

- $\widehat{\boldsymbol{W}}$ represents the within-imputation variance, representing sampling uncertainty
- $\widehat{\boldsymbol{B}}$ represents the between-imputation variance, representing the uncertainty in imputing the missing observations as well as the uncertainty in the estimation of ψ.
- Typically, m will be small: $m = 5, 10$ already yields a major improvement over single imputation.

Titanic simulation: MI results (500 data sets and $m = 5$)



Multiple imputation in R

- Several packages available: Amelia, VIM, mice...
- Different algorithms
  - Amelia: Bootstrapped EM algorithm
  - VIM: Iterative Robust Model-based Imputation (irmi)
  - mice: Chained equations algorithm (CEA)
- CEA has been found to work well in a variety of simulation studies (Schunk 2008; Drechsler and Rassler 2008; Giorgi et al. 2008)
- Area of active research

Observations and warnings

- Variables used to impute a missing outcome may themselves be incomplete
- Rows or columns in the data can be ordered, e.g., as with longitudinal studies
- Variables can be of different types (e.g., binary, unordered, ordered, continuous), thereby making the application of theoretically convenient models, such as the multivariate normal, inappropriate
- Imputation can create impossible combinations (e.g. pregnant fathers), or destroy deterministic relations in the data (e.g. sum scores)
- Imputations can be nonsensical (e.g. body temperature of the dead)

Titanic data: MI

- We impute each missing value 100 times and fitted the following model to every imputed data set

$$\text{logit}\left[P(Y=1|class, sex, age)\right] = \beta_0 + \beta_1 sex + \beta_2 class_2 + \beta_3 class_3 + \beta_4 age$$

- We obtained then 100 estimates for the parameters of interest with $k = 1, 2, \ldots, 100$

$$\widehat{\boldsymbol{\theta}}^{(k)} = (\widehat{\beta}_0^{(k)}, \widehat{\beta}_1^{(k)}, \widehat{\beta}_2^{(k)}, \widehat{\beta}_3^{(k)}, \widehat{\beta}_4^{(k)})$$

- We combined all these estimates using the Rubin's rules previously described.

Multiple imputation in R

```
> ###### Titanic multiple imputation
>
> ## Studying the patterns of missiness
>
> pattern=md.pattern(titanic.missing)
> pattern
    survived pclass sex age
628        1      1   1   1   0
  5        1      1   0   1   1
675        1      1   1   0   1
  5        1      1   0   0   2
           0      0  10 680 690
>
> pairs=md.pairs(titanic.missing)
> pairs
>
$rr
         survived pclass  sex age
survived     1313   1313 1303 633
pclass       1313   1313 1303 633
sex          1303   1303 1303 628
age           633    633  628 633

$rm
         survived pclass sex age
survived        0      0  10 680
pclass          0      0  10 680
sex             0      0   0 675
age             0      0   5   0

$mr
         survived pclass sex age
survived        0      0   0   0
pclass          0      0   0   0
sex            10     10   0   5
age           680    680 675   0

$mm
         survived pclass sex age
survived        0      0   0   0
pclass          0      0   0   0
sex             0      0  10   5
age             0      0   5 680
>
```

Multiple imputation in R

```
> ## Imputing the missing values
>
> imp <- mice(titanic.missing, m=100)
> imp
>
Multiply imputed data set
Call:
mice(data = titanic.missing, m = 100)
Number of multiple imputations:  100
Missing cells per column:
survived   pclass      sex      age
       0        0       10      680
Imputation methods:
survived   pclass      sex      age
      ""       ""    "pmm"    "pmm"
VisitSequence:
sex age
  3   4
PredictorMatrix:
         survived pclass sex age
survived        0      0   0   0
pclass          0      0   0   0
sex             1      1   0   1
age             1      1   1   0
Random generator seed value:  NA
>
## Imputations are generated according to the default method, which is, for numerical data, predictive
## mean matching (pmm) (Little 1988).
```


Diagnostic checking

- An important step in multiple imputation is to assess whether imputations are plausible
- Imputations should be values that could have been obtained had they not been missing
- Imputations should be close to the data
- Data values that are clearly impossible (e.g. negative counts, pregnant fathers) should not occur in the imputed data
- Imputations should respect relations between variables, and respect the appropriate amount of uncertainty about their *true* values
- Diagnostic checks on the imputed data provide a way to check the plausibility of the imputations

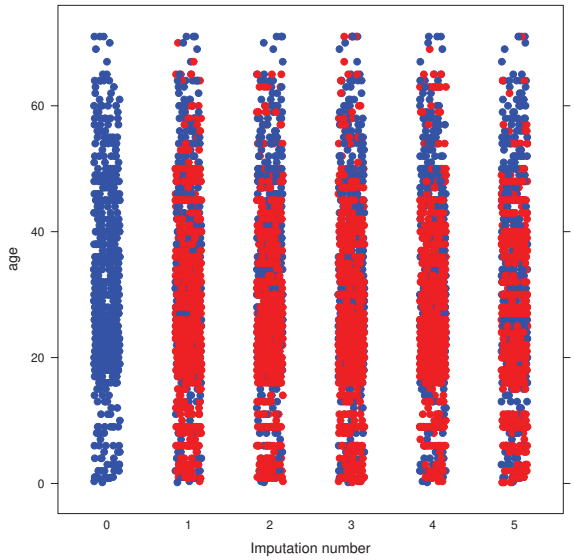
Diagnostic checking in R

```
> ## Imputed values for age. Each row corresponds to a missing entry in age.
> ## The columns contain the multiple imputations.
> imp$imp$age[1:10,1:5]
>
    1       2       3  4  5
13 60 27.0000 19.0000 55 22
14 57  0.9167 17.0000 48 26
15 47 28.0000 50.0000 47 31
30 28 28.0000 56.0000 55 40
33 22 37.0000 39.0000 24 30
36 50 50.0000 64.0000 61 27
41 30 34.0000  0.9167 34 38
46 62 46.0000 54.0000 36 39
47 61 58.0000 46.0000 61 24
53 45 37.0000 54.0000 21 23
>
> ## The complete data combine observed and imputed data.
> ## The first completed data set can be obtained as  (only first 10 passenger shown)
>
> complete(imp,1)[1:10,]
   survived pclass sex     age
1         1    1st   0 29.0000
2         0    1st   0  2.0000
3         0    1st   1 30.0000
4         0    1st   0 25.0000
5         1    1st   1  0.9167
6         1    1st   1 47.0000
7         1    1st   0 63.0000
8         0    1st   1 39.0000
9         1    1st   0 58.0000
10        0    1st   1 71.0000
>
```

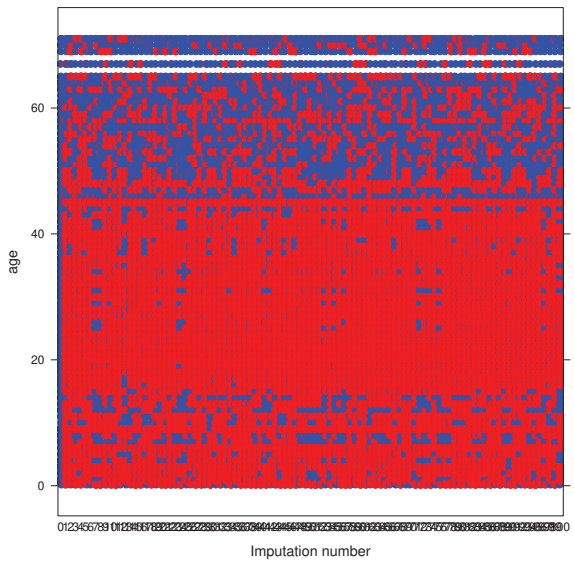

Diagnostic checking in R

```
> ## It is often useful to inspect the distributions of original and the imputed
> ## data. The complete() function extracts the original and the imputed data
> ## sets from the imp object as a long (row-stacked) matrix. The col vector
> ## separates the observed (blue) and imputed (red) data for age
>
> com <- complete(imp, "long", inc=T)
> col <- rep(c("blue","red")[1+as.numeric(is.na(imp$data$age))],101)
> stripplot(age~.imp, data=com, jit=TRUE, fac=0.8, col=col, pch=20, cex=1.4,
+           xlab="Imputation number")
>
```

Distributions: Original versus imputed data



Distributions: Original versus imputed data



Imputation methods

Method	Description	Scale type	Default
pmm	Predictive mean matching	numeric	Y
norm	Bayesian linear regression	numeric	
norm.nob	Linear regression, non-Bayesian	numeric	
mean	Unconditional mean imputation	numeric	
2l.norm	Two-level linear model	numeric	
logreg	Logistic regression	factor, 2 levels	Y
polyreg	Polytomous (unordered) regression	factor, >2 levels	Y
lda	Linear discriminant analysis	factor	
sample	Random sample from the observed data	any	

- The method argument of mice() specifies the imputation method per column and overrides the default
- Columns that need not be imputed have method "",i.e.,

```
imp <- mice(titanic.missing, meth = c("", "", "logreg", "pmm"), m=100)
```


Analysis of imputed data in R

```
> ## Analyzing the imputed data sets
>
> fit <- with(data=imp, exp=glm(survived ~ pclass + sex + age, family=binomial))
>
> ## Creating a data set with the results of all the analysis
>
> MI.matrix<-matrix(0,100,5)
> for(k in 1:100) MI.matrix[k,]<-coefficients(fit$analyses[[k]])
> MI.results=data.frame(Intercept=MI.matrix[,1], pclass2=MI.matrix[,2],
+          pclass3=MI.matrix[,3], sex=MI.matrix[,4], age=MI.matrix[,5])
> MI.results[1:10,]
>
   Intercept   pclass2   pclass3       sex         age
1   3.321512 -1.201354 -2.606778 -2.437407 -0.03492116
2   4.042564 -1.412543 -2.858506 -2.657812 -0.04863579
3   4.217690 -1.531627 -3.031196 -2.593078 -0.05211627
4   3.504774 -1.316043 -2.749440 -2.387495 -0.03891783
5   4.399160 -1.584609 -3.001377 -2.631284 -0.05634107
6   3.668436 -1.331814 -2.821121 -2.402105 -0.04331810
7   3.686304 -1.385195 -2.826104 -2.452432 -0.04270390
8   3.597697 -1.306929 -2.874065 -2.417954 -0.04120242
9   3.751935 -1.395021 -2.781783 -2.437738 -0.04433574
10  3.598338 -1.283901 -2.764597 -2.450878 -0.04034921
>
```

Analysis of imputed data in R

```
> ## Combining the results using Rubin's rule
>
> est <- pool(fit)
> summary(est)
                   est          se          t       df     Pr(>|t|)       lo 95       hi 95
(Intercept)  3.62632240 0.464628361   7.804781 212.6374 2.664535e-13  2.71045483  4.54218997
pclass2     -1.30813713 0.248053338  -5.273612 639.4154 1.832732e-07 -1.79523474 -0.82103951
pclass3     -2.76475931 0.262026202 -10.551461 474.9145 0.000000e+00 -3.27963337 -2.24988524
sex         -2.48033948 0.168420817 -14.727036 919.6779 0.000000e+00 -2.81087322 -2.14980575
age         -0.04111961 0.009669708  -4.252415 171.3019 3.470195e-05 -0.06020674 -0.02203248

             nmis       fmi    lambda
(Intercept)    NA 0.5505477 0.5463401
pclass2        NA 0.2382350 0.2358561
pclass3        NA 0.3160640 0.3131898
sex            10 0.1416097 0.1397450
age           680 0.6210089 0.6166097
>
> ## The column fmi contains the fraction of missing information, i.e. the proportion of the
> ## variability that is attributable to the uncertainty caused by the missing data.
>
```


Titanic results: CC+MI

		CC		MI	
Coefficient	Explanation	Estimate	Std. Error	Estimate	Std. Error
β_0	Intercept	4.43	0.470	3.63	0.465
β_1	sex	-3.09	0.241	-2.48	0.168
β_2	2nd	-1.47	0.282	-1.31	0.248
β_3	3rd	-2.79	0.339	-2.76	0.262
β_4	age	-0.05	0.009	-0.04	0.009

- Some differences in the estimates of β_1 (sex) and β_2 (2nd class indicator)
- Although p-values differ we get the same qualitative conclusions
- It is not always like this

Inverse probability weighting (IPW)

Suppose we have the following data

Group	A			B			C		
Response	1	1	1	2	2	2	3	3	3

then the average response is 2. However if we observed

Group	A			B			C		
Response	1	?	?	2	2	2	?	3	3

then the average response is $13/6 = 2.17$ which is biased.

Inverse probability weighting (IPW)

Suppose we have the following data

Group	A			B			C		
Response	1	?	?	2	2	2	?	3	3
$P(\text{Response})$		$\frac{1}{3}$			1			$\frac{2}{3}$	
$\frac{1}{P(\text{Response})}$		3			1			$\frac{3}{2}$	

Calculate weighted average

$$\frac{1 \cdot 3 + (2+2+2) \cdot 1 + (3+3) \cdot \frac{3}{2}}{3 + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}} = 2$$

Thus IPW has eliminated the biased. Notice that this example is MAR.

Titanic: Simulations II

Simulations mimicking Titanic data set

- Age simulated mimicking the original data.
- Gender: $sex \sim$ Bernoulli(0.5). For men $sex = 1$.
- Only two classes considered $class = 1$ indicating first class.
- Survival (Y) like in case study and

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = 2.18 + 1.93 \cdot class - 3.04 \cdot sex - 0.04 \cdot age$$

Titanic: The incomplete data

Generating the missing data

- 2500 datasets were generated each with 1000 passengers.
- Missing data created for age.
- The probability of age being missing depending on:
  - Class: First class less chance of missing age
  - Survival: Survivors less chance of missing age
  - Missing mechanism

$$\text{logit}\left[P(r = 0|class, Y)\right] = 2.11 - 1.5 \cdot class - 2.85 \cdot Y$$

$r = 0$ implies that age is missing.

Titanic: Simulations II

Analysis

Model:

$$\text{logit}\left[P(Y = 1|class, sex, age)\right] = \beta_0 + \beta_1 \cdot class + \beta_2 \cdot sex + \beta_3 \cdot age$$

Type of analysis

- Inverse probability weighting (IPW)

Inverse probability weighting (IPW)

Create the new variable r with $r = 0$ when age is missing and $r = 1$ when age is observed

Passenger	survived	class	sex	age	r
1	0	0	0	NA	0
2	0	1	0	30.44	1
3	1	1	1	26.60	1
4	0	0	0	NA	0
5	1	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	
197	1	0	1	28.67	1
198	1	0	1	28.88	1
199	0	1	1	22.77	1
200	0	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	

Titanic simulation: IPW

Recall that $r = 1$ if age is observed. One can then fit the model

$$\text{logit}\left[P(r = 1|\textit{class}, Y)\right] = \alpha_0 + \alpha_1 \cdot \textit{class} + \alpha_2 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

Passenger	survived	class	sex	age	r
1	0	0	0	NA	0
2	0	1	0	30.44	1
3	1	1	1	26.60	1
4	0	0	0	NA	0
5	1	0	0	NA	0
⋮	⋮	⋮	⋮	⋮	

Titanic simulation: IPW

The weight associated with subject i is

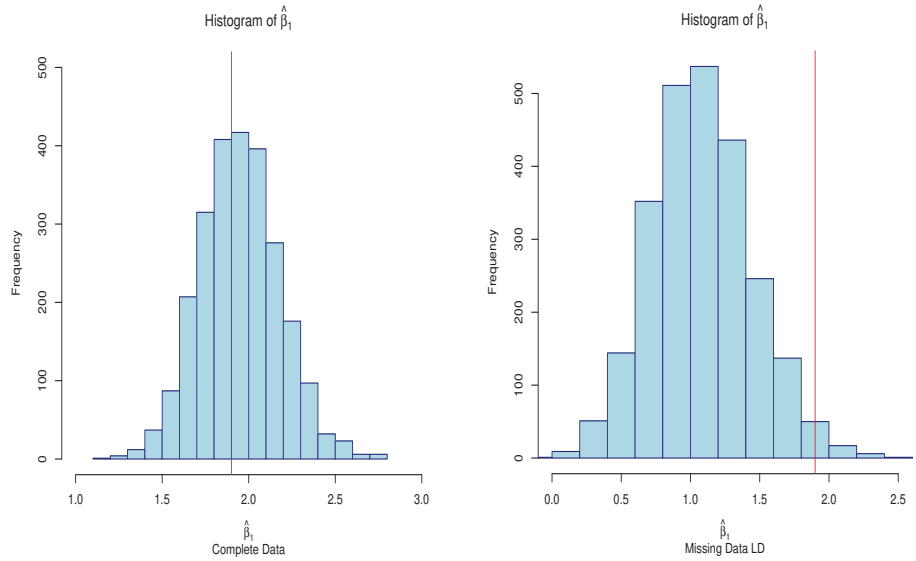
$$w_i = \frac{1}{P(r_i = 1|\textit{class}_i, \textit{sex}_i)} = 1 + \exp\left(1 + \hat{\alpha}_0 + \hat{\alpha}_1 \cdot \textit{class}_i + \hat{\alpha}_2 \cdot Y_i\right)$$

Data are again analyzed with model

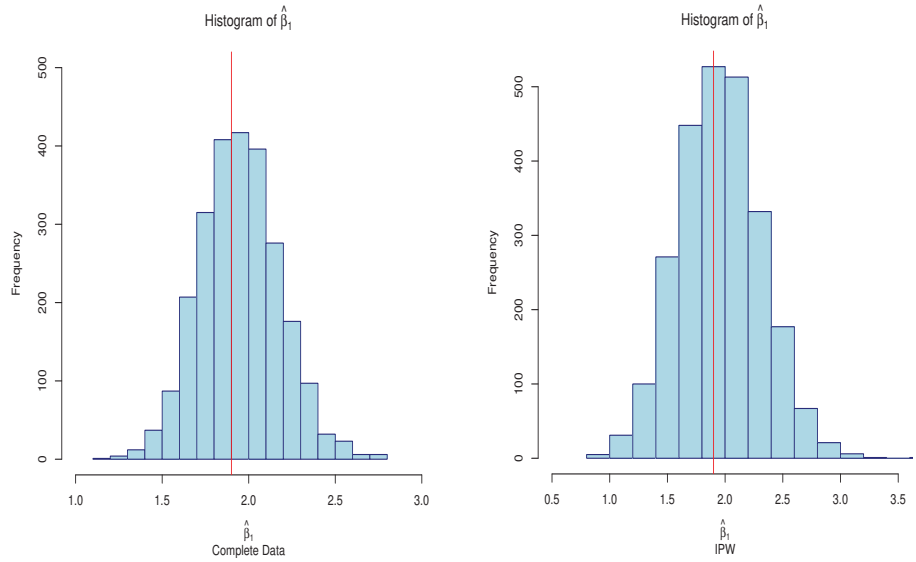
$$\text{logit}\left[P(Y = 1|\textit{class}, \textit{sex}, \textit{age})\right] = \beta_0 + \beta_1 \cdot \textit{class} + \beta_2 \cdot \textit{sex} + \beta_3 \cdot \textit{age}$$

but this time using a weighted logistic regression, i.e. passing the previous w_i weights to the fitting procedure.

Titanic simulation: Complete Case Analysis



Titanic simulation: IPW results



Titanic data: IPW

Let now $r = 1$ if age and sex are observed. One can then fit the model

$$\text{logit}\left[P(r = 1|\textit{class}, Y)\right] = \alpha_0 + \alpha_1 \cdot \textit{class}_2 + \alpha_2 \cdot \textit{class}_3 + \alpha_3 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$.

The weight associated with subject i is

$$\begin{aligned} w_i &= \frac{1}{P(r_i = 1|\textit{class}_{2i}, \textit{class}_{3i}, \textit{sex}_i)} \\ &= 1 + \exp\left(1 + \hat{\alpha}_0 + \hat{\alpha}_1 \cdot \textit{class}_{2i} + \hat{\alpha}_2 \cdot \textit{class}_{3i} + \hat{\alpha}_3 \cdot Y_i\right) \end{aligned}$$

Titanic data: IPW

Let now $r = 1$ if age and sex are observed. One can then fit the model

$$\text{logit}\left[P(r = 1|\textit{class}, Y)\right] = \alpha_0 + \alpha_1 \cdot \textit{class}_2 + \alpha_2 \cdot \textit{class}_3 + \alpha_3 \cdot Y$$

to get the estimates $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$.

Data are again analyzed with model

$$\text{logit}\left[P(Y = 1|\textit{class}, \textit{sex}, \textit{age})\right] = \beta_0 + \beta_1 \cdot \textit{sex} + \beta_2 \cdot \textit{class}_2 + \beta_3 \cdot \textit{class}_3 + \beta_4 \cdot \textit{age}$$

but this time using a weighted logistic regression, i.e. passing the previous w_i weights to the fitting procedure.

IPW in R

```
> ###### Titanic IPW
>
> ## Creating the missing data indicator variable r
>
> titanic.missing$r<-as.numeric(!is.na(titanic.missing$age))*as.numeric(!is.na(titanic.missing$sex))
> head(titanic.missing,15)
>
   survived pclass sex     age r        w
1         1    1st   0 29.0000 1 1.373464
2         0    1st   0  2.0000 1 1.526999
3         0    1st   1 30.0000 1 1.526999
4         0    1st   0 25.0000 1 1.526999
5         1    1st   1  0.9167 1 1.373464
6         1    1st   1 47.0000 1 1.373464
7         1    1st   0 63.0000 1 1.373464
8         0    1st   1 39.0000 1 1.526999
9         1    1st   0 58.0000 1 1.373464
10        0    1st   1 71.0000 1 1.526999
11        0    1st   1 47.0000 1 1.526999
12        1    1st   0 19.0000 1 1.373464
13        1    1st   0      NA 0 1.373464
14        1    1st   1      NA 0 1.373464
15        0    1st   1      NA 0 1.526999
>
```


IPW in R

```
> ## Fitting the logistic regression model to  calculate the probabilities of being complete
>
> titanic.ipw.glm<-glm(r ~ pclass + survived, data=titanic.missing,family=binomial)
> summary(titanic.ipw.glm)
>
Call:
glm(formula = r ~ pclass + survived, family = binomial, data = titanic.missing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7488  -0.7745  -0.7745   0.8119   1.6435

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6406     0.1447   4.426 9.58e-06 ***
pclass2nd     0.2999     0.1856   1.616   0.1062
pclass3rd    -1.6911     0.1559 -10.848  < 2e-16 ***
survived      0.3444     0.1377   2.502   0.0124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1817.7  on 1312  degrees of freedom
Residual deviance: 1538.4  on 1309  degrees of freedom
AIC: 1546.4

Number of Fisher Scoring iterations: 4
>
```

IPW in R

```
> ## Calculating the weights: Inverse Probabilities
>
> titanic.missing$w<-1/fitted(titanic.ipw.glm)
> head(titanic.missing,15)
>
   survived pclass sex     age r        w
1         1    1st   0 29.0000 1 1.373464
2         0    1st   0  2.0000 1 1.526999
3         0    1st   1 30.0000 1 1.526999
4         0    1st   0 25.0000 1 1.526999
5         1    1st   1  0.9167 1 1.373464
6         1    1st   1 47.0000 1 1.373464
7         1    1st   0 63.0000 1 1.373464
8         0    1st   1 39.0000 1 1.526999
9         1    1st   0 58.0000 1 1.373464
10        0    1st   1 71.0000 1 1.526999
11        0    1st   1 47.0000 1 1.526999
12        1    1st   0 19.0000 1 1.373464
13        1    1st   0      NA 0 1.373464
14        1    1st   1      NA 0 1.373464
15        0    1st   1      NA 0 1.526999
>
```


IPW in R

```
> titanic.results.ipw<- glm(survived ~ pclass + sex + age, data=titanic.missing, weights=titanic.missing$w,
+ family=binomial)
> summary(titanic.results.ipw)
>
Call:
glm(formula = survived ~ pclass + sex + age, family = binomial,
    data = titanic.missing, weights = titanic.missing$w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3652  -0.8523  -0.5784   0.7815   4.6110

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.766252   0.322451  11.680  < 2e-16 ***
pclass2nd   -1.294939   0.220977  -5.860 4.63e-09 ***
pclass3rd   -2.720643   0.221377 -12.290  < 2e-16 ***
sex         -2.659578   0.159116 -16.715  < 2e-16 ***
age         -0.041524   0.006168  -6.732 1.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1687.8  on 627  degrees of freedom
Residual deviance: 1112.7  on 623  degrees of freedom
  (685 observations deleted due to missingness)
AIC: 1081.6

Number of Fisher Scoring iterations: 4
>
```

Titanic data: CC, MI and IPW

		CC		IPW	
Coefficient	Explanation	Estimate	Std. Error	Estimate	Std. Error
β_0	Intercept	4.43	0.470	3.77	0.322
β_1	sex	-3.09	0.241	-2.66	0.159
β_2	2nd	-1.47	0.282	-1.29	0.221
β_3	3rd	-2.79	0.339	-2.72	0.221
β_4	age	-0.05	0.009	-0.04	0.006
		MI			
β_0	Intercept	3.63	0.465		
β_1	sex	-2.48	0.168		
β_2	2nd	-1.31	0.248		
β_3	3rd	-2.76	0.262		
β_4	age	-0.04	0.009		