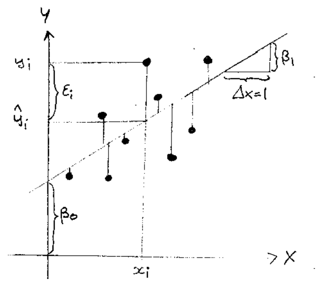
Linear regression

- Assuming that the relationship between the dependent and independent variables can be modeled by a straight line:

$$y = \beta_0 + \beta_1 x$$

the problem is to found out the values of the coefficients β_0 and β_1 so that the regression line fits best the data points (β_0 is the *intercept* and β_1 the *slope* of the regression line)



- The ***predicted value*** of Y for a given value of X is

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- The ***residual*** (error) is the difference between the observed and predicted value

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

- By definition, the observed value y_i is always equal to the sum of the predicted value $\hat{y}_i$ and the residual ε_i:

$$y_i = \hat{y}_i + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The ***method of the least squares*** to find out the values of the coefficients of the regression line is to minimize the sum of the squared vertical distance between the observed value y_i and the predicted value $\hat{y}_i$:

$$\min_{\beta_0,\beta_1} \sum_i \varepsilon_i^2 = \min_{\beta_0,\beta_1} \sum_i (yi - \hat{y}_i)^2$$
$$= \min_{\beta_0,\beta_1} \sum_i (yi - \beta_0 - \beta_1 x_i)^2$$

Least-square estimation

- To minimize the residual sum of squares SSE

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

we differentiate SSE w.r.t β_0 and β_1

$$\left.\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2\sum_i (yi - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2\sum_i x_i (yi - \beta_0 - \beta_1 x_i) = 0 \end{aligned}\right\} \text{Normal equations}$$

solving for β_0 and β_1 yields

$$\left.\begin{aligned} \beta_1 &= \frac{SXY}{SSX} \\ \beta_0 &= m_y - \beta_1 m_x \end{aligned}\right\} \text{Analytical solution}$$

Proof:

$$-2\sum_i (yi - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum_i yi - \beta_1 \sum_i x_i = N\beta_0 \Rightarrow \beta_0 = \frac{1}{N}\sum_i yi - \beta_1 \frac{1}{N}\sum_i x_i = m_y - \beta_1 m_y$$

$$-2\sum_i x_i (yi - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \beta_1 \sum_i x_i^2 = \sum_i x_i yi - \beta_0 \sum_i x_i = \sum_i x_i yi - (m_y - \beta_1 m_x)\sum_i x_i = \sum_i x_i yi - m_y \sum_i x_i + \beta_1 \frac{1}{N}\left(\sum_i x_i\right)^2$$

$$\Rightarrow \beta_1 \left(\sum_i x_i^2 - N{m_x}^2\right) = \sum_i x_i yi - N m_y m_x \Rightarrow \beta_1 = \frac{\sum_i x_i yi - N m_y m_x}{\sum_i x_i^2 - N{m_x}^2} = \frac{SXY}{SSX}$$

Multiple Regression

- In the multiple regression, we try to predict the value of the dependent variable Y on the basis of two or more independent variables (predictors) $\{X_1, X_2, \ldots, X_P\}$.
- The multiple line regression model

model parameters

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_P x_{Pi} + \varepsilon_i \quad (i = 1, \ldots, N)$$

where x_{1i} and x_{2i} represent the values of the predictor variables and y_i represents the value of the independent variable for the i^{th} observations, N represents the number of observations in the data set.

- The parameters of the model are the regression coefficients, β_1, β_2, ... β_p.
- The linear model has one random effect, the error term εi. The error term is assumed to follow a normal distribution $N(0, \sigma^2)$. Moreover, the error terms for the various observations are assumed to be uncorrelated ($\mathrm{cov}(y_i, y_j) = \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$).

Linear models in matrix form

In matrix form

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{p1} \\ x_{11} & x_{22} & & x_{p2} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{11} & x_{21} & \cdots & x_{pn} \end{bmatrix}}_{\text{design matrix X}} \overbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}}^{\text{model parameters}} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

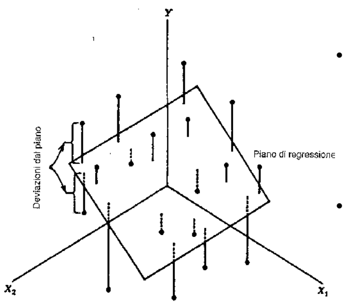
or where $\boldsymbol{y} = (y_1,..y_N)$ is the response vector, $\mathbf{X}$ is the model or design matrix, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$ is the vector of regression of coefficients and is the vector of errors (or residuals).

The vector of errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ is assumed to follow a n-variable multivariate-normal distribution with a n by n covariance matrix Σ

$$\Sigma = \begin{bmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1,\varepsilon_2) & \cdots & \mathrm{cov}(\varepsilon_1,\varepsilon_n) \\ \mathrm{cov}(\varepsilon_2,\varepsilon_1) & \mathrm{var}(\varepsilon_2) & & \mathrm{cov}(\varepsilon_2,\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(\varepsilon_n,\varepsilon_1) & \mathrm{cov}(\varepsilon_n,\varepsilon_2) & \cdots & \mathrm{var}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{12}^2 & \sigma_2^2 & & \sigma_{2n}^2 \\ \vdots & \vdots & & \vdots \\ \sigma_{1n}^2 & \sigma_{2n}^2 & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

The classic assumptions of homogeneity of the variances ($\mathrm{var}(\varepsilon_i) = \sigma^2$) and uncorrelated observations ($\mathrm{cov}(\varepsilon_i,\varepsilon_j) = 0$ for $i \neq j$) imply that $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I_n}$.

Geometric interpretation



- Any linear multiple regression involving p predictors can be represented graphically in p+1 dimensional space.
- For example, for two predictors, we can represent each observation (x_{1i}, x_{2i}, y_i) inside a three dimensional space. The value predicted by the multiple regression model lie in a plane:

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

- The residuals corresponds to the vertical distance between the data points and the plane (predicted values):

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

- The coefficients of the multiple regression plane minimize the deviations from the plane (residual or error sum of squares):

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

Linear model theory

- The parameters $\boldsymbol{\beta}$ of the linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

are obtained by that minimizing the sum of squares

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\varepsilon}\| = \min_{\boldsymbol{\beta}} \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\| \quad \text{where} \quad \|\boldsymbol{\varepsilon}\| = \sum_i \varepsilon_i^2$$

differentating w.r.t. β_i,equating by 0, and solving for β_i yields

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1} X' \mathbf{y}$$

This is not a good formula to compute $\boldsymbol{\beta}$ (other methods based on QR decomposition of X are actually used).

- The predicted values are

$$\hat{\mathbf{y}} = X\boldsymbol{\beta} = \underbrace{X(X'X)^{-1}X'}\mathbf{y}$$

projection operator
(= hat or influence matrix)

Properties of the hat matrix
- tr(A)=p (nb. of parameters)
- idempotence (AA=A)

lm

Syntax

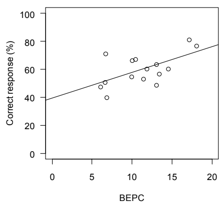
```
lm(formula, data, subset)
aov(formula, data, subset)
```

Usage

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although `aov` may provide a more convenient interface for these).

Linear regression example

```
# data set
data<-data.frame(
     BEPC=c(6.7,18,11.8,13,6.8,9.9,11.4,
          6.6,10,13,17.1,6,13.4,10.4,14.5),
     Percentage=c(71.1,77,60.4,48.7,39.8,54.9,53,
           50.6,66.5,63.4,81.1,47.5,56.9,67.2,60.2))
# plot
plot(data$BEPC,data$Percentage,xlim=c(0,20),ylim=c(0,100),
     las=1,xlab="BEPC",ylab="Correct response (%)")
```



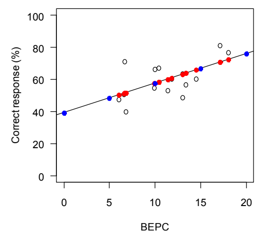

```
# simple linear regression model yi = b0 + b1*BEPCi + ei
fit<-lm(Percentage~BEPC,data)
# add regression line to the plot
abline(fit)
# The summary function test whether the coefficients are significantaly different from zero
summary(fit)
Call:
lm(formula = Percentage ~ BEPC, data = data)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.2246     8.0097   4.897 0.000292 ***
BEPC          1.8383     0.6785   2.709 0.017876 *
Residual standard error: 9.48 on 13 degrees of freedom
Multiple R-squared: 0.3609,     Adjusted R-squared: 0.3117
F-statistic:  7.34 on 1 and 13 DF,  p-value: 0.01788
```

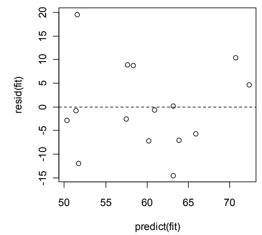
Tests of the regression coefficients

The test reported here is the so-called omnibus test. It compares the model with all predictors to a simpler model with only the intercept. In the case of a simple linear regression, this test is equivalent to the test of the slope

Regression diagnostic tools

```
# coefficient of the regression
coef(fit)
(Intercept)        BEPC
 39.224611    1.838261

# plot data
plot(data$BEPC,data$Percentage,xlim=c(0,20),ylim=c(0,100),
      las=1,xlab="BEPC",ylab="Correct response (%)")
# plot regression line
abline(fit)
# plot fitted values
points(data$BEPC,predict(fit),col="red",pch=16)

# predicted values for new observations
tmp<-data.frame(BEPC=seq(0,20,5))
tmp$Percentage<-predict(fit,newdata=tmp)
points(tmp$BEPC, tmp$Percentage,col="blue",pch=16)

# plot residuals agains predicted values
plot(predict(fit),resid(fit))
abline(h=0,lty=2)

# Standardized residual
standard(fit)
# Studentized residuals
rstudent(fit)
# Cook's distance
cooks.distance(fit)
# Leverage
hatvalues(fit1)
```





Testing the regression coefficients

In multiple regression, the usual test is to check whether the value of the coefficients is statistically different from zero. There is in fact one test per coefficient.

The default way to test the significance of one coeffecient is to compare the full model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_P x_{Pi} + \varepsilon_i$$

with a model that include all predictors except the tested one. For example, to test the coefficient β_1, the residual sum of squares RSS_{full} of the full model is compared to the residual sum of square RSS_0 of the model

$$y_i = \beta_0 \qquad + \beta_2 x_{2i} + \cdots + \beta_P x_{Pi} + \varepsilon_i$$

In other words, the default test will test whether the predictor x_1 explains some significant part of the variance *once all other predictors have already been included in the model.*

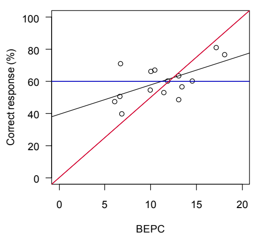
The F ratio

$$F = \frac{(RSS_{full} - RSS_0)/1}{RSS_{full}/df_{error}}$$

where the difference RSS_{full}-RSS_0 is the sum of square explained by the additional parameter in the full model. When testing the value of singular coefficient, the degree of freedom of the numerator is always 1 since the two model differ only by the exclusion of a single predictor.

Most software (including R) report a t value rather than a F ratio but the two tests are strictly equivalent because the F distribution $F_{1,N}$ with the first dof equal to 1 is distributed like the square of the t distribution t_N and the reported t values is equal to the square root of the F ratio.

Example




```
# full model
# yi = b0 + b1*BEPCi + ei
fit<-lm(Percentage~BEPC,data)
# the summary function will make the default test
# for the intercept and the slope
summary(fit)
Call:
lm(formula = Percentage ~ BEPC, data = data)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  39.2246     8.0097   4.897 0.000292
BEPC          1.8383     0.6785   2.709 0.017876
```

```
# The test for the intercept is equivalent to comparing
# the full model with a model without the intercept
#                     yi = b1*BEPCi + ei
fit1<-lm(Percentage~BEPC-1,data)
abline(fit1,col="red")
anova(fit1,fit)
Analysis of Variance Table
Model 1: Percentage ~ BEPC - 1
Model 2: Percentage ~ BEPC
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     14 3323.7
2     13 1168.4  1    2155.4 23.982 0.0002916 ***
# note that sqrt(23.982)=4.897

# The test for the slope is equivalent to comparing
# the full model with a model without the predictor
#                     yi = b0 + ei
fit2<-lm(Percentage~1,data)
abline(fit2,col="blue")
anova(fit2,fit)
Analysis of Variance Table
Model 1: Percentage ~ 1
Model 2: Percentage ~ BEPC
  Res.Df     RSS Df Sum of Sq    F  Pr(>F)
1     14 1828.04
2     13 1168.36  1    659.67 7.34 0.01788 *
# note that sqrt(7.340)=2.709
```

Other tests

note that the summary function applied to the model fit1 will test whether the slope passing through the origin is different from zero

```
>summary(fit1)
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
BEPC    5.002      0.337   14.84 5.85e-10 ***
```

In other words, in this case, the full model is

$$y_i = \beta_1 x_{1i} + \varepsilon_i$$

and the null model has no parameter

$$y_i = \varepsilon_i$$

Similarly, the summary function applied to the model fit2 will test whether the intercept of the intercept only model is different from zero

```
> summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    59.89       2.95    20.3  8.8e-12
```

In this case, the full model is

$$y_i = \beta_0 + \varepsilon_i$$

and the null model has no parameter

$$y_i = \varepsilon_i$$

These two tests are different from the tests that that estimated whether the intercept or slope of the linear regression are different from zero (see previous slide)

The above tests can be reproduced by fitting the model without any parameters

```
fit0<-lm(Percentage~0,data)
```

and comparing the resulting fit:

```
> anova(fit0,fit1)
Model 1: Percentage ~ 0
Model 2: Percentage ~ BEPC - 1
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     15 55624
2     14  3324  1     52301 220.30 5.852e-10
```

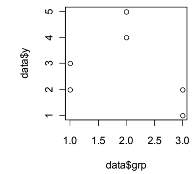
```
> anova(fit0,fit2)
Model 1: Percentage ~ 0
Model 2: Percentage ~ 1
  Res.Df   RSS Df Sum of Sq   F  Pr(>F)
1     15 55624
2     14  1828  1     53796 412 8.8e-12 ***
```


One-way ANOVA

Model

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

where y_{ij} is the jth observation of the ith group and μ_i is the avearge of the ith group



This model is often expressed in terms of the general mean μ and the effects α_i

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where, by definition, the effect is

$$\alpha_i = \mu_i - \mu$$

The sum of the effects is always zero:

$$\sum_i \alpha_i = 0$$

Proof:

$$\sum_i \alpha_i = \sum_i (\mu_i - \mu) = \left(\sum_i \mu_i\right) - k\mu = 0$$

```
# create data set with one factor (grp) with
# 3 levels and 2 observeation (y) per level
data<-data.frame(
            y=c(2,3,5,4,1,2),
            grp=c(1,1,2,2,3,3))
plot(data$grp,data$y)

# note aov makes a linear regression if
# the predictor is not a factor
fit<-aov(y ~ grp,data)
summary(fit)
            Df Sum Sq Mean Sq F value Pr(>F)
grp          1 1.0000  1.0000  0.4068 0.5583
Residuals    4 9.8333  2.4583

# one-way ANOVA
data$grp<-factor(data$grp)
fit<-aov(y ~ grp,data)
summary(fit)
            Df Sum Sq Mean Sq F value  Pr(>F)
grp          2 9.3333  4.6667  9.3333 0.05152 .
Residuals    3 1.5000  0.5000
```


ANOVA revisited

- The tests in the ANOVA can also been seen as a comparison between two models that correspond to the models where the null hypothesis is (simple model) and is not (full model) assumed.
- The F ratio corresponds the sum of the square explained by the additional parameters in the full model over the residual errors of the full model.

$$F = \frac{(RSS_{full} - RSS_{simple})/(df_{full} - df_{simple})}{RSS_{full} / df_{full}}$$

- For example, one-way ANOVAs are conducted to test whether there is a difference between the means of three or more groups. In other words, the null hypothesis is that all means are equal (H_0: $\mu_i = \mu$) or, equivalently, that the effects are null (H_0: $\alpha_i = 0$).
- In other words, the one-way ANOVA compares the variance explained by the model with only the intercept

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

with the variance explained by model where the means for each group can be fitted independently

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \text{or} \quad y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Simple model

$$y_{ij} = \beta_0 + \varepsilon_{ij}$$

in matrix form

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\beta_0] + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

The residual sum of squares is minimized when the parameter is equal to the general mean

$$\beta_0 = \mu.$$

Fit intercept only-model

```
> fit1<-lm(y ~ 1, data)
```

design matrix

```
> model.matrix(fit1)
  (Intercept)
1           1
2           1
3           1
4           1
5           1
6           1
```

design matrix

The esimated or fitted value of the parameter β_0 corresponds to the grand mean

```
> coef(fit1)
(Intercept)
   2.833333
```

Residual sum of square

```
> sum(resid(fit1)^2)
```

or

```
> deviance(fit1)
[1] 10.83333
```

Full model and F test

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

can be represented in matrix form

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

The columns of the design matrix contain **dummy variables** that indicate the group membership

$$y_i = \beta_1 x_{1i} + \beta_2 x_{1i} + \beta_3 x_{3i} + \varepsilon_i$$

where

$$x_{gi} = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in } g^{th} \text{ group } (i = g) \\ 0 & \text{otherwise} \end{cases}$$

It can be shown that the parameters β_i that minimize the residual sum of squares correspond to the group means μ_i.

This model can be fitted with the syntax

```
> fit2<-lm(y ~ grp - 1, data)
> model.matrix(fit2)
  grp1 grp2 grp3
1    1    0    0
2    1    0    0
3    0    1    0
4    0    1    0
5    0    0    1
6    0    0    1
> coef(fit2)
grp1 grp2 grp3
 2.5  4.5  1.5
> deviance(fit2)
[1] 1.5
```

design matrix

The coefficients of this fit correspond as expected to means of each group

residual sum of squares

This model can be compared to the previously fitted model with only the intercept

```
> anova(fit1,fit2)
Analysis of Variance Table
Model 1: y ~ 1
Model 2: y ~ grp - 1
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1      5 10.8333
2      3  1.5000  2    9.3333 9.3333 0.05152 .
```

df of the residual sum of square

residual sum of squares of the two models

number of additional parameters in the second model

$$F = \frac{(10.833 - 1.5)/2}{1.5/3} = 9.333$$

Other tests

Be careful when interpreting ANOVA tables.

```
>anova(fit2)
Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
grp        3 57.500  19.167  38.333 0.006829
Residuals  3  1.500   0.500
```

Using the anova function on the fit2 model does **not** test the equality of the means hypothesis (H_0: $\mu_i = \mu$). In this case, the null hypothesis assumedis that the three coefficients are equal to zero (H_0: $\mu_i = 0$)

In other words, the model

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

is compared to a model without coefficients

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

We can verify this by comparing both models explicitely (see right column)

The model without coefficients can be fitted

```
>fit0<-lm(y ~ 0, data)
> coef(fit0)
numeric(0)
```

Note that the residual sum of squares is the sum of the squares of the observed values

```
> deviance(fit0)
[1] 59
> sum(data$y^2)
[1] 59
```

The comparison between both models produces the desired result

```
anova(fit0,fit)
Analysis of Variance Table
Model 1: y ~ 0
Model 2: y ~ grp - 1
  Res.Df  RSS Df Sum of Sq      F   Pr(>F)
1      6 59.0
2      3  1.5  3      57.5 38.333 0.006829
```

$$F = \frac{(59 - 1.5)/3}{1.5/3} = 38.333$$

Contrasts

The model

$$y_{ij} = \beta_0 + \alpha_i + \varepsilon_{ij}$$

can be expressed in matrix form as

Intercept

Dummy variables indicating class memberships

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

This model cannot be fitted because the model is over-parametrized (design matrix is singular because columns are co-linear). One of the parameters must be removed. For example,

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

contrasts

With this design matrix, it is possible to show that the values of the parameters that minimize the residual sum of squares correspond to

$$\beta_0 = \mu_1$$
$$\beta_1 = \mu_2 - \mu_1$$
$$\beta_2 = \mu_3 - \mu_1$$

Note that the first coefficient (β_0) dost not correspond to average value (μ) because the contrasts are not orthogonal to the intercept column. In linear system theory, the term contrast is usually reserved orthogonal contrasts.

The gx(g-1) "contrast matrix" C specifies how to recode class membership:

$$\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{C}$$

Note that after recoding of the design matrix, testing the null hypothesis that the means are equal (H_0: $\mu_1 = \mu_2 = \mu_3$) corresponds to testing whehter the two last parameters are different from zero (H_0: $\beta_1 = \beta_2 = 0$).

R contrasts

By default, R uses the so call "treatment contrast" for factors and "polynomial contrast" for ordered factor

```
> contrasts(data$grp)
  2 3
1 0 0
2 1 0
3 0 1
```

The design matrix of the fit shows that these contrasts have been used

```
>fit<-aov(y ~ grp,data)
>model.matrix(fit) # design matrix
 (Intercept) grp2 grp3
1           1    0    0
2           1    0    0
3           1    1    0
4           1    1    0
5           1    0    1
6           1    0    1
> tapply(data$y,data$grp,mean) # group means
  1   2   3
2.5 4.5 1.5
> coef(fit) # coefficients
(Intercept)        grp2        grp3
        2.5         2.0        -1.0
```

As expected (see previous slide) the first parameter corresponds to the average of the first group and the other parameters correspond to the difference with this group

The contrast matrix can be defined explicitely. For example, to use the last group as the base group (the so-called SAS contrast)

```
contrasts(data$grp) <-matrix(c(1,0,0,0,1,0),3,2)
contrasts(data$grp)
  [,1] [,2]
1    1    0
2    0    1
3    0    0
```

R offers several functions that return contrast matrix

contrasts	function
treatment	`contr.treatment`
SAS	`contr.SAS`
helmert	`contr.helmert`
polynomial	`contr.poly`

Note that only helmert and polynomial contrasts define contrats orrthognal to the intercept column (this can be easily verified by checking that the sum of the elements in each contrast is zero).

```
contr.helmert(3)
  [,1] [,2]
1   -1   -1
2    1   -1
3    0    2
```

Two-way ANOVA

	B_1	B_2	B_3	Mean
A_1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\bullet}$
A_2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\bullet}$
Mean	$\mu_{\bullet 1}$	$\mu_{\bullet 2}$	$\mu_{\bullet 3}$	μ

- Let y_{ijk} be the k^{th} observation of the i^{th} level of factor A and j^{th} level of factor B.
- Let μ_{ij} be the population mean for the i^{th} level of factor A and j^{th} level of factor B (condition A_iB_j), let $\mu_{i\bullet}$ be the population mean in condition A_i, let $\mu_{\bullet j}$ be the population mean in condition B_j and let μ be the grand mean.
- By definition $\alpha_i = \mu_{i\bullet} - \mu$ is the effect of factor A and $\beta_j = \mu_{\bullet j} - \mu$ is the effect of factor B.

- The structural model of a two-way factorial ANOVA *without interaction* is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

- In absence of interaction, the mean value μ_{ij} in condition (A_iB_j) depends in a *additive* manner on the effect of each condition

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

- The complete model of the two-way factorial ANOVA is

$$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$

where $\alpha\beta_{ij} = \mu_{ij} - (\alpha_i + \beta_j + \mu) = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu$ is the interaction effect. The interaction effect represents the fact that the contribution of one factors depends on the value of the other factor in a non-additive way.

Sum of square types revisited

- In a two-way full factorial ANOVA, three different tests are performed that correspond to the following null hypotheses
  - *main effect of factor A* — H_0: $\alpha_i = 0$
  - *main effect for factor B* — H_0: $\beta_j = 0$
  - *Interaction effect* — H_0: $\alpha\beta_{ij} = 0$

  The null hypotheses only partially define the tests to be conducted.

- **Type I (sequential) SS**
  - *test of factor A:*
    $$y_{ijk} = \mu + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$
  - *test of factor B:*
    $$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
  - *test of interaction:*
    $$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
- **Type II (hierarchical) SS**
  - *test of factor A:*
    $$y_{ijk} = \mu + \beta_i + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \varepsilon_{ijk}$$
  - *test of factor B:*
    $$y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \varepsilon_{ijk}$$
  - *test of interaction:*
    $$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
- **Type III (marginal) SS**
  - *test of factor A:*
    $$y_{ijk} = \mu + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
  - *test of factor B:*
    $$y_{ijk} = \mu + \alpha_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$
  - *test of interaction:*
    $$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$
    $$y_{ijk} = \mu + \alpha_i + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}$$

Example

```
> visits<-read.table("visits.dat",header=TRUE)                  # read data
> visits$age<-ordered(visits$age,c("20-29","30-39","40-49",">50"))   # reorder factors
> v0<-visits[3:nrow(visits),]                                   # remove two first cases
```

Type I SS

```
> fit1<-lm(duration~1,v0)
> fit2<-lm(duration~disease,v0)
> fit3<-lm(duration~disease+age,v0)
> fit4<-lm(duration~disease*age,v0)

> anova(fit1,fit2)
Model 1: duration ~ 1
Model 2: duration ~ disease
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     77 5537.0
2     74 2697.4  3    2839.6 25.967 1.396e-11
> anova(fit2,fit3)
Model 1: duration ~ disease
Model 2: duration ~ disease + age
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     74 2697.4
2     71 1535.5  3    1161.9 17.908 9.34e-09
> anova(fit3,fit4)
Model 1: duration ~ disease + age
Model 2: duration ~ disease * age
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     71 1535.51
2  62   926.27  9    609.25 4.5311 0.0001331

> summary(aov(duration~disease*age,v0))
Response: duration
            Df  Sum Sq Mean Sq F value    Pr(>F)
disease      3 2839.64  946.55 63.3574 < 2.2e-16
age          3 1161.89  387.30 25.9238 5.464e-11
disease:age  9  609.25   67.69  4.5311 0.0001331
Residuals   62  926.27   14.94
```

Type III SS

```
> fit1<-lm(duration~age+disease:age,v0)
> fit2<-lm(duration~disease+disease:age,v0)
> fit3<-lm(duration~disease+age,v0)
> fit4<-lm(duration~disease*age,v0)

> anova(fit1,fit4)
Model 1: duration ~ age + disease:age
Model 2: duration ~ disease * age
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     62 926.27
2     62 926.27  0 1.137e-13
> anova(fit2,fit4)
Model 1: duration ~ disease + disease:age
Model 2: duration ~ disease * age
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     62 926.27
2     62 926.27  0 1.137e-13
> anova(fit3,fit4)
Model 1: duration ~ disease + age
Model 2: duration ~ disease * age
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     71 1535.51
2     62  926.27  9    609.25 4.5311 0.0001331

> library(car)
> Anova(aov(duration~disease+age,v0),type="III")
Anova Table (Type III tests)
Response: duration
             Sum Sq Df  F value    Pr(>F)
(Intercept) 29261.2  1 1353.001 < 2.2e-16 ***
disease      2928.9  3   45.142 < 2.2e-16 ***
age          1161.9  3   17.908  9.34e-09 ***
Residuals    1535.5 71
```