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 $\hat{y}_i$:

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

• 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 find out the values of the coefficients $\beta_0$ and $\beta_1$ so that the regression line fits best the data points ($\beta_0$ is the intercept and $\beta_1$ 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 $\epsilon_i = y_i - \hat{y}_i$

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

• To minimize the residual sum of squares $\text{SSE}$ we differentiate $\text{SSE}$ w.r.t $\beta_0$ and $\beta_1$

$\frac{\partial \text{SSE}}{\partial \beta_0} = -2 \sum (y_i - \hat{y}_i - \beta_0 - \beta_1 x_i) = 0$

$\frac{\partial \text{SSE}}{\partial \beta_1} = -2 \sum x_i (y_i - \hat{y}_i - \beta_0 - \beta_1 x_i) = 0$

solving for $\beta_0$ and $\beta_1$ yields

$\beta_0 = \frac{\sum y_i}{N} - \beta_1 \frac{\sum x_i}{N}$

$\beta_1 = \frac{\sum xy_i - \frac{\sum x_i \sum y_i}{N}}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}$

Proof:

- $\sum_{x_i < x_j} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{x_i < x_j} x_i y_i - \sum_{x_i < x_j} x_i \bar{y} - \sum_{x_i < x_j} \bar{x} y_i + \sum_{x_i < x_j} \bar{x} \bar{y}$

- $\sum_{x_i < x_j} x_i y_i - \sum_{x_i < x_j} x_i \bar{y} = \sum_{x_i < x_j} x_i (y_i - \bar{y})$

- $\sum_{x_i < x_j} x_i \bar{y} - \sum_{x_i < x_j} \bar{x} y_i = \sum_{x_i < x_j} \bar{x} (y_i - \bar{y})$

- $\sum_{x_i < x_j} \bar{x} \bar{y} = \frac{N}{2} \sum_{x_i < x_j} \bar{x} \bar{y}$

Linear regression

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) $\{X_1, X_2, ..., X_P\}$

• The multiple line regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_P x_{iP} + \epsilon_i$ ($i = 1, 2, ..., N$)

where $x_{ij}$, and $x_{ij}$ are the values of the predictor variables and $y_i$ the observed value of the independent variable for the $i$th observation, $N$ represents the number of observations in the data set.

• The parameters of the model are the regression coefficients, $\beta_1, \beta_2, ..., \beta_P$

• The linear model has one random effect, the error term $\epsilon_i$. The error term is assumed to follow a normal distribution $\mathcal{N}(0, \sigma^2)$. Moreover, the error terms for the various observations are assumed to be uncorrelated ($\text{cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$).

\[ \text{cov}(y, y') = \text{cov}(\epsilon, \epsilon') = 0, \ (\#) \]
Linear models in matrix form

In matrix form,

\[
\mathbf{y} = \mathbf{X}\hat{\mathbf{\beta}} + \mathbf{e}
\]

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

The vector of errors \( \mathbf{e} = (e_1, \ldots, e_n) \) is assumed to follow a \( n \)-variable multivariate-normal distribution with a \( n \times n \) covariance matrix \( \Sigma \)

\[
\Sigma = \begin{bmatrix}
\text{var}(e_1) & \text{cov}(e_1, e_2) & \cdots & \text{cov}(e_1, e_n) \\
\text{cov}(e_2, e_1) & \text{var}(e_2) & \cdots & \text{cov}(e_2, e_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{cov}(e_n, e_1) & \text{cov}(e_n, e_2) & \cdots & \text{var}(e_n)
\end{bmatrix}
\]

The classic assumptions of homogeneity of the variances (\( \text{var}(e_i) = \sigma^2 \)) and uncorrelated observations (\( \text{cov}(e_i, e_j) = 0 \) for \( i \neq j \)) imply that \( \Sigma = \sigma^2 \mathbf{I}_n \).

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):

\[
e_i = y_i - \hat{y}_i
\]

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

\[
SSE = \sum_{i} e_i^2 = \sum_{i} (y_i - \hat{y}_i)^2
\]

Linear model theory

- The parameters \( \mathbf{\beta} \) of the linear model

\[
\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{e}
\]

are obtained by that minimizing the sum of squares

\[
\min_{\hat{\mathbf{\beta}}} \| \mathbf{y} - \mathbf{X}\hat{\mathbf{\beta}} \| \quad \text{where} \quad \| \mathbf{e} \|^2 = \sum e_i^2
\]

differentiating w.r.t. \( \beta_i \), equating by 0, and solving for \( \beta_i \) yields

\[
\hat{\beta} = (\mathbf{X}^\prime\mathbf{X})^{-1}\mathbf{X}^\prime\mathbf{y}
\]

- The predicted values are

\[
\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{\beta}} = \mathbf{X}(\mathbf{X}^\prime\mathbf{X})^{-1}\mathbf{X}^\prime\mathbf{y}
\]

Properties of the hat matrix
- \( \text{tr}(\mathbf{H}) = \text{tr}(\mathbf{A}) = \text{nb. of parameters} \)
- idempotence (\( \mathbf{H}^2 = \mathbf{H} \))
- \( \text{tr}((\mathbf{X}^\prime\mathbf{X})^{-1}) = \text{p} \) (nb. of parameters)

Im

Syntax
- \texttt{lm(formula, data, subset)}
- \texttt{aov(formula, data, subset)}

Usage
\texttt{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 \texttt{aov} may provide a more convenient interface for these).
# Linear regression example

```r
# 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),
                   Linear regression example
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))

# single linear regression model y_i = \beta_0 + \beta_1 x_i + \epsilon_i
fit <- lm(Percentage ~ BEPC, data)
```

# plot
```r
plot(data$BEPC, data$Percentage, xlim = c(0,20), ylim = c(0,100),
     las = 1, xlab = "BEPC", ylab = "Correct response (%)")
```

# coefficient of the regression
```r
coef(fit) # Intercept \beta_0, \beta_1 BEPC
39.2246 1.8383
```

# plot data
```r
abline(fit) # plot regression line
```

# plot fitted values
```r
points(data$BEPC, predict(fit), col = "red", pch = 16)
```

# predicted values for new observations
```r
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
```r
plot(predict(fit), resid(fit)) # plot residuals
abline(h = 0, lty = 2)
```

# The summary function test whether the coefficients are significantly different from zero
```r
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
```

# Testing the regression coefficients

The F ratio

\[ F = \frac{(RSS_{\text{full}} - RSS_{\text{ref}})}{RSS_{\text{ref}}}/1 \]

where the difference \( RSS_{\text{full}} - RSS_{\text{ref}} \) 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 \( F \) with the first dof equal to 1 is distributed like the square of the \( t \) distribution with the residual degrees of freedom. 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, the test is equivalent to the test of the slope.

# Regression diagnostic tools

## Coefficient of the regression
```r
# fit <- lm(Percentage ~ BEPC, data)
# coef(fit) # Intercept \beta_0, \beta_1 BEPC
39.2246 1.8383
```

## Data
```r
plot(data$BEPC, data$Percentage, xlim = c(0,20), ylim = c(0,100),
     las = 1, xlab = "BEPC", ylab = "Correct response (%)")
```

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

## Tests of the regression coefficients

### Standardized residual
```r
rstandard(fit) # Standardized residual
```

### Studentized residuals
```r
rstudent(fit) # Studentized residuals
```

### Cook's distance
```r
cooks.distance(fit) # Cook's distance
```

### Hat values
```r
hatvalues(fit) # Hat values
```

### Leverage
```r
leverage(fit) # Leverage
```

# An Introduction to Statistics with R
Gabriel Baud-Bovy - IIT 2010
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.

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

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

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \]

and the null model has no parameter \( \beta_1 \)

\[ y_{ij} = \varepsilon_{ij} \]

These two tests are different from the tests 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

\[ \text{fitdev} \left< = \text{lm} \left( \text{Percentage} \sim 0 \right) \right. \]

and comparing the resulting fit:

\[ \text{anova} \left( \text{fitdev}, \text{fit1} \right) \]

\[ \text{anova} \left( \text{fitdev}, \text{fit2} \right) \]

\[ \text{Res.Df} \quad \text{RSS} \quad \text{DF} \quad \text{Sum of Sq} \quad \text{F} \quad \text{Pr(>F)} \]

\[ 1 \quad 15 \quad 55624 \quad 1 \quad 1828 \quad 412 \quad 8.8 \times 10^{-12} \quad * * * \]

\[ 1 \quad 14 \quad 3324 \quad 1 \quad 52301 \quad 220.3 \quad 5.85 \times 10^{-10} \quad * * * \]

One-way ANOVA

Model

\[ y_{ij} = \mu_i + \varepsilon_{ij} \]

where \( y_{ij} \) is the \( j \)th observation of the \( i \)th group and \( \mu_i \) is the average of the \( i \)th group

This model is often expressed in terms of the general mean \( \mu \) and the effects \( \alpha_i \)

\[ y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \]

where, by definition, the effect is

\[ \alpha_i = \mu_i - \mu \]

The sum of the effects is always zero:

\[ \sum \alpha_i = 0 \]

Proof:

\[ \sum \alpha_i = \sum (\mu_i - \mu) = \left( \sum \mu_i \right) - n \mu = 0 \]

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.

\[ F = \frac{RSS_{null} - RSS_{same}}{RSS_{null}/df_{null}} \]

- 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, 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 \( y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \)

Simple model

\[ y_{ij} = \beta_0 + \varepsilon_{ij} \]

in matrix form

\[
\begin{bmatrix}
  y_{11} \\
  y_{12} \\
  y_{13} \\
  y_{14} \\
  y_{15}
\end{bmatrix} =
\begin{bmatrix}
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0
\end{bmatrix} +
\begin{bmatrix}
  \varepsilon_{11} \\
  \varepsilon_{12} \\
  \varepsilon_{13} \\
  \varepsilon_{14} \\
  \varepsilon_{15}
\end{bmatrix}
\]

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

\[ \beta_0 = \mu. \]
**Full model and F test**

\[ y_i = \beta_0 + \beta_i + \epsilon_i \]

can be represented in matrix form

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  1 & \alpha_1 & \alpha_2 \\
  1 & \alpha_1 + \alpha_2 & \alpha_2 \\
  1 & \alpha_1 & \alpha_1 + \alpha_2 \\
  \vdots & \vdots & \vdots \\
  1 & \alpha_1 + \alpha_2 \end{bmatrix} 
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

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

\[ y_i = \beta_{i0} + \beta_{i1} + \epsilon_i \]

where

\[ \alpha_i = \begin{cases} 1 & \text{if } i^{th} \text{ obs is in } g^{th} \text{ group (} i = g \) } \\
0 & \text{otherwise}
\]

It can be shown that the parameters \( \beta_i \) that minimize the residual sum of squares correspond to the group means \( \mu_i \).

An Introduction to Statistics with R

Gabriel Baud-Bovy - IIT 2010

**Contrasts**

The model

\[ y_i = \beta_0 + \alpha_i + \epsilon_i \]

can be expressed in matrix form as

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  1 & 0 & 0 \\
  1 & 1 & 0 \\
  1 & 0 & 1 \\
  \vdots & \vdots & \vdots \\
  1 & 1 & 1
\end{bmatrix} 
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

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,

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  1 & \alpha_1 & \epsilon_1 \\
  1 & \alpha_1 + \alpha_2 & \epsilon_2 \\
  1 & \alpha_1 & \alpha_1 + \alpha_2 \\
  \vdots & \vdots & \vdots \\
  1 & \alpha_1 + \alpha_2 \end{bmatrix} 
\begin{bmatrix}
  \beta_0 \\
  \beta_1 \\
  \beta_2 \\
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

An Introduction to Statistics with R

Gabriel Baud-Bovy - IIT 2010

**Other tests**

Be careful when interpreting ANOVA tables.

\[
\begin{align*}
\text{anova(fit2)}
\text{Res.Df} & \quad \text{RSS} & \quad Df & \quad \text{Sum of Sq} & \quad F & \quad \text{Pr(>F)} \\
1 & \text{6} & 10.8333 & 2 & 9.3333 & 9.3333 & 0.05152 \\
\end{align*}
\]

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

In other words, the model

\[ y_i = \beta_0 + \epsilon_i \]

can be represented in matrix form

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n
\end{bmatrix} = 
\begin{bmatrix}
  1 & 1 & 1 \\
  1 & 1 & 1 \\
  1 & 1 & 1 \\
  \vdots & \vdots & \vdots \\
  1 & 1 & 1
\end{bmatrix} 
\begin{bmatrix}
  \beta_0 \\
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

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

\[
\text{deviance(fit0,fit)}
\]

The comparison between both models produces the desired result

\[
\begin{align*}
\text{anova(fit0,fit)}
\text{Res.Df} & \quad \text{RSS} & \quad Df & \quad \text{Sum of Sq} & \quad F & \quad \text{Pr(>F)} \\
1 & \text{6} & 3.6667 & 2 & 9.3333 & 9.3333 & 0.05152 \\
0 & \text{5} & 0.8333 & 1 & 2.5000 & 2.5000 & 0.12500 \\
\end{align*}
\]

The model without coefficients can be fitted

\[
\text{fitt0}<-	ext{lm}(-y \sim 1, \text{data)}
\]

The contrast matrix can be defined explicely.

\[
\text{contrasts(data$grp)} <- \text{matrix(c(1,0,0,0,1,0),3,2)}
\]

R offers several functions that return contrast matrix

\[
\text{contrasts(data$grp)} <- \text{matrix(c(1,0,0,0,1,0),3,2)}
\]

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

An Introduction to Statistics with R

Gabriel Baud-Bovy - IIT 2010

**R contrasts**

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

\[
\text{contrasts(data$grp)} <- \text{matrix(c(1,0,0,0,1,0),3,2)}
\]

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

\[
\text{anova(fit1,lfit2)}
\]

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

An Introduction to Statistics with R

Gabriel Baud-Bovy - IIT 2010
Two-way ANOVA

<table>
<thead>
<tr>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>B11</td>
<td>B12</td>
<td>B13</td>
</tr>
<tr>
<td>A2</td>
<td>B21</td>
<td>B22</td>
<td>B23</td>
</tr>
</tbody>
</table>

- The structural model of a two-way factorial ANOVA without interaction is
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]

- In absence of interaction, the mean value \( \mu_j \) in condition \((A, B)\) depends on the effect of each condition
  \[ \mu_j = \mu + \alpha_i + \beta_j \]

Let \( \mu_{ij} \) be the \( ij \)th observation of the \( j \)th level of factor A and \( i \)th level of factor B.

Let \( \mu_j \), \( \mu_i \), and \( \mu \) be the population mean for the \( j \)th level of factor A and \( i \)th and \( j \)th level of factor B (condition \( AB \)), let \( \mu_{ij} \) be the population mean in condition \( A_iB_j \), and let \( \mu \) be the grand mean.

By definition \( \alpha_i - \mu = \mu_i - \mu \) is the effect of factor A and \( \beta_j - \mu = \mu_j - \mu \) is the effect of factor B.

The complete model of the two-way factorial ANOVA is
\[ y_{ij} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ij} \]
where \( \alpha\beta_{ij} = \mu_{ij} - (\alpha_i + \beta_j + \mu) \) 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.

### Example

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

> # read data
> v0<-visits[3:nrow(visits),] # remove two first cases

Type I (sequential) SS
- test of factor A:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of factor B:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of interaction:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]

Type II (hierarchical) SS
- test of factor A:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of factor B:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of interaction:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]

Type III (marginal) SS
- test of factor A:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of factor B:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]
- test of interaction:
  \[ y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \]