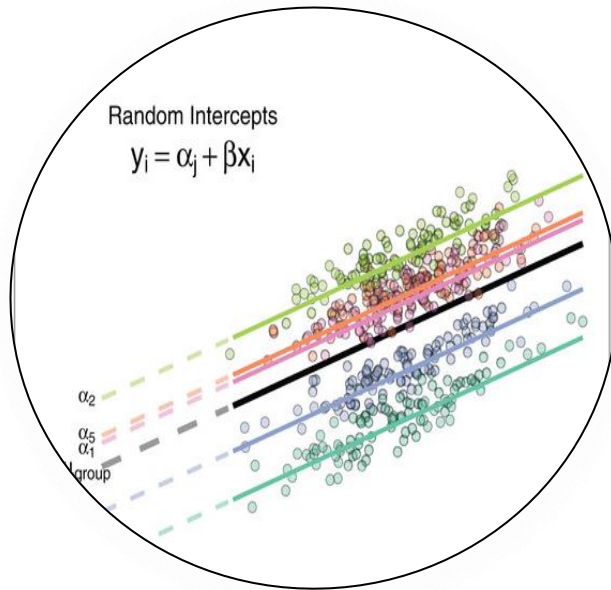


PhD course Mixed Linear Models

Session 1: Linear Models

Gerrit Gort
Biometris, Wageningen University

June 10-11-12, 2025



Practical information

- Location: Forum building (Droevendaalse steeg 2, Wageningen), room B0106
- Date: June 10-12, 2025
- Time
 - 9:00-17:00 h
 - breaks roughly at 10:45-11:00, 12:30-13:30, 15:15-15:30
 - coffee/tea and lunch included
- Alternation of lectures and practicals using R
 - Bring along your laptop and have R installed!

Programme Phd course Mixed Linear Models (MLM)

1. Refresher of Linear Models
2. Gentle introduction to Mixed Models
3. General theory of Mixed Models, examples of variance components models
4. Examples of variance components models (continued)
5. Estimation and testing in Mixed Models
6. Repeated measurements with examples in R

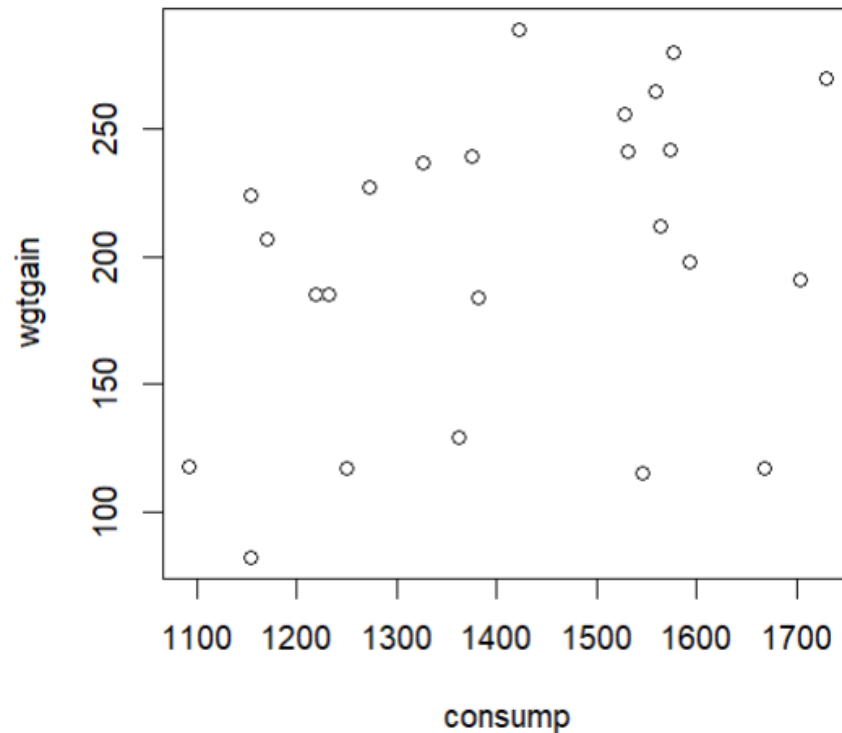
Models in Statistics

- **Model**: abstraction of a real phenomenon or process isolating the aspects relevant to a particular question
- **Statistical model**:
 - model with stochastic component, containing unknown parameters (constants), to be estimated from data.
 - describes distributional properties of response variables, thereby decomposing variability in known and unknown sources.
 - mathematical language used to describe reality
- **Example**: $y_i = \beta_0 + \beta_1 x_i + e_i$
- What is the name of this model?

Example 1

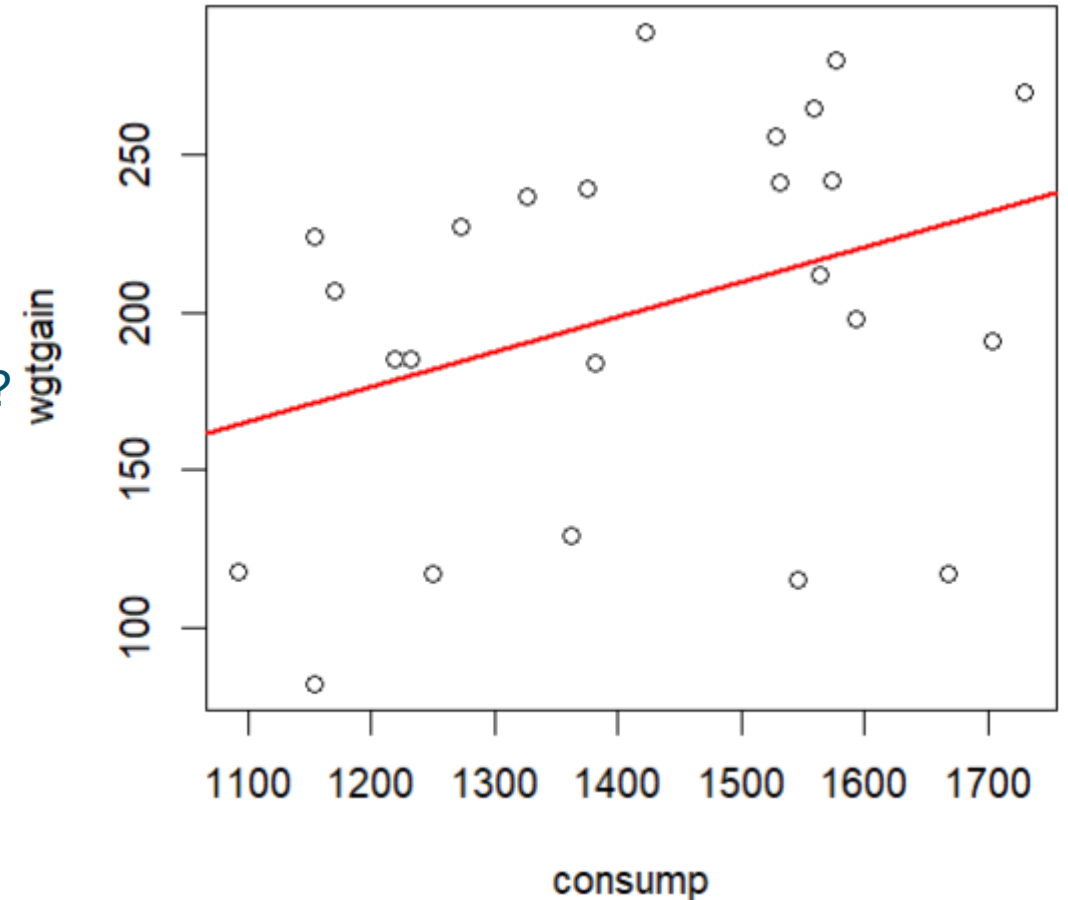
consump	wgtgain
1155	224
1326	237
1423	289
1559	265
1576	280
1528	256
1092	118
1154	82
1703	191
1250	117
1546	115
1667	117
1573	242
1381	184
1730	270
1363	129
1593	198
1564	212
1532	241
1375	239
1220	185
1170	207
1232	185
1273	227

- 24 guinea pigs consume given amounts of clover (in grams); weight gain (in grams) is measured
- How is weight gain related to amount of consumed clover?
- y_i = weight gain and x_i = amount of clover
- Scatter plot



Simple example of statistical model

- Example: **simple linear regression model**
 - statistical model $y_i = \beta_0 + \beta_1 x_i + e_i$ used to describe relationship between y_i (weight gain) and x_i (amount of clover)
 - line is “best fitting” straight line for these data: $\hat{\beta}_0 = 43.66$, $\hat{\beta}_1 = 0.11$
 - but which relationship is “true” relationship?
- First example of Linear Model



Parts of simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{Two parts:}$$

1) Variables:

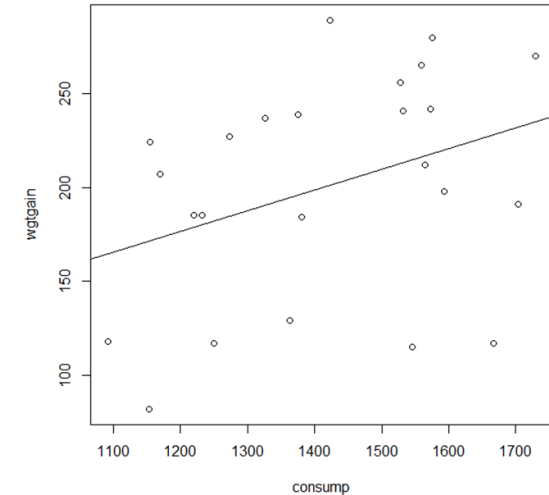
- variable y_i = **dependent variable** = **response variable** ; here: weight gain of animal i ; index i for 24 individual animals: here $i = 1, \dots, 24$
- variable x_i = **regressor** = **independent variable** = **explanatory variable** ; here: amount of consumed clover for animal i
- variable e_i (unobserved) = **error term** for animal i

2) Parameters = regression coefficients = unknown constants (in population)

- β_0 = **intercept** = expected value of y if $x = 0$
- β_1 = **slope** = expected change in y if $x = 0$ increases with 1 unit

■ Parameters are combined linearly → **Linear Model**

■ Random y_i (not known in advance), but we act as if we know x_i



Systematic and random part

- Linear model

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

response = systematic part + random part

- Systematic part: $\beta_0 + \beta_1 x_i = \mu_i = E(y_i)$ is expected value of y_i (in pop) given x_i

- Model assumptions for random part e_i

- e_i independent
- e_i normally distributed
- e_i constant variance
- e_i expected value zero

Identically, independently
distributed

- Model for e_i can be written as $e_i \sim N(0, \sigma^2), i.i.d.$

- What values of e_i are to be expected if e.g. $\sigma = 1$?

Model in matrix notation - 1

- Linear model for 24 guinea pigs ($i = 1, \dots, 24$)

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

- First guinea pig:

$$y_1 = \beta_0 + \beta_1 x_1 + e_1$$

- Second guinea pig:

$$y_2 = \beta_0 + \beta_1 x_2 + e_2$$

- Third guinea pig:

$$y_3 = \beta_0 + \beta_1 x_3 + e_3$$

.....

- Twenty-fourth guinea pig:

$$y_{24} = \beta_0 + \beta_1 x_{24} + e_{24}$$

Model in matrix notation - 2

- Linear model for 24 guinea pigs ($i = 1, \dots, 24$)

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

- Collect 24 weight gains into vector

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{23} \\ y_{24} \end{bmatrix}$$

- Likewise collect 24 consumed clover amounts into vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{23} \\ x_{24} \end{bmatrix}$$

- And 24 errors in error vector

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix}$$

- We also need vector with 24 ones (why and why green?)

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix}$$

- And finally small vector with regression coefficients

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Model in matrix notation - 3

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

- Piece it all together:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{23} \\ y_{24} \end{bmatrix} \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{23} \\ x_{24} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix}$$

matrix multiplication!



$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{23} \\ y_{24} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{23} \\ 1 & x_{24} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix} \quad \Rightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{23} \\ y_{24} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{23} \\ 1 & x_{24} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix}$$

X-matrix = model matrix = design matrix

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

Model for random part

- Errors collected in error vector

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix}$$

- Model assumptions for errors

- e_i independent
- e_i constant variance
- e_i normally distributed
- e_i expected value zero

Errors e_i assumed **constant variance**

- $\text{var}(e_i) = 0$

Errors e_i assumed **independent**

- e_i and e_j **independent** $\rightarrow e_i$ and e_j **uncorrelated**: correlation coefficient $\rho(e_i, e_j) = 0$ for any i and j
- If corr coef = 0, then **covariance** $\text{cov}(e_i, e_j) = 0$
- Recall that correlation is unscaled covariance:
 $\rho(e_i, e_j) = \text{cov}(e_i, e_j) / \sigma(e_i) \sigma(e_j)$
- Covariance measures strength of linear relationship between two random variables

Model for random part in matrix notation (1)

- Variances and covariances of errors are collected in **variance-covariance matrix V**

$$V = \text{Var}(\mathbf{e}) = \text{Var} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & \dots & 0 & \sigma^2 \end{bmatrix}$$

$\text{cov}(e_1, e_2) = 0$
all covariances zero,
uncorrelated errors

$\text{var}(e_{24}) = \sigma^2$
all error variances equal

- Variances on diagonal places, covariances on off-diagonal places.
- $\text{cov}(e_i, e_j) = \text{cov}(e_j, e_i) : V$ is symmetrical matrix

Model for random part in matrix notation (2)

$$V = \text{Var}(\mathbf{e}) = \text{Var} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{23} \\ e_{24} \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 & 0 \\ 0 & \sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 & 0 \\ 0 & 0 & \dots & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} = \sigma^2 \mathbf{I}_{24}$$

- \mathbf{I}_{24} is called unit-matrix (of order 24 for guinea pig example)
- Have solid understanding of variance-covariance matrix as this is key to mixed models: mixed models allow variance-covariance matrix to deviate from the unit matrix: it allows **analysis of dependent data!**

Simplest case possible . . . one sample case

■ Example 2

- In random sample of geese from large colony in Russian Arctic ($n=20$) weights of geese are measured. Observed weights (in grams):
1740, 1583, 1646, 2010, 1686, 1690, 1599, 1841, 1536, 1782, 1477, 1714, 1548, 1714, 1509, 1754, 1715, 1997, 1862, 1745
- We are interested in mean weight in the colony, not so much in this small sample. Name the mean weight in colony μ : μ is parameter of interest. What value might μ have?
- Let's name the weights y_i ($i = 1..20$); we believe that y_1, \dots, y_{20} form random sample from normal distribution with mean μ and variance σ^2 (with μ and σ^2 unknown).
- Written as statistical model: $y_i = \mu + e_i$ with $e_i \text{ iid } \sim N(0, \sigma^2)$
 - **Systematic part** very simple: μ (compare with $\beta_0 + \beta_1 x_i$)
 - **Random part** same as in example of simple linear regression!

■ Exercise 1

Write the model for this one sample situation in matrix notation.

Next case: two independent samples

■ Example 3: effect of lime on soil pH

- 10 soils, 5 without lime, 5 with lime, pH measured

```
# Soils without and with lime
twosamp <- read.table("lime2.dat", col.names=c("lime", "ph"))
# reads a raw ASCII datafile
twosamp
```

```
> twosamp
  lime  ph
1    N 5.735
2    N 5.770
3    N 5.730
4    N 5.735
5    N 5.750
6    Y 5.845
7    Y 5.880
8    Y 5.865
9    Y 5.875
10   Y 5.865
```

■ Statistical model: $y_{ij} = \mu_i + e_{ij}$ ($i = 1, 2$; $j = 1, \dots, 5$)

- Systematic part: without lime $E(y_{1j}) = \mu_1$ = expected soil pH for soils without lime
with lime $E(y_{2j}) = \mu_2$ = expected soil pH for soils with lime
- Random part: $e_i \sim iid N(0, \sigma^2) \rightarrow$ same as before!

Two independent samples: matrix notation

- Statistical model: $y_{ij} = \mu_i + e_{ij}$ ($i = 1, 2; j = 1, \dots, 5$)
- Use dummy variables x_1 and x_2 to code for two levels of lime:
 - $y = \mu_1 x_1 + \mu_2 x_2 + e$
 - linear model (in μ 's)!
 - dummies = dummy variables = indicator variables

lime	x1	x2	pH
N	1	0	5.735
N	1	0	5.770
N	1	0	5.730
N	1	0	5.735
N	1	0	5.750
Y	0	1	5.845
Y	0	1	5.880
Y	0	1	5.865
Y	0	1	5.875
Y	0	1	5.865

Matrix notation:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{24} \\ y_{25} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{21} \\ e_{22} \\ e_{23} \\ e_{24} \\ e_{25} \end{pmatrix} \quad \text{or} \quad y = X\beta + e$$

Two independent samples: random part in matrix notation

- As always assumptions for errors: $e_{ij} \sim N(0, \sigma^2)$ iid
 - so, independent (uncorrelated); constant variance
 - diagonal variance-covariance matrix

$$\text{Var}[\mathbf{e}] = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \sigma^2 I_{10}$$

Two independent samples: different parameterization

Exercise 2

Suppose we write the model for the pH of 5 soils without and 5 soils with lime in a different way:

$$y = \beta_0 + \beta_1 x_1 + e$$

with x_1 the same dummy variable as before.

Questions:

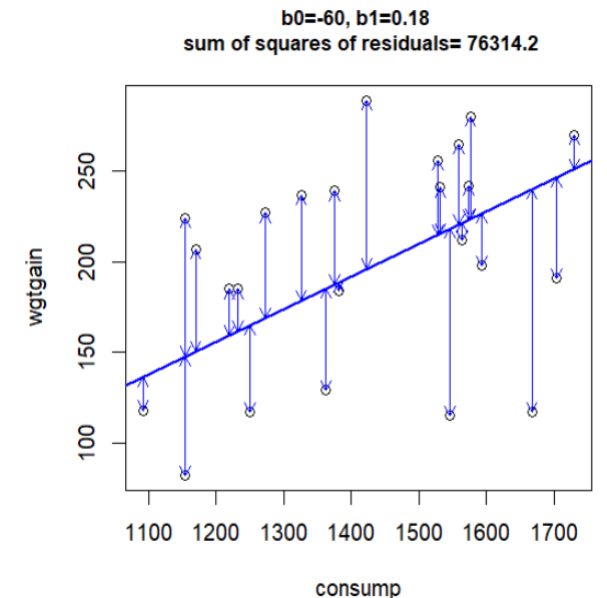
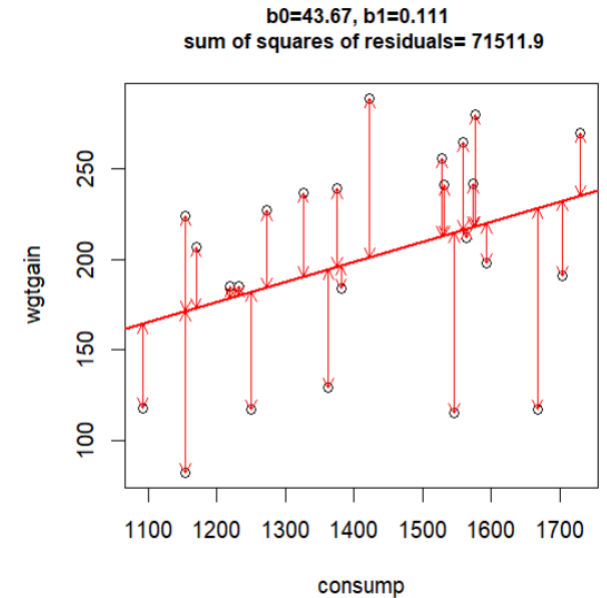
- Write down the model in matrix notation.
- Give the interpretation of the parameters β_0 and β_1 .
- This is the simple linear regression model! What are the usual names for the parameters β_0 and β_1 ?

Statistical inference

- Statistical model contains unknown parameters (constants).
- Given data, parameters are to be estimated: “statistical inference”
- Two types of statistical inference:
 - Estimation (point estimation and interval estimation: confidence intervals)
 - Hypothesis testing
- Estimation in Linear Models? Least Squares Estimation (LSE)
 - More general estimation principle is Maximum Likelihood (MLE)
 - LSE is special case of MLE (assuming normal distribution)
 - Later in Mixed Models MLE and Restricted Maximum Likelihood (REML)

Least Squares Estimation

- Linear model: y_i is sum of systematic part (e.g. $\beta_0 + \beta_1 x_i$) and random part (e_i)
- For which value of parameters (β 's) does model fit best? Let's call estimates b_0 and b_1 .
- Best-fitting = sum of squares of vertical distances between observations and line minimal.
- Points on regression line have y-coordinate $b_0 + b_1 x_i$:
predictions $\hat{y}_i = b_0 + b_1 x_i$
- Difference between observed and predicted y_i are **residuals**:
 $r_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$
- Least-squares estimation: find those values for b_0 and b_1 so that sum of squares of residuals $\sum_{i=1}^{24} r_i^2$ is **minimized**.



Example LSE

- Example 4 One independent sample. As in example 2, random sample of geese with weights y_1, \dots, y_{20} from a population with expected weight μ (which is parameter of interest). Give least-squares estimator of μ .

- Linear model: $y_i = \mu + e_i$
- Systematic part of model simply is μ , same for all n observations.
- Predicted values: $\hat{y}_i = \hat{\mu}$, residuals: $r_i = y_i - \hat{y}_i = y_i - \hat{\mu}$.
- LSE $\hat{\mu}$ (for μ): find $\hat{\mu}$ such that $\sum_{i=1}^{20} r_i^2 = \sum_{i=1}^{20} (y_i - \hat{\mu})^2$ is minimal.
- High school math: differentiate w.r.t. $\hat{\mu}$ and equate to 0:

$$\frac{df(\hat{\mu})}{d\hat{\mu}} = -2 \sum_{i=1}^{20} (y_i - \hat{\mu}) = 0$$
$$\sum_{i=1}^{20} (y_i - \hat{\mu}) = 0 \Leftrightarrow \sum_{i=1}^{20} y_i = 20\hat{\mu} \Leftrightarrow \hat{\mu} = \sum_{i=1}^{20} y_i / 20 \Rightarrow \hat{\mu} = \bar{y}_i$$

- In other words: LSE of population mean μ is simply sample mean! Best-fitting = sum of squares of vertical distances between observations and line minimal.
- Notice: for LSE normal distribution is not needed!

LSE in simple linear regression (1)

■ Example 5 Simple linear regression

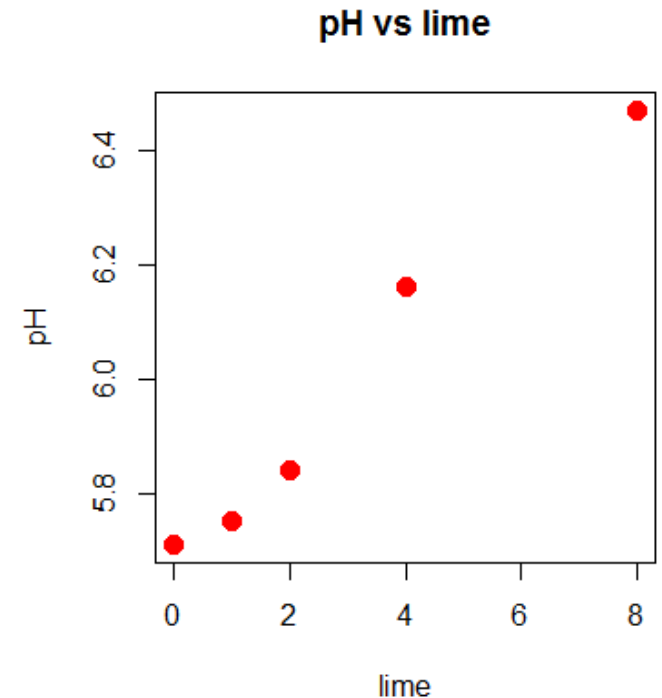
- Lime applied to soils, soil pH observed; lime rates 0, 1, 2, 4 and 8.
- Model: $y = \beta_0 + \beta_1 x_1 + e$
- What would be your guess for b_0 and b_1 ?

rate	pH
0	5.71
1	5.75
2	5.84
4	6.16
8	6.47

- Model in matrix notation:

$$y = X\beta + e$$

$$y = \begin{bmatrix} 5.71 \\ 5.75 \\ 5.84 \\ 6.16 \\ 6.47 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 8 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$



LSE in simple linear regression (2)

rate	pH
0	5.71
1	5.75
2	5.84
4	6.16
8	6.47

- What are LS estimators for β_0 and β_1 ?
- Same principle: minimize $\sum_{i=1}^{24} r_i^2 = \sum_{i=1}^{24} (y_i - (b_0 + b_1 x_i))^2$
- In matrix notation: minimize $f(b_0, b_1) = (y - Xb)^T(y - Xb)$
- Differentiating and equating to 0 results in **normal equations**:

matrix multiplication:
row \times column = sum of squares

$$X^T X b = X^T y$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 & 8 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 4 & 8 \end{bmatrix} \begin{bmatrix} 5.71 \\ 5.75 \\ 5.84 \\ 6.16 \\ 6.47 \end{bmatrix}$$

e.g. $15 = 1 \times 0 + 1 \times 1 + 1 \times 2 + 1 \times 4 + 1 \times 8$

e.g. $93.83 = 0 \times 5.71 + 1 \times 5.75 + 2 \times 5.84 + 4 \times 6.16 + 8 \times 6.47$

Essentially, solve system of (here) 2 linear equations:

$$\begin{cases} 5b_0 + 15b_1 = 29.93 \\ 15b_0 + 85b_1 = 93.83 \end{cases}$$

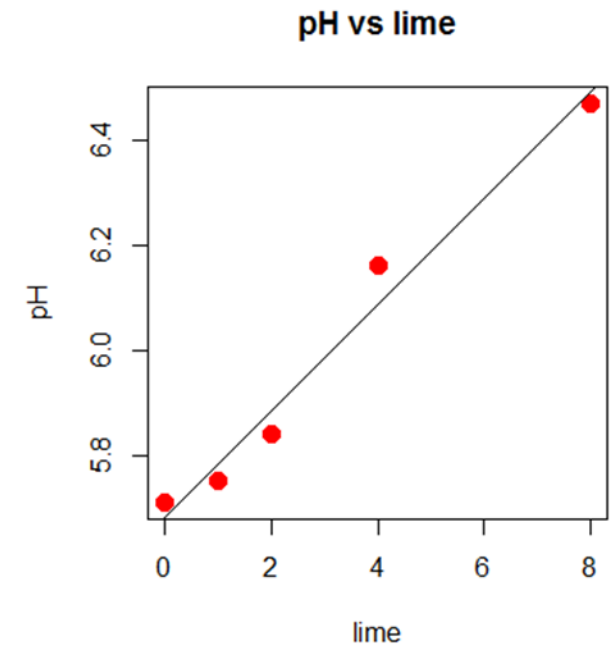
$$\begin{aligned} &\Leftrightarrow \begin{bmatrix} 5 & 15 \\ 15 & 85 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 29.93 \\ 93.83 \end{bmatrix} \\ &\Leftrightarrow \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 85 \end{bmatrix}^{-1} \begin{bmatrix} 29.93 \\ 93.83 \end{bmatrix} \\ &\Leftrightarrow \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \frac{1}{5 \times 85 - 15 \times 15} \begin{bmatrix} 85 & -15 \\ -15 & 5 \end{bmatrix} \begin{bmatrix} 29.93 \\ 93.83 \end{bmatrix} \\ &\Leftrightarrow \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} 0.425 & -0.075 \\ -0.075 & 0.025 \end{bmatrix} \begin{bmatrix} 29.93 \\ 93.83 \end{bmatrix} = \begin{bmatrix} 5.683 \\ 0.101 \end{bmatrix} \end{aligned}$$

LSE in linear models

- Least squares estimator is solution of **normal equations**: $b = (X^T X)^{-1} X^T y$
- Matrix $(X^T X)^{-1}$ plays important role in linear models.

- Exercise 3

- Use R to fit simple linear regression model.
- Locate least squares estimates of intercept and slope and compare with guesses.
- Plot pH versus the fitted regression line in the scatterplot
- Use functions in R to get the bits and pieces of the normal equations and solution: $X^T X$, $X^T y$, $(X^T X)^{-1}$, b



Regression output from R

```
> limepH <- data.frame(rate=c(0,1,2,4,8), pH=c(5.71,5.75,5.84,6.16,6.47) )
> regr.object <- lm(pH ~ rate, data=limepH) # Function lm for fitting linear model
> summary(regr.object)
Call: lm(formula = pH ~ rate, data = limepH)
```

Residuals:

1	2	3	4	5
0.027	-0.034	-0.045	0.073	-0.021

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.683000	0.037032	153.46	6.1e-07
rate	0.101000	0.008981	11.24	0.00151

Residual standard error: 0.0568 on 3 degrees of freedom
Multiple R-squared: 0.9768, Adjusted R-squared: 0.9691
F-statistic: 126.5 on 1 and 3 DF, p-value: 0.001508

```
> anova(regr.object)
```

Analysis of Variance Table

Response: pH

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rate	1	0.40804	0.40804	126.46	0.001508 **
Residuals	3	0.00968	0.00323		

Exercise 4 LSE in case of two independent samples

- Back to two independent samples of example 3: effect of lime on soil pH; 5 soils with lime, 5 soils without; 2 dummies coding for the two groups.

Written as linear model: $y = \mu_1 x_1 + \mu_2 x_2 + e$.

Regression without intercept!

- Fit model with R (using both dummies as regressors and without intercept). Find least squares estimates of μ_1 and μ_2 . What are they?
- Rewrite model as: $y = \beta_0 + \beta_1 x_1 + e$. Regression with intercept!
 - Fit model with R (using only the first dummy as regressor and with intercept). Find least squares estimates of β_0 and β_1 . What are they?

lime	x1	x2	pH
N	1	0	5.735
N	1	0	5.770
N	1	0	5.730
N	1	0	5.735
N	1	0	5.750
Y	0	1	5.845
Y	0	1	5.880
Y	0	1	5.865
Y	0	1	5.875
Y	0	1	5.865

Accuracy of estimation

- How well is parameter estimated by LS estimator?
- In estimation 2 types of inaccuracy: **bias** and **variance**:
standard error² of estimator = bias² + variance
- Least-squares estimators are **unbiased** (bias=0), so standard error² of estimator = variance of estimator

Standard error of mean

- Famous formula for standard error of mean:

$$se(\bar{y}) = \sigma / \sqrt{n}$$

- only for uncorrelated (independent) observations!
- by calculation rules for variances:

$$se(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \sqrt{\text{var}\left(\frac{1}{n} \sum_{i=1}^n y_i\right)} = \sqrt{\frac{1}{n^2} \text{var}\left(\sum_{i=1}^n y_i\right)} = \sqrt{\frac{1}{n^2} \sum_{i=1}^n \text{var } y_i} = \frac{1}{n} \sqrt{n\sigma^2} = \sigma / \sqrt{n}$$

“variance of sum” = “sum of variances”
only holds for uncorrelated data!

does **not** hold in mixed models!

Standard error of difference of means

- Two independent samples from two populations with sample sizes m and n :
 $y_{11}, \dots, y_{1m} \text{ iid } \sim N(\mu_1, \sigma^2)$ and $y_{21}, \dots, y_{2n} \text{ iid } \sim N(\mu_2, \sigma^2)$
 - variances assumed equal in two populations
 - interest in difference $\mu_1 - \mu_2$
 - LSE of $\mu_1 - \mu_2$ is $\bar{y}_{1.} - \bar{y}_{2.}$
- Standard error of difference of two means (for independent samples):

$$se(\bar{y}_{1.} - \bar{y}_{2.}) = \sigma \sqrt{1/m + 1/n}$$

- s.e.d. = $se(\bar{y}_{1.} - \bar{y}_{2.}) = \sqrt{var(\bar{y}_{1.} - \bar{y}_{2.})}$
 - with $var(\bar{y}_{1.} - \bar{y}_{2.}) = var(\bar{y}_{1.}) + var(\bar{y}_{2.})$ (independent samples!)
 - $= \sigma^2/m + \sigma^2/n$ (compare 1-sample situation; common variance of x_i and y_j)
 - $= \sigma^2(1/m + 1/n)$
 - so, $se(\bar{y}_{1.} - \bar{y}_{2.}) = \sigma \sqrt{1/m + 1/n}$

Accuracy of least squares estimator

- Recall: least squares estimator is $b = (X^T X)^{-1} X^T y$
- b is vector, containing individual coefficients, e.g. intercept, slope in case of simple linear regression
- Need **variance-covariance matrix of b** :

$$\text{Var}(b) = (X^T X)^{-1} \sigma^2$$

- variances of individual regression coefficients (e.g. b_0 and b_1) on the diagonal
- covariances between individual regression coefficients on off-diagonal places
- $se(b_1) = \sqrt{\text{var}(b_1)}$

needed e.g. for confidence intervals, hypothesis tests of individual coefficients (using t-distribution)

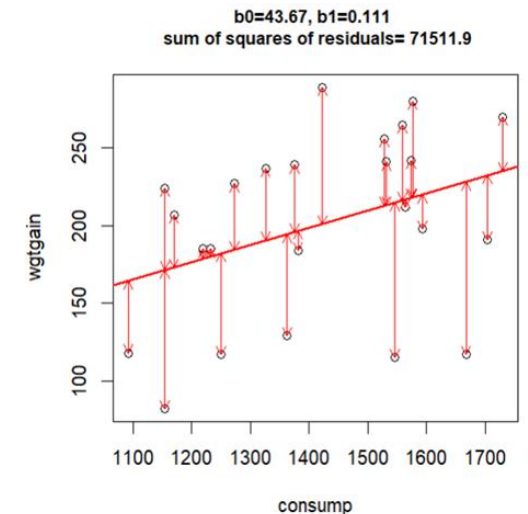
needed e.g. for linear combinations of coefficients (differences, predictions like $b_0 + 5b_1$)

Estimating error variance

- Least squares procedure delivers estimates of regression coefficients.
- Precisions needed for next steps in statistical inference (hypothesis tests, confidence intervals)!
- General formula: $Var(b) = (X^T X)^{-1} \sigma^2$
 - boils down to $se(\bar{y}) = \sqrt{\frac{1}{n}} \sigma$ (one-sample), $se(\bar{y}_1. - \bar{y}_2.) = \sqrt{\frac{1}{m} + \frac{1}{n}} \sigma$ (two independent samples)
 - need estimate of **error variance** σ^2 !
- Estimator of error variance σ^2 comes from (minimized) residual sum of squares from LS:
$$\hat{\sigma}^2 = SSE / dfe$$

(sort of) average squared deviation

 - residual degrees of freedom $dfe = n - p$
 - p is #parameters in model, e.g. simple LR: $p = 2$ (ic and slope)



Example 6 Standard errors of coefficients, error variance

```
> regr.object <- lm(pH ~ rate, data=limepH)
> summary(regr.object)
Call: lm(formula = pH ~ rate, data = limepH)
```

Residuals:

1	2	3	4	5
0.027	-0.034	-0.045	0.073	-0.021

Standard error of slope

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.683000	0.037032	153.46	6.1e-07
rate	0.101000	0.008981	11.24	0.00151

$\hat{\sigma}$

Residual standard error: 0.0568 on 3 degrees of freedom
Multiple R-squared: 0.9768, Adjusted R-squared: 0.9691
F-statistic: 126.5 on 1 and 3 DF, p-value: 0.001508

```
> anova(regr.object)
```

Analysis of Variance Table

Response: pH

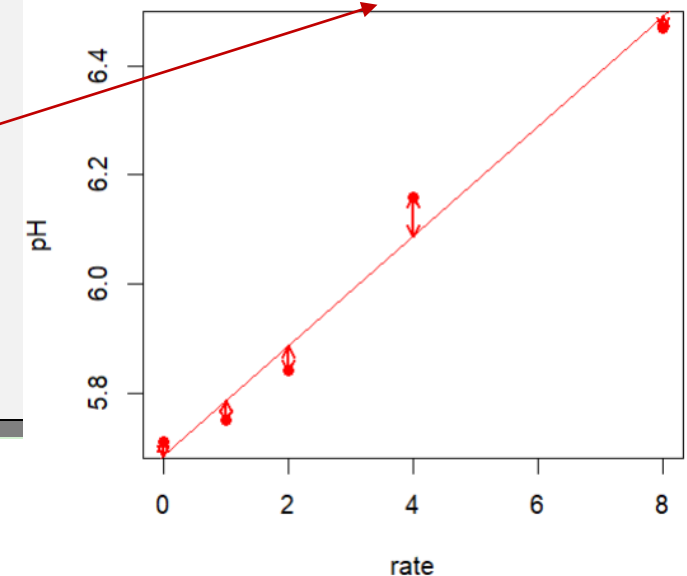
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rate	1	0.40804	0.40804	126.46	0.001508
Residuals	3	0.00968	0.00323		

Sum of Squares for Error (SSE)
(minimized in LSE)

$\hat{\sigma}^2 = \text{Mean Square Error} = \text{SSE} / (5 - 2)$

$dfe = \text{residual degrees of freedom} (n - p = 5 - 2 = 3)$

pH vs lime amount
sum of squares of residuals= 0.00968



Example 7 Statistical inference: t-test for slope in small steps

```
> regr.object <- lm(pH ~ rate, data=limepH)
> summary(regr.object)
Call: lm(formula = pH ~ rate, data = limepH)

Residuals:
    1     2     3     4     5 
0.027 -0.034 -0.045  0.073 -0.021 

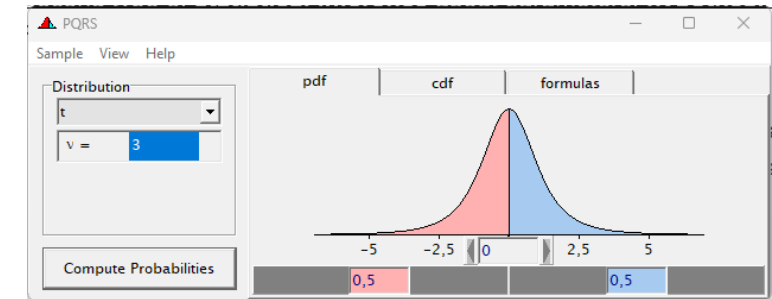
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.683000   0.037032   153.46  6.1e-07
rate          0.101000   0.008981   11.24  0.00151

Residual standard error: 0.0568 on 3 degrees of freedom
Multiple R-squared:  0.9768, Adjusted R-squared:  0.9691
F-statistic: 126.5 on 1 and 3 DF, p-value: 0.001508
```

1) $H_0: \beta_1 = 0$ versus $H_0: \beta_1 \neq 0$

2) Test statistic: $t = \frac{b_1 - 0}{se(b_1)}$

3) If H_0 true, $t \sim t_3$ -distribution



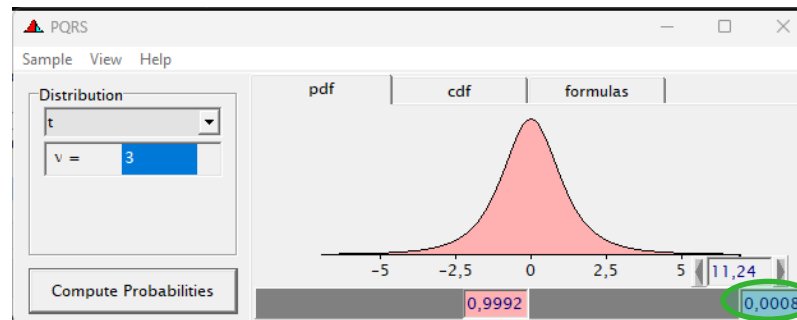
4) If H_0 false, larger or smaller values of t expected

5) Outcome of test statistic $t = \frac{0.101 - 0}{0.008981} = 11.24$

6) P-value $P = 2 \times P(t_3 \geq 11.24) = 0.00151$

7) Testing at $\alpha = 0.05$, $0.00151 < 0.05$ so reject H_0

8) Conclusion: slope deviates from zero



Example 8 Statistical inference: confidence interval of slope

```
> regr.object <- lm(pH ~ rate, data=limepH)
> summary(regr.object)
Call: lm(formula = pH ~ rate, data = limepH)

Residuals:
    1     2     3     4     5 
0.027 -0.034 -0.045  0.073 -0.021 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.683000   0.037032  153.46  6.1e-07
rate         0.101000   0.008981   11.24  0.00151

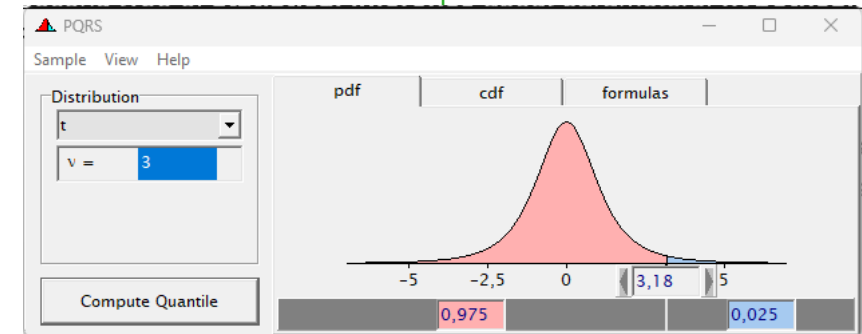
Residual standard error: 0.0568 on 3 degrees of freedom
Multiple R-squared:  0.9768, Adjusted R-squared: 0.9691
F-statistic: 126.5 on 1 and 3 DF, p-value: 0.001508

> confint(regr.object)
              2.5 %   97.5 %
(Intercept) 5.56515  5.8009
rate        0.07242  0.1296
```

- 95% confidence interval for β_1 :

$$b_1 \pm t_3(0.975) \times se(b_1)$$

$$0.101 \pm 3.18 \times 0.008981$$



- Lower bound for β_1 : 0.0724
- Upper bound for β_1 : 0.1296

Exercise 5 t-test for difference (2 independent samples)

Guinea pig example: guinea pigs were fed with clover from 4 types of soils. We select the guinea pigs fed with clover from the two loamy type of soils.

Question: is there a weight gain difference between guinea pigs fed with clover from the two types of loamy soils?

Two independent samples:

- Guinea pigs fed with "loamy1" clover: y_{11}, \dots, y_{16} ; $y_{1j} \sim N(\mu_1, \sigma^2)$ $j = 1, \dots, 6$
- Guinea pigs fed with "loamy2" clover: y_{21}, \dots, y_{26} ; $y_{2j} \sim N(\mu_2, \sigma^2)$ $j = 1, \dots, 6$

Parameter of interest: $\mu_1 - \mu_2$ = difference in expected weight gains

Testing situation: $H_0: \mu_1 - \mu_2 = 0$, $H_a: \mu_1 - \mu_2 \neq 0$

$$\text{Test statistic } t = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s\hat{e}d} = \frac{\bar{y}_1 - \bar{y}_2 - 0}{s_p \sqrt{\frac{1}{6} + \frac{1}{6}}}$$

$$\text{with } s_p^2 = \frac{1}{6+6-2} \left(\sum_{i=1}^6 (y_{1i} - \bar{y}_1)^2 + \sum_{j=1}^6 (y_{2j} - \bar{y}_2)^2 \right) \text{ ("pooled variance")}$$

Exercise 5 t-test for difference (2 independent samples), cont.

- Read the data from a csv-file (comma separated variable format) into the program.
- Select only the observations for guinea pigs fed with clover from loamy soils.
- Print the dataset on the screen.
- Do the independent samples t-test, assuming equal variances. Locate in the output, if present at all:
 - estimated difference in weight gain
 - s.e.d.
 - outcome of test statistic
 - degrees of freedom of t-distribution
 - P-value
- What is the conclusion?

```
> # Independent samples t-test
> t.test(wgtgain ~ soiltype, data=loamysoils, var.equal=T)

      Two Sample t-test
data:  wgtgain by soiltype
t = 2.3555, df = 10, p-value = 0.04026
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.846735 102.486598
sample estimates:
mean in group loam1 mean in group loam2
      258.5000      205.8333
```

Exercise 6 Independent samples t-test, using a linear model

We can do exactly the same t-test using software for linear models (in case of R: function `lm` = linear model). We now recognize that we have a linear model. Try it yourself and locate in the output the P-value that we saw earlier.

We have to tell the software that soiltype is **qualitative** (in R: treat soiltype as **factor**), we have to specify a model ← formula (in R: `wgtgain ~ soiltype`), and we want to see the estimated coefficients (in R: use `summary()` function).

For a factor, R will automatically make dummy variables and will put them in the model (i.e. they become columns in the X matrix). If you want to have full control yourself, you may make dummies yourself, and take these “manually” as regressors in the linear model

- Write down (again) the linear model that R uses for this situation.
- Check that you get identical results for the t-test of the slope as with previous approach. T-tests to compare means of independent samples can simply be obtained through linear model!

```
> # Check if soiltype is factor = qualitative explanatory variable:
> is.factor(loamysoils$soiltype)

> # Fit linear model
> lm.object <- lm(wgtgain ~ soiltype, data=loamysoils)
>
> # Ask for regression coefficients and their t-tests
> coef(summary(lm.object))
```

Paired samples t-test

- Sometimes pairs of observations are obtained: two observations originating from the same (experimental) unit.
- E.g. weight of animal before and after a treatment
- Data from the same pair are **not independent** → **independence assumption for error fails**
- Ordinary linear model cannot directly be employed.
- Simple solution: calculate difference for pair, as employ one-sample t-test for difference!
- More general approach is to employ **mixed model**, that allows observations to be correlated! See later.

Example 9 Comparing nested models with F-test

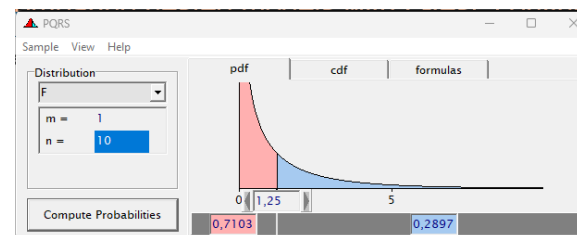
- Testing a null hypothesis can be seen as comparison of two **nested** models: **FULL MODEL** and **REDUCED MODEL**, induced from full model by assuming that null hypothesis is true.
- Compare weight gains in 2 groups of guinea pigs, fed with 2 types of clover.
 - Full model**: $y_{ij} = \mu_i + e_{ij}$ (two different μ 's for loamy soils type 1 and 2)
 - Reduced model**: $y_{ij} = \mu + e_{ij}$ (one common μ for the 2 groups)
- Compare the fit of the two models by comparing the SSE's (error sums of squares):
 - $SSE_{FM} = \sum_{i,j} (y_{ij} - \bar{y}_i)^2$
 - $SSE_{RM} = \sum_{i,j} (y_{ij} - \bar{y})^2$
- Note: $SSE_{FM} \leq SSE_{RM}$ (why?)
- Combine SSE_F and SSE_R into F-statistic: $F = \frac{(SSE_{RM} - SSE_{FM}) / (2 - 1)}{SSE_{FM} / (12 - 2)}$
- This procedure gives square of t-test statistic if we have 1 d.f. for numerator!
- F has F-distribution with 1 and 10 d.f.
- Normality of original Y's is needed.

$$H_0: \mu_1 = \mu_2$$

Difference of model fits of RM and FM

Difference of # model parameters under FM and RM

Pooled error variance s_p^2



Exercise 7 Comparing nested models with F-test

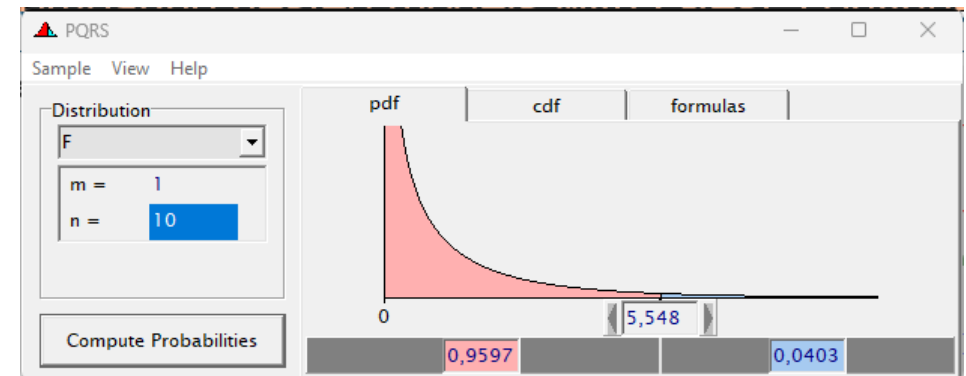
Test $H_0: \mu_1 = \mu_2$ vs $H_0: \mu_1 \neq \mu_2$ using F-test

- Specify models under H_0 and H_a .
- Give SSE_{RM} and SSE_{FM} using ANOVA tables from FM and RM.
- Construct F-test statistic by hand and check that F is square of t found earlier.
- Get P-value using e.g. PQRS.
- Use `anova()` function to compare the two nested models.

```
> anova(FM,RM)
Analysis of Variance Table

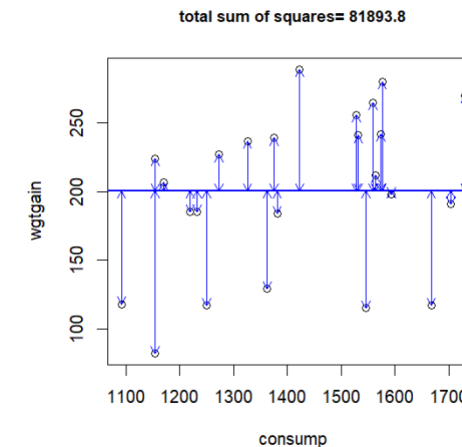
Model 1: wgtgain ~ soiltype
Model 2: wgtgain ~ 1
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     10 14998  1    -8321.3 5.5482 0.04026
2     11 23320  2    -8321.3 5.5482 0.04026
```

```
> # Fit FULL MODEL
> FM <- lm(wgtgain ~ soiltype, data=loamysoils)
> anova(FM)
Analysis of Variance Table
              Df Sum Sq Mean Sq F value Pr(>F)
soiltype      1  8321.3   8321.3   5.5482 0.04026
Residuals    10 14998.3   1499.8
>
> # Fit REDUCED MODEL = intercept-only model!!
> RM <- lm(wgtgain ~ 1, data=loamysoils)
> anova(RM)
Analysis of Variance Table
              Df Sum Sq Mean Sq F value Pr(>F)
Residuals    11  23320    2120
```

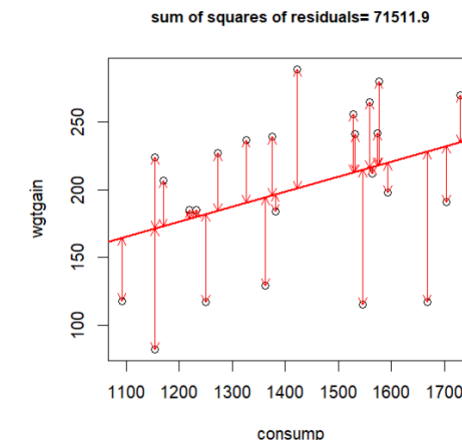


Partitioning of variability through sums of squares

- To judge how well the model explains the response, current model is compared with “null model”: model with only intercept (i.e. mean). The idea is to remove all explanatory variables so that only intercept remains.
- Intercept-only model is nested within the current model (if current model contains intercept)! So, we can use F-test to test all coefficients.
- Error sum of squares of intercept-only model is called **Total sum of squares** (around the mean)
$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$
 - SST sometimes called Corrected total sum of squares (corrected for mean), to discriminate from uncorrected total sum of squares
- Error sum of squares of current model (minimized in LS) is $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ with \hat{y}_i predictions from current model.



Guineapig
example 1



Partitioning of variability through sums of squares (2)

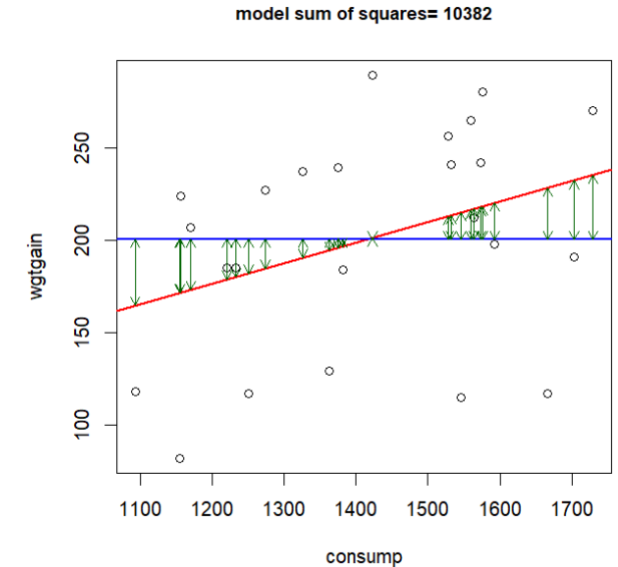
- Difference $SST - SSE$ is model sum of squares SSM or regression SS .
 - can also be calculated as $\sum(\hat{y}_i - \bar{y})^2$ (not straightforward)
- ANOVA table shows split of sums of squares:

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

$$SST = SSM + SSE$$

- ANOVA table contains:
 - source of variation
 - degrees of freedom (df)
 - sum of squares (SS)
 - mean square (MS = SS/df)
 - F statistic $F = MSM / MSE$
 - P-value

- R's function `anova` skips SST:



Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Corrected Model	1	10382	10382	3.19	0.088
Residuals	22	71512	3251		
Corrected Total	23	81894			

```
> anova(reg.result)
Analysis of Variance Table
```

```
Response: wgtgain
      Df Sum Sq Mean Sq F value Pr(>F)
consump  1  10382    10382   3.19  0.088 .
Residuals 22  71512     3251
```

R-square and R-square adjusted

- Coefficient of determination R^2 = fraction explained variance
 - Which part of variability of the response variable (around mean) is explained by the linear model?
 - $R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$
 - R^2 will never decrease if regressor is added, irrespective of importance of the regressor.
- Adjusted R^2
 - $R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{s^2_{current\ model}}{s^2_{intercept\ only}}$
 - p = number of regression coefficients
 - inclusion of a regressor into the model leads to increase of R-square adjusted **only** if the regressor is "important", in the sense that the estimate of the error variance s^2 decreases by inclusion of regressor.

Linear Models

- $y = X\beta + e$
- Assumptions systematic part
 - ✓ linear in parameters
- Assumptions random part
 - ✓ independent errors
 - ✓ constant variance (homoscedasticity)
 - ✓ normal distribution
- Linear models comprise (mainly historically grown):
 - Linear regression models (regressors quantitative)
 - Analysis of variance models (regressors qualitative)
 - Analysis of covariance models (both quantitative and qualitative regressors)

Linear models: simple linear regression

■ Example 10 Lime

- continuous response variable $y = \text{soil pH}$
- 1 quantitative regressor $x = \text{rate} = \text{amount of lime}$
- Model: $y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, \dots, 25)$
+ usual error assumptions: normal, constant variance, independent, mean zero

```
> limedata <- read.table("lime.prn", header=TRUE)
> limeAL <- limedata[limedata$lime=="AL",] # AL subset only
> limeAL.reg <- lm(pH ~ rate, data=limeAL) # Fit model
>
> anova(limeAL.reg) # gives ANOVA table
Analysis of Variance Table

Response: pH
          Df Sum Sq Mean Sq F value    Pr(>F)
rate        1  1.46804   1.46804    714.41 < 2.2e-16
Residuals   23  0.04726   0.00205

> # corrected total sum of squares: SST:
> (SST <- sum((limeAL$pH - mean(limeAL$pH))^2) )
[1] 1.515304
```

rate	pH
0	5.735
0	5.77
0	5.73
0	5.735
0	5.75
1	5.845
1	5.88
1	5.865
1	5.875
1	5.865
2	5.98
2	5.965
2	5.975
2	5.975
2	6.075
4	6.18
4	6.06
4	6.135
4	6.205
4	6.095
8	6.415
8	6.475
8	6.495
8	6.455
8	6.41

Linear models: simple linear regression (2)

- Understand all bits and pieces?

```
> summary(limeAL.reg)
```

```
Call:  
lm(formula = pH ~ rate, data = limeAL)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.06348	-0.03078	-0.00145	0.02287	0.12287

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.780775	0.013216	437.40	<2e-16
rate	0.085675	0.003205	26.73	<2e-16

```
Residual standard error: 0.04533 on 23 degrees of freedom
```

```
Multiple R-squared: 0.9688, Adjusted R-squared: 0.9675
```

```
F-statistic: 714.4 on 1 and 23 DF, p-value: < 2.2e-16
```

Linear models: check model assumptions

- Model: $y_i = \beta_0 + \beta_1 x_i + e_i$
- Assumptions random part: errors e_i assumed to
 - have constant variance
 - be independent
 - be normally distributed
- Assumptions systematic part: $\beta_0 + \beta_1 x_i$ is linear in x (rate)
(\rightarrow at any x error mean = 0)
- Violations particularly show in residuals: $r_i = y_i - \hat{y}_i = \text{observed } y - \text{predicted } y$
- Check residuals!

Regression diagnostics

- **Raw residuals:** $r_i = y_i - \hat{y}_i$ = observed y – predicted y
 - Residual tells whether model predicts individual observation well, extreme in y -space
- **Leverages:** h_i diagonal elements of “hat matrix”: $H = X(X^T X)^{-1} X^T$ ($Hy = \hat{y}$, “puts hat” on y)
 - Leverage tells whether individual observation is potentially influential, extreme in X -space; larger than $2 \times p/n$ is suspicious
- **Studentized residuals:** $sr_i = r_i / \sqrt{(1 - h_i)s^2}$
 - Studentized residuals have variance 1 (but are not independent), outside interval $(-3, 3)$ is suspicious
- Many more diagnostics, e.g. Cook’s distance

Residual plots

- Graphical checks:
 - scatterplot (studentized) residuals versus predicted value (no loudspeaker shape)
 - scatterplot (studentized) residuals versus regressor (no banana shape)
 - normal QQ-plot: plot residuals against corresponding quantiles of standard normal distribution
- Hardly residual checks available for independence assumption! Should follow from description of study design (random sampling, randomization)

Linear models: multiple linear regression

- Example 11 Milk yield
 - continuous response variable $y = \text{milk yield of cows}$
 - 2 quantitative regressor $x_1 = \text{feed intake}$, $x_2 = \text{energy density of feed}$
 - model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (i = 1, \dots, 50)$
 - + usual error assumptions: normal, constant variance, independent, mean zero
- Interpretation of regression coefficient changes: e.g. β_1 is expected change in y for unit change in x_1 **keeping the other regressors constant**

Linear models: multiple linear regression (2)

- Understand all bits and pieces?

```
> summary(FM)

Call:
lm(formula = milky ~ intake + density, data = my)

Residuals:
    Min       1Q   Median       3Q      Max
-0.80424 -0.35421 -0.07703  0.33352  0.91728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.59392    0.37086   4.298 8.61e-05
intake       0.03056    0.00236  12.950 < 2e-16
density     2.28566    0.23963   9.538 1.44e-12

Residual standard error: 0.4793 on 47 degrees of freedom
Multiple R-squared:  0.8462, Adjusted R-squared:  0.8397
F-statistic: 129.3 on 2 and 47 DF,  p-value: < 2.2e-16
```

Regression output: overall ANOVA table

- Program like SAS produces overall ANOVA table, showing combined explanatory contribution of intake and density.
- Such ANOVA table is not easily obtained in R: R splits model sum of squares directly into components. For the model SS we need to add the SS for intake and density.
- To get the model SS, we fit FM and RM, and compare these with the anova() function. Notice that the change in residual SS by going from RM to FM is 59.4, i.e. the model SS.
- To create an ANOVA table as SAS produces, some R coding is needed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	59.41564	29.70782	129.33	<.0001
Error	47	10.79577	0.22970		
Corrected Total	49	70.21140			

```
> FM <- lm(milky ~ intake + density, data=my)
> anova(FM)
```

Analysis of Variance Table

Response: milky

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
intake	1	38.519	38.519	167.694	< 2.2e-16
density	1	20.897	20.897	90.976	1.44e-12
Residuals	47	10.796	0.230		

```
> FM <- lm(milky ~ intake + density, data=my)
> RM <- lm(milky ~ 1, data=my)
> anova(RM, FM)
```

Model 1: milky ~ 1

Model 2: milky ~ intake + density

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	49	70.211				
2	47	10.796	2	59.416	129.33	< 2.2e-16

Partitioning model sum of squares in multiple linear regression

- Model sum of squares is now due to **multiple** regressors.
- But: what is the effect of **individual** regressors? Is it possible to split up the model sum of squares in components, corresponding to separate regressors?
- Yes, but unfortunately in different ways:
 - extra SS = sequential SS = **type I SS** (SAS terminology)
 - regressors enter model one after the other
 - change in SSE (or SSM) by entering a regressor is the extra SS of this regressor
 - order dependent and not unique!
 - sum of sequential SS's = model SS
 - partial SS = **type II SS** (SAS terminology)
 - individual regressors are dropped from the full model (i.e. corrected for all other regressors)
 - change in SSE by dropping a regressor from the full model is partial SS of this regressor
 - not order dependent and unique
 - but in general sum of partial SS \neq model SS
- If regressors are orthogonal (uncorrelated), sequential SS and partial SS's are identical.

Multicollinearity

- Correlated regressors: estimation of individual regression coefficients impaired: s.e. inflated
- $se_{b_i} = \sigma \sqrt{\frac{1}{\sum_j (x_{ij} - \bar{x}_i)^2} \frac{1}{(1 - R_i^2)}}$: s.e. increases if
 - error standard deviation σ increases
 - variance of individual x_i decreases
 - $1/(1 - R_i^2)$ increases, with R_i^2 the R^2 of the regression of x_i on all other x 's.
- **Variance Inflation Factor** $VIF = 1/(1 - R_i^2)$
 - No multicollinearity $VIF = 1 \Leftrightarrow R_i^2 = 0$, i.e. multiple correlation 0
 - If VIF is large (rule of thumb is >10), situation becomes problematic.
- Most extreme case: multiple correlation 1 or -1, complete confounding
- In experiments multicollinearity usually does not occur, as values of x_i are defined by experimenter \rightarrow correlations designed to be 0 or small.
- Be aware of the problem in observational studies.
- Multicollinearity in milkyield example?

Linear model: ANOVA model

- **ANOVA model** = Analysis of Variance model = LM containing qualitative regressors only: to compare means between groups
- Qualitative regressor = **factor** = classification variable = grouping variable
- Just LM: $y = X\beta + e$ the usual error assumptions!
- Model matrix X is different now: it (usually) contains **dummies**=indicator variables=0/1 variables, coding for factor levels
- In R, dummies will be created automatically if grouping variable is defined as **factor**
- ANOVA = regression on dummies
- Simplest ANOVA model: one-way ANOVA model = ANOVA model with single factor

Linear models: one-way ANOVA model

■ Example 12 Lime again

- continuous response variable $y = \text{soil pH}$
- 3 lime treatments: rate 1=no lime, rate 2=lime at low rate, rate 3=lime at high rate.
- Treat rate as qualitative regressor = factor
- Three groups with population pH means μ_1, μ_2, μ_3

■ Two ways to write model:

- **Means model** (parameters are means)
 - $y_{ij} = \mu_i + e_{ij} \quad (i = 1,2,3; j = 1,2,3,4,5)$
 - model without intercept
- **Effects model** (parameters are **effects**: differences between groups)
 - $y_{ij} = \mu + \alpha_i + e_{ij} \quad (i = 2,3; j = 1,2,3,4,5)$
 - model with intercept
 - default in `lm()`

rate	pH	x1	x2	x3
1	5.735	1	0	0
1	5.770	1	0	0
1	5.730	1	0	0
1	5.735	1	0	0
1	5.750	1	0	0
2	5.845	0	1	0
2	5.880	0	1	0
2	5.865	0	1	0
2	5.875	0	1	0
2	5.865	0	1	0
3	5.980	0	0	1
3	5.965	0	0	1
3	5.975	0	0	1
3	5.975	0	0	1
3	6.075	0	0	1

Model
matrix
 X

x0	x2	x3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1

Overparameterization

- Effects model $y_{ij} = \mu + \alpha_i + e_{ij}$ with 3 dummies is **overparameterized**: 4 parameters but only 3 groups!
- Restriction on parameters needed
 - corner stone restriction
 - $\alpha_1 = 0$ first level reference, default in R's `lm()`
 - obtained by removing dummy x1 (but any other could be removed too)
 - μ is mean of reference group (=group 1)
 - α_i is difference of mean of group i with reference group 1
 - sum-to-zero restriction
 - $\alpha_1 + \alpha_2 + \alpha_3 = 0$
 - obtained by using regressors with values 1,0,-1
 - μ is overall mean (mean of means)
 - α_i is difference of mean of group i with overall mean

x0	x2	x3
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	1	0
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	0	1

x0	x1	x2	x3
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1

z0	z2	z3
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	0	1
1	0	1
1	-1	-1
1	-1	-1
1	-1	-1
1	-1	-1
1	-1	-1

ANOVA table and null hypothesis of interest

- ANOVA table in one-way ANOVA (with k groups) shows split of SST into
 - SSM ("between group SS"), $k - 1$ d.f.
 - SSE ("within group SS") , $n - k$ d.f.
- Null hypothesis of interest: $H_0: \mu_1 = \mu_2 = \mu_3$ (means model) or $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$ (effects model)
- Handled with F-test: compare full model with reduced model (=intercept-only)

```
> limeAL$rate <- as.factor(limeAL$rate)
> # Treat rate as FACTOR = QUALITATIVE REGRESSOR!!
> levels(limeAL$rate) <- c("no","low","high")
>
> limeAL.oneway <- lm(pH ~ rate, data=limeAL)
> anova(limeAL.oneway)
Analysis of Variance Table

Response: pH
      Df Sum Sq Mean Sq F value    Pr(>F)
rate    2  0.15628  0.078140   92.748 5.032e-08
Residuals 12  0.01011  0.000842
>
> (SST <- sum((limeAL$pH-mean(limeAL$pH))^2))
[1] 0.16639
```

```
> coef(summary(limeAL.oneway))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.74400    0.01298  442.501  < 2e-16
ratelow      0.12200    0.01836   6.646 2.37e-05
ratehigh     0.25000    0.01836  13.618 1.17e-08
```

Post-hoc tests

- After a significant result from an F test in ANOVA, the next question is: which groups are different?
- Make **pairwise comparisons** between means.
- Different approaches (and a lot of controversy about which one to use when), e.g.:
 - LSD = least significant difference: simple t-tests between groups, no correction for multiple comparisons: $LSD = sed \times t_{dfe}(1 - \alpha)$
 - Tukey = “Experimentwise Error Rate” for all pairwise comparisons
 - Bonferroni = each comparison performed at α/k level, where k is number of comparisons to be made. Overall error rate not higher than α (, but may be much lower!)

Exercise 8 Granulated lime qualitative

a) Read granulated lime data into the program.

Treat rate as a qualitative regressor (though not very appropriate).

b) Fit one-way ANOVA model $y_{ij} = \mu + \alpha_i + e_{ij}$; $e_{ij} \sim N(0, \sigma^2)$

Interpret the results. Ask for and study parameter estimates.

c) Check assumptions using residual analysis: normality, equal variances. Extra: Levene's test for checking homoscedasticity (=equality of variances)

d) Calculate means. Try both ordinary means and estimated marginal means (=predicted means). We will return later to estimated marginal means.

e) Check using an appropriate post-hoc test which means are different.

Two-way and more-way ANOVA

- More than 1 treatment factor
- Factors may be crossed or nested:
 - **crossed** factors: factors A and B are crossed, if every level of A occurs with every level of B
 - **nested** factors: factor B is nested within factor A, if each level of B occurs only within one single level of A
- A factor which is **nested** within another factor, is often (but not always) a factor with random effects (i.e. random factor): we are not interested in the studied levels of this factor per se, but in the variation between the levels. In that case we treat the chosen levels as a random sample from a population.
- **Factorial design** = experimental design with crossed factors
- Means model (interaction model): each combination of factor levels mean of its own:
$$y_{ijk} = \mu_{ij} + e_{ijk} ; e_{ijk} \sim N(0, \sigma^2) \text{ as always}$$
- Effects model (equivalent!): $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} ; e_{ijk} \sim N(0, \sigma^2)$
 - heavily overparameterized → parameter restrictions needed!

Example 13: two-way ANOVA model

■ Again soil pH. Two crossed factors:

- **type of lime**: levels AL=Agricultural Lime, GL=Granular Lime
- **rate of lime** (treating it qualitatively!): levels: 1=0 units, 2=1 unit, 3=2 units, 4=4 units, 5=8 units

```
> limedata <- read.table("lime.prn", header=T)
> limedata$rate <- as.factor(limedata$rate)
>
> lime.anova <- lm(pH ~ lime + rate + lime:rate,
data=limedata)
>
> anova(lime.anova)
Analysis of Variance Table
Response: pH
      Df Sum Sq Mean Sq F value    Pr(>F)
lime    1  0.52429  0.52429  162.759 1.109e-15
rate    4  1.39846  0.34961  108.534 < 2.2e-16
lime:rate 4  0.31075  0.07769   24.117 3.316e-10
Residuals 40  0.12885  0.00322
> (SST <- sum((limedata$pH-mean(limedata$pH))^2 ))
[1] 2.362342 # total sum of squares
```


```
> levels(limedata$rate)
[1] "0" "1" "2" "4" "8"
>
> levels(limedata$lime)
[1] "AL" "GL"
>
> coef(summary(lime.anova))
Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.744  0.02538208  226.3013880 8.816313e-64
limeGL         -0.022  0.03589568   -0.6128871 5.434202e-01
rate1           0.122  0.03589568    3.3987375 1.544598e-03
rate2           0.250  0.03589568    6.9646260 2.100711e-08
rate4           0.391  0.03589568   10.8926750 1.549666e-13
rate8           0.706  0.03589568   19.6681037 3.585202e-22
limeGL:rate1   -0.056  0.05076416   -1.1031405 2.765589e-01
limeGL:rate2   -0.204  0.05076416   -4.0185831 2.514264e-04
limeGL:rate4   -0.199  0.05076416   -3.9200885 3.380006e-04
limeGL:rate8   -0.455  0.05076416   -8.9630163 4.113774e-11
```

Effect types in factorial designs

- Common parameter combinations of interest are:
 - **main effects** : e.g. $\mu_{1.} - \mu_{2.}$ difference between two levels of factor A, averaged over levels of B
 - **Interactions** : e.g. $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ "difference of differences"
- Other comparisons may be of interest too, e.g.
 - $\mu_{11} - \mu_{12}$ difference between two levels of factor A, averaged over levels of B
 - $\mu_{12} - \mu_{1.}$
 - "slices", e.g. comparison of rates per lime type
- Usual questions (to be translated to parameter contrasts):
 - Do factors Lime type and Rate interact?

If interaction unimportant:

 - Do Lime type differ (averaged over Rates)?
 - Do Rates differ (averaged over Lime type)?



Start with interaction!

Factorial design allows study of interaction

- Factors A and B **interact**: difference in mean response for two levels of one factor is **not** constant across levels of second factor.

- Example 14 lime: table of predicted means

	0	1	2	4	8
AL	$\hat{\mu}_{11} = 5.744$	$\hat{\mu}_{12} = 5.866$	$\hat{\mu}_{13} = 5.994$	$\hat{\mu}_{14} = 6.135$	$\hat{\mu}_{15} = 6.450$
GL	$\hat{\mu}_{21} = 5.722$	$\hat{\mu}_{22} = 5.788$	$\hat{\mu}_{23} = 5.768$	$\hat{\mu}_{24} = 5.914$	$\hat{\mu}_{25} = 5.973$

- pH difference AL-GL for level 0 of *rate*: $5.744 - 5.722 = 0.022$
- pH difference AL-GL for level 1 of *rate*: $5.866 - 5.788 = 0.078$

	0	1	2	4	8
AL-GL	$\hat{\mu}_{11} - \hat{\mu}_{21} = 0.022$	$\hat{\mu}_{12} - \hat{\mu}_{22} = 0.078$	$\hat{\mu}_{13} - \hat{\mu}_{23} = 0.226$	$\hat{\mu}_{14} - \hat{\mu}_{24} = 0.221$	$\hat{\mu}_{15} - \hat{\mu}_{25} = 0.477$

- pH differences AL-GL are not equal for the different levels of *rate*, so there is interaction.

- Interaction contrast**: compare pH - differences AL-GL for two levels of *rate*

e.g. $(\hat{\mu}_{11} - \hat{\mu}_{21}) - (\hat{\mu}_{12} - \hat{\mu}_{22}) = 0.022 - 0.078 = -0.056$

- "No interaction" → all interaction contrasts zero!



Partitioning SS in factorial designs

- LS estimation gives SSE. Difference of SSE and SST (intercept-only model) gives SSM: Model SS. What are contributions of main effects and interactions to SSM?
- SSM split into components:
 - main effect A:
 - main effect B
 - interactions $A*B$
- Each SS is difference in error SS of two nested models: go from FM to RM through H_0
- But which two models are compared e.g. to test main effect A?
 - e.g. $A+B+A:B \rightarrow B+A:B$ or $A+B \rightarrow B??$
 - to check an effect also specify the model the effect is part of!
 - order often matters!
- SS unique only for orthogonal cases

Leads to different **types of SS**:

- Type I = Sequential SS
- Type II = Partial SS, respecting marginality
- Type III = Partial, but not respecting marginality

Guidelines for testing

- Would sums of squares of individual term remain the same if we would change the order?
Sometimes they do, but more often they don't. They do if we have equal numbers of observations in the groups. But if they don't, which order of models would we need then?
- First test higher order terms, then lower order terms
- Remove effects from the model if they are not significant, but...
- Do not remove lower order terms from the model, if corresponding higher order terms are still present

Corresponds to Type II SS: respecting marginality
R's function `drop1()`, `Anova()`
But NOT `anova()` (type I SS!)

Exercise 9 Two-way ANOVA with interaction, balanced

Weight gains y_i of 18 calves are measured. Calves stem from 3 different bulls, fed using two types of feed. For each combination of bull and feed, weight gains of 3 calves are obtained. This gives a factorial design. Numbers of observations per combination of factor levels are equal, hence a balanced design.

- a) Read in data file `cattle-balanced.dat`. Columns: `feed`, `bull`, `wgtgain`.
- b) Calculate mean weight gain for all 2×3 feed-bull combinations.
- c) Fit the two-way ANOVA model with interaction.

Study the output produced:

- ANOVA table (what type(s) of sums of squares are these?)
- estimates of parameters (which restriction on parameters is used?)

Do feed and bull interact? Do you find significant main effects of bull and of feed?

- d) Make pairwise comparisons of weight gain between bulls and between levels of feed, using the Tukey method for multiple comparisons.
- e) Make profile plot to visualize the results.
- f) As the interaction is not-significant, we might fit an additive model (model without interaction). Make the profile plot again. Compare with earlier profile plot.

What is the difference in profile plots for two-way ANOVA models with and without interaction?

Exercise 10 Two-way ANOVA with interaction, unbalanced

Again, weight gains y_i of 18 calves are measured. Calves stem from 3 different bulls, fed using two types of food. This gives a factorial design. But now numbers of observations per combination of factor levels are unequal, resulting in an unbalanced design.

- a) Read in data. Columns: food, bull, weight gain.
- b) Fit the two-way ANOVA model with interaction. Study the output:
 - Do food and bull interact?
 - Are there main effects of bull and of food?
- c) Make a profile plot to visualize the interaction between food and bull.

We would like to know the marginal means for the 3 bulls, averaged over the two food types, i.e. estimate $\frac{1}{2} (\mu_{11} + \mu_{21})$, $\frac{1}{2} (\mu_{12} + \mu_{22})$ and $\frac{1}{2} (\mu_{13} + \mu_{23})$, so-called **Estimated Marginal Means** (=Least-Squares Means)

- d) Estimate these marginal means (use `emmeans()` function)

Compare with ordinary means. Explain the difference!

Create a estimated marginal means for bull 1 using `glht()` function by following the R-code.

- e) Compare the weight gain of calves from bull 1 compared to average of bull 2 and 3 *for food 1 only*.
- f) Test the “slice” of food*bull interaction at food type 1 only. What linear combination of parameters do you need to look at?
- g) Get type III SS for main effects food and bull and for interaction food*bull by specifying contrasts using function `linearHypothesis()`.

Experimental design and ANOVA

- Results from experiments are usually analyzed with ANOVA models.
- Commonly used experimental designs:
 - CRD = Completely Randomized Design
 - RCBD = Randomized Complete Block Design
 - BIBD = Balanced Incomplete Block Design
 - Latin square
- CRD
 - treatments allocated to experimental units completely at random
 - could be more than 1 treatment factor.
 - If more than 1 treatment factor, usually models with interactions are used
 - The ANOVA model to analyze the results would be a one-way ANOVA for a design with one treatment factor, two-way ANOVA with interaction for design with two crossed treatment factors, etc.
- RCBD
 - In a randomized complete block design RCBD randomization is restricted, because each treatment should occur at least once in each block.
 - Usually **additive model** is used, with additive effects of blocks and treatments (no interaction)

Experimental design and ANOVA, cont.

■ BIBD = Balanced Incomplete Block Design

- Block size too small to accommodate all treatments: incomplete blocks
- All pairs of treatments occur equally often within blocks
- E.g. 6 blocks, each of size 2, 4 treatments A, B, C, D:
block 1: A B, block 2: A C, block 3: A D, block 4: B C, block 5: B D, block 6: C D
- ANOVA: additive model with block effects and treatment effects

■ Latin square

- 3 factors, each with same number of levels
- Treat one factor as row, one factor as column; each level of third factor occurs exactly once in each row and in each column
- E.g. latin square of order 4

A B C D

D A B C

C D A B

B C D A

So, need only 16 observations.

- ANOVA: additive model with effects of rows, column and treatment.

Block factors and row / column factors may enter model with random effects → Recovery of Interblock Information, see later

Linear model: Analysis of Covariance = ANCOVA

- **Classical ANCOVA**: setting with treatment factor, but also measured **covariate = quantitative regressor**
by inclusion of covariate into model
 - Main question concerns treatment factor, but after “correction for covariate”
 - Inclusion of covariate in model decreases SSE, and, hopefully, MSE; in that case part of error variance is explained by covariate, so that test for treatment effect has more power
 - Inclusion of covariate in model ensures that treatments are compared at “fixed value of covariate”, correcting possible unbalancedness of covariate values over treatment groups
 - Covariate should not be influenced by treatment, because then we would be “explaining away” the effect of treatment by covariate!
- **Wider sense ANCOVA**: linear model with at least one qualitative explanatory variable (=factor) and at least one quantitative explanatory variable (=covariate)
- Case with single factor and single covariate, typical ANCOVA model is **parallel lines model**:

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + e_{ij}$$

Example 15: ANCOVA for peanut yield

y = seed yield of peanut plants; N=30 plants are used, in a CRD.
 $fert$ = fertilizer; 3 types: control, slow or fast release (C=1, S=2 or F=3)
 x = height of a plant at the start as measure of its development or health

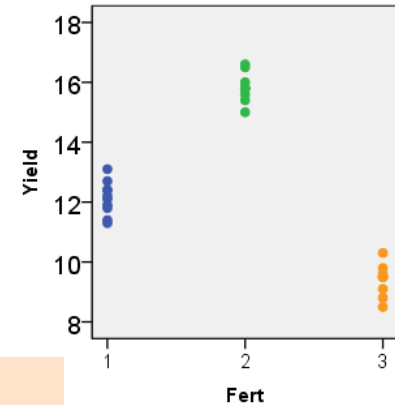
```
> aov.peanuts <- lm(yield ~ fert, data=peanuts)
> anova(aov.peanuts)
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
fert	2	207.074	103.537	391.64	< 2.2e-16 ***
Residuals	27	7.138	0.264		

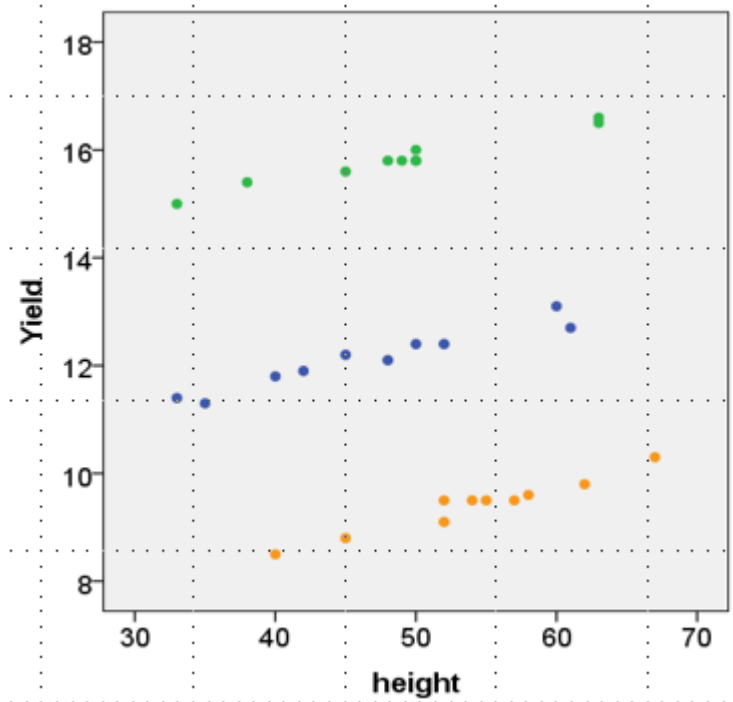
One-way ANOVA



“Correct” for difference in height, for more accurate comparison between C, S and F in 2 ways:

- 1) make blocks based on values of x .
- 2) analysis of covariance, a more sophisticated approach using assumptions about (linear) relationships between y and x for each of the treatments. To do so, x will be introduced in model as “covariate”.

Example 15: ANCOVA model for peanut yield



Three parallel lines:

3 intercepts (β_{01} , β_{02} , β_{03}) for C, F and S, and 1 slope β_1

$$y = \beta_{01} + \beta_1 x + \epsilon \quad (\text{group C})$$

$$y = \beta_{02} + \beta_1 x + \epsilon \quad (\text{group F})$$

$$y = \beta_{03} + \beta_1 x + \epsilon \quad (\text{group S})$$

Different parameterization for intercepts (effects model)
with corner stone restriction (in R):

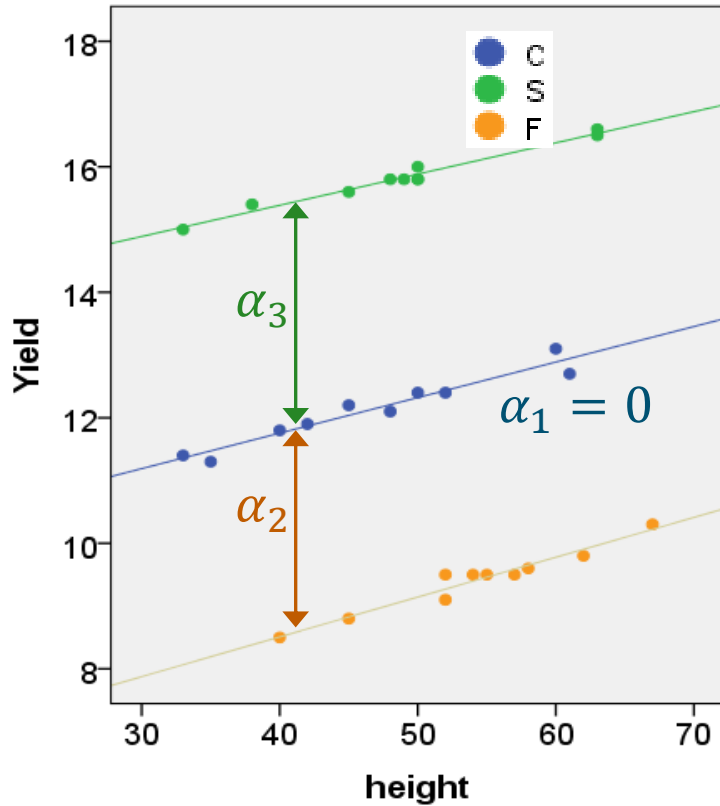
$$y_{ij} = \mu + \alpha_i + \beta_1 x_{ij} + \epsilon_{ij} \quad \text{splits into:}$$

$$y = \mu + \beta_1 x + \epsilon \quad (\text{group C})$$

$$y = \mu + \alpha_2 + \beta_1 x + \epsilon \quad (\text{group F})$$

$$y = \mu + \alpha_3 + \beta_1 x + \epsilon \quad (\text{group S})$$

Example 15: ANCOVA model for peanut yield



Use 3 intercepts to compare the means of the 3 treatments, corrected for height.

```
> anc.peanuts <- lm(yield ~ height + fert, data=peanuts)
> anova(anc.peanuts)
Analysis of Variance Table
Response: yield
      Df Sum Sq Mean Sq  F value    Pr(>F)
height  1   0.453   0.453    28.467 1.389e-05 ***
fert    2 213.346 106.673 6708.623 < 2.2e-16 ***
Residuals 26   0.413   0.016
> coef(anc.peanuts)
(Intercept) height  fertF  ferts
      9.5232  0.0559 -3.1351  3.5713
```

one-way ANOVA: $SSE=7.11$ $s_\varepsilon=0.518$
ANCOVA: $SSE=0.413$ $s_\varepsilon=0.126$

Adjusted treatment means in ANCOVA

- Adjusted treatment means are **predicted treatment means** of the response **at the average value of the covariate x** : $\hat{y}_{i,adj} = \mu + \alpha_i + \beta_1 \bar{x}$
- Differences between adjusted treatment means correspond to differences of response between treatments corrected for covariate (parameters α_i). These treatment effects are the quantities of direct interest in ANCOVA.
- If centered covariate $x_c = x - \bar{x}$ is used, the intercepts of the models $\mu, \mu + \alpha_2, \mu + \alpha_3$ are directly the adjusted treatment means, because the average of centered covariate is 0.

```
> summaryBy(yield+height ~ fert, data=peanuts, FUN=mean)
fert yield.mean height.mean
1    C      12.13      46.6
2    S      15.83      48.9
3    F       9.42      54.2
```

ordinary means

```
> emmeans(anc.peanuts, ~ fert)
fert emmean      SE df lower.CL upper.CL
C     12.315 0.04087 26   12.231   12.399
S     15.886 0.03997 26   15.804   15.968
F       9.179 0.04156 26    9.094    9.265
```

adjusted treatment means

Can you reconstruct the adjusted treatment means from the regression coefficients? To do so, you need the mean value of height (=49.9).

Some books on Mixed Models

