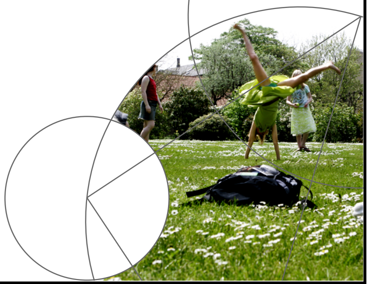
Faculty of Life Sciences

Multiple linear regression and two-way ANOVA

Ib Skovgaard and Claus Ekstrøm
E-mail: ekstrom@life.ku.dk



Program

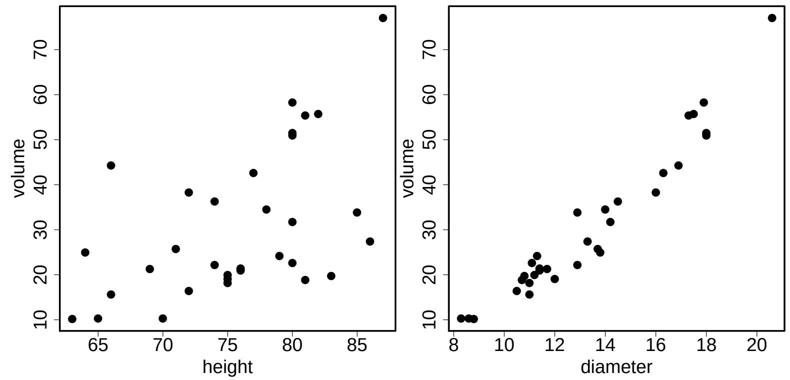
- Multiple linear regression
- Two-way analysis of variance
  - Multi-way ANOVA
- Relation between regression and ANOVA

Example — volume of cherry trees

Tree	Diameter	Height	Volume	Tree	Diameter	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Linear regression

Simple linear regression may describe the relation between two variables:



Linear regression

Regression of volume on height:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835 **
Height        1.5433     0.3839   4.021 0.000378 ***
```

Regression of volume on diameter:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
```

But what if both of the explanatory variables are needed in a good model for the volume?

Multiple regression

The multiple linear regression model with d explanatory variables is given as

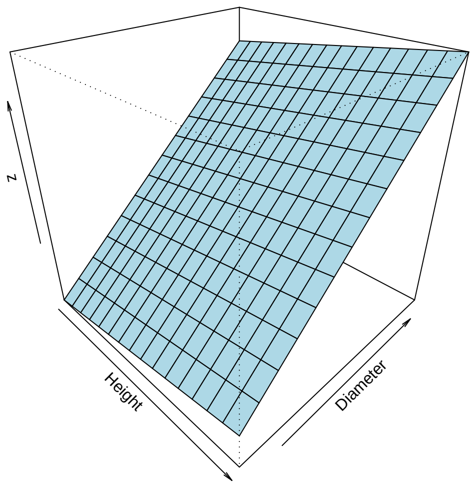
$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + e_i, \quad i = 1, \ldots, n,$$

where $e_i \sim N(0, \sigma^2)$.
It has the same form as the simple linear regression, but with extra explanatory variables.
Three parameters in the model for the mean:

- α intercept with the y-axis when $x_{i1} = \cdots = x_{id} = 0$
- β_1 and β_2 are the partial slopes, giving the y if the other explanatory variables are held constant.

The residual standard deviation, σ, also enters the model.

Graphical display of a multiple regression



Estimation and tests in multiple linear regression

You have learned all the tools already

We need the entire machinery from the previous weeks for estimation (least squares), test of hypotheses (F-tests), confidence- and prediction intervals and model validation.

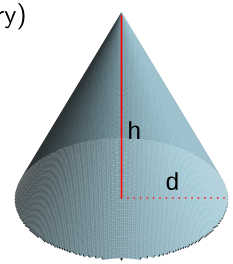
In R we use the function `lm(..)` for the multiple linear regression by appending extra terms to the model.

For example
`lm(Volume ~ Height + Girth)`

Transformation

If we model the tree as a cone with diameter d and height h, we may use the formula (from geometry)

$$v = \frac{\pi}{12} \cdot h \cdot d^2.$$



We replace the constants by parameters to get a more flexible model

$$v = c \cdot h^{\beta_1} \cdot d^{\beta_2},$$

By a log-transform we get

$$\log v_i = \alpha + \beta_1 \log h_i + \beta_2 \log d_i + e_i, \quad i = 1, \ldots, n$$

Polynomial regression

A special application of multiple linear regression is polynomial regression of order k

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i, \quad i = 1, \ldots, n,$$

May describe complicated relations between one variable and another.
Quadratic regression is polynomial regression of order 2

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n,$$

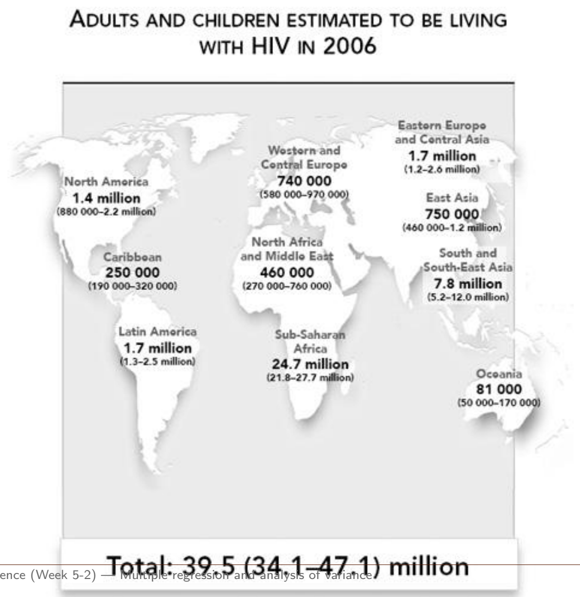
Note that it is the same explanatory variable, x, used in x_i and x_i^2.
Computations in R use the function `lm(..)`:

```
x2 <- x^2      # Defines a new variable
lm(y ~ x + x2)
```

or

```
lm(y ~ x + I(x^2))
```


Prevalence AIDS



Two-way analysis of variance

Two-way analysis of variance extends one-way ANOVA to more than one explanatory variable:

$$y_i = \alpha_{g(i)} + \beta_{h(i)} + e_i, \quad i = 1, \ldots, n,$$

where α and β are the parameters corresponding to the two categorical variables, while g and h define the "groups" for the two variables.
Two-way (and multi-way) ANOVA is handled in R by the function `lm(..)`.
Example:

```
lm(y ~ x1 + x2)
```

where `x1` and `x2` must be defined as factors; otherwise write

```
lm(y ~ factor(x1) + factor(x2))
```

Note that there are now more hypotheses to test (using `drop1()` in R)!

Example — yield of cabbage

Four fields were each divided into plots on which cabbage was grown. We want to investigate the effects of nitrogene applied in the form of calciumnitrate (C), ammoniumsulfate (A), nitrate (N) or control (K).

Yield	Field 1	Field 2	Field 3	Field 4
C	70.3	72.5	79.0	86.2
A	75.5	63.0	65.4	67.7
N	85.2	80.5	83.6	92.3
K	36.7	39.6	45.5	50.5

Inference in the two-way ANOVA

You have learned all the tools already

We need the entire machinery from the previous weeks for estimation (least squares), test of hypotheses (F-tests), confidence- and prediction intervals and model validation.

The only difference is the number of parameters/degrees of freedom, but we get those from the program anyway.

It is all linear regression

Linear regression and analysis of variance are the same model

Factors in the model may be recoded as explanatory variables in a multiple linear regression.
This means that the models may include quantitative as well as qualitative explanatory variable.

To write an ANOVA model as a regression we use dummy variable

$$x_{ij}^k = \begin{cases} 1 & \text{if observation } i \text{ belongs to category } j \text{ for variable } k \\ 0 & \text{ellers} \end{cases} .$$

Lecture summary: main points

- Multiple linear regression
  - What can we achieve by this model?
  - Interpretation, estimation and hypothesis testing
- Multi-way analysis of variance
  - What can we achieve by this model?
  - Interpretation, estimation and hypothesis testing
- It is all “linear regression”