Faculty of Life Sciences

Linear regression

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

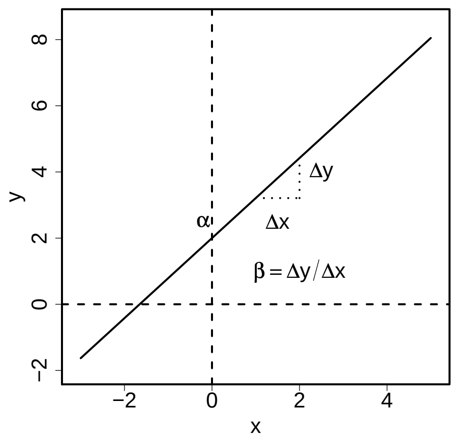
Program

- The straight line
- Fitting a line to data
  - The method of least squares
- Model validation
- The correlation coefficient

The straight line

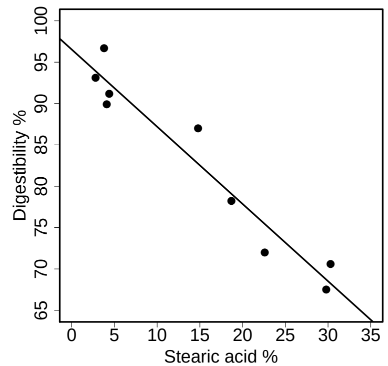
$$y = \alpha + \beta \cdot x$$

The slope is β and the intercept is α.



Example — Digestibility

% stearic acid	% digestibility
29.8	67.5
30.3	70.6
22.6	72.0
18.7	78.2
14.8	87.0
4.1	89.9
4.4	91.2
2.8	93.1
3.8	96.7



Residuals

Having (somehow) fitted a line,

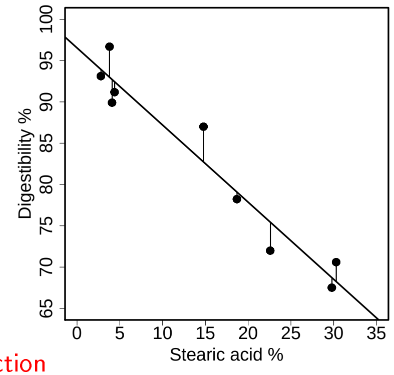
$$y = \hat{\alpha} + \hat{\beta} \cdot x$$

given by estimates $\hat{\alpha}$ and $\hat{\beta}$. The predicted value corresponding to any x is

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$$

The residual corresponding to any observation is the "model error" equal to the difference

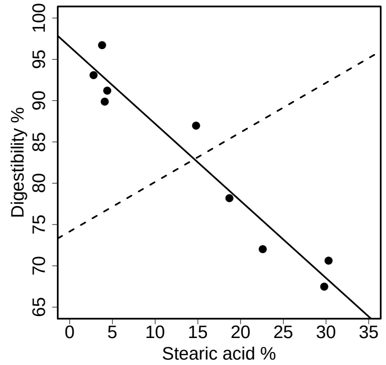
$r_i = y_i - \hat{y}_i =$ observation - model prediction



How do we find the best line?

- Residuals should be small (pos. or neg.)
- Gauss' solution: minimize the sum of squared residuals

$$r_1^2 + \ldots + r_n^2$$



Fitting a line

Best straight line

The least squares solution is the line that minimizes the sum of squared residuals.

Estimates:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (1)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}. \quad (2)$$

Example

i	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	29.8	67.5	15.211	-15.411	231.378	237.502	-234.420
2	30.3	70.6	15.711	-12.311	246.839	151.563	-193.421
3	22.6	72.0	8.011	-10.911	64.178	119.052	-87.410
4	18.7	78.2	4.111	-4.711	16.901	22.195	-19.368
5	14.8	87.0	0.211	4.089	0.045	16.719	0.863
6	4.1	89.9	-10.489	6.989	110.017	48.845	-73.306
7	4.4	91.2	-10.189	8.289	103.813	68.706	-84.455
8	2.8	93.1	-11.789	10.189	138.978	103.813	-120.116
9	3.8	96.7	-10.789	13.789	116.400	190.133	-148.767
Sum	131.3	746.2	0.000	0.000	1028.549	958.529	-960.399

Method of least squares: computations

Find α and β to make

$$\sum_i r_i^2 = \sum_i (y_i - \alpha - \beta \cdot x_i)^2$$

as small as possible.
Solve the equations

$$\begin{aligned} \frac{\partial f}{\partial \alpha} &= \sum_{i=1}^{n} \frac{\partial}{\partial \alpha}(y_i - \alpha - \beta \cdot x_i)^2 = \sum_{i=1}^{n} 2(y_i - \alpha - \beta \cdot x_i) \cdot (-1) \\ &= -2 \cdot (y_\bullet - n\alpha - \beta x_\bullet) = 0 \qquad (3) \\ \frac{\partial f}{\partial \beta} &= \sum_{i=1}^{n} 2(y_i - \alpha - \beta \cdot x_i) \cdot (-x_i) = 0 \qquad (4) \end{aligned}$$

Model validation

- Quantitative variables in pairs (x, y)
- Linear relation?
- Influential observations?
- x on y or y on x.
- Extra- and interpolation.

Transformations

Duckweed

Days	Leaves	Days	Leaves
0	100	7	918
1	127	8	1406
2	171	9	2150
3	233	10	2800
4	323	11	4140
5	452	12	5760
6	654	13	8250

What if data follow a curved relation?
Sometimes we can transform data to make the relation linear
Exponential growth model for population size at time t:

$$f(t) = c \cdot \exp(b \cdot t).$$

Taking logarithm on both sides we get

$$\log(f(t)) = \underbrace{\log c}_{\alpha} + \underbrace{b}_{\beta} \cdot t.$$

The correlation coefficient

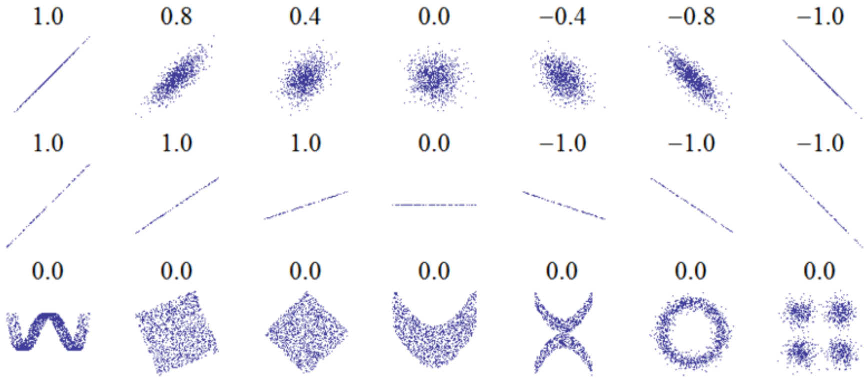
The correlation coefficient, ρ, quantifies how close the *linear* relation is between X and Y:

$$\hat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}}.$$

The correlation coefficient is always between -1 and 1, and it is

- 0 if there is no relation between x and y,
- 1 if the observations (x_i, y_i) are exactly on a line with positive slope,
- -1 if the observations (x_i, y_i) are exactly on a line with negative slope.

Correlation coefficient — Illustrations



Summary of main points from Chapter 2

- Linear regression (estimating a line)
  - Interpretation of the parameters
  - lm(..) in R: fitting the line
  - and reading the output
- Residuals
- Method of least squares
- The correlation coefficient
  - Definition, properties and interpretation
- Is a line the right model?
  - Transforming data to fit a line