Biostatistics

Correlation and linear regression

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit
University of Zurich

Correlation and linear regression

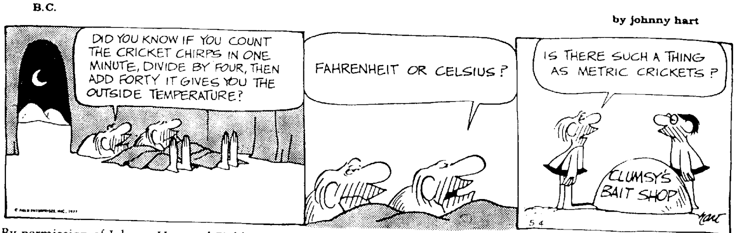
Analysis of the relation of two continuous variables (bivariate data).

Description of a non-deterministic relation between two continuous variables.

Problems:

1. How are two variables x and y related?
   (a) Relation of weight to height
   (b) Relation between body fat and bmi

2. Can variable y be predicted by means of variable x?



By permission of Johnny Hart and Field Enterprises, Inc.

Example

- Proportion of body fat modelled by age, weight, height, bmi, waist circumference, biceps circumference, wrist circumference, total $k = 7$ explanatory variables.
- Body fat: Measure for "health", measured by "weighing under water" (complicated).
- Goal: Predict body fat by means of quantities that are easier to measure.

$n = 241$ males aged between 22 and 81.

11 observations of the original data set are omitted: "outliers".

Penrose, K., Nelson, A. and Fisher, A. (1985), "Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques". Medicine and Science in Sports and Exercise, **17**(2), 189.

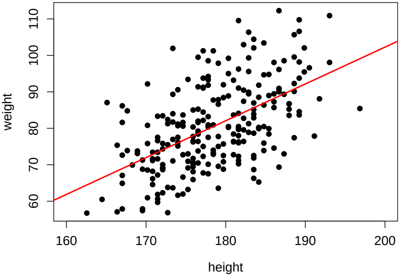
Bivariate data

- Observation of two continuous variables (x, y) for the same observation unit

  $\longrightarrow$ pairwise observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

  Example: Relation between weight and height for 241 men

- Every correlation or regression analysis should begin with a scatterplot



$\longrightarrow$ visual impression of a relation

Correlation

Pearson's product-moment correlation

- measures the strength of the linear relation, the linear coincidence, between x and y.

Covariance: $$\text{Cov}(x,y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

Variances: $$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

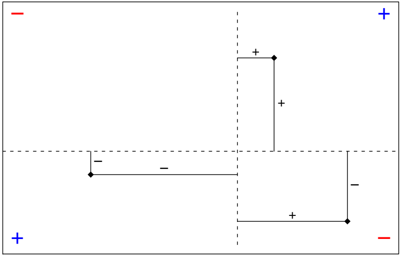
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Correlation: $$\boxed{r = \frac{s_{xy}}{s_x\, s_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}}$$

Correlation

Plausibility of the enumerator:

Correlation: $$r = \frac{s_{xy}}{s_x\, s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Plausibility of the denominator:

r is independent of the measuring unit.

Correlation

Properties:

$$-1 \leq r \leq 1$$

$$r = 1 \quad \longrightarrow \quad \text{deterministic positive linear relation between } x \text{ and } y$$

$$r = -1 \quad \longrightarrow \quad \text{deterministic negative linear relation between } x \text{ and } y$$

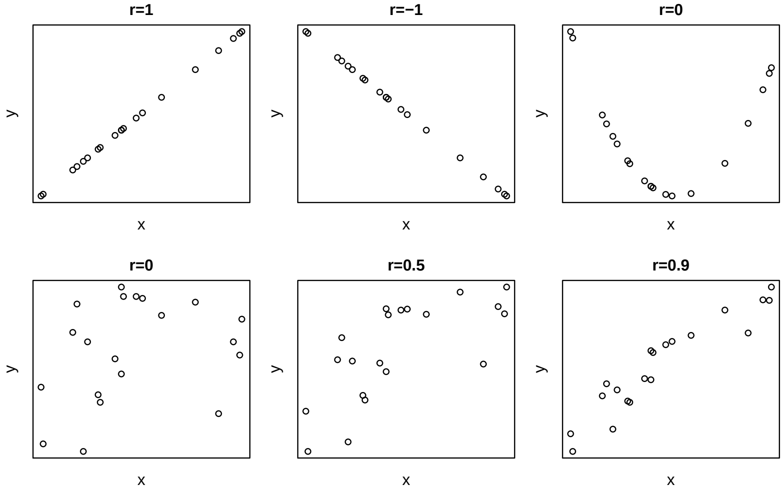
$$r = 0 \quad \longrightarrow \quad \text{no linear relation}$$

In general:

- Sign indicates direction of the relation
- Size indicates intensity of the relation

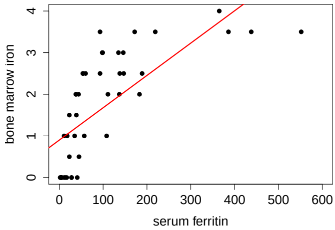
Correlation

Examples:



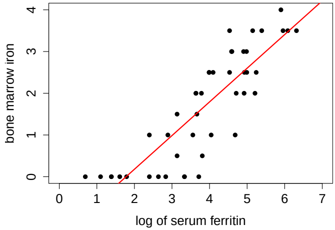
Correlation

Example: Relation between blood serum content of Ferritin and bone marrow content of iron.



$r = 0.72$

- Transformation to linear relation?
- Frequently a transformation to the normal distribution helps.



$r = 0.85$

Tests on linear relation

Exists a linear relation that is not caused by chance?

Scientific hypothesis: true correlation $\rho \neq 0$

Null hypothesis: true correlation $\rho = 0$

Assumptions:

- (x, y) jointly normally distributed
- pairs independent

Test quantity: $$\boxed{T = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}}$$

Tests on linear relation

Example: Relation of weight and body height for males.

$$n = 241, \qquad r = 0.55$$

$$\longrightarrow T = 7.9 > t_{239,0.975} = 1.97, p < 0.0001$$

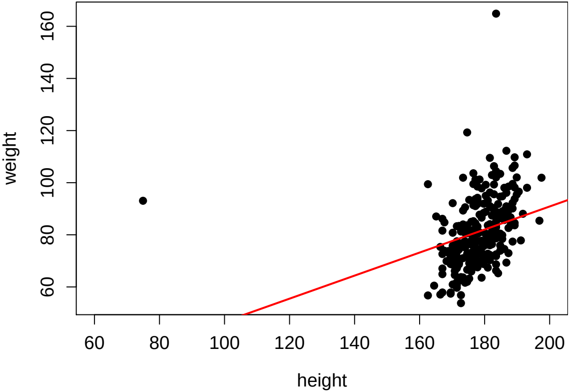
Confidence interval: Uses the so called Fisher's z-transformation leading to the approximative normal distribution

$$\rho \in (0.46, 0.64) \quad \text{with probability } 1 - \alpha = 0.95$$

Spearman's rank correlation

Treatment of outliers?

Testing without normal distribution?



$n = 252, r = 0.31, p < 0.0001$

Spearman's rank correlation

Idea: Similar to the Mann-Whitney test with ranks

Procedure:

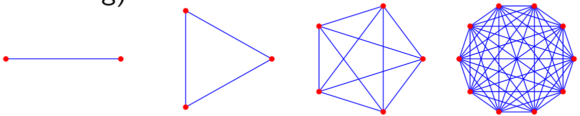
1. Order $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ separately by ranks
2. Compute the correlation for the ranks instead of for the observations

$\longrightarrow r_s = 0.52, p < 0.0001$

(correct data $(n = 241) : r_s = 0.55, p < 0.0001$)

Dangers when computing correlation

1. 10 variables → 45 possible correlations
   (problem of multiple testing)



Nb of variables	2	3	5	10
Nb of correlations	1	3	10	45
P(wrong signif.)	0.05	0.14	0.40	0.91

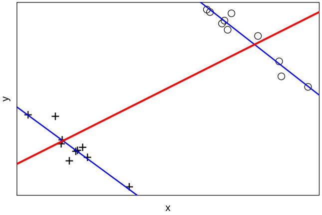
   Number of pairs increases rapidly with the number of variables.
   ⟶ increased probability of wrong significance

2. Spurious correlation across time (common trend)
   Example: Correlation of petrol price and divorce rate!

3. Extreme data points: outlier, “leverage points”

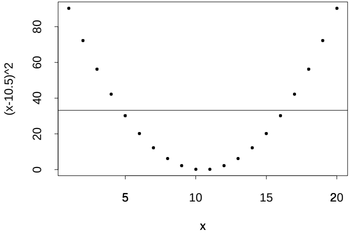
Dangers when computing correlation

4. Heterogeneity correlation (no or even opposed relation within the groups)



5. Confounding by a third variable
   Example: Number of storks and births in a district
   $\longrightarrow$ confounder variable: district size

6. Non-linear relations (strong relation, but $r = 0 \longrightarrow$ not meaningful)



Simple linear regression

Regression analysis = statistical analysis of the effect of one variable on others

$\longrightarrow$ directed relation

$x =$ independent variable, explanatory variable, predictor
(often not by chance: time, age, measurement point)

$y =$ dependent variable, outcome, response

Goal:

Do not only determine the strength and direction ($\nearrow$, $\searrow$) of the relation, but define a quantitative law (how does y change when x is changed).

Simple linear regression

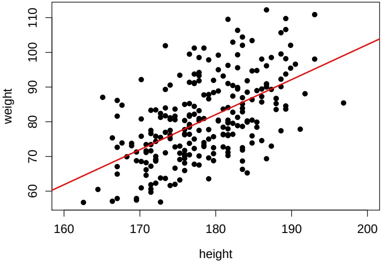
Example: Quantification of overweight.

Is weight a good measurement, is the "body mass index" (bmi = weight/height^2) better?

Regression:

y = weight, x = height

($n = 241$ men)



$$y = -99.66 + 1.01\,x, \quad r^2 = 0.31, \quad p < 0.0001$$

⇒ Body height is no good measurement for overweight

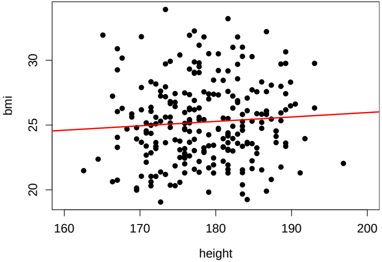
How heavy are males? $\bar{y} = 80.7$ kg, SD$= s_y = 11.8$ kg

How heavy are males of size 175 cm?

$$\hat{y} = -99.66 + 1.01 \times 175 = 77.0 \text{ kg}, \; s_e = 9.9 \text{ kg}$$

Simple linear regression

Regression:
y = bmi = weight/height^2,
x = height



$$y = 19.2 + 0.034\,x, \quad r^2 = 0.005, \quad p = 0.27$$

$\Rightarrow$ The bmi does not depend on body height and is therefore a better measurement for overweight

How heavy are males? $\bar{y} = 25.2$ kg/m^2, SD= $s_y = 3.1$ kg/m^2

How heavy are males of size 175 cm?
$\hat{y} = 19.2 + 0.034 \times 175 = 25.1$ kg/m^2, $s_e = 3.1$ kg/m^2

Statistical model for regression

$$\boxed{y_i = f(x_i) + \varepsilon_i \quad i = 1, \ldots, n}$$

$f =$ regression function; implies relation $x \mapsto y$; true course

$\varepsilon_i =$ unobservable, random variations (error; noise)

- ε_i independent
- mean(ε_i)$= 0$, variance(ε_i) $= \sigma^2 \leftarrow$ constant
- For tests and confidence intervals: ε_i normally distributed $\mathcal{N}(0, \sigma^2)$

Important special case: linear regression

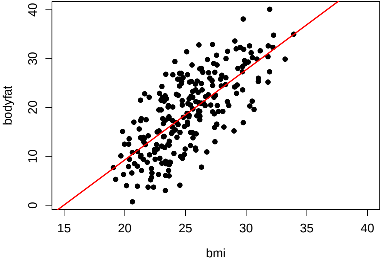
$$f(x) = a + bx$$

To determine ("estimate"): $a =$ intercept, $b =$ slope

Statistical model for regression

Example: Both percental body fat and bmi are measurements for overweight of males, but only bmi is easy to measure.

Regression:
y = body fat (in %),
x = bmi (in kg/m^2)

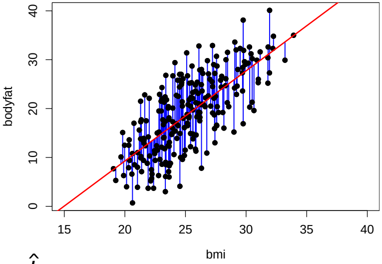


$$y = -27.6 + 1.84\,x, \quad r^2 = 0.52, \quad p < 0.0001$$

Interpretations:

- Men with a bmi of 25 kg/m^2 have 18% body fat on average.
- Men with an about 1 kg/m^2 increased bmi have 2% more body fat on average.

Method of least squares



Method to estimate a and b

Value on regression line at x_i: $\hat{y}_i = \hat{a} + \hat{b} x_i$

Choose parameter estimator, so that

$$S(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{is minimized}$$

$$\longrightarrow \text{Slope: } \hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r\,\frac{s_y}{s_x}; \quad \text{Intercept: } \hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

♣ Derivation of the formulas for $\hat{a}$ and $\hat{b}$

New parameterisation: $y - \bar{y} = \alpha + \beta(x - \bar{x})$

$$\begin{aligned} \longrightarrow a &= \alpha + \bar{y} - \beta\bar{x} \\ b &= \beta \end{aligned}$$

$$S(\alpha, \beta) = \sum_{i=1}^{n} \{(y_i - \bar{y}) - \alpha - \beta(x_i - \bar{x})\}^2$$

S is a quadratic function in (α, β)

- S has a unique minimum if there are at least two different values x_i.
- set the partial derivations equal to zero:

$$\frac{\partial S}{\partial \alpha} = 2 \sum \{(y_i - \bar{y}) - \alpha - \beta(x_i - \bar{x})\} \{-1\} = 0$$

$$\frac{\partial S}{\partial \beta} = 2 \sum \{(y_i - \bar{y}) - \alpha - \beta(x_i - \bar{x})\} \{-(x_i - \bar{x})\} = 0$$

♣ Derivation of the formulas for $\hat{a}$ and $\hat{b}$

⟶ Normal equations:

$$\begin{aligned} \alpha n + \beta \sum (x_i - \bar{x}) &= \sum (y_i - \bar{y}) = 0 \\ \alpha \sum (x_i - \bar{x}) + \beta \sum (x_i - \bar{x})^2 &= \sum (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

⟶ Solution:

$$\begin{aligned} \hat{\alpha} &= 0 \\ \hat{\beta} &= \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x} \end{aligned}$$

⟶

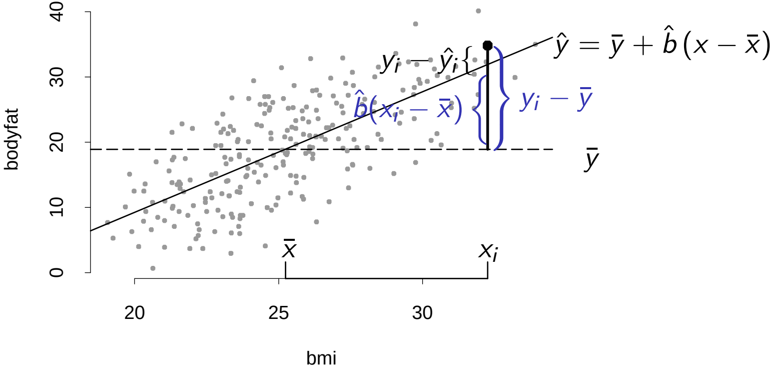
$$\begin{aligned} \hat{b} &= \hat{\beta} = r\,\frac{s_y}{s_x} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x} \end{aligned}$$

very intuitive regression equation: $\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$

Variance explained by regression

Question: How relevant is regression on x for y?

Statistically: How much variance of y is explained by the regression line, i.e. knowledge of x?



Variance explained by regression

Decomposition of the variance by regression:

$$\underbrace{y_i - \bar{y}}_{\text{observed}} = \underbrace{\left\{\hat{b}(x_i - \bar{x})\right\}}_{\text{explained}} + \underbrace{\left\{y_i - \bar{y} - \hat{b}(x_i - \bar{x})\right\}}_{\text{rest}}$$

$$\text{observed} = \text{explained} + \text{rest}$$

Square, sum up and divide by $(n-1)$:

$$s_y^2 = \hat{b}^2\, s_x^2 + s_{\text{res}}^2$$

mixed term $\hat{b}\, s_{x,\text{res}}$ disappears.

Variance explained by regression

"Explained" variance $\hat{b}^2 s_x^2$:

$$s^2_{\text{reg}} = \hat{b}^2\, s_x^2 = \left(r\, \frac{s_y}{s_x}\right)^2 s_x^2 = r^2\, s_y^2$$

$r^2 = \dfrac{s^2_{\text{reg}}}{s_y^2}$ = proportion of variance of y that is explained by x.

Residual variance: Variance that remains

$$s^2_{\mathrm{res}} = (1 - r^2)\, s_y^2\,, \qquad \hat{\sigma}^2 = s_e^2 = \frac{1}{n-2} \sum e_i^2 = \frac{n-1}{n-2}\, s^2_{\mathrm{res}}$$

Observations vary around the regression line with standard deviation

$$s_{\mathrm{res}} = \sqrt{1 - r^2}\, s_y$$

r		0.3	0.5	0.7	0.9	0.99
s_{res}/s_y	$= \sqrt{1 - r^2}$	0.95	0.87	0.71	0.44	0.14
Gain	$= 1 - \sqrt{1 - r^2}$	5%	13%	29%	56%	86%

Gain of the regression

- How heavy are males on average?

  Classical quantities: $\bar{y} = 80.7$ and $s_y = 11.8$

  $\Rightarrow$ Estimator: 80.7 kg

  $\Rightarrow$ Approx. 95% of the males weigh between $80.7 \pm 2 \times 11.8$ kg, i.e. between 57.1 and 104.3 kg

- How heavy are males of 175 cm on average?

  Regression: $\bar{y} = -99.7 + 1.01\, x$ and $s_{\text{res}} = 9.8$

  $\Rightarrow$ Estimator: $-99.7 + 1.01 \times 175 = 77.0$ kg

  $\Rightarrow$ Approx. 95% of the males of 175 cm weigh between $77.0 \pm 2 \times 9.8$ kg, i.e. between 57.4 and 96.6 kg

The regression model provides better estimators and a smaller confidence interval.

Gain: $1 - s_{\text{res}}/s_y = 1 - 9.8/11.8 = 17\%$ $\quad (r = 0.56)$

Gain of the regression

Is there a relation at all?

Scientific hypothesis: y changes with x ($b \neq 0$)

Null hypothesis: $b = 0$

if (x, y) normally distributed
$\longrightarrow$ same test as for correlation $\rho = 0$ (t–distribution)

In regression analysis:

- all analyses conditional on given values $x_1, \ldots, x_n$:
  ε_i independent $\mathcal{N}(0, \sigma^2)$

  $\longrightarrow$ simpler than analyses of correlation
  $\longrightarrow$ distribution of x negligible

- $\hat{b} \sim \mathcal{N}(b, SE(\hat{b}))\,, \qquad SE(\hat{b}) = \dfrac{\sigma}{s_x \sqrt{n-1}}$

Gain of the regression

Test quantity:

$$T = \hat{b}\,\frac{s_x\sqrt{n-1}}{\hat{\sigma}} \sim t_{n-2}$$

Comment: $\hat{\sigma}^2 = \frac{n-1}{n-2}\,(1-r^2)\,s_y^2$, $\hat{b} = r\,\frac{s_y}{s_x} \longrightarrow T = r\,\sqrt{\frac{n-2}{1-r^2}}$

Example: Body fat in dependence on bmi for 241 males.

Results R:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.617	2.939	-9.398	0.000
bmi	1.844	0.116	15.957	0.000

$r^2 = 0.52$

$\longrightarrow s_{res}/s_y = \sqrt{1-0.52} = 0.69 \longrightarrow$ Gain: 31%

♣ Confidence interval for b

Again conditional on the given values $x_1, \ldots, x_n$

$(1-\alpha)$ – confidence interval

$$\hat{b} \pm t_{1-\alpha/2} \, \frac{\hat{\sigma}}{s_x \sqrt{n-1}}$$

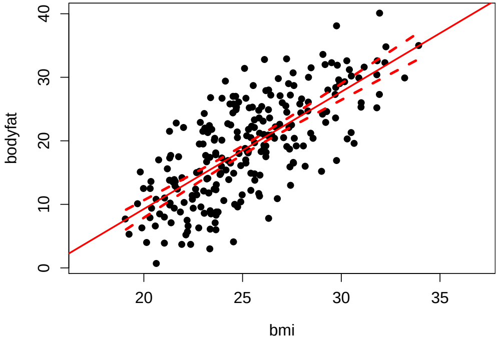
Confidence interval for the regression line

Consider the alternative parameterisation: $\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$

- The variances sum up since $\bar{y}$ and $\hat{b}$ are independent.

$\longrightarrow$ $(1-\alpha)$–confidence interval for the value of the regression line $y(x^\star)$ at $x = x^\star$:

$$\hat{a} + \hat{b}\,x^* \pm t_{1-\alpha/2}\,\hat{\sigma}\,\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_x^2(n-1)}}$$



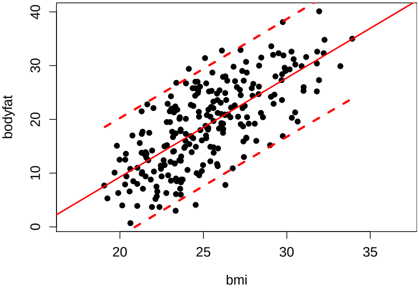
Prediction interval for y

Future observation $y^\star$ at $x = x^\star$

$$y^\star = \hat{y}(x^\star) + \varepsilon$$

$\longrightarrow$ $(1-\alpha)$–prediction interval for $y(x^\star)$:

$$\hat{a} + \hat{b}\,x^* \pm t_{1-\alpha/2}\,\hat{\sigma}\,\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_x^2(n-1)}}$$



- Prediction interval is much wider than the confidence interval

Multiple regression

Topics:

- Regression with several independent variables
  - Least squares estimation
  - Multiple coefficient of determination
  - Multiple and partial correlation
- Variable selection
- Residual analysis
  - Diagnostic possibilities

Multiple regression

Reasons for multiple regression analysis:

1. Eliminate potential effects of confounding variables in a study with one influencing variable.

   Example: A frequent confounder is age: y = blood pressure, x_1 = dose of antihypertensives, x_2 = age.

2. Investigate potential prognostic factors of which we are not sure whether they are important or redundant.

   Example: y = stenosis, x_1 = HDL, x_2 = LDL, x_3 = bmi, x_4 = smoking, x_5 = triglyceride.

3. Develop formulas for predictions based on explanatory variables.

   Example: y = adult height, x_1 = height as child, x_2 = height of the mother, x_3 = height of the father.

4. Study the influence of a variable x_1 on a variable y taking into account the influence of further variables $x_2, \ldots, x_k$.

Example: Prognostic factors for body fat

Number of observed males: $n = 241$

Dependent variable: bodyfat = percental body fat

We are interested in the influence of three independent variables:

- bmi in kg/m^2.
- waist circumference (abdomen) in cm.
- waist/hip-ratio.

Results of the univariate analyses of bodyfat based on bmi, abdomen and waist/hip-ratio with R:

Example: Prognostic factors for body fat

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.617	2.939	-9.398	0.000
bmi	1.844	0.116	15.957	0.000

BMI: $R^2 = 0.516$, $R^2_{\text{adj}} = 0.514$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-42.621	2.869	-14.855	0.000
abdomen	0.668	0.031	21.570	0.000

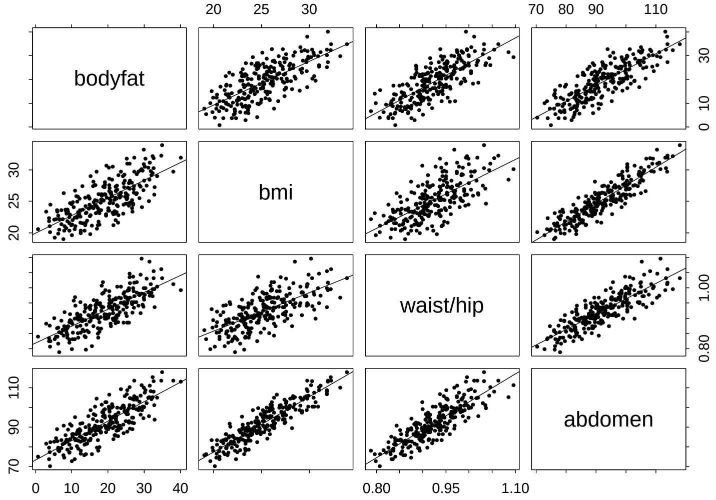
Abdomen: $R^2 = 0.661$, $R^2_{\text{adj}} = 0.659$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-78.066	5.318	-14.680	0.000
waist_hip_ratio	104.976	5.744	18.275	0.000

Waist/hip-ratio: $R^2 = 0.583$, $R^2_{\text{adj}} = 0.581$

Example: Prognostic factors for body fat

Pairwise-scatterplots:



Example: Prognostic factors for body fat

Multiple regression:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-60.045	5.365	-11.192	0.000
bmi	0.123	0.236	0.519	0.605
abdomen	0.438	0.105	4.183	0.000
waist_hip_ratio	38.468	10.262	3.749	0.000

$R^2 = 0.681$, $R^2_{\text{adj}} = 0.677$

Elimination of the non-significant variable bmi:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-59.294	5.158	-11.496	0.000
abdomen	0.484	0.057	8.526	0.000
waist_hip_ratio	36.455	9.486	3.843	0.000

$R^2 = 0.680$, $R^2_{\text{adj}} = 0.678$

Example: Prognostic factors for body fat

In general:

$$y = a + b_1 \; x_1 + b_2 \; x_2 + \ldots + \varepsilon$$

Estimation: $\downarrow \quad \downarrow \quad \downarrow \quad \downarrow \quad \downarrow$

$$\text{bodyfat} = -59.3 + 0.484 \; \text{abdomen} + 36.46 \; \text{waist/hip-ratio}$$

Statistical model

$$y_i = a + b_1\, x_{1i} + b_2\, x_{2i} + \ldots + b_k\, x_{ki} + \varepsilon_i \quad i = 1, \ldots, n$$

$a + b_1\, x_1 + b_2\, x_2 + \ldots + b_k\, x_k$ = regression function, response surface

ε_i = unobserved, random noise

- independent
- $\mathrm{E}(\varepsilon_i) = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2 \leftarrow$ constant

Procedure as in the case of the simple linear regression:

Least squares method:

Prediction: $\hat{y}_i = \hat{a} + \hat{b}_1\, x_{1i} + \ldots + \hat{b}_k\, x_{ki}$

Choose estimation of the parameters, so that

$$S(\hat{a}, \hat{b}_1, \ldots, \hat{b}_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{is minimized!}$$

Set partial derivatives equal to zero $\rightarrow$ normal equations.

Statistical model

For a clear illustration use a matrix formulation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$$

$\longrightarrow$ Statistical model: $\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$

Normal equations (for $a, b_1, \ldots, b_k$):

$$\mathbf{X}'\mathbf{X}\,\mathbf{b} = \mathbf{X}'\mathbf{y}$$

Remember: centered formulation for the simple linear regression:

$$\sum (x_i - \bar{x})^2 b = \sum (x_i - \bar{x})(y_i - \bar{y})$$

Generalisation of the correlation

Instead of one correlation we get a correlation matrix.

	bodyfat	bmi	waist_hip	abdomen	weight
bodyfat	1.000	0.000	0.000	0.000	0.000
bmi	0.718	1.000	0.000	0.000	0.000
waist_hip	0.763	0.678	1.000	0.000	0.000
abdomen	0.813	0.903	0.847	1.000	0.000
weight	0.600	0.867	0.540	0.865	1.000

Here the pairwise correlations are shown below the diagonal and the p–values above.

Generalisation of the correlation

How strong is the multiple linear relation?

Multiple coefficient of determination

$$R^2 = \frac{s^2_{\text{reg}}}{s^2_y} = \frac{\text{explained variance}}{\text{variance of } y} = 1 - \frac{s^2_{\text{res}}}{s^2_y}$$

Comment: $R^2 = (r_{y\hat{y}})^2$

$r_{y\hat{y}}$ is called multiple correlation coefficient

$=$ correlation between y and best linear combination of $x_1, \ldots, x_k$

Remember: R^2 is a measure for the goodness of a prediction:

- observations scatter around $\bar{y}$ with SD $= s_y$
- observations scatter around the prediction value $\hat{y}$ with $s_{\mathrm{res}} = \sqrt{1 - R^2}\, s_y \leq s_y$

Generalisation of the correlation

Example: $s_{\text{bodyfat}} = 8.0\,,\ R^2 = 0.68$

$\longrightarrow s_{\text{res}} = \sqrt{1 - 0.68} \times 8.0 = 4.5$

Warning: R^2 does not provide an unbiased estimation of the proportion of expected variance explained by regression (too optimistic).

Unbiased estimation of the residual variance:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} e_i^2 = \frac{n-1}{n-k-1} s_{\text{res}}^2$$

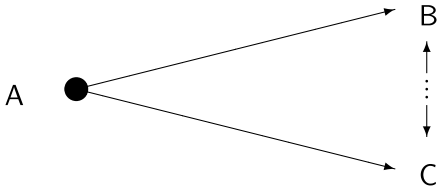
Unbiased estimation of the proportion of explained variance.

$$\boxed{R_{\mathsf{adj}}^2 = 1 - \frac{\hat{\sigma}^2}{s_y^2}}$$

♣ Partial correlation

Correlation coefficient between two variables whereby the remaining variables are kept constant.

$\longrightarrow$ Comparable statement as multiple regression coefficient



A is a "confounder" for the relation of B to C

♣ Partial correlation

Example: Relation of body fat proportion and weight for males.

A = abdomen, B = body fat, C = weight:

$$r_{\text{AB}} = 0.81, \qquad r_{\text{AC}} = 0.86, \qquad r_{\text{BC}} = 0.60$$

Are body fat proportion and weight related?

$$r_{BC.A} = \frac{r_{\text{BC}} - r_{\text{AB}} r_{\text{AC}}}{\sqrt{(1 - r_{\text{AB}}^2)(1 - r_{\text{AC}}^2)}} = -0.35$$

$\longrightarrow$ the sign of the correlation switches when the waist circumference is known.

Examination of hypotheses

(Null) hypotheses:

- There is no relation at all between $(x_1, \ldots, x_k)$ and y.
- A certain independent variable has no influence.
- A group of independent variables has no influence.
- The relation is linear and not quadratic.
- The influence of the independent variables is additive.

Condition: ε_i normally distributed

Linear hypotheses $\longrightarrow$ F-tests

Examination of hypotheses

Example:

Null hypothesis: true multiple correlation $R = 0$ (no relation at all).

Test quantity

$$T = \frac{R^2\,(n-k-1)}{1-R^2} \sim F_{1,n-k-1}$$

(Generalisation of the simple, linear case, since $F_{1,m} = t_m^2$)

♣ Variable selection

- Aspects:
  - simple model (without inessential variables)
  - include important variables
  - high prediction power
  - reproducibility of the results
- Procedure:
  - stepwise procedure
    - ⋆ forward
    - ⋆ backward
    - ⋆ stepwise
  - "best subset selection"
- Problem:
  - multi-collinearity $\longrightarrow$ instability

♣ Variable selection

Stepwise procedures: stepwise, forward, backward

- Dependent variable: y = bodyfat
- Independent variables: x = age, weight, body height, 10 body circumference measures, waist-hip ratio.

forward ($p = 0.05$)

step	included	R^2	R^2_{adj}	variable	p–value
1.	abdomen	.661	.659	abdomen	$<$.0001
2.	weight	.703	.700	abdomen	$<$.0001
				weight	$<$.0001
3.	wrist	.714	.711	abdomen	$<$.0001
				weight	.0004
				wrist	.002
4.	biceps	.718	.713	abdomen	$<$.0001
				weight	$<$.0001
				wrist	.001
				biceps	.08

backward: same result

Common model:

bodyfat = constant + abdomen + weight + wrist + error

♣ Variable selection

Keep in mind:

- The model of the multiple linear regression should be assessed according to the meaning and significance of the prediction variables and according to the proportion of explained variance R^2_{adj}.
- Stepwise p-values $\not\rightarrow$ significance
- If the forecast is important use AIC, GCV, BIC, ...

Residual analysis

- Examination of the assumptions of the regression analysis:
  - outliers, non-normal distribution
  - influential observations, leverage points
  - unequal variances
  - non-linearity
  - dependent observations
- graphical methods $\longleftrightarrow$ tests

Keep in mind:

There is no universally valid procedure for the examination of the assumptions of the regression analysis!

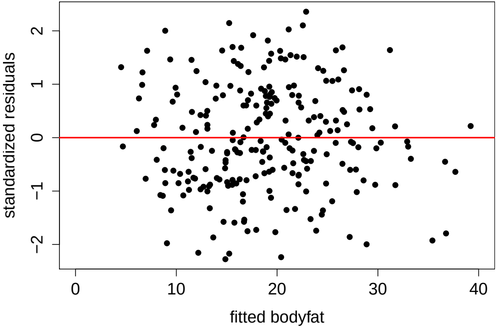
Residuals

Residual

observation - predicted value

Standardized residual

$$\frac{\text{residual}}{\text{sample standard deviation of the residuals}}$$



Residuals

Standardized residuals should be within -2 and 2. There should be no specific patterns.

Otherwise, check for

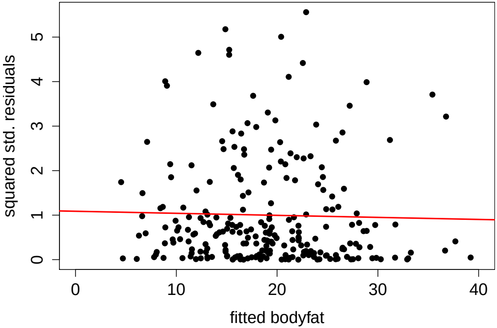
- outliers
- unequal variances
- non-normal distribution
- non-linearity
- important variable not included in the model

Remember:
"Pattern" should be interpretable in respect of contents and should be significant.

$\longrightarrow$ Non-parametric procedures

Variance stability

Plot squared standardized residuals against predicted target quantity.



H_0: Spearman's rank correlation coefficient $= 0 \longrightarrow p = 0.19$

Contraindications

- dependent measurements (e.g. for one person)
  Solution: Repeated-measures analysis
- variability dependent on measurement
  Solution:
  1. transformation
  2. weighted least-squares estimation
- skewed distribution
  Solution:
  1. transformation
  2. robust regression
- non-linear relation
  Solution:
  1. transformation
  2. non-linear regression

Non-linear and non-parametric regression

Non-linear regression:

Special case polynomial regression

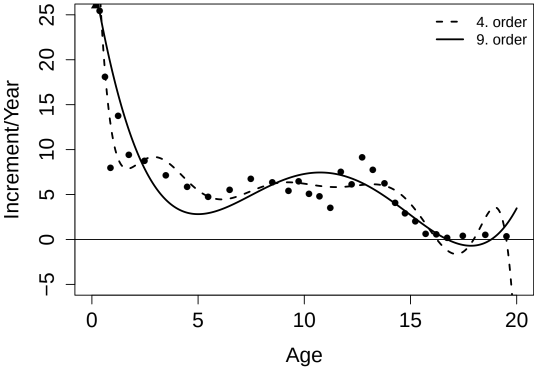
$=$ multiple linear regression
independent variable $(x - \bar{x}), (x - \bar{x})^2, \ldots, (x - \bar{x})^k$

Non-parametric regression:

- smoothing splines
- Gasser-Müller kernel estimator
- local linear estimator (LOWESS, LOESS)

Non-linear and non-parametric regression

Example: Growth data in form of increments



Polynomial 4. order: $R^2_{\text{adj}} = 0.76$

Polynomial 9. order: $R^2_{\text{adj}} = 0.93$

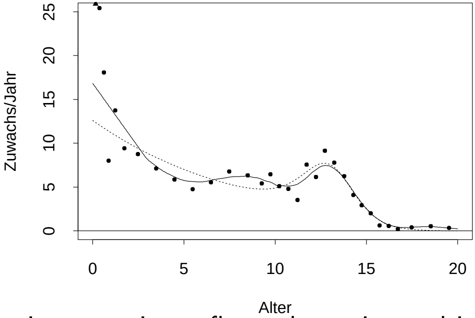
Non-linear and non-parametric regression

- Preece–Baines Modell (1978): $\cdots$

$$f(x) = a - \frac{4(a - f(b))}{[\exp\{c(x-b)\} + \exp\{d(x-b)\}]\,[1 + \exp\{e(x-b)\}]}$$

  – for increments the derivative is required.

- Gasser–Müller kernel estimator: —



- Non-parametric regression reflects dynamics and is better than the non-linear and polynomial regression.