Multiple Linear Regression

Janette Walde

janette.walde@uibk.ac.at

Department of Statistics
University of Innsbruck

Contents I

Contents II
Multiple linear regression

Non-random structure of the residuals

Origin of the word "Regression"

Sir Francis Galton (1822–1911), a famous geneticist, who studied the sizes of seeds and their offspring and the heights of fathers and their sons.

Tall fathers tend to have sons that are slightly smaller than the fathers. Sons of small fathers are on average larger than their fathers. He called this phenomenon "regression towards mediocrity".

A scatterplot of the heights of 1078 sons versus the heights of their fathers.
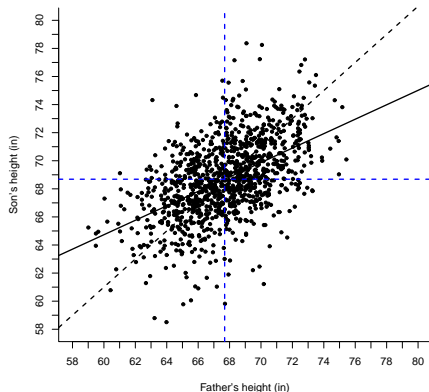
Note: $\bar{y}_S \approx \bar{x}_F + 1$.

However, the regression line $E[y] = b_0 + b_1 x$ shows a slope b_1 less than 1 (black line).

This "regression effect" must be taken into account in test-retest situations.

Suppose a group of preschool children are given an IQ test at age four and another test at age five. The results of the tests will certainly be correlated, and according to above, children who do poorly on the first test will tend to score higher on the second test. If, on the basis of the first test, low-scoring children were selected for supplementary educational assistance, their gains might be mistakenly attributed to the program. A comparable control group is needed in this situation to tighten up the experimental design.

Notation of the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$\text{or}$$
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, ..., n$$

β_0 ... population intercept

β_1 ... population slope,

measures the change in Y per unit change in X

ε_i ... random/unexplained error,

associated with the ith observation

Decomposition of y

- Linear functional form:

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

Decomposition of y

- ► Linear functional form:
  - ► Decomposition in a systematic component, explained by variable X

$$y_i = \underline{\beta_0 + \beta_1 x_i} + \varepsilon_i$$

Decomposition of y

- Linear functional form:
  - Decomposition in a systematic component, explained by variable X
  - and an unexplained component ε.

$$y_i = \beta_0 + \beta_1 x_i + \underline{\varepsilon_i}$$

Decomposition of y

- Linear functional form:
  - Decomposition in a systematic component, explained by variable X
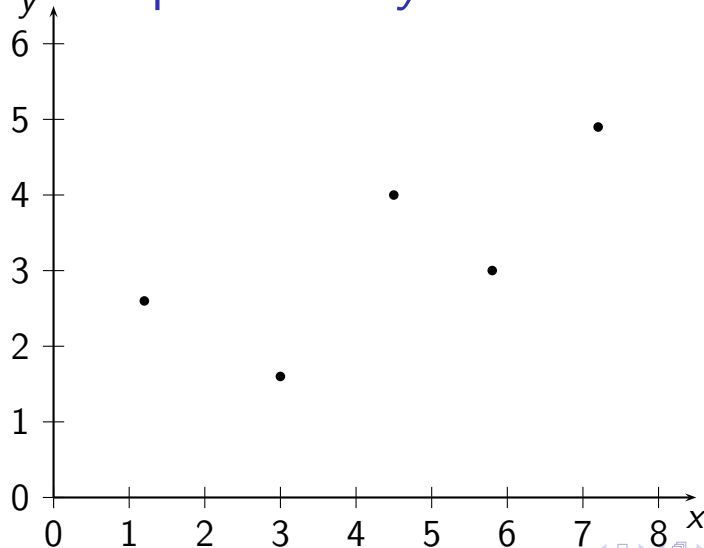 - and an unexplained component ε.

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

- The derivation

$$\beta_1 = \frac{dy}{dx}$$
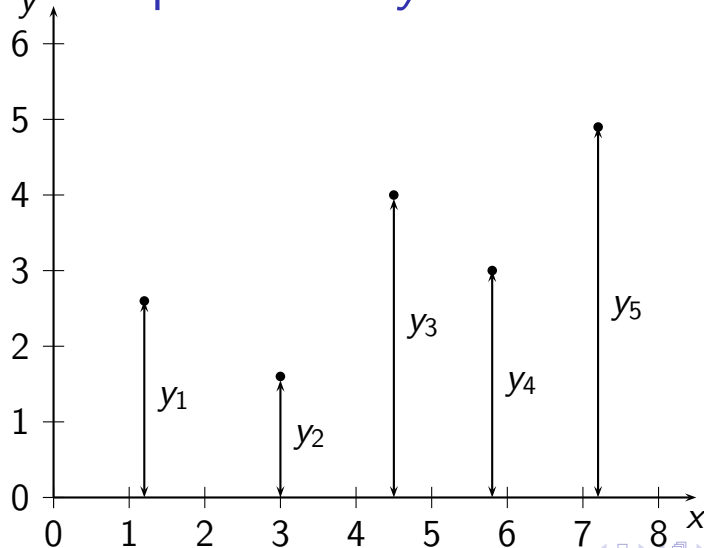
gives the marginal impact of a change in X.

Decomposition of y

Decomposition of y

Decomposition of y
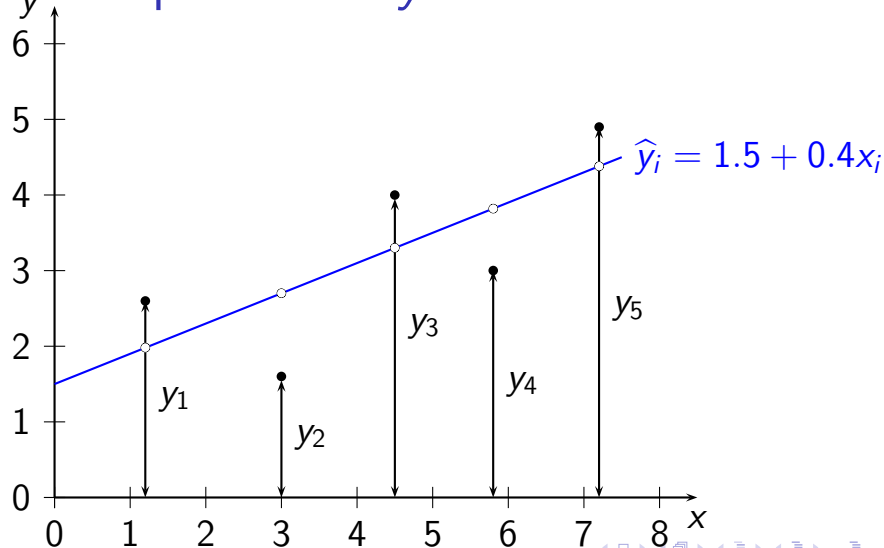
$$\widehat{y}_i = 1.5 + 0.4x_i$$

Decomposition of y
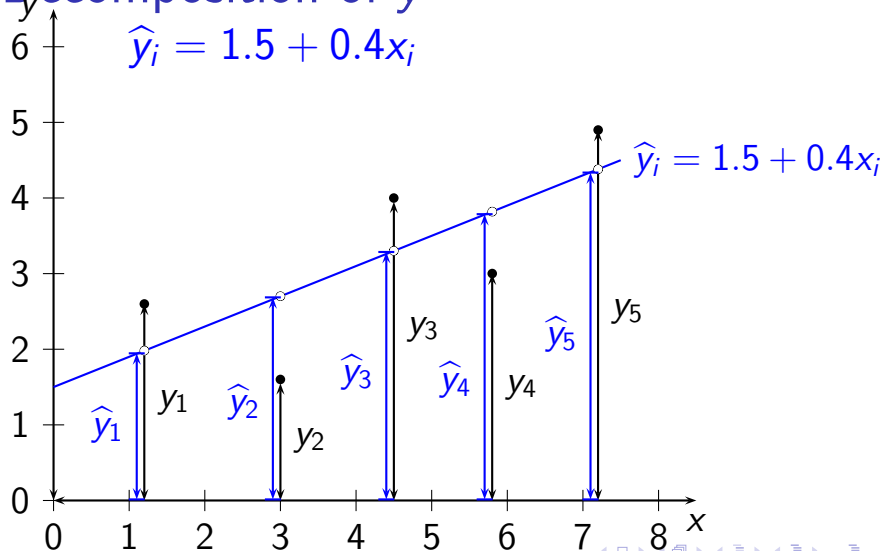
$$\widehat{y}_i = 1.5 + 0.4x_i$$

Decomposition of y

$$\widehat{y}_i = 1.5 + 0.4x_i$$
$$y_i = \widehat{y}_i + e_i$$
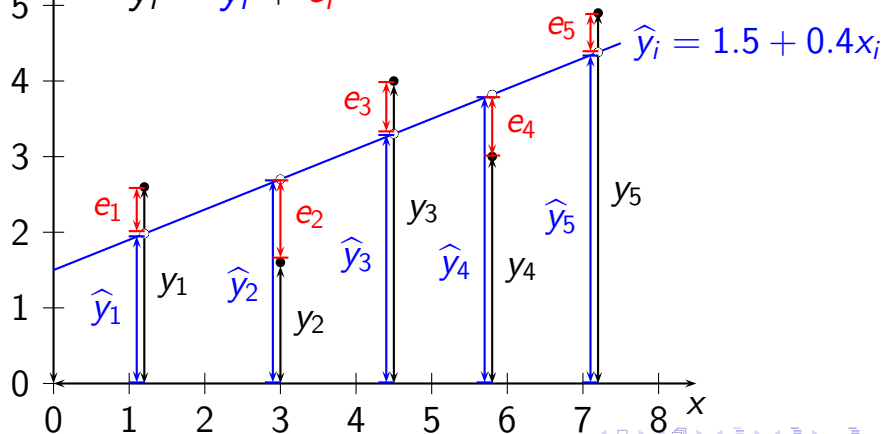
Computation of β_0 and β_1
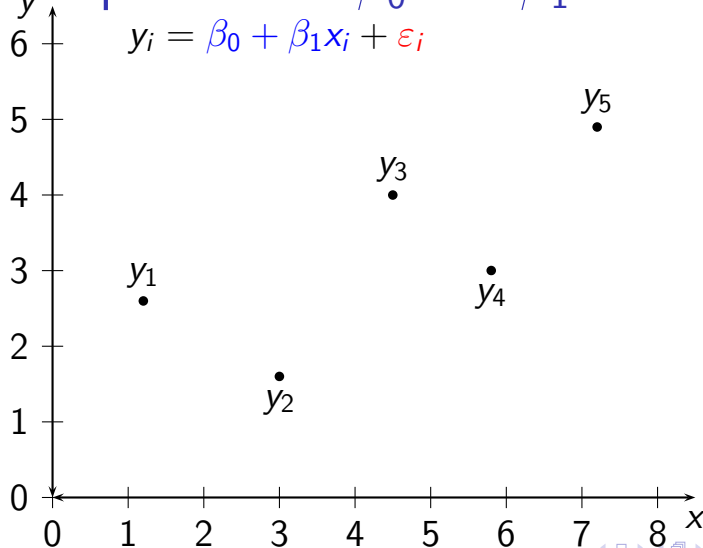
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Computation of β_0 and β_1
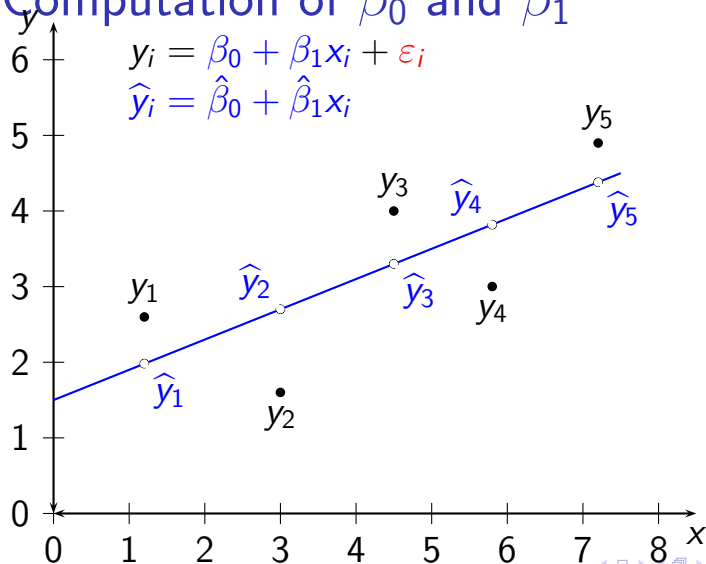
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Computation of β_0 and β_1

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

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$e_i = y_i - \widehat{y}_i$$

Computation of β_0 and β_1

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

$$\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$e_i = y_i - \widehat{y}_i$$

$$\min \sum_i e_i^2 = \sum_i (y_i - \widehat{y}_i)^2$$
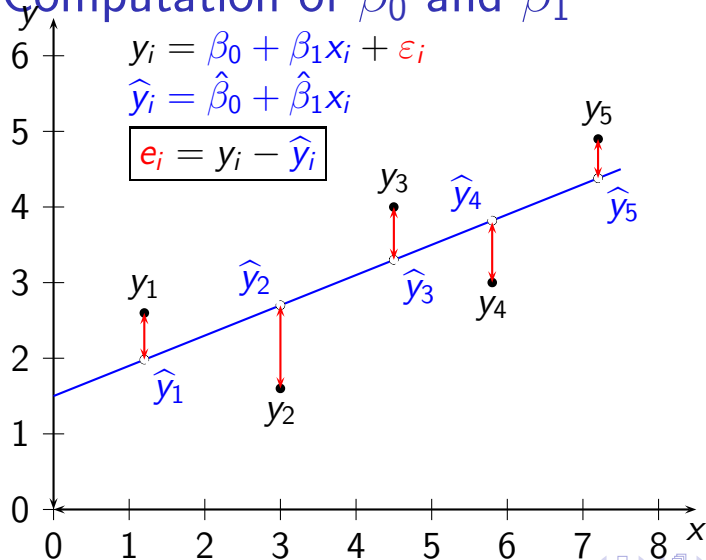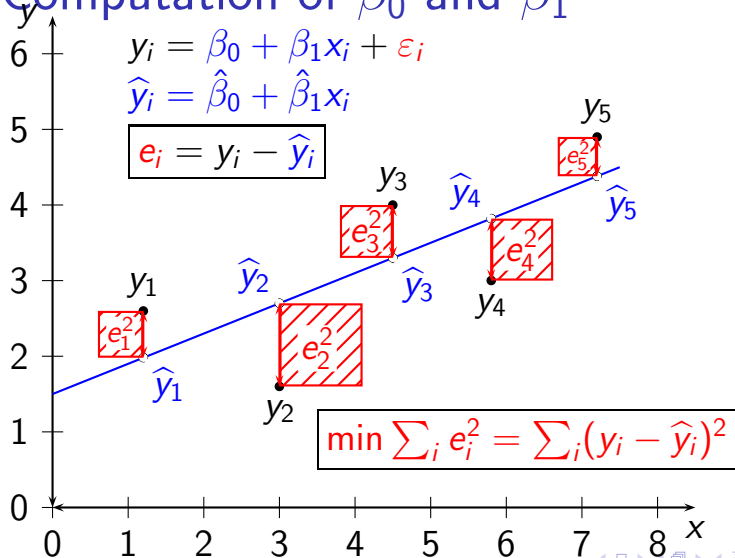
OLS estimate for the model parameters

	OLS estimate	Standard error
β_1	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$se_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}_\varepsilon}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
β_0	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$se_{\hat{\beta}_0} = \sqrt{\hat{\sigma}_\varepsilon \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
ε_i	$e_i = y_i - \hat{y}_i$	$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}e_i^2}$

Confidence intervals: $\hat{\beta}_1 \pm t_{\alpha, df=n-2} \cdot se_{\hat{\beta}_1}$

Hypothesis testing: $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$ with the test statistic $t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}}$ that is t-distributed with $df = n - 2$.

Coefficient of determination

A measure of association between Y and X is R^2 or the coefficient of determination which measures the proportion of the total variation in Y that is explained by its linear relationship with X.

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

Note:

- ▸ Do not compare models with different data transformations using R^2!
- ▸ A model with more explanatory variables will always have a higher R^2.

Anscombe data
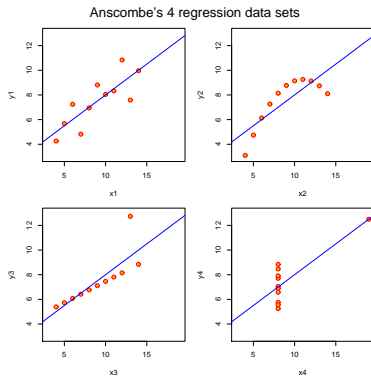
Same intercept, slope and confidence bands! $F-$statistics and $t-$values are significant, for all four $R^2 = 0.67$.

Reference: Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27, 17–21.

Assumptions of regression analysis

The four assumptions that allow the sample data to be used to estimate the population parameter and to do inferential statistics:

Model specification:
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

1. Normality
2. Homogeneity of variance
3. Independence and
4. Fixed X.

Assessing the important assumptions

- ▶ Normality assumption can be checked using a histogram of the residuals or a Q-Q plot.
- ▶ Homogeneity can be assessed by plotting the residuals against X to check for any increases (decreases) in the spread of residuals along the $x-$axis. This procedure can also assess model misspecification and model fit.
- ▶ Plotting residuals against fitted values can show increases in the spread for larger fitted values: a strong indicator for heterogeneity.

Assessing the important assumptions

- ▶ Plotting residuals against the dependent variable demonstrate the quality of the model.
- ▶ As to independence, for time series data residuals can be checked additionally by the auto-correlation function.
- ▶ For spatial correlation in the data residuals can be checked by a spatial plot or by computing Moran I.

Influential points

- Leverage is a tool that identifies observations that have rather extreme values for the explanatory variables and may potentially bias the regression.

$$leverage_i = h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

- Cook's distance statistic identifies single observations that are influential on all regression parameters

$$D_i = \frac{\sum_j \left(\hat{y}_j - \hat{y}_{j(i)}\right)}{p \cdot MSE}$$

p... number of fitted parameters in the model
MSE... mean square error of the regression model

- Jacknife procedure and plot changes in the parameters.

Studentized residuals

- The studentized residuals are defined as

$$\frac{(y_i - \hat{y}_i)}{\sqrt{\hat{\sigma}_\varepsilon^2 (1 - h_i)}}$$

- Taking the square root of the absolute values reduces the skewness and makes non-constant variance more noticeable.

Studentized deleted residuals

▸ To obtain the ith Studentized residual the regression model is applied on all data except for observation i, and the $\hat{\sigma}_\varepsilon$ is based on the $n - 1$ points (but not residual e_i or hat value h_i. These are based on the full data set.). If the ith studentized residual is much larger than a standardized residual, then this is an influential observation because the variance without this point is smaller.

Types of residuals for regression diagnostics

Residual	$e_i = y_i - \hat{y}_i$
Standardized residual	$\frac{e_i}{\hat{\sigma}_\varepsilon}$
Studentized residual	$\dfrac{e_i}{\hat{\sigma}_\varepsilon \sqrt{(1-h_i)}}$
Studentized deleted residual	$e_i \sqrt{\dfrac{n-1}{SS_{residual}(1-h_i)-e_i^2}}$

where h_i is the leverage for observation i.

Application

Using a Case Study published in "Analysing Ecological Data (Statistics for Biology and Health)" by A.F. Zuur, E.N. Ieno, G.M. Smith, Springer 2007, New York.

The Dutch governmental institute **RIKZ** started a research project whose aim was to find relationships between macrofauna of the intertidal area and abiotic variables (e.g., sediment composition, slope of the beach). Sampling was carried out in June 2002. Abundances of around 75 invertebrate species from 45 sites were measured on various beaches

Application, II

One of the collected variables was "**NAP**" which measured the height of the sample site compared with average sea level, and indicated the **time a site is under water**.

The species data was converted into a diversity index: **Shannon-Weaver index**. For these data the Shannon-Weaver index can also be seen as an indicator for the number of different species.

Application, III

Hypothesis: The tidal environment creates a harsh environment for the animals living there, and it is reasonable to assume that different species and species abundances will be found in beaches with different NAP values.

A simple starting point is therefore to compare species diversity (species richness) with the NAP values from different areas of the beach.

Scatterplot of the data
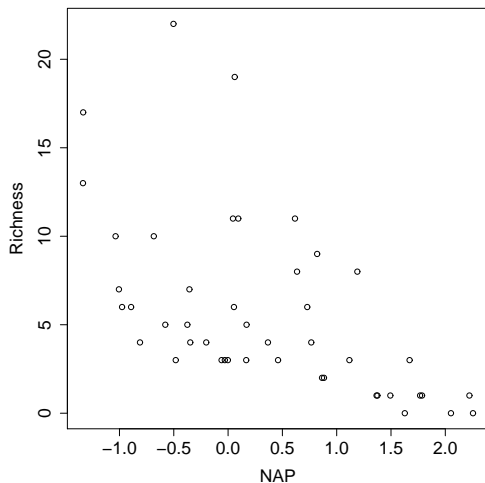
Regression line for the data

Regression line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
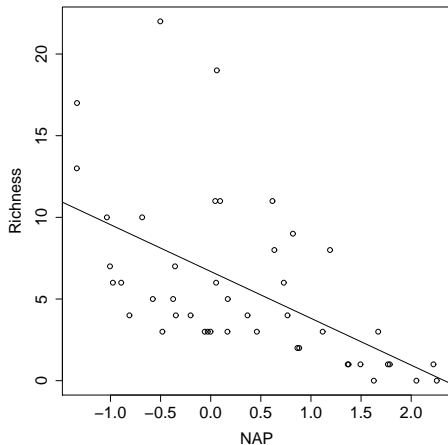
Model summary

Model	R	R^2	Adj. R^2	SE estimate
1	0.570	0.325	0.309	4.15993

ANOVA

	SS	df	MSS	F	Sig.
Regression	357.529	1	357.529	20.660	0.000
Residuals	744.115	43	17.305		
Total	1101.644	44			

Coefficients

	Reg. coeff.	Std. Error	Beta	T	Sig.
(constant)	6.686	0.658		10.164	0.000
NAP	-2.867	0.631	-0.570	-4.545	0.000

Residual diagnostic
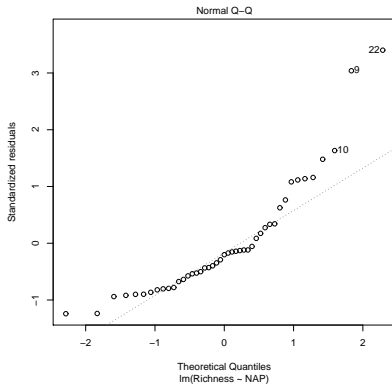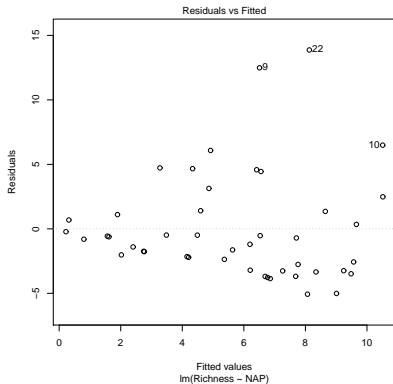
Leverage and Cook's distance statistic
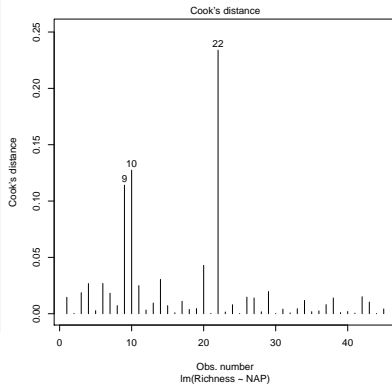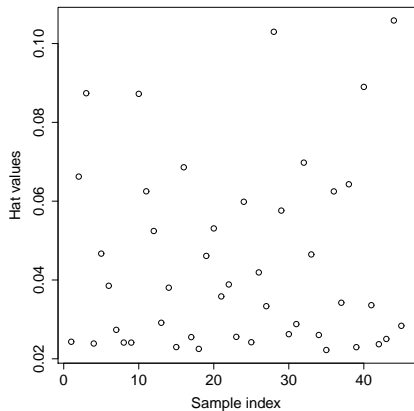
Jacknife procedure

Changes in intercept and slope
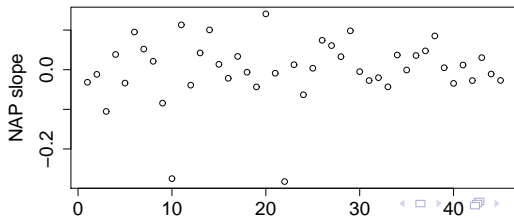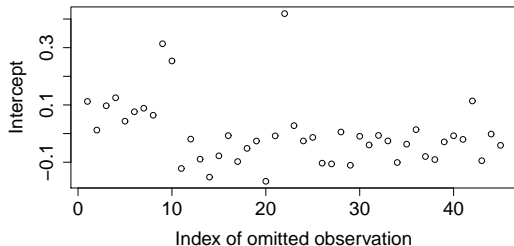
Plot of standardized and studentized residuals
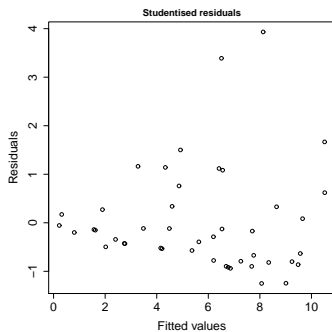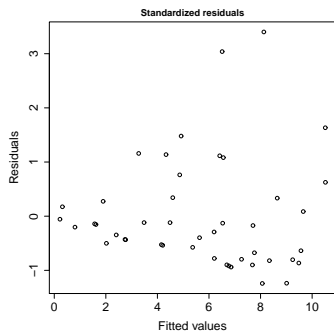
Multiple linear regression

$$y_i \;=\; \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \varepsilon_i$$

$$\varepsilon_i \text{ i.i.d. } N(0, \sigma_\varepsilon^2)$$

Application:

$$R_i \;=\; constant + \beta_1 NAP_i + \beta_2 Grainsize_i +$$
$$\beta_3 Humus_i + Week_i + \beta_4 Angle_i + \varepsilon_i$$

Residuals:

Min	1Q	Median	3Q	Max
-5.0454	-1.2865	-0.3314	0.7048	12.0917

Coefficients:

	Reg. coeff.	Std. Error	T	Sig.
(constant)	9.298448	7.967002	1.167	0.250629
angle2	0.016760	0.042934	0.390	0.698496
NAP	-2.274093	0.529411	-4.296	0.000121
grainsize	0.002249	0.021066	0.107	0.915570
humus	0.519686	8.703910	0.060	0.952710
factor(week)2	-7.065098	1.761492	-4.011	0.000282
factor(week)3	-5.719055	1.827616	-3.129	0.003411
factor(week)4	-1.481816	2.720089	-0.545	0.589182

— — —

Multiple R-squared: 0.679, Adjusted R-squared: 0.6182

F-statistic: 11.18 on 7 and 37 dF, Sig.: 0.000

Multiple Linear Regression
└ Multiple linear regression
  └ Standardized regression slopes

Standardized regression slopes

The value of $\hat{\beta}_i^*$ provides an estimate of the slope of the regression model that is independent of the units of X and Y and is useful in comparing regression slopes.

$$\hat{\beta}_i^* = \hat{\beta}_i \frac{s_X}{s_Y}$$

Multicollinearity

Multicollinearity is the phenomenon in which two or more
predictor variables in a multiple regression model are highly
correlated. In this situation the coefficient estimates may
change erratically in response to small changes in the model or
the data. Multicollinearity does not reduce the predictive
power or reliability of the model as a whole; it only affects
calculations regarding individual predictors. That is, a multiple
regression model with correlated predictors can indicate how
well the entire bundle of predictors predicts the outcome
variable, but it may not give valid results about any individual
predictor, or about which predictors are redundant with others.

Detecting multicollinearity I

- Tolerance for X_j is simply $1 - R_j^2$ from the OLS regression of X_j against the remaining $p - 1$ predictor variables. An approximate guide is to worry about tolerance values less than 0.1.

- Tolerance is sometimes expressed as the variance inflation factor (VIF) which is simply the inverse of tolerance.

Detecting multicollinearity II

► The principal components from the correlation matrix among the predictor variables can be extracted. Principal components with eigenvalues near zero indicate collinearity among the original predictor variables.

  ► Condition index > 30 indicates collinearity

  ► Eigenvalues < 0.5 indicate collinearity problems

Model selection

- Akaike information criteria
  $$AIC = n \cdot log(SS_{residual}) + 2(p+1) - n \cdot log(n)$$

- $AdjustedR^2 = 1 - \frac{SS_{residual}/(n-(p+1))}{ss_{total}/(n-1)}$

- Forward selection, backward selection, or a combination of forward and backward selection. Note:
  - multiple comparisons
  - collinearity
  - explorative analysis

Partial linear regression

$$Y_i = constant + \beta_1 X_i + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i$$

Pure X effect?

Filtering out the effects of W and Z:

1. $Y_i = constant + \beta_4 W_i + \beta_5 Z_i + \varepsilon_{Yi}$

2. $X_i = constant + \beta_6 W_i + \beta_7 Z_i + \varepsilon_{Xi}$
   ε_X can be seen as the information in X after filtering out the effects of W and Z.

Partial linear regression

3. $\varepsilon_{Yi} = \beta\varepsilon_{Xi} + noise_i$ This model shows the relationship between Y and X after partialling out the effect of W and Z. Hence, the regression model of ε_Y on ε_X shows the pure X effect.

$\Rightarrow$ Partial regression plot and interpret the significance of the slope.

Part correlation coefficient/decomposition of the variation

To obtain the variance components, Legendre and Legendre (1998) gave the following algorithm:

1. Apply the linear regression model $Y_i = constant + X_i\beta + W_i\gamma + noise_i$ and obtain R^2. This is equal to $[a + b + c]$, and $[d]$ is equal to $1 - [a + b + c]$.

2. Apply the linear regression model $Y_i = constant + X_i\tilde{\beta} + noise_i$ and obtain R^2. This gives $[a + b]$.

Part correlation coefficient/decomposition of the variation

3. Apply the linear regression model
   $Y_i = constant + W_i\tilde{\gamma} + noise_i$ and obtain R^2.
   This gives $[b + c]$.

The following computation gives
$[b] = [a + b] + [b + c] - [a + b + c]$.

Note: Semi-partial correlation coefficient
Note: Hierarchical partitioning

Variance partitioning for the RIKZ data

The following model was fitted

$$R_i = \text{constant} + \beta_1 NAP_i + \beta_2 Angle_i$$
$$+ \beta_3 Exposure_i + noise_i$$

All regression parameters are significantly different from zero at the 5% level.

What is the pure NAP effect?

Variance partitioning for the RIKZ data

1. $R_i = constant + \beta_1 NAP_i + \beta_2 Angle_i + \beta_3 Exposure_i + noise_i \rightarrow R^2 = 0.636 \rightarrow [a + b + c] = 0.637$ and $[d] = 1 - 0.637 = 0.364$.

2. $R_i = constant + \tilde{\beta}_1 NAP_i + noise_i \rightarrow R^2 = 0.325 \rightarrow [a + b] = 0.325$

3. $R_i = constant + \tilde{\beta}_2 An\tilde{g}le_i + \tilde{\beta}_3 Exposure_i + noise_i \rightarrow R^2 = 0.347 \rightarrow [b + c] = 0.344$

Hence, $[b] = 0.325 + 0.344 - 0.636 = 0.033$, and therefore $[a] = 0.292$ and $[c] = 0.311$.

The pure NAP effect is 29.2%, meaning that 29% of the variation in species richness is explained purely by NAP.

Non-random structure of the residuals

1. Apply a transformation.
2. Add other explanatory variables.
3. Add interactions.
4. Add non-linear terms of the explanatory variables (e.g., quadratic terms).
5. Use smoothing techniques like additive modelling.
6. Allow for different spread using generalized least squares (GLS).
7. Apply mixed modelling.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Example: body weight [kg] and brain weight [g] of 62 mammals species (and 3 dinosaurs)

64 mammals including 3 dinosaurs

Multiple Linear Regression
└─ Non-random structure of the residuals
   └─ Apply a transformation
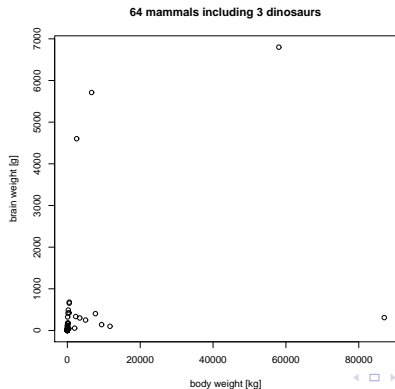
Example (cont.): Logarithmic scale

Multiple Linear Regression
└─ Non-random structure of the residuals
   └─ Apply a transformation
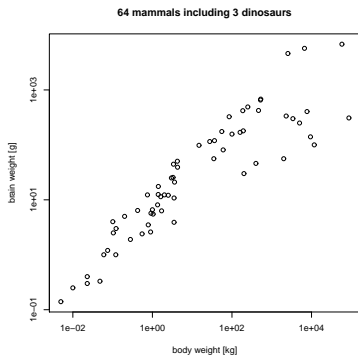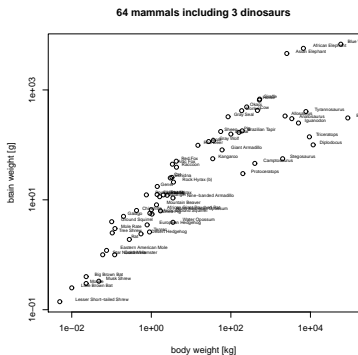
Example (cont.)

$brain\_weight = \beta_0 + \beta_1 body\_weight + \varepsilon$

Residuals:

Min	1Q	Median	3Q	Max
-633.8	-202.0	-193.6	-125.2	4679.0

Coefficients:

	Estimate	Std. Error	T	Sig.
(constant)	204.42722	111.15156	1.839	0.071
body_weight	0.12452	0.01462	8.518	0.000

− − −

Multiple R-squared: 0.5557, Adjusted R-squared: 0.5481

F: 72.55 on 1 and 58 df, Sig.: 0.000

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Example (cont.)

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation
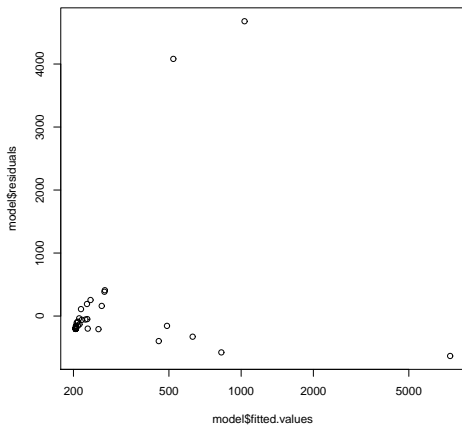
Variance-stabilizing transformation

We see that the residuals' variance depends on the fitted values (or the body weight): "heteroscedasticity".

The model assumes homoscedasticity, i.e. the random deviations must be (almost) independent of the explaining traits (body weight) and the fitted values.

Variance-stabilizing transformation: a logarithmic transformation scales body- and brain size to make deviations independent of variable.

Multiple Linear Regression
└─ Non-random structure of the residuals
  └─ Apply a transformation

Cause for heteroscedasticity

Actually not so surprising: An elephant's brain of
typically 5 kg can easily be 500 g lighter or heavier
from individual to individual. This cannot happen
for a mouse brain of typically 5 g. The latter will
rather also vary by 10%, i.e. 0.5 g. Thus, the
variance is not additive but rather multiplicative.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Cause for heteroscedasticity

brain mass = (expected brain mass) · random

We can convert this into something with additive randomness by taking the log:

log(brain mass) = log(expected brain mass) + log(random)

Multiple Linear Regression
└─Non-random structure of the residuals
　└─Apply a transformation

Example: Logarithmic transformation

$\log(\text{brain\_weight}) = \beta_0 + \beta_1 \log(\text{body\_weight}) + \varepsilon$

Residuals:

Min	1Q	Median	3Q	Max
-2.6500	-0.4957	0.0391	0.6776	1.6083

Coefficients:

	Estimate	Std. Error	T	Sig.
(constant)	1.93277	0.12570	15.38	0.000
log(body_weight)	0.62395	0.02982	20.93	0.000

− − −

Multiple R-squared: 0.883, Adjusted R-squared: 0.881

F: 437.9 on 1 and 58 df, Sig.: 0.000

Example (cont.)
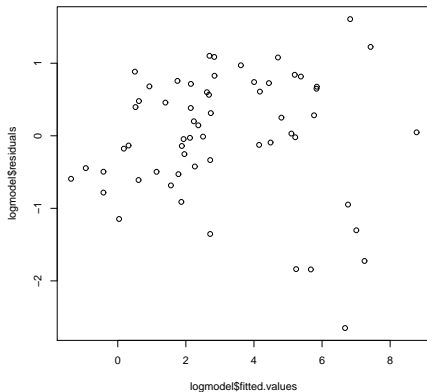
Aims of data transformations I

▸ Make the data and the model error terms closer to a normal distribution (i.e. make the distribution of the data symmetrical).

▸ Reduce any relationship between the mean and the variance (i.e. improve homogeneity of variances), often as a result of improving normality.

▸ Reduce the influence of outliers (extreme points), especially when they are at one end of the distribution.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Aims of data transformations II

- Improve linearity in regression analysis.
- Make effects that are multiplicative on the raw scale additive on the transformed scale, i.e. reduce the size of the interaction effects.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Types of transformations I

- ▶ The most common type of transformation is the power transformation: $Y \rightarrow Y^p$.
  - ▶ For data with right skew, the quare root transformation is applicable, particularly for data that are counts (Poisson distribution) and the variance is related to the mean.
  - ▶ $p = 0.33$, $p = 0.25$ etc. will be increasingly effective for data that are increasingly skewed. $p = 0.25$ are commonly used for abundance data in ecology when there are a lots of zeros and a few large values.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Types of transformations II

- ▸ A reciprocal transformation can also help although interpretation is a little bit difficult then order of values is reversed.

▸ Transforming data to logarithms ($log(Y + c)$) will also make right skewed distributions more symmetrical, especially when the mean is related to the standard deviation.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Types of transformations III

- If skewness is actually negative, i.e. the distribution has a long left tail, the variable can be reflected before transforming. Reflection simply involves creating a constant by adding one to the largest value in the sample and then subtracting each observation from this constant.

- These transformations can be considered part of the Box-Cox family of transformations: $\frac{Y^\lambda - 1}{\lambda}$, when $\lambda \neq 0$ (optimal λ) and for $\lambda = 0$ the transformation is $log(Y)$.

Multiple Linear Regression
└─Non-random structure of the residuals
   └─Apply a transformation

Types of transformations IV

▸ For percentages of proportions (between 0%
  and 100%) power transformations don't work
  because they change each end of the
  distribution differently. One common approach
  is to use the angular transformation,
  specifically the *arcsin* transformation.

Multiple Linear Regression
└─ Non-random structure of the residuals
  └─ Apply a transformation

Types of transformations

$y^2, y^3, \ldots$

$\log(x), \sqrt{x}, x^{0.3}, \ldots$

$x^2, x^3, \ldots$

$\log(y), \sqrt{y}, y^{0.3}, \ldots$
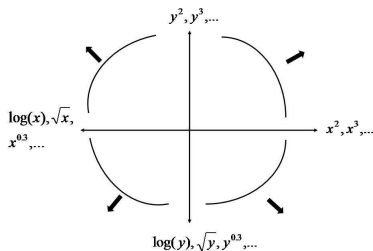
Mosteller and Tukey's bulging rule

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics.*

Addison-Wesley, Reading, MA.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

Changing hypothesis

- ▶ A transformation changing your response variable changes your formal null hypothesis.
- ▶ You might hypothesize that growth of plants varies with density and formalize that as:
  H_0: mean growth of plants at high density equals mean growth at low density.
- ▶ You are forced to log-transform your data, the null hypothesis becomes 'mean log-growth does not vary with density', or you might say that in the first case, growth is defined as mg of weight gained, whereas after log-transforming, growth is the log-mg weight gained.

Multiple Linear Regression
└─Non-random structure of the residuals
  └─Apply a transformation

A change in raw scale or a change in percentage scale?

Means for treatment and control groups for an experiment conducted at two times. (Artificial data)

| | Untransformed | | Log-transformed | |
	Time 1	Time 2	Time 1	Time 2
Control	10	50	1.000	1.699
Treatment	5	25	0.699	1.398

Transforming the data to a log scale changes the interpretation of the effect to a percentage change!

Multiple Linear Regression
└─ Non-random structure of the residuals
  └─ Allow for different spread using GLS

Allow for heterogeneity using variance structures

- $\varepsilon_i \sim N(0, X_i \cdot \sigma^2)$
- $\varepsilon_i \sim N(0, |X_i|^{2\delta} \cdot \sigma^2)$
- $\varepsilon_i \sim N(0, \sigma^2 \cdot exp(2\delta X_i))$
- $\varepsilon_i \sim N(0, \sigma_j^2)$

where δ is an unknown parameter.

The first three options allow for an increase (decrease) in residual variance depending on the values of the variance covariate X_i. The 4th option allows for different spread per level of a nominal variable.