Institute for Veterinary Epidemiology and Biostatistics (IVEB)
Department of Veterinary Medicine

Freie Universität Berlin

DRS Spring School – week 2

Model assumptions

PD Dr. Lorenz Gygax (HU Berlin)

Assumptions linear model

- Expected value:
  $\hat{Y}_i = a + b_1 * X_i^1 + b_2 * X_i^2 + b_3 * X_i^3 + \ldots$
- But, full model:
  $Y_i = a + b_1 * X_i^1 + b_2 * X_i^2 + b_3 * X_i^3 + \ldots + \varepsilon_i$
- $\varepsilon \sim N(\mu, \sigma^2)$, i.i.d. (or following a different distribution)
- independent identically distributed

↔ Assumptions of model

↔ If (clear) violations occur: e.g. p-values = «lottery numbers»

Independence

- Needs to be considered while planning an experiment!
  Not visible from the data!
- Is there evidence in the experimental design, that
  - one data point (residual) is informative for others, or
  - It is to be expected that groups of data are more similar within than between groups
- Indicators:
  - Temporal & spatial dependences (natural observations)
  - Individuality (e.g. **repeated measurements** on the same animal)
  - Grouping of measurements (blocks, farms, labs, groups)
- Good experiments have dependencies in their design!

Independence

- If there are dependencies in a data set and if these are not correctly dealt with, trouble is likely:
  - Degrees of freedom are overestimated
  - Statistical power is overestimated
  - There is an increase in the probability to reach a low p-value (and make an anti-conservative decision)
- All this is part of **pseudo-replication**!
  (one of the most serious statistical fallacies)
- On the other hand:
  True effects of interest may be lost in the (non-modelled) variability of block, groups, etc.

Independence

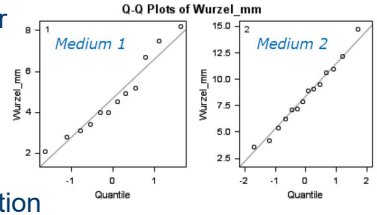
- Plan experiment with independent replicates
- Even better: learn to deal with dependencies (GLMMs)

Emergency solutions (**not recommended**):

- Averaging dependent data and evaluate means statistically. BUT:
  - Loss of information acquired with effort in time and money
  - Approach is also anti-conservative (SE < SD)
- Conduct multiple separate analyses with subsets of independent data. BUT:
  - The true question may not be answerable in doing so
  - Multiple testing (p-hacking).

What needs to be (normally) distributed?

- Often heard (and often tested):
  «The (raw) data have to be normally distributed»
- This is **NONSENSE**
- The assumptions are relevant for the residuals: $\varepsilon_i = Y_i - \hat{Y}_i$
  «Data given the model»
  (if structure based on explanatory variables has been deducted)
- This is the same as checking data for subgroups – disadvantage:
  Smaller sample size



- «identically» distributed
  - All residuals follow the same distribution
  - Variance of residuals is homogenous (homoscedastic)

Problems with statistically testing for normal distribution

- Cp. discussion on p-values
- (If at all) statistical tests are rejection tests
- low p-value «rejects» model (incl. H_0)
- a high p-value is NOT proof for H_0

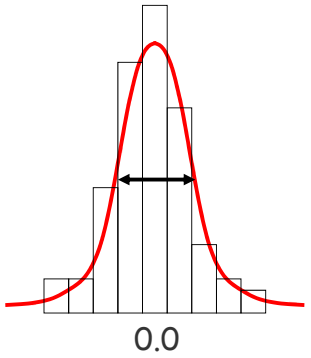
- If sample size is small:
  Testing for normal distribution will hardly ever detect a deviation («never» a low p-value)

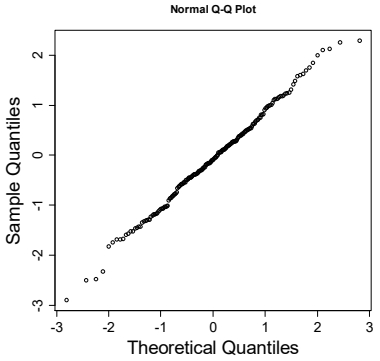
- If sample size is large:
  Testing for normal distribution will detect irrelevant deviations («always» a low p-value)

- Solution: graphical analysis of residuals

Checking distributional assumptions

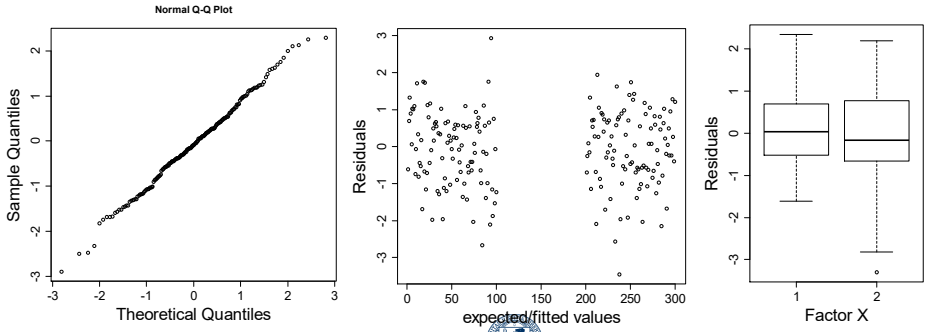
- Check residuals
- histogram quantile-quantile plot





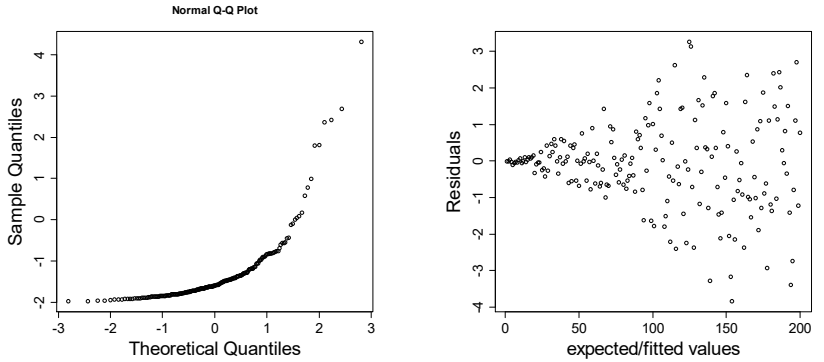
Checking distributional assumptions

- quantile-quantile Plot
- Tukey-Anscombe plot (residuals versus expected values)
- Residuals versus explanatory variables
- «Leverage» plot



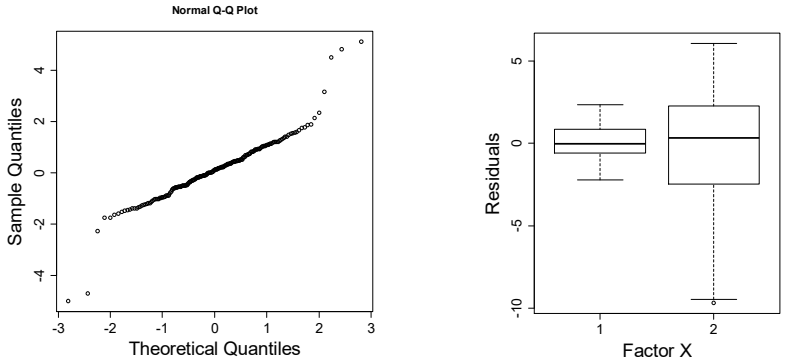
Typical deviations

- Long right tails in the distribution

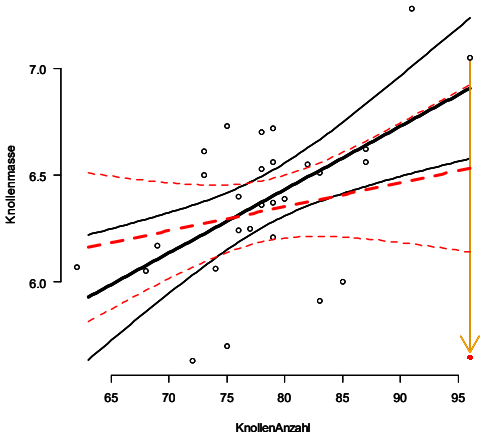


Typical deviations

- Outliers
- Inhomogeneity of variance («heteroscedasticity»):

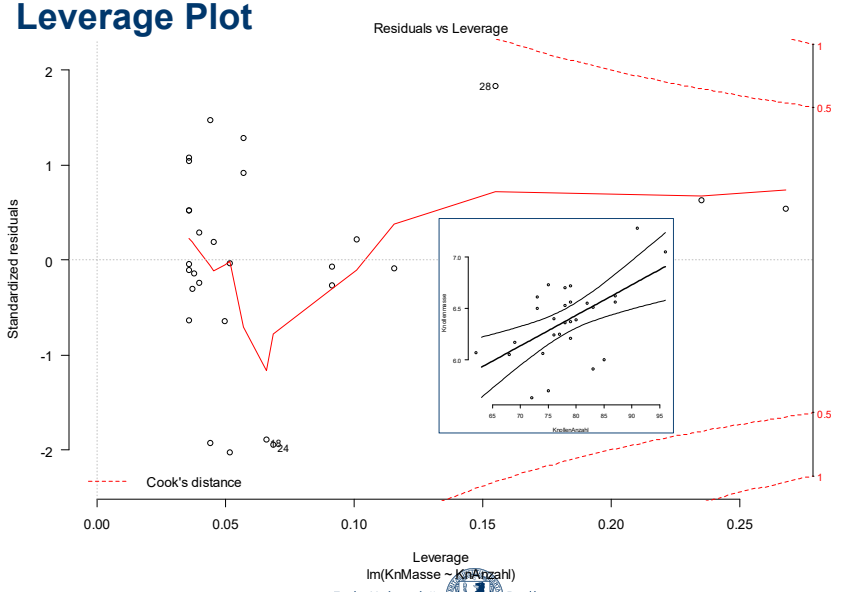


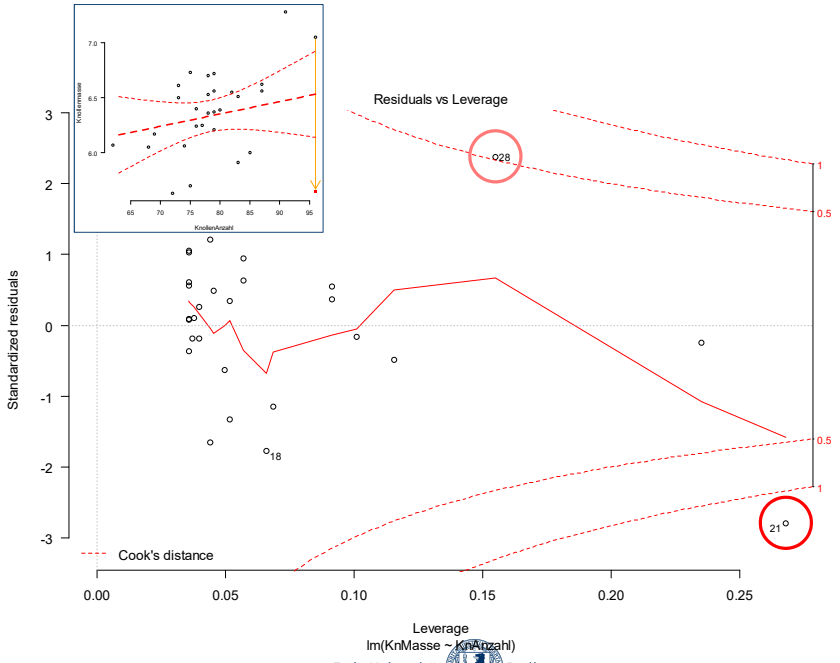
Leverage



Points with large Leverage («Hebelarm»)

Leverage Plot





What can be done if deviations occur?

- Use non-parametric tests (if experimental design simple)
- Omit outliers (?)
- Use transformations
- Use distributions other than the normal (generalised models)

Transformations & alternative distributions

What expectation do we have in respect to the distribution of the raw «data»?

	Tukey-first-aid	Alternative transformations	Alternative distributions
Amount	log (x)	log (x)	Log-normal
Count data	sqrt (x)	log (x + 0.5)	Poisson, Negative binomial, …
Proportions	asin (sqrt (x))	logit (x)	Bernoulli/Binomial Beta binomial, …

What happens due to the transformations

- How does the estimate of the outcome change if the explanatory variable changes by one unit?
- Untransformed model

  $Y_{x+1} = a + b (X + 1)$

  $Y_{x+1} = a + bX + b$

  $Y_{x+1} = Y_x + b$

  $Y_{x+1} - Y_x = b$

→ Effect of X is **additive**

→ Analogous argument to be used with factors

What happens due to the transformations

- Tukey First-aid
  - Count data

    $\text{sqrt}(Y_{x+1}) = a + b (X + 1)$

    $Y_{x+1} = (a + b (X + 1))^2$

    …
  - Proportions

    $\text{asin}(\text{sqrt}(Y_{x+1})) = a + b (X + 1)$

    $Y_{x+1} = \sin((a + b (X + 1))^2)$

    …

→ No simple interpretation of the parameter b

What happens due to the transformations

- Alternative transformations
  - Count data

    $\log(Y_{x+1}) = a + b (X + 1)$

    $Y_{x+1} = e^{a + bX + b}$

    $Y_{x+1} = e^a * e^{bX} * e^b$

    $Y_{x+1} = Y_x * e^b$

    $Y_{x+1} / Y_x = e^b$

  →Effect of X is **multiplicative**

What happens due to the transformations

- Alternative transformations
  - Proportions

    $\text{logit}(Y_{x+1}) = \log(Y_{x+1}/(1-Y_{x+1})) =$

    $\log(\text{odd}(Y_{x+1})) = a + b (X + 1)$

    $\text{odd}(Y_{x+1}) = e^{a + bX + b} = e^a * e^{bX} * e^b$

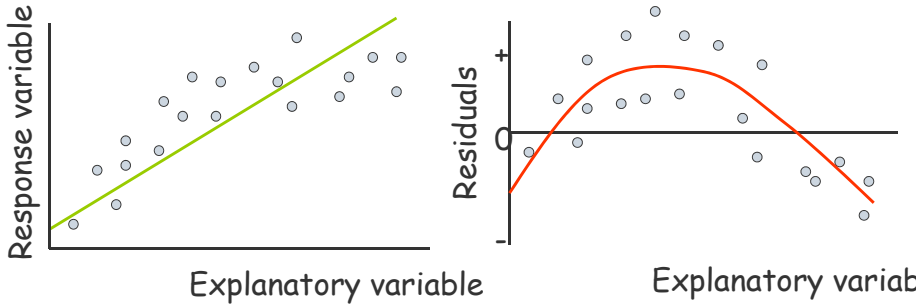
    $\text{odd}(Y_{x+1}) = \text{odd}(Y_x) * e^b$

    $\text{odd}(Y_{x+1}) / \text{odd}(Y_x) = e^b$

  →Effect of X can be expressed as an **«odds-ratio»**
  - Generating process?
    - Successes in number of trials (consider total number of trials, logistic model)
    - or e.g. proportion distance, length, time

Effects that had not been considered

- Plot residuals also versus potential explanatory variables
- Regression: does the assumption of linearity fit?



Final remarks assumptions

- Linear models are reasonably robust towards violations of distributional assumptions
- Central limit theorem:
  Expected value is always normally distributed if N large
- In a given sample:
  Clear violations should be dealt with

- For **each** (parametric) model that is analysed:
  - Are residuals independent?
  - Do the residuals follow a normal distribution closely?
  - Are residuals roughly homoscedastic?