




Know your data

Vesna Gagic



Data exploration

- 50% of the time spent on analysis
 - It is different from hypothesis testing and decisions about what models to test should be made a priori based on the researcher's biological understanding of the system
 - Data dredging (using aspects of a data exploration to search out patterns)- only as a guidance for future work
- 



Data exploration





- Check what are assumptions of the model you attempt to use
- Detailed data exploration to avoid wrong conclusions because of:

Type I error (discover a false effect)

Type II error (wrongly dismiss a variable)

Influential points

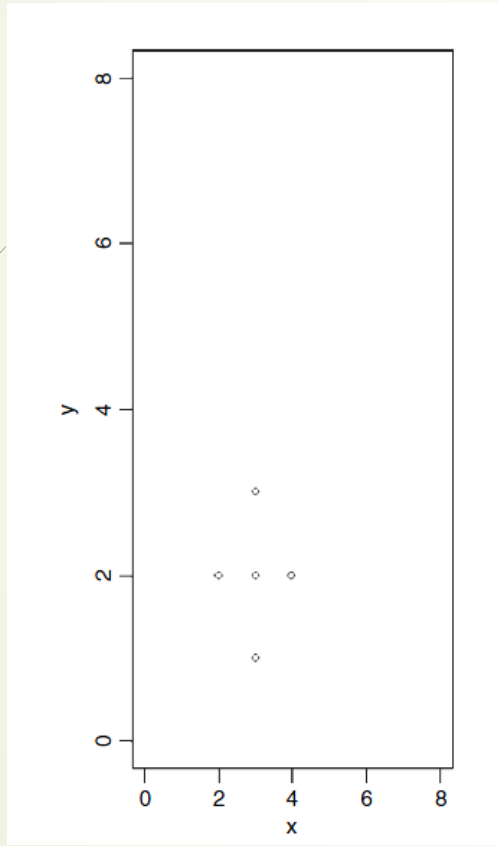
- The statistical literature, warns against certain tests and advocates graphical tools! (Montgomery & Peck 1992; Draper & Smith 1998, Quinn & Keough 2002)

- 
- 
- 1** Formulate biological hypothesis
Carry out experiment & collect data

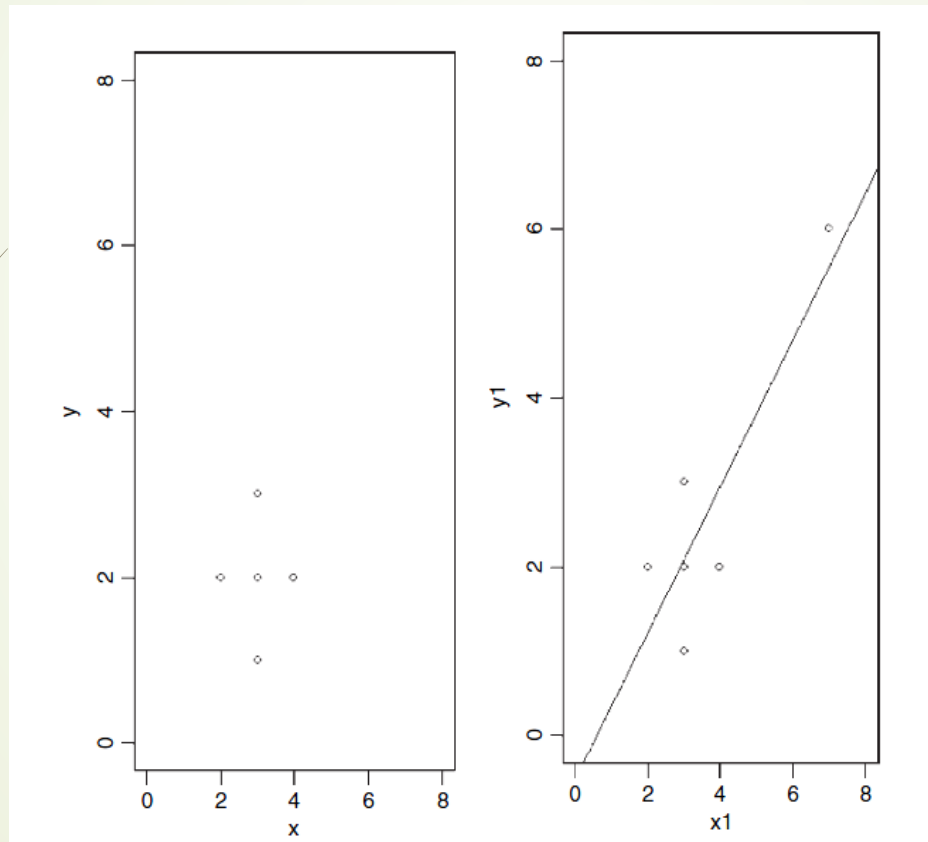
Data exploration

- | | | |
|----------|-----------------------------------|--|
| 2 | 1. Outliers Y & X | <i>boxplot & Cleveland dotplot</i> |
| | 2. Homogeneity Y | <i>conditional boxplot</i> |
| | 3. Normality Y | <i>histogram or QQ-plot</i> |
| | 4. Zero trouble Y | <i>frequency plot or corrgram</i> |
| | 5. Collinearity X | <i>VIF & scatterplots correlations & PCA</i> |
| | 6. Relationships Y & X | <i>(multi-panel) scatterplots conditional boxplots</i> |
| | 7. Interactions | <i>coplots</i> |
| | 8. Independence Y | <i>ACF & variogram plot Y versus time/space</i> |
- 3** Apply statistical model

Outliers, leverage and influential points



Outliers, leverage and influential points



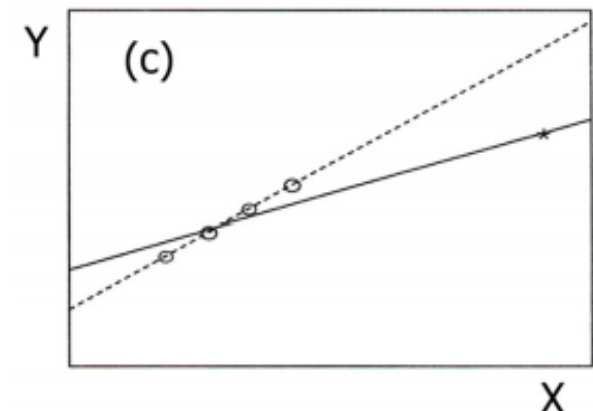
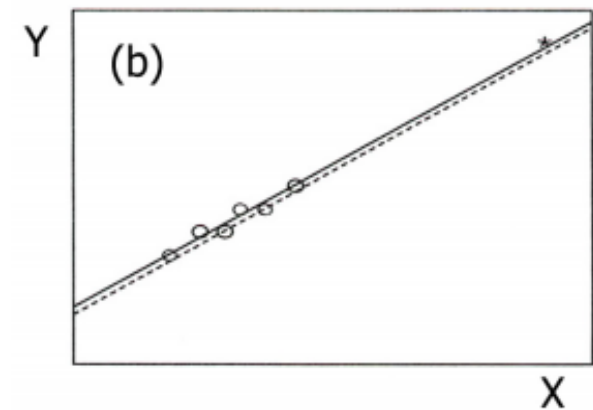
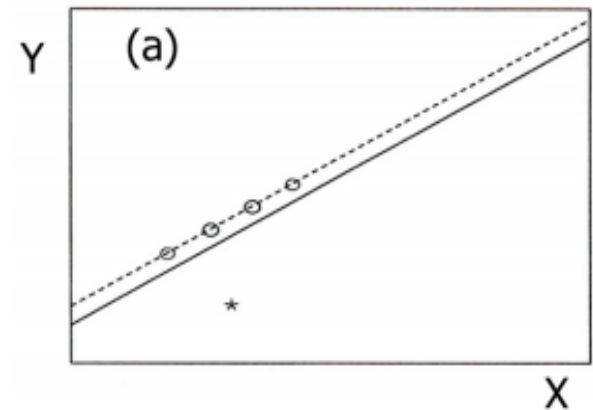



Outliers, leverage and influential points

- Outlier - observation that has a relatively large or small value compared to the majority of observations (**univariate outlier**) How to check it? -boxplot, Cleveland plot
- A **regression outlier** is an observation that has an unusual value of the dependent variable Y, conditional on its value of the independent variable X – In other words, for a regression outlier, neither the X nor the Y value is necessarily unusual on its own
- Note that a point may *appear* to be an outlier because of misspecification of the model!
- Small samples are especially vulnerable
- A regression outlier will have a large residual but not necessarily affect the regression slope coefficient
- **Leverage**-observations made at extreme values of the *independent variables* (it has potential to influence regression line) How to check it? – hat-value ($>2x$ average)
- **Influential point** - an observation that, if removed, significantly changes a statistical measure (How to check it? - Cook statistic in linear regression)
- Don't use analysis of residuals to look for influence! Why?
- Only when an observation has high leverage and is an outlier in terms of Y-value will it strongly influence the regression line

Types of Unusual Observations (4)

- **Figure (a): Outlier without influence.** Although its Y value is unusual given its X value, it has little influence on the regression line because it is in the middle of the X-range
- **Figure (b) High leverage** because it has a high value of X. However, because its value of Y puts it in line with the general pattern of the data it has **no influence**
- **Figure (c): Combination of discrepancy (unusual Y value) and leverage (unusual X value)** results in strong influence. When this case is deleted both the slope and intercept change dramatically.





How to deal with influential points?

-influential observations in response or explanatory variable

-Should you remove it?

-Make sure that the observation is correctly recorded

-Do the conclusions change when it's removed?


-Can we get more observations near that point?

How to deal with it if we don't want to exclude it?

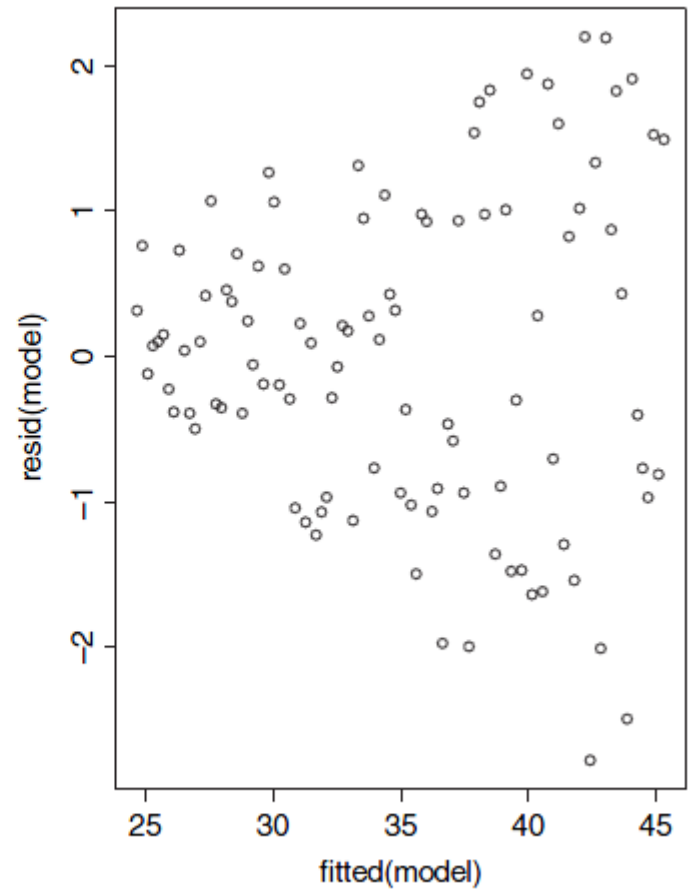
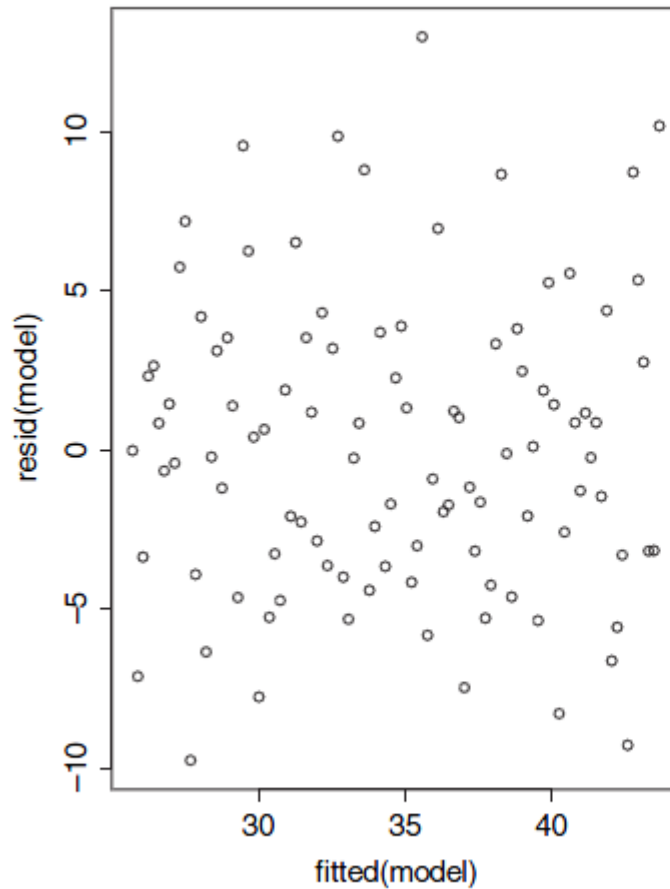
- Transformation, or choose different, robust test



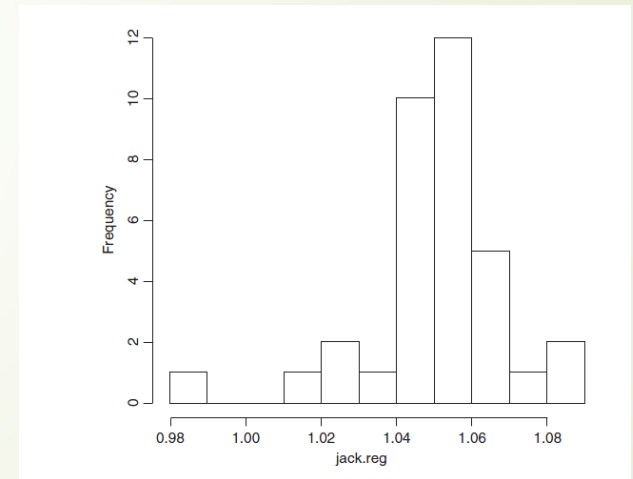
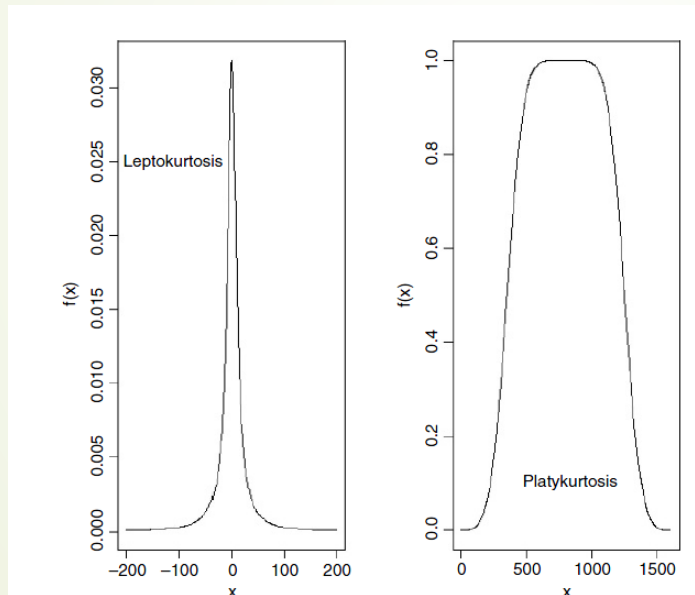
Homogeneity of variance

- Very important for majority of the analyses we do!
 - How to check it:
 - Plot residuals vs. fitted values and each variable
 - Conditional boxplots for the residuals
 - How to deal with it?
 - Transformation of the response
 - Variance function (better)
 - Using different tests
- 

Homogeneity of variance



Are the data normally distributed?





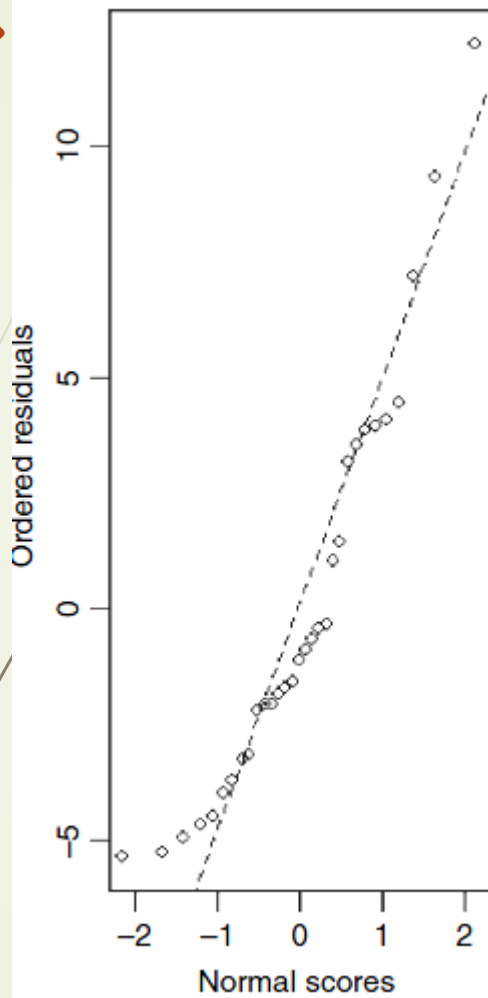
Are the data normally distributed?

- ▶ They may be skew (long tails to the left or right), or kurtotic (flatter or more pointy top)
- ▶ Most techniques are reasonably robust against violation of this assumption
- ▶ For small samples the power of the tests is low; and for larger data sets the tests are sensitive to small deviations
- ▶ How to check it?
 - histograms for the raw data if you do t-test
 - histograms of residuals or Q-Q ("Q" stands for quantile) plot if you do regression

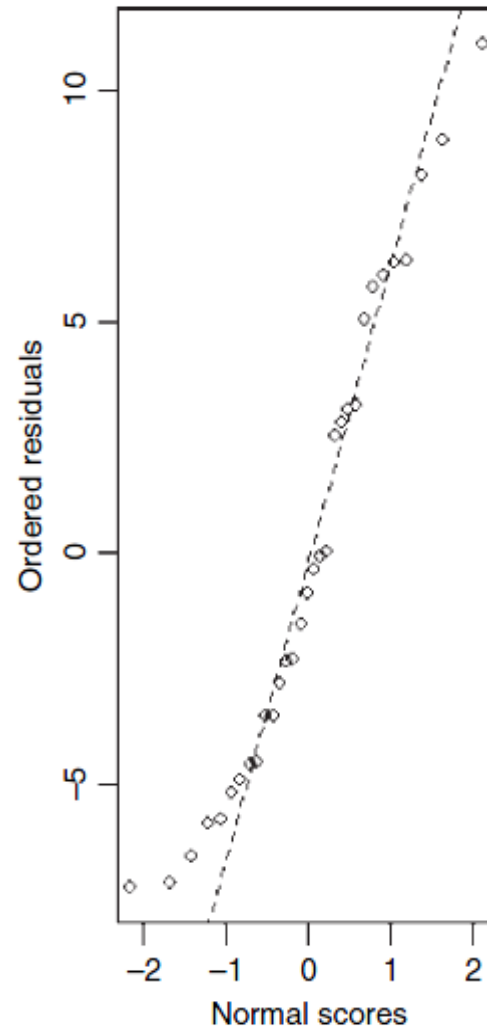
Note that quantiles are points taken at regular intervals from the cumulative distribution function (area under probability density function from $-\infty$ to x) of a random variable.

How to deal with it?

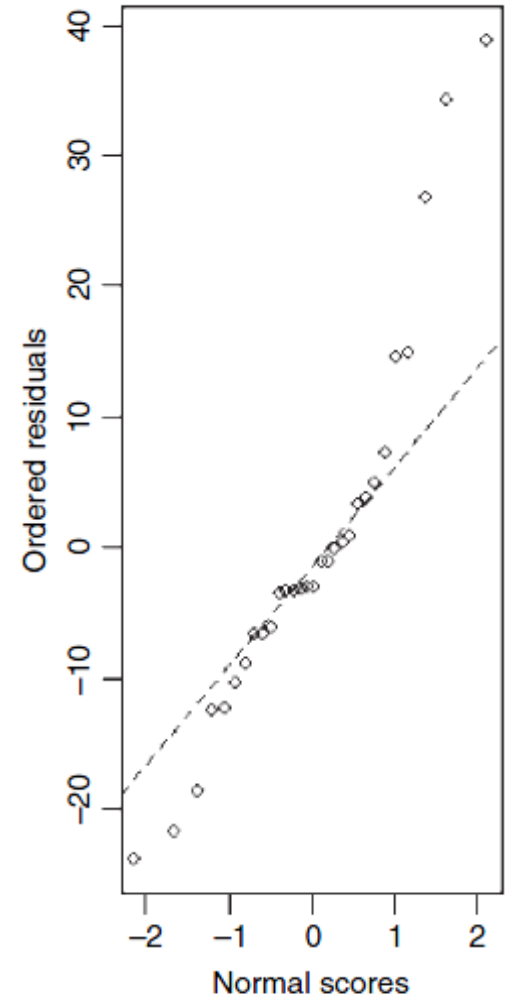
transformation



Binomial (J-shaped)



Uniform (S-shaped)



Gamma errors



Are there lots of zeroes in the data?

- Common in count data
- How to check it?
 - Histogram
- How to deal with it?
 - Use zero inflated GLMs

Is there collinearity among the covariates?

- Correlation between covariates
- Standard errors of the parameters are inflated –P-values get larger
 - Type II error
- How to check it?
 - Variance inflation factor (VIF) pairwise scatterplots comparing covariates, correlation coefficients
- How to deal with it?
 - Sequentially drop the covariate with the highest VIF until threshold is reached (10, 3, 2) or better, based on common sense or biological knowledge
 - Centering the input variables before fitting the model (Schielzeth 2010)
- Especially important when ecological signals are weak!

What are the relationships between Y and X variables?

- ▶ plotting the response variable vs. each covariate
- ▶ Note that the absence of clear patterns does not mean that there are no relationships; it just means that there are no clear two-way relationships.
- ▶ Is relationship linear?
- ▶ If not, you can use polynomial terms or different tests
- ▶ It is useful to center your data before analysis to increase interpretability of your coefficients (Schielzeth 2010)



Should we consider interactions?

- There should be a priori knowledge (hypothesis)
- How to check it?
 - Coplot
- If you have interaction, it is useful to center your data before analysis to increase interpretability of your coefficients , i.e. main effects (Schieleth 2010)
- Are the data balanced?
- In some cases the number of observations is very small or there are no data for some combinations!
- How to deal with it?
 - analyze only a subset of the data,
 - refrain from including certain interactions

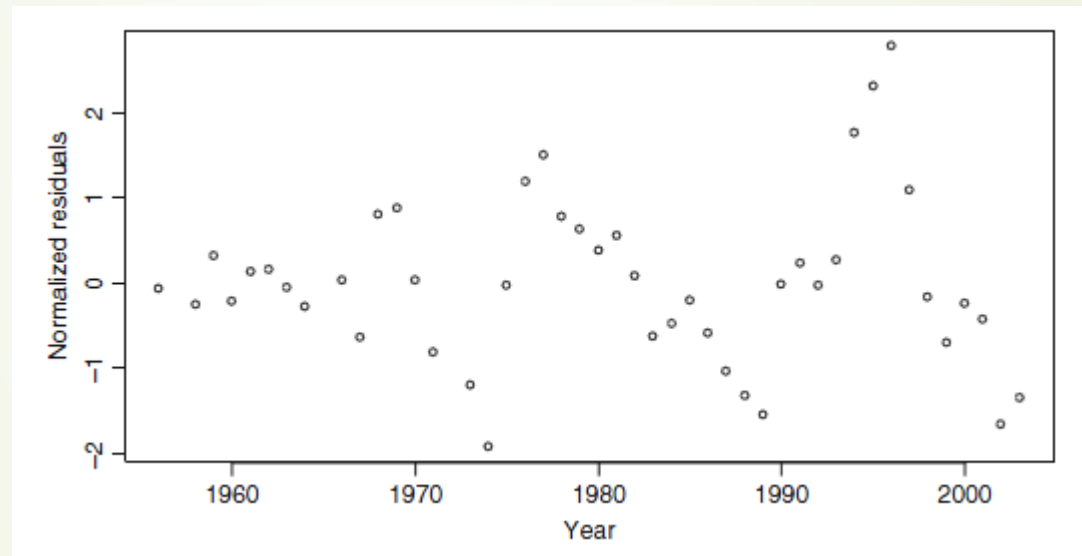


Are observations of the response variable independent?

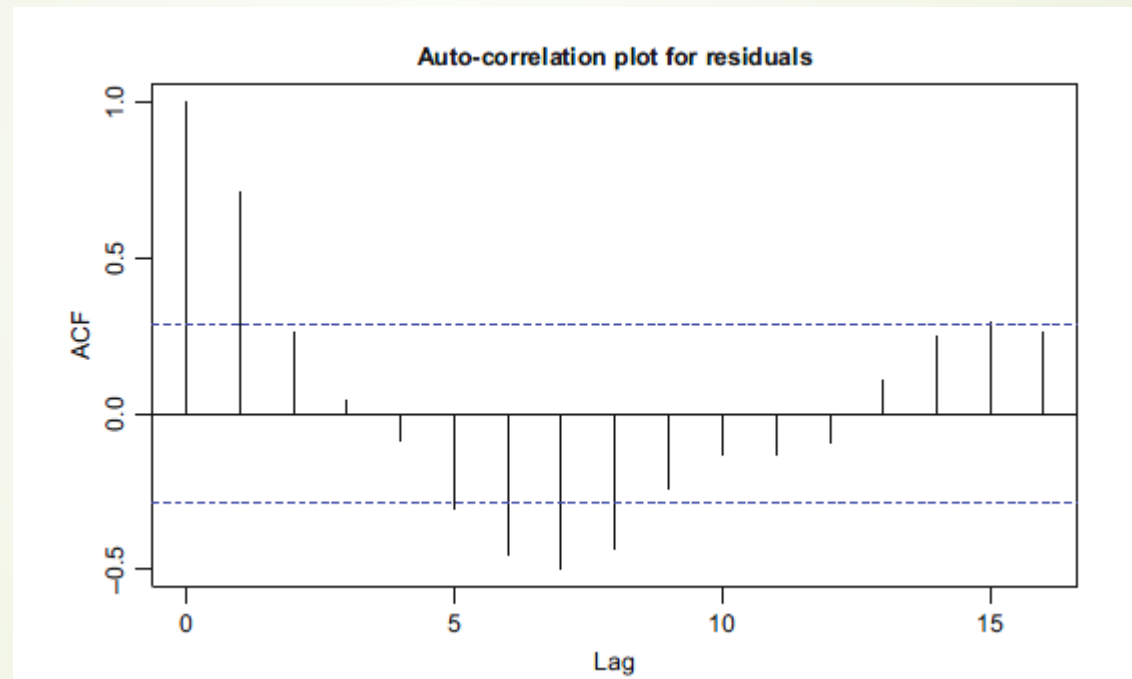
- Spatial or temporal relationships or e.g. multiple individuals of the same family (phylogenetic structure)
- violation may increase the type I error (rejecting the null hypothesis when it is true)
- dependence in the raw data before doing the analysis, and also the residuals afterwards
- How to deal with it?
 - mixed effects modelling
 - residual correlation structure using generalized leasts quares (gls) or lme

How to check it?

Plotting the response variable vs. time or spatial coordinates



Plot auto-correlation functions (ACF) for regularly spaced time series



Variograms for irregularly spaced time series and spatial data

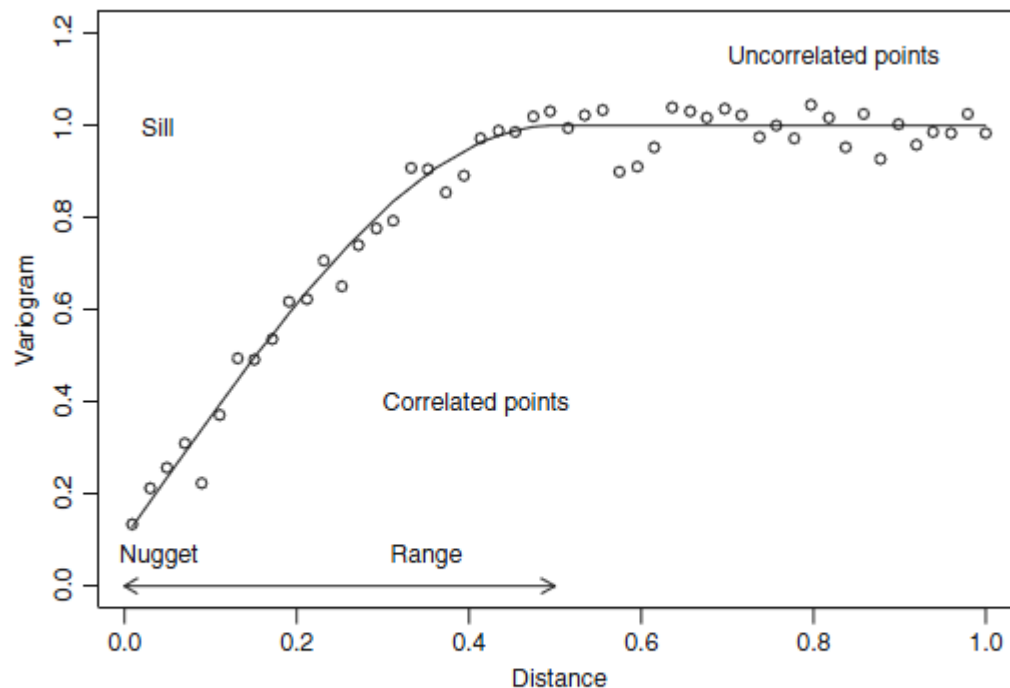


Fig. 7.2 Variogram with fitted line. The sill is the asymptotic value and the range is the distance where this value occurs. Pairs of points that have a distance larger than the range are uncorrelated. The nugget effect occurs if $\hat{\gamma}(h)$ is far from 0 for small h



Additional notes

- Not every data set requires each step
 - For example, PCA does not require normality
 - Some techniques can be used for several problems. For example, data transformation is used to reduce the effect of outliers, to stabilize the variance and to linearize relationships
- 