Statistics with Spa  ows

Lecture 16

Julia Schroeder

Julia.schroeder@imperial.ac.uk

Outline

- Model selection and simplification

Linear models

- Which variable is more important?
- Which should I leave in the model, and which not?
- When to use interactions?
- Do variables affect each other?

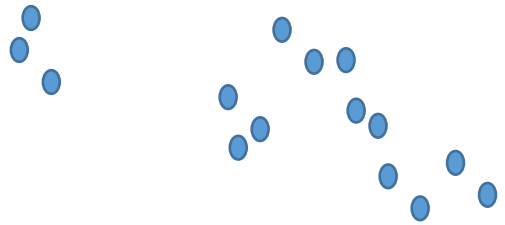
Linear models

- R^2 of model is not a good judge for which model is best
- Because the *fit* of a model increases with the number of parameters:

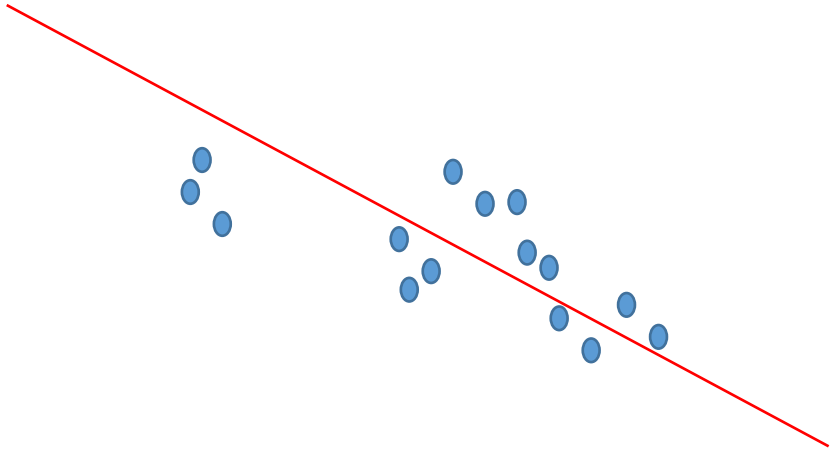
Fitting models to data

?

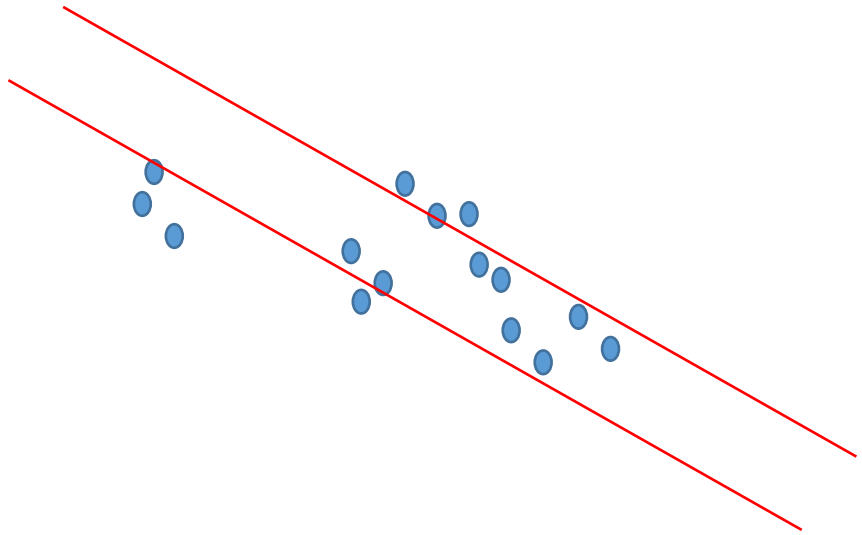
Fitting models to data



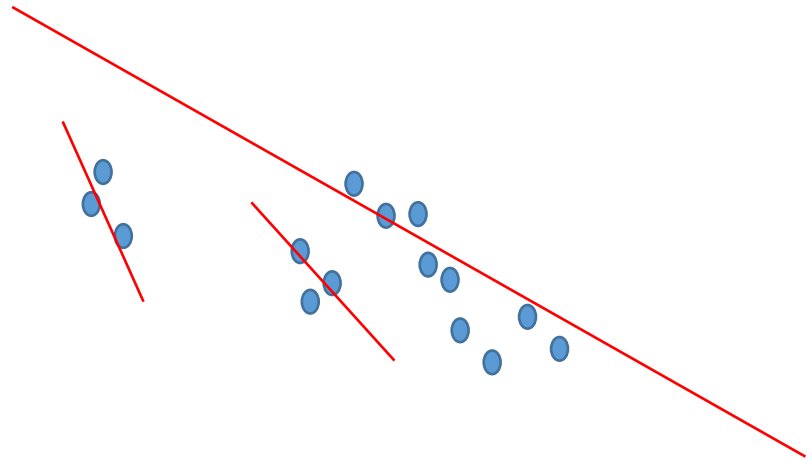
Fitting models to data



Fitting models to data

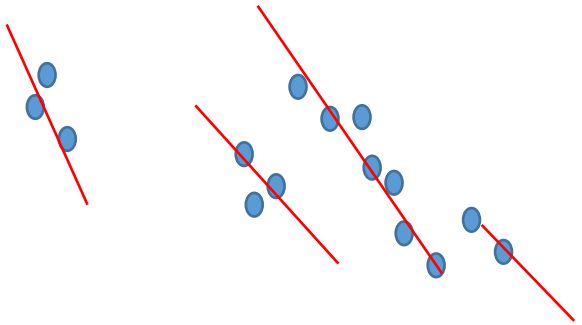


Fitting models to data



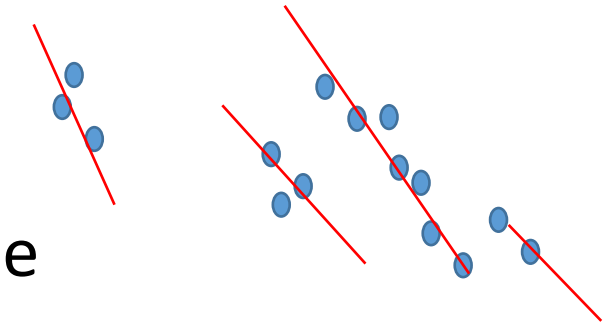
Fitting models to data

- Improving *fit* costs *df*'s
- And thus, statistical power



Fitting models to data

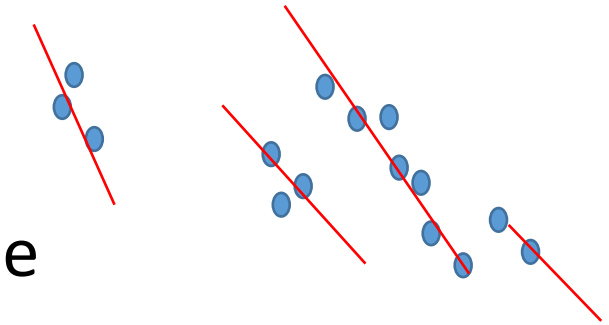
- Improving *fit* costs *df*'s
- And thus, statistical power



- Maximum number of parameters one can fit is number of datapoints

Fitting models to data

- Improving *fit* costs *df*'s
- And thus, statistical power



- Maximum number of parameters one can fit is number of datapoints

- overfitted model

Assessing *fit* of the model

- We need to find a compromise between *df* and how much variance it explains

Guidelines:

- Think before you run your model

Guidelines:

- Think before you run your model
- What is your question?

Guidelines:

- Think before you run your model
- What is your question?
- What is the response variable, what the explanatory variable?

Guidelines:

- Think before you run your model
- What is your question?
- What is the response variable, what the explanatory variable?
- Are there other variables that can affect the relationship you are investigating? If so, add them. Any interactions?

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant

Guidelines:

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant
- Run this model
- Examine it. Look at df's find out whether it's overparametrised.

Guidelines:

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant
- Run this model
- Examine it. Look at df's find out whether it's overparametrised.
- Look at your variables. First, remove interactions that are not significant

Guidelines:

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant
- Run this model
- Examine it. Look at df's find out whether it's overparametrised.
- Look at your variables. First, remove interactions that are not significant
- Look at the reduced model. Now remove main terms *if* not biologically required

Guidelines:

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant
- Run this model
- Examine it. Look at df's find out whether it's overparametrised.
- Look at your variables. First, remove interactions that are not significant
- Look at the reduced model. Now remove main terms *if* not biologically required
- Until you get your final model

Guidelines:

- Build a MAXIMAL model, that includes all variables that you consider biologically relevant, and all interactions that are biologically relevant
- Run this model
- Examine it. Look at df's find out whether it's overparametrised.
- Look at your variables. First, remove interactions that are not significant
- Look at the reduced model. Now remove main terms *if* not biologically required
- Until you get your final model
- Null model = all *b*s are 0.

Other ways

- Decide ahead and keep all in (good with Bayesian methods)

Other ways

- Decide ahead and keep all in (good with Bayesian methods)
- Use model selection information criterion like AIC, BIC, DIC ect. (ML methods)

Other ways

- Decide ahead and keep all in (good with Bayesian methods)
- Use model selection information criterion like AIC, BIC, DIC ect. (ML methods)
- Some people use strict backwards step-wise model selection
- Some people use forwards step-wise models selection (don’t do that)

Other ways

- Decide ahead and keep all in (good with Bayesian methods)
- Use model selection information criterion like AIC, BIC, DIC ect. (ML methods)
- Some people use strict backwards step-wise model selection
- Some people use forwards step-wise models selection (don’t do that)
- Using ANOVA

Other ways

- Decide ahead and keep all in (good with Bayesian methods)
- Use model selection information criterion like AIC, BIC, DIC ect. (ML methods)
- Some people use strict backwards step-wise model selection
- Some people use forwards step-wise models selection (don’t do that)
- Using ANOVA

- Important: keep the biology, and the parameter estimate in mind!

Other ways

- Decide ahead and keep all in (good with Bayesian methods)
- Use model selection information criterion like AIC, BIC, DIC ect. (ML methods)
- Some people use strict backwards step-wise model selection
- Some people use forwards step-wise models selection (don’t do that)
- Using ANOVA

- Important: keep the biology, and the parameter estimate in mind!
- **Keep parameters in if it makes biologically sense!**

Learning aim

- How to choose the *best* model

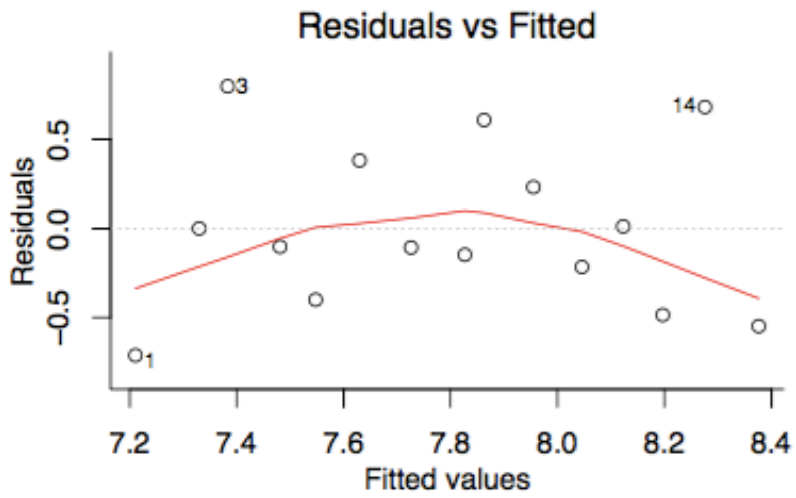
Model assumptions

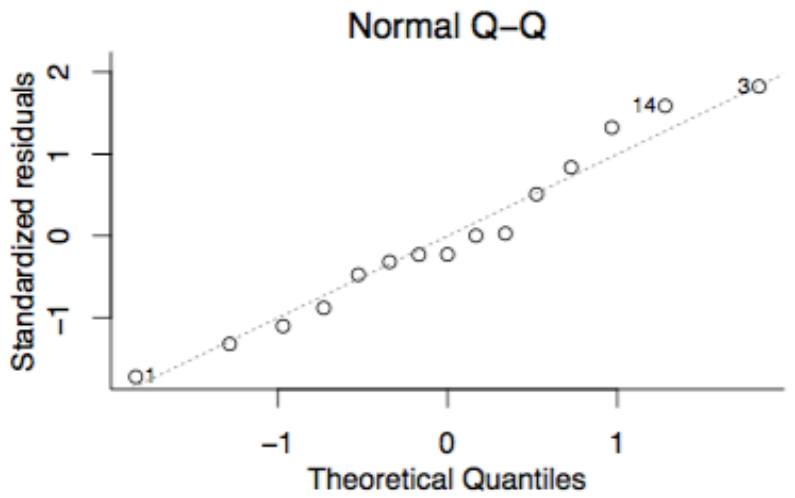
- Residuals are normally distributed

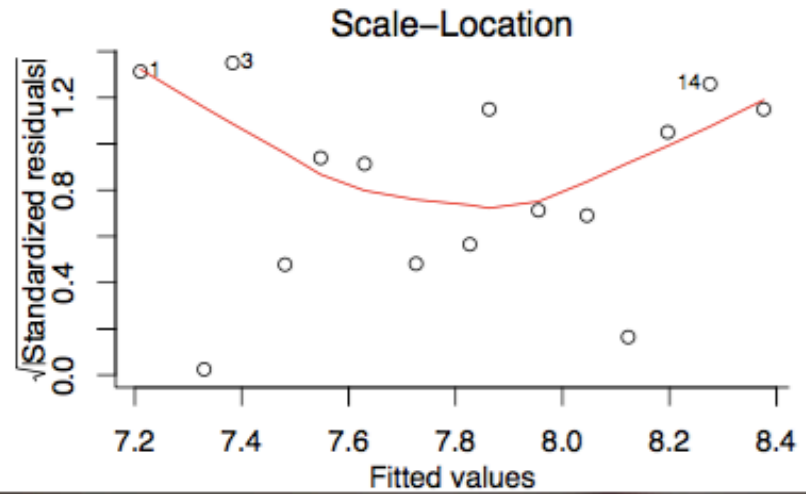
Check validity of that with `plot(model)`

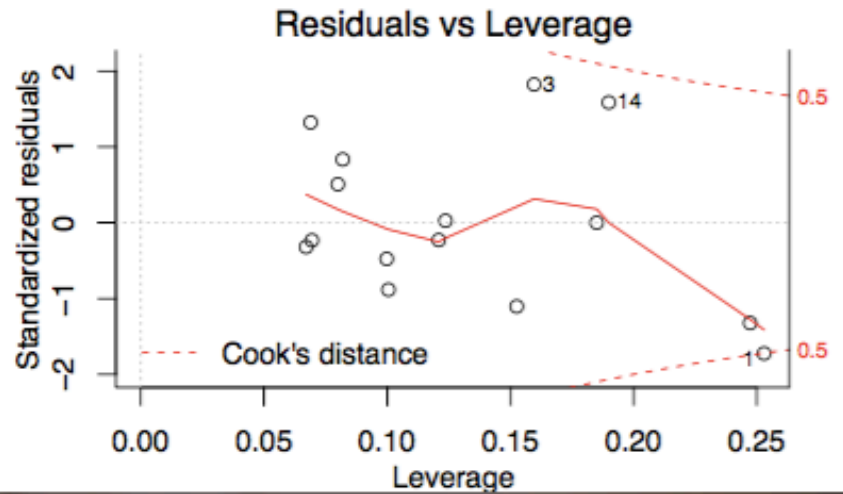
Diagnostic plots

```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```



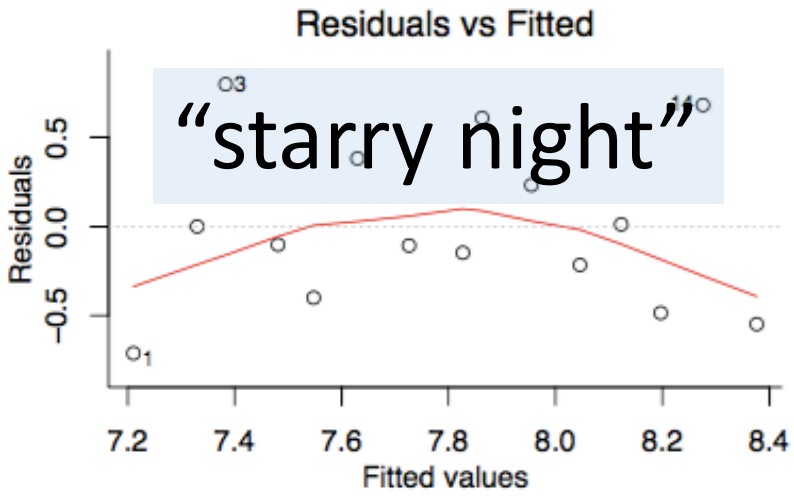




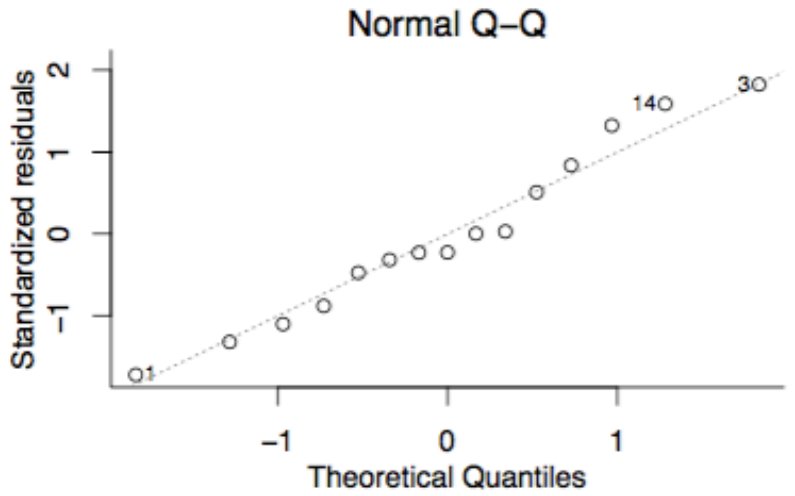


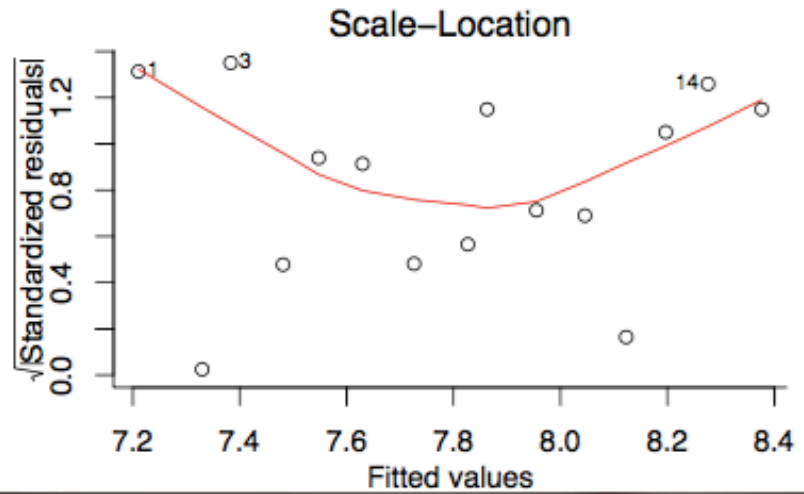
Diagnostic plots

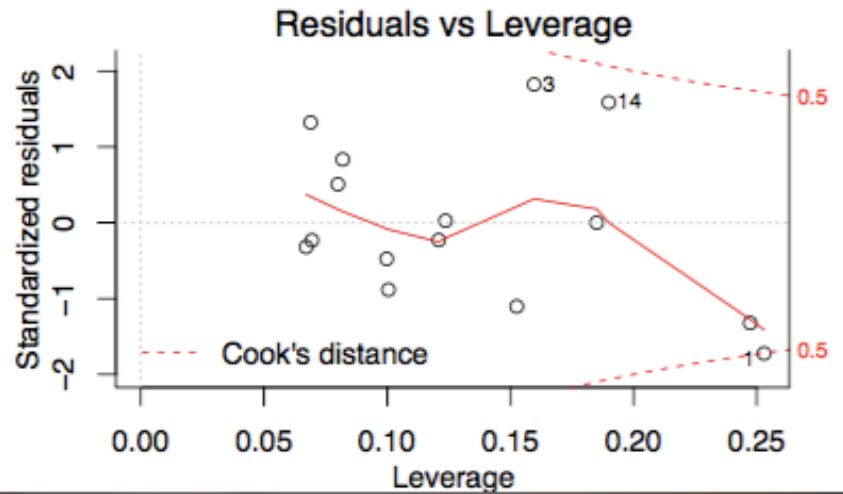
```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```



"starry night"

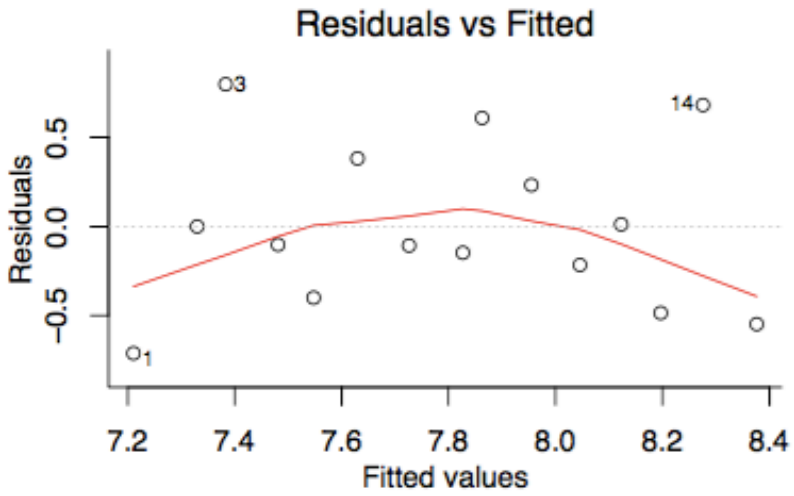


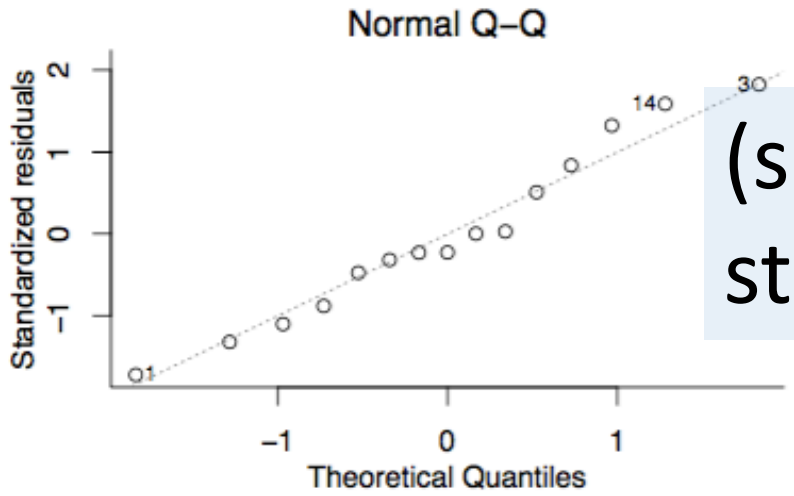




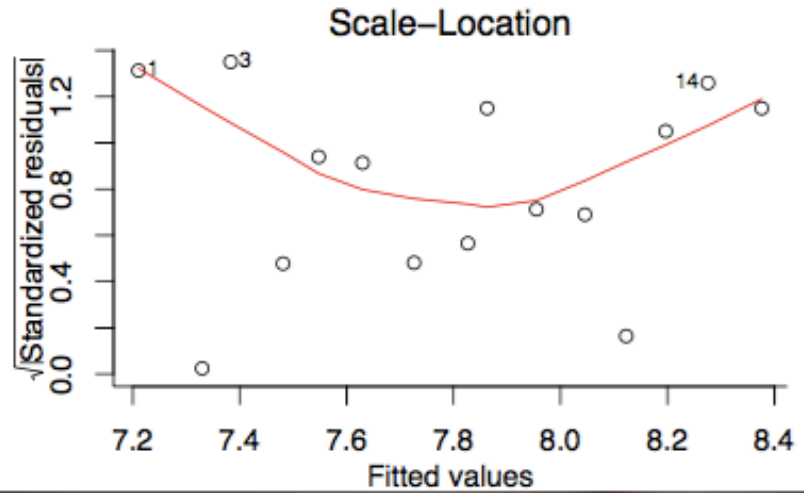
Diagnostic plots

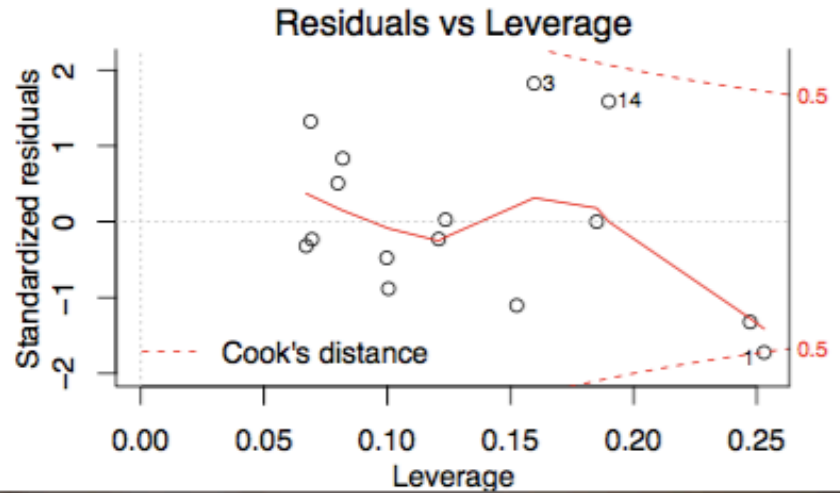
```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```





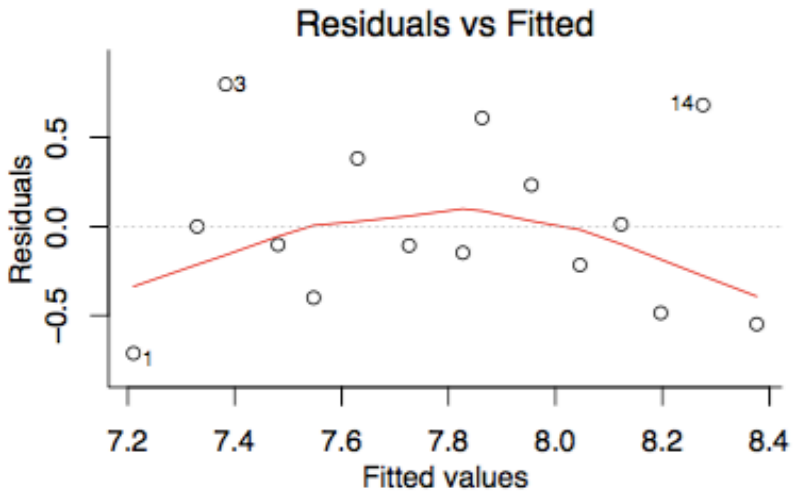
(somewhat) straight line

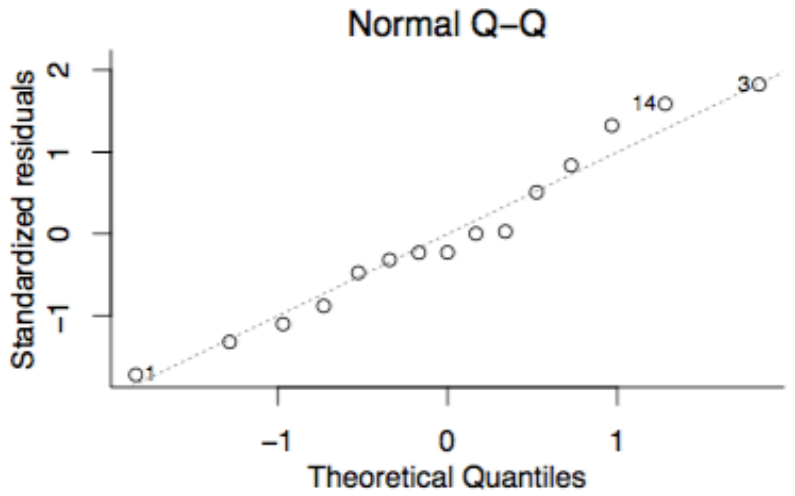


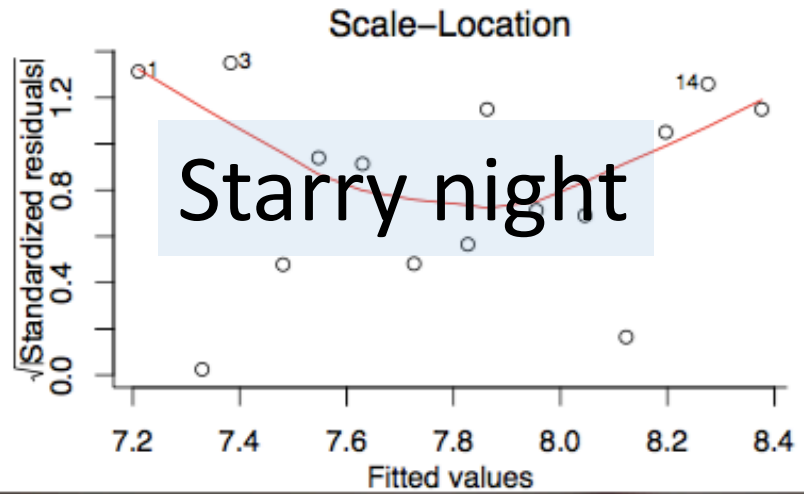


Diagnostic plots

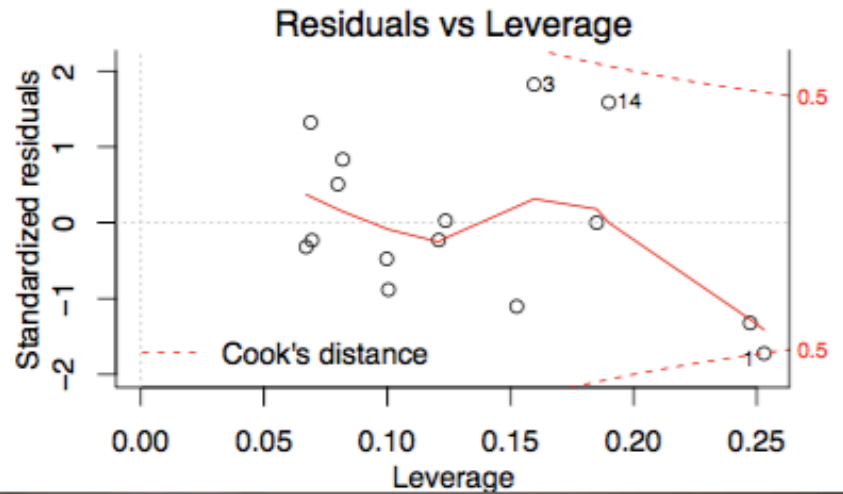
```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```





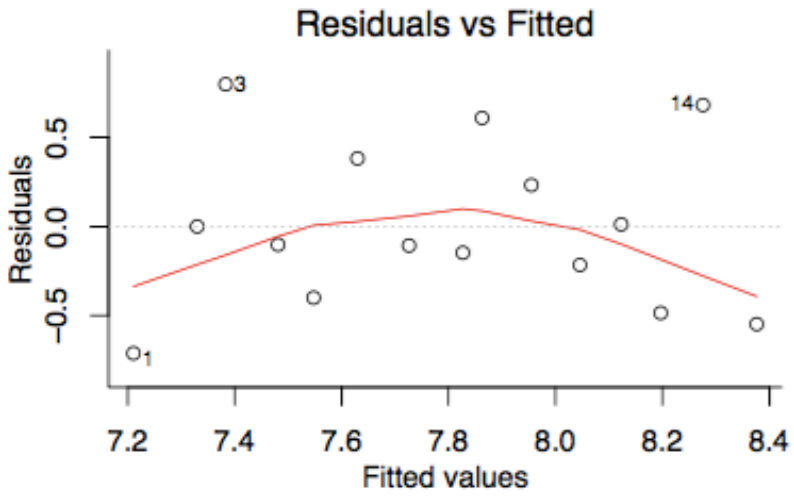


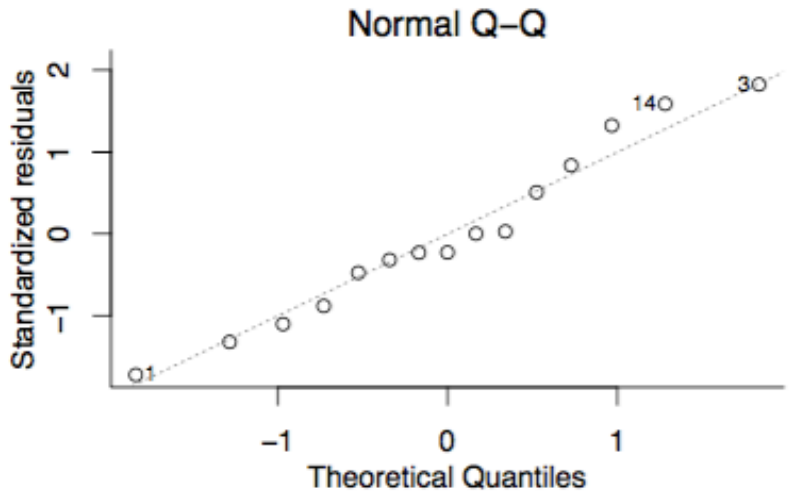
Starry night

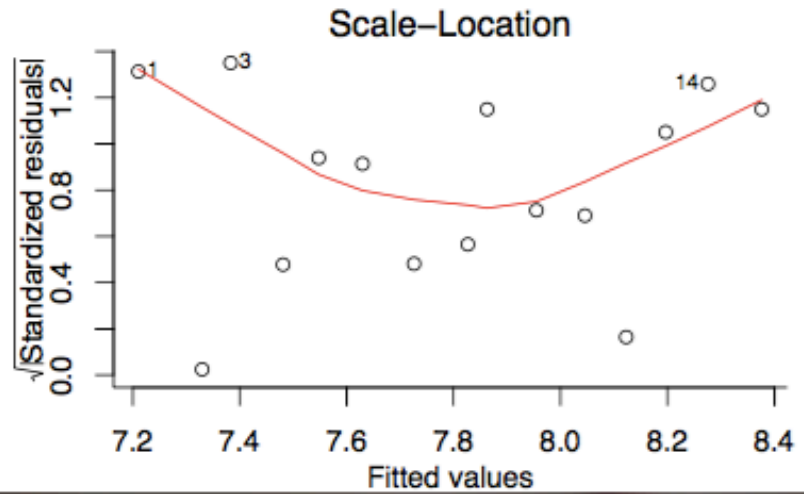


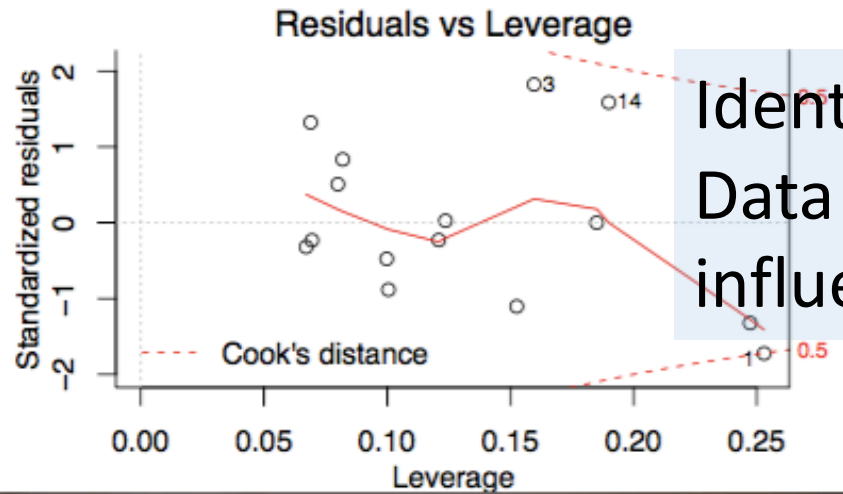
Diagnostic plots

```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```



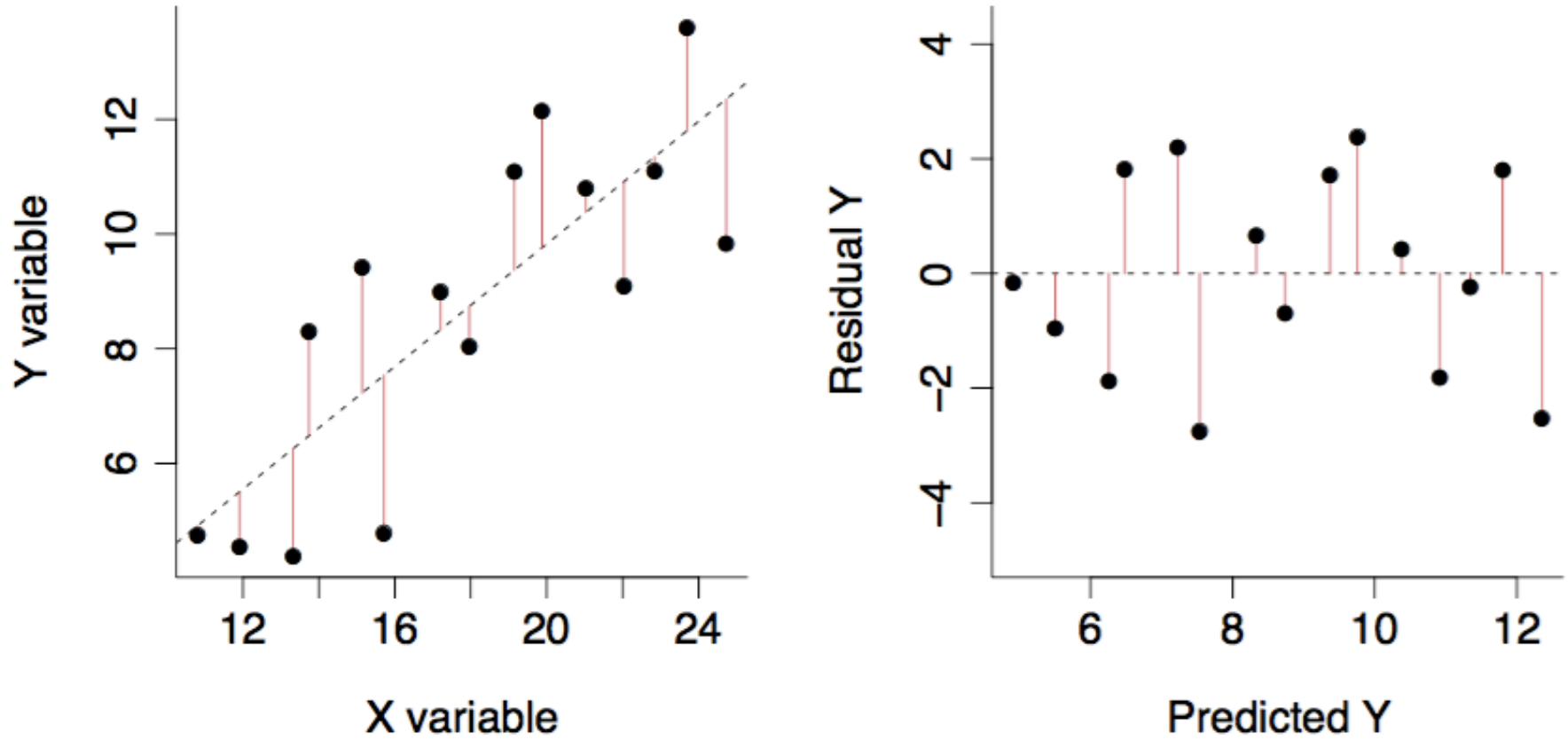




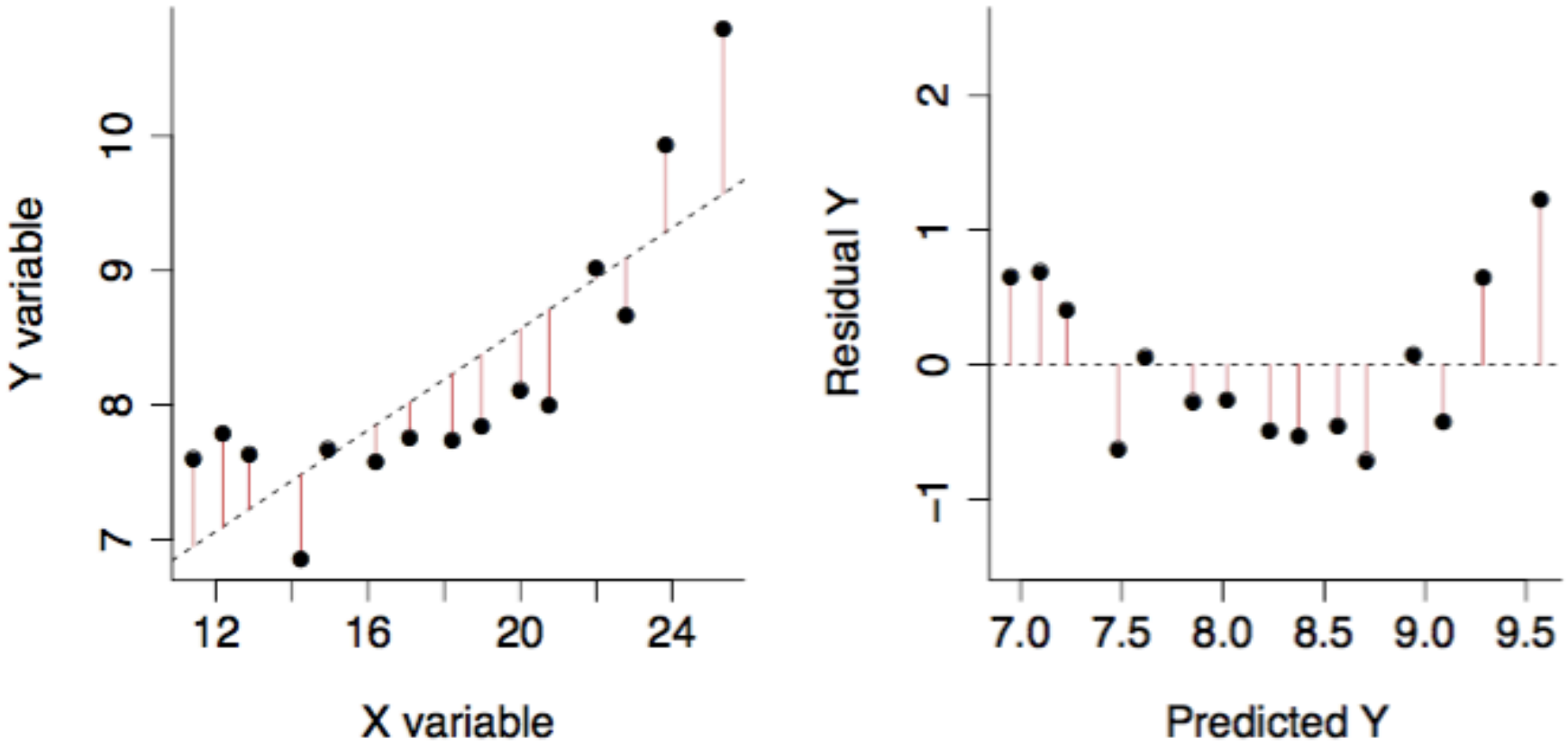


Identify outliers
Data with lots of influence

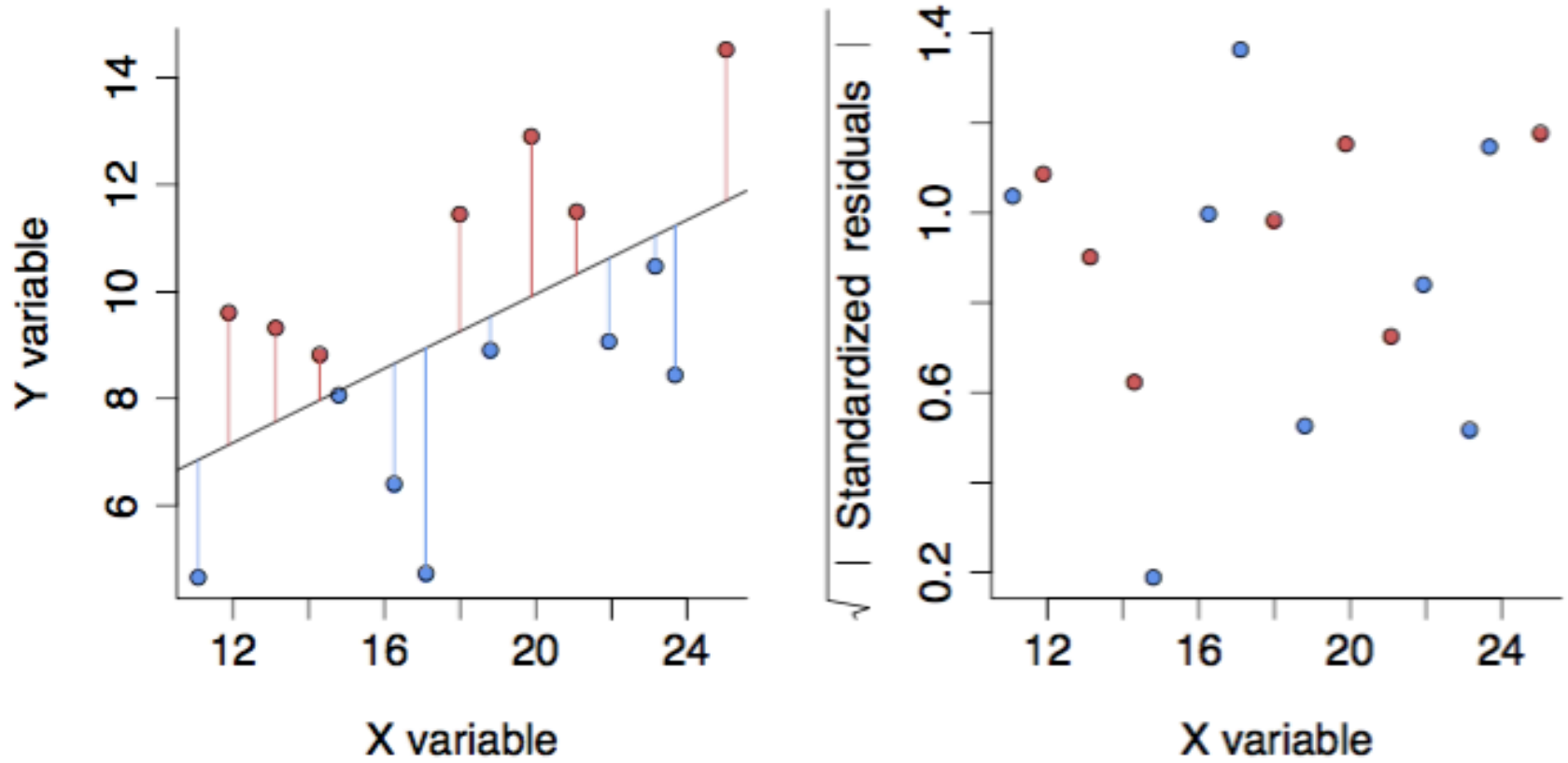
Residuals v fitted



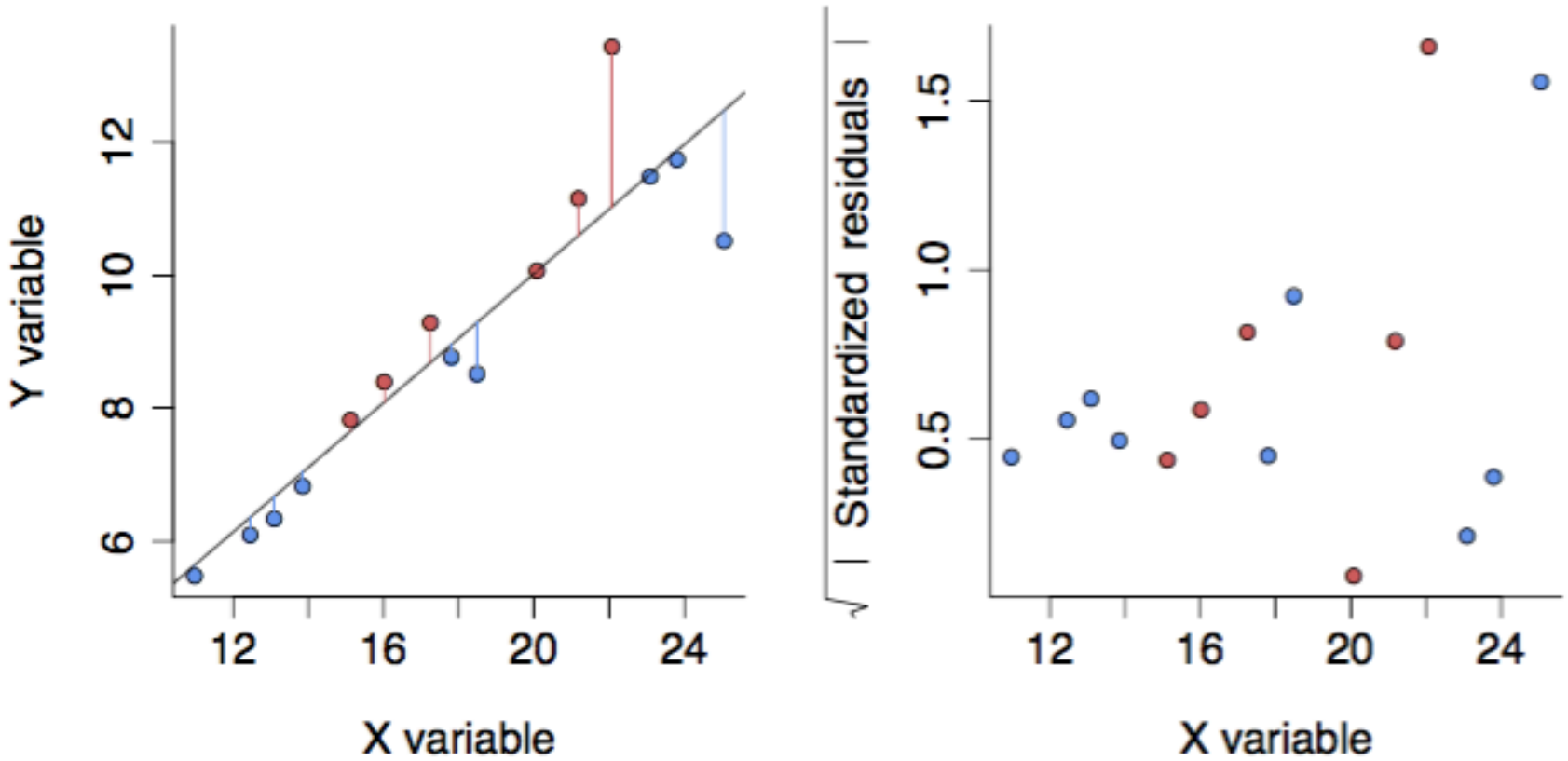
Residuals v fitted



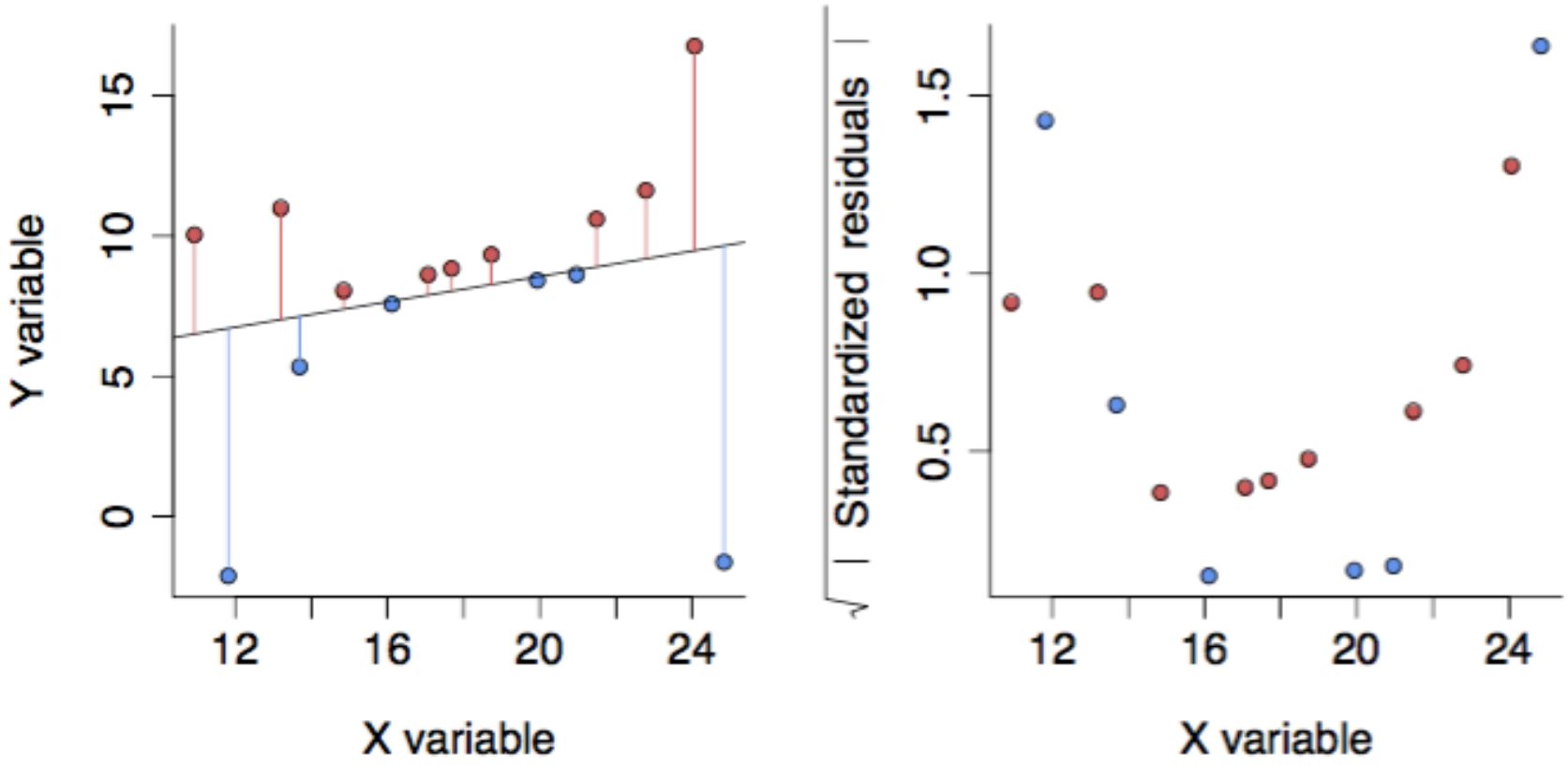
Scale Location plots



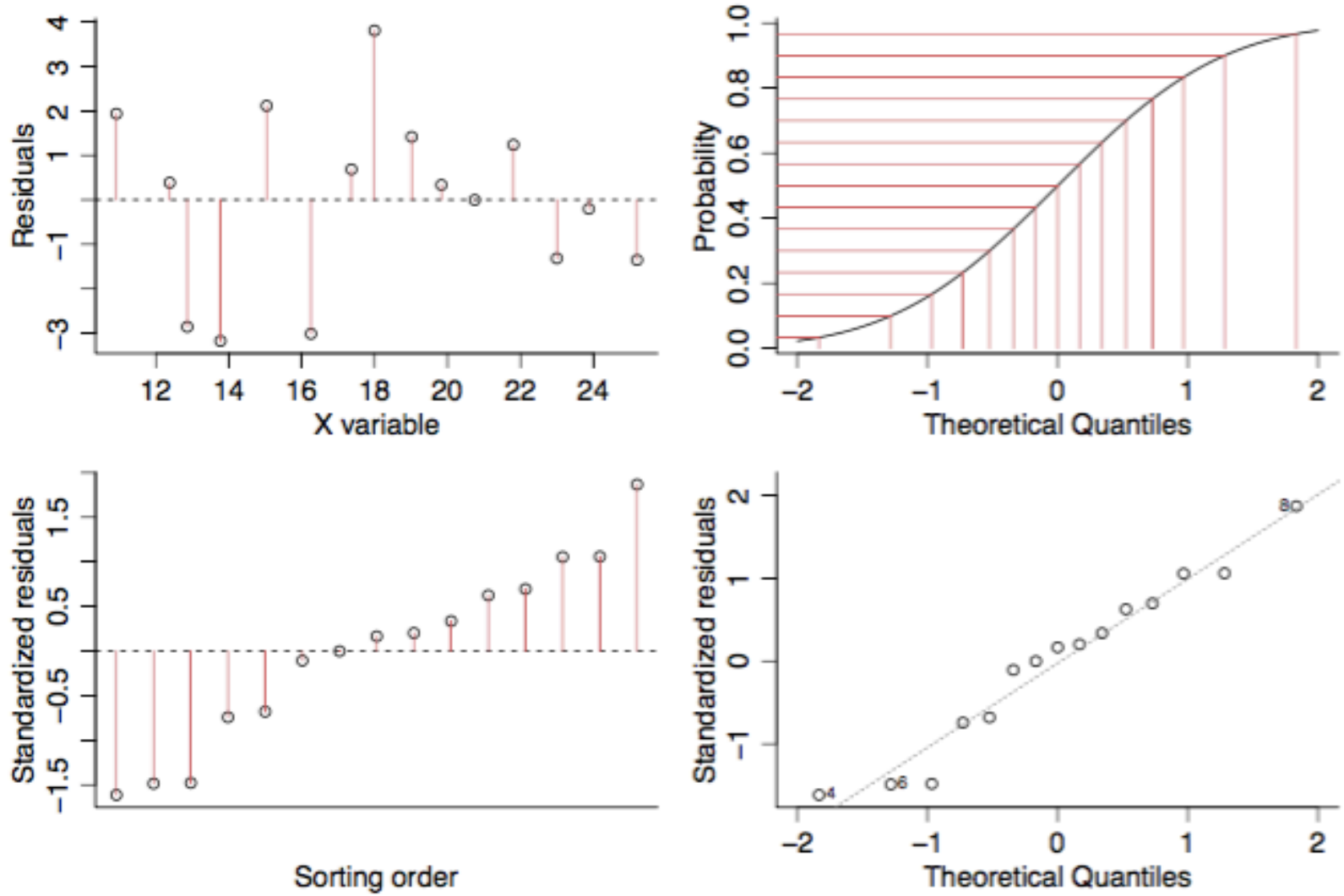
Scale Location plots



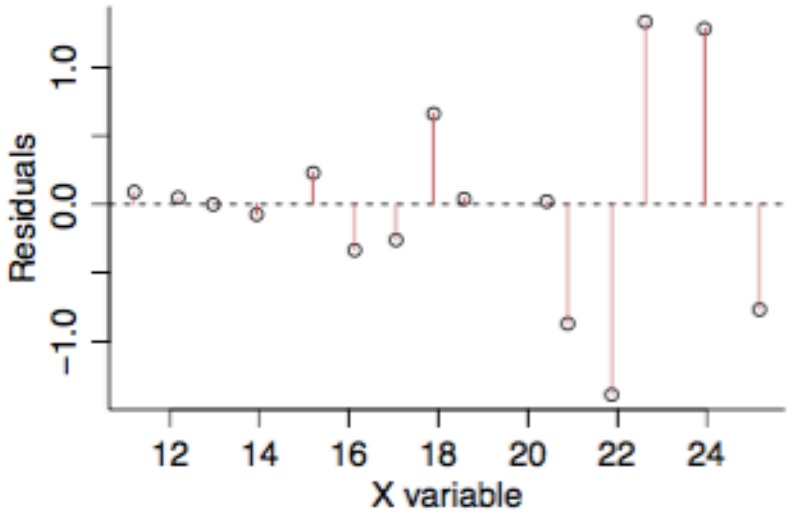
Scale Location plots

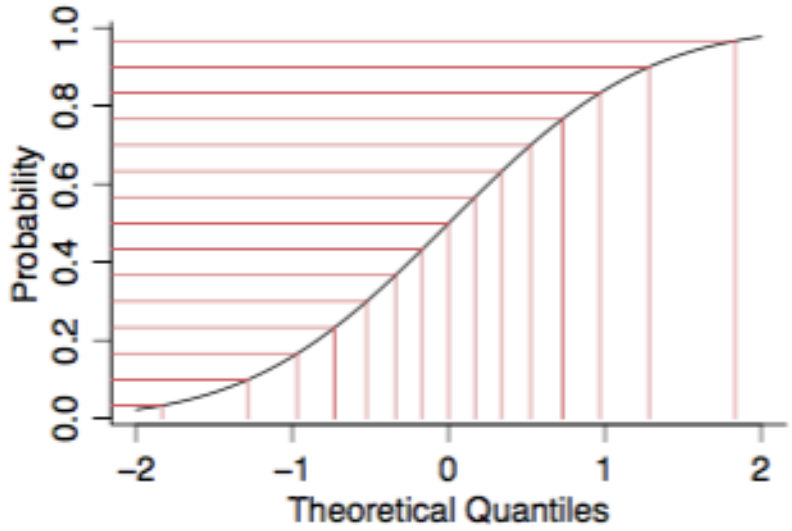


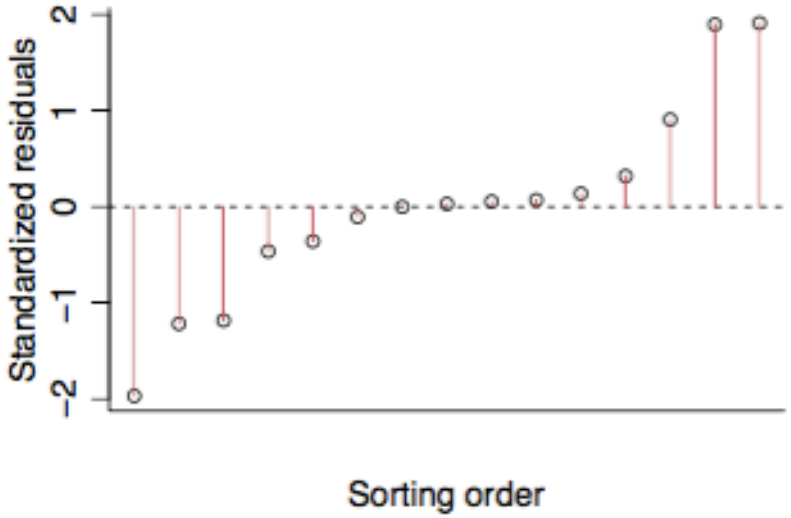
Normal Q-Q plots

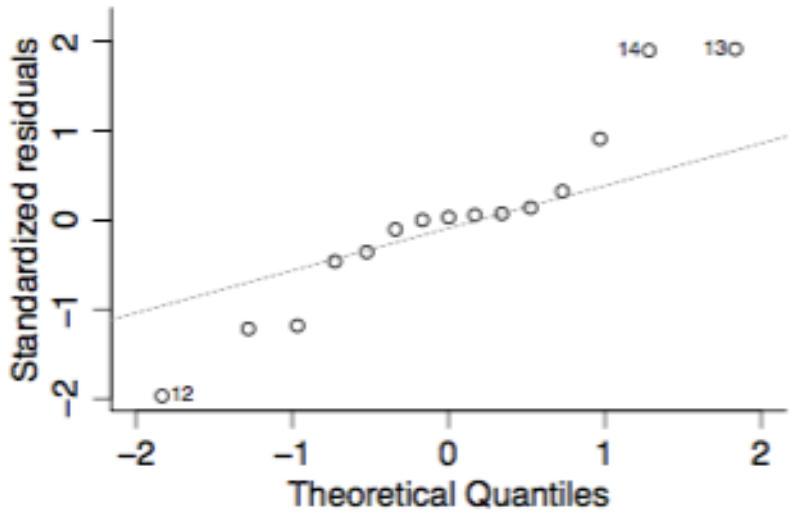


Normal Q-Q plots

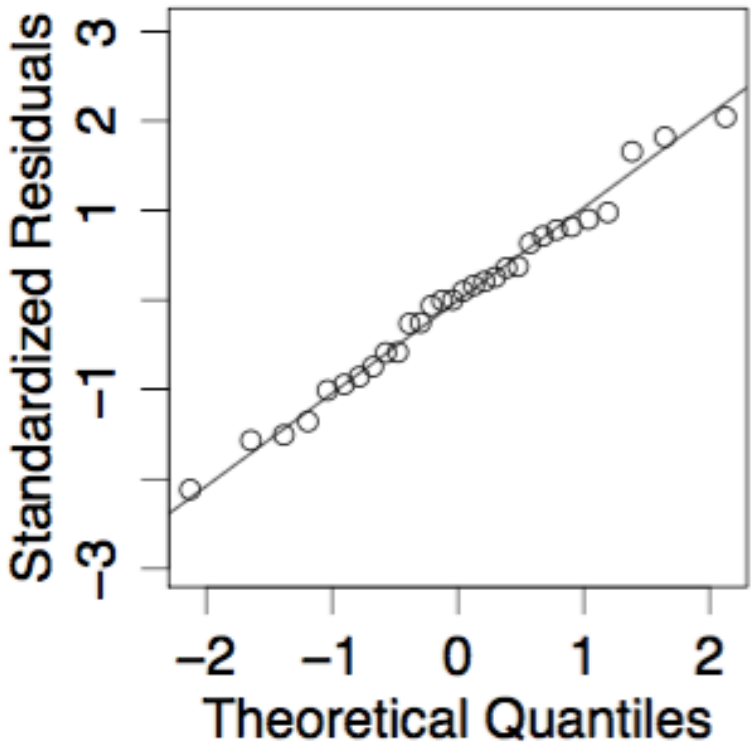


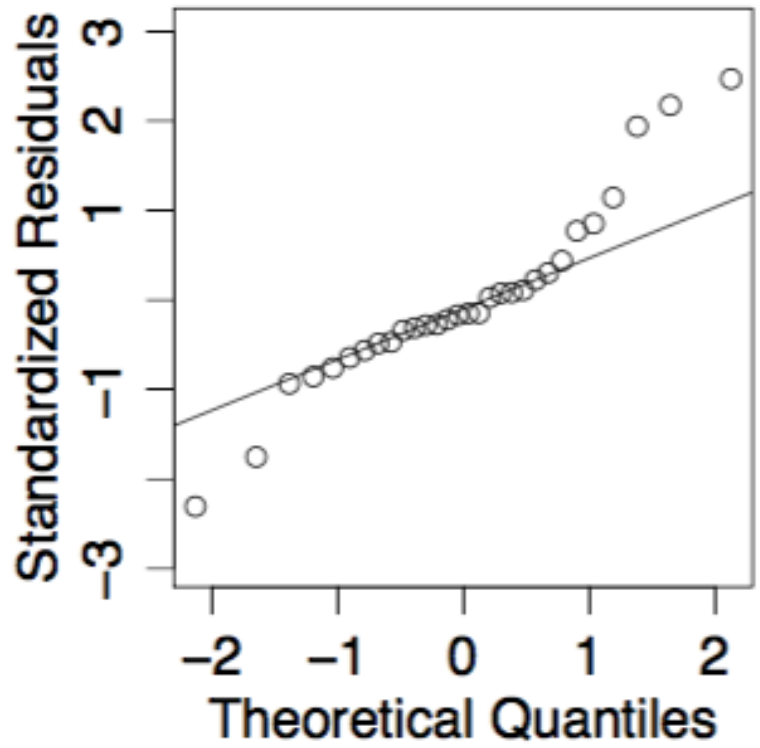






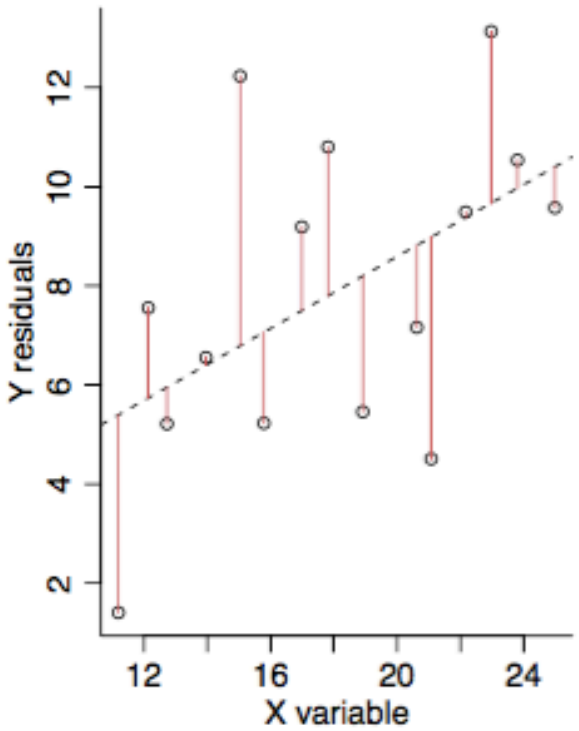
Normal Q-Q plots

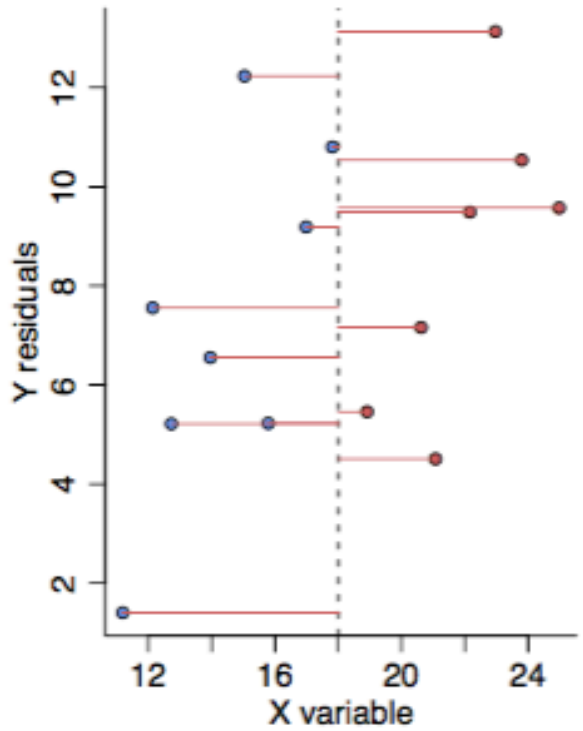


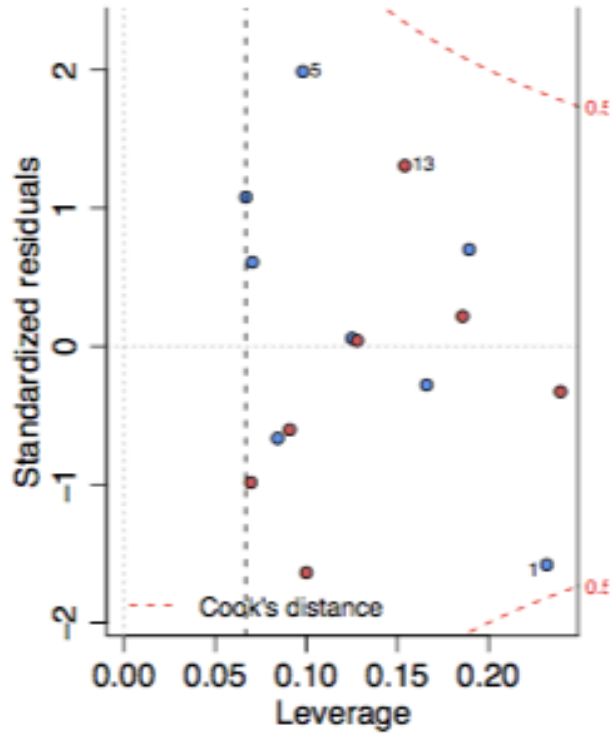


Residuals v Leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

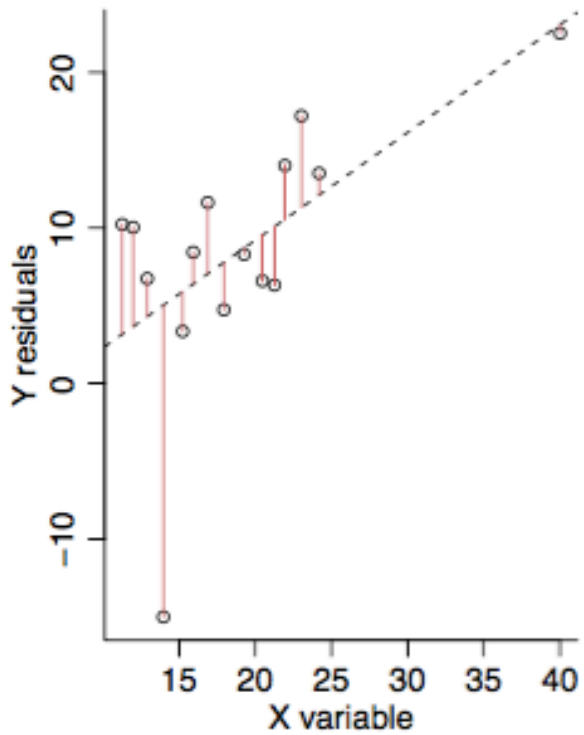


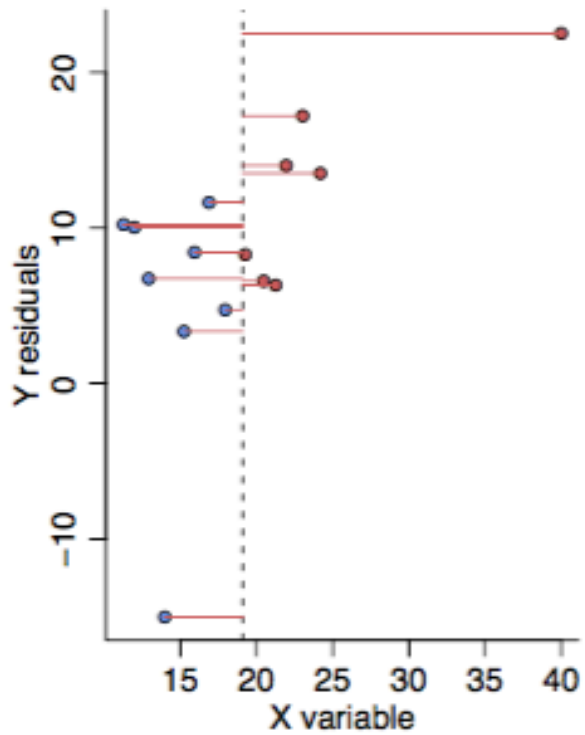


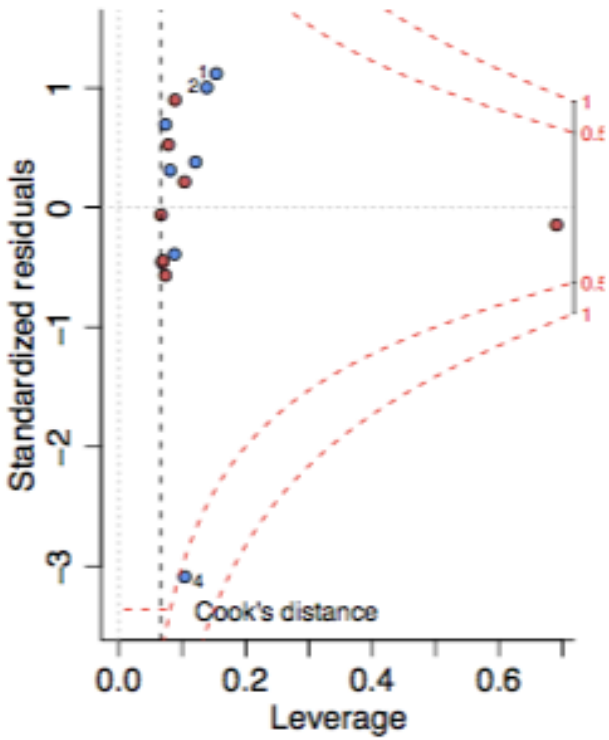


Residuals v Leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$







Outliers

- Are they wrongly measured?
- Are they biologically meaningful?

What to do if these plots are crap?

- First consider if your response is really a continuous variable

What to do if these plots are crap?

- First consider if your response is really a continuous variable
- If not → non-parametric tests, or GLMs

What to do if these plots are crap?

- First consider if your response is really a continuous variable
- If not → non-parametric tests, or GLMs
- Consider your units, check for typos and outliers

What to do if these plots are crap?

- First consider if your response is really a continuous variable
- If not → non-parametric tests, or GLMs
- Consider your units, check for typos and outliers
- Are the violations really bad?
- Use subsets to see how strong it affects your conclusions

What to do if these plots are crap?

- First consider if your response is really a continuous variable
- If not → non-parametric tests, or GLMs
- Consider your units, check for typos and outliers
- Are the violations really bad?
- Use subsets to see how strong it affects your conclusions

- Data transformation (we don’t like that a lot)

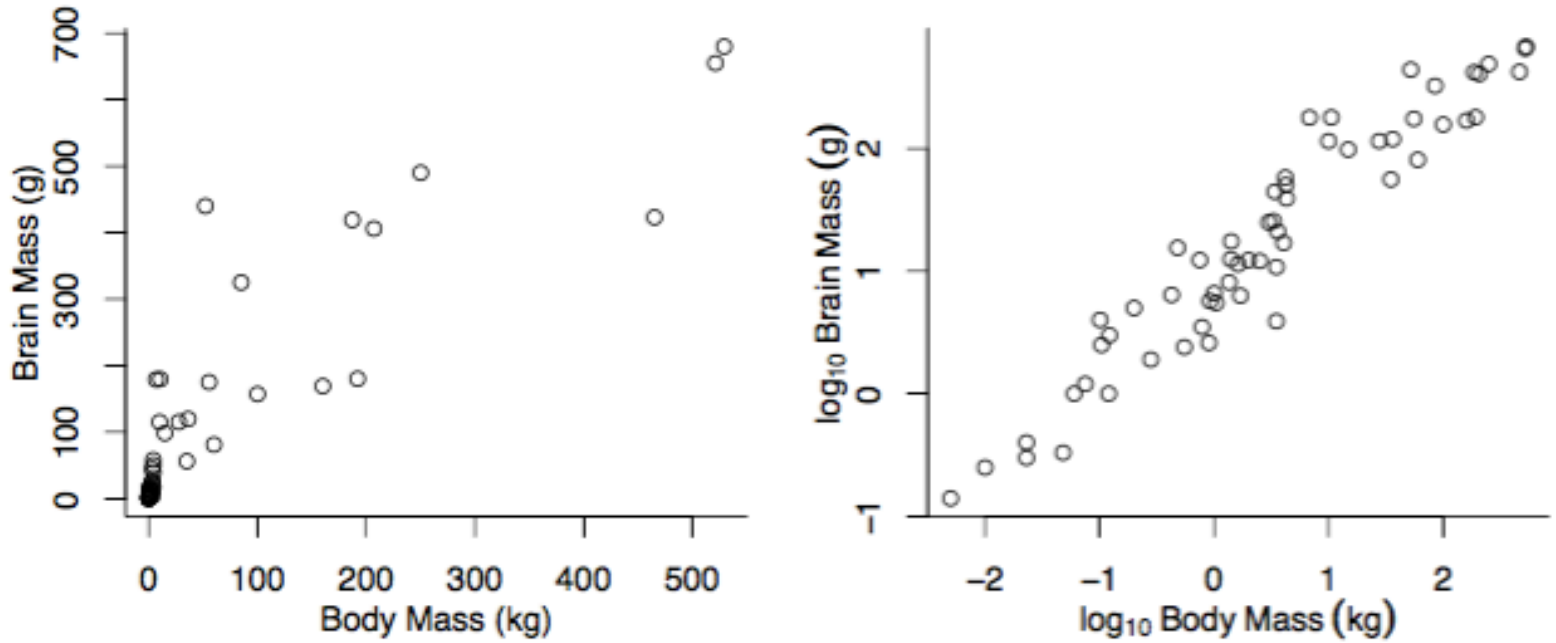
Transformation

Can we transform the data?

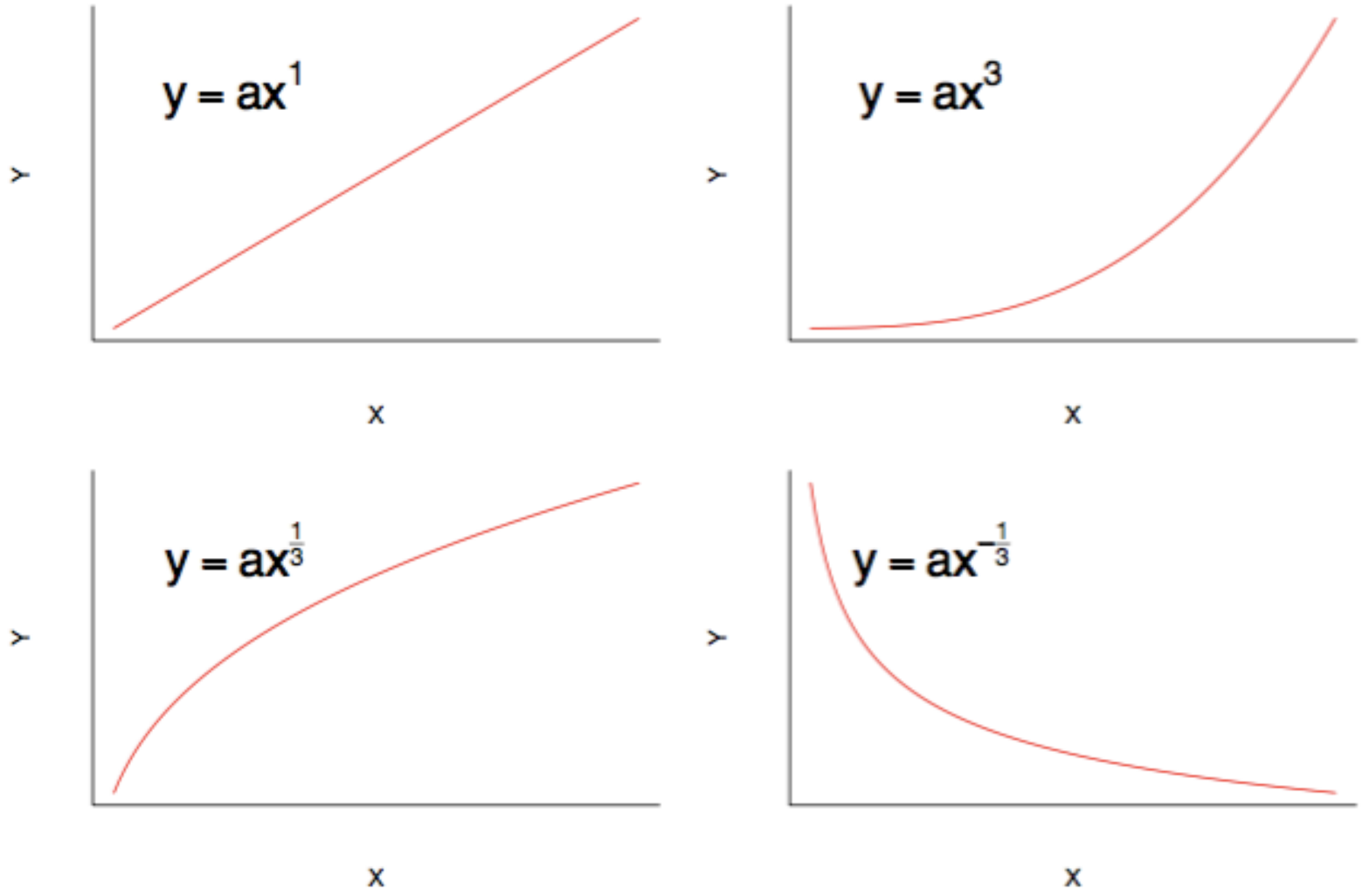
- Need to find a mathematical function that improves the model.
- The type of transformation may say something about the model.
- What process turns the data into a better linear model?

Power law

- $y = ax^b$
- $\log(y) = \log(a) + \log(x)\ b$

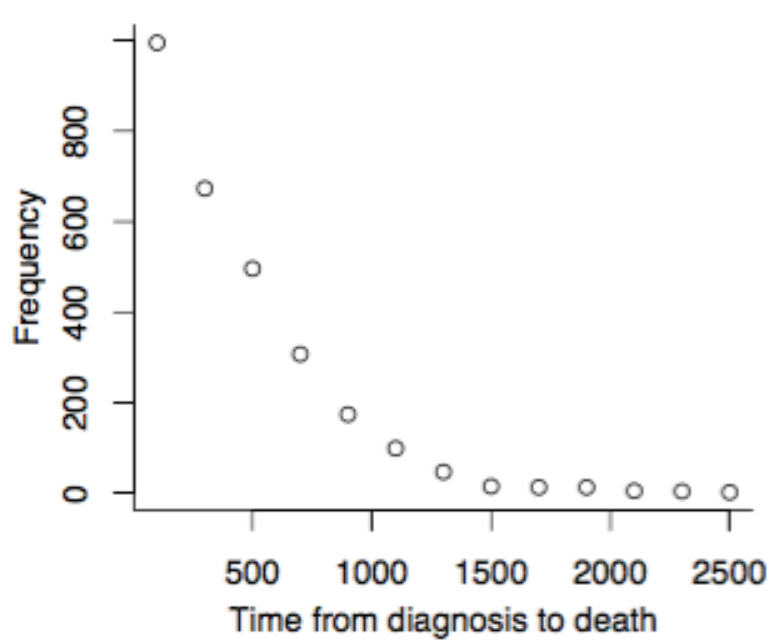


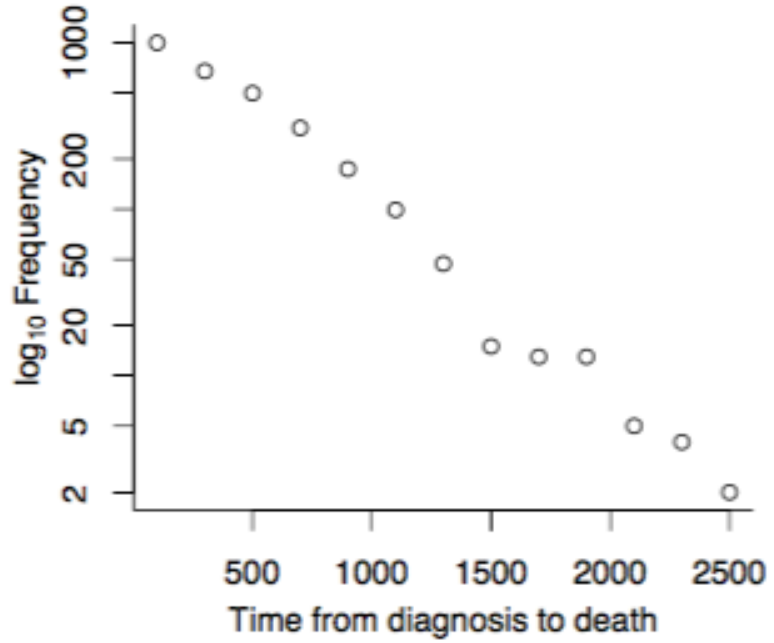
Power law



Exponential data

- $y = ae^{bx}$
- $\log(y) = \log(a) + bx$

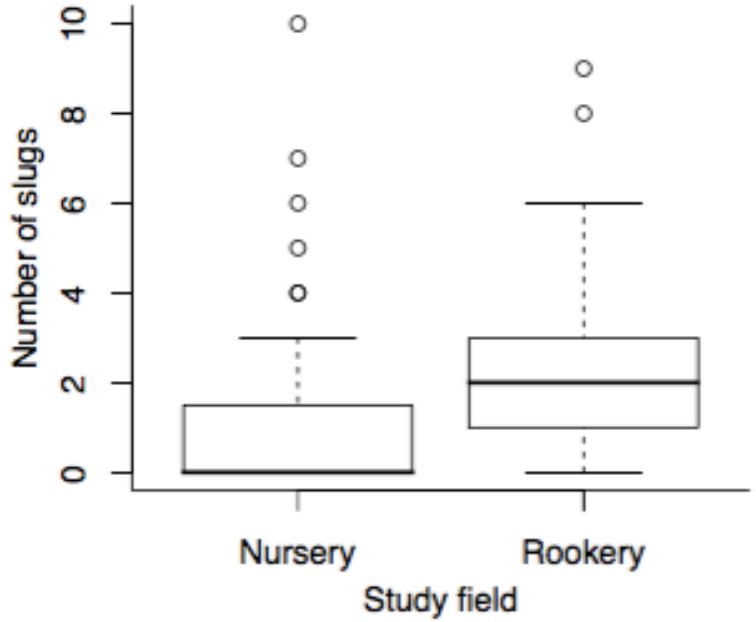


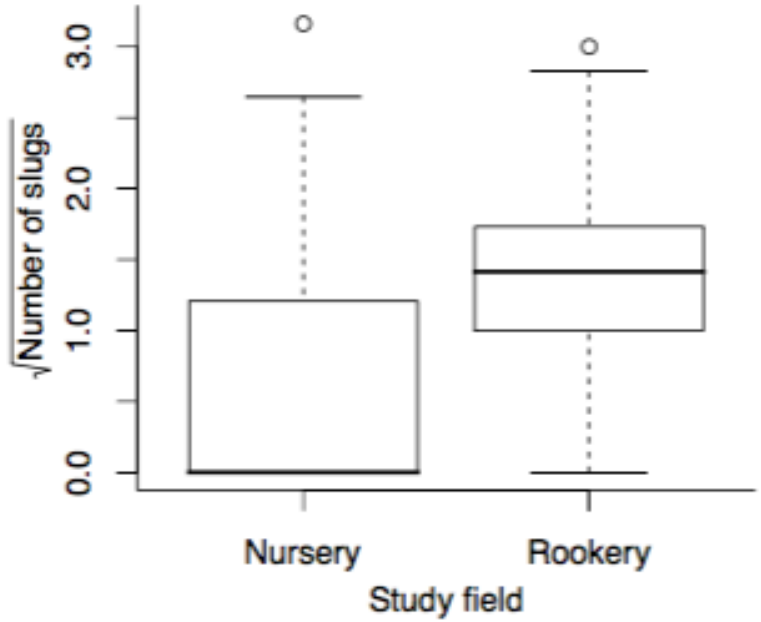


Square root transformation

Useful for simple analysis of count data

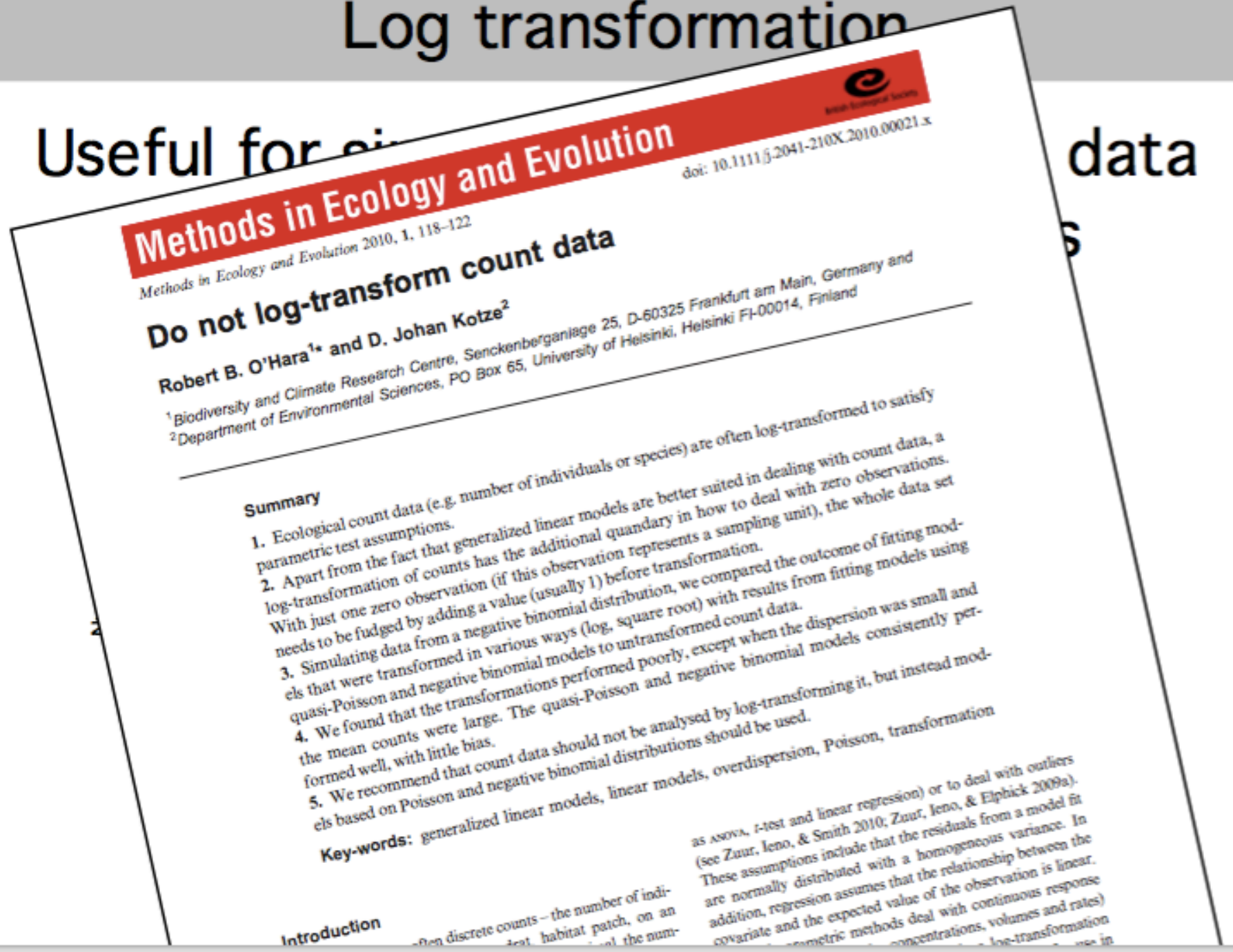
- many low and a few high counts





Log transformation

Useful for ... data



Methods in Ecology and Evolution

Methods in Ecology and Evolution 2010, 1, 118–122

doi: 10.1111/j.2041-210X.2010.00021.x

Do not log-transform count data

Robert B. O'Hara[1]* and D. Johan Kotze[2]

[1]Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany and [2]Department of Environmental Sciences, PO Box 65, University of Helsinki, Helsinki FI-00014, Finland

Summary

1. Ecological count data (e.g. number of individuals or species) are often log-transformed to satisfy parametric test assumptions.
2. Apart from the fact that generalized linear models are better suited in dealing with count data, a log-transformation of counts has the additional quandary in how to deal with zero observations. With just one zero observation (if this observation represents a sampling unit), the whole data set needs to be fudged by adding a value (usually 1) before transformation.
3. Simulating data from a negative binomial distribution, we compared the outcome of fitting models that were transformed in various ways (log, square root) with results from fitting models using quasi-Poisson and negative binomial models to untransformed count data.
4. We found that the transformations performed poorly, except when the dispersion was small and the mean counts were large. The quasi-Poisson and negative binomial models consistently performed well, with little bias.
5. We recommend that count data should not be analysed by log-transforming it, but instead models based on Poisson and negative binomial distributions should be used.

Key-words: generalized linear models, linear models, overdispersion, Poisson, transformation

Introduction

often discrete counts – the number of individuals, habitat patch, on an ... the num-

as ANOVA, *t*-test and linear regression) or to deal with outliers (see Zuur, Ieno, & Smith 2010; Zuur, Ieno, & Elphick 2009a). These assumptions include that the residuals from a model fit are normally distributed with a homogeneous variance. In addition, regression assumes that the relationship between the covariate and the expected value of the observation is linear. ... parametric methods deal with continuous response ... concentrations, volumes and rates) ... log-transformation ... use in

Hypothesis testing

- You can reject H_0, or accept H_0, but the latter does not mean the alternative hypothesis is not true!

Multiple hypothesis testing

- We test for something and accept an error of 5% (p=0.05).
- In 100 tests, 20 are wrong
- Can correct for this if wanted (Bonferroni – look it up!)

- Philosophy, and underlying *a priory* assumptions and knowledge

- Don’t use Bonferroni blindly. Always think about what you’re doing, and why, and how it relates to the data!

That’s after the model. But what’s with before?

- Step-by-step guide of how to run a model

That’s after the model. But what’s with before?

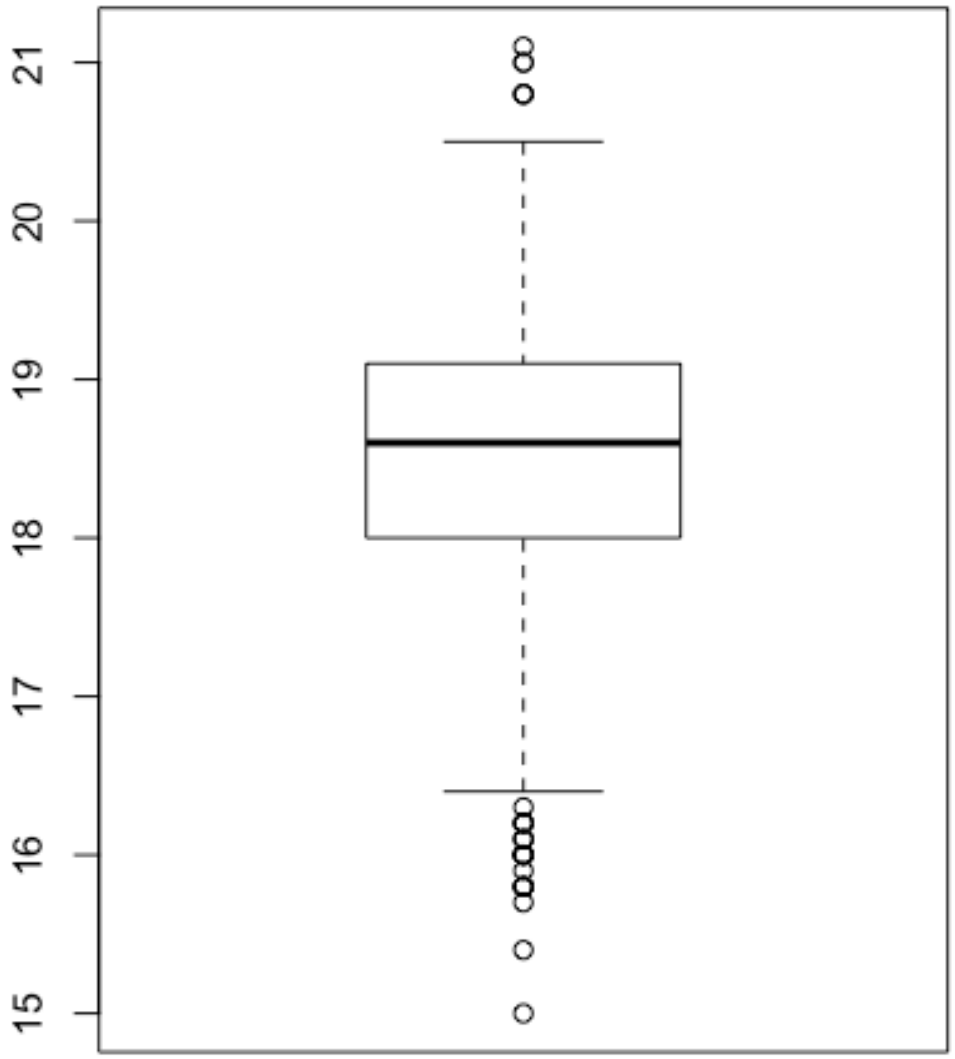
- Step-by-step guide of how to run a model
- First: visual inspection to see what you work with, to get accustomed with the data, to get a “feel” for whether assumptions are violated ect.

1 – are there outliers?

- Use boxplots

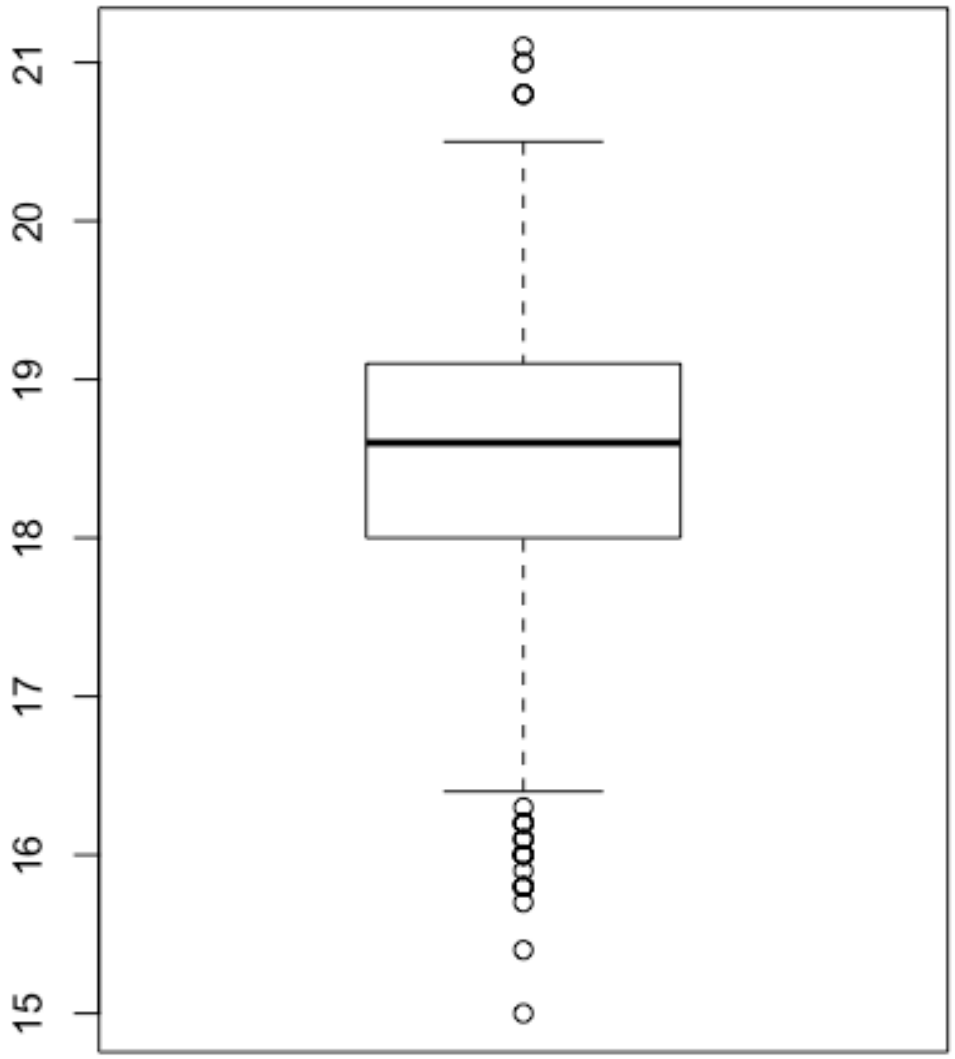
1 – are there outliers?

- Use boxplots



1 – are there outliers?

- Use boxplots
- Use biology arguments to exclude outliers

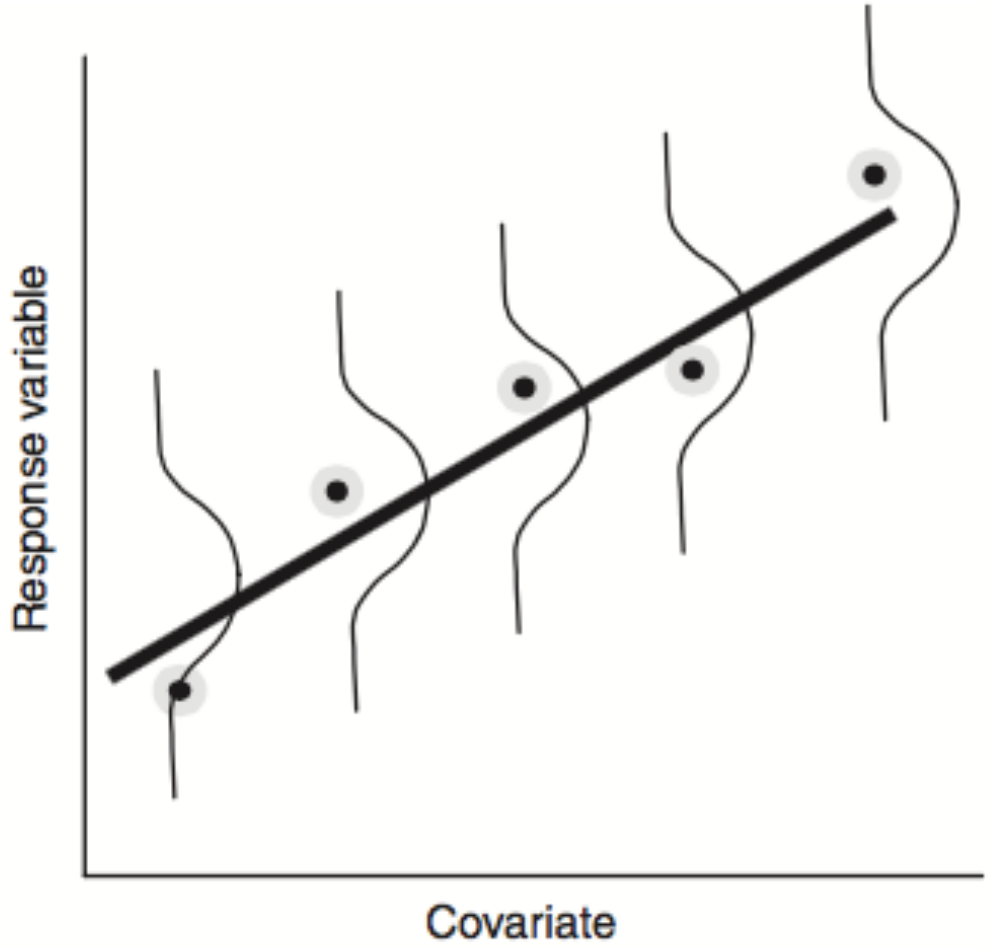


2 – Homogeneity of variances

- Per x_i for categorical x
- Per category for factors
- Residuals

2 – Homogeneity of variances

- Per x_i for categorical x
- Per category for factors
- Residuals

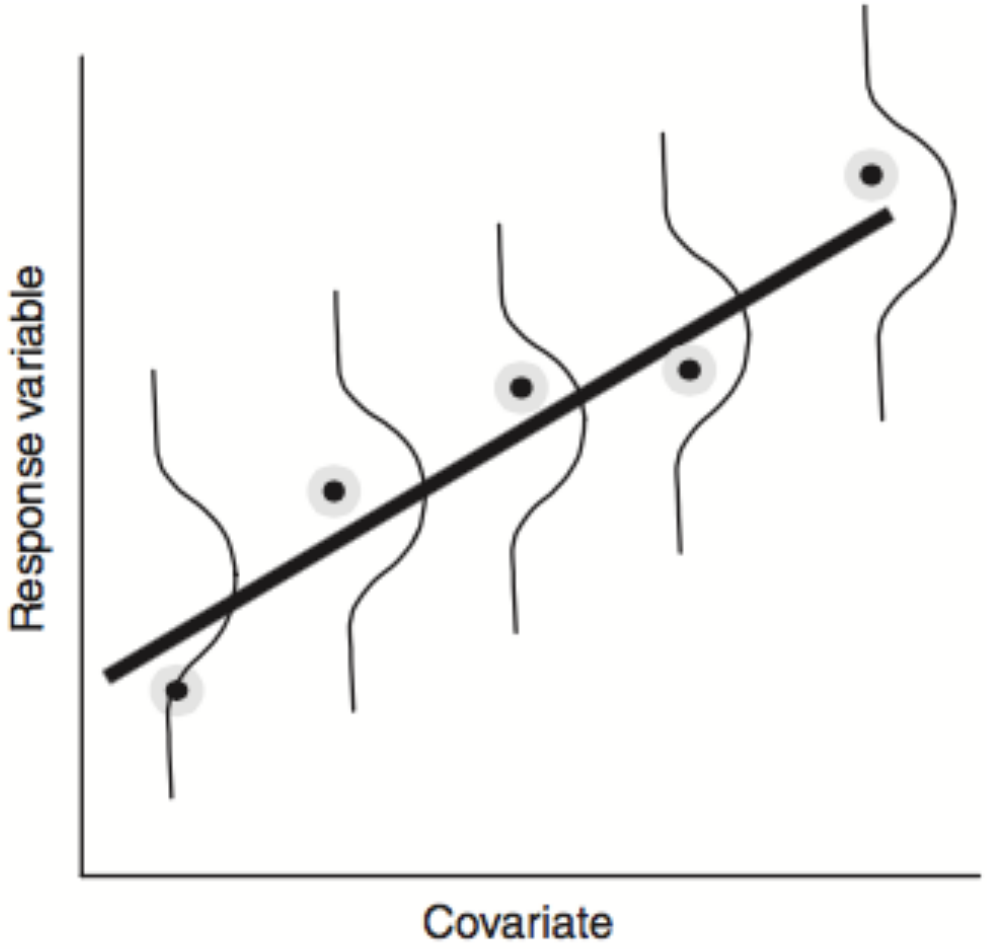


3 – Normally distributed data

- Use histograms
- LMs robust against some violation
- Consider transformations as a last resort

3 – Normally distributed data

- Use histograms
- LMs robust against some violation
- Consider transformations as a last resort
- Again, it’s most important in the residuals

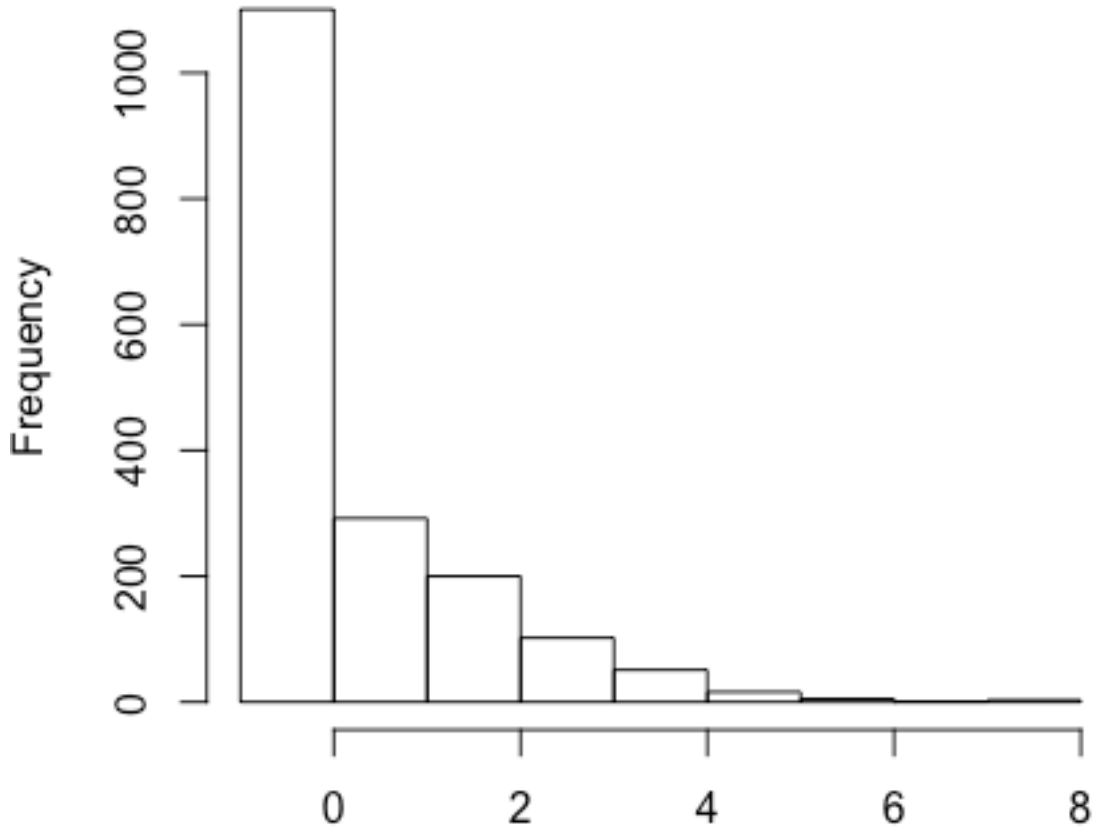


4 – Is your data zero-inflated?

- Use histograms to check for this

4 – Is your data zero-inflated?

- Use histograms to check for this
- More in GLM course

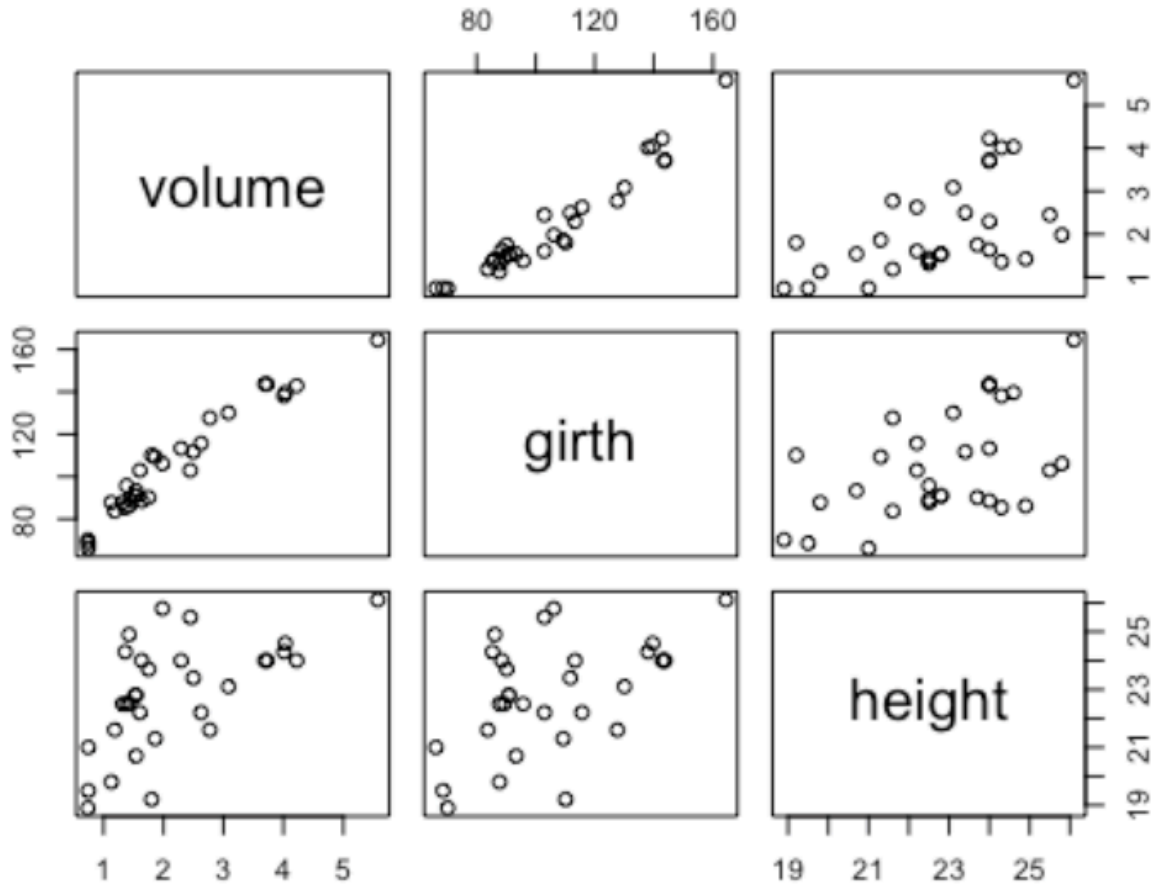


5 – Is there collinearity among the data?

- Collinearity = correlation between covariates

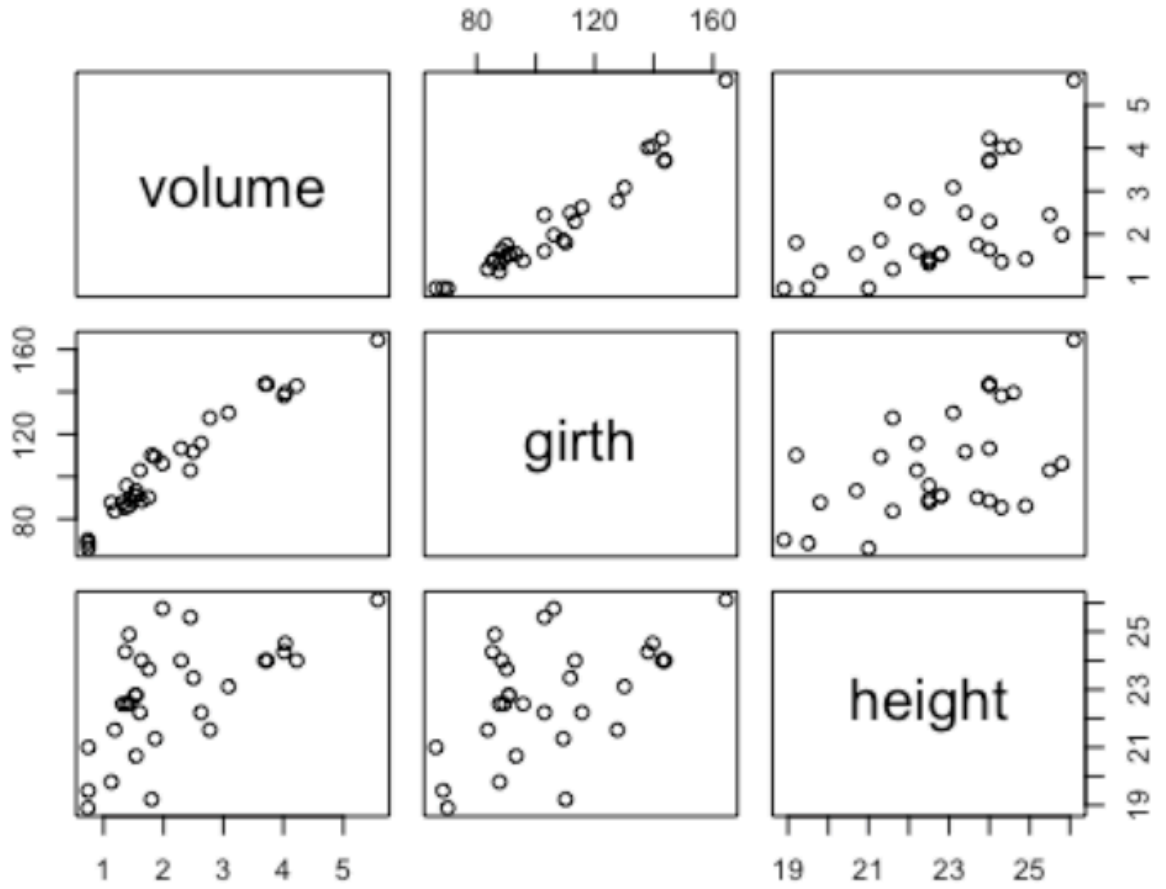
5 – Is there collinearity among the data?

- Collinearity = correlation between covariates



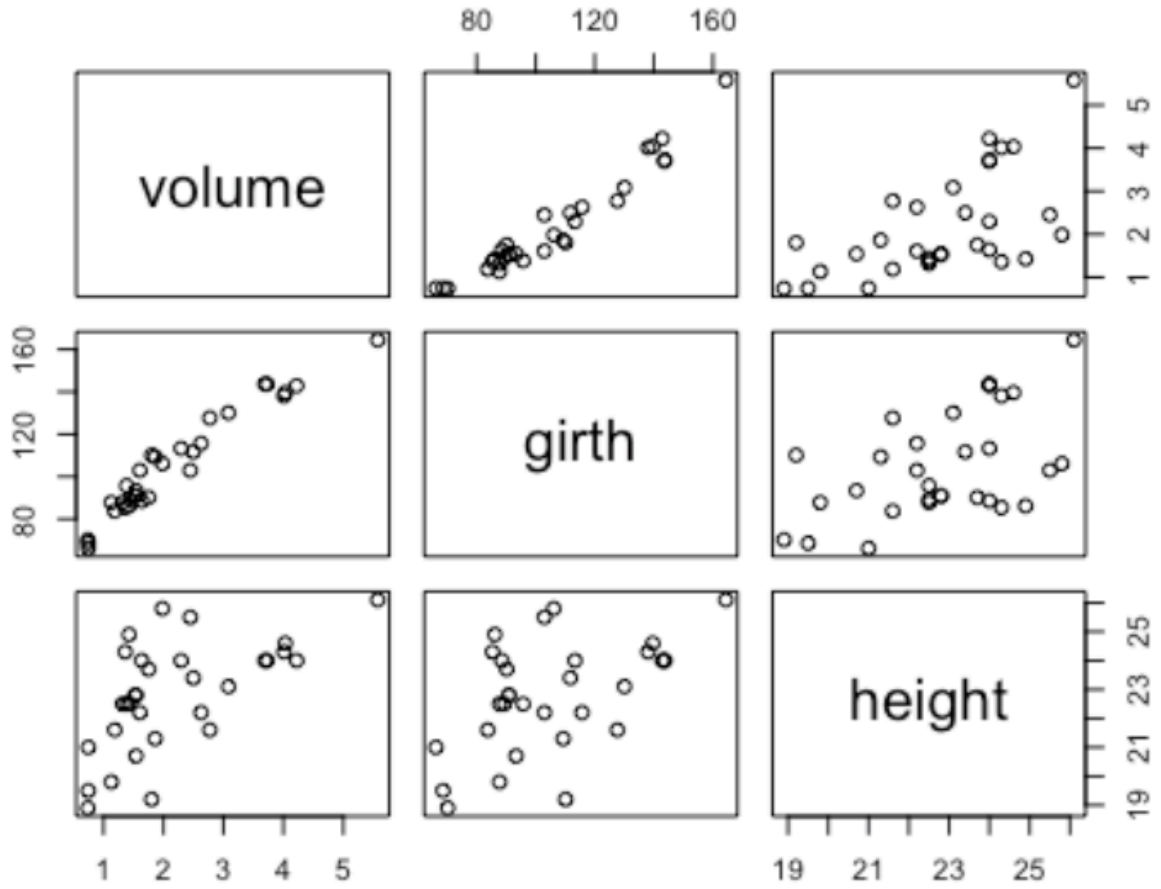
5 – Is there collinearity among the data?

- Collinearity = correlation between covariates
- Use *Variance Inflation Factor* to test for collinearity among covariates (more in HO)



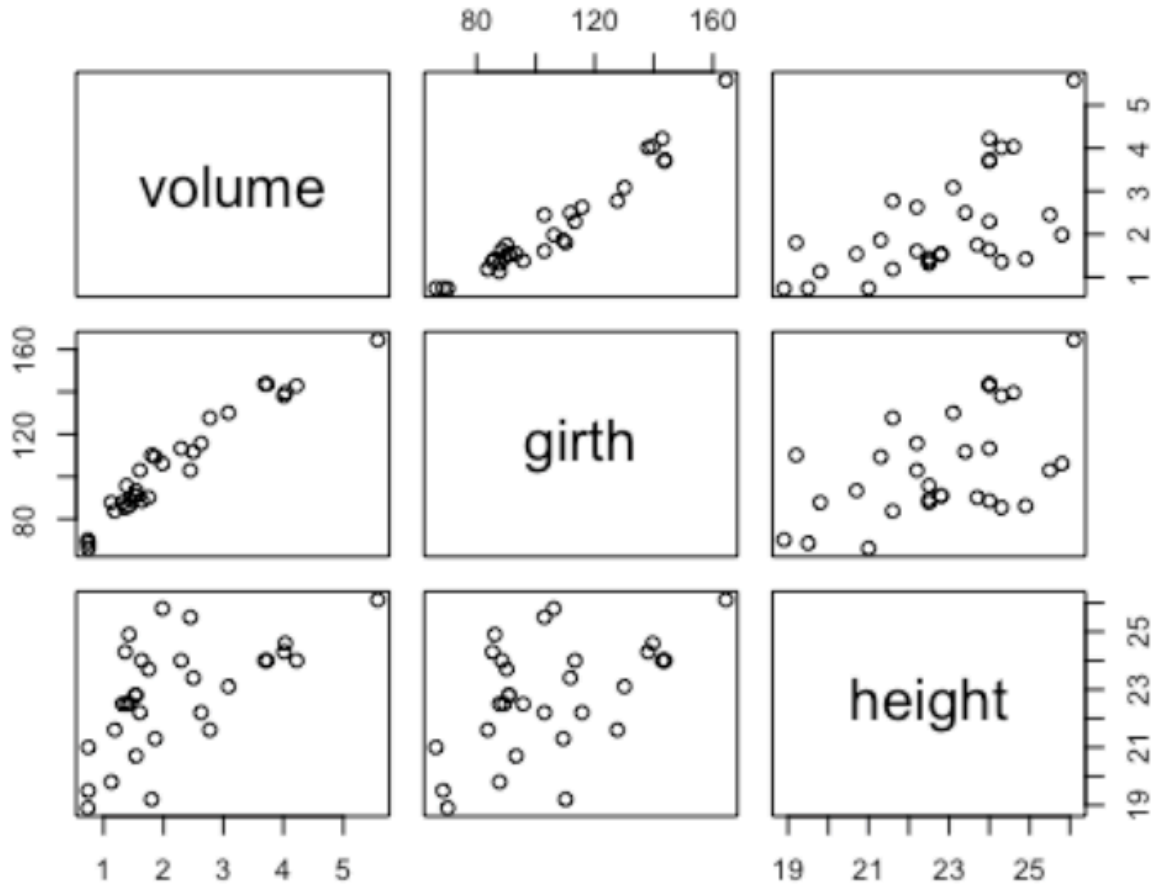
5 – Is there collinearity among the data?

- Collinearity = correlation between covariates
- Use *Variance Inflation Factor* to test for collinearity among covariates (more in HO)
- VIF > 10: reason to be concerned



5 – Is there collinearity among the data?

- Collinearity = correlation between covariates
- Use *Variance Inflation Factor* to test for collinearity among covariates (more in HO)
- VIF > 10: reason to be concerned
- VIF > 3: reason to be guarded



6 – Visually inspect the relationships of interest

- Plot x vs y

7 – consider which interactions you want to add

- Not only interactions, but also additional covariates you want to account for (time of season ect.)

8 – construct maximal model

- $Y \sim x_1 + x_2 + x_3 + \ldots$

- Consider biology! Re-think your question. Does the model map your question?

9 – simplify your model

- Use a method of your choice to simplify
- Make sure you can defend your simplification

10 – Decide on a final model

- Based on biologically sound justification

11 – run model validation

- `plot(model)`

- Starry skys and qq plots

12 – interpret your model

- This is where statistics don’t help you much
- Use your brains
- Use *all* the information you gained during the process of doing your analysis
- Interpret the model given the limitations (and potential violotion of assumptions) of your data
- Think biology!

Do it NOW!

- HO 16
- Work through the excercises. They use different datasets this time! Try to understand what goes on.

Do it NOW!

- 1) Run the timber model without the previously found outlier. See what your conclusions are. If you'd publish it, would you do it with outlier or without? Why?
- 2) Use the checklist (1-6) on the plant growth dataset. Fruit is the response, Root the covariate and Grazing the fixed factor. Specify your null model, the maximum model, and find the best model that includes and interaction between Root and Grazing. Do model validation.
- 3) Use the sparrow dataset to find out how much each structural measurement (tarsus, wing, bill) and sex affects body mass.