Linear regression

Sergey Rosbakh
University of Regensburg

Linear regression – useful links

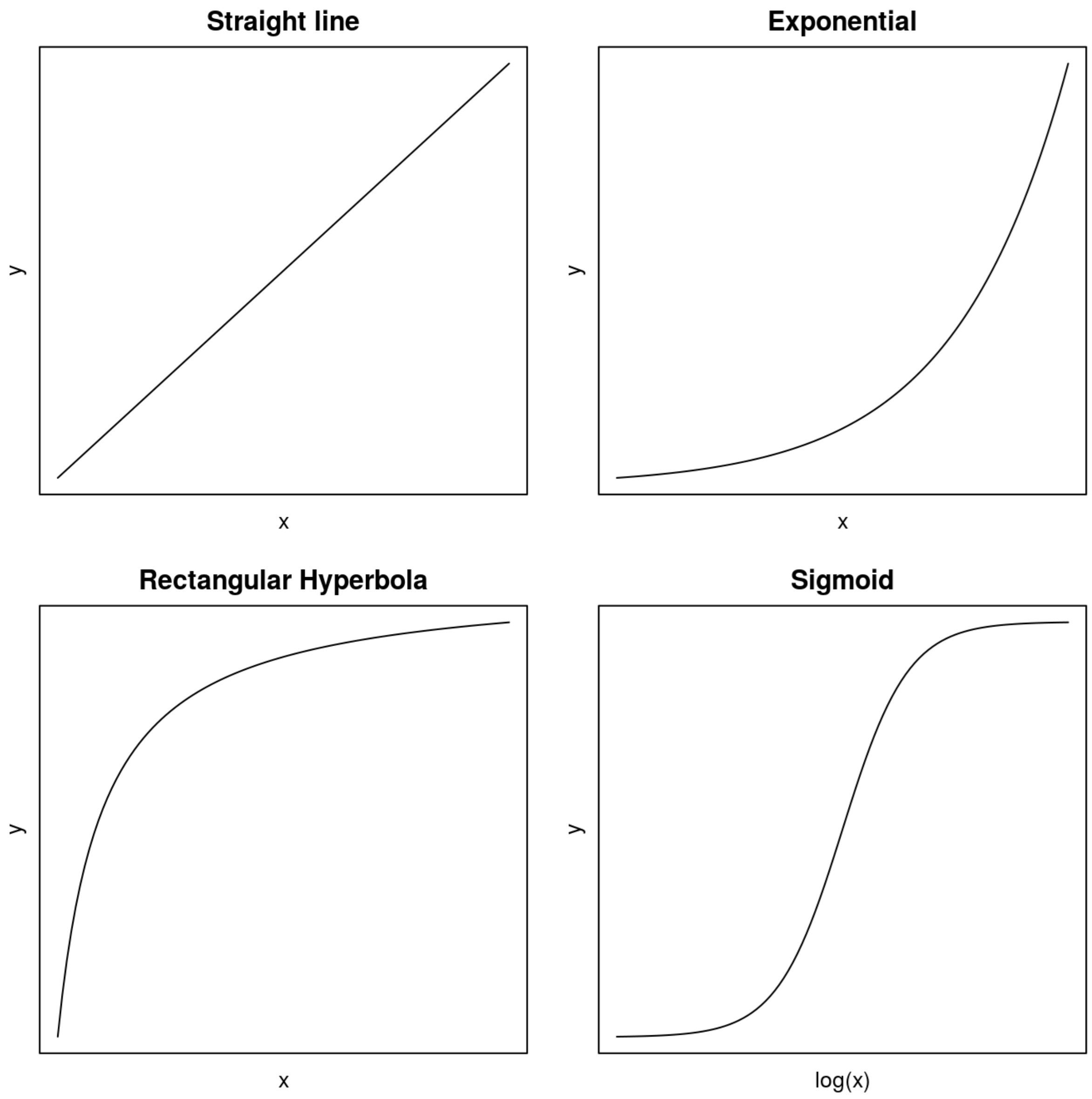
A few articles on linear regression:

- http://r-statistics.co/Linear-Regression.html
- http://r-statistics.co/Assumptions-of-Linear-Regression.html
- https://www.datacamp.com/community/tutorials/linear-regression-R
- http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/
- https://towardsdatascience.com/regression-analysis-linear-regression-239df26a94ac

Science is not much different from playing sorting box…

Science is not much different from playing sorting box…

Linear vs. non-linear relationships



www.rstats4ag.org

- A linear regression (LR) is a statistical model that analyzes the relationship between a **response** variable (**y**) and one (simple LR) or more (multiple LR) **explanatory** variables (**x**) and their interactions

- Mathematical model for simple (one variable) LR:

$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1} \overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

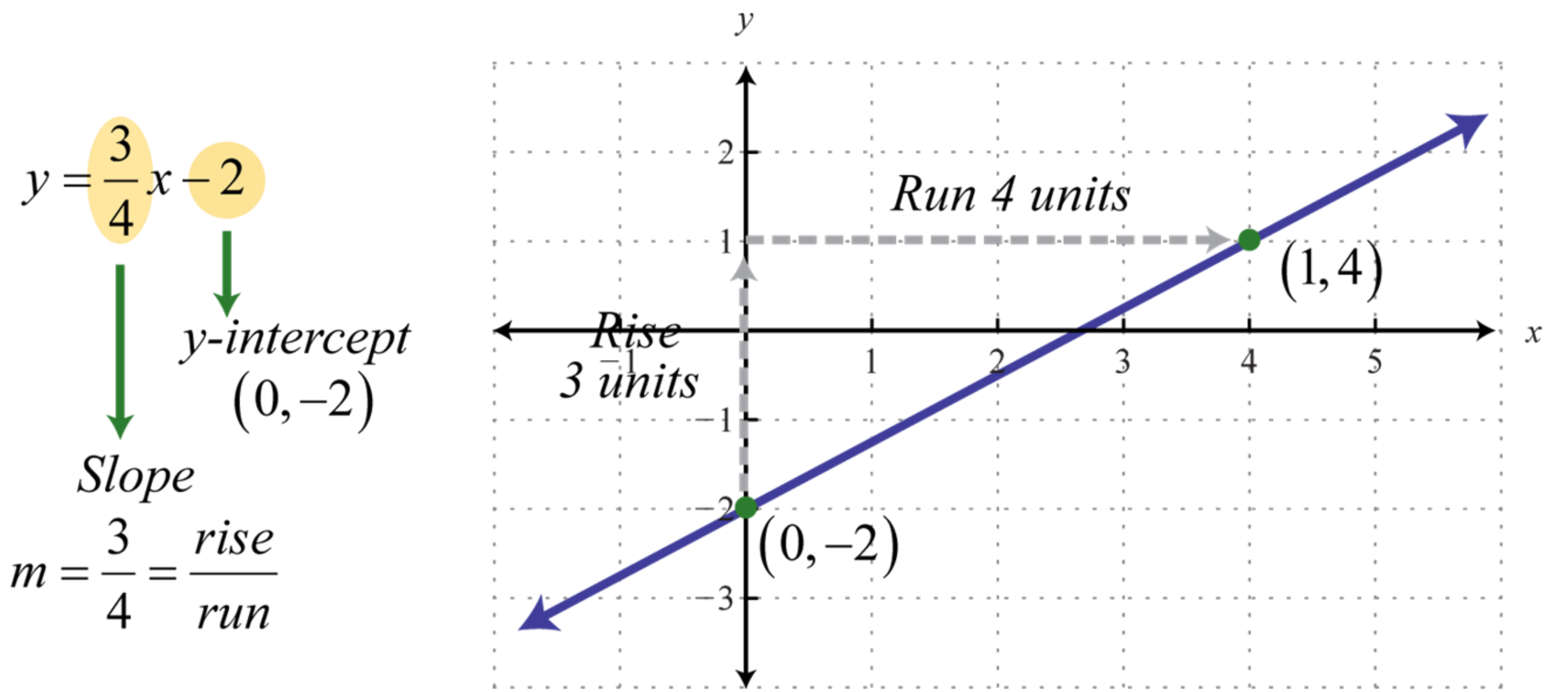
- This model predicts how y varies when x changes

- Beta coefficients are model **intercept** and **slope**. **Residual error** is the part of **y** the model cannot explain

- One of the most used statistical tool worldwide

www.towardsdatascience.com

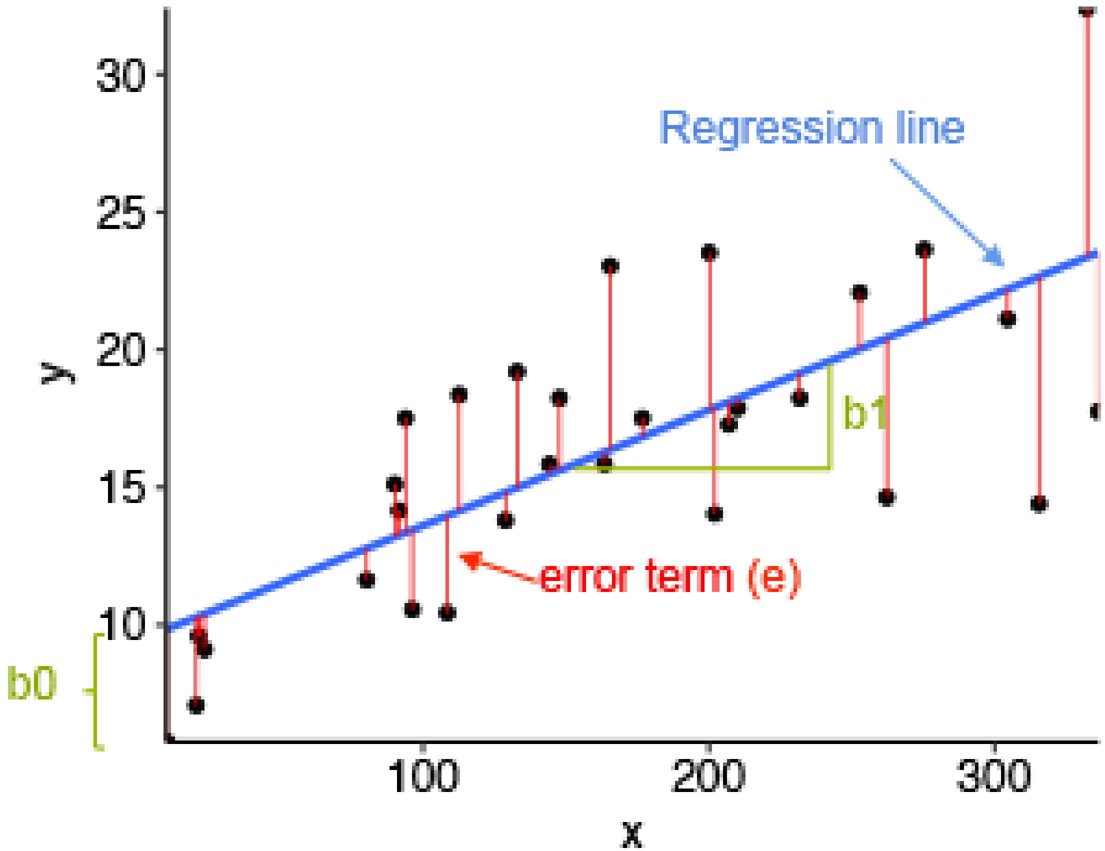
- From a geometrical point of view LR is a linear function



- **Intercept** – the point where the line croses the y-axis

- **Slope** – the rate of change in y when x varies

- From a computational point of view LR is a line with the lowest residual errors (the lowest sum of squared errors)



http://www.sthda.com

LR essentials:

- **Residual sum of squares** (**RSS**) - the sum of the squares of the residual errors. It is an universal metric of model 'geometric' fit (could be used in both linear and non-linear regressions). **The lower the better**.

- Least square regression or **ordinary least squares** (OLS) is the method for determining of the beta coefficients (b0 and b1) so that the **RSS** is as minimal as possible

- **Residual Standard Error** (**RSE**) - the average variation of points around the fitted regression line. This is another metric used to evaluate the overall quality of the fitted regression model. **The lower the better**.

- RSE is a measure of fit for LR. Ideally, it should be zero; in this case **y** can be predicted from **x**

Model summary - syntax

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,    Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

Model summary - coefficients

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- The regression beta coefficients, their standard errors and statistical significance from zero (t-test)

- If the coefficients are not significantly different from zero (p>0.05), then they are set to 0
  Intercept = 0 – the regression line crosses the y-axis at 0
  Slope = 0 – the regression line is parallel to x-axis

Model summary - interpretation

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- The relationship between MAT and elevation in our case is described by the equation: MAT = 9.967559 - 0.003779*Elevation
- Intercept 9.967559: MAT at elevation of 0 m a.s.l.
- Slope -0.003779: 1 meter of increasing elevation results in decreasing MAT by 0.003779 ^{o}C
- Global MAT decrease along elevational gradient is 0.6 ^{o}C/100 meters

Model summary – model accuracy I

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7181 -0.4539  0.1969  0.5750  0.8820

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,    Adjusted R-squared:  0.8467
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- A quick view of the distribution of the residuals
- The median should not deviate strongly from zero
- Minimum and maximum, as well as 1Q and 3Q should be roughly equal in absolute value

Model summary – model accuracy II

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- RSE: the average variation of the observations points around the fitted regression line. The lower the better; could be used to compare two models

- Overdispersion: the case when the model fails to explain the variation in the data. A quick test: RSS/degrees of freedom should be less than 1

Model summary – model accuracy III

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- R2 represents the proportion of information (i.e. variation) in the data that can be explained by the model. The closer to 1 the better. R2 = 1 means that all data points lay on the regression line (perfect fit)
- Important: including more variables into model leads to higher R2 – report adjusted R2 when working with multiple LRs

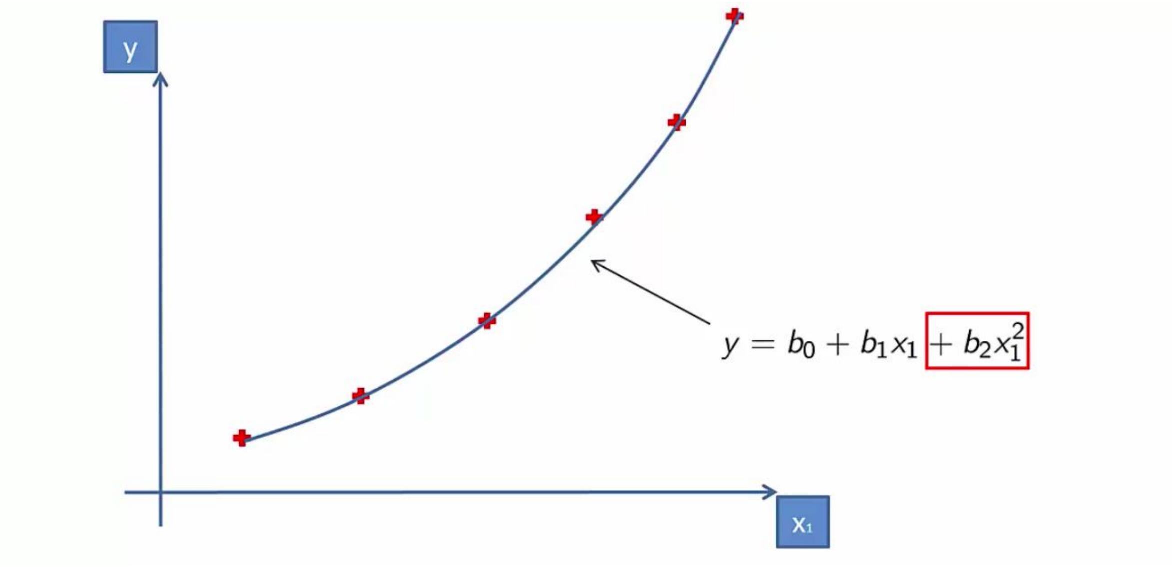
Model summary – model accuracy IV

```
Call:
lm(formula = MAT ~ Elevation, data = climdat2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7181 -0.4539  0.1969  0.5750  0.8820 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.967559   0.399323   24.96  < 2e-16 ***
Elevation   -0.003779   0.000327  -11.56 4.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6849 on 23 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8467 
F-statistic: 133.6 on 1 and 23 DF,  p-value: 4.668e-11
```

- The F-statistic gives the overall significance of the model. It assess whether at least one predictor variable has a non-zero coefficient (p – value). More important for multiple LRs.

Linear model assumptions

Assumption 1: The regression model is linear in parameters



- In layman's terms: the tested relationship should be linear
- Check the residual plots: there should be **NO** pattern
- Polynomial linear regressions could be calculated with the same lm () framework by including quadratic (I(Elevation^2)) or cubic (I(Elevation^3) terms

Linear model assumptions

Assumption 2: The mean of residuals is (close to) zero

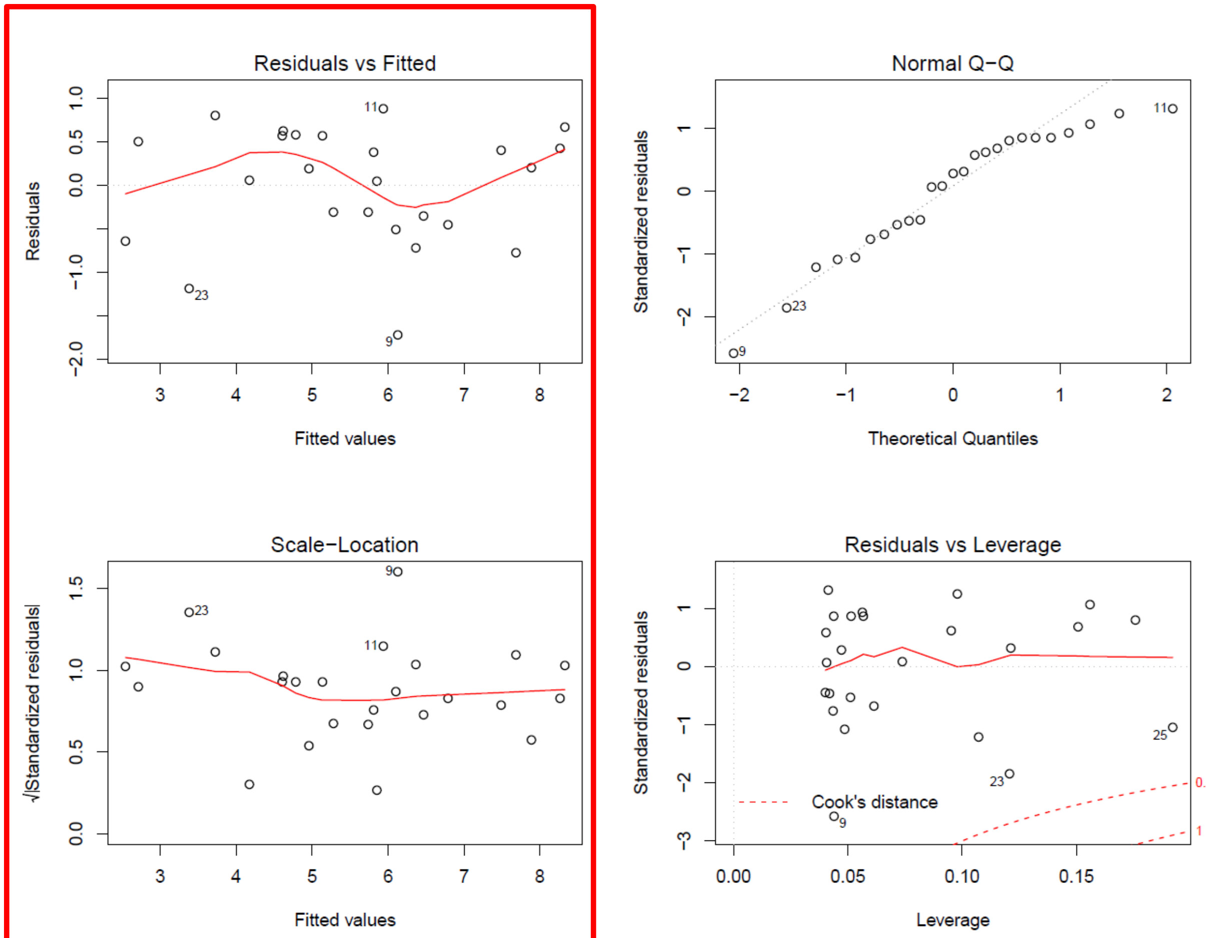
- mean(modelname$residuals)

[1] -1.493071e-17

Linear model assumptions

Assumption 3: Homoscedasticity of residuals or equal variance or residuals

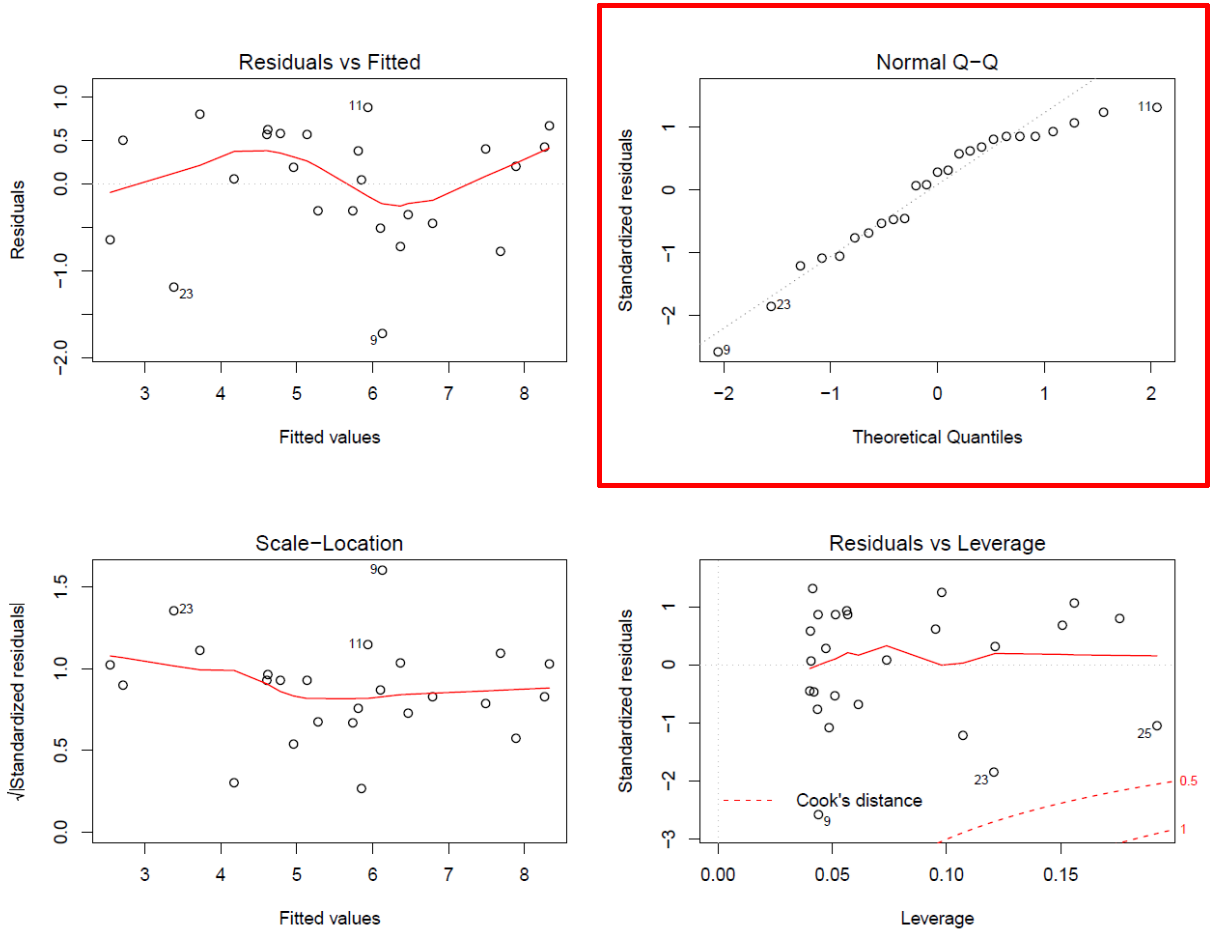
- The red line in the top-left and bottom-left graphs should be flat (or close to it). In the bottom-left graphs the values are standardized



Linear model assumptions

Assumption 4: Normality of residuals

- The **residuals** should be **normally** distributed. The common misconception is to check, if the variables ('data') are normally distributed (it is an assumption for ANOVA)



Linear model assumptions

Assumption 5: Lack of correlation between the explanatory variables and residuals.

- To check: run a correlation test cor.test ()
- Non-significant (>0.05) p-values and low correlation coefficients (the smaller the better) indicate lack of correlation

Linear model assumptions

Further assumptions (less relevant for our case study, but can be crucial for other studies):

- **Assumption 6**: there should be some variability in explanatory variables, i.e. the variability in explanatory values is positive. In our case the climate data should not origin from one elevation only, but from many different ones.
- To check: var ()

- **Assumption 7**: the number of observations must be greater than number of explanatory variables. Although a self-evident and intuitive assumption, it could be a problem in studies with very low number of replicates.
- The rule of thumb: 10 observations per explanatory variable in the regression. In our case data from 10 stations would be enough to estimate the Elevation ~ MAT relationship

- **Assumption 8**: Residuals should not be autocorrelated (a specific problem in **time-series** data).

Linear model assumptions

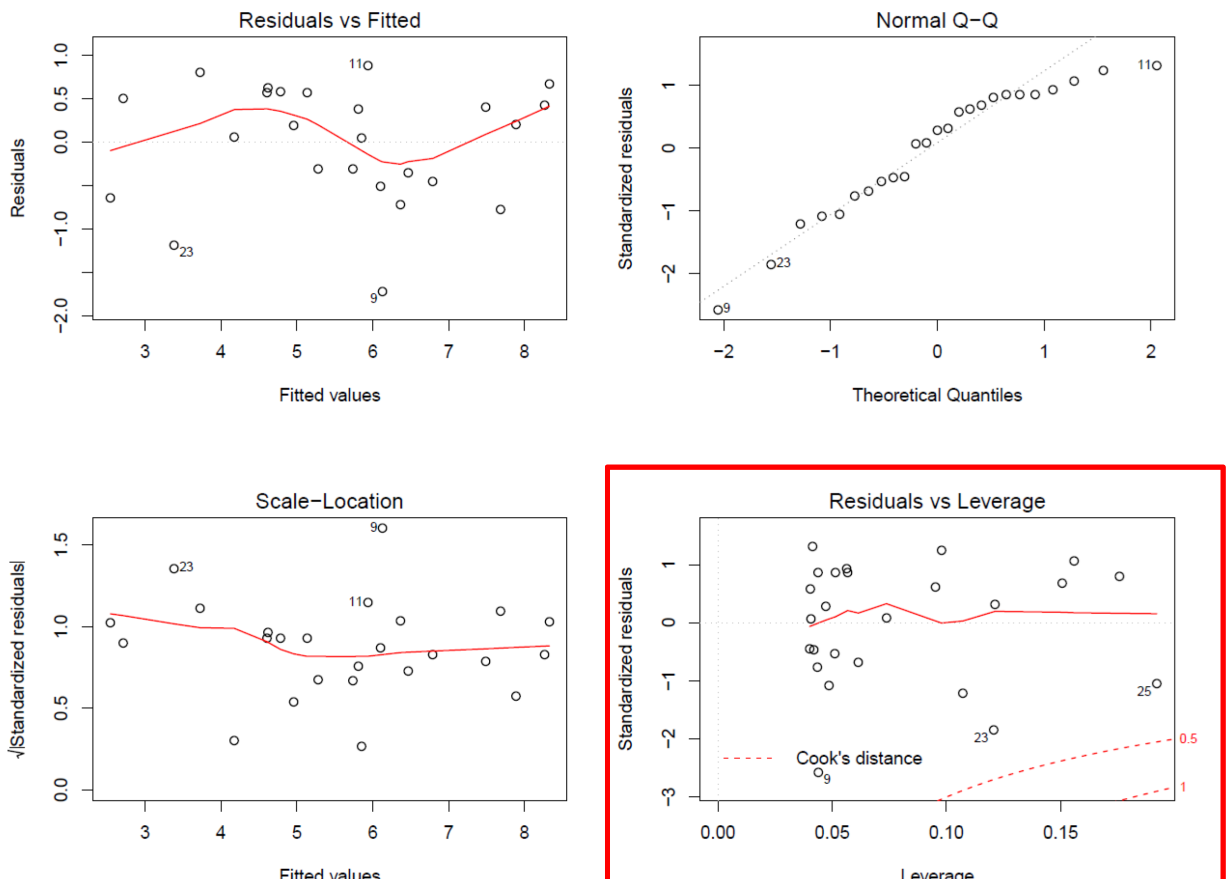
Further assumptions (less relevant for our case study, but can be crucial for other studies):

- **Assumption 9**: Multicollinearity: explanatory variables are correlated with each other
- To check: run a correlation test

- **Assumption 10**: Data points should be independent from each other (see the part on phylogeny)

Influential points

- The course of regression line through the ‘data cloud’ strongly depends on position of every single data point. Thus, some observations can have stronger influence than others
- You should take a very close look at such observations: are they outliers? measurement errors? wrong species? typos?



Generalized linear models

Sergey Rosbakh
University of Regensburg

Linear reression is a flexible framework

- **Group comparison** - only intercepts included into the model = ANOVA
- **Multiple regression** – several (independent) explanatory variables are included into the model
- Allows for accounting random effects – **linear mixed effect models**
- Can be adjusted to different cases with different residual distributions – **generalized linear models**
- Different types of explanatory variables can be used: numeric, categorical and binary data - **generalized linear models**

Useful links

- Wikipedia: https://en.wikipedia.org/wiki/Generalized_linear_model
- https://www.statmethods.net/advstats/glm.html
- https://www.theanalysisfactor.com/count-models-understanding-the-log-link-function/
- https://stats.stackexchange.com/questions/190763/how-to-decide-which-glm-family-to-use

- Logistic regression: https://www.theanalysisfactor.com/r-tutorial-glm1/
- Logistic regression2: https://www.mango-solutions.com/blog/an-intro-to-models-and-generalized-linear-models-in-r

Definition

- **Generalised Linear Mode (GLM)** is a flexible generalisation of an ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution (McCullagh and Nelder, 1982)

- GLMs are extensions of linear regression models that allow the dependent variable to be non-normal

- The word ‘linear’ in GLM does not necessarily require the linearity of a model. In fact, linear regression is just a special case that holds linearity

Dependent Variable

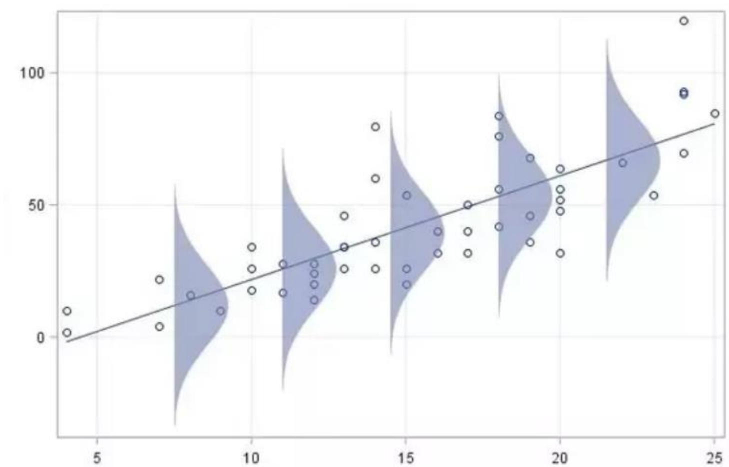
Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

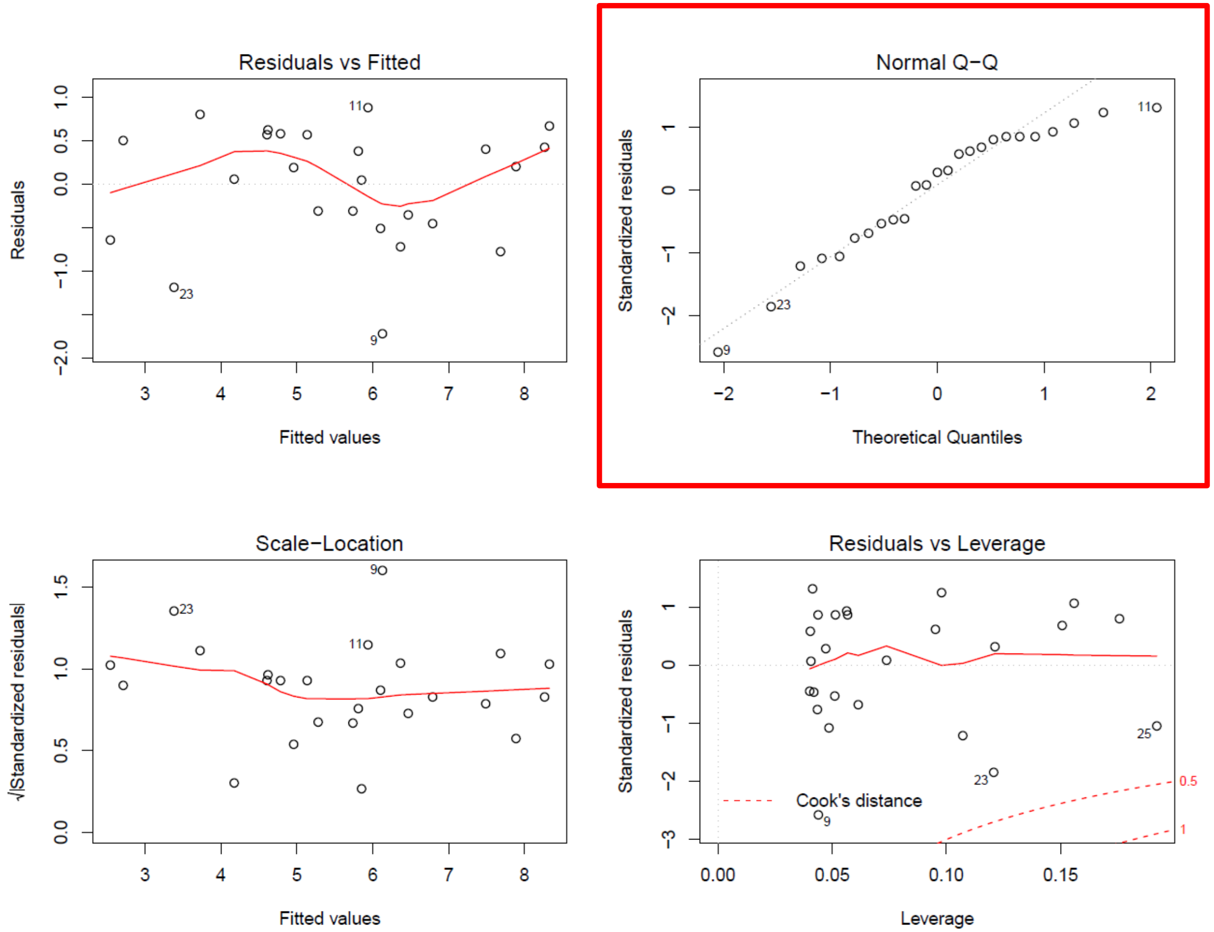


www.towardsdatascience.com

Linear model assumptions

Assumption 4: Normality of residuals

- The **residuals** should be **normally** distributed. The common misconception is to check, if the variables ('data') are normally distributed (it is an assumption for ANOVA)



GLM - link

- In R: glm(*formula*, family=*familytype* (link=*linkfunction*), data=)
- Technically, a glm is simply a linear model working with transformed data
- The family describes the error structure of the data, i.e. it ‘tells’ the software what kind of data you are dealing with
- Link types:

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

GLM - link

- A very short guide to the links:

Data type	Example	Family
Continuous	Temperature	gaussian
Continuous non-negative	Distance	Gamma; inverse.gaussian
Counts	Number of germinated seeds	poisson (mean is equal to variance)
		quasipoisson (mean is not equal to variance)
Binary (0 or 1)	Dispersal events ('yes' or 'no')	binomial
Probability (ranges from 0 to 1)	Germination percentage	binomial
Proportion (ranges from 0 to 1)	Proportion of dispersed seeds	binomial

GLM - assumptions

- The data **y1**, **y2**, …, **yn** are independently distributed
- The homogeneity of variance does not need to be satisfied. In fact, it is not even possible in many cases given the model structure, and overdispersion (when the observed variance is larger than what the model assumes) may be present
- Errors need to be independent but not normally distributed
- It uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS) to estimate the parameters and thus relies on large-sample approximations
- Goodness-of-fit measures rely on sufficiently large samples

GLM - overdispersion

- Overdispersion describes the observation that variation is higher than would be expected. Some distributions do not have a parameter to fit variability of the observation

- Overdispersion arises in different ways, most commonly through “clumping”

- The rule of thumb is that the ratio of deviance to df should be 1

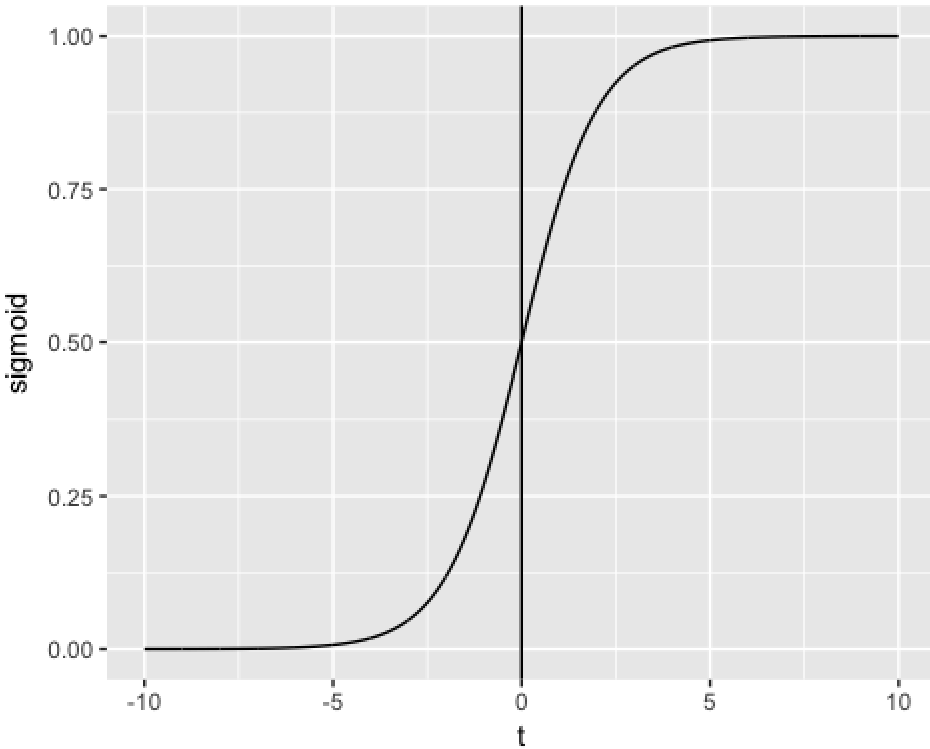
- Formal testing: package DHARMa

- How to deal with overdispersion:

1) Use quasi – families (no test for overdispersion available)
2) Use different distribution (e.g. negative binomial)
3) Observation-level random effects. Technically, a mixed-effect model with data IDs included as random factor.

Find out more: http://biometry.github.io/APES//LectureNotes/2016-JAGS/Overdispersion/OverdispersionJAGS.html

GLM – logistic regression



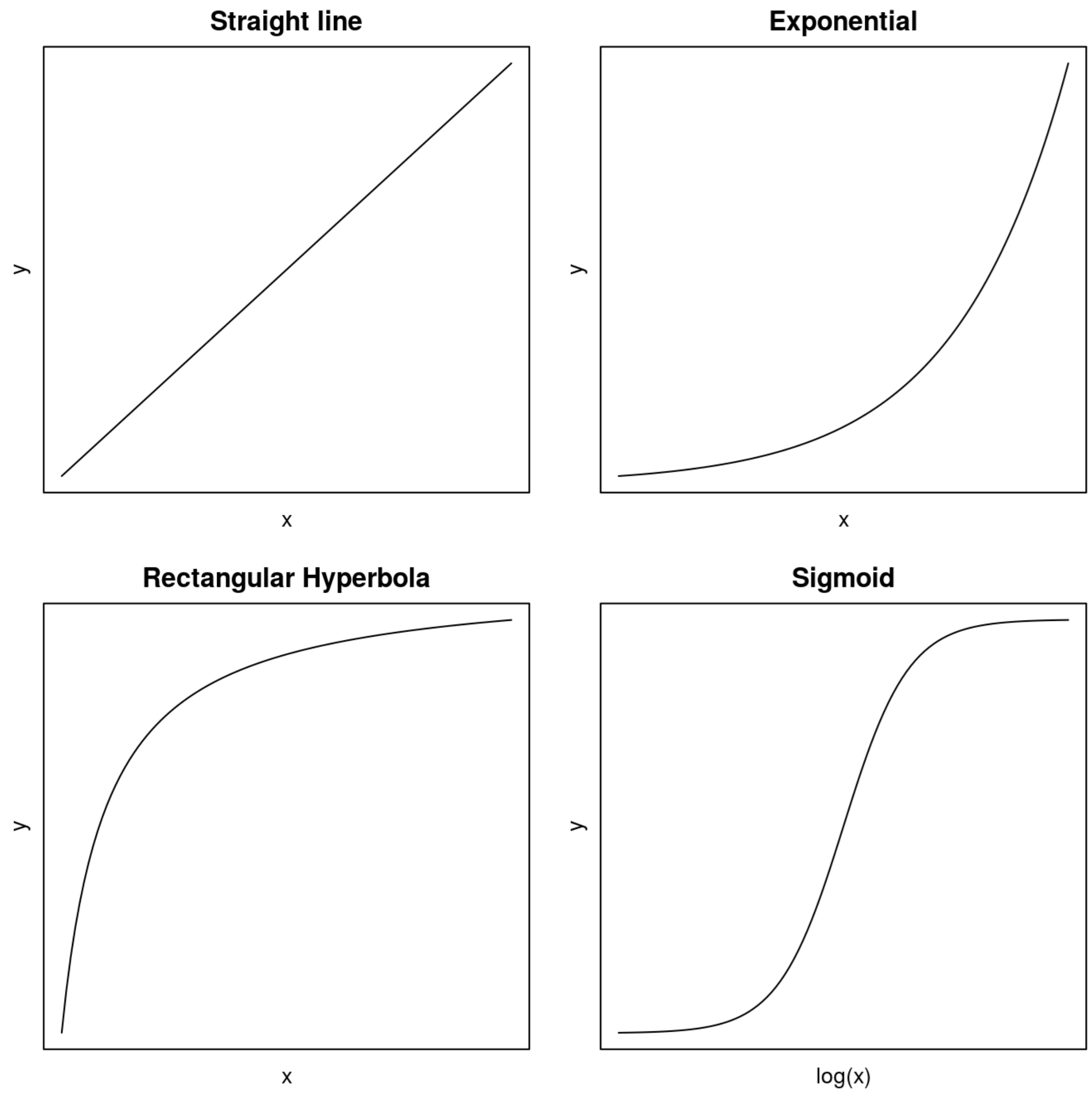
$$\sigma(t) = \frac{1}{1 + exp(-t)}$$

- logistic regression accepts only dichotomous (binary) input as a dependent variable
- The output is a probability of an event
- Estimates are logits
- No standard errors – confidence intervals

Non-linear regression

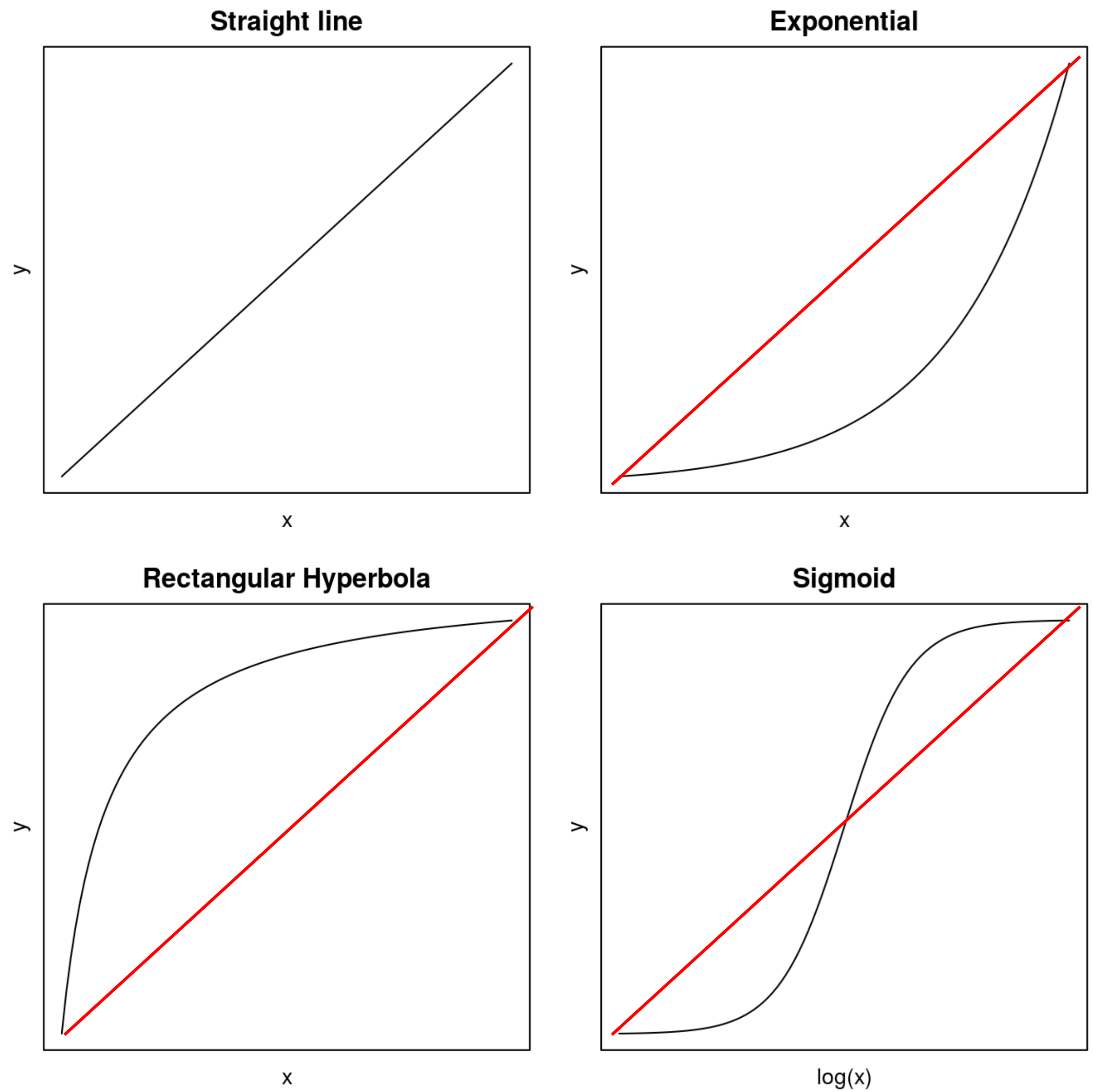
Sergey Rosbakh
University of Regensburg

Linear vs. non-linear relationships



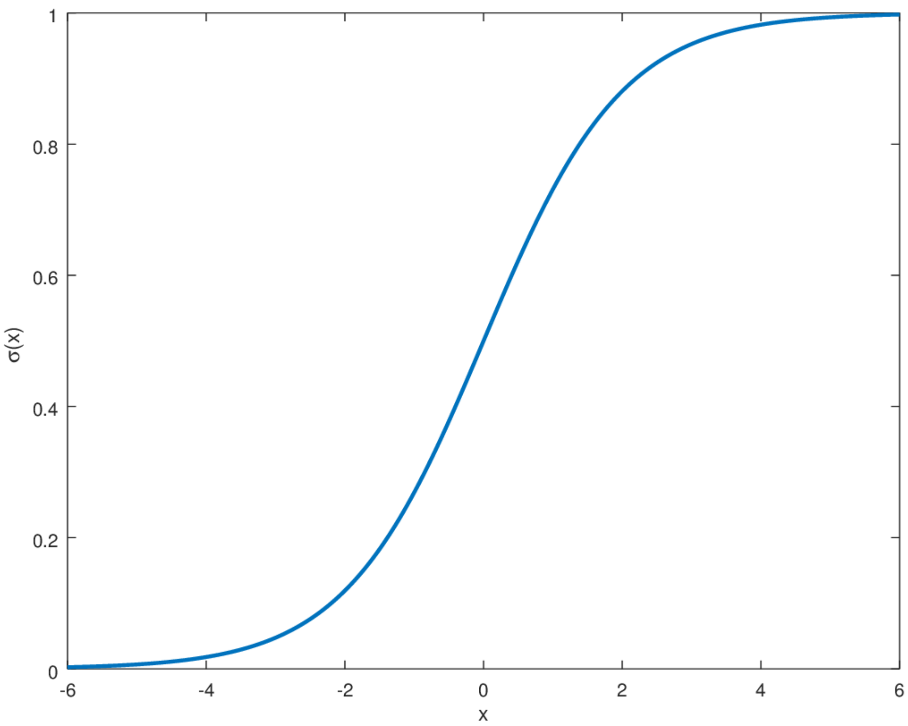
www.rstats4ag.org

Linear vs. non-linear relationships



Symmetric log-logistic model

$$f(x,(b,c,d,e)) = c + \frac{d-c}{1+\exp\{b(\log(x)-\log(e))\}}$$



y – response
C – lower limit (asymptote)
D – upper limit (asymptote)
B – slope
E – point of inflection (X50)

Revkin et al., 2008

Package drc (dose-response curves)

PLOS | ONE

- Engine: drm (y ~ x, fct=…)
- Available functions

RESEARCH ARTICLE

Dose-Response Analysis Using R

Christian Ritz[1]*, Florent Baty[2], Jens C. Streibig[3], Daniel Gerhard[4]

Table 1. List of model functions and corresponding names of some of the most important built-in models available in *drc*.

Model type	Model function (f)	Function in *drc*
Generalized log-logistic	$c + \frac{d-c}{(1+\exp(b(\log(x)-\log(e))))^f}$	`llogistic()`
Brain-Cousens	$c + \frac{d-c+fx}{1+\exp(b(\log(x)-\log(e)))}$	
Cedergreen-Ritz-Streibig	$c + \frac{d-c+f\exp(-1/(x^{\alpha}))}{1+\exp(b(\log(x)-\log(e)))}$	`cedergreen()`
(0 < α < 1 is usually fixed in advance)		
Log-logistic fractional polynomial	$c + \frac{d-c}{1+\exp(b(\log(x+1))^{\beta_1}+e(\log(x+1))^{\beta_2})}$	`fplogistic()`
Log-normal	$c+(d-c)\Phi(b(\log(x)-\log(e)))$	`lnormal()`
(Φ: distribution function for a normal distribution)		
Weibull I	$c+(d-c)\exp(-\exp(b(\log(x)-\log(e))))$	`weibull1()`
Weibull II	$c+(d-c)(1-\exp(-\exp(b(\log(x)-\log(e)))))$	`weibull2()`
Gamma	$c+(d-c)\tilde{\Gamma}(bx, e, 1)$	`gammadr()`
($\tilde{\Gamma}$: distribution function for a Γ distribution)		
Multistage	$c+(d-c)\exp(-b_1-b_2x-b_3x^2)$	`multi2()`
NEC	$c+(d-c)\exp(b(x-e))$ for $x > e$ and d otherwise	`NEC.4()`

doi:10.1371/journal.pone.0146021.t001

Weibull models are asymmetrical