LINEAR REGRESSION I
GENERALIZED LINEAR MODELS

Alejandro Ordonez

Assistant Professor - Department of Bioscience

Section for Ecoinformatics & Biodiversity

Center for Biodiversity Dynamics in a Changing World (BIOCHANGE)

LINEAR REGRESSION
WHAT WE WILL TALK ABOUT TODAY

- Basics of regressions – How this is done?
- Linear regression assumptions – What is needed?
- Collinearity – Which predictors should I use?
- Model selection – Reduce the number of variables used.
- Contrasting predictors – Which predictor is more important?
- Assessing my model – How good is you model?

Any questions?

Ready to start?

REGRESSION ANALYSIS

WHAT DOES "LINEAR" MEAN?

Definition 1

- A model of a **straight-line** relationship between the response (y_i) and a predictor (β_i) variables.
- This is the interpretation most biologists are familiar.

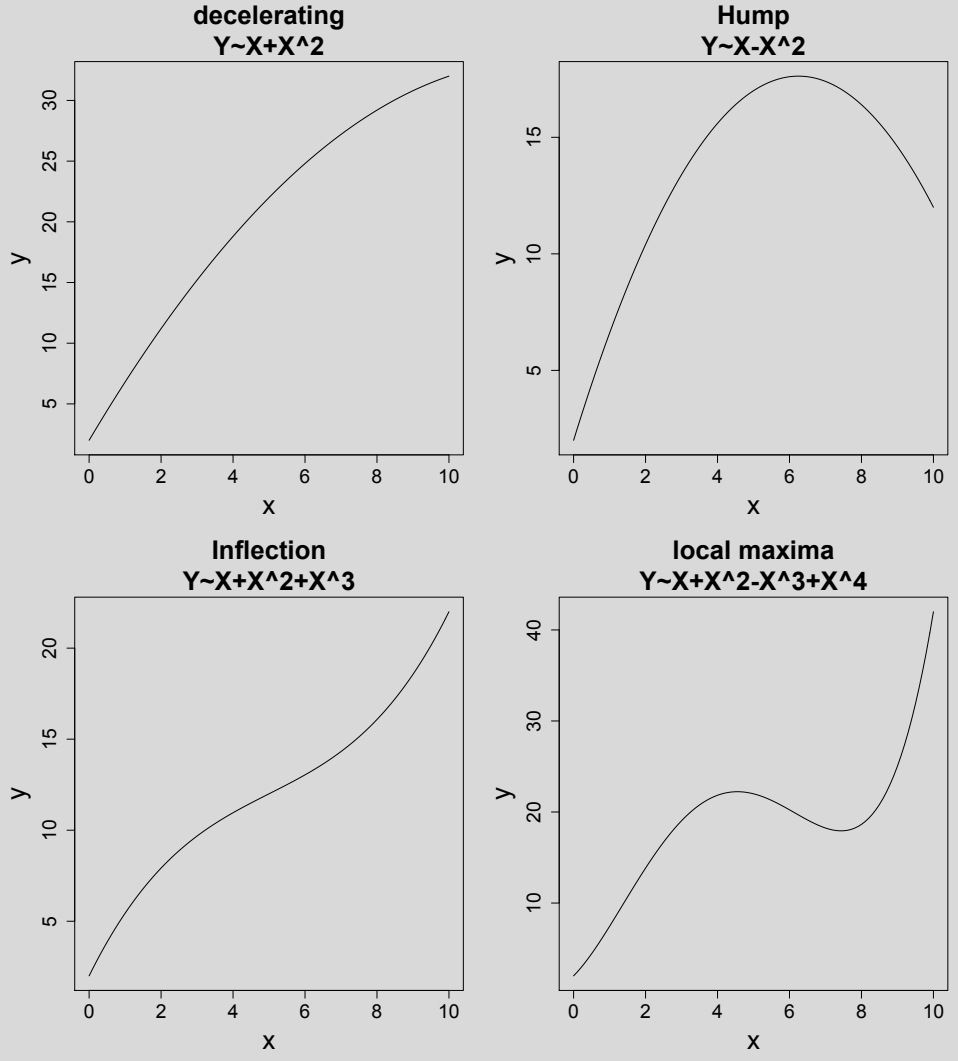
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

*The term "linear" refers to the combination of parameters, **NOT** the shape of the relationship.*

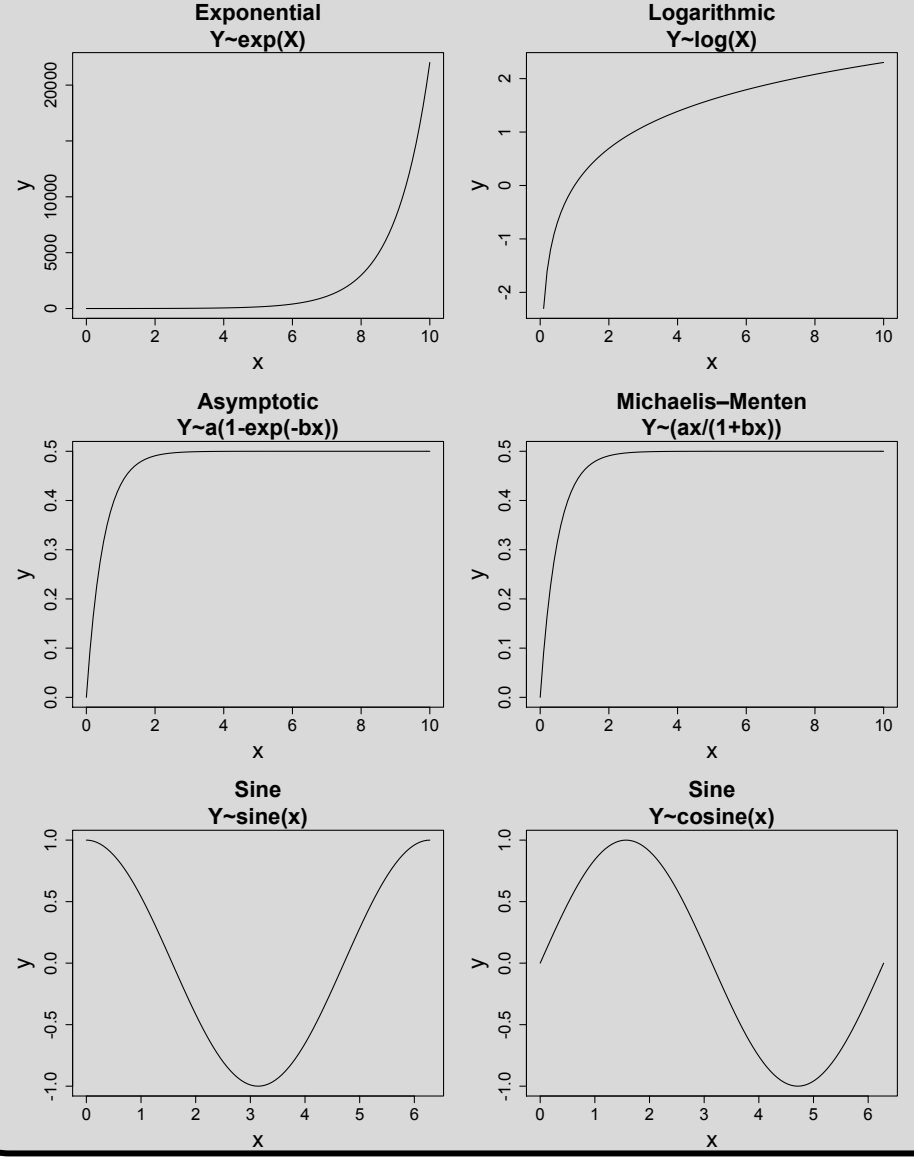
Definition 2

- A model where the variable of interest (y) is described by a **linear combination of a series of parameters** (regression slopes β_i, intercept β_0).
- That is **NO** parameter appears as:
  - an exponent [$e^{X_i\beta_1}$]
  - multiplied [$\beta_1 X_i \times \beta_2(1/X_i)$]
  - divided [$\beta_1 X_i/\beta_2+X_i$]

Linear models



Nonlinear models



ALEJANDRO ORDONEZ GLORIA
ASSISTANT PROFESSOR

SIMPLE LINEAR REGRESSION - ELEMENTS

The value of Y for the i^{th} observation when the predictor variable $X = x_i$.

The random or unexplained error associated with the i^{th} observation.

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

Population intercept (mean value of the probability distribution of Y when $x_i=0$).

Population slope and measures the change in Y per unit change in X.

MULTIPLE LINEAR REGRESSION - ELEMENTS

The value of Y for the i^{th} observation when the predictor variable $X = x_{i.}$

Population slope for Y on X_1 **holding the other X_{ij} constant.**

The random or unexplained error associated with the i^{th} observation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{K} + \beta_p x_{ip} + \varepsilon_i$$

Population intercept mean value of the probability distribution of Y when All $X_{ij} = 0$.

Population slope for Y on X_2 **holding the other X_{ij} constant.**

Population slope for Y on X_p **holding the other X_{ij} constant.**

MULTIPLE LINEAR REGRESSION - ELEMENTS

Partial Regression coefficients

A measure the expected change in the response variable (Y) associated with a one unit change in an independent variable (X_i) **holding the other independent variables ($X_{j \text{ to } p}$) constant.**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{K} + \beta_p x_{ip} + \varepsilon_i$$

Changing which and the number of predictors alters the "**relative**" effect of these.

So far so good?

Any questions?

Ready to continue?

LINEAR REGRESSION ASSUMPTIONS

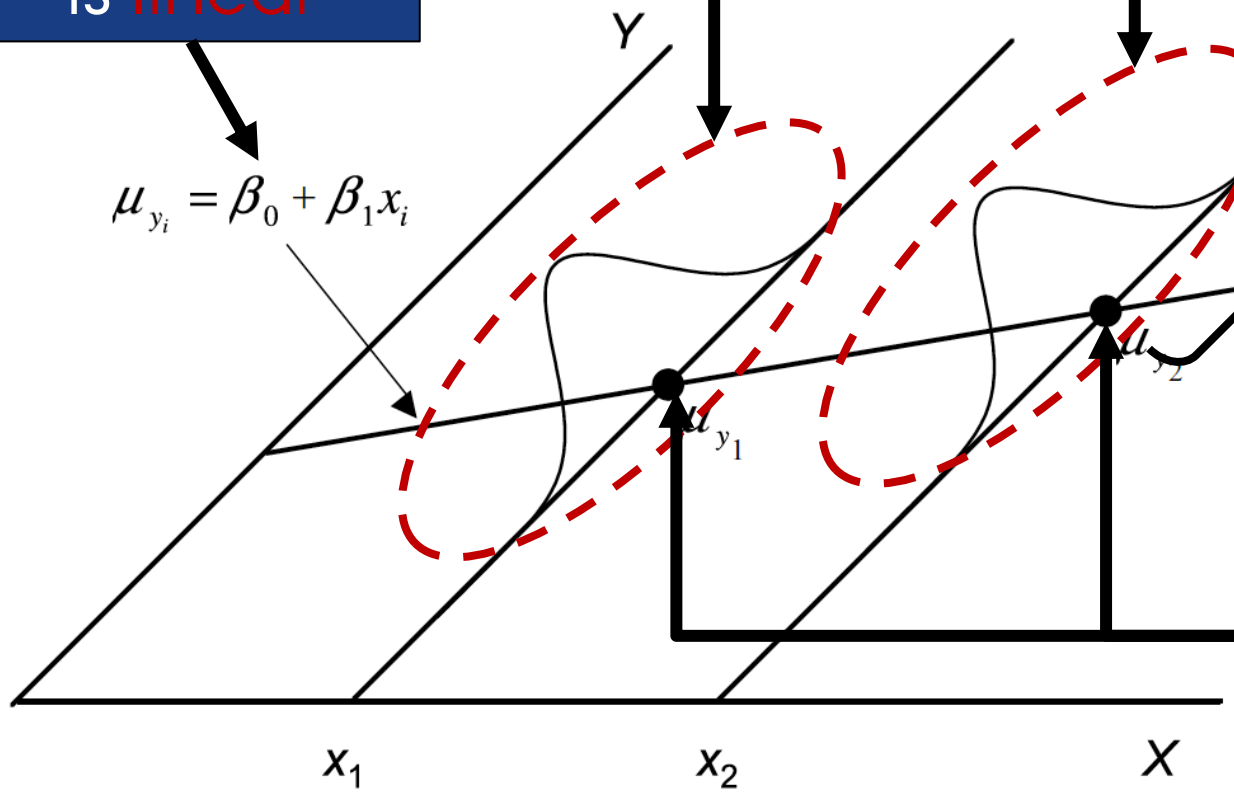
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{K} + \beta_p x_{ip} + \varepsilon_i$$

1. Relationship is linear.
2. Error terms ε_i have an equal variance for each x_{ij}.
3. Observations are independent.
4. Error terms ε_i are normally distributed.
5. No collinearity between predictor variables.

LINEAR REGRESSION ASSUMPTIONS

Relationship is linear

ε_i are normally distributed



ε_i have equal variance

Observations are independent

LINEAR REGRESSION ASSUMPTIONS

Questions to ask

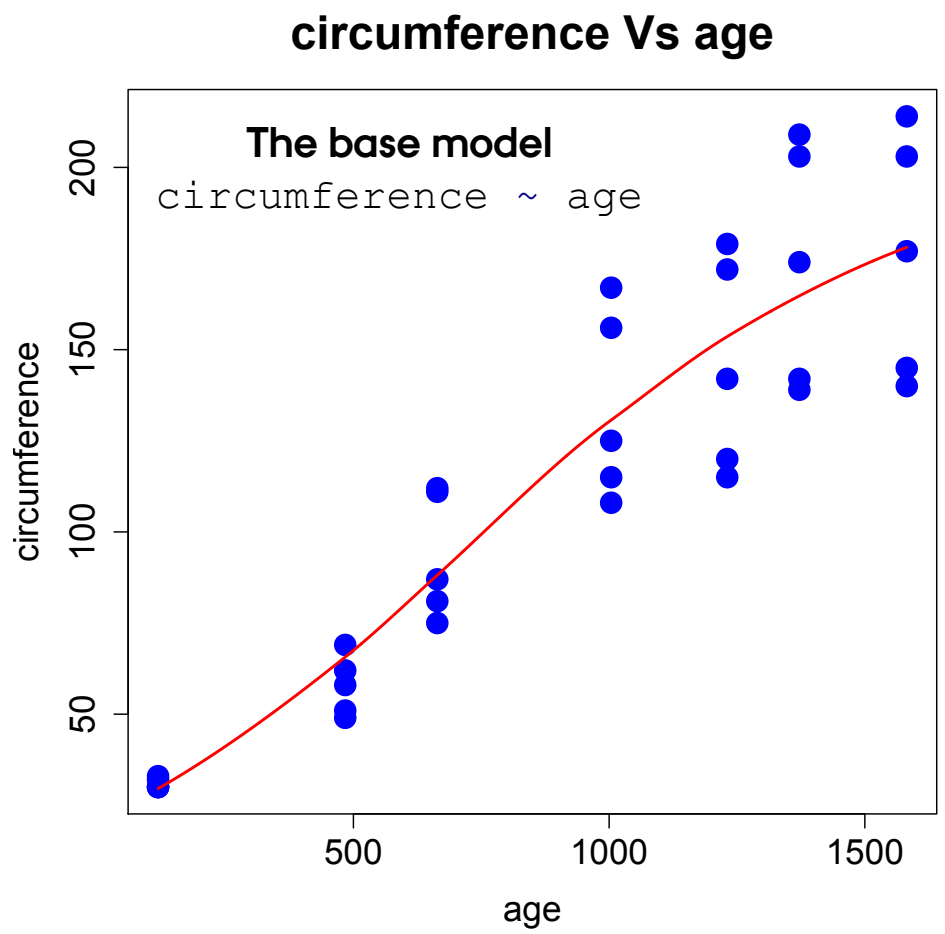
- Non-linear relation?
- Heteroscedastic in the **residuals**?
- Independence of variables?
- Non-normal **residuals**?
- Leverage of an observation?

Visual inspection of a object created using the `lm` function can provide the first pass assessment of these

- Collinear predictors?

Important for Multiple regressions

ASSUMPTIONS - LINEARITY

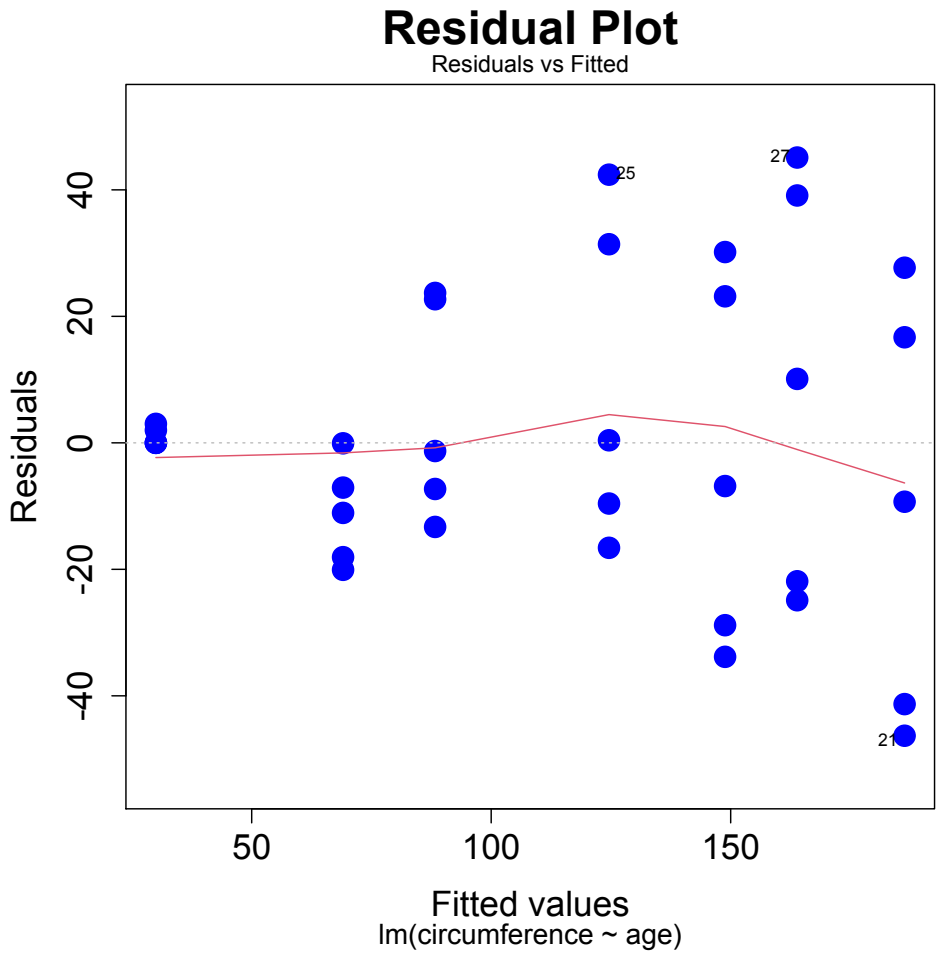


WHAT ARE WE TESTING?

Is the the relationship between dependent and independent variables is additive?

For this we use the model equation and exploratory analyses.

ASSUMPTIONS – HOMOSCEDASTICITY - 1



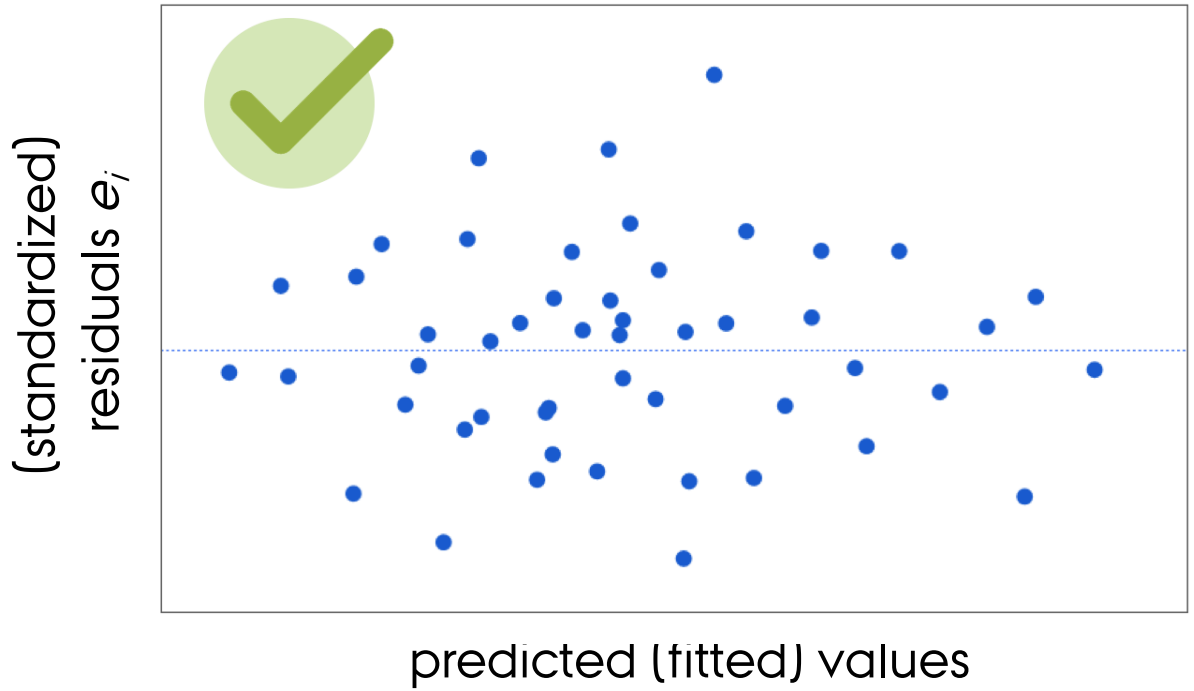
WHAT ARE WE TESTING

Is the **error terms** the same across the same across all values of the **predictions** of the model?

We see **Homoscedasticity** if the residuals appear to form an equal spread around the horizontal line without distance patterns.

ASSUMPTIONS – HOMOSCEDASTICITY - RESIDUAL PLOTS

Residual plot

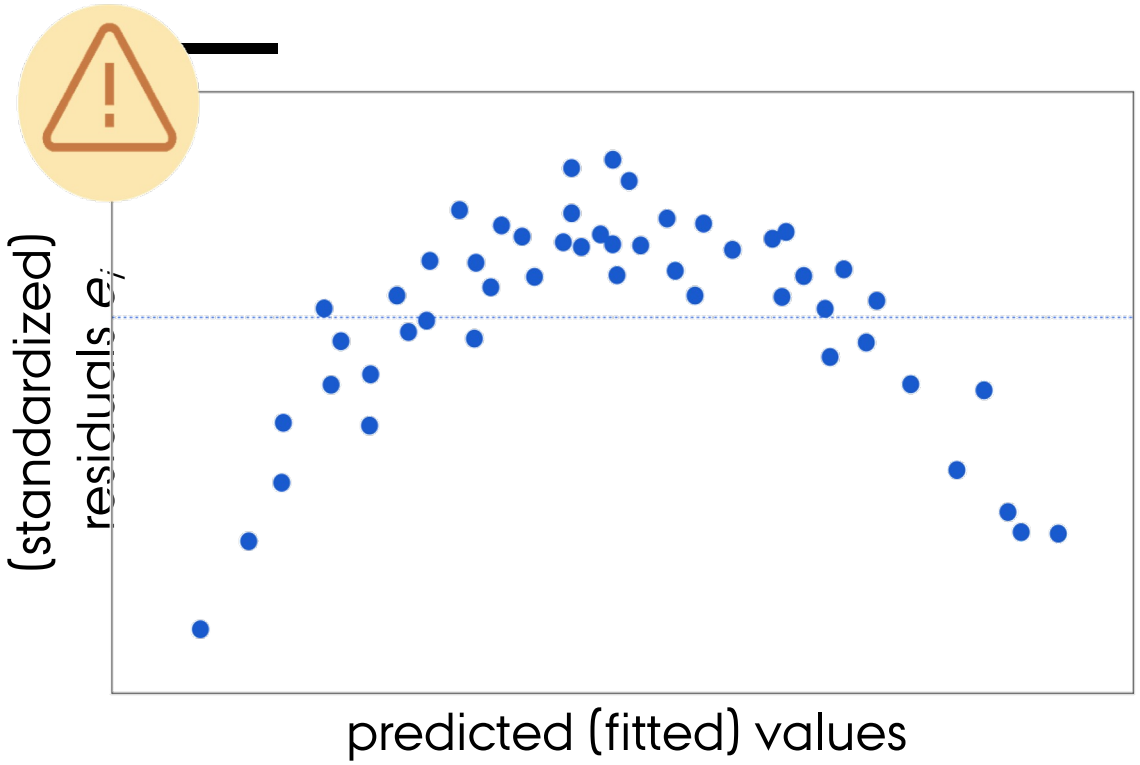


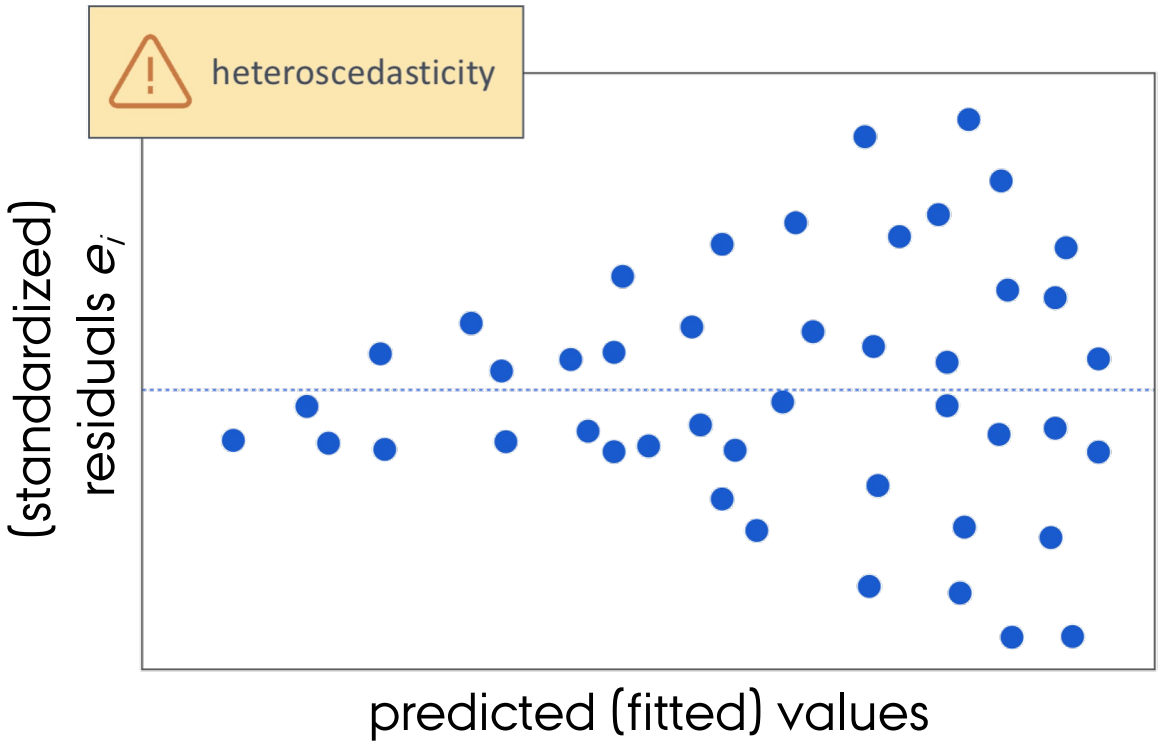
The most useful figure to test regression assumptions

As it helps detecting, heterogeneity of variances, non-linearity, and autocorrelation.

If the assumptions are met, the residuals will look like an unstructured cloud of points.

ASSUMPTIONS – HOMOSCEDASTICITY - RESIDUAL PLOTS

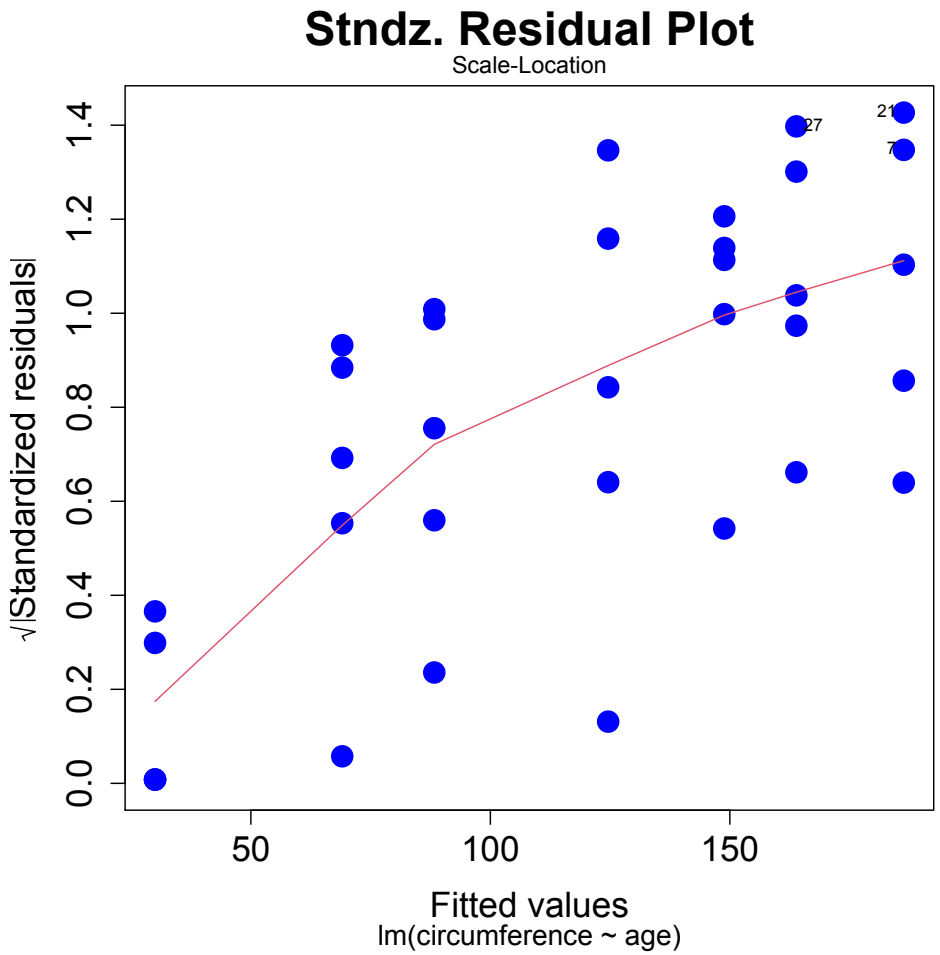




If the assumptions are NOT met

The residuals will look like a structured cloud of points.

ASSUMPTIONS – HOMOSCEDASTICITY - 2

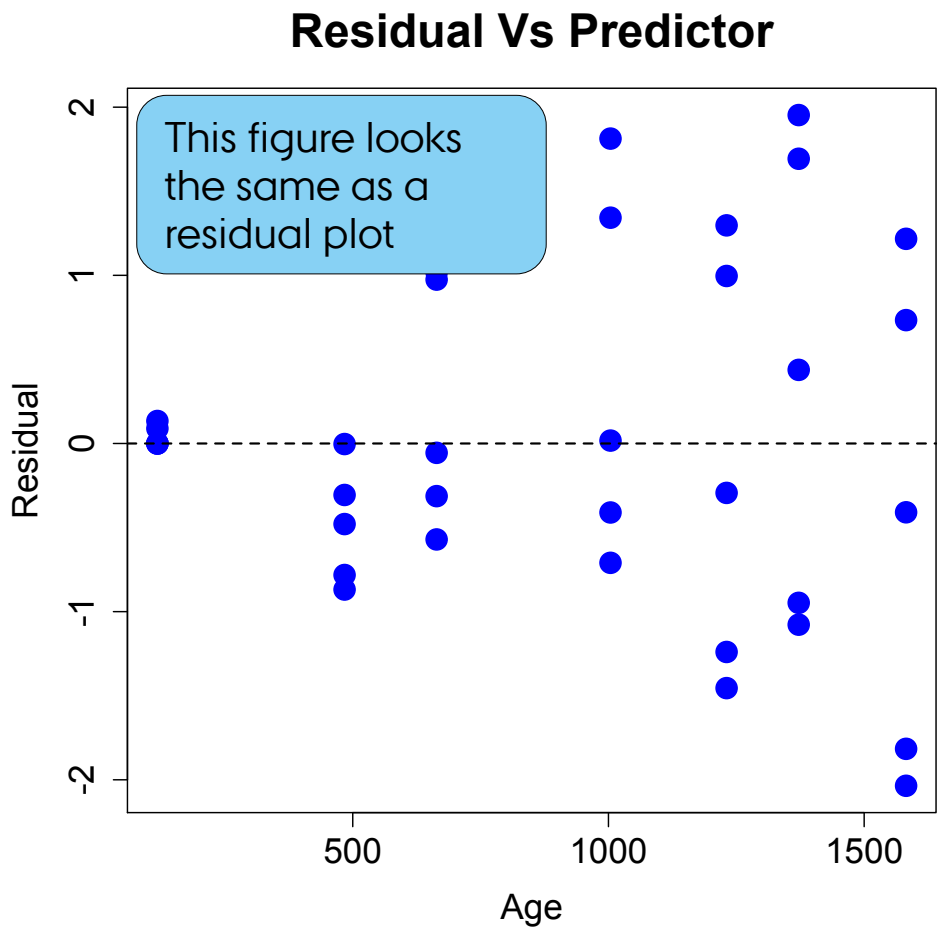


WHAT ARE WE TESTING

Are the **standardized error** terms the same across all values of the **predictions** of the model?

When using the Sqrt of Standardized residuals what we **DO NOT** want to see is a triangular shape.

ASSUMPTIONS - INDEPENDENCE

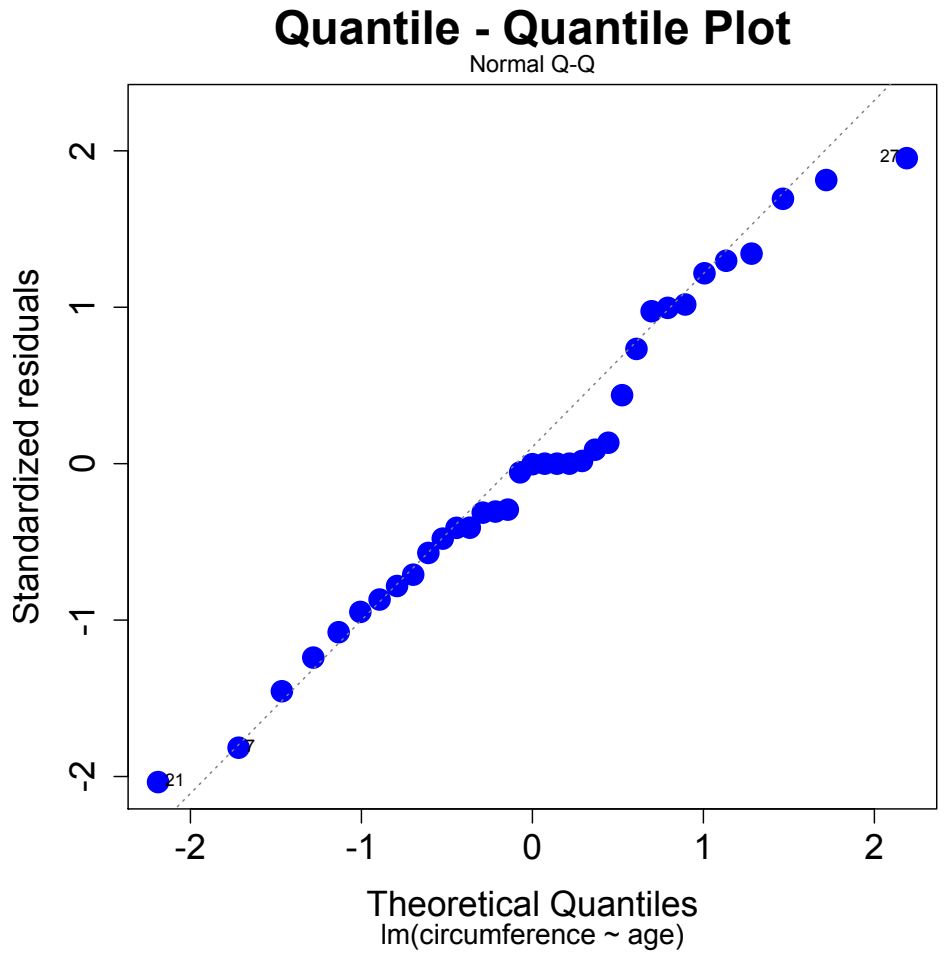


WHAT ARE WE TESTING?

Is there a pattern in the **residuals** as a function of the **predictor**?

As for Homoscedasticity, we are looking for no relation between these residuals and predictors.

ASSUMPTIONS - NORMAL DISTRIBUTION



WHAT ARE WE TESTING

Do the residual error terms meet the assumption of normal distribution?

The residuals are normally distributed if they are in a straight line.

SOLVING PROBLEMS WITH ASSUMPTIONS: TRANSFORMATIONS

power (e.g. square-root).
Logarithmic (e.g. L_n or Log_{10}).

→ **Box-Cox**

Logit.
arcsin square-root.

→ **Proportion data**

Why transform?
*to reduce or remove **non-normality, non-linearity, heteroscedasticity**.*

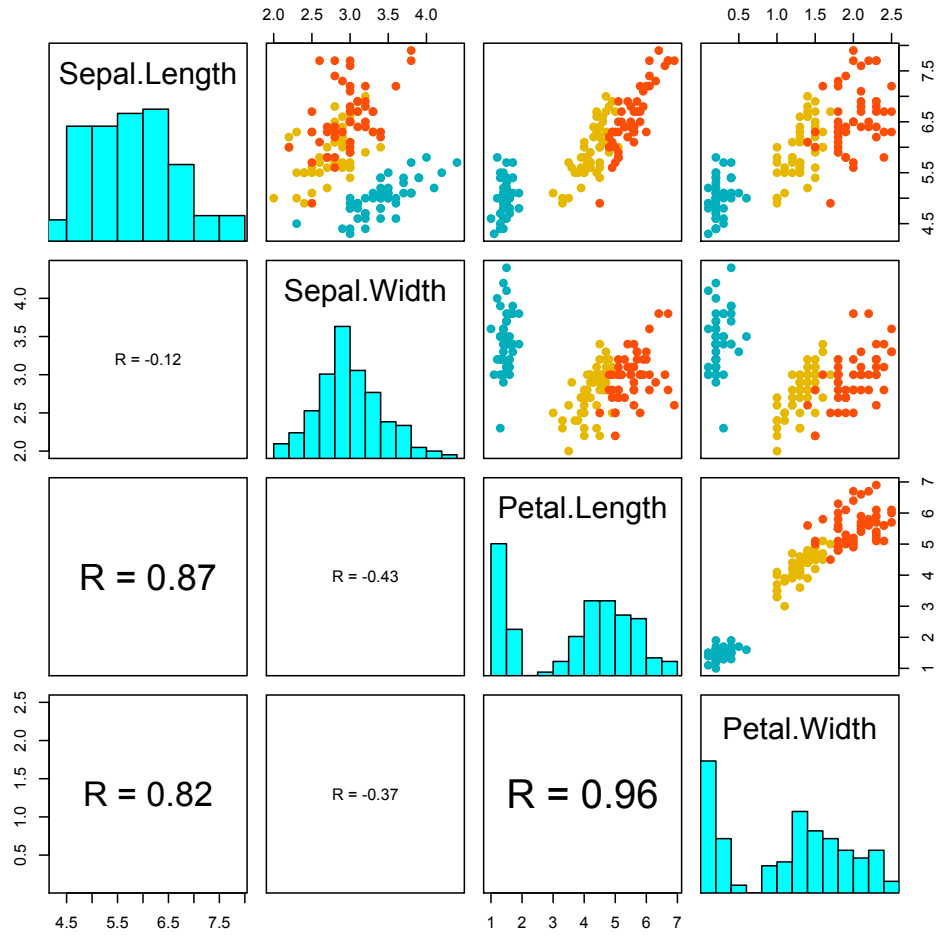
What to transform?
You manly want to transform your response variable.

So far so good?

Any questions?

Ready to continue?

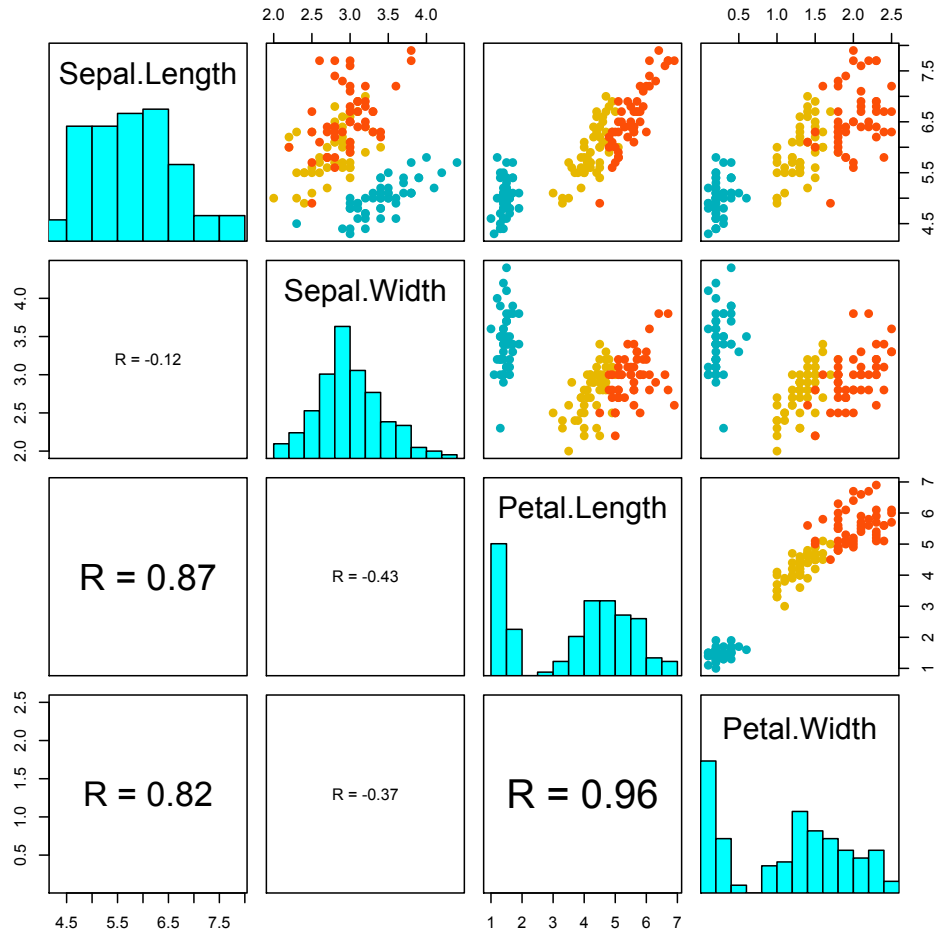
MULTIPLE REGRESSION ASSUMPTIONS – MULTICOLLINEARITY



WHAT ARE WE TESTING?
Are some predictors strongly correlated?

We have collinearity if one predictor can be predicted based on the others with a **substantial** degree of accuracy.

MULTIPLE REGRESSION ASSUMPTIONS – MULTICOLLINEARITY

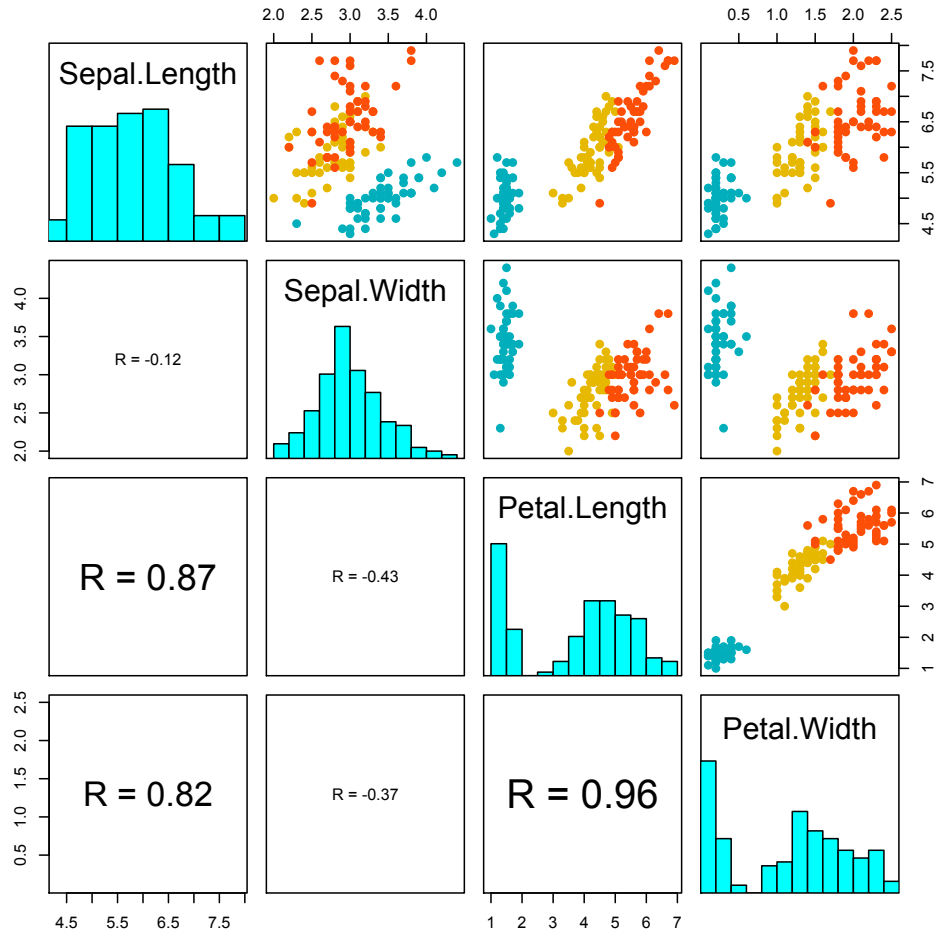


WHY THIS IS A PROBLEM?

- Inflate standard errors of partial regression coefficients.
- Partial regression coefficients change erratically by adding or removing a variable.

Multicollinearity refers to *individual predictors*
NOT
the reliability of the model as a whole.

MULTIPLE REGRESSION ASSUMPTIONS – MULTICOLLINEARITY



Testing for multicollinearity involves assessing the **redundancy of predictors** via ...

Pairwise correlations

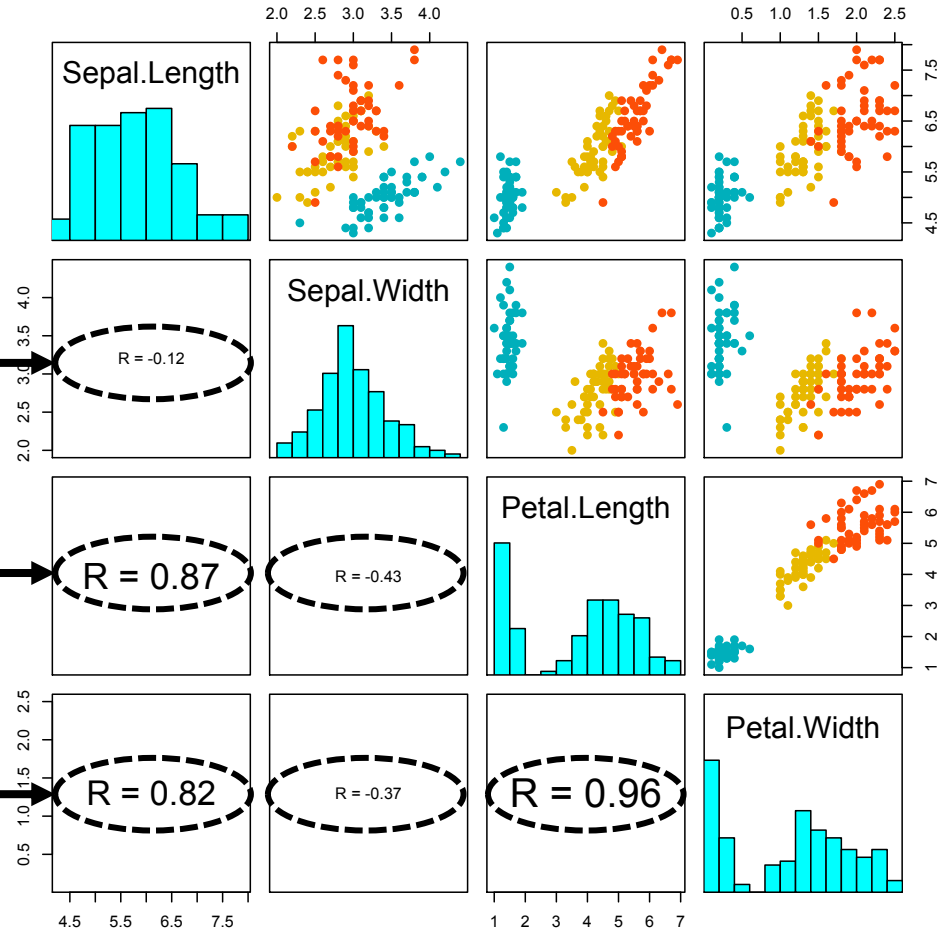
$X_i \sim X_j$

Tolerance / Variance inflation

$X_i \sim X_j + X_k + ... + X_n$

MULTIPLE REGRESSION ASSUMPTIONS – MULTICOLLINEARITY

Pairwise correlations



A rule of thumb is that values above 0.7 are strong correlations and hence imply strong Pairwise collinearity.

MULTIPLE REGRESSION ASSUMPTIONS – HOW TO ESTIMATE TOLERANCE?

- Is estimated by modelling each predictor as a function of all other predators.

$$X_i \sim X_j + X_k + ... + X_n$$

- The **Tolerance** for of each predictor is based on this regression and estimated as: $\mathbf{1 - R^2}$.

  Values below 0.1 are problematic.

- The inverse of the **Tolerance** is called **Variance inflation index**.

DEALING WITH COLLINEARITY

Simplest solution: Remove redundant variables.

Principal components regression: use the PCA-axes from the correlation matrix of the predictor variables as predictors ➔ Issue: ***How to read the partial regression coefficients?***

Ridge regression: small biasing constant is added to the normal equations that are solved to estimate the standardised regression coefficients ➔ **TO COMPLEX so avoid it.**

So far so good?

Any questions?

Ready to continue?

MODEL SELECTION – KEY TERMINOLOGY

Model	Interpretation
Saturated model	One parameter for every data point Fit: perfect Degrees of freedom: none Explanatory power of the model: none
Maximal model	Contains all (p) factors, interactions and covariates that might be of any interest. Many of the model's terms are likely to be insignificant Degrees of freedom: $n - p - 1$ Explanatory power of the model: it depends
Minimal adequate model	A simplified model with $0 \leq p' \leq p$ parameters Fit: less than the maximal model, but not significantly so Degrees of freedom: $n - p' - 1$ Explanatory power of the model: r^2 = SSR/SST
Null model	Just 1 parameter, the overall mean $\bar{y}$ Fit: none; SSE = SST Degrees of freedom: $n - 1$ Explanatory power of the model: none

DO I NEED THIS VARIABLE? - SIMPLE

```
Call:
lm(formula = Ozone ~ Solar.R + Wind + Temp, data = airquality)
Residuals:
    Min      1Q  Median      3Q     Max
-40.485 -14.219  -3.551  10.097  95.619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Solar.R       0.05982    0.02319   2.580  0.01124 *
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Temp          1.65209    0.25353   6.516 2.42e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6059, Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

Q1: Is predictor effect significant?

Only keep variables with significant effects

DO I NEED THIS VARIABLE? - COMPLEX

Step 1: Omit one variable (X_j) or a group of variables (X_j to X_n) from the model.

This is the reduced model, with n – 1 predictors.

Step 2: Compare the **reduced model** with a **full model**.

The **full model** has all the predictors.

You will use either:

A **log-likeihood** procedure to assess if there is information gain.

A **performance metric** to assess if adding a variable inprioved prediction performance

MODEL SELECTION

The goal is to:

*Define which combination of predictor variables (and their interactions) is the **best model**.*

Important principles:

- Occam's razor – *keep it simple (avoid overfitting)*
- Type I error – *chance goes up with multiple testing*
- Common sense – *model selection is an 'art'*

MODEL SELECTION
FORWARD PROCEDURE

Forward selection (addition)

- Start with only most significant predictor in the regression model (**lowest P**)
- Out of remaining predictors? → include the one which best explains residual variation (**lowest P**)
- Out of remaining predictors?... etc.
- **STOP** when no further predictors significant

MODEL SELECTION
BACKWARD PROCEDURE

Backward selection (elimination)

- Start with all predictors in the regression model
- Remove the least significant predictor (**highest P** in a partial F test), if non-significant
- remove next least significant predictor... etc.
- **STOP** when all remaining predictors significant

MODEL SELECTION
STEPWISE PROCEDURES

Stepwise selection (addition then elimination)

- Proceed as in forward selection ➔ **adding new predictors**
- Proceed as in forward Backward selection ➔ **remove predictors that are no longer significant**
- **STOP** when all predictors in model are significant and all those left out are not

So far so good?

Any questions?

Ready to continue?

IMPORTANCE OF PREDICTORS

How important is predictor X_j?

1. Look at P value from *t*-test or *F*- test
   (correct for multiple testing…?)

Only a qualitative assessment - ranking **NOT** how much will the response change by the increase of one unit in the predictor.

2. Look at the standardised effect size
   standardised partial regression coefficient:

Independent of scale of measurement

Now the coefficients is measured in units of SD

$$b_j^* = b_j \frac{s_{X_j}}{s_Y}$$

HOW GOOD IS MY MODEL?

Coefficient of multiple determination:

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

Proportion of the variance explained by regression model.

Multiple correlation coefficient:

$$r = \sqrt{r^2}$$

Between y and all x_j

Pearson's R

HOW GOOD IS MY MODEL?

Table 6.4 Criteria for selecting "best" fitting model in multiple linear regression. Formulae are for a specific model with p predictors included. Note that p excludes the intercept

Criterion	Formula
Adjusted r^2	$1 - \frac{SS_{Residual}/[n-(p+1)]}{SS_{Total}/(n-1)}$
Mallow's C_p	$\frac{\text{Reduced } SS_{Residual}}{\text{Full } MS_{Residual}} - [n - 2(p+1)]$
Akaike Information Criterion (AIC)	$n[\ln(SS_{Residual})] + 2(p+1) - n\ln(n)$
Schwarz Bayesian Information Criterion (BIC)	$n[\ln(SS_{Residual})] + (p+1)\ln(n) - n\ln(n)$

Adjusted R^2
Penalises the R^2 value based on the number of predictors

Mallow's C_p
Contrast based on the Sum of Squares between full and reduced model

AIC or BIC
Contrast the information in the full and reduced model

So far so good?

Any questions?

Ready to finish?

SUMMARY

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

The term "linear" refers to the combination of parameters, not the shape of the relationship.

- Always test of the assumptions to define if the linear model you build is "statistically correct".
- If you have multiple variables ➔ Check for **collinearity** and only retain meaningful variables.
- Always build the **Minimum Adequate Model (MAM)** based on a model selection procedure.
- For contrasting the "relative" effect of each variable use **standardised regression coefficients**.

AARHUS
UNIVERSITY