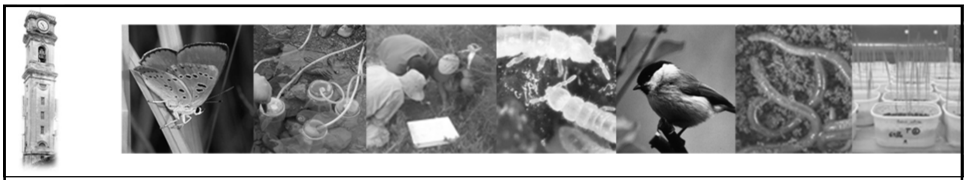


ANÁLISE DE DADOS ECOLÓGICOS

ISCED, Lubango, Março 2016

Soil Ecology & Ecotoxicology Lab

J. Paulo Sousa
Laboratory of Soil Ecology and Ecotoxicology
Centre for Functional Ecology
Universidade de Coimbra, Portugal
jps@zoo.uc.pt
http://cfe.uc.pt/paulosousa
http://www.facebook.com/labsolos

Tópicos do curso – Semana 1

1. **Revisão dos conceitos básicos em bioestatística**
   - testes de hipóteses para uma ou duas populações
   - análise de variância e desenho experimental
   - regressão linear simples e correlação
2. **Regressão linear múltipla**
   - exploração dos dados
   - avaliação de colinearidade entre variáveis explicativas
   - interação entre variáveis explicativas
   - interpretação dos resultados
3. **Modelos Lineares Generalizados**
   - GLM-Poisson
   - GLM-Logístico

Tópicos do curso – Semana 1

1. Revisão dos conceitos básicos em bioestatística
   - testes de hipóteses para uma ou duas populações
   - análise de variância e desenho experimental
   - regressão linear simples e correlação
2. Regressão linear múltipla
   - exploração dos dados
   - avaliação de colinearidade entre variáveis explicativas
   - interação entre variáveis explicativas
   - interpretação dos resultados
3. Modelos Lineares Generalizados
   - GLM-Poisson
   - GLM-Logístico

A mensagem para este bloco…

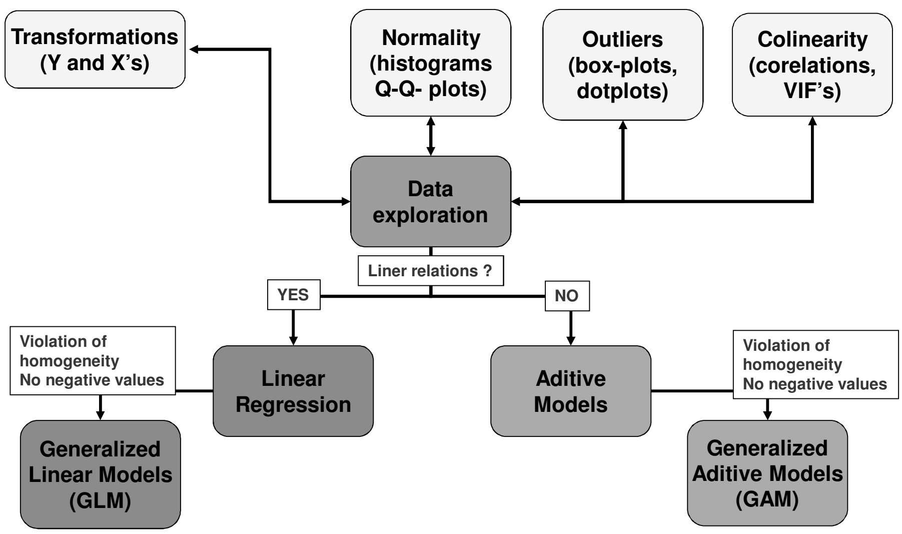
1. Gastem tempo a conhecer e a explorer os vossos dados…
2. Não confiem nos resultados após o primeiro clique…explorem vários modelos
3. *Sejam PACIENTES !*

Multiple regression revisited

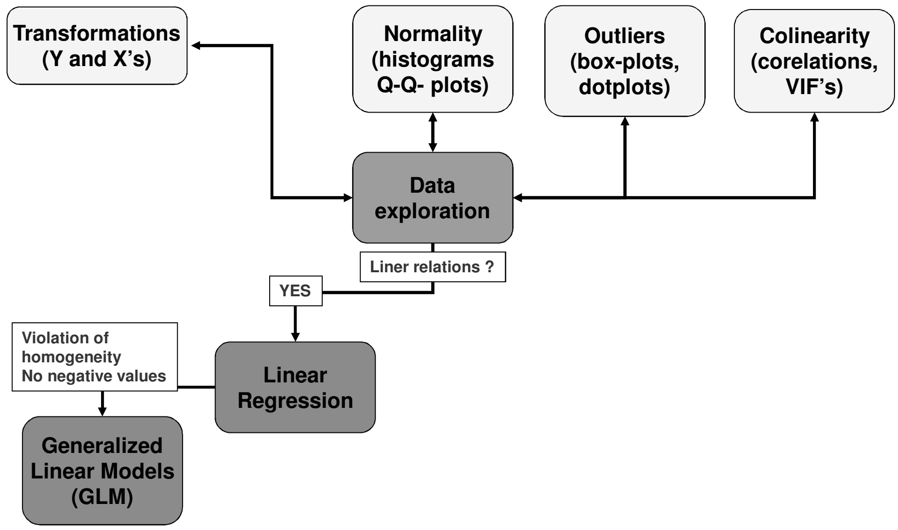
$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \varepsilon$$

1. **Check asumptions (normality and homoscedasticity) and transform data if necessary**
2. **Explore data regarding outliers, possible interactions between explanatory variables**
3. **Check for colinearity (tolerance, VIF values)**
4. **Perform regression and improve model according to the best fit (check significance values of βs, compare performance of models, check R2)**

Multiple regression & Co.



Generalized Linear Models



GLM (Poisson)

Models to use with non-normal BUT linear data:

- **Poisson Regression (count data, only positive values are possible)**

- **Linear relationship between response and explanatory variables is mantained via a Link function**

- **Poisson uses mostly the Loglink function:**

  $\mathrm{Log}(Y_i) = g(x)$;

  $g(x) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ (predictor function)

$$\boxed{Yi = e^{(a + b1X1 + b2X2 + \ldots + biXi)}}$$

GLM (Logistic) in Brodgar

Models to use with binary or proportion data:

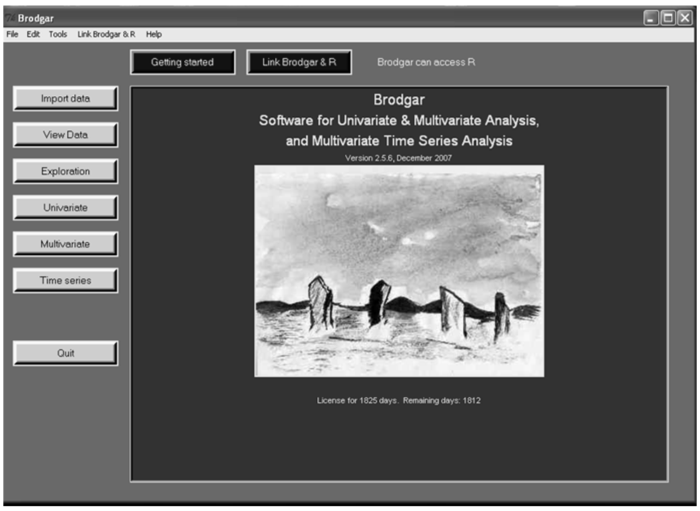
- **Logistic Regression (also only positive values are possible).**
- **Relationship between response and explanatory variables is mantained via a Link function**
- **Logistic uses mostly the Logit link function:**

  $Log(Y_i/(1-Y_i)) = g(x)$;

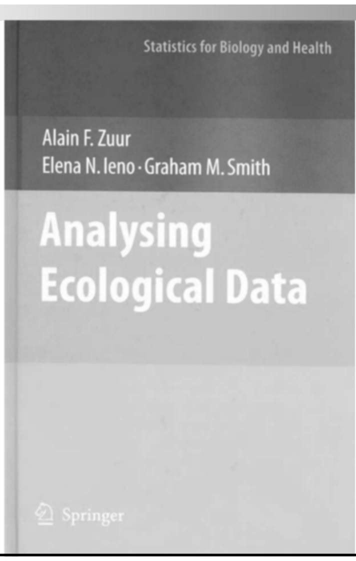
  $g(x) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ (predictor function)

$$Yi = e^{(a + b1X1 + b2X2 + \ldots + biXi)} / (1 + e^{(a + b1X1 + b2X2 + \ldots + biXi)})$$

SOFTWARE: Brodgar

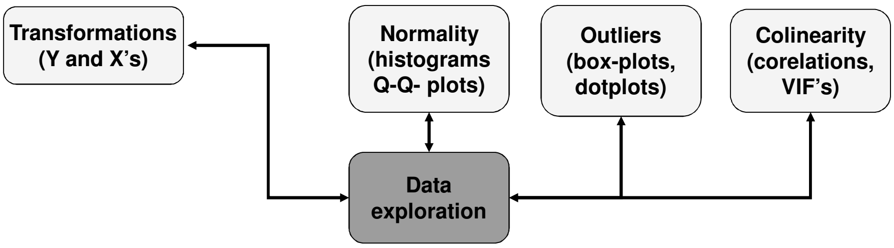


Key Reading



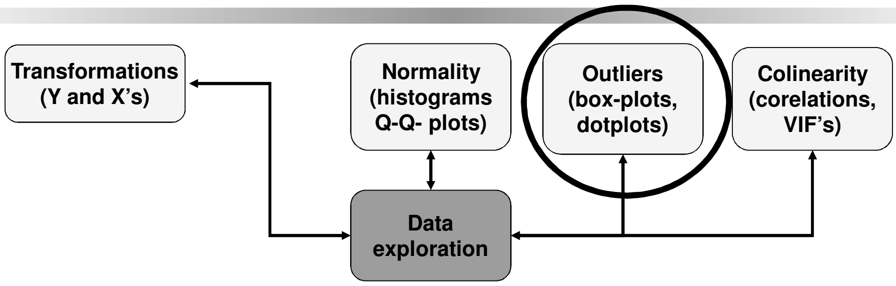
Zuur, A.F.; E.N. Ieno & G.M. Smith (2007). *Analysing Ecological Data*. Springer, NewYork, U.S.A.

Data Exploration



- Use dotplots to check data variation & outliers (transform data if necessary)
- Perform pair-plots to check for relations
- Calculate VIF's to check for colinearity
- Check for possible interactions (Coplots)
- Check normality via Q-Q- plots

Data Exploration – distribution & outliers

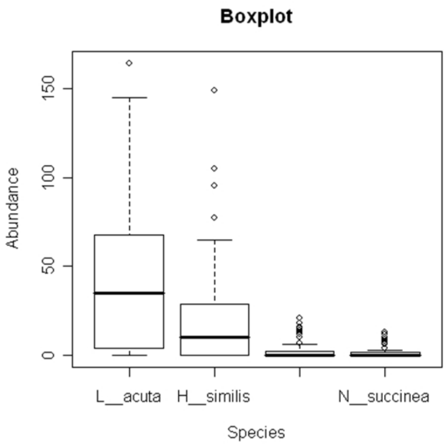


Box-plots vs. Dotplots

Boxplot: .../Data/Argentina.xls

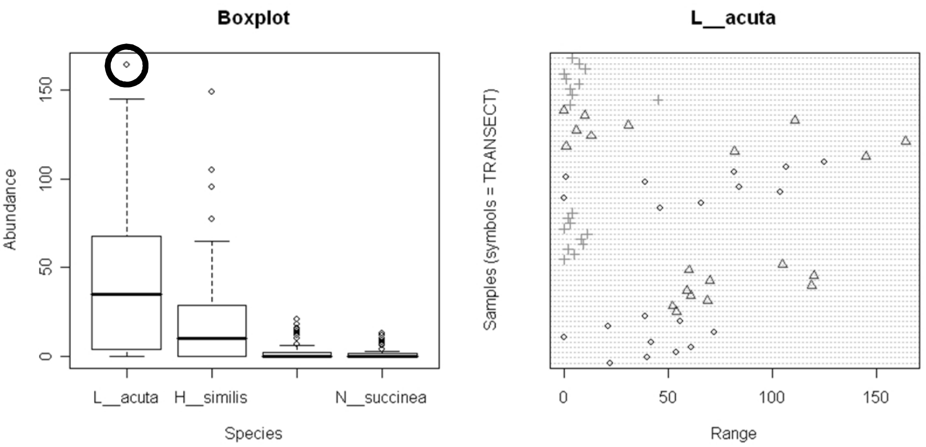
- **Abundance of 4 benthic spp; salt marsh in Argentina**
- **3 transects, each with 10 points**
- **Autumn (1) and Spring (2)**

Laeonereis acuta
Heteromastus similis
Uca uruguayensis
Neanthes succinea



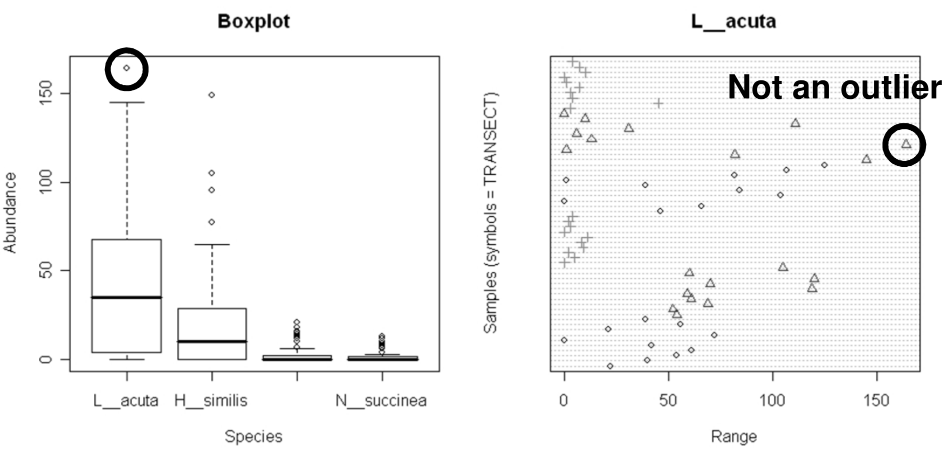
Dotplots

- Outliers? (by transept)



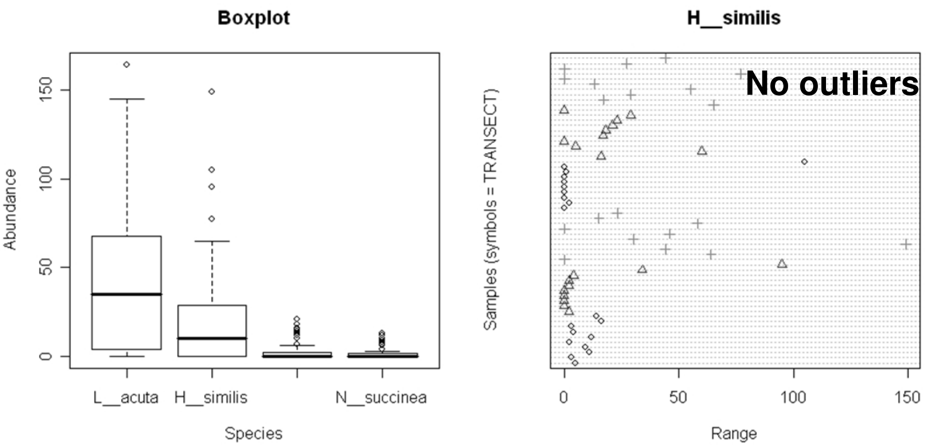
Dotplots

- Outliers?



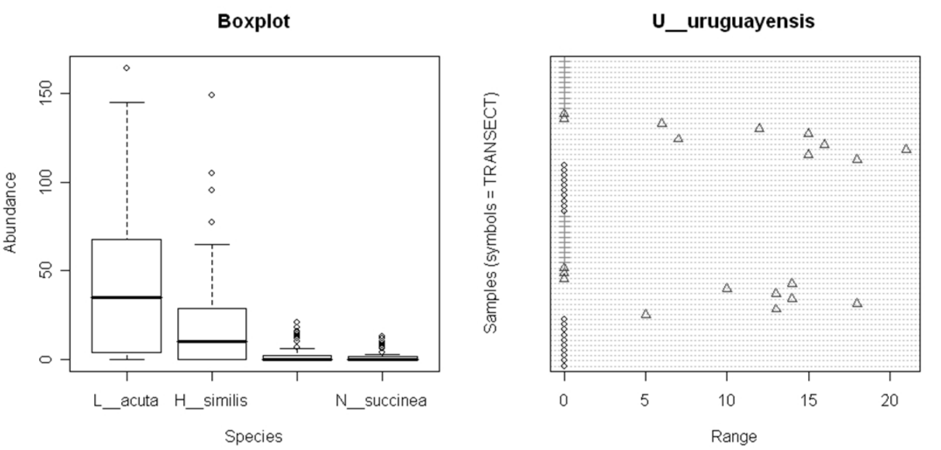
Dotplots

- Outliers?



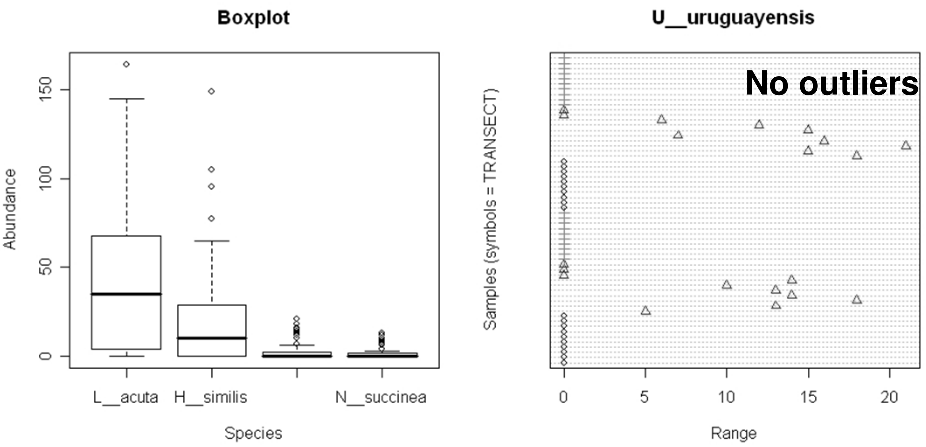
Dotplots

- Outliers?



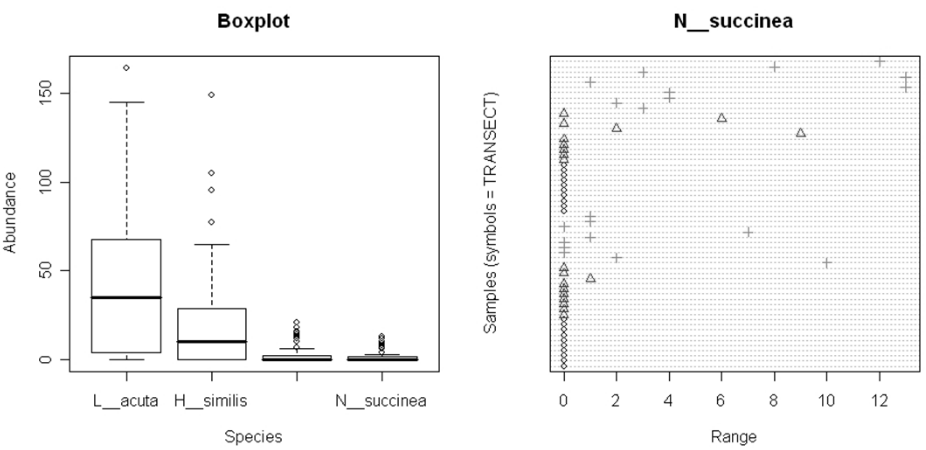
Dotplots

- Outliers?



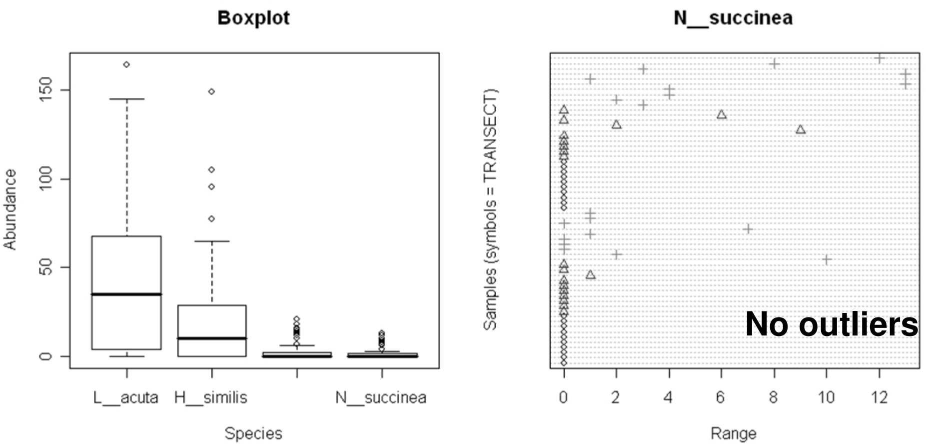
Dotplots

- Outliers?

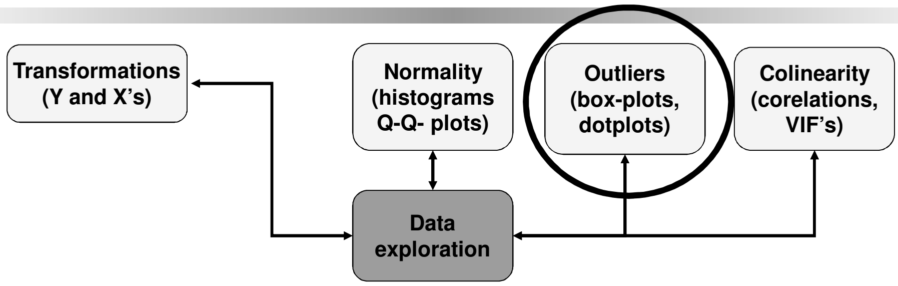


Dotplots

- Outliers?



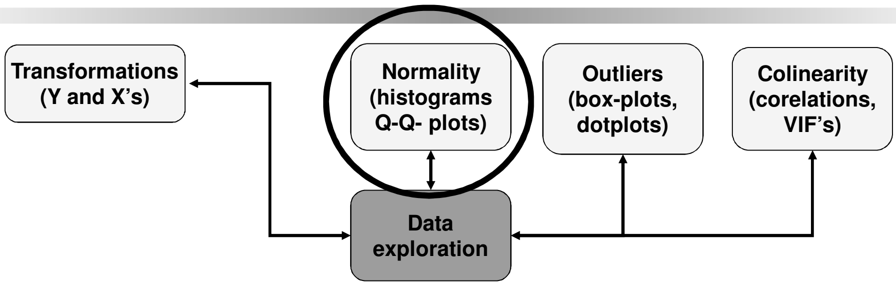
Data Exploration – distribution & outliers



Box-plots vs. Dotplots

- To check for outliers always compare results from both types of graphs.
- Box-plots only can misslead you

Data Exploration – distribution & outliers

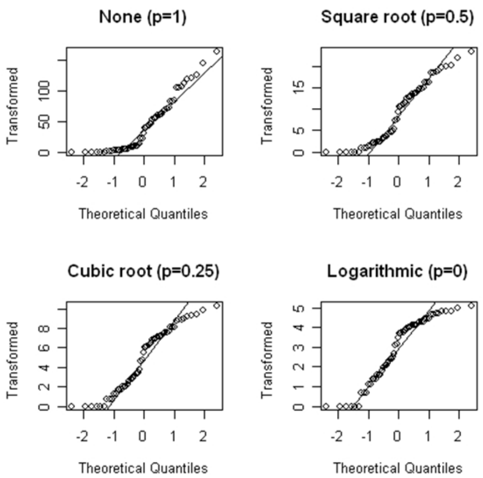


Histograms & Q-Q plots

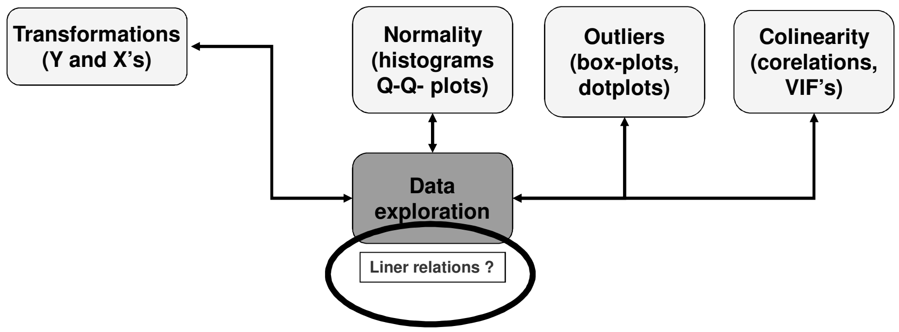
Data Exploration – normality

Q-Q Plots (Quantile-Quantile)

- **Normality**
  Q-Q plots compare the distribution of a given variable with a normal distribution
- **Abundance of *Laeonereis acuta***

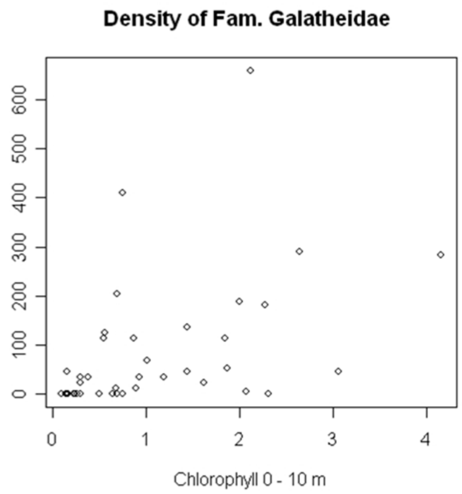


Data Exploration – linear relations?



Scatterplots: .../decapodNew.xls

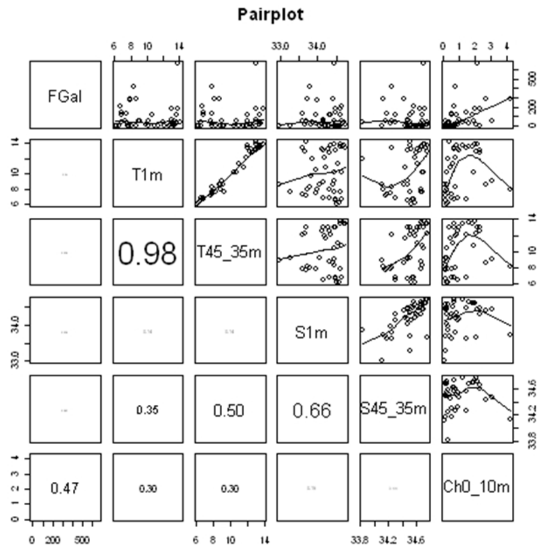
- Relationship BETWEEN variables – Relation? Linearity?
- Densities (families) planktonic decapods;
- Scotland
- 2 locations, 2 years;
- Temp. and salinity (1m and 35-45m) and chlorophyll (0-10m)



Pairplots: .../decapodNew.xls

- Relationships beetwen several variables
- Collinearity? (between explanatory variables)

(can reduce precision)



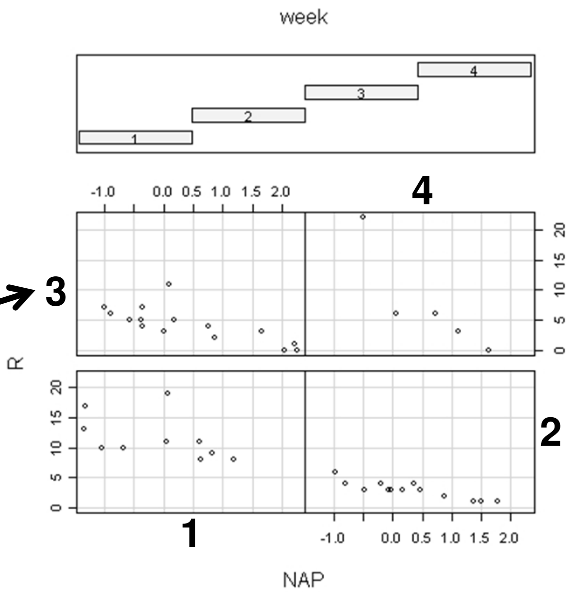
Coplots: .../Data/RIKZRichness.xls

- Coplot = Conditional scatterplot for several values of 1 or 2 other explanatory variables (nominal or continuous)

- RIKZ – Dutch governmental institute
- intertidal *benthos*
- 9 sandy beaches in The Netherlands
- 5 stations *per* beach (10 sub-replicates)
- 4 sampling times (4 sequential weeks)
- Station and beach slopes (“angles”)
- Exposure of the beach (waves, slope,…)
- Station NAP = reflects emersion time
- Salinity, temperature, grain size, organic matter,…

Coplots: /Data/RIKZRichness.xls

- **Specific richness *versus* NAP by week (nominal)**

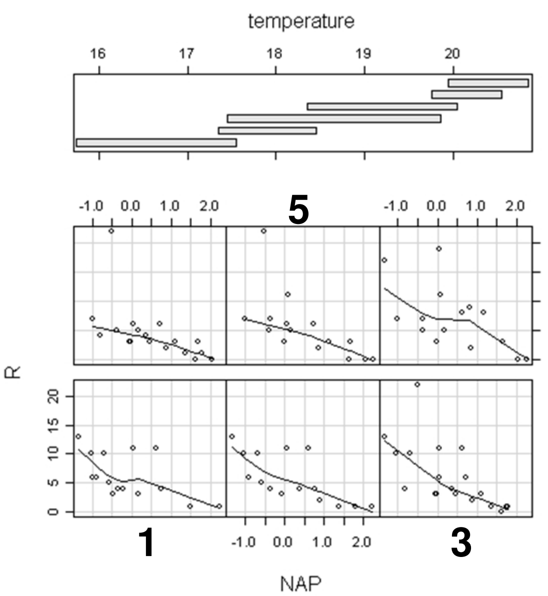


Week of sampling

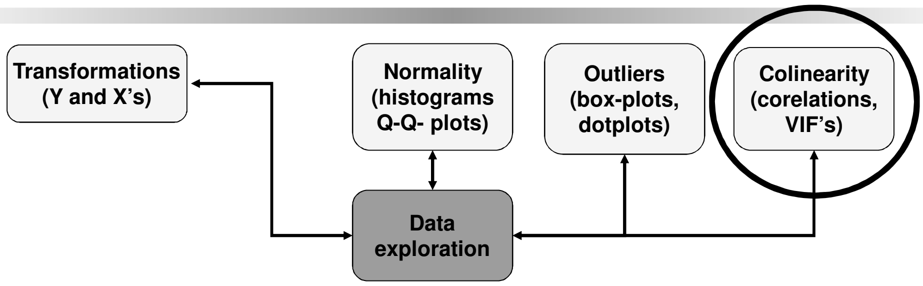
Also useful to check for interactions

Coplots: /Data/RIKZRichness.xls

- **Specific richness *versus* NAP by temperature (continuous)**
- **Some interception**
- ***Suggests to use temperature as an effect variable (or at least classes of temperature)***



Data Exploration – colinearity



VIF: .../Data/decapod.xls

- Variance Inflation Factor:

  $$VIF_i = 1 / (1- r_i^2)$$

  of the regression of each continuous explanatory variable with all others
  (TOLERANCE = 1/VIF)

- VIF: from 1 to ∞
  remove if VIF > 5 to 10

```
Correlations of the variables

         T1m T45_35m  S1m S45_35m Ch0_10m
T1m     1.000  0.983 0.142  0.4621  0.3147
T45_35m 0.983  1.000 0.139  0.4985  0.2956
S1m     0.142  0.139 1.000  0.6877  0.1355
S45_35m 0.462  0.498 0.688  1.0000  0.0532
Ch0_10m 0.315  0.296 0.136  0.0532  1.0000


Variance inflation factors

        GVIF
T1m     32.32
T45_35m 34.86
S1m      2.38
S45_35m  3.20
Ch0_10m  1.20
```


VIF: …/Data/decapod.xls

- **Variance Inflation Factor:**

  $$VIF_i = 1 / (1 - r_i^2)$$

  of the regression of each continuous explanatory variable with all others
  (TOLERANCE = 1/VIF)

- **VIF: from 1 to ∞**
  remove if VIF > 5 to 10
  REMOVE ONE OF THOSE

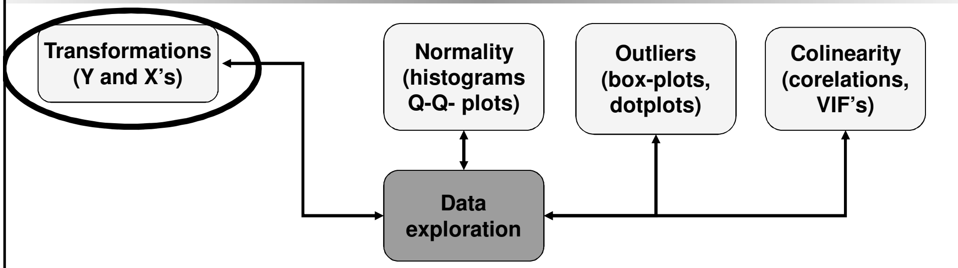
Correlations of the variables

	T1m	T45_35m	S1m	S45_35m	Ch0_10m
T1m	1.000	0.983	0.142	0.4621	0.3147
T45_35m	0.983	1.000	0.139	0.4985	0.2956
S1m	0.142	0.139	1.000	0.6877	0.1355
S45_35m	0.462	0.498	0.688	1.0000	0.0532
Ch0_10m	0.315	0.296	0.136	0.0532	1.0000

Variance inflation factors

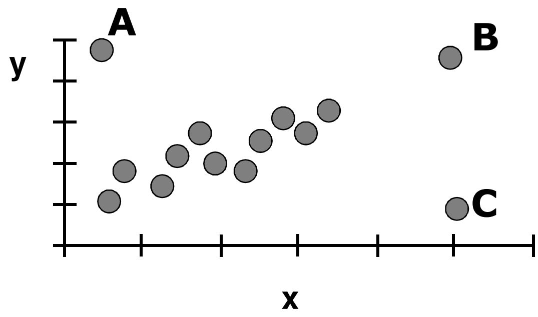
	GVIF
T1m	32.32
T45_35m	34.86
S1m	2.38
S45_35m	3.20
Ch0_10m	1.20

Data Exploration – Outliers and transformations



Outliers

- Outliers in the *x*-space, the *y*-space and the *xy*-space
- *A is what?*
- *B is what?*
- *C is what?*



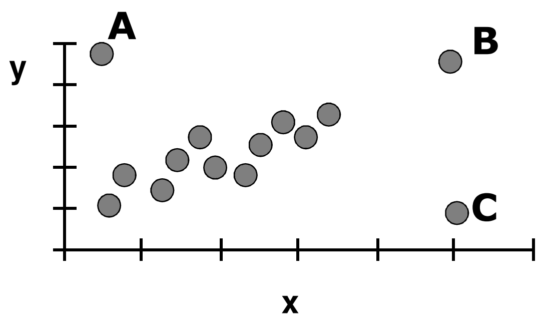
Outliers

- Outliers in the *x*-space, the *y*-space and the *xy*-space
- *A is what?*

Outlier in y space

Outlier in the xy space

- *B is what?*
- *C is what?*



Outliers

- Outliers in the *x*-space, the *y*-space and the *xy*-space
  - *A is what?*

Outlier in y space

Outlier in the xy space

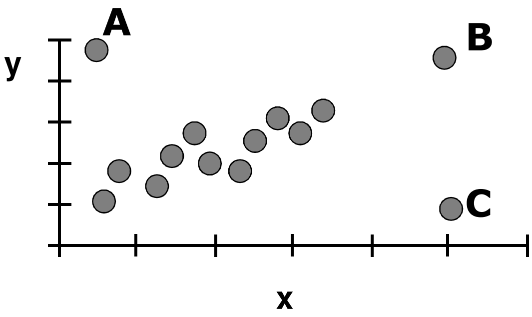
- *B is what?*

Outlier in the x space

Outlier in the y space

No outlier in xy space

- *C is what?*



Outliers

- Outliers in the *x*-space, the *y*-space and the *xy*-space
  - *A is what?*

Outlier in y space

Outlier in the xy space

- *B is what?*

Outlier in the x space

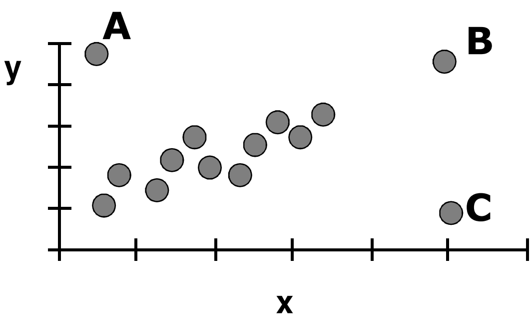
Outlier in the y space

No outlier in xy space

- *C is what?*

Outlier in the x space

Outlier in the xy space



Transformations

- Non normality, heterogeneity, outliers
- Non-linear relationships

- Can be applied to both response and explanatory variables
- Can be different to different variables in the same model

Transformations

- Logaritmic: log(y+1)

- Exponential: …$y^{1/3}$, $y^{1/2}$, y^2 , y^3 ,…

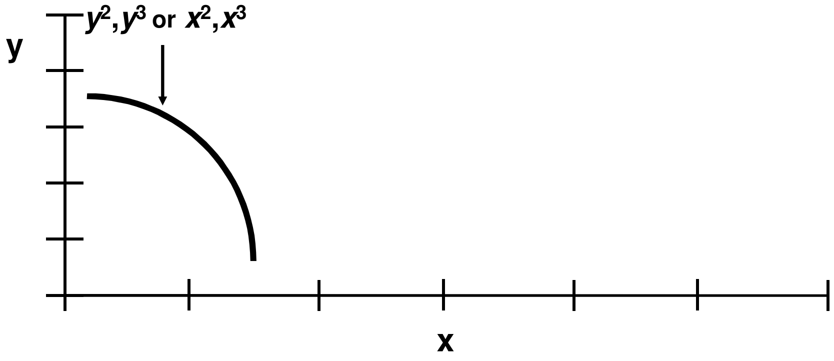
- Arcsin of square root (proportional data: 0 to 1)

- Ranking: 1st , 2nd, 3rd,… (more present *versus* less present)

- Transform to binary: 0 and 1 (presence/absence)

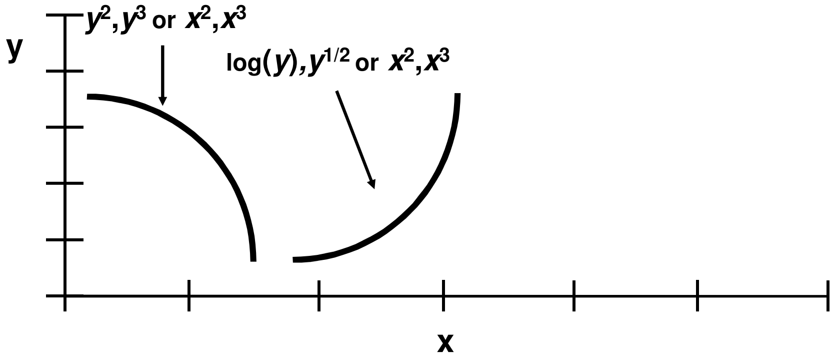
Transformations

- **Non linearity:**



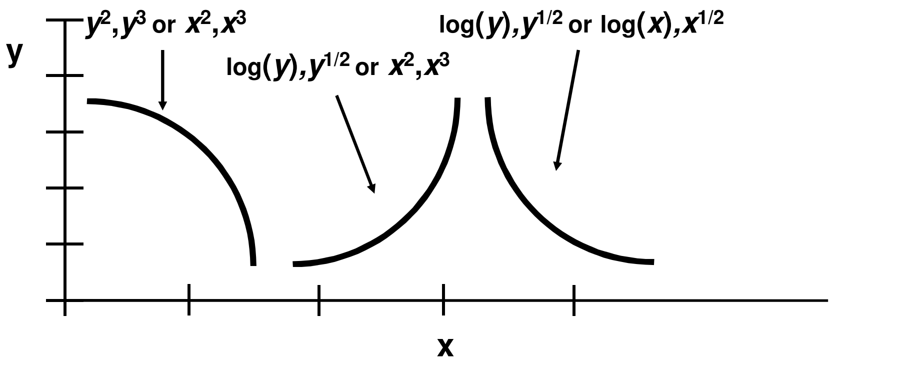
Transformations

- **Non linearity:**



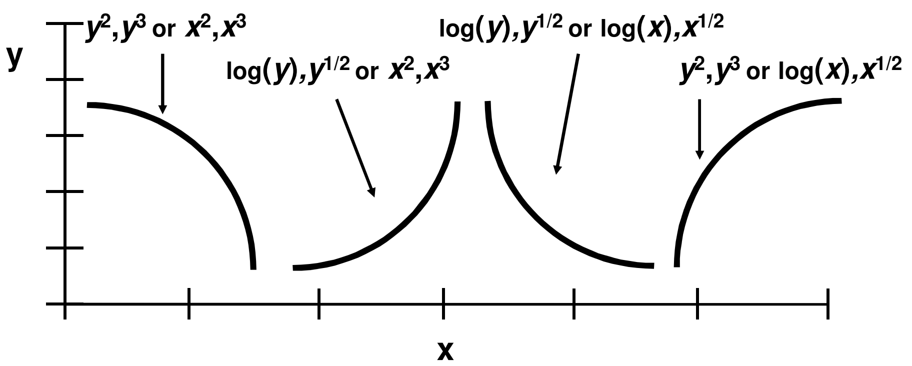
Transformations

- **Non linearity:**

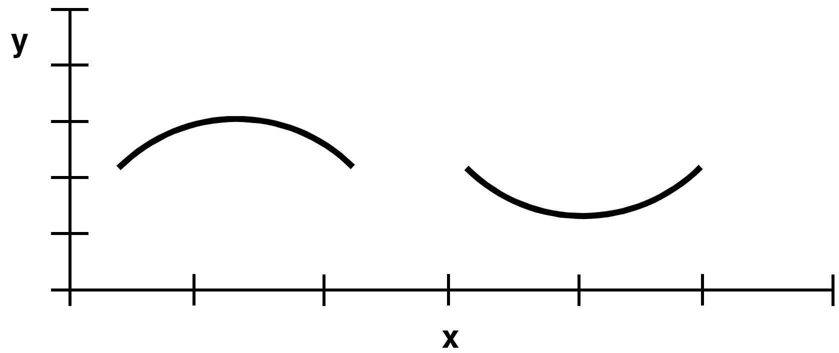


Transformations

- **Non linearity:**

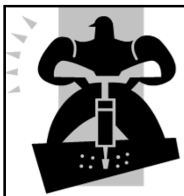


If…



- **… then quadratic regression:**

$$Y = a + bX + cX^2$$



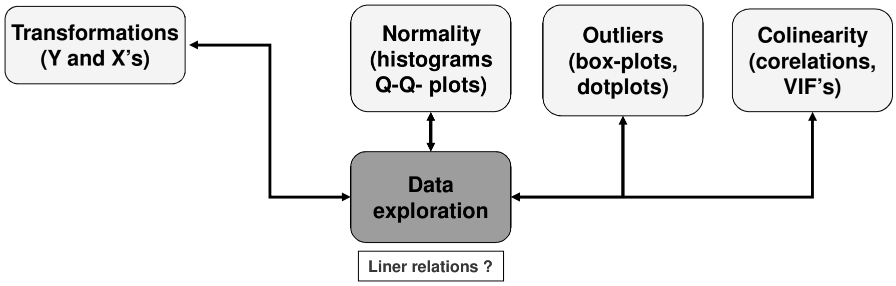
TASK for explorers… /Loyn.xls

The response variable ABUND is the density of birds in 56 forest patches (Australia).

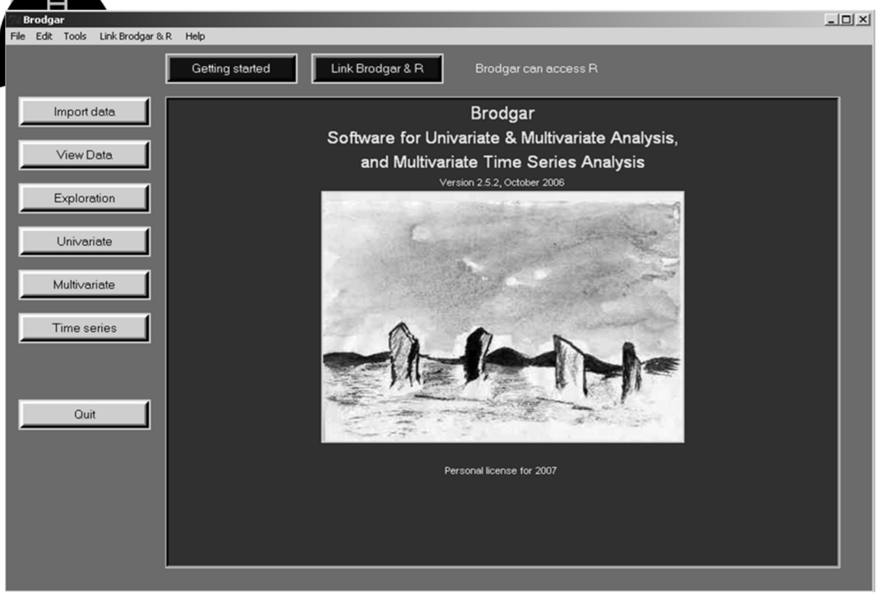
The explanatory variables are size of the forest patches (AREA), distance to the nearest forest patch (DIST), distance to the nearest larger forest patch (LDIST), year of isolation of the patch (YR.ISOL), agricultural grazing intensity at each patch (GRAZE = <u>nominal!</u>), and altitude (ALT).

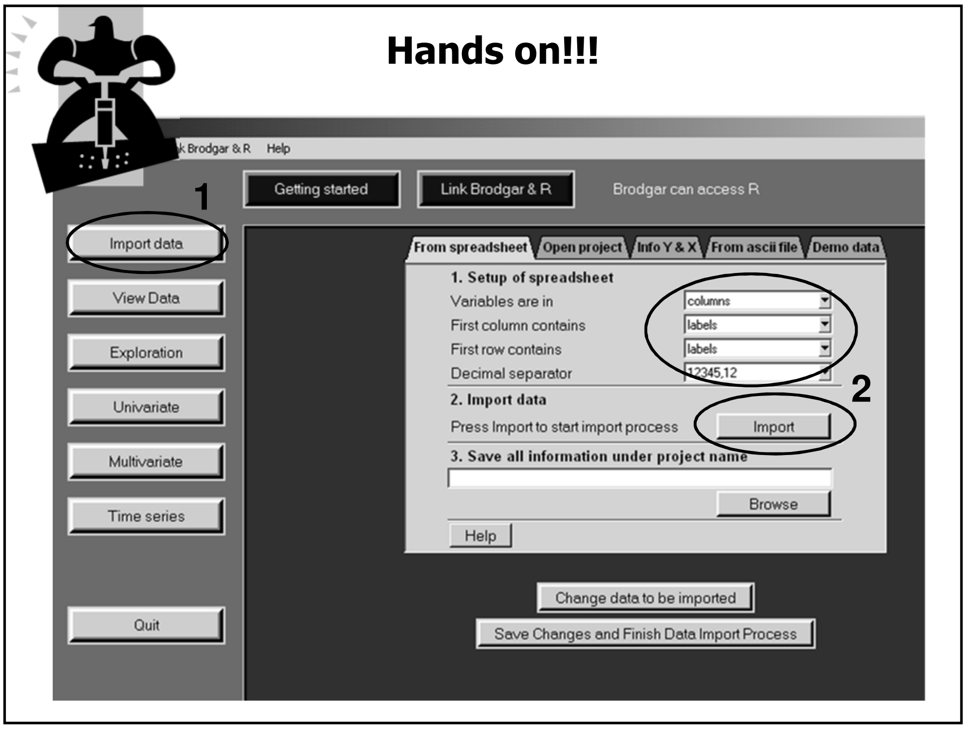
The underlying aim of the research is to find a relationship between bird densities and the explanatory variables....BUT NOW DATA EXPLORATION

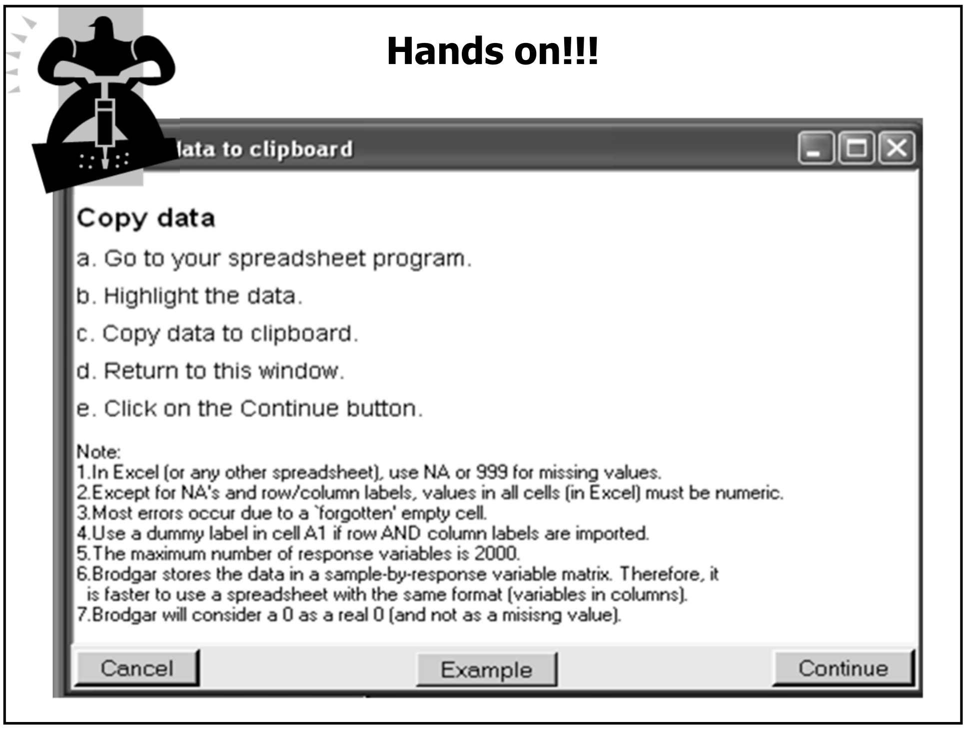
TASK for explorers… /Loyn.xls



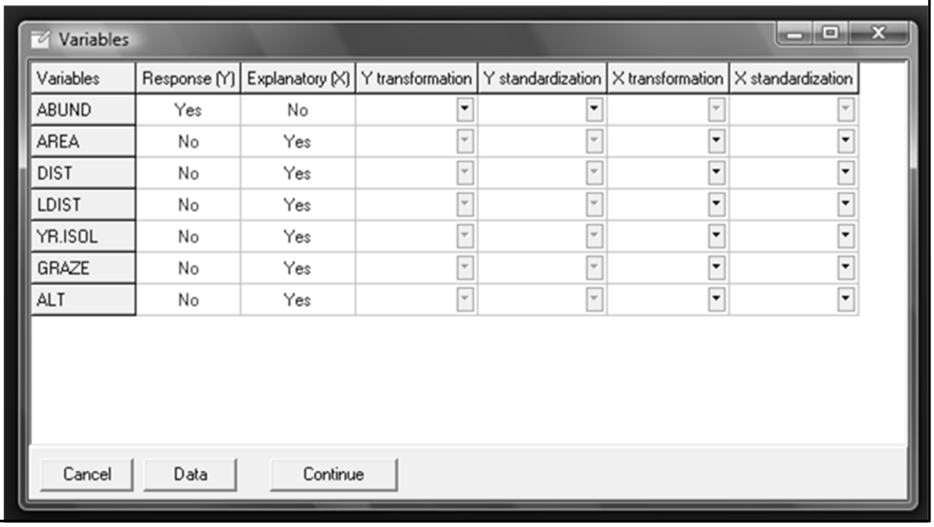
Hands on!!!



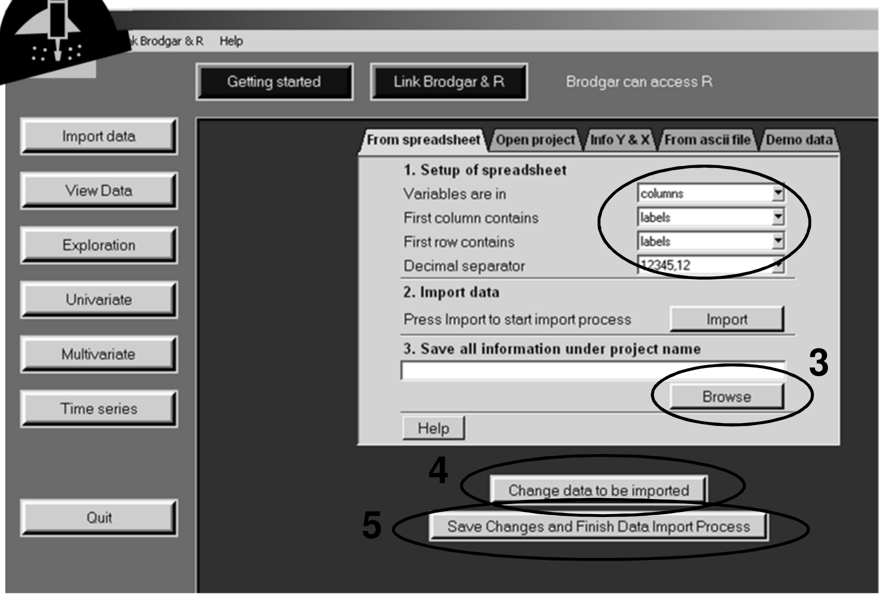

Hands on!!!
k Brodgar & R Help
Getting started
Link Brodgar & R
Brodgar can access R
1
Import data
View Data
Exploration
Univariate
Multivariate
Time series
Quit
From spreadsheet
Open project
Info Y & X
From ascii file
Demo data
1. Setup of spreadsheet
Variables are in columns
First column contains labels
First row contains labels
Decimal separator 12345,12
2
2. Import data
Press Import to start import process Import
3. Save all information under project name
Browse
Help
Change data to be imported
Save Changes and Finish Data Import Process


Hands on!!!
lata to clipboard
Copy data
a. Go to your spreadsheet program.
b. Highlight the data.
c. Copy data to clipboard.
d. Return to this window.
e. Click on the Continue button.
Note:
1.In Excel (or any other spreadsheet), use NA or 999 for missing values.
2.Except for NA's and row/column labels, values in all cells (in Excel) must be numeric.
3.Most errors occur due to a `forgotten' empty cell.
4.Use a dummy label in cell A1 if row AND column labels are imported.
5.The maximum number of response variables is 2000.
6.Brodgar stores the data in a sample-by-response variable matrix. Therefore, it is faster to use a spreadsheet with the same format (variables in columns).
7.Brodgar will consider a 0 as a real 0 (and not as a misisng value).
Cancel
Example
Continue

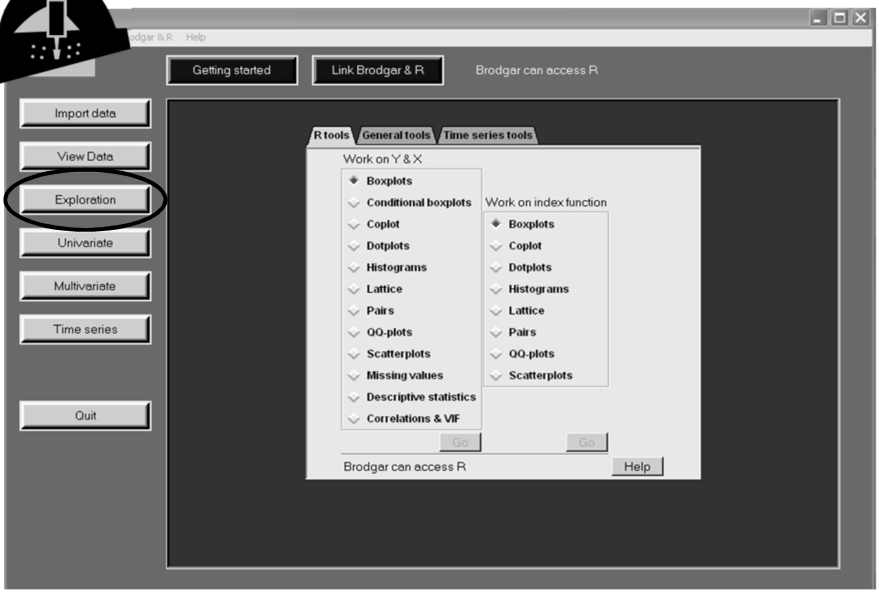
Hands on!!!



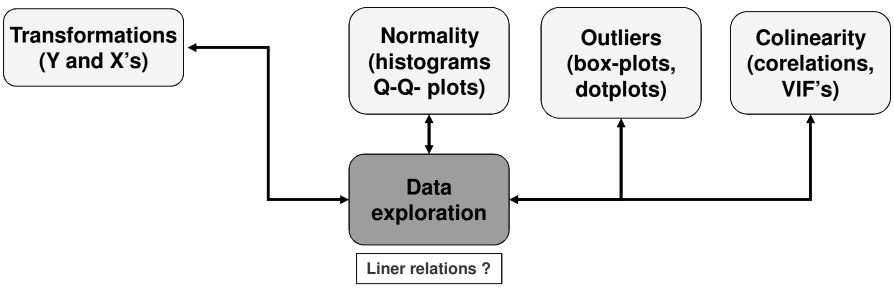
Hands on!!!



Hands on!!!



TASK for explorers… /Loyn.xls

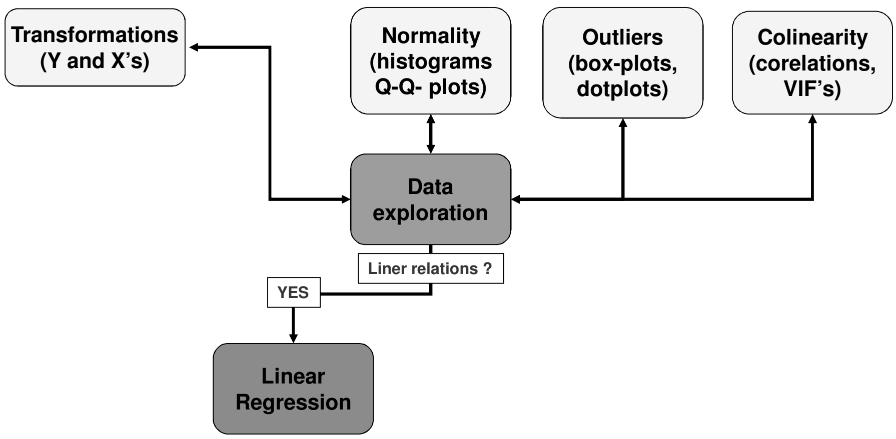


Always construct a log file with all your analyses and decisions

COURSE OUTLINE

1. Data Exploration
2. Linear Regression – Bivariate and Multiple
3. Generalised Linear Modelling
   Poisson
   Logistic

Bivariate & Multiple regression



Bivariate regression revisited

Population intercept -Y

Populatio n slope

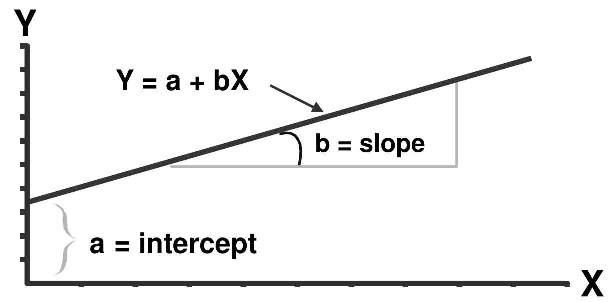
Rando m error

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Response variable (dependent)

Explanatory variable (independent)

Bivariate regression revisited

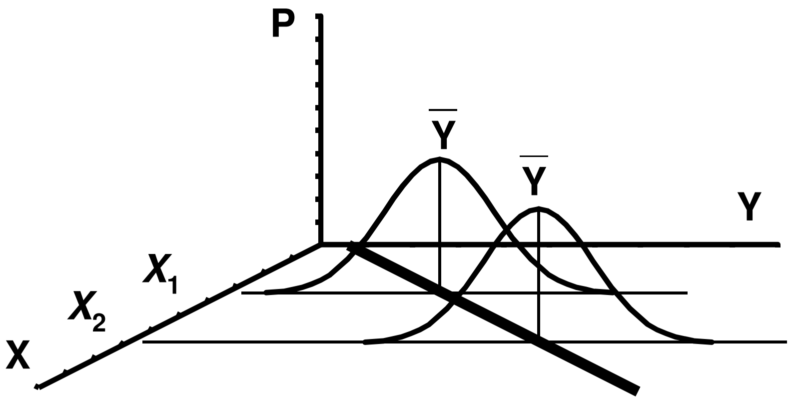


Bivariate regression-assumptions

- **Normality:** to each X_i, in the population, there is a serie of a normally distributed set of Y_i values
- **Homogeneity:** variances of these sets of Y_i values are equal
- **Independence:** Y values are independent of each other (sampling was made randomly and each sample was not influenced by previous or nearby sampling)
- **Fixed X:** X values are error-free

Bivariate regression-assumptions

- Normality & Homogeneity



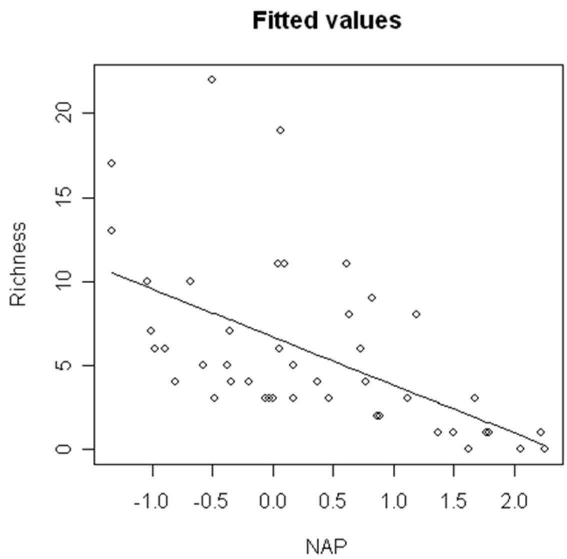
Normality & homogeneity

- **.../Data/RIKZ.xls**

- RIKZ – Dutch governmental institute
- intertidal *benthos*
- 9 sandy beaches in The Netherlands
- 5 stations *per* beach (10 sub-replicates)
- 4 sampling times (4 sequential weeks)
- Station and beach slopes (“angles”)
- Exposure of the beach (waves, slope,…)
- Station NAP = reflects emersion time (high NAP low immersion)
- Salinity, temperature, grain size, organic matter,…

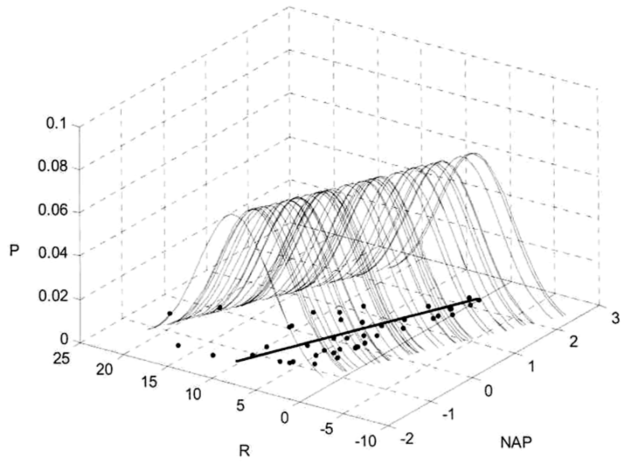
Bivariate linear regression

- Richness *versus* NAP



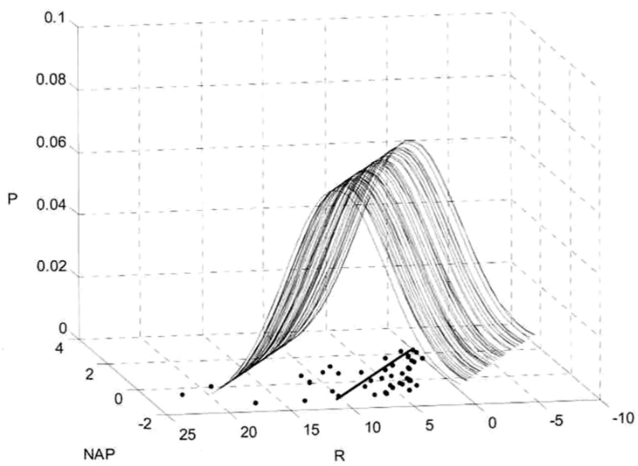
Normality & homogeneity

- **Richness**
  ***versus* NAP**



Normality & homogeneity

- **Richness**
  ***versus* NAP**



Variance components

Notation	Variance in	Sum of squared deviations of	Formula
SS_{total}	Y	Observed data from the mean	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
$SS_{regression}$	Y explained by X	Fitted values from the mean value	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
$SS_{residual}$	Y not explained by X	Observed values from fitted values	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

ANOVA

Source of variation	SS	df	MS	Expected MS
Regression	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	1	$\sum_{i=1}^{n}\frac{(\hat{Y}_i - \overline{Y})^2}{1}$	$\sigma_\varepsilon^2 + \beta^2\sum_{i=1}^{n}(X_i - \overline{X})^2$
Residual	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$		

$F = MS_{regression} / MS_{residual}$

ANOVA RIKZ: R *versus* NAP

Table 5.3. ANOVA table for the RIKZ data.

	df	SS	MS	F-value	P(>F)
NAP	1	357.53	357.53	20.66	<0.001
residuals	43	744.12	17.31		

Linear regression

- **Richness *versus* NAP**

```
############################################
#### LINEAR REGRESSION NUMERICAL OUTPUT ###
############################################

Model is given by f1:
Y1 ~ NAP

Residuals:
    Min      1Q  Median      3Q     Max
-5.0675 -2.7607 -0.8029  1.3534 13.8723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6857     0.6578  10.164 5.25e-13 ***
NAP          -2.8669     0.6307  -4.545 4.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.16 on 43 degrees of freedom
Multiple R-Squared: 0.3245,    Adjusted R-squared: 0.3088
F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05

Analysis of Variance Table

Response: Y1
          Df Sum Sq Mean Sq F value    Pr(>F)
NAP        1 357.53  357.53  20.660 4.418e-05 ***
Residuals 43 744.12   17.31
```

Model validation

r^2 = $SS_{regression}$ / SS_{total}

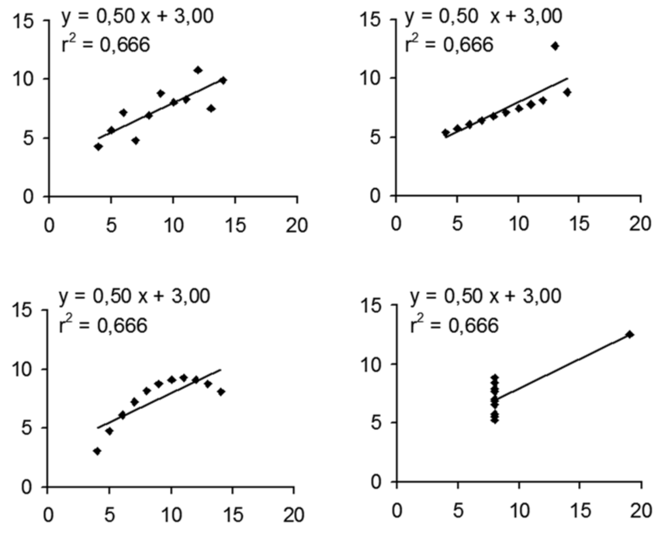
= 1 – ($SS_{residual}$ / SS_{total})

```
Residual standard error: 4.16 on 43 degrees of freedom
Multiple R-Squared: 0.3245,    Adjusted R-squared: 0.3088
F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05
```

$$r_{adj}^2 = 1 - \left[(1 - r^2) \times \frac{n-1}{n-m-1} \right]$$

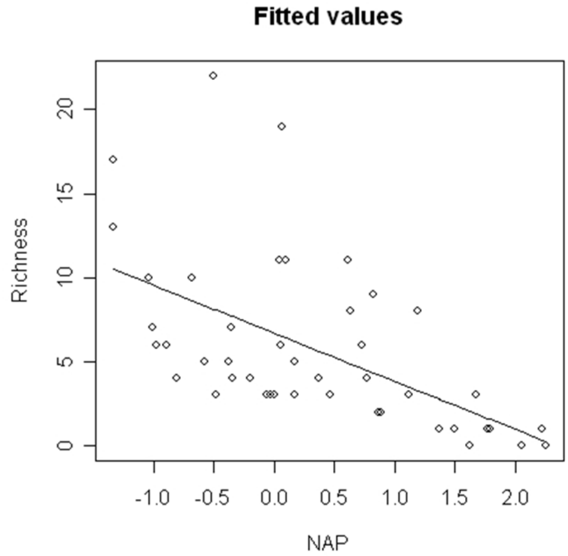
- Sample size (*n*)
- Number of explanatory variables (*m*)

Anscombe quartet



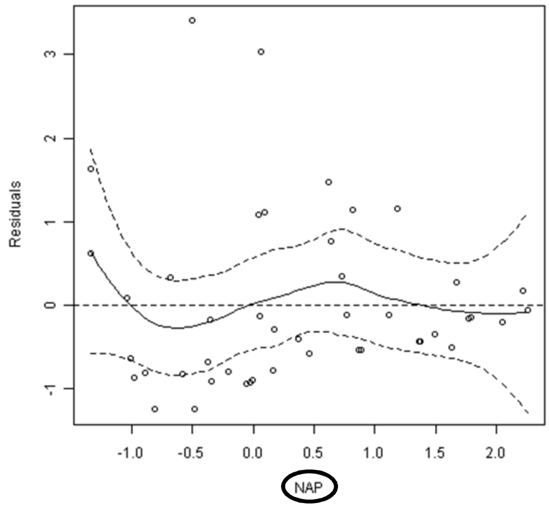
Assessing assumptions

- Richness *versus* NAP



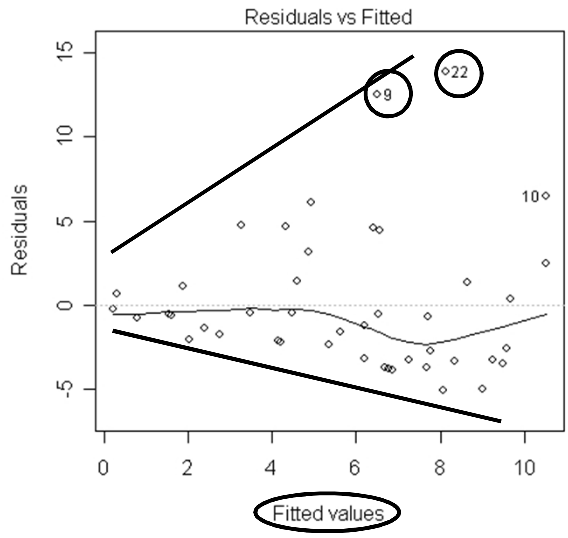
Assessing assumptions

- Residuals



Assessing assumptions

- **Residuals**



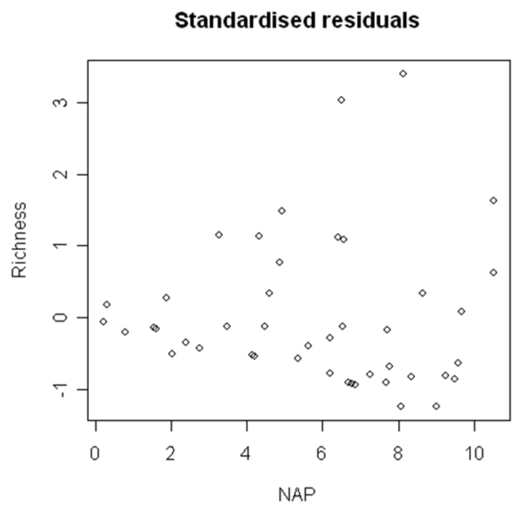
Assessing assumptions

Standardised residuals:

mean = 0,

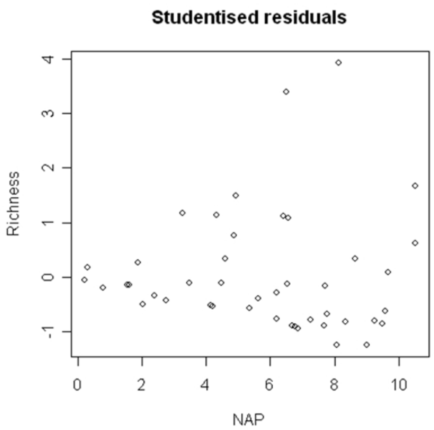
variance = 1

(if > 2 then… *outlier*?)



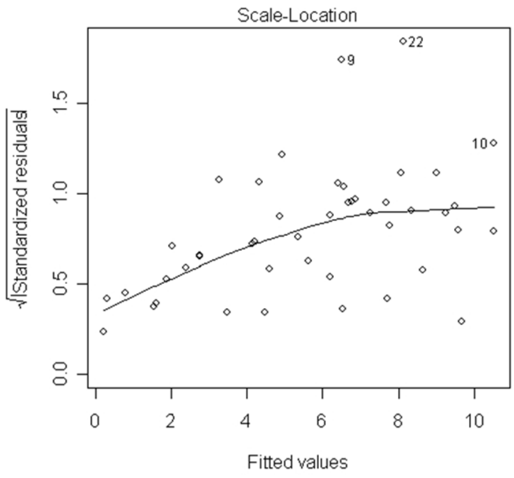
Assessing assumptions

Studentised residuals: same as standardised residuals but removing the respective observation (better visualisation)



Assessing assumptions

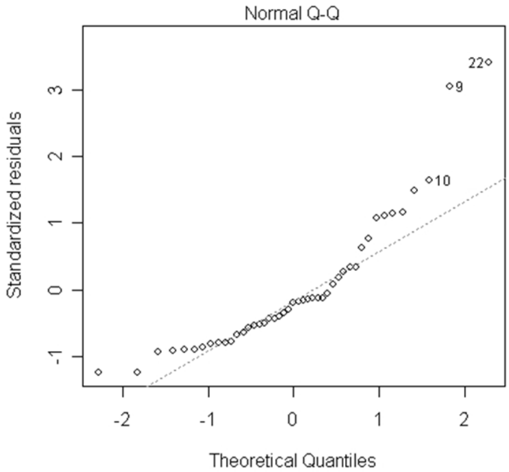
Square root of standardised residuals (better visualisation)



Assessing assumptions

Standardised residuals:

mean = 0, variance = 1 (if > 2 then… outlier?)

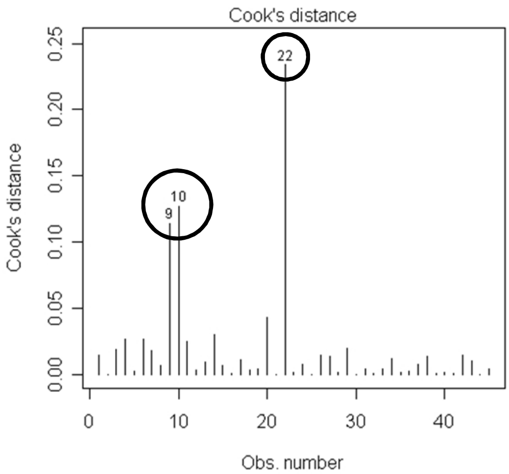


Outliers

Cook's distance:

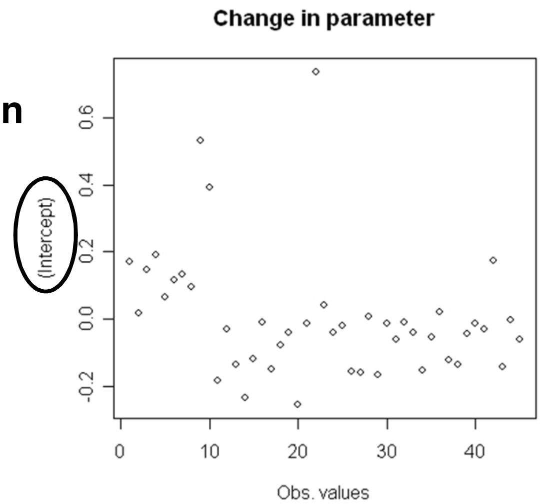
- **Influential points**

- **Outlier if >1 (remove it?)**



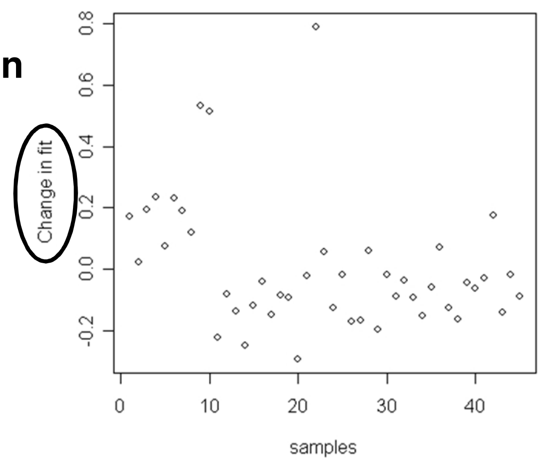
Outliers

Jackknife method:
change in regression
If removing
each observation
(Influential points!)



Outliers

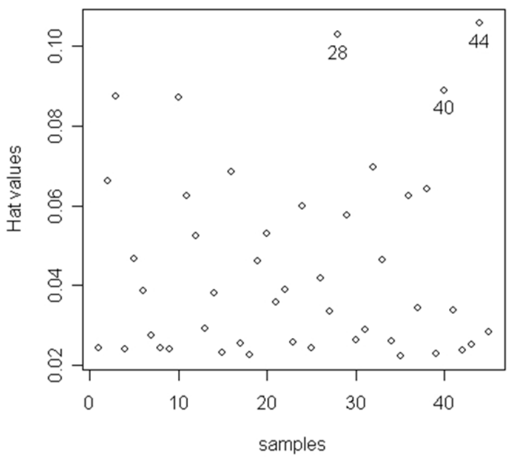
Jackknife method:
change in regression
If removing
each observation
(Influential points!)



Hat values

Hat values (high leverage):

- **Extreme values in *x* space**



Cook distance and hat values

Summarising,

leverage identifies extreme observations

And the Cook's distance detects points that are influential.

It is easier to justify omitting influential points (large Cook's distance) if they are extreme observations (these are points with a large leverage).

Final model presentation

$$R = 6.69(\pm 0.66) - 2.87(\pm 0.63)NAP$$
$$r^2 = 32.45\% \ (n = 45)$$

TASK … /Loyn.xls

Apply bivariate linear regression to model bird abundance as a function of AREA.

- **What is the fitted model?**
- **Are the parameters significant? Use two ways to assess this.**
- **How much variation do you explain?**
- **Apply a model validation; check all assumptions. Are there patterns in the residuals? Do you have normality and homogeneity?**
- **How many birds do you expect if AREA is 100?**

Multiple regression revisited

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \varepsilon$$

1. **Check asumptions (normality and homoscedasticity) and transform data if necessary**
2. **Explore data regarding outliers, possible interactions between explanatory variables**
3. **Check for colinearity (tolerance, VIF values)**
4. **Perform regression and improve model according to the best fit (check significance values of βs, compare performance of models, check R2)**

Multiple linear regression

RIKZ
Richness
versus
angle2
(beach slope)
NAP
grainsize
humus
week (nominal)

RIKZ
Richness
versus
angle2
(beach slope)
NAP
grainsize
humus
week (nominal)

```
Model is given by f1:
Y1 ~ 1 + angle2 + NAP + grainsize + humus + as.factor(week)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0454 -1.2865 -0.3314  0.7048 12.0917

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.298448   7.967002   1.167 0.250629
angle2            0.016760   0.042934   0.390 0.698496
NAP              -2.274093   0.529411  -4.296 0.000121 ***
grainsize         0.002249   0.021066   0.107 0.915570
humus             0.519686   8.703910   0.060 0.952710
as.factor(week)2 -7.065098   1.761492  -4.011 0.000282 ***
as.factor(week)3 -5.719055   1.827616  -3.129 0.003411 **
as.factor(week)4 -1.481816   2.720089  -0.545 0.589182
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.092 on 37 degrees of freedom
Multiple R-Squared: 0.679,      Adjusted R-squared: 0.6182
F-statistic: 11.18 on 7 and 37 DF,  p-value: 1.664e-07
```

week (nominal):
without week1
only:
week2, 3 e 4

To each level of
week
X =0 or X=1

```
Model is given by f1:
Y1 ~ 1 + angle2 + NAP + grainsize + humus + as.factor(week)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0454 -1.2865 -0.3314  0.7048 12.0917

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.298448   7.967002   1.167 0.250629
angle2            0.016760   0.042934   0.390 0.698496
NAP              -2.274093   0.529411  -4.296 0.000121 ***
grainsize         0.002249   0.021066   0.107 0.915570
humus             0.519686   8.703910   0.060 0.952710
as.factor(week)2 -7.065098   1.761492  -4.011 0.000282 ***
as.factor(week)3 -5.719055   1.827616  -3.129 0.003411 **
as.factor(week)4 -1.481816   2.720089  -0.545 0.589182
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.092 on 37 degrees of freedom
Multiple R-Squared: 0.679,      Adjusted R-squared: 0.6182
F-statistic: 11.18 on 7 and 37 DF,  p-value: 1.664e-07
```

Model selection

- Criteria: r^2_{adj} and AIC (Akaike Information Criteria)

$$AIC = n\log(SS_{residual}) + 2(m+1) - n\log(n)$$

- *Backwards* (removing…)

```
###############################################
####   OUTPUT OF SELECTION PROCEDURE   ####
The selection procedure cannot cope with missing values.
If you have missing values, re-apply linear regression but
remove rows with missing values.
Start:  AIC=108.78
Y1 ~ 1 + angle2 + NAP + grainsize + humus + as.factor(week)

                 Df Sum of Sq    RSS    AIC
- humus           1      0.03 353.70 106.78
- grainsize       1      0.11 353.77 106.79
- angle2          1      1.46 355.12 106.96
<none>                        353.66 108.78
- as.factor(week) 3    177.51 531.17 121.08
- NAP             1    176.37 530.03 124.98
```

Model selection

```
Start:  AIC=108.78
Y1 ~ 1 + angle2 + NAP + grainsize + humus + as.factor(week)

                 Df Sum of Sq    RSS    AIC
- humus           1      0.03 353.70 106.78
- grainsize       1      0.11 353.77 106.79
- angle2          1      1.46 355.12 106.96
<none>                        353.66 108.78
- as.factor(week) 3    177.51 531.17 121.08
- NAP             1    176.37 530.03 124.98

Step:  AIC=106.78
Y1 ~ angle2 + NAP + grainsize + as.factor(week)

                 Df Sum of Sq    RSS    AIC
- grainsize       1      0.12 353.82 104.80
- angle2          1      1.55 355.24 104.98
<none>                        353.70 106.78
- as.factor(week) 3    197.00 550.70 120.70
- NAP             1    180.31 534.01 123.32
```

Model selection

```
Step:  AIC=104.8
Y1 ~ angle2 + NAP + as.factor(week)

                  Df Sum of Sq    RSS    AIC
- angle2           1      3.19 357.00 103.20
<none>                         353.82 104.80
- NAP              1    213.45 567.26 124.04
- as.factor(week)  3    303.64 657.46 126.68

Step:  AIC=103.2
Y1 ~ NAP + as.factor(week)

                  Df Sum of Sq    RSS    AIC
<none>                         357.00 103.20
- NAP              1    210.33 567.33 122.04
- as.factor(week)  3    387.11 744.12 130.25
```

Model selection

```
Step:  AIC=103.2
Y1 ~ NAP + as.factor(week)

                  Df Sum of Sq    RSS    AIC
<none>                         357.00 103.20
- NAP              1    210.33 567.33 122.04
- as.factor(week)  3    387.11 744.12 130.25


Call:
lm(formula = Y1 ~ NAP + as.factor(week), data = datazz, weights = XW,     na.action = na.omit)

Coefficients:
    (Intercept)            NAP  as.factor(week)2  as.factor(week)3  as.factor(week)4
         11.368         -2.271            -7.625            -6.178            -2.594
```

Significance

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.3677     0.9459  12.017 7.48e-15 ***
NAP                -2.2708     0.4678  -4.854 1.88e-05 ***
as.factor(week)2   -7.6251     1.2491  -6.105 3.37e-07 ***
as.factor(week)3   -6.1780     1.2453  -4.961 1.34e-05 ***
as.factor(week)4   -2.5943     1.6694  -1.554    0.128
```

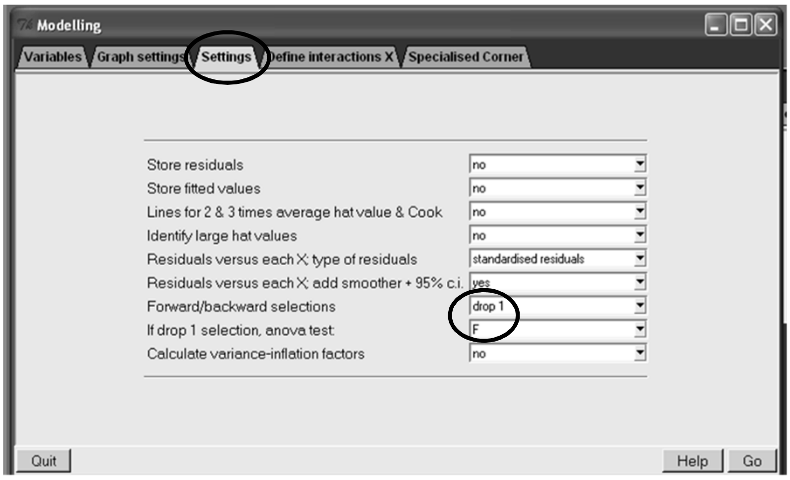
ANOVA

- **Each line of the ANOVA table is compared with the previous line:**
  - **Model 1:** $Y_i = \alpha + \varepsilon_i$
  - **Model 2:** $Y_i = \alpha + \beta_1 X_1 + \varepsilon_i$
  - **Model 3:** $Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$
  - **A.s.o.**

```
Model:
Y1 ~ 1 + NAP + as.factor(week)
                Df Sum of Sq    RSS    AIC F value    Pr(F)
<none>                       357.00 103.20
NAP              1    210.33 567.33 122.04  23.566 1.880e-05 ***
as.factor(week)  3    387.11 744.12 130.25  14.458 1.581e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA

- **drop 1 variable and F – ANOVA ok!**



ANOVA

- **drop 1 variable**
- **F**

```
Single term deletions

Model:
Y1 ~ 1 + NAP + as.factor(week)
                Df Sum of Sq    RSS    AIC F value     Pr(F)
<none>                       357.00 103.20
NAP              1    210.33 567.33 122.04  23.566 1.880e-05 ***
as.factor(week)  3    387.11 744.12 130.25  14.458 1.581e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

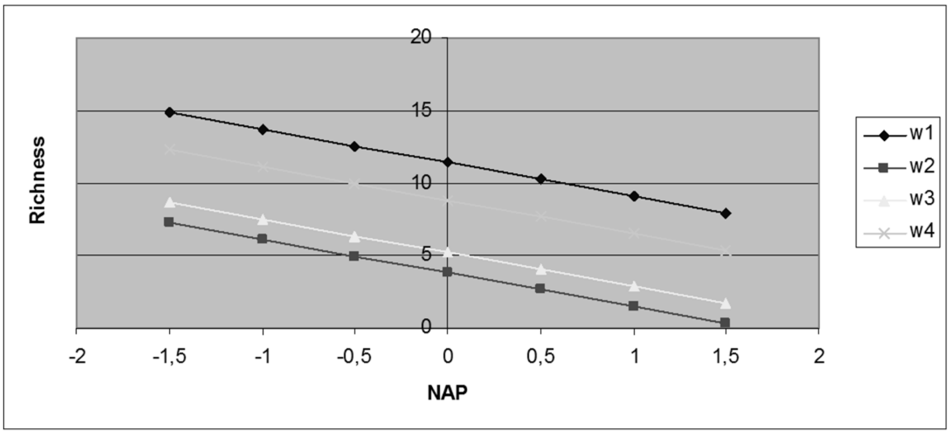
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.3677     0.9459  12.017 7.48e-15 ***
NAP                -2.2708     0.4678  -4.854 1.88e-05 ***
as.factor(week)2   -7.6251     1.2491  -6.105 3.37e-07 ***
as.factor(week)3   -6.1780     1.2453  -4.961 1.34e-05 ***
as.factor(week)4   -2.5943     1.6694  -1.554    0.128

Single term deletions

Model:
Y1 ~ 1 + NAP + as.factor(week)
                Df Sum of Sq    RSS    AIC F value     Pr(F)
<none>                        357.00 103.20
NAP              1    210.33 567.33 122.04  23.566 1.880e-05 ***
as.factor(week)  3    387.11 744.12 130.25  14.458 1.581e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.987 on 40 degrees of freedom
Multiple R-Squared: 0.6759,    Adjusted R-squared: 0.6435
F-statistic: 20.86 on 4 and 40 DF,  p-value: 2.369e-09
```

R = 11.4(±1.0) – 2.3(±0.5)NAP - 7,6(±1.3)week2 – 6.2(±1.3)week3 – 2.6(±1.7)week4

r_{adj}^2 = 64.4%; n = 45

Selected model

R* = 11,4(±1,0) - 2,3(±0,5)*NAP* - 7,6(±1,3)*week2* - 6,2(±1,3)*week3* - 2,6(±1,7)*week4



Without week:

- RIKZ
- Richness *versus* angle2 (beach slope) NAP grainsize humus

```
Model is given by f1:
Y1 ~ 1 + angle2 + NAP + grainsize + humus


Call:
lm(formula = f1, data = dataz, weights = XW, na.action = n

Residuals:
    Min      1Q  Median      3Q     Max
-4.6851 -2.1935 -0.4218  1.6753 13.2957

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.35322    5.71888   3.209  0.00262 **
angle2      -0.02277    0.02995  -0.760  0.45144
NAP         -2.90451    0.59068  -4.917 1.54e-05 ***
grainsize   -0.04012    0.01532  -2.619  0.01239 *
humus       11.77641    9.71057   1.213  0.23234
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.644 on 40 degrees of freedom
Multiple R-Squared: 0.5178,    Adjusted R-squared: 0.4696
F-statistic: 10.74 on 4 and 40 DF,  p-value: 5.237e-06
```

Without week:

- *Forwards* (adding…)

```
Start:  AIC=145.91
Y1 ~ 1

            Df Sum of Sq     RSS    AIC
+ NAP        1    357.53  744.12 130.25
+ humus      1    181.06  920.59 139.83
+ grainsize  1    154.07  947.58 141.13
+ angle2     1    124.86  976.78 142.49
<none>                   1101.64 145.91

Step:  AIC=130.25
Y1 ~ NAP

            Df Sum of Sq    RSS    AIC
+ grainsize  1    188.61 555.50 119.09
+ angle2     1     86.65 657.46 126.68
+ humus      1     80.67 663.45 127.09
<none>                   744.12 130.25

Step:  AIC=119.09
Y1 ~ NAP + grainsize

         Df Sum of Sq    RSS    AIC
<none>                555.50 119.09
+ humus   1     16.65 538.85 119.72
+ angle2  1      4.80 550.70 120.70
```

Without week:

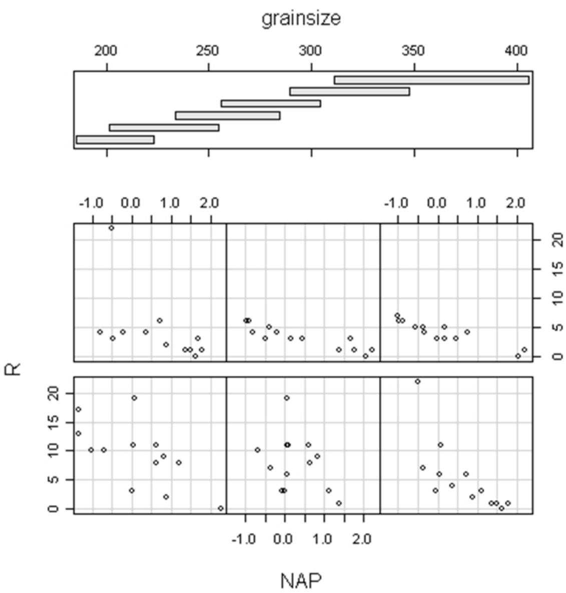
- **Final model**

```
Call:
lm(formula = Y1 ~ NAP + grainsize, (

Coefficients:
(Intercept)        NAP    grainsize
   16.50171   -3.00917     -0.03585
```

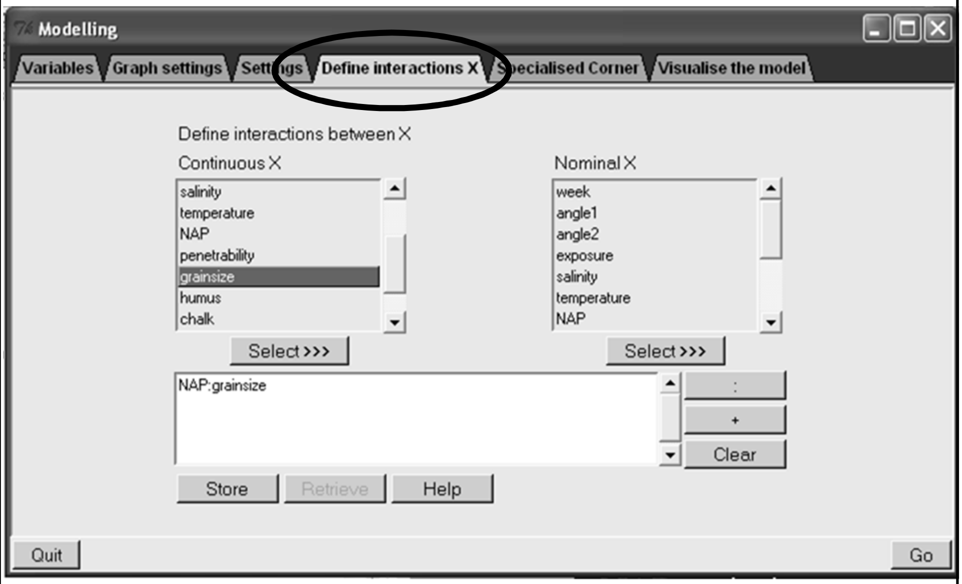
But:

- **Richness *versus* NAP by grainsize…**

(coplot!!!)



With the interaction:



With the interaction:

RIKZ
Richness
versus
NAP
grainsize
NAP:grainsize

```
Model is given by f1:
Y1 ~ 1 + NAP + grainsize + NAP:grainsize

Residuals:
    Min      1Q  Median      3Q     Max
-5.6225 -2.0147 -0.7107  1.4982 13.4201

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.952142   2.647940   6.402 1.16e-07 ***
NAP           -6.589228   2.554030  -2.580  0.01356 *
grainsize     -0.037385   0.009436  -3.962  0.00029 ***
NAP:grainsize  0.013351   0.009304   1.435  0.15891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.592 on 41 degrees of freedom
Multiple R-Squared: 0.5199,   Adjusted R-squared: 0.4847
F-statistic:  14.8 on 3 and 41 DF,  p-value: 1.125e-06
```

- ***Forwards*** **and** ***backwards***

```
Start:  AIC=145.91
Y1 ~ 1

            Df Sum of Sq     RSS    AIC
+ NAP        1    357.53  744.12 130.25
+ grainsize  1    154.07  947.58 141.13
<none>                   1101.64 145.91

Step:  AIC=130.25
Y1 ~ NAP

            Df Sum of Sq     RSS    AIC
+ grainsize  1    188.61  555.50 119.09
<none>                    744.12 130.25
- NAP        1    357.53 1101.64 145.91

Step:  AIC=119.09
Y1 ~ NAP + grainsize

                Df Sum of Sq    RSS    AIC
+ NAP:grainsize  1     26.56 528.94 118.89
<none>                       555.50 119.09
- grainsize      1    188.61 744.12 130.25
- NAP            1    392.08 947.58 141.13

Step:  AIC=118.89
Y1 ~ NAP + grainsize + NAP:grainsize

                Df Sum of Sq    RSS    AIC
<none>                       528.94 118.89
- NAP:grainsize  1     26.56 555.50 119.09
```

```
Call:
lm(formula = Y1 ~ NAP + grainsize + NAP:grainsize  d

Coefficients:
 (Intercept)          NAP    grainsize  NAP:grainsize
    16.95214     -6.58923     -0.03738        0.01335
```

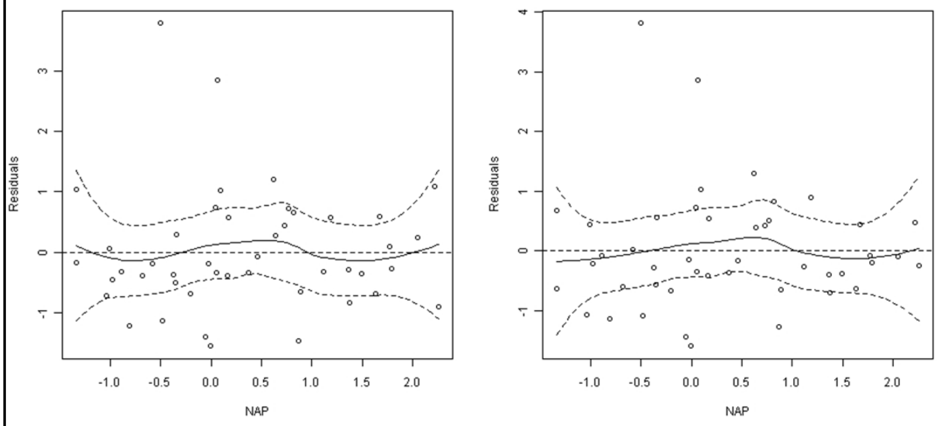
$R = 17{,}0(\pm 2{,}7) - 6{,}7(\pm 2{,}6)\,NAP - 0{,}037(\pm 0{,}009)\,grainsize + 0{,}013(\pm 0{,}009)\,NAP \times grainsize$

- **Interaction was included but no significant… therefore: <u>remove (?!)</u>… residuals?**

```
Analysis of Variance Table

Response: Y1
              Df Sum Sq Mean Sq F value    Pr(>F)
NAP            1 357.53  357.53 27.7134 4.774e-06 ***
grainsize      1 188.61  188.61 14.6202 0.0004395 ***
NAP:grainsize  1  26.56   26.56  2.0589 0.1589053
Residuals     41 528.94   12.90
```

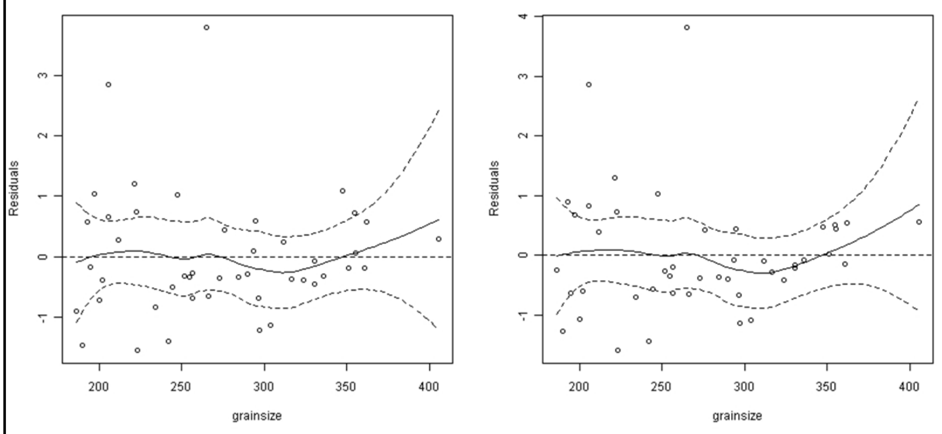
Model validation

NAP and grainsize **...and NAP:grainsize**



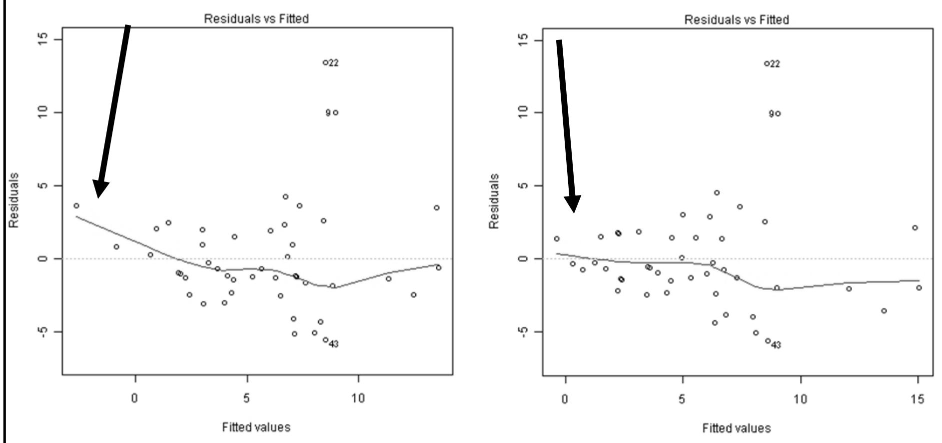
Model validation

NAP and grainsize **...and NAP:grainsize**



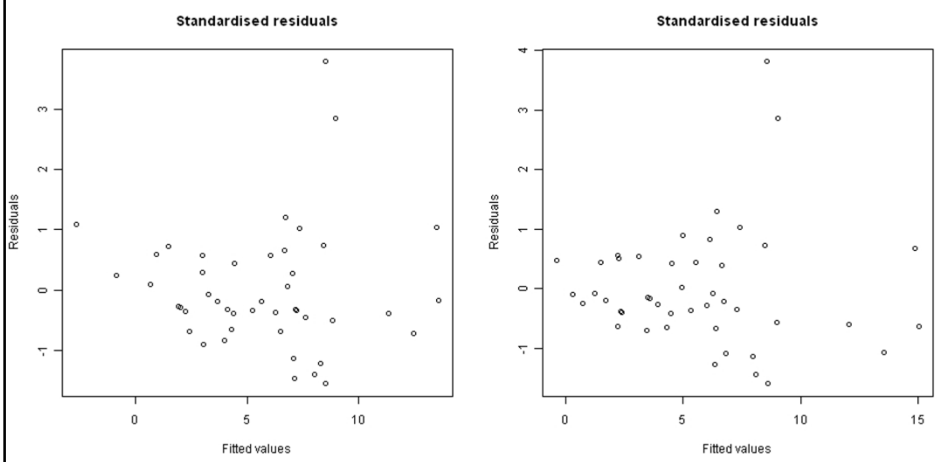
Model validation

NAP and grainsize **...and NAP:grainsize**



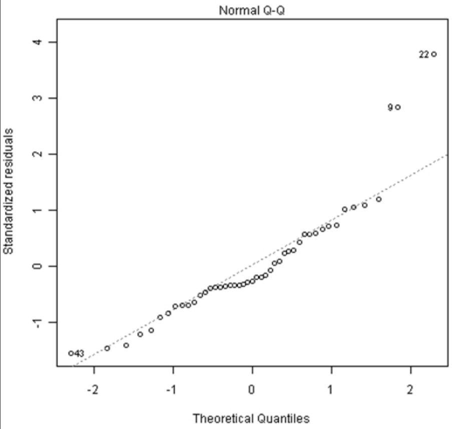
Model validation

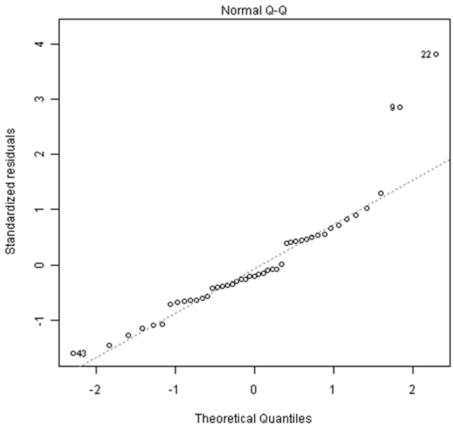
NAP and grainsize **...and NAP:grainsize**



Model validation

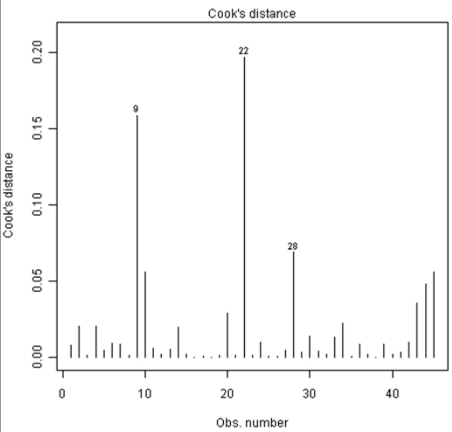
NAP and grainsize **...and NAP:grainsize**

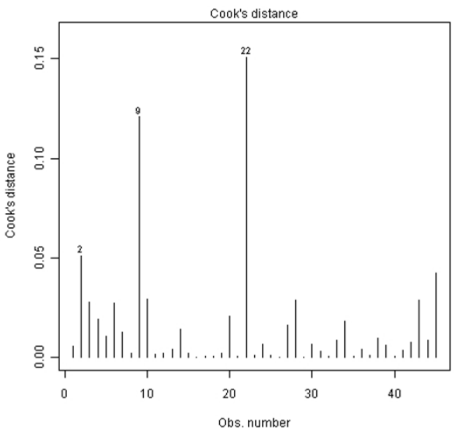




Model validation

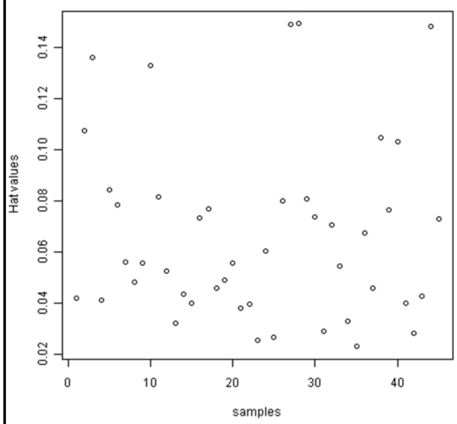
NAP and grainsize **...and NAP:grainsize**



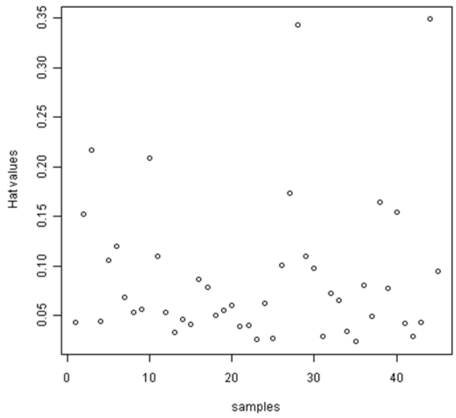


Model validation

NAP and grainsize



…and NAP:grainsize



TASK … /Loyn.xls

Apply multiple linear regression to model bird abundance …

With this wealth of potential pitfalls, ensuring that the scientist does not discover a false covariate effect (type I error), wrongly dismiss a model with a particular covariate (type II error) or produce results determined by only a few influential observations, requires that det[…] before any statistical analysis. The aim of this paper is to pro-

[…] In our experience, data exploration can take up to 50% of the time spent on analysis.

TASK … /Loyn.xls

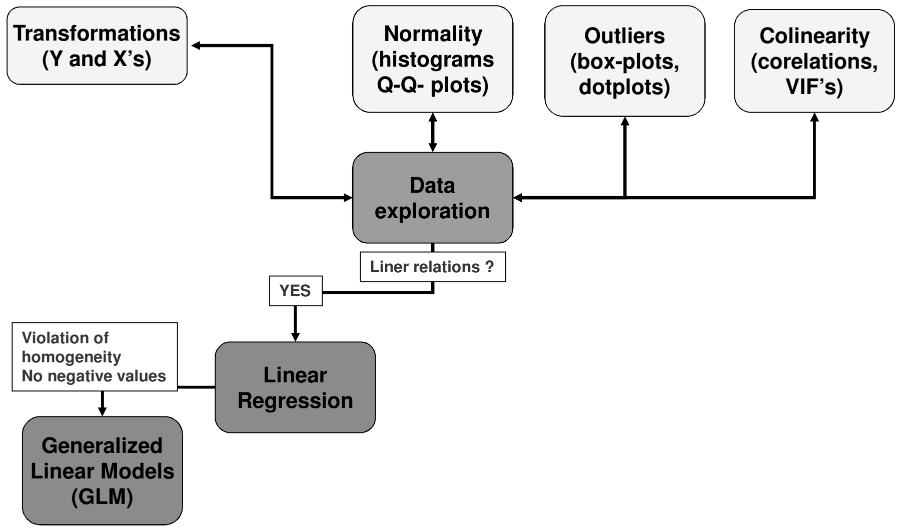
Reports should be correct, complete, clear and concise…

Regardless of the specific situation, the routine use and transparent reporting of systematic data exploration would improve the quality of ecological research and any applied recommendations that it produces.

COURSE OUTLINE

1. Data Exploration
2. Linear Regression – Bivariate and Multiple
3. Generalised Linear Modelling
   Poisson
   Logistic

Generalized Linear Models



Generalized Linear Models

- Parametric models: data belong to a specific distribution (normal, Poisson, binomial, gamma, negative binomial,…)

  Normal – may include negative values (regression)

  Poisson – count data,… (GLM Poisson)

  Binomial – binary, proportional data (GLM logistic)

Generalized Linear Models

- Different *link* functions relate the expected values of Y and the explanatory variables

- In linear regression:

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

id est:

$$Y_i = g(x_i)$$

Generalized Linear Models

- *Link* funtions:

Identity link: $\mu_i = g(x_i)$ (linear regression)

Always (*linear predictor function*):

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

Generalized Linear Models

- *Link* funtions:

Identity link: $\mu_i = g(x_i)$ (linear regression)

Log link: $\mu_i = e^{g(x_i)}$ or $\ln(\mu_i) = g(x_i)$ (GLM Poisson)

Always (*linear predictor function*):

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

Generalized Linear Models

- *Link* funtions:

Identity link: $\mu_i = g(x_i)$ (linear regression)

Log link: $\mu_i = e^{g(x_i)}$ or $\ln(\mu_i) = g(x_i)$ (GLM Poisson)

Logit link: $\ln[\mu_i/(1 - \mu_i)] = g(x_i)$ (GLM logistic)

Always (*linear predictor function*):

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

GLM Poisson

Log link: $\mu_i = e^{g(x_i)}$ or $\ln(\mu_i) = g(x_i)$

Always *(linear predictor function):*

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

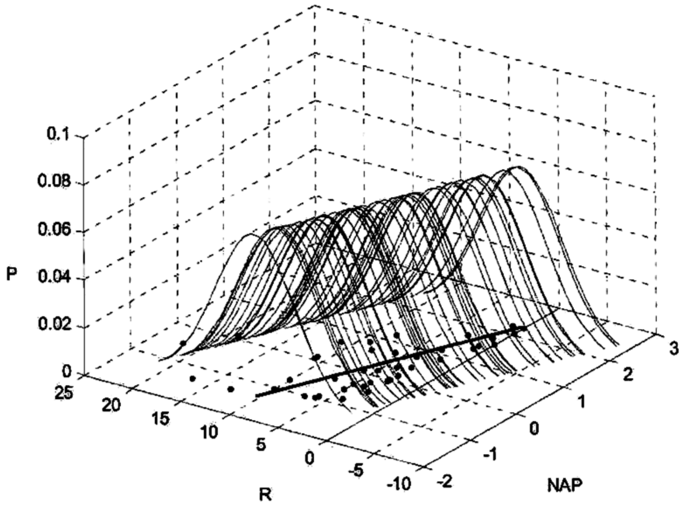
GLM Poisson

- **…/Data/RIKZ.xls**

- RIKZ – Dutch governmental institute
- intertidal *benthos*
- 9 sandy beaches in The Netherlands
- 5 stations *per* beach (10 sub-replicates)
- 4 sampling times (4 sequential weeks)
- Station and beach slopes (“angles”)
- Exposure of the beach (waves, slope,…)
- Station NAP = reflects emersion time
- Salinity, temperature, grain size, organic matter,…

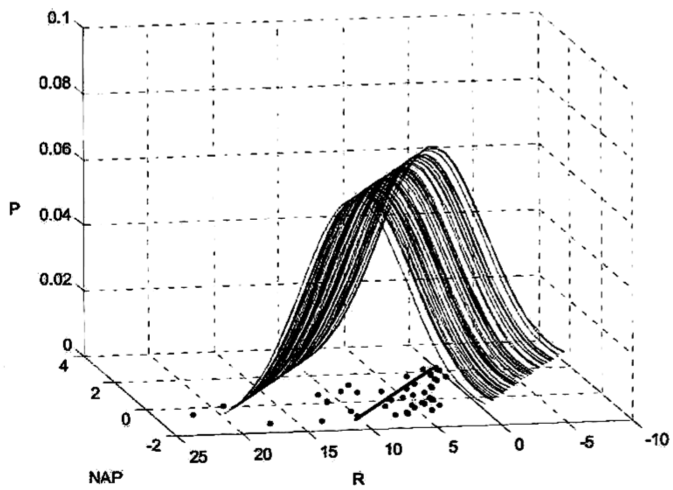
Normal distribution (or Gaussian)

- Richness *versus* NAP



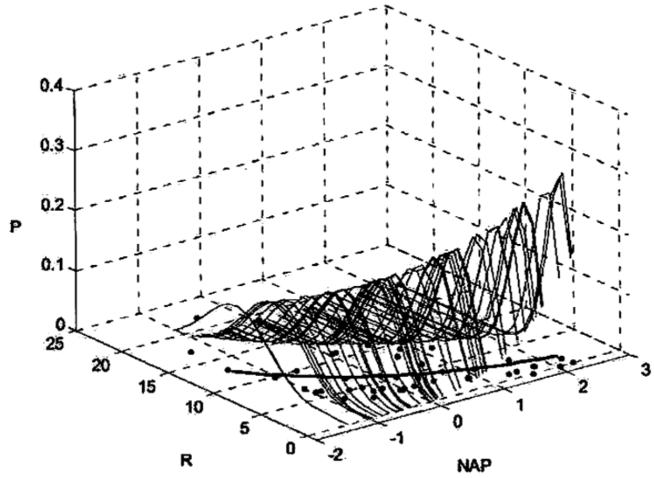
Normal distribution (or Gaussian)

- Richness *versus* NAP



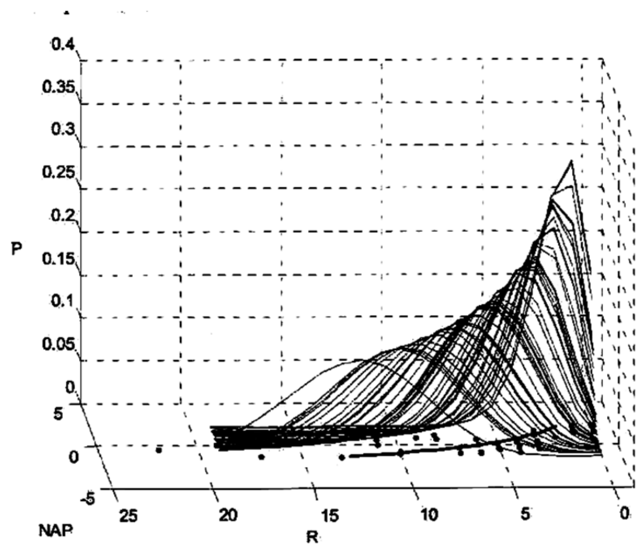
Poisson distribution

- Richness *versus* NAP



Poisson distribution

- Richness *versus* NAP



Linear regression

- **Richness *versus* NAP**

```
##########################################
#### LINEAR REGRESSION NUMERICAL OUTPUT ###
##########################################

Model is given by f1:
Y1 ~ NAP

Residuals:
    Min      1Q  Median      3Q     Max
-5.0675 -2.7607 -0.8029  1.3534 13.8723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.6857     0.6578  10.164 5.25e-13 ***
NAP          -2.8669     0.6307  -4.545 4.42e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.16 on 43 degrees of freedom
Multiple R-Squared: 0.3245,     Adjusted R-squared: 0.3088
F-statistic: 20.66 on 1 and 43 DF,  p-value: 4.418e-05

Analysis of Variance Table

Response: Y1
          Df Sum Sq Mean Sq F value    Pr(>F)
NAP        1 357.53  357.53  20.660 4.418e-05 ***
Residuals 43 744.12   17.31
```

Normal distrib. – variance

Notation	Variance in	Sum of squared deviations of	Formula
SS_{total}	Y	Observed data from the mean	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
$SS_{regression}$	Y explained by X	Fitted values from the mean value	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
$SS_{residual}$	Y not explained by X	Observed values from fitted values	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

[$r^2 = SS_{regression} / SS_{total} = (SS_{total} - SS_{residual}) / SS_{total}$]

GLM Poisson

- **Richness *versus* NAP**

```
Model is given by f1:
Y1 ~ 1 + NAP
Call:
glm(formula = Y1 ~ 1 + NAP, family = poisson(link = "log")
   weights = XW, na.action = na.omit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18
```

- ***Distribution:* Poisson**
- ***Link:* Log**

GLM Poisson

```
Model is given by f1:
Y1 ~ 1 + NAP
Call:
glm(formula = Y1 ~ 1 + NAP, family = poisson(link = "log")
   weights = XW, na.action = na.omit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18
```

[$r^2 = SS_{regression} / SS_{total} = (SS_{total} - SS_{residual}) / SS_{total}$]

Pseudo r^2* = (*null deviance - residual deviance*) / *null deviance

***Pseudo r^2* = (179,75 – 113,18) / 179,75 = 0,370 = 37,0%**

GLM Poisson

```
Model is given by f1:
Y1 ~ 1 + NAP
Call:
glm(formula = Y1 ~ 1 + NAP, family = poisson(link = "log")
    weights = XW, na.action = na.omit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18
```

Richness = $e^{1.79(\pm 0.06) - 0.56(\pm 0.07)NAP}$
Pseudo r^2 = 37.0%; n = 45

GLM Poisson

- ***Overdispersion***:

Caution (!) if > 5 (10?) then → QuasiPoisson

Ignoring *overdispersion* may result in accepting non relevant variables

```
Model is given by f1:
Y1 ~ 1 + NAP
Call:
glm(formula = Y1 ~ 1 + NAP, family = poisson(link = "log")
    weights = XW, na.action = na.omit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.79100    0.06329  28.297  < 2e-16 ***
NAP         -0.55597    0.07163  -7.762 8.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance: 113.18  on 43  degrees of freedom
AIC: 259.18

Number of Fisher Scoring iterations: 5

Deviance parameter                           = 113.18
n           (null degrees of freedom)        = 44
df.residual (residual degrees of freedom)    = 43
df          (n-df.residual)                  = 1

Overdispersion (Deviance/df.residual) =  2.63
```

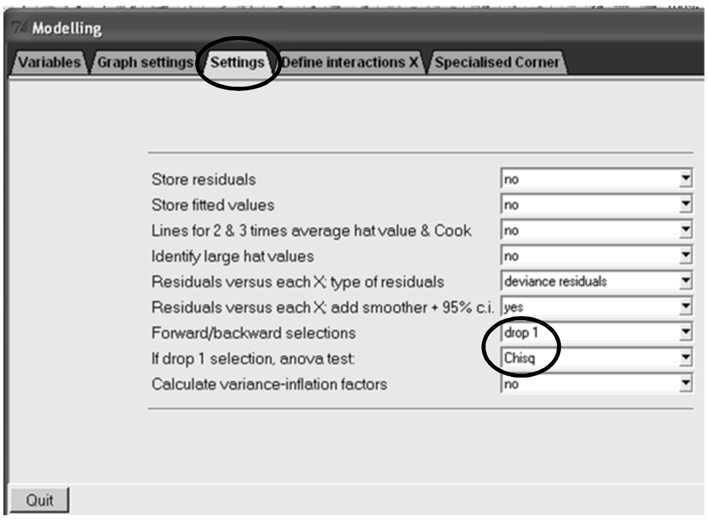
Multiple GLM

- *Distribution:* Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)
  exposure (nominal)

```
###############################################
#### NUMERICAL OUTPUT GLM            ####
###############################################
No weights  were used

Model is given by f1:
Y1 ~ 1 + NAP + as.factor(week) + as.factor(exposure)

Call:
glm(formula = Y1 ~ 1 + NAP + as.factor(week) + as.factor(exposure),
    family = poisson(link = "log"), data = dataz, weights = XW,
    na.action = na.omit)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.53136    0.12866  19.675  <2e-16 ***
NAP                   -0.48950    0.07449  -6.571   5e-11 ***
as.factor(week)2      -0.75723    0.35132  -2.155  0.0311 *
as.factor(week)3      -0.50717    0.21148  -2.398  0.0165 *
as.factor(week)4       0.12361    0.22617   0.547  0.5847
as.factor(exposure)10 -0.42602    0.19022  -2.240  0.0251 *
as.factor(exposure)11 -0.65481    0.33446  -1.958  0.0503 .

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.75  on 44  degrees of freedom
Residual deviance:  47.80  on 38  degrees of freedom
AIC: 203.8

Overdispersion (Deviance/df.residual) =  1.26
```

Multiple GLM - Drop 1, Chi square



Multiple GLM - Drop 1, Chi square

- *Distribution:* Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)
  ~~exposure (nomina~~l) ????

Model:
Y1 ~ 1 + NAP + as.factor(week) + as.factor(exposure)

	Df	Deviance	AIC	LRT	Pr(Chi)	
<none>		47.800	203.799			
NAP	1	93.460	247.459	45.660	1.407e-11	***
as.factor(week)	3	58.372	208.371	10.572	0.01428	*
as.factor(exposure)	2	53.466	205.464	5.666	0.05885	

Multiple GLM - Drop 1, Chi square

- *Distribution:* Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)
  ~~exposure (nomina~~l) ???? – compare residuals!

Model:
Y1 ~ 1 + NAP + as.factor(week) + as.factor(exposure)

	Df	Deviance	AIC	LRT	Pr(Chi)	
<none>		47.800	203.799			
NAP	1	93.460	247.459	45.660	1.407e-11	***
as.factor(week)	3	58.372	208.371	10.572	0.01428	*
as.factor(exposure)	2	53.466	205.464	5.666	0.05885	

Multiple GLM

- *Distribution:* Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)
  (exposure removed)

```
#############################################
#### NUMERICAL OUTPUT GLM           ####
#############################################
No weights  were used

Model is given by f1:
Y1 ~ 1 + NAP + as.factor(week)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.32603    0.09855  23.602  < 2e-16 ***
NAP               -0.44821    0.07313  -6.129 8.87e-10 ***
as.factor(week)2  -1.21144    0.18822  -6.436 1.22e-10 ***
as.factor(week)3  -0.80473    0.15963  -5.041 4.63e-07 ***
as.factor(week)4  -0.11102    0.19769  -0.562    0.574
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.753  on 44  degrees of freedom
Residual deviance:  53.466  on 40  degrees of freedom
AIC: 205.46

Overdispersion (Deviance/df.residual) =  1.34
```

Multiple GLM

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.32603    0.09855  23.602  < 2e-16 ***
NAP               -0.44821    0.07313  -6.129 8.87e-10 ***
as.factor(week)2  -1.21144    0.18822  -6.436 1.22e-10 ***
as.factor(week)3  -0.80473    0.15963  -5.041 4.63e-07 ***
as.factor(week)4  -0.11102    0.19769  -0.562    0.574
```

W1: $R = e^{2,33(\pm 0,10)\ -\ 0,45(\pm 0,07)NAP}$

W2: $R = e^{2,33(\pm 0,10)\ -\ 0,45(\pm 0,07)NAP\ -\ 1,21(\pm 0,19)}$

W3: $R = e^{2,33(\pm 0,10)\ -\ 0,45(\pm 0,07)NAP\ -\ 0,81(\pm 0,16)}$

W4: $R = e^{2,33(\pm 0,10)\ -\ 0,45(\pm 0,07)NAP\ -\ 0,11(\pm 0,20)}$

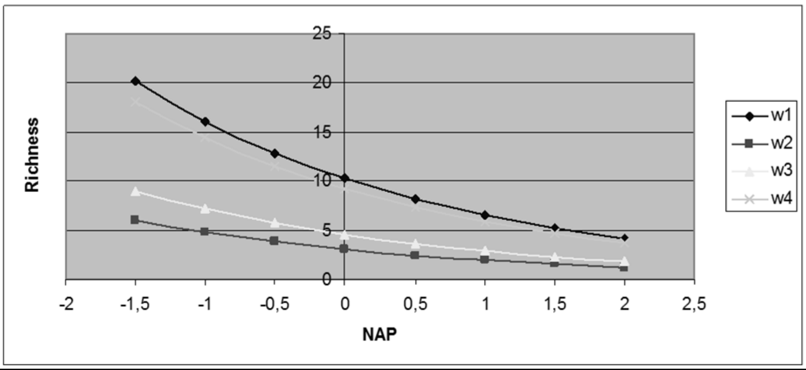
Multiple GLM

W1: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP}$

W2: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 1,21(\pm 0,19)}$

W3: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 0,81(\pm 0,16)}$

W4: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 0,11(\pm 0,20)}$



Multiple GLM

W1: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP}$

W2: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 1,21(\pm 0,19)}$

W3: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 0,81(\pm 0,16)}$

W4: $R = e^{2,33(\pm 0,10) - 0,45(\pm 0,07)NAP - 0,11(\pm 0,20)}$

```
Null deviance: 179.753  on 44  degrees of freedom
Residual deviance:  53.466  on 40  degrees of freedom
AIC: 205.46
```

pseudo r^2 = (null deviance - residual deviance) / null deviance

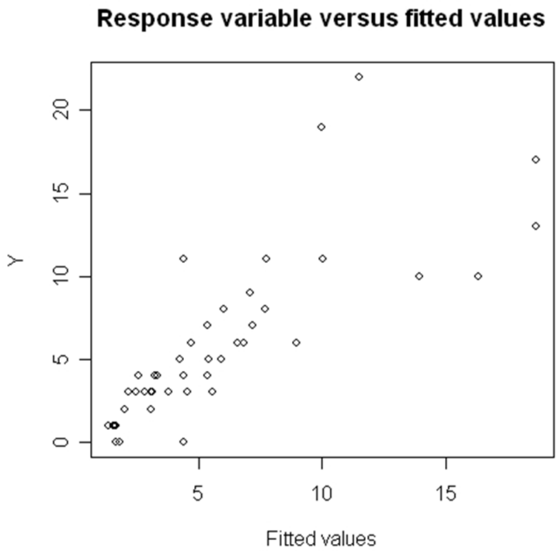
$$\text{pseudo } r^2_{adj} = 1 - \left[(1 - \text{pseudo } r^2) \times \frac{n-1}{n-m-1} \right]$$

- Sample size (n)
- Number of explanatory variables (m)

Multiple GLM

- *Distribution:* Poisson
- *Link:* Log

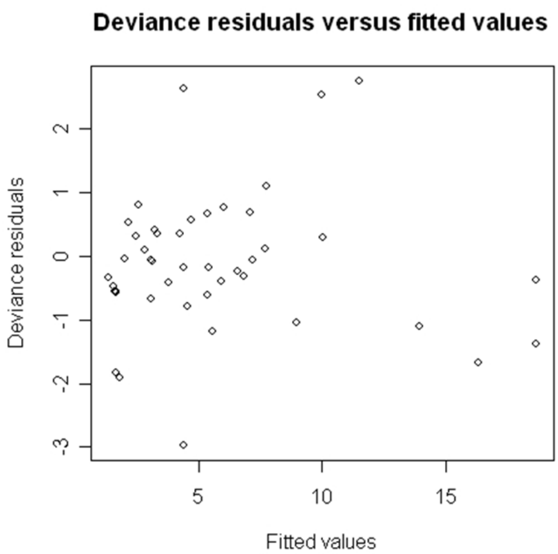
- Richness *versus* NAP week (nominal)



Multiple GLM

- *Distribution:* Poisson
- *Link:* Log

- Richness *versus* NAP week (nominal)

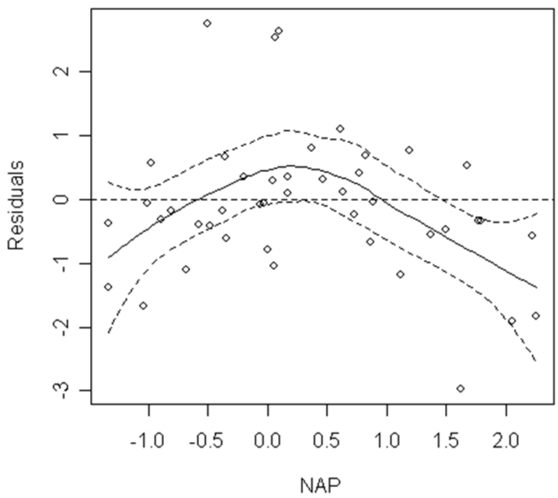


Multiple GLM

- *Distribution:*
  Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)

- No linearity!
  → GAM!!!!

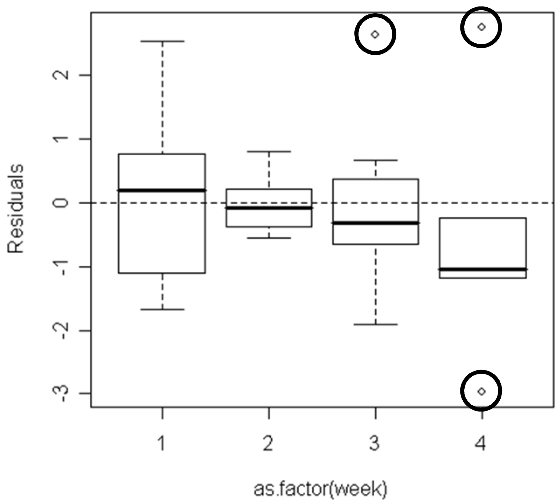


Multiple GLM

- *Distribution:*
  Poisson
- *Link:* Log

- Richness
  versus
  NAP
  week (nominal)

- Outliers



TASK … /Loyn.xls

Compare multiple linear regression and multiple GLM Poisson to model bird abundance …

GLM Logistic

Logit link: $\log[\mu_i / (1 - \mu_i)] = g(x_i)$

or

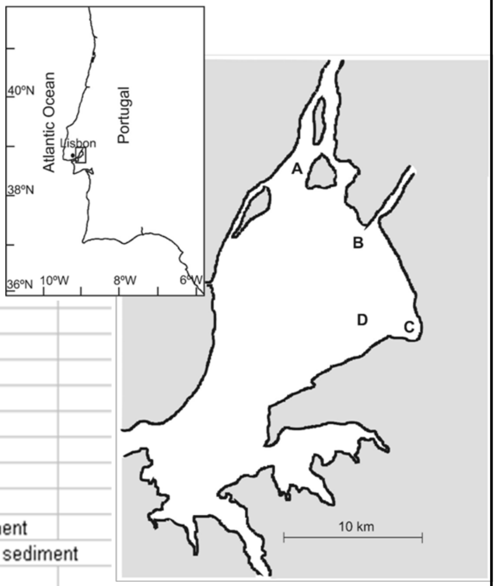
$$\mu_i = e^{g(x_i)} / (1 + e^{g(x_i)})$$

Always (linear predictor function):

$$g(x_i) = \alpha + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

…/Data/Soleasolea.xls

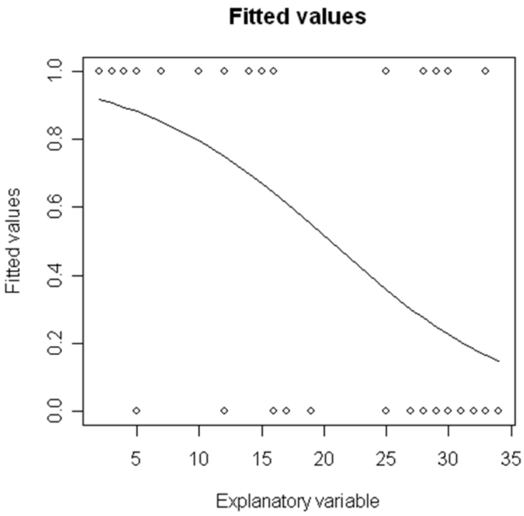
- Estuary of the Tagus river Portugal
- Presence/absence of *Solea solea*



2	season	1=spring,2=summer
3	month	
4	station	sampling station
5	depth	depth (m)
6	temp	temperature (ºC)
7	sal	salinity (ppt)
8	transp	water transparency (cm)
9	gravel	% gravel in the sediment
10	large sand	% large sand in the sediment
11	med fine sand	% medium and fine in the sediment
12	mud	% mud in the sediment
13	Solea solea	presence/absence of S. solea

GLM logistic

- Estuary of the Tagus river Portugal
- Presence/absence of *Solea solea*



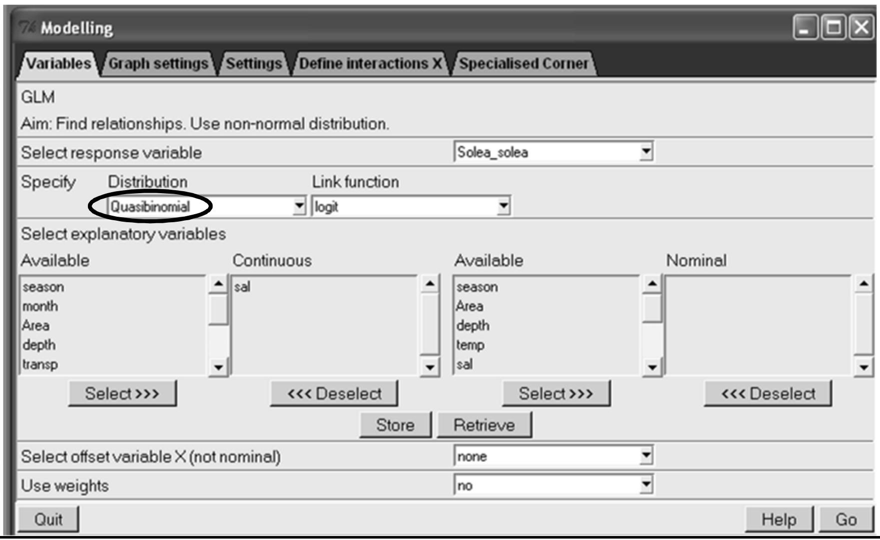
GLM logistic

- **Estuary of the Tagus river Portugal**
- **Presence/absence of *Solea solea***

```
Model is given by f1:
Y1 ~ 1 + sal


Call:
glm(formula = Y1 ~ 1 + sal, family = binomial(link = "logit"),
    data = dataz, weights = XW, na.action = na.omit)

Deviance Residuals:
   Min      1Q   Median     3Q     Max
-2.0674  -0.7146  -0.6362  0.7573  1.8996

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.66071    0.90167   2.951 0.003169 **
sal         -0.12985    0.03494  -3.716 0.000202 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87.492  on 64  degrees of freedom
Residual deviance: 68.560  on 63  degrees of freedom
AIC: 72.56

Number of Fisher Scoring iterations: 4

Deviance parameter                           = 68.56
n         (null degrees of freedom)          = 64
df.residual (residual degrees of freedom) = 63
df        (n-df.residual)                    = 1

Overdispersion (Deviance/df.residual) =  1.09
```

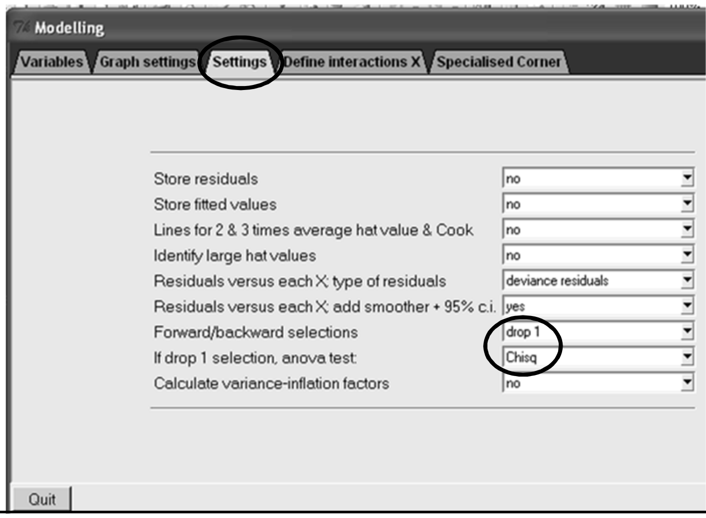
GLM logistic

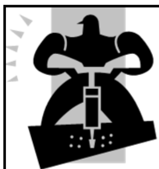
- **If *overdispersion* > ~5 → *Quasibinomial* distribution**



Multiple GLM logistic

→ **Drop 1, Chi square**





TASK **… /Frog.xls**

OBSERVED								
	Cu(ug/L)							
NaCl(g/L)	0	0,66	0,99	1,6	2,2	3,3	5	7,4
0	0,03	0,10	0,00	0,10	0,10	0,15	0,15	0,65
1,3	0,05	0,00	0,00	0,05	0,00	0,10	0,00	0,25
2	0,10	0,00	0,05	0,00	0,00	0,05	0,05	0,15
3	0,05	0,00	0,00	0,00	0,00	0,00	0,00	0,20
4,6	0,00	0,10	0,00	0,00	0,05	0,00	0,20	0,45
6,9	1,00	1,00	0,90	0,95	0,95	1,00	0,95	1,00
10,2	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

The objective is to obtain a model to predict embryo mortality at different combinations of NaCl and copper

Exercises for evaluation

Exercise 1
File: Med soils_GLM.xls
Aim: To find which soil parameters explain reproductive performance in *Eisenia andrei*

Exercise 2
File: Med soils_GLM.xls
Aim: To find which soil parameters explain avoidance behaviour in *Eisenia andrei*