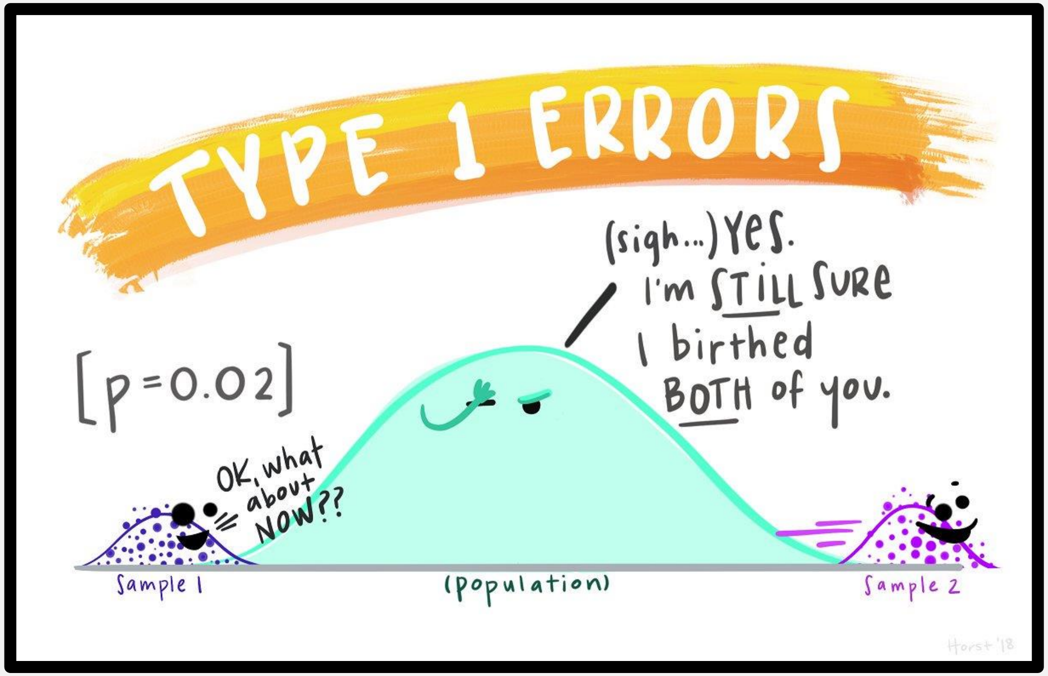
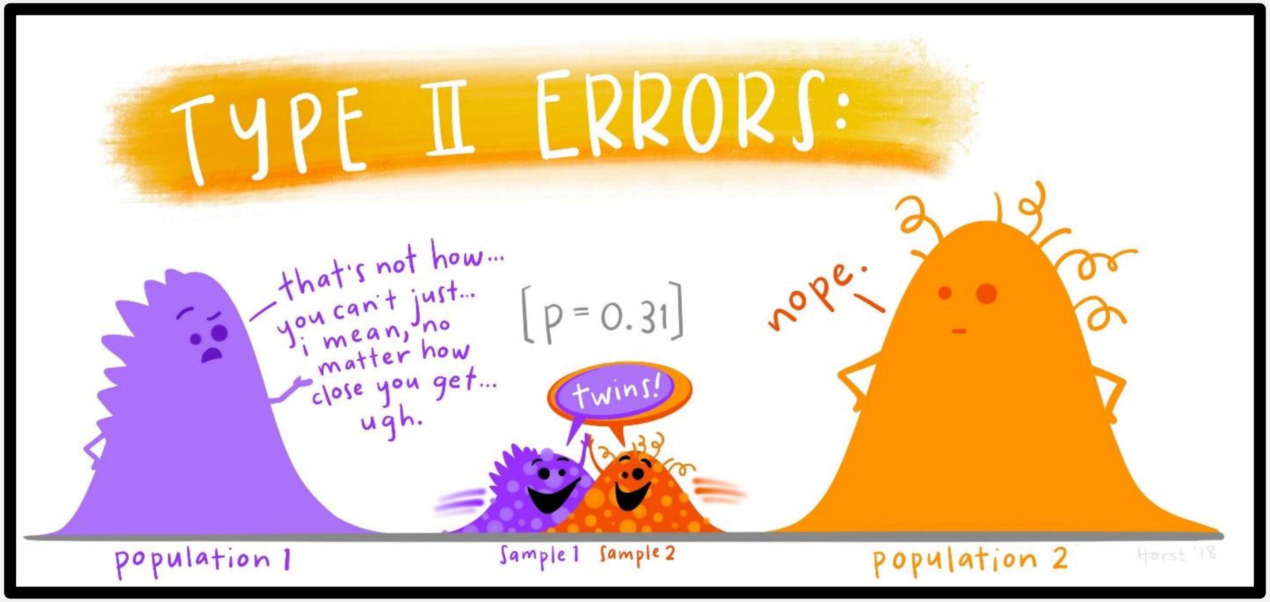
Modelação Ecológica

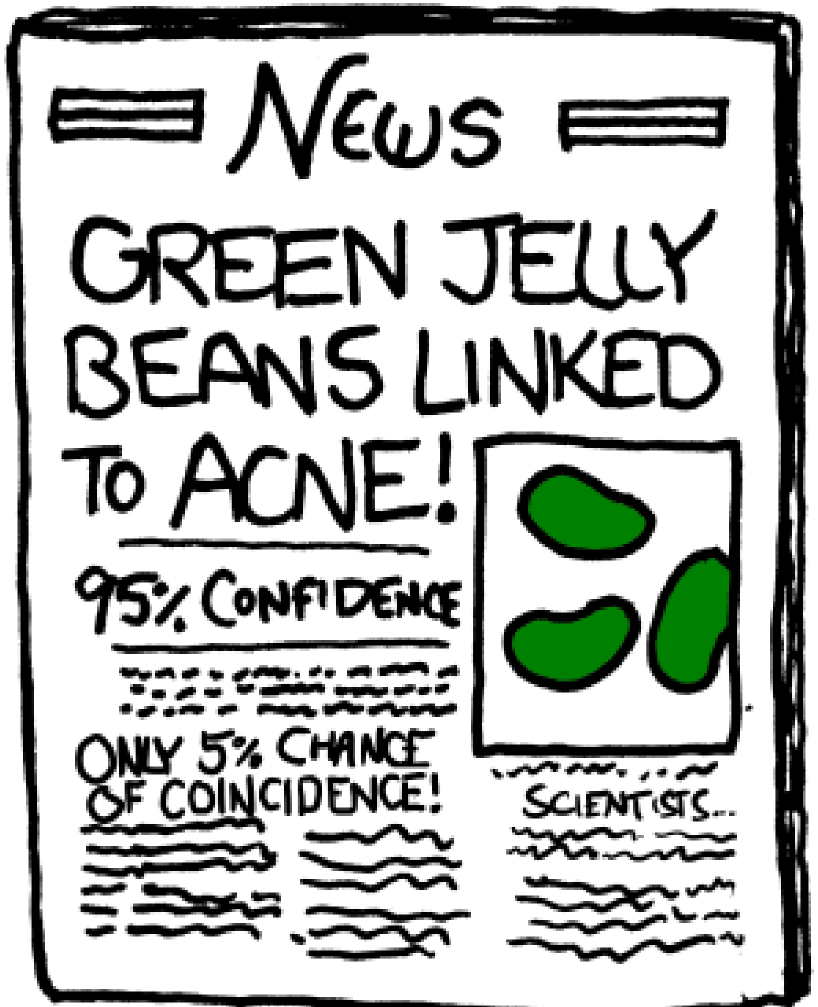
AULA 11

22 October 2019 – 14:00-16:30 – room 2.3.37

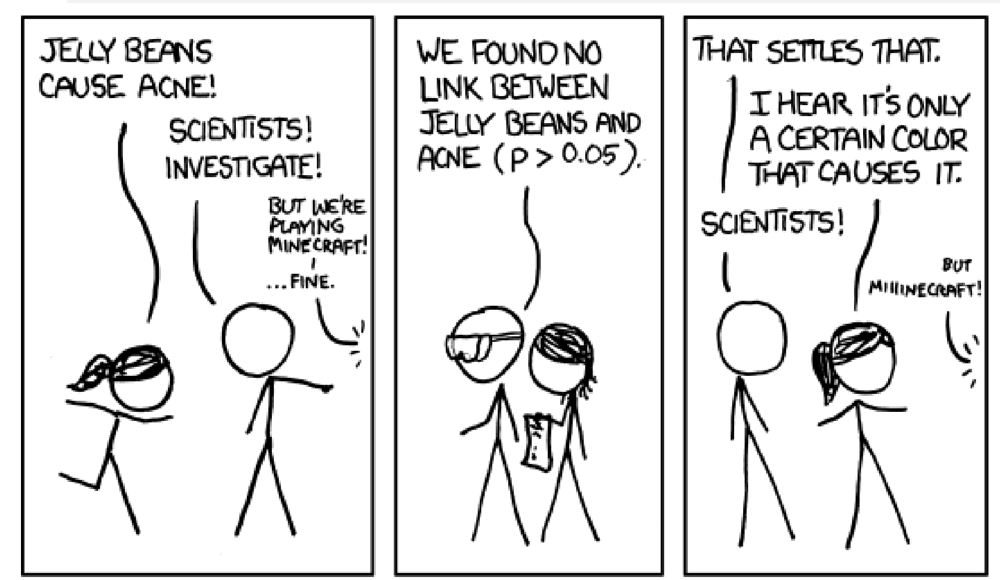
Tiago A. Marques


TYPE 1 ERRORS
(sigh...) Yes. I'm STILL sure I birthed BOTH of you.
[p=0.02]
OK, what about NOW??
Sample 1
(population)
Sample 2
Horst '18

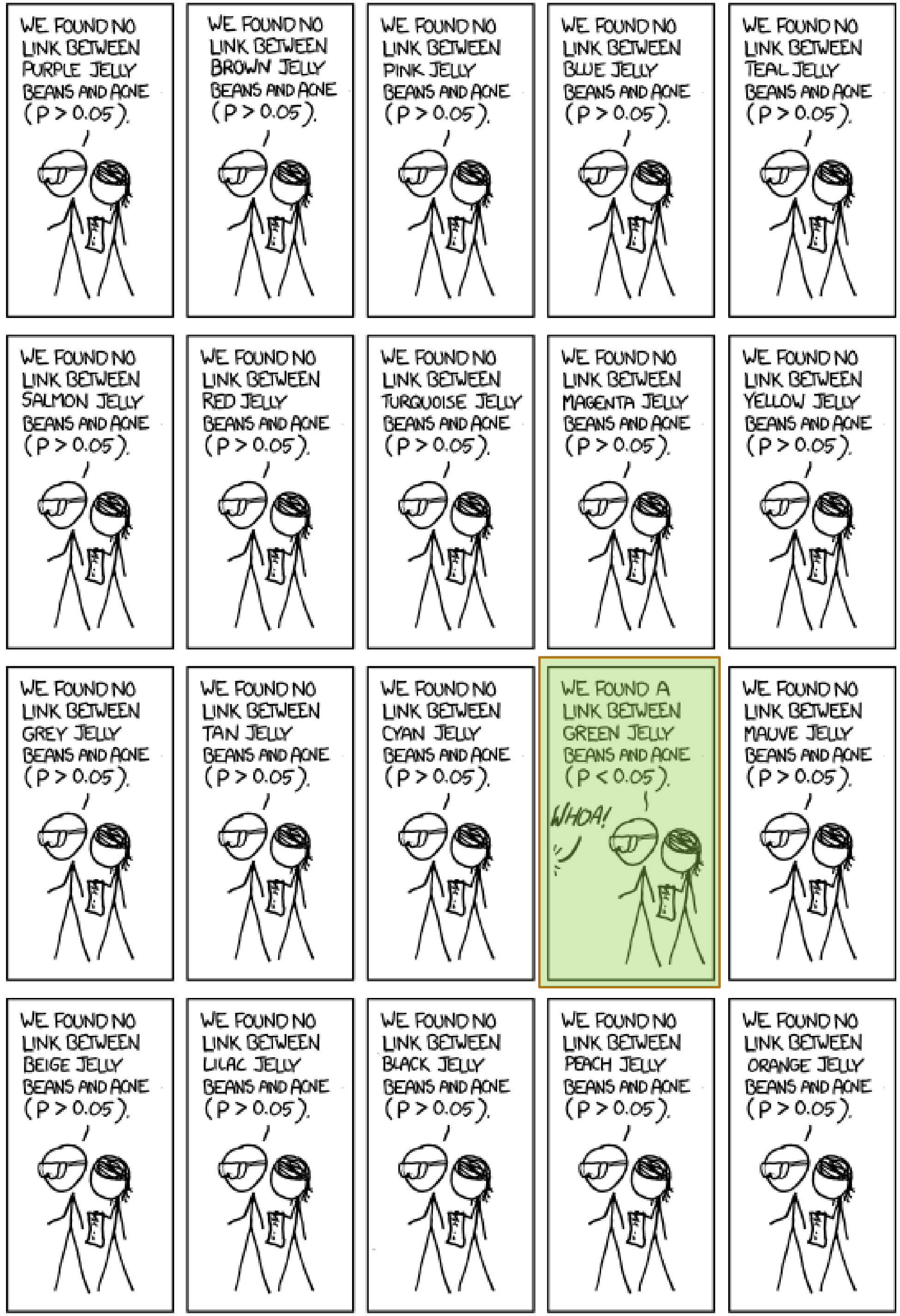

TYPE II ERRORS:
that's not how... you can't just... i mean, no matter how close you get... ugh.
[p=0.31]
nope.
twins!
population 1
Sample 1
Sample 2
population 2
Horst '18



https://xkcd.com/882/



Se não conhecem o xkcd.com… vale a pena explorar!



WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).

Editorial

Publication bias: What are the challenges and can they be overcome?

Ridha Joober, MD, PhD; Norbert Schmitz, PhD; Lawrence Annable, Dipstat; Patricia Boksa, PhD

Joober, Boksa — Douglas Mental Health University Institute and Department of Psychiatry, McGill University, Montréal, Que.; Schmitz, Annable — Department of Psychiatry, McGill University Health Centre, Montréal, Que.

Appearances to the mind are of four kinds.
Things either are what they appear to be;
Or they neither are, nor appear to be;
Or they are, and do not appear to be;
Or they are not, and yet appear to be.
Rightly to aim in all these cases
Is the wise man's task.
Epictetus, 2nd century AD

In the last few years, several meta-analyses[1–4] have reappraised the efficacy and safety of antidepressants and concluded that the therapeutic value of these drugs may have been significantly overestimated (see Ioannidis[5]). In some instances, the authors of these meta-analyses resorted to the United States' Freedom of Information Act to obtain unpublished data that, when included in meta-analyses with previously published data, reduced significantly the purported

that they are more likely to be considered for publication by editors, more favourably reviewed by peers and, once published, more likely to be cited. For editors, it is the competition for citation index and the financial survival of journals that makes it more attractive to publish positive findings.

Although publication bias has been documented in the literature for decades and its origins and consequences debated extensively, there is evidence suggesting that this bias is increasing. A recent investigation covering more than 4600 publications from different countries and disciplines found strong evidence for a steady and significant increase in publication bias over the years. The frequency of papers declaring significant statistical support for their a priori formulated hypotheses increased by 22% between 1990 and 2007 ($n = 4656$, $p < 0.001$). Psychology and psychiatry are among the disciplines in which this increase is highest ($p < 0.001$).[7] A

Adicionei ao Fenix

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3341407/

Gestão de Páginas

- Modelação Ecológica
  - Modelação Ecológica(Ecologia Marinha)
  - Modelação Ecológica(Ecologia e Gestão Ambiental)
  - Aulas
    - Aula1
    - Aula2
    - Aula3
    - Aula4
    - Aula5
    - Aula6
    - Aula7 08 10 2019
    - Aula8 09 10 2019
    - Aula9 15 10 2019
    - Aula10 16 10 2019
    - Aula11 22 10 2019
      - GLM example
      - Parques Eólicos
    - Aula12 23 10 2019
  - Outros Recursos
    - PDFs
    - R Cheat Sheets
    - Propostas de resolução de fichas de trabalho
  - Avaliação

+ Criar

PDFs

Página | Ficheiros 11 | Permissões | Link

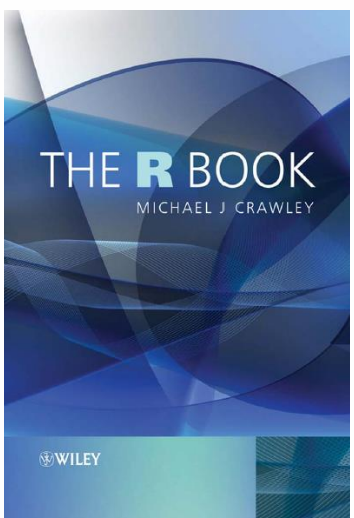
Adicionar Ficheiro

#	Nome
1	Modelling ecological systems in a changing world *Evans2012.pdf*
2	Norberg_et_al-2019 A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels *Norberg_et_al-2019-Ecological_Monographs.pdf*
3	The importance of stupidity in scientific research *Schwartz2008.pdf*
4	Ecological Models and Data in R *Bolker2007.pdf*
5	Numerical Ecology with R *Borcardetal2001EcologyUseR.pdf*
6	Introduction to Probability and Statistics Using R *IPSUR.pdf*
7	A Beginner's Guide to R *Zuuretal2009useR.pdf*
8	Analyzing Ecological Data *zuur_2007.pdf*
9	Mixed Effects Models And Extensions In Ecology With R *Zuur_Mixed-effects-models-and-extensions-in-ecology-with-R.pdf*
10	The R Book.pdf
11	Publication bias: What are the challenges and can they be overcome? *jpn-37-149.pdf*

NEW RESOURCE: THE R BOOK

- Modelação Ecológica
  - Modelação Ecológica(Ecologia Marinha)
  - Modelação Ecológica(Ecologia e Gestão Ambiental)
  - Aulas
  - Outros Recursos
    - PDFs
    - R Cheat Sheets
    - Propostas de resolução de fichas de trabalho
  - Avaliação

+ Criar



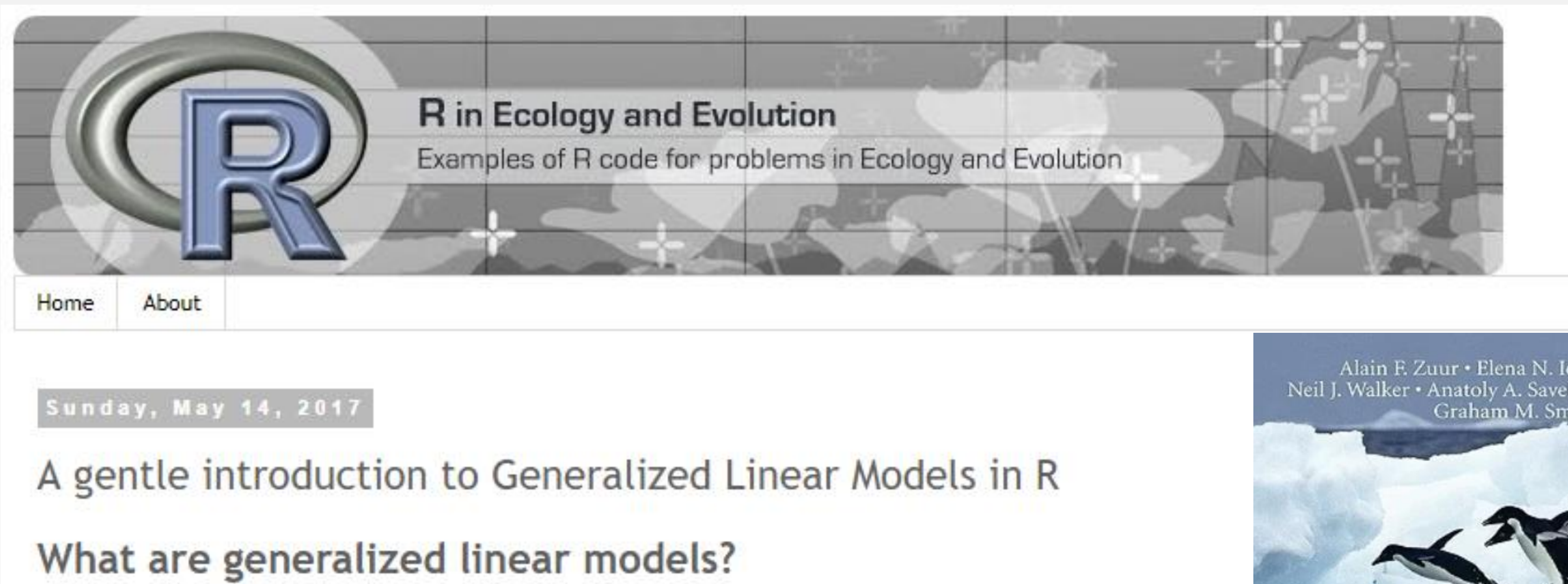
PDFs

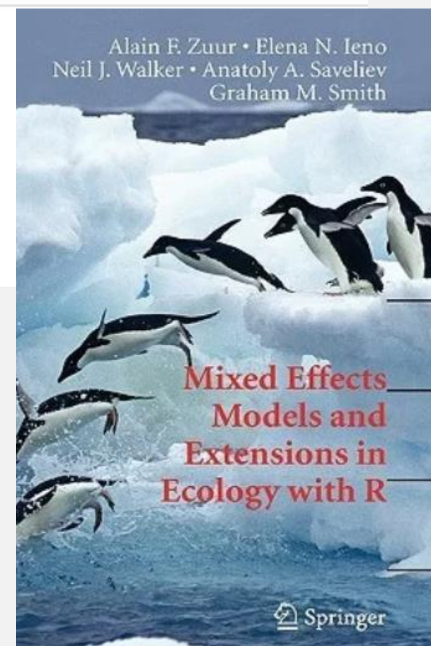
Página | Ficheiros 10 | Permissões | Link

Adicionar Ficheiro

#	Nome
1	Modelling ecological systems in a changing world *Evans2012.pdf*
2	Norberg_et_al-2019 A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels *Norberg_et_al-2019-Ecological_Monographs.pdf*
3	The importance of stupidity in scientific research *Schwartz2008.pdf*
4	Ecological Models and Data in R *Bolker2007.pdf*
5	Numerical Ecology with R *Borcardetal2001EcologyUseR.pdf*
6	Introduction to Probability and Statistics Using R *IPSUR.pdf*
7	A Beginner's Guide to R *Zuuretal2009useR.pdf*
8	Analyzing Ecological Data *zuur_2007.pdf*
9	Mixed Effects Models And Extensions In Ecology With R *Zuur_Mixed-effects-models-and-extensions-in-ecology-with-R.pdf*
10	The R Book.pdf

Generalized Linear Models
(continued!)



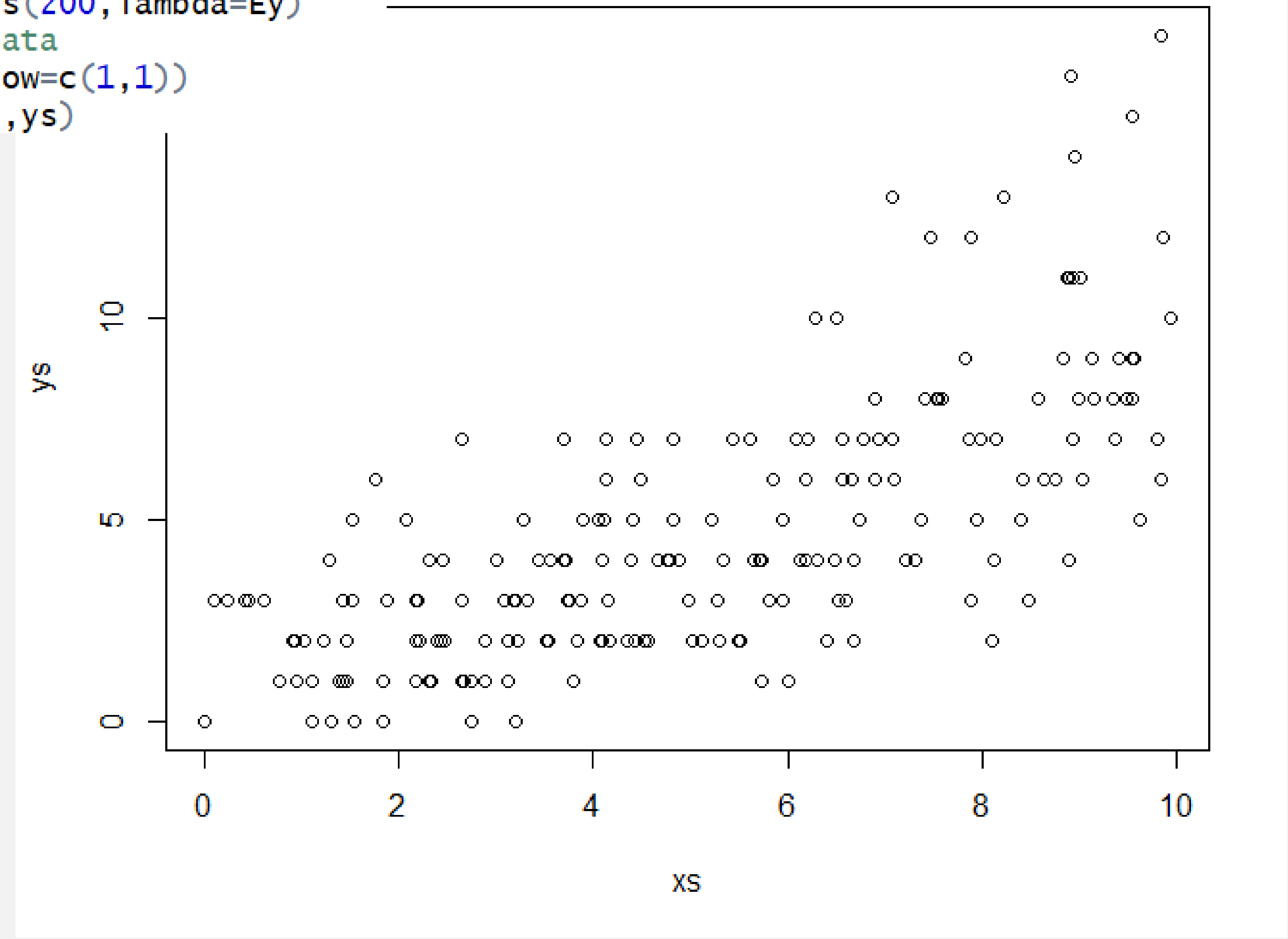


http://r-eco-evo.blogspot.com/2017/05/generalized-linear-models.html
http://spatialecology.weebly.com/r-code--data/category/glm

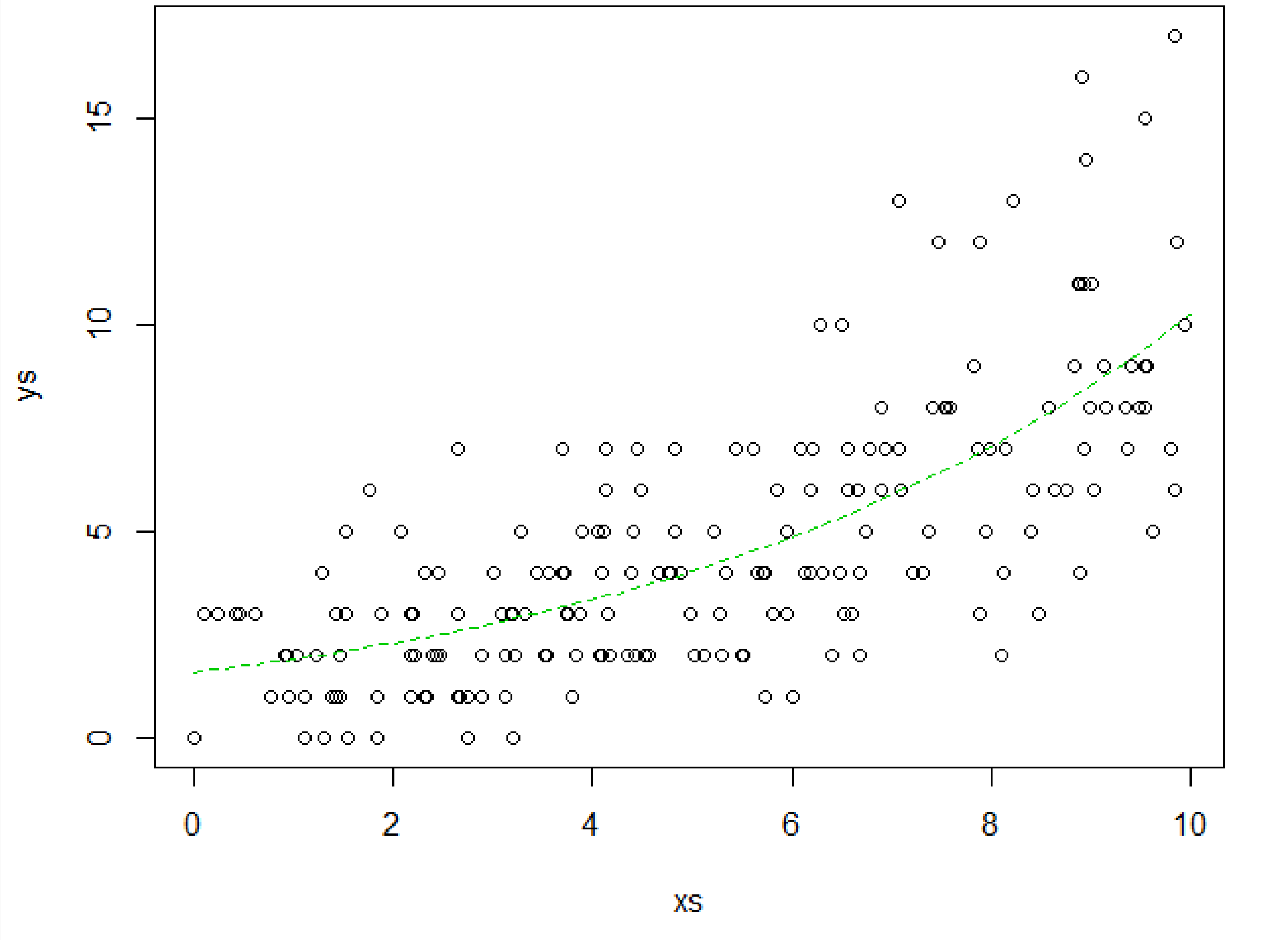
Working further on a GLM example

Como ver o modelo GLM estimado?

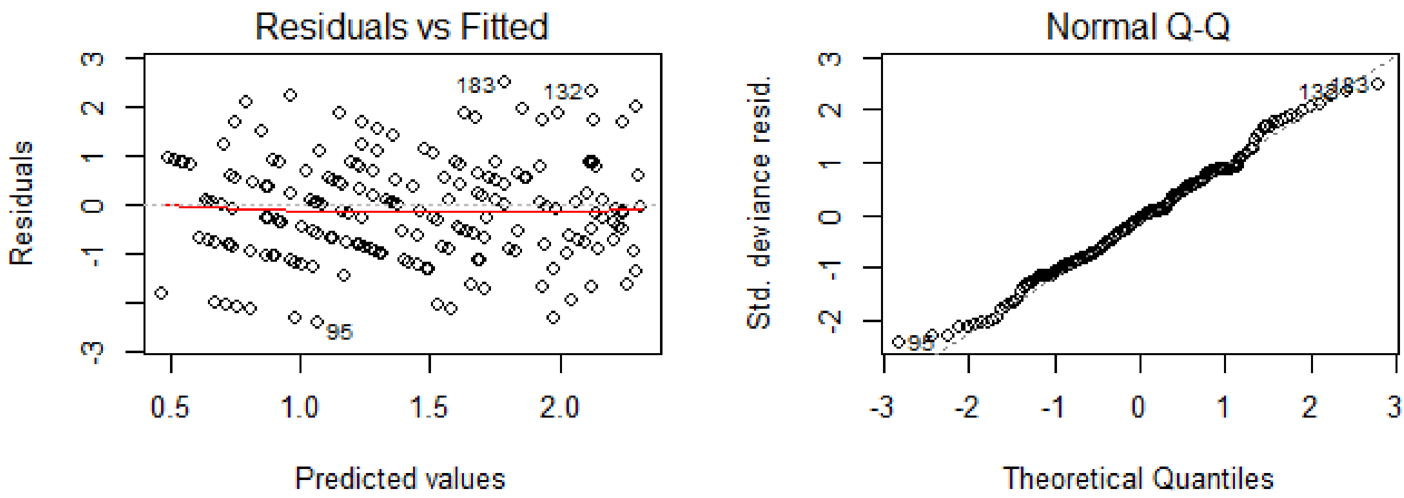
```
#creating data for a glm
set.seed(123)
#deifine the covariate
xs=runif(200,0,10)
#get the mean value
Ey=exp(0.4+0.2*xs)
#generate response
ys=rpois(200,lambda=Ey)
#plot data
par(mfrow=c(1,1))
plot(xs,ys)
```



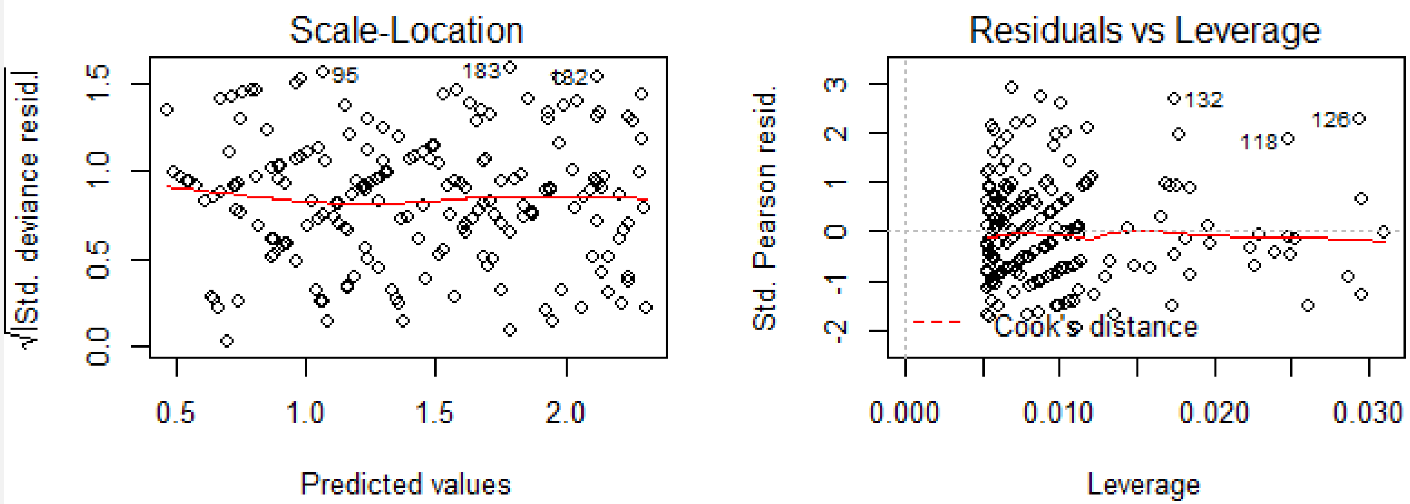
```
#fit model
glm1=glm(ys~xs,family=poisson(link=log))
#get new data for prediction
newxs=seq(0,10,by=0.1)
#predict
predglm1=predict(glm1,newdata=data.frame(xs=newxs),type="response")
#add fitted model
plot(xs,ys)
lines(newxs,predglm1,lty=2,col=3)
```



```
#diagnostics plot
par(mfrow=c(2,2))
plot(glm1)
```

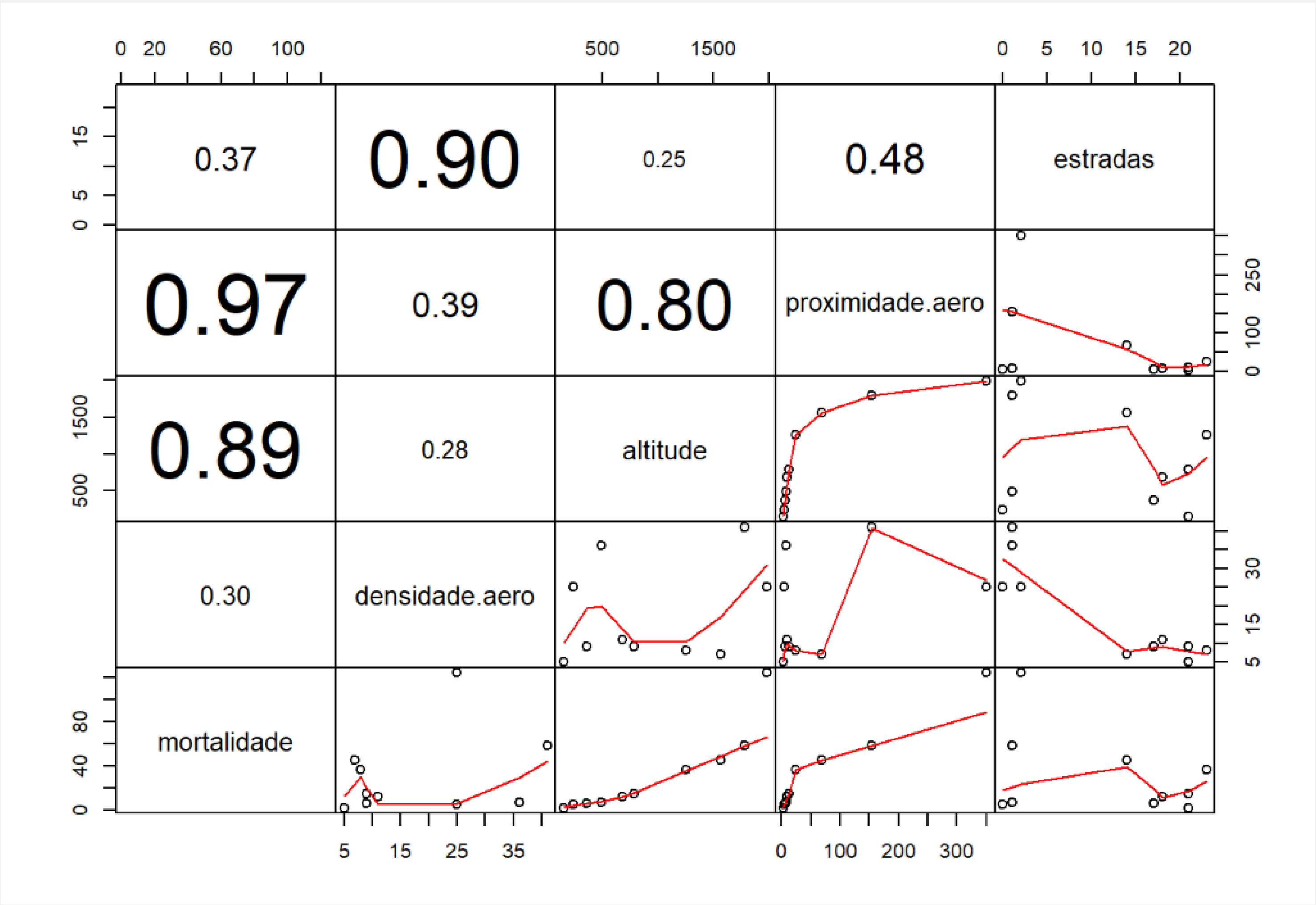


Note that now, even for residuals that are truly from a Poisson model, you get patterns in the residuals!



HANDS-ON GLM

Usando os dados “parqueseolicos.csv”, pasta FENIX “Parques Eólicos” - Explicar a mortalidade em função de 4 variáveis independentes



```r
regmul<-lm(mortalidade~., data=peol)
summary(regmul)
```

```
##
## Call:
## lm(formula = mortalidade ~ ., data = peol)
##
## Residuals:
##      1      2      3      4      5      6      7      8      9     10
## -2.146  1.872 -2.155  3.056 -1.753 -1.536  7.734 -2.491 -4.593  2.010
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.418208  10.009923   0.741 0.491957
## densidade.aero  -0.394911   0.303776  -1.300 0.250299
## altitude         0.018310   0.004458   4.108 0.009286 **
## proximidade.aero  0.252193   0.029437   8.567 0.000357 ***
## estradas        -0.228630   0.446504  -0.512 0.630427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.872 on 5 degrees of freedom
## Multiple R-squared:  0.9908, Adjusted R-squared:  0.9835
## F-statistic:   135 on 4 and 5 DF,  p-value: 2.804e-05
```

```r
library(MASS)
full.model<-lm(mortalidade~.,data=peol)
best.model<-stepAIC(full.model, direction="backward")
```

```
## Start:  AIC=34.74
## mortalidade ~ densidade.aero + altitude + proximidade.aero +
##     estradas
##
##                    Df Sum of Sq     RSS    AIC
## - estradas          1      6.22  124.90 33.249
## <none>                           118.68 34.738
## - densidade.aero    1     40.11  158.79 35.650
## - altitude          1    400.47  519.15 47.496
## - proximidade.aero  1   1742.14 1860.81 60.262
##
## Step:  AIC=33.25
## mortalidade ~ densidade.aero + altitude + proximidade.aero
##
##                    Df Sum of Sq     RSS    AIC
## <none>                           124.90 33.249
## - densidade.aero    1     85.32  210.22 36.456
## - altitude          1    435.26  560.16 46.256
## - proximidade.aero  1   2465.64 2590.54 61.570
```

```r
summary(best.model)
```

```
##
## Call:
## lm(formula = mortalidade ~ densidade.aero + altitude + proximidade.aero,
##     data = peol)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.248 -2.228 -1.510  2.738  7.114
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.665547   3.510035   0.759   0.4764
## densidade.aero   -0.255509   0.126211  -2.024   0.0893 .
## altitude          0.017361   0.003797   4.573   0.0038 **
## proximidade.aero  0.259738   0.023866  10.883 3.57e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.563 on 6 degrees of freedom
## Multiple R-squared:  0.9903, Adjusted R-squared:  0.9855
## F-statistic: 205.1 on 3 and 6 DF,  p-value: 1.963e-06
```

With such a small number of variables, we might just fit all the models and pick the best!

This can be done using function `bestglm` in package `bestglm` (note that while the model is a simple linear model, the function can cope with GLMs too!).

The first argument MUST be a matrix with all the variables, and the last column needs to be the response variable.

```
library(bestglm)
bestGLM=bestglm(Xy=peol[,c(2:5,1)],family = gaussian,IC="AIC",RequireFullEnumerationQ=TRUE)
```

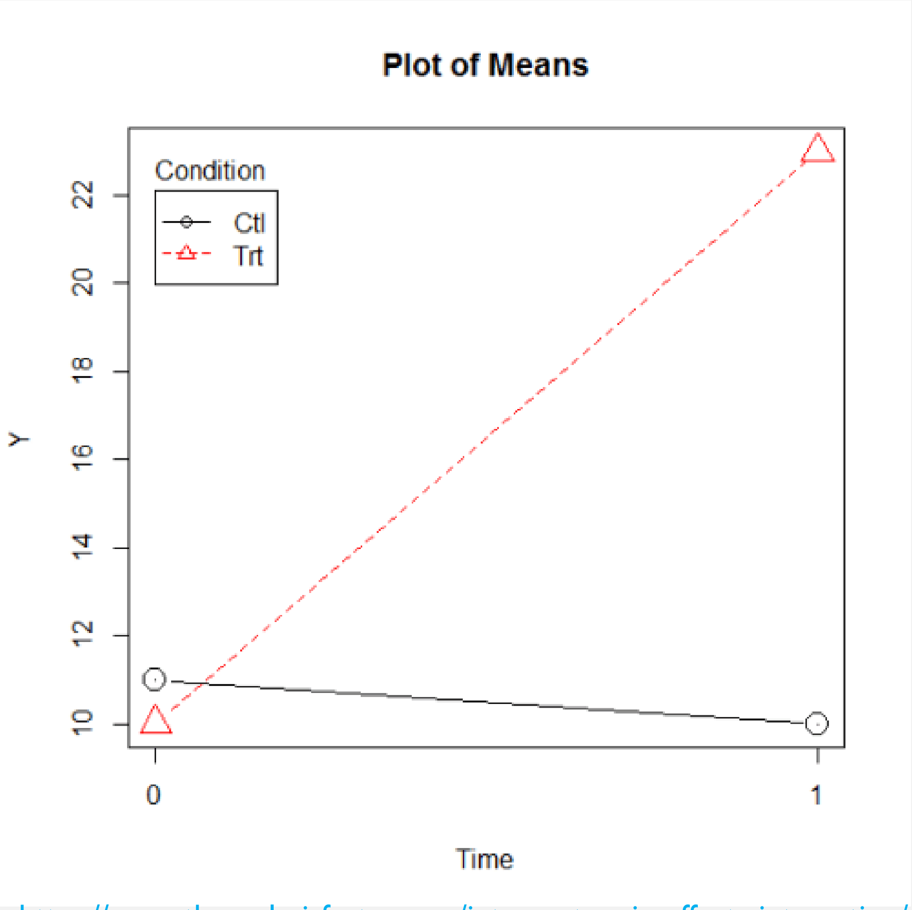
```
## Morgan-Tatar search RequireFullEnumerationQ=TRUE
```

```
bestGLM$BestModel
```

```
##
## Call:
## lm(formula = y ~ ., data = data.frame(Xy[, c(bestset[-1], FALSE),
##     drop = FALSE], y = y))
##
## Coefficients:
##      (Intercept)    densidade.aero          altitude  proximidade.aero
##          2.66555          -0.25551           0.01736           0.25974
```

In this case, we get the same result as the backwards procedure, which is reassuring, as for many variables, doing all the possible combinations becomes computationally intense.

Regression with Interactions



https://www.theanalysisfactor.com/interpret-main-effects-interaction/

In a regression context we are usually interested in explaining a response variable as a function of independent covariates

Response=function(var1 + var 2 + ….+ var k)

The default approach is to consider that these independent covariates act independently on the response, but…

Sometimes the effect of one covariate might depend on the level of a factor or on the value of a second variable that one is considering – this is called an interaction

As an example, the impact on the weight of a fish of a given type of food might be dependent on the temperature at which the fish is living. In such a case we would say there is an interaction between temperature and diet in the determination of a fish weight.

When defining a model in R, we represent an interaction term between variables A and B as A:B. If we want to run a model to explain Y that includes variables A and B and their interaction, we can use

Y~A+B+A:B

or equivalently

Y~A*B

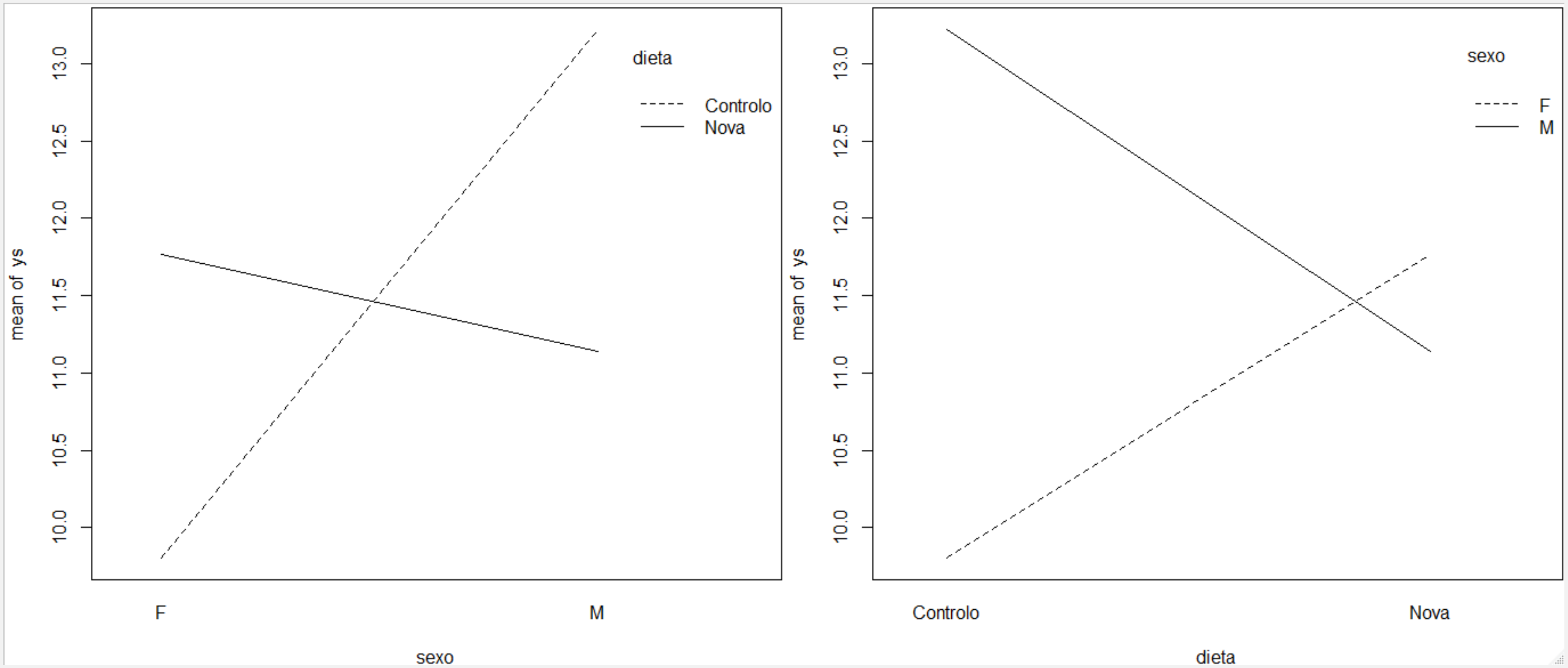
```
mod1=lm(ys~xs1+xs2+xs1:xs2)
mod2=lm(ys~xs1*xs2)
```

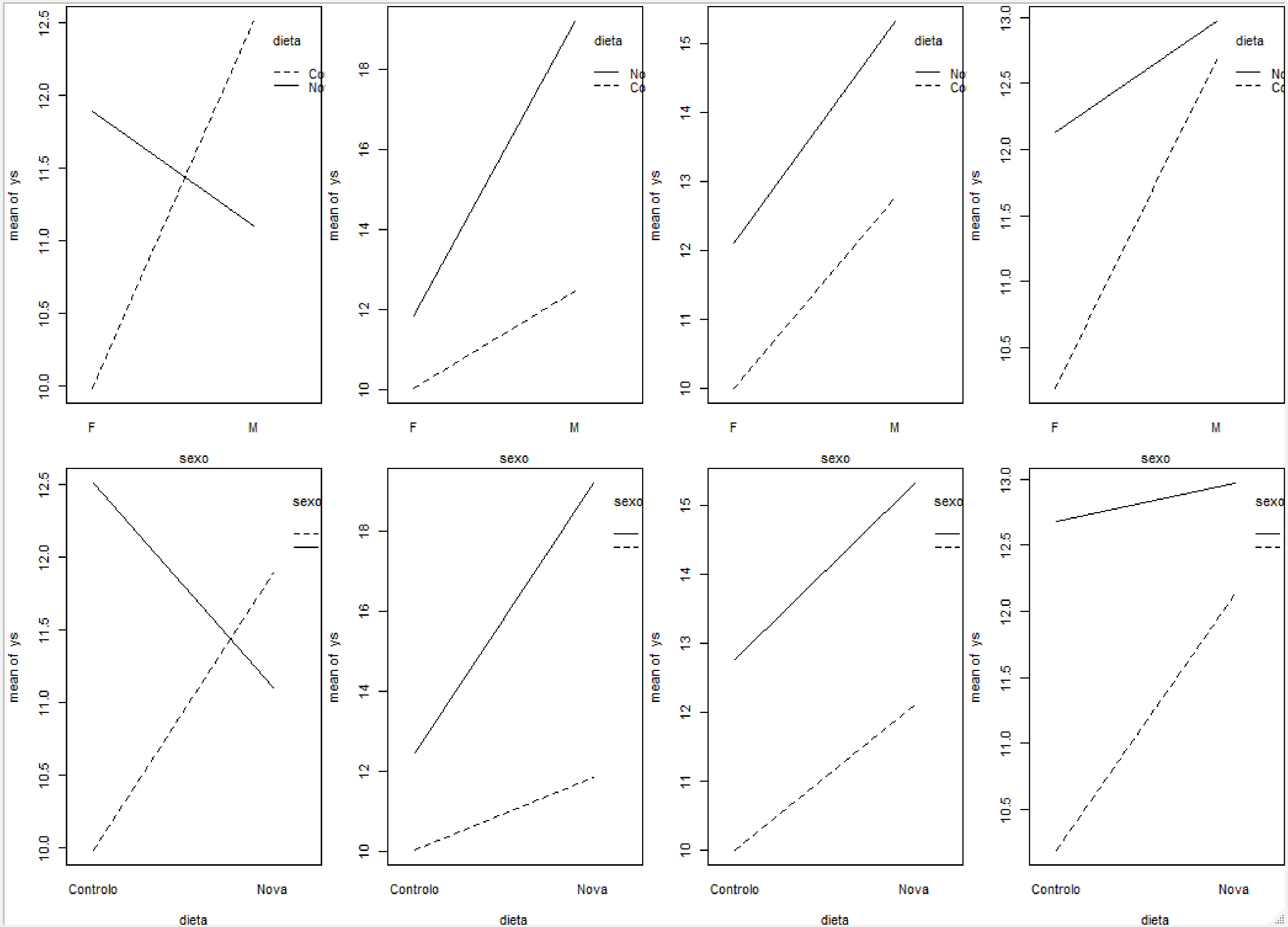
What is really an interaction?

```
#---------------------------------------------------------
#Interactions
#### with factor covariates
#---------------------------------------------------------
set.seed(123)
n=100
sexo=rep(c("M","F"),each=n)
dieta=rep(c("Controlo","Nova"),times=n)
ys=10+3*(sexo=="M")+2*(dieta=="Nova")-4*(sexo=="M")*(dieta=="Nova")+rnorm(2*n,mean=0,sd=2)
plot(ys~as.factor(paste0(sexo,dieta)))
lmSDi=lm(ys~sexo*dieta)
summary(lmSDi)

par(mfrow=c(1,2),mar=c(4,4,0.2,0.2))
interaction.plot(x.factor=sexo, trace.factor=dieta, response=ys)
interaction.plot(x.factor=dieta, trace.factor=sexo, response=ys)
```

The new diet helps females gain weight, but it actually makes males lighter! In other words, the new diet is not better or worse, it depends on the sex!



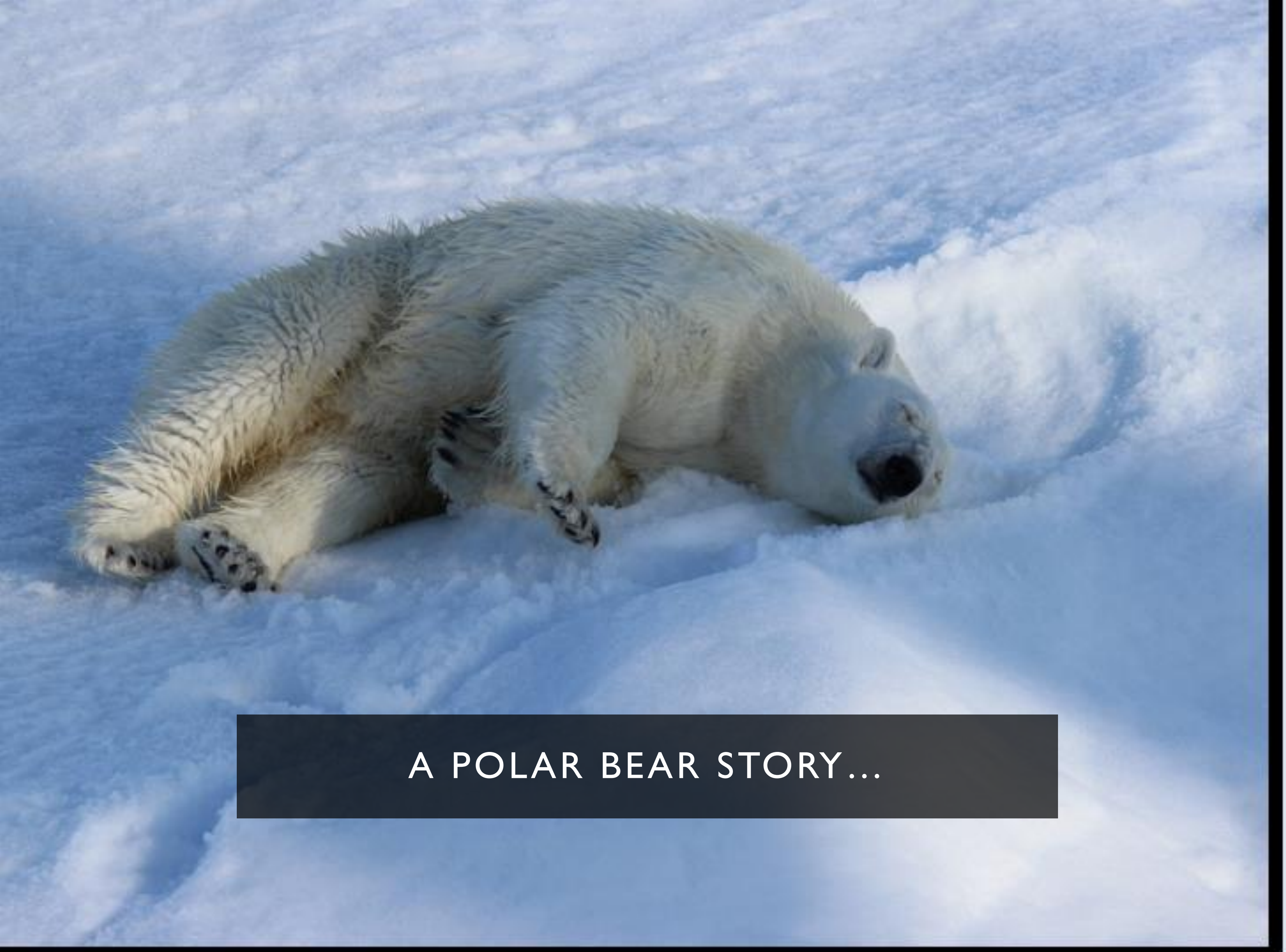


interaction

interaction

interaction

interaction


A POLAR BEAR STORY…


2015 Barents Sea
Polar Bear Survey
tiny
A photo selection by Tiago A. Marques


CO-PILOT

© Karen Lone 2015

Interactions for continuous covariates

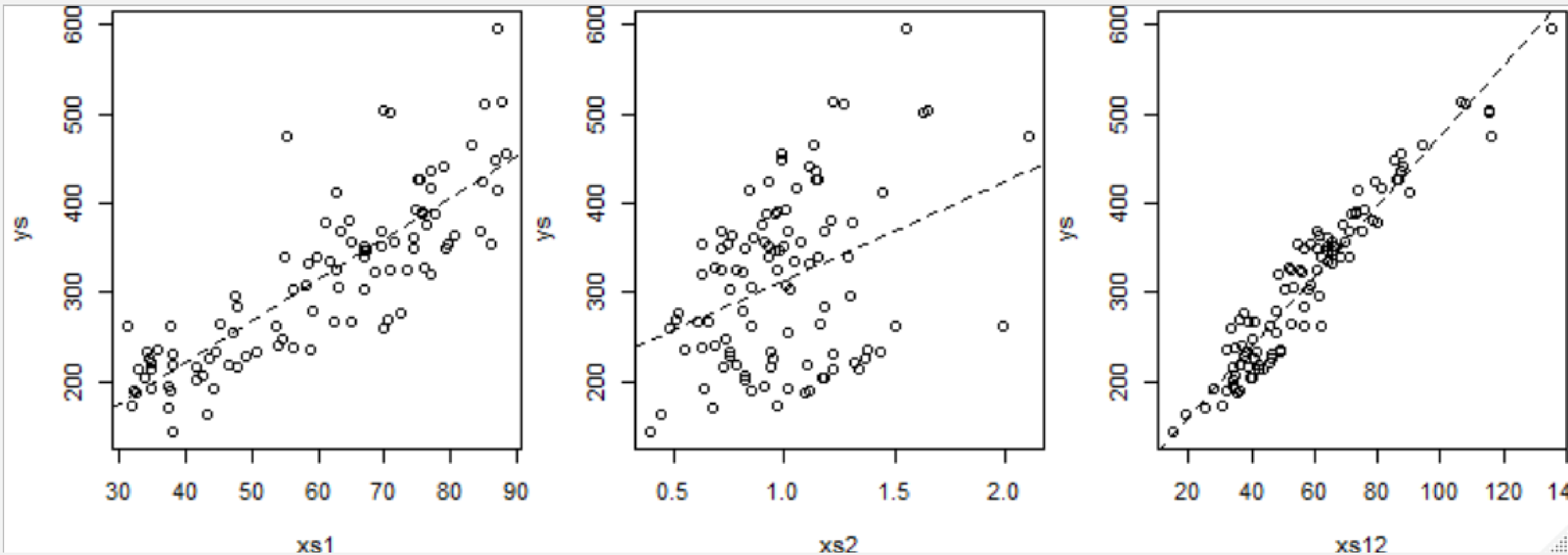
Aside: in fact, we already saw an example of an interaction when we were looking at a particular case of a regression model, the ANCOVA example, in the case of the ANCOVA with different slopes – that is an interaction effect, in which the slope of a relationship – i.e. the effect of one continuous covariate - depends on the level of a factor.

Now we can make it even more general if two or more quantitative covariates influence the response, but that response is dependent of the value of another covariate…

```r
#sample size
set.seed(121)
n=100
#get a response variable
xs1=runif(n,30,90)
#get a second variable
xs2=rgamma(n,10,10)
#define the linear predictor
ys=20+2*xs1-4*xs2+3*xs1*xs2+rnorm(n,2)

#to make it easier
xs12=xs1*xs2
```

```r
#ploting the data and partial models
par(mfrow=c(1,3),mar=c(4,4,0.2,0.2))
plot(xs1,ys)
abline(lm(ys~xs1),lty=2)
plot(xs2,ys)
abline(lm(ys~xs2),lty=2)
plot(xs12,ys)
abline(lm(ys~xs12),lty=2)
```



We can see that each of the variables per se, xs1, xs2 or their product xs12 (i.e. their interaction) could be relevant to explain the response variable! But… hey… xs2 seems to have a **positive** effect in the response!!!

The model implemented

Note these are estimates of true values of 20, 2, -4, 3

```
> summary(m1)

Call:
lm(formula = ys ~ xs1 + xs2 + xs1:xs2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7590 -0.6134  0.1005  0.6850  2.5857 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.21760    1.22479  18.956  < 2e-16 ***
xs1          1.97682    0.02030  97.368  < 2e-16 ***
xs2         -5.06833    1.09787  -4.617 1.21e-05 ***
xs1:xs2      3.01992    0.01876 160.986  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9462 on 96 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999 
F-statistic: 3.272e+05 on 3 and 96 DF,  p-value: < 2.2e-16
```

All terms are (not surprisingly given the plots) statistically significant!

Parameter standard errors: these depend on sample size and the true variability

Very high correlation – would probably never happen in practice!

Naturally the same happens with the overall test over the significance of regression models

```r
#-----------------------------------
#models with interaction
#model with interaction
m1=lm(ys~xs1+xs2+xs1:xs2)
#just the interaction term
m1B=lm(ys~xs1:xs2)
#same as m1
m1C=lm(ys~xs1*xs2)
#same as just the interaction term
m1D=lm(ys~xs12)
#-----------------------------------
#models without the interaction term
mxs1xs2=lm(ys~xs1+xs2)
mxs1=lm(ys~xs1)
mxs2=lm(ys~xs2)
```

```
> AIC(m1,mxs1,mxs2,mxs1xs2)
        df       AIC
m1       5  278.6349
mxs1     3 1076.8927
mxs2     3 1183.9971
mxs1xs2  4  836.8333
```

```
> summary(mxs2)

Call:
lm(formula = ys ~ xs2)

Residuals:
    Min      1Q  Median      3Q     Max 
-161.03  -80.04   11.24   63.49  221.62 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   201.58      30.27   6.659 1.61e-09 ***
xs2           111.57      29.18   3.824 0.000231 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 88.34 on 98 degrees of freedom
Multiple R-squared:  0.1299,	Adjusted R-squared:  0.121 
F-statistic: 14.62 on 1 and 98 DF,  p-value: 0.0002307
```

Significant **positive** effect on the response…

```
> summary(mxs1xs2)

Call:
lm(formula = ys ~ xs1 + xs2)

Residuals:
    Min      1Q  Median      3Q     Max 
-67.854  -6.369   0.761   7.561  52.010 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -156.08567    8.34374  -18.71   <2e-16 ***
xs1            5.11738    0.09208   55.57   <2e-16 ***
xs2          164.10708    5.20386   31.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.49 on 97 degrees of freedom
Multiple R-squared:  0.9735,	Adjusted R-squared:  0.973 
F-statistic:  1782 on 2 and 97 DF,  p-value: < 2.2e-16
```

Significant **positive** effect on the response…

Failing to include significant interactions could lead one to **wrongly** conclude that a variable that **in reality has a negative effect** on the response happens to have a significant **positive** impact on the response!

Therefore, exploring important interactions is important, and failing to include relevant ones might cause errors (while including spurious ones might also mask some real effects, see next slides!).

This was a 2-way interaction.

One can think about 3-way interactions or even higher order interactions, but no one can interpret those models anymore!

```
modK=lm(ys~xs1*xs2*xs3*xs4)
```

```r
#--------------------------------------------
# A 4 way interaction model
#but in reality there is only 1 second order interaction
#--------------------------------------------
set.seed(123)
#get a response variable
xs1=runif(n,30,90)
#get a second variable
xs2=rgamma(n,10,10)
#get a response variable
xs3=runif(n,3,6)
#get a second variable
xs4=rgamma(n,4,4)
#define the linear predictor
ys=20+2*xs1-4*xs2+3*xs1*xs2+xs3+xs4+rnorm(n,2)
modK=lm(ys~xs1*xs2*xs3*xs4)
modL=lm(ys~xs1+xs2+xs3+xs4+xs1:xs2)
summary(modK)
summary(modL)
```

```
> summary(modK)

Call:
lm(formula = ys ~ xs1 * xs2 * xs3 * xs4)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5364 -0.6935  0.0393  0.6770  3.2611 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     17.15188    7.52582   2.279   0.0233 *  
xs1              2.06364    0.11697  17.643   <2e-16 ***
xs2             -0.76162    7.13759  -0.107   0.9151    
xs3              2.44263    1.70557   1.432   0.1530    
xs4              5.43081    7.56823   0.718   0.4735    
xs1:xs2          2.96404    0.11227  26.401   <2e-16 ***
xs1:xs3         -0.02019    0.02628  -0.768   0.4429    
xs2:xs3         -1.06330    1.62602  -0.654   0.5136    
xs1:xs4         -0.04437    0.11307  -0.392   0.6950    
xs2:xs4         -3.36009    7.25366  -0.463   0.6435    
xs3:xs4         -1.36135    1.77582  -0.767   0.4438    
xs1:xs2:xs3      0.01356    0.02542   0.533   0.5941    
xs1:xs2:xs4      0.02494    0.10797   0.231   0.8175    
xs1:xs3:xs4      0.01629    0.02604   0.625   0.5321    
xs2:xs3:xs4      1.08330    1.71369   0.632   0.5277    
xs1:xs2:xs3:xs4 -0.01102    0.02513  -0.438   0.6613    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9978 on 349 degrees of freedom
Multiple R-squared:  0.9999,	Adjusted R-squared:  0.9999 
F-statistic: 2.422e+05 on 15 and 349 DF,  p-value: < 2.2e-16
```

Type II errors

Including interactions which are not real can mask the true influence of relevant variables!!

```
> summary(modL)

Call:
lm(formula = ys ~ xs1 + xs2 + xs3 + xs4 + xs1:xs2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.77396 -0.67394  0.02921  0.72956  3.10716 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21.899114   0.697442  31.399  < 2e-16 ***
xs1          2.000077   0.010031 199.380  < 2e-16 ***
xs2         -4.188672   0.616725  -6.792 4.61e-11 ***
xs3          1.024842   0.059860  17.121  < 2e-16 ***
xs4          1.105140   0.101134  10.927  < 2e-16 ***
xs1:xs2      3.001929   0.009804 306.204  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9907 on 359 degrees of freedom
Multiple R-squared:  0.9999,    Adjusted R-squared:  0.9999 
F-statistic: 7.372e+05 on 5 and 359 DF,  p-value: < 2.2e-16
```

```
> AIC(modL,modK)
      df      AIC
modL   7 1036.922
modK  17 1051.885
```