Modelação Ecológica

AULA 10

16 November 2019 – 14:00-16:30 – room 2.3.37

Tiago A. Marques

Journal of Animal Ecology
BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE | Full Access

Direct and indirect effects of regional and local climatic factors on trophic interactions in the Arctic tundra

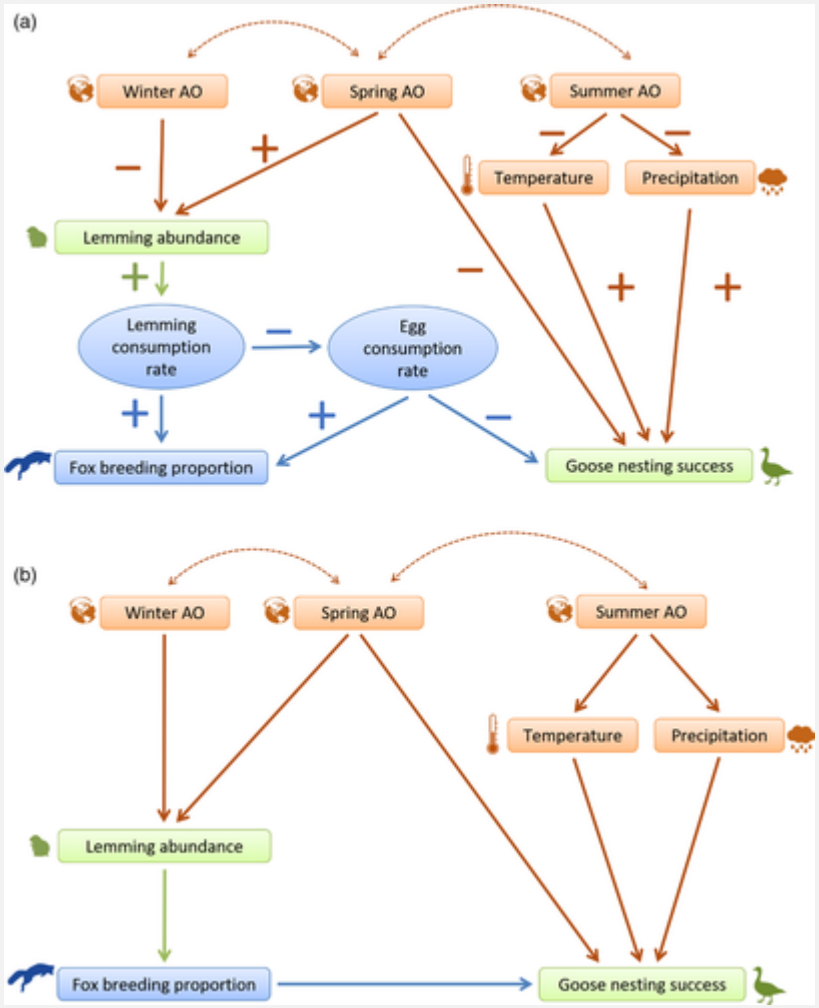
Claire-Cécile Juhasz, Bill Shipley, Gilles Gauthier, Nicolas Lecomte

First published: 20 September 2019 | https://doi.org/10.1111/1365-2656.13104

Texto Integral @ b-on

Gilles Gauthier and Nicolas Lecomte are senior authors.

Conceptualizing models helps to understand what one needs to do but also to explain to others what one has done!



Journal of Animal Ecology
BRITISH ECOLOGICAL SOCIETY

BIOLOGGING | Full Access

Identifying stationary phases in multivariate time series for highlighting behavioural modes and home range settlements

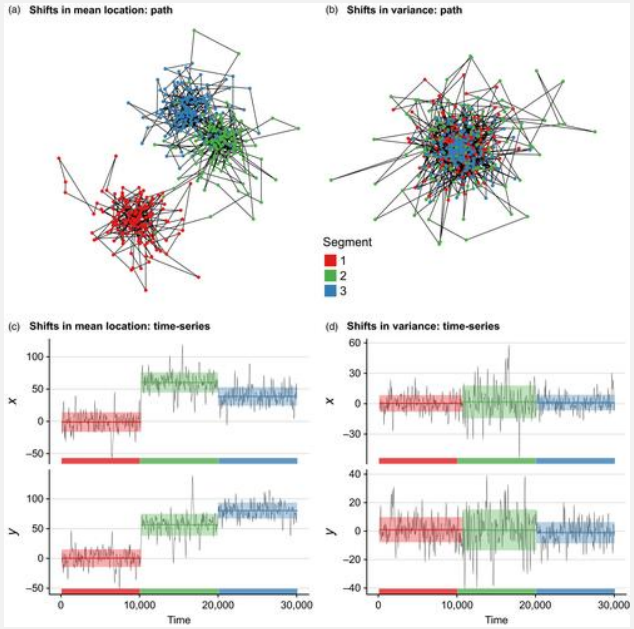
Rémi Patin, Marie-Pierre Etienne, Emilie Lebarbier, Simon Chamaillé-Jammes, Simon Benhamou

First published: 20 September 2019 | https://doi.org/10.1111/1365-2656.13105 | Cited by: 1

Texto Integral @ b-on

Abstract

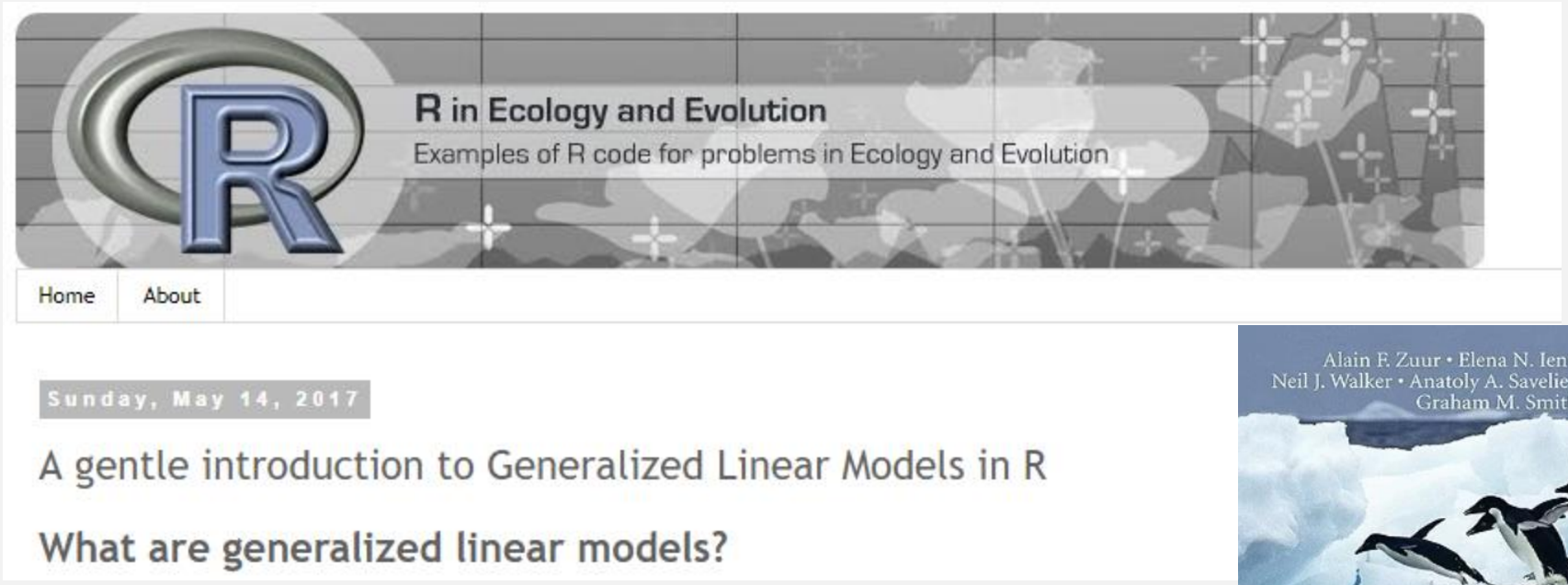
1. Recent advances in biologging open promising perspectives in the study of animal movements at numerous scales. It is now possible to record time series of animal locations and ancillary data (e.g. activity level derived from on-board accelerometers) over extended areas and long durations with a high spatial and temporal resolution. Such time series are often piecewise stationary, as the animal may alternate between different stationary phases (i.e. characterized by a specific mean and variance of some key parameter for limited periods). Identifying when these phases start and end is a critical first step to understand the dynamics of the underlying movement processes.
2. We introduce a new segmentation-clustering method we called segclust2d (available as a R package at cran.r-project.org/package=segclust2d). It can segment bivariate (or more generally multivariate) time series and possibly cluster the various segments obtained, corresponding to different phases assumed to be stationary. This method is easy to use, as it only requires specifying a minimum segment length (to prevent over-segmentation), based on biological rather than statistical considerations.

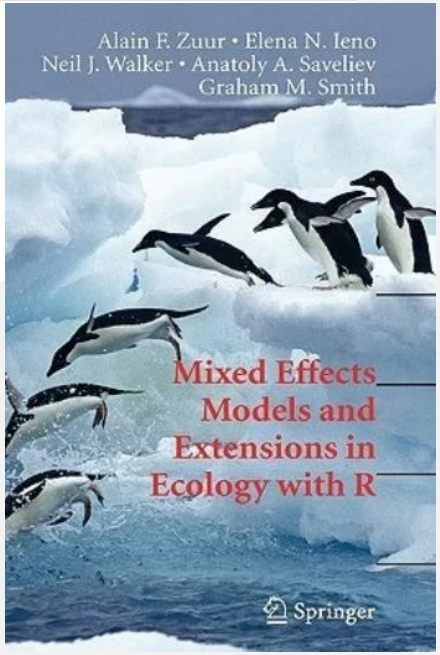


Theorethical-Rpackage-paper-20% worth work:
A new package to compare with the others I mentioned before about animal movement

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

How to choose the distribution family in a GLM (or GAM): it's related to the type of response data involved, particularly related to the values that the **response variable** can take.

- Continous data: Gaussian
- Continous data, strictly positive or with variance proportional to mean: Gamma
- Count data: Poisson
- Overdispersed count data (var>mean) : Negative Binomial
- Presence/Absence data: Binomial
- Underdispersed count data (mean>var, rare): Binomial
- Number of successes in *n* trials: Binomial

There are other even more general families, like Quasi-Poisson and Quasi-Binomial, or the Tweedie distribution (several of the above are special cases of the Tweedie).

Ver Hoef, J. M. & Boveng, P. L. 2007 Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology* **88**: 2766-2772

http://ugrad.stat.ubc.ca/R/library/statmod/html/tweedie.html

tweedie {statmod}

Tweedie Generalized Linear Models

Description

Produces a generalized linear model family object with any power variance function and any power link. Includes the Gaussian, Poisson, gamma and inverse-Gaussian families as special cases.

Usage

```
tweedie(var.power=0, link.power=1-var.power)
```

The variance power p characterizes the distribution of the responses y. The following are some special cases:

p	Response distribution
0	Normal
1	Poisson
(1, 2)	Compound Poisson, non-negative with mass at zero
2	Gamma
3	Inverse-Gaussian
> 2	Stable, with support on the positive reals

The name Tweedie has been associated with this family by Jørgensen in honour of M. C. K. Tweedie.

Examples

```
y <- rgamma(20,shape=5)
x <- 1:20
# Fit a poisson generalized linear model with identity link
glm(y~x,family=tweedie(var.power=1,link.power=1))

# Fit an inverse-Gaussion glm with log-link
glm(y~x,family=tweedie(var.power=3,link.power=0))
```

Implementing a GLM

The sequence of procedures to implement a GLM is, in general, relatively straightforward:

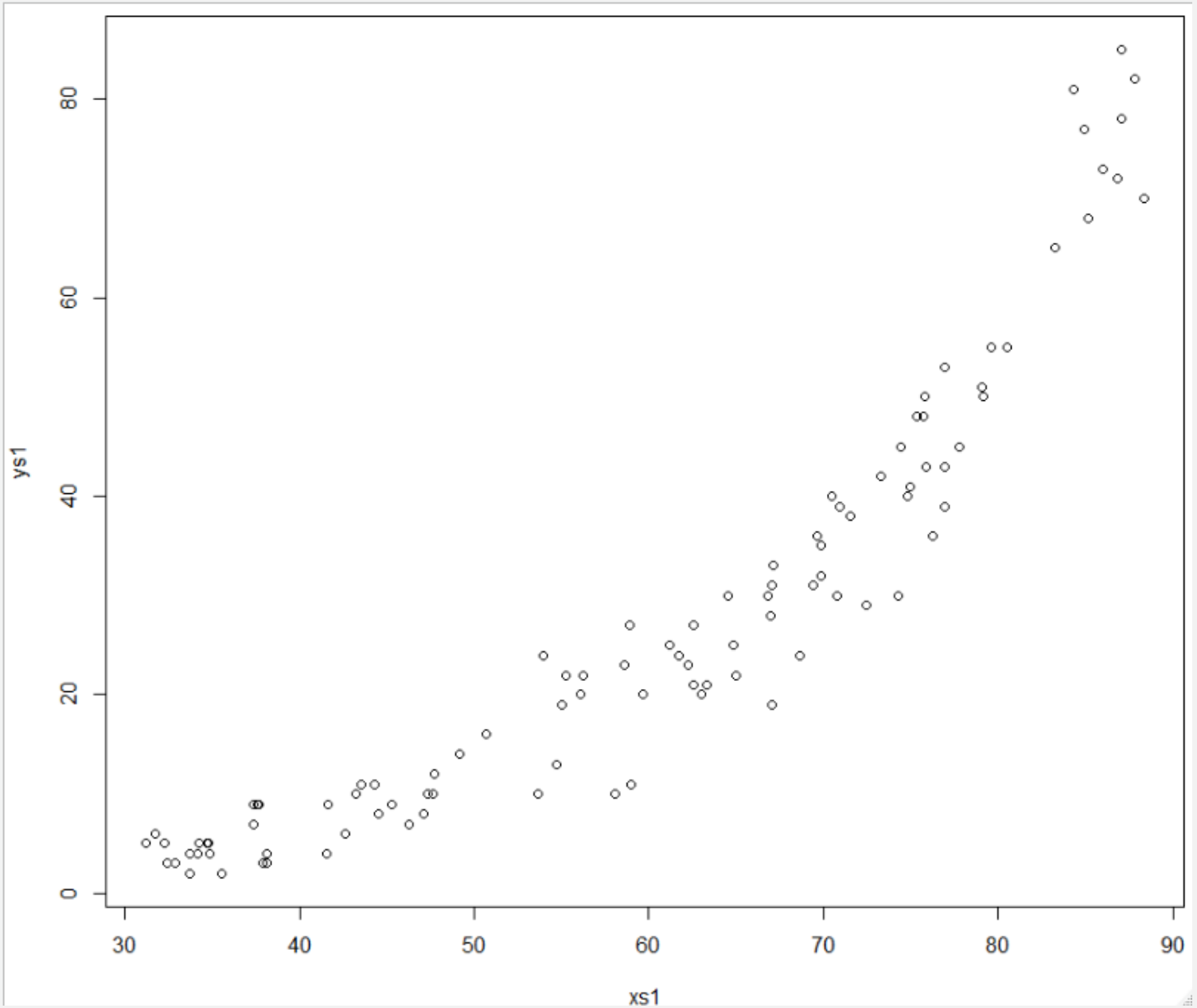
- Choose the specific GLM to consider
  - Distribution family
  - Link function
- Exploratory analysis and selection of independent variables
- Fit model and possibly sub-models
- Comparing sub-models
- Select final model for inferences
- Validation, Inferences, prediction

An example GLM

```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
lp1=0.01+0.05*xs1
Eys1=exp(lp1)

#get the response - a variable with distribution Poisson and mean value given by the linear predictor
ys1=rpois(n,Eys1)
```

The code above simulates and plots data that really comes from a model that is a Poisson GLM!

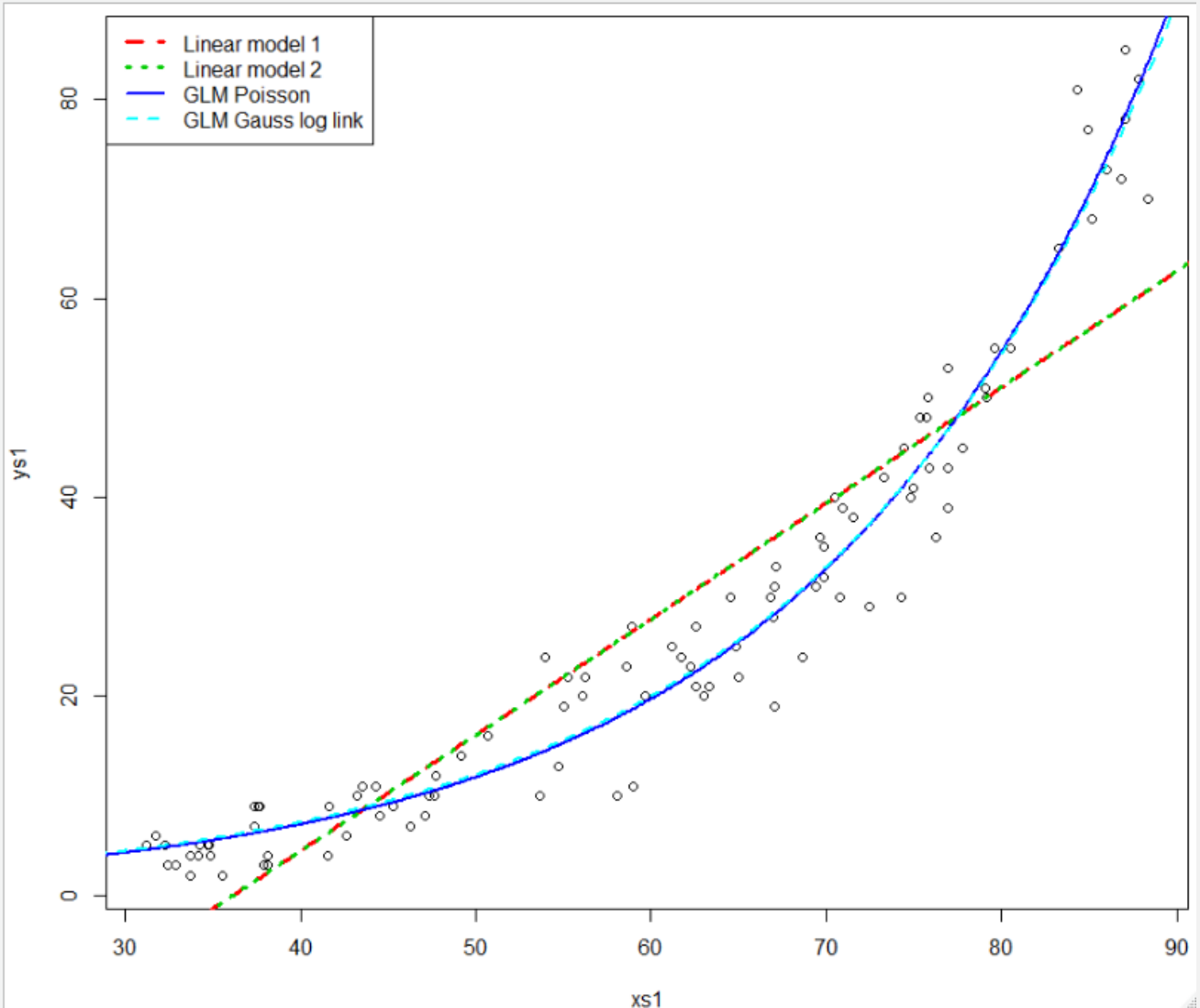
(look in FENIX for file "A10code.R" under file Aula10 16 10 2019 …. am I nice or what ?)



OK, now… lets fit models to the data!

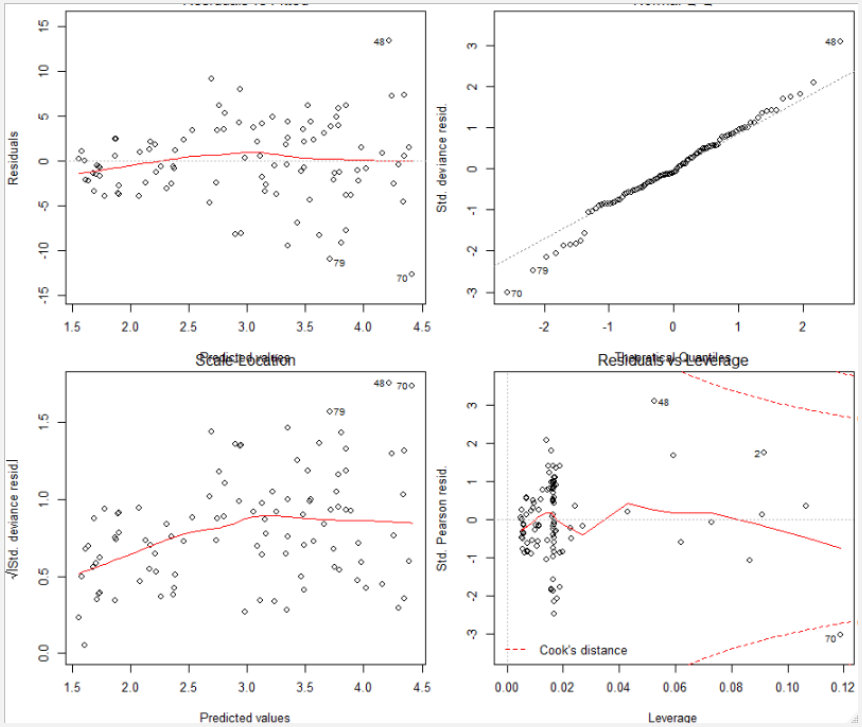
```r
#plot the data - then add fits
par(mfrow=c(1,1),mar=c(4,4,0.5,0.5))
plot(xs1,ys1)
#plot predictions
#get the new data object to predict at
seq.2.pred.at=seq(25,95,by=0.1)
novosdados=data.frame(xs1=seq.2.pred.at)
#linear model, version 1
preds=predict(lm1A,newdata = novosdados)
lines(seq.2.pred.at,preds,col=2,lty=2,lwd=3)
#linear model,version 2
preds=predict.glm(lm1B,newdata = novosdados)
lines(seq.2.pred.at,preds,col=3,lty=3,lwd=3)
#glm Poisson
preds=predict.glm(glmPoi1,newdata = novosdados,type = "response")
lines(seq.2.pred.at,preds,col=4,lwd=2)
#glm Gaussian with log link
preds=predict.glm(glmGau,newdata = novosdados,type = "response")
lines(seq.2.pred.at,preds,col=5,lwd=2,lty=2)
#the legend
legend("topleft",legend=c("Linear model 1","Linear model 2",
"GLM Poisson","GLM Gauss log link"),lwd=c(3,3,2,2),col=c(2,3,4,5),lty=c(2,3,1,2))
```

```r
#fit a linear model
lm1A=lm(ys1~xs1)
lm1B=glm(ys1~xs1)
#fit a Poisson regression
glmPoi1=glm(ys1~xs1,family=poisson(link="log"))
#fit a glm with a Gaussian but log link
glmGau=glm(ys1~xs1,family=gaussian(link="log"))
```

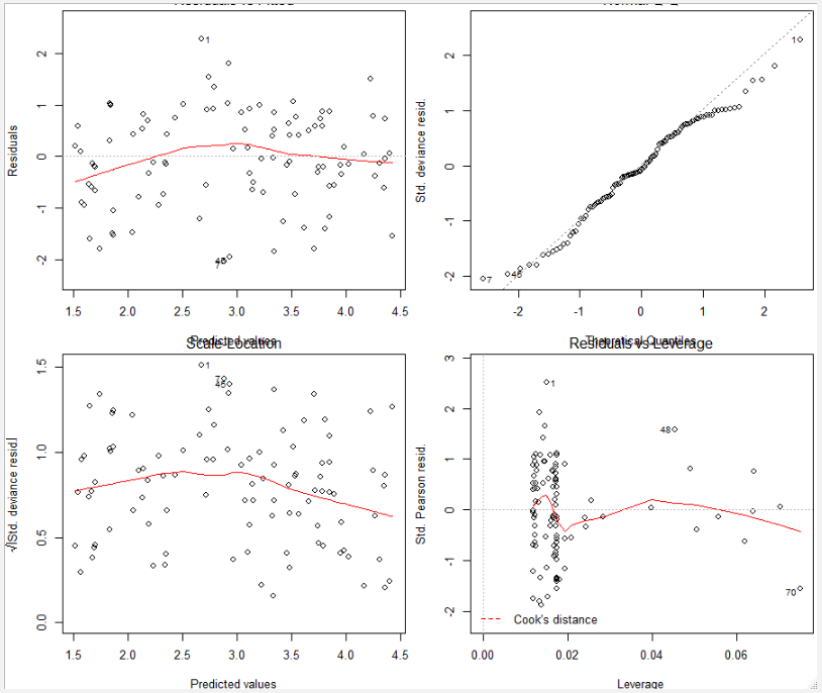


```
> #just for fun, now see that
> AIC(lm1A,lm1B,glmGau,glmPoi1)
        df      AIC
lm1A     3 718.4863
lm1B     3 718.4863
glmGau   3 586.4525
glmPoi1  2 564.2675
```

```
par(mfrow=c(2,2),mar=c(4,4,0.5,0.5))
plot(glmGau)
```




```
par(mfrow=c(2,2),mar=c(4,4,0.5,0.5))
plot(glmPoi1)
```



This is already about… model selection!

(stay tuned!)

MODEL ASSESSMENT
&
MODEL SELECTION

"Data don't make any sense,
we will have to resort to statistics."

Regression & GLM

We can compare different models based on several sets of tools:

	Advantages	Disadvantages
Goodness-of-fit,* R^2, *Deviance	Represents explained variance	Can be very dependent on under or over fitting
Hypothesis tests	Hierarchical structure	Reduced robustness and power; requires nested models
Information theory AIC, BIC	Looks for parsimony; balance complexity and fit	Generic / black box
Cross-validation	Assess prediction ability	Ad hoc / Computationally intensive

Example: *leave-one-out* cross validation: for each observation (1) we remove it from the sample, (2) re-fit the model (3) predict the observation that was left out (4) look at overall prediction errors

Regression & GLM

The deviance ... (em PT a desviância… say *whati* ?!?)

D = -2(LL proposed model – LL saturated model)
= 2(LL saturated model – LL proposed model)

is 2 times the log of the likelihood (a.k.a. *log-likelihood*, LL) ratio of **a model** compared to the corresponding **saturated model** (i.e. with as many parameters as observations!).

This difference reflects the quality of the fit

In GLMs we often compare the residual deviance with the model's deviance.

Regression & GLM

Null Deviance = 2(LL(Saturated Model) - LL(Null Model))
df = df_Sat - df_Null

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model))
df = df_Sat - df_Res

Saturated Model – model with as many parameters as observations.

Null Model – a single parameter model (i.e. a global mean).

Proposed Model – typically, it has *p* parameters (1 intercept + *p*-1 independent variables).

Regression & GLM

Null Deviance = 2(LL(Saturated Model) - LL(Null Model))

A small Null Deviance means that a model with a single parameter is good enough to explain the data

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model))

If the Residual Deviance is small, then the proposed model is a good descriptor of the data

H0: the two models (1 parameter vs. p parameter model) are equivalent

The difference between deviances is tested formally using a test statistic that follows a qui-squared distribution

(Null Deviance - Residual Deviance) has a qui-squared distribution with p degrees of freedom, under H0

Null Deviance = 2(LL(Saturated Model) - LL(Null Model)) df = N-1

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model)) df = N-p

Residual Deviance should be small (compared to the Null Deviance)

Null Deviance - Residual Deviance $\longrightarrow$ χ^2 df = (N-1)-(N-p)=p-1

Deviance test H0: simpler model is enough to describe the data

Null Deviance = 2(LL(Saturated Model) - LL(Null Model))

```
> summary(glmPoi1)

Call:
glm(formula = ys1 ~ xs1, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0385  -0.6290  -0.0771   0.6987   2.2716

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.071850   0.104375  -0.688    0.491
xs1          0.050922   0.001422  35.804   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1679.120  on 99  degrees of freedom
Residual deviance:   82.956  on 98  degrees of freedom
AIC: 564.27

Number of Fisher Scoring iterations: 4
```

Residual Deviance = 2(LL(Saturated Model) - LL(Proposed Model))

```
> glmPoi1s=summary(glmPoi1)
> 1-pchisq(glmPoi1s$null.deviance-glmPoi1s$deviance,glmPoi1s$df.null-glmPoi1s$df.residual)
[1] 0
```

AIC - Akaike Information Criterion (in R: AIC)

It is a measure of relative model fit (i.e. of parsimony), reflecting the quality of the fit to a given data set

AIC= 2k - 2LL (k=number of parameters, LL log likelihood)

One component penalizes complexity, the other evaluate how good the fit is

It's a good tool for model selection: the lower the AIC, the better the model

The AIC is based on Information theory – it provides a measure of the relative amount of information (contained in the data) is lost when a given model is used

AIC is not an absolute measure of fit, and therefore it is only good to compare between models – it provides no evidence about whether the best model of the set is really any good!

Aho et al. 2014 Model selection for ecologists: the worldviews of AIC and BIC *Ecology* **95**: 631-636

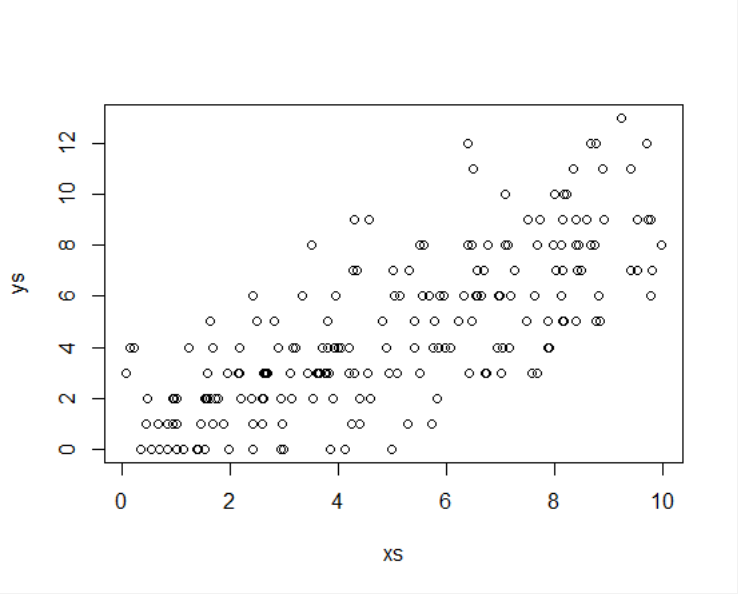
Model selection

Some possibilities:

- Forward selection
- Backward elimination
- Stepwise methods (also can be used using information theory methods e.g. `step AIC`)

A GLM example

```
> #creating data for a glm
> set.seed(12)
> #define the covariate
> xs=runif(200,0,10)
> #a covariate unrelated to the response
> zs=rnorm(200)
> #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 - good covariate
> glm1=glm(ys~xs,family=poisson(link=log))
> #fit model - useless covariate
> glm2=glm(ys~zs,family=poisson(link=log))
> #fit model - both covariates
> glm3=glm(ys~xs+zs,family=poisson(link=log))
```




```
> summary(glm1)

Call:
glm(formula = ys ~ xs, family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.85913  -0.83220  -0.08998   0.55724   2.46350

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.46171    0.08828    5.23  1.7e-07 ***
xs           0.18995    0.01284   14.80  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 464.28 on 199  degrees of freedom
Residual deviance: 228.19 on 198  degrees of freedom
AIC: 845.19

Number of Fisher Scoring iterations: 5
```

```
> summary(glm2)

Call:
glm(formula = ys ~ zs, family = poisson(link = log))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1282  -1.3700  -0.3197   0.9795   3.0871

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.55037    0.03258  47.589   <2e-16 ***
zs          -0.01926    0.03254  -0.592    0.554
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 464.28  on 199  degrees of freedom
Residual deviance: 463.93  on 198  degrees of freedom
AIC: 1080.9

Number of Fisher Scoring iterations: 5
```

```
> AIC(glm1,glm2)
     df       AIC
glm1  2  845.1904
glm2  2 1080.9384
```

```
> anova(glm1,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: ys

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   199     464.28
xs    1    236.1       198     228.19 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 1-pchisq(236.1,1)
[1] 0
```

Note function
anova
over a
glm
object does an
Analysis of Deviance

```
> anova(glm2,test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: ys

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   199     464.28
zs    1  0.34979       198     463.93   0.5542
> 1-pchisq(0.34979,1)
[1] 0.554232
```

Deviance test H0: simpler model is enough to describe the data

```
> anova(glm2,glm3)
Analysis of Deviance Table

Model 1: ys ~ zs
Model 2: ys ~ xs + zs
  Resid. Df Resid. Dev Df Deviance
1       198     463.93
2       197     227.33  1   236.61
> anova(glm1,glm3)
Analysis of Deviance Table

Model 1: ys ~ xs
Model 2: ys ~ xs + zs
  Resid. Df Resid. Dev Df Deviance
1       198     228.19
2       197     227.32  1  0.86061
>
> 1-pchisq(anova(glm2,glm3)$Deviance[2],anova(glm2,glm3)$Df[2])
[1] 0
> 1-pchisq(anova(glm1,glm3)$Deviance[2],anova(glm1,glm3)$Df[2])
[1] 0.3535677
```

reject

do not reject

```
> AIC(glm1,glm2,glm3)
      df       AIC
glm1   2  845.1904
glm2   2 1080.9384
glm3   3  846.3298
```

Regression & GLM

Há diversas metodologias de seleção de modelos.

Frequentemente, os modelos podem ter um número muito elevado de termos, devido às interações entre variáveis explicativas.

As metodologias mais utilizadas são procedimentos iterativos de avaliação multi-etápica, com base em critérios como o AIC ou indicadores equivalentes, uma vez que outros (e.g. R^2, deviance) são extremamente sensíveis em relação ao número de parâmetros e ao número de observações.

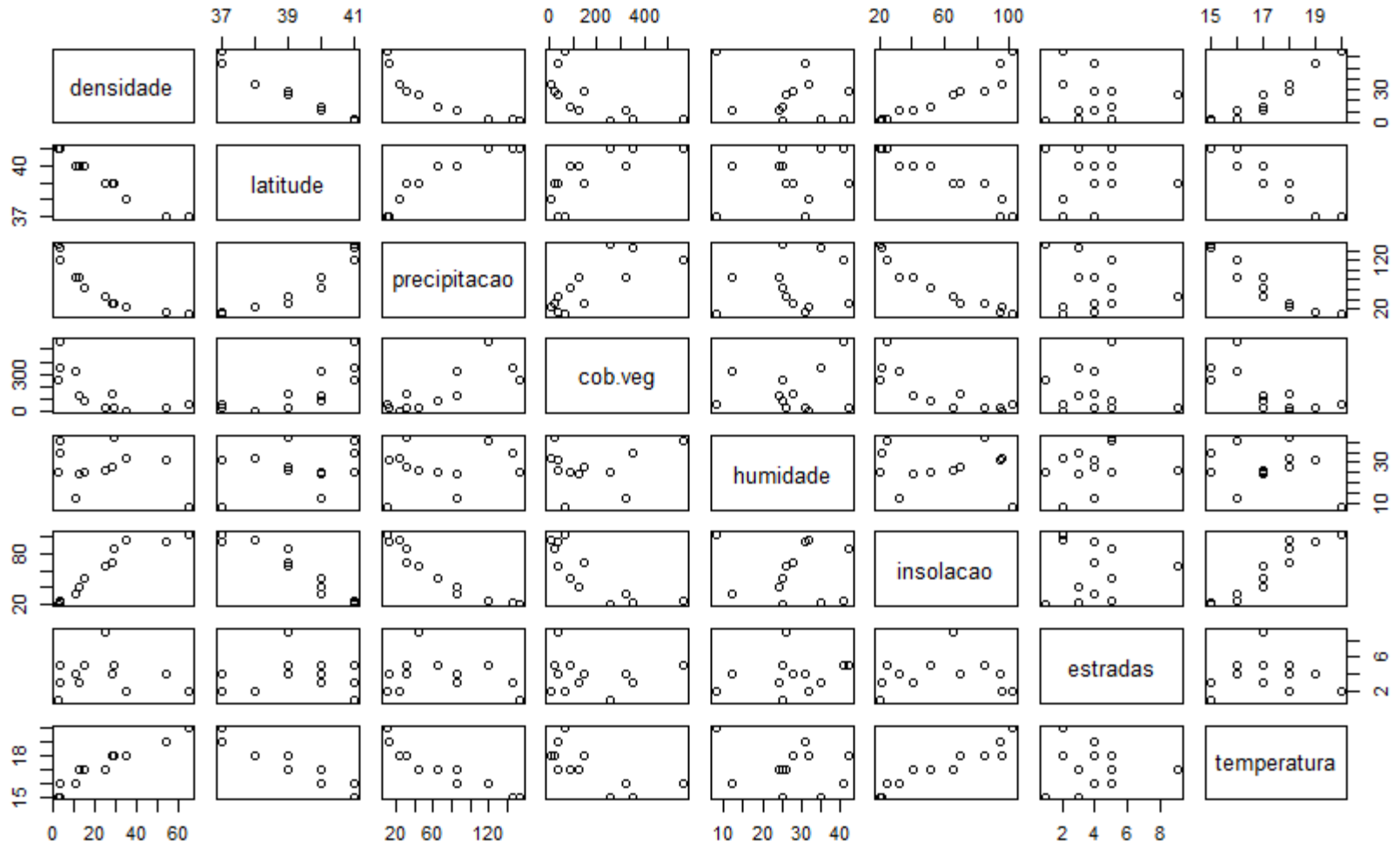
Regression & GLM

	densidade	latitude	precipitacao	cob.veg	humidade	insolacao	estradas	temperatura
1	2	41	152	256	25	21	1	15
2	3	41	145	354	35	22	3	15
3	3	41	120	564	41	25	5	16
4	11	40	86	324	12	32	4	16
5	12	40	85	125	24	41	3	17
6	15	40	65	90	25	51	5	17
7	25	39	45	35	26	65	9	17
8	28	39	32	145	28	70	4	18
9	29	39	31	25	42	85	5	18
10	35	38	24	10	32	95	2	18
11	54	37	12	38	31	94	4	19
12	65	37	10	64	8	102	2	20

Variável resposta, ou dependente, que tentamos explicar à custa das variáveis independentes

Regression & GLM

Antes de começar qualquer exercicio de modelação, explorar as relações entre as variáveis



Regression & GLM

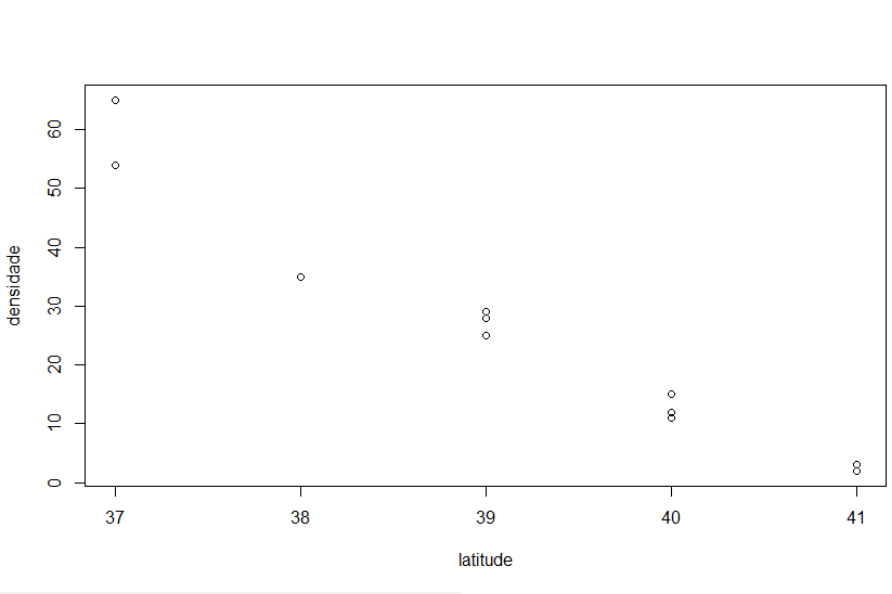
A multiple regression model

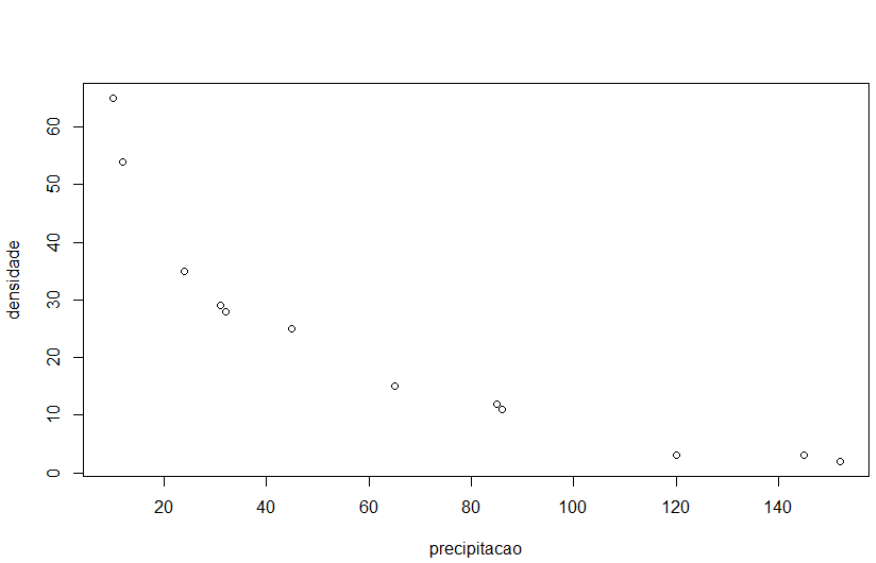
```
Call:
lm(formula = densidade ~ ., data = dens)

Residuals:
      1       2       3       4       5       6       7       8       9      10      11      12
 0.1644  1.9567 -1.6571  0.4803 -2.5667  1.0653 -1.0994  1.7283  1.2474 -2.5452  1.0432  0.1829

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 249.084140 116.048827   2.146   0.0984 .
latitude     -9.617558   2.526905  -3.806   0.0190 *
precipitacao  0.268172   0.081913   3.274   0.0307 *
cob.veg       0.008505   0.009097   0.935   0.4027
humidade     -0.215045   0.122166  -1.760   0.1532
insolacao     0.315824   0.180705   1.748   0.1554
estradas      1.042979   0.572918   1.820   0.1428
temperatura   6.788849   2.134940   3.180   0.0335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.641 on 4 degrees of freedom
Multiple R-squared:  0.9938,    Adjusted R-squared:  0.983
F-statistic: 91.59 on 7 and 4 DF,  p-value: 0.0002996

> |
```

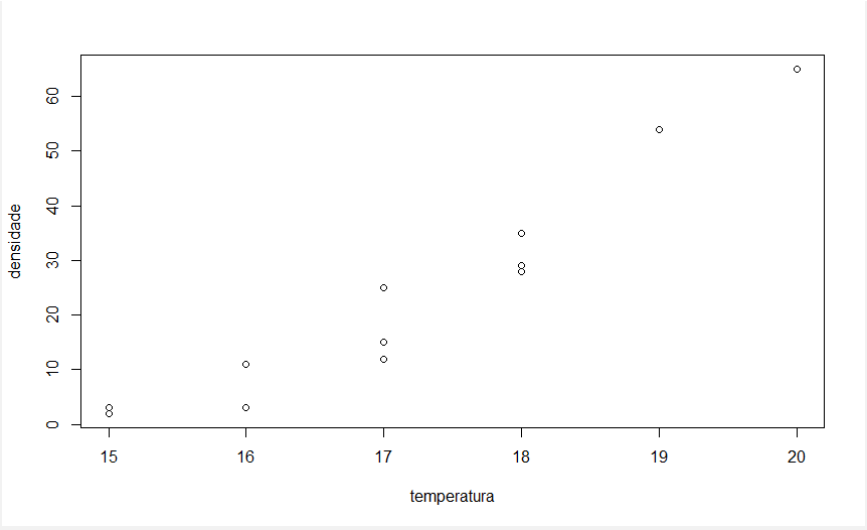
a densidade em função de todas as outras variáveis

Variáveis cujo coeficiente é significativamente diferente de 0, ou seja, que parecem influenciar a variável resposta

Regression & GLM

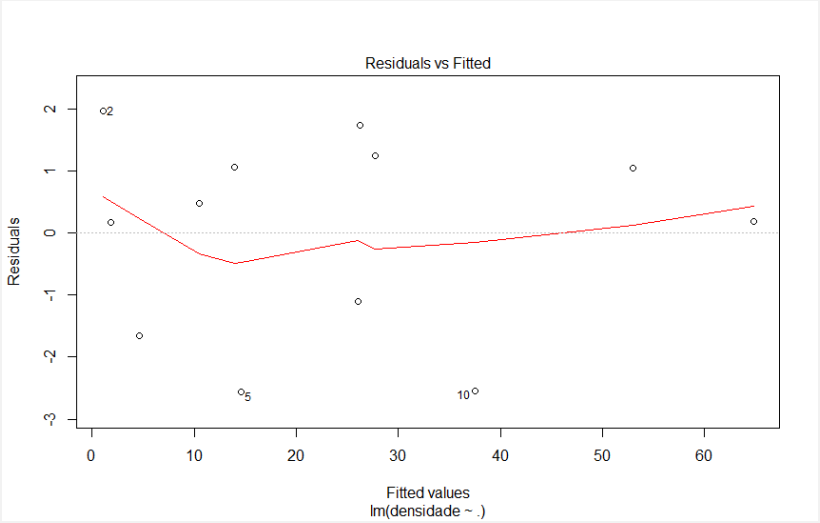


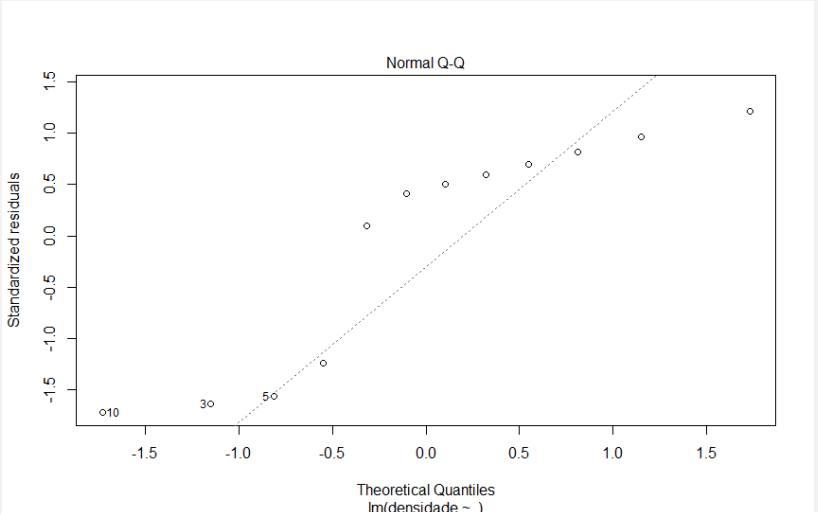


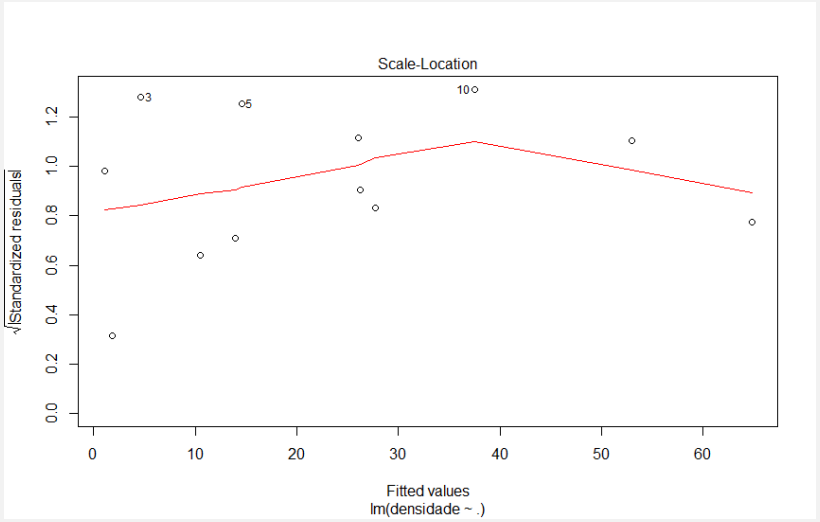


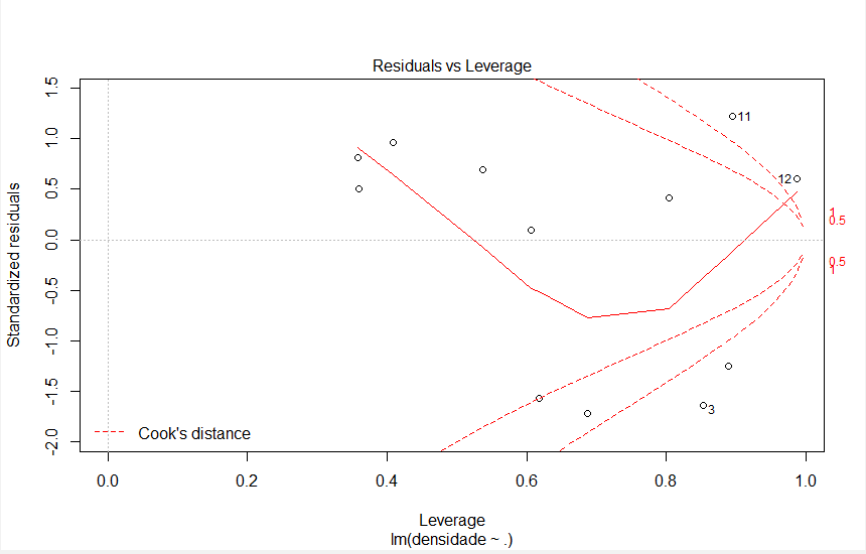
Regression & GLM

Diagnostic plots (obtêm-se fazendo o plot do modelo)

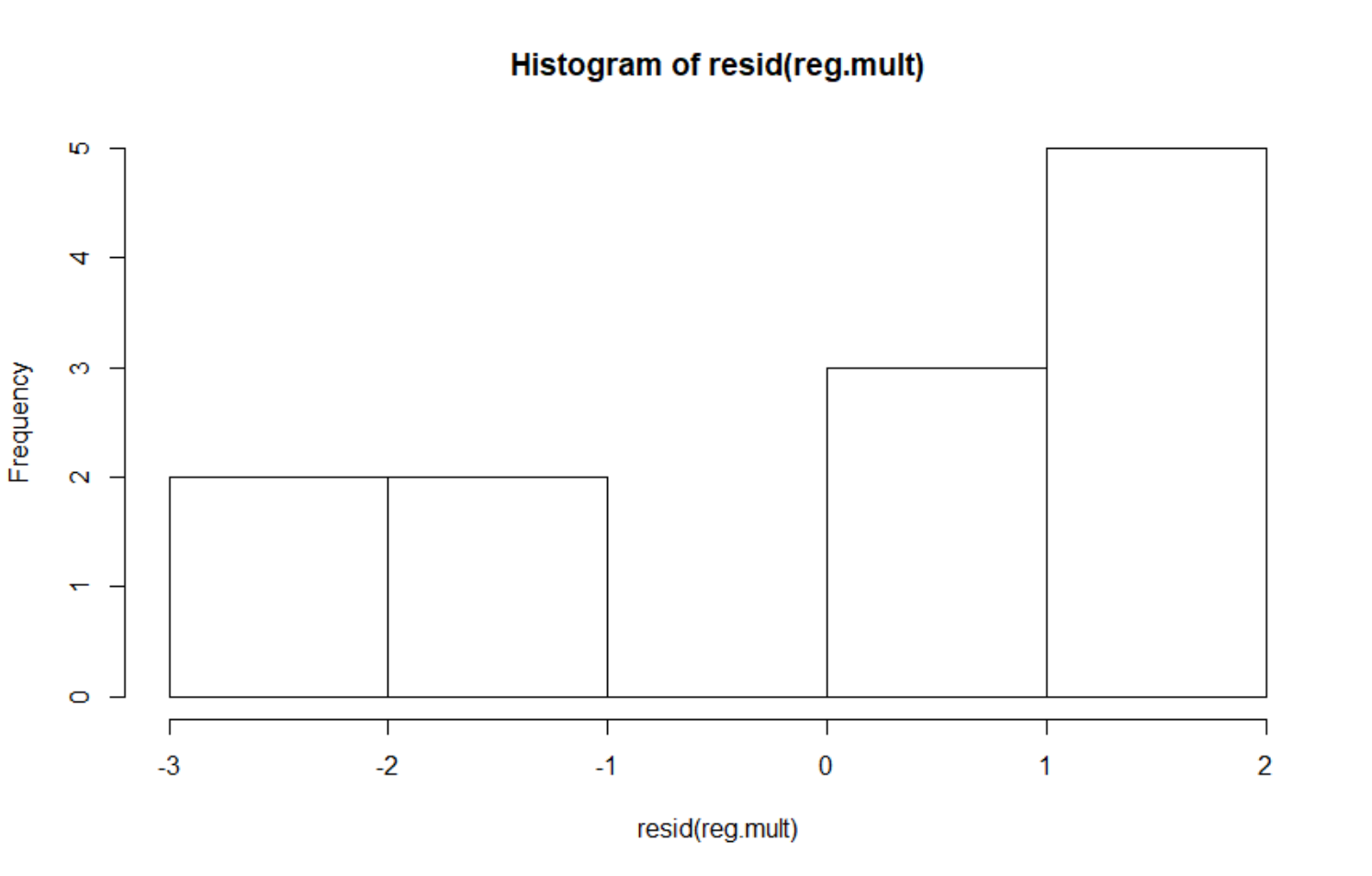








Histograma dos resíduos



regressão e MLG

Same data, now a Poisson GLM

```
Call:
glm(formula = densidade ~ ., family = poisson, data = dens)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10      11      12
-0.3845  0.2802  -0.2190  0.3128  0.1774  -0.2165  -0.0648  -0.1795  0.2760  -0.0831  -0.0069  0.0180

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.294842   7.845068   1.057   0.2904
latitude    -0.130397   0.156261  -0.834   0.4040
precipitacao -0.017535   0.008546  -2.052   0.0402 *
cob.veg     -0.000872   0.001231  -0.708   0.4787
humidade    -0.006591   0.008258  -0.798   0.4248
insolacao   -0.005802   0.013855  -0.419   0.6754
estradas     0.006784   0.046762   0.145   0.8847
temperatura  0.078208   0.182180   0.429   0.6677
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 194.24605  on 11  degrees of freedom
Residual deviance:   0.57038  on  4  degrees of freedom
AIC: 71.006

Number of Fisher Scoring iterations: 4

>
```

regressão e MLG