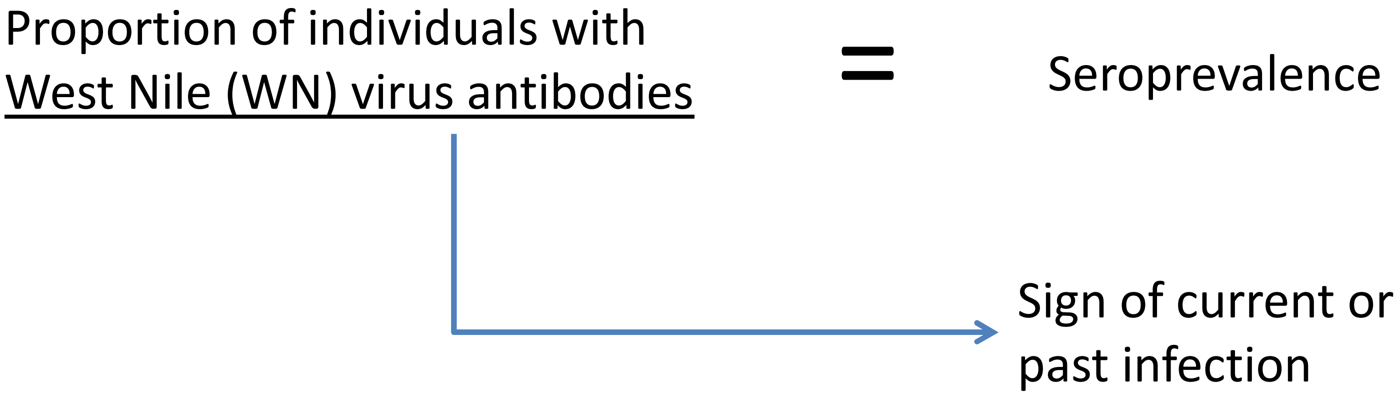
Generalized linear models (GLM)

Dr. Vladimir Grosbois
vladimir.grosbois@cirad.fr

CIRAD
UR AGIRs

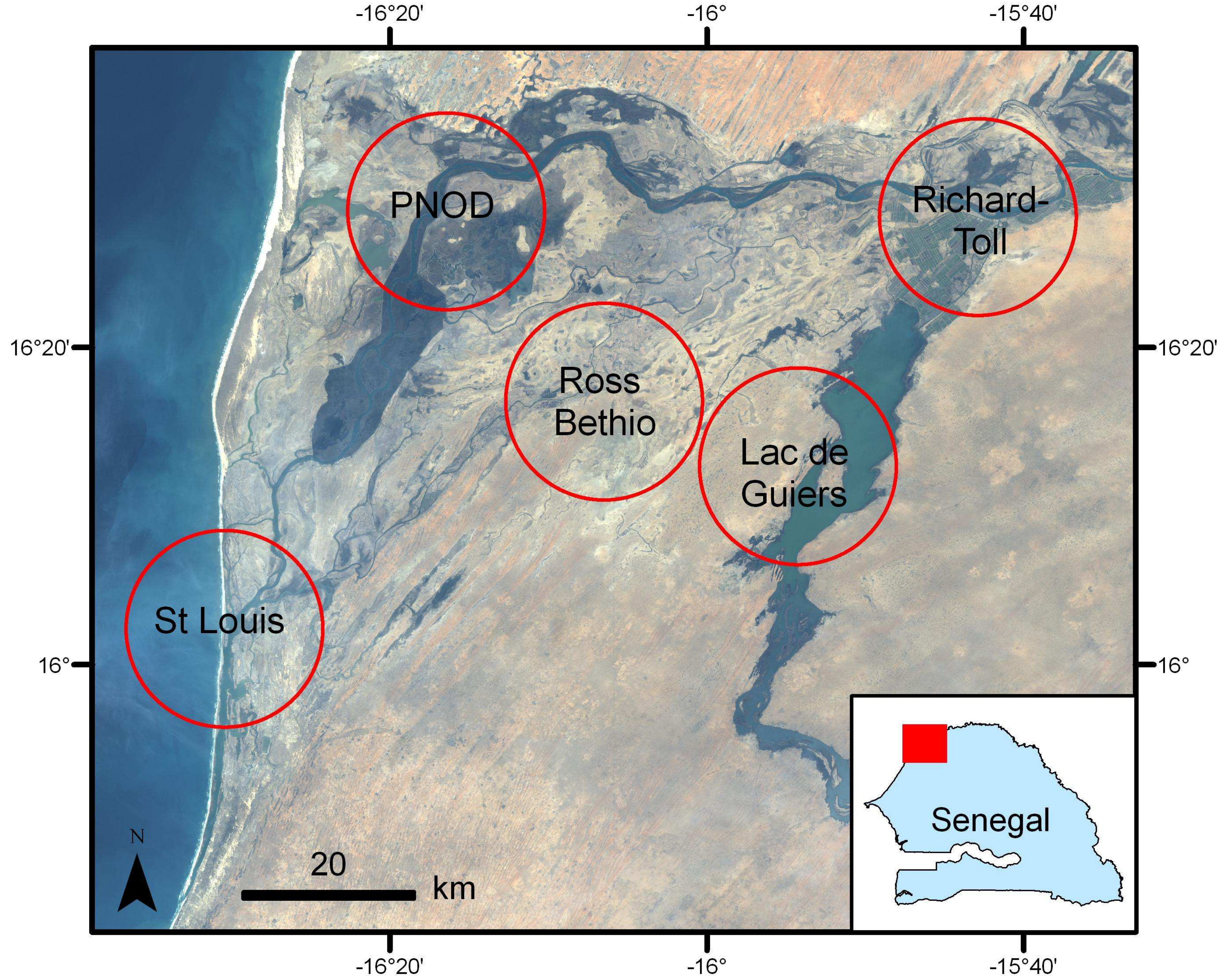
GLM: Application situation

- 2 categories of individuals in a population
- The variable we wish to model is the proportion of one of these two types
- Example: population of horses in Senegal



Proportion of individuals with West Nile (WN) virus antibodies = Seroprevalence

Sign of current or past infection


-16°20'
-16°
-15°40'
16°20'
16°
PNOD
Richard-
Toll
Ross
Bethio
Lac de
Guiers
St Louis
Senegal
N
20
km

The data

One sample of horses from that population (Senegal).
On each sampled horse, blood sample has been taken and WN antibodies have been searched for

Read the data

prevchev<-read.table ("prechev.csv",header=TRUE,sep=",",dec=".")

Look the data

summary(prevchev)

```
AGE              REGION      SALINITE              VILLAGE            POS              TOT
Min.   : 2.000   DJO:32   Min.   :-0.170000   Nguith        : 13   Min.   : 0.000   Min.   : 1.000
1st Qu.: 6.000   NGT:38   1st Qu.:-0.150000   Ross-bethio   : 13   1st Qu.: 1.000   1st Qu.: 1.000
Median : 8.000   RIT:43   Median :-0.090000   Tiguette      : 13   Median : 1.000   Median : 1.000
Mean   : 8.611   ROB:46   Mean   :-0.005657   Débi          : 12   Mean   : 1.586   Mean   : 1.854
3rd Qu.:10.000   STL:39   3rd Qu.:-0.040000   Gohou Mbathie: 12   3rd Qu.: 2.000   3rd Qu.: 2.000
Max.   :24.000            Max.   : 0.450000   Mbodiene      :  9   Max.   :12.000   Max.   :13.000
                                              (Other)       :126
```

132 villages

Data presentation

Look at the first 10 lines

head(prevchev,10)

```
   AGE REGION SALINITE      VILLAGE POS TOT
1    6    NGT    -0.15 Belel mbaye    1   1
2   11    NGT    -0.15 Belli bamdi    1   1
3    4    ROB    -0.17  Bissette 1    1   1
4   10    RIT    -0.09   Campement    1   1
5    7    NGT    -0.15 Darou salam    1   1
6    8    NGT    -0.15 Darou salam    1   1
7   10    NGT    -0.15 Darou salam    2   2
8    2    DJO    -0.04        Débi    2   4
9    3    DJO    -0.04        Débi    2   3
10   4    DJO    -0.04        Débi    4   8
```

Dimensions of the data table

dim(prevchev)

```
198    6
```

→ 6 variables

One statistical unit per age*village class(132 villages)

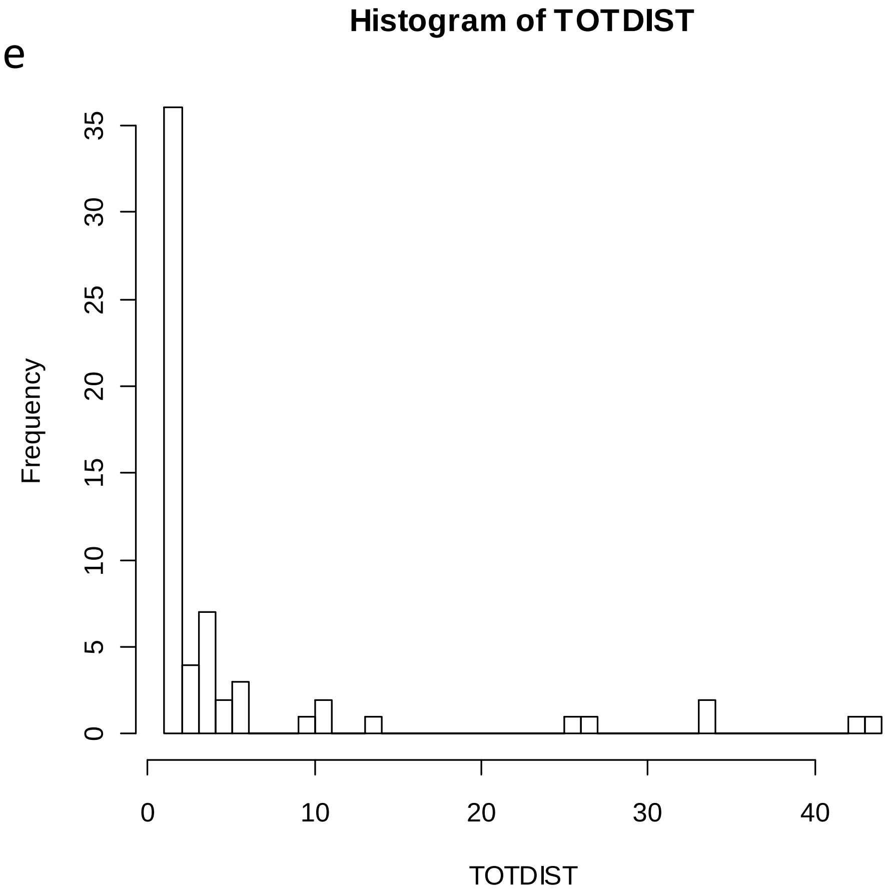
Description of the data: number of individuals sampled

Distribution of the number of horses sampled in a village
! Sometimes, more than one stat unit in a village (age classes)

```
TOTDIST<- tapply(prevchev$TOT,prevchev$VILLAGE,sum)
```

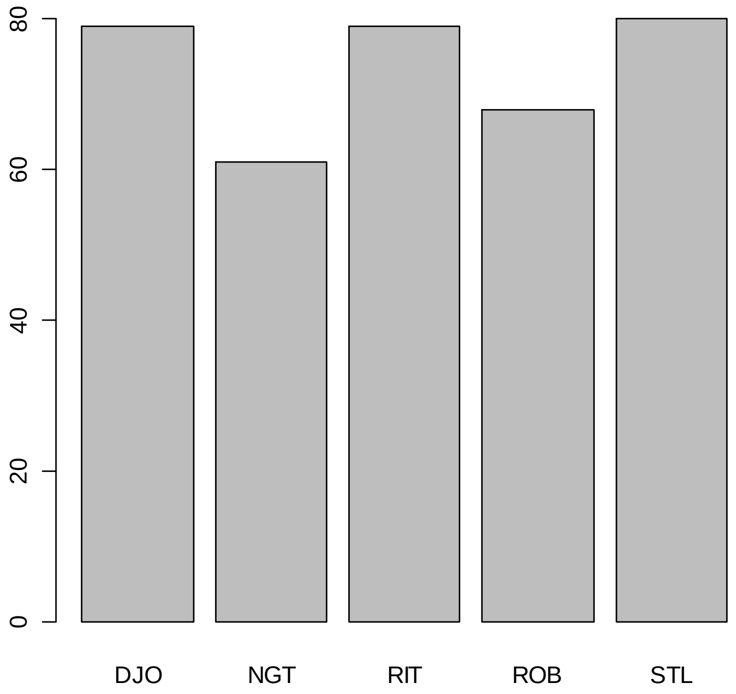
Generates a table that contains the sum of the variable TOT for the each modality of the variable VILLAGE

```
hist(TOTDIST,50)
```



Distribution of the number of horses sampled in each region

```
REGDISTCHEV<-tapply(prevchev$TOT,prevchev$REGION,sum)
 barplot(REGDISTCHEV)
```

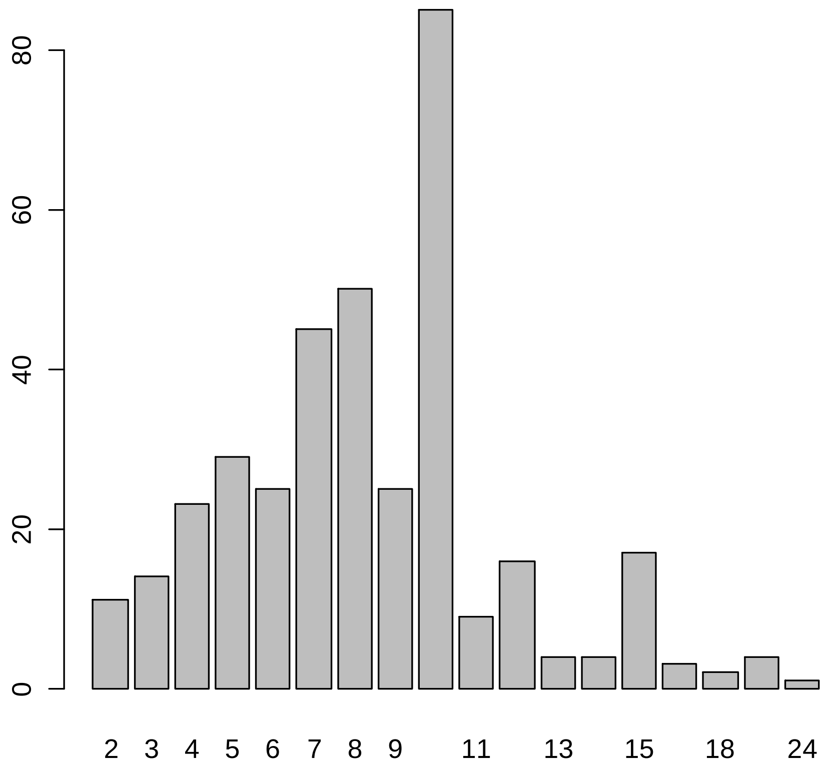


Description of the data: distribution / age

Distribution of the age of sampled horses

```
AGEDIST<- tapply(prevchev$TOT,prevchev$AGE,sum)
```

```
barplot(AGEDIST)
```



Aims of the study

- Estimate the proportion of horses with antibodies
- Determine the influence on the response variable (proportion of seropositive horses) of the explanatory variables:
    - Age
    - Region
    - Salinity

Type of model depend on types of response and explanatory variables

Response	Explanatory	Statistical model
Continuous	All continuous	Linear regression
Continuous	All categorical	Analysis of variance (ANOVA)
Continuous	Continuous and categorical	Analysis of covariance (ANCOVA)
Continuous	Any combination of continuous and/or categorical variables	Linear model (LM)
Categorical, Count, Probability, Proportion	Any combination of continuous and/or categorical variables	Generalized linear models (GLM)

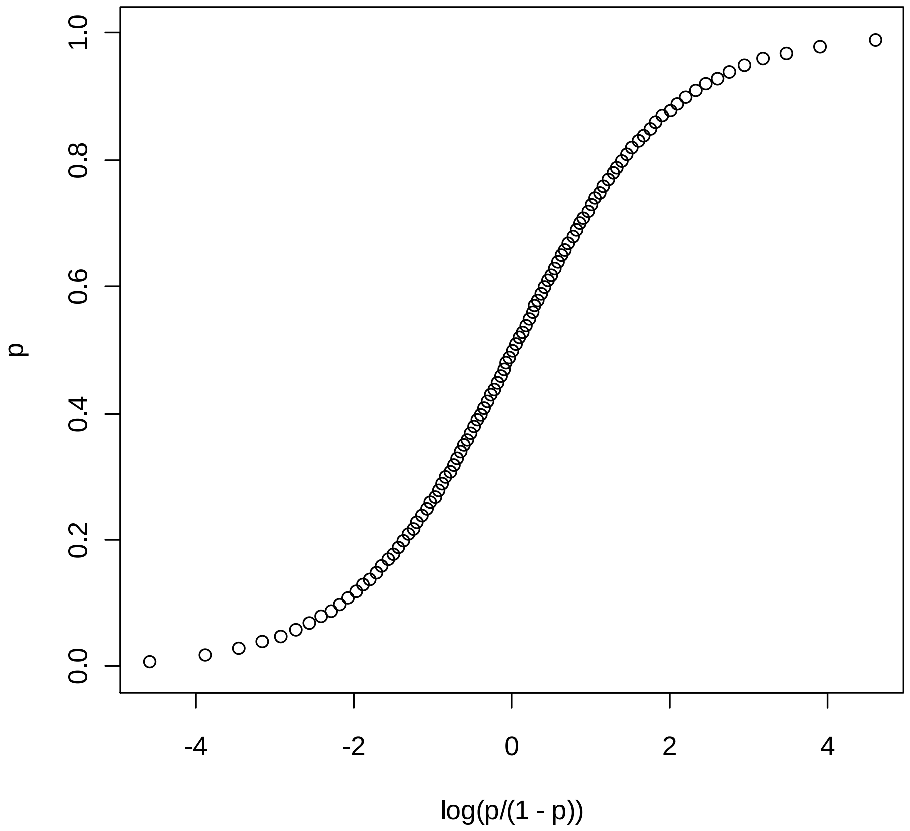
What type of model shall we use?

- A generalized linear model (GLM) characterized by :
  - A link function
  - A distribution law

Which link function ?

- The logit function
  - p is outcome proportion
  - Logit(p) = log(p/(1-p))

```
p=seq(0,1,0.01)
plot(log(p/(1-p)),p)
```

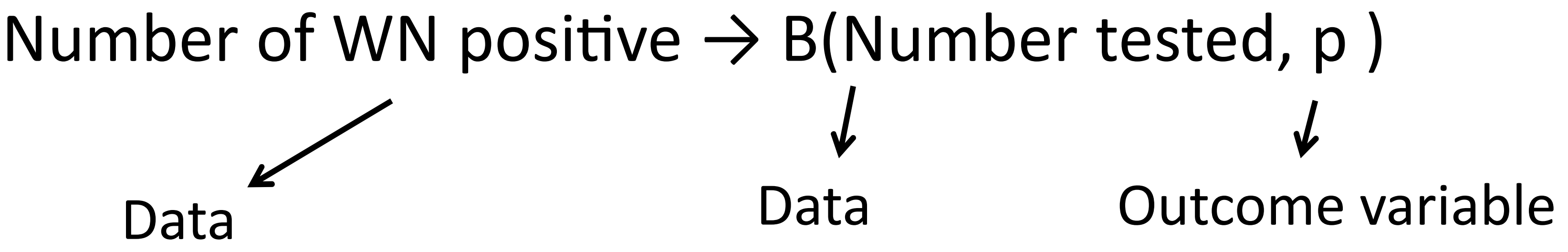


Interpretation of logit

- Logit(p) = log(p/(1-p))
  - Logit is the log of odd
  - Way of expressing probabilities originating from gambling vocabulary
  - A horse with an odd of 25 against 1 = 25 times more likely to loose the race than to win it
  - Scientists use p, gamblers p/(1-p)
  - With the logit function think in terms of log(p/(1-p)

What underlying distribution ?

- The binomial distribution
  - Classical example : numbers of 3s for 6 dice draws: B(6, 1/6)
  - Describe a number of events given:
    - The probability of the event
    - The number of trials (draws)
  - For modelling a proportion:
    - The data include the number of events
    - The data include the number of trials
    - We want to estimate the probability of an event



Number of WN positive → B(Number tested, p)

Data

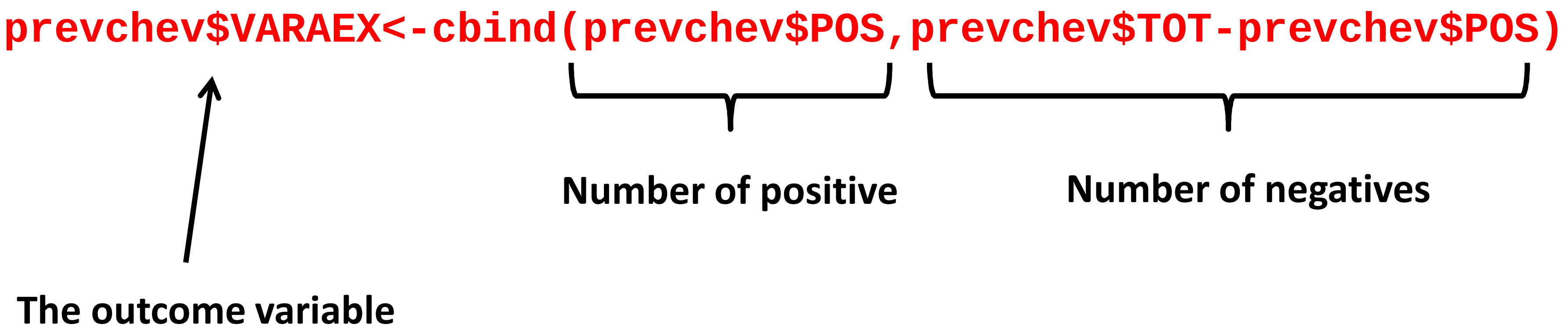
Data

Outcome variable

Outcome variable for a proportion

Outcome variable has 2 components:

- number of positive
- number of négatifs



```
prevchev$VARAEX<-cbind(prevchev$POS,prevchev$TOT-prevchev$POS)
```

Number of positive

Number of negatives

The outcome variable

Model syntax in R

In R the model is usually defined with a function including a formula as one argument

`aov(formula,options)`	Analysis of variance and covariance (ANOVA, ANCOVA)
`lm(formula,options)`	Regressions and linear models
`glm(formula,options)`	Generalized linear models

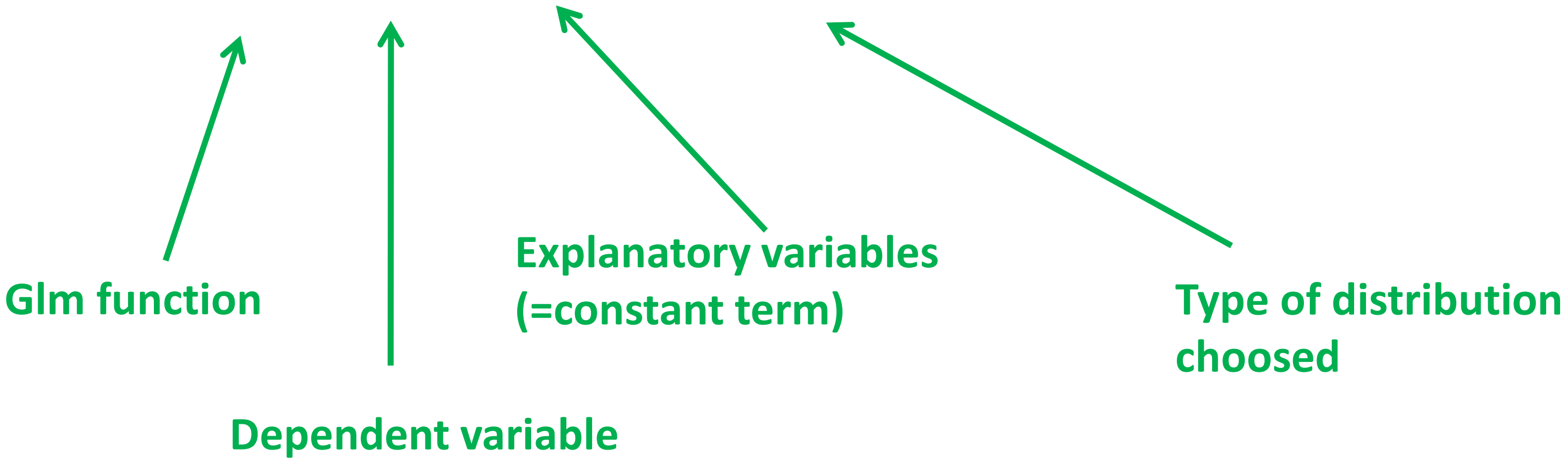
Formula syntax

~	separates response and explanatory variables
+	addition of an explanatory variable and b
a:b	interaction between a and b
a*b	equivalent to a+b+a:b
a/b	b is nest in a

First model: the null model

The null model is the simplest one can build: it considers the proportion as homogeneous

```
mod0<-glm(VARAEX~1, family=binomial, data=prevchev)
```



Glm function

Dependent variable

Explanatory variables (=constant term)

Type of distribution choosed

Interpretation of the outputs

summary(mod0)

```
Call:
glm(formula = VARAEX ~ 1, family = binomial, data = prevchev)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3744   0.5585   0.5585   0.5585   1.4777

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7791     0.1485   11.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 151.35  on 197  degrees of freedom
AIC: 207.93

Number of Fisher Scoring iterations: 4
```

Model description

Not important

Estimation (logit scale) and test of H_0: estimation = 0

Model fit: Deviance should not exceed the d° of freedom

How is the model fitted ?

Maximum likelihood method:

- Determine the value of the parameter that maximises the probability of the data

- Given the structure of the model (*i.e.* considering that the proportion WN positive individuals is homogeneous)

Interpretation of the outputs

summary(mod0)

```
Call:
glm(formula = VARAEX ~ 1, family = binomial, data = prevchev)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3744   0.5585   0.5585   0.5585   1.4777

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7791     0.1485   11.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 151.35  on 197  degrees of freedom
AIC: 207.93

Number of Fisher Scoring iterations: 4
```

Model description

Not important

Estimation (logit scale) and test of H_0: estimation = 0

Model fit

Interpretation of the parameter estimation

The estimation obtained is 1.7791

- It is the logit of the estimation of the proportion of WN positive individuals in a village under the hypothesis that this proportion is homogenous

- To obtain the proportion estimation,
one has to apply the inverse logit function

```
exp(1.7791)/(1+exp(1.7791))
```

```
0.8555857
```

or

```
fit<-fitted.values(mod0)
head(fit,5)
```

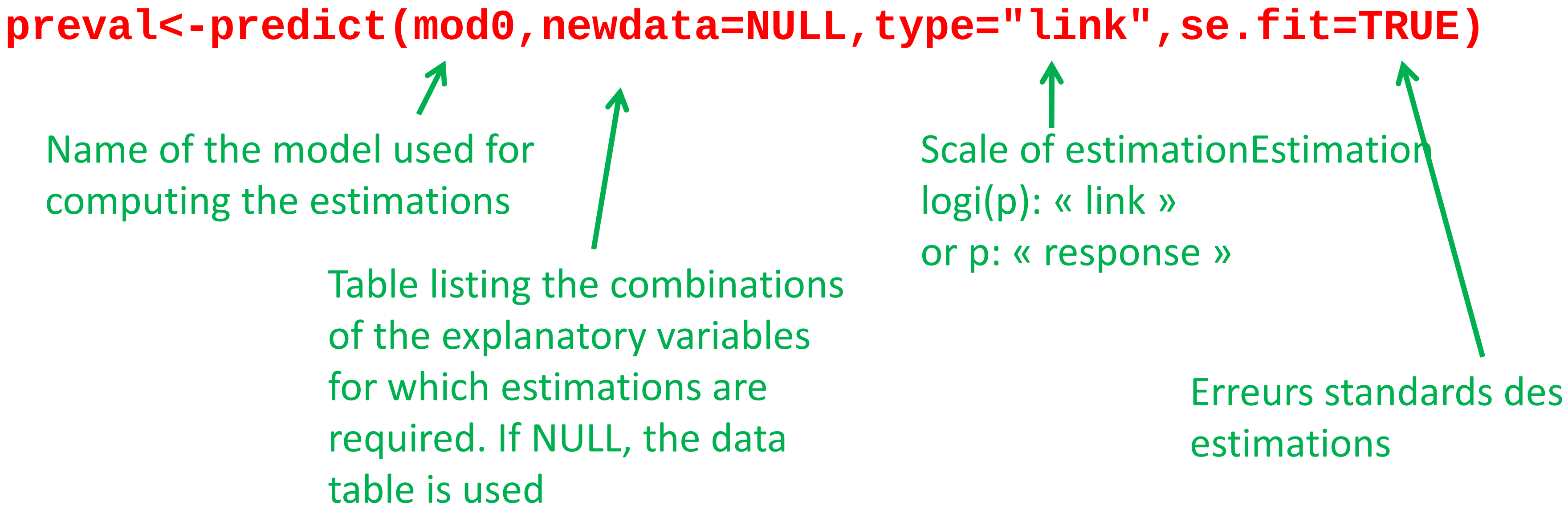
→ computes the value predicted by the model for each statistical unit (line) in the data table

```
1         2         3         4         5
0.8555858 0.8555858 0.8555858 0.8555858 0.8555858
```

Directly in proportion and not any more in Logit

What about the confidence interval !!!!!!!

Estimation of logit(p) with confidence interval



```
preval<-predict(mod0,newdata=NULL,type="link",se.fit=TRUE)
```

Name of the model used for computing the estimations

Table listing the combinations of the explanatory variables for which estimations are required. If NULL, the data table is used

Scale of estimationEstimation logi(p): « link » or p: « response »

Erreurs standards des estimations

Generates a 3 components list: $fit, $se.fit, $residual scale

```
head(preval$fit,5)
1.779101 1.779101 1.779101 1.779101 1.779101
```

```
head(preval$se.fit,5)
0.1485006 0.1485006 0.1485006 0.1485006 0.1485006
```

logit (p) is estimated at 1.7791 with a standard error of 0.1485

Estimation of p with confidence interval

logit (p) is estimated at 1.7791 with a standard error of 0.1485

- **The 95% confidence interval of logit(p) can be build**
  - **Lower limit:** `1.7791-1.96*0.1485` **1.488**
  - **Upper limit:** `1.7791+1.96*0.1485` **2.07**

To compute the confidence interval of the estimation of p

- **The inverse logit function is applied to the limits of the logit(p) IC**
  - **Estimation of p:** `exp(1.7791)/(1+exp(1.7791))` **0.856**
  - **lower limit:** `exp(1.488)/(1+exp(1.488))` **0.816**
  - **upper limit:** `exp(2.07)/(1+exp(2.07))` **0.888**

p is estimated at 0.856 with a 95% IC = [0.816; 0.888]

Interpretation of the outputs

summary(mod0)

```
Call:
glm(formula = VARAEX ~ 1, family = binomial, data = prevchev)
```

Model description

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3744   0.5585   0.5585   0.5585   1.4777
```

Not important

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7791     0.1485   11.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimation (logit scale) and test of H_0: estimation = 0

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 151.35  on 197  degrees of freedom
AIC: 207.93
```

Model fit

```
Number of Fisher Scoring iterations: 4
```

```
Null deviance:     151.35  on 197  degrees of freedom
Residual deviance: 151.35  on 197  degrees of freedom
AIC:               207.93
```

The deviance: Quantity of variation in the data unexplained by the model

- **The larger is the deviance, the larger is the quantity of unexplained variation**

- **Null deviance: the deviance of null model (the model in which the response variable is considered as homogeneous)**

- **Residual deviance: deviance of the current model (note that here null deviance = residual deviance because the current model is the null model)**

- **The degrees of freedom**
  = number of statistical units –number of parameters in the model

Model fit: residual deviance ≈ number of degrees of freedom

- **If residual deviance >> residual ddl**
  - **The model doesn’t contain any important explanatory variable**
  - **La chosen distribution (binomial) is not adapted**

```
Null deviance:     151.35  on 197  degrees of freedom
Residual deviance: 151.35  on 197  degrees of freedom
AIC:               207.93
```

AIC: is a measure of model quality in terms of quantity of explained variation and parameter number

- **For a given deviance, AIC selects the model with the lower number of parameters**

- **For a given number of parameters, AIC selects the model of lowest deviance**

- **The smaller the AIC, the best is the model**

- **A difference of 2 AIC points between 2 models is significant (the model with the lowest AIC is significantly better)**

A model to test the effect of a continuous explanatory variable : age

Now we add the effect of age in the model

```
mod1<-glm(VARAEX~1+AGE, family=binomial, data=prevchev)
```

Or

```
mod1<-update(mod0,~.+AGE)
```

```
Call:
glm(formula = VARAEX ~ AGE, family = binomial, data = prevchev)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8685   0.2642   0.4970   0.6448   1.5660

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.33691    0.39309   0.857 0.391390
AGE          0.18804    0.05155   3.647 0.000265 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 136.18  on 196  degrees of freedom
AIC: 194.75
```

Estimation of the effect of age. The proportion of WN positives increases with age.

The addition of the age affect results in a decrease of the deviance

Test of the effect of age

- Z-test on the coefficient: test of H_0 : coef(AGE)=0

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.33691    0.39309   0.857 0.391390
AGE          0.18804    0.05155   3.647 0.000265 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

P-value <0.05 .
Significant age effect

- Test of the likelihood ratio between mod0 and mod1

The deviance difference between two nested models ~ a χ^2 distribution
with nb of df = difference between the residual degrees of freedom of the two models

```
deviance(mod0)-deviance(mod1)        15.17363
df.residual(mod0)-df.residual(mod1)  1
1-pchisq(15.7,1)                      9.825205e-05
```

P-value <0.05
Significant age effect

- Comparaison of the AIC of mod1 and mod0

```
AIC(mod1)     194.75
AIC(mod0)     207.93
```

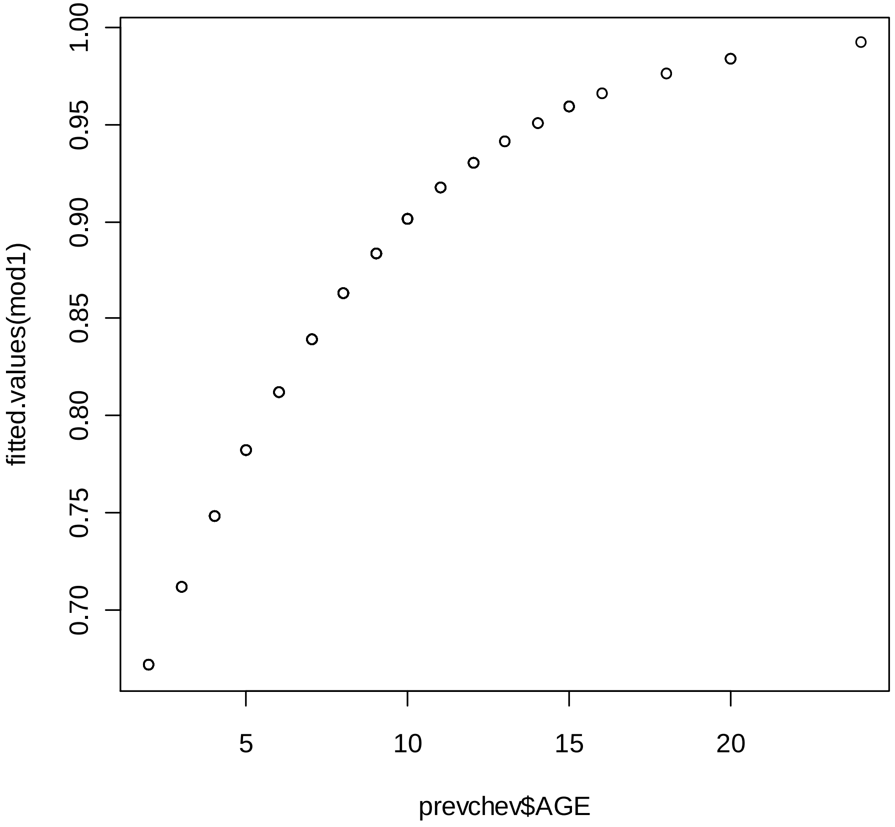
The model with age has a lower AIC. The age effect is significant

Representation of the age effect

```
plot(prevchev$AGE, fitted.values(mod1))
```

Age

Values predicted by the model including the effect of age



- **Non-linearity: the relationship is linear on the logit scale but not on the proportion scale**

- **With the logit link, the predicted values are not above 1**

A model to assess the effect of a categorical variable : region

```
mod2<-glm(VARAEX~1+REGION, family=binomial, data=prevchev)
```

Or

```
mod2<-update(mod0,~.+REGION)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5353     0.2946   5.211 1.88e-07 ***
REGIONNGT     1.1214     0.5953   1.884   0.0596 .
REGIONRIT     0.9634     0.5169   1.864   0.0624 .
REGIONROB     0.4796     0.4780   1.003   0.3157
REGIONSTL    -0.5023     0.3891  -1.291   0.1967
```

The region of DJO is used as a reference

Coefs quantify the difference in logit(p) between the focal region and the reference region. The test is H_0: no difference.

```
mod2<-glm(VARAEX~REGION-1, family=binomial, data=prevchev)
```

Or

```
mod2<-update(mod0,~.-1+REGION)
```

```
           Estimate Std. Error z value Pr(>|z|)
REGIONDJO    1.5353     0.2946   5.211 1.88e-07 ***
REGIONNGT    2.6568     0.5172   5.136 2.80e-07 ***
REGIONRIT    2.4987     0.4247   5.884 4.02e-09 ***
REGIONROB    2.0149     0.3764   5.353 8.64e-08 ***
REGIONSTL    1.0330     0.2541   4.065 4.80e-05 ***
```

Coefs give logit(p) for the focal region.

Test $\underline{H_0 : \text{logit}(p) = 0}$

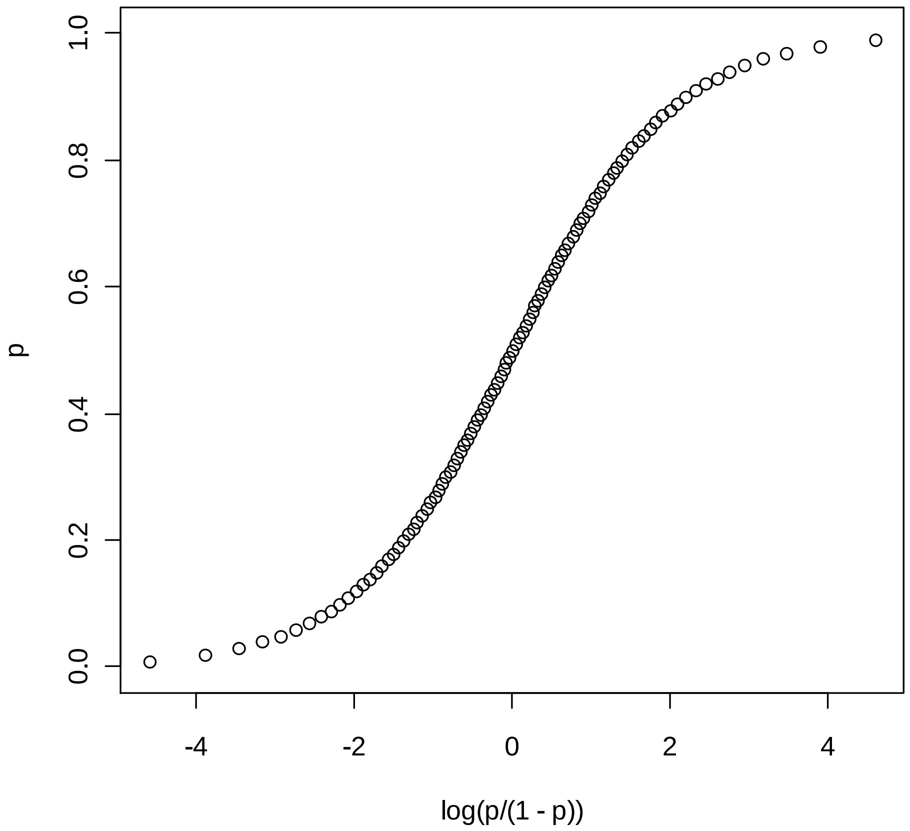
<u>it means p=0.5</u>

Usually not interesting

Which link function ?

- The logit function
  - p is outcome proportion
  - Logit(p) = log(p/(1-p))

```
p=seq(0,1,0.01)
plot(log(p/(1-p)),p)
```



Test of the region effect

- Z-test on the coefficient: not very usefull
  - Either a test of the difference with an arbitrarily determined reference region
  - Or a test of H_0 p=0.5

- Test of the likelihood ratio between mod0 and mod2

The deviance difference between two nested models ~ a χ^2 distribution
with nb of df = difference between the residual degrees of freedom of the two models

```
deviance(mod0)-deviance(mod2)          15.89853
df.residual(mod0)-df.residual(mod2)    4
1-pchisq(15.8985,4)                    0.003
```

P-value <0.05
Significant region effect

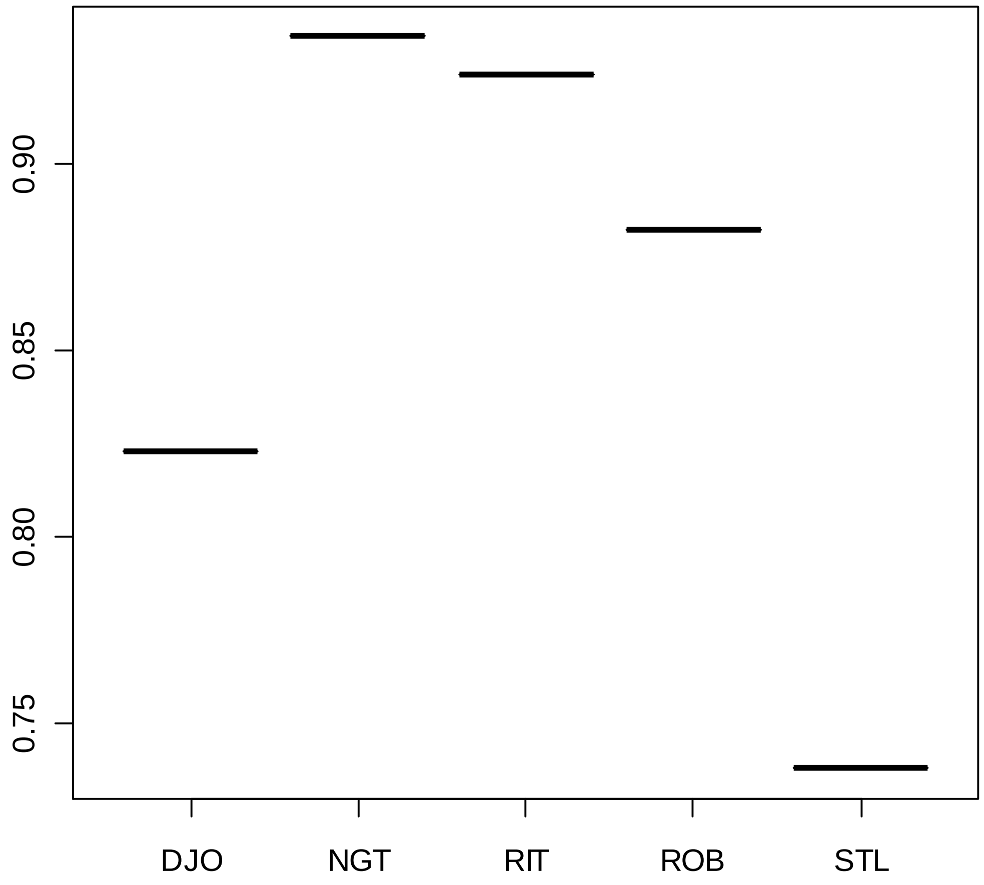
- Comparaison of the AIC of mod2 and mod0

```
AIC(mod1)      200.03
AIC(mod0)      207.93
```

AIC of the model including region is smaller. Region has a significant effect on the proportion of WN positive

Representation of the region effect

```
plot(previnddjo$REGION, fitted.values(mod2))
```



Modelling principles

We want to identify the model built with the available explanatory variables that provides the best possible description of the data

Modelling principles

Model	Interpretation
Saturated model	•Includes one parameter by data point. •Describe perfectly the data but is useless for inferring the mechanisms that generate variation in the response variable.
Maximal model	•Includes the effects of all the potential explanatory variables and all their interactions. •Usually used as a starting point for the model selection process
Minimal adequate model	**•Includes only the effects of the potential variables and of the interactions which removal results in a significant decrease in the fraction of explained variation** **•The description of the response variable retained**
Null model	•Includes only one parameter which represent the estimation of the response variable under the hypothesis that it is homogeneous in the population (no variation). •A kind of baseline model: models that do not explain more variation can be considered as irrelevant.

Modelling principles

We want to identify the model built with the available explanatory variable that provides the best possible description of the data

If we consider only AGE and REGION as potential explanatory variables,
The maximal model contains the effect of

- **AGE**
- **REGION**
- **And the interaction AGE*REGION**

The effect of age differs among regions

Modelling: building the maximal model

Maximal model

```
mod3<-glm(VARAEX~REGION+AGE+AGE*REGION, family=binomial, data=prevchev)
```

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)   
(Intercept)   -0.37072    0.67514  -0.549  0.58294   
AGE            0.29696    0.11079   2.680  0.00735 **
REGIONNGT      0.11548    1.57426   0.073  0.94152   
REGIONRIT      1.91226    1.47469   1.297  0.19473   
REGIONROB      0.74785    1.12760   0.663  0.50719   
REGIONSTL      1.17632    1.12666   1.044  0.29645   
AGE:REGIONNGT  0.10675    0.24481   0.436  0.66280   
AGE:REGIONRIT -0.17889    0.19490  -0.918  0.35871   
AGE:REGIONROB -0.08362    0.16524  -0.506  0.61283   
AGE:REGIONSTL -0.26914    0.15357  -1.753  0.07968 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 116.52  on 188  degrees of freedom
AIC: 191.1

Number of Fisher Scoring iterations: 6
```

Reference coefficient

Age coefficient

Region Coefficients

Interaction Coefficients

Modelling: reaching the minimum adequate model

We start from the maximal model and remove non significant effects

- **One can use Likelihood Ratio Tests to remove the non significant effects**
  - **Start by trying to remove the interactions**
  - **Do not remove a main effect when it is involved in an interaction**

- **One can use an automatic removal procedure based on AIC comparisons**

```
library(MASS)
stepAIC(mod3)
```

```
Start:  AIC=191.1
VARAEX ~ AGE + REGION + AGE * REGION

             Df Deviance    AIC
- AGE:REGION  4   121.17 187.75
<none>            116.53 191.10

Step:  AIC=187.75
VARAEX ~ AGE + REGION

         Df Deviance    AIC
<none>        121.17 187.75
- REGION  4   136.18 194.75
- AGE     1   135.45 200.03
```

The minimum adequate model includes the effects of AGE and REGION but not their interaction

Minimum adequate model = final model

```
modfin<-glm(VARAEX~AGE+REGION, family=binomial, data=prevchev)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.22085    0.44961   0.491 0.623276    
AGE          0.19282    0.05513   3.498 0.000469 ***
REGIONNGT    0.91045    0.60840   1.496 0.134534    
REGIONRIT    0.75867    0.53074   1.429 0.152871    
REGIONROB    0.29442    0.49332   0.597 0.550636    
REGIONSTL   -0.72290    0.41095  -1.759 0.078560 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 151.35  on 197  degrees of freedom
Residual deviance: 121.17  on 192  degrees of freedom
AIC: 187.75
```

Obtaining the estimations from the final model

Create a data frame including the combinations of levels of the explanatory variables for which we want to get an estimations of prevalence

```
newdata<-
expand.grid(AGE=seq(2,24,1),REGION=levels(prevchev$REGION))

newdata
```

```
      AGE REGION
1       2    DJO
2       3    DJO
3       4    DJO
4       5    DJO
5       6    DJO
6       7    DJO
7       8    DJO
8       9    DJO
9      10    DJO
10     11    DJO
11     12    DJO
12  …………..
```

Display the results of the final model

- Use the predict() function to obtain estimations from the final model for the combinations of levels of the explanatory variables

```
preval<-predict(modfin,newdata=newdata,type="link",se.fit=TRUE)
```

The model from which the estimations are required

The table including the combinations of levels of the explanatory variables

The predictions will be given on the logit scale

You want the standard errors of the estimations

```
str(preval)
```

```
List of 3
 $ fit           : Named num [1:115] 0.607 0.799 0.992 1.185 1.378
...
  ..- attr(*, "names")= chr [1:115] "1" "2" "3" "4" ...
 $ se.fit        : Named num [1:115] 0.376 0.346 0.324 0.309 0.304
...
  ..- attr(*, "names")= chr [1:115] "1" "2" "3" "4" ...
 $ residual.scale: num 1
```

Display the results of the final model

- **Add the predictions and confidence intervals in the newdata table**

```
newdata$pred<-exp(preval$fit)/(1+exp(preval$fit))
newdata$low<-exp(preval$fit-1.96*preval$se.fit)/(1+exp(preval$fit-1.96*preval$se.fit))
newdata$hig<-exp(preval$fit+1.96*preval$se.fit)/(1+exp(preval$fit+1.96*preval$se.fit))
```

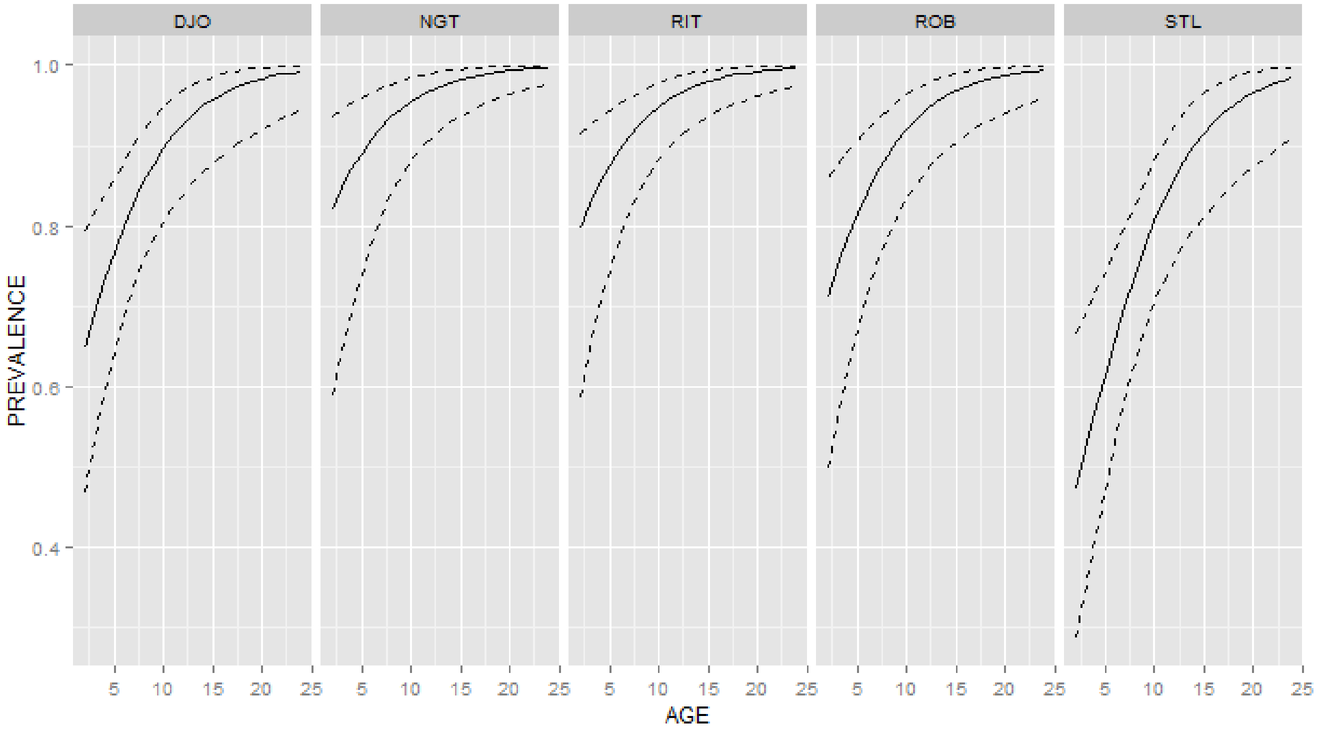
```
head(newdata)
```

```
 AGE REGION      pred       low       hig
1   2    DJO 0.6471421 0.4674587 0.7930405
2   3    DJO 0.6898298 0.5300237 0.8143319
3   4    DJO 0.7295117 0.5884477 0.8357227
4   5    DJO 0.7658401 0.6407583 0.8570843
5   6    DJO 0.7986362 0.6859485 0.8780763
6   7    DJO 0.8278712 0.7240064 0.8981474
```

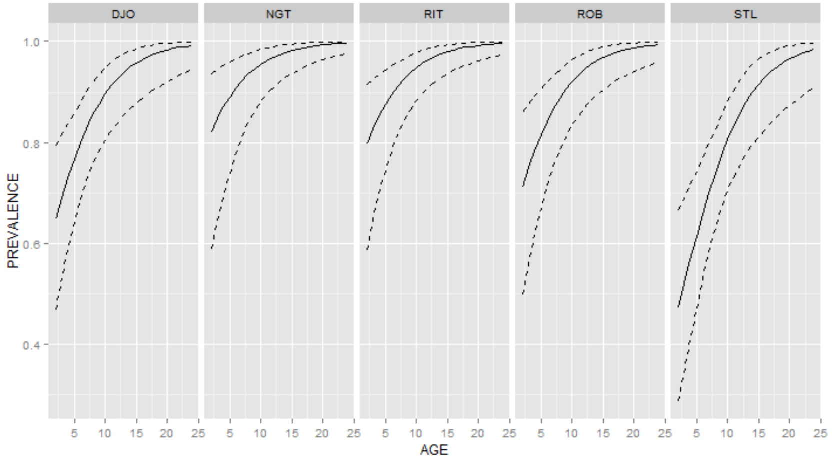
Graphic

(http://www.cookbook-r.com/Graphs/)

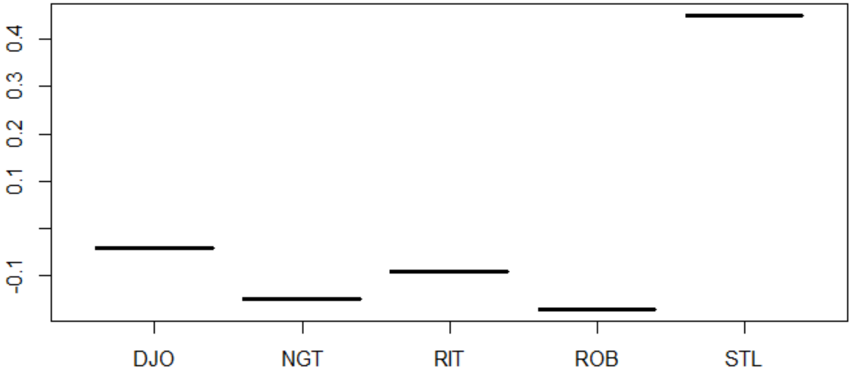
```
ibrary(ggplot2)
ggplot() + geom_line(data=newdata, aes(x=AGE,y=pred)) +
  facet_grid(.~REGION)+
  geom_line(data=newdata, aes(x=AGE,y=low),linetype=2) +
  geom_line(data=newdata, aes(x=AGE,y=hig),linetype=2) +
  xlab("AGE") + ylab("PREVALENCE")
```



Effect of salinity



```
boxplot(prevchev$SALINITE~prevchev$REGION)
```



Salinity provided at the regional scale, not at the village scale.
Lower prevalence in the region with the highest salinity

Displaying the results of the final model

Create a table including the combinations of the levels of the explanatory variables for which we want estimations from the final model

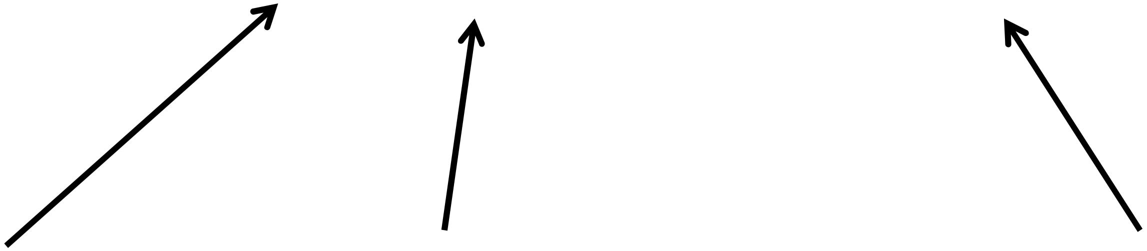
```
newdata<-as.data.frame(matrix(,nrow=5*23,ncol=2))
names(newdata)<-c("AGE","REGION")

newdata$AGE<-rep(seq(2,24,1),5)

newdata$REGION<-
c(rep("DJO",23),rep("NGT",23),rep("RIT",23),rep("ROB",23),rep("STL",23))

newdata
```

```
     AGE REGION
1      2    DJO
2      3    DJO
3      4    DJO
4      5    DJO
5      6    DJO
6      7    DJO
7      8    DJO
8      9    DJO
9     10    DJO
10    11    DJO
11    12    DJO
12  ……………
```

Displaying the results of the final model

- Use the predict function to obtain estimations model modfin for the combinations of the levels of the explanatory variables listed in newdata

```
preval<-predict(modfin,newdata=newdata,type="response")
```



The model used to obtain the predictions

The table containing the combinaitions of explanatory variable levels

The predictions are required on the p, not the logit(p) scale

- The predictions are pasted in the newdata table

```
newdata$pred<-preval
```

Displaying the results of the final model

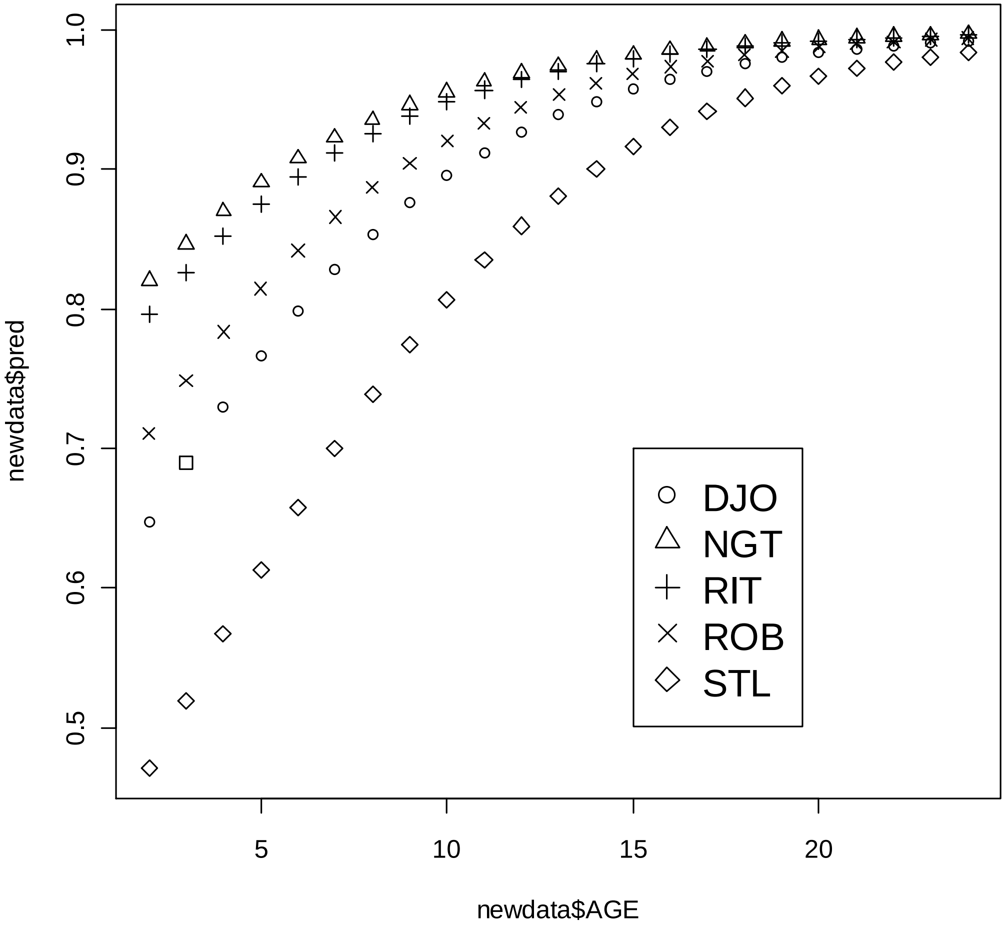
- **Crate an numeric equivalent of REGION (1 distinct digit for each region)**

```
newdata$regnum<-rep(1,115)
newdata$regnum<-replace(newdata$regnum,newdata$REGION=="NGT",2)
newdata$regnum<-replace(newdata$regnum,newdata$REGION=="RIT",3)
newdata$regnum<-replace(newdata$regnum,newdata$REGION=="ROB",4)
newdata$regnum<-replace(newdata$regnum,newdata$REGION=="STL",5)
```

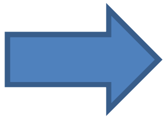
- **The predictions (y-axis) are displayed as a function AGE (x-axis) et of REGION (symbol, coded by regnum)**

```
plot(newdata$AGE,newdata$pred,pch=newdata$regnum)
legend(15,0.7,c("DJO","NGT","RIT","ROB","STL"),cex=1.5,pch=1:5)
```

Displaying the results of the final model



Additive effects of AGE and REGION



Les predicted lines are parallel on the logit(p) scale, but not on the p scale

Other GLMs

Type of data	GLM specifications
A number of events in a population of unknown size (number of cases of a disease) A number of events within a time period of a given length Size of a group	Link Function=log Distribution=Poisson Be careful for the output interpretation the link function is log. Otherwise, the same as for a proportion
A binary variable binaire (two possible outcomes) at the individual scale. Each line is an individual, the dependent variable can only take 2 values (coded 1 or 0)	Link function=logit Distribution=binomial Be careful: the outcome variable has only one component (not two components as when the outcome variable is a proportion) Otherwise, the same as for a proportion

```
glm(nbre~a+b+a*b, family=poisson, data=mydata)
glm(bin~a+b+a*b, family=binomial, data=mydata)
```