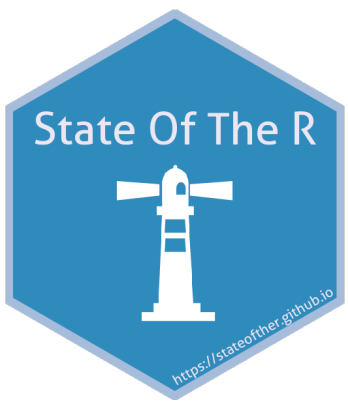
Statistical modelling for biological data with R

Day 4 - Generalized Linear Model (GLM)

Marie-Pierre Etienne

https://marieetienne.github.io

Novembre 2019



Plan

Plan

Les hypothèses du modèle linéaire

- Relation linéaire entre espérance de Y et les variables explicatives –> la prévision par un modèle linéaire peut produire des valeurs en dehors de l'ensemble admissible (cas des proportions)
- Normalité : Robuste à cette hypothèse mais si par exemple les observations sont issues d'une loi discrète, l'hypothèse de normalité n'est plus tenable
- Homoscédasticité, mais il peut arriver que la variance varie en fonction de la moyenne
- Indépendance

Plan

Rappel la régression simple

$$Y_k = \mu + \beta x_k + E_k.$$

Ou encore

$$Y_k \overset{i.i.d}{\sim} \mathcal{N}(\mu_k, \sigma^2)$$
$$E(Y_k) = \mu_k = \mu + \beta x_k.$$

La régression logisitique

Exemple : Evaluation d'une efficacité de captures.

- n individus marqués à la periode i.
- Après un temps x_i, on note Y_i le nombre d'individus marques recapturés

Objectif : estimer l'efficacité de la capture

Le modèle

$$Y_i = \sum_{k=1}^{n_i} Y_{ik} \overset{i.i.d}{\sim} \mathcal{B}(n_i, p_i)$$

$$logit(E(Y_{ik})) = logit(p_i) = \mu + \beta x_i$$

Le modèle linéaire généralisé

2 étapes de modélisation

- choix d'une loi pour Y

$$Y_i \overset{i.i.d}{\sim} \mathcal{L}(\theta_i)$$

- choix de la fonction de lien

$$g(\theta_i) = x_i\theta$$

Le plus souvent la fonction de lien choisie est la fonction de lien naturel (pour garantir de bonnes propriétés mathématiques).

Lois de probabilité et leur fonction de lien naturel

- Loi de Bernoulli et Binomiale, de paramètre p

$$logit(p)$$

- Loi de Poisson de paramètre λ

$$log(\lambda)$$

- Loi Binomiale Négative de paramètre p et r

$$log(p)$$

Des exemples

Remarques

Dans la partie modélisation, tout ce qui était possible dans le modèle linéaire reste possible ici :

Les variables explicatives peuvent être quantitatives ou qualitatives ou les deux.

Plan

Estimation par maximum de vraisemblance

Comme pour le modèle linéaire, mais pas de formule explicite

–> recours à un algorithme d'estimation (Vérifier la convergence)

Plan

Plan

Test sur un paramètre

Pas de loi explicite de l'estimateur : on a recours à ades approximations.

Quand n tend vers l'infini, T l'estimateur de θ:

$$T \sim_{n \to \infty} \mathcal{N}(\theta, Var_\theta)$$

Plan

La déviance

Le **modèle saturé, (Ms)** est le modèle comportant autant de paramètres que d'observations

La déviance d'un modèle M est définie comme

$$D_M = -2(\ell(\hat{\theta}, y, M) - \ell(\hat{\theta}_s, y, Ms))$$

C'est à dire -2 l'écart de vraisemblance entre le modèle saturé et le modèle M.

Remarque : la déviance joue le rôle de RSS dans le modèle linéaire et mesure la variabilité qui n'ets pas capturé par le modèle.

Test de comparaison de modèles emboîtés

Si M_1 est emboîté dans M_2,

Sous $H_0 : \{M_1 \text{ eq } M_2\}$,

$$D_{M_1} - D_{M_2} \sim \chi^2(DDL),$$

DDL étant le nombre de paramètres d'écart entre les deux modèles.

Plan

Plan

Les grenouilles à pattes rouges de Californie

- 237 points d'eau répertoriés
- Présence ou l'absence de grenouilles sauvages
- Longitude et Latitude pour chaque site
- Source de l'information (Museum, Literature, PersCom ou Field Note)

Objectif : Caractériser l'aire de répartition de cette espèce en étudiant comment varie la probabilité de trouver des grenouilles dans un point d'eau en fonction de la latitude et de la longitude.

Présentation

Les données sont disponibles sur grenouille

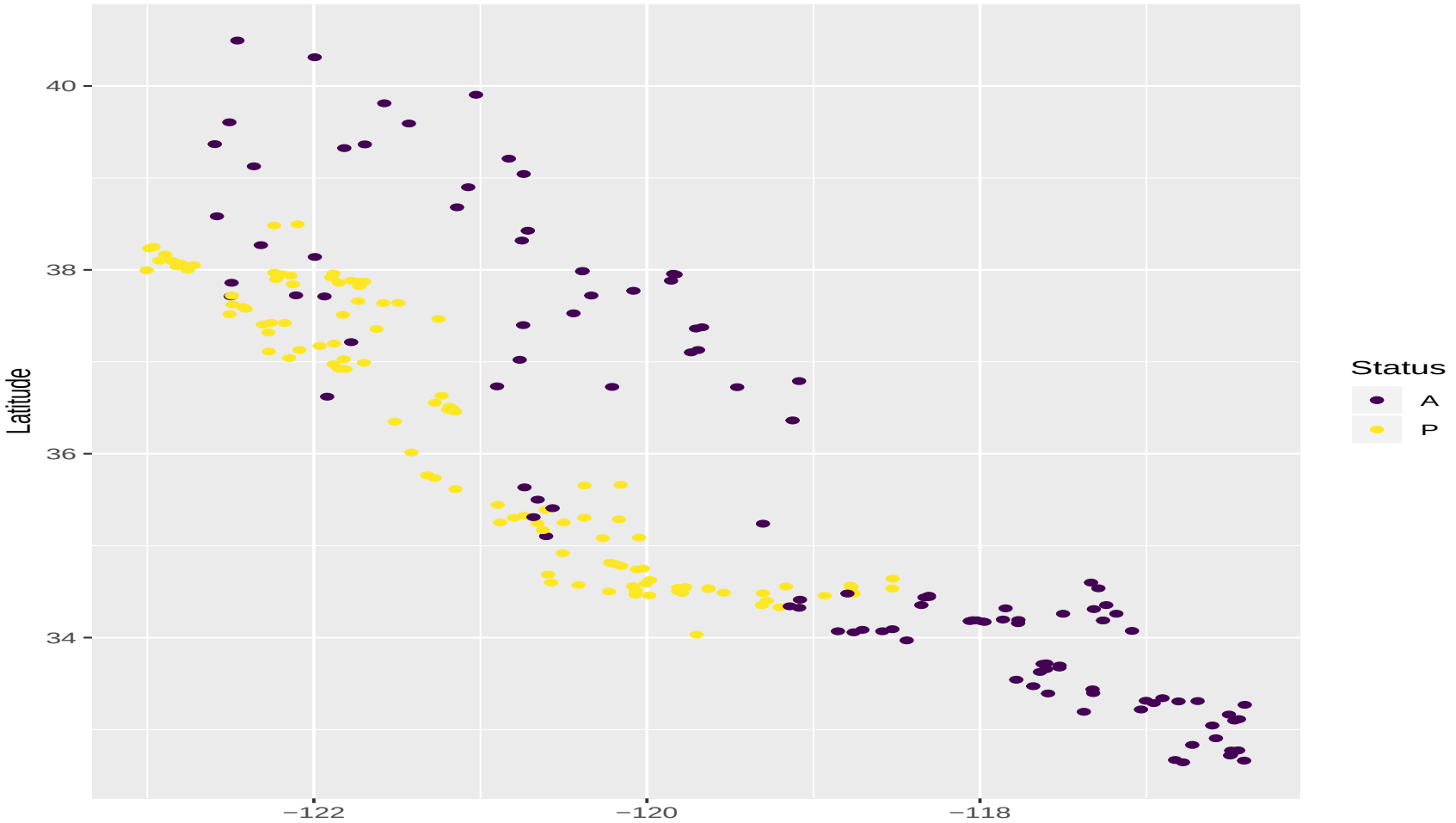
```
Grenouille <- read.table(file = 'https://marieetienne.github.io/dat
                         sep="",
                         header=TRUE)
n <- nrow(Grenouille)

summary(Grenouille)
```

```
##          Source       Source2   Status     Latitude       Longitu
##  FieldNote : 10   MVZ     :59   A:113   Min.   :32.64   Min.   :-
##  Literature:  7   Perscom :31   P:123   1st Qu.:34.39   1st Qu.:-
##  Museum    :188   LACM    :26           Median :35.47   Median :-
##  Perscom   : 31   CAS-SU  :15           Mean   :35.92   Mean   :-
##                   SDNHM   :15           3rd Qu.:37.72   3rd Qu.:-
##                   UMMZ    :12           Max.   :40.49   Max.   :-
##                   (Other):78
```

```
Grenouille %>%
  mutate(pres_bin = ifelse(Status=='A', 0, 1)) -> Grenouille
```

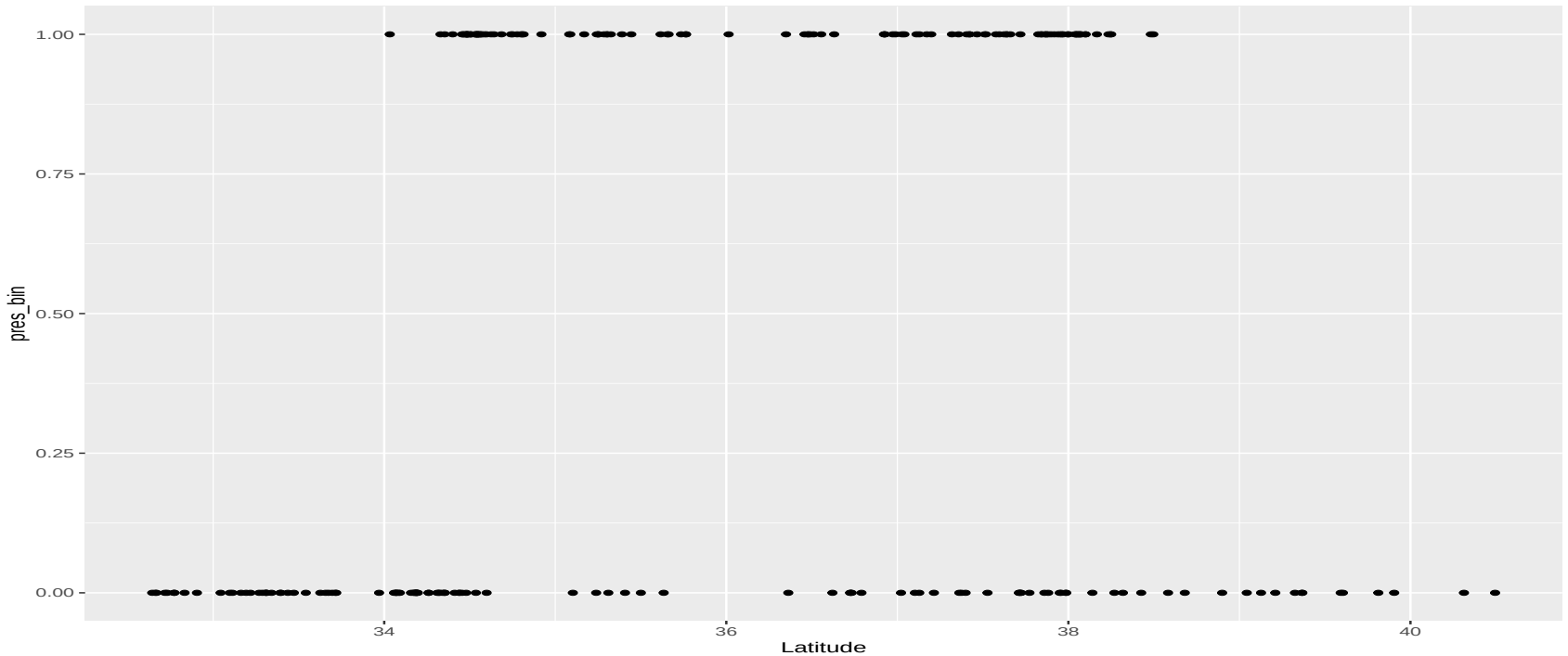
Les points de mesure

```
ggplot(data=Grenouille,
            aes(y = Latitude, x = Longitude, col = Status)) +
  geom_point() + coord_fixed() +
  scale_color_viridis_d()
```



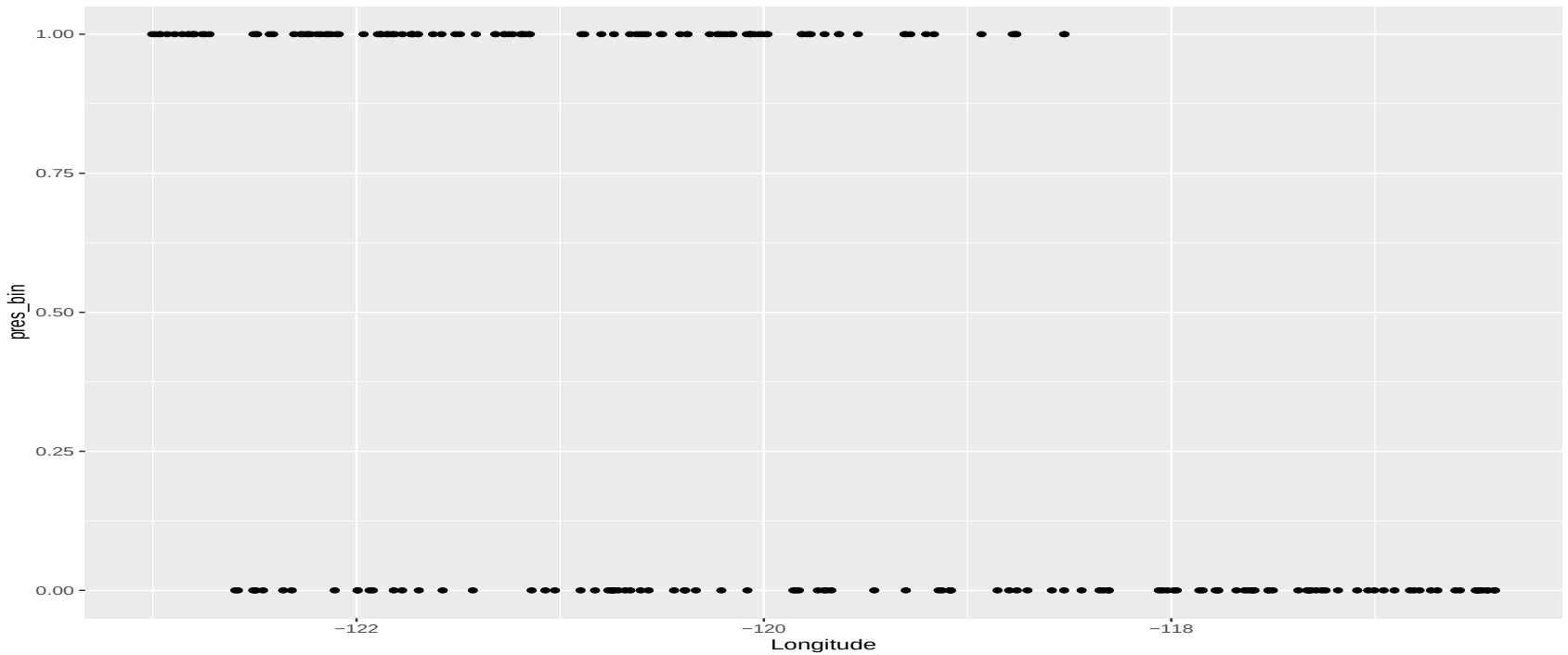
Visualisation de l'effet potentiel de la latitude

```
ggplot(data=Grenouille,
       aes(x=Latitude, y=pres_bin)) +
  geom_point()
```



Visualisation de l'effet potentiel de la longitude

```
ggplot(data=Grenouille,
            aes(x=Longitude, y=pres_bin)) +
  geom_point()
```



Etude de l'effet de la latitude et longitude

$$Y_k \sim \mathcal{B}(p_k), \quad logit(p_k) = \beta_0 + \beta_1 x_k^{(1)} + \beta_2 x_k^{(2)}.$$

```
glm0  <-  glm(Status ~ 1, family=binomial, data = Grenouille)

glm1  <-  glm(Status~Latitude, family=binomial, data = Grenouille)

glm2  <-  glm(Status~Longitude, family=binomial,
              data = Grenouille)

glm12  <-  glm(Status~Latitude+Longitude,
               family=binomial, data = Grenouille)
```

Y a t il un effet de la la latitude ou de la longitude

```
anova(glm0, glm12, test = 'Chisq')

## Analysis of Deviance Table
##
## Model 1: Status ~ 1
## Model 2: Status ~ Latitude + Longitude
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       235     326.74
## 2       233     146.81  2   179.93 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Décomposition des effets type I

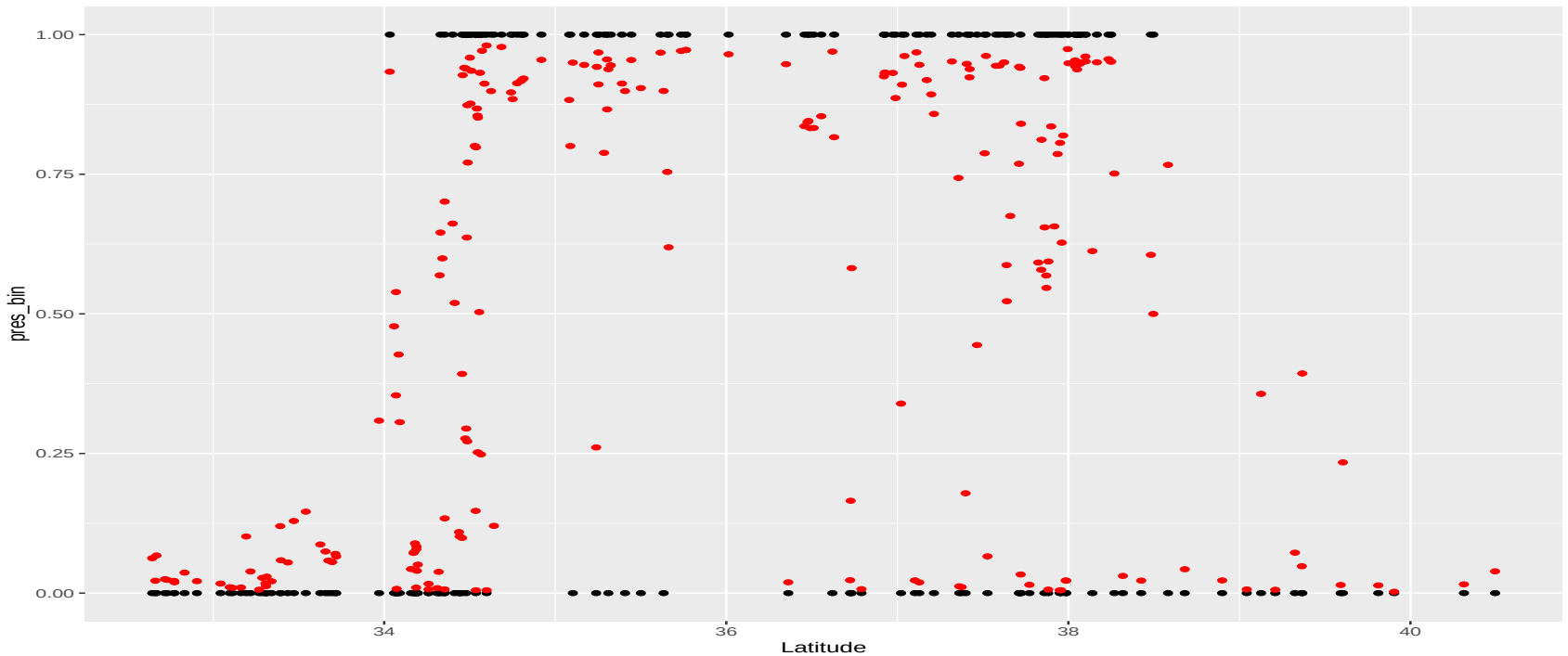
```
anova( glm12, test = 'Chisq')

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Status
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        235     326.74
## Latitude   1    8.929        234     317.81  0.002806 **
## Longitude  1  171.002        233     146.81 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Décomposition des effets type II

```
library(car)
Anova( glm12, test = 'LR')

## Analysis of Deviance Table (Type II tests)
##
## Response: Status
##           LR Chisq Df Pr(>Chisq)
## Latitude    100.04  1  < 2.2e-16 ***
## Longitude   171.00  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prédiction

```
Grenouille %>%
  mutate(pred = predict(glm12, Grenouille, type = 'response' )) %>%
ggplot(data = Grenouille) +
  geom_point(aes(x= Latitude, y = pres_bin) )   +
  geom_point(aes(x=Latitude, y = pred), col= 'red')
```

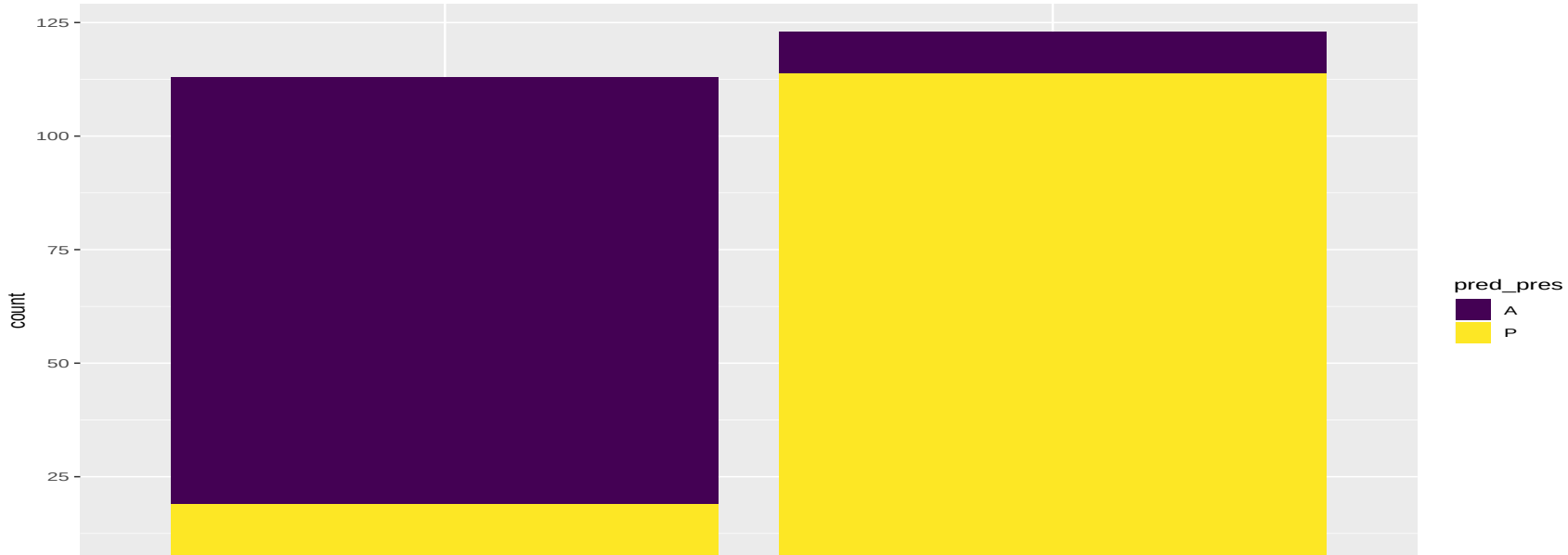


Prédiction - Proportion de bien classés

```
caret::confusionMatrix(Grenouille$pred_pres, Grenouille$Status)$tab
```

```
##            Reference
## Prediction   A   P
##          A  94   9
##          P  19 114
```

```
ggplot(data = Grenouille) +
  geom_bar(aes(x = Status, y = ..count.., fill = pred_pres) ) +
  scale_fill_viridis_d()
```



Références