UER Modélisation

Régression logistique

Thomas Ferté

26/02/2022

Rappels

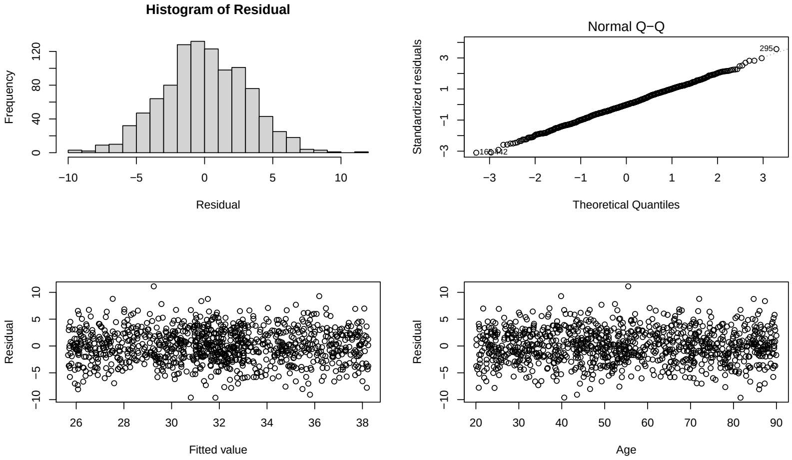
Montre moi tes données . . .

... je te montrerais mon modèle

$$\text{VO2max} = \alpha + \beta_1(\text{sexe}_{\text{homme}}) + \beta_2(\text{age}) + \epsilon \quad (1)$$

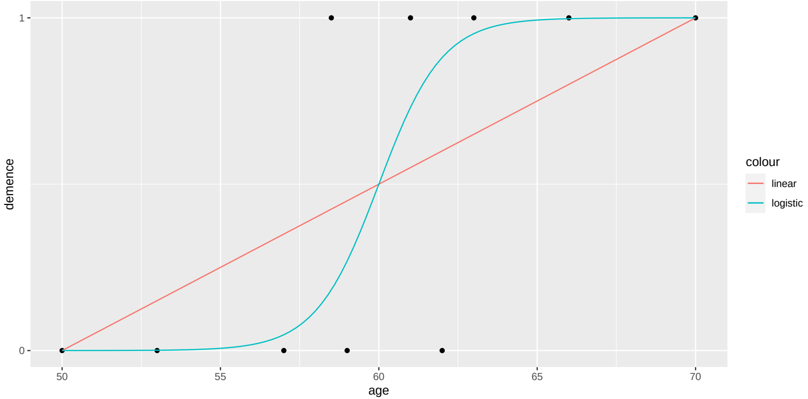
```
##
## Call:
## lm(formula = VO2max ~ sexe + age, data = dfstartsimple)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6417 -1.9696 -0.0389  2.2247 11.1178
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.027817   0.303415  115.44   <2e-16 ***
## sexehomme    5.308110   0.197394   26.89   <2e-16 ***
## age         -0.104089   0.004888  -21.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.118 on 997 degrees of freedom
## Multiple R-squared:  0.5466, Adjusted R-squared:  0.5457
## F-statistic:   601 on 2 and 997 DF,  p-value: < 2.2e-16
```

Et bien sur je vérifierai les hypothèses



Modèle statistique

Intuition



Linéaire : $f_{linear}(x) = \beta_0 + \beta_1 \times x = \eta(x)$

Logistique : $f_{logistic}(x) = P(Dementia = 1|x) = \frac{e^{\eta(x)}}{1+e^{\eta(x)}}$

Logit - exercice

On définit la fonction logit telle que : $Logit(x) = log(\frac{x}{1-x})$

Montrez que $Logit(\frac{e^{\eta}}{1+e^{\eta}}) = \eta$

NB : $log(e^a) = a$

Logit - solution

On définit la fonction logit telle que : $Logit(x) = log(\frac{x}{1-x})$

Montrez que $Logit(\frac{e^{\eta}}{1+e^{\eta}}) = \eta$

$$
\begin{aligned}
Logit(\frac{e^{\eta}}{1+e^{\eta}}) &= log(\frac{\frac{e^{\eta}}{1+e^{\eta}}}{1-\frac{e^{\eta}}{1+e^{\eta}}}) \\
&= log(\frac{\frac{e^{\eta}}{1+e^{\eta}}}{\frac{1+e^{\eta}}{1+e^{\eta}}-\frac{e^{\eta}}{1+e^{\eta}}}) \\
&= log(\frac{\frac{e^{\eta}}{1+e^{\eta}}}{\frac{1}{1+e^{\eta}}}) \\
&= log(e^{\eta}) \\
&= \eta
\end{aligned}
$$

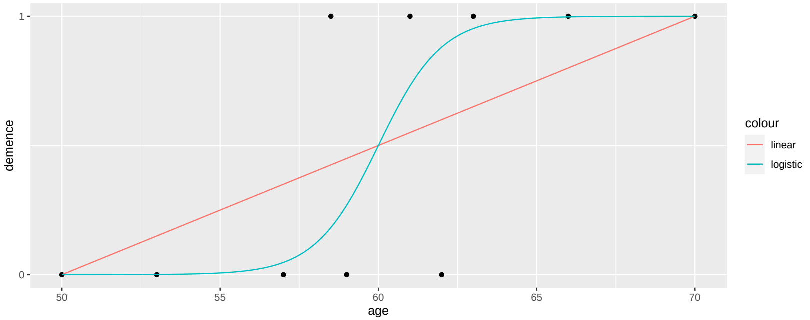
Spécification du modèle

2 spécifications équivalentes :

$$P(Y_i = 1|X_i) = \frac{e^{\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip}}}{1+e^{\beta_0+\beta_1 X_{i1}+...+\beta_p X_{ip}}}$$

$$Logit(P(Y_i = 1|X_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

Spécification du modèle - exemple démence



2 spécifications équivalentes :

$$P(Cognition_i = démence|age_i) = \frac{e^{\beta_0+\beta_1 age_i}}{1+e^{\beta_0+\beta_1 age_i}}$$

$$Logit(P(Cognition_i = démence|age_i)) = \beta_0 + \beta_1 age_i$$

Odd-ratio - exercice

A partir de l'expression suivante : $Logit(P(Y_i = 1|X_i)) = \beta_0 + \beta_1 X_i$

Montrez que : $RC = e^{\beta_1} = \frac{P(Y_i=1|X_i=1)/(1-P(Y_i=1|X_i=1))}{P(Y_i=1|X_i=0)/(1-P(Y_i=1|X_i=0))}$

PS : $log(a) - log(b) = log(a/b)$

Odd-ratio - solution

A partir de $Logit(P(Y_i = 1|X_i)) = \beta_0 + \beta_1 X_i$ on a :

$Logit(P(Y_i = 1|X_i = 0)) = \beta_0 + \beta_1 \times 0 = \beta_0$

$Logit(P(Y_i = 1|X_i = 1)) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$

En faisant une soustraction membre à membre on a :

$$\begin{aligned}
\beta_1 &= Logit(P(Y_i|X_i = 1)) - Logit(P(Y_i|X_i = 0)) \\
&= log(\frac{P(Y_i|X_i = 1)}{1 - P(Y_i|X_i = 1)}) - log(\frac{P(Y_i|X_i = 0)}{1 - P(Y_i|X_i = 0)}) \\
&= log(\frac{P(Y_i = 1|X_i = 1)/(1 - P(Y_i = 1|X_i = 1))}{P(Y_i = 1|X_i = 0)/(1 - P(Y_i = 1|X_i = 0))})
\end{aligned}$$

On a bien : $RC = e^{\beta_1} = \frac{P(Y_i=1|X_i=1)/(1-P(Y_i=1|X_i=1))}{P(Y_i=1|X_i=0)/(1-P(Y_i=1|X_i=0))}$

Odd-ratio - interprétation des coefficients

- β_0 : permet de calculer la probabilité chez les non-exposés égale à $e^{\beta_0}/(1+e^{\beta_0})$
- e^{β_1} : correspond au rapport de côte entre les exposés et les non-exposés (variable binaire) ou bien pour l'augmentation d'une unité d'une variable quantitative.

Spécification du modèle

Cadre général

On retrouve une formulation proche du modèle de régression linéaire :

$$Logit(P(Y_i = 1|X_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

Indicatrices

Les variables catégorielles à plusieurs modalités sont codées sous la forme d'indicatrice.

Exemple :

La survenue de cancer du poumon en fonction du statut tabagique codé *non fumeur*, *tabagisme actif*, *tabagisme passif* sécrira :

$$Logit(P(Cancer_i = 1|X_i)) = \beta_0 + \beta_1 TabagismeActif_i + \beta_2 TabagismePassif_i$$

e^{β_1} s'interprète comme le rapport de cote du cancer du poumon des fumeurs actifs par rapport aux noms fumeurs.

e^{β_2} s'interprète comme le rapport de cote du cancer du poumon des fumeurs passigs par rapport aux noms fumeurs.

Modification d'effet

Les modifications d'effet s'écrivent comme dans un modèle linéaire

Exemple :

La survenue de cancer du poumon dépend de la *consommation* en paquet-année avec un effet différent selon le *sexe* :

$$Logit(P(Cancer_i = 1|X_i)) = \\ \beta_0 + \beta_1 Consommation_i + \beta_2 Homme_i + \beta_3 Consommation_i \times Homme_i$$

e^{β_1} s'interprète comme le rapport de cote de l'augmentation de 1 paquet-année chez les femmes sur le risque de cancer du poumon.

$e^{\beta_1+\beta_3}$ s'interprète comme le rapport de cote de l'augmentation de 1 paquet-année chez les hommes sur le risque de cancer du poumon.

Facteur de confusion et choix des variables

Comme pour le modèle linéaire, les variables explicatives d'un modèle sont :

- L'exposition d'intérêt
- Ses éventuels modificateurs d'effet
- Les éventuels facteurs de confusion de la relation entre l'exposition et la maladie

Estimation du modèle et tests statistiques

Vraisemblance et maximum de vraisemblance - intuition (1)

On veut savoir quelle est la probabilité que le personnage de Sean Bean meurt dans un film. Pour cela on a répertorié tous les films dans lesquels il a joué et on a regardé s'il était ou non décédé.

film_id	death
1	0
2	0
3	0
4	0

Première méthode, on calcul simplement cette probabilité :

```
sum(df_SeanBean$death == 1)/nrow(df_SeanBean)
```

```
## [1] 0.32
```

Deuxième solution : le maximum de vraisemblance !

Vraisemblance et maximum de vraisemblance - intuition (2)

La vraisemblance correspond à la probabilité d'observer une réalisation particulière de l'échantillon pour une valeur des paramètre donnée.

Ici, on considère que les données suivent une loi de Bernouilli (pile ou face) de paramètre π

La vraisemblance pour un individu est π s'il a fait l'évenement et $1-\pi$ s'il n'a pas fait l'événement. On peut donc la noter :

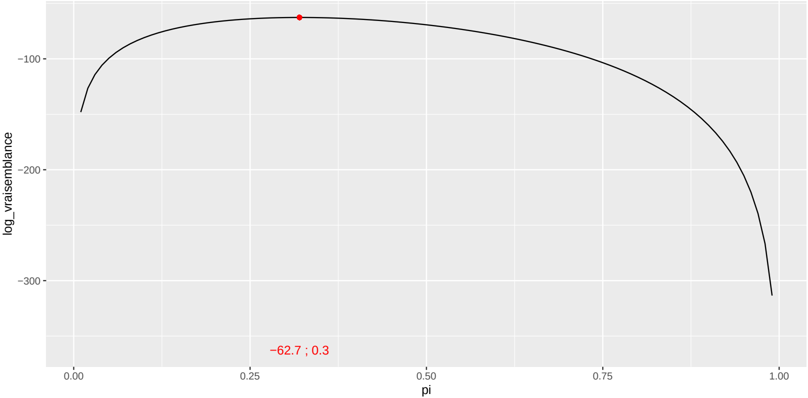
$$\mathcal{L}(\pi; y_i) = \pi^{y_i} \times (1-\pi)^{1-y_i}$$

La vraisemblance pour l'ensemble des individus est donc :

$$\mathcal{L}(\pi; y) = \mathcal{L}(\pi; y_1) \times ... \times \mathcal{L}(\pi; y_n) = \prod_{i=1}^{n} \mathcal{L}(\pi; y_n) = \prod_{i=1}^{n} \pi^{y_i} \times (1-\pi)^{1-y_i}$$

Vraisemblance et maximum de vraisemblance - intuition (3)

A partir de cela on peut faire un graphique montrant la vraisemblance en fonction de la valeur du paramètre π

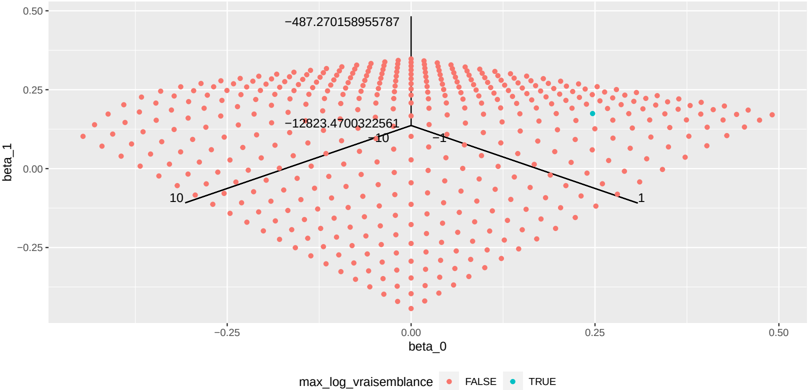


On retrouve la valeur de la première méthode.

Vraisemblance d'un modèle logistique - exemple démence

Même principe pour une régression logistique :

$\mathcal{L}(\pi; y) = \prod_{i=1}^{n} \pi^{y_i} \times (1-\pi)^{1-y_i}$ avec $\pi_i = \frac{e^{\beta_0+\beta_1 age_i}}{1+e^{\beta_0+\beta_1 age_i}}$



Intervalles de confiance

Les paramètres $\widehat{\beta}$ suivent une loi normale de variance $\widehat{var}(\widehat{\beta}_j)$ tel que l'intervalle de confiance au risque α est défini tel que :

$$[\widehat{\beta}_j \pm z_{\alpha/2}\sqrt{\widehat{var}(\widehat{\beta}_j)}]$$

Tests statistiques

- Test global : Rapport de vraisemblance (+++), Wald, Score
- Apport d'une variable : Wald (+++), Rapport de vraisemblance, Score
- Apport d'un ensemble de variables : Rapport de vraisemblance (+++), Wald, Score

Tests statistiques - Wald (un seul paramètre)

Soit $\widehat{\beta_1}$ l'estmateur du coefficient β_1 par le modèle et $SE_{\widehat{\beta_1}}$ sont erreur standard associée alors la statistique de test de wald est définie comme :

$Wald = \frac{\widehat{\beta_1}}{SE_{\widehat{\beta_1}}}$

et suit une loi du Chi-2 à 1 ddl.

Tests statistiques - Log-vraisemblance (comparaison de modèle)

Soit $m2$ le modèle complet et $m1$ le modèle restreint, le rapport de vraisemblance est défini comme :

$$RV = 2 \times (loglik(m2) - loglik(m1))$$

Dans ce cas RV suit une loi du Ch-2 avec un ddl égal à la différence du nombre de paramètres (β) entre les deux modèles.

Tests statistiques - Score (un seul paramètre)

Soit β le paramètre de la régression logisitique à tester. Soit $U(\beta)$ la dérivée première de la vraisemblance du modèle selon ce paramètre et $I(\beta)$ l'opposée de l'espérance de la dérivée seconde de la vraisemblance de ce paramètre. Alors la statistique du score est définie telle que :

$$Score = \frac{U(\beta)^2}{I(\beta)}$$

et suit une loi du Chi-2 à 1 ddl

Hypothèses

Hypothèses du modèle

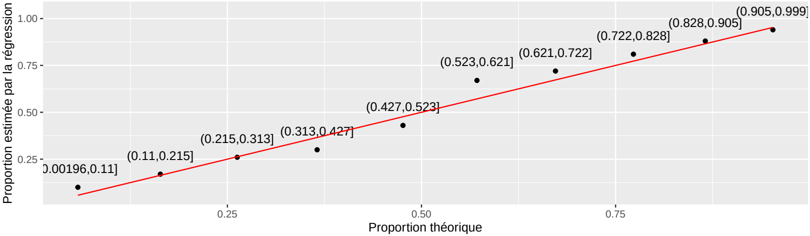
- Log-linéarité : comme pour le modèle de régression linéaire, il faut vérifier la log-linéarité des β pour les variables quantitatives.
- Indépendance des individus

Calibration - test Hosmer et Lemeshow

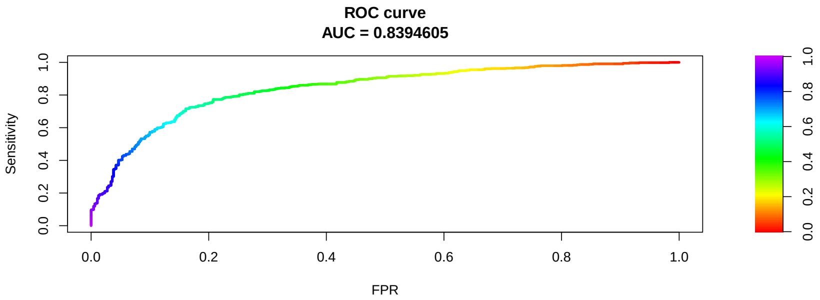
Compare la probabilité prédite et la proportion de réussite de l'outcome.

```
vec_proba <- runif(n = 1000, min = 0, max = 1)
vec_res <- rbinom(n = 1000, size = 1, prob = vec_proba)
hoslem <- generalhoslem::logitgof(vec_res, vec_proba)
hoslem
```

```
##
##  Hosmer and Lemeshow test (binary model)
##
## data:  vec_res, vec_proba
## X-squared = 12.391, df = 8, p-value = 0.1346
```



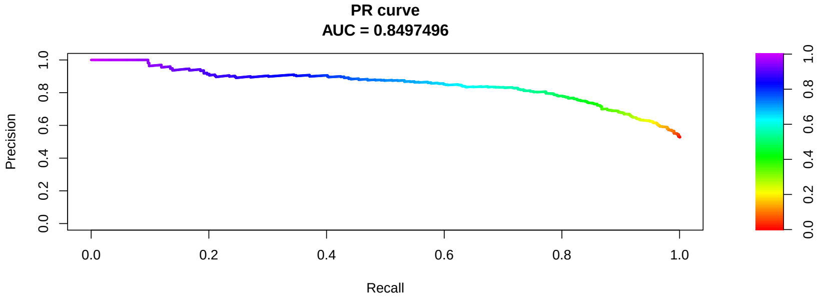
Performance - AUC



$$Se = \frac{VP}{VP+FN}$$

$$FPR = 1 - Sp = 1 - \frac{VN}{VN+FP}$$

Performance - AUPRC



$$Recall = Se = \frac{VP}{VP+FN}$$

$$Precision = VPP = \frac{VP}{VP+FP}$$

Exemple

Données

On s'intéresse au lien entre l'IMC et le décès. Pour cela, on a recueilli des informations sur l'âge des patients (en dizaine d'années) et sur leur IMC et on a recueilli leur statut vital à 6 mois.

age_10	IMC	deces
6.430541	22.16869	0
2.137705	38.04868	1
4.588950	21.64974	0
8.252873	23.21616	0
5.845522	29.52485	0
4.273867	19.22840	0

Spécification du modèle - exercice

On s'intéresse au lien entre l'IMC et le décès. Pour cela, on a recueilli des informations sur l'âge des patients et sur leur IMC et on a recueilli leur statut vital à 6 mois.

Ecrivez le modèle correspondant

Spécification du modèle - solution

$$Logit(P(Décès_i = 1 | IMC_i, Age_i)) = \beta_0 + \beta_1 Age_i + \beta_2 IMC_i$$

Fit du modèle

```r
fit <- glm(deces ~ age_10 + IMC, family = "binomial", data = df_reg_log)
summary(fit)
```

```
##
## Call:
## glm(formula = deces ~ age_10 + IMC, family = "binomial", data = df_reg_log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3144  -0.6114  -0.4096  -0.2567   2.9363
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.324358   0.168015  -37.64   <2e-16 ***
## age_10       0.233652   0.013366   17.48   <2e-16 ***
## IMC          0.109390   0.004278   25.57   <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: 8578.1  on 9999  degrees of freedom
## Residual deviance: 7504.8  on 9997  degrees of freedom
## AIC: 7510.8
##
## Number of Fisher Scoring iterations: 5
```

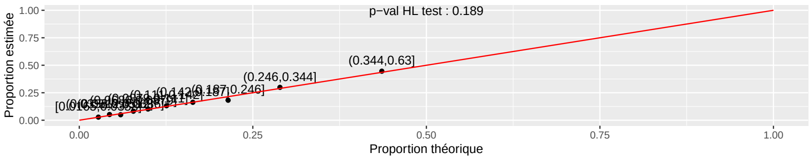
Hypothèses à vérifier - linéarité

```
library(mfp)
fp_reg_log <- mfp::mfp(formula = deces ~ fp(age_10) + fp(IMC), family = "binomial", data = df_reg_log)

glm_reg_log <- glm(fp_reg_log$formula, family = "binomial", data = df_reg_log)

summary(glm_reg_log)
```

```
##
## Call:
## glm(formula = fp_reg_log$formula, family = "binomial", data = df_reg_log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3835  -0.5956  -0.4092  -0.2696   2.8382
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.917746   0.124738  -39.42   <2e-16 ***
## I((IMC/10)^2)       0.195506   0.007338   26.64   <2e-16 ***
## I((age_10/10)^1)    2.359473   0.134294   17.57   <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: 8578.1  on 9999  degrees of freedom
## Residual deviance: 7484.2  on 9997  degrees of freedom
## AIC: 7490.2
##
```

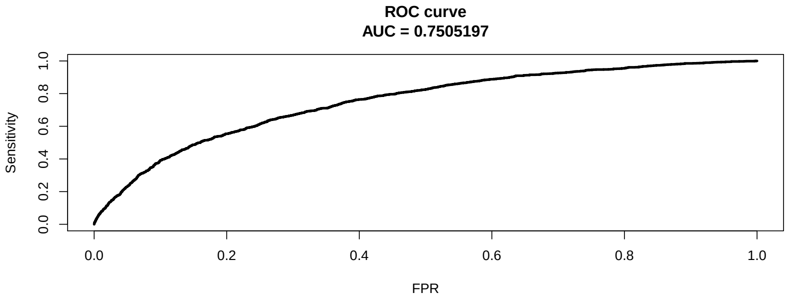
Calibration

```
hoslem_test <- generalhoslem::logitgof(df_reg_log$deces, fitted(fp_reg_log))

cbind(hoslem_test$expected, hoslem_test$observed) %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "group") %>%
  mutate(prop_theo = yhat1/(yhat0 + yhat1),
         prop_obs = y1/(y0 + y1)) %>%
  ggplot(mapping = aes(x = prop_theo, y = prop_obs, label = group)) +
  geom_point() +
  geom_text(nudge_y = +0.1) +
  geom_function(fun = function(x) x, color = "red") +
  annotate(label = paste0("p-val HL test : ", round(hoslem_test$p.value, 3)),
           x = 0.5, y = 1, geom = "text") +
  labs(x = "Proportion théorique", y = "Proportion estimée") +
  lims(x= c(0,1), y=c(0,1))
```



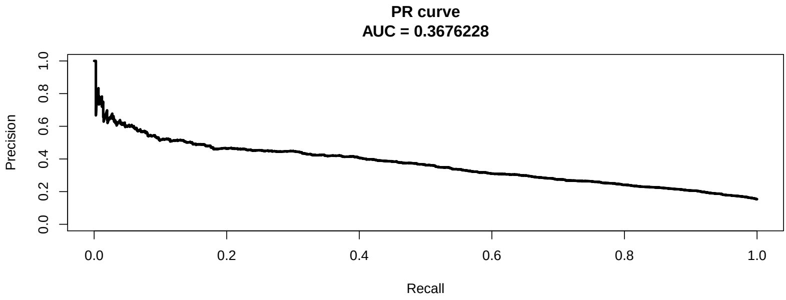
Performances ROC curve

```
roc_curve <- PRROC::roc.curve(scores.class0 = fitted(fp_reg_log),
                              weights.class0 = df_reg_log$deces, curve = TRUE)

plot(roc_curve, color = FALSE)
```



Performances PR curve

```
pr_curve <- PRROC::pr.curve(scores.class0 = fitted(fp_reg_log),
                            weights.class0 = df_reg_log$deces, curve = TRUE)

plot(pr_curve, color = FALSE)
```



Interprétation du modèle

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-4.918	0.125	-39.425	0	-5.165	-4.676
I((IMC/10)^2)	0.196	0.007	26.643	0	0.181	0.210
I((age_10/10)^1)	2.359	0.134	17.569	0	2.098	2.624

Prédiction

On peut se poser la question de la probabilité prédite par le modèle de faire un événement pour un individu de 28 ans avec un IMC à 18

```
df_new <- data.frame(IMC = 18,
                     age_10 = 2.8)
predict(glm_reg_log, df_new, type = "response")
```

```
##          1
## 0.02599142
```

Fin

Questions ?