

Econométrie des variables qualitatives, Roland Garros 2013

Min Zhu, Antoine Setif, Team MASSS

Présentation du sujet

- Que cherchons-nous à analyser ?
 - Les victoires au 1er tour à Roland Garros (RG).
Notre variable expliquée se nomme : **T1RG**.
- Avec quelles variables (explicatives) allons-nous chercher à expliquer ces victoires ?
 - Classement (**Ranking**), Meilleur Classement (**HighRank**)
 - **Age, Taille, Poids**
 - Droitier ou Gauchier (**Main**), Revers une main ou deux mains (**TypRev**)
 - Nombre de participations à RG (**NbrePartRG**)
- Il y a 128 participants (meilleurs joueurs mondiaux, qualifiés, invitations)

Visualisation des donnees

```
tennis[1:18, ]
```

##	T1RG	Ranking	HighRank	Age	Taille	Poids	Main	TypRev	NbrePartRG
## 1	1	1	1	26	188	80	1	2	8
## 2	0	58	42	22	180	67	1	2	1
## 3	0	51	32	28	183	80	1	2	2
## 4	1	83	83	23	182	75	2	2	0
## 5	0	171	158	26	182	76	1	2	0
## 6	1	324	310	19	185	81	1	2	0
## 7	0	74	48	29	182	76	2	2	7
## 8	1	28	26	22	188	77	1	1	2
## 9	0	24	13	25	180	71	1	2	3
## 10	1	60	20	30	185	82	1	2	9
## 11	0	61	27	20	193	77	1	2	3
## 12	1	54	26	31	198	85	1	1	9
## 13	0	88	36	27	183	79	1	1	6
## 14	1	76	33	29	180	74	1	2	5
## 15	0	127	125	19	198	90	2	2	0
## 16	1	19	16	29	178	70	1	1	8
## 17	1	14	2	35	188	88	1	1	11
## 18	0	98	84	23	186	73	1	2	3

Quelques caractéristiques statistiques

```
options(width = 70)
summary(tennis[, 2:9])
```

##	Ranking	HighRank	Age	Taille
##	Min. : 1.0	Min. : 1.0	Min. :18.0	Min. :174
##	1st Qu.: 34.8	1st Qu.: 15.8	1st Qu.:24.0	1st Qu.:182
##	Median : 68.5	Median : 36.5	Median :27.0	Median :185
##	Mean : 87.1	Mean : 48.8	Mean :26.8	Mean :186
##	3rd Qu.:103.2	3rd Qu.: 67.0	3rd Qu.:29.0	3rd Qu.:190
##	Max. :762.0	Max. :310.0	Max. :36.0	Max. :206
##	Poids	Main	TypRev	NbrePartRG
##	Min. : 67.0	Min. :1.00	Min. :1.00	Min. : 0.00
##	1st Qu.: 74.0	1st Qu.:1.00	1st Qu.:2.00	1st Qu.: 1.00
##	Median : 78.5	Median :1.00	Median :2.00	Median : 4.00
##	Mean : 79.1	Mean :1.15	Mean :1.78	Mean : 4.48
##	3rd Qu.: 83.0	3rd Qu.:1.00	3rd Qu.:2.00	3rd Qu.: 8.00
##	Max. :107.0	Max. :2.00	Max. :2.00	Max. :14.00

Premiers pas

- Cherchons à déterminer les variables significativement liées à notre variable. Pour cela, nous avons effectué une régression entre **T1RG** et les variables explicatives, une à une.

```
##          Coeff est. p-value
## Ranking    -0.006124 0.04101
## HighRank   -0.010046 0.02598
## Age         0.002330 0.96150
## Taille      0.034470 0.22430
## Poids       0.029541 0.27542
## Main       -0.123710 0.80374
## TypRev      -0.748129 0.09090
## NbrePartRG  0.116770 0.01936
```

- Ainsi, les variables les plus corrélées sont **Ranking**, **HighRank** et **NbrePartRG**.
- Notons que les signes des coefficients sont cohérents par rapport à ce que nous attendions.

Sélection de variables suivant le critère de l'AIC

```
mod.Nb <- glm(T1RG ~ NbrePartRG, family = binomial(logit), data = tennis)
mod.T <- glm(T1RG ~ ., family = binomial(logit), data = tennis)
step(mod.Nb, direction = "forward", scope = list(upper = formula(mod.T)))
```

AIC.jpg

- Comme l'âge n'est pas une variable significative avec **T1RG**, on décide de travailler avec le modèle suivant :

```
mod.F <- glm(T1RG ~ NbrePartRG + TypRev + HighRank, family = binomial(logit),
  data = tennis)
```

Test de déviance

- Notre modèle est-il correct ?
- Rappel du principe du test : H_1 : Le modèle n'est pas correct

```
1 - pchisq(mod.F$dev, df = 128 - 4)
## [1] 0.004761
```

On accepte H_1 : on peut dire que le modèle n'est pas correct...

- Cependant, constatons que :

```
1 - pchisq(mod.T$dev, df = 128 - 9)
## [1] 0.008812
```

Le modèle complet n'est pas non plus correct...

- Cela est probablement dû à la problématique posée (vainqueur du premier tour de RG).

Test de Hosmer-Lemeshow

- Rappel du principe du test : H_1 : Le modèle n'est pas correct

```
## Loading required package: ResourceSelection
## ResourceSelection 0.2-3 2013-06-18
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  tennis[, 1], mod.F$fitted.values
## X-squared = 8.333, df = 8, p-value = 0.4016
```

- Ici, on n'accepte pas H_1 .
Autrement dit, on ne peut pas dire que le modèle n'est pas correct.

Test de déviance - modèles emboîtés

- Rappel du principe du test : H_1 : mod.T est meilleur que mod.F.
- La statistique de la différence de déviance suit approximativement une loi de $khi^2(p_2 - p_1)$ sous H_0 .

```
1 - pchisq(mod.F$deviance - mod.T$deviance, df = 9 - 4)
## [1] 0.07647
```

- On ne peut pas accepter que mod.T soit meilleur que mod.F.
Autrement dit, le modèle choisi est significativement + informatif que le modèle complet.

Cote - Augmentation d'une unité

- Que se passe-t-il si le régresseur **NbrePartRG** augmente d'une unité ?

```
exp(mod.F$coeff[2])  
## NbrePartRG  
##      1.051
```

Réponse : La cote (rapport entre les probabilités de succès et d'échec) va être multipliée par 1.05.

Autrement dit, les chances de succès augmentent légèrement.

- Que se passe-t-il si le régresseur **HighRank** augmente d'une unité ?

```
exp(mod.F$coeff[4])  
## HighRank  
##      0.9932
```

Réponse : La cote va être multipliée par 0.9932387.

Les chances de succès diminuent légèrement.

Matrice de confusion

- Construisons la matrice de confusion du modèle avec un seuil à 50%.

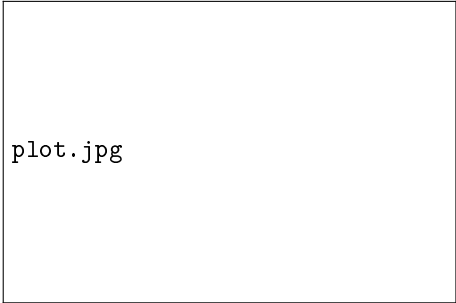
```
## $matconf
##
##      0  1
##  0 40 24
##  1 26 38
##
## $tbc
## [1] 0.6094
##
## $tvp
## [1] 0.625
##
## $tfp
## [1] 0.4062
```

- On constate que le taux de bonne classification est de 61%.
Le taux de vrais positifs est de 62.5% (pas très élevé).
Le taux de faux positifs est de 40.625% (trop élevé).

Courbe ROC

- Construisons la courbe ROC du modèle retenu (en noir) ainsi que la courbe du modèle complet (en rouge).

```
roc(mod.F, seq(0.11, 0.77, 0.001))  
roc(mod.T, seq(0.08, 0.89, 0.001), add = T)
```



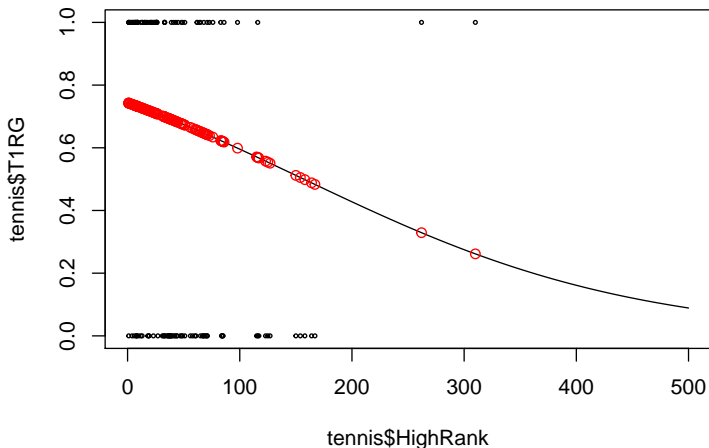
plot.jpg

- Globalement, la qualité de prédiction concernant le modèle complet est meilleure que le modèle retenu.

Graphique des prédictions (1)

- Comment évoluent les chances de succès si la seule variable pouvant varier est **HighRank** ?

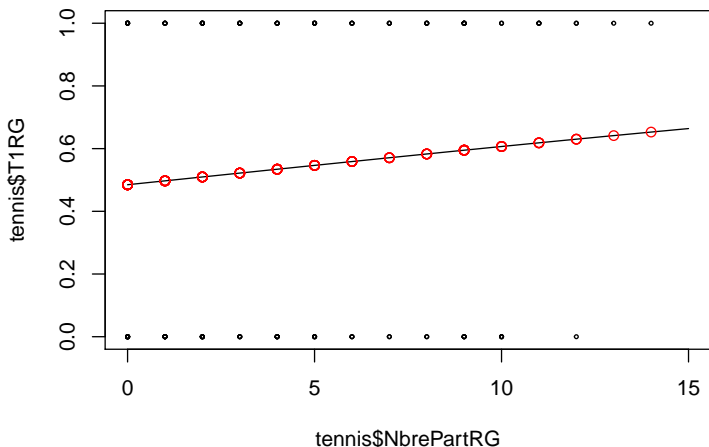
Proba de succès au T1 de RG, NbrePartRG=10, TypRev=1



Graphique des prédictions (2)

- Comment évoluent les chances de succès si la seule variable pouvant varier est **NbrePartRG** ?

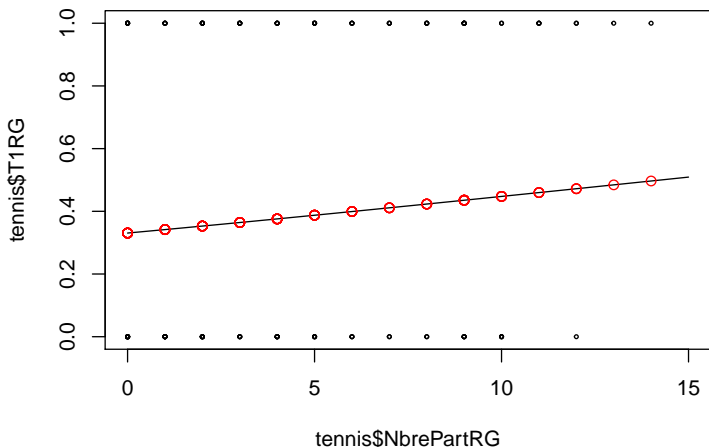
Proba de succès au T1 de RG, HighRank=5, TypRev=2



Graphique des prédictions (3)

- Comment évoluent les chances de succès si la seule variable pouvant varier est **NbrePartRG** ?

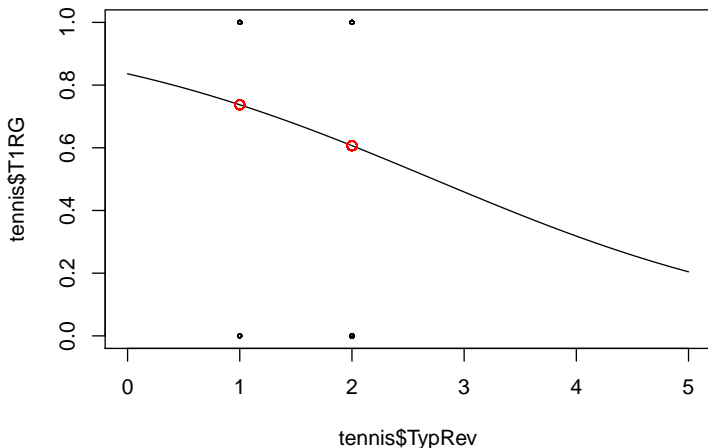
Proba de succès au T1 de RG, HighRank=100, TypRev=2



Graphique des prédictions (4)

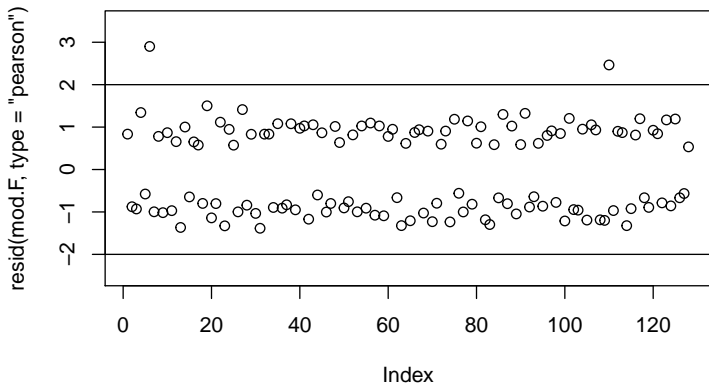
- Comment évoluent les chances de succès si la seule variable pouvant varier est **TypRev** ?

Proba de succès au T1 de RG, NbrePartRG=10, HighRank=5



Résidus de Pearson

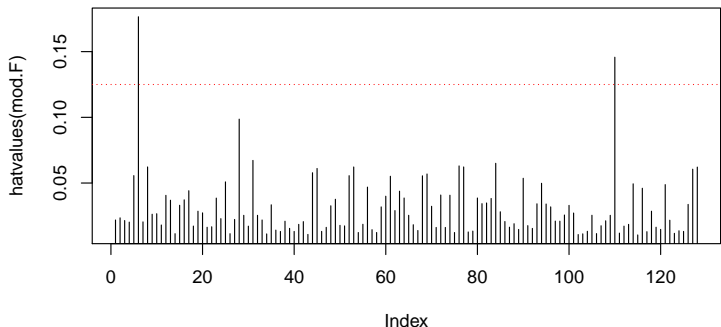
- Les résidus permettent de détecter des valeurs extrêmes et permettent de contrôler le modèle.



- Ici, seulement 2 valeurs sont en dehors des bornes $[-2;2] \Rightarrow$ OK !

Points leviers

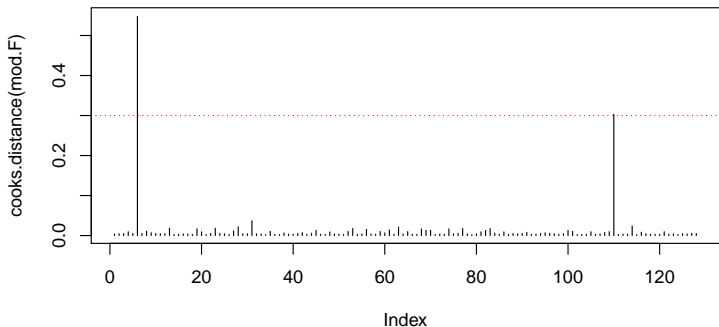
- Un point levier est un point qui participe à une hauteur importante à sa propre prédiction.



- Pour information, il s'agit de Lucas Pouille (19 ans, 324ème) et de Nick Kyrgios (18 ans, 262ème), qui ont tous les 2 gagné leur rencontre.

Points influents

- Un point influent est un point qui, quand il est supprimé, implique une grosse variation dans les estimations des paramètres.



- Merveille des merveilles, ce sont les mêmes individus !