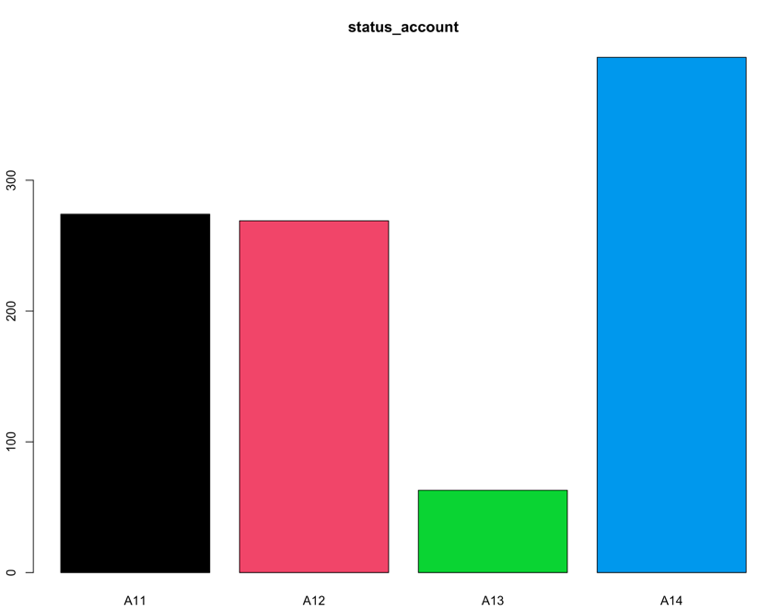
**Question 1**

Class est une variable qualitative binaire. C’est une variable à expliquer qui représente le défaut de crédit soit un bon (Good) risque crédit ou un mauvais risque de crédit (Bad). Et 0 signifie Good et 1 signifie Bad.

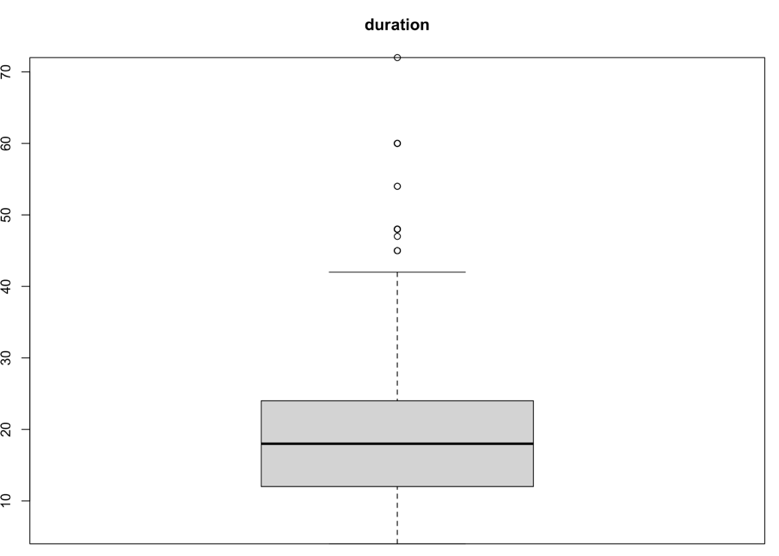


**Question 2**

Ce jeu de données contient 1000 observations et 20 variables comme la suite :

1. « status\_account » est une variable qualitative composée de 4 valeurs uniques :
   1. A11 avec une fréquence de 274
   2. A12 avec une fréquence de 269
   3. A13 avec une fréquence de 63
   4. A14 avec une fréquence de 394

Et on remarque que A14 est le mode



1. « duration » est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 4.0
   2. 1er Quantile : 12.0
   3. Médian : 18.0
   4. Moyenne : 20.9
   5. 3em Quantile. : 24.0
   6. Max. : 72.0
   7. Variance : 145.415
2. « history » est une variable qualitative composée de 5 valeurs uniques :
   1. A30 avec une fréquence de 40
   2. A31 avec une fréquence de 49
   3. A32 avec une fréquence de 530
   4. A33 avec une fréquence de 88
   5. A34 avec une fréquence de 293

Et on remarque que A32 est le mode

1. « purpose» est une variable qualitative composée de 6 valeurs uniques :
   1. A43 avec une fréquence de 280
   2. A40 avec une fréquence de 234
   3. A42 avec une fréquence de 181
   4. A41 avec une fréquence de 103
   5. A49 avec une fréquence de 97
   6. A46 avec une fréquence de 50

Et on remarque que A43 est le mode

1. « amount» est une variable quantitative continue avec les caractéristiques suivantes :
   1. Min. : 250
   2. 1er Quantile : 1366
   3. Médian : 2320
   4. Moyenne : 3271
   5. 3em Quantile. : 3972
   6. Max. : 18424
   7. Variance : 7967843.47
2. « savings» est une variable qualitative composée de 5 valeurs uniques :
   1. A61 avec une fréquence de 603
   2. A62 avec une fréquence de 103
   3. A63 avec une fréquence de 63
   4. A64 avec une fréquence de 48
   5. A65 avec une fréquence de 183

Et on remarque que A61 est le mode

1. « employment» est une variable qualitative composée de 5 valeurs uniques :
   1. A71 avec une fréquence de 62
   2. A72 avec une fréquence de 172
   3. A73 avec une fréquence de 339
   4. A74 avec une fréquence de 174
   5. A75 avec une fréquence de 253

Et on remarque que A73  est le mode

1. « amount» est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 1
   2. 1er Quantile : 2
   3. Médian : 3
   4. Moyenne : 2.973
   5. 3em Quantile. : 4
   6. Max. : 4
   7. Variance : 1.25
2. « status\_personnal» est une variable qualitative composée de 4 valeurs uniques :
   1. A91 avec une fréquence de 50
   2. A92 avec une fréquence de 310
   3. A93 avec une fréquence de 548
   4. A94 avec une fréquence de 92

Et on remarque que A93 est le mode

1. « other\_debtors» est une variable qualitative composée de 3 valeurs uniques :
   1. A101 avec une fréquence de 907
   2. A102 avec une fréquence de 41
   3. A103 avec une fréquence de 52

Et on remarque que A101 est le mode

1. « residence\_since » est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 1
   2. 1er Quantile : 2
   3. Médian : 3
   4. Moyenne : 2.873
   5. 3em Quantile. : 4
   6. Max. : 4
   7. Variance : 1.21
2. « property » est une variable qualitative composée de 4 valeurs uniques :
   1. A121 avec une fréquence de 282
   2. A122 avec une fréquence de 232
   3. A123 avec une fréquence de 332
   4. A124 avec une fréquence de 154

Et on remarque que A123 est le mode

1. « age » est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 19
   2. 1er Quantile : 27
   3. Médian : 33
   4. Moyenne : 35.55
   5. 3em Quantile. : 42
   6. Max. : 75
   7. Variance : 129.40
2. « other\_installment » est une variable qualitative composée de 3 valeurs uniques :
   1. A141 avec une fréquence de 139
   2. A142 avec une fréquence de 47
   3. A143 avec une fréquence de 814

Et on remarque que A143 est le mode

1. « housing » est une variable qualitative composée de 3 valeurs uniques :
   1. A151 avec une fréquence de 179
   2. A152 avec une fréquence de 713
   3. A153 avec une fréquence de 108

Et on remarque que A152 est le mode

1. « nb\_credits » est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 1
   2. 1er Quantile : 1
   3. Médian : 1
   4. Moyenne : 1.4
   5. 3em Quantile. : 2
   6. Max. : 4
   7. Variance : 0.33
2. « job» est une variable qualitative composée de 4 valeurs uniques :
   1. A171 avec une fréquence de 22
   2. A172 avec une fréquence de 200
   3. A173 avec une fréquence de 630
   4. A174 avec une fréquence de 148

Et on remarque que A173 est le mode

1. « nb\_credits » est une variable quantitative discrète avec les caractéristiques suivantes :
   1. Min. : 1
   2. 1er Quantile : 1
   3. Médian : 1
   4. Moyenne : 1.5
   5. 3em Quantile. : 2
   6. Max. : 2
   7. Variance : 0.13
2. « telephone » est une variable qualitative composée de 2 valeurs uniques :
   1. A191 avec une fréquence de 596
   2. A192 avec une fréquence de 404

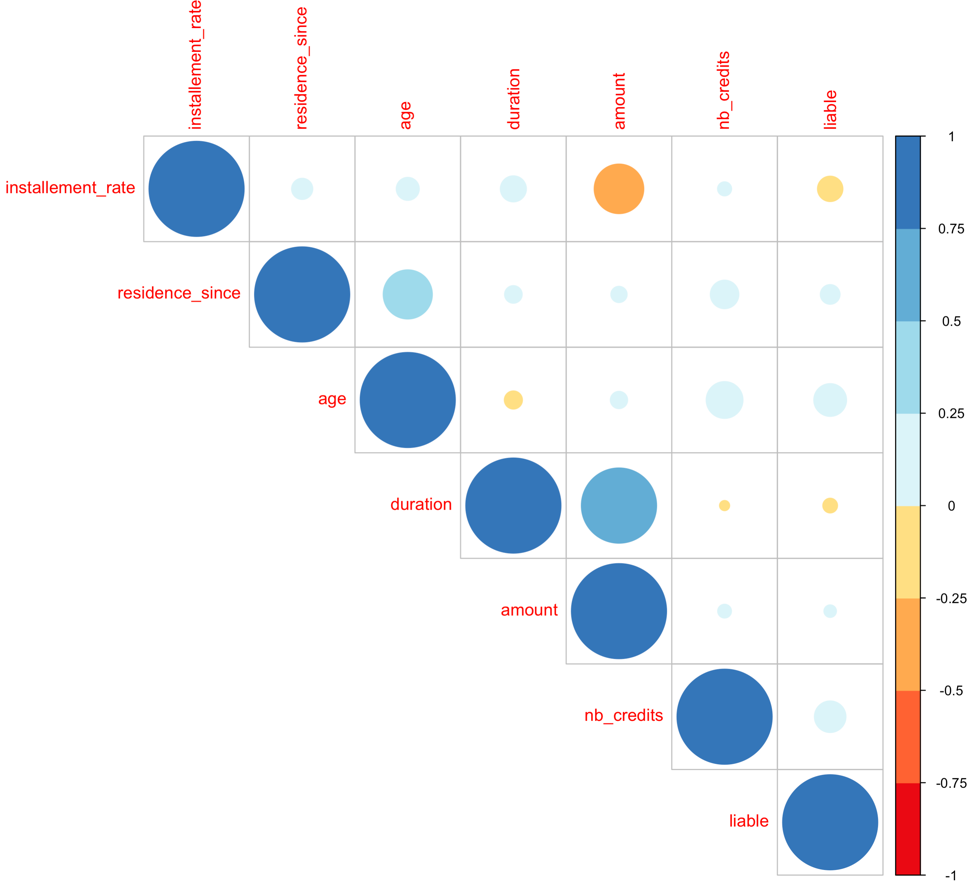
Et on remarque que A173 est le mode

1. « class » est une variable qualitative composée de 2 valeurs uniques :
   1. A171 avec une fréquence de 700
   2. A172 avec une fréquence de 300

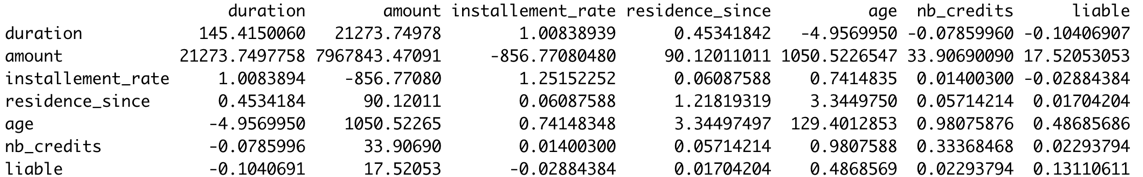
Et on remarque que A173 est le mode

**Note**: Une fonction s’appelle « decr » dans le code R où on peut trouver tous les graphiques de ce jeu de données.

La Corrélation entre les variables quantitatives :



On remarque que la variance est très grande pour duration, amount et age il sera préférable de réduire ces écarts en prenant la somme de la différence entre les valeurs d’une variable et la moyenne de cette variable en devisant par la variance de cette variable (équivalent à l’ACP normé) (Comme dans la méthode de LASSO)

****

**Question 3**

modReg <- glm( class ~ ., data = credit1, family = binomial )

**Question 4**

Si on fait summary(modReg) on trouve la suite :

Une image contenant table

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

On remarque que les variables suivantes particulièrement, ont des P-value très élevées : « residence\_since », « property », « job », « age », « liable », « telephone» elles ne sont pas statistiquement significatives et on peut douter de la fiabilité comme elles ont des P-value plus grandes que 0.05. Et c’est pourquoi le modèle ne semble pas parcimonieux. (Et comme leurs Z-value est entre -2 et 2)

**Note** : d’après de la table :

Une image contenant texte

Description générée automatiquement

On voit que le modèle est bien centré et symétrique.

**Deviance**

La différence entre « deviance nul » et « deviance résiduel » montre comment notre modèle se comporte par rapport au modèle nul (un modèle avec seulement l’intercepte). Plus cet écart est large, mieux c'est. Dans notre cas l’écart est de 319.85 (26.18%)

**McFadden's Pseudo R2**

Pseudo R2 prend une valeur supérieure ou égale à 0 et strictement inferieure à 1, avec une valeur plus proche de zéro indiquant que le modèle n'a pas de pouvoir prédictif et inversement si la valeur est plus proche de 1.

ll.null <- modReg$null.deviance/-2

ll.propse<- modReg$deviance/-2

Pseudo\_R2 <- (ll.null-ll.propse )/ll.null

Pseudo\_R2 # 0.26

Avec le modèle général on trouve que Pseudo R2 est de 0.26 donc un pouvoir prédictif moins que la moyenne.

**P-value pour chi2 de R2**

Chi2\_R2 <- 1- pchisq(2\*(ll.propse-ll.null),df=length(modReg$coefficients)-1)

Chi2\_R2 # 0

Comme Chi2\_R2 est 0 donc la relation entre « class » et les autres variables n'est pas due au hasard, et la valeur R2, 0.26 nous indique la taille de l'effet de cette relation.

**Note** : Il faut noter que certaines valeurs dans les variables : « purpose, history, other\_debtors job et other\_installment » ne sont pas bien représentées par rapport aux autres variables. Par exemple A410, A44 et A48 dans la variable purpose, ne sont pas de tout bien représentées :

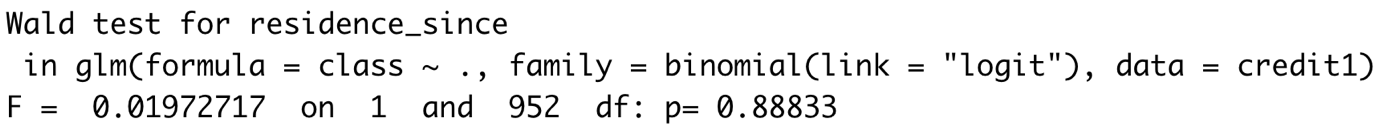
**Une image contenant texte

Description générée automatiquement**

**Wald Test**

Un test de Wald est utilisé pour évaluer la signification statistique de chaque coefficient dans le modèle et est calculé en prenant le rapport du carré du coefficient de régression au carré de l'erreur standard du coefficient. L'idée est de tester l'hypothèse que le coefficient d'une variable indépendante dans le modèle est significativement différent de zéro. Si le test ne rejette pas l'hypothèse nulle, cela suggère que la suppression de la variable du modèle ne nuira pas substantiellement à l'ajustement de ce modèle.

Par example:



Donc residence\_since a une P-value très élevé et son coefficient risque d’être zéro. Et il faut l’éliminer.

**Question 5**

Et pour trouver un modèle parcimonieux on doit sélectionner les variables significatives et qui contribuent le plus au modèle. D’après le Test de Wald, il est prudent d'éliminer certaines variables comme :

employment - age residence\_since - property - housing - nb\_credits- job - liable - telephone

Une image contenant table

Description générée automatiquement**Anova**

Si on exécute la fonction anova(modReg, test="Chisq")

En analysant le tableau, nous pouvons voir la baisse de l'écart lors de l'ajout de chaque variable une à la fois. Et en ajoutant :

* status\_account, duration, history, purpose, savings, employment, installement\_rate, status\_personnal, other\_debtors, other\_installment

Réduit considérablement la déviance résiduelle et les autres variables semblent moins importantes.

Pour trouver un bon modèle prédictif on doit diviser le jeu de donnée sous Train et Test :

* Train 80% de données
* Test 20% de données

Car il ne faut pas avoir un modèle parfait que pour ces données (Overfitting) on a besoin d’être un peu plus sûr que ce modèle va marcher pour des nouvelles données et c’est pourquoi on teste et on peut utiliser l’accuracy (la qualité ou l'état d'être correct) comme un critère d’un bon modèle.

Nous avons créé une fonction qui s’appelle « Trying\_all » et qui prend en entrée les données, la « link function » et le nombre d’itération et qui return data.frame de 16 colonnes et ces colonnes sont regroupées en 4 techniques dans lesquelles on va trouver le bon modèle final. Et chaque technique possède 4 critères ().

De plus il y a 100 lignes, car nous avons initialisé d’une manière aléatoire 100 fois Train data et Test Data pour avoir une vision globale et modéliser sur Test Data et tester toutes les variables. De plus on a implémenté 3 méthodes de « link function : logit, probit , cloglog »

En bref :

Nous avons utilisé trois méthodes de « link function » :

- link function de logit

- link function  de probit

- link function  de cloglog

Et dans chaque méthode il y a quatre techniques pour modéliser et les techniques sont :

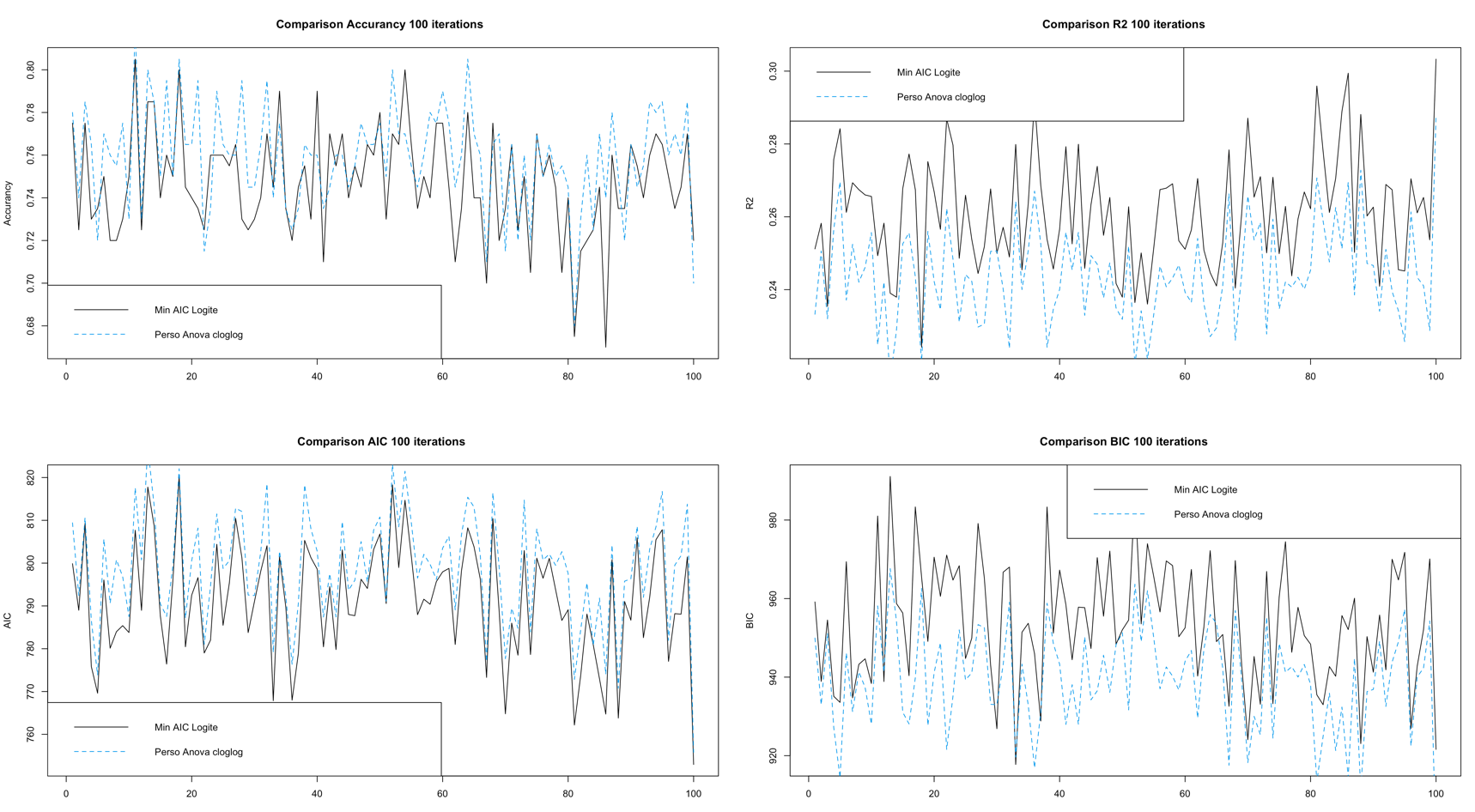
1. Toutes les variables d’une manière générale
2. Stepwise pour minimiser l'AIC avec une direction forward and backward sélection
3. Stepwise pour minimiser la BIC avec une direction forward and backward sélection
4. Méthode personnelle élaborée grâce à Anova pour sélectionner toutes les variables qui ont une Déviance la plus élevée et qui sont significatives d'après Wald test.

Pour comparer on prend en compte 4 critères :

* AIC
* Pseudo R2
* Accuracy
* BIC

Mais avant de comparer nous nous sommes inspirés de la simulation de monte Carlo. Donc nous avons décidé de simuler 100 fois d’une manière aléatoire les 80% de données qui permettent d’établir le modèle et utiliser 20% restantes pour prédire et tester. Pour modéliser 100 fois les 4 techniques (chaque fois avec des données différentes d’une manière aléatoire) la sortie de la fonction est les 4 critères \* 4 techniques = 16 colonnes.

Ensuite on prend la Moyenne, la Quantile "97.5%" et la Quantile "2.5%" de ces critères : d'AIC, de Pseudo R2, d’Accuracy et de BIC pour chaque technique de modélisation et pour les 3 méthodes de link functions de 100 simulations pour comparer et choisir la méthode et la technique idéale pour sélectionner les variables .

Après avoir analysé et observé par rapport à la Moyenne, la Quantile "97.5% et la Quantile "2.5%" on trouve deux techniques intéressantes : celle de stepwise-min-AIC avec link fucntion de type « logit » et celle d’anova-perso avec un link function de type « cloglog »

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step-min-AIC logit | AIC | Pseudo R2 | Accuracy | BIC |
| Moyenne | 791.3831 | 0.2612 | 0.7472 | 953.6112 |
| Quantile "97.5%" | 816.3110 | 0.2941 | 0.7952 | 983.3337 |
| Quantile "2.5%" | 764.2374 | 0.2362 | 0.7024 | 923.5422 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Anova-perso- cloglog | AIC | Pseudo R2 | Accuracy | BIC |
| Moyenne | 798.7325 | 0.2442 | 0.7590 | 939.2708 |
| Quantile "97.5%" | 821.7801 | 0.2702 | 0.8026 | 962.3185 |
| Quantile "2.5%" | 773.3210 | 0.2206 | 0.7124 | 913.8594 |

On remarque l’AIC et Pseudo R2 sont un peu meilleurs pour le modèle Step-min-AIC logit que le modèle Anova-perso- cloglog et inversement pour l’ Accuracy et BIC. Mais comme on ne cherche pas forcement que le modèle soit parfait uniquement pour ces données on va donner l’Accuracy la priorité maximale puis Pseudo R2 puis AIC et BIC. Donc on va choisir le modèle Anova-perso qui a une link fonction de type cloglog et ses variables sont les suivantes :

* status\_account
* duration
* history
* purpose
* savings
* installement\_rate
* tatus\_personnal
* other\_debtors
* other\_installment

Et toutes ces variables sont significatives et contribuent le plus au modèle.

**Question 6**

Y∼Ber(p), 0≤p≤1, if Y={1,with probability p,0,with probability 1−p,

Y∼Ber(p), 0≤p≤1, if Y={1,with probability p,0,with probability 1−p,

Y appertain à Class

On sait que chaque

Donc si

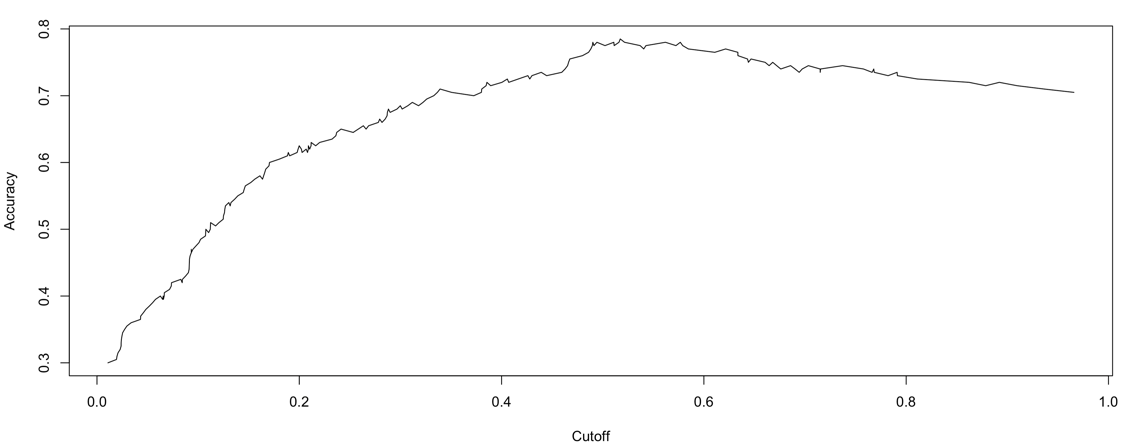
Et

Mais class est de taille donc on a (en ajoutant Independent donc est la même que )

On sait que est complètement déterminé par donc on a :

Et en régression logistique on a :

Trouver la meilleure Cutoff pour améliorer la prédation :

Lors du développement de modèles de prédiction, la mesure la plus critique concerne l'efficacité du modèle dans la prédiction de la variable cible sur les observations hors échantillon. Le processus implique l'utilisation des estimations du modèle pour prédire les valeurs de l'ensemble d'apprentissage normalisé. Une image contenant texte

Description générée automatiquement