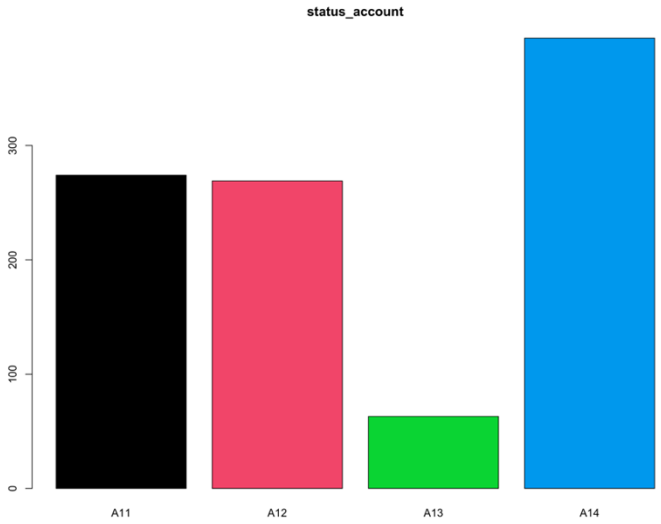
Question 1

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.

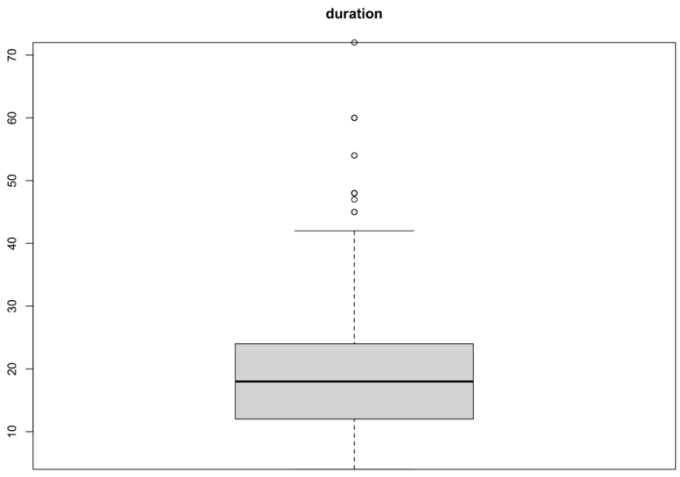
Question 2



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 :
   a. A11 avec une fréquence de 274
   b. A12 avec une fréquence de 269
   c. A13 avec une fréquence de 63
   d. A14 avec une fréquence de 394

   Et on remarque que A14 est le mode



2- « duration » est une variable quantitative discrète avec les caractéristiques suivantes :
   a. Min. : 4.0
   b. 1er Quantile : 12.0
   c. Médian : 18.0
   d. Moyenne : 20.9
   e. 3em Quantile. : 24.0
   f. Max. : 72.0
   g. Variance : 145.415

3- « history » est une variable qualitative composée de 5 valeurs uniques :
   a. A30 avec une fréquence de 40
   b. A31 avec une fréquence de 49
   c. A32 avec une fréquence de 530
   d. A33 avec une fréquence de 88
   e. A34 avec une fréquence de 293

   Et on remarque que A32 est le mode

4- « purpose» est une variable qualitative composée de 6 valeurs uniques :
   a. A43 avec une fréquence de 280
   b. A40 avec une fréquence de 234
   c. A42 avec une fréquence de 181
   d. A41 avec une fréquence de 103
   e. A49 avec une fréquence de 97

f. A46 avec une fréquence de 50

Et on remarque que A43 est le mode

5- « amount» est une variable quantitative continue avec les caractéristiques suivantes :
   a. Min. : 250
   b. 1er Quantile : 1366
   c. Médian : 2320
   d. Moyenne : 3271
   e. 3em Quantile. : 3972
   f. Max. : 18424
   g. Variance : 7967843.47

6- « savings» est une variable qualitative composée de 5 valeurs uniques :
   a. A61 avec une fréquence de 603
   b. A62 avec une fréquence de 103
   c. A63 avec une fréquence de 63
   d. A64 avec une fréquence de 48
   e. A65 avec une fréquence de 183

   Et on remarque que A61 est le mode

7- « employment» est une variable qualitative composée de 5 valeurs uniques :
   a. A71 avec une fréquence de 62
   b. A72 avec une fréquence de 172
   c. A73 avec une fréquence de 339
   d. A74 avec une fréquence de 174
   e. A75 avec une fréquence de 253

   Et on remarque que A73 est le mode

8- « amount» est une variable quantitative discrète avec les caractéristiques suivantes :
   a. Min. : 1
   b. 1er Quantile : 2
   c. Médian : 3
   d. Moyenne : 2.973
   e. 3em Quantile. : 4
   f. Max. : 4
   g. Variance : 1.25

9- « status_personnal» est une variable qualitative composée de 4 valeurs uniques :
   a. A91 avec une fréquence de 50
   b. A92 avec une fréquence de 310
   c. A93 avec une fréquence de 548
   d. A94 avec une fréquence de 92

   Et on remarque que A93 est le mode

10- « other_debtors» est une variable qualitative composée de 3 valeurs uniques :
   a. A101 avec une fréquence de 907
   b. A102 avec une fréquence de 41

c. A103 avec une fréquence de 52

Et on remarque que A101 est le mode

11- « residence_since » est une variable quantitative discrète avec les caractéristiques suivantes :

a. Min. : 1
b. 1er Quantile : 2
c. Médian : 3
d. Moyenne : 2.873
e. 3em Quantile. : 4
f. Max. : 4
g. Variance : 1.21

12- « property » est une variable qualitative composée de 4 valeurs uniques :

a. A121 avec une fréquence de 282
b. A122 avec une fréquence de 232
c. A123 avec une fréquence de 332
d. A124 avec une fréquence de 154

Et on remarque que A123 est le mode

13- « age » est une variable quantitative discrète avec les caractéristiques suivantes :

a. Min. : 19
b. 1er Quantile : 27
c. Médian : 33
d. Moyenne : 35.55
e. 3em Quantile. : 42
f. Max. : 75
g. Variance : 129.40

14- « other_installment » est une variable qualitative composée de 3 valeurs uniques :

a. A141 avec une fréquence de 139
b. A142 avec une fréquence de 47
c. A143 avec une fréquence de 814

Et on remarque que A143 est le mode

15- « housing » est une variable qualitative composée de 3 valeurs uniques :

a. A151 avec une fréquence de 179
b. A152 avec une fréquence de 713
c. A153 avec une fréquence de 108

Et on remarque que A152 est le mode

16- « nb_credits » est une variable quantitative discrète avec les caractéristiques suivantes :

a. Min. : 1
b. 1er Quantile : 1
c. Médian : 1
d. Moyenne : 1.4

e. 3em Quantile. : 2
f. Max. : 4
g. Variance : 0.33

17- « job» est une variable qualitative composée de 4 valeurs uniques :
a. A171 avec une fréquence de 22
b. A172 avec une fréquence de 200
c. A173 avec une fréquence de 630
d. A174 avec une fréquence de 148
Et on remarque que A173 est le mode

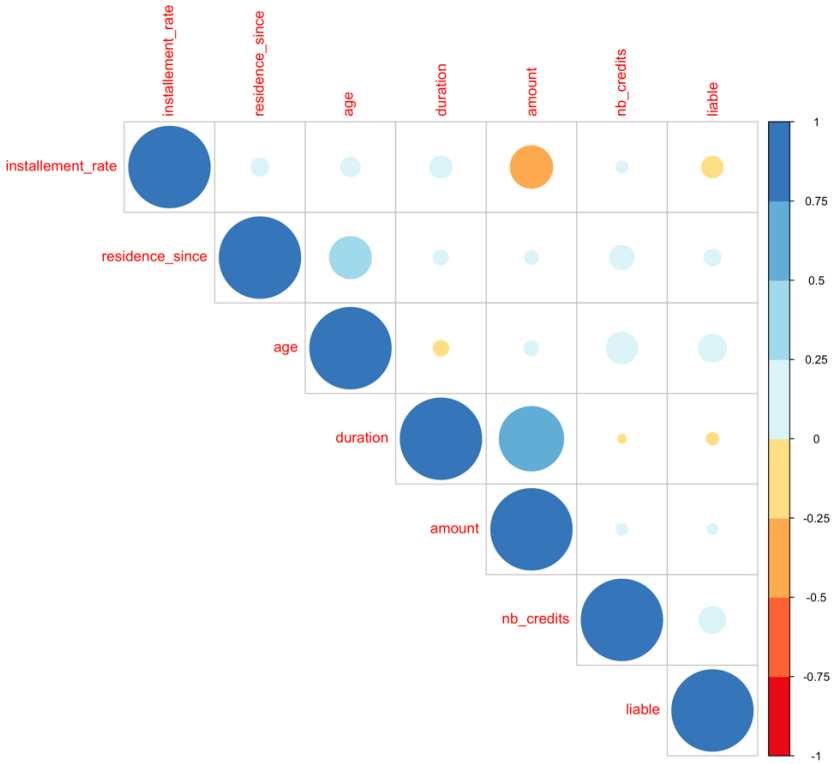
18- « nb_credits » est une variable quantitative discrète avec les caractéristiques suivantes :
a. Min. : 1
b. 1er Quantile : 1
c. Médian : 1
d. Moyenne : 1.5
e. 3em Quantile. : 2
f. Max. : 2
g. Variance : 0.13

19- « telephone » est une variable qualitative composée de 2 valeurs uniques :
a. A191 avec une fréquence de 596
b. A192 avec une fréquence de 404
Et on remarque que A173 est le mode

20- « class » est une variable qualitative composée de 2 valeurs uniques :
a. A171 avec une fréquence de 700
b. 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)

	duration	amount	installement_rate	residence_since	age	nb_credits	liable
duration	145.4150060	21273.74978	1.00838939	0.45341842	-4.9569950	-0.07859960	-0.10406907
amount	21273.7497758	7967843.47091	-856.77080480	90.12011011	1050.5226547	33.90690090	17.52053053
installement_rate	1.0083894	-856.77080	1.25152252	0.06087588	0.7414835	0.01400300	-0.02884384
residence_since	0.4534184	90.12011	0.06087588	1.21819319	3.3449750	0.05714214	0.01704204
age	-4.9569950	1050.52265	0.74148348	3.34497497	129.4012853	0.98075876	0.48685686
nb_credits	-0.0785996	33.90690	0.01400300	0.05714214	0.9807588	0.33368468	0.02293794
liable	-0.1040691	17.52053	-0.02884384	0.01704204	0.4868569	0.02293794	0.13110611

Question 3

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

Question 4

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

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.688e-01  1.076e+00   0.157 0.875351
status_accountA12    -3.238e-01  2.157e-01  -1.501 0.133320
status_accountA13    -9.252e-01  3.677e-01  -2.516 0.011854 *
status_accountA14    -1.685e+00  2.311e-01  -7.291 3.07e-13 ***
duration              2.965e-02  9.272e-03   3.198 0.001384 **
historyA31            2.212e-01  5.443e-01   0.406 0.684375
historyA32           -5.121e-01  4.276e-01  -1.198 0.231008
historyA33           -7.838e-01  4.696e-01  -1.669 0.095050 .
historyA34           -1.410e+00  4.378e-01  -3.220 0.001281 **
purposeA41           -1.635e+00  3.749e-01  -4.361 1.30e-05 ***
purposeA410          -1.534e+00  8.005e-01  -1.917 0.055280 .
purposeA42           -7.355e-01  2.587e-01  -2.843 0.004467 **
purposeA43           -8.410e-01  2.457e-01  -3.424 0.000618 ***
purposeA44           -4.477e-01  7.618e-01  -0.588 0.556760
purposeA45           -1.692e-01  5.453e-01  -0.310 0.756392
purposeA46            9.472e-02  3.957e-01   0.239 0.810804
purposeA48           -1.951e+00  1.207e+00  -1.616 0.106089
purposeA49           -6.829e-01  3.321e-01  -2.056 0.039764 *
amount                1.232e-04  4.446e-05   2.771 0.005581 **
savingsA62           -3.684e-01  2.847e-01  -1.294 0.195679
savingsA63           -3.851e-01  4.008e-01  -0.961 0.336560
savingsA64           -1.330e+00  5.237e-01  -2.540 0.011100 *
savingsA65           -9.662e-01  2.601e-01  -3.715 0.000203 ***
employmentA72        -1.246e-01  4.253e-01  -0.293 0.769543
employmentA73        -2.290e-01  4.096e-01  -0.559 0.576113
employmentA74        -8.664e-01  4.444e-01  -1.950 0.051226 .
employmentA75        -2.975e-01  4.130e-01  -0.720 0.471307
installement_rate     3.347e-01  8.784e-02   3.810 0.000139 ***
status_personnalA92  -2.553e-01  3.856e-01  -0.662 0.507904
status_personnalA93  -8.150e-01  3.791e-01  -2.150 0.031563 *
status_personnalA94  -3.857e-01  4.530e-01  -0.852 0.394473
other_debtorsA102     3.766e-01  4.032e-01   0.934 0.350283
other_debtorsA103    -1.057e+00  4.251e-01  -2.487 0.012874 *
residence_since       1.207e-02  8.592e-02   0.140 0.888302
propertyA122          2.713e-01  2.523e-01   1.075 0.282192
propertyA123          2.195e-01  2.350e-01   0.934 0.350367
propertyA124          7.121e-01  4.205e-01   1.694 0.090348 .
age                  -1.488e-02  9.216e-03  -1.614 0.106506
other_installmentA142 -9.305e-02  4.120e-01  -0.226 0.821318
other_installmentA143 -6.242e-01  2.389e-01  -2.613 0.008966 **
housingA152          -4.143e-01  2.319e-01  -1.786 0.074051 .
housingA153          -6.097e-01  4.714e-01  -1.294 0.195819
nb_credits            2.925e-01  1.894e-01   1.544 0.122558
jobA172               5.076e-01  6.721e-01   0.755 0.450119
jobA173               5.550e-01  6.482e-01   0.856 0.391896
jobA174               4.634e-01  6.570e-01   0.705 0.480549
liable                2.550e-01  2.481e-01   1.028 0.304128
telephoneA192        -2.730e-01  2.007e-01  -1.360 0.173840

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  901.88  on 952  degrees of freedom
AIC: 997.88

Number of Fisher Scoring iterations: 5
```

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 :

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3324  -0.7020  -0.3783   0.7198   2.6073
```

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 :

```
 [1] "purpose"
      credit1[, g]
 class A40 A41 A410 A42 A43 A44 A45 A46 A48 A49
     0 145  86    7 123 218   8  14  28   8  63
     1  89  17    5  58  62   4   8  22   1  34
```

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:

```
Wald test for residence_since
 in glm(formula = class ~ ., family = binomial(link = "logit"), data = credit1)
F =  0.01972717  on  1  and  952  df: p= 0.88833
```

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

Anova

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

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			999	1221.73		
status_account	3	131.336	996	1090.39	< 2.2e-16	***
duration	1	38.497	995	1051.90	5.485e-10	***
history	4	29.311	991	1022.58	6.759e-06	***
purpose	9	33.509	982	989.08	0.0001089	***
amount	1	1.504	981	987.57	0.2200237	
savings	4	19.068	977	968.50	0.0007623	***
employment	4	12.496	973	956.01	0.0140190	*
installement_rate	1	11.907	972	944.10	0.0005594	***
status_personnal	3	9.459	969	934.64	0.0237759	*
other_debtors	2	8.137	967	926.51	0.0171062	*
residence_since	1	0.155	966	926.35	0.6934956	
property	3	2.520	963	923.83	0.4717369	
age	1	3.725	962	920.11	0.0536071	.
other_installment	2	8.357	960	911.75	0.0153184	*
housing	2	3.517	958	908.23	0.1723184	
nb_credits	1	2.328	957	905.90	0.1270420	
job	3	1.110	954	904.79	0.7745757	
liable	1	1.049	953	903.74	0.3057414	
telephone	1	1.863	952	901.88	0.1722319	

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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 ($\frac{16\ colonnes}{4\ critères} = 4\ techniques$).
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 :

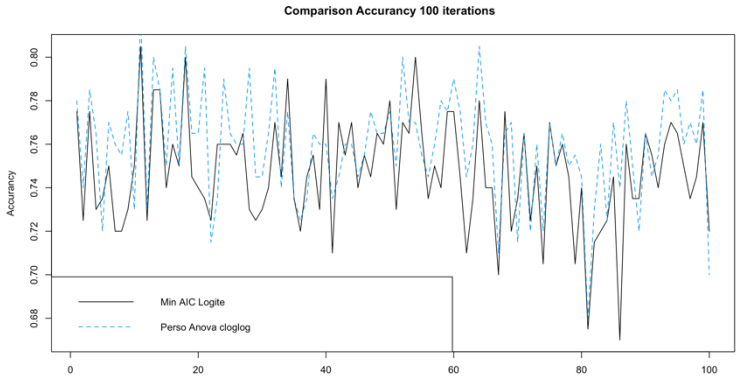
1- Toutes les variables d'une manière générale
2- Stepwise pour minimiser l'AIC avec une direction forward and backward sélection
3- Stepwise pour minimiser la BIC avec une direction forward and backward sélection
4- Méthode personnelle élaborée grâce à Anova pour sélectionner toutes les variables qui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

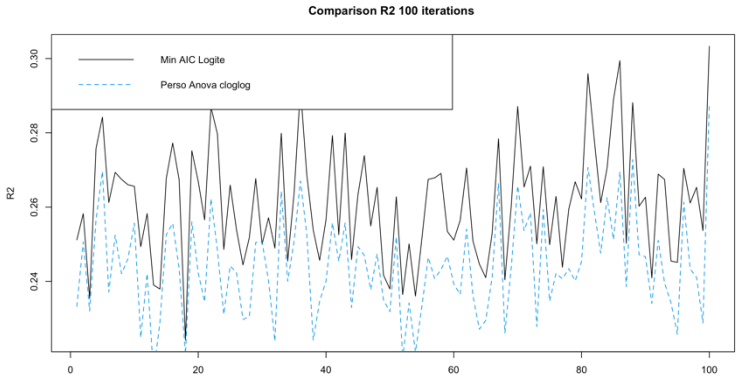
Pour comparer on prend en compte 4 critères :

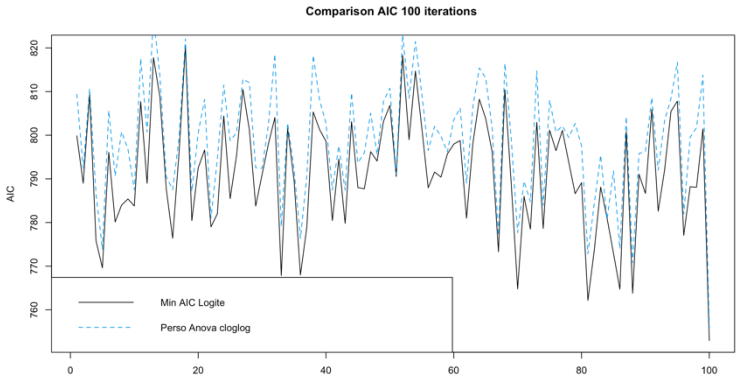
- AIC
- Pseudo R2
- Accuracy
- BIC

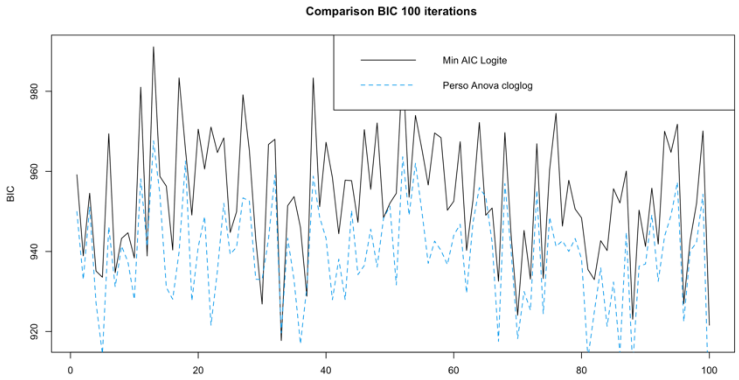
Mais avant de comparer nous nous sommes inspirés de la simulation de monte Carlo. Donc nous avons décidé de simuler 100 fois d'une manière aléatoire les 80% de données qui permettent d'établir le modèle et utiliser 20% restantes pour prédire et tester. Pour modéliser 100 fois les 4 techniques (chaque 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 appertain à Class
On sait que chaque $Y \sim Ber, 0 \leq p \leq 1\ si$:

$$Y = \begin{cases} 1, & avec\ probability\ p, \\ 0, avec\ probability\ 1-p, \end{cases}$$

Donc si $P[Y=1] = 1\ et\ P[Y=0] = 1-p,$
Et $P[Y=y] = p^y(1-p)^{1-y}, y = 0{,}1$

Mais class est de taille $n = 1000$ donc on a $class \sim Bi(n,p)$ (en ajoutant n Independent $Ber(p)$ donc $Ber(p)$ est la même que $Bi(1,p)$)

On sait que Y est complètement déterminé par p donc on a :

- $\mathrm{E}[Y] = p\,.1 + (1-p).0 = p$
- $\mathrm{V}[Y] = p\,(1-p)$

Et en régression logistique on a :

$$p(x) = P[Y=1|X=x] = E[Y|X=x]$$

$$E[p(tout\ x)] = E\Big[\,\mathrm{E}[Class = 0\,|X=x]\Big] = \frac{700}{1000} = 0.7$$

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.

```
> print(c(Accuracy = acc, Cutoff = cut))
  Accuracy Cutoff.476
 0.7850000  0.5173151
```