BE3 – Plans d'expériences et régression logistique

- **Exercice 1**

On étudie dans cet exercice les facteurs influençant l'apparition de défauts lors de la fabrication de plaques de silicium. Le défaut de planéité peut être influencé par six facteurs potentiels : Ltime, Ltemps, Lpress, Ctemp, Ctime, Catmos.

1.1 On commence par étudier le plan d'expérience mis en place pour étudier l'influence des différents facteurs sur l'apparition de défauts de planéité. Les résultats sont consignés dans le fichier *silicium.txt*, la variable *Camber* rend compte de la cambrure de la plaque de silicium obtenue et donc de l'apparition d'un défaut de planéité.

6 variables explicatives sont testées à un niveau 1 ou -1 en un jeu de 64 expériences (=2^6 expériences). On comprend que les valeurs des variables sont normalisées, par exemple pour Ltemp 1 peut signifier qu'on est à température maximale et -1 à température minimale.

Le plan d'expérience est conçu idéalement pour minimiser la variance des estimateurs beta. De ce fait, avec X la matrice qui répertorie les 64 expériences en ligne avec les variables explicatives en colonne, on étudie la matrice ^tXX calculée avec R : celle ci est une matrice diagonale avec uniquement la valeur 64 en chacune de ses valeurs diagonales :

	Ltemp	Ltime	Lpress	Ctemp	Ctime	Catmos
Ltemp	64	0	0	0	0	0
Ltime	0	64	0	0	0	0
Lpress	0	0	64	0	0	0
Ctemp	0	0	0	64	0	0
Ctime	0	0	0	0	64	0
Catmos	0	0	0	0	0	64

Ainsi les colonnes de ^tXX sont orthogonales entre elles et on pourrait penser que l'on a à faire à un plan d'expérience orthogonal complet.

Or, chaque expérience est « répétée quatre fois » selon l'énoncé, on observe bien un bloc de 16 (=2^4 expériences) expériences différentes répété quatre fois.

De ce fait, **le plan d'expérience est un plan d'expérience fractionnaire 2^{6-2} de résolution 4.**

En étudiant les 16 premières lignes, on note que : (A.B.C = E) et (A.C.D = F). Autrement dit, Ctemp est une interaction triple de Ltemp, Ltime et Lpress , et Catmos est une interaction triple de Ltemp, Lpress, Ctemp. De ce fait, les effets principaux se confondent tous avec des interactions triples. Or, comme dans le cours, on fait l'hypothèse que les interactions triples sont d'influence inenvisageable et donc négligeables. De ce fait, **les effets principaux peuvent être estimés sans confusion. Cependant, ce n'est pas le cas des effets des interactions doubles**.

1.2 Pour analyser la variance d'un modèle additif de la variable Camber en fonction des 6 facteurs, on effectue une ANOVA de type III dont les résultats sont présentés ci dessous :

```
Anova Table (Type III tests)

Response: Camber
            Sum Sq Df  F value    Pr(>F)
(Intercept) 732950  1 498.5563 < 2.2e-16 ***
Ltemp        24219  1  16.4740 0.0001520 ***
Ltime          535  1   0.3638 0.5488232
Lpress       50232  1  34.1681 2.574e-07 ***
Ctemp         3235  1   2.2003 0.1434922
Ctime        19010  1  12.9304 0.0006762 ***
Catmos       96023  1  65.3150 5.028e-11 ***
Residuals    83798 57
```

On effectue un test à 5 %, les variables Ltime et Ctemp sont celles qui n'ont pas une influence significative dans le modèle.

1.3 On présente ci-dessous le summary du modèle en régression linéaire :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.016      4.793  22.328  < 2e-16 ***
Ltemp         19.453      4.793   4.059 0.000152 ***
Ltime          2.891      4.793   0.603 0.548823
Lpress        28.016      4.793   5.845 2.57e-07 ***
Ctemp         -7.109      4.793  -1.483 0.143492
Ctime        -17.234      4.793  -3.596 0.000676 ***
Catmos       -38.734      4.793  -8.082 5.03e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.34 on 57 degrees of freedom
Multiple R-squared:  0.6975,    Adjusted R-squared:  0.6657
F-statistic: 21.91 on 6 and 57 DF,  p-value: 3.566e-13
```

La statistique testée est une statistique de Fischer sur 6 variables à 57 (64 expériences moins (6+1) facteurs libres) degrés de liberté. La p-valeur obtenue est très inférieure à 5 % et Ltime et Ctemp sont à nouveau sans influence notable dans la régression linéaire calculée.

L'erreur standard peut être calculée à la main avec la formule

$$Std\ Error = (\hat{\sigma}^2 . {}^tXX^{-1})^{1/2}$$

Et dans notre cas la valeur ${}^tXX^{-1}$ vaut 1/64. De ce fait, la valeur de l'erreur standard peut se retrouver en calculant $\hat{\sigma}/(64^{1/2})$. L'estimation de sigma est obtenue en vert dans le summary et vaut 38,34. P**ar le calcul, on obtient une erreur standard de 4,7925, en accord avec la valeur de 4,793 du summary.**

1.4 On modélise à nouveau la variable Camber par régression linéaire en ne prenant en compte cette fois ci que les quatre variable explicatives : Ltemp, Lpress, Ctime et Catmos. Le summary obtenu est le suivant :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.016      4.816  22.222  < 2e-16 ***
Ltemp         19.453      4.816   4.040 0.000157 ***
Lpress        28.016      4.816   5.818 2.59e-07 ***
Ctime        -17.234      4.816  -3.579 0.000698 ***
Catmos       -38.734      4.816  -8.043 4.62e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.53 on 59 degrees of freedom
Multiple R-squared:  0.6839,    Adjusted R-squared:  0.6625
F-statistic: 31.92 on 4 and 59 DF,  p-value: 3.711e-14
```

Les quatre coefficients obtenus avec ce modèle sont strictement similaires aux quatre coefficients associées aux mêmes variables explicatives (identifiés en jaune précédemment) dans le modèle précédent qui prenait en compte les six variables explicatives. De ce fait, **en supprimant deux variables explicatives, aucun potentiel effet d'interaction n'a été perdu.** Ceci valide la suppression des deux variables Ltime et Ctemp.

1.5 Pour minimiser la courbure de la plaque de silicium dans notre modèle, on veut d'une part maximiser l'influence des facteurs à influence négative sur la courbure, et d'autre part minimiser l'influence des facteurs à influence positive sur la courbure. Ainsi, on va se placer en 1 pour Ctime et Catmos et en -1 pour Ltemp et Lpress. Autrement dit, **les conditions expérimentales pour minimiser la courbure sont : Ctime 17,5 secondes ; Catmos 20°C ; Ltemp 55 °C et Lpress 5bars.**

On cherche maintenant un intervalle de confiance pour la courbure moyenne en ce point de fonctionnement optimal.

```
> data_new
  Ltemp Lpress Ctime Catmos
1    -1     -1     1      1
> predict(lm.sili3, data_new, interval = "confidence", level =0.95)
       fit      lwr      upr
1 3.578125 -17.9689 25.12515
```

Avec les conditions initiales évoquées précédemment, R situe *Camber* entre la lower value de -17,97 et la upper value de 25,12. En valeur absolue (cambrure toujours positive) et en arrondissant on obtient un intervalle [0 ; 25,2]. On souhaite que cet intervalle soit un intervalle de confiance à 95 %.

Connaissant la forme $[\bar{x} - 2\sigma(X)/\sqrt{n} ; \bar{x}+2\sigma(X)/\sqrt{n}]$ de l'intervalle,et l'erreur standard dans notre modèle qui vaut 38,53, notre intervalle de confiance à 95 % sera : ***Camber* ∈ [0 ; 102,06]**

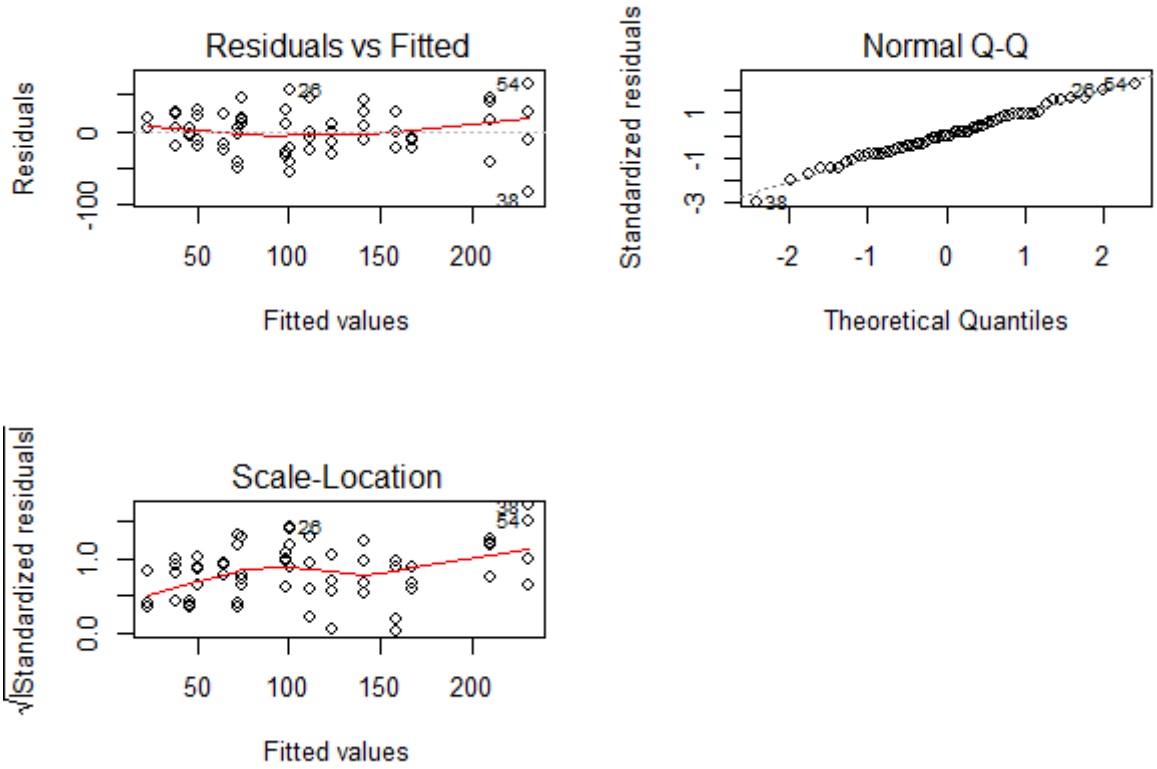
On étudie enfin l'influence dans notre modèle d'une augmentation de 5°C de la température de laminage de la plaque (soit +0,5 en valeurs normalisées par rapport à l'intervalle [55°C;75°C]).

```
> data_augment = data.frame (Ltemp = -0.5, Lpress=-1, Ctime=1, Catmos=1)
> predict(lm.sili3, data_augment)
       1
13.30469
```

Précédemment on avait en les points minimum une courbure attendue de 3,57. Ici, après une augmentation de 5°C on obtient une courbure de 13,30. Ainsi, **une telle augmentation de la température de laminage aboutit à une augmentation de 9,73 de la courbure**.

1.6 L'hypothèse d'une dépendance uniquement linéaire en les quatre variables explicatives Ltemp, Lpress, Ctime et Catmos n'est pas parfaitement vérifiée. En effet, le modèle de régression obtenu a un coefficient de régression R^2=0,6839 assez faible.

De plus, on plot la régression linéaire comme montré en haut de page suivante. Les résidus ne sont pas uniformément répartis, ce qui confirme l'**existence de dépendances qui restent à expliquer**.



1.7 On peut supposer qu'un meilleur modèle prendrait en compte les **interactions doubles entre les variables**. On met en œuvre un tel modèle et on obtient le summary suivant :

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          107.016      3.889  27.520  < 2e-16 ***
Ltemp                 19.453      3.889   5.003 6.58e-06 ***
Lpress                28.016      3.889   7.204 2.11e-09 ***
Ctime                -17.234      3.889  -4.432 4.72e-05 ***
Catmos               -38.734      3.889  -9.961 9.65e-14 ***
I(Ltemp * Lpress)     11.203      3.889   2.881  0.00571 **
I(Ltemp * Ctime)       9.078      3.889   2.335  0.02339 *
I(Ltemp * Catmos)     -9.859      3.889  -2.535  0.01422 *
I(Lpress * Ctime)      9.578      3.889   2.463  0.01706 *
I(Lpress * Catmos)    -6.109      3.889  -1.571  0.12212
I(Catmos * Ctime)     11.516      3.889   2.961  0.00458 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.11 on 53 degrees of freedom
Multiple R-squared:  0.8149,    Adjusted R-squared:  0.7799
F-statistic: 23.33 on 10 and 53 DF,  p-value: 5.193e-16
```

Au vu de ces tests, l'interaction double Lpress x Catmos n'a pas d'influence sur la courbure finale de la plaque, mais les cinq autres interactions doubles ont toute une influence significative sur la courbure de la plaque. Il est donc important de prendre ces interactions en compte pour un meilleur modèle.

Avec ce modèle, on obtient un coefficient de corrélation de R^2=0,8149 bien meilleur que dans le modèle linéaire simple proposé précédemment.

On évalue ce nouveau modèle en le point de fonctionnement minimal du modèle précédent. On obtient une courbure moyenne prévisionnelle de 23,61 dans un intervalle de confiance à 95 % de [0;111].

```
> predict(lm.interact, data_new, interval = "confidence", level =0.95)
       fit       lwr      upr
1 23.60938 -2.259084 49.47783
```

Notons que cet i**ntervalle de confiance s'applique à la valeur de l'estimation de la courbure obtenu mais ne prend pas en compte le bruit** qui pourrait s'ajouter à cette expérience.

- **Exercice 2**

2.1 Le fichier *neuralgia.txt* décrit un ensemble de 60 patients soignés pour névralgie avec chacun quatre variables explicatives : l'âge (variabe à modalités mutliples), le sexe, (variable à deux modalités) le traitement suivi (variable à trois modalités) et la durée (variable à modalités multiples). La variable étudiée est *Pain* et représente la souffrance du patient : 0 s'il ne souffre pas et 1 s'il souffre. On visualise les premières lignes de ce fichier importé dans R ci-dessous:

```
> head(neuralgia)
  Treatment Sex Age Duration Pain
1         P   F  68        1    0
2         B   M  74       16    0
3         P   F  67       30    0
4         P   M  66       26    1
5         B   F  67       28    0
6         B   F  77       16    0
```

2.2 On partage le fichier des tests en **deux parties en vue d'un apprentissage par régression logistique**. Le fichier d'apprentissage contient 80 % des données, soit les données de 48 patients. Le fichier de test contient les 20 % (12 patients) restants.

2.3 On réalise une régression logistique sur l'ensemble d'apprentissage pour prédire la variable *Pain*. Alors, l'événement que l'on modélise est « Il y a *Pain* » et en notant π(x) la probabilité de cet événement, le **modèle logit consiste à écrire cette probabilité sous la forme** :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Où les Betas sont les paramètres que l'on cherche à estimer.

La fonction logit est : p → log(p/(1-p)) et on va l'appliquer à cette probabilité.

2.4 On réalise plusieurs tests successifs sur notre modèle logit pour révéler l'influence de certains facteurs.

La première analyse est une **anova avec un test = « Chisq », ie un test du Chi carré**. On obtient les résultats affichés en haut de page suivante :

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Pain

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         59     81.503
Treatment  2  14.0230        57     67.480 0.0009015 ***
Sex        1   7.5945        56     59.886 0.0058545 **
Age        1  11.1182        55     48.767 0.0008548 ***
Duration   1   0.0317        54     48.736 0.8586632
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Avec le test du Chi carré, la durée du traitement n’a aucune influence sur la douleur du patient. Le traitement employé, le sexe et l’âge du patient ont eux une influence.

On effectue ensuite une **anova de type III avec pour test un test de maximum de vraisemblance** (test= »LR ») . On obtient :

```
Analysis of Deviance Table (Type III tests)

Response: Pain
          LR Chisq Df Pr(>Chisq)
Treatment  19.6882  2  5.306e-05 ***
Sex         6.3000  1    0.012074 *
Age        10.4769  1    0.001209 **
Duration    0.0317  1    0.858663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dans ce cas, la durée du traitement n’a toujours pas d’influence et le traitement employé, l’âge et le sexe du patient ont une influence. Notons néanmoins que l’influence estimée de ces trois variables explicatives est différente du test précédent. Ici, le type de traitement appliqué a une plus forte influence ainsi que l’âge, le sexe ayant une influence du même ordre de grandeur.

Enfin, on effectue une Anova avec un test de Wald-Wolfowitz. On obtient :

```
Analysis of Deviance Table (Type III tests)

Response: Pain
             Df   Chisq Pr(>Chisq)
(Intercept)  1  8.4018   0.003749 **
Treatment    2 12.5334   0.001898 **
Sex          1  5.2954   0.021382 *
Age          1  7.2990   0.006899 **
Duration     1  0.0315   0.859054
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ici, même conclusion sur la non influence de la durée du traitement. De la même manière, les influences estimées du type de traitement, de l’âge et du sexe du patient sont différentes des analyses précédentes.

On visualise ci dessous le summary du modèle logit :

```
glm(formula = Pain ~ Treatment + Sex + Age + Duration, family = binomial(link = "logit"),
    data = neuralgia)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7638  -0.5904  -0.1952   0.6151   2.3153

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.588282   7.102883  -2.899  0.00375 **
TreatmentB   -0.526853   0.937025  -0.562  0.57394
TreatmentP    3.181690   1.016021   3.132  0.00174 **
SexM          1.832202   0.796206   2.301  0.02138 *
Age           0.262093   0.097012   2.702  0.00690 **
Duration     -0.005859   0.032992  -0.178  0.85905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 81.503  on 59  degrees of freedom
Residual deviance: 48.736  on 54  degrees of freedom
AIC: 60.736

Number of Fisher Scoring iterations: 5
```

A la lecture de ce summary, on conclut encore une fois sur la large non-influence (en rouge) de la durée du traitement sur la douleur ressentie par le patient.

Le sexe et l’âge du patient sont tout deux influents. Concernant le traitement, on peut préciser à la lecture du summary que le traitement B n’a aucune influence sur *Pain* alors que le traitement P en a une grande (en vert).

2.5 On effectue maintenant une **procédure forward pour construire le modèle basé sur les variables les plus significatives.** Celui ci fonctionne en ajoutant à chaque itération une variable au modèle. La variable ajoutée sera celle qui minimisera le critère AIC. On arrête les itérations quand toutes les variables auront été ajoutées ou bien quand toutes les variables restantes à ajouter dépassent un certain seuil (5%) du critère.

Dans notre cas, on obtient les itérations et le résultat (en vert) affichés ci dessous :

```
Start:  AIC=83.5
Pain ~ 1

           Df Deviance    AIC
+ Treatment  2   67.480 73.480
+ Age        1   73.056 77.056
+ Sex        1   75.849 79.849
<none>           81.503 83.503
+ Duration   1   79.886 83.886

Step:  AIC=73.48
Pain ~ Treatment

           Df Deviance    AIC
+ Age        1   55.044 63.044
+ Sex        1   59.886 67.886
<none>           67.480 73.480
+ Duration   1   66.688 74.688
```

```
Step:  AIC=63.04
Pain ~ Treatment + Age

           Df Deviance    AIC
+ Sex        1   48.767 58.767
<none>           55.044 63.044
+ Duration   1   55.036 65.036

Step:  AIC=58.77
Pain ~ Treatment + Age + Sex

           Df Deviance    AIC
<none>           48.767 58.767
+ Duration   1   48.736 60.736
```

Le modèle obtenu par cette procédure confirme les analyses faites précédemment. Il sélectionne les variables explicatives *Traitement, Age* et *Sex* avec *Traitement* qui a la plus grande influence car sélectionné en premier.

2.6 Pour le modèle logit, la matrice de confusion obtenue est présentée ci dessous :

```
> table(liste_predict>0.5, liste_test$Pain)

       0 1
FALSE 6 2
TRUE  0 4
```

Avec ce modèle appliqué au fichier de tests comprenant 12 patients, 10 sont bien classés mais 2 sont déclarés sans douleur alors qu'ils ressentent de la douleur. Ainsi, avec ce modèle, **16,7 % des instances seront a priori mal classées**.

Le modèle obtenu à la suite de la procédure forward, après prédiction sur le fichier des 20 % patients tests, donne la matrice de confusion suivante :

```
> table(liste_predictforward>0.5, liste_test$Pain)

       0 1
FALSE 6 2
TRUE  0 4
```

Même si les modèles sont strictement différents en les influences qu'ils affectent aux différentes variables explicatives, on obtient la même matrice de confusion où 16,7 % des instances sont mal classées.

2.7 On choisit dans cette question d'utiliser uniquement le modèle issu de la procédure forward. On choisit 50 fois d'affilée un ensemble au hasard de 80 % des patients (ie 48 personnes) et on teste à chaque itération la sensibilité des qualités prédictives sur le fichier test des patients restants (20 % - 12 patients). Dans ce cas 80/20 on obtient les erreurs suivantes :

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
   2    2    3    3    2    2    3    2    3     3     2     1     3     3     3     2     3     3     2     3     3     3
[,23] [,24] [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43]
   5     2     1     2     2     4     2     0     4     1     1     4     1     0     3     1     1     1     3     2     3
[,44] [,45] [,46] [,47] [,48] [,49] [,50]
   3     3     3     0     4     0     2
```

Le nombre de patients mal classés oscille entre 0 et 4 personnes mal classés sur 12 au total. En moyenne 19,8 % des patients seront mal classés.

Dans le cas d'un **apprentissage sur 60 % des patients** et d'un test sur 40 % des patients on obtient qu'en moyenne 14,5% des patients seront mal classés.

Dans le cas d'un **apprentissage sur 70 % des patients** et d'un test sur 30 % des patients on obtient qu'en moyenne 21% des patients seront mal classés

Dans le cas d'un **apprentissage sur 90 % des patients** et d'un test sur 30 % des patients on obtient qu'en moyenne 0 % des patients seront mal classés.

Ainsi, le modèle avec le moins d'erreurs de classement des patients en moyenne est celui effectuant un apprentissage sur 90% des patients. Notons que cette conclusion est très relatives puisque l'ensemble de patients sur lequel nous travaillons est petit (60 patients) et rend les conclusions sur les pourcentages moyens d'erreurs moins fiables (un pourcentage sur 12 cas de test est peu significatif…). De la même manière, le nombre de prédictions (50) effectuées dans chaque cas d'apprentissage est faible et abaisse la significativité des résultats obtenus.

2.8 Pour un modèle plus complexe, on prend en compte les interactions des différentes variables explicatives. En l'occurence, on étudie maintenant dans R : *glm = (Treatment + Sex + Age + Duration)^2*