Projet Recherche : introduction à l'analyse de données de survie

Douâa Gratet, Berivane Samou, Wassim Ben Youssef

27 Mai 2016

Table des matières

Introduction

En médecine, en biologie ou en industrie on s'intéresse souvent à des durées, qui caractérisent principalement le temps qui s'écoule entre un instant de départ et un événement précis tel que la mort d'un individu ou la défaillance d'une machine. Cette durée est appelée durée de survie. Nous pouvons illustrer cette durée par un exemple en médecine : un individu atteint d'une maladie commence à prendre un traitement. L'individu est ensuite suivi jusqu'à l'instant de son décès. L'instant où l'individu a commencé le traitement est l'instant initial et l'instant de son décès est l'instant final, la durée entre ces deux instants correspond à la durée de survie. Cependant, lors de ce type d'études effectués sur plusieurs individus, certaines personnes sont amenés à disparaître de l'étude pour différentes raisons : déménagement, mort pour une raison qui n'a aucun lien avec la maladie … Dans ce cas là, l'instant final doit être considéré différemment que pour les autres individus car il ne correspond pas à la mort de l'individu à cause de la maladie : on appelle ce type de données des données censures. Une base de données contenant ce type de données ne peut pas être sujet à une analyse avec des outils classiques de statistique et nécessite une approche plus particulière que nous allons expliquer dans ce rapport.

Dans un premier temps, nous allons décrire les données sur lesquelles nous allons mener notre étude. Nous nous intéressons à la base de données "Veteran" du package "Survival", qui répertorie des données sur 137 individus atteints d'un cancer du poumon sur lesquelles les médecins testent deux types de traitement. La variable à expliquer est la variable "time" qui est la durée entre l'instant initial (le début de l'étude) et la mort (ou la disparition dans le cas d'une donnée censure) correspondant à l'instant final. On dispose de 7 autres variables explicatives : la variable binaire "status" qui va mentionner si la valeur "time" de l'individu est une donnée censurée, la variable binaire "trt" qui indique quel type de traitement prend l'individu (le traitement standard ou le traitement test), la variable "celltype" qui précise le type de cellules touchées ("squamous", "smallcell", "adeno" ou "large"), la variable quantitative Karno qui prend une valeur entre 0 et 100 et qui indique le niveau de vie et l'hygiène de vie (avec 0 le moins bon et 100 le meilleur), la variable « âge » qui indique l'âge de l'individu en années, la variable binaire « prior » qui indique si l'individu suit une thérapie prioritaire ou non et enfin la variable quantitative « diagtime » qui indique le temps en mois qui s'est écoulé depuis le dernier diagnostique.

Préliminaires : Etude théorique des données de survie

Pour faire une étude sur une base de données contenant des données censurées, on ne peut pas simplement les supprimer pour ne pas qu'elles posent de problème, cela entrainerai une perte importante d'information. Il faut donc utiliser des outils adaptés à ce type d'étude. Notre objectif est d'étudier la loi du temps de survie qui est régie par 5 fonctions principales :

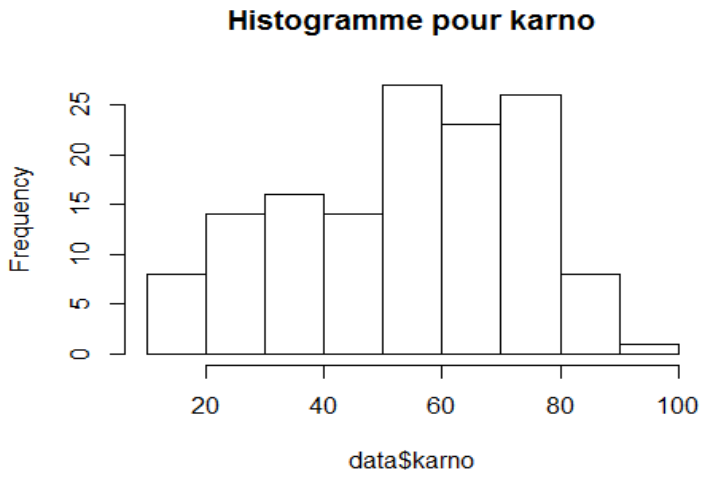
- La fonction de survie, qui représente la probabilité de survivre jusqu'à l'instant t, en notant T l'instant de décès $S(t) = P(T > t)$
- La fonction de répartition $F(t) = P(T \leq t)$
- La fonction de densité f(t) avec $F(t) = \int f(x)dx$
- Le risque instantané de décès $h(t) = lim(dt -> 0)1/dt\ P(t < T \leq t + dt \mid T > t)$
- Le risque cumulé de décès $H(t) = \int h(x)dx$

Nous retrouvons également les relations suivantes entre ces différentes fonctions :

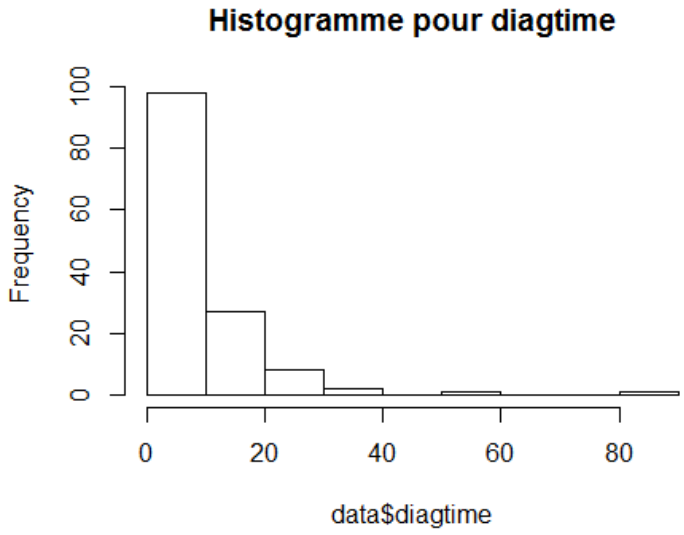
- $F(t) = 1 - S(t)$
- $f(t) = -S'(t)$
- $H(t) = -ln(S(t))$

Première partie : analyse classique avec les outils habituels

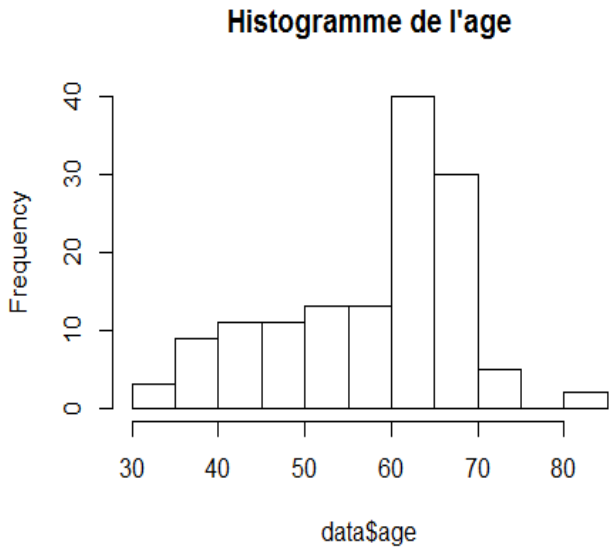
Pour commencer, nous allons effectuer une analyse classique des données en observant la répartition de chacune des variables à l'aide d'histogrammes.



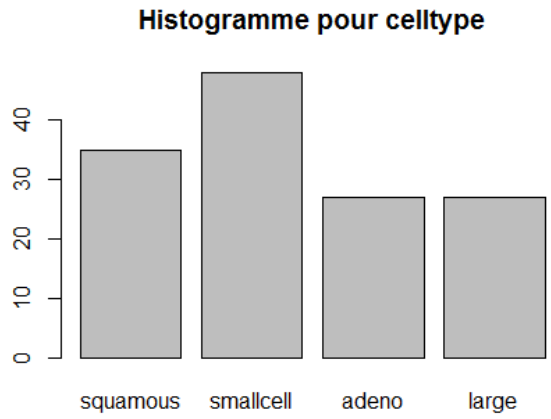
La partie la plus importante des individus se situe entre 50 et 80 pour le score de Karnofsky. Entre 20 et 50, c'est réparti quasi équitablement alors qu'entre 80 et 100 il n'y a pas beaucoup d'individus.



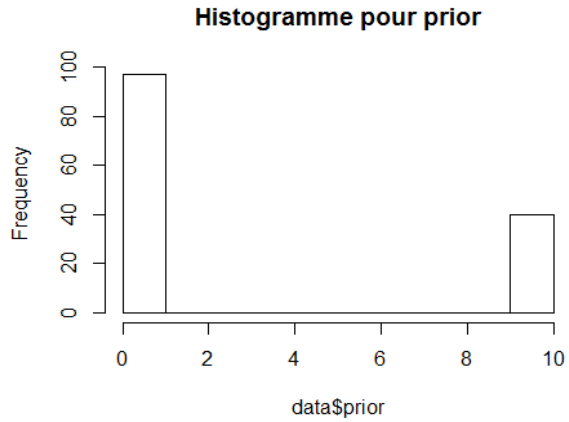
La majorité des individus ont un diagtime entre 0 et 10 mois. Plus on avance dans le temps moins il y a de personnes.



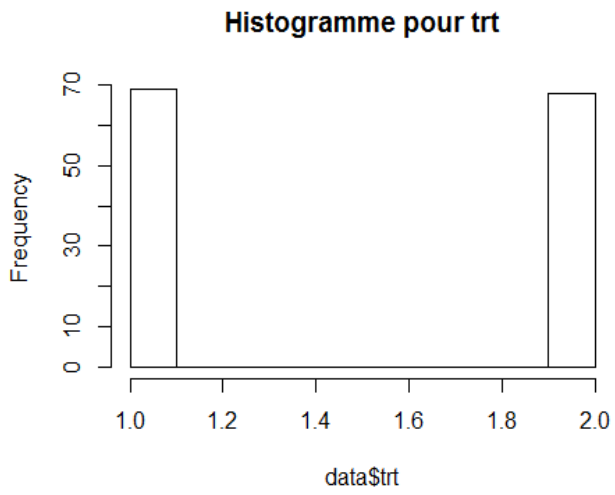
La majorité des personnes atteintes se situe entre 60 et 70 ans. Néanmoins, le nombre de personnes atteintes entre 35 et 60 ans n'est pas négligeable.



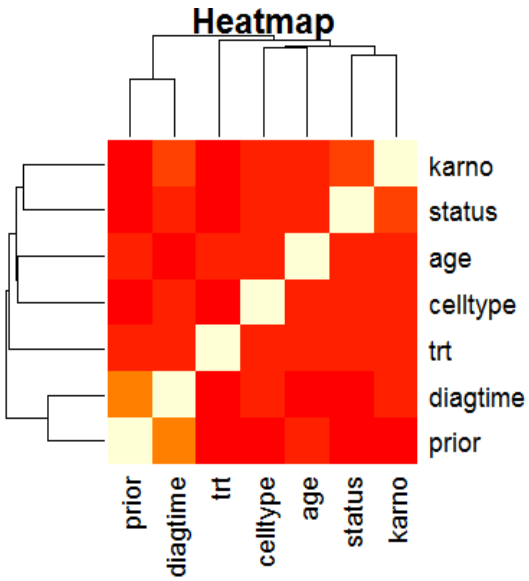
Les types de cellule "squamous et smallcell" des personnes atteintes représentent le plus grand taux.



La grande majorité des individus n'ont pas suivi une thérapie prioritaire.



Le nombre de personnes qui suivent le traitement est équivalent au nombre de personnes qui ne le suivent pas.



D'après le graphe de corrélation, on remarque que seul les variables diagtime et prior sont un peu corrélées.

Deuxième partie : les tests non paramétriques qui ne font pas d'hypothèse sur la distribution des échantillons

Les estimateurs non-paramétriques :

Pour obtenir la fonction de survie de données censurées, qui est la probabilité qu'un individu soit encore vivant à l'instant t, les outils usuels d'analyse ne sont pas adaptés. On utilise donc une autre approche avec des estimateurs non-paramétriques.

Les estimateurs non-paramétriques sont utilisés pour estimer un modèle qui n'est pas décrit par un nombre fini de paramètres, dans notre cas nous l'utilisons car nous n'avons pas toutes les informations sur un paramètre.

Estimateur de Kaplan-Meier :

L'estimateur de Kaplan-Meier permet de déterminer la courbe de survie à partir des données empiriques. Il calcule les probabilités instantanées Si de survivre au-delà du temps Ti sachant que l'on a survécu jusqu'en T_i en utilisant D_i le nombre d'événement (décès, censure…) à l'instant T_i et N_i le nombre de patient ayant survécu jusqu'à T_i, R_i le nombre d'individus encore vivant à l'instant T_i (individus à risque) :

$$\widehat{S_i} = 1 - \frac{D_i}{R_i}$$

Au final, on a :

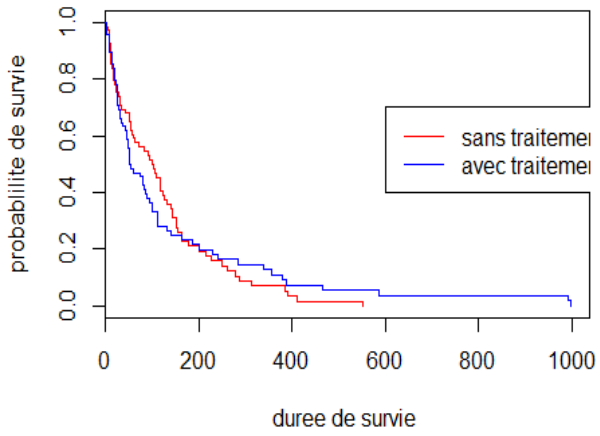
$$\hat{S}(t) = \begin{cases} 1 \; si \; t < t_1 \\ \prod_{i:\, t_i<t} (1 - \frac{D_i}{R_i}) \; si \; t_1 < t < t_D \\ indéterminé \; si \; t > t_D \; et \; C_D > 0 \end{cases}$$

Avec ci le nombre de censures à droite qui se produisent à ti sachant que les censures se font après les événements.

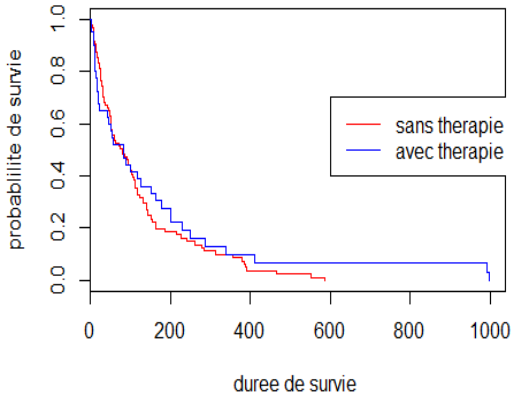
```
##   [1]  72  411  228  126  118   10   82  110  314  100+  42    8  144
25+
##  [15]  11   30  384    4   54   13  123+  97+ 153   59  117   16  151   22
##  [29]  56   21   18  139   20   31   52  287   18   51  122   27   54    7
##  [43]  63  392   10    8   92   35  117  132   12  162    3   95  177  162
##  [57] 216  553  278   12  260  200  156  182+ 143  105  103  250  100  999
##  [71] 112   87+ 231+ 242  991  111    1  587  389   33   25  357  467  201
##  [85]   1   30   44  283   15   25  103+  21   13   87    2   20    7   24
##  [99]  99    8   99   61   25   95   80   51   29   24   18   83+  31   51
```

```
## [113]  90   52   73    8   36   48    7  140  186   84   19   45   80   52
## [127] 164   19   53   15   43  340  133  111  231  378   49
```

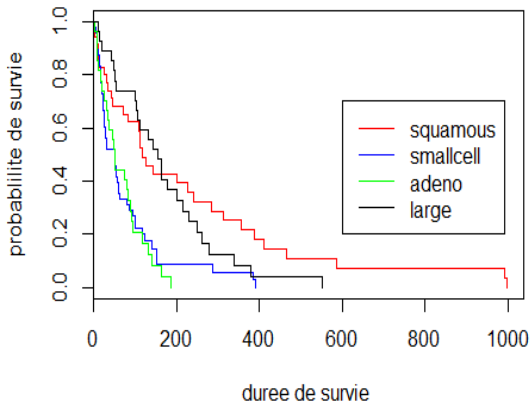
On commence par créer un tableau qui nous permettra de sélectionner les variables censurées grâce à la variable status. On voit dans ce tableau que les données censurées sont suivies d'une croix.



Ci-dessus la courbe de survie en fonction du traitement tracée à l'aide de l'estimateur de Kaplan-Meier, elle nous montre l'évolution de la probabilité de survie selon si les individus suivent le traitement ou non. On remarque qu'au début il n'y a pas de grande différence entre les 2 catégories mais qu'à la fin les personnes ayant suivi un traitement survivent plus longtemps que ceux ne l'ayant pas pris. Le traitement semble n'avoir d'effet qu'à très long terme.



Au début, les personnes ne suivant pas de thérapie ont quasiment la même probabilité de survie que ceux qui la suivent. Plus tard, les personnes suivant une thérapie ont une légère meilleure chance de survie mais qui n'est pas très significative.



Les individus avec les cellules "adeno" et "smallcell" ont une probabilité presque équivalente au début mais qui diffère légèrement à la fin, pareil pour ceux avec les cellules "squamous" et "large". Cependant, les individus avec la cellule "squamous" semblent avoir plus de chances de survie que les autres.

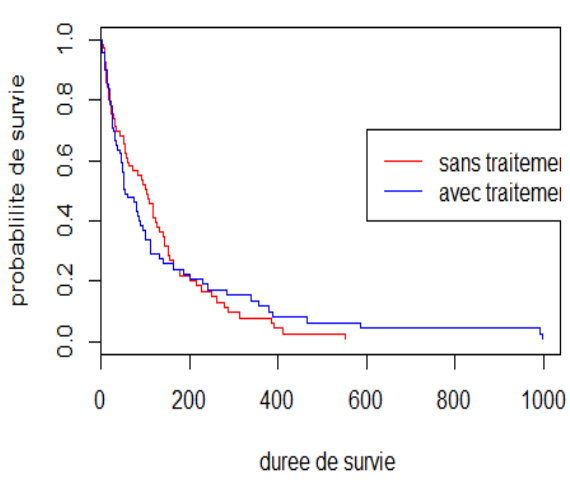
Estimateur de Nelson Aalen :

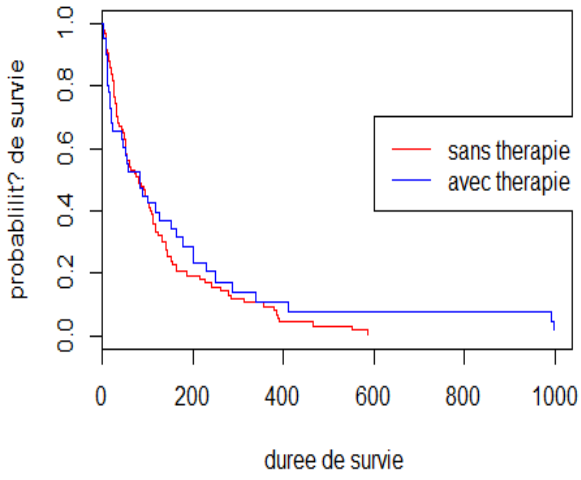
L'estimateur de Nelson-Aalen cherche l'estimateur cumulé de risque de décès, ou taux de décès cumulatif. Cet estimateur est calculé par :

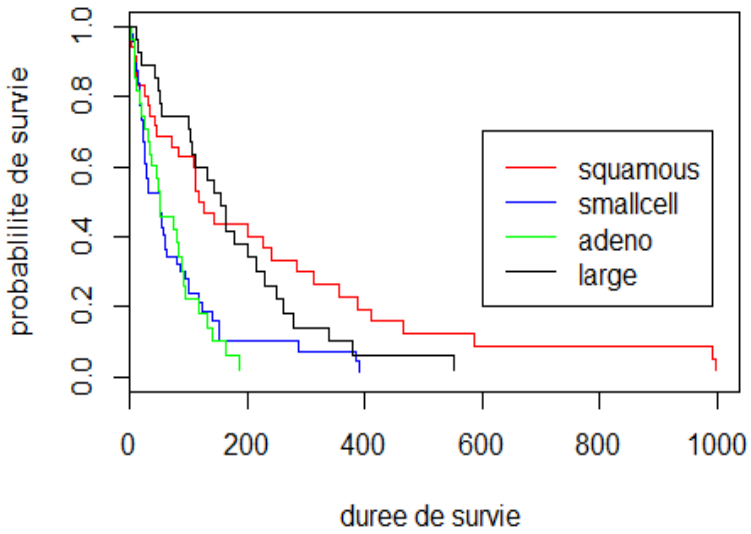
$$\widehat{H}(t) = \begin{cases} 0 \; si \; t < t_1 \\ \sum_{i:t_i} \frac{D_i}{R_i} \; si \; t_1 < t < \; t_D \\ indéterminé \; si \; t > t_D \; et \; C_D > 0 \end{cases}$$

Il estime ensuite la fonction de survie par :

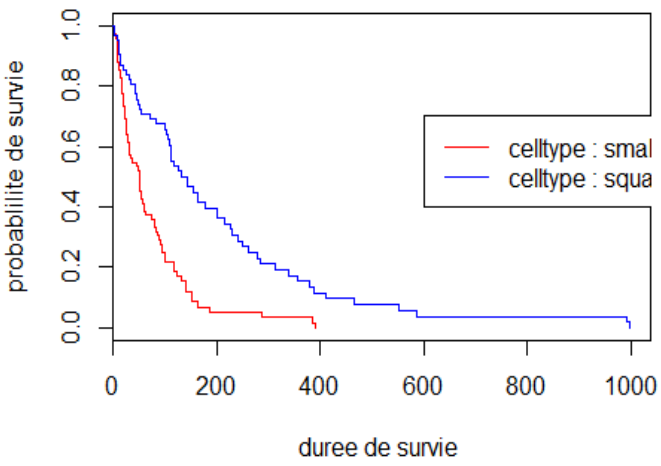
$$\tilde{S}(t) \quad = \quad \exp[-\tilde{H}(t)]$$







Les courbes ci-dessus ont été tracées par la méthode de l'estimateur de Nelson Aalen, elles sont très similaires aux précédentes. Donc nous pouvons faire les mêmes conclusions que précédemment.



On a vu que pour la variable qualitative celltype, on pouvait distinguer 2 groupes de courbes grâce aux courbes obtenues avec l'estimateur de Kaplan-Meier. Donc nous pouvons les fusionner en 2 groupes afin d'obtenir une variable binaire qui sera plus avantageuse lors des prochaines études que nous allons aborder. On regroupe "smallcell" et "adeno" qui avaient une moins bonne probabilité de survie et "squamous" et "large" dont la probabilité de survie est plutôt bonne.

Test de log-rank :

Ces tests servent à étudier l'importance des instances d'une variable sur la fonction de survie. On va donc chercher à savoir si les estimations de la fonction de survie par rapport aux différentes instances d'une variable sont significativement différentes ou non. Si elles sont différentes, cela signifie que la survie des individus dépend significativement des valeurs prises par cette variable, et donc que la variable a une importance pour la prédiction d'un modèle de survie. Nous utiliserons ces tests uniquement sur des variables avec deux instances. Par exemple, lorsque nous faisons le test de Wilcoxon sur la variable prior, nous obtenons une p-value assez proche de 1, ce qui signifie que le fait d'être en thérapie prioritaire ou non n'a pas vraiment d'influence sur la durée de survie des individus. Nous étudions ici deux tests de vraisemblances : le test de Wilcoxon et le test du log-rank.

Le test du log-rank va comparer deux courbes tracées par l'estimateur de Kaplan-Meier. Ces deux courbes correspondent à l'estimation de la fonction de survie par une variable binaire : chacune des courbes représente la fonction de survie en fonction d'une instance de la variable. Nous prenons comme hypothèse H0 que les deux courbes sont diférentes. Ne pouvant pas avoir de réel preuve, le test consistera à chercher si l'on peut rejeter cette hypothèse ou non. Ainsi, si on a une p-value supérieure à 5%, nous rejetterons cette hypothèse et dans l'autre cas, cela signifiera qu'on a de grandes chances de penser que les courbes ne sont pas similaires et donc que la variable étudiée a un rôle important par rapport

à la fonction de survie et à la survie de l'individu. Il est construit en comparant le nombre d'événements prédit à l'instant t avec le nombre d'événements observé à cet instant. Pour cela, nous prenons deux groupes d'individus, le premier vérifiant la première instance de la variable et le second vérifiant la deuxième instance. Nous notons Y_{1i} le nombre de sujets à risques (nombre de personnes qui ne sont pas encore décédées) pour le groupe 1 à l'instant i et Y_{2i} ce même nombre pour le groupe 2, avec $Y_i = Y_{1i} + Y_{2i}$. Notons également O_{1i} le nombre d'événements (disparition, décès d'un individu) observés à l'instant i pour le groupe 1 et O_{2i} pour le groupe 2 et $O_i = O_{1i} + O_{2i}$. Sous l'hypothèse H0, O_{1i} suis une distribution hypergéométrique de paramètres O_i, N_{1i} et N_i, la distribution attendue si les deux groupes ont la même fonction de survie et la même courbe de survie serait alors $E_{1i} = \frac{O_i}{N_i} N_{1i}$ de variance $Var_i = \frac{O_i\left(N_{1i}/N_i\right)(1-N_{1i}/N_i)(N_j - O_j)}{N_j - 1}$ La statistique du log-rank comparant E_{1i} et O_{1i} se calcule alors par $Z = \frac{\sum_{i=1}^{N}(O_{1i} - E_{1i})}{\sqrt{\sum_{i=1}^{N} Var_i}}$

Sous l'hypothèse H0, cette statistique pour le groupe 1 et le groupe 2 suit une loi normale. Ainsi, on rejette l'hypothèse H0 si la valeur de Z est supérieure à l'alpha quantile de la distribution normale, comme pour un test du Chi-2.

Ce test de log-rank analyse ainsi les courbes dans leur globalité. Il peut être significatif même si les deux courbes se rejoignent en fin de suivi, aboutissant ainsi au même nombre de décès dans chaque groupe. Cependant le test perd de son efficacité lorsque les deux courbes n'évoluent pas de façon proportionnelle, en particulier lorsque les courbes se croisent. L'analyse visuelle des courbes doit donc toujours accompagner l'interprétation d'un test du Log-rank.

```
## Call:
## survdiff(formula = survie ~ trt, data = data)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69       64     64.5   0.00388   0.00823
## trt=2 68       64     63.5   0.00394   0.00823
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.928
```

La p-value est très proche de 1 donc on peut rejeter l'hypothèse H0, on ne détecte donc pas de différence majeure entre les deux courbes tracées par l'estimateur de Kaplan-Meier sur la variable trt. Cela signifie que les deux instances de la variable trt ont quasiment le même effet sur la durée de survie des individus, cette variable n'affecte donc pas significativement la durée de survie des individus.

```
## Call:
## survdiff(formula = survie ~ prior, data = data)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0  97       91     87.4     0.150     0.501
## prior=10 40       37     40.6     0.323     0.501
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.479
```

La p-value est grande, donc il n'y aucune différence entre les deux instances de la variable prior, elles n'affectent pas la variable à expliquer, qui est le temps de survie de l'individu.

```
## Call:
## survdiff(formula = survie ~ celltype2, data = data)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype2=0 75       71     45.8     13.87      24.5
## celltype2=1 62       57     82.2      7.73      24.5
##
##  Chisq= 24.5  on 1 degrees of freedom, p= 7.34e-07
```

Pour la nouvelle variable celltype2 qu'on a créé la p-value est très petite, donc la valeur à expliquer prend des valeurs bien différentes en fonction des valeurs de celltype2, la variable celltype2 a donc une vraie influence pour la survie des individus celltype2 est donc la meilleure variable explicative parmi les 3.

Test de Wilcoxon :

Le test de Wilcoxon est similaire au test du log-rank mais il est plus sensible aux valeurs situées en début de courbes.

```
## Call:
## survdiff(formula = survie ~ trt, data = data, rho = 1)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69     32.2     35.4     0.279     0.871
## trt=2 68     35.2     32.1     0.308     0.871
##
##  Chisq= 0.9  on 1 degrees of freedom, p= 0.351
```

La p-value est très grande, donc on a la même conclusion qu'avec le test de log-rank.

```
## Call:
## survdiff(formula = survie ~ prior, data = data, rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0  97     47.6     48.2   0.00694    0.0366
## prior=10 40     19.8     19.3   0.01736    0.0366
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.848
```

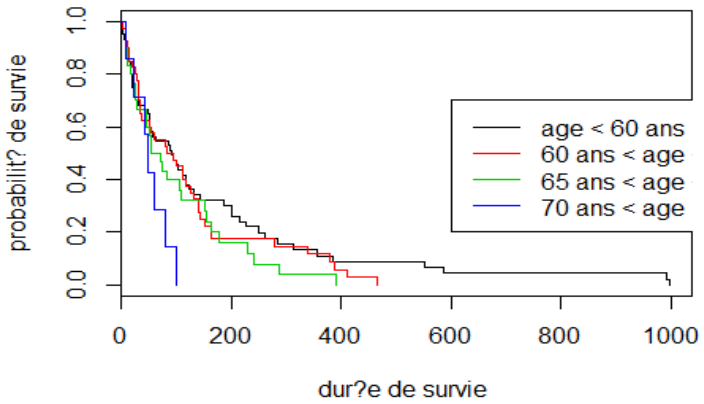
La p-value est grande, donc on a la même conclusion qu'avec le test de log-rank.

```
## Call:
## survdiff(formula = survie ~ celltype2, data = data, rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype2=0 75     44.5     29.8      7.23      19.5
## celltype2=1 62     23.0     37.6      5.73      19.5
##
##  Chisq= 19.5  on 1 degrees of freedom, p= 1e-05
```

La p-value est très petite, donc on a la même conclusion qu'avec le test de log-rank.

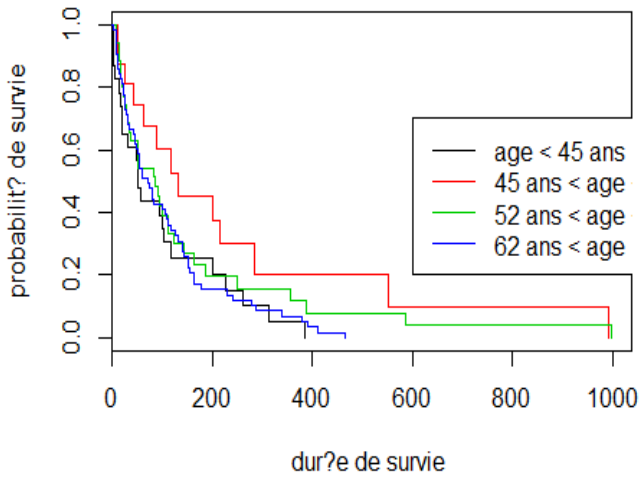
Regroupement des variables quantitatives :

La variable âge est quantitative, donc pour mieux l'exploiter on va essayer d'en obtenir une variable quantitative binaire en regroupant les individus en 2 classes. Pour commencer, nous séparons les individus en tranche d'âge :

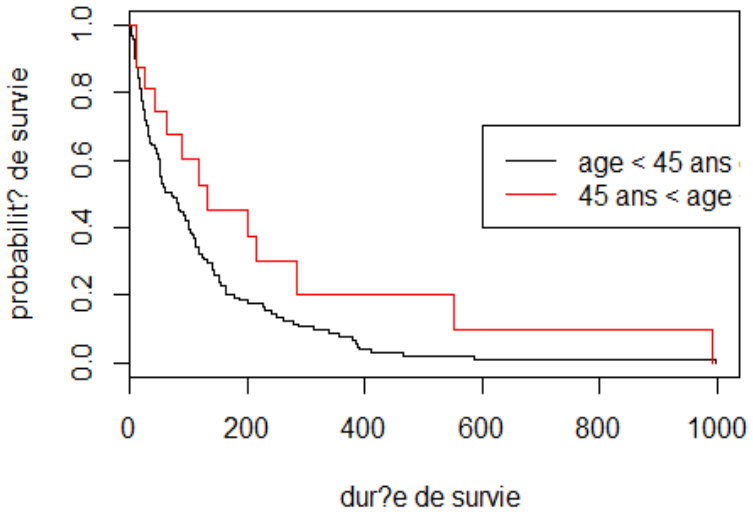


```
## Call:
## survdiff(formula = survie ~ data$age3)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age3=1 60        54    61.95    1.0207   2.08945
## data$age3=2 40        39    38.65    0.0031   0.00456
## data$age3=3 30        28    24.03    0.6561   0.82897
## data$age3=4  7         7     3.36    3.9272   4.15703
##
##  Chisq= 5.9  on 3 degrees of freedom, p= 0.117
```

D'après le graphe, on remarque que les courbes 3 de survie des personnes agées entre 60 et 65 ans, 65 et 70 ans et plus de 70 ans sont assez proches. Par la suite, nous faisons plusieurs essais en modifiant ces intervalles pour rapprocher plus clairement certaines courbes.



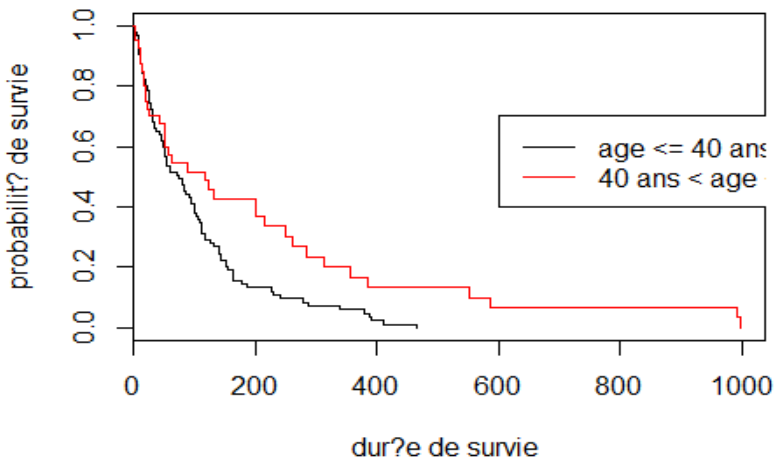
On observe ici que trois courbes sont très proches : la noire, la bleue et la verte. On choisit donc de les fusionner.




```
## Call:
## survdiff(formula = survie ~ data$age4)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age4=0 121      115    107.7     0.496      3.25
## data$age4=1  16       13     20.3     2.632      3.25
##
##  Chisq= 3.2  on 1 degrees of freedom, p= 0.0716
```

```
## Call:
## survdiff(formula = survie ~ data$age4, rho = 1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age4=0 121    61.92    58.22     0.236      2.67
## data$age4=1  16     5.53     9.23     1.486      2.67
##
##  Chisq= 2.7  on 1 degrees of freedom, p= 0.102
```

A présent que nous avons deux courbes, nous effectuons les tests du log-rank et de Wilcoxon sur celles-ci. Les p-values obtenues pour les deux tests sont supérieures à 5% ce qui signifie que la répartition que nous avons faite n'est pas très efficace et donc que la nouvelle variable créée n'est pas très utile pour prédire la survie des individus. En effectuant d'autres tests, on fait un nouveau regroupement de variables qui lui aussi n'est pas concluant au vu des p-values du test de Wilcoxon :

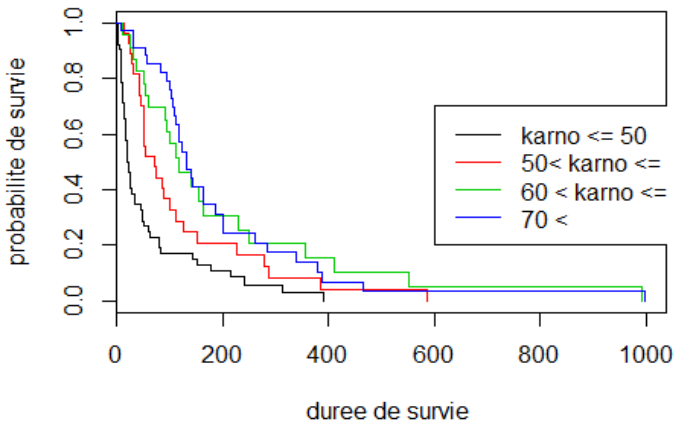



```
## Call:
## survdiff(formula = survie ~ age5, data = data, rho = 0)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## age5=0 97       93     79.7      2.21      6.51
## age5=1 40       35     48.3      3.64      6.51
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.0107
```

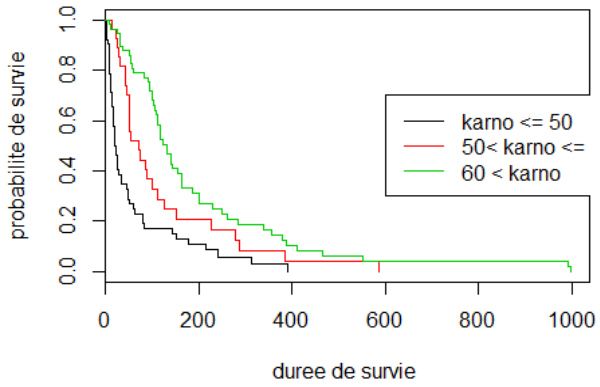
```
## Call:
## survdiff(formula = survie ~ age5, data = data, rho = 1)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## age5=0 97     50.5     46.2     0.392      1.91
## age5=1 40     17.0     21.2     0.853      1.91
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.167
```

A noter qu'ici, les deux courbes passent le test du log-rank mais pas le test de Wilcoxon qui est plus sensible aux premières valeurs. Ainsi, nous choisissons de ne pas utiliser de variable classant les individus en deux catégories selon leur âge.

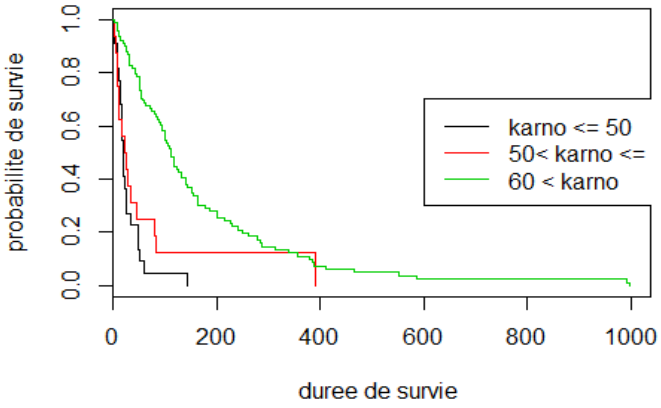
A présent, nous effectuons la même étude sur la variable Karno. A première vue, cette variable, qui quantifie le niveau et l'hygiène de vie, semblerait pertinente pour effectuer une étude de la survie des individus, le mode d'une vie d'une personne malade étant très important pour l'évolution de celui-ci. Il semble donc pertinent de vouloir classer les individus en deux classes : les personnes avec un niveau de vie assez faible et ceux ayant un niveau de vie plutôt bon.



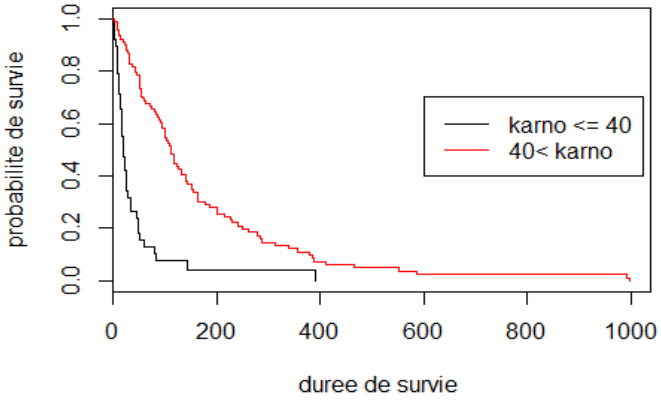
D'après le graphe, on remarque que les courbes bleu et verte sont quasi similaires donc on va les regrouper pour la suite.



On remarque qu'on ne peut pas regrouper les trois courbes, donc on va changer les bornes de l'intervalle pour voir si l'on peut avoir un modèle plus performant.



On remarque que les courbes de survie rouge et noire sont assez proches, donc on peut les fusionner.



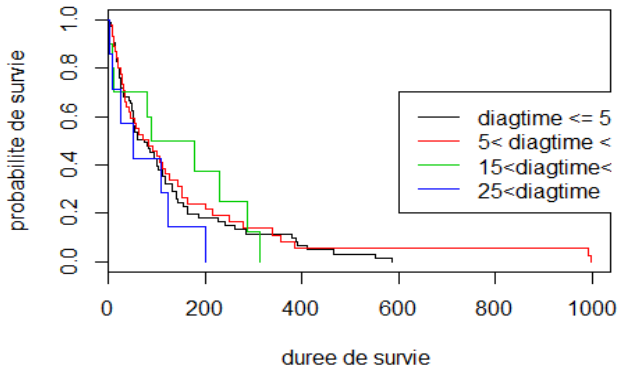
On obtient 2 courbes bien distinctes. On conclut que les individus avec un indice de Karnofsky supérieur à 40 ont plus de chance de survivre.

```
## Call:
## survdiff(formula = survie ~ data$karno5)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## data$karno5=0 38       37     14.1     36.94      44.5
## data$karno5=1 99       91    113.9      4.59      44.5
##
##  Chisq= 44.5  on 1 degrees of freedom, p= 2.55e-11
```

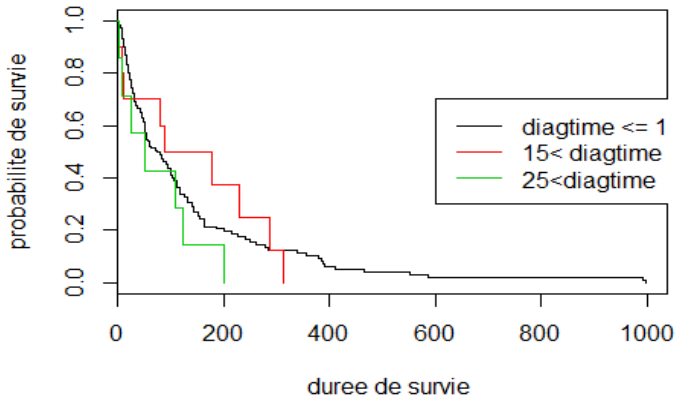
```
## Call:
## survdiff(formula = survie ~ data$karno5, rho = 1)
##
```

```
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## data$karno5=0  38     28.3     9.85     34.70      55.2
## data$karno5=1  99     39.1    57.59      5.94      55.2
##
##  Chisq= 55.2  on 1 degrees of freedom, p= 1.1e-13
```

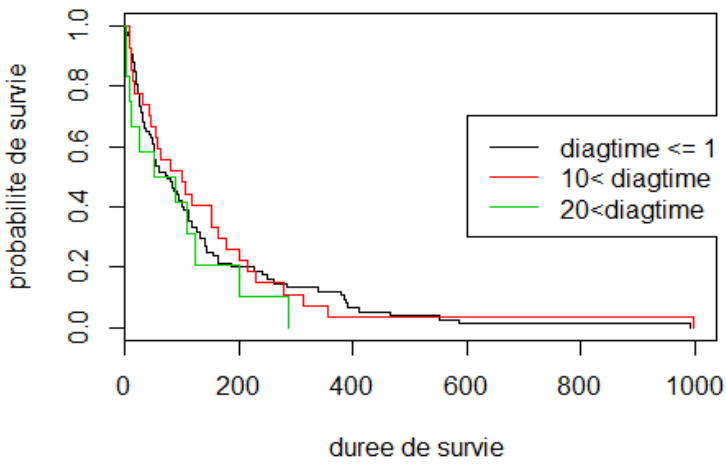
On effectue un test de log-rank et de wilcoxon sur ces 2 courbes, et on obtient des p-values très petites donc on rejette l'hypothèse qu'elles soient proches. Cela signifie que notre regroupement des variables karno semble efficace et nous permet d'obtenir une variable assez pertinente pour la prédiction de la survie.



On remarque que les courbes rouge et noire sont quasi identiques donc on peut les fusionner.



On obtient 3 courbes distinctes, on essaie d'obtenir 2 courbes proches en modifiant les bornes de l'intervalle.



Après plusieurs essais, on n'arrive pas à fractionner la variable en 2 groupes distincts. Donc on garde la variable de base.

Modèle de Cox :

C'est une classe de modèles de survie en statistiques. Les modèles de survie étudient le temps écoulé avant qu'un événement ne survienne. Historiquement, dans le modèle de Cox, cet événement est le décès de l'individu, c'est pourquoi on parle généralement de survie et de décès. Au cours des années, l'utilisation du modèle s'est étendue à d'autres situations, l'événement peut donc être de quelconque nature : il peut s'agir de la récidive d'une maladie, ou à l'inverse d'une guérison. D'un point de vue statistique, la nature de l'événement n'est bien sûr pas importante, il s'agira alors d'interpréter les coefficients en conséquence.

Le modèle de Cox exprime la fonction de risque instantané de décès λ (on peut aussi trouver les appellations suivantes : fonction de risque, taux de panne, taux de fiabilité, force de mortalité, taux de risque...) en fonction du temps t et des covariables $X_1 \ldots X_n$. On a alors :

$$\lambda\left(t, X_1, \cdots, X_n\right) = \lambda_0\left(t\right) exp\left(\Sigma_{i=1}^{n} \beta_i X_i\right)$$

Où $\lambda_0(t)$: le risque de base, il correspond au risque instantané de décès lorsque toutes les covariables sont nulles.

Et βi : des constantes inconnues représentant les paramètres de régression.

Pour utiliser ce modèle, on utilise la fonction Coxph en R.

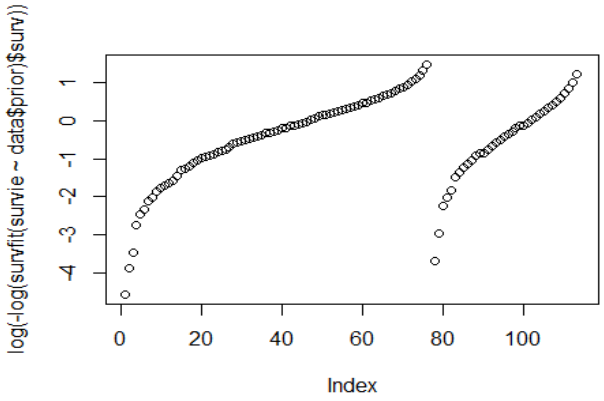
Pour effectuer le modèle de Cox, on crée de nouvelles variables à partir des interactions entre les variables binaires. Nous commençons par faire un premier modèle sans tenir compte de la variable Karno5, en prenant toutes les variables, puis nous faisons une séléction de variables à l'aide de stepAIC avec la méthode Backward qui donnait de meilleurs résultats :

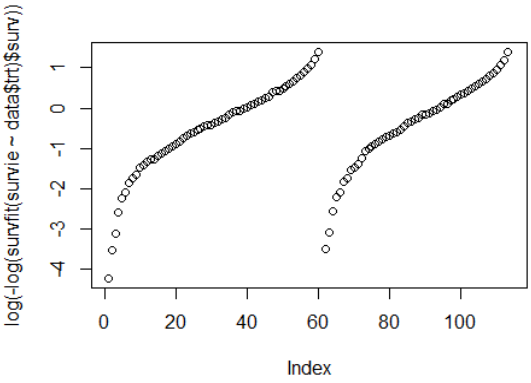
```
## Start:  AIC=964.68
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$diagtime + data$karno + data$trt_celltype2 + data$trt_prior +
##     data$prior_celltype2
##
##                         Df    AIC
## - data$diagtime          1 962.68
## - data$prior_celltype2   1 962.69
## - data$trt_celltype2     1 963.87
## - data$prior             1 964.09
## <none>                     964.68
## - data$age               1 964.74
## - data$trt_prior         1 966.05
## - data$celltype2         1 966.95
## - data$trt               1 970.23
## - data$karno             1 995.59
##
## Step:  AIC=962.68
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_celltype2 + data$trt_prior +
data$prior_celltype2
##
##                         Df    AIC
## - data$prior_celltype2   1 960.69
## - data$trt_celltype2     1 961.91
## - data$prior             1 962.41
## <none>                     962.68
## - data$age               1 962.75
## - data$trt_prior         1 964.23
## - data$celltype2         1 965.03
## - data$trt               1 968.24
## - data$karno             1 994.85
##
## Step:  AIC=960.69
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_celltype2 + data$trt_prior
##
##                        Df    AIC
## - data$trt_celltype2    1 959.92
## <none>                    960.69
## - data$age              1 960.81
## - data$prior            1 961.06
## - data$trt_prior        1 962.29
## - data$celltype2        1 964.35
## - data$trt              1 966.26
## - data$karno            1 992.98
##
## Step:  AIC=959.92
```

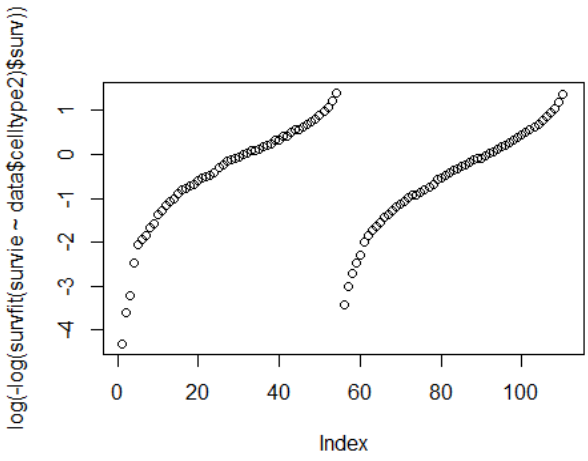
```
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_prior
##
##                   Df    AIC
## <none>               959.92
## - data$age         1 960.20
## - data$prior       1 960.73
## - data$trt_prior   1 962.40
## - data$trt         1 964.51
## - data$celltype2   1 974.34
## - data$karno       1 993.29
```

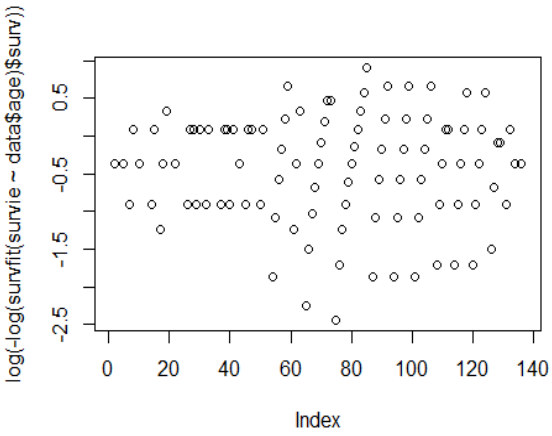
```
## Call:
## coxph(formula = survie ~ data$prior + data$trt + data$celltype2 +
##     data$age + data$karno + data$trt_prior)
##
##                     coef exp(coef) se(coef)     z       p
## data$prior       0.48346   1.62167  0.28047  1.72   0.085
## data$trt         0.59704   1.81673  0.23365  2.56   0.011
## data$celltype2 -0.81595   0.44222  0.20348 -4.01 6.1e-05
## data$age       -0.01492   0.98519  0.00973 -1.53   0.125
## data$karno     -0.03249   0.96803  0.00546 -5.95 2.7e-09
## data$trt_prior -0.89397   0.40903  0.42456 -2.11   0.035
##
## Likelihood ratio test=63  on 6 df, p=1.11e-11
## n= 137, number of events= 128
```

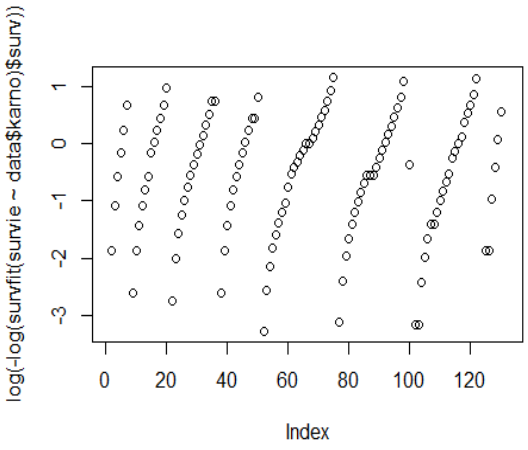
Pour la séléction de variables, nous avons utilisé le stepAIC sur le modèle de cox avec la méthode "backward", il nous indique que les meilleures variables explicatives sont : prior, trt, celltype2, age, karno et trt_prior. A présent nous allons tester si ces variables peuvent effectivement être utilisées dans notre modèle de Cox.

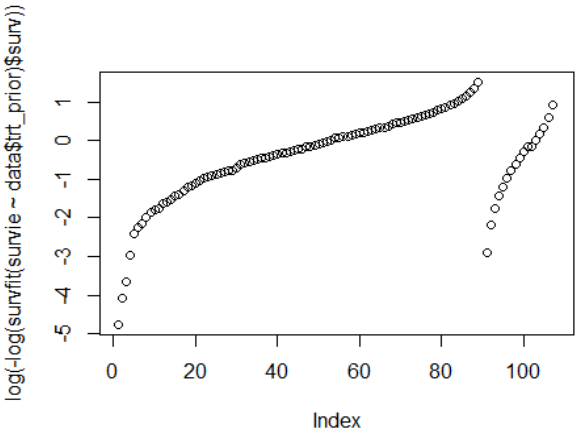












Ces graphes représentent le log du risque instantané cumulé de décès pour chaque variable. Quand les courbes sont parallèles, cela signifie que le modèle de Cox est bon sur ces variables. Or, on observe que ce n'est pas le cas pour les variables : âge et trt_prior.

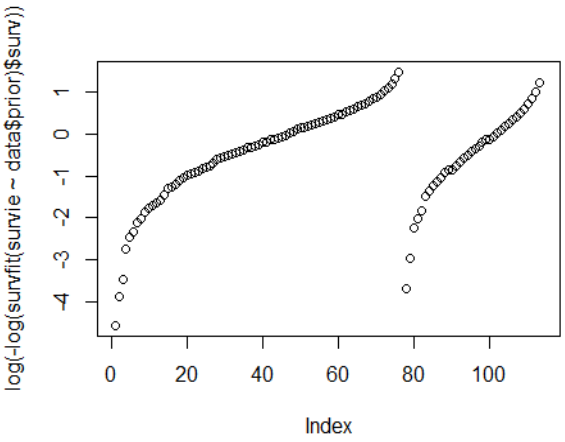
A présent, nous tenons en compte cette variable.

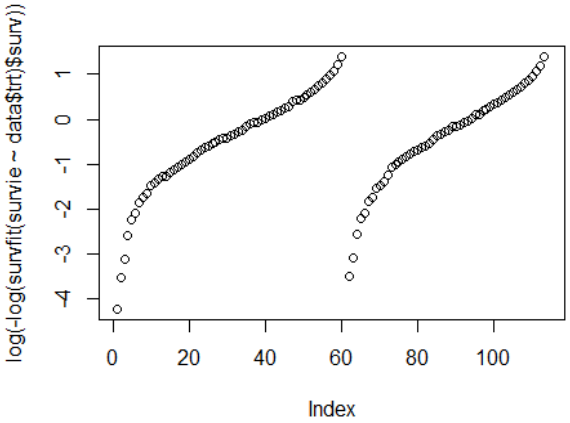
```
## Start:  AIC=963.07
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$diagtime + data$karno5 + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_celltype2 + data$trt_prior +
##     data$prior_celltype2
##
##                      Df    AIC
## - data$celltype2        1 961.08
## - data$prior_celltype2   1 961.19
## - data$diagtime         1 961.23
## - data$karno5           1 961.28
## - data$age              1 961.45
## - data$trt_celltype2    1 961.54
## <none>                  963.07
## - data$trt_prior        1 964.18
## - data$karno5_celltype2  1 964.70
## - data$karno5_trt       1 965.30
## - data$prior            1 968.76
## - data$karno5_prior     1 969.63
## - data$trt              1 971.54
##
## Step:  AIC=961.08
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5 +
##     data$karno5_trt + data$karno5_prior + data$karno5_celltype2 +
##     data$trt_celltype2 + data$trt_prior + data$prior_celltype2
##
##                      Df    AIC
## - data$prior_celltype2   1 959.21
## - data$diagtime         1 959.24
## - data$karno5           1 959.28
## - data$age              1 959.45
## - data$trt_celltype2    1 959.77
## <none>                  961.08
## - data$trt_prior        1 962.38
## - data$karno5_trt       1 963.31
## - data$prior            1 966.92
## - data$karno5_prior     1 968.09
## - data$trt              1 969.61
## - data$karno5_celltype2  1 969.64
##
## Step:  AIC=959.21
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5 +
##     data$karno5_trt + data$karno5_prior + data$karno5_celltype2 +
##     data$trt_celltype2 + data$trt_prior
```

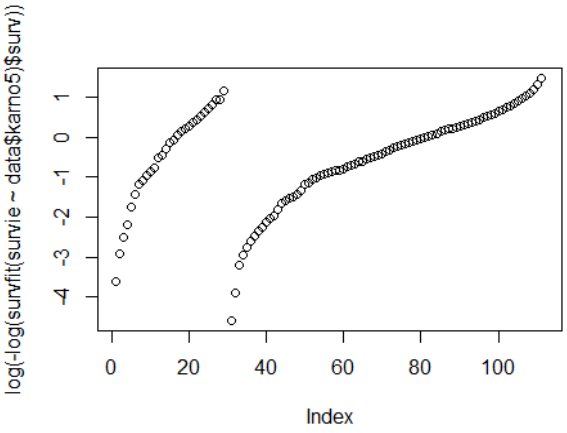
```
##
##                   Df    AIC
## - data$karno5          1 957.49
## - data$diagtime        1 957.49
## - data$age             1 957.59
## - data$trt_celltype2   1 957.87
## <none>                 959.21
## - data$trt_prior       1 960.48
## - data$karno5_trt      1 961.33
## - data$karno5_prior    1 966.11
## - data$trt             1 967.62
## - data$prior           1 967.81
## - data$karno5_celltype2  1 968.02
##
## Step:  AIC=957.49
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_celltype2 +
##     data$trt_prior
##
##                   Df    AIC
## - data$age             1 955.71
## - data$diagtime        1 955.90
## - data$trt_celltype2   1 955.99
## <none>                 957.49
## - data$trt_prior       1 958.72
## - data$karno5_trt      1 964.84
## - data$karno5_prior    1 967.60
## - data$prior           1 968.03
## - data$trt             1 969.04
## - data$karno5_celltype2  1 969.70
##
## Step:  AIC=955.71
## survie ~ data$prior + data$trt + data$diagtime + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_celltype2 +
##     data$trt_prior
##
##                   Df    AIC
## - data$diagtime        1 954.14
## - data$trt_celltype2   1 954.16
## <none>                 955.71
## - data$trt_prior       1 956.72
## - data$karno5_trt      1 963.00
## - data$karno5_prior    1 965.73
## - data$prior           1 966.03
## - data$trt             1 967.07
## - data$karno5_celltype2  1 967.77
```

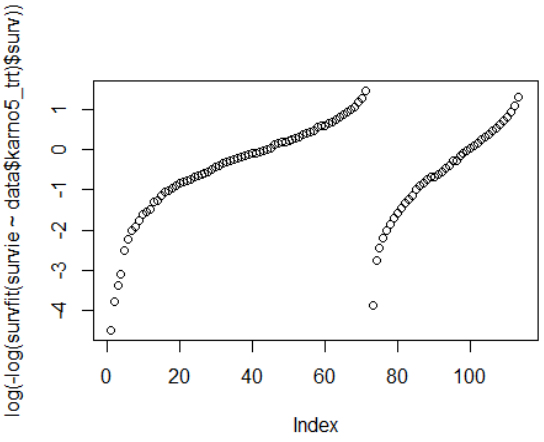
```
##
## Step:  AIC=954.14
## survie ~ data$prior + data$trt + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_celltype2 + data$trt_prior
##
##                    Df    AIC
## - data$trt_celltype2     1 952.50
## <none>                   954.14
## - data$trt_prior         1 955.93
## - data$karno5_trt        1 961.00
## - data$karno5_prior      1 963.76
## - data$prior             1 964.21
## - data$trt               1 965.08
## - data$karno5_celltype2  1 966.45
##
## Step:  AIC=952.5
## survie ~ data$prior + data$trt + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_prior
##
##                    Df    AIC
## <none>                   952.50
## - data$trt_prior         1 954.39
## - data$karno5_trt        1 959.68
## - data$karno5_prior      1 962.11
## - data$prior             1 962.70
## - data$trt               1 963.37
## - data$karno5_celltype2  1 972.10

## Call:
## coxph(formula = survie ~ data$prior + data$trt + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_prior)
##
##                      coef exp(coef) se(coef)     z       p
## data$prior          1.518     4.562    0.393  3.86 0.00011
## data$trt            1.246     3.478    0.327  3.81 0.00014
## data$karno5_trt      -1.036     0.355    0.329 -3.15 0.00164
## data$karno5_prior    -1.448     0.235    0.401 -3.61 0.00030
## data$karno5_celltype2 -1.040     0.354    0.228 -4.56 5.1e-06
## data$trt_prior       -0.807     0.446    0.412 -1.96 0.04985
##
## Likelihood ratio test=70.4  on 6 df, p=3.38e-13
## n= 137, number of events= 128
```

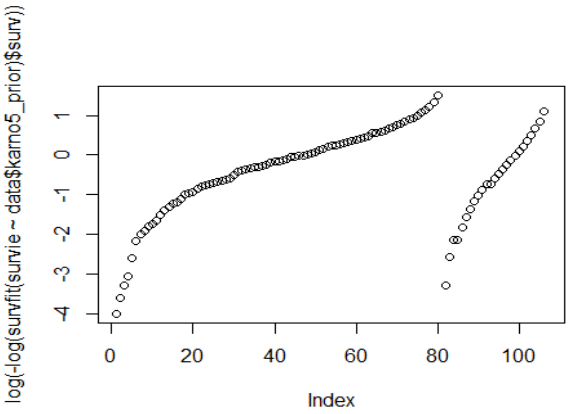
En comptant la variable karno5, la sélection des variables avec stepAIC change. On obtient alors les variables suivantes : prior, trt, karno5, karno5_trt, karno5_prior, karno5_celltype2 et trt_prior. En effet, la disparition de la variable celltype2, qui est importante, semble étrange.

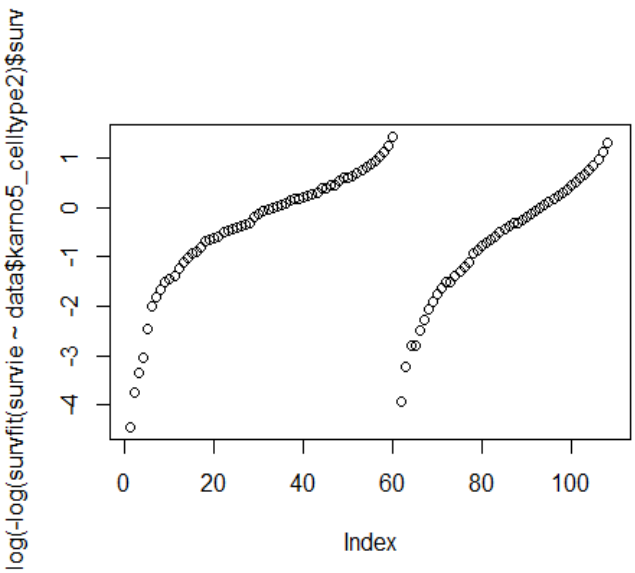


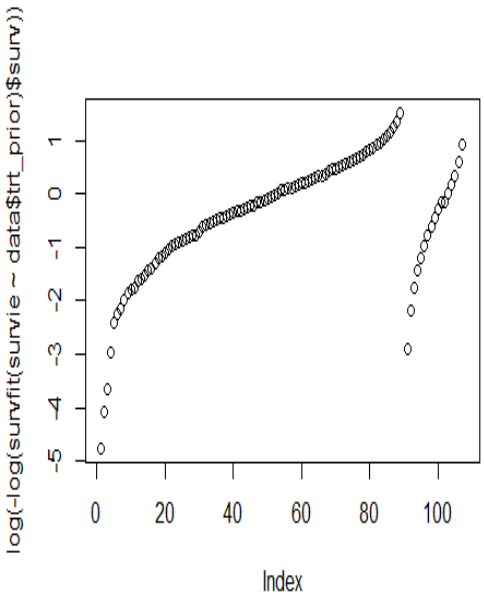












On remarque que les courbes pour karno5 et karno5_prior ne sont pas parallèles donc le modèle de cox n'est pas adapté pour elles.

Conclusion

Les temps de survie mesurés à partir d'une origine appropriée ont deux caractéristiques. La première est qu'ils sont positifs et tels qu'une hypothèse de normalité n'est généralement pas raisonnable en raison d'une asymétrie prononcée. La seconde est structurelle et tient au fait que pour certains individus l'évènement étudié ne se produit pas pendant la période d'observation et en conséquence certaines données sont censurées. Cette censure à droite est la plus courante mais n'est pas la seule censure que l'on peut rencontrer avec des données de survie.

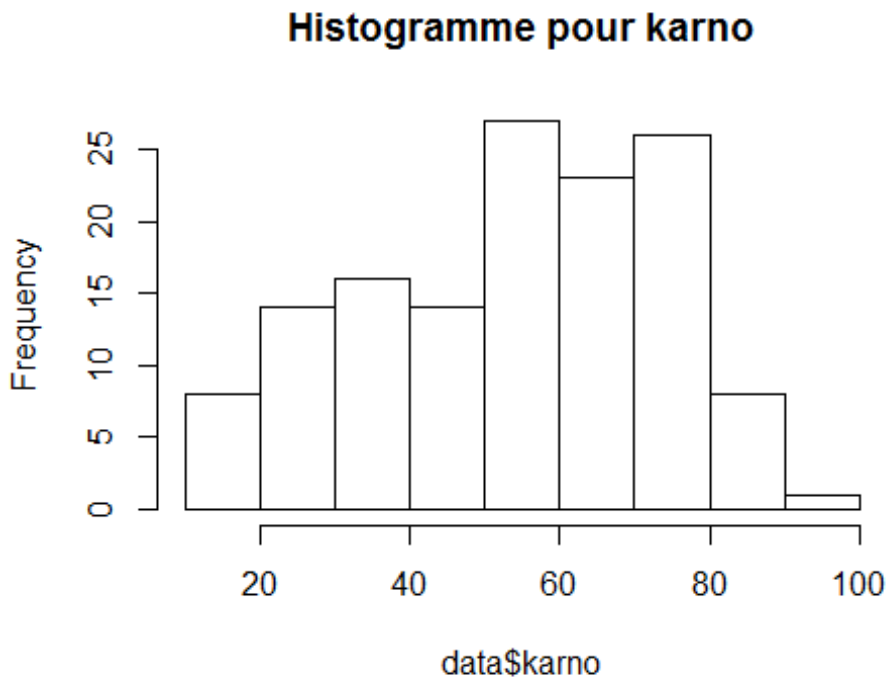
Au cours de ce projet, nous avons appris que les méthodes de régression connues diffèrent selon le type de données que nous avons. En effet, pour les données de survie la création d'un modèle prédictif les estimateurs sont différents. L'estimateur de la fonction de survie le plus utilisé lorsqu'aucune hypothèse ne veut être faite sur la distribution des temps de survie est l'estimateur de Kaplan-Meier. Si Kaplan-Meier est utile pour estimer une fonction de survie, on peut être intéressé par l'estimation d'autres fonctions qui caractérisent la distribution des temps d'évènements. Nous traiterons donc de l'estimation de la fonction de risque cumulé, avec l'estimateur de Nelson-Aalen. Le modèle de Cox est l'approche la plus populaire dans l'analyse des modèles de survie. Ce modèle ne requiert pas la formulation d'une hypothèse de distribution des temps de survie mais l'estimation des paramètres du modèle et tout particulièrement des coefficients des variables explicatives passe par la maximisation d'une fonction de vraisemblance dite partielle.

Références

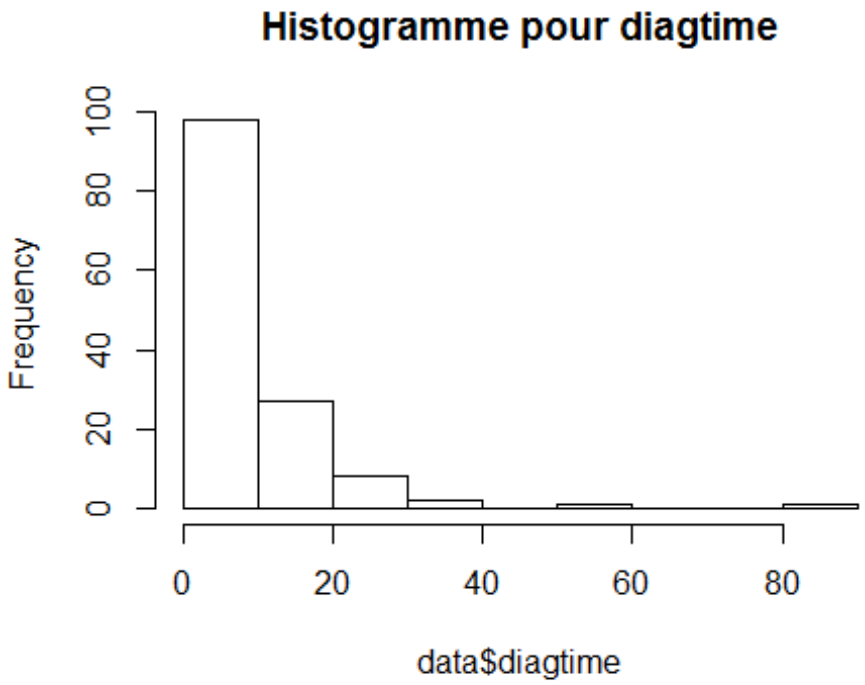
- https://www.fun-mooc.fr/c4x/UPSUD/42001S02/asset/MOOC_Cours_18_Survie1_V2.pdf
- http://iml.univ-mrs.fr/~reboul/duree62011.pdf
- http://iml.univ-mrs.fr/~reboul/R-survie.pdf
- http://zoonek2.free.fr/UNIX/48_R_2004/19.html
- http://emmanuel.duguet.free.fr/poly_duree_2011_v4.pdf

Annexes

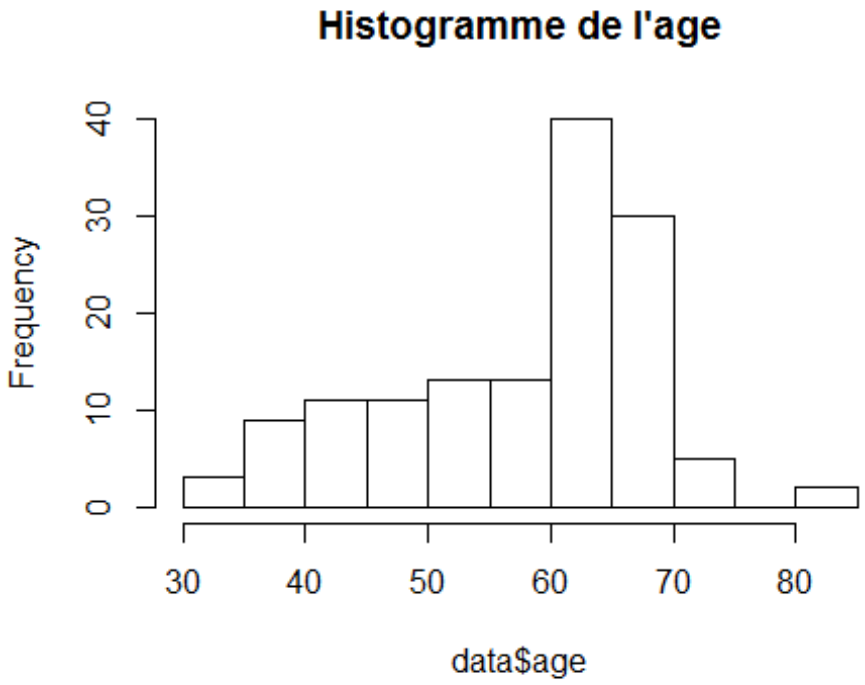
```
#install.packages("survival")
library(survival)

## Warning: package 'survival' was built under R version 3.2.5

data <- veteran
data$prior <- as.numeric(data$prior)
data$trt <- as.numeric(data$trt)
attach(data)

hist(data$karno, main="Histogramme pour karno")
```



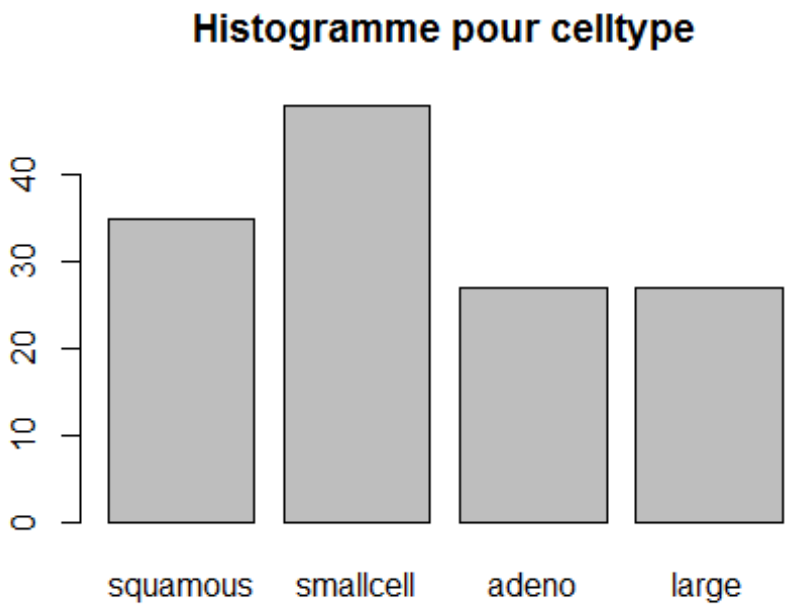

```
hist(data$diagtime, main="Histogramme pour diagtime")
```



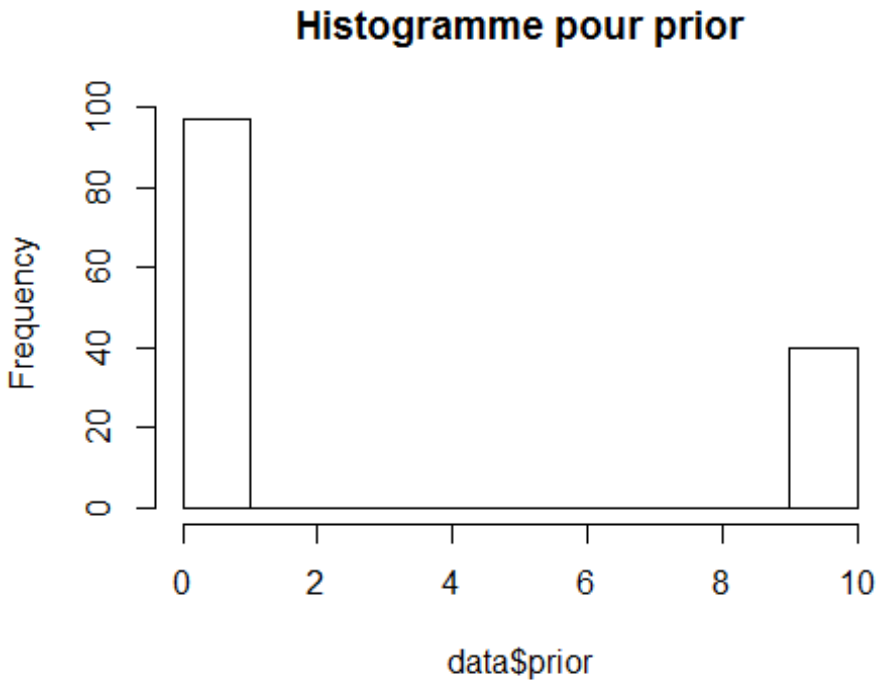

```
hist(data$age, main="Histogramme de l'age")
```



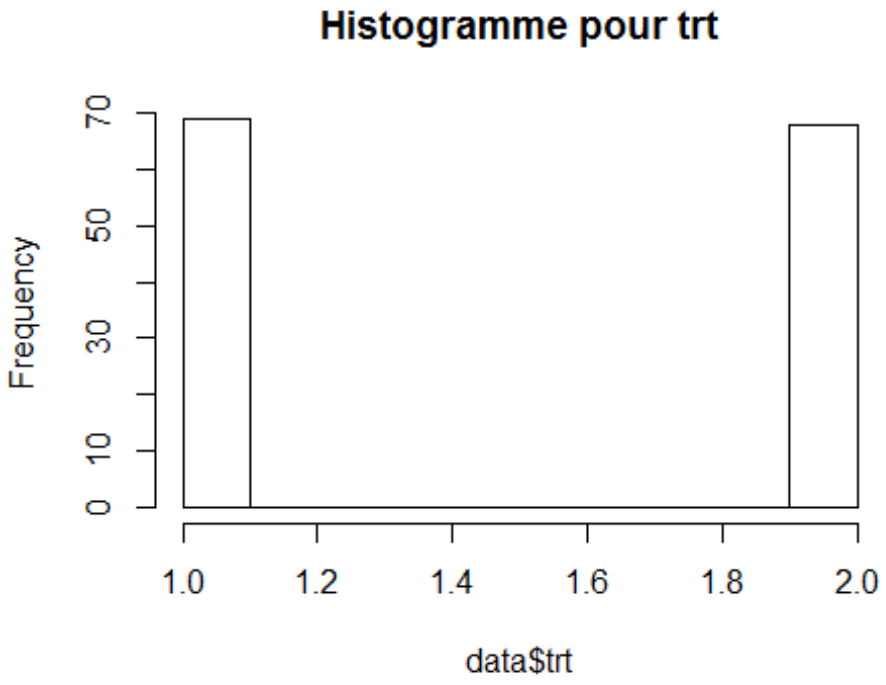

```
plot(data$celltype, main="Histogramme pour celltype")
```



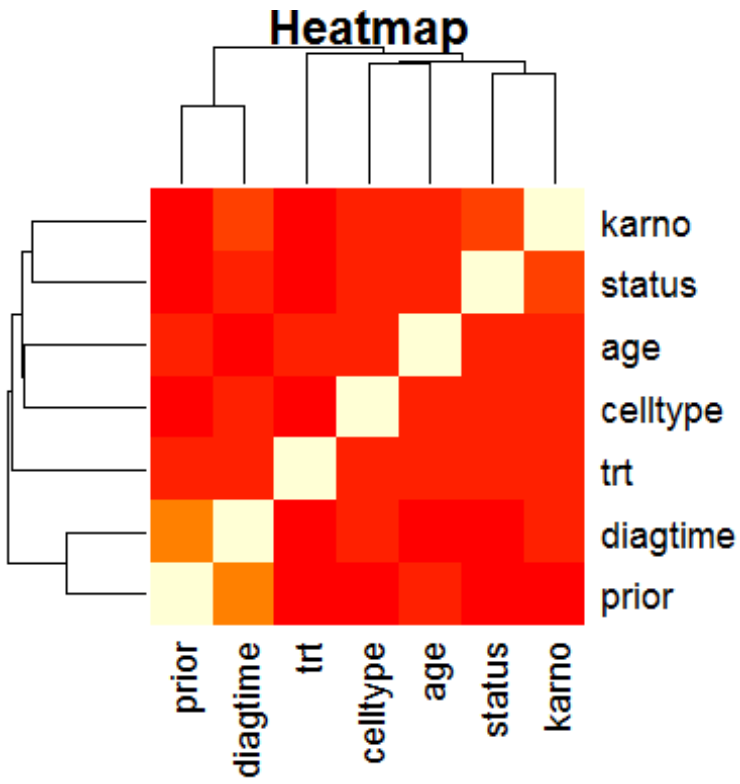

```
data$celltype <- as.numeric(data$celltype)
hist(data$prior, main="Histogramme pour prior")
```



```
hist(data$trt, main="Histogramme pour trt")
```



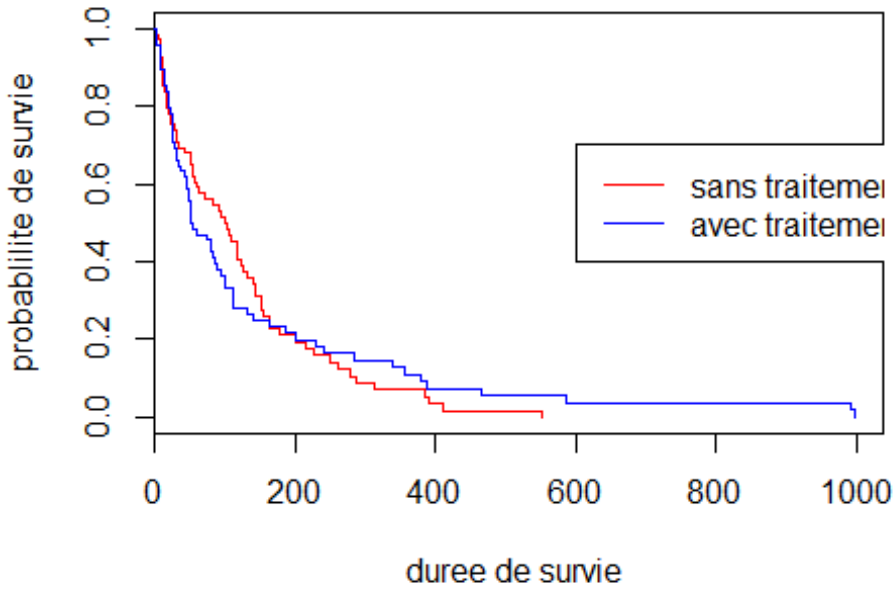

```
data_sanstime <- data
data_sanstime$time <- NULL
are.factor  <-sapply(data_sanstime, is.factor)
heatmap(abs(cor(data_sanstime[, !are.factor])),main="Heatmap")
```



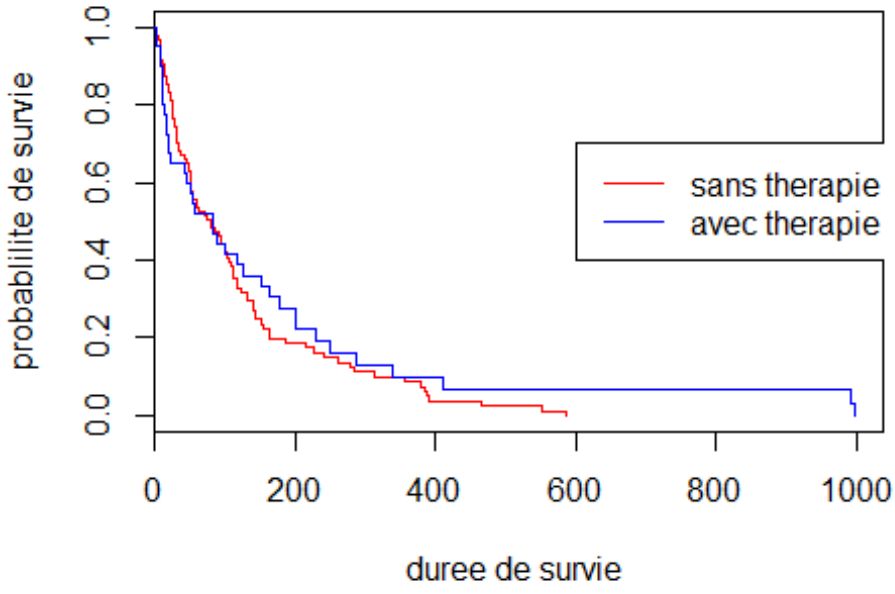

```
data$status <- as.numeric(data$status)

survie <- Surv(data$time,data$status)
survie
```

```
##  [1]  72 411 228 126 118  10  82 110 314 100+ 42   8 144  25+
## [15]  11  30 384   4  54  13 123+ 97+ 153  59 117  16 151  22
## [29]  56  21  18 139  20  31  52 287  18  51 122  27  54   7
## [43]  63 392  10   8  92  35 117 132  12 162   3  95 177 162
## [57] 216 553 278  12 260 200 156 182+ 143 105 103 250 100 999
## [71] 112  87+ 231+ 242 991 111   1 587 389  33  25 357 467 201
## [85]   1  30  44 283  15  25 103+ 21  13  87   2  20   7  24
## [99]  99   8  99  61  25  95  80  51  29  24  18  83+ 31  51
## [113]  90  52  73   8  36  48   7 140 186  84  19  45  80  52
## [127] 164  19  53  15  43 340 133 111 231 378  49
```

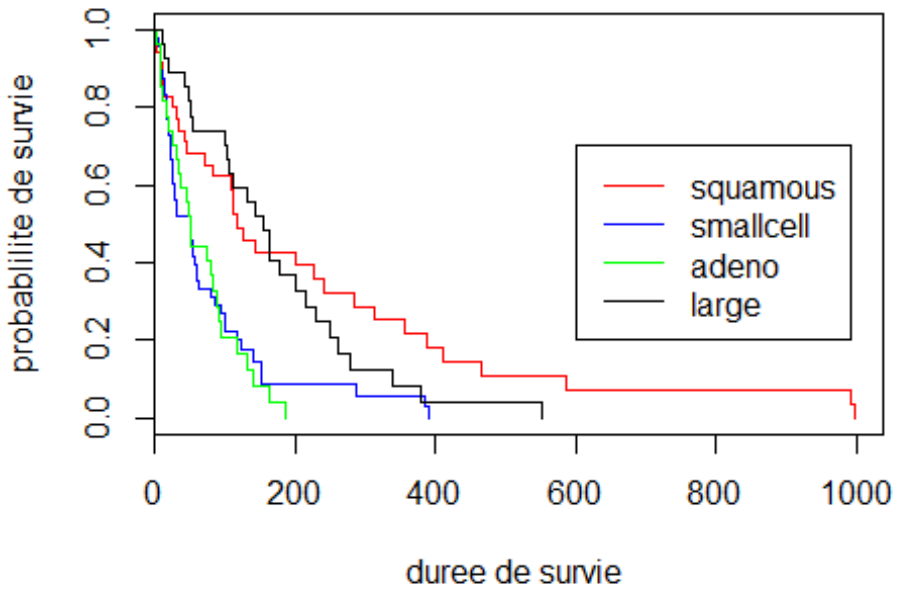
```
plot(survfit(survie~data$trt), xlab="duree de survie ", ylab = "probablilite de survie", col=c
('red','blue'))
legend(600,0.7,c("sans traitement","avec traitement"),lty=c(1,1),col=c("red","blue"))
```



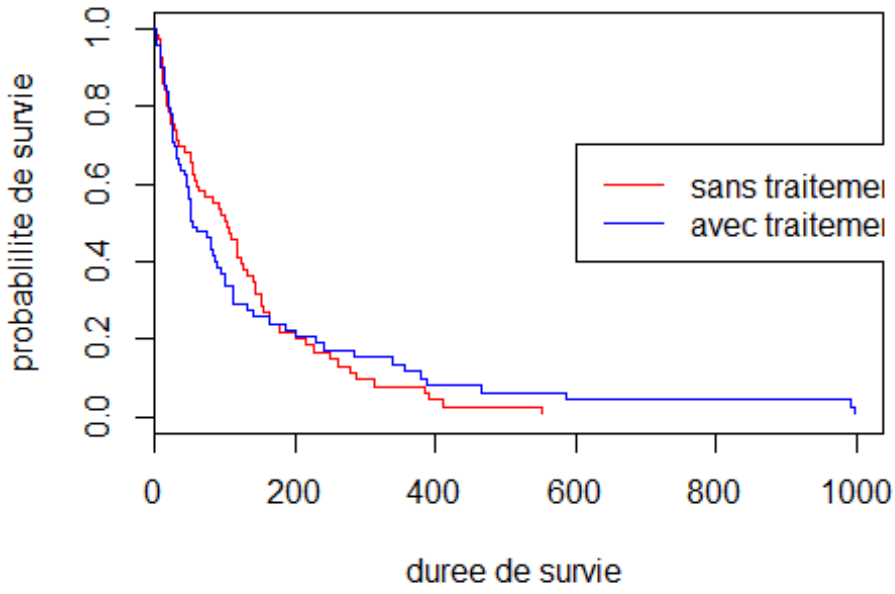

```
plot(survfit(survie~data$prior), xlab="duree de survie ", ylab = "probablilite de survie", col=c('red','blue'))
legend(600,0.7,c("sans therapie","avec therapie"),lty=c(1,1),col=c("red","blue"))
```



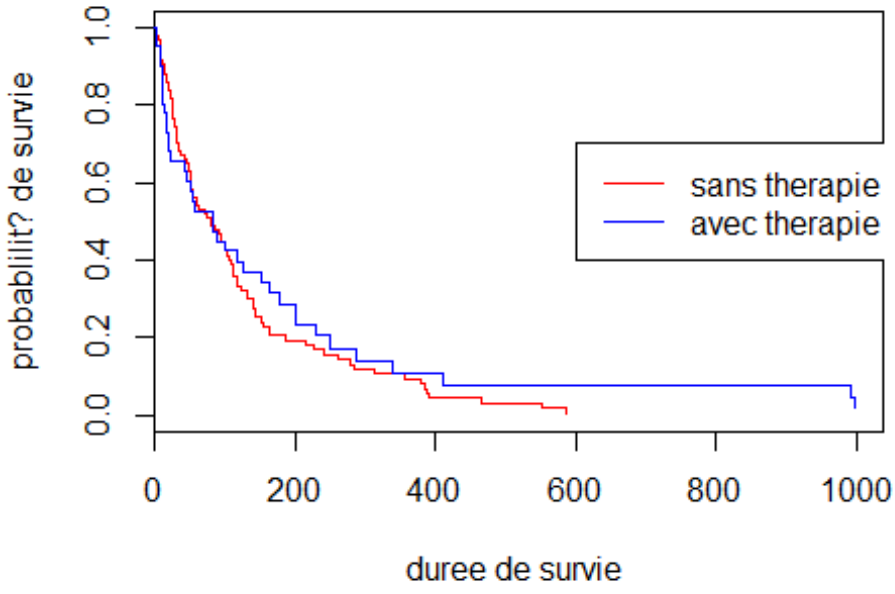
```
plot(survfit(survie~data$celltype), xlab="duree de survie ", ylab = "probablilite de survie",
col=c('red','blue','green','black'))
legend(600,0.7,c("squamous","smallcell","adeno","large"),lty=c(1,1),col=c("red","blue","gre
en","black"))
```



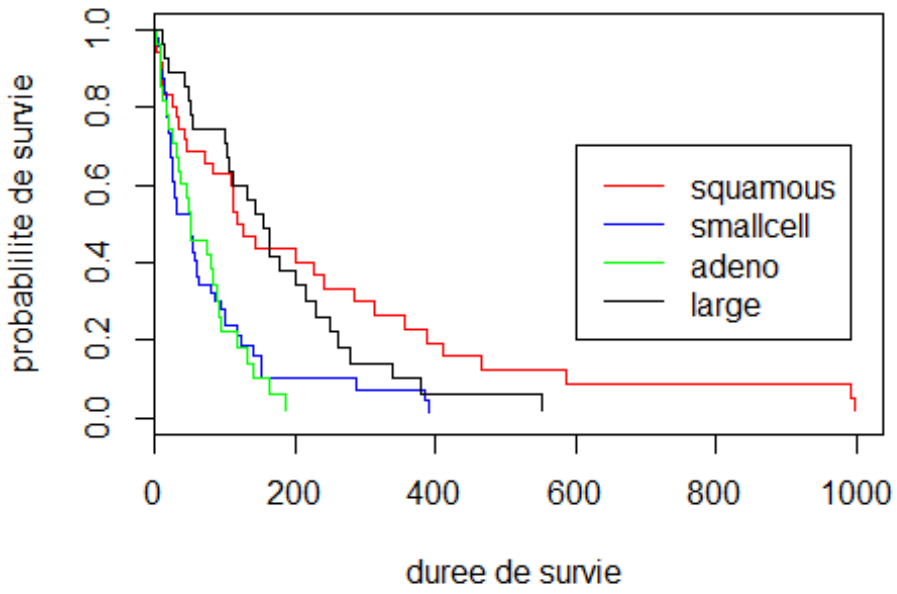

```
plot(survfit(survie~data$trt,type="fleming-harrington"), xlab="duree de survie ", ylab = "p
robablilite de survie", col=c('red','blue'))
legend(600,0.7,c("sans traitement","avec traitement"),lty=c(1,1),col=c("red","blue"))
```



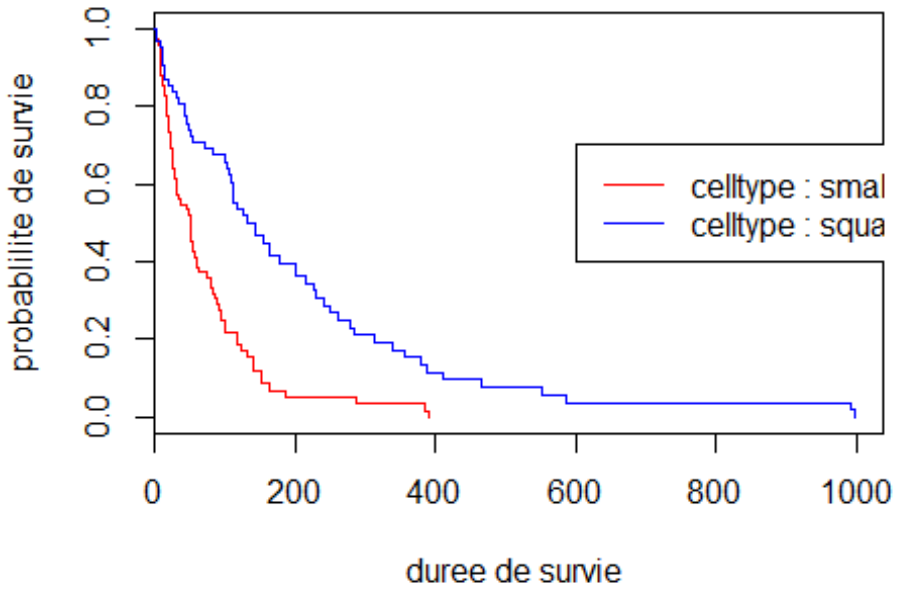

```
plot(survfit(survie~data$prior,type="fleming-harrington"), xlab="duree de survie ", ylab =
"probablilit? de survie", col=c('red','blue'))
legend(600,0.7,c("sans therapie","avec therapie"),lty=c(1,1),col=c("red","blue"))
```



```
plot(survfit(survie~data$celltype,type="fleming-harrington"), xlab="duree de survie ", yla
b = "probablilite de survie", col=c('red','blue','green','black'))
legend(600,0.7,c("squamous","smallcell","adeno","large"),lty=c(1,1),col=c("red","blue","gre
en","black"))
```




```
data$celltype2 <- data$celltype
data$celltype2[data$celltype2==2]<-0
data$celltype2[data$celltype2==3]<-0
data$celltype2[data$celltype2==4]<-1
plot(survfit(survie~data$celltype2), xlab="duree de survie ", ylab = "probablilite de survie
", col=c('red','blue'))
legend(600,0.7,c("celltype : smallcell or adeno","celltype : squamous or large"),lty=c(1,1),c
ol=c("red","blue"))
```




```
survdiff(survie~trt,data=data)
```

```
## Call:
## survdiff(formula = survie ~ trt, data = data)
##
##       N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69       64     64.5   0.00388   0.00823
## trt=2 68       64     63.5   0.00394   0.00823
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.928
```

```
survdiff(survie~prior,data=data)
```

```
## Call:
## survdiff(formula = survie ~ prior, data = data)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0  97       91     87.4     0.150     0.501
## prior=10 40       37     40.6     0.323     0.501
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.479
```

```
survdiff(survie~celltype2,data=data)
```

```
## Call:
## survdiff(formula = survie ~ celltype2, data = data)
```

```
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype2=0 75      71     45.8     13.87     24.5
## celltype2=1 62      57     82.2      7.73     24.5
##
##  Chisq= 24.5  on 1 degrees of freedom, p= 7.34e-07
```

```
survdiff(survie~trt,data=data,rho=1)
```

```
## Call:
## survdiff(formula = survie ~ trt, data = data, rho = 1)
##
##       N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=1 69     32.2     35.4     0.279     0.871
## trt=2 68     35.2     32.1     0.308     0.871
##
##  Chisq= 0.9  on 1 degrees of freedom, p= 0.351
```

```
survdiff(survie~prior,data=data,rho=1)
```

```
## Call:
## survdiff(formula = survie ~ prior, data = data, rho = 1)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## prior=0  97     47.6     48.2   0.00694    0.0366
## prior=10 40     19.8     19.3   0.01736    0.0366
##
##  Chisq= 0  on 1 degrees of freedom, p= 0.848
```

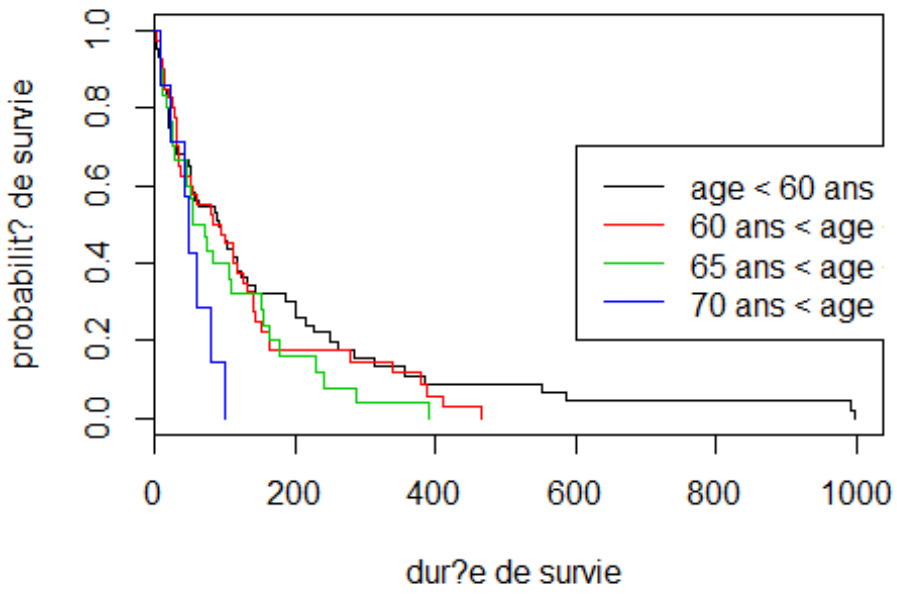
```
survdiff(survie~celltype2,data=data,rho=1)
```

```
## Call:
## survdiff(formula = survie ~ celltype2, data = data, rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## celltype2=0 75     44.5     29.8     7.23     19.5
## celltype2=1 62     23.0     37.6     5.73     19.5
##
##  Chisq= 19.5  on 1 degrees of freedom, p= 1e-05
```

```
data$age3 <- data$age
data$age3[data$age3<=60] <- 1
data$age3[60<data$age3&data$age3<=65] <- 2
data$age3[65<data$age3&data$age3<=70] <- 3
data$age3[70<data$age3] <- 4
```

```
surv.fit3 <- survfit(survie~data$age3)
plot(surv.fit3,col="blue",xlab="dur?e de survie ",ylab = "probabilit? de survie")
lines(surv.fit3,col=c(1,2,3,4))
```

```
legend(600,0.7,c("age < 60 ans","60 ans < age < 65 ans","65 ans < age < 70 ans", "70 ans < a
ge"),lty=c(1,1),col=c(1,2,3,4))
```

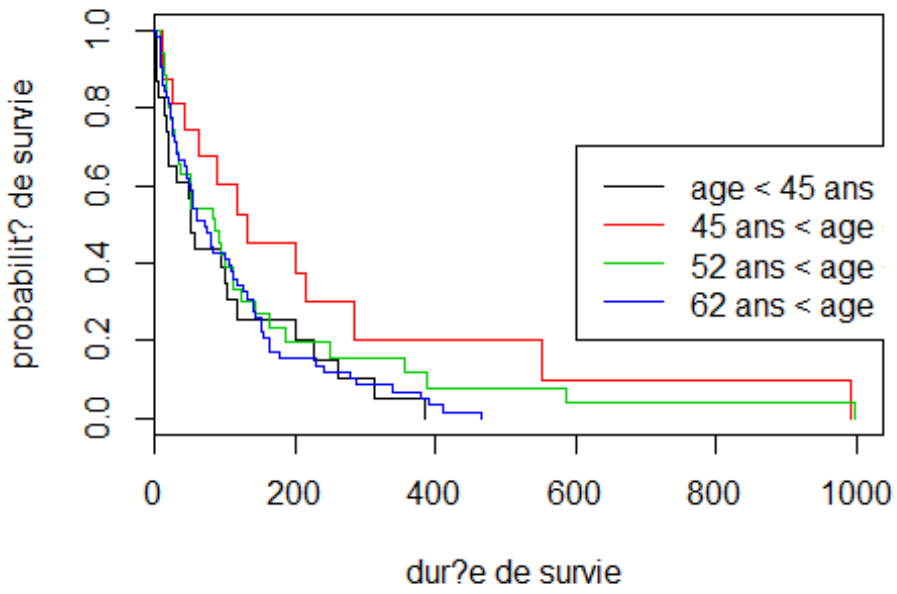



```
survdiff(survie~data$age3)
```

```
## Call:
## survdiff(formula = survie ~ data$age3)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age3=1 60    54    61.95   1.0207   2.08945
## data$age3=2 40    39    38.65   0.0031   0.00456
## data$age3=3 30    28    24.03   0.6561   0.82897
## data$age3=4  7     7     3.36   3.9272   4.15703
##
##  Chisq= 5.9  on 3 degrees of freedom, p= 0.117
```

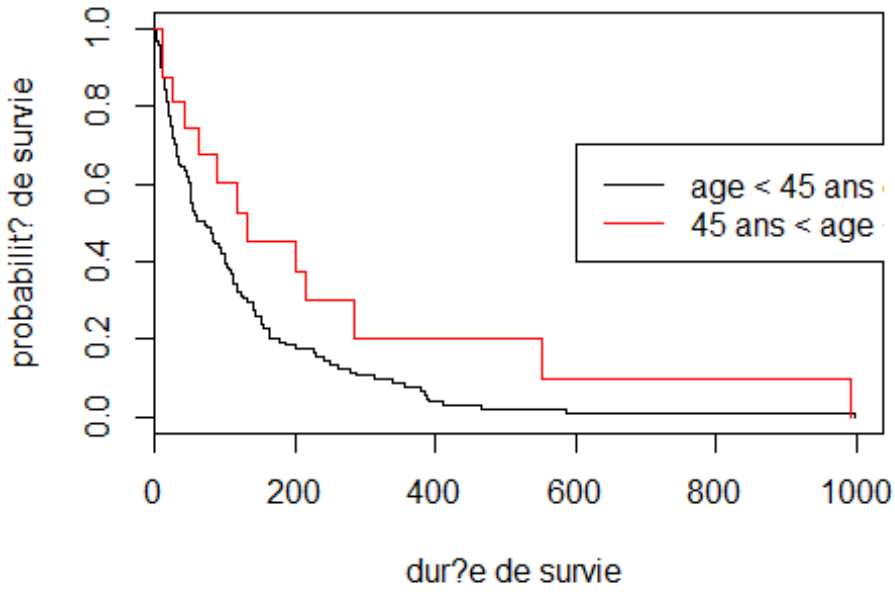
```
data$age3 <- data$age
data$age3[data$age3<=45] <- 1
data$age3[45<data$age3&data$age3<=52] <- 2
data$age3[52<data$age3&data$age3<=62] <- 3
data$age3[62<data$age3] <- 4
surv.fit3 <- survfit(survie~data$age3)
plot(surv.fit3,col="blue",xlab="dur?e de survie ",ylab = "probabilit? de survie")
lines(surv.fit3,col=c(1,2,3,4))
```

```
legend(600,0.7,c("age < 45 ans","45 ans < age < 52 ans","52 ans < age < 62 ans", "62 ans < a
ge"),lty=c(1,1),col=c(1,2,3,4))
```




```
data$age4 <- data$age
data$age4[data$age4<=45] <- 0
data$age4[52<data$age4]<-0
data$age4[45<data$age4&data$age4<=52] <- 1
```

```
plot(survfit(survie~data$age4),xlab="dur?e de survie ",ylab = "probabilit? de survie")
lines(survfit(survie~data$age4),col=c(1,2))
legend(600,0.7,c("age < 45 ans et age > 52 ans","45 ans < age < 52 ans"),lty=c(1,1),col=c(1,
2))
```




```
survdiff(survie~data$age4)
```

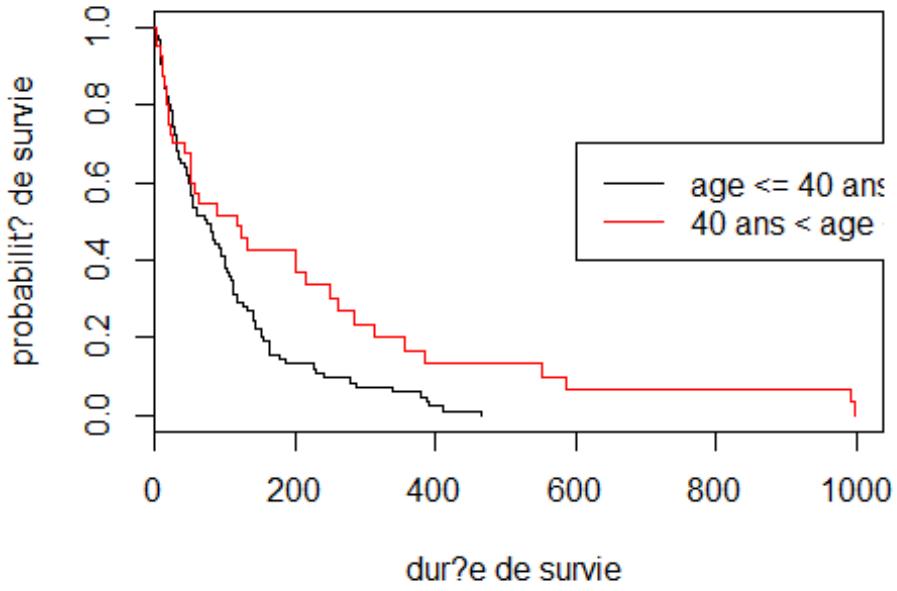
```
## Call:
## survdiff(formula = survie ~ data$age4)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age4=0 121      115    107.7     0.496      3.25
## data$age4=1  16       13     20.3     2.632      3.25
##
##  Chisq= 3.2  on 1 degrees of freedom, p= 0.0716
```

```
survdiff(survie~data$age4,rho=1)
```

```
## Call:
## survdiff(formula = survie ~ data$age4, rho = 1)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## data$age4=0 121    61.92    58.22     0.236      2.67
## data$age4=1  16     5.53     9.23     1.486      2.67
##
##  Chisq= 2.7  on 1 degrees of freedom, p= 0.102
```

```
data$age5 <- data$age
data$age5[data$age5<=40] <- 0
data$age5[40<data$age5&data$age5<=58] <- 1
data$age5[58<data$age5] <- 0
```

```
plot(survfit(survie~data$age5),xlab="dur?e de survie ",ylab = "probabilit? de survie")
lines(survfit(survie~data$age5),col=c(1,2))
legend(600,0.7,c("age <= 40 ans et >58 ans","40 ans < age < 58 ans"),lty=c(1,1),col=c(1,2))
```



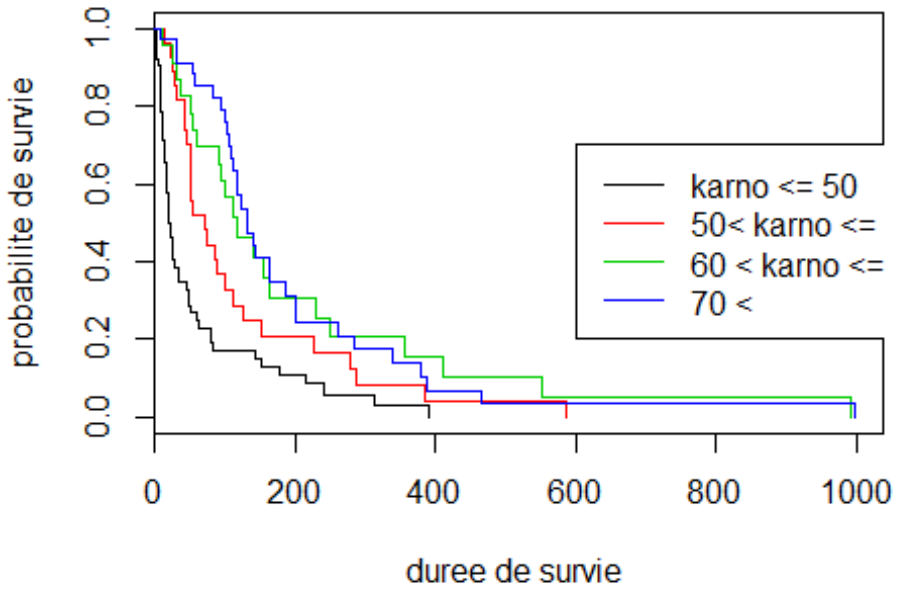

```
survdiff(survie~age5,data=data,rho=0)
```

```
## Call:
## survdiff(formula = survie ~ age5, data = data, rho = 0)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## age5=0 97       93     79.7     2.21     6.51
## age5=1 40       35     48.3     3.64     6.51
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.0107
```

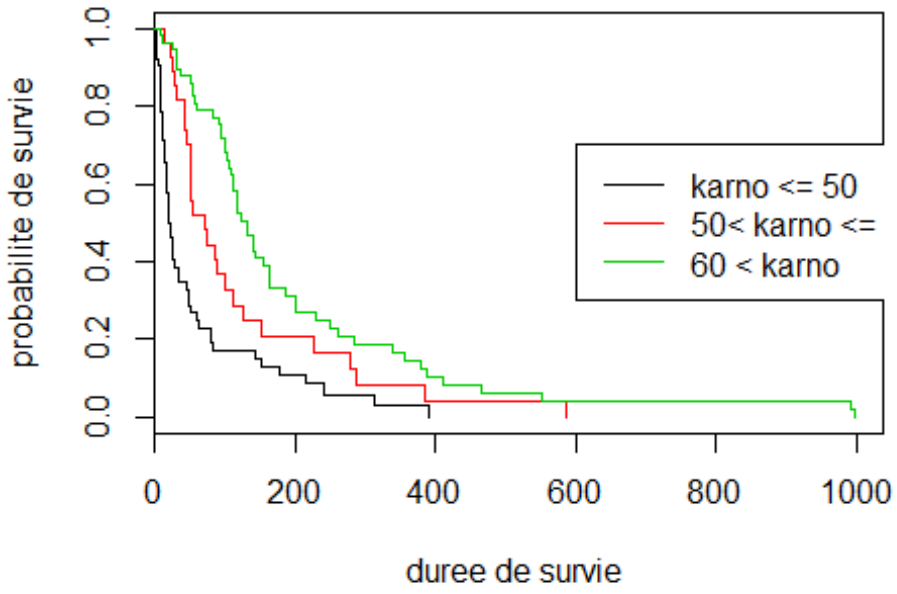
```
survdiff(survie~age5,data=data,rho=1)
```

```
## Call:
## survdiff(formula = survie ~ age5, data = data, rho = 1)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## age5=0 97     50.5     46.2    0.392     1.91
## age5=1 40     17.0     21.2    0.853     1.91
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.167
```

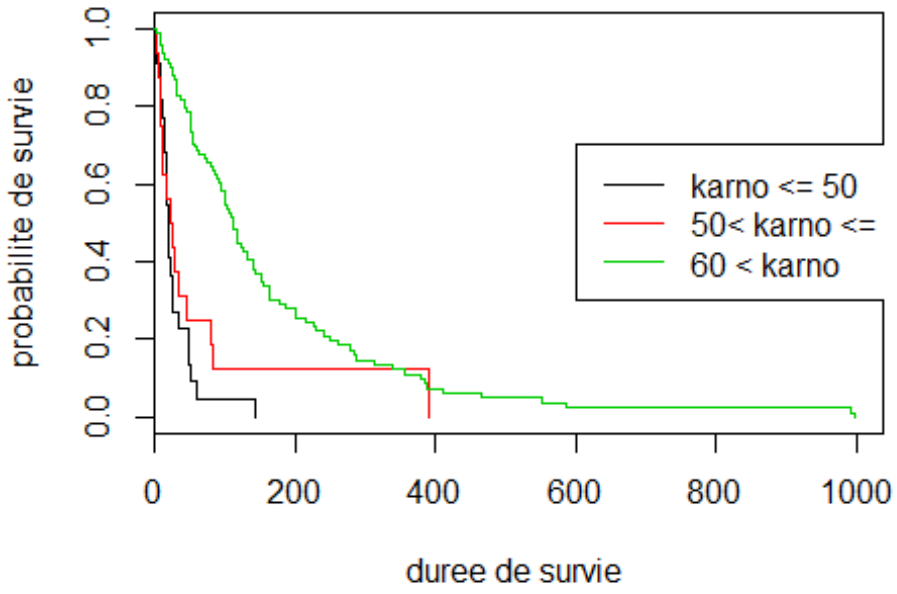
```
data$karno2 <- data$karno
data$karno2[data$karno2<=50] <- 0
data$karno2[50<data$karno2&data$karno2<=60] <- 1
data$karno2[60<data$karno2&data$karno2<=70] <- 2
data$karno2[70<data$karno2] <- 3

plot(survfit(survie~data$karno2),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$karno2),col=c(1,2,3,4))
legend(600,0.7,c("karno <= 50","50< karno <= 60 ","60 < karno <= 70 ", "70 < "),lty=c(1,1),c
ol=c(1,2,3,4))
```



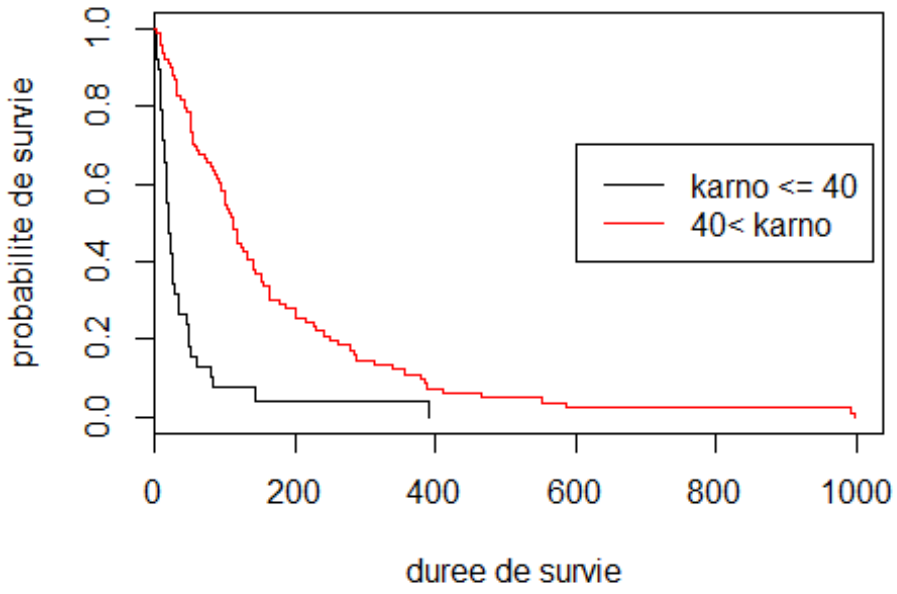

```
data$karno3 <- data$karno
data$karno3[data$karno3<=50]<-0
data$karno3[50<data$karno3&data$karno3<=60] <- 1
data$karno3[60<data$karno3] <- 2

plot(survfit(survie~data$karno3),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$karno3),col=c(1,2,3,4))
legend(600,0.7,c("karno <= 50","50< karno <= 60 ","60 < karno "),lty=c(1,1),col=c(1,2,3))
```




```
data$karno4 <- data$karno
data$karno4[data$karno4<=30]<-0
data$karno4[30<data$karno4&data$karno4<=40] <- 1
data$karno4[40<data$karno4] <- 2

plot(survfit(survie~data$karno4),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$karno4),col=c(1,2,3,4))
legend(600,0.7,c("karno <= 50","50< karno <= 60 ","60 < karno "),lty=c(1,1),col=c(1,2,3))
```




```
data$karno5 <- data$karno
data$karno5[data$karno5<=40]<-0
data$karno5[40<data$karno5] <- 1

plot(survfit(survie~data$karno5),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$karno5),col=c(1,2,3,4))
legend(600,0.7,c("karno <= 40","40< karno "),lty=c(1,1),col=c(1,2))
```




```
survdiff(survie~data$karno5)
```

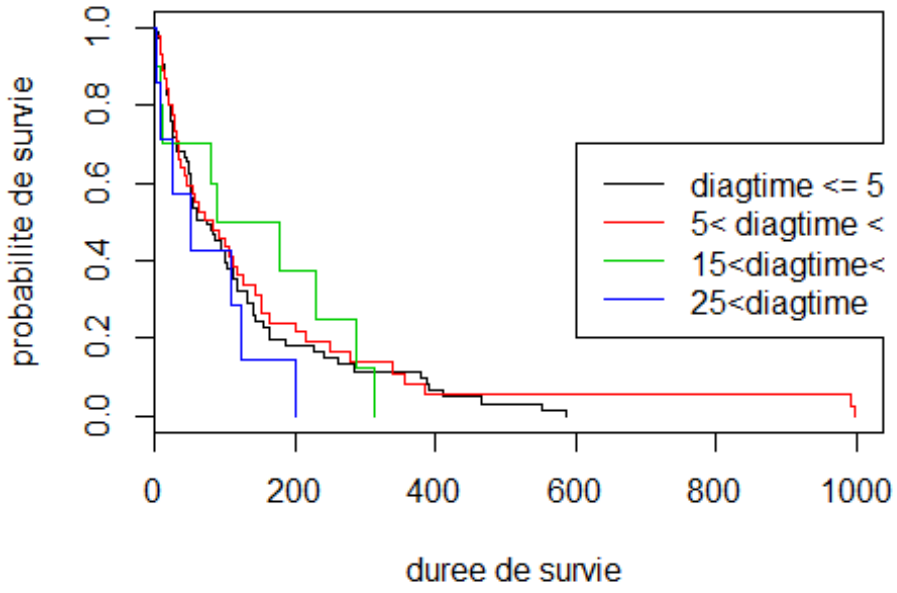
```
## Call:
## survdiff(formula = survie ~ data$karno5)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$karno5=0 38       37     14.1     36.94     44.5
## data$karno5=1 99       91    113.9      4.59     44.5
##
##  Chisq= 44.5  on 1 degrees of freedom, p= 2.55e-11
```

```
survdiff(survie~data$karno5,rho=1)
```

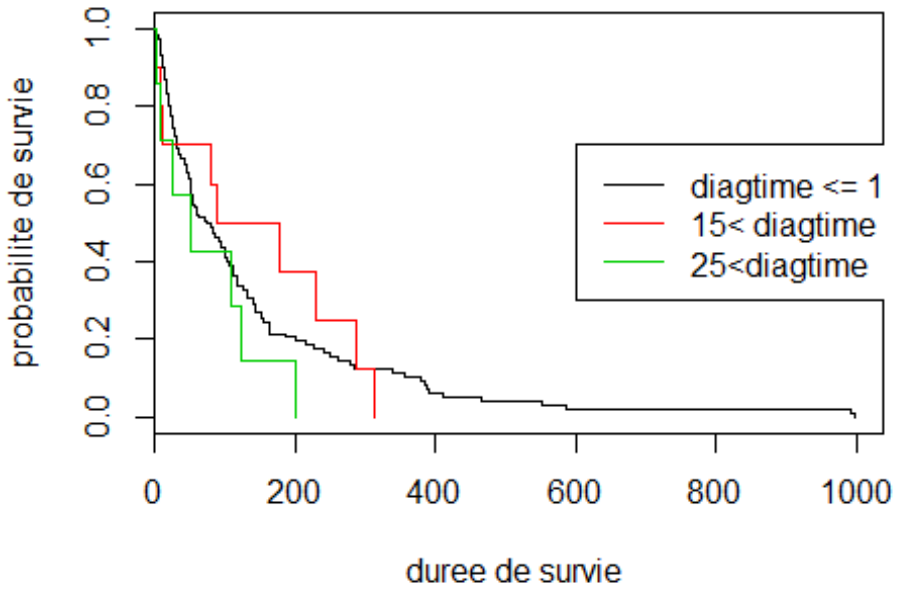
```
## Call:
## survdiff(formula = survie ~ data$karno5, rho = 1)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## data$karno5=0 38     28.3     9.85     34.70     55.2
## data$karno5=1 99     39.1    57.59      5.94     55.2
##
##  Chisq= 55.2  on 1 degrees of freedom, p= 1.1e-13
```

```
data$diag2 <- data$diagtime
data$diag2[data$diag2<=5] <- 0
data$diag2[5<data$diag2&data$diag2<=15] <- 1
data$diag2[15<data$diag2&data$diag2<=25] <- 2
```

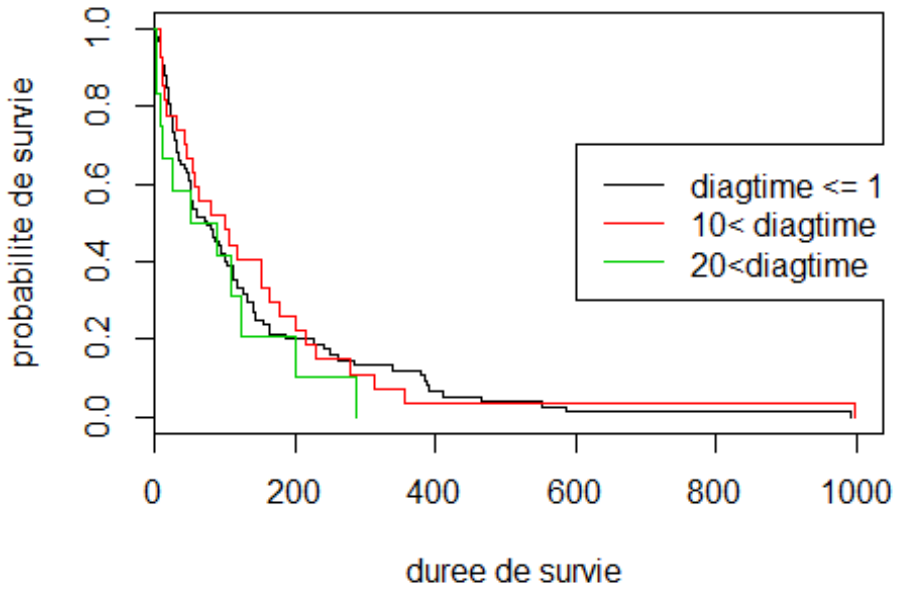
```
data$diag2[25<data$diag2] <- 3

plot(survfit(survie~data$diag2),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$diag2),col=c(1,2,3,4))
legend(600,0.7,c("diagtime <= 5","5< diagtime <= 15 ","15<diagtime<=25","25<diagtime"),l
ty=c(1,1),col=c(1,2,3,4))
```




```
data$diag3 <- data$diagtime
data$diag3[data$diag3<=15] <- 0
data$diag3[15<data$diag3&data$diag3<=25] <- 1
data$diag3[25<data$diag3] <- 2

plot(survfit(survie~data$diag3),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$diag3),col=c(1,2,3,4))
legend(600,0.7,c("diagtime <= 15","15< diagtime <= 25 ","25<diagtime"),lty=c(1,1),col=c(1
,2,3))
```




```
data$diag4 <- data$diagtime
data$diag4[data$diag4<=10] <- 0
data$diag4[10<data$diag4&data$diag4<=20] <- 1
data$diag4[20<data$diag4] <- 2

plot(survfit(survie~data$diag4),xlab="duree de survie ",ylab = "probabilite de survie")
lines(survfit(survie~data$diag4),col=c(1,2,3,4))
legend(600,0.7,c("diagtime <= 10","10< diagtime <= 20 ","20<diagtime"),lty=c(1,1),col=c(1
,2,3))
```

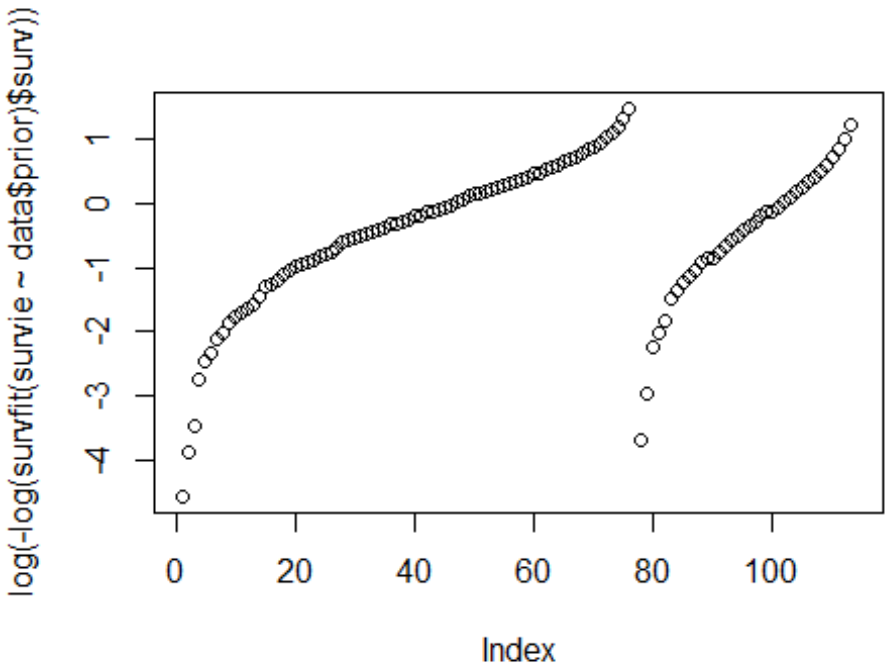



```
data$prior[data$prior==10] <- 1
data$trt[data$trt==1] <- 0
data$trt[data$trt==2] <- 1

data$trt_prior <- data$trt*data$prior
data$trt_celltype2 <- data$trt*data$celltype2
data$prior_celltype2 <- data$prior*data$celltype2
data$karno5_celltype2 <- data$karno5*data$celltype2
data$karno5_trt <- data$karno5*data$trt
data$karno5_prior <- data$karno5*data$prior

library("stats")
library("MASS")

cox <- coxph(survie~data$prior+data$trt+data$celltype2+data$age+data$diagtime+data$
karno+data$trt_celltype2+data$trt_prior+data$prior_celltype2)

stepAIC(cox,scope=list(lower=.~1,upper=.~.),direction="backward")

## Start:  AIC=964.68
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$diagtime + data$karno + data$trt_celltype2 + data$trt_prior +
##     data$prior_celltype2
##
##                  Df    AIC
```

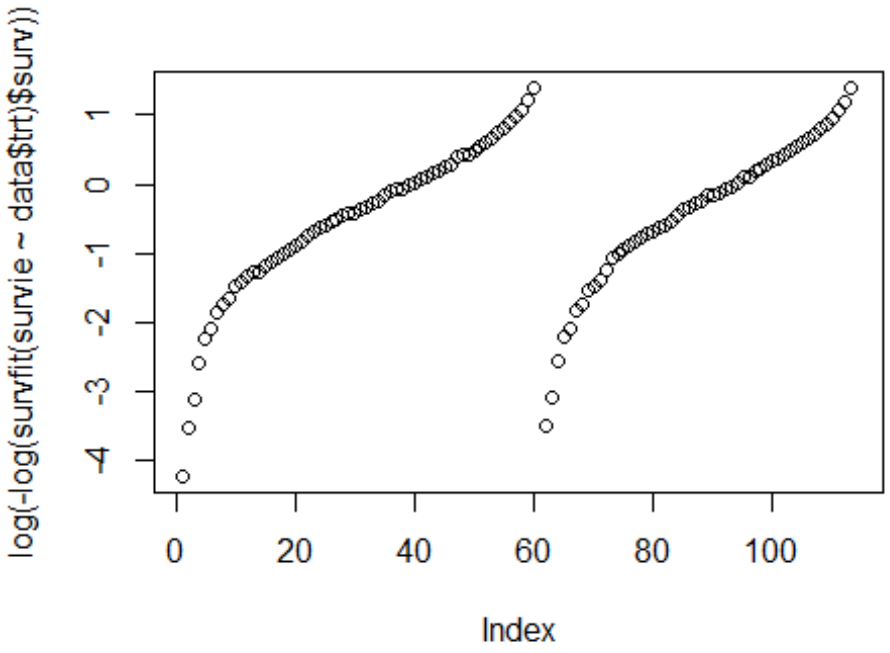
```
## - data$diagtime        1 962.68
## - data$prior_celltype2  1 962.69
## - data$trt_celltype2    1 963.87
## - data$prior            1 964.09
## <none>                  964.68
## - data$age              1 964.74
## - data$trt_prior        1 966.05
## - data$celltype2        1 966.95
## - data$trt              1 970.23
## - data$karno            1 995.59
##
## Step:  AIC=962.68
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_celltype2 + data$trt_prior + data$prior_celltype2
##
##                   Df    AIC
## - data$prior_celltype2  1 960.69
## - data$trt_celltype2    1 961.91
## - data$prior            1 962.41
## <none>                  962.68
## - data$age              1 962.75
## - data$trt_prior        1 964.23
## - data$celltype2        1 965.03
## - data$trt              1 968.24
## - data$karno            1 994.85
##
## Step:  AIC=960.69
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_celltype2 + data$trt_prior
##
##                   Df    AIC
## - data$trt_celltype2  1 959.92
## <none>                960.69
## - data$age            1 960.81
## - data$prior          1 961.06
## - data$trt_prior      1 962.29
## - data$celltype2      1 964.35
## - data$trt            1 966.26
## - data$karno          1 992.98
##
## Step:  AIC=959.92
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$karno + data$trt_prior
##
##               Df    AIC
## <none>           959.92
```

```
## - data$age       1 960.20
## - data$prior     1 960.73
## - data$trt_prior  1 962.40
## - data$trt       1 964.51
## - data$celltype2  1 974.34
## - data$karno     1 993.29

## Call:
## coxph(formula = survie ~ data$prior + data$trt + data$celltype2 +
##     data$age + data$karno + data$trt_prior)
##
##                  coef exp(coef) se(coef)     z      p
## data$prior     0.48346   1.62167  0.28047  1.72   0.085
## data$trt       0.59704   1.81673  0.23365  2.56   0.011
## data$celltype2 -0.81595   0.44222  0.20348 -4.01 6.1e-05
## data$age       -0.01492   0.98519  0.00973 -1.53   0.125
## data$karno     -0.03249   0.96803  0.00546 -5.95 2.7e-09
## data$trt_prior -0.89397   0.40903  0.42456 -2.11   0.035
##
## Likelihood ratio test=63  on 6 df, p=1.11e-11
## n= 137, number of events= 128
```

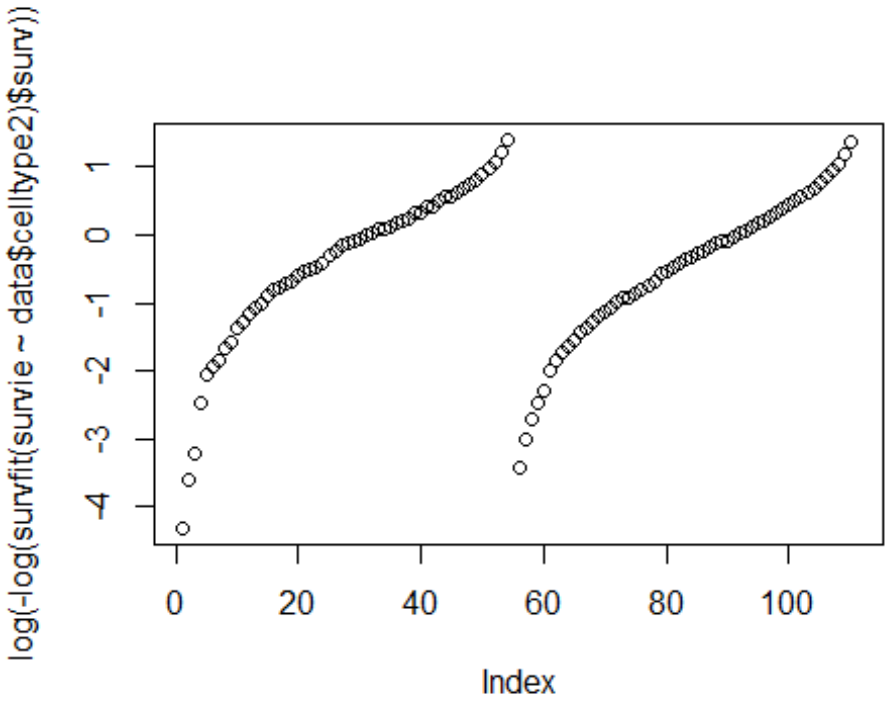
```
plot(log(-log(survfit(survie~data$prior)$surv)))
```



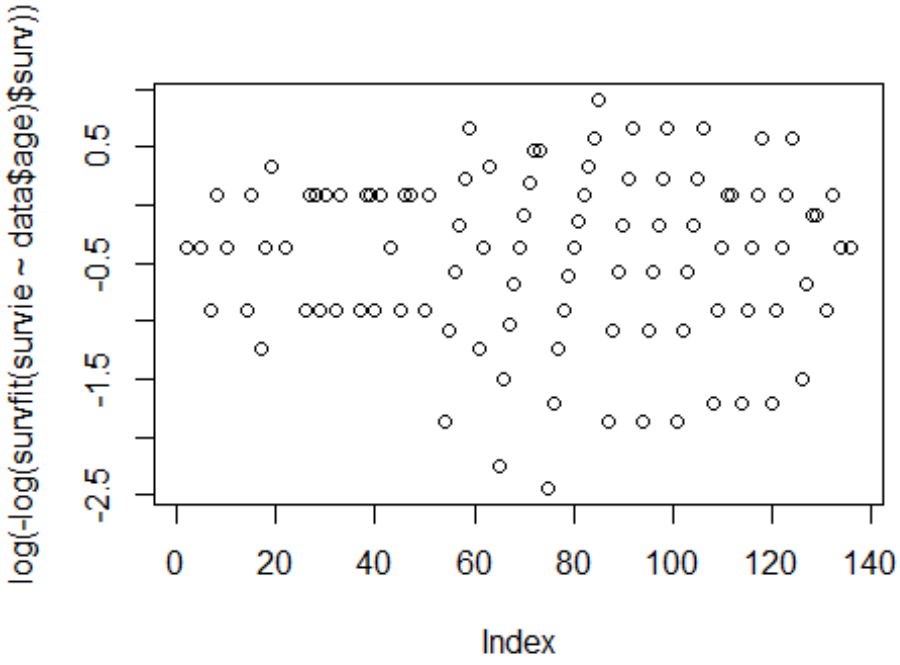

```
plot(log(-log(survfit(survie~data$trt)$surv)))
```



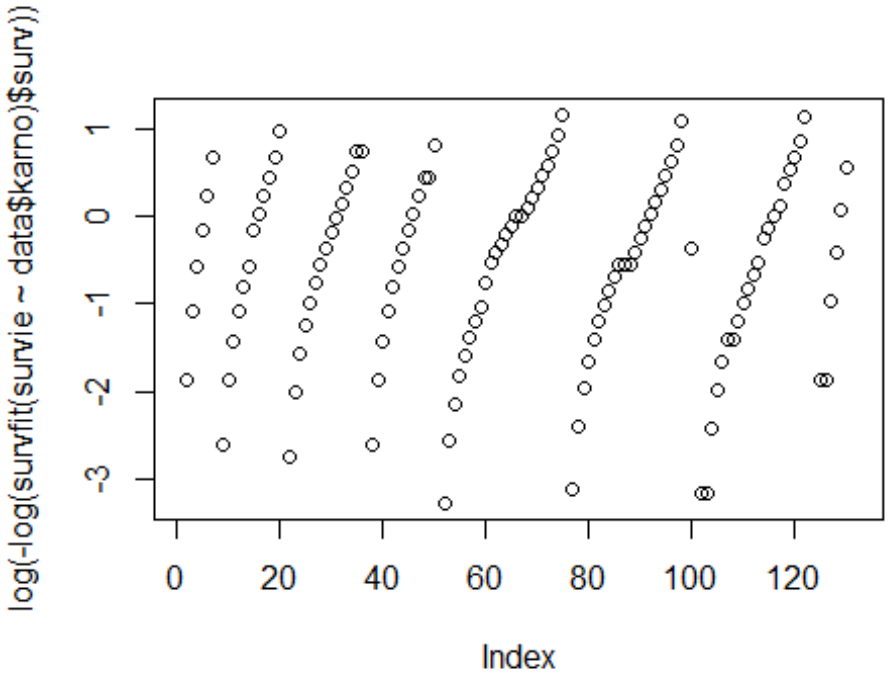

```
plot(log(-log(survfit(survie~data$celltype2)$surv)))
```



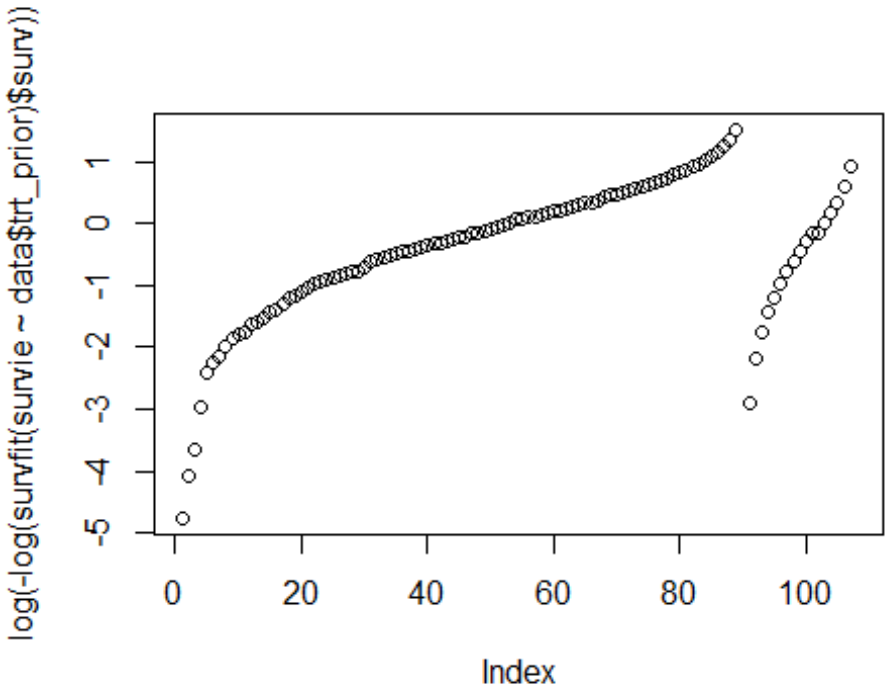

```
plot(log(-log(survfit(survie~data$age)$surv)))
```




```
plot(log(-log(survfit(survie~data$karno)$surv)))
```




```
plot(log(-log(survfit(survie~data$trt_prior)$surv)))
```




```
cox2<- coxph(survie~data$prior+data$trt+data$celltype2+data$age+data$diagtime+data$
karno5+data$karno5_trt+data$karno5_prior+data$karno5_celltype2+data$trt_celltype2+d
ata$trt_prior+data$prior_celltype2)

stepAIC(cox2,scope=list(lower=.~1,upper=.~.),direction="backward")

## Start:  AIC=963.07
## survie ~ data$prior + data$trt + data$celltype2 + data$age +
##     data$diagtime + data$karno5 + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_celltype2 + data$trt_prior +
##     data$prior_celltype2
##
##                      Df   AIC
## - data$celltype2       1 961.08
## - data$prior_celltype2   1 961.19
## - data$diagtime        1 961.23
## - data$karno5          1 961.28
## - data$age             1 961.45
## - data$trt_celltype2     1 961.54
## <none>                   963.07
## - data$trt_prior       1 964.18
## - data$karno5_celltype2  1 964.70
## - data$karno5_trt      1 965.30
## - data$prior           1 968.76
## - data$karno5_prior      1 969.63
```

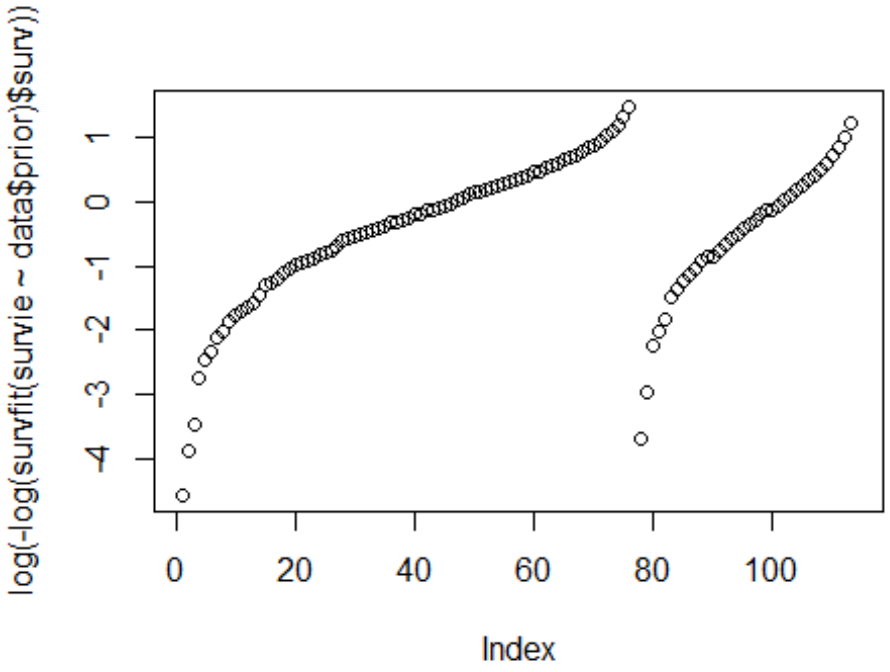
```
## - data$trt            1 971.54
##
## Step:  AIC=961.08
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5 +
##     data$karno5_trt + data$karno5_prior + data$karno5_celltype2 +
##     data$trt_celltype2 + data$trt_prior + data$prior_celltype2
##
##                    Df    AIC
## - data$prior_celltype2   1 959.21
## - data$diagtime          1 959.24
## - data$karno5            1 959.28
## - data$age               1 959.45
## - data$trt_celltype2     1 959.77
## <none>                   961.08
## - data$trt_prior         1 962.38
## - data$karno5_trt        1 963.31
## - data$prior             1 966.92
## - data$karno5_prior      1 968.09
## - data$trt               1 969.61
## - data$karno5_celltype2  1 969.64
##
## Step:  AIC=959.21
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5 +
##     data$karno5_trt + data$karno5_prior + data$karno5_celltype2 +
##     data$trt_celltype2 + data$trt_prior
##
##                    Df    AIC
## - data$karno5            1 957.49
## - data$diagtime          1 957.49
## - data$age               1 957.59
## - data$trt_celltype2     1 957.87
## <none>                   959.21
## - data$trt_prior         1 960.48
## - data$karno5_trt        1 961.33
## - data$karno5_prior      1 966.11
## - data$trt               1 967.62
## - data$prior             1 967.81
## - data$karno5_celltype2  1 968.02
##
## Step:  AIC=957.49
## survie ~ data$prior + data$trt + data$age + data$diagtime + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_celltype2 +
##     data$trt_prior
##
##                    Df    AIC
## - data$age               1 955.71
```

```
## - data$diagtime        1 955.90
## - data$trt_celltype2    1 955.99
## <none>                  957.49
## - data$trt_prior        1 958.72
## - data$karno5_trt       1 964.84
## - data$karno5_prior     1 967.60
## - data$prior            1 968.03
## - data$trt              1 969.04
## - data$karno5_celltype2  1 969.70
##
## Step:  AIC=955.71
## survie ~ data$prior + data$trt + data$diagtime + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_celltype2 +
##     data$trt_prior
##
##                   Df    AIC
## - data$diagtime        1 954.14
## - data$trt_celltype2    1 954.16
## <none>                  955.71
## - data$trt_prior        1 956.72
## - data$karno5_trt       1 963.00
## - data$karno5_prior     1 965.73
## - data$prior            1 966.03
## - data$trt              1 967.07
## - data$karno5_celltype2  1 967.77
##
## Step:  AIC=954.14
## survie ~ data$prior + data$trt + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_celltype2 + data$trt_prior
##
##                   Df    AIC
## - data$trt_celltype2    1 952.50
## <none>                  954.14
## - data$trt_prior        1 955.93
## - data$karno5_trt       1 961.00
## - data$karno5_prior     1 963.76
## - data$prior            1 964.21
## - data$trt              1 965.08
## - data$karno5_celltype2  1 966.45
##
## Step:  AIC=952.5
## survie ~ data$prior + data$trt + data$karno5_trt + data$karno5_prior +
##     data$karno5_celltype2 + data$trt_prior
##
##                   Df    AIC
## <none>                  952.50
```

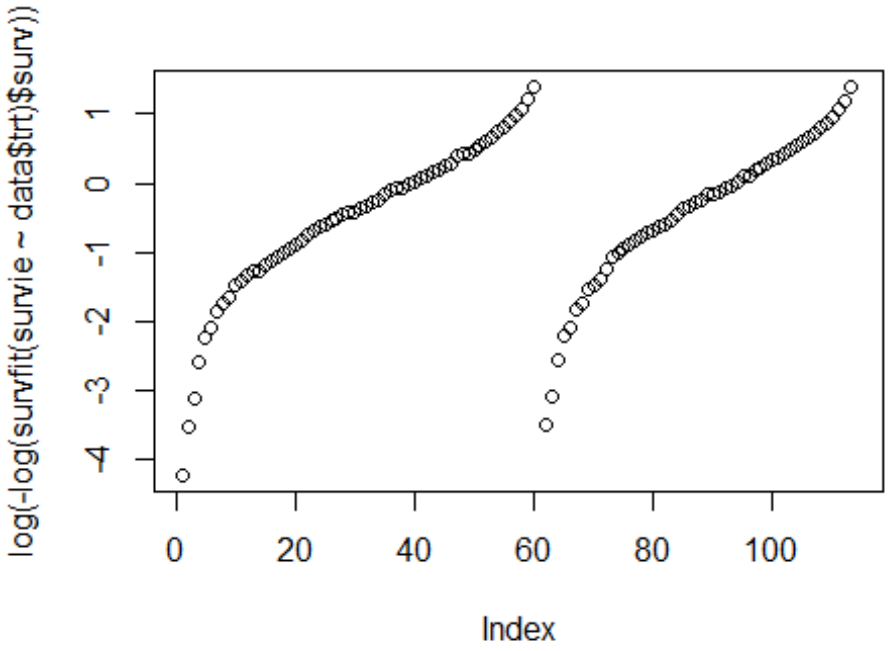
```
## - data$trt_prior        1 954.39
## - data$karno5_trt       1 959.68
## - data$karno5_prior     1 962.11
## - data$prior            1 962.70
## - data$trt              1 963.37
## - data$karno5_celltype2  1 972.10

## Call:
## coxph(formula = survie ~ data$prior + data$trt + data$karno5_trt +
##     data$karno5_prior + data$karno5_celltype2 + data$trt_prior)
##
##                          coef exp(coef) se(coef)     z      p
## data$prior             1.518     4.562    0.393  3.86 0.00011
## data$trt               1.246     3.478    0.327  3.81 0.00014
## data$karno5_trt       -1.036     0.355    0.329 -3.15 0.00164
## data$karno5_prior     -1.448     0.235    0.401 -3.61 0.00030
## data$karno5_celltype2 -1.040     0.354    0.228 -4.56 5.1e-06
## data$trt_prior        -0.807     0.446    0.412 -1.96 0.04985
##
## Likelihood ratio test=70.4  on 6 df, p=3.38e-13
## n= 137, number of events= 128
```

En comptant la variable karno5, la sélection des variables avec stepAIC change. On obtient alors les variables suivantes : prior, trt, karno5, karno5_trt, karno5_prior, karno5_celltype2 et trt_prior. En effet, la disparition de la variable celltype2, qui est importante, semble étrange.

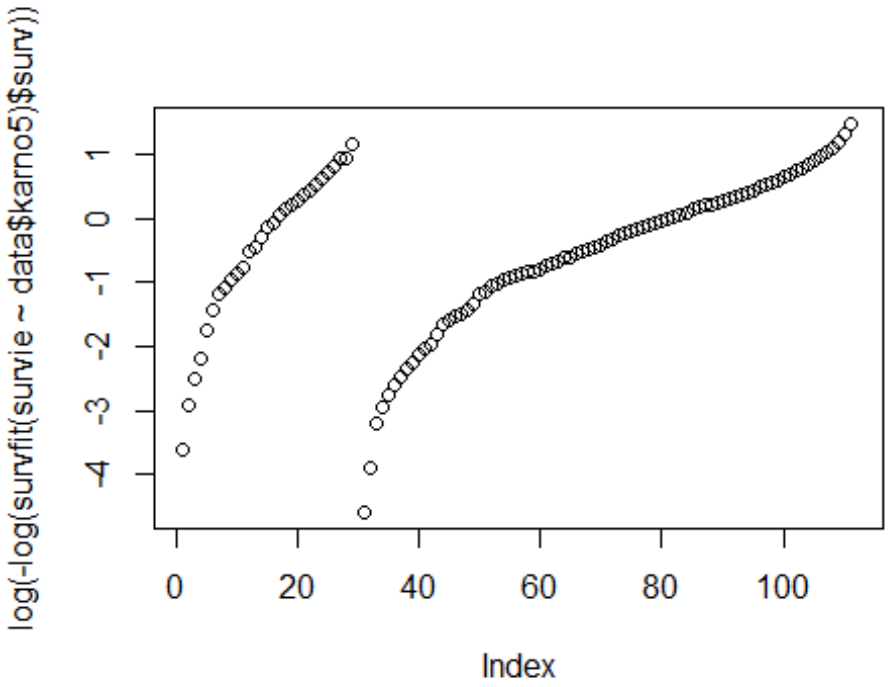
```
plot(log(-log(survfit(survie~data$prior)$surv)))
```



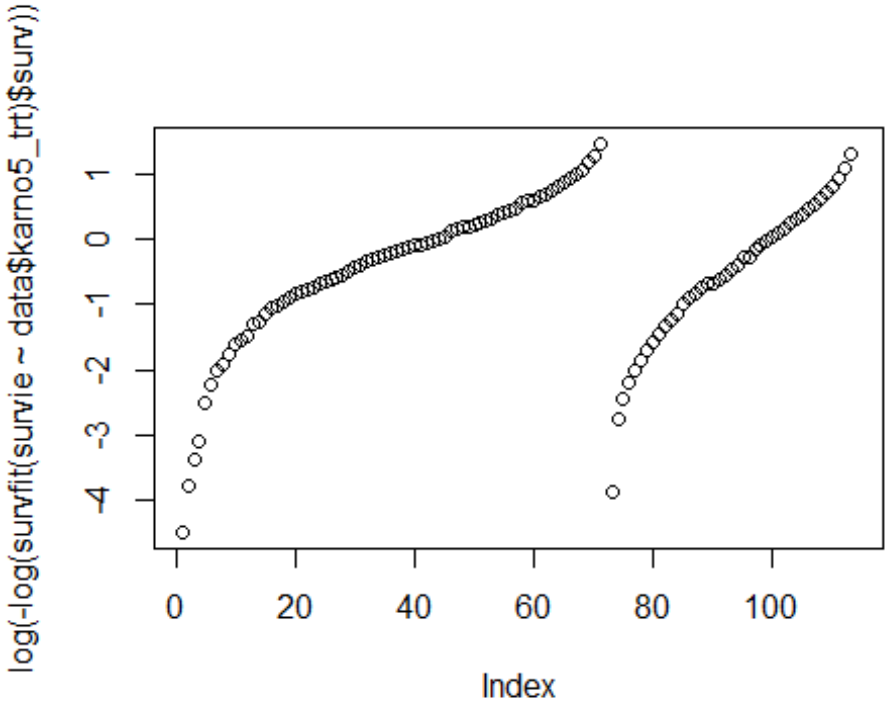

```
plot(log(-log(survfit(survie~data$trt)$surv)))
```



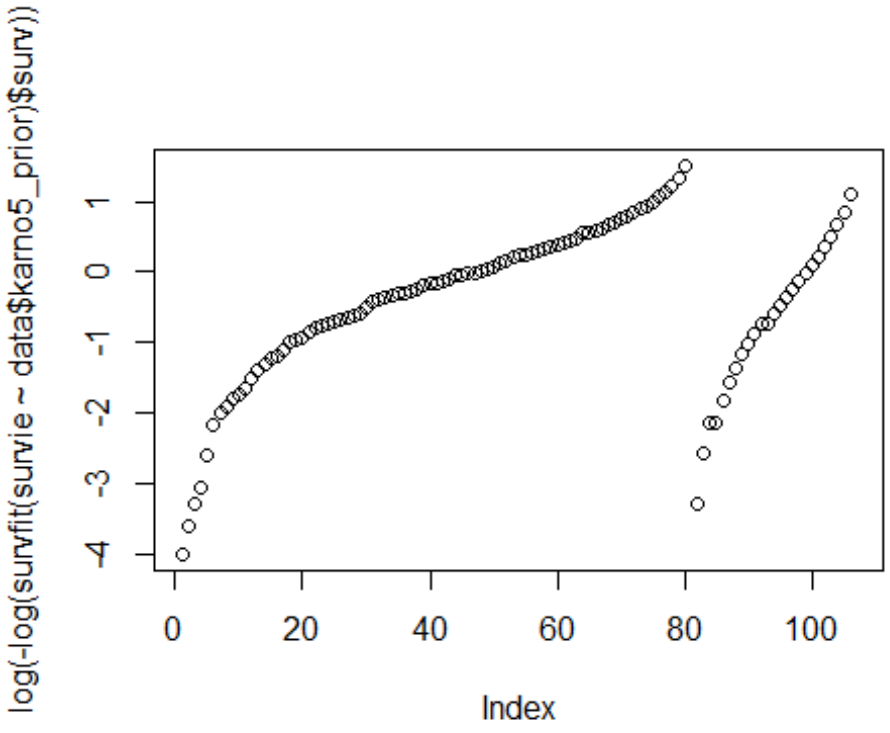

```
plot(log(-log(survfit(survie~data$karno5)$surv)))
```



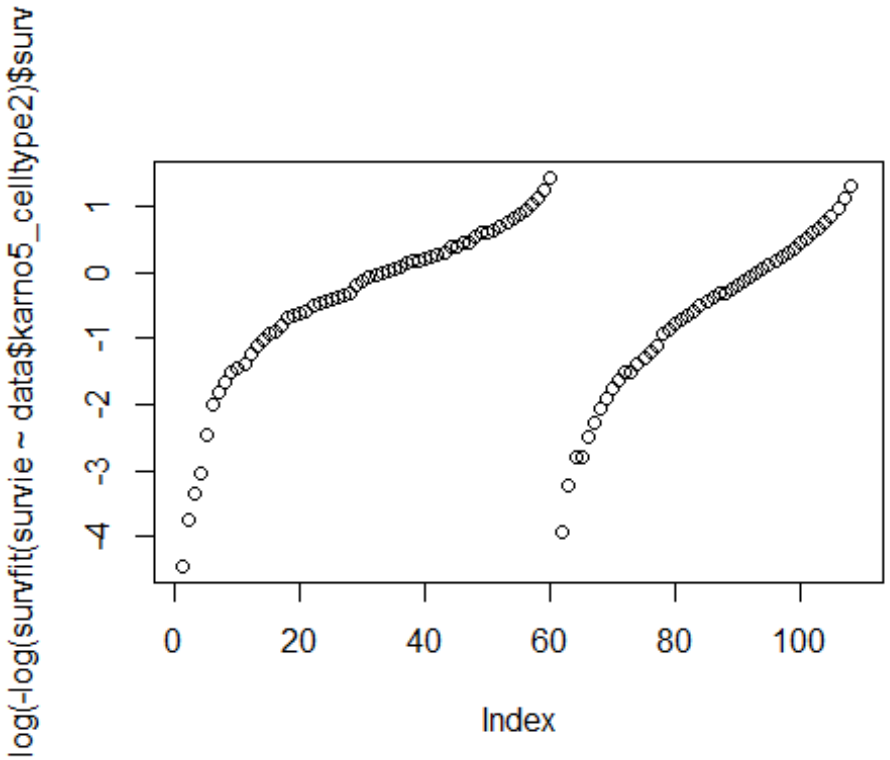

```
plot(log(-log(survfit(survie~data$karno5_trt)$surv)))
```



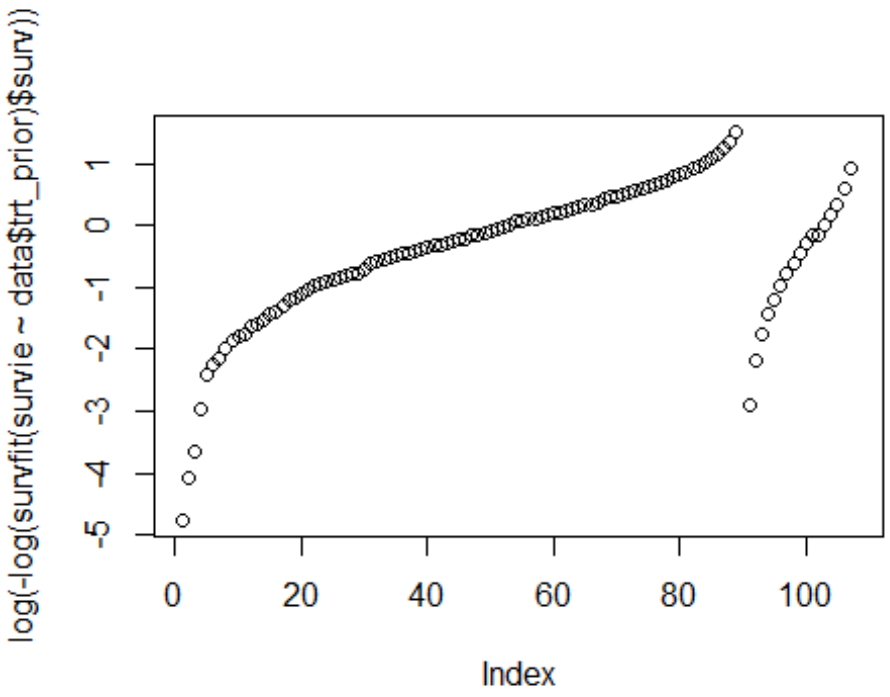

```
plot(log(-log(survfit(survie~data$karno5_prior)$surv)))
```




```
plot(log(-log(survfit(survie~data$karno5_celltype2)$surv)))
```




```
plot(log(-log(survfit(survie~data$trt_prior)$surv)))
```


log(-log(survfit(survie ~ data$trt_prior)$surv))
-5
-4
-3
-2
-1
0
1
0
20
40
60
80
100
Index