Université de Toulouse III

BE Statistique & Santé

Infection par le VIH : Causes d'arrêts de traitement

Auteurs :
M. Ismail Addou
M. Axel Bellec
M. Joseph Meunier

Enseignant :
Mme. Cécile Chouquet

Table des matières

Présentation des objectifs et du contexte de l'étude **5**

Données 5

Objectifs du projet 5

Analyse **7**

1 Description de la population étudiée **8**

1.1 Analyse univariée 8

1.2 Caractéristique des patients 8

1.3 Traitement et son suivi 8

1.4 Contamination 9

1.5 État initial du patient 9

1.6 État du patient à l'abandon du traitement ou fin de suivi . . 10

1.7 Motif d'arrêt du traitement 11

2 Traitement des données manquantes **12**

2.1 Visualisation des valeurs manquantes 12

2.1.1 Suppression des enregistrements avec trop de valeurs manquantes 14

2.2 Stratégie d'interpolation avec apprentissage statistique 14

3 Relation entre l'arrêt du traitement et les caractéristiques des patients **16**

3.1 Préparation des données 16

3.2 Description des données 16

3.2.1 Facteur : Type de traitement 16

3.2.2 Facteur : Âge 17
3.2.3 Facteur : Observance 17
3.2.4 Facteur : Sida 17
3.2.5 Facteur : Sexe 17
3.2.6 Facteur : Mode de contamination 18
3.2.7 Facteur : cd4b1 et cvb1 18
3.2.8 Facteur : cd4b2 et cvb2 18
3.2.9 Conclusion 18
3.3 Analyse de la variance 19
3.4 Régression logistique sur l'arrêt du traitement 19

4 Relation entre le motif d'arrêt du traitement et les caractéristiques des patients 22
4.1 Caractéristique des patients 22
4.2 Traitement et son suivi 23
4.3 Contamination 24
4.4 État initial du patient 25
4.5 État du patient à l'abandon du traitement ou fin de suivi . . 27
4.6 Régression logistique sur les motifs d'arrêt 28
4.6.1 Intolérance/Toxicité 28
4.6.2 Echec thérapeutique 29
4.6.3 Problème d'observance 31
4.6.4 Simplification de traitement 31
4.6.5 Autres 33

5 Modélisation longitudinale par des méthodes d'analyse de survie 35
5.1 Définition de la variable arrêt/censure et du délai correspondant 35
5.2 Survie globale 35
5.2.1 Estimations par la méthode de Kaplan-Meier 35
5.2.2 Représentation de la courbe de survie 36
5.3 Estimation de la survie suivant le type de traitement 37
5.3.1 Estimation de la survie par la méthode de Kaplan-Meier 37
5.3.2 Représentation graphique 38

5.4 Test du Log-Rank 38
5.5 Modèles de Cox 40

Conclusion **45**

Présentation des objectifs et du contexte de l'étude

Données

Les données à notre disposition proviennent de la cohorte NADIS (New AIDS Data Information System). C'est une cohorte multicentrique alimentée par un dossier médical informatisé (DMI). Les données (socio-démographiques et cliniques) sont saisies par les soignants en temps réel lors des consultations ou des hospitalisations. Ce DMI est utilisé dans 30 centres hospitaliers français de la métropole, des Antilles Françaises et de la Guyane.

Ont été inclus, dans l'étude, 1136 patients infectés par le VIH, âgés d'au moins 18 ans, naïfs d'antirétroviraux et initiant une première ligne de HAART entre le 1^{er} janvier 2000 et le 30 juin 2008 dans l'un des 8 sites utilisant NADIS et fournissant des données de qualitée contrôlée. Les variables retenues concernent l'état civil du patient, la maladie et son évolution, le traitement reçu et l'arrêt de ce traitement.

Objectifs du projet

L'objectif général de ce projet est d'étudier les arrêts de traitement survenant au cours du suivi, leurs motifs et les facteurs pouvant favoriser ou empêcher leur survenue. Différents critères concernant l'arrêt du traitement seront considérés dans l'analyse statistique : arrêt pour toutes causes (présence/absence), arrêt pour cause d'intolérance (présence/absence), arrêt pour échec immuno-virologique (présence/absence), et cetera. L'arrêt est défini comme toute interruption d'un ou de plusieurs médicaments de la 1ère ligne, sans tenir compte des variations de posologies.

Ce projet s'appuyant sur des données réelles, il faudra prendre en considération la présence de valeurs manquantes.

Analyse

Notre plan d'analyse sera décomposé en 4 majeures parties. Dans un premier temps nous décrirons la population étudiée. Ensuite nous expliquerons notre de démarche de traitement des valeurs manquantes. Après nous étudierons la relation entre l'arrêt du traitement par un patient suivant ses caractéristiques. In fine, nous mettrons en place des méthodes d'analyse de survie pour étudier la survenue d'un arrêt de traitement au cours du temps.

Partie 1

Description de la population étudiée

1.1 Analyse univariée

1.2 Caractéristique des patients

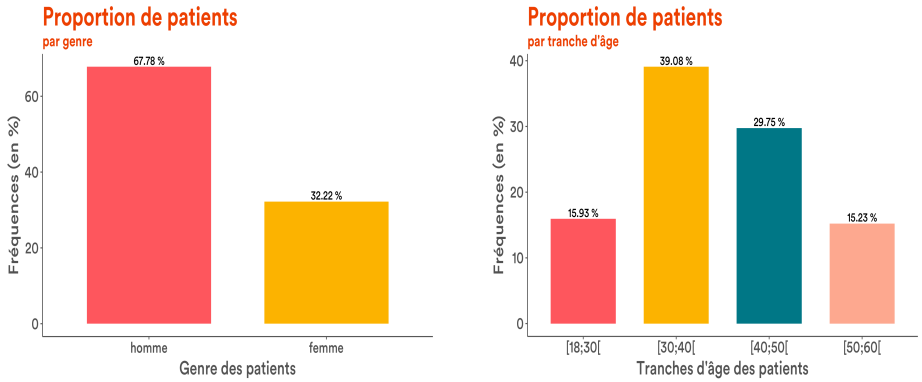

FIGURE 1.1 – Proportion de patients selon le genre et selon la tranche d'âge

Dans la population étudiée, deux tiers des patients sont des hommes, 770 contre 366 femmes. Leur âge varie de 18 à 60 ans. La tranche d'âge modale est $[30; 40]$, elle représente près de 2 patients sur 5 (39% des patients).

1.3 Traitement et son suivi

35% des patients ont reçu le traitement 2IN+1IP, un tier le 2IN+INN. Près de 4 patients sur 5, 79.4%, ne suivent pas leur traitement correctement.

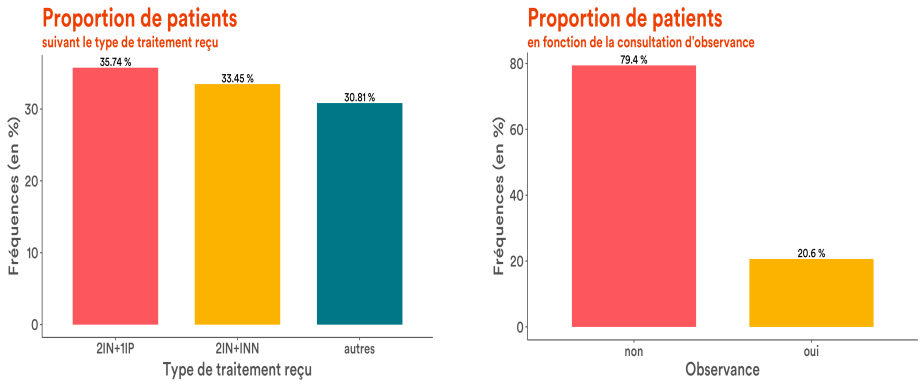

FIGURE 1.2 – Proportion de patients selon le type de traitement et selon l'observance

1.4 Contamination

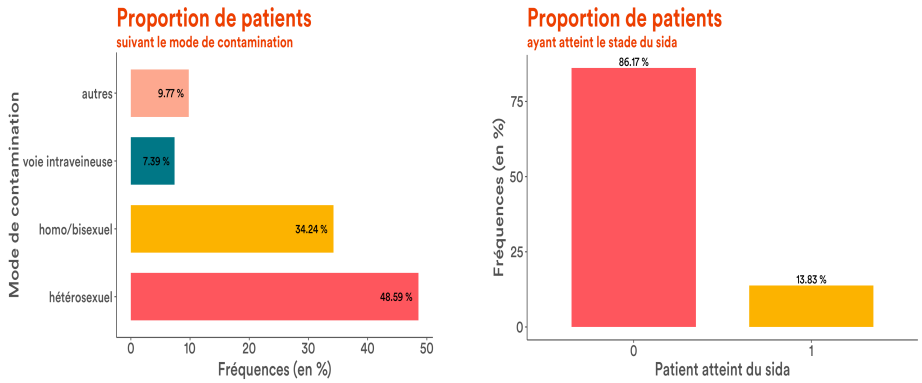

FIGURE 1.3 – Proportion de patients selon le type de contamination et le stade de la maladie

Plus de quatre patients sur cinq sont contaminés par voie sexuelle (49% hétérosexuel et 34% homo/bisexuel). Un peu plus de 7% par voie intraveineuse. Pour près de 4% des patients on ne connaît pas leur stade d'avancement. Parmi ceux dont on connaît le stade, 14% des patients atteints par le VIH ont atteint le stade du SIDA.

1.5 État initial du patient

Au début du traitement, pour 18% des patients, on n'a pas le niveau de *cd*4 et pour plus d'un patient sur quatre, on n'a pas de valeur de la charge virale.

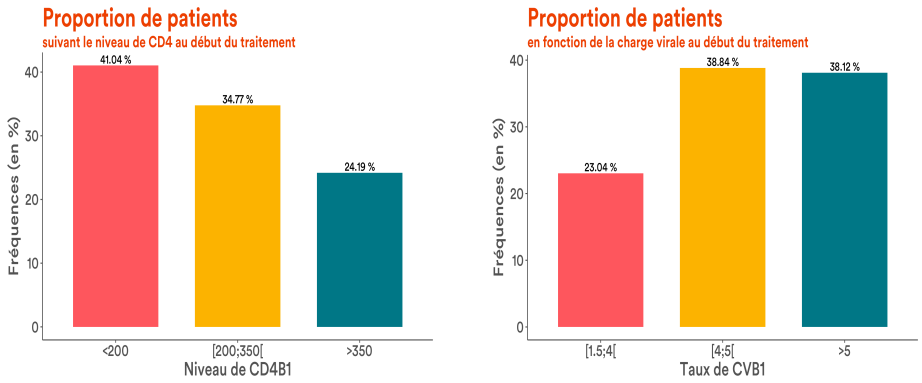

FIGURE 1.4 – Proportion de patients selon le taux de CD4B1 et de CVB1

Lorsque le niveau de CD4 est connu au début du traitement, deux sur cinq ont leur niveau de CD4 inférieur à 200. Un tiers a un niveau compris entre 200 et 350. Et un quart au-dessus de 350. Quand le taux est connu, plus de trois patients sur quatre à une charge virale élevée, 39% entre 4 et 5, 38% au-dessus de 5. Moins d'un quart des patients a une charge inférieure à 4.

1.6 État du patient à l'abandon du traitement ou fin de suivi

A la fin du traitement, pour 10% des patients, on n'a pas le niveau de '*cd*4' et pour environ 15% des patients on n'a pas de valeur de la charge virale.

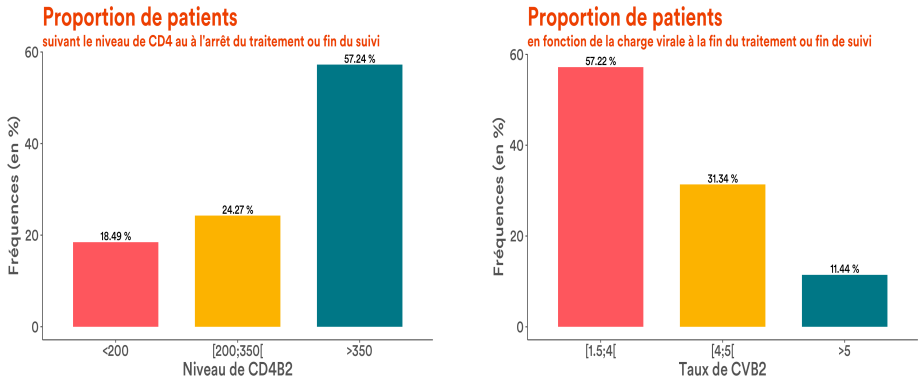

FIGURE 1.5 – Proportion de patients selon le taux de CD4B2 et de CVB2

Quand le niveau est connu, plus d'un patient sur deux, 57%, à un niveau de CD4 supérieur à 350 à la fin de leur suivi. Près d'un patient sur quatre

à un niveau entre 200 et 350 de CD4. Et 18% des patients ont un niveau inférieur à 200 CD4. Quand la charge virale du patient à la fin du traitement est connue, 57% ont une charge inférieure à 1.7, 30% entre 1.7 et 4. Et 11% d'entre eux a une charge supérieure à 4.

1.7 Motif d'arrêt du traitement

Seulement 32% des patients ont continué leur traitement, au moins jusqu'au 30 juin 2008.

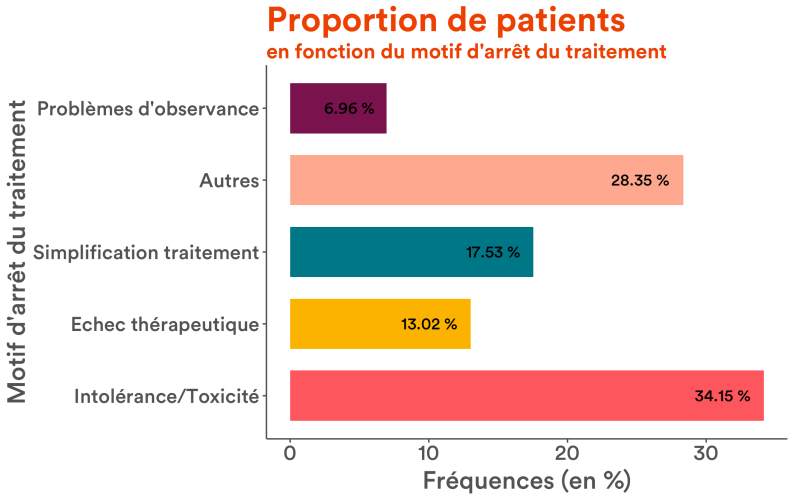

FIGURE 1.6 – Proportion de patients en fonction du motif d'arrêt de traitement

La première cause d'arrêt du traitement est l'intolérance ou la toxicité, 34%. 18% des patients ont un traitement simplifié. Pour 13% des patients l'arrêt est dû à un échec thérapeutique.

Le profil type d'un patient est un homme ayant entre 30 et 40 ans, ayant été contaminé par voie sexuelle (hétérosexuelle) mais n'ayant pas atteint le stade du sida. Il a reçu le traitement `IN+1IP` mais ne le suit pas correctement. Son niveau de CD4 passe de moins de 200 à plus de 350 durant le traitement, sa charge virale diminue, passant de 4-5 à moins de 1,7. Il a tendance à arrêter son traitement pour intolérance/toxicité.

Partie 2

Traitement des données manquantes

Notre jeu de données est un cas d'étude réel et contient par conséquent des valeurs manquantes. Nous décidons d'étudier le nombre de valeurs manquantes pour chaque variable du jeu de données.

2.1 Visualisation des valeurs manquantes

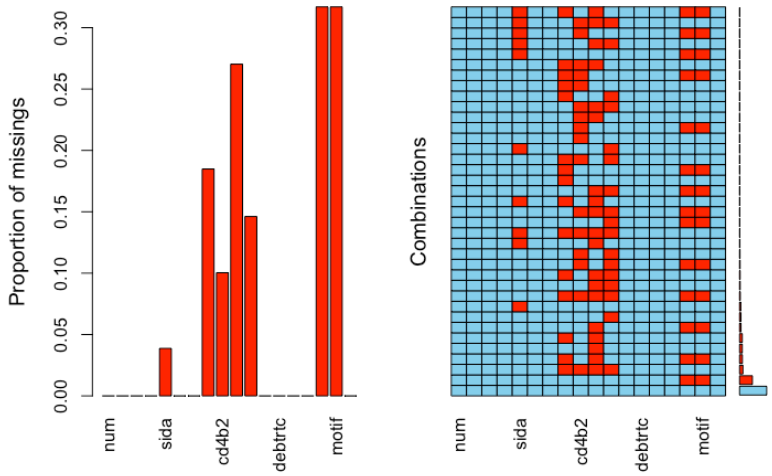

FIGURE 2.1 – Combinaison de valeurs manquantes

Avec l'aide de ce type de visualisation nous pouvons facilement interpréter les valeurs manquantes présentes dans notre jeu de données. Nous constatons que nous avons beaucoup de valeurs manquantes pour les va-

riables `d_arret` et `motif` (33% des valeurs ne sont pas définies). Ceci n'est pas alarmant car la majorité des individus n'ont pas mis fin à leur traitement. C'est donc normal d'avoir des valeurs manquantes les concernant. Il faut davantage s'intéresser aux variables concernant la concentration de CD4 et la charge virale à l'initiation et à l'arrêt du traitement (ou fin du suivi sans arrêt). Pour aller plus loin dans notre analyse, il serait intéressant de relancer le même procédé mais en ne conservant que les variables avec des données manquantes.

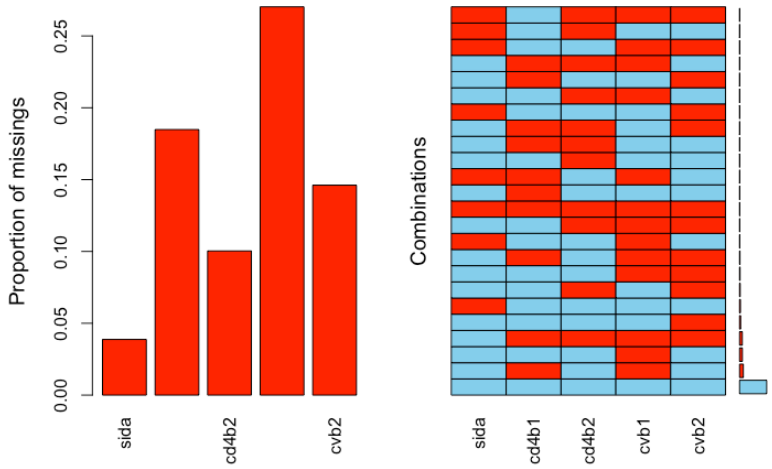

FIGURE 2.2 – Combinaison de valeurs manquantes pour les variables les plus importantes

Notre intuition était basée sur le fait qu'il pourrait y exister des combinaisons entre les valeurs manquantes. Par exemple nous pensions que si un individu avait des valeurs manquantes pour la variable `cd4b1` alors il aurait plus de chances d'avoir des valeurs manquantes pour la variable `cd4b2`. On constate ici que les proportions sont très faibles (visibles à droite sur le graphique des combinaison) et qu'elles ne nous permettent pas de généraliser notre intuition.

Afin de traiter ces données manquantes, plusieurs stratégies sont possibles :

— Supprimer les lignes avec des données manquantes
— Interpoler une classe pour chaque variable factorielle avec donnée manquante

2.1.1 Suppression des enregistrements avec trop de valeurs manquantes

On décide de supprimer les lignes ou il y a un trop grand nombre de combinaisons de valeurs manquantes pour les facteurs :

— `sida`
— `cd4b1`
— `cd4b2`
— `cvb1`
— `cvb2`

On décide de retirer également les enregistrements pour lesquels nous avons 4 modalités manquantes.

Avec ce procédé nous avons retiré 255 observations, il nous en reste donc 881 enregistrements.

Nous avons d'abord décidé de remplacer les valeurs pour le facteur `sida`. Environ 4% des enregistrements contiennent des valeurs manquantes pour le facteur `sida`.

Nous pensons qu'il est possible de faire un lien entre la date de début du traitement et la date de séropositivité (début connu de l'infection). En effet, on peut penser que plus un individu séropositif a tardé à prendre son traitement, plus le charge virale pourrait se développer et le taux de CD4 diminuer. Après une démarche de visualisation, nous ne pouvons pas confirmer notre hypothèse. Nous n'avons pas par exemple un large nombre d'individus atteints du sida avec un delta faible (nombre de jours entre le début de la séropositivité et la date de début du traitement). Notre intuition n'était donc pas fondée.

2.2 Stratégie d'interpolation avec apprentissage statistique

Il nous faut donc nous tourner vers une stratégie d'apprentissage statistique.

On décide de mettre en place un apprentissage statistique pour remplacer les modalités du facteur `sida`.

Nous avons mis en place une validation croisée de 10 plis avec pour classifieur un arbre de décision.

Pour entraîner notre classifieur, nous nous sommes basés sur les features suivantes :

— Consultation d'observance
— Type de traitement

— Arrêt du traitement (variable construite)
— Année de début du traitement
— CD4 à l'initiation du traitement
— CD4 à l'arrêt du traitement ou à la fin du suivi dans arrêt
— Charge virale à l'initiation du traitement
— Charge virale à l'arrêt du traitement ou à la fin du suivi dans arrêt

Le nombre de patients malades du sida ne représente a peine 20% du jeu de données. Pour ne pas faire de sur-apprentissage, nous avons créé un jeu de données d'apprentissage composé d'autant de patients atteints du sida que de patients non malades. Notre classifieur a ensuite prédit les valeurs pour chacunes de nos valeurs manquantes de la variable sida. Nous avons obtenu une assez bonne précision de 75%.

Nous avons adopté la même stratégie d'apprentissage pour les 4 variables `cd4b1`, `cd4b2`, `cvb1` et `cvb2`. Néanmoins, nos classifieurs sont beaucoup moins performants (précision inférieure à 50%). On ne peut pas donc remplacer les valeurs manquantes pertinemment. Nous avons donc décidé de supprimer les lignes avec des valeurs manquantes pour les variables `cd4b1`, `cd4b2`, `cvb1` et `cvb2`.

Nous avons donc conservé près des deux tiers des enregistrements (767 observations).

Partie 3

Relation entre l'arrêt du traitement et les caractéristiques des patients

3.1 Préparation des données

Dans un premier temps nous construisons un variable booléene qui vaut `TRUE` si le patient en question a arrêté son traitement et `FALSE` sinon.

3.2 Description des données

Avant le traitement des données manquantes nous avions 776 sur 1136 individus qui arrêtent leurs traitements soient 68% de la cohorte. Le traitement des données a engendré la suppression dun nombre de ligne ou il yavait beaucoup dattributs non renseignés à la fois. Après suppression, nous avons 527 sur 767 individus qui arrêtent leurs traitements soient 68% de la population. On remarque que la proportion est donc similaire.

Dans un premier temps, nous allons analyser l'indépendance des facteurs par rapport aux arrêts de traitement chez les différents patients.

3.2.1 Facteur : Type de traitement

On observe que presque 86% des individus qui ont suivi le traitement 2 ont arrêté, d'où l'hypothèse que le nouveau traitement (traitement 2) est moins efficace que les premiers, et entraîne de plus en plus d'arrêts. D'après le test de chi-2, on observe que la p-value *ll* 5% (p-value= 7.197e-11), alors l'hypothèse de l'indépendance entre le type du traitement appliqué sur le

patient et la variable réponse booléenne `is_arret` est rejetée.

3.2.2 Facteur : Âge

Dans notre cas aucun patient n'appartient à la tranche d'âge de plus de 60 ans donc on ignore la tranche dâge 4. Le facteur âge na apparemment pas beaucoup deffet vis à vis du pourcentage des individus qui arrêtent le traitement, on remarque une légère hausse du pourcentage darrêt pour les individus âgés par comparaison aux individus moins âgés. Le test de χ^2 appuie notre hypothèse de départ puisque on a p-value = 0.1768 donc on rejette lhypothèse de lindépendance par rapport à la variable âge.

3.2.3 Facteur : Observance

70% des individus qui ont arrêté le traitement sont marqués comme des patients qui ne respectent pas la posologie de leurs traitement, et 76% des individus qui nont pas arrêté le traitement ne respectent pas cette posologie non plus, on ne peut tirer aucune conclusion de cette observation. Et visiblement, la variable posologie nest pas une variable explicative, puisque selon le test de chi2 le facteur dobservance est indépendant de la variable réponse arrêt du traitement (p-value $>$ 5%).

74,1% des individus qui ont arrêté le traitement sont marqués comme des patients qui ne respectent pas la posologie de leurs traitement, et 67.5% des individus qui nont pas arrêté le traitement ne respectent pas cette posologie non plus, on ne peut pas tirer de conclusion à partir de cette observation. Et visiblement, la variable posologie nest pas une variable explicative, puisque selon le test de chi2 le facteur dobservance est indépendant de la variable réponse arrêt du traitement puisque p-value = 0.06731 (p-value $>$ 5%).

3.2.4 Facteur : Sida

81,5% des gens qui ont arrêté le traitement ne sont pas encore passé à la phase du sida, mais aussi 90,8% des gens qui suivent toujours le traitement ne sont pas encore passé à cette phase. On fait un test de Chi2 pour voir est ce quil y a une indépendance entre larrêt du traitement et le facteur sida, et le résultat est p-value $<$ 5% (p-value = 0.001527), on rejette donc l'hypothèse de l'indépendance entre le SIDA et l'arrêt du traitement.

3.2.5 Facteur : Sexe

Le facteur sexe nest pas significatif puisque la p-value $>$ 5% (p-value = 0.5292) on accepte lhypothèse de lindépendance selon le test de chi2, et on

conclut que larrêt du traitement agit indépendemment du sexe.

3.2.6 Facteur : Mode de contamination

La plupart des individus de la population étudiée ont été contaminés selon le mode de contamination 0 et 1, c'est à dire à travers des rapports hétérosexuels où homo/bisexuels. Et les modes de contamination 2 et 3 sont rares par rapport aux 0 et 1. D'un autre côté, selon le test de chi2 la p-value $< 5\%$ (p-value = 0.002978) on rejette donc lhypothèse de lindépendance.

3.2.7 Facteur : cd4b1 et cvb1

La valeur de l'indicateur cd4b1 est le taux des lymphocytes dans le sang du patient à l'initiation du traitement. Quand on applique le test de chi2 sur ce taux, on a p-value $< 5\%$ (p-value = 0.04877) c'est tout juste significatif, d'où on rejette l'hypothèse de l'indépendance. Par contre pour le taux du cvb1 qui est le taux de la charge virale dans le sang du patient, avec une p-value $> 5\%$ (p-value = 0.5393), on accepte l'hypothèse de l'indépendance. On peut conclure que les taux de mesures du début de traitement cd4b1 et cvb1 sont presque indépendants de l'arrêt du traitement d'un patient, avec un effet léger par rapport au taux initial des lymphocytes chez l'individu.

3.2.8 Facteur : cd4b2 et cvb2

La valeur de l'indicateur de cd4b2 et cvb2 sont respectivement les taux des lymphocytes et la charge virale chez le patient dans l'arrêt du traitement, selon le test de chi2 on remarque qu'elles ont des p-value «$< 5\%$ respectivement p-value = 1.216e-10 et p-value $<$ 2.2e-16, d'où on rejette l'hypothèse de l'indépendance, justement car ce sont les indicateurs dont se base l'expert pour arrêter le traitement au patient.

3.2.9 Conclusion

Selon le test de Chi-2 les facteurs suivants sont des facteurs indépendants de l'arrêt du traitement chez les patients en question :

— `age`
— `observ`
— `sexe`
— `cvb1`

3.3 Analyse de la variance

Dans la section suivante nous allons effectuer une analyse sur la variance pour décerner les facteurs qui ont plus d'effet sur la variable réponse arrêt du traitement. Le résultat de cette analyse montre que des variables sont plus significatifs que d'autres, et on va les préciser par ordre décroissant de la variable la plus significative avec la plus petite p-value à la moins significative.

No	facteur	p-value	Signif.
1	cvb2	<2e-16	
2	typetrt	1.31e-12	***
3	cd4b2	6.57e-11	
4	sida	0.00103	**
5	cd4b1	0.0488	*
6	observ	0.0554	.
7	age	0.177	
8	sexe	0.476	
9	cvb1	0.54	

Conclusion : Selon l'analyse de la variance, on conclut que type de traitement est le facteur le plus significatif après le cvb2 avec une p-value = $1.31e-12 \lll 5\%$.

3.4 Régression logistique sur l'arrêt du traitement

Dans cette partie nous appliquons la régression logistique sur l'arrêt du traitement par rapport aux variables typetrt, age, sida, sexe, conta, cd4b1, cd4b2, cvb1 et cvb2. Les résultats de cette analyse montrent que les facteurs age, sexe, cd4b1 et cvb1 n'ont pas d'effet significatif sur l'arrêt ou non du traitement.

Alors les facteurs significatifs sont le type du traitement (typetrt), sida et le mode de contamination (conta). Par contre les facteurs cd4b2 et cvb2 sont les plus significatifs et c'est normal puisque ce sont les indicateurs qui poussent le médecin d'arrêter le traitement à un individu.

D'un autre côté le type de traitement 2 a un écart très significatif par rapport aux traitements 1 et 2, avec une p-value d'écart (3.34e-07 «< 5%), par contre le mode de contamination 3 et la valeur 1 pour le facteur sida ont des effets significatifs par rapport aux autres niveaux (p-value $< 5\%$).

Sélection de variables explicatives :
À présent nous allons sélectionner automatiquement les variables explicatives de notre modèle selon une démarche descendante. Le résultat de cette

sélection est le suivant :

facteur	p-value	Signif.
typetrt	4.286e-15	***
sida	0.0004381	***
conta	0.0025757	**
cd4b2	8.435e-09	***
cvb2	3.439e-16	***

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Intolerance/Toxicite' AUC = 0.7917457.

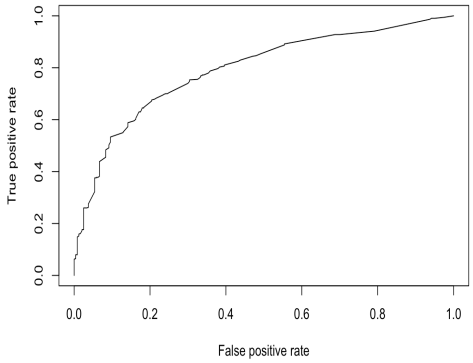

FIGURE 3.1 – Courbe ROC du modèle de la régression logistique sur l'arrêt du traitement

L'analyse suivante nous aidera à décerner les valeurs des facteurs qui mènent un individu à arrêter son traitement. On observe que suivre le type de traitement 2 multiplie le risque d'arrêter le traitement par 1,22 par rapport au traitement 0. Si l'individu est passé à la phase SIDA il a 10% plus de chance d'arrêter le traitement que s'il n'en a pas, le type de contamination 2 multiplie la chance d'arrêter le traitement par 1,07 et la contamination 3 multiplie par 0,86, mais il faut prendre en considération que les individus dont le mode de contamination est 2 où 3 sont rares. D'un autre côté si le taux de cd4b2 appartient à la tranche 0 (cela veut dire que son taux de lymphocyte est à moins de 200) il a 20% plus de chance d'arrêter que s'il appartient à la tranche 2 (taux de lymphocytes + de 350), et vice versa pour le taux de la charge viral cvb2, où l'individu a 30% plus de chance d'arrêter le traitement s'il appartient à la phase 2.

Le tableau suivant illustre les différents moyennes des effets entre les niveaux de facteurs avec des intervalles de confiances de risque 5%.

	OR	2.5 %	97.5 %
(Intercept)	1.8360690	1.6623309	2.0279654
typetrt1	0.9958740	0.9260702	1.0709394
typetrt2	1.2220874	1.1357210	1.3150217
sida1	1.0906835	1.0006928	1.1887670
conta1	1.0388967	0.9732881	1.1089279
conta2	1.0756646	0.9478888	1.2206646
conta3	0.8668001	0.7799432	0.9633297
cd4b21	0.9150426	0.8344320	1.0034405
cd4b22	0.8374774	0.7684945	0.9126524
cvb21	1.2970639	1.2129629	1.3869960
cvb22	1.3228827	1.2018301	1.4561281

Partie 4

Relation entre le motif d'arrêt du traitement et les caractéristiques des patients

Nous allons maintenant traiter uniquement les données des patients ayant arrêté leur traitement. Cela a pour objectif de remarquer si les caractéristiques des patients sont les mêmes suivant le motif d'arrêt. Et si pour chacun d'en eux une population a risque existe.

4.1 Caractéristique des patients

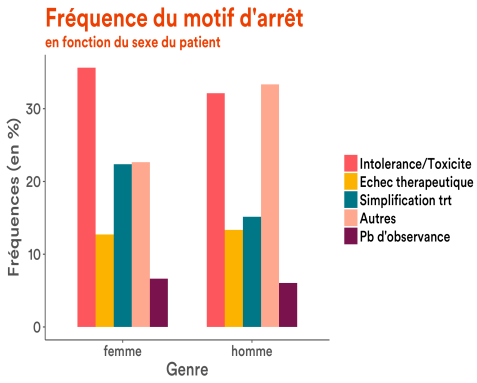

FIGURE 4.1 – Proportion de patients selon le genre et selon le sexe

Les femmes arrêtent leur traitement plus souvent pour des raisons particulières (autres) que les hommes, 34% contre 25% et elles ont moins souvent une simplification de leur traitement. Sinon que le patient soit une femme ou un homme les proportions d'arrêts sont presque identique quelque soit le

motif. Lorsque l'on test l'indépendance entre le sexe et le motif d'arrêt, on obtient une p-value bien supérieure à 0,05 avec 0,3943. Le motif le sexe du patient sont indépendants.

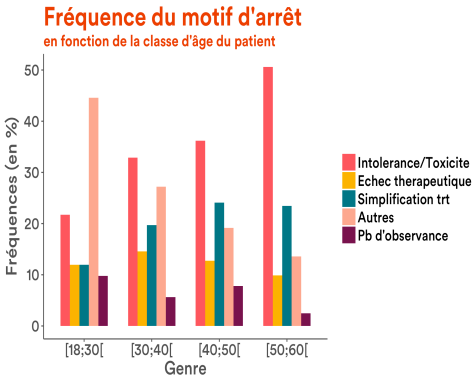

FIGURE 4.2 – Proportion de patients selon le genre et selon la tranche d'âge

Les patients jeunes, 18-30 ans, arrêtent bien moins souvent pour une intolérance au traitement contrairement aux plus âgés, 50-60 ans (25% des jeunes contre 49% des patients âgés). Par contre on peut remarquer 12% des jeunes arrêtent le traitement car ils ne le suivent pas correctement contre moins de 2% des 50-60 ans. Un patient sur cinq de plus de 40 ans arrêtent le suivi par une simplification du traitement, chez les moins de 30 ans moins de 10% des arrêts sont dû à une simplification. Les 30-40 qui arrêtent le font plus souvent pour un échec thérapeutique que la moyenne contrairement aux jeunes. Les deux jeunes sur cinq ne font plus le suivi pour d'autres raisons, moins de 15% du temps chez les 50-60 ans, contre un patient sur quatre en moyenne. Les deux variables sont significativement liées, p-value = 0,0001497.

4.2 Traitement et son suivi

Les traitements 2IN+1IP et 2IN+INN ont une moins bonne tolérance, +5 points et +7 points par rapport à la moyenne, contrairement aux autres traitements. Pour le traitement 2IN+INN, on retrouve peu d'échec thérapeutique, -5 points. Les patients prenant un autre traitement arrêtent eux plus souvent dû à un échec thérapeutique, +5 points. A l'aide du test du Chi-deux on remarque un lien significatif entre le motif et le type de traitement reçu, p-value = 0,02.

Les patients qui suivent leurs traitements correctement arrêtent plus souvent le suivi dû à une intolérance ou simplification du traitement mais moins souvent suite à un échec thérapeutique. L'observance d'un patient est très significativement liée au motif d'arrêt, p-value = 5,023e-05.

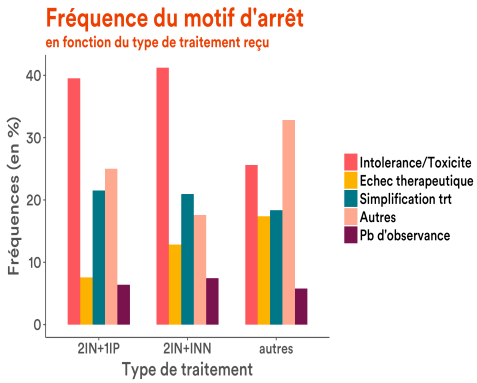

FIGURE 4.3 – Proportion de patients selon le motif d'arrêt par type de traitement

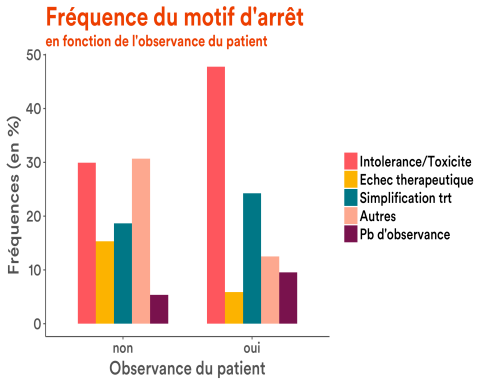

FIGURE 4.4 – Proportion de patients selon le motif d'arrêt par observance ou non

4.3 Contamination

Les patients contaminés par toxico voie intraveineuse arrêtent plus souvent par un problème d'observance, 8 points au dessus de la moyenne, mais bien moins souvent que l'ensemble par simplification du traitement, 12% des cas contre 20%. Les patients contaminés par d'autres voies sont arrêtés plus souvent dû a une intolérance, 42% des cas contre 34% en moyenne et pour des problèmes d'observance, 10% de leur arrêt contre $6,5\%$ en moyenne. En revanche ils stop leur traitement, plus rarement a cause d'une simplification de ce traitement. Pour tester le lien entre le motif d'arrêt et le type de contamination, nous devons, par manque d'effectifs, associer les groupes toxico intraveineuse et autres. Les variables ne sont pas significativement liées, p-value = 0,07984.

Les patients ayant atteint le stade du sida, on arrêtait plus souvent dû a

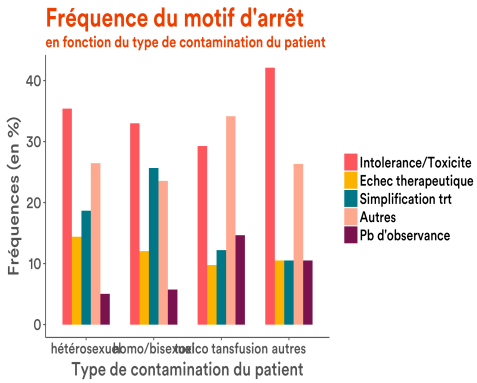

FIGURE 4.5 – Proportion de patients selon le motif d'arrêt en fonction du type de contamination

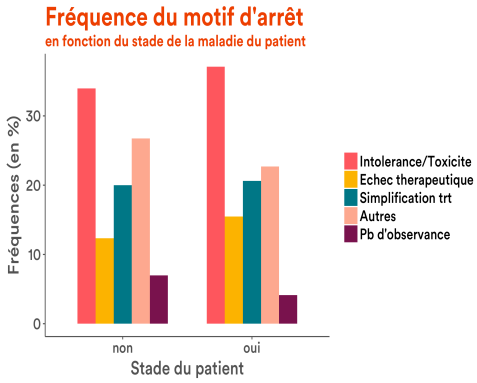

FIGURE 4.6 – Proportion de patients selon le motif d'arrêt suivant le stade de la maladie

une intolérance ou un échec thérapeutique, respectivement 3.4 et 2.5 points au-dessus de la moyenne. Mais globalement quelques soit le stade de la maladie, il n'y a pas de grande différence dans le motif d'arrêt. Cela se vérifie avec le test du chi-deux, p-value = 0.9684.

4.4 État initial du patient

Les patients ayant un niveau de CD4 supérieur à 350 ont plus souvent régulièrement arrêté pour d'autres motifs, +8 points, et un peu moins d'échec thérapeutique. Mais dans l'ensemble, il n'y a pas d'écart sur la proportion de motif d'arrêt pour chaque niveau de CD4 au début du traitement. Visuellement on ne retrouve pas de lien entre le niveau de CD4 initial et le motif d'arrêt. Cela est vérifié avec le test du chi-deux, p-value = 0.6472.

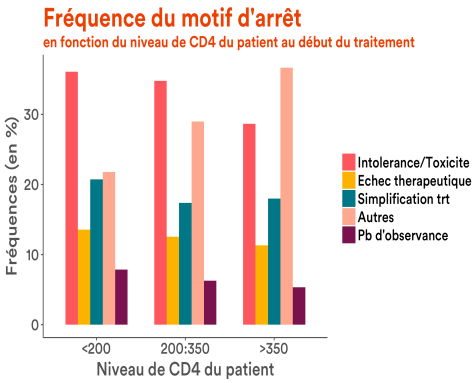

FIGURE 4.7 – Proportion de patients selon le motif d'arrêt en fonction du niveau de CD4 initial

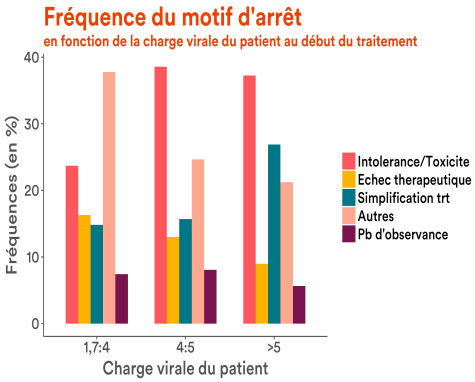

FIGURE 4.8 – Proportion de patients selon le motif d'arrêt suivant la charge virale initiale

Les patients ayant eu une charge virale "faible" au début du traitement ont été plus tolérants au traitement et ont eux leur traitement simplifier moins souvent que la moyenne, respectivement -9.3 et -6,3 points de moins que la moyenne. Par contre il y a eu plus d'échec thérapeutique, +4,1 points. Les patients avec une charge virale très élevés ont eux plus souvent une simplification de leur traitement, +7 points, et moins d'échec thérapeutique, -3 points. La charge virale au début du traitement est significativement liée au motif d'arrêt, p-value = 0,01046.

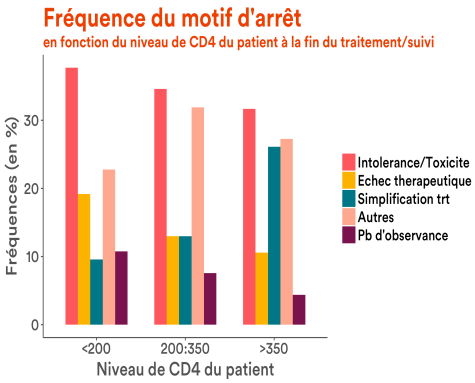

FIGURE 4.9 – Proportion de patients selon le motif d'arrêt en fonction du niveau de CD4 final

4.5 État du patient à l'abandon du traitement ou fin de suivi

Les patients ayant un niveau de CD4 inférieur à 200 ont eu plus souvent un échec thérapeutique, +6%, ou un problème d'observance, +3%, par contre ils ont près de deux fois moins souvent une simplification du traitement. A l'inverse des patients ayant eu leur niveau de CD4 qui est au-dessus de 350 où 27,6% ont une simplification du traitement contre 20,1% en moyenne. Les patients, qui ont leur niveau de CD4 entre 200 et 350, ont plus souvent arrêté suite à une intolérance, 4 points au-dessus de la moyenne. Ils ont aussi eux moins de simplification de leur traitement, 6 points de moins que la moyenne. Le test du chi-deux nous permet de valider le lien entre ces deux variables, p-value = 0,01059.

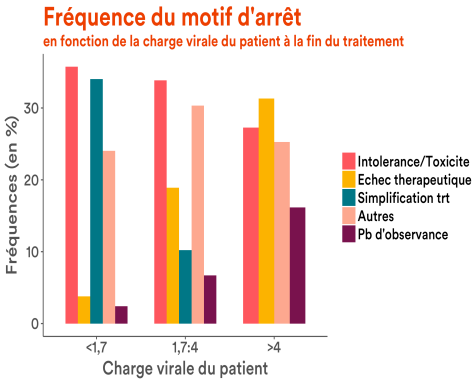

FIGURE 4.10 – Proportion de patients selon le motif d'arrêt suivant la charge virale finale

Les patients qui ont une charge virale inférieure à 1,7 ont arrêté plus

souvent que la moyenne due à une simplification du traitement, 16 points de plus que la moyenne. A l'inverse, ils ont rarement arrêté à cause d'un échec thérapeutique et pour un problème d'observance, 8 et 3 points de moins que la moyenne. On retrouve le cas inverse pour les patients ayant une charge virale supérieure à 4. Ils ont un grand nombre d'échec thérapeutique et de problème d'observance, +15 et +10 points par rapport à la moyenne, et moins de simplification, -19 points (il n'y en a pas). Les patients ayant une charge virale entre 1,7 et 4 ont moins souvent arrêté dû à une simplification du traitement, 11% contre 20% en moyenne mais plus souvent pour d'autres raisons, 31% contre 26%. Ces deux variables sont significativement liées avec une p-value $<$ 2,2e-16.

4.6 Régression logistique sur les motifs d'arrêt

Dans cette partie nous allons prendre en compte que les individus qui ont arrêté le traitement et puis étudier les facteurs ont favorisé l'arrêt du traitement chez deux pour chaque motif.

Voici les différents motifs d'arrêt qui ont été renseigné pour l'ensemble des individus.

— Intolérance/Toxicité
— Simplification traitement
— Échec thérapeutique
— Problème d'observance
— Autres

A ce niveau nous appliquerons les régressions logistiques pour chaque niveau de motif.

4.6.1 Intolérance/Toxicité

Dans cette partie nous établirons la regression logistique sur le motif 'Intolerance/Toxicite' en se basant sur la nouvelle colonne dont les valeurs sont de type booléan qui prennent la valeur TRUE si le motif est 'Intolerance/Toxicite'.

Le résultat de cette analyse montre que le facteur age est fortement lié à l'arret avec motif 'Intolerance/Toxicite' avec une p-value *lll* 5% très significative (p-value = 0.0009237). Puis les facteurs type traitement 'typetrt' (p-value = 0.0017076) et observance 'observ' (p-value = 0.0046916) viennent en second lieu avec des p-value significatives « 5%.

D'un autre côté on remarque que la tranche d'âge 3 est très significatives avec une p-value d'écart «$<$ 5% (p-value= 0.00039), d'autres parts le type de traitement 2 a un écart très significatif par rapport aux traitements 1

et 0, avec une p-value d'écart ($0.02151 < 5\%$), et le respect des doses noté observ=1 a une p-value d'écart significative « 5% (p-value=0.0046916).

A présent, on sélectionne automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante, en ne gardant que les facteurs qui impactent sur la variable réponse 'Intolerance/Toxicite', et le résultat de cette sélection du modèle complet donne en sortie l'âge, l'observance, le type du traitement et cvb1.

Le résultat de cette démarche est comme suit :

Facteur	p-value	signif.
typetrt	0.0016386	**
age	0.0008784	***
observ	0.0044861	**

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Intolerance/Toxicite' AUC = 0.648495

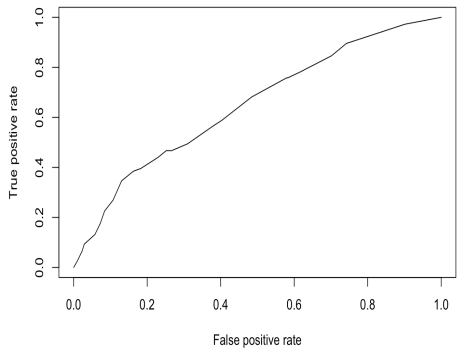

FIGURE 4.11 – Courbe ROC du modèle de la regression logistique pour le motif Intolérance/Toxicité

4.6.2 Echec thérapeutique

Dans cette partie nous établirons la régression logistique sur le motif 'Echec thérapeutique' en se basant sur la nouvelle colonne dont les valeurs sont de type booléen qui prennent la valeur TRUE si le motif est 'Echec thérapeutique'.

Le résultat de cette analyse montre que le facteur d'observance (observ) est fortement lié à l'arrêt du traitement des individus avec motif 'Echec thérapeutique' avec une p-value *ll* 5% très significative (p-value = 0.006522). Puis le facteur type du traitement (typetrt) qui a une p-value significative <

5% (p-value= 0.012664), et les deux mesures cd4b2 et cvb2 qui sont aussi très significatives avec des p-value respectivement égales à 0.026283 et 1.15e-07.

D'un autre côté on remarque que le type de traitement 2 a un écart significatif par rapport au traitement 0 avec une p-value d'écart (p-value = 0.043582 « 5%), l'observance de valeur 1 a un effet très significatif avec une p-value « 5% (p-value = 0.004129), et le cvb2 est significatif pour le niveau 1 et 2 avec des p-values très significatifs respectivement 0.000158 et 1.63e-07, c'est à dire que le niveau 1 engendre dans beaucoup de cas l'arrêt du traitement, et c'est encore plus visible quand le cvb2 est dans la phase 2 ce qui appuie notre hypothèse que le cvb2 est un indicateurs qui engendre l'arrêt du traitement certainement dans le niveau 2.

On sélectionne automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante, en ne gardant que les facteurs qui impactent sur la variable réponse 'Echec thérapeutique', et le résultat de cette sélection du modèle complet donne en sortie le type de traitement (typetrt), le cd4b2 , l'observance (observ) et le cvb2 avec les p-value indiqués selon le tableau suivant.

Facteur	p-value	signif.
typetrt	0.012402	*
observ	0.009662	**
cd4b2	0.009478	**
cvb2	6.119e-07	***

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Echec thérapeutique' AUC = 0.76 qui est un bon résultat pour notre modèle.

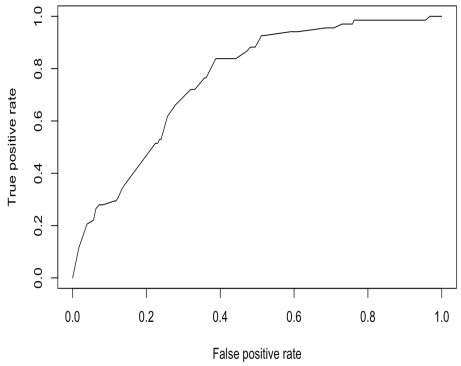

FIGURE 4.12 – Courbe ROC du modèle de la régression logistique pour le motif Échec thérapeutique

4.6.3 Problème d'observance

Dans cette partie nous établirons la régression logistique sur le motif 'Problème d'observance' en se basant sur la nouvelle colonne dont les valeurs sont de type booléen qui prennent la valeur TRUE si la valeur de la colonne motif d'arrêt est bien 'Pb d'observance'.

Le résultat de cette analyse montre que seul le facteur 'observ' (l'observance) qui est significativement lié aux arrêts de traitements dus aux problèmes d'observances avec une p-value $< 5\%$ (0.0490314).

D'un autre côté on remarque que la tranche d'âge 1 et 3 ont un écart significatif par rapport à l'âge 0 avec des p-value d'écart respectivement (p-value $= 0.0283 < 5\%$) et (p-value $= 0.0477 < 5\%$), et le mode de contamination 2 aussi qui correspond aux individus qui se sont contaminé par voie intraveineuse due à la toxicomanie avec une p-value $= 0.0188 < 5\%$.

On sélectionne automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante, en ne gardant que les facteurs qui impactent sur la variable réponse 'Pb d'observance', et le résultat de cette sélection du modèle complet donne en sortie le résultat suivant.

Facteur	p-value	signif.
observ	0.08028	.
conta	0.05611	.
cvb2	6.779e-05	***

On déduit qu'il n'y a pas un facteur qui est lié significativement à ce motif et qui explique l'arrêt due à un problème d'observance, à part le cvb2 qui n'est enfait qu'une mesure selon notre hypothèse de départ.

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Echec thérapeutique' AUC $= 0.6903412$.

4.6.4 Simplification de traitement

Dans cette partie nous établirons la regression logistique sur le motif 'Simplification traitement' en se basant sur la nouvelle colonne dont les valeurs sont de type booléan qui prennent la valeur TRUE si le motif est 'Simplification trt'.

Le résultat de cette analyse montre qu'aucun des facteurs n'est significatif mis à part les deux mesures cd4b2 et cvb2 qui sont très significatives avec les p-value respectives 1.144e-08 et 4.164e-13.

Finalement on essaie de sélectionner automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante, en ne

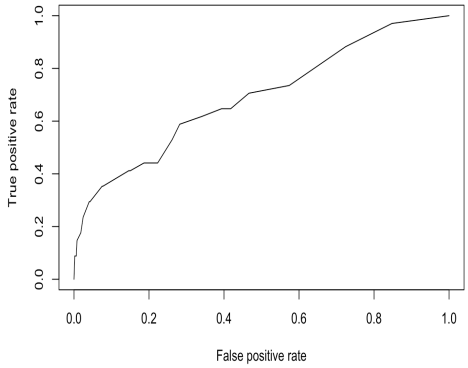

FIGURE 4.13 – Courbe ROC du modèle de la régression logistique pour le motif Échec thérapeutique

gardant que les facteurs qui impactent sur la variable réponse 'Simplification trt', et le résultat de cette selection du modèle complet donne en sortie le résultat suivant.

cd4b1	0.26651	
cd4b2	7.222e-10	***
cvb1	0.04546	*
cvb2	3.486e-13	***

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Simplification trt' AUC =0.792408 qui est un résultat très bon pour notre modèle.

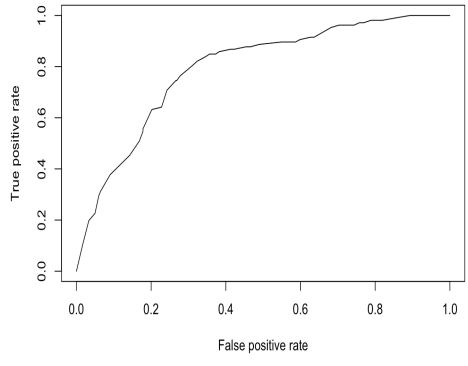

FIGURE 4.14 – Courbe ROC du modèle de la régression logistique pour le motif Simplification Traitement

4.6.5 Autres

Dans cette partie nous établirons la regression logistique sur le motif 'Autres' en se basant sur la nouvelle colonne dont les valeurs sont de type booléan qui prennent la valeur TRUE si le motif est 'Autres'.

Le résultat de cette analyse montre que c'est le facteur âge qui agit essentiellement sur l'arrêt du traitement quand le motif n'est pas identifiable et noté 'Autres', avec une p-value = 3.23e-06 «< 5%, et cela peut s'expliquer par le fait que les individus agées peuvent avoir des complications de santé non cités dans notre modèle qui peuvent agir à l'encontre des attentes du traitement, et conduire à l'arrêt du traitement au final. Le type de traitement aussi a un effet significatif avec une p-value « 5% (p-value = 0.003165) et l'observance avec une p-value « 5% (p-value = 0.001864).

Finalement on essaie de sélectionner automatiquement les variables à partir d'un modèle dit "complet" selon une démarche descendante, en ne gardant que les facteurs qui impactent sur la variable réponse 'Autres', et le résultat de cette sélection du modèle complet donne en sortie le résultat suivant.

```
                        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                      526    101.385
motif_arrets_vih$typetrt  2   2.0409       524     99.344  0.003143 **
motif_arrets_vih$age      3   5.0075       521     94.337 3.176e-06 ***
motif_arrets_vih$observ   1   1.7953       520     92.541  0.001452 **
motif_arrets_vih$cd4b1    2   0.8114       518     91.730  0.101177
```

Avec des effets dont la significativité augmente avec l'augmentation de l'age.

```
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.41967    0.05683   7.385 6.13e-13 ***
motif_arrets_vih$typetrt1 -0.09467    0.04847  -1.953 0.051343 .
motif_arrets_vih$typetrt2  0.03829    0.04428   0.865 0.387601
motif_arrets_vih$age1     -0.14631    0.05287  -2.767 0.005852 **
motif_arrets_vih$age2     -0.22135    0.05735  -3.860 0.000128 ***
motif_arrets_vih$age3     -0.27092    0.06473  -4.186 3.34e-05 ***
motif_arrets_vih$observ1  -0.12872    0.04347  -2.961 0.003206 **
motif_arrets_vih$cd4b11    0.08136    0.04304   1.890 0.059250 .
motif_arrets_vih$cd4b12    0.08195    0.04941   1.659 0.097815 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

La figure suivante représente la courbe ROC qui sert à évaluer notre modèle en calculant l'aire sous la courbe AUC. Pour notre motif 'Autres'

AUC =0.784 qui est un résultat très bon pour notre modèle.

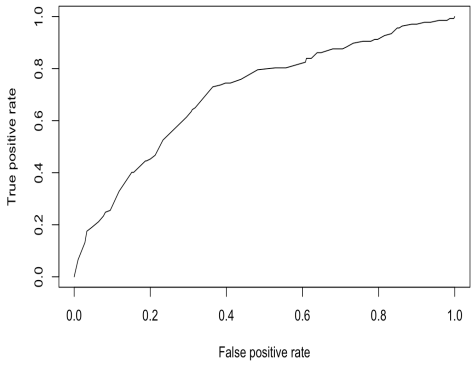

FIGURE 4.15 – Courbe ROC du modèle de la régression logistique pour le motif Échec thérapeutique

Partie 5

Modélisation longitudinale par des méthodes d'analyse de survie

5.1 Définition de la variable arrêt/censure et du délai correspondant

Notre objectif est ici d'étudier la survenue d'évènements au cours du temps. Cela implique d'estimer la probabilité de survenue d'un évènement au cours du temps et également évaluer l'impact d'un facteur sur la survenue de cet évènement. Afin de mettre en place une modélisation longitudinale par des méthodes d'analyse de survie, nous construisons notre jeu de données. On se base sur notre jeu de données nettoyé des valeurs manquantes. Puis nous calculons le nombre de jours entre la date de fin de traitement si elle existe sinon date de point (⇒ date à laquelle l'étude se termine : fin du recueil des informations) et la date de début du traitement. Enfin, nous effectuons sur ce dataset un tri croissant selon le delta de jours. Notre variable évènement sera donc représentée par un arrêt de traitement.

5.2 Survie globale

5.2.1 Estimations par la méthode de Kaplan-Meier

Nous estimons la survie selon la méthode de Kaplan-Meier. Cette méthode estime la survie à chaque survenue d'évênements et tient compte de la date exacte de l'évênement.

```
Call: survfit(formula = Surv(df$duree, df$arret_trt) ~ 1)

      n  events  median 0.95LCL 0.95UCL
    767     527     396     355     479
```

Sur nos 767 patients, 527 ont arrêté leur traitement. Un patient sur deux arrête son traitement au bout de 13 mois, plus précisément 396 jours.

```
    time n.risk n.event  survival  1-survival     std.err lower 95% CI upper 95% CI
1      0    767       1 0.9986962 0.001303781 0.001304632   0.99614578    1.0000000
2     18    727       1 0.9503061 0.049693887 0.008269661   0.93502748    0.9658344
3     30    691       7 0.9003265 0.099673451 0.012050774   0.87931082    0.9218446
4     44    645       3 0.8514149 0.148585091 0.015151756   0.82650227    0.8770785
5     70    601       4 0.7982222 0.201777829 0.018266680   0.77014972    0.8273179
6     99    552       3 0.7480986 0.251901402 0.021152100   0.71771850    0.7797646
7    134    504       1 0.6999959 0.300004146 0.023951960   0.66789397    0.7336407
8    187    463       3 0.6481849 0.351815127 0.027085649   0.61467216    0.6835247
9    257    411       1 0.6007244 0.399275571 0.030149900   0.56625448    0.6372927
10   323    360       2 0.5496928 0.450307235 0.033782254   0.51447528    0.5873210
11   394    322       3 0.5001443 0.499855731 0.037688489   0.46453110    0.5384877
12   519    266       1 0.4500976 0.549902438 0.042189395   0.41437637    0.4888981
13   691    215       1 0.4007846 0.599215375 0.047568765   0.36510724    0.4399483
14   841    175       2 0.3503610 0.649639012 0.054411404   0.31492031    0.3897901
15  1007    133       2 0.2980460 0.701953967 0.063703751   0.26306230    0.3376821
16  1246    100       1 0.2499045 0.750095450 0.074955550   0.21576047    0.2894519
17  1560     67       1 0.1989486 0.801051397 0.091615263   0.16624854    0.2380806
18  1957     32       1 0.1487723 0.851227671 0.121756493   0.11718792    0.1888693
19  2419     17       1 0.1253275 0.874672548 0.150377866   0.09333496    0.1682860
```

On remarque un arrêt rapide des premiers patients, 10% au bout d'un mois. Un arrêt sur quatre a lieu dans les 3 premiers mois. Un patient sur deux va arrête son traitement après 13 mois. Au bout de 40 mois seulement un patient sur quatre continuera son traitement. Un patient sur huit va aller au delà de 80 mois. A travers ces chiffres, on voit clairement que de nombreux patients vont vite arrêter mais ceux qui arrivent à tenir auront tendance à suivre leur traitement beaucoup plus longtemps. Avec le temps, la population étudiée diminue (censures) et engendre une augmentation du risque d'erreurs, mesurable via l'écart-type et qui impacte les intervalles de confiance.

5.2.2 Représentation de la courbe de survie

La courbe représente bien les remarques que nous avons faites précédemment. On retrouve une croissance très forte du nombre d'arrêts au début,

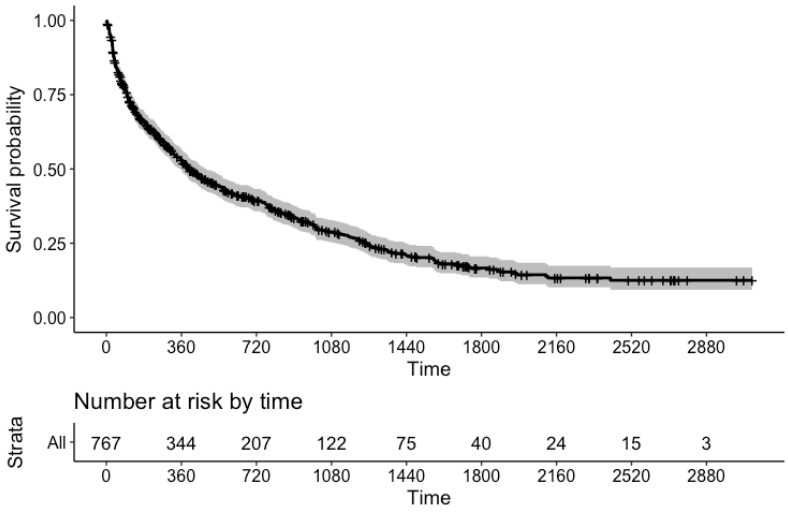

FIGURE 5.1 – Courbe de survie globale

puis un éloignement des arrêts avec le temps. Ceux qui tiennent leur traitement durant les 3-4 premières années auront tendance à tenir très longtemps après.

5.3 Estimation de la survie suivant le type de traitement

5.3.1 Estimation de la survie par la méthode de Kaplan-Meier

Notre but est ici de comparer les 3 types de traitement.

```
Call: survfit(formula = Surv(df$duree, df$arret_trt) ~ df$typetrt)

                n events median 0.95LCL 0.95UCL
df$typetrt=0 268    172    392     322     512
df$typetrt=1 259    148    805     556    1002
df$typetrt=2 240    207    255     181     335
```

Nous n'avons pas exactement le même nombre d'évênements pour chaque type de traitement mais ils sont tout de même assez proches. On note qu'il y a eu plus d'évênements d'arrêts de traitement pour les patients ayant suivi le traitement "autres" (mono, bi, quadrithérapie).

Quelque soit le type de traitement, on voit que la tendance est la même, beaucoup d'arrêts rapidement. Les patients prenant un autre type

de traitement auront une tendance à arrêter plus rapidement au début mais cela s'équilibre avec le temps. A l'inverse les patients prenant le traitement `2IN+INN` vont avoir tendance à moins arrêter leur traitement que l'ensemble sur les premiers mois mais vont avoir une seconde période de ou il va y avoir beaucoup plus d'arrêts entre 30 et 40 mois de traitement. Pour tous les traitements, au bout d'environ un mois, entre 27 et 41 jours, 10% des patients auront arrêté. La perte du premier quart des patients à lieu après à peine 53 jours pour les autres traitements contre 140 jours pour le traitement `2IN+INN`. L'écart se creuse pour le temps d'arrêt médian, environ 255 jours pour les autres traitements, 391 jours pour `2IN+1IP` et 797 jours pour `2IN+INN`. Sur la période entre le 800^{me} et 1200^{me} jour le traitement `2IP+INN` perd 10% de patients en plus que les 2 autres type de traitement. On retrouve 80% d'arrêt après environ 1300 jours pour le traitements 2IN+1IP et autres, et 1600 jours pour `2IP+INN`. Au final quelque soit le type de traitement on se retrouve avec un taux de survie assez proche.

5.3.2 Représentation graphique

La fonction de survie représente la probabilité pour qu'un patient n'est pas arrêté son traitement après un délai t à compter d'un instant de référence, ou encore la proportion de personne poursuivant leur traitement après un délai t. Ici, le délai t, correspond à la date de début de traitement pour chaque patient.

On représente la survie de Kaplan-Meier avec en abscisse la durée de suivi et en ordonnée l'estimation de la fonction de survie. On note que le taux d'arrêt de traitement est quasiment identique pour les 3 traitements mais sur les 3 premières années l'arrêt est plus tardif pour le traitement 1 (`2IN+INN`).

5.4 Test du Log-Rank

Pour montrer que le type de traitement a un lien avec la survie, nous mettons en oeuvre le test du *Log-Rank*. C'est un test de comparaison entre courbes de survie. Si 2 courbes de survie sont égales, le nombre d'arrêts de traitement devrait survenir au même rythme dans les 3 groupes. L'idée de base du test du Log-Rank est de comparer le nombre d'évènements observés et le nombre d'évènements attendus sous l'hypothèse nulle d'égalité de la survie dans les 3 groupes.

L'hyptohèse nulle est l'égalité des probabilités de survie dans les 3

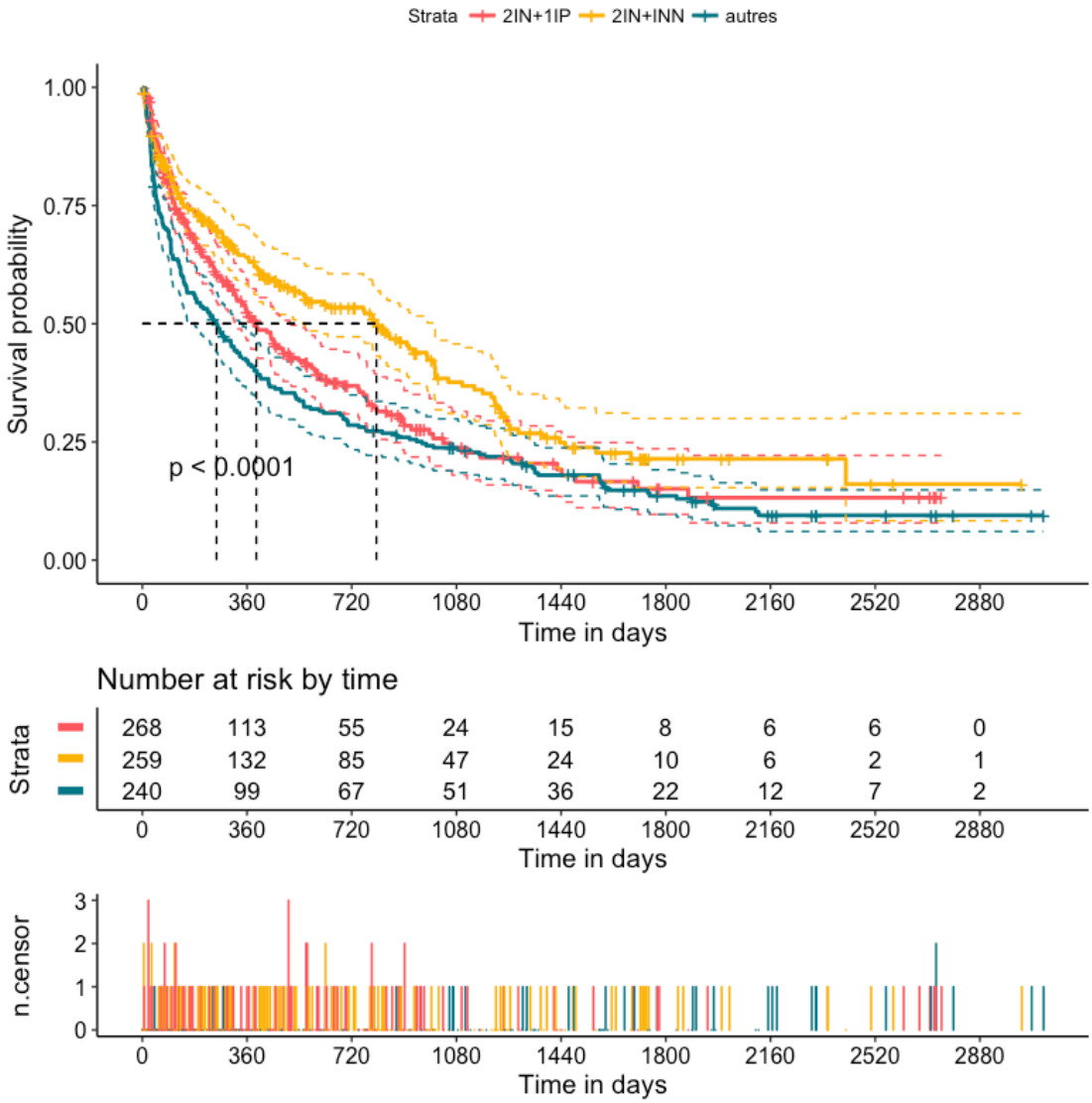

FIGURE 5.2 – Courbe de survie en fonction du type de traitement

groupes :

$$H_0 : S_{trt\_0}(t) = S_{trt\_1}(t) = S_{trt\_2}(t)$$

On cherche si la différence observée entre les courbes de survie dans les groupes `trt_0`, `trt_1` et `trt_2` permet de rejeter l'hypothèse nulle.

```
Call:
survdiff(formula = Surv(df$duree, df$arret_trt) ~ df$typetrt)
```

```
                N Observed Expected (O-E)^2/E (O-E)^2/V
df$typetrt=0 268      172      168    0.0899     0.134
df$typetrt=1 259      148      191    9.5390    15.033
df$typetrt=2 240      207      168    8.9281    13.353

 Chisq= 18.7  on 2 degrees of freedom, p= 8.54e-05
```

Ici on rejette H_0 l'hypothèse d'égalité des courbes de survie et on conclut à une différence significative des délais de survenue de l'évènement dans les 3 groupes. On note que la plus grosse contribution vient du traitement 1 suivi du traitement 2. Cela signifie qu'il y a de gros écarts entre les effectifs observés et les effectifs attendus (théoriques). On conclut que la survenue d'un arrêt de traitement est plus tardive lorsque les patients prennent le traitement 1. Cela confirme ce que nous avons pu voir sur le graphique représentant la fonction de survie.

On décide de mettre en oeuvre le test du log-rank pour tous les facteurs possibles et de ne retenir que les facteurs qui ont des p-values significatives. On ne prend pas les variables qui concernent la fin du traitement ou du suivi comme `cd4b2` ou `cvb2`.

```
[1] "typetrt: 1e-04"
[1] "sexe: 0.061"
[1] "sida: 3e-04"
[1] "observ: 0.15"
[1] "conta: 0.0049"
[1] "cd4b1: 0.1939"
[1] "cvb1: 0.793"
[1] "age: 0.2549"
```

Ici nous constatons que seules les facteurs 'typetrt', 'sida', et 'conta' sont significatif au seuil $\alpha = 5\%$.

5.5 Modèles de Cox

Le modèle de Cox permet d'exprimer le risque instantané de survenue de l'évènement en fonction de l'instant t et des variables explicatives X^j. Il permet donc d'étudier l'effet éventuel de plusieurs covariables sur la survie. En effectuant le test du Log-Rank sur chaque facteur nous en avons conclu que nous conserverions seulement 3 covariables explicatives.

```
Analysis of Deviance Table
```

```
 Cox model: response is Surv(df$duree, df$arret_trt)
Terms added sequentially (first to last)

            loglik  Chisq Df Pr(>|Chi|)
NULL       -3104.0
df$typetrt -3094.6 18.952  2  7.667e-05 ***
df$sida    -3089.4 10.378  1   0.001275 **
df$conta   -3083.8 11.161  3   0.010889 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

On note ici que les facteurs `typetrt`, `sida` sont très significatifs (p-value $\ll 0.01$) sur le risque instantané d'arrêt de traitement. La covariable `conta` est elle aussi significative.

```
Call:
coxph(formula = Surv(df$duree, df$arret_trt) ~ df$typetrt + df$sida +
    df$conta)

  n= 767, number of events= 527

               coef exp(coef) se(coef)      z Pr(>|z|)
df$typetrt1 -0.22246   0.80055  0.11408 -1.950  0.05117 .
df$typetrt2  0.23369   1.26326  0.10564  2.212  0.02695 *
df$sida1     0.35020   1.41936  0.11434  3.063  0.00219 **
df$conta1    0.01520   1.01531  0.09627  0.158  0.87455
df$conta2    0.34395   1.41051  0.17088  2.013  0.04413 *
df$conta3   -0.39986   0.67041  0.17406 -2.297  0.02161 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

            exp(coef) exp(-coef) lower .95 upper .95
df$typetrt1    0.8005     1.2491    0.6402     1.001
df$typetrt2    1.2633     0.7916    1.0270     1.554
df$sida1       1.4194     0.7045    1.1344     1.776
df$conta1      1.0153     0.9849    0.8407     1.226
df$conta2      1.4105     0.7090    1.0091     1.972
df$conta3      0.6704     1.4916    0.4766     0.943

Concordance= 0.583  (se = 0.014 )
Rsquare= 0.051   (max possible= 1 )
Likelihood ratio test= 40.49  on 6 df,   p=3.648e-07
Wald test            = 40.96  on 6 df,   p=2.949e-07
Score (logrank) test = 41.72  on 6 df,   p=2.085e-07
```

Nous allons ici décrire les facteurs et leurs modalités significatives en commençant par ceux les plus significatifs. Pour chaque *p-value* liée à la statistique Z, on va regarder si elle est inférieure à 5% pour vérifier que l'hypothèse que $\beta_{coeff} = 0$ puisse être rejetée. Si ce n'est pas le cas, on ne pourra pas rejeter H_0 au seuil $\alpha = 5\%$ et la modalité du facteur ne sera pas considérée.

Tout d'abord, on note que $e^{\beta_{sida_1}}$ est supérieur à 1, ceci indique que des valeurs élevées de la $sida\_1$ sont associées à un risque instantané d'arrêt de traitement plus élevé. Le risque instantané d'arrêt est multiplié par 1.41 lorsque que l'on compare le patient a le sida. Concernant le type de traitement, le traitement 1 n'est pas significatif, mais le traitement 2 l'est. Nous pouvons donc interpréter l'exponentielle du coefficient associé. Comme cette valeur est supérieure à 1, cela indique que des valeurs élevées du traitement "autres" sont associées à un risque instantané d'arrêt de traitement lui aussi plus élevé. Ce constat est similaire pour la modalité 2 du facteur contamination. Le risque instantané d'arrêt de traitement est multiplié par 1.41 lorsque le patient est contaminé par voie intraveineuse par rapport à un patient contaminé de manière hétérosexuelle. Lorsque les patients sont contaminés par un autre mode (transfusion, hémophilie), le risque instantané d'arrêt de traitement est plus faible ($\times 0.67$).

Nous cherchons à voir si l'hypothèse des risques proportionnels est raisonnable ou non. Nous avons mis en évidence un effet : cet effet est il constant au cours du temps ?

Nous testons le modèle des risques proportionnels de Cox :

```
                rho   chisq       p
df$typetrt1 -0.02964  0.4629 0.49628
df$typetrt2 -0.11212  6.7856 0.00919
df$sida1     0.00627  0.0210 0.88485
df$conta1    0.05926  1.8543 0.17328
df$conta2    0.01159  0.0734 0.78650
df$conta3   -0.01942  0.1992 0.65535
GLOBAL             NA 10.9207 0.09086
```

Cette matrice contient une ligne pour chaque variable et une ligne pour le test global. La matrice contient le coefficient de corrélation entre la durée de survie transformée et les résidus de Schoenfeld, un $\mathcal{X}^2$, et une p-value. Pour le test global il n'y a pas de corrélation appropriée, c'est pourquoi nous avons un "NA". Le test global de validité du modèle des risques proportionnels de Cox conduit à rejeter H_0 : certaines covariables on un effet dépendant du temps. C'est le type de traitement 2 qui conduit à rejeter cette hypothèse (p-value $< 5\%$).

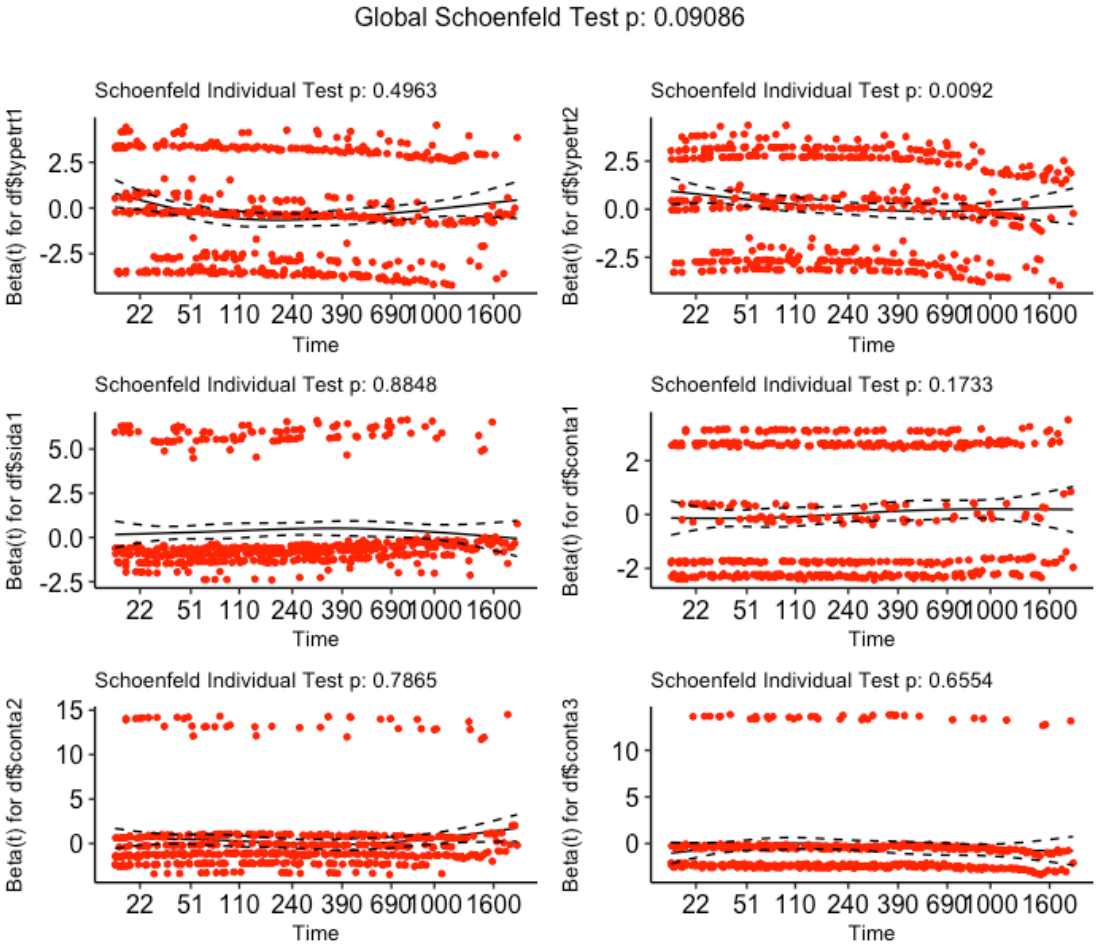

FIGURE 5.3 – Résidus au cours du temps

En abscisse nous avons la durée de suivi et en ordonnée nous avons les valeurs des résidus de Schoenfeld.

Tout éloignement de lhorizontale traduit un effet dépendant du temps. Leffet du traitement 2 décroit linéairement avec le temps, il n'est donc pas toujours constant. Les autres effets ont lair plutôt fixes.

Nous avons essayé ensuite de mettre en place un modèle de sélection descendante pour choisir les facteurs les plus significatifs et cette méthode de sélection n'a retiré aucun des 3 facteurs déjà présents.

Conclusion

Après traitement des valeurs manquantes, nous avons fais une analyse qui vise à savoir quels sont les facteurs qui participent à larrêt du traitement chez les patients qui ont suivi cette expérience. Et selon les différents test que nous avons appliqué sur les différents facteurs on conclut que le type de traitement, le sida et le mode de contamination sont les facteurs qui influent dune façon majeure sur notre évènement, en précisant que le type de traitement 2 qui a été appliqué sur les patients en 2004 augmente la probabilité darrêt chez les individus que les autres traitements, dun côté et dun autre côté si lindividu passe de la phase séropositive à la phase sida il a encore plus de chance darrêter que les autres, également pour les modes de contamination toxico par voie intraveineuse et autres (transfusé, hémophile, ...). Après une analyse par rapports aux motifs darrêts, lanalyse est plus pointue et a donné les résultats suivants :

- le motif 'Intolérance/ Toxicité' est lié par les facteurs type du traitement, la tranche dâge et lobservance
- le motif 'Echec thérapeutique' cest le type du traitement, lobservance le cd4b2 et le cvb2 qui ont une liaison avec larrêt ou non du traitement chez les individus
- Le motif 'Problèmes d'observances' n'est pas lié à l'observance comme on l'attend, mais il est lié juste aux taux de lymphocytes et la charge virale chez les individus à la fin du traitement.
- Le motif 'Simplification traitement' est pareil au motif précédent, en plus du taux de lymphocytes en début de traitement.
- Le motif 'Autres' regroupe l'ensemble des arrêts qui n'ont pas été expliqués par les experts, et c'est généralement due à la tranche d'âge qui favorise l'arrêt pour les plus âgés, le type du traitement et l'observance.

D'un autre côté, on conclut que la dépendance est très forte entre les événement et les facteurs cd4b2 et cvb2, qui sont la charge virale et le taux de lymphocytes à l'arrêt du traitement, cela relève la grande possibilité qu'il soient des indicateurs qui poussent l'expert à arrêter le traitement chez les patients. Alors même si on trouve une grande dépendance entre ces facteurs et notre événement, nous les considérons pas comme des variables explicatives pour notre modèle.

L'analyse longitudinale par des méthodes d'analyse de survie ont été très efficaces. Nous avons réussi a obtenir des graphiques contenant beaucoup d'informations tout en restant très pertinents. Ces visualisations nous ont permis de mieux concevoir les courbes de survies. Nous avons mis en place également différents modèles de Cox, toutefois sans prendre en compte les interactions entre covariables explicatives. Cette méthode nous a permis d'exprimer un risque instantané de survenue d'un évênement en fonction du temps de variables explicatives. Nous avons construit un modèle robuste composé des facteurs liés au sida, au type de traitement et au mode de conta-

mination. Cela a permis de mettre en évidence l'influence de ces facteurs sur la survenue des arrêts au cours du temps.