

Université Paris Saclay

Année universitaire 2024-2025

**Master 2 méthodologie et statistiques en
recherche biomédicale**

Devoir 2 (TP Hamilton)

Alexandre DEGACHI
(groupe 2)

Table des matières

I – Préparation des données des variables	3
A – Fusion des tableaux.....	3
B – Valeurs manquantes et aberrantes.....	3
II – Question 1 : validation de l'échelle de Hamilton à J0 et J56.....	4
A – Distribution des niveaux de réponse pour chaque item	4
1 – Echelle de Hamilton à J0	4
2 – Echelle de Hamilton à J56	6
3 – Matrices de corrélation à J0 et J56.....	8
B – Vérification de l'unidimensionnalité de l'échelle de Hamilton	9
C – Calcul du coefficient de Cronbach.....	12
D – Validité convergente et divergente	12
III – Question 2 : comparaison du score brut de Hamilton entre les groupes	14
A – Score brut de Hamilton par groupe.....	14
1 – Score brut de Hamilton à J0.....	14
2 – Evolution du score brut de Hamilton dans chaque groupe	14
B – Comparaison du score de Hamilton entre J0 et J56 dans chaque groupe, après imputation des données manquantes selon la méthode LOCF	15
C – Comparaison du score de Hamilton entre J0 et J56 dans chaque groupe, après gestion des données manquantes par un modèle linéaire mixte.....	18
IV – Question 3 : comparaison de la réponse au traitement par une analyse de survie	21
V – Script R.....	23

I – Préparation des données des variables

A – Fusion des tableaux

Nous disposons des données de 146 patients atteints de dépression répartis en 2 groupes (0 et 1). 2 fichiers fournissent les résultats d'une auto-évaluation (SCL90) et d'une hétéro-évaluation (Hamilton) de la dépression.

L'évaluation de ces 2 scores a été réalisée de façon répétée à J0, J4, J7, J14, J21, J28, J42, J56.

Nous modifions les variables erronément considérées comme numériques par R en variables catégorielles (notamment NUMERO et GROUPE).

Nous fusionnons les colonnes HAMD16A et HAMD16B du fichier hdrs en une colonne HAMD16 (le tableau de contingence de ces 2 variables permet de voir que si la valeur est manquante dans une colonne, elle est non manquante dans l'autre, à l'exception d'une ligne pour laquelle la valeur manque dans les 2 colonnes).

Enfin nous fusionnons fusionner les tableaux Hamilton et SCL90 avec le tableau donnant le groupe de chaque patient.

B – Valeurs manquantes et aberrantes

Dans le fichier Hamilton, une seule donnée manque pour chacun des items, correspondant toutes à la visite J7 de l'individu 128.

Dans le fichier SCL90, toutes les questions comportent des données manquantes, jusqu'à 1,45% de données manquantes (aux questions 69 et 84).

Il n'y a aucun doublon (défini par le même numéro et la même visite sur 2 lignes différentes) dans les tableaux Hamilton et SCL90.

Il n'y a pas de valeur aberrante dans le tableau Hamilton : les maximums de chaque item sont conformes aux valeurs attendues.

Dans le tableau SCL90, certaines réponses aux items sont supérieures à 4, nous les remplaçons par des données manquantes.

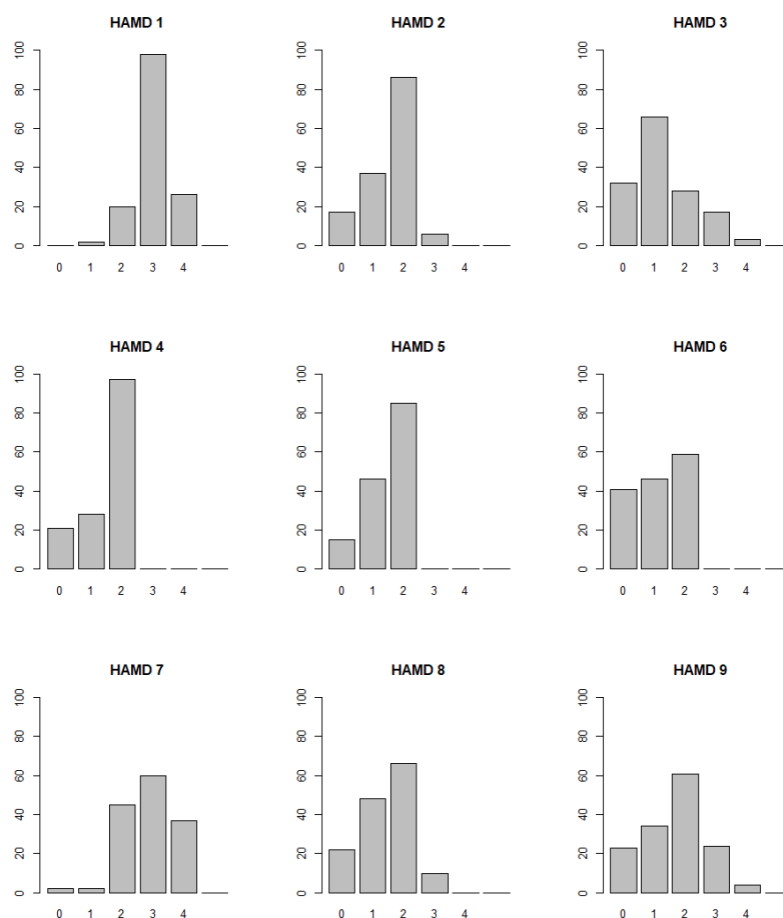
Naturellement, le pourcentage de valeurs manquantes a augmenté mais reste à 1,55% (pour la question 84).

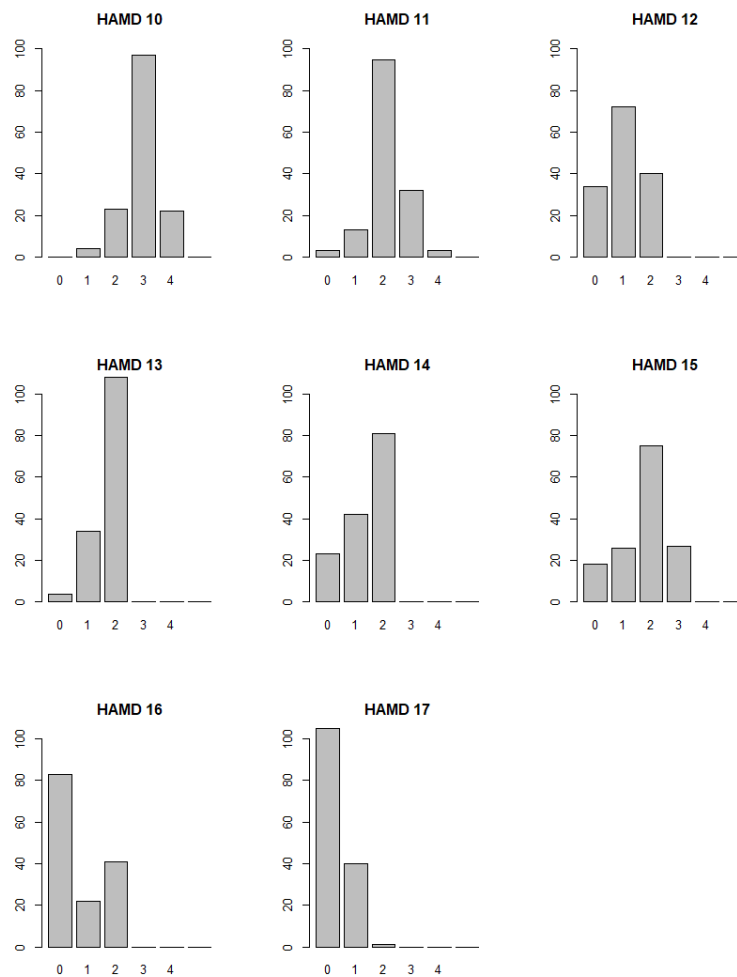
II – Question 1 : validation de l'échelle de Hamilton à J0 et J56

A – Distribution des niveaux de réponse pour chaque item

1 – Echelle de Hamilton à J0

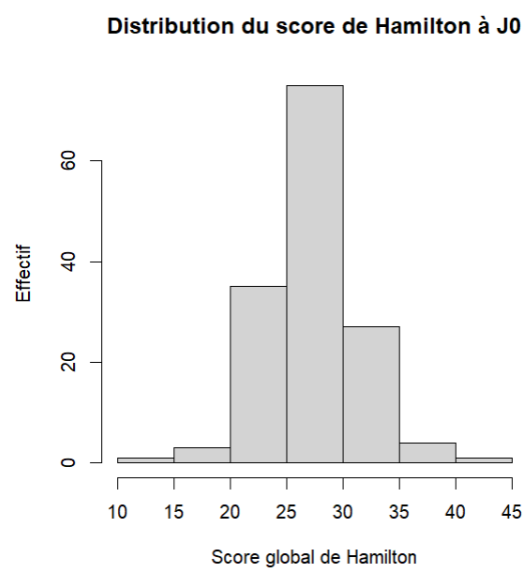
Figure 1 : diagrammes en barres représentant la distribution des réponses à J0 pour chaque item de l'échelle de Hamilton (avec données manquantes)





A la visite J0, on observe un effet plancher pour les items 16 et 17.

Figure 2 : histogramme de la distribution du score brut de Hamilton à J0

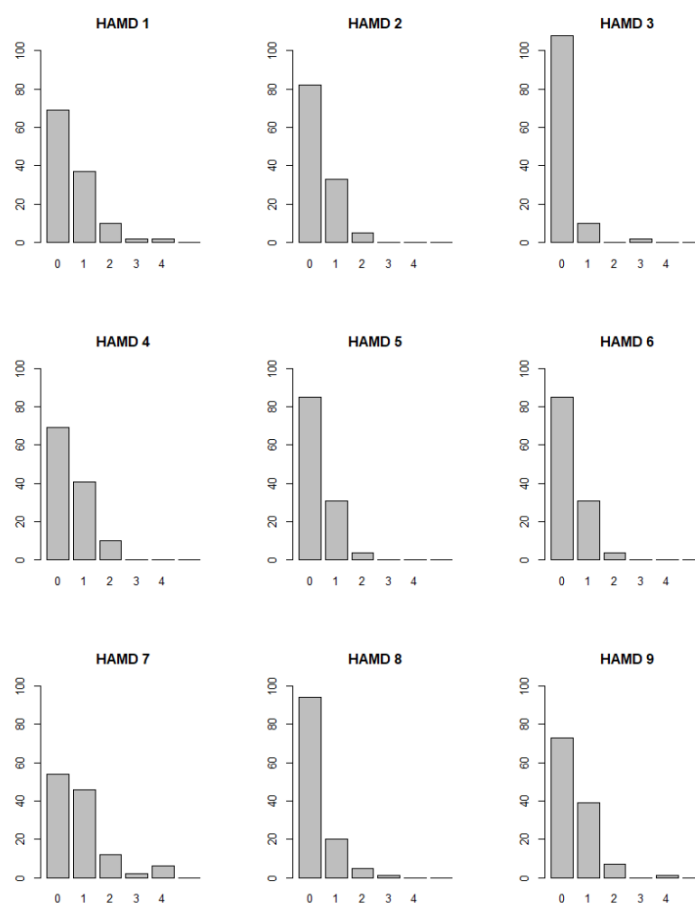


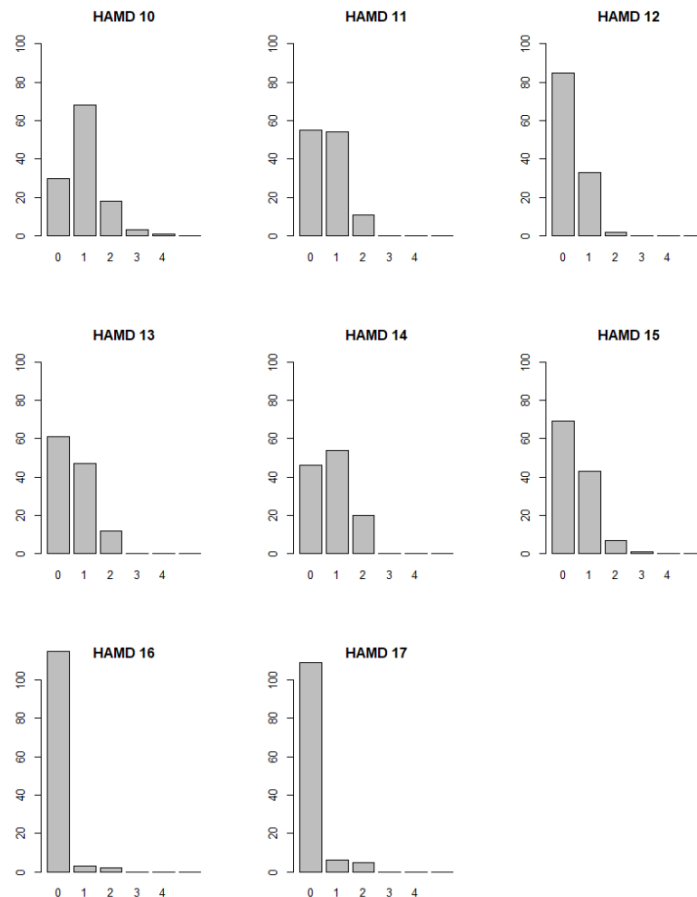
	Effectif	Moyenne	Ecart-type	Minimum	Maximum	1 ^{er} quartile	Médiane	3 ^e quartile
Score brut J0	146	27,9	4,2	13	44	25	27	30

Nous disposons des données de 146 patients atteints de dépression répartis en 2 groupes (0 et 1).

2 – Echelle de Hamilton à J56

Figure 3 : diagrammes en barres représentant la distribution des réponses à J56 pour chaque item de l'échelle de Hamilton (avec données manquantes)

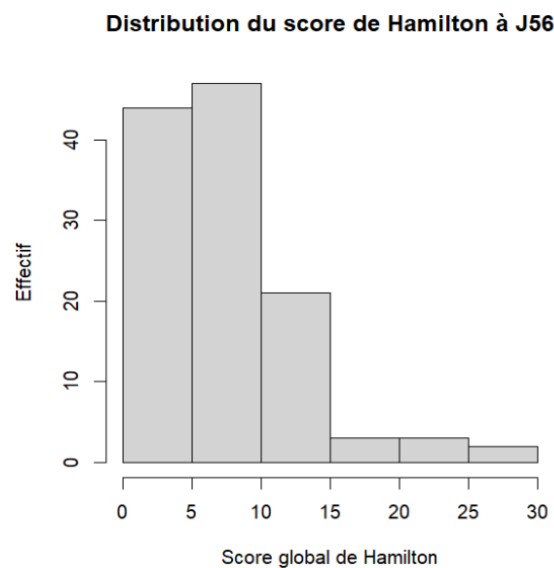




A la visite J56, on observe un effet plancher pour de nombreux items (tous sauf 10 et 14).

On peut supposer que la symptomatologie dépressive diminue avec le temps. Ainsi l'échelle de Hamilton manque de capacité à discriminer des symptomatologies dépressives légères ou nulles (effet plancher pour de nombreux items). Cette échelle de Hamilton semble mieux adaptée lorsque la symptomatologie est plus floride.

Figure 4 : histogramme de la distribution du score brut de Hamilton à J56



	Effectif	Moyenne	Ecart-type	Minimum	Maximum	1 ^{er} quartile	Médiane	3 ^e quartile
Score brut J56	120	7,8	5,5	0	27	4	7,5	10

3 – Matrices de corrélation à J0 et J56

Les corrélations entre les items sont globalement faibles à J0.

Les corrélations entre les items sont plus élevées à J56, cela est certainement à l'effet plancher observé pour la plupart des items (tous les items tendent vers 0).

Figure 5 : matrice de corrélation entre les items du score de Hamilton à J0

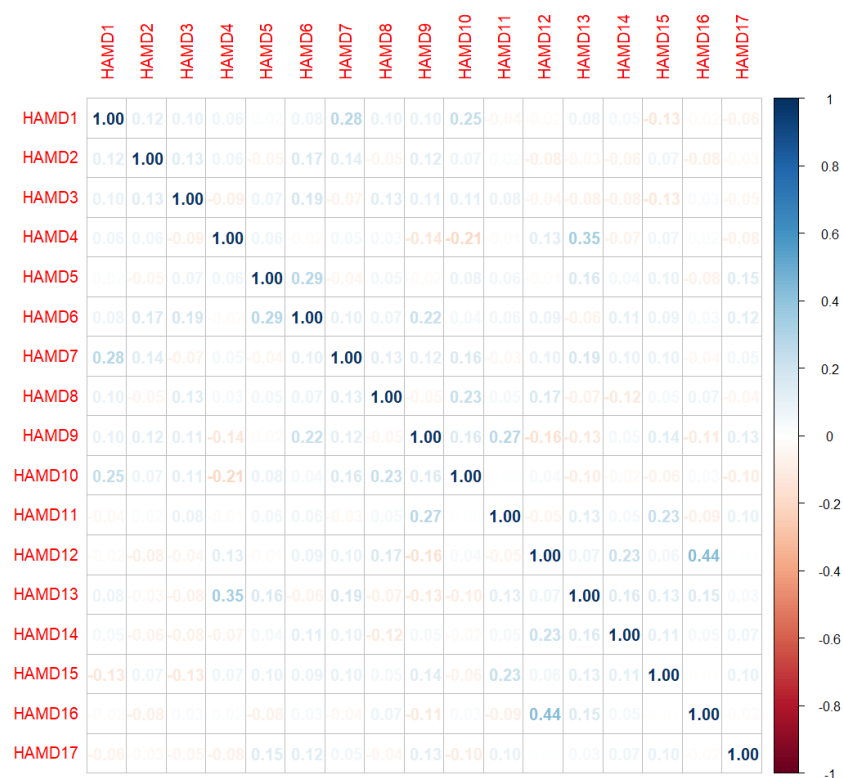
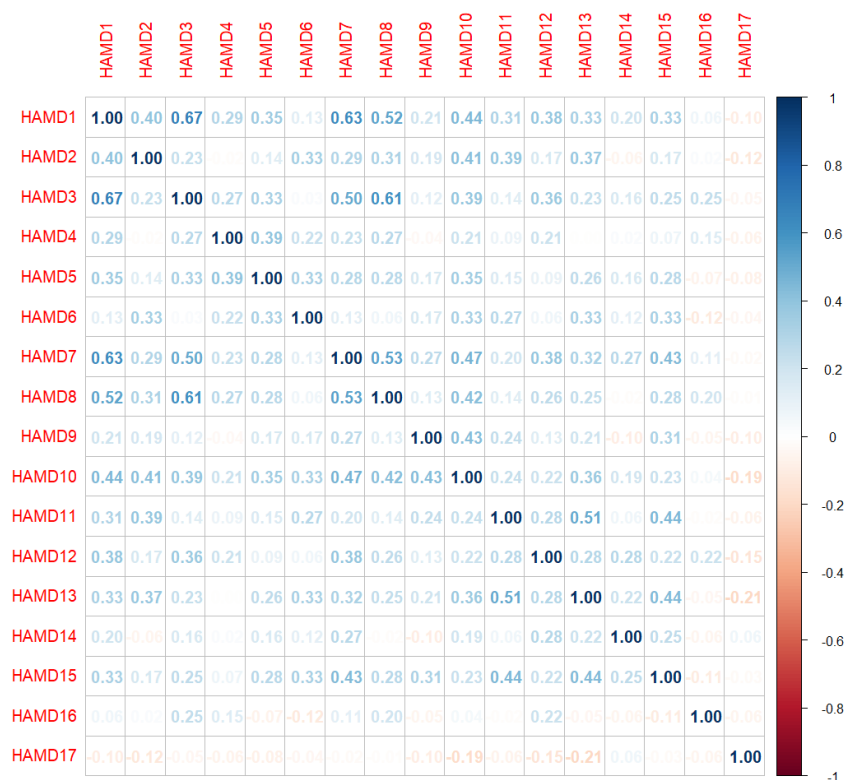


Figure 6 : matrice de corrélation entre les items du score de Hamilton à J56



B – Vérification de l'unidimensionnalité de l'échelle de Hamilton

Figure 7 : diagramme des valeurs propres de l'analyse en composantes principales portant sur les items du score de Hamilton à J0

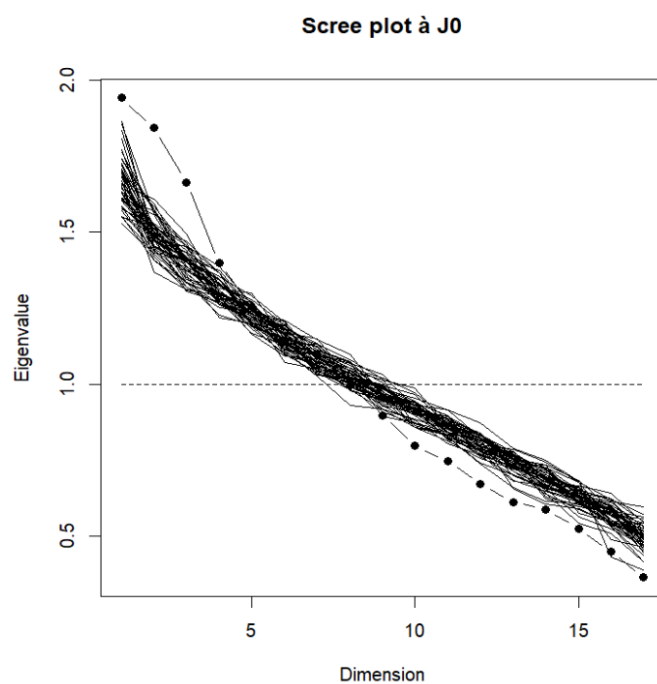
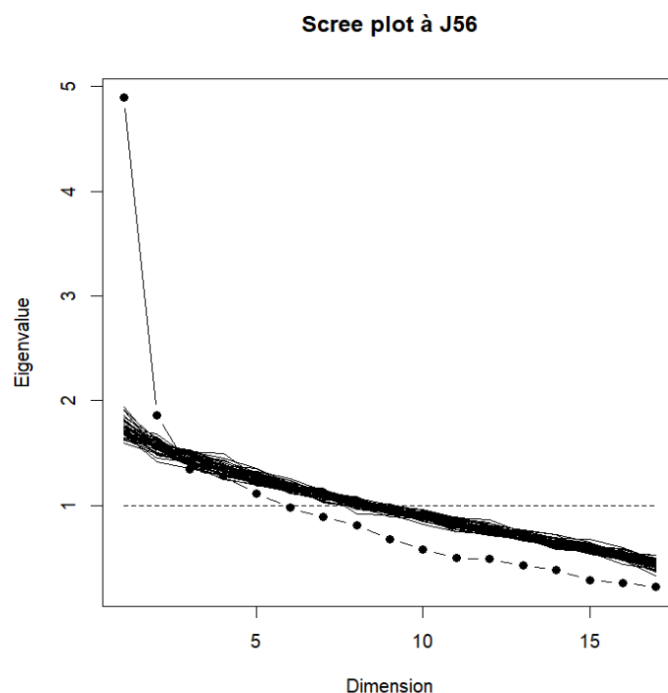


Figure 8 : diagramme des valeurs propres de l'analyse en composantes principales portant sur les items du score de Hamilton à J56



A J0, l'échelle de Hamilton ne peut pas être considérée comme unidimensionnelle : 3 dimensions ont une valeur propre supérieure à celle obtenue sur des données simulées. Par contre, à J56, l'échelle de Hamilton semble bien être unidimensionnelle : la première composante principale a une valeur propre nettement supérieure à celle des autres composantes.

Nous réalisons une analyse factorielle avec rotation varimax en ajoutant 3 variables latentes pour le score à J0 et une variable latente pour le score à J56.

Table 1 : contribution des facteurs à la variance de la réponse à chaque item du score de Hamilton évalué à J0 (analyse factorielle avec rotation varimax à 3 facteurs)

	Facteur 1	Facteur 2	Facteur 3
HAMD1 : humeur dépressive	0.006	0.028	0.263
HAMD2 : sentiments de culpabilité	-0.057	-0.076	0.240
HAMD3 : suicide	-0.021	-0.120	0.207
HAMD4 : insomnie du début de nuit	0.110	0.359	-0.050
HAMD5 : insomnie du milieu de nuit	0.004	0.118	0.240
HAMD6 : insomnie du matin	0.132	-0.155	0.427
HAMD7 : travail et activités	0.124	0.121	0.320
HAMD8 : ralentissement	0.183	-0.105	0.101
HAMD9 : agitation	-0.107	-0.228	0.501
HAMD10 : anxiété psychique	0.066	-0.152	0.255
HAMD11 : anxiété somatique	-0.029	0.070	0.316

HAMD12 : symptômes somatiques gastro-intestinaux	0.992	0.044	-0.091
HAMD13 : symptômes somatiques généraux	0.046	0.977	0.194
HAMD14 : symptômes génitaux	0.238	0.120	0.147
HAMD15 : hypocondrie	0.076	0.089	0.226
HAMD16 : perte de poids	0.424	0.155	-0.100
HAMD17 : prise de conscience	0.022	-0.004	0.176

A J0, l'analyse factorielle à 3 facteurs met en évidence 3 variables expliquant 21,9% de la variance des réponses aux items.

Le facteur 1 est le plus associé avec les items 8, 12, 14, 16 (principalement des symptômes somatiques).

Le facteur 2 est le plus associé avec les items 4 et 13 (relatifs à de l'asthénie).

Le facteur 3 est le plus associé avec les items restants (symptômes dépressifs psychiatriques proprement dits).

L'échelle de Hamilton semble donc être tri-dimensionnelle à J0.

Table 2 : contribution du facteur latent à la variance de la réponse à chaque item du score de Hamilton évalué à J56 (analyse factorielle avec rotation varimax à un facteur)

	Facteur 1
HAMD1 : humeur dépressive	0.802
HAMD2 : sentiments de culpabilité	0.468
HAMD3 : suicide	0.706
HAMD4 : insomnie du début de nuit	0.333
HAMD5 : insomnie du milieu de nuit	0.456
HAMD6 : insomnie du matin	0.282
HAMD7 : travail et activités	0.745
HAMD8 : ralentissement	0.657
HAMD9 : agitation	0.328
HAMD10 : anxiété psychique	0.624
HAMD11 : anxiété somatique	0.401
HAMD12 : symptômes somatiques gastro-intestinaux	0.471
HAMD13 : symptômes somatiques généraux	0.495
HAMD14 : symptômes génitaux	0.252
HAMD15 : hypocondrie	0.494
HAMD16 : perte de poids	0.125
HAMD17 : prise de conscience	-0.132

En dehors de l'item 17, tous les items semblent bien relever de la même dimension.

A J56, il existe bien un facteur commun aux 16 items sur 17 qui explique une partie de la variance de la réponse à l'item. L'échelle de Hamilton est donc unidimensionnelle à J56.

C – Calcul du coefficient de Cronbach

Le coefficient α de Cronbach permet ici d'estimer la fidélité de l'échelle au concept de dépression (il s'agit en quelque sorte du carré du coefficient de corrélation entre les réponses aux 17 items de l'échelle de Hamilton et le score « réel » de dépression pris comme le score qu'obtiendrait un patient s'il pouvait répondre à tous les items possibles relatifs à la dépression).

Nous estimons l'intervalle de coefficient autour du coefficient α par un bootstrap (l'utilisation d'un bootstrap est licite avec plus d'une centaine d'observations supposées indépendantes).

A la visite J0 : $\alpha = 0,46$; intervalle de confiance à 95% = [0,22 ; 0,60].

A la visite J56 : $\alpha = 0,82$; intervalle de confiance à 95% = [0,75 ; 0,86].

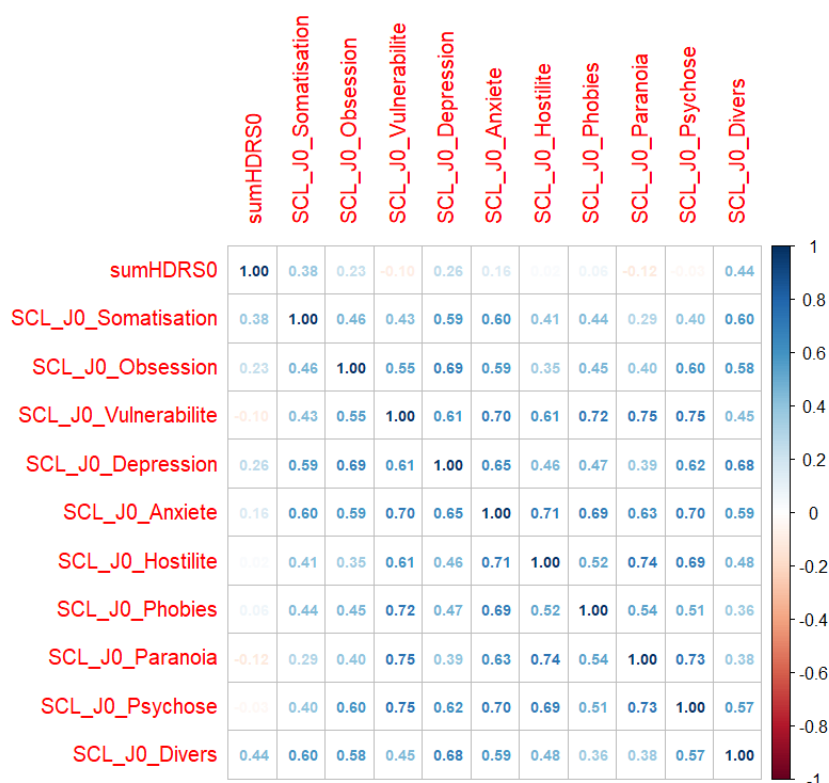
La fidélité de l'échelle est moyenne à J0, elle est très bonne à J56.

D – Validité convergente et divergente

Nous allons utiliser l'échelle SCL90 que les mêmes patients ont également passée.

Nous décomposons l'échelle SCL90 en 10 sous-scores correspondant chacun à une dimension différente.

Figure 9 : matrice de corrélation entre le score brut de l'échelle de Hamilton et les sous-dimensions de l'échelle SCL90, évaluées chez des patients à J0



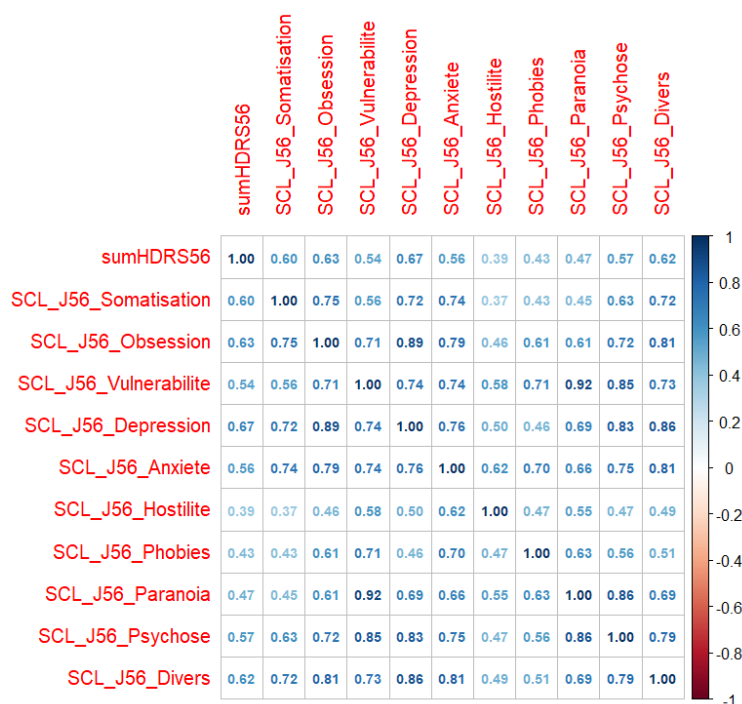
A la visite J0, les corrélations sont moyennes à faibles entre le score brut de Hamilton et chacune des sous-dimensions de l'échelle SCL90.

La corrélation est la plus élevée entre score brut de Hamilton et sous-dimensions « symptômes divers » et « somatisation » de la SCL90. La corrélation entre Hamilton brut et sous-dimension « dépression » de la SCL90 est de 0,26.

Au contraire, il y a une corrélation inverse entre Hamilton brut et les sous-dimensions de vulnérabilité, paranoïa et psychose.

Au total, à J0, il n'y a pas d'argument fort en faveur d'une validité convergente ni divergente de l'échelle de Hamilton.

Figure 10 : matrice de corrélation entre le score brut de l'échelle de Hamilton et les sous-dimensions de l'échelle SCL90, évaluées chez des patients à J56



A la visite J56, les corrélations sont bonnes et relativement homogènes entre le score brut de Hamilton et chacune des sous-dimensions de l'échelle SCL90.

La corrélation est la plus élevée entre score brut de Hamilton et la sous-dimension « dépression » de la SCL90 (0,67).

A nouveau, à J56, il n'y a pas d'argument fort en faveur d'une validité convergente ni divergente de l'échelle de Hamilton.

III – Question 2 : comparaison du score brut de Hamilton entre les groupes

Nous transformons le fichier long contenant le score brut de Hamilton à chaque visite en fichier large (une ligne par patient).

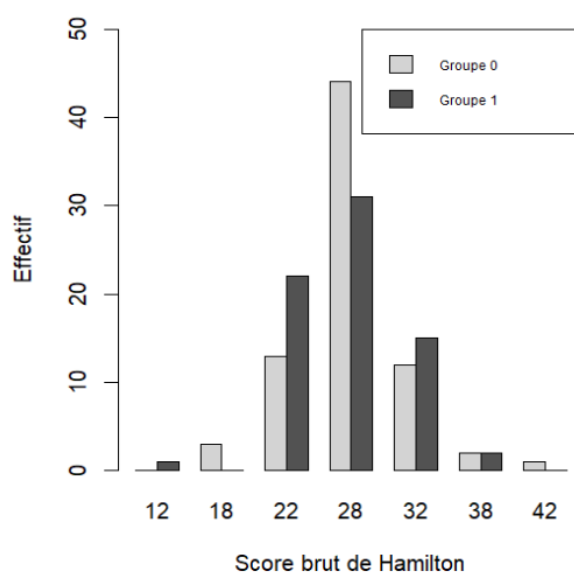
A – Score brut de Hamilton par groupe

1 – Score brut de Hamilton à J0

Table 3 : statistiques descriptives du score de Hamilton à J0 par groupe de patients

	Effectif	Moyenne	Ecart-type	Minimum	Maximum	1 ^{er} quartile	Médiane	3 ^e quartile
Groupe 0	75	28,1	4,4	16	44	26	27	30
Groupe 1	71	27,8	4,0	13	38	25	27	30

Figure 11 : histogramme du score brut de Hamilton à J0 dans chaque groupe de patients



Les distributions du score de Hamilton sont similaires entre les deux groupes à J0.

2 – Evolution du score brut de Hamilton dans chaque groupe

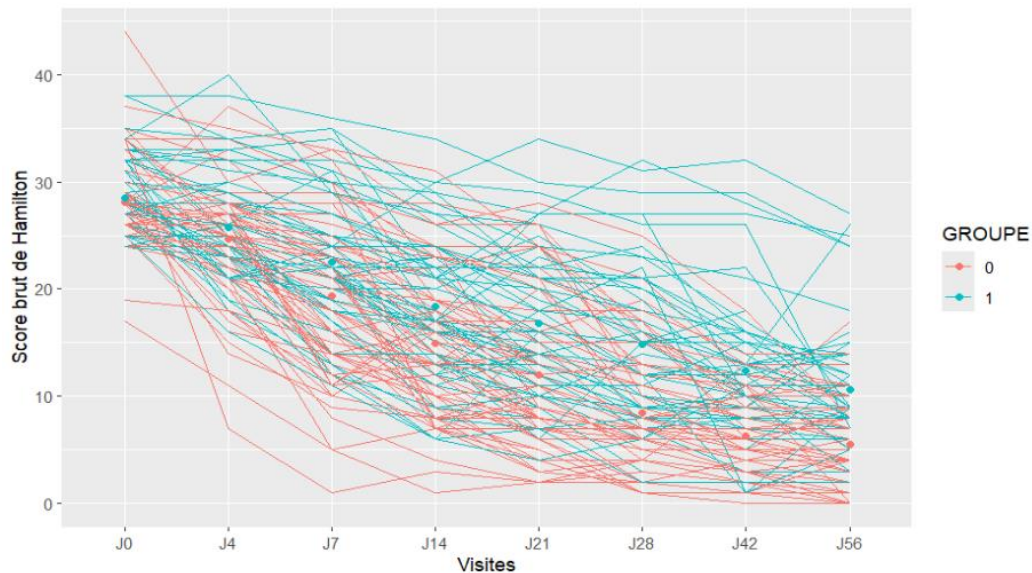
Table 4 : nombre d'observations disponibles à chaque visite, par groupe de patients

	Visite J0	Visite J4	Visite J7	Visite J14	Visite J21	Visite J28	Visite J42	Visite J56
Groupe 0	75	72	71	70	69	68	65	64
Groupe 1	71	67	65	63	60	59	57	56

A la visite J56, les visites manquent pour 11 patients sur 75 dans le groupe 0 (14,7%), contre 15 patients sur 56 dans le groupe 1 (26,8%).

Il y a plus de perdus de vue dans le groupe 1, ce qui est susceptible de biaiser les résultats.

Figure 12 : diagramme en fagot de la trajectoire individuelle du score de Hamilton durant l'étude, selon le groupe des patients



Le score de Hamilton baisse dans chaque groupe. Il semble que la trajectoire de quelques patients, plus dans le groupe 1 que dans le groupe 0, soit marquée par un score stable voire en augmentation aux dernières visites.

B – Comparaison du score de Hamilton entre J0 et J56 dans chaque groupe, après imputation des données manquantes selon la méthode LOCF

Nous allons comparer la différence moyenne du score brut de Hamilton entre J0 et J56 entre les 2 groupes.

Nous imputons les données manquantes aux visites postérieures à J0 par la méthode LOCF (last observation carried forward).

Figure 13 : diagramme en fagot de la trajectoire individuelle du score de Hamilton durant l'étude, selon le groupe des patients, après imputation des données manquantes par méthode LOCF

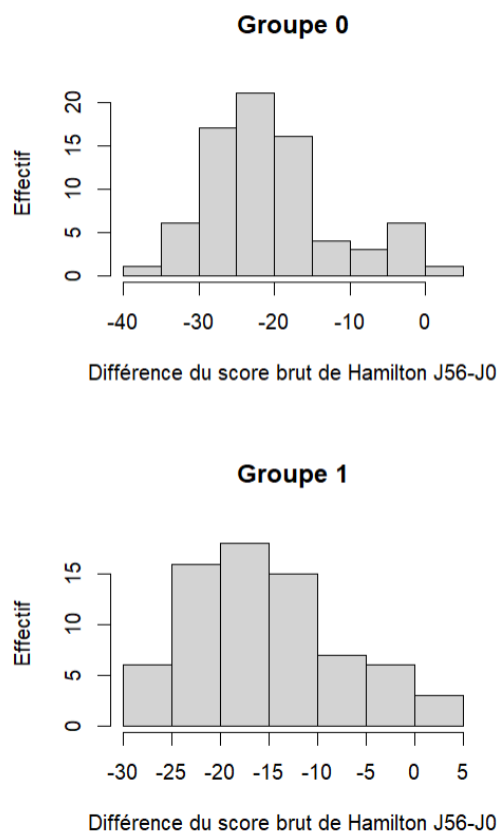


Table 5 : statistiques descriptives du score de Hamilton à J56 par groupe de patients (données non imputées et imputées par méthode LOCF)

J56 non imputé	Données disponibles	Moyenne	Ecart-type	Minimum	Maximum	1 ^{er} quartile	Médiane	3 ^e quartile
Groupe 0	64	5,5	3,9	0	17	3	4	8
Groupe 1	56	10,5	5,9	2	27	7,8	9	13

J56 imputé LOCF	Données disponibles	Moyenne	Ecart-type	Minimum	Maximum	1 ^{er} quartile	Médiane	3 ^e quartile
Groupe 0	75	8,1	8,1	0	42	3	6	10
Groupe 1	71	13	7,7	2	30	8	11	17

Figure 14 : histogramme de la distribution de la différence du score brut de Hamilton J56–J0 (données imputées selon la méthode LOCF), selon le groupe de patients



Nous comparons la moyenne de la différence de score brut de Hamilton J56 – J0 (avec données imputées) entre les deux groupes par un test t de Student sur données indépendantes :

- Groupe 0 : moyenne -19,9 ; écart-type 8,7
- Groupe 1 : moyenne -14,8 ; écart-type 8,2

Le test de Student est significatif au risque 5% ($p = 0,0003$).

L'intervalle de confiance de la différence entre les 2 groupes (groupe 0 – groupe 1) est [-7,9 ; -2,4].

Les conditions de validité du test t sont satisfaites : indépendance des observations, distribution d'allure normale des variables (et dans tous les cas bien plus de 30 observations par groupe), variances égales dans les 2 groupes.

Donc le score brut de Hamilton évolue significativement plus favorablement dans le groupe 0 que dans le groupe 1 à J56.

C – Comparaison du score de Hamilton entre J0 et J56 dans chaque groupe, après gestion des données manquantes par un modèle linéaire mixte

Dans tous les cas, la variable à expliquer est le score brut de Hamilton.

Les variables prises en effet fixe sont le groupe de patient et la visite.

Le patient (numéro) est pris en effet aléatoire (ordonnée à l'origine aléatoire, pente commune entre les patients).

- Modèle 1 : pas d'interaction entre groupe et visite, visite considérée comme variable catégorielle
- Modèle 2 : interaction entre groupe et visite, visite considérée comme variable catégorielle
- Modèle 3 : pas d'interaction entre groupe et visite, visite considérée comme variable quantitative
- Modèle 4 : interaction entre groupe et visite, visite considérée comme variable quantitative

Modèle	Visite	Interaction groupe* visite	BIC du modèle	Effet fixe du groupe (groupe 0 en référence)		Effet fixe de la visite	Effets aléatoires	
				Coefficient ± écart- type (IC95%)	p-value du test de nullité du coefficient	Coefficient ± écart-type (IC95%)	Ecart- type de l'effet patient	Ecart- type résiduel
1	quali	non	6379	2,9 ± 0,8 (1,3 ; 4,4)	0,0004	-7,1 ± 0,5 (-8,1 ; -6,2)	4,49	4,18
2	quali	oui	6358	Non interprétable directement avec un terme d'interaction	Interaction entre groupe et visite (J56) : p = 1,8.10 ⁻⁶	Non interprétable directement avec un terme d'interaction	4,49	4,06
3	quanti	non	6671	3,0 ± 0,8 (1,4 ; 4,6)	0,0002	-0,35 ± 0,008 (-0,37 ; - 0,34)	4,39	4,98
4	quanti	oui	6656	Non interprétable directement avec un terme d'interaction	Interaction entre groupe et visite : p = 0,19	Non interprétable directement avec un terme d'interaction	4,02	8,15

Commentaires techniques :

1) Le test de nullité du coefficient de l'effet fixe du groupe (1 par rapport à 0) a été réalisé avec la méthode d'ajustement du modèle ML (maximum de vraisemblance).

Toutes les autres variances et intervalles de confiance à 95%, ainsi que le BIC, sont fournis par la méthode d'ajustement REML (maximum de vraisemblance restreint).

2) Les résultats présentés ci-dessus sont ceux fournis par la fonction lmer du package lme4 de R.

Nous avons également réalisé ces modèles avec la fonction lme du package nlme. Les résultats sont superposables, nous ne les rapportons pas par souci de clarté.

3) Pour le coefficient de l'effet fixe de la visite, lorsque la visite était prise comme variable qualitative, nous indiquons le coefficient de l'effet de la modalité « visite J56 » (visite J0 prise comme référence).

Interprétation :

On retient le codage de la visite comme variable catégorielle et non pas comme variable continue (BIC plus faible des modèles 1 et 2).

A la visite J56, le score de Hamilton est plus élevé dans le groupe 1 que dans le groupe 0 (intervalle de confiance à 95% de la différence groupe 1 – groupe 0 : [1,3 ; 4,4], $p = 0,0004$).

Il existe une interaction statistiquement significative entre groupe et visite.

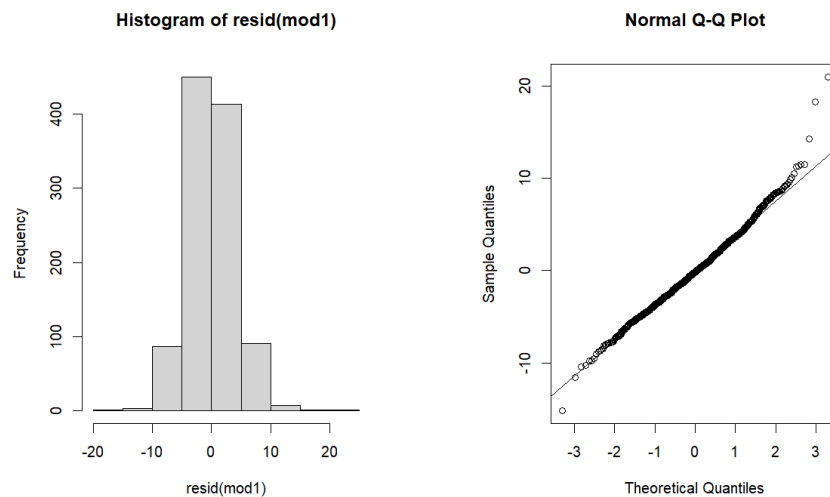
Ainsi le coefficient de l'effet fixe du groupe (différence groupe 1 – groupe 0) est -0,3 (IC95% [-2,2 ; 1,7], $p = 0,79$) après prise en compte de l'interaction entre groupe de visite. Cela revient à dire que les groupes ne sont pas différents à J0. Le temps fait baisser le score de Hamilton dans les 2 groupes, mais davantage dans le groupe 0 que dans le groupe 1.

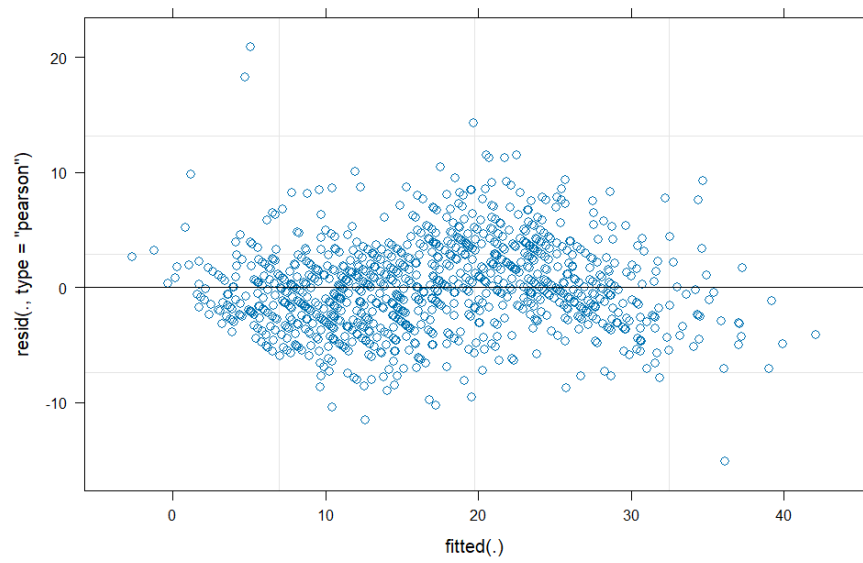
Conditions de validité :

Pour chacun des modèles testés, on a bien la nullité moyenne des résidus, la normalité des résidus et l'indépendance des résidus par rapport aux valeurs de la variable à expliquer.

Par souci de clarté, on ne représente la vérification graphique des conditions de validité que pour le modèle 1.

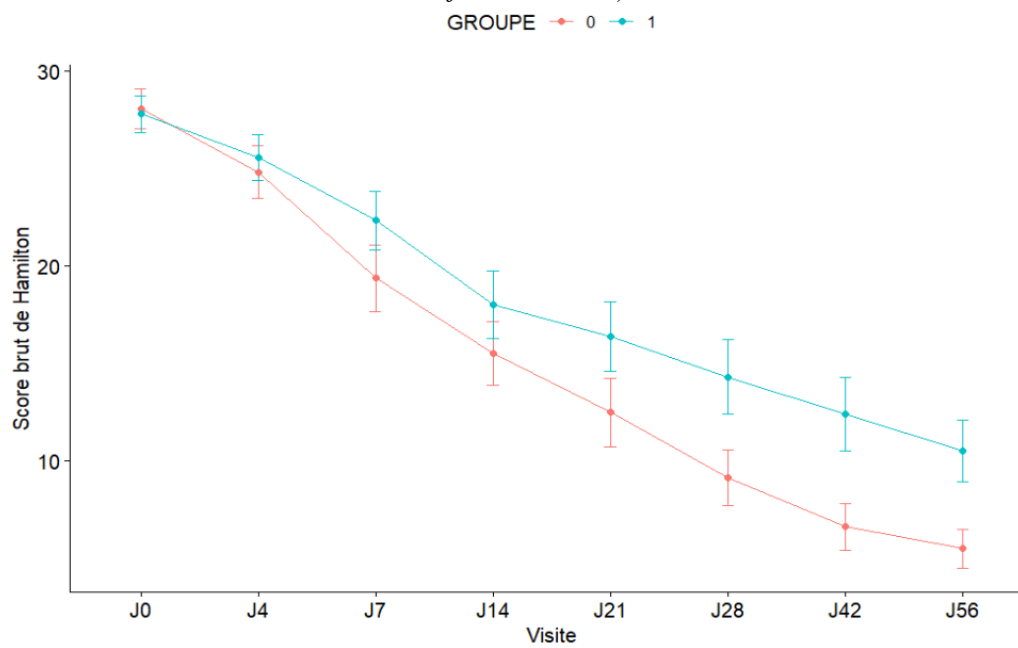
Figure 15 : histogramme des résidus, diagramme quantile-quantile des résidus, diagramme des résidus en fonction des valeurs prédites pour le modèle 1





De plus, les variables « temps » et « groupe » ont un effet linéaire sur le score brut de Hamilton.

Figure 16 : évolution du score de Hamilton en fonction de la visite (moyenne \pm intervalle de confiance à 95%)



IV – Question 3 : comparaison de la réponse au traitement par une analyse de survie

Nous comparons maintenant le temps jusqu'à réponse au traitement entre les deux groupes par une analyse de survie.

L'évènement (« décès ») est défini par la diminution d'au moins 50% du score brut de Hamilton par rapport au score à la visite J0.

Les causes de censures sont les perdus de vue (données manquantes à une visite entre J4 et J56 puis à toutes les visites suivantes) et les exclus-vivants (données disponibles de J0 à J56, sujets n'ayant pas encore présenté l'évènement à J56).

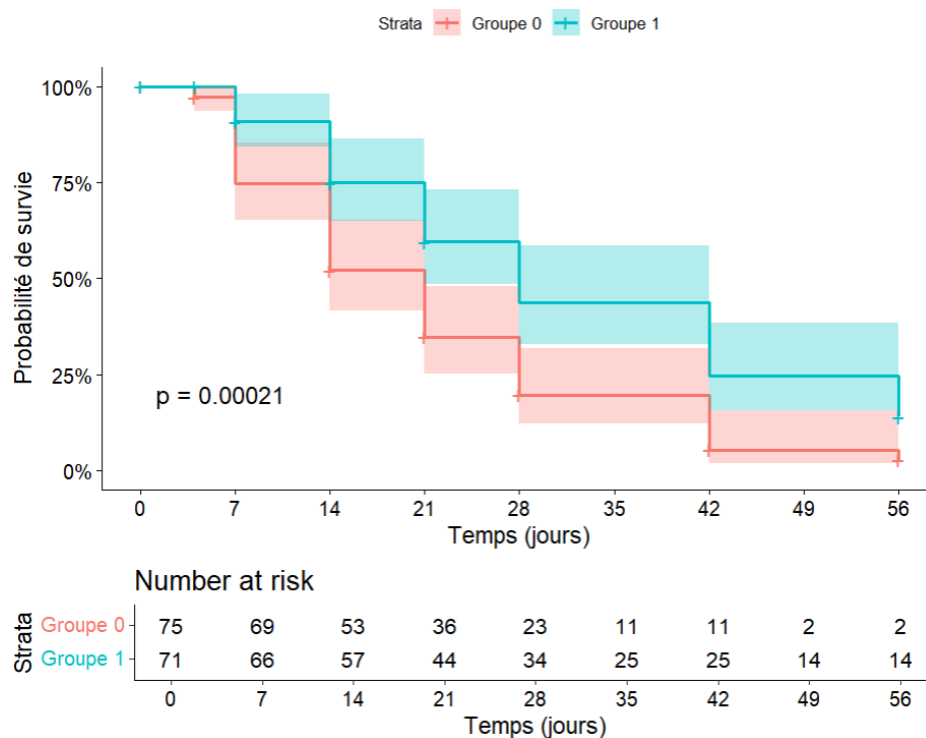
Nous constatons dans un premier temps que 114 individus présentent l'évènement et 32 individus sont censurés.

Cependant, l'observation du nombre de ces « censures » fait apparaître une discordance avec le nombre de données manquantes à chaque visite.

Nous observons que :

- la visite de l'individu 92 est manquante à J4 alors qu'il est présent à J7 et aux visites suivantes. Nous imputons la valeur manquante à J4 par celle disponible à J0 pour cet individu.
- la visite de l'individu 170 est manquante à J7 alors qu'il est présent à J14 et aux visites suivantes. Nous imputons la valeur manquante à J7 par celle disponible à J4 pour cet individu.

Figure 17 : fonctions de survie (baisse du score brut de Hamilton d'au moins 50% par rapport à J0) par groupe de traitement, estimées par la méthode de Kaplan-Meier. Les zones colorées autour des courbes correspondent à l'intervalle à 95% de la fonction de survie.



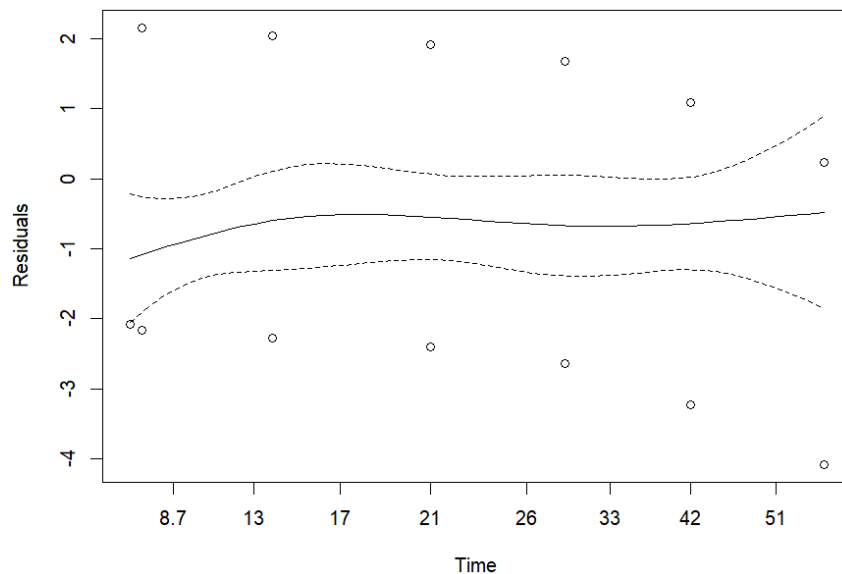
Le test du log-rank entre les fonctions de survie de chaque groupe est statistiquement significatif au risque 5% ($p = 0,0002$).

La probabilité instantanée de diminution du score brut de Hamilton de 50% par rapport à J0 (succès du traitement) est supérieure dans le groupe 0, par rapport au groupe 1.

Conditions de validité du test du log-rank de comparaison du hazard ratio à 1 :

- indépendance des observations, ce qu'on suppose ici (considérations méthodologiques)
- le processus de censure est non informatif (indépendance entre durée de survie et temps jusqu'à la censure), ce qui est improbable ici (il y a plus de perdus de vue dans le groupe 1, le moins efficace, ce qui peut être lié justement à la faible efficacité du traitement 1)
- risques proportionnels. On peut raisonnablement considérer que cette hypothèse est vérifiée ici au vu de la figure 17. On représente tout de même les résidus du modèle de Cox en fonction du temps, qui sont constants au cours du temps.

Figure 18 : graphique des résidus du modèle de Cox en fonction du temps (lignes pointillées à ± 2 erreurs-type)



V – Script R

```
library(readxl)
groupe<-read_excel("../outils groupe.xls")
hdrs<-read_excel("../outils hdrs.xls")
scl<-read_excel("../outils autoeval.xls")
View(groupe)
View(hdrs)
View(scl)
```

```
library(prettyR)
```

```
#Changement du type de variables
summary(groupe)
groupe$NUMERO<-as.factor(groupe$NUMERO)
groupe$GROUPE<-as.factor(groupe$GROUPE)
summary(groupe)
```

```
hdrs$NUMERO<-as.factor(hdrs$NUMERO)
hdrs$VISIT<-as.factor(hdrs$VISIT)
summary(hdrs)
```

```
scl$NUMERO<-as.factor(scl$NUMERO)
scl$VISIT<-as.factor(scl$VISIT)
summary(scl)
```

```
#Fusion des colonnes HAMD16A et HAMD16B
table(hdrs$HAMD16A,hdrs$HAMD16B,useNA="always")
hdrs$HAMD16<-ifelse(!is.na(hdrs$HAMD16A),hdrs$HAMD16A,hdrs$HAMD16B)
```

```
table(hdrs$HAMD16,hdrs$HAMD16A,useNA="always")#vérification de la fusion
table(hdrs$HAMD16,hdrs$HAMD16B,useNA="always")
```

```
colnames(hdrs)
hdrs<-hdrs[,-c(18,19)]#On retire les colonnes HAMD16A et HAMD16B.
colnames(hdrs)
```

```
hdrs<-hdrs[,c("NUMERO","VISIT",colHAMD)]#réorganisation des colonnes pour les ranger
de HAMD1 à HAMD17
colnames(hdrs)
```

```
#Data management : fusion des tableaux scl et hdrs avec le fichier groupe grâce au numéro de
patient
hdrs<-merge(hdrs,groupe,by="NUMERO")
scl<-merge(scl,groupe,by="NUMERO")
```

```
#Données manquantes
hdrs$NUMERO[is.na(hdrs$HAMD1)]
```

```

hdrs$VISIT[is.na(hdrs$HAMD1)]

sumNA<-data.frame(matrix(ncol=2,nrow=0))
colnames(sumNA)<-c("Question SCL90","Pourcentage de données manquantes")

for (i in c(3:92)) {
  res<-round(100*sum(is.na(scl[,i]))/length(scl[,i]),3)
  sumNA<-rbind.data.frame(sumNA,data.frame(i,res))
}
sumNA
max(sumNA[,2])
min(sumNA[,2])
which((sumNA[,2]==max(sumNA[,2])))

#Vérification de l'absence de doublon dans chaque tableau
hdrs$test_doublon<-paste(hdrs$NUMERO,hdrs$VISIT)#on concatène les colonnes
NUMERO et VISIT pour voir si une visite pour un même patient est bien unique dans le
tableau
duplicated(hdrs$test_doublon)

scl$test_doublon<-paste(scl$NUMERO,scl$VISIT)
duplicated(scl$test_doublon)

#Vérification de l'absence de valeur aberrante
summary(hdrs)
summary(scl)
for(i in c(3:92)){
  scl[,i]<-ifelse(scl[,i]>4,NA,scl[,i])
}
summary(scl)

for (i in c(3:92)) {#on calcule à nouveau le pourcentage de données manquantes
  res<-round(100*sum(is.na(scl[,i]))/length(scl[,i]),3)
  sumNA<-rbind.data.frame(sumNA,data.frame(i,res))
}
sumNA
max(sumNA[,2])
min(sumNA[,2])
which((sumNA[,2]==max(sumNA[,2])))

###Question 1 : validation de l'échelle de Hamilton
j0hdrs<-subset(hdrs,hdrs$VISIT=="J0")
summary(j0hdrs)
j56hdrs<-subset(hdrs,hdrs$VISIT=="J56")
summary(j56hdrs)

par(mfrow=c(3,3))
for (i in c(3:11)){

```



```
barplot(table(factor(j0hdrs[,i],levels=0:4),useNA="always"),ylim=c(0,100),main=paste("HAM",i-2))
}
```

```
par(mfrow=c(3,3))
for (i in c(12:19)){
```

```
barplot(table(factor(j0hdrs[,i],levels=0:4),useNA="always"),ylim=c(0,100),main=paste("HAM",i-2))
}
```

```
par(mfrow=c(3,3))
for (i in c(3:11)){
```

```
barplot(table(factor(j56hdrs[,i],levels=0:4),useNA="always"),ylim=c(0,100),main=paste("HAM",i-2))
}
```

```
par(mfrow=c(3,3))
for (i in c(12:19)){
```

```
barplot(table(factor(j56hdrs[,i],levels=0:4),useNA="always"),ylim=c(0,100),main=paste("HAM",i-2))
}
```

Calcul du score global de Hamilton et représentation en histogramme

```
par(mfrow=c(1,1))
j0hdrs$sumHDRS0<-apply(j0hdrs[,3:19],MARGIN=1,sum)
hist(j0hdrs$sumHDRS0,main="Distribution du score de Hamilton à J0",ylab="Effectif",xlab="Score global de Hamilton")
describe(j0hdrs$sumHDRS0,num.desc=c("mean","sd","median","min","max","valid.n"))
```

```
j56hdrs$sumHDRS56<-apply(j56hdrs[,3:19],MARGIN=1,sum)
hist(j56hdrs$sumHDRS56,main="Distribution du score de Hamilton à J56",ylab="Effectif",xlab="Score global de Hamilton")
describe(j56hdrs$sumHDRS56,num.desc=c("mean","sd","median","min","max","valid.n"))
```

Matrice de corrélation entre items

```
library(corrplot)
corj0<-cor(j0hdrs[,3:19])
corrplot(corj0,method='number')
corj56<-cor(j56hdrs[,3:19])
corrplot(corj56,method='number')
```

Diagramme des valeurs propres

```
library(psy)
mdspca(j0hdrs[,3:19])
mdspca(j56hdrs[,3:19])
scree.plot(j0hdrs[,3:19],simu=50,title="Scree plot à J0")
```

```

scree.plot(j56hdrs[,3:19],simu=50,title="Scree plot à J56")

###Analyse factorielle
print(factanal(na.omit(j0hdrs[,3:19]), factors = 3), cutoff = 0)
print(factanal(na.omit(j56hdrs[,3:19]), factors = 1), cutoff = 0)

###Coefficient de Cronbach avec son intervalle de confiance (bootstrap)
library(boot)
cronbach.boot<-function(data,i){cronbach(data[i,])[[3]]}

data<-j0hdrs[,3:19]
cronbach(j0hdrs[,3:19])
bootj0<-boot(data,cronbach.boot,10000)
quantile(bootj0$t,c(0.025,0.975))

data<-j56hdrs[,3:19]
cronbach(j56hdrs[,3:19])
bootj56<-boot(data,cronbach.boot,10000)
quantile(bootj56$t,c(0.025,0.975))

###Validité convergente et divergente (échelle SCL90)
j0scl<-subset(scl,scl$VISIT=="J0")
summary(j0scl)
j56scl<-subset(scl,scl$VISIT=="J56")
summary(j56scl)

j0hdrs_short<-data.frame(j0hdrs[,c(1,22)])#tableau HDRS plus court pour ne fusionner que la
colonne sumHDRS
j56hdrs_short<-data.frame(j56hdrs[,c(1,22)])
summary(j0hdrs_short)
summary(j56hdrs_short)

j0scl<-merge(j0scl,j0hdrs_short,by="NUMERO")
j56scl<-merge(j56scl,j56hdrs_short,by="NUMERO")
summary(j0scl)
summary(j56scl)

j0scl$SCL_J0_Somatisation<-
apply(j0scl[,c(3,6,14,29,44,50,51,54,55,58,60,42)],MARGIN=1,sum)
j0scl$SCL_J0_Obsession<-apply(j0scl[,c(11,12,30,40,5,47,48,53,57,67)],MARGIN=1,sum)
j0scl$SCL_J0_Vulnerabilite<-apply(j0scl[,c(8,23,36,38,39,43,63,71,75)],MARGIN=1,sum)
j0scl$SCL_J0_Depression<-
apply(j0scl[,c(7,16,17,22,24,28,31,32,33,34,56,73,81)],MARGIN=1,sum)
j0scl$SCL_J0_Anxiété<-apply(j0scl[,c(4,19,25,35,41,59,74,80,82,88)],MARGIN=1,sum)
j0scl$SCL_J0_Hostilite<-apply(j0scl[,c(13,25,65,69,76,83)],MARGIN=1,sum)
j0scl$SCL_J0_Phobies<-apply(j0scl[,c(15,27,49,72,77,84,52)],MARGIN=1,sum)
j0scl$SCL_J0_Paranoia<-apply(j0scl[,c(10,20,45,70,78,85)],MARGIN=1,sum)
j0scl$SCL_J0_Psychose<-apply(j0scl[,c(9,18,37,64,79,86,87,89,92,90)],MARGIN=1,sum)
j0scl$SCL_J0_Divers<-apply(j0scl[,c(21,46,61,62,66,68,91)],MARGIN=1,sum)

```

```

j56scl$SCL_J56_Somatisation<-
apply(j56scl[,c(3,6,14,29,44,50,51,54,55,58,60,42)],MARGIN=1,sum)
j56scl$SCL_J56_Obsession<-
apply(j56scl[,c(11,12,30,40,5,47,48,53,57,67)],MARGIN=1,sum)
j56scl$SCL_J56_Vulnerabilite<-
apply(j56scl[,c(8,23,36,38,39,43,63,71,75)],MARGIN=1,sum)
j56scl$SCL_J56_Depression<-
apply(j56scl[,c(7,16,17,22,24,28,31,32,33,34,56,73,81)],MARGIN=1,sum)
j56scl$SCL_J56_Anxiete<-apply(j56scl[,c(4,19,25,35,41,59,74,80,82,88)],MARGIN=1,sum)
j56scl$SCL_J56_Hostilite<-apply(j56scl[,c(13,25,65,69,76,83)],MARGIN=1,sum)
j56scl$SCL_J56_Phobies<-apply(j56scl[,c(15,27,49,72,77,84,52)],MARGIN=1,sum)
j56scl$SCL_J56_Paranoia<-apply(j56scl[,c(10,20,45,70,78,85)],MARGIN=1,sum)
j56scl$SCL_J56_Psychose<-
apply(j56scl[,c(9,18,37,64,79,86,87,89,92,90)],MARGIN=1,sum)
j56scl$SCL_J56_Divers<-apply(j56scl[,c(21,46,61,62,66,68,91)],MARGIN=1,sum)

corj0_SCL<-cor(j0scl[,95:105],use="complete.obs")
corrplot(corj0_SCL,method='number',number.cex=0.7)
corj56_SCL<-cor(j56scl[,95:105],use="complete.obs")
corrplot(corj56_SCL,method='number',number.cex=0.7)

```

#Question 2

###Transformation en fichier large

```

lhdrs<-lhdrs[,-21]#je retire la colonne "test_doublon" qui m'avait servi précédemment
lhdrs$sumHDRS<-apply(lhdrs[,3:19],MARGIN=1,sum)#calcul du score brut de Hamilton à
chaque visite
lhdrs<-lhdrs[,-(3:19)]#simplification du tableau long pour ne garder que le score brut
dim(lhdrs)
colnames(lhdrs)
lhdrs<-
reshape(data=lhdrs,idvar=c("NUMERO","GROUPE"),timevar="VISIT",direction="wide")
dim(lhdrs)
colnames(lhdrs)

```

Description du score brut à J0 selon les groupes

```

describe(lhdrs$sumHDRS.J0[lhdrs$GROUPE==0],num.desc=c("mean","sd","median","min",
"max","valid.n"))
describe(lhdrs$sumHDRS.J0[lhdrs$GROUPE==1],num.desc=c("mean","sd","median","min",
"max","valid.n"))
summary(lhdrs$sumHDRS.J0[lhdrs$GROUPE==0])
summary(lhdrs$sumHDRS.J0[lhdrs$GROUPE==1])

```

library(plotrix)#histogramme du score à J0 selon les groupes

```

histogramme<-
list(lhdrs$sumHDRS.J0[lhdrs$GROUPE==0],lhdrs$sumHDRS.J0[lhdrs$GROUPE==1])
par(mar = c(5,5,5,5))

```

```

histo<-multhist(histogramme,ylab="Effectif",xlab="Score brut de
Hamilton",ylim=c(0,50),col=c("lightgrey","gray32"))
legend("topright",legend=paste("Groupe",0:1),fill=c("lightgrey","gray32"),cex=.6)

table(!is.na(lhdrs[,3]),lhdrs[,2])#nombre de données non manquantes à chaque visite
table(!is.na(lhdrs[,4]),lhdrs[,2])
table(!is.na(lhdrs[,5]),lhdrs[,2])
table(!is.na(lhdrs[,6]),lhdrs[,2])
table(!is.na(lhdrs[,7]),lhdrs[,2])
table(!is.na(lhdrs[,8]),lhdrs[,2])
table(!is.na(lhdrs[,9]),lhdrs[,2])
table(!is.na(lhdrs[,10]),lhdrs[,2])

library(ggplot2)#spaghetti plot
library(GGally)
ggparcoord(
  data = lhdrs,
  columns=3:10,
  groupColumn=2,
  missing="exclude",
  scale="globalminmax")
+labs(x="Visites",y="Score brut de Hamilton")
+scale_x_discrete(labels=c("J0","J4","J7","J14","J21","J28","J42","J56"))
+stat_summary(fun=mean,group=2,geom="point")

###Imputation des données manquantes par LOCF
locf_hdrs<-lhdrs
sum(is.na(locf_hdrs))
#na.locf ne donne pas de bons résultats sur un fichier wide, je le repasse en format long.
locf_hdrs<-reshape(data=locf_hdrs,idvar=c("NUMERO","GROUPE"),
  varying=3:10,
  times=c("J0","J4","J7","J14","J21","J28","J42","J56"),
  direction="long")
dim(locf_hdrs)#il y a bien 1168 lignes = 146 patients * 8 visites
head(locf_hdrs)
sum(is.na(locf_hdrs))
locf_hdrs<-locf_hdrs[order(locf_hdrs$NUMERO),]#tri dans l'ordre des numéros (les visites
sont déjà dans l'ordre)

library(zoo)
locf_hdrs<-na.locf(locf_hdrs)#je fais une vérification visuelle sur quelques patients ayant des
NA
View(locf_hdrs)
sum(is.na(locf_hdrs))

locf_hdrs<-
reshape(data=locf_hdrs,idvar=c("NUMERO","GROUPE"),timevar="time",direction="wide")
dim(locf_hdrs)
View(locf_hdrs)

```

```
ggparcoord(
  data = locf_hdrs,
  columns=3:10,
  groupColumn=2,
  missing="exclude",
  scale="globalminmax")+labs(x="Visites",y="Score brut de
Hamilton")+scale_x_discrete(labels=c("J0","J4","J7","J14","J21","J28","J42","J56"))+stat_su
mmmary(fun=mean,group=2,geom="point")
```

```
describe(locf_hdrs$sumHDRS.J0[locf_hdrs$GROUPE==0],num.desc=c("mean","sd","media
n","min","max","valid.n"))#Stat descriptive sur J56 imputée
summary(locf_hdrs$sumHDRS.J0[locf_hdrs$GROUPE==0])
describe(locf_hdrs$sumHDRS.J0[locf_hdrs$GROUPE==1],num.desc=c("mean","sd","media
n","min","max","valid.n"))
summary(locf_hdrs$sumHDRS.J0[locf_hdrs$GROUPE==1])
describe(locf_hdrs$sumHDRS.J56[locf_hdrs$GROUPE==0],num.desc=c("mean","sd","medi
an","min","max","valid.n"))
summary(locf_hdrs$sumHDRS.J56[locf_hdrs$GROUPE==0])
describe(locf_hdrs$sumHDRS.J56[locf_hdrs$GROUPE==1],num.desc=c("mean","sd","medi
an","min","max","valid.n"))
summary(locf_hdrs$sumHDRS.J56[locf_hdrs$GROUPE==1])
describe(lhdrs$sumHDRS.J56[lhdrs$GROUPE==0],num.desc=c("mean","sd","median","min
","max","valid.n"))
summary(lhdrs$sumHDRS.J56[lhdrs$GROUPE==0])
describe(lhdrs$sumHDRS.J56[lhdrs$GROUPE==1],num.desc=c("mean","sd","median","min
","max","valid.n"))
summary(lhdrs$sumHDRS.J56[lhdrs$GROUPE==1])
```

```
locf_hdrs$deltaHDRS<-locf_hdrs$sumHDRS.J56-locf_hdrs$sumHDRS.J0#calcul du delta
J56-J0
```

```
par(mfrow=c(2,1))
histo2<-hist(locf_hdrs$deltaHDRS[locf_hdrs$GROUPE==0],main="Groupe
0",ylab="Effectif",xlab="Différence du score brut de Hamilton J56-J0")
histo3<-hist(locf_hdrs$deltaHDRS[locf_hdrs$GROUPE==1],main="Groupe
1",ylab="Effectif",xlab="Différence du score brut de Hamilton J56-J0")
```

```
sd(locf_hdrs$deltaHDRS[locf_hdrs$GROUPE==0])#test de Student
sd(locf_hdrs$deltaHDRS[locf_hdrs$GROUPE==1])
t.test(locf_hdrs$deltaHDRS~locf_hdrs$GROUPE,var.equal=TRUE)
```

```
#### Modèle mixte
```

```
library(lme4)
library(nlme)
library(lmerTest)
```

```
mixte_hdrs<-lhdrs
mixte_hdrs<-reshape(data=mixte_hdrs,idvar=c("NUMERO","GROUPE"),
  varying=3:10,
  times=c("J0","J4","J7","J14","J21","J28","J42","J56"),
```

```

        direction="long")
dim(mixte_hdrs)
colnames(mixte_hdrs)

typeof(mixte_hdrs$time)
mixte_hdrs$time<-as.factor(mixte_hdrs$time)#transformation de time en factor
typeof(mixte_hdrs$time)
mixte_hdrs$time<-
factor(mixte_hdrs$time,levels=c("J0","J4","J7","J14","J21","J28","J42","J56"))#modalités de
la variable time rangées par ordre croissant
levels(mixte_hdrs$time)

mixte_hdrs$timenum<-mixte_hdrs$time
mixte_hdrs$timenum<-
substr(mixte_hdrs$timenum,2,nchar(as.character(mixte_hdrs$timenum)))#création d'une
autre variable time, cette fois numérique (je retire le J devant le jour de la visite)
mixte_hdrs$timenum<-as.numeric(mixte_hdrs$timenum)

mod1<-lmer(sumHDRS~GROUPE+time+(1|NUMERO),REML=TRUE,data=mixte_hdrs)
summary(mod1)
BIC(mod1)
confint(mod1)
par(mfrow=c(1,2))
hist(resid(mod1))
qqnorm(resid(mod1))
qqline(resid(mod1))
par(mfrow=c(1,1))
plot(mod1)

mod2<-
lmer(sumHDRS~GROUPE+time+(1|NUMERO),REML=FALSE,data=mixte_hdrs)#pour test
de la nullité des coefficients (méthode ML)
summary(mod2)
hist(resid(mod2))
qqnorm(resid(mod2))
qqline(resid(mod2))
plot(mod2)

mod3<-
lme(sumHDRS~GROUPE+time,random=~1|NUMERO,method="REML",na.action=na.omit,
data=mixte_hdrs)#même modèle que le 1 ci-dessus, mais avec la fonction lme
summary(mod3)
hist(resid(mod3))
qqnorm(resid(mod3))
qqline(resid(mod3))
plot(mod3)

```

```

mod4<-
lme(sumHDRS~GROUPE+time,random=~1|NUMERO,method="ML",na.action=na.omit,dat
a=mixte_hdrs)#même modèle que le 2 ci-dessus, mais avec la fonction lme
summary(mod4)
hist(resid(mod4))
qqnorm(resid(mod4))
qqline(resid(mod4))
plot(mod4)

```

contrasts(mixte_hdrs\$time)#vérification des modalités de codage de la variable catégorielle time (avant d'ajouter le terme d'interaction)

```

mod5<-
lmer(sumHDRS~GROUPE*time+(1|NUMERO),REML=TRUE,data=mixte_hdrs)#modèle
avec interaction
summary(mod5)
BIC(mod5)
confint(mod5)
hist(resid(mod5))
qqnorm(resid(mod5))
qqline(resid(mod5))
plot(mod5)

```

```

mod6<-
lmer(sumHDRS~GROUPE*time+(1|NUMERO),REML=FALSE,data=mixte_hdrs)#modèle
avec interaction
summary(mod6)
hist(resid(mod6))
qqnorm(resid(mod6))
qqline(resid(mod6))
plot(mod6)

```

```

mod7<-
lme(sumHDRS~GROUPE*time,random=~1|NUMERO,method="REML",na.action=na.omit,
data=mixte_hdrs)#même modèle que le 5 ci-dessus
summary(mod7)
hist(resid(mod7))
qqnorm(resid(mod7))
qqline(resid(mod7))
plot(mod7)

```

```

mod8<-
lme(sumHDRS~GROUPE*time,random=~1|NUMERO,method="ML",na.action=na.omit,dat
a=mixte_hdrs)#même modèle que le 6 ci-dessus
summary(mod8)
hist(resid(mod8))
qqnorm(resid(mod8))
qqline(resid(mod8))
plot(mod8)

```

```

mod9<-
lmer(sumHDRS~GROUPE+timenum+(1|NUMERO),REML=TRUE,data=mixte_hdrs)#avec
visite en variable quantitative
summary(mod9)
BIC(mod9)
confint(mod9)
hist(resid(mod9))
qqnorm(resid(mod9))
qqline(resid(mod9))
plot(mod9)

```

```

mod10<-
lmer(sumHDRS~GROUPE+timenum+(1|NUMERO),REML=FALSE,data=mixte_hdrs)
summary(mod10)
hist(resid(mod10))
qqnorm(resid(mod10))
qqline(resid(mod10))
plot(mod10)

```

```

mod11<-
lme(sumHDRS~GROUPE+timenum,random=~1|NUMERO,method="REML",na.action=na.
omit,data=mixte_hdrs)#même modèle que le 9 ci-dessus
summary(mod11)
hist(resid(mod11))
qqnorm(resid(mod11))
qqline(resid(mod11))
plot(mod11)

```

```

mod12<-
lme(sumHDRS~GROUPE+timenum,random=~1|NUMERO,method="ML",na.action=na.omi
t,data=mixte_hdrs)#même modèle que le 10 ci-dessus
summary(mod12)
hist(resid(mod12))
qqnorm(resid(mod12))
qqline(resid(mod12))
plot(mod12)

```

```

mod13<-
lmer(sumHDRS~GROUPE*timenum+(1|NUMERO),REML=TRUE,data=mixte_hdrs)#avec
visite en variable quantitative, terme d'interaction
summary(mod13)
BIC(mod13)
confint(mod13)
hist(resid(mod13))
qqnorm(resid(mod13))
qqline(resid(mod13))
plot(mod13)

```

```

mod14<-
lmer(sumHDRS~GROUPE*timenum+(1|NUMERO),REML=FALSE,data=mixte_hdrs)

```



```

summary(mod14)
hist(resid(mod14))
qqnorm(resid(mod14))
qqline(resid(mod14))
plot(mod14)

mod15<-
lme(sumHDRS~GROUPE*timenum,random=~1|NUMERO,method="REML",na.action=na.omit,data=mixte_hdrs)#même modèle que le 13 ci-dessus
summary(mod15)
hist(resid(mod15))
qqnorm(resid(mod15))
qqline(resid(mod15))
plot(mod15)

mod16<-
lme(sumHDRS~GROUPE*timenum,random=~1|NUMERO,method="ML",na.action=na.omit,data=mixte_hdrs)#même modèle que le 16 ci-dessus
summary(mod16)
hist(resid(mod16))
qqnorm(resid(mod16))
qqline(resid(mod16))
plot(mod16)

####Meanplot du score de Hamilton
library(ggpubr)
ggline(mixte_hdrs,x="time",y="sumHDRS",group="GROUPE",color="GROUPE",add=c("mean_ci"),xlab="Visite",ylab="Score brut de Hamilton")

```

```

#Question 3 : analyse de survie
library(survival)

```

```

lhdrs_surv<-lhdrs
lhdrs_surv$temps_suivi<-lhdrs_surv$sumHDRS.J0#création d'une variable "temps de suivi"
lhdrs_surv$evenement<-lhdrs_surv$sumHDRS.J0#création d'une variable événement : 1 si censure, 2 si décès

##Pour chaque ligne de lhdrs_surv, la boucle vérifie d'abord que le score à J4 est NA.
##Si TRUE, alors c'est une censure à J0 : temps_suivi=0 et evenement=1.
##Si FALSE, alors la boucle vérifie que le score à J4 est inférieur à la moitié du score à J0.
##Si TRUE, alors c'est un événement à J4 : temps_suivi=4 et evenement=2.
##Si FALSE, alors on vérifie le statut du score à J7 (NA ? si non inférieur à la moitié du score à J0 ?), et ainsi de suite
for (i in 1:nrow(lhdrs_surv)) {
  if (is.na(lhdrs_surv$sumHDRS.J4[i])) {
    lhdrs_surv$temps_suivi[i] <- 0
    lhdrs_surv$evenement[i] <- 1
  }
}

```

```

} else if (lhdrs_surv$sumHDRS.J4[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 4
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J7[i])) {
  lhdrs_surv$temps_suiwi[i] <- 4
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J7[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 7
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J14[i])) {
  lhdrs_surv$temps_suiwi[i] <- 7
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J14[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 14
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J21[i])) {
  lhdrs_surv$temps_suiwi[i] <- 14
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J21[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 21
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J28[i])) {
  lhdrs_surv$temps_suiwi[i] <- 21
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J28[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 28
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J42[i])) {
  lhdrs_surv$temps_suiwi[i] <- 28
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J42[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 42
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J56[i])) {
  lhdrs_surv$temps_suiwi[i] <- 42
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J56[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suiwi[i] <- 56
  lhdrs_surv$evenement[i] <- 2
} else {
  lhdrs_surv$temps_suiwi[i] <- 56
  lhdrs_surv$evenement[i] <- 1
}
}
}

```

Surv(lhdrs_surv\$temps_suiwi,lhdrs_surv\$evenement)
table(lhdrs_surv\$evenement,lhdrs_surv\$temps_suiwi)#on a ainsi le nombre de décès et de
censures à chaque visite

##Vérification du nombre de données manquantes dans chaque colonne

```

sumNAcol<-data.frame(matrix(ncol=2,nrow=0))
for (i in c(3:10)) {
  res<-sum(is.na(lhdrs_surv[,i]))
  sumNAcol<-rbind.data.frame(sumNAcol,data.frame(i,res))
}
sumNAcol

##On observe 2 discordances :
###La boucle entraîne 4 censures à J4, or il n'y a que 3 données manquantes de plus à J7 par rapport à J4.
###Donc un individu a été erronément censuré à J4 alors qu'il "réapparaît" à J7.
lhdrs_surv$NUMERO[which(is.na(lhdrs_surv$sumHDRS.J4)&!is.na(lhdrs_surv$sumHDRS.J7))]
###Il s'agit de l'individu 92.
###Puisqu'il ne s'agit pas vraiment d'une censure, nous décidons d'imputer la valeur manquante à J4 par celle de J0.
lhdrs_surv$sumHDRS.J4[lhdrs_surv$NUMERO==92]<-
lhdrs_surv$sumHDRS.J0[lhdrs_surv$NUMERO==92]
###On détecte une 2e discordance :
lhdrs_surv$NUMERO[which(is.na(lhdrs_surv$sumHDRS.J7)&!is.na(lhdrs_surv$sumHDRS.J14))]
###Il s'agit de l'individu 170.
###Nous imputons sa valeur manquante à J7 par celle de J4.
lhdrs_surv$sumHDRS.J7[lhdrs_surv$NUMERO==170]<-
lhdrs_surv$sumHDRS.J4[lhdrs_surv$NUMERO==170]

###Ainsi fait, on refait tourner la boucle pour les indicatrices d'événement/censure et le délai d'observation.
for (i in 1:nrow(lhdrs_surv)) {
  if (is.na(lhdrs_surv$sumHDRS.J4[i])) {
    lhdrs_surv$temps_suivi[i] <- 0
    lhdrs_surv$evenement[i] <- 1
  } else if (lhdrs_surv$sumHDRS.J4[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
    lhdrs_surv$temps_suivi[i] <- 4
    lhdrs_surv$evenement[i] <- 2
  } else if (is.na(lhdrs_surv$sumHDRS.J7[i])) {
    lhdrs_surv$temps_suivi[i] <- 4
    lhdrs_surv$evenement[i] <- 1
  } else if (lhdrs_surv$sumHDRS.J7[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
    lhdrs_surv$temps_suivi[i] <- 7
    lhdrs_surv$evenement[i] <- 2
  } else if (is.na(lhdrs_surv$sumHDRS.J14[i])) {
    lhdrs_surv$temps_suivi[i] <- 7
    lhdrs_surv$evenement[i] <- 1
  } else if (lhdrs_surv$sumHDRS.J14[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
    lhdrs_surv$temps_suivi[i] <- 14
    lhdrs_surv$evenement[i] <- 2
  } else if (is.na(lhdrs_surv$sumHDRS.J21[i])) {
    lhdrs_surv$temps_suivi[i] <- 14
    lhdrs_surv$evenement[i] <- 1
  }
}

```

```

} else if (lhdrs_surv$sumHDRS.J21[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suivi[i] <- 21
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J28[i])) {
  lhdrs_surv$temps_suivi[i] <- 21
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J28[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suivi[i] <- 28
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J42[i])) {
  lhdrs_surv$temps_suivi[i] <- 28
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J42[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suivi[i] <- 42
  lhdrs_surv$evenement[i] <- 2
} else if (is.na(lhdrs_surv$sumHDRS.J56[i])) {
  lhdrs_surv$temps_suivi[i] <- 42
  lhdrs_surv$evenement[i] <- 1
} else if (lhdrs_surv$sumHDRS.J56[i] <= 0.5 * lhdrs_surv$sumHDRS.J0[i]) {
  lhdrs_surv$temps_suivi[i] <- 56
  lhdrs_surv$evenement[i] <- 2
} else {
  lhdrs_surv$temps_suivi[i] <- 56
  lhdrs_surv$evenement[i] <- 1
}
}
}

```

```

Surv(lhdrs_surv$temps_suivi,lhdrs_surv$evenement)
survfit(Surv(lhdrs_surv$temps_suivi,lhdrs_surv$evenement) ~ lhdrs_surv$GROUPE)
survie<-survfit(Surv(lhdrs_surv$temps_suivi,lhdrs_surv$evenement) ~ lhdrs_surv$GROUPE)

```

```

library(survminer)
ggssurvplot(survie,data=lhdrs_surv,conf.int=.95,surv.scale="percent",
  xlab="Temps (jours)",ylab="Probabilité de survie",
  pval=TRUE,risk.table=TRUE,
  legend.labs=c("Groupe 0","Groupe 1"),
  xlim=c(0,56),break.time.by=7)

```

```

survdiff(Surv(lhdrs_surv$temps_suivi,lhdrs_surv$evenement) ~ lhdrs_surv$GROUPE)

```

```

####Vérification de l'hypothèse des risques proportionnels
cox_mod<-coxph(Surv(lhdrs_surv$temps_suivi,lhdrs_surv$evenement) ~
lhdrs_surv$GROUPE)
cox_mod
cox.zph(cox_mod)
plot(cox.zph(cox_mod),ylab="Residuals")

```