LSTAT2210 Project

Tytgat Alexandre

March 7, 2021

1 Introduction

Dans ce rapport, plusieurs concepts et techniques de modèles linéaires avancés sont mises en pratique dans un cas concret. Ce dernier se base sur un jeu de donnée récolté par des chercheurs s'intéressant à l'effet de privation de sommeil sur le temps de réaction des patients. Ces patients ont suivi un régime de trois heures de sommeil pendant 9 jours (le premier jour, les patients dorment sans restriction). Ces patients ont été séparés dans deux groupes de traitement : l'un où un placebo leur a été administré (traitement A) et l'autre dans lequel ils ont reçu un traitement expérimental (traitement B). Ce rapport est structuré de la façon suivante : Tout d'abord, une présentation suivi d'une analyse des données est proposée. Ensuite, sept modèles de régression linéaire de complexité croissante seront présentés. Pour chacun d'entre eux, les estimations des paramètres seront effectués via le logiciel SAS et les résultats principaux seront présentés. Enfin, une résumé des résultats principaux sera offert ainsi qu'une analyse plus approfondie du modèle le plus performant.

2 Analyse des données

Une description des variables considérées pour cette étude est présentée dans la Table 1.

Variable	Description
Days	jour par rapport au cycle (jour 0: sommeil normal, 1-9 jours suivants)
Subject	identifiant du patient
age	âge du patient
treatment	traitement A ou B
Reaction	temps de réaction en ms

Table 1: Description des variables.

Commençons par considérer l'impact potentiel du traitement sur le temps de réaction. Le panel (a) de la Figure 1 semble indiquer que les individus ayant reçu le traitement A, c'est à dire le placebo, on un temps de réaction plus élevé que ceux ayant pris le traitement B. Cependant, on observe aussi dans le panel (b) de la même figure que les individus du groupe A sont en moyenne plus vieux que ceux du groupe B. Il est donc possible que ce soit cette différence d'âge qui explique la différence d'effet entre les deux groupes plutôt que le traitement A ou B. Le panel (b) de la Figure 2 renforce cette hypothèse. En effet, il est clair que en moyenne plus un individus est âgé, plus son temps de réaction augmente. Ceci colle parfaitement avec nos attentes. En effet, il est connu qu'à partir d'un certain âge, les capacités mentales d'un individu régressent et en particulier, le temps de réaction à tendance à augmenter.

Quant à la variable `Days`, celle-ci est représentée via son effet sur le temps de réaction dans le panel (a) de la Figure 2. Il semblerait, au même titre que la variable `age`, qu'il existe une relation linéaire avec `Reaction`. Cela serait cohérent avec notre intuition car nous nous attendons à ce que à mesure que le manque de sommeil s'accumule au fil des jours, le temps de réaction des sujets devraient augmenter.

Mis à part l'analyse de la corrélation entre les variables, il est aussi intéressant de regarder de plus près la nature de la variable réponse. C'est dans ce but que sont présentés les deux graphes de la Figure 3. Celui du panel (a) montre un diagramme QQ de comparaison de la distribution de la variable `Reaction` avec une loi normale. Nous observons un bon alignement avec la droite théorique entre les quantiles -1 et 1. Au delà de ceux-ci, les points sont de plus en plus éloignés, en particulier à l'extrémité de gauche. Quant au panel (b), un histogramme de la variable réponse `Reaction` est représenté. Nous observons une distribution centrée aux alentours de 275 avec une légère asymétrie, comme le suggère également le diagramme QQ. Cette analyse visuelle indique que le temps de réaction peut être modélisé par une loi normale, bien qu'il n'est pas possible

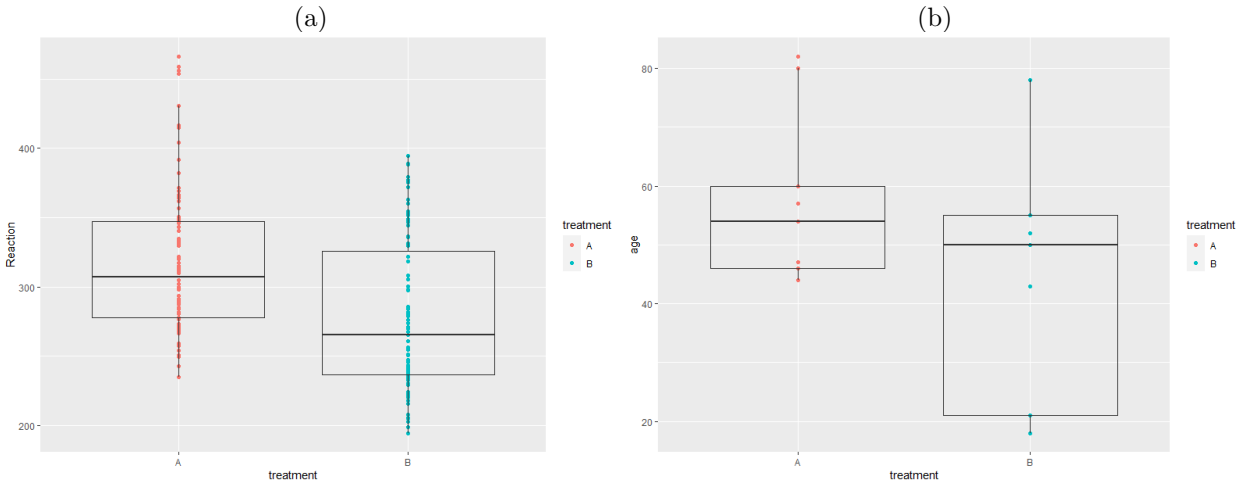

Figure 1: (a) Boxplot de `treatment` vs. `Reaction`. (b) Boxplot de `treatment` vs. `age`.

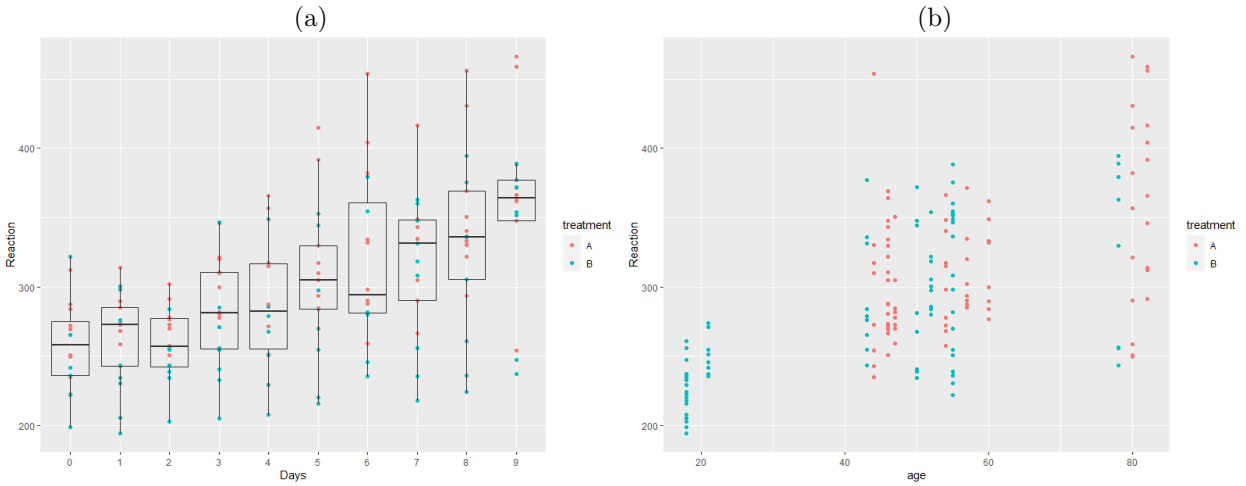

Figure 2: (a) Boxplot de `Days` vs. `Reaction` time. (b) Graphe de `age` vs. `Reaction`.

d'affirmer avec certitude que cette loi est la mieux appropriée. Toutefois, il sera montré par la suite, lors de l'estimation des modèles, que ce choix est justifié à posteriori.

Enfin, notons que les mesures pour certains jours de quelques sujets sont manquantes. Celles-ci sont supposées être manquante de manière aléatoire et leur traitement sera détaillé par la suite lorsque cela sera nécessaire.

3 Modèles

Dans cette section, plusieurs modèles statistiques seront présentés, prenant en compte de différentes façon les variables présentées dans la Table 1, afin d'offrir une évaluation de leur contribution respectives sur le temps de réaction. En particulier, l'intérêt sera porté sur l'impact du traitement. Les modèles considérés sont de complexité croissante et se basent systématiquement sur le meilleur modèle obtenu jusqu'à présent (à l'exception du premier bien entendu).

Afin d'évaluer les performances de chaque modèle, les critères de qualité BIC et l'AICc (plus adapté que l'AIC étant donné le faible nombre des données) ont été choisis.

Enfin, les valeurs obtenues ont toutes été arrondies à la troisième décimale près, à l'exception des p-valeurs. De plus, les intervalles de confiance ont tous été estimés avec $\alpha = 0.05$.

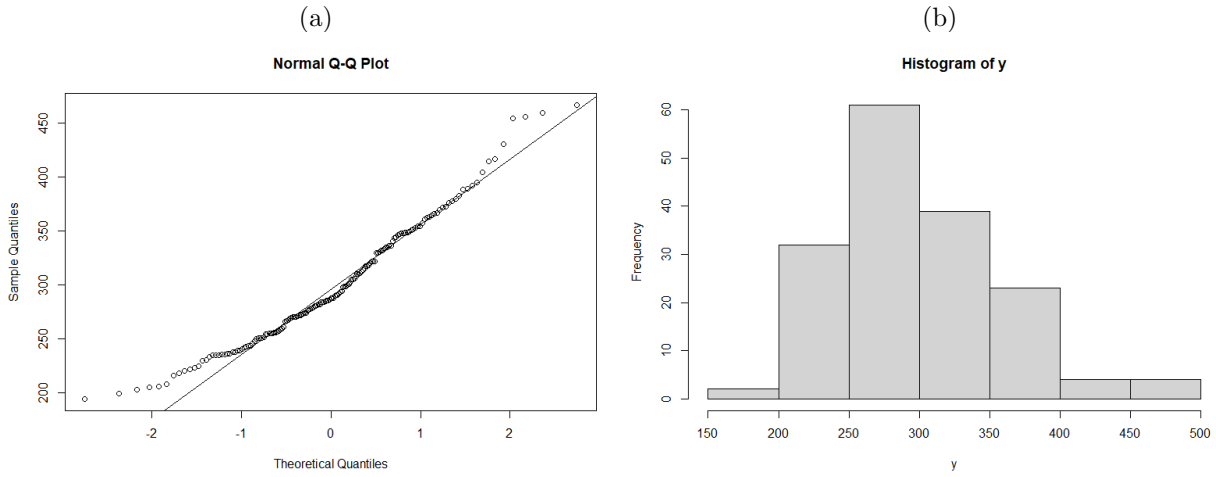

Figure 3: Visualisation de l'hypothèse de normalité de la variable réponse. (a) Diagramme QQ `Reaction` (b) Histogramme de `Reaction`.

3.1 Modèle d'ANOVA de la variable traitement

Dans ce premier modèle, seul l'effet du traitement est inclus en tant qu'un effet fixe. Ceci correspond donc à un modèle d'ANOVA, un test de différence de moyenne pour les individus dans le groupe A et ceux du groupe B. L'équation du modèle est donc la suivante :

$$Y_{ij} = \beta_0 + \beta_1 \tau_i + \varepsilon_{ij}.$$

avec τ_i le traitement ($\tau_i = 0$ pour le traitement B, et $\tau_i = 1$ pour le A) et $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ correspondent aux résidus de l'observation j.

Les résultats principaux sont présentés dans les Tables 2 et 3. On observe que le traitement est considéré comme un effet significatif. A priori, on ne peut donc pas rejeter l'hypothèse nulle. Ceci n'est pas surprenant étant donné la différence déjà bien marquée sur le panel (a) de la Figure 1. Cependant, ce modèle n'est pas très réaliste car il ne tient pas compte du fait que les sujets sont corrélés avec eux même dans le temps. Cet aspect sera pris en compte dans tous les modèles qui suivront.

Enfin, notons que ce modèle possède un score de BIC=1798.3 et un score de AICc=1789.2.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	279.18	5.899	<0.0001	267.54	290.83
Traitement A	38.534	8.368	<0.0001	22.011	55.057

Table 2: Estimations des effets fixes du modèle 1

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
Résidus σ	2888.12	317.97	<0.0001

Table 3: Estimations des paramètres de covariance du modèle 1

3.2 Modèle avec effet aléatoire

Comme première amélioration au modèle précédent, la corrélation présente entre les mêmes sujets au cours du temps de l'expérience est prise en compte. Pour cela, un effet aléatoire est ajouté au modèle précédent. L'équation devient alors :

$$Y_{ij} = \beta_0 + \beta_1 \tau_i + s_j + \varepsilon_{ij}.$$

avec s_j l'effet aléatoire associé au sujet j où $s_j \sim \mathcal{N}(0, \sigma_s^2)$.

Premièrement, les résultats obtenus pour l'estimation des effets fixes sont présentés dans la Table 4. A nouveau, le traitement est considéré comme un effet significatif (au seuil $\alpha = 0.05$). Les estimations des valeurs pour les coefficients sont proches de celles du premier modèle. Toutefois, les erreurs types sont plus élevées et les intervalles de confiance sont plus larges.

Ensuite, la Table 5 présente les résultats concernant les estimations des paramètres de covariance. Nous observons que le test de Wald confirme que ceux-ci sont significatifs. D'une part, les résidus sont bien inférieurs à ceux du premier modèle, ce qui traduit une meilleur capacité de ce modèle à expliquer les observations. D'autre part, la variance estimée pour les sujets est fort élevée.

Enfin, ce modèle possède les scores suivant : BIC=1761.9 et AICc=1758.6. Ces valeurs sont inférieures , et donc meilleures, à celles du premier modèle. Ces résultats ne sont pas surprenant étant donné que ce modèle représente mieux à la réalité des données.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	280.32	11.139	<0.0001	256.71	303.94
Traitement A	36.371	15.762	0.0224	5.222	67.52

Table 4: Estimations des effets fixes du modèle 2

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
Subject σ_s	900.62	372.53	0.0078
Résidus σ	1980.79	230.97	<0.0001

Table 5: Estimations des paramètres de covariance du modèle 2

3.3 Modèle avec structure de covariance AR(1)

Plutôt que de modéliser la covariance des données à travers un effet aléatoire, ce modèle impose une structure prédéfinies à la matrice des résidus. Initialement, il aurait été intéressant de considérer une structure non spécifiée afin d'observer les estimations des paramètres de covariance et ainsi identifier quelle structure pourrait mieux convenir par la suite. Malheureusement, il n'est alors pas possible d'estimer les paramètres du modèle obtenu car la matrice Hessienne obtenue n'est pas définie positive. C'est pourquoi, une structure de type AR(1) est considérée à la place. Celle-ci considère une corrélation de la forme :

$$corr(\varepsilon_{k'}, \varepsilon_k) = \rho^{|k'-k|}$$

Cette structure est particulièrement pertinente dans ce cas étant donné qu'elle modélise une corrélation temporelle : plus les observations sont éloignées dans le temps, moins elles sont corrélées. C'est exactement la situation des sujets dont les mesures sont prises au cours des 10 jours de l'expérience.

Tout d'abord, la Table 6 présente les estimations obtenues pour les effets fixes. Par rapport au premier et au second modèles, les valeurs des coefficients sont proches, les erreurs types plus élevées et les intervalles plus larges. De plus, l'effet du traitement est rejeté au seuil de 5%, ce qui n'était pas le cas des deux modèles précédents.

En ce qui concerne les paramètres de covariance, les résultats de ceux-ci sont reportés dans la Table 7. Les valeurs obtenues des tests de Wald confirment la pertinence des estimations. Concernant la variance des résidus, celle-ci est encore plus élevée que celle du premier modèle. L'ajout de variables explicatives devrait parer à ce problème.

Enfin, ce modèle obtient un BIC=1659.6 et AICc=1656.2 ce qui est une amélioration significative par rapport aux deux modèles précédents. En effet, cela signifie que ce modèle capture mieux la réalité des données pour une complexité (nombre de paramètres) légèrement supérieur au premier modèle et égale à celle du second. En somme, ce modèle sera dorénavant le modèle de base sur lequel seront effectués les

comparaisons suivantes.

En outre, d'autres structures existent pour modéliser la corrélation entre les données. A titre d'exemple, une structure plus complexe que AR(1) est celle de Toeplitz. Celle-ci considère que les éléments des diagonales descendantes de gauche à droite sont égaux. La même procédure a été conduite en considérant cette structure, cependant elle n'a pas été retenue car les résultats obtenus sont inférieures (BIC=1668.9, AICc=1660.2). Cela signifie qu'une structure plus complexe que AR(1) n'est pas nécessaire pour obtenir un modèle bien ajusté aux données.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	282.33	14.268	<.0001	252.08	312.57
Traitement A	36.834	20.209	0.0871	-6.0079	79.676

Table 6: Estimations des effets fixes du modèle 3

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
AR(1) ρ	0.815	0.044	<0.0001
Résidus σ	3368.05	737.52	<0.0001

Table 7: Estimations des paramètres de covariance du modèle 3

3.4 Modèle avec l'âge comme effet fixe

A présent, la variable `age` est incluse comme effet fixe dans le modèle précédent. Ce choix est motivé par l'analyse visuelle réalisée dans la Section 2. En effet, il semblerait que cet effet aie un impact sur le temps de réaction.

En premier lieu, considérons les résultats obtenus pour les effets fixes dans la Table 8. Comme dans le modèle précédent, le traitement n'est pas considéré comme un effet significatif. D'autant plus que la p-valeur obtenue est largement plus élevée. En revanche, l'effet estimé de l'âge est très significatif quant à lui. Ce modèle suggère donc que l'âge seul, et pas le traitement, est nécessaire pour prédire le temps de réaction d'un individus. De plus, le signe positif indique que le temps de réaction est plus élevé à mesure que le sujet est âgé, ce qui est consistent avec nos attentes.

En second lieu, comme précédemment les estimations des paramètres de covariance sont pertinentes comme il peut être observé dans la Table 9. Sans surprise, la variance estimées des résidus est moindre avec l'ajout d'une variable explicative.

Par ailleurs, ce modèle possède un BIC=1649.9 et un AICc=1645.4 ce qui est une amélioration par rapport au modèle précédent. Cette amélioration est obtenne grâce à l'inclusion d'un seul paramètre additionnel (l'âge) ce qui est une faible différence de complexité.

Enfin, un modèle basé sur celui-ci avec en plus l'inclusion de l'interaction entre `age` et `treatment` fut aussi considéré mais pas retenu car ses scores étaient inférieurs (BIC=1652.7 et AICc=1647.9). Par conséquent, ce quatrième modèle prenant en compte la variable `age` sera utilisé comme modèle de base pour construire le prochain modèle.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	203.26	21.617	<.0001	157.18	249.33
Traitement A	9.985	16.236	0.5478	-24.621	44.59
Age	1.828	0.442	0.0009	0.886	2.77

Table 8: Estimations des effets fixes du modèle 4

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
AR(1) ρ	0.724	0.057	<0.0001
Résidus σ	2286.44	435.30	<0.0001

Table 9: Estimations des paramètres de covariance du modèle 4

3.5 Modèle avec le jour comme effet fixe

Dans la même logique qui nous a incité à inclure l'effet de l'âge dans la sous-section précédente, nous considérons l'inclusion de la variable `Days` dans le modèle. En effet, le panel (a) de la Figure 2 semblait indiquer une relation linéaire entre les jours et le temps de réaction des patients.

Bien que cette variable ne possède que dix valeurs, elle n'est pas considérée comme une variable catégorielle. D'une part, car cela a plus de sens étant donné qu'il n'y a pas réellement un nombre de jour limité. D'autre part, car cela simplifie le modèle et qu'un ajout de complexité trop élevé n'est pas désirable. Rappelons que certaines données sont manquantes pour certains jours chez quelques individus. Le traitement choisi pour permettre à SAS d'en tenir compte est de créer un dupliqua catégorielle de la variable en question, et d'inclure cette nouvelle variable dans REPEATED. La variable continue quant à elle est incluse dans MODEL comme il se doit.

Premièrement, la Table 10 offre un aperçu des résultats obtenus pour les effets fixes. Comme attendu, le nombre de jours est considéré comme un effet pertinent pour prédire le temps de réaction, au même titre que l'âge des sujets de l'expérience. Le signe positif du coefficient indique bien que le temps de réaction augmente à mesure que le temps passe comme cela était attendu. A nouveau, l'effet du traitement est rejeté. Nous observons tout même une baisse de l'estimation de l'ordonnée, probablement causée par la valeur relativement importante du coefficient associé à la variable `Days`.

Ensuite, la Table 11 présente un résumé des résultats obtenus pour les paramètres de covariance. Nous observons que la variance estimée pour les résidus est largement inférieure à celle du modèle précédent. Nous pouvons en déduire que la variable `Days` explique particulièrement bien les observations faites sur la variable réponse. Ceci devrait en théorie se traduire par des bons scores BIC et AICc.

Clairement, ce modèle est supérieur au précédent. Les scores respectifs sont BIC=1597.9 et AICc=1593.1, meilleurs donc que ceux du quatrième modèle. Ce modèle est donc selectionné comme le nouveau modèle de base.

En outre, l'ajout de l'interaction entre les variables `Treatment` et `Days` fut considéré mais rejeté car ses performances sont moindres (BIC=1600 , AICc=1594.4).

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	159.07	13.704	<0.0001	129.86	188.28
Traitement A	12.891	9.393	0.1901	-7.13	32.913
Age	1.731	0.258	<0.0001	1.189	2.28
Days	10.425	1.217	<0.0001	8.02	12.83

Table 10: Estimations des effets fixes du modèle 5

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
AR(1) ρ	0.558	0.067	<0.0001
Résidus σ	1158.79	166.74	<0.0001

Table 11: Estimations des paramètres de covariance du modèle 5

3.6 Modèle avec interaction entre l'âge et le jour

La Figure 4 ci-dessous, présente l'évolution du temps de réaction en fonction du jour où la mesure a été faites. De plus, les mesures appartenant à des individus du même âge sont groupées et une couleur leur

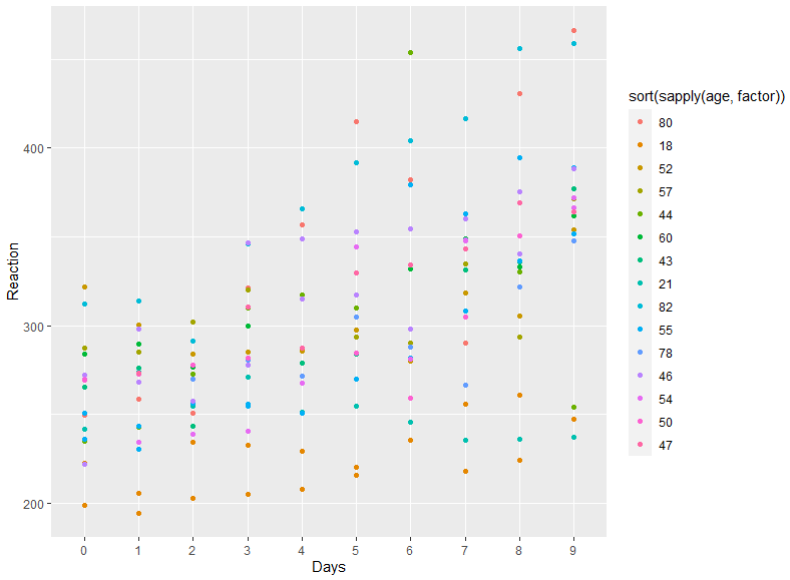

Figure 4: Evolution de `Reaction` en fonction de `Days` pour chaque groupe d'âge de la variable `age`.

est associée. A première vue, il semblerait ue les pentes de chaque groupe d'âge sont différentes. Cette observation va donc être analysée en considérant un terme d'interaction entre les variables `Days` et `age`.

Tout d'abord, les résultats concernant les effets fixes sont présentés dans la Table 12. Le traitement est encore une fois rejeté comme effet significatif. Par ailleurs, les variables `age` et `Days` sont également rejetées lorsque leur interaction est considérée, qui est quant à lui un effet significatif. Cela peut se traduire simplement par une bonne capture de l'effet de base par l'ordonnée. Ainsi, le jour et l'âge jouent tout les deux un rôle majeur dans les observations faites sur le temps de réaction. Notons tout de même que l'estimation du coefficient de leur interaction est relativement faible. Cela ne doit pas trop surprendre étant donné que les valeurs des observations de l'interaction sont quant à elles fort élevées. En outre, le coefficient associé à l'interaction est positif ce qui n'est pas surprenant étant donné les obsevations faites sur la Figure 4. Remarquons aussi le changement de signe du coefficient de la variable `age`.

En second lieu, la Table 13 montre que l'estimation de la variance des résidus est encore réduite par rapport au modèle précédent. Il en est de même pour le paramètre de corrélation ρ qui décroît systématiquement à chaque nouveau modèle de base.

En dernier lieu, les scores obtenus sont les suivants : BIC=1581.9, AICc=1576.3. En conséquence, ce modèle est considéré comme le nouveau modèle de base.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	220.93	17.41	<0.0001	183.80	258.06
Traitement A	13.77	7.754	0.096	-2.757	30.297
Age	0.478	0.341	0.1811	-0.248	1.203
Days	-3.16	3.091	0.3082	-9.269	2.949
Days*Age	0.271	0.058	<0.0001	0.157	0.386

Table 12: Estimations des effets fixes du modèle 6

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
AR(1) ρ	0.472	0.0716	<0.0001
Résidus σ	931.84	123.39	<0.0001

Table 13: Estimations des paramètres de covariance du modèle 6

3.7 Modèle avec corrélation différente selon le traitement

Pour terminer, rappelons que l'intérêt de cette étude est de vérifier l'impact du traitement en particulier sur le temps de réaction des individus. Jusqu'à présent, son effet a été rejeté par la plupart des modèles considérés. Toutefois, cet effet est peut-être tout de même significatif dans la population mais pas dans nos modèles car ceux-ci ne sont pas assez réaliste. Par exemple, il se pourrait que les individus appartenant à un groupe de traitement différent s'associent d'une manière où d'une autre (dû à un manquement au niveau du protocole de l'expérience) ce qui impliquerait un effet différent sur les deux groupes. Un tel effet peut être explicitement estimé en considérant une corrélation différentes pour ces deux groupes, ce qui n'a pas été fait dans les modèles précédents.

Pour commencer, la Table 14 présente les résultats pour les effets fixes. La p-valeur obtenue pour le traitement est encore plus élevée que celle du sixième modèle. La conclusion que le traitement n'est pas significatif semble donc fort plausible. A part cela, très peu de changements sont observés par rapport au modèle précédent.

Ensuite, la Table 15 présente les résultats pour les paramètres de covariance. Tout d'abord, l'hypothèse selon laquelle la corrélation est différentes entre les deux groupes est confirmée par les résultats du test de Wald et les estimations différentes pour ρ_A et ρ_B (0.287 pour le groupe A et 0.701 pour le groupe B). Ainsi, l'effet d'appartenir à un groupe de traitement semble être un effet plus significatif que le traitement lui même.

Pour terminer, les scores sont les meilleurs obtenus de tous les modèles considérés : BIC=1570.0, AICc=1563.1. Cela renforce encore une fois l'hypothèse de départ selon laquelle les individus appartenant à un groupe du traitement sont corrélés entre eux mais pas de la même façon selon le groupe.

Effet	Valeur du coefficient	SE	p-valeur	Limite inférieure	Limite supérieure
Ordonnée	221.65	18.061	<0.0001	183.16	260.15
Traitement A	12.671	8.296	0.1475	-5.011	30.353
Age	0.477	0.346	0.1883	-0.261	1.215
Days	-3.032	3.079	0.3264	-9.118	3.054
Days*Age	0.269	0.0573	<0.0001	0.155	0.382

Table 14: Estimations des effets fixes du modèle 7

Paramètre de cov.	Estimation	SE	Pr>Z (Wald test)
AR(1) Treatment (A) ρ_A	0.287	0.116	0.0132
Résidus Treatment (A) σ_A	1019.87	171.38	<0.0001
AR(1) Treatment (B) ρ_B	0.701	0.077	<0.0001
Résidus Treatment (B) σ_B	865.8	208.73	<0.0001

Table 15: Estimations des paramètres de covariance du modèle 7

4 Résultats

Dans cette section, un résumé des performances des modèles et une analyse du meilleur modèle sont présentés.

La Table 16 reprend l'ensemble des scores obtenus pour chacun des modèles considérés. Une nette amélioration est observée entre le premier modèle, le plus basique, et le dernier, le plus complexe. Cet addition de complexité est justifiée par un bien meilleur ajustement aux observations. Dans la suite, nous concentrons l'analyse sur ce dernier modèle étant donné qu'il est le meilleur selon les critères choisis.

La Figure 5 est obtenue sur base du dernier modèle. Elle présente trois graphes afin de vérifier les hypothèses classiques d'un modèle de régression suivant une loi normale. D'abord, les panels (a) et (b) permettent de confirmer l'hypothèse de départ selon laquelle la variable réponse suit une loi normale. En effet, l'histogramme du panel (a) imite assez bien une la courbe d'une loi normale et le diagramme QQ du

Modèle.	BIC	AICc
Anova	1798.3	1789.9
+ Effet aléatoire	1761.9	1758.6
+ AR(1)	1659.6	1656.2
+ âge	1649.9	1645.4
+ jour	1600	1594.4
+ interaction	1581.9	1576.3
+ diff. corr.	1570	1563.1

Table 16: Valeurs des critères BIC et AICc obtenus pour chaque modèles.

panel (b) montre un alignement très proche de la droite théorique. Ensuite, le graphe présenté au panel (c) montre que les résidus semblent avoir une variante constante (propriété d'homoscédasticité). En conséquence, il semblerait que les hypothèses soient respectées. Pour s'en assurer, il serait préférable d'utiliser des tests plus formels mais nous nous contenterons de cette analyse visuelle par soucis de simplicité.

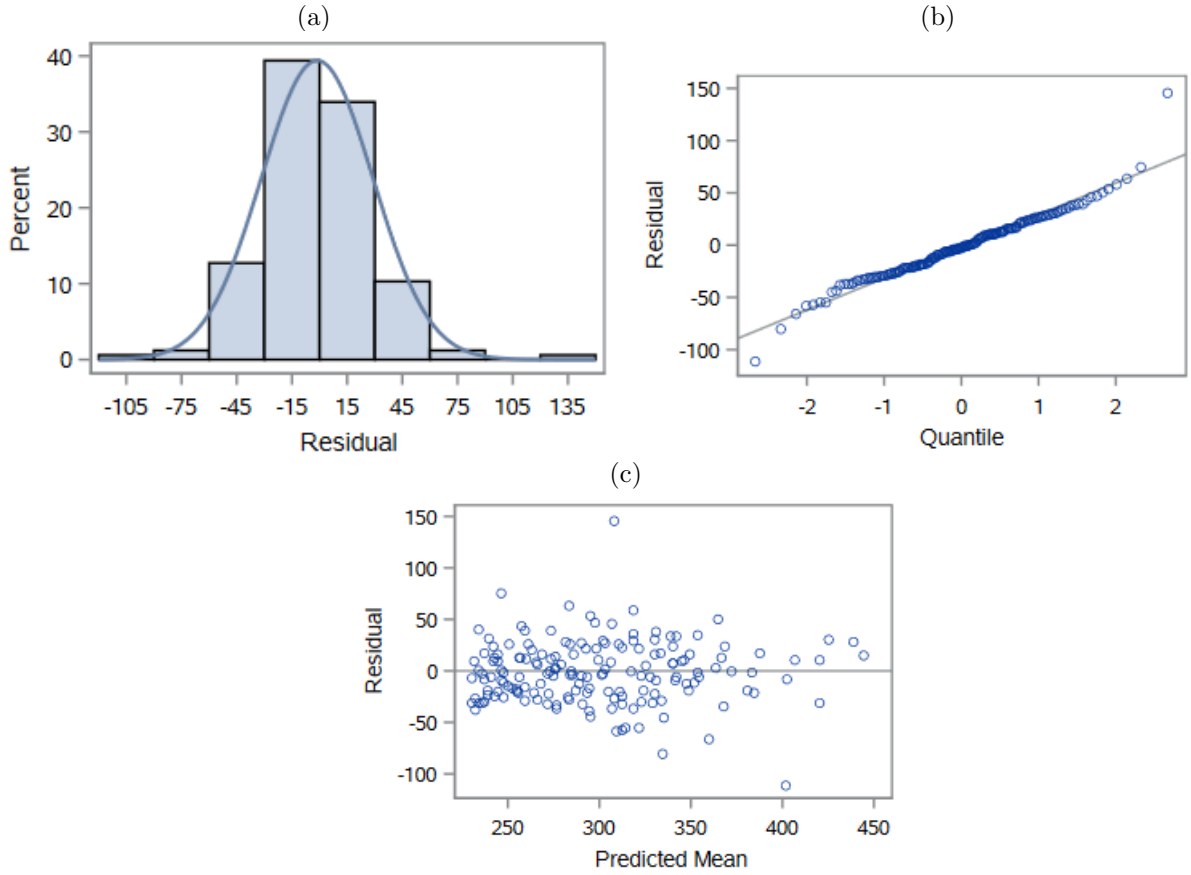

Figure 5: Vérification visuelle des hypothèses classiques : normalité des résidus et homoscédasticité (a) Histogramme des résidus (b) Diagramme QQ des résidus (c) Résidus vs. moyennes prédites.

5 Conclusion

Tout au long de ce rapport, sept modèles ont été présentés et pour chacun d'entre eux leurs paramètres ont été estimés. Ceci nous a permis de les comparer selon les critères BIC et AICc, équilibrant les aspects de complexité et de bon ajustement aux données. Au final, la plupart, en particulier les meilleurs, des modèles arrivent tous à la conclusion que le traitement n'a pas d'effet sur le temps de réaction des individus. En effet, il semblerait que l'âge des personnes ainsi que le nombre de jours depuis le début de la réduction du temps de sommeil jouent un rôle dominant pour expliquer l'évolution du temps de réaction. En outre, il

semblerait que les individus appartenant au même groupe de traitement soient corrélés entre eux et que cette corrélation soit différente entre les deux groupes. Pour expliquer cette dernière observation, une analyse du protocole d'expérience devrait être effectuée. Enfin, le dernier modèle considéré, le plus complet, semble respecter les hypothèses classiques comme il se doit.

6 Annexe

6.1 Code SAS

```
/* Generated Code (IMPORT) */
/* Source File: sleepstudy_treat_miss.csv */
/* Source Path: /home/u44436243 */
/* Code generated on: 1/2/21, 4:57 PM */

%web_drop_table(WORK.SLEEP);


FILENAME REFFILE '/home/u44436243/sleepstudy_treat_miss.csv';

PROC IMPORT DATAFILE=REFFILE
  DBMS=CSV
  OUT=WORK.SLEEP;
  GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.SLEEP; RUN;

%web_open_table(WORK.SLEEP);

/* copy Days to use when Days is considered not categorial */
DATA SLEEP;
SET SLEEP;
DaysC = put(Days,2.);
RUN;



/* Modèle 1 : ANOVA pour le traitement */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest;
title 'Modele 1';
CLASS treatment;
MODEL Reaction = treatment  / SOLUTION OUTP=previsions RESIDUAL CL;
RUN;
ODS graphics off;


/* Modèle 2 : Subject comme effet aléatoire */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 2';
CLASS Subject treatment;
MODEL Reaction = treatment / SOLUTION OUTP=previsions RESIDUAL CL;
RANDOM Subject / G;
REPEATED / TYPE = SIMPLE SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;


/* Modèle 3 : AR(1)*/
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 3';
CLASS Subject treatment;
MODEL Reaction = treatment / SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED / TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;

/* Modèle 3 bis : TOEP*/
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 3 bis';
CLASS Subject treatment;
MODEL Reaction = treatment / SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED / TYPE = TOEP  SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;

/* Modèle 3bisbis: UN */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 3bis';
CLASS Subject treatment;
MODEL Reaction = treatment / SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED / TYPE = UN SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;
```

```
/* Modèle 4: + age */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 4';
CLASS Subject treatment;
MODEL Reaction = treatment age/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED / TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;


/* Modèle 4bis : interaction age et traitement rejetée*/
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 4bis';
CLASS Subject treatment;
MODEL Reaction = treatment age age*treatment/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED / TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;


/* Modèle 5 : + Days*/
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 5';
CLASS Subject treatment DaysC;
MODEL Reaction = treatment age Days/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED DaysC/ TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;

/* Modèle 5bis : Days treatment interaction rejetée */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 5bis';
CLASS Subject treatment DaysC;
MODEL Reaction = treatment age Days Days*treatment/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED DaysC/ TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;


/* Modèle 6 : + interaction age*jour */
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 6';
CLASS Subject treatment DaysC;
MODEL Reaction = treatment age Days Days*age/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED DaysC/ TYPE = AR(1) SUBJECT = Subject R RCORR;
RUN;
ODS graphics off;

/* Modèle 7 : correlation differente selon le traitement*/
ODS graphics on;
PROC MIXED data = SLEEP method = ml covtest IC;
title 'Modele 7';
CLASS Subject treatment DaysC;
MODEL Reaction = treatment age Days Days*age/ SOLUTION OUTP=previsions RESIDUAL CL;
REPEATED DaysC/ TYPE = AR(1) SUBJECT = Subject Group=treatment R=1,11 RCORR=1,11;
RUN;
ODS graphics off;
```

6.2 Code R

```
#############################################################
#                        PROJET MLA                         #
#############################################################
library(lme4)
library(AppliedPredictiveModeling)
library(MASS)
library(tree)
library(plyr)
library(glmnet)
library(ggplot2)
library(ggpubr)
transparentTheme(trans = .5)
library(caret)

data = read.table("C:/Users/alext/OneDrive - UCL/UCL/DATA/Q3/LSTAT2210 - Modles Linaires Avancs/Projet
    /sleepstudy_treat_miss.csv", header=TRUE, sep = ",")
data$Subject <- sapply(data$Subject, factor)
#data$age <- sapply(data$age, factor)
data$Days <- sapply(data$Days, factor)
```

```r
data$treatment <- sapply(data$treatment, factor)
y = data$Reaction
qqnorm(y)
qqline(y)
hist(y, breaks = 10)



# 1) Effet du traitement sur le temps de  r action
ggplot(data = data, mapping = aes(x = treatment, y = Reaction)) +
  geom_point(alpha = 1, aes(color = treatment))  + geom_boxplot(alpha = 0)

# 2) Effet du temps (nbr de jours avec peu de sommeil) sur le temps de  r action
ggplot(data = data, mapping = aes(x = Days, y = Reaction)) +
  geom_point(alpha = 1, aes(color = treatment))  + geom_boxplot(alpha = 0)

# 3) Effet de l'age sur le temps de  r action
ggplot(data = data, mapping = aes(x = age, y = Reaction)) +
  geom_point(alpha = 1, aes(color = treatment))

# 4) Age des individus dans chacun des groupes de traitement
ggplot(data = data, mapping = aes(x = treatment, y = age)) +
  geom_point(alpha = 1, aes(color = treatment)) + geom_boxplot(alpha = 0)

# 5) Evolution du temps de  r action  durant le temps pour chaque groupe d'age d'individus.
ggplot(data = data, mapping = aes(x = Days, y = Reaction)) +
  geom_point(alpha = 1, aes(color = sort(sapply(age, factor))))
```