Multivariate Analysis of Variance
(MANOVA)

Julien Parfait Bidias Assala

2024-09-01

Background

Problème

Dans une étude médicale, un chercheur souhaite examiner l'effet d'un traitement particulier sur plusieurs paramètres biologiques chez les patients atteints de diabète de type 2. Le traitement en question est une nouvelle thérapie qui pourrait potentiellement améliorer le contrôle glycémique, la tension artérielle et la fonction rénale chez ces patients. Pour évaluer l'efficacité du traitement, les patients sont divisés en deux groupes : un groupe de traitement et un groupe témoin.

Background

Les mesures suivantes sont prises après une période de suivi de 6 mois :

1. *HbA*1*c* (%) : Un indicateur du contrôle glycémique à long terme.
2. Pression artérielle systolique (mmHg) *pa_systolique* : Un indicateur clé de la santé cardiovasculaire.
3. Créatinine sérique (mg/dL) *créatinine_s*: Un marqueur de la fonction rénale.

Background

Comment un modèle MANOVA peut résoudre cette Problématique ?

L'utilisation d'une ANOVA à un facteur pour chaque variable pourrait montrer si le traitement a un effet sur chaque paramètre indépendamment. Cependant, cette approche ne tient pas compte des corrélations potentielles entre les différentes variables dépendantes. Par exemple, une amélioration de la fonction rénale pourrait être liée à une meilleure gestion de la glycémie, et ces relations pourraient être cruciales pour comprendre l'effet global du traitement. Tout ceci justifie le recours à un modèle plus adapté permettant de prendre en compte ces constats.

Background

Intérêt du modèle

Le modèle MANOVA (Multivariate Analysis of Variance) à un facteur permet de tester simultanément l'effet du traitement sur plusieurs variables dépendantes en tenant compte des corrélations entre elles. Il fournit une vue d'ensemble sur l'efficacité du traitement en considérant toutes les variables biologiques étudiées comme un ensemble, ce qui augmente la puissance statistique et diminue le risque d'erreurs de type I liées à des tests multiples.

Présentation des données et méthodologie

Notre base de données contient au total 60 observations (lignes) et 5 variables (colonnes). La variable **groupe** permet de classer les patients en deux catégories : ceux qui ont reçu le traitement et ceux qui ne l'ont pas reçu. L'ID permet d'identifier chaque patient, et les autres variables correspondent aux caractéristiques susmentionnées qui ont été relevées sur chaque patient après 6 mois.

Table 1: Entête de la base de données

Id	Groupe	HbA1c	pa_systolique	creatinine_s
1	Traitement	7.2	130	1.1
2	Traitement	6.8	125	1.0
3	Traitement	7.0	128	1.2
4	Traitement	6.9	132	1.1
5	Traitement	6.7	126	1.0

Méthodologie

Le modèle MANOVA que nous estimons s'écrit comme suit :

Nous avons trois variables réponses : HbA1c (Y_1), pression artérielle systolique (Y_2), et créatinine sérique (Y_3). Le modèle MANOVA peut être exprimé de la manière suivante :

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\varepsilon}_{ij}$$

où : $\mathbf{Y}_{ij} = \begin{pmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \end{pmatrix}$ représente le vecteur des observations pour les trois variables réponses pour l'individu i dans le groupe j,

Modèle MANOVA à un facteur

- $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$ est le vecteur des moyennes globales pour chacune des variables réponses,
- $\boldsymbol{\tau}_j = \begin{pmatrix} \tau_{1j} \\ \tau_{2j} \\ \tau_{3j} \end{pmatrix}$ est le vecteur des effets du groupe j pour chacune des variables réponses (c'est-à-dire la différence entre la moyenne du groupe j et la moyenne globale pour chaque variable),
- $\boldsymbol{\varepsilon}_{ij} = \begin{pmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \\ \varepsilon_{3ij} \end{pmatrix}$ est le vecteur des erreurs aléatoires associées à l'individu i dans le groupe j, avec $\boldsymbol{\varepsilon}_{ij} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \sum \right)$.

Écriture du Modèle MANOVA

Par exemple, la matrice de réponses pour tous les Groupes

$$\mathbf{Y_{ij}} = \begin{pmatrix} \mathbf{Y}_{11} & \mathbf{Y}_{12} & \mathbf{Y}_{13} \\ \mathbf{Y}_{21} & \mathbf{Y}_{22} & \mathbf{Y}_{23} \\ \mathbf{Y}_{31} & \mathbf{Y}_{32} & \mathbf{Y}_{33} \\ \mathbf{Y}_{41} & \mathbf{Y}_{42} & \mathbf{Y}_{43} \end{pmatrix}$$

où chaque colonne $\mathbf{Y}_{ij}$ représente les mesures pour un individu spécifique dans un groupe donné.

Écriture Mathématique du Modèle MANOVA

Matrice des Erreurs :

$$\varepsilon_{ij} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \\ \epsilon_{41} & \epsilon_{42} & \epsilon_{43} \end{pmatrix}$$

En résumé, la MANOVA analyse comment les vecteurs des réponses $\mathbf{Y}_{ij}$ varient entre les groupes, en tenant compte simultanément de toutes les variables réponses.

Technique d'estimation

Une fois les paramètres estimés, l'analyse MANOVA consiste à tester les hypothèses suivantes :

- **Hypothèse nulle (H_0)** : Les moyennes des variables réponses sont égales entre les groupes.
- **Hypothèse alternative (H_A)** : Au moins un groupe a une moyenne différente.

Pour tester ces hypothèses, on utilise diverses statistiques de test basées sur les matrices de variance-covariance. Par exemple la statistique de Pillai qui est une mesure de la distance multivariée entre les groupes, la statistique de Wilks, la statistique de Hotelling-Lawley et la statistique de Roy.

Tableau d'analyse de la variance

Table 2: Tableau d'Analyse de la Variance Multivariée

Source	**Df**	**Test stat**	**F-Stat**	**Num Df**	**Den Df**	**p-value**
Groupe	1	Valeur	Valeur	Valeur	Valeur	Valeur

Avec Source : la source de variation dans le modèle, ici **Groupe**. Df : Degrés de liberté associés à la source de variation. Test Stat : Statistique de test (Pillai's Trace, etc.). F-stat: Valeur de la statistique F pour le test de MANOVA. Num Df et Den Df : Degrés de liberté du numérateur et du dénominateur. p-value : Valeur p associée au test.

Hypothèses de bases pour le modèle MANOVA

Hypothèses de base du modèle :

- Chaque participant doit être assigné à un seul groupe, sans que les observations d'un groupe soient liées à celles d'un autre. Les mesures répétées sur les mêmes individus ne sont pas autorisées. Les échantillons doivent être choisis de manière totalement aléatoire : c'est l'hypothèse d'indépendance des observations.
- Taille d'échantillon adéquate : le nombre d'observations n dans chaque groupe doit être supérieur au nombre de variables-réponses (variables dépendantes).
- Absence de valeurs aberrantes.
- Normalité multivariée.

Hypothèses de bases pour le modèle MANOVA

Hypothèses de base du modèle (suite) :

- Absence de multicolinéarité entre les variables de réponses (variables dépendantes). Elle ne doivent pas être trop corrélées les unes aux autres.
- Linéarité entre toutes les variables de résultat (variables explicatives) pour chaque groupe.
- Homogénéité des variances.
- Homogénéité des matrices de variance-covariance

Résultats des Analyses descriptives

Table 3: Statistiques descriptives

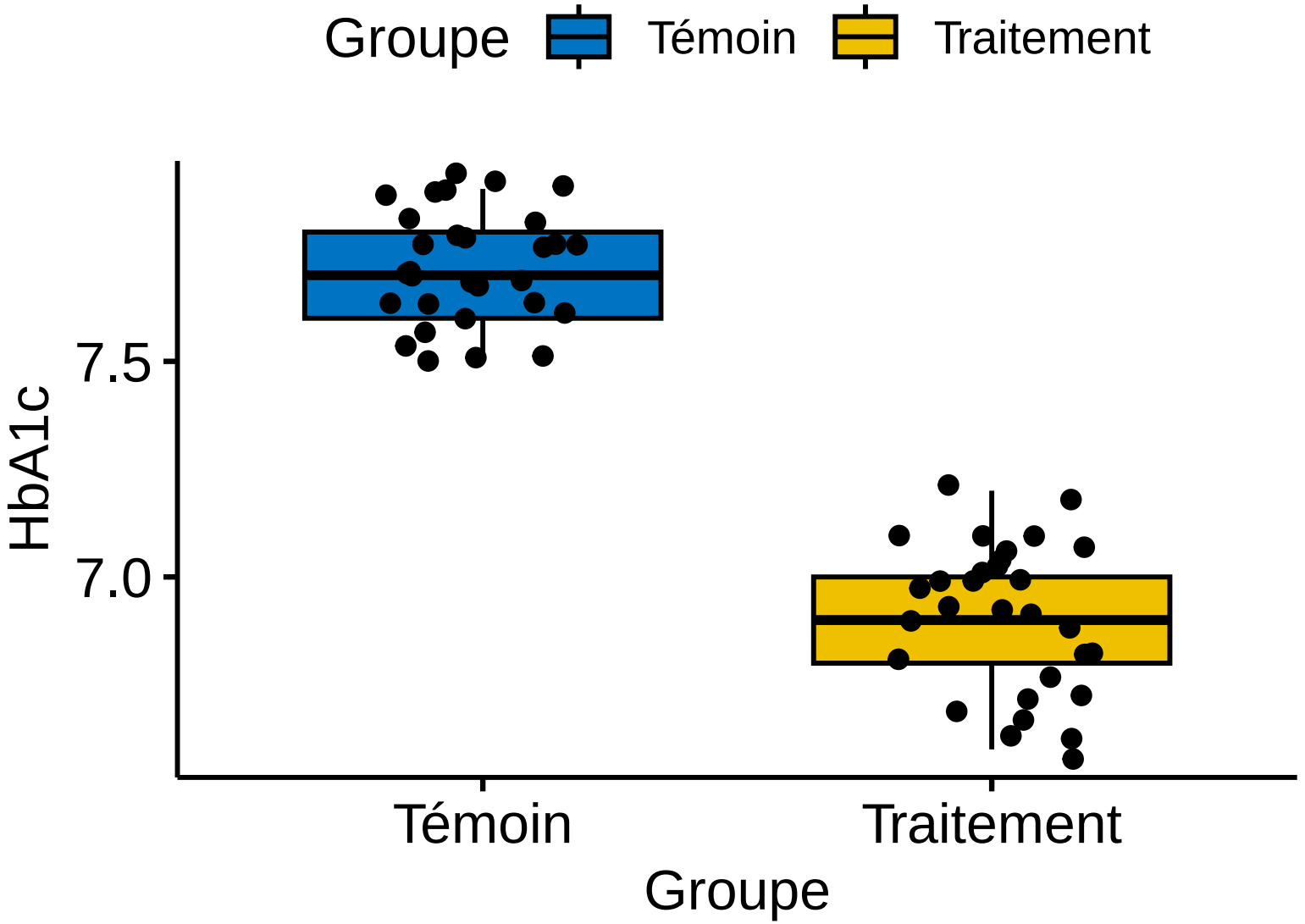
Groupe	variable	n	min	max	median	mean	sd
Témoin	HbA1c	30	7.5	7.9	7.7	7.72	0.14
Témoin	pa_systolique	30	135.0	145.0	140.0	140.00	2.88
Témoin	creatinine_s	30	1.3	1.6	1.4	1.43	0.11
Traitement	HbA1c	30	6.6	7.2	6.9	6.91	0.18
Traitement	pa_systolique	30	125.0	132.0	129.0	128.67	2.09
Traitement	creatinine_s	30	1.0	1.2	1.1	1.09	0.08

Commentaires

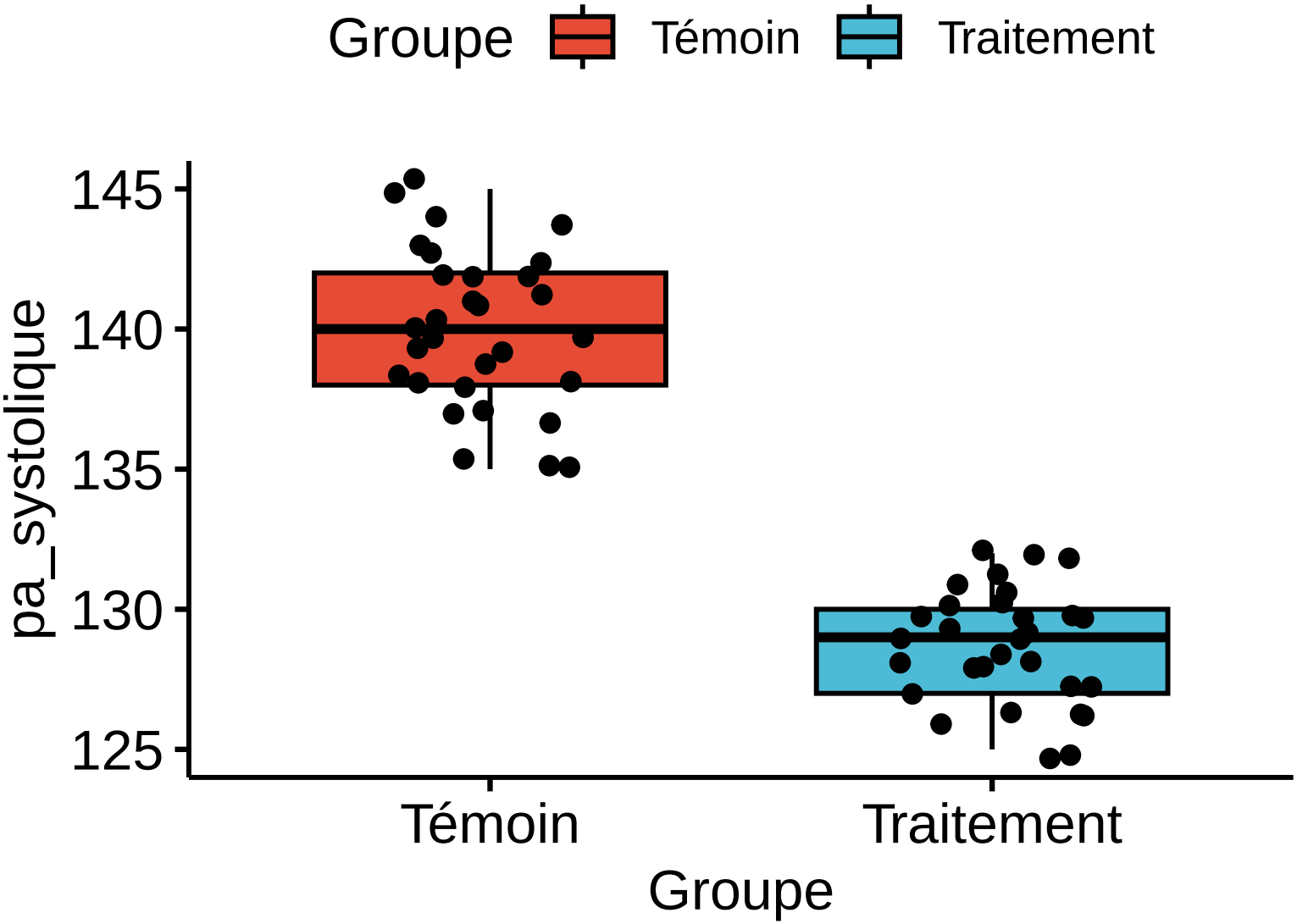
Pour chaque variable prise dans chaque groupe, on peut constater que le nombre d'observations est effectivement de 60. Nous avons ainsi un nombre d'observations dans chaque groupe supérieur au nombre de variables de réponse (3 variables), ce qui permet de vérifier l'hypothèse d'indépendance des observations. L'analyse de la médiane et de la moyenne pour chaque groupe révèle des valeurs relativement proches.

En conséquence, nous avons une présomption de symétrie des distributions de chaque variable, quel que soit le groupe, et le problème des outliers uni-variés ne se pose donc pas. Enfin, avec des écarts-types faibles, on peut affirmer qu'il y a une faible variabilité des observations autour de la moyenne.

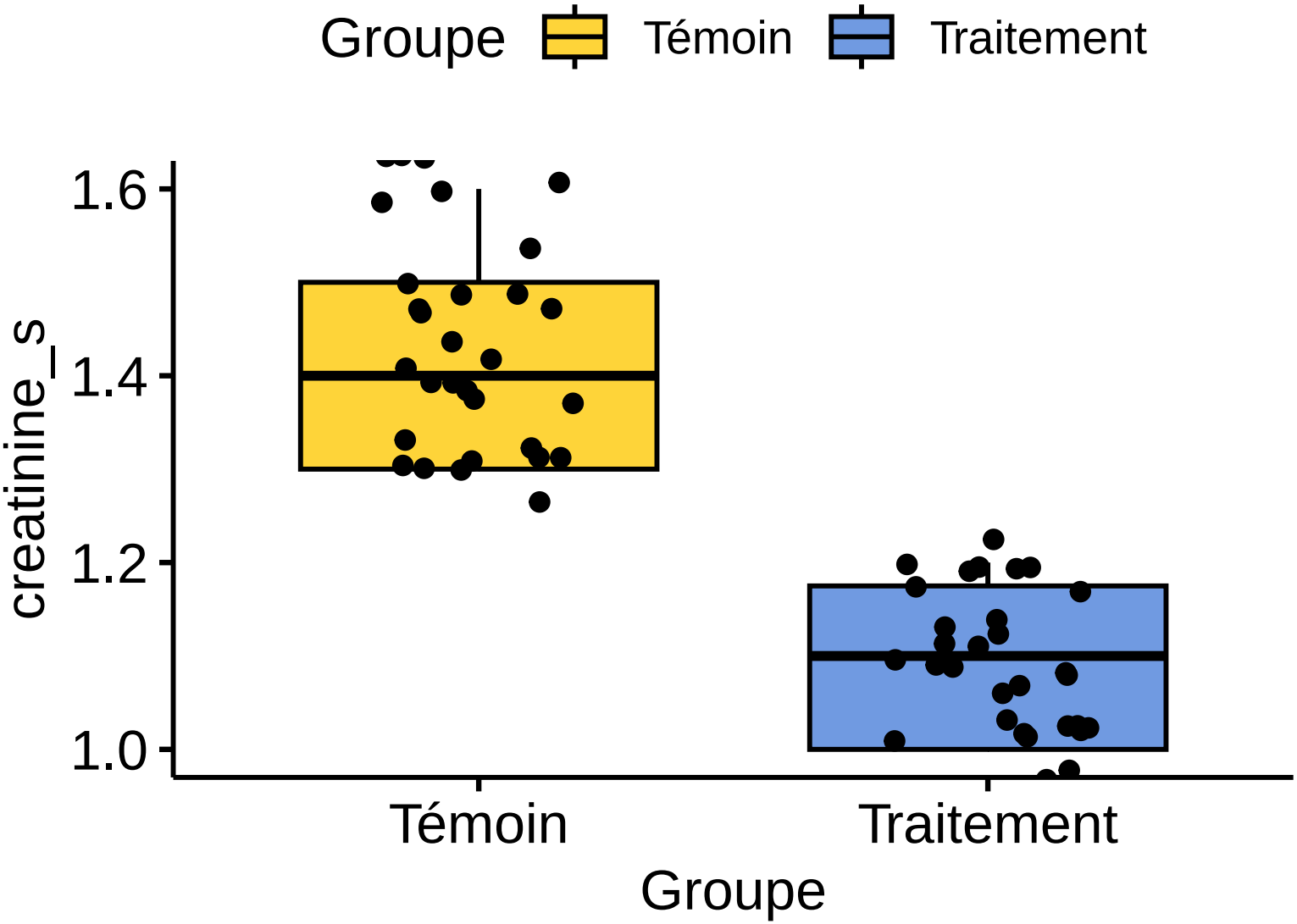
Analyses graphiques



Analyses graphiques



Box-plots et identification des valeurs extrêmes



Commentaires

- L'analyse des graphiques ci-dessus renforce l'idée qu'en plus de l'absence d'outliers uni-variés, on pourrait également s'attendre à l'absence d'outliers multivariés. Ces hypothèses sont donc vérifiées. Le traitement des outliers ne se pose pas. La dispersion des observations sur les box-plots ne révèle aucun point extrême (aucune valeur d'une variable de la base ne semble s'écarter des autres).
- Enfin, on peut noter des écarts médians considérables entre les groupes témoins et de traitement pour chaque paramètre biologique.

Normalité uni-variée

À l'exception des variables Pression artérielle systolique et HbA1c (dans le groupe traitement), qui ont une p-value supérieure au seuil de 5%, les autres variables ont des p-values inférieures à ce seuil. Ainsi, elles ne vérifient pas l'hypothèse de normalité uni-variée. Il reste à tester la normalité multivariée.

Table 4: Normalité multivariée

Groupe	variable	statistic	p
Témoin	HbA1c	0.90057	0.00867
Traitement	HbA1c	0.93925	0.08685
Témoin	creatinine_s	0.85577	0.00082
Traitement	creatinine_s	0.80700	0.00009
Témoin	pa_systolique	0.96436	0.39832
Traitement	pa_systolique	0.94918	0.16073

Normalité multivariée

Avec une p-value de 0,168 supérieur au seuil de 5% on peut conclure à l'hypothèse de normalité multivariée.

Table 5: Normalité multivariée

statistic	p.value
0.9712593	0.1681154

Test de corrélations

L'analyse des corrélations permet de tester l'hypothèse de multi-colinéarité. Tel que susmentionné, les variables dépendantes (de réponses) ne doivent pas être fortement corrélées. Une corrélation supérieure à 0,9 est une indication de la multicollinéarité, ce qui est problématique pour réaliser le modèle MANOVA. Les résultats du tableau ci-dessous révèlent qu'il n'y a pas de multicollinéarité, selon le coefficient de corrélation de Pearson ($r = 0,86, p < 0,0001$).

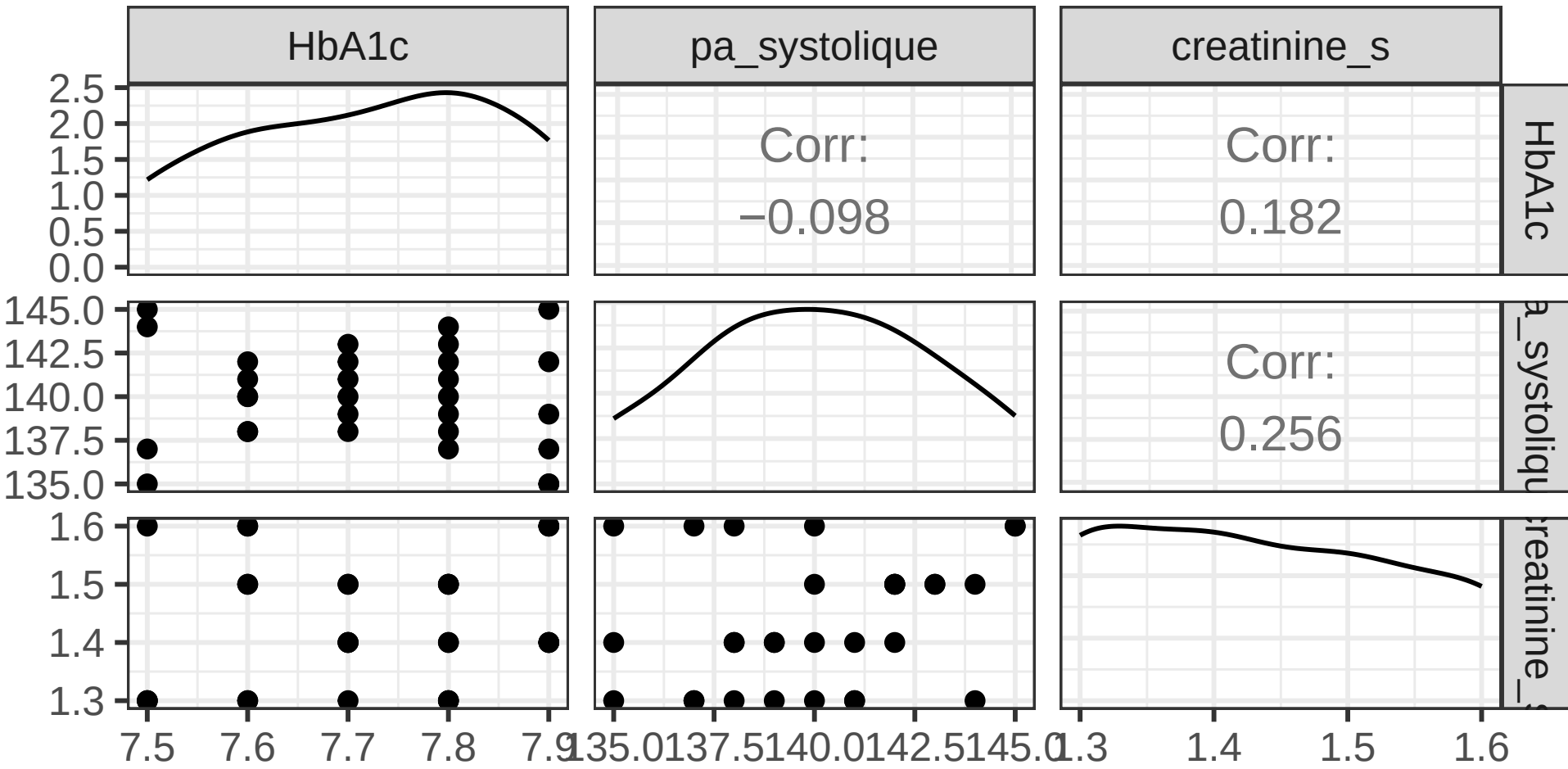
Test de corrélations

Table 6: Corrélations

var1	var2	cor	statistic	p	conf.low	conf.high	method
HbA1c	HbA1c	1.00	Inf	0	1.000	1.000	Pearson
HbA1c	pa_systolique	0.86	13.113	0	0.783	0.917	Pearson
HbA1c	creatinine_s	0.88	14.390	0	0.812	0.929	Pearson
pa_systolique	HbA1c	0.86	13.113	0	0.783	0.917	Pearson
pa_systolique	pa_systolique	1.00	Inf	0	1.000	1.000	Pearson
pa_systolique	creatinine_s	0.86	12.879	0	0.777	0.915	Pearson
creatinine_s	HbA1c	0.88	14.390	0	0.812	0.929	Pearson
creatinine_s	pa_systolique	0.86	12.879	0	0.777	0.915	Pearson
creatinine_s	creatinine_s	1.00	Inf	0	1.000	1.000	Pearson

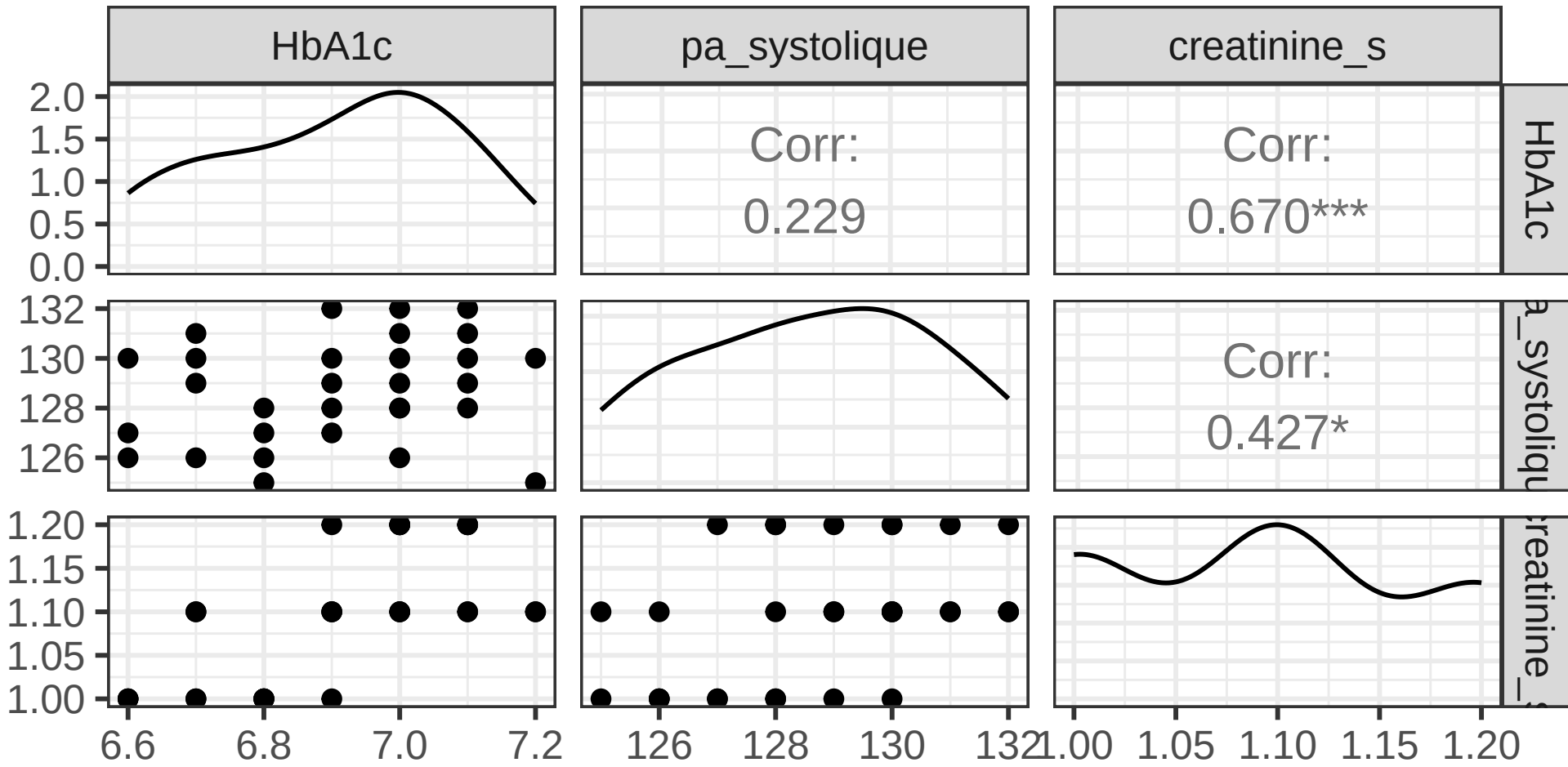
Hypothèse de linéarité

```
[[1]]
```



Hypothèse de linéarité

```
[[1]]
```



Commentaires

La relation par paire entre les variables de réponse doit être linéaire pour chaque groupe. En cas de non-linéarité, plusieurs options s'offrent à nous :

- Transformer ou supprimer les variables de réponse concernées.
- Poursuivre l'analyse malgré une légère perte de puissance.

Dans le cas présent, on observe que cette hypothèse n'est totalement vérifiée dans les deux graphes. Nous opterons donc pour la seconde option, en continuant l'analyse tout en acceptant une légère perte de puissance dans les tests.

Hypothèse d'homogénéité des covariances

La p-value associée au test de Box est de 0,0053, ce qui est inférieur au seuil de signification habituel de 5% (0,05). Cela signifie que nous rejetons l'hypothèse nulle d'homogénéité des matrices de covariance. Cette violation de l'hypothèse peut affecter la validité des résultats du modèle MANOVA, car celui-ci suppose que cette condition est remplie. Cependant, nous utilisons une méthode MANOVA plus robuste, telle que le test de Pillai.

Table 7: résultats du test

statistic	p.value	parameter	method
18.43198	0.0052387	6	Box's M-test for Homogeneity of Covariance Matrices

Hypothèse d'homogénéité de la variance

Les variables *HbA1c* et *pa_systolique* ont des p-values supérieures au seuil de 5%, ce qui indique une homogénéité des variances avec parcimonie.

Table 8: Test d'homogénéité de la variance

variable	df1	df2	statistic	p
HbA1c	1	58	2.252	0.139
creatinine_s	1	58	4.589	0.036
pa_systolique	1	58	2.688	0.107

Réalisation de MANOVA à un facteur

Les résultats du tableau d'analyse de la variance révèlent que le facteur **Groupe** est significatif dans le modèle MANOVA, avec une p-value extrêmement faible ($<$ 2.2e-16), bien en dessous du seuil de signification habituel de 0,05. Cela indique que le groupe a un effet statistiquement significatif sur les variables dépendantes prises ensemble.

La statistique de test (Pillai's Trace)) de $0,92$ (très proche de 1) suggère un effet important du groupe sur les variables dépendantes.La valeur F approximative de $213,96$, associée à une p-value quasi nulle, confirme l'importance de l'effet du groupe.

Table 9: Résultats du test MANOVA - Statistique de Pillai

Term	Df	Test.Stat	Approx.F	Num.Df	Den.Df	Pr..F.
Groupe	1	0.91976	213.96	3	56	$<$ 2.2e-16***

Conclusion

En résumé, les résultats suggèrent que le traitement a un effet positif significatif sur les paramètres biologiques. Cela renforce l'hypothèse que le traitement pourrait être efficace pour améliorer la santé des patients diabétiques. La nouvelle thérapie est donc bénéfique.