Projet Biostatistique

Kuassi Pierre DOVODJI - Evencia MICHONDARD - José BANKOLE

2024-11-12

Contents

1 Introduction

Les pathologies orthopédiques comme la hernie discale et le spondylolisthésis posent de gros défis pour la santé publique. Ils affectent la mobilité, la qualité de vie des patients et engendrent des coûts élevés pour les soins. Il est donc important de mieux comprendre les facteurs biomécaniques et anatomiques qui contribuent à ces pathologies.
Nous commencerons par des analyses exploratoires des différentes mesures effectuées, puis nous étudierons les liens entre ces mesures et les phatologies à l'aide de graphiques, de tests statistiques et de modèles.

2 Objectifs

L'objectif est d'identifier si des mesures telles que l'incidence pelvienne, la lordose lombaire ou le degré de spondylolisthésis prédisent la présence d'une pathologie diagnostiquée.

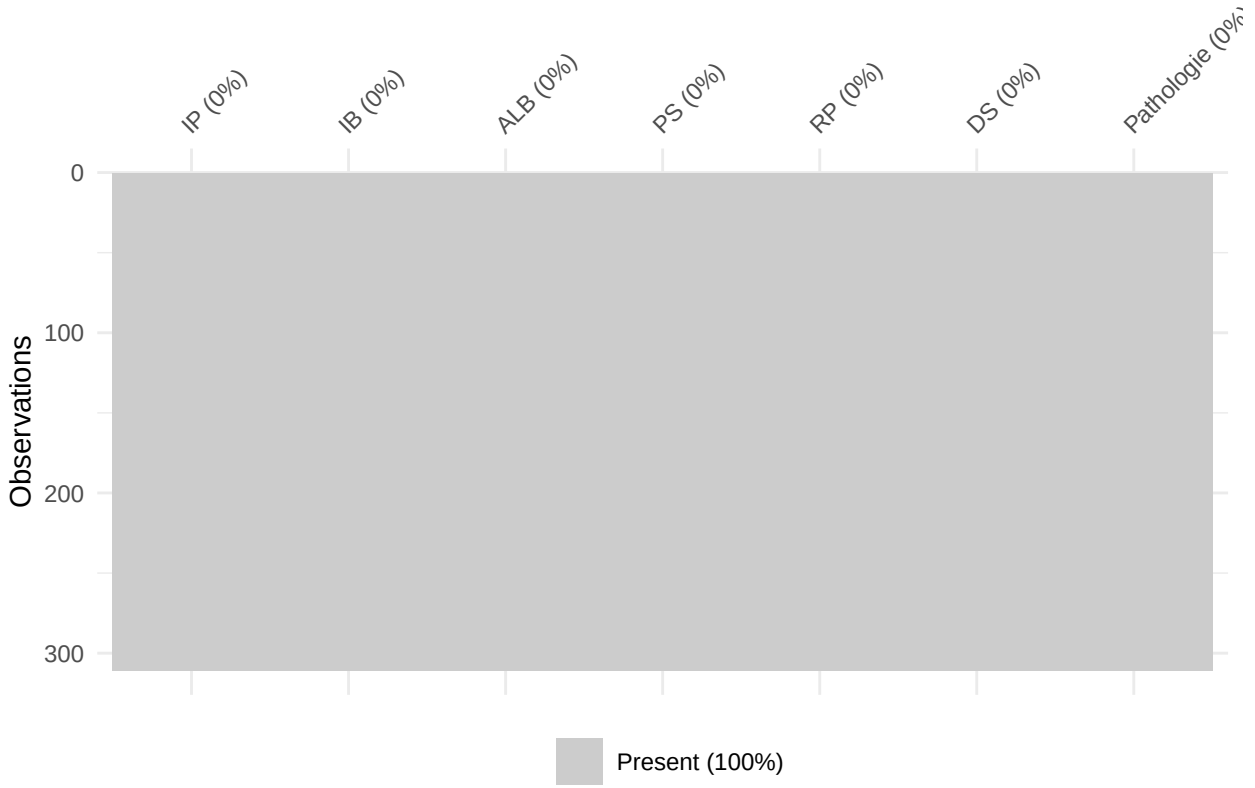
3 Description de la base de données

La base de données comporte 310 observations et 7 variables dont 6 variables quantitatives et une qualitative qui est notre variable d'intérèt.

- Les variables numériques :
  - **IP**: Incidence Pelvienne du patient
  - **IB**: Inclinaison du Bassin
  - **ALB**: Angle de Lordose Lombaire
  - **PS**: Pente Sacrée(mesure numérique qui décrit l'orientation de la partie supérieure du sacrum).
  - **RP**: Rayon Pelvien
  - **DS**: Degré de Spondylolisthesis
- Le variable qualitative :
  - **Pathologie** : Indique la classification clinique du patient ('Normal', 'Hernie', ou 'Spondylolisthesis').

```
## 'data.frame':    310 obs. of  7 variables:
##  $ IP        : num  63 39.1 68.8 69.3 49.7 ...
##  $ IB        : num  22.55 10.06 22.22 24.65 9.65 ...
##  $ ALB       : num  39.6 25 50.1 44.3 28.3 ...
##  $ PS        : num  40.5 29 46.6 44.6 40.1 ...
##  $ RP        : num  98.7 114.4 106 101.9 108.2 ...
##  $ DS        : num  -0.254 4.564 -3.53 11.212 7.919 ...
##  $ Pathologie: chr  "Hernie" "Hernie" "Hernie" "Hernie" ...
```

4 Les valeurs manquantes



L'analyse de la base de données montre qu'elle est complète et ne contient aucune valeur manquante, ce qui garantit l'intégrité des données pour les analyses.

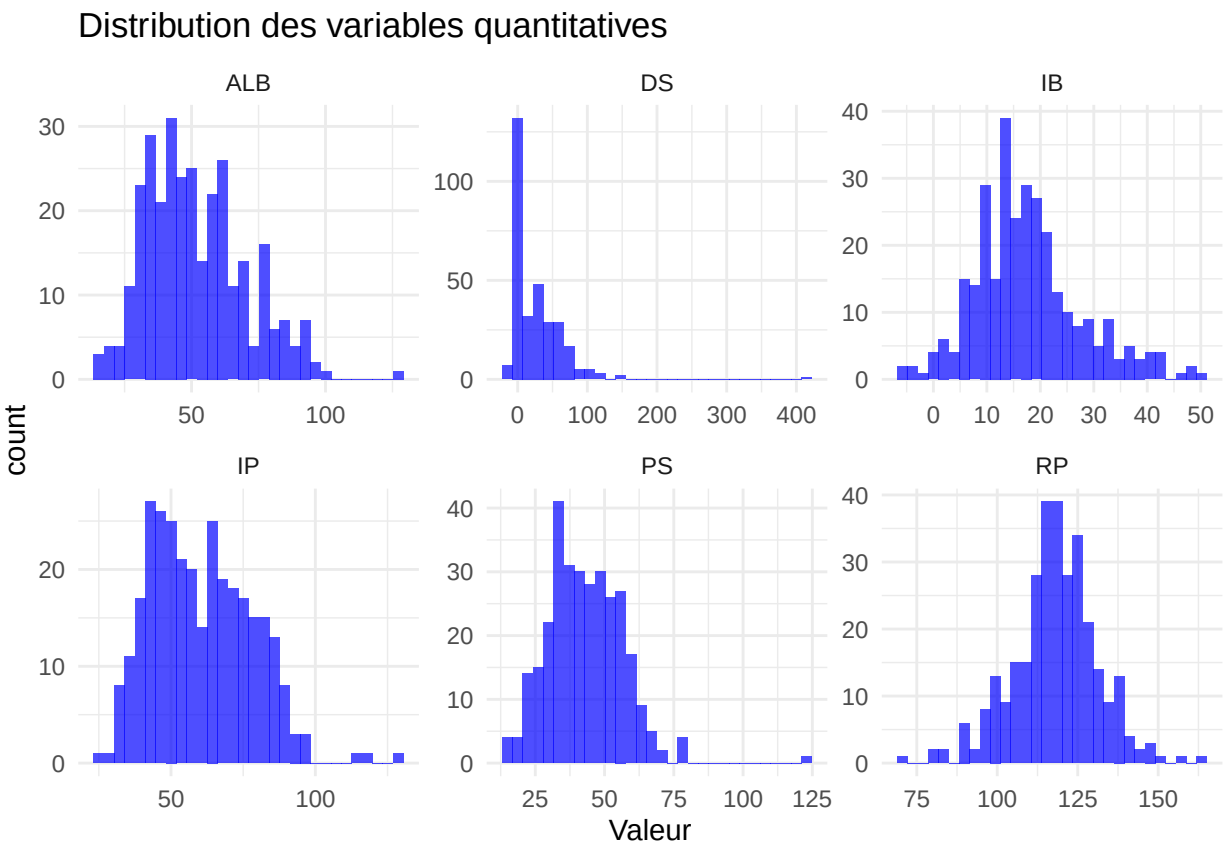
5 Statistique descriptives

A cette étape, nous donnons une première vue de la répartition des données. ## Résumé statistiques

```
##        IP               IB              ALB               PS        
##  Min.   : 26.15   Min.   :-6.555   Min.   : 14.00   Min.   : 13.37  
##  1st Qu.: 46.43   1st Qu.:10.667   1st Qu.: 37.00   1st Qu.: 33.35  
##  Median : 58.69   Median :16.358   Median : 49.56   Median : 42.40  
##  Mean   : 60.50   Mean   :17.543   Mean   : 51.93   Mean   : 42.95  
##  3rd Qu.: 72.88   3rd Qu.:22.120   3rd Qu.: 63.00   3rd Qu.: 52.70  
##  Max.   :129.83   Max.   :49.432   Max.   :125.74   Max.   :121.43  
##        RP               DS                    Pathologie
##  Min.   : 70.08   Min.   :-11.058   Hernie           : 60
##  1st Qu.:110.71   1st Qu.:  1.604   Normal           :100
##  Median :118.27   Median : 11.768   Spondylolisthesis:150
##  Mean   :117.92   Mean   : 26.297                        
##  3rd Qu.:125.47   3rd Qu.: 41.287                        
##  Max.   :163.07   Max.   :418.543
```

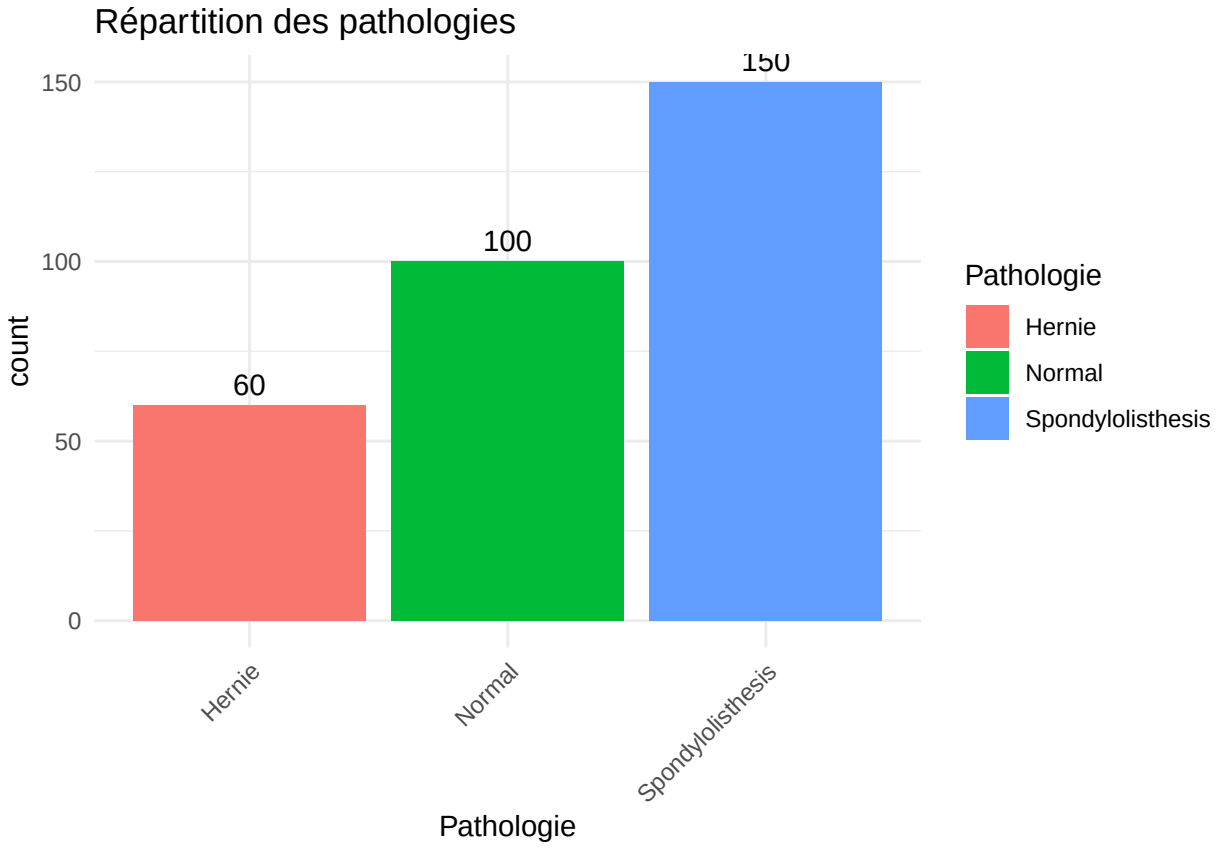
Pour la variable DS, certaines valeurs semblent anormales. Nous la gardons pour l'instant, mais nous pourrions la supprimer si nous obtenons plus d'informations.

5.1 Distribution des variables quantitatives



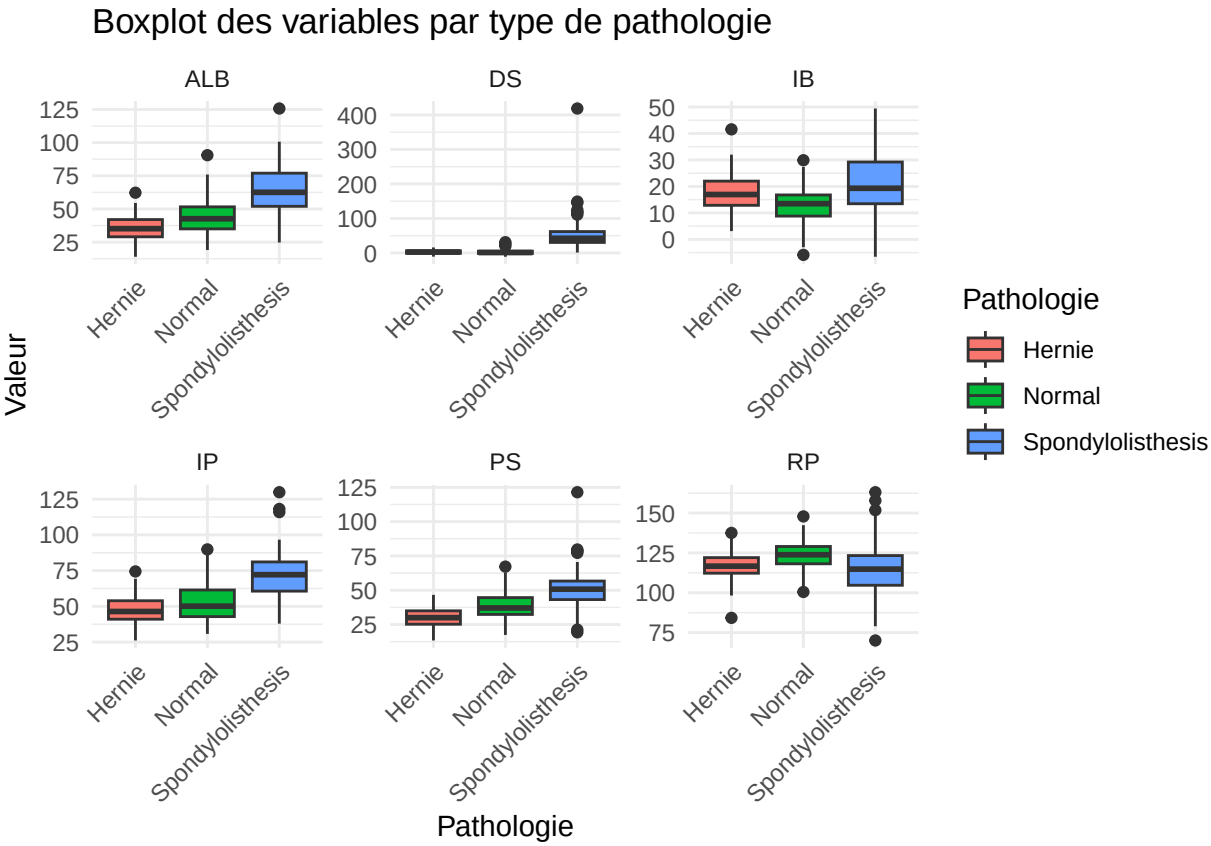
La distribution des variables de la base de données semble majoritairement normale. Cependant, la variable *DS* se distingue par une distribution qui s'apparente davantage à une exponentielle, ce qui pourrait indiquer une asymétrie ou une concentration des valeurs vers un extrême.

5.2 Barplot de la variable quantitative



Les effectifs des trois catégories de pathologie sont inégalement répartis : 60 observations pour la Hernie, 100 pour la catégorie normale et 150 pour la Spondylolisthesis, qui est la plus représentée. Cette différence de taille d'échantillon peut influencer les analyses statistiques. De plus, la disparité des effectifs pourrait biaiser certaines conclusions, rendant nécessaire un ajustement par des méthodes comme le rééchantillonnage.

5.3 Boxplot des variables par type de pathologie



L'analyse des boxplots montre que les variables:

- **ALB, IP, PS** : La médiane et l'étendue des valeurs sont plus élevées pour les patients atteints de **Spondylolisthesis** que pour les autres groupes.
- **DS** : Cette variable présente une distribution fortement asymétrique, avec des valeurs très faibles pour **Hernie** et **Normal**, mais des valeurs beaucoup plus dispersées pour **Spondylolisthesis**.
- **IB et RP** : Les distributions semblent plus homogènes entre les pathologies, bien que **Spondylolisthesis** affiche une plus grande dispersion.

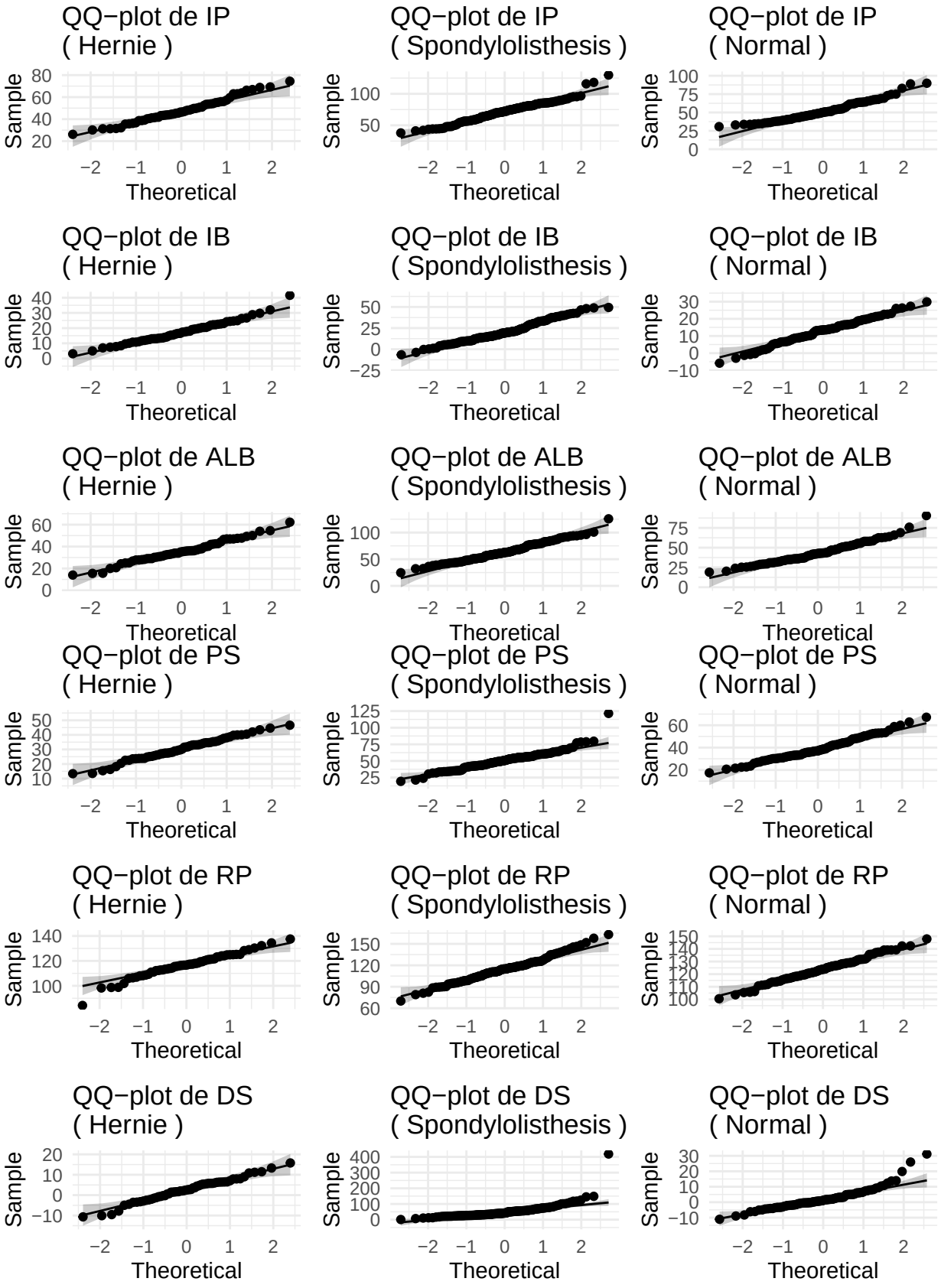
Conclusion:

Certaines variables, comme **DS** et **ALB**, montrent des différences marquées entre les pathologies, ce qui pourrait les rendre intéressantes pour la classification des patients.

5.4 Test statistique

5.4.1 Vérifiecation la normalité des données

Utilisons le qqplot pour chaque variable et chaque groupe de pathologie pour vérifié la normalité.



On remarque que pour chaque groupe de pathologie, toutes les variables suivent une distribution normale.

5.4.2 Test de Levene

Nous allons utilser le test de Levene pour vérifier si la variabilité au sein des groupes est homogène.

5.4.2.1 Hypothèses pour le test de Levene :

- **Hypothèse nulle (H) :**
  Il n'y a pas de différence significative dans les variances entre les groupes. Autrement dit, les variances sont égales pour toutes les catégories de la variable `Pathologie`.

$$H_0 : \sigma_h^2 = \sigma_n^2 = \sigma_s^2$$

  où $\sigma_h^2 = \sigma_n^2 = \sigma_s^2$ sont les variances des groupes définis par la variable `Pathologie` c'est à dire `Normal`,`Hernie`, `Spondylolisthesis`.

- **Hypothèse alternative (H) :**

  Il existe une différence significative dans les variances entre au moins deux groupes (la variabilité au sein des groupes n'est pas homogène).

$$H_1 : \text{Au moins deux variances sont différentes.}$$

Variables	P-value	Commentaire
IP	0.0210397	Différence significative entre les groupes
IB	0.0000003	Différence significative entre les groupes
ALB	0.0000736	Différence significative entre les groupes
PS	0.0251733	Différence significative entre les groupes
RP	0.0000030	Différence significative entre les groupes
DS	0.0000000	Différence significative entre les groupes

Étant donné que nous observons une variabilité entre les variances des groupes pour toutes les variables, ce qui viole l'une des hypothèses fondamentales de l'ANOVA classique, nous ne pourrons pas utiliser ce test pour comparer les moyennes. Par conséquent, nous allons recourir au test de **Kruskal-Wallis**, également appelé **ANOVA unidirectionnelle sur rangs**, qui est une alternative non paramétrique et ne nécessite pas l'hypothèse d'homogénéité des variances.

5.4.2.2 Test Kruskal-Wallis

5.4.2.2.1 Hypothèses pour le test de Kruskal-Wallis :

- **Hypothèse nulle (H) :**
  Il n'y a pas de différence significative dans les distributions des groupes. Autrement dit, les différentes catégories de la variable `Pathologie` proviennent de populations ayant des distributions similaires (les médianes des groupes sont égales).

$$H_0 : \text{Les distributions des groupes sont identiques.}$$

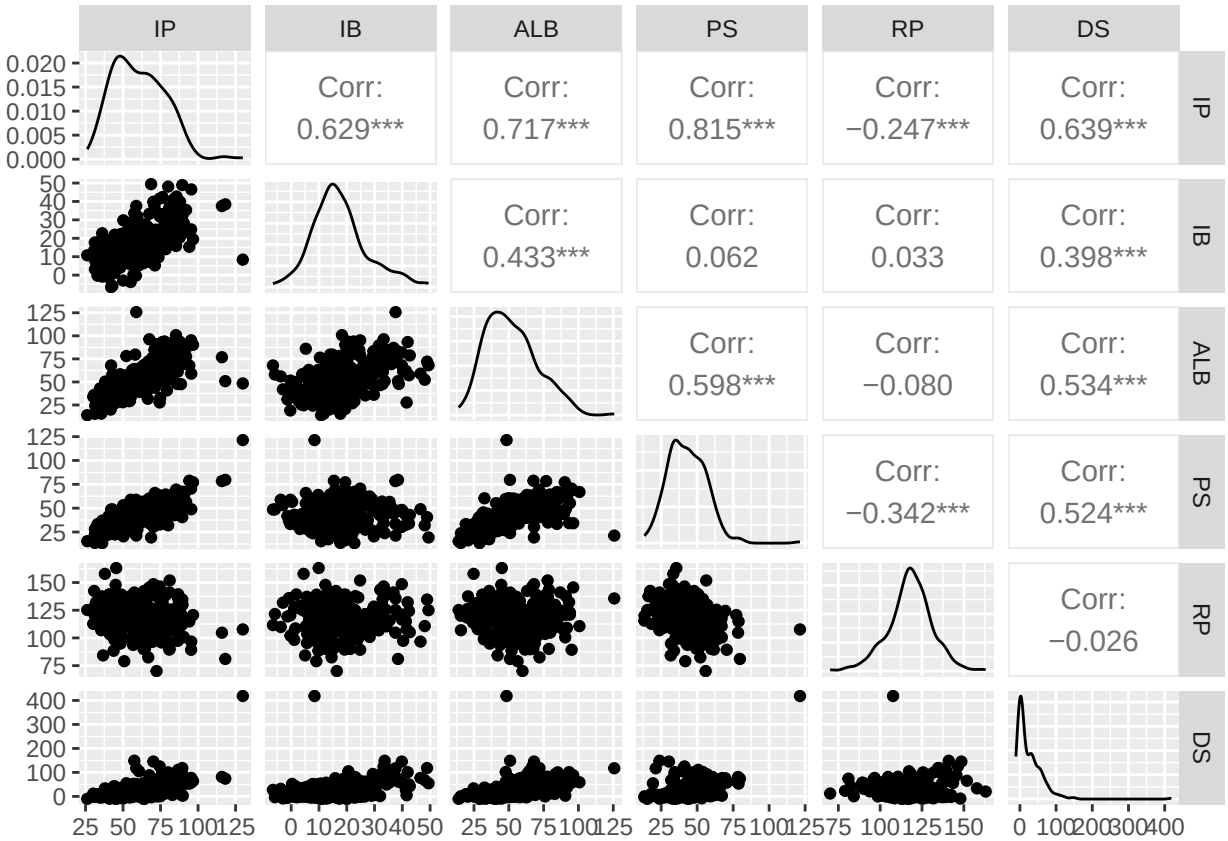
- **Hypothèse alternative (H) :**
  Il existe au moins une différence significative dans les distributions des groupes. Cela signifie que les groupes ne proviennent pas tous de populations ayant des distributions similaires (les médianes des groupes ne sont pas égales).

H_1 : Au moins une des distributions des groupes est différente.

Variables	P-value	Interprétation
IP	0.0210397	Différence significative entre les groupes
IB	0.0000003	Différence significative entre les groupes
ALB	0.0000736	Différence significative entre les groupes
PS	0.0251733	Différence significative entre les groupes
RP	0.0000030	Différence significative entre les groupes
DS	0.0000000	Différence significative entre les groupes

Conclusion générale :

Étant donné que la **p-value** pour toutes les variables est inférieure à 0.05, nous rejetons l'hypothèse nulle pour chaque test de Kruskal-Wallis, ce qui signifie qu'il existe des **différences significatives** entre les distributions des groupes définis par `Pathologie`. En d'autres termes, chaque variable montre des variations notables entre les différentes pathologies, ce qui suggère que les groupes de `Pathologie` ont des comportements distincts pour toutes les métriques mesurées. Ce que nous allons essayer de vérifier avec quelques modèles statistiques.



En analysant le nuage de points et la matrice de corrélation, nous constatons une forte relation positive entre l'incidence pelvienne et deux variables clés : l'angle de lordose lombaire et la pente sacrée. Dans notre cas, la corrélation est d'autant forte entre l'incidence pelvienne et la pente sacrée et de l'ordre de 0.81.

Cette observation est confirmée par la littérature scientifique. Par exemple, l'article "Valeur physiologique des paramètres pelviens et rachidiens : étude chez 300 sujets asymptomatiques" publié sur EM-Consulte

met en évidence une corrélation significative entre la pente sacrée et l'incidence pelvienne (r = 0,8). Aussi, on trouve dans la littérature la relation $IP = PS + VP$.
Ainsi, la question que l'on se pose à cette étape est la conséquence que cela aura si l'on gardait les deux variables dans la suite de notre étude.

Nous réalisons une première étude en conservant l'ensemble des variables. Ensuite, nous menons deux autres analyses, chacune excluant successivement l'une des variables.
La variable d'intérèt concerne les différentes pathologies que l'on remarque en s'intéressant aux variables explicatives. Comme pathologies, nous avons **Hernie**, **Normal** et **Spondylolisthesis**.
En se basant sur la gravité des pathologies (du moins au plus problématique), un ordre naturel se dessine:
- Normal: Aucune douleur.
- Hernie: Souvent douloureux mais pouvant être traité.
- Spondylolisthésis: Plus grave, pouvant nécessiter une intervention chirurgicale si sévère.

Conclusion: Si on suppose que "Spondylolisthésis" est une évolution plus grave que "Hernie" et que "Normal" est l'état initial, alors une régression polytomique ordonnée est plus appropriée.
Cependant, si on s'intéresse à leur origine:
- Normal: Aucune atteinte.
- Hernie: Problème mécanique.
- Spondylolisthésis: Atteinte structurelle.

Conclusion: En s'intéressant à leur origine les trois catégories sont sans ordre précis. Ainsi, une régression polytomique non ordonnée est conseillée.

6 Modèles

6.1 Données d'apprentissage et de test

Nous divisons le jeu de donnée en deux parties:
- Une partie entrainement qui permettra d'entrainer les modèles
- Une partie test pour évaluer la qualité des modèles.

6.2 Modèle logistique ordonné

6.2.1 Avec toutes les variables explicatives

```
## Call:
## polr(formula = Pathologie ~ IP + IB + ALB + PS + RP + DS, data = train,
##     Hess = TRUE)
##
## Coefficients:
##          Value Std. Error    t value
## IP  -1.275e+07    0.01866 -6.834e+08
## IB   1.275e+07    0.02010  6.343e+08
## ALB -2.723e-02    0.02310 -1.179e+00
## PS   1.275e+07    0.01951  6.537e+08
## RP  -6.951e-02    0.02147 -3.237e+00
## DS   2.390e-01    0.02802  8.527e+00
##
## Intercepts:
##                            Value         Std. Error    t value
## Normal|Hernie              -9.039600e+00  3.029300e+00 -2.984100e+00
## Hernie|Spondylolisthesis -6.083600e+00  2.968200e+00 -2.049600e+00
##
## Residual Deviance: 182.05
## AIC: 198.05
```

- **Interpretation**: Le modèle pose un problème car certains coefficients, comme ceux de `IP`, `IB` et `PS`, sont **beaucoup trop grands** de l'ordre de 10^7. Cela peut indiquer un **problème de forte corrélation entre les variables**. À l'inverse, `ALB` a un **effet très faible et non significatif**, ce qui signifie qu'elle n'a probablement pas d'impact sur `Pathologie`. Même si `RP` et `DS` semblent avoir un effet significatif, les valeurs extrêmes des autres coefficients rendent le modèle **instable et difficile à interpréter**. Cela confirme nos conclusions précédentes. Pour améliorer le modèle, nous allons tester des versions sans inclure `IP` et `PS` en même temps, car nous avons vu qu'elles sont très corrélées. Cela aidera à réduire les problèmes de multicolinéarité et à rendre le modèle plus fiable.

6.2.2 Modèle avec la variable `PS` sans `IP`

```
##                              Value Std. Error   t value      p value
## IB                        0.05695358 0.03028851  1.880369 6.005777e-02
## ALB                      -0.02786173 0.02311174 -1.205523 2.280016e-01
## PS                       -0.04051478 0.03048538 -1.328990 1.838512e-01
## RP                       -0.06932421 0.02147284 -3.228460 1.244585e-03
## DS                        0.23834596 0.02800731  8.510135 1.737311e-17
## Normal|Hernie            -9.01385770 3.02970599 -2.975159 2.928365e-03
## Hernie|Spondylolisthesis -6.06154393 2.96947731 -2.041283 4.122269e-02
```

On constate déjà qu'en retirant la varible `IP` les estimations des paramètres semblent être bonne.

Interprétation :

En analysant les coéficients et les p_values, on peut déduire :

-`IB` a un effet légèrement positif mais faiblement significatif , ce qui signifie qu'une augmentation de IB pourrait légèrement diminuer la gravité de la pathologie, mais l'effet reste incertain.

- `ALB` a un effet négatif faible et non significatif, ce qui indique qu'il n'a probablement pas d'impact notable sur Pathologie.

-`PS` a un effet négatif mais non significatif , ce qui suggère qu'il n'influence pas fortement la sévérité de la pathologie.

-`RP` a un effet significativement négatif, ce qui signifie que plus `RP` est élevé, plus la pathologie est sévère.

-`DS` est hautement significatif et positif, ce qui indique que plus `DS` diminue, plus la pathologie est grave.

Intervale de confiances

```
##            2.5 %      97.5 %
## IB  -0.002269883  0.11768260
## ALB -0.074318450  0.01686169
## PS  -0.101756837  0.01847899
## RP  -0.113275074 -0.02892359
## DS   0.187550284  0.29792908
```

Les intervalles de confiance nous donnent une idée de l'incertitude des effets estimés :

- `RP` a un intervalle entièrement négatif (`[-0.113, -0.029]`) qui contient le paramètre estimé, confirmant son effet aggravant. Autrement dit, toutes choses égales par ailleurs, une augmentation de `RP (Rayon pelvien)` de 10 mm augmente le risque de changer de catégorie de pathologie d'au plus 3.
- `DS` a un intervalle entièrement positif (`[0.188, 0.298]`) qui contient le paramètre estimé, confirmant son effet protecteur. Autrement dit toutes choses égales par ailleurs, une augmentation de `DS (Degré de spondylolisthesis)` de 5° dimunie le risque de gravité de la pathologie d'au moins 0.22.

- IB, ALB, et PS ont des intervalles qui incluent **zéro**. De plus $ exp(-coef) 1 $ ce qui signifie que une augmentation de leur mesure n'a pas d'effet sur le changement de la gravité de la maladie toutes choses égales par ailleurs.

Conclusion

Toutes choses égales par ailleurs, les résultats montrent que **RP augmente la gravité de la pathologie**, tandis que **DS la diminue fortement**. En revanche, les effets de IB, ALB et PS restent **incertains ou faibles**.

Prévison

```
## [1] "Accuracy: 79.49 %"
```

	Normal	Hernie	Spondylolisthesis
Normal	23	8	2
Hernie	4	9	1
Spondylolisthesis	1	0	30

On constate que:

- Le modèle a une exactitude de 79.49 %, ce qui signifie que le modèle prédit correctement 79.49 % des observations de test.
- Le modèle semble bien fonctionner pour la classe Spondylolisthesis avec une grande précision, mais il a un certain taux d'erreur pour les classes "Normal" et "Hernie", avec des confusions entre ces deux classes.
- La classe "Hernie" souffre d'un nombre significatif de faux positifs.

6.2.3 Modèle avec la variable IP sans PS

```
##                              Value Std. Error   t value      p value
## IP                     -0.04052020 0.03048614 -1.329135 1.838033e-01
## IB                      0.09747369 0.03719852  2.620365 8.783568e-03
## ALB                    -0.02785388 0.02311273 -1.205132 2.281525e-01
## RP                     -0.06932192 0.02147222 -3.228447 1.244644e-03
## DS                      0.23834413 0.02800705  8.510146 1.737139e-17
## Normal|Hernie          -9.01352439 3.02960146 -2.975152 2.928435e-03
## Hernie|Spondylolisthesis -6.06104646 2.96938673 -2.041178 4.123315e-02
```

Interprétation : La(es) variable(s):

- **Inclinaison du bassin (IB)** : A un effet positif significatif, suggérant qu'une augmentation de l'inclinaison du bassin diminue le risque de changer de catégorie de pathologie.
- **Rayon pelvien (RP)** : A un effet négatif significatif, indiquant que lorsque le rayon pelvien augmente, diminue la risque de changer de catégorie de pathologie.
- **Degré de spondylolisthesis (DS)** : A un effet positif très significatif, ce qui signifie qu'un degré plus élevé de spondylolisthesis est associé à une pathologie moins sévère.
- **Incidence pelvienne (IP)** et **Angle de lordose lombaire (ALB)** n'ont pas d'effet significatif sur la pathologie.

Intervalle de confinance

```
##           2.5 %      97.5 %
## IP  -0.10175683  0.01847900
## IB   0.02579040  0.17233820
```

```
## ALB -0.07431842  0.01686167
## RP  -0.11327506 -0.02892361
## DS   0.18755030  0.29792907
```

Les intervalles de confiances nous confime des commentaires:

- **RP (Rayon pelvien)** : L'intervalle de confiance pour `RP` est entièrement négatif ([**-0.113, -0.029**]) qui contient le paramètre estimé, ce qui indique qu'une augmentation du rayon pelvien est associée à une probabilité élevée de passer à une pathologie plus grave. Autrement dit, toutes choses égales par ailleurs, une augmentation du rayon pelvien de 10mm augmente d'au moins 3 le risque de progression vers une pathologie plus sévère.
- **DS (Degré de spondylolisthesis)** : L'intervalle de confiance pour `DS` est entièrement positif ([**0.188, 0.298**]) qui contient le paramètre estimé, ce qui signifie qu'une augmentation du degré de spondylolisthesis est associée à une probabilité élevée de passer à une pathologie moins grave. Autrement dit, toutes choses égales par ailleurs, une augmentation du degré de spondylolisthesis de 5° diminue d'au moins 0.23 le risque de progression vers une pathologie plus sévère.
- **IB (Inclinaison du bassin)**, **ALB (Angle de lordose lombaire)** et **PS (Pente sacrée)** : Les intervalles de confiance pour ces variables incluent **zéro**, ce qui signifie qu'elles n'ont **pas d'effet significatif** sur la gravité de la pathologie. Ces variables ne sont pas statistiquement significatives dans ce modèle, ce qui suggère qu'elles n'influencent pas de manière mesurable le passage entre les catégories de pathologie. Aussi en l'intervalle de confiance de **ALB**, on constate que, toutes choses égales par ailleurs une augmentation de **ALB** n'a quasiment pas d'effet sur le changement de catégorie de la pathologie.

Prévison

```
## [1] "Accuracy: 79.49 %"
```

	Normal	Hernie	Spondylolisthesis
Normal	23	8	2
Hernie	4	9	1
Spondylolisthesis	1	0	30

On constate que:

- Le modèle a une exactitude de 79.49 %, ce qui signifie que le modèle prédit correctement 79.49 % des observations de test.
- Le modèle semble bien fonctionner pour la classe Spondylolisthesis avec une grande précision, mais il a un certain taux d'erreur pour les classes "Normal" et "Hernie", avec des confusions entre ces deux classes.
- La classe "Hernie" souffre d'un nombre significatif de faux positifs (classées comme "Normal").

6.2.4 Résumé - conclusion

Les variables `IP`(Incidence pelvienne du patient) et `PS`(Pente sacrée) n'ont pas montré d'impact significatif sur le modèle, comme le confirme l'intervalle de confiance des coefficients, qui inclut zéro pour ces deux variables. Cela signifie que, dans ce contexte particulier, l'inclusion de ces variables ne contribue pas à améliorer la capacité prédictive du modèle. De plus, les performances des modèles avec et sans ces variables sont identiques sur l'ensemble de test, ce qui suggère qu'elles sont redondantes ou non significatives pour la tâche de classification. Nous allons donc supprimer les deux variables du modèle.

6.2.5 Modèle sans les variables `IP` et `PS`

```
##                                Value Std. Error   t value      p value
## IB                          0.06764935 0.02933412  2.306166 2.110135e-02
## ALB                        -0.05001791 0.01637649 -3.054251 2.256231e-03
## RP                         -0.05978308 0.01988896 -3.005842 2.648465e-03
## DS                          0.23450006 0.02790116  8.404670 4.290673e-17
## Normal|Hernie              -7.18491064 2.64129575 -2.720222 6.523811e-03
## Hernie|Spondylolisthesis -4.26783180 2.59537330 -1.644400 1.000936e-01
```

```
##           2.5 %      97.5 %
## IB   0.01086662  0.12688888
## ALB -0.08388261 -0.01916350
## RP  -0.10053806 -0.02236664
## DS   0.18403714  0.29399547
```

- **Interpretation**: En analysant les intervalles de confiance et les p_values, on déduit que la variables:
  - **IB (Inclinaison du bassin)** : Cette variable a un effet positif significatif sur la progression de la pathologie. Cela signifie que toutes choses égales par ailleurs, une augmentation de l'inclinaison du bassin de 5° va entrainer une diminution du risque de transition vers une pathologie plus faible d'au plus 0.95.
  - **ALB (Angle de lordose lombaire)** : Cette variable a un effet négatif , indiquant que toutes choses égales par ailleurs, un angle de lordose plus élevé est associé à une augmentation du risque de progression vers une pathologie plus grave.
  - **RP (Rayon pelvien)** : Cette variable a un effet négatif significatif. Cela suggère, toutes choses égale par ailleurs, un rayon pelvien plus grand est lié à une probabilité plus élevée de progression vers une pathologie plus sévère.
  - **DS (Degré de spondylolisthesis)** : Cette variable montre un effet positif très significatif , indiquant, toutes choses égales par ailleurs, un degré plus élevé de spondylolisthesis diminue le risque de progression vers une pathologie plus grave.

En résumé, les variables **IB**, **ALB**, **RP** et **DS** ont un impact significatif sur la progression de la pathologie, tandis que la comparaison entre **Hernie** et **Spondylolisthesis** n'est pas significative.

En résumé, les principales variables influençant la pathologie sont l'inclinaison du bassin, le rayon pelvien, et le degré de spondylolisthesis, tandis que l'incidence pelvienne et l'angle de lordose lombaire n'ont pas d'impact important dans ce modèle.

```
## [1] "Accuracy: 74.36 %"
```

	Normal	Hernie	Spondylolisthesis
Normal	20	9	2
Hernie	7	8	1
Spondylolisthesis	1	0	30

La précision globale a diminué de 79.49% à 74.36%, ce qui suggère que la nouvelle configuration du modèle, même avec des ajustements sur les variables, a légèrement dégradé les performances de classification, surtout dans la matrice de confusion entre `Normal` et `Hernie`. Cependant, la classification de `Spondylolisthesis` semble robuste avec aucune confusion.

6.3 Modèles polytomiques non ordonnés

Dans cette section, nous considérons les différentes pathologies comme des catégories sans ordre spécifique. Étant donné que nous avons précédemment observé que l'inclusion simultanée des variables IP et PS, fortement corrélées, entraîne de mauvaises estimations des paramètres du modèle, nous construirons des modèles en excluant ces deux variables.

Nous mettons en place le modèle

```
## # weights:  18 (10 variable)
## initial  value 254.878051
## iter  10 value 102.072193
## iter  20 value 68.446502
## iter  30 value 67.799271
## iter  40 value 67.458930
## final  value 67.458642
## converged
```

```
## Call:
## multinom(formula = Pathologie ~ IB + ALB + RP + DS, data = train)
##
## Coefficients:
##                   (Intercept)        IB          ALB          RP         DS
## Hernie               10.35044 0.1554930 -0.102930562 -0.07706455 0.03127583
## Spondylolisthesis    -2.23241 0.1101382 -0.001353361 -0.04158491 0.36908685
##
## Std. Errors:
##                   (Intercept)         IB        ALB         RP         DS
## Hernie               3.814377 0.04356194 0.02510922 0.02899300 0.04267845
## Spondylolisthesis    5.319389 0.08227537 0.03912663 0.03234475 0.07714135
##
## Residual Deviance: 134.9173
## AIC: 154.9173
```

```
## , , Hernie
##
##                    2.5 %      97.5 %
## (Intercept)  2.87439433 17.82647901
## IB           0.07011317  0.24087284
## ALB         -0.15214373 -0.05371739
## RP          -0.13388979 -0.02023931
## DS          -0.05237239  0.11492406
##
## , , Spondylolisthesis
##
##                     2.5 %     97.5 %
## (Intercept) -12.65822078 8.19339994
## IB           -0.05111860 0.27139492
## ALB          -0.07804014 0.07533342
## RP           -0.10497945 0.02180963
## DS            0.21789258 0.52028112
```

Une des particularités de la régression logistique multinomiale est qu'elle produit une série de coefficients pour chaque modalité de la variable d'intérêt (sauf la modalité de référence). Ici, nous aurons donc une série de coefficients pour celles et ceux qui dont la pathologie est Hernie (comparés à la modalité Normal) et une série de coefficients pour celles et ceux qui dont la pathologie est Spondylolisthesis (comparés aux

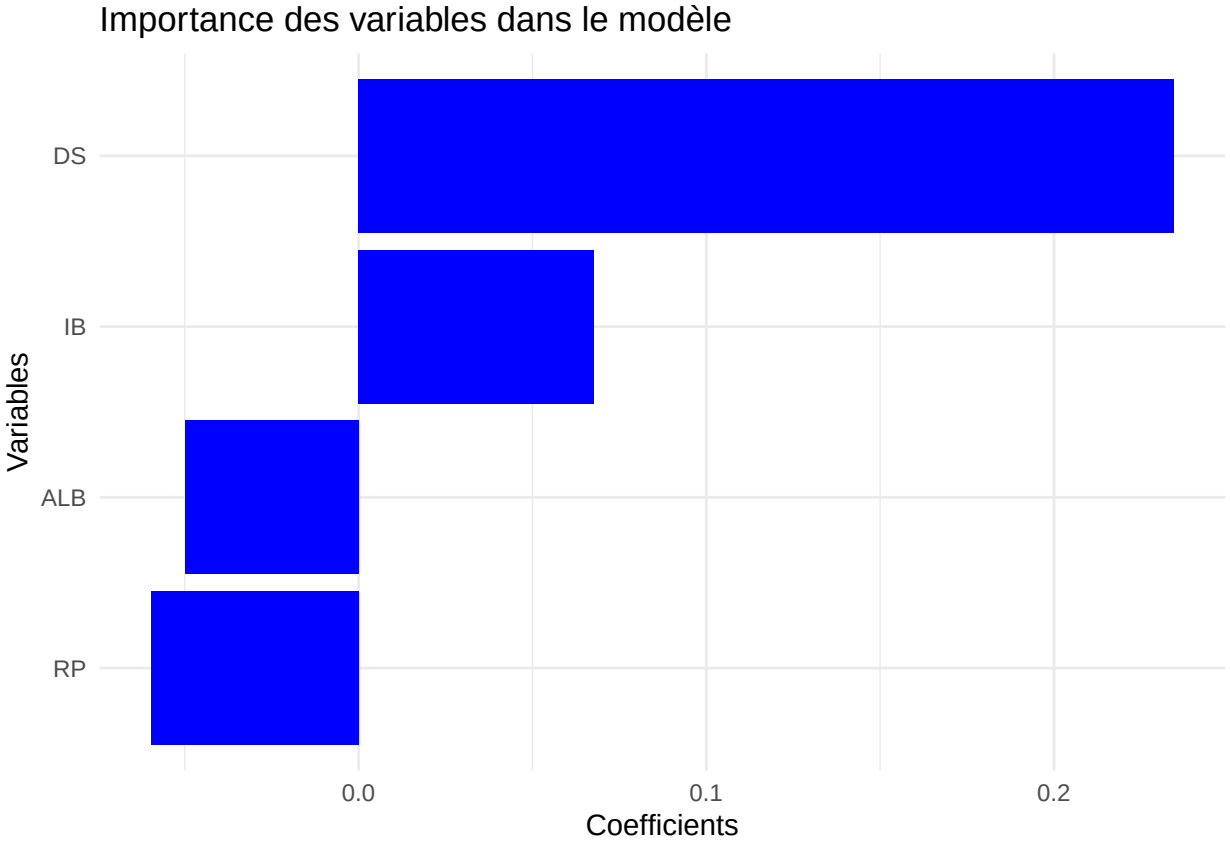
aussi à la modalité Normal).

En analysant les intervalles de confiance des coefficients, nous constatons que l'interprétation des effets varie en fonction de la pathologie considérée. Or, les coefficients devraient refléter le même effet en fixant une modalité de référence. Cette incohérence suggère que traiter les différentes pathologies comme des catégories sans ordre n'est pas l'approche la plus appropriée.

	AIC	Exactitude
Modèle2	195.7925	0.7948718
Modèle3	195.7925	0.7948718
Modèle4	195.5947	0.7435897

Pour des raisons de simplicité et en raison de l'impact moindre des variables **IP** et **PS**, le modèle 4 (sans les variables **IP** et **PS**) a été choisi comme le meilleur modèle, bien que tous les modèles présentent des précisions similaires.

6.4 Importances des variables



Le **degré de spondylolisthesis (DS)** est la variable la plus déterminante dans la progression de la pathologie, tandis que **IB, ALB et RP** ont un effet plus modéré. Cela suggère que DS devrait être un facteur clé à surveiller dans l'évaluation du risque. L'influence des autres variables reste pertinente, mais leur impact est moins prononcé.

7 Conclusion

En conclusion, cette étude a permis d'analyser les facteurs influençant la progression des pathologies vertébrales à l'aide d'une régression polytomique ordinale. Les résultats montrent que le **degré de spondylolisthesis (DS)** est le facteur le plus déterminant, tandis que d'autres variables comme **IB, ALB et RP** ont un impact plus modéré. Le choix du modèle optimal a été guidé par la simplicité et la significativité des variables, aboutissant à la sélection du modèle excluant **IP et PS**. Bien que tous les modèles testés présentent des performances similaires, cette approche permet une meilleure interprétation clinique et une utilisation plus efficace des variables clés pour le diagnostic et la prise en charge des patients.

8 Annexe : Code R

```r
knitr::opts_chunk$set(echo=FALSE, warning = FALSE, message = FALSE, results = "hide")
# Importation des library
library(tidyverse)
library(nnet)
library(ggplot2)
library(corrplot)
library(pROC)
library(MASS)
library(GGally)
library(MASS)
library(nnet)
# importation de la base de données
data <- read.csv("Orthopédie/Orthopédie.csv")
# inspection de la data
str(data)
# visualisation des valeurs manquantes
naniar::vis_miss(data)
data$Pathologie<- as.factor(data$Pathologie )
summary(data)
data %>%
  dplyr::select(-Pathologie) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Valeur") %>%
  ggplot(aes(x = Valeur)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  facet_wrap(~Variable, scales = "free") +
  theme_minimal() +
  labs(title = "Distribution des variables quantitatives")

ggplot(data, aes(x=Pathologie, fill=Pathologie)) +
  geom_bar() +
  geom_text(stat='count', aes(label=..count..), vjust=-0.5) +
  theme_minimal() +
  labs(title = "Répartition des pathologies") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
data %>%
  pivot_longer(cols = c("IP", "IB", "ALB", "PS", "RP", "DS"),
               names_to = "Variable", values_to = "Valeur") %>%
  ggplot(aes(x = Pathologie, y = Valeur, fill = Pathologie)) +
  geom_boxplot() +
  facet_wrap(~Variable, scales = "free") +
  theme_minimal() +
  labs(title = "Boxplot des variables par type de pathologie")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
library(ggpubr)
library(gridExtra)
library(grid)

variables <- colnames(data)[sapply(data, is.numeric)] # Variables numériques
plots <- list()
index <- 1
```

```r
for (var in variables) {
  for (patho in unique(data$Pathologie)) {
    subset_data <- data[data$Pathologie == patho, var, drop = FALSE]

    p <- ggqqplot(subset_data[[var]]) +
      ggtitle(paste("QQ-plot de", var, "\n(", patho, ")")) +
      theme_minimal()

    plots[[index]] <- p
    index <- index + 1
  }
}

total_plots <- length(plots)
half_plots <- ceiling(total_plots / 2)  # Moitié des graphiques

# Première page
grid.arrange(grobs = plots[1:half_plots], nrow = 3, ncol = 3)
# Seconde page
grid.arrange(grobs = plots[(half_plots + 1):total_plots], nrow = 3, ncol = 3)

library(car)

# Créer un tableau vide pour stocker les résultats
results <- data.frame(Métrique = character(), P_value = numeric(), Commentaire = character(),
                      stringsAsFactors = FALSE)

data$Pathologie <- as.factor(data$Pathologie)
# Appliquer le test de Levene et stocker les résultats

for (var in variables) {
  # Effectuer le test de Levene
  test_result <- leveneTest(as.formula(paste(var, "~ Pathologie")), data=data)

  #Extraire la p-value
  p_value <- test_result$`Pr(>F)`[1]

  if (p_value < 0.05) {
    commentaire <- "Différence significative entre les groupes"
  }
  else {
    commentaire <- "Pas de différence significative"
  }

  results <- rbind(results, data.frame(Métrique = var, P_value = p_value, Commentaire = commentaire))
}

# Afficher les résultats avec knitr::kable
knitr::kable(results, col.names = c("Variables", "P-value", "Commentaire"), format = "markdown")
# Créer un tableau vide pour stocker les résultats
results_krust <- data.frame(Métrique = character(), P_value = numeric(), Commentaire = character(),
                      stringsAsFactors = FALSE)
```

```r
data$Pathologie <- as.factor(data$Pathologie)

# Appliquer le test de Kruskal-Wallis et stocker les résultats
for (var in variables) {
  # Effectuer le test de Kruskal-Wallis
  test_result <- kruskal.test(as.formula(paste(var, "~ Pathologie")), data=data)

  # Extraire la p-value
  p_value <- test_result$p.value

  # Ajouter un commentaire basé sur la p-value
  if (p_value < 0.05) {
    commentaire <- "Différence significative entre les groupes"
  } else {
    commentaire <- "Pas de différence significative"
  }

  # Ajouter le résultat au tableau
  results_krust <- rbind(results, data.frame(Métrique = var, P_value = p_value, Commentaire = commentai
}

# Afficher les résultats avec knitr::kable
knitr::kable(results_krust[1:6,], col.names = c("Variables", "P-value", "Interprétation"), format = "ma

data$Pathologie<-factor(data$Pathologie,
       levels = c("Normal", "Hernie", "Spondylolisthesis"), ordered = TRUE)
ggpairs(data[,1:6] )
set.seed(2)
n<- sample(nrow(data), nrow(data)*0.75)
train<- data[n,]
test<- data[-n,]
train1 <- train %>% dplyr::select(-IP)
train2<- train %>% dplyr::select(-PS)
rec1<- polr(Pathologie~IP+IB+ALB+PS+RP+DS, data=train, Hess = TRUE)
summary(rec1)
rec2<- polr(Pathologie~IB+ALB+PS+RP+DS, data=train1, Hess = TRUE)
ctable2 <- coef(summary(rec2))
p2 <- pnorm(abs(ctable2[, "t value"]), lower.tail = FALSE) * 2
ctable2 <- cbind(ctable2, "p value" = p2)
ctable2
# Intervalle de confiance
confint(rec2)
pred2 <- predict(rec2, newdata = test)
conf_matrix2<- table(Predicted = pred2, Actual = test$Pathologie)

#Calcul de l'exactitude
accuracy2 <- sum(diag(conf_matrix2)) / sum(conf_matrix2)
print(paste("Accuracy:", round(accuracy2 * 100, 2), "%"))
knitr::kable(conf_matrix2,format = "markdown")
rec3<- polr(Pathologie~IP+IB+ALB+RP+DS, data=train2, Hess = TRUE)
ctable3 <- coef(summary(rec3))
p3 <- pnorm(abs(ctable3[, "t value"]), lower.tail = FALSE) * 2
ctable3 <- cbind(ctable3, "p value" = p3)
```

```r
ctable3

confint(rec3)
pred3 <- predict(rec3, newdata = test)
conf_matrix <- table(Predicted = pred3, Actual = test$Pathologie)

#Calcul de l'exactitude
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
print(paste("Accuracy:", round(accuracy * 100, 2), "%"))
knitr::kable(conf_matrix,format = "markdown")
rec4<- polr(Pathologie~IB+ALB+RP+DS, data=train, Hess = TRUE)
ctable4 <- coef(summary(rec4))
p4 <- pnorm(abs(ctable4[, "t value"]), lower.tail = FALSE) * 2
ctable4 <- cbind(ctable4, "p value" = p4)
ctable4
confint(rec4)

pred4 <- predict(rec4, newdata = test)
conf_matrix4 <- table(Predicted = pred4, Actual = test$Pathologie)

#Calcul de l'exactitude
accuracy4 <- sum(diag(conf_matrix4)) / sum(conf_matrix4)
print(paste("Accuracy:", round(accuracy4 * 100, 2), "%"))
knitr::kable(conf_matrix4,format = "markdown")
rec5 <- multinom(Pathologie ~ IB + ALB  + RP + DS, data = train)
summary(rec5)
confint(rec5)
resume <- rbind(AIC(rec2), AIC(rec3),AIC(rec4))
resume <- cbind(resume, rbind(accuracy2, accuracy, accuracy4))
rownames(resume) <- c("Modèle2", "Modèle3","Modèle4")
knitr::kable(resume,  col.names=c('AIC', 'Exactitude'), format = "markdown")
importance <- data.frame(Variable = c('IB', 'ALB','RP','DS'),
                         Coefficient = coef(rec4))
ggplot(importance, aes(x = reorder(Variable, Coefficient), y = Coefficient)) +
  geom_bar(stat = "identity", fill = "blue") +
  coord_flip() +
  theme_minimal() +
  labs(title = "Importance des variables dans le modèle",
       x = "Variables",
       y = "Coefficients")
```