Athérosclérose

Perrine Warter

2024-11-22

Introduction du sujet

Cette analyse porte sur l'athérosclérose et les principaux facteurs de risque. L'athérosclérose est l'une des principales causes de décès dans les pays développés, touchant majoritairement les hommes après 35 ans et les femmes après 45 ans. Cette maladie se caractérise par un épaississement et une perte d'élasticité des parois internes des artères, pouvant entraîner un infarctus du myocarde. La paroi des artères est composée de trois dont l'épaisseur de l'intima-média est un indicateur clé de l'athérosclérose. Nous chercherons à quantifier l'impact des variables explicatives via une régression linéaire multiple

Le but de cette analyse est de voir parmis les variables suivantes : le sexe, l'âge, la taille, le poids, le sport, l'alcool et le tabac (en étudiant la quantité de paquets fumés par an) laquelle ou lesquelles ont un réel impact sur l'athérosclérose (en mesurant la taille de l'intima-média : indicateur clé dans cette maladie).

Data Frame Summary

intima

Dimensions: 110 x 9
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	SEXE [factor]	1. homme 2. femme	53 (48.2%) 57 (51.8%)		110 (100.0%)	0 (0.0%)
2	AGE [integer]	Mean (sd) : 39.5 (11.2) min ≤ med ≤ max: 22 ≤ 39.5 ≤ 64 IQR (CV) : 18 (0.3)	40 distinct values		110 (100.0%)	0 (0.0%)
3	taille [integer]	Mean (sd) : 168.7 (9.4) min ≤ med ≤ max: 150 ≤ 169 ≤ 187 IQR (CV) : 16 (0.1)	32 distinct values		110 (100.0%)	0 (0.0%)
4	poids [integer]	Mean (sd) : 67.8 (13.4) min ≤ med ≤ max: 43 ≤ 68 ≤ 115 IQR (CV) : 18.8 (0.2)	41 distinct values		110 (100.0%)	0 (0.0%)
5	tabac [factor]	1. ne fume pas 2. a arrêté de fumer 3. fume	72 (65.5%) 18 (16.4%) 20 (18.2%)		110 (100.0%)	0 (0.0%)
6	paqan [integer]	Mean (sd) : 13.9 (11) min ≤ med ≤ max: 1 ≤ 10 ≤ 45 IQR (CV) : 15 (0.8)	16 distinct values		38 (34.5%)	72 (65.5%)
7	SPORT [factor]	1. non 2. oui	61 (55.5%) 49 (44.5%)		110 (100.0%)	0 (0.0%)
8	mesure [numeric]	Mean (sd) : 0.5 (0.1) min ≤ med ≤ max: 0.4 ≤ 0.5 ≤ 0.8 IQR (CV) : 0.1 (0.2)	31 distinct values		110 (100.0%)	0 (0.0%)
9	alcool [factor]	1. ne boit pas 2. boit occasionnellement 3. boit régulièrement	23 (20.9%) 71 (64.5%) 16 (14.5%)		110 (100.0%)	0 (0.0%)

Generated by summarytools 1.0.1 (R version 4.3.2)
2024-12-04

Cette étude porte sur un échantillon de 110 sujets : 53 hommes et 57 femmes. L'âge médian est de 40 ans, la taille médiane est de 169 cm et le poids médian de 68 kg. Concernant le tabac, 65 % ne fument pas, 18 % fument, et l'exposition (nombre de paquet par an) médiane est de 10. Côté activité physique, 45 % pratiquent un sport. Pour l'alcool, 65 % consomment occasionnellement, 21 % ne boivent pas, et 15 % boivent régulièrement.

Régression linéaire multiple

Pour l'analyse statistique, un seuil de significativité de 0,05 % (p < 0.0005) a été retenu pour déterminer si les relations observées étaient statistiquement significatives."

Sélection automatique des variables

```
## # A tibble: 11 × 5
##    term                          estimate std.error statistic     p.value
##    <chr>                            <dbl>     <dbl>     <dbl>       <dbl>
##  1 (Intercept)                    0.226     0.183      1.78    0.0788
##  2 SEXEFemme                      0.00360   0.0192     0.187   0.852
##  3 AGE                            0.00416   0.000747   5.57    0.000000215
##  4 taille                        -0.000547  0.00109   -0.503   0.616
##  5 poids                          0.00175   0.000659   2.65    0.00929
##  6 tabaca arrêté de fumer         0.0428    0.0264     1.62    0.108
##  7 tabacfume                      0.00989   0.0242     0.409   0.683
##  8 paqan                         -0.00117   0.00114   -1.01    0.315
##  9 SPORTOui                      -0.000300  0.0142    -0.0212  0.983
## 10 alcoolboit occasionnellement  -0.00758   0.0178    -0.427   0.671
## 11 alcoolboit régulièrement      -0.0271    0.0271    -0.997   0.321
```

Cette fonction nous permet d'identifier quelle variable à un effet sur la mesure de l'intima-média et donc un impact sur la maladie. En analysant la p-value, nous observons que la variable âge (Pr < 0,05) a un effet significatif sur la mesure de l'intima-média. La p-value du poids étant aussi inférieure à 0,05, cette variable a aussi un impact sur la mesure de l'intima média. Les autres variables présentant des p-values > 0,05 n'ont que peu d'impact sur notre variable mesure.

Modèle ajusté

```
## Start:  AIC=-583.87
## mesure ~ AGE + poids
##
##          Df Sum of Sq     RSS     AIC
## <none>                 0.51582 -583.87
## - poids   1  0.052943 0.56876 -575.13
## - AGE     1  0.259555 0.77537 -541.04
```

```
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1      NA       NA       107  0.5158194 -583.8727
```

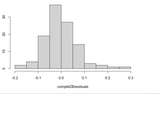
```
## lm(formula = mesure ~ AGE + poids, data = intima)
```

Nous réalisons ensuite une selection des variables par rapport au critère de l'AIC afin d'obtenir le modèle statistique le plus simple tout en expliquant bien la variable dépendante (mesure). Plus l'AIC est bas, plus le modèle s'ajuste aux données. Nous remarquons que seules les variables âge et poids restent présentes validant ainsi nos premières hypothèses, ces variables ont un effet significatif sur notre variable mesure de l'intima-média.

Graphique du modèle ajusté

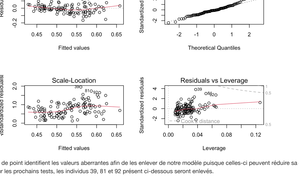
Ce graphique affiche les coefficients estimés de mon modèle ajusté. Nous pouvons voir que l'âge et le poids ont un effet positif et un impact important sur notre variable dépendante. Ce qui signifie que plus l'âge et le poids augmentent, plus la mesure de l'intima-média augmente. Cependant les autres variables n'impactent pas la variable dépendante.

Normalité




```
##
##  Shapiro-Wilk normality test
##
## data:  complet2$residuals
## W = 0.95832, p-value = 0.001674
```

En réalisant un histogramme des résidus et le test de Shapiro-Wilk, nous souhaitons vérifier la normalité des résidus permettant d'évaluer la qualité du modèle. Nous obtenons une p-value < 0,05 signifiant que les résidus ne suivent pas une loi normale. Cela pourrait être due à des valeurs aberrantes.



Ces quatres nuages de point identifient les valeurs aberrantes afin de les enlever de notre modèle puisque celles-ci peuvent réduire sa capacité prédictive. Ainsi pour les prochains tests, les individus 39, 81 et 92 présent ci-dessous seront enlevés.

```
##      SEXEE AGE taille poids      tabac paqan SPORT mesure       alcool
## 39 Femme  53   155    50 ne fume pas    0   non   0.82 boit occasionnellement
## 81 Femme  59   156    50 ne fume pas    0   non   0.79 boit occasionnellement
## 92 homme  45   175   100 ne fume pas    0   non   0.79 boit occasionnellement
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_filtered
## W = 0.99095, p-value = 0.6595
```

En enlevant les valeurs aberrantes nous obtenons une p-value > 0,05 nous permettant de dire que nos données suivent une loi normale.

Selection automatique des variables du modèle sans valeurs aberrantes

```
## Start:  AIC=-603.01
## mesure ~ AGE + poids + tabac
##
##          Df Sum of Sq     RSS     AIC
## <none>                 0.24775 -603.01
## - tabac   2  0.015889 0.26364 -602.23
## - poids   1  0.044164 0.29191 -592.22
## - AGE     1  0.147766 0.49551 -567.12
```

```
##   Step Df Deviance Resid. Df Resid. Dev      AIC
## 1      NA       NA       102  0.2477492 -603.014
```

```
## lm(formula = mesure ~ AGE + poids + tabac, data = intima_reduit)
```

Nous réalisons à nouveau une selection des variables par rapport au critère de l'AIC cette fois sans les valeurs aberrantes. Nous remarquons que les variables âge et poids restent présentes cependant la variable tabac qui n'était pas présente dans notre premier modèle est maintenant présente, signifiant donc que cette variable à un effet significatif sur notre variable mesure de l'intima-média.

Modèle ajusté sans valeurs aberrantes.

En comparant le modèle avec les valeurs aberrantes et le modèle sans, nous observons que ces valeurs avaient un fort impact sur notre modèle. En effet les 3 individus étaient des non-fumeurs mais présentaient des mesures de l'intima-média énorme modifiant donc nos prédictions. Nous pouvons ainsi continuer notre analyse sur ce modèle ajusté sans valeurs aberrantes.

Multicolinéarité

```
## `geom_smooth()` using formula = 'y
## ~ x'
```

Une multicolinéarité peut affecter la stabilité et la fiabilité du modèle, il est donc nécessaire de la vérifier. La ligne de tendance est plate, ce qui indique qu'il n'y a pas de relation apparente entre le sexe, le poids et le tabac. Cependant il est nécessaire d'utiliser des tests statistiques adaptés afin de valider ces observations.

```
##            GVIF Df GVIF^(1/(2*Df))
## AGE    1.123109  1        1.060295
## poids 1.039661  1        1.019736
## tabac 1.168620  2        1.039725
```

Un VIF proche de 1 (ici 1.001981) signifie qu'il n'y a pas de multicolinéarité significative. Cela veut dire que l'âge et le poids ne sont pas fortement corrélés avec les autres variables du modèle de manière problématique. Je peux donc continuer à travailler avec les variables telles quelles.

Estimation des coefficient du modèle ajusté sans valeurs aberrantes.

```
## # A tibble: 5 × 7
##   term                estimate std.error statistic  p.value conf.low conf.high
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)          0.266     0.0374      7.11  1.63e-10  1.92e-1   0.340
## 2 AGE                  0.00359   0.000545    6.58  2.03e- 9  2.51e-3   0.00467
## 3 poids                0.00160   0.000444    3.60  4.95e- 4  7.17e-4   0.00248
## 4 tabaca arrêté de fum… 0.0332   0.0156      2.12  3.64e- 2  2.15e-3   0.0642
## 5 tabacfume            0.0140    0.0159      0.883 3.79e- 1 -1.75e-2   0.0456
```

Nous obtenons donc les coefficients de nos variables âge, poids et tabac, de plus nous pouvons voir que l'erreur standard reste très faible signifiant que cette estimation est précise. Ces coefficients nous permettent de prédire notre variable mesure de l'intima-média pour chaque individu

```
##          1          2          3          4          5          6
## 0.5292853 0.5053526 0.5899542 0.5711659 0.5422819 0.5439699
```

Voici les premières valeurs prédites par notre modèle linéaire. Nous pouvons ainsi mesurer les residus, soit la différence entre la mesure réelle et la valeur prédite pour voir l'erreur de prédiction pour cette observation, et nous obtenons une différence de 2%, ce qui reste très faible.

Prédiction de la variable dépendante avec un modèle ajusté.

```
## [1] "La valeur PRESS du modèle complet est : 0.0554631294317875 et celui du modèle réduit : 0.0035042459259830
## 1"
```

Le test PRESS permet de voir quel modèle est le plus précis dans ses prédictions, avec de moins grands écarts entre les valeurs observées et les valeurs prédites. Ici on observe une différence entre le modèle complet et le modèle ajusté sans les valeurs aberrantes, ce test permet de savoir que nous pouvons utiliser le modèle (modèle plus simple) et qu'il a un léger avantage en termes de capacité prédictive, montrant ainsi que le modèle peut bien prédire de nouvelles observations.

```
##        fit       lwr       upr
## 1 0.355503 0.4373924 0.4733234
```

Nous pouvons enfin prédire les valeurs de la variable mesure de l'intima-média, par exemple pour un homme de 50 ans mesurant 170cm, pesant 69kg ne fumant pas, ne buvant pas et faisant du sport nous obtenons un valeur prédictive de la mesure de l'intima-média de 0,58mm avec des bornes de l'intervalle de prédiction allant de 0,44 à 0,67mm.

Conclusion

L'analyse de la régression linéaire multiple réalisée sur les données de l'athérosclérose montre que certaines variables ont un effet significatif sur l'épaisseur de l'intima-média, qui est un marqueur clé de l'athérosclérose. Plus précisément, l'âge, le poids et la consommation de tabac apparaissent comme des facteurs significatifs. Nous pouvons aussi conclure que l'arrêt de tabac peut être associé à un risque accru de l'athérosclérose. D'autre part, des variables telles que l'activité physique, l'alcool ou le sexe n'ont pas montré d'effet statistiquement significatif dans ce modèle, ce qui peut suggérer qu'elles n'ont pas un impact aussi direct sur l'épaisseur de l'intima-média dans cet échantillon particulier.