UVSQ université PARIS-SACLAY

ISTY
Institut des Sciences et Techniques des Yvelines
CAMPUS DE MANTES EN YVELINES
CAMPUS DE SAINT-QUENTIN-EN-YVELINES

Projet Big Data

Préparé par : Rochdi DARDOR
IATIC4
2023/2024

Table des matières :

INTRODUCTION :

Les méthodes statistiques telles que la régression linéaire et l'analyse en composantes principales (ACP) sont des outils essentiels pour explorer et comprendre les relations complexes entre les variables dans les ensembles de données. Ce rapport applique ces techniques à deux jeux de données distincts : le jeu de données Davis et le jeu de données ANSUR II, afin d'analyser la relation entre la taille et le poids.

La régression linéaire simple permet de modéliser la relation entre deux variables, tandis que l'ACP est une technique de réduction de la dimensionnalité qui facilite l'exploration de la structure des données multidimensionnelles. Dans chaque analyse, nous mettons en évidence les étapes clés, allant de la visualisation initiale des données à l'interprétation des résultats des modèles statistiques.

Dans la première partie de ce rapport, nous utilisons la régression linéaire simple sur le jeu de données Davis pour examiner la relation entre le poids et la taille. Nous décrivons en détail le processus d'analyse, de la visualisation des données à l'interprétation des coefficients de régression, en soulignant les défis rencontrés et les découvertes réalisées.

Ensuite, nous appliquons la régression linéaire au jeu de données ANSUR II pour étudier la relation entre la stature et le poids chez les hommes. En évaluant les résultats, nous analysons la significativité de la relation entre ces variables. Enfin, nous utilisons l'ACP sur le jeu de données ANSUR II pour explorer la structure sous-jacente des mesures anthropométriques. Nous analysons également les graphiques des individus et des variables afin d'identifier les schémas de variation et les regroupements potentiels au sein des données.

Ce rapport a pour objectif d'extraire des informations pertinentes à partir des données en utilisant deux approches analytiques. Notre but est de mieux comprendre les relations entre les variables étudiées en appliquant ces techniques statistiques aux deux jeux de données distincts.

Régression linéaire simple :

La régression linéaire simple est une méthode statistique utilisée pour modéliser la relation entre une variable indépendante (ou explicative) et une variable dépendante (ou expliquée). Cette méthode est choisie pour sa capacité à fournir une compréhension claire et quantifiable de la relation linéaire entre deux variables.
Le principal intérêt de la régression linéaire simple réside dans sa capacité à estimer l'impact de la variable explicative sur la variable expliquée. En ajustant une droite qui minimise les écarts entre les valeurs observées et les valeurs prédites, cette méthode permet de déterminer comment les changements dans la variable indépendante se traduisent par des changements dans la variable dépendante.

Avant de procéder à l'analyse de régression, il est important de souligner deux aspects qui guideront notre analyse : la visualisation des données et les valeurs (métriques) sur lesquelles reposeront nos interprétations.

la visualisation des données et offre plusieurs avantages essentiels :

- **Identification des anomalies** : La visualisation permet de détecter des valeurs aberrantes ou des erreurs de saisie qui pourraient biaiser les résultats de l'analyse. Par exemple, un diagramme de dispersion (scatter plot) peut révéler des points qui ne suivent pas la tendance générale des données.
- **Détection des relations potentielles** : La visualisation aide à identifier des relations linéaires potentielles entre la variable explicative et la variable expliquée. Un graphe initial peut montrer si une relation linéaire semble plausible, justifiant ainsi l'utilisation de la régression linéaire.
- **Vérification des hypothèses** : Les graphiques permettent également de vérifier certaines hypothèses de base de la régression linéaire, comme la linéarité et l'homogénéité des variances.

La régression produit plusieurs valeurs métriques fondamentales pour l'interprétation des résultats :

1. **Coefficients de régression (Estimate)** : Le coefficient de la variable explicative indique de combien la variable dépendante change en moyenne pour chaque unité de changement de la variable indépendante. Par exemple, un coefficient positif suggère une relation positive entre les variables.
2. **Erreur standard (Std. Error)** : L'erreur standard mesure la précision de l'estimation des coefficients. Une petite erreur standard indique une estimation plus précise.
3. **Valeur t (t value)** : La valeur t est utilisée pour tester l'hypothèse nulle selon laquelle le coefficient est nul. Une valeur t élevée indique que le coefficient est significativement différent de zéro.
4. **Valeur p (Pr(>|t|))** : La valeur p indique la probabilité d'obtenir un résultat aussi extrême que celui observé sous l'hypothèse nulle. Une valeur p faible (inférieure à 0,05) suggère que le coefficient est statistiquement significatif.
5. **Coefficient de détermination** R^2 : Le coefficient de détermination mesure la proportion de variance de la variable dépendante expliquée par le modèle. Un R^2 élevé indique que le modèle explique bien la variabilité des données.
6. R^2 **Ajusté** : le R^2 ajusté prend en compte le nombre de prédicteurs dans le modèle et ajuste le R^2 en conséquence. Cela est particulièrement utile lorsque plusieurs variables explicatives sont utilisées.

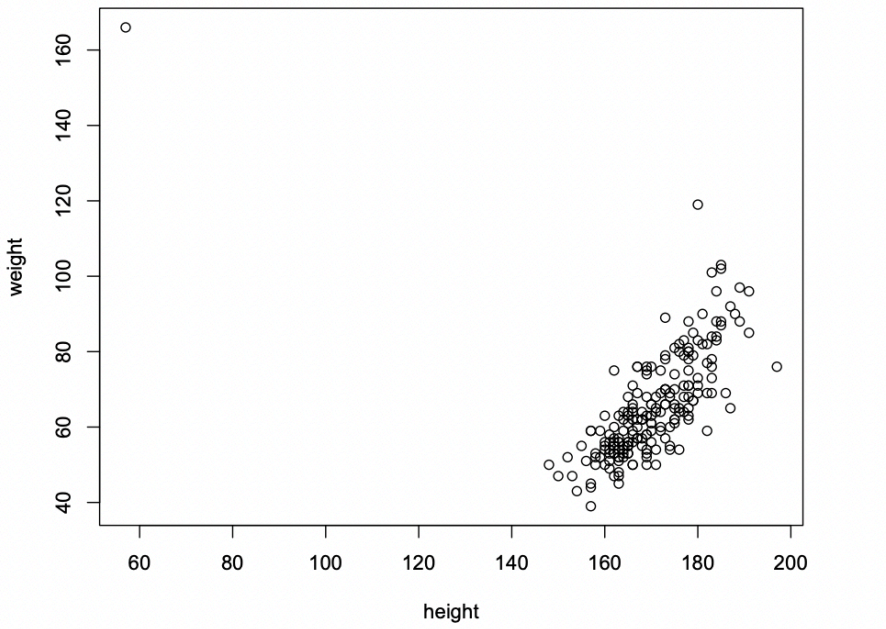
7. **Statistique F et valeur p associée** : La statistique F teste la significativité globale du modèle. Une valeur F élevée, combinée à une valeur p faible, suggère que le modèle est globalement significatif et explique bien la variation de la variable dépendante.

Présentation du jeu de données Davis :

Pour illustrer l'application de la régression linéaire simple, nous utiliserons le jeu de données Davis. Ce jeu de données contient des informations sur des individus, notamment le sexe, le poids, la taille, le poids rapporté et la taille rapportée. Pour cette analyse, nous nous concentrerons sur les variables poids (weight) et taille (height). Le but de cette analyse est d'expliquer le poids en fonction de la taille. Le jeu de données Davis est particulièrement pertinent pour cette étude car il fournit une base empirique pour explorer la relation entre ces deux variables, permettant ainsi de démontrer l'utilité et l'application pratique de la régression linéaire simple.

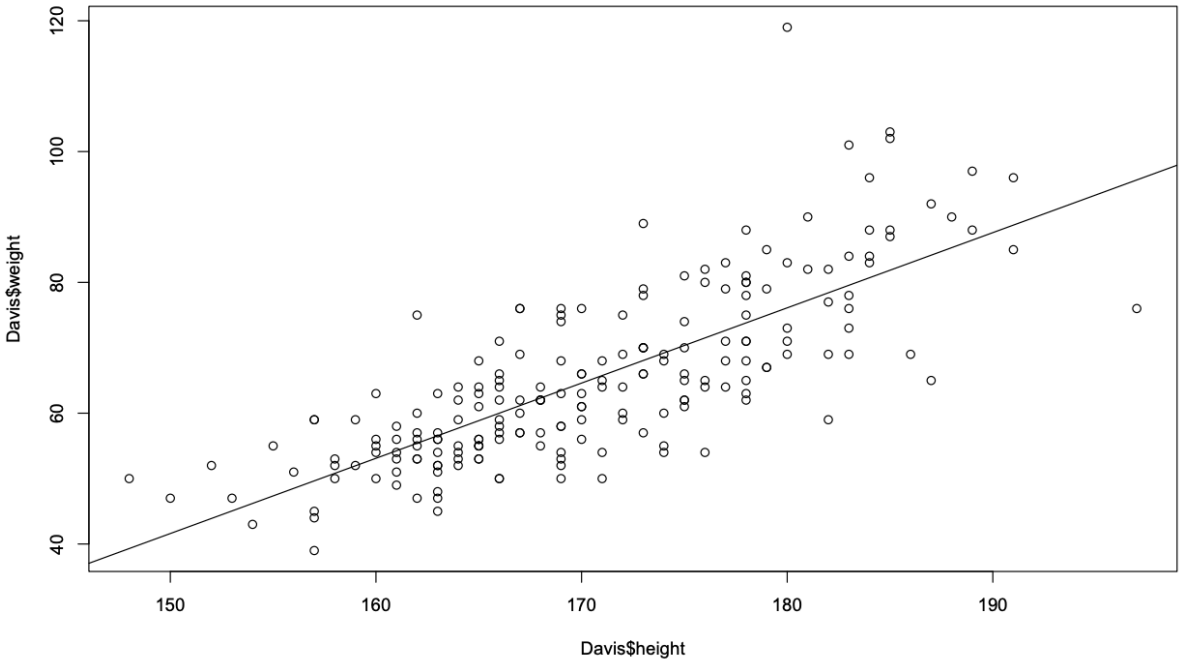
Graphique de régression linéaire initiale :

Pour commencer l'analyse, nous avons visualisé les données du jeu de données Davis afin de comprendre la relation entre le poids (weight) et la taille (height).
En analysant le graphique initial montrant la relation entre le poids (weight) et la taille (height), nous avons remarqué la présence d'une valeur aberrante. Cette observation anormale peut fausser les résultats de notre analyse. Pour identifier cette valeur aberrante, nous avons utilisé la fonction **which.max(weight)** et avons



(Figure 1 : Graphe initiale)

découvert qu'elle correspond à l'observation numéro 12, qui affiche un poids de 166 kg pour une taille de 57 cm, ce qui est manifestement erroné.
Il est évident que cette erreur résulte d'une inversion entre le poids et la taille. En corrigeant cette inversion, la nouvelle valeur devient plus cohérente avec les autres observations du jeu de données. Après cette correction, nous avons mis à jour le graphe pour refléter la relation corrigée entre le poids et la taille.



(Figure 2: Graphique corrigé montrant le poids en fonction de la taille après suppression de la valeur aberrante)

La Figure 2 montre le graphe corrigé, où l'on observe une relation linéaire plus claire et cohérente. Ce graphique confirme également la tendance générale indiquant qu'à mesure que la taille augmente, le poids tend également à augmenter, suggérant une relation positive entre ces deux variables.

Maintenant, nous allons effectuer une analyse de régression linéaire pour explorer plus en détail la relation entre le poids et la taille.

```
Residuals:
    Min      1Q  Median      3Q     Max
-19.658  -5.381  -0.555   4.807  42.894

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -130.91040   11.52792  -11.36   <2e-16 ***
Davis$height    1.15009    0.06749   17.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.505 on 198 degrees of freedom
Multiple R-squared:  0.5946,    Adjusted R-squared:  0.5925
F-statistic: 290.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

Interprétation :

- **Intercept** : L'intercept est de -130.91040 avec une erreur standard de 11.52792. Le t value associé est très faible (t = -11.36) avec une valeur p pratiquement nulle (< 2e-16), indiquant que cet intercept est statistiquement significatif. Cela signifie que lorsque la variable explicative (taille) est nulle, la variable expliquée (poids) est d'environ -130.91 kg. Il convient de noter que dans la pratique, cette valeur n'a pas de sens car il n'existe pas de taille nulle.
- **height** : Le coefficient de la taille est de 1.15009 avec une erreur standard de 0.06749. Le t-value élevé (17.04) et la valeur p quasi nulle indiquent que la taille est un prédicteur significatif du poids. Ainsi, pour chaque unité d'augmentation de la taille, le poids augmente en moyenne de 1.15 kg.
- **Le coefficient de détermination R^2 :** Le coefficient de détermination est de 0.5946, ce qui signifie que 59.46 % de la variation du poids est expliquée par la taille. Cela indique que la taille est un facteur majeur dans l'explication du poids, bien qu'une partie de la variabilité du poids reste non expliquée par ce modèle.
- **Statistique F et valeur p associée :** La statistique F est de 290.4 avec une valeur p très faible, indiquant une significativité globale du modèle.

Ainsi, L'équation de régression obtenue est :

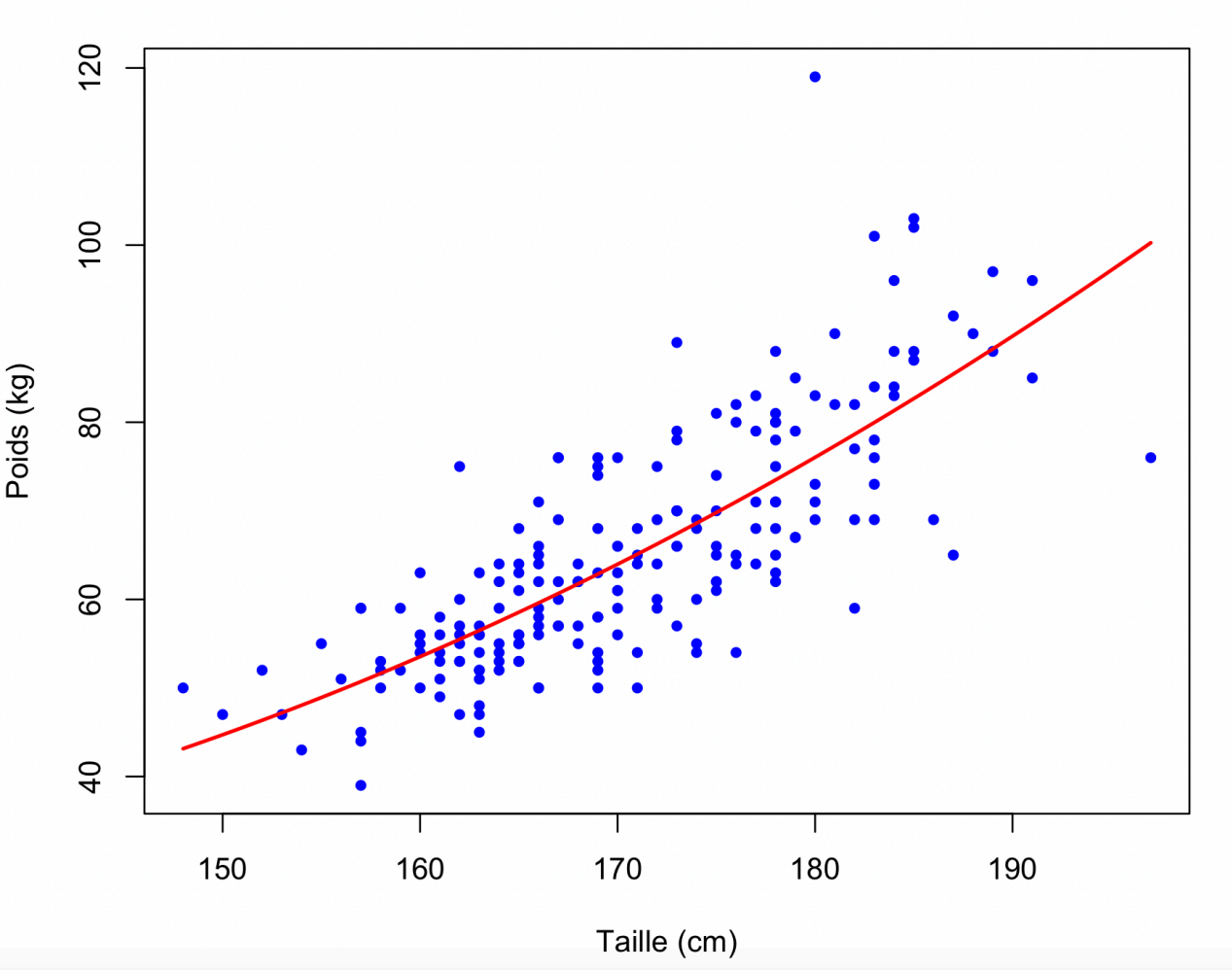
$$Poids = -130.91 + 1.15 \times Taille$$

En somme, notre analyse de régression linéaire simple a révélé une relation significative entre le poids et la taille Cependant, une observation minutieuse du graphique a révélé une tendance des points à suivre une courbe plutôt qu'une droite, suggérant la présence de non-linéarités dans la relation.

Ainsi, pour une compréhension plus approfondie, nous envisagerons dans la section suivante l'utilisation d'un modèle de régression polynomiale pour mieux capturer cette complexité et explorer davantage la relation entre le poids et la taille.

Régression Polynomiale :

L'approche de la régression polynomiale vise à modéliser des relations plus complexes entre les variables en introduisant des termes polynomiaux dans le modèle. Contrairement à la régression linéaire, qui suppose une relation linéaire entre les variables, la régression polynomiale permet de capturer des formes de relations non linéaires telles que des courbes ou des pics. En ajoutant des termes polynomiaux, tels que X^2 ou X^3 au modèle, cette méthode offre une flexibilité accrue pour mieux ajuster les données observées, ce qui peut conduire à une compréhension plus approfondie des relations entre les variables.



(Figure 3 : Graphique de Régression Polynomiale)

Analyse de Graphique de Régression Polynomiale :

En observant le graphique, on remarque que la courbe de régression polynomiale suit étroitement la distribution des points de données. Cela indique que le modèle polynomial est bien ajusté aux données, capturant ainsi de manière plus précise la relation entre le poids et la taille. En comparaison avec le modèle précédent, cette observation suggère que la modélisation polynomiale offre une représentation plus judicieux de la relation entre ces variables

```
Residuals:
    Min      1Q  Median      3Q     Max
-24.265  -5.159  -0.499   4.549  42.965

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.117140 175.246872   0.611    0.542
height       -1.632719   2.045524  -0.798    0.426
h2            0.008111   0.005959   1.361    0.175

Residual standard error: 8.486 on 197 degrees of freedom
Multiple R-squared:  0.5983,    Adjusted R-squared:  0.5943
F-statistic: 146.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

Interprétation :

- **Intercept :** L'intercept est de 107.117140. Cela représente la valeur estimée du poids lorsque la taille est nulle. Cependant, il est important de noter que cette interprétation n'a pas de sens pratique car une taille de zéro n'existe pas dans la réalité. Ce coefficient n'est pas statistiquement significatif (p-value = 0.542), ce qui signifie qu'il n'est pas différent de zéro de manière significative.

- **height** : Le coefficient de la hauteur est de -1.632719. Ce coefficient n'est pas statistiquement significatif (p-value = 0.426), suggérant que la relation linéaire entre la taille et le poids n'est pas significative dans ce modèle.

- **h2 (taille carrée) :** Le coefficient du terme quadratique est de 0.008111. Ce coefficient n'est pas statistiquement significatif non plus (p-value = 0.175), suggérant qu'il n'y a pas de relation quadratique significative entre la taille et le poids dans ce modèle.

- **Multiple** R^2 : La valeur de 0.5983 indique que 59.83% de la variance du poids est expliquée par le modèle polynomial. Cela signifie qu'une part importante de la variation du poids est capturée par ce modèle, mais qu'il reste encore une proportion non négligeable de la variation qui n'est pas expliquée par ce modèle.

- R^2 **ajusté :** Le R2 ajusté est de 0.5943. Il s'agit d'une version corrigée du R2 qui tient compte du nombre de variables explicatives et du nombre d'observations. Dans ce cas, le R2 ajusté est légèrement inférieur au multiple R2, ce qui est attendu
- **Statistique F et valeur p associée :** La statistique F est évaluée à 146.7, avec une valeur p très faible. Cela suggère une significativité globale du modèle polynomial.

En résumé,Bien que ce modèle ait démontré une capacité à expliquer une part significative de la variation du poids, l'insignifiance statistique de certains coefficients soulève des interrogations quant à sa pertinence complète. Par conséquent, la nécessité d'explorer d'autres modèles reste primordiale pour obtenir une compréhension plus précise et robuste de cette relation complexe.

Synthèse :

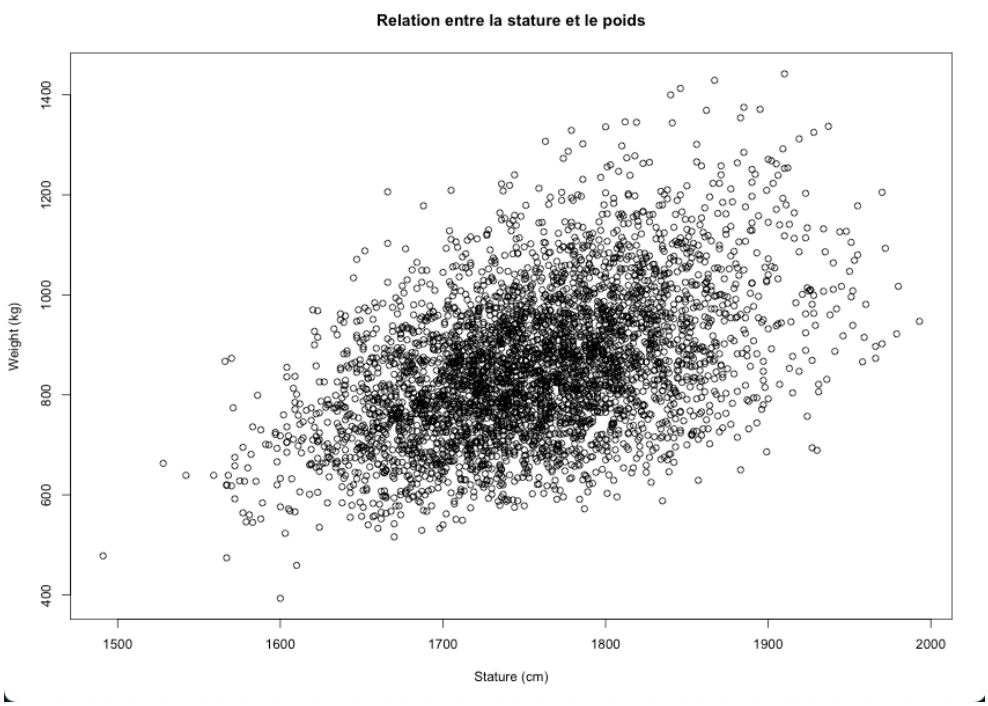
L'analyse comparative des modèles de régression linéaire simple et polynomiale appliqués aux données de Davis a révélé des insights significatifs. Bien que la régression polynomiale semble mieux expliquer la variance du poids, les coefficients associés aux termes polynomiaux manquent de signification statistique, suggérant une absence de relation significative entre ces termes et le poids. En revanche, le modèle de régression linéaire simple a démontré des coefficients significatifs, indiquant une relation significative entre la taille et le poids. Ainsi, pour une compréhension plus approfondie et fiable de la relation entre la taille et le poids, il est essentiel d'explorer d'autres modèles et approches analytiques.

Présentation du jeu de données ANSUR II :

Le jeu de données ANSUR II contient des mesures anthropométriques de plus de 6 000 membres adultes de l'armée américaine, avec un total de 105 variables. Ces mesures comprennent la stature, le poids, et une variété de dimensions corporelles telles que la circonférence du poignet et la longueur de la main. En plus des données physiques, des informations démographiques telles que le genre, l'âge, et le poids sont également incluses.

Dans cette section, notre objectif était d'explorer la relation entre la stature et le poids chez les hommes à travers une régression linéaire. Nous avons utilisé des données anthropométriques de l'US Army, nous concentrant spécifiquement sur les variables de poids en kilogrammes et de stature en centimètres. L'analyse visait à déterminer si la stature pouvait être un facteur significatif pour expliquer le poids des individus.

Graphique de régression linéaire :



(Figure 4 : Graphique de la régression linéaire)

Le graphique de la régression linéaire entre la stature et le poids révèle une relation linéaire entre ces deux variables. Il met en évidence une tendance générale où le poids semble augmenter avec la taille, suggérant ainsi une relation positive entre ces deux aspects anthropométriques.

```
Residuals:
    Min      1Q  Median      3Q     Max
-343.88  -86.95   -6.83   79.50  470.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -853.08972   50.41611  -16.92   <2e-16 ***
stature        0.97273    0.02869   33.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125.6 on 4080 degrees of freedom
Multiple R-squared:  0.2199,    Adjusted R-squared:  0.2197
F-statistic:  1150 on 1 and 4080 DF,  p-value: < 2.2e-16
```

Interprétation :

- **Intercept :** L'intercept est de -853.08972. dans le modèle de régression linéaire représente le poids estimé lorsque la stature est nulle. Cependant, dans le contexte anthropométrique, cette valeur n'a pas de signification pratique. elle est statistiquement significative avec un p-value très faible, indiquant une relation significative entre la stature et le poids.
- **stature** :Le coefficient de la stature (0.97273) indique que, en moyenne, chaque centimètre supplémentaire de taille est associé à une augmentation de 0.97273 kg du poids. Ce coefficient est également hautement significatif (p-value < 2e-16), confirmant ainsi une relation linéaire positive forte entre la taille et le poids.

- **Multiple** R^2 :Le coefficient de détermination (Multiple R-squared) de 0.2199 indique que 21.99% de la variance du poids est expliquée par la stature. Bien que cette proportion soit significative, cela signifie également qu'une grande partie de la variabilité du poids reste inexpliquée par la taille seule.

- **Statistique F et valeur p associée :** La statistique F est évaluée à 1150, avec une valeur p très faible. Cela suggère que l'ensemble du modèle de régression est statistiquement significatif. Cela confirme que la relation entre la stature et le poids est globalement significative, ce qui rend le modèle utile pour expliquer la variation observée dans le poids.

Ainsi, L'équation de la régression linéaire obtenue est la suivante :

$$Poids = -853.08972 + 0.97273 \times stature$$

Cette équation indique que pour chaque augmentation d'un centimètre de la stature, le poids augmente en moyenne de 0.97273 kg.

En résumé, nos résultats démontrent une relation positive et significative entre la taille et le poids chez les hommes. Cependant, le coefficient de détermination Multiple R-squared indique que d'autres facteurs non inclus dans ce modèle ont également une influence sur le poids. Afin d'améliorer la précision de notre modèle, il serait judicieux d'explorer d'autres variables anthropométriques ou d'envisager des transformations non linéaires.

L'Analyse en Composantes Principales (ACP):

L'Analyse en Composantes Principales (ACP) est une technique statistique utilisée pour réduire la dimensionnalité d'un ensemble de données tout en conservant autant que possible la variabilité présente dans celles-ci. Elle transforme des variables corrélées en un plus petit nombre de variables non corrélées appelées composantes principales. Le but de l'ACP est de simplifier l'analyse des données complexes, de faciliter la visualisation et de révéler les structures sous-jacentes.
Le graphique des individus représente les observations projetées dans l'espace des composantes principales, permettant de visualiser les similarités et différences entre les individus. Le graphique des variables montre les variables originales dans le même espace, illustrant leur contribution aux composantes principales et leurs relations. Ces graphiques sont essentiels pour interpréter les résultats de l'ACP, en identifiant les principaux facteurs de variation et les regroupements d'individus ou de variables.

Pour réaliser l'analyse en composantes principales (ACP), nous utilisons le jeu de données ANSUR. Et afin de se concentrer sur les mesures pertinentes, certaines variables doivent être exclues, notamment les identifiants individuels et les informations non pertinentes telles que le genre, la date, le lieu de naissance, etc. Les variables exclues sont : "subjectid", "Gender", "Date", "Installation", "Component", "Branch", "PrimaryMOS", "SubjectsBirthLocation", "SubjectNumericRace", "Ethnicity", "DODRace", "WritingPreference", "Age", "Heightin", et "Weightlbs".

Après avoir identifié et exclu ces variables, nous calculons la quantité de variance expliquée par chaque composante pour comprendre l'importance relative de chaque dimension. Ensuite, nous procédons à l'ACP pour explorer la structure et les relations sous-jacentes des données anthropométriques.

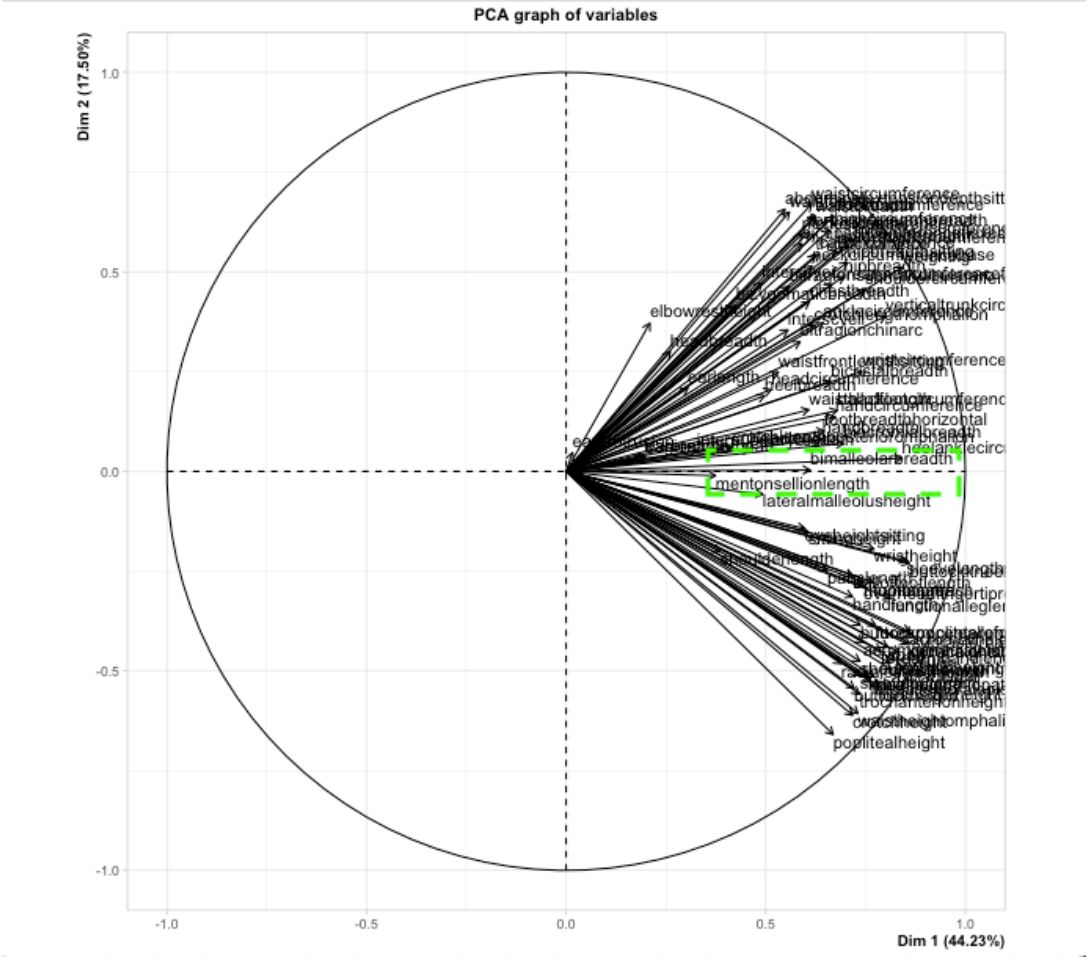
```
PCA(X = colonnes_pertinentes)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance              41.138  16.272   3.912   3.021   2.125   1.969   1.755
% of var.             44.235  17.497   4.207   3.248   2.285   2.117   1.887
Cumulative % of var.  44.235  61.731  65.938  69.186  71.471  73.588  75.476
                       Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
Variance               1.383   1.325   1.241   1.155   1.080   0.867   0.858
% of var.              1.487   1.424   1.334   1.242   1.161   0.932   0.922
Cumulative % of var.  76.963  78.387  79.721  80.962  82.124  83.056  83.978
```

Interprétation :

Les deux premières dimensions de l'ACP expliquent environ 61,73% de la variance totale, ce qui est significatif. Cela indique que ces dimensions capturent une part importante des schémas de variation les plus significatifs dans les données. Les valeurs propres décroissantes suggèrent également que la variance est concentrée principalement dans les premières composantes, ce qui justifie l'utilisation de ces dimensions pour réduire la dimensionnalité des données.

Graphique des variables :

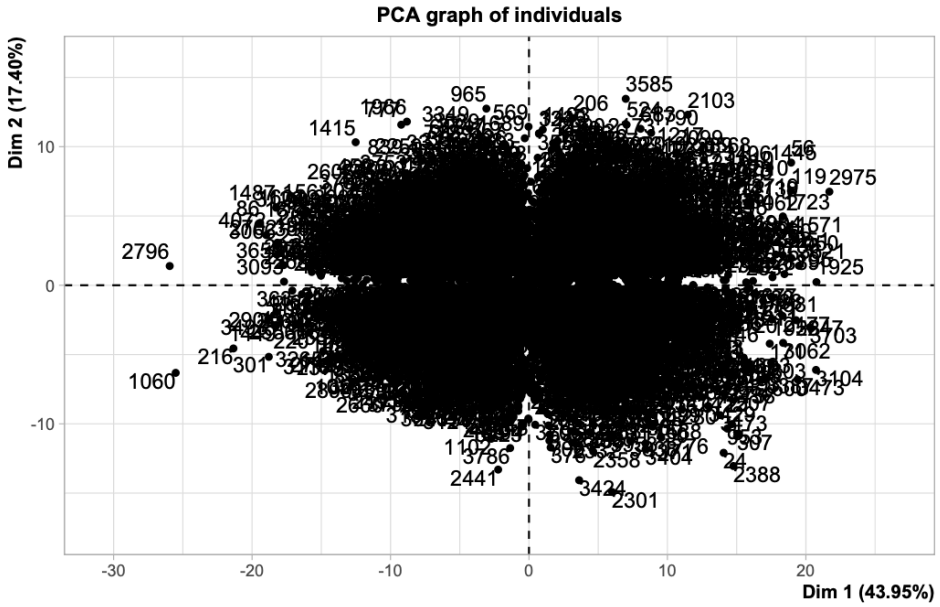


(Figure 5 : Graph des variables)

Interprétation :

L'analyse du graphe des variables révèle des relations complexes entre les différentes mesures anthropométriques. Dans la première dimension, toutes les variables présentent une corrélation positive, indiquant que les individus avec des coordonnées élevées sur cette dimension ont tendance à avoir des mesures anthropométriques élevées, et vice versa. Parmi ces variables, "mentonsellionlength" et "bimalleolarbreadth" se distinguent par leur forte corrélation avec cette première dimension (Figure 5). Quant à la deuxième dimension, les variables se regroupent en deux ensembles distincts sur le graphe des variables. Cela suggère une séparation entre les individus ayant des indices anthropométriques élevés pour le premier groupe de variables et des indices plus faibles pour le deuxième groupe, et vice versa. La proximité des variables sur le graphe des variables indique des similarités ou des corrélations entre ces dernières, offrant ainsi un aperçu des schémas de variation complexes dans les données anthropométriques.

Graphique des individus :



(Figure 6 : Graph des individus)

Interprétation :

Ce graphique présente la répartition des individus sur les deux premières dimensions principales résultant de l'Analyse en Composantes Principales (ACP). Les deux dimensions principales capturent environ 61.82% de la variance totale, ce qui est significatif pour une compréhension globale de la structure des données.

Les points, chacun associé à un individu unique, sont majoritairement concentrés autour de l'origine, indiquant que la plupart des individus sont proches de la moyenne dans ces nouvelles dimensions. Les points éloignés de l'origine représentent des individus avec des mesures anthropométriques extrêmes : des coordonnées élevées sur les axes principaux indiquent des individus ayant des mesures corporelles plus élevées que la moyenne, et vice versa.

Au sein de chaque groupe principal de données, on remarque la formation de sous-groupes distincts. Cela indique que les individus de chaque groupe possèdent des caractéristiques morphologiques semblables qui les différencient des autres sous-groupes. La répartition symétrique des points autour de l'origine suggère que les deux premières dimensions principales capturent des variations importantes dans les deux directions sans créer de clusters distincts. Toutefois, l'identification de sous-groupes révèle des différences subtiles mais significatives dans les mesures corporelles de la population étudiée.
Les coordonnées élevées sur les deux axes principaux pour certains groupes montrent qu'ils possèdent des mesures anthropométriques supérieures à la moyenne. Inversement, les groupes avec des coordonnées faibles ont des mesures inférieures. À l'intérieur de ces groupes, les sous-groupes se démarquent par leurs caractéristiques morphologiques distinctes, ce qui révèle des similitudes internes et des différences externes notables.

Cette visualisation fournit une vue d'ensemble précieuse des variations principales dans les données, facilitant ainsi l'identification des dimensions clés et des anomalies potentiellement intéressantes pour des études futures. Elle aide également à comprendre les similitudes et les différences morphologiques entre les individus, ainsi qu'à identifier les groupes et sous-groupes présentant des caractéristiques spécifiques.
En résume, ce graphique issu de l'ACP permet non seulement de réduire la complexité des données, mais aussi de révéler des patterns significatifs et des relations complexes entre les variables corporelles, offrant ainsi une meilleure compréhension de la structure des données anthropométriques étudiées.

Synthèse :

En conclusion, l'Analyse en Composantes Principales (ACP) appliquée aux données anthropométriques de l'ANSUR a permis de réduire la dimensionnalité de l'ensemble de données tout en conservant une grande partie de sa variabilité. Les deux premières dimensions de l'ACP ont capturé une part importante des schémas de variation les plus significatifs, offrant ainsi un aperçu des relations sous-jacentes entre les mesures anthropométriques. L'interprétation du graphe des variables a révélé des corrélations complexes entre différentes variables, mettant en évidence des schémas de variation et des regroupements potentiels dans les données.

CONCLUSION :

En guise de conclusion, ce projet souligne l'importance cruciale des méthodes statistiques telles que la régression linéaire et l'analyse en composantes principales (ACP) dans l'étude des relations entre les variables. À travers l'analyse de la corrélation entre le poids et la taille dans deux ensembles de données distincts, nous avons conduit des analyses approfondies, utilisé des outils visuels pour identifier des tendances, établi des modèles pour quantifier ces relations, et interprété les résultats obtenus. Cette démarche met en évidence la valeur et la puissance des outils analytiques dans l'exploration des données, fournissant ainsi un aperçu précieux des relations complexes entre les variables.