Projet Régression linéaire multiple

Université de Paris Nanterre

Anissa Goulif - Radwan SARMINI DET SATOUF - Sirine Gabriel

March 29, 2021

Jeu de données

Données bodyfat.

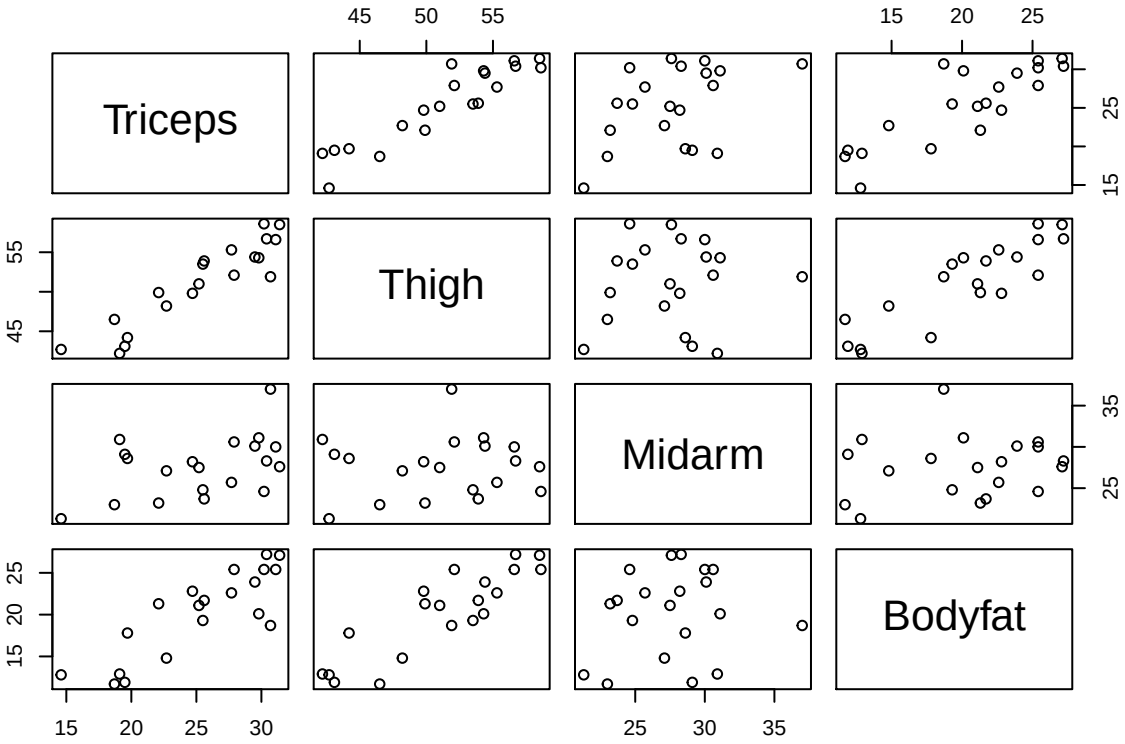
Nous travaillons dans cette partie sur le jeu de données bodyfat, disponible sur la plateforme Cours en Ligne. Le jeu de données est tiré de la page : https://online.stat.psu.edu/stat501/lesson/5/5.5 sur laquelle il est décrit ainsi : For a sample of n = 20 individuals, we have measurements of y = body fat, x1 = triceps skinfold thickness, x2= thigh circumference, and x3 = midarm circumference.

```
## 'data.frame':    20 obs. of  4 variables:
##  $ Triceps: num  19.5 24.7 30.7 29.8 19.1 25.6 31.4 27.9 22.1 25.5 ...
##  $ Thigh  : num  43.1 49.8 51.9 54.3 42.2 53.9 58.5 52.1 49.9 53.5 ...
##  $ Midarm : num  29.1 28.2 37 31.1 30.9 23.7 27.6 30.6 23.2 24.8 ...
##  $ Bodyfat: num  11.9 22.8 18.7 20.1 12.9 21.7 27.1 25.4 21.3 19.3 ...
```

Question 1

L'option header=T dans l'appel R à read.table sert à dire à R que les données ont déjà des noms de columns (variables) et pas besoin d'installer des noms de columns par défaut.

Question 2



Cette représentation graphique est un type de graphique ou de diagramme mathématique utilisant des coordonnées cartésiennes pour afficher les valeurs de deux variables typiques d'un ensemble de données. Les données sont affichées sous forme d'une collection de points, chacun ayant la valeur d'une variable déterminant la position sur l'axe horizontal et la valeur de l'autre variable déterminant la position sur l'axe vertical. Cette fonction permet en fait de représenter les corrélations

entre les variables 2 à 2.

On voit des corrélations positives assez fortes entre :

- Triceps et Thigh
- Triceps et Bodyfat
- Thigh et Bodyfat

Nous pouvons confirmer cela à l'aide du tableau donnant les corrélations :

```
##           Triceps     Thigh    Midarm   Bodyfat
## Triceps 1.0000000 0.9238425 0.4577772 0.8432654
## Thigh   0.9238425 1.0000000 0.0846675 0.8780896
## Midarm  0.4577772 0.0846675 1.0000000 0.1424440
## Bodyfat 0.8432654 0.8780896 0.1424440 1.0000000
```

Les autres graphes représentent des points dispersés de manière insignificante.

Question 3

Ces modèles statistiques correspondent à des modèles de regression lineaire.

LM est utilisé pour ajuster des modèles linéaires. Il peut être utilisé pour effectuer une régression, une analyse de variance à une seule strate et une analyse de covariance.

- Résidus : Cette section résume les résidus, l'erreur entre la prédiction du modèle et les résultats réels.

Les résidus les plus petits sont les meilleurs.

Coefficients : Pour chaque variable et l'intercept, un poids est produit et ce poids a d'autres attributs comme l'erreur standard, une valeur de t-test et la signification.

- Estimation : C'est le poids donné à la variable.

Les coefficients sont comme suit:

```
## $coefficients
## (Intercept)     Triceps       Thigh      Midarm
##  117.084695    4.334092   -2.856848   -2.186060
```

```
##
## Call:
## lm(formula = Bodyfat ~ Triceps + Thigh + Midarm, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## Triceps        4.334      3.016   1.437    0.170
## Thigh         -2.857      2.582  -1.106    0.285
```

```
## Midarm        -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 on 16 degrees of freedom
## Multiple R-squared:  0.8014, Adjusted R-squared:  0.7641
## F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Erreur standard : indique avec quelle précision l'estimation a été mesurée. Elle n'est vraiment utile que pour le calcul de la valeur t. Valeur t et Pr(>[t]) : La valeur t est calculée en prenant le coefficient divisé par l'erreur standard. Elle est ensuite utilisée pour tester si le coefficient est significativement différent de zéro ou non. Si elle n'est pas significative, ' le coefficient n'apporte rien au modèle et peut être abandonné ou étudié plus avant. Pr(>|t|) est le niveau de signification.'

- Mesures de performance : Trois ensembles de mesures sont fournis. Erreur standard résiduelle : Il s'agit de l'écart type des résidus. Plus elle est petite, c'est mieux. R-carré multiple / ajusté : Pour une seule variable, la distinction n'a pas vraiment d'importance. Le R-carré indique la quantité de variance expliquée par le modèle. Le R-carré ajusté prend en compte le nombre de variables et est plus utile pour les régressions multiples.
- F-Statistique : Le test F vérifie si le poids d'au moins une variable est significativement différent de zéro. Il s'agit d'un test global pour aider à évaluer un modèle. Si la valeur p n'est pas significative (par exemple, supérieure à 0,05), alors votre modèle ne fait essentiellement rien.

Les modèles fit1 et fit2 sont les mêmes, à la différence de fit2, dans le modèle fit1 les variables explicatives de la variable bodyfat sont explicitées (fit1 : on a choisi à ma main les variables et fit2 : on utilise "." pour sélectionner toutes les variables). Ce sont des modèles linéaires multiples décrivant la variable bodyfat par les autres variables des données (Triceps, Thigh, Midarm). Pour ces modèles le coefficient de détermination R^2 est de 0.7641. Ce coefficient est un indicateur de qualité du modèle. Ainsi avec un R^2 de 0.7641, les modèles fit1 et fit2 sont capables de déterminer 76,41% de la distribution des points.

Formulation matricielle : Ici la variable bodyfat est la variable à expliquer. Nous noterons Y le vecteur de taille 20 qui représentera la variable réponse, la matrice X de taille (20x4), les variables explicatives sont ici Triceps, Thigh et Midarm, la première colonne de X contient le vecteur de taille 20 constitué de 1 et les autres colonnes sont les valeurs de Triceps, Thigh et Midarm. Nous noterons A le vecteur de taille 4 constitué des paramètres du modèle, enfin nous noterons E le vecteur de taille 20 des résidus. Le modèle s'écrit donc matriciellement : Y = XA + E.

Le modèle fit0 est un modèle dont la variable réponse est bodyfat mais pour pour lequel il n'y a pas de variable explicative prise en compte. La seule sortie sera l'intercept. Ce n'est pas un modèle pertinant et interessant à étudier. La forme matricielle de ce modèle serait Y = X.Intercept + E, où X est un vecteur de taille 20 composé de 1, Intercept(= 20.195) le paramètre du modèle fitO et E le vecteur de taille 20 des résidus.

Question 4

L'équation de prédiction selon fit1 est : $Bodyfat = 4.334 * Triceps - 2.857 * Thigh - 2.186 * Midarm + 117.085$.

On implémente donc sous R la fonction réalisant ces prédictions (voir code R).

Question 5

- "Residual standard error: 2.48 on ?? degrees of freedom" Ici la valeur manquante est 16, il s'agit du degrés de liberté. Ce dernier est donné par : (n-p), où n est le nombre d'observations (ici n=20) et p est le nombre de variables (ici p=4).
- "F-statistic: 21.52 on ?? and ?? DF, p-value: 7.343e-06" Les valeurs manquantes ici sont 3 et 16 qui sont les degrés de liberté pour le test de Fisher étant ((p-1),(n-p)).
- la t value pour Thigh est donnée par Estimate/Std. Error de Thigh, la valeur manquante ici est donc :

```
## [1] -1.106441
```

- Pr(>|t|) pour Triceps est donné par 2*pt(abs(t), ddl, lower.tail = FALSE), la valeur manquante est donc :

```
## [1] 0.1699111
```

Question 6

Pour calculer la matrice $(X^tX)^{-1}$ comme sur le site :

Tout d'abord on calcule Mean standard errors avec 4 degrés de liberté vu qu'ils ont 4 variables :

```
## [1] 6.150306
```

On divise la matrice de covariance par le MSE, on trouve :

```
##             (Intercept)    Triceps       Thigh      Midarm
## (Intercept)  1618.86721 48.8102522 -41.8487041 -25.7987855
## Triceps        48.81025  1.4785133  -1.2648388  -0.7785022
## Thigh         -41.84870 -1.2648388   1.0839791   0.6657581
## Midarm        -25.79879 -0.7785022   0.6657581   0.4139009
```

Voila donc la matrice $(X^tX)^{-1}$.

Question 7

Pour recalculer les valeurs fournies dans la colonne Std. Error de l'objet fit1, on calcule, pour chaque valeur, la racine carrée de chaque coefficient correspondant dans la matrice variance covariance de fit1 (en utilisant donc la fonction vcov). On prend alors, par exemple pour la valeur Std. Error de l'intercept, la racine carrée de la valeur [1,1] de vcov (fit1).

```
## [1] 99.7824
```

```
## [1] 3.015511
```

```
## [1] 2.582015
```

```
## [1] 1.595499
```

On retrouve bien les valeurs fournies dans la colonne std. error de l'objet fit1.

Question 8

Thigh semble jouer un rôle significatif car il y a un correlation positive significative entre Thigh et Bodyfat :

```
## [1] 0.8780896
```

De plus si on construit un modele de regression linéaire simple entre Bodyfat et Thigh :

```
##
## Call:
## lm(formula = Bodyfat ~ Thigh, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4949 -1.5671  0.1241  1.3362  4.4084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.6345     5.6574  -4.178 0.000566 ***
## Thigh         0.8565     0.1100   7.786  3.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 18 degrees of freedom
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.7583
## F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07
```

On remarque que R-squared = 0.771 donc 77.1% de données de Thigh explique Bodyfat et avec p-value « 0.05 où on est amené à rejeter l'hypothèse d'indépendance.

On en conclut que la variable Thigh joue un role significatif.

Question 9

On va tout d'abord utiliser la méthode stepwise pour voir quelles variables sont plus significatives par rapport à d'autres. Penchons nous en premier lieu sur la méthode ascendante de la méthode stepwise, c'est-à-dire qu'on part en premier lieu d'un modèle sans variable (ici fit0) pour y ajouter au fur et à mesure des variables qui diminuent le critère (ici l'AIC). La méthode s'arrête lorsque l'ajout d'une nouvelle variable ne diminue plus le critère (l'AIC ici).

```
## Start:  AIC=66.19
## Bodyfat ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + Thigh    1    381.97 113.42 38.708
## + Triceps  1    352.27 143.12 43.359
## <none>                 495.39 66.192
## + Midarm   1     10.05 485.34 67.782
##
```

```
## Step:  AIC=38.71
## Bodyfat ~ Thigh
##
##           Df Sum of Sq    RSS    AIC
## <none>                 113.42 38.708
## + Triceps  1    3.4729 109.95 40.086
## + Midarm   1    2.3139 111.11 40.296

##
## Call:
## lm(formula = Bodyfat ~ Thigh, data = dat)
##
## Coefficients:
## (Intercept)        Thigh
##    -23.6345       0.8565
```

Ici on remarque donc que le modèle s'arrête après l'ajout de thigh. R considère alors que l'ajout de triceps ou de midarm va faire augmenter l'AIC (donc que ces variables vont "détériorer" le modèle car moins significative).

Testons maintenant la méthode déscendante. En effet, pour cette méthode on part d'un modèle modèle complet, donc un modèle avec toutes les variables (ici fit1) et R enlève au fur et à mesure la variable qui diminue le plus le critère AIC. Le programme s'arrête lorsqu'on ne diminue plus le critère.

```
## Start:  AIC=39.87
## Bodyfat ~ Triceps + Thigh + Midarm
##
##           Df Sum of Sq     RSS    AIC
## - Thigh    1    7.5293 105.934 39.342
## <none>                  98.405 39.867
## - Midarm   1   11.5459 109.951 40.086
## - Triceps  1   12.7049 111.110 40.296
##
## Step:  AIC=39.34
## Bodyfat ~ Triceps + Midarm
##
##           Df Sum of Sq    RSS    AIC
## <none>                 105.93 39.342
## - Midarm   1     37.19 143.12 43.359
## - Triceps  1    379.40 485.34 67.782

##
## Call:
## lm(formula = Bodyfat ~ Triceps + Midarm, data = dat)
##
## Coefficients:
## (Intercept)      Triceps       Midarm
##      6.7916       1.0006      -0.4314
```

Encore une fois ici on en conclue que ni triceps ni midarm ne diminue le critère d'AIC significative-

ment par rapport à thigh.

Grace à ces deux méthodes que nous venons de réaliser, nous pouvons en conclure que triceps et midarm ne semblent pas jouer un rôle siginificatif.

Question 10

Le modèle

On remarque, grace à la question précédente, que les modèle ayant le critère AIC le plus faible (donc le meilleur modèle au sens de l'AIC) est le modèle avec uniquement Thigh comme variable (AIC = 38.71).

Interessons-nous à ce modèle :

```
##
## Call:
## lm(formula = Bodyfat ~ Thigh, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4949 -1.5671  0.1241  1.3362  4.4084
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.6345     5.6574  -4.178 0.000566 ***
## Thigh         0.8565     0.1100   7.786  3.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 18 degrees of freedom
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.7583
## F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07
```

Les variables sont plutôt bien pour la P value et 75.83% des données de Thigh expliquent Bodyfat avec 2.51 de residual standard error.

Et pour valider notre modèle, nous avons choisi de tester le modèle. Nous allons diviser notre jeu de données : 85% de données pour entraîner le modele et 15% pour tester le modèle (cf code R). Pour se faire et pour être sûr de notre choix, on choisit de réaliser le test 10 fois en prenant 10 fois des données aléatoirement dans notre jeu de données (pour ne pas faire le test sur les memes données sachant que nous prenons 85% pour entrainer et 15% pour tester). De plus,nous allons faire le test sur 2 modèles : le modèle comprenant uniquement thigh et le modèle comprenant uniquement midarm et triceps. Pour comparer nos tests on utilise la valeur de la MSE (mean square error).

Resultats

```
## [1] "modèle Thigh : 1.50462000952166"
## [1] "modèle Triceps Midarm : 1.51733827627084"
## [1] "  "
## [1] "modèle Thigh : 1.10097558554294"
```

```
## [1] "modèle Triceps Midarm : 1.48585304897833"
## [1] "  "
## [1] "modèle Thigh : 0.397313457374898"
## [1] "modèle Triceps Midarm : 0.316114694511624"
## [1] "  "
## [1] "modèle Thigh : 1.09147016645875"
## [1] "modèle Triceps Midarm : 1.11129759621598"
## [1] "  "
## [1] "modèle Thigh : 0.689398482959585"
## [1] "modèle Triceps Midarm : 0.825554222009351"
## [1] "  "
## [1] "modèle Thigh : 0.89750040975377"
## [1] "modèle Triceps Midarm : 1.08953409090954"
## [1] "  "
## [1] "modèle Thigh : 1.43641110716472"
## [1] "modèle Triceps Midarm : 1.4353442676611"
## [1] "  "
## [1] "modèle Thigh : 1.40767066056203"
## [1] "modèle Triceps Midarm : 1.37657578361842"
## [1] "  "
## [1] "modèle Thigh : 0.497832609706941"
## [1] "modèle Triceps Midarm : 0.580770524260564"
## [1] "  "
## [1] "modèle Thigh : 1.0890554199575"
## [1] "modèle Triceps Midarm : 1.20336199948328"
## [1] "  "

## [1] 3
```

On remarque que la MSE est globalement toujours plus faible avec le modèle comprenant uniquement thigh. De plus, le chiffre à la fin fait référence à un compteur que nous avons mis en place pour compter le nombre de fois où le modèle thigh est moins bon que le modèle avec triceps et midarm au sens de la MSE. Globalement, ce chiffre est généralement égale à 3 (il arrive qu'il atteigne 5). Ces tests nous font permette alors de dire, une fois de plus, que le modèle avec uniquement thigh est globalement meilleur.

Pour conclure, nous choisissons alors le modèle avec uniquement thigh, donc le modele : Bodyfat ~ Thigh , (r squared bon, AIC le plus faible et MSE faible).