

---

## TP 3 Modèle linéaire

---

Les données considérées dans ce TP sont contenues dans le fichier `fitness.csv`. Pour chaque individu, on a mesuré lors de séances de sport les 7 variables suivantes

- `age`
- `weight` : poids de l'individu
- `oxy` : consommation d'oxygène
- `runtime` : temps de l'effort
- `rstpulse` : mesure de pulsation cardiaque avant effort
- `runpulse` : mesure de pulsation cardiaque moyenne pendant l'effort
- `maxpulse` : mesure de pulsation cardiaque maximale pendant l'effort

On va s'intéresser à la relation entre la mesure `oxy` et les autres.


## Préliminaires et visualisations

1. Commencez par charger les données et afficher les résumés à l'aide des fonctions `str` et `summary`.
2. On peut observer visuellement l'ensemble des interactions entre les différentes variables du jeu de donnée via la fonction `plot` :

```
plot(fitness)
```

Si on souhaite s'intéresser uniquement à l'influence de chaque variable sur la variable `oxy`, la commande suivante suffit

```
plot(oxy ~ ., data = fitness)
```

Il vous sera alors proposé de parcourir tous les graphiques, en pressant . On peut également observer une interaction en particulier, par exemple `oxy` en fonction de `maxsp` :

```
plot(oxy ~ maxpulse, data = fitness)
```

3. Calculer la matrice de corrélation du jeu de données avec la fonction `cor`, puis représenter la graphiquement avec la fonction `corrplot` (nécessite l'installation du package `corrplot`). Commentez le graphique ainsi obtenu.

## 1 Régression linéaire multiple

On va à présent considérer un modèle avec  $p$  prédicteurs pour exprimer la variable `oxy` en fonction des  $p = 6$  autres variables, c'est-à-dire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

où  $Y$  est un vecteur colonne représentant les  $n$  mesures de `oxy`, les  $X_j$  sont des vecteurs colonnes représentant les prédicteurs, et  $\varepsilon$  est un vecteur gaussien de taille  $n$  centré autour du vecteur nul et de matrice de covariance  $\sigma^2 I_n$ . Sous forme matricielle, le modèle s'écrit donc

$$Y = X\beta + \varepsilon$$

où  $X$  est une matrice avec  $n$  lignes et  $p+1$  colonnes : la 1ère colonne ne contenant que des 1 (le terme “d’intercept”) et les autres colonnes correspondant aux  $X_j$ , et  $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ .

Le produit matriciel entre 2 matrices s’effectue dans R avec l’opérateur `%*%`, l’inversion d’une matrice carrée via la fonction `solve`, et la transposée est simplement obtenue avec `t`

```
M%*%N ## produit matriciel entre la matrice M et la matrice N
solve(M) ## inverse de la matrice M
t(M) ## transposée de la matrice M
```

1. Calculer dans R l’estimation du vecteur  $\beta$ . Ces estimations sont-elles cohérentes avec les diagrammes obtenus en début de TP?
2. Calculer dans R les valeurs ajustées au jeu de données, ainsi que les résidus de la régression.
3. Calculer l’estimation de  $\sigma^2$  ainsi que celle de la matrice de covariance de l’estimateur du vecteur  $\beta$ . Repérer en particulier les estimations de la variance des estimateurs de chaque  $\beta_j$ .
4. Calculer les réalisations des intervalles de confiance au niveau 95% de chaque  $\beta_i$ .

On va chercher à s’intéresser l’influence des prédicteurs sur `oxy`. Pour cela on va procéder en 2 étapes

5. Éprouvez dans R dans un premier temps l’hypothèse globale (1) à partir des résultats de l’Exercice 1. Que peut-on conclure?
6. Effectuer à présent les tests associées aux hypothèses (2) à partir des résultats de l’Exercice 1. Que peut-on conclure?

L’ajustement d’un modèle linéaire dans R se fait avec la fonction `lm`. Si on dispose d’un jeu de données `df` contenant des colonnes nommées `VarY` et `VarX1` et `VarX2`, la régression de `VarY` en fonction de `VarX1` et `VarX2` se fait simplement comme ceci

```
lm(VarY ~ VarX1+VarX2, data = df)
```

Si le résultat est stocké dans une variable, on peut alors obtenir différentes informations sur le modèle ainsi construit (voir possibilités avec la fonction `names`), comme par exemple

```
reg = lm(VarY ~ VarX1+VarX2, data = df) # régression linéaire de VarY relativement à VarX1 et VarX2
reg$coefficients # estimations des coefficients de la régression
```

On peut aussi afficher un résumé de la régression avec la commande

```
summary(reg)
```

Pour effectuer une régression de `VarY` par rapport à toutes les autres variables :

```
lm(VarY ~ ., data = df)
```

Dans le cas d’une régression simple, on peut visualiser la droite de régression à l’aide de la commande `abline`

```
plot(VarY ~ VarX, data = df) ## Diagramme de VarY en fonction de VarX
abline(reg) ## Ajout de la droite de régression calculée ci-dessus
```

La fonction `lm` permet en outre d’accéder à différents outils pour vérifier les hypothèses de modélisations.

7. Commencer par afficher les graphiques suivants (supposant que `reg` contient votre modèle ajusté)

```
plot(reg, which = 1)
plot(reg, which = 3)
```

Que représente ces 2 graphiques? Justifier alors qu’ils peuvent servir à évaluer l’hypothèse d’homoscédasticité du modèle.

8. On peut également tester cette hypothèse avec le test de Breusch-Pagan, dont l’hypothèse nulle est l’homoscédasticité, disponible avec la fonction `bptest` du package `lmtest` avec l’usage `bptest(reg)`. Appliquer ce test et conclure.

9. On s'intéresse maintenant à l'hypothèse de normalité des résidus. Proposez un outil graphique ainsi qu'un test statistique pour éprouver cette hypothèse.

## 2 Sélection de variable

Un dernier travail à effectuer lorsqu'on ajuste un modèle multiple est la sélection de variable : il s'agit de réduire la taille du modèle (i.e. le nombre de covariables) typiquement pour éviter tout sur-ajustement du modèle aux données.

### 2.1 Tests entre modèle emboîtés

Les tests entre modèle emboîtés sont un premier moyen d'effectuer une réduction de la taille du modèle.

1. Quelles variables pourraient-on souhaiter retirer du modèle ?
2. Appliquer un test entre modèle emboîtés (q1 de l'Exercice 2) afin de déterminer si oui non il semblerait que ces variables pourraient bien être retirées du modèle.

Néanmoins, se baser sur ce genre de test pour sélectionner des variables n'est pas réellement approprié (q2 et 3 de l'Exercice 2)...

### 2.2 Sélection par pénalisation

Plutôt que de se baser uniquement sur la vraisemblance pour sélectionner les variables, une possibilité repose sur une pénalisation de la vraisemblance, c'est-à-dire chercher à minimiser des quantités de la forme

$$C = -2 \log(\mathcal{L}^*) + 2p\phi(n)$$

où  $\mathcal{L}^*$  est la vraisemblance du modèle maximisée,  $p$  le nombre de variables explicatives considérées et  $\phi$  une fonction. Ainsi minimiser  $C$  permet de trouver un compromis ajustement (via  $\mathcal{L}^*$ ) et complexité ( $p\phi(n)$ ) du modèle.

L'un des critères de sélection de variable pour les modèles linéaires est le critère AIC ("Akaike's Information Criterion"), défini par :

$$AIC = -2 \log(\mathcal{L}^*) + 2k$$

Le critère AIC d'un modèle de régression `reg` préalablement calculé via `lm` peut être obtenu avec la fonction `extractAIC`

```
extractAIC(reg)
```

Le premier élément concerne le nombre de variables explicatives considérées, et le deuxième élément la valeur de l'AIC. La sélection de variable consiste simplement l'identification de variables dont le retrait du modèle fera baisser l'AIC.

1. Sachant que

— Retirer une variable d'un modèle se fait via la fonction `update`

```
reg.new = update(reg, . ~ . -VarZ) ## Retirer la variable VarZ du modèle reg
```

— On peut renseigner une formule de type  $y \sim \tilde{x}$  (nécessaire pour la fonction `update`) via la fonction `formula`

```
formula("y ~ x")
```

— On peut combiner chaîne de caractères et valeurs numériques avec la fonction `paste` :

```
a = 2
paste("a est égal à", a, sep = " ")
```

effectuez la sélection de la première variable à retirer du modèle.

2. Implémenter la réduction totale du modèle.
3. La réduction totale du modèle se fait dans R via la fonction `step`

```
auto.reg = step(reg, direction = "backward")
```

Considérez alors le modèle réduit par le critère AIC et effectuez à nouveau l'étude des résidus. Que constate-t-on?

### 3 Travail théorique

#### Exercice 1 Modèle multiple

On se place dans le contexte de la Section 1, et les hypothèses de simulation associées. **Remarque** : les résultats classiques et potentiellement très utiles sur les vecteurs gaussiens sont donnés en annexe.

1. Rappeler les expressions de l'estimateur  $\hat{\beta}$  de  $\beta$ , de l'estimateur  $\hat{\sigma}^2$  de  $\sigma^2$ , de l'estimateur  $\hat{V}$  de la matrice de covariance de  $\hat{\beta}$ , et de l'estimateur de la variance de  $\hat{\beta}_j$ , pour  $j = 1, \dots, p$ .
2. Vérifier que  $\hat{\beta}$  est un estimateur sans biais de  $\beta$ . Quelle est sa loi? Quelle est la loi de chaque  $\hat{\beta}_j$ ?
3. Montrer que  $\hat{\beta}$  et  $\hat{\sigma}^2$  sont indépendants.
4. Montrer que  $(n - p - 1) \frac{\hat{\sigma}^2}{\sigma^2}$  suit la loi  $\chi^2(n - p - 1)$ .
5. En déduire l'expression des intervalles de confiance au niveau 95% de chaque  $\beta_j$ , pour  $j = 1, \dots, p$ .
6. On souhaite à présent éprouver l'hypothèse globale suivante

$$\mathcal{H}_0 : \{\beta_1 = 0, \dots, \beta_p = 0\} \quad \text{contre} \quad \mathcal{H}_1 : \{\exists j \in \{1, \dots, p\} \text{ t.q. } \beta_j \neq 0\} \quad (1)$$

La prise de décision pour ce test s'effectue grâce à la statistique

$$T = \frac{\|\hat{Y} - \bar{Y}\mathbf{1}_n\|^2}{p} \frac{n - p - 1}{\|Y - \hat{Y}\|^2}$$

dont on va étudier la loi sous  $\mathcal{H}_0$ . On note  $X_{1:p}$  la matrice  $X$  sans la première colonne  $X_0$ . On note également  $\beta_{1:p}$  (resp.  $\hat{\beta}_{1:p}$ ) le vecteur  $\beta$  (resp.  $\hat{\beta}$ ) auquel on a retiré l'élément  $\beta_0$  (resp.  $\hat{\beta}_0$ ). Et on admet que

$$\frac{\|(I_n - P_0)X_{1:p}(\hat{\beta}_{1:p} - \beta_{1:p})\|^2}{\sigma^2} \sim \chi^2(p)$$

où  $P_0$  est la matrice de projection orthogonale sur le sous-espace engendré par le vecteur colonne  $X_0$ .

- (a) Montrer que  $\hat{Y} - \bar{Y}\mathbf{1}_n = \hat{Y} - P_0Y = (I_n - P_0)X_{1:p}\hat{\beta}_{1:p}$ .
- (b) En déduire la loi sous l'hypothèse  $\mathcal{H}_0$  de

$$\frac{\|\hat{Y} - \bar{Y}\mathbf{1}_n\|^2}{\sigma^2}$$

puis que  $T$  suit sous  $\mathcal{H}_0$  la loi de Fisher à  $p$  et  $n - p - 1$  degrés de libertés,

- (c) On admet que la région de rejet à un seuil  $\alpha$  est de la forme  $\{T > c_\alpha\}$ . En déduire la  $p$ -valeur du test.
7. On souhaite tester à présent pour chaque  $j = 1, \dots, p$

$$\mathcal{H}_0^{(j)} : \{\beta_j = 0\} \quad \text{contre} \quad \mathcal{H}_1^{(j)} : \{\beta_j \neq 0\} \quad (2)$$

Pour cela on va se baser sur la statistique de test

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\hat{v}_j}}$$

Quelle est sa loi sous  $\mathcal{H}_0^{(j)}$  vraie? En déduire la  $p$ -valeur associée à chacun des  $p$  tests.

## Exercice 2      Test entre modèles emboîtés

1. Rappeler (sans démonstration) la construction d'un test entre modèles emboîtés, en précisant bien les hypothèses  $\mathcal{H}_0$  et  $\mathcal{H}_1$ , la statistique de test, sa loi sous  $\mathcal{H}_0$  ainsi que la règle de décision.
2. Rappeler le lien entre ce test et les vraisemblances des 2 modèles.
3. Après avoir rappelé le lien entre taille du modèle et ajustement, expliquer en quoi se baser sur un tel test pour sélectionner des variables peut ne pas être adapté.