

Introduction à R Commander

Régression linéaire et Markdown

Philippe Grosjean & Guyliann Engels

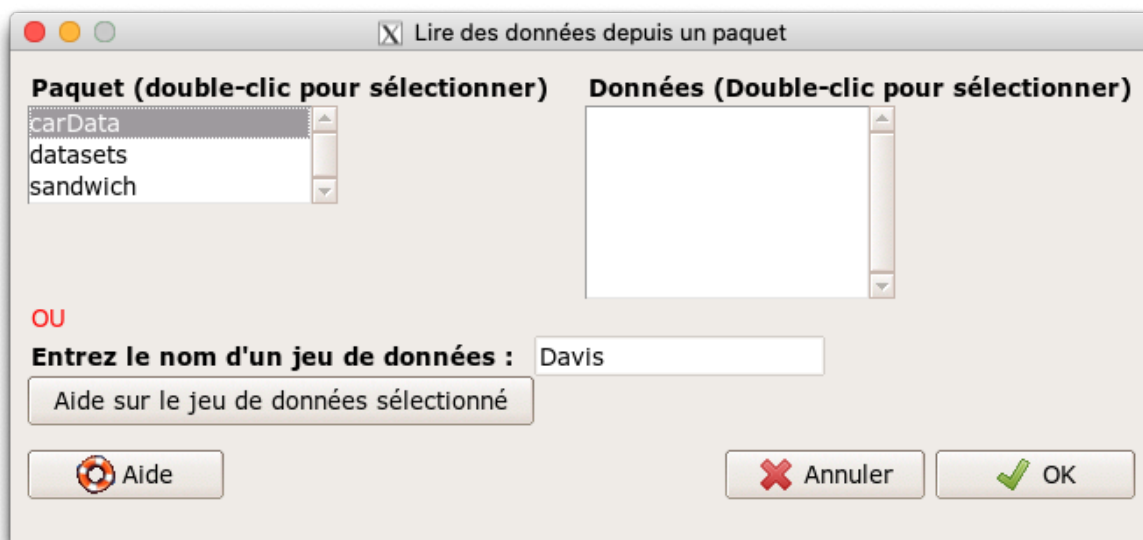
2021-10-28

Ce document est une contribution à STAT for U.

STAT **for** **U**

1 Se préparer...

Démarrez R Commander et chargez le jeu de données **Davis** du package `{carData}`. Vous devez donc faire **Données -> Données dans les paquets -> Lire des données depuis un paquet attaché...** Tant que vous êtes dans cette boîte de dialogue, affichez le fichier d'aide associé à ce jeu de données pour le découvrir. Ensuite, ouvrez le jeu de données (bouton **OK**).



Prenez l'habitude de **visualiser** votre tableau de données juste après son importation pour vérifier qu'il se présente comme prévu (bouton **Visualiser** dans la barre d'outils).

Pour les analyses statistiques plus complexes comme les modèles linéaires, linéaires généralisés, etc., R travaille en plusieurs étapes :

1. Création d'un objet qui contient toutes les informations nécessaires relatives au modèle,
2. Appel à des fonctions spécialisées (on les appelle des **méthodes** de l'objet) pour obtenir plus d'information sur l'analyse en cours,
3. Utilisations de graphiques également **spécialisés**

Donc, contrairement à d'autres logiciels de statistiques qui crachent plusieurs pages de résultats sur tout ce qui pourrait éventuellement vous être utile dans le contexte de votre analyse (oui, SPSS par exemple), vous devez apprendre à aller piocher les items qui vous intéressent par vous-mêmes dans R et R Commander.

Cette approche est voulue et a démontré son intérêt, car il faut être bien conscient des éléments nécessaires à l'analyse et au diagnostic du modèle... et donc, il faut comprendre ce que l'on fait pour s'en sortir avec R !

2 Description et visualisation des données

La description des données est toujours la première étape à réaliser. Nous ferons une description numérique et une visualisation graphique, toutes deux adaptées à l'étude qui nous intéresse, à savoir, **un modèle linéaire est-il pertinent pour représenter la masse (variable `weight`, en kg) par rapport à la taille (height en cm) chez l'adulte, ou en tous cas dans la population ciblée par l'étude, éventuellement en fonction du sexe (variable `sex`, avec les modalités F et M).**

2.1 Description numérique

- Le plus simple est d'utiliser **Statistiques -> Résumés -> Jeu de données actif**. Vous obtenez ceci :

Sortie



Soumettre

```
> data(Davis)
> summary(Davis)
sex      weight      height      repwt      repht
F:112   Min.    : 39.0   Min.    : 57.0   Min.    : 41.00   Min.    :148.0
M: 88   1st Qu.: 55.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:160.5
        Median : 63.0   Median :169.5   Median : 63.00   Median :168.0
        Mean   : 65.8   Mean   :170.0   Mean   : 65.62   Mean   :168.5
        3rd Qu.: 74.0   3rd Qu.:177.2   3rd Qu.: 73.50   3rd Qu.:175.0
        Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0
        NA's   :17     NA's   :17
```

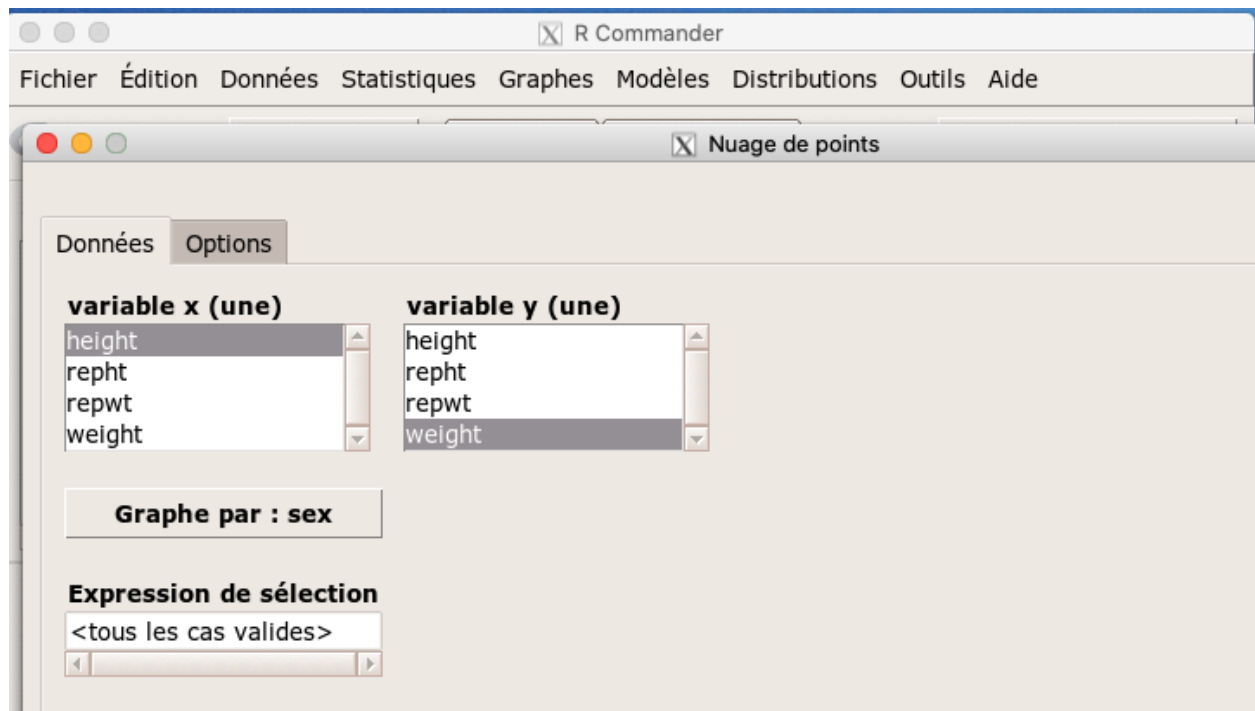
Nous pouvons voir, pour les variables quantitative, les cinq nombres (min, 1er quartile, médiane, 3ème quartile et max), ainsi que la moyenne et une indication éventuelle du nombre de valeurs manquantes (NA's. Pour les variables qualitatives, nous avons les proportions pour chaque modalité). Sachant que `repwt` et `repht` sont les masses et tailles indiquées par les sujets, ne voyez-vous pas déjà une anomalie ici ?

Si on s'intéresse à la relation linéaire entre deux variables, nous pouvons plutôt commencer par calculer le coefficient de corrélation linéaire de Pearson entre ces variables. Faites **Statistiques -> Résumés -> Matrice de corrélation...** Ensuite, vous sélectionnez les quatre variables (utilisation de **Shift** en sélectionnant). Prenez un moment pour analyser ce tableau.

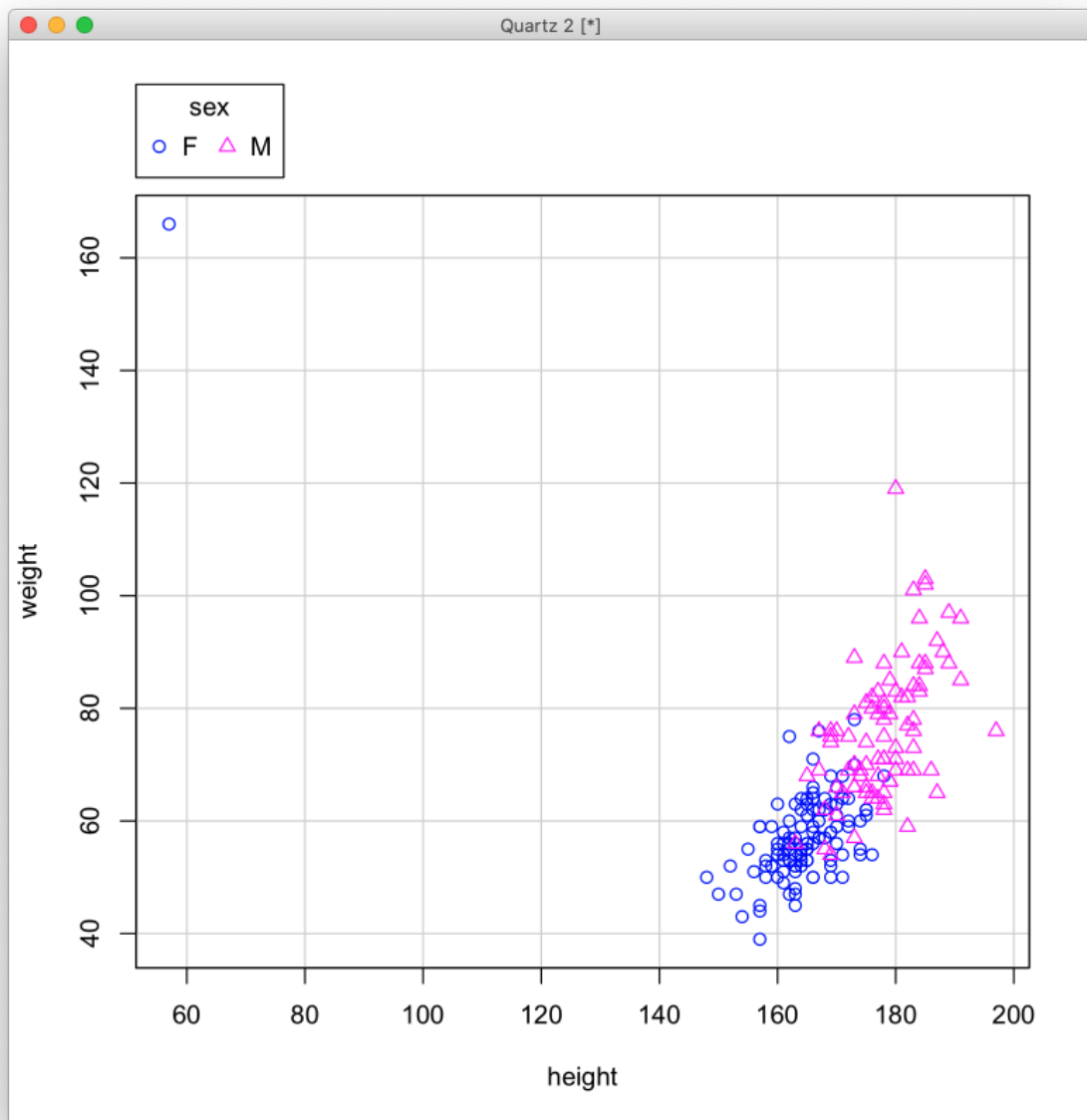
```
> library(e1071, pos=18)
> cor(Davis[,c("height", "repht", "repwt", "weight")], use="complete")
      height  repht  repwt  weight
height 1.0000000 0.7391662 0.6037367 0.1542575
repht  0.7391662 1.0000000 0.7618604 0.6314352
repwt  0.6037367 0.7618604 1.0000000 0.8353758
weight 0.1542575 0.6314352 0.8353758 1.0000000
```

2.2 Visualisation des données

Plusieurs graphiques différents sont utilisables ici, mais le plus pertinent est le nuage de points. Allez dans le menu **Graphes -> Nuage de points**. Sélectionnez `height` comme variable x, et `weight` comme variable y, cliquez sur **Graphique par groupe...** et sélectionnez `sex`, puis cliquez sur le bouton **OK**.

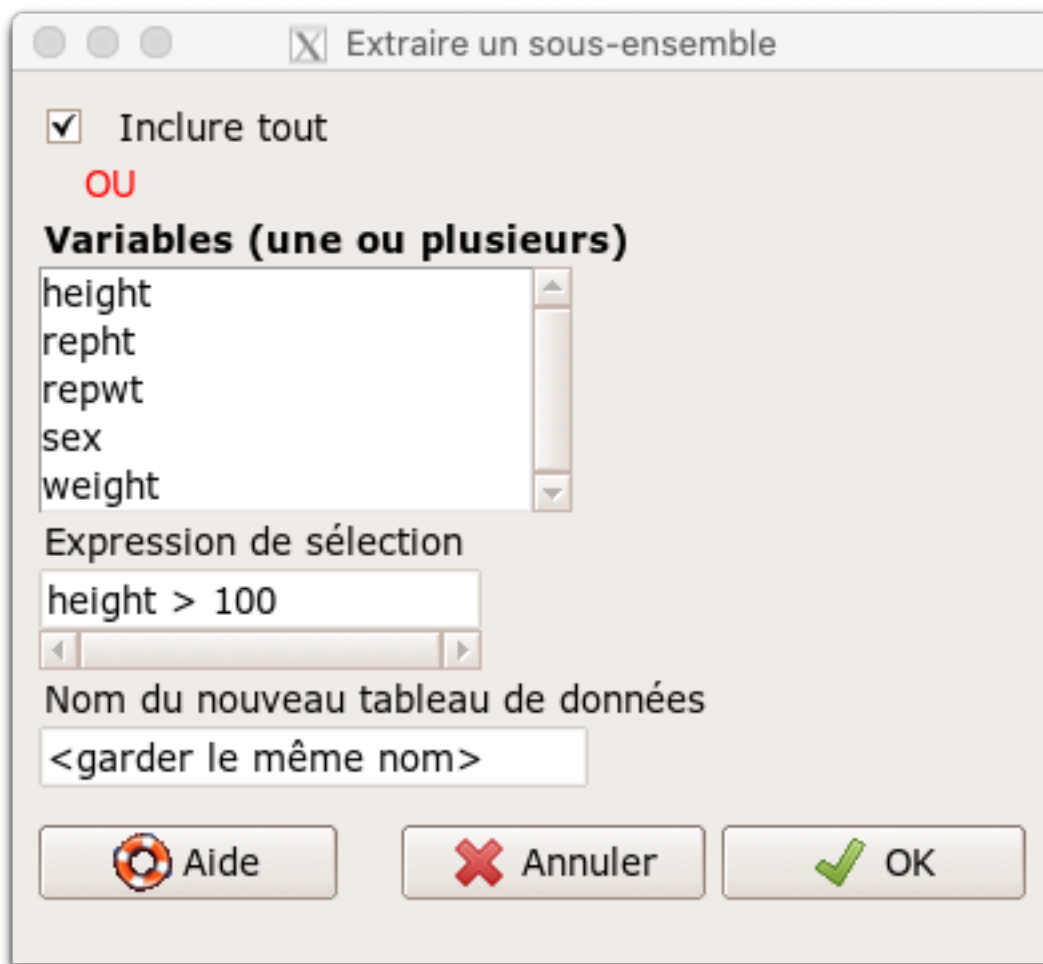


Vous devez obtenir le graphique suivant :

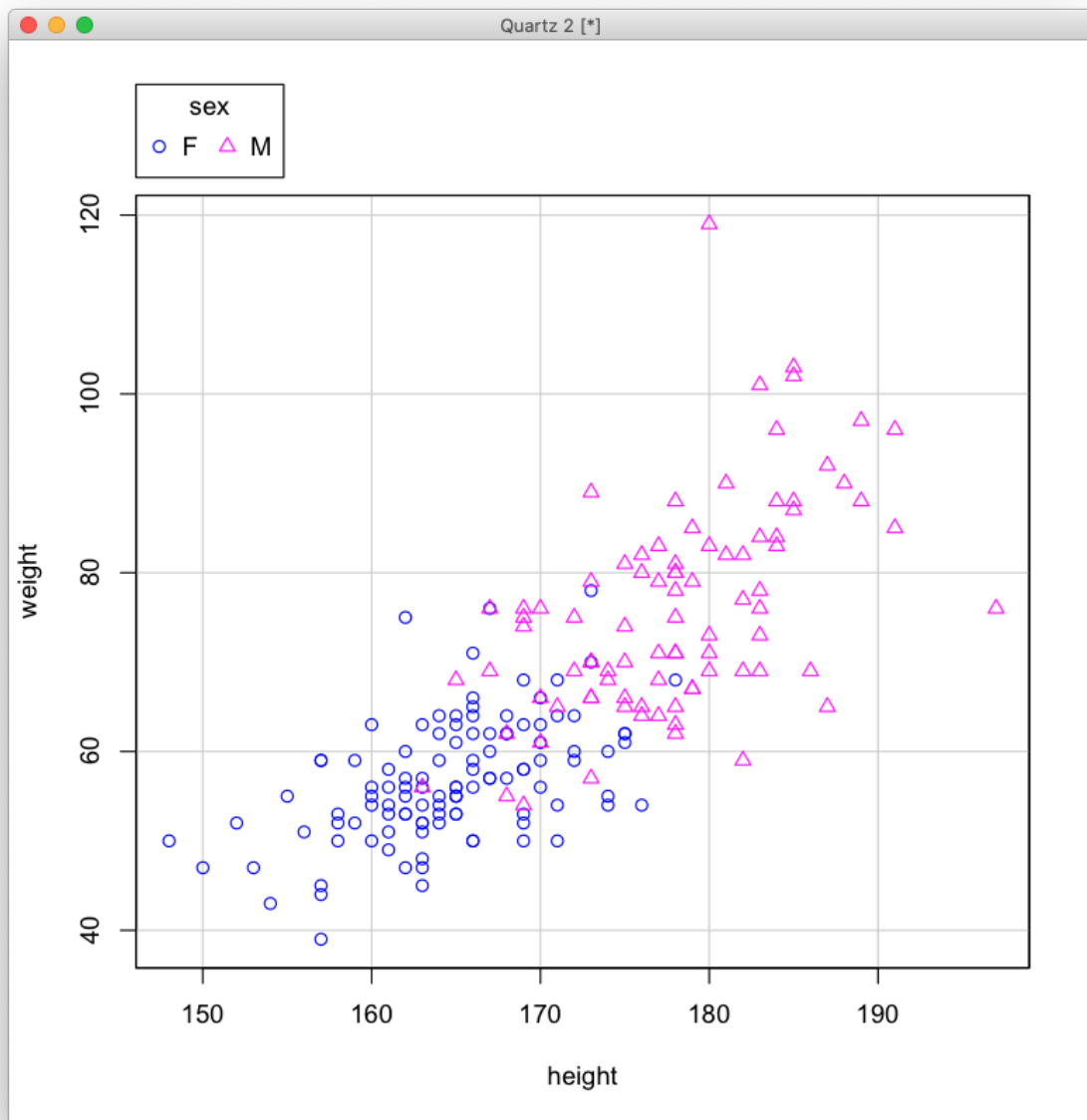


Vous notez immédiatement la présence d'un point anormal ! Il semblez que la masse et la teille aient été inversés pour une femme car 165kg pour une taille de 58cm, ce n'est pas vraiment possible. Après vérification, nous décidons donc d'éliminer ce point. Tentez de le faire par vous-mêmes avant de lire plus loin.

Nous allons dans le menu **Données** -> **Jeu de données actif** -> **Sous-ensemble...** et nous rentrons par exemple **height > 100** dans la case **Expression de sélection** pour éliminer cet individu adulte soi-disant de 58cm de haut. Cliquez sur **OK** et confirmez le remplacement du tableau. **Nous ne devons pas oublier dans notre publication d'expliquer cette étape.**



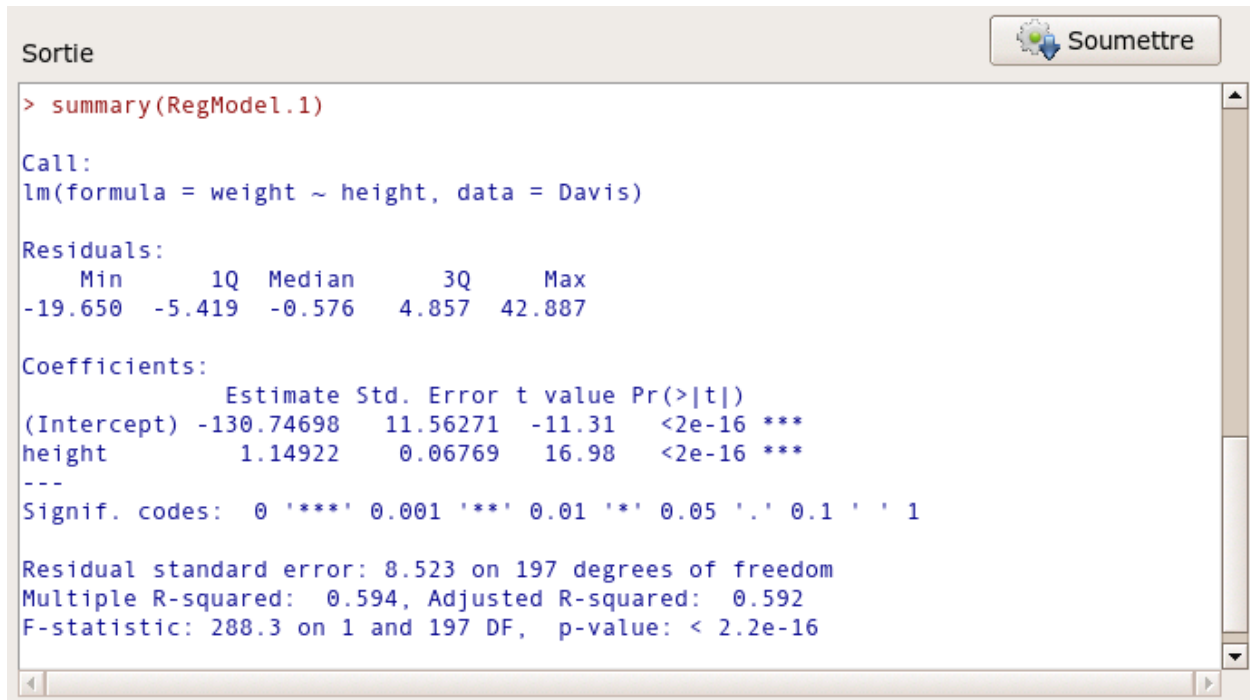
Ensuite, nous refaisons une description numérique et graphique du jeu de données nettoyé. Notez comme le coefficient de corrélation de Pearson est passé d'un très maigre 0.154 à un bien meilleur 0.768 après cette opération. Voici le nouveau graphique avec les mêmes options que précédemment :



Ce que nous observons nous encourage à tenter une régression linéaire dans ces données. soit un modèle unique pour les hommes et les femmes, soit un modèle plus complexe qui tiendrait compte des différences en fonction du sexe.

3 Régression linéaire

La première étape consiste à créer le modèle dans R (R commander visualise automatiquement un résumé numérique du modèle par la suite). Allez dans **Statistiques -> Ajustement de modèles -> Régression linéaire...** Sélectionnez la variable **weight** comme réponse et **height** comme explicative et puis **OK**. Vous voyez un résumé du modèle (son contenu est riche, mais nous ne pouvons le détailler ici. Je vous renvoie au cours de Science des Données Biologiques, ici: <https://wp.sciviews.org/sdd-umons2/?iframe=wp.sciviews.org/sdd-umons2-2021/r%25C3%25A9gression-lin%25C3%25A9aire-simple.html%23r%25C3%25A9sum%25C3%25A9-avec-summary> et ici: [https://wp.sciviews.org/sdd-umons2/?iframe=wp.sciviews.org/sdd-umons2-](https://wp.sciviews.org/sdd-umons2/?iframe=wp.sciviews.org/sdd-umons2-2021/r%25C3%25A9gression-lin%25C3%25A9aire-simple.html%23r%25C3%25A9sum%25C3%25A9-avec-summary)



The screenshot shows the 'Sortie' (Output) window of R Commander. At the top right is a button labeled 'Soumettre' (Submit) with a gear icon. The main area contains the output of the command `> summary(RegModel.1)`. The output includes the call to `lm()`, the residuals summary (Min, 1Q, Median, 3Q, Max), the coefficients table with estimates, standard errors, t-values, and p-values, and summary statistics like R-squared and F-statistic.

```
> summary(RegModel.1)

Call:
lm(formula = weight ~ height, data = Davis)

Residuals:
    Min       1Q   Median       3Q      Max
-19.650  -5.419  -0.576   4.857  42.887

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -130.74698   11.56271  -11.31  <2e-16 ***
height       1.14922    0.06769   16.98  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.523 on 197 degrees of freedom
Multiple R-squared:  0.594, Adjusted R-squared:  0.592
F-statistic: 288.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

Même avec un R^2 assez faible de 0.59, notre modèle est significatif (ANOVA résumé à la dernière ligne et test de Student sur les deux paramètres). Notre modèle est : $\text{weight (kg)} = 1.15 * \text{height (cm)} - 131$.

Pour avoir d'autres informations sur notre modèle, rappelons-nous qu'il faut les demander explicitement dans R. Tout cela se trouve dans le menu **Modèle** de R Commander. Par exemple, le tableau de l'ANOVA complète liée au modèle qui nous dit s'il est significatif ou pas peut-être obtenue à l'aide de **Modèles -> Tests d'hypothèses -> Table d'ANOVA...** Gardez les options proposées et cliquez sur le bouton **OK** dans la boîte de dialogue qui s'ouvre. Vous obtenez ceci :


```
Sortie Soumettre

(Intercept) -130.74698  11.56271 -11.31 <2e-16 ***
height      1.14922    0.06769  16.98 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.523 on 197 degrees of freedom
Multiple R-squared:  0.594, Adjusted R-squared:  0.592
F-statistic: 288.3 on 1 and 197 DF, p-value: < 2.2e-16

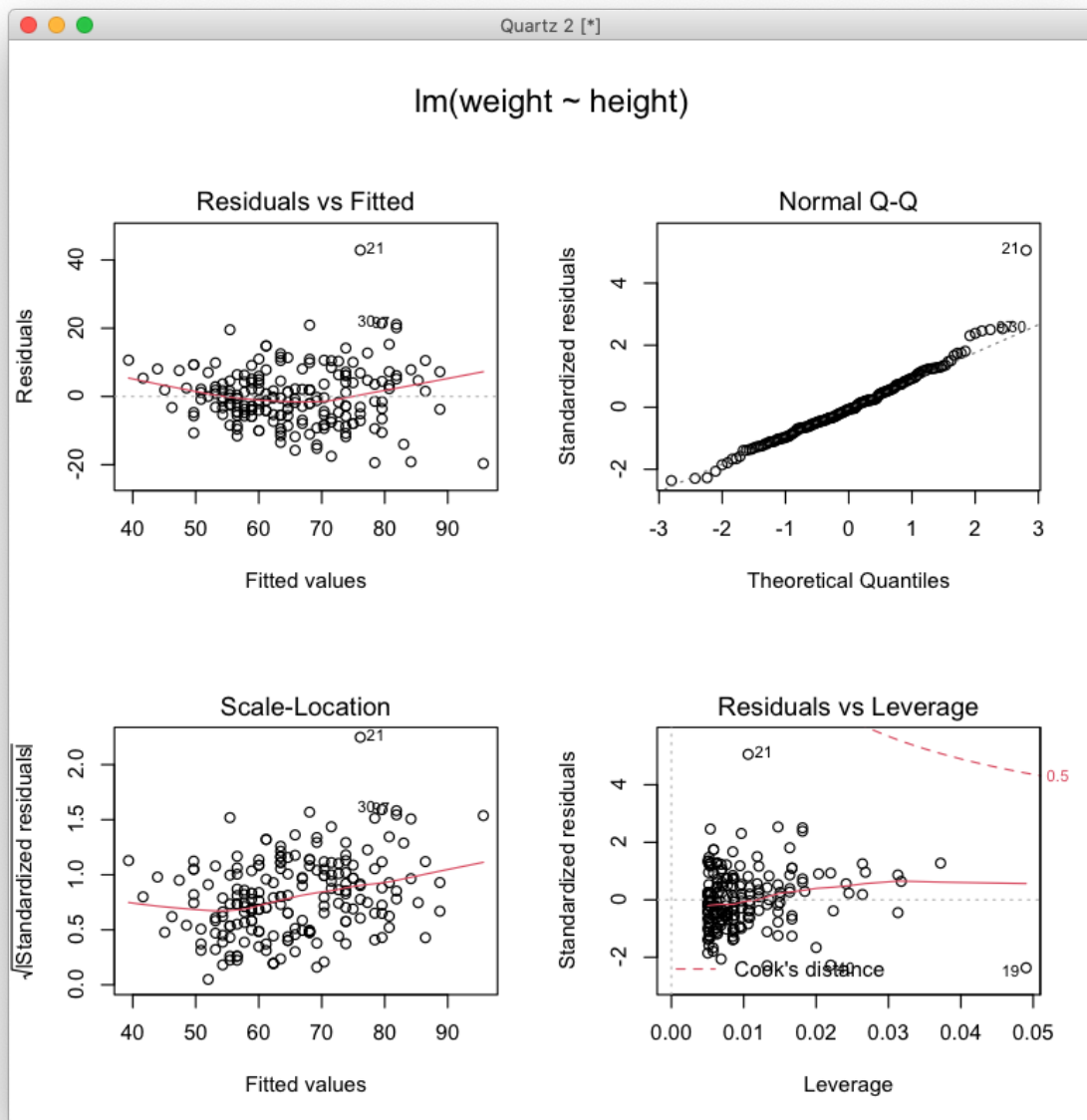
> Anova(RegModel1, type="II")
Anova Table (Type II tests)

Response: weight
      Sum Sq Df F value    Pr(>F)
height  20942  1  288.25 < 2.2e-16 ***
Residuals 14312 197
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vous voyez bien que la dernière ligne du résumé de notre modèle correspond à l'ANOVA du modèle. Nous pouvons calculer bien d'autres choses (explorez le menu **Modèles**). Par exemple, nous pouvons calculer le critère d'Akaike pour ce modèle (un critère qui permet de comparer différents modèles ajustés dans le même jeu de données et de choisir celui qui présente la valeur d'AIC la plus faible, éventuellement). Faites **Modèles** -> **Critère d'information d'Akaike (AIC)**. Vous devriez obtenir la valeur de 1421.

3.1 Analyse des résidus

Les résidus du modèle (la partie non expliquée) doivent avoir une distribution Normale et une variance homogène le long de la droite. Ce sont deux parmi les conditions d'application de la régression linéaire par les moindres carrés. Il faut aussi que le nuage de points soit linéaire, qu'il n'y ait pas de valeurs extrêmes suspectes, pas de valeurs trop influentes, etc. R et R Commander vous propose une série de graphiques pour diagnostiquer tout cela dans une troisième étape. Il s'agit des graphiques d'**analyse des résidus**. Allez dans le menu **Modèles** -> **Graphes** -> **Diagnostics graphiques**. Vous obtenez ceci :



Ces quatre graphiques vous permettent déjà de détecter pas mal de problèmes potentiels. Leur description va au delà de cette petite démonstration, mais vous pouvez avoir plus d'informations et un exemple d'utilisation dans le cours de science des données biologiques ici : <https://wp.sciviews.org/sdd-umons2/?iframe=wp.sciviews.org/sdd-umons2-2021/outils-de-diagnostic.html>.

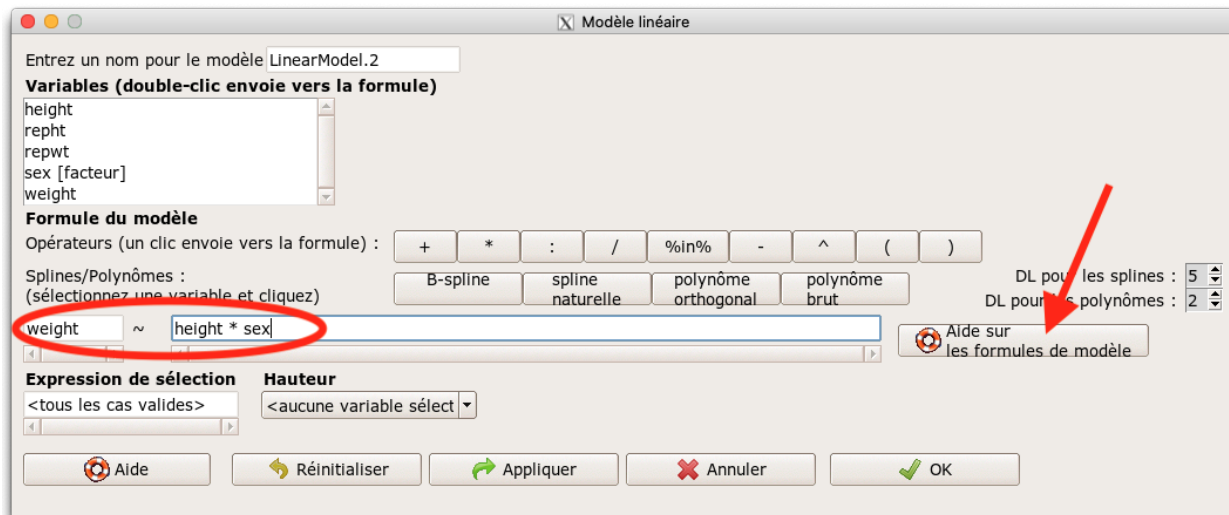
En gros, dans le cas présent, nous ne détectons pas d'anomalies particulière, si ce n'est un point extrême (la ligne 21 du tableau) mais non suspect. De plus, la linearité du nuage de points n'est pas parfaite, mais pas trop gênante.

3.2 Modèle linéaire plus complexe

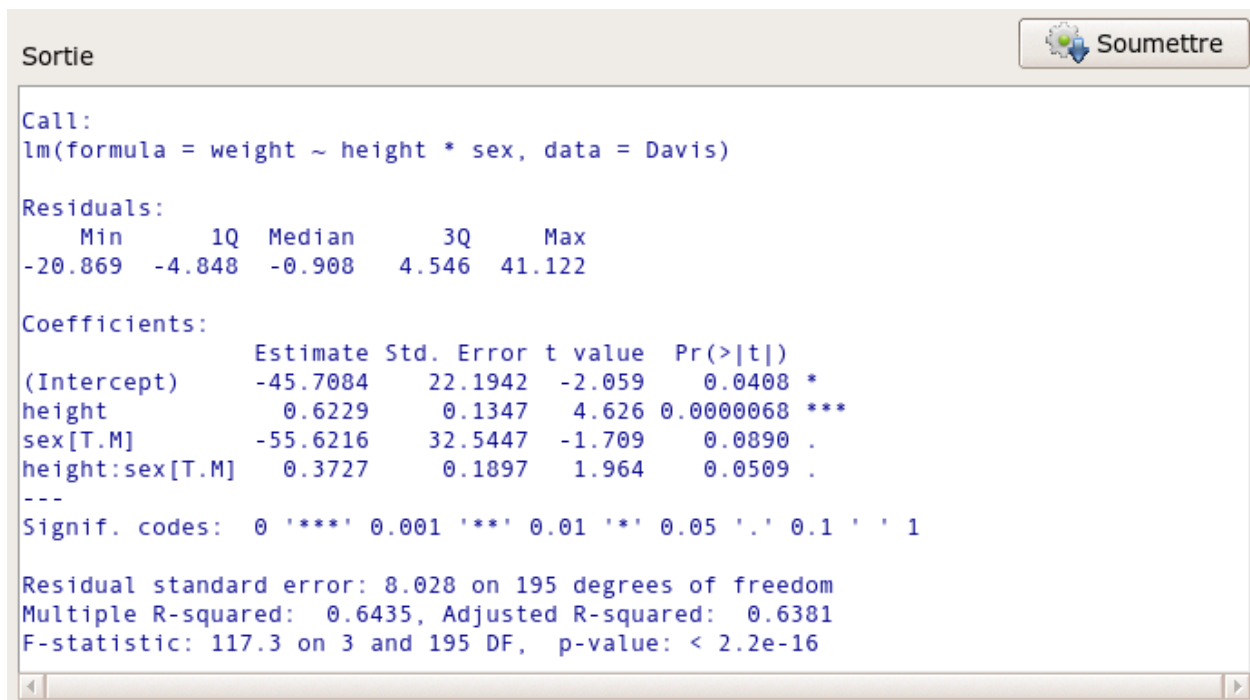
Nous allons terminer cette exploration rapide des fonctionnalités de modélisation de R Commander en réalisant un **modèle linéaire** plus complexe qui tiendra compte de différences entre hommes et femmes. Pour cela, nous allons dans le menu **Statistiques -> Ajustement de modèles -> Modèle linéaire....**

Ici, l'interface est beaucoup plus compliquée, à l'image des possibilités immenses de ce type de modèle.

Nous devons spécifier ici une **formule** qui va représenter la relation entre les variables dans notre modèle. La régression linéaire simple réalisée plus haut correspond au modèle `weight ~ height` qui se lit “weight en fonction de height”. Cette formule est déjà rentrée. Nous allons ajouter une seconde variable au modèle : `sex` en écrivant `weight ~ height * sex` (vous avez un bouton qui vous renvoie à la page d'aide qui explique comment élaborer une formule pour un modèle linéaire dans la boîte de dialogue) :



Cliquez **OK**. Votre modèle est calculé et son résumé est affiché.



```
Call:
lm(formula = weight ~ height * sex, data = Davis)

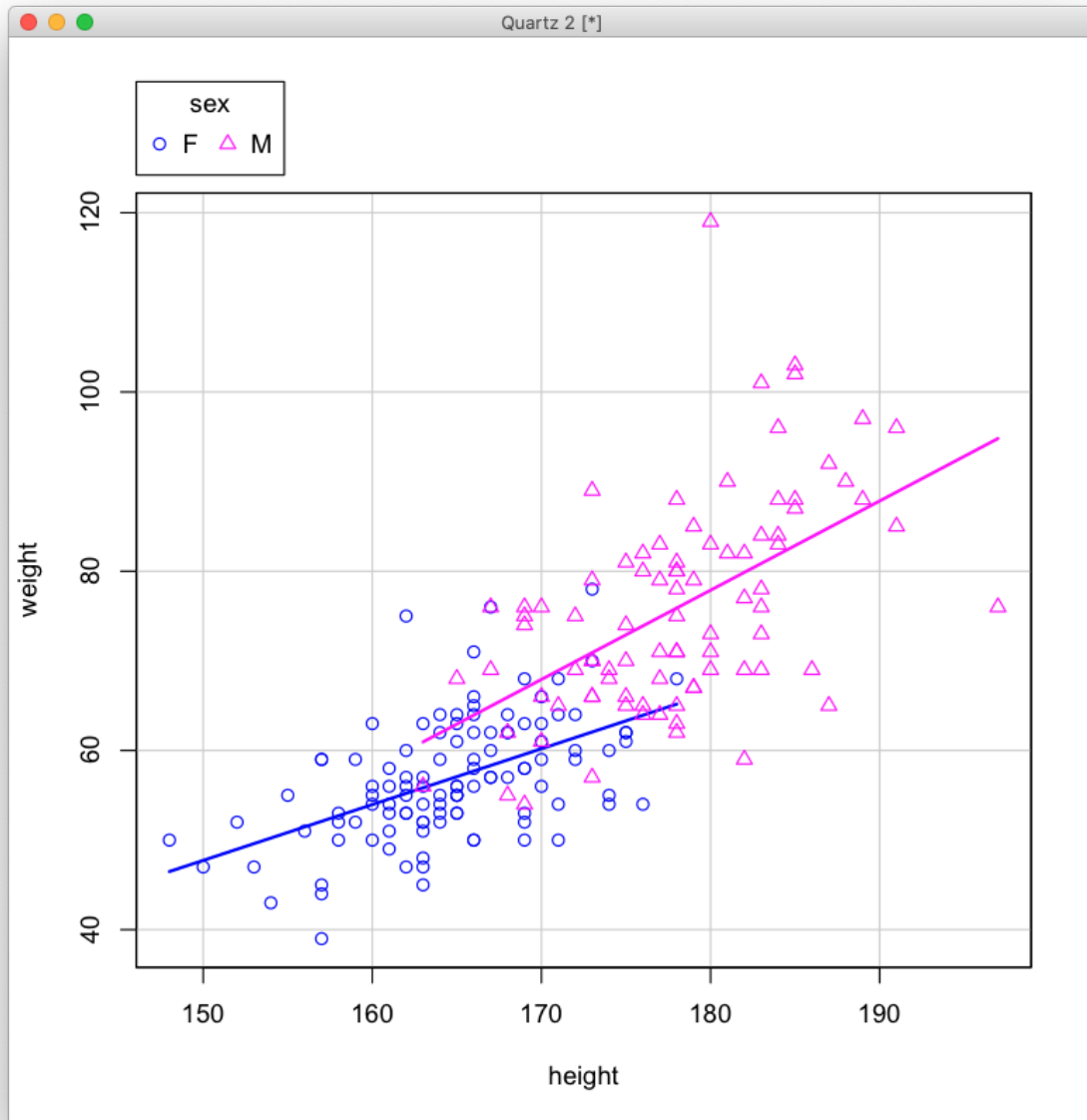
Residuals:
    Min       1Q   Median       3Q      Max
-20.869  -4.848  -0.908   4.546  41.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -45.7084    22.1942  -2.059   0.0408 *
height         0.6229     0.1347   4.626 0.0000068 ***
sex[T.M]     -55.6216    32.5447  -1.709   0.0890 .
height:sex[T.M]  0.3727     0.1897   1.964   0.0509 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.028 on 195 degrees of freedom
Multiple R-squared:  0.6435, Adjusted R-squared:  0.6381
F-statistic: 117.3 on 3 and 195 DF, p-value: < 2.2e-16
```

Ici aussi, vous êtes supposé utiliser les différents outils de diagnostic pour étudier votre modèle et vérifier s'il répond bien aux hypothèses de départ de normalité et d'homogénéité de la variance des résidus, etc. Les mêmes outils que précédemment sont aussi utilisables ici. Par exemple, calculons le critère d'Akaike, ce qui

donne 1400, soit une valeur légèrement plus basse que pour le modèle plus simple qui ne prend pas en compte les différences homme - femme. Il faudrait l'analyser plus en détails, mais il semble que ce nouveau modèle est pertinent. Pour le visualiser, vous pouvez faire **Graphes -> Nuage de points...**, sélectionner **height** comme x, **weight** comme y, et **sex** pour le groupe. Ensuite dans l'onglet **Options**, vous cochez **ligne des moindres carrés**. Vous obtenez ceci :



Nous n'avons fait que survoler les possibilités en matière de modélisation du logiciel. De plus, des addins existent pour aller encore plus loin. L'important est ici de réaliser les deux points suivants :

1. R et R Commander ne crachent pas tous les résultats possibles et imaginables en relation avec un modèle. Vous devez aller les chercher *sélectivement* par vous-mêmes (dans le menu **Modèles** dans le cas de R Commander).
2. Pour les modèles plus complexes, R et R Commander utilisent une interface particulière dite "formule" qui permet de spécifier très finement la relation entre les différentes variables dans le modèle.

4 R Markdown

Jusqu'ici, nous n'avons utilisé que le **Script R** qui reprend les différentes commandes à exécuter. Vous pouvez le sauvegarder et le rejouer quand vous voulez à l'aide du bouton **Soumettre**. C'est pratique car cela permet de tracer les analyses réalisées, voire de les reprendre dans un autre contexte si vous avez plusieurs jeux de données similaires à analyser.

L'onglet **R Markdown** donne accès à une autre fonctionnalité qui permet de créer des rapports formatés en HTML, PDF, Word, Powerpoint, etc. Le principe est le suivant. Vous avez trois zones distinctes dans le document :

1. Un entête dit "YAML" qui spécifie quoi faire avec le document (par exemple le compiler en un fichier PDF) et indique des données générales comme le titre, la date, l'auteur,
2. Des zones de texte libre qui peut être formaté à l'aide de balises **Markdown** (voir ici : <https://rmarkdown.rstudio.com/> et ici : <https://rmarkdown.rstudio.com/lesson-8.html>)
3. Des **chunks** qui contiennent du code R. Lors de la compilation du document, ces chunks sont exécutés et remplacés par le résultat, par exemple, un graphique ou un tableau.

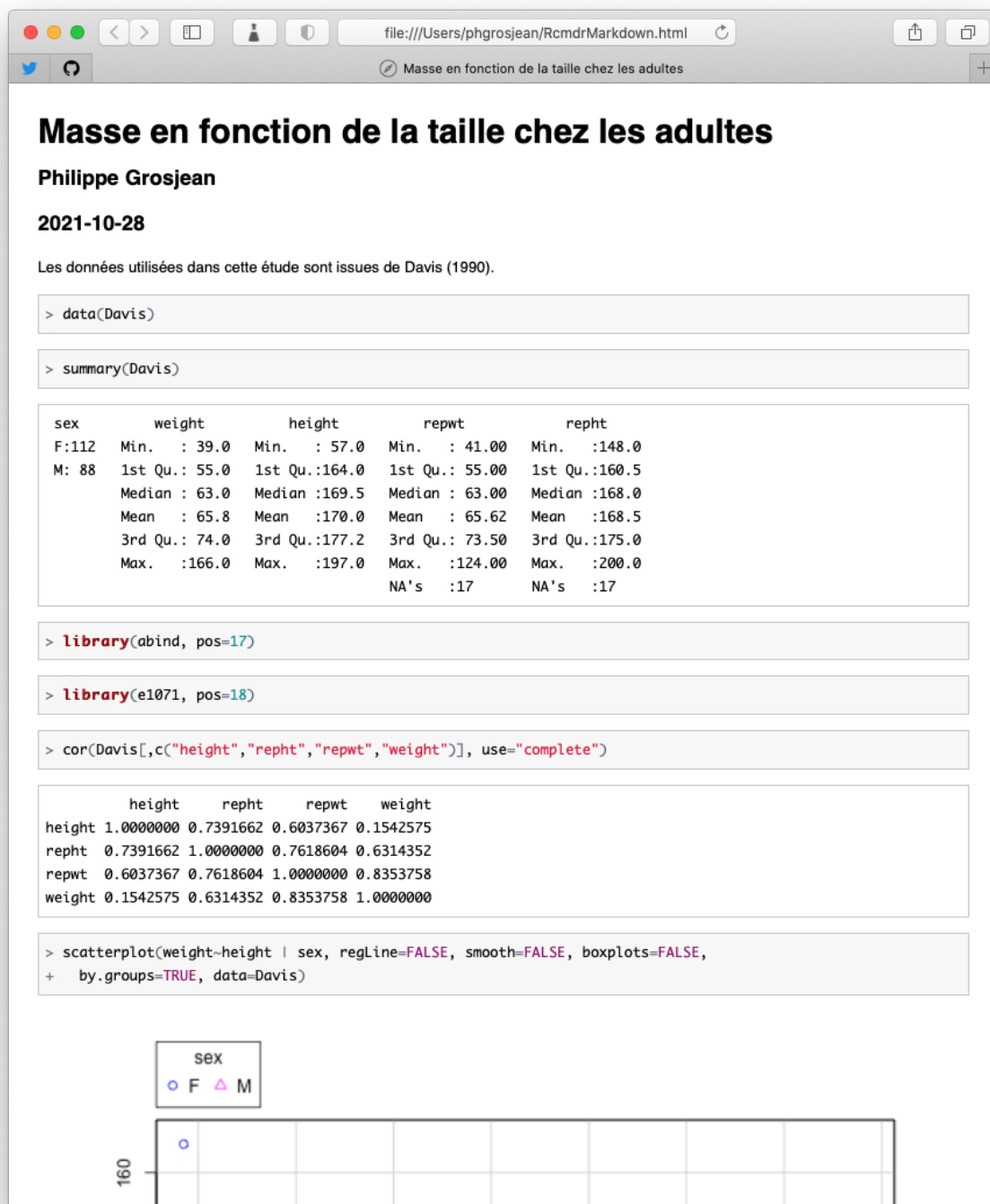
Ainsi, les documents R Markdown permettent d'effectuer les calculs statistiques et les graphiques *directement dans le même document que celui qui génère le rapport*. Visualisez la petite vidéo sur cette page pour une autre explication : <https://rmarkdown.rstudio.com/lesson-1.html>.

4.1 Rapport de nos analyses

R Commander propose une version très light de R Markdown dans l'onglet correspondant. Par exemple, pour générer un rapport des analyses que nous venons de faire, allez dans l'onglet **R Markdown** et cliquez sur le bouton **Générer un rapport**. Vous verrez une page HTML apparaître avec le résultat de vos analyses.

Vous pouvez naturellement éditer la partie texte (et même les "chunks") de ce rapport. Changez le titre en "Masse en fonction de la taille chez les adultes". Indiquez votre nom à la place de "Your Name" et écrivez une petite phrase d'introduction juste après "### 2021-10-28", par exemple, "Les données utilisées dans cette étude sont issues de Davis (1990).".

Régénérez votre rapport, vous devez obtenir quelque chose comme ceci :



Markdown propose de nombreux formatages, et vous pouvez même gérer et générer la bibliographie de votre rapport au format de différentes revues. Pour sauvegarder votre rapport, vous faites **Fichier -> Enregistrer un fichier R Markdown...**, donc de manière similaire à l'enregistrement d'un script R (mais avec extension **.Rmd** au lieu de **.R** pour le script). Tout comme votre script, vous pouvez recharger un document R Markdown et le "rejouer" pour recréer votre rapport.

4.2 R Markdown dans R Studio (d  mo)

R Commander est un bon logiciel pour un utilisateur d  butant ou occasionnel, mais si vous vous lancez dans des fonctions plus avanc  es (telles que justement les documents R Markdown), vous allez rapidement   tre emb  t  s par les limitations du logiciel. Dans ce cas, tout en conservant les bonnes habitudes de R qui est sous-jacent aux deux logiciels, vous pouvez passer sur **RStudio** (voir <https://www.rstudio.com/>). L  , les possibilit  s de manipulation, de formatage et de sortie    partir de R Markdown sont d  cupl  es.

D  monstration de l'ouverture du fichier R Markdown sauvegard      partir de R Commander et g  n  ration de diff  rents formats de sortie.

5 Exercices suppl  mentaires

Pour les mod  les :

- Effectuez l'analyse des r  sidus pour votre mod  le complet qui tient compte du sexe.
- Explorez d'autres types de mod  les    partir de **Statistiques -> Ajustement de mod  les**.
- Explorer les formatages Markdown dans une zone de texte    l'int  rieur de votre rapport, par exemple, pour mettre en italique, vous encadrez le passage d'une ast  risque avant et apr  s. Pour le mettre en gras, vous utilisez deux ast  risques avant et apr  s. Pour un titre, vous le pr  c  dez de un    six di  ses ("hashtags" #) suivi d'une espace et placez ce titre sur sa propre ligne encadr  e de lignes vides. Pour plus d'id  es de formatage, voyez l'aide-m  moire de Markdown ici : <https://github.com/rstudio/cheatsheets/raw/master/rmarkdown.pdf>.