



1 Objectif

« [swirl](#) » : [package R pour l'autoformation à R \(plus généralement à la statistique\)](#).

L'autoformation est un des leviers essentiels de notre progression. Je le dis souvent à mes étudiants, le ratio entre ce que j'ai appris par moi-même et ce qu'on m'a appris s'accroît chaque jour. Il ne peut pas en être autrement, nous ne pouvons pas rester indéfiniment sur les bancs de l'école. Certes, les formations sont indispensables, je ne dis pas le contraire, elles nous inculquent les fondamentaux, les schémas de compréhension, les bonnes pratiques, les compétences essentielles. Mais par la suite, il nous appartient de faire fructifier par nous-même notre capital de connaissances et de savoir-faire.

Surtout que nous sommes dans une période faste. Naguère, nous devions passer des heures à la bibliothèque pour se documenter. Aujourd'hui, nous accédons quasiment à tout en restant chez soi, assis devant son ordinateur. La gageure n'est plus la rareté de l'information mais sa surabondance. On peut s'y perdre facilement. Il faut s'appuyer sur les bonnes références, sur les bons outils.

J'essaie de contribuer moi-même à mon échelle à travers les supports de cours et les tutoriels. Personnellement, je crois à la force de l'écrit. Mais je conçois que la vidéo puisse être un vecteur intéressant également. Je passe beaucoup de temps à en visionner. Dans les deux cas, l'interaction est implicite. L'apprenant doit faire l'effort de prendre du recul à bon escient pour essayer par lui-même de reproduire la démarche. Une voie alternative est le tutoriel où on le guide explicitement avec des systèmes de questions-réponses et des exercices contrôlés.

Le package [swirl](#) fait partie de cette seconde catégorie. Une ancienne du Master SISE qui assure des formations sous R me l'a indiqué. Je la remercie pour cela. Malgré le temps que je passe sur le web à scruter ce qui se fait, beaucoup de choses m'échappent. Elle m'indiquait qu'elle l'utilisait avec succès pour ses séances de cours. Bien évidemment que je me suis rué dessus pour voir un peu ce qu'il en était, et le parti que je pourrais en tirer pour mes propres enseignements.

Dans ce tutoriel, nous essaierons de faire le tour de l'outil pédagogique en montrant son mode de fonctionnement sur un exemple de séance consacrée à la régression linéaire multiple. Nous en profiterons pour regarder un peu sous le capot et voir les éléments constitutifs des cours.



2 Installation, chargement et lancement

Même si nous fonctionnons en mode console, il est plus avantageux de travailler sous l'EDI [RStudio](#) (EDI : environnement de développement intégré) car le package charge parfois du code R à étudier ou à exécuter dans l'éditeur. La visualisation est plus facile.

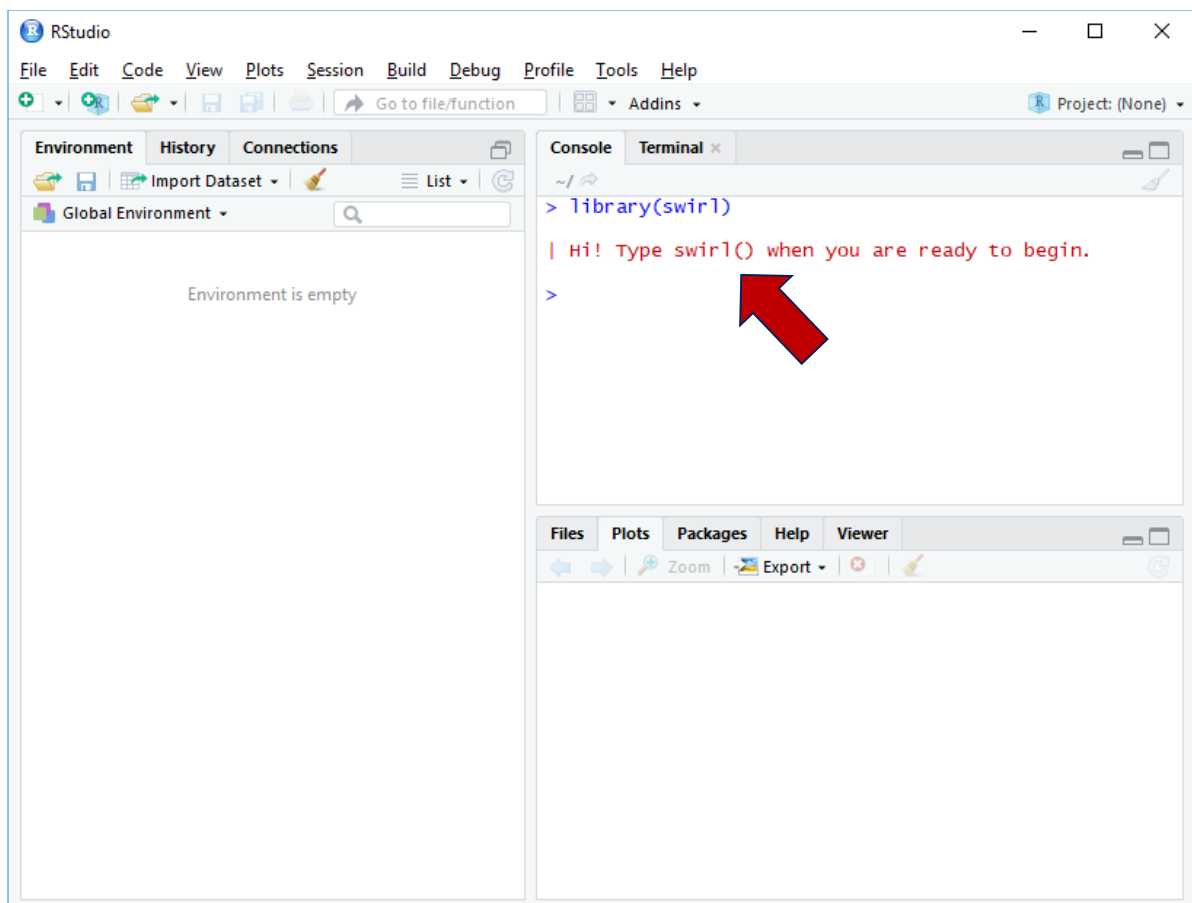
Nous lançons RStudio donc. Nous installons le package (à ne faire qu'une seule fois).

```
#installation du package
install.package('swirl')
```

Nous le chargeons ensuite avec la commande `library()`

```
#chargement du package en mémoire
library(swirl)
```

L'outil nous invite à entrer la commande **swirl()** pour démarrer.



Ce que nous faisons. L'outil fournit plusieurs indications, en particulier les commandes clés qui nous permettront d'interagir avec l'application. Puis il nous demande le cours que nous souhaitons suivre.



```
> swirl()

| Welcome to swirl! Please sign in. If you've
| been here before, use the same name as you did
| then. If you are new, call yourself something
| unique.

What shall I call you? Ricco

| Thanks, Ricco. Let's cover a couple of quick
| housekeeping items before we begin our first
| lesson. First of all, you should know that when
| you see '...', that means you should press
| Enter when you are done reading and ready to
| continue.

... <-- That's your cue to press Enter to continue

| Also, when you see 'ANSWER:', the R prompt (>),
| or when you are asked to select from a list,
| that means it's your turn to enter a response,
| then press Enter to continue.

Select 1, 2, or 3 and press Enter

1: Continue.
2: Proceed.
3: Let's get going!

Selection: 1

| You can exit swirl and return to the R prompt
| (>) at any time by pressing the Esc key. If you
| are already at the prompt, type bye() to exit
| and save your progress. When you exit properly,
| you'll see a short message letting you know
| you've done so.

| When you are at the R prompt (>):
| -- Typing skip() allows you to skip the current
| question.
| -- Typing play() lets you experiment with R on
| your own; swirl will ignore what you do...
| -- UNTIL you type nxt() which will regain
| swirl's attention.
| -- Typing bye() causes swirl to exit. Your
| progress will be saved.
| -- Typing main() returns you to swirl's main
| menu.
| -- Typing info() displays these options again.

| Let's get started!

...

| To begin, you must install a course. I can
```



```
| install a course for you from the internet, or  
| I can send you to a web page  
| (https://github.com/swirldev/swirl_courses)  
| which will provide course options and  
| directions for installing courses yourself. (If  
| you are not connected to the internet, type 0  
| to exit.)  
  
1: R Programming: The basics of programming in R  
2: Regression Models: The basics of regression  
modeling in R  
3: Statistical Inference: The basics of statistical  
inference in R  
4: Exploratory Data Analysis: The basics of exploring  
data in R  
5: Don't install anything for me. I'll do it myself.  
  
Selection: 2
```

Nous optons pour la Régression (2). Son contenu est automatiquement chargé à partir du dépôt.

Nous sommes prêts pour suivre un premier enseignement.

3 S'initier à la régression linéaire avec « swirl »

3.1 Suivre un chapitre de cours

Via le menu de démarrage qui apparaît, nous demandons à suivre le cours de Régression.

```
|=====| 100%  
  
| Course installed successfully!  
  
| Please choose a course, or type 0 to exit swirl.  
  
1: Regression Models  
2: Take me to the swirl course repository!  
  
Selection: 1
```

Un autre menu nous demande le chapitre que nous souhaitons explorer.

```
| Please choose a lesson, or type 0 to return to course  
| menu.  
  
1: Introduction  
2: Residuals  
3: Least Squares Estimation  
4: Residual Variation  
5: Introduction to Multivariable Regression
```



- 6: MultiVar Examples
- 7: MultiVar Examples2
- 8: MultiVar Examples3
- 9: Residuals Diagnostics and Variation
- 10: Variance Inflation Factors
- 11: Overfitting and Underfitting
- 12: Binary Outcomes
- 13: Count Outcomes

Selection: 6

Nous sommes un peu spectateurs dans certains cours, surtout dans les parties introductives où sont exposés les concepts fondamentaux. Ça n'a pas trop d'intérêt dans le cadre de ce tutoriel. Nous choisissons plutôt le **chapitre 6** où nous devons traiter un exemple avec un certain niveau d'interactivité.

The screenshot displays the RStudio interface. The editor window shows R code for a function named `makeIms` and a simulation function `rgp1`. The console window shows the execution of `makeIms()` and the loading of lesson dependencies, including packages `datasets`, `stats`, and `graphics`. The environment window shows the `swiss` dataset with 47 observations and 6 variables, and the `makeIms`, `rgp1`, and `rgp2` functions. The viewer window displays a data visualization titled "Swiss data" consisting of a 5x5 grid of scatter plots showing relationships between variables: Fertility, Agriculture, Examination, Education, Catholic, and Infant.Mortality.



La base de données « **swiss** » avec 47 observations et 6 variables est directement chargée en mémoire. Du code source R apparaît dans l'éditeur. Un graphique croisant les variables deux à deux (nuages de points) est affiché. L'intérêt de travailler sous RStudio apparaît pleinement ici.

3.2 Accès à la documentation

Autre aspect important, un texte nous invitant à accéder à la documentation en ligne est affiché dans la console.

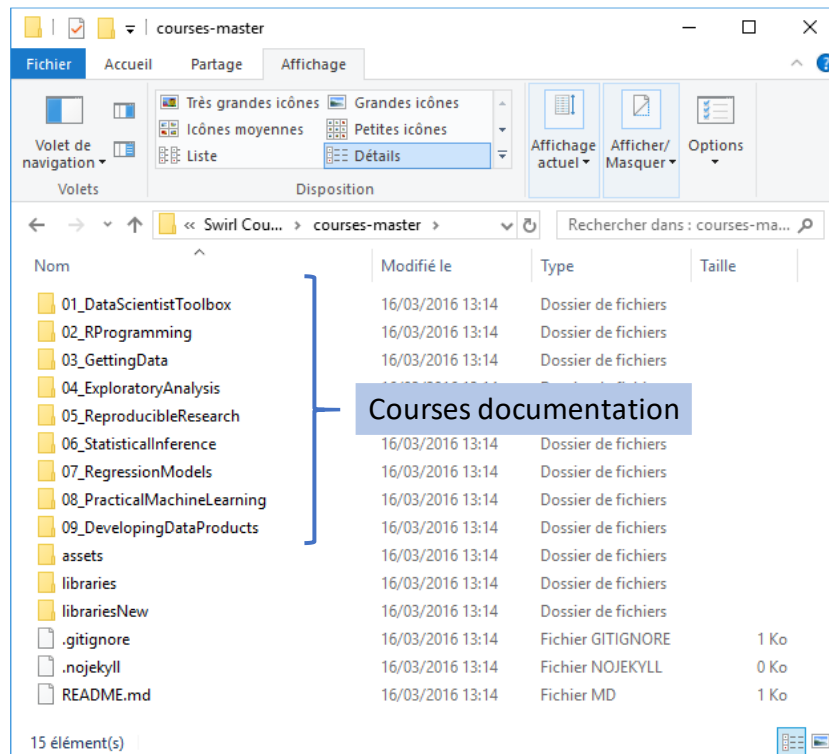
```
| MultiVar_Examples. (Slides for this and other Data  
| Science courses may be found at github  
| https://github.com/DataScienceSpecialization/courses.  
| If you care to use them, they must be downloaded as a  
| zip file and viewed locally. This lesson corresponds to  
| Regression_Models/02_02_multivariateExamples.)
```

Nous récupérons ladite documentation sur le dépôt GitHub.

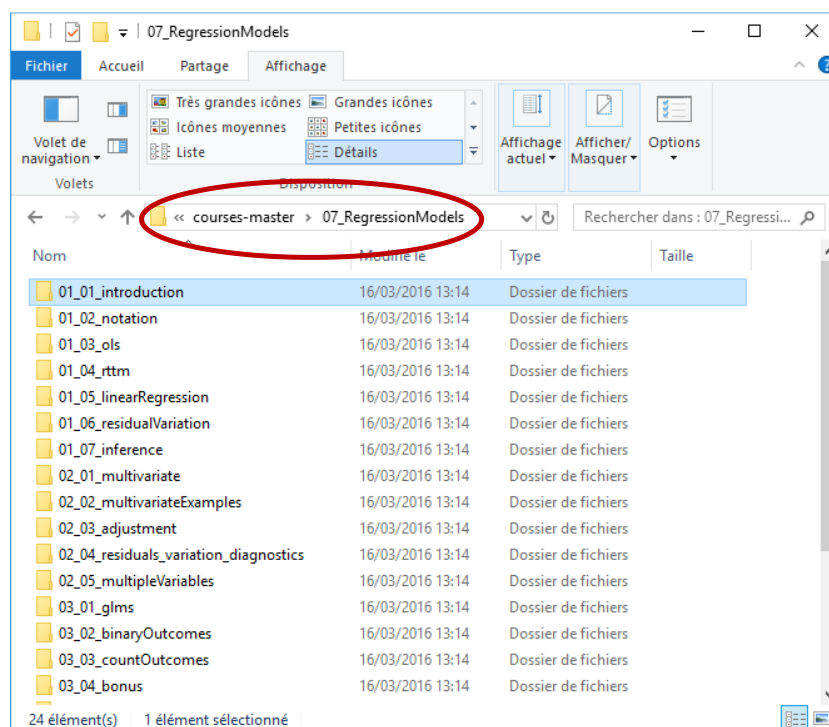
The screenshot shows the GitHub repository page for `DataScienceSpecialization / courses`. The page includes a navigation bar with 'Features', 'Business', 'Explore', 'Marketplace', and 'Pricing'. Below the navigation bar, there are statistics for the repository: 261 commits, 4 branches, 0 releases, and 13 contributors. A 'Clone or download' button is visible, and a dropdown menu is open showing 'Clone with HTTPS' and 'Download ZIP' options. A red arrow points to the 'Download ZIP' button.



L'archive « **courses-master.zip** » est conséquente. Elle correspond à la documentation de l'ensemble des enseignements.

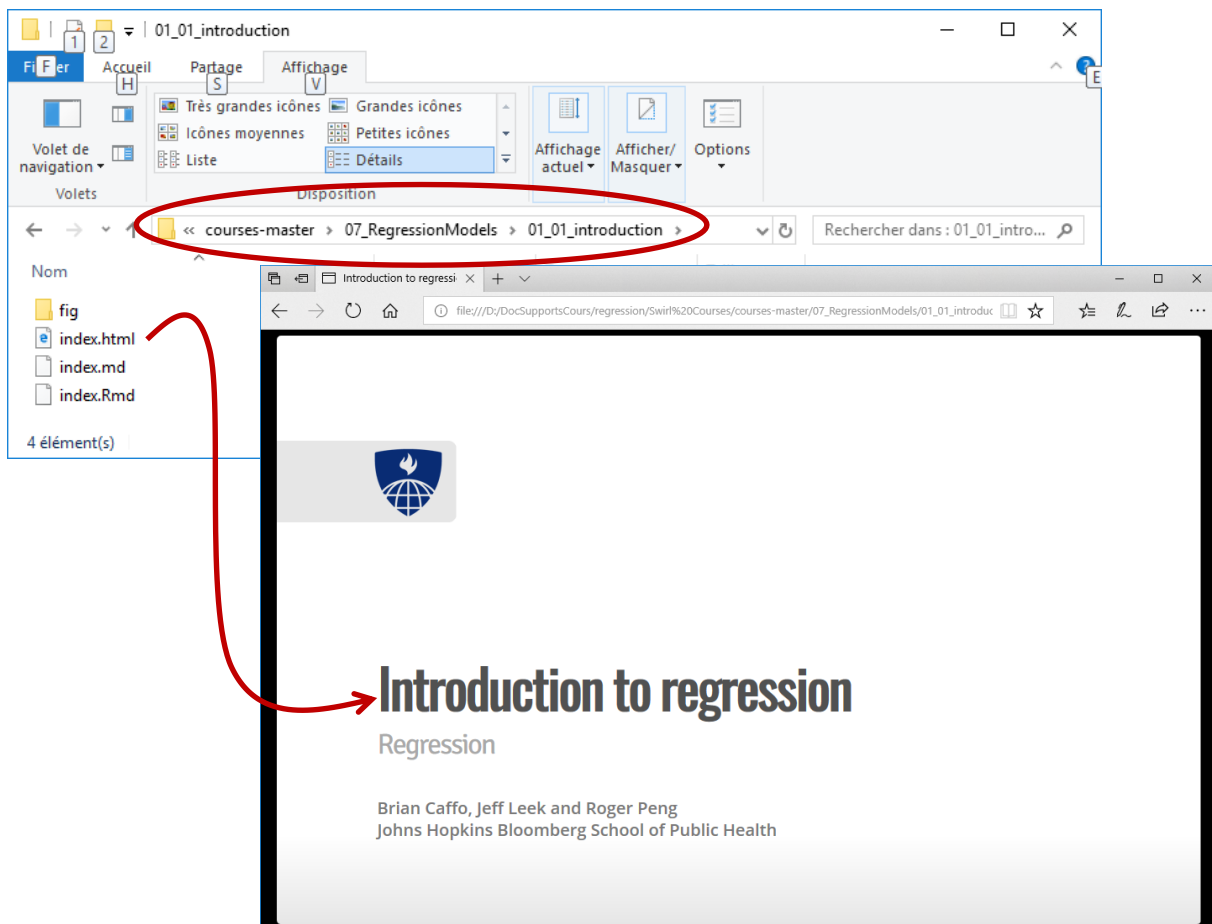


Pour la régression par exemple, nous disposons des supports pour chaque chapitre.





Si l'on s'intéresse au chapitre d'introduction...



... un support de cours de 15 pages est à notre disposition.

Remarque : Très honnêtement, rien que pour la documentation qu'il propose, et je l'ai largement parcourue, ce package vaut le détour.

3.3 Suite du chapitre

Continuons notre tutoriel.

```
|== | 4%
| In this lesson, we'll look at some examples of
| regression models with more than one variable. We'll
| begin with the Swiss data which we've taken the liberty
| to load for you. This data is part of R's datasets
| package. It was gathered in 1888, a time of demographic
| change in Switzerland, and measured six quantities in
| 47 French-speaking provinces of Switzerland. We used
| the code from the slides (the R function pairs) to
| display here a 6 by 6 array of scatterplots showing
| pairwise relationships between the variables. All of
```




```
| the variables, except for fertility, are proportions of
| population. For example, "Examination" shows the
| percentage of draftees receiving the highest mark on an
| army exam, and "Education" the percentage of draftees
| with education beyond primary school.
```

```
...
```

```
|====                                     | 9%
| From the plot, which is NOT one of the factors
| measured?
```

- 1: Infant Mortality
- 2: Catholic
- 3: Fertility
- 4: Obesity

```
Selection: 4
```

On nous demande d'indiquer la variable qui n'est pas présente dans le dataset. Il s'agit d'Obesity, qui est la 4^{ème} dans la liste proposée. Nous le précisons.

```
| You're the best!
```

```
|=====                                     | 13%
| First, use the R function lm to generate the linear
| model "all" in which Fertility is the variable
| dependent on all the others. Use the R shorthand "." to
| represent the five independent variables in the formula
| passed to lm. Remember the data is "swiss".
```

```
>
```

Oui, c'est la bonne réponse et nous sommes les meilleurs. C'est toujours agréable à attendre.

L'outil nous demande ensuite de saisir à la console la commande permettant d'expliquer **Fertility** en fonction des autres variables à l'aide de **lm()**. Je fais exprès de me tromper.

```
> all <- lm(Fertility ~ .)
Error in terms.formula(formula, data = data) :
  '.' dans la formule et pas d'argument 'data'
```

J'ai omis de préciser le paramètre **data** visiblement. Corrigeons cela.

```
> all <- lm(Fertility ~ ., data = swiss)
```

```
| That's a job well done!
```

```
|=====                                     | 17%
```



| Now look at the summary of the linear model all.

Ça va mieux. On nous demande d'afficher le **summary()** du modèle produit.

```
summary(all)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
    Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.91518   10.70604   6.250 1.91e-07 ***
Agriculture  -0.17211    0.07030  -2.448 0.01873 *
Examination  -0.25801    0.25388  -1.016 0.31546
Education    -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic      0.10412    0.03526   2.953 0.00519 **
Infant.Mortality 1.07705    0.38172   2.822 0.00734 **
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10

| Nice work!

|=====| 22%
| Recall that the Estimates are the coefficients of the
| independent variables of the linear model (all of which
| are percentages) and they reflect an estimated change
| in the dependent variable (fertility) when the
| corresponding independent variable changes. So, for
| every 1% increase in percent of males involved in
| agriculture as an occupation we expect a .17 decrease
| in fertility, holding all the other variables constant;
| for every 1% increase in Catholicism, we expect a .10
| increase in fertility, holding all other variables
| constant.

...

```

Notre interlocuteur n'est pas avare d'encouragements.

Il nous décrit par la suite l'interprétation des coefficients du modèle. Puis il nous demande à quel niveau de risque α est significatif le coefficient de la variable **Agriculture**.



```
|=====| 26%
| The "*" at the far end of the row indicates that the
| influence of Agriculture on Fertility is significant.
| At what alpha level is the t-test of Agriculture
| significant?

1: R doesn't say
2: 0.01
3: 0.1
4: 0.05
```

Dans la sortie de **summary()** ci-dessus, nous notons que sa p-value est 0.01873. Le coefficient serait donc significatif à 5%. Nous entrons la valeur 4 (pour le choix n°4).

```
Selection: 4

| Your dedication is inspiring!

|=====| 30%
| Now generate the summary of another linear model (don't
| store it in a new variable) in which Fertility depends
| only on agriculture.
```

...

Etc., etc. Je m'arrêterais ici en ce qui me concerne. Je note quand même que nous assez bien guidés. D'autant plus que la documentation fournie (section 3.2) permet de progresser pas à pas avec une certaine sérénité.

4 Monter son propre cours

4.1 Structure d'un cours pour swirl

Sur le site web de l'éditeur, on nous dit qu'on peut d'apprendre ([Learn](#)) ou enseigner ([Teach](#)). En tant que formateur, on pourrait donc monter notre propre cours et le faire exécuter par l'application. Encore faut-il savoir le faire en accord avec le cahier des charges du package.

Un tutoriel est accessible sur le site web de l'éditeur (<http://swirlstats.com/instructors.html>). Un ensemble de fichiers permettent de définir une leçon (<http://swirlstats.com/swirlify/structure.html>).

En particulier :

- [lesson.yaml](#) qui contient le texte à afficher, ainsi que les questions avec les réponses attendues ;

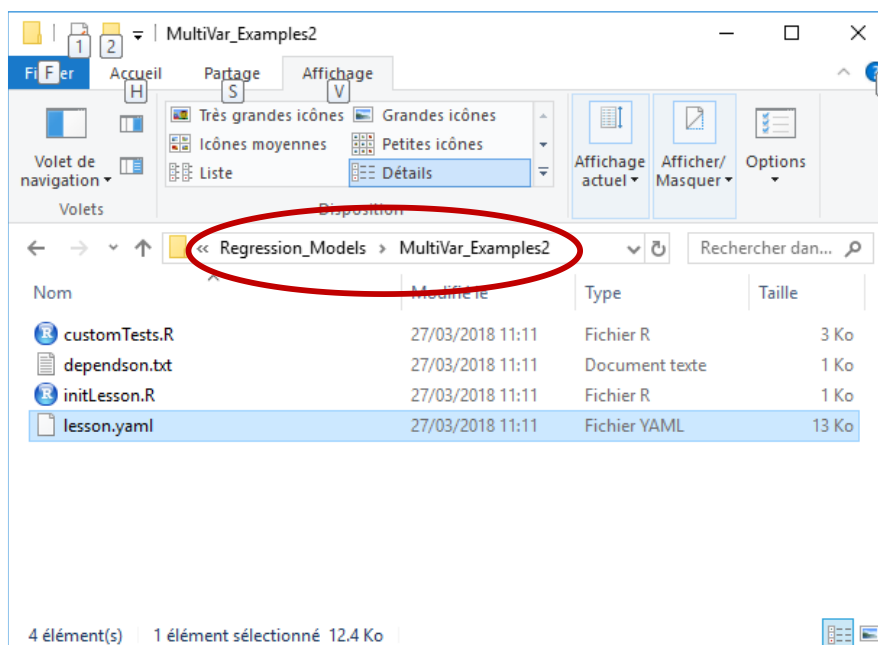


- `initLesson.R` qui contient le script R à exécuter automatiquement lors du démarrage du cours, il peut être pratique par exemple pour charger les données ;
- `dependson.txt` qui énumère les packages R qui sont nécessaires au bon déroulement du cours, swirl peut même les installer le cas échéant ;
- `customTests.R` qui définit les éventuels tests additionnels à mettre en place pour les réponses des apprenants.

D'autres fichiers contenant du code à exécuter ou des données illustratives peuvent être également présents dans le dossier.

4.2 Exemple

Pour le chapitre « Second exemple de la régression linéaire multiple (Regression Models / Multivar_Exemples2) », voici ce que nous trouvons dans le répertoire de la leçon à l'installation du package (les fichiers relatifs au package sont normalement installés dans le dossier Library de votre installation de R si on y dispose des droits en écriture, sinon il faudra chercher dans les dossiers personnels de l'utilisateur courant) :



L'écriture des leçons dans le fichier « `.yaml` » obéit à un formalisme spécifique. Voici les premières lignes du fichier `lesson.yaml` pour ce chapitre.



```
- Class: meta
Course: Regression_Models
Lesson: MultiVar_Examples2
Author: Swirl Coders
Type: Standard
Organization: Johns Hopkins Bloomberg School of Public Health
Version: 2.2.0

- Class: text
Output: "MultiVar_Examples2. (Slides for this and other Data Science courses may
be found at github https://github.com/DataScienceSpecialization/courses. If you
care to use them, they must be downloaded as a zip file and viewed locally. This
lesson corresponds to Regression_Models/02_02_multivariateExamples.)"

- Class: text
Output: This is the second lesson in which we'll look at some regression models
with more than one independent variable. We'll begin with the InsectSprays data
which we've taken the liberty to load for you. This data is part of R's datasets
package. It shows the effectiveness of different insect sprays. We've used the
code from the slides to show you a boxplot of the data.

- Class: exact_question
Output: How many Insect Sprays are in this dataset?
CorrectAnswer: 6
AnswerTests: omnitest(correctVal=6)
Hint: How many letters are in the sequence of letters A, B, C, D, E, F ?

- Class: text_question
Output: From the boxplot, which spray has the largest median?
CorrectAnswer: B
AnswerTests: omnitest(correctVal='B')
Hint: The median of each spray is indicated by the thick horizontal line in each
box. Which spray has its median furthest from 0?

...
```



Chaque item peut correspondre à un texte indicatif simple (`class : text`), soit à un texte associé à une question (`class : ..._question`). Auquel cas, la bonne réponse doit être précisée (`CorrectAnswer` et `AnswerTests`). Une indication permettant de guider l'utilisateur peut être fournie (`Hint`).

Ce ne sont là que quelques exemples. Le formateur dispose d'une large palette d'outils pour organiser son cours et interagir avec l'utilisateur (<http://swirlstats.com/swirlify/writing.html>).

5 Conclusion

Les documents pédagogiques pour l'autoformation peuvent prendre différentes formes. Dans ce tutoriel, nous avons expérimenté le package « **swirl** » pour R qui introduit l'interactivité dans le processus d'apprentissage. Au-delà des cours actuellement disponibles, je note surtout qu'il est relativement facile de monter des leçons avec cet outil (le principe est simple, après il faut rédiger le contenu bien sûr). Je sens que ça va me donner des idées...

6 Références

Swirl : Learn R, in R -- <http://swirlstats.com/>