

ECONOMETRICS AND PROGRAMMATION: PROGRAMMING SUSTAINABLE DEVELOPMENT TOPICS

N. EL MEKKAOUI & Y. FOROUHESHFAR

MASTER 111
SEPT. 2026

Dauphine
UNIVERSITÉ



**PARIS
SCIENCES
& LETTRES**

2EME PARTIE DU COURS

ANALYSE DE DONNÉES ET PROGRAMMATION EN R

Université Paris-Dauphine
Master 111
Yeganeh Forouheshfar
Septembre 2026



PLAN

Download course material from here:
<https://forouheshfar.com/teaching/>

Jour1

- Présentation et Installation de R
- Fonctionnement
- Importation des données
- Analyse statistique

Jour2

- Régressions et tests statistiques
- R Markdown



PRÉSENTATION DE R

HISTOIRE

- Le projet R naît en 1993 comme un projet de recherche de [Ross Ihaka](#) et Robert Gentleman à l'université d'Auckland (Nouvelle-Zélande)
- En 1996 : le projet devient open source
- En 2000: La version R 1.0.0, première version officielle du langage R, est publiée
- En 2003: l'équipe de développement crée la *R Foundation for Statistical Computing*
- Aujourd'hui : R est un langage mature, activement maintenu, avec un vaste écosystème de packages.
- Dernière version : juin 2026 (une nouvelle version chaque année)



LES AVANTAGES DE R

- R est une option viable et performante pour l'analyse de données
- R est **libre et gratuit** donc le seul coût est celui d'apprentissage
- Un logiciel multiplateforme (Linux, Mac et Windows)
- Compatible avec C++, Java et Python
- Possède, un système de documentation intégré très bien conçu.
- Un outil puissant pour mener des analyses de donnée, Machine Learning et programmation.
- R est largement utilisé par les académiques et dans les secteurs de la santé et de la finance et aussi par les compagnies (Uber, Google, Airbnb, Facebook, etc.)
- 1^{er} outil utilisé dans « Data Science »

LES INCONVÉNIENTS DE R

- R nécessite des machines relativement puissantes pour travailler sur de grosses base de données de plusieurs milliers d'individus
- La qualité du résultat dépend de la qualité du code, de la documentation et des choix économétriques.
- Les très gros volumes de données peuvent nécessiter des outils adaptés (data.table, Arrow, bases SQL, etc.).
- Les erreurs sont parfois peu intuitives au début : apprendre à lire les messages d'erreur fait partie du cours.
- Dans R, par défaut, l'affichage est réduit au minimum, et c'est l'utilisateur qui demande à voir des résultats supplémentaires ou plus détaillés.

LES SOURCES



CRAN
[Mirrors](#)
[What's new?](#)
[Search](#)
[CRAN Team](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Task Views](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)

Donations
[Donate](#)



• [Le CRAN: The Comprehensive R Archive Network](#)

• <https://cran.r-project.org/>

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) ([Debian](#), [Fedora/Redhat](#), [Ubuntu](#))
- [Download R for macOS](#)
- [Download R for Windows](#)



R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2026-06-24, Happy Hop) [R-4.6.1.tar.gz](#), read [what's new](#) in the latest version.
- The CRAN directory [src/base-prerelease](#) contains R alpha, beta, and rc releases as daily snapshots in time periods before a planned release.
- Between releases, the same directory [src/base-prerelease](#) contains snapshots of current patched and development versions. Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Alternatively, daily snapshots are [available here](#).
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#).



INSTALLATION DE R



[HTTPS://CLOUD.R-PROJECT.ORG/](https://cloud.r-project.org/)

L'INSTALLATION DE R EST TRÈS FACILE, IL SUFFIT DE SUIVRE LES INSTRUCTIONS

[HTTPS://POSIT.CO/DOWNLOAD/RSTUDIO-DESKTOP/](https://posit.co/download/rstudio-desktop/)





Introduction

Getting Started

Guide



Accessibility



User Interface



Code



Data



Tools



Productivity



Computational
Documents



Deploy



Package Development



Environments



Troubleshooting



Reference



Keyboard Shortcuts

Environment variables

RStudio IDE

Direct download links for RStudio Desktop 2026.08.2, the version this documentation describes:

Platform	Download
Windows 11	RStudio-2026.08.2-200.exe
Windows 11 (zip archive)	RStudio-2026.08.2-200.zip
macOS 14+	RStudio-2026.08.2-200.dmg
Ubuntu 22 / 24 / 26, Debian 13	rstudio-2026.08.2-200-amd64.deb
Ubuntu / Debian (tar.gz)	rstudio-2026.08.2-200-amd64-debian.tar.gz
RHEL 9 / 10, Fedora	rstudio-2026.08.2-200-x86_64.rpm
RHEL / Fedora (tar.gz)	rstudio-2026.08.2-200-x86_64-fedora.tar.gz

On this page

About this guide

Intended audience

Commercial Edition

Install Links

Open Source

Commercial

**Direct Downloads (Open
Source)**

Direct Downloads
(Commercial)

Edit this page

View source

https://docs.posit.co/ide/user/?_gl=1*10jz99n*_ga*NDE5MDU1NDk5LjE3ODg5ODY2MjU.*_ga_G020NYQMP9*cZ_E3ODg5OTA0MzUkbzEkZzAkDE3ODg5OTA0NDEkajYwJGwwJGgw*_gcl_au*MTg2NTQxNzg5LjE3ODg5OTA0NDE.#rstudio-ide-oss-downloads

- Rstudio est un environnement de travail pour R,

- Mais y en a d'autres par exemple ESS:

<http://ess.r-project.org/index.php?Section=download>



Une fenêtre typique de RStudio

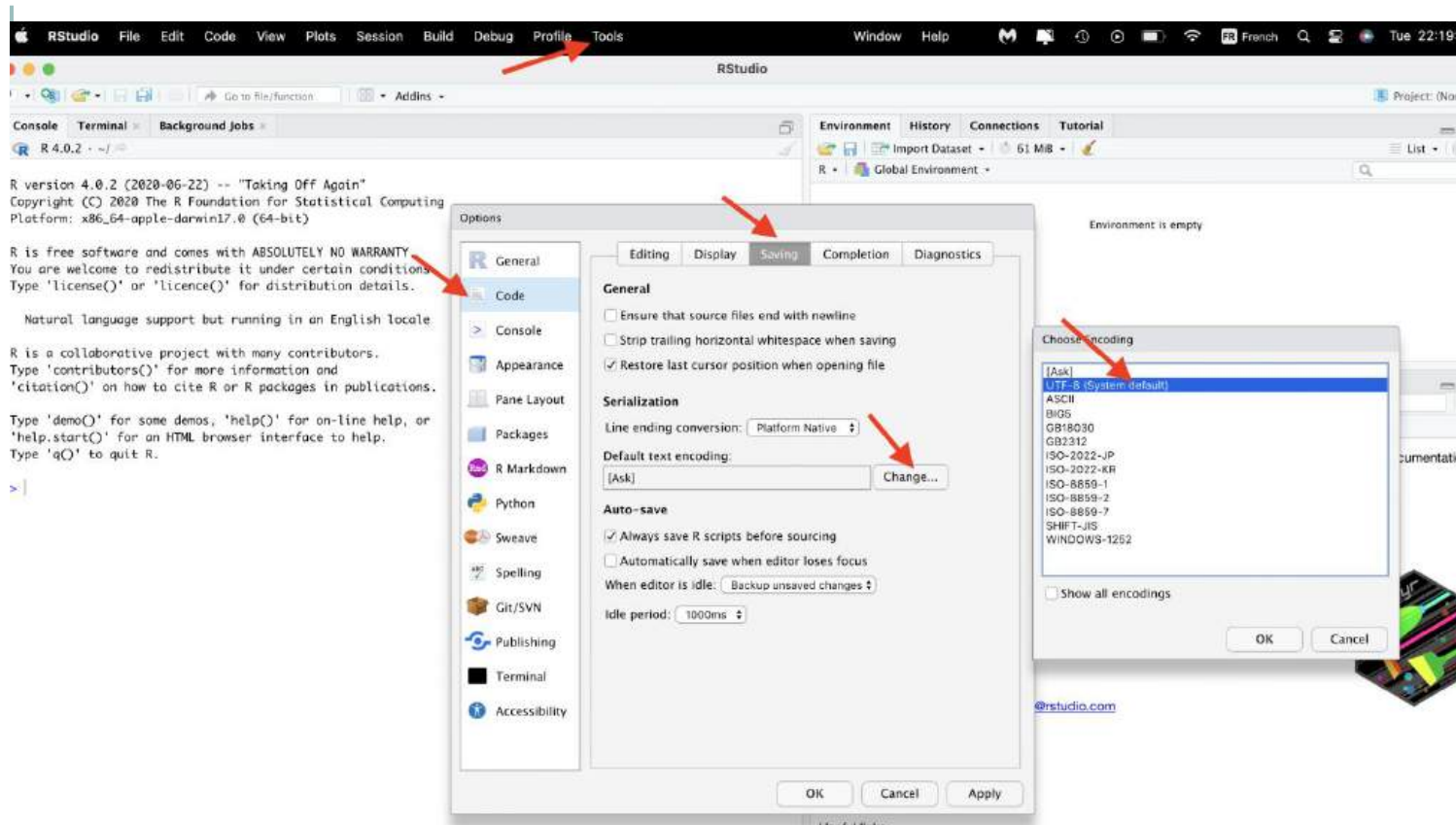
The screenshot shows the RStudio interface with four red numbers highlighting specific components:

- 1** points to the **Source** editor, which contains R code for installing and loading packages. The code includes comments and function calls like `library(tidyverse)`, `library(rmarkdown)`, `install.packages("lme4")`, `library(lme4)`, `install.packages("emmeans")`, `library(emmeans)`, and `install.packages("effects")`.
- 2** points to the **Console** pane, which shows the output of the R script. It displays the download progress for the `lme4` and `emmeans` packages, including the URL and file size.
- 3** points to the **Environment** pane, which shows the objects in the current R session. It lists `data1` and `dataB`, both with 7 observations and 0 variables.
- 4** points to the **Script** editor, which shows the R code being executed. The code includes `read.table`, `read.csv`, `read.csv2`, `read.delim`, and `read.delim2` functions.

A VOUS DE VOIR

- Installation R et Rstudio
- Ouvrir Rstudio
- Familiarisation avec l'interface

Manipulation dans les préférences → pour avoir les accents





FONCTIONNEMENT

LES DIFFÉRENTES AIDES

- L'aide en ligne:
Pour l'afficher directement dans l'interface, on tape
 `help()`
 `?`
- Pour l'afficher dans un navigateur au format HTML, on tape:
 `help.start()`
- L'aide sur le CRAN:
 `RsiteSearch("nom de la commande")`

LES OBJETS

- Le langage R est dit “orienté objet”/ *Object oriented*
- Nous manipulons des objets : variables, vecteurs, tableaux, modèles, graphiques...
- Pour chaque objet, trois questions :
 - ← Quel est son type ? numeric, character, logical...
 - ← Quelle est sa structure / classe ? vector, matrix, data.frame, list...
 - ← Que contient-il et quelle est sa dimension ?
- R supporte plusieurs systèmes orientés objet, mais pour ce cours nous nous concentrons surtout sur la manipulation des objets.

LE TYPE D'UN OBJET

- Objet vide : null
- Objet booléen que nous notons logique avec les valeurs soit TRUE/T soit FALSE/F : logical
- Nombre réel que nous notons numérique : numeric
- Nombre complexe que nous notons complexe : complex
- Chaîne de caractères que nous notons caractère : character

STRUCTURE D'UN OBJET DANS R/ CLASSE

- **Vecteur** — une dimension, un seul type de données.
 - **Matrice** — deux dimensions, un seul type.
 - **Array** — extension multidimensionnelle de la matrice.
 - **data.frame** — colonnes de même longueur pouvant avoir des types différents ; structure centrale pour l'économétrie.
 - **Liste** — collection ordonnée d'objets potentiellement très différents.
-
- **Réflexe** : utilisez `str(objet)`, `class(objet)`, `dim(objet)` pour comprendre ce que vous manipulez.

POUR CONSTRUIRE UN VECTEUR:

- **<-** : opérateur qui désigne des valeur d'une variable.
- **c** : Construction par la fonction collecteur « c »
- Construction par l'opérateur séquence « : »
- Construction par la fonction séquence « seq »
- Construction par la fonction « rep »

FONCTIONS

- **Une fonction** est un objet R. Un grand nombre de fonctions sont prédéfinies dans R, cependant, il est possible de créer ses propres fonctions.
- Une fonction admet des arguments en entrée et retourne un résultat en sortie.

`FunctionName(arg.1 = value.1, arg.2 = value.2, ..., arg.n = value.n)`

- R dispose de fonctions pour faire des analyses basiques : `summary()`, ...

FONCTIONS MATHÉMATIQUES

Commande R	Résultat	Exemple
<code>sqrt(x)</code>	$\sqrt{x}$	<code>sqrt(4) => 2</code>
<code>abs(x)</code>	$ x $	<code>abs(-4) => 4</code>
<code>log(x,base=y)</code>	$\log(x)$	<code>log(1,base=0) => 0</code>
<code>log(x)</code>	$\ln(x)$	<code>log(1) => 0</code>
<code>factorial(x)</code>	$x!$	<code>factorial(3) => 6</code>
<code>floor(x)</code>	entier inférieur	<code>floor(1.1) => 1</code>
<code>ceiling(x)</code>	entier supérieur	<code>ceiling(1.1) => 2</code>
<code>round(x)</code>	entier le + proche	<code>round(1.1) => 1</code>
<code>min(x)</code>	élément minimum	<code>min(c(8,1,5)) => 1</code>
<code>max(x)</code>	élément maximum	<code>max(c(8,1,5)) => 8</code>
<code>a%%n</code>	modulo a mod n	<code>7%%3 => 1</code>

FONCTIONS STATISTIQUES

Commande R	Résultat
mean(x)	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
sum(x)	$\sum_{i=1}^n x_i$
prod(x)	$\prod_{i=1}^n x_i$
quantile(x,probs=p)	y tel que $P(X < y) = p$
median(x)	médiane
var(x)	$s'^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
sd(x)	écart-type corrigé s'
IQR(x)	écart interquartile Q3-Q1
summary(x)	min,Q1,Q2, $\bar{x}$,Q3,max

PACKAGES

- Un package ajoute des fonctions, jeux de données ou méthodes à R.
- 2 types de packages:
 - ← Base: déjà disponible, il suffit de les appeler ...
 - ← Contributed: il faut d'abord les télécharger ensuite les appeler.
- Installer une seule fois : `install.packages("nom_du_package")`
- Charger à chaque nouvelle session : `library(nom_du_package)`
- `pacman::p_load()` peut installer si nécessaire puis charger plusieurs package
- Bonne pratique : au début d'un script, charger uniquement les packages réellement nécessaires.



() = run something

[] = select something

[[]] = extract one thing

{ } = group commands

Example:

AVANT EXO 1 — LES BRIQUES DE BASE DE R

- À maîtriser : objets, affectation `<-`, types et structures de données.
- Commandes : `typeof()`, `class()`, `is.vector()`, `matrix()`, `as.data.frame()`, `list()`, `rm()`.
- Comprendre la différence entre vecteur, matrice, `data.frame` et liste.
- Packages : installer une fois, charger à chaque session.

Question flash : avant d'exécuter une ligne, quel objet pensez-vous qu'elle va créer ?

EXO 1 — LES BASES DE R

- Ouvrez “EXO1.R” dans RStudio.
- Exécutez le code par blocs et prédisez le résultat avant de lancer chaque commande.
- Objectif : comprendre ce qu’est un objet R et savoir inspecter sa structure.

IMPORTATION DES DONNÉES

IMPORTATION DES DONNÉES

- Formats des données:

- ← CSV

- ← TXT

- ← XLSX

- ↓ Utilisation du package “rio”

WORKING DIRECTORY & CHEMINS DE FICHIERS

- Working directory : le dossier de référence à partir duquel R cherche et enregistre les fichiers.
- `getwd()` affiche le dossier courant ; `setwd("...")` permet de le modifier.
- Chemin absolu : adresse complète du fichier (ex. `C:/Users/...` ou `/Users/...`).
- Chemin relatif : adresse à partir du dossier du projet (ex. `data/donnees.xlsx`) — plus portable et reproductible.
- Bonne pratique : créer un projet RStudio et organiser les fichiers dans des dossiers `data/`, `scripts/` et `outputs/`.

Réflexe en cas d'erreur "file not found" : vérifiez d'abord `getwd()`, le nom du fichier et son chemin. `file.exists("data/donnees.xlsx")` permet de tester le chemin.

AVANT EXO 2 — IMPORTER SANS PERDRE LE CONTRÔLE

- Formats : CSV, TXT, XLSX — mêmes données, méthodes d'import différentes.
- Workflow : importer → head() / View() → str() → vérifier types et noms de variables.
- Accès à une variable : **data\$variable**.
- Renommer une colonne sans modifier par erreur le reste du jeu de données.

Piège à éviter : réussir l'importation ne signifie pas que les variables ont été lues avec le bon type.

EXO 2 — IMPORTER ET INSPECTER DES DONNÉES

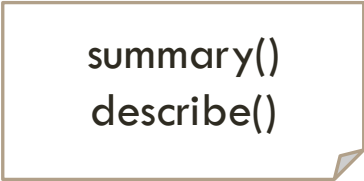
- Ouvrez “EXO2.R” dans RStudio.
- Importez plusieurs formats et vérifiez immédiatement la structure des variables.
- Objectif : obtenir un jeu de données proprement chargé et compréhensible.



ANALYSE STATISTIQUE

ANALYSE STATISTIQUE : COMMENCER PAR COMPRENDRE LES DONNÉES

- Identifier le type de chaque variable : quantitative, catégorielle, binaire.
- Examiner tendance centrale, dispersion, valeurs extrêmes et données manquantes.
- `summary()` / `describe()` pour un premier diagnostic.
- Corrélation : mesure une association linéaire, pas une causalité.
- Toujours associer les statistiques descriptives à une visualisation adaptée.



```
summary()  
describe()
```

CHOISIR UN GRAPHIQUE EN FONCTION DE LA QUESTION

- Distribution d'une variable quantitative → **hist**
- Variable catégorielle → **barplot**
- Quantitative × catégorielle → **boxplot**
- Quantitative × quantitative → **scatterplot**
- **ggplot2** : construire le graphique couche par couche et rendre le code reproductible.
- Question à se poser : que doit comprendre le lecteur en 5 secondes ?

AVANT EXO 3 — DÉCRIRE, VISUALISER, COMPARER

- Commencer par le type de variable : il détermine le résumé et le graphique pertinents.
- `summary()` / `describe()` : position, dispersion, minimum, maximum, valeurs atypiques.
- Histogramme, barplot, boxplot ou scatterplot : un graphique = une question.
- T-test : H_0 = égalité des moyennes ; interpréter statistique, p-value et intervalle de confiance.

Piège à éviter : $p < 0,05$ ne vous dit ni la taille de la différence ni son importance économique.

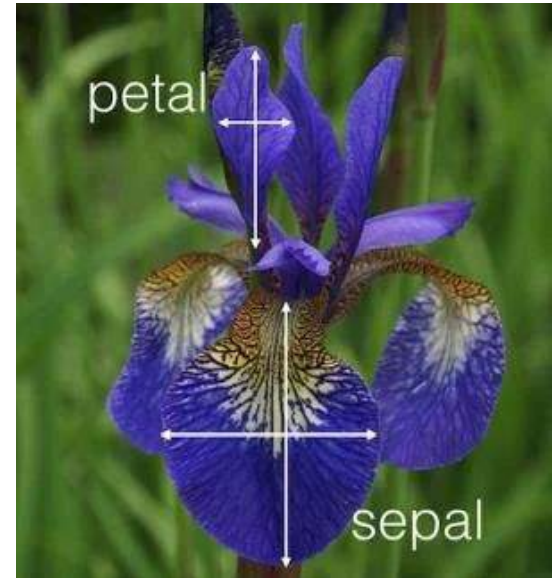
LE T-TEST : COMPARER DEUX MOYENNES

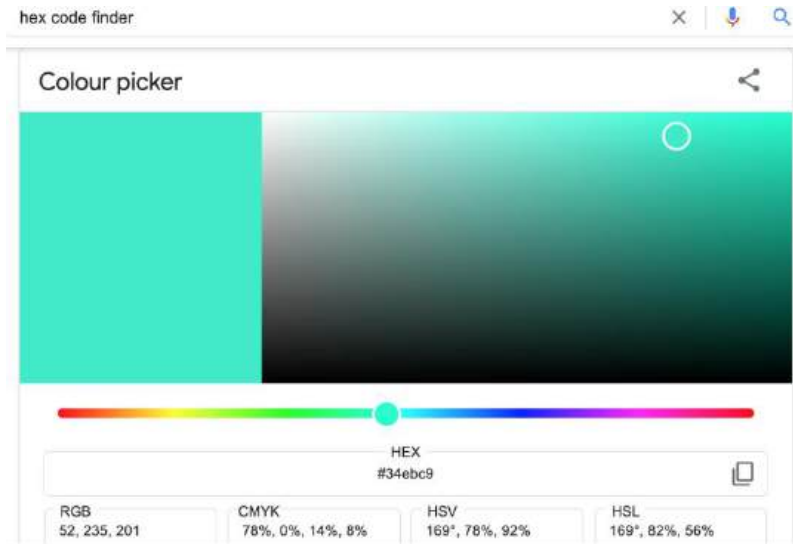
- Question : l'écart entre deux moyennes reflète-t-il une différence statistiquement détectable ou le hasard d'échantillonnage ?
- Hypothèses : $H_0 : \mu_1 = \mu_2$ | $H_1 : \mu_1 \neq \mu_2$. Le t-stat rapporte la différence observée à son erreur standard.
- À lire ensemble : différence estimée, intervalle de confiance à 95 %, statistique t et p-value.
- Dans R : `t.test(y ~ groupe, data = df)`. Par défaut, R utilise le test de Welch : les variances ne sont pas supposées égales.
- Conditions : Y quantitative, observations indépendantes et deux groupes ; données appariées $\rightarrow$ `paired = TRUE`.

À retenir : significativité statistique $\neq$ importance économique. Une p-value $> 0,05$ ne prouve pas que les deux moyennes sont identiques.

EXO 3 — STATISTIQUES DESCRIPTIVES, GRAPHIQUES ET T-TEST

- Ouvrez “EXO3.R” dans RStudio.
- Décrivez les données, choisissez les graphiques adaptés et réalisez un t-test.
- Objectif : produire une courte lecture économique des résultats, pas seulement du code.





pch arguments

□ 0	○ 1	△ 2	+ 3	× 4
◇ 5	▽ 6	⊠ 7	* 8	⬠ 9
⊕ 10	⊗ 11	⊞ 12	⊗ 13	⊞ 14
■ 15	● 16	▲ 17	◆ 18	● 19
● 20	● 21	■ 22	◆ 23	▲ 24
▼ 25				

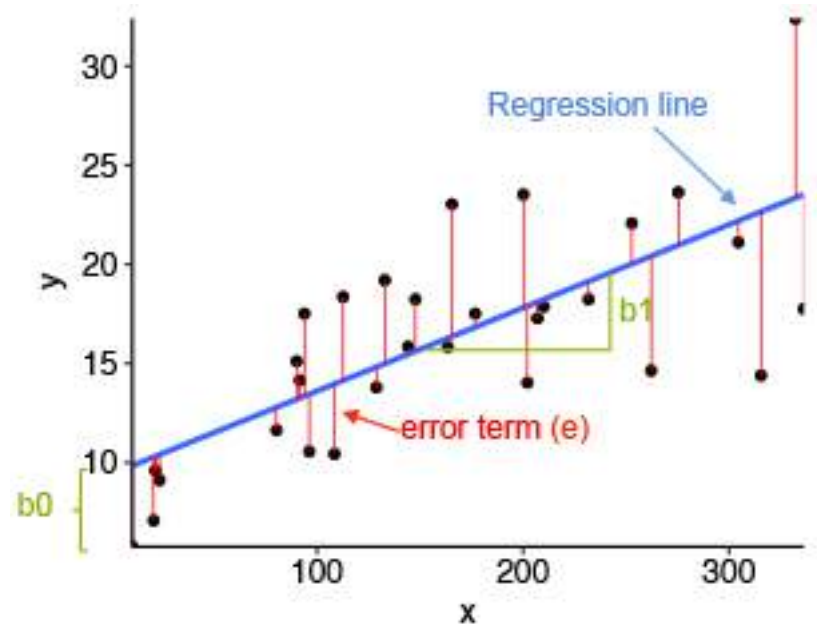
And a lot more!



REGRESSIONS ET TESTS STATISTIQUES

LE MODÈLE LINÉAIRE : RELIER UNE QUESTION ÉCONOMIQUE À UNE ÉQUATION

- Modèle simple : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- β_1 mesure la variation moyenne de Y associée à une unité supplémentaire de X .
- Modèle multiple :
 $Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + \varepsilon_i$
- Interprétation : β_j est un effet partiel — “toutes choses égales par ailleurs” dans le modèle.
- Une régression estime une association conditionnelle ; elle n'établit pas automatiquement une causalité.



EXO 4 — QUE VOULONS-NOUS ANALYSER ?

- Données : USJudgeRatings — 43 juges, évalués par des avocats sur 12 dimensions.
- $Y = RTEN$: évaluation du caractère « digne d'être maintenu en fonction ». $X = \text{contacts (CONT), intégrité, comportement, diligence, gestion des dossiers, décisions, préparation, connaissance du droit, etc.}$
- Question : quelles dimensions de l'évaluation sont associées à une meilleure note de maintien en fonction ?
- Stratégie : $RTEN \sim \text{CONT} \rightarrow \text{modèle multiple} \rightarrow \text{comparer les coefficients} \rightarrow \text{examiner intervalles de confiance, résidus et test de Breusch-Pagan.}$

À la fin : traduire un coefficient OLS en langage clair et distinguer association, significativité statistique et validité de l'inférence.

EXO 4 — PREMIÈRE RÉGRESSION LINÉAIRE

- Ouvrez “EXO4.R” dans RStudio.
- Estimez un modèle simple puis multiple avec `lm()`.
- Interprétez coefficients, incertitude et diagnostics de résidus.

`[]` **square brackets** → used to select/subset elements

`</>` R

```
x[1]           # first element of x
x[2:5]         # elements 2 to 5
data[1, 2]     # row 1, column 2
data[, 2]      # all rows, column 2
```

AVANT EXO 5 — CONSTRUIRE UN WORKFLOW ÉCONOMÉTRIQUE COMPLET

- 1. Comprendre les variables avant le modèle.
- 2. Examiner distributions, corrélations et graphiques.
- 3. Estimer une spécification de référence puis une alternative justifiée.
- 4. Vérifier multicolinéarité, cohérence des signes et robustesse de l'interprétation.
- 5. Traduire les sorties R en conclusions substantives.

Réflexe économiste : une variable ne doit pas être supprimée uniquement parce que sa p-value est élevée.

EXO 5 — QUE VOULONS-NOUS ANALYSER ?

- Données : Fair (1977). Y = nombre d'aventures extraconjugales au cours de l'année précédente.
- X : genre, âge, années de mariage, enfants, religiosité, éducation, occupation et satisfaction conjugale.
- Questions : quels facteurs sont associés à Y ? Les moyennes diffèrent-elles selon le genre ou la présence d'enfants ?
- Stratégie : statistiques descriptives + corrélations + graphiques → modèle complet → modèle réduit → t-tests.

AVANT EXO 6 — POURQUOI UTILISER DES LOGARITHMES ?

- Les logs peuvent linéariser une relation non linéaire et réduire l'influence des très grandes valeurs.
- Log-level et level-log : l'interprétation change ; ne pas appliquer mécaniquement la règle de l'élasticité
- `l()` dans une formule R permet d'utiliser une transformation arithmétique comme X^2 .
- `ggplot2` : données + `aes()` + `geom_*()` + couches.

Piège à éviter : `log(0)` et `log(valeur négative)` ne sont pas définis dans ce contexte.

EXO 6 — LOG-LINÉARISATION, RÉGRESSION ET VISUALISATION

- Utilisez le jeu de données Journals.
- Explorez la non-linéarité et comparez niveaux vs logarithmes.
- Estimez une relation log-log et interprétez l'élasticité.
- Créez puis exportez un graphique avec ggplot2.

Most commonly: $y \sim x$ means
“y is explained by x”

Usage: Specification of plots,
models, etc.

AVANT EXO 7 — LIRE UNE ÉQUATION DE SALAIRE

- $\log(\text{wage})$ comme variable dépendante : les coefficients continus s'interprètent approximativement en %.
- education : rendement associé à une année supplémentaire, conditionnellement aux autres variables.
- $\text{experience} + \text{experience}^2$: le rendement de l'expérience peut varier avec le niveau d'expérience.
- Variable catégorielle : le coefficient compare une catégorie à une catégorie de référence.
- Avec un terme quadratique, l'effet marginal de l'expérience dépend de $\beta_2 + 2\beta_3 \cdot \text{experience}$.

$$\begin{aligned} \log(\text{wage}) = & \beta_1 + \beta_2 \text{experience} + \beta_3 \text{experience}^2 \\ & + \beta_4 \text{education} + \beta_5 \text{ethnicity} + \varepsilon \end{aligned}$$

EXO 7 — UNE ÉQUATION DE SALAIRE AVEC CPS1988

- Chargez `data("CPS1988", package = "AER")`.
- Explorez `salaire`, `éducation`, `expérience`, `temps partiel` et `ethnicité`.
- Lancer la regression suivante et expliquez pourquoi le terme quadratique change l'interprétation de l'expérience.

$$\begin{aligned}\log(\text{wage}) = & \beta_1 + \beta_2 \text{experience} + \beta_3 \text{experience}^2 \\ & + \beta_4 \text{education} + \beta_5 \text{ethnicity} + \varepsilon\end{aligned}$$

AVANT EXO 8 — QUAND Y EST BINAIRE

- Si $Y \in \{0,1\}$, nous modélisons une probabilité conditionnelle $P(Y=1 | X)$.
- Logit et Probit maintiennent les probabilités prédites entre 0 et 1.
- R : `glm(..., family = binomial(link = "logit"))` ou `link = "probit"`.
- Le signe du coefficient indique le sens de l'association ; sa valeur brute n'est pas un changement direct en points de probabilité.
- Pour communiquer : privilégier probabilités prédites et/ou effets marginaux.

Piège à éviter : comparer directement la taille des coefficients Logit et Probit comme s'ils avaient la même échelle.

LE MODÈLE PROBIT — INTUITION ET INTERPRÉTATION

- Quand Y est binaire, le Probit relie les variables X à la probabilité que $Y = 1$.
- Variable latente :
$$Y_i^* = X_i\beta + \varepsilon_i \quad \text{avec } \varepsilon_i \sim N(0,1).$$
- Règle d'observation : $Y_i = 1$ si $Y_i^* > 0$; sinon $Y_i = 0$.
- Donc : $P(Y_i = 1 \mid X_i) = \Phi(X_i\beta)$, où Φ est la fonction de répartition de la loi normale standard.
- Interprétation : le signe de β_k indique le sens de l'effet, mais β_k n'est pas directement un effet en points de probabilité.
- Effet marginal : $\partial P(Y=1 \mid X) / \partial X_k = \phi(X\beta)\beta_k$; il dépend donc des valeurs de X .
- R : `glm(Y ~ X1 + X2, family = binomial(link = "probit"), data = df)`

À retenir : dans un modèle Probit, privilégier les probabilités prédites et les effets marginaux pour communiquer les résultats.

EXO 8 — MODÈLES LOGIT ET PROBIT

- Ouvrez infidelity_probit.xlsx.
- Identifiez la variable dépendante binaire et les variables explicatives.
- Estimez Logit et Probit puis comparez les probabilités prédites.
- Objectif : distinguer coefficient du modèle et effet sur la probabilité.

AVANT EXO 9 — DE L'API AU JEU DE DONNÉES EXPLOITABLE

- Une API ou «Application Programming Interface» (en français : interface de programmation applicative) permet d'interroger une base publique directement depuis R.
- Toujours documenter : indicateur, définition, unité, pays, période et source.
- Vérifier les données manquantes et les changements de couverture temporelle.
- Pour un panel : une ligne correspond généralement à une unité i et une période t .
- Avant toute régression, visualiser l'évolution de quelques pays et comparer la couverture.

EXO 9 — IMPORTER DES DONNÉES DE LA BANQUE MONDIALE

- Importer automatiquement des indicateurs depuis la Banque mondiale.
- Identifier pays, année, indicateur, unité et données manquantes.
- Transformer les observations en structure pays $\times$ temps (panel)
- Objectif : préparer un jeu de données prêt pour l'analyse en panel.

QU'EST-CE QU'UNE DONNÉE DE PANEL ?

Un panel observe les mêmes unités i (pays, entreprises, ménages...) à plusieurs dates t .

Il combine dimension transversale et dimension temporelle : Y_{it} , X_{it} .

N = nombre d'unités ; T = nombre de périodes.

Panel équilibré : chaque unité est observée à toutes les périodes.

Panel déséquilibré : certaines observations unité-temps sont absentes.

POOLED OLS : LE POINT DE DÉPART

- $Y_{it} = \beta_0 + \beta X_{it} + u_{it}$
- Le pooled OLS empile toutes les observations et ignore l'hétérogénéité spécifique aux unités.
- Il sert de benchmark, mais peut être biaisé si des caractéristiques inobservées propres à chaque unité sont corrélées avec X .

Le panel contient plus d'information, mais pooled OLS ne l'exploite pas complètement.

MODÈLE À EFFETS FIXES (FIXED EFFECTS, FE)

- $Y_{it} = \alpha_i + \beta X_{it} + u_{it}$
- Chaque unité possède sa propre constante α_i .
- FE neutralise les caractéristiques inobservées invariantes dans le temps, même si elles sont corrélées avec X .
- Intuition : on compare essentiellement une unité à elle-même au cours du temps — variation within.
- Les variables parfaitement constantes dans le temps ne sont pas identifiées séparément des effets fixes individuels.

Question clé : X varie-t-il suffisamment dans le temps au sein d'une même unité ?

EFFETS FIXES TEMPORELS ET TWO-WAY FIXED EFFECTS

- Effets fixes individuels α_i : contrôlent les facteurs propres à chaque unité et invariants dans le temps.
- Effets fixes temporels λ_t : contrôlent les chocs communs à toutes les unités à une date donnée.
- Two-way FE : $Y_{it} = \alpha_i + \lambda_t + \beta X_{it} + u_{it}$.
- Exemples de chocs temporels : récession mondiale, pandémie, réforme commune, changement de prix international.

Question : quel choc commun à toutes les entreprises pourrait sinon être attribué à tort à X ?

MODÈLE À EFFETS ALÉATOIRES (RANDOM EFFECTS, RE)

- $Y_{it} = \beta_0 + \beta X_{it} + \alpha_i + u_{it}$
- RE traite l'hétérogénéité individuelle α_i comme une composante aléatoire.
- Hypothèse clé : $\text{Cov}(\alpha_i, X_{it}) = 0$.
- RE exploite variation within et between et permet d'estimer des variables invariantes dans le temps.
- Cette hypothèse doit être économiquement crédible, pas seulement statistiquement commode.

FE autorise la corrélation α_i - X_{it} ; RE suppose cette corrélation nulle.

LE TEST DE HAUSMAN : FE OU RE ?

- H_0 : les effets individuels sont non corrélés avec les régresseurs $\rightarrow$ RE est cohérent.
- H_1 : les effets individuels sont corrélés avec les régresseurs $\rightarrow$ RE devient incohérent ; FE reste approprié.
- p-value significative $\rightarrow$ rejeter H_0 : privilégier FE.
- p-value non significative $\rightarrow$ on ne rejette pas H_0 ; cela ne prouve pas que RE est “vrai”.
- Toujours combiner le test avec le raisonnement économique.

Un test est une information pour décider, pas un substitut à la réflexion sur les hypothèses.

VARIATION WITHIN ET BETWEEN

- Within : comment une même unité change au cours du temps.
- Between : comment les niveaux moyens diffèrent entre unités.
- FE identifie β à partir de la variation within.
- RE combine information within et between sous son hypothèse d'exogénéité.
- Avant d'estimer : vérifier N, T, couverture temporelle, données manquantes et variation des variables.
- R : `pdata.frame()`, `pdim()`, `summary()` avec le package `plm`.

Si X varie très peu within, un modèle FE aura peu d'information pour estimer son effet.

EXO 10

- **Travailler avec des données de panel**
 - **Étape 1** : Charger le jeu de données *Grunfeld* (véritables données de panel).
 - **Étape 2** : Déclarer la structure de panel avec `pdata.frame`, en indiquant à R quelle variable correspond à l'individu (entreprise) et laquelle correspond au temps (année).
 - **Étape 3** : Estimer les modèles :
- **Moindres carrés ordinaires groupés (Pooled OLS)** : Ignore la nature en panel.
- **Effets fixes (FE)** : Contrôle les caractéristiques inobservées propres à chaque entreprise.
- **Effets aléatoires (RE)** : Suppose que les effets spécifiques aux entreprises sont aléatoires.
- **Étape 4** : Utiliser le test de Hausman pour choisir entre FE et RE.

LES MODÈLES DE RÉGRESSION EN PANEL

1) Régression OLS simple

- Ne tient pas compte de l'hétérogénéité individuelle.
- Risque de biais si les effets spécifiques sont corrélés avec les régresseurs.
- Sert souvent de point de comparaison, mais elle est moins adaptée aux données de panel.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + u_{it}$$

Modèle à effets fixes (Fixed Effects, FE)

- Hypothèse : les spécificités propres à chaque individu (pays, entreprise, ménage) sont corrélées avec les variables explicatives.
- Méthode : chaque individu dispose de sa propre constante, ce qui permet de neutraliser les caractéristiques inobservées qui ne varient pas dans le temps.
- Avantage : contrôle du biais dû aux variables inobservées corrélées avec les régresseurs.

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$$

Modèle à effets aléatoires (Random Effects, RE)

- Hypothèse : les spécificités individuelles sont **non corrélées** avec les variables explicatives.
- Méthode : on considère l'hétérogénéité comme une composante aléatoire de l'erreur.
- Avantage : permet d'exploiter à la fois la variation intra-individuelle (dans le temps) et inter-individuelle (entre individus).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \mu_i + u_{it}$$

LE TEST DE HAUSMAN

Le **test de Hausman** est utilisé pour choisir entre le modèle à effets fixes et le modèle à effets aléatoires :

- **Hypothèse nulle (H_0)** : les effets individuels sont non corrélés avec les régresseurs → le modèle à effets aléatoires est efficace et non biaisé.
- **Hypothèse alternative (H_1)** : les effets individuels sont corrélés avec les régresseurs → le modèle à effets fixes est préférable.

En pratique :

- Si la statistique de Hausman est significative → on rejette H_0 et on retient le modèle **à effets fixes**.
- Si elle n'est pas significative → on peut utiliser le modèle **à effets aléatoires**.

STATISTIQUES DESCRIPTIVES DE PANEL

Variance intra-individuelle et inter-individuelle : très spécifiques aux données de panel.

- La variation **intra-individuelle** mesure l'évolution d'un même individu dans le temps.
- La variation **inter-individuelle** compare les différences entre individus.

```
library(plm)  
pdim(Grunfeld) # Variance intra/inter (dimensions du panel)
```

Interprétation

- Si la variation **intra-individuelle** est élevée → un modèle à effets fixes sera pertinent pour capturer les dynamiques temporelles propres à chaque individu.
- Si la variation est surtout **inter-individuelle** et que l'hypothèse d'absence de corrélation tient, un modèle à effets aléatoires pourra être plus efficace.

R MARKDOWN AND REPORTING

- **Qu'est-ce que R Markdown ?**

Un fichier .Rmd combine texte, code R, tableaux et graphiques dans un document unique.

- Un format de fichier utilisé pour créer des documents dynamiques en intégrant du code R, des textes en markdown, des visualisations et plus encore.

Permet de générer des documents au format HTML, PDF, Word, et autres.

- **Pourquoi utiliser R Markdown ?**

- Combiner texte, code, et visualisation dans un même document.
- Reproductibilité des analyses.
- Facilité de partage et de présentation des résultats.
- Possibilité de mettre à jour les rapports automatiquement via scripts.

STRUCTURE

&

CREATION

- **Header (En-tête YAML)**

- Définition du titre, de l'auteur, du format de sortie, etc.

- **Texte en Markdown**

- Sections, listes, images, liens, etc.

- **Champs de code (Code Chunks)**

- Sections de code R (ou autres langages) entourées par ```{r}`.

- **Résultats**

- Affichage des résultats du code, comme les tables ou graphiques.

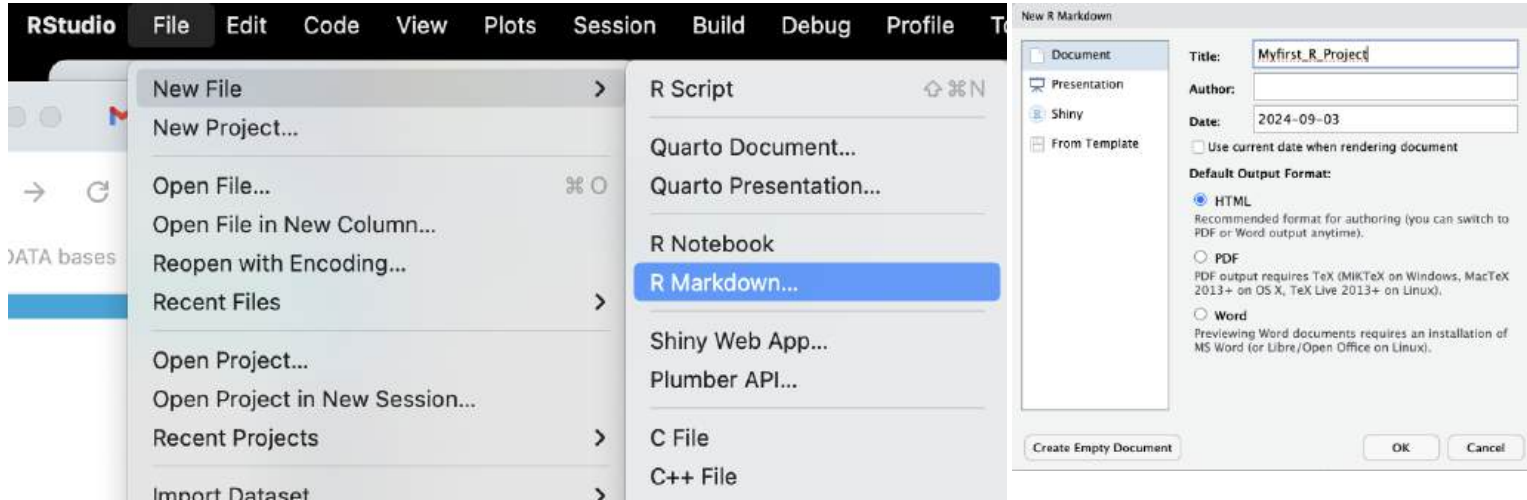
- **1. Créer un fichier .Rmd** En utilisant RStudio ou un éditeur de texte compatible.

- **2. Écrire le texte et le code** Combinez vos analyses et explications dans le même fichier.

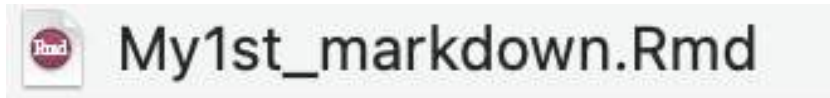
- **3. Knit le document** Utilisez le bouton "Knit" dans RStudio pour générer la sortie (HTML, PDF, Word, etc.).

- **4. Révision et partage** Relisez et partagez le document final.

UN R MARKDOWN POUR EX03



Ouvrez:



Create PDF and Word files:

<https://www.youtube.com/watch?v=vWGGIMHkhyY>

LES OUTILS D'IA UTILES POUR LA RECHERCHE

1. Recherche bibliographique et revue de littérature

- *Perplexity* : synthèse rapide d'articles et de concepts
- *Elicit* : aide à trouver et résumer des papiers scientifiques
- *Connected Papers*, *ResearchRabbit* : exploration de réseaux de publications
- *NotebookLM* (Google) : analyse et synthèse de documents longs
- *Anara* : lecture critique, organisation et analyse de la littérature scientifique

• 2. Rédaction et édition scientifique

- QuillBot, Trinko,...

• 3. Prompt engineering: <https://academy.openai.com/home>

SOURCES:

AN INTRODUCTION TO R
COMPLETE INTRODUCTION TO BASE R. MY RECOMMENDED PLACE TO GET STARTED.

[HTTPS://CRAN.R-PROJECT.ORG/DOC/MANUALS/R-RELEASE/R-INTRO.PDF](https://cran.r-project.org/doc/manuals/r-release/r-intro.pdf)

[HTTPS://CRAN.R-PROJECT.ORG/OTHER-DOCS.HTML](https://cran.r-project.org/other-docs.html)

[HTTPS://RCOMPANION.ORG/HANDBOOK/](https://rcompanion.org/handbook/)

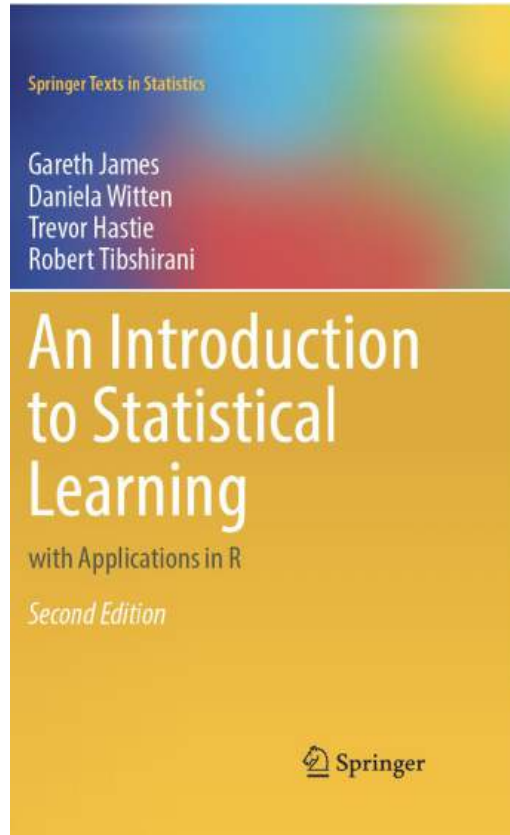
[HTTPS://IQSS.GITHUB.IO/DSS-WORKSHOPS/RINTRO.HTML](https://iqss.github.io/dss-workshops/rintro.html)

[HTTP://USERS.STAT.UMN.EDU/~GEYER/RC/](http://users.stat.umn.edu/~geyer/rc/)

[HTTPS://WWW.ECONOMETRICS-WITH-R.ORG/](https://www.econometrics-with-r.org/)

[HTTPS://WWW.PRINCETON.EDU/~OTORRES/](https://www.princeton.edu/~otorres/)

<https://www.statlearning.com/>



ADVANCED LEVEL

- **R for Data Science**

Introduction to data analysis using R, focused on the tidyverse packages. If your goal is to find a substitute for Stata, start here.

<http://r4ds.had.co.nz/>

- **Advanced R**

In-depth discussion of programming in R. Read later, if you want to become a good R programmer.

<https://adv-r.hadley.nz/>

DATA VISUALISATION

- **Data Visualization - A Practical Introduction**
Textbook on data visualization, using ggplot2. <http://socviz.co/>
- **ggplot2 - Elegant Graphics for Data Analysis**
In depth discussion of R-package for data vizualization.
<http://moderngraphics11.pbworks.com/f/ggplot2-Book09hWickham.pdf>

A serene sunset scene over a calm ocean. The sun is a bright, glowing orb just above the horizon, casting a warm, orange and pink glow across the sky. The sky is filled with soft, wispy clouds. In the distance, a pier or breakwater structure is visible on the horizon. The water in the foreground shows gentle ripples and small waves.

MERCI POUR VOTRE ATTENTION!

QUESTIONS?

YEGANEH.FAR@GMAIL.COM