TP1 : Rappels

A. L. N'Guessan, V. Roca & W. Heyse

Le 15 Septembre 2025

1 RMarkdown

1.1 Introduction à RMarkdown

1.1.1 L'intérêt de RMarkdown

Pourquoi utiliser RMarkdown ?

- **Rapports reproductibles** : L'un des plus grands avantages de RMarkdown est qu'il permet de créer des documents reproductibles. Tout ce que vous faites dans R (analyses, visualisations, etc.) peut être documenté et les résultats peuvent être facilement partagés sous forme de rapports, sans avoir besoin de copier-coller des graphiques ou des sorties de code.
- **Intégration du code et du texte** : RMarkdown permet d'intégrer du code R directement dans le texte, ce qui facilite la combinaison des analyses et des commentaires.
- **Sortie flexible** : Avec RMarkdown, vous pouvez générer des documents dans différents formats (HTML, PDF, Word, etc.), ce qui vous donne la flexibilité de choisir la présentation qui convient le mieux à votre audience.
- **Facilité de mise à jour** : En cas de modification du code ou des données, il suffit de recompiler le document pour mettre à jour automatiquement le rapport avec les nouvelles analyses ou visualisations.

Application pratique : Imaginez que vous devez rendre un rapport de projet avec des analyses statistiques. Sans RMarkdown, vous devrez exporter chaque graphique, copier les résultats dans un document Word, puis les annoter. Avec RMarkdown, tout cela est intégré : vous écrivez votre code et votre texte ensemble, et le document final est généré d'un seul coup.

A titre d'exemple, le document que vous êtes en train de lire a été généré avec RMarkdown.

1.1.2 Généralités sur RMarkdown

1.1.2.1 Type de fichier

Les fichiers RMarkdown ont l'extension `.Rmd`. Ce sont des fichiers texte simples qui peuvent être édités dans n'importe quel éditeur de texte, mais ils sont généralement créés et modifiés dans RStudio.

1.1.2.2 Création d'un fichier RMarkdown

1. **Ouvrir RStudio**.
2. **Créer un nouveau fichier RMarkdown** :
   - Allez dans `File > New File > R Markdown...`.
   - Une fenêtre s'ouvre vous demandant d'entrer un titre pour le document, le nom de l'auteur, et de choisir le format de sortie (HTML, PDF, Word). Sélectionnez `HTML` ou `PDF` pour ce TP.
   - Cliquez sur `OK` pour créer le fichier. Un fichier pré-rempli, avec quelques exemples devrait s'ouvrir.
3. Sauvez votre fichier dans un dossier.

1.1.2.3 Compilation d'un fichier RMarkdown

Une fois votre fichier `.Rmd` créé, vous pouvez le compiler (ou "knit") pour générer le document final.

Cliquez sur le bouton **Knit** en haut de l'éditeur RMarkdown dans RStudio (Pelote de laine violette) ou alors avec le raccourci clavier `Ctrl+Shift+k` ou `Cmd+Shift+k`. Un fichier HTML (ou autre format choisi) sera créé et affiché dans le navigateur par défaut. Vous devez considérer le document RMarkdown comme un environnement à part c'est-à-dire que pour travailler sur des données, le document doit charger les données. Le fait d'avoir chargé les données dans votre environnement de travail n'est pas suffisant à ce que le document compile correctement.

Questions :

1. Créez un nouveau fichier RMarkdown dans RStudio avec votre nom comme auteur, `HTML` comme format de sortie et `Rapport TP1` comme titre.
2. Compilez le fichier en cliquant sur `Knit`. Que remarquez-vous dans le document `HTML` généré ?

1.1.3 Principales commandes en RMarkdown

1.1.3.1 Sections et titres

Les titres sont créés en utilisant le symbole `#`. Plus il y a de `#`, plus le niveau du titre est bas.

`Exemple :`

```
# Titre de niveau 1
## Titre de niveau 2
### Titre de niveau 3
```

1.1.3.2 Mise en forme du texte

- **Italique** : Utilisez `*italique*` ou `_italique_`.
- **Gras** : Utilisez `**gras**` ou `__gras__`.
- **Listes** :
  - Liste à puces : Utilisez `-` ou `*`.
  - Liste numérotée : Utilisez `1.`.

  Exemple :

  ```
  - Élément de liste à puces
  - Un autre élément

  1. Premier élément
  2. Deuxième élément
  ```

  Attention, l'espace entre l'élément de liste (`-` ou `1.`) et le texte est important !

1.1.3.3 Chunks de code

Un **chunk** de code est une section de votre fichier `.Rmd` où vous pouvez écrire du code R. Ils sont délimités comme suit, avec des options disponibles pour en contrôler le comportement (cf. section suivante).

Exemple :

```
` ` ` {r nom_du_chunk}
dim(mtcars) #Code quelconque qui sera exécuté et dont le résultat sera dans le document compilé
` ` `
```

Attention, ` sont des backticks (touche 7 du clavier) **les espaces entre les backticks** qui apparaissent ci-dessus **ne doivent pas être présents**.

Le `nom_du_chunk` n'est pas obligatoire mais permet de vite repérer les erreurs le cas échéant. Chaque chunk doit avoir un nom unique.

Vous pouvez créer un nouveau chunk de code au clic-bouton en cliquant sur l'icône "+c" verte tout à droite du bouton "knit" ou avec le raccourci `Ctrl+Shift+i` ou `Cmd+Option+i`.

1.1.3.4 Options des Chunks de Code

Dans RMarkdown, chaque chunk de code peut être configuré avec des options spécifiques pour personnaliser l'affichage du code et des résultats.

Voici quelques-unes des options les plus couramment utilisées.

1. **Afficher ou cacher le code :**
   - `echo = TRUE` **(par défaut)** : Affiche le code R dans le document final.
   - `echo = FALSE` : Masque le code R dans le document, mais affiche les résultats du code.

Exemple :

```
` ` ` {r echo=FALSE}
  summary(mtcars)
` ` `
```

2. **Afficher ou cacher les résultats :**
   - `results = 'markup'` **(par défaut)** : Affiche les résultats du code de manière standard.
   - `results = 'hide'` : Masque les résultats du code.

   Exemple :

```
  ` ` ` {r results='hide'}
    summary(mtcars)
  ` ` `
```

3. **Afficher ou cacher les messages et les avertissements :**
   - `message = TRUE` **(par défaut)** : Affiche les messages générés par le code R.
   - `message = FALSE` : Masque les messages.
   - `warning = TRUE` **(par défaut)** : Affiche les avertissements générés par le code R.
   - `warning = FALSE` : Masque les avertissements.

   Exemple :

```
      ` ` ` {r message=FALSE, warning=FALSE}
        library(ggplot2)  # Ce chunk ne montrera pas les messages ou avertissements
      ` ` `
```

4. **Afficher ou cacher les graphiques :**
   - `fig.show = 'asis'` **(par défaut)** : Affiche les graphiques produits par le code.
   - `fig.show = 'hide'` : Masque les graphiques.

   Exemple :

```
      ` ` ` {r fig.show='hide'}
        plot(mtcars$wt, mtcars$mpg) #La figure ne sera pas montrée dans le document final
      ` ` `
```

5. Ajuster la taille des graphiques :

- `fig.width` et `fig.height` : Permettent de définir la largeur et la hauteur des graphiques.

  Exemple :

  ```
  ```{r fig.width=6, fig.height=4}
  plot(mtcars$wt, mtcars$mpg)
  ```
  ```

6. Exécution du code :

- `eval = TRUE` **(par défaut)** : Exécute le code dans le chunk.
- `eval = FALSE` : Empêche l'exécution du code, affichant uniquement le code R sans résultats.

  Exemple :

  ```
  ```{r eval=FALSE}
  summary(mtcars) # Ce code ne sera pas exécuté
  ```
  ```

1.1.3.5 Code Inline

Le code *inline* permet d'insérer du code R dans une phrase ou un paragraphe de texte comme suit :

Exemple :

```
La moyenne de mpg est `r mean(mtcars$mpg)`.
```

1.1.4 Fiche récapitulative

En plus des éléments de base décrits ci-dessus, RMarkdown comporte énormément d'autres options et de méthodes de mise en page différentes. Les plus simples sont résumées sur cette fiche, vos recherches personelles vous permettrons d'en découvrir bien d'autres encore !

Questions :

1. Le jeu de données `mtcars` est déjà chargé dans `R`. Créez une première partie intitulée "Analyse descriptive" avec une titre de premier niveau.
2. Rédigez une très courte présentation de ce jeu de données (vous pourrez trouver des informations sur ce jeu de données avec `?mtcars`) dans laquelle vous inclurez, à l'aide de code *inline* le nombre de lignes et de colonnes de ce jeu de données.
3. Ajoutez un chunk de code qui génère un graphique de la consommation de carburant en fonction de la puissance des voitures. Faites en sorte que le graphique soit visible mais pas le code le générant.
4. Expérimentez avec l'option `fig.width` et `fig.height` en ajustant la taille d'un graphique pour qu'il soit plus large que par défaut.
5. Affichez un chunk pour montrer un exemple de code permettant déterminer le nombre de modèles de voitures ayant 8 cylindres sans l'exécuter.

Pour la suite du TP, structurez votre document avec des parties, des listes, ...

2 Analyse des corrélations linéaires

2.1 Rappels de cours

Utilité de la Corrélation Linéaire : La corrélation linéaire permet de mesurer la force et la direction de la relation linéaire entre deux variables quantitatives. C'est un outil statistique essentiel pour comprendre comment une variable peut influencer une autre. La méthode de calcul de la corrélation linéaire la plus utilisée est la méthode de Pearson.

Formule du Coefficient de Corrélation de Pearson : Le coefficient de corrélation de Pearson, noté r, se calcule à l'aide de la formule suivante :

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Où :

- x_i et y_i sont les valeurs des variables X et Y.
- $\bar{x}$ et $\bar{y}$ sont les moyennes des variables X et Y.
- n est le nombre d'observations.

Interprétation du Coefficient de Corrélation :

- $r = 1$: Corrélation positive parfaite (les variables augmentent ensemble).
- $r = -1$: Corrélation négative parfaite (une variable augmente tandis que l'autre diminue).
- $r = 0$: Aucune corrélation linéaire.
- **Attention** : La corrélation mesure uniquement la relation linéaire. Une corrélation de zéro n'indique pas nécessairement l'absence de toute relation entre les variables.

Test de Corrélation de Pearson

Le test de corrélation de Pearson permet de déterminer si la corrélation observée entre deux variables quantitatives est statistiquement significative.

Hypothèses du test :

- Hypothèse nulle H_0 : Il n'y a pas de corrélation linéaire entre les deux variables ($r = 0$).
- Hypothèse alternative H_1 : Il existe une corrélation linéaire entre les deux variables ($r \neq 0$).

Interprétation des résultats :

- Coefficient de corrélation r : Le coefficient calculé comme précédemment.
- p-value : Si la p-value est inférieure à un seuil α (généralement 0,05), on rejette l'hypothèse nulle, indiquant que la corrélation est statistiquement significative.
- Intervalle de confiance : L'intervalle de confiance pour le coefficient de corrélation indique la précision de l'estimation de r.

2.2 Questions

2.2.1 Calcul des corrélations

1. Calculer la corrélation entre `mpg` (consommation) et `wt` (poids).
2. Interprétez le résultat obtenu. Quelle est la nature de la relation entre le poids des voitures et leur consommation ?
3. A l'aide de la fonction `cor.test`, déterminez si la corrélation de ces deux variables est statistiquement significative au seuil de 5%.

2.2.2 Calcul des matrices de corrélations

1. A l'aide de la fonction `cor`, calculez la matrice de corrélation de l'ensemble des variables du jeu de données. Les résultats vous semblent-ils facilement lisibles ?
2. Avec la fonction `corrplot`, disponible dans le package éponyme, tracez un graphe de cette matrice de corrélation. Commentez les corrélations.
3. On propose le code suivant, que permet de faire ce code ?

```
library(psych)
cor.matrix <- corr.test(mtcars)
corrplot(cor.matrix$r, p.mat = cor.matrix$p, insig = "p-value", order = "AOE")
```

3 Régression Linéaire

3.1 Rappels de cours

La régression linéaire est une méthode statistique utilisée pour modéliser la relation entre une variable dépendante Y (variable à expliquer) et une ou plusieurs variables indépendantes $X_1, X_2, \ldots, X_n$ (variables explicatives). Son objectif principal est de prédire la valeur de Y à partir des valeurs de X et de comprendre comment Y varie en fonction des X.

3.1.1 Régression Linéaire

Dans le cas le plus simple, où l'on cherche à expliquer Y par une seule variable X, la régression linéaire se représente par la formule suivante :

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Où :

- Y est la variable dépendante.
- X est la variable indépendante.
- β_0 est l'ordonnée à l'origine (intercept) : la valeur prédite de Y lorsque $X = 0$.
- β_1 est le coefficient de régression : il représente l'effet moyen d'une unité de variation de X sur Y.
- ϵ est le terme d'erreur aléatoire qui capte les variations de Y non expliquées par X.

Formule pour plusieurs variables explicatives :

Dans le cas d'une régression linéaire multiple, la formule devient :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

3.1.2 Indicateurs de Qualité du Modèle

Coefficient de détermination R^2 :

Le R^2 mesure la proportion de la variance totale de la variable dépendante Y expliquée par le modèle. C'est un indicateur de la qualité de l'ajustement du modèle.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Où :

- y_i est la valeur observée de la variable dépendante.
- $\hat{y}_i$ est la valeur prédite par le modèle.
- $\bar{y}$ est la moyenne des valeurs observées de Y.
- n est le nombre d'observations.

Il varie entre 0 et 1 et s'interprète de la manière suivante :

- $R^2 = 0$ signifie que le modèle n'explique aucune variance de Y.
- $R^2 = 1$ signifie que le modèle explique toute la variance de Y. Un R^2 élevé signifie que le modèle est performant pour expliquer les variations de Y.

Coefficient de Détermination Ajusté $R^2_{ajusté}$:

Le $R^2_{ajusté}$ est une version ajustée de R^2 qui prend en compte le nombre de prédicteurs dans le modèle. Il est particulièrement utile dans la régression linéaire multiple car il pénalise l'ajout de variables non pertinentes.

$$R^2_{ajusté} = 1 - \left(\frac{(1 - R^2)(n-1)}{n - p - 1} \right)$$

Où :

- n est le nombre d'observations.
- p est le nombre de variables explicatives dans le modèle.
- R^2 est le coefficient de détermination non ajusté.

Interprétation :

- $R^2_{ajusté}$ est toujours inférieur ou égal à R^2.
- Un $R^2_{ajusté}$ élevé signifie que le modèle explique bien la variance de Y tout en pénalisant la complexité inutile (c'est-à-dire l'ajout de variables explicatives qui n'améliorent pas substantiellement le modèle).

3.1.3 Tests sur les Coefficients du Modèle

Dans le contexte de la régression linéaire, il est essentiel de tester si les coefficients $\beta_1, \beta_2, \ldots, \beta_n$ sont significativement différents de zéro, ce qui indiquerait qu'il existe une relation statistiquement significative entre les variables explicatives et la variable dépendante.

Test de significativité des coefficients :

- Hypothèse nulle H_0: $\beta_i = 0$ (le coefficient n'a pas d'effet sur Y).
- Hypothèse alternative H_1: $\beta_i \neq 0$ (le coefficient a un effet sur Y).

3.1.4 Test de Significativité Globale du Modèle

Le test de significativité globale du modèle (test de Fisher) permet de déterminer si le modèle de régression linéaire, dans son ensemble, explique une proportion significative de la variance de la variable dépendante Y. En d'autres termes, il teste si au moins un des coefficients de régression β_i (autre que l'ordonnée à l'origine β_0) est significativement différent de zéro.

Hypothèses du test :

- Hypothèse nulle H_0 : Tous les coefficients de régression sont égaux à zéro ($\beta_1 = \beta_2 = \cdots = \beta_n = 0$), ce qui signifie que le modèle n'a pas de pouvoir explicatif.
- Hypothèse alternative H_1 : Au moins un des coefficients β_i est différent de zéro ($\exists i : \beta_i \neq 0$), indiquant que le modèle a un pouvoir explicatif.

La statistique F est calculée à partir des carrés moyens de la régression et des carrés moyens des erreurs :

$$F = \frac{MSC}{MSE} = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}}$$

Où :

- MSC (Mean Square of the Regression) est le carré moyen de la régression.
- MSE (Mean Square of the Error) est le carré moyen des erreurs.
- SSR (Sum of Squares of the Regression) est la somme des carrés expliqués par le modèle.
- SSE (Sum of Squares of the Error) est la somme des carrés des résidus (erreurs).
- p est le nombre de variables explicatives dans le modèle.
- n est le nombre total d'observations.

3.1.5 Limites de la Régression Linéaire

1. Hypothèses fortes : La régression linéaire repose sur plusieurs hypothèses, notamment la linéarité, l'indépendance des erreurs, l'homoscédasticité (variance constante des erreurs), et la normalité des résidus. Si ces hypothèses ne sont pas respectées, les résultats peuvent être biaisés.

2. Sensibilité aux outliers : Les points de données aberrants (outliers) peuvent avoir un effet disproportionné sur la ligne de régression et altérer les résultats.

3. Multicolinéarité : Comme mentionné précédemment, la multicolinéarité peut rendre l'interprétation des coefficients difficile.

4. Extrapolation risquée : La régression linéaire permet de faire des prédictions dans la plage des données observées, mais extrapoler au-delà de cette plage peut mener à des prédictions erronées.

3.2 Questions

1. Ajustement d'un Modèle de Régression Linéaire :

- **a)** Ajustez un modèle de régression linéaire multiple pour prédire la consommation de carburant (`mpg`) à partir du poids (`wt`), de la puissance (`hp`), et du rapport de pont (`drat`).
- **b)** Donnez l'équation du modèle ajusté. Quelle est la signification des coefficients associés à `wt`, `hp`, et `drat` ?

2. Qualité du Modèle :

- **a)** Calculez le coefficient de détermination R^2 et le coefficient de détermination ajusté $R^2_{ajusté}$ pour ce modèle. Comment interprétez-vous ces valeurs ?
- **b)** Le modèle explique-t-il bien la variance de la consommation de carburant ? Justifiez votre réponse en vous basant sur les valeurs de R^2 et $R^2_{ajusté}$.

3. Significativité des Coefficients :

- **a)** Pour chaque coefficient de régression (c'est-à-dire les variables `wt`, `hp`, et `drat`), testez l'hypothèse nulle selon laquelle le coefficient est égal à zéro. Quelles sont les conclusions de ces tests ?

Indications : Observez les p-values associées aux coefficients dans la sortie du résumé du modèle.

- **b)** Lequel de ces prédicteurs semble avoir l'impact le plus significatif sur la consommation de carburant ? Comment le savez-vous ?

4. Test de Significativité Globale du Modèle :

- **a)** Examinez la statistique F et la p-value associée pour le test de significativité globale du modèle. Le modèle est-il pertinant ?
- **b)** Que peut-on conclure sur la capacité globale du modèle à prédire la consommation de carburant ?

5. Limites du Modèle de Régression Linéaire :

- **a)** Quelles autres variables du jeu de données `mtcars` pourriez-vous envisager d'ajouter au modèle pour améliorer sa performance ?
- **b)** Tracez les résidus du modèle. Ces résidus vérifient-ils les hypothèses du modèle ? On pourra vérifier leur normalité à l'aide du test de Shapiro-Wilk.