Modèle de régression linéaire

Zineb Ziad Elsa Catteau Safia Zaari Jabri

Sommaire

Sommaire

Introduction

Le principe est donc qu'à partir d'une série d'observations (les points bleus), on souhaite déterminer la droite (la courbe rouge) qui passe au plus près des points.

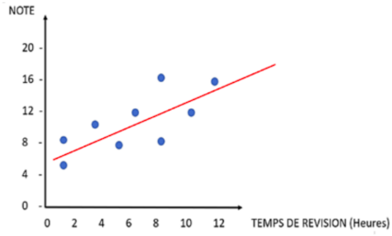


Figure: Exemple tiré du site ipgrade.com

La régression linéaire simple ou multiple a pour but de modéliser la relation linéaire entre une ou des variables explicatives et une variable à expliquer.

Sommaire

Régression linéaire simple

Le modèle de régression linéaire simple s'écrit :

$$Y_i = a\,x_i + b + \varepsilon_i$$

On cherche à savoir s'il existe une relation fonctionnelle entre la variable explicative x_i et la variable réponse y_i

On considère alors le risque empirique défini par :

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2$$

On va donc vouloir minimiser ce risque pour des fonctions du type g du type $g(x) = ax + b$.

Régression linéaire simple : Bruit

Avant de se lancer dans une régression linéaire, il est important de savoir si celle-ci sera pertinente. Pour cela, il faut d'abord vérifier si le bruit est normalement distribué. On examine donc si les résidus suivent une loi normale.

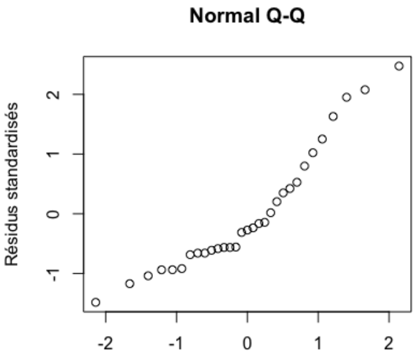
$$\hat{\varepsilon}_{i,\mathrm{sd}} = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_n \sqrt{1 - h_i}}$$

où h_i est le i-ème terme diagonal de la matrice de projection, et où la matrice des observations $\mathbf{X}$ est :

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

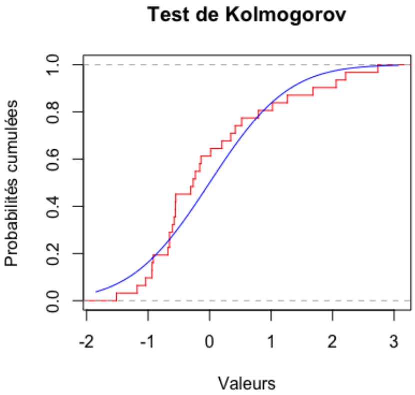
Régression linéaire simple: Bruit

A l'aide d'un graphique QQ (quantiles-quantiles) on compare les quantiles théoriques d'une loi normale avec les quantiles empiriques (observés) de nos données.



Régression linéaire simple: Bruit

On peut aussi effectuer un test d'adéquation de Kolmogorov afin de verifier l'hypothèse de gaussianité sur les résidus studentisés qui suivent une loi de Student à (n-3) degrés de liberté



Régression linéaire simple: Bruit

```
        Exact one-sample Kolmogorov-Smirnov test

data:  residus_st
D = 0.16832, p-value = 0.3079
alternative hypothesis: two-sided
```

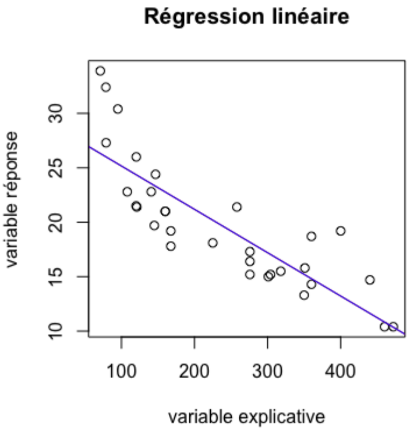
Ici, comme notre p-valeur est assez grande, au dessus de 0,05, on considère bien qur les observations peuvent être modélisées par la loi théorique considérée. Maintenant que l'on considère bien que le bruit est gaussien il est donc possible dans le cas de nos observations, d'effectuer une régression lineaire simple.

Principe des moindres carrés correspond à la minimisation du risque empirique permet donc de déterminer les estimateurs $\hat{a}_n$ de a et $\hat{b}_n$ de b.

$$\boxed{\hat{a}_n = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{Y}_n \bar{x}_n}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$$

$$\boxed{\hat{b}_n = \bar{Y}_n - \hat{a}_n \, \bar{x}_n}$$

Régression linéaire simple: Visualisation



Régression linéaire simple

```
Call:
lm(formula = Yi ~ xi)

Residuals:
   Min      1Q Median      3Q     Max
-4.671 -2.065 -0.862  1.473  7.586

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.145800   1.268635  22.974  < 2e-16 ***
xi          -0.039825   0.004791  -8.313 3.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.219 on 29 degrees of freedom
Multiple R-squared:  0.7044,    Adjusted R-squared:  0.6942
F-statistic: 69.11 on 1 and 29 DF,  p-value: 3.652e-09
```

Figure: Informations graphique régression linéaire simple

Maintenant vérifions que notre régression linéaire est bien valide.

Régression linéaire simple : Coefficient de détermination

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

Plus R^2 est proche de 1, plus la régression linéaire est une bonne modélisation. Ici nous avons un R^2=0,7044.

Régression linéaire simple : Test du paramètre a

Test d'hypothèse pour le paramètre a :

$$H_0 : a = 0 \quad \text{vs} \quad H_1 : a \neq 0$$

Sous H_0, la statistique de test est :

$$T_a = \frac{\hat{a}_n}{\hat{\sigma}_n \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$$

p-valeur $= \mathrm{P}_{H_0}\left(|T| > |T_{a,\text{obs}}|\right)$

Décision basée sur la p-valeur :

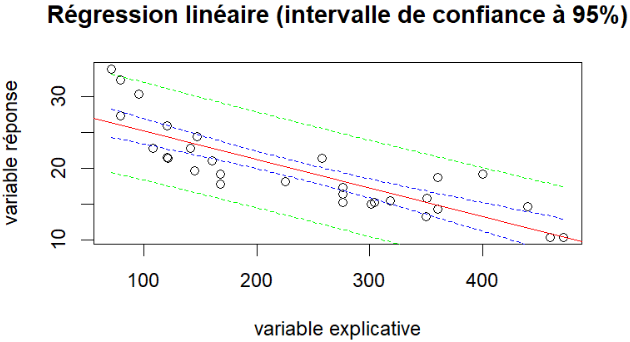
- Si la p-valeur $< \alpha$, alors on décide H_1.
- Si la p-valeur $> \alpha$, alors on ne rejette pas H_0.

On pose α=0,05 et on obtient une p-valeur égale à 1.825833e-09 donc on décide H1

Régression linéaire simple: intervalle de confiance

Si on note x_{new} une nouvelle observation de la variable explicative, une prévision de la variable réponse est donnée par : $\hat{y}_{\text{new}} = \hat{a}_n \cdot x_{\text{new}} + \hat{b}_n$.

Cependant, les estimateurs $\hat{a}_n$ et $\hat{b}_n$ dépendent du jeu de données d'apprentissage. Pour en tenir compte, on affiche l'intervalle de confiance (en vert) .



Sommaire

Régression linéaire multiple

Principe :

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_p x_i^{(p)} + \varepsilon_i$$

- Dans cette écriture, les x_i ne sont pas des quantités aléatoires (fixées à l'avance).
- Les seules quantités aléatoires sont les Y_i et les ε_i.
- Si l'on suppose l'hypothèse de gaussianité du bruit, les variables ε_i sont des variables aléatoires indépendantes et de même loi.

Vérification de l'hypothèse de gaussianité du bruit

On vérifie l'hypothèse de gaussianité du bruit de la même manière que pour la régression linéaire simple.
Pour les résidus standardisés, on utilise la formule suivante :

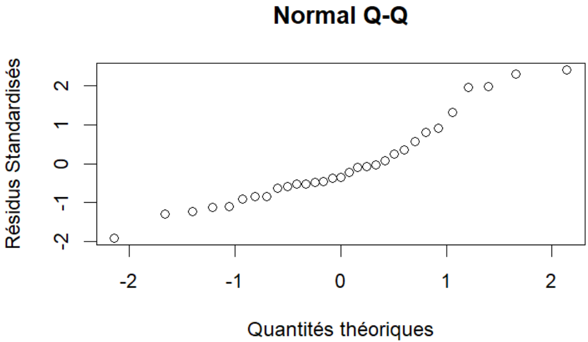
$$\hat{\varepsilon}_{i,\mathrm{sd}} = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}_n \cdot \sqrt{1 - h_i}}$$

avec h_i qui est le i-ème terme diagonal de la matrice $\mathbb{X} \cdot (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, en vérifiant au préalable que $\mathbb{X}'\mathbb{X}$ est bien inversible.
On a alors :

$$\hat{\varepsilon}_{i,\mathrm{sd}} \xrightarrow{(\mathcal{L})} \mathcal{N}(0, 1)$$

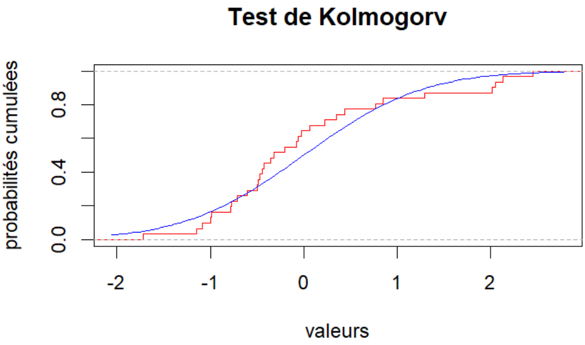
Vérification de l'hypothèse de gaussianité du bruit

Afin de valider l'hypothèse de gaussianité du bruit, on vérifie que les résidus standardisés suivent approximativement une loi normale standard à l'aide d'un graphe `qqnorm`.



Vérification de l'hypothèse de gaussianité du bruit

Pour obtenir les résidus studentisés, nous avons alors utilisé la fonction `rstudent`. Puis, on a vérifié l'hypothèse de gaussianité du bruit en vérifiant que les résidus suivent une loi de Student à $n - \text{rang}(\mathbb{X}) - 1$ degrés de liberté, en réalisant un test d'adéquation de Kolmogorov.



Test de significativité globale du modèle

On souhaite tester l'utilité globale du modèle de régression multiple :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{ou} \quad H_1 : \text{au moins un } \beta_j \neq 0$$

$$\mathbf{Y} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \mathcal{U}$$

La statistique de test est :

$$F = \frac{\|\hat{Y} - \bar{Y}_n\|^2/(r-1)}{\|Y - \bar{Y}_n\|^2/(n-r)} \quad \overset{H_0}{\sim} \quad \mathcal{F}(r-1, n-r)$$

Test de significativité globale du modèle

Pour cela on calcule la p-valeur de F

- Si la p-valeur $< \alpha$, alors on décide H_1.
- Si la p-valeur $> \alpha$, alors on ne rejette pas H_0.

Dans notre cas on obtiens une p-valeur égale à 1.346396e-07, ainsi notre modèle est bien significatif.

Intervalle de confiance

Si on note x_{new} une nouvelle observation de la variable explicative, une prévision de la variable réponse est donnée par :

$$\hat{y}_{\text{new}} = \begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix} \hat{\beta}_n.$$

On a choisi de prendre comme x_{new}, la ligne qu'on avait enlevée avec la commande `A[-set1]`, pour que cela corresponde bien à une nouvelle observation.

Intervalle de confiance

Un intervalle de confiance pour la prédiction de Y pour une nouvelle valeur x_{new} de la variable explicative, de niveau de confiance $100(1-\alpha)\%$, est :

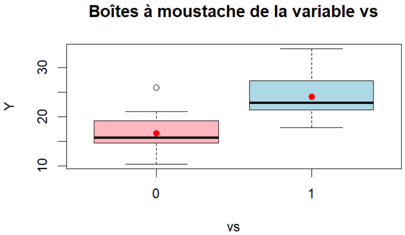
$$\begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix} \hat{\beta}_n \pm \hat{\sigma}_n \sqrt{\begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix} (\mathbb{X}'\mathbb{X})^{-1} \begin{pmatrix} 1 & x_{\text{new}} \end{pmatrix}'} \cdot t_{1-\alpha/2, n-\text{rang}(X)},$$

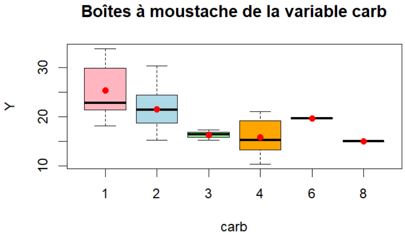
Au préalable, on a vérifié que $\mathbb{X}'\mathbb{X}$ est inversible. Notre programme affiche une valeur pour la variable réponse qui est de 27.70598 avec un intervalle de confiance à 95% : $[23.72286, 31.68909]$.

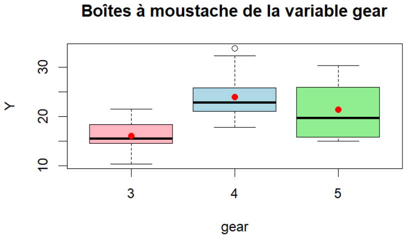
Analyse de la variance à un facteur

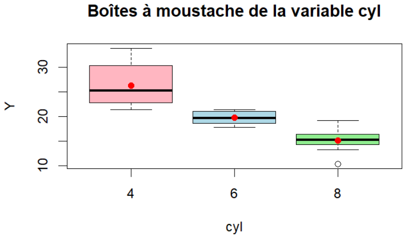
On effectue une représentation graphique pour chacune des variables qualitatives qui fait intervenir des boîtes à moustaches. Dans notre jeu de données, nous avons 4 variables qualitatives : `vs`, `carb`, `gear` et `cyl`. De plus, il est intéressant de faire apparaître la moyenne sur les boîtes à moustaches afin de voir si le facteur à une influence sur la variable réponse.

Analyse de la variance à un facteur









Le cadre mathématique du modèle d'analyse de la variance à un facteur est le suivant :

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \mu + \sum_{k=1}^{I} \mu_k \cdot \mathbf{1}_{\{x_i = a_k\}} + \epsilon_i$$

avec :

- $a_1, \dots, a_I$: les différentes modalités de la variable explicative
- μ : l'effet commun
- μ_k : effet spécifique de la modalité a_k
- $\forall i \in \{1, \dots, n\}, \quad \mathbb{E}[\epsilon_i] = 0$
- $\forall i \in \{1, \dots, n\}, \quad \mathbb{V}[\epsilon_i] = \sigma^2$
- $\forall i, k \in \{1, \dots, n\}, i \neq k, \quad \mathrm{cov}(\epsilon_i, \epsilon_k) = 0$

Analyse de la variance à un facteur

Pour mener l'analyse de la variance à un facteur, puisque $\mathbb{X}'\mathbb{X}$ n'est pas une matrice inversible, il faut donc extraire une base au sein de la famille des colonnes de X.
Il y a deux choix privilégier :

- On supprime la première colonne de $\mathbb{X}$ ce qui revient à faire l'hypothèse $\mu = 0$ (choix mathématique)
- On supprime la seconde colonne de $\mathbb{X}$ ce qui revient à faire l'hypothèse $\mu_1 = 0$ (choix pratique)

Pour tester l'influence ou non du facteur, on va procéder à un test.

Analyse de la variance à un facteur

Cas $\mu = 0$: Les hypothèses du test sont :

- $\mathcal{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_I$
- $\mathcal{H}_1$: ce n'est pas le cas

Cas $\mu_1 = 0$: Les hypothèses du test sont :

- $\mathcal{H}_0 : \mu_2 = \ldots = \mu_I = 0$
- $\mathcal{H}_1$: ce n'est pas le cas

Ces deux tests sont identiques.

On utilise la commande `summary(aov(lm(`$Y \sim$ variable`)))` qui nous donne un tableau où il y a la donnée `Pr(>F)` qui correspond à la p-valeur. Ainsi:

- Si p-valeur < 0.05 : on rejette $\mathcal{H}_0$ et on retient $\mathcal{H}_1$
- Si p-valeur ≥ 0.05 : on ne rejette pas $\mathcal{H}_0$

En appliquant cette commande à nos quatres variables qualitatives, on obtient ces quatres tableaux.

Analyse de la variance à un facteur

```
          Df Sum Sq Mean Sq F value   Pr(>F)    
vs         1  423.6   423.6   20.72 8.79e-05 ***
Residuals 29  592.8    20.4                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) Variable vs

```
          Df Sum Sq Mean Sq F value  Pr(>F)   
carb       5  462.0   92.39   4.166 0.00685 **
Residuals 25  554.4   22.18                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(b) Variable carb

```
          Df Sum Sq Mean Sq F value   Pr(>F)    
gear       2  411.1  205.54   9.509 0.000706 ***
Residuals 28  605.3   21.62                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(c) Variable gear

```
          Df Sum Sq Mean Sq F value   Pr(>F)    
cyl        2  730.4   365.2   35.77 1.94e-08 ***
Residuals 28  285.9    10.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) Variable cyl

Pour chaque variable on a p-valeur $< 0{,}05$ donc on rejette $\mathcal{H}_0$.

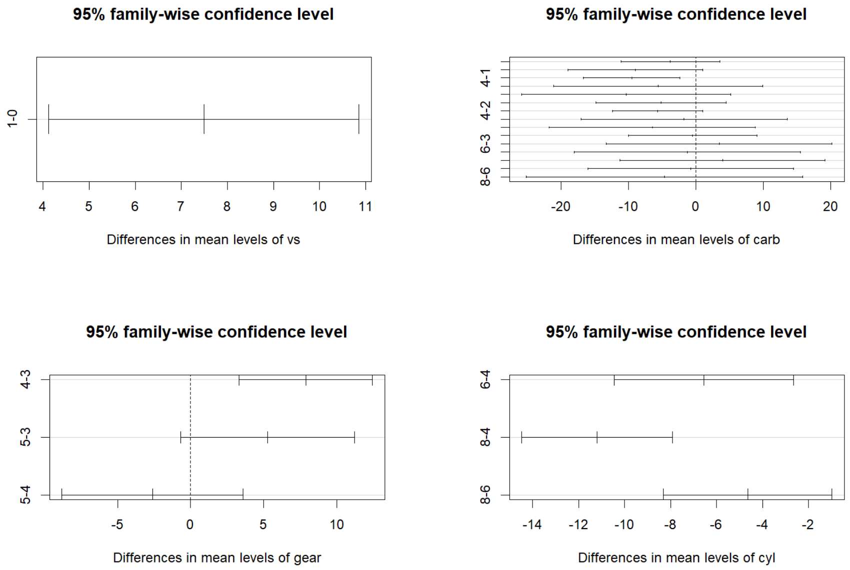
Analyse de la variance à un facteur

Ensuite, on utilise le test TukeyHSD qui permet d'identifier quelles paires de groupes diffèrent significativement.
Le résultat contient pour chaque paire :

- la différence de moyennes entre les groupes,
- l'intervalle de confiance de cette différence,
- la p-valeur ajustée : si elle est $< 0{,}05$, la différence est significative

Quand on affiche le graphe correspondant on obtient les intervalles de confiance des différences de moyennes entre les groupes. Si l'intervalle de confiance ne contient pas 0, la différence est significative.

Analyse de la variance à un facteur

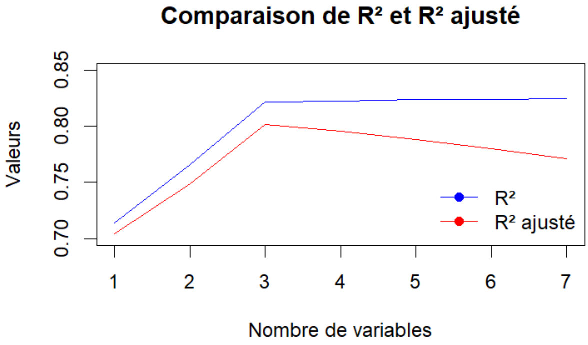


Sélection de variables

Il existe différentes stratégies pour la sélection de variables. Nous avons choisi de le faire en fonction du R_a^2, c'est-à-dire que nous effectuons une sélection de variables afin d'obtenir le meilleur R_a^2. Pour cela, nous avons utilisé la bibliothèque `leaps` du logiciel R et notamment la fonction `regsubsets` qui permet de tester toutes les combinaisons possibles et de trouver les meilleures sous-sélections pour modéliser notre variable Y.

Sélection de variables

Ainsi, on trace le graphe qui compare les valeur de R^2 et R_a^2 en fonction du nombre de variables. On obtient alors ce graphe :



On peut alors observer que le meilleur R_a^2 est obtenu avec 3 variables.

Sélection de variables

Pour connaître les 3 variables qui nous permettent d'obtenir le meilleur R_a^2, on peut utiliser la commande `summary(regsubsets(...))$which` qui est une matrice boléenne qui nous donne les variables sélectionnés.

	(Intercept)	vs1	carb	qsec	disp	drat	gear	cyl
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
2	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
3	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
4	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
5	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE
6	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Dans notre cas il faut regarder la ligne numéro 3 qui correspond au nombre de variable qui maximise R_a^2.

Sommaire

Conclusion

- La régression linéaire est un outil fondamental en statistique permettant de modéliser la relation entre une une variable réponse qui est quantitative, et une ou plusieurs variables explicatives elles aussi quantitatives.
- Dans le cas de la **régression linéaire simple**, il y a qu'une seule variable quantitative, le modèle est facile à interpréter, mais limité.
- Dans le cas de la **régression linéaire multiple** on utilise plusieurs variables quantitatives,, offrant une modélisation plus riche et réaliste.
- Dans les deux cas, il est essentiel de vérifier les hypothèses du modèle pour garantir la validité des résultats.

Merci pour votre attention !