UNIVERSITE DE PARAKOU

ECOLE NATIONALE DE STATISTIQUE, DE PLANIFICATION ET DE DEMOGRAPHIE (ENSPD/UP)

Département de Statistique Appliquée

METHODES DE REGRESSION PARAMETRIQUE

MASTER 1

07 02 2025

Dr. DICKO Aliou

Objectif général de l’ECU

La formation vise à initier les apprenants aux méthodes de régression paramétrique dans l’analyse statistique des données

Objectifs spécifiques de l’ECU

À la fin de ce cours, vous serez capable de :

- Comprendre les principes fondamentaux des méthodes de régression paramétrique.
- Savoir formuler un modèle de régression adapté à une problématique donnée.
- Identifier les hypothèses sous-jacentes et les conditions de validité des modèles.
- Interpréter les coefficients et les sorties des modèles paramétriques.
- Diagnostiquer et évaluer la performance des modèles.
- Appliquer les méthodes de régression paramétrique à des cas concrets à l'aide d'outil logiciel (R).

Prérequis

- Bonne connaissance en statistiques descriptive et inférentielle
- Bonne connaissance en informatique (notamment le logiciel Excel).

Contenu de la formation

1. Description
2. Définition de quelques concepts
3. Exemples d'application des méthodes de régression paramétrique
4. Hypothèses sur les modèles linéaires et les termes d'erreur
5. Procédure de test de régression
6. Relations linéaire et non linéaire
7. Modèle théorique de regression
8. Régression linéaire simple
9. Régression linéaire multiple
10. Limites des régressions linéaires
11. Extension des modèles linéaires aux modèles spécialisés:
    - Régression polynomiale
    - Régression logistique
    - Régression Poisson

Méthodes d’enseignement/apprentissage

- Cours théorique ;
- Travaux pratiques ;
- Travaux dirigés ;
- TPE (Etude de cas, exposé).

Lieu d’apprentissage

- Ecole/salle de cours ;
- Salle informatique.

Matériel pédagogique

- Projecteur ;
- Ordinateur ;
- Note de cours ;
- Tableau et craie.

Compétences générales

- Prendre des initiatives et décisions ;
- Croire que l’on peut surmonter tous les obstacles ;
- Mettre en œuvre des ressources organisationnelles pour produire des résultats ;
- Analyser les problèmes pour y trouver des solutions ;
- Apprendre.
- Ne pas remettre à plus tard ;
- Être une personne fiable et responsable ;
- Rédiger des rapports.

Mode d’évaluation

- Evaluation formative ;
- Devoir de table et oral.

1. Description

- La régression est l'un des outils les plus fondamentaux en analyse statistique et en science des données.

- Elle permet d'explorer, de modéliser et de prédire la relation entre une variable cible, appelée variable dépendante, et une ou plusieurs variables explicatives, appelées variables indépendantes.

- Parmi les approches disponibles, la régression **paramétrique** se distingue par sa simplicité, sa rigueur et sa large utilisation dans différents domaines.

- La régression paramétrique repose sur une hypothèse clé : la relation entre les variables peut être décrite par une fonction mathématique dont la forme est prédéfinie (linéaire, polynomiale, logarithmique, etc.).

1. Description

- Ce cours, intitulé **"Méthodes de régression paramétrique"**, a pour objectif de vous fournir une compréhension approfondie des techniques paramétriques de modélisation et d'interprétation des données.

- Il met l'accent sur les bases théoriques, les étapes pratiques et les applications concrètes des modèles de régression paramétrique, notamment dans des domaines tels que l'agriculture, l'économie, la santé, ou encore les sciences sociales.

2. Définition de quelques notions

2.1. Régression :

- C'est une technique statistique permettant de modéliser la relation entre une variable à expliquer y et une ou plusieurs variables explicatives X_1, X_2 , … , Xp.

- Y est encore appelé variable dépendante tandis que X_1, X_2 , … , Xp sont qualifiés de variables indépendantes.

- On distingue les régressions linéaires (simple et multiple) et les régressions non linéaires (simple et multiple).

2. Définition de quelques notions

2.2. Régressions linéaire et non linéaire :

- Une régression est dite linéaire lorsque la relation entre la variable dépendante et les variables explicatives est linéaire. Au cas contraire, elle est dite régression non linéaire.

2.3. Régressions simple et multiple :

- Lorsqu'il s'agit d'une seule variable explicative, on parle de régression simple.
- Par contre, lorsque la variable dépendante est expliquée par plusieurs variables indépendantes, on parle de régression multiple.
- Si de plus, la relation entre la variable dépendante et la (les) variable(s) explicative(s) est linéaire, on parle de régression linéaire simple ou multiple.

2. Définition de quelques notions

2.4. Régressions paramétrique et non paramétrique :

2.4.1. Régressions paramétrique

- **Définition** : Ce type de régression suppose que la relation entre les variables dépend d'une forme fonctionnelle prédéfinie (par exemple, linéaire, polynomial, exponentiel, etc.). On estime les paramètres de cette fonction à partir des données.

- **Hypothèses fortes** :
  - La forme du modèle est fixée (par exemple, $y=\beta_0+\beta_1 x+\varepsilon$ pour une régression linéaire simple).
  - Les erreurs résiduelles suivent souvent une distribution normale et indépendante.

- **Exemples** : Régression linéaire, régression logistique, régression polynomiale.

2. Définition de quelques notions

2.4.2. Régression non paramétrique :

- **Définition** : Ce type de régression ne suppose pas de forme fonctionnelle fixe pour la relation entre les variables. Il utilise les données pour estimer directement cette relation de manière flexible.

- **Hypothèses faibles** :
  - Pas besoin de spécifier à l'avance une forme fonctionnelle particulière.
  - Les seules hypothèses concernent généralement la régularité ou la continuité de la fonction à estimer.

- **Exemples** : Régression par k-plus proches voisins (k-NN), méthodes basées sur les splines, régression locale (loess), forêts aléatoires (random forest), modèles à noyau (kernel regression).

3. Thèmes d’exposés :

- Transformations de variables en régressions : Transformations logarithmique, carré et inverse
- Transformation Box-Cox et applications
- Effet des transformations sur l’hétéroscédasticité et la normalité des résidus
- Analyse des résidus : détection des erreurs et des anomalies
- Critères de sélection des modèles : AIC, BIC et application
- Test de multicolinéarité et solutions (VIF, PCA, Ridge)
- Régression polynomiale : applications et limites
- Régression logistique : fondements et applications
- Comparaison entre régression de Poisson et régression binomiale négative
- Comparaison entre régression de Poisson et régression quasi-poisson
- Régression quantile : une alternative à la régression linéaire
- Régression log-linéaire : applications en analyse des tableaux de contingence
- Régression spline : une approche flexible pour modéliser des relations non linéaires
- Impact des valeurs aberrantes sur les modèles de régression et stratégies de traitement
- Méthodes d'estimation des paramètres en régression

4. Exemple d'application de régression paramétrique:

- Prédire le rendement d'une culture en fonction des facteurs climatiques
- Déterminer la relation qui pourrait exister entre l'âge et la pression sanguine
- Expliquer le prix d'une voiture en fonction du poids de la voiture et de sa consommation en carburant.
- Un agronome peut chercher à prédire le rendement d'une culture en fonction de l'apport en engrais et de la pluviométrie,
- Un économiste peut vouloir estimer l'effet d'une politique publique sur le revenu des ménages.
- Étudier l'impact de la quantité d'engrais (X) sur le rendement agricole (Y)
- Analyser la relation entre le nombre d'heures d'étude (X) et la note obtenue à un examen (Y).
- Prédire le prix d'un bien immobilier (Y) en fonction de sa surface (X).

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.1. Hypothèses sur le modèle linéaire

- Le modèle de régression linéaire peut être exprimé comme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

où :

- y est la variable dépendante (ou cible),
- x_1, x_2,…, x_k sont les variables explicatives (ou prédicteurs),
- β_0, β_1,…,,β_k sont les coefficients à estimer,
- ε est le terme d'erreur.

Les hypothèses suivantes s'appliquent au modèle :

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.1.1. Relation linéaire entre y et les x_i

- On suppose que la relation entre la variable dépendante y et chaque variable explicative x_i est **linéaire**.

- Cela signifie que les prédicteurs x_i entrent dans le modèle de manière additive et proportionnelle.

- Si cette hypothèse est violée, des transformations des variables (e.g., logarithme, carré) ou des modèles non linéaires peuvent être nécessaires.

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.1.2. Les variables explicatives ne doivent pas être fortement corrélées (absence de multicolinéarité)

- La multicolinéarité, ou forte corrélation entre les variables explicatives, peut rendre difficile l'estimation précise des coefficients.
- Pour détecter la multicolinéarité, on utilise :
  - Le **facteur d'inflation de la variance (VIF)**.
  - Les **matrices de corrélation**.

5.1.3. Spécification correcte du modèle

- Toutes les variables pertinentes doivent être incluses, et les variables non pertinentes doivent être exclues.
- Une mauvaise spécification peut conduire à des estimations biaisées (biais d'omission ou de surajout).

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.2. Hypothèses sur les termes d'erreur (ε)

5.2.1. Moyenne des erreurs égale à zéro

$E(\varepsilon)=0$

Cette hypothèse garantit que le modèle est correctement centré, c'est-à-dire qu'en moyenne, les prédictions du modèle ne sont pas biaisées par les erreurs.

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.2.2. Homoscédasticité (variance constante des erreurs)

$$\text{Var}(\varepsilon) = \sigma^2 \text{ (constante pour tous les xi).}$$

- Les termes d'erreur doivent avoir une variance constante, quelle que soit la valeur des variables explicatives. En cas de **hétéroscédasticité**, les prédictions et les tests statistiques peuvent être biaisés.

- Détection :
  - Graphique des résidus vs valeurs prédites.
  - Tests statistiques : Test de Breusch-Pagan, test de White.

- Solutions en cas de violation : Transformation des variables (e.g., logarithme), utilisation de modèles robustes.

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.2.3. Indépendance des erreurs

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \text{ pour tout } i \pm j$$

- Les erreurs doivent être indépendantes les unes des autres.
- Cette hypothèse est souvent violée dans les données temporelles ou spatiales, où les observations successives peuvent être corrélées (**autocorrélation**).

- Détection :
  - Test de Durbin-Watson.
  - Analyse des résidus en fonction du temps.

- Solutions : Inclusion d'effets temporels ou spatiaux dans le modèle, ou utilisation de modèles adaptés (e.g., régressions avec erreurs auto-corrélées).

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.2.4. Normalité des erreurs

$$E \sim N(0,\sigma^2)$$

- Les erreurs doivent suivre une distribution normale pour que les tests d'hypothèse (comme le test t ou F) soient valides.

- Détection :
  - Histogramme ou diagramme de probabilité normale (Q-Q plot) des résidus.
  - Tests statistiques : Test de Shapiro-Wilk, test de Kolmogorov-Smirnov.

- Solutions en cas de violation : Transformation des variables ou utilisation de tests non paramétriques.

5. Hypothèses sur le modèle linéaire et termes d'erreur

5.3. Résumé des hypothèses (rappel mnémotechnique : LINER)

L : Linearity (relation linéaire entre y et les x_i).

I : Independence (indépendance des erreurs).

N : Normality (normalité des erreurs).

E : Equal variance (homoscédasticité des erreurs).

R : Right specification (bonne spécification du modèle).

- Ces hypothèses jouent un rôle crucial dans l'interprétation des coefficients et la validité des prédictions.
- Il est donc essentiel de les vérifier à l'aide d'outils graphiques et de tests avant d'utiliser les résultats du modèle.

6. Procédure de test de régression

- Avant de procéder à l'analyse d'un modèle de régression, il est important de vérifier que la variable indépendante et la variable dépendante présentent une relation significative.
- Cette relation peut être explorée à l'aide d'une analyse préliminaire (par exemple, un diagramme de dispersion ou une matrice de corrélation).

- Soit l'hypothèse nulle :

H_0 : « Il n'existe aucune relation entre la démographie et la demande en logements » (Dans ce cas, le coefficient de pente b=0).

6. Procédure de test de régression

- **Description des variables :**

- Variable indépendante : démographie (facteur explicatif).

- Variable dépendante : demande en logements (variable cible).

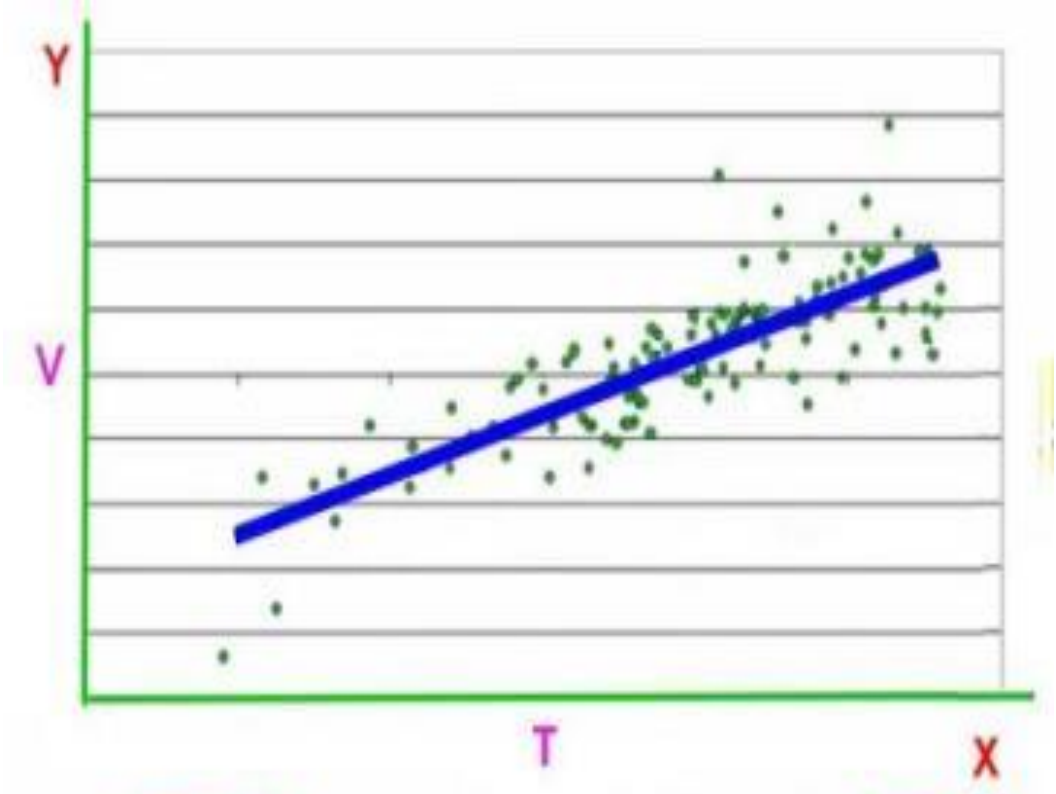
- **Décision selon les hypothèses :**

- Si **H_0 est rejetée**, cela signifie qu'il existe une relation statistiquement significative entre la démographie et la demande en logements. Dans ce cas, le coefficient de pente $b \neq 0$.

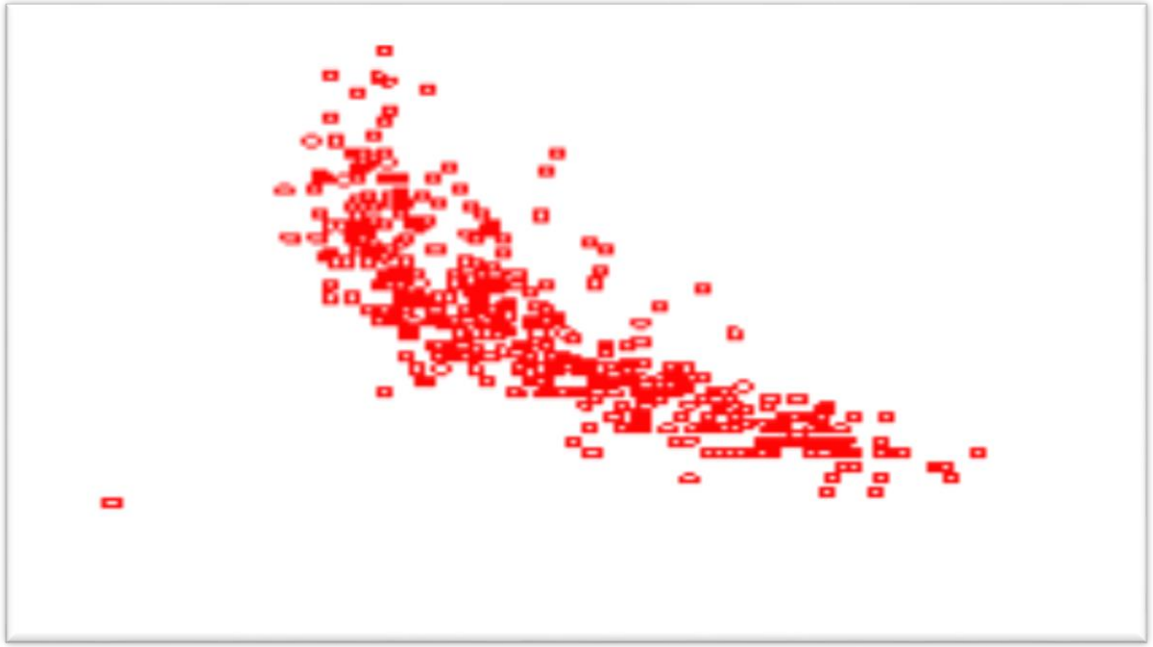
- Si **H0 est acceptée**, cela indique que la pente est égale à zéro ($b = 0$), et donc, il n'existe une relation linéaire entre les deux variables.

7. Relation linéaire et non linéaire

- **Deux exemples de relation linéaire**

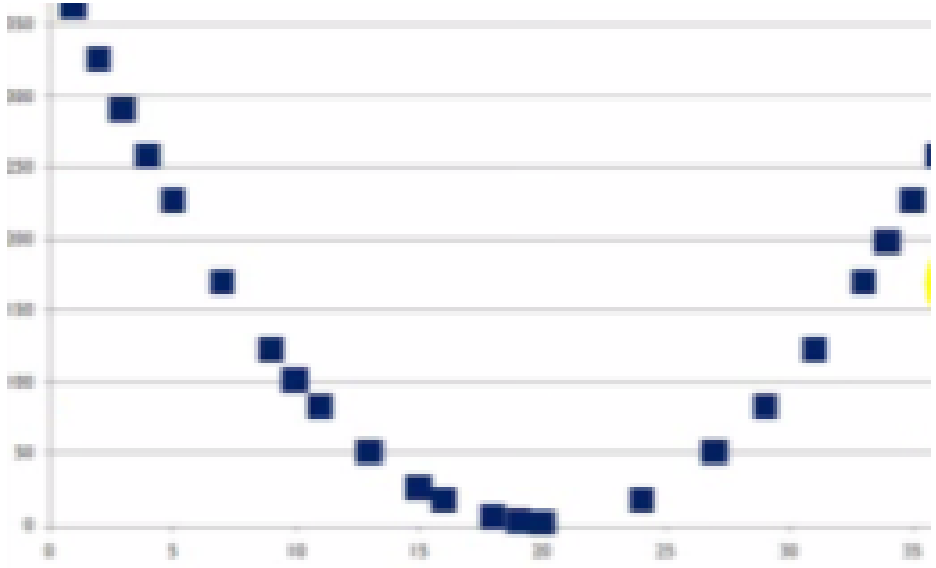


La vitesse (V) est en partie en relation avec le temps (T). Plus le temps nécessaire pour parcourir une distance est important plus la vitesse du véhicule sera grande.



7. Relation linéaire et non linéaire

- **Deux exemples de relation non linéaire**

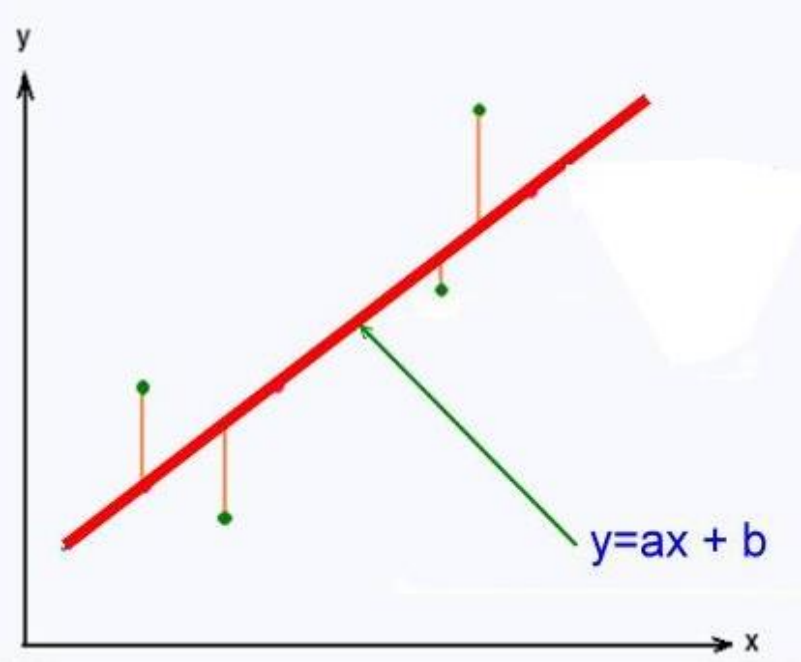


Ici, les observations (nuage de point en verte) suivent presque une ligne droite. La ligne bleue, qui exprime le meilleur ajustement des valeurs observées, est la régression. Cependant, cette droite de régression n'exprime pas parfaitement la position parfaite des différentes observations, il y a toujours une erreur (ε), car il existe une certaine distance entre les valeurs observées et les valeurs calculées qui constituent la ligne bleue de régression. Pour cela, il faut introduire ε dans l'équation y=ax+b.

y= ax + b + ε constitue uniquement une prédiction. D'où x (la variable indépendante) ne dépend pas totalement de y (la variable dépendante) et qu'il y a uniquement des preuves qui attestent de l'existence d'une relation entre les 2 variables.

7. Relation linéaire et non linéaire

Les valeurs observées et valeurs calculées



Valeurs calculées

Valeurs observées

Les valeurs des observations sont distantes par rapport à la droite de régression (ou droite des moindres carrés) .
La droite de régression est constituée de l'ensemble des valeurs calculées à partir des observations.

8. Modèle théorique

Un **modèle théorique de régression linéaire** repose sur la relation entre une variable dépendante Y et une ou plusieurs variables indépendantes Xi. Il se formalise ainsi :

8.1. Formulation Générale

- Pour une régression linéaire simple avec une seule variable explicative X :

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

- Pour une régression linéaire multiple avec p variables explicatives $X_1, X_2,...,X_p$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

où :

- Y est la variable dépendante (réponse),
- Xi sont les variables explicatives (prédicteurs),
- β0 est l'ordonnée à l'origine (intercept),
- βi sont les coefficients de régression qui mesurent l'effet des Xi sur Y,
- ε est le terme d'erreur, supposé suivre une distribution normale $N(0, \sigma^2)$.

8. Modèle théorique

8.2. Estimation des Coefficients

- Les coefficients β sont généralement estimés par la **méthode des moindres carrés ordinaires (MCO)**, qui minimise la somme des carrés des résidus :

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

où X est la matrice des variables explicatives et Y est le vecteur des valeurs observées.

8.3. Validation du Modèle

- Une fois le modèle ajusté, il est essentiel de le valider en examinant :
- **Le coefficient de détermination R^2** : mesure la proportion de variance expliquée par le modèle.
- **L'analyse des résidus** : permet de vérifier les hypothèses du modèle.
- **Les tests de significativité des coefficients** : via des tests de Student (t) et de Fisher (F).

9. Régression linéaire simple

9. Régression linéaire simple

- La régression et la corrélation conviennent pour détecter une relation linéaire entre variables.

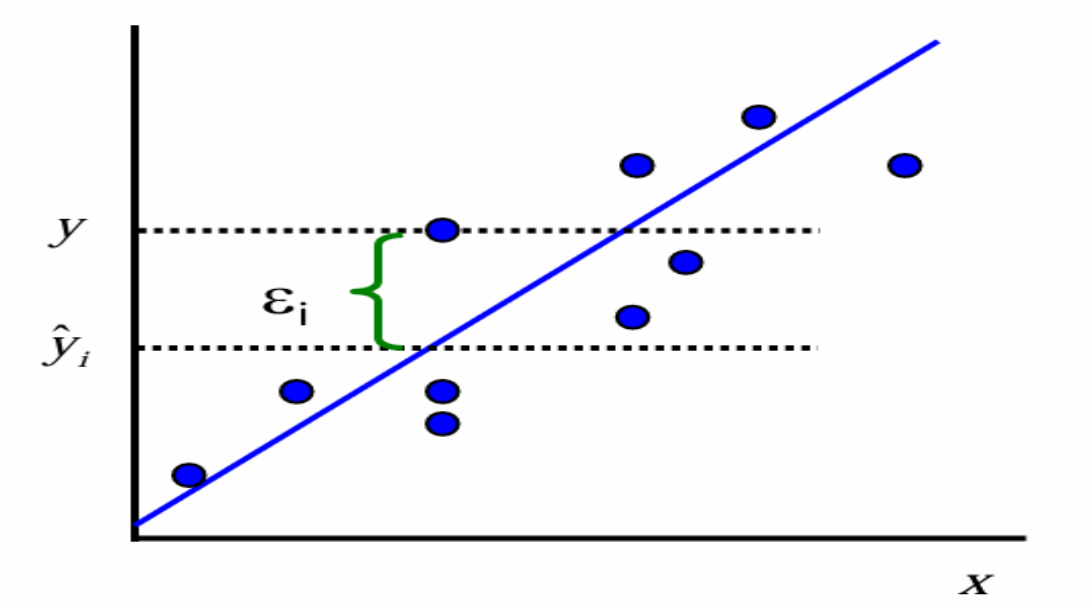
- Donc, la régression vise aussi à ***analyser l'association*** entre une **variable dépendante(*variable prédicteur*)** et une ou plusieurs ***variables indépendantes (variable à prédire)*** **et à prédire la variable dépendante si la variable indépendante est connue .**

9. Régression linéaire simple

9.1. Illustration

• L'écart entre une observation et la droite de régression, encore appelé résidu est donné par la relation:

$$\varepsilon_i = y_i - \hat{y}_i$$



La régression linéaire simple est un modèle statistique qui établit une relation entre une variable dépendante Y et une seule variable explicative X.

9. Régression linéaire simple

9.2. Formulation du Modèle

Le modèle s'écrit sous la forme :

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

où :

- Y : variable dépendante (ce qu'on cherche à prédire),
- X : variable explicative (prédicteur),
- β_0 : intercept (valeur de Y lorsque X=0),
- β1 : coefficient de régression (pente de la droite de régression),
- ε : terme d'erreur (écart entre les valeurs observées et la droite de régression), supposé suivre une distribution normale N (0, σ^2).

9. Régression linéaire simple

9.3. Hypothèses du Modèle

Pour que la régression linéaire simple soit valide, certaines hypothèses doivent être respectées :

- Linéarité : Y et X ont une relation linéaire.
- Indépendance des erreurs : Les erreurs ε sont indépendantes.
- Homoscedasticité : La variance des erreurs est constante.
- Normalité des erreurs : ε suit une distribution normale N $(0,\sigma^2)$.

9. Régression linéaire simple

9.4. Estimation des Paramètres

- Les coefficients β0 et β1 sont estimés par la méthode des moindres carrés ordinaires (MCO), qui minimise la somme des carrés des résidus :

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

où $\bar{X}$ et $\bar{Y}$ sont les moyennes de X et Y.

9. Régression linéaire simple

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

Plus R^2 est proche de 1, plus le modèle explique bien les variations de Y.

- **Test de significativité des coefficients** :

- Test de Student (t) pour vérifier si $\beta 1 \neq 0$.

- Test de Fisher (F) pour évaluer l'ensemble du modèle.

- **Analyse des résidus** : Permet de vérifier la normalité et l'homoscédasticité des erreurs.

9. Régression linéaire simple

9.6. Exemple d'estimation des paramètres et d'évaluation du modèle de régression linéaire simple

9.6.1. Données d'exemple

- Supposons que nous étudions la relation entre le **rendement du maïs (Y, en tonnes par hectare)** et la **quantité d'engrais appliquée (X, en kg/ha)**.

Engrais (X, kg/ha)	Rendement (Y, t/ha)
50	1.2
60	1.4
70	1.6
80	1.8
90	2.0

9. Régression linéaire simple

9.6. Exemple d'estimation des paramètres et d'évaluation du modèle de régression linéaire simple

Nous allons estimer les coefficients du modèle :

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

où :

- Y est la variable dépendante (le rendement du maïs en tonnes/ha),
- X est la variable indépendante (la quantité d'engrais appliquée en kg/ha),
- β_0 est l'ordonnée à l'origine (intercept),
- β_1 est le coefficient de régression (l'effet de l'engrais sur le rendement),
- ε est le terme d'erreur aléatoire.

9. Régression linéaire simple

9.6.2. Estimation des Paramètres

9.6.2.1. Calcul des Moyennes

$$\bar{X} = \frac{50 + 60 + 70 + 80 + 90}{5} = 70$$

$$\bar{Y} = \frac{1.2 + 1.4 + 1.6 + 1.8 + 2.0}{5} = 1.6$$

9.6.2.2. Calcul de la pente $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$= \frac{(50-70)(1.2-1.6) + (60-70)(1.4-1.6) + (70-70)(1.6-1.6) + (80-70)(1.8-1.6) + (90-70)(2.0-1.6)}{(50-70)^2 + (60-70)^2 + (70-70)^2 + (80-70)^2 + (90-70)^2}$$

$$= \frac{(-20 \times -0.4) + (-10 \times -0.2) + (0 \times 0) + (10 \times 0.2) + (20 \times 0.4)}{(-20)^2 + (-10)^2 + 0^2 + 10^2 + 20^2}$$

$$= \frac{8 + 2 + 0 + 2 + 8}{400 + 100 + 0 + 100 + 400} = \frac{20}{1000} = 0.02$$

9. Régression linéaire simple

9.6.2.3. Calcul de l'intercept $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$= 1.6 - (0.02 \times 70) = 1.6 - 1.4 = 0.2$$

Donc, **l'équation estimée du modèle est :**

$$\hat{Y} = 0.2 + 0.02X$$

9.6.3. Évaluation du Modèle : Coefficient de Détermination R^2

On calcule d'abord la somme des carrés totaux SCT, la somme des carrés des résidus SCR, et la somme des carrés expliquée SCE.

$$SCT = \sum (Y_i - \bar{Y})^2$$

$$= (1.2 - 1.6)^2 + (1.4 - 1.6)^2 + (1.6 - 1.6)^2 + (1.8 - 1.6)^2 + (2.0 - 1.6)^2$$

$$= 0.16 + 0.04 + 0 + 0.04 + 0.16 = 0.4$$

9. Régression linéaire simple

Ensuite, on calcule la somme des carrés des résidus :

$$SCR = \sum (Y_i - \hat{Y}_i)^2$$

On calcule les valeurs prédites $\hat{Y}_i = 0.2 + 0.02X_i$:

X_i	Y_i	$\hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$
50	1.2	0.2 + 0.02(50) = 1.2	0
60	1.4	0.2 + 0.02(60) = 1.4	0
70	1.6	0.2 + 0.02(70) = 1.6	0
80	1.8	0.2 + 0.02(80) = 1.8	0
90	2.0	0.2 + 0.02(90) = 2.0	0

9. Régression linéaire simple

$$SCR = 0$$

$$SCE = SCT - SCR = 0.4 - 0 = 0.4$$

$$R^2 = \frac{SCE}{SCT} = \frac{0.4}{0.4} = 1$$

R^2=1 signifie que le modèle explique **100%** de la variance de Y (ce qui est rare en réalité et souvent signe d'un ajustement parfait aux données d'échantillon).

9.6.4. Interprétation des Résultats

- $\hat{\beta}_0 = 0.2$: Lorsque la quantité d'engrais est **0 kg/ha**, le rendement estimé est **0.2 t/ha**.
- $\hat{\beta}_1 = 0.02$: Pour **chaque augmentation de 1 kg/ha d'engrais**, le rendement augmente de **0.02 t/ha**.
- $R^2 = 1$: Le modèle explique **parfaitement** la relation entre l'engrais et le rendement dans cet échantillon.

9. Régression linéaire simple

9.6.5. Conclusion

- Ce modèle simple montre une relation linéaire entre la quantité d'engrais et le rendement du maïs.

Dans une **application réelle**, il serait essentiel de :

- Tester l'hypothèse de normalité des résidus,
- Vérifier si R^2 reste élevé avec **d'autres données** (validation croisée),
- Vérifier s'il y a d'autres variables influençant Y (pluie, type de sol…).

9. Régression linéaire simple

9.7. Autre méthode : Méthode matricielle

9.7.1. Notation matricielle

En notation matricielle, on peut réécrire ce modèle sous la forme :

$$Y=X_{\beta} + \varepsilon$$

où :

- Y est un **vecteur colonne** des valeurs observées de la variable dépendante (rendement),
- X est la **matrice de conception**, contenant une colonne de 1 (pour l'intercept) et une colonne avec les valeurs de la variable indépendante (quantité d'engrais),
- β est le **vecteur des coefficients** à estimer (β0 et β1),
- ε est le **vecteur des erreurs aléatoires**.
- Nous devons estimer β0 et β1.

9. Régression linéaire simple

9.7.2. Construction des matrices

- La matrice X est construite en ajoutant une colonne de 1 (pour β0) :

$$X = \begin{bmatrix} 1 & 50 \\ 1 & 60 \\ 1 & 70 \\ 1 & 80 \\ 1 & 90 \end{bmatrix}$$

9. Régression linéaire simple

9.7.2. Construction des matrices

Le **vecteur Y** est donné par :

$$Y = \begin{bmatrix} 1.2 \\ 1.4 \\ 1.6 \\ 1.8 \\ 2.0 \end{bmatrix}$$

Le **vecteur des coefficients β** est :

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

9. Régression linéaire simple

9.7.3. Estimation des coefficients avec la formule matricielle

Les **moindres carrés ordinaires (MCO)** permettent d'estimer les coefficients $\hat{\beta}$ en minimisant la somme des erreurs quadratiques. La **formule matricielle** pour estimer $\hat{\beta}$ est :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

où :

- X^TX est la **transposée** de la matrice X,
- $(X^TX)^{-1}$ est **l'inverse de la matrice** X^TX
- X^TY est le **produit matriciel** entre X^T et Y.

9. Régression linéaire simple

Étape 1 : Calcul de X^TX

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 70 & 80 & 90 \end{bmatrix} \times \begin{bmatrix} 1 & 50 \\ 1 & 60 \\ 1 & 70 \\ 1 & 80 \\ 1 & 90 \end{bmatrix}$$

Effectuons le produit matriciel

$$X^T X = \begin{bmatrix} 5 & 350 \\ 350 & 25500 \end{bmatrix}$$

9. Régression linéaire simple

Étape 2 : Calcul de X^TY

$$X^TY = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 70 & 80 & 90 \end{bmatrix} \times \begin{bmatrix} 1.2 \\ 1.4 \\ 1.6 \\ 1.8 \\ 2.0 \end{bmatrix}$$

Effectuons le produit matriciel :

$$X^TY = \begin{bmatrix} 5.0 \\ 355.0 \end{bmatrix}$$

9. Régression linéaire simple

Étape 3 : Calcul de $(X^T X)^{-1}$

La matrice inverse de $X^T X$ est :

$$(X^T X)^{-1} = \begin{bmatrix} 1.02 & -0.014 \\ -0.014 & 0.0002 \end{bmatrix}$$

Étape 4 : Calcul de $\hat{\beta}$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\beta} = \begin{bmatrix} 1.02 & -0.014 \\ -0.014 & 0.0002 \end{bmatrix} \times \begin{bmatrix} 5.0 \\ 355.0 \end{bmatrix}$$

Effectuons la multiplication :

$$\hat{\beta} = \begin{bmatrix} 0.2 \\ 0.02 \end{bmatrix}$$

9. Régression linéaire simple

9.7.4. Conclusion : Équation de régression finale

Nous avons estimé les coefficients :

- $\hat{\beta}_0 = 0.2$ (intercept),
- $\hat{\beta}_1 = 0.02$ (effet de l'engrais sur le rendement).

L'équation de régression estimée est donc :

$$\hat{Y} = 0.2 + 0.02X$$

9. Régression linéaire simple

9.8. Application dans R

On cherche à expliquer les variations de la variable quantitative Y (Rendement) par la variable explicative X également quantitative (Quantité d'Engrais).

9.8.1. Importation et affichage des données

RLS<-read.table(file.choose(), header=T)

attach(RLS)

RLS

```
> RLS
  Engrais Rendement
1      50       1.2
2      60       1.4
3      70       1.6
4      80       1.8
5      90       2.0
>
```

9. Régression linéaire simple

9.8.2. Ajustement sur les données

9.8.2.1. Inspection graphique

plot(Rendement~Engrais,xlab="Engrais", ylab="Rendement")

Nous observons une tendance d'augmentation du rendement avec la quantité d'engrai

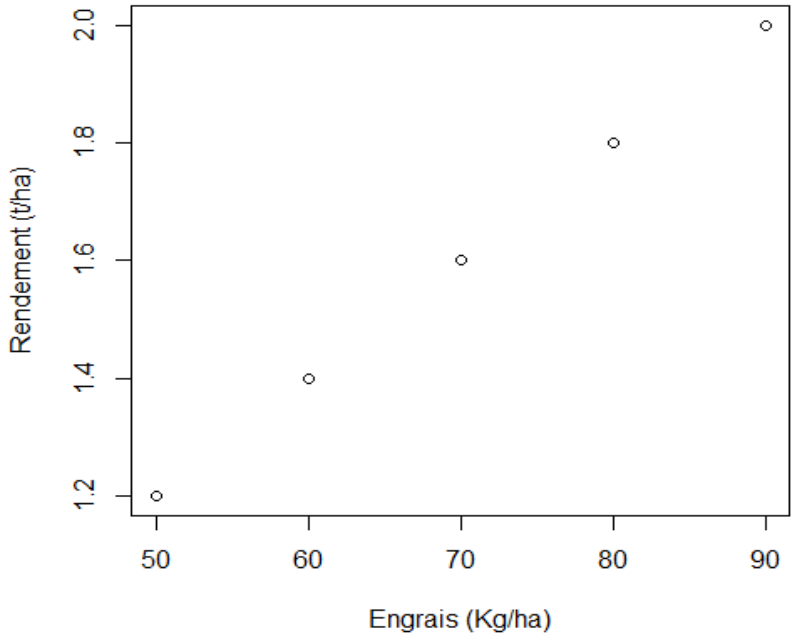


Fig. Nuage de points du rendement (en t/ha) versus la quantité d'engrais (en kg/ha).

9. Régression linéaire simple

9.8.3. Estimation des paramètres de régression a et b.

Le modèle de régression est:

modele1<- lm(Rendement~Engrais)

Modele1

- a = 0.20
- b= 0.02

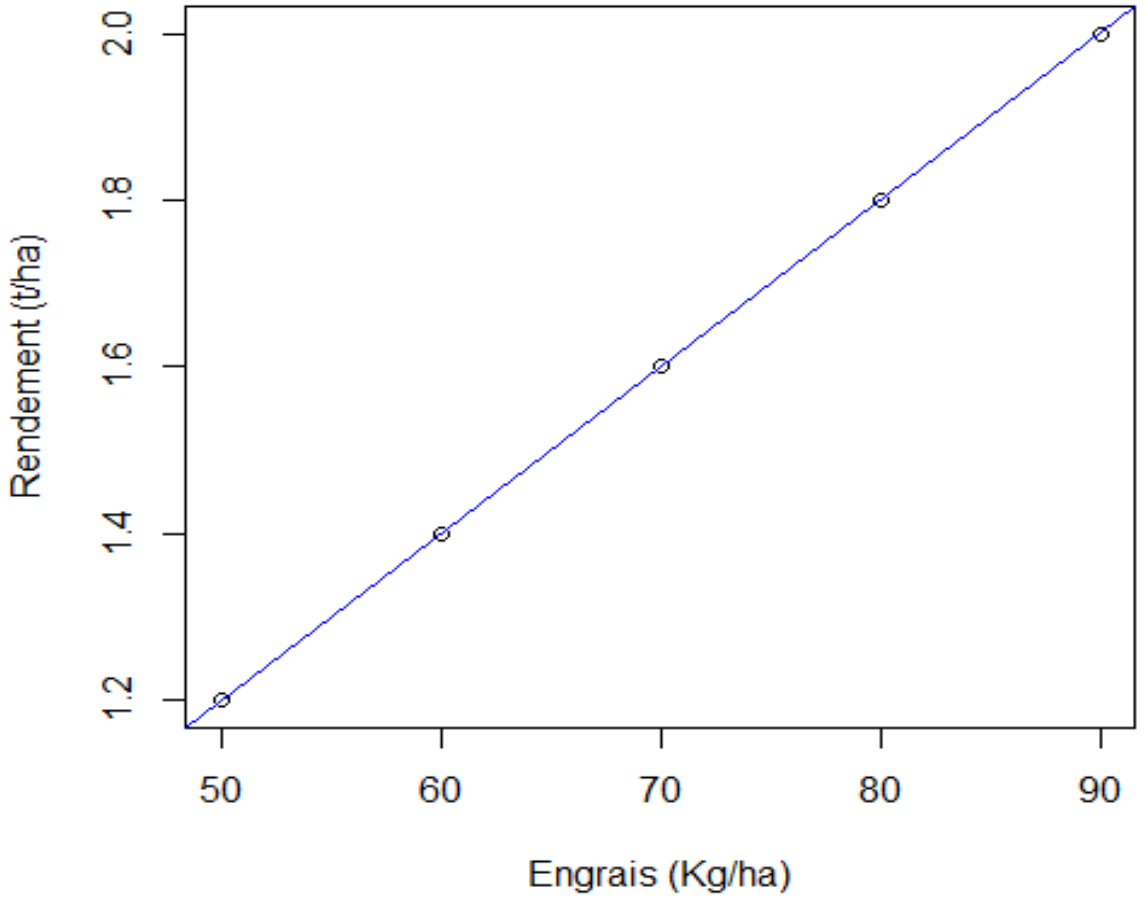
```
> modele1<- lm(Rendement~Engrais)
> modele1

Call:
lm(formula = Rendement ~ Engrais)

Coefficients:
(Intercept)      Engrais
       0.20         0.02
```

9. Régression linéaire simple

9.8.4. Représenter la droite de régression sur le nuage de points.

plot(Rendement~Engrais,xlab="Engrais (Kg/ha)", ylab="Rendement (t/ha)")

abline(modele1,col="blue")



9. Régression linéaire simple

9.8.5. Conditions d'application:

Pour que le modèle soit validé il faut que:

• les résidus soient indépendants et normalement distribués de moyenne nulle et de variance constante;

• la relation entre x et y est linéaire;

• il n'y a pas d'erreur de mesure sur x ;

• le modèle soit globalement significatif

• significativité des coefficients

9. Régression linéaire simple

9.8.6. Validation du modèle

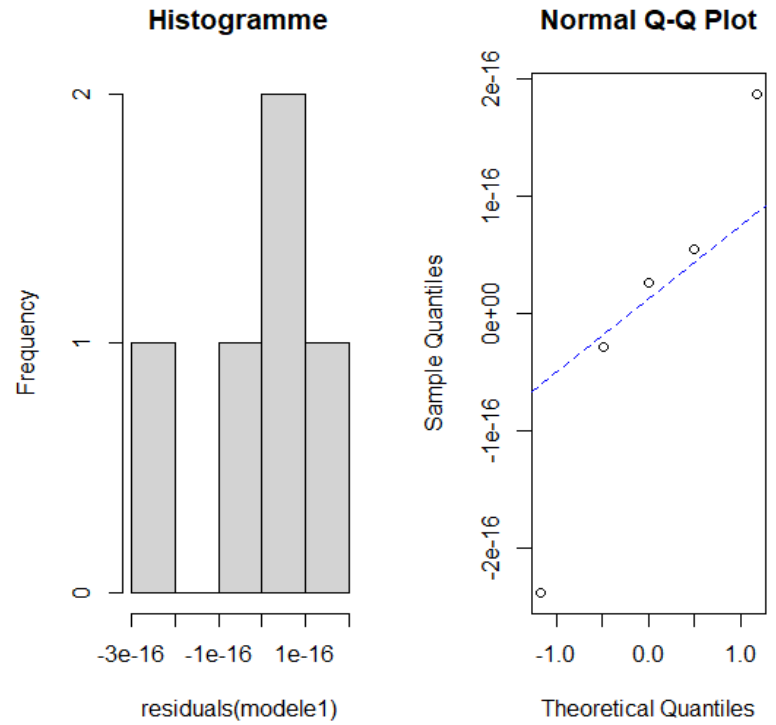
9.8.6.1. Test de normalité des résidus

- **Tracé de l'histogramme des résidus pour détecter la non-normalité.**

Le graphique QQ-plot est une autre approche.

On peut également appliquer le **test de Shapiro-Wilk**

```
par(mfrow=c(1,2))
hist(residuals(modele1), main="Histogramme")
qqnorm(resid(modele1))
qqline(resid(modele1),lty=2,col="blue")
```



Le QQ-plot suggère des erreurs normales puisque les quantiles observés et les quantiles théoriques (obtenus si la distribution est normale) forment une droite.

9. Régression linéaire simple

Test de Shapiro-Wilk (équivalent de Ryan-Joiner)

shapiro.test(resid(modele1))

```
> shapiro.test(resid(modelel))

        Shapiro-Wilk normality test

data:  resid(modelel)
W = 0.95284, p-value = 0.7575
```

Le test ne permet pas de conclure à la non-normalité des erreurs.

Conclusion: L'hypothèse de normalité ne sera donc pas remise en question.

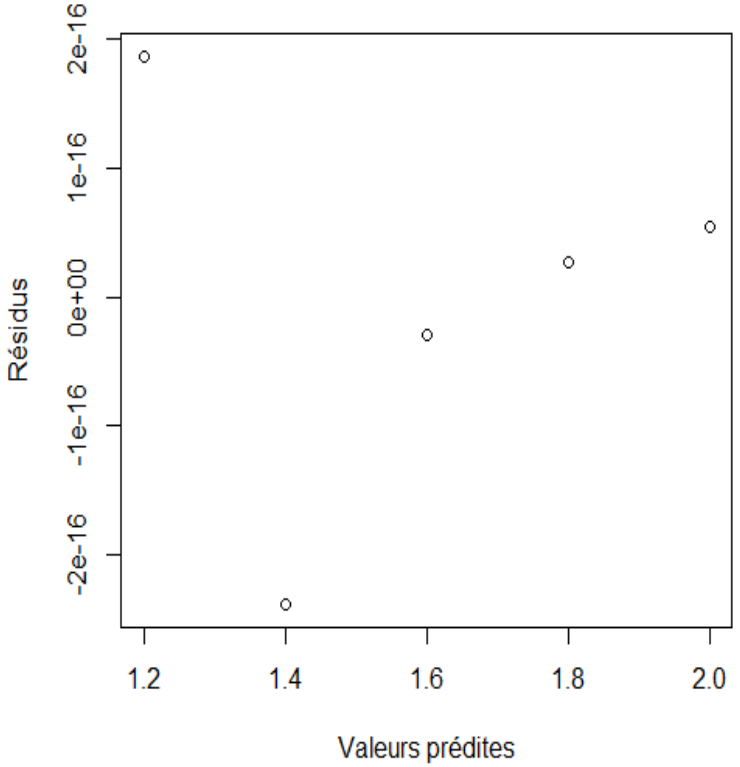
9. Régression linéaire simple

9.8.6.2. Homogénéité des résidus

Le nuage de résidus est correctement répartis et symétrique autour de l'axe des abscisses, les conditions du modèle ne semblent donc pas invalidées

Examen graphique

plot(residuals(modele1)~fitted(modele1), xlab="Valeurs prédite



Le nuage de résidus est correctement répartis et symétrique autour de l'axe des abscisses, les conditions du modèle ne semblent donc pas invalidées.

9. Régression linéaire simple

Test de Breush-Pagan

library(lmtest)

bptest(modele1)

```
> library(lmtest)
> bptest(modelel)

        studentized Breusch-Pagan test

data:  modelel
BP = 2.7656, df = 1, p-value = 0.09631
```

- P-value =0.09631 >0.05 donc H0 est acceptée
- Les résidus sont homogènes

9. Régression linéaire simple

9.8.6.3. Linéarité de la relation

resettest(Rendement ~ Engrais, power=2:3, type="regressor") (Test reset de non linéarité)

```
> resettest(Rendement~Engrais, power=2:3, type="regressor")

        RESET test

data:  Rendement ~ Engrais
RESET = 2.4017, df1 = 2, df2 = 1, p-value = 0.4151
```

- P-value =0 .4151 > 0.05 donc H0 est acceptée
- La condition de linéarité est respectée.

9. Régression linéaire simple

9.8.6.4. Test de signification globale du modèle

anova(modele1)

```
> anova(modele1)
Analysis of Variance Table

Response: Rendement
          Df Sum Sq Mean Sq    F value    Pr(>F)
Engrais    1    0.4     0.4 1.2462e+31 < 2.2e-16 ***
Residuals  3    0.0     0.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Message d'avis :
Dans anova.lm(modele1) :
  Les tests F d'ANOVA sur un ajustement pratiquement parfait ne sont pas fiables
```

Le modèle est globalement significatif

9. Régression linéaire simple

9.8.6.5. Test de significativité de a et de b

result<- summary(modelel)

Result

```
> result

Call:
lm(formula = Rendement ~ Engrais)

Residuals:
          1           2           3           4           5
 1.869e-16 -2.384e-16 -2.938e-17  2.642e-17  5.447e-17

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 2.000e-01  4.046e-16 4.943e+14   <2e-16 ***
Engrais     2.000e-02  5.666e-18 3.530e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.792e-16 on 3 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.246e+31 on 1 and 3 DF,  p-value: < 2.2e-16
```

a et b sont significatifs

9. Régression linéaire simple

9.8.6.6. Test d'autocorrélation des résidus: test de Durbin-Waston

But: tester l'indépendance entre les résidus de régression (présence d'autoccorélation entre les résidus)

library(lmtest)

dwtest(BWT ~ LWT, alternative="greater")

```
> library(lmtest)
> dwtest(modelel)

        Durbin-Watson test

data:  modelel
DW = 2.3726, p-value = 0.3771
alternative hypothesis: true autocorrelation is greater than 0
```

P > 0.05 donc les résidus sont dépendants (auto-corrélés)

10. Régression linéaire multiple

10.1. Rappel

• La régression linéaire simple présente des insuffisances. En effet, pour une expérience factorielle où plusieurs facteurs sont évalués à la fois, la **régression linéaire simple** ne peut être appliquée.

• Dans ces conditions, la régression prenant en compte plusieurs variables indépendantes, en mesurant leurs effets sur une variable dépendante est appelée **régression multiple**.

• Lorsque toutes les variables indépendantes sont supposées affecter la variable dépendante de façon linéaire et indépendamment l'une de l'autre, on parle de **régression linéaire multiple**.

• La **régression linéaire multiple** est une extension de la régression linéaire simple où la variable dépendante (Y) est expliquée par **plusieurs** variables indépendantes (X_1, X_2,..., X_p).

10. Régression linéaire multiple

10.2. Exemples d'application

- Agriculture **: Étudier l'impact de plusieurs facteurs (**quantité d'engrais, irrigation, type de sol**) sur le** rendement agricole**.**

- Économie **: Prédire le** chiffre d'affaires **d'une entreprise en fonction de plusieurs variables (**publicité, prix, saisonnalité**).**

- Immobilier **: Estimer le** prix d'un bien immobilier **en fonction de sa** surface, nombre de chambres, localisation, année de construction**.**

10. Régression linéaire multiple

10.3. Modèle Mathématique

L'équation générale d'un modèle de régression linéaire multiple est :

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

où :

- Y : Variable dépendante (réponse).
- X_1, X_2,..., Xp: Variables explicatives.
- β_0 : **Intercept** (valeur de Y quand toutes les variables explicatives valent 0).
- β1, β2,..., βp: **Coefficients de régression**, qui mesurent l'effet de chaque Xi sur Y.
- ε : **Erreur**, représentant les facteurs non expliqués par le modèle.

10. Régression linéaire multiple

10.4. Hypothèses du Modèle

Pour que la régression linéaire multiple soit valide, les hypothèses suivantes doivent être respectées :

- **Linéarité** : La relation entre Y et chaque X_i est linéaire.
- **Indépendance des erreurs** : Les erreurs doivent être indépendantes entre elles.
- **Homoscedasticité** : La variance des erreurs doit être constante pour toutes les valeurs des X_i.
- **Normalité des erreurs** : Les erreurs doivent suivre une distribution normale N $(0, \sigma^2)$.
- **Absence de multicolinéarité** : Les variables explicatives ne doivent pas être fortement corrélées entre elles.

10. Régression linéaire multiple

10.5. Estimation des Paramètres

- Les coefficients β0, β1,...,βp sont estimés en utilisant la **méthode des moindres carrés ordinaires (MCO)**, qui minimise la somme des carrés des erreurs :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

où :

- X est la matrice des valeurs des variables explicatives.
- Y est le vecteur des valeurs observées de la variable dépendante.
- $\hat{\beta}$ est le vecteur des coefficients estimés.

10. Régression linéaire multiple

10.6. Évaluation du Modèle

10.6.1. Coefficient de Détermination R^2

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

- R^2 mesure la proportion de la variance de Y expliquée par les Xi.
- R^2 proche de **1** indique un bon ajustement du modèle.

10.6.2. Coefficient R^2 ajusté

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

Prend en compte le nombre de variables explicatives pour éviter le surajustement.

10. Régression linéaire multiple

10.7. Tests de Significativité

- **Test de Fisher (F)** : Vérifie si au moins une variable explicative a un effet significatif sur Y.
- **Tests de Student (t)** : Vérifie la significativité individuelle des coefficients βi .

10.8. Analyse des Résidus

- **Normalité** : Test de Shapiro-Wilk ou histogramme des résidus.
- **Homoscedasticité** : Graphique des résidus versus valeurs ajustées.
- **Autocorrélation** : Test de Durbin-Watson.

10. Régression linéaire multiple

10.9. Exemple d'Application

Données : Facteurs influençant le rendement du maïs

Engrais (X1) kg/ha	Irrigation (X2) mm	Type de sol (X3) (1 = argileux, 0 = sableux)	Rendement (Y) t/ha
50	100	1	1.8
60	120	0	2.0
70	110	1	2.3
80	130	1	2.7
90	140	0	2.9

10. Régression linéaire multiple

10.9.1. Estimation des Coefficients

10.9.1.1. Définition du modèle de régression linéaire

Nous avons le modèle suivant :

$$Y=\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

où :

- Y est le rendement (t/ha) (variable dépendante),
- X1 est la quantité d'engrais (kg/ha) (variable indépendante),
- X2 est la quantité d'irrigation (mm) (variable indépendante),
- X3 est le type de sol (1 = argileux, 0 = sableux) (variable indépendante),
- β0, β1, β2, β3 sont les coefficients que nous devons estimer.

10. Régression linéaire multiple

10.9.1.2. Forme Matricielle du Modèle

L'équation de régression linéaire multiple peut être exprimée sous forme matricielle :

$$Y = X\beta + \varepsilon$$

où :

- Y est le vecteur des valeurs observées de la variable dépendante Y,
- X est la matrice des valeurs des variables explicatives avec une première colonne de 1 pour l'intercept,
- β est le vecteur des coefficients inconnus à estimer,
- ε est le vecteur des erreurs.
- Nous avons :

10. Régression linéaire multiple

$$X = \begin{bmatrix} 1 & 50 & 100 & 1 \\ 1 & 60 & 120 & 0 \\ 1 & 70 & 110 & 1 \\ 1 & 80 & 130 & 1 \\ 1 & 90 & 140 & 0 \end{bmatrix} \qquad Y = \begin{bmatrix} 1.8 \\ 2.0 \\ 2.3 \\ 2.7 \\ 2.9 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

10. Régression linéaire multiple

10.9.1.3. Estimation des coefficients avec la méthode des moindres carrés

- **Formule des Moindres Carrés**

Les coefficients β sont estimés en utilisant la formule des moindres carrés ordinaires (MCO) :

où :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

- X^T est la transposée de la matrice X,
- $(X^T X)^{-1}$ est l'inverse du produit $X^T X$,
- $X^T Y$ est le produit de X^T et Y.

10.9.1.4. Calcul Étape par Étape

- Nous allons maintenant effectuer les calculs.
- **Calcul de** $X^T X$
- On transpose X

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 70 & 80 & 90 \\ 100 & 120 & 110 & 130 & 140 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

10. Régression linéaire multiple

Puis, on calcule X^TX:

$$X^TX = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 70 & 80 & 90 \\ 100 & 120 & 110 & 130 & 140 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 50 & 100 & 1 \\ 1 & 60 & 120 & 0 \\ 1 & 70 & 110 & 1 \\ 1 & 80 & 130 & 1 \\ 1 & 90 & 140 & 0 \end{bmatrix}$$

Calcul de X^TY

$$X^TY = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 70 & 80 & 90 \\ 100 & 120 & 110 & 130 & 140 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1.8 \\ 2.0 \\ 2.3 \\ 2.7 \\ 2.9 \end{bmatrix}$$

Calcul de $(X^TX)^{-1}$

- L'inverse de X^TX peut être trouvée par calcul matriciel

Calcul de $\hat{\beta}$

En multipliant $(X^TX)^{-1}$ par X^TY, on obtient :

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -0.298 \\ 0.024 \\ 0.0075 \\ 0.139 \end{bmatrix}$$

10. Régression linéaire multiple

10.9.2. Interprétation des coefficients

$$\hat{Y} = -0.298 + 0.024X_1 + 0.0075X_2 + 0.139X_3$$

- $\beta_0 = -0.298$: C'est l'intercept, représentant le rendement estimé lorsque toutes les variables explicatives sont nulles.
- $\beta_1 = 0.024$: Pour chaque kg/ha supplémentaire d'engrais, le rendement augmente en moyenne de 0.024 t/ha.
- $\beta_2 = 0.0075$: Pour chaque mm supplémentaire d'irrigation, le rendement augmente en moyenne de 0.0075 t/ha.
- $\beta_3 = 0.139$: Lorsque le sol est argileux ($X_3 = 1$), le rendement augmente en moyenne de 0.139 t/ha par rapport à un sol sableux ($X_3 = 0$).

10. Régression linéaire multiple

10.9.3. Prédiction pour un nouveau champ

10.9.3.1. Méthode linéaire

Si on applique **75 kg/ha d'engrais**, **125 mm d'irrigation** sur un **sol argileux** (X3=1) :

$$\hat{Y} = -0.298 + 0.024X_1 + 0.0075X_2 + 0.139X_3$$

Nous allons maintenant remplacer les valeurs des variables par celles fournies :

- $X_1 = 75$ (engrais en kg/ha)
- $X_2 = 125$ (irrigation en mm)
- $X_3 = 1$ (sol argileux)

Nous effectuons le calcul suivant :

$$\hat{Y} = -0.298 + (0.024 \times 75) + (0.0075 \times 125) + (0.139 \times 1)$$

Décomposons :

$$\hat{Y} = -0.298 + 1.8 + 0.9375 + 0.139$$

$$\hat{Y} = 2.5515$$

Arrondi à deux décimales, cela donne **2.55 t/ha**.

10. Régression linéaire multiple

10.9.3.2. Méthode matricielle (utilisée dans le calcul informatique)

La prédiction peut également être obtenue en multipliant le vecteur des coefficients β par le vecteur des nouvelles valeurs Xnew :

$$Y_{\text{pred}} = X_{\text{new}} \cdot B$$

où : $X_{\text{new}} = \begin{bmatrix} 1 & 75 & 125 & 1 \end{bmatrix}$

$$B = \begin{bmatrix} -0.298 \\ 0.024 \\ 0.0075 \\ 0.139 \end{bmatrix}$$

Le produit matriciel donne : $Y_{\text{pred}} = (1 \times -0.298) + (75 \times 0.024) + (125 \times 0.0075) + (1 \times 0.139)$

Ce qui donne le même résultat : **2.55 t/ha**.

10. Régression linéaire multiple

10.9.3.3. Calcul des valeurs prédites

D'abord, calculons les valeurs prédites $\hat{Y}_i$ à partir de notre modèle de régression pour chaque observation.

$$\hat{Y}_i$$

- **Modèle de régression :**

Le **coefficient de détermination R^2** et le **coefficient de détermination ajusté $R^2_{ajusté}$** sont des mesures de la qualité d'ajustement du modèle.

- **Coefficient de détermination R^2 :**

Il mesure la proportion de la variance de Y expliquée par le modèle : $R^2 = 1 - \frac{SSE}{SST}$

où :

- $SSE = \sum(Y_{observé} - Y_{prédit})^2$ (Somme des Erreurs au Carré)
- $SST = \sum(Y_{observé} - \bar{Y})^2$ (Somme Totale des Carrés)

$$SSE = \sum(Y_i - \hat{Y}_i)^2 \qquad SST = \sum(Y_i - \bar{Y})^2$$

- $\bar{Y}$ est la moyenne des valeurs observées de Y.

- R^2 indique la proportion de la variance de Y qui est expliquée par le modèle.

10. Régression linéaire multiple

- **Coefficient de détermination ajusté R^2_{ajuste} :**

Il prend en compte le nombre de variables dans le modèle pour éviter le surajustement :

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

où :

- n est le nombre d'observations,
- p est le nombre de variables explicatives (hors constante).

- **Étapes des calculs**

Les coefficients du modèle de régression sont :

$$\hat{Y} = -0.298 + 0.024X_1 + 0.0075X_2 + 0.139X_3$$

10. Régression linéaire multiple

Étape 1 : Calcul des prédictions ($\hat{Y}$)

Nous allons maintenant calculer R^2 et R^2_{ajuste}.

Y observé	$\hat{Y}$ prédit
1.8	1.799
2.0	2.002
2.3	2.301
2.7	2.698
2.9	2.900

Étape 2 : Calcul de la moyenne des Y observés ($\bar{Y}$)

$$\bar{Y} = \frac{1.8 + 2.0 + 2.3 + 2.7 + 2.9}{5} = \frac{11.7}{5} = 2.34$$

10. Régression linéaire multiple

Étape 3 : Calcul de SST

$$SST = \sum (Y_i - \bar{Y})^2$$

$$SST = (1.8 - 2.34)^2 + (2.0 - 2.34)^2 + (2.3 - 2.34)^2 + (2.7 - 2.34)^2 + (2.9 - 2.34)^2$$

$$SST = (-0.54)^2 + (-0.34)^2 + (-0.04)^2 + (0.36)^2 + (0.56)^2$$

$$SST = 0.2916 + 0.1156 + 0.0016 + 0.1296 + 0.3136 = 0.852$$

Étape 4 : Calcul de SSE

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

$$SSE = (1.8 - 1.799)^2 + (2.0 - 2.002)^2 + (2.3 - 2.301)^2 + (2.7 - 2.698)^2 + (2.9 - 2.900)^2$$

$$SSE = (0.001)^2 + (-0.002)^2 + (-0.001)^2 + (0.002)^2 + (0)^2$$

$$SSE = 0.000001 + 0.000004 + 0.000001 + 0.000004 + 0 = 0.002$$

10. Régression linéaire multiple

Étape 5 : Calcul de R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{0.002}{0.852}$$

$$R^2 = 1 - 0.0024 = 0.998$$

Donc $R^2 = 0.998$ (**99.76%**), ce qui indique que le modèle explique très bien les variations du rendement.

10. Régression linéaire multiple

Étape 6 : Calcul de $R^2_{\text{ajusté}}$

Nous avons :

- $n = 5$ (nombre d'observations),
- $p = 3$ (nombre de variables explicatives : Engrais, Irrigation, Type de sol).

$$R^2_{\text{ajusté}} = 1 - \left(\frac{(1 - 0.998)(5 - 1)}{5 - 3 - 1} \right)$$

$$R^2_{\text{ajusté}} = 1 - \left(\frac{(0.002 \times 4)}{1} \right)$$

$$R^2_{\text{ajusté}} = 1 - 0.008 = 0.990$$

Donc $R^2_{\text{ajusté}} = 0.990$ (99.04%), ce qui confirme que même en tenant compte du nombre de variables, le modèle reste excellent.

10. Régression linéaire multiple

- **Interprétation des résultats:**

· **Coefficient de détermination R^2 : 0.998**

→ Cela signifie que **99.76 %** de la variance du rendement est expliquée par le modèle.

· **Coefficient de détermination ajusté R^2_{ajuste} : 0.990**

→ Cela ajuste R^2 en tenant compte du nombre de variables explicatives et indique que **99.04 %** de la variance est expliquée après correction du nombre de variables.

- Ces valeurs montrent que le modèle est très performant.

10. Régression linéaire multiple

10.10. **Application DANS R**

- **10.10.1. Importation et affichage des données**

RLM<-read.table(file.choose(),header=T)

attach(RLM)

RLM

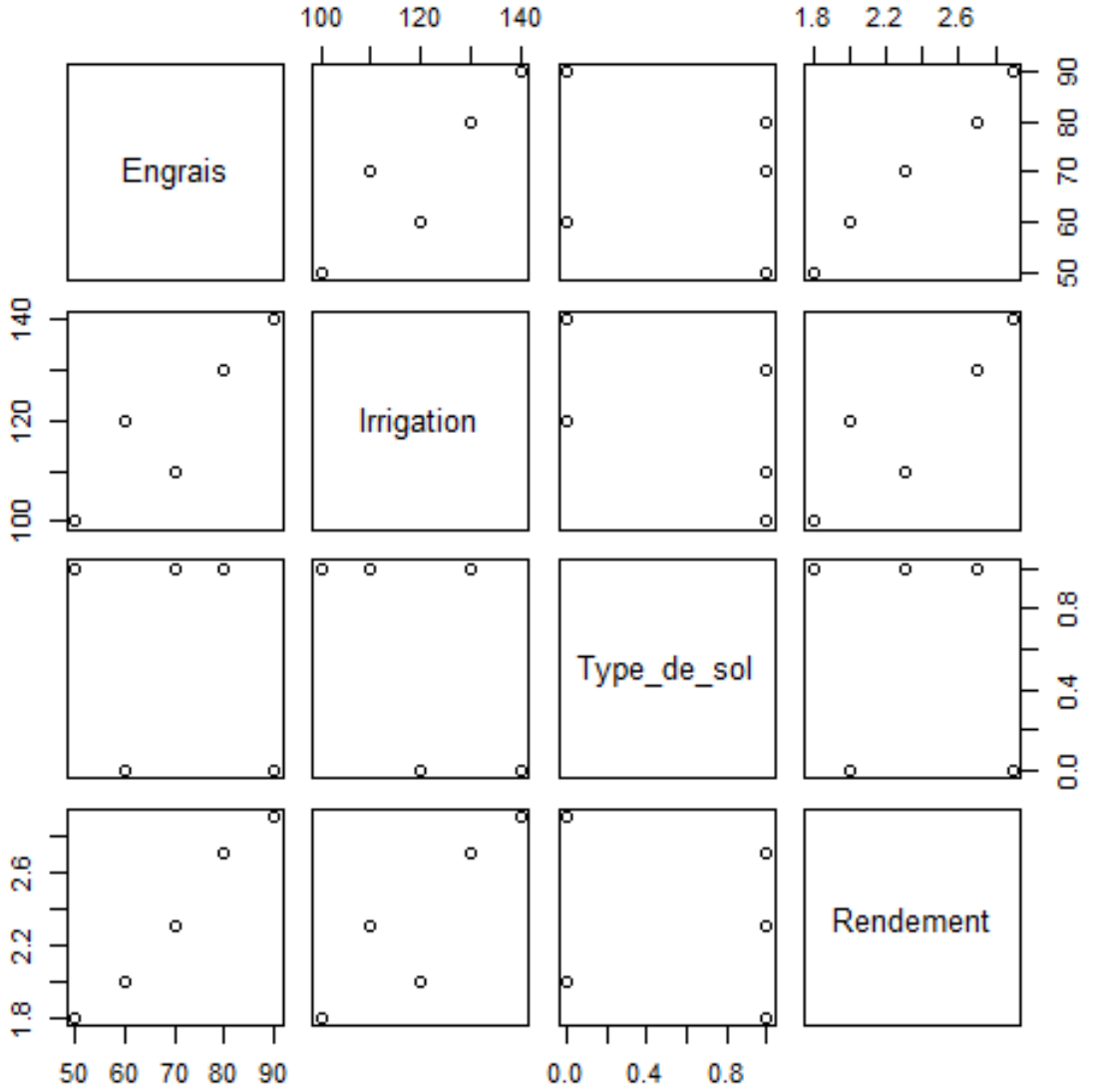
```
> RLM
  Engrais Irrigation Type_de_sol Rendement
1      50        100           1       1.8
2      60        120           0       2.0
3      70        110           1       2.3
4      80        130           1       2.7
5      90        140           0       2.9
```

10. Régression linéaire multiple

10.10.2. Ajustement sur les données

- **Inspection graphique**

pairs(RLM) # Diagramme de dispersion.



- Fig. Diagramme de dispersion de toutes les paires de variables

10. Régression linéaire multiple

10.10.3. Estimation des paramètres

modele2 <- lm(Rendement~Engrais+Irrigation+Type_de_sol)

Modele2

```
> modele2 <- lm(Rendement~Engrais+Irrigation+Type_de_sol)
> modele2

Call:
lm(formula = Rendement ~ Engrais + Irrigation + Type_de_sol)

Coefficients:
(Intercept)      Engrais   Irrigation  Type_de_sol
   -0.29773      0.02364      0.00750      0.13864
```

- **Conditions d'application**: Mêmes conditions que régression linéaire simple (normalité, homoscédasticité, résidus indépendants, significativité des coefficients et de modèle global).
- **Condition supplémentaire**: Colinéarité: Existence d'une relation linéaire entre une variable explicative et les autres.

10. Régression linéaire multiple

10.10.4. Validation du modèle

10.10.4.1. Test de normalité des résidus

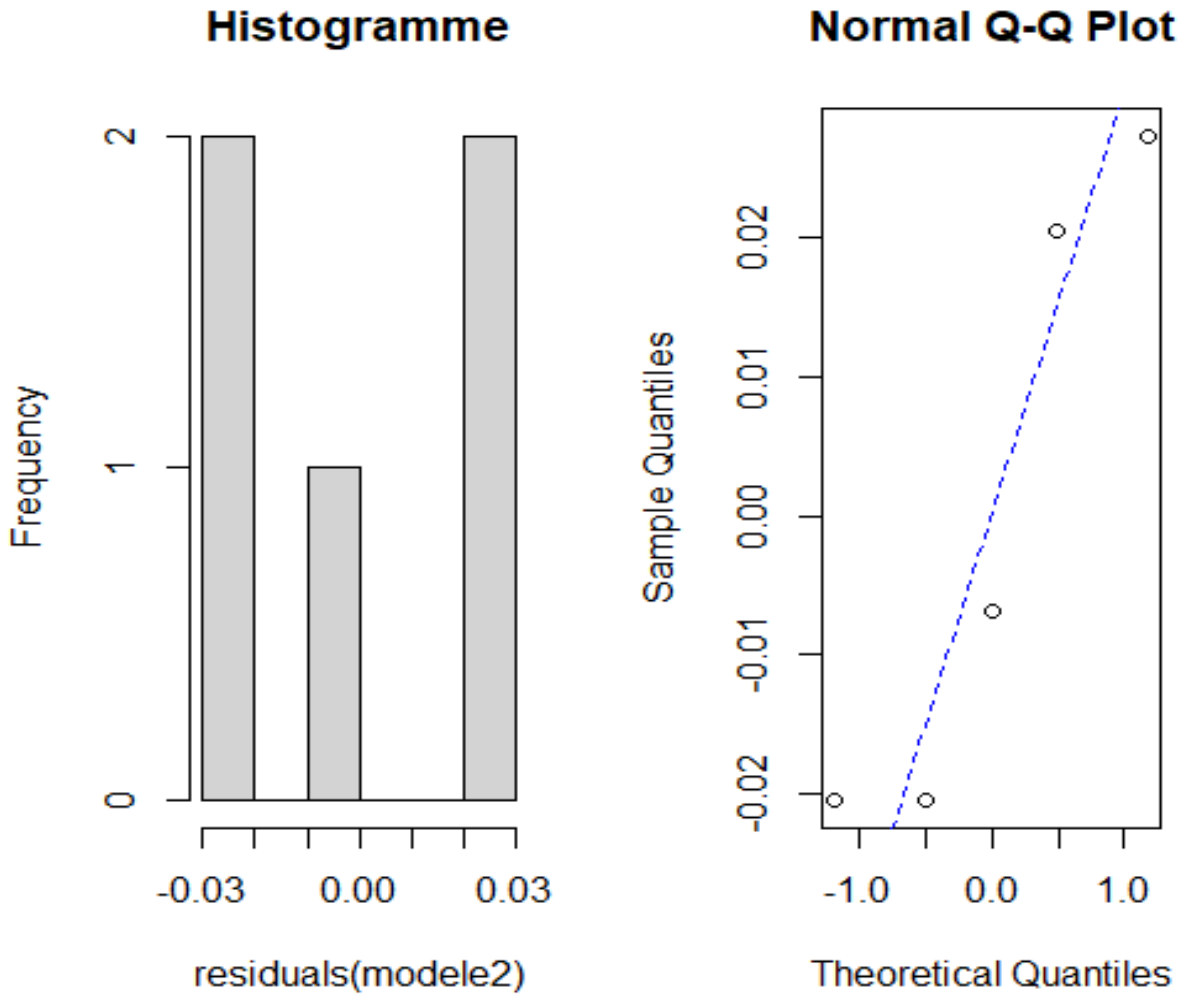
- ***Tracé de l'histogramme des résidus pour détecter la non-normalité.***

Le graphique QQ-plot est une autre approche.

On peut également appliquer le test de Shapiro-Wilk

```
par(mfrow=c(1,2))
hist(residuals(modele2), main="Histogramme")
qqnorm(resid(modele2))
qqline(resid(modele2),lty=2,col="blue")
```

10. Régression linéaire multiple



Le QQ-plot suggère des erreurs normales puisque les quantiles observés et les quantiles théoriques (obtenus si la distribution est normale) forment une droite.

10. Régression linéaire multiple

- **Test de Shapiro-Wilk (équivalent de Ryan-Joiner)**

shapiro.test(resid(modele2))

```
> shapiro.test(resid(modele2))

        Shapiro-Wilk normality test

data:  resid(modele2)
W = 0.84586, p-value = 0.1818
```

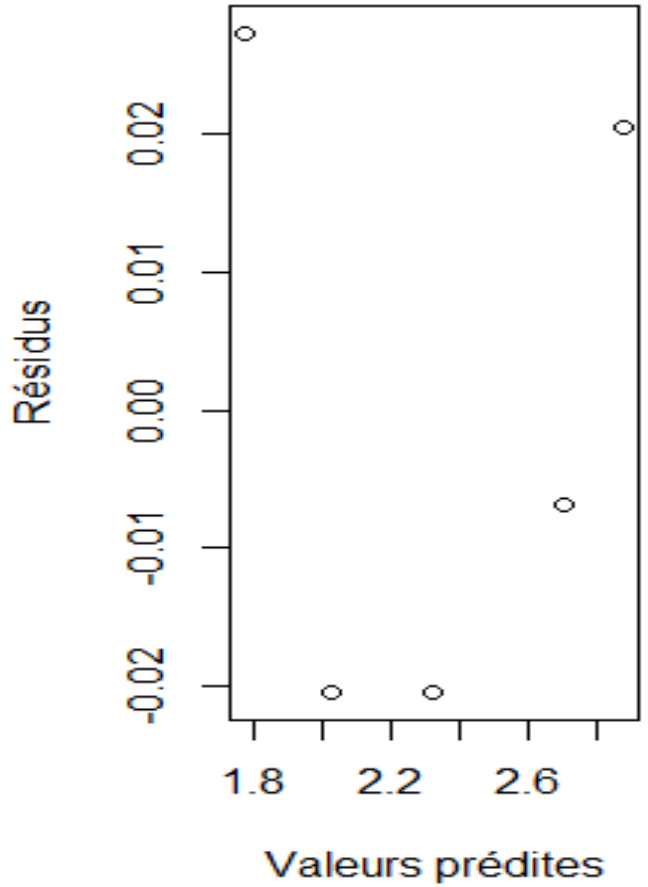
- Le test ne permet pas de conclure à la non-normalité des erreurs.
- **Conclusion**: L'hypothèse de normalité ne sera donc pas remise en question.

10. Régression linéaire multiple

10.10.4.2. Homogénéité des résidus

- **Examen graphique**

plot(residuals(modele2)~fitted(modele2), xlab="Valeurs prédites",y



Le nuage de résidus est correctement répartis et symétrique autour de l'axe des abscisses, les conditions du modèle ne semblent donc pas invalidées.

10. Régression linéaire multiple

- **Test de Breush-Pagan**

library(lmtest)

bptest(modele2)

```
> library(lmtest)
> bptest(modele2)

        studentized Breusch-Pagan test

data:  modele2
BP = 3.6944, df = 3, p-value = 0.2964
```

P-value =0.2964 > 0.05 donc H0 est acceptée

Les résidus sont alors homogènes

- **Test d'autocorrélation des résidus: test de Durbin-Waston**

But: tester l'indépendance entre les résidus de régression (présence d'autocorrélation entre les résidus)

library(lmtest)

dwtest(modele2)

P <0.05 donc les résidus sont indépendants

```
> library(lmtest)
> dwtest(modele2)

        Durbin-Watson test

data:  modele2
DW = 1.5682, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

10. Régression linéaire multiple

10.10.4.3. Linéarité de la relation

resettest(modele2, power=2:3, type="regressor") (Test reset de non linéarité)

```
> resettest(modele2, power=2:3, type="regressor") #(Test reset de non l

        RESET test

data:  modele2
RESET = -Inf, df1 = 6, df2 = -5, p-value = NA

Message d'avis :
Dans pf(reset, df1, df2, lower.tail = FALSE) : Production de NaN
```

10. Régression linéaire multiple

10.10.4.4. Etude de la colinéarité

- Calcul du facteur d'inflation de la variance (VIF)

```
> vif(modele2)
    Engrais  Irrigation Type_de_sol
   9.090909   12.500000    2.590909
```

- **VIF = 1** indique l'absence de relations entre les prédicteurs; **VIF > 1** indique que les prédicteurs sont corrélés. Pour des valeurs de **VIF > 5 ou 10**, les coefficients de régression sont alors mal estimés.

10. Régression linéaire multiple

10.10.4.5. Test de significativité

- **Test de significativité des coefficients**

#result2<- summary(modele1)

result2<- summary(modele2)

result2

```
> result<- summary(modele2) # Présentation des résultats.
> result

Call:
lm(formula = Rendement ~ Engrais + Irrigation + Type_de_sol)

Residuals:
        1         2         3         4         5
 0.027273 -0.020455 -0.020455 -0.006818  0.020455

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.297727   0.372637  -0.799    0.571
Engrais      0.023636   0.004312   5.481    0.115
Irrigation   0.007500   0.005056   1.483    0.378
Type_de_sol  0.138636   0.066455   2.086    0.285

Residual standard error: 0.04523 on 1 degrees of freedom
Multiple R-squared:  0.9976,    Adjusted R-squared:  0.9904
F-statistic: 138.5 on 3 and 1 DF,  p-value: 0.06236
```

10. Régression linéaire multiple

- **Test de signification globale du modèle**

```
> anova(modele2)
Analysis of Variance Table

Response: Rendement
             Df  Sum Sq Mean Sq  F value  Pr(>F)
Engrais       1 0.84100 0.84100 411.1556 0.03137 *
Irrigation    1 0.00005 0.00005   0.0257 0.89874
Type_de_sol   1 0.00890 0.00890   4.3520 0.28456
Residuals     1 0.00205 0.00205
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11. Limites et perspectives des méthodes paramétriques

- Hypothèses restrictives : solutions alternatives (modèles non paramétriques).
- Utilisation conjointe avec d’autres méthodes (e.g., apprentissage automatique).
- Amélioration des performances par des techniques de régularisation (Lasso, Ridge).

12. Extensions des modèles linéaires aux modèles spécialisés : Exposés

- Transformations de variables en régressions : Transformations logarithmique, carré et inverse
- Transformation Box-Cox et applications
- Effet des transformations sur l'hétéroscédasticité et la normalité des résidus
- Analyse des résidus : détection des erreurs et des anomalies
- Critères de sélection des modèles : AIC, BIC et application
- Test de multicolinéarité et solutions (VIF, PCA, Ridge)
- Régression polynomiale : applications et limites
- Régression logistique : fondements et applications
- Comparaison entre régression de Poisson et régression binomiale négative
- Comparaison entre régression de Poisson et régression quasi-poisson
- Régression quantile : une alternative à la régression linéaire
- Régression log-linéaire : applications en analyse des tableaux de contingence
- Régression spline : une approche flexible pour modéliser des relations non linéaires
- Impact des valeurs aberrantes sur les modèles de régression et stratégies de traitement
- Méthodes d'estimation des paramètres en régression