SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

13 mars 2024

Plan de la séance

Plan de la séance

- Corrélation
- Régression linéaire

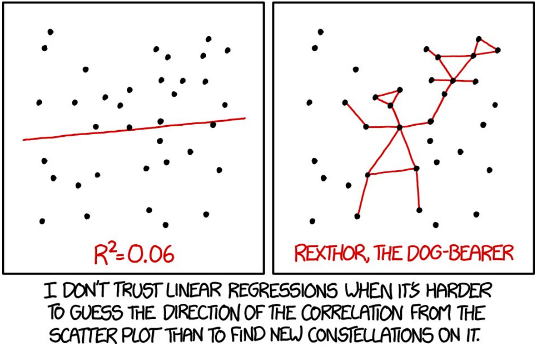


Figure 1: Brise-glace

Corrélation

Corrélation

La corrélation de variables aléatoires est une mesure qui quantifie le degré auquel deux variables aléatoires varient ensemble.

- Si les variations des deux variables montrent une tendance à se produire ensemble, on dit qu'elles sont positivement corrélées.
- Si une variable a tendance à augmenter quand l'autre diminue, elles sont négativement corrélées.

Corrélation

La corrélation de variables aléatoires est une mesure qui quantifie le degré auquel deux variables aléatoires varient ensemble.

- Si les variations des deux variables montrent une tendance à se produire ensemble, on dit qu'elles sont positivement corrélées.
- Si une variable a tendance à augmenter quand l'autre diminue, elles sont négativement corrélées.

La corrélation est souvent mesurée par un **coefficient** qui varie entre -1 et 1.

Un coefficient de 1 indique une corrélation positive parfaite, -1 indique une corrélation négative parfaite, et 0 indique l'absence de corrélation.

Corrélation de Pearson (Paramétrique)

Définition: La corrélation de Pearson, également connue sous le nom de coefficient de corrélation produit-moment de Pearson, évalue la **relation linéaire** entre deux variables quantitatives.

Caractéristiques: Valeurs entre -1 et 1.

Utilisation: Préférable lorsque les deux variables sont **normalement distribuées** et la relation est supposée être linéaire.

Corrélation de Pearson (Paramétrique)

Formule: Corrélation de Pearson = (Covariance de X et Y) / (Écart-type de X * Écart-type de Y).

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

où r_{xy} est le coefficient de corrélation de Pearson entre les variables x et y, x_i et y_i sont les valeurs des variables, et $\bar{x}$ et $\bar{y}$ sont les moyennes de x et y, respectivement.

Corrélation de Spearman (Non-Paramétrique)

Définition: La corrélation de Spearman, ou le coefficient de **rang** de Spearman, est utilisée pour mesurer la force et la direction de l'association entre deux variables classées.

Caractéristiques: Également évaluée entre -1 et 1. Moins sensible aux valeurs aberrantes.

Utilisation: Appropriée lorsque les données ne sont pas normalement distribuées ou lorsqu'on examine des relations non linéaires.

Corrélation de Spearman (Non-Paramétrique)

Formule: Corrélation de Spearman = 1 - (6 * Somme des carrés des différences de rang) / (n(n^2 - 1)).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

où ρ est le coefficient de corrélation de Spearman, d_i est la différence entre les rangs des i-èmes valeurs de x et y, et n est le nombre de paires de données.

Rappel

Il est crucial de se rappeler que la corrélation ne signifie pas causalité.

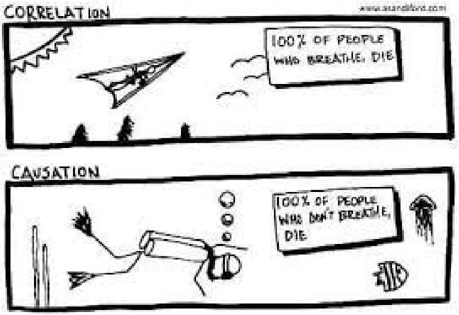


Figure 2: Corrélation vs. causalité

Analyse avec R : Données

Base de données “Pima Indian Diabetes”

Analyse avec R : Données

Base de données “Pima Indian Diabetes”

Test de normalité de Shapiro-Wilk

```
## Glucose : p-value = 1.720326e-11
## BloodPressure : p-value = 9.45138e-05
## SkinThickness : p-value = 1.775691e-09
## Insulin : p-value = 1.698218e-21
## BMI : p-value = 8.557785e-09
## Age : p-value = 2.402274e-24
```

Analyse avec R : Données

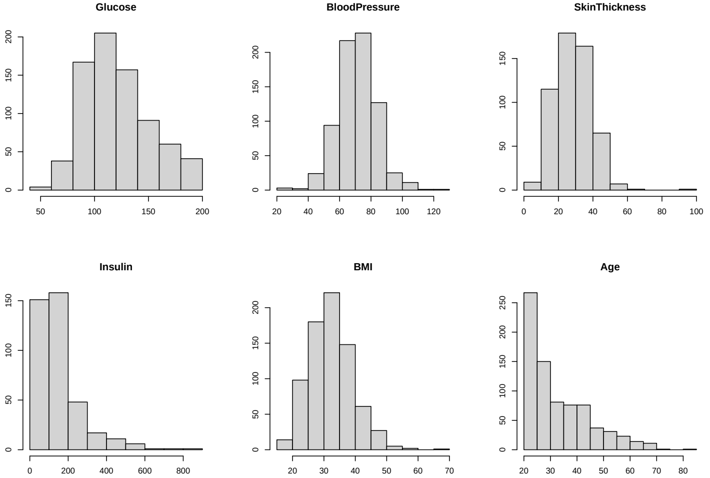
Base de données "Pima Indian Diabetes"

Test de normalité de Shapiro-Wilk

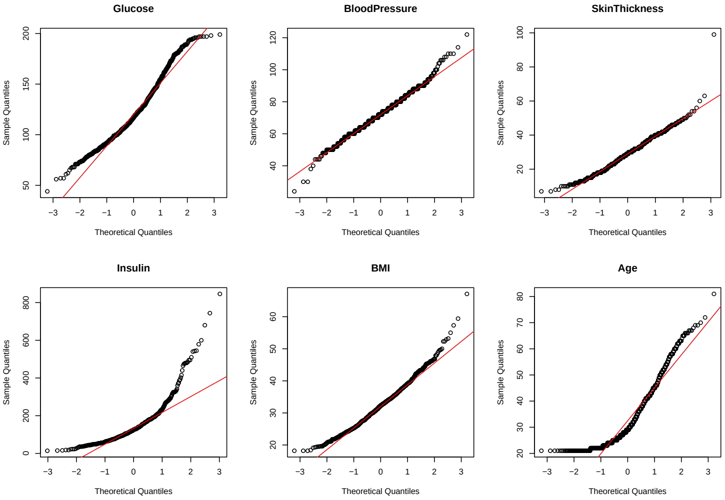
```
## Glucose : p-value = 1.720326e-11
## BloodPressure : p-value = 9.45138e-05
## SkinThickness : p-value = 1.775691e-09
## Insulin : p-value = 1.698218e-21
## BMI : p-value = 8.557785e-09
## Age : p-value = 2.402274e-24
```

Les données ne sont pas normalement distribuées. Il faut donc utiliser la corrélation de Spearman.

Analyse avec R : Histogrammes



Analyse avec R : Graphiques Q-Q



Analyse avec R : Corrélation de Spearman

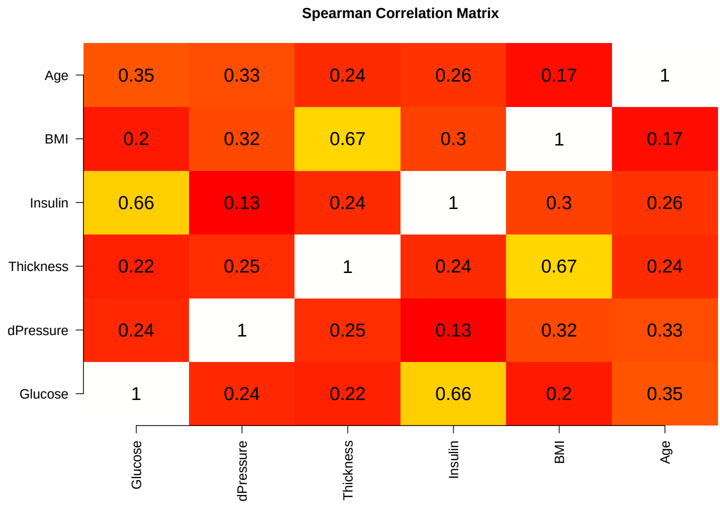
```r
spearman_correlation_matrix <-
  cor(diabetes_subset,
      use="complete.obs",
      method="spearman")
```

La fonction `cor(diabetes_subset)` calcule les coefficients de corrélation pour toutes les paires de variables dans la base de données `diabetes_subset`.

L'argument `method="spearman"` spécifie que le coefficient de corrélation de rang de Spearman doit être utilisé.

L'argument `use="complete.obs"` indique à R d'utiliser uniquement des cas complets (c'est-à-dire des lignes sans aucune valeur NA).

Analyse avec R : Corrélation de Spearman



- Assez forte corrélation positive entre `SkinThickness` et `BMI`, et entre `Glucose` et `Insulin`. Toutefois, . . .

Rappel

Le fait que deux variables soient fortement corrélées ne démontre pas que l'une est la cause de l'autre.

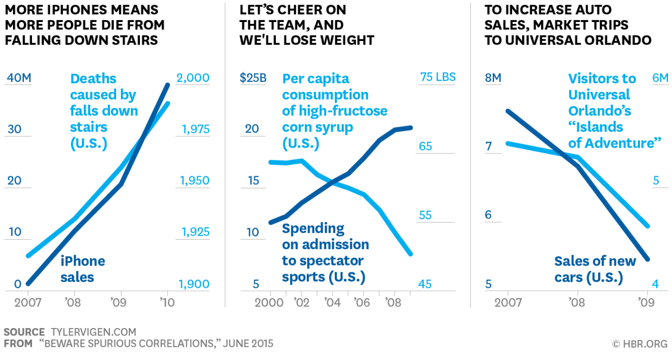
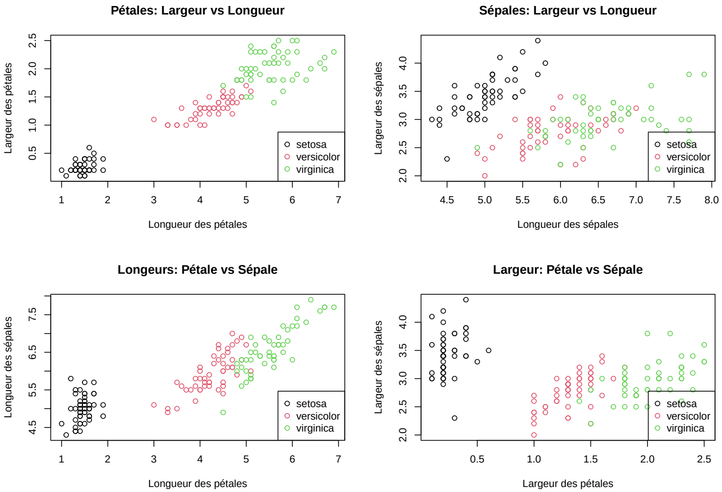


Figure 3: Corrélation fallacieuse

Analyse avec R : Visualisation de données

Iris - nuage de points (“scatter plots” en anglais)



Analyse avec R : Distribution normale

Test de normalité de Shapiro-Wilk

```
## Sepal.Length : p-value = 0.01018
## Sepal.Width : p-value = 0.10115
## Petal.Length : p-value = 0
## Petal.Width : p-value = 0
```

Analyse avec R : Distribution normale

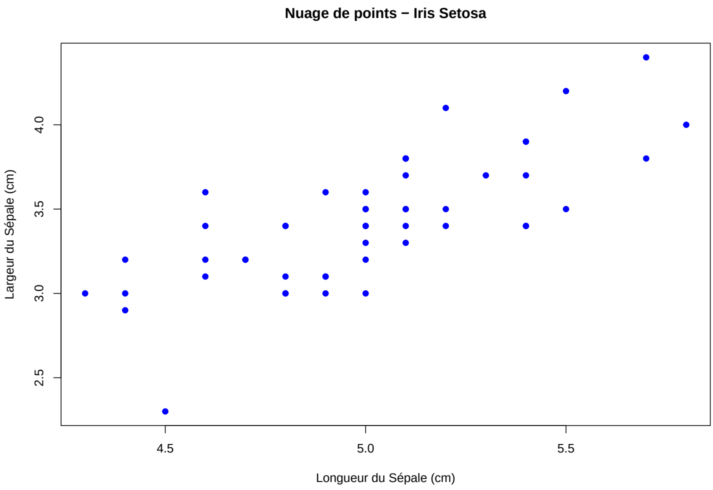
Test de normalité de Shapiro-Wilk

```
## Sepal.Length : p-value = 0.01018
## Sepal.Width : p-value = 0.10115
## Petal.Length : p-value = 0
## Petal.Width : p-value = 0
```

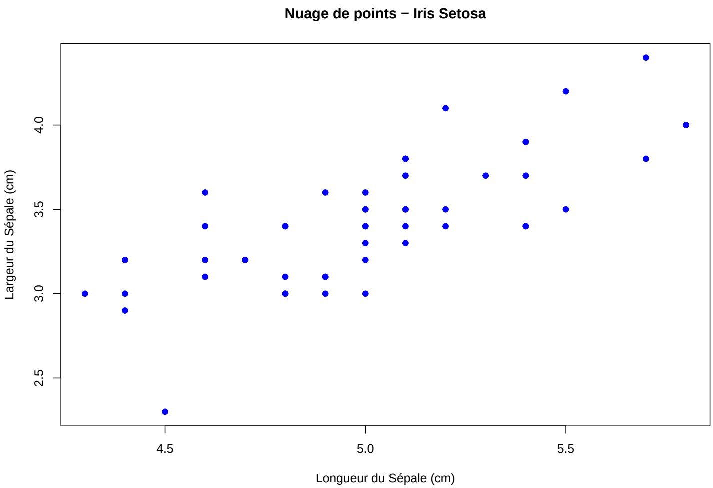
En incluant uniquement l'espèce de `setosa`

```
## Sepal.Length : p-value = 0.45951
## Sepal.Width : p-value = 0.27153
## Petal.Length : p-value = 0.05481
## Petal.Width : p-value = 0
```

Analyse avec R : Rélation linéaire



Analyse avec R : Rélation linéaire



Ainsi, nous démontrerons le calcul de corrélation de Pearson pour la longueur et la largeur des sépales de `setosa`.

Analyse avec R : Corrélation de Pearson

La méthode par défaut pour `cor()` est le coefficient de corrélation de Pearson.

La fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` calcule le coefficient de corrélation de Pearson entre deux variables `x` et `y`.

```r
cor(iris_setosa$Sepal.Length, iris_setosa$Sepal.Width)
```

```
## [1] 0.7425467
```

Analyse avec R : Corrélation de Pearson

La méthode par défaut pour `cor()` est le coefficient de corrélation de Pearson.

La fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` calcule le coefficient de corrélation de Pearson entre deux variables `x` et `y`.

```
cor(iris_setosa$Sepal.Length, iris_setosa$Sepal.Width)
```

```
## [1] 0.7425467
```

- Le coefficient d'environ 0,74 suggère qu'à mesure que l'une des variables (longueur ou largeur) augmente, l'autre variable a tendance à augmenter également, et cette relation est relativement forte. Cependant, ...

Rappel

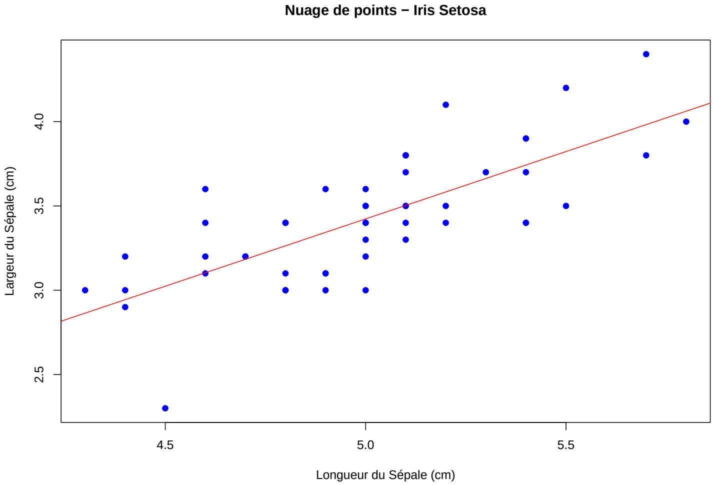
L'existence d'une corrélation entre deux variables n'implique pas une relation de cause à effet.

Figure 4: Corrélation, et non causalité

Régression linéaire simple

Régression linéaire simple

Alors, peut-on connaître ou deviner la largeur des sépales si l'on a déjà mesuré la longueur des sépales ?



Relation entre la corrélation de Pearson et la régression linéaire simple

But

- **Corrélation de Pearson** mesure la force et la direction de la relation linéaire entre deux variables.
- **Régression linéaire** explique une variable (réponse) en fonction de la valeur d'une autre (prédicteur).

Relation entre la corrélation de Pearson et la régression linéaire simple

But

- **Corrélation de Pearson** mesure la force et la direction de la relation linéaire entre deux variables.
- **Régression linéaire** explique une variable (réponse) en fonction de la valeur d'une autre (prédicteur).

Résultat

- **Coefficient de corrélation** (`r`) varie de -1 à 1.
- **Régression linéaire** fournit une équation de la forme : $y = \beta_0 + \beta_1 x + \epsilon$

Régression linéaire simple : Modèle

La régression linéaire simple d'une variable dépendante y sur une variable indépendante x est modélisée par:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Intercept (β_0) est l'ordonnée à l'origine ou la valeur attendue de y quand $x = 0$.
- Pente (β_1) est la pente de la ligne de régression indiquant le changement attendu dans y pour une augmentation d'une unité de x.
- Terme d'erreur (ϵ) est la variation non expliquée.

Régression linéaire simple : Estimation des coefficients

L'objectif est d'estimer les coefficients β_0 et β_1 à partir des données. Cela se fait généralement en utilisant la méthode des **moindres carrés**, qui minimise la somme des différences au carré entre les valeurs observées et les valeurs prédites par le modèle.

Régression linéaire simple : Estimation des coefficients

L'objectif est d'estimer les coefficients β_0 et β_1 à partir des données. Cela se fait généralement en utilisant la méthode des **moindres carrés**, qui minimise la somme des différences au carré entre les valeurs observées et les valeurs prédites par le modèle.

- Calcul des résidus : $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
  - ϵ_i est le résidu pour l'observation i
  - y_i est la valeur observée
  - $(\beta_0 + \beta_1 x_i)$ est la valeur prédite par le modèle

Régression linéaire simple : Estimation des coefficients

L'objectif est d'estimer les coefficients β_0 et β_1 à partir des données. Cela se fait généralement en utilisant la méthode des **moindres carrés**, qui minimise la somme des différences au carré entre les valeurs observées et les valeurs prédites par le modèle.

- Calcul des résidus : $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
  - ϵ_i est le résidu pour l'observation i
  - y_i est la valeur observée
  - $(\beta_0 + \beta_1 x_i)$ est la valeur prédite par le modèle

- Minimisation de la a somme des carrés des résidus : $\sum \epsilon_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$
- Trouve les coefficients β_0 et β_1 qui minimisent cette somme

Régression linéaire simple : Conditions

- **Linéarité** : La relation entre $\mathtt{x}$ et $\mathtt{y}$ est linéaire.
- **Indépendance** : Les observations sont indépendantes.
- **Homoscédasticité** : La variance du terme d'erreur ϵ est constante pour tous les niveaux de $\mathtt{x}$.
- **Normalité des erreurs** : Les termes d'erreur ϵ sont normalement distribués (important pour faire des inférences sur les coefficients).

Régression linéaire simple : Inférence

- **Tests d'Hypothèse** pour la pente (β_1) en utilisant un test t, si elle est significativement différente de zéro.

Régression linéaire simple : Inférence

- **Tests d'Hypothèse** pour la pente (β_1) en utilisant un test t, si elle est significativement différente de zéro.
- **Intervalles de Confiance** pour les coefficients β_0 et β_1.

Régression linéaire simple : Inférence

- **Tests d'Hypothèse** pour la pente (β_1) en utilisant un test t, si elle est significativement différente de zéro.
- **Intervalles de Confiance** pour les coefficients β_0 et β_1.
- **Coefficient de Détermination (R^2)** indique la qualité de l'ajustement du modèle aux données.
  - Est équivalent du carré du coefficient de corrélation de Pearson (r^2)
  - Estime la proportion de la variance dans la variable dépendante y qui peut être prédite ou inférée à partir de la variable indépendante x.

Relation entre la corrélation de Pearson et la régression linéaire simple (retour)

Force de l'association : La corrélation de Pearson fournit une mesure de la force de la relation linéaire, ce qui est crucial pour décider si la régression linéaire est appropriée.

Relation entre la corrélation de Pearson et la régression linéaire simple (retour)

Force de l'association : La corrélation de Pearson fournit une mesure de la force de la relation linéaire, ce qui est crucial pour décider si la régression linéaire est appropriée.

Direction et Pente : Le signe du coefficient de corrélation de Pearson `r` indique la direction de la relation (+ ou -), qui correspond à la pente dans la régression linéaire.

Relation entre la corrélation de Pearson et la régression linéaire simple (retour)

Force de l'association : La corrélation de Pearson fournit une mesure de la force de la relation linéaire, ce qui est crucial pour décider si la régression linéaire est appropriée.

Direction et Pente : Le signe du coefficient de corrélation de Pearson `r` indique la direction de la relation (+ ou -), qui correspond à la pente dans la régression linéaire.

Variance expliquée : Dans une régression linéaire simple avec un seul prédicteur, le carré du coefficient de corrélation de Pearson (r^2) est égal à la statistique R^2 en régression, représentant la proportion de la variance dans la variable dépendante expliquée par la variable indépendante.

Analyse avec R : Modèle

Iris

La pente de la régression / la direction de la rélation :
Le coefficient de x est-il positif ou négatif ?

```
model <- lm(Sepal.Width ~ Sepal.Length,
            data = iris_setosa)
model
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = iris_setos
##
## Coefficients:
##  (Intercept)  Sepal.Length
##      -0.5694        0.7985
```

Analyse avec R : Tests d'Hypothèse

La valeur p du β_1 (coefficient de x) est-elle inférieure à 0,05 ?

```
summary(model)
```

```
##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = iris_setos
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72394 -0.18273 -0.00306  0.15738  0.51709
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.5694     0.5217  -1.091    0.281
## Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
## ---
```

Analyse avec R : Intervalles de Confiance

Les intervalles de confiance couvrent-ils les valeurs négatives, 0, ou positives... ou toutes ?

```r
confint(model, level = 0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept)  -1.6184048 0.4795395
## Sepal.Length  0.5894925 1.0075641
```

Analyse avec R : R^2

La force de l'association : R^2 est-elle supérieure à 0,06 ?

```r
summary(model)$r.squared
```

```
## [1] 0.5513756
```

Analyse avec R : R^2

La force de l'association : R^2 est-elle supérieure à 0,06 ?

```
summary(model)$r.squared
```

```
## [1] 0.5513756
```

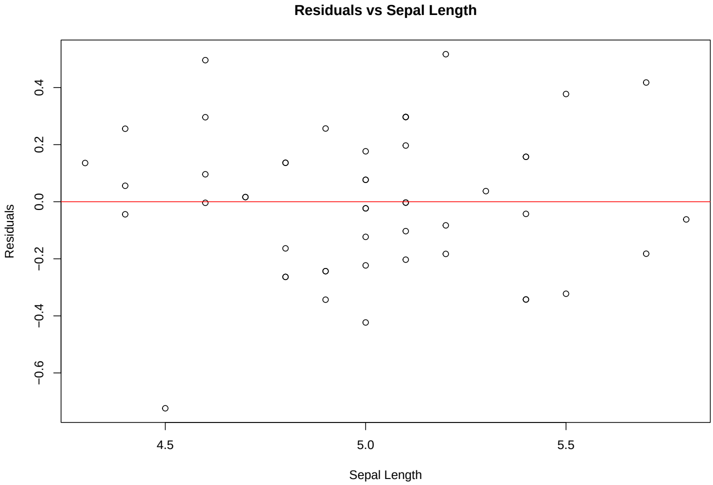
Et le carré du coefficient de corrélation de Pearson est . . .

```
cor(iris_setosa$Sepal.Length, iris_setosa$Sepal.Width)^2
```

```
## [1] 0.5513756
```

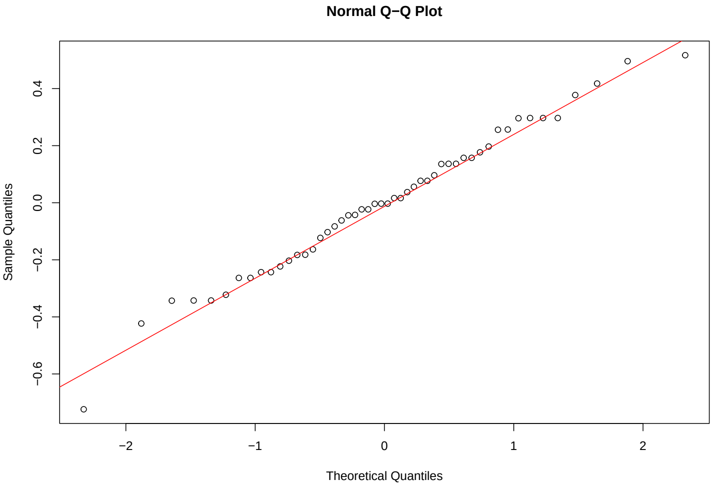
Analyse avec R : Homoscédasticité

La variance de ϵ ou `model$residuals` est-elle constante pour tous les niveaux de `x` ?



Analyse avec R : Normalité des Erreurs

La valeur p du test Shapiro-Wilk des résidus du modèle (ϵ ou `model$residuals`) est 0.8459357.



Régression linéaire multiple

Confondeurs