SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

13 mars 2024

Plan de la séance

Plan de la séance

- Corrélation
- Régression linéaire

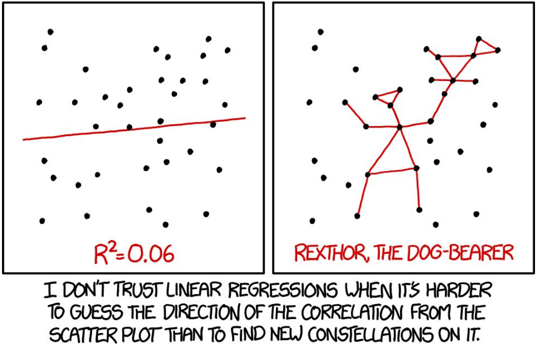


Figure 1: Brise-glace

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

Fonction R : `cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))`

Corrélation

Fonction R : `cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))`

Lorsqu'il manque des valeurs dans l'ensemble de données, il faut bien choisir l'argument `use` parmi `everythng`, `all.obs`, `complete.obs`, `na.or.complete`, ou `pairwise.complete.obs`.

Corrélation

Fonction R : `cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))`

Lorsqu'il manque des valeurs dans l'ensemble de données, il faut bien choisir l'argument `use` parmi `everythng`, `all.obs`, `complete.obs`, `na.or.complete`, ou `pairwise.complete.obs`.

`method = c("pearson", "kendall", "spearman")`

Avec quelle méthode est-ce qu'on va utiliser ?

Corrélation de Pearson (Paramétrique)

La corrélation de Pearson évalue la **relation linéaire** entre deux variables quantitatives.

Elle est utilisée lorsque les deux variables sont **normalement distribuées** et la relation est supposée être linéaire.

Corrélation de Pearson (Paramétrique)

La corrélation de Pearson évalue la **relation linéaire** entre deux variables quantitatives.

Elle est utilisée lorsque les deux variables sont **normalement distribuées** et la relation est supposée être linéaire.

= (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)

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.

Elle est moins sensible aux valeurs aberrantes.

Corrélation de Spearman (Non-Paramétrique)

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.

Elle est moins sensible aux valeurs aberrantes.

= 1 - (6 * Somme des carrés des différences de rang) / (n(nˆ2 - 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.

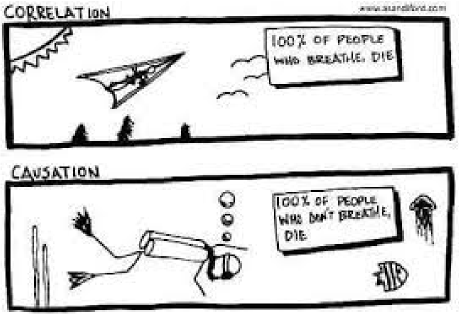


Figure 2: Corrélation vs. causalité

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

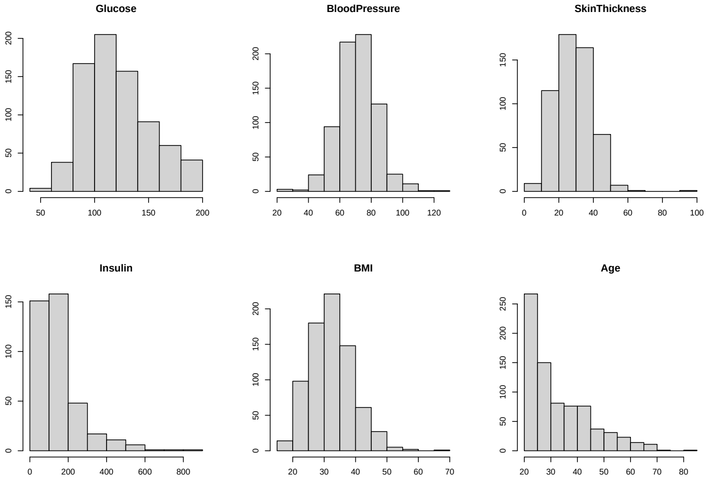
Base de données “Pima Indian Diabetes”

Test de normalité de Shapiro-Wilk

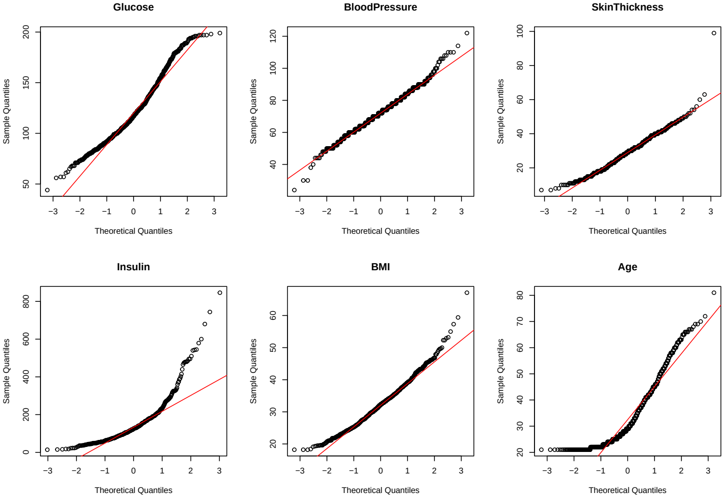
```
## Glucose : p-value = 1.720326e-11
## BloodPressure : p-value = 9.45138e-05
## SkinThickness : p-value = 1.775691e-09
## Insulin : p-value = 1.698218e-21
## BMI : p-value = 8.557785e-09
## Age : p-value = 2.402274e-24
```

Les données ne sont pas normalement distribuées. On va donc utiliser la corrélation de Spearman.

Analyse avec R : Histogrammes



Analyse avec R : Graphiques Q-Q



Analyse avec R : Corrélation de Spearman

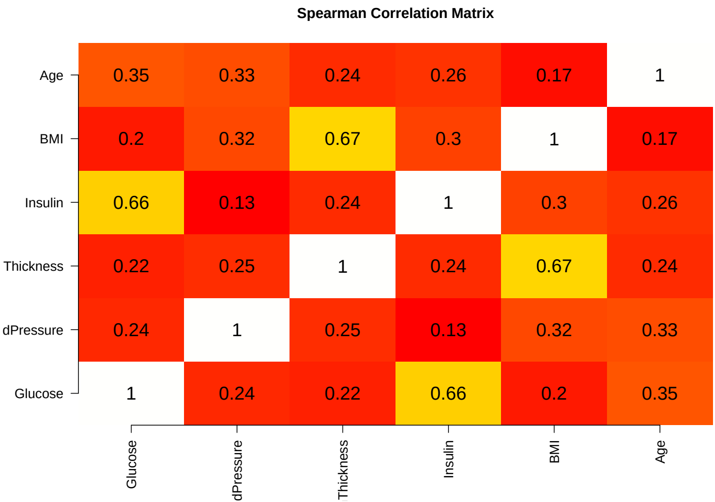
```
spearman_correlation_matrix <-
  cor(diabetes_subset,
      use="complete.obs",
      method="spearman")
```

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.

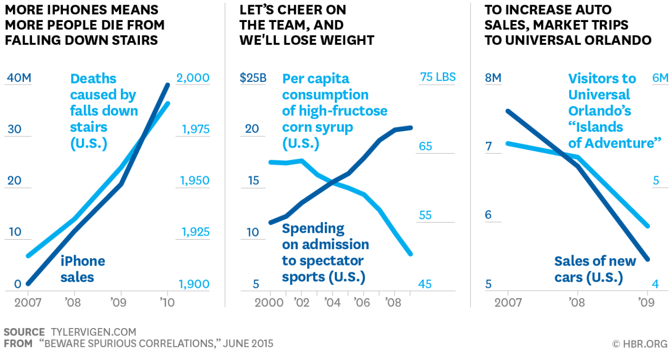
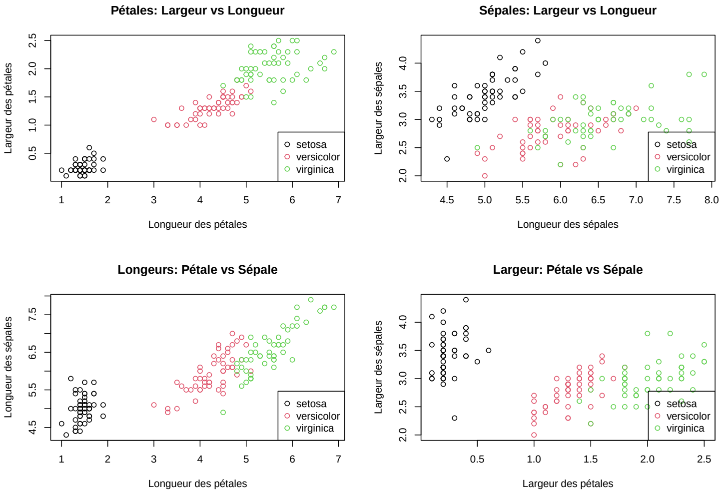


Figure 3: Corrélation fallacieuse

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

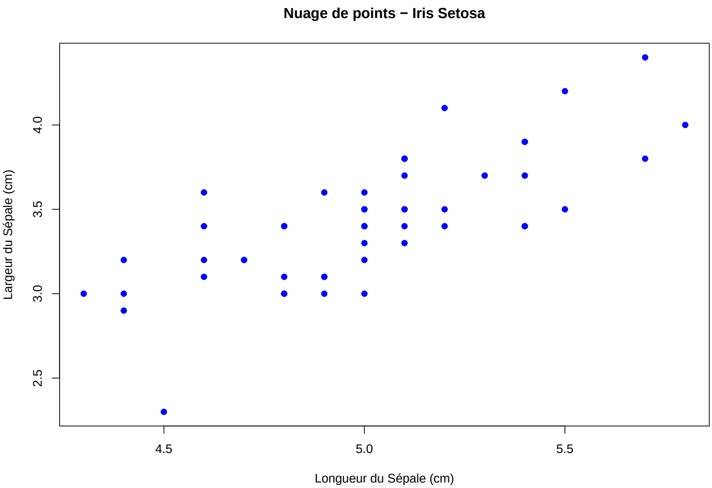
Test de normalité de Shapiro-Wilk

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

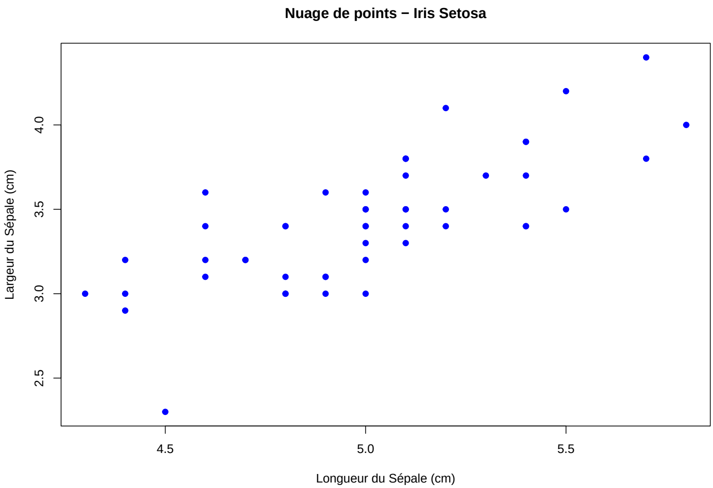
En incluant uniquement l'espèce de `setosa`

```
## 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

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

Lorsque les variables `x`et `y` sont normalement distribuées, on utilise la fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` pour calculer le coefficient de corrélation de Pearson entre deux variables `x` et `y`.

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

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

Analyse avec R : Corrélation

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

Lorsque les variables `x`et `y` sont normalement distribuées, on utilise la fonction `cor(x, y)` ou `cor(x, y, method = "pearson")` pour calculer 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

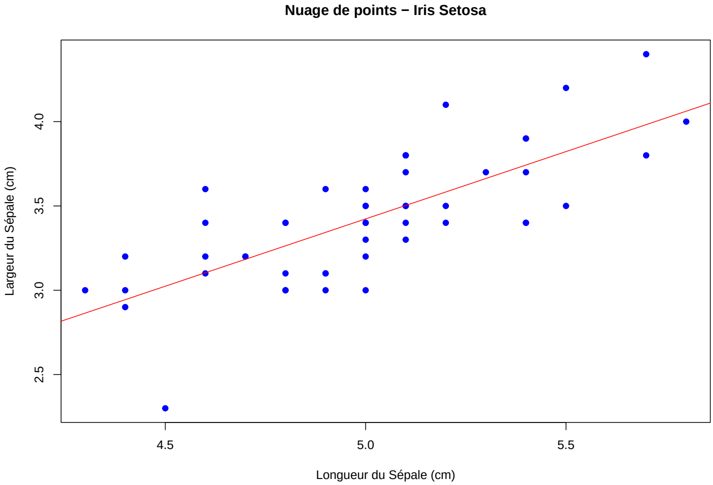
L'existence d'une corrélation entre deux variables n'implique pas une relation de cause à effet.

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

- Minimise 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 x et y est linéaire.
- **Indépendance** : Les observations sont indépendantes.
- **Homoscédasticité** : La variance du terme d'erreur ϵ est constante pour tous les niveaux de 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 ?

```
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 ?

```
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
```

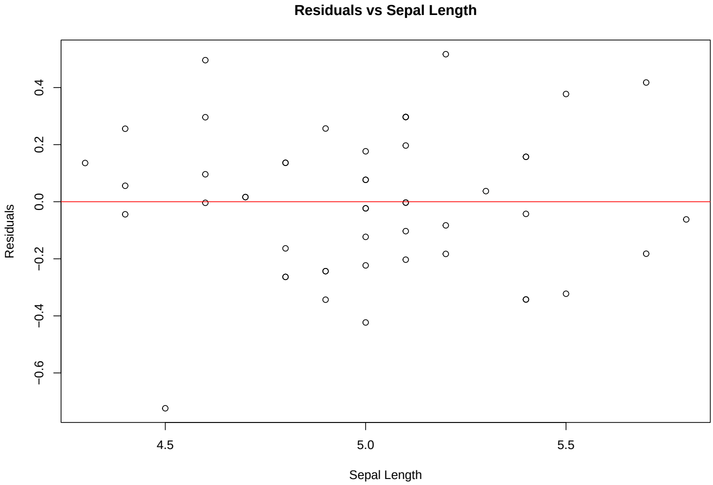
Et le carré du coefficient de corrélation de Pearson est . . .

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

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

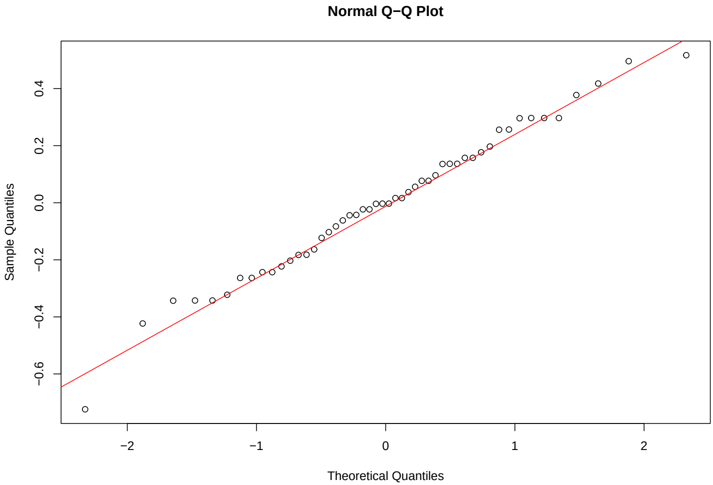
Analyse avec R : Homoscédasticité

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 0.85.



Régression linéaire multiple

Régression linéaire multiple

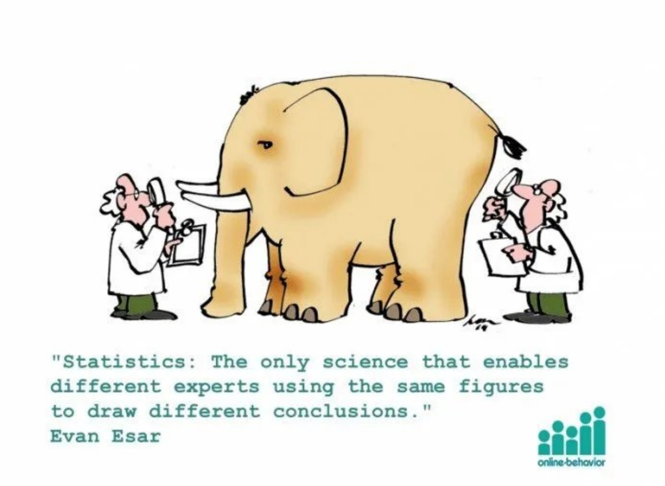


Figure 5: Brise-glace

Régression linéaire multiple

Représentation du modèle

- Rélation linéaire entre Y et plusieurs $X_1, X_2, ..., X_k$
- Formule : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

Régression linéaire multiple

Représentation du modèle

- Rélation linéaire entre Y et plusieurs $X_1, X_2, ..., X_k$
- Formule : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

Estimation des coefficients

- Utilisation de la méthode des moindres carrés.
- Minimisation de la somme des carrés des résidus.

Régression linéaire multiple

Représentation du modèle

- Rélation linéaire entre Y et plusieurs $X_1, X_2, ..., X_k$
- Formule : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

Estimation des coefficients

- Utilisation de la méthode des moindres carrés.
- Minimisation de la somme des carrés des résidus.

Interprétation des coefficients

- β_0 : Valeur de Y lorsque toutes les X sont nulles
- $\beta_1, \beta_2, ..., \beta_k$: Effet de chaque X sur Y

Régression linéaire multiple : Conditions

- Linéarité
- Indépendance
- Homoscédasticité
- Normalité des résidus
- **Absence de multicollinéarité** : Les variables indépendantes ne doivent pas être trop fortement corrélées entre elles.

Régression linéaire multiple : Inférence

- Des **tests d'hypothèse** sont effectués pour déterminer si les coefficients sont significativement différents de zéro, ce qui indique que le prédicteur correspondant a un effet statistiquement significatif sur la variable dépendante.

Régression linéaire multiple : Inférence

- Des **tests d'hypothèse** sont effectués pour déterminer si les coefficients sont significativement différents de zéro, ce qui indique que le prédicteur correspondant a un effet statistiquement significatif sur la variable dépendante.

- Des **intervalles de confiance** peuvent être construits pour les coefficients afin d'estimer leur précision.

Régression linéaire multiple : Inférence

- Des **tests d'hypothèse** sont effectués pour déterminer si les coefficients sont significativement différents de zéro, ce qui indique que le prédicteur correspondant a un effet statistiquement significatif sur la variable dépendante.

- Des **intervalles de confiance** peuvent être construits pour les coefficients afin d'estimer leur précision.

Évaluation et ajustement du modèle

- **R^2** mesure la proportion de la variance de la variable dépendante expliquée par les variables indépendantes.
- **R^2 ajusté** est également utilisé, en particulier lors de la comparaison de modèles avec un nombre différent de prédicteurs.

Régression linéaire multiple : R^2

Limitation du R^2

Un problème avec le R^2 est qu'il peut augmenter simplement en ajoutant plus de variables indépendantes au modèle, qu'elles soient significatives ou non. Cela peut conduire à un modèle surajusté (**overfitting**).

Régression linéaire multiple : R^2

Limitation du R^2

Un problème avec le R^2 est qu'il peut augmenter simplement en ajoutant plus de variables indépendantes au modèle, qu'elles soient significatives ou non. Cela peut conduire à un modèle surajusté (**overfitting**).

Pour surmonter cette limitation . . .

Le R^2 ajusté modifie le R^2 pour prendre en compte le nombre de prédicteurs dans le modèle.

Régression linéaire multiple : R^2 ajusté

Le R^2 ajusté est calculé comme suit :

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

où :

- n est le nombre d'observations.
- p est le nombre de variables indépendantes.

Régression linéaire multiple : R² ajusté

Le R^2 ajusté est calculé comme suit :

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

où :

- n est le nombre d'observations.
- p est le nombre de variables indépendantes.

Le R^2 ajusté peut être inférieur au R^2, et contrairement au R^2, il ne va pas automatiquement augmenter avec l'ajout de nouvelles variables.

Analyse avec R : Données et modèle

Base de données “Pima Indian Diabetes”

Notez que les données ne sont pas normalement distribuées. Cependant, à des fins de démonstration uniquement, on va utiliser les variables `Glucose`, `SkinThickness`, `Insulin`, `BMI` et `Age` pour prédire `BloodPressure`.

```
model1 <-
  lm(BloodPressure ~ Glucose + SkinThickness +
       Insulin + BMI + Age, data = data_filtered)
```

Analyse avec R : Données et modèle

Base de données “Pima Indian Diabetes”

Notez que les données ne sont pas normalement distribuées. Cependant, à des fins de démonstration uniquement, on va utiliser les variables `Glucose`, `SkinThickness`, `Insulin`, `BMI` et `Age` pour prédire `BloodPressure`.

```
model1 <-
  lm(BloodPressure ~ Glucose + SkinThickness +
       Insulin + BMI + Age, data = data_filtered)
```

Observez l’estimation et la valeur p pour chacun des β’s.

Analyse avec R : Test d'hypothèse

```
##
## Call:
## lm(formula = BloodPressure ~ Glucose + SkinThickness + Ins
##     BMI + Age, data = data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.158  -7.335  -0.615   7.864  30.349
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.971521   3.640612  10.979  < 2e-16 ***
## Glucose        0.045378   0.023977   1.893   0.0592 .
## SkinThickness -0.011414   0.074334  -0.154   0.8780
## Insulin       -0.009158   0.006013  -1.523   0.1285
## BMI            0.513417   0.111437   4.607 5.55e-06 ***
## Age            0.320817   0.060805   5.276 2.20e-07 ***
```

Analyse avec R : Interprétation, évaluation et ajustement

Les contributions de `SkinThickness` et `Insulin` au modèle ne sont pas significatives.

Analyse avec R : Interprétation, évaluation et ajustement

Les contributions de `SkinThickness` et `Insulin` au modèle ne sont pas significatives.

Selon l'analyse de corrélation ...

- `Glucose` et `Insulin` sont fortement corrélés
- `SkinThickness` et `BMI` sont fortement corrélés.

On peut choisir d'éliminer `SkinThickness` et `Insulin`.

```
model2 <-
  lm(BloodPressure ~ Glucose + BMI + Age,
     data = data_filtered)
```

Analyse avec R : Test d'hypothèse

```
##
## Call:
## lm(formula = BloodPressure ~ Glucose + BMI + Age, data = d
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -48.860  -7.141  -0.529   7.741  30.456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.64177    3.44230  12.097  < 2e-16 ***
## Glucose      0.02584    0.02029   1.273    0.204
## BMI          0.48534    0.08389   5.785 1.49e-08 ***
## Age          0.31734    0.06018   5.273 2.23e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
```

Analyse avec R : Intervalles de Confiance

```
##                    2.5 %      97.5 %
## (Intercept) 34.87386876 48.40967373
## Glucose     -0.01405801  0.06574061
## BMI          0.32039173  0.65027842
## Age          0.19902480  0.43566163
```

On a observé que les intervalles de confiance des variables significatives (`BMI` et `Age`) ne couvrent pas 0, mais celui de la variable insignifiante (`Glucose`) le fait.

Analyse avec R : R² et R² ajusté

```
round(summary(model1)$r.squared,5)
```

```
## [1] 0.17919
```

```
round(summary(model1)$adj.r.squared,5)
```

```
## [1] 0.16856
```

```
round(summary(model2)$r.squared,5)
```

```
## [1] 0.17422
```

```
round(summary(model2)$adj.r.squared,5)
```

```
## [1] 0.16783
```

Travaux pratiques

Travaux pratiques

Continuez à travailler avec la base de données “Pima Indian Diabetes”.

Définissez une variable dépendante et utilisez le reste comme variables indépendantes pour créer un modèle de régression linéaire.

Évaluez et ajustez le modèle pour atteindre un bon R^2.

Quelles variables apportent une contribution significative au modèle et dans quelle direction ?