SYS865 Inférence statistique avec programmation R

Ornwipa Thamsuwan

20 mars 2024

Plan de la séance

Plan de la séance

- Régression logistique
- Confondeur

Récap et matière à réfléxion

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

R code

```
data <- read.csv("diabetes.csv")
selected_columns <- data[, 2:6]
rows_with_zero <- apply(selected_columns, 1,
                        function(x) any(x == 0))
data_cleaned <- data[!rows_with_zero, ]
names(data_cleaned)[
  names(data_cleaned) ==
    "DiabetesPedigreeFunction"] <- "DbtPdgFunc"
```

Recap : Modèle complète

```
model_full <- lm(Outcome ~ Pregnancies + Glucose +
                   BloodPressure + SkinThickness +
                   Insulin + BMI + DbtPdgFunc + Age,
                 data = data_cleaned)
round(summary(model_full)$coefficients, 4)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.1027     0.1436 -7.6806   0.0000
## Pregnancies       0.0130     0.0084  1.5486   0.1223
## Glucose           0.0064     0.0008  7.8550   0.0000
## BloodPressure     0.0001     0.0017  0.0316   0.9748
## SkinThickness     0.0017     0.0025  0.6652   0.5063
## Insulin          -0.0001     0.0002 -0.6031   0.5468
## BMI               0.0093     0.0039  2.3907   0.0173
## DbtPdgFunc        0.1572     0.0580  2.7083   0.0071
## Age               0.0059     0.0028  2.1090   0.0356
```

Recap : Modèle ajusté

En supprimant les variables non importantes `BloodPressure`, `SkinThickness` et `Insulin` ...

```
model_reduced <- lm(Outcome ~ Pregnancies + Glucose
                            BMI + DbtPdgFunc + Age,
                          data = data_cleaned)
round(summary(model_reduced)$coefficients, 4)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.0908     0.1174 -9.2876   0.0000
## Pregnancies      0.0136     0.0083  1.6387   0.1021
## Glucose          0.0062     0.0007  8.9698   0.0000
## BMI              0.0108     0.0029  3.7636   0.0002
## DbtPdgFunc       0.1578     0.0574  2.7483   0.0063
## Age              0.0059     0.0027  2.1739   0.0303
```

Recap : Comparaison des modèles par R^2 et R^2 ajusté

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

```
## [1] 0.3457734
```

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

```
## [1] 0.3443796
```

Recap : Comparaison des modèles par R^2 et R^2 ajusté

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

```
## [1] 0.3457734
```

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

```
## [1] 0.3443796
```

```r
summary(model_full)$adj.r.squared
```

```
## [1] 0.3321081
```

```r
summary(model_reduced)$adj.r.squared
```

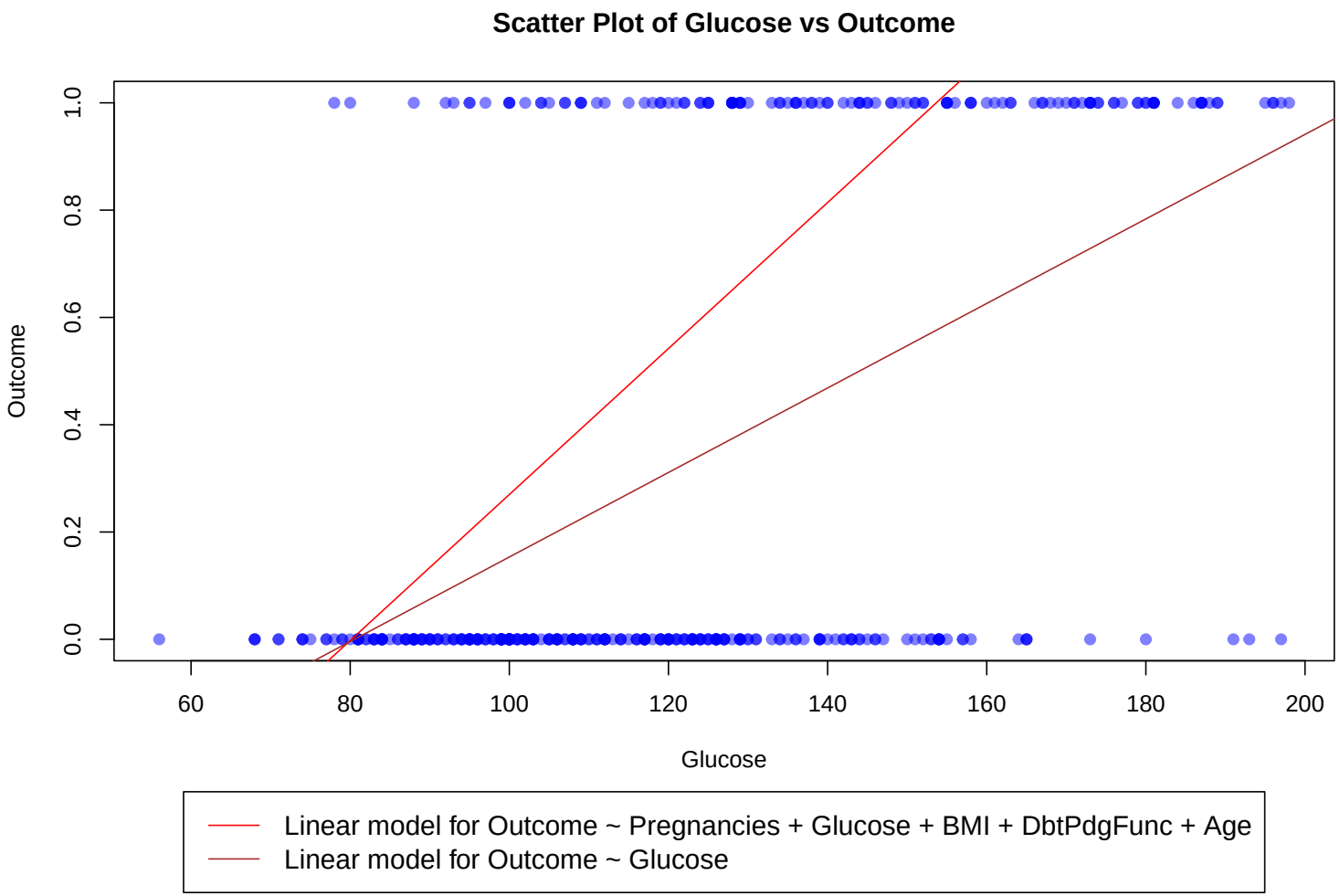
```
## [1] 0.3358872
```

Recap : Intervalles de confiance de β's

```
round(confint(model_reduced, level = 0.95), 4)
```

```
##                 2.5 %  97.5 %
## (Intercept) -1.3217 -0.8599
## Pregnancies -0.0027  0.0299
## Glucose      0.0048  0.0075
## BMI          0.0051  0.0164
## DbtPdgFunc   0.0449  0.2708
## Age          0.0006  0.0113
```

Recap : Visualisation des résultats



La réponse (`y` ou `Outcome`) n'est pas une variable continues, mais binaire, soit 0 ou 1 et non une valeur intermédiare.

Attention!

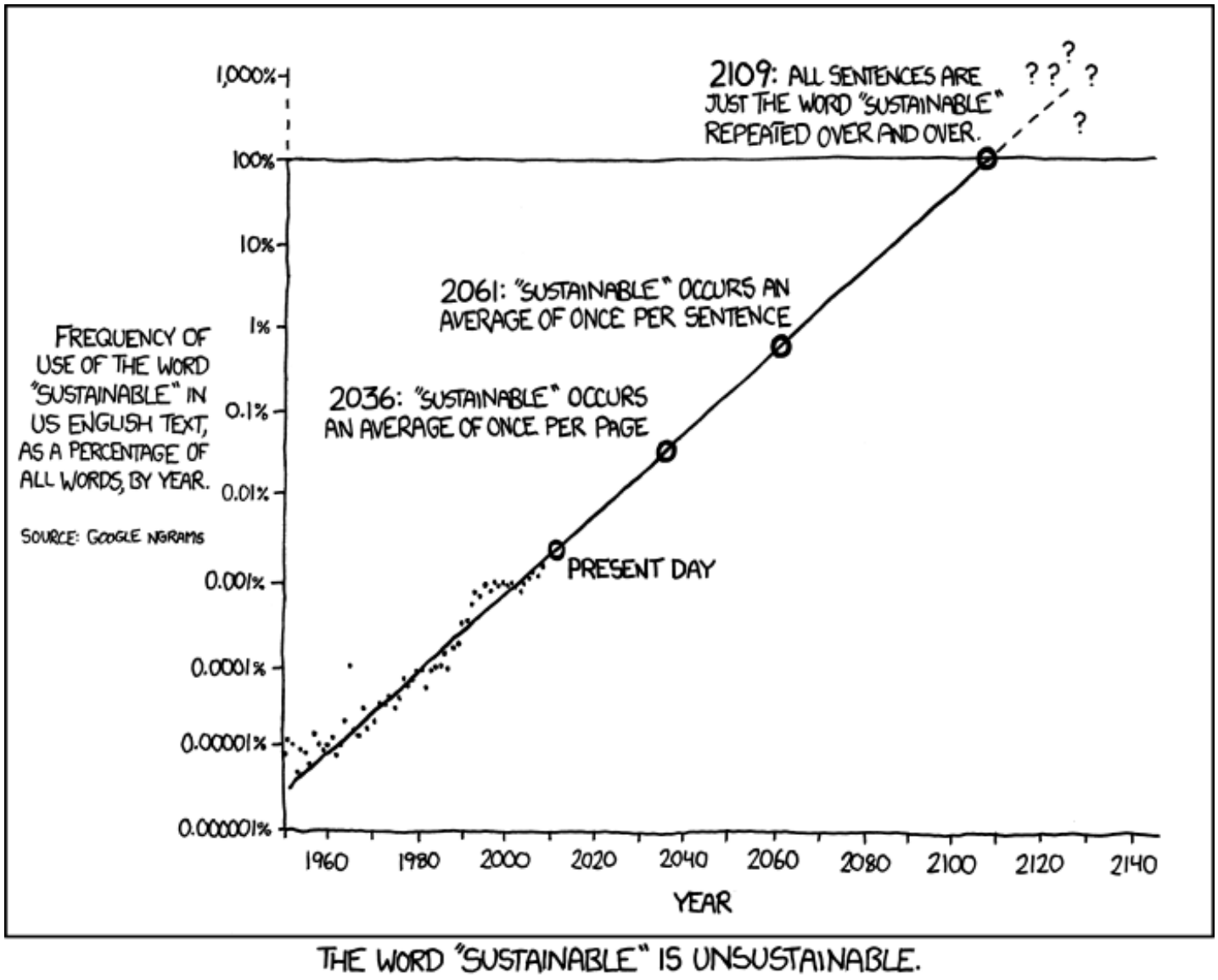
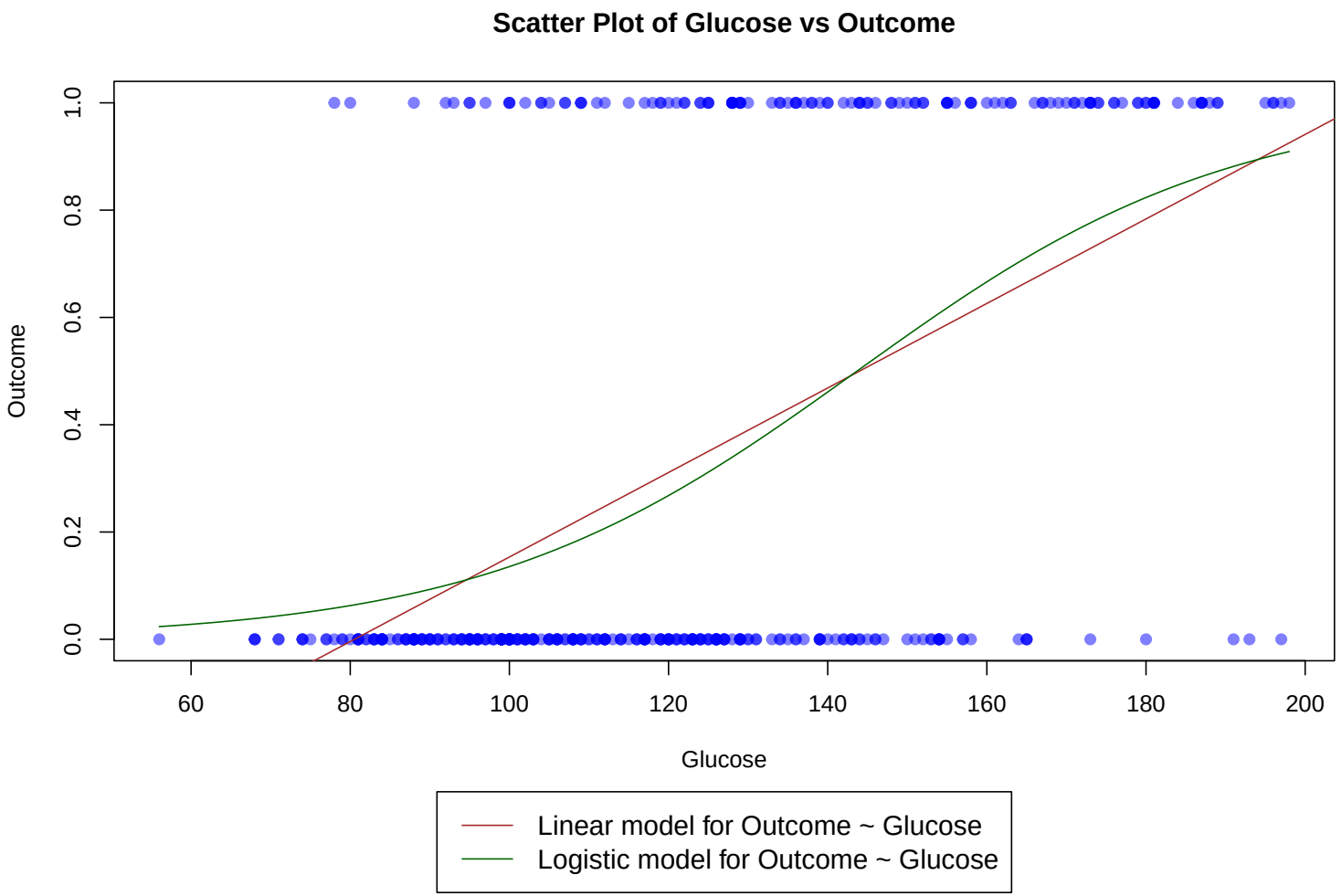


Figure 1: Extrapolation - "Sustainable is unsustainable."

Recap : Visualisation des résultats



Une alternative est la régression logistique, fournissant un résultat sous forme de probabilité que y soit 0 ou 1.

Introduction au nouveau sujet

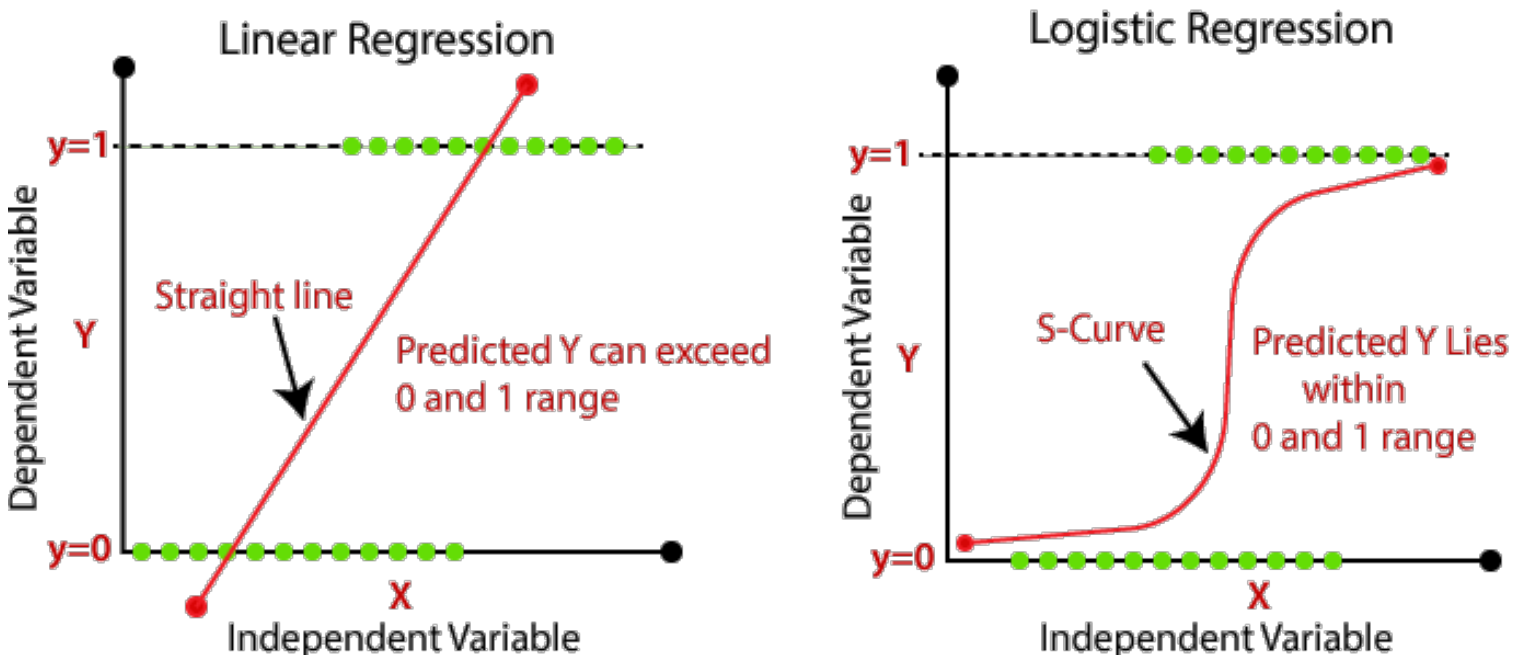


Figure 2: Régression linéaire vs. logistique

Régression logistique

Régression logistique

La régression logistique modélise la probabilité d'un résultat binaire basée sur une ou plusieurs variables prédictives. Cela est particulièrement utile lorsque la variable dépendante ne peut prendre que deux résultats possibles (succès ou échec).

Régression logistique

La régression logistique modélise la probabilité d'un résultat binaire basée sur une ou plusieurs variables prédictives. Cela est particulièrement utile lorsque la variable dépendante ne peut prendre que deux résultats possibles (succès ou échec).

Le modèle de régression logistique est basé sur **la fonction logit, le logarithme naturel du rapport de cotes**.

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

- p est la probabilité d'une des issues,
- $X_1, X_2, ..., X_k$ sont les variables prédictives.
- $\beta_1, \beta_2, ..., \beta_k$ représentent le changement dans le log des cotes de l'issue pour un changement unitaire dans les variables prédictives.

Régression logistique

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Régression logistique

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Méthode d'Estimation

- Les coefficients sont estimés en utilisant l'Estimation du Maximum de Vraisemblance (MLE).
- Cette méthode trouve les coefficients qui maximisent la vraisemblance d'observer les données de l'échantillon.

Régression logistique

Inférence sur les Coefficients : Les tests d'hypothèse sur $\beta_1, \beta_2, ..., \beta_k$ sont réalisées pour déterminer si les prédicteurs sont significativement associés à l'issue.

Méthode d'Estimation

- Les coefficients sont estimés en utilisant l'Estimation du Maximum de Vraisemblance (MLE).
- Cette méthode trouve les coefficients qui maximisent la vraisemblance d'observer les données de l'échantillon.

Interprétation en Rapport de Cotes : Un rapport de cotes supérieur à 1 indique une augmentation des cotes de l'issue avec une augmentation unitaire du prédicteur, et vice versa.

Critères d'information

Critère d'Information d'Akaike (AIC)

- L'AIC est une mesure de la qualité relative d'un modèle statistique pour un ensemble de données.
- Basé sur le concept d'**entropie d'information**, l'AIC offre un équilibre entre la complexité du modèle (nombre de paramètres) et l'adéquation du modèle.
- Formule de l'AIC : $AIC = 2k - 2\ln(L)$
  - k est le nombre de paramètres dans le modèle et
  - L est la vraisemblance du modèle.
- Une valeur AIC plus basse indique un meilleur modèle.
- L'AIC pénalise les modèles pour leur complexité, aidant ainsi à éviter le surajustement.
- Lors de la comparaison de modèles, la valeur absolue de l'AIC n'est pas aussi importante que la différence entre les valeurs AIC de différents modèles.
- Des modèles avec un AIC différant de plus de 2 sont généralement considérés comme ayant des preuves substantielles contre le modèle avec l'AIC le plus élevé.

Critère d'Information Bayésien (BIC)

- Le BIC est similaire à l'AIC mais introduit une pénalité plus forte pour le nombre de paramètres dans le modèle.
- Le BIC est dérivé de la probabilité bayésienne et utilisé pour la sélection de modèles.
- Formule du BIC : $BIC = \ln(n)k - 2\ln(L)$
  - n est le nombre d'observations,
  - k est le nombre de paramètres, et
  - L est la vraisemblance du modèle.
- Comme l'AIC, une valeur BIC plus basse indique un meilleur modèle.
- Le BIC a tendance à pénaliser plus lourdement la complexité que l'AIC, surtout à mesure que la taille de l'échantillon augmente.
- La règle de décision pour comparer les modèles avec le BIC est similaire à l'AIC.
- Une différence de 6 ou plus est considérée comme une preuve forte contre le modèle avec le BIC le plus élevé.

AIC vs BIC

- L'AIC est plus axé sur la recherche du modèle qui explique le mieux les données, tandis que le BIC tente de trouver le véritable modèle parmi l'ensemble des candidats.
- En pratique, l'AIC peut être meilleur pour les modèles prédictifs, tandis que le BIC est plus utile pour la modélisation explicative, en particulier dans de grands ensembles de données.
- Il est essentiel de se rappeler que l'AIC et le BIC sont des mesures comparatives, utiles pour comparer différents modèles sur le même ensemble de données, mais pas pour évaluer la qualité absolue d'un modèle unique isolément.

Récap et matière à réfléxion

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

Récap et matière à réfléxion

Base de données “Pima Indian Diabetes”

- Variable dépendante : `Outcome`
- Variables indépendantes : `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` et `Age`

Corrélations modérées entre variables indépendantes

```
cor(data_cleaned$Pregnancies, data_cleaned$Age)
```

```
## [1] 0.6796085
```

```
cor(data_cleaned$Glucose, data_cleaned$Insulin)
```

```
## [1] 0.581223
```

```
cor(data_cleaned$SkinThickness, data_cleaned$BMI)
```

```
## [1] 0.6643549
```

Recap :

Confondeur

Travaux pratiques