الكلية متعددة التخصصات - ورزازات
FACULTE POLYDISCIPLINAIRE DE OUARZAZATE

AIC

L'Akaike Information Criterion

Présenter par:

- MOKNIA Youssef

Encadrer par:

- Mr. EL OUALI Morad

Plan

Introduction

Les Fondations Théoriques de l’AIC

Comparaison entre l'AIC et le BIC

Exemple pratique

Conclusion

Introduction

Hirotugu Akaike

1927 - 2009

Statistical Mathematician

1973

AIC est largement utilisé pour évaluer et comparer différents modèles statistiques, en prenant en compte à la fois **la qualité de l'ajustement** et **la simplicité** du modèle.

Les Fondations Théoriques de l’AIC

Formule Générale de l'AIC

L'AIC est défini par la formule suivante :

$$AIC = 2k - 2\ln(\mathcal{L}(\hat{\theta}))$$

où :

- $\ln\left(\mathcal{L}(\hat{\theta})\right)$ est **la vraisemblance maximale**, qui mesure à quel point le modèle s'ajuste bien aux données.
- k est **le nombre de paramètres** dans le modèle, ce qui représente sa complexité

*Plus **l'AIC** est faible, meilleur est le modèle pour les données.*

Les Fondations Théoriques de l’AIC

Explication de la formule :

Nombre de paramètres

- Le terme $\mathbf{k}$ représente le nombre de paramètres dans notre modèle. Un modèle trop complexe, avec beaucoup de paramètres, pourrait **s'ajuster** parfaitement aux données d'entraînement mais échouer à généraliser sur de nouvelles données.

  Exemple : *Si pour prédire la température, on ajoute des dizaines de facteurs (heure, humidité, vent, etc.), on risque de faire un modèle très complexe, mais qui ne sera pas capable de bien prédire la température d’autres jours.*

- $\mathbf{2k}$ C'est la partie qui pénalise la complexité du modèle. Plus il y a de paramètres dans le modèle, plus cette valeur augmente. Cela empêche un modèle d’ajouter trop de paramètres juste pour mieux s’ajuster aux données.

Les Fondations Théoriques de l'AIC

Explication de la formule :

La vraisemblance

La vraisemblance mesure à quel point le modèle s'ajuste bien aux données observées. $\mathcal{L}$ Plus est grande, mieux le modèle explique les données.

Pour un ensemble de données $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n})$, la vraisemblance est donnée par :

$$L(\theta) = P(X|\theta)$$

C'est la probabilité d'observer **les données $\boldsymbol{X}$**, sachant **les paramètres $\boldsymbol{\theta}$**

Pourquoi utiliser la log-vraisemblance ln(L) :

Nous utilisons souvent le logarithme de la vraisemblance pour simplifier les calculs. La vraisemblance peut être un nombre très petit (surtout si on a beaucoup de données), et travailler avec son logarithme permet de simplifier les opérations et d'éviter les erreurs numériques.

Les Fondations Théoriques de l’AIC

Problème de l’AIC :

- *Lorsque* **la taille de l'échantillon** (n) *est petit par rapport au* ***nombre de paramètres*** k*, l'AIC peut sous-estimer la pénalité pour la complexité du modèle.*
- *Cela conduit souvent à sélectionner des modèles trop complexes, car la vraisemblance maximale peut être surévaluée pour des petits échantillons.*

Les Fondations Théoriques de l’AIC

AIC corrigé (AICc)

Pour pallier ce problème, une version corrigée appelée **AICc** a été introduite. **L'AICc** ajuste la pénalité pour la complexité en fonction de la taille de l'échantillon, avec une formule donnée par :

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}$$

- n : taille de l'échantillon.
- k : nombre de paramètres estimés.

*Cette correction est particulièrement importante lorsque n est proche de k. Si $n \gg k$, le terme correctif devient négligeable, et **l'AICc** converge vers l'AIC classique.*

Comparaison entre l'AIC et le BIC

Maintenant que nous avons compris l'**AIC**, il est important de le comparer à un autre critère très utilisé en modélisation statistique : le **BIC**, ou **Critère d'Information Bayésien**.

Comparaison entre l'AIC et le BIC

Formule du BIC

Comme l'AIC, le BIC a une formule simple :

$$BIC = k \cdot \ln(n) \ - 2 \ln(\mathcal{L}(\hat{\theta}))$$

n est le nombre d'observation dans les données.

- *Comparé à l'AIC, le BIC inclut le terme $\boldsymbol{ln(n)}$, qui est directement lié au nombre d'observations n.*
- *Ce terme augmente la pénalisation lorsque le nombre d'observations est élevé, ce qui signifie que le BIC tend à préférer des modèles plus simples que l'AIC lorsque la taille des données augmente.*

Quand utiliser le AIC ?

- Lorsque nous voulons minimiser le risque de **sous-ajustement (underfitting).**
- Lorsque nous travaillons avec un nombre modéré d'observations et que nous sommes prêts à accepter des modèles un peu plus complexes.

BIC ?

Quand utiliser le BIC ?

- Lorsque nous avons un **grand nombre d'observations** et que nous souhaitons fortement pénaliser les modèles trop complexes.
- Lorsque notre priorité est de minimiser le risque de **sur-ajustement (overfitting).** En choisissant des modèles plus simples

Exemple pratique

Contexte de l'exemple

Nous avons un jeu de données avec les variables suivantes :

- **Prix** (en dirhams)
- **Publicité** (en dirhams)
- **Emplacement** (qualité de l'emplacement, sur une échelle de 1 à 5)
- **Ventes** (en unités, la variable cible que nous voulons prédire)

Jeu de données (fictif)

Prix	Publicité	Emplacement	Ventes
200	1500	3	500
220	1600	4	530
180	1200	2	450
250	1800	5	600
230	1700	4	550
210	1400	3	520
190	1300	2	480

Exemple pratique

Contexte de l'exemple

Nous voulons comparer deux modèles de régression linéaire pour prédire les **ventes** à partir de différentes combinaisons de variables explicatives :

1

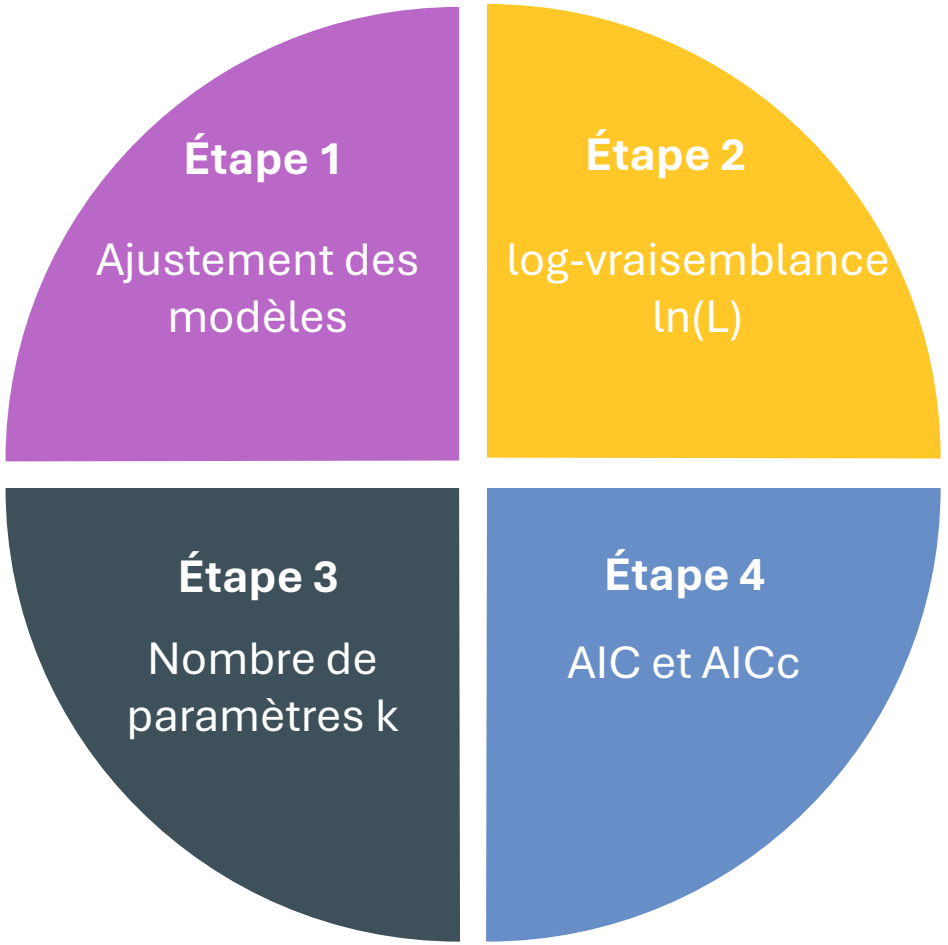
$$Ventes = f(Prix)$$

2

$$Ventes = f(Prix, Publicité, Emplacement)$$

Exemple pratique

Étapes de la sélection du modèle avec AIC



Exemple pratique

Ajustement des modèles

On commence par ajuster les deux modèles avec nos données.

- **Modèle 1** : Régression simple où les **ventes** dépendent uniquement du **prix**.

$$\text{Ventes} = \beta_0 + \beta_1 \cdot \text{Prix}$$

- **Modèle 2** : Régression multiple où les **ventes** dépendent du **prix**, de la **publicité**, et de l'**emplacement**.

$$\text{Ventes} = \beta_0 + \beta_1 \cdot \text{Prix} + \beta_2 \cdot \text{Publicité} + \beta_3 \cdot \text{Emplacement}$$

Exemple pratique

Calcul de la log-vraisemblance ln(L)

En **régression linéaire**, on fait souvent l'hypothèse que les erreurs résiduelles (la différence entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$) suivent une **distribution normale**.

La fonction de densité de probabilité d'une variable aléatoire suivant une loi normale est donnée par la formule suivante :

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Où :

- x est la variable.
- μ est la moyenne (ou espérance) de la distribution.
- σ^2 est la variance de la distribution.
- $\exp$ représente la fonction exponentielle.

Exemple pratique

Calcul de la log-vraisemblance ln(L)

Si les erreurs suivent une loi normale, la **probabilité conditionnelle** de l'observation y_i donnée l'entrée x_i et les paramètres du modèle θ est donnée par :

$$P(y_i|x_i,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right)$$

Cela représente la probabilité que le modèle prédise la valeur $\mathbf{y_i}$ avec un certain **écart $\mathbf{y_i - \hat{y}_i}$**, sachant que cet écart suit une distribution normale avec une **variance $\mathbf{\sigma^2}$**.

Exemple pratique

Calcul de la log-vraisemblance ln(L)

La vraisemblance totale pour un ensemble de données est obtenue en multipliant ces probabilités pour chaque observation i. La fonction de vraisemblance devient donc :

$$L(\theta) = \prod_{i=1}^{n} P(y_i|x_i, \theta)$$

Avec

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right)$$

La **log-vraisemblance** s'écrit :

$$\ln(L(\theta)) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma^2}\right)\right)$$

Simplifie

$$\ln(L(\theta)) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Exemple pratique

Calcul de la log-vraisemblance ln(L)

❖ **Modele 1 :**

Les prévisions obtenues pour le Modèle 1 sont les suivantes :

Ventes observées	Ventes prédites $\hat{y}_i$	Erreurs résiduelles $(y_i - \hat{y}_i)$
500	510	-10
530	525	5
450	460	-10
600	590	10
550	540	10
520	515	5
480	470	10

Exemple pratique

Calcul de la log-vraisemblance ln(L)

❖ **Modele 1 :**

Supposons que σ^2 =25. Nous calculons la log-vraisemblance :

$$\ln(L(\theta_1)) = -\frac{7}{2}\ln(2\pi \times 25) - \frac{1}{2 \times 25}\sum_{i=1}^{7}(y_i - \hat{y}_i)^2$$

En calculant la somme des carrés des résidus :

$$\sum_{i=1}^{7}(y_i - \hat{y}_i)^2 = (-10)^2 + (5)^2 + (-10)^2 + (10)^2 + (10)^2 + (5)^2 + (10)^2 = 500$$

Ainsi, la log-vraisemblance devient :

$$\ln(L(\theta_1)) = -\frac{7}{2}\ln(2\pi \times 25) - \frac{500}{50}$$

Exemple pratique

Calcul de la log-vraisemblance ln(L)

❖ **Modèle 1**:

$$\ln(L(\theta_1)) \approx -27.70$$

On faire la même chose pour le **Modèle 2.**

❖ **Modèle 2**:

$$\ln(L(\theta_2)) \approx -23.70$$

*Cela signifie que le **Modèle 2** a une vraisemblance plus élevée, ce qui indique qu'il ajuste mieux les données que le Modèle 1.*

Exemple pratique

Calcul du nombre de paramètres k

Le nombre de paramètres inclut le coefficient d'intercept (β_0) et les coefficients pour chaque variable explicative

1. **Modèle 1** : $Ventes = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot Prix$
   - Il y a 1 paramètre pour le coefficient du Prix.
   - Plus, 1 paramètre pour l'ordonnée à l'origine (l'intercept).
   - Donc, $\boldsymbol{k_1 = 2}$.

2. **Modèle 2** : $Ventes = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot Prix + \boldsymbol{\beta_2} \cdot Publicité + \boldsymbol{\beta_3} \cdot Emplacement$
   - Il y a 3 paramètres pour les coefficients du Prix, de la Publicité, et de l'Emplacement.
   - Plus, 1 paramètre pour l'ordonnée à l'origine (l'intercept).
   - Donc, $\boldsymbol{k_2 = 4}$.

Exemple pratique

Calcul de l'AIC

Donc La valeur de l'AIC pour chaque modèle est :

1. **Modèle 1**

$$AIC_1 = 2 \times 2 - 2 \times (-27.70) = 4 + 55.40 = 59.40$$

2. **Modèle 2**

$$AIC_2 = 2 \times 4 - 2 \times (-23.70) = 8 + 47.40 = 55.40$$

*Le **Modèle 2** a un AIC plus faible **(55,40)**, donc c'est le modèle préféré selon l'AIC, car il atteint un meilleur compromis entre ajustement et complexité.*

Exemple pratique

Calcul de l'AIC corrige

Donc La valeur de l'AICc pour chaque modèle est :

1. **Modèle 1**

$AIC_1 = 59.40$

$k_1 = 2$ (1 paramètre pour le Prix + 1 intercept)

$n = 7$

$$AICc_1 = 59.40 + \frac{2 \times 2 \times (2+1)}{7-2-1}$$

$$AICc_1 = 59.40 + \frac{12}{4} = 59.40 + 3 = 62.40$$

Exemple pratique

Calcul de l'AIC corrige

Donc La valeur de l'AICc pour chaque modèle est :

2. Modèle 2

- $AIC_2 = 55.40$
- $k_2 = 4$ (3 paramètres pour Prix, Publicité, Emplacement + 1 intercept)
- $n = 7$

$$AICc_2 = 55.40 + \frac{2 \times 4 \times (4+1)}{7-4-1}$$

$$AICc_2 = 55.40 + \frac{40}{2} = 55.40 + 20 = 75.40$$

Exemple pratique

Calcul de l’AIC corrige

Donc :

- **AICc du Modèle 1 = 62.40**
- **AICc du Modèle 2 = 75.40**

*Bien que **l'AIC** initial favorise **le Modèle 2**, après correction pour petits échantillons, **le Modèle 1** devient meilleur, car **l'AICc** pénalise davantage le modèle plus complexe avec peu d'observations.*

Conclusion

L'AIC est un outil puissant pour comparer des modèles en tenant compte à la fois de ***l'ajustement*** *et de* ***la complexité****. En pénalisant les modèles trop complexes, il nous aide à éviter* ***le sur-ajustement****, tout en favorisant ceux qui expliquent bien les données avec un minimum de paramètres.*

Merci Pour Votre Attention

MOKNIA Youssef