الكلية متعددة التخصصات-ورزازات
ⵜⴰⵖⵔⵉⵎⵜ ⵜⴰⵣⵜⵔⴰⴼⵜ – ⵡⵔⵣⴰⵣⴰⵜ
FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE
UNIVERSITÉ IBN ZOHR - AGADIR

Choix de modèle via le critère

Cp de Mallow

Présenté par : Salma EL GOUDE **Sous la direction de :** M.Mourad EL OUALI

Plan

01-Introduction

Le Cp de Mallows est un critère essentiel en statistique, utilisé pour évaluer la qualité d'un modèle de régression. Conçu par Colin Lingwood Mallows en 1973, cet indicateur aide à trouver un équilibre entre la complexité d'un modèle, mesurée par le nombre de paramètres qu'il utilise, et sa capacité à fournir des prédictions précises. Dans le contexte de la sélection de modèles, où plusieurs variables prédictives peuvent être utilisées pour expliquer un résultat, le Cp de Mallows permet de comparer différents modèles afin d'identifier celui qui optimise cet équilibre.

02-Cp comme estimation de Γp

Le meilleur modèle est simplement celui qui minimise Γp. On sait même que la valeur théorique minimale de Γp est égale au nombre de paramètres k+1.

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}\sigma^2_{\hat{y}_i} + \sum_{i=1}^{n}\left[E(\hat{y}_i) - E(y_i)\right]^2\right\}$$

Remarque :

Γp repose sur des quantités théoriques souvent inconnues, comme σ², $\sigma^2_{\hat{y}_i}$, et Γp lui-même. Comme ces paramètres grecs représentent des quantités de population, il est impossible de connaître leur valeur exacte.

C'est ici qu'intervient le critère Cp de Mallows : il permet d'estimer Γp en utilisant les données disponibles.

02-Cp comme estimation de Γp

✳ Calcul avec σ2

Si la variance de la population σ² est connue, on peut estimer Γp à l'aide de Cp :

$$C_p = k + 1 + \frac{(MSE_k - \sigma^2)(n - k - 1)}{\sigma^2}$$

où :

- k est le nombre de prédicteurs dans le modèle.
- MSE_k est l'erreur quadratique moyenne du modèle avec k prédicteurs. Il est représenté comme suit :

$$MSE = \frac{SSE}{n-k-1} = \frac{\sum(y_i - \hat{y}_i)^2}{n-k-1}$$

- σ² est la variance de la population, que nous ne connaissons généralement pas.
- n est le nombre total d'observations.

02-Cp comme estimation de Γp

✳ Calcul avec σ² inconnue :

Mais comme σ² est inconnu, on l'estime en utilisant MSE_{all}, l'erreur quadratique moyenne du modèle complet avec tous les prédicteurs candidats. Cela donne :

$$C_p = k + 1 + \frac{(MSE_k - MSE_{\text{all}}) \cdot (n - k - 1)}{MSE_{\text{all}}}$$

$$= k + 1 + \frac{(n - k - 1) \cdot MSE_k}{MSE_{\text{all}}} - (n - k - 1)$$

$$= \frac{SSE_k}{MSE_{\text{all}}} + 2(k + 1) - n$$

Où

- SSE_k est la somme des carrés des résidus pour le modèle avec k prédicteurs.

$$SSE_k = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i,k})^2$$

03-Formule générale de Cp

La statistique Cp pour un modèle de régression, en fonction du nombre de prédicteurs P sélectionnés parmi un ensemble de K prédicteurs, est définie par la formule suivante :

$$C_p = \frac{SSE_p}{S^2} - N + 2(P+1)$$

où :

- SSE_p est la somme des carrés des résidus pour le modèle avec P prédicteurs.
- S² est l'estimation de la variance des erreurs pour le modèle complet (celui avec K prédicteurs),

$$S^2 = \frac{1}{N-K}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

- N est la taille de l'échantillon
- P est le nombre de prédicteurs inclus dans le modèle.

04-Formule alternative de Cp

Définition alternative de Cp qui repose sur une approche légèrement différente, mais qui conserve le même objectif : sélectionner un modèle de régression qui minimise l'erreur de prédiction tout en évitant la surcomplexité. Elle est définie comme suit :

$$C_p = \frac{1}{n}(\mathrm{RSS} + 2p\hat{\sigma}^2)$$

Où :

- RSS est la somme des carrés des résidus sur l'échantillon de données d'entraînement,
- p est le nombre de prédicteurs dans le modèle,
- $\hat{\sigma}^2$ est une estimation de la variance associée à chaque réponse dans le modèle linéaire, estimée à partir d'un modèle contenant tous les prédicteurs.

05-Utilisation du critère Cp

✳ pour la sélection des ”meilleurs” modèles

Pour choisir le bon modèle selon Cp, il suffit d'analyser ses valeurs et leur proximité par rapport à k+1, avec k le nombre de prédicteurs dans le modèle.

- Modèles avec un faible Cp : Les modèles présentant une faible valeur de Cp suggèrent une variation totale estimée plus faible des réponses prédites, ce qui indique un modèle plus précis.

05-Utilisation du critère Cp

✳ pour la sélection des ”meilleurs” modèles

Pour choisir le bon modèle selon Cp, il suffit d'analyser ses valeurs et leur proximité par rapport à k+1, avec k le nombre de prédicteurs dans le modèle.

- Interprétation des valeurs de Cp :
  - ✳ Lorsque Cp est proche de k +1, le biais du modèle est considéré comme faible, ce qui signifie que le modèle atteint un bon équilibre entre la variance et le biais.
  - ✳ Si Cp est significativement supérieur à k + 1, cela indique un biais élevé, ce qui peut suggérer une mauvaise sélection des prédicteurs ou un surajustement.
  - ✳ Lorsque Cp est inférieur à k + 1, cela est souvent dû à une erreur d’échantillonnage et doit être interprété comme une absence de biais significatif.

05-Utilisation du critère Cp

✳ pour la sélection des ”meilleurs” modèles

Pour choisir le bon modèle selon Cp, il suffit d'analyser ses valeurs et leur proximité par rapport à k+1, avec k le nombre de prédicteurs dans le modèle.

- Pour le modèle complet, c’est-à-dire celui qui inclut tous les prédicteurs candidats, Cp est toujours égal à k + 1. Par conséquent, Cp ne doit pas être utilisé pour évaluer le modèle complet, car son but est d’identifier des sous-ensembles de prédicteurs optimaux.

06-Stratégie pour l'utilisation du critère Cp

Pour sélectionner les "meilleurs" modèles à l'aide du critère Cp, il est recommandé de suivre cette stratégie :

01

Identifier les sous-ensembles de prédicteurs dont Cp est proche de k+1: Ces modèles minimisent le biais, ce qui suggère une bonne adéquation entre le modèle et les données.

02

Ne pas choisir le modèle complet en fonction de Cp: Le modèle complet aura toujours Cp égal à k + 1, ce qui n'apporte pas d'information utile pour la sélection. Le modèle complet peut inclure des prédicteurs superflus qui augmentent la complexité sans amélioration significative de la performance prédictive.

03

Si tous les modèles, sauf le modèle complet, présentent un Cp élevé, cela peut indiquer qu'il manque des prédicteurs essentiels dans l'analyse. Dans ce cas, il est conseillé de réexaminer les variables pour identifier celles qui pourraient être omises.

06-Stratégie pour l'utilisation du critère Cp

Pour sélectionner les "meilleurs" modèles à l'aide du critère Cp, il est recommandé de suivre cette stratégie :

04

Lorsque plusieurs modèles ont des valeurs de Cp proches de k + 1, il convient de sélectionner le modèle avec la plus faible valeur de Cp. Cela assure une minimisation conjointe du biais et de la variance.

05

Si plusieurs modèles ont un Cp faible et proche de k + 1, privilégiez le modèle le plus simple ou celui qui répond le mieux aux exigences de votre recherche. Les modèles plus simples sont souvent préférés, car ils sont plus faciles à interpréter et à généraliser.

07-Limitations du critère Cp

Bien que le critère Cp soit un outil puissant pour la sélection de modèles, il présente certaines limitations importantes à prendre en compte.

- **Validité pour de grands échantillons** : Le critère **Cp** est plus fiable avec de grands échantillons. Pour les petits échantillons, l'estimation de la variance des erreurs peut être biaisée, ce qui rend les résultats moins précis.

- **Problèmes complexes de sélection de variables : Cp** n'est pas adapté aux problèmes où plusieurs modèles complexes sont testés, comme la sélection de variables. Il peut ne pas gérer correctement les interactions ou corrélations entre les variables.

8-Conclusion

En conclusion, le critère **Cp de Mallows** constitue une méthode efficace pour la sélection de modèles, en équilibrant la complexité et la précision de la prédiction. Toutefois, comme nous l'avons vu, il présente des limitations, notamment sa validité restreinte aux grands échantillons et ses difficultés à traiter des modèles complexes. Il est essentiel de compléter cette approche avec d'autres méthodes de sélection de modèles, telles que l'**AIC** ou le **BIC**, pour obtenir une évaluation plus robuste et précise des modèles. Ainsi, un choix éclairé de la méthode de sélection permet d'optimiser les performances du modèle tout en évitant le surajustement.

Références

https ://online.stat.psu.edu/stat462/node/197/
https ://www.youtube.com/watch?v=Clql44fHLfM
https ://www.youtube.com/watch?v=rqFpWc-x0dM
https ://www.youtube.com/watchv=C3eMZYJXwk8t=229s
https://en.wikipedia.org/wiki/Mallows%27s_Cp#cite_note-Giraud-6

Merci

pour votre attention