MAT 3775
Analyse de la régression

Chapitre 5
Sélection de modèles

P. Boily (uOttawa)

Session d'hiver – 2023

Aperçu

5 – Sélection de modèles

Avec des ensembles de données et des situations réelles raisonnables, nous pouvons souvent construire des dizaines (voire des centaines) de modèles liés à un scénario spécifique.

Lorsque la plupart de ces modèles sont “alignés” les uns avec les autres (c’est-à-dire qu’ils donnent des résultats semblables), choisir le modèle le plus simple est généralement la meilleure approche.

En pratique, nous pouvons également atteindre un point de **rendements décroissants** – inclure plus de variables dans le modèle pourrait ne pas donner un meilleur pouvoir prédictif.

Comment choisir ”le” modèle avec lequel travailler ?

5.1 – Préliminaires

Un modèle linéaire $Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ est considéré comme une approximation de la fonction de régression (**non nécessairement linéaire**)

$$y = f(\mathbf{x}) = \mathrm{E}\left\{Y \mid (X_1, \ldots, X_p) = \mathbf{x}\right\}.$$

Dans le cadre des moindres carrés, nous supposons une **relation linéaire** entre la réponse Y et les prédicteurs $X_1, \ldots, X_p$:

$$\mathbf{b} = \arg\min_{\boldsymbol{\beta}} \{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\}.$$

Mais la précision de la prédiction souffre lorsque $p > n$; l'interprétabilité du modèle peut être améliorée en supprimant **les caractéristiques non pertinentes** (c'est-à-dire en **réduisant** p).

Il y a 3 classes de méthodes pour ce faire :

- les **méthodes de réduction/régularisation** (hors sujet pour le cours, cf. section 4.6) ;
- la **réduction de la dimension**, dans laquelle nous projetons les p prédicteurs sur une variété $\mathcal{H}$, avec $\dim(\mathcal{H}) = M \ll p$, et
- la **sélection des sous-ensembles**, où nous identifions un sous-ensemble de prédicteurs p pour lesquels il y a une (forte) association avec la réponse, et nous ajustons un modèle à cet ensemble réduit en utilisant le cadre des moindres carrés – étant donné p prédicteurs (dont certains peuvent être des termes d'interaction, des variables binaires, des puissances, etc.), il existe 2^p modèles qui peuvent être ajustés sur les données. Lequel de ces modèles est le **meilleur** ?

5.2 – Sélection du meilleur sous-ensemble

Dans l'approche **sélection du meilleur sous-ensemble** (BSS), la recherche du meilleur modèle est généralement composée de 3 étapes :

1. soit $\mathcal{M}_0$ le **modèle nul** (sans prédicteur) qui prédit simplement la moyenne de l'échantillon pour toutes les observations ;

2. pour $k = 1, \ldots, p$ (et tant que le modèle peut être ajusté) :

   (a) ajuster **chaque** modèle qui contient k prédicteurs (il y en a $\binom{p}{k}$) ;
   (b) choisir le modèle $\mathcal{M}_k$ ayant la **plus petite** SSE (**plus grand** R^2) ;

3. sélectionner un **unique** modèle à partir de $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ en utilisant C_p (AIC), BIC, R_a^2, ou toute autre métrique appropriée.

Nous ne pouvons pas utiliser SSE ou R^2 comme métriques à la dernière étape, car nous choisirions toujours $\mathcal{M}_p$ (puisque SSE **décroît de façon monotone** avec k et R^2 **augmente de façon monotone** avec k).

L'approche BSS est simple, mais avec 2^p modèles à essayer, elle devient **infaisable sur le plan des calculs** lorsque p est élevé ($p > 40$, disons).

De plus, lorsque p est élevé, les chances de trouver un modèle qui fonctionne **bien** selon l'étape 3 mais **mal** pour de nouvelles observations **augmentent**, ce qui conduit au **sur-ajustement** et à la **haute variance** des estimations.

Nous supposons que tous les modèles sont des modèles de moindres carrés, mais les algorithmes de sélection de sous-ensembles peuvent être utilisés pour d'autres familles de méthodes ; il suffit de disposer d'estimations d'erreurs de **formation** (2b) et de **validation** (3) appropriées.

5.3 – Sélection par étapes

La **sélection par étapes** (SS) tente de surmonter ce défi en ne considérant qu'un **ensemble restreint** de modèles. La **sélection par étapes par l'avant** (FSS) commence par le **modèle nul** $\mathcal{M}_0$ et ajoute des prédicteurs un par un jusqu'à ce qu'on atteigne le **modèle complet** $\mathcal{M}_p$:

1. soit $\mathcal{M}_0$ le **modèle nul** ;

2. pour $k = 0, \ldots, p-1$ (et tant que le modèle peut être ajusté) :
   (a) considérer les $p - k$ modèles qui ajoutent un **seul prédicteur** à $\mathcal{M}_k$;
   (b) choisir le modèle $\mathcal{M}_{k+1}$ ayant la **plus petite** SSE (**plus grand** R^2) ;

3. sélectionner un **unique** modèle à partir de $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ en utilisant C_p (AIC), BIC, R_a^2, ou toute autre métrique appropriée.

La **sélection par étapes rétrograde** (aussi BSS, malheureusement) fonctionne dans l'autre sens, en commençant par le **modèle complet** $\mathcal{M}_p$ et en supprimant les prédicteurs un par un jusqu'à ce qu'on atteigne le **modèle nul** $\mathcal{M}_0$:

1. soit $\mathcal{M}_p$ le **modéle complet** ;

2. pour $k = p, \ldots, 1$ (et tant que le modèle peut être ajusté) :

   (a) considérer les k modèles qui suppriment un **seul prédicteur** de $\mathcal{M}_k$;
   (b) choisir le modèle $\mathcal{M}_{k-1}$ ayant la **plus petite** SSE (**plus grand** R^2) ;

3. sélectionner un **unique** modèle à partir de $\{\mathcal{M}_0, \ldots, \mathcal{M}_p\}$ en utilisant C_p (AIC), BIC, R_a^2, ou toute autre métrique appropriée.

L'avantage computationnel de SS par rapport à 5.2 est évident : au lieu d'ajuster 2^p modèles, SS n'en requiert que

$$1 + p + (p-1) + \cdots + 2 + 1 = \frac{p^2 + p + 2}{2}.$$

Bien qu'il n'y ait aucune garantie que le "**meilleur**" modèle (parmi les 2^p) se retrouve dans les modèles SS, SS peut être utilisé dans des contextes où p est **trop grand** pour que l'autre approche soit réalisable en termes de calcul.

Pour les moindres carrés, **BSS** ne fonctionne que si $p \leq n$; si $p > n$, seule **FSS** est viable.

Les méthodes de **sélection hybride** (HS) tentent d'imiter la sélection du meilleur sous-ensemble tout en maintenant les calcul dans une plage gérable, un peu comme dans la SS.

5.4 – Mesures d'ajustement

En général, nous utilisons l'une des **statistiques d'ajustement** suivantes :

- le coefficient C_p **de Mallow**
- le **critère d'information d'Akaike** (AIC)
- le **critère d'information bayésien** (BIC), ou
- le **coefficient de détermination ajusté** R_a^2.

Les trois premiers doivent être **minimisés**, le dernier doit être **maximisé**.

Ces **statistiques d'ajustement** requierent les quantités suivantes :

- n, p, et $d = p + 2$
- $\hat{\sigma}^2$, l'estimation de $\sigma^2\{\varepsilon\}$;
- SSE et SST.

Le coefficient C_p **de Mallow** est

$$C_p = \frac{1}{n}(\mathrm{SSE} + 2d\hat{\sigma}^2) = \frac{1}{n}\mathrm{SSE} + \underbrace{\frac{2d\hat{\sigma}^2}{n}}_{\text{ajustment}}.$$

Plus d augmente, plus le terme d'ajustement augmente ; si $\hat{\sigma}^2$ est une estimation sans biais de $\sigma^2\{\varepsilon\}$, C_p est une estimation sans biais de MSE.

Le **critère d'information d'Akaike** (AIC) est

$$\text{AIC} = -2 \ln L + \underbrace{2d}_{\text{adjustment}},$$

où L est la valeur maximisée de la fonction de vraisemblance pour le modèle estimé. Si les erreurs suivent une **loi normale**, cela revient à maximiser

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2}{2\hat{\sigma}^2}\right) = \frac{1}{(2\pi)^{n/2}\hat{\sigma}^n} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2\right),$$

ou, en prenant le logarithme,

$$\ln L = \text{constante} - \frac{1}{2\hat{\sigma}^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

d'où

$$\arg\max_{\boldsymbol{\beta}}\{\ln L(\boldsymbol{\beta})\} = \arg\min_{\boldsymbol{\beta}}\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2\}.$$

Cependant,

$$\begin{aligned}\text{AIC} &= -2\ln L + 2d = \text{constante} + \frac{1}{\hat{\sigma}^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2d \\ &= \text{constante} + \frac{\text{SSE}}{\hat{\sigma}^2} + 2d \\ &= \text{constante} + \frac{n}{\hat{\sigma}^2}\cdot\frac{1}{n}\left(\text{SSE} + 2d\hat{\sigma}^2\right) = \text{constante} + \frac{n}{\hat{\sigma}^2}C_p.\end{aligned}$$

De toute évidence, lorsque la structure d'erreur est normale, **minimiser** AIC revient à **minimiser** C_p.

Le **critère d'information bayésien** utilise un terme d'ajustement différent :

$$\text{BIC} = \frac{1}{n}(\text{SSE} + d\hat{\sigma}^2 \ln n) = \frac{1}{n}\text{SSE} + \underbrace{d\hat{\sigma}^2 \frac{\ln n}{n}}_{\text{ajustment}}.$$

Cet ajustement pénalise les modèles ayant un nombre **élevé** de prédicteurs ; **minimiser** BIC aboutit à la sélection de modèles comportant moins de variables que ceux obtenus en **minimisant** C_p, en général.

Le **coefficient de détermination ajusté** R^2_a d'un modèle à $k-$paramètres est

$$R^2_{a,k} = 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k-1}.$$

Maximiser $R^2_{a,k}$ **minimise** $\frac{\text{SSE}}{n-k-1}$, et pénalise les **variables inutiles**.

TL;DR : si p est le $\#$ de paramètres dans le **modèle complet** (F), on cherche un **modèle réduit** (R) bien ajusté à k paramètres.

1. **Critère R^2 :** pour chaque sous-ensemble de prédicteurs, on calcule R^2; on trouve un sous-ensemble à k prédicteurs tel que R^2_k ne change pas de manière significative lorsque l'on augmente k.

2. **Critère R^2_a :** pour chaque sous-ensemble de prédicteurs, on calcule R^2_a; on trouve un sous-ensemble à k prédicteurs qui maximise les R^2_a.

3. **Critère C_p de Mallow :** pour chaque sous-ensemble de prédicteurs, on calcule $C_p = \frac{\text{SSE}_k}{\text{MSE}(F)} - (n - 2k)$; on cherche un sous-ensemble à k prédicteurs tel que C_p est petit et près de k (ce critère pourrait produire de nombreux modèles réduits appropriés).

Exemple : pour un certain ensemble de données avec trois prédicteurs, nous obtenons les C_p et R_p^2 de Mallow correspondants pour tous les sous-ensembles des prédicteurs.

p	C_p	R_p^2	Variables dans le modèle
4	4.0000	0.8548	X_1, X_2, X_3
3	22.4041	0.7527	X_1, X_2
3	29.1518	0.7189	X_1, X_3
2	42.3306	0.6429	X_1
3	52.8666	0.6002	X_2, X_3
2	81.6508	0.4461	X_2
2	146.8485	0.1197	X_3

En utilisant le critère C_p de Mallow ou le critère R^2, pouvons-nous trouver de “bons” modèles réduits ?

Solution : à part pour la première sélection (qui s'avère être le modèle complet, et non un modèle réduit), aucun des C_p n'est vraiment petit et près de p, donc le critère C_p de Mallow a peu de chances d'être utile.

Pour l'autre critère, nous avons

p	R^2 le plus élevé
2	0.6429
3	0.7527
4	0.8548

En passant de $p = 2$ à $p = 3$, la différence est $0.7527 - 0.6429 = 0.1098$;
En passant de $p = 3$ à $p = 4$, la différence est $0.8548 - 0.7527 = 0.1021$.
La seconde est légèrement plus petite que la première ; si nous devions absolument choisir un modèle réduit, nous devrions opter pour le modèle pour lequel $R^2 = 0.7527$ (celui avec les variables X_1 et X_2).