Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

OLS #14 (sélection de variables)

Sélection de variables

- ξ un sous-ensemble d'indices $\subseteq \{1, \ldots, p\}$ de cardinal $|\xi|$.
- $\mathbf{X}_\xi$ sous-matrice des covariables $\mathbf{x}_j, j \in \xi$.
- Dans le modèle ξ sélectionnant $|\xi|$ variables, les paramètres associés sont notés $\boldsymbol{\beta}_\xi$.
- $[\hat{\boldsymbol{\beta}}]_\xi$: coordonnées ξ du vecteur $\hat{\boldsymbol{\beta}}$; $[\hat{\boldsymbol{\beta}}]_\xi \neq \hat{\boldsymbol{\beta}}_\xi$ sauf si $\mathcal{V}(\mathbf{X}_\xi) \perp \mathcal{V}(\mathbf{X}_{\xi^c})$.
- soit une nouvelle observation $\mathbf{x}'^* = (\mathbf{x}'^*_\xi, \mathbf{x}'^*_{\xi^c})$, on note $\hat{Y}^* = \mathbf{x}'^* \hat{\boldsymbol{\beta}}$ et $\hat{Y}^*_\xi = \mathbf{x}'^*_\xi \hat{\boldsymbol{\beta}}_\xi$.
- si n^* nouvelles observations: $\hat{\mathbf{Y}}^* = \mathbf{X}^* \hat{\boldsymbol{\beta}}$ et $\hat{\mathbf{Y}}^*_\xi = \mathbf{X}^*_\xi \hat{\boldsymbol{\beta}}_\xi$.

Sélection de variables

- supposons disposer de $p = 3$ covariables et que le vrai modèle soit le modèle linéaire homoscédastique suivant:

$$\mathbf{Y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\varepsilon} = \mathbf{X}_{12}\boldsymbol{\beta}_{12} + \boldsymbol{\varepsilon}, \quad \xi = \{1, 2\}.$$

(pour alléger on supprime les "," et les {}, ainsi $\mathbf{X}_{12} = \mathbf{X}_{\{1,2\}}$).

- 7 modèles sont potentiellement envisageables: $\xi = \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}$.
- choix incorrect= trop peu ou trop de covariables!
- Examinons simplement ce qui se passe si $\xi = \{1\}$

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top \mathbf{Y}, \quad \hat{\mathbf{Y}}_1 = \mathcal{P}_{\mathbf{X}_1}\mathbf{Y}, \quad \hat{\sigma}_1^2 = \|\mathcal{P}_{\mathbf{X}_1^\perp}\mathbf{Y}\|^2/(n-1).$$

Sélection de variables

- $\mathbb{E}\mathbf{Y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 = \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\beta}_2\mathbf{X}_2$ donc

$$\mathbb{E}\hat{\boldsymbol{\beta}}_1 = \beta_1 + (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2\beta_2$$

$$\mathbb{E}\hat{Y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathcal{P}_{\mathbf{X}_1}\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbb{E}Y - \mathcal{P}_{\mathbf{X}_1^\perp}\mathbf{X}_2\beta_2.$$

- et pour l'estimateur de la variance

$$\begin{aligned}\mathbb{E}\hat{\sigma}_1^2 &= \frac{1}{n-1}\mathbb{E}\mathrm{tr}\left(\mathbf{Y}^\top\mathcal{P}_{\mathbf{X}_1^\perp}\mathbf{Y}\right) = \frac{1}{n-1}\mathrm{tr}\left(\mathcal{P}_{\mathbf{X}_1^\perp}\mathbb{E}(\mathbf{Y}\mathbf{Y}^\top)\right)\\ &= \frac{1}{n-1}\mathrm{tr}\left(\mathcal{P}_{\mathbf{X}_1^\perp}\left(\sigma^2\mathbb{I}_n + \mathbb{E}(\mathbf{Y})\mathbb{E}(\mathbf{Y})^\top\right)\right)\\ &= \sigma^2 + \frac{1}{n-1}\boldsymbol{\beta}_{12}^\top\mathbf{X}_{12}^\top\mathcal{P}_{\mathbf{X}_1^\perp}\mathbf{X}_{12}\boldsymbol{\beta}_{12}\\ &= \sigma^2 + \frac{1}{n-1}\beta_2^2\|\mathcal{P}_{\mathbf{X}_1^\perp}\mathbf{X}_2\|^2.\end{aligned}$$

- $\hat{\boldsymbol{\beta}}_\xi$ et $\hat{Y}_\xi$ sont en général biaisés.
- $\hat{\sigma}_\xi^2$ est en général positivement biaisé.

Sélection de variables

modèle	estimations	propriétés
$Y_1 = X_1\beta_1 + \varepsilon$	$\hat{Y}_1 = X_1\hat{\beta}_1$ $\hat{\sigma}_1^2 = \frac{\\|P_{X_1^\perp}Y\\|^2}{n-1}$	$B(\hat{Y}_1) = P_{X_1^\perp}X_2\beta_2$ $B(\hat{\sigma}_1^2) = \frac{1}{n-1}\beta_2^2\\|P_{X_1^\perp}X_2\\|^2$
$Y = X_2\beta_2 + \varepsilon$	$\hat{Y}_2 = X_2\hat{\beta}_2$ $\hat{\sigma}_2^2 = \frac{\\|P_{X_2^\perp}Y\\|^2}{n-1}$	$B(\hat{Y}_2) = P_{X_2^\perp}X_1\beta_1$ $B(\hat{\sigma}_2^2) = \frac{1}{n-1}\beta_1^2\\|P_{X_2^\perp}X_1\\|^2$
$Y = X_3\beta_3 + \varepsilon$	$\hat{Y}_3 = X_3\hat{\beta}_3$ $\hat{\sigma}_3^2 = \frac{\\|P_{X_3^\perp}Y\\|^2}{n-1}$	$B(\hat{Y}_3) = P_{X_3^\perp}X_{12}\beta_{12}$ $B(\hat{\sigma}_3^2) = \frac{1}{n-1}\beta_{12}'X_{12}'P_{X_{12}^\perp}X_{12}\beta_{12}$
$Y = X_{12}\beta_{12} + \varepsilon$	$\hat{Y}_{12} = X_{12}\beta_{12}$ $\hat{\sigma}_{12}^2 = \frac{\\|P_{X_{12}^\perp}Y\\|^2}{n-2}$	$B(\hat{Y}_{12}) = 0$ $B(\hat{\sigma}_{12}^2) = 0$
$Y = X_{13}\beta_{13} + \varepsilon$	$\hat{Y}_{13} = X_{13}\hat{\beta}_{13}$ $\hat{\sigma}_{13}^2 = \frac{\\|P_{X_{13}^\perp}Y\\|^2}{n-2}$	$B(\hat{Y}_{13}) = P_{X_{13}^\perp}X_{12}\beta_{12}$ $B(\hat{\sigma}_{13}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{13}^\perp}X_{12}\beta_{12}$
$Y = X_{23}\beta_{23} + \varepsilon$	$\hat{Y}_{23} = X_{23}\hat{\beta}_{23}$ $\hat{\sigma}_{23}^2 = \frac{\\|P_{X_{23}^\perp}Y\\|^2}{n-2}$	$B(\hat{Y}_{23}) = P_{X_{23}^\perp}X_{12}\beta_{12}$ $B(\sigma_{23}^2) = \frac{1}{n-2}\beta_{12}'X_{12}'P_{X_{23}^\perp}X_{12}\beta_{12}$
$Y = X_{123}\beta_{123} + \varepsilon$	$\hat{Y}_{123} = X_{123}\hat{\beta}_{123}$ $\hat{\sigma}_{123}^2 = \frac{\\|P_{X_{123}^\perp}Y\\|^2}{n-3}$	$B(\hat{Y}_{123}) = 0$ $B(\hat{\sigma}_{123}^2) = 0$

(via Cornillon & Matzner-Løober (2007)

Sélection de variables

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_1) = \sigma^2(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}, \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}_{12}) = \sigma^2\begin{pmatrix}\mathbf{X}_1^\top\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{X}_2\\ & \mathbf{X}_2^\top\mathbf{X}_2\end{pmatrix}^{-1}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{123}) = \sigma^2\begin{pmatrix}\mathbf{X}_1^\top\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{X}_2 & \mathbf{X}_1^\top\mathbf{X}_3\\ & \mathbf{X}_2^\top\mathbf{X}_2 & \mathbf{X}_2^\top\mathbf{X}_3\\ & & \mathbf{X}_3^\top\mathbf{X}_3\end{pmatrix}^{-1}.$$

De même

$$\mathrm{Var}(\hat{\mathbf{Y}}_1) = \sigma^2\mathcal{P}_{\mathbf{X}_1}, \quad \mathrm{Var}(\hat{\mathbf{Y}}_{12}) = \sigma^2\mathcal{P}_{\mathbf{X}_{12}} = \sigma^2\mathcal{P}_{\mathbf{X}_1} + \sigma^2\mathcal{P}_{\mathbf{X}_1^\perp\cap\mathbf{X}_2},$$

$$\mathrm{Var}(\hat{\mathbf{Y}}_{123}) = \sigma^2\mathcal{P}_{\mathbf{X}_1} + \sigma^2\mathcal{P}_{\mathbf{X}_1^\perp\cap\mathbf{X}_{23}}.$$

- $\mathrm{Var}([\hat{\boldsymbol{\beta}}]_\xi) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_\xi)$ est une matrice semi-définie positive.
- $\mathrm{Var}(\hat{\mathbf{Y}}) \geq \mathrm{Var}(\hat{\mathbf{Y}}_\xi)$.

Sélection de variables

$$\text{EQM}(\hat{\theta}) = \text{MSE}(\hat{\theta}) = \mathbb{E}((\hat{\theta} - \theta)^2) = (\mathbb{E}(\hat{\theta}) - \theta))^2 + \text{Var}(\hat{\theta})$$

(quantitité pertinente également si $\hat{\theta}$ est un vecteur, mais on regardera $\text{tr}(\text{EQM}(\hat{\theta}))$)

- $$\text{tr}(\text{EQM}(\hat{\mathbf{Y}}_\xi)) = |\xi|\sigma^2 + \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}\boldsymbol{\beta}\|^2.$$
- (si nouvelles données sont indépendantes des observations)

$$\text{tr}(\text{EQMP}(\hat{\mathbf{Y}}_\xi^*)) = n^*\sigma^2 + \text{tr}(\text{EQM}(\mathbf{X}_\xi^*\hat{\boldsymbol{\beta}}_\xi))$$

Sélection de variables

- Concernant $\hat{\mathbf{Y}}_\xi$

$$\mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_\xi)) = |\xi|\sigma^2 + \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}_{12}\boldsymbol{\beta}_{12}\|^2, \ \ \xi = 1, 2, 3, 23, 13,$$

$$\mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_{12})) = 2\sigma^2 \text{ et } \mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_{123})) = 3\sigma^2.$$

  (si quantités calculables, on pourrait sélectionner, selon σ^2 un modèle qui ne soit pas le bon modèle mais pour lequel l'EQM plus faible)

- Concernant $\hat{\mathbf{Y}}^*_\xi$

$$\mathrm{tr}(\mathrm{EQMP}(\hat{\mathbf{Y}}^*_\xi)) = (n^*+|\xi|)\sigma^2+\|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}^*_{12}\boldsymbol{\beta}_{12}\|^2, \ \ \xi = 1, 2, 3, 13, 23,$$

$$\mathrm{tr}(\mathrm{EQMP}(\hat{\mathbf{Y}}^*_{12})) = (n^*+2)\sigma^2 \text{ et } \mathrm{tr}(\mathrm{EQMP}(\hat{\mathbf{Y}}^*_{123})) = (n^*+3)\sigma^2.$$

- On peut montrer que $\mathrm{tr}(\mathrm{EQMP}(\hat{\mathbf{Y}}_\xi))$ est un très mauvais critère, qui sélectionne le modèle ayant le plus de covariables!

Sélection de variables

- si les modèles concurents sont emboîtés les uns dans les autres, il est possible d'utiliser une procédure de test.
- notation: modèle ξ à $|\xi|$ variables et ξ_{+1} modèle ξ auquel on a rajouté une variable.
- dire que le modèle ξ est le bon $\leftrightarrow \mathbb{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X}_\xi)$.
- $\text{SCR}(\xi)$ somme des carrés des résidus du modèle ξ, $= \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{Y}\|^2$.

On veut tester $H_0 : \mathbb{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X}_\xi)$ contre $H_1 :\in \mathcal{V}(\mathbf{X}_{\xi_{+1}})$ au seuil α. La statistique de test est

$$F = \frac{\text{SCR}(\xi) - \text{SCR}(\xi_{+1})}{\hat{\sigma}_\bullet^2} \text{ où } \hat{\sigma}_a^2 = \frac{\text{SCR}(\xi_{+1})}{n - |\xi| - 1} \text{ ou } \hat{\sigma}_b^2 = \frac{\text{SCR}(2:p)}{n-p}.$$

Le modèle ξ est rejeté au profit de ξ_{+1} si $f_{obs} > f_{1-\alpha,1,n-d_\bullet-1}$ avec $d_a = |\xi|$ et $d_b = p$

- $R^2(\xi) = 1 - \mathrm{SCR}(\xi)/\|\mathbf{Y} - \overline{Y}\|^2$.
- $R^2(\xi)$ croît avec $|\xi|$ (lorsque les modèles sont emboîtés).
- ξ_1, ξ_2, $|\xi_1| = |\xi_2|$, pertinent de comparer $R^2(\xi_1)$ à $R^2(\xi_2)$.
- Maximiser $R_a^2(\xi)$ revient à minimiser $\mathrm{SCR}(\xi)$ puisque

$$R_a^2(\xi) = 1 - \frac{n-1}{n-|\xi|}(1 - R^2(\xi)) = 1 - \frac{n-1}{\|\mathbf{Y} - \overline{Y}\|^2}\,\frac{\mathrm{SCR}(\xi)}{n-|\xi|}.$$

Le $C_p(\xi)$ de Mallows d'un modèle à $|\xi|$ variables explicatives est

$$C_p(\xi) = \frac{\mathrm{SCR}(\xi)}{\hat{\sigma}^2} - n + 2|\xi|, \quad \text{où } \hat{\sigma}^2 = \mathrm{SCR}(\{2:p\})/(n-p)$$

$$\begin{aligned}\mathbb{E}(\mathrm{SCR}(\xi)) &= \mathbb{E}\|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{Y}\|^2 = \mathbb{E}\|\mathcal{P}_{\mathbf{X}_\xi^\perp}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\|^2 \\ &= \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}\boldsymbol{\beta}\| + \sigma^2\mathrm{tr}(\mathcal{P}_{\mathbf{X}_\xi^\perp}) = \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}\boldsymbol{\beta}\| + (n-|\xi|)\sigma^2.\end{aligned}$$

Rappelons que $\mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_\xi)) = \|\mathcal{P}_{\mathbf{X}_\xi^\perp}\mathbf{X}\boldsymbol{\beta}\|^2 + |\xi|\sigma^2$ implique

$$\mathbb{E}\left(\hat{\sigma}^2 C_p(\xi)\right) = \mathrm{tr}(\mathrm{EQM}(\hat{\mathbf{Y}}_\xi)).$$

- Pour que $\hat{\sigma}^2 C_p$ soit un bon estimateur de l'EQM, il faut que l'estimation des paramètres et le choix des modèles ne dépendent pas des données, ce qui est rarement le cas.
- L'estimateur du C_p est biaisé: *biais de sélection*.

Rappel: sous l'hypothèse de normalité du bruit additif, la log-vraisemblance du modèle à ξ variables explicatives évaluée à l'EMV vaut

$$\log \mathcal{L}(\xi) = -\frac{n}{2} \log \frac{\mathrm{SCR}(\xi)}{n} - \frac{n}{2} (1 + \log 2\pi) \, .$$

- Maximiser la vraisemblance revient à minimiser $\mathrm{SCR}(\xi)$, un terme de pénalisation doit être introduit de la forme

$$-2 \log \mathcal{L}(\xi) + 2|\xi| f(n)$$

où $f(n)$ est une pénalisation dépendant de n.

- Critère d'information d'Akaike (1973),

$$\mathrm{AIC}(\xi) = -2\log\mathcal{L}(\xi) + 2|\xi| = n\log\frac{\mathrm{SCR}(\xi)}{n} + 2|\xi| + cte.$$

(autrement dit $f(n) = 1$).

- Bayesain Information Criterion (Schwarz, 1978)

$$\mathrm{BIC}(\xi) = -2\log\mathcal{L}(\xi) + |\xi|\log n = n\log\frac{\mathrm{SCR}(\xi)}{n} + |\xi|\log n + cte.$$

(autrement dit $f(n) = .5\log n$).

- dès que $n > 7$, $\log(n) > 2$ donc le critère BIC a tendance à sélectionner des modèles plus petits que l'AIC.

- Par les tests de Fisher, on conservera ξ si (en notant $f = f_{.95,1,n-|\xi|-1}$) si

$$\frac{\text{SCR}(\xi) - \text{SCR}(\xi_{+1})}{\text{SCR}(\xi_{+1})/(n - |\xi| - 1)} < f \simeq 4.$$

- Pour le R_a^2

$$R_a^2(\xi) > R_a^2(\xi_{+1}) \Longleftrightarrow \frac{\text{SCR}(\xi) - \text{SCR}(\xi_{+1})}{\text{SCR}(\xi_{+1})/(n - |\xi| - 1)} < 1$$

- Pour le C_p

$$C_p(\xi) < C_p(\xi_{+1}) \Longleftrightarrow \frac{\text{SCR}(\xi) - \text{SCR}(\xi_{+1})}{\text{SCR}(2:p)/(n - p)} < 2$$

- Pour les critères de vrais. pénalisés: $A\text{IC}$,

$$A\text{IC}(\xi) < A\text{IC}(\xi_{+1}) \Longleftrightarrow \frac{\text{SCR}(\xi) - \text{SCR}(\xi_{+1})}{\text{SCR}(\xi_{+1})/(n - |\xi| - 1)} \leq 2f(n)\left(1 - \frac{|\xi| + 1}{n}\right)$$