

Sélection de modèles en régression

David Causeur

Laboratoire de Mathématiques Appliquées

Agrocampus Rennes

IRMAR CNRS UMR 6625

<http://www.agrocampus-rennes.fr/math/causeur/>

Plan du cours

- 1 **Problématique**
Prédiction du taux de muscle de carcasses de porcs
Objectifs
- 2 **Choix de variables**
Critères de comparaison de modèles
Sélection de modèles
- 3 **Réduction de la dimension**
Régression biaisée
Méthodes à rang réduit
- 4 **Bilan et perspectives**

Classement de carcasses de porcs

TMP = Critère d'évaluation d'une carcasse de porc

- Mesure du TMP par dissection totale de la carcasse
→ mesuré indirectement par des épaisseurs tissulaires
- Différents types d'instruments de mesure

Sondes invasives



Scanners



Prédiction du TMP

Construire une équation de prédiction du **TMP** [ci-après Y]
par des mesures indirectes (**épaisseurs tissulaires**)
[ci-après x_1, x_2, \dots, x_p]

Modèle de régression linéaire

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon,$$

où

- p est le nombre d'épaisseurs tissulaires
- ε est l'erreur résiduelle d'écart-type σ

Estimation des paramètres du modèle

Échantillonnage

Individus	Y	x_1	x_2	\dots	x_p
1	Y_1	x_{11}	x_{12}	\dots	x_{1p}
2	Y_2	x_{21}	x_{22}	\dots	x_{2p}
\vdots	\vdots	\vdots	\vdots		\vdots
n	Y_n	x_{n1}	x_{n2}	\dots	x_{np}

Estimation des paramètres du modèle

Échantillonnage

Individus	Y	x_1	x_2	...	x_p	Données AutoFOM
1	58.02	0.55	0.57	...	0.41	$p = 137$
2	47.39	0.38	0.43	...	0.47	$n = 118$
⋮	⋮	⋮	⋮		⋮	
n	54.44	0.55	0.56	...	0.53	

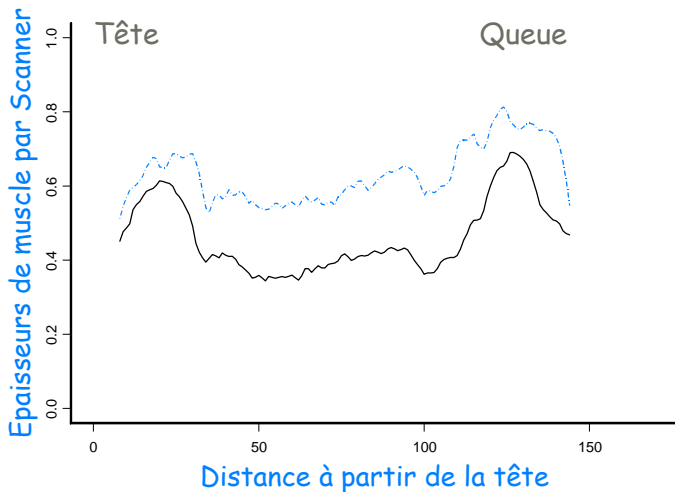
Deux situations

Scanner

- $p = 137, n = 118$
- Comment prédire au mieux ?

Estimation des paramètres du modèle

Données Scanner



Estimation des paramètres du modèle

Échantillonnage

Individus	Y	x_1	x_2	...	x_p	Données AutoFOM
1	58.02	0.55	0.57	...	0.41	$p = 137$
2	47.39	0.38	0.43	...	0.47	$n = 118$
⋮	⋮	⋮	⋮		⋮	
n	54.44	0.55	0.56	...	0.53	

Deux situations

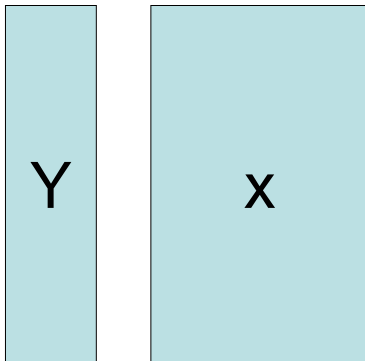
Scanner

- $p = 137, n = 118$
- Comment prédire au mieux ?

Sonde invasive

- $p = 11, n = 60$
- Quels sites d'application pour prédire ?

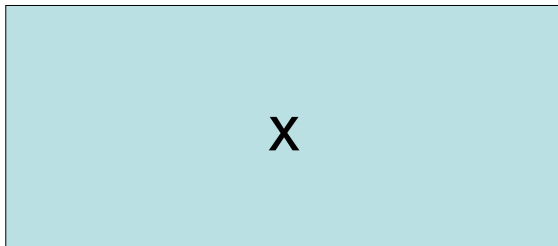
Exploration des données *Sondes invasives*



Exploration des données *Sondes invasives*

	TMP	CHMUSCLE	FRMUSCLE	MU23DCFR	MU34DCFR	G1CGM	GR34VLFR	CHGRAS	FRGRAS	G2CGM	GR34DCFR	GR23DCFR
TMP	1	0.45	0.43	0.34	0.28	-0.71	-0.81	-0.7	-0.72	-0.76	-0.78	-0.84
CHMUSCLE	0.45	1	0.97	0.68	0.62	0.01	-0.15	-0.21	-0.25	-0.11	-0.17	-0.26
FRMUSCLE	0.43	0.97	1	0.64	0.59	0.01	-0.14	-0.18	-0.21	-0.09	-0.14	-0.22
MU23DCFR	0.34	0.68	0.64	1	0.9	0.11	-0.06	-0.03	-0.08	-0.1	-0.12	-0.22
MU34DCFR	0.28	0.62	0.59	0.9	1	0.15	-0.07	-0.01	-0.03	-0.02	-0.09	-0.13
G1CGM	-0.71	0.01	0.01	0.11	0.15	1	0.9	0.72	0.72	0.86	0.81	0.81
CHGRAS	-0.7	-0.21	-0.18	-0.03	-0.01	0.72	0.74	1	0.98	0.76	0.76	0.77
FRGRAS	-0.72	-0.25	-0.21	-0.08	-0.03	0.72	0.73	0.98	1	0.75	0.75	0.77
GR34VLFR	-0.81	-0.15	-0.14	-0.06	-0.07	0.9	1	0.74	0.73	0.82	0.83	0.83
G2CGM	-0.76	-0.11	-0.09	-0.1	-0.02	0.86	0.82	0.76	0.75	1	0.93	0.93
GR34DCFR	-0.78	-0.17	-0.14	-0.12	-0.09	0.81	0.83	0.76	0.75	0.93	1	0.95
GR23DCFR	-0.84	-0.26	-0.22	-0.22	-0.13	0.81	0.83	0.77	0.77	0.93	0.95	1

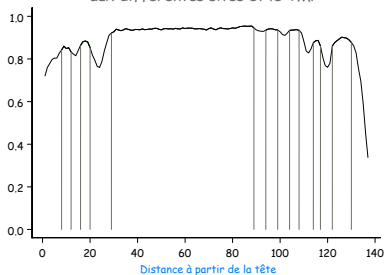
Exploration des données *Scanner*



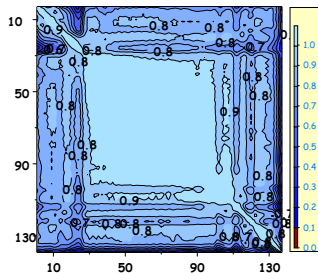


Exploration des données *Scanner*

Corrélations entre les mesures
aux différentes sites et le TMP



Corrélations inter-sites de mesures



Selection de modèles : objectifs

Réduction de l'information prédictrice

- **Objectif cognitif** : *En quels sites appliquer la sonde ?*
 - Choix de variables
 - Sélection d'un sous-modèle du modèle complet
- **Objectif de précision** : *Comment prédire au mieux ?*
 - Minimiser l'erreur de prédiction
 - Tenir compte de la redondance de l'information prédictrice

Plan du cours

- 1 Problématique
Prédiction du taux de muscle de carcasses de porcs
Objectifs
- 2 Choix de variables
Critères de comparaison de modèles
Sélection de modèles
- 3 Réduction de la dimension
Régression biaisée
Méthodes à rang réduit
- 4 Bilan et perspectives

Comparaison de modèles

Problème : comparer le modèle complet

$$\mathcal{M}_p : Y = \beta_0^{(p)} + \beta_1^{(p)} x_1 + \dots + \beta_p^{(p)} x_p + \varepsilon^{(p)}$$

et un sous-modèle construit sur q variables ($q < p$)

$$\mathcal{M}_q : Y = \beta_0^{(q)} + \beta_{i_1}^{(q)} x_{i_1} + \dots + \beta_{i_q}^{(q)} x_{i_q} + \varepsilon^{(q)}$$

Comparaison de modèles

Problème : comparer le modèle complet

$$\mathcal{M}_p : Y = \beta_0^{(p)} + \beta_1^{(p)} x_1 + \dots + \beta_p^{(p)} x_p + \varepsilon^{(p)}$$

et un sous-modèle construit sur q variables ($q < p$)

$$\mathcal{M}_q : Y = \beta_0^{(q)} + \beta_{i_1}^{(q)} x_{i_1} + \dots + \beta_{i_q}^{(q)} x_{i_q} + \varepsilon^{(q)}$$

Quel critère de comparaison ?

- Qualité d'ajustement (SCER, R^2)
- Test d'hypothèses

$$\begin{cases} H_0 : \mathcal{M}_p \text{ "équivalent à" } \mathcal{M}_q \\ H_1 : \mathcal{M}_p \text{ "meilleur que" } \mathcal{M}_q \end{cases}$$

- Précision de la prédiction

Qualité d'ajustement

R^2 : mesure de la qualité d'ajustement du modèle \mathcal{M}_q

$$\begin{aligned}
 R_q^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SCEM}_q}{\text{SCET}} \\
 &= \frac{\text{SCET} - \text{SCER}_q}{\text{SCET}} = 1 - \frac{\text{SCER}_q}{\text{SCET}}
 \end{aligned}$$

Qualité d'ajustement

R^2 : mesure de la qualité d'ajustement du modèle \mathcal{M}_q

$$\begin{aligned} R_q^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SCEM}_q}{\text{SCET}} \\ &= \frac{\text{SCET} - \text{SCER}_q}{\text{SCET}} = 1 - \frac{\text{SCER}_q}{\text{SCET}} \end{aligned}$$

Attention : $R_p^2 > R_q^2$

Qualité d'ajustement

R^2 : mesure de la qualité d'ajustement du modèle \mathcal{M}_q

$$\begin{aligned} R_q^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SCEM}_q}{\text{SCET}} \\ &= \frac{\text{SCET} - \text{SCER}_q}{\text{SCET}} = 1 - \frac{\text{SCER}_q}{\text{SCET}} \end{aligned}$$

Attention : $R_p^2 > R_q^2$

$R'^2 = R^2$ ajusté

$$R_q'^2 = 1 - \frac{\text{SCER}_q / (n - q - 1)}{\text{SCET} / (n - 1)}$$

Autres critères mesurant la qualité de l'ajustement

Pénalisation de SCER

- Critère d'Akaike (AIC)

$$AIC \propto n \log \left[\frac{SCER_q}{n} \right] + 2(q + 1)$$

- Critère d'information Bayésienne (BIC)

$$BIC \propto n \log \left[\frac{SCER_q}{n} \right] + (q + 1) \log(n)$$

Mesure de la redondance entre \mathcal{M}_p et \mathcal{M}_q

$$\begin{cases} H_0 & : \beta_j^p = 0 \text{ pour } j \notin \{i_1, i_2, \dots, i_q\} \\ H_1 & : \beta_j^p \neq 0 \text{ pour au moins un } j \text{ dans } \{i_1, i_2, \dots, i_q\} \end{cases}$$

Equation d'analyse de la variance de \mathcal{M}_q

$$SCET = SCEM_q + SCER_q$$

Mesure de la redondance entre \mathcal{M}_p et \mathcal{M}_q

$$\begin{cases} H_0 & : \beta_j^p = 0 \text{ pour } j \notin \{i_1, i_2, \dots, i_q\} \\ H_1 & : \beta_j^p \neq 0 \text{ pour au moins un } j \text{ dans } \{i_1, i_2, \dots, i_q\} \end{cases}$$

Equation d'analyse de la variance de \mathcal{M}_q

$$\begin{aligned} SCET &= SCEM_q + SCER_q \\ &= SCEM_q + (SCER_q - SCER_p) + SCER_p \end{aligned}$$

Mesure de la redondance entre \mathcal{M}_p et \mathcal{M}_q

$$\begin{cases} H_0 : \beta_j^p = 0 \text{ pour } j \notin \{i_1, i_2, \dots, i_q\} \\ H_1 : \beta_j^p \neq 0 \text{ pour au moins un } j \text{ dans } \{i_1, i_2, \dots, i_q\} \end{cases}$$

Equation d'analyse de la variance de \mathcal{M}_q

$$\begin{aligned} SCET &= SCEM_q + SCER_q \\ &= \underbrace{SCEM_q + (SCER_q - SCER_p)}_{SCEM_p} + SCER_p \end{aligned}$$

Mesure de la redondance entre \mathcal{M}_p et \mathcal{M}_q

$$\begin{cases} H_0 : \beta_j^p = 0 \text{ pour } j \notin \{i_1, i_2, \dots, i_q\} \\ H_1 : \beta_j^p \neq 0 \text{ pour au moins un } j \text{ dans } \{i_1, i_2, \dots, i_q\} \end{cases}$$

Equation d'analyse de la variance de \mathcal{M}_q

$$\begin{aligned} SCET &= SCEM_q + SCER_q \\ &= SCEM_q + \underbrace{(SCER_q - SCER_p)}_{SCEM_{p|q}} + SCER_p \end{aligned}$$

Mesure de la redondance entre \mathcal{M}_p et \mathcal{M}_q

$$\begin{cases} H_0 : \beta_j^p = 0 \text{ pour } j \notin \{i_1, i_2, \dots, i_q\} \\ H_1 : \beta_j^p \neq 0 \text{ pour au moins un } j \text{ dans } \{i_1, i_2, \dots, i_q\} \end{cases}$$

Equation d'analyse de la variance de \mathcal{M}_q

$$\begin{aligned} SCET &= SCEM_q + SCER_q \\ &= SCEM_q + \underbrace{(SCER_q - SCER_p)}_{SCEM_{p|q}} + SCER_p \end{aligned}$$

$$n - 1 = q + (p - q) + (n - p - 1)$$

Test de comparaison de modèles

Test de Fisher

$$F_{p|q} = \frac{\frac{SCER_q - SCER_p}{p-q}}{\frac{SCER_p}{n-p-1}} = \frac{\frac{R_p^2 - R_q^2}{p-q}}{\frac{1 - R_p^2}{n-p-1}} \underset{H_0}{\sim} \mathcal{F}_{p-q, n-p-1}$$

→ Probabilité critique : p_f

Test de comparaison de modèles

Test de Fisher

$$F_{p|q} = \frac{\frac{SCER_q - SCER_p}{p-q}}{\frac{SCER_p}{n-p-1}} = \frac{\frac{R_p^2 - R_q^2}{p-q}}{\frac{1-R_p^2}{n-p-1}} \underset{H_0}{\sim} \mathcal{F}_{p-q, n-p-1}$$

→ Probabilité critique : p_f

$\mathcal{M}_p > \mathcal{M}_q$ si

$$[F_{p|q} \geq f_\alpha] \Leftrightarrow \left[R_p^2 - R_q^2 \geq \frac{p-q}{n-p-1} f_\alpha (1 - R_p^2) \right]$$

Test de comparaison de modèles

Test de Fisher

$$F_{p|q} = \frac{\frac{SCER_q - SCER_p}{p-q}}{\frac{SCER_p}{n-p-1}} = \frac{\frac{R_p^2 - R_q^2}{p-q}}{\frac{1-R_p^2}{n-p-1}} \sim_{H_0} \mathcal{F}_{p-q, n-p-1}$$

→ Probabilité critique : p_f

$\mathcal{M}_p > \mathcal{M}_q$ si

$$[F_{p|q} \geq f_\alpha] \Leftrightarrow \left[R_p^2 - R_q^2 \geq \frac{p-q}{n-p-1} f_\alpha (1 - R_p^2) \right]$$

Données "Sonde invasive" : $R_p^2 = 0.83$, $\alpha = 0.05$

$$\mathcal{M}_p > \mathcal{M}_{q=4} \text{ si } R_p^2 - R_q^2 \geq 0.06$$

Test de comparaison de modèles

Test de Fisher

$$F_{p|q} = \frac{\frac{SCER_q - SCER_p}{p-q}}{\frac{SCER_p}{n-p-1}} = \frac{\frac{R_p^2 - R_q^2}{p-q}}{\frac{1 - R_p^2}{n-p-1}} \sim H_0 \mathcal{F}_{p-q, n-p-1}$$

→ Probabilité critique : p_f

$\mathcal{M}_p > \mathcal{M}_q$ si

$$[F_{p|q} \geq f_\alpha] \Leftrightarrow \left[R_p^2 - R_q^2 \geq \frac{p-q}{n-p-1} f_\alpha (1 - R_p^2) \right]$$

Données "Sonde invasive" : $R_p^2 = 0.83$, $\alpha = 0.05$

$$\begin{aligned} \mathcal{M}_p > \mathcal{M}_{q=4} \text{ si } R_p^2 - R_q^2 &\geq 0.06 \\ R_q^2 &\leq 0.77 \end{aligned}$$

Précision de la prédiction

Erreur de prédiction : $Y_0 - \hat{Y}_0$

- $\mathbb{E}(Y_0 - \hat{Y}_0) = 0$
- Mesure de la précision :

$$\sigma_0^2 = \text{Var}(Y_0 - \hat{Y}_0) = \mathbb{E} \left[(Y_0 - \hat{Y}_0)^2 \right]$$

Précision de la prédiction

Erreur de prédiction : $Y_0 - \hat{Y}_0$

- $\mathbb{E}(Y_0 - \hat{Y}_0) = 0$
- Mesure de la précision :

$$\sigma_0^2 = \text{Var}(Y_0 - \hat{Y}_0) = \mathbb{E} \left[(Y_0 - \hat{Y}_0)^2 \right]$$

Estimation de la variance de prédiction

- n erreurs de prédiction $Y_i - \hat{Y}_{-i}$
- Écart quadratique moyen de prédiction (EQMP) :

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^n \left[Y_i - \hat{Y}_{-i} \right]^2$$

Précision de la prédiction

Erreur de prédiction : $Y_0 - \hat{Y}_0$

- $\mathbb{E}(Y_0 - \hat{Y}_0) = 0$
- Mesure de la précision :

$$\sigma_0^2 = \text{Var}(Y_0 - \hat{Y}_0) = \mathbb{E}[(Y_0 - \hat{Y}_0)^2]$$

Estimation de la variance de prédiction

- n erreurs de prédiction $Y_i - \hat{Y}_{-i}$
- Écart quadratique moyen de prédiction (EQMP) :

$$\hat{\sigma}_0^2 = \frac{1}{n} \underbrace{\sum_{i=1}^n [Y_i - \hat{Y}_{-i}]^2}_{\text{PRESS}}$$

Recherche du meilleur sous-modèle

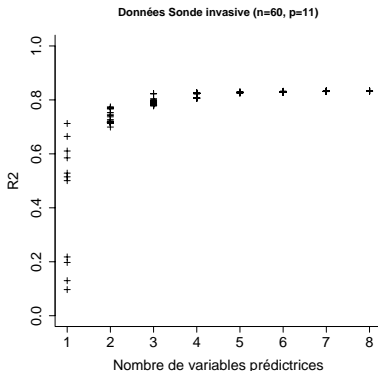
Nombre de sous-modèles du modèle complet : 2^p

- $p = 11$: 2048 sous-modèles

Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

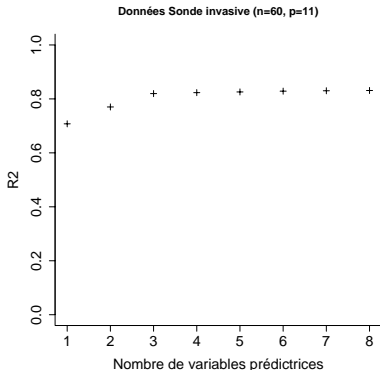
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

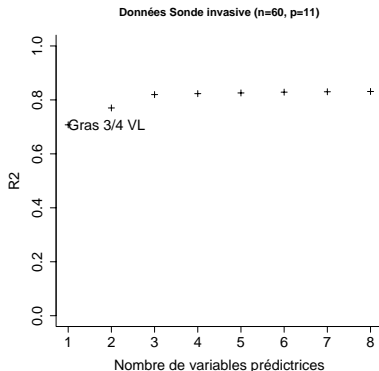
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

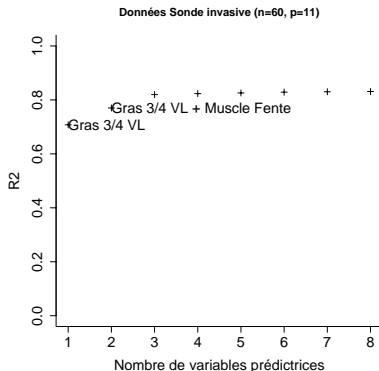
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

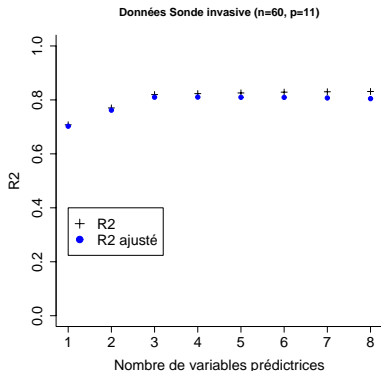
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

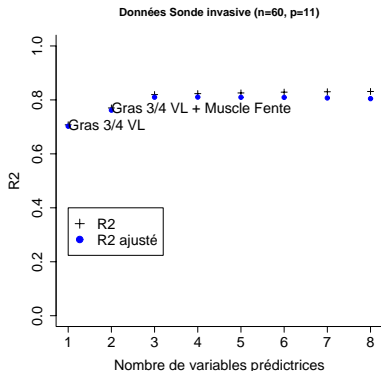
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

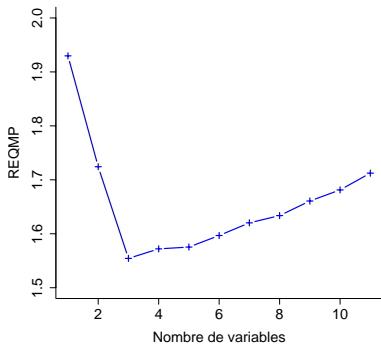
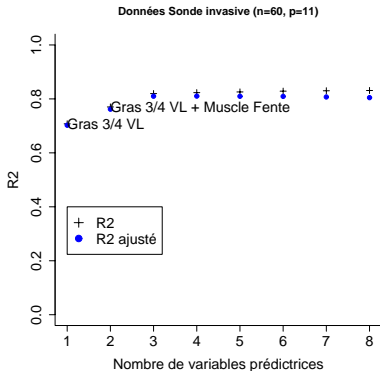
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

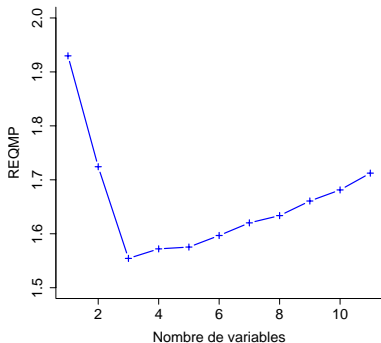
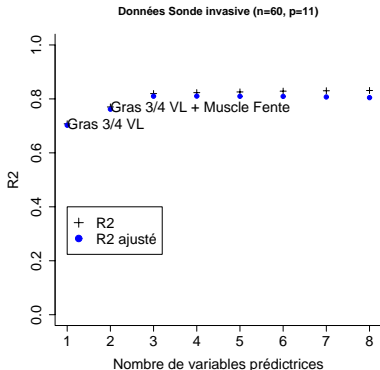
- $p = 11$: 2048 sous-modèles



Recherche du meilleur sous-modèle

Nombre de sous-modèles du modèle complet : 2^p

- $p = 11$: 2048 sous-modèles



- $p = 137$: $1.7e41$ sous-modèles

Sélection pas à pas

Sélection descendante

- **Modèle de départ : \mathcal{M}_p**
 - Comparaison de \mathcal{M}_p à tous les \mathcal{M}_{p-1} possibles
 - Sélection du modèle \mathcal{M}_{p-1}^* optimisant le critère
- **Modèle de départ : \mathcal{M}_{p-1}^***
- ...
- **Modèle de départ : \mathcal{M}_{k+1}^***
 - Comparaison de \mathcal{M}_{k+1}^* à tous les \mathcal{M}_k possibles
 - Arrêt si, pour tous les \mathcal{M}_k , le critère n'est pas amélioré

Sélection pas à pas

Sélection ascendante

- **Modèle de départ : \mathcal{M}_0**
 - Comparaison de \mathcal{M}_0 à tous les \mathcal{M}_1 possibles
 - Sélection du modèle \mathcal{M}_1^* optimisant le critère
- **Modèle de départ : \mathcal{M}_1^***
- ...
- **Modèle de départ : \mathcal{M}_{k-1}^***
 - Comparaison de \mathcal{M}_{k-1}^* à tous les \mathcal{M}_k possibles
 - Arrêt si, pour tous les \mathcal{M}_k , le critère n'est pas amélioré

Sélection ascendante sur la base de l'EQMP

Suivi de la sélection

- Modèle de départ : \mathcal{M}_0 : *REQMP = 5.39*
- Meilleur sous-modèle à une variable (X_{92}) : *REQMP = 1.61*
- Meilleur sous-modèle à 2 variables ($X_{92} + X_{134}$) : *REQMP = 1.35*
- ...
- Meilleur sous-modèle à 11 variables ($X_{92} + X_{134} + \dots + X_{85}$) : *REQMP = 1.02*

Sélection ascendante sur la base de l'EQMP

Suivi de la sélection

- Modèle de départ : \mathcal{M}_0 : *REQMP = 5.39*
- Meilleur sous-modèle à une variable (X_{92}) : *REQMP = 1.61*
- Meilleur sous-modèle à 2 variables ($X_{92} + X_{134}$) : *REQMP = 1.35*
- ...
- Meilleur sous-modèle à 11 variables ($X_{92} + X_{134} + \dots + X_{85}$) : *REQMP = 1.02*

Remarques

- Exploration d'une infime fraction des sous-modèles
au plus $p + (p - 1) + (p - 2) + \dots + 1 = \frac{p(p+1)}{2}$
- Intérêt cognitif de la sélection limité

Plan du cours

- 1 **Problématique**
Prédiction du taux de muscle de carcasses de porcs
Objectifs
- 2 **Choix de variables**
Critères de comparaison de modèles
Sélection de modèles
- 3 **Réduction de la dimension**
Régression biaisée
Méthodes à rang réduit
- 4 **Bilan et perspectives**

Méthode des moindres carrés lorsque $p > n$

Cas $p = 2$

$$\hat{\beta} = \mathbf{S}_{xx}^{-1} \mathbf{s}_{xy},$$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \sigma^2 \mathbf{S}_{xx}^{-1} = \sigma^2 \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{s_1^2 s_2^2 (1 - r_{12}^2)} \begin{bmatrix} s_2^2 & -s_{12} \\ -s_{12} & s_1^2 \end{bmatrix} \end{aligned}$$

Méthode des moindres carrés lorsque $p > n$

Cas $p = 2$

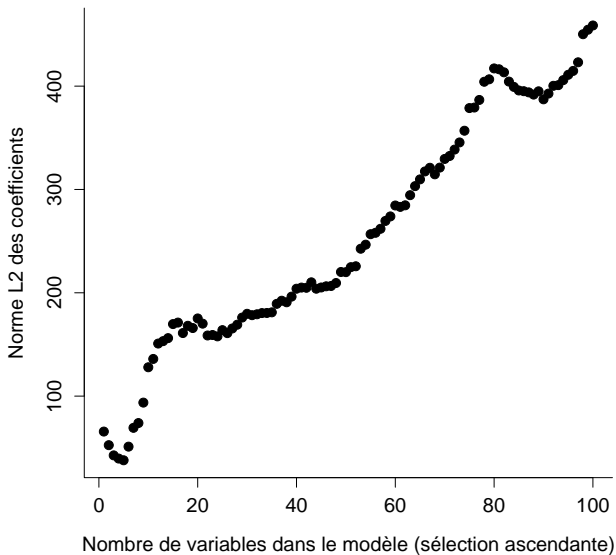
$$\hat{\beta} = \mathbf{S}_{xx}^{-1} \mathbf{s}_{xy},$$

$$\text{Var}(\hat{\beta}) = \sigma^2 \mathbf{S}_{xx}^{-1} = \sigma^2 \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}^{-1}$$

$$= \frac{\sigma^2}{s_1^2 s_2^2 (1 - r_{12}^2)} \begin{bmatrix} s_2^2 & -s_{12} \\ -s_{12} & s_1^2 \end{bmatrix}$$

Variance forte si $r_{12}^2 \approx 1$

Méthode des moindres carrés lorsque $p > n$



Méthode des moindres carrés lorsque $p > n$

Cas $p = 2$

$$\hat{\beta} = \mathbf{S}_{xx}^{-1} \mathbf{s}_{xy},$$

$$\text{Var}(\hat{\beta}) = \sigma^2 \mathbf{S}_{xx}^{-1} = \sigma^2 \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix}^{-1}$$

$$= \frac{\sigma^2}{s_1^2 s_2^2 (1 - r_{12}^2)} \begin{bmatrix} s_2^2 & -s_{12} \\ -s_{12} & s_1^2 \end{bmatrix}$$

Variance forte si $r_{12}^2 \approx 1$

Explosion des valeurs estimées des coefficients

Moindres carrés pénalisés : régression *ridge*

Critère à minimiser sous contrainte

$$SC(\beta) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2,$$

$$\text{avec } \sum_{i=1}^n \beta_i^2 = \|\beta\|_2^2 \leq \kappa$$

Moindres carrés pénalisés : régression *ridge*

Critère à minimiser sous contrainte (formulation équivalente)

$$SC(\beta) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2,$$

$$\text{avec } \sum_{i=1}^n \beta_i^2 = \|\beta\|_2^2 = \kappa$$

Moindres carrés pénalisés : régression *ridge*

Critère à minimiser sous contrainte (formulation équivalente)

$$SC(\beta) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2,$$

$$\text{avec } \sum_{i=1}^n \beta_i^2 = \|\beta\|_2^2 = \kappa$$

Critère à minimiser (multiplicateur de Lagrange)

$$SC(\beta; \lambda) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2$$

$$+ \lambda \|\beta\|_2^2$$

Troc "biais-variance"

Estimateur Ridge

$$\hat{\beta}_\lambda = [\mathbf{S}_{xx} + \lambda I_p]^{-1} \mathbf{s}_{xy} \quad \hat{\beta}_\lambda = \frac{\mathbf{s}_{xy}}{\mathbf{s}_x^2 + \lambda}$$

Troc "biais-variance"

Estimateur Ridge

$$\hat{\beta}_\lambda = [\mathbf{S}_{xx} + \lambda I_p]^{-1} \mathbf{s}_{xy} \quad \hat{\beta}_\lambda = \frac{\mathbf{s}_{xy}}{\mathbf{s}_x^2 + \lambda}$$

Propriétés

- Estimateur biaisé si $\lambda \neq 0$: $\mathbf{b}_\lambda = \mathbb{E} [\hat{\beta}_\lambda - \beta] \neq 0$
- Estimateur de variance plus faible que $\hat{\beta}$:

$$V_\lambda = \text{Var} [\hat{\beta}_\lambda] \quad \text{Var}(\hat{\beta}_\lambda) = \sigma^2 \frac{\mathbf{s}_x^2}{(\mathbf{s}_x^2 + \lambda)^2}$$

Troc "biais-variance"

Estimateur Ridge

$$\hat{\beta}_\lambda = [\mathbf{S}_{XX} + \lambda I_p]^{-1} \mathbf{s}_{XY} \quad \hat{\beta}_\lambda = \frac{\mathbf{s}_{XY}}{\mathbf{s}_X^2 + \lambda}$$

Propriétés

- Estimateur biaisé si $\lambda \neq 0$: $b_\lambda = \mathbb{E} [\hat{\beta}_\lambda - \beta] \neq 0$
- Estimateur de variance plus faible que $\hat{\beta}$:

$$V_\lambda = \text{Var} [\hat{\beta}_\lambda] \quad \text{Var}(\hat{\beta}_\lambda) = \sigma^2 \frac{\mathbf{s}_X^2}{(\mathbf{s}_X^2 + \lambda)^2}$$

Échange biais-variance

Augmentation du biais
 Réduction de la variance
 →

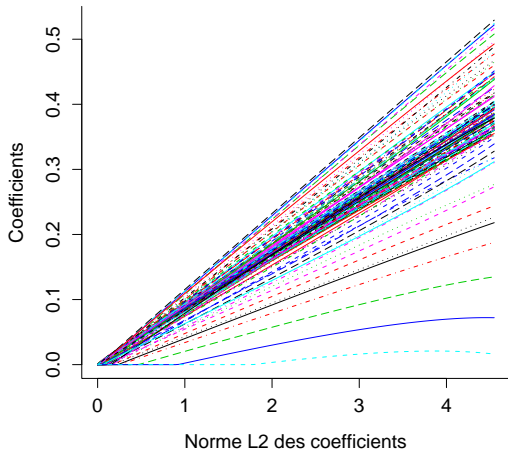
Biais nul, variance forte

0

λ

Troc "biais-variance"

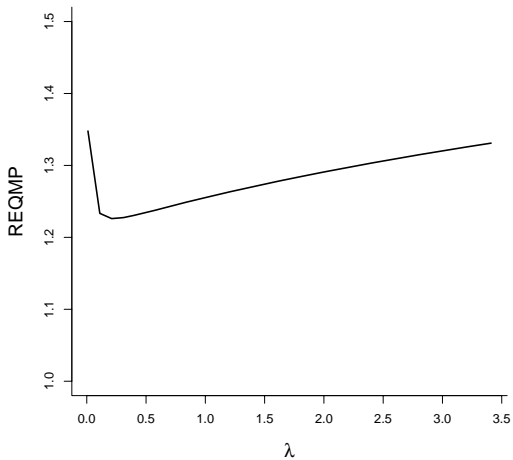
Rétrécissement des valeurs des coefficients



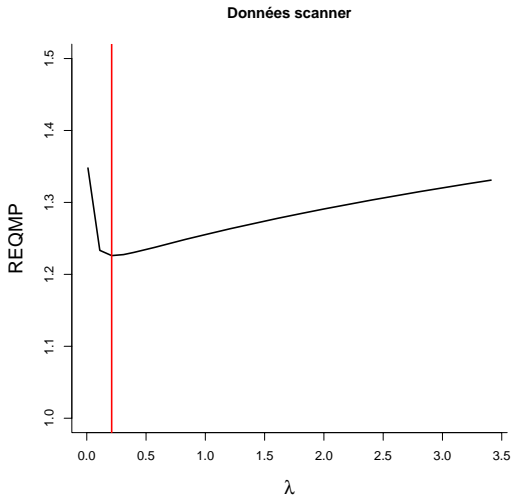


Choix de λ

Données scanner



Choix de λ



Rétrécissement + Sélection : régression *LASSO*

Critère à minimiser sous contrainte

$$SC(\beta) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2,$$

$$\text{avec } \sum_{i=1}^n |\beta_i| = \|\beta\|_1 \leq \kappa$$

Rétrécissement + Sélection : régression *LASSO*

Critère à minimiser sous contrainte

$$SC(\beta) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2,$$

$$\text{avec } \sum_{i=1}^n |\beta_i| = \|\beta\|_1 \leq \kappa$$

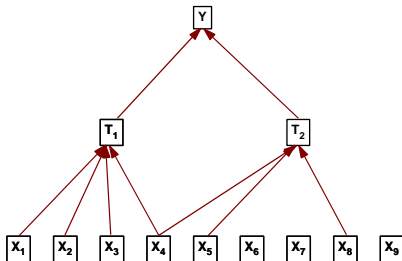
Critère à minimiser sous contrainte (formulation équivalente)

$$SC(\beta; \lambda) = \sum_{i=1}^n (Y_i - \bar{Y} - \beta_1[x_i^{(1)} - \bar{x}^{(1)}] - \dots - \beta_p[x_i^{(p)} - \bar{x}^{(p)}])^2$$

$$+ \lambda \|\beta\|_1$$

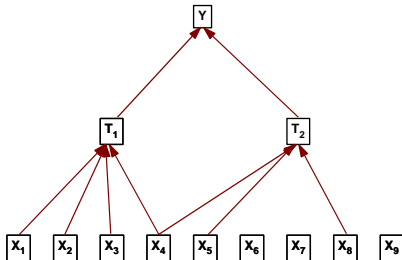
Dimensionnalité d'un modèle de régression ($p > n$)

Variables latentes



Dimensionnalité d'un modèle de régression ($p > n$)

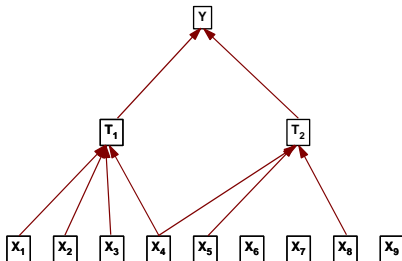
Variables latentes



Dimensionnalité : nombre de variables latentes

Dimensionnalité d'un modèle de régression ($p > n$)

Variables latentes



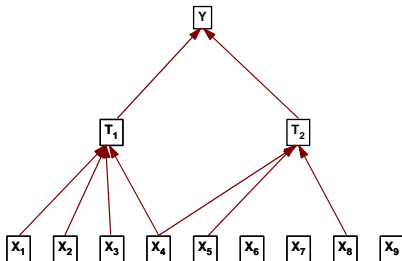
Dimensionnalité : nombre de variables latentes

Modèles de variables latentes

$$T_k = \alpha_{k1} \frac{x_1 - \bar{x}_1}{s_1} + \dots + \alpha_{kp} \frac{x_p - \bar{x}_p}{s_p}, \text{ avec } \alpha_{k1}^2 + \dots + \alpha_{kp}^2 = 1$$

Dimensionnalité d'un modèle de régression ($p > n$)

Variables latentes



Dimensionnalité : nombre de variables latentes

Modèles de variables latentes

$$T_k = \alpha_{k1} \frac{x_1 - \bar{x}_1}{s_1} + \dots + \alpha_{kp} \frac{x_p - \bar{x}_p}{s_p}, \text{ avec } \alpha_{k1}^2 + \dots + \alpha_{kp}^2 = 1$$

Extraction des variables latentes

2 méthodes d'extraction des variables latentes

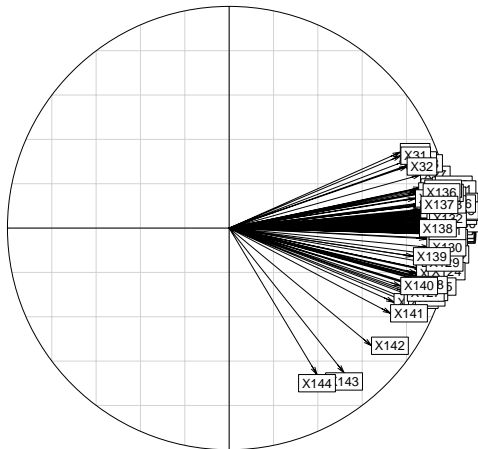
- **Régression sur composantes principales (PCR)**
 - $\alpha^{(1)}$ choisi tel que $\text{Var}(T_1)$ maximale
 - $\alpha^{(2)}$ choisi tel que $\text{Cov}(T_1, T_2) = 0$ et $\text{Var}(T_2)$ maximale
 - ...
 - $\alpha^{(k)}$ choisi tel que $\text{Cov}(T_k, T_j) = 0, j < k$ et $\text{Var}(T_k)$ maximale

Extraction des variables latentes

2 méthodes d'extraction des variables latentes

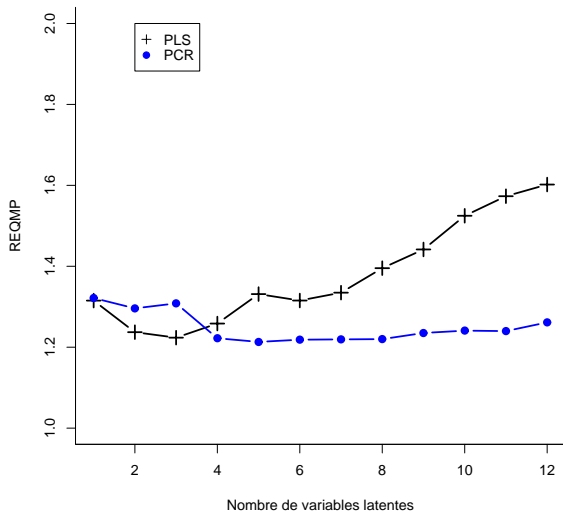
- **Régression sur composantes principales (PCR)**
 - $\alpha^{(1)}$ choisi tel que $\text{Var}(T_1)$ maximale
 - $\alpha^{(2)}$ choisi tel que $\text{Cov}(T_1, T_2) = 0$ et $\text{Var}(T_2)$ maximale
 - ...
 - $\alpha^{(k)}$ choisi tel que $\text{Cov}(T_k, T_j) = 0, j < k$ et $\text{Var}(T_k)$ maximale
- **Régression par moindres carrés partiels (PLS)**
 - $\alpha^{(1)}$ choisi tel que $\text{Cov}^2(T_1, Y)$ maximale
 - $\alpha^{(2)}$ choisi tel que $\text{Cov}(T_1, T_2) = 0$ et $\text{Cov}^2(T_2, Y)$ maximale
 - ...
 - $\alpha^{(k)}$ choisi tel que $\text{Cov}(T_k, T_j) = 0, j < k$ et $\text{Cov}^2(T_k, Y)$ maximale

Interprétation des variables latentes



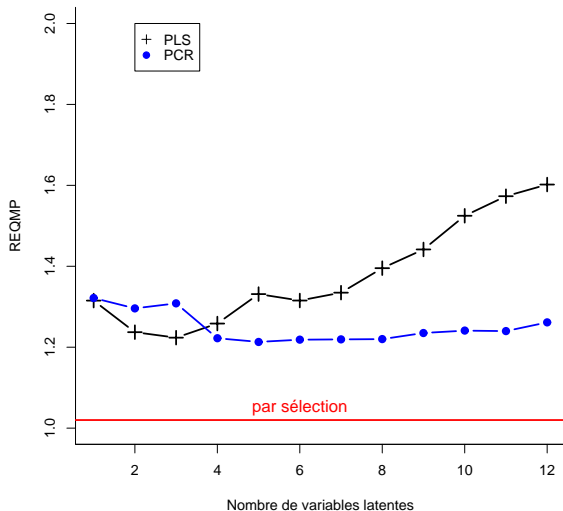
Comparaison des performances de prédiction

Données scanner



Comparaison des performances de prédiction

Données scanner



Plan du cours

- 1 Problématique
Prédiction du taux de muscle de carcasses de porcs
Objectifs
- 2 Choix de variables
Critères de comparaison de modèles
Sélection de modèles
- 3 Réduction de la dimension
Régression biaisée
Méthodes à rang réduit
- 4 Bilan et perspectives

Bilan et perspectives

Sélection de modèles

- Choix de variables
- Réduction de la dimension

... dans les prochaines séances de TD

- Modélisation de la rentabilité d'élevages par les caractéristiques de l'exploitation (`fonction step`)
- Modélisation de la teneur en matière sèche par des SPIR (`packages PLS + glmnet`)

... dans les prochains cours

- Lorsque $x_i = x(s_i)$... autres méthodes de réduction de la dimension
- Extension dans le cas où Y est qualitative