Colinéarité et Sélection de Variables

Régression Linéaire Multiple

Ricco RAKOTOMALALA

PLAN

1. Problème de la sélection de variables
2. Colinéarité et ses conséquences
3. Détection de la colinéarité
4. Répondre à la colinéarité : la sélection de variables

Le problème de la sélection de variables

Quelles variables conserver dans la régression ?

Global results

Endogenous attribute	Consommation
Examples	27
R^2	0.929520
Adjusted-R^2	0.916706
Sigma error	0.651169
F-Test (4,22)	72.5365 (0.000000)

La régression est d'excellente qualité

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	123.0278	4	30.7570	72.5365	0.0000
Residual	9.3285	22	0.4240		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(22)	p-value
Intercept	1.838006	0.793367	2.316716	0.030220
Prix	0.000034	0.000045	0.752738	0.459587
Cylindrée	0.001208	0.000722	1.672661	0.108557
Puissance	-0.003742	0.015030	-0.248956	0.805704
Poids	0.003728	0.001300	2.868568	0.008926

Régression « CONSO » sans les données atypiques

A 5%, seul « poids » semble significatif, les autres ne semblent pas pertinentes. Deux raisons possibles à cela :
(1) La variable n'a aucun lien avec la variable à prédire
(2) La variable est liée avec Y, mais elle est gênée (redondante) avec une (ou plusieurs) des autres variables exogènes, qui elle même peut ne pas paraître significative → **colinéarité**

Conséquences de la colinéarité

Pourquoi la colinéarité (la corrélation entre exogènes) est un problème ?

Colinéarité parfaite $rang(X'X) < (p+1) \Rightarrow (X'X)^{-1}$ n'existe pas. Calcul des coefficients de la régression impossible.

Colinéarité forte $\det(X'X) \approx 0 \Rightarrow (X'X)^{-1} = \frac{1}{\det(X'X)}[com(A)]'$ contient de très grandes valeurs

Les valeurs de la matrice de variance covariance des coefficients $\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\varepsilon}(X'X)^{-1}$ sont exagérées, notamment l'estimation de la variance des paramètres

Quelles conséquences ?

- Variances des estimateurs exagérées
- Les t de Student sont sous-estimés, les variables ne paraissent pas significatives (cf. cylindrée, puissance)
- Les valeurs/signes des coefficients sont contradictoires, ne concordent pas avec les connaissances du domaine (puissance est reliée négativement avec la consommation ????)
- Les résultats sont très instables, l'adjonction ou la suppression de quelques observations modifient fortement les valeurs et les signes des coefficients

→ Lecture des résultats très périlleuse

Détection de la colinéarité

Règle de Klein et autres règles très simples

1ère règle très simple : Il existe au moins 2 variables X_{j1} et X_{j2} telles que

$$\left| r_{j_1, j_2} \right| > 0.8$$

Cette règle ne tient pas compte des résultats (caractéristiques) de la régression.

2ème règle (dite de Klein) : Il existe au moins 2 variables X_{j1} et X_{j2} telles que

$$r^2_{j_1, j_2} \geq R^2$$

- R^2 est le coefficient de détermination de la régression (= 0.9295 dans notre exemple)
- En réalité, la situation est problématique dès que les valeurs sont comparables

Régression « CONSO »

Il n'y a que des problèmes dans ce fichier, en particulier (poids, prix) et (puissance, cylindrée)

Matrice des corrélations croisées

	prix	cylindree	puissance	poids
prix	1	0.92	0.93	0.95
cylindree	0.92	1	0.96	0.86
puissance	0.93	0.96	1	0.85
poids	0.95	0.86	0.85	1

Matrice des corrélations croisées au carré

	prix	cylindree	puissance	poids
prix	1	0.84	0.86	0.90
cylindree	0.84	1	0.91	0.74
puissance	0.86	0.91	1	0.73
poids	0.90	0.74	0.73	1

3ème règle (Cohérence des signes) : Le signe de la corrélation brute endogène /exogène devrait être le même que le signe du coefficient de la régression.

$$\mathrm{sgn}(r_{Y,X_j}) = \mathrm{sgn}(\hat{a}_j)$$

Régression « CONSO »

(1) Toutes les variables sont fortement liées avec l'endogène (les tests de la significativité de la régression ne disaient pas ça du tout)
(2) Manifestement, il y a un contre-sens sur « puissance »

	a_j	r_y,x
prix	0.000034	0.942597
cylindree	0.001208	0.908790
puissance	-0.003742	0.888304
poids	0.003728	0.944740

Détection de la colinéarité

Facteur d'inflation de la variance et tolérance

Plus loin que l'étude des corrélations des exogènes deux à deux → analyser la multicolinéarité

→ Coefficient de détermination de la régression de la variable X_j avec les (p-1) autres variables R_j^2

VIF (Facteur d'inflation de la variance)

$$v_j = \frac{1}{1 - R_j^2}$$

Problème dès que $R^2 \approx 1$ → $v_j \approx +oo$

Pourquoi VIF ?

$$V(\hat{a}_j) = \frac{\hat{\sigma}_\varepsilon}{n} \times v_j$$

$v_j \approx +oo$ → $V(a_j) \approx +oo$ (et t de Student ≈ 0)

Valeurs critiques du VIF $v_j \geq 4$, d'autres sources proposent 5 ou même 10; le plus important est d'identifier des groupes de variables à VIF élevés.

Calcul pratique : Effectuer p régression peut être lourd (p élevé et beaucoup d'observations), on peut lire le VIF sur la diagonale principale de la matrice inverse de la matrice des corrélations.

Matrice des corrélations croisées

	prix	cylindree	puissance	poids
prix	1	0.92	0.93	0.95
cylindree	0.92	1	0.96	0.86
puissance	0.93	0.96	1	0.85
poids	0.95	0.86	0.85	1

$^{-1}$ =

inverse matrice de corrélation

	PRIX	CYLINDREE	PUISSANC	POIDS
PRIX	19.79	-1.45	-7.51	-11.09
CYLINDREE	-1.45	12.87	-9.80	-1.36
PUISSANC	-7.51	-9.80	14.89	2.86
POIDS	-11.09	-1.36	2.86	10.23

Que des problèmes !!!

Résoudre les problèmes de colinéarité

Régression RIDGE
Régression sur facteurs de l'ACP
Régression PLS
Etc.

Sélection de variables → Éjecter les variables non-pertinentes et comprendre/détecter les variables redondantes : l'interprétation des résultats ne peut qu'en bénéficier

Sélection de variables

Sélection par optimisation

Principe : Trouver la combinaison de q ($q \leq p$) variables qui maximise un critère de qualité de la régression

Pourquoi pas le R^2

- Le R^2 indique la variance expliquée par le modèle, il semble tout indiqué.
- Mais le R^2 augmente mécaniquement avec le nombre de variables → phénomène de sur-apprentissage [(p+1) = n → R^2 = 1, même si variables totalement farfelues]
- Pénaliser l'aptitude à coller aux données (SCR faible) par la complexité (q élevé)
- Le R^2 n'est valable que si l'on compare des modèles de même complexité

Le R^2 corrigé

$$\bar{R}^2 = 1 - \frac{CMR}{CMT} = 1 - \frac{SCR/(n-q-1)}{SCT/(n-1)} = 1 - \frac{n-1}{n-q-1}(1-R^2)$$

Trop permissif c.-à-d. Favorise les modèles avec beaucoup de variables

Critères AIC et BIC (Akaike et Schwartz)

$$AIC = n \times \ln\frac{SCR}{n} + 2 \times (q+1)$$

$$BIC = n \times \ln\frac{SCR}{n} + \ln(n) \times (q+1)$$

« Vraie » formule de AIC

$$AIC = n \times \left[\ln(2\pi e) + \ln\left(\frac{SCR}{n}\right)\right] + 2 \times (q+1)$$

Objectif : Minimiser AIC ou BIC

Le critère BIC est le plus restrictif (favorise les solutions avec peu de variables)

Sélection de variables

Optimiser l'AIC

Recherche exhaustive : Tester toutes les combinaisons de q ($q \leq p$) variables qui minimise AIC

Problème : Il y a (2^p-1) régressions à évaluer → c'est prohibitif !!!

Recherche pas-à-pas : Forward (Ø puis adjonctions successives) ou Backward (Toutes puis éliminations successives) → jusqu'à la solution optimale

Backward – Régression « CONSO »

Étape	Modèle courant (avec constante)	AIC	AIC si suppression d'une variable
1	Conso = f(prix+cylindrée+puissance+poids+cte)	-18.69	Puissance : -20.62 Prix : -20.01 Cylindrée : -17.46 Poids : -12.16
2	Conso = f(prix+cylindrée+poids+cte)	-20.62	Prix : -21.99 Cylindrée : -17.6 Poids : -13.34
3	Conso = f(cylindrée+poids) (FORWARD fournit le même résultat)	-21.99	Cylindrée : -13.30 Poids : -0.28

Le critère AIC ne tient pas compte explicitement de la redondance (colinéarité) entre les variables. Il le fait implicitement en mettant en balance l'amélioration de l'ajustement (SCR) avec l'augmentation de la complexité (q) : est-ce qu'une variable supplémentaire dans le modèle emmène de l'information complémentaire pertinente

Sélection de variables

Utiliser le F partiel de Fisher

Principe : S'appuyer sur les propriétés inférentielles de la régression

- Ajouter une variable si le t de Student (ou t^2 = F-partiel) dans la régression passe le seuil critique à 5% (1%, etc.)
- Supprimer une variable si le t de Student est en deçà du seuil critique

Forward (à 5%)

Étape	Modèle courant (avec constante)	R^2	F-partiel = t^2 (p-value) si ajout de…
1	Conso = f(cte)	-	Poids : 207.63 (0.0000) Prix : 199.19 (0.0000) Cylindrée : 118.60 (0.0000) Puissance : 93.53 (0.0000)
2	Conso = f(cte + poids)	0.8925	Cylindrée : 11.6 (0.0023) Puissance : 7.42 (0.0118) Prix : 6.32 (0.0190)
3	Conso = f(cte + poids + cylindrée)	0.9277	Prix : 0.53 (0.4721) Puissance : 0.01 (0.9288)

Backward (à 5%)

Étape	Modèle courant (avec constante)	R^2	F-partiel = t^2 (p-value) dans la rég.
1	Conso = f(prix+cylindrée+puissance+poids+cte)	0.9295	Puissance : 0.0620 (0.8057) Prix : 0.5666 (0.4596) Cylindrée : 2.7978 (0.1086) Poids : 8.2287 (0.0089)
2	Conso = f(prix+cylindrée+poids+cte)	0.9293	Prix : 0.5344 (0.4721) Cylindrée : 4.6779 (0.0412) Poids :9.4345 (0.0054)
3	Conso = f(cylindrée+poids+cte)	0.9277	Cylindrée : 11.6631 (0.0023) Poids : 33.7761 (0.0000)

STEPWISE

Mixer Forward et Backward. Basé toujours sur le F-Partiel.

→ Vérifier que l'adjonction d'une variable ne provoque pas la suppression d'une variable déjà introduite

→ Ainsi de suite jusqu'à convergence (plus d'ajout ni de retrait possible)

Sélection de variables

Sélection STAGEWISE

Principe : Méthode FORWARD. Choisir la variable qui explique le mieux la fraction de Y non-expliquée par les variables déjà introduites. On parle de corrélation semi-partielle.

Algorithme :

(0) Commencer par une sélection vide
(1) e = Y
(2) Choisir la variable $\mathbf{X_a}$ la plus corrélée avec **e**. Si significative au sens du t^2 (à 5%, ou 1%…), introduire; sinon STOP.
(3) Calculer la part de Y non expliquée par les variables déjà sélectionnées → le résidu : e = Y – (a0+a1.X_a+…)
(4) Retour en (2)

Test t :

$$t_q = \frac{r}{\sqrt{\frac{1-r^2}{n-(q+1)}}} \equiv \Im[n-(q+1)]$$

Attention aux degrés de liberté pour le calcul du t^2 lors du test de significativité de la corrélation à l'étape q (q-1 variables déjà choisies)

Données « CONSO »

Étape 1

X	r
Poids	**0.9447**
Prix	0.9426
Cylindrée	0.9088
Puissance	0.8883

e = conso – (1.0353 + 0.0068 x poids)

$$t = \frac{0.9447}{\sqrt{(1-0.9447^2)/(27-2)}} = 14.41$$

Étape 2

X	r
Cylindrée	*0.2908*
Puissance	0.2544
Prix	0.1471
Poids	0.0000

$$t = \frac{0.2908}{\sqrt{(1-0.2908^2)/(27-3)}} = 1.4891 < 2.06 = t_{0.975}(24)$$

Seule la variable « Poids » est sélectionnée.

Sélection de variables
Corrélation partielle

Principe : Mesure le lien entre 2 variables (Y,X), après avoir retranché l'effet d'une tierce variable Z (sur Y **et** X). On parle de corrélation partielle.

Définition de la corrélation partielle YX.Z

$$r_{YX.Z} = \frac{r_{YX} - r_{YZ}.r_{XZ}}{\sqrt{(1-r_{YZ}^2)(1-r_{XZ}^2)}}$$

Test de significativité

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-3}}} \equiv \Im(n-3)$$

Données « CONSO »

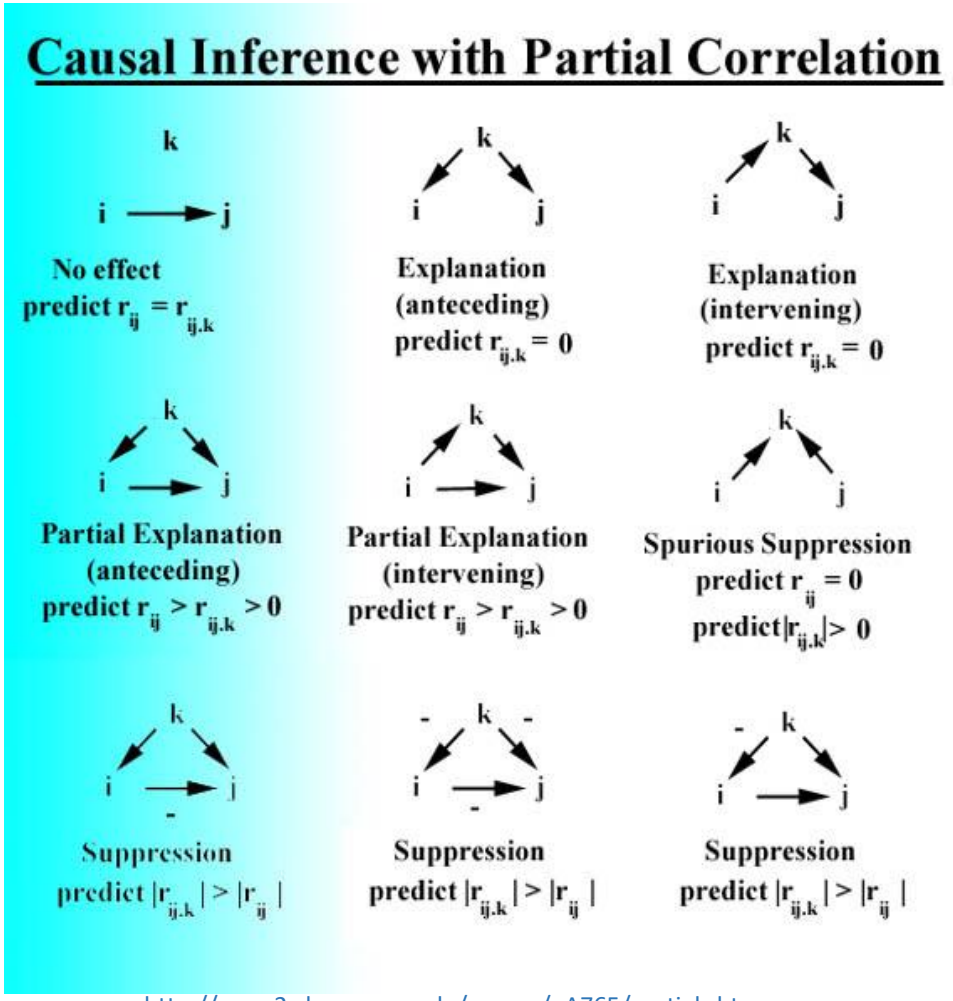
$$r_{conso-puissance.cylindree} = \frac{0.8883 - 0.9088 \times 0.9559}{\sqrt{(1-0.9088^2)(1-0.9559^2)}} = 0.16$$

avec

$$t = \frac{0.16}{\sqrt{\frac{1-0.16^2}{27-3}}} = 0.7940 < 2.06 = t_{0.975}(27-3)$$

Si on contrôle l'effet de la cylindrée (à cylindrée égale), la puissance ne pèse pas sur la consommation.

Typologie des corrélations partielles



http://www2.chass.ncsu.edu/garson/pA765/partialr.htm

Sélection de variables

Corrélation partielle d'ordre > 1

Principe : Mesure le lien entre 2 variables (Y,X), après avoir retranché l'effet de plusieurs variables Z1, Z2, …, Zq.

Approche n° 1 : Développer une définition récursive de la corrélation partielle

$$r_{YX.Z_1Z_2} = \frac{r_{YX.Z_1} - r_{YZ_2.Z_1}.r_{XZ_2.Z_1}}{\sqrt{(1-r^2_{YZ_2.Z_1})(1-r^2_{XZ_2.Z_1})}}$$

Compliquée à manipuler dès que « q » augmente

Approche n° 2 : Exploiter les résidus de la régression

$$\left.\begin{aligned} e_1 &= y - (\hat{a}_0 + \hat{a}_1 z_1 + \cdots + \hat{a}_q z_q) \\ e_2 &= x - (\hat{b}_0 + \hat{b}_1 z_1 + \cdots + \hat{b}_q z_q) \end{aligned}\right\} \Rightarrow$$

Corrélation partielle = corrélation brute entre les résidus

$$r_{YX.Z_1\cdots Z_q} = r_{e_1e_2} \quad \text{avec}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-(q+2)}}} \equiv \Im[n-(q+2)]$$

Données « CONSO »

$$r_{conso,\,puissance.cylindree,\,poids} = ?$$

Calculer les 2 résidus des régressions

$$\left.\begin{aligned} e_1 &= conso - (+1.3923 + 0.0045.poids + 0.0013.cylindree) \\ e_2 &= puissance - (-15.834 + 0.0117.poids + 0.0444.cylindree) \end{aligned}\right\}$$

Former : $r_{e_1e_2} = 0.0188$ et $t = \frac{0.0188}{\sqrt{\frac{1-0.0188^2}{27-(2+2)}}} = 0.0903 < 2.07 = t_{0.975}(27-4)$

Récapitulatif des corrélations brutes et partielles

Variables	r
Conso,Puis	0.89
Conso,Puis/Cylind	0.16
Conso,Puis/Cylind,Poids	0.02

A cylindrée et poids égal, la consommation n'est absolument pas liée à la puissance.

Sélection de variables

Basée sur la corrélation partielle

Principe : Méthode FORWARD. Détecter la variable maximisant la corrélation partielle avec Y (au regard des variables déjà sélectionnées). La sélectionner si la corrélation est significative. Etc.

Processus de sélection « CONSO »

Forward Selection Process

partial corr. F (p-value)	Step 1	Step 2	Step 3
d.f.	25	24	23
r(Y,Xj*/Xj1,Xj2...)	Poids : 0.9447	Cylindrée : 0.5719	-
R^2	0.8925	0.9277	-
Prix	0.9426 199.19 (0.0000)	0.4567 6.32 (0.0190)	0.1507 0.53 (0.4721)
Cylindrée	0.9088 118.60 (0.0000)	0.5719 11.66 (0.0023)	-
Puissance	0.8883 93.53 (0.0000)	0.4859 7.42 (0.0118)	0.0188 0.01 (0.9288)
Poids	0.9447 207.63 (0.0000)	-	-

Après avoir enlevé l'effet de « Poids » sur l'ensemble des variables (dont l'endogène)

Équivalence avec la méthode basée sur le t^2 de Student (F-Partiel)

Modèle courant (avec constante)	R^2	F-partiel = t^2 (p-value) si ajout de…
Conso = f(cte)	-	Poids : 207.63 (0.0000) Prix : 199.19 (0.0000) Cylindrée : 118.60 (0.0000) Puissance : 93.53 (0.0000)
Conso = f(cte + poids)	0.8925	Cylindrée : 11.6 (0.0023) Puissance : 7.42 (0.0118) Prix : 6.32 (0.0190)
Conso = f(cte + poids + cylindrée)	0.9277	Prix : 0.53 (0.4721) Puissance : 0.01 (0.9288)

La technique basée sur le t^2 repose (en réalité) sur la notion de corrélation partielle.

Conclusion

La colinéarité peut fausser complètement l'interprétation des coefficients de la régression.

Il faut la détecter. Il faut la traiter.

Parmi les traitements possibles : la sélection de variables.
D'autant plus intéressante qu'elle aide à l'interprétation des résultats en mettant en avant les variables les plus intéressantes.

Attention, ce ne sont que des procédures automatiques. Elles peuvent proposer des solutions différentes. Ils faut les voir comme des scénarios que l'on soumet (fait valider par) à l'expertise du domaine.

Bibliographie

En ligne

R. Rakotomalala, « Pratique de la Régression Linéaire Multiple – Diagnostic et sélection de variables ». Support de cours.
http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

R. Rakotomalala. Portail.
http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

Wikipédia.
http://fr.wikipedia.org/wiki/Régression_linéaire_multiple

Ouvrages

M. Tenenhaus, « Statistique – Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.

R. Bourbonnais, « Econométrie – Manuel et exercices corrigés », Dunod, 1998.

Y. Dodge, V. Rousson, « Analyse de régression appliquée », Dunod, 2004.