EXEMPLE QUI ILLUSTRE LE COURS « CHOIX DU MODÈLE » (COURS 3)

FRÉDÉRIC BERTRAND ET MYRIAM MAUMY

On va traiter cet exemple issu du livre « Analyse de régression appliquée » de Yadolah Dodge, Édition Dunod.

Pour illustrer la procédure exhaustive, considérons les données du tableau suivant relatives dans les quartiers de Chicago, où $FIRE$ correspond aux nombres d'incendies pour 1 000 ménages par quartier i de Chicago en 1975. La variable Y correspond au logarithme de $FIRE$ et X_1, X_2 et X_3 correspondent respectivement à la proportion d'habitants construites avant 1940, au nombre de vols et au revenu médian pour le quartier i.

Quartier i	$FIRE$	X_1 x_{i1}	X_2 x_{i2}	X_3 x_{i3}	Y y_i
1	6.2	0.604	29	11.744	1.825
2	9.5	0.765	44	9.323	2.251
3	10.5	0.735	36	9.948	2.351
4	7.7	0.669	37	10.656	2.041
5	8.6	0.814	53	9.730	2.152
6	34.1	0.526	68	8.231	3.529
7	11.0	0.426	75	21.480	2.398
8	6.9	0.735	18	11.104	1.932
9	7.3	0.901	31	10.694	1.988
10	15.1	0.898	25	9.631	2.715
11	29.1	0.827	34	7.995	3.371
12	2.2	0.402	14	13.722	0.788
13	5.7	0.279	11	16.250	1.740
14	2.0	0.077	11	13.686	0.693
15	2.5	0.638	22	12.405	0.916
16	7.7	0.669	37	10.656	2.041
17	8.6	0.814	53	9.730	2.152
18	34.1	0.526	68	8.231	3.529
19	11.0	0.426	75	21.480	2.398
20	6.9	0.735	18	11.104	1.932
21	7.3	0.901	31	10.694	1.988
22	15.1	0.898	25	9.631	2.715
23	29.1	0.827	34	7.995	3.371
24	2.2	0.402	14	13.722	0.788
25	18.5	0.783	22	8.014	2.918
26	23.3	0.790	29	8.177	3.148
27	12.2	0.480	46	8.212	2.501
28	5.6	0.715	23	11.230	1.723
29	21.8	0.731	4	8.330	3.082
30	21.6	0.650	31	5.583	3.073
31	9.0	0.754	39	8.564	2.197
32	3.6	0.208	15	12.102	1.281
33	5.0	0.618	32	11.876	1.609
34	28.6	0.781	27	9.742	3.353
35	17.4	0.686	32	7.520	2.856
36	11.3	0.734	34	7.388	2.425
37	3.4	0.020	17	13.842	1.224
38	11.9	0.570	46	11.040	2.477
39	10.5	0.559	42	10.332	2.351
40	10.7	0.675	43	10.908	2.370
41	10.8	0.580	34	11.156	2.380
42	4.8	0.152	19	13.323	1.569
43	10.4	0.408	25	12.960	2.342
44	15.6	0.578	28	11.260	2.747
45	7.0	0.124	3	10.080	1.946
46	7.1	0.492	23	11.428	1.960
47	4.9	0.466	27	13.731	1.589

Indications de choix de modèle

Comme il y a 3 variables explicatives (X_1, X_2 et X_3), on a $2^3 = 8$ modèles possibles. Appliquons à présent la procédure présentée dans le cours.

Répartissons pour cela ces 8 équations en 4 ensembles :

- **Premier ensemble : l'ensemble A.** Il contient l'unique équation sans variable explicative, c'est-à-dire :
$$Y = \beta_0 + \varepsilon.$$
- **Deuxième ensemble : l'ensemble B.** Il contient les 3 équations avec une seule variable explicative, c'est-à-dire :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon \\ Y &= \beta_0 + \beta_2 X_2 + \varepsilon \\ Y &= \beta_0 + \beta_3 X_3 + \varepsilon. \end{aligned}$$
- **Troisième ensemble : l'ensemble C.** Il contient les 3 équations avec deux variables explicatives, c'est-à-dire :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \\ Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \end{aligned}$$
- **Quatrième ensemble : l'ensemble D.** Il contient l'unique équation avec trois variables explicatives, c'est-à-dire le modèle complet :
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Critère du R^2

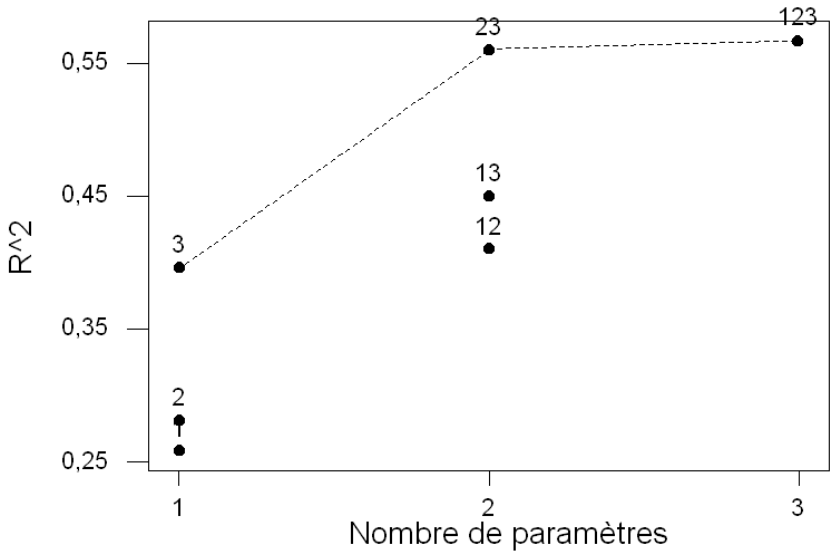
Ordonnons les équations à l'intérieur des ensembles B et C selon la valeur du coefficient de détermination multiple R^2.

1. Ensemble B :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon \quad R^2 = 25,8\% \\ Y &= \beta_0 + \beta_2 X_2 + \varepsilon \quad R^2 = 28,2\% \\ Y &= \beta_0 + \beta_3 X_3 + \varepsilon \quad R^2 = 39,6\%. \end{aligned}$$

2. Ensemble C :
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad R^2 = 41,0\% \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \quad R^2 = 45,0\% \\ Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad R^2 = 56,0\%. \end{aligned}$$

Notons en outre que pour le modèle complet, nous avons un coefficient de détermination multiple $R^2 = 56,7\%$ (alors que pour le modèle constant, nous avons évidemment un coefficient de détermination multiple $R^2 = 0\%$). L'introduction de la seule variable X_3 permet d'obtenir un coefficient de détermination multiple $R^2 = 39,6\%$. Par ailleurs, le fait d'avoir dans le modèle simultanément X_2 et X_3 nous donne un coefficient de détermination multiple $R^2 = 56\%$ proche

du coefficient de détermination multiple R^2 du modèle complet. L'introduction supplémentaire de X_1 ne permet donc pas au coefficient de détermination multiple R^2 de s'améliorer notablement puisque le coefficient de détermination multiple R^2 passe seulement de 56% à $56,7\%$. Par conséquent, l'ajout de la variable X_1 quand X_2 et X_3 sont déjà incluses dans l'équation de régression n'améliore guère le modèle. Ceci est illustré par la figure ci-dessous où le nombre de paramètres p est représenté en abscisse et la valeur du coefficient de détermination multiple R^2 en ordonnée. Les points représentent le modèle. Les chiffres à côté de chaque point correspondent aux variables incluses dans le modèle. Cette approche graphique est pratique pour visualiser les résultats.



En observant la position des points, on remarque que le modèle à une variable explicative ($p = 2$) qui explique le mieux Y, est le modèle :

$$Y = \beta_0 + \beta_3 X_3.$$

L'introduction de X_2 dans le modèle comprenant déjà X_3 entraîne une hausse du coefficient de détermination multiple R^2 pour l'amener à 56%. Ce gain s'exprime sur le graphe par une pente élevée pour la droite en pointillé entre le modèle symbolisé par **3** (pour X_3) et celui symbolisé par **23** (pour X_2 et X_3). Le passage du modèle **23** au modèle complet symbolisé par **123** n'apporte pratiquement aucune contribution supplémentaire au coefficient de détermination multiple R^2. En effet, la droite entre **23** et **123** présente une pente relativement faible. Par conséquent, en se basant sur le critère du coefficient de détermination multiple R^2, l'équation comprenant les variables X_2 et X_3 s'avère être la meilleure, compte tenu de l'ensemble des variables considérées.

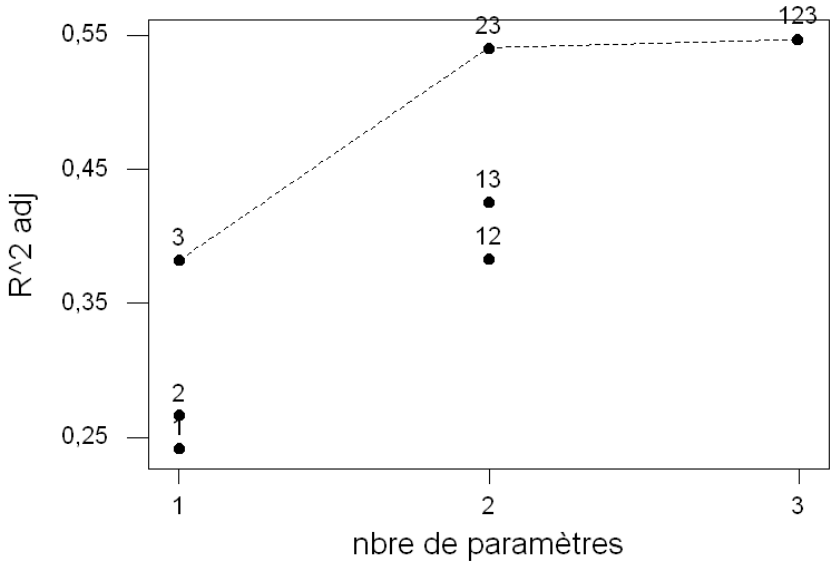
Critère du R^2 ajusté

Le tableau ci-dessous donne le coefficient de détermination multiple R^2 et le coefficient de détermination multiple ajusté R^2 pour tous les modèles possibles (les modèles sont symbolisés de la même manière que ci-dessus)

Tableau présentant le R^2 et le R^2 ajusté.

Modèle	R^2	R^2_{aj}
3	0.396	0.382
2	0.281	0.266
1	0.258	0.241
23	0.560	0.540
13	0.450	0.425
12	0.410	0.383
123	0.567	0.547

Pour chaque modèle, on remarque que la valeur du coefficient de détermination multiple ajusté R^2_{aj} ne diffère que faiblement de celle du coefficient de détermination multiple R^2. De plus, on constate que le passsage d'une équation comprenant un seul terme explicatif à une équation en comprenant deux, entraîne un accroissement non négligeable de la valeur de R^2_{aj}. En appliquant strictement le critère du R^2_{aj}, on est orienté vers le modèle complet pour lequel on a un R^2_{aj} maximal de $54,7\%$.
On résume ces informations dans le graphique suivant :



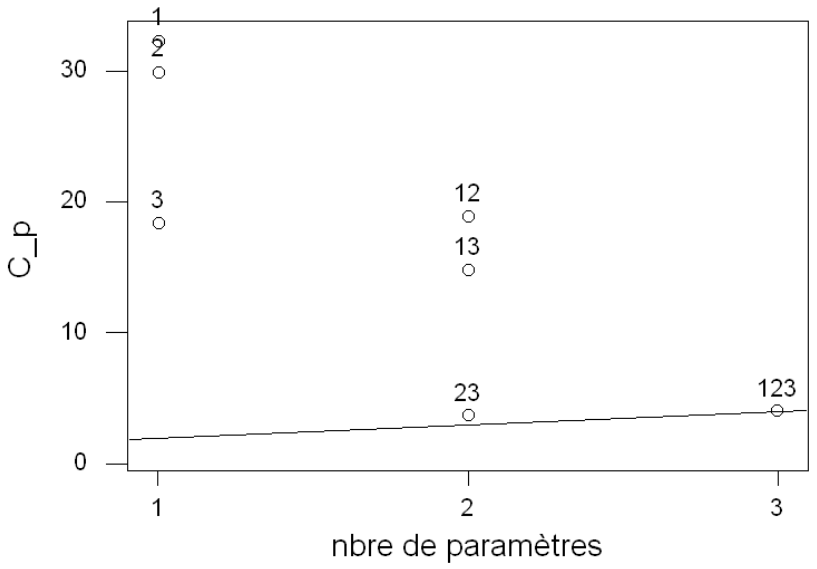
Critère du C_p de Mallows

Le troisième critère, le C_p de Mallows, permet de sélectionner la meilleure équation en fonction du biais du modèle. On a le tableau suivant :

Tableau présentant le R^2, le R^2 ajusté et le C_p de Mallows.

Modèle	R^2	R^2_{aj}	C_p
3	0.396	0.382	$18,354$
2	0.281	0.266	$29,937$
1	0.258	0.241	$32,356$
23	0.560	0.540	$3,681$
13	0.450	0.425	$14,817$
12	0.410	0.383	$18,936$
123	0.567	0.547	$4,000$

Dans un modèle comprenant une seule variable explicative, deux paramètres sont à estimer. De ce fait C_p devrait approcher la valeur de $p = 2$ ce qui n'est pas le cas dans notre exemple. Un modèle comprenant deux variables explicatives devrait restituer un C_p proche de 3. En se référant au tableau ci-dessus, on obtient une valeur proche de 3 pour le modèle sumbolisé par **23**. La valeur de $C_p = 4$ obtenue pour le modèle complet n'est pas intéressante puisque le critère de sélection associé au C_p ne s'apllique pasau modèle complet. On trace alors sur un même graphique la droite $C_p = p$ et les différentes valeurs de C_p obtenues pour les modèles **1**, **2**, **3**, **12**, **13**, **23** et **123**.



On voit alors que le modèle sélectionné par le critère du C_p est le modèle **23**.