EXEMPLE DÉTAILLÉ QUI ILLUSTRE LE COURS 4 "CHOIX DE MODÈLE"

FRÉDÉRIC BERTRAND ET MYRIAM MAUMY

Cet exemple est issu du livre de Yadolah Dodge intitulé *Analyse de régression appliquée*, éditions Dunod.

Pour illustrer le cours 4 "Choix de modèle", considérons les données du tableau suivant (Source : Andrews et Herzberg (1985)) relatives dans les quartiers de Chicago, où $FIRE$ correspond aux nombres d'incendies pour 1 000 ménages par quartier i de Chicago en 1975.

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.

1. A l'aide de la procédure de choix de modèle adéquate étudier la liaison entre la variable $Fire$ et les variables X_1, X_2, X_3. Cet ajustement est-il satisfaisant ? Détailler soigneusement la méthode et les critères de décision utilisés.

2. A l'aide de la procédure de choix de modèle adéquate étudier la liaison entre la variable $Y = \ln(Fire)$ et les variables X_1, X_2, X_3. Cet ajustement est-il désormais satisfaisant ? Détailler soigneusement la méthode et les critères de décision utilisés.

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	0.114	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 $2^3 = 8$ moèles 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 2 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 3 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 :**
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \varepsilon \quad R^2 = 26,2\% \\ Y &= \beta_0 + \beta_2 X_2 + \varepsilon \quad R^2 = 19,4\% \\ Y &= \beta_0 + \beta_3 X_3 + \varepsilon \quad R^2 = 25,8\% \end{aligned}$$

2. **Ensemble C :**
$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad R^2 = 37,5\% \\ Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \quad R^2 = 34,2\% \\ Y &= \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad R^2 = 51,7\% \end{aligned}$$

Notons en outre que pour le modèle complet, nous avons $R^2 = 53,2\%$ (alors que pour le modèle constant, nous avons évidemment $R^2 = 0\%$). L'introduction de la seule variable X_1 permet d'obtenir un $R^2 = 26,2\%$. Par ailleurs, le fait d'avoir dans le modèle simultanément X_2 et X_3 nous donne un $R^2 = 51,7\%$ proche du R^2 du modèle complet. L'introduction supplémentaire de X_1 ne permet donc pas au R^2 de s'améliorer notablement puisque R^2 passe seulement de $51,7\%$ à $53,2\%$. 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 le tableau ci-dessous où le nombre de paramètres p est représenté dans la première colonne et la valeur de R^2 dans a suivante. Les points représentent le modèle. Les croix dans les dernières colonnes correspondent aux variables incluses dans le modèle.

Vars	R^2	R^2_{ajust}	C_p	S	X – 1	X – 2	X – 3
1	26,2	24,6	24,8	0,62752	X		
1	25,8	24,2	25,2	0,62922			X
1	19,4	17,6	31,1	0,65603		X	
2	51,7	49,5	3,4	0,51360		X	X
2	37,5	34,6	16,5	0,58424	X	X	
2	34,2	31,2	19,5	0,59937	X		X
3	53,2	50,0	4,0	0,51119	X	X	X

On construit alors a représentation graphique suivante : le nombre de paramètres p est représenté en abscisse et la valeur de 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 R^2 pour l'amener à $51,7\%$. 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 R^2. En effet, la droite entre 23 et 123 présente une pente relativement faible. Par conséquent, en se basant sur le critère de 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 R^2 ajusté pour tous les modèles possibles.

Pour chaque modèle, on remarque que la valeur R^2_{aj} ne diffère que faiblement de celle de R^2. De plus, on constate que le passsage d'une équation comprenant un seul terme explicatif à une équation en comprenant 2, entraîne un accroissement non négligeable de la valeur 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 $50,0\%$.

Critère du C_p
Le tableau ci-dessous donne le C_p pour tous les modèles possibles.

On constate que seul le modèle contenant X_2 et X_3 présente un C_p proche de p (3,4 proche de 3), le cas du modèle complet étant à exclure lors d'un choix à l'aide d'un critère basé sur un C_p puisque l'on sait que $C_p = p$.