Régression multiple

Tables des matières :

Objectifs

Le but de la régression multiple est d'expliquer au mieux la variabilité d'une variable Y par celle d'une série de variables X1, X2, X3 etc….

Le modèle linéaire se construit sur une équation qui oriente une droite dans un hyper-espace :

$Y = B0 + B1X1 + B2X2 + B3X3 + \ldots.$

le terme B0 permettant un écart entre le point Y=0 et l'origine du système d'axe des X.

Le modèle polynomial remplace les différentes variables X par une seule variable élevée à la seconde, troisième, quatrième … puissance. Cette équation permet de tracer une courbe qui pourra avoir autant de points d'inflexion qu'il y a de degrés au polynôme :

$Y = B0 + B1X + B2X^2 + B3X^3 + \ldots.$

le terme B0 permettant un écart entre le point Y=0 et l'origine du système d'axe des X.

Principe

Le critère qui permet de déterminer le meilleur jeu de valeurs pour les paramètres B0, B1, B2, B3 … est la maximisation du coefficient de détermination

$$R^2 = \text{Var(Y modélisé)/ Var(Y observé)}$$

ou, ce qui revient au même,la minimisation globale des écarts entre les valeurs de Y modélisées et observées .

Le principe de la méthode consiste à développer l'équation de la somme des carrés des écarts entre les observations et le modèle, pour Y, (SCERy) et d'en calculer la dérivée partielle par rapport à chacun des paramètres. Le minimum de la fonction SCERy correspond au point où toutes les dérivées partielles sont nulles.

Ce système d'équation peut être résolu de façon analytique par le calcul matriciel suivant :

$$\mathbf{b = (X'X)^{-1}X'Y}$$

X est une matrice de genre n x p comprenant l'ensemble des n valeurs de X1, X2 … Xp-1. On ajoute à X une colonne constante (Xp =1 partout) pour estimer le paramètre libre B0.

Y est un vecteur colonne n x 1 comprenant l'ensemble des n valeurs de Y

Le produit matriciel combine donc les genres suivants :

$$p \times n \sim n \times p \sim p \times n \sim n \times 1$$

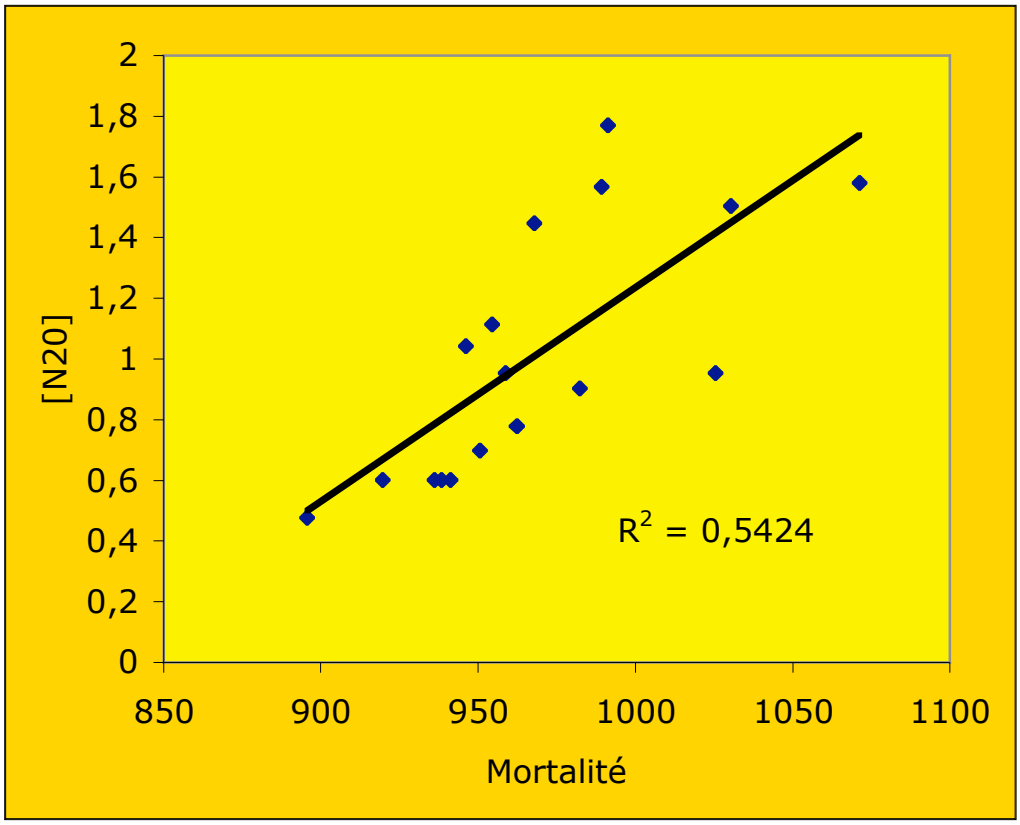
Il est défini et produit le vecteur colonne **b** de genre p x 1, qui reprend la valeur des p paramètres Bo, B1, B2, … B p-1.

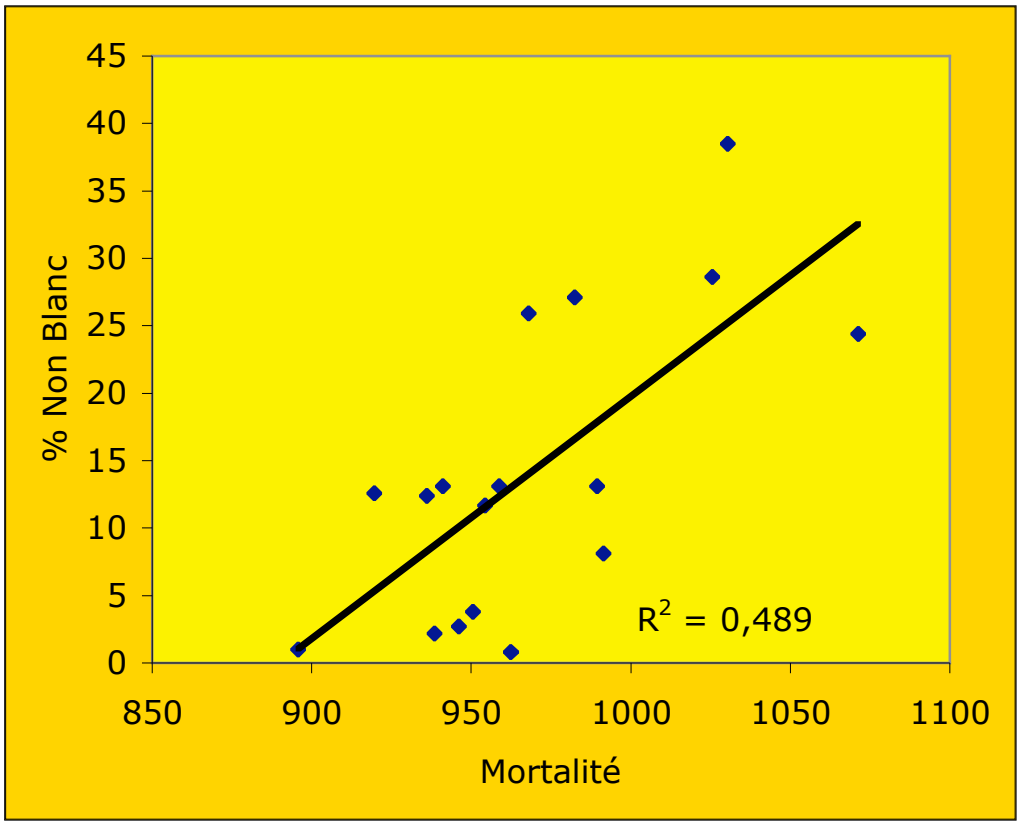
Exemple de modèle linéaire

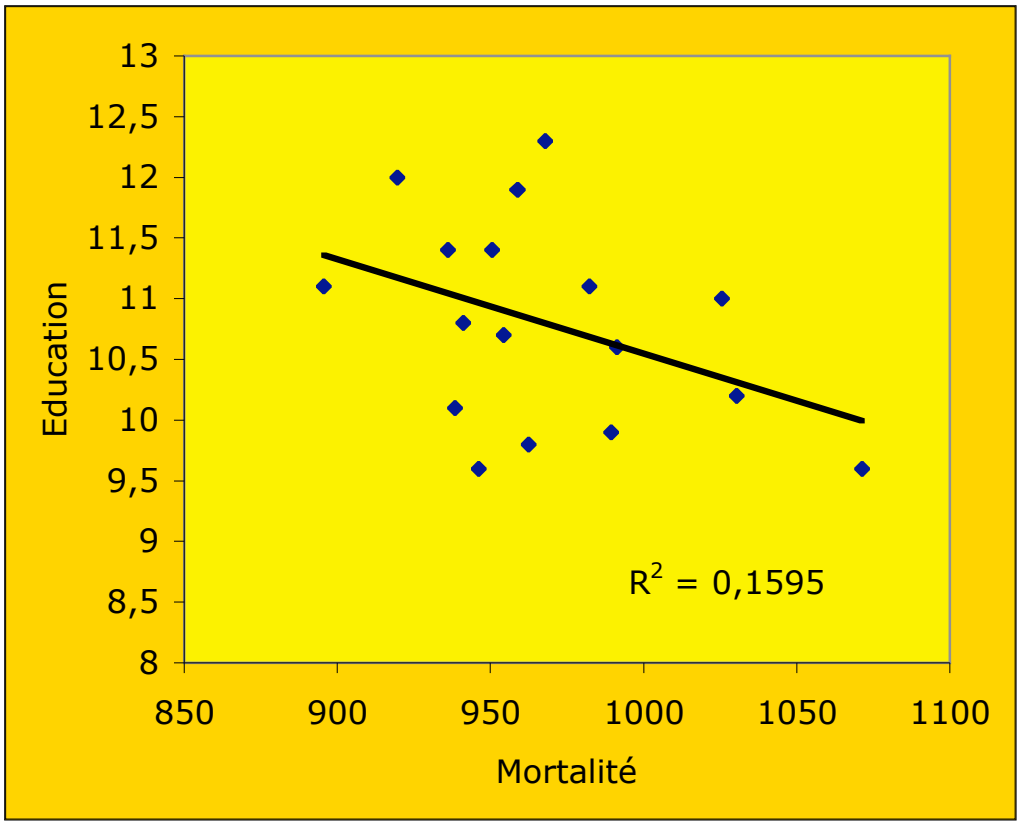
La mortalité, la composition raciale ,un polluant atmosphérique (protoxyde d'azote, N2O) et le niveau d'éducation ont été relevés dans quelques villes américaine (subset des données de http://lib.stat.cmu.edu/DASL/Datafiles/SMSA.html)

Ville USA	Mortalité	%NonBlanc	N2O	Education
Allentown, Bethlehem, PA-NJ	962,35	0,80	0,78	9,8
Atlanta, GA	982,29	27,10	0,90	11,1
Baltimore, MD	1071,29	24,40	1,58	9,6
Birmingham, AL	1030,38	38,50	1,51	10,2
Columbus, OH	958,84	13,10	0,95	11,9
Flint, MI	941,18	13,10	0,60	10,8
Dayton-Springfield, OH	936,23	12,40	0,60	11,4
Kansas City, MO	919,73	12,60	0,60	12
Louisville, KY-IN	989,26	13,10	1,57	9,9
Pittsburgh, PA	991,29	8,10	1,77	10,6
Providence, RI	938,5	2,20	0,60	10,1
Richmond-Petersburg, VA	1025,5	28,60	0,95	11
Syracuse, NY	950,67	3,80	0,70	11,4
Washington, DC-MD-VA	967,8	25,90	1,45	12,3
Reading, PA	946,19	2,70	1,04	9,6
Worcester, MA	895,7	1,00	0,48	11,1
Youngstown-Warren, OH	954,44	11,70	1,11	10,7

L'analyse à deux variables produit les résultats suivants :






13
12,5
12
11,5
11
10,5
10
9,5
9
8,5
8
Education
850
900
950
1000
1050
1100
Mortalité
R2 = 0,1595

L'analyse en régression multiple se base sur les matrices suivantes :

Y	X			
962,35	1	0,80	0,78	9,8
982,29	1	27,10	0,90	11,1
1071,29	1	24,40	1,58	9,6
1030,38	1	38,50	1,51	10,2
958,84	1	13,10	0,95	11,9
941,18	1	13,10	0,60	10,8
936,23	1	12,40	0,60	11,4
919,73	1	12,60	0,60	12
989,26	1	13,10	1,57	9,9
991,29	1	8,10	1,77	10,6
938,50	1	2,20	0,60	10,1
1025,50	1	28,60	0,95	11
950,67	1	3,80	0,70	11,4
967,80	1	25,90	1,45	12,3
946,19	1	2,70	1,04	9,6
895,70	1	1,00	0,48	11,1
954,44	1	11,70	1,11	10,7

Et produit le vecteur **b** calculé suivant:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

par la fonction excel :

```
=PRODUITMAT(INVERSEMAT(PRODUITMAT(TRANSPOSE(X);X));
  PRODUITMAT(TRANSPOSE(X);Y))
```

b0	1109,57
b1	36,0395
b2	2,2969
b3	-19,456

Les valeurs prédites par le modèle sont calculées en appliquant l'équation:

Y mod = 1109,57 + 36,04 X1 + 2,29 X2 -19,45 X3

Y	observé	modélisé	écarts
Allentown, Bethlehem, PA-NJ	962,35	948,79	13,56
Atlanta, GA	982,29	988,4	-6,11
Baltimore, MD	1071,29	1035,77	35,52
Birmingham, AL	1030,38	1053,8	-23,42
Columbus, OH	958,84	942,53	16,31
Flint, MI	941,18	951,23	-10,05
Dayton-Springfield, OH	936,23	937,95	-1,72
Kansas City, MO	919,73	926,74	-7,01
Louisville, KY-IN	989,26	1003,56	-14,3
Pittsburgh, PA	991,29	985,76	5,53
Providence, RI	938,5	939,82	-1,32
Richmond-Petersburg, VA	1025,5	995,64	29,86
Syracuse, NY	950,67	921,69	28,98
Washington, DC-MD-VA	967,8	981,91	-14,11
Reading, PA	946,19	966,53	-20,34
Worcester, MA	895,7	913,1	-17,4
Youngstown-Warren, OH	954,44	968,41	-13,97
Variance	1893,44	= 1547,85	+ 345,59
R2 =	1547,85	/ 1893,44	= 0,82

Soit une forte augmentation de la valeur prédictive pour la fonction des variables prises ensemble.

La régression pas à pas (stepwise)

La question qui suit généralement l'approche par la régression multiple est de choisir parmi les variables X le plus petit nombre d'entre elles qui expliquent au mieux la variabilité de Y.

Une méthode courante est une régression itérative qui inclut d'abord dans le modèle la variable qui propose le meilleur coefficient de détermination. Ensuite, celle qui améliore le plus le coefficient de détermination et ainsi de suite.

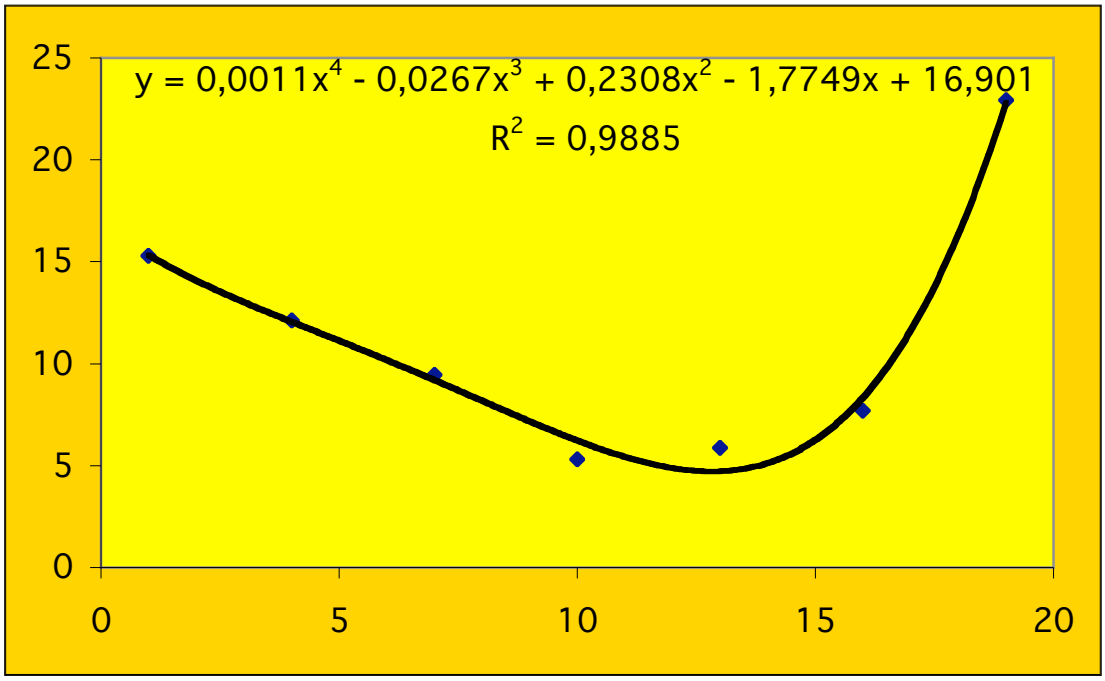
Alternativement, toutes les variables sont entrées dans le modèle et les variables sont progressivement exclues, en fonction de celles qui contribuent le moins au modèle.

Il faut noter que la seconde variable qui entre dans le modèle n'est pas forcément celle qui présente, à elle seule, le second meilleur coefficient de détermination avec Y. Sinon, la solution serait triviale. En effet, X1 et X2 peuvent être très corrélées, voire quasi redondantes. Dans ce cas la qualité du modèle ne sera pas améliorée. C'est donc la variable qui contribue le plus à réduire la variabilité résiduelle, du modèle en voie d'élaboration qui sera sélectionnée à chaque étape.

La régression en Excel

La solution **b = (X'X)$^{-1}$X'Y** est mise en pratique par le logiciel Excel pour la régression linéaire simple et la régression polynomiale.

X	Y
1	14,26
4	13,54
7	10,98
10	4,94
13	1,72
16	9,19
19	24,81



Exemple de régression polynomiale
(menu Graphique« Ajouter une courbe de tendance »)

Aucune solution n'est proposée pour le modèle linéaire multiple, mais le calcul se réalise facilement à l'aide des fonctions TRANSPOSE( ), PRODUITMAT( ) et INVERSEMAT( ).

Les modèles logarithmiques , exponentiels et puissance sont calculés par la régression linéaire simple via la transformation de X et/ou de Y en log, la solution étant retransformée en anti-log.

Le logiciel Excel ne réalise pas de régression non linéaire proprement dite. Les fonctions sigmoïdes, multiples exponentielles, Michaëlis Menten … n'ont pas de solution analytique et doivent être réalisée par un logiciel qui propose un algorithme de minimisation numérique. A noter que même les fonctions linéarisables (exponentielle, puissance…) tirent avantage de cette approche.

Pour faire de l'inférence, notamment pour obtenir l'intervalle de confiance des paramètres, il est préférable d'utiliser un logiciel statistique plus sophistiqué que le tableur.

Limites de l'interprétation

Les principes et les mises en garde concernant les limites de cette approche sont développés au module 20 dans le cadre de l'équation la plus simple Y = B0 + B1X1 . Ce sont les points spécifiques à la généralisation du modèle qui seront abordés ici.

Les précautions suivantes doivent être prises pour interpréter les résultats :

Plus on complexifie le modèle, plus la variabilité résiduelle peut être – apparemment- expliquée. Le nombre d'observations doit être relativement grand par rapport au nombre de variables incluses dans le modèle. Bien qu'il n'existe aucune règle absolue en cette matière on se référera au minimum à la règle empirique n > 2p

Les coefficients sont délicats à interpréter. En effet, B1 donne la variation de X1 correspondant à l'augmentation d'une unité de Y, pour autant que X2 reste constant. En pratique, cela est irréaliste car X1 est généralement corrélé à X2.

Les relations bivariées doivent préalablement être explorées graphiquement. La présence de données extrêmes ou aberrantes, la non linéarité de certaines relation, les écarts systématiques au modèle sont susceptible d'affecter grandement les résultats.

Les conditions de linéarité étant souvent précaires et limitées à un domaine de X, l'extrapolation des résultats est toujours hasardeuse.

La régression polynomiale produit un modèle très « plastique » qui interpolent bien les points mais ne possède aucune valeur d'extrapolation. La valeur des paramètres ne peut pas être associée à une explication structurelle du phénomène décrit.