CODAGE DES VARIABLES DANS LES MODÈLES STATISTIQUES

Il existe plusieurs raisons pour utiliser un codage des variables dans le développement des modèles statistiques. Les principales raisons sont :

- **Les variables catégoriques doivent être transformées en variables numériques avec un codage neutre permettant leur utilisation dans les équations.**
- **Les grandeurs des coefficients des équations de prédiction sont comparables car les variables codées varient entre -1 et +1 sur une échelle identique qui ne tient pas en compte les unités spécifiques de chacune des variables originelles. Il est alors possible de comparer la contribution relative de leur impact sur la variable de réponse. Cela n'est pas le cas lorsque les variables ne sont pas codées.**
- **L'intercepte β_0 (effet général) peut s'interpréter comme la valeur de la réponse au centre de l'espace expérimental.**
- **Les colinéarités sont amoindries et, dans certains cas, elles sont éliminées. En particulier, lors de l'ajustement de polynôme du second degré.**
- **Les inexactitudes dans les calculs attribuables aux erreurs d'arrondies sont amoindries.**
- **Les modélisations de la réponse avec les variables codées ou avec les variables non codées sont équivalentes car elles donnent lieu :**
  - **au même tableau d'analyse de la variance (ANOVA);**
  - **au même test de signification (test-t);**
  - **aux mêmes coefficients de détermination R^2 et R_{ajust}^2;**
  - **aux mêmes valeurs prédites;**
  - **à la même analyse de résidus.**
- **La seule différence concerne les coefficients de régression : ils sont différents dans le modèle avec les variables originelles et dans le modèle avec les variables codées.**
- **Les coefficients des équations avec les variables originelles ne peuvent pas être comparés car ils tiennent en compte les unités dans lesquelles les variables s'expriment.**

A) cas d'un facteur quantitatif X variant dans l'intervalle [a, b]

Posons $a = \min(X)$ $b = \max(X)$

W est la variable de codage associée à X

$$W = (X - c) / d$$

où $c = (a + b)\ 2$ **: point milieu de l'intervalle** $[a, b]$

$d = (b - a)/2$ **: demi-longueur de l'intervalle** $[a, b]$

l'intervalle de variation de la variable W est [-1, + 1].

Le choix de c est commode car le « centre » du nouvel espace expérimental de la variable W est 0.

Le choix de d pour définir le diviseur est arbitraire mais commode.

Il est possible de faire un autre choix mais il est important que toutes les variables codées varient dans un intervalle identique.

Dans le cas où l'on dispose d'une série d'observations $x_1, x_2, \ldots, x_n$ on peut utiliser le codage provenant de la forme centrée-réduite de X définit par

$$W = (X - Xbar) / ET(X)$$

avec $Xbar = (1/n) \sum x_i$ $\quad ET(X) = [\ (1/(n-1)) \sum (x_i - Xbar)^2\ ]^{0,5}$

Cas particuliers

Si la variable prend seulement les valeurs a et b, alors W prend les valeurs -1 et +1.
Si le facteur prend seulement les valeurs a, c, b, alors W prend les valeurs, -1, 0, +1

B) cas d'un facteur qualitatif Z variant à 2 modalités m1 et m2

La variable Z est remplacée par variable U définit par :

$U = -1$ si $Z = m1$
$U = +1$ si $Z = m2$

L'assignation de m1 à -1 est arbitraire. On aurait pu utiliser m_2.
La variable de codage U mesure l'effet (différence) sur la réponse Y entre la modalité m1 et la modalité m2.

C) cas d'une variable catégorique Z variant à k modalités ($k \geq 3$) : codage à effet

Posons $m_1, m_2, \ldots, m_k$ les k modalités de Z.

On fait la création de de k – 1 variables $U_1, U_2, \ldots, U_{k-1}$ dont les valeurs sont -1, 0, 1.

On choisit une modalité de référence disons m_k. Ce choix est arbitraire.

La définition des variables de codage $U_1, U_2, \ldots, U_{k-1}$ est :

$U_1 = 1$ si $Z = m_1$ $\quad U_1 = 0$ si $Z = m_2, m_3, \ldots, m_{k-1}$ $\quad U_1 = -1$ si $Z = m_k$

$U_2 = 0$ si $Z = m_1$ $\quad U_2 = 1$ si $Z = m_2$ $\quad U_2 = 0$ si $Z = m_3, \ldots, m_{k-1}$
$U_2 = -1$ si $Z = m_k$

……..

$U_{k-1} = 0$ si $Z = m_1, m_2, \ldots, m_{k-2}$ $\quad U_{k-1} = 1$ si $Z = m_{k-1}$ $\quad U_{k-1} = -1$ si $Z = m_k$

On peut écrire la définition des variables $U_1, U_2, \ldots, U_{k-1}$ sous forme matricielle

Z	U_1	U_2	$U_3 \ldots U_{k-2}$	U_{k-1}
m_1	1	0	0 …… 0	0
m_2	0	1	0 …… 0	0
……				
m_{k-1}	0	0	0 …… 0	1
m_k	-1	-1	-1 ….. -1	-1

On peut aussi écrire la relation entre Z et les nouvelles variables $U_1, U_2, \ldots, U_{k-1}$

Z	m_1	m_2	$m_3 \ldots m_{k-1}$	m_k
U_1	1	0	0 …… 0	-1
U_2	0	1	0 …… 0	-1
……				
U_{k-1}	0	0	0 1	-1

Ce codage porte le nom de *codage à effet.*
Le codage B) plus haut (k=2) est un cas particulier du codage C).
Une variable catégorique Z avec plusieurs k modalités fait la création de (k-1) variables de codage U. Il est donc recommandé d'agréger des modalités afin de limiter le nombre des variables de codage possède un grand nombre de modalités.

Les variables de codage U correspondant à des variables catégoriques Z peuvent intervenir dans des modèles avec des effets d'interaction de type XZ impliquant des variables catégoriques Z et des continues X. On utilisera des termes XU. Toutefois les variables de codage ne peuvent pas avoir d'effet quadratique car $U^2 = 1$

Exemple d'un codage à effet

Dans une étude statistique, une variable catégorique Z prend les 5 modalités suivantes :

Z : municipalité, occupant, utilité, autre, contracteur

On a fait la création d'un codage à effet de Z en définissant 4 variables U_1, U_2, U_3, U_4 à valeurs -1, 0, 1 en utilisant la modalité *contracteur* comme la modalité de référence. On aurait pu choisir toute autre modalité.
Le tableau suivant indique les valeurs prises par ces 4 variables de codage.

Z	U_1	U_2	U_3	U_4	Y : réponse
municipalité	1	0	0	0	y_1
occupant	0	1	0	0	y_2
utilité	0	0	1	0	y_3
autre	0	0	0	1	y_4
contracteur	-1	-1	-1	-1	y_5

Le modèle entre Y et Z prend alors la forme

$$Y = \beta_0 + \beta_1 * U_1 + \beta_2 * U_2 + \beta_3 * U_3 + \beta_4 * U_4$$

Par exemple, la variable U_1 mesure l'effet (différence) β_1 sur la variable de réponse Y entre la modalité *municipalité* et la modalité *contracteur*.

Pour mesurer l'effet sur Y entre 2 modalités, par exemple, entre *occupant* et *utilité,* on fait la différence entre leurs coefficients respectifs : $\beta_2 - \beta_3$

D) cas d'une variable catégorique Z variant à k modalités ($k \geq 3$) : codage disjonctif complet

On fait la création de k variables U1, U2, …, Uk chacune prenant la valeur 1 pour la chaque modalité spécifique et 0 pour les autres modalités.

Exemple : Z variable catégorique avec 5 modalités : m1, m2, m3, m4, m5
Le tableau suivant indique les valeurs prises par ces 5 variables de codage.

Z	U_1	U_2	U_3	U_4	U_5	Y : réponse
municipalité	1	0	0	0	0	y_1
occupant	0	1	0	0	0	y_2
utilité	0	0	1	0	0	y_3
autre	0	0	0	1	0	y_4
contracteur	-0	0	0	0	1	y_5

Mais cela fait la création d'un problème de multicolinéarité : U1 + U2 + U3 + U4 + U5 = 1

E) codage disjonctif complet modifié

on retient seulement 4 des 5 variables, disons U1, U2, U3 U4

examen du modèle : $Y = \beta_0 + \beta_1 * U_1 + \beta_2 * U_2 + \beta_3 * U_3 + \beta_4 * U_4$

Si $U1 = U2 + U3 = U4 = 0$ **alors** $Y = \beta_0$

l'effet général β_0 est confondu avec celui de la modalité m5.
Cela n'est pas intéressant car on préfère un effet général qui soit indépendant des variables (facteurs) X

CONCLUSION

le codage à effet est préférable ; c'est celui qui est choisi dans STATISTICA.