

UTILISATION DE VARIABLES MUETTES (DUMMY)

Source : D.N.Gujarati, "Basic Econometrics", Third Ed., McGraw Hill, 1995

Le concept

Les variables muettes (ou binaires ou "dummy") permettent la prise en compte dans des modèles de régression de phénomènes qualitatifs (grèves, les saisons, ...). Elles enrichissent considérablement la portée de l'analyse de régression. Le principe de base est simple : il s'agit de coder en 0/1 l'événement en question.

Régression avec une variable quantitative et une variable qualitative avec deux classes

Le modèle est le suivant : $Y_i = a + b_1 D_i + b_2 X_i + u_i$ avec $D_i = 0$ ou $= 1$. L'interprétation du modèle est bien la suivante :

- $E(Y_i | X_i, D_i = 0) = a + b_2 X_i$
- $E(Y_i | X_i, D_i = 1) = a + b_1 + b_2 X_i$

Le modèle permet donc bien de capter un changement d'ordonnée à l'origine.

Régression avec une variable quantitative et une variable qualitative avec plus de deux classes

Le modèle est le suivant : $Y_i = a + b_1 D_{1i} + b_2 D_{2i} + b_3 X_i + u_i$ avec D_{1i}
et $D_{2i} = 0$ ou $= 1$.

Les variables muettes D_{1i} et D_{2i} sont mutuellement exclusives et permettent de coder l'existence de trois classes ($D_{1i}=1$ et $D_{2i}=0$ pour la classe 1, $D_{1i}=0$ et $D_{2i}=1$ pour la classe 2 et $D_{1i}=0$ et $D_{2i}=0$ pour la classe 3). Le modèle permet bien de capter les changements d'ordonnée à l'origine associés aux changements de classe.

Régression avec une variable quantitative et deux variables qualitatives

L'approche suivie est la même qu'au point précédant. Les variables muettes codent cette fois les modalités de variables qualitatives différentes. Les phénomènes captés sont toujours des changements d'ordonnée à l'origine. On peut, en prenant les produits des variables muettes prises deux à deux, capter les interdépendances entre variables qualitatives. Les produits de variables permettent alors de capter les changements de pentes (cfr infra).

Comparaison de deux modèles de régression (changement structurel)

Lorsque l'on étudie un phénomène dans le temps, la nature des relations peut se modifier suite à des changements structurels. On pensera, par exemple, à l'impact de la seconde guerre mondiale dans l'étude des phénomènes économiques.

EXEMPLE : étude de la relation entre épargne et revenu

DONNEES

Temps	Epargne	Revenu	Dt
1946	0.36	8.8	1
1947	0.21	9.4	1
1948	0.08	10	1
1949	0.2	10.6	1
1950	0.1	11	1
1951	0.12	11.9	1
1952	0.41	12.7	1
1953	0.5	13.5	1
1954	0.43	14.3	1
1955	0.59	15.1	0
1956	0.9	16.7	0
1957	0.95	17.7	0
1958	0.82	18.6	0
1959	1.04	19.7	0
1960	1.53	21.1	0
1961	1.94	22.8	0
1962	1.75	23.9	0
1963	1.99	25.2	0

MODELE 1 : $Y_t = a + bX_t + e_t$

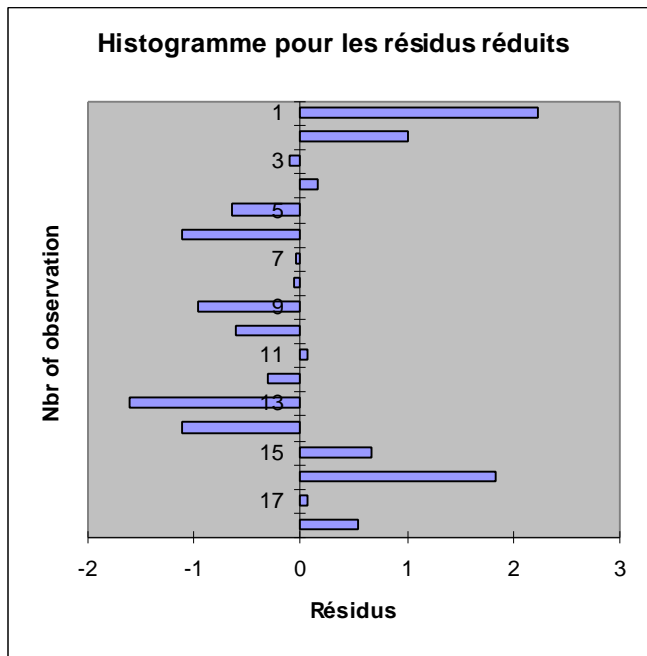
	Degrés de	Somme de	Carrés mo:	F de	Fishe	Pr > F
Modèle	1	6.4646	6.4646	184.7851		0.0001
Résidus	16	0.5598	0.0350			
Total	17	7.0244				

Analyse du modèle (Type I SS) :

Source	DF	Type I SS	arrés moy	F de Fisher	Pr. >F
Revenu	1	6.4646	6.4646	184.7851	0.0001

Paramètres de la régression et statistiques correspondantes :

	Valeur	Ecart-type	de Student	lité corresp	de l'interva	de l'intervalle à 95%
Constante	-1.08055	0.1433	-7.5390	0.0001	-1.3844	-0.7767
Revenu	0.117915	0.0087	13.5936	0.0001	0.0995	0.1363



MODELE 2 : $Y_t = a + bD_t + gX_t + c(D_t X_t) + e_t$

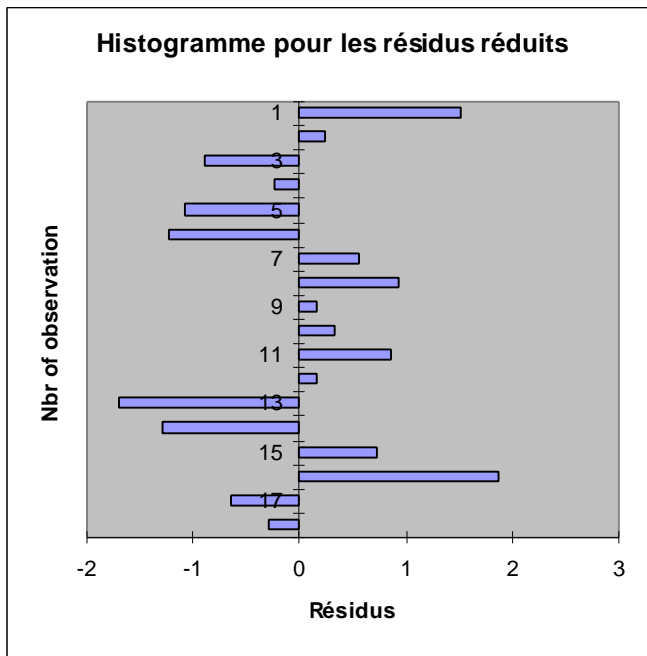
	Degrés de	Somme de Carrés	mo	F de Fisher	Pr > F
Modèle	3	6.6884	2.2295	92.8970	0.0001
Résidus	14	0.3360	0.0240		
Total	17	7.0244			

Analyse du modèle (Type I SS) :

Source	DF	Type I SS	arrés moy	F de Fisher	Pr. >F
Revenu	1	6.4646	6.4646	269.3672	0.0001
Dt	1	0.0060	0.0060	0.2499	0.6249
Interaction	1	0.2178	0.2178	9.0738	0.0093

Paramètres de la régression et statistiques correspondantes :

	Valeur	Ecart-type	de Student	ité corresp	de l'interva	de l'interva
Constante	-1.67825	0.3260	-5.1478	0.0001	-2.3775	-0.9790
Revenu	0.147203	0.0160	9.1865	0.0001	0.1128	0.1816
Dt	1.412004	0.4674	3.0210	0.0092	0.4096	2.4145
Interaction	-0.10018	0.0333	-3.0123	0.0093	-0.1715	-0.0288



Modélisation d'un facteur saisonnier

Une solution pour évaluer la composante saisonnière d'une série temporelle est l'utilisation, dans un modèle de régression, de variables binaires. Si la composante en question est trimestrielle, le modèle prend la forme suivante :

$$Y_t = a_1 + a_2 D_{1,t} + a_3 D_{2,t} + a_4 D_{3,t} + b_1 X_t + u_t$$

où $D_{1,t}$ est égal à 1 pour le 1^o trimestre et 0 sinon, $D_{2,t}$ est égal à 1 pour le 2^o trimestre et 0 sinon et $D_{3,t}$ est égal à 1 pour le 3^o trimestre et 0 sinon.

Il ne faut bien entendu pas introduire de variable binaire pour le 4^o trimestre (cette variable serait une combinaison linéaire des trois précédentes). Rien n'empêche de n'introduire que l'une ou l'autre des variables binaires si l'on estime que le phénomène saisonnier ne se manifeste que pour l'autre ou l'autre des saisons. Les coefficients attachés aux variables binaires (les a_2 , a_3 , a_4) mesurent l'impact de l'effet saisonnier sur la relation entre la variable dépendante et les autres variables explicatives par rapport à la situation de référence (le 4^o trimestre dans le cas présent). On notera enfin que l'on peut prendre en compte une interaction entre la tendance et l'effet saisonnier (effet de croissance de ce dernier dans le temps) en introduisant dans le modèle les produits des variables binaires par la variable temps.

Régression par morceaux

Les variables binaires permettent également de modéliser les effets de seuil dans les relations linéaires. Le modèle utilisé est cette fois :

$$Y_i = a_1 + b_1 X_i + b_2 (X_i - X^*) D_i + u_i$$

où X_i est la variable explicative, X^* est le seuil à partir duquel la relation entre la variable explicative et la variable dépendante se modifie, D_i prend pour valeur 0 si $X_i < X^*$ et 1 sinon. La forme de la relation modélisée est la suivante :

