Si on revient à la théorie des opinions, ce résultat n'a rien d'étonnant. Les positions relatives occupées par les individus et les groupes sociaux ne peuvent évoluer de façon significative dans une brève période. Pour qu'il y ait évolution, il faudrait que les rôles objectivement joués par les individus aient été redistribués, et que cette redistribution ait affecté la perception que l'on a d'autrui. Mais si l'individu se situe bien vis-à-vis des autres toujours en gros de la même façon, il exprime sa différence au moyen de conflits dont la signification peut changer rapidement. Cette signification émerge d'un débat entre acteurs parties prenantes qui est réglé par une logique d'intérêts : si le rôle de certains des acteurs se trouve modifié, les intérêts en jeu ne sont plus les mêmes, le débat se transforme, les significations évoluent. Or une telle modification est intervenue à l'évidence lors du changement de majorité présidentielle de mai 1981 : on ne disait plus la même chose après, quand on remettait en cause les institutions relevant de l'Etat ou quand on exprimait un doute sur le bien fondé d'un projet qui ne pouvait se réaliser sans sa volonté. Le constat sur la stabilité des positions sociales sur le plan principal n'est donc pas une découverte : il conforte simplement la démarche qui a été empruntée. En effectuant une rotation dans le Ciel qui va modifier en conséquence la disposition des points moyens sur la Terre, on cherche à répondre à la question : ce que l'on veut dire à travers les conflits étant à peu près constant, comment le dit-on et pourquoi le dit-on ainsi à chaque époque ?

9

TRAITEMENTS D'ENQUETES PAR MODELES

Bernard Burtschy

Ecole Nationale Supérieure des Télécommunications
Paris

9.1 Introduction

L'analyse par tableaux croisés est une méthode traditionnelle et encore très largement utilisée pour traiter les enquêtes. Même si elle apparaît à certains quelque peu désuète, est-elle obsolète pour autant ? Oui certainement si on considère les grandes enquêtes sociologiques où de nombreuses variables interagissent entre elles et où on connaît, en fait, relativement peu de choses, soit parce que le domaine est relativement complexe, soit tout simplement parce qu'il n'a jamais été vraiment étudié. Les méthodes de l'analyse des données, dans leur acception francophone (analyses descriptives multidimensionnelles) avec les méthodes factorielles et/ou encore les techniques de classification, sont utiles dans ce genre de situations : ce serait une erreur de ne pas les utiliser.

On aurait cependant tort de réduire toutes les enquêtes à ce type : il en existe bien d'autres variétés. Dans bien des domaines, on n'en est plus à chercher à comprendre le phénomène ; mais il est plutôt question de vérifier des hypothèses de travail, de valider et de confirmer des comportements, voire d'agir. Ces enquêtes ont souvent un questionnaire réduit avec, au plus, quelques dizaines de questions fortement structurées. L'enquête sert essentiellement à quantifier des comportements et à valider des hypothèses issues d'un modèle.

Dans cette philosophie, l'une des étapes ultimes est de concevoir le plan de sondage comme un véritable plan d'expérience. Par exemple, dans une enquête d'épidémiologie médicale chez les enfants, on sait depuis longtemps que certains paramètres sanguins varient fortement par sexe et par âge : on peut certes retrouver une nouvelle fois ce fait déjà largement connu et documenté.

Mais afin d'aller plus loin, il est préférable d'effectuer l'enquête par strate (cf. chapitre 2), en utilisant comme variables de stratification le sexe et

l'âge. Le traitement de l'enquête, très dépendant de ces deux paramètres, ne pourra s'effectuer sans introduire, a priori, un modèle qui permettra de rendre les autres variables indépendantes du sexe et de l'âge : sans cette précaution, toute méthode d'analyse des données fournirait des résultats décevants. L'introduction d'un modèle est donc souvent indispensable pour pouvoir s'affranchir de l'influence de certaines variables dominantes.

Cette utilisation de modèle est d'autant plus indispensable et d'ailleurs naturelle que de nombreuses enquêtes sont conçues, avant tout, pour expliquer un phénomène précis réduit parfois à une seule question et qu'on cherche à quantifier très précisément les variations de ce phénomène. Cette absolue nécessité de quantification est souvent aussi à l'origine de l'utilisation de ces méthodes.

9.2 Les différentes méthodes d'analyse par modèles

Ces méthodes interviennent dès que l'on dispose d'une variable à expliquer et d'une ou plusieurs variables explicatives. Les techniques diffèrent sensiblement selon la nature statistique des variables mises en jeu.

Il est courant de distinguer schématiquement des variables quantitatives et des variables qualitatives. Avec les premières, il est d'usage de procéder par calcul de moyennes et de variances ; avec les secondes, on préfère utiliser des techniques basées sur des comptages. Cette distinction, classique dans l'ensemble de la statistique, a des conséquences importantes.

Du fait de cette différence de traitement, les techniques statistiques utilisées sont radicalement différentes ce qui implique des méthodes différentes. On peut, à première vue, organiser ces méthodes qui distinguent variables à expliquer et variables explicatives, selon le tableau 1 :

Tableau 1

Configuration des variables à expliquer et des variables explicatives

		Variable à expliquer	
		Quantitative	Qualitative
Variable explicative	Quantitative	QT x QT	QT x QL
	Qualitative	QL x QT	QL x QL

Dès que l'on dispose d'une variable à expliquer quantitative, on se situe dans le cadre théorique du modèle linéaire. Le cas le plus simple est représenté par une variable à expliquer quantitative et une variable explicative quantitative elle aussi (cas QT x QT). Il est aisé (et utile) de faire un graphique à l'aide de ces deux variables ; l'importance de l'éventuelle liaison linéaire est donnée par l'indice de corrélation linéaire. Il est enfin possible d'en déduire un "modèle" linéaire, s'il s'applique, sous la forme suivante :

$$Y = a X + b$$

Il s'agit de la classique (et inusable) régression linéaire simple. Ce modèle peut se compliquer par l'adjonction d'autres variables explicatives $X_1, X_2, ..., X_p$ sous la forme de la régression multiple à p variables :

$$Y = a_1 X_1 + a_2 X_2 + ... + a_p X_p.$$

Une énorme littérature a été consacrée à ce sujet sous le terme de régression ou de modèle linéaire. Le succès de la méthode s'explique par le fait qu'elle répond au besoin inextinguible de quantification. Soulignons toutefois que si la méthode a été extensivement adaptée au traitement d'enquête sous le nom d'analyse des dépendances (nous la décrirons dans la section 4) son utilisation est loin d'être aisée sur le plan technique, contrairement à une fausse apparence (Mosteller et Tukey, 1977).

Rappelons que cette méthode s'étend aux variables explicatives qualitatives (QT x QL) sous le nom d'analyse de variance ainsi qu'au mélange de variables explicatives quantitatives et qualitatives sous le nom d'analyse de covariance.

Enfin l'explication d'une variable peut aussi se représenter sous la forme d'un arbre hiérarchique sous le nom de segmentation. La segmentation avec une variable à expliquer quantitative est due à Sonquist et Morgan (1964) et est connue sous le nom d'A.I.D. (Automatic Interation Detection). Les méthodes de segmentation se présentent sous la forme externe d'un arbre et non sous la forme d'un modèle analytique : elles ne sont citées ici que pour mémoire. Malgré la grande popularité de ces méthodes due à des raisons historiques et à une grande facilité d'interprétation, il faut tout de même souligner que la plupart des données d'enquêtes sont rarement quantitatives mais plutôt qualitatives.

Le cas le plus courant est représenté par l'explication d'une variable dichotomique (0/1) : on veut, par exemple, expliquer la présence ou l'absence d'une maladie. L'application directe de la régression linéaire fournirait des valeurs comprises entre 0 et 1 qui peuvent s'apparenter à la probabilité d'avoir la maladie, mais aussi des valeurs négatives ou

supérieures à 1 qu'il est difficile d'interpréter. La régression logistique a pour objet de pallier ce type d'inconvénient.

La régression logistique peut s'étendre, au prix d'une généralisation délicate, à l'explication de variables qualitatives à plus de deux modalités en utilisant les ressources du modèle multinomial. Les variables explicatives peuvent être soit qualitatives (QL x QL), soit quantitatives (QT x QL). Il est donc tout à fait possible de quantifier, par cette méthode, l'impact de diverses variables sur une variable qualitative : la théorie générale, esquissée dans la section 5, est connue sous le nom de modèle log-linéaire.

L'explication d'une variable qualitative peut également donner lieu à une structure d'arbre en utilisant non pas les propriétés du modèle logistique mais celles du χ^2 (Bouroche et Tenenhaus, 1970). Ces méthodes progressent beaucoup actuellement depuis la sortie de l'ouvrage de Breiman et al. (1984). Elles sont cependant hors du champ de ce chapitre restreint aux modèles au sens classique du terme (donnant lieu à une formulation analytique). Ces méthodes sont directement en compétition avec l'analyse discriminante (Celeux, 1990 ; Tomassone et al., 1988).

Les diverses méthodes "explicatives" sont résumées dans le tableau 2, qu'elles produisent un modèle explicite ou non.

Dans la suite de ce chapitre, nous ne nous intéresserons qu'à celles qui produisent un modèle explicite, c'est-à-dire celles fondées sur le modèle linéaire (explication d'une variable quantitative) et celles fondées sur le modèle log-linéaire (explication d'une variable qualitative).

Il faut souligner pour l'utilisateur que ces méthodes brillent certes par leur sophistication théorique mais qu'elles font aussi preuve régulièrement d'inconstance par leur manque de robustesse et qu'il vaut souvent mieux démarrer par quelques simples tableaux croisés avant de se lancer dans un ambitieux modèle.

9.3 Tables de contingence

Depuis les travaux historiques de Durkheim (1897) et de Lazarsfeld (1948), la table de contingence est devenue un des principaux outils du traitement d'enquête. Pourquoi une telle popularité ?

Tout simplement parce que la table de contingence, le fameux tableau croisé, permet immédiatement d'illustrer, sous la forme d'une matrice, la réponse à une question simple.

Prenons par exemple la question : "le fait qu'une cabine téléphonique soit sale, transforme-t-il l'utilisateur de la cabine en client mécontent ?"

Tableau 2 : Quelques méthodes d'analyse par modèle.

	Variable à expliquer quantitative	Variable à expliquer qualitative : 2 modalités	Variable à expliquer qualitative : > 2 modalités ordonnées	Variable à expliquer qualitative : > 2 modalités quelconques
Variables explicatives quantitatives	*Régression* *Analyse de dépendance* *Segmentation AID*	*Rég. logistique* *Discrimination* *Segmentation AID*	*Log-linéaire* * *Segmentation* χ^2 *	*Multinomial* *Discrimination* *Segmentation* χ^2
Variables explicatives qualitatives	*Analyse de variance* *Analyse de dépendance* *Segmentation AID*	*Rég. logistique* *Segmentation AID*	*Log-linéaire* * *Segmentation* χ^2 *	*Log-linéaire* *Multinomial* *Segmentation* χ^2 *Discrimination*
Variables explicatives mixtes	*Analyse de covariance* *Analyse de dépendance* *Segmentation AID*			

* : Extensions des méthodes

La réponse est dans le tableau suivant :

		Client mécontent	
		oui	non
Cabine sale	oui	82	1 629
	non	56	2 158

Les utilisateurs de tableaux croisés calculent divers pourcentages qu'on imagine assez facilement : pourcentages de cabines sales, pourcentages de personnes mécontentes, etc...

L'examen "à l'œil" du tableau ne permet cependant pas de répondre quantitativement à la question fondamentale que l'on s'était posée, à savoir : est-ce que le fait qu'une cabine soit sale est lié au nombre de mécontents ? Pour répondre à cette question, on peut utiliser deux critères : le risque relatif et l'odds ratio.

9.3.1 Le risque relatif

Notons la table de contingence précédente sous la forme plus générale suivante :

		Variable à expliquer		
		oui	non	
Facteur d'influence	oui	a	b	a+b
	non	c	d	c+d
		a+c	b+d	t

Le risque d'être mécontent (variable à expliquer) chez ceux qui sont dans une cabine sale (oui au facteur d'influence ou encore appelé facteur de risque) est de a / (a+b) alors que chez ceux qui n'ont pas le facteur de risque (ils ne sont pas dans une cabine sale), il est de c / (c+d). Le risque relatif est le rapport entre les deux :

$$RR = \frac{a}{(a+b)} / \frac{c}{(c+d)} = \frac{a\,(c+d)}{c\,(a+b)}$$

Dans le cas précédent, la proportion de mécontents dans les cabines téléphoniques sales est de :

$$\frac{82}{82+1629} = 4{,}79\ \%,$$

alors qu'elle est de :

$$\frac{56}{56+2158} = 2{,}52\ \%$$

dans les cabines propres.

Le risque relatif est de 1,90. On remarque que la comparaison des deux chiffres 4,79 et 2,52 est au moins aussi instructive que leur rapport (1,90).

Dans ce cas précis, l'objectif de toute organisation est, on peut le supposer, d'avoir le moins de mécontents possible ou plutôt, dans une perspective d'action, de le réduire de plus en plus. Le terme a+b peut alors être approximé par b ; il en est de même pour c+d avec d.

$$\frac{a\,/\,(a+b)}{c\,/\,(c+d)} \approx \frac{a\,/\,b}{c\,/\,d} = \frac{a\,d}{b\,c}$$

Prenons un autre exemple légèrement différent issu du domaine biomédical qui a une longue pratique de ces ratios et où ils sont abondamment décrits (Schlesselman, 1982). On cherche à savoir si le fait de fumer est un facteur de risque de cancer. Pour tenter de répondre à cette question, on relèvera la proportion de fumeurs dans une population atteinte de la maladie que l'on comparera à la proportion issue d'une population témoin.

On a le tableau suivant :

		Cancer	*Témoin*	
Fumeur	oui	a	b	m_1
	non	c	d	m_2
		n_1	n_2	n

Ce tableau a généralement été constitué à partir de deux échantillons différents : la population des personnes atteintes de cancer et une population témoin. Les effectifs totaux $n_1 = (a+c)$ et $n_2 = (b+d)$ ne dépendent que des tailles des échantillons respectifs. Souvent ils sont choisis à peu près égaux en taille.

Les rapports a / b et a / (a+b) sont par conséquent peu interprétables. Il en est de même des rapports c / d et c / (c+d). Le risque relatif n'est pas, dans ce cas, vraiment pertinent.

9.3.2 L'odds ratio

L'odds ratio (OR) est donné par le rapport :

$$OR = \frac{a\,d}{b\,c}$$

On a remarqué précédemment qu'il était, sous certaines conditions, une approximation du risque relatif. La différence entre les deux notions de risque relatif et d'odds ratio est familière aux parieurs. Ainsi un odds ratio

:st équivalent à un pari de 2 contre 1 alors que le risque relatif est :quivalent à une probabilité de 2 sur 3.

)n remarquera que l'odds ratio n'a pas le défaut du risque relatif oncernant les échantillons séparés car il peut se calculer indifféremment n ligne et en colonne.

$$\frac{a / c}{b / d} = \frac{a}{c}\left(\frac{d}{b}\right) = \frac{a}{b}\left(\frac{d}{c}\right) = \frac{a / b}{c / d}$$

nfin l'interversion de lignes ou de colonnes conduit tout au plus à calculer)n inverse.

3.3 Interprétation des risques relatifs et odds ratio

es deux indicateurs sont des mesures d'association entre facteurs de risque)u d'influence) et facteurs à expliquer. Ils varient de 0 à l'infini avec, nsiblement, la même interprétation. Une valeur supérieure à 1 indique un sque positif, une valeur inférieure à 1 un risque négatif et une valeur gale à 0 l'indépendance entre risque et variable à expliquer.

: calcul de la variance de l'OR s'effectue par l'approximation de Wolf :

$$\text{Var}(\ln \hat{OR}) = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$$

ette formule est à la base du calcul des tailles d'échantillons, des uissances de tests (Woodward, 1992). On calcule un intervalle de nfiance approché avec t = 1,96 pour p = 0,95.

: χ^2 est une mesure classique de l'association de deux variables dans une)le de contingence. Appliqué aux tables 2×2, il devient

$$\chi^2 = \frac{(ad - bc)^2 n}{n_1 n_2 m_1 m_2}$$

ı correction de continuité de Yates qui permet une meilleure proximation du test exact de Fisher est donnée par :

$$\chi^2_c = \left(\frac{(|ad - bc| - 1/2n)^2 n}{n_1 n_2 m_1 m_2}\right)$$

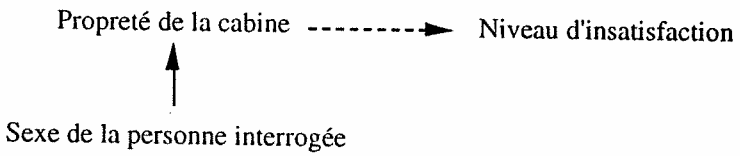
ıpplication ou la non application de cette continuité a fait l'objet d'une ondante littérature. Comparé à l'odds ratio, on remarquera que le χ^2 est

beaucoup moins intuitif. Il a bien sûr l'avantage de pouvoir se généraliser à des tableaux supérieurs au cas 2×2.

Les praticiens cependant, en raison de l'extrême simplicité de l'odds ratio, préfèrent souvent se ramener à ces fameuses tables 2×2. La comparaison entre elles de tables plus générales n'est d'ailleurs pas si simple en statistique rigoureuse.

9.3.4 Variables de contrôle

Dès que l'on analyse les relations entre deux variables, on se pose naturellement la question de savoir si la relation n'est pas sous l'influence d'une troisième variable. Supposons que l'on veuille vérifier si le sexe de la personne interrogée a une influence sur le taux d'insatisfaction des utilisateurs de cabine téléphonique. On sait que la propreté de la cabine a une influence directe sur le taux d'insatisfaction. Par ailleurs, les femmes sont beaucoup plus sensibles que les hommes à la propreté de la cabine. Du seul fait de ces deux relations, on peut trouver une relation apparente entre le sexe de la personne interrogée et le niveau d'insatisfaction.



Pour s'affranchir de cet effet pervers, il faut tester la relation sexe × niveau d'insatisfaction dans les cabines propres d'une part, et dans celles qui ne le sont pas, d'autre part. Dans le cas des tables 2×2, Mantel et Haenszell (1959) proposent une méthode très simple. Supposons que l'on ait stratifié la population selon une ou plusieurs variables en k groupes. Dans le groupe i, on aura la table suivante :

		Cas contrôle		
Risque	oui	a_i	b_i	m_{1i}
	non	c_i	d_i	m_{2i}
		n_{1i}	n_{2i}	n_i

L'odds ratio, corrigé de l'effet de la stratification, sera le suivant :

$$OR_{mh} = \frac{\sum_{i=1}^{k} (a_i d_i / n_i)}{\sum_{i=1}^{k} (b_i c_i / n_i)}$$

On peut appliquer le même raisonnement dans le cas du χ^2 en définissant les quantités suivantes :

$$E(a_i) = \frac{n_{1i}\, m_{1i}}{n_i}$$

$$V(a_i) = \frac{n_{1i}\, n_{2i}\, m_{1i}\, m_{2i}}{n_i^2\,(n_i - 1)}$$

$$\chi^2_{mh} = \frac{1}{\sum V(a_i)} \left[\left| \sum a_i - \sum E(a_i) \right| - \frac{1}{2} \right]^2$$

Ici encore, l'avantage de ces méthodes est de permettre des ajustements d'effets de manière intuitive qui tout en étant rigoureux, n'utilisent pas un impressionnant arsenal statistique ; accessoirement, ces méthodes utilisent des techniques de statistique exacte (Mantel, 1987).

Leur seul inconvénient est de se fonder sur des tables 2x2 ou, plus généralement, sur des tables 2xq, (on a deux groupes à comparer). On remarquera qu'il est possible d'avoir la même approche dans le cadre du χ^2 (Mantel, 1963). D'ailleurs de nombreuses méthodes de segmentation utilisent aussi cette approche.

Sur un plan purement théorique, cette restriction est fâcheuse, et l'on aurait envie de généraliser l'approche à des tables pxq (Agresti, 1992). On y reviendra plus tard, mais soulignons que, pour le praticien, cette manière de procéder n'est nullement gênante, bien au contraire, car elle est très intuitive ; d'autant que la comparaison simultanée de plus de deux groupes n'est pas si simple que cela.

9.4 L'analyse des structures causales

Un des enseignements de la section précédente est de montrer que souvent, on ne cherche pas tant à mesurer l'association entre deux variables qu'à expliquer une variable par une autre, avec d'un côté une "cause" possible (la cabine n'est pas propre) et de l'autre côté un effet (le client n'est pas satisfait).

De par la nature et les caractéristiques des outils statistiques mis en œuvre, on distingue très nettement le cas où la variable à expliquer est quantitative, du cas où elle est qualitative (cf. section 2).

Lorsque la variable à expliquer est quantitative, on se servira peu ou prou du modèle linéaire et plus particulièrement de la régression, de l'analyse de la variance et de la covariance. Cette méthode, poussée dans ses retranchements conduit à l'analyse des dépendances (*path analysis*).

Lorsque la variable à expliquer est qualitative, l'analyse des structures "causales" s'effectue à l'aide de tableaux de contingences. Un cas intermédiaire, important en pratique, est celui où la variable à expliquer a deux modalités : la régression logistique est alors l'outil adéquat.

L'étape ultime de la recherche des relations entre variables est souvent désignée en sciences humaines par l'expression "analyse de la causalité" (terminologie trompeuse, car la causalité n'est qu'une présomption). Elle s'effectue, dans le cas d'une variable à expliquer à plus de deux modalités, selon le modèle log-linéaire (Agresti, 1990) avec les mêmes principes que l'analyse des dépendances.

Comme les variables qualitatives sont les plus nombreuses dans les enquêtes, on s'attachera surtout à expliciter ici cette dernière démarche d'autant que le modèle linéaire a été abondamment décrit dans de nombreux manuels (mais son application aux données d'enquêtes est rarement décrite).
Indépendamment de la manière dont sont mesurées les relations, toutes ces méthodes cherchent à déterminer la structure causale des variables observées.

Lorsque ces variables sont nombreuses, l'analyse est souvent très difficile et il vaut mieux utiliser l'analyse de données. Il est important de souligner que ces méthodes sont surtout adaptées au cas où l'on a relativement peu de variables (de l'ordre de la dizaine tout au plus), ce qui est en accord avec la plupart des enquêtes dont l'objectif est la quantification.

9.4.1 Les structures causales simples

Partons du cas simple où on a une variable à expliquer Y et des variables explicatives $X_1, X_2, \ldots, X_p$.

Une structure causale est dite simple (Boudon, 1966) lorsque aucune variable explicative ne dépend causalement d'aucune autre variable explicative (figure 1).

Dans ce cadre, on distingue les structures simples sans effets d'interaction des structures simples avec effets d'interaction (figure 2). On dit qu'il y a effet d'interaction entre deux variables lorsque l'action de chacune des deux variables sur une variable dépendante, dépend elle-même de l'autre variable.

Il doit être clair dans cette section que la causalité n'est qu'une "présomption a priori de lien causal", lien causal qu'aucune méthode statistique ne pourra d'ailleurs établir avec certitude.

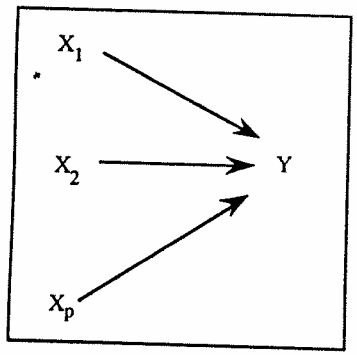


Figure 1

Structure causale simple sans interaction

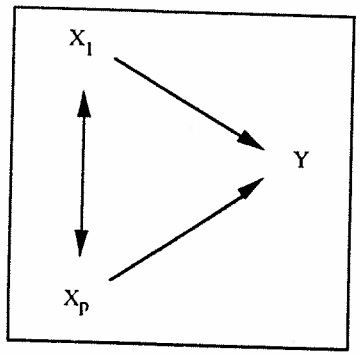


Figure 2

Structure causale simple avec interaction

)n trouve souvent les structures simples dans les cas de situations l'expérimentation dans lesquelles il est possible de répartir aléatoirement es effets de variables qui ne sont pas causalement dépendantes : c'est ouvent le cas dans les recueils de mesures biomédicaux (et plus articulièrement épidémiologiques), pharmaceutiques, agronomiques, etc.

:appelons-en rapidement le cadre.

)n pourrait ainsi dire que la satisfaction Y d'un utilisateur d'une cabine éléphonique dépend des quatre variables suivantes :

X_1 : la propreté de la cabine

X_2 : la qualité de la transmission

X_3 : la longueur de la file d'attente

X_4 : des conditions météorologiques

)n peut considérer, à première vue, que ces variables sont indépendantes s unes des autres et écrire un modèle sous la forme :

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4 + \varepsilon$$

ı ε représenterait les effets qui ne sont pas dans ce modèle.

4.2 Les structures causales complexes

ans une structure complexe, il existe au moins une variable explicative ıi dépend causalement (simple *présomption a priori* de causalité) d'une tre variable explicative.

Par exemple, le fait d'être dans une cabine téléphonique sans porte (fréquent dans les gares) induit une qualité d'écoute moindre et donc une insatisfaction plus grande.

Cabine téléphonique sans porte ---> qualité d'écoute ---> Insatisfaction

On peut quelquefois substituer une structure simple à une structure complexe. Comme dans un plan d'expérience, on égalisera la distribution des facteurs en annexes dans la population des cabines avec porte et sans porte en éliminant aléatoirement un certain nombre d'individus. On revient ainsi à une structure simple.

Sur cet exemple, on voit que la transformation par un plan d'observation approprié permet, dans certains cas, de transformer une structure complexe en une structure simple. Ce n'est cependant pas toujours possible, surtout lorsque l'on ne cherche pas à vérifier un effet, mais à l'interpréter. C'est en particulier le cas lorsque l'on a affaire à une structure complexe élémentaire (figure 3) :

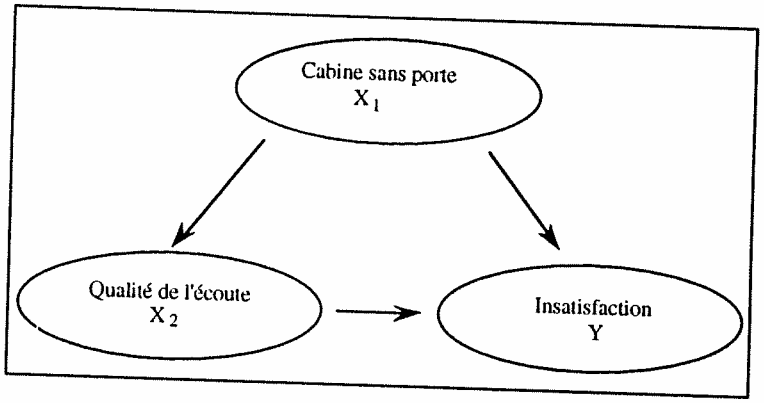


Figure 3

Exemple de structure causale complexe

L'écriture d'un modèle comme dans le cas précédent conduirait aux deux équations suivantes :

$$X_2 = a_1 X_1 + e_1$$

$$Y = a_1 X_1 + a_2 X_2 + e_2$$

Dans le cas d'une structure causale simple, les coefficients avaient une interprétation physique (encore faut-il être très prudent sur la linéarité, la

robustesse et la "clôture" du système). Dans le cas d'une structure causale complexe, l'interprétation des coefficients est beaucoup moins évidente.

9.5 Modèle linéaire

Dans le cas d'une structure causale simple, on dispose d'une variable à expliquer Y et d'une ou plusieurs variables explicatives. Le cas le plus simple à considérer est celui où toutes les variables sont quantitatives, c'est-à-dire lorsque l'on peut très prosaïquement calculer une moyenne.

Le cas est très fréquent dans le domaine biomédical ; il l'est beaucoup moins dans les enquêtes sociologiques. Dans ce dernier cas, si les utilisateurs veulent à tout prix utiliser le modèle linéaire, ils transforment, dans le questionnaire, toutes les questions en échelle d'attitude (tout à fait d'accord, assez d'accord...) ou sous forme de note. Même si cette manière de faire est quelquefois contestable et a donné lieu à une abondante littérature, elle est couramment pratiquée. Elle permet de fournir des résultats, certes peut-être pas aussi complets que par d'autres méthodes, mais rapidement et sans trop grands investissements.

9.5.1 Régression et théorème de Gauss-Markov

L'outil de base du modèle linéaire est la bonne vieille régression abondamment décrite dans de nombreux manuels (voir Saporta, 1990 pour un exposé exemplaire et compact ou Tomassone et al, 1983, pour un exposé très complet). Rappelons en les grandes lignes. Dans la régression simple, il s'agit d'estimer les coefficients a et b tels que :

$$Y = aX + b$$

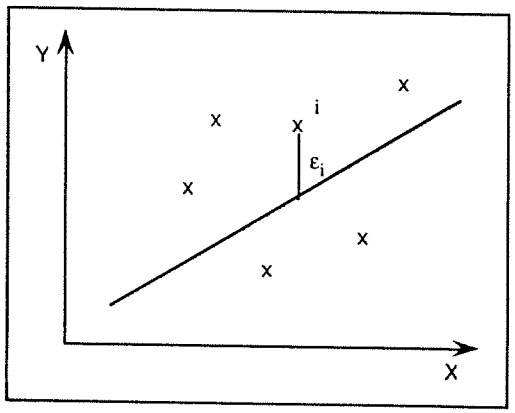


Figure 4

Ajustement selon les moindres carrés

Il est classique de le faire selon la méthode des "moindres carrés" c'est-à-dire minimiser le carré de la distance ε_i de chaque point i à la droite recherchée (figure 4).

Malgré une théorie mathématique achevée (le théorème de Gauss-Markov), il ne faut pas en ignorer les sérieuses limites qui restreignent notablement son utilisation pratique et conduisent à des utilisations parfois discutables. Ces limites prennent un relief particulier dans les traitements statistiques d'enquêtes.

a) La linéarité

D'emblée, on impose la linéarité de la relation alors que, souvent, on ne sait rien de la nature de la relation. Il s'agit donc d'une hypothèse de travail qu'il faut tester sérieusement à partir des résidus. Ainsi par exemple, pense-t-on sérieusement que la consommation de chewing-gum est une fonction linéaire de l'âge (figure 5) ? Comme ce n'est probablement pas le cas, il faut soit changer de modèle, soit transformer les variables.

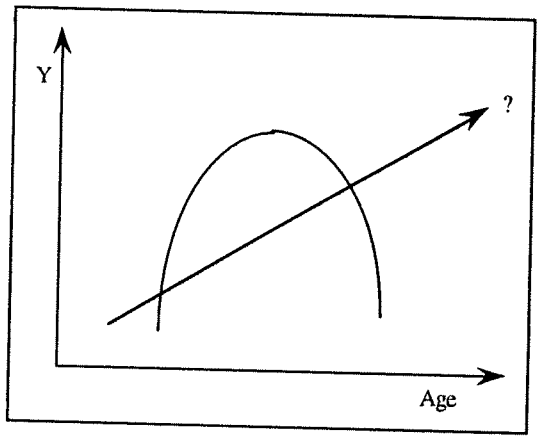


Figure 5

Exemple d'erreur de spécification

La régression peut malheureusement fournir des résultats "valides" selon les critères couramment employés (validité des coefficients F de Fisher) et être de peu d'intérêt et même fournir des résultats trompeurs.

Un bon graphique (Mosteller et Tukey, 1977) permet de remédier simplement à la plupart des erreurs rencontrées. On peut aussi transformer les variables (Aitkinson, 1985).

b) La sensibilité aux points aberrants

L'estimation utilisée est, en général, l'estimation des moindres carrés. Ceci a pour conséquence qu'un individu atypique prend une importance d'autant plus énorme que l'on met son résidu au carré (figure 6). Combien de régressions ne tiennent que sur quelques individus ! Une méthode efficace de recherche des observations dites "influentes" est indispensable surtout dans les enquêtes (Besley and al., 1980).

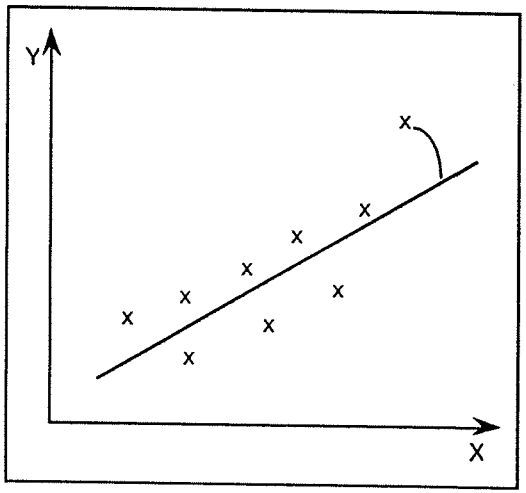


Figure 6

Importance des points irréguliers

Ces méthodes de détection des observations influentes utilisent toutes des indicateurs de mesure d'influence ; ceux-ci peuvent s'exprimer comme une fonction du résidu e_i de l'individu i et son effet h_i qui est le ième élément diagonal de la matrice :

$$H = X\ (X'\ X)^{-1}\ X'$$

X étant la matrice dont les colonnes sont constituées par les observations des variables explicatives (exogènes) du modèle.

Le principe de ces méthodes est de rechercher le ou les individus qui contribuent le plus à la formation de la droite, c'est-à-dire celui ou ceux qui ont un indicateur d'influence élevé.

On peut aussi utiliser des estimateurs robustes qu'il faut adapter au cas particulier de la régression. Il faut cependant souligner que, en pratique, les résultats diffèrent sensiblement selon la fonction robuste choisie (Chaffai, 1989).

c) Les erreurs ont-elles même variance ?

Une des hypothèses habituelles lors de l'utilisation du modèle linéaire général est l'homoscédasticité des erreurs, c'est-à-dire l'hypothèse que la variance des erreurs est constante.

Ce n'est pas le cas, par exemple, lorsqu'on analyse une consommation d'un produit avec le revenu, la variance croissant proportionnellement avec ce dernier (cf. figure 7).

Une transformation logarithmique ou, plus généralement une transformation de Box-Cox (1964, 1982) peut y remettre bon ordre. Soulignons quand même que la classe des transformations de Box et Cox est loin d'être exhaustive et qu'elle n'est pas toujours d'utilisation facile. Elle permet toutefois de résoudre la plupart des cas rencontrés en pratique.

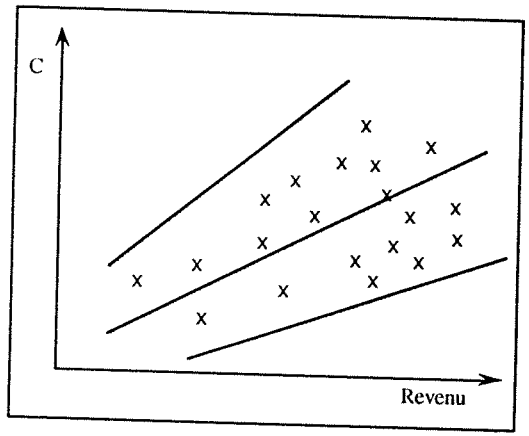


Figure 7

Hétéroscédasticité

d) Les variables explicatives sont-elles connues sans erreurs ?

Cette hypothèse, un des fondements du théorème de Gauss-Markov, est totalement irréaliste dans le cas des enquêtes. Même des variables telles que l'âge qui, en principe, ne devraient poser aucun problème, sont connues avec des erreurs importantes qu'il est difficile de considérer comme aléatoires. Le seraient-elles, elles contreviendraient tout de même aux hypothèses du théorème.

On sait que dans ce cas les estimateurs des moindres carrés sont inconsistants. Les moyens de contourner cette difficulté ne sont, malheureusement, pas simples.

9.5.2 Régression multiple

Lorsque l'on dispose de plusieurs variables explicatives ou exogènes X_1, X_1... X_p, quantitatives, on peut généraliser la régression simple sous la forme :

$$Y = \beta_1 X_1 + ... + \beta_p X_p + e$$

Si l'on dispose des relevés de p variables sur n individus, on a la disposition suivante :

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & ... & X_{1p} \\ X_{21} & X_{22} & ... & X_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ X_{n1} & X_{n2} & ... & X_{np} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_p \end{bmatrix}$$

Si on note **Y** le vecteur de la variable à expliquer, la matrice des variables explicatives, β le vecteur des coefficients et **e** le vecteur des résidus, on peut noter l'équation sous forme vectorielle :

$$\mathbf{Y = X \beta + e}$$

L'estimation des coefficients par la théorie des moindres carrés est donnée par :

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

Aux dangers précédemment soulignés, on en aperçoit immédiatement un nouveau : la non inversion ou la difficulté d'inversion de la matrice $\mathbf{X}' \mathbf{X}$.

La non-inversion est rare : elle n'intervient que si une variable explicative peut s'exprimer comme une combinaison linéaire d'autres variables.

La difficulté d'inversion est plus courante. Elle intervient lorsque plusieurs variables sont corrélées entre elles (cf. Malinvaud, 1964 ; Besley and al., 1980). Elle conduit à de mauvaises estimations des coefficients. Celles-ci sont alors elles-mêmes autocorrélées, et ont des écarts-types importants.

On peut éviter ce problème en protégeant l'inversion par une procédure numérique, en sélectionnant les variables (Furnival, 1971) ou encore en travaillant sur une base orthogonale (régression sur composantes principales).

9.5.3 Analyse de variance et de covariance

L'analyse de variance est un cas particulier de la régression qui apparaît lorsque la variable explicative X est nominale alors que la variable à expliquer Y reste quantitative.

C'est par exemple le cas lorsque l'on veut expliquer la satisfaction d'un utilisateur du téléphone public selon le type de cabine (avec carte, à pièces, sans porte). On transforme classiquement, comme en analyse de données, la variable nominale en trois variables dichotomiques (présence/absence). Il faut quelque peu amender ce principe pour tenir compte du fait que la matrice **X** ainsi engendrée conduira à une matrice $\mathbf{X}'\mathbf{X}$ non inversible.

L'analyse de covariance s'utilise lorsque certaines variables explicatives à expliquer sont quantitatives alors que d'autres sont nominales.

Ces deux techniques forment la base des plans d'expériences qui sont beaucoup utilisés en agronomie et dans le domaine biomédical, mais aussi dans les études de marché et plus particulièrement les tests de produits.

9.5.4 L'analyse canonique

Dans l'analyse canonique, on dispose de plusieurs variables à expliquer alors qu'il n'y en avait qu'une seule dans la régression multiple. On peut la voir comme une extension de la régression multiple.

Notons Y le vecteur des variables à expliquer (Y_1, ... Y_p) et X le vecteur des variables explicatives (X_1, ..., X_q).

L'analyse canonique revient à chercher simultanément deux combinaisons linéaires : la combinaison linéaire des variables X et la combinaison linéaire des variables Y qui ont une corrélation maximale.

Ici, X et Y jouent des rôles symétriques. On parlera cependant de variables à expliquer et de variables explicatives.

Pour illustrer l'analyse canonique, prenons le même exemple que précédemment avec comme variables à expliquer Y l'ensemble des variables de satisfaction (Y_1 = qualité d'écoute , Y_2 = propreté de la cabine, etc.).

Les variables explicatives X sont constituées par des variables socio-démographiques ou des caractéristiques de cabine (X_1 = Age, X_2 = Nombre d'appels dans la semaine, etc.).

La meilleure corrélation est fournie par les combinaisons linéaires des variables suivantes :

$$F_1 = 0{,}27\ X_1 + 0{,}42\ X_2 + \ldots$$

$$G_1 = 0{,}32\ Y_1 + 0{,}61\ Y_2 + \ldots$$

Le coefficient de corrélation entre F_1 et G_1 qui est de 0,67 est nommé premier coefficient de corrélation canonique. Ce résultat peut être représenté en analyse des dépendances (cf. plus bas) par le diagramme suivant :

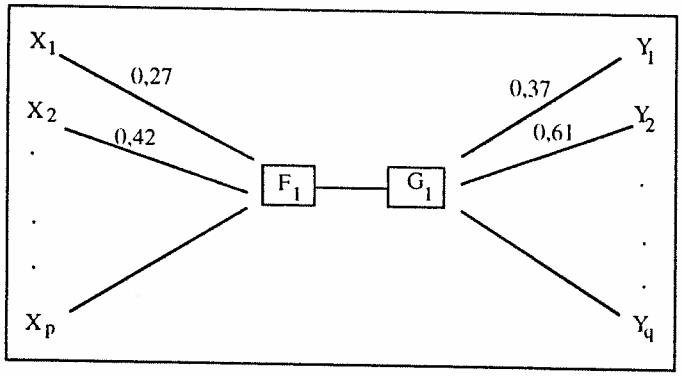


Figure 8

Structure "causale"d'une analyse canonique

L'analyse canonique, qui peut se voir comme une généralisation de la régression multiple, est aussi une méthode unificatrice de l'analyse de données (Caillez et Pagès, 1976) : elle contient en effet comme cas particulier l'analyse discriminante et l'analyse des correspondances.

Si de nombreux auteurs ont, à juste titre, souligné les difficultés d'interprétation de l'analyse canonique, ils ne l'ont que rarement mis dans son élément le plus favorable qui est celui de la régression multiple et de l'analyse de dépendance.

9.5.5 L'analyse discriminante

En analyse discriminante (cf. par exemple Celeux, 1991 ; Tomassone et al., 1987 ; McLachlan, 1992), on distingue a priori deux ou plusieurs groupes d'individus sur lesquels on mesure un ensemble de variables communes.

Pour reprendre une nouvelle fois l'exemple précédent avec ses variables explicatives X (Age, Nombre d'appels, etc.), on pourrait partager l'ensemble des individus en deux groupes avec d'un côté les globalement satisfaits et de l'autre les globalement non satisfaits.

L'analyse discriminante (ici à deux groupes) va chercher à mettre en évidence ce qui différencie ces deux groupes. On pourrait faire un pas supplémentaire en "prédisant" uniquement à partir de ces variables X, à quel groupe appartient un individu particulier.

Il est facile de se ramener à l'analyse précédente en créant une variable Y sous forme dichotomique dans le cas de deux groupes ou plusieurs variables (0/1) dans le cas de plus de deux groupes.

L'analyse discriminante est ainsi une application directe de l'analyse canonique (plus de deux groupes) et de la régression multiple (deux groupes).

L'application de l'analyse discriminante dans les enquêtes ne diffère donc que peu de ces deux dernières techniques avec d'ailleurs les mêmes contraintes d'application. En particulier, comme on s'appuie sur une combinaison *linéaire* des variables, les transformations des variables jouent un rôle crucial.

9.5.6 L'analyse des dépendances

La terminologie de l'analyse des dépendances a beaucoup fluctué au cours du temps. Initialement, il s'agissait d'une simple technique de décomposition de la corrélation (Boudon, 1966, 1970). Puis elle est finalement devenue la méthode, basée sur le modèle linéaire, de recherche des relations entre les variables (Duncan, 1975). Pour pouvoir utiliser efficacement cette méthode, il faut disposer d'un bon logiciel de régression multiple et restreindre sérieusement le nombre de variables à analyser. En quoi consiste-t-elle ?

Rappelons d'abord que les coefficients d'une régression multiple ont une interprétation très simple mais très importante : comme toute modification de la variable X se répercute sur Y proportionnellement au coefficient, le coefficient mesure l'importance de cet effet.

Lorsqu'il faut deux équations comme dans une structure causale simple avec interaction, l'effet total sur une variable est la somme d'un effet direct et d'un ou plusieurs effets indirects. L'avantage de cette décomposition est de pouvoir relativiser l'importance de chaque variable.

Notons $(X_1, \ldots, X_p)$ les variables endogènes ou expliquées et $(Z_1, \ldots, Z_q)$ les variables exogènes ou explicatives. Le modèle peut s'écrire sous la forme d'un ensemble d'équations :

$$\begin{aligned}
X_1 &= && a_{12} X_2 + \ldots + a_{1p} X_p + b_{11} Z_1 + \ldots b_{1q} Z_q \\
X_2 &= a_{21} X_1 && + \ldots + a_{2p} X_p + b_{21} Z_1 + \ldots b_{2q} Z_q \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
X_p &= a_{p1} X_1 + \ldots + a_{pp-1} X_{p-1} && + b_{p1} Z_1 + \ldots b_{pq} Z_q
\end{aligned}$$

Il est clair que ce modèle est purement théorique car son estimation, dans le cas le plus général, serait des plus problématiques en raison des nombreuses multicollinéarités. On décrit d'habitude ce modèle comme étant "complet non récursif" et il est défini par un ensemble de p équations structurelles.

S'il était possible d'estimer ce modèle, le coefficient a_{ij} représenterait l'effet direct de la variable endogène X_j sur la variable endogène X_i. Le coefficient b_{ij} représenterait l'effet de la variable exogène Z_j sur la variable endogène X_i.

Pour pouvoir développer un modèle réaliste au sens de son estimation, il faut imposer des hypothèses très fortes ; l'hypothèse classique est l'hypothèse de récursivité, qui permettra d'estimer une équation après l'autre avec une régression linéaire. Cette hypothèse implique que le modèle soit hiérarchique c'est-à-dire qu'il est possible de réordonner les variables X telles que $a_{ij} = 0$ quand $i < j$.

La première variable est ainsi uniquement déterminée par les variables exogènes ; la seconde variable est déterminée par la première et les exogènes. La troisième variable l'est par la première, la deuxième et les exogènes, etc.

Prenons un exemple simple avec deux variables endogènes X_1 et X_2 et une variable exogène Z_1.

On aura le système de deux équations :

$$X_1 = b_{11} Z_1 + e_1$$
$$X_2 = a_{21} X_1 + b_{21} Z_1 + e_2$$

On voit tout de suite l'intérêt de l'approche : on peut utiliser les moindres carrés ordinaires pour pouvoir estimer les coefficients en traitant une équation après l'autre. Mais on voit aussi son inconvénient : il faut pouvoir ordonner correctement les variables. Il y a aussi une interrogation plus fondamentale : le problème traité peut-il se formaliser en un problème récursif ?

Evacuons pour le moment ces deux questions qui souvent ne se posent pas dans la pratique car on peut hiérarchiser les variables surtout si elles sont peu nombreuses. D'autant que l'on dispose souvent d'une seule variable à expliquer.

La stratégie d'estimation est la suivante dans le cas, par exemple, de cinq variables :

a) Lancer la régression $X_1 = f(X_2, X_3, X_4, X_5)$. Les coefficients normalisés fournissent les dépendances recherchées du chemin E_1.

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5$$

α_2 est une estimation du coefficient a_{12}.

b) Lancer la régression $X_2 = f(X_3, X_4, X_5)$. Les coefficients fournissent les dépendances recherchées à partir du chemin E_2. Et ainsi de suite.

On peut alors quantifier le diagramme de la figure 9.

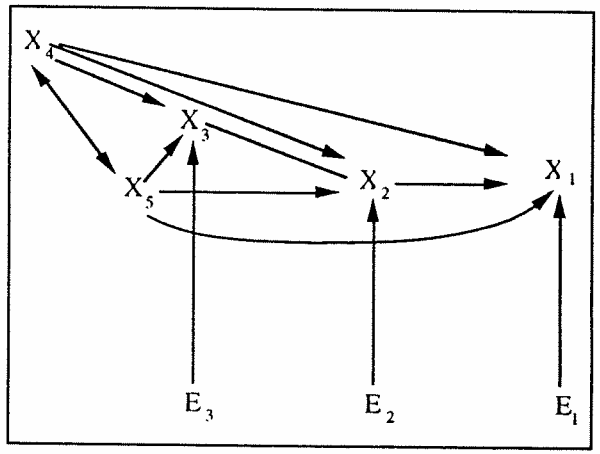


Figure 9

Relations de dépendance

Pour revenir à la question posée précédemment sur la hiérarchisation des variables, il est possible de s'en sortir sans hiérarchisation a priori, en utilisant extensivement les techniques de sélection de variables. A la limite ultime, on pourrait tester tous les cas dans le cas d'une équation (méthode de Furnival, 1971) en se définissant un critère de sélection. Si les variables ne sont pas trop nombreuses, cette approche est possible.

Cette méthode pourrait d'ailleurs très bien ne pas faire apparaître un modèle récursif. Le modèle n'est en particulier pas récursif si deux variables ont des effets réciproques de causalité. Ce problème est d'autant plus difficile que la corrélation, qui est la méthode statistique la plus usitée pour mesurer le degré de liaison entre deux variables, est symétrique et n'indique donc pas le sens d'une éventuelle causalité.

Si on est persuadé de l'existence d'au moins une relation bidirectionnelle, la technique des moindres carrés ordinaires (MCO) est tout à fait inadaptée : il faut utiliser des techniques plus élaborées telles que les doubles moindres carrés, les méthodes à information limitée ou complète, etc. Toutes ces méthodes sont discutées dans les manuels d'économétrie pour la résolution

des modèles à équations simultanées (cf. Malinvaud, op. cit.). On conseillera cependant au praticien de se ramener à un diagramme récursif, qu'il soit hiérarchique ou non. La technique est, dans ce cas, très au point (MacDonald et Doreian, 1977).

Cette méthode peut se généraliser en utilisant des combinaisons de variables à l'aide d'une analyse factorielle (modèle LISREL de Jöreskog, 1979).

9.6 Les modèles log-linéaires

Les enquêtes contiennent en général de nombreuses variables qualitatives. L'examen des relations entre ces variables s'effectue en construisant des tableaux croisés et, plus généralement, des tableaux de contingence multiples. Une manière simple de traiter ces tableaux de contingence est de les réduire sous la forme de tableaux 2x2 et d'utiliser la technique des odds ratio exposée précédemment (cf. § 9.2).

Cette réduction drastique à des tableaux 2x2 n'est pas toujours possible ni d'ailleurs souhaitable. Le modèle log-linéaire fournit un cadre théorique unifié de l'analyse des tables de contingence multiples. Initialement proposé par Birch (1963), ce modèle a été développé par Cox (1972) pour l'explication des variables dichotomiques et par Goodman (1971) pour la décomposition des tables de contingence. Une généralisation de ces modèles peut être trouvée dans Agresti (1988).

Proche du modèle log-linéaire, la régression logistique est une méthode qui permet de modéliser la relation entre une variable qualitative à deux modalités et des variables qui peuvent être quantitatives ou qualitatives.

9.6.1 La régression logistique

La variable Y est une variable qui prend deux modalités Y = 1 et Y = 0. Reprenons l'exemple précédent de la satisfaction des utilisateurs des cabines téléphoniques en expliquant cette satisfaction (Y = 1 pour les satisfaits et Y = 0 pour les mécontents) en fonction de la chute de pluie exprimée en millimètres. L'application directe de la régression linéaire aurait conduit à l'équation suivante :

$$\text{Satisfaction (0/1)} = 0{,}94 - 0{,}02 \times \text{Pluie}$$

La modélisation sous forme 0/1 fournit l'équivalent d'une probabilité. La chute d'un millimètre de pluie fait baisser la satisfaction de 2 %. Mais il y a un problème à l'une des extrémités : en effet, une chute de pluie importante

pourrait conduire à une satisfaction négative. Pour éviter ces petits ennuis à la borne 0, on peut utiliser le modèle suivant :

$$\text{Log } Y = a + b\,X$$

Il n'y a, fort heureusement pas de chute de pluie négative. Mais d'autres applications pourraient conduire à l'estimation d'une probabilité supérieure à 1. On utilisera en fait la formulation suivante :

$$\text{Log } \frac{Y}{1 - Y} = a + b\,X$$

L'intégration de cette expression fournit la fonction logistique :

$$Y = \frac{1}{1 + e^{-(a + b\,X)}}$$

La relation entre la variable Y et la variable X est directement liée à l'odds ratio (OR), défini dans la section 2. Si la variable X est codée en deux modalités (0/1), le modèle logistique s'écrit :

$$P(Y = 1) = \frac{1}{1 + e^{-(a + b)}} = P_1 \qquad \text{si } X = 1$$

$$P(Y = 1) = \frac{1}{1 + e^{-a}} = P_0 \qquad \text{si } X = 0$$

L'odds ratio s'écrit :

$$OR = \frac{P_1 / (1 - P_1)}{P_0 / (1 - P_0)} = e^{b}$$

Ce modèle se généralise aisément à plusieurs variables $X_1, X_2, \ldots, X_p$.

$$P(Y = 1 / X_1 \ldots X_p) = \frac{1}{1 + \exp - (a + \sum_{i=1}^{p} b_i X_i)}$$

Comme dans la régression multiple, le coefficient b_i de la variable X_i dépend de la présence des autres variables. L'intérêt, dans une enquête, de pouvoir utiliser plusieurs variables, est de pouvoir prendre en compte des facteurs de confusion.

Ainsi par exemple, si la satisfaction des clients d'une cabine téléphonique dépend de la structure de la population (sexe, age), il est possible

"d'ajuster" ces deux variables pour pouvoir les rendre constantes et ainsi mesurer l'impact des autres variables.

Cette procédure généralise l'approche de Mantel et Haenszell dans le cadre de tables 2x2 en lui apportant beaucoup plus de souplesse mais peut être aussi un peu plus d'opacité. Il faut cependant souligner que toute variable explicative nominale à k modalités, où k est supérieur à 2, doit être transformée en (k - 1) variables à 2 modalités. Les variables explicatives quantitatives peuvent être utilisées telles quelles.

L'estimation des paramètres du modèle logistique n'est pas une opération triviale. Elle s'effectue classiquement par la méthode du maximum de vraisemblance. Il existe bien entendu un ensemble de tests pour valider la régression logistique et les coefficients. On peut les consulter dans l'ouvrage de Hosmer et Lemeshow (1989).

9.6.2 *Le modèle polytomique*

La régression logistique, malgré son taux d'utilisation croissant dans les traitements d'enquêtes, possède une limitation sérieuse. Elle ne permet l'expliquer que deux modalités dichotomiques. Lorsque l'on souhaite expliquer une variable à trois modalités ou plus, il est toujours possible de décomposer cette variable en plusieurs variables à deux modalités et l'enchaîner les modèles logit associés. Il s'agit d'une approche pragmatique qui demande un savoir-faire non négligeable en particulier dans la spécification du modèle.

Une autre voie, plus prometteuse sur le plan théorique, est d'expliquer directement cette variable à plus de deux modalités. Si la variable explicative est elle aussi nominale, nous sommes en présence d'un tableau croisé (ou table de contingence).

Notons classiquement X_{ij} l'intersection de la ligne i et de la colonne j ainsi que $X_{i.}$ et $X_{.j}$ les deux marges du tableau. Dans l'hypothèse d'indépendance, on aura à l'intersection de la ligne i et de la colonne j, la valeur F_{ij} reconstituée à partir des marges.

$$F_{ij} = \frac{X_{i.}\, X_{j}}{n}$$

Notons p le nombre de lignes, q le nombre de colonnes et n la somme totale du tableau (= X..). En prenant le logarithme, on obtient :

$$\log F_{ij} = \log X_{i.} + \log X_{.j} - \log n$$

Cette équation donne l'idée de la forme d'un modèle log-linéaire dans le cas de l'hypothèse d'indépendance :

$$\log F_{ij} = m + m_1(i) + m_2(j)$$

Avec :

$$m = \frac{1}{pq} \sum_i \sum_j \log F_{ij}$$

$$m_1(i) = \frac{1}{q} \sum_j \log F_{ij} - m$$

$$m_2(j) = \frac{1}{p} \sum_i \log F_{ij} - m$$

Ce modèle est très proche de celui de la classique analyse de variance à deux variables : m représente la moyenne générale des logarithmes, $m_1(i)$ la différence moyenne générale et celle du niveau i de la première variable. $m_1(i)$ représente l'impact principal du logarithme de la fréquence de la modalité i de la première variable. Il en est de même pour $m_2(j)$ pour la seconde variable.

Comme dans l'analyse de variance, on a les contraintes suivantes sur les coefficients :

$$\sum_i m_1(i) = \sum_j m_2(j) = 0$$

Continuant l'analogie avec l'analyse de variance, on peut introduire un terme d'interaction :

$$\log F_{ij} = m + m_1(i) + m_2(j) + m_{12}(ij)$$

Il faut cependant souligner que, contrairement à l'analyse de variance, le modèle log-linéaire ne distingue pas, dans sa version classique, une variable dépendante et une variable indépendante : le concept de terme d'interaction ne s'y applique donc pas de la même manière.

Il y a bien entendu des contraintes sur les coefficients d'interaction :

$$\sum_i m_{12}(ij) = \sum_j m_{12}(ij) = 0$$

Ce modèle général est appelé modèle saturé. L'objectif est de trouver le modèle qui s'ajuste le mieux aux données observées. A cet effet, chaque modèle est accompagné de son degré de liberté (tableau 3, où V_1 et V_2 désignent les effets des lignes et des colonnes de la table de contingence).

Tableau 3

Emboîtement des modèles log-linéaires

P	Termes	Nombre de paramètres	Description
1	m	1	V_1 et V_2 nuls
2	$m + m_1 (i)$	$1 + (p - 1)$	V_2 nul
3	$m + m_2 (j)$	$1 + (q - 1)$	V_1 nul
4	$m + m_1 (i) + m_2 (j)$	$1 + (p - 1) + (q - 1)$	V_1 et V_2 indépendants
5	$m + m_1 (i) + m_2 (j) + m_{12} (ij)$	$1 + (p - 1) + (q - 1) + (p - 1)(q - 1) = pq$	V_1 et V_2 dépendants (modèle saturé)

.es tests d'adéquation se font en général par maximisation de la raisemblance de l'échantillon. Une méthode des moindres carrés pondérés, lus proche du modèle linéaire, a également été proposée (cf. Christensen, 990).

our un modèle 2x2, on peut écrire le modèle sous forme matricielle selon ı configuration suivante :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} \text{Log } F_{11} \\ \text{Log } F_{12} \\ \text{Log } F_{21} \\ \text{Log } F_{22} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} m \\ m_1 (1) \\ m_1 (2) \\ m_2 (1) \\ m_2 (2) \\ m_{12} (11) \\ m_{12} (12) \\ m_{12} (21) \\ m_{12} (22) \end{bmatrix}$$

n peut re-paramétrer ce modèle en tenant compte des contraintes sur les ıefficients :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \begin{pmatrix} \text{Log } F_{11} \\ \text{Log } F_{12} \\ \text{Log } F_{21} \\ \text{Log } F_{22} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

ec $b_1 = m$ $b_2 = m_1 (1) - m_1 (2)$
$b_3 = m_2(1) - m_2(2)$ $b_4 = m_{12} (11) - m_{12} (12)$

ıns le cas particulier des tables 2x2, on retrouve les résultats de la gression logistique. Le modèle se généralise lorsqu'il y a trois variables plus.

Avant l'utilisation récente des modèles log-linéaires, le traitement d'enquête consistait souvent en l'analyse d'une pléiade de tableaux croisés. Cette approche est dangereuse en raison de l'intervention, éventuellement cachée d'une troisième variable : cette situation est particulièrement surprenante dans le cas du paradoxe de Simpson (cf. par exemple Christensen, 1990), où l'utilisation des seules marges conduit à des conclusions contradictoires.

On peut, bien sûr, multiplier les tables de contingence multiples : avec cinq variables à quatre modalités, on a ainsi 1 024 cases (4^5), ce qui est aussi le cas avec 10 variables dichotomiques (2^{10}). Mais les individus enquêtés risquent fort de manquer pour remplir toutes ces cases.

Pour utiliser une telle approche, il faut donc limiter sérieusement le nombre de questions et le nombre de modalités : il faut utiliser une procédure de sélection efficace (Goodman, 1971). Le critère d'ajustement utilisé est, en règle générale, le χ^2 qui est un critère symétrique. Il est possible, en toute rigueur, de prendre un critère asymétrique (Agresti, 1992), mais il n'est jamais proposé dans les logiciels, reculant peut être devant la complexité des calculs nécessaires pour l'estimation des paramètres.

En pratique, pour utiliser efficacement les modèles log-linéaires, il ne faudrait pas dépasser dix variables (les procédures de sélection sont fragiles) et se ramener, dans la mesure du possible, à des variables dichotomiques. Il faudra ne pas accorder une attention excessive aux critères de significativité statistique en raison du nombre de paramètres à estimer (Koehler, 1986). Finalement, il faut toujours résumer les conclusions en termes de fréquences ou de proportions dans les cases. Si on est familier des odds ratio, ils sont d'un bon secours pour quantifier et présenter les résultats.

Enfin, il faut rappeler qu'il existe des passerelles théoriques importantes entre les modèles log-linéaires et d'une part l'analyse des correspondances (cf. par exemple Goodman, 1986), d'autre part l'analyse des dépendances, dans les cadres des *modèles graphiques* (cf. Whittaker, 1990 ; pour une présentation en français : Fine, 1992 ; pour une synthèse récente : Wermuth et Cox, 1992).

9.7 Intérêt et limites des méthodes basées sur le modèle linéaire

Après une période où le modèle linéaire a été sans doute trop appliqué et a donné lieu à quelques abus, il semble que les utilisateurs font preuve de plus de discernement. Les conditions d'application sont en effet si restrictives qu'elles sont rarement réunies, et il faut un tour de main non négligeable pour s'en servir efficacement.

Le modèle log-linéaire souffre surtout d'être trop peu connu. Moins ancien que le modèle linéaire, il ne fait pas partie en France du parcours obligé de l'apprentissage de la statistique. A tort sans doute, car il est le pendant naturel du modèle linéaire pour les variables qualitatives qui sont légions dans les enquêtes.

A partir du moment où l'on considère (à juste titre) qu'utiliser le modèle linéaire ou le modèle log-linéaire avec ses techniques dérivées n'est ni plus ni moins compliqué qu'une autre méthode d'analyse de données, on peut en attendre de réelles satisfactions dans le domaine qui est le sien.

Pour reprendre l'exemple de la satisfaction de l'usager du téléphone qui nous a beaucoup servi dans ce chapitre, il faut pouvoir quantifier l'impact de telle ou telle mesure qui sera prise. Quel sera, par exemple, l'impact prévisible du nettoyage systématique de toutes les cabines de téléphone sur l'indice de satisfaction ? Lorsque la décision sera effectivement prise de nettoyer toutes les cabines, la vague d'enquête suivante permettra de confirmer ou d'infirmer l'impact prévu : ce garde-fou, outre le fait qu'il rend le statisticien très modeste car l'erreur est si vite arrivée, demande une vérification systématique de tous les outils et de toutes les hypothèses.

A ce niveau, il ne s'agit plus de l'outil simple et galvaudé de l'initiation à la statistique mais un outil relativement sophistiqué qui le rend indispensable, et à vrai dire peu concurrencé pour l'action. S'il n'est pas tout à fait le seul outil à permettre d'expliquer des variables, il est le seul à pouvoir traiter les plans d'enquêtes un peu complexes, à permettre d'éliminer simplement les variables influentes et surtout, il est l'un des rares outils disponibles pour donner des conclusions chiffrées.

10

TRAITEMENT DES QUESTIONS OUVERTES

Ludovic Lebart

Centre National de la Recherche Scientifique
Ecole Nationale Supérieure des Télécommunications
Paris

10.1 Questions ouvertes et questions fermées

Les outils et méthodes présentés aux chapitres précédents concernaient tous le traitement de variables numériques, ordinales ou nominales. Les variables nominales étudiées résultaient de la codification de réponses à des questions fermées, c'est-à-dire à des questions dont les réponses sont prévues à l'avance (cf. chapitre 3, section 3.3).

Dans un certain nombre de situations, qui seront évoquées plus bas, il peut être intéressant au contraire de laisser ouvertes certaines questions, dont les réponses se présenteront donc sous forme de textes de longueurs variables. Le traitement de ce type d'information est évidemment complexe. Ce chapitre doit montrer quelle aide les outils de calcul et les méthodes statistiques peuvent apporter à l'analyse de ces *réponses libres*.

On rappellera auparavant quelques uns des problèmes posés par la rédaction des libellés des questions dans les questionnaires d'enquêtes.

10.1.1 Le libellé des questions

Le libellé d'une question joue un rôle fondamental : il est très difficile de trouver deux libellés distincts, pour deux questions fermées dont les contenus sont similaires, donnant les mêmes résultats en termes de pourcentages.

La sensibilité des pourcentages de réponses vis-à-vis des libellés peut d'ailleurs être à l'origine de "manipulations d'opinion", dans la mesure où elle permet de moduler les taux de réponses à propos d'un même thème, les résultats faisant l'objet d'une publication sélective.