REVUE DE STATISTIQUE APPLIQUÉE

F. DERRIENNIC
P. DUCIMETIERE

Comparaison de deux méthodes statistiques d'analyse des taux de mortalité

Revue de statistique appliquée, tome 25, n° 4 (1977), p. 37-44
<http://www.numdam.org/item?id=RSA_1977__25_4_37_0>

COMPARAISON DE DEUX MÉTHODES STATISTIQUES D'ANALYSE DES TAUX DE MORTALITÉ

F. DERRIENNIC [1] ET P. DUCIMETIERE [2]

RESUME

L'analyse des taux de mortalité pour une cause de décès donnée selon les modalités conjointes de plusieurs facteurs ne peut pas être effectuée par les techniques habituelles d'analyse de variance, car les taux sont calculés sur des effectifs généralement très variables.

Pour résoudre ce problème, le modèle logistique classique (COX) permet d'analyser les effets principaux des facteurs. Récemment un modèle additif (EL SHAARAWI) a été proposé conduisant à une analyse de variance non orthogonale.

Le but de ce travail est de comparer les résultats obtenus par ces deux types d'approche, l'estimation des paramètres du modèle logistique étant effectué comme un cas particulier d'une méthode plus générale d'estimation dans les tableaux de contingence développée par HUBER et LELLOUCH.

L'analyse de la mortalité par différentes causes en France en 1968 chez les hommes d'âge compris entre 45 et 64 ans en fonction de leur catégorie socio-professionnelle et de leur région de domicile est choisie comme exemple.

Les statistiques nationales de mortalité demeurent un outil très utilisé, bien qu'imparfait, dans les études de santé. La recherche d'éventuels facteurs d'environnement, démographiques, socio-économiques... conduit souvent à comparer les taux de mortalité générale ou par causes spécifiques dans des sous-groupes de la population, définis par des critères multiples : régions géographiques, tranches d'âge, époques...

Classiquement, le modèle logistique (COX 1) permet d'analyser les effets des facteurs sur les taux de décès et dans le cas particulier de taux petits par rapport à l'unité, le modèle est en pratique de type multiplicatif. Récemment, leur analyse par un modèle additif a été proposée (EL SHAARAWI [3]) conduisant à l'estimation des paramètres dans une analyse de variance non orthogonale.

Le but de la présente note est de comparer les résultats obtenus par ces deux types de méthode, l'estimation des paramètres du modèle logistique étant effectuée comme un cas particulier d'une méthode plus générale d'estimation dans les tableaux de contingence développée par HUBER et LELLOUCH [4].

L'analyse de la mortalité par différentes causes en France en 1968 chez les hommes d'âge compris entre 45 et 64 ans, en fonction de leur catégorie socio-professionnelle et de la région géographique est choisie comme exemple.

---

(1) Centre de Recherche INSERM 44, Chemin de Ronde 78110 Le Vesinet.

(2) Unité de Recherches Statistiques INSERM, 16 Bis A.V. Paul-Vaillant Couturier 94800 Villejuif.

Mots-clés : **Estimation. Mortalité. Modèle additif. Modèle multiplicatif.**

MATERIEL ET METHODES

1) Données de base

Les données sont établies dans le cadre d'une étude sur la mortalité cardiovasculaire en France et sont décrites complètement dans ce dernier travail (DERRIENNIC et DUCIMETIERE [2]).

La population française de référence est celle de 1968 (année du dernier recensement disponible), masculine âgée de 44 à 65 ans, catégorie choisie pour limiter l'imprécision dans la codification des causes de décès et dans la répartition des sujets selon les catégories socioprofessionnelles.

9 catégories socio-professionnelles sont constituées(*). Les sujets dits inactifs (16 % des hommes du groupe d'âge) sont exclus car peu comparables aux sujets en activité.

19 régions géographiques sont étudiées après exclusion de 3 régions (Pays de la Loire, Provence - Côte d'Azur et Corse) pour lesquelles le pourcentage de causes de décès mal définies ou non spécifiées est particulièrement important (12 à 21% contre 4 à 7 % pour les autres).

En dehors de la mortalité toutes causes, différents groupes de causes de décès sont constitués à titre d'exemple : maladies cardiaques, cérébro-vasculaires, tumeurs, maladies respiratoires, maladies digestives, accidents et morts violentes. Leur définition précise est donnée dans [2].

2) Modèle multiplicatif

Ce modèle s'introduit aisément comme cas particulier de l'estimation de distributions multinomiales (4). Considérons en effet les trois variables polychotomiques suivantes appliquées à l'ensemble de la population étudiée :

X_1 : sujet vivant ou décédé (2 modalités)

X_2 : catégorie socio-professionnelle du sujet (9 modalités)

X_3 : région géographique du sujet (19 modalités)

Le modèle le plus général donnant la probabilité conjointe p_{ijk} sans faire intervenir l'intéraction d'ordre 3 entre les variables peut s'écrire :

$$p_{ijk} = \text{Prob}\,(X_1 = i, X_2 = j, X_3 = k) = C\, f_{X_1 X_2}(i,j)\, f_{X_1 X_3}(i,k)\, f_{X_2 X_3}(j,k)$$

où C est une constante telle que $\sum_{ijk} p_{ijk} = 1$

Les deux premiers termes f expriment d'une certaine manière la variation du taux de décès selon la catégorie socio-professionnelle et la région, le troisième exprime la liaison qui existe entre ces deux variables.

(*) d'après les définitions de l'Institut National de la Statistique et des Etudes Economiques.

Les estimations $\hat{p}_{ijk}$ sont obtenues par la maximisation de la vraisemblance des observations du tableau de contingence par une méthode itérative qui utilise le fait que les statistiques exhaustives pour l'estimation du modèle sont ici les marginales d'ordre 2.

Le rapport $\frac{p_{2jk}}{p_{1jk}}$ s'écrit $\frac{f_{X_1X_2}(2,j)}{f_{X_1X_2}(1,j)} \quad \frac{f_{X_1X_3}(2,k)}{f_{X_1X_3}(1,k)}$

c'est-à-dire $\phi_j \psi_k$ avec des notations évidentes.

Le taux de décès $t_{jk} = \frac{p_{2jk}}{p_{1jk} + p_{2jk}}$ est ainsi mis sous la forme logistique générale $\text{Log} \frac{t_{jk}}{1 - t_{jk}} = \text{Log}\, \phi_j + \text{Log}\, \psi_k$

et comme $t_{jk} \ll 1$ le modèle peut s'écrire : $\boxed{t_{jk} = \phi_j \psi_k}$ (1)

qui exprime le taux de décès dans la catégorie (j , k) sous forme multiplicative. Cependant les estimations t_{jk} obtenues à partir des p_{ijk} ne permettent pas une détermination unique des ϕ_j et ψ_k et il est nécessaire pour cela de poser des contraintes.

De manière à comparer aisément les estimations obtenues, à celles du modèle additif, on posera $\hat{t}_{jk} = \hat{A}\, \hat{\phi}_j\, \hat{\psi}_j$ où les $\hat{\phi}_j$ et $\hat{\psi}_k$ vérifient la double contrainte :

$$\sum_j n_{j.}\, \phi_j = \sum_k n_{.k}\, \psi_k = n.. \qquad (n_{j.} = \sum_i n_{ji}\,,\ n_{.k} = \sum_1 n_{ik}\ ,\ n.. = \sum_{ij} n_{ij})$$

Les estimations $\hat{A}$, $\hat{\phi}_j$ et $\hat{\psi}_k$ sont alors entièrement déterminées (c.f. Annexe 1) à partir des taux estimés $\hat{t}_{jk}$.

3) Modèle additif

Il s'écrit $t_{jk} = \mu + a_j + b_k + \epsilon_{jk}$. ϵ_{jk} est distribué normalement de variance $\frac{\sigma^2}{n_{jk}}$ où n_{jk} est l'effectif de la catégorie (j , k) et σ^2 une quantité constante.

Les estimations $\hat{a}_j$ et $\hat{b}_k$ sont obtenues par la méthode du maximum de vraisemblance sous la double contrainte :

$$\sum_j n_{j.}\, \hat{a}_j = \sum_k n_{.k}\, \hat{b}_k = 0$$

$\hat{\mu}$ est le taux de décès estimé sur l'ensemble de la population.

La méthode d'estimation et de calcul des paramètres est décrite par EL SHAARAWI et Coll. [3].

4) Comparaison des deux modèles

Le choix des contraintes que vérifient les estimateurs conduit à comparer les paramètres $\hat{A}$ et $\hat{\mu}$ d'une part et $\hat{\phi}_j$ (resp. $\hat{\psi}_k$) à $\frac{\hat{\mu} + \hat{a}_j}{\hat{\mu}}$ resp. $\frac{\hat{\mu} + \hat{b}_k}{\hat{\mu}}$ d'autre part (cf. Annexe 2). Ceci peut être interprété comme la comparaison des estimations d'un "taux relatif de mortalité" propre à la catégorie socio-professionnelle j (ou à la région k) par rapport à celui de la population générale. Par ailleurs, l'adéquation relative de chacun des modèles aux observations peut être jugée en calculant dans chaque cas, le tableau des effectifs des décédés dans chaque catégorie (j , k) tableau qui est à comparer au tableau observé par le calcul d'un χ^2. En théorie, les tableaux calculé et observé des effectifs des survivants devraient intervenir mais leur contribution à la valeur du χ^2 est négligeable.

RESULTATS

Les résultats complets concernant l'analyse de la mortalité toutes causes dans la population sont donnés à titre d'exemple. Les estimations $\hat{\mu}$ et $\hat{A}$ sont respectivement 1189 et 1205 pour 100 000. Le tableau 1 indique pour chaque catégorie socio-professionnelle les valeurs de $\hat{\phi}_j$ et $\frac{\hat{\mu} + \hat{a}_j}{\hat{\mu}}$ pour la mortalité toutes causes. Bien que les "taux relatifs de mortalité" varient de 0,60 (Cadres supérieurs) à 1,35 (Ouvriers qualifiés), leurs estimations par les deux modèles sont pratiquement identiques.

TABLE 1

Comparaison des paramètres estimés par le modèle additif (1) et multiplicatif (2)
(catégories socio-professionnelles, mortalité toutes causes)

	(1) $\frac{\hat{\mu} + \hat{a}_j}{\hat{\mu}}$	(2) $\hat{\phi}_j$
Ouvriers qualifiés	1.35	1.35
Salariés agricoles	1.30	1.30
Employés	1.18	1.19
Artisans, commerçants	1.03	1.03
Ouvriers	0.98	0.97
Agriculteurs	0.90	0.90
Cadres moyens	0.79	0.79
Industriels	0.78	0.78
Cadres supérieurs	0.61	0.60

Il en est de même pour les variations de mortalité selon la région géographique (Tableau 2).

TABLE 2

Comparaison des paramètres estimés par le modèle additif (1) et multiplicatif (2)
(régions géographiques, mortalité toutes causes)

	(1) $\frac{\hat{\mu} + \hat{b}_k}{\hat{\mu}}$	(2) $\hat{\psi}_k$
Bretagne	1.55	1.56
Alsace	1.50	1.52
Limousin	1.30	1.31
Lorraine	1.12	1.12
Basse-Normandie	1.11	1.11
Nord	1.04	1.04
Rhône-Alpes	1.02	1.02
Franche-Comté	1.02	1.01
Auvergne	1.00	1.00
Picardie	0.96	0.96
Bourgogne	0.96	0.96
Haute-Normandie	0.93	0.93
Région Parisienne	0.91	0.91
Champagne	0.91	0.91
Aquitaine	0.89	0.88
Midi-Pyrénées	0.87	0.86
Poitou-Charentes	0.85	0.85
Centre	0.84	0.84
Languedoc-Roussillon	0.77	0.77

Les valeurs du χ^2 d'adéquation aux modèles sont données au Tableau 3 pour la mortalité toutes causes et pour les différentes grandes catégories de causes de décès.

TABLE 3

Comparaison des valeurs du χ^2 d'adéquation du modèle additif (1) et multiplicatif (2) aux observations des nombres de décès
(mortalité par causes spécifiques et toutes causes)

	(1)	(2)
maladies cardiovasculaires	301	290
maladies cérébrovasculaires	220	212
tumeurs	390	338
maladies du système respiratoire (*)	276	241
Maladies du système digestif(*)	286	236
accidents, mort violente	309	263
toutes causes	1 028	910

(*) à l'exclusion des tumeurs

Elles sont dans tous les cas plus petites pour le modèle multiplicatif, le gain étant de l'ordre de 4 à 17 % par rapport au modèle additif. On peut remarquer qu'en attribuant à ces χ^2 un nombre de degrés de liberté approximatif égal à (9-1) (19-1) = 144, ils sont cependant tous très significatifs, ce qui implique une intéraction entre le facteur géographique et socio-professionnel dans leurs effets sur les taux de mortalité non prise en compte dans les deux modèles.

DISCUSSION

L'analyse simultanée des taux de décès dans des sous-groupes de la population définis par le croisement de deux facteurs ne peut être effectuée par les techniques habituelles d'analyse de variance compte-tenu de la variabilité des effectifs de ces sous-groupes.

Les paramètres du modèle multiplicatif sans intéraction : $t_{jk} = A\ \phi_j\ \psi_k$ présentent la particularité d'être interprétables en termes de "risques relatifs" dès que le paramètre $\hat{A}$ peut être assimilé à un taux de mortalité dans l'ensemble de la population par analogie avec la règle de multiplication des probabilités.

Or l'examen des formules (1) de l'Appendice donnant les estimations des paramètres montre que $\hat{A}$ représente bien un taux de mortalité moyen mais recalculé dans une population théorique où les effectifs de chaque sous-groupe (j , k) est $\frac{n_{j.}\ n_{.k}}{n_{..}}$ c'est-à-dire pour laquelle les deux facteurs sont distribués indépendamment, conditionnellement aux totaux marginaux.

L'écart entre les estimations $\hat{A}$ et $\hat{\mu}$ reflète donc l'effet de la liaison entre les deux facteurs dans la population. Cet écart est ici très faible.

Les paramètres du modèle additif sans intéraction $t_{ik} = \mu + a_j + b_k$ ne présentent pas (en dehors de μ) d'interprétation simple en termes de probabilités de décès et le modèle doit être considéré sous l'angle purement opérationnel. Dans l'exemple étudié dans le présent travail, les estimations qu'il fournit sont quasiment identiques à celles du modèle multiplicatif.

Il présente cependant l'avantage de permettre l'établissement de tests de signification des effets des facteurs et de leurs intervalles de confiance, simplement à partir de l'analyse de variance (EL SHAARAWI [3]). Dans le cas du modèle multiplicatif. des calculs similaires peuvent être conduits à partir de la théorie du maximum de vraisemblance mais sont plus lourds.

L'adéquation des deux modèles aux observations de la mortalité pour différentes causes apparaît à l'avantage du modèle multiplicatif bien que ce dernier ne permette pas de faire disparaître l'intéraction entre la catégorie socio-professionnelle et la région géographique dans leurs effets sur les taux de mortalité étudiés.

En conclusion, en ce qui concerne l'estimation des effets dûs aux facteurs, les deux modèles ne sont guère discernables mais le modèle additif apparaît plus simple à mettre en œuvre sur le plan des calculs. Le modèle multiplicatif s'ajuste mieux aux observations et peut se révéler utile lorsque l'on s'intéresse particulièrement à l'effet de l'intéraction entre les facteurs.

REFERENCES

[1] COX D.R. – The analysis of binary data. Methuen and C° London, 1970.

[2] DERRIENNIC F., DUCIMETIERE P. – La mortalité cardiaque des Français actifs d'âge moyen selon leur catégorie socio-professionnelle et leur région de domicile. *Rev. d'Epid. Santé Publique,* 1977 (à paraître).

[3] EL SHAARAWI A.H., CHERRY W.H., FORBES W.F., PRENTICE R.L. – A statistical model for studying regional differences in observed mortality rates, and its application to Ontorio during 1964-1968. *J. Chron. Dis.*, 29, 311-330, 1976.

[4] HUBER C., LELLOUCH J. – Estimation dans les tableaux de contingence à un grand nombre d'entrées. *Int. Stat. Rev.*, 42, 193-203, 1974.

Nous remercions Mr J. LELLOUCH pour ses nombreux conseils.

ANNEXE

1) Estimation des paramètres du modèle multiplicatif

Posons :

$$\hat{A} = \frac{1}{n..} \sum_{r,s} \frac{n_{r.}\, n_{.s}}{n..}\, \hat{t}_{rs}$$

$$\hat{\phi}_j = \frac{1}{A} \sum_{s} \frac{n_{.s}}{n..}\, \hat{t}_{js} \qquad (1)$$

$$\hat{\psi}_k = \frac{1}{A} \sum_{r} \frac{n_{r.}}{n..}\, \hat{t}_{rk}$$

calculés à partir des taux de décès estimés $\hat{t}_{jk}$

Ces quantités sont solutions du systèmes : $\forall_{jk} : \hat{t}_{jk} = \hat{A}\, \hat{\phi}_j\, \hat{\psi}_k$

en effet :

$$\hat{A}\, \hat{\phi}_j\, \hat{\psi}_k = \frac{1}{\hat{A}} \frac{\sum_{s} n_{.s}\, \hat{t}_{js} \sum n_{r.}\, \hat{t}_{rk}}{n^2_{..}} = \frac{\sum_{rs} n_{r.}\, n_{.s}\, \hat{t}_{js}\, \hat{t}_{rk}}{\sum_{rs} n_{r.}\, n_{.s}\, \hat{t}_{rs}} = \hat{t}_{jk}$$

car les estimations $\hat{t}$ vérifient $\forall_{r,s} : \hat{t}_{js}\, \hat{t}_{rk} = \hat{t}_{jk}\, \hat{t}_{rs}$ ainsi que l'impose la forme du modèle.

D'autre part, elles vérifient les contraintes car :

$$\sum_j n_{j.}\,\hat{\phi}_j = \frac{1}{A}\,\frac{\sum_j n_{j.} \sum_s n_{.s}\,\hat{t}_{js}}{n..} = n..$$

$$\sum_k n_{.k}\,\hat{\psi}_k = \frac{1}{A}\,\frac{\sum_k n_{.k} \sum n_{r.}\,\hat{t}_{rk}}{n..} = n..$$

2) Relations entre les estimations des paramètres des deux modèles

Soit $\hat{t}_{rs}$ (resp. $\tilde{t}_{rs}$) l'estimation du taux de décès dans la catégorie (r , s) par le modèle multiplicatif (resp. additif). En écrivant $\hat{t}_{rs}$ dans les formules (1) sous la forme $(\hat{t}_{rs} - \tilde{t}_{rs}) + \tilde{t}_{rs}$ et remplaçant le deuxième terme par $\hat{\mu} + \hat{a}_r + \hat{b}_s$, les contraintes que vérifient les quantités $\hat{a}_r$ et $\hat{b}_s$ conduisent aux relations :

$$\hat{A} = \hat{\mu} + \frac{1}{n..}\sum_{rs} \frac{n_{r.}\,n_{.s}}{n..}\,(\hat{t}_{rs} - \tilde{t}_{rs})$$

$$\hat{\phi}_j = \frac{\hat{\mu}}{\hat{A}}\,\frac{(\hat{\mu} + \hat{a}_j)}{\hat{\mu}} + \frac{1}{\hat{A}}\sum_s \frac{n_{.s}\,(\hat{t}_{js} - \tilde{t}_{js})}{n..}$$

$$\hat{\psi}_k = \frac{\hat{\mu}}{\hat{A}}\,\frac{(\hat{\mu} + \hat{b}_k)}{\hat{\mu}} + \frac{1}{\hat{A}}\sum_r \frac{n_{r.}\,(\hat{t}_{rk} - \tilde{t}_{rk})}{n..}$$

Sous cette forme, on s'aperçoit que $\hat{A}$ peut être assimilé à $\hat{\mu}$ et $\hat{\phi}_j$ (resp. $\hat{\psi}_k$) à $\frac{\hat{\mu} + \hat{a}_j}{\hat{\mu}}$ $\left(\text{resp. } \frac{\hat{\mu} + \hat{b}_k}{\hat{\mu}}\right)$ dès que les termes correctifs, moyennes pondérées des écarts entre les estimations des taux de décès par les deux modèles sont petits.