Régression linéaire symbolique avec variables taxonomiques.

Filipe Afonso*·**, Lynne Billard***
Edwin Diday*

*Ceremade/Université Paris 9 Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16, France.
afonso@ceremade.dauphine.fr
diday@ceremade.dauphine.fr
**Lamsade/ Université Paris 9 Dauphine
***Department of statistics/University of Georgia
Athens, 30602, USA.
lynne@stat.uga.edu

Résumé. Le présent papier concerne l'extension des méthodes classiques de régression linéaire aux cas des données symboliques et fait suite à de précédents travaux de Billard et Diday sur la régression linéaire avec variables intervalles et histogrammes. Dans ce papier, nous présentons des méthodes de régression avec variables taxonomiques. Les variables taxonomiques sont des variables organisées en arbre exprimant plusieurs niveaux de généralité (les villes sont regroupées en régions qui sont elles-mêmes regroupées en pays). La méthode proposée sera testée sur données simulées. Finalement, nous observerons que ces méthodes nous permettent d'utiliser la régression linéaire pour étudier des concepts et pour réduire le nombre de données afin d'améliorer les résultats obtenus par rapport à une régression classique.

1 Introduction

Dans la pratique, nous sommes souvent intéressés par l'étude de groupes d'individus plutôt que les individus statistiques eux-mêmes. Aussi, les bases de données atteignent des masses considérables d'observations et la réduction du nombre de données peut faciliter les études. Dans ces deux cas une agrégation des données va nous amener à manipuler des variables qui ne sont pas à valeurs uniques. Nous obtenons par exemple des intervalles, des histogrammes et des diagrammes. De plus, ces variables peuvent être organisées par des taxonomies ou des hiérarchies. Les données ainsi constituées sont appelées données symboliques (Billard et Diday 2003, Bock et Diday 2000). Des travaux sur l'extension des méthodes de régression linéaire aux cas des variables intervalles et histogrammes ont déjà été entrepris dans (Billard et Diday 2000 et 2002). Dans ce papier, nous nous intéressons aux variables taxonomiques (Voir aussi Afonso et al 2003).

2 Problématique

En régression linéaire, nous voulons expliquer une variable dépendante Y à partir de k variables explicatives $X_1,\ldots, X_k$ sous la forme d'un modèle linéaire $Y= a+b_1X_1+\ldots+b_kX_k+\varepsilon = \beta X+\varepsilon$ où ε constitue le résidu. Dans la théorie classique, nous calculons le vecteur optimal β^*

des poids en minimisant l'erreur. Nous obtenons la formule $\beta^*=(X'X)^{-1}X'Y$.

$\Rightarrow$ Nous allons, dans notre étude, proposer les matrices X et Y à régresser.

3 Régression linéaire avec variables taxonomiques

Les variables taxonomiques sont des variables organisées en arbre exprimant plusieurs niveaux de généralité. Par exemple, des régions démographiques sont divisées en villes au plus bas niveau de l'arbre qui seront regroupées en régions au 2ème niveau et en pays au 3ème. Nous pouvons aussi avoir des nuances de couleurs regroupées en tonalités(FIG. 1).

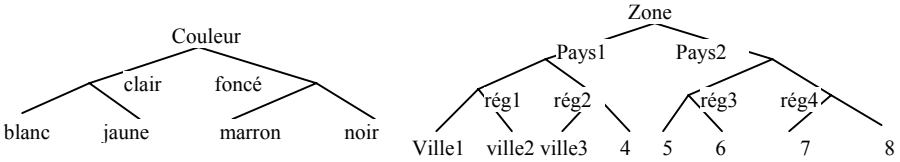

FIG. 1 –*taxonomies couleur et zone.*

3.1 Données de départ

Nous disposons de données où cohabitent des valeurs sur différents niveaux (TAB 1).

Nous avons l'information sur la tonalité mais pas sur la couleur

Individu	Couleur	Zone	revenu
1	jaune	ville1	4400
2	noir	ville 2	4000
3	blanc	ville 3	3500
4	noir	ville 3	2000
5	clair	region1	2500
6	jaune	pays1	2800
7	noir	ville 5	3800
8	marron	region3	3000
9	jaune	ville 7	4200
10	clair	ville 8	3800

Nous avons l'information sur la région ou sur le pays mais pas sur la ville.

TAB 1 – *Données avec variables taxonomiques.*

3.2 Méthode de régression à partir de ces données

Nous allons définir une équation propre à chaque niveau de la taxonomie avec tous les individus. Pour chaque niveau, nous allons faire une régression après avoir augmenté la généralité des données de niveaux inférieurs et diminué la généralité des données de niveaux supérieurs. Pour diminuer la généralité des valeurs de niveaux supérieurs, nous affecterons un poids à chaque fils f. Nous pouvons prendre comme poids w(f) = 1/m (m = nombre de fils), $w(f) = P_f$ = proportion de la modalité f dans les données ou bien tout P_f tel que $\Sigma P_f=1$. La qualité de la régression sera donc dépendante des poids choisis. Nous faisons la régression après avoir décomposé les variables en modalités prenant leur valeur dans l'intervalle [0,1]. Il nous faudra enlever une modalité à chaque variable afin d'inverser (X'X). Nous présentons les résultats obtenus avec les données du paragraphe 3.1.

Pour la variable « Zone », au niveau 1 (TAB 2A): nous faisons la régression au niveau des villes après avoir remplacé Region1 par ½ville1 ½ville2, Region2 par ½ville3 ½ville4,

…, pays1 par ¼ville1 ¼ville2 ¼ville3 ¼ville4 et pays2 par ¼ville5 ¼ville6 ¼ville7 ¼ville8. Au niveau 2 (TAB 2B), nous faisons la régression au niveau des régions après avoir agrégé les villes à leur région (ville3 devient region2…) et remplacé pays1 par ½Region1 ½Region2 et pays2 par ½Region3 ½Region4. Au niveau 3 (TAB 2C), nous faisons la régression au niveau des pays après avoir agrégé les villes et les régions à leur pays (ville3 est remplacé par pays1 et region3 par pays2…). Nous faisons de même avec "Couleur".

I	blanc	jaune	marron	noir	ville1	ville2	ville3	ville4	ville5	ville6	ville7	ville8
1	0	1	0	0	1	0	0	0	0	0	0	0
2	0	0	0	1	0	1	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	0	0
4	1	0	0	1	0	0	1	0	0	0	0	0
5	1/2	1/2	0	0	1/2	1/2	0	0	0	0	0	0
6	0	1	0	0	1/4	1/4	1/4	1/4	0	0	0	0
7	0	0	0	1	0	0	0	0	1	0	0	0
8	0	0	1	0	0	0	0	0	1/2	1/2	0	0
9	0	1	0	0	0	0	0	0	0	0	1	0
10	1/2	1/2	0	0	0	0	0	0	0	0	0	1

TAB 2A – *Matrice à régresser au niveau 1.*

I	clair	foncé	Region1	Region2	Region3	Region4
1	1	0	1	0	0	0
2	0	1	1	0	0	0
3	1	0	0	1	0	0
4	0	1	0	1	0	0
5	1	0	1	0	0	0
6	1	0	1/2	1/2	0	0
7	0	1	0	0	1	0
8	0	1	0	0	1	0
9	1	0	0	0	0	1
10	1	0	0	0	0	1

I	clair	foncé	pays1	pays2
1	1	0	1	0
2	0	1	1	0
3	1	0	1	0
4	0	1	1	0
5	1	0	1	0
6	1	0	1	0
7	0	1	0	1
8	0	1	0	1
9	1	0	0	1
10	1	0	0	1

TAB 2B, 2C – *Matrice à régresser aux niveaux 2 et 3.*

3.3 Test de la méthode sur données simulées

Nous allons comparer les différents modèles à l'aide d'un exemple sur données simulées à partir de statistiques réelles. Pour le premier test, nous avons n = 10 000 individus extraits de la population américaine; une variable à expliquer « revenus » ; deux variables taxonomiques explicatives « Zone » (4 niveaux) et « Travail » (2 niveaux).

Dans un deuxième test, nous additionnons d'autres variables aux variables taxonomiques. Ainsi, à l'exemple précédent, nous ajoutons sept variables continues classiques « age », « glucose », « cholestérol », « hémoglobine », « hématocrite », « globule rouge », « globule blanc » et 3 variables nominales classiques « couleur », « age groupe » et « diabète ».

Nous allons utiliser le coefficient de détermination R^2 pour tester la qualité de la méthode : $R^2 = \Sigma(y_i^*-y_{moy})^2/\Sigma(y_i-y_{moy})^2 = SSE/SST \rightarrow 1$ quand $SSE \rightarrow SST$ (Y = variable dépendante, Y* = prédictions de Y calculées avec l'équation, y_{moy} = moyenne des y_i).

Nous allons faire plusieurs régressions en remplaçant progressivement et aléatoirement des données à chaque niveau de la hiérarchie par des données de plus haut niveau afin de voir la sensibilité de cette méthode au manque d'information. Ainsi, dans un premier temps toutes les données de plus bas niveaux sont utilisées (TAB 3 colonne données

supprimées=1), dans un deuxième temps nous agrégeons un petit nombre de données (12% au niveau 1 de la taxonomie, 6% au niveau 2 et 3% au niveau 3, colonne données supprimées=2) et finalement nous supprimons beaucoup de données au niveau 1 et 2 (40%, 20%, 10% aux niveaux 1, 2, 3 respectivement, colonne données supprimées=3).

		Test 1			Test 2		
Niveau:	Données supprimées:	1	2	3	1	2	3
1	R^2	0,33	0,31	0,25	0,86	0,84	0,78
2	R^2	0,16	0,16	0,14	0,72	0,72	0,70
3	R^2	0,05	0,05	0,05	0,59	0,59	0,59
4	R^2	0,03	0,03	0,03	0,56	0,56	0,56

TAB 3 – *Coefficients de détermination R^2 en fonction de l'augmentation de la généralité des données (1, 2, 3) pour les tests 1 et 2.*

Nous observons que les coefficients de détermination restent assez stables même lorsque nous supprimons un nombre important de données. En effet R^2 passe de 0.33 à 0.25 pour le test 1 et de 0.86 à 0.78 pour le test 2 lorsque 40% des données de niveau 1 sont supprimées. De plus, nous constatons que le niveau 1 est le meilleur, que le niveau 2 est convenable et qu'il rattrape légèrement le niveau 1 lorsqu'il manque beaucoup de données à ce plus bas niveau. Enfin, dans cet exemple, les niveaux 3 et 4 donnent de très mauvais résultats. Par conséquent, en exploitant la définition de la taxonomie, nous allons pouvoir montrer qu'un niveau de la taxonomie est explicatif mais que les niveaux au-dessus ne le sont pas forcément. Au contraire, un niveau supérieur pourra être intéressant s'il manque beaucoup de données aux premiers niveaux.

4 Le module SREG

Ces différentes méthodes ont été implantées dans le logiciel SODAS du programme européen d'analyse de données symboliques ASSO (Voir le site http:// www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm). Ce module SREG fournit des méthodes et des tests pour la régression linéaire multiple avec les variables symboliques intervalles, histogrammes, taxonomiques, hiérarchiques et diagrammes.

5 Applications de la régression symbolique

La régression symbolique pourra être utilisée pour étudier des concepts. Par exemple, nous pouvons étudier des groupes d'individus selon l'âge, la couleur et la région d'habitation. Ainsi, nous créons les concepts age × couleur × zone. Ces concepts dépendent des centres d'intérêt de celui qui fait l'étude ou provient directement de la définition des données. Par exemple, chaque enregistrement d'une base de données appartenant à une caisse d'assurance maladie peut faire référence à une ligne de remboursement pour un assuré mais nous ne nous intéressons pas à l'étude des individus « ligne de remboursement » mais à l'étude des assurés. Nous obtenons donc des données symboliques en agrégeant toutes les « lignes de remboursement » de chaque assuré. De plus, nous pouvons réduire la taille des données en construisant les concepts à partir de la variable dépendante. Par exemple, avec les données du point 3.3, si nous voulons étudier la variable « cholestérol » alors nous pourrons créer les concepts sur cette variable qui seront en fait des intervalles de cholestérol. Pour ne pas perdre les taxonomies, nous pourrons créer les concepts cholestérol × Zone ×

Travail. Si nous voulons étudier plusieurs variables dépendantes, nous créerons les concepts variable dépendante 1 × variable dépendante 2 ×….
Ainsi, à partir des données du point 3.3, nous calculons les régressions simples classiques des variables quantitatives « hématocrite » et « hémoglobine » et taxonomique « travail » sur la variable dépendante « cholestérol ». Nous présentons les tests de Fisher F de validation des régressions et les coefficients de détermination R² dans le tableau TAB 5A. Nous observons que le Test de Fisher accepte aisément les régressions simples avec les variables prédictives « hématocrite» et « hémoglobine » ce qui semble contradictoire avec les coefficients de déterminations qui sont très faibles (R^2_{hemat}=0.009 et R^2_{hemo}=0.015). En effet, en régression linéaire, un nombre important d'observations engendre énormément de bruit ce qui rend difficile l'interprétation. La variable « travail » est refusée. Nous allons comparer les résultats obtenus lorsque nous créons 28 × 7 = 196 concepts à partir de 28 intervalles de cholestérol ([57,99.9], [100,109.9], [110,119.9], [120,129.9], [130,134.9], [135,139.9], …, [230,234.9], [235,244.9], [245,259.9], [260,281]) et des 7 modalités de la variable « travail » afin de pouvoir conserver la taxonomie. Au 2ème niveau X_2 de la taxonomie nous avons 3 modalités pour le travail, 1=plein temps, 2=mi-temps, 3=pas de travail. Au premier niveau X_1, pour X_2 =1 et X_2 =2, nous avons quatre fils relatifs au nombre de semaines de travail. Après la création de ces concepts, la variable quantitative « hématocrite » devient une variable intervalle (grâce au module DB2SO de Sodas). De plus, nous créons une variable « hémogroupe » à partir de 10 intervalles d'hématocrite ([10,11], [11.1,11.4], ..., [13.9,14.2], [14.3,15.2]) qui deviendra une variable à valeurs histogrammes après la création des concepts. Finalement, la variable « travail » est toujours une variable taxonomique. Nous présentons un extrait des données initiales TAB 4A et leur transformation en données symboliques TAB 4B. Nous observons TAB 5B que Les régressions simples avec les variables « hématocrite » et « hémogroupe » sur le « cholestérol » sont également acceptées alors que la régression avec « travail » est refusée. De plus, les résultats sont beaucoup plus faciles à interpréter et semblent avoir plus de sens (R^2_{hemat}=0.42 et R^2_{hemo}=0.40). Finalement, en présence d'un nombre considérable d'individus dans la régression, le test de Fisher a tendance à accepter toutes les régressions. Après réductions des données, nous n'avons plus ce même problème.

i	concepts	cholestérol	X_2	X_1	X=travail	hématocrite	hémogroupe
1	[150,154.9] × 11	151	1	1	11	35	[13.9,14.2]
2	[150,154.9] × 11	153	1	1	11	32	[13.9,14.2]
3	[150,154.9] × 11	154	1	1	11	31	[13.5,13.8]
4	[180,184.9] × 12	180	1	2	12	43	[11,11.4]
5	[180,184.9] × 12	184	1	2	12	46	[11.5 ,11.8]

TAB 4A – *Extrait des données initiales classiques.*

concepts	cholestérol	travail	hématocrite	hémogroupe
[150,154.9] × 11	[151,154]	11	[31,35]	1/3[13.5,13.8], 2/3[13.9,14.2]
[180,184.9] × 12	[180,184]	12	[43,46]	½[11,11.4], ½[11.5 ,11.8]

TAB 4B – *Tables des concepts obtenus à partir de TAB 4A.*

Variables explicatives	Rapport de variances F	Quantile f(0.95)	R²
Hématocrite	90.84	4.17	0.009
Hémoglobine	156.87	4.17	0.015
Travail (niveau 1)	0.82	2.42	0.0005

TAB 5A – *Tests de Fisher et coefficients de détermination pour les régressions classiques simples. Variable dépendante : « cholestérol ».*

Variables explicatives	Rapport de variances F	Quantile f(0.95)	R^2
Hématocrite	34.5	4.17	0.42
Hémogroupe	32.3	4.17	0.40
Travail (niveau 1)	0.05	2.42	0.007

TAB 5B – *Tests de Fisher et coefficients de détermination pour les régressions symboliques simples. Variable dépendante : « cholestérol ».*

6 Conclusions et perspectives

Ces nouvelles méthodes ont permis d'étendre la régression linéaire aux cas des variables taxonomiques. Cependant, ce travail peut être complété suivant différents axes. Nous pouvons étendre la régression linéaire à d'autres variables et rechercher des tests spécifiques à la régression symbolique. Aussi, nous pouvons étendre ces méthodes à d'autres régressions comme la régression après analyse des correspondances et la régression bayésienne.

Finalement, nous avons vu que les résultats obtenus avec la régression symbolique étaient plus faciles à interpréter et avaient peut être plus de sens que les résultats de la régression classique pour une quantité importante de données. Cependant, la mise en place d'une rigueur mathématique afin de soutenir ces observations reste un problème ouvert.

Références

Afonso F., Billard L., Diday E. (2003), Extension des Méthodes de Régression Linéaire aux cas des Variables Symboliques Taxonomiques et Hiérarchiques, Actes des XXXVémes journées de Statistique, SFDS Lyon 2003,Vol. 1, pp 89-92.

Billard L., Diday E. (2003), From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis, Journal of the American Statistical Association, 98, pp470-487

Billard L., Diday E. (2002), Symbolic Regression Analysis, In Classification, Clustering and Data Analysis; Bock et al. Eds, Springer-Verlag; pp 281-288.

Billard L., Diday E. (2000), Regression Analysis for Interval-Valued Data, In Data analysis, Classification and Related Methods, Kiers et al. Eds, Springer-Verlag, pp 369-374.

Bock HH., Diday E. (2000), Analysis Data Sets: Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag.

Summary

This work deals with the extension of classical linear regression to symbolic data and constitutes a continuation of previous papers from Billard and Diday on linear regression with interval and histogram-valued data. In this paper, we present methods to regress taxonomic variables. Taxonomic variables are variables organized in a tree with several levels of generality (for example, the towns are aggregated up to their regions, the regions are aggregated up to their country…). All the methods are tested with a data set simulated from real statistics. Moreover, we note that with these methods, we can use symbolic linear regression in order to study concepts and also in order to summarize the initial data set with the intention of improving the results obtained with classical regression.