Série de Fiches Méthodologiques en Recherche et en Rédaction Scientifique
Fiche Méthodologique n°6:
Conduire une analyse multi variée par régression logistique
Conduct a multivariate analysis by logistic regression

إجراء تحليل متعدد المتغيرات عن طريق الانحدار اللوجستي

Kamel Ben Salem[1], Ahmed Ben Abdelaziz[1,2], Réseau Maghrébin : Pédagogie-Recherche-Publication (PRP2S)

1. Réseau Maghrébin Pédagogie-Recherche-Publication (PRP2S)
2. Laboratoire de Recherche LR19SP01. Université de Sousse. Tunisie

Cette série…

Le Réseau Maghrébin PRP2S et la Rédaction de la revue « La Tunisie Médicale » ont l'honneur de vous présenter, régulièrement à partir du numéro de janvier 2020, une série des fiches techniques en épidémiologie et en bio statistique. Ces fiches méthodologiques décrivent, d'une manière standardisée, les modes d'usage des concepts, des outils et des méthodes utilisés lors des différentes phases de la rédaction médicale scientifique depuis la phase de la recherche documentaire jusqu'à la phase de la communication médicale scientifique.
Cette série est rédigée par des experts de méthodologie de recherche dans les universités du Grand Maghreb et les facultés sœurs au Nord de la Méditerranée. Chaque fiche répond à trois questions essentielles (Quoi ? Pourquoi ? Comment) du concept étudié, en se basant sur un article publié dans la revue Tunis Med.
Le coordinateur de la série « Fiches Méthodologiques »
Professeur Ahmed Ben Abdelaziz (*Président du Réseau Maghrébin PRP2S*)
ahmedbenabdelaziz.prp2s@gmail.com

Série des Fiches méthodologiques
Sommaire

Correspondance

Kamel Ben Salem
Email : kbsalem@gmail.com

ETUDE DE CAS

En 2010, l'équipe du Professeur Dziri a conduit une étude épidémiologique sur les facteurs alimentaires prédisposant au risque de cancers colorectaux (CCR). Les chercheurs ont comparé un groupe de 32 patients ayant un CCR à un groupe témoin de 61 malades. Ils ont procédé dans un premier temps à une analyse uni variée ayant dégagé 12 facteurs (p< 0,05) influençant le risque de CCR (âge, sexe, origine géographique, anémie, tabagisme, sport, marche, charcuterie, lait, fruits, huile crue et fritures). Tous ces facteurs associés ont été introduits dans un modèle de régression logistique afin d'identifier les facteurs indépendants influençant le risque de CCP. Selon le tableau I, cette analyse multi variée n'a retenu que trois facteurs: l'âge 40/60 ans (ORa: 5,15 ; IC 95% [2,3–11,4]), la charcuterie consommation fréquente/ consommation rare (ORa: 5,1 ; IC 95% [1,4-18,5]), le lait consommation rare/consommation fréquente (ORa: 7,07 ; IC 95% [1,4-35,6]). Les auteurs ont conclu que le jeune âge et la consommation fréquente de charcuterie ont été des facteurs de risque de CCR alors que la consommation fréquente de lait a été un facteur de protection.

Tableau 1. Analyse multi variée des facteurs de risque des cancers colorectaux (Tunisie, 2010) [1]

	B	SE	OR	IC (95%)	P
Age 40/60ans	- 0,082	0,020	5,15	(2,3 11,4)	$< 10^{-4}$
Charcuterie	-1,631	0,655	5,1	(1,4 18,5)	0,013
Cons fqt / Cons rare					
Lait	+1,956	0,825	7,07	(1,4 35,6)	0,018
Cons rare / Cons fqt					

OR : Odds ratio ; IC : intervalle de confiance ; cons fqt : consommation fréquente ; cons rare : consommation rare

Quizz

1. Dans une étude multi variée par régression logistique binaire, la variable d'intérêt (à expliquer ou à prédire) est une variable qualificative dichotomique

    1. Vrai
    2. Faux

2. La régression logistique permet le contrôle des facteurs de confusion

    1. Vrai
    2. Faux

3. Dans une étude multi variée par régression logistique, le risque est calculé par un *Odds Ratio ajusté* (ORa)

    1. Vrai
    2. Faux

INTRODUCTION

L'exercice médical impose, au quotidien du médecin, la recherche de facteurs associés à un évènement de santé pour justifier un diagnostic, faire un pronostic et implanter des mesures préventives. L'épidémiologie analytique et la statistique, sciences qui permettent de pondérer et de tester la relation entre des variables explicatives et un évènement de santé, sont des outils fondamentaux comme aide à la prise de décision des professionnels de santé. Par ailleurs, un état de santé donné est rarement expliqué par un seul facteur. La nécessité d'expliquer cet état dans ses multiples dimensions, prenant en compte plusieurs facteurs, pose le problème de sa modélisation (représentation simplifiée d'une réalité complexe). Cette modélisation est possible en se basant sur des approches mathématiques probabilistes. La régression logistique répond à cet impératif d'analyse gérant simultanément plusieurs variables pour expliquer un évènement dichotomique. L'objectif de cette fiche méthodologique est de décrire la méthode épidémiologique d'analyse multi variée par régression logistique, des conditions de son application et l'interprétation de ses extrants dans les situations les plus fréquentes de la recherche en sciences de la santé [2].

LA RÉGRESSION LOGISTIQUE: POURQUOI ?

La majorité des phénomènes de santé se présentent sous forme dichotomique binaire, être malade ou ne pas l'être, avoir une complication ou ne pas l'avoir … en codant la présence de l'évènement étudié égal à «1», la fréquence de cette modalité peut être interprétée comme une probabilité. Cette approche nous permet de:

- **Prédire,** à l'aide des variables indépendantes «$\boldsymbol{X_i}$» (qualitatives et/ou quantitatives) caractérisant l'évènement de santé, la probabilité de l'observer (ou ne pas l'observer). Le modèle final retenu sera ainsi un **modèle prédictif.**
- **Déterminer** quelles sont les variables indépendantes «$\boldsymbol{X_i}$» (qualitatives et/ou quantitatives), qui expliquent de façon indépendante, la probabilité d'observer (ou ne pas observer) l'évènement étudié. Le modèle final retenu sera **descriptif** des caractéristiques **propres** à cet évènement, en les pondérant.

LA RÉGRESSION LOGISTIQUE: C'EST QUOI ?

La fonction logistique permet de modéliser les réponses binaires non linéaires dont l'intervalle des solutions est compris entre [0-1]. Cette fonction s'écrit sous forme :

$P(Y = 1/X_i) = \frac{e^{\beta_0+\sum \beta_i X_i}}{1+e^{\beta_0+\sum \beta_i X_i}}$ où :

- *Y* représente la variable dépendante à décrire ou à pronostiquer,
- β_i les coefficients associés aux variables explicatives X_i .

Ces coefficients, une fois calculés, se présentent sous forme «e^{β} », sont des *odds-ratio (OR)* quand la variable est qualitative. L'*odds* est le rapport de la probabilité de survenue d'un évènement divisée par l'évènement contraire, soit $odds = \frac{p}{1-p}$. Pour les variables quantitatives, ces coefficients dépendent de leur unité. La transformation logarithmique (logarithme naturel) de *l'odds*, appelée *«Logit»*, simplifie son écriture sous forme : $Logit(P) = \beta_0 + \sum \beta_i X_i$. Cette transformation logarithmique permettra de calculer la vraisemblance *(V)* du modèle qui est la probabilité d'observer cet échantillon. On appellera par ailleurs *déviance,* la quantité $-2Ln(V)$. Les logiciels statistiques permettent d'estimer ces différents coefficients par la méthode du maximum de vraisemblance et de calculer leur Intervalle de Confiance à 95%. Par ailleurs la quantité $2Ln(\frac{V_1}{V_2})$ suit une loi de *Chi2* à un degré de liberté (*ddl)* avec :

- V_1 : Vraisemblance du modèle à *k* modalités
- V_2 : Vraisemblance du modèle à *k-1* modalités.

Ce rapport peut également s'écrire sous forme d'une différence, il permet de tester l'effet sur le modèle de l'ajout ou le retrait d'une variable. Les vraisemblances de deux modèles, dits emboités, diffèrent statistiquement, la variable a son poids sinon elle peut être exclue du modèle.

LA RÉGRESSION LOGISTIQUE : COMMENT ?

1. Champs d'application

La régression logistique peut s'appliquer aux études épidémiologiques de cohorte, transversales et cas témoins; cependant pour ces dernières, le modèle retenu ne peut être que descriptif. Les études Cas/Témoins ne permettent pas la prédiction. Elle n'a aucune exigence sur la distribution de la variable; la normalité par exemple

n'est pas une condition nécessaire. La seule contrainte, est d'avoir simultanément au minimum une dizaine de réponses pour toutes les variables retenues (une cinquantaine pour certains auteurs) afin de garantir une puissance suffisante aux tests statistiques.

2. Le codage des variables

Il est impératif de coder la variable dépendante (évènement à expliquer): 0 si l'évènement est absent, 1 si évènement présent. Les variables indépendantes (explicatives) qualitatives doivent être dichotomisées autant que possible comme la variable dépendante (0/1). Les variables qualitatives à plus de deux modalités *(m)* doivent être signalées au logiciel d'analyse. Il se chargera de les stratifier en *(m-1)* modalités indicatrices dont sera prise comme référence. Les variables quantitatives peuvent être incluses dans le modèle en leur qualité ou dichotomisées selon une valeur seuil ayant une signification clinique. Elles seront ainsi traitées comme des variables qualitatives.

3. Construction et choix du modèle

La construction du modèle est la dernière étape de l'analyse. Les variables statistiquement significatives (généralement au seuil de 0,05), associées à la variable à expliquer lors de l'analyse uni variée, sont théoriquement «candidates» au modèle. Nous rappelons que l'analyse uni variée nous calcule également la force d'association entre la variable dépendante et la variable explicative *par l'OR brut.* Le seuil de signification des variables candidates peut aller jusqu'à un seuil de signification de 0,25 afin de rechercher d'éventuels facteurs de confusion ou des interactions entre deux variables. Cependant toutes les variables répondant à ces conditions ne sont pas automatiquement retenues. Deux règles fondamentales sont à respecter, la **parcimonie et la non redondance.** La parcimonie consiste à ne retenir que les variables cliniquement pertinentes, une revue de la littérature du problème étudiée est ainsi primordiale. La redondance s'applique pour les variables fortement corrélées (exemple ne pas retenir dans un même modèle le poids, la taille et l'indice de masse corporelle qui n'est que le rapport entre eux).

La modélisation peut se faire de deux façons.

La méthode descendante (pas à pas). Elle consiste, dans une première étape à prendre toutes les variables retenues et à réaliser l'analyse sur un modèle dit saturé à *k* variables. Puis, dans une deuxième étape, à soustraire une à une du modèle, la variable la moins significative (le *p* le plus élevé) et tester le nouveau modèle à *k-1* variable au modèle saturé à *k* variables par le rapport de vraisemblance qui suit une loi *de Chi2* à 1 *ddl*. Si la différence entre les deux variances est significative, la variable a son poids, si non elle peut être définitivement retirée. Nous pouvons également tester la variable par le test de Wald. Il nous calcule un *chi2* à un *ddl;* cette quantité est le carré du coefficient estimé par le modèle divisé par sa variance. Cette opération sera répétée de la même façon jusqu'à obtenir un modèle ayant des variables statistiquement significatives et indépendamment associées à l'évènement étudié. Les *OR* ainsi obtenus sont des *OR ajustés.*

La méthode ascendante (pas à pas). La deuxième façon d'opérer et de commencer par un modèle à une seule variable et d'ajouter une à une les variables sélectionnées et de tester au fur et à mesure les modèles emboités selon les mêmes principes.

4. Qualité du modèle

Ouf, vous n'êtes pas au bout de vos peines !. Le modèle construit doit avoir des qualités métrologiques acceptables. Sa validité dot être appréciée au moins par le test de Hosmer Lemshow au seuil de 0,1, que le logiciel se charge de son calcul. Ce test évalue son adéquation à décrire, le plus fidèlement possible, l'évènement étudié en fonction des variables retenues. Dans un modèle adéquat, les valeurs prédites seront proches des valeurs observées, reflet d'un ajustement correct.

5. Application sous «SPSS»

Les données de l'exemple suivant sont tirées d'une étude africaine non publiée sur les facteurs associés au HIV. Pour des raisons pédagogiques, nous choisirons la régression logistique ascendante pas à pas afin d'expliquer les étapes de la régression logistique en général. La variable à expliquer est la présence ou non de cas de HIV. L'analyse uni variée a tenu compte de variables comme les « antécédents de transfusion », les « scarifications » et l'usage de « préservatifs » entre autre. Dans un premier

temps nous allons inclure dans le modèle la notion de transfusion. Le logiciel nous donne les résultats suivants:

Tableau 2. Historiques des itérations dans les sorties d'un exemple de régression logistique, sur le logiciel SPSS

Historique des itérations			
Itération		-2log-vraisemblance Constante	Coefficients
Etape 0	1	855,169	-1,008
	2	853,688	-1,107
	3	853,687	-1,109
	4	853,687	-1,109

Ce premier tableau nous donne la déviance de départ d'un modèle ne contenant aucune variable. Il est basé uniquement sur la constante soit -2Log_vraisemblance= 853,687. Nous rappelons que cette valeur indique la quantité d'informations non retenues par le modèle ; Par conséquent l'ajout de variables fait baisser cette valeur.

L'ajout de la variable **transfusion** donne une déviance plus faible soit :

Tableau 3a. Récapitulatif des modèles dans les sorties d'un exemple de régression logistique, sur le logiciel SPSS

Récapitulatif des modèles			
Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	849,087[a]	,006	,009

Le tableau suivant est un *Chi2* à *un ddl*; sa valeur est la différence des deux déviances (853,687 - 849,087 = 4,6). L'ajout de la variable transfusion change de façon significative, le modèle (p=0,032)

Tableau 3b. Tests de spécification du modèle dans les sorties d'un exemple de régression logistique, sur le logiciel SPSS

Tests de spécification du modèle				
		Khi-Chi-deux	ddl	Sig.
Etape 1	Etape	4,600	1	,032
	Bloc	4,600	1	,032
	Modèle	4,600	1	,032

ddl : degré de liberté
Sig : degré de signification

Le test de Wald s'applique à la variable, ici p=0,027. La variable «transfusion» est associée de façon statistiquement significative et indépendante à la maladie HIV. En plus, le logiciel pondère cette association par l'OR (Exp B) avec son Intervalle de Confiance à 95% soit 2,212 (1,094 - 4,471).

Tableau 4. Test de Wald dans un modèle de régression logistique, sur le logiciel SPSS

	A	E.S.	Wald	d.d.l	Sig	Exp(B)	IC 95%
Transfusion	0,794	0,359	4,888	1	0,027	2,212	1,094
constante	-1,151	0,087	175,978	1	0,000	0,316	4,471

A : Constante du modèle
ES : Erreur Standard,
ddl : degré de liberté,
Sig: degré de signification statistique,
Exp(B): Exponentiel : *Odds Ratio*$_a$,
IC: Intervalle de Confiance à 95% autour de l'*OR*$_a$

Ajoutons maintenant la variable «scarification»; les nouveaux résultats montrent que celle-ci n'apportent pas d'amélioration au modèle. La différence entre les deux déviances est non significative (p=0,076) et le test de Wald confirme ce résultat ; le p (0,508) associé à la variable scarification est non significatif. Ainsi en introduisant une à une les variables et avec la même procédure on retiendra le modèle le plus approprié. Sa qualité sera testée par le test de Hosmer Lemshow au seuil de 0,1.

Tableau 5. Sorties SPSS du modèle de régression logistique, sur le logiciel SPSS

Récapitulatif des modèles			
Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	842,946[a]	,007	,010

Tests de spécification du modèle				
		Khi-Chi-deux	ddl	Sig.
Etape 1	Etape	5,148	2	,076
	Bloc	5,148	2	,076
	Modèle	5,148	2	,076

Variables dans l'équation							
		A	E.S.	Wald	ddl	Sig.	Exp(B)
Etape	Transfusion	,806	,359	5,034	1	,025	2,239
	Scarifcation	-,306	,463	,438	1	,508	,736
	Constante	-,866	,455	3,619	1	,057	,421

CONCLUSION

La régression logistique est une technique d'analyse statistique multi variée permettant d'identifier les facteurs explicatifs ou prédictifs d'un phénomène de santé, en contrôlant les variables de confusion associes à ce phénomène. Elle est spécifique aux variables dépendantes, qualificatives et dichotomiques, sans interférence avec le temps. D'autres techniques d'analyse multi variées sont indiquées pour les variables dépendantes, quantitatives (régression multiple) ou liées au temps (modèle de Cox). Des nouvelles fiches méthodologiques présenteront le mode d'emploi de ces deux approches épidémiologiques et statistiques.

Réponses aux quizz

1. Vrai

2. Vrai

3. Vrai

L'essentiel à retenir

- La régression logistique est une technique d'analyse multi variée permettant le contrôle des variables de confusion et l'identification des facteurs indépendamment associés à la variable à expliquer ou à prédire
- Dans une régression logistique, la variable dépendante (évènement à expliquer ou à prédire) doit être codé en 0/1 (0 si l'évènement est absent, 1 si évènement présent), de même pour les variables qualificatives indépendantes (explicatives) de préférence.
- La régression logistique par la méthode descendante (pas à pas) consiste à intégrer toutes les variables retenues puis à soustraire une à une du modèle, la variable la moins significative.
- Dans une étude multi variée par régression logistique, les *Odds Ratio* calculés sont des OR ajustés en fonction des autres variables incluses dans le modèle (facteurs de confusion)
- La validité du modèle construit par régression logistique est appréciée par le test de Hosmer Lemshow. Plus les valeurs prédites seront proches des valeurs observées, plus ce modèle est adéquat

Pour en savoir plus

1. Guesmi F, Zoghlami A, Sghaier D, Nouira R, Dziri C. Les facteurs alimentaires prédisposant au risque de cancers colorectaux: étude épidémiologique prospective. Tunis Med 2010; 88(3):184-9.

2. Bouyer J. La régression logistique en épidémiologie. Partie II. Rev Epidemiol Sante Publique 1991; 39(2):183-96.