RESCUT : un modèle prédictif

1. Principe.

On dispose d'un échantillon avec :

- une variable à expliquer (dépendante) numérique,
- des variables explicatives (indépendantes) booléennes.

Il s'agit de construire un modèle réalisant cette reconstitution avec une bonne approximation. La méthode RESCUT est une méthode originale, différente des méthodes de régression habituelles, avec les caractéristiques suivantes :

- **grande simplicité d'usage**
- **autorisation de pondération individuelle**
- **indifférence quant aux données manquantes** : la méthode fonctionne avec les données effectivement présentes par individu.

Son principe est le suivant : on ne cherche pas à minimiser une fonction d'écart entre la variable observée et la variable prédite, mais on cherche à minimiser les écarts entre les moyennes de ces deux variables dans les n cellules correspondant aux n variables indépendantes: par exemple, si la première variable indépendante correspond aux hommes, on cherchera à minimiser pour les hommes l'écart entre les moyennes des variables observée et prédite, et de même pour toutes les variables indépendantes.

Ainsi posé, le problème est analogue aux problèmes de redressement d'échantillon, c'est pour cela que la méthode RESCUT s'inspire d'une des méthodes utilisées pour ce genre de problèmes.
La méthode s'étend facilement au **cas plus général** suivant :

- **la variable à expliquer** peut être de type **numérique ou nominal**, y compris multiple.
- **les variables explicatives** peuvent être **nominales, simples ou multiples, ou texte**, dans ces 3 cas avec une liste de modalités associée. Les non réponses sont admises.

2. Méthode.

2.1 Principe.

Soit Y la variable dépendante, et (x_1 ,….,x_n) les variables indépendantes. (x_j =0 ou 1).
La méthode consiste à rechercher des « utilités » U_j , (coefficients analogues à des coefficients de régression) à associer à chaque variable :

Pour chaque individu i, la valeur prédite $\mathbf{Z_i}$ sera :

$$\mathbf{Z_i = \Sigma_j(U_j\ x_{ij})/\Sigma\ x_{ij}} \qquad (1)$$

On choisit les utilités U_j de manière à minimiser les écarts entre les moyennes de Y et de Z pour toutes les populations telles que $x_j = 1$:
On cherche donc à minimiser les $\boldsymbol{\varepsilon_j}$ de chaque variable j :

$$\mathbf{\Sigma_i(Z_i\ x_{ij})/\Sigma\ x_{ij} = \Sigma_i(Y_i\ x_{ij})/\Sigma\ x_{ij} + \varepsilon_j}$$

La méthode RESCUT, créée dans ce but, conduit à avoir, dans la plupart des cas, des valeurs très faibles pour les $\boldsymbol{\varepsilon_j}$.

2.2 Justification.

- S'il n'a qu'une variable explicative, par exemple SEXE (0=Hommes, 1=Femmes), il est clair que la meilleure prédiction est de prendre pour les hommes, leur moyenne observée, et pour les femmes, leur moyenne observée (on pourrait aussi prendre dans certains cas la médiane, mais cela n'est pas généralisable).

- S'il y a plusieurs variables explicatives, on pourrait être tenté de suivre le même procédé, en appliquant à chaque individu la moyenne de la cellule élémentaire à laquelle il appartient, définie par la configuration de son vecteur (x_1 ,….,x_n) ; mais cela est en fait inapplicable, le nombre de telles cellules étant très grand (2^n), il y a énormément de cellules vides, empêchant toute prédiction, et d'autres avec très peu de représentants, avec donc des résultats peu significatifs.

- Il faut donc se contenter de considérer les moyennes dans les marges, c'est-à-dire indépendamment pour chaque variable.

2.3 Algorithme.

Soit y la variable dépendante, et (x_1 ,….,x_n) les variables indépendantes. (x_j =0 ou 1).
Le processus est itératif, il conduit à rechercher des « utilités », (coefficients analogues à des coefficients de régression) à associer à chaque variable :

1) on calcule d'abord la moyenne de y pour chaque variable j, la moyenne étant calculée sur tous les individus i ayant $x_j = 1$

$$\mathbf{Y_j = \Sigma_i (y_i\ x_{ij}) / \Sigma\ x_{ij}}$$

2) on initialise les coefficients recherché U_j à Y_j

3) on calcule pour chaque individu i la valeur prédite z_i ainsi :

$$\mathbf{z_i = \Sigma_j (U_j\ x_{ij}) / \Sigma\ x_{ij}}$$

4) on calcule la moyenne de z pour chaque variable j :

$$\mathbf{Z_j = \Sigma_i\ (z_i\ x_{ij}) / \Sigma\ x_{ij}}$$

5) on remplace U_j par : $\mathbf{U_j . Y_j / Z_j}$

6) on itère à partir de 3) : le processus converge rapidement.

7) le résultat est donc constitué par les utilités $\mathbf{U_j}$ qu'on doit associer à chaque modalité x_j à l'aide de la formule 3).

8) Pour obtenir la valeur de Y pour un nouvel individu, on prend la moyenne des utilités $\mathbf{U_j}$ correspondant aux modalités $x_j = 1$, ce qui est équivalent à une équation de régression.

2.4 Remarques.

1) Les utilités U_j ainsi calculées ont vocation à représenter l'influence propre de la variable j, alors que les moyennes Y_j calculées initialement sont entachées de l'influence des autres modalités, compte-tenu de l'hétérogénéité de la distribution des variables.
2) L'équation (1) ainsi obtenue ne comprend ni terme constant, ni coefficients négatifs.
3) Contrairement à une équation de régression, Z_i est calculée comme la moyenne des U_j et non comme leur somme, cela est nécessaire car **le nombre de variables telles que $x_j = 1$ n'est pas constant d'un individu à l'autre** (il peut y avoir des non réponses dans les variables originelles par exemple).
4) Contrairement à une équation de régression, la méthode ne cherche pas à minimiser les **écarts individuels** entre valeurs observées et calculées, mais les écarts **entre les moyennes** observées et calculées dans chaque population telle que x_j =1. On ne cherche donc pas tant une reconstitution individuelle qu'une **reconstitution statistique de la variable dépendante dans ses effets** au niveau des variables indépendantes. Cela ne doit pas être choquant, dans la mesure où l'on considère la valeur de la variable dépendante non pas comme une valeur certaine, mais comme une **probabilité**.
5) **Si les coefficients d'utilité sont nuls** pour certaines modalités des variables explicatives, (alors que ces modalités sont non vides), le modèle ne permet pas alors d'ajuster correctement les individus correspondant à ces modalités, ce qui entraîne que la moyenne calculée pour ces modalités sera plus forte que la moyenne observée, ce sera également le cas pour la moyenne générale.
6) Si un individu n'a aucune variable avec une valeur 1, il n'y a pas de valeur résultat pour la variable à expliquer, ce qui est normal.
7) S'il n'y a qu'une variable explicative, le résultat obtenu est manifestement le seul résultat possible.
8) Si on soupçonne une influence non négligeable des **interactions entre 2 variables** indépendantes, rien n'empêche de les remplacer par une troisième variable, construite comme le croisement des 2 premières.

2.5 Extension au cas où la variable dépendante est booléenne.

Ayant obtenu une valeur Z pour un individu, on effectue un tirage aléatoire entre 0 et 1 : si le résultat est plus petit que Z, on affecte la valeur 1, sinon on prend 0 comme résultat.

2.6 Extension au cas où la variable dépendante est nominale ou multiple.

On fait tourner le modèle sur toutes les modalités de ces variables, on obtient ainsi des probabilités pour chacune d'elles. Des tirages aléatoires effectués en fonction de ces probabilités permettent de trouver les valeurs de variables nominales ou multiples.

2.7 Extension au cas où les variables indépendantes sont nominales.

On utilise la forme disjonctive, c'est à dire qu'on remplace chaque variable indépendante nominale en autant de variables booléennes qu'elle a de modalités(que cette variable indépendante soit simple ou multiple).

2.8 Cas général. Le cas général auquel s'applique RESCUT est donc le suivant:

- **la variable à expliquer** peut être de type numérique ou nominal, y compris multiple.
- **les variables explicatives** doivent être nominales, simples ou multiples, ou texte, dans ces 3 cas avec une liste de modalités associée. Les non réponses sont admises.

3. Applications.

3.1 **Problèmes de scoring :** la méthode RESCUT a été appliquée avec succès à de nombreux problèmes de scoring :
- étude de l'absentéisme dans une grande administration
- étude de l'efficacité d'une politique de lutte contre le chômage
- étude du risque de défaillance dans le remboursement d'un crédit
- étude de risque de désabonnement pour un opérateur téléphonique

3.2 **Données manquantes** : RESCUT donne de bien meilleurs résultats que les méthodes consistant à prendre une valeur moyenne, ou une valeur au hasard, etc. Sa facilité de mise en œuvre est de plus un atout décisif.

3.3 **Fusion de données :** afin d'enrichir un fichier client par exemple avec les données recueillies par enquête sur un échantillon.

3.4 **Simulations :** lorsqu'on a une enquête fournissant la note de satisfaction globale pour un produit ou un service, et les notes pour plusieurs items (rapidité, qualité, accueil, etc...), on peut utiliser RESCUT pour reconstruire la note globale à partir des notes partielles, puis opérer des simulations pour trouver l'évolution de la note globale en fonction de l'évolution contrôlée des notes partielles successivement : on dispose alors d'un outil indiquant les points à améliorer et le résultat attendu.

4. Comparaison avec d'autres méthodes.

La méthode RESCUT a été appliquée à une étude de satisfaction « points de vente », conjointement à 2 autres méthodes. Cette étude comportait :
- une note d'appréciation globale (0 ou 1)
- des notes d'appréciation portant sur onze critères.

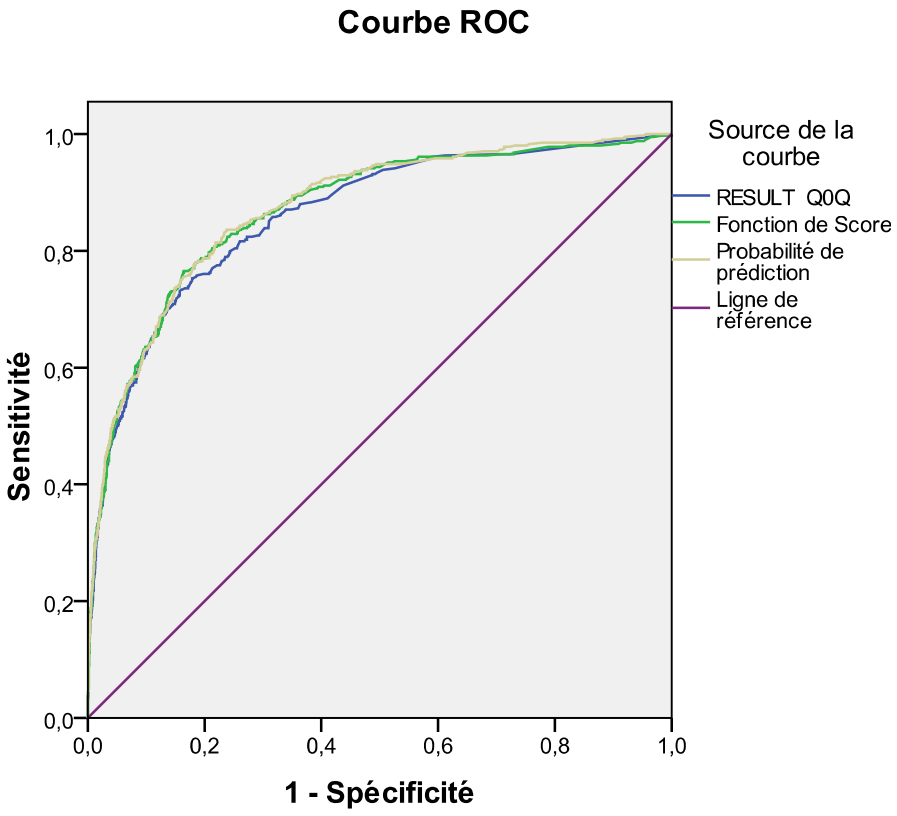
On avait une étude d'apprentissage pour construire le modèle, et une étude de test pour l'appliquer (environ 4000 individus dans les 2 cas). Les méthodes concurrentes étaient :

- une **régression logistique**
- la **méthode DISQUAL** (analyse discriminante sur coordonnées factorielles).

Les résultats obtenus par RESCUT ont été très voisins de ceux obtenus par ces deux autres méthodes.

Si on compare les courbes ROC: les AUC sont presque les mêmes . Cela pour l'échantillon global: la séparation apprentissage test donne à peu près la même chose. Il n'y a pas de différence significative entre les intervalles de confiance.

Comparaison avec Disqual et régression logistique



Les segments diagonaux sont générés par des liaisons.

AUC

Variable(s) de résultats tests	Zone
RESCUT	,861
Score Disqual	,869
Probas régression logistique	,873