École Nationale d'Ingénieurs de Tunis

Gestion de Risque de Crédits

Élève :
Mayara Elatrach

Enseignant :
Mhamed Gaiji

10 janvier 2020

Table des matières

Introduction **2**

1 Présentation de la base de données **2**

2 La modélisation **2**
2.1 Le choix du modèle 2
2.2 Algorithme de Newton 4
2.3 Test de Wald 4
2.4 Tests artistiques 5

3 Ajustement de la base de données **5**

4 Répartition logistique **6**

5 Implémentation du modèle **7**

6 Conclusion **8**

Problématique du TP et la Régression Logistique

Introduction

Notre problème est de modéliser le scoring crédit, pour pouvoir donner des crédits, par une régression non linéaire. On va estimer les paramètres de cette régression en maximisant la vraisemblance. Par la suite on va choisir les meilleures variables explicatives. Et pour valider le modèle on applique plusieurs tests (pseudo R^2, Pvalue et ratio de vraisemblance).

1 Présentation de la base de données

La base de données qu'on a travaillée avec est composée de 4 000 échantillons et cinq variables.
Les échantillons sont des entreprises et les variables sont des ratios décisionnels qui sont :

R1 $= \frac{Fondderoulement}{Actiftotal}$

R2 $= \frac{Bnficenondistribu}{ActifTotal}$

R3 $= \frac{Bienavantimptsetchargesfinanciers}{Actiftotal}$

R4 $= \frac{Capitalisationbancaire}{PassifTotal}$

R5 $= \frac{Ventes}{ActifTotal}$

2 La modélisation

Le score recherché résume l'information disponible, sur le client, dans des facteurs qui affectent la probabilité de défaut. Ce score est donné par l'équation suivante :

$$score_i = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 \quad (1)$$

avec :
i est le numéro du client
$(X_1, X_2, X_3, X_4, X_5)$: facteurs explicatives
$(b_0, b_1, b_2, b_3, b_4, b_5)$: paramètres à estimer qui sont indépendants de i

2.1 Le choix du modèle

Pour obtenir ce score nous avons utilisé la régression logistique.
La régression logistique est un modèle multivarié qui permet d'expliquer, sous forme de probabilité, la relation entre une variable qualitative dépendante Y qui est binaire Y=0 ou 1, et une ou plusieurs variables indépendantes X_i qui peuvent être quantitatives ou qualitatives.
Le modèle fournit la probabilité qu'un événement se produise ou non et les variables indépendantes X_i sont celles susceptibles d'influencer la survenue ou non de l'événement.

La fonction logistique est donnée par l'équation 2.2 et de densité f_x donnée par 2.3.

$$F(x) = \frac{1}{1 + exp(-x)} \tag{2}$$

$$f_x = \frac{exp(-x)}{(1 + exp(-x))^2} \tag{3}$$

avec E(x) = 0 et V(x) = $\frac{pi^2}{3}$.
Pour estimer le vecteur b des coefficients on a utilisé le principe de vraisemblance $f_x(x, \theta)$.

$$L(X_1, X_2, X_3, X_4, X_5, \theta) = \prod_1^n f_x(x_i, \theta) \tag{4}$$

Supposant que P_i est la probabilité de défaut pour le client i et (1-P_i) est la probabilité de non défaut et cette probabilité est donnée par la fonction 2.5.

$$P_i = F(b^T x_i) \tag{5}$$

La vraisemblance suit la loi de Bernouilli et devient égale à :

$$L(b) = \prod_1^n L_i(b) = \prod_1^n P_i^{y_i}(1 - P_i)^{1-y_i} \tag{6}$$

$y_i = \begin{cases} 1 & \text{si le client fait défaut} \\ 0 & \text{sinon} \end{cases}$

Ainsi on applique le ln et on obtient :

$$\ln(L(b)) = \sum_{i=0}^{n} [y_i \ln(F(b^T X_i)) + (1 - y_i) \ln(1 - F(b^T X_i))] \tag{7}$$

On passe maintenant à la dérivation de la fonction $\ln(L(b))$:

$$\frac{\partial(\ln(L(b)))}{\partial t_j} = \sum_{i=1}^{n} y_i \frac{F'(b^T X_i)}{F(b^T X_i)} X_{ij} - (1 - y_i) \frac{F'(b^T X_i)}{1 - F(b^T X_i)} X_{ij} \tag{8}$$

$Or \begin{cases} \frac{F'}{F} = & \text{1-F} \\ \frac{F'}{1-F} = & \text{F} \end{cases}$

L'équation 1.8 devient alors :

$$\frac{\partial(\ln(L(b)))}{\partial t_j} = \sum_{i=1}^{n} y_i(1 - F(b^T X_i) X_{ij} - (1 - y_i) F(b^T X_i) X_{ij} \tag{9}$$

$$\begin{aligned} \nabla(\ln(b)) &= \sum_{i=1}^{n} y_i(1 - F(b^T X_i) X_i - (1 - y_i) F(b^T X_i) X_i \\ &= \sum_{i=1}^{n} (y_i - F(b^T X_i)) X_i \end{aligned} \tag{10}$$

2.2 Algorithme de Newton

$$\begin{cases} x_0 &= donnee \\ x_{n+1} &= x_n - \dfrac{f(x_n)}{f'(x_n)} \end{cases}$$

D'après Taylor on a

$f(x) \simeq f(x_0) - f'(x_0) * (x - x_0)$

$x = x_0 - \dfrac{f(x_0)}{f'(x_0)}$

Algorithme de Newton pour résoudre
$\bigtriangledown_b ln(L(b)) = 0_{\Re^6}$
Soit $b^{(0)}$ donnée
$b^{(n+1)} = b^{(n)} - (\dfrac{\partial^2 ln(L(b^{(n)})}{\partial b \partial b'})^{-1} * \bigtriangledown_b ln(L(b^{(n)}))$
$(\dfrac{\partial^2 ln(L(b^{(n)})}{\partial b \partial b'})$ est inversible car la Hessienne est définie positive.
Critère d'arrêt :
$\left\| b^{(n+1)} - b^{(n)} \right\|_2 < \xi$

2.3 Test de Wald

$t_j = \dfrac{b_j - \tilde{b_j}}{SE(b_j)} = \dfrac{b_j}{SE(b_j)}$
$t_j \sim Student(N - K)$
$N \simeq 4000$
$t_j \sim \mathcal{N}(0, 1)$
Si $p - value < 0.01$ On rejette l'hypothèse de nulleté donc R_j est significatif.

$p - value = 2 * (1 - F_{\mathcal{N}(0,1}(|t_j|))$
$SE(b_j) = \sqrt{(Var(\dfrac{\partial lnL(b)}{\partial b_j}))^{-1}}$
Le meilleur estimateur est celui qui a la variance la plus petite. Il est efficace
$V(T_n) = (I_n(\theta))^{-1}$ donc meilleur estimateur.
$I_X(\theta) = -E(\dfrac{\partial^2 lnf_X(x,\theta)}{\partial \Theta_i \partial \Theta_j})_{1 \leq i,j \leq n}$
$= E(\dfrac{\partial lnf_\Theta(\theta,x)}{\partial \Theta_i} * \dfrac{\partial lnf_\Theta(\theta,x)}{\partial \Theta_j})_{1 \leq i,j \leq n}$
$SE(b_{ij}) = \sqrt{(Var(\dfrac{\partial lnL(b)}{\partial b_j}))^{-1}}$
$= \sqrt{-E(\dfrac{\partial^2 lnL(b)}{\partial b_i \partial b_j})_{1 \leq i,j \leq n}}$

2.4 Tests artistiques

Notre modèle est un modèle de régression non linéaire, donc on utilise les tests artistiques de pseudo R et de rapport de vraisemblance (LR : Likelyhood Ratio).
• Pseudo R^2
Pseudo $R^2 = 1 - log(L)/log(L_0)$

Où

$$\begin{cases} L & \text{vraisemblance de modèle à 6 paramètres.} \\ L_0 & \text{vraisemblance de modèle contenant un seul paramètre } b_0. \end{cases} \tag{11}$$

Plus Pseudo R^2 est proche de 1, plus on dit le modèle à 6 paramètres est bien ajusté.
Si les valeurs prédictives R_i augmentent les pouvoirs explicatifs du modèle, la valeur du log de vraisemblance $log(L)$ sera proche de 0 $\text{Ln}(\text{L}_0) = ? \Rightarrow log(L_0) = ?$
• Test de rapport de vraisemblance LR
Il permet de tester l'hypothèse de nullité des coefficients b_i c.à.d. permet de tester l'hypothèse que toute les variables explicatives jointes n'ajoutent pas un pouvoir au modèle.
$LR = 2(log(L) - log(L_0)) = 2log(L/L_0)$
Plus LR est grande plus on est sur de rejeter l'hypothèse de nullité des paramètres.

$$LR \sim X^2(n) \tag{12}$$

n : nombre de restrictions imposées Dans le cas L_5 et L_0
$N = 6 - 1 = 5$

3 Ajustement de la base de données

Avant de travailler avec les données, on va les ajuster à l'aide de la technique de Winsorisation pour éliminer les valeurs abérrantes.
La technique de Winsorisation consiste à remplacer les valeurs extremes par des valeurs plus modérés.
En effet, pour un niveau α on remplacr la valeur de la distribution au dessous du centile α (respectivement au dessus du centile $1 - \alpha$.
Le niveau de winsorisation peut etre fixé séparément par chaque indicateur.
En comparant les sommaires des données avant et après winsorisation on obtient les deux tableaux suivants.

	WC/TA	RE/TA	EBIT/TA	ME/TL	S/TA
Moyenne	0,143	0,210	0,052	1,954	0,304
Mediane	0,117	0,219	0,052	1,136	0,261
Ecart type	0,12759815	0,20812097	0,01819561	1,51271926	0,13845368
Coef Assymetrie	-1,01453171	-2,55489151	-4,84436164	7,75072812	4,48100441
Kurtosis	17,6822686	17,441196	85,9974266	103,127537	71,2152414
Min	-2	-3	-1	0	0
Max	1	2	0	61	5
Quantile 0,5	-0,33405335	-1,74031666	-0,04702038	0,05222764	0,06149793
Quantile 1	-0,1730141	-0,941112	-0,0165243	0,0740331	0,070169
Quantile 1,5	-0,13156222	-0,73749912	-0,00787069	0,0904338	0,07624661
Quantile 95	0,44170254	0,65087794	0,09216085	5,6051545	0,68152365
Quantile 99	0,57714753	0,89730099	0,12106126	14,8215476	1,05514503
Quantile 99,5	0,63841571	0,94146514	0,13537314	19,0897394	1,13045844

FIGURE 1 – Sommaire des données avant Winsorisation

	WC/TA	RE/TA	EBIT/TA	ME/TL	S/TA
Moyenne	0,144	0,217	0,052	1,871	0,302
Mediane	0,117	0,219	0,052	1,136	0,261
Ecart type	0,12471559	0,19860558	0,01722089	1,39746654	0,13595661
Coef Assymetrie	0,03183932	-0,9525027	0,1350102	3,29766685	1,68154545
Kurtosis	5,54585305	3,2023598	1,0988754	13,4786476	3,42269499
Min	-2	-1	0	0	0
Max	1	1	0	14	1
Quantile 0,5	-0,15395065	-0,92362929	-0,01620374	0,0784782	0,07030681
Quantile 1	-0,14400926	-0,92362929	-0,01620374	0,0784782	0,07030681
Quantile 1,5	-0,13156222	-0,73749912	-0,00787069	0,0904338	0,07624661
Quantile 95	0,43964772	0,65087794	0,09216085	5,6051545	0,68152365
Quantile 99	0,56111459	0,89680162	0,12060919	14,4445781	1,05376212
Quantile 99,5	0,57699606	0,89680162	0,12060919	14,4445781	1,05376212

FIGURE 2 – Sommaire des données après Winsorisation

D'après ces deux tableaux on peut constater que les valeurs du coefficient d'asymétrie et du kurtosis ont diminués significativement. Les valeurs sont donc plus homogènes.

4 Répartition logistique

La fonction de répartition logistique est une fonction continue bornée entre 0 et 1. Elle peut représenter une probabilité.
Elle devient presque linéaire à partir d'un certain score, et puis elle se stabilise pour les grands scores (à partir de ce seuil le décideur n'a pas intérêt à étudier ce cas, il est clair que ce client n'est pas capable de rembourser le crédit) de meme pour les très petit scores : il est presque sur que le client est capable de rembourser le crédit.

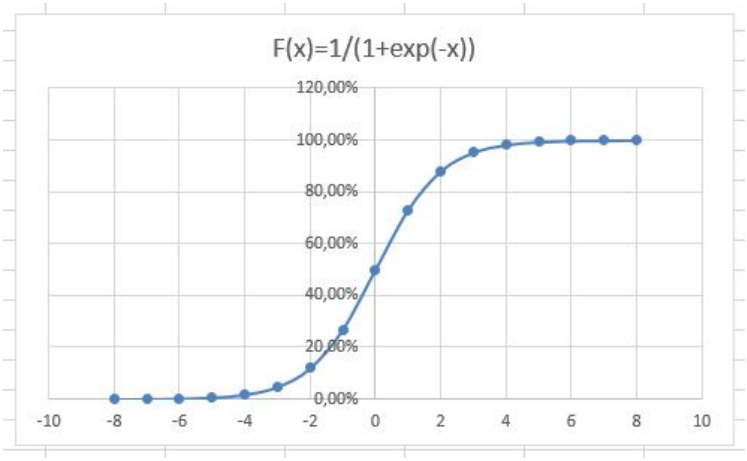

FIGURE 3 – Fonction de répartition logistique

5 Implémentation du modèle

Pour implémenter le modèle, on applique un code d'une fonction logit qui calcule les scores. Pour faire on doit :
Initialiser $b_0 := (1, 1, 1, 1, 1, 1)$ pour supposer que tous les variables sont pertinentes.
Faire un boucle dans laquelle on calcule la différentielle de la logarithmique de b, la matrice Hessienne pour calculer le nouveaux b. La boucle s'arrete au critère d'arret.
Si on applique cet algorithme aux données winsorisés on va obtenir b. Ensuite en applique les tests d'hypothèses en on abtient ce tableaux :

	CONST	WC/TA	RE/TA	EBIT/TA	ME/TL	S/TA
b	-2,40372243	0,51746796	-3,16803734	-24,4805473	-1,07513209	1,33745995
SE(b)	0,33426951	0,85397166	0,41065882	6,26641527	0,29166155	0,61016817
t	-7,19097123	0,60595449	-7,71452397	-3,90662704	-3,68623185	2,19195301
p-value	6,4326E-13	0,54454497	1,2212E-14	9,3593E-05	0,0002276	0,0283829
Pseudo R^2 / # iter	0,25451194	10	#N/A	#N/A	#N/A	#N/A
LR-test / p-value	183,553853	9,3182E-38	#N/A	#N/A	#N/A	#N/A
lnL / lnL_0	-268,822764	-360,59969	#N/A	#N/A	#N/A	#N/A

FIGURE 4 – Résultat de la fonction logit

Elimination des variables non significatives en se basant sur la p-value (entre modele 1 et 2 , on néglige ceux avec p-value>0.01)
Donc on doit éliminer les deux variables WC/TA et S/TA car ils ont des p-values qui sont respectivement 0.54 et 0.02.

Après élimination des variables on obtient enfin le deuxième modèle :

		CONST	RE/TA	EBIT/TA	ME/TL
	b	-2.03701615	-3.29088173	-21.2591582	-1.10707021
	SE(b)	0.29245482	0.41040077	6.11638749	0.29008518
	t	-6.96523376	-8.01870266	-3.47577034	-3.81636254
	p-value	3.2785E-12	1.1102E-15	0.00050939	0.00013543
	Pseudo R^2 / # iter	0.24693845	10	#N/A	#N/A
	LR-test / p-value	178.091856	2.2779E-38	#N/A	#N/A
	lnL / lnL_0	-271.553762	-360.59969	#N/A	#N/A

FIGURE 5 – Résultat du deuxième modèle

Toutes les variables ont des p-values qui sont inféreires à 0.01 donc le deuxième modèle est notre modèle final.

6 Conclusion

Au cours de ce TP nous avons introduit le modèle de scoring. Nous avons ajusté nos données avec la technique de winsorisation. Nous avons calculés des paramètres du modèle à l'aide de la fonction logit. Nous avons enfin appliqué des tests statistiques pour fixer le modèle final.