ESGFINANCE

APPLICATIONS FINANCIERES SOUS EXCEL VBA

4ESGF OPTION BIG DATA ET DATASCIENCE EN FINANCE

ERIC DUCROS

PROJET

APPLICATION DE LA REGRESSION LOGISTIQUE A LA MODELISATION DE LA PROBABILITE DE DEFAUT DES ENTREPRISES

TABLE DES MATIERES

APPLICATION DE LA REGRESSION LOGISTIQUE A LA MODELISATION DE LA PROBABILITE DE DEFAUT DES ENTREPRISES

I. PRESENTATION DE LA PROBLEMATIQUE

1) Le défaut

Le risque de crédit est un des sujets les plus étudiés en finance quantitative. Ce sujet a pris une importance décisive après la crise des *subprimes* Une entreprise fait défaut lorsqu'elle est incapable d'honorer le remboursement de tout ou partie de sa dette. Il est donc primordial pour toute établissement de crédit d'estimer la probabilité qu'une entreprise aura de faire défaut afin de :

- décider d'octroyer ou non le prêt demandé par l'entreprise ;
- si le prêt est octroyé alors le taux d'intérêt qui sera facturé par la banque sera une fonction croissante de la probabilité de défaut.

Le processus par lequel sont déterminés les facteurs permettant de quantifier la probabilité de défaut d'une entreprise se nomme la notation de crédit (le *credit scoring* ou *credit rating* en anglais). Les principales agences de notation comme Standard & Poor's, Moody's et Fitch ont pour objectif d'évaluer la solvabilité (capacité à rembourser un emprunt) d'une entreprise. Un des éléments de cette évaluation consistera à évaluer la probabilité de défaut d'une entreprise qui permettra avec d'autres éléments qualitatifs ou quantitatifs de décerner une note à l'entreprise. Cette note aura un impact décisif sur le coût de son endettement, plus la note est haute, plus les taux facturés seront bas et inversement.

Modéliser la probabilité de défaut est donc un élément incontournable en finance de marché. La question qui se pose alors est de déterminer quels sont les facteurs permettant d'expliquer cette probabilité de défaut.

2) Les facteurs potentiellement explicatifs de la probabilité de défaut

Ces facteurs ont jusqu'ici été recherchés dans les ratios issus de l'analyse financière classique. A partir du bilan et du compte de résultat de l'entreprise il est possible de calculer plusieurs ratios susceptibles de donner un éclairage sur la santé financière de l'entreprise et plus précisément sur sa solvabilité.

Altman (1968) fût l'un des pionniers de la notation de crédit même s'il cherchait initialement à prédire la probabilité de faillite et non le simple défaut de remboursement d'une échéance. En utilisant une analyse discriminante multiple (méthode très proche de la régression linéaire multiple) appliquée à un échantillon de sociétés ayant fait faillite ou pas[1], il a estimé la fonction Z-score suivante :

$$Z = 0{,}012X_1 + 0{,}014X_2 + 0{,}033X_3 + 0{,}006X_4 + 0{,}999X_5 \quad (1)$$

Avec :

- X_1 = Besoin en Fonds de Roulement d'exploitation / total de l'actif ;
- X_2 = Réserves / total de l'actif ;
- X_3 = Résultat d'exploitation / total de l'actif ;
- X_4 = Capitalisation boursière / valeur comptable de la dette financière ;
- X_5 = Chiffre d'affaires / total de l'actif.

Utilisation de l'équation :
L'application la méthode Z-Score d'Altman sur une entreprise se fait en deux temps :

- Calcul des 5 ratios ;
- Calcul du Z-Score avec l'équation (1) appliquée aux ratios précédemment calculés ;
- Classement de la société en fonction du résultat Z trouvé.

[1] L'estimation de l'équation du Z-score a été conduite sur un échantillon de 66 entreprises américaines sur la période 1946 – 1965. La moitié des entreprises de l'échantillon ont fait faillite durant la période d'étude.

L'équation (1) s'utilise comme celle d'une régression linéaire, il suffit d'insérer dans (1) la valeur des ratios calculés à partir d'un bilan pour obtenir le Z-score. Supposons qu'on ait obtenu les valeurs suivantes pour une société appelée LOGITRON :

- $X_1 = 0{,}35$
- $X_2 = 0{,}12$
- $X_3 = 0{,}45$
- $X_4 = 0{,}33$
- $X_5 = 0{,}5$

Le Z-score de LOGITRON sera donné par :

$$Z = 0{,}012 \times 0{,}35 + 0{,}014 \times 0{,}12 + 0{,}033 \times 0{,}45 + 0{,}006 \times 0{,}33 + 0{,}999 \times 0{,}50 \approx 0{,}5222$$

Une fois le Z-score calculé, Altman fournit une règle d'interprétation résumée dans le tableau ci-dessous :

Z-score	Faillite d'ici deux ans
< 1.81	Oui
> 1.81 et <2.675	Oui avec une forte probabilité
> 2.675 et <2.99	Non avec une forte probabilité
> 2.99	Non

Le Z-score de LOGITRON est largement inférieur à 1.81 impliquant une faillite certaine d'ici deux ans d'après Altman.

Depuis cet article fondateur, la méthodologie a beaucoup évolué. On trouvera une revue de la littérature exhaustive sur les différentes méthodologies chez Balcaen et Ooghe (2006). Après l'analyse discriminante multiple, les chercheurs ont proposé d'utiliser des modèles de régression logistique. Depuis la fin des années 90, certains auteurs ont tenté d'utiliser les réseaux de neurones artificiels ou les arbres de décision. Imtiaze et Brimicombre (2017) présentent une comparaison des performances des différents outils. On observe que la simple régression logistique donne de très bons résultats souvent meilleurs que les réseaux de neurones.

Il est important de noter que quelle que soit la méthode utilisée, les données de base sont toujours les mêmes. Il faut disposer au départ d'un ensemble de ratios d'analyse financière susceptibles d'avoir un effet sur la probabilité de défaut pour un groupe d'entreprise ayant fait défaut et pour un autre groupe d'entreprise n'ayant pas fait défaut. A partir de ces données il sera possible d'entraîner un modèle qui permettra de sélectionner les ratios importants ayant un effet significatif sur la probabilité de défaut et de déterminer s'ils ont un effet positif ou négatif sur cette dernière.

L'objet de ce projet est de montrer comment estimer une équation du type de celle d'Altman mais en utilisant la méthode de la régression logistique. Cette technique permet d'expliquer une variable dépendante binaire à partir de variables indépendantes numériques ou binaires. En *credit scoring* la variable dépendante sera codée 1 si l'entreprise a fait défaut et 0 sinon.

Travail à faire

1. ***A partir des documents de synthèse de la société IASCORE (voir annexe 1 ci-dessous), calculer les ratios d'analyse financière utilisés par Altman dans son étude. On vous précise que la société est cotée en bourse et que la valeur moyenne du titre IASCORE sur le mois de décembre 2017 était de 320 €. L'annexe 2 vous rappelle la définition et le calcul du besoin en fonds de roulement d'exploitation.***

2. ***En utilisant les ratios précédemment calculés et le modèle d'Altman, déterminer le Z-score de la société IASCORE et conclure sur sa probabilité de faire faillite dans les deux ans.***

3. ***Télécharger les données financières d'une dizaine de sociétés cotées sur Bloomberg pour l'année 2018 et calculer les mêmes ratios. Assembler le tout dans une base de données avec une ligne pour chaque***

entreprise. Insérer une colonne dans laquelle sera calculée le Z-score grâce à l'équation d'Altman. Il faudra aussi inclure une colonne avec le code ISIN de chaque société.

4. ***Quelles sont les critiques que l'on peut adresser au modèle du Z-score d'Altman ?***

ANNEXE 1 : documents de synthèse de IASCORE au 31/12/2017

COMPTE DE RESULTAT 31/12/2017			
Charges			**Produits**
Achats MP	5 500.00	Chiffre d'affaires	30 000.00
Charges externes	2 000.00	Autres produits d'exploitation	5 000.00
Salaires	8 000.00		
Impôts et taxes	800.00		
Amortissement et dépréciation	3 000.00		
Charges d'exploitation	**19 300.00**	**Produits d'exploitation**	**35 000.00**
Charges financières	**3 000.00**	**Produits financiers**	**250.00**
Charges exceptionnelles	**500.00**	**Produits exceptionnels**	**1 000.00**
RCAI	**13 450.00**		
Impôt sur les sociétés	3 766.00		
Résultat net	**9 684.00**		
TOTAL CHARGES	**36 250.00**	**TOTAL PRODUITS**	**36 250.00**

BILAN AU 31/12/2017			
Actif			**Passif**
Immobilisation incorporelles	2 000.00	Capital social (1 000 actions)	10 000.00
Immobilisation corporelles	15 000.00	Réserves	4 000.00
Immobilisation financières	800.00	Résultat	9 684.00
Actif immobilisé	**17 800.00**	**Capitaux propres**	**23 684.00**
Stocks	4 000.00	Dettes financières	35 000.00
Clients	8 500.00	Fournisseurs	6 000.00
Autres créances	4 500.00	Autres dettes	2 000.00
Valeur Mobilières de Placement	2000	Dettes fiscales et sociales	116.00
Banque	30000		
TOTAL ACTIF	**66 800.00**	**TOTAL PASSIF**	**66 800.00**

ANNEXE 2 : définition et calcul du BFRE

Le besoin en fonds de roulement d'exploitation (BFRE dans la suite du texte) se calcul de la façon suivante :

BFRE de l'année N = Stock N + Créances d'exploitation N – Dettes d'exploitation N

Le BFR représente le montant des ressources financières que l'entreprise doit avancer pour que le cycle d'exploitation puisse se dérouler normalement. Il s'agit donc d'un besoin d'argent et il est absolument indispensable d'en maîtriser le niveau sous peine d'avoir de graves problèmes de trésorerie.

II. METHODOLOGIE DE LA REGRESSION LOGISTIQUE

1) Régression logistique et solveur Excel

a) Présentation du problème

Afin de présenter la méthode et l'intuition sous-jacente nous allons explorer un cas simple avec une seule variable explicative. Lorsque la variable à expliquer ne peut prendre que deux valeurs, 0 ou 1 et qu'on trace le nuage de points on se retrouve avec le genre de graphique de la figure 1 ci-dessous. On voit tout de suite que la méthode des moindres carrés va avoir beaucoup mois de sens même s'il est toujours possible de l'appliquer.

Vous trouverez dans la feuille « **Logit_simple** » du fichier Excel « **Projet_Credit_Scoring.xlsm** » le même nuage de points tracés à partir de 4 000 observations. Pour chaque observation on a relevé la valeur du ratio résultat d'exploitation divisé par le total de l'actif de la firme (colonne REX/TA) et la valeur de la variable binaire DEFAUT qui est égale à 1 lorsque l'entreprise a fait défaut et 0 sinon. La figure 2 ci-dessous présente un extrait du fichier Excel.

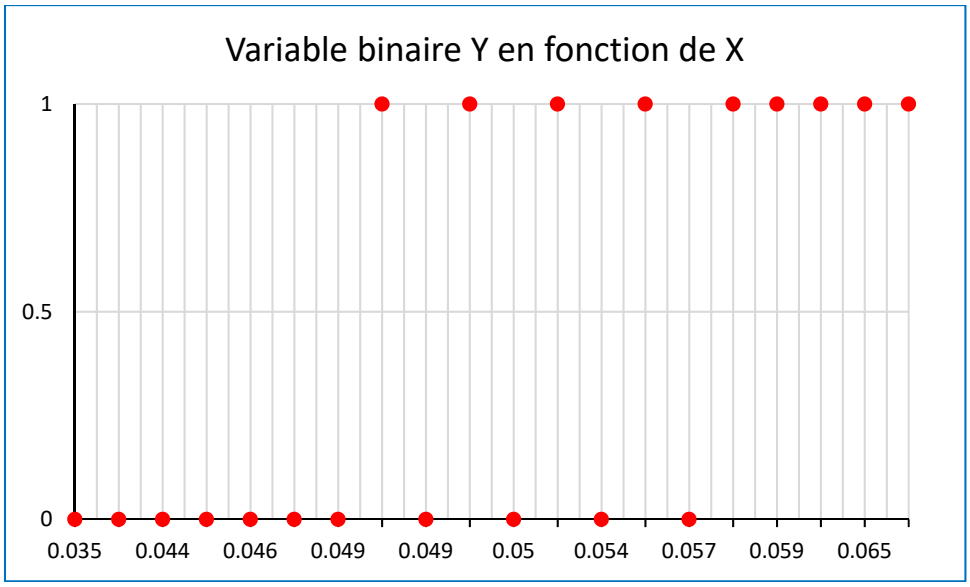
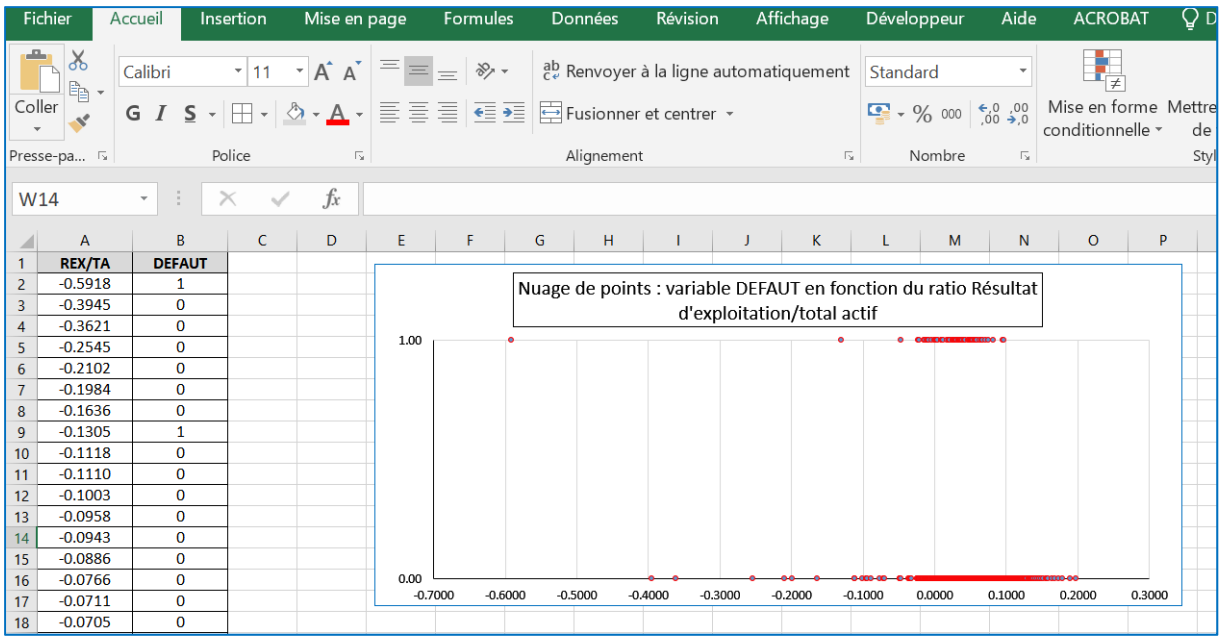


Figure 1



	A	B
1	REX/TA	DEFAUT
2	-0.5918	1
3	-0.3945	0
4	-0.3621	0
5	-0.2545	0
6	-0.2102	0
7	-0.1984	0
8	-0.1636	0
9	-0.1305	1
10	-0.1118	0
11	-0.1110	0
12	-0.1003	0
13	-0.0958	0
14	-0.0943	0
15	-0.0886	0
16	-0.0766	0
17	-0.0711	0
18	-0.0705	0

Figure 2

En revanche, si on cumule les fréquences de la variable DEFAUT sur un découpage en classe de la variable REX/TA on obtient les données suivantes :

Intervalle de valeur	Fréquence	Fréquence cumulée
[- ∞ ; - 0.025[	4.17%	4.17%
[- 0.025 ; 0[	12.50%	16.67%
[0 ; 0.025[	16.67%	33.33%
[0.025 ; 0.05[	40.28%	73.61%
[0.05 ; 0.075[	22.22%	95.83%
[0.075 ; + ∞[	4.17%	100.00%

Tableau 1

Si on trace le nuage de point correspondant aux données du tableau 1 ci-dessus alors on obtient le graphique caractéristique de la figure 3 ci-dessous :

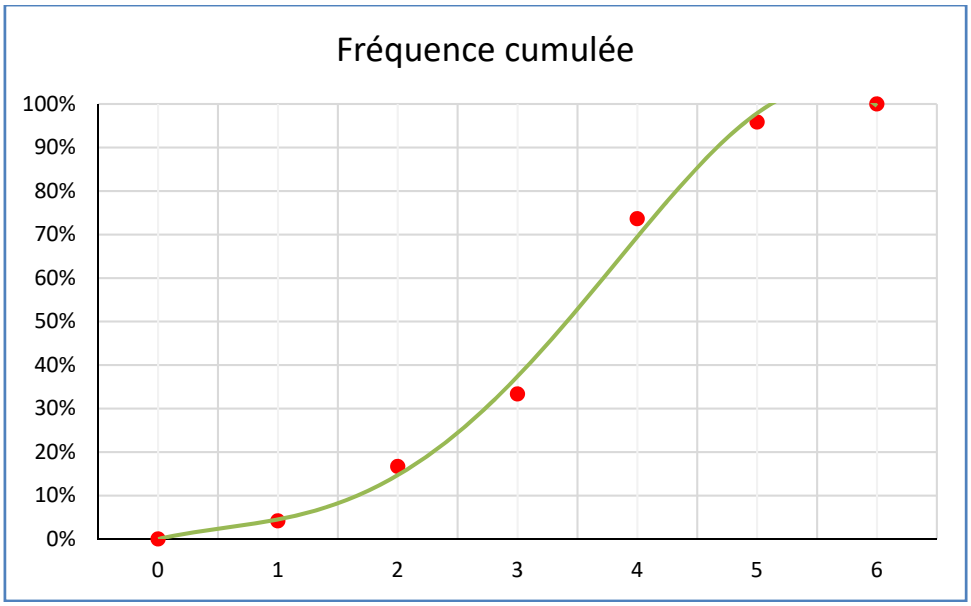


Figure 3

La courbe de la figure 3 ci-dessus montre que plus le ratio REX/TA augmente plus la proportion de défaut augmente. La forme de la courbe est dite sigmoïde (en forme de S). Cette observation nous incite à utiliser la fonction de répartition de la loi logistique pour modéliser les problèmes où la variable dépendante est binaire.

La fonction de répartition de la loi logistique est donnée par l'expression :

$$f(x)=\frac{1}{1+e^{-x}} \qquad (2)$$

Cette expression peut encore s'écrire : $$f(x)=\frac{e^{x}}{1+e^{x}} \qquad (2)'$$

On voit bien sur la figure 4 ci-dessous que la fonction logistique possède une forme tout à fait adéquate pour modéliser le lien entre la fréquence cumulée d'une variable binaire prenant les valeur 0 ou 1 et une variable *x*.

b) Estimation avec le solveur Excel à partir d'un échantillon de 20 entreprises et trois variables explicatives.

Nous voulons étudier l'impact de trois variables explicatives $X = (1, X_1, X_2, X_3)$; avec X_1, X_2 et X_3, trois ratios

d'analyse financière ; sur une variable dépendante *Y* à valeur dans {0,1} modélisant le défaut d'une entreprise (*Y* = 1 si l'entreprise a fait défaut et 0 sinon).

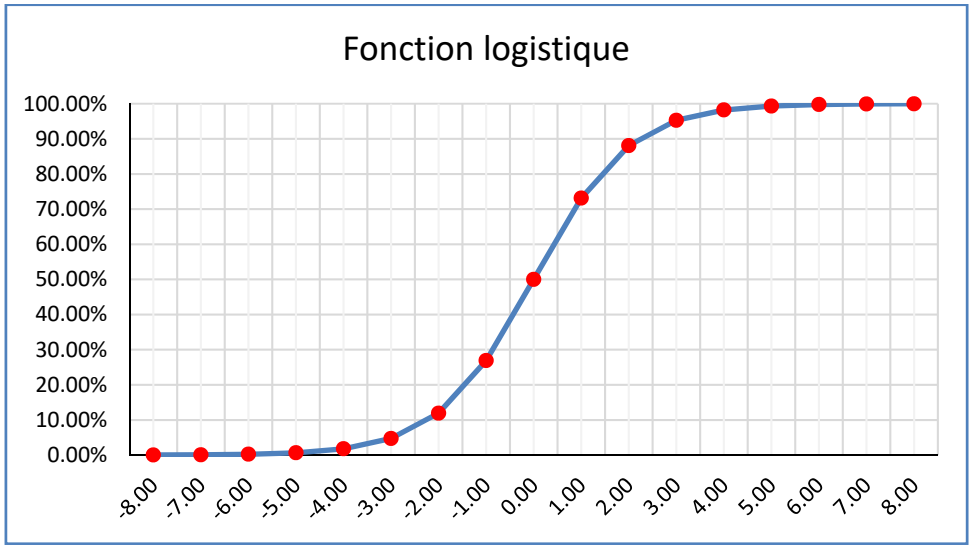


Figure 4

Pour arriver à nos fins nous allons utiliser un échantillon de taille *n*=20 entreprises (onglet « **Estimation_coeff_solveur** » du fichier Excel « **Projet_Credit_Scoring.xlsm** ») :

Entreprise i	Constante	X_1 BFR/TA	X_2 RES/TA	X_3 REX/TA	*Y* DEFAUT
1	1	0.028	3.196	0.284	1
2	1	-0.022	0.138	0.697	1
3	1	0.004	0.085	0.189	1
4	1	-0.131	0.834	0.094	1
5	1	0.022	0.047	0.232	1
6	1	0.043	0.956	0.335	0
7	1	0.052	1.065	0.335	0
8	1	0.027	0.804	0.246	0
9	1	0.030	0.387	0.253	0
10	1	0.032	0.792	0.276	0
11	1	-0.021	0.743	0.143	0
12	1	0.046	1.143	0.207	0
13	1	0.039	1.705	0.192	0
14	1	0.073	1.572	0.164	0
15	1	0.040	3.370	0.086	0
16	1	0.043	3.127	0.116	0
17	1	0.040	0.171	0.165	0
18	1	0.043	0.220	0.154	0
19	1	0.034	0.333	0.142	0
20	1	0.037	0.585	0.159	0

Nous avons ajouté une colonne de 1 pour l'estimation de la constante exactement comme dans le modèle de régression par les moindres carrés.

Nous devons estimer $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ tel que la probabilité de faire défaut $P(Y = 1|X) = p_i$ d'une entreprise *i* soit donnée par l'égalité :

$$p_i = \frac{e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}}}{1+e^{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}}} \quad (3)$$

Pour estimer le vecteur β nous allons utiliser la méthode du maximum de vraisemblance qui consiste à maximiser la fonction de vraisemblance *V* définie sur un échantillon de taille *n* par :

$$V = P(Y = y_1) \times P(Y = y_2) \times \ldots \times P(Y = y_n) \quad (4)$$

avec $P(Y=y_i)$ la loi de probabilité de la variable aléatoire pour l'individu *i*. La fonction de vraisemblance associe à un échantillon la probabilité qu'il se réalise en supposant que les observations sont indépendantes. Remarquons qu'il est souvent plus facile de maximiser le logarithme de la fonction de vraisemblance.

Travail à faire

1. ***Déterminer la loi de probabilité suivie par Y pour une entreprise i.***

2. ***Utiliser le résultat précédent pour démontrer que le logarithme de la fonction de vraisemblance (4) peut s'écrire selon l'expression :***

$$\ln(V) = \sum_{i=1}^{20} \left[y_i . \ln(p_i) + (1 - y_i) . \ln(1 - p_i) \right] \quad (5)$$

3. ***Renseigner les colonnes vides de la feuille « Estimation_coeff_solveur » et calculer la valeur du logarithme de la fonction de vraisemblance dans la cellule I23.***

4. ***Utiliser le solveur Excel avec comme cellule cible I23 et comme variables de d'optimisation le vecteur $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$ en plage L2:L5 avec comme valeurs initiales $\beta = (0, 0, 0, 0)'$. Il faut trouver la solution $\beta = (-4.01, -82.17, 0.32, 19.11)'$.***

5. ***Donner l'expression de l'équation estimée et commenter chacun des coefficients.***

2) Maximisation de la fonction de vraisemblance par la méthode de Newton-Raphson.

La maximisation de la vraisemblance se fait généralement par la méthode de Newton-Raphson. Dans cette partie nous allons présenter cet algorithme à partir d'un exemple simple avec une fonction à une variable. Ensuite, nous appliquerons la méthode sur une fonction de plusieurs variables. Finalement nous verrons comment transposer cette méthode pour maximiser la fonction de vraisemblance de l'équation (5). Cela implique de calculer la dérivée première et la dérivée seconde de la fonction à maximiser. Dans le cas d'une fonction à plusieurs variables nous devrons alors calculer les dérivées partielles premières et secondes par rapport à l'ensemble des variables ce qui nous amènera à définir le gradient (vecteur des dérivées partielles premières par rapport à chaque variable) et la hessienne (matrice des dérivées partielles secondes par rapport à chaque variable) de la fonction de vraisemblance.

a) Méthode de Newton-Raphson pour optimiser une fonction f(x).

La méthode de Newton-Raphson est un algorithme permettant de trouver une solution de l'équation $f(x)=0$. Il consiste à appliquer l'équation de récurrence suivante en partant d'un point $x=x_0$:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (6)$$

où $f'(x)$ désigne la dérivée de *f*.

Le processus itératif est arrêté lorsque $|x_{n+1} - x_n| < \varepsilon_1$ ou encore lorsque $|f(x_{n+1}) - f(x_n)| < \varepsilon_2$ avec ε_1 et ε_2, la précision désirée pour chaque test d'arrêt.

Lorsqu'on veut optimiser une fonction, c'est-à-dire trouver un maximum ou un minimum local, alors on appliquera le schéma suivant en partant d'un point $x=x_0$:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)} \quad (7)$$

où $f'(x)$ et $f''(x)$ désignent respectivement la dérivée première et la dérivée seconde de f. Les tests d'arrêt sont identiques. Nous supposons évidemment que les conditions de premier et second ordre sont satisfaites.

Travail à faire

Soit la fonction suivante définie sur $\mathbb{R}$:

$$f(x) = -x^6 + 2x^5 \quad (8)$$

L'objet de cette question est de déterminer :

$$x^* = \underset{x\in[0;2]}{\arg\max}\, f(x)$$

1. ***Démontrer que l'équation (6) implique l'équation (7) lorsqu'il s'agit de réaliser une optimisation sans contrainte.***

2. ***La figure 5 ci-dessous donne le graphique de la relation (8) :***

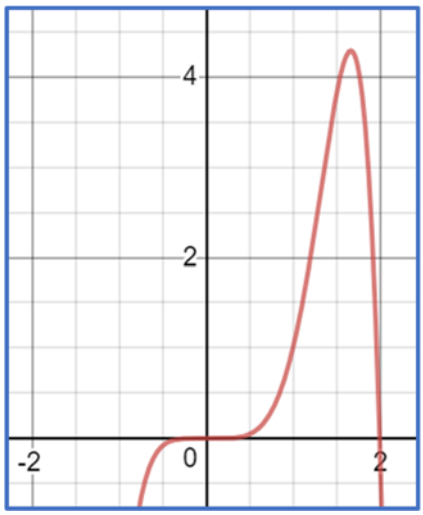


Figure 5

Evaluer graphiquement la valeur de x*.

3. ***Ensuite, rendez-vous dans la feuille « Méthode_Newton_Raphson » du classeur Excel joint à ce projet pour appliquer la méthode de Newton-Raphson afin de déterminer x* avec une précision ε = 0.0001. La mise en œuvre de l'algorithme se fera en renseignant le tableau avec comme valeur initiale x_0 = 4. Une ligne du tableau correspond à une itération. A partir de n=1 la valeur de x_n est déterminée grâce à l'équation (7) appliquée avec les valeurs de f' et f'' calculées dans la ligne précédente. Mettre en vert la première ligne du tableau où le critère de convergence est atteint.***

b) Méthode de Newton-Raphson pour optimiser une fonction de plusieurs variables $f(\beta_1, \beta_2, \ldots, \beta_p)$.

L'algorithme de Newton-Raphson peut aussi s'appliquer à l'optimisation de fonctions de plusieurs variables. L'équations (7) devient dans ce cas :

$$\beta_{n+1} = \beta_n - \frac{\nabla f(\beta_n)}{\nabla^2 f(\beta_n)} \quad (9)$$

avec $\nabla f(\beta_n)$ le vecteur des dérivées partielles premières appelé gradient de f :

$$\nabla f(\beta_n) = \begin{pmatrix} \frac{\partial f(\beta_{n0})}{\partial \beta_{n0}} \\ \frac{\partial f(\beta_{n1})}{\partial \beta_{n1}} \\ \vdots \\ \frac{\partial f(\beta_{np})}{\partial \beta_{np}} \end{pmatrix} \quad (10)$$

et $\nabla^2 f(\beta_n)$ la matrice des dérivées partielles secondes de f appelée hessienne de f :

$$\nabla^2 f(\beta_n) = \begin{pmatrix} \frac{\partial^2 f(\beta_{n0})}{\partial^2 \beta_{n0}} & \frac{\partial^2 f(\beta_{n0})}{\partial \beta_{n0}\, \partial \beta_{n1}} & \cdots & \frac{\partial^2 f(\beta_{n0})}{\partial \beta_{n0}\, \partial \beta_{np}} \\ \frac{\partial^2 f(\beta_{n1})}{\partial \beta_{n1}\, \partial \beta_{n0}} & \frac{\partial^2 f(\beta_{n1})}{\partial^2 \beta_{n1}} & \cdots & \frac{\partial^2 f(\beta_{n1})}{\partial \beta_{n1}\, \partial \beta_{np}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\beta_{np})}{\partial \beta_{np}\, \partial \beta_{n0}} & \frac{\partial^2 f(\beta_{np})}{\partial \beta_{np}\, \partial \beta_{n1}} & \cdots & \frac{\partial^2 f(\beta_{np})}{\partial^2 \beta_{np}} \end{pmatrix} \quad (11)$$

où β_n est le vecteur des variables de f à l'étape n de l'algorithme :

$$\beta_n = \begin{pmatrix} \beta_{n0} \\ \beta_{n1} \\ \vdots \\ \beta_{np} \end{pmatrix}$$

Etant donné que $\nabla f(\beta_n)$ est un vecteur et que $\nabla^2 f(\beta_n)$ est une matrice carrée, la division de l'un par l'autre dans l'équation (9) ne peut se faire directement. Pour réaliser ce calcul il faut passer par l'inverse de la matrice hessienne ce qui amène :

$$\beta_{n+1} = \beta_n - \left[\nabla^2 f(\beta_0)\right]^{-1} \times \nabla f(\beta_0) \quad (12)$$

On utilisera cette dernière équation dans la mise en œuvre de la méthode de Newton-Raphson sur une fonction de plusieurs variables.

Travail à faire

Soit la fonction de deux variables suivante définie sur $\mathbb{R}^2$:

$$f(x,y)=\frac{1}{e^{(x+0.5)^2+(y-0.8)^2}} \quad (13)$$

L'objet de cette partie est de déterminer :

$$(x^*,y^*)=\arg\max_{(x,y)} f(x)$$

avec (x,y) ∈ [-2 ;0]×[0 ;2].

1. ***Montrer que le vecteur gradient de f et que la matrice hessienne de f sont :***

$$\nabla f(x,y)=\begin{pmatrix} -\frac{(2x+1)}{e^{(x+0.5)^2+(y-0.8)^2}} \\ \frac{(1.6-2y)}{e^{(x+0.5)^2+(y-0.8)^2}} \end{pmatrix} \quad (14)$$

et

$$\nabla^2 f(x,y)=\begin{pmatrix} \frac{(2x+1)^2-2}{e^{(x+0.5)^2+(y-0.8)^2}} & \frac{(2x+1)(2y-1.6)}{e^{(x+0.5)^2+(y-0.8)^2}} \\ \frac{(2x+1)(2y-1.6)}{e^{(x+0.5)^2+(y-0.8)^2}} & \frac{(2y-1.6)^2}{e^{(x+0.5)^2+(y-0.8)^2}} \end{pmatrix} \quad (15)$$

2. ***La figure 6 ci-dessous montre le graphique en trois dimensions de la surface définie par l'équation (13).***

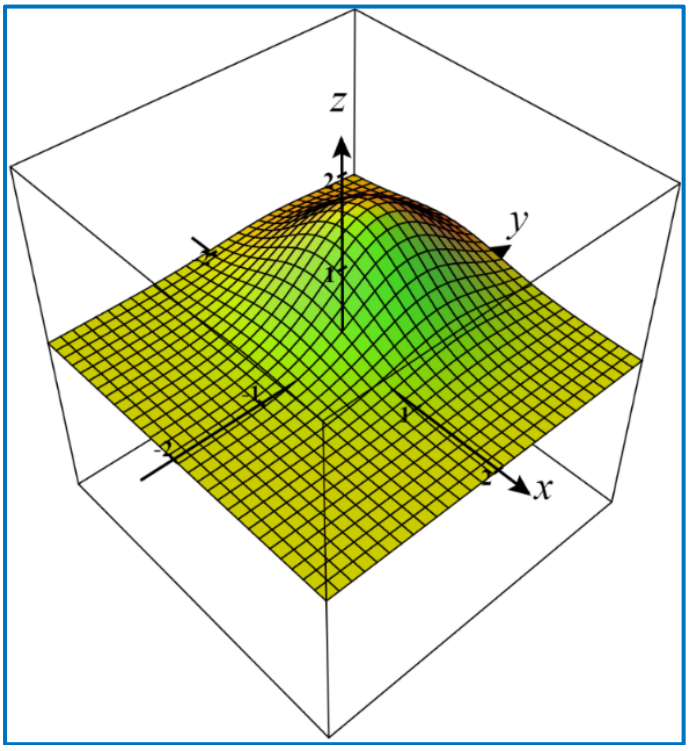


Figure 6

Evaluer graphiquement la valeur de (x*, y*) (le fichier Excel contient le même graphe qu'il est possible d'agrandir pour mieux estimer la solution).

3. ***Appliquer la méthode de Newton-Raphson en renseignant les tableaux de la feuille « Méthode_ Newton_Raphson_2 » du fichier Excel « Projet_Credit_Scoring.xlsm ». Les étapes à suivre sont les suivantes :***
   - ***On note $z_n = (x_n, y_n)'$ le vecteur des variables de f à l'étape n de l'optimisation. Prendre comme valeur initiale $z_0 = (-0.7;1.2)'$.***
   - ***Pour chaque itération indiquée dans la feuille Excel, calculer dans les cellules prévues à cet effet les valeurs de $f(z_n)$, $\nabla f(z_n)$ et de $\nabla^2 f(z_n)$.***
   - ***A partir de n=1, la valeur de $f(z_n)$ est déterminée à chaque étape en appliquant la relation (12) sur les éléments déterminés à l'itération précédente.***
   - ***On appliquera le critère de convergence suivant :***

$$\left|f\left(x_{n+1}\right)-f\left(x_n\right)\right|<\varepsilon$$

avec $\varepsilon=1.10^{-15}$. Le test d'arrêt sera codé dans les cellules prévues à cet effet par une fonction SI() qui affichera « OUI » lorsque le critère d'arrêt sera validé et « NON » sinon.

c) Maximum de vraisemblance et méthode de Newton-Raphson pour la régression logistique.

On suppose qu'on a un échantillon de *n* individus avec *p* variables explicatives que l'on représente de manière classique dans la matrice *X* suivante :

$$X=\begin{pmatrix}1 & x_{11} & \dots & x_{1p}\\ 1 & x_{21} & \dots & x_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & x_{n1} & \dots & x_{np}\end{pmatrix} \quad (16)$$

Pour chaque individu de l'échantillon, nous avons collecté les valeurs d'une variable Y_i prenant la valeur 1 si l'individu *i* présente le caractère étudié et 0 sinon. On note *Y* le vecteur des observations y_i :

$$Y=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix} \quad (17)$$

En reprenant l'équation (5) démontrée précédemment et en généralisant à notre échantillon de *n* individus, nous pouvons écrire la log-vraisemblance :

$$L(\beta)=\ln(V)=\sum_{i=1}^{n}\left[y_i.\ln(p_i)+(1-y_i).\ln(1-p_i)\right] \quad (18)$$

avec p_i la probabilité pour l'individu *i* de présenter le caractère étudié en l'occurrence le défaut :

$$p_i=\frac{e^{X_i\beta}}{1+e^{X_i\beta}} \quad (19)$$

où $\beta = (\beta_0\ \beta_1 \dots \beta_p)'$ représente les coefficients à estimer et $X_i = (1, x_{i1}, \dots, x_{ip})$ le vecteur des variables observées pour l'individu *i*.

L'estimation des coefficients de la régression par la méthode du maximum de vraisemblance va donc se faire en appliquant l'algorithme de Newton-Raphson. De façon plus formelle, l'estimation de $\beta = (\beta_0\ \beta_1 \dots \beta_p)'$ va se faire

en appliquant le schéma itératif suivant à partir de valeurs initiales $\beta = \beta_0$:

$$\beta_{n+1} = \beta_n - \left[\nabla^2 L(\beta_0)\right]^{-1} \times \nabla L(\beta_0) \quad (20)$$

où $\nabla L(\beta)$ et $\nabla^2 L(\beta)$ désignent respectivement le vecteur gradient et la matrice hessienne de la fonction de vraisemblance $L(\beta)$ de l'équation (18).

On démontre que $\nabla L(\beta)$ peut s'écrire :

$$\nabla L(\beta) = X'(Y - P) \quad (21)$$

avec :

$$P = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} \quad (22)$$

où p_i correspond à l'équation (19).

On démontre aussi que $\nabla^2 L(\beta)$ peut s'écrire :

$$\nabla^2 L(\beta) = X' W_\beta X \quad (23)$$

avec W_β la matrice définie comme suit :

$$W_\beta = \begin{pmatrix} -p_1(1-p_1) & 0 & \dots & 0 \\ 0 & -p_2(1-p_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -p_n(1-p_n) \end{pmatrix} \quad (24)$$

avec p_i tel que défini par l'équation (19).

On démontre que le gradient est strictement concave ce qui assure la convergence en une solution unique de l'algorithme de Newton-Raphson. En générale on initialise le processus en fixant $\beta = \beta_0 = (0, 0, \dots, 0)'$.

Travail à faire

Le but de cette partie est d'appliquer la méthode décrite dans les équations (16) à (24) sur un échantillon de 6 entreprises. L'échantillon est réduit afin de rendre possible le calcul des estimateurs avec les formules matricielles d'Excel. Les données se trouvent dans la feuille « Estimation_coeff_Newton_Raphson » du fichier Excel qui accompagne ce projet :

Entreprise *i*	Constante	X_1 BFR/TA	X_2 RES/TA	X_3 REX/TA	Y DEFAUT
1	1	0.028	3.196	0.284	1
2	1	-0.022	0.138	0.697	1
3	1	0.004	0.085	0.189	1
4	1	0.043	0.956	0.335	0
5	1	0.052	1.065	0.335	0
6	1	0.027	0.804	0.246	0

Comme précédemment, la variable Y prend la valeur 1 si l'entreprise a fait défaut et 0 sinon. Les variables X_1, X_2 et X_3 correspondent à 3 ratios d'analyse financière utilisés par Altman (1968).

Dans les questions qui suivent il faudra prendre soin d'insérer des $ dans le codage des formules de manières à pouvoir copier et coller tous les tableaux de l'itération 0 dans les suivantes jusqu'à arriver à la convergence de l'algorithme.

1. *Coder les équations (18) et (19) en renseignant la matrice en D13:E18 et la cellule E20 :*

	A	B	C	D	E
9					
10	**Fonction LOGIT - Initialisation :**		**0**		
11					
12	**β**			**$p_i=1/(1-e^{-(xiβ)})$**	**$L_i(β)$**
13	**$β_0$**	0			
14	**$β_1$**	0			
15	**$β_2$**	0			
16	**$β_3$**	0			
17					
18					
19					
20				**Fonction L(β)**	

Figure 7

Utiliser une formule matricielle pour coder l'équation (19) en D13:D18. Pour l'équation (18) il faut coder la vraisemblance d'un individu et faire la somme des vraisemblances individuelles dans la cellule E20.

2. *Coder le vecteur gradient de l'équation (21) dans la plage de cellules G13:G16 en utilisant des formules matricielles :*

	G
12	**∇(L(β))**
13	
14	
15	
16	
17	
18	
19	**Convergence**
20	n.a

Figure 8

3. *Coder la matrice $W_β$ définie à l'équation (24) dans la plage de cellule N13:S18 :*

	M	N	O	P	Q	R	S
11							
12		**$W_β$**					
13							
14							
15							
16							
17							
18							
19							

Figure 9

4. ***Coder la matrice hessienne de l'équation (23) dans la plage de cellule I13:S16 :***

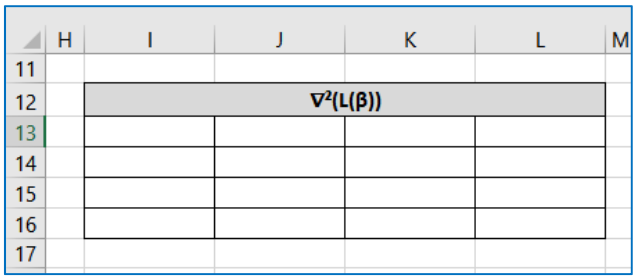


Figure 10

5. ***Appliquer la relation de récurrence de l'équation (20) dans la plage de cellule B26:B29 :***

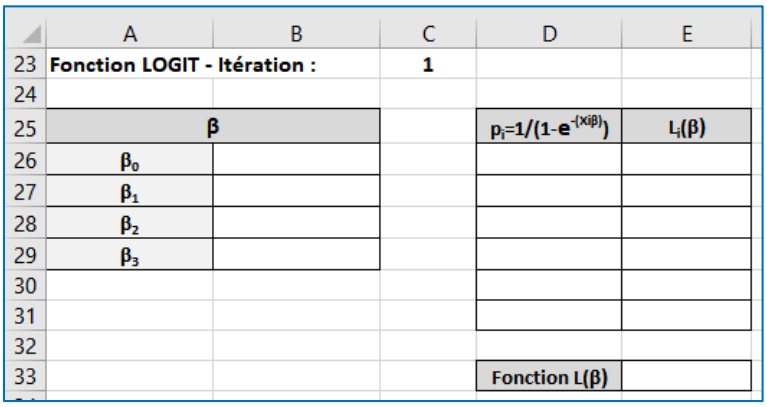


Figure 11

6. ***A partir de là, il faut copier et coller les formules des étapes précédentes afin d'automatiser les calculs. Une fois calculée la log-vraisemblance L(β) en cellule E33, coder le critère de convergence suivant :***

$$\left|L\left(\beta_{n+1}\right)-L\left(\beta_{n}\right)\right|<\varepsilon$$

avec ε = 1.10^{-11}. Le test d'arrêt sera codé dans les cellules prévues à cet effet par une fonction SI() qui affichera « OUI » lorsque le critère d'arrêt sera validé et « NON » sinon.

7. ***Copier et coller toutes vos formules jusqu'à ce que la cellule du critère de convergence affiche « OUI ». Si tout est bien paramétré vous devriez converger à l'itération 28.***

III. CODAGE D'UNE NOUVELLE FONCTION EXCEL LOGIT() POUR ESTIMER LA PROBABILITE DE DEFAUT

Dans cette dernière partie, il s'agit d'appliquer les connaissances acquises dans les parties précédentes pour coder une fonction *LOGIT()* permettant d'automatiser les calculs effectués dans Excel pour un échantillon de n'importe quelle taille *n* et pour le nombre de variables *p* désiré.

Travail à faire :

1. ***Coder en VBA la fonction LOGIT (Var_Y, Var_X). Dans Var-Y, l'utilisateur entrera le vecteur des observations Y qui ne contient que des 0 et des 1 et dans Var_X, l'utilisateur entrera la matrice des variables explicatives X.***

> *La colonne de 1 de la matrice X (équation 16) doit être ajoutée par la fonction comme pour la régression linéaire vue en cours. L'utilisateur ne sélectionne que les variables X_{ij} présentes initialement dans les données de base.*

2. ***Une fois la fonction codée et opérationnelle, on vous demande de l'appliquer à des données réelles collectées sur 830 entreprises sur la période 2005 – 2014 soit au total 4 000 entreprises-années. Les données se trouvent dans la feuille « Credit_Scoring » du fichier Excel « Projet_Credit_Scoring.xlsm ». Les ratios retenus dans l'analyse sont les mêmes que ceux proposés par Altman (1968) :***
   - ***X_1 : BFR/TA ==> Besoin en Fonds de Roulement d'exploitation / total de l'actif ;***
   - ***X_2 : RES/TA ==> Réserves / total de l'actif ;***
   - ***X_3 : REX/TA ==> Résultat d'exploitation / total de l'actif ;***
   - ***X_4 : CAPI/VCD ==> Capitalisation boursière / valeur comptable des dettes ;***
   - ***X_5 : CA/TA ==> Chiffre d'affaires / total de l'actif.***

   Donner les valeurs de $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_4, \widehat{\beta}_5,)'$ estimés sur les 4 000 observations de la feuille « Credit_Scoring » grâce à la fonction LOGIT(). Interpréter les résultats.

3. ***Comparer les coefficients et les signes donnés par la fonction LOGIT() avec ceux estimés par Altman en 1968. Commenter les différences.***

4. ***Reprendre les données téléchargées sur Bloomberg pour le travail demandé en première partie. Ajouter une colonne « Proba défaut Logit » à cotés du Z-Score d'Altman dans laquelle vous calculerez la probabilité de défaut pour chaque entreprise en appliquant l'équation (19) :***

$$p_i = \frac{e^{X_i \hat{\beta}}}{1 + e^{X_i \hat{\beta}}} \quad (19)$$

   où $\widehat{\beta}$ représente le vecteur des coefficients estimés grâce à la fonction LOGIT() et $X_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5},)$ le vecteur des ratios observés pour l'entreprise i.

5. ***Comparer les classifications données par les deux méthodes sur vos données.***

6. ***Proposer des améliorations de la fonction LOGIT.***

IV. BIBLIOGRAPHIE

Altman, E., 1968, « Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy », *The Journal of Finance*, Vol. 23, No. 4, pp. 589-609.

Balcaen, S. et Ooghe, H., 2006, « 35 years of studies on business failure : an overview of the classic statistical methodologies and their related problems », *The British Accounting Review*, Vol. 38, No. 1, pp. 63-93.

Imtiaz, S. et Brimicombe, A. J., 2017, « A Better Comparison Summary of Credit Scoring Classification », *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 7, pp. 1-4.