Modélisations dynamiques avec le tableur

Philippe Langenaken(*)

Dans les études supérieures de type économique, les calculs d'optimisations sont nombreux. Par ailleurs, les problèmes économiques sont fréquemment illustrés par des tableaux de données et des graphiques. Pour s'en convaincre, il suffit de parcourir une revue spécialisée ou les pages éco des quotidiens.

Mais nos étudiants qui abordent la première des cinq années d'études dans notre école de commerce ont souvent, pour seul bagage mathématique, des connaissances de formules et de routines de calcul qui leur ont permis dans l'enseignement secondaire de résoudre des exercices types aux données parfois simplistes. Il leur manque une « culture » graphique qui devrait leur permettre de mieux appréhender l'applicabilité des objets mathématiques aux problèmes que leurs études leur feront rencontrer.

Les recherches d'extremums sont difficiles à faire passer. Les étudiants ne *voient* pas bien ce qu'ils font, ils calculent souvent en automates en confondant les notions : valeurs de x et de $f(x)$, sens des extremums, ils omettent les calculs aux bornes des domaines, … Il est vrai que les outils didactiques classiques (tableau et craie, calculatrice) ne permettent d'effectuer des représentations qu'en un temps assez long.

Concepteurs d'un gros cours en ligne de 90 heures par année, nous avons été amenés à nous poser la question de l'utilisation par nos étudiants des ressources mêmes de l'ordinateur sur lequel ils apprennent, leur permettant d'acquérir peut-être plus aisément les notions mathématiques par des manipulations dynamiques.

Nous avons montré lors de l'atelier « Visualisations de cas d'optimisations » à Caen en octobre 2005 quelques réalisations, dont des approximations de dérivées de fonctions et une approche graphique de la recherche de coût unitaire minimum. Nous détaillons ici un exercice de modélisation qui permet de redécouvrir par la pratique diverses propriétés des études de fonctions et des calculs d'extremums.

Cette application utilise à fond la puissance de calcul de l'ordinateur : nous y verrons qu'une variation dynamique d'un paramètre engendre le recalcul de milliers de valeurs en une fraction de seconde.

1. Concepts fondamentaux

1.1. Tables de mortalité

Les **tables de mortalité** sont créées à partir des **taux de mortalité** par âge x qu'on notera q_x, mesurés à partir des données démographiques. On en déduit des tables de

(*) Haute École Francisco Ferrer – IREM de Bruxelles.
philippe.langenaken@brunette.brucity.be

survie qui reprennent les effectifs de population théoriques l_x par âge x sur base d'une population initiale donnée (en général 100 000 ou 1 000 000), avec

$$l_{x+1} = l_x(1-q_x).$$

On trouve ainsi sur le site de l'INED[1] la table de mortalité française 2000-2002. Les analyses se basent sur deux ans, entre deux dates anniversaires de l'échantillon :

Age x	Sexe masculin			Sexe féminin			Les deux sexes		
	Lx	Q(x, x+1)	Ex	Lx	Q(x, x+1)	Ex	Lx	Q(x, x+1)	Ex
0	100.000	489	75,51	100.000	384	82,9	100.000	438	79,11
1	99.511	38	74,87	99.616	33	82,22	99.562	35	78,46
2	99.473	27	73,9	99.583	21	81,24	99.527	25	77,49
...	...	...	...	...	...	...	...	...	...

On introduit le taux de survie p_x à l'âge x tel que

$$p_x + q_x = 1,$$

ce qui nous donne :

$$l_{x+1} = l_x \cdot p_x.$$

1.2. Modèles

La modélisation des tables de mortalité a pour but de donner une base théorique permettant aux assureurs de calculer leurs risques : risques pour l'assurance-vie ou pour l'assurance-décès.

La plupart des modèles se basent sur des hypothèses simplificatrices qui permettent de faciliter les calculs en omettant certains paramètres négligeables ou difficilement traitables dans la recherche de fonctions continues qui « collent » le mieux aux expérimentations. On fera entre autres ici l'hypothèse de statique : dans le modèle, le taux de survie devient une probabilité de survie, et la probabilité de survie pendant un an dans k ans pour quelqu'un d'âge x aujourd'hui est égale à la probabilité de survie pendant un an de quelqu'un âgé aujourd'hui de $x + k$ ans.

Les données des tables de mortalité mènent à des modèles de fonctions de survie que nous noterons L(x). Il existe plusieurs modèles utilisés. Nous en détaillerons un ci-dessous.

1.3. Taux instantané de décès

Pour un modèle de fonction de survie continue L(x) à l'âge x, nous pouvons introduire le **taux de décès instantané** à l'âge x défini comme suit :

$$\mu(x) = \lim_{\Delta x \to 0} \frac{\dfrac{\mathrm{L}(x) - \mathrm{L}(x+\Delta x)}{\Delta x}}{\mathrm{L}(x)}.$$

En transformant, nous obtenons :

(1) Institut national d'études démographiques,
http://www.ined.fr/population-en-chiffres/france/index.html
Q(x ; x + 1) vaut $q_x \cdot 10^5$; E_x est l'espérance de vie à l'âge x.

$$\mu(x) = -\lim_{\Delta x \to 0} \frac{\dfrac{\mathrm{L}(x+\Delta x) - \mathrm{L}(x)}{\Delta x}}{\mathrm{L}(x)} = -\frac{\mathrm{L}'(x)}{\mathrm{L}(x)} = -(\ln \mathrm{L}(X))'.$$

2. Modèle de Makeham

2.1. Description

William Matthew Makeham est le fondateur de l'actuariat moderne. Le modèle de Makeham (ou Gompertz-Makeham[2]) part d'un taux instantané de décès défini par un risque exponentiel lié au vieillissement auquel s'ajoute une constante liée au risque accidentel :

$$\mu(x) = \mathrm{A} + \alpha c^x.$$

α est le risque initial de la population et c le coefficient d'aggravation du taux de décès par année. On a bien entendu $\mathrm{A} > 0$ et $c > 1$.

Sachant que

$$\mu(x) = -(\ln \mathrm{L}(x))' = \mathrm{A} + \alpha c^x,$$

nous obtenons

$$\ln \mathrm{L}(x) = -\mathrm{A}\,x - \frac{\alpha}{\ln c} c^x + \ln k$$

et

$$\mathrm{L}(x) = k \cdot \mathrm{e}^{-\mathrm{A}x} \cdot e^{-\alpha \frac{c^x}{\ln c}}$$
$$= k \cdot s^x \cdot g^{c^x}$$

où $s = \mathrm{e}^{-\mathrm{A}}$ peut être vu comme la probabilité de ne pas mourir par accident, et avec par ailleurs $g = \mathrm{e}^{-\frac{\alpha}{\ln c}}$ (ou $\alpha = -\ln g \cdot \ln c$) et $k = \dfrac{\mathrm{L}(0)}{g}$.

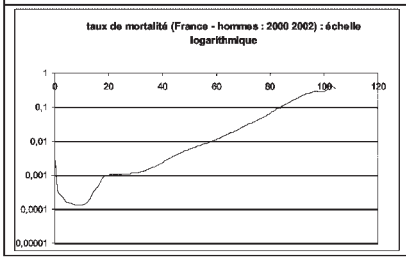
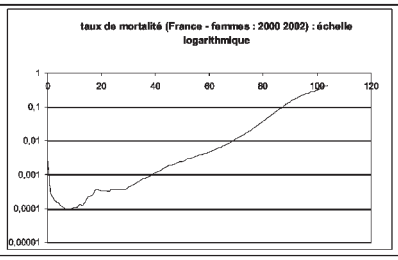
2.2. Ajustement au modèle

Le modèle de Makeham est imposé en Belgique pour la tarification des primes d'assurances vie et décès. Nous nous sommes intéressés à l'application du modèle de Makeham à la population française dont nous avons recueilli les données sur le site de l'INED.

Pour rechercher l'ajustement idéal, nous utiliserons le tableur Excel, présent sur pratiquement tous les micro-ordinateurs. Les données nous permettent aisément de calculer le taux de mortalité q_x.

Voici, tracés avec Excel, les graphiques associés aux taux de mortalité en France en 2000-2002, respectivement pour les hommes et pour les femmes :

(2) Le modèle de Gompertz (1825) définit $\mu(x) = \alpha c^x$. Makeham introduit en 1862 la composante en risque accidentel A.

Sexe masculin	Sexe féminin
	

Ces représentations en échelle semi-logarithmique permettent d'observer une certaine mortalité infantile, puis une bosse autour de 20 ans, sensible en particulier chez les hommes. Elle peut s'expliquer par les accidents de la route et les suicides chez les jeunes.

2.2.1. Transformations

L'outil « Courbe de tendance » dans un graphique d'Excel fonctionne pour des approximations affines, polynomiales, exponentielles ou logarithmiques. Afin de pouvoir utiliser cet outil, nous allons quelque peu manipuler les données.

Nous savons que

$$l_{x+1} = l_x \cdot p_x$$

ou

$$\frac{1}{p_x} = \frac{l_x}{l_{x+1}}.$$

Nous allons tenter d'ajuster les valeurs du modèle $\frac{\mathrm{L}_x}{\mathrm{L}_{x+1}}$ avec $\frac{1}{p_x}$:

$$\frac{\mathrm{L}_x}{\mathrm{L}_{x+1}} = \frac{ks^x g^{c^x}}{ks^{x+1} g^{c^{x+1}}} = \frac{1}{s} g^{c^x - c^{x+1}} = \frac{1}{s} g^{c^x(1-c)}.$$

Pour obtenir une expression ajustable avec Excel, nous prenons le logarithme :

$$\ln\left(\frac{1}{p_x}\right) \approx -\ln s + c^x(c-1)\ln\left(\frac{1}{g}\right),$$

$$\ln\left(\frac{1}{p_x}\right) + \ln s \approx c^x(c-1)\ln\left(\frac{1}{g}\right),$$

$$\ln\left(\frac{1}{p_x}\right) - \mathrm{A} \approx c^x(c-1)\ln\left(\frac{1}{g}\right),$$

puisque dans le modèle de Makeham, nous avons défini $s = \mathrm{e}^{-\mathrm{A}}$, et comme la composante accidentelle A est strictement positive, la probabilité de ne pas mourir par accident s, sera strictement inférieure à 1.

2.2.2. Ajustement avec Excel

$c^x(c-1)\ln\left(\frac{1}{g}\right)$ est une fonction exponentielle et nous pouvons maintenant utiliser Excel pour effectuer l'ajustement. Mais nous ne connaissons pas s. Nous allons le faire varier et déterminer la valeur de s pour laquelle la courbe de tendance aura le meilleur coefficient de détermination R^2. Nous avons choisi l'ajustement « linéaire » sur les logarithmes des valeurs (colonne F). Les valeurs négatives obtenues pour la fourchette d'âges de 0 à 18 ans ont été éliminées. Cela ne nous gêne guère puisque le propre d'un modèle est d'être local, et que le public qui intéresse les assureurs est un public de plus de 18 ans.

Les fichiers Excel de l'application du modèle de Makeham à la population française (hommes et femmes) peuvent être téléchargés sur

http://www.brunette.brucity.be/ferrer/UERMath/activ/200510Caen.

Examinons la table de mortalité française pour les hommes. Nous faisons varier s grâce à une barre de défilement, par pas de 0,000 01.

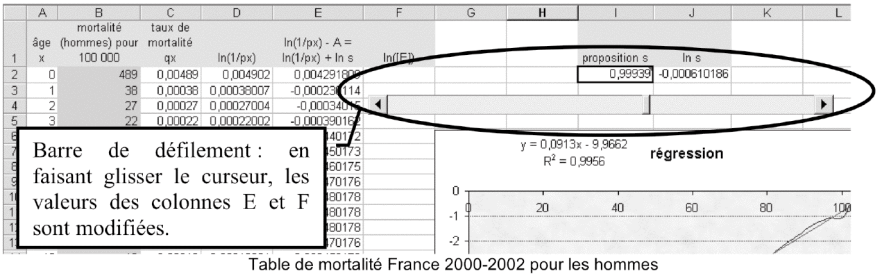
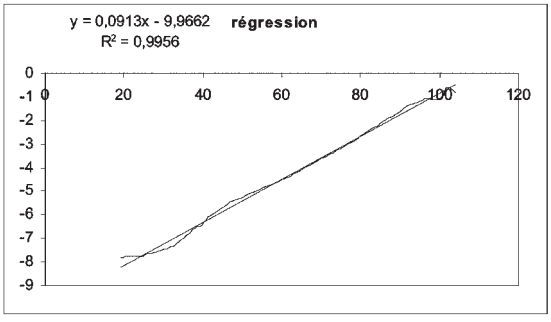


Table de mortalité France 2000-2002 pour les hommes

À chaque changement de valeur, tous les calculs sont réeffectués et le graphique actualisé. En faisant diminuer progressivement s à partir de la valeur 1 à l'aide de la glissière, on observe la valeur du coefficient de détermination R^2 sur le graphique. R^2 commence à croître, de plus en plus lentement, puis se stabilise à 0,995 6 pour plusieurs valeurs, avant de décroître lorsqu'on continue à faire diminuer s. Nous avons donc un maximum local. C'est l'occasion de faire observer les propriétés de la fonction autour de ce maximum : la fonction R^2 croît, le taux d'accroissement diminue, tend vers zéro…, puis la fonction R^2 décroît.



Ajustement optimal (France 2000-2002 hommes)

En continuant à faire varier *s*, on découvre qu'il n'existe pas d'autre maximum. Mais nous avons 11 valeurs de *s* pour lesquelles R^2 vaut la valeur maximum 0,995 6. Une idée serait de prendre la valeur du milieu, mais nous préférons vérifier en calculant de manière plus précise, toujours avec Excel, les paramètres de la droite de régression et le coefficient de détermination. Nous découvrons alors que le maximum est atteint pour $s = 0{,}999\,39$.

L'équation de la droite ($y = 0{,}091\,301\,36x - 9{,}966\,168\,25$) nous permet alors de calculer les valeurs suivantes pour le modèle :

A	0,00061019
α	4,4851E-05
g	0,99950888
c	1,09559912
k	100049,136

2.3. Éléments de calcul

Quelques petits rappels de statistiques s'imposent ici pour appréhender la quantité de calculs opérés par le tableur pour chaque changement de valeur *s*.

La droite de régression est celle obtenue par la méthode dite des moindres carrés, introduite par Gauss en 1802. Pour un ensemble d'observations (*x*,*y*), on obtient son équation en minimisant les sommes des carrés

$$\sum (y_r - y)^2$$

des différences entre les valeurs observées *y* et les valeurs « estimées » $y_r = ax + b$.

La minimisation de cette somme de carrés nous fournit pour la droite un coefficient directeur *a* égal au quotient de la covariance des *x* et *y* par la variance des *x*.

$$a = \frac{\text{cov}(x, y)}{s_x^2}.$$

L'ordonnée à l'origine *b* de la droite de régression est donnée par :

$$b = \overline{y} - a\overline{x}$$

où $\overline{x}$ et $\overline{y}$ représentent les moyennes respectives des *n* valeurs *x* et *y*, c'est-à-dire

$\frac{1}{n}\sum x$ et $\frac{1}{n}\sum y$.

La variance s_x^2 vaut

$$\frac{1}{n}\sum (x - \overline{x})^2$$

ou, par un calcul simple,

$$\frac{n\sum x^2 - \left(\sum x\right)^2}{n^2}$$

qu'on retrouve comme définition de la fonction VAR.P dans Excel.

La covariance cov(x,y) vaut

$$\frac{1}{n}\sum\left((x-\bar{x})(y-\bar{y})\right)=\frac{n\sum(xy)-\sum x\sum y}{n^2}.$$

Les deux paramètres a et b de la droite de régression sont fournis dans Excel par la fonction DROITEREG.

Une covariance positive indique une tendance des valeurs observées x et y à varier dans le même sens. On obtiendra dans ce cas une droite de régression à coefficient directeur positif. Une covariance négative indique une tendance à varier en sens opposés, et le coefficient directeur de la droite de régression sera négatif.

Une covariance proche de zéro indique soit l'indépendance des informations x et y observées, soit une dépendance de type non linéaire ou en tout cas non représentable par une stricte croissance ou une stricte décroissance. Dans tous ces cas, l'ajustement par une droite de régression n'est pas adéquat. Pour mesurer la qualité de l'ajustement des observations (x,y) par la droite de régression, on utilise le coefficient de détermination R^2.

Il est défini par

$$R^2=\frac{\left(\text{cov}(x,y)\right)^2}{s_x^2\ s_y^2}=\frac{\left(n\sum(xy)-\sum x\sum y\right)^2}{\left(n\sum x^2-\left(\sum x\right)^2\right)\left(n\sum y^2-\left(\sum y\right)^2\right)}.$$

On montre aisément que

$$R^2=\frac{s_y^2-\sum(y_r-y)^2}{s_y^2}=1-\frac{\sum(y_r-y)^2}{\sum(y-\bar{y})^2}$$

avec $y_r = ax + b$ donnés par l'équation de la droite de régression.

Maximiser le coefficient de détermination revient donc à minimiser le total des carrés des différences de valeurs $(y_r - y)$ par rapport au modèle.

Un coefficient de détermination égal à 1 indique une corrélation parfaite entre les observations et les valeurs déterminées par la droite (aucune différence entre les valeurs y estimées et réelles). À l'inverse, un coefficient de détermination proche de zéro indique que l'équation de régression ne peut servir à déterminer une valeur y.

Dans Excel, on utilisera la fonction COEFFICIENT.DETERMINATION.

2.4. Utilisation pratique du modèle

En Belgique, le législateur impose le modèle de Makeham pour la tarification des primes d'assurances vie et décès. Ce faisant, il refuse la prise en compte de la plus grande probabilité de décès chez les hommes jeunes qui se voit par la « bosse » dans le graphique des taux de mortalité en coordonnées semi-logarithmiques vus plus haut. La tarification imposée se base toujours actuellement sur les mesures de 1986-1988.

Le Moniteur (Journal officiel belge) du 31 décembre 1992 fournit les paramètres s, g et c pour les hommes et les femmes :

	Hommes	Femmes
s	0,999 441 703	0,999 669 730
g	0,999 624 664	0,999 935 634
c	1,103 798 111	1,119 312 877

Il impose, « *moyennant une marge de sécurité multiplicative et additive pour ce qui concerne les opérations de genre décès, une marge de sécurité correspondant à un accroissement attendu de la longévité et à une correction d'antisélection pour les opérations de genre vie* », quatre séries de valeurs pour les opérations de genre vie ou décès, pour les hommes et les femmes :

	Vie Hommes	Vie Femmes	Décès Hommes	Décès Femmes
k	1 000 266,63	1 000 048,56	1 000 450,59	1 000 097,39
s	0,999 441 704	0,999 669 731	0,999 106 876	0,999 257 048
g	0,999 733 441	0,999 951 440	0,999 549 614	0,999 902 624
c	1,101 077 536	1,116 792 454	1,103 798 111	1,118 239 062

Ces paramètres de la formule $L(x) = k \cdot s^x \cdot g^{c^x}$ pour 1 000 000 de naissances permettent de calculer le paramètre de risque accidentel A et le risque initial de la population α à l'aide des formules vues plus haut.

Pour un exposé détaillé du modèle de Makeham, la critique de ce modèle imposé, ainsi que l'approche par d'autres modèles, le lecteur se référera à l'ouvrage de Daniel Justens et Laurence Hulinelles « Théories actuarielles », publié aux Éditions du Céfal en 2003 (ISBN 287130136-0).

3. Bilan

Dans le cadre de la résolution d'un problème a priori fort complexe (une modélisation par exponentielle d'exponentielle), nous avons abordé dynamiquement diverses propriétés des fonctions, en particulier dans la recherche du coefficient de détermination optimal pour la régression paramétrique. Les notions de dérivée, de croissance et décroissance, la manipulation des exponentielles et logarithmes ont été vues ici dans un cadre concret.

L'utilisation des concepts sur des données réelles a un effet motivant sur les étudiants, susceptible de mieux ancrer la connaissance des différents objets mathématiques dans les esprits.

Le calcul, laissé au logiciel, permet de se concentrer sur le sens des notions abordées.

Nous avons travaillé graphiquement : après avoir utilisé des échelles logarithmiques pour les taux de mortalité, nous avons, au gré des développements, ramené le problème à une recherche de droite de régression … à *des* recherches de droites de régression. C'est ici que l'ordinateur joue un rôle essentiel, en permettant d'actualiser en une fraction de seconde des centaines de données et leur représentation graphique.

Reste à faire acquérir aux étudiants cette « culture de la représentation » qui leur manque souvent.