De Gompertz à Verhulst :
une petite histoire illustrée de la fonction logistique
Daniel Justens

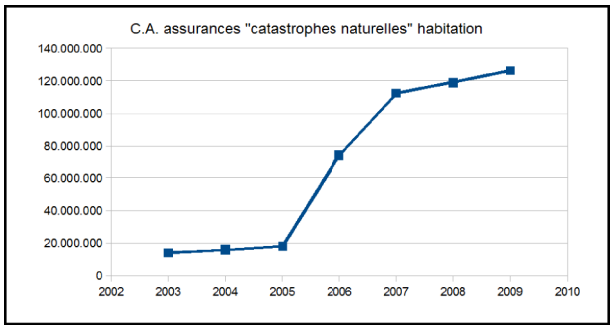
Chapeau
Alors que les résultats des secteurs classiques des assurances affichent des progressions calquées sur les indices de prix sans véritable extension possible de marchés saturés, de nouveaux produits d'assurances sont proposés actuellement, dont le chiffre d'affaire initialement très faible, a explosé dans la première décennie de ce siècle, pour atteindre ensuite une certaine stabilisation. Comment modéliser ce type de progression lente, puis explosive, puis ralentie à nouveau. Différents modèles sont en concurrence pour ce genre de comportement, comme celui de Gompertz ou ceux de Verhulst (fonction logistique).

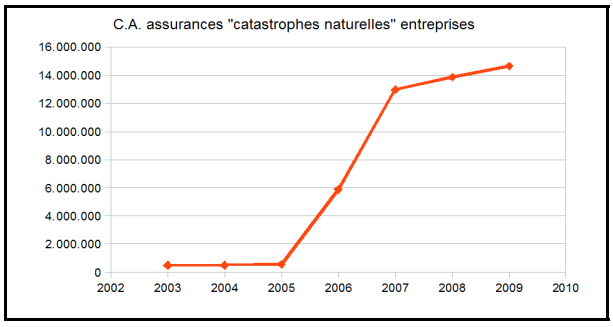
Un nouveau secteur d'assurances : les « catastrophes naturelles »

Les chiffres d'affaires de la plupart des contrats classiques d'assurances sont arrivés à un stade de saturation. Leur progression suit en gros l'évolution des prix. Aucune percée significative ne peut plus être espérée dans ces domaines. Et c'est bien naturel. Toutefois, c'est dans ces secteurs stables que se perçoivent les chiffres d'affaire les plus importants et surtout les plus réguliers. Il s'agit donc de produits nécessaires et assurant l'avenir des entreprises d'assurances. Mais ces derniers ne répondent plus nécessairement à toutes les attentes des citoyens et des entreprises. Il convient donc de se pencher sur de nouveaux produits et d'anticiper le développement de nouveaux marchés. Voyons quels furent dans la première décennie de ce siècle les résultats cumulés en termes de perception de primes des assurances dites « catastrophes naturelles » en Belgique pour les particuliers et les entreprises.

	habitation AR catastrophes naturelles	AR entreprises catastrophes nat
2003	14.212.942	492.539
2004	16.063.333	502.283
2005	18.024.188	568.989
2006	74.178.520	5.881.418
2007	112.314.800	12.989.455
2008	118.931.020	13.875.916
2009	126.491.964	14.652.567

Si les ordres de grandeurs sont différents, les deux chroniques présentent une structure d'évolution comparable, comme on peut le voir dans les deux graphiques ci-dessous dont l'allure générale porte le nom de « sigmoïde ».





Encadré
Benjamin Gompertz (1779 - 1865) fut un mathématicien anglais d'origine hollandaise. Il fut autodidacte car on lui refusa l'entrée à l'université en raison de ses origines juives. Il se forma en découvrant les travaux de Newton et de Mac-Laurin. La qualité de ses résultats fut néanmoins reconnue et il fut reçu *fellow* de la *Royal Society* en 1819. Il est connu des actuaires pour son modèle viager (1825) présentant les taux instantanés de mortalité comme une fonction exponentielle ce qui conduit à une représentation du nombre de survivants affectant l'allure d'une exponentielle d'exponentielle, ouvrant ainsi les portes de l'actuariat moderne.
Fin encadré

Courbes de Gompertz

Plusieurs tentatives ont été faites pour modéliser et formaliser ce type de comportement. La première est due au mathématicien Gompertz (voir encadré). Formellement, la fonction qu'il a proposée est du type :

$$f(t) = e^{-\alpha\beta^t+\gamma} = K\,e^{-\alpha\beta^t}$$

Dans cette expression, α est supposé strictement positif et, dans le cas qui nous intéresse, β doit être compris entre 0 et 1, de façon à obtenir le type de fonction sigmoïde espéré. On vérifie que la fonction est strictement croissante et possède une asymptote horizontale en $e^{\gamma} = K$ pour t tendant vers l'infini. Un ajustement aux observations est possible en utilisant la régression paramétrique. Pour toute valeur de K raisonnable, c'est-à-dire strictement supérieure à toutes les

observations, on peut procéder à une régression linéaire simple en passant deux fois aux coordonnées logarithmiques. Chaque modèle linéaire possède une détermination qui mesure la puissance explicative du modèle en terme de diminution en pourcentage de la variance. On retient alors la valeur de K qui maximise cette détermination. La relation idéale peut s'écrire pour tout couple d'observations (t_i, y_i) :

$$y_i = K\, e^{-\alpha\beta^{t_i}}$$

Les éventuels écarts au modèle seront pris en compte plus loin dans la procédure. Un premier passage aux logarithmes livre :

$$\ln(y_i) - \ln(K) = -\alpha\beta^{t_i} \quad \text{ou} \quad \ln(K) - \ln(y_i) = \alpha\beta^{t_i}$$

Le second :

$$\ln[\ln(K) - \ln(y_i)] = \ln(\alpha) + \ln(\beta)\, t_i$$

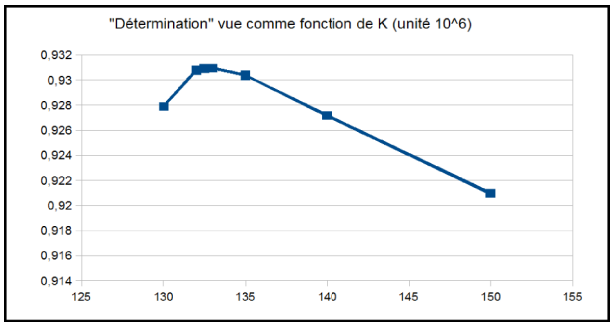
Ce sont les écarts entre ces deux quantités que la régression linéaire simple minimise. Ils prennent la forme :

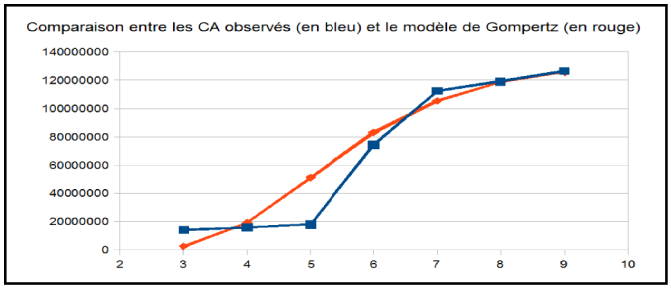
$$e_i = \ln[\ln(K) - \ln(y_i)] - \ln(\alpha) - \ln(\beta)\, t_i$$

K étant strictement supérieur à toute observation, la différence de logarithmes est toujours strictement positive ce qui autorise le second passage au logarithme. Il reste à minimiser

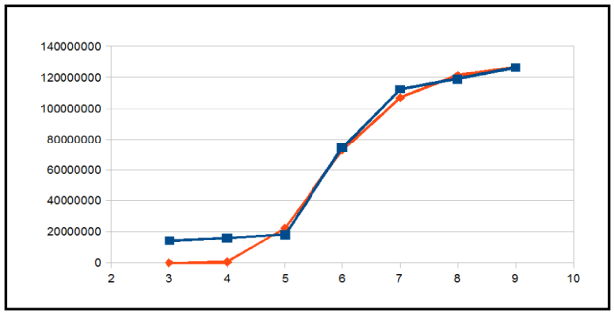
$$\sum_{i=1}^{n} e_i^2$$

en se souvenant évidemment des torsions successives de l'axes des ordonnées qui a conduit à la linéarisation. Chaque régression livre une valeur α et β et également une détermination qui quantifie la puissance explicative du modèle avec toutes les réserves à apporter à cette affirmation : travail avec la variance, torsions multiples de l'axe des ordonnées et méthode non robuste d'ajustement. Dans le cadre de notre exemple, le maximum de détermination s'obtient pour K = 133.000.000. La fonction détermination D(K) est représentée ci-contre et présente bien un maximum pour cette valeur.





Le modèle est-il pour autant satisfaisant. Une rapide comparaison graphique nous montre que non (origine des temps : 2000). En fait les deux premières observations quasiment identiques à la troisième font effet de levier.



En « oubliant » ces deux premières observations, on obtient par le même procédé un modèle presque parfait pour les observations prises en compte mais sous-estimant drastiquement les résultats antérieurs. On retient ici la valeur *K* = 129.000.000.

Le modèle de Verhulst : la « vraie » courbe logistique

Encadré

Pierre-François Verhulst (1804 - 1849) est un mathématicien belge élève de Quetelet. La lecture de Malthus, lui inspira une succession de modèles (en 1838, en 1845 et en 1847) d'évolution de populations animales et humaines au moyen de fonctions non exponentielles, en optant pour l'introduction d'une contrainte de saturation. C'est en 1845 qu'il nomme la fonction obtenue *courbe logistique* sans fournir de justification à cette appellation. Verhulst ajusta plusieurs fois ses modèles aux populations belges et françaises avec des succès divers. La courbe logistique, utilisée dans l'étude des populations fut redécouverte en 1920 par les statisticiens et biologistes Raymond Pearl (1879 - 1940) et Lowell Jacob Reed (1886-1966) qui ne créditèrent Verhulst de la paternité de la découverte qu'en 1922. Le terme de *logistique* tomba dans l'oubli et ne réapparut qu'en 1924 dans une correspondance entre George Yule et Reed. On retrouve des traces de l'utilisation cette courbe en chimie

dans un inventaire (1929) de Joseph Berkson (1899 - 1982). C'est ce même Berkson qui ajustera certaines de ses observations au moyen de la fonction logistique en introduisant la fonction « logit ».
Fin encadré

Le mathématicien belge Verhulst proposa un modèle descriptif comme solution d'une équation différentielle supposée décrire l'évolution d'une population soumise à deux types de flux : d'une part l'apparition de nouveaux individus, d'autre part la disparition des individus existant et cela en proportion de la population. Verhulst entendait décrire une population *P(t)* soumise à deux contraintes contradictoires : une tendance à l'expansion au taux *r* et une sensibilité à la saturation qui contrarie cette tendance :

$$\frac{dP}{dt} = r\,P\left(1 - \frac{P}{K}\right)$$

L'équation se résout sous la condition d'existence d'une population initiale $P(0) = P_0$. *r* représente un taux de croissance tempéré en fonction du rapport entre le niveau de population déjà atteint et une certaine capacité porteuse notée *K*. On vérifie en effet que la population croît quand elle est inférieure à K et décroît lorsqu'elle est supérieure, ce qui justifie intuitivement cette appellation. Mais le modèle est bien plus riche et peut être appliqué à de multiples applications avec succès. Le changement de variable $P = 1/f$ (qui induit $dP = -df/f^2$) transforme cette équation différentielle en :

$$\frac{df}{dt} = r\left(\frac{1}{K} - f\right)$$

On tombe ici bien agréablement sur une équation différentielle ordinaire linéaire non homogène dont la solution est (vérification par simple dérivation) :

$$f(t) = Ce^{-rt} + \frac{1}{K}$$

La constante multiplicative *C*, issue du processus d'intégration sera déterminée à partir des conditions initiales. On en tire d'abord l'expression de la fonction *P(t)* :

$$P(t) = \frac{1}{f(t)} = K\frac{1}{1 + (C\,K)e^{-rt}}$$

pour ensuite déterminer *C* (ou *CK* ce qui revient au même), en tenant compte de $P(0) = P_0$ ce qui donne

$$P_0 = \frac{K}{1+CK} \quad \text{et donc} \quad C = \frac{K - P_0}{K\,P_0} \quad \text{ou} \quad CK = \frac{K - P_0}{P_0} = \frac{K}{P_0} - 1$$

On en vient enfin à l'expression générale de la fonction logistique de Verhulst :

$$P(t) = K\frac{1}{1 + \left(\frac{K}{P_0} - 1\right)e^{-rt}}$$

ou encore en modélisant plus spécifiquement le rapport entre la population et sa capacité porteuse :

$$\frac{P(t)}{K} = \frac{1}{1 + \left(\frac{K}{P_0} - 1\right)e^{-rt}}$$

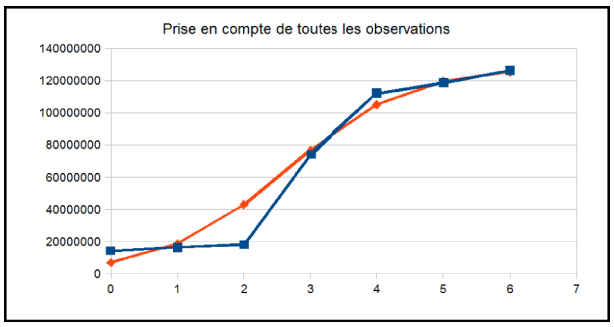
L'ajustement de ce type de fonctions aux observations se fait au moyen de la fonction *logit*, définie par :

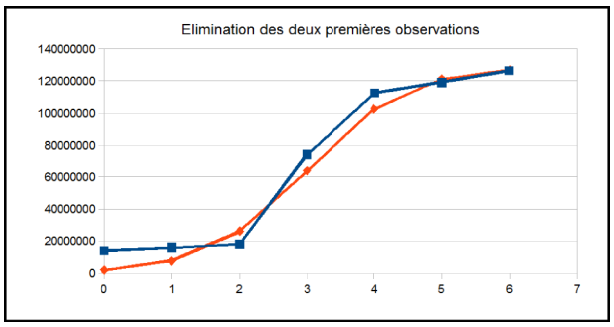
$$logit(x) = \ln\left(\frac{x}{1-x}\right)$$

Appliquée à notre rapport, l'opérateur *logit* transforme la fonction logistique en une fonction du premier degré qui permet un ajustement linéaire paramétrique comme plus haut. On peut légitimement s'interroger sur les torsions que la transformation logit fait subir aux observations et sur la nature des « erreurs » prises en compte lors de la procédure de régression linéaire. On a :

$$logit\left(\frac{P(t)}{K}\right) = \ln\left[\frac{e^{rt}}{\frac{K}{P_0} - 1}\right] = rt - \ln\left(\frac{K}{P_0} - 1\right)$$

Ici le maximum de détermination s'obtient pour *K*=129.000.000, qui avait déjà convenu à notre modèle Gompertz 2. Voyons ce que donne cet ajustement à nos chiffres d'affaires en optant pour les mêmes hypothèses que plus haut : prise en compte de toutes les observations ou élimination des deux premières :





Comme toujours, il y a un monde de différence entre les tentatives de modélisation élémentaires, adoptant une forme structurelle unique, tentant une explication interprétée globale et le réel, mouvant, multiple. L'univers économique subit un grand nombre d'influences dont les effets au cours du temps sont loin d'être constants. Il semble donc illusoire de donner aux modèles un pouvoir prédictif qu'ils ne peuvent avoir, au moins à long terme. De plus, les deux modèles présentés reposent sur une même hypothèse à savoir l'existence d'un niveau de saturation absolu K. Dans le cadre de notre exemple, nous avons rappelé que les observations en termes de perceptions de primes pour les contrats classiques suivaient grosso-modo une exponentielle à taux raisonnable, de l'ordre du taux d'inflation. Il est donc probable, même en marché saturé qu'une telle évolution sera encore observée pour les nouveaux produits. Il conviendrait donc d'adapter les modèles en fonction de cette observation.