STATISTIQUE MATHÉMATIQUE - MS ENSAI
TP2 : Buy Til You Die! VVC et modèle Pareto/NBD en marketing.
Sept. 2025

1 Hypothèses du modèle

La VVC (valeur vie client) ou CLV (Customer Lifetime Value) modélise le comportement d'achat des clients dans une enseigne donnée et est une grandeur d'importance en marketing. C'est aussi un indicateur prédictif qui peut également servir à valoriser une entreprise à partir de son portefeuille de clients. Elle est utilisée en téléphonie, en banque, en assurance, etc. et est une notion essentielle dans les problématiques d'acquisition de clients, de fidélisation ou de prévision de l'attrition.

Précisément, la VVC représente la somme des profits actualisés attendus en moyenne sur la durée de vie d'un client, c'est à dire la quantité d'argent qu'il va rapporter à l'entreprise durant la période où il sera client de l'enseigne, dans le cadre d'une relation non contractuelle (le client est libre de ne plus acheter dans l'enseigne dès qu'il le souhaite). Le vendeur a donc intérêt à fidéliser les clients ayant une VVC élevée. Dans ce modèle, le vendeur ne connait pas et ne peut pas observer l'espérance de vie du client. Il ne voit que son dernier achat, mais ne sait pas quand le client a cessé de vouloir acheter dans son enseigne, ni pourquoi. Ce modèle est crédible dans le cas d'achats de produits de grande consommation, par exemple, pour lesquels les coûts de changement sont faibles. Dans la grande distribution, l'utilisation de cartes de fidélité apporte de plus en plus d'information et implique une gestion du client orientée vers le long terme. La VVC est alors un instrument de mesure permettant de cibler les clients potentiellement les plus intéressants. Quoiqu'il en soit, la prévision de la probabilité de défection et de la durée de vie du client est une donnée essentielle en marketing.

Il existe plusieurs formules permettant de calculer la VVC ν. La plus simple est

$$\nu = \mathrm{VC} \times \mathrm{T} = \mathrm{VMC} \times \mathrm{F} \times \mathrm{T} \tag{1}$$

Où VC est la valeur du client, VMC sa valeur moyenne, F la fréquence d'achat et T l'espérance de vie résiduelle du client.

Mais cette formule ne prend pas en compte l'évolution dans le temps du comportement du client. Une formule moins grossière de la VVC est

$$\nu = \sum_{t=1}^{\mathrm{T}} \frac{f_t \times m_t}{(1+d)^t} \tag{2}$$

où f_t est la fréquence d'achat moyenne à la date t et m_t la valeur moyenne des achats réalisés pendant cette période. d représente l'inflation et la formule détermine la valeur présente des achats futurs en neutralisant cette inflation.

Le premier problème des deux formules précédentes est que chacune des quantités en jeu est inconnue et doit être être estimée à partir des données. Le second est que la VVC n'a d'intérêt que si elle est calculée de façon agrégée pour l'ensemble des clients, alors qu'elle définie par des données individuelles, variant beaucoup d'un client à l'autre. Il est nécessaire d'en chercher une expression globale sur l'ensemble des clients, prenant en compte cette variabilité.

Si le sujet vous intéresse, vous pouvez approfondir le contenu du TP en lisant les articles [3, 4, 2, 1].

Dans toute la suite, une barre verticale | signifie « sachant que ».

2 Calculs préliminaires

2.1 La loi binomiale négative

On rappelle que la loi binomiale négative de paramètres $r > 1$ et $p \in]0,1[$ est la loi discrète sur $\mathbb{N}$ définie par

$$\mathbb{P}[\mathrm{X} = k] = \frac{\Gamma(r+k)}{\Gamma(r)k!} p^r q^k \tag{3}$$

Lorsque r est un nombre entier, on rappelle que $\Gamma(r+k) = (r+k-1)!$ et $\Gamma(r) = (r-1)!$. Cette loi donne le nombre d'échecs X nécessaires avant d'obtenir r succès dans une succession d'expériences de Bernoulli indépendantes avec une probabilité de succès p (la probabilité d'échec est notée $q = 1-p$).

1°. On suppose r entier. On a alors $\Gamma(r+k) = (r+k-1)!$ et $\Gamma(r) = (r-1)!$.Démontrer que la fonction génératrice des moments d'une loi binomiale négative est

$$\mathrm{G}(t) = \left(\frac{p}{1-qt}\right)^r \tag{4}$$

En déduire que

$$\mathbb{E}[\mathrm{X}] = \frac{rq}{p} \text{ et } \mathbb{V}(\mathrm{X}) = \frac{rq}{p^2} \tag{5}$$

2°. Démontrer que les estimateurs des moments de la loi binomiale négative sont donnés par

$$\widehat{r} = \frac{\overline{\mathrm{X}}^2}{\mathrm{S}^2 - \overline{\mathrm{X}}} \text{ et } \widehat{p} = \frac{\overline{\mathrm{X}}}{\mathrm{S}^2} \tag{6}$$

où $\overline{\mathrm{X}}$ et S^2 sont respectivement la moyenne et la variance empirique d'un échantillon.

3°. Déterminer une équation dont la résolution numérique permettrait de trouver les estimateurs du maximum de vraisemblance de la loi binomiale négative.

4°. Soit X une loi de Poisson dont le paramètre Λ est aléatoire et suit une loi gamma de paramètres r, θ et de densité

$$f_\Lambda(\lambda) = \frac{\lambda^{r-1}}{\Gamma(r)\theta^r} e^{-\lambda/\theta} \mathbb{1}_{[0,\infty[}(\lambda) \tag{7}$$

Démontrer que la loi de X est une loi binomiale négative dont on déterminera le paramètre p en fonction de θ. On parle alors de mélange Gamma-Poisson.

2.2 La loi de Pareto

Économiste et sociologue italien né à Paris en 1848, Vilfredo Pareto est à l'origine de la loi de probabilité que présentons ici. Alors titulaire de la chaire d'économie politique de l'université de Lausanne (il succède à Léon Walras), Pareto s'intéresse à la distribution et à la répartition des revenus dans les différents pays d'Europe.

Disposant des données fiscales pour la France, l'Angleterre, la Suisse, l'Italie, la Russie et la Prusse, il remarque que les inégalités de revenus varient fortement d'un pays à l'autre, mais il met également en lumière une régularité statistique remarquable, vérifiée dans tous les pays pour lesquels il dispose de données. Dans son « essai sur la courbe de la répartition de la richesse » publié en 1896, il écrit : « nous indiquerons par x un certain revenu, et par N le nombre de contribuables ayant un revenu supérieur à x (...). Traçons deux axes (AB) et (AC). Sur (AB) portons les logarithmes de x, sur (AC) les logarithmes de N. Il ressort une relation tout à fait linéaire. » De ce constat empirique, l'auteur en déduit la relation mathématique suivante :

$$\log(\mathrm{N}) = \mathrm{B} - \alpha \times \log(x) \Leftrightarrow \mathrm{N} = \frac{\mathrm{A}}{x^\alpha} \tag{8}$$

Avec $\mathrm{B} = \log(\mathrm{A})$. Finalement, selon Pareto, le pourcentage de la population dont la richesse est supérieure à une valeur x est toujours proportionnelle à $\mathrm{A} \div x^\alpha$. C'est le paramètre α qui varie entre les différents pays et explique des différences dans la distribution des revenus.

Aujourd'hui, la loi de Pareto est encore couramment utilisée en économie ou en sociologie pour étudier les inégalités de revenus dans nos sociétés. Elle a également fait l'objet de multiples applications en gestion des risques, actuariat, dans le domaine du management des entreprises ou dans la gestion des flux de données sur internet.

La densité d'une loi de Pareto $\mathscr{P}(\alpha, c)$ est donnée par

$$f(x) = \frac{\alpha c^\alpha}{x^{\alpha+1}} \mathbb{1}_{[c,+\infty[}(x) \tag{9}$$

5°. Déterminer sa fonction de répartition.

6°. Calculer $\mathbb{E}[\mathrm{X}]$ et $\mathbb{V}(\mathrm{X})$.

7°. Soit $\mathrm{X} = (\mathrm{X}_1, ..., \mathrm{X}_n)$ un n-échantillon de Pareto. La notation $\rightsquigarrow$ signifie « converge en loi lorsque n tend vers l'infini ». On pose

$$\mathrm{F}(x) = \mathbb{P}[\mathrm{X}_i \leq x] \tag{10}$$

$\forall\ i = 1, ..., n$. Si $\mathrm{M}_n = \max(\mathrm{X}_1, ..., \mathrm{X}_n)$ et si F_n est la fonction de répartition de M_n, déterminer le lien entre F_n et F.

8°. Démontrer que l'estimateur de c par la méthode du maximum de vraisemblance est

$$\widehat{c} = \min_{i=1}^{n} \mathrm{X}_i \tag{11}$$

et déterminer sa loi.

9°. On suppose maintenant c connu. Démontrer que l'estimateur du maximum de vraisemblance de α est

$$\widehat{\alpha}_n = \left(\frac{1}{n} \sum_{i=1}^{n} \ln \frac{\mathrm{X}_i}{c} \right)^{-1} \tag{12}$$

10°. Montrer que $\mathrm{Y}_i = \ln(\mathrm{X}_i / c)$ suit une loi exponentielle de paramètre α.

11°. Déterminer la loi de

$$\mathrm{T}_n = \sum_{i=1}^{n} \ln \frac{\mathrm{X}_i}{c} \tag{13}$$

12°. Déterminer l'espérance et la variance de $\widehat{\alpha}_n$ et en déduire un estimateur sans biais $\alpha_n^\star$ de α. Calculer sa variance.

13°. Montrer que T_n est une statistique exhaustive et que cette statistique est complète. En déduire que $\alpha_n^\star$ est l'estimateur VUMSB de α.

14°. Montrer que $\widehat{\alpha}_n$ et $\alpha_n^\star$ sont des estimateurs consistants de α. Déterminer la loi limite de $\sqrt{n}(\widehat{\alpha}_n - \alpha)$ et $\sqrt{n}(\alpha_n^\star - \alpha)$.

15°. Montrer que l'estimateur des moments de α (lorsque $\alpha > 2$) est

$$\overline{\alpha}_n = \frac{\overline{\mathrm{X}}}{\overline{\mathrm{X}} - c} \tag{14}$$

où $\overline{\mathrm{X}}$ est la moyenne empirique de l'échantillon. Calculer la loi limite de $\sqrt{n}(\overline{\alpha}_n - \alpha)$.

16°. Soit τ une variable aléatoire de loi (conditionnelle) exponentielle dont le paramètre λ est aléatoire et suit une loi gamma de paramètres (s, β). Démontrer que la loi de τ est une loi de Pareto dont on précisera les paramètres.

2.3 Étude théorique du modèle NBD

Le modèle de calcul de la VVC le plus simple et le plus ancien est le modèle NBD (loi binomiale négative), proposé par Ehrenberg en 1959. Il se fonde sur deux hypothèses :

• Le nombre d'achats X_t réalisés par un client durant une période $]0, t]$ est une variable aléatoire de Poisson de paramètre λt :

$$\mathbb{P}[X_t = x] = \frac{(\lambda t)^x}{x!} e^{-\lambda t} \tag{15}$$

$x \in \mathbb{N}$.

17°. Calculer $\mathbb{E}[X_t|\lambda, t]$ et $\mathbb{V}[X_t|\lambda, t]$.

Le modèle suppose que les achats se font sans mémoire et se produisent indépendamment les uns des autres. On estime que cette hypothèse est crédible pour des biens de grande consommation, peu onéreux.

18°. Démontrer que le temps $\Delta_t = t_k - t_{k-1}$ entre deux achats $k-1$ et k, suit une loi exponentielle dont on précisera le paramètre et que la probabilité qu'un achat ait lieu à l'instant t_k sachant que le précédent a eu lieu à l'instant t_{k-1} a pour densité

$$f(t_k|t_{k-1}, \lambda) = \lambda e^{-\lambda \Delta_t} \mathbb{1}_{]0,\infty[}(t_k) \mathbb{1}_{[0,\infty[}(t_{k-1}) \tag{16}$$

Le paramètre λ représente la fréquence d'achat (le nombre d'achat par unité de temps).

• On suppose que ce paramètre est en fait une variable aléatoire Λ dont les réalisations λ sont des fréquences qui varient en fonction des clients. On le modélise Λ par une loi gamma de densité

$$f_\Lambda(\lambda) = f(\lambda) = \frac{\alpha^r}{\Gamma(r)} \lambda^{r-1} e^{-\alpha\lambda} \mathbb{1}_{[0,\infty[}(\lambda) \tag{17}$$

Le choix de la loi Gamma vient de sa souplesse et du fait qu'elle rende compte de plusieurs niveaux d'hétérogénéité dans le comportement des clients.

19°. Montrer que

$$\mathbb{E}[\Lambda|r, \alpha] = \frac{r}{\alpha} \text{ et } \mathbb{V}[\Lambda|r, \alpha] = \frac{r}{\alpha^2} \tag{18}$$

Le coefficient r est un indicateur d'homogénéité de la fréquence d'achat. Le fait que λ ne varie pas dans le temps signifie que les marchés sont stationnaires.

20°. Montrer que la loi de X_t sachant r et α (et donc sachant $\Lambda = \lambda$) vérifie :

$$\mathbb{P}[X_t = x|r, \alpha] = \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha + t}\right)^r \left(\frac{t}{\alpha + t}\right)^x \tag{19}$$

En déduire que l'espérance du nombre d'achats réalisés durant une période de durée t vaut

$$\mathbb{E}[X_t|t, r, \alpha] = \frac{rt}{\alpha} \tag{20}$$

2.4 Le modèle Pareto/NBD

En 1987, Schmittlein, Morrison et Colombo ont proposé un modèle plus réaliste modélisant l'attrition du client à l'aide d'une variable aléatoire. Ce modèle est compatible avec la méthode d'analyse client marketing appelée RFM (récence, fréquence, montant). Le modèle repose sur cinq hypothèses :

• Lorsqu'il est actif, le nombre d'achats d'un client durant une durée t suit une loi de Poisson de paramètre λt .

• λ est aléatoire et suit une loi gamma de paramètre d'échelle α et de paramètre de forme r.

• Chaque client a une espérance de vie (qui représente la durée de consommation dans l'enseigne) de durée τ. Cette durée est aléatoire et suit une loi exponentielle de paramètre μ, dont la densité est donnée par

$$g(\tau) = \mu e^{-\mu\tau} \mathbb{1}_{[0,\infty[}(\tau) \tag{21}$$

• μ est également aléatoire et suit une loi gamma de paramètre de forme s et de paramètre d'échelle β.

En toute rigueur, on devrait noter μ majuscule cette variable aléatoire et μ la valeur d'une observation (et c'est la même chose pour λ et Λ). En fait, on notera à la fois μ la variable aléatoire et sa réalisation et il faudra faire attention au contexte. Celui-ci sera dans tous les cas précisé dans l'évènement par la barre verticale | ; sachant μ signifiera une réalisation de μ (idem pour λ). La densité de μ est donc donnée par :

$$h(\mu) = \frac{\beta^s}{\Gamma(s)} \mu^{s-1} e^{-\beta\mu} \mathbb{1}_{[0,\infty[}(\mu) \tag{22}$$

• Les variables aléatoires λ et μ sont indépendantes.

On a ainsi :

$$\mathbb{P}[X_T = x|\lambda, \tau > T] = e^{-\lambda T} \frac{(\lambda T)^x}{x!} \tag{23}$$

et l'on en déduit :

$$\mathbb{P}[X_T = x|r, \alpha, \tau > T] = \frac{\Gamma(r+x)}{\Gamma(r)x!} \left(\frac{\alpha}{\alpha + T}\right)^r \left(\frac{T}{\alpha + T}\right)^x \tag{24}$$

Autrement dit, X_T suit une loi une loi binomiale négative. Il est important de noter la condition $[\tau > T]$ (le client est encore actif à T) car l'espérance de vie n'est pas observée. On a aussi :

$$g(\tau|s, \beta) = \frac{s}{\beta} \left(\frac{\beta}{\beta + \tau}\right)^{s+1} \mathbb{1}_{[0,\infty[}(\tau) \tag{25}$$

Autrement dit, τ suit une loi de Pareto et $\mathbb{E}[\tau|s, \beta] = \beta/(s-1)$, avec $s > 1$.

Le comportement observé de chaque client c peut se résumer par un triplet $c = [X = x, t_x, T]$ représentant le nombre $X = x$ d'achats observés pendant la durée $]0, T]$ et la date $t_x \in]0, T]$ du dernier achat (qui dépend de x). Fréquence (modélisée par $X = x$) et récence (modélisée par t_x) forment une statistique exhaustive pour prédire le comportment futur d'un client. En marketing, on appelle cela l'analyse RFM (recency, frequency, monetary value).

Les grandeurs d'intérêt, à partir desquelles on pourra calculer la VVC, sont alors :

• $\mathbb{E}[X_T]$, nombre moyen d'achats d'un client pendant une période de durée T.

• $\mathbb{P}[\tau > T|X = x, t_x, T]$, probabilité qu'un client soit actif, sachant $[X = x, t_x, T]$. Nous noterons parfois A (pour actif) l'évènement $[\tau > T]$.

• $\mathbb{E}[Y_t|X = x, t_x, T]$, nombre moyen d'achats dans la période future $]T, T + t]$ pour un client c caractérisé par un comportement observé $[X = x, t_x, T]$.

Ces grandeurs, qui sont des données individuelles, devront ensuite être agrégées pour calculer une CLV globale à l'ensemble des clients.

Les questions suivantes sont plus difficiles. Les résultats peuvent être admis si vous souhaitez passer directement aux simulations et aux données réelles.

21°. Selon que le client est toujours actif à la fin de la période d'observation (c'est à dire $\tau > T$) ou pas, montrer que la vraisemblance du paramètre Λ s'écrit :

$$L(\lambda|t_1, ..., t_x, T, \tau > T] = \lambda^x e^{-\lambda T} \quad (26)$$

$$L(\lambda|t_1, ..., t_x, T, \tau \leq T] = \lambda^x e^{-\lambda \tau} \quad (27)$$

La donnée de $t_1, ..., t_x, T$ est équivalente à la donnée $[X = x, t_x, T]$ dès lors que $t_x = 0$ quand $x = 0$.

22°. Déduire que

$$L(\lambda, \mu|X = x, t_x, T) = L(r, \alpha, s, \beta|X = x, t_x, T] = \frac{\lambda^x \mu}{\lambda + \mu} e^{-(\lambda+\mu)t} + \frac{\lambda^{x+1}}{\lambda + \mu} e^{-(\lambda+\mu)T}$$

L'étape suivante consiste à supprimer le conditionnement en λ et μ. On admet alors qu'en intégrant par rapport à ses deux paramètres, la vraisemblance peut s'écrire sous la forme

$$L(r, \alpha, s, \beta|X = x, t_x, T] = \int_0^{+\infty} \int_0^{+\infty} L(\lambda, \mu|X = x, t_x, T) f(\mu) f_\Lambda(\lambda) d\lambda d\mu = \frac{\Gamma(r + x)\alpha^r \beta^s}{\Gamma(r)} \times \left[\frac{1}{(\alpha + T)^{r+x}(\beta + T)^s} + \left(\frac{s}{r + s + x} \right) A_0 \right]$$

où A_0 est une fonction gaussienne hypergéométrique dépendant de tous les paramètres, dont nous ne préciserons pas la forme exacte.

23°. Le nombre moyen d'achats pendant la période $]0, T]$ est λT si $\tau > T$. Par contre, si $\tau \leq T$, le nombre moyen d'achats est $\lambda\tau$. En prenant en compte ces deux cas, montrer que

$$\mathbb{E}[X_T|\lambda, \mu] = \frac{\lambda}{\mu} \left(1 - e^{-\mu t}\right) \quad (28)$$

En intégrant ensuite par rapport à λ et μ, montrer que :

$$\mathbb{E}[X_T] = \mathbb{E}[X_T|r, \alpha, s, \beta] = \frac{r\beta}{\alpha(s-1)} \left[1 - \left(\frac{\beta}{\beta + T} \right)^{s-1} \right] \quad (29)$$

24°. En utilisant le théorème de Bayes, montrer que

$$\mathbb{P}[\tau > T|\lambda, \mu, X = x, t_x, T] = \frac{L(\lambda|x, T, \tau > T)\mathbb{P}[\tau > T|\mu]}{L(\lambda, \mu|X = x, t_x, T)} = \frac{\lambda^x e^{-(\lambda+\mu)T}}{L(\lambda, \mu|X = x, t_x, T)}$$

puis en déduire que

$$\mathbb{P}[\tau > T|\lambda, \mu, X = x, t_x, T] = \frac{1}{1 + \frac{\mu}{\lambda+\mu} \times \left[e^{(\lambda+\mu)(T-t_x)} - 1 \right]}$$

Pour un client choisi au hasard, dont le comportement d'achat passé est $[X = x, t_x, T]$, montrer que la probabilité qu'il soit toujours actif à T est :

$$\mathbb{P}[A|r, \alpha, s, \beta, X = x, t_x, T] = \quad (30)$$

$$\int_0^{+\infty} \int_0^{+\infty} \mathbb{P}[A|\lambda, \mu, X = x, t_x, T] f(\lambda, \mu) d\lambda d\mu \quad (31)$$

L'évènement $A = [\tau > T]$ signifie que le client est actif à la date T et $f(\lambda, \mu)$ est la densité jointe du couple de variables aléatoires λ et μ sachant les paramètres (r, α, s, β) et l'historique de consommation $[X = x, t_x, T]$ du client. On admet que cette probabilité peut se mettre sous la forme suivante :

$$\mathbb{P}[A|r, \alpha, s, \beta, X = x, t, T] = \quad (32)$$

$$\left[1 + \left(\frac{s}{r + s + x} \right) (\alpha + T)^{r+x} (\beta + T)^s A_0 \right]^{-1} \quad (33)$$

où A_0 est une fonction gaussienne hypergéométrique dont nous ne donnerons pas la form exacte.

25°. Au moment d'agréger les résultats de tous les clients, on estime les 4 paramètres (r, α, s, β) du modèle par la méthode du maximum de vraisemblance en utilisant la fonction de log-vraisemblance de tout l'échantillon des N clients :

$$l(r, \alpha, s, \beta) = \sum_{i=1}^{N} \ln L(r, \alpha, s, \beta|X_i = x_i, t_{x_i}, T_i)$$

Il reste finalement à déduire une estimation de la VVC. La formule exacte prenant en compte l'aspect probabiliste et agrégé du modèle serait de la forme :

$$\mathbb{E}[CLV] = \int_0^{+\infty} \mathbb{E}[X_t|\tau > t]\mathbb{P}[\tau > t]\delta(t) dt \quad (34)$$

$\mathbb{E}[X_t|\tau > t]$ représente la valeur du client. Pour nous, c'est le nombre d'achats, qu'il faudrait éventuellement multiplier par la valeur moyenne des achats pour un modèle prenant en compte cette grandeur (partie M de RFM).

$\mathbb{P}[\tau > t]$ est la probabilité qu'un client soit toujours actif à t.
$\delta(t)$ est la valeur présente de l'argent disponible au temps t. Ce facteur intègre donc l'inflation.

Mais cette formule doit être approchée pour prendre en compte l'aspect discret des dates d'achats et de façon générale des statistiques de transactions. On peut poser :

$$\mathbb{E}[\text{CLV}] = \sum_{t=0}^{+\infty} \mathbb{E}[X_t | \tau > t] \mathbb{P}[\tau > t] \delta(t) dt \qquad (35)$$

Finalement, la VVC peut-être approchée par la formule suivante (en convenant toujours que l'on s'intéresse uniquement à la fréquence des achats et pas à leur montant) :

$$\text{CLV} = \sum_{t=0}^{+\infty} \frac{\mathbb{E}[Y_t - Y_{t-1} | r, \alpha, s, \beta, X = x, t_x, T]}{(1+d)^t} \qquad (36)$$

Pas de panique concernant les formules précédentes : elles sont admises et tous les calculs les concernant seront effectués de façon numérique en utilisant des fonctions du package BTYD sous R ou lifetimes sous Python.

3 Étude empirique

3.1 Données simulées

Cette première partie empirique se propose de modéliser les lois de probabilités introduite dans la section précédente. Les librairies Python utiles sont les suivantes : `numpy`, `pandas`, `matplotlib.pyplot`, `scipy`, `scipy.stats`.

On va simuler le comportement de N clients effectuant des achats durant deux ans, avec un découpage de temps hebdomadaire de 104 semaines. Chaque client c sera caractérisé par une date d'arrivée T_0 correspondant à son premier achat, et par les paramètres τ et X_t définis précédemment via le modèle Pareto / NBD.

1°. Nous supposons que les nouveaux clients arrivent dans l'enseigne de façon uniforme au cours de ces deux ans. À l'aide de la fonction `uniform` de `scipy.stats`, créer un vecteur `start` de N coordonnées représentant les différentes dates d'entrées des N clients. En déduire un vecteur T dont les coordonnées T_c indiquent, pour chaque client c, la période entre la date T_0 du premier achat et la date commune de fin d'observation de tous les clients, à l'issue des deux ans.

2°. Déterminer un vecteur μ représentant la réalisation de N tirages d'une loi gamma (r, α) puis en déduire un vecteur τ dont les coordonnées sont les espérances de vie des N clients.

3°. De la même façon, déterminer un vecteur λ représentant la réalisation de N tirages d'une loi gamma (s, β), dont les coordonnées sont différentes valeurs de fréquence d'achat de chacun des N clients.

4°. Pour un client c donné, l'intervalle de temps entre deux achats consécutifs suit une loi exponentielle de paramètre λ_c (λ_c est une des coordonnées du vecteur λ et correspond à une réalisation d'une simulation de la loi $\Gamma(r, \alpha)$). Afin de déterminer le nombre d'achats effectués durant une durée donnée, écrire une fonction générant les durées aléatoires entre les achats consécutifs et comptant le nombre d'achats effectués par un client entre sa date d'entrée T_0 et la date T. La fonction devra retourner comme valeur l'instant t du dernier achat ainsi que le nombre d'achats effectués entre T_0 (propre à chaque client) et T (date de fin d'observation commune à tous les clients).

5°. Sachant μ et λ, la VVC se calcule de la façon suivante : l'espérance de vie résiduelle δ du client (sa « Remaining Lifetime Value » ou RLV) est la différence entre son espérance de vie et son âge (l'âge est la différence entre l'instant présent et la date T_0). La VVC est égale au nombre d'achats effectués durant cette espérance de vie résiduelle (on rappelle qu'on ne tient pas compte de la valeur de chaque achat). On la simule par une loi de Poisson X de paramètre $\lambda\delta$. En donnant plusieurs valeurs de plus en plus grandes à N, simuler des espérances de vie résiduelles et des VVC.

6°. Visualiser les histogrammes de différents vecteurs τ, λ et δ, reconnaître leurs lois et estimer leurs paramètres.

7°. Estimer les paramètres des lois binomiales négatives et des lois de Pareto obtenues.

3.2 Données réelles : utilisation de la librairie Python BTYD

Toutes les fonctions permettant la modélisation du modèle BTYD ("Buy Til You Die") sont implémentées dans la librairie `btyd` (lien) qu'il faut installer à l'aide de l'instruction `pip install btyd`.

Les données à étudier proviennent d'une base appelée CDNOW qui contient les achats d'un échantillon de 2357 clients de la société de vente en ligne CDNOW, sur deux périodes consécutives de 39 semaines (fichiers `cdnow_data.xls` et fichier `p2x`). L'unité de temps est donc la semaine.

Pour chaque client c, on dispose donc de l'information $[X = x, t_x, T]$ représentant le nombre d'achats effectués, la date du dernier achat et le temps d'observation entre son premier achat et la fin des 39 premières semaines.

8°. Décrire les données, les visualiser et les mettre en forme en utilisant les fonctions du package `BTYD` («Buy til you die»). Vous pouvez vous inspirer du document « BTYD - a Walkthrough » qui décrit les différentes fonctions disponibles et les modalités

d'utilisation.

9°. Estimer numériquement les paramètres (r, α, s, β) du modèle Pareto / NBD par la méthode du maximum de vraisemblance.

10°. Estimer les paramètres des lois binomiales négatives et de Pareto sous-jacentes au modèle, par la méthodes des moments. Comparer avec les résultats de la question précédente.

11°. À l'aide des données des 39 premières semaines, donner les prévisions d'achats pour les dernières 39 semaines et vérifier ces prévisions avec les données empiriques.

3.3 Données réelles en Python

La bibliothèque équivalente à BTYD en Python s'appelle lifetimes. Après l'avoir installée et avoir invoqué le package, vous pouvez compléter le notebook initié en séance de TP et répondre aux mêmes questions que la partie R.

Références

[1] McCarthy Daniel and Wadsworth Edward. Buy til you die - a walktrhough. https://cran.r-project.org/web/packages/BTYD/vignettes/BTYD-walkthrough.pdf.

[2] Lukasz Dziurzynski, Wadsworth Edward, and McCarthy Daniel. Btyd : Implementing buy til you die models. http://CRAN.R-project.org/package=BTYD.Rpackageversion2.4.

[3] Peter S. Fader and Bruce G. S. Hardie. A note on deriving the pareto/nbd model and related expressions. 2005.

[4] David C. Schmittlein, Donald G. Morrison, and Richard Colombo. Counting your customers : Who-are they and what will they do next? *Manage. Sci.*, 33(1) :124, jan 1987.