ECONOMETRIE DES DONNEES DE SURVIE

MASTER 2 ECONOMETRIE ET STATISTIQUE APPLIQUEE

ANALYSE DES DETERMINANTS DE RENOUVELLEMENT DE L'ACHAT D'UN PRODUIT

Enseignant

Prof Jude EGGOH

Etudiants

Fataï OSSENI

Achille ADJIKPE

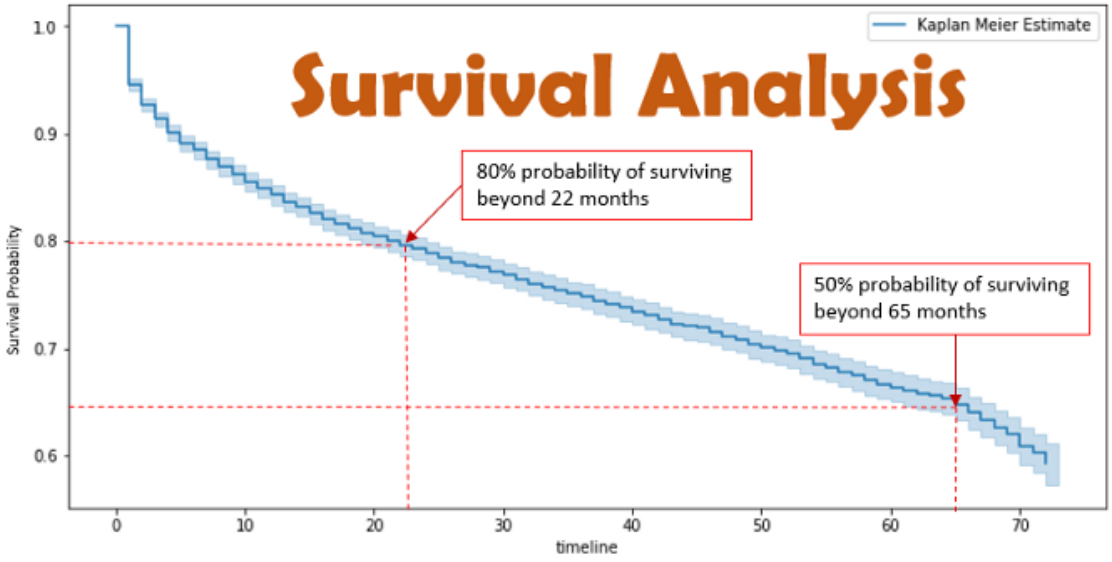
Table des matières

1 Résumé

L'objectif de notre étude est d'analyser les déterminants de renouvellement de l'achat d'un produit et d'étudier l'influence de certains covariables sur cet évènement d'intérêt. Dans toutes les approches (non paramétrique, semi paramétrique et paramétrique) les variables telles que enfant, sexe, couple influencent positivement la survie des individus c'est-à-dire contribuent à accroitre le risque de renouvellement de l'achat des individus. Par contre la variable resid n'est pas significative. Par conséquent, on a un meilleur ciblage des variables pour des actions plus efficaces en vue d'une augmentation du renouvellement de l'achat des clients.

Mots clés : Estimateur de Kaplan Meir, résidu de Cox-Snell, résidus de Schoenfeld, fonction survie, fonction de risque.

2 INTRODUCTION



Le suivi des clients des structures commerciales est très importante pour la pérennité de ces structures. Face aux risques de faire faillite, plusieurs sociétés commerciales cherchent à accroitre leurs chiffres d'affaires. Ces structures adoptent plusieurs techniques plus sophistiquées pour prédire le temps de faillite et aussi la durée de désabonnement des clients aux produits et services qu'elles offrent. Au nombre de ces techniques, figurent les lignes de vie. En effet, l'analyse de survie est un ensemble d'approches statistiques utilisées pour déterminer le temps nécessaire pour qu'un événement d'intérêt se produise. Ainsi, il s'agit non seulement d'estimer la durée de renouvellement des clients à l'achat d'un produit dans une structure commerciale mais aussi d'analyser les interactions entre les covariables sur le temps de renouvellement.

L'objectif de notre étude est d'analyser les déterminants de renouvellement de l'achat d'un produit et d'étudier l'influence de certains covariables sur cet évènement d'intérêt.

Pour ce faire, nous allons adopter un plan en quatre parties. D'abord dans la première partie, nous présenterons les variables en jeu et analyserons les données, ensuite dans la deuxième partie, nous ferons une analyse non paramétrique de la survie et enfin la troisième et la quatrième partie, seront respectivement consacrées à l'analyse semi paramétrique et paramétrique

3 Première partie : Présentation des variables et analyse des données

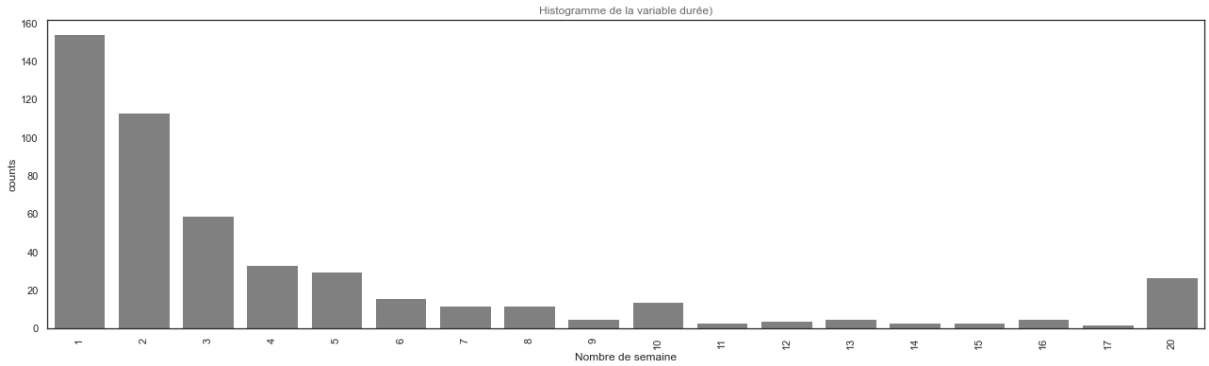
Nous disposons d'un échantillon de 500 clients d'une structure commerciale et qui sont suivis durant 20 semaines suivants leurs premiers achats de renouvellement.

3.1 Analyse exploratoire de la base

3.1.1 Variable Durée

La variable durée correspond au nombre de semaine séparant le premier achat de renouvellement des clients.

La figure ci-après montre la repartions des clients en fonction de la durée de renouvellement de l'achat.



Source : Réalisée par les auteurs

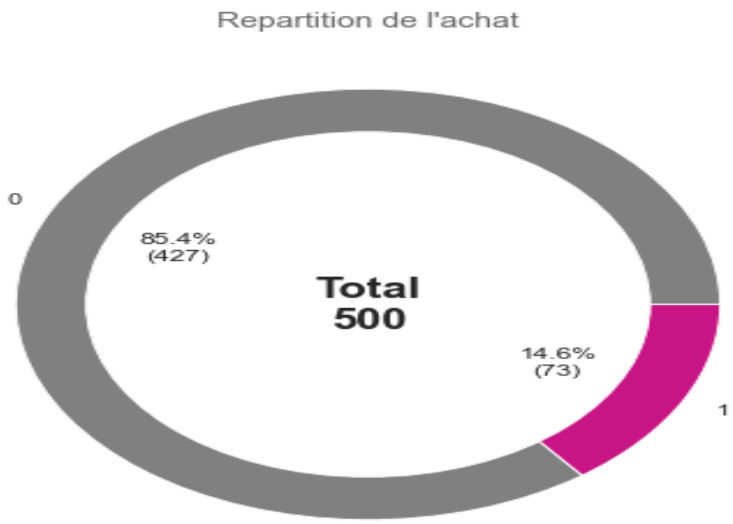
Figure1 : Répartition des individus en fonction de la durée

Cette figure montre que plusieurs clients ont renouvelé leurs achats dans la première semaine après leurs premiers achats. En revanche, une infime partie des clients ont renouvelé leurs achats entre la sixième et la vingtième semaine après le premier achat.

3.1.2 Variable achat

C'est une variable indicatrice qui vaut 1 si l'achat a été renouvelé et 0 sinon.

La figure suivante montre la répartition de la variable achat.



Source : Réalisée par les auteurs

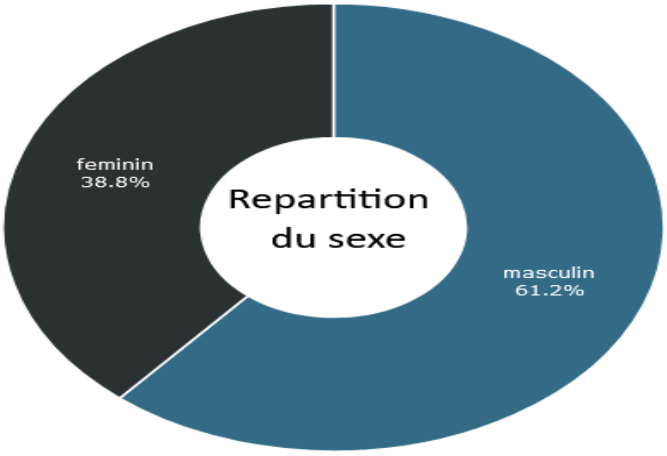
Figure2 : Répartition de l'achat

La figure montre que 85,4% des clients ont renouvelé leurs achats et 14,6% d'entre eux n'ont pas pu le faire. On conclut que 14,6% des clients ont une censure déterministe et cette censure est à droite.

3.1.3 Variable sexe

C'est une variable binaire qui indique le sexe de l'individu.

La figure ci-après présente la répartition du sexe.



Source : Réalisée par les auteurs

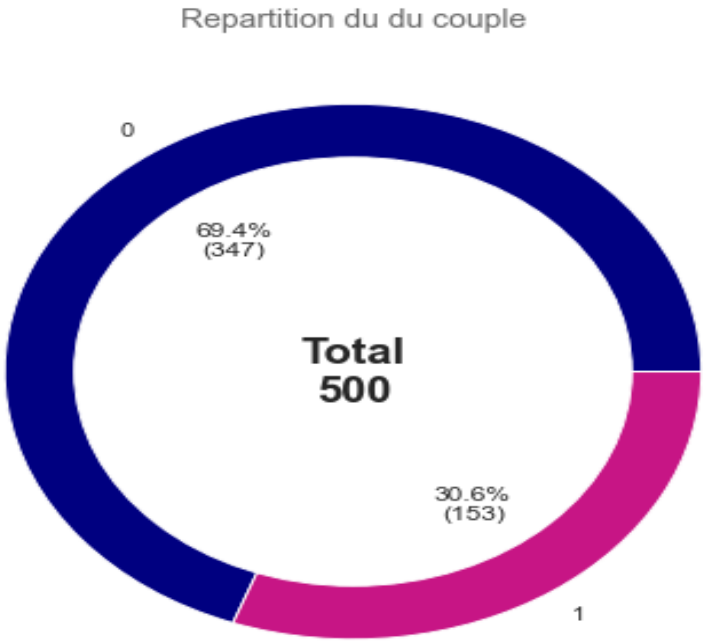
Figure3 : répartition du sexe

Cette figure montre que 61,2% des individus du jeu de donnée sont des femmes et 38,8% sont des hommes. On conclut que les femmes achètent majoritairement ce produit.

3.1.4 Variable Couple

C'est une variable indicatrice qui traduit le statut familial des individus et dont nous avons codé 0 si la personne est célibataire ou vie seule, 1 si elle est en couple.

Le graphique suivant présente la répartition de la variable couple.



Source : Réalisée par les auteurs

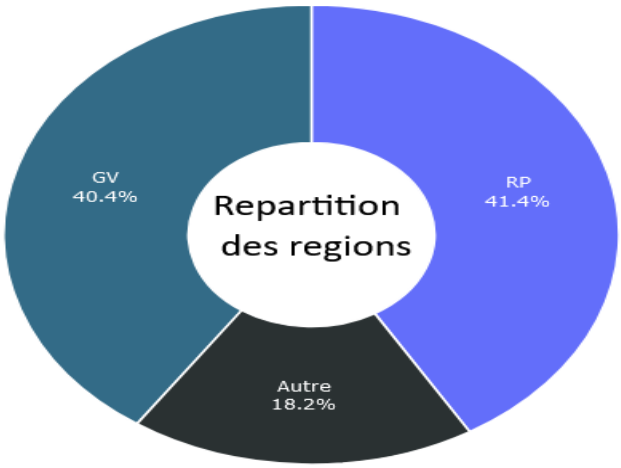
Figure4 : répartition de la variable couple

Cette figure montre que 69,4% des individus de notre base sont des célibataires et 30,6% des individus vivent en couple. On conclut que les célibataires sont fortement représentées.

3.1.5 Variable Resid

C'est une variable catégorielle indiquant le lieu de résidence des clients et qui est codée 'RP' si région parisienne, 'GV' si grandes agglomérations, 'Autre' sinon.

La figure ci-après présente la répartition de la variable Resid.



Source : Réalisée par les auteurs

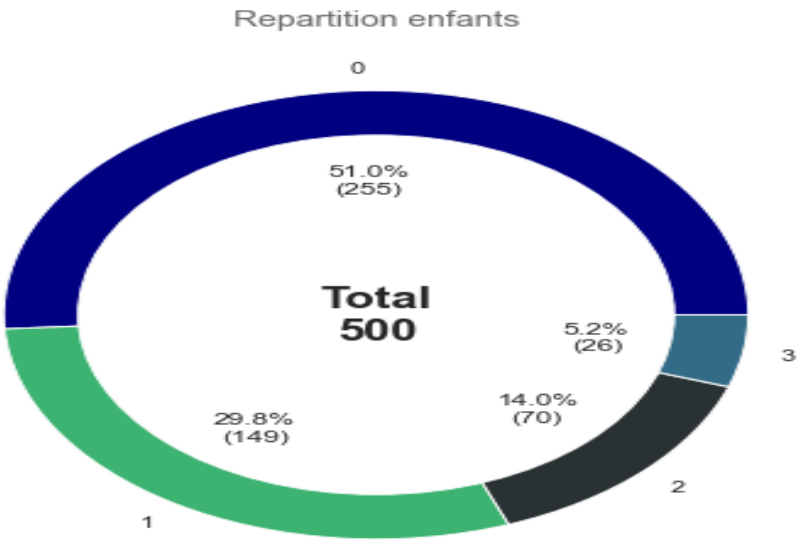
Figure5 : répartition de la variable Resid

La figure montre que 40,4% des clients vivent dans les aggromerations,41,4% des clients vivent dans la région parisienne et 18,2% vivent dans d'autres régions. On conclut que les régions les plus représentées dans cette structure commerciale sont essentiellement la région parisienne et les grandes agglomérations.

3.1.6 Variable enfant

C'est une variable catégorielle qui indique le nombre d'enfants des individus et qui est codée 0 si pas d'enfant, 1 si 1 enfant, 2 si 2 enfants, 3 si 3 enfants et plus.

Le graphique suivant présente la répartition de la variable **enfant**



Source : Réalisée par les auteurs

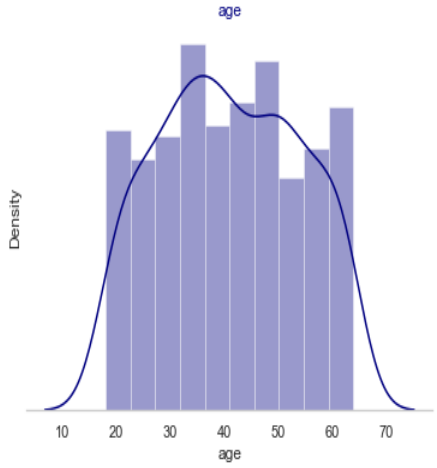
Figure6 : répartition des enfants

La figure montre que 51% des clients n'ont pas d'enfant,29,8% des individus ont un enfant,14% des clients ont deux enfants et 5,2% des clients ont trois et plus enfants. On conclut qu'il y a une forte représentativité des modalités à savoir 0 et 1.

3.1.7 Variable age

C'est une variable qui indique l'âge des individus de notre base et qui est codée en année

3.1.8 Distribution de l'âge

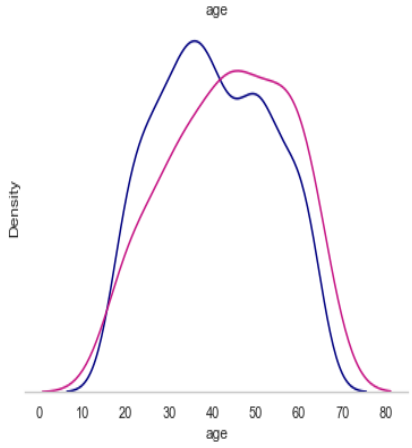


Source : Réalisée par les auteurs

Figure7 : Analyse de distribution de l'âge

Cette figure suggère que l'âge est non normalement distribué. Et donc le renouvellement de l'achat est fonction d'une tranche d'âge.

3.1.9 Distribution conditionnelle l'achat sachant la variable age

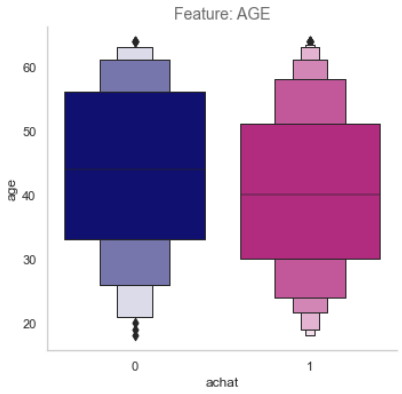


Source : Réalisée par les auteurs

Figure8 : Distribution conditionnelle

Cette figure suggère trois catégories d'âge susceptible d'influencer la variable achat. On distingue la catégorie [0,20[la catégorie [20,50[et la catégorie [50,80[.

3.1.10 Boxplot de l'âge conditionnellement à l'achat.



Source : Réalisée par les auteurs

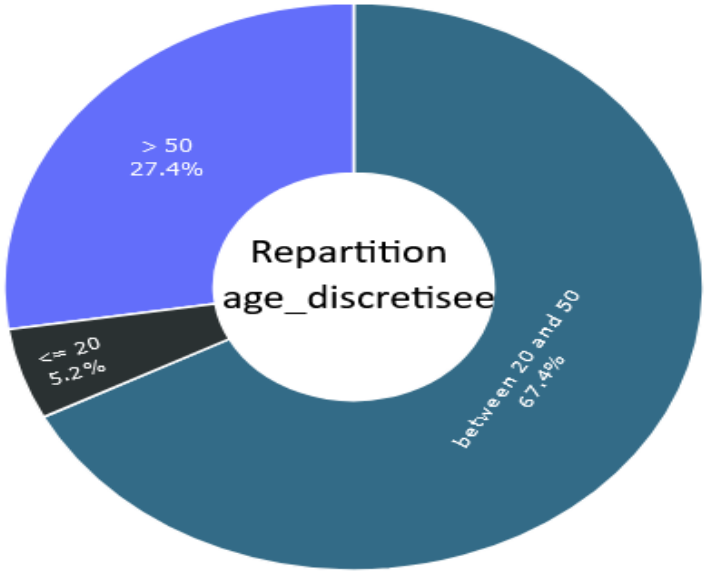
Figure9 : Boite à moustache de l'âge en fonction de l'achat.

La figure montre que les médianes et les variances de l'age conditionnellement à l'achat sont différentes. Donc l'âge peut contribuer à expliquer la variable **achat.**

3.2 Création de la variable age_discretisee

On a constaté au vue des analyses précédentes qu'il existe trois catégories d'âges qui influencent la variable **achat**. En effet, nous avons décidé de remplacer la variable **age** de la base de donnée par **age_discretisee** qui prend en compte les trois catégories d'âge mentionnées. La variable **age_discretisee** est obtenue à travers une discrétisation supervisée dont la variable cible est **achat.**

La figure suivante présente la répartition de la nouvelle variable **age_discretisee.**



Source : Réalisée par les auteurs

Figure10 : répartition de la nouvelle variable age_discretisee.

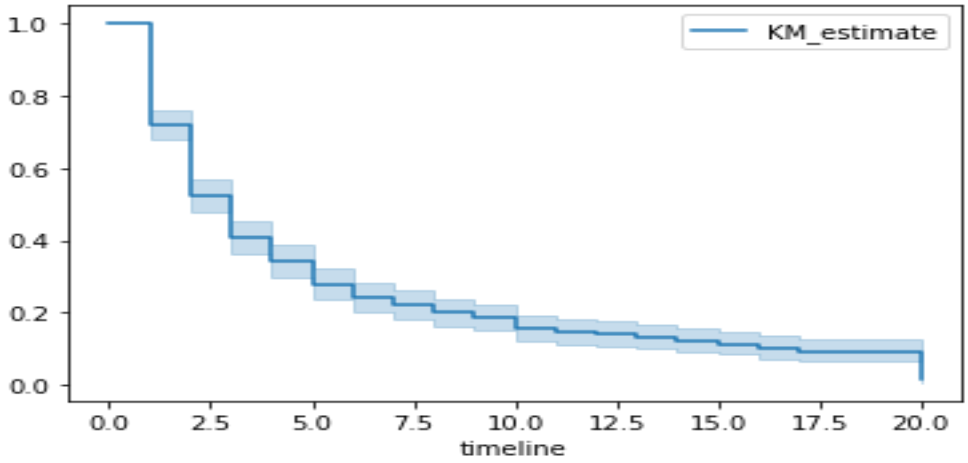
La figure montre que 27,4% des clients ont un age supérieur à 50 ans, 67,4% ont un age compris entre 20ans et 50ans et seulement 5,2% ont un age inferieur à 20ans.On conclut que les catégories d'âge les plus représentées sont essentiellement ceux dont les âges sont supérieurs à 50ans et ceux dont les âges sont compris entre 20ans et 50ans.

4 Deuxième partie : Approche non paramétrique : analyse de la survie et mesure de l'influence des variables explicatives.

4.1 Estimateur de Kaplan-Meier de la fonction de la survie

L'estimateur de Kaplan – Meier est une statistique non paramétrique utilisée pour estimer la fonction de survie (probabilité qu'une personne survit) à partir des données sur la durée de vie. L'estimateur de Kaplan-Meier de la fonction de survie donne une valeur de 0.525 pour le dernier individu non censuré. Ce résultat signifie que la probabilité qu'un individu connaisse l'événement c'est-à-dire renouvelé l'achat à la 20ème semaine est de 52,5%.

4.1.1 Présentation graphique de Kaplan-Meier

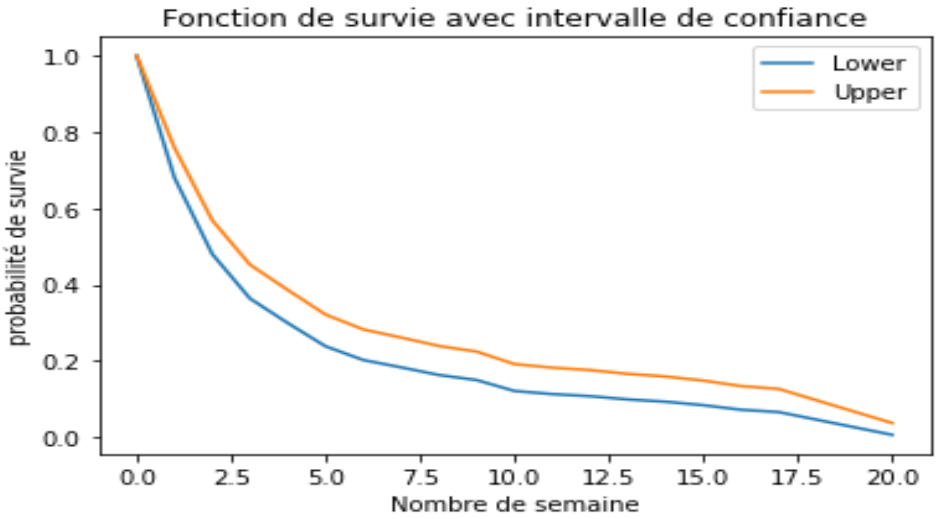


Source : Réalisée par les auteurs

Figure11 : Graphique de la fonction de survie de Kaplan-Meier.

La figure montre que à mesure que la chronologie augmente, la probabilité de survie diminue pour un client. En plus, la figure montre qu'en moyenne, un client a renouvelé son achat trois semaine après le premier achat.

4.1.2 Présentation graphique de la probabilité de survie avec intervalle de confiance



Source : Réalisée par les auteurs

Figure12 : Fonction de survie avec intervalle de confiance

La figure montre la précision et la fiabilité de l'estimateur Kaplan Meier. On conclut que, l'estimateur est globalement stable et constant dans la prédiction de survie.

4.2 Estimation des taux de risque à l'aide de Nelson-Aalen

Les fonctions de survie sont un excellent moyen de résumer et de visualiser le jeu de données de survie, mais ce n'est pas le seul moyen. Si nous sommes curieux de connaître la fonction de danger **(h(t))** d'une population, nous ne pouvons malheureusement pas transformer l'estimation de Kaplan Meier (les statistiques ne fonctionnent pas si bien). Heureusement, il existe un estimateur non paramétrique approprié de la fonction de danger cumulatif :

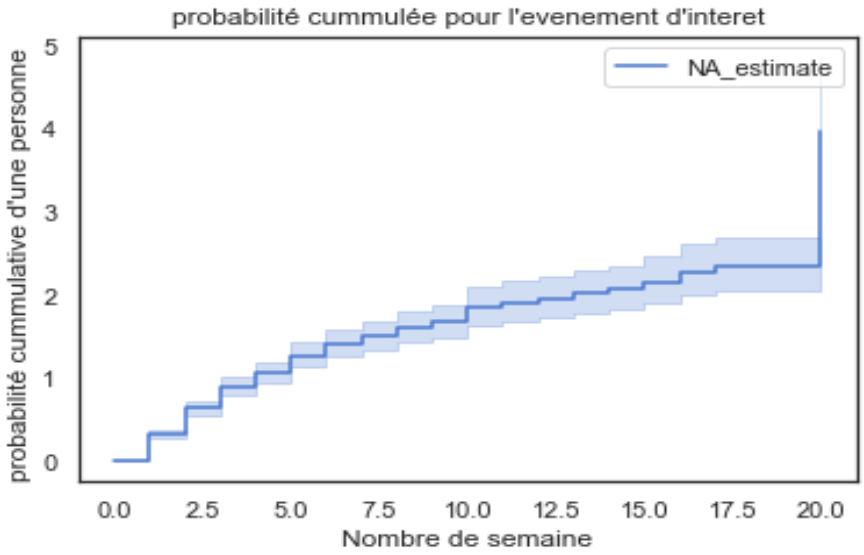
$$\mathbf{H(t)=\int_0^t \lambda(z)\,dz}$$

L'estimateur de cette quantité est appelé l'estimateur de Nelson Aalen :

$$\mathbf{H(t)=\sum_{ti\le t} \frac{di}{ni}}$$

Nous pouvons visualiser les informations agrégées sur la survie en utilisant la fonction de risque de Nelson-**Aalen h (t).** La fonction de hasard **h (t)** nous donne la probabilité qu'un sujet observé au temps t ait un événement d'intérêt (renouvellement de l'achat) à ce moment.

4.2.1 Présentation graphique de Nelson-Aalen

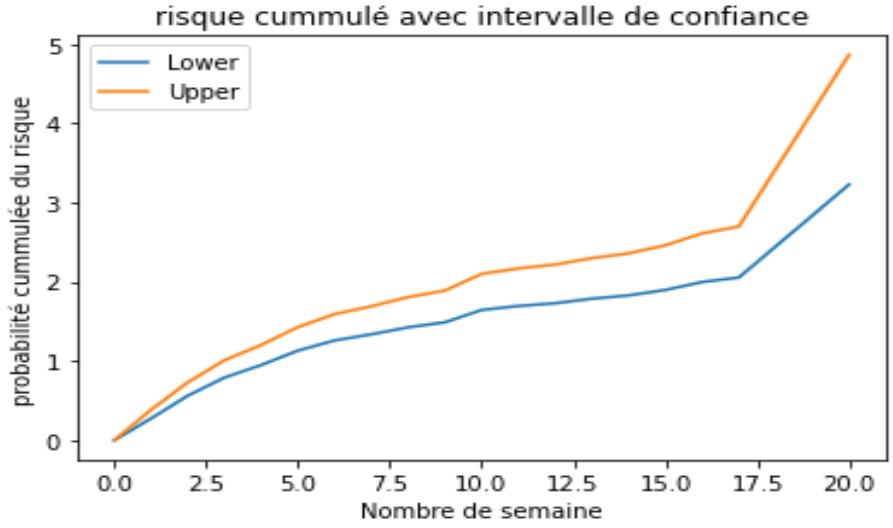


Source : Réalisée par les auteurs

Figure13 : Graphique du taux de risque

La figure montre que à mesure que le nombre de semaine de survie augmente, le risque de renouvellement augmente.

4.2.2 **Présentation graphique** de la probabilité de danger cumulée avec intervalle de confiance



Source : Réalisée par les auteurs

Figure14 : Fonction de risque avec intervalle de confiance

La figure montre que l'estimateur Nelson-Aalen est précis au début et n'est pas stable au cour du temps et pose donc un problème de fiabilité des prédictions.

4.3 Mesure de l'impact des différentes variables explicatives sur la survie

Il sera question dans cette partie de mesurer l'impact de chaque variable explicative sur la survie. Pour les variables qualitatives nous allons constituer des strates d'individus selon les modalités de la variable explicative et ensuite procéder à la comparaison des courbes de survie à l'aide de tests d'hypothèse.

4.3.1 Variable sexe

Il faudra examiner dans cette partie si la variable sexe influence significativement la survie ou pas ; c'est-à-dire si les individus de sexe masculin ont une survie plus longue que ceux du sexe féminin. **Hypothèse nulle** : L'hypothèse nulle indique qu'il n'y a pas de différence significative entre les groupes étudiés. S'il y a une différence significative entre ces groupes, nous devons rejeter notre hypothèse nulle.

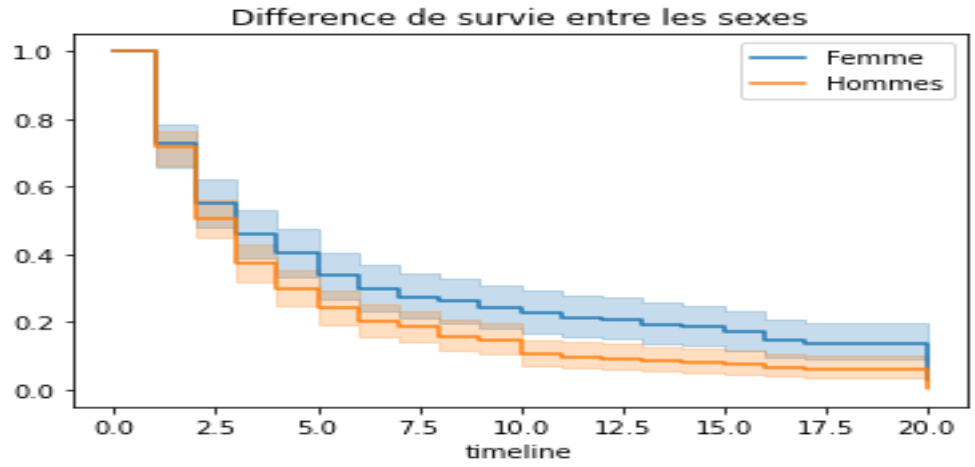
Tableau 1 : Tests d'égalité de la fonction de survie des strates de la variable sexe

Les tests d'hypothèses Log-Rank, Wilcoxon, Tarone-Ware, Peto et Fleming-Harrington donnent les résultats suivants.

Tests	Chi-deux	ddl	Pr> Chi-deux
Log-Rank	7,57	1	0,00699
Wilcoxon	2,47	1	0,1160
Tarone-Ware	4,39	1	0,0361
Peto	3,30	1	0,0694
Fleming-Harrington	10,05	1	0,0015

Notons d'abord que dans ce cas de figure les tests de **Log-Rank, wilcoxon de Tarone-Ware, de Peto, de Fleming-Harrington** donnent approximativement les mêmes résultats.

Ainsi, au seuil d'erreur de 10% quatre tests sur cinq rejettent l'hypothèse nulle d'égalité de la distribution des fonctions de la survie ; donc la fonction de survie des individus de sexe masculin est différentes de celle de sexe féminin. On conclut que le sexe a un effet significatif sur le renouvellement de l'achat**.** Ce résultat est mis en exergue par le graphique illustrant la survie des strates puisqu'il y a un écart entre les courbes de survie



Source : Réalisée par les auteurs

Figure15 : Comparaison de la survie des strates de la variable sexe

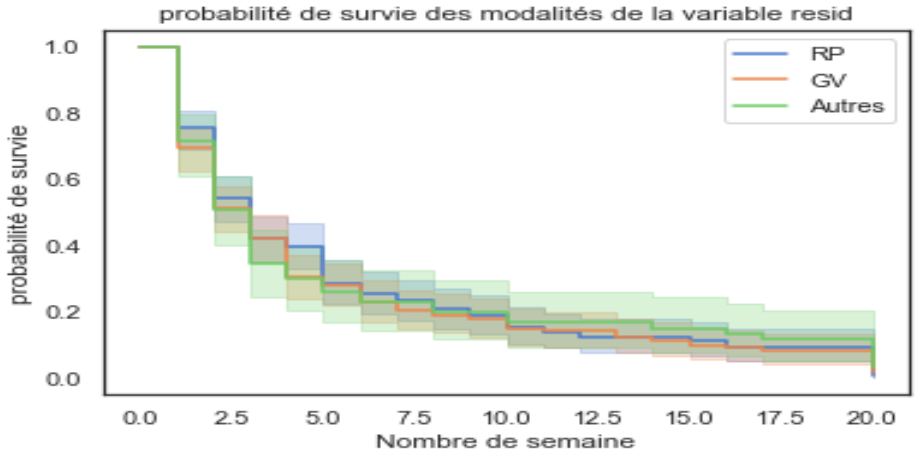
4.3.2 Variable Resid

Il est question de voir si le lieu de résidence des individus a un impact significatif sur l'achat c'est-à-dire si la région parisienne ou les grandes agglomérations ont une survie plus longue que les autres régions vice versa. Pour s'en persuader il faudra réaliser les tests multivariés d'égalité des distributions de la fonction de survie entre les strates :

Tableau 2 : Tests d'égalité de la fonction de survie des strates de la variable resid

Tests	Chi-deux	ddl	Pr> Chi-deux
Log-Rank	0,52	2	**0,77**
Wilcoxon	1,31	2	**0,52**
Tarone-Ware	0,98	2	**0,61**
Peto	1,27	2	**0,53**
Fleming-Harrington	0,30	2	**0,86**

L'analyse du tableau ci-dessus donne des p-value relativement élevée pour les statistiques de **Log-Rank, wilcoxon,** Tarone-**Ware, Peto, Fleming-Harrington**. Donc la probabilité de rejeter à tort l'hypothèse nulle d'égalité des courbes de survie est grande ; ce qui nous amène à accepter H0 au seuil d'erreur de 5%. Par conséquent la variable **Resid** n'influence pas significativement la survie des individus au seuil d'erreur de 5% et donc peut être considérée comme non déterminante dans le renouvellement de l'achat des clients. Ce résultat peut être confirmé par l'examen du graphique ci-dessous qui montre une parfaite superposition des courbes de survie des différents strates de la variable **Resid**



Source : Réalisée par les auteurs

Figure16 : Comparaison de la survie des strates de la variable Resid

4.3.3 Variable enfant

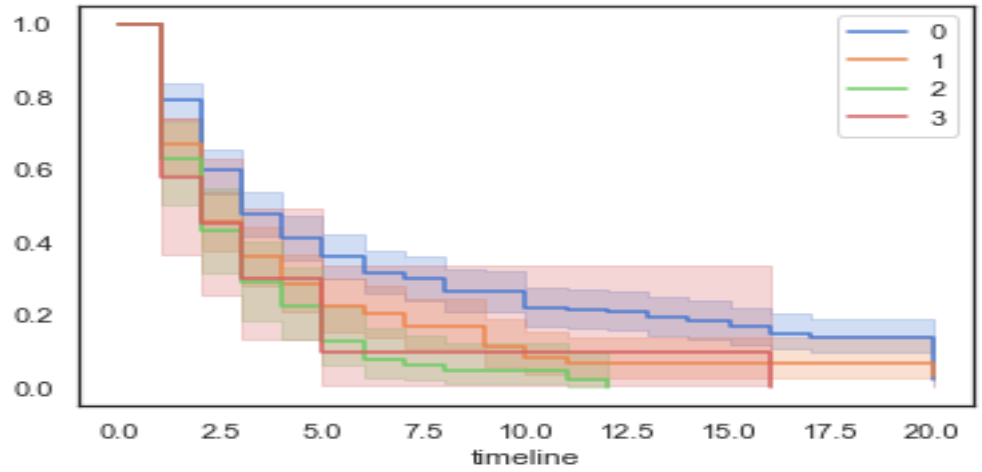
Il s'agit ici comme dans les cas précédents de voir si la variable **enfant** influence la survie c'est-à-dire si la survie des individus dépend des modalités

de la variable enfant. Pour ce faire, on va procéder à des tests multivariés d'égalité de la fonction de survie des strates :

Tableau 3: Tests d'égalité de la fonction de survie des strates de la variable enfant

Tests	Chi-deux	ddl	Pr> Chi-deux
Log-Rank	27,71	2	<0.005
Wilcoxon	20,70	2	<0.005
Tarone-Ware	23,84	2	<0.005
Peto	22,63	2	<0.005
Fleming-Harrington	19,65	2	<0.005

Au seuil de 5%, les cinq tests à savoir Log-Rank, Wilcoxon, Tarone-Ware, Peto, Fleming-Harrington rejettent l'hypothèse nulle d'égalité des fonctions de survie des quatre strates. On conclut que la variable enfant influence significativement le renouvellement de l'achat. Ce résultat peut être confirmé par l'examen du graphique ci-dessous qui montre une séparation des courbes de survie des individus.



Source : Réalisée par les auteurs

Figure17 : Comparaison de la survie des states de la variable enfant

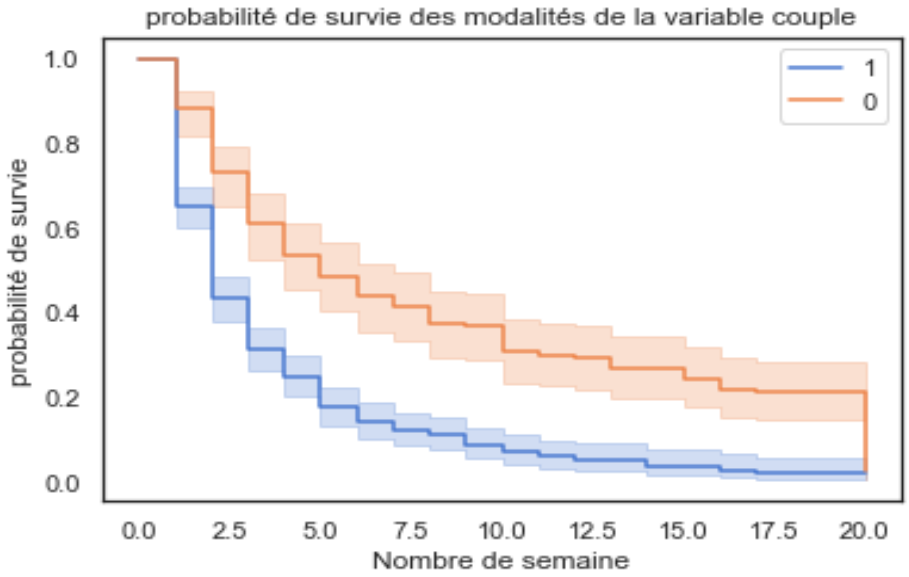
4.3.4 Variable couple

Il s'agit ici de tester l'importance de la variable couple en comparant la survie des individus célibataires et ceux qui sont en couple. Les résultats des tests de comparaison des strates sont présentés dans le tableau ci-après :

Tableau 4 : Tests d'égalité de la fonction de survie des strates de la variable couple

Tests	Chi-deux	ddl	Pr> Chi-deux
Log-Rank	58,91	1	<0.005
Wilcoxon	52,43	1	<0.005
Tarone-Ware	57,67	1	<0.005
Peto	56,06	1	<0.005
Fleming-Harrington	31,16	1	<0.005

Les tests de Log-Rank, Wilcoxon, Tarone-Ware, Peto, Fleming-Harrington rejettent à 5% l'hypothèse nulle d'égalité des courbes de survie constituée sur la base de la variable couple; par conséquent nous pouvons conclure que la variable couple impacte significativement la survie des clients . Ce résultat est confirmé par le graphique ci-après qui montre un écart réel entre la courbe de survie des individus célibataires et ceux qui sont en couple. En effet, la courbe de survie des individus célibataire est au-dessus des individus qui sont en couple ; preuve que la survie est plus longue pour les célibataires que les mariés.



Source : Réalisée par les auteurs

Figure18 : Comparaison de la survie des states de la variable couple

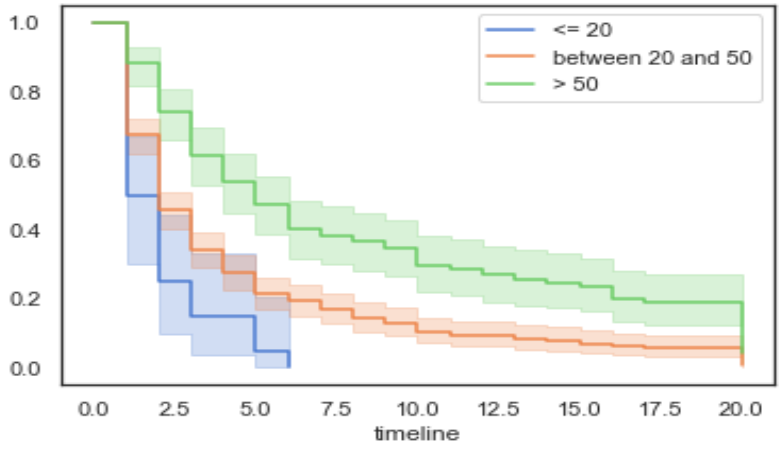
4.3.5 Variable age_discretisee

Il s'agit d'analyser l'influence de la nouvelle crée age_discretisee (Variable crée à partir de la variable age pour faciliter l'interprétation de cette variable) sur les fonctions de survie des strates. Les tests d'égalité des strates sont consignés dans le tableau suivant :

Tableau 5 : Tests d'égalité de la fonction de survie des strates de la variable age_discretisee

Tests	Chi-deux	ddl	Pr> Chi-deux
Log-Rank	49,07	2	<0.005
Wilcoxon	48,92	2	<0.005
Tarone-Ware	51.31	2	<0.005
Peto	50,27	2	<0.005
Fleming-Harrington	21,57	2	<0.005

Les résultats des tests statistiques nous amènent à rejeter l'hypothèse nulle au seuil d'erreur de 5%. Par conséquent, la variable age_discretisee serait significativement explicative du renouvellement de l'achat des clients au seuil d'erreur de 5%. L'examen graphique des courbes de survie révèle que la survie des individus dont l'âge est supérieur à 50ans est plus longue que les autres tranches d'âge.



Source : Réalisée par les auteurs

Figure19 : Comparaison de la survie des states de la variable age_discretisee

5 Troisième partie : Approche semi paramétrique

5.1 Impact des différentes variables explicatives

Il serait question dans cette partie de mesurer l'impact des différentes variables explicatives sur la fonction de risque.

La spécification de la fonction de risque proportionnel est :

$$H(t/x) = h_o(t)r(x)$$

Où $h_0(t)$ est le risque de base et **r(x) le** risque lié à chaque individu à travers les variables explicatives. On pose :

$$r(x)=exp(x'\beta)$$

Puisque le risque doit être positif pour toutes les valeurs des explicatives et des coefficients, il est postulé que la fonction de risque est sous la forme d'une exponentielle.

Tableau 6 : Analyse de la vraisemblance des coefficients estimés pour le premier modèle

Variables	Coefficients estimés	Chi-Square	Pr > ChiSq	Ratio de risque	Mesure de l'exactitude prédictive (La concordance)
age	-0,04	-10,57	<0.005	0,96	
resid	0,04	0,60	0,55	1,04	
enfant	0,12	1,88	0,06	1,12	0,73
couple	0,96	7,49	<0.005	2,61	

Avant de procéder à l'analyse et à l'interprétation des coefficients nous allons procéder à l'étude de la significativité globale du modèle et des coefficients.

Au seuil d'erreur de 10%, la variable **resid** n'est pas significatif. Nous allons donc le retirer et procéder à une ré estimation du modèle.

L'indice de concordance du modèle qui évalue la précision du classement du temps prédit est de 0,73. Ainsi le modèle fait mieux plus que le résultat attendu des prédictions aléatoires**.** En effet, tous les coefficients sont globalement différents de zéro.

5.2 Estimation du modèle à l'aide des variables significatives au seuil d'erreur de 10% dans la première estimation.

Dans cette partie, les variables qui seront prises en compte
dans le modèle sont : age, couple et enfant. Les résultats de la nouvelle estimation sont :

Tableau 7 : Analyse de la vraisemblance des coefficients estimés pour le deuxième modèle

Variables	Coefficients estimés	Chi-Square	Pr > ChiSq	Ratio de risque	Mesure de l'exactitude prédictive (La concordance)
Age	-0,04	-10,56	<0.005	0,96	
Enfant	0,11	1,83	0,07	1,12	
Couple	0,96	7,53	<0.005	2,61	0,73

On remarque qu'au seuil d'erreur de 8% tous les coefficients des variables sont significatifs. De plus le taux de meilleur classement de temps prédit est meilleur que la prédiction aléatoire.

Remarquons que le taux de meilleur classement pour les deux modèles est le même.
Cependant nous retenons le dernier modèle car il est parcimonieux

5.3 Interprétation des coefficients

5.3.1 Variable age

Le signe du coefficient de la variable **age** est négatif ; ceci implique que la variable age influence négativement le risque. Le rapport de risque (HR) est de 0,96, ce qui indique une forte relation entre l'âge des clients et une diminution du risque de renouvellement de l'achat. Par exemple, en maintenant les autres covariables constantes, une personne dont le niveau de l'âge est élevé réduit le risque d'un facteur de 0,96, soit de 4%.

5.3.2 Variable enfant

Le signe du coefficient de la variable enfant est positif ; donc la variable enfant affecte positivement le risque des individus. le rapport de risque (HR) est de 1,12, ce qui indique une forte relation entre la variable **enfan**t et le risque accru de renouvellement de l'achat. Par exemple, en maintenant les autres covariables constantes, être un foyer deux enfant accroit le risque d'un facteur de 1,12, soit de 12%.

5.3.3 Variable sexe

Le signe du coefficient de la variable sexe est positif ; ce qui signifie que, comme la variable age, la variable sexe influence positivement le risque des individus. Le rapport de risque (HR) est 1,42, ce qui indique une forte relation entre le sexe des clients et un accroissement du risque de renouvellement. Toutes choses égales par ailleurs, être une femme (sexe = 0) accroit le risque d'un facteur de 1,42, soit 42%.

5.3.4 Variable Couple

Le signe du coefficient de la variable couple est positif ; donc la variable couple affecte positivement le risque des individus. Le rapport de risque (HR) est 2,61, ce qui indique une forte relation entre la couple et le risque accru de renouvellement. Par exemple, toutes choses égales par ailleurs, être un célibataire accroit le risque d'un facteur de 2,61 soit de 161%.

5.4 Analyse des résidus et validation du modèle

5.4.1 Les résidus de Schoenfeld mis à l'échelle

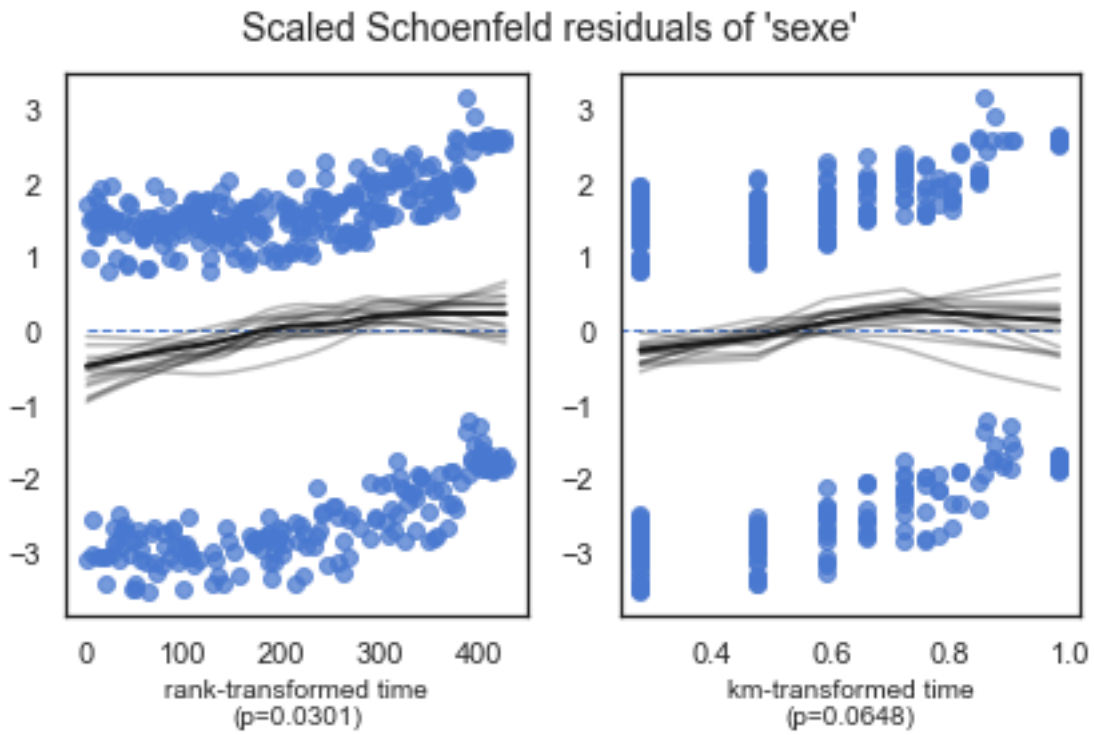
Les résultats d'un test statistique visant à tester les coefficients variants dans le temps sont présentés en premier dans les résidus de Schoenfeld. Un coefficient variant dans le temps implique l'influence d'une covariable par rapport aux changements de base au fil du temps. Cela implique une violation de l'hypothèse de risque proportionnel. Pour chaque variable, nous transformons le temps quatre fois (ce sont des transformations courantes du temps à effectuer). Si les lignes de vie rejettent la valeur nulle (autrement dit, les lignes de *vie* rejettent que le coefficient ne varie pas dans le temps), nous le signalons à l'utilisateur. Le tableau suivant présente les résultats d'un test statistique qui teste les coefficients variant dans le temps.

5.4.2 Test statistique des résidus de Schoenfeld

Tableau 8 : Analyse de résidus de Schoenfeld

Variables	Chi-Square	dll	Pr > ChiSq
age	1,12	1	0,29
couple	0,65	1	0,42
enfant	0,24	1	0,63
sexe	4,72	1	0,03

Au seuil d'erreur de 10%, nous décidons de rejeter l'hypothèse nulle d'invariance du coefficient de la variable sexe dans le temps. On conclut que la variable sexe a violé l'hypothèse de risque proportionnelle. Ce résultat est confirmé par les graphiques résiduels ci-après qui montrent que l'effet de sexe est croissant avec le temps.



Source : Réalisée par les auteurs

Figure20 : graphiques résiduels du sexe

5.5 Prise en compte de la variable sexe qui varie dans le temps

Comme mentionné dans les précédentes analyses, la variable sexe est une variable particulière qui change de valeur au cours du temps. Compte tenu de son caractère particulier nous ne l'avons introduit dans le modèle qu'au dernier moment afin de mesurer son impact sur le renouvellement de l'achat. Les résultats d'estimation après ajout de la variable sexe dans le modèle se présente comme suit :

Tableau 9 : Analyse de la vraisemblance des coefficients estimés du modèle

Variables	Coefficients estimés	Chi-Square	Pr > ChiSq	Ratio de risque	Mesure de l'exactitude prédictive (La concordance)
Age	-0,04	-10,56	<0.005	0,96	0,73
Enfant	0,11	1,83	0,07	1,12	
Sexe	0,35	3,48	<0.005	1,42	
Couple	0,96	7,53	<0.005	2,61	

Au seuil d'erreur de 10%, tous les coefficients sont significatifs et on remarque que l'introduction de la variable sexe dans le modèle n'a pas fondamentalement pas affecté la valeur des coefficients de toutes les variables significatives à savoir age, enfant, couple. Compte tenu du signe de la valeur obtenue pour le coefficient de la variable sexe, on peut dire que cette dernière influence positivement le risque. Le rapport de risque (HR) est 1,42, ce qui indique une forte relation entre le sexe des clients et un accroissement du risque de renouvellement. Toutes choses égales par ailleurs, être une homme (sexe = 1) accroit le risque d'un facteur de 1,42, soit 42%.

6 Quatrième partie : Analyse paramétrique de la survie

6.1 Modèles de temps de défaillance accéléré(AFT)

Dans l'approche paramétrique, il s'agira de spécifier la distribution des temps de survie des individus et ensuite d'introduire dans le modèle les variables explicatives en vue de mesurer leur impact sur le temps de survie des individus qui sont ici le nombre de semaine séparant le premier achat et le renouvellement. En effet, les modélisations paramétriques de la survie sont souvent faites à l'aide des Modèles de temps de défaillance accéléré(AFT)

Supposons que nous ayons deux populations, A et B, avec des fonctions de survie différentes, **SA(t) et SB(t)** et qu'elles soient liées par un taux de défaillance accéléré, λ le lambda :

$$\mathbf{SA(t)=SB(\frac{t}{\lambda})}$$

Cela peut être interprété comme ralentissant ou accélérant le déplacement le long de la fonction de survie. Ce modèle a d'autres propriétés intéressantes : le temps de survie moyen de la

population B est λ(lambda) multiplié par le temps de survie moyen de la population A. De même avec le temps de survie *médian*. Plus généralement, nous pouvons modéliser le lambda en fonction des covariables disponibles, c'est-à-dire :

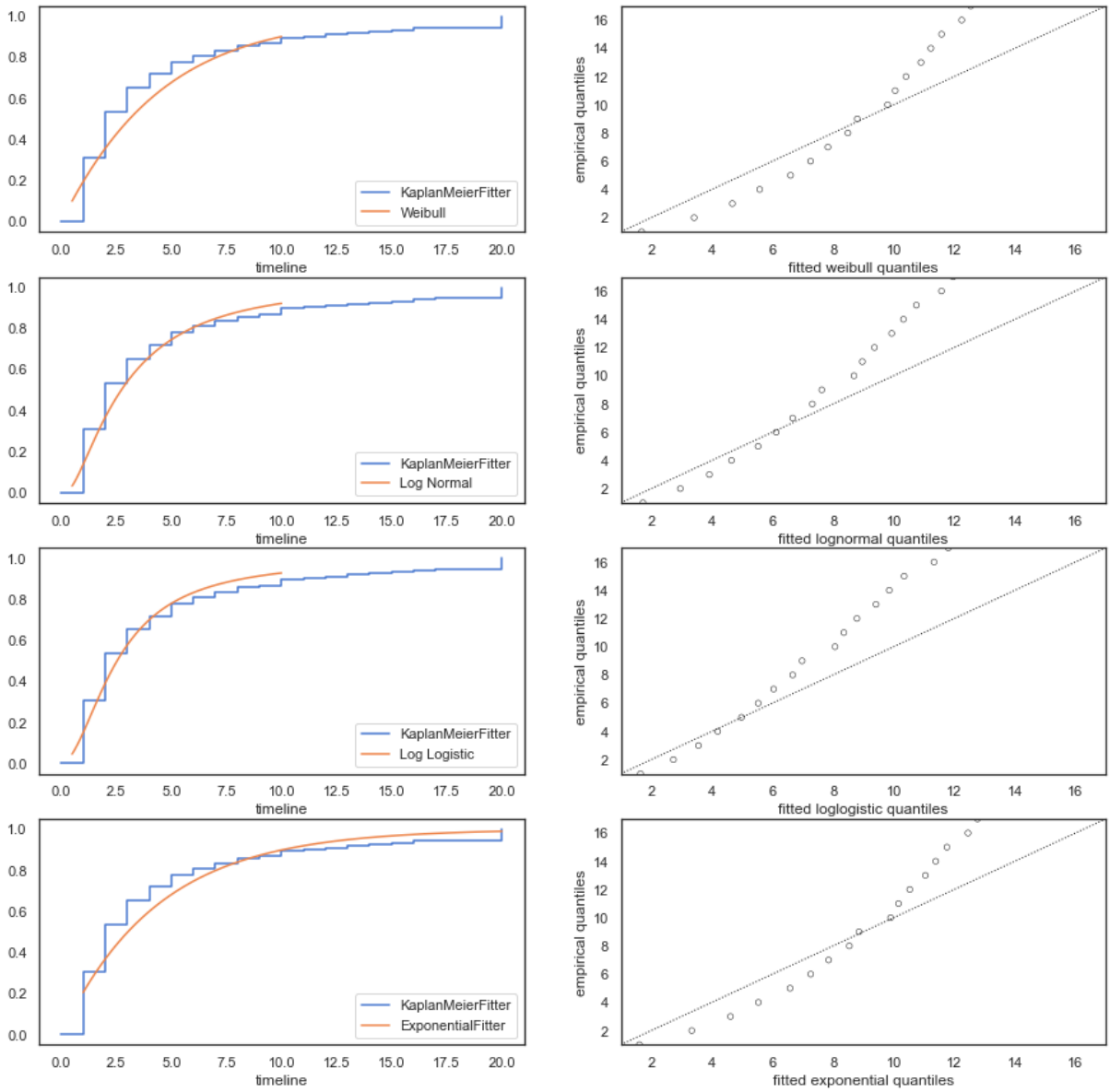
$$\mathbf{SA(t)=SB\left(\frac{t}{\lambda}\right)}$$

$$\mathbf{\lambda(x)=exp(b0+\sum_{i=1}^{n} bixi)}$$

Ce modèle peut accélérer ou ralentir les temps de défaillance en fonction des covariables des sujets. Une autre caractéristique intéressante de ceci est la facilité d'interprétation des coefficients: une augmentation unitaire de xi signifie que le temps de survie moyen / médian change d'un facteur de exp(bi). En effet, le choix du modèle dépendra beaucoup de la distribution spécifique des temps de survie. Ainsi, en raison de la présence de covariables, nous allons précéder à l'analyse des résidus après estimation afin de mieux spécifier la distribution à retenir.

6.2 Analyse des résidus après estimation

Par cette analyse on espère approximer la distribution conditionnelle des temps de survie à partir de l'analyse du résidu issu de l'estimation. L'une des méthodes les plus utilisées est l'analyse des résidus de Cox-Snell. En effet, quel que soit la distribution spécifiée dans l'approximation du modèle, les résidus de Cox-Snell construits à partir de la fonction de répartition des temps de survie estimés, suivent une distribution exponentielle standard de paramètre lambda(λ) qui est une constante. La méthode revient donc à faire le graphique des résidus de Cox-Snell Pour ce faire, nous avons procédé à l'estimation en utilisant plusieurs distributions à savoir : la Gamma, la Weibull, le Log-Normale, la Log-Logistique et l'Exponentielle. Nous nous sommes rendu compte que c'est
la distribution Log-Normale qui donne une représentation des résidus de Cox-Snell plus proche d'une droite.



Source : Réalisée par les auteurs

Figure21 : Représentation des résidus de Cox-Snell

Au vue des graphiques présentés dans l'analyse précédente, la distribution Log-Normal nous semble la mieux appropriée. Une fois la distribution des temps de survie choisie, il ne nous reste plus qu'à interpréter les coefficients des variables estimés dans le modèle.

6.3 Présentation des résultats et interprétation des coefficients

Tableau 10 : Analyse de la vraisemblance des coefficients estimés du modèle

variables	Coefficients estimés	Chi-Square	Pr > ChiSq	Ratio de risque
intercept	0,52	3.74	<0.005	1.69
age	0,03	11.62	<0.005	0.48
couple	-0,73	-8.45	<0.005	0.48
enfant	-0,08	-1.69	0.09	0.93
resid	-0,02	-0.44	0.66	0.98
sexe	-0,24	-3.28	<0.005	0,79

On constate que les variables **enfant et resid** ne sont pas significatives au seuil d'erreur de 10% comme dans le cas de l'approche semi paramétrique. Nous allons donc les extraire du modèle et procéder à une estimation à nouveau.

Tableau 11 : Analyse de la vraisemblance des coefficients estimés pour le deuxième modèle

Variables	Coefficients estimés	Chi-Square	Pr > ChiSq	Ratio de risque
Intercept	0,50	3.64	<0.005	1.65
Age	0,03	11.56	<0.005	1,03
Couple	-0,8	-10.57	<0.005	0.45
Sexe	-0,24	-3.23	<0.005	0,79

A un risque de première espèce, de 10%, les variables qui ressortent significativement influentes sur la fonction de survie, sont l'âge, le couple et le sexe. Ainsi, le signe de chaque coefficient indique l'impact positif ou négatif de la variable sur la survie. Par conséquent, le fait d'être célibataire influencerait le renouvellement de l'achat à la baisse c'est-à-dire allonge la durée de renouvellement. Plus l'individu est âgé, plus il renouvelle l'achat. L'âge accroit le risque de renouvellement de l'achat à la hausse et donc réduit la durée de survie. La variable sexe influence aussi le renouvellement à la baisse et donc allonge le risque de survie.

Il faudra faire une remarque intéressante qui va en concordance avec les résultats obtenus dans l'approche semi paramétrique : en effet les variables significatives dont les coefficients avaient un signe positif dans l'approche semi paramétrique ont maintenant un signe négatif dans l'approche paramétrique. Ce qui signifie simplement que les variables qui impactent positivement la survie affectent négativement le risque.

7 Conclusion : Recommandations de politique

D'abord, si nous allons à une comparaison des résultats obtenus dans les différentes approches, nous pouvons conclure que l'approche non paramétrique et semi paramétrique donnent les mêmes résultats parce que les variables qui sont apparues significatives dans l'explication de survie dans la première sont également déterminantes dans la mesure du risque dans la seconde.

En nous référant à l'approche non paramétrique, la seule différence qu'on remarque est que la variables enfant qui était significative dans les approche non paramétrique et semi paramétrique ne l'est pas dans l'approche paramétrique Ceci nous permet de conclure que d'une manière ou d'une autre, que les approches non paramétriques et semi paramétriques permettent de valider les résultats de l'approche paramétrique. Enfin, à la lumière de cette étude, il serait intéressant de tirer de judicieuses conclusions qui serviront à éclairer les choix de politique commerciales de la structure d'augmentation du renouvellement des clients. En effet, un certain nombre de variables se sont révélées très significatives dans l'explication du renouvellement de l'achat. Dans toutes les approches (non paramétrique, semi paramétrique et paramétrique) les variables telles que enfant, sexe, couple et l'âge influencent positivement la survie des individus c'est-à-dire contribuent à accroitre le risque de renouvellement de l'achat des individus. Ainsi, les actions suivantes pourraient être entreprises par les structures si elles veulent accroitre le renouvellement de l'achat.

- Faire le marketing clientèle en faisant des faveurs aux célibataires car ils ont une durée de survie comparativement très élevé c'est-à-dire une durée de renouvellement elevée
- Accorder une priorité aux femmes pendant la politique de ciblage client puisqu'elles ont une durée de survie aussi longue et un risque très faible
- Prendre soins des clients dont les âges sont compris entre 21ans et 50ans et ceux dont les âges sont supérieurs à 50ans puisque ces catégories d'âges ont une durée de survie très longue.

Enfin, la variable resid qui à priori pouvait paraître déterminante dans l'explication du renouvellement s'est révélée non significative. Par conséquent, des actions sur cette variable ne pourront pas, compte tenu des résultats que nous avons obtenus, influencer le risque. D'où toute l'importance de cette étude qui permet

une meilleure identification des variables pertinentes afin de mettre en œuvre des actions plus efficaces en vue d'une augmentation du renouvellement de l'achat des clients.