Modélisation du risque arrêt de travail

# Modèle de durée : censures et troncatures

Pour modéliser le risque arrêt de travail sur notre portefeuille nous avons procédé à un travail de nettoyage et de cohérence minutieux sur nos données, cela est primordial pour avoir une modélisation qui soit la plus fiable possible. Cependant, l’information que nous avons actuellement dans notre base de données de sinistres arrêts de travail reste encore biaisée à cause de la nature des arrêts (cf. *Non Parametric Estimation from Incomplete Observation* E.L. Kaplan et P.Meier [ajouter la ref à la bio]) et plus précisément de leur durée et de la période d’observation, on parle d’arrêts tronqués et/ou censurés, nous présenterons les types de censures et troncatures présentes dans nos données qui sont la troncature à gauche et la censure à droite de type I.

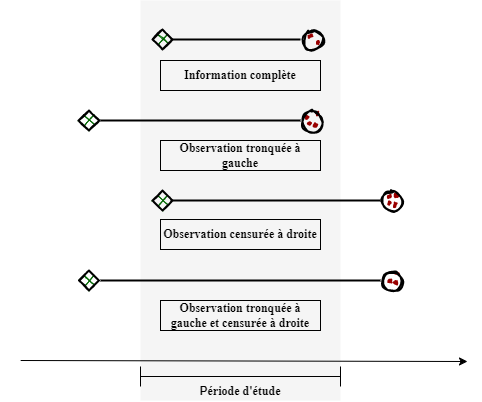


Figure : Schéma représentant les phénomènes de censures et de troncatures

## Censure à droite :

Certains arrêts de travail ne sont pas observés jusqu'au bout, on parle de censure à droite de l’information sur la durée de l'arrêt de travail, bien que nous ayons une date de fin d'observation dans nos données, la période réelle d’arrêts est plus longue que celle que nous captons dans nos données.

Considérons un échantillon de durée de survie qui dans notre cas représenterait la durée réelle des arrêts de travail dans notre portefeuille, ce que l’on observe dans notre base de données sont en fait les observations avec :

La variable D indique si l’observation est censurée ou non, et C représente la date de censure, dans la pratique, nous observons trois types de censures :

Nous observons deux causes principales de censures : la censure causée par l'observation sur un intervalle de temps : tout arrêt de travail continuant au-delà de fin 2022 est censuré dû à la date d’extraction des données, dans ce cas =31/12/2022. La deuxième cause est la censure pour départ à la retraite, =date de départ à la retraite de l’assuré.

La troisième et dernière cause traitée en tant que censure dans cette étude est la sortie au bout des 3 ans d'incapacité.

## Troncature à gauche :

La troncature gauche se produit lorsque la variable d'intérêt n'est pas observable en dessous d'un certain seuil c. Contrairement à la censure, la troncature signifie que l'information concernant les observations en dehors de cette plage est entièrement perdue, car on ne dispose d'aucune donnée concernant ces observations. En revanche, dans le cas de la censure, on sait qu'il existe une information, mais on ne dispose que d'une limite supérieure ou inférieure pour cette information.

La troncature peut être observée dans différentes situations, par exemple lors d'une migration informatique où seuls les sinistres en cours ont été transférés dans la nouvelle base de données, entraînant la perte d'informations sur les sinistres de durée plus courte pour les mêmes événements. Un autre exemple est celui du contrat d'arrêt de travail avec une franchise : les arrêts de durée inférieure à la franchise ne sont pas observés, ce qui signifie qu'aucune information n'est disponible sur eux.

Nous avons sélectionné pour notre étude les observations qui ont eu lieu durant la période d’étude, c’est-à-dire celles qui ne sont ni tronquées ni censurées, ainsi que celles qui présentent une information partielle, les observations censurées à droite.

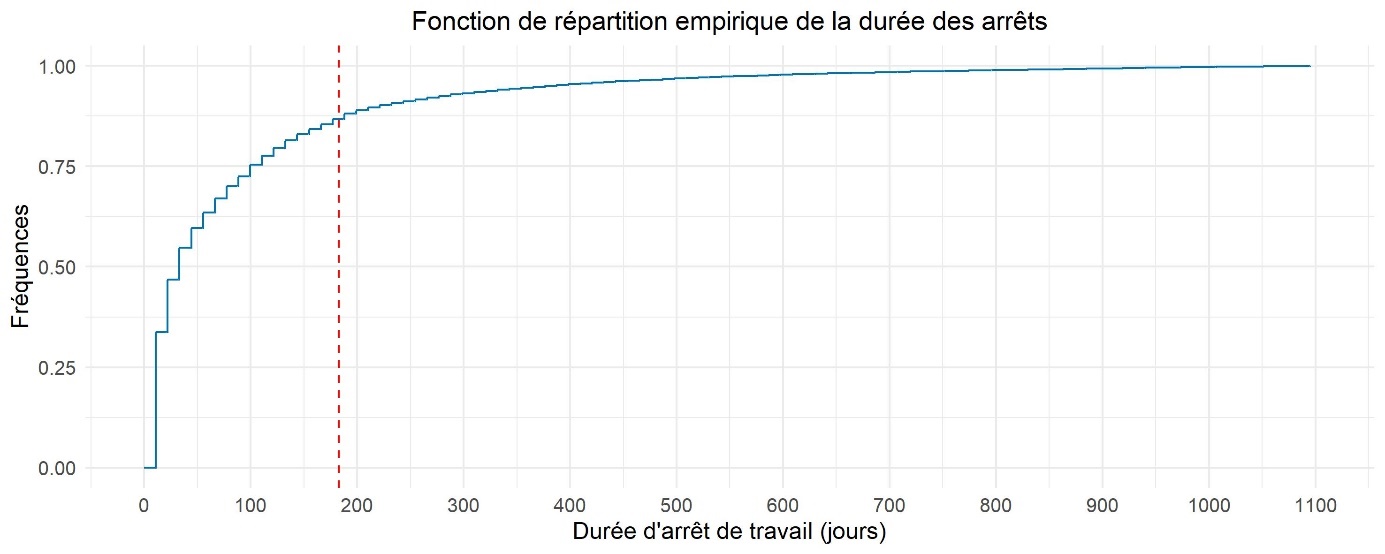
## Modélisation de la durée des arrêts de travail

### Etude de la distribution de la durée des arrêts de travail

L’idée de cette partie est de déterminer la loi continue qui s’ajustera au mieux à nos données, afin de calibrer au mieux nos modèles linéaires généralisés. Pour ce faire nous allons comparer la distribution de nos données avec la fonction de répartition de plusieurs lois usuelles, en estimant les paramètres des lois.

#### Comparaison des fonctions de répartition

Soit la fonction de répartition empirique de la durée d’incapacité dont le tracé se trouve ci-dessous, nous nous intéresserons aussi à la proportion des arrêts durant moins d’une demi-année afin de savoir s’il est intéressant ou non d’effectuer deux modélisations différentes pour les arrêts courts (< 6 mois) et les arrêts longs



Nous observons que plus de 80% des arrêts durent moins de 6 mois (représenté par l’axe vertical en point tillés sur le graphe), ce seuil est très important pour la sécurité sociale, en effet pour que l’indemnisation d’un arrêt de travail puisse dépasser 6 mois, il faut impérativement le service médical de l’Assurance Maladie donne son accord et que l’état de santé de la personne arrêtée le justifie. Nous verrons de notre côté dans la suite s’il est pertinent ou non mathématiquement d’affiner nos modèles en effectuant cette séparation des arrêts.

#### Estimation des paramètres de lois :

Les lois usuelles que nous allons utiliser pour la comparaison avec notre distribution sont les lois suivantes :

* Loi Gamma
* Loi Log-Normal
* Loi Inverse-Gaussienne
* Loi Exponentielle

Nous estimerons les paramètres de ces lois à partir des moyennes et des variances empiriques de nos échantillons que nous calculerons grâce aux estimateurs sans biais suivants :

et

Avec un vecteur de variables iid de même loi que la variable à expliquer

Les paramètres des lois sont estimés de la façon suivante :

* Loi Gamma  :

et

* Loi Log-Normal  :

et

Avec et respectivement les estimateurs de la moyenne et la variance de la variable durée à expliquer composé par la fonction ln.

* Loi Inverse-Gaussienne  :

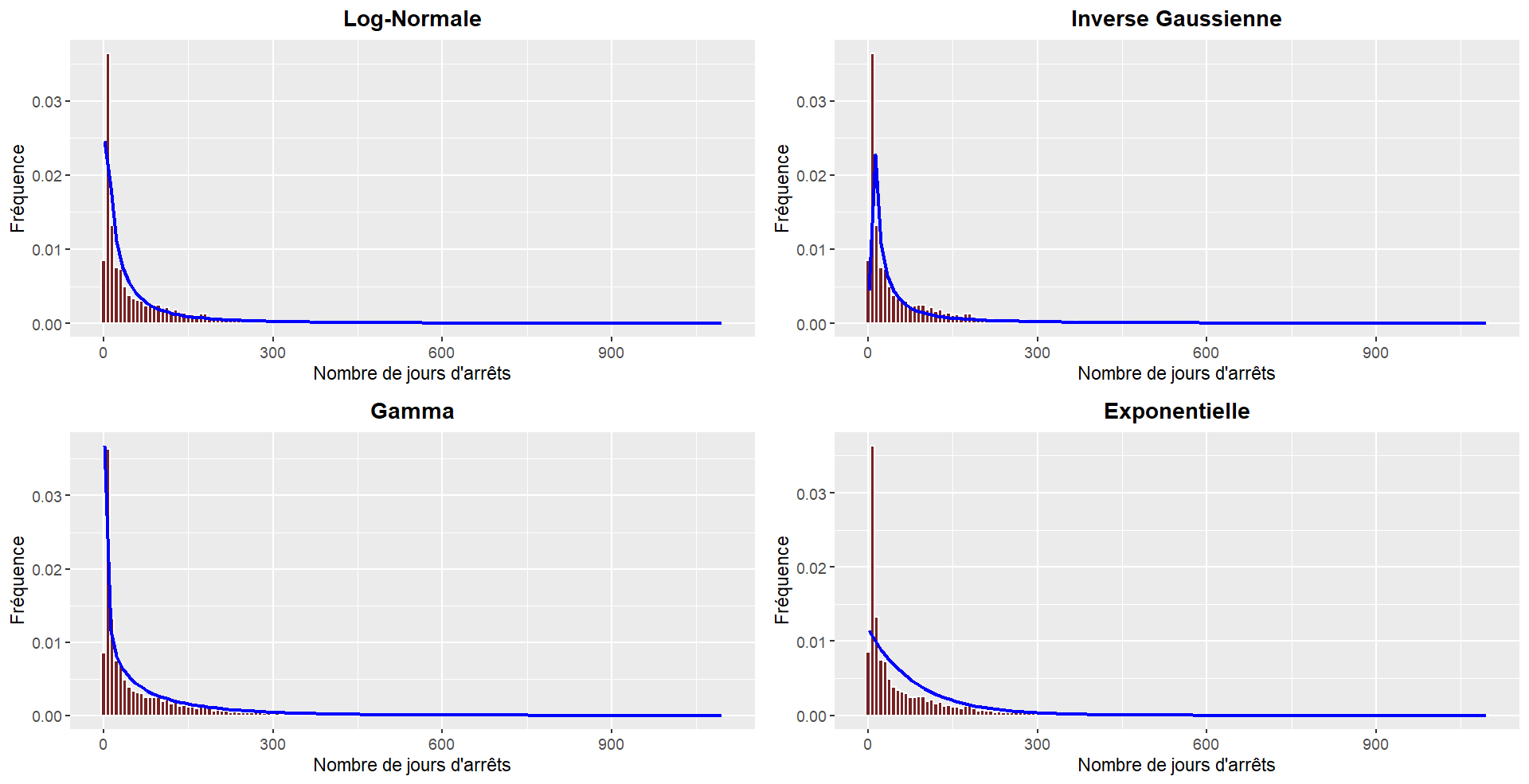
et

* Loi Exponentielle  :

#### Analyse graphique et comparaison des lois

##### Comparaison des densités

Nous nous baserons sur une analyse graphique dans un premier temps pour avoir une idée de la loi qui s’ajuste le mieux à nos données, pour ce faire nous allons nous baser sur l’histogramme des durées d’arrêts d’incapacité.



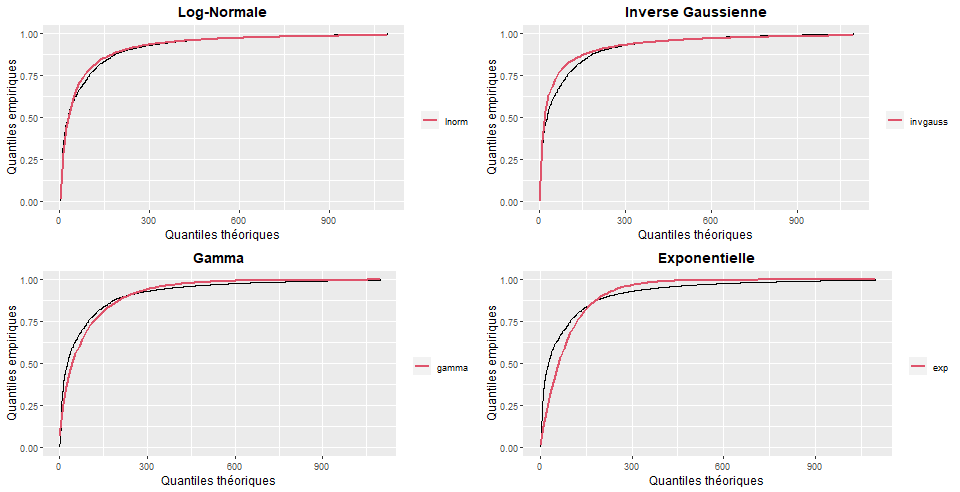
Nous avons tracé en bleu les densités des lois testées, on remarque que la loi gamma et la loi inverse gaussienne s’ajustent correctement à nos données. Cela reste une première impression graphique, nous confirmerons ou infirmerons cela dans la suite grâce à nos modèles linéaires généralisés.

##### Comparaison des QQ-plot

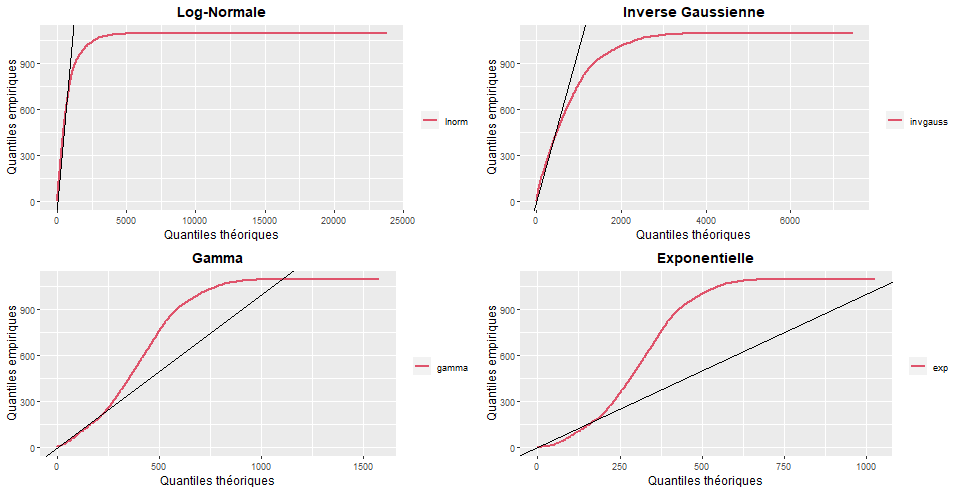
Le QQ-plot (Quantile to Quantile Plot) est une méthode qui permet d’étudier graphiquement des données avec les hypothèses d’une loi en comparant les quantiles empiriques et théoriques.

Pour expliquer cette méthode, reprenons la notation précédentes pour décrire notre échantillon de données triées par ordre croissant de la variable à expliquer (=durée des arrêts), la fonction de répartition de la loi usuelle testée. Nous nous basons sur l’hypothèse d’adéquation suivante :

Dans un premier temps, nous allons comparer les fonctions de répartitions calibrées avec les paramètres estimés avec la fonction de répartition empirique de nos données :



Puis les QQ-plot des lois usuelles considérées :



Nous remarquons que les lois Log-Normale et Inverse Gaussienne s’ajustent particulièrement bien à la variable à expliquer. La loi Gamma reste cela dit intéressante nous l’utiliserons elle aussi dans la suite de l’étude alors que nous excluons la loi exponentielle de notre étude aux vues de cette analyse graphique.

### Modélisation des arrêts grâce aux modèles linéaires généralisés

#### Aspect théorique des modèles linéaires généralisés

Les modèles linéaires généralisés ont été introduits par Nelder et Wedderburn en 1972, ils constituent la base de référence pour modéliser l’effet des variables de segmentation sur un tarif en assurance. Ces modèles sont adaptés à de nombreuses problématiques courante dans le domaine de la statistique et de l’actuariat.

##### Intérêt des modèles linéaires généralisés

Les modèles linéaires généralisés sont des modèles régulièrement utilisés en assurance que ce soit en assurance santé (remboursements soins, frais d’hospitalisation), en assurance auto (dommage matériel, vol, …), en assurance MRH (incendie, vol, dégâts des eaux, …).

Les GLM permettent de :

* Modéliser des réponses diverses
* Intégrer toute type d’information exogène susceptible d’influer sur la variable dépendante (réponse Y)
* Quantifier l’impact des facteurs de risque X (sens/intensité)
* Résidus hétéroscédastiques (la loi varie par profil)

Cependant, leur mise en place nécessite d’introduire deux hypothèses fondamentales :

* Les données que l’on cherche à expliquer sont indépendants entre elles
* Les variables explicatives X sont indépendantes deux à deux

Ces hypothèses sont très importantes notamment car l’un des intérêts des modèles linéaires généralisés en assurance est la tarification et donc le calcul d’une prime pure, calculée généralement de la façon suivante avec une approche fréquence-sinistre :

#### Principe des GLM

##### Composant d’un GLM :

Les différents composants d’un modèle linéaire généralisé sont :

* **La loi de la réponse aléatoire Yi :** par hypothèse la distribution de cette loi appartient à la famille exponentielle

Rappel : La densité de probabilité d’une loi appartenant à la famille exponentielle s’écrit de la façon suivante :

Avec :

* Avec θ le paramètre naturel de la dispersion,
* φ le paramètre de dispersion, b une fonction définie sur deux fois dérivable et de dérivée première injective,
* c une fonction définie sur

De plus on a en particulier avec les lois de cette famille :

ainsi que

* **Le prédicteur** noté ηi et défini tel que est linéaire et déterministe : il est calculé via les facteurs de risques explicatifs
* **La fonction de lien g :** elle est par définition monotone, dérivable et inversible. Elle est définie et utilisée en pratique de la façon suivante :

En pratique, il faut adapter la fonction de lien au domaine de définition de Y, voici un tableau qui présente les différentes fonction lien utilisées en pratique :

|  |  |  |
| --- | --- | --- |
| Loi | Lien canonique | Moyenne |
| Normale | Identité : |  |
| Binomiale | Logit : |  |
| Poisson | Log : |  |
| Gamma | Inverse : |  |
| Inverse-Gaussienne | Inverse² : |  |

La fonction de lien canonique est la fonction lien qui associe la moyenne µ au paramètre canonique θ. Elle est telle que :

A présent, présentons quelques modèles fréquemment utilisés dans le contexte assurantiel :

##### Modèle gaussien :

Dans le cas d’un échantillon gaussien, les densités d’une famille de lois s’écrivent :

Et appartiennent à la famille exponentielle en posant :

On remarque avec la première égalité que la famille gaussienne se met sous forme canonique

##### Modèle de Poisson :

Le modèle de Poisson est un type de modèle de régression utilisé pour modéliser des variables de réponse qui suivent une distribution de Poisson. La distribution de Poisson est souvent utilisée pour modéliser des variables qui comptent le nombre d'événements qui se produisent au cours d'une période donnée, comme le nombre de sinistres d'assurance sur une période donnée.

Le modèle GLM de Poisson utilise une fonction de lien logarithmique pour lier la variable de réponse à un ensemble de variables explicatives (également appelées variables prédictives ou variables indépendantes). Cela signifie que la relation entre la variable de réponse et les variables explicatives est modélisée comme une régression linéaire dans le logarithme de la variable de réponse. Cela permet de modéliser des variables de réponse qui peuvent prendre des valeurs entières uniquement et qui ont des variances qui augmentent avec la moyenne.

On considère n variables indépendantes Yi de loi de Poisson de paramètre µi = E(Yi). Les Yi sont par exemple les effectifs d’une table de contingence. Ces variables admettent pour densités :

Avec :

Et donc la fonction de lien canonique de ce modèle est la fonction logarithme népérien.

##### Modèle Log-Normal :

Le modèle lognormal est un type de modèle de régression utilisé pour modéliser des variables de réponse qui suivent une distribution lognormale. La distribution lognormale est souvent utilisée pour modéliser des variables qui ont des valeurs positives et sont asymétriques, comme les revenus ou les tailles.

Le modèle GLM lognormal utilise une fonction de lien logarithmique pour lier la variable de réponse à un ensemble de variables explicatives (également appelées variables prédictives ou variables indépendantes). Cela signifie que la relation entre la variable de réponse et les variables explicatives est modélisée comme une régression linéaire dans le logarithme naturel de la variable de réponse. Cela permet de modéliser des variables de réponse qui ont des valeurs positives et des variances qui augmentent avec la moyenne.

On considère n variables indépendantes Yi de loi lognormale de paramètre µi et σ^2. Les Yi sont par exemple les revenus d'un échantillon. Ces variables admettent pour densités :

Et donc la fonction de lien canonique de ce modèle est la fonction logarithme népérien.

##### Modèle Logistique :

Le modèle logistique est un modèle de régression utilisé pour modéliser des variables de réponse binaires, c'est-à-dire des variables qui prennent seulement deux valeurs possibles, comme oui ou non, vrai ou faux, succès ou échec. Il est souvent utilisé en sciences sociales, en sciences de la santé et en marketing pour prédire la probabilité d'un événement.

Ce modèle utilise une fonction de lien logistique pour lier la variable de réponse binaire à un ensemble de variables explicatives (également appelées variables prédictives ou variables indépendantes). Cette fonction de lien logistique permet de modéliser la probabilité de la variable de réponse en fonction des variables explicatives en transformant la réponse binaire en une probabilité qui varie de 0 à 1.

La fonction de lien canonique pour le modèle logistique est la fonction logit, présentée ci-dessus. La fonction logit transforme la probabilité de la variable de réponse en une variable continue qui varie de moins l'infini à plus l'infini. Cette variable continue est ensuite modélisée comme une régression linéaire des variables explicatives.

La fonction de densité de probabilité pour le modèle logistique est donnée par la formule suivante :

où y est la variable de réponse binaire prenant les valeurs 0 ou 1, x est le vecteur de variables explicatives et p est la probabilité que y soit égal à 1 pour les valeurs de x données.

Le modèle logistique est souvent utilisé pour prédire la probabilité d'un événement binaire, comme la probabilité qu'un patient développe une maladie, la probabilité qu'un client achète un produit, ou la probabilité qu'un électeur vote pour un candidat donné.

#### Validation d’un modèle GLM :

Pour valider le modèle linéaire généralisé étudié, il existe les différentes méthodes suivantes :

##### Validation croisée (k-fold)

En apprentissage automatique, la validation croisée k-fold est une technique essentielle pour évaluer les performances d'un modèle et estimer son erreur. Cette méthode implique de diviser aléatoirement un ensemble de données en k sous-ensembles de taille égale. Chaque sous-ensemble est utilisé une fois comme ensemble de validation tandis que les autres sous-ensembles sont utilisés pour entraîner le modèle.

Cette procédure est répétée k fois pour que chaque sous-ensemble soit utilisé une fois comme ensemble de validation. Ainsi, on obtient k estimateurs différents du modèle, chacun entraîné sur des données différentes. En moyennant ces k estimateurs, on obtient l'estimateur final, qui est utilisé pour évaluer les performances du modèle en calculant l'erreur moyenne sur l'ensemble de validation.

La validation croisée k-fold est particulièrement avantageuse car elle permet d'estimer l'erreur du modèle de manière fiable et robuste, même avec un petit ensemble de données. Cela est particulièrement important dans les cas où l'on dispose de peu de données et où il est nécessaire de maximiser la précision des estimations. Ainsi, la validation croisée k-fold est une technique incontournable pour l'évaluation des modèles d'apprentissage automatique.

##### Validation de la significativité globale du modèle

Pour valider la significativité globale du modèle, il est judicieux de travailler avec des critères de qualité d’ajustement du modèle comme la déviance ou la statistique de Pearson que nous définissons ci-dessous :

* La déviance :

Pour définir la déviance du modèle, nous introduisons la notion de modèle saturé qui est le modèle qui possède autant de paramètres que d’observations. Il est caractérisé par l’égalité entre les facteurs estimés et les observations, c’est-à-dire : . Puis nous comparerons ce modèle avec le modèle d’étude pour calculer la déviance.

Il faut aussi introduire la log-vraisemblance pour calculer la déviance :

Soit Y = le vecteur à expliquer (appartenant à la famille exponentielle) dont la densité s’écrit sous la forme :

La log-vraisemblance de ce modèle s’écrit :

Nous utiliserons comme mesure de la qualité d’ajustement du modèle la statistique du rapport de vraisemblance suivante :

ou

Car le modèle décrit bien les données lorsque L ≃

Enfin, on définit la statistique de déviance réduite ou normalisée de la façon suivante :

La déviance non réduite est :

De plus, , c’est-à-dire que le modèle est de mauvaise qualité si

* La statistique de Pearson :

La statistique du test de significativité globale de Pearson est définie de la façon suivante :

Comme pour la Déviance, nous avons

##### Validation de la significativité individuelle des coefficients de la régression

Pour valider ou non le modèle en étudiant la significativité individuelle des coefficients de la régressions, nous pouvons utiliser des tests d’hypothèses sur les paramètres, pour cela nous introduirons l’hypothèse H0 de la façon suivante :

Avec : **C** une matrice connue possédant q lignes et r un ensemble de valeurs testées.

Ensuite nous pourrons considérer trois approches pour valider ou non cette hypothèse :

* Test du rapport de vraisemblance :

Nous avons défini le rapport de vraisemblance précédemment : mais dans ce cas nous n’utiliserons pas un modèle saturé mais un modèle sans contrainte de vraisemblance et un modèle sous contrainte de vraisemblance.

Ainsi nous le rapport de vraisemblance s’exprimera :

La statistique de ce test est :

* Test de Wald :

Pour définir la statistique du test de Wald, il faut dans un premier temps définir l’estimateur du maximum de vraisemblance de , .

Avec W la matrice diagonale de pondération définie telle que :

Et donc avec les notations précédentes, sous H0,

Enfin, la statistique du test de Wald est définie comme :

En pratique, nous pouvons tester un seul paramètre :

On a alors, et .

La statistique du test de Wald est dans ce cas égale à :

* Test du Score :

Ce test est basé sur la dérivée de la fonction log-vraisemblance définie précédemment, cette dérivée est appelée le Score, on la défini en pratique dans ce cas par :

Avec G une matrice diagonale composée des éléments g’( et W une matrice d’éléments . De plus, on peut montrer que et

Et enfin la statistique du test du score est donnée par :

##### Etude des résidus

Pour évaluer l’erreur du modèle, nous pouvons calculer les résidus du modèle de différentes manières, les méthodes que l’on utilisera dans cette étude sont les principales utilisées pour calculer des résidus, les résidus de Pearson et les résidus de déviances définis tels que :

* Les résidus de Pearson :

Le résidu de Pearson est défini comme :

* Les résidus de Déviance :

On peut considérer que chaque observation yi contribue à une quantité di à la déviance (), le résidu de déviance est défini comme :

La somme des carrés des résidus dans les deux cas est asymptotiquement une statistique du Khi-2 à n-p-1 degrés de liberté.

#### Application à nos données

### Construction de loi de maintien en arrêt de travail

Cette sous-partie a pour objectif de modéliser la durée de maintien en arrêt de travail à l’aide de tables de maintien que l’on construira grâce aux données en notre possession.

#### Rappels théoriques sur les modèles de durée

**La durée de maintien en incapacité** estla variable aléatoire représente le temps passé en état d'incapacité d'un individu à partir de l'âge x. Cette durée est définie sur l'espace probabiliste .

La fonction de répartition de la durée d’arrêt que nous avons utilisé précédemment, appelée fonction de répartition de , c’est une mesure de probabilité définie sur la tribu borélienne , elle donne la probabilité de sortir de l'état d'incapacité avant un certain moment t.

La fonction de répartition étant dérivable presque partout pour la mesure de Lebesgue, notons sa dérivée, appelée densité telle que :

**La fonction de survie de**  est définie par :

Elle correspond à la probabilité que la durée de l’arrêt de travail d’un individu d’âge x soit supérieure à t.

**La fonction de survie conditionnelle de**  est définie par :

Elle correspond à la probabilité que la durée de l’arrêt de travail actif depuis une durée u, d’un individu d’âge x, s’étale encore sur une période t.

**La fonction de hasard** est définie par :

Elle correspond à la probabilité de sortie instantanée de l’état d’incapacité. Cette fonction permet d'analyser si et quand un individu sort de l'état d'incapacité depuis le début de l'étude, et est donc essentielle pour l'analyse des durées de survie. En étudiant les variations de cette fonction, on peut identifier les moments où le risque de sortie de l'état d'incapacité est élevé pour l'individu, ainsi que l'évolution de ce risque au fil du temps. La fonction de hasard s'exprime formellement comme la limite d'un ratio, et est parfois appelée le taux de hasard instantané.

De plus,

C’est grâce à cette propriété que nous allons pouvoir estimer la loi de , en estimant la fonction de hasard de la loi.

**Le taux de sortie** est défini par :

Ce rapport correspond à la probabilité qu’un individu en arrêt de travail à l’instant t sorte de l’état d’incapacité entre l’instant et l’instant . On en déduit la probabilité de maintien dans l’état tel que :

Ici nous ne considérons pas les causes de sorties de l’état d’incapacité qui sont le rétablissement, le passage en invalidité ou le décès car nous sommes dans une démarche de calcul des prestations liées au maintien en incapacité.

Nous allons à présent modéliser la loi via une estimation de la fonction de survie grâce à la méthode de Kaplan-Meier qui nous permettra d’obtenir les taux de sorties bruts que nous lisserons grâce à un lissage de Whittaker-Henderson que nous présenterons au préalable.

#### Estimation de la fonction de survie : estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est un estimateur de la fonction de survie non paramétrique qui peut être utilisé en présence de censures et troncatures. L'estimateur de Kaplan-Meier est basé sur l'estimation des taux instantanés de survie, ce qui justifie son utilisation historique dans la tarification et le provisionnement liés à l'arrêt de travail.

Dans ce mémoire, nous présentons une version simplifiée de l'estimateur en raison de la présence de doublons dus aux données discrètes. Nous effectuerons nos calculs avec un pas quotidien, l’estimation se fait âge par âge de la façon suivante :

avec l'échantillon sous risque à l'âge x juste avant le jour t et le nombre de sorties d'incapacité pour l'âge x et l'instant (par reprise d'activité, passage en invalidité, décès ou fin de garanties).

L’estimateur de Kaplan-Meier présente les propriétés suivantes :

* Unique estimateur cohérent de la fonction de survie,
* Estimateur du maximum de vraisemblance généralisée,
* C’est un estimateur robuste à condition que la survie et la censure n'aient aucune discontinuité commune,
* Il vérifie un théorème de normalité asymptotique à condition que la survie et la censure soient telles que leur fonction de répartition soient indépendantes et n'aient aucune discontinuité commune.

On obtient ainsi une estimation de la fonction de survie pour tous les âges :

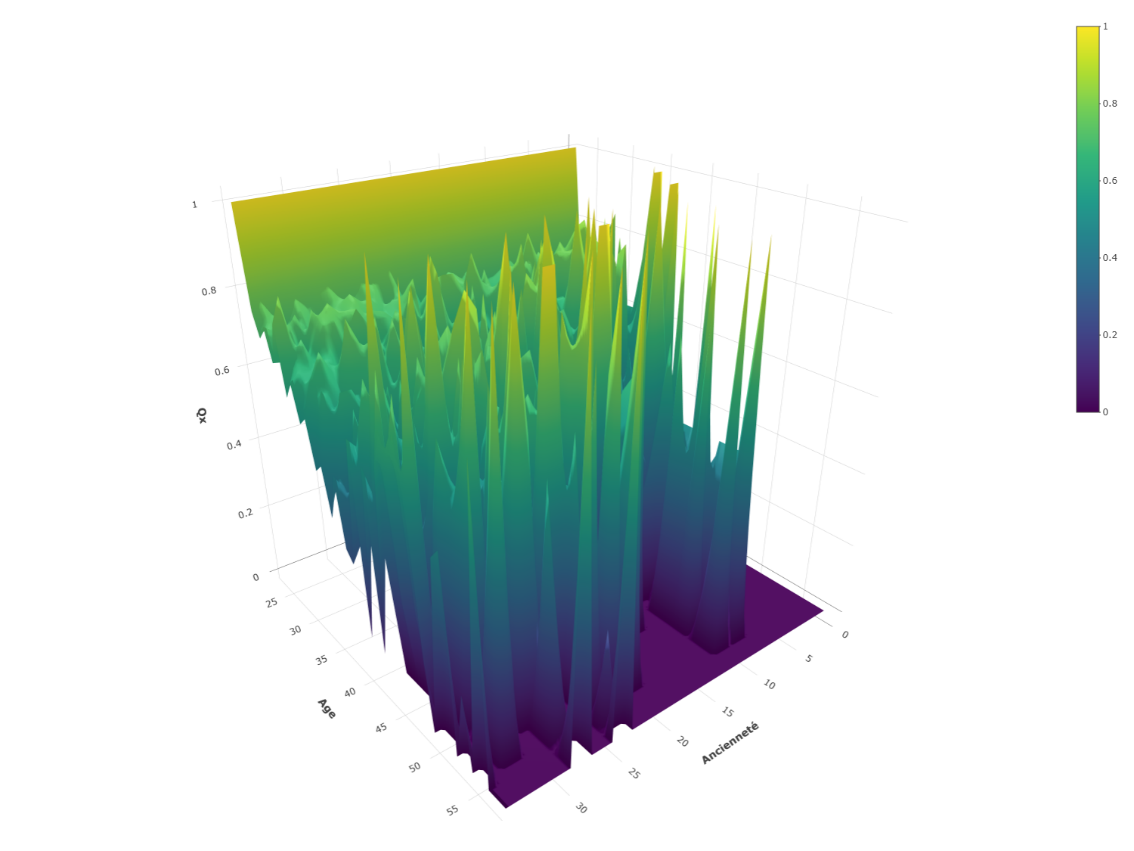


Figure : Taux bruts Qxt pour le maintien en incapacité

#### Lissage des taux bruts – Méthode de Whittaker-Henderson

La méthode de Whittaker Henderson est une méthode de lissage non paramétrique qui permet de lisser des données en combinant un critère de fidélité et un critère de régularité. La fidélité fait référence à la capacité du modèle de s'ajuster aux données observées, tandis que la régularité vise à éviter un lissage excessif ou une sur-adaptation aux données.

Pour l'appliquer, on cherche à minimiser la somme de ces deux critères pour obtenir le lissage optimal des données. Cette méthode est similaire à un lissage bayésien, où les données sont considérées comme une réalisation d'une distribution de probabilité sous-jacente.

La méthode de Whittaker-Henderson est couramment utilisée pour le lissage de données unidimensionnelles telles que les tables de mortalité. Cependant, pour des données bidimensionnelles telles que la problématique du maintien en incapacité, l'application successive de deux séries de lissages unidimensionnels ne suffit pas car elle ne capture pas les dépendances entre les deux composantes des taux de sortie.

Ainsi, on cherche à appliquer un lissage à deux dimensions pour obtenir un lissage plus précis et plus adéquat pour les données bidimensionnelles. Cependant, l'adaptation de la méthode de Whittaker-Henderson dans ce cas pose des problèmes pratiques plus importants que théoriques, nécessitant des ajustements et des améliorations pour une utilisation efficace.

##### Application au maintien en incapacité

On pose :

* : l’âge de l’assuré à la survenance
* : les taux bruts estimés selon Kaplan-Meier
* : les taux ajustés estimés avec le lissage de Whittaker-Henderson
* : les poids proportionnels aux effectifs sous risque

Notons  l’opérateur différence d’ordre n suivant :

La méthode de Whittaker-Henderson utilise deux critères :

* **Le critère de fidélité** utilisé est un critère des moindres carrés ordinaires pondérés qui quantifie la qualité de l’ajustement, noté F et défini par :
* **Les critères de régularité** utilisés sont sous la forme de sommes d’opérateurs de différence, calculés pour chaque âge et chaque ancienneté, ils sont définis par :
  + Critère de régularité verticale :
  + Critère de régularité horizontale :

L’enjeu de cette optimisation est de minimiser la fonction , pour ce faire nous utiliserons la résolution matricielle suivante :

* Soit U un vecteur colonne de dimension (, réorganisant les taux bruts   tels que :
* Soit V le vecteur colonne des taux ajustés
* Soit W la matrice de poids de dimensions
* Soit la matrice définie par : , est la matrice à m-zq lignes et m colonnes contenant les coefficients binomiaux d’ordre z affectés de leur signe
* De même la matrice définie par : de dimension (

La fonction M peut alors s’écrire :

Soit X tel que :

Alors :

Avec un terme indépendant de X

La fonction M est alors minimale lorsque

Les valeurs lissées via cette méthode s’obtiennent alors en posant :

La méthode de Whittaker-Henderson appliquée à nos données conduit à la représentation suivante des taux de sortie :

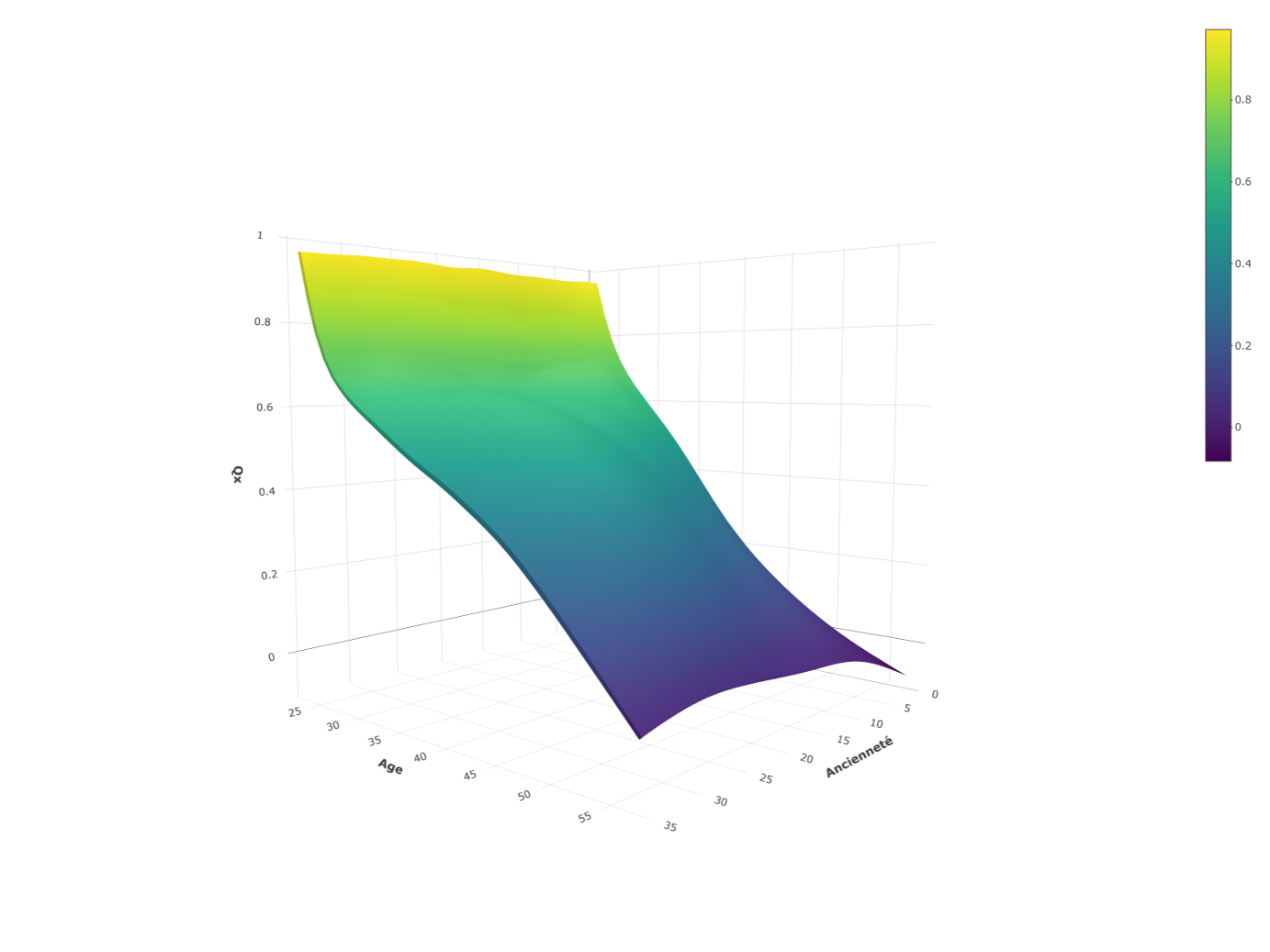


Figure : Taux lissés grâce à la méthode de Whittaker-Henderson

##### Validation du lissage