UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES ADDITIFS GÉNÉRALISÉS DANS LA MODÉLISATION DE L'IMPACT DU KILOMÉTRAGE ET DE L'EXPOSITION AU RISQUE EN ASSURANCE AUTOMOBILE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

STEVEN CÔTÉ

AOÛT 2016

REMERCIEMENTS

Je tiens d'abord à remercier Jean-Philippe Boucher, mon directeur de recherche. Merci de m'avoir confié ce projet ambitieux qui m'aura permis d'atteindre un niveau insoupçonné de compréhension de concepts mathématiques. Merci pour ton implication dans l'obtention du stage de recherche en entreprise que j'ai effectué au cours de ma maîtrise. Merci pour tous les commentaires et suggestions qui m'auront permis d'améliorer la qualité du contenu de ce mémoire. Merci pour ton support financier. Et surtout, merci de toujours avoir cru en moi.

Je tiens aussi à remercier tous les membres de ma famille, particulièrement mes parents et mon frère Bryan. Ils ont été d'un support immense à certains moments et n'ont jamais cessé de m'encourager au cours des années qu'ont duré mes études universitaires. Je vous suis extrêmement reconnaissant, merci pour tout.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Bien que proposés pour la première fois en 1986, les modèles additifs généralisés, communément connus sous l'acronyme GAM, sont encore très peu utilisés en pratique en actuariat. Ceci s'explique en grande partie par la simplicité et l'efficacité des modèles linéaires généralisés (GLM). Une problématique sous-jacente à l'utilisation d'un modèle traditionnel GLM pour la fréquence de sinistres est le traitement réservé à l'exposition au risque. Puisque les primes d'assurance sont généralement acquises de façon proportionnelle à la durée contractuelle, deux assurés ayant des profils identiques ne seraient pas considérés comme des risques équivalents dans le cas où leurs contrats respectifs ont des durées différentes. Pour répondre à cette problématique, ce mémoire présente comment les splines cubiques employées dans la construction des modèles additifs généralisés permettent une très grande flexibilité pour ajuster des données. Ces modèles, appliqués à des données d'assurance, permettent non seulement de traiter plus efficacement la durée d'exposition, mais aussi d'étudier l'impact conjoint de celle-ci avec le kilométrage exact parcouru sur le risque de sinistres automobiles. L'analyse montre que les GAM sont en mesure de capter des tendances dans les données que les modèles classiques actuellement utilisés en assurance ne réussissent pas à déceler. Finalement, une structure tarifaire simple *Pay-As-You-Drive* (PAYD) est proposée.

Mots-clés : modèle additif généralisé, modèle linéaire généralisé, GAM, GLM, spline cubique, modèle de comptage, modèle de fréquence, tarification, tarification automobile, actuariat, assurance, réclamation, inférence statistique.

CHAPITRE I

INTRODUCTION

Dans le monde de l'assurance, le travail d'un actuaire est de quantifier le plus justement possible le risque financier que représente un client ou un potentiel client. Avant même de songer à comment y parvenir, il est crucial de posséder des données sur lesquelles travailler. Bien que celles-ci se font rares dans certains segments de l'assurance « incendie, accidents et risques divers » (IARD), par exemple des données sur les sinistres résultant d'inondations en assurance habitation, ce n'est heureusement pas le cas en assurance automobile.

Pour donner une idée, le Groupement des assureurs automobiles du Québec (GAA), organisme qui s'occupe notamment de la cueillette et de la compilation de données statistiques de l'ensemble des assureurs québécois, révèle dans son Plan statistique automobile de 2014 [1] que 4 934 565 voitures de tourisme étaient assurées pour la responsabilité civile (chapitre A) au Québec en 2014. Pour ce chapitre, qui inclut principalement les sinistres payés aux véhicules des assurés non responsables d'une collision qui s'est produite au Québec, une fréquence de 3.68 % fut enregistrée en 2014, ce qui équivaut exactement à 181 403 sinistres. À l'échelle mondiale, l'organisation WardsAuto (Sousanis, 2011) a estimé que plus d'un milliard de véhicules étaient en circulation dans le monde en 2010. Ces chiffres impressionnants

1. Rapport sommaire 2014, GAA. https ://www.gaa.qc.ca/documents/99991A2F-S.pdf

démontrent bien toute la pertinence et l'importance du travail de l'actuaire en assurance auto, c'est-à-dire de bien quantifier ces risques pour offrir la meilleure protection aux clients sans mettre en danger la pérennité de la compagnie.

Malgré l'abondance générale de données en assurance automobile, rares sont les assureurs qui possèdent des renseignements fiables sur l'utilisation de chaque véhicule assuré. La plupart des assureurs utilisent aujourd'hui une estimation du kilométrage annuel comme variable de tarification. Par contre, cette estimation est habituellement fournie par les clients. De plus, les compagnies d'assurance définissent généralement un nombre faible de classes qui segmentent la variable du kilométrage (Schwartz, 2004).

Depuis quelques années, de plus en plus d'assureurs offrent un produit d'assurance qui, de différentes façons, va incorporer des données sur les habitudes de conduite. Ces données peuvent être recueillies de différentes façons, par exemple via une application mobile ou encore un système GPS (*Global Positioning System*) installé directement dans la voiture.

L'objectif de ce mémoire est de montrer de quelle façon des modèles statistiques avancés, soient les modèles additifs généralisés (GAM), peuvent être extrêmement utiles pour comprendre et analyser l'impact qu'ont certaines variables liées à la conduite sur le risque d'accident automobile. Dans ce mémoire, l'analyse sera portée sur le kilométrage exact parcouru et la durée d'exposition au risque.

Dans un premier temps, ce présent chapitre couvrira les notions de base de l'assurance automobile et rappellera la théorie sur les modèles linéaires généralisés (GLM). Puis, le chapitre 2 introduira et couvrira exhaustivement la théorie sur les modèles additifs généralisés. Le chapitre 3 appliquera directement les concepts développés au chapitre 2 dans un contexte actuariel où des données recueillies par GPS serviront à la construction d'une structure tarifaire en assurance automobile.

Finalement, le chapitre 4 concluera le mémoire.

1.1 Prime pure et autres définitions préalables

Cette section présentera quelques définitions du jargon de l'assurance générale et mettra la table au reste du chapitre.

Exposition au risque : L'exposition au risque correspond à la durée pendant laquelle une personne a pu bénéficier d'une couverture d'assurance achetée auprès d'un assureur. En assurance auto, les contrats sont généralement d'une durée d'un an.

Fréquence : La fréquence est définie comme étant le nombre de sinistres par unité d'exposition.

Sévérité : La sévérité est définie comme le coût moyen d'un sinistre.

Prime pure : La prime pure correspond au coût moyen d'assurance par unité d'exposition.

Il est très simple de montrer que la prime pure n'est qu'en fait la fréquence multipliée par la sévérité :

$$\begin{aligned} \text{Prime pure} &= \frac{\text{Coûts totaux encourus}}{\text{Nb. unités d'exposition}} \\ &= \frac{\text{Nb. de sinistres}}{\text{Nb. unités d'exposition}} \times \frac{\text{Coûts totaux encourus}}{\text{Nb. de sinistres}} \\ &= \text{Fréquence} \times \text{Sévérité}. \end{aligned} \tag{1.1}$$

Un des mandats des actuaires oeuvrant au sein d'une compagnie d'assurance est donc de déterminer la prime pure pour chaque assuré. Une des manières d'y parvenir serait de calculer (1.1) pour chacun des assurés. Une telle pratique supposerait par contre que le passé est exactement garant du futur, ce qui n'est pas très sensé. De plus, les gens qui soumettent une réclamation à leur assureur représentent

chaque année une minorité parmi l'ensemble du portefeuille d'assurés, ce qui fait que des problèmes de données seraient rencontrés.

Pour ces raisons, on associe la prime pure à une variable aléatoire. Classiquement, l'espérance mathématique de la prime pure est modélisée à l'aide de modèles mathématiques qui permettent, une fois ajustés, de déterminer une prédiction pour la prime pure.

Espérance mathématique : L'espérance mathématique est la formalisation probabiliste du concept de moyenne arithmétique. Concrètement, l'espérance d'une variable aléatoire S, notée $\mathbb{E}[S]$, est la valeur moyenne que l'on attendrait si l'on pouvait répéter la même expérience aléatoire un très grand nombre de fois.

Si l'on suppose que la fréquence, notée par Y, et la sévérité, notée par X, sont des variables aléatoires indépendantes, on a alors que

$$\mathbb{E}[S] = \mathbb{E}[Y]\,\mathbb{E}[X],$$

où $\mathbb{E}[S]$ est l'espérance mathématique de la prime pure.

Dans ce mémoire, on ne s'intéressera exclusivement qu'à la fréquence et à sa modélisation. Ainsi, pour le reste de l'ouvrage, lorsque le mot « prime » est utilisé, il faut garder en tête que l'on est en train de parler de $\mathbb{E}[Y]$. Très souvent, en pratique, l'hypothèse que la fréquence de réclamations suive une loi de Poisson est effectuée.

Loi de Poisson : La loi de Poisson est une loi de probabilité discrète qui a été proposée par le mathématicien français Siméon-Denis Poisson en 1837 dans son ouvrage « Recherches sur la probabilité des jugements en matière criminelle et en

matière civile ». La fonction de probabilité de Poisson est définie comme suit :

$$\Pr[Y = y] = \begin{cases} \frac{\lambda^y e^{-\lambda}}{y!} & \text{si } y = 0, 1, 2, ... \\ 0 & \text{sinon,} \end{cases} \tag{1.2}$$

où λ est un nombre réel strictement positif et dit paramètre de la distribution de Poisson.

Il peut être démontré que l'espérance mathématique et la variance de la loi de Poisson sont égales à la valeur de son paramètre. Ainsi, si $Y \sim Poisson(\lambda)$, on a l'égalité suivante :

$$\mathbb{E}[Y] = Var[Y] = \lambda.$$

1.2 Segmentation

La segmentation en assurance est définie comme le processus par lequel l'assureur distingue les risques qu'il accepte d'assurer. Cette opération est nécessaire, car elle permet de créer des classes homogènes de risque et de leur appliquer un traitement commun et adéquat. Par exemple, les assureurs distinguent généralement les hommes des femmes, les véhicules sports des véhicules familiaux, etc.

Une segmentation efficace permet aux assureurs de se protéger contre l'antisélection, phénomène qui peut menacer la viabilité de n'importe quel joueur du monde de l'assurance. On parle d'antisélection lorsqu'un assuré détient des informations sur son propre risque d'accident qu'un assureur ignore. Cette asymétrie dans l'information peut se traduire en pertes importantes pour un assureur qui en est victime.

Imaginons un exemple simple où deux assureurs, A et B, offrent une couverture pour dommages matériels survenus à la suite d'un accident d'auto. L'assureur A offre la couverture à un coût de 100 dollars. Du côté de l'assureur B, le coût

est aussi de 100 dollars, mais une surcharge de 10% est applicable si la voiture principale est une voiture sportive. L'assureur B considère donc que conduire une voiture sportive représente un risque d'accident plus élevé que conduire une voiture standard. Dans un exemple comme celui-ci, un assuré ayant une voiture sport qui magasine sa prime d'assurance choisira l'assureur A en raison du prix inférieur. L'assureur A attirera ainsi beaucoup de conducteurs de voitures sportives et leur chargera une prime d'assurance qui ne réflète pas le « vrai » risque financier que ces assurés représentent. À terme, la compagnie A subira de grosses pertes et pourrait devoir se retirer du marché, à moins bien sûr qu'elle ne revoit sa segmentation et tarification en général.

La prochaine section montrera de quelle façon les modèles linéaires généralisés permettent d'intégrer la segmentation des risques dans une distribution de probabilité choisie pour expliquer un phénomène aléatoire.

1.3 Modèles linéaires généralisés

Les modèles linéaires généralisés (GLM), proposés par Nelder et Wedderburn (1972), sont une généralisation du concept de régression linéaire. Dans un modèle de régression linéaire, on tente d'expliquer la valeur d'une quantité inconnue (variable dépendante ou réponse) par une combinaison linéaire de d'autres facteurs (variables indépendantes ou explicatives). La combinaison linéaire implique que la variation de la variable réponse est directement proportionnelle à la variation d'une variable explicative. De plus, le cadre classique de la régression linéaire suppose que la variable réponse suit une distribution normale (Wood 2006, section 1.1.3).

Or, il arrive souvent en pratique que ces hypothèses ne soient pas appropriées. L'idée derrière les GLM est justement de relaxer ces hypothèses. D'une part, le

cadre théorique des GLM permet de supposer une loi de probabilité autre que la loi normale pour la variable dépendante. D'autre part, l'introduction d'une fonction de lien permet de modéliser une transformation de l'espérance de la variable réponse.

1.3.1 Structure générale

Concrètement, un modèle linéaire généralisé a la structure

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_l x_{li} \\ &= \mathbf{X}_i \boldsymbol{\beta}, \end{aligned} \tag{1.3}$$

où $\mu_i \equiv \mathbb{E}[Y_i]$, $\mathbf{X}_i$ est la i-ème ligne d'une matrice de design formée par les variables explicatives x_j, $j = 1, ..., l$, et $\boldsymbol{\beta}$ est un vecteur de paramètres à estimer. La fonction g est appelée fonction de lien et permet de lier la moyenne de la distribution que suit Y_i au prédicteur linéaire. Concernant la distribution de probabilité, celle-ci doit appartenir à la famille exponentielle linéaire.

1.3.2 Famille exponentielle linéaire

Pour toute variable aléatoire Y, la famille exponentielle linéaire correspond à un ensemble de distributions dont la loi de probabilité (discrète ou continue) peut être exprimée sous la forme

$$f_\theta(y) = \exp\left[\{y\theta - b(\theta)\}/a(\phi) + c(y, \phi)\right], \tag{1.4}$$

où b, a et c sont des fonctions arbitraires. De plus, ϕ est le paramètre canonique alors que θ est le paramètre de dispersion. Des développements peuvent être faits pour obtenir des formulations générales pour l'espérance et la variance d'une distribution appartenant à la famille exponentielle linéaire (Wood 2006, section 2.1.1).

En effet, on a que

$$\mathbb{E}[Y] = b'(\theta) \tag{1.5}$$

et

$$\begin{aligned} Var[Y] &= b''(\theta)a(\phi) \\ &= b''(\theta)\phi/\omega \\ &= V(\mu)\phi, \end{aligned} \tag{1.6}$$

où $b'(\theta)$ et $b''(\theta)$ sont les dérivées premières et secondes de b par rapport à θ. De plus, $\mu \equiv \mathbb{E}[Y]$, $a(\phi) = \phi/\omega$ et $V(\mu) = b''(\theta)/\omega$, où ω est une constante connue (habituellement 1).

Parmi les distributions notoires appartenant à la famille exponentielle linéaire, notons les loi normale, gamma, binomiale et Poisson.

1.3.3 Estimation par maximum de vraisemblance

L'estimation du vecteur de paramètres $\boldsymbol{\beta}$ dans (1.3) est effectuée par maximum de vraisemblance. L'intuition derrière cette technique d'estimation est que l'on va chercher le vecteur $\boldsymbol{\beta}$ qui maximisera la densité conjointe que la variable réponse prenne chaque valeur observée dans les données de modélisation, dépendamment de la loi de probabilité choisie et de l'ensemble des valeurs observées pour les variables explicatives.

Donc, en supposant un contexte où $Y_i \sim f_{\theta_i}(y_i)$, où l'on a des observations pour une variable aléatoire Y qui suit une loi de probabilité appartenant à la famille exponentielle, la fonction de vraisemblance est définie comme

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i).$$

Généralement, la fonction de vraisemblance est maximisée via maximisation de la fonction de log-vraisemblance, pour des raisons de simplicité. Ainsi, on a pour fonction de log-vraisemblance

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)] \\ &= \sum_{i=1}^{n} \left[y_i\theta_i - b_i(\theta_i)\right]/a_i(\phi) + c_i(y_i,\phi). \end{aligned} \tag{1.7}$$

En posant $a_i(\phi) = \phi/\omega_i$, l'estimation de $\boldsymbol{\beta}$ se fait en égalant à 0 les dérivées partielles de (1.7) par rapport à chaque élément β_j constituant le vecteur $\boldsymbol{\beta}$:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} \right) = 0.$$

En utilisant le principe des dérivations en chaîne et les définitions (1.5) et (1.6), on trouve

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \tag{1.8}$$

qui est la condition de premier ordre à résoudre pour l'estimation de $\boldsymbol{\beta}$ par maximum de vraisemblance.

CHAPITRE II

INTRODUCTION AUX MODÈLES ADDITIFS GÉNÉRALISÉS

2.1 Idée générale

Les modèles additifs généralisés (GAM), d'abord proposés par Hastie et Tibshirani (1986), sont une extension des modèles linéaires généralisés (GLM). La différence se retrouve au niveau du prédicteur linéaire. Dans le cas du GLM, on retrouvait la structure générale définie par l'équation (1.3). On souhaitait alors expliquer l'espérance d'une variable réponse Y_i directement par les valeurs que peuvent prendre les variables explicatives x_j, $j = 1, ..., l$.

Quant à eux, les modèles additifs généralisés permettent une plus grande flexibilité que les modèles linéaires généralisés. Le prédicteur reste sous forme linéaire, mais on souhaite désormais expliquer la variable étudiée par des fonctions de variables explicatives. Ces fonctions représentent une manière de traiter la relation non linéaire que peuvent avoir certaines variables avec la variable à expliquer. On a ainsi une façon plus souple et plus générale d'expliquer la dépendance entre la moyenne μ_i de la variable Y_i et les variables explicatives x_j. Une structure possible d'un modèle GAM pourrait donc être

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}), \tag{2.1}$$

où les f_j sont des fonctions des variables x_j. La partie droite de l'équation (2.1)

est dite non-paramétrique. Les fonctions impliquées dans la modélisation devront bien entendu être estimées.

Le GAM représente un outil très intéressant pour étudier le comportement de certains facteurs sur une variable d'étude. Il est d'autant plus utile lorsque la relation est soupçonnée d'être non linéaire : il permet d'aller au-delà de la relation paramétrique. En fait, une approche non-paramétrique laisse davantage « parler » les données. Cependant, le GAM apporte son lot de complexité. Une des premières interrogations à survenir concerne la définition des différentes fonctions f_j présentes dans (2.1). Dans la prochaine section, nous verrons pourquoi les splines seront utiles à cet effet.

2.2 Introduction aux splines

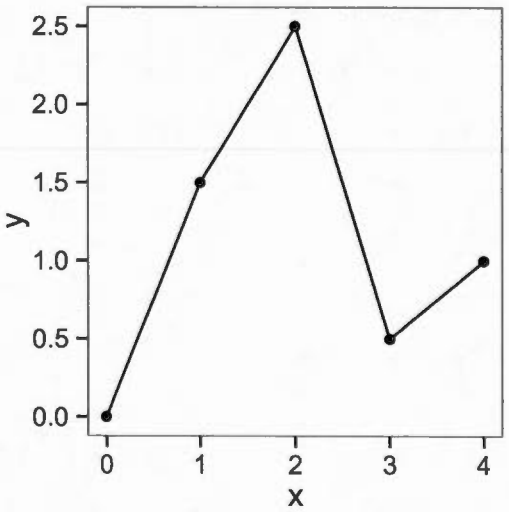
Pour comprendre de quelle façon on construit un modèle additif généralisé, il est important de savoir ce que sont les splines. Essentiellement, une spline est une fonction polynomiale définie par morceaux. Il existe plusieurs types de splines, mais nous nous concentrerons surtout sur les splines cubiques. On nomme celles-ci de cette façon puisque dans leur caractérisation, le polynôme dont le degré est le plus élevé en est un de degré 3. Mais avant d'aller au degré 3, passons par le degré 1.

2.2.1 Spline linéaire d'interpolation

Supposons que nous avons un jeu de données $\{x_j; y_j\}$, où $j = 1, 2, ..., n$ et $x_j < x_{j+1}$. Nous sommes intéressés à trouver une fonction s qui interpolera les points $(x_j; y_j)$ de telle sorte que $s(x_j) = y_j \ \forall j$. Autrement dit, le but du processus est que la fonction passe par l'ensemble des points. Une solution évidente serait de relier tous les points ensemble à l'aide de lignes droites.

Par exemple, supposons que nous avons les couples suivants : (0;0), (1;1.5), (2;2.5), (3;0.5) et (4;1). La figure 2.1 est la représentation graphique de la « spline linéaire d'interpolation » [1] construite à l'aide de ces données.

Figure 2.1: Spline linéaire d'interpolation



Plus formellement, la spline linéaire d'interpolation est définie par la fonction par morceaux suivante :

$$s(x) = \frac{(x_{j+1} - x)}{h_j} s(x_j) + \frac{(x - x_j)}{h_j} s(x_{j+1}) \text{ si } x_j \leq x \leq x_{j+1}, \tag{2.2}$$

où $h_j = x_{j+1} - x_j$. Une telle approche, par sa simplicité, peut valoir son pesant d'or. Par contre, si les différents couples $(x_j; y_j)$ ont été tirés d'une fonction f qui, graphiquement, affiche une allure courbée, alors nécessairement la distance $|s(x) - f(x)| \geq 0$ pour toutes les valeurs possibles de x. Une manière plus efficace de résoudre notre problème initial est donc d'utiliser la « spline cubique

1. Traduction libre de l'expression *linear interpolating spline*.

d'interpolation » [2].

2.2.2 Spline cubique d'interpolation

L'interpolation par spline cubique va en général nous permettre de mieux approximer une fonction lisse sous-jacente à un ensemble de données que l'interpolation par spline linéaire. En fait, le but est que les dérivées premières et secondes de la fonction par morceaux construite soient continues autant dans un intervalle spécifié qu'aux noeuds d'interpolation. Ces conditions auront pour résultat d'induire une courbure continue à la spline. Comme notre objectif est toujours d'avoir une courbe qui passe par tous les points du jeu de données, les noeuds d'interpolation correspondent ici aux valeurs x_j des points $(x_j; y_j)$.

Pour la construction, on va donc développer la formule d'interpolation de manière à ce que la dérivée seconde de la spline varie de façon linéaire dans un intervalle donné. En procédant ainsi, on s'assure de respecter la condition stipulant que la dérivée seconde doit être continue. Or, on a que

$$s''(x) = \frac{(x_{j+1} - x)}{h_j} s''(x_j) + \frac{(x - x_j)}{h_j} s''(x_{j+1}) \text{ si } x_j \leq x \leq x_{j+1}, \tag{2.3}$$

où h_j est défini de la même façon que pour l'équation (2.2). Comme on veut $s(x)$, on doit intégrer deux fois $s''(x)$. Pour la première intégrale, on a

$$\begin{aligned} s'(x) &= \int s''(x)dx \\ &= \int \left[\frac{(x_{j+1} - x)}{h_j} s''(x_j) + \frac{(x - x_j)}{h_j} s''(x_{j+1}) \right] dx \\ &= \frac{-(x_{j+1} - x)^2}{2h_j} s''(x_j) + \frac{(x - x_j)^2}{2h_j} s''(x_{j+1}) + C_1, \end{aligned} \tag{2.4}$$

2. Traduction libre de l'expression *cubic interpolating spline*.

où C_1 est une première constante d'intégration. On intègre de nouveau de telle sorte que

$$\begin{aligned} s(x) &= \int s'(x)dx \\ &= \int \left[\frac{-(x_{j+1}-x)^2}{2h_j} s''(x_j) + \frac{(x-x_j)^2}{2h_j} s''(x_{j+1}) + C_1 \right] dx \\ &= \frac{(x_{j+1}-x)^3}{6h_j} s''(x_j) + \frac{(x-x_j)^3}{6h_j} s''(x_{j+1}) + C_1 x + C_2, \end{aligned} \tag{2.5}$$

où C_2 est une deuxième constante d'intégration. Il ne nous reste plus qu'à trouver les constantes d'intégration pour avoir une expression pour $s(x)$. Afin de faire ceci, on utilise les deux conditions aux bornes suivantes :

- $s(x_j) = y_j$;
- $s(x_{j+1}) = y_{j+1}$.

Nous obtenons ainsi ces deux équations à résoudre pour C_1 et C_2 :

$$\begin{aligned} y_j &= \frac{(x_{j+1}-x_j)^3}{6h_j} s''(x_j) + C_1 x_j + C_2; \\ y_{j+1} &= \frac{(x_{j+1}-x_j)^3}{6h_j} s''(x_{j+1}) + C_1 x_{j+1} + C_2. \end{aligned}$$

En se rappelant que $h_j = x_{j+1} - x_j$, on trouve que

$$\begin{aligned} C_1 &= \frac{y_{j+1}}{h_j} - \frac{y_j}{h_j} + s''(x_j)\frac{h_j}{6} - s''(x_{j+1})\frac{h_j}{6}; \\ C_2 &= \frac{y_j x_{j+1}}{h_j} - \frac{y_{j+1}x_j}{h_j} - s''(x_j)\frac{x_{j+1}h_j}{6} + s''(x_{j+1})\frac{x_{j+1}h_j}{6} - s''(x_{j+1})\frac{h_j^2}{6}. \end{aligned}$$

Maintenant, en insérant les expressions obtenues pour les constantes d'intégration dans (2.5) et en effectuant quelques manipulations algébriques, on obtient la formule suivante pour la spline cubique d'interpolation :

$$s(x) = a_j^-(x)y_j + a_j^+(x)y_{j+1} + c_j^-(x)s''(x_j) + c_j^+(x)s''(x_{j+1}) \text{ si } x_j \le x \le x_{j+1}, \tag{2.6}$$

où

$$a_j^-(x) = (x_{j+1} - x)/h_j \qquad c_j^-(x) = \left[(x_{j+1} - x)^3/h_j - h_j(x_{j+1} - x)\right]/6$$
$$a_j^+(x) = (x - x_j)/h_j \qquad c_j^+(x) = \left[(x - x_j)^3/h_j - h_j(x - x_j)\right]/6.$$

L'équation (2.6) spécifie complètement la courbe recherchée en fonction des valeurs que prennent les dérivées premières et secondes aux noeuds d'interpolation. Le seul problème est qu'en pratique, on ne connait pas les valeurs des dérivées secondes $s''(x_j)$. Ceci constitue une différence par rapport à la formule (2.2) que nous avions pour la spline linéaire d'interpolation où l'on n'avait aucun inconnu. Pour contourner ce problème, on va se servir de la condition inhérente aux splines cubiques (évoquée en début de section) qui mentionne que les dérivées premières doivent être continues aux noeuds séparant deux intervalles adjacents. Mathématiquement, prenons les deux intervalles $[x_j, x_{j+1}]$ et $[x_{j+1}, x_{j+2}]$, on doit avoir que l'équation (2.4) évaluée au point x_{j+1} retourne la même valeur pour les deux intervalles. On aura donc que

$$s'(x_{j+1})_{[x_j, x_{j+1}]} = s'(x_{j+1})_{[x_{j+1}, x_{j+2}]}$$

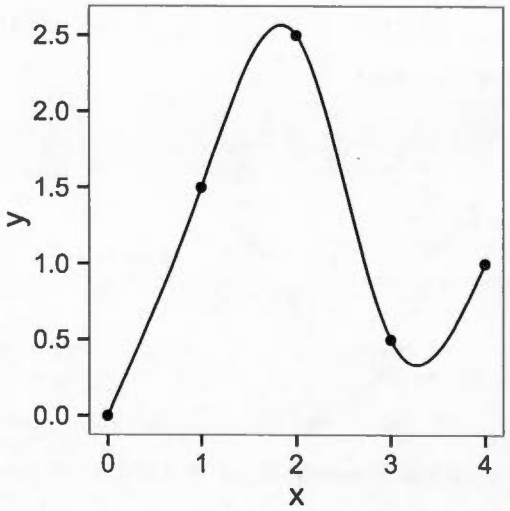
$$\frac{h_j}{6}s''(x_j) + \frac{(x_{j+2} - x_j)}{3}s''(x_{j+1}) + \frac{h_{j+1}}{6}s''(x_{j+2}) = \frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j}, \quad (2.7)$$

pour $j = 1, 2, ..., n-2$. Or, on se retrouve avec un système de $n-2$ équations avec n inconnus à résoudre. Ceux-ci correspondent aux $s''(x_j)$, où $j = 1, 2, ..., n$. Afin d'avoir une solution unique, il nous faut émettre deux autres conditions. Typiquement, elles concernent les bornes x_1 et x_n. La façon la plus commune de procéder est d'imposer que $s''(x_1) = s''(x_n) = 0$. Lorsque ces conditions sont appliquées, la spline que l'on obtient pour résultat se nomme « spline cubique naturelle ».

Exemple : spline cubique naturelle d'interpolation

Reprenons le même exemple de la page 13. Nous avions alors les couples de données (0;0), (1;1.5), (2;2.5), (3;0.5) et (4;1). Avec l'aide du résultat (2.7), on est en mesure d'écrire un système de trois équations avec trois inconnus à résoudre. Les inconnus correspondent aux dérivées secondes évaluées aux noeuds (en excluant les deux noeuds extrêmes, où les dérivées secondes sont nulles). Une fois le système résolu, chaque section peut être écrite sous la forme d'une fonction polynomiale cubique à l'aide de l'équation (2.6). La spline cubique naturelle résultante est définie par l'ensemble de ces fonctions polynomiales cubiques. Tous les détails de cet exemple sont présentés à l'annexe A.1. La figure 2.2 montre à quoi ressemble visuellement la solution de cet exemple.

Figure 2.2: Spline cubique naturelle



2.2.3 Spline cubique d'ajustement

Dans la section 2.2.2, on a vu comment interpoler des données à l'aide d'une spline cubique naturelle. Dépendamment du contexte dans lequel on travaille, ou de l'objectif que l'on souhaite atteindre, l'interpolation n'est parfois pas satisfaisante. Dans plusieurs cas, on va plutôt vouloir trouver une fonction qui va effectuer un lissage (passer « au travers » des points). Par exemple, en statistique, les données sont souvent mesurées avec une « erreur ». Or, si l'on cherche une fonction d'interpolation, on reproduit en quelque sorte les erreurs puisque l'on ne fait que répliquer les données. On voudrait, à l'inverse, faire notre possible pour les éliminer. Dans le cas où l'on travaille avec des données exactes, disons le cours d'une action sur un certain horizon de temps, on s'intéressera rarement aux variations très locales, mais bien à la tendance générale.

Ainsi, on souhaite estimer une courbe qui va répondre à deux objectifs :

1. Approcher le plus possible les points de notre jeu de données ;
2. Prioriser la variation générale aux variations locales de notre variable à l'étude (Lissage).

Maintenant, il faut trouver comment gérer le fait d'avoir deux objectifs contradictoires. La solution consiste à développer un critère à minimiser qui va permettre de trouver une fonction dite d'ajustement ou de lissage (*smooth function*).

Supposons toujours que nous avons les couples de coordonnées $(x_j; y_j)$, où $j = 1, 2, ..., n$ et $x_j < x_{j+1}$. Admettons également g, la fonction de lissage que nous cherchons. On va relaxer la condition associée à l'interpolation (discutée à la page 15) qui imposerait ici que $g(x_j) = y_j$ et plutôt traiter les $g(x_j)$ comme n paramètres à estimer.

Pour le premier objectif, soit d'obtenir une courbe qui se rapproche des données,

on va utiliser la somme des carrés des différences entre les points observés et la fonction g. Pour le deuxième objectif, soit d'effectuer un lissage des données, on va instaurer une « pénalité pour irrégularité » (*roughness penalty*). La forme générale du critère d'ajustement va donc s'apparenter à

$$\sum_{j=1}^{n} \{y_j - g(x_j)\}^2 + \lambda \cdot \text{pénalité}, \tag{2.8}$$

où λ est appelé paramètre de lissage (*smoothing parameter*). Pour le moment, il est considéré arbitraire. Le paramètre de lissage est utilisé pour donner plus d'importance à l'un ou l'autre des objectifs décrits plus haut.

Nous avons maintenant établi (2.8), mais il reste à définir plus précisément g et la pénalité pour irrégularité pour être en mesure de faire quoi que ce soit.

Une façon intuitive de quantifier le niveau d'irrégularité d'une fonction doublement dérivable g, définie sur son domaine $[x_1, x_n]$, est de calculer l'intégrale du carré de sa dérivée seconde, soit

$$\int_{x_1}^{x_n} g''(x)^2 dx.$$

En fait, si une fonction oscille beaucoup, donc monte et descend à plusieurs reprises, alors plus sa dérivée première subit d'importantes variations. Ce faisant, plus la dérivée seconde doit être grande, puisque celle-ci quantifie la courbure qu'induisent les oscillations à la fonction. Comme on s'intéresse à l'ampleur de la courbure, et non pas à la concavité ou la convexité de la fonction à un certain point, on utilise le carré de la dérivée seconde. Ainsi, avec cette approche pour la pénalité, (2.8) devient

$$\sum_{j=1}^{n} \{y_j - g(x_j)\}^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 dx. \tag{2.9}$$

Maintenant, il reste à définir la forme de la fonction g. Une définition pour g s'impose d'elle-même : la spline cubique naturelle. En effet, les qualités en matière

d'interpolation de celle-ci en font une option idéale dans la mesure où l'on doit minimiser (2.9).

Pour comprendre pourquoi, **revenons momentanément dans un contexte d'interpolation.** Parmi toutes les fonctions continues sur $[x_1, x_n]$ dont les dérivées premières sont continues, et qui de plus interpolent l'ensemble des couples $(x_j; y_j)$, $s(x)$, définie par (2.6) et soumise aux conditions $s''(x_1) = s''(x_n) = 0$, est la fonction la plus lisse qui soit au sens de la minimisation de

$$J(g) = \int_{x_1}^{x_n} g''(x)^2 dx.$$

Démonstration

La preuve est tirée de Wood (2006), section 4.1.1. Supposons $\widetilde{g}(x)$ une fonction d'interpolation quelconque pour le jeu de données $\{x_j; y_j\}$, mais autre que $s(x)$. Définissons $h(x) = \widetilde{g}(x) - s(x)$ et réécrivons $J(\widetilde{g})$ pour avoir $J(s)$:

$$\begin{aligned}\int_{x_1}^{x_n} \widetilde{g}''(x)^2 dx &= \int_{x_1}^{x_n} \{s''(x) + h''(x)\}^2 \, dx \\ &= \int_{x_1}^{x_n} s''(x)^2 dx + 2\int_{x_1}^{x_n} s''(x)h''(x)dx + \int_{x_1}^{x_n} h''(x)^2 dx.\end{aligned}$$

En intégrant le deuxième terme par parties, on obtient

$$\begin{aligned}\int_{x_1}^{x_n} s''(x)h''(x)dx &= s''(x_n)h'(x_n) - s''(x_1)h'(x_1) - \int_{x_1}^{x_n} s'''(x)h'(x)dx \\ &= -\int_{x_1}^{x_n} s'''(x)h'(x)dx \\ &= -\sum_{j=1}^{n-1} s'''(x_j^+) \int_{x_j}^{x_{j+1}} h'(x)dx \\ &= -\sum_{j=1}^{n-1} s'''(x_j^+) \{h(x_{j+1}) - h(x_j)\} \\ &= 0.\end{aligned}$$

L'égalité entre la première et la deuxième ligne s'explique par les conditions aux bornes $s''(x_1) = s''(x_n) = 0$ qu'implique la spline cubique naturelle. Dans la troisième ligne, $s'''(x_j^+)$ représente la dérivée troisième sur l'intervalle $[x_j, x_{j+1}]$, qui est une valeur constante étant donné que $s(x)$ est constituée de polynômes cubiques. De plus, par la définition de $h(x)$, qui est la différence entre les fonctions d'interpolation $\widetilde{g}(x)$ et $s(x)$, $h(x_j)$, pour $j = 1, ..., n$, a nécessairement une valeur nulle, ce qui valide la dernière égalité.

Comme le deuxième terme disparaît, on a donc que

$$\int_{x_1}^{x_n} \widetilde{g}''(x)^2 dx = \int_{x_1}^{x_n} s''(x)^2 dx + \int_{x_1}^{x_n} h''(x)^2 dx \geq \int_{x_1}^{x_n} s''(x)^2 dx,$$

avec une égalité possible uniquement lorsque $h''(x) = 0$ pour tout x, $x_1 < x < x_n$. Puisque $h(x_1) = h(x_n) = 0$, on peut dire de façon équivalente que l'égalité ne peut avoir lieu que lorsque $h(x) = 0$ pour $x_1 < x < x_n$. L'unique circonstance où l'on observera une telle chose est lorsque $\widetilde{g}(x) = s(x)$.

Par conséquent, on vient de démontrer que toute fonction d'interpolation $\widetilde{g}(x)$ non identique à $s(x)$ aura une plus grande valeur pour le carré de sa dérivée seconde. En d'autres mots, ceci fait de la spline cubique naturelle la fonction d'interpolation la plus lisse qui soit. ■

Maintenant, **revenons dans un contexte d'ajustement** où l'on souhaite minimiser (2.9). On cherche une définition pour la fonction de lissage $g(x)$.

Supposons une fonction $g(x)$, qui minimise effectivement (2.9). On peut alors décider d'interpoler tous les points $(x_j, g(x_j))$ à l'aide d'une spline cubique $s(x)$, définie par (2.6). Par conséquent, $g(x)$ et $s(x)$ produiront la même valeur pour le terme de gauche dans le critère d'ajustement (2.9), qui est en fait la somme des carrés des résidus. Quant au terme de droite, c'est-à-dire celui qui caractérise la pénalité pour irrégularité, $s(x)$ aura nécessairement une valeur moindre pour le

carré de sa dérivée seconde que $g(x)$ en raison de la propriété qui a fait l'objet de la démonstration précédente. On fait donc face à une contradiction : $g(x)$ ne peut être la fonction qui répond à notre objectif de minimiser (2.9). En réalité, la seule manière possible que ce soit le cas est d'avoir que $g(x) = s(x)$. Ainsi, on constate que la spline cubique naturelle est la solution qui découle « naturellement » de la spécification même du critère d'ajustement.

Dans cette section, on a établi l'équation (2.9) qui permet, lorsque minimisée, de lisser des données. On a également vu que la meilleure définition possible pour la fonction $g(x)$ est celle de la spline cubique naturelle. Étant donné que l'on n'est plus dans un contexte d'interpolation, mais bien de lissage, on parlera alors de spline cubique d'ajustement [3].

Le seul problème que l'on a désormais, c'est qu'il existe autant de paramètres à estimer qu'il y a de données. En effet, rappelons que dans (2.9), les valeurs pour $g(x_j)$, $j = 1, ..., n$, sont des paramètres inconnus à estimer. Évidemment, il y a un coût important associé aux calculs numériques permettant de trouver les meilleures estimations possible. En pratique, pour limiter le nombre de paramètres, on va utiliser la « régression pénalisée par spline cubique » [4] qui représente un bon compromis entre la rétention des propriétés des splines et l'efficacité des calculs numériques.

Les détails théoriques concernant la régression pénalisée par spline cubique ainsi que son rôle dans la construction d'un modèle additif généralisé seront couverts dans les prochaines sections.

---

3. Traduction libre de l'expression *cubic smoothing spline.*

4. Traduction libre de l'expression *penalized cubic regression spline.*

2.3 Construction d'un modèle additif généralisé

2.3.1 Fonction de lissage univariée

L'idée générale derrière l'utilisation d'un modèle GAM a été introduite à la section 2.1 à l'aide de l'équation (2.1). Celle-ci correspondait à une formulation possible pour un modèle GAM et impliquait alors trois fonctions de lissage univariées, soient $f_1(x_1)$, $f_2(x_2)$ et $f_3(x_3)$.

Les splines sont généralement utilisées pour définir les fonctions de lissage dans un GAM. Dans notre cas, l'intérêt continuera d'être porté sur les splines cubiques. Ainsi, nous représenterons toute fonction de lissage univariée par une spline cubique d'ajustement.

Pour faciliter le processus d'estimation des paramètres, il sera nécessaire d'exprimer chaque fonction de lissage univariée sous la forme linéaire suivante :

$$f(x) = \sum_{k=1}^{q} b_k(x)\beta_k, \tag{2.10}$$

où $\boldsymbol{\beta}$ est le vecteur de paramètres à estimer et $b_k(x)$ sont des fonctions créées par la « base par spline cubique » (*cubic spline basis*) de dimension q utilisée. En procédant de cette façon, des techniques similaires utilisées pour l'estimation d'un GLM pourront être utilisées plus tard pour l'estimation d'un GAM.

La prochaine section montrera donc en détails les étapes permettant d'exprimer une spline cubique d'ajustement sous la forme (2.10) puis d'estimer ses paramètres.

2.3.2 Régression pénalisée par spline cubique

Imaginons que l'on souhaite expliquer une variable aléatoire Y par une fonction d'une autre variable explicative, disons X. Des couples de valeurs observées

$(x_i ; y_i)$, $i = 1, ..., n$, constituent les données sur lesquelles on souhaite effectuer une régression pénalisée par spline cubique.

L'idée est de définir une spline cubique d'ajustement pour la variable X, mais de limiter le nombre de paramètres en divisant les données observées pour la variable X en sections. Les deux valeurs extrêmes ainsi que les points séparant deux sections sont appelés « noeuds » (*knots*). Généralement, les données sont divisées en utilisant les quantiles. Une des manières de paramétriser la spline cubique est de le faire en fonction des valeurs qu'elle prend aux noeuds. Au final, chaque section accueillera son polynôme cubique ajusté.

Supposons donc une spline cubique d'ajustement $s(x)$, définie similairement à (2.6). La différence provient du fait que l'on n'interpole plus les données, on a plutôt des paramètres qui correspondent aux valeurs inconnues que prend la spline à ces noeuds. Admettons donc q noeuds, soient w_j pour $j = 1, ..., q$. On pose $\beta_j = s(w_j)$ et $\delta_j = s''(w_j)$ de telle sorte que l'on va avoir que

$$s(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \text{ si } w_j \leq x \leq w_{j+1}, \qquad (2.11)$$

où les fonctions de base a_j^-, a_j^+, c_j^- et c_j^+ sont définies à la page 26 dans le tableau 2.1. Les paramètres β_j et δ_j sont inconnus et à estimer. Cependant, puisque l'on souhaite éventuellement écrire (2.11) sous la forme de l'équation (2.10), il serait plus commode de réexprimer (2.11) en fonction des β_j seulement.

Pour ce faire, on va utiliser deux conditions sous-jacentes à la définition (2.11) de la spline cubique. La première est que sa dérivée, évaluée au noeud w_j, doit être égale pour le polynôme cubique situé à gauche et à droite de ce dernier. La deuxième est que l'on a une spline cubique naturelle, ainsi les dérivées secondes aux extrémités, soient aux noeuds w_1 et w_q, sont nulles.

On dérive alors (2.11), et on obtient

$$s'(x) = -\frac{\beta_j}{h_j} + \frac{\beta_{j+1}}{h_j} + \delta_j\left(\frac{h_j}{6} - \frac{3(w_{j+1}-x)^2}{6h_j}\right) + \delta_{j+1}\left(\frac{3(x-w_j)^2}{6h_j} - \frac{h_j}{6}\right),$$

où $h_j = w_{j+1} - w_j$. La première condition stipule que

$$s'(w_{j+1})_{[w_j,w_{j+1}]} = s'(w_{j+1})_{[w_{j+1},w_{j+2}]}$$

$$-\frac{\beta_j}{h_j} + \frac{\beta_{j+1}}{h_j} + \delta_j\frac{h_j}{6} + \delta_{j+1}\frac{3h_j}{6} - \delta_{j+1}\frac{h_j}{6} = -\frac{\beta_{j+1}}{h_{j+1}} + \frac{\beta_{j+2}}{h_{j+1}} - \delta_{j+1}\frac{3h_{j+1}}{6} + \delta_{j+1}\frac{h_{j+1}}{6} - \delta_{j+2}\frac{h_{j+1}}{6}.$$

Après de simples manipulations algébriques, on obtient que

$$\frac{1}{h_j}\beta_j - \left(\frac{1}{h_j} + \frac{1}{h_{j+1}}\right)\beta_{j+1} + \frac{1}{h_{j+1}}\beta_{j+2} = \frac{h_j}{6}\delta_j + \left(\frac{h_j}{3} + \frac{h_{j+1}}{3}\right)\delta_{j+1} + \frac{h_{j+1}}{6}\delta_{j+2}.$$

Si l'on répète cette égalité pour $j = 1, ..., q-2$, on a en fait que

$$\mathbf{D}\boldsymbol{\beta} = \mathbf{B}\boldsymbol{\delta}^-, \tag{2.12}$$

où $\boldsymbol{\delta}^- = (\delta_2, ..., \delta_{q-1})^\top$, $\mathbf{B}$ et $\mathbf{D}$ étant définies dans le tableau 2.2. Si l'on multiplie chaque côté de l'équation (2.12) par $\mathbf{B}^{-1}$, on obtient

$$\mathbf{B}^{-1}\mathbf{D}\boldsymbol{\beta} = \boldsymbol{\delta}^-.$$

En définissant $\mathbf{F}^- = \mathbf{B}^{-1}\mathbf{D}$ et

$$\mathbf{F} = \begin{bmatrix} \mathbf{0} \\ \mathbf{F}^- \\ \mathbf{0} \end{bmatrix},$$

où $\mathbf{0}$ est un vecteur ligne constitué uniquement de zéros, on a que $\boldsymbol{\delta} = \mathbf{F}\boldsymbol{\beta}$. Les deux lignes nulles de $\mathbf{F}$ sont nécessaires, car $\delta_1 = \delta_q = 0$. La formule (2.11) peut maintenant être réécrite uniquement en termes de $\boldsymbol{\beta}$:

$$s(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\mathbf{F}_j\boldsymbol{\beta} + c_j^+(x)\mathbf{F}_{j+1}\boldsymbol{\beta} \text{ si } w_j \le x \le w_{j+1}. \tag{2.13}$$

Fonctions de base pour spline cubique d'ajustement	
$a_j^-(x) = (w_{j+1} - x)/h_j$	$c_j^-(x) = \left[(w_{j+1} - x)^3/h_j - h_j(w_{j+1} - x)\right]/6$
$a_j^+(x) = (x - w_j)/h_j$	$c_j^+(x) = \left[(x - w_j)^3/h_j - h_j(x - w_j)\right]/6$

Tableau 2.1: Fonctions de base. $h_j = w_{j+1} - w_j$.

$k = 1, ..., q-2$		$k = 1, ..., q-3$
B	**D**	**B**
$B_{k,k} = (h_k + h_{k+1})/3$	$D_{k,k} = 1/h_k$	$B_{k,k+1} = h_{k+1}/6$
	$D_{k,k+1} = -1/h_k - 1/h_{k+1}$	$B_{k+1,k} = h_{k+1}/6$
	$D_{k,k+2} = 1/h_{k+1}$	

Tableau 2.2: Éléments non nuls des matrices **B** et **D**

On souhaite désormais exprimer (2.13) sous la forme d'une combinaison linéaire, soit

$$s(x) = \sum_{k=1}^{q} b_k(x)\beta_k,$$

où q est la dimension de la base par spline (le nombre de noeuds) et $b_k(x)$ sont de nouvelles fonctions définies de manière à avoir une équivalence exacte avec (2.13). Le défi est de parvenir à trouver ces fonctions $b_k(x)$.

Pour ce faire, prenons un cas général. Supposons que nous avons une certaine variable observée, disons x. Cinq noeuds, $w_1, ..., w_5$, sont choisis. Le premier et le dernier correspondent aux valeurs extrêmes que prend la variable x. Les trois autres sont les quartiles de la distribution.

Maintenant, admettons une valeur pour x qui se trouve entre le deuxième et le troisième noeud ($w_j \leq x \leq w_{j+1}$, $j = 2$). Le truc est de développer l'expression

(2.13) pour y retrouver l'ensemble des cinq paramètres :

$$s(x) = a_j^-(x)\beta_2 + a_j^+(x)\beta_3 + c_j^-(x)\left[F_{j,1}\beta_1 + F_{j,2}\beta_2 + F_{j,3}\beta_3 + F_{j,4}\beta_4 + F_{j,5}\beta_5\right] + c_j^+(x)\left[F_{j+1,1}\beta_1 + F_{j+1,2}\beta_2 + F_{j+1,3}\beta_3 + F_{j+1,4}\beta_4 + F_{j+1,5}\beta_5\right].$$

Par souci de clarté, seulement les indices j associés aux β ont été remplacés. On poursuit le développement en effectuant quelques regroupements de telle sorte que l'on a

$$\begin{aligned} s(x) = &\left[c_j^-(x)F_{j,1} + c_j^+(x)F_{j+1,1}\right]\beta_1 + \left[c_j^-(x)F_{j,2} + c_j^+(x)F_{j+1,2} + a_j^-(x)\right]\beta_2 \\ &+ \left[c_j^-(x)F_{j,3} + c_j^+(x)F_{j+1,3} + a_j^+(x)\right]\beta_3 + \left[c_j^-(x)F_{j,4} + c_j^+(x)F_{j+1,4}\right]\beta_4 \\ &+ \left[c_j^-(x)F_{j,5} + c_j^+(x)F_{j+1,5}\right]\beta_5. \end{aligned}$$

Si l'on faisait le même exercice pour toutes les sections entre deux noeuds, on retrouverait des résultats similaires pour chacune. Ceci nous permet enfin de redéfinir complètement la spline cubique de la façon suivante :

$$s(x) = \sum_{k=1}^{q} b_k(x)\beta_k, \tag{2.14}$$

où

$$b_k(x) = \begin{cases} c_j^-(x)F_{j,k} + c_j^+(x)F_{j+1,k} + a_j^+(x) & \text{si } k = j+1 \\ c_j^-(x)F_{j,k} + c_j^+(x)F_{j+1,k} + a_j^-(x) & \text{si } k = j \\ c_j^-(x)F_{j,k} + c_j^+(x)F_{j+1,k} & \text{sinon,} \end{cases}$$

et où l'indice j est déterminé selon la section dans laquelle se trouve la valeur x ($w_j \leq x \leq w_{j+1}$).

Avec le résultat précédent, en admettant une série de valeurs connues pour une variable x, il est possible d'évaluer la spline à ces valeurs et ainsi créer une matrice de design qui servira à l'évaluation du vecteur de paramètres $\boldsymbol{\beta}$.

Il nous faut désormais introduire la pénalité associée à la spline cubique, qui est définie par

$$\int s''(x)^2 dx = \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}, \tag{2.15}$$

où $s(x)$ est définie par (2.13) ou (2.14) et $\mathbf{S} = \mathbf{D}^\top \mathbf{B}^{-1} \mathbf{D}$. Communément, $\mathbf{S}$ est appelée « matrice de pénalité » (*penalty matrix*). Sans faire la démonstration complète du résultat, qui peut être trouvée dans Wood 2006 (annexe B.4, exercice 2), voici l'essentiel de la démarche :

1. Dériver deux fois (2.13) et réexprimer le résultat sous forme d'une combinaison linéaire ;
2. À l'aide des développements de l'étape 1, il peut être démontré que $\int s''(x)^2 dx = \boldsymbol{\delta}^{-\top} \mathbf{B} \boldsymbol{\delta}^{-}$;
3. Finalement, en utilisant (2.12), il est facile de montrer que l'expression $\boldsymbol{\delta}^{-\top} \mathbf{B} \boldsymbol{\delta}^{-}$ est égale à $\boldsymbol{\beta}^\top \mathbf{D}^\top \mathbf{B}^{-1} \mathbf{D} \boldsymbol{\beta}$.

Maintenant que la base par spline cubique et sa pénalité associée sont développées, le pont entre la théorie présentée sur la spline cubique d'ajustement (section 2.2.3) et celle présentée ici sur la régression par spline cubique peut être clairement établi. Pour fins de rappel, nous avions développé le critère à minimiser suivant dans le cadre des splines cubiques d'ajustement :

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int_{x_1}^{x_n} g''(x)^2 dx.$$

On a vu que la solution pour la forme de $g(x)$ qui émerge de la spécification même de ce critère est la spline cubique naturelle. Le problème que l'on avait alors était que nous avions autant de paramètres $g(x_i)$ à estimer qu'il y avait de données. La régression pénalisée par spline cubique corrige ce problème, puisque l'on va poser

$$g(x_i) = \mathbf{X}_i \boldsymbol{\beta},$$

où $\mathbf{X}_i$ est la i-ième ligne de la matrice de design $\mathbf{X}$ construite avec la définition (2.14). De façon plus précise, on a

$$\mathbf{X}_i = [b_1(x_i), b_2(x_i), ..., b_q(x_i)]\,, \tag{2.16}$$

où q est la dimension choisie de la base par spline. Ceci nous permet de réduire considérablement le nombre de paramètres à estimer.

En admettant un jeu de données comprenant 10 000 observations, on se retrouve ainsi à devoir estimer q paramètres β au lieu de 10 000 paramètres. Généralement, q est un nombre faible, mais choisi suffisamment grand pour couvrir correctement l'ensemble des données. Par exemple, nous pourrions avoir 10 noeuds. Il ne semble pas y avoir de consensus auprès de la communauté scientifique au sujet de la détermination du nombre de noeuds optimal. Le choix du nombre de noeuds reste donc une partie importante du processus de modélisation et peut dépendre du domaine d'application. Une fois choisi, ce n'est plus le nombre de noeuds, mais bien le paramètre de lissage λ qui fera varier l'allure générale de la courbe ajustée.

Nous pouvons désormais énoncer le critère à minimiser pour ajuster des données à l'aide d'une régression pénalisée par spline cubique. En se servant des développements précédents et de (2.15), on obtient

$$\sum_{i=1}^{n} \{y_i - \mathbf{X}_i\boldsymbol{\beta}\}^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}.$$

Le critère d'ajustement peut être réécrit exclusivement en forme matricielle de la façon suivante :

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta}. \tag{2.17}$$

Nous sommes maintenant intéressés à trouver l'estimateur pour $\boldsymbol{\beta}$ qui résulte de la minimisation du critère (2.17). Pour ce faire, il faut dériver ce dernier par rapport à $\boldsymbol{\beta}$ et égaler les équations obtenues à 0.

Après quelques manipulations algébriques, on a l'équivalence suivante :

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} = \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta}.$$

Ensuite, pour la dérivation, on a

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}\left[\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta}\right] &= -2\mathbf{X}^\top\mathbf{y} + 2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta} \\ &= 2\left[\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta} - \mathbf{X}^\top\mathbf{y}\right].\end{aligned}$$

En posant le dernier résultat égal à $\mathbf{0}$, on obtient notre estimateur pour $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{2.18}$$

Dans cette section, nous avons considéré la valeur de λ connue. Son estimation sera couverte à la section 2.3.3. Il est aussi très important de mentionner que les développements théoriques qui ont mené à l'estimateur (2.18) ne sont valides que dans un contexte de **régression gaussienne non additive**.

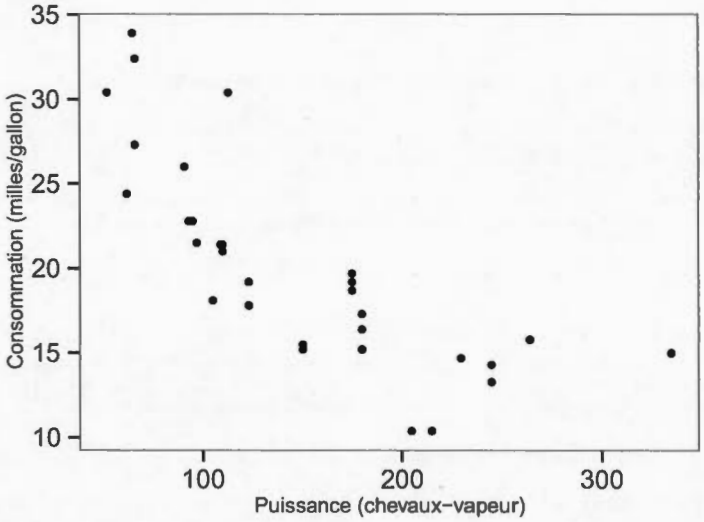
En effet, la normalité est supposée. Nous verrons plus tard comment procéder lorsque nous supposons que les données suivent d'autres lois appartenant à la famille exponentielle linéaire (se référer à la section 1.3.2 pour plus de détails sur cette famille). De plus, le prédicteur linéaire n'était jusqu'ici composé que d'un seul terme non-paramétrique. Nous verrons que lorsqu'un autre terme (paramétrique ou non-paramétrique) est additionné dans le prédicteur linéaire, des problèmes d'identification de modèle surviennent.

Exemple : régression pénalisée par spline cubique

Pour démontrer explicitement de quelle façon une régression par spline cubique fonctionne, nous allons utiliser la base de données *mtcars*, disponible gratuitement avec le logiciel de programmation statistique R. Cette base de données recense des informations sur 32 modèles de voitures datant du début des années 1970. Pour notre exemple, nous allons uniquement utiliser les données concernant la consommation d'essence ainsi que la puissance des voitures.

On s'intéresse donc à ajuster les données sur la consommation d'essence (y) uniquement en fonction de la puissance (x) à l'aide d'une régression pénalisée par spline cubique. Concrètement, on souhaite obtenir une courbe lisse qui va bien représenter la distribution illustrée par la figure 2.3.

Figure 2.3: Distribution de la consommation de 32 modèles de voitures



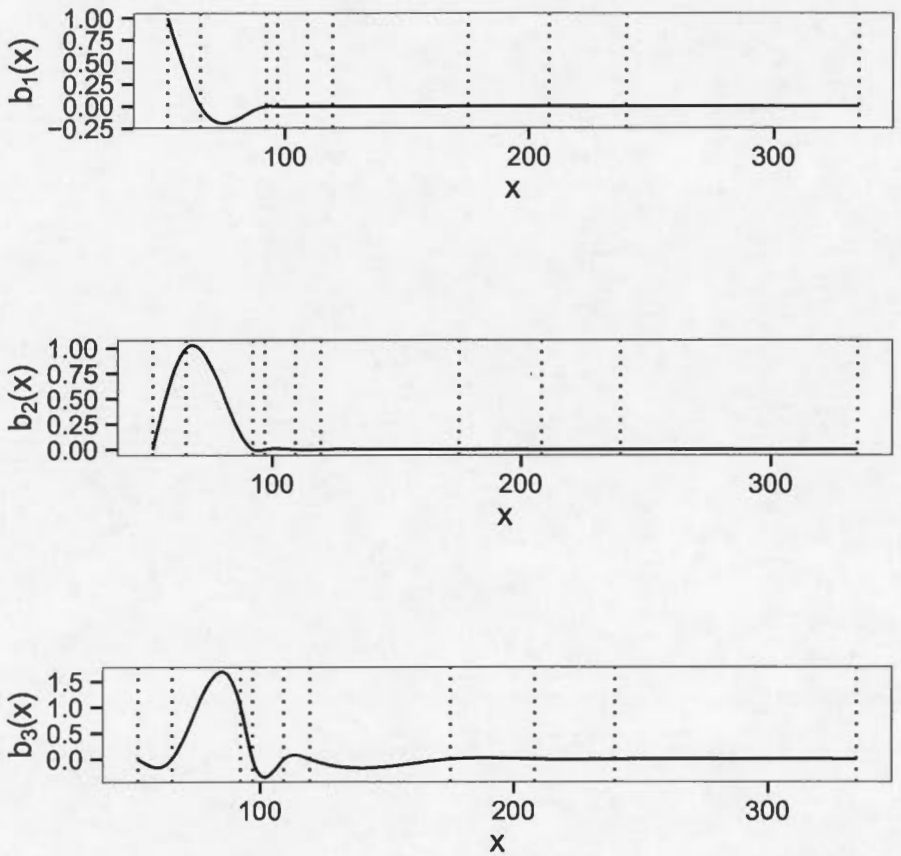
Pour $i = 1, ..., 32$, on considère le modèle suivant :

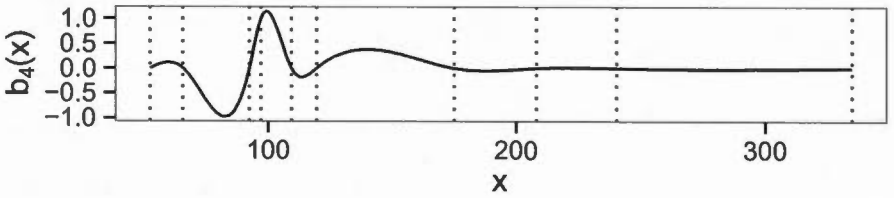
$$y_i = f(x_i) + \epsilon_i, \tag{2.19}$$

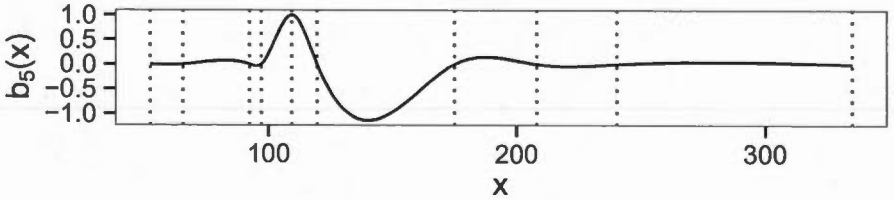
où f est la spline cubique à estimer et $\epsilon_i \sim N(0, \sigma^2)$. Par conséquent, on admet la normalité et l'estimateur (2.18) pourra être utilisé.

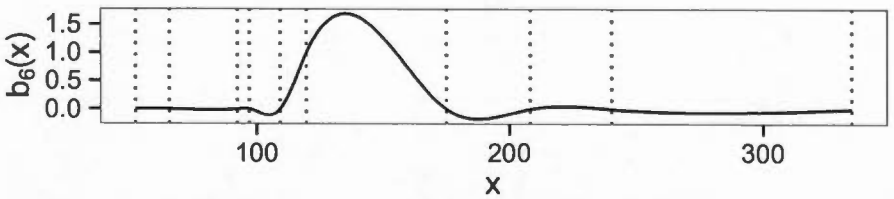
La première des choses à faire est de séparer la distribution des chevaux-vapeur en sections à l'aide de noeuds. Nous allons en choisir dix, dont le premier et le dernier correspondent aux valeurs extrêmes observées pour les chevaux-vapeur. Les huit autres noeuds seront les huit centiles qui permettent de séparer les données en onze parties égales. Par la suite, en utilisant la définition (2.14), on est en mesure de construire les fonctions de base $b_k(x)$, pour $k = 1, ..., 10$. Celles-ci sont illustrées graphiquement à la figure 2.4. Les lignes pointillées correspondent aux noeuds.

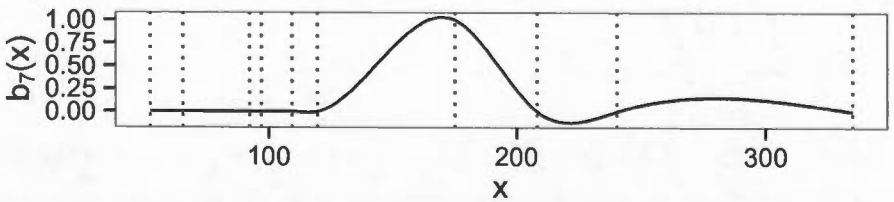
Figure 2.4: Exemple sur la consommation d'essence - Fonctions de base

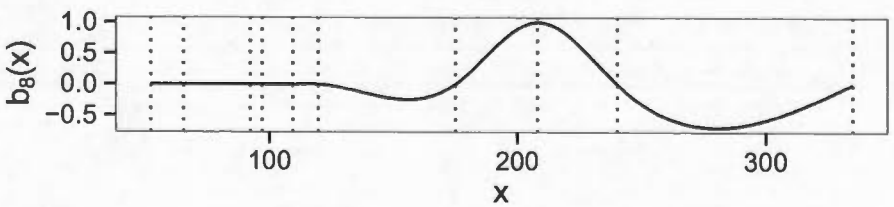


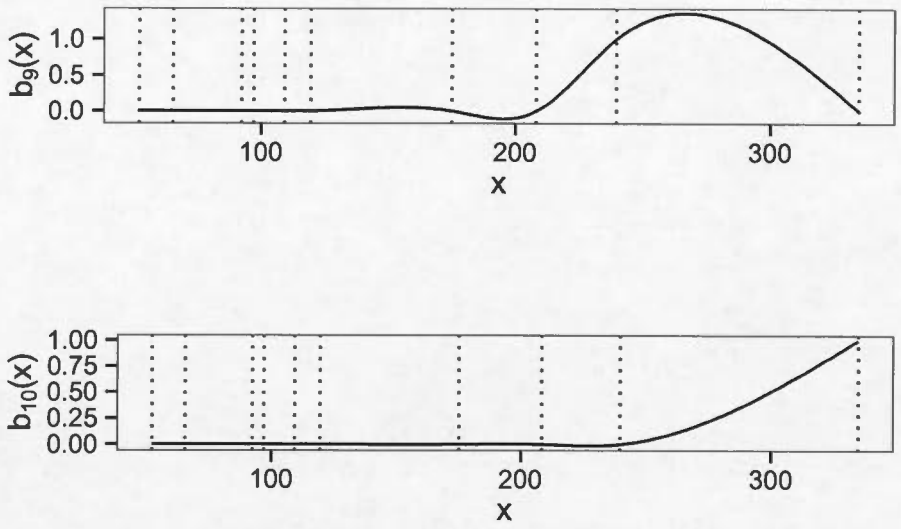

b4(x)
1.0
0.5
0.0
−0.5
−1.0
100
200
300
x


b5(x)
1.0
0.5
0.0
−0.5
−1.0
100
200
300
x


b6(x)
1.5
1.0
0.5
0.0
100
200
300
x


b7(x)
1.00
0.75
0.50
0.25
0.00
100
200
300
x

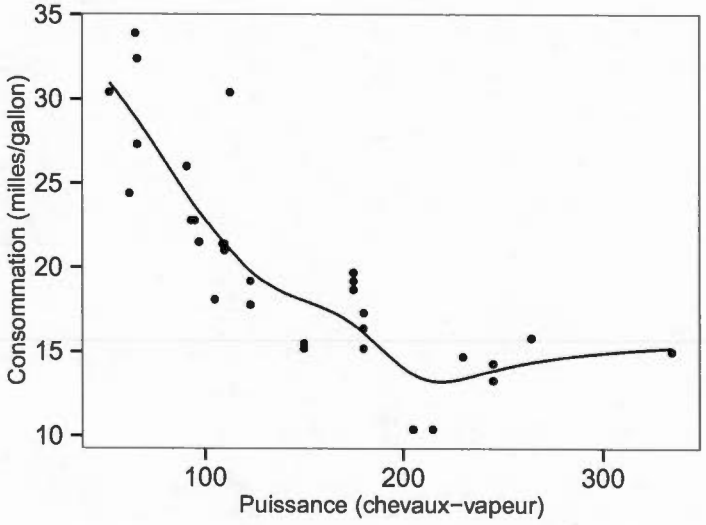

b8(x)
1.0
0.5
0.0
−0.5
100
200
300
x



En choisissant une valeur pour λ et en utilisant le résultat (2.18), on est en mesure de calculer les valeurs estimées pour les paramètres β_k. Chacun de ceux-ci va multiplier sa fonction de base $b_k(x)$ associée. Finalement, toutes les fonctions résultantes vont être additionnées ensemble pour ne donner qu'une fonction finale, c'est-à-dire la courbe ajustée que l'on recherche. La figure 2.5 illustre le résultat obtenu avec un $\lambda = 10\,000$.

Les données utilisées pour l'exemple, de même que les matrices $\mathbf{B}$, $\mathbf{D}$, $\mathbf{F}$ et $\mathbf{X}$, toutes nécessaires pour l'estimation de $\hat{\boldsymbol{\beta}}$, sont présentées à l'annexe A.2. Le vecteur de paramètres estimés $\hat{\boldsymbol{\beta}}$ est également donné.

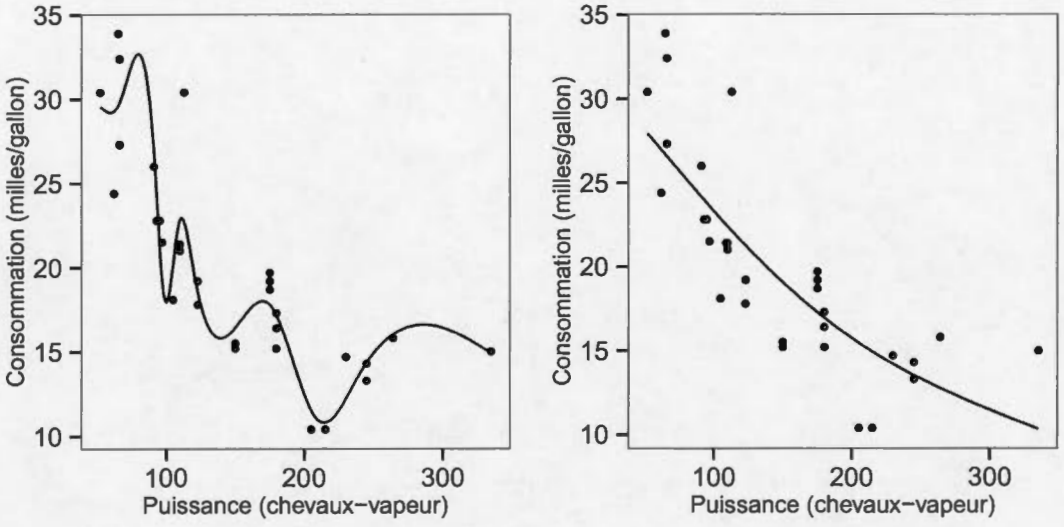
Figure 2.5: Exemple sur la consommation d'essence - Courbe ajustée



2.3.3 Estimation du paramètre de lissage

Le paramètre de lissage est important, car c'est lui qui est responsable de l'importance donnée à chacun des objectifs visés par la minimisation du critère d'ajustement défini par (2.17). En fait, si $\lambda = 0$, la pénalité disparaît et nous aurons une courbe ajustée très inégale qui se rapprochera le plus possible des données. À l'inverse, si $\lambda \to \infty$, toute l'importance est portée sur la minimisation des oscillations pour l'ajustement, si bien que nous obtiendrons une ligne droite. La figure 2.6 reprend l'exemple sur la consommation d'essence de la section précédente pour illustrer ce phénomène.

Figure 2.6: Effet du paramètre de lissage sur l'ajustement



(a) $\lambda = 0.5$ (b) $\lambda = 2\,000\,000$

L'estimation du paramètre de lissage est donc une étape considérable, car nous cherchons le λ qui composera de manière optimale avec les deux objectifs contradictoires. Pour arriver à estimer efficacement le paramètre de lissage, nous allons faire appel à la validation croisée (*cross validation*). Le principe est simple : choisir λ en fonction de la qualité des prédictions sur des données qui n'ont pas servi à la modélisation.

On définit le score de validation croisée ordinaire (OCV – *ordinary cross validation score*) comme étant

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2, \tag{2.20}$$

où $\hat{f}_i^{[-i]}$ est la i-ème valeur prédite générée par le modèle préalablement ajusté sans l'observation y_i. En remplaçant y_i par l'expression donnée par (2.19) dans

(2.20), on obtient alors

$$\begin{aligned}\nu_o &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - (\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2,\end{aligned}$$

où $f_i \equiv f(x_i)$. Puisque l'on sait que $\mathbb{E}[\epsilon_i] = 0$ et que ϵ_i et $\hat{f}_i^{[-i]}$ sont indépendants, on a que

$$\mathbb{E}[\nu_o] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\right] + \sigma^2. \tag{2.21}$$

De plus, on a que $\hat{f}^{[-i]} \approx \hat{f}$, avec une égalité possible dans la limite où le nombre de données est suffisamment grand. Par conséquent, on a que

$$\mathbb{E}[\nu_o] \approx \mathbb{E}[M] + \sigma^2, \tag{2.22}$$

où $M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$. L'équation (2.22) représente l'erreur attendue pour la prédiction (EPE – *expected prediction error*). Dans le meilleur des mondes, on voudrait choisir λ de manière à minimiser M. Comme f, la vraie fonction que l'on souhaite estimer, n'est pas connue, on ne peut utiliser directement M. Or, une solution est de choisir le λ qui minimise $\mathbb{E}[\nu_o]$ (EPE). Par les développements précédents, on voit que minimiser ν_o, le score de validation croisée ordinaire, remplit cet objectif.

Il est cependant plutôt laborieux de devoir ajuster n modèles tour à tour, chacun laissant à l'écart une observation, afin de pouvoir estimer le paramètre de lissage. Heureusement, il peut être démontré que

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1 - A_{ii})^2, \tag{2.23}$$

où $\mathbf{A}$ est la matrice chapeau résultante du modèle produisant $\hat{f}$, soit celui ajusté avec toutes les observations. Explicitement, avec la définition (2.18), on trouve

aisément que

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top. \tag{2.24}$$

En pratique, les poids $1-A_{ii}$ sont fréquemment remplacés par le poids moyen $tr(\mathbf{I}-\mathbf{A})/n$. Ceci mène au score de validation croisée généralisé (GCV – *generalized cross validation score*) :

$$\nu_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{\left[tr(\mathbf{I}-\mathbf{A})\right]^2}. \tag{2.25}$$

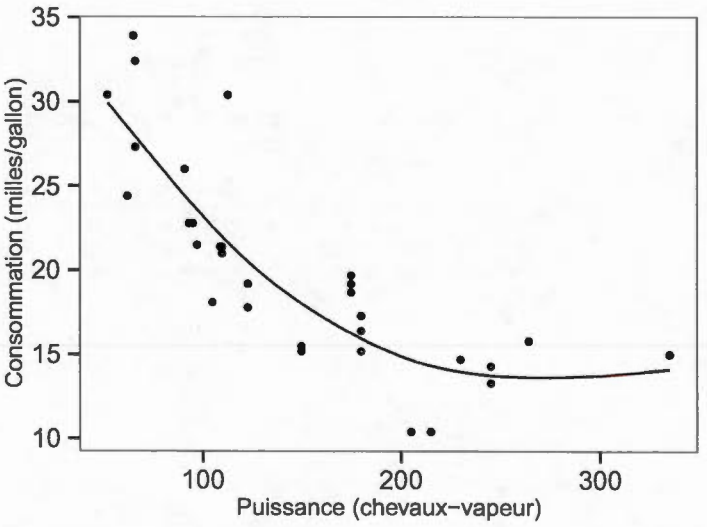
Le score GCV est souvent plus employé que le score OCV en raison de sa propriété d'invariance. En utilisant le poids moyen $tr(\mathbf{I}-\mathbf{A})/n$ au lieu des poids individuels dans la défintion (2.25), on donne à chaque observation la même importance dans la détermination de la valeur ν_g. Par contre, si l'on utilise la validation croisée ordinaire, le score ν_o défini par (2.23) est alors impacté de façon importante par les observations les plus influentes, c'est-à-dire celles dont les valeurs A_{ii} associées sont les plus grandes.

Exemple : Détermination du paramètre de lissage optimal

Reprenons une nouvelle fois notre exemple sur la consommation d'essence de la page 31. On s'intéresse maintenant à trouver le λ optimal. Une façon simple de procéder, mais numériquement pas très efficace, est de déterminer une grille de valeurs possibles pour λ. Ensuite, pour chaque valeur de la grille, on ajuste les données avec une régression pénalisée par spline cubique et l'on calcule le score GCV associé. Finalement, parmi toutes ces modélisations, on va choisir celle dont la valeur pour le score GCV s'avère la plus faible.

En programmant une routine numérique qui effectue cet exercice, on trouve que $\lambda \approx 223\,000$ est la valeur optimale. La figure 2.7 illustre le résultat de la modélisation.

Figure 2.7: Exemple sur la consommation d'essence - Courbe ajustée optimale



2.3.4 Ajustement d'un modèle additif généralisé

Avec les notions abordées jusqu'à maintenant dans ce chapitre, nous allons être en mesure de définir plus précisément ce qu'est un modèle additif généralisé. Nous verrons également comment estimer un tel modèle.

L'équation (2.1) a introduit une structure possible pour un GAM. Celle-ci était additive et constituée de plusieurs termes non-paramétriques. Bien que la régression pénalisée par spline cubique discutée à la section 2.3.2 ne faisait intervenir qu'un seul terme non-paramétrique, les développements théoriques présentés sur ce type de régression seront très utiles pour la construction d'un modèle additif généralisé.

L'attrait principal des modèles GAM est la flexibilité que procurent les termes non-paramétriques sur l'ajustement de données. Jusqu'à maintenant, nous n'avons parlé que de fonctions de lissage univariées, mais il est possible d'ajouter des fonc-

tions multivariées dans le prédicteur linéaire du GAM. À ce sujet, nous verrons à la section 2.3.5 comment les fonctions de lissage bivariées sont construites et comment elles permettent d'intégrer une dépendance entre deux variables explicatives.

Notons qu'il est aussi possible d'inclure des termes paramétriques dans le prédicteur linéaire. Les modèles GAM sont donc considérés comme une approche de modélisation semi-paramétrique. Dans un souci de garder les choses simples, l'équation (2.1) ne comptait aucun terme paramétrique et aucun terme non-paramétrique qui dépendait de plus d'une variable.

De façon plus générale, un GAM modélisant une variable aléatoire Y_i peut avoir la structure suivante :

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}),$$

où $\mu_i \equiv \mathbb{E}[Y_i]$ et Y_i suit une distribution quelconque appartenant à la famille exponentielle linéaire. Celle-ci et le rôle de la fonction de lien g ont été discutés à la section 1.3 qui introduisait les modèles linéaires généralisés.

Quant à $\mathbf{X}_i^*$, il s'agit de la i-ème ligne de la matrice de design créée uniquement à partir des termes paramétriques du modèle, et $\boldsymbol{\theta}$ est le vecteur de paramètres associé. $f_1(x_{1i})$ et $f_2(x_{2i})$ sont des fonctions de lissage univariées et $f_3(x_{3i}, x_{4i})$ en est une bivariée.

Pour le moment, excluons $f_3(x_{3i}, x_{4i})$ et construisons un modèle GAM dont la structure est la suivante :

$$g(\mu_i) = \mathbf{X}_i^* \boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}). \tag{2.26}$$

Chacune des fonctions de lissage univariées, soient $f_1(x_1)$ et $f_2(x_2)$, seront des splines cubiques d'ajustement construites en utilisant exactement la même mé-

canique suivie aux sections 2.3.1 et 2.3.2. En utilisant l'équation (2.14), chaque fonction peut être exprimée sous la forme suivante :

$$f_j(x_j) = \sum_{k=1}^{q_j} b_{jk}(x_j)\beta_{jk}, \tag{2.27}$$

où β_{jk} sont les paramètres à estimer. Les fonctions de base $b_{jk}(x_j)$ sont équivalentes aux fonctions $b_k(x)$ définies par la formule (2.14). L'indice j dans $b_{jk}(x_j)$ ne sert ici qu'à identifier la variable explicative. Donc, avec (2.14), on est en mesure de créer une matrice de design $\tilde{\mathbf{X}}_j$ à partir du vecteur de valeurs observées pour la variable x_j. En posant $\tilde{\boldsymbol{\beta}}_j = \left[\beta_{j1}, \beta_{j2}, ..., \beta_{jq_j}\right]^\top$, on a

$$\mathbf{f}_j = \tilde{\mathbf{X}}_j \tilde{\boldsymbol{\beta}}_j, \tag{2.28}$$

où $\mathbf{f}_j$ est un vecteur dont les éléments correspondent aux valeurs $f_j(x_{ji})$. Pour $i = 1, ..., n$, x_{ji} est la i-ème valeur observée pour la variable explicative x_j.

Jusqu'ici, rien n'est nouveau par rapport à ce que l'on a déjà vu dans la section sur la régression pénalisée par spline cubique (section 2.3.2). Tout ce que nous avons fait, c'est associer une fonction de lissage univariée à une spline cubique et exprimer celle-ci sous la forme de l'équation (2.27). Le processus a simplement été répété deux fois, car (2.26) contient deux fonctions de lissage univariées. Chacune est identifiée par l'indice j. Le tilde est ajouté à la notation, mais l'équivalence est totale avec la notation utilisée dans les sections précédentes.

Par contre, nous faisons face à un nouveau problème. Le modèle représenté par (2.26) n'est pas identifiable. Classiquement, on parle de « problème d'identification » de modèle. La cause vient du fait d'avoir des composantes non-paramétriques additives non restreintes. Par exemple, en admettant deux termes non-paramétriques, disons $f_1(x_1)$ et $f_2(x_2)$, le même résultat pour la somme de ces deux termes peut être obtenu par différents vecteurs $\tilde{\boldsymbol{\beta}}_1$ et $\tilde{\boldsymbol{\beta}}_2$. Le même problème survient lorsqu'un

« terme d'interception » [5] est inclus dans le prédicteur linéaire. Ceci fait en sorte que lors du processus d'estimation du modèle, nous aurons un problème puisque l'on ne sera pas en mesure d'identifier les vecteurs de paramètres $\tilde{\boldsymbol{\beta}}_1$ et $\tilde{\boldsymbol{\beta}}_2$.

Alors, pour rendre le modèle (2.26) identifiable, nous allons devoir soumettre chaque terme non-paramétrique à une contrainte de centralité. Concrètement, on va imposer que la somme des éléments de chaque vecteur $\mathbf{f}_j$ soit nulle, c'est-à-dire que l'équation suivante soit vérifiée pour tout j :

$$\mathbf{1}^\top \tilde{\mathbf{X}}_j \tilde{\boldsymbol{\beta}}_j = 0. \tag{2.29}$$

Pour s'assurer que le modèle respecte cette contrainte, une manière de procéder est d'utiliser la décomposition QR. Celle-ci va nous permettre de reparamétriser chaque terme $\mathbf{f}_j$ (équation 2.28) en un nouveau vecteur de paramètres $\boldsymbol{\beta}_j$ associé à une nouvelle matrice $\mathbf{X}_j$. À partir d'ici, lorsque le tilde est présent dans la notation, il fait référence aux anciennes matrices et vecteurs de paramètres. Lorsqu'il est absent, on parle des nouvelles matrices et vecteurs créés à la suite de la reparamétrisation. Il est également à noter que la reparamétrisation par la décomposition QR fera perdre un paramètre à chacun des termes. Ainsi, chaque nouveau vecteur $\boldsymbol{\beta}_j$ comprendra $q_j - 1$ paramètres.

Posons d'abord $\mathbf{C} = \mathbf{1}^\top \tilde{\mathbf{X}}_j$. En algèbre linéaire, la décomposition QR de $\mathbf{C}^\top$ est définie comme

$$\mathbf{C}^\top = \mathbf{QR},$$

où $\mathbf{Q}$ est une matrice orthogonale de dimensions $q_j \times q_j$ et $\mathbf{R}$ une matrice triangulaire supérieure de dimensions $q_j \times 1$. $\mathbf{Q}$ peut être partitionnée de telle sorte que $\mathbf{Q} \equiv (\mathbf{D} : \mathbf{Z})$, où $\mathbf{Z}$ est également une matrice orthogonale, mais de dimensions $q_j \times (q_j - 1)$.

---

5. Traduction libre de l'expression *intercept*.

En posant que

$$\tilde{\boldsymbol{\beta}}_j = \mathbf{Z}\boldsymbol{\beta}_j,$$

la contrainte de centralité (2.29) sera toujours respectée, et ce pour n'importe quel vecteur de nouveaux paramètres $\boldsymbol{\beta}_j$. En effet, on peut le voir avec le développement qui suit :

$$\mathbf{1}^\top \tilde{\mathbf{X}}_j \tilde{\boldsymbol{\beta}}_j = \mathbf{C}\tilde{\boldsymbol{\beta}}_j = \mathbf{R}^\top \mathbf{Q}^\top \mathbf{Z}\boldsymbol{\beta}_j = \mathbf{R}^\top \begin{bmatrix} \mathbf{D}^\top \\ \mathbf{Z}^\top \end{bmatrix} \mathbf{Z}\boldsymbol{\beta}_j = \mathbf{R}^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \boldsymbol{\beta}_j = 0.$$

La reparamétrisation $\tilde{\boldsymbol{\beta}}_j = \mathbf{Z}\boldsymbol{\beta}_j$ fait en sorte que l'on obtient également une nouvelles matrice de design. Celle-ci correspond à l'originale multipliée par la matrice orthogonale $\mathbf{Z}$, soit $\mathbf{X}_j = \tilde{\mathbf{X}}_j \mathbf{Z}$.

À la suite de cette reparamétrisation, nécessaire pour des raisons d'identification, la formule du GAM (2.26) peut maintenant être réécrite en utilisant les nouvelles matrices et vecteurs de paramètres :

$$g(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}, \tag{2.30}$$

où $\mathbf{X} = [\mathbf{X}^* : \mathbf{X_1} : \mathbf{X_2}]$ et $\boldsymbol{\beta}^\top = \left[\boldsymbol{\theta}^\top : \boldsymbol{\beta}_1^\top : \boldsymbol{\beta}_2^\top\right]$.

On voit clairement que (2.30) n'est qu'un GLM. Sachant cela, la fonction de log-vraisemblance associée, disons $l(\boldsymbol{\beta})$, peut être écrite. Bien qu'il est possible d'estimer les paramètres $\boldsymbol{\beta}$ par une maximisation de la fonction de vraisemblance, il existe une forte chance d'observer un surajustement (*overfitting*). En procédant ainsi, les fonctions f_j estimées risquent d'être très oscillantes. Évidemment, dans une perspective d'inférence, le surajustement n'est jamais souhaitable. Par conséquent, les GAM sont généralement estimés par une « maximisation de la fonction de vraisemblance pénalisée » (*penalized likelihood maximization*).

Le concept de pénalité est exactement le même que celui présenté à la section 2.2.3, c'est-à-dire que la pénalité permet, dépendamment de la valeur du paramètre de lissage, de conjuguer avec les deux objectifs d'ajustement. D'un côté, on veut que chaque fonction f_j représente le plus possible les données, mais d'un autre, on veut effectuer un lissage pour capter la tendance générale des observations.

Comme chaque terme non-paramétrique est une spline cubique à estimer, la pénalité utilisée pour quantifier l'irrégularité de chaque fonction f_j sera définie par l'équation (2.15), soit $\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{S}}_j \tilde{\boldsymbol{\beta}}_j$. Pour des raisons de cohérence, la notation a été modifiée, mais l'équivalence est respectée par rapport à l'équation (2.15).

La reparamétrisation $\tilde{\boldsymbol{\beta}}_j = \mathbf{Z}\boldsymbol{\beta}_j$, nécessaire afin de respecter la contrainte de centralité (2.29), va également avoir un impact sur la forme de la pénalité. En effet, on déduit aisément que

$$\tilde{\boldsymbol{\beta}}_j^\top \tilde{\mathbf{S}}_j \tilde{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j^\top \bar{\mathbf{S}}_j \boldsymbol{\beta}_j, \tag{2.31}$$

où $\bar{\mathbf{S}}_j = \mathbf{Z}^\top \tilde{\mathbf{S}}_j \mathbf{Z}$. Il est pratique de réécrire la pénalité en fonction du vecteur global de paramètres $\boldsymbol{\beta}$, de sorte que (2.31) devient $\boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}$. Quant à $\mathbf{S}_j$, il s'agit en fait de la matrice $\bar{\mathbf{S}}_j$ à laquelle on a ajouté des valeurs nulles de façon à ce que $\boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta} \equiv \boldsymbol{\beta}_j^\top \bar{\mathbf{S}}_j \boldsymbol{\beta}_j$.

On peut maintenant définir $l_p(\boldsymbol{\beta})$, la fonction de log-vraisemblance pénalisée :

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2} \sum_j \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}, \tag{2.32}$$

où $l(\boldsymbol{\beta})$ est définie par (1.7) et les λ_j sont les paramètres de lissage. Ici, nous les considérons connus. Par contre, ils sont habituellement estimés en utilisant le score de validation croisée généralisé (GCV). La description de ce score et la démonstration de son utilité sont discutés à la section 2.3.3.

Pour estimer le vecteur de coefficients $\boldsymbol{\beta}$ d'un GAM, la fonction de log-vraisemblance

pénalisée $l_p(\boldsymbol{\beta})$ doit être maximisée. Pour ce faire, il faut dériver l'équation (2.32) par rapport à chaque élément β_k constituant $\boldsymbol{\beta}$ et égaler le résultat à 0.

D'abord, (2.32) peut être réécrite sous la forme suivante :

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta},$$

où $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. On dérive ensuite par rapport à β_k :

$$\frac{\partial l_p}{\partial \beta_k} = \frac{\partial l}{\partial \beta_k} - [\mathbf{S}\boldsymbol{\beta}]_k = \frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_k} - [\mathbf{S}\boldsymbol{\beta}]_k = 0, \tag{2.33}$$

où $[\mathbf{S}\boldsymbol{\beta}]_k$ représente la k-ème entrée du vecteur formé par $\mathbf{S}\boldsymbol{\beta}$. L'équation (2.33) est donc la condition de premier ordre à résoudre pour l'estimation de $\boldsymbol{\beta}$ par maximum de vraisemblance pénalisée.

Wood (2006) propose dans sa section 4.3 une méthode itérative qui permet numériquement de résoudre (2.33). Celle-ci est connue sous le nom de *Penalized Iteratively Re-weighted Least Squares* (P-IRLS) et est une version modifiée de la méthode *Iteratively Re-weighted Least Squares* (IRLS) proposée par Nelder et Wedderburn (1972). La librairie du logiciel de programmation statistique R *Mixed GAM Computation Vehicle* (mgcv) implémente cette technique d'estimation et permet d'ajuster facilement des modèles additifs généralisés.

2.3.5 Base de lissage par produit tensoriel

Jusqu'à maintenant, nous avons vu de quelle façon construire et ajuster un modèle additif généralisé avec des fonctions de lissage univariées. Par contre, il arrive souvent dans une analyse statistique que l'on souhaite tester des interactions potentielles entre deux ou plusieurs variables. Ainsi, dans cette section, nous verrons comment inclure une fonction de lissage bivariée dans la construction d'un GAM.

L'emploi d'une « base de lissage par produit tensoriel » [6] apporte une très grande flexibilité au GAM, puisqu'elle permet d'expliquer une variable réponse par des fonctions impliquant plusieurs variables explicatives.

Bien que la base par produit tensoriel sera ici expliquée dans le cas de la construction d'une fonction de lissage bivariée, il n'y a aucune limite quant au nombre possible de variables. Par exemple, si l'on croit qu'un certain phénomène peut être expliqué par une interaction de cinq facteurs à la fois, la base par produit tensoriel permet de procéder à l'élaboration d'une fonction qui traitera cette information.

L'idée générale est que l'on va utiliser des bases de lissage marginales et les combiner de telle sorte que l'on va construire une fonction multivariée. Supposons que l'on a deux variables, soient x et z. Supposons également que nous avons les fonctions de lissage univariées f_x et f_z, qui sont représentées à l'aide d'une base par spline cubique de la façon suivante :

$$f_x(x) = \sum_{k=1}^{K} a_k(x)\alpha_k \text{ et } f_z(z) = \sum_{l=1}^{L} d_l(z)\delta_l, \tag{2.34}$$

où α_k et δ_l sont des paramètres, $a_k(x)$ et $d_l(z)$ sont des fonctions de base connues (voir l'équation 2.14 pour la définition explicite des fonctions de base pour une spline cubique).

On souhaite donc créer une fonction f_{xz}. Une manière de procéder consiste à prendre la fonction f_x et de la convertir en une fonction lisse impliquant x et z. Précisément, il suffit de permettre aux paramètres α_k, dans la définition de f_x, de varier avec z. En utilisant la même base par spline cubique avec laquelle la

6. Traduction libre de l'expression *tensor product smoothing base*.

fonction f_z est définie, on peut définir une nouvelle fonction $\alpha_k(z)$ comme suit :

$$\alpha_k(z) = \sum_{l=1}^{L} d_l(z)\beta_{kl}.$$

En insérant $\alpha_k(z)$ dans f_x, on obtient ainsi notre nouvelle fonction f_{xz} :

$$f_{xz}(x,z) = \sum_{k=1}^{K}\sum_{l=1}^{L} a_k(x)d_l(z)\beta_{kl}. \tag{2.35}$$

En ayant établi (2.35), il est maintenant possible de créer une matrice de design $\mathbf{X}$ à partir des vecteurs de valeurs observées pour les variables x et z. En considérant un vecteur de paramètres $\boldsymbol{\beta}$, où les entrées correspondent aux β_{kl}, (2.35) peut être exprimée sous la forme

$$\mathbf{f} = \mathbf{X}\boldsymbol{\beta}.$$

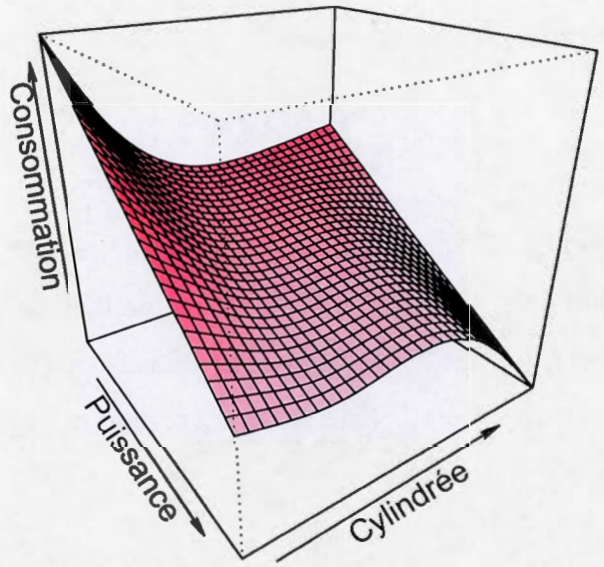
Il peut paraître difficile d'établir la matrice $\mathbf{X}$ en ne visualisant que l'équation (2.35), mais une relation simple et très utile en découle directement. En fait, il faut d'abord trouver $\mathbf{X}_x$ et $\mathbf{X}_z$, qui sont les matrices de design qui proviennent du lissage par spline cubique des variables individuelles x et z. Autrement dit, elles sont créées à partir des fonctions de base $a_k(x)$ et $d_l(z)$ que l'on retrouve dans les représentations de f_x et f_z (équation 2.34).

Ensuite, la i-ème ligne de $\mathbf{X}$ se trouve tout simplement de cette façon :

$$\mathbf{X}_i = \mathbf{X}_{xi} \otimes \mathbf{X}_{zi},$$

où $\otimes$ est le produit de Kronecker.

Figure 2.8: Fonction de lissage bivariée construite par produit tensoriel



La figure 2.8 illustre un exemple d'une fonction de lissage bivariée construite à l'aide d'une base par produit tensoriel. La figure reprend les données de l'exemple sur la consommation d'essence (se référer à la page 31). Ici, on ajoute cependant une variable, soit la cylindrée des moteurs. En utilisant d'abord deux bases par spline cubique individuelles de dimensions 5, le produit tensoriel permet ensuite de combiner celles-ci et de construire une fonction de lissage f(puissance, cylindrée) avec laquelle on souhaite expliquer la consommation d'essence des voitures. La figure 2.8 présente la courbe ajustée finale.

Par contre, avant d'arriver à estimer une telle fonction, ll nous reste à introduire la forme que prennent les pénalités lorsque l'on utilise une base par produit tensoriel pour représenter des fonctions lisses de plusieurs variables. Encore une fois, nous nous limiterons à deux variables, mais les résultats se généralisent facilement pour

un nombre de variables plus élevé.

Rappelons d'abord que pour développer f_{xz}, définie par l'équation (2.35), nous sommes partis des définitions suivantes pour f_x et f_z :

$$f_x(x) = \sum_{k=1}^{K} a_k(x)\alpha_k \text{ et } f_z(z) = \sum_{l=1}^{L} d_l(z)\delta_l.$$

Dans la section 2.3.2, nous avons vu sous quelle forme se présente la pénalité pour une spline cubique. L'intégrale du carré de la dérivée seconde, mesure jugée adéquate pour quantifier l'irrégularité d'une fonction, a pour résultat (2.15) dans le cas d'une spline cubique. Considérant cela, on peut définir les pénalités respectives pour f_x et f_z comme suit :

$$J_x(f_x) = \int \left(\frac{\partial f_x}{\partial x^2}\right)^2 dx = \boldsymbol{\alpha}^\top \mathbf{S}_x \boldsymbol{\alpha} \text{ et } J_z(f_z) = \int \left(\frac{\partial f_z}{\partial z^2}\right)^2 dz = \boldsymbol{\delta}^\top \mathbf{S}_z \boldsymbol{\delta},$$

où $\mathbf{S}_x$ et $\mathbf{S}_z$ sont des matrices de pénalité dont les éléments sont connus.

Considérons maintenant $f_{x|z}(x)$, qui est $f_{xz}(x, z)$ où z a cependant été fixé. Ceci fait en sorte qu'il s'agit désormais d'une fonction de x seulement. De façon similaire, considérons également $f_{z|x}(z)$.

Dès lors, en utilisant ces fonctions univariées, on peut définir une expression pour la pénalité dans le cas d'une fonction de lissage bivariée :

$$J(f_{xz}) = \lambda_x \int_z J_x(f_{x|z})dz + \lambda_z \int_x J_z(f_{z|x})dx, \tag{2.36}$$

où les paramètres λ_x et λ_z contrôlent l'importance donnée au lissage dans chaque direction. Ils permettent également à la valeur pour la pénalité de ne pas être influée par une variable plus qu'une autre. Les paramètres de lissage vont ainsi rendre la pénalité insensible aux différents ordres de grandeur que l'on pourrait observer chez les variables explicatives.

Au premier coup d'oeil, (2.36) ne paraît pas évident, mais il est plus facile d'interpréter l'équation avec une illustration. Prenons par exemple la fonction de lissage ajustée de la figure 2.8. Associons celle-ci à f_{xz}, la variable x à la puissance, et z à la cylindrée.

On fixe une valeur pour la cylindrée (z), ce qui fait que l'on obtient une fonction de la puissance (x). Celle-ci est $f_{x|z}(x)$. On calcule alors l'intégrale du carré de la dérivée seconde de cette fonction pour quantifier son irrégularité. On répète alors le processus pour toutes les valeurs possibles de z. Comme on s'intéresse à l'irrégularité de f_{xz} dans la direction de x, on somme toutes les valeurs obtenues. Cette somme correspond tout simplement à

$$\int_z J_x(f_{x|z})dz.$$

Maintenant, on peut écrire que

$$f_{x|z}(x) = \sum_{k=1}^{K} a_k(x)\alpha_k(z).$$

Rappelons qu'il s'agit simplement de f_{xz}, définie par (2.35), où la valeur de la variable z a été fixée.

La pénalité pour $f_{x|z}(x)$, soit $J_x(f_{x|z})$, est donc

$$J_x(f_{x|z}) = \boldsymbol{\alpha}(z)^\top \mathbf{S}_x \boldsymbol{\alpha}(z).$$

Ultimement, on souhaite estimer le vecteur de paramètres $\boldsymbol{\beta}$, c'est-à-dire les paramètres β_{kl} présents dans f_{xz}. Ainsi, il serait avantageux de pouvoir exprimer $\boldsymbol{\alpha}(z)$ en fonction de $\boldsymbol{\beta}$.

En fait, il est possible de définir une matrice de coefficients connue $\mathbf{M}_z$ de sorte que $\boldsymbol{\alpha}(z) = \mathbf{M}_z\boldsymbol{\beta}$. Concrètement, $\mathbf{M}_z$ ne contient que des valeurs nulles et les

différentes valeurs que la base par spline cubique $d_l(z)$ peut prendre pour une valeur donnée de z, $l = 1, ..., L$. Or, on a que

$$J_x(f_{x|z}) = \boldsymbol{\alpha}(z)^\top \mathbf{S}_x \boldsymbol{\alpha}(z) = \boldsymbol{\beta}^\top \mathbf{M}_z^\top \mathbf{S}_x \mathbf{M}_z \boldsymbol{\beta}.$$

Par conséquent,

$$\int_z J_x(f_{x|z}) dz = \boldsymbol{\beta}^\top \int_z \mathbf{M}_z^\top \mathbf{S}_x \mathbf{M}_z dz \boldsymbol{\beta}.$$

Cette dernière intégrale peut être évaluée numériquement. Des développements similaires peuvent être faits pour l'autre composante de $J(f_{xz})$, soit $\int_x J_z(f_{z|x}) dx$.

Dans ce mémoire, étant donné que les splines cubiques ont été définies comme des fonctions dont les paramètres à estimer sont les valeurs que prennent ces fonctions aux noeuds, une approximation des intégrales peut être faite et permet donc d'éviter de les effectuer explicitement.

Par exemple, il est possible de faire l'approximation suivante :

$$\int_z J_x(f_{x|z}) dz \approx h \sum_l J_x(f_{x|z_l}),$$

où h est une constante liée à l'espace qui existe entre les différents noeuds z_l, $l = 1, ..., L$.

Il peut ensuite être démontré que

$$\sum_l J_x(f_{x|z_l}) = \boldsymbol{\beta}^\top \bar{\mathbf{S}}_x \boldsymbol{\beta} = J_x^*(f_{xz}),$$

où $\bar{\mathbf{S}}_x = \mathbf{S}_x \otimes \mathbf{I}_L$, et $\mathbf{I}_L$ la matrice identité de rang L.

Similairement, on a que

$$\sum_k J_z(f_{z|x_k}) = \boldsymbol{\beta}^\top \bar{\mathbf{S}}_z \boldsymbol{\beta} = J_z^*(f_{xz}),$$

où $\bar{\mathbf{S}}_z = \mathbf{I}_K \otimes \mathbf{S}_z$, et $\mathbf{I}_K$ la matrice identité de rang K.

Finalement, on est en mesure de fournir une approximation raisonnable pour l'équation (2.36) :

$$J(f_{xz}) \approx J^*(f_{xz}) = \lambda_x J_x^*(f_{xz}) + \lambda_z J_z^*(f_{xz}), \tag{2.37}$$

où chaque constante h est absorbée dans le paramètre λ correspondant.

Maintenant que nous pouvons construire une matrice de design et les matrices de pénalité associées, un GAM impliquant une fonction de lissage bivariée construite par produit tensoriel peut être ajusté en suivant les mêmes étapes vues à la section 2.3.4.

CHAPITRE III

APPLICATION À L'ASSURANCE AUTOMOBILE

3.1 Introduction

Dans ce chapitre, les modèles additifs généralisés, dont le cadre théorique a été le sujet du chapitre 2, seront appliqués dans un contexte réel d'assurance automobile.

Introduite à la section 1.2, la segmentation est un processus très important pour une compagnie d'assurance. Une segmentation efficace permet aux actuaires oeuvrant pour une compagnie d'assurance de calculer une prime qui se veut la plus juste possible pour chaque assuré. Par contre, le processus de segmentation est limité par la qualité des données : il est difficile de bien segmenter une variable lorsque l'on ne possède pas de données fiables pour celle-ci. Cette problématique a longtemps affecté la variable caractérisant le niveau d'utilisation d'un véhicule assuré. Heureusement, de nouvelles technologies permettent désormais de quantifier précisément l'utilisation faite d'un véhicule en assurance automobile.

Les modèles additifs généralisés seront donc employés pour analyser exhaustivement l'impact du degré d'utilisation d'un véhicule sur le risque d'accident automobile. Une composante temporelle, soit la durée d'exposition au risque, sera également considérée.

3.1.1 Importance de l'utilisation du véhicule

Il est évident que l'utilisation faite du véhicule est un facteur important dans le risque d'accident. En effet, on s'attend à ce qu'une personne qui conduit beaucoup ait une forte propension à avoir un accident automobile par rapport à une personne qui utilise sa voiture occasionnellement. Or, une façon de quantifier cette utilisation du véhicule est tout simplement de considérer le nombre de kilomètres parcourus annuellement. Plusieurs études ont, par le passé, démontré qu'il existe un lien important entre le nombre de kilomètres parcourus et le risque d'être impliqué dans un accident automobile.

Par exemple, Vickrey (1968) critique durement la tarification forfaitaire en assurance automobile, car la prime calculée dans un tel système ignore une variable intimement liée au risque d'accident, soit l'ampleur de l'utilisation de la voiture. Ceci a pour conséquence de créer des « injustices » au niveau des primes d'assurance chargées aux assurés. En fait, un tel système de tarification est solidaire, c'est-à-dire que certains assurés paient une prime plus élevée pour compenser la prime trop peu élevée de certains autres assurés.

Les analyses de Lourens *et al.* (1999) indiquent que peu importe l'âge d'un assuré, une augmentation du kilométrage parcouru annuellement provoque une augmentation du nombre d'occasions où il sera impliqué dans un sinistre automobile. Litman (2005) parvient à un constat similaire et expose une relation positive et non linéaire entre le nombre d'accidents encourus annuellement et le kilométrage annuel parcouru. De surcroît, peu importe la responsabilité ou non des assurés dans ces accidents, leur fréquence de sinistres tend à augmenter avec le kilométrage annuel.

Bordoff et Noel (2008) corroborent la relation non proportionnelle avancée par Litman (2005) entre le kilométrage et le nombre d'accidents par année. Ils sou-

tiennent que ceci est observable lorsque la relation est étudiée sur un ensemble d'assurés à la fois (données agrégées). Lorsqu'un individu est étudié de façon isolée, ils ajoutent cependant que la relation est proportionnelle en raison du fait que les caractéristiques du risque d'un conducteur spécifique ne changent pas en fonction du kilométrage parcouru. Par exemple, peu importe l'ampleur de l'utilisation du véhicule, un individu qui réduit son kilométrage de 10% réduira vraisemblablement son risque d'accident de 10%.

Encore aujourd'hui, la plupart des assureurs ne possèdent pas de données exactes quant au kilométrage parcouru annuellement, mais plutôt une estimation fournie directement par l'assuré. Dans la plupart des cas, ces informations sont inexactes. À titre d'exemple, une compagnie d'assurance américaine, *The Travelers Company*, a déjà déclaré que 60 à 70% des véhicules qu'elle assurait étaient catégorisés comme des véhicules parcourant moins de 7500 milles par année (12 070 km), alors que la réelle distance moyenne parcourue par véhicule était plutôt de l'ordre des 12 000 milles par année, soit 19 312 kilomètres (Butler *et al.*, 1988). Généralement, il semble être avantageux pour l'assuré de mentir au sujet de son kilométrage, car aucune vérification périodique de l'odomètre n'est faite. Il est donc difficile pour les assureurs de prouver la malhonnêteté de certains assurés. Dans le cas où une réclamation est soumise et qu'une vérification de l'odomètre révèle un problème, certaines pénalités financières peuvent être appliquées. Toutefois, celles-ci ne semblent pas avoir un effet très dissuasif.

Il est donc difficile de construire un modèle de tarification basé sur l'utilisation du véhicule qui soit équitable. Vickrey (1968) émet quelques propositions qui visent à réformer le produit d'assurance automobile pour qu'il soit dorénavant tarifé en fonction de l'utilisation du véhicule. Ce type d'assurance est connu sous le nom *Pay-As-You-Drive Insurance* (PAYD). Parmi les suggestions, notons le *insured gasoline*, où le coût de l'assurance est compris dans le coût de l'essence. Notons

également le *insured tires*, où l'achat de pneus chez un concessionnaire associé à un assureur permet une couverture en cas d'accident. En effectuant ces changements, Vickrey soutient que la tarification serait plus efficace et pourrait avoir des effets bénéfiques : inciter les gens à moins utiliser leur automobile, permettre une réduction des coûts d'assurance en général, etc. Bien que ces propositions soient intéressantes, il est difficile d'implanter ce genre de système en pratique. Malgré tout, elles auront servi à lancer le débat autour des produits d'assurance PAYD.

3.1.2 *Pay-As-You-Drive* : trois structures suggérées

Une structure possible pour le PAYD, discutée par Litman (2011), est de considérer le kilométrage comme critère de tarification dans le calcul des primes d'assurance. On parle alors de ***Mileage Rate Factor*** (MRF). Certaines compagnies offrent un rabais à la fin de l'année lorsque le kilométrage parcouru d'un client se situe au-dessous d'une valeur de référence. La principale critique du MRF vient du fait que cette structure dépend des estimés de kilométrage fournis par les assurés au moment de la souscription. Comme discuté précédemment, ces estimés sont généralement plus faibles qu'ils ne devraient l'être, or évidemment la précision actuarielle des primes calculées s'en trouve impactée. Pour améliorer le MRF, Litman suggère d'introduire des audits de manière à pouvoir ajuster les primes en cours d'année et/ou en fin d'année.

Litman (2011) discute également de la structure tarifaire ***Per-Mile Premiums*** (PMP). Celle-ci change l'unité d'exposition traditionnelle, soit la durée du contrat (généralement d'une année), pour une unité de distance (dans ce mémoire, le kilomètre est l'unité de distance privilégiée). Donc, l'assuré paie un coût pour chaque unité de distance parcourue. En prenant en considération d'autres critères de tarification dans le calcul des primes, on verra alors les conducteurs les plus à

risque payer plus cher pour chaque kilomètre par rapport à ceux moins à risque. En raison d'études précédentes, dont Litman (2005), qui soulèvent que le nombre d'accidents n'est pas proportionnel avec la distance parcourue, l'auteur propose un coût d'assurance qui décroît avec le nombre de kilomètres. Essentiellement, les assurés paient à l'avance pour un certain nombre de kilomètres qu'ils estiment parcourir pendant une année. Puis, à la fin du contrat, moment où leur prime d'assurance est calculée, ils reçoivent un crédit ou une surcharge. Pour que la révision soit juste, un audit de l'odomètre est réalisé au renouvellement de chaque contrat d'assurance.

Avec l'avènement de nouvelles technologies, il est maintenant possible d'installer des systèmes GPS (*Global Positioning System*) dans les voitures. Ces derniers permettent d'obtenir le kilométrage exact que parcourent les assurés. D'autres données télémétriques peuvent également être recueillies : vitesse, heure et endroit où la conduite est effectuée, nombre d'accélérations et de freinages brusques, etc. Une nouvelle formule d'assurance PAYD, de plus en plus offerte par les compagnies d'assurance, est connue sous le nom de ***GPS-Based Pricing*** (Litman, 2011). Cette structure d'assurance permet une plus grande justesse des primes, car celles-ci varient directement en fonction du kilométrage effectué. De plus, certains assureurs intègrent d'autres éléments à leur produit PAYD tels que l'endroit et le moment de la journée où le véhicule est utilisé. Par exemple, la conduite de nuit mènera généralement à une prime plus dispendieuse qu'une conduite effectuée majoritairement le jour. Similairement, la conduite en zone urbaine produira un coût d'assurance plus important que celle en zone rurale. Comme la possibilité d'installer des systèmes GPS dans les voitures pour en extraire des données est très récente, le *GPS-Based Pricing* est en grande partie basé sur des intuitions (par exemple, plus grand risque d'avoir un accident de nuit que de jour). Bien que Jun *et al.* (2007) ont réussi à démontrer le potentiel des données sur la conduite

recueillies par GPS, des études statistiques exhaustives et poussées sur le sujet n'ont pas encore été publiées pour appuyer ces prétentions de charger plus cher pour certains comportements routiers.

Il est important de noter qu'autant la tarification *Per-Mile Premiums* que *GPS-Based Pricing* sont des options possibles à choisir pour les clients de compagnies d'assurance qui les offrent. Ainsi, le client doit consentir à ce type de tarification s'il veut s'en prévaloir. Même si plusieurs études discutent du bien-fondé d'offrir une assurance PAYD (Buxbaum 2006, Bordoff et Noel 2008), il existe encore des gens qui sont réfractaires à ce genre de produit d'assurance. Selon Litman (2011), environ 25 à 50% des assurés seraient prêts à choisir une option PMP si leur compagnie d'assurance offrait ce type d'assurance, avec un pourcentage qui s'accentuerait au fil des ans. Par opposition, une option *GPS-Based Pricing* n'attirerait seulement que de 2 à 5% des polices courantes. Ce faible pourcentage s'explique en grande partie par le sentiment d'atteinte à la vie privée que provoque l'installation de systèmes GPS dans les voitures. À ce sujet, Iqbal et Lim (2006) proposent une façon d'implémenter une assurance PAYD qui s'appuie sur la méthode de tarification par GPS, mais qui supprime complètement le côté intrusif de celle-ci.

3.1.3 Nouveau potentiel pour la recherche

L'installation de systèmes GPS dans les voitures, maintenant possible, crée un énorme potentiel pour la recherche. Cette avancée technologique permet de mieux comprendre le risque d'accident automobile, puisque des données précises relatives à une multitude de facteurs de risque sont collectées.

Dans ce mémoire, nous étudierons l'impact qu'ont le kilométrage et la durée d'exposition sur le risque de sinistre automobile. Pour ce faire, nous utiliserons les

modèles additifs généralisés. On retrouve dans la littérature quelques papiers qui discutent de ces relations. Boucher *et al.* (2013) ont démontré qu'inclure le kilométrage (données exactes) via une variable offset généralisée dans un GLM Poisson pouvait très bien expliquer la relation existante entre le kilométrage annuel et la fréquence de sinistres. De leur côté, Ayuso *et al.* (2014) ont modélisé le temps d'attente avant un premier sinistre ainsi que le nombre de kilomètres parcourus avant un premier sinistre pour des assurés ayant souscrit à une assurance de type *GPS-Based Pricing*. Pour ce faire, ils ont notamment utilisé une loi de Weibull.

L'objectif de ce chapitre est donc d'ajouter à ces recherches en présentant de quelle façon les GAM peuvent être un outil extrêmement utile pour extraire de l'information d'une base de données. Cependant, dans une perspective d'implantation pratique, les GAM sont plus ou moins envisageables. Ceci est dû en majeure partie à la simplicité et l'efficacité des GLM. Normalement, en tarification automobile, une prime est calculée pour un assuré de référence et des relativités sont appliquées pour corriger la prime en fonction des caractéristiques du risque des assurés. La partie non-paramétrique du GAM ne permet pas directement une telle pratique. C'est pourquoi un autre but de cette recherche est de montrer de quelle façon les résultats des GAM peuvent être répliqués à l'aide d'une structure tarifaire simple.

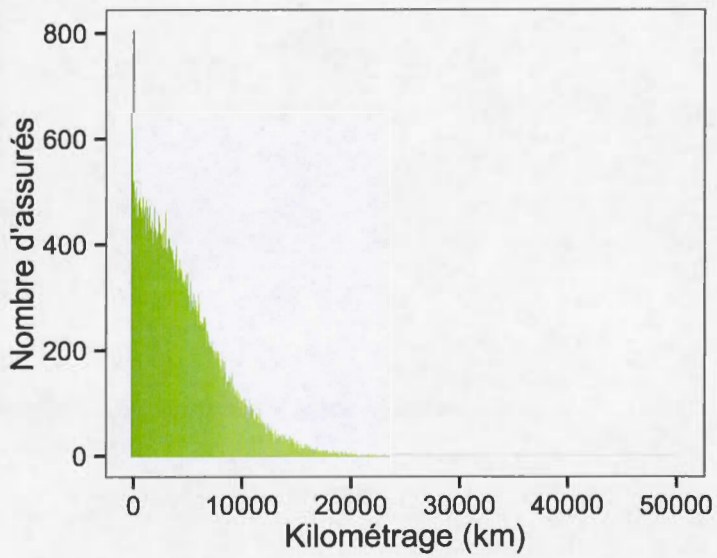
3.2 Données et statistiques descriptives

La base de données utilisée a été fournie par un assureur espagnol. Elle est constituée de 71 489 contrats d'assurance automobile PAYD, tous effectifs au courant de l'année calendaire 2011. Pour chaque contrat, plusieurs informations sont disponibles.

3.2.1 Kilométrage, durée d'exposition et nombre de réclamations

D'abord, le kilométrage exact, transmis par systèmes GPS installés dans les voitures, est disponible pour chaque assuré. Cette variable en est une clé dans cette étude, car nous sommes certains de son exactitude. Comme il a été mentionné précédemment, en temps normal, seulement une estimation du kilométrage annuel est fournie par l'assuré au début du processus d'achat d'assurance.

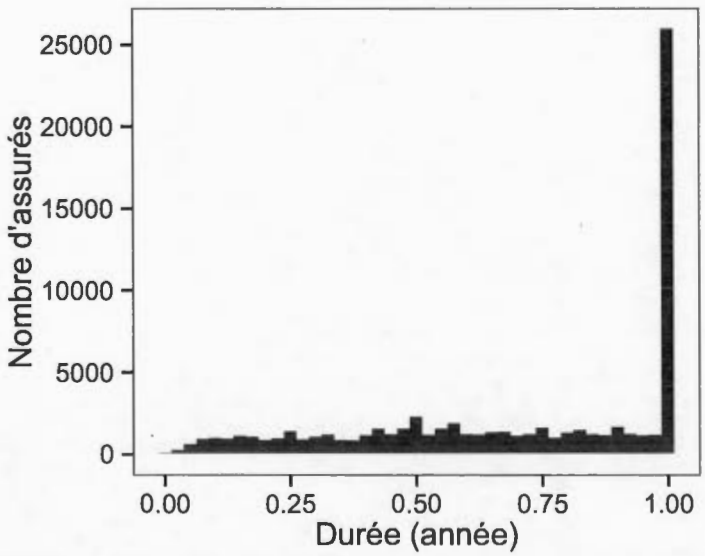
Figure 3.1: Distribution du kilométrage pour les contrats PAYD effectifs en 2011



La figure 3.1 montre sous quelle forme se présente la distribution du kilométrage parmi tous les assurés ayant un contrat d'assurance PAYD en force au cours de l'année 2011. L'histogramme est construit avec des bandes de largeur équivalente à 50 kilomètres. On y remarque une tendance générale : plus le kilométrage augmente, plus la densité d'assurés diminue. D'ailleurs, ceci peut surprendre lorsque l'on remarque que les premières bandes correspondent à un très faible kilométrage. D'abord, il faut se rappeler que tous les contrats sont considérés ici, incluant autant ceux couvrant toute l'année 2011 que ceux ne couvrant que quelques jours.

Le tableau 3.2 de la page 63 présente quelques statistiques descriptives pour la variable du kilométrage (km). On peut notamment y voir que 75% des conducteurs ont roulé moins de 6950 kilomètres durant leur période d'exposition. De plus, la moyenne et l'écart-type sont respectivement égaux à 4890 et 3978 kilomètres. À titre de comparaison, si l'on isole les assurés dont la durée observée est d'une année complète, le kilométrage moyen se situe alors à 7160 kilomètres.

Figure 3.2: Distribution de la durée observée des contrats PAYD pour l'année 2011



Pour ce qui est de la durée d'exposition, on observe une certaine variabilité à ce niveau. La figure 3.2 illustre celle-ci. L'histogramme est construit avec des bandes de largeur équivalente à 0.05 an, soit environ 18 jours. On peut y voir qu'un peu plus de 25 000 assurés ont été observés durant une année complète, ce qui correspond à 35% du nombre total de clients sous contrat durant l'année 2011. Par conséquent, 65% des gens ont eu une durée d'exposition plus petite qu'une année. La durée moyenne est de 0.706 année, c'est-à-dire environ 258 jours. Quant à l'écart-type, il est de 0.305 année, soit environ 111 jours. Le tableau 3.2 présente

des statistiques descriptives pour la durée d'exposition des véhicules assurés (voir variable d).

Dans la base de données fournie par l'assureur espagnol, les réclamations sont divisées en quatre types. Chaque sinistre est répertorié à la fois selon la nature des dommages observés et la responsabilité ou non de l'assuré dans l'occurence dudit sinistre. On a donc quatre catégories :

1. Accident responsable avec dommages matériels ($nb1$) ;
2. Accident non responsable avec dommages matériels ($nb2$) ;
3. Accident responsable avec lésions corporelles ($nb3$) ;
4. Accident non responsable avec lésions corporelles ($nb4$).

On observe très peu de réclamations liées à des accidents impliquant des lésions corporelles. Pour la suite du chapitre, la variable $nb2$ a été choisie comme variable d'étude, puisque c'est le type d'accidents qui a été le plus fréquemment observé en 2011. Il est cependant simple de généraliser les résultats obtenus pour les variables $nb1$, $nb3$ et $nb4$.

Les tableaux 3.1 et 3.2 donnent une idée de la distribution du nombre de réclamations soumises en 2011 pour des dommages matériels causés par des accidents non responsables ($nb2$). On peut y voir que près de 93% des assurés n'ont rapporté aucun incident. Ainsi, tout près de 7% des gens ont réclamé à une ou plusieurs reprises. Notons que la plupart des gens ayant un incident à leur dossier n'en ont effectivement qu'un seul, alors que moins de 1% des assurés ont été impliqués dans plus d'un sinistre en 2011.

	nb2	
Nombre de récl.	Nombre d'assurés	Pourcentage
0	66 371	92.841
1	4824	6.748
2	283	0.396
3	10	0.014
4	1	0.001
	71 489	100

Tableau 3.1: Distribution du nombre de réclamations de type *nb2*

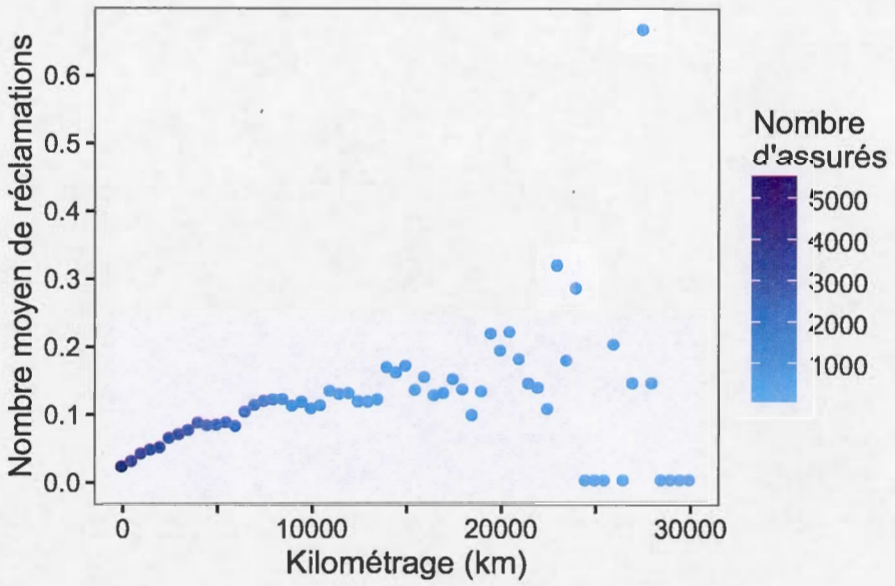
					k^{e} centile		
Variable	Moyenne	Écart-type	Min	Max	$k = 25$	$k = 50$	$k = 75$
km	4889.98	3978.50	0.005	50 035.56	1836.25	4018.32	6949.45
d	0.706	0.305	0.003	1	0.468	0.795	1
nb2	0.076	0.281	0	4	0	0	0

Tableau 3.2: Statistiques descriptives pour les variables *km*, *d* et *nb2*

Il est également intéressant de regarder empiriquement comment se comporte le nombre de réclamations en fonction de nos deux principales variables explicatives d'intérêt : le kilométrage parcouru et la durée d'exposition.

Les graphes 3.3 et 3.4 ont été construits en calculant des moyennes sur le nombre de réclamations pour des groupes d'assurés. Pour la figure 3.3, les assurés ont été regroupés sur une base de 500 kilomètres. Dans le cas de la figure 3.4, les assurés ont plutôt été regroupés sur une base de 0.1 année, ce qui équivaut environ à 37

Figure 3.3: Distribution observée du nombre de réclamations selon le kilométrage

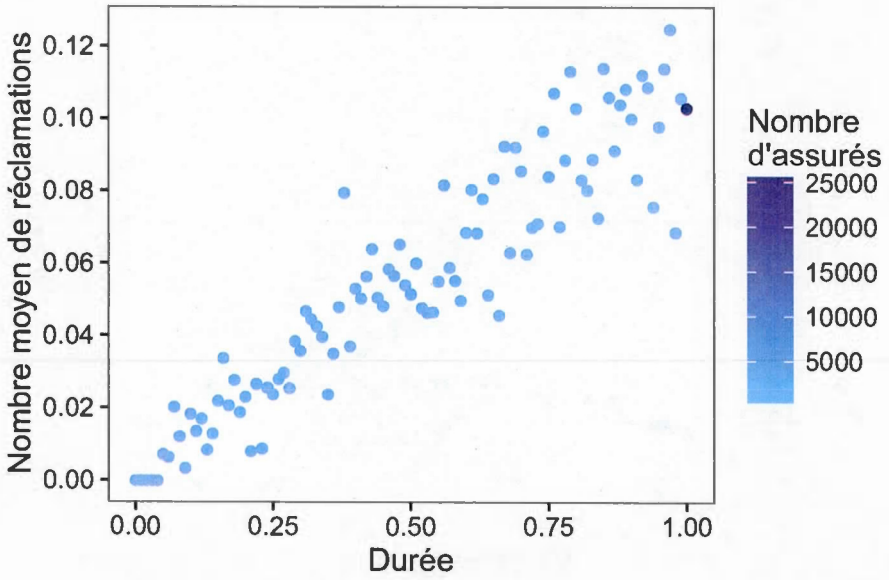


jours.

On peut notamment voir dans la figure 3.3 que bien qu'il y ait une relation positive entre le nombre de réclamations et le kilométrage parcouru, on n'est toutefois pas en présence d'une relation linéaire. L'effet du kilométrage sur la propension à réclamer pour un sinistre semble s'amenuiser plus le kilométrage augmente. Néanmoins, il faut faire preuve de prudence ici, car il ne s'agit que d'une analyse descriptive. Comme nous l'a montré la figure 3.1, une très grande concentration des assurés ont parcouru moins de 10 000 kilomètres durant leur période d'exposition. Ainsi, la densité d'observations que l'on retrouve dans chaque point du graphe 3.3 est variable.

Pour ce qui est de la figure 3.4, on est dans ce cas-ci en mesure d'apercevoir une relation qui semble s'approcher d'une relation linéaire entre le nombre de réclamations et la durée d'exposition. Encore une fois, il faut garder en tête que chaque point du graphique ne représente pas le même nombre d'observations. On

Figure 3.4: Distribution observée du nombre de réclamations selon la durée



y voit d'ailleurs que plus du tiers des assurés ont été observés sur une durée d'un an.

3.2.2 Autres caractéristiques du risque

La base de données contient également d'autres informations intéressantes quant au profil des gens qui ont acheté une protection PAYD offerte par la compagnie d'assurance espagnole. Les caractéristiques recueillies sont les suivantes :

- Âge de l'assuré (*age*) ;
- Âge du véhicule (*ageveh*) ;
- Sexe de l'assuré (*sexe*) ;
- Type de stationnement (*stn*).

Les tableaux 3.3 et 3.4 présentent des statistiques descriptives sur ces quatre variables. On peut notamment constater que les assurés sont majoritairement de

jeunes adultes. La moyenne d'âge est de 26 ans avec un écart-type d'environ 3 ans. Pour ce qui est de l'âge du véhicule, on voit grâce à la moyenne de près de 8 ans que l'assureur assure une flotte de véhicules relativement vieux. Le fait que le portefeuille d'assurés soit constitué majoritairement de jeunes explique peut-être pourquoi les véhicules sont en général âgés (les jeunes ont généralement moins de ressources financières que les gens plus vieux). On peut aussi observer que les données présentent pratiquement une parité au niveau de la distribution du sexe. Finalement, la grande majorité des assurés ont accès à un stationnement privé.

					k^{e} centile		
Variable	Moyenne	Écart-type	Min	Max	$k = 25$	$k = 50$	$k = 75$
age	25.97	3.17	18.00	37.00	23.00	26.00	29.00
ageveh	7.91	4.55	0.00	34.00	4.00	7.00	11.00

Tableau 3.3: Statistiques descriptives pour l'âge de l'assuré et l'âge du véhicule

	sexe		*stn*	
	Homme	Femme	Extérieur	Garage privé
Proportion d'assurés	53.7%	46.3%	23.3%	76.7%

Tableau 3.4: Fréquence des modalités pour le sexe et le type de stationnement

Dans la prochaine section, le nombre de réclamations sera modélisé à l'aide de différents modèles additifs généralisés. Une telle approche permettra une meilleure compréhension de l'impact qu'ont le kilométrage parcouru et la durée d'exposition sur le nombre de sinistres encourus par les assurés. Pour la suite du chapitre, les données seront séparées en deux bases distinctes : les données d'entraînement et

les données de validation. La première base servira uniquement à la modélisation, alors que la suivante sera plutôt utilisée à des fins prédictives. La division des données est faite tout à fait aléatoirement : 5000 observations ont été tirées pour faire partie de la base de validation, alors que la balance servira à la construction des différents modèles.

3.3 Modélisation avec modèles additifs généralisés

3.3.1 Modélisation avec splines cubiques indépendantes

D'abord, nous modéliserons le nombre de réclamations des assurés à l'aide d'un modèle additif généralisé où des splines cubiques indépendantes seront ajustées pour le kilométrage parcouru (km) et la durée du contrat d'assurance (d). Il s'agit d'un point de départ et il sera intéressant de voir par la suite de quelle façon l'ajout d'une interaction entre les deux variables changera les résultats.

Nous allons supposer que N_i, le nombre de sinistres ($nb2$) rapportés par l'assuré i, suit une loi de Poisson de moyenne μ_i. Un lien multiplicatif est utilisé pour lier la moyenne au prédicteur du modèle. Ainsi, le GAM décrit ici peut être représenté par l'équation suivante :

$$log(\mu_i) = \beta_0 + f_1(km_i) + f_2(d_i), \tag{3.1}$$

où β_0 est l'intercept du modèle. Les fonctions f_1 et f_2 sont des splines cubiques définies par (2.14). Pour la paramétrisation, 7 et 3 noeuds ont été choisis pour f_1 et f_2 respectivement. Le choix du nombre de noeuds est un processus manuel qui dépend du degré de flexibilité souhaité. Néanmoins, c'est une étape importante, car un nombre trop faible de noeuds produira un ajustement qui risque de ne pas capter des tendances importantes dans les données, tandis que trop de noeuds

peut mener à du surajustement. Pour le reste du mémoire, le modèle défini par (3.1) sera appelé modèle 3.1.

Les tableaux 3.5, 3.6 et 3.7 affichent les résultats de la modélisation.

	Estimé	Écart-type	Valeur t	Pr(>|t|)
$\hat{\beta}_0$	-2.7352	0.0171	-160.17	$< 2\times10^{-16}$

Tableau 3.5: Résultats pour la partie paramétrique du modèle 3.1

	EDF	Valeur F	valeur-p
$\hat{f}_1(km)$	4.30	55.61	$< 2\times10^{-16}$
$\hat{f}_2(d)$	1.95	81.53	$< 2\times10^{-16}$

Tableau 3.6: Résultats pour la partie non-paramétrique du modèle 3.1

GCV
0.38412

Tableau 3.7: GCV pour le modèle 3.1

Avant de discuter davantage des résultats, voici quelques précisions préalables :

1. *Effective degrees of freedom* (EDF) : La notion de degrés de liberté doit être adaptée dans le contexte des modèles additifs généralisés. Si les paramètres de lissage sont nuls, alors le nombre de degrés de liberté pour une fonction de lissage est simplement son nombre de paramètres à estimer moins un (contrainte que la fonction doit sommer à 1 pour une observation donnée). Dans le cas où les paramètres de lissage sont non nuls, le nombre de degrés

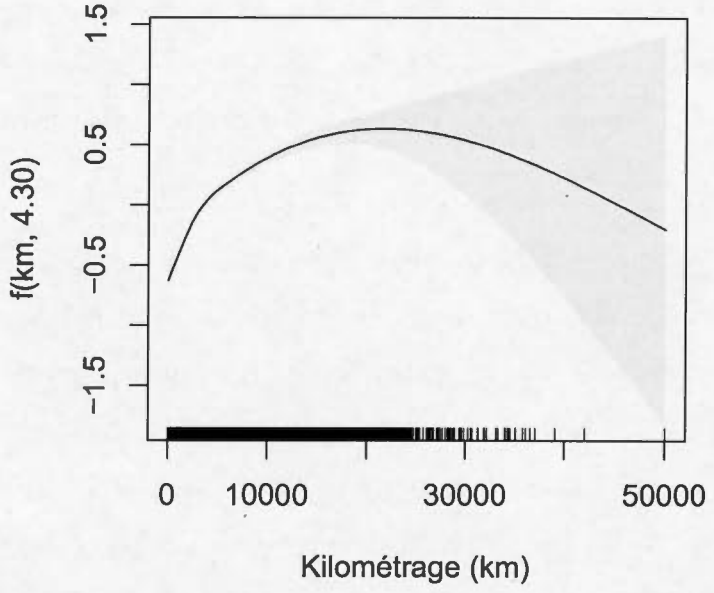
de liberté sera forcément réduit et l'on parlera alors de degrés de liberté effectifs pour quantifier la flexibilité d'une fonction de lissage ou d'un modèle en général. Mathématiquement, le nombre de EDF correspond à $tr(\mathbf{A})$, où $\mathbf{A}$ est la définition de la matrice chapeau (2.24) pour un GAM.

2. Valeur F : Statistique résultante d'un test similaire à un test de Wald pour investiguer la significativité des termes non-paramétriques. Pour plus de détails, se référer à Wood (2006), section 4.8.5.
3. GCV : Score de validation croisée généralisé, discuté au chapitre 2 et défini par l'équation (2.25).

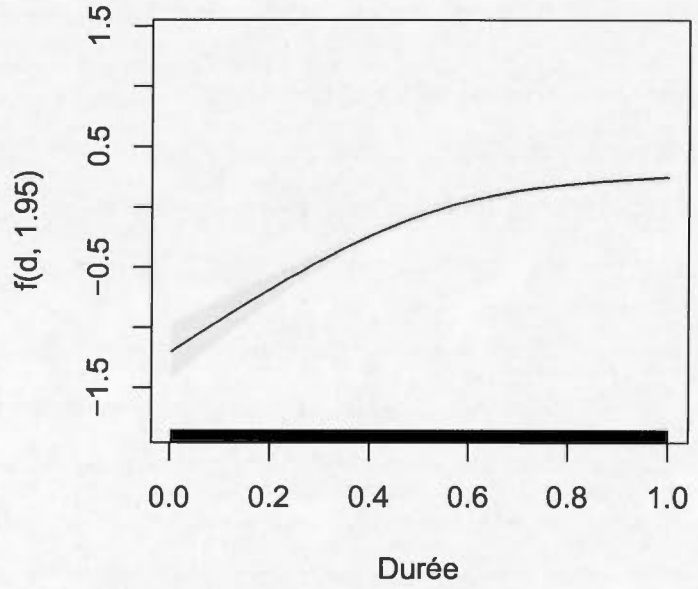
Les valeurs-p très faibles retrouvées dans les tableaux 3.5 et 3.6 montrent qu'autant la partie paramétrique que celle non-paramétrique du modèle 3.1 sont très significatives. Pour ce qui est du score GCV, la valeur de 0.38412 ne nous permet pas à ce stade-ci de statuer quoi que ce soit. Le score GCV est une statistique qui ne prend son sens que lorsque plusieurs modèles sont comparés.

La figure 3.5 présente visuellement les fonctions de lissage ajustées obtenues à la suite de l'estimation du modèle 3.1. Les courbes noires retrouvées dans les deux illustrations correspondent aux valeurs prédites pour chacune des fonctions. Les zones grises correspondent quant à elles à des intervalles de confiance à 95 % sur les prédictions. Au bas des graphes, la densité des observations qui ont servi à la modélisation est affichée.

Figure 3.5: Modèle 3.1 - Fonctions de lissage ajustées



(a) $\hat{f}_1(km)$



(b) $\hat{f}_2(d)$

Dans la figure 3.5a, on peut voir que le kilométrage a un effet sur $\hat{f}_1(km)$ qui croît rapidement dans les 10 000 premiers kilomètres. Par la suite, l'effet continue de croître plus lentement pour se stabiliser vers 20 000 kilomètres. On remarque également que $\hat{f}_1(km)$ finit par décroître, mais la confiance en les valeurs prédites devient très faible étant donné le peu de contrats observés ayant un très grand nombre de kilomètres parcourus.

En somme, la figure 3.5a montre bien la relation non proportionnelle entre le kilométrage et le nombre d'accidents. Le fait que la pente de la courbe diminue au fur et à mesure que la valeur du nombre de kilomètres augmente tend à confirmer qu'il existe des facteurs qui atténuent le risque d'accident des conducteurs aguerris par rapport aux conducteurs occasionnels. Une interprétation intuitive de ce résultat est que les gens qui conduisent beaucoup développent de meilleurs réflexes et aptitudes de conduite. De plus, une grande proportion du kilométrage des grands utilisateurs est effectué sur l'autoroute, là où les accidents sont moins fréquents qu'en zone urbaine. D'autres analyses et recherches qui permettraient de confirmer pourquoi on observe ce résultat seraient très intéressantes.

Dans le cas de la figure 3.5b, on voit que la durée d'exposition a un effet pratiquement linéaire sur $\hat{f}_2(d)$ dans les premiers six mois. Par la suite, la durée continue d'avoir un effet positif, mais de façon moins prononcée, pour terminer par n'avoir pratiquement plus d'impact notable après plus de 10 mois d'observation.

Les constats tirés de la figure 3.5b sont en contradiction avec ce que font actuellement la majorité des assureurs, c'est-à-dire supposer que le nombre d'accidents est directement proportionnel avec la durée d'exposition. Ainsi, toutes autres caractéristiques du risque étant égales par ailleurs, dire qu'un assuré observé sur un an est un risque deux fois plus important qu'un assuré observé sur 6 mois est inexact. Ceci est en ligne avec les résultats obtenus par Boucher et Denuit (2007).

3.3.2 Modélisation avec une base par produit tensoriel

Pour le modèle de la section précédente, le kilométrage effectué avec le véhicule et la durée d'exposition au risque ont été considérés comme variables explicatives via la construction de fonctions de lissage ayant pour bases des splines cubiques. Celles-ci ont été paramétrisées de façon complètement indépendante pour ensuite être estimées.

Nous allons maintenant voir de quelle manière l'ajout d'une interaction entre le kilométrage et la durée durant laquelle celui-ci est parcouru change les résultats obtenus par rapport au modèle 3.1. Nous utiliserons les mêmes outils pour la modélisation que précédemment, à savoir les modèles additifs généralisés et les splines cubiques. La différence vient du fait qu'au lieu d'utiliser deux splines cubiques indépendantes pour inclure le kilométrage et la durée dans le modèle, nous allons employer une base de lissage par produit tensoriel. Grossièrement, l'idée est d'emboîter les bases marginales par spline cubique l'une dans l'autre. Pour plus de détails théoriques sur la base par produit tensoriel, il suffit de se référer à la section 2.3.5.

À l'instar du modèle 3.1, seulement le kilométrage parcouru et la durée d'exposition seront les variables qui expliquent celle du nombre de sinistres ($nb2$) rapportés par chaque assuré. Cette dernière, qu'on dénote N_i, suit toujours une loi de Poisson de moyenne μ_i. Un lien multiplicatif est utilisé pour lier la moyenne au prédicteur linéaire du modèle. Comme mentionné précédemment, la base par produit tensoriel est utilisée pour construire la fonction bivariée qui sera employée dans le GAM. Concrètement, celui-ci est représenté par l'équation qui suit :

$$log(\mu_i) = \beta_0 + f(km_i, d_i), \tag{3.2}$$

où β_0 est l'intercept du modèle. La fonction $f(km, d)$ est définie par (2.35). Celle-ci

a été paramétrisée en admettant 7 noeuds pour la distribution du kilométrage et 3 noeuds pour celle de la durée d'exposition. Pour le reste du mémoire, le modèle défini par (3.2) sera dénommé modèle 3.2. Les tableaux 3.8, 3.9 et 3.10 affichent les résultats de la modélisation.

	Estimé	Écart-type	Valeur t	Pr(>|t|)
$\hat{\beta}_0$	-2.7402	0.0174	-157.3	$< 2\times10^{-16}$

Tableau 3.8: Résultats pour la partie paramétrique du modèle 3.2

	EDF	Valeur F	valeur-p
$\hat{f}(km, d)$	13.69	62.56	$< 2\times10^{-16}$

Tableau 3.9: Résultats pour la partie non-paramétrique du modèle 3.2

GCV
0.38403

Tableau 3.10: GCV pour le modèle 3.2

Les valeurs-p extrêmement faibles que l'on retrouve dans les tableaux 3.8 et 3.9 montrent que les parties paramétrique et non-paramétrique du modèle 3.2 sont toutes deux très significatives dans l'explication du nombre de réclamations des assurés.

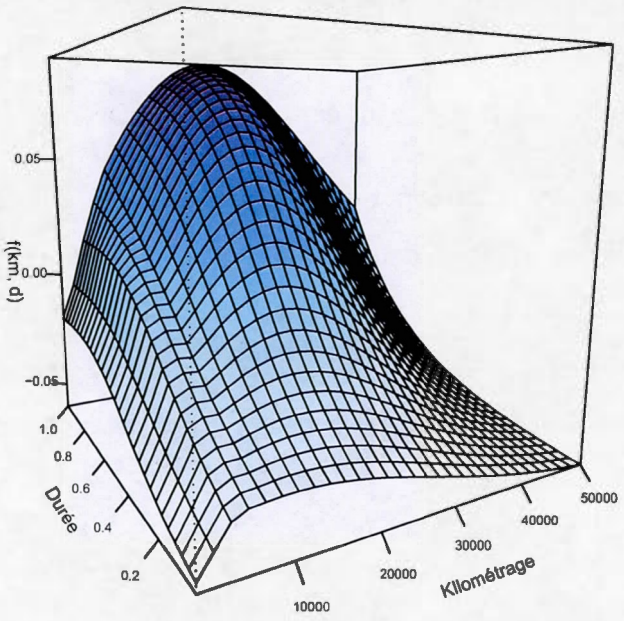
Le score GCV de 0.38403 suggère une minime amélioration par rapport au modèle 3.1 qui lui produisait un score GCV de 0.38412. Autrement dit, la flexibilité ajoutée par l'utilisation de la base par produit tensoriel (13.69 degrés de liberté effectifs

versus un total de 6.25 pour le modèle 3.1) a permis un ajustement légèrement meilleur des données de réclamations.

L'illustration 3.6 présente la surface construite par les prédictions produites par la fonction estimée $\hat{f}(km, d)$ du modèle 3.2 pour tout couple possible (km,d).

Au niveau des observations intéressantes, on peut remarquer que le kilométrage a un impact considérable pour les premiers 10 000 à 15 000 kilomètres sur $\hat{f}(km, d)$. Par la suite, l'impact s'estompe progressivement. En ce qui concerne la durée, on voit qu'elle semble impacter de manière positive et assez constante la surface prédite.

Figure 3.6: Modèle 3.2 - Fonction de lissage $\hat{f}(km, d)$ ajustée



Évidemment, dans le cas où un très grand kilométrage est enregistré, étant donné le peu d'observations disponibles à ce niveau, on s'attend à ce que les prédictions deviennent beaucoup plus volatiles. La prochaine section permettra de bien visualiser la variabilité des résultats, car une comparaison des prédictions du nombre de sinistres produites par les deux GAM ajustés jusqu'ici sera effectuée.

3.3.3 Analyse comparative

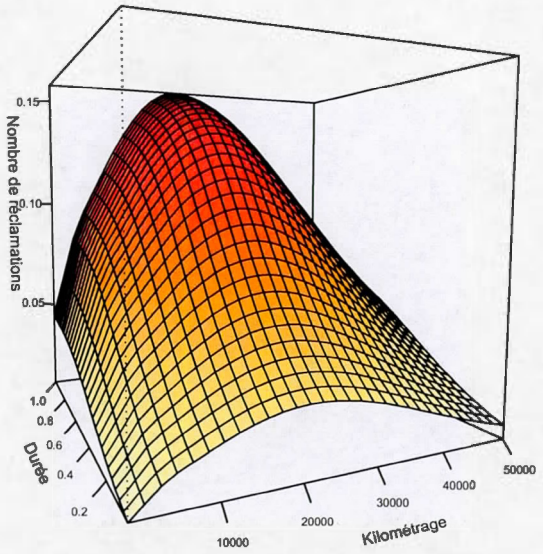
L'ajustement de deux modèles additifs généralisés a été présenté dans les sections 3.3.1 et 3.3.2 pour expliquer la fréquence du nombre d'accidents automobiles des assurés en fonction du kilométrage parcouru et de la durée d'exposition.

Le premier (modèle 3.1) était défini avec deux fonctions de lissage indépendantes pour le kilométrage et la durée. Ces fonctions étaient des splines cubiques d'ajustement dont les paramètres ont été estimés. Le deuxième (modèle 3.2) était quant à lui défini avec une fonction de lissage bivariée construite par produit tensoriel des deux splines cubiques utilisées dans le premier modèle.

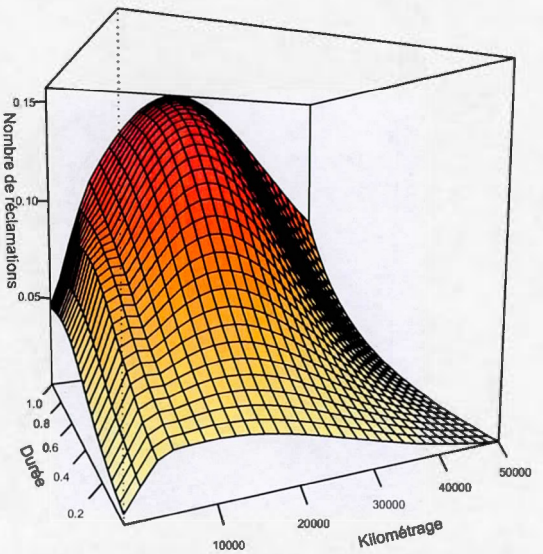
Puisque les mêmes bases par spline cubique ont été utilisées pour définir la base par produit tensoriel, les deux modèles sont directement comparables. Dans cette situation, l'équation (3.1) est « strictement incluse » (*strictly nested*) dans (3.2). En d'autres mots, la fonction $f(km, d)$ du modèle 3.2 pourrait en théorie équivaloir à n'importe quelle valeur possible de la somme définie par $f_1(km)+f_2(d)$ du modèle 3.1.

Les tableaux 3.7 et 3.10 nous ont indiqué que l'ajustement des deux modèles était très similaire selon le critère GCV. De façon plus concrète, les figures 3.7, 3.8 et 3.9 comparent les surfaces créées à partir des prédictions de la fréquence de réclamations des deux différents modèles.

Figure 3.7: Modèle 3.1 vs Modèle 3.2 (1)

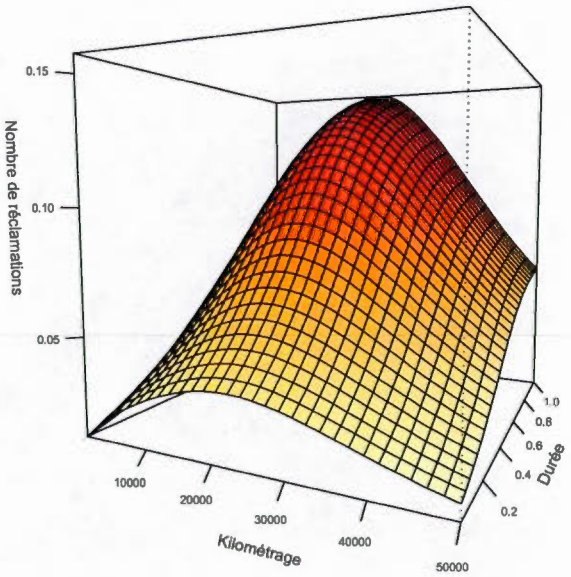


(a) **Modèle 3.1** - Surface des prédictions (angle 1)

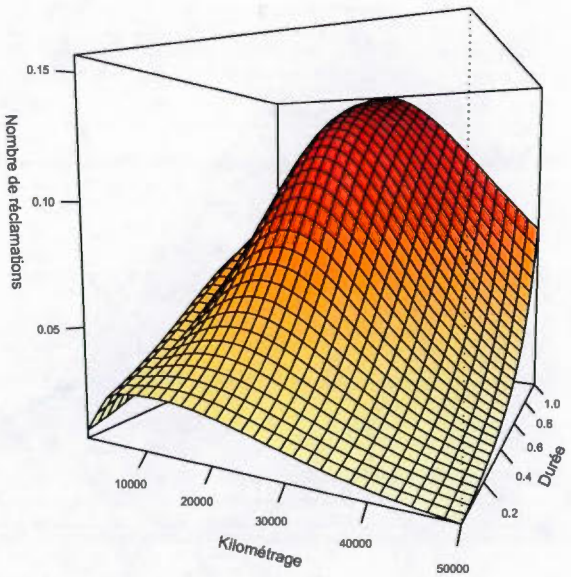


(b) **Modèle 3.2** - Surface des prédictions (angle 1)

Figure 3.8: Modèle 3.1 vs Modèle 3.2 (2)

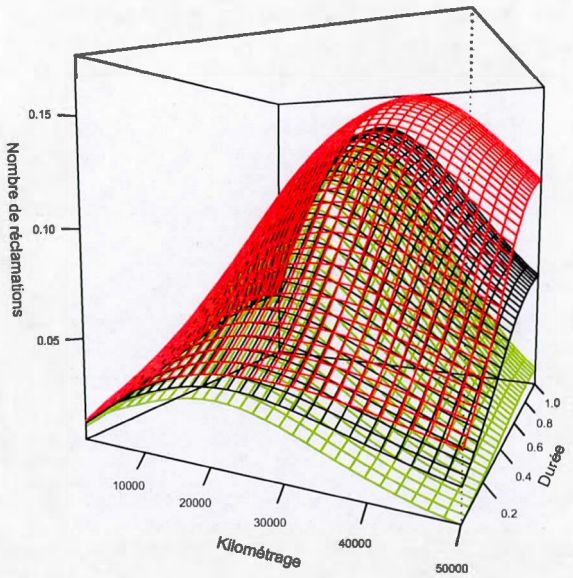


(a) **Modèle 3.1** - Surface des prédictions (angle 2)

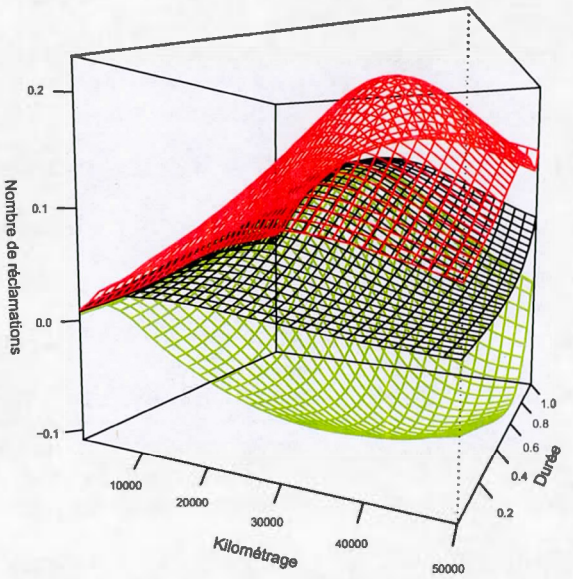


(b) **Modèle 3.2** - Surface des prédictions (angle 2)

Figure 3.9: Modèle 3.1 vs Modèle 3.2 (3)



(a) **Modèle 3.1** - Prédictions ± un écart-type



(b) **Modèle 3.2** - Prédictions ± un écart-type

D'abord, le fait que la surface affichée à la figure 3.7b soit de forme identique à celle de la figure 3.6 n'a rien de surprenant. Comme aucune variable explicative autre que le kilométrage annuel et la durée d'exposition n'est prise en considération dans le modèle 3.2, la surface des prédictions de la fréquence de sinistres se retrouve en fait à être exactement la fonction estimée $\hat{f}(km, d)$ de l'équation (3.2), mais translatée de $e^{-2.7402} = 0.0646$ unités vers le haut.

Maintenant, la première observation que l'on fait en analysant les figures 3.7 et 3.8 sont qu'elles sont très similaires. Ceci était prévisible étant donné la proximité observée des valeurs GCV des deux modèles. On peut cependant voir que la surface prédictive du modèle 3.1 est plus lisse que celle du modèle 3.2. Ce constat est dû à la flexibilité supplémentaire qu'ajoute la base par produit tensoriel par rapport à l'emploi de deux bases par spline cubique indépendantes. Le nombre de dégrés de liberté effectifs des deux modèles vient appuyer ceci : 6.25 degrés de liberté effectifs pour le modèle 3.1 versus un total de 13.69 pour le modèle 3.2.

À la vue du graphe 3.7b, cette flexibilité ajoutée semble surtout se manifester dans les prédictions de la fréquence pour les assurés effectuant entre 0 et 10 000 kilomètres. Les premiers quelques milliers de kilomètres parcourus semblent impacter plus fortement le nombre prédit d'accidents comparativement à ce que l'on voit à la figure 3.7a. Puis, une inflexion dans la surface semble survenir autour des 5000 kilomètres, dénotant un changement dans la tendance du nombre de réclamations. Il serait intéressant d'ajuster le modèle 3.2 sur des données du même assureur, mais pour une année calendaire différente. Ceci nous permettrait possiblement d'expliquer la cause de cette inflexion. Il est possible qu'une raison logique soit derrière le phénomène, mais il se peut également qu'on observe un léger surajustement ici.

En ce qui concerne la variabilité des prédictions des deux modèles, la figure 3.9

permet de visualiser rapidement que le modèle 3.1 produit des résultats beaucoup plus stables que le modèle 3.2. Pour chacun des graphes, la surface rouge correspond aux valeurs prédites ajoutées de leur écart-type, alors que celle de couleur verte correspond plutôt aux valeurs prédites auxquelles on a retranché leur écart-type. Encore une fois, la différence entre les résultats des deux modèles s'explique par la plus grande flexibilité du modèle 3.2. Concrètement, c'est l'ajout de paramètres à estimer qui provoque cette plus grande variabilité. En fait, 9 paramètres doivent être estimés pour le modèle 3.1, alors que c'est plutôt 21 paramètres qui doivent l'être dans le cas du modèle 3.2. Ainsi, plus il y a de paramètres à estimer, plus la variabilité des valeurs prédites sera grande.

Pour une visualisation plus facile des résultats des modélisations réalisées aux sections 3.3.1 et 3.3.2, les graphes à trois dimensions présentés dans les pages précédentes ont été découpés par tranche à l'annexe B. Pour chacun des modèles, une des deux variables explicatives (kilométrage parcouru ou durée d'exposition) a été fixée à différents niveaux et l'autre a été laissée continûment variable. Le comportement du nombre de réclamations prédit peut donc y être étudié d'une perspective différente.

Dans les figures B.1 et B.3, les assurés ont été regroupés sur une base de 1000 kilomètres. Dans le cas des figures B.2 et B.4, les assurés ont plutôt été regroupés sur une base de 0.1 an. De plus, chose que l'on ne pouvait pas voir dans les illustrations 3.7, 3.8 et 3.9, la densité des observations pour chaque paire (kilométrage, durée) y est également affichée. Cependant, pour des soins d'esthétique seulement, la gradation de l'axe représentant le nombre d'assurés a été limité à 200 ou 500 assurés. Malgré cela, on peut tout de même avoir une excellente idée de la distribution du nombre d'assurés pour chaque scénario.

Les figures B.1 et B.3 sont en général très similaires. Dans les quelques différences,

on peut notamment souligner que pour les premiers niveaux de kilométrage, le modèle 3.1 semble capturer une relation linéaire entre le nombre d'accidents et la durée d'exposition. Du côté du modèle 3.2, on peut voir que le nombre d'accidents prédit semble atteindre des plateaux après une observation d'environ 9 mois lorsque 4000 kilomètres ou moins sont parcourus.

Finalement, en raison de l'inflexion évoquée plus tôt dans la surface produite par le modèle 3.2, les figures B.2 et B.4 présentent des résultats qui partagent moins de similitudes. Par contre, celle-ci disparaît à partir d'une durée de 0.9 année et le modèle 3.2 présente conséquemment des résultats à toutes fins pratiques identiques à ceux du modèle 3.1 pour une durée d'exposition de 0.9 et 1 an.

3.4 Modèles linéaires généralisés ou modèles additifs généralisés ?

Le but de cette section est d'effectuer un comparatif entre la méthodologie classique utilisée en pratique par les compagnies d'assurance pour modéliser la fréquence de sinistres et les modèles présentés à la section 3.3. La plupart du temps, un modèle linéaire généralisé qui suppose un nombre de sinistres suivant un processus de Poisson est privilégié par les actuaires pour construire un modèle de fréquence dans une perspective de tarification automobile.

Dans un premier temps, c'est ce qui sera fait dans cette section. Par la suite, les données de validation, c'est-à-dire 5000 observations qui ont été mis de côté tout au long de ce chapitre, seront utilisées pour établir un comparatif entre les deux modèles additifs généralisés ajustés précédemment et le modèle classique GLM.

3.4.1 Ajustement d'un GLM Poisson aux données d'assurance

Au même titre que les modèles 3.1 et 3.2, seulement le kilométrage parcouru (km) et la durée d'exposition (d) seront pris en compte pour expliquer le nombre de sinistres ($nb2$) rapportés par chaque assuré. Nous allons segmenter la variable du kilométrage en 6 classes parmi lesquelles une sera choisie comme classe référence. Ainsi, 5 régresseurs sur le kilométrage seront ajoutés au modèle. La durée d'exposition au risque sera quant à elle traitée comme variable offset. Le but ici est vraiment de répliquer un modèle GLM Poisson qui se veut actuellement la norme dans les compagnies d'assurance.

Nous allons supposer que le nombre de sinistres N_i suit une loi de Poisson de moyenne μ_i. Un lien multiplicatif est utilisé pour lier la moyenne au prédicteur linéaire du modèle. Concrètement, celui-ci est représenté par l'équation suivante :

$$log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + log(d_i) \qquad (3.3)$$

où β_0 est l'intercept du modèle. Associons pour la suite le modèle défini par (3.3) au modèle dit 3.3. Les variables binaires x_j, $j = 1, ..., 5$, utilisées dans la modélisation sont décrites au tableau 3.11.

Variable	Description
x_1	Prend la valeur 1 si $km \leq 1000$
x_2	Prend la valeur 1 si $5000 < km \leq 10\,000$
x_3	Prend la valeur 1 si $10\,000 < km \leq 15\,000$
x_4	Prend la valeur 1 si $15\,000 < km \leq 20\,000$
x_5	Prend la valeur 1 si $km > 20\,000$

Tableau 3.11: Variables binaires utilisées pour la segmentation du kilométrage

L'assuré qui parcourt entre 1000 et 5000 kilomètres est considéré comme la référence dans le modèle. Les tableaux 3.12 et 3.19 affichent les résultats de la modélisation.

On peut notamment y voir que tous les paramètres du modèle sont significatifs. De plus, l'estimé $\hat{\beta}_1$ est le seul à être négatif, ce qui veut dire que pour une durée fixe, seulement les assurés parcourant moins de 1000 kilomètres auront une prime plus faible que la prime référence.

	Estimé	Écart-type	Valeur t	Pr(>|t|)
$\hat{\beta}_0$	-2.3568	0.0242	-97.33	$< 2\times10^{-16}$
$\hat{\beta}_1$	-0.2201	0.0727	-3.03	2×10^{-3}
$\hat{\beta}_2$	0.1989	0.0338	5.89	4×10^{-9}
$\hat{\beta}_3$	0.3426	0.0463	7.40	1×10^{-13}
$\hat{\beta}_4$	0.4734	0.0837	5.66	2×10^{-8}
$\hat{\beta}_5$	0.5777	0.1506	3.84	1×10^{-4}

Tableau 3.12: Résultats de l'estimation des paramètres du modèle 3.3

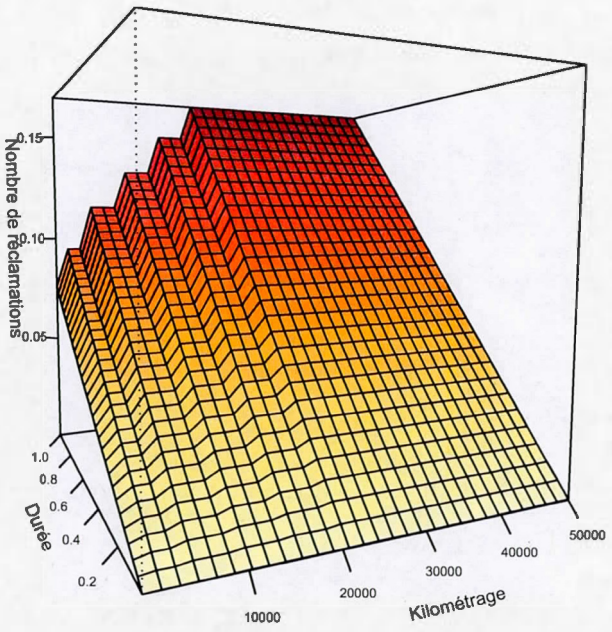
GCV
0.38640

Tableau 3.13: GCV pour le modèle 3.3

La statistique GCV a également été calculée et est de 0.38640. Si l'on compare cette valeur avec celles présentées aux tableaux 3.7 (0.38412) et 3.10 (0.38403) pour les modèles GAM ajustés à la section 3.3, on voit que modèle GLM Poisson offre un ajustement un peu moins bon des données observées.

La figure 3.10 montre la surface des prédictions produite par l'ajustement du modèle 3.3.

Figure 3.10: Modèle 3.3 - Surface des prédictions



3.4.2 Comparaison d'une tarification classique GLM vs GAM

Si les modèles additifs généralisés estimés à la section 3.3 semblaient offrir un meilleur ajustement des données d'assurance automobile que le modèle linéaire généralisé estimé à la section 3.4.1, une excellente façon d'évaluer la performance réelle d'un modèle (et ainsi éviter le risque de surajustement) est de le mettre à l'épreuve sur un jeu de données dit de validation.

C'est dans ce but précis que 5000 observations provenant des données présentées

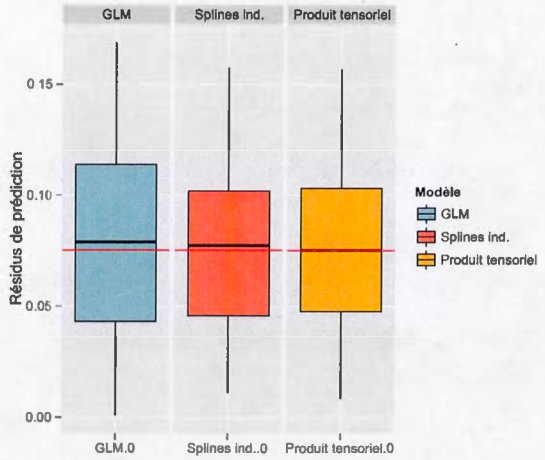
à la section 3.2 ont été exclues jusqu'ici de toute modélisation. Chacune de ces observations provenant des données de validation a donc été utilisée par les trois modèles pour produire une prédiction. Ensuite, un résidu de prédiction, qui est la différence absolue entre le nombre de sinistres prédit et le nombre de sinistres observé, a été calculé.

La figure 3.11 montre à l'aide de boîtes à moustaches la distribution des résidus de prédiction pour les trois modèles ajustés au cours du présent chapitre. Dans chaque sous-figure, on retrouve de gauche à droite respectivement les résultats pour les modèles 3.3 (GLM Poisson), 3.1 (GAM avec splines cubiques indépendantes) et 3.2 (GAM avec base par produit tensoriel).

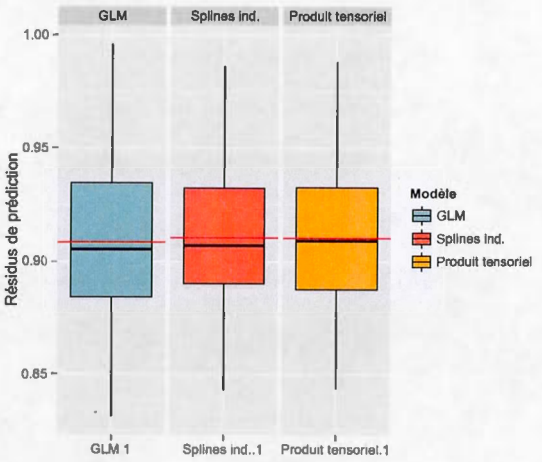
Les points suivants expliquent comment traduire l'information contenue dans chacune des boîtes à moustaches :

1. L'extrémité de la moustache inférieure correspond au plus petit résidu de prédiction ;
2. La bordure inférieure de la boîte (partie colorée) correspond au 25e centile de la distribution des résidus ;
3. La ligne horizontale noire dans la boîte correspond au 50e centile (médiane) de la distribution des résidus ;
4. La ligne horizontale rouge qui traverse la boîte correspond à la moyenne arithmétique des résidus ;
5. La bordure supérieure de la boîte correspond au 75e centile de la distribution des résidus ;
6. L'extrémité de la moustache supérieure correspond au plus grand résidu de prédiction.

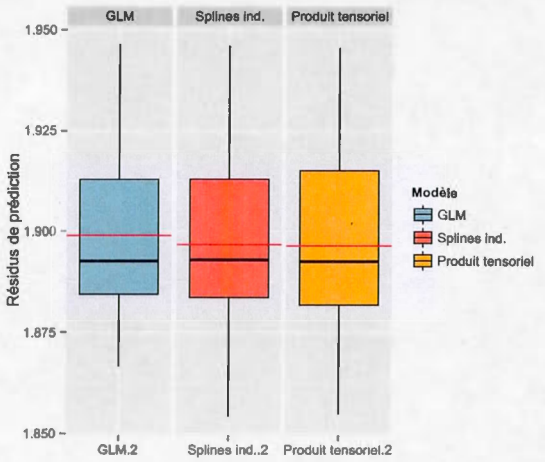
Figure 3.11: Comparaison des résidus de prédiction



(a) 0 sinistre (4644 observations)



(b) 1 sinistre (332 observations)



(c) 2 sinistres (24 observations)

Les figures 3.11a, 3.11b et 3.11c illustrent à quel point les trois modèles ont un pouvoir prédictif similaire. Malgré le fait que le GLM soit moins flexible dans sa définition que les deux GAM, le GLM fait un très bon travail. Il se permet même

d'être le meilleur modèle au niveau de la moyenne des résidus pour les assurés n'ayant eu qu'un seul sinistre (voir les lignes rouges horizontales à la figure 3.11b).

Pour ce qui est des assurés n'ayant eu aucun sinistre, c'est-à-dire la grande majorité des 5000 observations, les moyennes des résidus sont pratiquement égales pour tous les modèles. Toutefois, on voit que les deux modèles GAM produisent des prédictions qui sont moins volatiles que celles du GLM. Enfin, pour les 24 assurés qui ont eu deux accidents au cours de leur période d'exposition, les modèles GAM semblent encore une fois avoir un léger avantage par rapport au modèle GLM.

Cette analyse des résidus de prédiction démontre que peu importe le modèle qui serait appliqué en pratique, aucune grande divergence dans les résultats globaux d'une compagnie d'assurance ne serait observée (en prenant pour acquis évidemment que l'on charge la prime pure).

Jusqu'à maintenant, aucune variable explicative autre que le kilométrage et la durée d'exposition n'a été considérée. Un exercice intéressant serait de comparer les trois modèles lorsque l'on ajoute d'autres variables dans leur prédicteur linéaire. La section 3.2.2 a présenté les autres variables disponibles dans la base de données. Rappelons que celles-ci étaient l'âge de l'assuré, l'âge du véhicule, le sexe de l'assuré ainsi que le type de stationnement.

Après avoir tenté d'intégrer toutes ces variables dans chacun des modèles, seulement une s'avère être significative dans l'explication du nombre d'accidents : l'âge du conducteur.

Les modèles 3.1, 3.2 et 3.3 ont donc tous été modifiés pour inclure la variable d'âge (*age*). Celle-ci a été segmentée de la façon suivante :

Variable	Description
x_6	Prend la valeur 1 si $age \leq 25$
x_7	Prend la valeur 1 si $25 < age \leq 30$

Tableau 3.14: Variables binaires utilisées pour la segmentation de l'âge

On suppose que l'assuré de référence est âgé de plus de 30 ans. Les équations (3.4), (3.5) et (3.6) définissent les trois nouveaux modèles ajustés :

$$log(\mu_i) = \beta_0 + \beta_6 x_{6i} + \beta_7 x_{7i} + f_1(km_i) + f_2(d_i), \tag{3.4}$$

$$log(\mu_i) = \beta_0 + \beta_6 x_{6i} + \beta_7 x_{7i} + f(km_i, d_i), \tag{3.5}$$

$$log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + log(d_i). \tag{3.6}$$

Deux régresseurs (β_6 et β_7) pour la variable d'âge ont été ajoutés dans chacune des structures de nos anciens modèles. Quant aux variables binaires x_j, $j = 1, ..., 5$, elles ont les mêmes définitions que celles énumérées au tableau 3.11.

Comme d'habitude, associons le numéro de chaque équation ci-dessus au nom du modèle. Le tableau suivant présente les estimations et écarts-types obtenus pour les nouveaux paramètres β_6 et β_7 pour chacun des modèles :

	Modèle 3.4	Modèle 3.5	Modèle 3.6
$\hat{\beta}_6$	0.3917 (0.0597)	0.3906 (0.0598)	0.4260 (0.0592)
$\hat{\beta}_7$	0.1655 (0.0602)	0.1647 (0.0603)	0.1942 (0.0597)

Tableau 3.15: Résultats de l'estimation des paramètres associés à l'âge

Encore une fois, le tableau 3.15 fait foi de l'ajustement très similaire des deux modèles additifs généralisés. Tous les modèles présentent un estimé $\hat{\beta}_6$ plus élevé que

$\hat{\beta}_7$, ce qui fait du sens intuitivement étant donné que c'est la classe où se trouve les plus jeunes assurés. Ceci signifie que les jeunes âgés de 25 ans ou moins représentent un risque plus élevé d'accident que ceux âgés de 26 à 30 ans inclusivement. Comme les valeurs estimées des deux paramètres sont positives, on déduit que les assurés de plus de 30 ans sont ceux qui représentent le risque le plus faible. Il est important de préciser que ces raisonnements ne sont faits que sur la base de l'âge en supposant des valeurs fixes pour le kilométrage et la durée.

Les statistiques GCV des modèles 3.4, 3.5 et 3.6 sont respectivement de 0.38285, 0.38277 et 0.38499. Fait intéressant, les estimés $\hat{\beta}_6$ et $\hat{\beta}_7$ sont plus élevés pour le GLM que pour les GAM. Donc, malgré un ajustement qui semble similaire selon le GCV, on observe tout de même des différences entre les GAM et le GLM pour les valeurs estimées des coefficients associés à l'âge. Il y a donc forcément aussi des divergences dans le traitement réservé au kilométrage et à la durée par les deux types de modèles.

Les tableaux 3.16, 3.17 et 3.18 prouvent que c'est bien le cas. Chaque tableau affiche, pour un groupe d'âge spécifique, les primes générées par les trois modèles pour cinq profils $(km; d)$. On remarque que les primes produites par les GAM (modèle 3.4 et modèle 3.5) sont similaires. On voit aussi quelque chose de très intéressant : le GLM (modèle 3.6) offre toujours des primes beaucoup plus basses pour les trois premiers profils que les GAM. Ce constat reste vrai pour tous les groupes d'âge. Selon toute vraisemblance, les GAM semblent permettre ici une meilleure précision actuarielle que le modèle classique largement utilisé en pratique. Toutefois, comme l'exercice n'a été réalisé que sur 15 profils différents, il serait pertinent de pousser davantage l'analyse pour s'en convaincre.

	Prime		
$(km\,;\,d)$	Modèle 3.4	Modèle 3.5	Modèle 3.6
(3500 ; 0.35)	0.0519	0.0571	0.0379
(4500 ; 0.50)	0.0748	0.0778	0.0542
(9000 ; 0.65)	0.1147	0.1058	0.0860
(15 500 ; 0.90)	0.1654	0.1638	0.1594
(19 000 ; 1.00)	0.1803	0.1791	0.1771

Tableau 3.16: Primes pour assurés âgés de 25 ans et moins

	Prime		
$(km\,;\,d)$	Modèle 3.4	Modèle 3.5	Modèle 3.6
(3500 ; 0.35)	0.0414	0.0456	0.0301
(4500 ; 0.50)	0.0596	0.0621	0.0430
(9000 ; 0.65)	0.0915	0.0844	0.0682
(15 500 ; 0.90)	0.1319	0.1306	0.1264
(19 000 ; 1.00)	0.1438	0.1429	0.1405

Tableau 3.17: Primes pour assurés âgés de 26 à 30 ans

	Prime		
$(km\,;\,d)$	Modèle 3.4	Modèle 3.5	Modèle 3.6
(3500 ; 0.35)	0.0351	0.0386	0.0248
(4500 ; 0.50)	0.0505	0.0527	0.0354
(9000 ; 0.65)	0.0775	0.0716	0.0562
(15 500 ; 0.90)	0.1118	0.1108	0.1041
(19 000 ; 1.00)	0.1219	0.1212	0.1157

Tableau 3.18: Primes pour assurés âgés de plus de 30 ans (référence)

3.5 Structure tarifaire simple *Pay-As-You-Drive* (PAYD)

Un des gros avantages des modèles linéaires généralisés avec lien multiplicatif est leur simplicité d'utilisation en pratique. De plus, l'interprétation des résultats de modélisation est très intuitive. Une prime de référence est calculée, puis des relativités sont appliquées pour ajuster à la hausse ou à la baisse la prime en fonction des caractéristiques du risque d'un assuré ou groupe d'assurés.

Bien que ce ne soit pas aussi direct qu'avec un GLM, il est possible de répliquer avec un système classique de relativités les résultats d'un modèle GAM construit avec une base par spline cubique. Ceci est possible en raison du fait que l'équation (2.13) peut être réécrite sous la forme linéaire de l'équation (2.14).

Pour démontrer comment construire une table tarifaire simple à l'aide des résultats de l'ajustement d'un modèle additif généralisé, prenons l'exemple du modèle 3.1 :

$$log(\mu) = \beta_0 + f_1(km) + f_2(d),$$

où μ est la moyenne du nombre de sinistres et β_0 l'intercept du modèle. Quant à $f_1(km)$ et $f_2(d)$, il s'agit de splines cubiques, définies par l'équation (2.14),

prenant respectivement pour variable indépendante le kilométrage parcouru et la durée d'exposition au risque.

Voici les trois étapes qui permettent le développement d'une structure tarifaire à l'aide des résultats du modèle 3.1 présentés à la section 3.3.1 :

1. La prime de référence est égale à $\exp(\hat{\beta}_0) = \exp(-2.7352)$;
2. Les relativités pour les valeurs choisies de kilométrage x_j, $j = 1, ..., m$, sont égales à $\exp(\hat{f}_1(x_j))$;
3. Les relativités pour les valeurs choisies de durée y_k, $k = 1, ..., n$, sont égales à $\exp(\hat{f}_2(y_k))$.

Admettons que nous souhaitons bâtir une tarification qui segmente le kilométrage parcouru à tous les 500 kilomètres pour chaque intervalle de 0.05 an. Ces valeurs sont arbitraires et modifiables, mais évidemment plus la segmentation est fine, meilleure la réplication des prédictions obtenues directement avec le modèle GAM sera.

Donc, comme la valeur la plus élevée qui a été enregistrée pour le kilométrage est de 50035 kilomètres, on aura $m = 102$ relativités associées à la variable km. Similairement, puisque la durée maximale d'exposition en 2011 est de 1 (année complète), on aura $n = 20$ relativités associées à la variable d. En tout, cela fait un total de 2040 relativités uniques associées à chaque combinaison possible (km ; d).

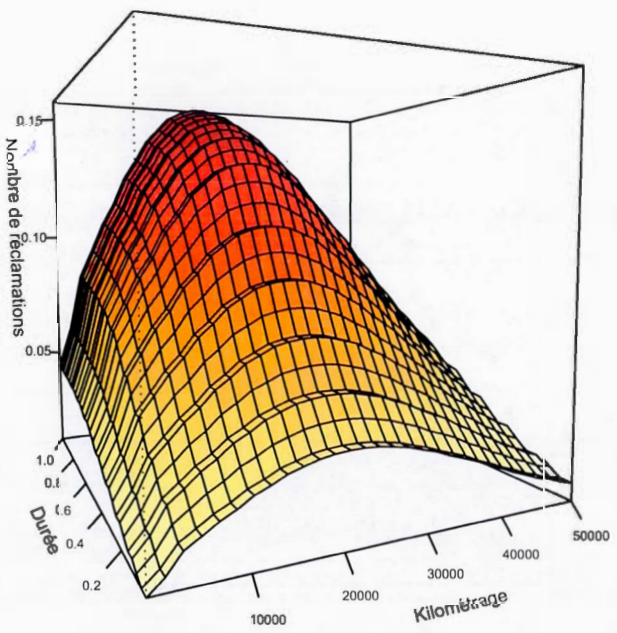
Le tableau 3.19 recense les résultats d'une telle structure tarifaire pour les cinq profils introduits à la section précédente. La prime correspond à la fréquence prédite et est calculée en multipliant la relativité totale (relativité km x relativité d) par la prime référence $\exp(-2.7352)$.

$(km\,;\,d)$	Relativité km	Relativité d	Relativité totale	Prime
(3500 ; 0.35)	0.9975	0.7144	0.7127	0.0462
(4500 ; 0.50)	1.0944	0.9341	1.0223	0.0663
(9000 ; 0.65)	1.4154	1.1059	1.5652	0.1016
(15 500 ; 0.90)	1.7713	1.2540	2.2213	0.1441
(19 000 ; 1.00)	1.8665	1.2851	2.3988	0.1556

Tableau 3.19: Structure tarifaire simple PAYD

La figure 3.12 fait état de la précision avec laquelle cette structure tarifaire réplique la figure 3.7a, qui est la surface des prédictions que produisait le modèle 3.1.

Figure 3.12: Structure tarifaire PAYD - Surface des prédictions



La même méthodologie peut être appliquée pour le modèle GAM construit avec une base par produit tensoriel, soit le modèle 3.2 :

$$log(\mu) = \beta_0 + f(km, d).$$

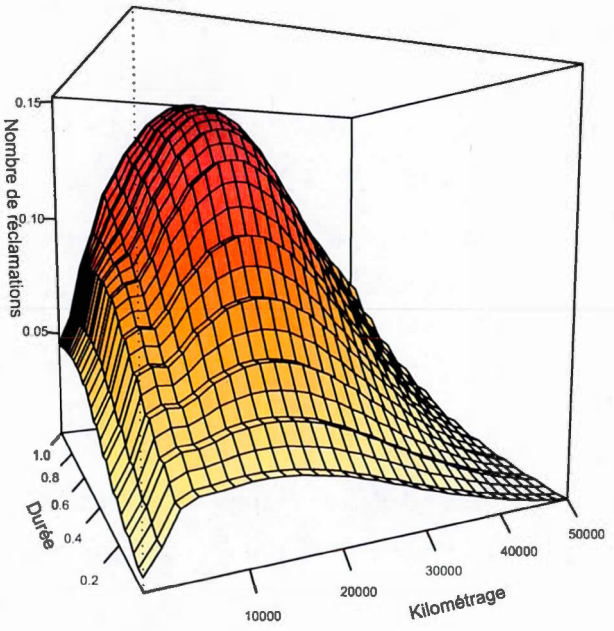
La seule différence dans ce cas-ci est que nous allons obtenir directement des relativités pour des couples $(km; d)$. Celles-ci sont égales à $\exp(\hat{f}(km, d))$. Quant à la prime de référence, est elle égale à $\exp(-2.7401)$. Le tableau 3.20 affiche les résultats pour les profils présentés plus tôt au tableau 3.19. On peut d'ailleurs y voir que les primes calculées sont très similaires à celles du tableau 3.19.

Finalement, la figure 3.13 témoigne une nouvelle fois de la justesse avec laquelle cette méthodologie parvient à répliquer les résultats de modélisation d'un GAM (voir figure 3.7b) tout en adoptant une structure tarifaire classique avec relativités.

$(km\,;\,d)$	Relativité $(km\,;\,d)$	Prime
(3500 ; 0.35)	0.7904	0.0510
(4500 ; 0.50)	1.0691	0.0690
(9000 ; 0.65)	1.4565	0.0940
(15 500 ; 0.90)	2.2103	0.1427
(19 000 ; 1.00)	2.3428	0.1513

Tableau 3.20: Structure tarifaire simple PAYD (2)

Figure 3.13: Structure tarifaire PAYD - Surface des prédictions (2)



CHAPITRE IV

CONCLUSION

Dans les prochaines années, nous assisterons probablement à une petite révolution dans le monde de l'assurance IARD. De nouvelles technologies sont désormais choses d'actualité et elles viendront sans doute changer le visage de l'assurance à moyen terme. Il suffit de taper *Internet of Things (IoT)* dans un moteur de recherche web pour constater à quel point la technologie s'est transcendée depuis quelques années. Les exemples ne manquent pas : pile de détecteur de fumée connectée qui envoie un courriel ou un message texte lorsqu'elle est presque vide de charge, voitures pourvues de systèmes d'assistance à la conduite (*Advanced Driver Assistance Systems*) qui prennent littéralement le contrôle du véhicule pour prévenir les accidents, etc. Les assureurs doivent donc déjà commencer à se préparer à intégrer ces nouvelles réalités dans leurs produits d'assurance.

L'assurance automobile, étant très réglementée, verra probablement son développement technologique ralenti. Néanmoins, depuis quelques années, un nouveau type d'assurance PAYD est né grâce à la récente possibilité d'installer des systèmes GPS directement dans les voitures : *GPS-Based Pricing*. Ces GPS transmettent beaucoup de données et cela représente un défi pour les compagnies d'assurance de savoir de quelle façon les gérer et les utiliser efficacement.

Or, c'est pour apporter des éléments de réponse à la question « Comment utiliser

ces données ? » que ce mémoire a été écrit. Au chapitre 2, les modèles additifs généralisés (GAM) ont été introduits. Les GAM offrent une grande flexibilité qui est permise grâce à l'introduction de fonctions de lissage dans la définition du prédicteur linéaire. Il a notamment été démontré que l'utilisation d'une spline cubique comme base de lissage était un choix judicieux dans une perspective d'ajustement. La construction et l'estimation d'un modèle additif généralisé ont également été couverts exhaustivement.

Le chapitre 3 s'est concentré sur l'application en assurance automobile des concepts développés au chapitre 2. Une base de données d'un assureur majeur en Espagne, qui offre un produit d'assurance auto PAYD, a été utilisée. Entre autres, on pouvait retrouver dans ces données le nombre exact de kilomètres parcourus (recueilli par GPS) pour chacun des assurés. Des modèles additifs généralisés ont été employés pour modéliser l'impact du kilométrage parcouru et de la durée d'exposition sur le risque d'accident automobile.

Dans un premier temps, la modélisation avec splines cubiques indépendantes a mis en évidence la relation non proportionnelle entre le kilométrage et le nombre d'accidents. Il a aussi été observé que le nombre d'accidents semble atteindre un plateau après un certain kilométrage, ce qui suggère que les gens qui conduisent beaucoup développent de meilleures aptitudes de conduite. De plus, les résultats de la modélisation sont en contradiction avec ce qui se fait couramment en pratique, c'est-à-dire supposer que le nombre de sinistres est proportionnel avec la durée d'exposition au risque.

Dans un deuxième temps, la modélisation avec une base par produit tensoriel, qui admet une interaction entre les variables du kilométrage et de la durée, s'est révélée être statistiquement un peu meilleure que la modélisation avec splines cubiques indépendantes. Cependant, l'amélioration sur les prédictions est légère et celles-

ci sont beaucoup plus variables. Étant donné la proximité des résultats avec le modèle GAM discuté au paragraphe précédent, les interprétations des résultats pour le GAM construit par produit tensoriel sont les mêmes.

Le chapitre 3 présente également la comparaison entre un modèle de tarification largement utilisé en pratique et les deux modèles GAM ajustés au cours du chapitre. Une analyse de résidus de prédiction effectuée sur des données de validation a montré que les performances globales des trois modèles étaient similaires. Par contre, l'analyse de l'apport de la variable d'âge dans les trois modèles a clairement indiqué que le modèle classique traitait certains profils d'individus différemment par rapport aux GAM. La flexibilité des GAM permet de constater qu'il y a possiblement une valeur à les employer dans un monde pratique. Des études futures qui détermineraient la réelle valeur (profits, meilleure satisfaction des clients, etc.) d'implanter une tarification basée sur des modèles GAM seraient très pertinentes.

Enfin, le chapitre 3 termine justement en proposant une structure tarifaire simple qui permet d'utiliser les résultats de modélisation des GAM. Ceux-ci sont souvent difficiles à interpréter et à appliquer en pratique. L'idée générale est de revenir à une structure classique où une prime de référence est multipliée par des relativités associées au kilométrage parcouru et à la durée d'exposition.

Dans de futures recherches, il serait très intéressant de pousser l'analyse des modèles additifs généralisés en incluant d'autres données télémétriques sur la conduite, comme le nombre de freinages et d'accélérations brusques, le moment de la journée et le type d'endroit où la conduite est effectuée, etc. Il serait aussi pertinent de généraliser les résultats présentés dans ce mémoire pour tous les types de réclamations (voir page 62). La dépendance entre les différents types d'accidents pourrait aussi être étudiée. Enfin, les *Generalized Additive Models for Location, Scale and Shape* (GAMLSS), proposés par Rigby et Stasinopoulos (2005), pour-

raient également être considérés. Il s'agit de modèles semi-paramétriques, tout comme les GAM, mais une différence notable vient du fait que le choix de la distribution de probabilité pour la variable à expliquer n'est plus limité aux distributions appartenant à la famille exponentielle linéaire. De plus, chaque paramètre de la loi de probabilité choisie peut être modélisé additivement avec des fonctions de variables explicatives.

ANNEXE A

EXEMPLES DU CHAPITRE 2 - COMPLÉMENT

A.1 Spline cubique d'interpolation

Nous avons les 5 couples de coordonnées suivants : (0 ;0), (1 ;1.5), (2 ;2.5), (3 ;0.5) et (4 ;1). Associons $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$ aux noeuds d'interpolation. Posons également $y_1 = 0$, $y_2 = 1.5$, $y_3 = 2.5$, $y_4 = 0.5$, $y_5 = 1$, les valeurs observées aux différents noeuds.

Le système d'équations suivant doit être résolu pour trouver une fonction par morceaux $s(x)$ dont la forme est définie par l'équation (2.6) :

$$s'(x_2)_{[x_1,x_2]} = s'(x_2)_{[x_2,x_3]}$$
$$s'(x_3)_{[x_2,x_3]} = s'(x_3)_{[x_3,x_4]}$$
$$s'(x_4)_{[x_3,x_4]} = s'(x_4)_{[x_4,x_5]}$$

Avec le résultat (2.7) de la page 16, on est en mesure de développer le système d'équations de la façon suivante :

$$\frac{1}{6}s''(x_1) + \frac{2}{3}s''(x_2) + \frac{1}{6}s''(x_3) = -0.5$$
$$\frac{1}{6}s''(x_2) + \frac{2}{3}s''(x_3) + \frac{1}{6}s''(x_4) = -3$$
$$\frac{1}{6}s''(x_3) + \frac{2}{3}s''(x_4) + \frac{1}{6}s''(x_5) = 2.5$$

En sachant que $s''(x_1) = 0$ et $s''(x_5) = 0$ et en effectuant quelques manipulations algébriques entre les 3 équations précédentes, on trouve que

$$s''(x_2) = 0.75$$
$$s''(x_3) = -6$$
$$s''(x_4) = 5.25$$

Avec l'équation (2.6), on peut maintenant définir une expression pour les 4 polynômes cubiques qu'accueillera chaque intervalle délimité par les noeuds d'interpolation. L'ensemble de ces polynômes définit complètement la fonction par morceaux $s(x)$ que l'on cherche, soit la spline cubique naturelle d'interpolation suivante :

$$s(x) = \begin{cases} 0.125x^3 + 1.375x & \text{si } 0 \leq x \leq 1 \\ 0.125(2-x)^3 - (x-1)^3 + 2.125x - 0.75 & \text{si } 1 \leq x \leq 2 \\ 0.875(x-2)^3 - (3-x)^3 - 3.875x + 11.25 & \text{si } 2 \leq x \leq 3 \\ 0.875(4-x)^3 + 1.375x - 4.5 & \text{si } 3 \leq x \leq 4 \end{cases}$$

La figure 2.2 de la page 17 illustre graphiquement $s(x)$.

A.2 Régression pénalisée par spline cubique

Associons $\mathbf{x}$ et $\mathbf{y}$ les vecteurs des observations pour la puissance (chevaux-vapeur) et la consommation d'essence (milles/gallon) :

$$\mathbf{x} = \begin{bmatrix} 110 \\ 110 \\ 93 \\ 110 \\ 175 \\ 105 \\ 245 \\ 62 \\ 95 \\ 123 \\ 123 \\ 180 \\ 180 \\ 180 \\ 205 \\ 215 \\ 230 \\ 66 \\ 52 \\ 65 \\ 97 \\ 150 \\ 150 \\ 245 \\ 175 \\ 66 \\ 91 \\ 113 \\ 264 \\ 175 \\ 335 \\ 109 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 21.0 \\ 21.0 \\ 22.8 \\ 21.4 \\ 18.7 \\ 18.1 \\ 14.3 \\ 24.4 \\ 22.8 \\ 19.2 \\ 17.8 \\ 16.4 \\ 17.3 \\ 15.2 \\ 10.4 \\ 10.4 \\ 14.7 \\ 32.4 \\ 30.4 \\ 33.9 \\ 21.5 \\ 15.5 \\ 15.2 \\ 13.3 \\ 19.2 \\ 27.3 \\ 26.0 \\ 30.4 \\ 15.8 \\ 19.7 \\ 15.0 \\ 21.4 \end{bmatrix}.$$

On choisit $q = 10$ noeuds pour la distribution des chevaux-vapeur : $[w_1 \cdots w_q] = \begin{bmatrix} 52.00 & 65.33 & 92.33 & 97.00 & 109.33 & 119.67 & 175.00 & 208.33 & 240.00 & 335.00 \end{bmatrix}$.

En utilisant l'équation (2.14), on peut construire la matrice de design $\mathbf{X}$ du modèle dont la forme est définie par (2.16). Pour ce faire, on a besoin de la matrice $\mathbf{F}$. Rappelons sa définition qui se trouve à la page 25 :

$$\mathbf{F} = \begin{bmatrix} \mathbf{0} \\ \mathbf{F}^- \\ \mathbf{0} \end{bmatrix},$$

où $\mathbf{F}^- = \mathbf{B}^{-1}\mathbf{D}$. Les matrices $\mathbf{B}$ et $\mathbf{D}$ sont définies dans le tableau 2.2. Voici les résultats que l'on obtient pour les matrices $\mathbf{B}$, $\mathbf{D}$, $\mathbf{F}$ et $\mathbf{X}$:

$$\mathbf{B} = \begin{bmatrix} 13.44444 & 4.50000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 4.50000 & 10.55556 & 0.77778 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.77778 & 5.66667 & 2.05556 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 2.05556 & 7.55556 & 1.72222 & 0.00000 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 1.72222 & 21.88889 & 9.22222 & 0.00000 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 9.22222 & 29.55556 & 5.55556 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 5.55556 & 21.66667 & 5.27778 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 5.27778 & 42.22222 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix}
0.07500 & -0.11204 & 0.03704 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.03704 & -0.25132 & 0.21429 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.21429 & -0.29537 & 0.08108 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.08108 & -0.17786 & 0.09677 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.09677 & -0.11485 & 0.01807 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.01807 & -0.04807 & 0.03000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.03000 & -0.06158 & 0.03158 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.03158 & -0.04211 & 0.01053
\end{bmatrix},$$

$$\mathbf{F} = \begin{bmatrix}
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00652 & -0.01113 & 0.01386 & -0.00985 & 0.00076 & -0.00017 & 0.00001 & -0.00000 & 0.00000 & -0.00000 \\
-0.00281 & 0.00835 & -0.03319 & 0.02943 & -0.00227 & 0.00051 & -0.00002 & 0.00001 & -0.00000 & 0.00000 \\
0.00043 & -0.00127 & 0.04712 & -0.06688 & 0.02644 & -0.00596 & 0.00018 & -0.00008 & 0.00002 & -0.00000 \\
-0.00012 & 0.00035 & -0.01309 & 0.02954 & -0.03258 & 0.01623 & -0.00050 & 0.00021 & -0.00004 & 0.00000 \\
0.00001 & -0.00003 & 0.00120 & -0.00270 & 0.00811 & -0.00789 & 0.00198 & -0.00084 & 0.00017 & -0.00001 \\
-0.00000 & 0.00001 & -0.00039 & 0.00089 & -0.00266 & 0.00323 & -0.00264 & 0.00196 & -0.00040 & 0.00001 \\
0.00000 & -0.00000 & 0.00010 & -0.00023 & 0.00070 & -0.00085 & 0.00213 & -0.00364 & 0.00186 & -0.00007 \\
-0.00000 & 0.00000 & -0.00001 & 0.00003 & -0.00009 & 0.00011 & -0.00027 & 0.00120 & -0.00123 & 0.00026 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000
\end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix}
0.00024 & -0.00070 & 0.02585 & -0.05833 & 0.99395 & 0.03980 & -0.00122 & 0.00052 & -0.00011 & 0.00000 \\
0.00024 & -0.00070 & 0.02585 & -0.05833 & 0.99395 & 0.03980 & -0.00122 & 0.00052 & -0.00011 & 0.00000 \\
0.00210 & -0.00624 & 0.86061 & 0.15254 & -0.01155 & 0.00260 & -0.00008 & 0.00003 & -0.00001 & 0.00000 \\
0.00024 & -0.00070 & 0.02585 & -0.05833 & 0.99395 & 0.03980 & -0.00122 & 0.00052 & -0.00011 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00000 & 0.00000 & 0.00000 & 0.00000 \\
-0.00221 & 0.00658 & -0.24320 & 0.59213 & 0.75256 & -0.10806 & 0.00334 & -0.00142 & 0.00029 & -0.00001 \\
0.00002 & -0.00005 & 0.00190 & -0.00428 & 0.01285 & -0.01561 & 0.03884 & -0.17566 & 1.12699 & 0.01501 \\
0.18662 & 0.85818 & -0.13479 & 0.09576 & -0.00739 & 0.00167 & -0.00005 & 0.00002 & -0.00000 & 0.00000 \\
0.00297 & -0.00882 & 0.40490 & 0.62748 & -0.03404 & 0.00767 & -0.00024 & 0.00010 & -0.00002 & 0.00000 \\
-0.00050 & 0.00149 & -0.05497 & 0.12403 & -0.37271 & 1.28263 & 0.03032 & -0.01283 & 0.00264 & -0.00009 \\
-0.00050 & 0.00149 & -0.05497 & 0.12403 & -0.37271 & 1.28263 & 0.03032 & -0.01283 & 0.00264 & -0.00009 \\
0.00013 & -0.00039 & 0.01433 & -0.03233 & 0.09715 & -0.11799 & 0.90772 & 0.16308 & -0.03288 & 0.00117 \\
0.00013 & -0.00039 & 0.01433 & -0.03233 & 0.09715 & -0.11799 & 0.90772 & 0.16308 & -0.03288 & 0.00117 \\
0.00013 & -0.00039 & 0.01433 & -0.03233 & 0.09715 & -0.11799 & 0.90772 & 0.16308 & -0.03288 & 0.00117 \\
0.00004 & -0.00011 & 0.00391 & -0.00882 & 0.02651 & -0.03219 & 0.08111 & 0.97924 & -0.05151 & 0.00184 \\
-0.00004 & 0.00013 & -0.00472 & 0.01066 & -0.03203 & 0.03890 & -0.09680 & 0.92988 & 0.15939 & -0.00536 \\
-0.00004 & 0.00011 & -0.00414 & 0.00935 & -0.02809 & 0.03412 & -0.08491 & 0.41551 & 0.67060 & -0.01251 \\
-0.02925 & 1.01459 & 0.04408 & -0.03130 & 0.00242 & -0.00054 & 0.00002 & -0.00001 & 0.00000 & -0.00000 \\
1.00000 & 0.00000 & 0.00000 & -0.00000 & 0.00000 & -0.00000 & 0.00000 & -0.00000 & 0.00000 & -0.00000 \\
0.01570 & 0.99087 & -0.01978 & 0.01405 & -0.00108 & 0.00024 & -0.00001 & 0.00000 & -0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 1.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 \\
-0.00130 & 0.00385 & -0.14255 & 0.32165 & -0.96657 & 1.26618 & 0.70240 & -0.22916 & 0.04718 & -0.00168 \\
-0.00130 & 0.00385 & -0.14255 & 0.32165 & -0.96657 & 1.26618 & 0.70240 & -0.22916 & 0.04718 & -0.00168 \\
0.00002 & -0.00005 & 0.00190 & -0.00428 & 0.01285 & -0.01561 & 0.03884 & -0.17566 & 1.12699 & 0.01501 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00000 & 0.00000 & 0.00000 & 0.00000 \\
-0.02925 & 1.01459 & 0.04408 & -0.03130 & 0.00242 & -0.00054 & 0.00002 & -0.00001 & 0.00000 & -0.00000 \\
-0.00775 & 0.02312 & 1.23692 & -0.26845 & 0.02073 & -0.00467 & 0.00014 & -0.00006 & 0.00001 & -0.00000 \\
0.00074 & -0.00219 & 0.08115 & -0.18311 & 0.81879 & 0.28961 & -0.00756 & 0.00322 & -0.00066 & 0.00002 \\
0.00006 & -0.00017 & 0.00644 & -0.01453 & 0.04366 & -0.05303 & 0.13196 & -0.59684 & 1.35766 & 0.12479 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00000 & 0.00000 & 0.00000 & 0.00000 \\
0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 1.00000 \\
-0.00014 & 0.00041 & -0.01504 & 0.03396 & 0.99772 & -0.01726 & 0.00053 & -0.00023 & 0.00005 & -0.00000
\end{bmatrix}.$$

Il faut maintenant déterminer la matrice de pénalité $\mathbf{S}$. Celle-ci correspond à $\mathbf{S} = \mathbf{D}^\top \mathbf{B}^{-1} \mathbf{D}$ (voir équation 2.15). De plus, pour cet exemple, on fixe $\lambda = 10\,000$. On obtient ainsi

$$\lambda\mathbf{S} = \begin{bmatrix}
4.8895 & -8.3452 & 10.3985 & -7.3873 & 0.5704 & -0.1285 & 0.0040 & -0.0017 & 0.0003 & -0.0000 \\
-8.3452 & 15.5574 & -27.8269 & 21.9346 & -1.6937 & 0.3816 & -0.0118 & 0.0050 & -0.0010 & 0.0000 \\
10.3985 & -27.8269 & 189.5260 & -220.9179 & 62.6469 & -14.1134 & 0.4359 & -0.1858 & 0.0382 & -0.0014 \\
-7.3873 & 21.9346 & -220.9179 & 284.5487 & -109.3767 & 31.8462 & -0.9836 & 0.4192 & -0.0863 & 0.0031 \\
0.5704 & -1.6937 & 62.6469 & -109.3767 & 87.2268 & -41.3197 & 2.9556 & -1.2597 & 0.2593 & -0.0093 \\
-0.1285 & 0.3816 & -14.1134 & 31.8462 & -41.3197 & 25.3432 & -3.2356 & 1.5300 & -0.3150 & 0.0112 \\
0.0040 & -0.0118 & 0.4359 & -0.9836 & 2.9556 & -3.2356 & 2.2668 & -2.1871 & 0.7837 & -0.0280 \\
-0.0017 & 0.0050 & -0.1858 & 0.4192 & -1.2597 & 1.5300 & -2.1871 & 3.2088 & -1.6553 & 0.1266 \\
0.0003 & -0.0010 & 0.0382 & -0.0863 & 0.2593 & -0.3150 & 0.7837 & -1.6553 & 1.1054 & -0.1295 \\
-0.0000 & 0.0000 & -0.0014 & 0.0031 & -0.0093 & 0.0112 & -0.0280 & 0.1266 & -0.1295 & 0.0271
\end{bmatrix}.$$

On a maintenant tout ce qui est nécessaire pour calculer le vecteur $\hat{\boldsymbol{\beta}}$ dont les éléments correspondent aux paramètres estimés de la régression par spline cubique.

Avec l'équation (2.18) qui définit $\hat{\boldsymbol{\beta}}$, on trouve que

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 29.94 & 28.03 & 24.24 & 23.62 & 22.07 & 20.88 & 16.27 & 14.56 & 13.83 & 14.13 \end{bmatrix}.$$

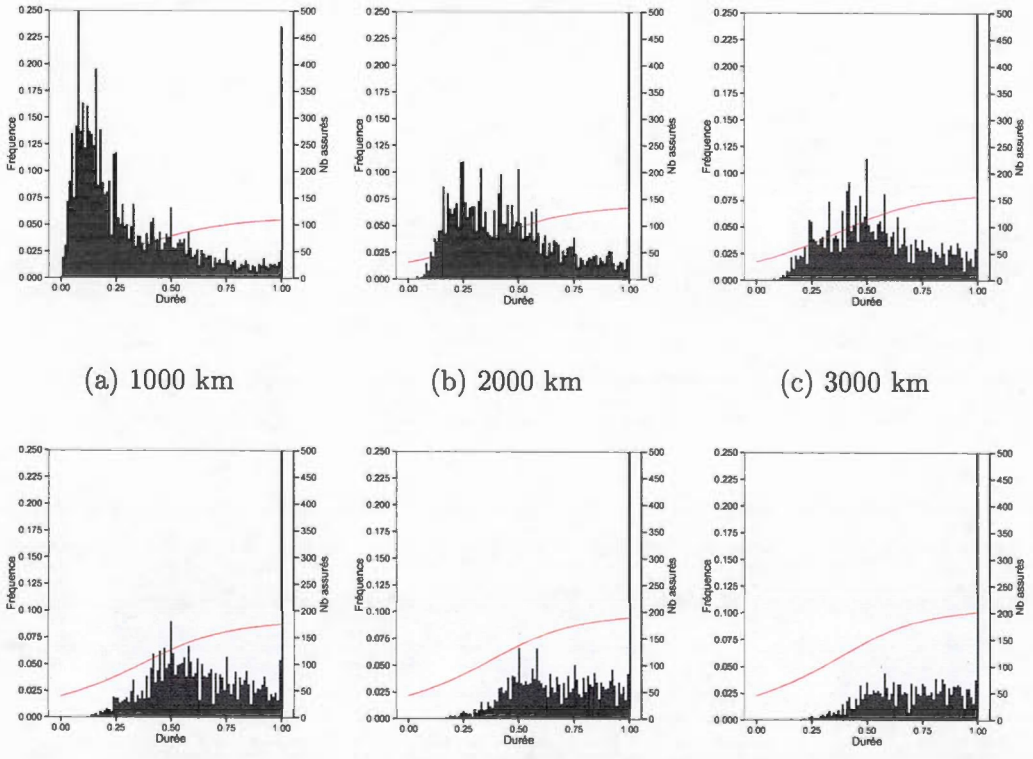
On a désormais une spline cubique complètement spécifiée qui s'exprime sous la forme de l'équation (2.14). Chaque fonction de base illustrée graphiquement à la figure 2.4 est multipliée par son paramètre estimé $\hat{\beta}_k$. Finalement, toutes les nouvelles courbes sont sommées pour ne donner qu'une seule courbe qui correspond à la spline cubique ajustée. Celle-ci est illustrée à la figure 2.5.

ANNEXE B

ANALYSE GRAPHIQUE

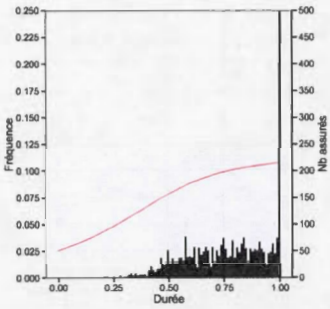
B.1 Analyse graphique de la surface prédite par le modèle 3.1

Figure B.1: Décomposition par tranche (fréquence x durée) de la figure 3.7a

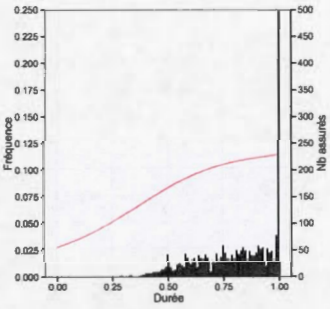


(a) 1000 km (b) 2000 km (c) 3000 km

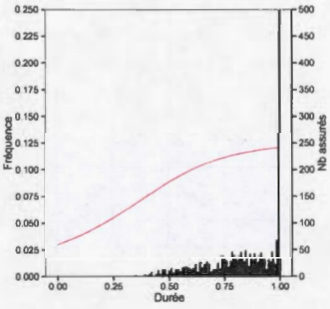
(d) 4000 km (e) 5000 km (f) 6000 km



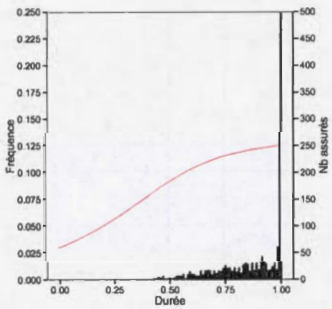
(g) 7000 km



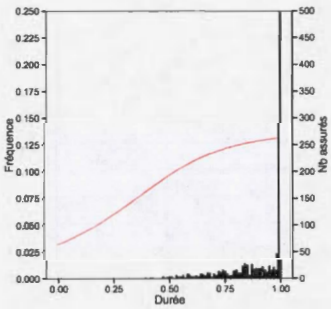
(h) 8000 km



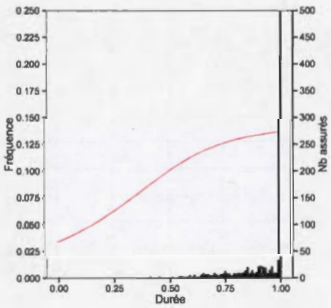
(i) 9000 km



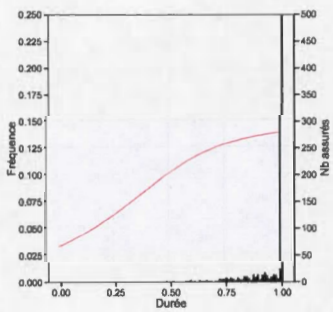
(j) 10 000 km



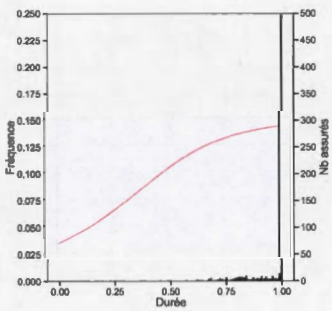
(k) 11 000 km



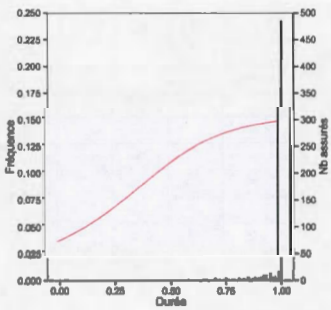
(l) 12 000 km



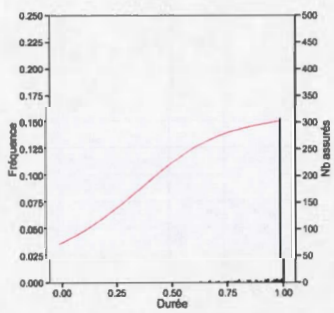
(m) 13 000 km



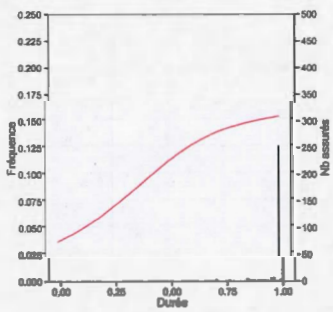
(n) 14 000 km



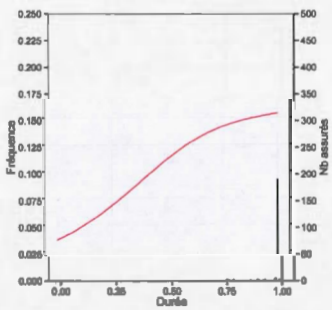
(o) 15 000 km



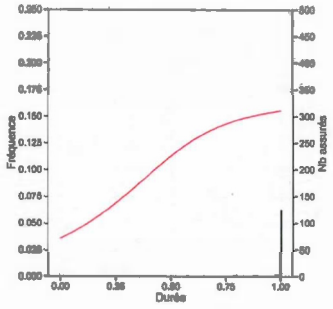
(p) 16 000 km



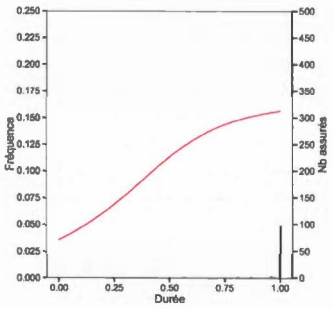
(q) 17 000 km



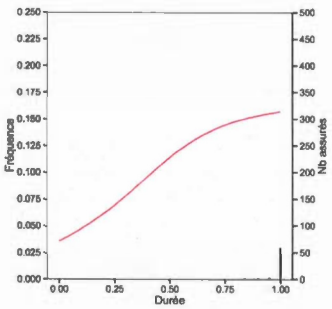
(r) 18 000 km



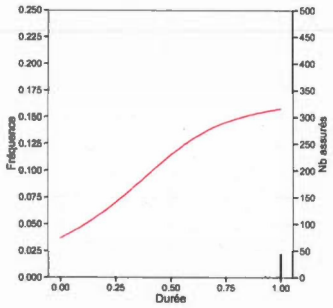
(s) 19 000 km



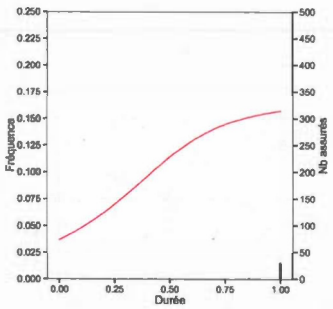
(t) 20 000 km



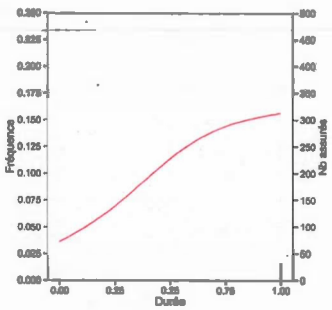
(u) 21 000 km



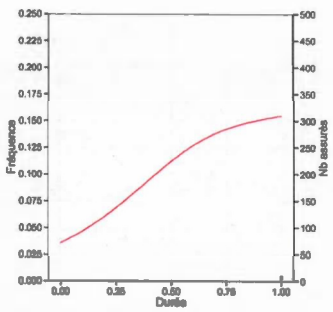
(v) 22 000 km



(w) 23 000 km

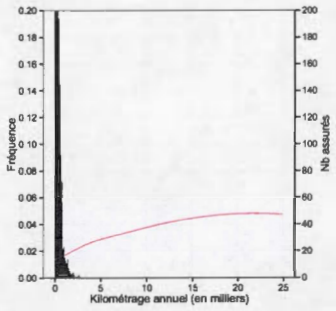


(x) 24 000 km

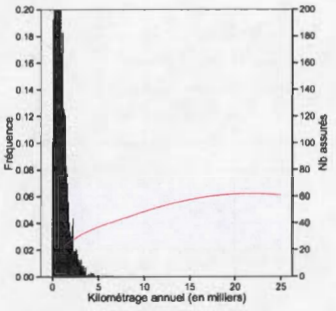


(y) 25 000 km

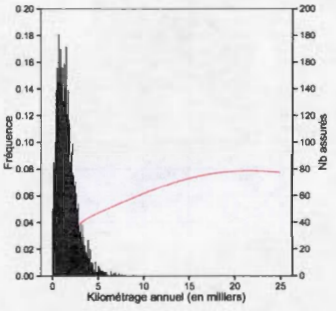
Figure B.2: Décomposition par tranche (fréquence x kilométrage) de la figure 3.7a



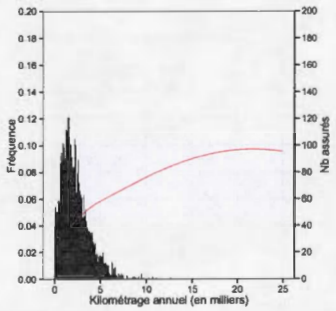
(a) 0.1 an



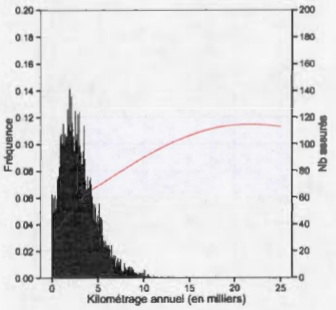
(b) 0.2 an



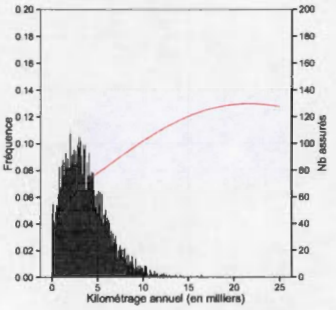
(c) 0.3 an



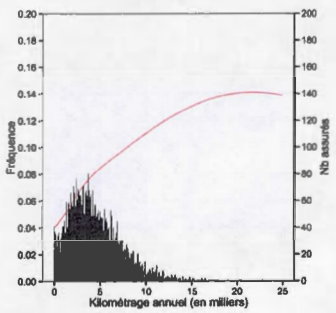
(d) 0.4 an



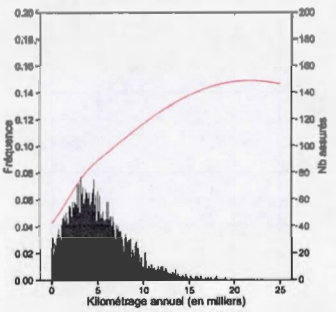
(e) 0.5 an



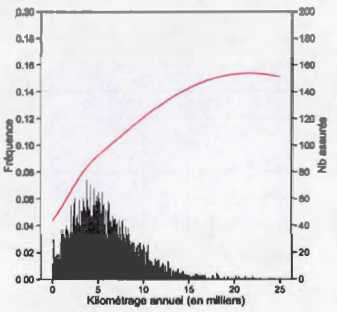
(f) 0.6 an



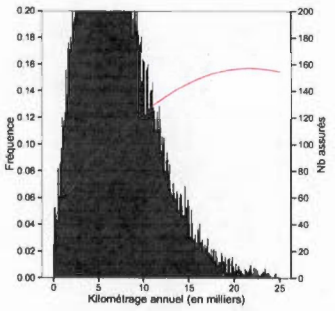
(g) 0.7 an



(h) 0.8 an



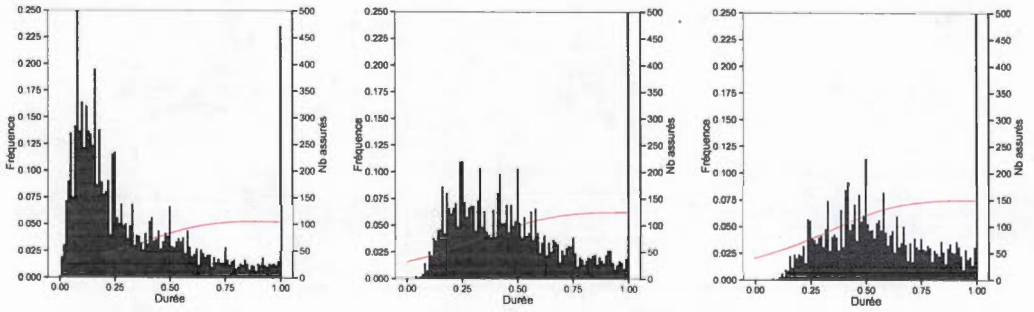
(i) 0.9 an



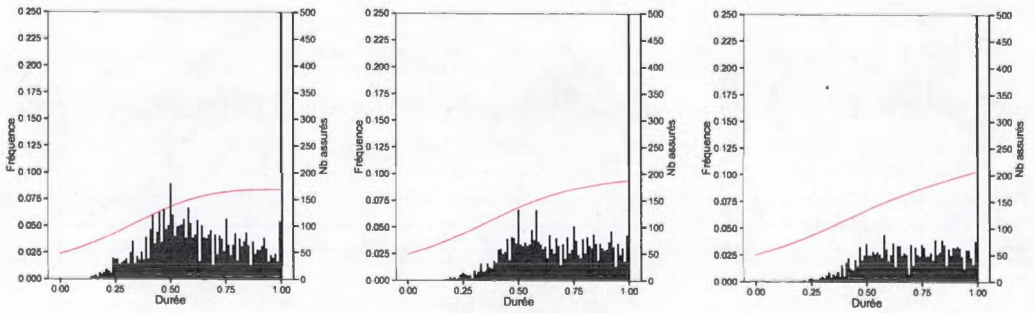
(j) 1 an

B.2 Analyse graphique de la surface prédite par le modèle 3.2

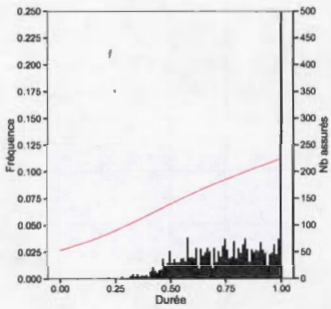
Figure B.3: Décomposition par tranche (fréquence x durée) de la figure 3.7b



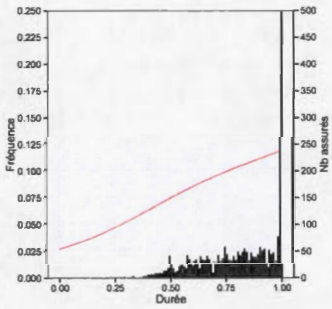
(a) 1000 km (b) 2000 km (c) 3000 km



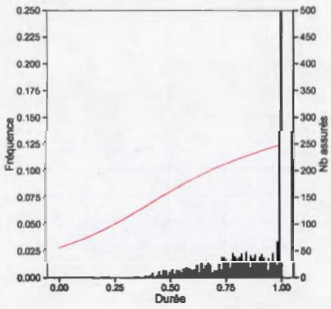
(d) 4000 km (e) 5000 km (f) 6000 km



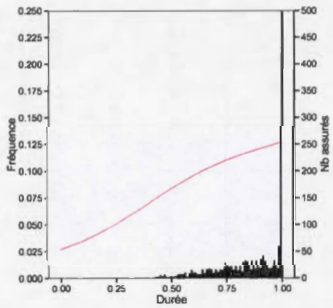
(g) 7000 km



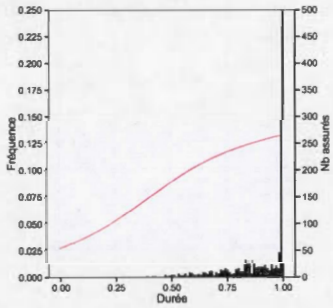
(h) 8000 km



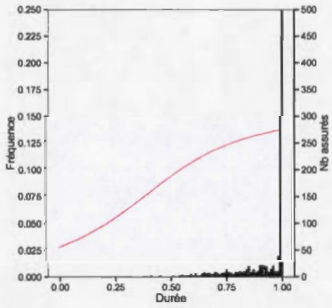
(i) 9000 km



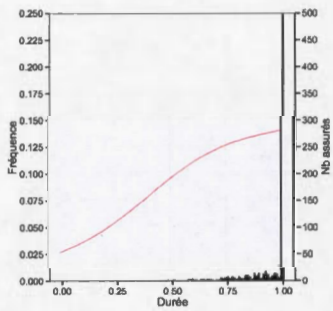
(j) 10 000 km



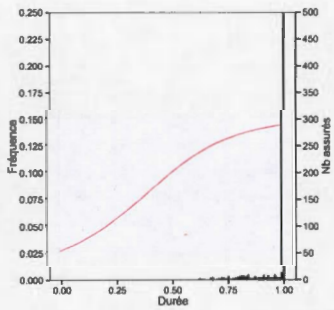
(k) 11 000 km



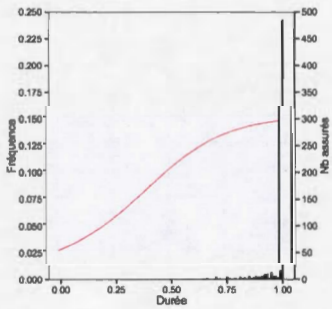
(l) 12 000 km



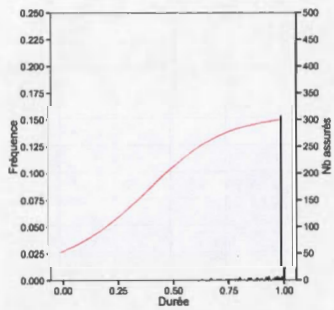
(m) 13 000 km



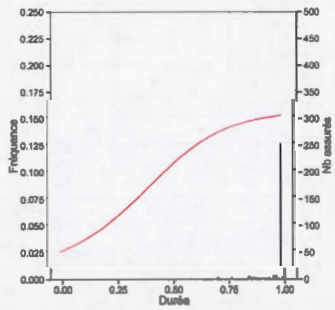
(n) 14 000 km



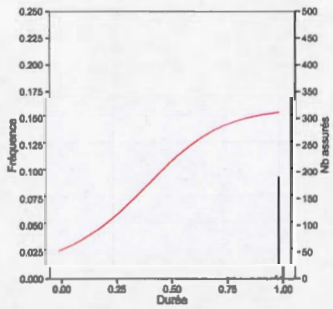
(o) 15 000 km



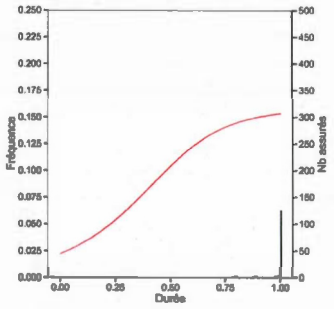
(p) 16 000 km



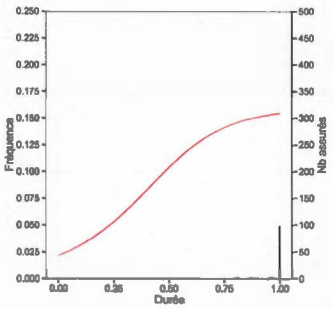
(q) 17 000 km



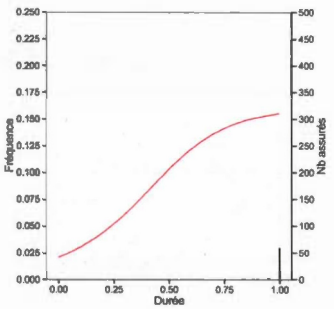
(r) 18 000 km



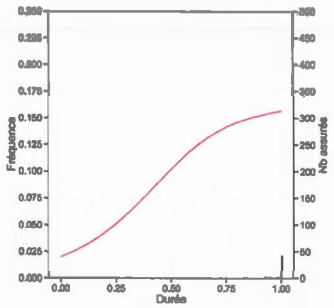
(s) 19 000 km



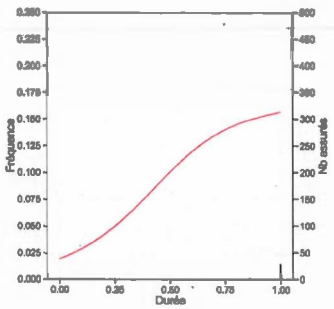
(t) 20 000 km



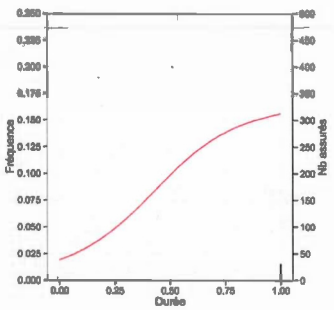
(u) 21 000 km



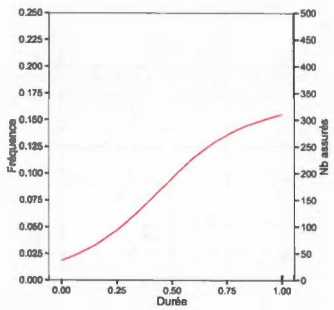
(v) 22 000 km



(w) 23 000 km

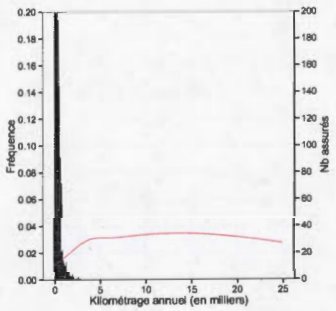


(x) 24 000 km

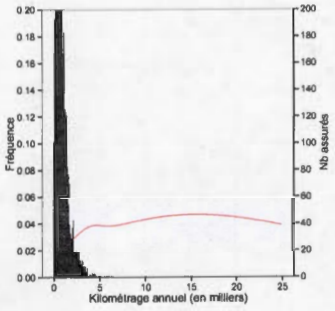


(y) 25 000 km

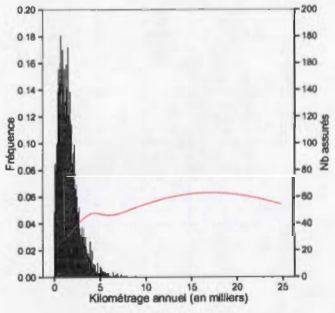
Figure B.4: Décomposition par tranche (fréquence x kilométrage) de la figure 3.7b



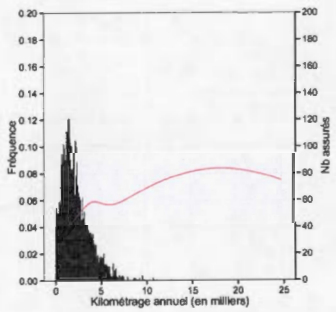
(a) 0.1 an



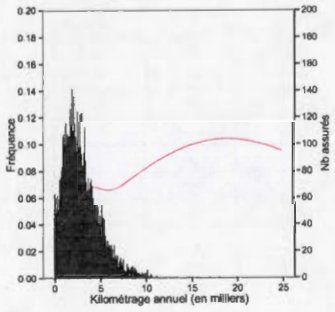
(b) 0.2 an



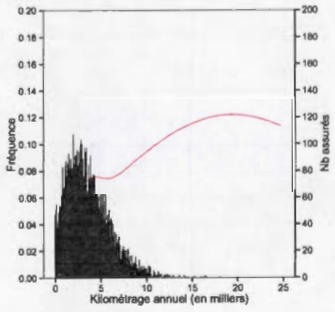
(c) 0.3 an



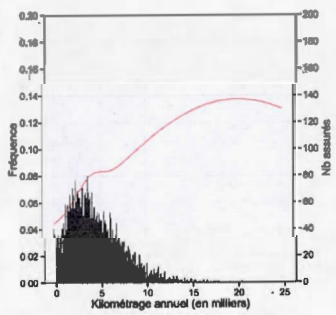
(d) 0.4 an



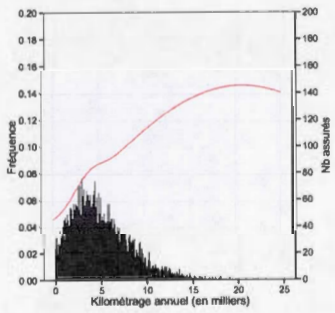
(e) 0.5 an



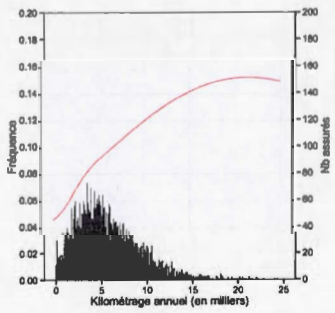
(f) 0.6 an



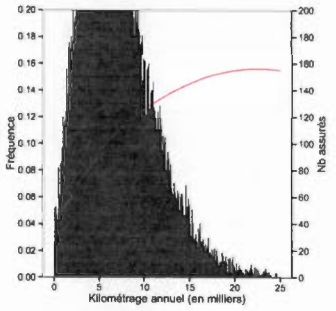
(g) 0.7 an



(h) 0.8 an



(i) 0.9 an



(j) 1 an

RÉFÉRENCES

Ayuso, M., Guillén, M. et Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, *73*, 125–131.

Bordoff, J. et Noel, P. (2008). Pay-as-you-drive auto insurance : A simple way to reduce driving-related harms and increase equity. *Hamilton Project Discussion Paper*.

Boucher, J.-P. et Denuit, M. (2007). Duration dependence models for claim counts. *Blätter der DGVFM*, *28*(1), 29–45.

Boucher, J.-P., Pérez-Marín, A. M. et Santolino, M. (2013). Pay-as-you-drive insurance : the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles, 3ª Época*, *19*, 135–154.

Butler, P., Butler, T. et Williams, L. L. (1988). *Sex-Divided Mileage, Accident, and Insurance Cost Data Show That Auto Insurers Overcharge Most Women*. National Assoc. of Insurance Commissioners.

Buxbaum, J. N. (2006). *Mileage-Based User Fee Demonstration Project : Pay-As-You-Drive Experimental Findings*. Rapport technique.

Clark, M. (2013). Generalized additive models : getting started with additive models in r. Dernier accès en décembre 2015. Récupéré de `https://www3.nd.edu/~mclark19/learn/GAMS.pdf/`

Green, P. et Silverman, B. (1993). *Nonparametric Regression and Generalized Linear Models : A roughness penalty approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. IMA Volumes in Mathematics and Its Applications. Springer.

Gu, C. et Kim, Y.-J. (2002). Penalized likelihood regression : general formulation and efficient approximation. *Canadian Journal of Statistics*, *30*(4), 619–628.

Hastie, T. et Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 297–310.

Hastie, T. et Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Iqbal, M. U. et Lim, S. (2006). A privacy preserving gps-based pay-as-you-drive insurance scheme. Dans *Symposium on GPS/GNSS (IGNSS2006)*.

Jun, J., Ogle, J. et Guensler, R. (2007). Relationships between crash involvement and temporal-spatial driving behavior activity patterns : use of data for vehicles with global positioning systems. *Transportation Research Record : Journal of the Transportation Research Board*, (2019), 246–255.

Litman, T. (2005). Pay-as-you-drive pricing and insurance regulatory objectives. *Journal of Insurance Regulation*, *23*(3), 35.

Litman, T. (2011). Pay-as-you-drive insurance : recommendations for implementation. *Victoria Transport Policy Institute (www.vtpi.org)*.

Lourens, P. F., Vissers, J. A. et Jessurun, M. (1999). Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. *Accident Analysis & Prevention*, *31*(5), 593–597.

Nelder, J. A. et Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, *135*, 370–384.

Ohlsson, E. et Johansson, B. (2010). *Non-life insurance pricing with generalized linear models*. Springer Science & Business Media.

Poisson, S. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile : précédées des règles générales du calcul des probabilités*. Bachelier.

Rigby, R. A. et Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, *54*(3), 507–554.

Ruppert, D., Wand, M. P. et Carroll, R. J. (2003). *Semiparametric regression*. Numéro 12. Cambridge university press.

Schwartz, A. (2004). Evaluating insurer compliance with proposition 103 : Promoting fair and affordable auto insurance pricing in california. *University of California-Berkeley*.

Sousanis, J. (2011). World vehicle population tops 1 billion units. Dernier accès en décembre 2015. Récupéré de `http://wardsauto.com/news-analysis/world-vehicle-population-tops-1-billion-units/`

Vickrey, W. (1968). Automobile accidents, tort law, externalities, and insurance : an economist's critique. *Law and Contemporary Problems*, 464–487.

Wood, S. (2006). *Generalized Additive Models : An Introduction with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.