Université Claude Bernard – Lyon 1

INSTITUT DE SCIENCE FINANCIERE ET D'ASSURANCES

Mémoire présenté
devant l'Institut de Science Financière et d'Assurances
pour l'obtention du diplôme d'Actuaire de l'Université de Lyon

le 19 octobre 2011

Par : Déborah BARGAIN

Titre: Modélisation du taux d'absentéisme en entreprise

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Membre du jury de l'Institut des Actuaires

Catherine PIGEON

Membres du jury I.S.F.A.

M. Jean Claude AUGROS
M. Alexis BIENVENÜE
M. Areski COUSIN
Mme Diana DOROBANTU
Mme Anne EYRAUD-LOISEL
M. Nicolas LEBOISNE
M. Stéphane LOISEL
Mlle Esterina MASIELLO
Mme Véronique MAUME-DESCHAMPS
M. Frédéric PLANCHET
M. François QUITTARD-PINON
Mme Béatrice REY-FOURNIER
M. Pierre RIBEREAU
M. Christian-Yann ROBERT
M. Didier RULLIERE
M. Pierre THEROND

Secrétariat

Mme Marie-Claude MOUCHON

Bibliothèque :

Mme Michèle SONNIER

Entreprise :

WINTER & Associés

Directeur de mémoire en entreprise :
Julien JACQUEMIN

Invité :

Autorisation de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Signature du candidat

50 Avenue Tony Garnier
69366 Lyon Cedex 07

MOTS CLES

Absentéisme en entreprise, Modèles Linéaires Généralisés, Taux d'absentéisme, Sélection de variables, Variables explicatives.

RESUME

Historiquement, le phénomène de l'absentéisme est en hausse constante, globalement depuis l'implantation progressive des 35 heures. Il est donc important de comprendre le phénomène, de repérer objectivement les populations et les situations à risque pour envisager des actions positives de réduction des coûts du phénomène.

Après avoir rappelé les caractéristiques de l'absentéisme en entreprise en France, en présentant notamment les coûts que cela engendre, les causes de l'absentéisme ainsi que les moyens d'y remédier, ce mémoire propose de modéliser l'absentéisme par un Modèle Linéaire Généralisé, au sein d'une entreprise de plus de 1 000 salariés dans un premier temps, puis sur près de 200 entreprises réparties sur toute la France dans une seconde partie.

Il s'agit donc de rappeler le principe de ce type de modèle, puis de le mettre en application pour déterminer le taux d'absentéisme d'un salarié puis d'une entreprise. Ce mémoire présente d'abord les caractéristiques des salariés et entreprises en question puis décrit toutes les étapes du processus, du choix des modèles et des variables à la détermination du taux d'absentéisme.

Cette étude a posteriori nous permettra de calculer un taux d'absentéisme pour une entreprise en fonction de ces caractéristiques et ainsi permettre la mise en place des actions de gestion et de réduction de l'absentéisme adaptées.

KEY WORDS

Absenteeism, Generalized Linear Models, Absenteeism rate, Selection of variables, Explanatory variables.

ABSTRACT

Historically, the phenomenon of absenteeism is rising steadily, overall since the gradual implementation of 35 hours. Therefore, it is important to understand the phenomenon, identify objectively the populations at risk and consider positive action to reduce costs of the phenomenon.

After reviewing the characteristics of absenteeism in companies in France, particularly with the costs that this entails, causes of absenteeism and how to remedy it, this paper proposes to model absenteeism by a Generalized Linear Model in a company with more than 1 000 employees initially, then about 200 companies spread across France in the second part.

It is therefore to recall the principle of this type of model, then implement it to determine the rate of absenteeism of an employee and a company. This thesis first presents the characteristics of workers and enterprises in question and then describes all phases of the process, the choice of models and variables to determine the rate of absenteeism.

This study retrospectively will allow us to calculate a rate of absenteeism for a company based on these characteristics and allow the implementation of management actions and reduced absenteeism adapted.

REMERCIEMENTS

Mes premiers remerciements s'adressent à Julien JACQUEMIN pour m'avoir accueilli au sein de son équipe au Cabinet WINTER & Associés, et m'avoir confié ce sujet de mémoire qui m'a appris et intéressé énormément.

Je remercie également Sidonie G. de m'avoir fait profiter de ses connaissances concernant l'Absentéisme.

Je remercie très vivement Frédéric PLANCHET pour sa grande disponibilité et ses nombreux conseils tant pratiques que théoriques. Il a été une aide indispensable dans mes travaux.

SOMMAIRE

INTRODUCTION

Les entreprises sont confrontées à une problématique à laquelle elles portent de plus en plus d'attention, l'absentéisme. L'effet des 35 heures n'a pas été celui escompté, au contraire, les absences au travail n'ont cessées d'augmenter suite à cette diminution du temps de travail[1]. Cela a eu pour conséquence une augmentation de la pression au travail car il est souvent demandé de travailler autant en 35 heures qu'en 39 heures, mais également la volonté grandissante de laisser plus de place aux loisirs.

L'absentéisme représente un enjeu pour les entreprises en termes d'organisation et de coût. En effet, une personne absente ne peut tenir son poste et cause une désorganisation plus ou moins importante dans l'entreprise. De plus, les coûts engendrés par l'absence sont nombreux, tels que liés à une baisse de productivité ou encore du fait du remplacement du salarié absent par une personne extérieure.

Les entreprises peuvent être accompagnées dans leur gestion et réduction de l'absentéisme. Suite à la capitalisation grandissante des données d'absentéisme, elles peuvent évaluer leur niveau d'absentéisme par rapport à un secteur d'activité, une zone géographique et se comparent à la moyenne nationale. L'intérêt est alors de savoir quelles variables influencent sur les absences et ainsi avoir une démarche adaptée de réduction de l'absentéisme.

Le Modèle Linéaire Généralisé est une technique permettant de connaitre l'influence de variables explicatives sur une variable à expliquer. Nous proposons dans ce mémoire d'utiliser cette technique pour décrire les caractéristiques des absences des gens absents et modéliser le taux d'absentéisme en entreprise.

Dans une première partie, nous fixerons le contexte en présentant les caractéristiques de l'absentéisme en France. Nous rappellerons ensuite la théorie mathématique nécessaire pour la mise en place d'un Modèle Linéaire Généralisé. Nous appliquerons cette théorie au cas d'une entreprise dans un premier temps puis à l'ensemble des entreprises participantes à un sondage réalisé sur le sujet en 2008 et 2009.

[1] Article « Absentéisme : La France championne d'Europe », Le Figaro, 5 mai 2008

PARTIE 1 : CONTEXTE DE L'ETUDE

Dans cette partie, nous allons dans un premier temps présenter ce qu'est l'absentéisme, et comment on le mesure. Nous verrons ensuite ses origines, ses causes, qui sont multiples, ainsi que les coûts que cela engendre, que ce soit pour les employeurs, les assurances ou encore l'économie nationale. Enfin, nous verrons les moyens permettant de réduire ce taux notamment au niveau des conditions de travail et du management.

CHAPITRE 1. QU'EST-CE QUE L'ABSENTEISME ?

Dans ce chapitre, on se propose de définir ce qu'est l'absentéisme et comment on le mesure.

Section 1.1 Définitions

Selon l'Encyclopédie Universalis, l'absentéisme représente le fait de ne pas être présent, d'être absent d'une fonction, d'un travail. Définir l'absentéisme comme le « non-présentéisme » au travail revient à intégrer les absences liées aux repos hebdomadaires, aux congés annuels et aux autorisations diverses ainsi que toutes les absences pour maladie et maternité. Si ces différents types d'absence se manifestent toutes par la non présence du travailleur, elles n'ont pas la même signification sociale. Toutes ne doivent pas être prises en compte dans l'étude de ce phénomène.

A contrario, la définition donnée par Le Petit Robert n'est pas non plus satisfaisante car trop restrictive. Le sens donné par cet ouvrage à l'absentéisme est le suivant, il s'agit « de l'absence d'un salarié de son lieu de travail, non justifié par un motif légal ». L'application stricte de ce critère a pour effet de sortir du champ de l'absentéisme toutes les causes d'arrêt de travail pour motifs légaux puisqu'elles sont justifiables.

La définition la plus complète de ce phénomène complexe qu'est l'absentéisme pourrait être celle-ci : l'absentéisme correspond, sur une période donnée, à la somme des absences physiques individuelles non comprises dans le cycle de travail, hormis les périodes de formation et de représentation syndicale, qui se traduisent par la non réalisation des tâches normalement attendues. Faire référence à une notion de période laisse entendre que l'absentéisme est un phénomène qui se reconduit dans le temps.

La proposition de définition de LETEURTRE [1991][2] est intéressante car elle cible le phénomène sur les absences qui sont susceptibles de gêner le fonctionnement de l'institution et pour lesquelles la mise en œuvre d'un plan d'action est envisageable. Selon cet auteur le terme absentéisme recouvrirait « les absences pouvant, sans que la certitude ne soit jamais faite, révéler un comportement de fuite devant le travail résultant, soit d'un rejet de celui-ci (cause objective liée au travail où à son organisation), soit d'un arbitrage entre obligations ressenties de sens contraire (exigence morale d'aller au travail ou exigence morale affective de garder son enfant malade) ou encore de désirs de sens contraires (désir de retrouver le groupe et désir de se retrouver dans un univers autre que celui du travail) ».

[2] H. Leteurtre « Audit de l'absentéisme du personnel hospitalier » Berger-Levrault 1991

Dans les organisations, la définition privilégiée s'appuie souvent sur des contraintes administratives. En général, on y sépare deux groupes d'absences : les absences justifiées ou autorisées et les absences non justifiées ou non autorisées. Cette division s'appuie sur le principe que les absences non justifiées sont celles sur lesquelles on doit faire les principaux efforts car du fait de leur importance, peuvent tout à la fois être préjudiciable au fonctionnement régulier d'une structure et symptôme de malaise social.

La notion de taux d'absentéisme reste encore controversée dans les organisations, ceci dû au fait que les chercheurs mesurent encore les absences de plusieurs manières. Plusieurs subdivisent les absences en deux types : les absences volontaires et les absences involontaires. L'importance de cette distinction tient au fait que la plupart des absences volontaires sont évitables. Selon certains auteurs, ces absences durent en général moins de trois jours consécutifs et sont potentiellement contrôlables. Les absences involontaires sont celles qui résultent d'une maladie invalidante, d'un deuil ou d'une maladie grave d'un membre de la famille ; on les qualifie d'inévitables et leur durée va souvent au-delà de trois jours consécutifs.

Section 1.2 Mesure de l'absentéisme

Le taux d'absentéisme pour une entreprise, rapport des durées d'absences d'un groupe au total des durées normalement travaillées par le groupe, est l'indicateur le plus pertinent, bien qu'approximatif, puisqu'il n'individualise pas la mesure de l'absence, ni le nombre de jours d'arrêt par travailleur arrêté.

Il existe un taux d'absentéisme dit « structurel », incompressible. Il est généralement estimé entre 4 et 6 %, mais il est nécessaire d'étalonner ce chiffre selon l'entreprise, en fonction de sa taille, de sa composition (sexe et âge des salariés notamment) et de son activité (service ou production, pénibilité du travail, etc.). L'absentéisme ne devient un problème pour une structure qu'à partir d'un certain niveau.
Il est important de noter que le taux d'absentéisme connaît des variations au sein d'une organisation en fonction des saisons, de la demande de prestations, de la politique managériale en matière d'absentéisme et d'autres évènements tels que les périodes de fête. Les périodes d'activités intenses au sein des organisations pourraient également occasionner plus d'absence chez les employés, à cause de la surcharge de travail. Toutefois, ce phénomène peut être contrebalancé par des contrôles plus serrés des absences par l'administration.

A cause du manque de consensus entre les auteurs sur la définition du mot « absence », on recommande GALLOIS [2005][3] tout simplement d'être clair sur les catégories d'absence qui composent la mesure des taux d'absentéisme. A ce propos, Gallois propose que soient retenus au numérateur de ce ratio : les absences pour maladie, pour

[3] P. Gallois « L'absentéisme : Comprendre et agir » LIAISONS 2005

accident de travail, les absences de maternité et les absences non autorisées à l'instar des retards. Ceci suppose une exclusion du numérateur, des absences prévisibles telles que les congés payés, les congés parentaux, la formation professionnelle ainsi que des absences qui ont un caractère exceptionnel (heures de débrayage ou de grèves).

L'absentéisme en milieu de travail peut être perçu comme un simple coût que l'organisation cherche à minimiser ou alors comme un indicateur pertinent de mesure du climat social (GALLOIS [2005]). En raison de cette dichotomie, les éléments retenus dans le décompte des absences que l'organisation aura enregistrées au cours de la période considérée, peuvent varier suivant que ces données sont au service d'une analyse des coûts ou au service d'une étude des ressources humaines.

CHAPITRE 2. COUTS DE L'ABSENTEISME

Les différentes parties, gouvernements, employeurs, employés, compagnies d'assurances et société en tant que telle, portent chacune les coûts liés à l'absentéisme et à la santé. Le travailleur (ainsi que les personnes à sa charge) voit souvent son revenu se réduire en conséquences des absences liées à la santé, en particulier lorsque celle-ci se prolongent. Il peut devoir faire face à des dépenses supplémentaires, par exemple pour payer des services ou des équipements médicaux, et souffrir d'une perte de bien-être sous forme physique ou morale. De plus, l'absentéisme répété ou prolongé peut entraîner la perte de l'emploi ou perturber les relations avec les collègues ou les supérieurs.

Les employeurs sont gênés par le caractère imprévisible de ces absences, qui obligent à des ajustements de plannings ou à la prise de mesures pour remplacer le travailleur manquant. De plus, l'absentéisme au travail augmente les frais de la société (indemnités journalières, versements en sus des règlementations, productivité perdue, qualité inférieure, etc.) d'où un effet négatif sur sa position concurrentielle. Il est généralement estimé que 1 % de taux d'absentéisme coûte entre 1 % et 4 % de la masse salariale (MARTORY, [2009]). Ce qui permet d'établir le coût de l'absentéisme pour la collectivité chaque année à environ 25 milliards d'euros[4].

Les compagnies d'assurance garantissent souvent à la fois le risque d'absentéisme et la santé des travailleurs et de leurs familles. Elles doivent généralement payer les indemnités dues à l'absence et les coûts induits par la santé des employés.

L'absentéisme au travail nuit également à l'économie nationale puisqu'il débouche sur une perte de production potentielle imputable à la réduction de la main-d'œuvre disponible, et sur un accroissement des coûts des traitements médicaux et des dépenses de sécurité sociale. Les gouvernements ont donc intérêt à entretenir un bas niveau d'absentéisme et à limiter les dépenses de santé et de sécurité sociale, les coûts induits par les invalidités et les retraites précoces étant élevés. Il importe pour la société que ses membres puissent travailler dans un milieu sain jusqu'à l'âge de la retraite.

[4] Article « L'absentéisme en entreprise », Chambre de commerce et d'industrie de Paris, Val-de-Marne, N°49-Janvier 2009

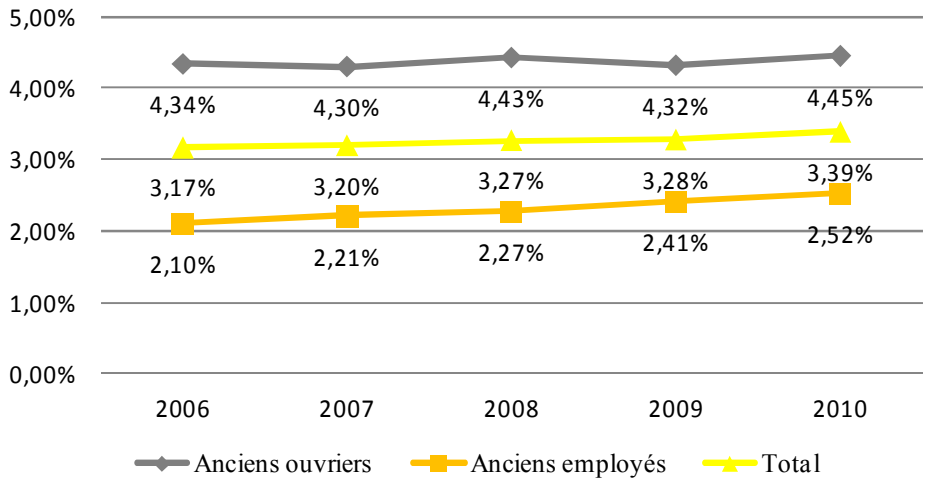
CHAPITRE 3. CAUSES DE L'ABSENTEISME

Une fois les absences identifiées et leurs proportions connues, il faut rechercher quelles en sont les causes, seul moyen pour mettre en place des mesures qui permettront de les limiter. Selon les entreprises et leurs spécificités, les causes varient de manière importante.

Historiquement, le phénomène de l'absentéisme est en hausse constante, globalement depuis l'implantation progressive des 35 heures (Gouvernement Jospin et Loi Aubry). En effet, les salariés doivent faire en 35 heures ce qu'ils faisaient en 39 heures afin de rester dans les objectifs au niveau de la compétitivité des entreprises. Ce qui implique plus de pression, de stress. On doit faire plus, plus vite, avec toujours plus de qualité, toujours plus de services clients…et pour un coût de production de plus en plus bas.

Le graphique suivant présente l'évolution du taux d'absentéisme entre 2006 et 2010 selon le statut. On constate globalement une légère augmentation du taux d'absentéisme sur la période observée. En effet, le taux global est passé de 3,2% en 2006 à 3,4% en 2010[5].

Evolution de l'absentéisme depuis 2006



Les origines de l'absentéisme sont nombreuses : il faut distinguer l'absentéisme lié à la santé individuelle, celui lié à l'environnement professionnel ou l'environnement familial, mais aussi l'absentéisme « de confort » (abusif). Les accidents du travail ne constituent pas les seules causes d'absentéisme, les maladies professionnelles en sont un des facteurs récurrents. Le stress professionnel progresse par ailleurs comme l'un des risques d'absentéisme majeurs auquel sont confrontées les entreprises. Le vieillissement de la population a également un impact sur le taux d'absentéisme. On peut constater des causes d'absentéisme liées à l'environnement familial, et principalement pour des problèmes d'organisation (garde des enfants, gestion des tâches quotidiennes, déplacement

[5] Pour plus de précisions voir le site http://www.observatoire-absenteisme.public.lu/chiffres_cles/Absenteisme_maladie_2010.pdf

professionnel…). Enfin, l'absentéisme dit « de confort » ou abusif n'est pas la cause la plus importante mais ne peut être évacué.

La mauvaise santé est la raison principale pour laquelle les travailleurs s'absentent de leur travail. Cependant, mauvaise santé ne rime pas nécessairement avec absence du travail. Bien que les employés souffrant de problèmes de santé soient en général plus souvent et plus longtemps absents du travail que leurs collègues « en bonne santé », certains d'entre eux ne s'absentent guère plus que leurs collègues. De plus, les actions visant à diminuer l'absentéisme sur le lieu de travail n'ont pas toutes un effet sur la santé. Une entreprise peut par exemple tenter de restreindre l'absentéisme au travail en durcissant les procédés de vérification et en intensifiant les contrôles sur les employés absents.

La problématique de l'absentéisme n'est pas la même pour toutes les entreprises. Les facteurs explicatifs mesurables sont nombreux et selon la structure de l'entreprise sont plus ou moins présents. On peut regrouper ces facteurs en trois catégories :

- Les facteurs inhérents à l'entreprise et à son environnement : l'effectif, le secteur d'activité et le secteur géographique.
- Les facteurs inhérents au salarié qui affectent son comportement : le sexe, l'âge, l'ancienneté, la catégorie socioprofessionnelle, le niveau de qualification et la situation familiale.
- Les facteurs externes, sur lesquels l'entreprise n'a pas de pouvoir : les jours de la semaine, les épidémies, évènements sportifs et autres.

Arrêts maladies, congés sabbatiques, petits retards répétés…les causes d'absentéisme sont multiples et on préférera toujours un traitement différencié, au cas par cas, à des mesures générales. On ne traitera pas de la même façon le salarié régulièrement absent qui subit « discrètement » des séances de chimiothérapie pour soigner son cancer et celui qui prend chaque année ses quatre jours pour faire du ski …

L'absentéisme reflète en partie la pénibilité du travail ou le climat social : les ouvriers (qualifiés ou non) se sont toujours absentés davantage que les cadres, tous secteurs confondus. Quand la pression psychologique exercée par les clients ou par les collègues engendre une dose de stress et de souffrance telle que les arrêts maladie pour dépression se multiplient, il y a à l'évidence une responsabilité du management.
L'absentéisme est un indicateur « climatique ». S'il augmente subitement dans tel service ou dans tel atelier, il faut se demander avec quoi cette hausse peut être corrélée : l'arrivée d'un nouveau responsable, la mise en œuvre d'une nouvelle organisation…

Pour certains évènements familiaux, le droit du travail prévoit des autorisations d'absence, notamment pour le mariage, chaque naissance, le décès d'un enfant, le mariage d'un enfant, etc. Par ailleurs, certaines conventions collectives prévoient des autorisations d'absence pour la rentrée scolaire. Il est donc à noter que pour les entreprises dont la convention collective prévoit des autorisations d'absence non prévu par le code du travail, le nombre de jours d'absence sera plus important.

Pour résumer, on peut donc les regrouper en trois catégories principales, cette liste étant loin d'être exhaustive :

- Les conditions de travail et d'organisation : pénibilité des conditions de travail, amplitude trop élevée des journées de travail, problèmes liés à la répartition des tâches entre les salariés, changements d'horaires fréquents, etc. ;
- L'implication des salariés et l'ambiance de travail : absentéisme de mécontentement suite à une décision défavorable de l'employeur, inquiétude de certains salariés liée aux changements de technologie, absence de responsabilisation du personnel de production liée au remplacement trop facile des absents, trop forte pression dans le travail, absence d'opportunité d'évolution de carrière, vieillissement de la population et apparition de pathologies plus longue à soigner, etc. ;
- L'influence du management : management trop souple face aux abus et trop permissif sur la sanction des retards, manque d'implication des chefs d'équipes dans la lutte contre l'absentéisme, absence de valorisation des tâches, etc.

CHAPITRE 4. LES SOLUTIONS

Quelle que soit la cause de l'absentéisme, il y a toujours des raisons à une absence. Il devient pertinent d'agir au cas par cas. Et de réintégrer la dimension individuelle de l'absentéisme dans l'organisation collective. C'est le moyen de crever l'abcès en cas de malaise, d'oser parler d'un problème qu'un salarié n'ose pas aborder…Il est important de débanaliser l'absence dans l'entreprise mais aussi de déculpabiliser. Ainsi, pour parvenir à lutter efficacement contre l'absentéisme, il est nécessaire de procéder à une analyse préalable de la situation de l'entreprise. Il faut réfléchir aux causes du phénomène pour espérer ensuite trouver les moyens d'action les plus adaptés.

Section 4.1. Les pistes à explorer

Lorsque l'absentéisme est lié à un problème de santé, il est nécessaire de prévoir, d'anticiper. Notamment pour ce qui concerne les infections saisonnières ; ainsi certaines entreprises ont mis en œuvre une politique de financement du vaccin contre la grippe, pour que leurs salariés échappent à ce virus saisonnier. Dans le cas de maladies chroniques ou de pathologies plus lourdes, il faut aider les personnes et dialoguer avec elles : autrement dit, mieux connaître ses salariés. Pour ce qui concerne la gestion du stress professionnel, l'instauration d'un dialogue constructif avec le médecin du travail doit favoriser une meilleure compréhension des pressions éprouvées par les salariés.

Lorsque c'est lié à l'environnement professionnel interne à l'entreprise il est nécessaire de travailler à partir d'un axe directeur : débanaliser l'absentéisme.
L'amélioration des conditions de travail reste un levier majeur de réduction des absences. Mais n'oublions pas la perspective d'un véritable plan de carrière qui est une source de motivation, tout particulièrement pour ceux qui viennent d'être embauchés. Concernant le « présentéisme » : le salarié est payé en fonction de ses heures de travail effectives. Ainsi les primes d'assiduité proposées par certaines sociétés, mais qui n'ont pas fait réellement leurs preuves : elles coûtent cher à l'entreprise et pénalisent les salariés contraints de s'absenter pour des raisons justifiées. L'accompagnement et le suivi des salariés restent donc les choix pédagogiques à privilégier. Pour ce qui est de l'absentéisme « de confort », l'entreprise peut aussi avoir recours à une solution de contrôle, comme la contre visite médicale.
La loi 78-49 de 1978 permet à tout employeur de faire effectuer un contrôle médical de ses salariés en arrêt de travail, dès lors qu'il complète l'indemnisation de la sécurité sociale. Toutes les entreprises, quelle que soit leur taille, sont concernées par cette loi. En effet, elles ont l'obligation de compléter l'indemnisation de la Caisse Primaire d'Assurance Maladie (CPAM) en cas d'arrêt de travail, dès lors que le salarié a plus de trois ans d'ancienneté. Certaines conventions collectives instaurent même le complément obligatoire dès la 2ème, voire la 1ère année.

Le contrôle est effectué au domicile du salarié ou au cabinet du médecin (selon le régime de sorties autorisées dont bénéficie le salarié). Le salarié a l'obligation de se soumettre au contrôle du médecin mandaté, qui va juger de la validité de l'arrêt au moment du contrôle, de la date de reprise de travail et de l'utilité de prévoir une prolongation.
Dès lors que le médecin juge que l'état de santé du salarié permet la reprise du travail, l'employeur peut exiger qu'il reprenne son travail. Si celui-ci refuse, l'employeur est en droit de lui supprimer le complément de salaire, à compter de la date du constat fait par le médecin contrôleur.
De la même façon, l'employeur est en droit de supprimer le complément de salaire à tout salarié absent lors du contrôle, qui ne s'est pas rendu à la convocation, absent lors du contrôle ou dont l'adresse, incomplète, n'a pas permis au médecin d'effectuer la contre-visite médicale. Là encore, l'arrêt du complément de salaire se fait à partir de la date du rapport du médecin.
Comme la loi le permet désormais, l'employeur peut transmettre le résultat du contrôle au service médical de la CPAM, qui peut décider de supprimer son indemnisation ou encore réaliser lui aussi un contrôle.

En résumé, les pistes à explorer sont les suivantes :

- **Les conditions de travail** : utiliser l'entretien annuel pour évoquer les conditions de travail avec le salarié, réaliser des audits sécurité et développer des formations pour réduire les accidents du travail, etc. ;
- **L'organisation du temps de travail** : veiller à la répartition des tâches entre les salariés, autoriser le fractionnement des jours de réduction du temps de travail (RTT) en demi-journées, effectuer un suivi rigoureux des absences durant les périodes de forte activité, etc. ;
- **L'ambiance de travail et l'implication des salariés** : informer davantage pour responsabiliser et sensibiliser les salariés (l'absentéisme est l'affaire de tous et influe sur la charge de travail de chacun), vérifier que le report de charge de travail d'un salarié absent ne soit pas fait systématiquement au détriment des mêmes salariés, etc. ;
- **Le management** : former et encadrer davantage autour de cette question (salariés mais également hiérarchie intermédiaire), pratiquer l'entretien de retour après maladie (pour faciliter la réintégration du salarié, mais aussi faire remonter certaines problématiques liées au poste de travail, aux relations avec les collègues et qui expliqueraient indirectement les absences du salarié), mettre en place des contrats d'objectifs, sensibiliser en amont les délégués du personnel, etc.

Section 4.2. Ce qui donne peu de résultats

- **La prime de présentéisme** : Appelée également prime d'assiduité, cette prime est versée aux salariés pour les encourager à ne pas s'absenter. Elle présente souvent des effets pervers. Ainsi, dans de nombreux cas, le fait d'avoir quelques absences fait perdre au salarié le bénéfice de la prime. Perdue pour perdue, certains salariés décident

alors de s'absenter encore davantage, allant ainsi à l'encontre de l'objectif recherché. De plus, ce type de prime doit être particulièrement bien conçu afin d'éviter tout risque juridique de discrimination par rapport à l'état de santé. Les évènements à l'origine de l'absence devront entraîner de la même manière une réduction ou une perte de la prime.

- **Le recours massif aux contre-visites** : Il est possible de faire appel à un organisme spécialisé pour vérifier si l'arrêt de travail d'un salarié est médicalement justifié. Si le salarié n'est pas à son domicile ou si le médecin estime que celui-ci est en état de reprendre le travail, l'employeur est alors en droit de cesser tout maintien de salaire, à compter du jour du contrôle et jusqu'à la fin de l'arrêt. Cependant, dans la pratique, on constate que, dans la majeure partie des cas, le médecin contrôleur confirme le diagnostic initial de son confrère. Par ailleurs, le contrôle médical demeure onéreux pour l'entreprise. Enfin, si le médecin estime l'arrêt de travail sans fondement, l'employeur n'a pas pour autant le droit de prendre une sanction à l'encontre du salarié. C'est pourquoi il peut être intéressant d'utiliser cette technique, mais seulement à bon escient et afin de montrer que l'entreprise se soucie de l'absentéisme. Cette mesure dissuasive ne constitue pas un remède miracle et doit être utilisée en parallèle à d'autres actions.

CHAPITRE 5. QUELQUES DONNEES

Une enquête réalisée en juillet 2008, par un Cabinet de conseil travaillant sur ce sujet, permet de dégager que le secteur des services, les PME et l'Est de la France affichent les taux d'absentéisme les plus forts. Si le taux d'absentéisme global est stable, cette étude révèle une augmentation des absences liées aux accidents de travail et maladies professionnelles. Un résultat qu'il convient de lier également à la reconnaissance d'un plus grand nombre de maladies professionnelles et d'une meilleure connaissance de la législation de la part des collaborateurs. Selon les Directeurs des Ressources Humaines interrogés, le contexte économique des organisations et le climat social sont également des facteurs favorisant les arrêts maladies (impactant sur l'absentéisme), directement en lien avec la souffrance au travail. En effet, les organisations dont la situation économique s'est dégradée au cours de ces dernières années ont des taux d'absentéisme élevés.
Si l'on retenait le seul critère du coût de maintien de salaire, 1 % d'absentéisme représente selon les organisations de 0,10 % à 1,68 % de la masse salariale. Sans parler des coûts supplémentaires, dits de régulation (recours à des équipes de remplacement, prise en charge des délais de carence…) et de perturbations (retards, problèmes qualité…) difficilement mesurables. Aussi, les entreprises se basent sur l'équation « 1 % d'absentéisme équivaut à 1 % de la masse salariale ». Or cette équation s'avère « simpliste » car elle ne prend pas en compte de nombreux paramètres décisifs sur l'absentéisme comme les conditions de maintien de salaire, le choix de remplacement des collaborateurs absents ou bien la gestion administrative des absences.

D'autre part, la tendance générale observée ces dernières années concernant les périodes d'absentéisme dans les entreprises est qu'elles sont plus courtes mais de plus en plus nombreuses. Or, les petites absences sont les plus pénalisantes car les plus difficiles à prévoir et à remplacer.

PARTIE 2 : UTILISATION DES MODELES LINEAIRES GENERALISES

Nous souhaitons modéliser le taux d'absentéisme en entreprise par un modèle linéaire généralisé dans le but de prédire un taux d'absentéisme propre à chaque salarié au sein d'une entreprise dans un premier temps, puis propre à chaque entreprise participante aux sondages sur le sujet réalisés en 2008 et 2009. Ces questionnaires ont été lancés pour la première fois en France par un Cabinet de conseil travaillant sur le sujet afin de permettre une capitalisation significative de données d'absentéisme et de participer à la sensibilisation des acteurs aux enjeux sociaux et économiques liés à ce phénomène. Nous proposons dans cette partie de rappeler les principaux aspects théoriques et techniques de cette modélisation.

CHAPITRE 1. LE MODELE LINEAIRE GENERALISE

Dans une première partie nous établirons ce qu'est un modèle linéaire généralisé avec ses hypothèses et les principales distributions de la famille exponentielle. Mais tout d'abord un petit retour en arrière concernant l'apparition de ces modèles.

Section 1.1. Un peu d'histoire

Les modèles linéaires généralisés ont fait leur apparition en 1972 (Nelder et Wedderburn [1972]). Ils sont adaptés à de nombreuses problématiques et sont d'utilisation courante dans le domaine de la statistique et de l'actuariat.
La théorie des modèles linéaires généralisés bénéficie d'un avantage par rapport aux modèles linéaires classiques : le caractère normal de la variable à expliquer Y n'est plus imposé, seule l'appartenance à une famille exponentielle est indispensable.

Section 1.2. Distributions de la famille exponentielle naturelle

Les distributions de la famille exponentielle naturelle sont indispensables pour la mise en place d'un modèle linéaire généralisé. Nous allons donc présenter ici les principales distributions qui sont généralement utilisées.

1.2.1. Forme générale

Soit Y une variable aléatoire et y une observation de Y.
La loi de probabilité de Y appartient à la famille exponentielle naturelle si et seulement si sa densité peut se mettre sous la forme :

$$f(y;\theta;\phi) = \exp\left\{\frac{\theta y - b\theta}{a(\phi)} + c(y;\phi)\right\}$$

où :

- a est une fonction non nulle définie sur R ;
- b est une fonction définie sur R, deux fois dérivable ;
- c est une fonction définie sur R^2 ;
- θ paramètre canonique ou paramètre de la moyenne ;
- φ paramètre de dispersion.

De nombreuses distributions classiques appartiennent à cette famille. Dans la famille exponentielle, l'espérance et la variance de la variable aléatoire Y s'expriment aisément en fonction des différents paramètres.

L'espérance de Y s'écrit alors $\mu = \mathrm{E}(Y) = b'(\theta)$

On introduit la fonction de lien canonique g_c telle que : $\theta = g_c(\mu)$

La variance de Y s'écrit : $Var(Y) = b''(\theta) * a(\varphi)$

La fonction variance se définit par : $b''(\theta) = V(\mu)$

Où les notations ' et '' indiquent les dérivées premières et secondes par rapport à θ

1.2.2. Exemples

Les lois normales, gamma, exponentielle, inverse gaussienne, Bernoulli, binomiale, ou de Poisson sont des lois appartenant à la famille des lois exponentielles.

1) Loi normale

Soit Y une variable aléatoire suivant une loi normale d'espérance μ et de variance σ². Y est à valeurs réelles.

Sa fonction de densité est :

$$f_{\mu,\sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(y-\mu)^2)}{2\sigma^2} \right\}$$

qui peut être mis sous la forme :

$$f_{\mu,\sigma}(y) = \exp\left\{ \frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)}{2} \right\}$$

ainsi la loi gaussienne appartient à la famille exponentielle naturelle avec :

- $\theta = \mu$
- $\varphi = \sigma^2$
- $a(\varphi) = \varphi$
- $b(\theta) = \theta^2/2$
- $c(y;\varphi) = -\frac{1}{2}\left[\frac{y^2}{\varphi} + \ln(2\pi\varphi) \right]$

2) Loi de Poisson

Soit Y une variable aléatoire suivant une loi de Poisson de paramètre λ. Y est à valeurs discrètes.

Sa fonction de densité est de la forme :

$$P(Y = y) = \exp(-\lambda)\frac{\lambda^y}{y!}$$

qui peut être mis sous la forme :

$$P(Y = y) = \exp(y \ln \lambda - \lambda - \ln(y!))$$

ainsi la loi de Poisson appartient à la famille exponentielle naturelle avec :

- $\theta = \ln(\lambda)$
- $\varphi = 1$
- $a(\varphi) = 1$
- $b(\theta) = \exp(\theta)$
- $c(y ; \varphi) = -\ln(y !)$

3) Loi Gamma

Soit Y une variable aléatoire suivant une loi Gamma de paramètres υ et $\frac{\mu}{\upsilon}$ (tous deux strictement positifs).
La densité s'écrit :

$$f(y) = \frac{1}{\Gamma(\upsilon)} \left(\frac{\mu}{\upsilon} \right)^{\upsilon} y^{\upsilon - 1} \exp(-\frac{\mu}{\upsilon} y)$$

où $\Gamma(x) = \int_0^{\infty} e^{-u} u^{x-1} du$

La fonction de densité peut également être mise sous la forme :

$$f(y) = \exp\left(\frac{-\frac{\mu}{\upsilon^2} y + \ln \frac{\mu}{\upsilon^2}}{\frac{1}{\upsilon}} + c\left(y, \upsilon\right) \right)$$

Ainsi la loi gamma appartient à la famille exponentielle naturelle avec :

- $\theta = -\frac{\mu}{\upsilon^2}$
- $\varphi = \frac{1}{\upsilon}$
- $a(\varphi) = \varphi$
- $b(\theta) = \ln \frac{-1}{\theta}$

Section 1.3. Hypothèses et principe du Modèle Linéaire Généralisé

1.3.1. Hypothèses

Le modèle linéaire généralisé part du même principe que celui du modèle linéaire simple. La différence est qu'au lieu de modéliser la variable à expliquer directement, c'est une fonction de l'espérance de cette variable (appelée fonction lien) qui est modélisée.

Une variable aléatoire Y relève du modèle linéaire généralisé si la loi de Y sachant $\{X_1 = x_1, ..., X_n = x_n\}$ est telle que :

- Il existe une fonction lien g strictement monotone de R dans R et des coefficients $(\alpha_0, ..., \alpha_n)$ tels que :

$$g\left(E[Y]\right) = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

- La loi de probabilité de Y doit appartenir à la famille exponentielle naturelle.

Les paramètres $(\alpha_0, ..., \alpha_n)$ sont les coefficients de régression et la quantité g(E[Y]) est le prédicateur linéaire.
L'estimation des paramètres de ces lois se fait par maximum de vraisemblance et le choix de la distribution pour le modèle linéaire généralisé est fait en observant les données.

1.3.2. Fonctions de lien

Les fonctions de lien classiques sont les suivantes :

- Fonction identité : g : z -> z
- Fonction logarithme : g : z -> ln(z)
- Fonction inverse : g : z -> 1/z
- Fonction logit : g : z -> $\ln\left(\frac{z}{1-z}\right)$
- Fonction probit : g : z -> φ(z)
  où φ est la fonction de répartition de la loi N(0,1)

Il y a deux cas où le choix n'est pas laissé à l'opérateur pour la fonction de lien. Ce sont les cas du modèle additif ou du modèle multiplicatif. Dans le cas d'un modèle additif, il faudra opter pour une fonction de lien identité et pour un modèle multiplicatif pour une fonction de lien logarithme.
En effet, si la fonction de lien est l'identité, on a :

$$E[Y] = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

ou encore : $E[Y] = \sum_{i=1}^{n} \beta_i$

avec :

- $\beta_0 = \alpha_0$
- $\beta_i = \alpha_i X_i$ pour tout i entre 1 et n

Il s'agit bien d'un modèle additif.
Si la fonction de lien est le logarithme, on a :

$$E[Y] = \exp\left(\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i\right)$$

ou encore :

$$E[Y] = \prod_{i=1}^{n} \beta_i$$

avec :

- $\beta_0 = \exp(\alpha_0)$
- $\beta_i = \exp(\alpha_i X_i)$ pour tout i entre 1 et n.

Il s'agit bien d'un modèle multiplicatif.
En choisissant le couple densité de la variable à expliquer/fonction de lien adéquat, le modèle linéaire généralisé correspond à des modèles statistiques classiques.

1.3.3. Estimation des paramètres

L'estimation des paramètres du modèle linéaire généralisé se fait par maximum de vraisemblance dont nous rappelons ici la méthode.
Rappelons tout d'abord que la vraisemblance est, par définition, un produit de fonctions de densité. Pour en déterminer le maximum, il suffit de déterminer la valeur du paramètre de la fonction qui l'annule tout en gardant la dérivée seconde négative. Pour des raisons pratiques, on préfère dériver le logarithme de la vraisemblance. On dérive ainsi une somme plutôt qu'un produit. De plus, comme la fonction logarithme est strictement croissante, maximiser le logarithme de la fonction équivaut à maximiser la fonction.
Concrètement, prenons l'exemple d'une loi exponentielle de paramètre θ et de densité f. Soit Y une variable aléatoire suivant cette loi et $(y_1,...,y_p)$ p observations de cette variable.

Rappelons que notre modèle est de la forme :

$$g(\mu) = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

avec μ = E[Y]
Nous avons de plus :

$$\theta = g_c(\mu)$$

On en déduit :

$$\theta = g_c\left(g^{-1}\left(\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i\right)\right)$$

Si on note la fonction de vraisemblance :

$$L(y_1, y_2, ..., y_p, \theta) = \prod_{i=1}^{n} f(y_i, \theta)$$

Cette fonction de vraisemblance s'écrit également :

$$L(y_1, y_2, ..., y_p, \alpha_0, \alpha_1, ..., \alpha_n) = \prod_{i=1}^{p} f\left(y_i, g_c\left(g^{-1}\left(\alpha_0 + \sum_{j=1}^{n} \alpha_j x_j\right)\right)\right)$$

L'équation à résoudre sera :

$$\frac{\partial \ln L(y_1, y_2, ..., y_p, \alpha_0, \alpha_1, ..., \alpha_n)}{\partial \alpha} = \sum_{i=1}^{p} \frac{\partial \ln f(y_i, \alpha_0, \alpha_1, ..., \alpha_n)}{\partial \alpha} = 0$$

On obtient ainsi $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, ..., \hat{\alpha}_n)$ les estimations des paramètres du modèle.
Ces estimations nous seront données par la suite par le logiciel R.

Section 1.4. Sélection de variables

1.4.1. Notions de robustesse et de précision d'un modèle

Notre problème considère une variable à expliquer et un ensemble de n variables explicatives. Il s'agit de déterminer un sous-ensemble de l'ensemble des variables explicatives réalisant un compromis entre le souhait que le modèle sélectionné contienne peu de paramètres et le souhait que ce modèle bénéficie d'un pouvoir explicatif suffisant.
Plus un modèle contient de variables explicatives, plus il est précis mais moins il est robuste. A l'inverse, moins un modèle à de variables explicatives, plus il est robuste mais moins il est précis. En effet, l'ajout d'une nouvelle variable explicative apporte des informations supplémentaires sur la variable à expliquer mais impose de fait une contrainte supplémentaire au modèle.

Il s'agira donc, dans notre étude, de déterminer la combinaison de variables explicatives qui permettra d'obtenir le meilleur compromis entre précision et robustesse.

1.4.2. Les méthodes du type stepwise

La méthode choisie pour rechercher la combinaison optimale est une méthode de type stepwise : *forward* ou *backward*.

Rappelons dans un premier temps les définitions du R² et de la déviance. Soient $(y_1, ..., y_n)$ n réalisations d'une variable aléatoire Y. $\hat{y}_i$ est la valeur estimée de la i^{ème} observation et $\overline{y}$ est la moyenne empirique. Alors, mathématiquement, la formule du coefficient de détermination R² est la suivante :

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

Il représente la proportion de variance totale expliquée par le modèle. Plus il est proche de 1, plus le modèle explique la variance totale.
En regardant le R² ajusté plutôt que le R², on obtient un lien entre les variables sélectionnés et le nombre de variables. Le meilleur compromis est alors la solution où le R² ajusté est maximum.

La formule de la déviance est la suivante :

$$D = \varphi \times D^*$$

où φ est le paramètre de dispersion et D* la déviance normalisée, définit comme suit :

$$D^* = -2\left[L(y,\mu) - L(y,\hat{\mu})\right]$$

avec L fonction de log-vraisemblance de la variable Y et $\hat{\mu}$ l'estimateur de l'espérance de cette variable.
Concrètement, elle représente l'écart entre la loi supposée pour la variable à expliquer et la loi qu'elle suit réellement.
Pour un modèle linéaire classique, la variable introduite dans le modèle est celle qui lui apporte le plus grand gain en R^2. Pour un modèle linéaire généralisé, c'est la variable qui entraîne la plus grande diminution de la déviance.
L'introduction des variables est stoppée à partir du moment où leur effet sur le modèle n'est plus significatif, c'est-à-dire que le gain en R^2 ou la perte en déviance sont jugés négligeables.

Dans la méthode *forward*, il s'agit de rechercher la variable la plus significative au sens du R^2 pour un modèle linéaire classique ou au sens de la déviance pour un modèle linéaire généralisé. Partant de ce modèle à un facteur, nous cherchons ensuite la variable qui, associée à la première explique le mieux la sinistralité et ainsi de suite.

Dans la méthode *backward*, il s'agit de démarrer avec le modèle complet (c'est-à-dire toutes les variables ayant un effet significatif sur le risque) puis de retirer la variable la moins significative, autrement dit celle dont l'élimination entraine la plus faible diminution de R^2 ou la plus faible augmentation de la déviance.

Les résultats des deux méthodes sont ensuite comparés pour déterminer l'ordre d'importance des variables explicatives dans le modèle complet, en tenant compte désormais de l'influence des variables déjà présentes dans le modèle.

1.4.3. Validité d'un modèle

I. Tests d'hypothèses

1) Test de signification d'un coefficient β_j

Sous l'hypothèse nulle, le coefficient β_j vaut 0, ce qui signifie que la variable explicative associée ne devrait pas être dans le modèle car elle n'influence pas la variable à expliquer.
Dans R, on utilisera la fonction summary pour tester la significativité des coefficients.

2) Test de Fisher global

Sous l'hypothèse nulle, $\beta_1 = \beta_2 = ... = \beta_p = 0$.
Dans R, on utilisera la fonction anova pour tester la significativité du modèle.

II. Analyse des résidus

Pour valider un modèle, une méthode classique consiste à analyser les résidus issus de ce modèle.
Dans un modèle linéaire classique, la variable à expliquer est décomposée en une partie explicative et une partie résiduelle supposée vérifier des propriétés spécifiques, notamment la normalité.
Dans un modèle linéaire généralisé, il n'y a pas ce type de décomposition. Il est donc plus difficile d'apprécier la validité des hypothèses formulées aussi bien sur le modèle lui-même que sur la loi des observations.

Trois types de résidus construits sur l'écart entre l'observé et l'estimé :

- Résidus normalisés ;
- Résidus standardisés ;
- Résidus studentisés.

La validité du modèle est jugée bonne (mais pas certaine) si les résidus observés se situent autour de l'axe des abcsisses et avec une variance constante, ils sont dits homoscédastiques.
D'autre part, concernant les résidus studentisés, on pourra dire qu'il n'y a pas de valeurs aberrantes, s'ils se situent dans l'intervalle [-2 ;2] en ordonnée.

De plus, le test de Durbin-Watson permet de tester l'indépendance des résidus. Sous l'hypothèse alternative, les résidus ne sont pas indépendants et suivent un processus auto-régressif d'ordre 1.

Conclusion : Nous allons, par la suite, sélectionner les variables selon une méthode de type stepwise, la méthode *backward*, afin d'obtenir le meilleur compromis entre robustesse et précision, puis tester les différentes familles de lois, pour ensuite estimer les paramètres du modèle par maximum de vraisemblance, et terminer en vérifiant si le modèle établi est valide ou non.

CHAPITRE 2. TESTS NON PARAMETRIQUES DE COMPARAISON D'ECHANTILLON

On s'intéresse dans ce mémoire à un phénomène de durée car notre variable à expliquer, le taux d'absentéisme, a pour particularité d'être générée par une variable aléatoire positive. D'autre part, nos données de durée utilisent des variables explicatives exogènes : l'âge, l'ancienneté, le sexe, la catégorie socioprofessionnelle, etc.

Dans notre étude, les tests non paramétriques de comparaison d'échantillons vont nous servir à comparer deux populations pour savoir si la loi de durée sous-jacente est la même. Si on a plusieurs critères, on va les comparer deux à deux et on peut alors ensuite regrouper les modalités qui conduisent à des lois de durée identiques. L'idée est de mettre en évidence une hétérogénéité puis, dans un second temps, de modéliser cette hétérogénéité.

Plus généralement, ces tests sont utiles lorsqu'on dispose de deux échantillons indépendants, éventuellement censurés et que l'on souhaite tester l'hypothèse nulle d'égalité des fonctions de survie dans les deux échantillons.

Section 2.1. Principe des tests de rang[6]

On dispose de deux séries d'observations, E_1 et E_2, de tailles respectives n_1 et n_2 ; on note $n = n_1 + n_2$; on range la séquence des valeurs observées $(x_1,...,x_n)$ par ordre croissant :

$$x_{(1)} < ... < x_{(n)}$$

Une statistique linéaire de rang a pour principe d'attribuer une pondération α_i à l'observation $x_{(i)}$ de rang i dans le classement commun des deux échantillons. On construit alors deux statistiques :

$$R_1 = \sum_{i \in E_1} \alpha_i \text{ et } R_2 = \sum_{i \in E_2} \alpha_i$$

Comme $R_1 + R_2 = \sum_{i=1}^{n} \alpha_i$, qui est connue et déterministe, il est indifférent de travailler sur l'une ou l'autre des statistiques ; en pratique on retient celle associée à l'échantillon le plus petit.
En choisissant $\alpha_i = i$, on obtient le test de Wilcoxon ; le test de Savage est quant à lui associé au choix $\alpha_i = 1 - \sum_{j=1}^{i} \frac{1}{n-j+1}$.
Le choix d'un test plutôt que d'un autre peut être guidé par la forme de l'alternative, en retenant le test (localement) le plus puissant pour une alternative donnée[7].

[6] F. Planchet, Modèle de durée, cours ISFA
[7] Pour des développements sur le sujet se reporter à CAPERAA et VAN CUTSEM [1988].

Section 2.2. Adaptation des tests de rang au cas censuré

L'adaptation des tests précédents au cas censuré conduit à introduire la suite ordonnée des instants de sorties observés (non censurés) dans l'échantillon commun, que l'on notera $t_1 < ... < t_N$. A chaque instant t_i, on désigne par d_{ij} le nombre de sorties et r_{ij} l'effectif sous risque dans le groupe j. L'effectif sous risque est calculé avant les sorties en t_i, de sorte que la population restante après t_i sont en nombre $r_{ij} - d_{ij}$. On peut synthétiser cela dans le tableau suivant :

Tableau n°1 : *Synthèse des tests de rang*

	Sorties en t_i	Restants après t_i	Total
Groupe n°1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Groupe n°2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
Ensemble	d_i	$r_i - d_i$	r_i

Sous l'hypothèse nulle d'égalité des distributions de survie dans les deux groupes, à chaque instant on doit avoir égalité des proportions de sorties dans les deux groupes, ce qui a pour conséquence l'indépendance des lignes et des colonnes dans le tableau ci-dessus. On est donc dans le cas d'un tableau de contingence à marges fixées, et alors la variable aléatoire d_{ij} est distribuée selon une loi hypergéométrique $H\left(r_i, d_i, \frac{r_{ij}}{r_i}\right)$ (puisqu'on compte le nombre de sorties dans le groupe n°j choisis parmi les d_i sorties totales, la probabilité d'appartenance au groupe n°j étant $p = \frac{r_{ij}}{r_i}$ et la taille de la population étant r_i).

On en conclut que l'espérance et la variance de d_{ij} : $E\left(d_{ij}\right) = d_i \frac{r_{ij}}{r_i}$ et $V\left(d_{ij}\right) = d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$.

Ces observations conduisent à construire des statistiques fondées sur des sommes pondérées des $d_{ij} - E(d_{ij})$, qui sont asymptotiquement gaussiennes. En notant (w_i) les pondérations retenues, on utilise finalement des statistiques de la forme :

$$\phi_j = \frac{\left[\sum_{i=1}^{N} w_i \left(d_{ij} - d_i \frac{r_{ij}}{r_i}\right)\right]^2}{\sum_{i=1}^{N} w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}}$$

qui suit asymptotiquement un $\chi^2(1)$. Dans la suite, on notera $\sigma^2 = \sum_{i=1}^{N} w_i^2 d_i \frac{r_i - d_i}{r_i - 1} \frac{r_{i1} r_{i2}}{r_i^2}$

2.2.1. *Le test du log-rank*

Le test du log-rank est le test le plus commun pour comparer des courbes de survie en présence de données censurées. C'est un test non paramétrique, il n'est donc pas nécessaire de connaître les distributions des fonctions de survie que l'on cherche à comparer.

Le choix le plus simple que l'on puisse imaginer pour les pondérations est $w_i = 1$, il conduit au test dit du « log-rank ». Dans ce cas le numérateur de la statistique de test φ_j est le carré de la différence entre le nombre de sorties observés et le nombre de sorties théoriques, sous l'hypothèse nulle :

$$\varphi_j = \frac{\left(D_j^{th} - D_j^{obs}\right)^2}{\sigma^2}$$

Ce test généralise au cas de données censurées le test de Savage. On peut noter que sous l'hypothèse nulle $D_1^{obs} + D_2^{obs} = D_1^{th} + D_2^{th}$, en d'autres termes la valeur de la statistique de test ne dépend pas du groupe sur lequel on l'évalue. La forme de la statistique suggère la formule approchée suivante :

$$\phi = \frac{\left(D_1^{th} - D_1^{obs}\right)^2}{D_1^{th}} + \frac{\left(D_2^{th} - D_2^{obs}\right)^2}{D_2^{th}}$$

dont on peut montrer qu'elle est inférieure à celle du log-rank (cf. PETO et PETO [1972]). Sa forme évoque celle d'un Khi-2 d'ajustement usuel. Le test du log-rank est le test le plus couramment employé et celui que nous allons utiliser par la suite.

2.2.2. *Le test de Gehan*

Gehan propose de retenir $w_i = r_i$, ce qui conduit à pondérer plus fortement les sorties les plus précoces. Ce test généralise au cas de données censurées le test de Wilcoxon. La statistique de test n'admet pas d'expression simplifiée comme dans le cas du log-rank. Il présente l'inconvénient de dépendre assez fortement de la distribution de censure.

PARTIE 3 : ETUDES PRELIMINAIRES

Nous allons dans cette partie appliquer à nos données les différents aspects théoriques et techniques vus dans la partie précédente.
Dans un premier temps, nous analysons les données récupérées, étudions la variable à expliquer ainsi que les variables explicatives, et enfin nous testons les modèles linéaires généralisés dans le but de trouver le modèle adapté à nos données.

CHAPITRE 1. LES DONNEES

Les données fournies sont des données brutes issues d'un sondage sur le sujet, lancé en 2009 par un Cabinet de conseil. Nous disposons de données individuelles pour trois entreprises, et nous allons traiter les données concernant l'entreprise pour laquelle nous avons le plus d'individus.
Avant de commencer tout travail sur ces données brutes, il convient de les retraiter afin de disposer d'une base de travail convenable.

Section 1.1 Présentation des données

Les données par entreprise fournies sont des listes établies sous excel couvrant trois années : 2007, 2008 et 2009.

1.1.1. Liste des salariés de l'entreprise

Nous disposons de la liste de l'ensemble des salariés d'une entreprise. Pour chacun d'entre eux, il est fourni des informations d'ordre général telles que l'âge, l'ancienneté, le sexe, le type de contrat, etc. Les données se présentent de la façon suivante :

Tableau n°2 : *Données concernant les salariés*

Matricule	Nom	Prénom	Catégorie âge	Catégorie ancienneté	Sexe	Contrat	Statut	Métier	Secteur	Région
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…

1.1.2. Liste des absences

Concernant les absences, nous avons les informations suivantes :

Tableau n°3 : *Données concernant les absences*

Matricule	Année	Durée absence	Date début	Date fin
…	…	…	…	…
…	…	…	…	…

La date de début est la date de début d'arrêt réelle, et la date de fin est la date de fin réelle.

Il est important de préciser que nous disposons de la liste des personnes qui ont au moins une absence au cours des trois années étudiées, 2007, 2008 et 2009. On ne calcule donc pas un taux d'absentéisme puisqu'on ne dispose pas des salariés qui n'ont jamais été absents sur cette période.

Pour obtenir un taux d'absentéisme, il faut raisonner en deux temps, il faut que l'absence survienne et il faut savoir combien de temps elle dure. Sur la base du produit fréquence (taux d'incidence) * coût (durée de l'absence), on obtient le taux d'absentéisme qui mesure le pourcentage de jours perdus du fait des absences.

Or, nous ne disposons d'aucune information sur l'exposition, c'est-à-dire que nous ne disposons d'aucune information concernant les personnes qui sont susceptibles d'être absentes mais qui ne le sont pas nécessairement. Il faut alors dans cette partie se limiter à décrire les caractéristiques des absences des gens absents, sans prétendre mesurer la charge de l'absentéisme pour l'entreprise.

Section 1.2 Contrôle des données

Nous disposons de 13 122 lignes d'absences observées sur ces trois années.
Avant d'effectuer des calculs sur ces données, il convient d'en vérifier préalablement la validité. Les différents points pris en compte pour exploiter le fichier sont listés ci-dessous :

- Doublons

La période d'observation étant de trois années, nous avons des individus présents plusieurs années. De plus, chaque absence étant comptabilisé, il se peut qu'on ait des doublons par année au niveau des individus quand une personne a été absente plus d'une fois dans la même année. La manière la plus simple de s'y prendre est alors de regrouper les périodes d'absence de chaque individu sur une année pour avoir une ligne par individu chaque année, et cela afin de pouvoir calculer un taux d'absentéisme annuel. En toute rigueur il faudrait regrouper les absences associées à une même cause pour un individu. En regroupant toutes les absences on mélange un peu deux choses : la survenance d'une absence et sa durée, mais on mesure quelque chose qui a du sens, la durée globale d'absence sur une période donnée. On aurait pu faire le choix de ne pas les regrouper, mais alors il aurait fallu analyser aussi l'incidence pour reconstituer une durée globale, ce qui est plus compliqué.
Nous supprimons 9 197 lignes sur les trois années.

- Date de début d'absence

Nous vérifions que la date de début d'absence qui a été saisie dans le fichier est bien antérieure à la date de fin d'absence. Une inversion entre les dates au moment de la saisie est toujours possible. Nous ne trouvons ici aucune erreur concernant la chronologie des dates.

- Durée

Nous vérifions que la durée de l'absence correspond bien à la différence entre la date d'entrée et la date de sortie de l'absence. C'est bien le cas lorsque l'entreprise n'a pas précisé s'il

s'agissait de demi-journées d'absence ou de journées entières. Dans le cas contraire, les demi-journées d'absences sont prises en compte.

Il nous reste finalement 3 925 lignes à étudier.

- Récapitulatif des traitements réalisés sur le fichier

Les corrections apportées aux données initiales permettent d'obtenir une base plus saine pour notre étude. Il faut cependant garder à l'esprit que notre travail portera dans une certaine proportion sur des observations retraitées.

Tableau n°4 - *Synthèse des traitements*

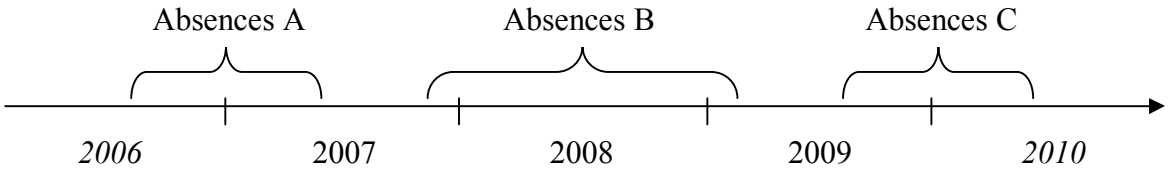
	Nombre de données	Volume par rapport au volume initial
Nombres de lignes initiales	13 122	
Nombre de lignes supprimés	9 197	70,1 %
Nombre de lignes finales	3 925	29,9 %

CHAPITRE 2. LA PERIODE D'OBSERVATION

La période d'observation joue en général un rôle capital car la proportion de censures et de troncatures en dépend. La durée d'observation fixée doit réaliser le meilleur compromis possible entre une durée assez longue pour disposer d'une quantité importante de données, mais aussi être relativement courte pour assurer l'homogénéité des conditions d'observations.

Nous disposons de données d'absences sur la période de 2007 à 2009 :

- des absences de « type A » ayant débutés avant le 01/01/2007 et se terminant entre 2007 et 2009 ;
- des absences de « type B » débutant et finissant entre le 01/01/2007 et le 31/12/2009 ;
- des absences de « type C » débutant entre 2007 et 2009 et non arrêtées au 31/12/2009.



Section 2.1 Date de début d'observation

Aucune question ne se pose concernant les absences de « type B » car elles débutent après le 01/01/2007. Il convient d'examiner les absences de « type A ». Si l'observation débute le 26/12/2006, date de début la plus ancienne, alors toutes les absences recensées dans notre fichier sont bien incluses dans la période d'observation. Cependant, les absences allant du 26/12/2006 au 31/12/2006 sont inconnues, et nos calculs seraient basés sur des données incomplètes, ce qui n'est pas envisageable.
Nous faisons donc débuter notre période d'observation au 01/01/2007.

Section 2.2 Date de fin d'observation

Il convient maintenant d'examiner les absences de « type C ». Si la date de fin d'observation était le 31/01/2010, date de fin d'absence la plus éloignée, alors toutes les absences recensées dans les données seraient bien incluses dans la période d'observation. Cependant, les absences se déroulant entre le 01/01/2010 et le 31/01/2010 seraient basés sur des données incomplètes comme précédemment.
Notre date de fin de période est donc le 31/12/2009.

Section 2.3 Arrêts tronqués ou censurés

La période d'observation étant fixée, nous devons maintenant modifier certaines données.

2.3.1 Censures et troncatures

Pour quelques salariés de notre étude, nous observons uniquement la durée d'absence incluse dans la période d'observation. Pour ces individus, toute l'information n'est pas observable : nous parlons de censures ou de troncatures suivant le cas.

Soit $(X_1,...,X_n)$ notre échantillon de durées d'absences.

- Censure de type 1 : censure fixe

On dit qu'il y a censure à droite pour notre échantillon, si au lieu d'observer directement les durées d'absences $(X_1,...,X_n)$, nous observons $(T_1,D_1),...,(T_n,D_n)$ avec :

$T_i = \min(X_i, C)$ et $D_i = 1$ si $X_i \leq C$,

0 sinon.

Nous observons donc la fin de durée d'absence uniquement si elle a lieu avant C.

Les absences de « type C » sont donc des absences censurées à droite puisque nous ne les observons pas au-delà du 31/12/2009. Nous dénombrons au total 9 absences censurées de ce type dans notre fichier d'étude, ce qui représente 0,07 % des absences.

- Troncature

On dit qu'il y a troncature à gauche lorsque la variable étudiée n'est pas observable lorsqu'elle est inférieure à un seuil C.
La troncature est différente de la censure dans le sens où lorsqu'une variable est tronquée, nous perdons complètement l'information en dehors de la période d'observation. Dans le cas de la censure au contraire, nous savons qu'il y une information, mais nous ne connaissons pas sa valeur précise.

Les absences de « type A » sont donc des absences qui sont tronquées car nous ne les observons pas avant le 01/01/2007. Nous dénombrons au total 2 absences tronquées de ce type dans notre fichier d'étude, ce qui représente 0,02 % des absences.

- Modification de la durée des absences

Les durées renseignées dans les données fournies ne sont pas restreintes à la période d'observation. Il convient donc pour les absences censurées et tronquées de ramener la durée renseignée à la durée effectivement observable.

CHAPITRE 3. ETUDE STATISTIQUE

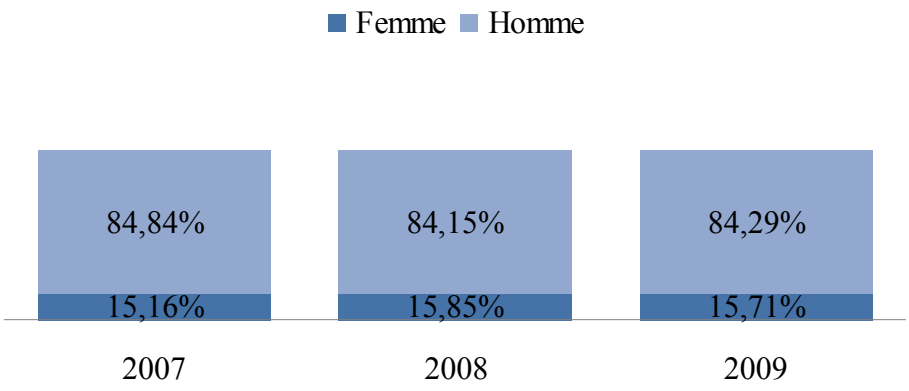
Ce chapitre sera consacré à l'étude statistique de la population d'absents ainsi qu'à l'étude des absences.

Section 3.1 Statistiques démographiques

Etudions les principales caractéristiques démographiques de notre population.

3.1.1. Répartition Homme / Femme

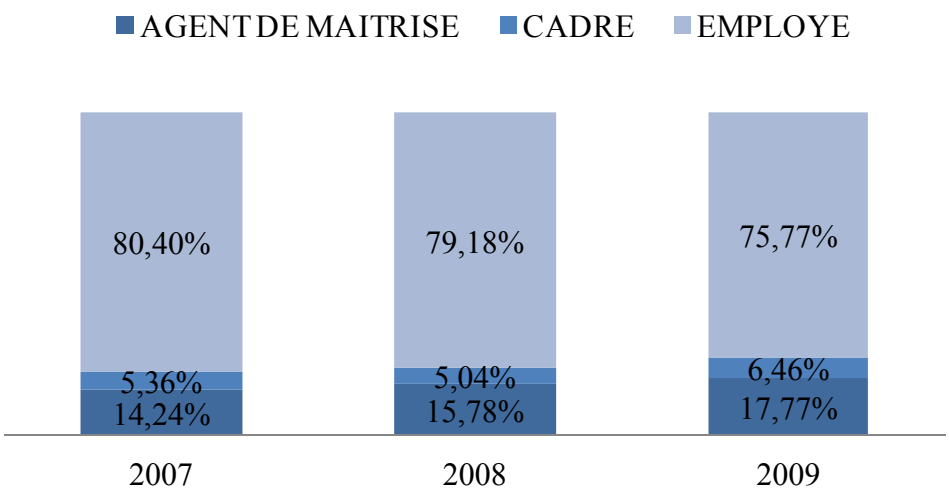
Figure 1 - *Répartition par sexe*



Globalement, notre population est composée essentiellement d'hommes, avec une tendance à une légère augmentation de l'importance des femmes.

3.1.2. *Répartition par catégorie socio-professionnelle*

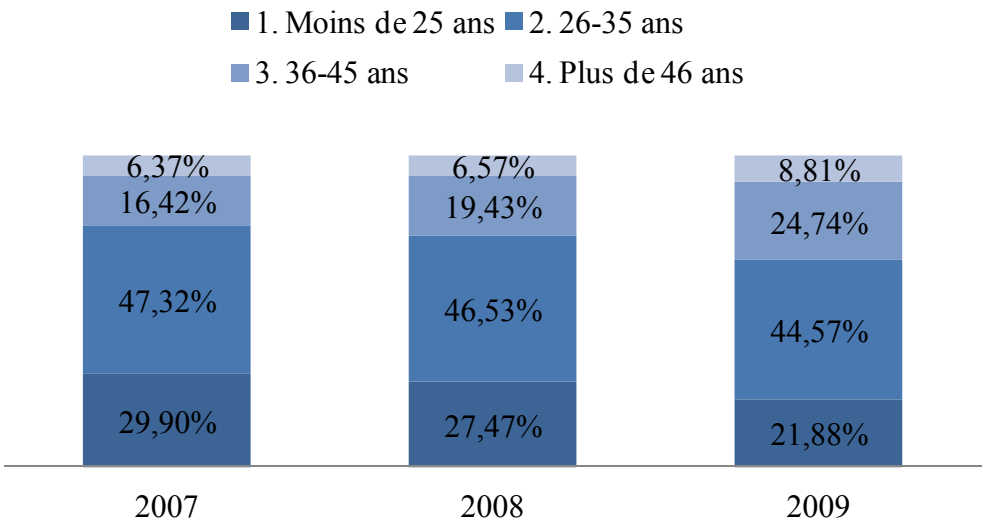
***Figure 2 -** Répartition par CSP*



Sur les trois années, plus des trois quarts de la population traitée est non cadre.

3.1.3. *Répartition en fonction de l'âge*

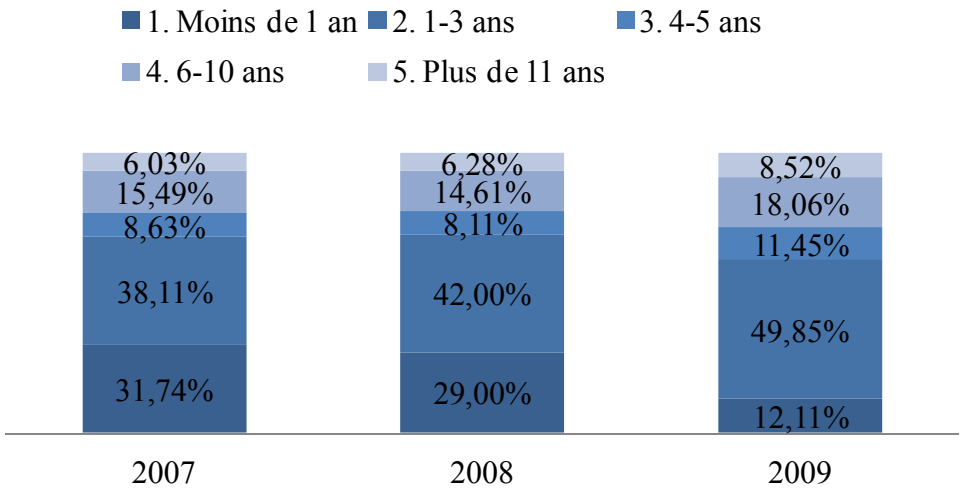
***Figure 3 -** Répartition en fonction de l'âge*



On constate que plus de la moitié de notre population a moins 35 ans. Cette tendance est à la baisse sur les trois années, et la catégorie des personnes ayant entre 36 et 45 ans augmente d'environ 8% entre 2007 et 2009.

3.1.4. *Répartition en fonction de l'ancienneté*

Figure 4 - *Répartition en fonction de l'ancienneté*



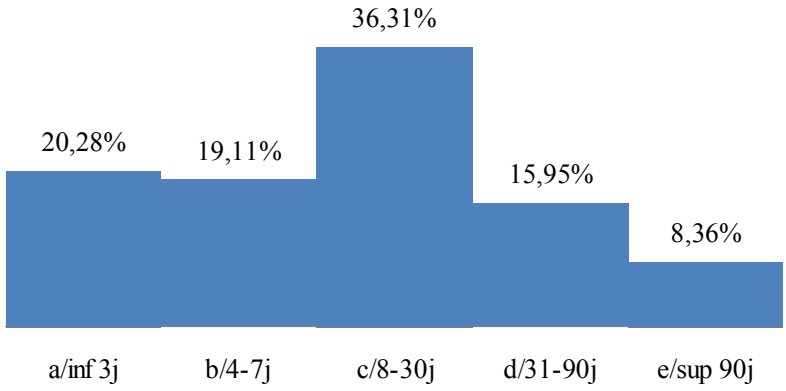
Nous constatons que la majorité des personnes ont moins de trois ans d'ancienneté au sein de l'entreprise. Entre 2007 et 2009, les personnes avec plusieurs années d'ancienneté se substituent aux personnes avec très peu d'ancienneté (moins d'un an).

Nos données d'absence ont donc été établies sur la base d'une population composée de 85 % d'hommes et pour laquelle près de 85 % des salariés sont non cadres. L'ensemble de ces caractéristiques est stable sur la période d'observation.

Section 3.2 Description des absences

Nous constatons que notre individu de référence est un homme âgé entre 26 et 35 ans, employé, en CDI et avec une ancienneté comprise entre 1 et 3 ans.

Figure 5 - *Répartition du nombre de jours d'absence*



D'autre part, un peu plus d'un quart de notre population s'absente entre 8 et 30 jours au cours de la même année.

CHAPITRE 4. ANALYSE DESCRIPTIVE DE LA VARIABLE A EXPLIQUER

Cette première étape a pour but principal de dégager les lois continues permettant d'ajuster convenablement la variable TauxAbs, ainsi que la détermination des variables explicatives significatives. Il n'est pas question ici de quantifier l'impact de ces variables, il s'agit simplement de se faire une idée des variables qui devront être intégrées dans le modèle. Le logiciel utilisé pour cette étude est R.

Section 4.1. Méthode de calcul de la variable TauxAbs

Il n'existe aucune définition légale du calcul du taux d'absentéisme. Cependant, il est possible de définir le taux d'absentéisme par la formule de calcul suivante qui amène à s'interroger sur trois principaux axes :

$$\text{Taux d'absentéisme} = \frac{\text{Temps d'absence pendant une période une période P} * 100}{\text{Temps travaillé pendant P}}$$

- Quelle unité de temps choisir : heures, jours ouvrés, ouvrables, calendaires
- Que doit-on inclure dans le temps travaillé ? congés payés, RTT

Nous avons retenu comme unité de temps les jours calendaires soit tous les jours de la semaine y compris les jours fériés. Par conséquent, tous les jours de l'année sont inclus dans le temps travaillé, les congés payés et RTT non ôtés. La formule de calcul devient la suivante :

$$\text{Taux d'absentéisme} = \frac{\text{Nombre de jours calendaires d'absence sur l'année} * 100}{\text{Nombre de jours calendaires sur l'année}}$$

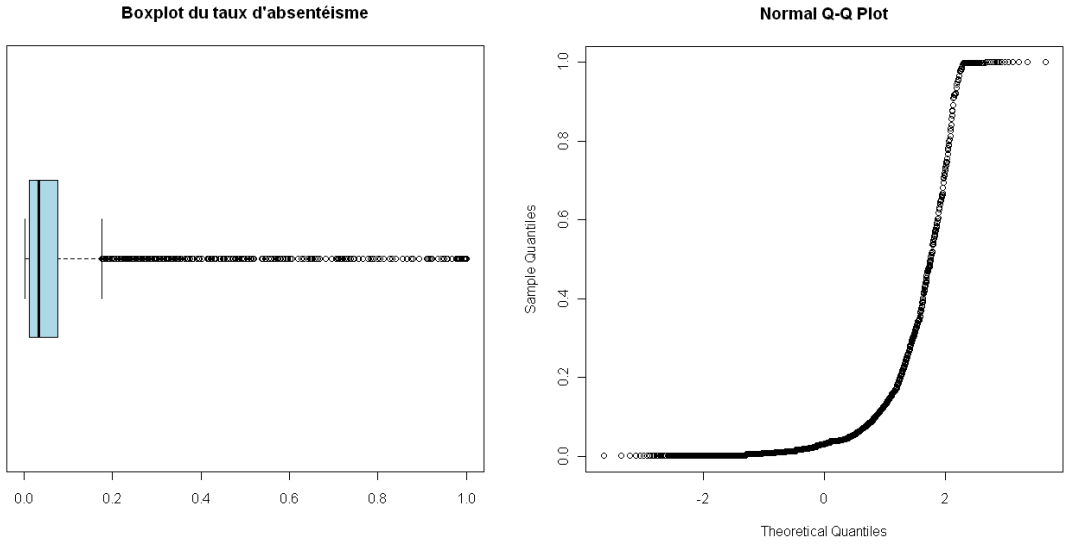
Cependant, rappelons que nous disposons uniquement de la liste des personnes qui au moins une absence au cours des trois années étudiées, 2007, 2008 et 2009. On ne calcule donc pas un taux d'absentéisme puisqu'on ne dispose pas des salariés qui n'ont jamais été absents sur cette période.

Dans cette partie, nous nous limitons donc à décrire les caractéristiques des absences des gens absents, sans prétendre mesurer la charge de l'absentéisme pour l'entreprise.

Section 4.2. Quelques graphiques

A partir de nos données, nous avons pu tracer à l'aide du logiciel R le boxplot ainsi que le qq-plot de notre variable à expliquer, le taux d'absentéisme :

Figure 6 *– Le taux d'absentéisme*



A partir du boxplot, on remarque que 50 % de nos individus ont un taux d'absentéisme relativement faible, compris entre 0 et 10 % environ. De plus, à partir du qq-plot, on observe clairement que le taux d'absentéisme ne suit pas une distribution normale.

Section 4.3. Recherche de lois compatibles avec nos données

3.2.1. Comparaison des fonctions de répartition

Notons $\hat{F}$ la fonction de répartition empirique de notre variable à expliquer. Afin de déterminer les lois susceptibles d'ajuster cette variable, nous allons comparer $\hat{F}$ avec les fonctions de répartition des lois Gamma(n,γ), Exp(λ), LN(μ,σ²) et IG(η,θ).
Les paramètres n, γ, λ, μ, σ², η, θ sont estimés à partir des moyennes et variances empiriques de TauxAbs.

Notons $\overline{X}$ et S² les estimateurs sans biais de E[TauxAbs] et Var[TauxAbs] et $\overline{x}$, s² leurs réalisations à partir de l'échantillon de données TauxAbs.

On rappelle que $\overline{X}$ et S² sont donnés par :

$$\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m} \text{ et } S^2 = \frac{\sum_{i=1}^{m} \left(X_i - \overline{X} \right)^2}{m-1}$$

Où $X_1, X_2, ..., X_m$ sont des variables aléatoires identiquement distribuées de même loi que TauxAbs.

Notons également $\overline{X}_{\log}$ et $S^2_{\log}$ les estimateurs de E[ln(TauxAbs)] et Var[ln(Tauxabs)] et $\overline{x}_{\log}$ et $s^2_{\log}$ leurs réalisations.

Les paramètres des lois théoriques sont alors estimés par la méthode des moments par :

$$\hat{n} = \frac{\overline{x}^2}{s^2} \text{ et } \hat{\gamma} = \frac{\overline{x}}{s^2}$$

$$\hat{\lambda} = \frac{1}{\overline{x}}$$

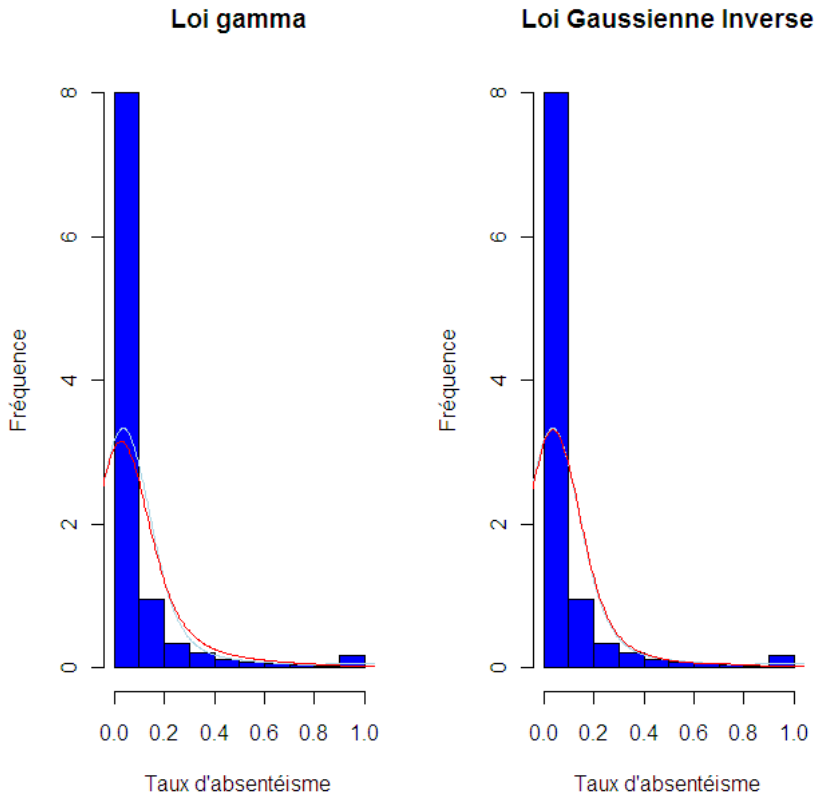
$$\hat{\mu} = \overline{x}_{\log} \text{ et } \hat{\sigma}^2 = s^2_{\log}$$

$$\hat{\eta} = \overline{x} \text{ et } \hat{\theta} = \frac{\overline{x}^3}{s^2}$$

Notre variable à expliquer étant continue, nous allons utiliser les familles de lois gamma et inverse gaussienne.

Afin de voir à quelle loi notre variable se rapproche le plus, nous avons fait les histogrammes suivants (qui permettent d'obtenir l'allure de la densité de la loi) :

***Figure 7** – Histogrammes du taux d'absentéisme*



avec en rouge la densité de la loi testée, et en bleu clair l'allure de la densité de la loi de notre variable à expliquer. On remarque que la loi inverse gaussienne est la loi qui s'en rapproche le plus.
A la vue de ces graphiques, nous pouvons constater que plusieurs salariés ont été absents durant toute une année.

3.2.1. Ajustement q-q plot (Quantile to Quantile Plot)

Cette méthode basée sur la comparaison des quantiles empiriques et théoriques permet de tester l'hypothèse d'adéquation d'une variable à une famille de lois donnée. Cependant, contrairement aux tests statistiques d'adéquation usuels tels que Kolmogorov-Smirnov ou le test d'adéquation du Chi 2, cette méthode ne permet pas de valider ou de rejeter une hypothèse H_0.
Ici, il s'agit uniquement d'étudier graphiquement la compatibilité des données avec une hypothèse de loi.

Notons $\left(x_{(1)}, x_{(2)}, ..., x_{(m)}\right)$ la statistique d'ordre de l'échantillon de notre variable à expliquer et notons F_θ la fonction de répartition de la loi à tester.

Sous l'hypothèse d'adéquation, on doit avoir :

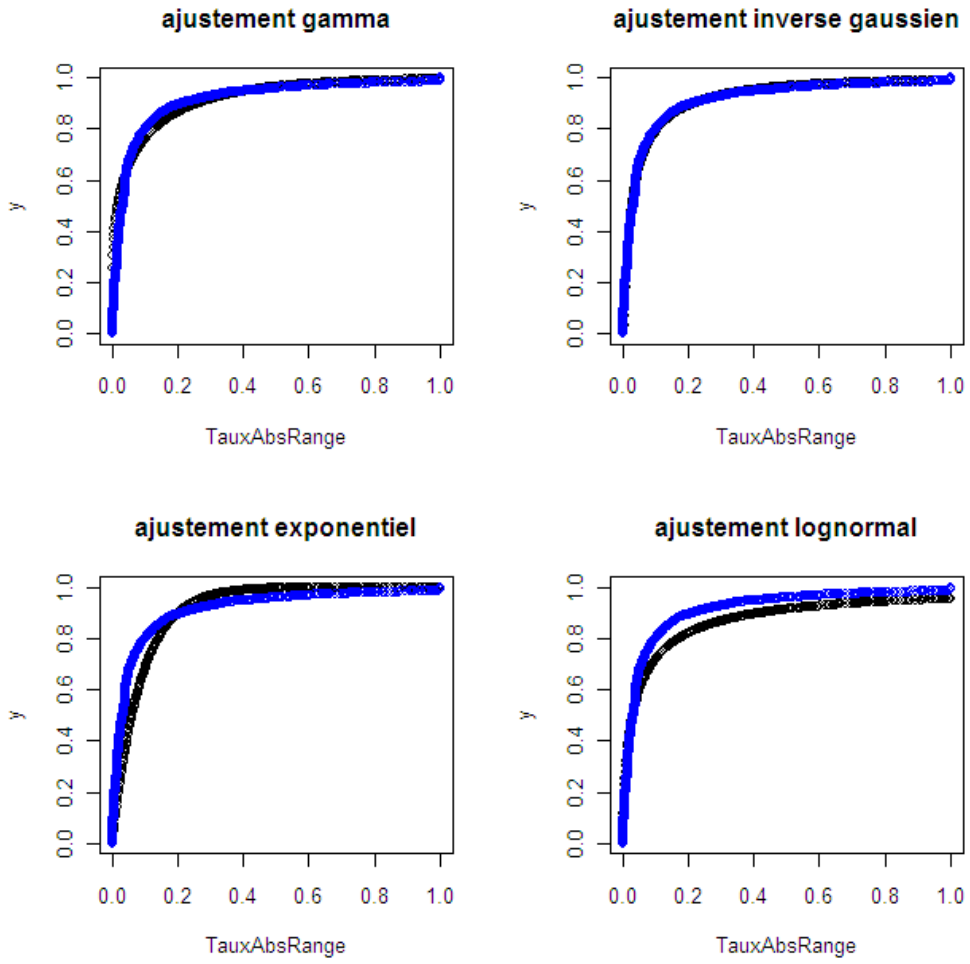
$$F_\theta\left(x_{(i)}\right) \approx \frac{i}{m}, \forall i \in \{1, 2, ..., m\}$$

Graphiquement, cela signifie que les points $\left(x_{(i)}, F_\theta^{-1}\left(\frac{i}{m}\right)\right)_{i=1,...,m}$ doivent être sensiblement alignés sur une droite de pente strictement positive et passant par l'origine.

De la même manière que pour la comparaison des fonctions de répartition, nous allons afficher le nuage de points correspondant à chacune des lois théoriques après en avoir estimé les paramètres.

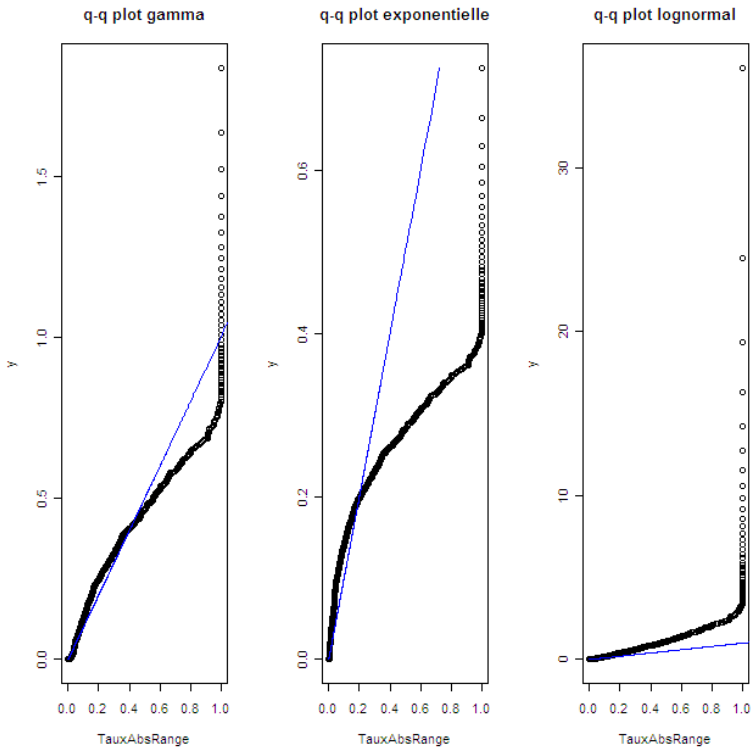
Dans un premier temps figure la comparaison des fonctions de répartition empirique et théorique. La fonction de répartition empirique de la variable à expliquer est en bleu et la fonction de répartition théorique est en noir.

Figure 8 – *Ajustements du taux d'absentéisme*



Ensuite, on affiche pour chaque loi en noir les points $\left(x_{(i)}, F_\theta^{-1}\left(\frac{i}{m}\right)\right)_{i=1,\ldots,m}$ et en bleu la première bissectrice.

Figure 9 *– Q-q plot du taux d'absentéisme*



On remarque clairement que les lois gamma et inverse gaussienne sont les lois qui ajustent le mieux notre variable à expliquer.
On ne peut cependant pas affirmer que la loi gamma ajuste particulièrement bien les données car le nuage de points q-q plot n'est pas confondu avec la première bissectrice.
En revanche, les lois exponentielles et log-normales sont à exclure. Ni l'ajustement des fonctions de répartition, ni celui des quantiles n'est satisfaisant pour ces deux lois.

Section 4.4. Etude visuelle de l'impact des variables qualitatives

Les données dont nous disposons nous conduisent à retenir a priori les variables explicatives suivantes :

- Sexe
- Age
- Ancienneté
- Type de contrat (CDI, CDD)
- Catégorie socio-professionnelle (cadre, employé, ouvrier)

2.3.1. *Paramétrage des variables explicatives*

Nous segmentons nos variables explicatives en différentes classes. Une fois cette segmentation effectuée, on obtient des variables qualitatives, ce qui nécessite un traitement particulier.

Prenons l'exemple de deux variables explicatives qualitatives notées A et B prenant respectivement leurs valeurs à travers a et b modalités. On écrit donc les matrices des variables A et B sous la forme de vecteurs binaires contenant respectivement a et b éléments.
Si l'on considère un individu quelconque, l'élément i d'une variable explicative prend la valeur 1 si l'individu prend la modalité i. Par exemple, si l'on considère la variable Age segmentée en quatre modalités : moins de 25 ans, de 26 à 35 ans (modalité de référence), de 36 à 45 ans et plus de 46 ans. La matrice d'un individu âgé de 40 ans pour cette variable sera (0, 0, 1, 0).

Dans le cas où l'on considère plusieurs variables explicatives, la matrice des explicatives sera la concaténation des matrices des différentes variables.
Un individu sera donc représenté par un vecteur prenant les valeurs 0 ou 1 selon les valeurs des variables explicatives lui correspondant.

La modélisation que nous retiendrons pour l'étude est quelque peu différente puisque nous ajouterons un Intercept à la matrice des variables explicatives, c'est-à-dire un vecteur dont les valeurs seront égales à 1. Un individu de référence sera alors choisi comme étant celui présentant les caractéristiques les plus répandues dans la population.
La matrice d'une variable explicative ne comporte plus autant d'éléments que de modalités de la variable, mais un élément de moins puisque ne sont représentées que les modalités alternatives à la modalité de référence.

De plus, nous optons pour un modèle sans interactions, notamment par souci de simplicité. Et nous allons voir par la suite si le choix de cette hypothèse nous donne des résultats corrects.

Pour l'individu de référence, toutes les valeurs du vecteur servant à le caractériser seront égales à 0. Pour un individu quelconque, chaque modalité variant de la modalité de référence sera caractérisé par la présence d'un 1 dans le vecteur.

Pour expliquer ce codage, prenons l'exemple d'une tarification à deux variables explicatives prenant les modalités suivantes :

- Le sexe : Masculin (modalité de référence) et Féminin
- L'âge : Moins de 35 ans, de 35 à 60 ans (modalité de référence), et plus de 60 ans.

Un individu de sexe masculin et âgé de 33 ans sera donc représenté pour l'intercept par (1), pour la variable sexe par le vecteur (0), pour la variable âge par le vecteur (1,0).
La matrice des explicatives s'écrit donc sous la forme (1,0,1,0).

Le prédicteur linéaire associé est de la forme : $g\left(E\left[Y\right]\right)=\alpha_0+\sum_{i=1}^{n}\alpha_i X_i$

Le α_0 est l'ordonnée à l'origine (ou intercept). Il est associé à l'individu de référence pour lequel tous les X_i sont égaux à 0. De ce fait, un coefficient $\alpha_j > 0$ indique un facteur aggravant l'absentéisme par rapport à l'individu de référence. Au contraire, un coefficient $\alpha_j < 0$ indiquera un facteur améliorant l'absentéisme par rapport à un individu de référence.

Si nous reprenons l'exemple précédent d'une étude à deux variables explicatives Sexe et Age présentant respectivement 2 et 3 modalités, le prédicteur linéaire s'écrit sous la forme :

$$\eta = \alpha X, \text{ avec } \alpha = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} \text{ le vecteur des paramètres,}$$

$$\text{et } X = \begin{pmatrix} \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix} & (X_1) & (X_2) \end{pmatrix} \text{ la matrice des explicatives.}$$

Le premier vecteur de X est l'intercept, composé exclusivement de 1, la matrice X_1 représente la variable Sexe et la matrice X_2 la variable Age.
On peut alors écrire le prédicteur linéaire pour une observation i sous la forme :

$$\eta_i = \alpha_0 + \alpha_1 x_{1_{i1}} + \alpha_2 x_{2_{i1}} + \alpha_3 x_{2_{i2}}$$

Où $x_{1_{i1}} = \begin{cases} 1 \text{ si l'individu i est de sexe féminin} \\ 0 \text{ sinon} \end{cases}$

$x_{2_{i1}} = \begin{cases} 1 \text{ si l'individu est i a moins de 35 ans} \\ 0 \text{ sinon} \end{cases}$

$x_{2_{i2}} = \begin{cases} 1 \text{ si l'individu est i a plus de 60 ans} \\ 0 \text{ sinon} \end{cases}$

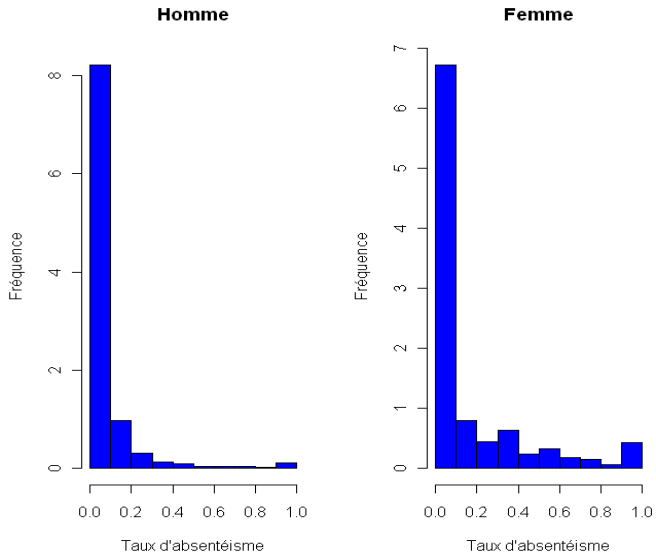
L'espérance de la variable aléatoire Yi, notée μi, peut ainsi se mettre sous la forme :

$$\mu_i = g^{-1}(\eta_i)$$

Pour chaque modalité de ces variables, nous allons comparer les histogrammes (qui permettent d'obtenir l'allure de la densité de la loi).

2.3.2. *La variable Sexe*

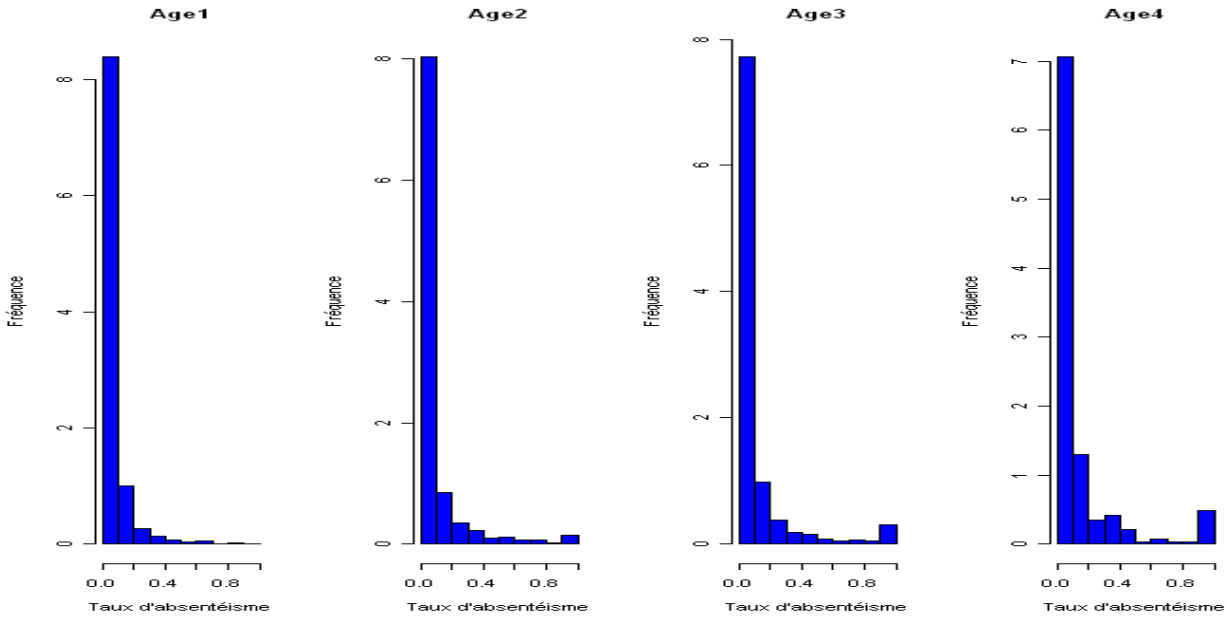
Figure 10 *– Histogrammes de la variable Sexe*



L'allure de la densité de la loi du taux d'absentéisme est légèrement différente, l'influence du sexe sur notre variable à expliquer est significative au vue de ces observations. La moyenne du taux d'absentéisme des femmes est le double de celle des hommes. On observe sur le graphique que les absences des femmes sont de plus longue durée que celles des hommes.

2.3.3. *La variable Age*

Figure 11 *– Histogrammes de la variable Age*

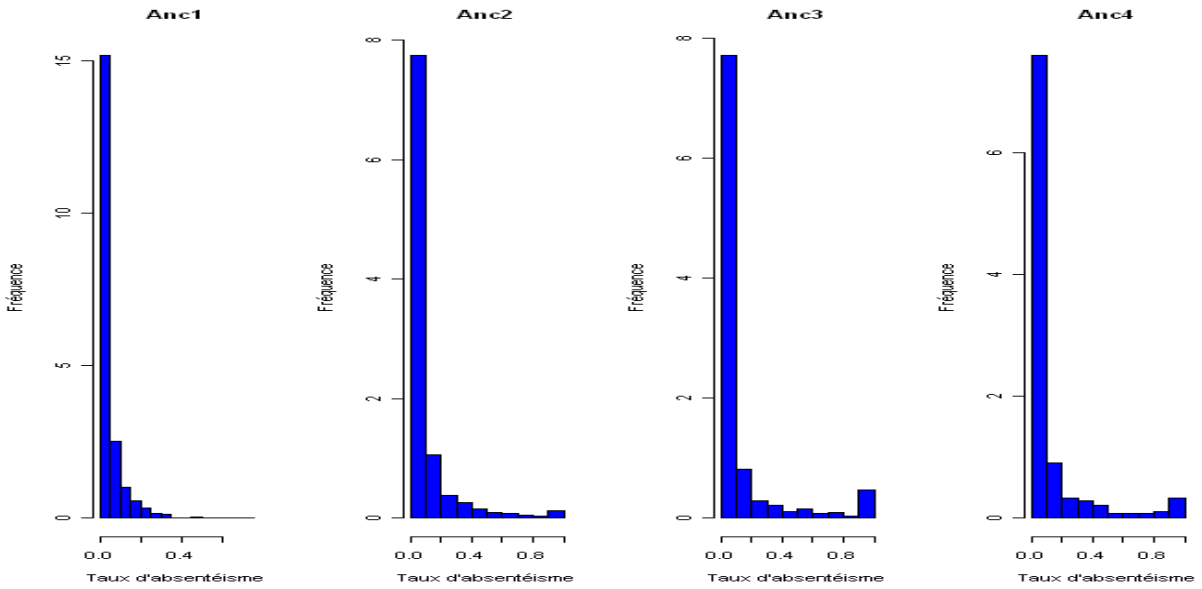


La variable explicative Age est ventilée en 4 modalités : Age1 (moins de 25 ans), Age2 (entre 26 et 35 ans), Age3 (entre 36 et 45 ans) et Age4 (plus de 46 ans).

On observe de très légères variations du taux d'absentéisme selon la catégorie d'âge observée, et notamment une augmentation du taux d'absentéisme avec l'âge. Le taux d'absentéisme des plus âgés est le double du taux d'absentéisme des plus jeunes. Ce qui s'explique en partie par le fait que les plus âgés ont des durées d'absences plus longues que les plus jeunes.

2.3.4. La variable Ancienneté

Figure 12 *– Histogrammes de la variable Ancienneté*

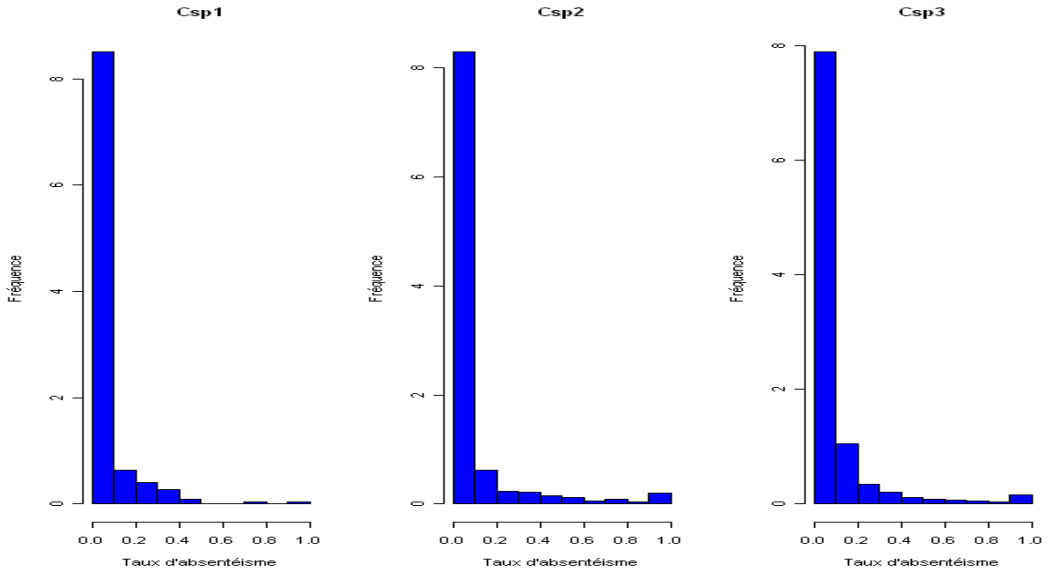


De même que précédemment, cette variable explicative est ventilée en 4 modalités : Anc1 (moins de 1 an), Anc2 (entre 2 et 5 ans), Anc3 (entre 6 et 10 ans) et Anc4 (plus de 11 ans).

On observe une augmentation du taux d'absentéisme avec l'ancienneté avec une certaine stabilité entre les deux dernières catégories d'ancienneté, c'est-à-dire à partir de six ans d'ancienneté.

2.3.5. *La variable Catégorie socio-professionnelle*

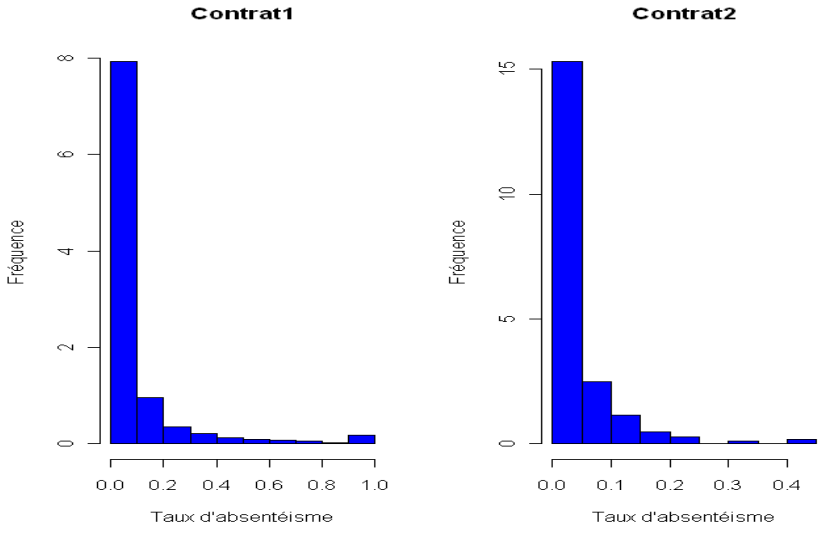
Figure 13 *– Histogrammes de la variable Catégorie socio-professionnelle*



Ici, on remarque que le taux d'absentéisme des cadres (Csp1) est plus faible de 50% que le taux d'absentéisme des non cadre (Csp2 et Csp3).

2.3.6. *La variable Contrat*

Figure 14 *– Histogrammes de la variable Contrat*



Le taux d'absentéisme des personnes en CDI (Contrat1) est le double des personnes en CDD (Contrat2). Ce résultat doit être nuancé du fait du faible volume de données sur ces contrats.

2.3.7. Remarque

L'analyse présentée ci-dessus permet simplement de dégager des tendances puisque nous posons comme hypothèse qu'il n'existe pas d'éventuelles corrélations entre les facteurs, ce qui bien entendu ne correspond pas à la réalité car des corrélations existent entre les variables explicatives comme par exemple l'ancienneté et le type de contrat, l'ancienneté et la catégorie socioprofessionnelle dans certaines entreprises ou encore l'âge et l'ancienneté.

CHAPITRE 5. MISE EN ŒUVRE DES MODELES LINEAIRES GENERALISES

Dans ce chapitre, nous allons mettre en œuvre les Modèles Linéaires Généralisés sur les données de l'entreprise que nous avons choisie.

Section 5.1. Sur l'ensemble des données

4.1.1. Le test du log rank

Rappelons que le test du log rank compare les estimations des fonctions de hasard des deux échantillons à chaque temps d'évènement observé. Sous l'hypothèse nulle, les deux courbes de survie ne sont pas différentes. Ici, seuls les facteurs qualitatifs peuvent être pris en compte.
Pour chaque variable explicative, on fixe une modalité comme la population de référence arbitraire, par exemple celle contenant le plus de monde. On va ensuite prendre chaque modalité et tester si sa distribution de durée est égale à celle de la population de référence. On peut alors faire une liste de modalités à regrouper avec la modalité de référence.
Le souci qu'on peut rencontrer est que le test de « A a la même distribution de durée que R » et « B a la même distribution de durée que R » passent alors que « A a la même distribution que B » ne passe pas.

1) La variable Catégorie socioprofessionnelle

On choisit comme modalité de référence la Csp3 (ouvrier/employé) et on la compare aux autres modalités, Csp1 et Csp2. Sur R, on utilise la fonction survdiff pour ce test.

Tableau n°5 *– Résultats du test du log rank pour la variable CSP*

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Csp1	221	221	175	12,263	13,7
Csp3	3 076	3 076	3 122	0,686	13,7

Valeur de la statistique = 13.7 et p-valeur = 0.000216

Tableau n°6 *– Résultats du test du log rank pour la variable CSP*

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Csp2	628	628	596	1,751	2,22
Csp3	3 076	3 076	3 108	0,336	2,22

Valeur de la statistique = 2.2 et p-valeur = 0.137

où : N est le nombre de personnes dans chaque groupe
Observed le nombre de personnes observé dans chaque groupe
Expected le nombre de personnes attendu dans chaque groupe
Chisq la statistique du chi deux pour un test d'égalité

On constate que l'on peut regrouper la Csp2 avec la Csp3 car notre p-valeur est supérieure à 5% donc on accepte l'hypothèse nulle : Csp3 et Csp2 ont la même distribution de durée.
On a donc deux modalités : cadre et non cadre.

2) La variable Age

Age2 (26-35 ans) est notre variable de référence. On a les résultats suivants :

Tableau n°7 *– Résultats du test du log rank pour la variable Age*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Age1	1 031	1 031	909	16,4	25,8
Age2	1 809	1 809	1 931	7,7	25,8

Valeur de la statistique = 25.8 et p-valeur = 3.76e-07

Tableau n°8 *– Résultats du test du log rank pour la variable Age*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Age2	1 809	1 809	1 772	0,782	2,61
Age3	799	799	836	1,657	2,61

Valeur de la statistique = 2.6 et p-valeur = 0.106

Tableau n°9 *– Résultats du test du log rank pour la variable Age*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Age2	1 809	1 809	1 769	0,924	6,4
Age4	286	286	326	5,006	6,4

Valeur de la statistique = 6.4 et p-valeur = 0.0114

On regroupe la variable Age3 avec notre variable de référence.
On a trois modalités : moins de 25 ans, entre 26 et 45 ans, et plus de 46 ans.

3) La variable Ancienneté

Anc2 est notre variable de référence (1-5 ans). On a les résultats suivants :

Tableau n°10 *– Résultats du test du log rank pour la variable Ancienneté*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Anc1	941	941	699	84,1	119
Anc2	2 079	2 079	2 321	25,3	119

Valeur de la statistique = 119 et p-valeur = 0

Tableau n°11 *– Résultats du test du log rank pour la variable Ancienneté*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Anc2	2 079	2 079	2 038	0,83	3,62
Anc3	631	631	672	2,51	3,62

Valeur de la statistique = 3.6 et p-valeur = 0.057

Tableau n°12 *– Résultats du test du log rank pour la variable Ancienneté*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Anc2	2 079	2 079	2 071	0,029	0,257
Anc4	274	274	282	0,213	0,257

Valeur de la statistique = 0.3 et p-valeur = 0.612

On regroupe nos modalités Anc3 et Anc4 avec Anc2.
Nos modalités finales sont : moins d'1 an et plus d'1 an.

4) La variable Sexe

Tableau n°13 *– Résultats du test du log rank pour la variable Sexe*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
S1	3 313	3 313	3 092	15,8	82,5
S2	612	612	833	58,7	82,5

Valeur de la statistique = 82.5 et p-valeur = 0

5) La variable Contrat

Tableau n°14 *– Résultats du test du log rank pour la variable Contrat*

	N	Observed	Expected	(O-E)²/E	(O-E)²/V
Contrat1	3 715	3 715	3 780	1,11	32,1
Contrat2	210	210	145	28,99	32,1

Valeur de la statistique = 32.1 et p-valeur = 1.47e-08

Concernant les variables Sexe et Contrat, il n'y a aucune modification à effectuer d'après le test du log rank.

4.1.2. Les modèles linéaires généralisés

1) Choix du modèle

Le but de la présente étude va être de modéliser les absences, à partir de variables explicatives. Ces dernières sont le sexe de l'individu, la catégorie d'âge dans laquelle il se trouve, son ancienneté au sein de l'entreprise, sa catégorie socioprofessionnelle, et le type de son contrat. Ces variables explicatives étant toutes qualitatives, il est nécessaire de créer des variables binaires afin d'utiliser les modèles linéaires généralisés.

Dans la partie précédente, on a pu sélectionner les lois continues permettant de modéliser la variable à expliquer (loi gamma / inverse gaussienne). On va tester dans cette partie les modèles linéaires généralisés afin de déterminer la meilleure modélisation et le vecteur des paramètres du modèle.

Tableau n°15 *– Choix du modèle*

Famille	Fonction de lien	AIC
Gamma	Log	- 12 676
Gamma	Inverse	- 12 687
Gamma	Identité	- 12 634
Inverse gaussienne	Inverse	- 13 583

Il apparait que le modèle le plus pertinent à l'égard du critère de minimisation de l'AIC est le modèle avec pour famille la famille inverse gaussienne et la fonction de lien inverse. En effet, on a AIC= -13 583.

Rappelons que l'AIC (Akaike Information Criterion) nous permet d'évaluer la bonne adéquation d'un modèle et surtout de comparer plusieurs modèles entre eux. L'AIC utilise le maximum de vraisemblance, mais en pénalisant les modèles comportant trop de variables.
Sa formulation est la suivante : AIC = -2 ln $L(\theta) + 2k$ avec k le nombre de paramètres.
Le modèle à retenir est celui qui montre l'AIC le plus faible.

Il est habituel de présenter ce critère avec le BIC de Schwarz, qui pénalise davantage le sur paramétrage.

L'ajustement est convenable si $\frac{D}{n-p-1}$ n'est pas beaucoup plus grande que 1, ce qui est le cas uniquement pour le modèle Gamma avec fonction de lien inverse.

Nous allons plutôt utiliser le critère AIC car il tient compte du nombre de paramètres du modèle, ce qui n'est pas le cas de la déviance.

Dans un premier temps, on observe que les familles les plus adaptées à nos données sont la famille gaussienne inverse et gamma avec fonction de lien inverse. Cependant, il est nécessaire d'utiliser la méthode *backward* afin de ne garder dans notre modèle que les variables explicatives significatives.

2) Sélection des variables par la méthode *backward*

Nous avons, sur R, les résultats suivants :

Tableau n°16 *– Résultats de la méthode backward*

	Coefficients estimés	**Erreur standard**	**Statistique**	**P-value**
Intercept	11.7971	0.4844	24.356	< 2e-16
S2	-6.2287	0.7648	-8.144	5.11e-16
Age1	2.5718	0.9627	2.671	0.007587
Age4	-3.1798	1.0283	-3.092	0.002000
Anc1	9.9049	1.0822	9.152	< 2e-16
Csp1	6.3653	1.7551	3.627	0.000291
Contrat2	5.5315	2.3168	2.388	0.017006

On remarque que toutes nos modalités sont significatives car toutes les p-value sont inférieures à 5%, ce qui signifie qu'on rejette l'hypothèse nulle que les modalités n'ont pas d'impact sur la variable à expliquer. D'autre part, on remarque ici que les modalités S2 (sexe féminin) et Age4 (plus de 46 ans) ont tendance à jouer favorablement sur le taux d'absentéisme, ce qui n'est pas le cas des autres modalités.

Section 5.2. Sur les absences de courte durée

On va maintenant distinguer les absences de courte durée des absences de longue durée en considérant qu'au-delà de 181 jours (soit 6 mois), une absence est considérée comme être de longue durée.

Cette distinction réside dans le fait que les entreprises peuvent agir sur les absences de courte durée et qu'il est souvent difficile d'agir sur des absences de longue durée. Il est donc intéressant d'effectuer cette distinction. La définition des absences de longue durée peut varier d'une entreprise à une autre mais 6 mois est souvent considéré comme la limite.
On supprime 156 lignes de données. Nous disposons désormais de 3 769 lignes d'absence.

4.2.1. *Le test du log rank*

1) La variable Age

Ici, c'est Age2 (26-35 ans) notre variable de référence. Après avoir effectué des tests comme précédemment, on décide de regrouper la variable Age3 et Age4 avec notre variable de référence.
Nous avons deux modalités : moins de 25 ans, et plus de 26 ans.

2) La variable Ancienneté

Anc2 est notre variable de référence (1-5 ans), et on regroupe la modalité Anc4.
Nos modalités finales sont : moins d'1 an, entre 1 et 5 ans, et plus de six ans.

Concernant les variables Sexe, Catégorie socioprofessionnelle et Contrat, il n'y a aucune modification à effectuer d'après le test du log rank.

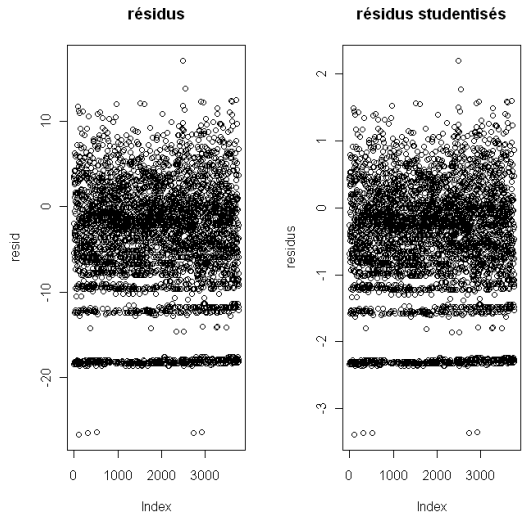
4.2.2. *Les modèles linéaires généralisés*

Tout comme dans la partie précédente, il apparait que le modèle le plus pertinent à l'égard du critère de minimisation de l'AIC est le modèle avec pour loi la loi gaussienne inverse avec fonction de lien inverse (AIC = - 14 515).
Comme précédemment, nous utilisons la méthode *backward* afin de ne garder dans notre modèle uniquement les variables explicatives significatives. En l'occurrence on les garde toutes.
La qualité d'ajustement n'est toujours pas convenable pour les modèles avec loi inverse gaussienne.

4.2.3. *Analyse des résidus*

Le test de Durbin-Watson nous indique que les résidus sont indépendants car D = 1,87 est proche de 2.
On regarde maintenant nos résidus afin de valider notre modèle :

Figure n°17 *– Résidus du modèle*



Les données ayant des résidus studentisés en dehors de l'intervalle [-2;2] sont considérées comme étant aberrantes. Afin de corriger ces résidus mal placés correspondant probablement à des valeurs atypiques de la durée d'absence, on choisit de supprimer les lignes avec un taux d'absentéisme inférieur à 1% considérant que les individus s'absentent au moins 3 jours dans l'année pour quelques raisons que ce soit. On réitère donc notre étude sur la base de ces données.

Section 5.3. Sur les absences de courte durée sans valeurs aberrantes

4.3.1. *Le test du log rank*

1) La variable Age

Age2 (26-35 ans) est notre variable de référence. Après avoir effectué des tests comme précédemment, on décide de regrouper la variable Age1 et Age3 avec notre variable de référence.
Maintenant, on a deux modalités : moins de 45 ans, et plus de 46 ans.

2) La variable Ancienneté

Anc2 est notre variable de référence (1-5 ans), et on regroupe la modalité Anc4.
Nos modalités finales sont : moins d'1 an, entre 1 et 5 ans, et plus de six ans.

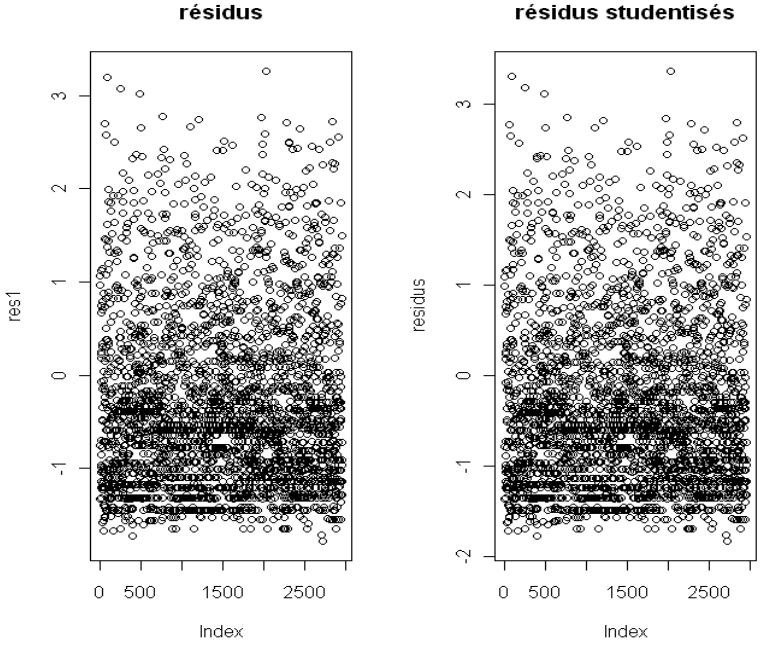
3) La variable Csp

Csp1 et Csp2 suivent la même distribution de durée que la variable de référence Csp3.

Concernant les variables Sexe et Contrat, il n'y a aucune modification à effectuer d'après le test du log rank.

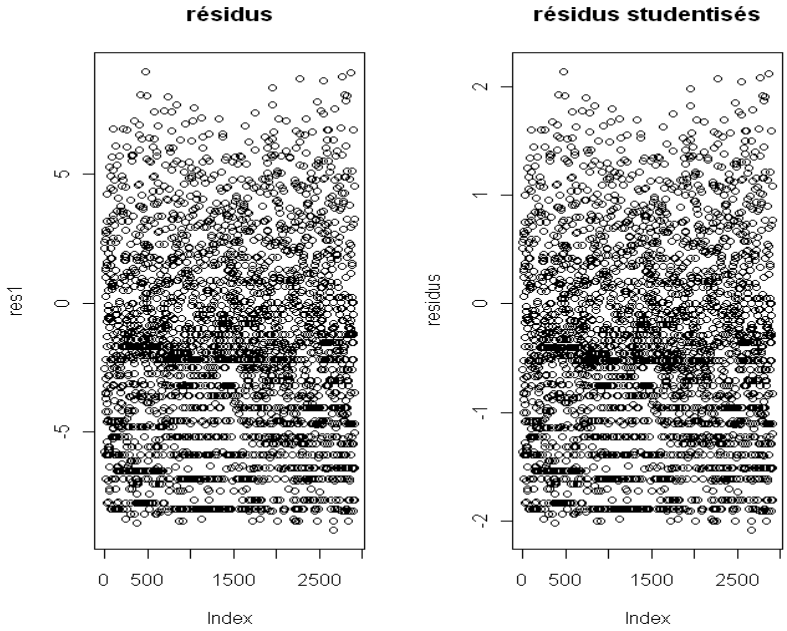
En testant les modèles linéaires généralisés, on remarque que notre variable explicative Contrat n'est pas significative. On l'enlève donc du modèle.
Finalement, on garde comme modèle le modèle inverse gaussien avec fonction de lien inverse comme précédemment. On pourrait également choisir la fonction de lien log car les AIC sont égaux pour ces deux modèles.

Figure n°18 *– Résidus du modèle*



On observe encore des valeurs aberrantes (valeurs supérieures à 2). On supprime alors les taux d'absentéisme supérieurs à 40 % et on a cette fois les résidus suivants :

Figure n°19 *– Résidus du modèle final*



Nos résidus sont centrés, homoscédastiques et on n'a plus de valeurs aberrantes, ce qui valide notre modèle.

On a, sur R, les résultats suivants :

Tableau n°17 *– Coefficients estimés des variables explicatives du modèle*

	Coefficients estimés	**Erreur standard**	**Statistique**	**P-value**
Intercept	14,8543	0,397	37,412	< 2e-16
S2	- 4,2109	0,7572	- 5,561	2.93e-08
Age4	- 3.5271	1,0442	- 3.378	0.00074
Anc1	3,4943	0,7614	4,589	4.64e-06
Anc3	2,5714	0,8496	3,027	0.00249

Comme dans notre section 4.4., les modalités S2 (sexe féminin) et Age4 (plus de 46 ans) ont tendance à augmenter le taux d'absentéisme, ce qui n'est pas le cas des autres modalités Anc1 (moins de 1 an d'ancienneté) et Anc3 (plus de 6 ans d'ancienneté).
En d'autres termes, les femmes ainsi que les salariés de plus de 46 ans auront tendance à être plus absents de l'entreprise que les autres. D'autre part, les salariés avec peu ou beaucoup d'ancienneté auront tendance à être moins absents de l'entreprise que les autres salariés.

Section 5.4. Prévisions

Rappelons que le modèle le plus adapté à nos données est le modèle inverse gaussien avec fonction de lien inverse. Cela signifie que l'on a :

$$\frac{1}{E(Y)} = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

Ou encore :

$$E(Y) = \frac{1}{\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i}$$

On peut donc calculer le taux d'absentéisme estimé :

$$TauxAbs = \frac{1}{14.8543 - 4.2109 * 1_{S2} - 3.5271 * 1_{Age4} + 3.4943 * 1_{Anc1} + 2.5714 * 1_{Anc3}}$$

avec les modalités suivantes :

Tableau n°18 *– Rappel de la signification des modalités*

Modalité	Equivaut à
Age2	Moins de 45 ans
Age4	Plus de 46 ans
S1	Homme
S2	Femme
Anc1	Moins d'1 an
Anc2	Entre 1 et 5 ans
Anc3	Plus de 6 ans

Prenons l'exemple d'un homme âgé de plus de 46 ans avec plus de 6 ans d'ancienneté. D'après notre modèle, son taux d'absentéisme prédit est alors le suivant :

$$TauxAbs = \frac{1}{14.8543 - 3.5271 + 2.5714} = 7.19\%$$

Maintenant que notre modèle est stabilisé, nous pouvons l'utiliser sur nos données initiales afin faire des prédictions de taux d'absentéisme. On a en exemple les résultats suivants :

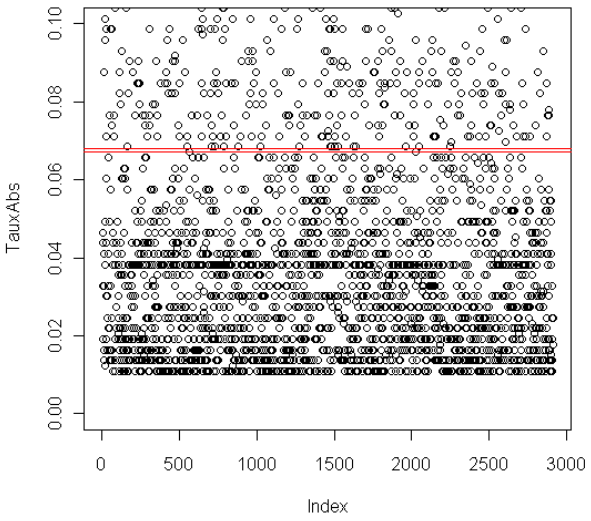
Tableau n°19 *– Exemple d'erreurs de prévision*

TauxAbs	Prévision	Résidus
30,8 %	14,1 %	16,7 %
33,1 %	14,1 %	19,1 %
4,4 %	9,4 %	-5,0 %
1,9 %	9,4 %	-7,5 %
13,1 %	8,8 %	4,3 %
4,7 %	8,8 %	-4,2 %
20,0 %	8,8 %	11,2 %
3,3 %	8,8 %	-5,5 %

L'erreur de prévision s'écrit : $\varepsilon^P = Y - \hat{Y}^P$ et est centrée. En effet, en faisant la moyenne de nos erreurs de prévision, on obtient -0.004%.
On constate que nos prévisions ne sont pas justes, mais que sur l'ensemble de notre population, la moyenne de nos résidus est proche de 0.

Il peut être intéressant de déterminer des intervalles de confiance autour des valeurs prédites et de positionner dedans les valeurs observées. La fonction t.test du langage R fournit la moyenne d'un vecteur ainsi que son intervalle de confiance.
On a par exemple les résultats suivants : t = 49.1965, df = 2912, p-value < 2.2e-16
95 percent confidence interval : 0.06481927 0.07020063
mean of x : 0.06750995

Figure 20 *– Intervalle de confiance des valeurs prédites*



Il est utile de préciser que lorsqu'il y a plusieurs variables, des interactions sont possibles entre ces dernières. En toute rigueur, il aurait fallu tester dans cette partie les modèles avec interactions, mais nous avons choisi de ne pas le faire par souci de simplicité d'une part et aussi parce que les résultats obtenus avec les modèles sans interactions sont corrects.

PARTIE 4 : PARTICIPANTS AU BAROMETRE DE L'ABSENTEISME

Nous souhaitons désormais mettre en œuvre notre étude aux entreprises participantes aux sondages réalisés en 2008 et 2009, lancé par un Cabinet de conseil.

Ces questionnaires ont été diffusés nationalement à des établissements de plus de 150 collaborateurs des secteurs privés et publics, avec pour objectif de permettre aux entreprises et collectivités d'évaluer leur niveau d'absentéisme et de permettre une capitalisation significative de données d'absentéisme.

Notre objectif à l'issue de cette étude est de pouvoir estimer un taux d'absentéisme propre à chaque entreprise en fonction de certains critères. Cela aboutirait à la mise en place de solutions selon le niveau d'absentéisme au sein d'une entreprise, dans le but de diminuer cet absentéisme.

CHAPITRE 1 : PRESENTATION DES DONNEES

Dans ce chapitre, nous exposons le contexte et la méthodologie de ces sondages.

Section 1.1. Contexte

Les informations disponibles sur l'absentéisme en France étant insuffisantes et non systématiques pour agir de façon globale et cohérente, un Cabinet de conseil a lancé en 2008 un sondage sur le sujet, pour la première fois en France.
Fort de ce succès, ce sondage est renouvelé en 2009. Il permet aux entreprises et collectivités d'évaluer leur niveau d'absentéisme par rapport à un secteur d'activité, une zone géographique et de se comparer à la moyenne nationale.
Ce sondage permet une capitalisation significative de données d'absentéisme et participe à la sensibilisation des acteurs aux enjeux sociaux et économiques liés à ce phénomène.

Section 1.2. Profil des participants

1) Caractéristiques des répondants

193 entreprises et collectivités (après retraitement) ont participés au sondage en 2009 contre 205 en 2008. Cependant, la base des répondants en 2009 représente 410 284 collaborateurs, soit le double de l'édition précédente.
D'autre part, près de 38 % des répondants ont renouvelé leur participation à ce sondage en 2009.

2) Contrôle des données

Dans le cadre de cette étude, il était nécessaire d'avoir le plus de données possibles. En ce sens, nous avons regroupé celles obtenues en 2008 et 2009, en supprimant les doublons dus à la présence de certaines entreprises sur les deux années. Lorsqu'une entreprise a participé aux sondages 2008 et 2009, nous avons choisis de supprimer les données de 2008 et de conserver celles de 2009, afin de garder les données les plus récentes, tout en vérifiant qu'il n'y a pas de différences aberrantes de taux d'absentéisme entre 2008 et 2009.

Nous avons reçu 205 et 254 lignes de données provenant de la base des sondages 2008 et 2009 respectivement, soit 459 lignes de données pour les deux années.

Tableau n°20 - *Synthèse des traitements*

	Nombre de données	Volume par rapport au volume initial
Nombres de lignes initiales	459	
Nombre de lignes supprimés	243	52,9 %
Nombre de lignes finales	216	47,1 %

Nous avons supprimé 243 lignes pour cause de doublons ou de manque d'information.

Les causes de suppression de lignes de données sont principalement les suivantes :

- Le manque de données : pour certaines entreprises nous n'avions pas les informations suffisantes pour calculer le taux d'absentéisme et pour d'autres il nous manquait des informations indispensables pour notre étude, telles que les variables explicatives du taux d'absentéisme que verrons par la suite.
- Les valeurs aberrantes : après discussion avec le Cabinet de conseil, nous avons considéré qu'une valeur est aberrante lorsqu'on a un taux d'absentéisme inférieur à 1 % ou supérieur à 15 %. Nous avons donc exclus ces lignes de données de nos travaux.
- Les entreprises présentes dans les sondages 2008 et 2009 : comme évoqué précédemment, nous avons choisi de garder que les lignes de données 2009 des entreprises présentes les deux années afin de ne pas avoir de doublons qui fausseraient notre étude.

CHAPITRE 2 : ETUDE STATISTIQUE

Section 2.1. Répartition des répondants en fonction du secteur géographique

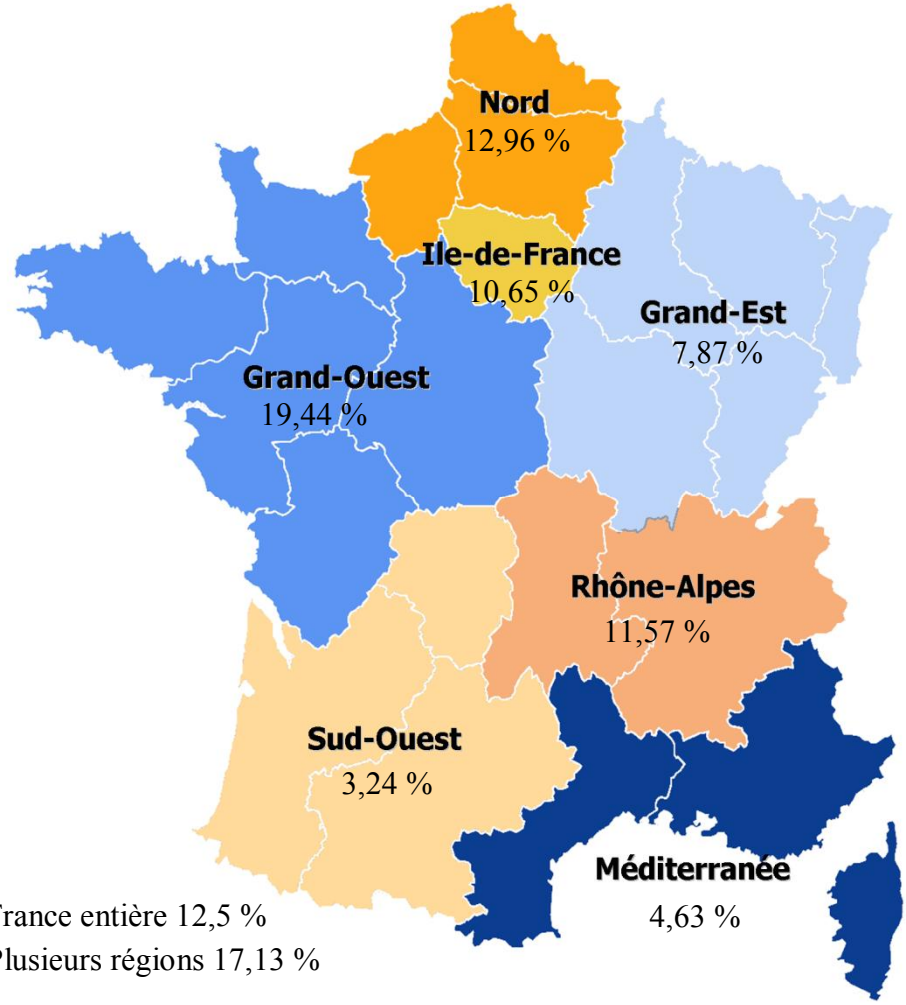
Pour effectuer cette répartition, la France a été découpée en 7 secteurs géographiques :

- Nord
- Grand Ouest
- Grand Est
- Ile-de-France
- Rhône-Alpes
- Sud-ouest
- Méditerranée.

De plus, certains participants ayant répondu pour plusieurs établissements situés dans différentes régions, deux autres secteurs ont été ajoutés :

- « France entière », pour les organisations structurées en réseau sur tout le territoire ;
- « Plusieurs régions », pour les organisations situées sur 2 régions différentes ou plus.

Figure 21 *– Répartition des répondants en fonction du secteur géographique*

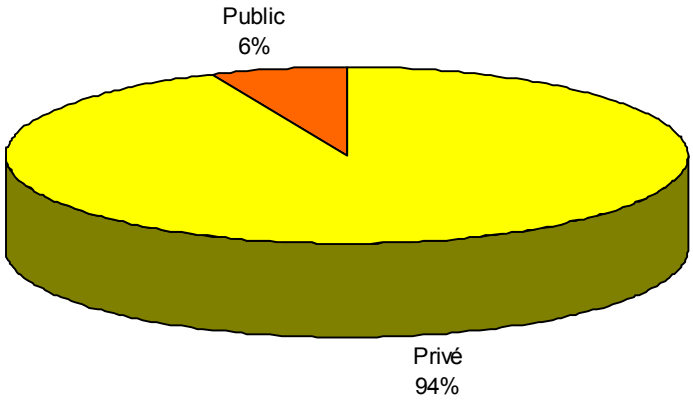


France entière 12,5 %

Plusieurs régions 17,13 %

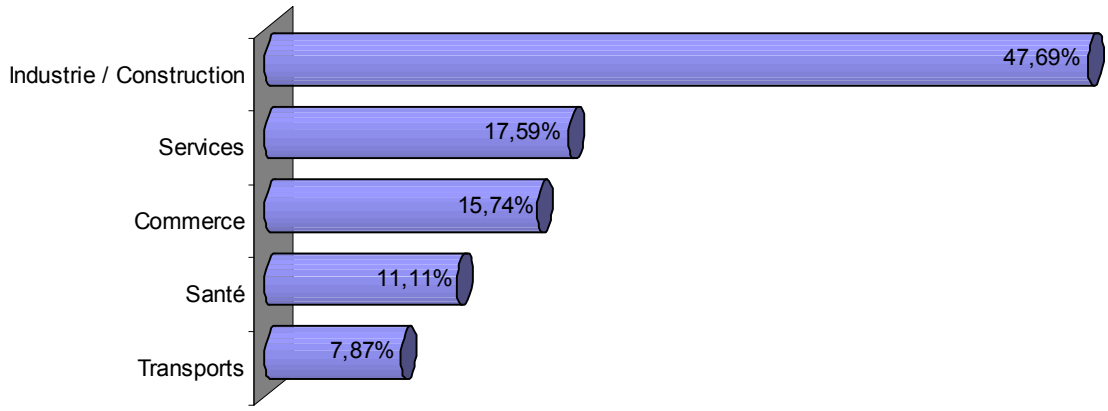
Section 2.2. Analyse de l'activité des établissements

Figure 22 *– Répartition des participants par secteur privé / public*



94 % des répondants font partie du secteur privé. Bien que les établissements publics soient faiblement représentés, leurs effectifs sont significatifs. En effet, des structures publiques d'effectifs importants ont répondu aux sondages.

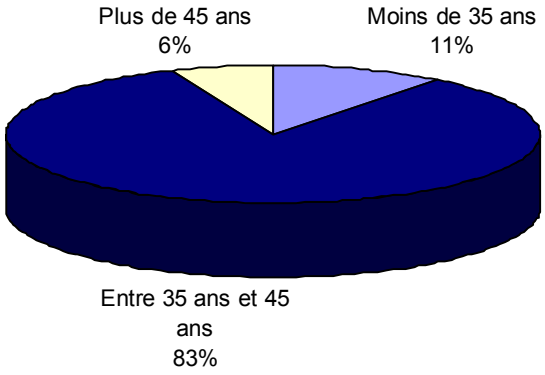
Figure 23 *– Répartition des participants par secteur d'activité*



Près de la moitié des répondants (48 %) sont issus du secteur de l'Industrie et de la Construction. Cette surreprésentation peut s'expliquer par l'existence fréquente de problématiques d'absentéisme dans ce domaine d'activité. Le Transport est le secteur le moins représenté (8 %). Les entités de ce secteur se sentent moins concernées par la problématique d'absentéisme et portent leur attention majoritairement sur la prévention des risques professionnels.

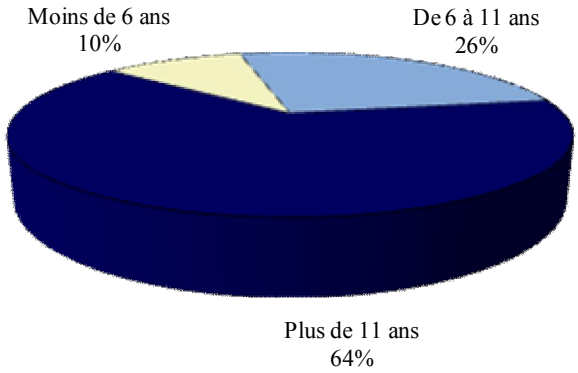
Section 2.3. Analyse de la population salariée

Figure 24 *– Répartition des répondants par âge moyen*



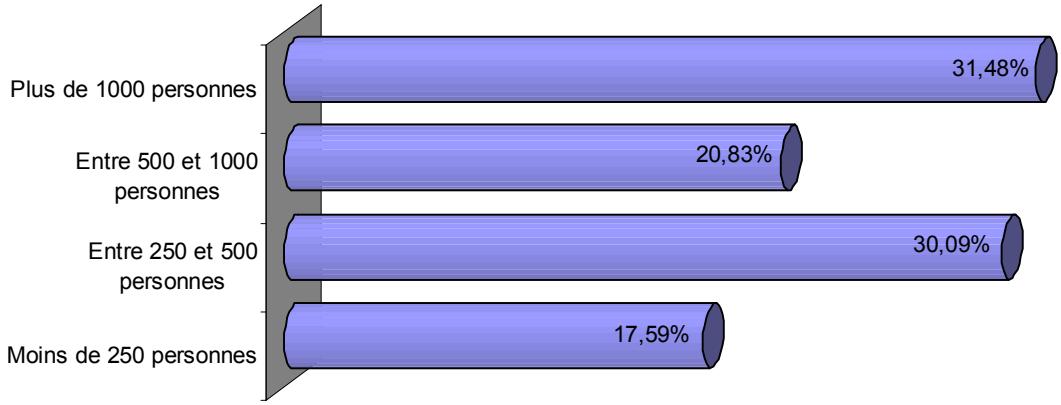
L'âge moyen des effectifs de plus de trois quarts des entreprises ayant répondu aux sondages est compris entre 35 ans et 45 ans.

Figure 25 *– Répartition des répondants par ancienneté moyenne*



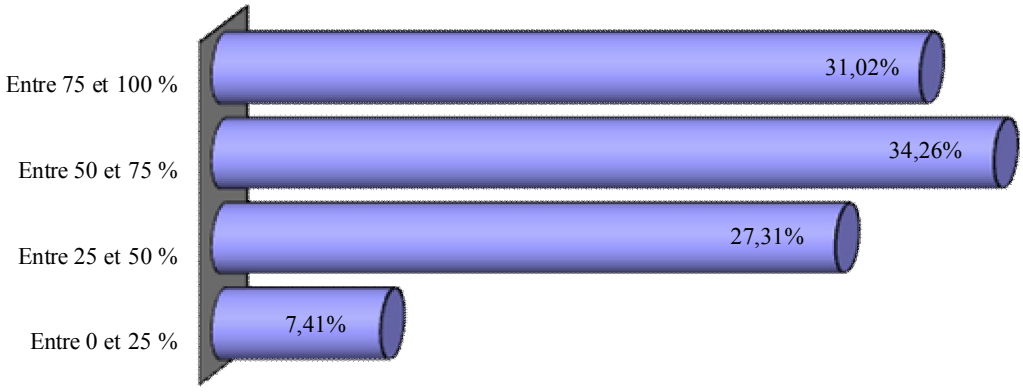
L'ancienneté moyenne des effectifs pour plus de 60 % des entreprises ayant répondu aux sondages dépasse 11 ans.

Figure 26 *– Répartition des participants par effectifs*



La composition de l'échantillon en fonction de leurs effectifs est relativement homogène. Les entités de moins de 250 personnes ont participé aux sondages à hauteur de 18 % et celles de plus de 1 000 salariés représentent un tiers du panel.

Figure 27 *– Répartition des participants selon la proportion homme/femme*



Pour plus de 65% des répondants, leurs effectifs sont constitués de plus de 50 % d'hommes.

CHAPITRE 3 : ANALYSE DESCRIPTIVE DE LA VARIABLE A EXPLIQUER

Cette étape a pour but principal de dégager les lois continues permettant d'ajuster convenablement la variable à expliquer TauxAbs, ainsi que la détermination des variables explicatives significatives.

Section 3.1. Méthode de calcul de la variable TauxAbs

Comme dans la partie précédente, nous avons retenu comme unité de temps les jours calendaires soit tous les jours de la semaine y compris les jours fériés. Par conséquent, tous les jours de l'année sont inclus dans le temps travaillé, les congés payés et RTT non ôtés. La formule de calcul devient la suivante :

$$\text{Taux d'absentéisme} = \frac{\text{Nombre de jours calendaires d'absence sur l'année} * 100}{\text{Nombre de jours calendaires sur l'année} * \text{Effectif moyen}}$$

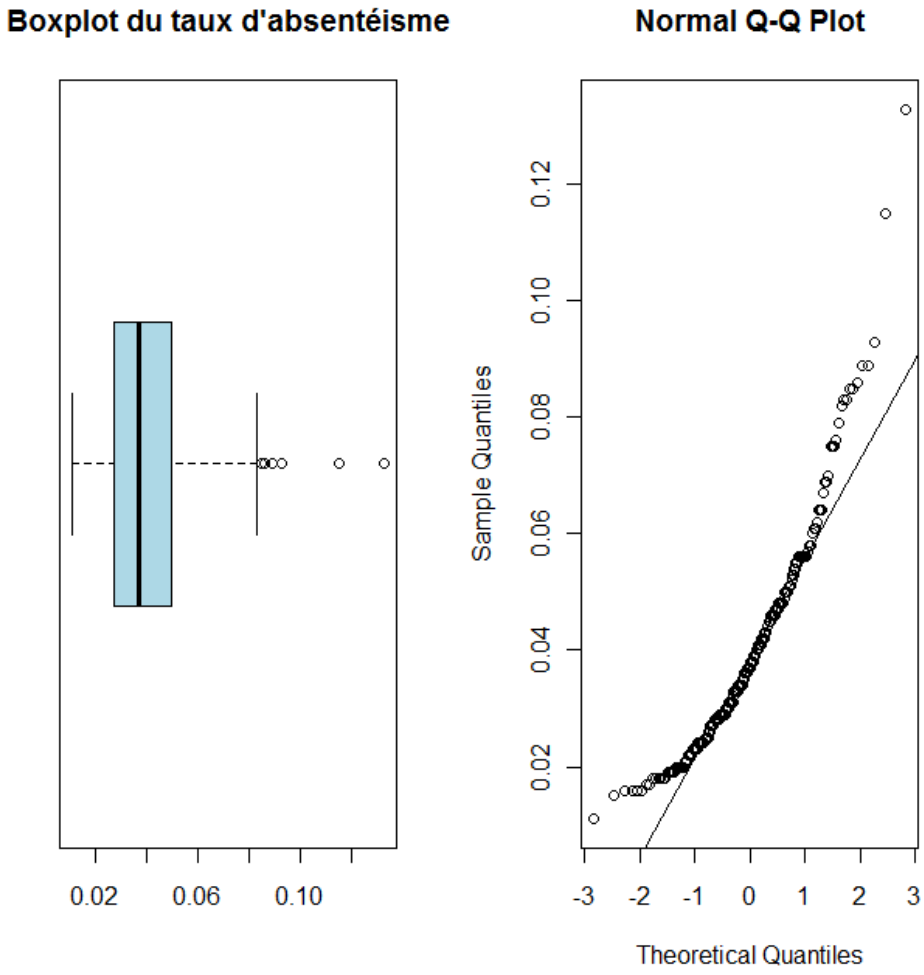
Dans la partie précédente, il nous était impossible de calculer un taux d'absentéisme du fait du manque de données concernant les salariés jamais été absents sur notre période d'observation. Nous nous limitions donc à décrire les caractéristiques des absences des gens absents, sans prétendre mesurer la charge de l'absentéisme pour l'entreprise.

Dans cette partie, nous disposons du nombre de jours d'absences sur l'année pour chaque entreprise participante aux sondages, nous pouvons donc modéliser la charge de l'absentéisme pour une entreprise.

Section 3.2. Quelques graphiques

A partir de nos données, nous avons tracé le boxpplot ainsi que le qq-plot de notre variable à expliquer, le taux d'absentéisme :

Figure 28 *– Le taux d'absentéisme*



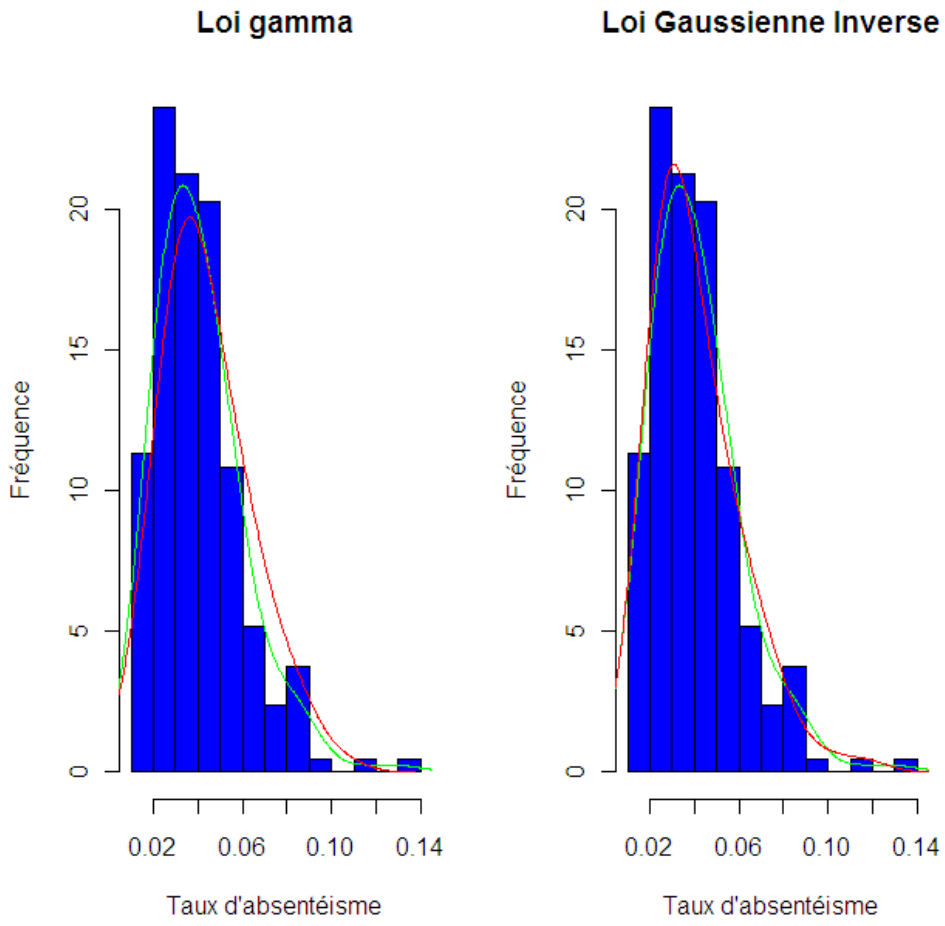
A partir du boxplot, on remarque que 50% des participants aux sondages ont un taux d'absentéisme relativement faible, compris entre 0 et 6% environ, ce qui est plus faible que dans la partie précédente, où notre étude se basait sur une seule entreprise. Le fait d'étudier l'absentéisme au niveau national sur plusieurs entreprises permet de lisser les résultats. On peut parler d'homogénéisation du risque, le risque ici étant l'absentéisme en entreprise.

D'autre part, à partir du qq-plot, on observe clairement que le taux d'absentéisme suit davantage une distribution normale que dans la partie précédente.

Section 3.3. Recherche de lois compatibles avec nos données

Notre variable à expliquer étant continue, nous allons nous intéresser aux familles de lois gamma et inverse gaussienne.
Afin de voir à quelle loi notre variable se rapproche le plus, nous avons tracé l'histogramme du taux d'absentéisme (qui permet d'obtenir l'allure de la densité de la loi) :

Figure 29 *– Histogrammes du taux d'absentéisme*



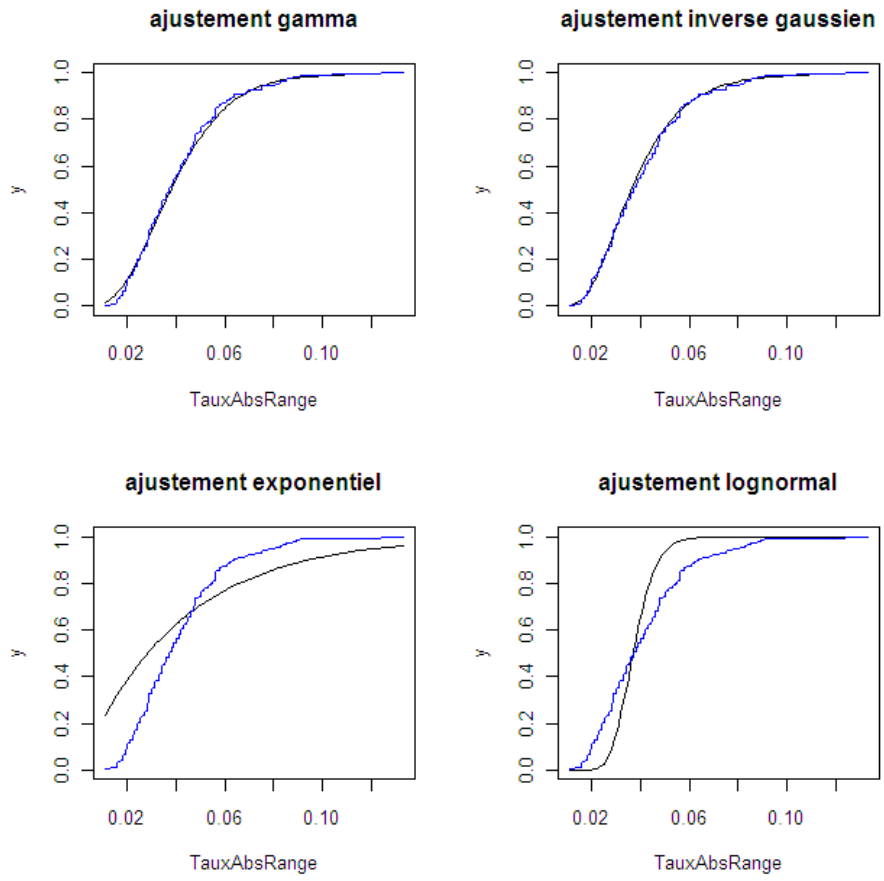
avec en rouge la densité des lois testées et en vert l'allure de la densité de la loi de notre variable à expliquer.

Tout comme dans notre précédente étude, on remarque que la loi inverse gaussienne se rapproche plus de notre variable à expliquer que la loi gamma.

De la même manière que pour la comparaison des fonctions de répartition, nous allons afficher le nuage de points correspondant à chacune des lois théoriques après en avoir estimé les paramètres.

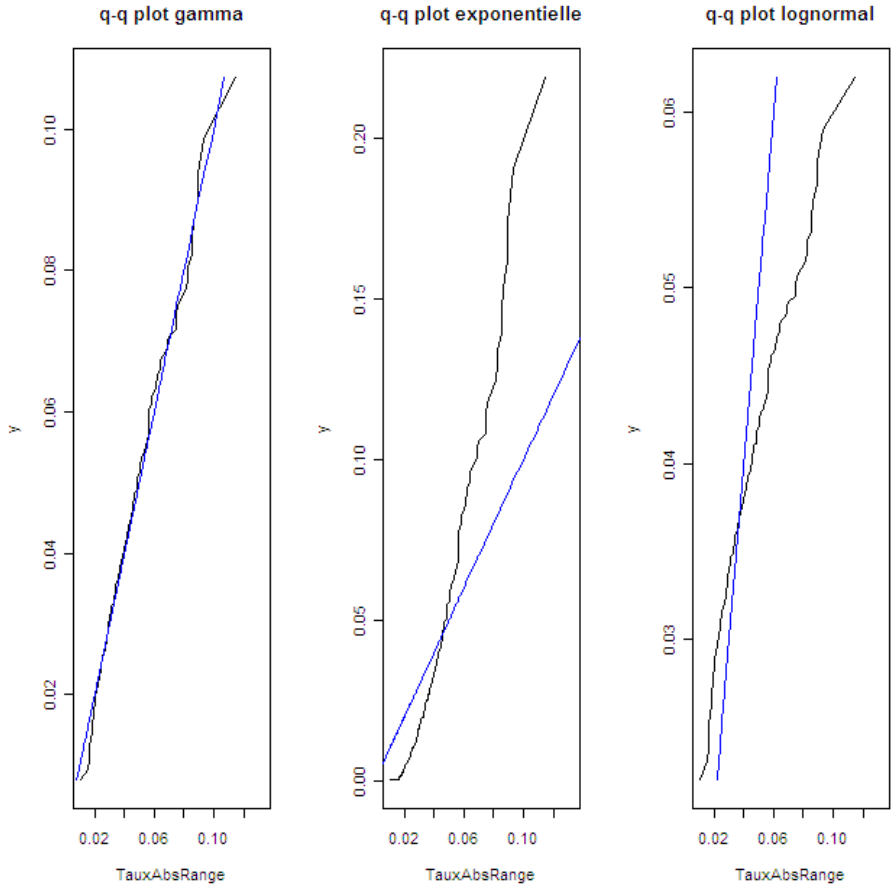
Dans un premier temps figure la comparaison des fonctions de répartition empirique et théorique. La fonction de répartition empirique de la variable à expliquer est en bleu et la fonction de répartition théorique est en noir.

Figure 30 *– Ajustements du taux d'absentéisme*



On remarque que ces courbes sont moins continues en comparaison aux graphiques similaires de la partie 3, ce qui est essentiellement due au nombre de données moins important dans cette partie.

Figure 31 *– Q-q plot du taux d'absentéisme*



On peut ici remarquer que la loi gamma ajuste particulièrement bien les données car le nuage de points q-q plot est relativement bien confondu avec la première bissectrice, ce qui n'est pas le cas concernant les autres lois.
Par conséquent, comme dans la partie précédente, nous allons exclure les lois exponentielles et log-normales de notre étude.

Section 3.4. Étude visuelle de l'impact des variables qualitatives

Les données dont nous disposons nous conduisent à retenir a priori les variables explicatives suivantes :

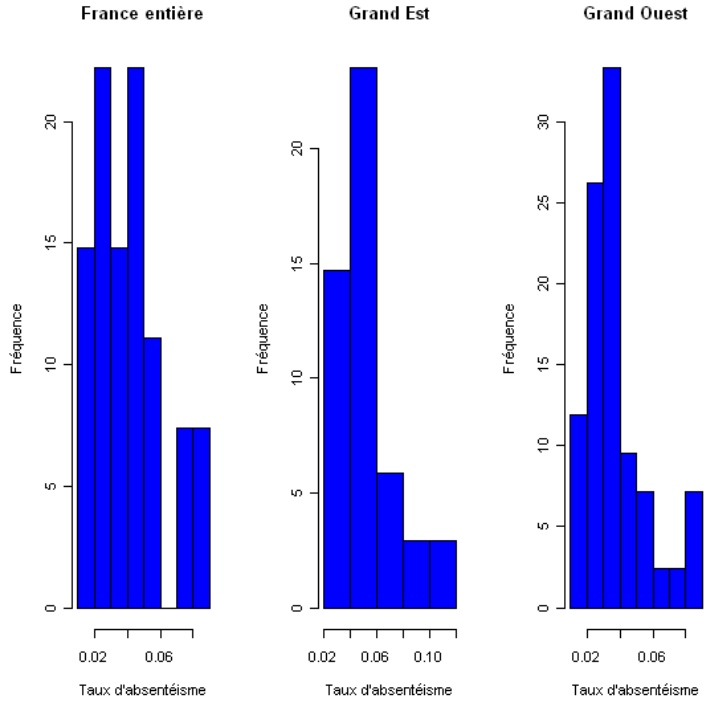
- La région
- Le secteur d'activité
- Le statut
- L'âge moyen
- L'ancienneté moyenne
- L'effectif
- La proportion d'hommes

3.4.1. *Paramétrage des variables explicatives*

Nous segmentons nos variables explicatives en différentes classes. Une fois cette segmentation effectuée, on obtient des variables qualitatives, ce qui nécessite un traitement particulier.
Nous retenons les mêmes hypothèses que dans la partie précédente.
Pour chaque modalité de ces variables, nous allons comparer les histogrammes, qui permettent d'obtenir l'allure de la densité de la loi.

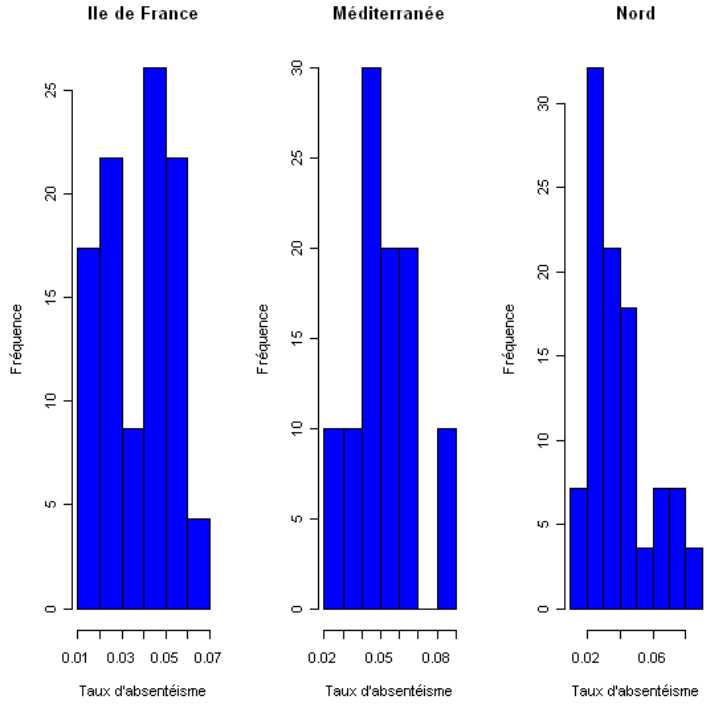
3.4.2. *La variable Région*

Figure 32 *– Histogrammes de la variable Région*



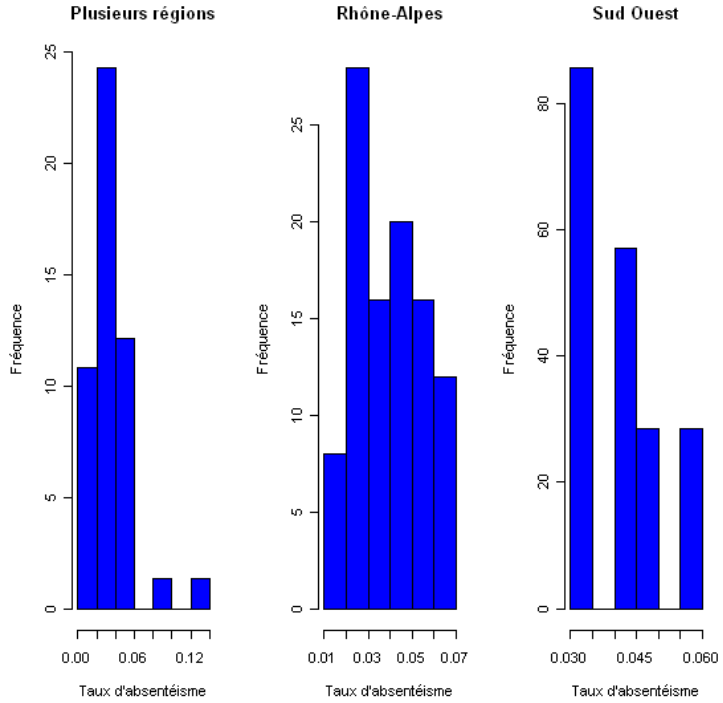
La région Grand Est a un plus fort taux d'absentéisme (5,28 %) que la Région Grand Ouest (3,85 %) et que l'ensemble des régions françaises réunies (4,18 %).

Figure 33 *– Histogrammes de la variable Région*



La Méditerranée présente un plus fort taux d'absentéisme (5,23 %) que l'Ile de France (3,90 %) et le Nord (4,01 %).

Figure 34 *– Histogrammes de la variable Région*

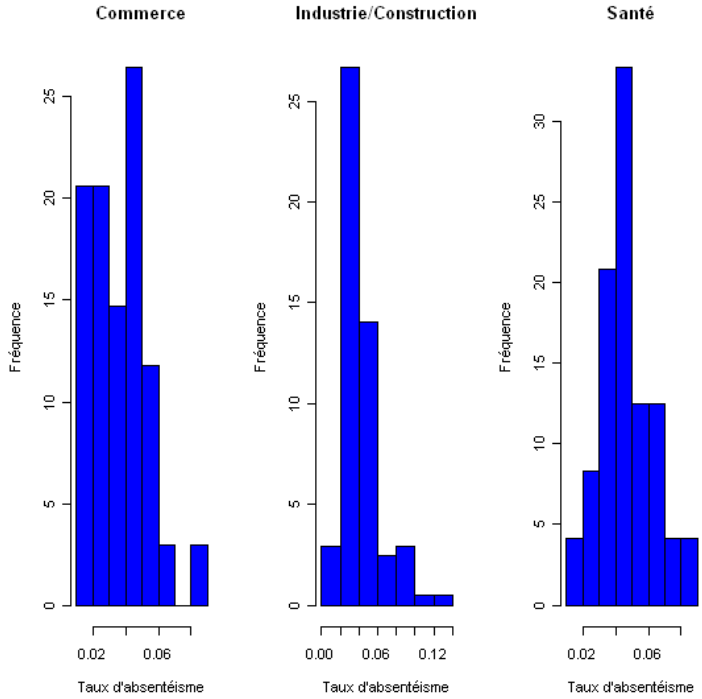


Le Sud Ouest présente un plus fort taux d'absentéisme (4,07 %) que le Rhône-Alpes (4,02 %) et que plusieurs régions réunies (3,63 %).

L'ensemble de ces histogrammes nous permet de conclure que les entreprises implantées dans le Sud et l'Est de la France présentent un taux d'absentéisme plus élevés. D'autre part, une entreprise implantée dans plusieurs régions aura un plus faible taux d'absentéisme que si elle est présente dans une seule région.

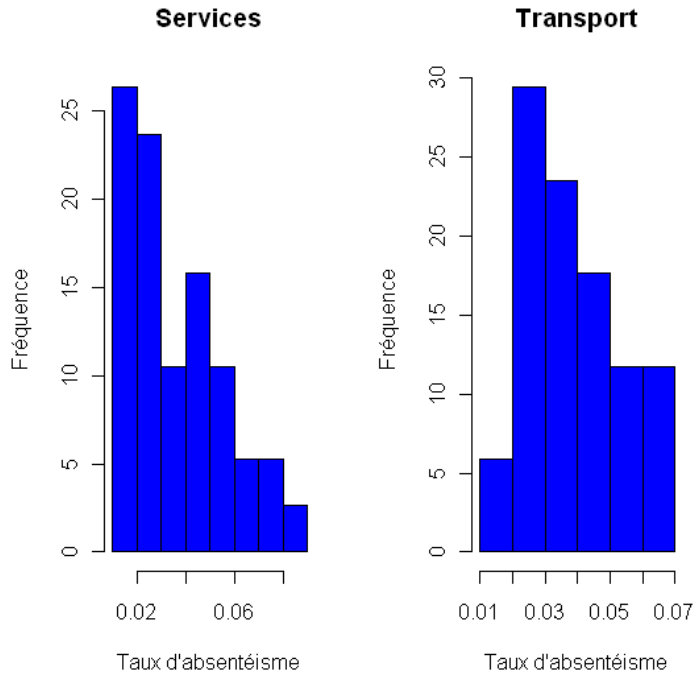
3.4.3. La variable Secteur d'activité

Figure 35 *– Histogrammes de la variable Secteur d'activité*



Le secteur de la santé (4,68 %) est plus touché par l'absentéisme que les secteurs du commerce (3,78 %) et de l'industrie/construction (4,21 %).

Figure 36 *– Histogrammes de la variable Secteur d'activité*

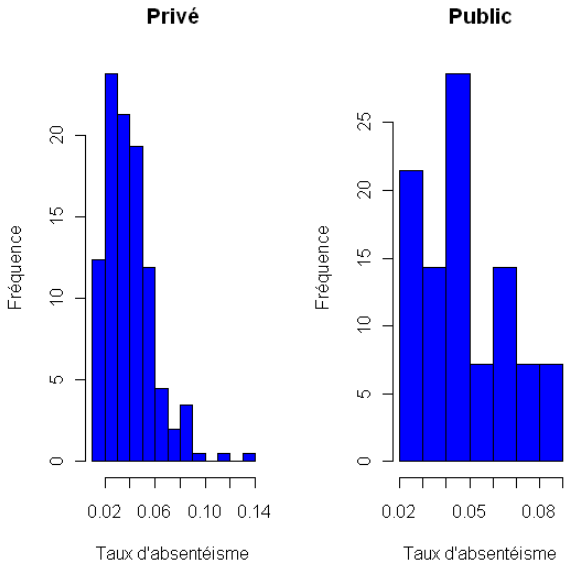


Le secteur des transports (3,88 %) est quant à lui légèrement plus touché par l'absentéisme que le secteur des services (3,74 %).

On note que le secteur d'activité le plus touché par l'absentéisme en entreprise est le secteur de la santé. Ce secteur concentre une forte population salariée féminine puisqu'il emploie 73 % de femmes.
D'autre part, les secteurs de l'industrie/construction et du transport font partie des secteurs les plus touchés par l'absentéisme. Or, des disparités dans la répartition entre les hommes et les femmes au travail perdurent dans ces secteurs où moins d'un salarié sur cinq est une femme.
On constate finalement que lorsque les effectifs hommes/femmes ne s'équilibrent pas, l'absentéisme est plus important.

3.4.4. *La variable Statut*

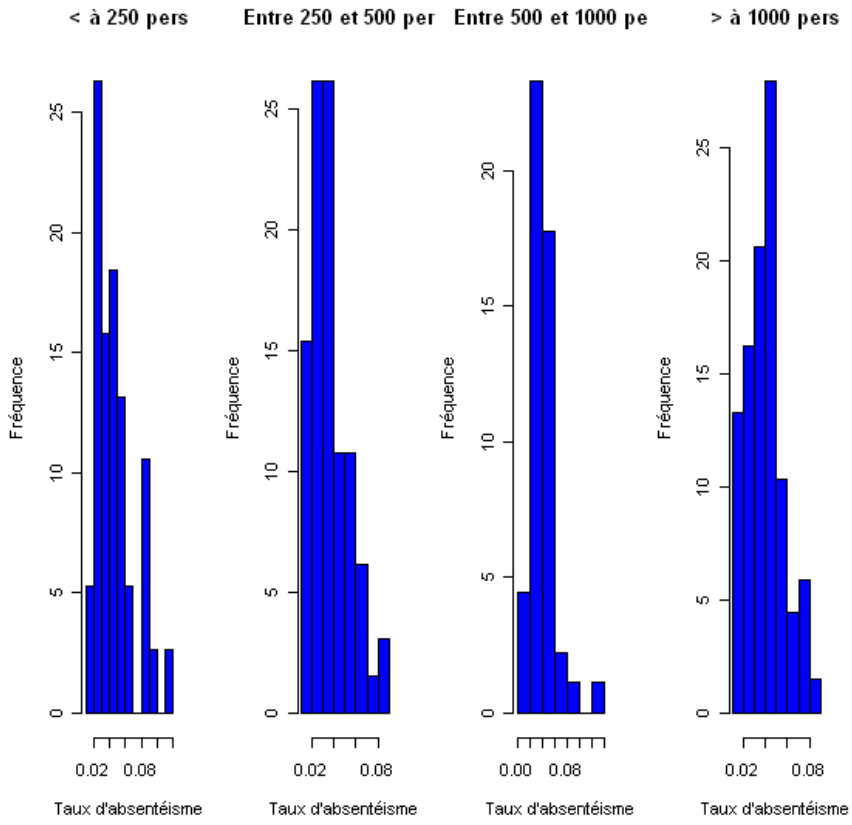
Figure 37 *– Histogrammes de la variable Statut*



Le secteur public (4,81 %) est davantage confronté à l'absentéisme que le secteur privé (4,03 %).

3.4.5. *La variable Effectifs*

Figure 38 *– Histogrammes de la variable Effectifs*

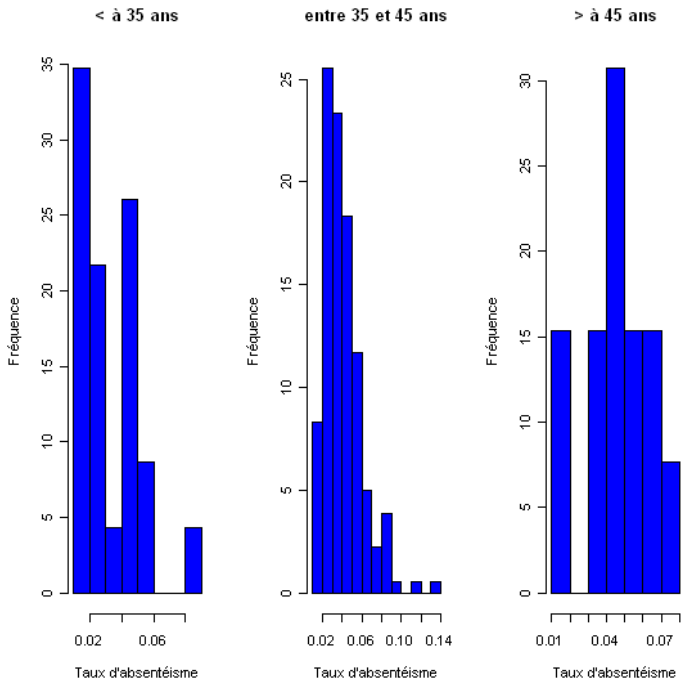


Les entreprises avec des effectifs inférieurs à 250 personnes (4,74 %) sont davantage confrontées à l'absentéisme que les entreprises avec des effectifs supérieurs. En effet, les entreprises de 250 à 500 personnes ont un taux d'absentéisme de 3,76%, celles de 500 à 1 000 personnes ont un taux de 4,06 % et les entreprises avec des effectifs supérieurs à 1 000 personnes un taux de 4,04 %.

Il est important de noter que la taille de l'entreprise est à mettre en lien avec l'égalité entre hommes et femmes. Plus les effectifs sont importants, plus les effectifs hommes/femmes s'équilibrent.
Les entreprises dont l'effectif est supérieur à 1 000 personnes atteignent l'équilibre entre hommes et femmes. A contrario, les entreprises de moins de 500 salariés comptent un peu plus d'un quart de femmes dans leurs effectifs.
Là encore, on constate donc que lorsque les effectifs hommes/femmes ne s'équilibrent pas, l'absentéisme en entreprise est plus important.

3.4.6. La variable Age moyen

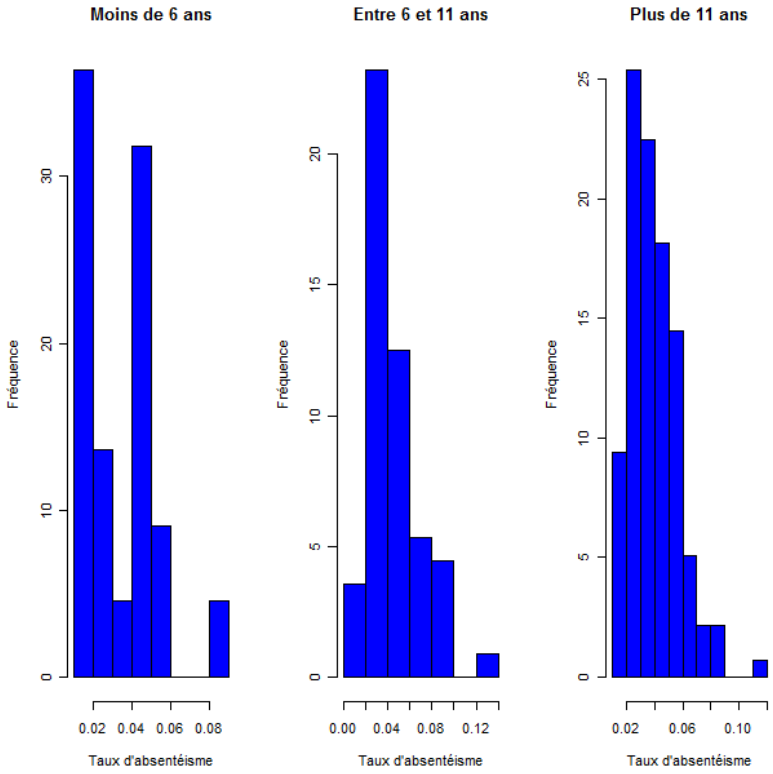
Figure 39 *– Histogrammes de la variable Age moyen*



On constate que plus l'âge moyen des effectifs d'une entreprise est élevé, plus celle-ci est confrontée à un fort taux d'absentéisme. En effet, pour un âge moyen inférieur à 35 ans, on a un taux d'absentéisme moyen de 3,36 %. Pour un âge moyen compris entre 35 et 45 ans, on a un taux de 4,13 %. Et lorsque l'âge moyen est supérieur à 45 ans, le taux d'absentéisme moyen est de 4,70 %.

3.4.7. *La variable Ancienneté moyenne*

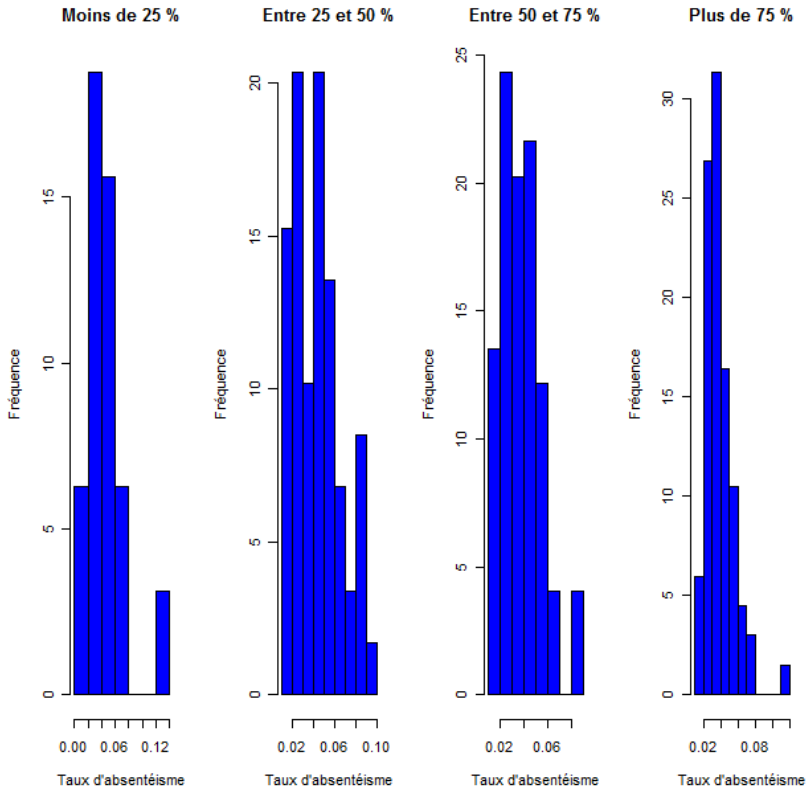
Figure 40 *– Histogrammes de la variable Ancienneté moyenne*



On constate que lorsque l'ancienneté moyenne des effectifs d'une entreprise est supérieure à 6 ans, celle-ci est confrontée à un plus fort taux d'absentéisme. En effet, pour une ancienneté moyenne inférieure à 6 ans, on a un taux d'absentéisme moyen de 3,52 %. Pour une ancienneté moyenne comprise entre 6 et 11 ans, on a un taux de 4,44 %. Et lorsque l'ancienneté moyenne est supérieure à 11 ans, le taux d'absentéisme moyen est de 4,03 %.

3.4.8. *La variable Proportion d'hommes*

Figure 41 *– Histogrammes de la variable Ancienneté moyenne*



Les entreprises avec moins de 25 % d'hommes (4,63 %) sont davantage confrontées à l'absentéisme que les entreprises avec une proportion d'hommes plus importante. En effet, les entreprises ayant entre 25 et 50 % d'hommes ont un taux d'absentéisme de 4,44 %, celles ayant entre 50 et 75 % d'hommes ont un taux de 3,87 % et les entreprises avec plus de 75 % d'hommes un taux de 3,88 %.

CHAPITRE 4. MISE EN ŒUVRE DES MODELES LINEAIRES GENERALISES

Dans ce chapitre, nous mettons en œuvre les modèles linéaires généralisés aux données issues des sondages 2008 et 2009.

Section 4.1. Sur l'ensemble des données

4.1.1. Le test du log rank

Rappelons que le test du log rank compare les estimations des fonctions de hasard de deux échantillons à chaque temps d'évènement observé. Sous l'hypothèse nulle, les deux courbes de survie ne sont pas différentes. Ici, seuls les facteurs qualitatifs peuvent être pris en compte.
Pour chaque variable explicative, on fixe une modalité comme la population de référence, par exemple celle contenant le plus de monde. On prend ensuite chaque modalité et on teste si sa distribution de durée est égale à celle de la population de référence. On peut alors faire une liste de modalités à regrouper avec la modalité de référence.

1) La variable Région

On choisit comme modalité de référence Reg3 (Grand Ouest) et on la compare aux autres modalités, les huit autres régions. Sur R, on utilise la fonction survdiff pour faire ce test. Ci-après, un exemple de résultat :

Tableau n°21 *– Résultats du test du log rank pour la variable CSP*

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Reg1	27	27	28,8	0,1125	0,206
Reg3	42	42	40,2	0,0806	0,206

Valeur de la statistique = 0,2 et p-value = 0,65

où : N est le nombre de personnes dans chaque groupe
Observed le nombre de personnes observé dans chaque groupe
Expected le nombre de personnes attendu dans chaque groupe
Chisq la statistique du chi deux pour un test d'égalité

On constate que l'on peut regrouper la Reg1 avec la Reg3 car notre p-value est supérieure à 5 % donc on accepte l'hypothèse nulle : Reg3 et Reg1 ont la même distribution de durée.

Nous réitérons ces tests pour l'ensemble des régions et à l'issue de ces tests, nous regroupons toutes les régions avec la région 3 à l'exception de la région Grand Est.
On a finalement deux modalités : Grand Est et Autres régions.

2) La variable Secteur d'activité

Ici, Sec2 (Industrie/Construction) est notre variable de référence. Après avoir effectué nos tests, toutes les variables peuvent être regroupées avec notre variable de référence.
Nous retirons donc cette variable de nos variables explicatives car elle ne présente plus qu'une seule modalité.

3) La variable Statut

Stat1 est notre variable de référence (Privé), et on regroupe la modalité Stat2 (Public) avec Stat1.
Comme précédemment, nous retirons cette variable de nos variables explicatives.

4) La variable Effectifs

Notre variable de référence est Ef4 (Plus de 1 000 personnes). A l'issue des tests, on regroupe Ef2 et Ef3 avec notre variable de référence.
On a finalement deux modalités : Moins de 250 personnes et Plus de 250 personnes.

5) La variable Age Moyen

Age2 (Entre 35 et 45 ans) est la variable de référence. On regroupe Age2 et Age3 ensemble.
On a deux modalités : Moins de 35 ans et Plus de 35 ans.

6) La variable Ancienneté Moyenne

Anc3 est notre variable de référence (Plus de 11 ans).
A l'issue des tests, nous retirons cette variable de nos variables explicatives.

7) La variable Proportion d'hommes

Prop3 est notre variable de référence (Entre 50 et 75 %).
Comme précédemment, à l'issue des tests, nous retirons cette variable de nos variables explicatives.

A l'issue de nos tests du log rank, nous retenons les trois variables explicatives suivantes :
Régions : reg1 (France entière à l'exception de la région Grand Est) et reg2 (Grand Est),
Effectifs : ef1 (moins de 250 pers) et ef2 (plus de 250 personnes),
Age moyen : age1 (moins de 35 ans) et age2 (plus de 35 ans).

4.1.2. Les modèles linéaires généralisés

1) Choix du modèle

Comme dans la partie précédente, il apparait que le modèle le plus pertinent à l'égard du critère de minimisation de l'AIC est le modèle de famille gaussienne inverse avec fonction de lien inverse.

Tableau n°22 *– Choix du modèle*

Famille	Fonction de lien	AIC
Gamma	Log	- 1 157,6
Gamma	Inverse	- 1 157,9
Gamma	Identité	- 1 157,3
Inverse gaussienne	Inverse	- 1 164,3

Il est important de noter que quel que soit le modèle testé, la variable explicative Effectifs n'est pas significative. On la retire donc de notre modèle.
Comme précédemment, nous utilisons la méthode *backward* afin de garder dans notre modèle uniquement les variables explicatives significatives.

3) Sélection des variables par la méthode *backward*

On a, sur R, les résultats suivants :

Tableau n°23 *– Résultats de la méthode backward*

	Coefficients estimés	**Erreur standard**	**Statistique**	**P-value**
Intercept	24,6113	0,8501	28,952	< 2e-16
Reg2	- 6,1775	2,5596	- 2,413	0,0166
Age1	5,7976	2,7348	2,120	0,0352

On remarque que toutes nos modalités sont significatives car l'ensemble des p-value est inférieure à 5 %, ce qui signifie qu'on rejette l'hypothèse nulle : les modalités n'ont pas d'impact sur la variable à expliquer.

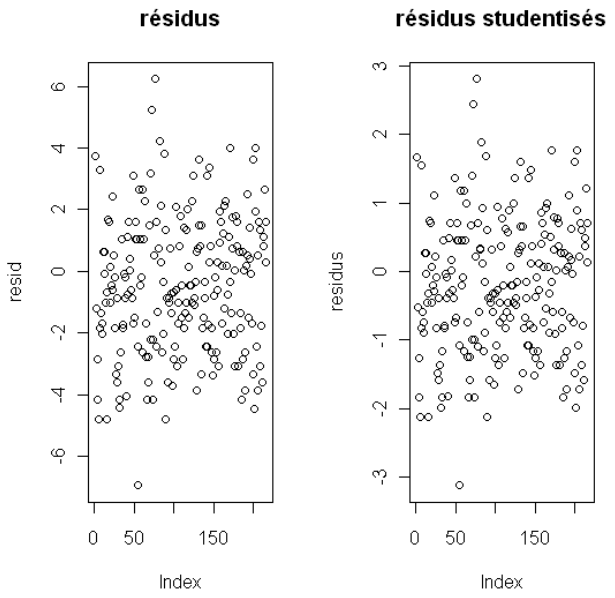
D'autre part, on remarque ici la modalité Reg2 (Grand Est) a tendance à augmenter le taux d'absentéisme, ce qui n'est pas le cas de que la modalité Age1 (Moins de 35 ans). Autrement dit, une entreprise avec une population salariée jeune sera moins exposée à l'absentéisme en entreprise. A l'inverse, une entreprise implantée dans l'Est de la France sera davantage confrontée à l'absentéisme en entreprise.

4.1.3. *Analyse des résidus*

Le test de Durbin-Watson nous indique que les résidus sont indépendants car D=1,9499 est proche de 2.

On regarde maintenant nos résidus afin de valider notre modèle :

Figure 42 *– Histogrammes de la variable Age moyen*



Nos résidus sont centrés, homoscédastiques et on compte uniquement trois valeurs aberrantes, ce qui valide notre modèle.

Section 4.2. Prévisions

Rappelons que le modèle le plus adapté à nos données est le modèle inverse gaussien avec fonction de lien inverse. Cela signifie que l'on a :

$$\frac{1}{E(Y)} = \alpha_0 + \sum_{i=1}^{n} \alpha_i X_i$$

Ou encore :

$$E(Y) = \frac{1}{\alpha_0 + \sum_{i=1}^{n} \alpha_i X_i}$$

On peut donc calculer le taux d'absentéisme estimé :

$$TauxAbs = \frac{1}{24{,}6113 - 6{,}1775 * 1_{Reg2} + 5{,}7976 * 1_{Age1}}$$

avec les modalités suivantes :

Tableau n°24 *– Rappel de la signification des modalités*

Modalité	Equivaut à
Reg1	France entière à l'exception de la région Grand Est
Reg2	Grand Est
Age1	Moins de 35 ans
Age2	Plus de 35 ans

Prenons l'exemple d'une entreprise implantée en Bretagne avec une population salariée dont l'âge moyen n'excède pas 35 ans. Son taux d'absentéisme prédit est alors égal à :

$$TauxAbs = \frac{1}{24{,}6113 + 5{,}7976} = 3{,}29\%$$

Maintenant que notre modèle est stabilisé, nous pouvons l'utiliser sur nos données initiales afin de faire des prédictions de taux d'absentéisme. On a en exemple les résultats suivants :

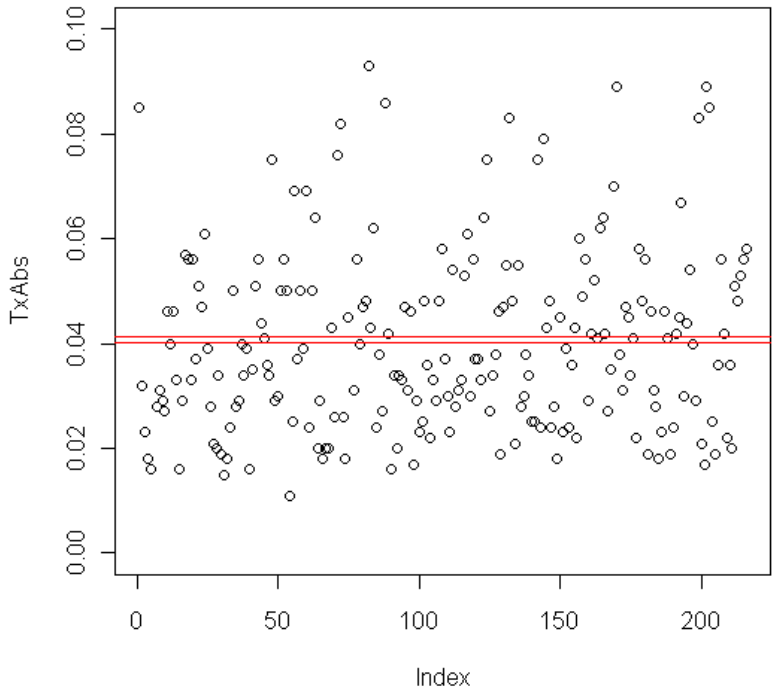
Tableau n°25 *– Exemple d'erreurs de prévision*

TauxAbs	Prévision	Résidus
2,9 %	3,3 %	-0,3 %
5,7 %	4,1 %	1,7 %
5,6 %	4,1 %	1,5 %
3,3 %	4,1 %	-0,8 %
5,6 %	5,4 %	0,2 %
3,7 %	4,1 %	-0,4 %
5,1 %	3,3 %	1,8 %

L'erreur de prévision s'écrit : $\varepsilon^P = Y - \hat{Y}^P$ et est centrée. En effet, en faisant la moyenne de nos erreurs de prévision, on obtient 0 %.

Il peut être intéressant de déterminer des intervalles de confiance autour des valeurs prédites et de positionner dedans les valeurs observées. La fonction t.test du langage R fournit la moyenne d'un vecteur ainsi que son intervalle de confiance.
On a par exemple les résultats suivants :
95 percent confidence interval : 0.04024918 0.04141098

Figure 39 *– Intervalle de confiance des valeurs prédites*



Il est utile de préciser que lorsqu'il y a plusieurs variables, des interactions sont possibles entre ces dernières. En toute rigueur, il aurait fallu tester dans ce mémoire les modèles avec interactions, mais nous avons choisi de ne pas le faire par souci de simplicité d'une part et aussi parce que les résultats obtenus avec les modèles sans interactions sont corrects.

CONCLUSION

L'absentéisme est un phénomène coûteux qui nécessite d'être pris en compte par les entreprises. Il n'est que partiellement réductible mais les absences incompressibles peuvent être gérées.

L'objet de ce mémoire a été de modéliser l'absentéisme en entreprise. L'utilisation des modèles linéaires généralisés nous a permis d'identifier, parmi les données disponibles celles qui influencent le plus sur l'absentéisme et de quantifier cette influence. Afin de permettre l'utilisation des modèles linéaires généralisés, l'absentéisme a été modélisé par des lois de probabilités usuelles pour au final n'en retenir qu'une. L'individu de référence a été choisi comme ayant des caractéristiques moyennes sur les variables explicatives retenues.

Une première étape a tout d'abord été de modéliser l'absentéisme au sein d'une entreprise et d'identifier l'influence des variables explicatives, telles que l'âge, l'ancienneté, le sexe, la catégorie socioprofessionnelle et le type de contrat, sur l'absentéisme. Cependant, ne disposant d'aucune information sur l'exposition, nous nous sommes limités dans cette partie à décrire les caractéristiques des absences des gens absents.

Dans un second temps, nous avons étudié les données concernant près de 200 entreprises réparties sur toute la France, avec pour objectif de calculer un taux d'absentéisme par entreprise selon diverses variables explicatives, comme notamment l'âge moyen dans l'entreprise, le secteur d'activité de l'entreprise, son statut, le nombre de salariés ainsi que la(les) région(s) où elles sont implantées. Contrairement à la première étape, nous avons ici pu mesurer la charge de l'absentéisme pour une entreprise.

Il pourrait être intéressant de mener cette étude avec davantage d'entreprises réparties sur toute la France dans le but d'affiner le modèle obtenu. Ceci implique donc un travail au niveau des sociétés de conseil mais aussi une coopération des entreprises. Il serait également intéressant d'étudier l'absentéisme des salariés au sein d'une entreprise, ce que nous n'avons pu faire dans ce mémoire faute d'informations suffisantes.

L'utilisation de tels modèles pourrait permettre de prédire un taux d'absentéisme pour une entreprise en fonction de ses caractéristiques et ainsi mettre en place des actions de gestion et de réduction de l'absentéisme adaptées. D'un point de vue commercial, prédire un taux d'absentéisme plus faible que ce qu'il n'est réellement à partir du modèle trouvé, pourrait signifier qu'il est possible de faire diminuer le taux d'absentéisme de cette entreprise.

BIBLIOGRAPHIE

BALLESTEROS S. [10 juin 2008] « Le modèle linéaire généralisé avec R : fonction glm() », SEMIN-R du MNHN.

CAPERAA et VAN CUTSEM [1988] « Méthodes et modèles en statistique non paramétrique : exposé fondamental », Presses Université Laval.

DROESBEKE J.J., LEJEUNE M., SAPORTA G. [2005] « Modèles statistiques pour données qualitatives », Editions Technip.

GALLOIS P. [2005] « L'absentéisme : Comprendre et agir », Editions Liaisons.

LAGADEC F. [2009] « Tarification d'un contrat complémentaire santé par un Modèle Linéaire Généralisé », Mémoire d'actuariat EURIA.

LETEURTRE H. [1991] « Audit de l'absentéisme du personnel hospitalier », Berger-Levrault.

MASIELLO E. [2010] « Les Modèles Linéaires Généralisés », Cours ISFA.

NELDER et WEDDERBURN [1972] "Generalized linear models", Journal of the Royal Statistical Society series A.

PIQUES C. [2004] « Vers la mise en place d'une stratégie de maîtrise de l'absentéisme au Centre Hospitalier de Montauban », Mémoire de l'ENSP.

PLANCHET F. [2009] « Modèles de durée », Cours ISFA.

POYET E. [2008] « Modélisation de l'effet consommateur en santé et application à la tarification », Mémoire d'actuariat ISFA.

RENAUD S., BELOUT A., ROCHELEAU I. [1999] « Les politiques de gestion de l'absence et leurs impacts sur l'absentéisme au travail », Ecole de relations industrielles, Université de Montréal.

ANNEXES

ANNEXE 1 : CODES R

1) Test du log rank

On détaille ici un exemple de codage du test du log rank sous le logiciel R pour tester quelle(s) modalité(s) regrouper avec la modalité de référence d'une variable explicative. Dans notre exemple, nous testons une modalité de la variable explicative « Proportion d'hommes au sein de l'entreprise » avec la modalité de référence.

```
donnees<-read.table("C:/Desktop/Mémoire/Données/Test log rank/PropHommes/Données
2008&2009 R Prop1.txt",header=TRUE,dec=",")
TxAbs<-donnees[,1]
Regions<-donnees[,2]
Secteur<-donnees[,3]
Statut<-donnees[,4]
Effectif<-donnees[,5]
AgeMoyen<-donnees[,6]
PropHommes<-donnees[,7]
AncMoyenne<-donnees[,8]
survdiff(formula=Surv(TxAbs)~PropHommes,data=donnees)
```

2) Méthode backward

On détaille ici un exemple de codage d'utilisation de la méthode *backward* dont le but est de garder dans notre modèle uniquement les variables explicatives significatives.

```
# Procédure ‘‘backward’’
# Partir du modèle complet :
pr.glm = glm(TxAbs~Regions+AgeMoyen+Effectifs,data=donnees,
subset=(TxAbs>0)&(TxAbs<1),family=inverse.gaussian(link="inverse"))
anova(pr.glm,test="Chisq")
# Procédure automatique descendante
pr.step <- stepAIC(pr.glm, trace = FALSE)
pr.step$anova # variables supprimées
```

ANNEXE 2 : RESULTATS R DU MODELE PARTIE 3

1) Significativité des variables

```
Call:
glm(formula = TauxAbs ~ Sexe + CatAge + CatAncienneté, family = inverse.gaussian(link =
"inverse"),
   data = donnees, subset = (TauxAbs > 0) & (TauxAbs < 1))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.8117  -4.6999  -2.0133   0.8024   8.9945

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      14.8543     0.3970  37.412  < 2e-16 ***
SexeS2           -4.2109     0.7572  -5.561 2.93e-08 ***
CatAgeAge4       -3.5271     1.0442  -3.378  0.00074 ***
CatAncienneténc1   3.4943     0.7614   4.589 4.64e-06 ***
CatAncienneténc3   2.5714     0.8496   3.027  0.00249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for inverse.gaussian family taken to be 16.98553)

   Null deviance: 53273  on 2912  degrees of freedom
Residual deviance: 52027  on 2908  degrees of freedom
AIC: -10801

Number of Fisher Scoring iterations: 2
```

2) Anova :

Analysis of Deviance Table

Model: inverse.gaussian, link: inverse

Response: TauxAbs

Terms added sequentially (first to last)

```
             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                         2912      53273
Sexe          1   588.38      2911      52685 3.966e-09 ***
CatAge        1   223.11      2910      52462 0.0002898 ***
CatAncienneté 2   434.53      2908      52027 2.785e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3) **Procédure *backward* :**

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
TauxAbs ~ Sexe + CatAge + CatAncienneté

Final Model:
TauxAbs ~ Sexe + CatAge + CatAncienneté

```
  Step Df Deviance Resid. Df Resid. Dev       AIC
1                    2908    52027.28 -10801.47
```

ANNEXE 3 : RESULTATS R DU MODELE PARTIE 4

1) Significativité des variables

```
Call:
glm(formula = TxAbs ~ Regions + AgeMoyen, family = inverse.gaussian(link = "inverse"),
    data = donnees, subset = (TxAbs > 0) & (TxAbs < 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9534  -2.0286  -0.5327   1.0311   6.2335

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.6113     0.8501  28.952   <2e-16 ***
Regionsreg2   -6.1775     2.5596  -2.413   0.0166 *
AgeMoyenage1   5.7976     2.7348   2.120   0.0352 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for inverse.gaussian family taken to be 5.2751)

    Null deviance: 1156.2  on 215  degrees of freedom
Residual deviance: 1101.6  on 213  degrees of freedom
AIC: -1164.2

Number of Fisher Scoring iterations: 2
```

2) Anova :

```
Analysis of Deviance Table

Model: inverse.gaussian, link: inverse

Response: TxAbs

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      215     1156.2
Regions  1   30.894       214     1125.3   0.01552 *
```

AgeMoyen 1 23.677 213 1101.6 0.03412 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

3) Procédure *backward* :

Stepwise Model Path
Analysis of Deviance Table

Initial Model:
TxAbs ~ Regions + AgeMoyen

Final Model:
TxAbs ~ Regions + AgeMoyen

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
1			213	1101.619	-1164.164