

PROJET GLM

Prédiction des coûts médicaux

ESILV – ACT3

Léon-Paul Dufour, Alix Gleizes, Julien Glon, Max Prugnaud, Achille Robin



Introduction

Dans le cadre de notre cours de *Generalized Linear Models* (GM) en M1 Actuariat à l'ESILV, nous avons mené un projet visant à appliquer concrètement les méthodes étudiées en cours à un jeu de données réels. L'objectif est de comprendre comment différents facteurs individuels influencent les dépenses médicales, et de construire un modèle prédictif cohérent et interprétable.

Le dataset utilisé rassemble plusieurs variables socio-démographiques et comportementales ; l'âge, le sexe, l'indice de masse corporelle (BMI), le statut de fumeur ou non-fumeur, le nombre d'enfants à charge ainsi que la région de résidence. La variable cible correspond au montant des dépenses médicales annuelles pour chaque individu. Ces données présentent des caractéristiques typiques des coûts d'assurance : montant strictement positifs, forte asymétrie, dispersion importante et variance croissante avec la moyenne.

La problématique de ce projet est donc la suivante : Peut-on prédire le montant des dépenses médicales à partir de ces caractéristiques individuelles et quels sont les facteurs qui influencent le plus ces coûts ? Pour répondre à cette question, l'utilisation d'un GLM s'impose. Ce type de modèle est particulièrement adapté aux données positives et asymétriques, comme les charges médicales, et permet en outre une interprétation directe des effets des variables explicatives.

Ce projet s'inscrit ainsi pleinement dans une démarche actuarielle : analyser un risque, sélectionner un modèle adapté et interpréter les effets des variables pour mieux comprendre les déterminants du coût médical.

Sommaire

I - Description du dataset.....	4
II - Choix du GLM	5
III - Ajustement & sélection du modèle.....	6
IV - Interprétation.....	7
V - Comparaison de modèle	8
VI - Conclusion	9

I - Description du dataset

Le jeu de données utilisés dans ce projet est le fichier `medical_insurance.csv`, chargé et exploré dans notre notebook Python. Il contient 2772 individus et 7 variables initiales. Les variables sont les suivantes :

- Age : âge de l'individu (entier)
- Sex : sexe, codé *male ou female*
- Bmi : indice de masse corporelle (float)
- Children : nombre d'enfants à charge (entier)
- Smoker : statut de fumeur (*yes ou no*)
- Region : région d'habitation (*northwest, northeast, southeast, southwest*)
- Charges : dépenses médicales annuelles (float), notre variable cible

Le code montre que le dataset ne contient aucune valeur manquante et que les types sont cohérents avec une analyse GLM.

Les statistiques descriptives affichées indiquent que l'âge moyen est d'environ 39 ans, le BMI moyen de 30.7, et les charges médicales variant fortement, avec une moyenne autour de 13 261\$ et un maximum supérieur à 63 000\$. Le statut de fumeur présente un déséquilibre marqué (564 fumeurs contre 2208 non-fumeurs), ce qui constitue déjà un signal de variabilité importante.

Afin de visualiser la distribution de la variable cible, nous avons représenté un boxplot et un histogramme des charges. Ces deux graphiques confirment la forte asymétrie des dépenses médicales, avec une longue traîne à droite et de nombreux outliers. Cette structure est typique de données de coûts et justifie l'utilisation d'un modèle Gamma avec lien logarithmique.

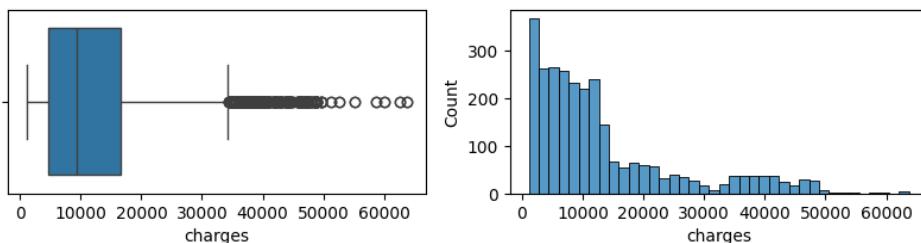


Figure 1 - Boxplot et histogramme des charges

Enfin, la matrice de corrélation entre les variables numériques (age, bmi, children, charges) montre des corrélations faibles à modérées. On observe notamment une corrélation positive entre l'âge et les charges (0.30) et entre le BMI et les charges (0.20), tandis que le nombre d'enfants n'est que faiblement lié aux dépenses. Cette matrice confirme l'absence de multicolinéarité problématique et soutient l'utilisation d'un GLM sur l'ensemble de ces variables.

Enfin, le preprocessing inclut un One-Hot Encoding des variables catégorielles (sex, smoker, region), créant 6 variables indicatrices telles que sex_male, smoker_yes ou region_northwest, intégrant dans un dataframe final prêt pour la modélisation GLM.

II - Choix du GLM

Nous avons d'abord ajusté un modèle linéaire classique afin d'évaluer s'il pouvait constituer une base satisfaisante pour la prédiction des dépenses médicales. L'analyse de certains composants nous ont montré que le modèle ne respectait pas les hypothèses fondamentales de la régression linéaire. Nous avons abandonné le modèles linéaire.

Le choix d'un modèle linéaire généralisé repose sur la nature de la variable réponse *Charges*, qui correspond à des coûts médicaux individuels. Ce type de variable présente trois caractéristiques essentielles :

- (i) Strictement positive
- (ii) Continue
- (iii) Distribution fortement asymétrique, avec de nombreux outliers et une longue traîne à droite.

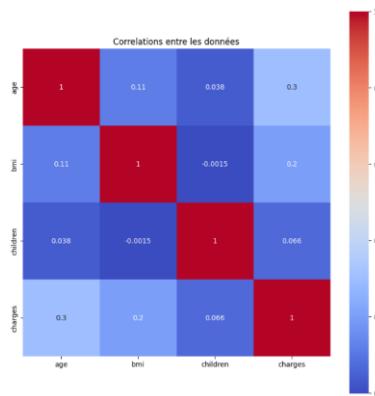


Figure 2 - Matrice de corrélation

Commented [RA1]: J'aurais ajouté qu'on a d'abord essayé un modèle linéaire puis qu'on l'a enlevé (renvoie vers ce que t'as écrit après)

Commented [GA2R1]: OK

Ces caractéristiques rendent la régression linéaire classique inadaptée : les hypothèses de normalité et d'homoscédasticité des résidus sont violées, comme l'ont confirmé les tests

préliminaires (Shapiro-Wilk, Kolmogorov-Smirnov) et les graphes Q-Q. Le code montre également une variance qui augmente avec la moyenne, ce qui est typique des données de coûts.

Dans ce contexte, les modèles linéaires généralisés (GLM) sont particulièrement adaptés.

La famille Gamma a été retenue car elle est conçue pour les variables continues positives. Elle modélise une variance proportionnelle au carré de la moyenne, ce qui correspond bien à la structure du dataset. Gamma est couramment utilisée en assurance santé et en actuariat pour la tarification des dépenses médicales.

Le lien logarithmique a été choisi car il garantit des prédictions strictement positives. De plus, il permet d'interpréter les coefficients en termes d'effets multiplicatifs. Pour finir, il stabilise la variance et rend la relation entre prédicteurs et moyenne plus linéaire.

C'est pour cela qu'on utilise un GLM Gamma avec lien log comme modèle principal.

III - Ajustement & sélection du modèle

L'ajustement du modèle a débuté par l'estimation d'un GLM Gamma avec lien logarithmique intégrant l'ensemble des variables explicatives, incluant les variables numériques (*age*, *bmi*, *children*) ainsi que les variables catégorielles transformées par One-Hot-Encoding (*sex*, *smoker* et les différentes régions). Le résumé du modèle complet montre que certaines variables apparaissent clairement significatives, notamment l'âge, le BMI et surtout le statut de fumeur, qui exerce un effet très marqué sur le niveau des charges médicales. À l'inverse, les variables régionales

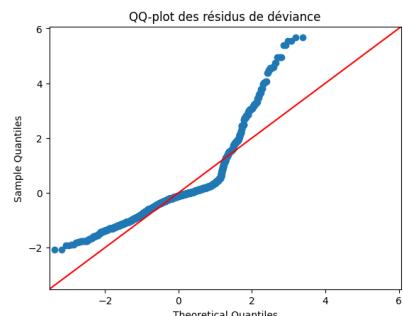


Figure 3 - QQ plot des résidus de déviance

modélise une variance proportionnelle au carré de

la moyenne, ce qui correspond bien à la structure du dataset. Gamma est couramment utilisée en assurance santé et en actuariat pour la tarification des dépenses médicales.

Le lien logarithmique a été choisi car il garantit des prédictions strictement positives. De plus, il permet d'interpréter les coefficients en termes d'effets multiplicatifs. Pour finir, il stabilise la variance et rend la relation entre prédicteurs et moyenne plus linéaire.

C'est pour cela qu'on utilise un GLM Gamma avec lien log comme modèle principal.

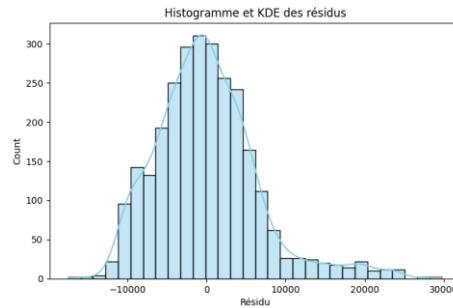


Figure 4 - Histogramme et KDE des résidus

et le nombre d'enfants ne semblent pas contribuer substantiellement à l'explication de la variabilité de la dépense.

Afin d'obtenir un modèle plus harmonieux, une procédure de sélection basée sur l'AIC a été appliquée. Cette sélection backward et forward identifie un sous-modèle optimisant le compromis entre qualité d'ajustement et simplicité :

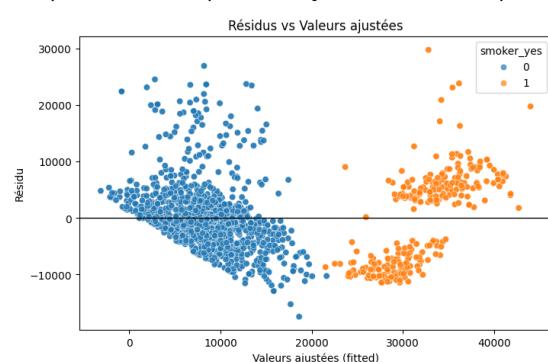


Figure 5 - Résidus vs valeurs ajustées

seules les variables *age*, *bmi*, et *smoker_yes* sont conservées. Le retrait des autres prédicteurs n'entraîne aucune dégradation significative de la qualité statistique du modèle. Cela montre que ces variables initialement incluses ne participaient pas réellement à améliorer la prédictivité. Les diagnostics réalisés sur ce modèle final, notamment l'analyse des résidus, confirment sa bonne spécification. Les résidus ne présentent pas de structure particulière, la variance suit un comportement compatible avec une distribution Gamma, et aucun point excessivement influent n'est détecté. Dans l'ensemble, le modèle retenu apparaît robuste, stable et statistiquement cohérent avec les caractéristiques du dataset.

Par ailleurs, compte tenu de la présence d'observations extrêmes dans les charges médicales, une analyse complémentaire a été menée à l'aide d'une distribution de Pareto généralisée (GPD) afin d'examiner la queue de distribution. Cette approche, couramment utilisée en modélisation des valeurs extrêmes, permet d'évaluer la pertinence d'un ajustement spécifique pour les très grands coûts. Les résultats suggèrent que, même si les outliers sont marqués, leur comportement reste compatible avec une queue lourde modérée, ce qui confirme que le GLM Gamma capture correctement la structure générale des données sans nécessiter un modèle séparé pour les extrêmes.

IV – Interprétation

L'interprétation d'un modèle Gamma avec lien logarithmique repose sur une lecture multiplicative des coefficients. Chaque coefficient traduit l'effet d'une variable sur la charge médicale moyenne après exponentiation, ce qui permet d'apprecier l'impact relatif des facteurs de risque. Dans ce modèle, l'âge exerce un effet positif et significatif : chaque année supplémentaire conduit à une augmentation progressive de la dépense attendue, ce qui reflète la dégradation naturelle de l'état de santé au fil du temps. Le BMI, également significatif, traduit l'influence des risques métaboliques et cardiovasculaires associés au

Commented [RA3]: Pas sex_male du coup

Commented [GA4R3]: ok

Commented [RA5]: Ajoute une ou deux phrases sur Pareto Généralisée pour les outliers, j'ai fait quelques lignes dessus dans le notebook

Commented [GA6RS]: ok

surpoids ; une valeur plus élevée du BMI correspond structurellement à une dépense médicale plus importante.

Le sexe apparaît dans le modèle, mais son effet n'est pas statistiquement significatif. Autrement dit, une fois contrôlés les autres facteurs comme l'âge, le BMI et le statut de fumeur, la différence moyenne de dépenses médicales entre hommes et femmes n'est pas suffisamment marquée pour être considérée comme un déterminant robuste du coût médical. Le coefficient associé à *sex_male* suggère une légère variation, mais celle-ci reste trop faible et trop incertaine pour être interprétée comme un véritable effet causal ou prédictif. L'effet le plus marquant demeure toutefois celui du statut de fumeur. Le coefficient associé à *smoker_yes* est nettement plus élevé que les autres paramètres et, une fois exponentié, conduit à un facteur multiplicatif supérieur à 2. Cela signifie que les fumeurs, dans ce dataset, engendrent des dépenses médicales en moyenne deux à trois fois plus importantes que les non-fumeurs, ce qui en fait le déterminant principal du modèle.

Les variables régionales et le nombre d'enfants, exclues du modèle final, ne montraient quant à elles aucun impact significatif sur la charge, confirmant qu'elles ne constituent pas des facteurs explicatifs pertinents après contrôle des effets principaux.

V - Comparaison de modèles

Plusieurs modèles ont été ajustés au cours de l'analyse afin de déterminer celui offrant la meilleure combinaison de performance et d'interprétabilité. La régression linéaire classique a servi de point de départ, mais les tests statistiques menés sur les résidus (notamment Shapiro-Wilk, Kolmogorov-Smirnov et Breusch-Pagan) ont immédiatement montré que les hypothèses de normalité et d'homoscédasticité étaient largement violées. Les résidus présentaient une structure non aléatoire ainsi qu'une variance clairement croissante, confirmant l'inadéquation de l'OLS pour modéliser des dépenses médicales.

Commented [RA7]: Nickel

La sélection de variables appliquée au modèle linéaire permettait d'identifier les facteurs principaux mais ne résolvait pas les problèmes structurels liés à la distribution de la variable réponse, ce qui rendait cette approche insuffisante. Le GLM Gamma log complet offrait déjà un ajustement nettement supérieur à celui du modèle linéaire, mais incluait des variables redondantes ou non significatives. Le modèle Gamma log final, obtenu après sélection par AIC, présentait un AIC plus faible, des coefficients mieux identifiés et une meilleure stabilité des prédictions. À ce titre, il représente le modèle statistiquement optimal dans le cadre de cette analyse.

Un modèle XGBoost a également été ajusté sur le dataset afin d'évaluer si une approche non paramétrique pouvait améliorer la performance prédictive. Celui-ci obtenait une amélioration du RMSE mais au prix d'une perte totale d'interprétabilité. Dans un

Commented [RA8]: En vrai l'amélioration est quand même flagrante notamment sur le R2 très proche de 1 (risque d'overfitting) mais oui on perd complètement l'explicabilité

Commented [GA9R8]: ok

contexte académique et assurantiel où l'explicabilité des modèles est essentielle, le GLM Gamma apparaît donc comme la solution la plus appropriée, même si des techniques de machine learning peuvent être utilement mobilisées pour un objectif purement prédictif. Par ailleurs, pour mieux appréhender le comportement des observations extrêmes, le GLM peut être complété par une analyse de queue au moyen d'un modèle de Pareto généralisée, permettant ainsi de caractériser de manière plus précise la distribution des coûts les plus élevés tout en conservant une structure explicative globale robuste.

VI – Conclusion

L'analyse menée sur le dataset *medical_insurance* montre qu'il est effectivement possible de prédire le montant des dépenses médicales à partir de caractéristiques individuelles simples. L'utilisation d'un modèle linéaire généralisé de type Gamma avec lien logarithmique s'avère particulièrement adaptée à la structure des données, caractérisées par des coûts strictement positifs, fortement asymétriques et présentant une variance croissante avec la moyenne. Après sélection des variables par AIC, le modèle final conserve trois déterminants majeurs : l'âge, le BMI et surtout le statut de fumeur.

Parmi ces facteurs, le statut de fumeur apparaît comme le plus influent, avec un effet multiplicatif très marqué sur les charges médicales attendues. Le BMI et l'âge jouent également un rôle significatif et cohérent avec la littérature médicale : un surpoids ou un vieillissement progressif s'accompagne d'une augmentation des dépenses. Les variables régionales et le nombre d'enfants n'ont pas montré d'impact substantiel une fois les autres covariables contrôlées.

Ainsi, les résultats confirment qu'un modèle explicatif et prédictif solide peut être construit à partir de ces informations, tout en conservant une interprétabilité essentielle dans un cadre assurantiel. Pour une analyse plus exhaustive et pour l'ensemble des visualisations complémentaires (diagnostics, distributions, comparaisons de modèles et graphiques interactifs), le lecteur est invité à consulter le notebook Python associé, où la totalité des traitements et figures est présentée en détail.

Commented [RA10]: Nickel

Commented [RA11]: Ajoute un truc du genre : couplée à un modèle de pareto généralisée

Commented [GA12R11]: ok