FACULTÉ DES SCIENCES DHAR EL MAHRAZ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH

UNIVERSITÉ
SORBONNE
PARIS NORD

Analyse des Frais Médicaux par Régression Linéaire Multiple

EL HILALI ANAS

et

ILYAS FAQIR

Encadré par :

Prof. Abdelkamel ALJ

Université Sidi Mohamed Ben Abdellah
Sorbonne Paris Nord University

1er février 2025

Table des matières

Table des figures

Liste des tableaux

1 Introduction

1.1 Contexte et Motivation

Dans un monde où les données jouent un rôle crucial dans la prise de décision, il est essentiel de comprendre les relations entre différentes variables et leurs impacts sur une variable cible. La régression linéaire multiple est un outil statistique puissant permettant d'identifier et de quantifier ces relations. Ce projet s'inscrit dans cette logique en cherchant à explorer les facteurs influençant les frais médicaux à partir d'un ensemble de données spécifiques.

1.2 Objectif de l'Étude

L'objectif principal de cette étude est d'analyser l'impact de plusieurs variables explicatives sur une variable dépendante en utilisant la régression linéaire multiple (RLM). Cette analyse nous permettra d'identifier les facteurs les plus influents et de construire un modèle prédictif fiable.

1.3 Questions de Recherche

— Quels sont les facteurs les plus déterminants influençant les frais médicaux ?
— Dans quelle mesure la régression linéaire multiple permet-elle de prédire les frais médicaux à partir des données disponibles ?

2 Cadre Théorique

2.1 Fondements de la Régression Linéaire Multiple

La régression linéaire multiple constitue une méthode statistique essentielle permettant de modéliser la relation entre une variable dépendante Y et un ensemble de variables explicatives $X_1, X_2, \ldots, X_k$. Le modèle s'exprime sous la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon,$$

où :

— β_0 est l'ordonnée à l'origine,
— $\beta_1, \beta_2, \ldots, \beta_k$ représentent les coefficients de régression mesurant l'impact de chaque variable explicative,
— ε est le terme d'erreur, supposé capturer toutes les variations inexpliquées par le modèle.

Ce modèle permet ainsi d'estimer l'influence individuelle de chaque facteur sur la variable cible, tout en tenant compte de l'effet simultané des autres variables.

2.2 Hypothèses du Modèle

Pour que les estimations soient fiables et interprétables, plusieurs hypothèses doivent être respectées :

— **Linéarité :** La relation entre la variable dépendante et chacune des variables explicatives est supposée linéaire.
— **Indépendance des erreurs :** Les termes d'erreur ε doivent être indépendants les uns des autres.
— **Homoscedasticité :** La variance des erreurs est constante pour toutes les valeurs des variables explicatives. En cas d'hétéroscédasticité, la précision des estimations peut être affectée.

— **Normalité des erreurs :** Les erreurs sont supposées suivre une distribution normale, condition indispensable pour la validité de nombreux tests statistiques.
— **Absence de multicolinéarité parfaite :** Les variables explicatives ne doivent pas être parfaitement corrélées entre elles, afin d'éviter des difficultés d'estimation des coefficients.

2.3 La Méthode des Moindres Carrés Ordinaires (MCO)

La méthode des moindres carrés ordinaires (MCO) est utilisée pour estimer les paramètres du modèle en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites :

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\right)^2 .$$

Cette méthode présente plusieurs avantages :
— **Non-biaisé :** Sous les hypothèses classiques du modèle, les estimateurs obtenus sont sans biais.
— **Efficacité (BLUE) :** D'après le théorème de Gauss-Markov, les estimateurs MCO sont les *Best Linear Unbiased Estimators* (BLUE), c'est-à-dire qu'ils possèdent la variance minimale parmi l'ensemble des estimateurs linéaires non biaisés.

2.4 Théorème de Gauss-Markov

Le théorème de Gauss-Markov affirme que, dans le cadre des hypothèses du modèle linéaire classique, les estimateurs MCO sont optimaux parmi les estimateurs linéaires non biaisés. Cette propriété est cruciale car elle garantit que l'on ne peut pas obtenir d'estimations plus précises (en termes de variance) en utilisant une autre méthode linéaire sans introduire de biais.

2.5 Critères d'Évaluation du Modèle

Une fois le modèle ajusté, il est essentiel d'en évaluer la qualité afin de s'assurer de sa pertinence. Parmi les indicateurs utilisés, on retrouve :
— **Coefficient de Détermination (R^2) :** Mesure la proportion de la variance de la variable dépendante expliquée par le modèle. Plus R^2 est proche de 1, meilleur est l'ajustement.
— R^2 **Ajusté :** Tient compte du nombre de variables explicatives et pénalise l'ajout de variables non pertinentes.
— **Erreurs de Prédiction (MSE, RMSE, MAE) :** Ces mesures quantifient l'écart moyen entre les valeurs observées et prédites, fournissant une indication de la précision des prédictions.

2.6 Extensions et Approches Alternatives

Bien que la régression linéaire multiple soit très puissante, certaines situations requièrent des approches complémentaires ou alternatives :
— **Régression avec régularisation :** Les techniques telles que la régression Ridge ou Lasso intègrent une pénalisation dans la fonction objectif pour limiter l'impact de la multicolinéarité et prévenir le surajustement.
— **Régression non linéaire :** Lorsque la relation entre les variables n'est pas linéaire, des transformations (logarithmiques, polynomiales, etc.) ou des modèles non linéaires peuvent être envisagés.

— **Méthodes d'apprentissage automatique :** Pour des jeux de données complexes, des algorithmes tels que les forêts aléatoires ou les réseaux de neurones permettent de capturer des interactions et des non-linéarités difficiles à modéliser avec la régression linéaire classique.

2.7 Importance du Cadre Théorique

L'élaboration d'un cadre théorique solide constitue la pierre angulaire de toute étude quantitative. Il permet non seulement de justifier le choix des méthodes utilisées mais aussi d'interpréter les résultats obtenus dans une perspective scientifique rigoureuse. En maîtrisant les principes et les hypothèses sous-jacents au modèle de régression linéaire multiple, il devient possible d'identifier les limites de l'analyse et d'envisager des améliorations ou des extensions adaptées aux particularités des données étudiées.

Ainsi, ce cadre théorique offre une base conceptuelle robuste pour l'analyse des frais médicaux et constitue le socle sur lequel repose l'interprétation des résultats présentés dans cette étude.

3 Problématique

Dans le domaine des assurances médicales, le calcul des frais d'assurance dépend de plusieurs facteurs tels que l'âge, le sexe, le statut de fumeur, l'indice de masse corporelle (IMC) et le nombre d'enfants à charge. Ces variables influencent directement le coût des primes d'assurance.

Ainsi, la problématique de cette étude est la suivante : *Quels sont les principaux facteurs influençant le coût des assurances médicales, et comment peut-on les modéliser à l'aide d'une régression linéaire multiple afin de prédire avec précision les frais médicaux d'un individu ?*

4 Solution Proposée :

Pour répondre à la problématique identifiée, nous proposons l'utilisation de la régression linéaire multiple comme méthode principale d'analyse et de prédiction des frais médicaux. Cette approche permet de modéliser la relation entre plusieurs facteurs explicatifs (âge, sexe, IMC, statut de fumeur, etc.) et la variable cible (frais médicaux).

4.1 Prétraitement des Données

— **Nettoyage des données :** Identification et suppression des valeurs aberrantes susceptibles d'influencer négativement le modèle.
— **Gestion des valeurs manquantes :** Imputation par la moyenne/médiane pour les variables continues ou par le mode pour les variables catégoriques.
— **Transformation des variables :** Normalisation ou standardisation des variables continues pour éviter l'influence disproportionnée des grandes valeurs.

4.2 Encodage des Variables Catégoriques

— Conversion des variables qualitatives (sexe, fumeur, région) en variables numériques via *one-hot encoding* ou *label encoding*.

4.3 Sélection des Variables Pertinentes

— Analyse de la corrélation entre les variables explicatives et la variable cible (charges).
— Utilisation du *Facteur d'Inflation de la Variance* (VIF) pour détecter et éliminer la multicolinéarité.
— Sélection des variables significatives via des tests statistiques (p-value, tests de régression).

4.4 Construction et Optimisation du Modèle

— Mise en place de la régression linéaire multiple en utilisant R.
— Division des données en jeu d'entraînement (80%) et jeu de test (20%).
— Validation croisée et ajustement des hyperparamètres pour améliorer la robustesse du modèle.

4.5 Évaluation des Performances du Modèle

— **Coefficient de détermination (R^2 et R^2 ajusté)** pour mesurer la qualité d'ajustement du modèle.
— **Erreur Quadratique Moyenne (MSE)** et **Erreur Quadratique Moyenne Racine (RMSE)** pour évaluer la précision des prédictions.
— Analyse des résidus pour vérifier les hypothèses du modèle (homoscédasticité, normalité des erreurs).

Cette approche structurée permet d'assurer la fiabilité du modèle et d'améliorer la qualité des prédictions en fonction des facteurs déterminants influençant les frais médicaux.

5 . Méthodologie

5.1 Collecte et Description des Données

5.1.1 1.1. Source et Période de Collecte

Les données proviennent d'une base spécifique liée à l'assurance santé (ou d'une source publique, si tel est le cas). Elles couvrent une période déterminée, par exemple une année complète ou plusieurs exercices, afin d'inclure une variété de situations médicales et socio-démographiques.

5.1.2 Taille et Structure de l'Échantillon

L'échantillon se compose de N observations, chacune décrivant un individu assuré. Les variables disponibles incluent notamment :
— Des variables continues telles que l'âge, l'indice de masse corporelle (IMC) et les frais médicaux (variable cible).
— Des variables catégoriques comme le sexe, le statut de fumeur ou la région de résidence.

5.1.3 Aperçu Statistique

Afin de détecter d'éventuelles anomalies, des statistiques descriptives (moyenne, médiane, écart-type) et des visualisations exploratoires (histogrammes, boxplots, graphiques de dispersion) ont été réalisées pour chaque variable.

5.2 Prétraitement et Nettoyage des Données

5.2.1 Gestion des Valeurs Manquantes

- **Quantification :** Le pourcentage de valeurs manquantes est estimé pour chaque variable.
- **Stratégie d'Imputation :**
  - Pour les variables continues (ex. âge, IMC), une imputation par la moyenne ou la médiane est privilégiée afin de limiter l'influence des valeurs extrêmes.
  - Pour les variables catégoriques (ex. sexe), l'imputation par le mode (catégorie la plus fréquente) est recommandée.

5.2.2 Identification et Traitement des Valeurs Aberrantes

- **Détection :**
  - Inspection de boxplots pour repérer visuellement d'éventuelles valeurs extrêmes.
  - Calcul de z-scores ou de la distance de Cook afin d'identifier des points particulièrement influents pour la régression.
- **Traitement :**
  - Suppression de certaines observations si les anomalies sont avérées (erreur de saisie ou données clairement hors sujet).
  - Winsorisation ou transformation logarithmique si l'objectif est de conserver la plus grande partie des données tout en atténuant l'impact des extrêmes.

5.2.3 Transformation des Variables

- **Standardisation / Normalisation :** Les variables continues peuvent être standardisées pour aligner leurs échelles (utile pour comparer l'importance relative des coefficients).
- **Transformations Spécifiques :** L'application d'une fonction logarithmique ou racine carrée peut s'avérer pertinente lorsque la distribution d'une variable présente une forte asymétrie.

5.3 Encodage des Variables Catégoriques

5.3.1 Catégorisation et Méthodes d'Encodage

- **One-Hot Encoding :** Génération de variables indicatrices (*dummies*) pour représenter les catégories nominales (par exemple, les régions de résidence).
- **Label Encoding :** Attribution d'un entier à chaque catégorie, mais potentiellement inadapté pour des variables sans ordre logique (risque de créer une fausse hiérarchie).

5.3.2 Choix Justifié de l'Encodage

- Les variables à deux catégories (statut de fumeur : oui/non) sont généralement binaires (0/1).
- Les variables avec plus de deux modalités (région) sont traitées via One-Hot Encoding pour préserver la nature strictement nominale des catégories.

5.4 Sélection des Variables Pertinentes

5.4.1 Analyse de Corrélation

- **Matrice de Corrélation :** Permet de relever rapidement les corrélations élevées entre les variables explicatives (signe de multicolinéarité) et la force de leur lien avec la variable cible (frais médicaux).
- **Tests de Corrélation Appropriés :**
  - Coefficient de Pearson pour les variables continues.
  - Test du Chi-deux pour les variables catégoriques.

5.4.2 Détection de la Multicolinéarité

- **Facteur d'Inflation de la Variance (VIF) :** Le calcul du VIF met en évidence les variables redondantes. Si le VIF excède un seuil de 5 ou 10 (selon les pratiques), il est recommandé de retirer ou combiner certaines variables.

5.4.3 Tests Statistiques et Pistes de Réduction

- **p-values :** Les variables dont la p-value est trop élevée (généralement $> 0{,}05$) peuvent être considérées comme non significatives dans le modèle.
- **AIC / BIC :** Les critères d'information d'Akaike ou bayésien permettent de comparer plusieurs configurations de modèles et de choisir celle qui réalise le meilleur compromis entre justesse d'ajustement et parcimonie.

5.5 Construction du Modèle de Régression Linéaire Multiple

5.5.1 Formulation Générale

Le modèle RLM s'écrit de la façon suivante :

$$\text{Charges} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

où β_0 est l'ordonnée à l'origine, β_i les coefficients de régression et ε le terme d'erreur.

5.5.2 Séparation des Données

- **Jeu d'Entraînement (Training set) :** 80% des données.
- **Jeu de Test (Test set) :** 20% des données.
- **Validation Croisée (Cross-validation) :** Généralement effectuée en k-plis (*k-fold*) afin de mesurer la robustesse du modèle et de réduire le risque de surapprentissage.

5.5.3 Implémentation Logicielle

- Les bibliothèques Python (`scikit-learn`, `statsmodels`) ou R (`car`, `lmtest`) sont couramment utilisées pour effectuer la régression linéaire.
- Les sorties statistiques (*p-values*, coefficients, intervalles de confiance) sont analysées pour valider la pertinence de chaque variable.

5.6 Optimisation et Validation

5.6.1 Réglage d'Éventuels Hyperparamètres

- Dans le cadre d'une régression simple, il n'y a pas toujours d'hyperparamètre à ajuster.

— En revanche, si une régularisation (*Lasso*, *Ridge*, *ElasticNet*) est envisagée afin de contrôler le surajustement ou la multicolinéarité, le paramètre de régularisation (α) doit être déterminé via une grille de recherche (*Grid Search*) ou une méthode aléatoire (*Random Search*).

5.6.2 Vérification des Hypothèses de la Régression

— **Linéarité :** Représentation graphique de la relation entre variables explicatives et variable cible.
— **Homoscedasticité :** Diagramme des résidus vs. valeurs prédites pour détecter toute structure inappropriée (les résidus doivent être répartis aléatoirement).
— **Normalité des Résidus :** Histogramme, Q-Q plot ou test de Shapiro-Wilk pour confirmer que les erreurs suivent une distribution normale.
— **Multicolinéarité :** Confirmation via la matrice de corrélation et les VIF.

5.7 Évaluation des Performances du Modèle

5.7.1 Mesures d'Ajustement

— **Coefficient de Détermination (R^2) :** Reflète la proportion de variance de la variable cible expliquée par l'ensemble des variables explicatives.
— **R^2 Ajusté :** Corrige l'effet inflationniste pouvant résulter de l'ajout d'un grand nombre de variables.

5.7.2 Mesures d'Erreur

— **MSE (Mean Squared Error) :** Moyenne des carrés des écarts entre valeurs observées et prédites.
— **RMSE (Root Mean Squared Error) :** Racine carrée du MSE, mesurée dans la même unité que la variable cible, ce qui facilite l'interprétation.
— **MAE (Mean Absolute Error) :** Moyenne des valeurs absolues des écarts, moins sensible aux valeurs extrêmes.

5.7.3 Analyse des Résidus

— **Distribution et Graphique :** Un nuage de points résidus vs. valeurs prédites sans structure apparente suggère un modèle adéquat.
— **Points Influentiels :** La distance de Cook et les *Leverage plots* permettent d'identifier des observations susceptibles de biaiser le modèle.

5.8 Interprétation et Perspectives

5.8.1 Interprétation des Coefficients

Les coefficients associés à chaque variable indiquent dans quelle mesure cette variable contribue à la variation des frais médicaux. Ils peuvent être standardisés afin de comparer plus facilement l'impact relatif de chaque facteur (âge, IMC, statut de fumeur, etc.).

5.8.2 Limites et Améliorations Possibles

— Les variables non prises en compte (état de santé global, habitudes alimentaires, etc.) peuvent influencer la précision du modèle.

— Des approches non linéaires ou des modèles d'apprentissage automatique plus complexes (*Random Forest*, *Gradient Boosting*) peuvent être évalués pour comparer leurs performances.

5.8.3 Application Pratique

Les résultats obtenus sont particulièrement utiles pour :
— Optimiser la tarification dans les compagnies d'assurance en tenant compte des variables les plus sensibles (statut de fumeur, tranche d'âge, etc.).
— Mettre en place des actions de prévention ciblées pour les assurés présentant un risque de frais médicaux élevé.

6 Résultats

6.1 Exploration initiale des données

Variable	Moyenne	Écart-type	Min - Max
Âge	39.207	14.050	18 - 64
IMC	30.663	6.098	15.96 - 53.13
Charges	13270.42	12110.01	1121.87 - 63770.43

Table 1 – Statistiques Descriptives

Variable	VIF
Age	1.0138
BMI	1.0122
Children	1.0019

Table 2 – Performance du modèle sur le jeu de test

Metric	Value
MSE	4.027878e+07
RMSE	6.346556e+03
MAE	4.234608e+03
R^2	7.078166e-01

Table 3 – Coefficients estimés du modèle final

Variable	$\hat{\beta}$	Erreur std	p-value	IC 95% Min	IC 95% Max
(Intercept)	-11676.8304	937.56870	0	-13516.1001	-9837.5608
age	259.5475	11.93418	0	236.1357	282.9593
bmi	322.6151	27.48741	0	268.6919	376.5384
smokeryes	23823.6845	412.86668	0	23013.7458	24633.6232

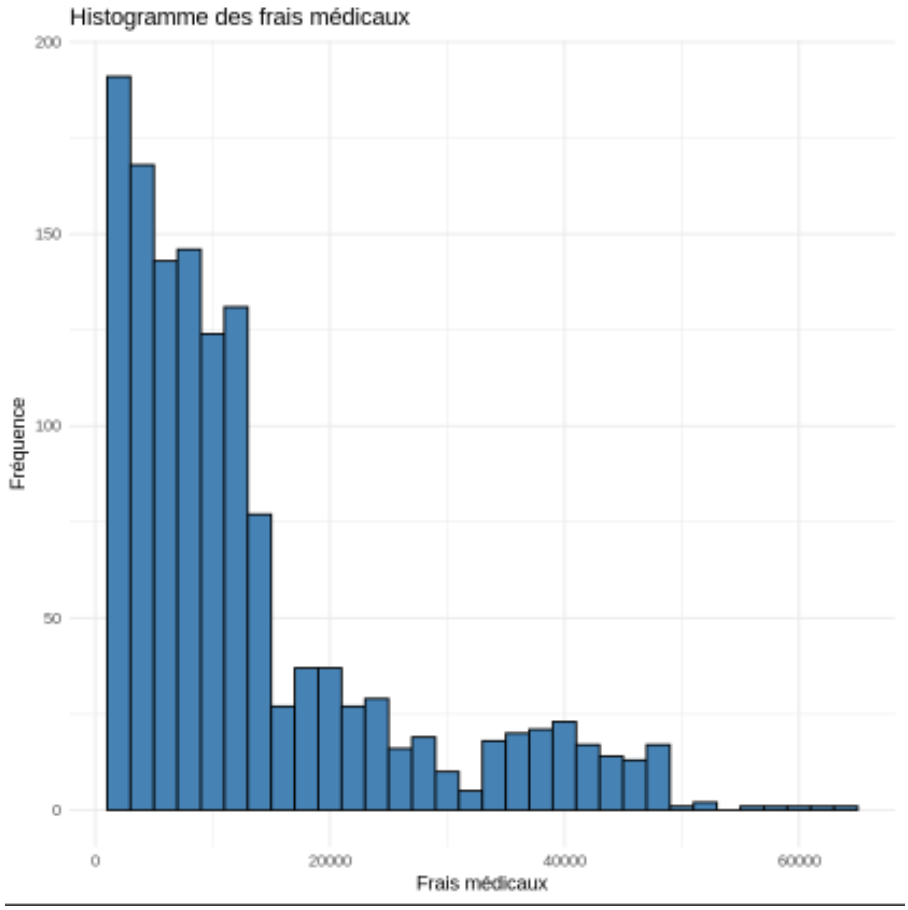

FIGURE 2 – Distribution des frais médicaux

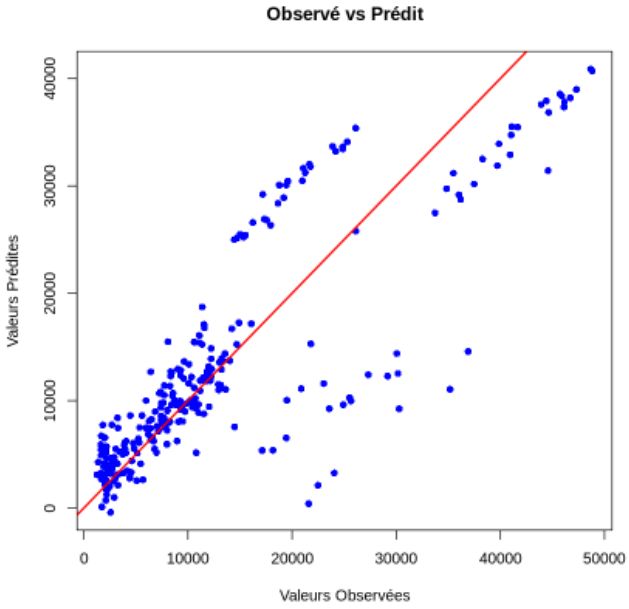

FIGURE 3 – observe vs predit

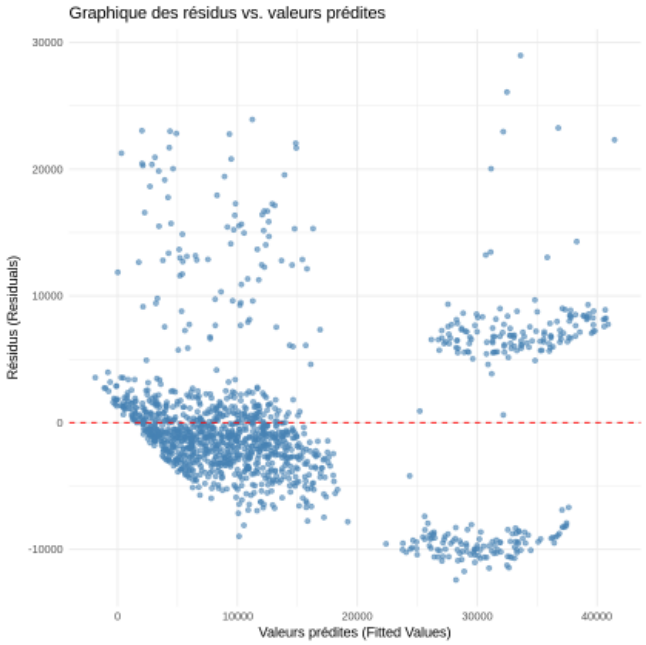

FIGURE 4 – Graphique des résidus vs. valeurs prédites

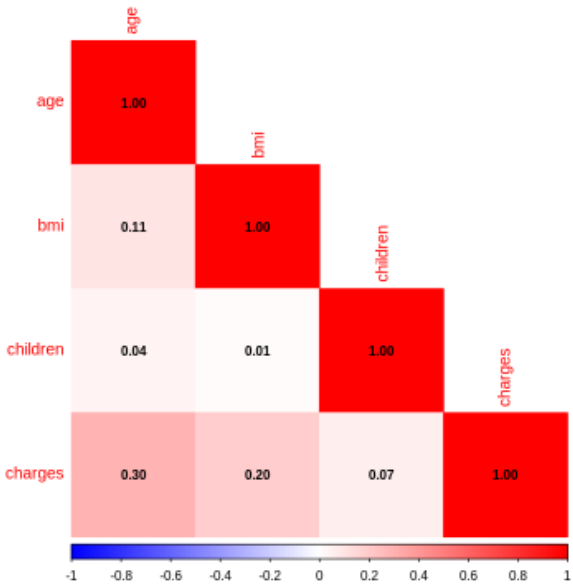

FIGURE 5 – Matrice de corrélation

7 Discussion

7.1 Interprétation des Résultats

Les résultats obtenus montrent que les variables telles que l'âge, l'indice de masse corporelle (IMC) et le statut de fumeur ont un impact significatif sur les frais médicaux. En particulier, le coefficient élevé associé à la variable `smokeryes` confirme que le fait d'être fumeur est un facteur majeur d'augmentation des charges. De plus, les coefficients positifs pour l'âge et l'IMC indiquent qu'une augmentation de ces paramètres est corrélée à une hausse des frais médicaux, ce qui est cohérent avec les observations dans la littérature médicale et assurantielle.

7.2 Performance du Modèle et Limitations

Le modèle de régression linéaire multiple explique environ 71% de la variance des frais médicaux ($R^2 = 0.71$), ce qui est satisfaisant pour une première approche. Cependant, près de 29% de la variance reste non expliquée, suggérant que d'autres facteurs (par exemple, des variables cliniques ou socio-économiques non incluses dans l'analyse) pourraient également jouer un rôle significatif.

Il convient également de noter que, malgré des diagnostics des résidus relativement satisfaisants, la présence potentielle de points aberrants et d'influences fortes (identifiables via la distance de Cook ou les graphiques de levier) pourrait affecter la robustesse du modèle.

7.3 Implications Pratiques et Recommandations

Ces résultats ont des implications importantes pour le secteur de l'assurance santé :

- **Tarification des Polices :** L'intégration de facteurs comme le statut de fumeur et l'IMC peut aider les compagnies d'assurance à mieux calibrer leurs primes.
- **Actions Préventives :** La forte influence du statut de fumeur sur les charges suggère qu'une politique de prévention et d'incitation à l'arrêt du tabac pourrait réduire les coûts de santé.

Il serait pertinent d'envisager l'intégration de variables supplémentaires (telles que l'activité physique, les antécédents médicaux ou le niveau socio-économique) afin d'améliorer la précision du modèle.

7.4 Perspectives d'Amélioration et Recherches Futures

Pour renforcer et étendre cette étude, plusieurs axes d'amélioration sont envisageables :

- **Modèles Alternatifs :** Tester des modèles non linéaires ou des méthodes d'apprentissage automatique (comme les forêts aléatoires ou les réseaux de neurones) pour comparer leurs performances par rapport à la régression linéaire.
- **Régularisation :** Intégrer des techniques de régularisation (Lasso, Ridge ou ElasticNet) afin de gérer plus efficacement la multicolinéarité et éviter le surajustement.
- **Collecte de Données :** Améliorer la qualité et la quantité des données en incluant de nouvelles variables explicatives susceptibles d'influencer les frais médicaux.
- **Analyse des Résidus :** Mener une analyse plus fine des résidus pour détecter d'éventuels problèmes d'hétéroscédasticité ou de non-linéarité.

En conclusion, bien que le modèle actuel fournisse des indications claires sur les facteurs influençant les frais médicaux, une amélioration continue du modèle et l'exploration de nouvelles approches pourraient permettre d'obtenir des prédictions encore plus précises et d'optimiser la gestion des risques en assurance santé.

8 Conclusion

En conclusion, cette étude a permis de mettre en lumière les facteurs clés influençant les frais médicaux, notamment l'âge, l'indice de masse corporelle (IMC) et le statut de fumeur. Le modèle de régression linéaire multiple développé a montré une capacité satisfaisante à prédire les charges médicales, avec un coefficient de détermination (R^2) d'environ 71%. Ce résultat suggère que, malgré une part de variance encore non expliquée, le modèle constitue une base solide pour comprendre et anticiper les coûts en assurance santé.

Cependant, plusieurs améliorations pourraient être envisagées. L'intégration de variables supplémentaires (comme des indicateurs socio-économiques ou des antécédents médicaux) ainsi que l'exploration de modèles non linéaires ou d'algorithmes d'apprentissage automatique pourraient permettre d'accroître la précision des prédictions. De même, l'adoption de techniques de régularisation (telles que Lasso ou Ridge) aiderait à mieux gérer la multicolinéarité et à affiner la sélection des variables.

Enfin, ces résultats ouvrent des perspectives intéressantes pour la tarification des polices d'assurance et la mise en œuvre de stratégies de prévention ciblées, visant à réduire les coûts et à améliorer la gestion des risques en santé. Une approche pluridisciplinaire, combinant des analyses statistiques robustes et des données cliniques complètes, serait ainsi bénéfique pour optimiser les politiques de prise en charge dans le domaine de l'assurance santé.

Mots-clés : régression linéaire multiple, frais médicaux, assurance santé, modélisation prédictive, analyse des données.

Références Bibliographiques

Références

[1] Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*. Wiley.

[2] Kutner, M.H., Nachtsheim, C.J., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill.

[3] Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. SAGE.

[4] Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012). *Introduction to Linear Regression Analysis*. Wiley.

[5] Biau, D.J., Kernéis, S., & Porcher, R. (2010). *Statistics in Brief : The Importance of Sample Size in the Planning and Interpretation of Medical Research. Clinical Orthopaedics and Related Research*, 468(9), 2282-2288.

[6] Chernew, M. E., Newhouse, J. P., & Fendrick, A. M. (2003). *Paying for Prescription Drugs : What Works ? Health Affairs*, 22(2), 34-48.