Rapport de Statistiques Bayésiennes
Influence de variables et prédiction de coûts médicaux

Noémie Turmel, Assmaa Alsamadi, Aymen Lakhyar,
Achraf Azalmad, Aurélien Henriques

M2DS
2023-2024

Table des matières

1 Introduction

Nous analysons le jeu de données "insurance" qui rend compte du coût de frais médicaux de patients en fonction de variables telles que le nombre d'enfant, l'Indice de masse corporelle ou le fait de fumer en utilisant la méthode bayésienne. Nous commençons par une analyse exploratoire des variables afin d'évaluer la possibilité de faire une régression du coût. Nous créeons ensuite cette régression bayésienne que nous comparons à une régression fréquentiste. Nous utilisons alors des critères d'évaluation afin d'apprécier la performance de la régression. Enfin, nous concluons sur les questions tendancielles entre grandeurs qui émanent de ce jeu de données.

Le jeu de données que nous utilisons correspond à 1338 données générées à partir de données démographiques du US Census Bureau. On note plusieurs choses à partir de ce-ci : premièrement, un millier de données permet d'entrevoir que nous pourrons faire une régression robuste. En effet, la division en sous-groupe de données en fonction des caractéristiques ne devrait pas tomber sous un effectif de l'ordre de la centaine, ce qui reste en pratique cohérent avec la possibilité de définir une distribution postérieure. Nous analyserons par la suite la distribution des classes et des variables quantitatives afin d'avoir un aperçu complet de cette répartition. De plus, les données présentées sont synthétiques, ce qui indique que la nature réelle des tendances devra être mis en perspective. En particulier, nous remarquerons des résultats contre-intuitifs qui ne permettent pas a priori de statuer sur une hypothèse physique sous-jacente.

Les variables, au nombre de sept, sont quantitatives (coût médical, âge, IMC et nombre d'enfants) ou qualitatives (région, sexe et fumeur/non-fumeur). Pour ces dernières, nous utiliserons une conversion de type dummy-coding / one-hot encoding pour les considérer dans la régression. Le but est alors de pouvoir prédire le coût médical d'un client avec une certaine confiance en fonction de ces variables pour une entreprise d'assurance. Intuitivement, ces variables semblent être corrélées aux frais médicaux, bien qu'on imagine que la région n'a pas une influence aussi importante que les autres variables. En général, d'autres variables peuvent être considérées dans ce type de régression comme les antécédents médicaux et l'historique famillial.

Ici, on se bornera à évaluer l'influence des variables en se posant 3 questions :

-Les variables médicales comme l'âge, l'imc et le fait de fumer sont-elles influentes dans l'augmentation du coût des frais médicaux ?

-Y'a-t-il une différence régionale dans la facturation des frais médicaux aux clients ?

-Une assurance peut-elle prédire les frais médicaux facturés en fonction de l'ensemble ou d'un sous-ensemble des variables proposées ?

2 Analyse Exploratoire Des Variables

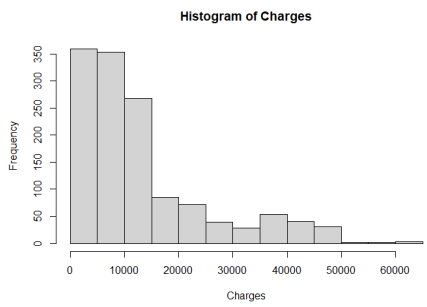
2.1 Description Des Données

Les 1338 patients sont répartis d'une façon quasi-équitable entre femmme et homme. De plus les fumeurs représentent 79% des patients. Le tableau 1 ci-dessous montre la répartion de données selon le nombre d'enfant. Les patients qui ont un nombre d'enfant supérieur ou égal à 3 sont regroupés ensemble dans la suite de l'étude.

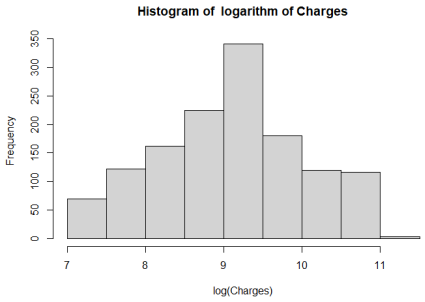
Nombre d'enfant	**0**	**1**	**2**	**3**	4	5
Pourcentage	43%	24%	18%	12%	2%	1%

Table 1 – Répartition du nombre d'enfants

Concernant la variable cible, les frais médicaux, on obtient une moyenne de 13270 et un écart-type de 12110. Ceci montre que les données sont extremement dispersées. De plus, la figure 1a montre que la densité marginale des charge ressemble à la densité Gamma ou une fonction (k-exp(x)). On peut tracer la distribution logarithmique des charge (figure 1b) pour le vérifier, ce qui nous permet d'obtenir une distribution proche d'une gaussienne. De ce fait, on utilisera cette transformation logarithmique par la suite dans la régression.



(a) histogramme des coûts



(b) Histogramme des log-coûts médicaux

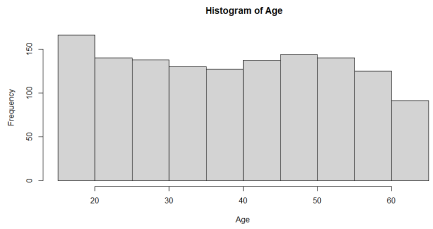
FIGURE 1 – Répartition des charges par effectif

Le tableau 2 montre que l'âge moyen des patients est 39.21 ans, pour un écart-type de 14.05 ans, montrant une forte dispersion. De plus, la figure 2a montre une quasi-répartition de la proportion pour tous les age.

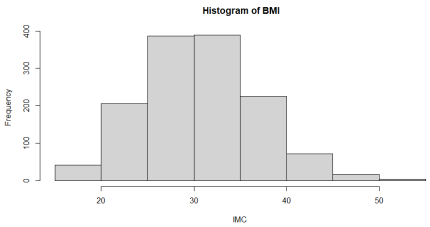
La moyenne d'IMC est 30.66 avec un max de 54.13 et la figure 2b montre que la majorité ont un score IMC entre 20 et 40, reparti selon une distribution qu'on considère gaussienne.

Variable	**Moyenne**	**Ecart-type**	**Médiane**	**Min**	**Max**
Age	39.21	14.05	39	18	64
BMI	30.66	6.1	30.4	15.96	53.13
Charges	13270	12110	9382	1122	63770

TABLE 2 – Statistiques des variables quantitatives



(a) Histogramme de l'âge



(b) Histogramme de l'IMC

FIGURE 2 – Répartitions des données pour les variables quantitatives

2.2 Critique de la qualité de génération des données

On peut déjà se permettre quelques commentaires sur la qualité de la génération de nos données vis-à-vis des statistiques et distributions précédemment descrites. On constate sur la Figure 2a que la répartition de l'âge est plutôt uniforme et en cohérence avec la pyramide des âges américaine [1]. En revanche, la partinence de la génération des données se gâte lorsqu'il s'agit de l'IMC, en effet d'après la Figure 2b on aurait une prévalence de l'obésité, c'est à dire un IMC supérieur à 30, environ à 50% pour une médiane à 30.4%. Bien que l'obésité et le surpoids soit un problème majeur aux Etats-Unis, la prévalence est au maximum estimée à 41.9% [2], maximum considérant toute la population américaine et toutes les catégories d'âge. Ainsi rien qu'avec ces statistiques basiques non approfondies on conclue déjà que les individus de notre dataset ont une tendance à l'obésité, plus que dans la population générale.

2.3 Influence des variables sur la variable cible

Dans la suite on va étudier l'effet des différentes variables sur la charge médical.

Pour le genre du patient, les box plots ci-dessous (Figure 3) montrent que les charges médianes de deux groupes (hommes et femmes) sont presque égales mais le troisième quartiles des hommmes (q=2000) est plus grand de celui des femmmes (q=1700). De plus, le upper whisker du groupe des hommes est plus élevé que chez les femmes. Pour étudier l'hypothèse que les hommes ont des charges médicales plus élevées que les femmes, on doit estimer la probabilité que les hommes aient une charge médicale plus élevée que les femmes. Pour cela une approximation de type MonteCarlo est réalisée.

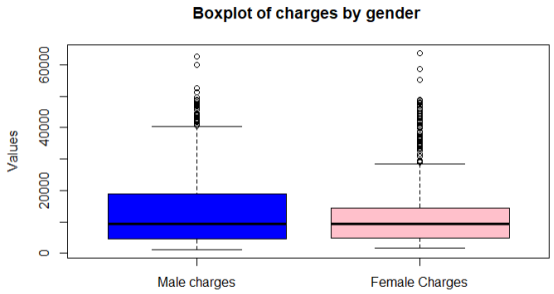


FIGURE 3 – Box Plots des coûts en fonction du sexe

La figure 4 montre que la densité des charges pour les femmes et les hommes resemble a un loi $\Gamma(\alpha, \beta)$. Les parametres de loi $\Gamma_m(\alpha_m, \beta_m)$ des hommes et $\Gamma_f(\alpha_f, \beta_f)$

des femmes sont estimés avec $\alpha_i = (\frac{\text{moyenne des charges de groupe i}}{\text{ecart type des charges de groupe i}})^2$ et $\beta_i = \frac{\text{moyenne des charges de groupe i}}{\text{ecart type des charges de groupe i}^2}$ où $i \in \{$ hommes, femmes$\}$. Le prior des charges médicales est supposé de densité $\Gamma(1, 0.5)$. Une approximation par Montecarlo de taille 1000 nous donne une probabilité de 0.579 que le coût des hommes soit plus élevé. Ainsi, on conclut par cette analyse unimodale que les hommes ont tendance à avoir un coût médical plus élevé.

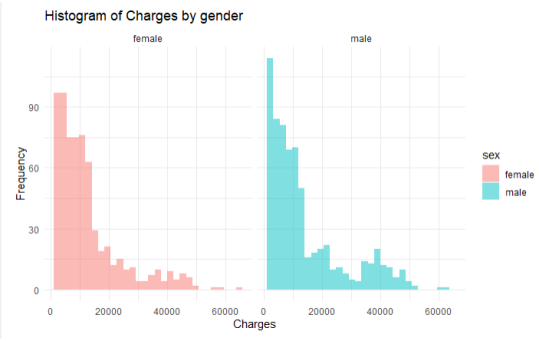
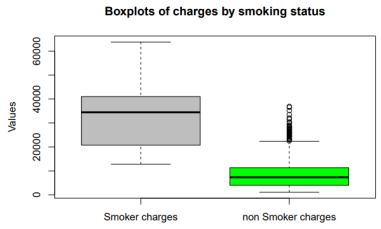
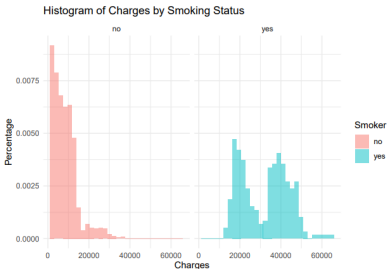


FIGURE 4 – Histogramme des coûts par sexe

Le groupe des fumeurs qui représentent 80% des patients, a un coût médian de 34456 $. Cette valeur est quatre fois plus plus élevée que celle des non-fumeurs de 7345$. (figure 5a). De plus la distribution des coûts des fumeurs ont une densité différente de celle des non-fumeur (figure 5b). Dans ce cas une approximation par méthode Monte-Carlo n'est pas valable, car les non-fumeur ont une densité bimodal. Ceci montre que d'autres covariables sont nécessaires pour expliquer les charges dans le groupe des fumeurs.



(a) Box-Plot des charges en fonction du statut fumeur/non



(b) Histogramme des charges

FIGURE 5 – Statistiques des charges en fonction du fait de fumer

Les nuages de points dans la figure 6 représentent les frais médicaux par rapport à l'IMC, sur lesquels on représente en couleur le statut de fumeur ou non-fumeur. Ces nuages de points montrent que les fumeur avec un IMC plus grand que 30 ont un coût médical élevé et montre une cission, ce qui peut expliquer les charges élevées chez les fumeur. De plus, la plupart de non fumeur ont une charge entre 0 et 1400 dans le cas non-fumeur, ce qui montre l'importance de considérer la corrélation entre les deux variables Dans ce but, Une variable binaire est crée, pour regrouper les données selon l'IMC, ceux avec un IMC plus grand que 30 ou plus petit que 30. Cette cission en groupes d'IMC inférieur et supérieur à 30 sera conservée dans notre analyse futur, en correspondance avce l'utilisation faite dans la littérature et la proposition de l'auteur du jeu de données. Une autre variable catégorielle de 4 classes a aussi éte créée de la façon suivante : les gens qui ont un IMC plus petit que 18.6 (Anoréxie), les gens qui ont un IMC entre 18.6 et 25 (Bonne santé), ceux entre 25 et 30 sont (surpoids). Et ceux qui restent (obesité). Nous effectuerons aussi la régression en utilisant cette variable pour nous assurer que la première cission est la meilleure.

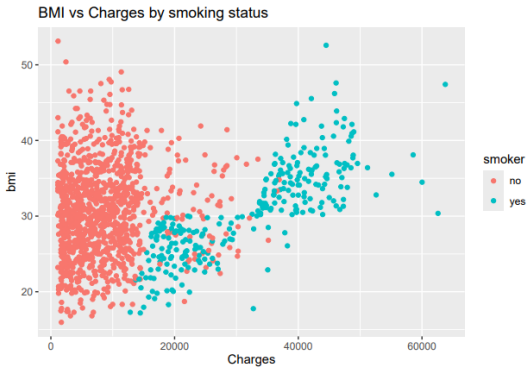
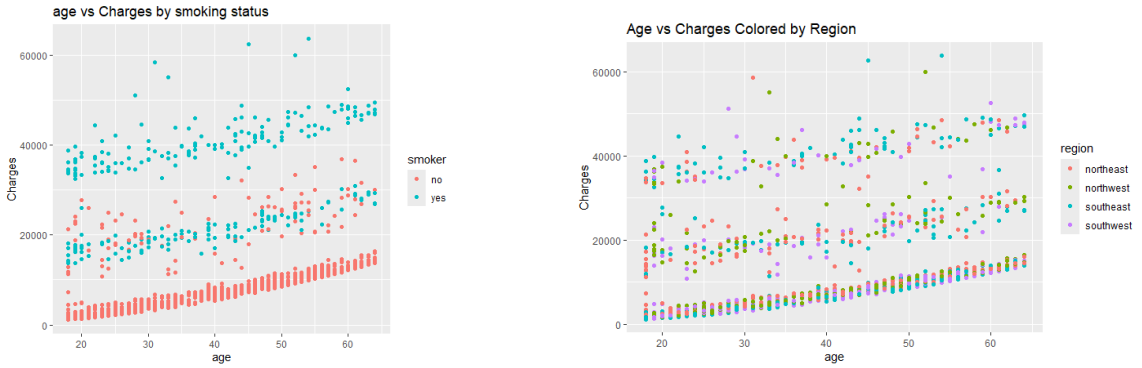


FIGURE 6 – IMC en foncion des charges

Pour voir si l'âge a une influence sur les charges médicales, on trace le graphe de l'un en fonction de l'autre. On trouve des courbes de "niveaux" distincts (courbe d'ordonnée à l'origine commençant à environ 0, 16000 et 38000) qu'on tente de découvrir en représentant différentes variables catégorielles en couleur (Figure 7a). Nous constatons que le fait de fumer peut influencer les charges en fonction de l'âge, notamment que les niveaux supérieur et inférieur de ces courbes correspondent clairement à fumeur et non-fumeur, mais nous n'avons pas trouver une manière significative d'attribuer la courbe médiane à une catégorie particulière de patients. Nous tentons

d'examiner l'influence de la région dans les courbes des frais médicaux en fonction de l'âge (Figure 7b) mais aucune tendance notable ne se dégage de ce-ci. En vue de son caractère particulier, nous discutons de la variable région individuellement.



(a) (b)

FIGURE 7 – Coûts médicaux en fonction de l'âge, en tentant d'apprécier l'influence de la région et du fait de fumer sur les courbes. On constate une dépendance de ces courbes au fait de fumer mais il est beaucoup plus complexe de déterminer une tendance régionale dans celles-ci.

2.4 Cas de l'influence régionale dans le coût

Nous souhaitons comprendre l'importance de la variable de localité. Cette variable a en effet une place particulière puisque l'effet sur les charges médicales est intuitivement moins directe que les autres variables. Il est donc fondamental de bien comprendre sa structure afin de savoir si l'on peut la considérer dans les régressions.

Comme nous l'avons vu précédemment, les distributions des frais médicaux en fonction de l'âge différencié par région ne permettent pas de conclure à une influence explicit. On peut cependant développer une intuition de celle-ci en regardant une carte du pays :

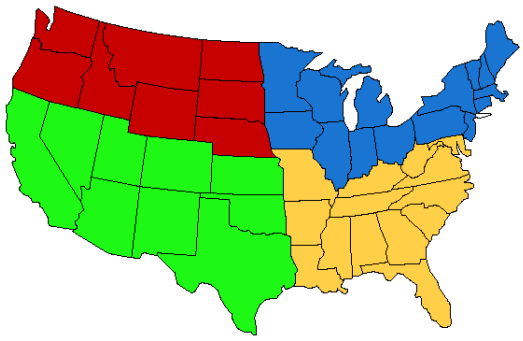


FIGURE 8 – Séparation des états-unis en régions. Notons que la répartition ici ne correspond pas exactement à celle utilisée dans la classification du dataset en l'absence de précision de la méthodologie des données.

Avec prudence, on retient seulement que la division en région reprend une tendance en terme de niveaux de services et économiques, notamment que la region southeast coincide avec le deep south, une région des états-unis marquée par un taux d'obésité important et un accès aux soins réduit, qui peut se traduire par des opérations plus lourdes et plus onéreuses [3].

La distribution de coût (figure 9) permet de s'en apercevoir au travers d'un second pic marqué pour cette région, dans lequel l'effectif entre 38000 et 50000 est aussi important que pour les trois autres régions réunies. Notre régression peut donc prendre en compte cette particularité pour information.

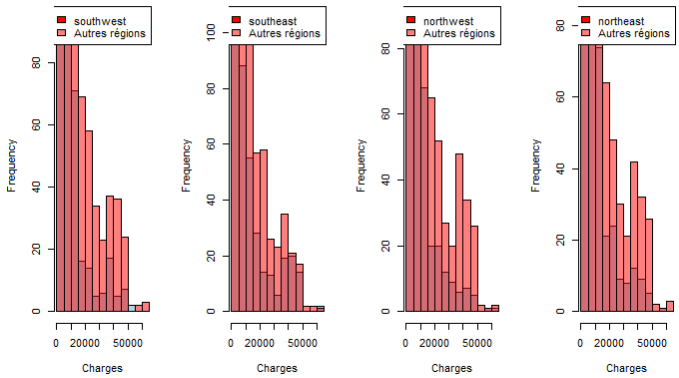


FIGURE 9 – Comparaison des distributions régionales vs autres régions

Les autres régions sont a priori homogènes au niveau sociétal donc nous ne disposons pas de plus d'informations que ce qui est disponbile dans la distribution des effectifs, le box plot suivant permet d'en rendre compte.

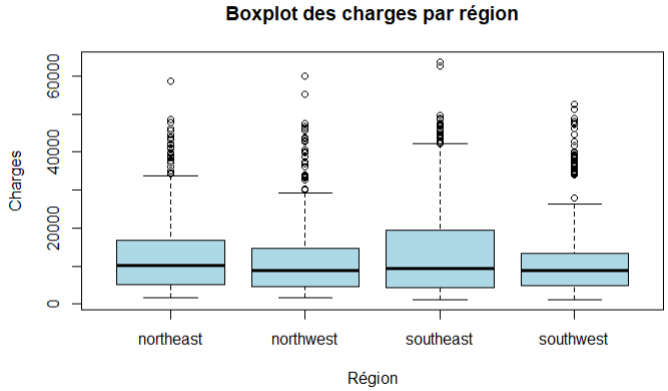


FIGURE 10

3 Régression Bayésienne

3.1 Analyse Préliminaire

Pour commencer, on utilise un modèle simple géneralisé de régression bayésienne. On a utilisé stan glm du package rstanarm avec les priors déja implementés dans le modèle. Stanglm utilise une famille peu informative de distribution pour les priors. Pour un model $y = a + b_1x_1 + \ddots + b_px_p$. Chaque $b_k \sim \mathcal{N}(0, 2.5\frac{sd(y)}{sd(x_k)})$ et $a \sim \mathcal{N}(mean(y), 2.5sd(y))$.

Ce modèle est utilisé pour estimer les effets de fumer sur les frais médicaux, pour lesquels on comparera les coefficient médians des groupes. Ainsi, différentes variables ont été ajoutées au modèle.

Dans un premier temps, on a comparé le coefficient de régression des fumeurs et non-fumeurs en utilisant la regression bayésienne avec le logarithme des charges comme variable cible. Ce choix est motivé par les explications précédentes sur la distribution des charges. Le model stan glm(log(charges) $\sim$ fumeur) montre que l'intercept est 8.79 avec une incertitude de 0.02, et le fait d'être fumeur augmente le

logarithme de charge de 1.52 avec 0.05 d'incertitude avec une erreur résiduel de 0.69. Ceci montre l'influence du fait de fumer sur la charge.

```
            Median MAD_SD
(Intercept) 8.79   0.02
smokeryes   1.52   0.05

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.69   0.01
```

FIGURE 11 – Regression bayésienne des log-charges avec le fait de fumer

Ensuite, nous ajoutons au modèle la variable continue de l'IMC et l'on compare le modèle obtenu avec celui de l'IMC en tant que variable catégorielle, par le dummy-coding cité plus haut. L'intercept et les coeficients des fumeurs dans les 3 modèles sont similaires. Mais on remarque que dans la figure 12a, l'IMC est presque sans influence sur le coût mais en classifiant les patients selon celui-ci, on montre qu'en réalité plus le taux d'IMC est élevé plus les charge augmentent (figure 12b). De plus un IMC de moins de 18 a un influence négative sur les charges avec une incertitude de 0.16 et une erreur résiduel est de 0.68 (Figure 12c). Moralement, une anoréxie concorderait donc avec une baisse des coûts médicaux

```
------
            Median MAD_SD
(Intercept) 8.19   0.10
smokeryes   1.51   0.05
bmi         0.02   0.00

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.68   0.01
```

(a) Regression avec IMC continue

```
------
                          Median MAD_SD
(Intercept)               8.67   0.03
smokeryes                 1.52   0.05
factor(bmi2)more then30 0.22   0.04

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.68   0.01
```

(b) Regression avec IMC en 2 classes

```
                        Median MAD_SD
(Intercept)              8.60   0.05
smokeryes                1.52   0.04
factor(bmi4)Obesity      0.29   0.05
factor(bmi4)Overweight   0.12   0.06
factor(bmi4)Underweight -0.25   0.16

Auxiliary parameter(s):
      Median MAD_SD
sigma 0.68   0.01
```

(c) Regression avec IMC en 4 classes

FIGURE 12 – Regression des log charges avec l'IMC et le statutvfumeur

Pour finir, on s'intéresse à l'effet du nombre d'enfant et du sexe des patients sur les log-charge (figure 13b). Le nombre d'enfant par patient a un effet similaire à l'imc où la log-charge augmente de 0.2 avec l'augmentation du nombre d'enfant à 3 ou plus pour une incertitude de 0.04 et une erreur résiduel presque égale à celle de tout autre modèle. Pour le sexe du patient, la log-charge diminue de 0.18 si c'est un homme avec une incertitude de 0.05. Notons donc qu'a priori, le sexe n'augmente pas drastiquement les coûts médicaux.

	Median	MAD_SD
(Intercept)	8.81	0.04
smokeryes	1.51	0.05
factor(grouped_children)Three or more	0.20	0.06
factor(grouped_children)Two	0.15	0.06
factor(grouped_children)Zero	-0.18	0.05

Auxiliary parameter(s):

	Median	MAD_SD
sigma	0.67	0.01

(a) Regression avec nombre d'enfant

	Median	MAD_SD
(Intercept)	8.83	0.03
smokeryes	1.52	0.05
sexmale	-0.08	0.04

Auxiliary parameter(s):

	Median	MAD_SD
sigma	0.69	0.01

(b) Regression avec variable genre

	Median	MAD_SD
(Intercept)	8.85	0.04
smokeryes	1.52	0.05
factor(grouped_children)Three or more	0.20	0.06
factor(grouped_children)Two	0.14	0.05
factor(grouped_children)Zero	-0.18	0.05
sexmale	-0.09	0.04

Auxiliary parameter(s):

	Median	MAD_SD
sigma	0.67	0.01

(c) Regression avec nombre d'enfant et genre

FIGURE 13 – Regression de log charges

En conclusion, ces différents modèles permettent de supposer qu'être fumeur et avoir un IMC et un nombre d'enfant élevé ont des effets plus important sur les charges en base logarithmique que les autres variables. On déduit ceci avec un modèle bayésien peu informative. Cependant, nous souhaitons creuser ces déductions avec un modèle plus informative sur les variables à influence importante et comparer la performance des modèles

3.2 Modèle complet

Le modèle final intègre une combinaison de variables sélectionnées sur la base des insights obtenus à partir des modèles précédents.

```
model_7<- stan_glm(log(charges) ~ age + `factor(bmi2)more then30` + children +
`factor(smoker)yes`, data = train_data, family = gaussian(), prior = normal(0, 2), chains = 4, seed =
12345)
```

FIGURE 14 – Modèle complet

3.2.1 Comparaison des modèles

Après l'implémentation de différents modèles de régression bayésienne visant à prédire les charges médicales basées sur divers facteurs, l'étape suivante consiste à évaluer et comparer la performance prédictive de ces modèles. Pour ce faire, nous utilisons la validation croisée leave-one-out (LOO-CV) comme méthode d'évaluation. Cette approche est essentielle pour identifier le modèle qui fournit les estimations les plus précises sur des données non observées.

Le LOO-CV est une technique de validation croisée où, pour un ensemble de données de taille n, chaque observation est utilisée une fois comme donnée de test tandis que les n moins 1 autres observations constituent l'ensemble d'apprentissage. Cette méthode est particulièrement utile pour les modèles bayésiens, car elle permet d'estimer la performance du modèle sur de nouvelles données sans nécessiter de grands ensembles de données distincts pour l'entraînement et le test.

Dans le cadre de la bibliothèque loo en R, trois paramètres clés issus du LOO-CV sont particulièrement pertinents pour l'évaluation des modèles :

— **elpdloo** : Indique la capacité prédictive d'un modèle sur de nouvelles données. Plus elle est élevée, mieux c'est.
— **ploo** : Reflète la complexité du modèle en termes de nombre de paramètres effectifs. Une valeur plus élevée peut suggérer un risque de surajustement.
— **looic** : Un critère d'information basé sur elpdloo, où des valeurs plus basses sont meilleures, similaire à l'AIC mais pour la validation croisée LOO.

Modèle	elpd_loo	p_loo	looic
Modèle 1	-1397.1	2.4	2794.3
Modèle 2	-1395.7	3.4	2791
Modèle 3	-1366.8	4.8	2733.7
Modèle 4	-1364.8	5.6	2729.7
Modèle 5	-888.6	9.1	1777.2
Modèle 6	-1380.3	3.3	2760.5
Modèle 7	-669.5	8.3	1339

FIGURE 15 – Comparaison des modèles

Les résultats du LOO-CV montrent des variations significatives dans la performance prédictive entre les différents modèles, comme indiqué par leurs scores LOOIC. Le Modèle 5 et le Modèle 7 se distinguent particulièrement avec des scores LOOIC nettement inférieurs à ceux des autres modèles, ce qui indique une meilleure performance prédictive. Cependant, le Modèle 7 avec un LOOIC de 1339 et un PLOO de 8.3, présente la meilleure performance prédictive parmi tous les modèles considérés. Ce résultat suggère que les variables et les interactions incluses dans le Modèle 7 fournissent une combinaison particulièrement informative pour estimer les charges médicales.

Il est important de noter que le PLOO représente le nombre effectif de paramètres ou la complexité du modèle dans le contexte de la validation croisée LOO. Un équilibre doit être trouvé entre la précision prédictive du modèle (comme indiqué par le LOOIC) et sa complexité (comme indiqué par le PLOO). Le Modèle 7, tout en étant relativement simple avec un PLOO de 8.3, offre une excellente performance prédictive, ce qui en fait un choix robuste pour la prédiction des charges médicales.

3.2.2 Analyse du modèle choisi

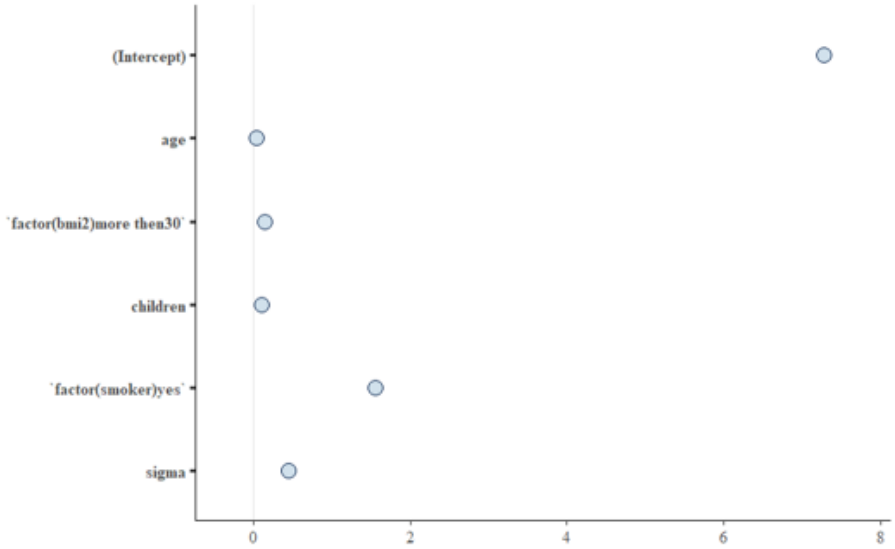


FIGURE 16 – Résultats du modèle. Notons que la transformation logarithmique écrase les coefficients des variables âge, nombre d'enfant, et imc.

On observe les coefficients et performances du modèle représentés dans la figure

16, qui montrent clairement les effets estimés des paramètres du Modèle 7. Il est manifeste que l'âge et le statut de fumeur ont des intervalles de crédibilité ne chevauchant pas zéro, ce qui indique que ces facteurs exercent une influence significative sur les charges médicales, même après prise en compte de l'effet des autres variables dans le modèle. De manière intéressante, les estimations pour l'imc et le nombre d'enfant révèlent également des effets significatifs.

La valeur de sigma, reflétant la variabilité des charges médicales non expliquée par les prédicteurs, est également estimée avec une précision raisonnable. Ces résultats fournissent une validation convaincante que le Modèle 7.

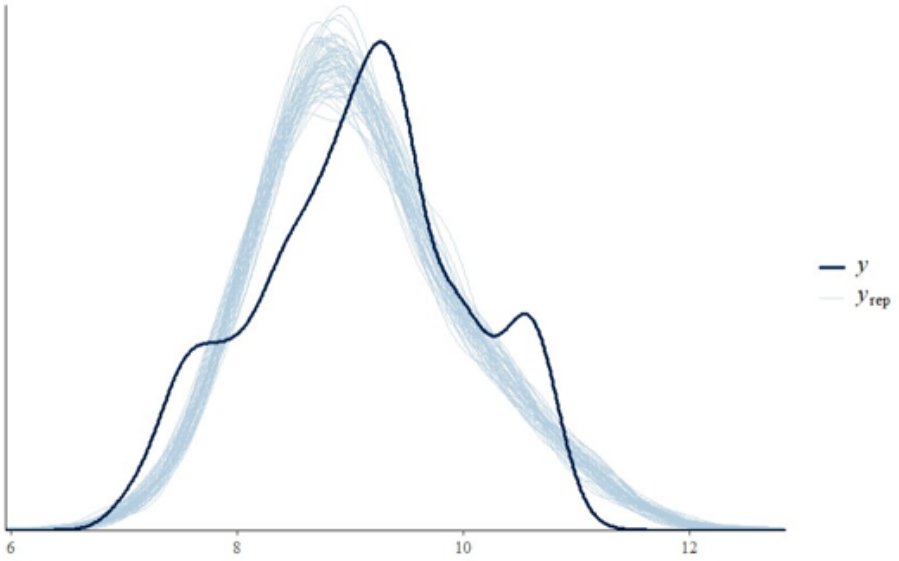


FIGURE 17 – Contrôle prédictif

La procédure de construction du postérieur, illustrée ci-dessus, offre une évaluation intuitive et puissante de la performance de notre modèle bayésien. Cette méthode, obtenue par la commande ppcheck sur notre modèle, génère des répliques de données simulées basées sur les distributions postérieures des paramètres du modèle. Comme indiqué par les lignes bleues claires, qui représentent une série de prédictions générées par le modèle, nous observons une adéquation avec la courbe noire, qui symbolise les données réelles observées. Cette congruence suggère que le modèle capture non seulement la tendance des données mais également leur dispersion, ce qui est essentiel pour la précision des prédictions dans les données médicales où la variabilité est inhérente.

3.2.3 Evaluation du pouvoir prédictif

L'évaluation des prédictions de notre modèle bayésien sélectionné, appliqué à un ensemble de données de test, révèle des insights précieux sur sa performance prédictive. Le tableau ci-dessous affiche un extrait de la comparaison entre les charges médicales actuelles et les prédictions moyennes générées à partir de la distribution postérieure.

Les mesures d'erreur pour les prédictions de notre modèle révèlent des écarts négligeables entre les valeurs prédites et les charges médicales réelles. Avec un MAE et un RMSE faibles (figure 19), il est évident que le modèle restitue convenablement la structure.

Le MRE particulièrement bas suggère que le modèle est bien ajusté et n'a pas besoin d'être modifié ou d'avoir des variables ajoutées. Ces résultats indiquent que le modèle actuel est déjà performant et ne nécessite pas de modifications majeures pour capturer avec précision la réalité des charges médicales

```
# Calcul de l'erreur absolue moyenne (MAE)
MAE <- mean(abs(log(test_data$charge) - predicted_means2))
print(paste("MAE:", MAE))

# Calcul de l'erreur quadratique moyenne (MSE)
MSE <- mean((log(test_data$charge) - predicted_means2)^2)
print(paste("MSE:", MSE))

# Calcul de l'erreur quadratique moyenne racine (RMSE)
RMSE <- sqrt(MSE)
print(paste("RMSE:", RMSE))

# Calcul de l'erreur relative moyenne (MRE)
MRE <- mean(abs((log(test_data$charge) - predicted_means2) / log(test_data$charge)) * 100)
print(paste("MRE:", MRE))
```

```
[1] "MAE: 0.28369559508535"
[1] "MSE: 0.196676525108429"
[1] "RMSE: 0.443482271470269"
[1] "MRE: 3.18360798992488"
```

FIGURE 18 – Calcul des erreurs

3.3 Comparaison avec une régression fréquentiste

Pour nous donner un aperçu de la performance de cette régression bayésienne, nous comparons celle-ci avec une régression fréquentiste. Nous utilisons un modèle linéaire simple avec comme variable cible les log-charges et comme features les va-

riables sélectionnées plus tôt. La regréssion présente un R^2 de 0.76 et une tendance à la sous-estimation du prix avec une médiane de -0.05 sur la prédiction de log-coût, ce qui donne donc un facteur de sous-estimation du prix réel de $\cdot e^{-0.05} - 1$.

```
data$bmi_binaire_30 <- ifelse(data$bmi >= 30, 1, 0)

data$logcharges <- log(data$charges)

model <- lm(logcharges ~ age  + children  + sex +
bmi_binaire_30+smoker, data = data)

summary(model)
```

```
Call:
lm(formula = logcharges ~ age + children + sex + bmi_binaire_30 +
    smoker, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92866 -0.19768 -0.05080  0.07652  2.09866

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.269196   0.041210 176.392  < 2e-16 ***
age            0.034808   0.000875  39.779  < 2e-16 ***
children       0.101820   0.010157  10.025  < 2e-16 ***
sexmale       -0.075426   0.024563  -3.071  0.00218 **
bmi_binaire_30 0.139656   0.024612   5.674 1.71e-08 ***
smokeryes      1.550832   0.030395  51.023  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4472 on 1332 degrees of freedom
Multiple R-squared:  0.7643,    Adjusted R-squared:  0.7635
F-statistic: 864.1 on 5 and 1332 DF,  p-value: < 2.2e-16
```

FIGURE 19 – Performances de la régression fréquentiste (en base logarithmique) basée sur la meilleure séléction de variables.

L'écart entre le prix prédit et le prix réel est donc de 5% pour la médiane. On peut comparer les coefficients de régression avec les médianes des distributions de coefficients des régression bayésienne. Dans notre cas, on constate que les deux méthodes convergent vers la même régression :

Les coefficients des variables quantitatives sont proches, avec un écart relatif de l'ordre de 1-3 %. Sur la variable fumeur, la précision atteint même une précision

Variable	Coeff fréquentiste	Coeff bayésien	différence relative en %
Age	0.03481	0.03359	3.6
BMI	0.13966	0.14207	1.7
fumeur	1.55083	1.551	0.01
enfants	0.10183	0.09932	2.5

TABLE 3 – Coefficients de régressions bayésiennes et fréquentistes

accrue, avec seulement 0.01 % de différence entre les deux régressions. Cela nous permet donc de renforcer la confiance que nous avons dans le précédent modèle et dans nos choix de priors.

Nous n'avons ici pas représenter la variable région dans la régression, celle-ci ne permet pas de correctement distinguer les coûts médicaux. En effet, bien que nous étions partis du postulat qu'il pouvait y avoir une influence, autant la régression bayésienne que fréquentiste nous a montré le contraire. Les coefficients obtenus pour celle-ci étaient d'ailleurs positifs lorsque l'on utilisait comme valeur de réference "southeast" et des variables binaires pour les autres régions. Ceci est en contradiction avec ce que nous avons vu dans l'analyse des variables qui indiquait une tendance à ce que les coûts soient plus élevés pour la region "southeast" et donc des coeffiencts négatifs pour les variables binaires représerntant les autres régions.

Nous utilisons maintenant des critères de performances afin d'évaluer objectivement les régressions réalisées.

3.4 Qualité des régressions

On a pu constater précemment qu'avec un bon prior, nos régressions bayésiennes et fréquentistes sont plutôt équivalentes puisqu'elles conduisent à des coefficients similaires. Ainsi, nos variables Table 3 ont un effet positif sur notre coût médical, c'est à dire que les facteurs age, BIM>30, être fumeur et avoir des enfants ont une tendance à augmenter la charge.

Il s'agit des coefficients fittant au mieux nos régressions linéaires. Néanmoins, reste à regarder la qualité de nos régressions. L'approche usuelle fréquentiste est de mesurer le RMSE et R^2, sur un set de train où on été entrainé nos modèles et un set de test pour vérifier la généralisation possible du modèle à de nouvelles données. Une approche visuelle courante est aussi de plot le nuage de points, nos charges en abscisse en fonction de nos charge prédite par le modèle linéaire en ordonné, de tracer $y = x$ pour estimer déjà visuellement à quel point nos données prédites sont proches des

vraies valeurs, c'est à dire le nuage de point autour de la droite identité. Concidérant que nous avons les même coefficients pour la régression bayésienne et fréquentistes il est inutile de faire ce travail 2 fois. Ainsi d'après Figure 19 nous obtenons un R^2 sans même split train test de 0.7 ce qui n'est pas fameux, elle n'est pas foncièrement mauvaise car on est relativement proche de 1, néanmoins nous en sommes encore loin. Pour ce qui est du RMSE, selon Figure 18, il avoisine 0.44 pour des données qui sont peut être centrées et réduites, mais l'absence de communication claire à ce sujet nous empêche de conclure s'il s'agit d'un bon ou d'un mauvais score. Si les données n'ont pas été normalisées, ce résultat pourrait être considéré comme très satisfaisant, étant donné que nos coûts s'échelonnent dans les milliers de dollars. En revanche, si les données ont effectivement été centrées et réduites, ce score de RMSE serait alors jugé très insuffisant.

Concernant les plots, à cause d'un manque de temps, il nous a été impossible de communiquer efficacement, de sorte à transmettre les données à la personne chargée d'évaluer les régressions. De ce fait cette évaluation des modèles restera superflue.

4 Conclusion

Pour conclure, on constate qu'avec un choix judicieux de prior, nos modèles bayésiens linéaires peuvent égaler, voire surpasser, les performances des régressions fréquentistes. Cependant, le jeu de données s'avère relativement limité en termes de covariables pour une estimation efficace du coût associé à un patient. En effet, l'inclusion de variables telles que l'activité professionnelle et les antécédents médicaux aurait été pertinente pour affiner l'analyse.

Quant à la question évoquée sur la région comme facteur de différence dans les charges médicales, nous avons pu constater que ce jeu de données généré artificiellement n'a pas été suffisamment bien construit pour être réellement représentatif des disparités régionales induites par des histoires différentes, telles que la proportion d'obèses, qui est plus élevée dans certains états.

De surcroît, les contraintes de temps imposées par ce projet ont compliqué l'approfondissement de notre étude. Dans ce laps de temps limité, il a été ardu de se familiariser avec les concepts non abordés dans le cours, ainsi que de se familiariser avec les outils spécifiques à la programmation bayésienne que nous avons dû totalement découvrir. Par ce fait, nous avons dû nous contenter d'une approche relativement simple, tant pour le développement des modèles que pour l'évaluation de leurs performances.

5 Bibliographie

1 : https ://perspective.usherbrooke.ca/bilan/servlet/BMPagePyramide/USA/2018/

2 : https ://www.cdc.gov/obesity/data/adult.html

3 : https ://fr.wikipedia.org/wiki/Sud$_profond$