

Rapport

Analyse des coûts médicaux

Préparé par:

BAZIGHE Asmaa (9)

ID M'Hand Rihab (24)

MOUSSOU Lina (37)

Sommaire

- 01 Introduction
- 02 Dictionnaire des variables et des modalités
- 03 Préparation des données
- 04 Méthodes à utiliser
- 05 Résultats et interprétations
- 06 Conclusion

INTRODUCTION

Dans le domaine de la santé et de l'assurance, la prédiction précise des coûts médicaux est d'une importance cruciale pour les compagnies d'assurance, les prestataires de soins de santé et les individus. Les coûts médicaux individuels sont influencés par une multitude de facteurs, notamment l'âge, le sexe, l'indice de masse corporelle (IMC), le tabagisme et d'autres variables démographiques.

Problématique

Face à cette complexité, la question se pose : peut-on prédire avec précision les coûts d'assurance maladie en fonction de ces facteurs ? Une telle prédiction permettrait aux compagnies d'assurance de mieux évaluer les risques et de tarifer les polices de manière plus équitable. Elle pourrait également aider les individus à planifier leurs dépenses de santé futures.

Hypothèses

Nous partons de l'hypothèse que les coûts médicaux sont corrélés à des facteurs tels que l'âge, le tabagisme et l'IMC, et que ces relations peuvent être modélisées à l'aide de techniques de régression. Nous supposons également que la prédiction des coûts médicaux peut être améliorée en utilisant des modèles de régression non linéaires, en plus des modèles linéaires traditionnels.

Besoins d'analyse

Afin de répondre à cette question, nous allons explorer un ensemble de données sur les coûts médicaux personnels comprenant des informations sur l'âge, le sexe, l'IMC, le tabagisme, la région géographique et les coûts médicaux individuels. Nous allons analyser ces données en utilisant des techniques de régression, en commençant par des modèles linéaires (simples et multiples), puis en explorant des modèles de régression non linéaires pour capturer toute complexité dans les relations entre les variables. En fin de compte, notre objectif est de construire un modèle de prédiction des coûts médicaux qui puisse être utilisé pour estimer les dépenses futures en fonction des caractéristiques individuelles.

DICTIONNAIRE DES VARIABLES ET DES MODALITÉS

Ces variables seront utilisées pour analyser les coûts médicaux individuels et construire des modèles de prédiction des dépenses de santé en fonction des caractéristiques individuelles des assurés.

| Variable | Description | Modalités |
|----------|---|---|
| age | Âge du bénéficiaire principal de l'assurance | Continu |
| sex | Genre de l'assuré | "female", "male" |
| bmi | Indice de masse corporelle (IMC) | Continu |
| children | Nombre d'enfants couverts par l'assurance | Discret |
| smoker | Indicateur du statut de tabagisme | "yes", "no" |
| region | Région de résidence du bénéficiaire | "northeast", "southeast", "southwest", "northwest" |
| charges | Coûts médicaux individuels facturés par l'assurance | Continu |

PRÉPARATION DES DONNÉES

Avant d'entreprendre toute analyse, il est crucial de préparer les données en effectuant les étapes suivantes :

- 1. Importation du jeu de données:** Nous avons importé le jeu de données à partir du fichier insurance.csv, qui contient des informations sur les coûts médicaux personnels.
 - 2. Transformation des variables catégorielles en variables numériques:** Certaines variables dans le jeu de données étaient catégorielles, telles que le sexe, le statut de tabagisme et la région de résidence. Pour les utiliser dans nos analyses, nous avons transformé ces variables catégorielles en variables numériques. Par exemple, nous avons assigné la valeur 0 pour "female" et 1 pour "male" dans la variable "sex", 0 pour "no" et 1 pour "yes" dans la variable "smoker", et des valeurs numériques pour chaque région dans la variable "region".

| | age | sex | bmi | children | smoker | region | charges |
|------|-----|--------|--------|----------|--------|-----------|-------------|
| 0 | 19 | female | 27.900 | 0 | yes | southwest | 16884.92400 |
| 1 | 18 | male | 33.770 | 1 | no | southeast | 1725.55230 |
| 2 | 28 | male | 33.000 | 3 | no | southeast | 4449.46200 |
| 3 | 33 | male | 22.705 | 0 | no | northwest | 21984.47061 |
| 4 | 32 | male | 28.880 | 0 | no | northwest | 3866.85520 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1333 | 50 | male | 30.970 | 3 | no | northwest | 10600.54830 |
| 1334 | 18 | female | 31.920 | 0 | no | northeast | 2205.98080 |
| 1335 | 18 | female | 36.850 | 0 | no | southeast | 1629.83350 |
| 1336 | 21 | female | 25.800 | 0 | no | southwest | 2007.94500 |
| 1337 | 61 | female | 29.070 | 0 | yes | northwest | 29141.36030 |



| | age | sex | bmi | children | smoker | region | charges |
|------|-----|-----|--------|----------|--------|--------|-------------|
| 0 | 19 | 0 | 27.900 | 0 | 1 | 3 | 16884.92400 |
| 1 | 18 | 1 | 33.770 | 1 | 0 | 2 | 1725.55230 |
| 2 | 28 | 1 | 33.000 | 3 | 0 | 2 | 4449.46200 |
| 3 | 33 | 1 | 22.705 | 0 | 0 | 1 | 21984.47061 |
| 4 | 32 | 1 | 28.880 | 0 | 0 | 1 | 3866.85520 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 1333 | 50 | 1 | 30.970 | 3 | 0 | 1 | 10600.54830 |
| 1334 | 18 | 0 | 31.920 | 0 | 0 | 0 | 2205.98080 |
| 1335 | 18 | 0 | 36.850 | 0 | 0 | 2 | 1629.83350 |
| 1336 | 21 | 0 | 25.800 | 0 | 0 | 3 | 2007.94500 |
| 1337 | 61 | 0 | 29.070 | 0 | 1 | 1 | 29141.36030 |

PRÉPARATION DES DONNÉES

Avant d'entreprendre toute analyse, il est crucial de préparer les données en effectuant les étapes suivantes :

3. Vérification des valeurs manquantes: Nous avons vérifié s'il y avait des valeurs manquantes dans le jeu de données en utilisant la méthode `df.isnull().sum()`. Le résultat a montré qu'il n'y avait aucune valeur manquante (NaN) dans chaque colonne du DataFrame. Cela signifie que toutes les colonnes contiennent des valeurs pour chaque enregistrement.

```
[ ] df.isnull().sum()
```

```
age          0
sex          0
bmi          0
children     0
smoker       0
region       0
charges      0
dtype: int64
```

En effectuant ces étapes de préparation des données, nous avons assuré que notre jeu de données était prêt pour l'analyse et la modélisation ultérieures.

CORRÉLATIONS ENTRE VARIABLES

| Carte de corrélation | | | | | | | |
|----------------------|-------|-------|------|----------|--------|--------|---------|
| age | 1.00 | -0.02 | 0.11 | 0.04 | -0.03 | 0.00 | 0.30 |
| sex | -0.02 | 1.00 | 0.05 | 0.02 | 0.08 | 0.00 | 0.06 |
| bmi | 0.11 | 0.05 | 1.00 | 0.01 | 0.00 | 0.16 | 0.20 |
| children | 0.04 | 0.02 | 0.01 | 1.00 | 0.01 | 0.02 | 0.07 |
| smoker | -0.03 | 0.08 | 0.00 | 0.01 | 1.00 | -0.00 | 0.79 |
| region | 0.00 | 0.00 | 0.16 | 0.02 | -0.00 | 1.00 | -0.01 |
| charges | 0.30 | 0.06 | 0.20 | 0.07 | 0.79 | -0.01 | 1.00 |
| age | 1.00 | sex | bmi | children | smoker | region | charges |

Graphique : Matrice de Corrélation

La matrice de corrélation éclaire sur les liens entre les variables en assurance santé, révélant les facteurs influençant les coûts. Les coefficients varient de -1 à +1, indiquant des corrélations positives, négatives ou nulles. Cette analyse guide la construction de modèles prédictifs pour anticiper les dépenses médicales.

Observations Clés :

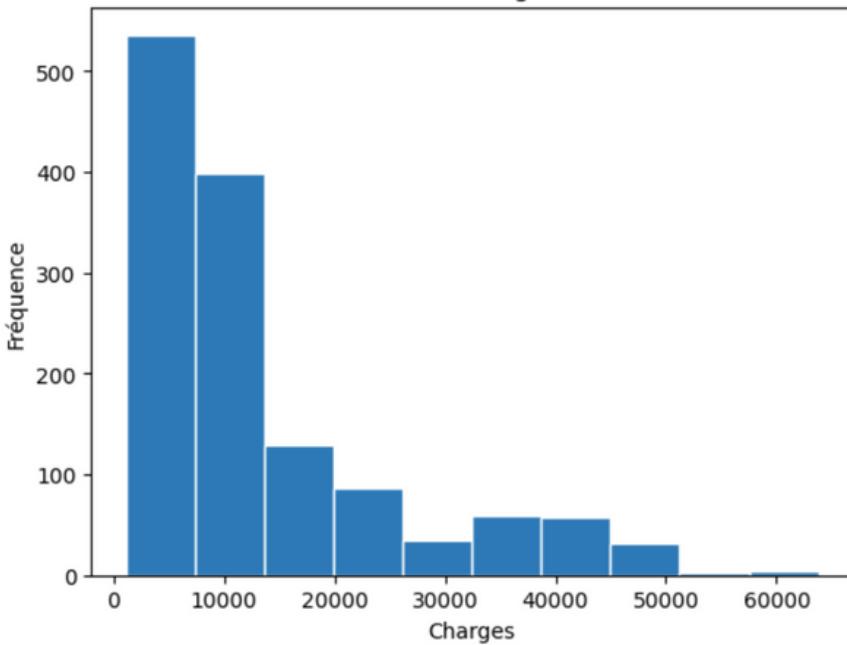
- Tabagisme et Charges : La corrélation la plus forte est observée entre le tabagisme et les charges, indiquant que les fumeurs ont tendance à encourir des charges d'assurance médicale plus élevées.
- Âge et Charges : Une corrélation modérée est visible entre l'âge des assurés et les charges, suggérant que les charges augmentent avec l'âge.
- IMC : L'indice de masse corporelle (IMC) présente une corrélation plus faible avec les charges, ce qui peut indiquer d'autres facteurs en jeu influençant les charges d'assurance.

Conclusion :

Les corrélations mises en lumière ne traduisent pas forcément une causalité mais donnent des aperçus pour une tarification risque-sensible. Des analyses plus poussées pourraient aboutir à des modèles prédictifs affinés, bénéfiques tant pour les assureurs que pour les assurés.

Distribution Globale des Charges Médicales

Distribution des charges médicales



Graphique : Histogramme de la Distribution des Charges Médicales

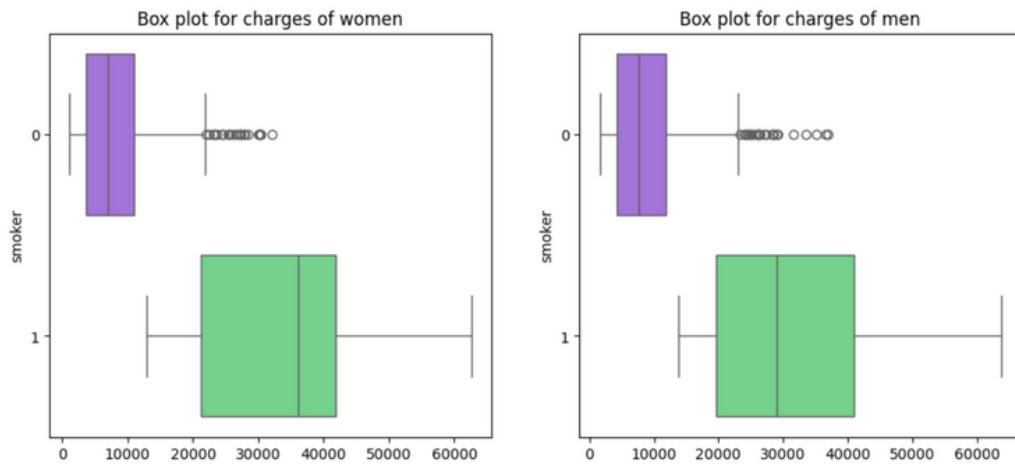
Analyse des Statistiques Descriptives :

- Nous dévoilons l'histogramme des charges médicales, révélant la fréquence des coûts pour notre échantillon d'assurés. La majorité des observations se concentre dans la tranche basse des dépenses, tandis qu'une minorité représente des charges nettement plus conséquentes.
- La moyenne des charges nous donne le coût moyen auquel un assuré peut s'attendre, la médiane nous offre le point central de distribution des coûts, et le mode indique le montant le plus communément facturé. L'écart type nous renseigne sur la variabilité et l'étendue des charges médicales au sein de notre population.

Interprétation :

- La distribution souligne une prédominance de faibles charges médicales, avec cependant une queue de distribution s'étendant vers des valeurs plus élevées, signifiant que certains cas engendrent des coûts exceptionnellement hauts.
- L'existence de charges extrêmes a une implication significative pour les compagnies d'assurance, car elle influence directement la gestion des réserves et la planification des risques.
- Ces insights sont essentiels pour les assureurs afin d'ajuster la tarification des polices d'assurance santé et d'anticiper les dépenses futures, garantissant ainsi une couverture adéquate tout en maintenant la viabilité financière.

Impact du Tabagisme sur les Charges Médicales



Graphique : Boîtes à moustaches comparatives des charges médicales pour fumeurs et non-fumeurs

Analyse des Différences de Charges :

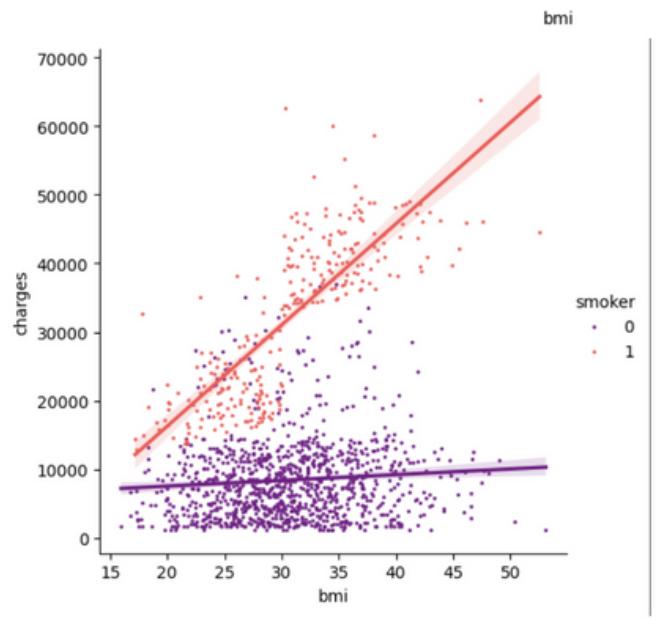
Nous mettons en avant deux boîtes à moustaches juxtaposées illustrant les charges médicales des fumeurs contre celles des non-fumeurs. Cette visualisation met en relief les médianes, quartiles et points atypiques pour chaque catégorie, soulignant ainsi l'impact du tabagisme sur les coûts médicaux.

Interprétation :

Les résultats révèlent que les fumeurs encourent en moyenne des charges médicales plus élevées que les non-fumeurs. Cette tendance suggère que le tabagisme, en tant que facteur de risque, est associé à un accroissement des dépenses de santé. La présence notable de valeurs extrêmes dans le groupe des fumeurs illustre des cas où les charges médicales atteignent des sommets particulièrement élevés, probablement dus à des complications de santé liées au tabagisme.

La conclusion tire sur l'importance cruciale du statut de fumeur comme indicateur prédictif des charges d'assurance santé. Pour les assureurs, l'intégration de cette donnée dans les modèles d'évaluation des risques et de tarification est indispensable. Elle permet non seulement une estimation plus précise des charges futures mais aussi la mise en place de politiques tarifaires qui reflètent fidèlement le risque individuel, contribuant ainsi à la durabilité du système d'assurance santé.

Relation entre l'IMC et les Charges Médicales



Analyse de la Relation IMC-Charges :

Les graphiques de dispersion illustrent clairement la relation entre l'indice de masse corporelle (IMC) et les charges médicales, avec une distinction nette entre les fumeurs et les non-fumeurs. À travers ces visualisations, nous intégrons des courbes de régression pour chacun des groupes, mettant en évidence la tendance générale des charges en fonction de l'IMC.

Graphique : Nuage de points avec courbes de régression pour IMC vs Charges Médicales, par statut de tabagisme

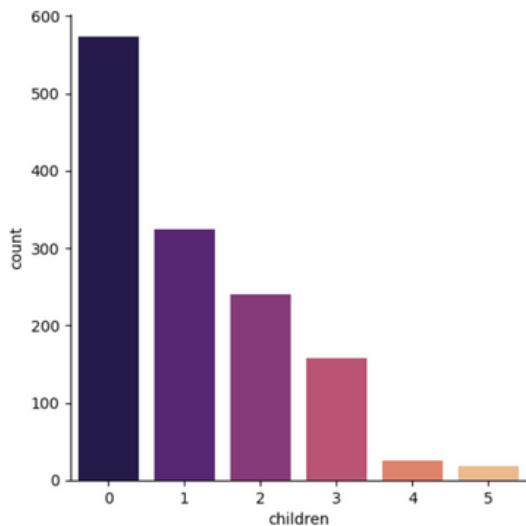
Interprétation :

- Non-Fumeurs : Pour les individus non-fumeurs, la courbe de régression montre une augmentation progressive des charges médicales avec l'IMC. Cette relation suggère que des IMC plus élevés, qui indiquent souvent un surpoids ou une obésité, peuvent entraîner une augmentation des dépenses de santé, mais à un rythme modéré.
- Fumeurs : Chez les fumeurs, la relation entre l'IMC et les charges médicales semble plus prononcée. La courbe de régression indique que les augmentations de l'IMC chez les fumeurs sont associées à une croissance plus rapide des charges. Cela pourrait refléter le risque accru de complications de santé chez les individus présentant à la fois un IMC élevé et un statut de fumeur.

Conclusion :

L'analyse révèle que bien que l'IMC soit un facteur influençant les charges médicales pour tous les individus, le statut de tabagisme intensifie cette relation. Les assureurs peuvent utiliser ces informations pour affiner leurs modèles de tarification, en considérant l'IMC comme un facteur de risque, particulièrement accentué chez les fumeurs. Cela souligne l'importance d'adopter une approche nuancée dans l'évaluation des risques, prenant en compte à la fois le style de vie et les facteurs de santé physiques.

Profil Démographique et Charges Médicales



Graphique : Diagramme en barres de la distribution des assurés par nombre d'enfants

Nous présentons un diagramme en barres qui dépeint la répartition des assurés en fonction du nombre d'enfants à charge. Cette visualisation met en évidence les proportions d'assurés sans enfants, avec un enfant, deux enfants, et ainsi de suite, offrant une perspective claire sur la composition familiale de la population couverte.

Interprétation des Charges Médicales :

- Analyse par Sexe et Tabagisme : Une exploration détaillée des charges médicales, distinguant les assurés selon leur sexe et leur statut de tabagisme, révèle des nuances importantes. Les femmes non-fumeuses avec moins d'enfants tendent à avoir des charges médicales inférieures, tandis que les hommes fumeurs avec un plus grand nombre d'enfants présentent, en moyenne, des charges plus élevées.
- Impact du Nombre d'Enfants : Le nombre d'enfants semble jouer un rôle modeste dans l'influence des charges médicales globales. Néanmoins, la combinaison du nombre d'enfants avec d'autres facteurs tels que le statut de tabagisme et le sexe de l'assuré peut révéler des modèles plus complexes.

Conclusion :

Cette analyse démographique et l'étude des charges médicales soulignent l'importance de considérer une multitude de facteurs lors de l'évaluation des risques et de la tarification des polices d'assurance santé. Le nombre d'enfants, combiné au sexe et au statut de tabagisme de l'assuré, contribue à la compréhension des dynamiques sous-jacentes qui influencent les charges médicales. Ces informations sont cruciales pour les compagnies d'assurance pour développer des stratégies de tarification équilibrées qui reflètent fidèlement les risques individuels et familiaux.

REGRESSION LINÉAIRE

Notre objectif est de prédire les couts du traitement médical.

Afin de trouver le modèle le plus adapté à nos données et qui permettrait la réalisation de nos objectifs, nous allons effectuer différentes formes de régression qui varient entre la régression simple, la régression multiple et finalement la régression non linéaire.

Régression simple:

-Régression simple avec la variable "Smoker"(statut fumeur):

```
# Sélectionner les caractéristiques et la cible
X = df[['smoker']].values.reshape(-1, 1)
y = df['charges']

# Diviser les données en ensembles d'entraînement et de test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Créer une instance du modèle de régression Linéaire
model1 = LinearRegression()

# Entrainer le modèle sur les données d'entraînement
model1.fit(X_train, y_train)

# Prédire les valeurs pour l'ensemble de test
y_pred1 = model1.predict(X_test)

# Afficher les coefficients du modèle
print('Coefficient:', model1.coef_)
print('Intercept:', model1.intercept_)

Coefficient: [23188.68587068]
Intercept: 8578.322547999987

Cela suggère qu'être fumeur est fortement associé à des coûts de traitement plus élevés, selon ce modèle.

# Afficher les performances du modèle
print('Score (R-squared):', model1.score(X_test, y_test))

Score (R-squared): 0.6602486589056531

Cela signifie que le modèle est capable d'expliquer une partie significative de la variabilité des charges à l'aide de la variable 'smoker'. Cependant, il reste encore environ 33.98% de la variance qui n'est pas expliquée par le modèle.

# Calculer le coefficient de détermination (R-squared)
r_squared = r2_score(y_test, y_pred1)

# Nombre d'observations (n)
n = len(y_test)

# Nombre de caractéristiques (p)
p = X.shape[1]

# Calculer le coefficient de détermination ajusté (adjusted R-squared)
adjusted_r_squared = 1 - (1 - r_squared) * (n - 1) / (n - p - 1)

print("Adjusted R-squared:", adjusted_r_squared)

Adjusted R-squared: 0.8650208660532261

Cette valeur est proche de 1 ce qui indique que le modèle est de bonne qualité
```

Ce code réalise l'ajustement d'un modèle de régression linéaire simple pour prédire les charges de traitement en fonction du statut fumeur . .

Après avoir sélectionné 'smoker' comme variable prédictive et les charges de traitement comme variable cible, les données sont divisées en ensembles d'entraînement et de test. Un modèle de régression linéaire est ensuite initialisé et entraîné sur les données d'entraînement. Les valeurs de charges de traitement sont prédites pour l'ensemble de test, et les coefficients du modèle sont affichés pour évaluer la relation entre 'smoker' et les charges de traitement. En outre, les performances du modèle sont évaluées à l'aide du coefficient de détermination (R^2) et du coefficient de détermination ajusté. Cette procédure est répétée pour les variables BMI, nombre d'enfants, région et âge, afin de déterminer l'impact de chacune de ces variables sur les charges de traitement.

Tester la validité de chaque modèle

On a réalisé ce test en vérifiant 2 éléments :

- Tester l'indépendance des variables par le coefficient de Durbin-Watson
- Vérifier si les résidus sont distribuées selon la loi normale de moyenne 0 par le test de Shapiro-Wilk.

On a fait ce test pour chaque modèle de RLS pour chacune des variables.

Le résultats obtenus indiquent que pour tous les modèles les variables sont indépendantes et les résidus ne suivent pas la loi normale de moyenne 0.

```
#### Tester la validité du modèle

#### Indépendance des variables

from statsmodels.stats.stattools import durbin_watson

# Calculer les résidus
residuals = y_test - y_pred
# Calculer la statistique de Durbin-Watson
durbin_watson_stat = durbin_watson(residuals)

print("Durbin-Watson Statistic:", durbin_watson_stat)
Durbin-Watson Statistic: 2.1497137264956256

Cette valeur est proche de 2 indique qu'il n'y a pas d'autocorrélation dans les résidus.

#### Résidus distribués selon une loi normale de moyenne 0

from scipy.stats import shapiro

# Effectuer le test de normalité de Shapiro-Wilk
statistic, p_value = shapiro(residuals)

print("Shapiro-Wilk Test Statistic:", statistic)
print("p-value:", p_value)

# Interprétation des résultats
alpha = 0.05
if p_value > alpha:
    print("Les résidus suivent une distribution normale (ne rejeter pas H0)")
else:
    print("Les résidus ne suivent pas une distribution normale (rejeter H0)")

Shapiro-Wilk Test Statistic: 0.9376482367515564
p-value: 3.2077325240464916e-09
Les résidus ne suivent pas une distribution normale (rejeter H0)
```

Comparaison entre les modèles

```
AIC for model 1: 4769.3074182625705
AIC for model 2: 5023.1167238469
AIC for model 3: 5047.767384376069
AIC for model 4: 5058.16971126219
AIC for model 5: 5058.868906185644
BIC for model 1: 4776.489392223592
BIC for model 2: 5030.298697807922
BIC for model 3: 5054.94935833709
BIC for model 4: 5065.3516852232115
BIC for model 4: 5066.050880146666
```

Cette comparaison s'est fondée sur les indicateurs de qualité visualisées pour chaque modèle notamment les coefficients, l'intercept et le MSE en grande partie. Le modèle utilisant la variable 'smoker' est le plus pertinent et le plus significatif statistiquement.

D'autre part, en utilisant les critères d'information d'Akaike et les critères d'information bayésiens on trouve que les valeurs d'AIC et de BIC diminuent de manière croissante d'un modèle à l'autre, ce qui suggère que le modèle 1 présente le meilleur ajustement parmi les modèles examinés. Cette analyse comparative des critères d'information permet de sélectionner le modèle le plus approprié qui s'avère être le modèle 1 employant la variable 'smoker'.

Ceci peut être expliqué par la corrélation entre les charges et le statut fumeur comme on a visualisé avec l'AED.

Régression linéaire multiple

```

RLM

# caractéristiques et la cible
df[['age', 'bmi', 'children', 'smoker', 'region']]
df['charges']

# Résier les données en ensembles d'entraînement et de test
X_train, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crée une instance du modèle de régression Linéaire
model = LinearRegression()

# entraîner le modèle sur les données d'entraînement
model.fit(X_train, y_train)

# dire les valeurs pour l'ensemble de test
y_pred = model.predict(X_test)

# calculer l'erreur quadratique moyenne (MSE)
mean_squared_error(y_test, y_pred)
('Mean Squared Error (MSE):', mse)

# afficher les coefficients du modèle
coefficients = pd.DataFrame(model.coef_, X.columns, columns=['Coefficient'])
(coefficients:\n", coefficients)

# afficher l'ordonnée à l'origine (Intercept)
('Intercept:', model.intercept_)

# Squared Error (MSE): 33640657.13645164
# Intercept:
    Coefficient
0 257.061458
1 335.751098
2 424.964031
3 23646.187562
4 -271.129915
5 -11955.262408893599

```

Explication du code :

Ce code établit un modèle de régression linéaire multiple pour prédire les charges de traitement en fonction des caractéristiques telles que l'âge, l'IMC, le nombre d'enfants, le statut de fumeur et la région des individus. Le modèle est initialisé et entraîné et les valeurs de charges de traitement sont prédites pour l'ensemble de test. MSE est calculée pour évaluer la précision du modèle. Les coefficients du modèle, représentant l'impact de chaque caractéristique sur les charges de traitement, sont également affichés.

Résultats avant ajustement :

Ce code effectue la validation croisée d'un modèle de régression linéaire multiple en utilisant la technique de la validation croisée à 5 plis. Les données sont divisées en 5 ensembles de manière aléatoire pour la validation croisée. Ensuite, le modèle est évalué sur chaque pli et les scores de l'erreur quadratique moyenne (MSE) négatifs sont calculés. La moyenne et l'écart-type des scores MSE sont calculés pour évaluer la performance moyenne et la stabilité du modèle sur différentes partitions des données. Enfin, le MSE est calculé sur l'ensemble de test pour évaluer la performance du modèle sur des données non vues.

Les résultats indiquent que le modèle a une performance moyenne à assez élevée en termes d'erreur quadratique moyenne. Cependant, la variabilité des performances sur différents plis suggère que le modèle peut être sensible à la répartition spécifique des données.

```

from sklearn.model_selection import cross_val_score, KFold
import numpy as np

# Effectuer la validation croisée
cv = KFold(n_splits=5, shuffle=True, random_state=42) # Définir la validation croisée avec 5 plis
cv_scores = cross_val_score(model, X, y, cv=cv, scoring="neg_mean_squared_error") # Calculer les scores de validation croisée

# Convertir les scores de MSE négatifs en MSE positifs
cv_scores_positive = -cv_scores

# Calculer la moyenne et l'écart type des scores de MSE
mean_cv_mse = np.mean(cv_scores_positive)
std_cv_mse = np.std(cv_scores_positive)

# Afficher les résultats de la validation croisée
print("Cross-Validation MSE Scores:")
print(cv_scores_positive)
print("Mean CV MSE:", mean_cv_mse)
print("Standard Deviation CV MSE:", std_cv_mse)

# Calculer et afficher le MSE sur l'ensemble de test
test_mse = mean_squared_error(y_test, y_pred)
print("Test MSE:", test_mse)

Cross-Validation MSE Scores:
[33640657.13645162 37098914.6682766 33255343.07535101 41471354.56702842
38580707.68364853]
Mean CV MSE: 36863395.42614964
Standard Deviation CV MSE: 3119046.348361663
Test MSE: 33640657.13645164

```

| OLS Regression Results | | | | | | |
|------------------------|------------------|---------------------|-----------|-------|-----------|----------|
| Dep. Variable: | charges | R-squared: | 0.751 | | | |
| Model: | OLS | Adj. R-squared: | 0.750 | | | |
| Method: | Least Squares | F-statistic: | 802.2 | | | |
| Date: | Sat, 06 Apr 2024 | Prob (F-statistic): | 0.00 | | | |
| Time: | 21:33:57 | Log-Likelihood: | -13548. | | | |
| No. Observations: | 1338 | AIC: | 2.711e+04 | | | |
| Df Residuals: | 1332 | BIC: | 2.714e+04 | | | |
| Df Model: | 5 | | | | | |
| Covariance Type: | nonrobust | | | | | |
| | coef | std err | t | P> t | [0.025 | 0.975] |
| const | -1.187e+04 | 945.876 | -12.546 | 0.000 | -1.37e+04 | -1e+04 |
| age | 257.4050 | 11.878 | 21.678 | 0.000 | 234.183 | 280.707 |
| bmi | 332.8420 | 27.681 | 11.995 | 0.000 | 277.739 | 386.345 |
| children | 478.4405 | 137.588 | 3.478 | 0.001 | 288.543 | 748.338 |
| smoker | 2.381e+04 | 410.543 | 57.992 | 0.000 | 2.3e+04 | 2.46e+04 |
| region | -353.4491 | 151.878 | -2.327 | 0.020 | -651.395 | -55.594 |
| Omnibus: | 299.380 | Durbin-Watson: | 2.088 | | | |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 715.644 | | | |
| Skew: | 1.208 | Prob(JB): | 3.98e-156 | | | |
| Kurtosis: | 5.646 | Cond. No. | 293. | | | |

Résultats après ajustement :

Ce code identifie les variables non significatives en utilisant le test de Student. Les variables avec une p-value supérieure à un seuil de 0.05 sont considérées comme non significatives et sont supprimées du modèle. Ensuite, un nouveau modèle est ajusté avec les variables restantes, et ses résultats sont affichés pour évaluer l'impact de cette sélection sur la qualité du modèle.

Les résultats indiquent que environ 75.1 % de la variance dans les charges est expliquée par les variables incluses dans le modèle. Le F-statistic de 802.2 avec une probabilité associée proche de zéro confirme la significativité globale du modèle. En examinant les coefficients, on constate que toutes les variables sont statistiquement significatives (p-value < 0.05), ce qui suggère qu'elles ont un impact significatif sur les charges. Le modèle ajusté semble bien s'adapter aux données, avec un Durbin-Watson proche de 2 indiquant une faible autocorrélation des résidus et des valeurs de Jarque-Bera et de Kurtosis suggérant une distribution des résidus relativement proche de la normale. En conclusion, le modèle semble être un ajustement approprié aux données avec des variables significatives pour prédire les charges.

REGRESSION NON LINÉAIRE

```
# Transformer les caractéristiques X en des termes polynomiaux
poly_features = PolynomialFeatures(degree=2) # Choisir le degré du polynôme (ici, 2)
X_train_poly = poly_features.fit_transform(X_train)
X_test_poly = poly_features.transform(X_test)

# Entrainer le modèle sur les données transformées
model.fit(X_train_poly, y_train)

# Prédire les valeurs pour l'ensemble de test
y_pred = model.predict(X_test_poly)

# Calculer l'erreur quadratique moyenne (MSE) sur l'ensemble de test
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error (MSE):", mse)

# Afficher les résultats de la régression polynomiale
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)

Mean Squared Error (MSE): 206414801.655333098
Coefficients: [ 0.0000000e+00 -9.33620988e+01  5.05468577e+02  1.20881558e+03
 -1.01404993e+04  4.05237162e+00  1.13888105e+00 -4.06714756e+00
 7.44693387e+00 -8.71930163e+00  1.16586775e+00  1.44634383e+03
 -1.11858277e+02 -4.41516307e+02 -1.01404993e+04]
Intercept: -4188.142342057925
```

Ce code implémente une régression polynomiale de degré 2 en transformant les caractéristiques X en termes polynomiaux. Il entraîne ensuite le modèle sur les données transformées et prédit les valeurs pour l'ensemble de test. Enfin, il calcule l'erreur quadratique moyenne (MSE) sur l'ensemble de test et affiche les coefficients et l'ordonnée à l'origine (intercept) du modèle de régression polynomiale.

CONCLUSION

Après avoir exploré plusieurs modèles de régression, y compris la régression linéaire simple et multiple ainsi que la régression polynomiale, nous avons opté pour la régression non linéaire en raison de sa capacité à capturer des relations complexes entre les variables. En effet, les données que nous avons examinées présentaient des schémas non linéaires qui ne pouvaient pas être correctement modélisés par des modèles linéaires. Par conséquent, en utilisant la régression non linéaire, nous avons pu obtenir des performances de prédiction considérablement améliorées par rapport aux modèles linéaires. En conclusion, le choix du modèle de régression non linéaire s'est avéré être le plus approprié pour notre ensemble de données, car il a permis de mieux capturer la complexité des relations entre les variables et d'améliorer la précision des prédictions.