Contents

1. 準備

1.1 背景

このデータセットは、個人の属性（年齢、性別、BMI など）と、その人が支払った年間医療費（保険金請求額）を含んでいます。私たちの目標は、これらの属性から医療費を予測するモデルを構築し、どの属性が医療費に強く影響を与えるのかを明らかにすることです。

データセットは Kaggle で公開されているものを使用します。

https://www.kaggle.com/datasets/mirichoi0218/insurance

1.2 使用データとライブラリ

- **データファイル**: `insurance.csv`
- **使用ライブラリ (例)**: `numpy`, `pandas`, `seaborn`, `matplotlib`, `japanize-matplotlib`, `statsmodels`, `scikit-learn`

1.3 データの読み込み

Google Colab に `insurance.csv` ファイルをアップロードし、`pandas` で DataFrame として読み込みます。

```
import pandas as pd
import io
from google.colab import files

# Google 環境でのファイルアップロードColab
uploaded = files.upload()
file_name = next(iter(uploaded))
df = pd.read_csv(io.BytesIO(uploaded[file_name]))
```

Part I
探索的データ分析 (EDA)

2. データの概要把握

データ分析の最初のステップは、データを深く理解することです。データセットの構造、基本的な統計量、変数間の関係性を可視化によって明らかにします。

(1) データの確認

https://www.kaggle.com/datasets/mirichoi0218/insurance のページで各列（変数）の意味を確認してください。その後、head() メソッドを使ってデータの最初の 5 行を表示し、各列にどのようなデータが格納されているかを確認・説明してください。

(2) データ構造の確認

info() メソッドを使い、各列のデータ型と欠損値の有無を確認してください。結果をレポートに示し、欠損値がないことを確認した上で、データが分析に適した状態か考察してください。

(3) 要約統計量の確認

describe() メソッドを使い、数値データ列の基本的な統計量（平均、標準偏差、四分位数など）を算出してください。特に、charges（医療費）の平均値と中央値（50%タイル）に大きな差がある点に注目し、どのような分布をしているか推測してみましょう。

2.1 データの可視化

(4) 変数間の関係の概観

seaborn の pairplot 関数を使い、データセット内の数値変数間の関係性を一覧で表示してください。(注: データ件数が多いため、描画に少し時間がかかる場合があります。) この図から、目的変数である charges（医療費）と他の変数（特に age, bmi）との間にどのような関係が見られるか、全体的な傾向を説明してください。

(5) 喫煙と医療費の関係

喫煙の有無が医療費に与える影響は大きいと予想されます。seaborn の boxplot（箱ひげ図）を使い、喫煙者と非喫煙者で医療費の分布がどのように異なるか可視化してください。この図から読み取れることを具体的に説明してください。

(6) 年齢・喫煙と医療費の関係

(4) で見た年齢と医療費の関係を、喫煙の有無で層別化して詳しく見てみましょう。seaborn の scatterplot（散布図）で、横軸を age、縦軸を charges とし、hue='smoker' を指定して喫煙の有無で点の色を分けてプロットしてください。この図から、年齢と医療費の関係が喫煙者と非喫煙者でどのように異なるか考察してください。

(8) 地域の分布

`region`（地域）ごとのデータ件数に偏りがないか確認します。`seaborn` の `countplot` で可視化し、各地域のデータがほぼ均等に含まれていることを確認してください。

(追加) そのほかの特徴

(4) で描いた結果や、そのほかの結果から、このデータについて何か特筆すべきことがあればここに記述してください。

Part II
予測モデルの構築と評価

3. データ前処理の考え方：カテゴリ変数の扱い

私たちが構築する線形回帰モデルは、数式に基づいて予測を行います。例えば、年齢と BMI だけで医療費を予測するモデルは、以下の数式で表せます。

$$\text{charges} = (\beta_1 \times \text{age}) + (\beta_2 \times \text{BMI}) + \beta_0$$

この式が示す通り、モデルが扱う変数はすべて**数値**である必要があります。コンピュータは`'female'`や`'southwest'`といった文字列を直接計算に使うことはできません。

この問題を解決するのが**ダミー変数 (Dummy Variables)** です。これは、カテゴリ変数を「Yes(1) / No(0)」で表現される複数の数値変数に変換する手法です。

例えば、4 つのカテゴリ（'northeast', 'northwest', 'southeast', 'southwest'）を持つ `region` 列は、以下のように 3 つのダミー変数に変換されます。（1 つは基準として残します）

- `region_northwest`: この人は北西地域の人ですか？ → Yes(1)/No(0)
- `region_southeast`: この人は南東地域の人ですか？ → Yes(1)/No(0)
- `region_southwest`: この人は南西地域の人ですか？ → Yes(1)/No(0)

この変換により、モデルは各地域が（基準となる `northeast` に比べて）医療費にどれだけ影響を与えるかを、それぞれ独立した係数として学習できるようになります。

補足　後ほど使う `statsmodels` というライブラリでは、モデル式の中に `C(sex)` のように書くだけで、このダミー変数への変換を自動的に内部で行ってくれます。そのため、**今回は手動でダミー変数に変換するコードを書く必要はありません。** しかし、モデルの裏側で何が行われているかを理解することは、結果を正しく解釈するために非常に重要です。

4. ベースラインモデルの構築

4.1 データセットの分割

モデルの性能を正しく評価するため、データをモデル学習用の「訓練データ」と、性能評価用の「テストデータ」に分割します。`scikit-learn` の `train_test_split` 関数を使い、元の DataFrame を訓練データ（80%）とテストデータ（20%）に分割してください。結果の再現性を確保するため、`random_state=42` と指定します。

```
from sklearn.model_selection import train_test_split

df_train, df_test = train_test_split(df, test_size=0.2, random_state=42)
```

4.2 モデルの学習と結果の解釈

`statsmodels` を使い、線形回帰モデルを構築します。

```
import statsmodels.formula.api as smf

# 1. モデル式を定義
# カテゴリ変数は C() で囲むことで、statsmodels が自動でダミー変数化してくれる
formula = 'charges ~ age + bmi + children + C(sex) + C(smoker) + C(region)'

# 2. 訓練データでモデルを学習
model = smf.ols(formula, data=df_train).fit()

# 3. 学習結果のサマリーを表示
print(model.summary())
```

(8) モデル結果の解釈

(a) 表示されたサマリー表から**決定係数** (R^2) を読み取り、このモデルが訓練データの医療費のばらつきをどの程度説明できているか考察してください。

(b) **回帰係数** (`coef`) の表を確認し、各変数が医療費に与える影響を説明してください。例えば、「年齢が 1 歳上がると医療費は平均で約 XX ドル増加/減少する」や「喫煙者は非喫煙者に比べて医療費が平均で約 XX ドル高い/低い」のように、具体的な数値を挙げて説明しましょう。

(9) 予測性能の評価（テストデータ）

モデルが未知のデータに対しても同様の性能を発揮できるか（汎化性能）を確認します。学習済みモデルを使い、**テストデータ**（`df_test`）に対する予測を行い、以下の 2 つの指標を計算してモデルの性能

を評価してください。

- **決定係数 (R^2)**: 訓練データの結果と比較してどう変化したか。
- **RMSE (Root Mean Squared Error)**: 予測誤差の平均的な大きさを表す指標。

5. モデルの改善：目的変数の対数変換

EDA の (4) で見たように、charges の分布は右に長く裾を引いており、一部に非常に高額なデータが存在します。このような場合、目的変数を対数変換することで、分布が正規分布に近くなり、モデルの仮定（特に誤差の正規性や等分散性）が満たされやすくなることで、性能が向上することがあります。

対数モデルの構築

まず、訓練データとテストデータの両方に、charges を対数変換した新しい列 log_charges を作成します。np.log1p は $log(1+x)$ を計算し、0 を安全に扱えます。

```
import numpy as np
df_train['log_charges'] = np.log1p(df_train['charges'])
df_test['log_charges'] = np.log1p(df_test['charges'])

# 新しいモデル式を定義
formula_log = 'log_charges ~ age + bmi + children + C(sex) + C(smoker) + C(
    region)'

# 対数モデルを学習
model_log = smf.ols(formula_log, data=df_train).fit()

# 結果のサマリーを表示
print(model_log.summary())
```

サマリー表から、この対数モデルの**修正済み決定係数 (Adj. R-squared)** が、前のベースラインモデルと比較してどのように変化したか確認してください。

(11) 残差プロットによるモデル診断

モデルの仮定が満たされているか、残差プロットで確認します。

(a) **ベースラインモデル**について、横軸に**予測値**、縦軸に**残差**をとった散布図を描画してください。

(b) **対数モデル**についても同様に、横軸に**予測値 (対数スケール)**、縦軸に**残差**をとった散布図を描画してください。

(c) 2 つのプロットを比較し、対数変換によって残差のばらつき（ラッパ状の広がりなど）が改善されたか、誤差の等分散性が満たされていると言えるか、あなたの意見を述べてください。

6. モデルの改善：交互作用項の追加

EDA の (6) で、年齢と医療費の関係は喫煙の有無によって大きく異なることが示唆されました。これは、年齢の効果が喫煙ステータスによって変わる、すなわち「交互作用」がある可能性を示しています。この効果をモデルに組み込みます。

(12) 交互作用モデルの構築

‘age‘と ‘C(smoker)‘の交互作用項 (‘age:C(smoker)‘) をモデル式に追加して、新しいモデルを学習させてください。

```
formula_interaction = 'log_charges ~ age + bmi + children + C(sex) + C(
    smoker) + C(region) + age:C(smoker)'

model_interaction = smf.ols(formula_interaction, data=df_train).fit()

print(model_interaction.summary())
```

このモデルの**修正済み決定係数**を、(10) の対数モデルと比較し、交互作用項の追加がモデルの当てはまりを改善したか評価してください。また、追加された交互作用項 ‘age:C(smoker)[T.yes]‘ の係数が何を意味するのか考察してください。

7. 最終モデルの選択と評価

(13) 最終モデルの選択

これまで構築した 3 つのモデル（ベースライン、対数、交互作用項付き対数）の性能を、以下の観点から総合的に比較し、どれを最終モデルとして採用すべきか、理由とともに論じてください。

- **訓練データでの当てはまりの良さ**: 修正済み決定係数
- **テストデータでの汎化性能**: RMSE

(14) 最終モデルによる性能評価

(13) で選択した最終モデルを使い、**テストデータ**に対する予測性能を RMSE で最終評価します。対数スケールで予測した場合、予測値を `np.expm1()` で元のドル単位のスケールに戻してから、実際の `charges` と比較して RMSE を計算してください。この最終的な RMSE が、モデルの平均的な予測誤差となります。

Part III
統計的推測による不確実性の評価

これまでは、手元のデータを「真実」として最適な予測モデルを探求してきました。しかし、現実のデータ分析では、手元のデータはより大きな「母集団」から得られた一部の「標本」に過ぎません。ここでは、その「標本から得た結論がどの程度確からしいのか」という不確実性を評価します。ここでは、`insurance.csv` の全データ（1338 件）に対して、ブートストラップ法を用いて推定された値の不確実性を評価します。

7.1　標本と母集団：統計量のばらつき

(15) 母平均と標本平均

全データ（母集団）の `charges` の平均値（母平均 μ）を計算してください。次に、この母集団から大きさ $n = 200$ の標本を 1 回だけ無作為に抽出し、その標本平均 $\bar{x}$ を計算してください。両者の値はなぜ一致しないのかを説明しなさい。

(16) 標本平均の分布（中心極限定理）

(15) の標本抽出（$n = 200$）と標本平均の計算を 1000 回繰り返します。得られた 1000 個の標本平均の分布をヒストグラムで可視化してください。この分布の中心が母平均に近くなることを確認し、なぜ多くの標本平均が μ の周りに集まるのか、中心極限定理の概念を調べて、その考え方に触れつつ説明をしなさい。

7.2　ブートストラップ法による回帰係数の区間推定

私たちが計算した、回帰係数も、あくまで「手元の標本から計算された推定値」です。もし、全く別の標本が手に入っていたら、係数はどの程度変動するのでしょうか？この不確実性を、ブートストラップ法を用いて評価します。

(17) 復元抽出による回帰係数の計算

全データ（母集団）から、$n = 1338$ の標本を**無作為に 1 つ**復元抽出してください。この標本に対し、(14) のモデルの選択で選んだ **モデル式**を当てはめ、`C(smoker)[T.yes]` の係数を計算しなさい。

(18) 複数回の復元抽出による回帰係数の計算

全データ（母集団）から、$n = 1338$ の標本を**無作為に 1 つ**復元抽出を 1000 回行い、`C(smoker)[T.yes]` の係数を計算し、その平均と分散を計算した結果を書きなさい。また、回帰係数の分布をヒストグラムで可視化しなさい。

(19) 回帰係数のブートストラップ信頼区間の計算

1000 個の係数を小さい順に並べ、2.5 パーセンタイル値と 97.5 パーセンタイル値を求めることで、**ブートストラップ 95%信頼区間**を算出しなさい。その上で、この 95%信頼区間が何を意味するのか説明しなさい。

(20) 回帰係数のブートストラップ信頼区間に基づく推論

(19) で計算した区間に「0」が含まれているかを確認し、その結果から「喫煙が医療費に与える影響は統計的に有意と言えるか」、その結論の不確実性（推定値のばらつき）も考慮に入れて論じなさい。