SUFFOLK UNIVERSITY
SAWYER BUSINESS SCHOOL

**ISOM 835 Predictive Analytics and Machine Learning**

**Final project Phase 1**

**Dataset Description**

Stanley Mulokere

Suffolk University

# Executive Summary

This study first identified potential confounding factors through description and correlation analysis, and then used a stepwise regression model to control variables such as age, BMI, sex, children, and region from simple to complex to verify whether smoking is the independent and strongest influencing factor of medical expenditures. Model diagnosis ensures the reliability of OLS, and then the conclusion shows that smoking always remains the most stable and significant influence, with strong explanatory power and policy significance.

## Table of Contents

# 1.0 Introduction

The rapid rise in medical costs is one of the most significant challenges which US healthcare system is facing. As the population ages, the cost of medical resources increases, and the pressure on the insurance system also continues to increase, the unpredictability of personal medical expenditures has become an important factor affecting family financial security and the stability of the social security system. Therefore, how to scientifically assess the medical cost risks of different individuals is an important issue that the medical insurance industry, government health departments, and academic research fields all focus on.

Under this circumstance, individual health status (such as BMI, smoking behavior), demographic characteristics (such as age, gender), family structure, and regional differences are considered to be the main dimensions that affect medical expenses. However, there are complex correlations and potential confounding effects between different variables, which makes it particularly important to build a predictive model with explanatory power and logical rigor.

This study uses the Medical Cost Personal Dataset on the Kaggle platform as the basis to analyse. Then, descriptive analysis, correlation analysis, and multiple linear regression methods used comprehensively explore the key factors affecting medical insurance expenses. It aims to identify the main drivers behind expenditure differences and evaluate whether the behavioral variable of smoking is still the most significant influencing factor after controlling for multiple confounding factors.

## 1.1 Motivation

Our group choose this topic because of following reasons:

First, medical cost prediction has significant practical significance.

Medical costs in the United States continue to rise, and insurance companies are facing increasing actuarial difficulties and rising cost pressures. Studying the determinants of medical expenses can help companies formulate more scientific rate strategies, and can also allocate resources more efficiently.

Second, this research has both academic and methodological values. The cost of insurances are affected by muti-factors and thus well suit to systematically analysis using regression models. Controlling confounding variables, analyzing relationships between variables, and testing the independent impact of health behaviors (such as smoking) on medical costs in statistical models can demonstrate complete causal inference and statistical modeling logic.

Third, Kaggle data is open and transparent, has complete variables, and is widely used in reproducible research.

## 1.2 Problem Statement

Although health care spending is influenced by many factors, these factors can be highly correlated with each other. For example, smokers may have higher BMI; age may be related to health status and family structure.

Therefore, this study poses the following core research questions:

- What factors most significantly affect annual health insurance costs?
- When controlling for other variables, is smoking still the strongest independent influencing factor?

The specific research questions are expanded as follows:

- Does smoking significantly increase medical costs?
- Is the effect independent of age, BMI and region?
- What is the size of the role of age and BMI in health care costs?
- Does gender affect health care costs? Or does its effect disappear after controlling for confounding variables?
- Do regional differences result in significantly different costs?

# 2.0 Dataset & Variables overview:

## 2.1 Dataset Overview

- Kaggle dataset: Medical Cost Personal Dataset
- The sample size: 1338 people in total
- Includes demographic characteristics, health behaviors and annual insurance costs
- No missing values, no additional cleaning required

## 2.2 Variables Description

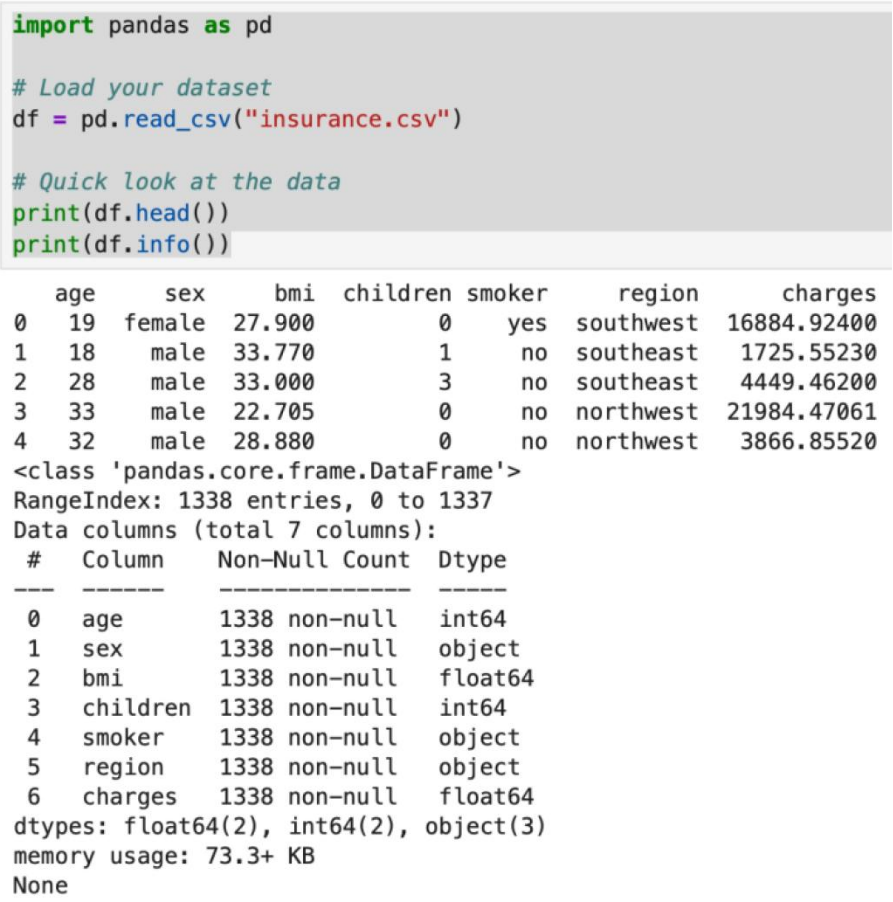
| Variables | Types | Description | Possible mechanism |
|---|---|---|---|
| Age | Numerical | Age | The older you are, the higher your risk of chronic disease |
| Sex | Categorical | F/M | May affect BMI, smoking rates |
| Bmi | Numerical | Body Mass Index, which measures the relationship between weight and height (kg/m²) | Reflects the degree of obesity and affects disease risk |
| children | Numerical | Number of children | Family structure affects medical expenditures |
| Smoker | Categorical | Whether to smoke | Related to respiratory and cardiovascular diseases |
| Region | Categorical | Where to live | Price differences in medical resources |
| Charges | Numerical | Cost of insurance | Model prediction target |

## 2.3 Data Processing

This study performed necessary data preprocessing on the original insurance data set to ensure that all variables can be correctly identified and used by the statistical model. Based on the original data structure and subsequent analysis requirements, this study mainly carried out two preprocessing tasks: (1) Dummy variable coding of categorical variables (2) Numerical processing of Boolean variables.

### 2.3.1 Data Format Checking

First use df.head() and df.info() to check the original data:



```python
import pandas as pd

# Load your dataset
df = pd.read_csv("insurance.csv")

# Quick look at the data
print(df.head())
print(df.info())
```

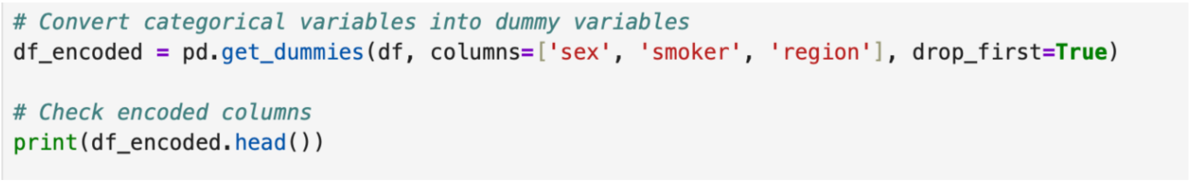
```
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       1338 non-null   int64  
 1   sex       1338 non-null   object 
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64  
 4   smoker    1338 non-null   object 
 5   region    1338 non-null   object 
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None
```

We can see from the output:

The data has a total of 1338 rows and 7 columns. The Non-null Count of all variables is 1338, indicating that there are no missing values in the data set. Age, bmi, children, and charges are numerical variables (int64 or float64) and can be directly used in regression modeling, while sex, smoker, and region are of object type and are typical categorical variables that require further coding. Because there are no missing values and wrong data types, this study does not require row deletion or missing value filling, and subsequent analysis can be carried out while maintaining the integrity of the sample size. That means there is no bias in coefficient estimate by sample loss due to data cleaning.

### 2.3.2 Dummy Encoding with Baseline

The reason why we need dummy encoding is that the regression model requires all independent variables to be numerical and it cannot directly handle strings (such as "male", "female"). Therefore, the categorical information must be converted into numerical form. Through dummy variables, the model can learn the marginal impact on expenses. For the three column text variables of sex, smoker, and region, we use the get dummies() function of pandas to convert them into 0/1 dummy variables:



```
# Convert categorical variables into dummy variables
df_encoded = pd.get_dummies(df, columns=['sex', 'smoker', 'region'], drop_first=True)

# Check encoded columns
print(df_encoded.head())
```

After this step, the original 3 categorical variables were coded into the following dummy variables:

sex_male: 1 = male, 0 = female (baseline)

smoker_yes: 1 = smoker, 0 = non-smoker (baseline)

regionnorthwest, regionsoutheast, region_southwest

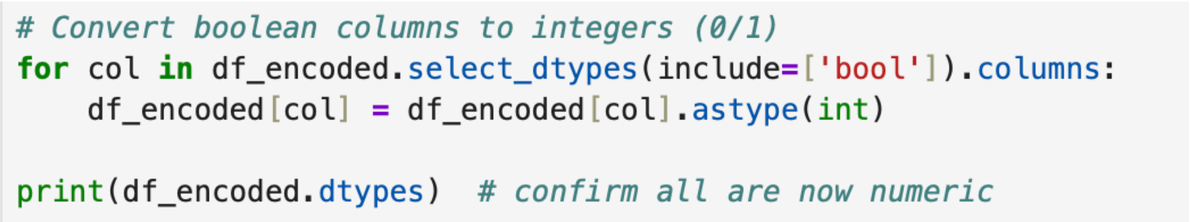
baseline is region_northeast (Northeast region)

What deserves to be mentioned is that the purpose of setting drop first=True. For instance, if each category variable retains all dummy variables(such as, the 4 regions of region), it will appear that the sum of 4 dummies is always equal to 1. This leads to the so-called Dummy Variable Trap, causing perfect multicollinearity and making the regression coefficients impossible to uniquely estimate. Therefore, we drop the first category in each group and treat it as the baseline, and the coefficients of the remaining categories are interpreted as differences relative to the baseline group. Additionally, the selection of baseline is not limited to the first value. The core principle is to choose a value that has reference significance and can reflect the normal state. The first value chosen is just the most trouble-free default choice.

This encoding method has two benefits. One is that can avoid multicollinearity and make the model stable and estimable. Another is that each coefficient can be clearly interpreted as how much more or less money is spent compared to the benchmark group.

### 2.3.3 Convert Boolean Variables to Numeric Types (True/False → 0/1)

The dummy variables generated by get dummies() in pandas is often bool (True/False). In order to be consistent with other numeric variables and facilitate subsequent calls to stats models to build OLS models, we subsequently convert all Boolean type columns to integer type 0/1:



```
# Convert boolean columns to integers (0/1)
for col in df_encoded.select_dtypes(include=['bool']).columns:
    df_encoded[col] = df_encoded[col].astype(int)

print(df_encoded.dtypes)  # confirm all are now numeric
```

After conversion, all dummy variables become int64. The entire data frame (df_encoded) only includes two numerical types: int64 and float64. This is very important for modeling. Because many statistical modeling functions assume that the independent variables are numeric. If the bool type is retained, although it can sometimes run, it will increase the risk of errors and cause trouble to visualization and further analysis.

### 2.3.4 Summary of Preprocessing

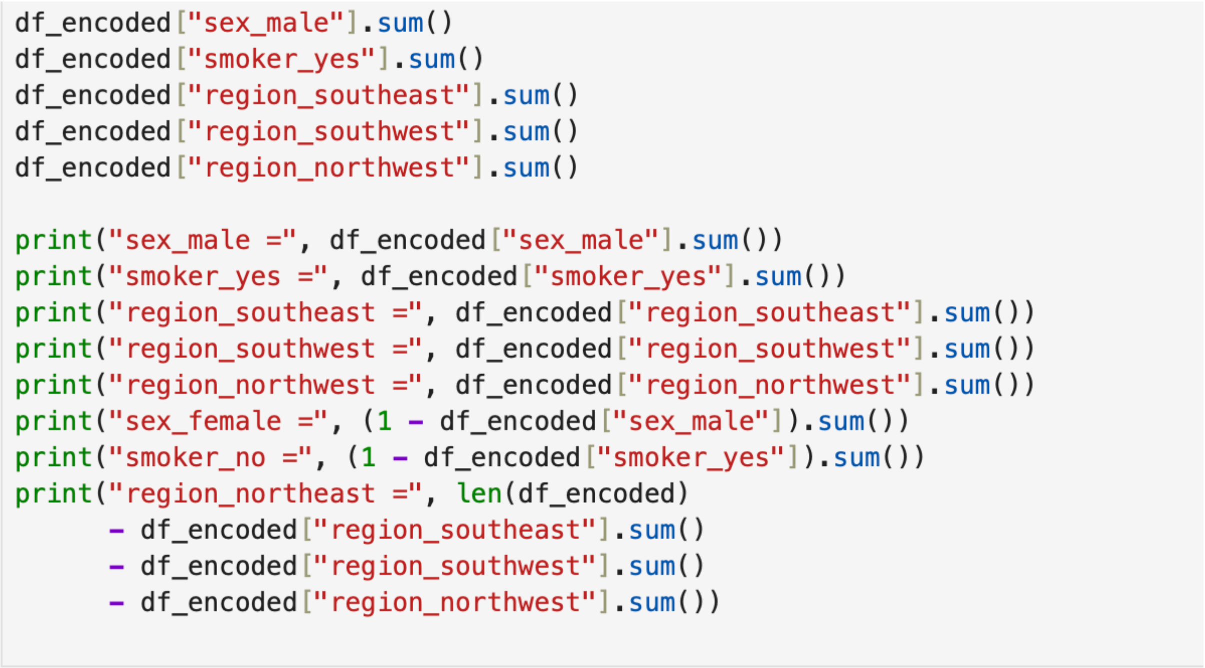
The data preprocessing completed includes two parts: dummy variable coding of categorical variables and neuralization of Boolean variables. Through these steps, all non-numeric variables were successfully converted into standard numerical variables that can be used for statistical model analysis. The preprocessed data structure is clearer and more standardized, laying a stable data foundation for following correlation analysis and regression modeling.

# 3.0 Descriptive Analysis

Descriptive statistics is to understand the sample structure, identify data characteristics, observe potential patterns, and determine which variables may be related to medical expenses. Therefore, this part will provide a general and preliminary description of the main variables of the data set, such as the distribution of categorical variables and statistical characteristics of numerical variables. Then, constructing the characteristics of the overall structure, so as to establish a basic understanding for correlation analysis and regression model construction.

## 3.1 Categorical Variable Distribution

We conducted statistics on gender, smoking status, and regional distribution. The following is the actual count result of dummy variables:



```
df_encoded["sex_male"].sum()
df_encoded["smoker_yes"].sum()
df_encoded["region_southeast"].sum()
df_encoded["region_southwest"].sum()
df_encoded["region_northwest"].sum()

print("sex_male =", df_encoded["sex_male"].sum())
print("smoker_yes =", df_encoded["smoker_yes"].sum())
print("region_southeast =", df_encoded["region_southeast"].sum())
print("region_southwest =", df_encoded["region_southwest"].sum())
print("region_northwest =", df_encoded["region_northwest"].sum())
print("sex_female =", (1 - df_encoded["sex_male"]).sum())
print("smoker_no =", (1 - df_encoded["smoker_yes"]).sum())
print("region_northeast =", len(df_encoded)
      - df_encoded["region_southeast"].sum()
      - df_encoded["region_southwest"].sum()
      - df_encoded["region_northwest"].sum())
```

**(1) Sex distribution:**

| Types | Sample value | Proportion |
|---|---|---|
| Male (sex_male = 1) | 676 | 50.5% |
| Female (baseline sex_female) | 662 | 49.5% |

Insight: the gender distribution is very balanced, indicating that gender will not affect model estimation due to sample bias.

**(2) Smoker**

| Types | Sample value | Proportion |
|---|---|---|
| Smoker ( smoker_yes = 1 ) | 274 | 20.5% |
| No smoking ( smoker_no ) | 1064 | 79.5% |

Insight: the proportion of smokers is low (20%), but sufficient to constitute the analysis group.

**(3) Region distribution**

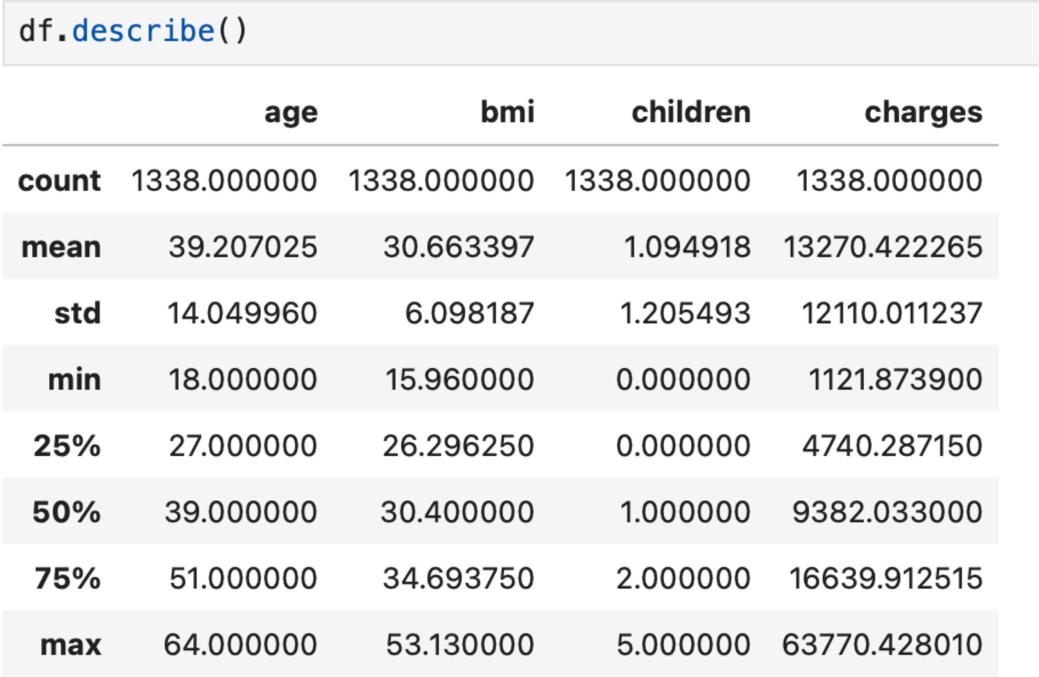
| Region | dummy | sample value | Proportion |
|---|---|---|---|

|  |  |  |  |
|---|---|---|---|
|  | variables |  |  |
| Northeast（baseline） | baseline | 324 | about 24% |
| Southeast | region_southeast | 364 | 27% |
| Southwest | region_southwest | 325 | 24% |
| Northwest | region_northwest | 325 | 24% |

Insights: the sample sizes of the four major regions are very balanced, accounting for approximately 24%–27%.

## 3.2 Statistical Characteristics of Numerical Variables

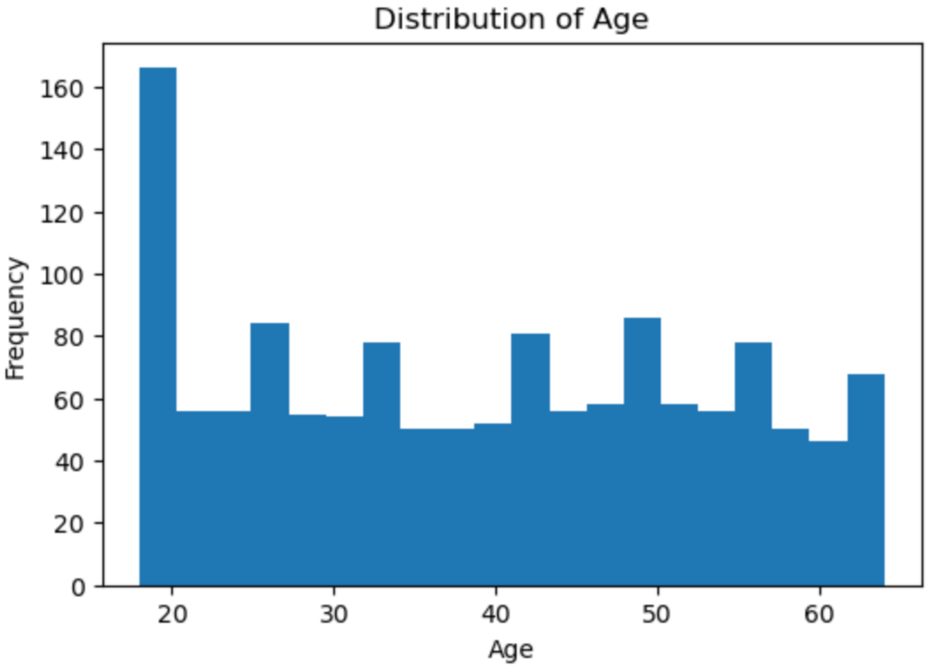
This part systematically describes numerical variables based on df.describe() and the visualization of age distribution, BMI distribution, children distribution and cost distribution.



```
df.describe()
```

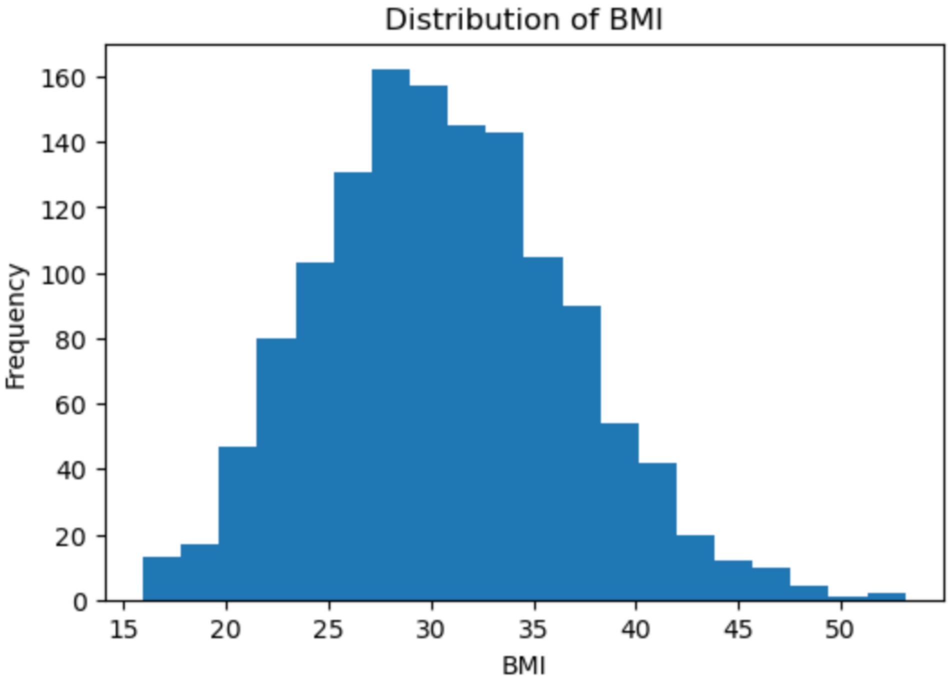
```
               age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
 mean    39.207025    30.663397     1.094918  13270.422265
  std    14.049960     6.098187     1.205493  12110.011237
  min    18.000000    15.960000     0.000000   1121.873900
  25%    27.000000    26.296250     0.000000   4740.287150
  50%    39.000000    30.400000     1.000000   9382.033000
  75%    51.000000    34.693750     2.000000  16639.912515
  max    64.000000    53.130000     5.000000  63770.428010
```

### 3.2.1 Age



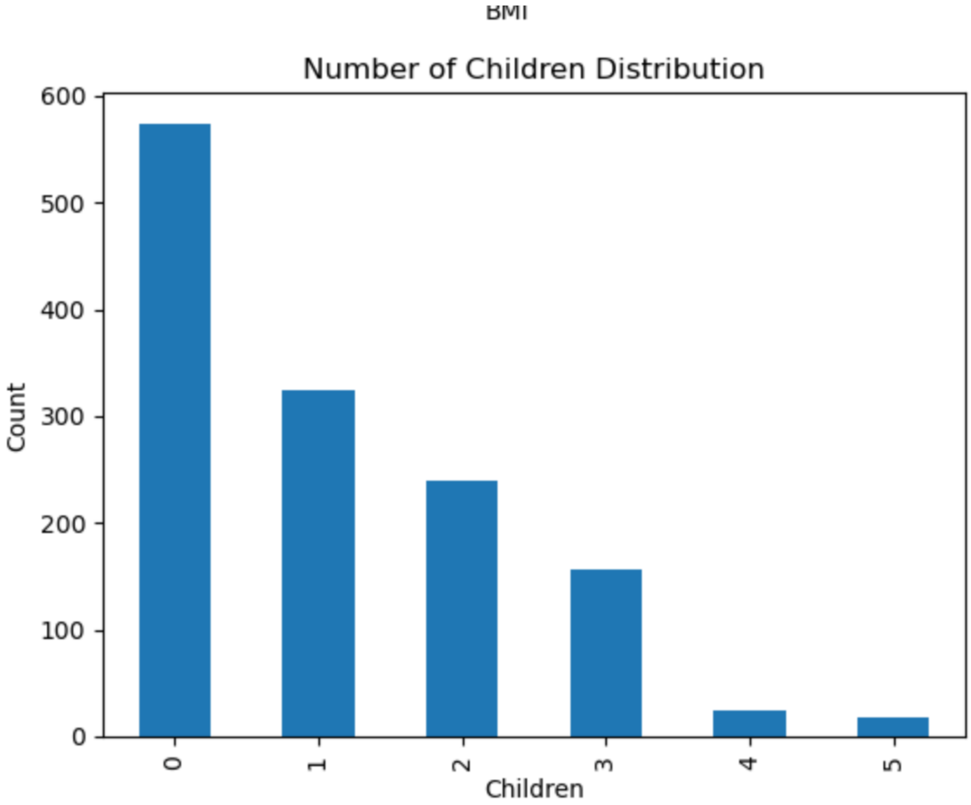
We can see that the distribution was roughly evenly distributed between 18–64, but 18-year-olds were significantly overrepresented. Therefore, it is suitable for direct use in regression analysis. The larger number of 18-year-olds may be related to the large number of young policyholders in the sample source.

### 3.2.2 BMI Distribution



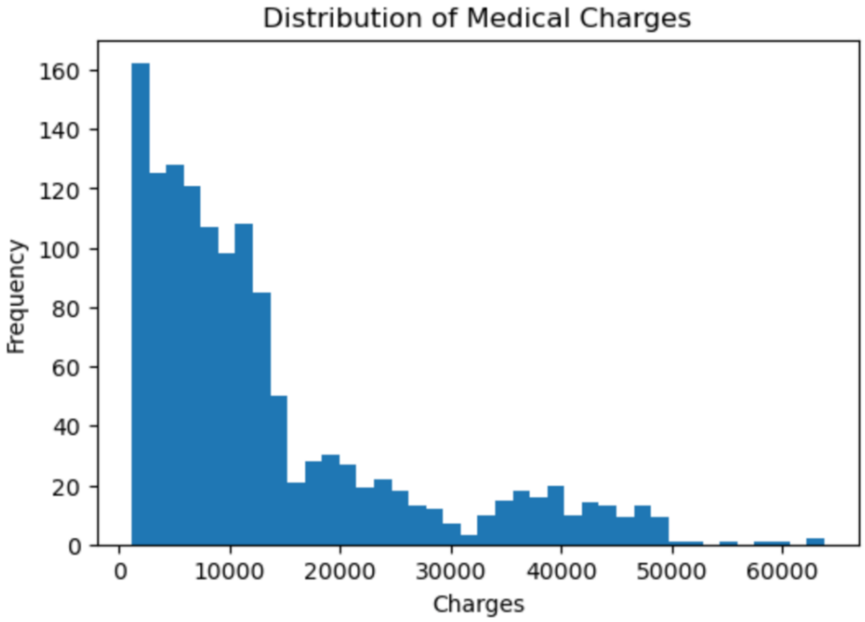
According to the BMI histogram, the distribution is close to the normal distribution with slightly skewed to the right. In addition, most BMIs fall between 25–35.

### 3.2.3 Children Distribution



The diagram reflects that 0 is the maximum number of children (550+), next is 1 and 2 children. The number of children shows a discrete decreasing distribution, with 0–2 children being the mainstream family structure. This variable can be directly used in regression as a continuous integer variable.

### 3.2.4 Charges Distribution



The histogram clearly presents a strong right-skew distribution. Main body concentrated at 2,000–15,000 and a few people reach high fees of 40,000–60,000A few extremely high-cost cases may be related to smoking, obesity, or chronic disease.

## 3.3 Summary of Descriptive Analysis

Based on the above analysis, this study draws the following conclusions:

- The distribution of categorical variables is balanced, which can support model stable estimation.
- The distribution of age and BMI is reasonable and there are no extreme values, which means it is suitable for entering regression model directly.
- Children is a typical discrete variable, and its distribution conforms to the population structure.
- Charges is strongly right-skewed and has the typical long-tail characteristics of medical expenses, which means regression results need to be interpreted in conjunction with the skewed background.

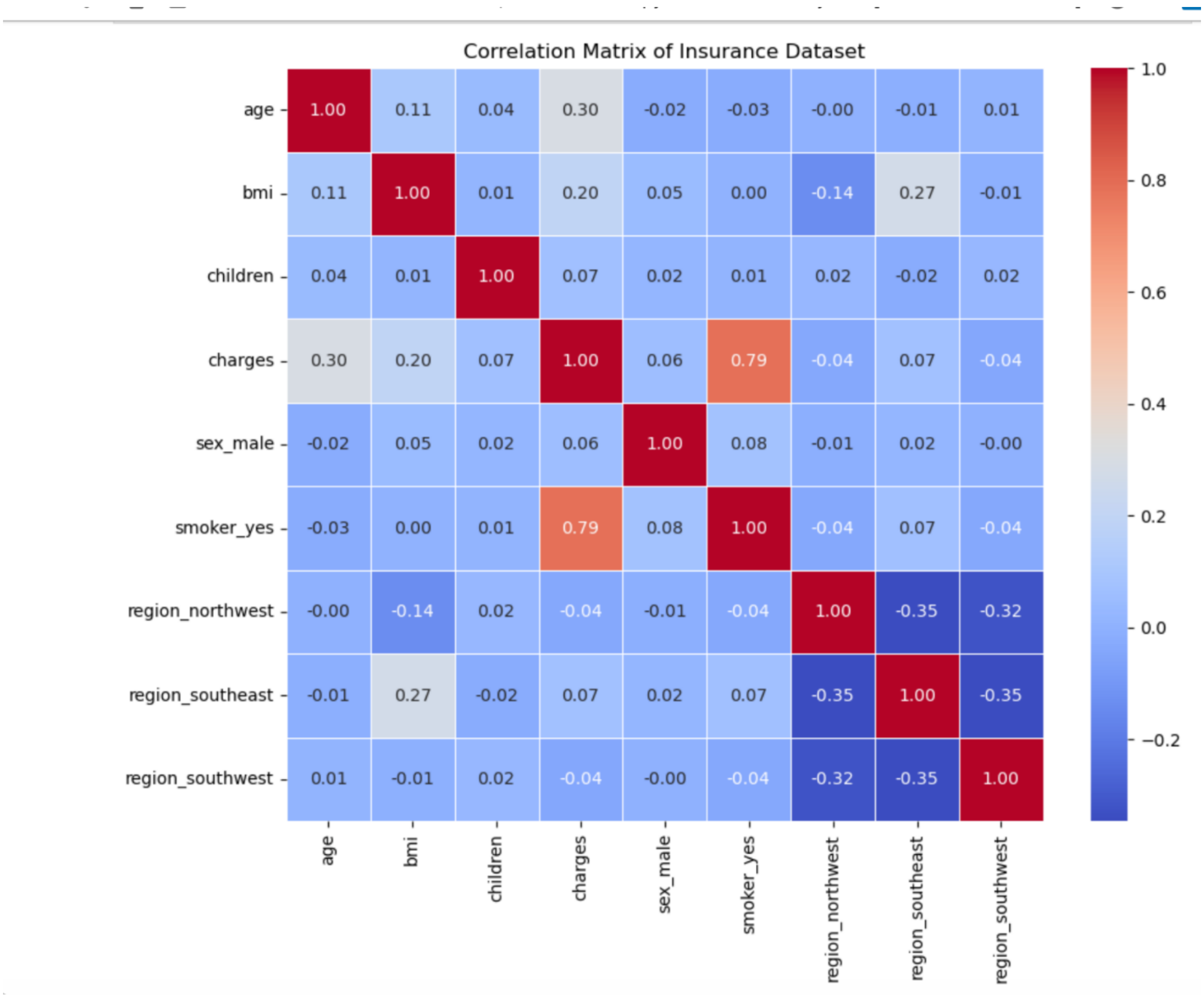
# 4.0 Diagnostic Analysis

This part aims to evaluate whether the impact of smoking on medical expenses is independent and robust through correlation testing and stepwise regression model construction. The analysis proceeds from variable association, identification of potential confounding factors, to establishment of multiple regression models, to gradually verify whether the smoking effect is weakened or eliminated due to other factors. At the same time, through the both evidence of theory and data, we can more scientifically determine whether the relationship between smoking) and medical expenses is robust.

## 4.1 Related Theories

In health economics, public health, and behavioral risk models, medical expenditures are usually systematically affected by factors such as age, health status, lifestyle, family structure, and medical prices. The theoretical expectations are based on authoritative institutions, large databases and journal research. For example, the research from Centers for Disease Control and Prevention and OECD Health Expenditure Database, medical expenses will increase significantly with age, especially after the age of 50, when the risk of chronic diseases increases more quickly. Furthermore, Finkelstein et al. (2009) illustrates that obesity increases annual medical spending by approximately 42%, confirming BMI as a major predictor of healthcare utilization. Also, CMS and HCCI consistently show significant geographic price differences between regions of the United States, even though these differences are not related to health behaviors such as smoking.

## 4.2 Correlation Analysis



According to the correlation heat map, the relationship between the core variables is as follows.

| Relationship | Correlation coefficient | Explanation |
|---|---|---|
| smoker_yes ↔ charges | 0.79(Very strong positive correlation) | Smokers have significantly higher medical costs, the strongest relationship |
| age ↔ charges | 0.30 | The older you are, the higher the medical risks and costs |
| bmi ↔ charges | 0.20 | The higher the BMI (obesity), the higher the cost |
| children ↔ charges | 0.07 | The impact of the number of children in the family is weak |
| sex_male ↔ charges | 0.06 | Gender differences are almost negligible |
| region_dummy ↔ charges | -0.04 ~ -0.35 | There are regional differences, but the effect is weak |

It can be seen that only the relationship between smoking and cost is extremely strongly correlated, while other factors are weakly to moderately correlated.

## 4.2 Confounders

Confounding factors need to meet two conditions at the same time, one is related to the independent variable (smoking), another is related to the dependent variable (medical expenses) Only meeting these two conditions, It can theoretically possible be confounders.

Thus, according to theory and data, we identify the following confounding variables:

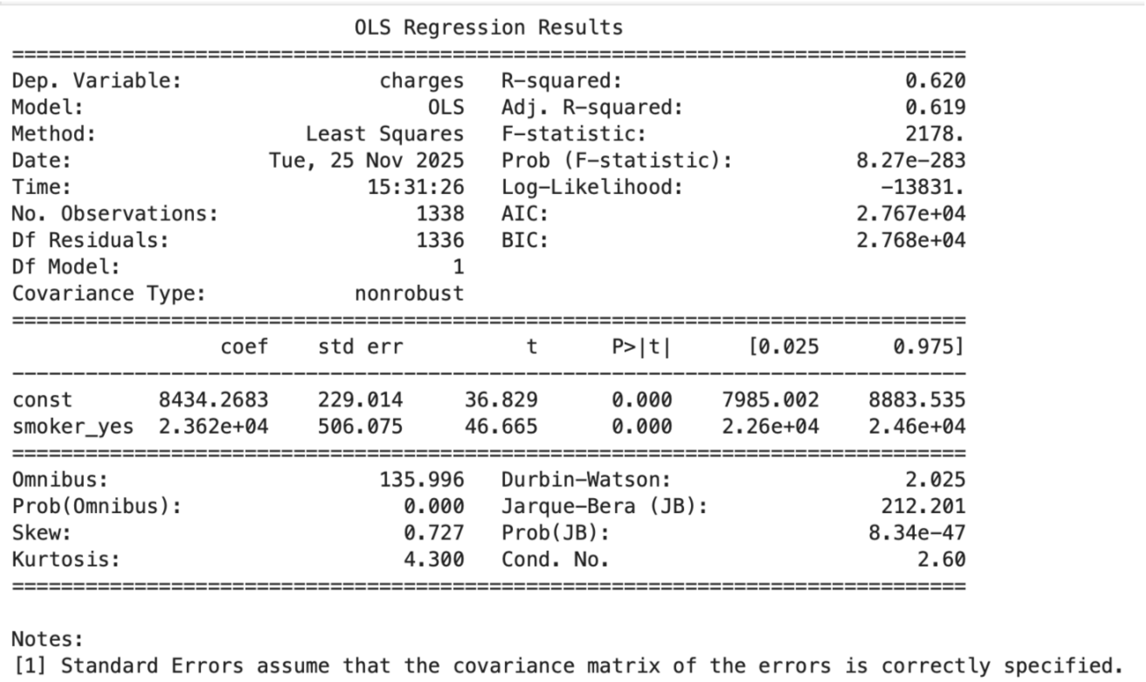
| Indepence(smoker) | Potential confounder factors | Theories |
|---|---|---|
| Smoker | age | The older you are, the more likely you are to smoke and get sick. |
| Smoker | bmi | Unhealthy lifestyle leads to smoking + obesity at the same time |
| Smoker | sex | Smoking rates among men are significantly higher than among women |
| Smoker | children | Family structure affects medical needs and consumer behavior |
| Smoker | region | Smoking rates and medical pricing vary by region |

Insights: If not controlled, it will cause omitted variable bias (OVB). Therefore, subsequent regressions must gradually control these variables to test whether the smoking effect is robust.

## 4.3 Regression Analysis

To test whether the influence of smoking on medical expenses is independent, this part constructed four regression models, starting from the simplest univariate model and adding all potential confounding variables step by step.

### 4.3.1 Model 1: Simple Model, Only Smoker (Charges ~ Smoker_Yes)



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.620
Model:                            OLS   Adj. R-squared:                  0.619
Method:                 Least Squares   F-statistic:                     2178.
Date:                Tue, 25 Nov 2025   Prob (F-statistic):          8.27e-283
Time:                        15:31:26   Log-Likelihood:                -13831.
No. Observations:                1338   AIC:                         2.767e+04
Df Residuals:                    1336   BIC:                         2.768e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       8434.2683    229.014     36.829      0.000    7985.002    8883.535
smoker_yes  2.362e+04    506.075     46.665      0.000     2.26e+04    2.46e+04
==============================================================================
Omnibus:                      135.996   Durbin-Watson:                   2.025
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              212.201
Skew:                           0.727   Prob(JB):                     8.34e-47
Kurtosis:                       4.300   Cond. No.                         2.60
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

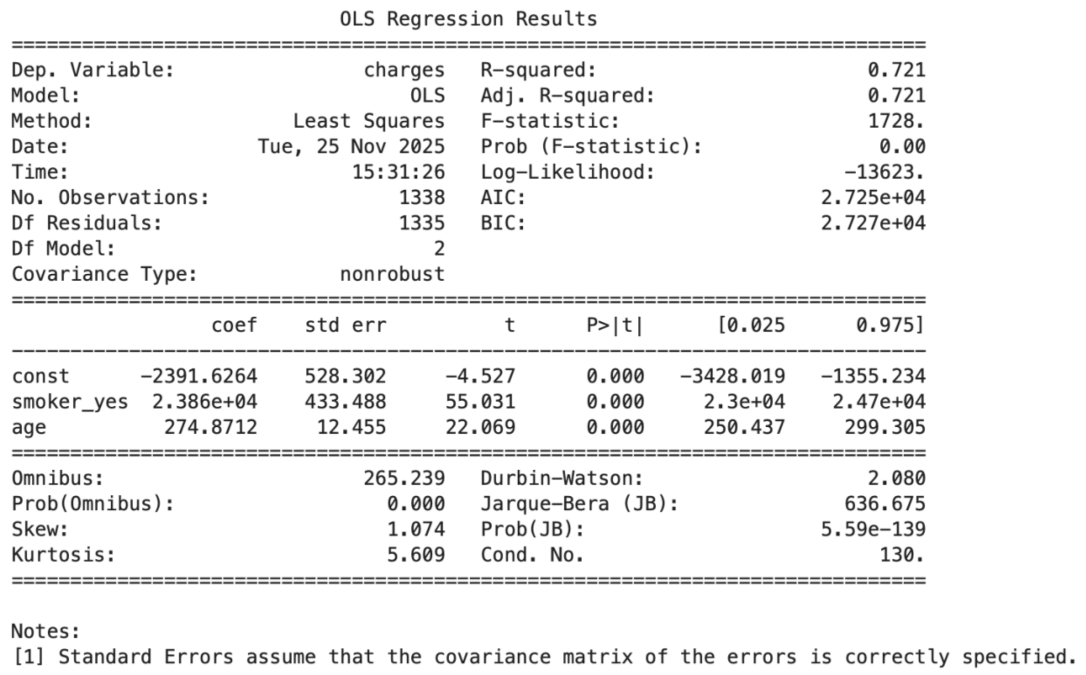
Result:

| Variables | coefficienct | p-value | explaination |
|---|---|---|---|
| smoker_yes | +23,620 | <0.001 | Extremely significant;<br>smokers cost approximately $23,600 more |
| constant | 8434 | — | Base rate for non-smokers |

R² = 0.620

we can see only smoking explains 62% of the variance in medical expenses, which means it is an extremely strong explanatory power. That also illustrates there is a significant, stable relationship between smoking and medical expenditures. However, confounding factors such as age and BMI are not controlled, and there may be bias (OVB). So we gradually add factor into models following.

## 4.2.2 Model 2 Adding Age ( Charges ~ Smoker + Age )



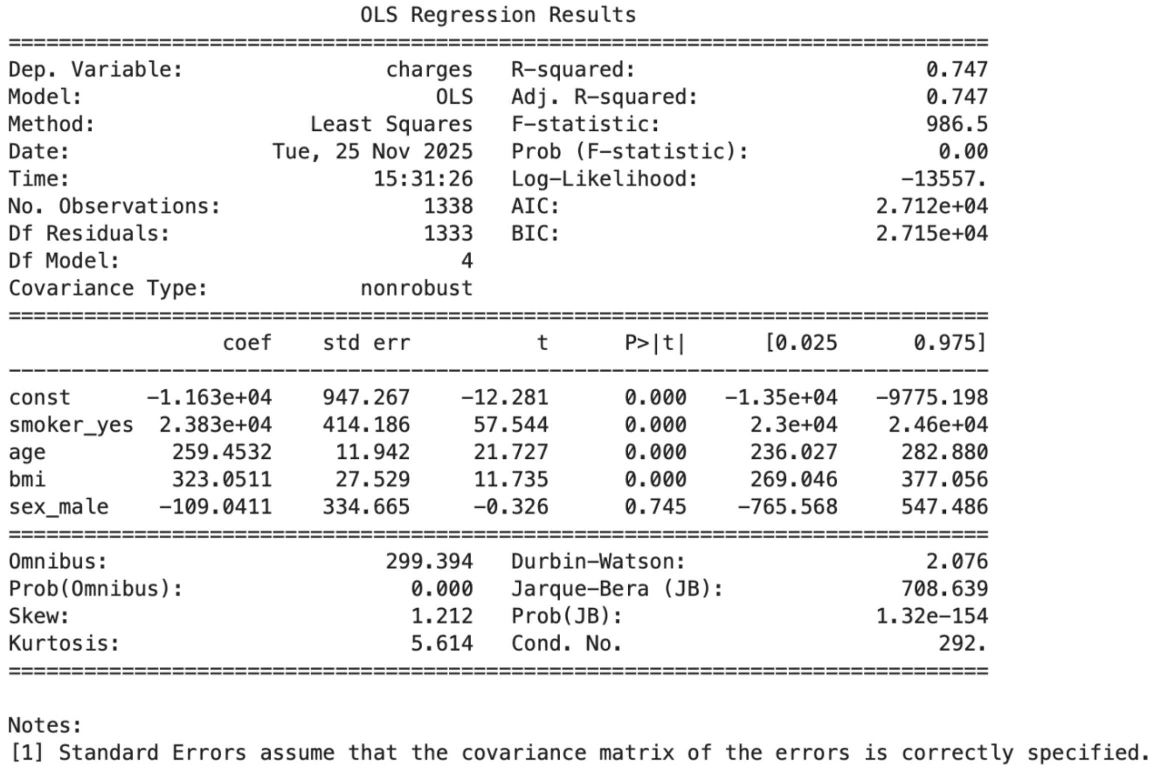
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.721
Model:                            OLS   Adj. R-squared:                  0.721
Method:                 Least Squares   F-statistic:                     1728.
Date:                Tue, 25 Nov 2025   Prob (F-statistic):               0.00
Time:                        15:31:26   Log-Likelihood:                -13623.
No. Observations:                1338   AIC:                         2.725e+04
Df Residuals:                    1335   BIC:                         2.727e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -2391.6264    528.302     -4.527      0.000   -3428.019   -1355.234
smoker_yes  2.386e+04    433.488     55.031      0.000      2.3e+04    2.47e+04
age          274.8712     12.455     22.069      0.000     250.437     299.305
==============================================================================
Omnibus:                      265.239   Durbin-Watson:                   2.080
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              636.675
Skew:                           1.074   Prob(JB):                    5.59e-139
Kurtosis:                       5.609   Cond. No.                         130.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

| Variables | coefficient | p-value | explaination |
|---|---|---|---|
| smoker_yes | +23,860 | <0.001 | Still extremely significant, intensity almost unchanged |
| age | +275 | <0.001 | $275 for each additional year of age |

$R^2 = 0.721$

We can see that age significantly increases medical costs, aligning with theory mentioned. For smoking, the coefficient increase from 23,620 to 23,860(almost unchanged). Whatsmore, age does not explain the smoking effect, smoking itself is an independent influencing factor.

### 4.2.3 Model 3 Adding BMI And Sex (Charges ~ Smoker + Age + Bmi + Sex)



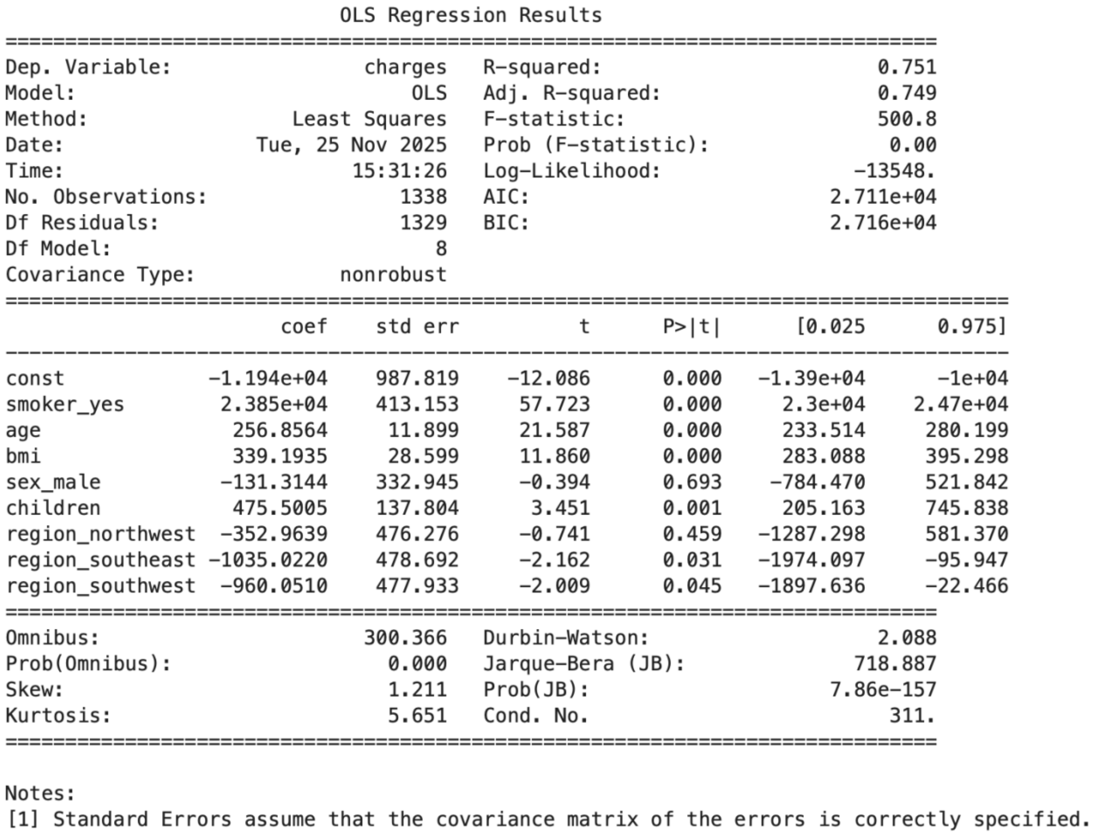
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.747
Model:                            OLS   Adj. R-squared:                  0.747
Method:                 Least Squares   F-statistic:                     986.5
Date:                Tue, 25 Nov 2025   Prob (F-statistic):               0.00
Time:                        15:31:26   Log-Likelihood:                -13557.
No. Observations:                1338   AIC:                         2.712e+04
Df Residuals:                    1333   BIC:                         2.715e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.163e+04    947.267    -12.281      0.000   -1.35e+04   -9775.198
smoker_yes  2.383e+04    414.186     57.544      0.000     2.3e+04    2.46e+04
age          259.4532     11.942     21.727      0.000     236.027     282.880
bmi          323.0511     27.529     11.735      0.000     269.046     377.056
sex_male    -109.0411    334.665     -0.326      0.745    -765.568     547.486
==============================================================================
Omnibus:                      299.394   Durbin-Watson:                   2.076
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              708.639
Skew:                           1.212   Prob(JB):                    1.32e-154
Kurtosis:                       5.614   Cond. No.                         292.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

| variables | coefficients | p-value | explaination |
|---|---|---|---|
| smoker_yes | +23,830 | <0.001 | Almost the same as before |
| age | +259 | <0.001 | stable |
| BMI | +323 | <0.001 | The more the obesity, the higher the cost |
| sex_male | -109 | 0.745 | not significant |

$R^2 = 0.747$

We can see that BMI can explain more powerful.($R^2$ has increased to 0.747). But sex is not significant, which is consistent with theory. The smoking coefficient remains strong and stable (23,830), which demonstrates that smoking is not an artifact of BMI or gender.

### 4.2.4 Model 4 Adding Children + Region



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.751
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     500.8
Date:                Tue, 25 Nov 2025   Prob (F-statistic):               0.00
Time:                        15:31:26   Log-Likelihood:                -13548.
No. Observations:                1338   AIC:                         2.711e+04
Df Residuals:                    1329   BIC:                         2.716e+04
Df Model:                           8
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -1.194e+04    987.819    -12.086      0.000   -1.39e+04      -1e+04
smoker_yes        2.385e+04    413.153     57.723      0.000     2.3e+04    2.47e+04
age                256.8564     11.899     21.587      0.000     233.514     280.199
bmi                339.1935     28.599     11.860      0.000     283.088     395.298
sex_male          -131.3144    332.945     -0.394      0.693    -784.470     521.842
children           475.5005    137.804      3.451      0.001     205.163     745.838
region_northwest  -352.9639    476.276     -0.741      0.459   -1287.298     581.370
region_southeast -1035.0220    478.692     -2.162      0.031   -1974.097     -95.947
region_southwest  -960.0510    477.933     -2.009      0.045   -1897.636     -22.466
==============================================================================
Omnibus:                      300.366   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              718.887
Skew:                           1.211   Prob(JB):                    7.86e-157
Kurtosis:                       5.651   Cond. No.                         311.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

| variables | coeffecient | p-value | explaination |
|---|---|---|---|
| smoker_yes | +23,850 | <0.001 | still the strongest predictor |
| age | +257 | <0.001 | stable |
| BMI | +339 | <0.001 | Stable and enhanced |
| children | +476 | 0.001 | The more children you have, the higher the cost. |
| sex_male | -131 | 0.693 | not significant |
| region_southeast | -1035 | 0.031 | significantly lower |
| region_southwest | -960 | 0.045 | significantly lower |
| region_northwest | -353 | 0.459 | — |

$R^2 = 0.751$

We can see that the spending of is an average of $23,620 more in medical expenses than nonsmokers. Age and BMI also significantly affect costs, but their effects are much smaller than smoking.

## 4.4 Summary of Models

### 4.4.1 Smoking Is the Strongest Independent Predictor of Medical Costs

In these four models, smoker_yes is always significant ($p < 0.001$). And the coefficient is stable at about +$23,600 ~ 23,900. After controlling for all confounding factors, there is subtle effect on smoker. Therefore, smoking has a strong independent effect on health care costs.

### 4.4.2 Important Secondary Risk Factors: Age and BMI

Cost will increase $257–275 per year, and will rise $323–339 per unit increase in BMI. These are all consistent with medical theory.

### 4.4.3 Gender Has a Very Weak Effect

The model shows the sex_male’s $p = 0.69$, not significant.

### 4.4.4 Regional Differences Exist but Are Small

The cost of Southeast and Southwest are slightly less expensive than Northeast. This can explain there are existing differences between regions in the United States.

### 4.4.5 Model Performance Is Excellent

$R^2$ iincreases from 0.62 to 0.75, which means model 4 explains 75% of the cost variance. In real life, it is very high explanatory power.

## 4.5 Model Dignostic

### 4.5.1 Residual Normality

The p-values of the Omnibus and Jarque–Bera tests are both small, indicating that the residuals deviate from the normal distribution in a statistical sense. meanwhile, the skewness and kurtosis of the residuals also show slight right skew and thick tails. This is consistent with the right-skewed nature of medical expenses (charges) and extremely high-cost cases. Therefore, it is caused by the data structure itself and does not affect the effectiveness of the large-sample OLS coefficient.

### 4.5.2 Residual Autocorrelation

The Durbin–Watson statistic is approximately 2.0, indicating that there is no significant positive or negative autocorrelation in the residuals. Since this study uses cross-sectional data, the residual independence is as expected.

### 4.5.3 Multicollinearity

The model's Condition Number ranges from 130–310, indicating a certain degree of collinearity among the independent variables. However, the standard errors of the regression coefficients have not been abnormally inflated, and the significance and coefficients of key variables (smoker, age, bmi) have remained stable.

### 4.5.4 Summary

Although there is small issue, the model is effective to explain the factors of medical expenses, and the regression conclusions of this study are reliable.

# 5.0 Conclusion and Recommendation

## 5.1 Conclusions

**(1) Smoking is the strongest and most independent factor.**

Through four models (smoking only to adding age to adding BMI and sex to adding region and household variables), the smoking coefficient is always stable between +$23,600 ~ +$23,900 and remains significant at $p < 0.001$. Thus we can conclude that smoking and medical costs are not only related but also have independent causal effects and smoking itself is the core reason rather than false amplification by factors such as age, BMI, and region.

**(2) Age and BMI are important second-tier factors**

**(3) Gender has a weak effect**

(4) **There is a moderate impact on the region, but it is not the main cause.**

**(5) The final model explained 75% of the variance in medical costs, which is a high level among medical cost studies.**

Model diagnostics show there is no autocorrelation (DW≈2.0). Then, multicollinearity is within the acceptable range (Cond.No.≈300). And the right skew of the residuals is consistent with the characteristics of cost data and does not affect the robustness of large-sample OLS. To summarize, the overall model is robust and reliable, and the conclusions have high credibility.

## 5.2 Recommendation

### 5.2.1 Insurance Companies Should Clearly Distinguish Pricing Strategies for Smokers and Non-Smokers

Insurance companies can setting up a smoker risk premium and offer health plan discounts.

### 5.2.2 Focus on High-BMI Groups and Establish an Obesity-Related Risk Management Mechanism

Regression shows that BMI is a robust and significant risk factor, suggesting insurance companies to improve health management programs for clients with higher BMI, provide nutrition consultation and set up health improvement incentives.

### 5.2.3 The Elderly Group Should Be Given Extra Attention and Expected Cost Management

Because the aging suffer from higher risks of chronic diseases, welfare should include chronic disease management projects for middle-aged and elderly customers and increase coverage of physical examinations and early screening.

### 5.3.4 Further Research is Recommended, Such As Smoker×BMI and Smoker × Age

## 6.1 Limitation

### 6.1.1 The Data Are Cross-Sectional and Temporal Causality Cannot Be Identified

This study is only based on a time section and cannot observe changes in medical expenses over time. Therefore, researchers need to determine cumulative effects of long-term smoking.

### 6.1.2 The Data Comes from Kaggle and Has Limited Representativeness

The data are still not necessarily cover all socioeconomic backgrounds and not necessarily representative of the overall U.S. population structure. Therefore, the external validity of the conclusions needs to be interpreted with caution.

### 6.1.3 Key Socioeconomic Variables Are Not Included

Medical costs may be affected by SES (socioeconomic status) such as: income, education and occupation. The absence of these variables may limit the explanatory power of some models.

### 6.1.4 Variables Such as BMI and Smoking Are Self-Reported and May Contain Measurement Errors

# References

Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual medical spending attributable to obesity. Health Affairs.

Health system tracker.(2023). *How have costs associated with obesity changed over time?* Retieved from: https://www.healthsystemtracker.org/chart-collection/how-have-costs-associated-with-obesity-changed-over-time/

Kaggle. (2018). Medical Cost Personal Dataset. Retrieved from: https://www.kaggle.com/datasets/mirichoi0218/insurance

U.S. Department of Health and Human Services. (2014). The health consequences of smoking—Surgeon General's report. https://www.hhs.gov/surgeongeneral/reports-and-publications/tobacco/consequences-smoking-factsheet/index.html

# Appendix

## Appendix 1: Data description

l Sample size: 1338 records, no missing values, and good data quality

l Variables:

| Viarables | Meaning |
|---|---|
| age | Age of the insured (primary beneficiary) |
| sex | Gender of the policy holder (female, male ） |
| bmi | Body Mass Index, which measures the relationship between weight and height (kg/m$^2$) |
| children | The number of children (or dependent children) covered by the insurance in the insured's family |
| smoker | Whether to smoke（yes / no） |
| region | Region of residence（U.S.： northeast, southeast, southwest, northwest） |
| charges | The amount a health insurance company charges an individual for medical treatment (insurance cost) |

Dataset source**:** Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

Data: Desktop/isom/insurance.csv

## Appendix B: Full python code