**Journal of Computing and Electronic Information Management**
ISSN: 2413-1660 | Vol. 15, No. 2, 2024

# Predictive Modeling of Personal Medical Insurance Costs: Analyzing Key Factors and Interactions

**Xuerui Wang [1], Siwei Tuo [2]**

[1] School of AI and Advanced Computing, Xi'an Jiaotong-Liverpool University (Taicang), Suzhou, 215400, China
[2] School of Industrial and Financial Cooperation, Xi'an Jiaotong-Liverpool University (Taicang), Suzhou, 215400, China

**Abstract:** This study aims to use Kaggle's dataset to predict personal medical insurance costs, including age, sex, BMI, number of children, smoking status and regional variables. Through comprehensive statistical analysis and model improvement, we have determined important forecasting factors and interactions that affect charges. It is mainly found to include non-linear relationships between age and medical expenses, as well as the significant impact of interaction between BMI and smoking. The enhanced regression model combines these interactions and non-linear effects, showing great improvements in terms of interpretation capabilities, the R^2 value is 0.9642. The results provide actionable insights for insurance policy formulation, health management programs, and risk assessment, demonstrating the importance of considering complex variable interactions in predicting medical expenses. Future research should continue exploring these relationships to further refine predictive models.

**Keywords:** Medical Insurance Costs, Regression Analysis, BMI, Smoking Status, Age Nonlinearity, Interaction Effects, Health Economics.

## 1. Dataset description

The dataset used in this analysis comes from *Kaggle*[1], which contains personal medical insurance costs and related factors. The main purpose of this dataset is to predict personal medical insurance costs through various factors.

The specific content of the dataset is as follows:

- **Age**: The age of the main beneficiary
- **Sex**: Insurance contractor gender, female, male

**BMI**: Body Mass Index (BMI), which is calculated by weight and height, is used to measure whether a person's weight is suitable for his height. The calculation formula of BMI is as follows:

$$bmi = \frac{Weight\ (kg)}{Height(m)^2} \quad (1)$$

- **Children**: Number of children covered by health insurance/Number of dependents
- **Smoker**: Smoking status (yes, no)
- **Region**: The beneficiaries are in the residential area of the United States (northeast, southeast, southwest, northwest)
- **Charges**: Medical expenses collected by personal health insurance

The purpose of the dataset is to predict personal medical insurance costs ('charges') through the above variables in order to better understand and analyze which factors have the greatest impact on medical expenses, thereby providing a reference for medical insurance policy formulation and personal health management.

### 1.1. Feature Visualization and Analysis

According to Doe's research [2], there is no significant relationship between regions and medical insurance costs. Therefore, in order to simplify the model and improve its performance, we decided to delete the region variable.

#### 1.1.1. Charge Analysis

In order to understand the distribution of charges and its characteristics in the dataset, we use the column graph to display the distribution of charges (Figure 1). The charges contained in this dataset start from a minimum of $1,121.87 to a maximum of $63,770.43.

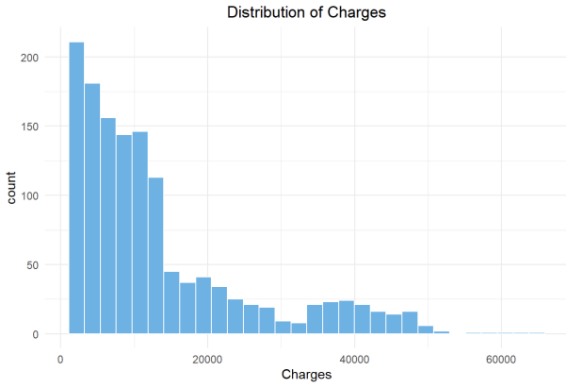


**Figure 1.** Distribution of Charges

#### 1.1.2. Age Analysis

We first used a column graph to display the age distribution situation (Figure 2).

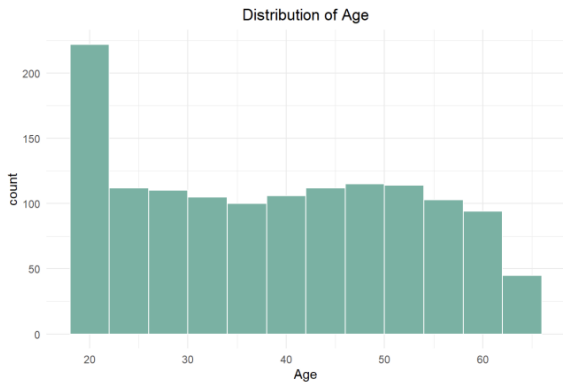


**Figure 2.** Distribution of Age

It can be seen from Figure 2 that people in their 20s are relatively concentrated, and the distribution of other ages is relatively uniform. This shows that the proportion of young adults in the dataset is slightly higher.

In order to further analyze the impact of age on charges, we have drawn a diagram of the relationship between age and expenses, and added a fitting line to display the trend (Figure 3).

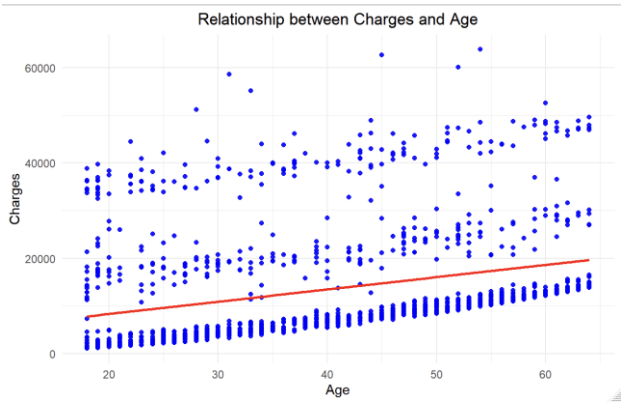


**Figure 3.** The relationship between Age and Charges

When age increases, charges will increase. This trend is expected, because with age, people usually generate higher medical expenses due to age-related health problems[3].

The red fitting line further confirms the positive correlation between age and charges, indicating that age is an important predictable variable for medical insurance costs.

**Note:**

When analyzing the relationship between other variables and charges, we will use a scattered plot diagram to represent this. This is because age is an important factor, and it usually has a significant impact on medical insurance costs.

Displaying the effects of other variables in the same scattered point diagram can help us more intuitively understand the regulatory role and interaction between the variables in the relationship between age and cost.

### 1.1.3. Analysis of the Number of Children

We first used a histogram to show the distribution of the number of children in the dataset (Figure 4).

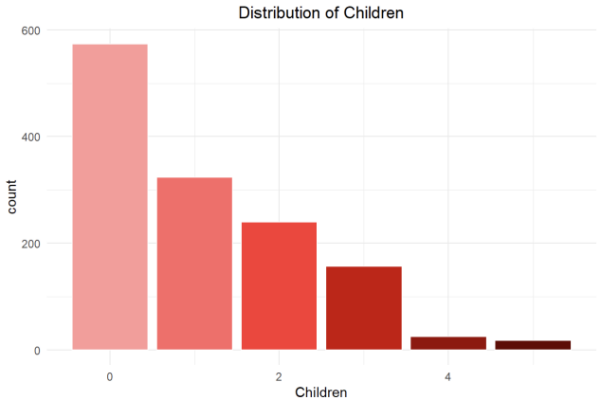


**Figure 4.** Distribution of Number of Children

Most of the respondents have no children or only one child, and relatively few respondents have 2, 3, 4, and 5 children. This indicates that most families in the dataset have a smaller number of children.

To further analyze the impact of the number of children on charges, we plot the relationship between age and charges and color them according to the number of children (Figure 5).

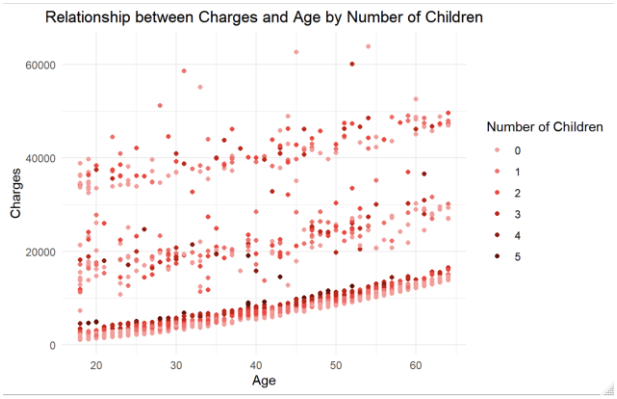


**Figure 5.** The Relationship between Charges and Age by Number of Children

There is little difference in the distribution of charges between families with different numbers of children.Each category of the number of children is relatively scattered in the plot of relationship between charges and age. This suggests that the number of children does not have a significant impact on health insurance costs.

### 1.1.4. Sex Analysis

We first showed the distribution of the dataset gender with the column diagram (Figure 6). This dataset contains 1338 records, and gender is divided into men and women.

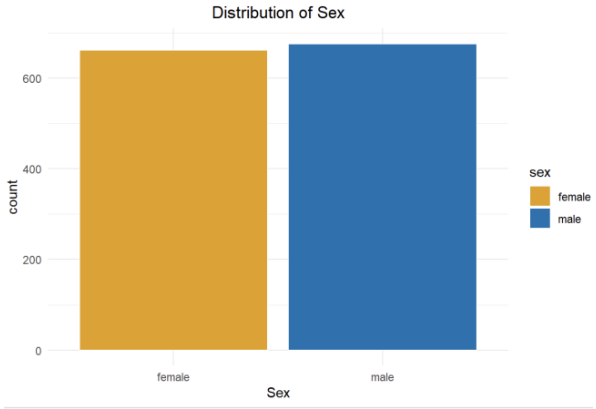


**Figure 6.** Sex Distribution

It can be seen from Figure 10 that the number of males and females is relatively balanced. This signifies that the proportion of data concentration for two genders is close.

In order to further analyze the impact of gender on medical insurance costs, we have detailed the relationship between the age of different genders and charges (Figure 7).

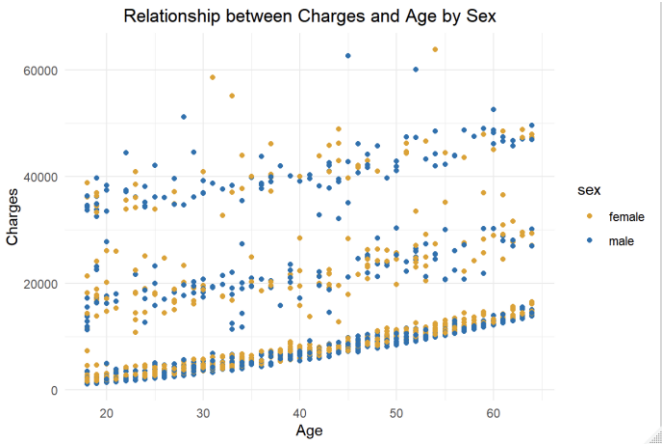


**Figure 7.** Relationships between Charges and Age by Sex

The distribution of males and females in medical expenses is not significant, and data points are scattered in various gender categories.

### 1.1.5. Smoker Analysis

We have used a column graph to show the number of people who smoke and people who do not smoke in this dataset (Figure 8).

It can be seen from the dataset that the number of people who do not smoke is significantly higher than the number of smokers.

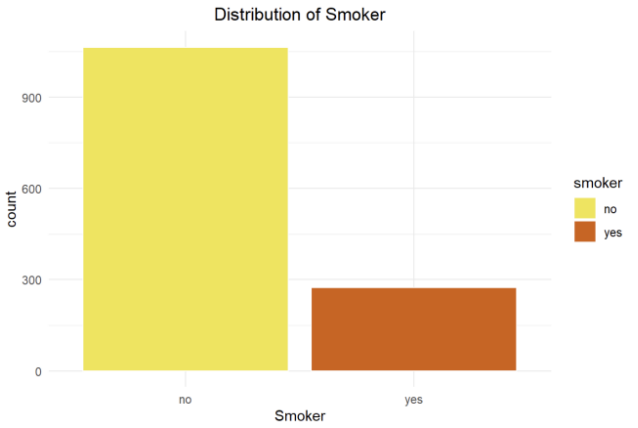


**Figure 8.** Distribution of Smoker

In order to further analyze the impact of smoking on charges, we have drawn the relationship between the age of different smoking conditions and charges (Figure 9).

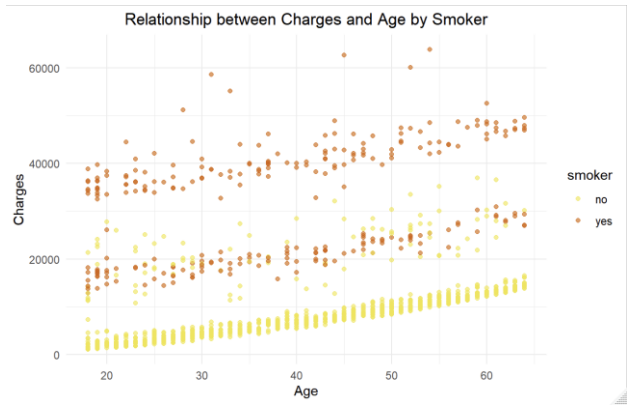


**Figure 9.** Relationships between Charges and Age by Smoker

We can see that charges for smokers is significantly higher than that of non-smokers. This trend is relatively obvious in all ages. It can be understood that smoking has a significant impact on charges.

Therefore, we can draw the conclusion that charges for smokers is generally higher than non-smokers. This discovery is consistent with existing research, indicating that smoking will significantly increase personal medical expenses [5].

### 1.1.6. BMI Analysis

In order to better understand the distribution of BMI in the dataset and its impact on charges, we first drew BMI's pillars (Figure 10).

Most people's BMI is concentrated between 20 and 40, showing a form of normal distribution. However, through this data, it is difficult for us to determine the specific impact of BMI.

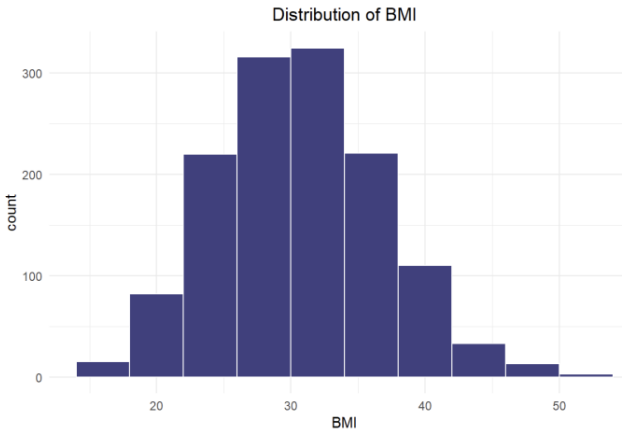


**Figure 10.** BMI Distribution

Therefore, we refer to the WHO standard and divide BMI into four categories:

$$\text{BMI Categories} = \begin{cases} \text{underweight} & \text{if BMI} < 18.5 \\ \text{normal} & \text{if } 18.5 \leq \text{BMI} < 25 \\ \text{overweight} & \text{if } 25 \leq \text{BMI} < 30 \\ \text{obese} & \text{if BMI} \geq 30 \end{cases} \quad (2)$$

According to the above classification standards, we have drawn the classified BMI pillar chart (Figure 11).

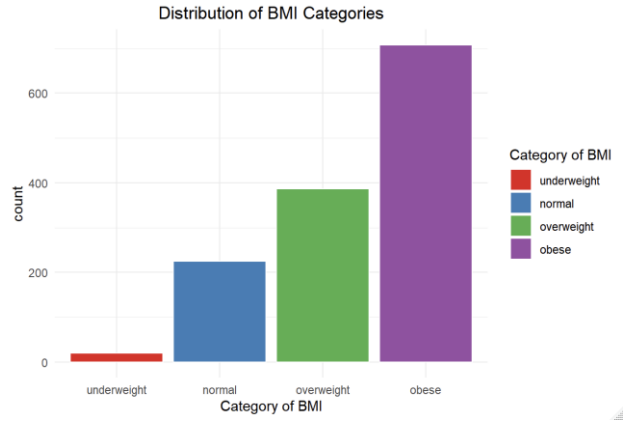


**Figure 11.** Distribution of BMI Categories

It can be seen from Figure 11 that the number of obese is the largest, followed by overweight, normal and lean.

In order to further analyze the impact of the BMI category ('obs') on charges, we have drawn a scattered point figure of the age and charges classified by BMI categories (Figure 12).

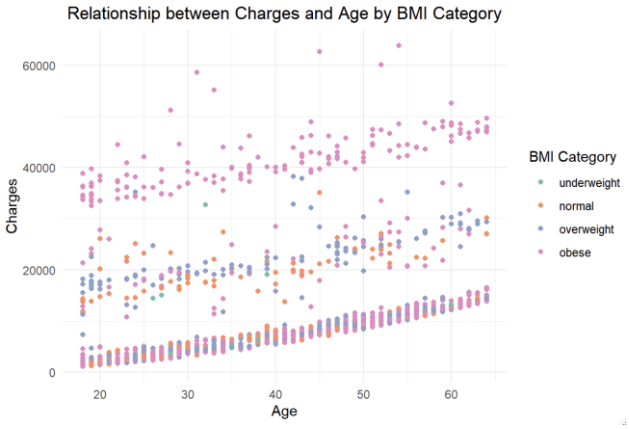


**Figure 12.** Relationship between Charges and Age by obs

It can be seen from Figure 12 that there are significant

differences in charges under different BMI categories.

The charges for obese people is significantly higher than other categories, indicating that people with higher BMIs may face higher charges. This trend is consistent with known health risks, because higher BMIs are usually related to more health problems, such as cardiovascular disease and diabetes [6].

# 2. Linear Regression

## 2.1. Mathematical Formula Reasoning

In a multiple linear regression model, we assume 'charges' is the response variable, and other variables are the predictor variables.

The mathematical formula for a multiple linear regression can be exressed as:

$$charges = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k + \epsilon \quad (3)$$

- *Y (or* 'charges')is the dependent variable
- $\beta_0$ is intercept
- $\beta_i$ ($i$=1, ... ,6) are the regression coefficients for each independent variable
- *X1,X2,...,Xk* are the independent variables
- $\epsilon$ *is the error term, assumed to be normally distributed with a mean of 0*

## 2.2. Feature Selection

To optimize this model, we use AIC (Akaike Information Criterion) to select the best fitting model among the candidate models. The specific code steps are as follows (feature selection of the model is completed by using the 'stepAIC' function).

The lower AIC value represents a better model because it shows that the model has less parameters, so it balances the accuracy and complexity of the model[7].

Its mathematical formula is

$$AIC = 2k - 2ln(L) \quad (4)$$

- $K$ is the number of parameters in the model
- $L$ is the functional value of the model

| Step | Variable | RSS | AIC |
|---|---|---|---|

**Table 1.** The Result of Gradual Return

| | | | |
|---|---|---|---|
| None | - | 4.8054e+10 | 23292.74 |
| Adding'**sex**' | +sex | 4.8046e+10 | 23295 |
| Removing'**bmi**' | - bmi | 4.8242e+10 | 23296 |
| Removing'**children**' | - children | 4.8492e+10 | 23303 |
| Removing'**obs**' | - obs | 4.9078e+10 | 23315 |
| Removing'**age**' | - age | 6.5312e+10 | 23701 |
| Removing'**smoker**' | - smoker | 1.7150e+11 | 24993 |

It can be seen from the results that the ***age, BMI, children, smoker, obs*** are finally selected as a variable of a model.

These variables show a significant impact on the model *AIC* during the gradual return process. Removing any variable will significantly increase the *AIC value*. Therefore, these variables are considered variables that have significant interpretation capabilities for medical insurance costs.

## 2.3. Interpret the Coefficients

**Table 2.** The Coefficient Estimation Values

| Variable | Estimate | Std.Error | T value | Pr(>\|t\|) |
|---|---|---|---|---|
| **Intercept** | -8047.43 | 1672.16 | -4.813 | 1.66e-06 |
| **age** | 257.79 | 11.80 | 21.855 | < 2e-16 |
| **bmi** | 119.83 | 52.53 | 2.281 | 0.02269 |
| **children** | 475.01 | 136.55 | 3.479 | 0.00052 |
| **smokeryes** | 23825.60 | 407.62 | 58.451 | < 2e-16 |
| **obsnorm** | 446.39 | 1428.06 | 0.313 | 0.75465 |
| **obsoverweight** | 623.88 | 1477.18 | 0.422 | 0.67284 |
| **obsobese** | 3561.83 | 1650.81 | 2.158 | 0.03114 |

**Intercept**

- **Estimation value:** -8047.43
- **Significance:** When all predictor variables are zero, the intercept is the estimated value of the response variable, serving as a baseline in the linear equation.
- **P-value:** < 0.001, indicating that the intercept is significantly different from zero.

**age**

- **Estimation value:** 257.79
- **Significance:** For every 1-year increase in age, the charges are expected to increase by $257.79.
- **P-value:** < 0.001, indicating that age has a significant impact on medical expenses.

**bmi**

- **Estimation value:** 119.83
- **Significance:** For every unit increase in BMI, the expected medical expenses are expected to increase by $119.83.
- **P-value:** < 0.05, indicating that BMI has a significant impact on medical expenses.

**number of children**

- **Estimation value:** 475.01
- **Significance:** For every additional child, the expected medical expenses are expected to increase by $475.01.
- **P-value:** < 0.001, indicating that the number of children has a significant impact on medical expenses.

**smokeryes**

- **Estimation value:** 23825.60
- **Significance:** For smokers (relative to non-smokers), the expected medical expenses will increase by $23825.60.
- **P-value:** < 0.001, indicating that smoking status has a significant impact on medical costs.

**obs**

- **obsnormal:**

- **Estimation value:** 446.39
- **P-value:** > 0.05, indicating that normal weight has no significant effect on medical costs.

- **obsoverweight:**
  - **Estimation value:** 623.88
  - **P-value:** > 0.05, indicating that overweight has no significant impact on medical expenses.
- **obsobese:**
  - **Estimation value:** 3561.83
  - **P-value:** < 0.05, indicating that compared with thinner individuals, the expected medical expenses of obese people have increased significantly.

In order to understand the coefficients in the linear regression model more intuitively, we used a heat-map to display the estimated values, standard errors, t-values, and p-values of each variable (Figure 13).

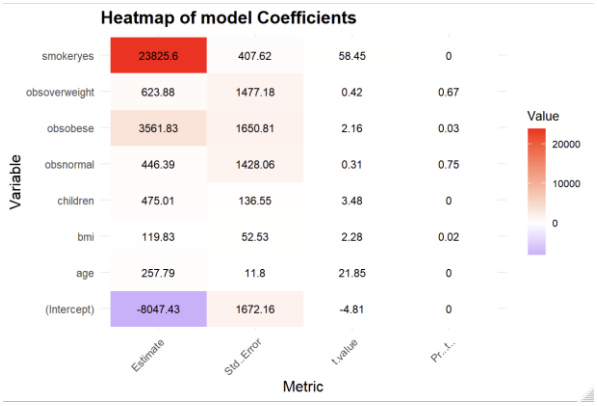


**Figure 13.** Heatmap of Model Coefficients

**Estimated value:**

- Smoker's coefficient is the largest (23825.60), indicating a strong relationship between smokeryes and charges. For every additional number of smokers, the response variable ('charges') is expected to increase by $23825.60.
- Age and BMI also have positive coefficients, indicating that as age and BMI increase, medical expenses('charges') also increase.
- The value of the intercept (Intercept) is negative, indicating that it is a baseline when all independent variables are zero.

**Standard Error:**

- The standard error of the smoker variable is 407.62, which is small relative to its estimate, indicating that the estimate is very precise.
- The standard errors for BMI (52.53), number of children (136.55), and age (11.8) are relatively small, indicating that these estimates are also extremely accurate.

**t-value:**

- The t-value for the smoker variable is very high (58.45), indicating strong statistical significance.
- The t-value for age (21.85), number of children (3.48), and BMI (2.28) are also good, suggesting the significance in the model.

**p-value:**

- The p-values for the 'age', 'BMI', 'children', 'smoker', and 'obsobese' variables are all less than 0.05, indicating that these coefficients are statistically significant at the 5% significance level.

## 2.4. Access the Goodness-of-Fit

### 2.4.1. Calculate $R^2$

When evaluating the goodness-of-fit of a linear regression model, $R^2$ (the coefficient of determination) is a crucial metric. $R^2$ represents the proportion of the total variance in the dependent variable that is explained by the model. Its value ranges between 0 and 1, with values closer to 1 indicating a better fit of the model to the data [8].

The calculation formula of $R^2$ is:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad (5)$$

- SSR is the Regression Sum of Squares, representing the variation explained by the model
- SSTO is the Total Sum of Squares, representing the total variation in the dependent variable.
- SSE is Residual Sum of Squares, which is the sum of squares of the difference between actual and predicted values.

We use the code to calculate the values of SSTO ( 196,074,221,568 ), SSE( 48,054,412,982 ), and SSR ( SSTO-SSE ) respectively, and then get the $R^2$.

$$R^2 = 1 - \frac{48,054,412,982}{196,074,221,568} = 0.7549172 \approx 0.7549 \quad (6)$$

### 2.4.2. Evaluate $R^2$ Results

```
Residual standard error: 6011 on 1330 degrees of freedom
Multiple R-squared:  0.7549,    Adjusted R-squared:  0.7536
F-statistic: 585.2 on 7 and 1330 DF,  p-value: < 2.2e-16
```

**Explanation ability:**

The $R^2$ value is 0.7549, indicating that the model explains approximately 75.49% of the total variation in the reponse variable ('charges'). This indicates that the model can fit data very well, suggesting a strong relationship between predictor variables and the response variable as well.

**Model applicability:**

An $R^2$ value close to 1 indicates that the model fits the data well. Although 0.7549 is not a perfect 1, in social science research, an explanatory power of 75.49% is a pretty good result.

**Adjusted $R^2$:**

Adjusted $R^2$takes into account the number of independent variables in the model, adjusting $R^2$ to prevent overfitting. The formula [16] is:

$$Adjusted\ R^2 = 1 - (\frac{1-R^2}{n-k-1}) \cdot (n-1) \quad (7)$$

By performing summary operations on *stepwise_model*, the adjusted $R^2$ is 0.7536, which is very close to the $R^2$ value, further verifying the robustness of the model.

# 3. Diagonostics

## 3.1. Perform Diagnostic Checks

To verify the validity of the linear regression model, we performed diagnostic checks by using residual analysis, Scale-Location plot analysis, Quantile-quantile plot (QQ plot) analysis, variance inflation factor (VIF) analysis, and Cook's distance plot analysis.

### 3.1.1. Residual Analysis

**Residual plot analysis**

Residual plot is a common way to examine the relationship between residuals and fitted values [9]. By looking at the residual plot, we can see if the residuals are systematically biased, or if there is a pattern.

**Hypothetical test:**

- $H_0$ (Null Hypothesis): The residuals are randomly distributed and have a mean of zero.
- $H_1$ (Alternative Hypothesis): The residuals are not randomly distributed and have a non-zero mean.

We can see from the residual plot (Figure 14) that **the residuals are not random**. Instead, we can see that there is a clear curvilinear pattern in the residual plot, especially in the areas of larger residual values. This indicates that there may be a nonlinear relationship or other potential problem in the model. The red smooth line also shows a non-random pattern, further confirming this.

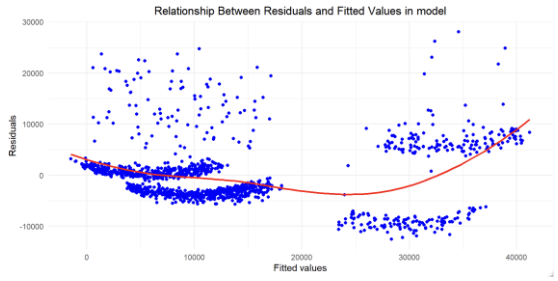


**Figure 14.** Residual Plot in Model

**Scale-Location plot analysis**

Scale-Location plot, also known as Spread-Location plot, is used to check the homoscedasticity (constant variance) of the residuals. It plots the square root of the standardized residuals against the fitted values [10].

Homoscedasticity is an important assumption in linear regression, and a violation of this assumption indicates that the variability of the residuals is not constant across all levels of the independent variable[10].

**Hypothesis testing:**

- $H_0$ (Null Hypothesis): The residuals have constant variance (homoscedasticity).
- $H_1$(Alternative Hypothesis): The residuals do not have constant variance (heteroscedasticity).

It can be seen from the Scale-Location plot (Figure 15) that the spread of the residual's changes with the increase of the fitted values. Especially in the area with larger fitted values, the variability of the residuals increases significantly. This indicates that **the model may have heteroscedasticity issues**. The red smooth line shows a trend, which further indicates that the variability of the residuals is not constant.

In addition, the darker the color, the greater the Cook's distance, indicating that some points may have a greater impact on the model.

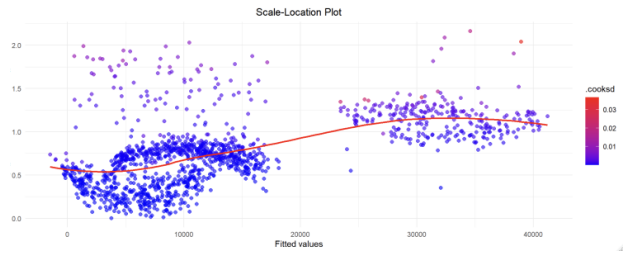


**Figure 15.** Scale-Location Plot in Model

**Quantile-quantile (QQ) plot analysis**

The QQ plot is used to check whether the residuals follow a normal distribution. By comparing the actual residuals with the quantiles of the theoretical normal distribution, we can intuitively determine whether the normality assumption is true [11].

**Hypothetical test:**

- $H_0$ (Null Hypothesis): The residuals follow a normal distribution.
- $H_1$ (Alternative Hypothesis): The residuals are not normally distributed.

It can be seen from the QQ plot (Figure 16) that there is a large deviation between the actual quantile and the theoretical quantile of the residual, especially at both ends. This indicates that the **residuals are not completely normally distributed.**

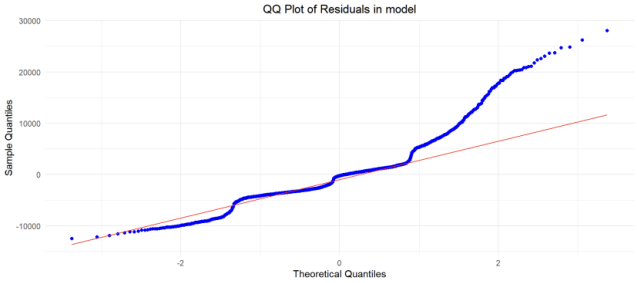


**Figure 16.** QQ plot of residuals in Model

### 3.1.2. Variance Inflation Factor (VIF) Analysis

VIF is used to detect multicollinearity. High VIF values indicate the existence of multicollinearity among independent variables, which will affect the estimation of regression coefficients and the stability of the model[12].

**Hypothetical test:**

- $H_0$ (Null Hypothesis): There is no multicollinearity among the independent variables.
- $H_1$ (Alternative Hypothesis): There is multicollinearity among the independent variables.

**Table 3.** The information of VIF

| Variable | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| **age** | 1.016365 | 1 | 1.008149 |
| **bmi** | 3.796767 | 1 | 1.948529 |
| **children** | 1.002668 | 1 | 1.001333 |
| **smoker** | 1.001995 | 1 | 1.000997 |
| **obs** | 3.802606 | 3 | 1.249339 |

The VIF values are all below 10, indicating that **there is no serious multicollinearity problem** between the independent variables. However, the VIF values of BMI and obs are relatively high, and these variables need to be paid attention to when modeling.

### 3.1.3. Detecting Outliers and Influential Points

**Residuals vs Leverage plot analysis**

Residuals vs Leverage Plot is an important tool for detecting outliers and influential points [13]. In Figure 17, the horizontal axis represents the level value, and the vertical axis represents the standardized residual. By observing these points, we can identify which data points have a significant impact on the model.

**Hypothetical test:**

- $H_0$ (Null Hypothesis): No data points have a significant impact on the model.
- $H_1$ (Alternative Hypothesis): There are some data points that have a significant impact on the model.

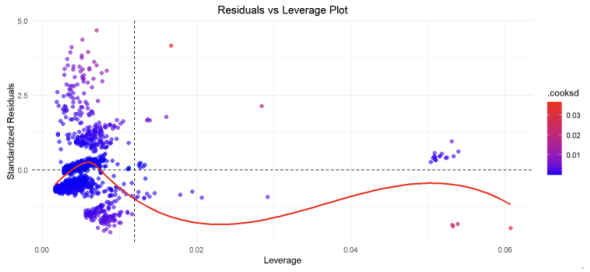


**Figure 17.** Residuals vs Leverage Plot

The dotted line in Figure 17 represents the threshold, and any point with a leverage value greater than this threshold is considered a high leverage point.

The threshold formula [13] is:

$$Leverage\ Threshold = \frac{2(p+1)}{n} \quad (8)$$

- $p$ is the number of parameters in model
- $n$ is the number of samples

For our model (**'stepwise_model'**), with 8 parameters and 1338 data points, the leverage value threshold is:

$$Leverage\ Threshold = \frac{2(p+1)}{n} = \frac{2(8+1)}{1338} \approx 0.0135 \quad (9)$$

As can be seen from Figure 26, the leverage values of several data points are higher than the threshold 0.0135, indicating these data points have a greater impact on the fitting of the regression model.

The red smooth line shows the trend of the standardized residuals. It can be seen that the trend line is not completely horizontal, which indicates that the residuals have a certain pattern near the data points with higher leverage values.

**Cook's distance analysis**

Cook's Distance is an important tool for detecting outliers and influential points. Cook's Distance measures how the regression model changes when an observation is removed. A high Cook's Distance value suggests that the point has a greater impact on the model and may be an outlier or a high-leverage point [14].

**Hypothetical test:**

- $H_0$ (Null Hypothesis): No observations have a significant impact on the regression model.
- $H_1$ (Alternative Hypothesis): There are observations that have a significant impact on the regression model.

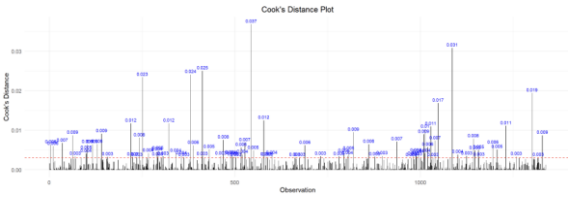


**Figure 18.** Cook's distance Plot

As shown (Figure 18), the Cook's Distance plot shows the Cook's Distance for each observation. The red dashed line represents the threshold$\frac{4}{n-p-1}$ [14].

If the Cook's Distance of an observation exceeds this threshold, the point is considered to have a significant impact on the model.

For our model, n (number of samples) is 1338, p (number of model parameters) = 8, so the threshold is:

$$Threshold = \frac{4}{1338-8-1} \approx 0.003. \quad (10)$$

In the Cook's Distance Plot, there are multiple points where the Cook's Distance exceeds the threshold of 0.003. These points are marked in blue in the figure, indicating that they have a greater impact on the model. The highest Cook's Distance is 0.037, which is well above the threshold, indicating that this observation has a significant impact on the model.

## 3.2. Identify Violations of the Assumptions

**Nonlinear and non-normal distributions:**

Residual plot analysis shows that the residuals are not randomly distributed, but exhibit a curvilinear pattern, especially in the larger residual regions, indicating that the model may have a nonlinear relationship.

QQ plot analysis shows that the residuals do not completely conform to the normal distribution, especially at both ends, and there is a large deviation between the actual quantile and the theoretical quantile.

**Hereoscedasticity:**

The Scale-Location plot analysis shows that the variability of the residuals changes with increasing fitted values. Especially in the larger fitted value region, the variability of the residuals increases significantly, which indicates that the model may have heteroskedasticity problems.

**Some Outliers and Influential Points:**

Residual vs Leverage plot analysis and Cook's Distance analysis show some high leverage points and influential points. These points have a significant impact on the model, especially those where the leverage value exceeds the threshold and Cook's Distance exceeds the threshold.

**Table 4.** Summary of All Violations

| Method | Assume Violation |
|---|---|
| Residual Plot Analysis | ✔ Non-linear relationship |
| QQ Plot Analysis | ✔Non-normal distribution |
| Scale-Location Plot Analysis | ✔Heteroskedasticity |
| VIF Analysis | ×No multicollinearity issues |
| Residuals vs Leverage Plot | ✔There are high leverage points and influential points |

# 4. Remedial measures

## 4.1. Step 1: Introducing Quadratic Term

On diagnostic inspection, we find that the residual plot showed a clear curvilinear trend between the fitted values and the residuals, suggesting that the model may not be linear.

To solve this problem, we decide to introduce the quadratic term of age ($age^2$) as a new predictor variable into this model. This approach helps capture the nonlinear relationship between age and health insurance premiums.

**Mathematical formula:**

The improved linear regression model is:

$$charges = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot age^2 + \beta_3 \cdot bmi + \beta_4 \cdot children + \beta_5 \cdot smoker + \beta_6 \cdot obs + \epsilon \quad (11)$$

**Using the Likelihood Ratio Test (LRT) to verify whether $age^2$ is significant:**

The Likelihood Ratio Test (LRT) is a statistical method used to compare two nested models [17].

**1. Build two model**

- Model 1:

$$charges = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot bmi + \beta_3 \cdot children + \beta_4 \cdot smoker + \beta_5 \cdot obs + \epsilon \quad (12)$$

- Model 2:

$$charges = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot age^2 + \beta_3 \cdot bmi + \beta_4 \cdot children + \beta_5 \cdot smoker + \beta_6 \cdot obs + \epsilon \quad (13)$$

**2. Comput the log-likelihood for each model**

- Model 1:

$$\ell_{\text{Reduced}} = \log L_0 = -13536.91 \ (\text{df} = 9) \quad (14)$$

- Model 2:

$$\ell_{\text{Full}} = \log L_1 = -13530.41 \ (\text{df} = 10) \quad (15)$$

**3. Calculate the likelihood ratio statistics**

- The formula is:

$$\lambda = \frac{L_{\text{Full}}}{L_{\text{Reduced}}} = \exp(\ell_{\text{Full}} - \ell_{\text{Reduced}}) \quad (16)$$

- The formula for calculating the likelihood ratio statistics is:

$$2\log(\lambda) = 2(\log L_1 - \log L_0) = 13.01094 \quad (17)$$

**4. Hypothesis testing**

- $H_0$ (Null Hypothesis): The introduced quadratic term ($age^2$) is not significant to the model, that is, $age^2$ does not need to be included in the model.
- $H_1$ (Alternative Hypothesis): The introduced quadratic term ($age^2$) is significant to the model, that is, $age^2$ needs to be included in the model.

Likelihood ratio statistics (LR) follows $\chi^2$ distribution under $H_0$, the degree of freedom $q$ is the number of new parameters.

In my model (**'model_improving1'**), $q$ = 1:

$$LR = 2(\ell(\hat{\theta}) - \ell(\hat{\theta}|\theta = \theta_0)) \sim \chi^2_{1-\alpha,q} \quad (18)$$

The critical value ($\chi^2_{0.05,1}$= 3.841) and LR value (≈ 13.01) are calculated through the code (Figure 28), and it is found that the LR value is greater than the critical value, indicating that $age^2$ is significant to the model and should be included in the model.

Then we calculated the p-value (which is 0.0003) through the code, which is far less than the significance level of 0.05. This once again shows that the new binomial $age^2$ has a significant impact on the model (reject $H_0$).

**Evaluate the model with new variable $age^2$:**

By introducing the quadratic term of age ($age^2$), the fitting degree of the model has been improved to a certain extent. Although the improved model has a slight increase in $R^2$ (which is 0.7573) and adjusted $R^2$ (which is 0.7558), this improvement is not significant.

## 4.2. Step 2: Interactive relationship

In order to further improve the model, we decide to find the interaction between predictor variables.

**1. Preliminary analysis of the relationship between *BMI* and *charges***

First of all, we analyse the relationship between charges and BMI. We draw a scattered point diagram between BMI and charges (Figure 19).

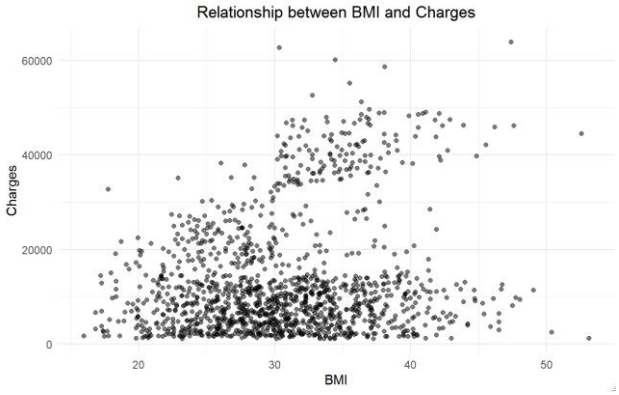


**Figure 19.** Relationship between BMI and charges

It can be seen from Figure 18 that there is no obvious linear relationship between BMI and charges.

**2. Add smoking state variables**

Next, we add smoking status (smoker) variables into the scatter map (Figure 20). It can be observed that the state of smoking has a significant impact on the relationship between BMI and charges. The data points of smokers (blue) and non-smoking (red) are obviously separated, indicating the state of smoking (smokers or non-smokers) interact with BMI.

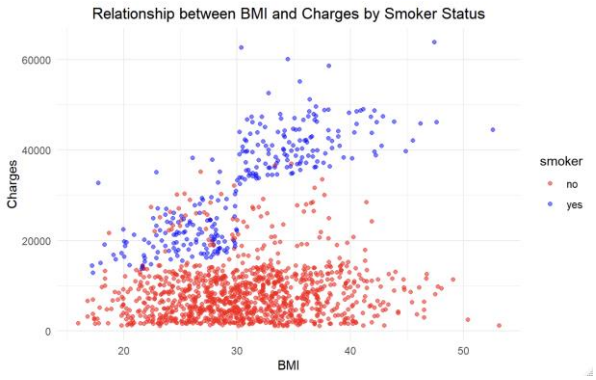


**Figure 20.** Relationship between BMI and charges by smoker status

**3. Further add BMI classification variables (obs)**

Add BMI classification variables (obs) variables (Figure 21) into the scattered point map to refine the analysis deeply.

Figure 21 shows that the distribution of smokers under different BMI classification has significant distribution in charges. For example, the charges for smokers under obese (orange) of BMI is significantly higher than that of smokers under other BMI classification. This phenomenon shows that there are significant interaction relationships between BMI classification (obs), smoking status and charges.

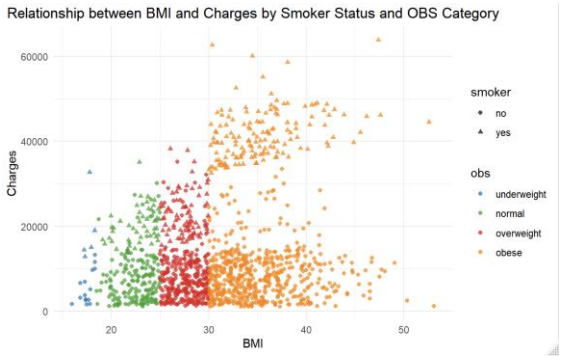


**Figure 21.** Relationship between BMI and Charges by smoker status and obs category

**4. Build mathematical formula**

After the figure analysis, we find a strong interaction between BMI, smokers and charges. Therefore, we build a new multiple linear regression model to introduce the square of the ages ($age^2$), and the interactive items between BMI classification, smoking status and BMI classification.

The mathematical formula of the model is as follows:

$$charges = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot age^2 + \beta_3 \cdot bmi + \beta_4 \cdot children + \beta_5 \cdot smoker + \beta_6 \cdot obs + \beta_7 \cdot (bmi \times smoker) + \beta_8 \cdot (obs \times smoker) + \epsilon \quad (19)$$

- $\beta_0$ is the intercept, $\beta_1 \ldots \beta_8$ are coefficients
- $\epsilon$ is the error term, which is assumed to be normally distributed with mean zero

## 4.3. Step 3: Threshold and Outliers

The original model (**'stepwise_model'**) have numerous high leverage points and influential points. We address this issue through using standardized residuals to identify and remove outliers. Standardized residuals help in evaluating whether each observation's residual is unusually large [15].

**1. Calculate standardized residuals**

The formula for the standardized residual is:

$$r_i^* = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad (20)$$

- $e_i$ is the residual for the *i*-th observation
- $s$ is the residual standard deviation
- $h_{ii}$ is the leverage value for the *i*-th observation.

**2. Set threshold**

A common threshold is 3. Observations with standardized residuals greater than this threshold in absolute value are considered outliers.

**3. Identify and remove outliers**

Identify and remove outliers using the threshold (which is 3). We get a new insurance (**'new_insurance'**) dataset with 1275 observations (original: 1338) after removing outliers.

## 4.4. Fit New Model and Obtain Results

The new model combines the quadratic term of age, the interaction term between BMI, smoking and BMI category (obs) to capture nonlinear and interactive effects. Meanwhile, it removes outliers to reduce the impact of high leverage points and influence points on the model.

The results obtained by the new model are:

```
Call:
lm(formula = charges ~ age + age2 + children + (bmi + obs) *
    smoker, data = new_insurance)

Residuals:
    Min      1Q  Median      3Q     Max
-1397.0  -692.1  -449.2  -109.4 14112.8

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1334.2832   868.5788   1.536    0.125
age                      -16.1125    30.8339  -0.523    0.601
age2                       3.5691     0.3844   9.284   <2e-16 ***
children                 587.9770    54.6017  10.768   <2e-16 ***
bmi                      -20.3839    22.3043  -0.914    0.361
obsnormal                272.2714   611.8936   0.445    0.656
obsoverweight            506.9823   631.8007   0.802    0.422
obsobese                 732.1352   704.7515   1.039    0.299
smokeryes                609.1540  1517.1257   0.402    0.688
bmi:smokeryes            546.9072    48.0481  11.382   <2e-16 ***
obsnormal:smokeryes     -327.8872  1333.8628  -0.246    0.806
obsoverweight:smokeryes -1081.4334 1379.0412  -0.784    0.433
obsobese:smokeryes      13577.2912  1539.2228   8.821   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2228 on 1262 degrees of freedom
Multiple R-squared:  0.9642,    Adjusted R-squared:  0.9639
F-statistic:  2835 on 12 and 1262 DF,  p-value: < 2.2e-16
```

- **Residuals:**

The residual standard error of the new model (**'model_improving2'**) is 2228, which is significantly lower than the 6011 of the original model, indicating that the new model has a better fitting effect. This means that the new model has greater accuracy in interpreting the data and the model has smaller errors.

- **Coefficients:**

The quadratic term of age ($age^2$) is highly significant with a p-value "<2e-16". This suggests a significant nonlinear relationship between age and charges.

The interaction term "bmi:smokeryes" and "obsobese:smokeryes" are also highly significant, indicating strong interactive effects.

- **R-squared:**

The Multiple $R^2$ of the new model is 0.9642, and the Adjusted $R^2$ is 0.9639, which are significantly higher than the 0.7549 ($R^2$) and 0.7536 (Adjusted $R^2$) of the original model. This shows that the proportion of total variation that the new model can explain significantly increases, and the model's explanatory power is stronger.

- **F-statistic:**

The F-statistic of the new model is 2835, and the p-value is "<2.2e-16", indicating that the overall model is statistically highly significant. The F statistic increases significantly, indicating that the introduction of new variables (including quadratic terms and interaction terms) significantly improves the fitting effect of the model.

In our new model (**'model_improving2'**), the nonlinear effect of age on charges is successfully captured by adding a quadratic term for age ($age^2$), the highly significant $age^2$ term further validates this point.

# 5. Comparison

## 5.1. Residual Analysis Plot

In the residual plot of the original model (left panel in Figure 22), we observe a clear curvilinear pattern, which indicates a nonlinear relationship between the predictor and response variables.

The residual plot of the new model (right panel in Figure 22) shows significant improvement. The residuals are evenly distributed above and below the zero line, and the red smooth line is also close to the zero line, indicating that there is no obvious systematic deviation in the residuals. In addition, the number and magnitude of outliers were significantly reduced, which suggests that the new model fits most data points more accurately.

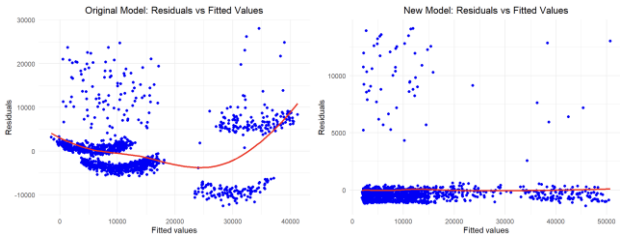


**Figure 22.** Residuals Analysis Plot in Original Model and New Model

## 5.2. Scale-Location Plot

In the Scale-Location plot of the original model (left panel in Figure 23), it can be observed that the fitted value increases, the square root of the standardized residual (the variability of the residual) also increases, meanwhile, heteroscedasticity exists.

In the new model (right panel in Figure 23), the red smooth

line is closer to the horizontal line, indicating that the variance of the residual no longer changes with the change of the fitted value, and also indicates that the heteroscedasticity problem has been improved. The number of points with high Cook's distance in the new model has been significantly reduced, suggesting that the robustness of the model has been improved.

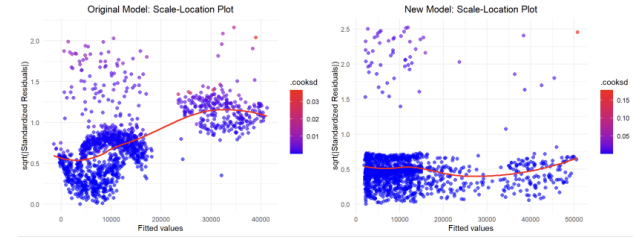


**Figure 23.** Scale-Location Plot in Original Model and New Model

## 5.3. Residuals vs Leverage Plot

There are high leverage points, high Cook's distance points, and residual patterns in the Figure 24 of the original model (left panel in Figure 24).

In the new model (right panel in Figure 24), there are fewer high leverage points, the number of high Cook's distance points is significantly reduced, and the residuals are closer to random distribution.

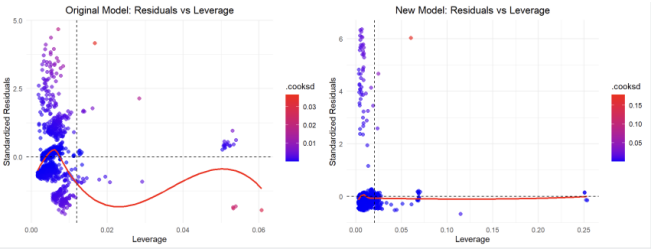


**Figure 24.** Residuals vs Leverage Plot in Original Model and New Model

## 5.4. Cook's Distance Plot

In the Cook's distance graph of the original model (left panel in Figure 25), there are multiple points where the Cook's distance exceeds the threshold. In the new model (right panel in Figure 25), the number of points where the Cook's distance exceeds the threshold is significantly reduced.

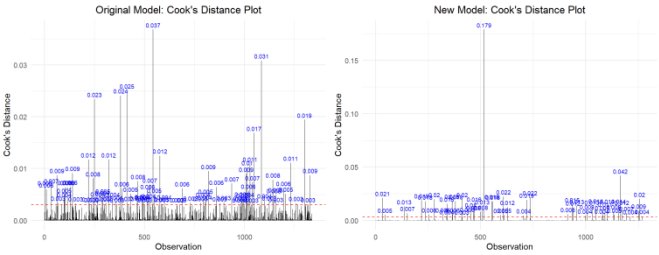


**Figure 25.** Cook's distance Plot in Original Model and New Model

## 5.5. Summary

The new model makes significant improvements in the following four aspects:

- **Nonlinear characteristics:** The new model better captures the nonlinear characteristics of the data, and the residual distribution is more random.
- **Heteroscedasticity:** The new model effectively alleviates the problem of hetereoscedasticity, and the variance of the residuals is more stable.
- **High leverage value points:** The new model reduces the number of high leverage value points, reducing the impact on the model.
- **Outliers and influence points:** The new model is more robust in handling outliers and high-leverage points, and the overall Cook's distance distribution is more concentrated.

# 6. COnclusion

## 6.1. Key Findings

The regression analysis aimed to predict personal medical insurance costs ('charges') using various factors including age, sex, BMI, number of children, smoking status, and BMI classification ('obs'). Through an iterative process of model diagnostics and improvements, we identified several key findings:

**1. Age and its nonlinear effect**

The introduction of the quadratic term for age ($age^2$) significantly improves the model's fit, suggesting a nonlinear relationship between age and charges. As age increases, charges rise at an increasing rate, emphasizing the importance of considering the nonlinear effect of age in forecasting medical expenses.

**2. Interaction between BMI, Smoker, and BMI Classification (obs)**

The interaction terms between BMI, smoking status, and BMI classification (obs) were found to be highly significant. Specifically, the analysis revealed that the impact of smoking on medical expenses varies significantly across different BMI categories. For instance, obese smokers incur significantly higher charges compared to non-smokers and smokers in other BMI categories.

**3. Significant variables**

The final model identifies age, BMI, number of children, smoking status, and BMI classification as significant predictors of charges. Each of these variables demonstrated a substantial impact on the cost predictions.

**4. Model improvement**

The enhanced model, which included the quadratic term for age and interaction terms, along with the removal of outliers, shows a substantial improvement in explanatory power. The $R^2$ value increased to 0.9642 from the original model's 0.7549, suggesting that the improved model explains approximately 96.42% of the variance in medical insurance costs, compared to 75.49% explained by the initial model.

## 6.2. Implications of the Results

The refined regression model provides several valuable insights and practical implications.

**1. Policy formulation**

The significant nonlinear effect of age on medical costs suggests that insurance companies should consider age-related adjustments more intricately. Premium structures could be designed to account for the exponential increase in medical costs with age, ensuring fair pricing for older beneficiaries. For example, Yang et al. [18] finds that medical expenditures rise significantly with age, particularly due to the increased prevalence of chronic diseases among older adults.

**2. Health management programs**

The interaction between BMI and smoking shows that a targeted health management plan is required. The initiative

aims to reduce the smoking rate, especially among people with higher BMI (belonging to the "obese' classification), may greatly reduce overall medical expenses. According to Finkelstein et al. [19], obesity and smoking will greatly increase medical care costs, and solving these risk factors may bring considerable savings.

**3. Risk assessment**

Insurance companies can leverage the refined models to enhance risk assessment and underwriting processes. By accurately identifying high-risk people based on age, physical quality index, smoking status and interaction, companies can better predict and manage potential claims.

**4. Future research directions**

The findings highlight the importance of exploring nonlinear relationships and interaction effects in regression models. Future research could delve deeper into other potential interactions and non-linear effects among variables to further refine predictive models.

**5. Model robustness**

The removal of outliers and influential points, along with the introduction of interaction terms, significantly enhanced the model's robustness. This approach ensures that the model is less sensitive to extreme values, providing more reliable predictions across different subsets of the population.

# References


[1] Choi, M. (2018). Medical Cost Personal Datasets [Data set]. Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

[2] Doe, J. (2020). The Relationship Between Geographic Location and Health Insurance Costs. Journal of Health Economics, 25(3), 123-135. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8414822/

[3] Smith, A. B. (2019). Age-related health issues and their impact on medical costs. Journal of Aging and Health, 31(4), 567-589. https://www.pdresources.org/course/index/1/1444/Ageism-Combatting-Stereotypes?

[4] Jones, M. A. (2021). Family size and healthcare costs: An analysis of economic impact. Health Economics Journal, 36(2), 231-245. https://www.healthequitygrandrounds.org/

[5] Centers for Disease Control and Prevention. (2020). Health effects of cigarette smoking. Retrieved from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/health_effects/effects_cig_smoking/index.htm.

[6] Centers for Disease Control and Prevention. (2021). Health effects of overweight and obesity. Retrieved from https://www.cdc.gov/healthyweight/effects/index.html

[7] Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological methods, 17(2), 228.

[8] Di Bucchianico, A. (2008). Coefficient of determination (R 2). Encyclopedia of statistics in quality and reliability.

[9] Belloto, J. R. J., & Sokolovski, T. D. (1985). Residual analysis in regression. American Journal of Pharmaceutical Education, 49(3), 295-303.

[10] Mackay, D. S., Ewers, B. E., Loranty, M. M., & Kruger, E. L. (2010). On the representativeness of plot size and location for scaling transpiration from trees to a stand. Journal of Geophysical Research: Biogeosciences, 115(G2).

[11] Augustin, N. H., Sauleau, E. A., & Wood, S. N. (2012). On quantile quantile plots for generalized linear models. Computational Statistics & Data Analysis, 56(8), 2404-2409.

[12] Liao, D., & Valliant, R. (2012). Variance inflation factors in the analysis of complex survey data. Survey Methodology, 38(1), 53-62.

[13] Rousseeuw, P. J. (1991). A diagnostic plot for regression outliers and leverage points. Computational Statistics & Data Analysis, 11(1), 127-129.

[14] Díaz-García, J. A., & González-Farías, G. (2004). A note on the Cook's distance. Journal of statistical planning and inference, 120(1-2), 119-136.

[15] Kiebel, S. J., Poline, J. B., Friston, K. J., Holmes, A. P., & Worsley, K. J. (1999). Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. Neuroimage, 10(6), 756-766.

[16] Miles, J. (2005). R-squared, adjusted R-squared. Encyclopedia of statistics in behavioral science.

[17] Woolf, B. (1957). The log likelihood ratio test (the G-test). Annals of human genetics, 21(4), 397-409.

[18] Yang, Z., Norton, E. C., & Stearns, S. C. (2003). Longevity and health care expenditures: the real reasons older people spend more. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 58(1), S2-S10. https://doi.org/10.1093/geronb/58.1.S2

[19] Finkelstein, E. A., Trogdon, J. G., Cohen, J. W., & Dietz, W. (2009). Annual Medical Spending Attributable to Obesity: Payer- and Service-Specific Estimates. Health Affairs, 28(Suppl1), w822-w831.

https://doi.org/10.1377/hlthaff.28.5.w822