The Price of Health:
Multiple Linear Regression Analysis of Medical Charges Based on Insurance Data

Wenxin Wang, Lingzhi Pan, Xinrui Wang, Zihan Geng, Chen Yang

1. INTRODUCTION

1.1. MOTIVATION

1.1.1. Context

Everyone cares about their health, and healthcare costs are a must expense for every one of us. So understanding the factors that influence healthcare costs is crucial for not only policymakers but also individuals.

Through analyzing the related dataset, we can gain insights into which factors contribute the most to healthcare expenses or try to find some unique relations between these variables.

This analysis is important because it can also inform insurance companies about risk factors that lead to higher charges, guiding pricing strategies for premiums. It can also help in identifying high-risk populations that might benefit from targeted interventions, ultimately helping reduce healthcare costs and improve overall public health.

The project investigates the influence of personal characteristics (e.g., age, gender, body mass index, etc.) on medical insurance charges. This analysis is primarily applied in the health insurance sector to better understand the impact of individual health factors on insurance costs. It aims to optimize premium structures and develop fair pricing strategies.

1.1.2. Problem

The main problem is the complexity and non-linear relationships among factors affecting medical insurance charges. This project focuses on identifying and quantifying the impacts of age, BMI, smoking status, having children, and geographic region on insurance costs by building an accurate regression model.

1.1.3. Challenges

The challenges in this study include:

Complex variable interactions: Some variables like smoking status and BMI, may interact and jointly influence insurance charges, which needs us to find interaction models.
Non-linear relationships: Many variables (e.g., age, BMI) may have non-linear effects on costs, requiring higher-order terms to be added to the model.

Outliers in data: Insurance charge data may contain outliers (e.g., extremely high medical expenses), which need careful handling to avoid skewing the results.

Statistical assumptions: Ensuring that the regression model adheres to statistical assumptions such as normality and homoscedasticity of residuals is crucial; violations may compromise the validity of the findings.

1.2. OBJECTIVES

1.2.1. Overview

Our intent of this project is to develop a predictive model for estimating individual insurance costs using linear regression. By analyzing various features, such as age, BMI, presence of children, smoking status, and region, the project aims to identify the key factors that significantly influence insurance charges. Additionally, the model explores possible interactions and non-linear relationships between these variables to improve predictive accuracy.

The main objectives are to determine which variables have a substantial impact on insurance costs and to construct an optimized model that can effectively estimate these costs. This is achieved by building several regression models, including additive, interaction, and higher-order models, and iteratively improving the fit of the model. Furthermore, the project involves validating key assumptions of linear regression, such as linearity, homoscedasticity, and normality of residuals, to ensure the robustness of the model.

Ultimately, the purpose of this group project is to create a reliable tool for predicting insurance costs, which can support insurance companies or related organizations in cost estimation and risk assessment.

1.2.2. Goals & Research Questions

Goals:

1. Identify Significant Predictors: Determine which factors, such as age, BMI, smoking status, or number of children, significantly influence individual insurance costs.
2. Optimize Predictive Model: Develop and refine a linear regression model that accurately predicts insurance costs, using various approaches such as including interaction terms and higher-order variables.
3. Model Evaluation and Validation: Validate the developed model against key linear regression assumptions, such as linearity, homoscedasticity, and normality of residuals, to ensure its reliability and accuracy.
4. Data Visualization: Use data visualization techniques to understand the distribution of variables, identify relationships among them, and support model interpretation.

Research Questions:

1. Which factors most significantly affect individual insurance costs? Are age, BMI, smoking status, and other variables all equally important in determining insurance charges?
2. How do the relationships between predictors and insurance costs vary? Are there non-linear relationships between certain predictors (e.g., age or BMI) and insurance costs?
3. Do any interaction effects exist between variables that influence insurance costs? For example, does smoking have a stronger impact on insurance costs for individuals with a higher BMI?

4. How well does the model fit the data, and does it meet linear regression assumptions? Does the model satisfy assumptions of linearity, homoscedasticity, and normality, and how can violations be addressed?
5. How can visual analytics support the understanding of relationships in the data? What visual insights can be obtained from scatter plots, residual plots, and other forms of graphical analysis to guide model improvement?

2. METHODOLOGY

2.1 Data

The dataset used in this project was sourced from Kaggle at https://www.kaggle.com/datasets/mirichoi0218/insurance/data. It contains information on individual insurance charges along with demographic and health-related attributes. The dataset comprises 1,338 rows and 7 attributes. A brief description of each attribute is as follows:

- **Age**: The age of the individual (numeric).
- **Sex**: Gender of the individual (male or female).
- **BMI**: Body Mass Index, a measure of body fat based on height and weight (numeric).
- **Children**: Indicates if the individual has children (yes or no).
- **Smoker**: Indicates if the individual is a smoker (yes or no).
- **Region**: The residential region in the U.S. (northeast, northwest, southeast, southwest).
- **Charges**: Medical insurance charges billed to the individual (numeric).

2.2 Approach

We employed statistical modeling and data visualization to analyze the relationship between insurance charges and the predictor variables. Two types of models were evaluated:

1. **Additive Model**: Considers only the main effects of the predictors.
2. **Interaction Model**: Includes interaction terms and higher-order terms to capture complex relationships.
3. **Assumption Test**: Check multicollinearity, linearity, normality, homoscedasticity and outliers assumptions.

This approach works well because it allows us to systematically identify significant predictors, evaluate their relationships with the response variable, and ensure model assumptions are satisfied. Additionally, visualization techniques help uncover patterns and validate model fits.

2.3 Workflow

The workflow for this project includes the following steps:

1. **Data Preprocessing**:
    - Load the dataset and clean the data.
    - Convert categorical variables into factors.
2. **Exploratory Data Analysis (EDA)**:
    - Visualize the relationships between variables.

    - Identify potential nonlinearities or interactions.
3. **Model Development**:
    - Fit an additive model to capture the main effects.
    - Evaluate interaction terms and higher-order variables.
4. **Model Validation**:
    - Check assumptions such as multicollinearity, normality, and homoscedasticity.
    - Use metrics like Adjusted R-squared and Residual Standard Error (RSE) for model comparison.
5. **Result Interpretation**:
    - Derive insights based on significant predictors.
    - Summarize the model equations.
6. **Report Preparation**:
    - Present findings with relevant plots and statistical summaries.

The hardest step is ensuring the model satisfies assumptions (e.g., normality and homoscedasticity). If these assumptions are violated, potential solutions include:

- Transforming the data (e.g., log transformation).
- Adding or removing interaction terms or higher-order variables.
- Using alternative models such as generalized linear models (GLMs).

2.4 Contributions

The workload distribution among the five team members was as follows:

Model Development (Lingzhi Pan):
Built and tested the initial regression models, including additive and interaction models.
Analyzed the significance of variables and refined the model structure.

Higher-Order Term Analysis (Xinrui Wang):
Investigated the impact of higher-order terms. Improved the model's accuracy by evaluating adjustments and testing significance.

Visualization and Statistical Validation (Wenxin Wang):

Generated plots, including residuals, fitted values, and variable interactions, to evaluate model assumptions. Performed statistical tests to ensure validity.

Outlier and Influence Analysis (Zihan Geng):
Identified influential data points and leverage values using Cook's distance and leverage metrics.
Assessed their impact on model performance and ensured robust results.

Report Writing and Presentation (Chen Yang):
Compiled all analyses, visualizations, and interpretations into a comprehensive report. Prepared and delivered the final presentation, summarizing findings and conclusions.

All team members participated in discussions, provided feedback on each stage, and collectively ensured the quality of the final report. Collaboration was key in addressing challenges during model development and assumption testing.

3 MAIN RESULTS OF THE ANALYSIS

3.1 Models

3.1.1 Initial Model

Firstly, we will find the additive model which only contains the main effects. We first fit the full model and observe the analysis results.

```
model1 <- lm(charges~age+factor(sex)+bmi+factor(child)+factor(smoker)+factor(region))
stepmod=ols_step_both_p(model1,p_enter = 0.05, p_remove = 0.1, details=TRUE)
summary(stepmod$model)
```

From the output, we find the model applied Stepwise Regression Procedure with p_enter=0.05 and p_remove=0.1 is:

$$\hat{y} = -12149.61 + 257.94x_{age} + 321.73x_{bmi} + 990.74x_{child} + 23810.76x_{smoker}$$

We can consider that sex has no influence on insurance charges. Dummy variable region_northwest has t_cal=-0.739 and p-value=0.46029>0.05. We can not reject H_0:β_(region_n orthwest)=0. But other region dummy variables have P-value<0.05. So we keep the region variable. We can get a reduced model:

```
model1_reduce <- lm(charges~age+bmi+factor(child)+factor(smoker)+factor(region))
summary(model1_reduce)
```

From the output, only the dummy variable region_northwest has t_cal=-0.737 and p-value= 0.4611>0.05. Other variables have p-value<0.05.
We can try to delete region variables and use ANOVA to test the model before and after deletion.

```
model1_nonregion <- lm(charges~age+bmi+factor(child)+factor(smoker))
summary(model1_nonregion)
anova(model1_nonregion, model1_reduce)
```

From the ANOVA output, F_cal=2.1383 with df=3,1330 and p-value=0.09361. We can not reject H_0:β_region=0. We can consider the region does not influence insurance charges. We can delete the region variable. Therefore, we can consider that age, BMI, child, and smoker have a significant influence on insurance charges.

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(child) + factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12160.1  -2978.6   -970.7   1488.1  29556.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -12149.61     948.52 -12.809  < 2e-16 ***
## age                  257.94      11.91  21.653  < 2e-16 ***
## bmi                  321.73      27.41  11.738  < 2e-16 ***
## factor(child)1       990.74     335.96   2.949  0.00324 **
## factor(smoker)yes  23810.76     411.70  57.835  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6075 on 1333 degrees of freedom
## Multiple R-squared:  0.7491, Adjusted R-squared:  0.7484
## F-statistic:   995 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The final additive model is:

$$\hat{y} = -12149.61 + 257.94x_{age} + 321.73x_{bmi} + 990.74x_{child} + 23810.76x_{smoker}$$

where y is insurance charges;
x_age is the individual's age;
x_bmi is the body mass index;
x_child is a dummy variable. x_child=1 is have children; x_child=0 is no children;
x_smoker is a dummy variable. x_smoker=1 is smoker; x_smoker=0 is non-smoker;

After we find the final additive model, we will check the multicollinearity.

imcdiag(model1_nonregion, method="VIF")

```
##
## Call:
## imcdiag(mod = model1_nonregion, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                      VIF detection
## age               1.0149         0
## bmi               1.0122         0
## factor(child)1    1.0025         0
## factor(smoker)yes 1.0008         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
```

All VIFs are close to 1, which means that there is very weak multicollinearity, but it is not severe enough to warrant corrective measures. Therefore, for these main effects we do not need to consider the multicollinearity.

3.1.2 Interaction Model

After we find the additive model and main effects, we consider adding some interaction terms into the model. First, we add all the interaction variables and observe their significance.

```
model_in <- lm(charges~(age+bmi+factor(child)+factor(smoker))^2)
summary(model_in)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + factor(child) + factor(smoker))^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13690.6  -2056.0  -1248.6   -310.8  29372.1
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -8.727e+02  2.180e+03  -0.400   0.6889
## age                              2.020e+02  4.998e+01   4.042 5.61e-05 ***
## bmi                             -5.717e+01  6.976e+01  -0.820   0.4126
## factor(child)1                   1.755e+03  1.527e+03   1.149   0.2508
## factor(smoker)yes               -1.960e+04  1.869e+03 -10.486  < 2e-16 ***
## age:bmi                          1.989e+00  1.564e+00   1.272   0.2036
## age:factor(child)1               5.556e+00  1.950e+01   0.285   0.7757
## age:factor(smoker)yes           -6.733e-02  2.395e+01  -0.003   0.9978
## bmi:factor(child)1              -2.530e+01  4.479e+01  -0.565   0.5723
## bmi:factor(smoker)yes            1.437e+03  5.331e+01  26.953  < 2e-16 ***
## factor(child)1:factor(smoker)yes -1.174e+03  6.736e+02  -1.743   0.0816 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4886 on 1327 degrees of freedom
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.8372
## F-statistic: 688.6 on 10 and 1327 DF,  p-value: < 2.2e-16
```

From the output, only the interaction term bmi×smoker has t_cal=26.953 and p-value<0.05. We can reject H_0:β_(bmi×smoker)=0. We can consider the interaction term bmi×smoker to have a significant influence on insurance charges. After deleting other interaction terms, the reduced interaction model is:

```
model_in1 <- lm(charges ~ age+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))
summary(model_in1)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + factor(child) + factor(smoker) +
##     bmi * factor(smoker))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14992.4  -2033.6  -1242.8   -372.9  29880.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2753.520    838.589  -3.284 0.001052 **
## age                      265.192      9.585  27.667  < 2e-16 ***
## bmi                        6.423     24.950   0.257 0.796894
## factor(child)1           960.640    270.224   3.555 0.000391 ***
## factor(smoker)yes     -20082.183   1659.601 -12.101  < 2e-16 ***
## bmi:factor(smoker)yes   1430.143     52.987  26.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4886 on 1332 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.8372
## F-statistic:  1376 on 5 and 1332 DF,  p-value: < 2.2e-16
```

From this output about the reduced interaction model, all the interaction terms are significant. We can compare the R^2_{Ryx} and RSE between additive and interaction models.

data.frame(Model = c("additive", "interaction"),
AdjRsq=c(summary(model1_nonregion)$adj.r.squared,summary(model_in1)$adj.r.squared),
RSE=c(summary(model1_nonregion)$sigma,summary(model_in1)$sigma))

```
##         Model    AdjRsq      RSE
## 1    additive 0.7483611 6074.821
## 2 interaction 0.8372067 4886.105
```

The interaction model has a greater R^2_{Ryx} =0.8372067 and a smaller RSE=4886.105. We prefer the interaction model. We also compare the additive and interaction models by ANOVA.

anova(model1_nonregion, model_in1)

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + factor(child) + factor(smoker)
## Model 2: charges ~ age + bmi + factor(child) + factor(smoker) + bmi *
##     factor(smoker)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1333 4.9192e+10
## 2   1332 3.1800e+10  1 1.7392e+10 728.49 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Source	SS	df	MS	F	p-value
Regression	1.7392×10^{10}	1	1.7392×10^{10}	728.49	0
Residual	3.18×10^{10}	1332	2.3874×10^{7}		
Total	4.9192×10^{10}	1333			

From the ANOVA table, F_cal = 728.49 with df= 1,1332 and p − value < 0.05. We can reject H_0 : β_bmi×smoker = 0. Therefore, interaction model is better to fit insurance charges which is:

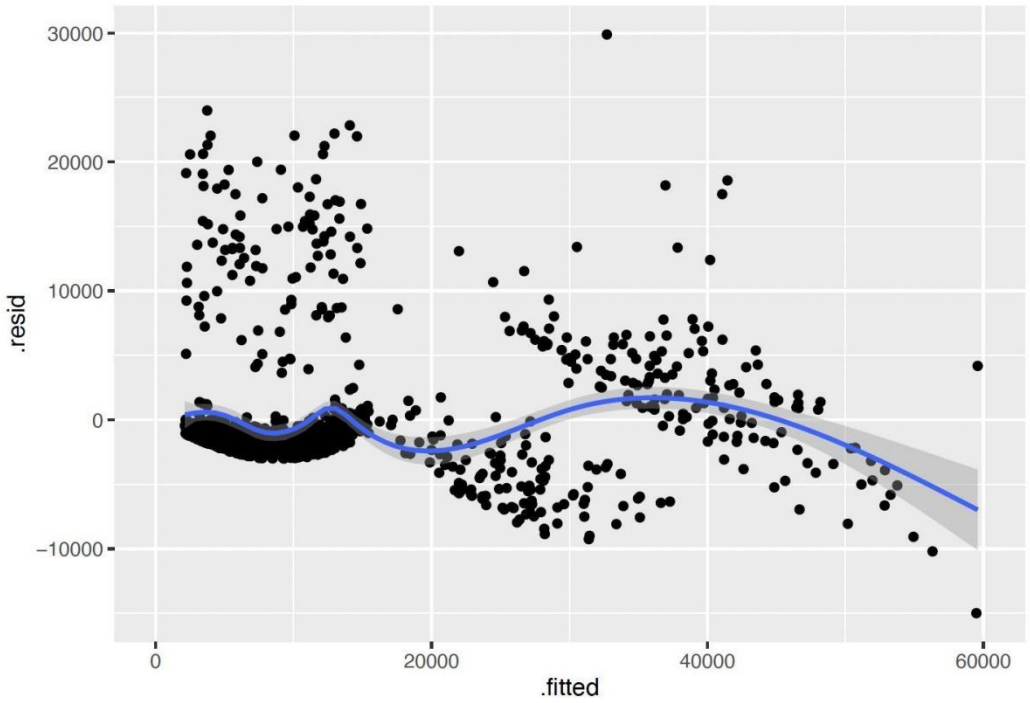
$$\hat{y} = -2753.520 + 265.192x_{age} + 6.423x_{bmi} + 960.640x_{child} - 20082.183x_{smoker} + 1430.143x_{bmi}x_{smoker}$$

where y is insurance charges;
x_age is the individual's age;
x_bmi is the body mass index;
x_child is a dummy variable. x_child=1 is have children; x_child=0 is no children;
x_smoker is a dummy variable. x_smoker=1 is smoker; x_smoker=0 is non-smoker;

3.1.3 Assumption

After we find the best-fit model, we need to check the assumptions for the model. First, we check the linearity assumption:

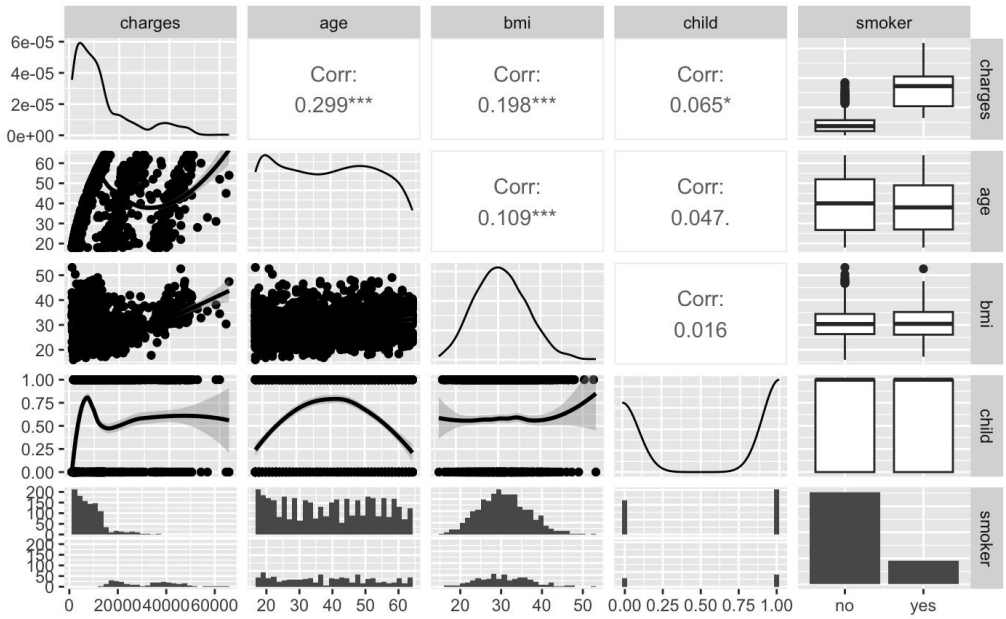
ggplot(model_in1, aes(x=.fitted, y=.resid)) +geom_point() +geom_smooth()+xlim(0,60000)



geom_hline(yintercept = 0)

There seems to be some sort of pattern happening with our residuals. So we will use 'ggpairs' to draw scatter plots and see if there are higher-order variables.

```
df1 <- df[c('charges','age','bmi','child','smoker')]
ggpairs(df1,lower = list(continuous = "smooth_loess", combo =
 "facethist", discrete = "facetbar", na = "na"))
```



From the scatter plots, the relationship between age and charges looks like it might be quadratic or something higher. At the same time, the relationship between BMI and charges looks like it might be quadratic or something higher. We try to add a higher-order variable of age:

*model_age2 <- lm(charges~age+I(age^2)+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))*

model_age3 <-

*lm(charges~age+I(age^2)+I(age^3)+bmi+factor(child)+factor(smoker)+bmi*factor(smoker))*

```
##   Model    AdjRsq      RSE
## 1   age 0.8372067 4886.105
## 2  age2 0.8397518 4847.760
## 3  age3 0.8396350 4849.527
```

model_bmi2 <-

*lm(charges~age+I(age^2)+bmi+I(bmi^2)+factor(child)+factor(smoker)+bmi*factor(smoker))*

model_bmi3 <-

*lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+factor(child)+factor(smoker)+bmi*factor(smoker))*

summary(model_bmi3)

model_bmi4 <-

*lm(charges~age+I(age^2)+bmi+I(bmi^2)+I(bmi^3)+I(bmi^4)+factor(child)+factor(smoker)+bmi*factor(smoker))*

```
##   Model    AdjRsq      RSE
## 1   bmi 0.8397518 4847.760
## 2  bmi2 0.8410058 4828.755
## 3  bmi3 0.8428621 4800.483
## 4  bmi4 0.8431882 4795.500
```

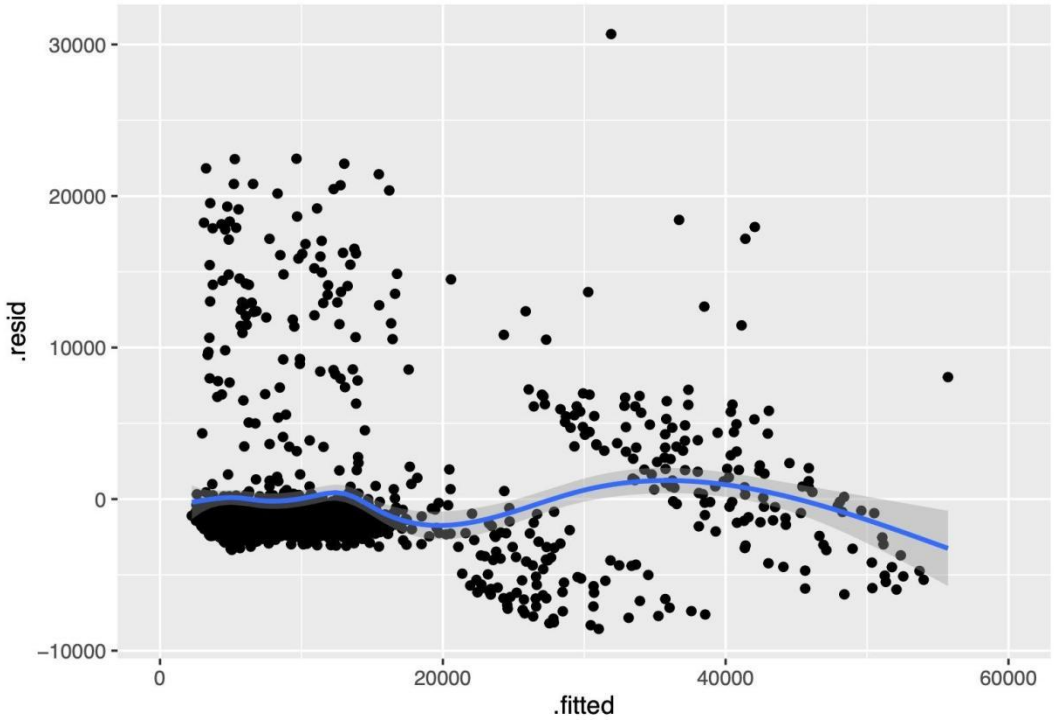
From the table above, we can find that RSE becomes smaller as the order of BMI increases, and the smallest R^2_{RyX} = 0.8431882 and is found when we add. But the of is greater than 0.05 which is not significant. Therefore, we decide to add and into the model. The final model after adding high-order variables is as follows:

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child} - 20320x_{smoker} + 1439x_{bmi}x_{smoker}$$

where y is insurance charges;

is x_age the individual's age;
is x_bmi the body mass index;
is x_child a dummy variable. x_child=1 is have children; x_child=0 is no children;
is x_smoker a dummy variable. x_smoker=1 is smoker; x_smoker=0 is non-smoker;

ggplot(model_bmi3, aes(x=.fitted, y=.resid)) +geom_point() + geom_smooth()+xlim(0,60000)



It looks a little smoother than it did before we increased the higher order.
After adding higher-order variables, we test this model for homogeneity of variance and normality assumptions. We use the Breusch-Pagan test to test the equal variance assumption:
H_0: heteroscedasticity is not present
H_1: heteroscedasticity is present

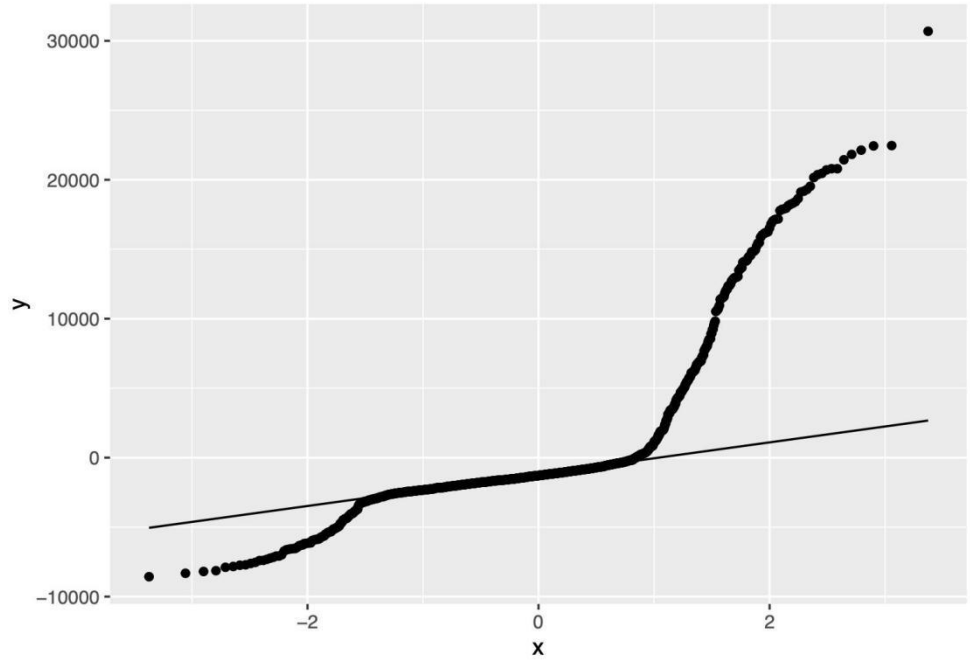
bptest(model_bmi3)

```
##
##  studentized Breusch-Pagan test
##
## data:  model_bmi3
## BP = 9.4923, df = 8, p-value = 0.3025
```

From the output, BP=9.4923 and p-value=0.3025>0.05, we can not reject H_0. Therefore, we can conclude that this model does meet the equal variance assumption.

Then, we draw qqplot and use the Shapiro-Wilk test (S-W) to test the normality assumption:
H_0:the sample data are significantly normally distributed
H_1:the sample data are not significantly normally distributed

ggplot(df, aes(sample=model_bmi3$residuals)) + stat_qq() + stat_qq_line()



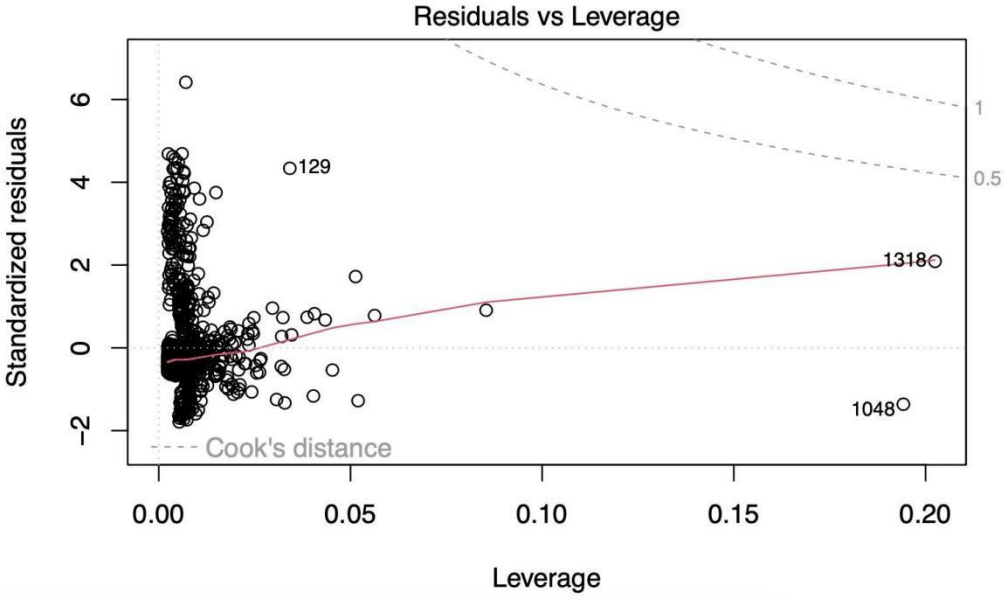
shapiro.test(residuals(model_bmi3))

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_bmi3)
## W = 0.66184, p-value < 2.2e-16
```

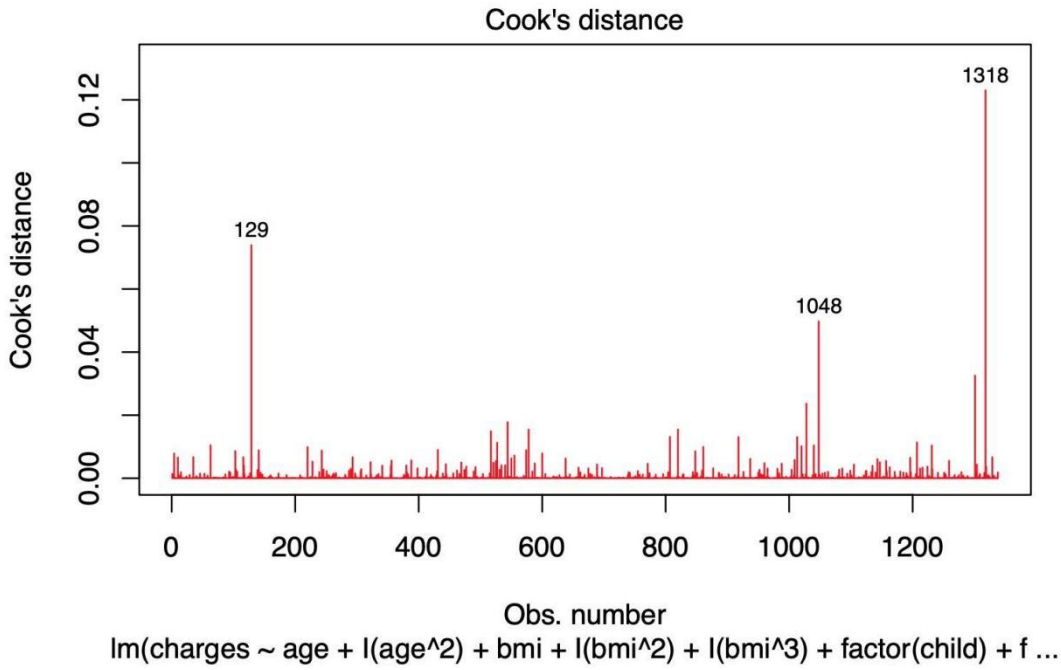
From the output, the residual data do not have a normal distribution.W=0.66184 and p-value=0.05. We can reject H_0. We do not have normality.

We can find the outlier points by Cooks distance:

plot(model_bmi3,which=5)



plot(model_bmi3,pch=18,col="red",which=c(4))

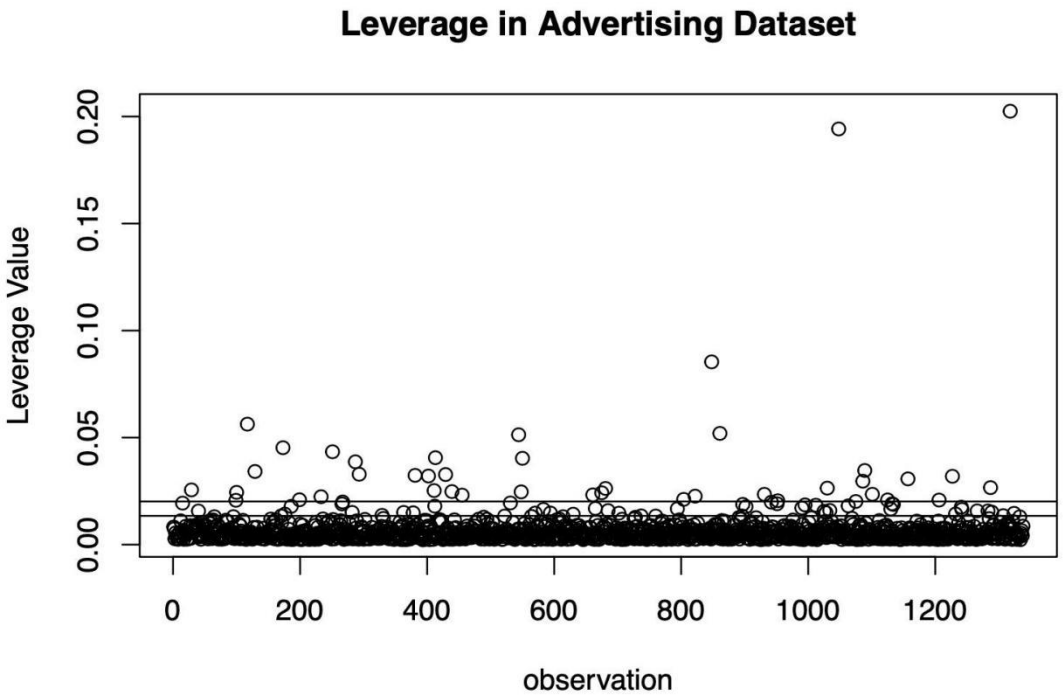


From the plots above, we can see that data points 129, 1048, and 1318 have a larger cook's distance than other points but smaller than 0.5 and all cases are well inside of the Cook's distance lines. Therefore, there is no influential case. Then, we can find the Leverage points.

```
lev=hatvalues(model_bmi3)
p = length(coef(model_bmi3))
n = nrow(df)
outlier3p = lev[lev>(3*p/n)]
print(outlier3p)
```

```
##          29          99         100         117         129         173         199
## 0.02557650 0.02071208 0.02442002 0.05628873 0.03417414 0.04528710 0.02095753
##         233         251         287         293         381         402         411
## 0.02238290 0.04338883 0.03869951 0.03287624 0.03236663 0.03206159 0.02525940
##         413         429         439         455         544         548         550
## 0.04060996 0.03275058 0.02478133 0.02314994 0.05133024 0.02464643 0.04032560
##         661         675         681         804         822         848         861
## 0.02325163 0.02422694 0.02626306 0.02114614 0.02264806 0.08537652 0.05197711
##         931         952        1030        1048        1086        1089        1101
## 0.02351193 0.02052329 0.02639497 0.19419925 0.02964513 0.03464202 0.02353288
##        1125        1157        1206        1227        1287        1318
## 0.02089923 0.03072183 0.02084025 0.03194576 0.02663214 0.20246511
```

```
plot(rownames(df),lev, main = "Leverage in Advertising Dataset", xlab="observation",
     ylab = "Leverage Value")
abline(h = 3 *p/n, lty = 1)
```



There are 41 leverage points with leverage values > 3p/n.

3.1.4 Final Model

Therefore, our final model is:

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child} - 20320x_{smoker} + 1439x_{bmi}x_{smoker}$$

where y is insurance charges;
x_age is the individual's age;
x_bmi is the body mass index;
x_child is a dummy variable. x_child=1 is have children; x_child=0 is no children;
x_smoker is a dummy variable. x_smoker=1 is smoker; x_smoker=0 is non-smoker;
The interpretations of the final model:
For families non-smokers, the model is:

$$\hat{y} = 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child}$$

which is:

$$\hat{y} = \begin{cases} 29260 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,with child} \\ 27770 - 42.73x_{age} + 3.853x_{age}^2 - 2788x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 & \text{,without child} \end{cases}$$

The insurance charges (y) vary with age (x_age). Both linear (x_age) and quadratic (x^2_age) terms are included in the model, capturing a nonlinear relationship. Since the samples are all older than 18, in this case, the decrease trend from age 0 to 5 would not show up, only the increase in charge gradually goes from slow to fast with age, reflecting the higher medical risks associated with aging.

The insurance charges (y) change with BMI (x_bmi) in a nonlinear pattern, as captured by the linear (x_bmi), quadratic (x^2_bmi), and cubic (x^3_bmi) terms in the model. The negative linear coefficient β_bmi=-2788 suggests that charges decrease slightly at lower BMI levels. However, the positive quadratic coefficient β_bmi^2=97.78 reflects an upward trend in charges as BMI increases.

The small negative cubic coefficient (β_bmi^3=-1.093) further moderates this increase at extremely high BMI levels, reflecting the complex relationship between BMI and medical costs. This pattern aligns with the understanding that both underweight and overweight individuals are at greater health risks, leading to higher insurance charges.

The insurance charges (y) vary about without or with children (x_child). The average insurance charge is 1490 units higher with children than without children when other variables are constant. This indicates that having children has a positive effect on insurance charges.

For families with smokers the model is:

$$\hat{y} = 27770 - 20320 - 42.73x_{age} + 3.853x_{age}^2 + (1439 - 2788)x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3 + 1490x_{child}$$

which is:

$$\hat{y} = \begin{cases} 8940 - 42.73x_{age} + 3.853x_{age}^2 - 1349x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3, \text{with child} \\ 7450 - 42.73x_{age} + 3.853x_{age}^2 - 1349x_{bmi} + 97.78x_{bmi}^2 - 1.093x_{bmi}^3, \text{without child} \end{cases}$$

The relationship and coefficient of age variable (x_age) between smoking and non-smoking do not change. Both linear (x_age) and quadratic (x^2_age) terms are included in the model, capturing a nonlinear relationship. The charges decrease slightly with linear coefficient (β_age=-42.7318) until the ages reach 6, but because all the sample ages are older than 18, they only show the increasing trend, and increase more steeply for older individuals as the quadratic coefficient (β_age^2=3.8525).

The relationship of the bmi variable (x_bmi) between smoking and non-smoking does not change. But part of the coefficient has changed. The negative linear coefficient increase to (β_bmi=-1349) suggests that charges decrease more slightly at lower BMI levels than non-smoker. However, the positive quadratic coefficient (β_bmi^2=97.78) reflects an upward trend in charges as BMI increases. The small negative cubic coefficient (β_bmi^3=-1.093) further moderates this increase at extremely high BMI levels.

The insurance charges (y) vary about without or with children (x_child). The average insurance charge is still 1490 units higher with children than without children, when other variables are constant.

Smoking and non-smoking change the relationship and coefficients between age (x_age), child (x_child) variables and insurance charges. But it will change the intercept and the linear coefficient of bmi. The intercept is reduced and the coefficient on bmi is increased.

3.2 Results

The analysis found that **age**, **BMI**, **having children**, and **smoking status** are significant predictors of insurance costs. The specific findings are as follows:

1. **Significant Predictors**:
    - **Age**: When considering individuals aged 18 and above (x_age ⩾ 18), there is a positive relationship between age and insurance costs. Costs increase slowly at first and then rise more rapidly with age, reflecting the growing medical risks associated with older individuals.
    - **BMI**: The relationship is non-linear, with linear, quadratic, and cubic terms showing significance. Both very low and high BMI levels are associated with higher insurance costs, aligning with health risks at these extremes.

    - **Children**: Having children increases insurance costs by an average of 1,490 units compared to those without children.
    - **Smoking**: Significantly raises insurance charges, reflecting the high medical risks associated with smoking.
2. **Insignificant Predictors**:
    - **Gender**: No significant impact on insurance costs, likely due to dataset characteristics or uniform pricing policies.
    - **Region**: While some regions had significant effects, the northwest region did not show a meaningful impact, contrary to expectations about regional healthcare cost differences.
3. **Interactions**:
    - The interaction between **BMI and smoking** significantly influences costs, with smokers showing steeper cost increases at higher BMI levels compared to non-smokers.

3.3 Practical Significance

By performing a multiple regression analysis on the medical insurance cost dataset provided by Kaggle, we can gain insights into how various factors influence individual medical insurance expenses. This analysis has the following practical significance:

1. **Identifying Key Influencing Factors**

Determine which factors (e.g., age, BMI, smoking status) significantly affect medical expenses. This can help insurance companies consider these variables when setting premiums, leading to more accurate pricing strategies.

2. **Personalized Premium Pricing**

By quantifying the impact of each factor on medical costs, insurance companies can design personalized premium rates for different customer groups, improving fairness and competitiveness in pricing.

3. **Risk Assessment and Management**

Understanding how lifestyle factors such as smoking and obesity influence medical expenses helps insurers assess customer health risks and develop corresponding health management programs to reduce future claims.

4. **Market Segmentation and Marketing Strategies**

By analyzing the expense patterns of different demographic groups, insurers can conduct market segmentation and develop targeted marketing strategies to enhance customer satisfaction and market share.

5. **Policy Making and Health Advocacy**

Analyzing the impact of lifestyle factors on medical expenses provides valuable insights for insurers and policymakers to advocate for healthier lifestyles, ultimately reducing overall healthcare costs.

In summary, conducting a multiple regression analysis on this dataset not only aids insurance companies in optimizing pricing and risk management but also offers data-driven support for public health policy development.

4 CONCLUSION

In the current project, the approach shows strengths in stepwise model development and diagnostics. By building models iteratively—from additive to interaction and higher-order models—the team could better understand the impact of each variable on insurance costs and enhance model interpretability. The inclusion of interaction terms and higher-order variables helped capture complex non-linear relationships, improving the model fit. Moreover, diagnostic tools like residual plots, Breusch-Pagan tests, and Shapiro-Wilk tests were employed to ensure the model's statistical soundness and reliability.

5 REFERENCES

1. Miri Choi, Medical Cost Personal Datasets , 2017 , https://www.kaggle.com/datasets/mirichoi0218/insurance
2. Danika Lipman, Statistical Modelling with Data: MULTIPLE LINEAR REGRESSION, 2023, based on content created by Thuntida Ngamkham
3. Danika Lipman, Statistical Modelling with Data: MULTIPLE LINEAR REGRESSION PART 2, 2023, based on content created by Thuntida Ngamkham
4. Danika Lipman, Statistical Modelling with Data: MULTIPLE LINEAR REGRESSION PART 3, 2023, based on content created by Thuntida Ngamkham
5. Danika Lipman, Statistical Modelling with Data: MULTIPLE LINEAR REGRESSION PART 4, 2023, based on content created by Thuntida Ngamkham

End of Project Report