Investigation of the Relation Between Annual Hospital Bills and Demographic Information

Alan Wei, Connor Johst, Jacqueline Liang, Muxi Jin

2022-08-10

1. Introduction

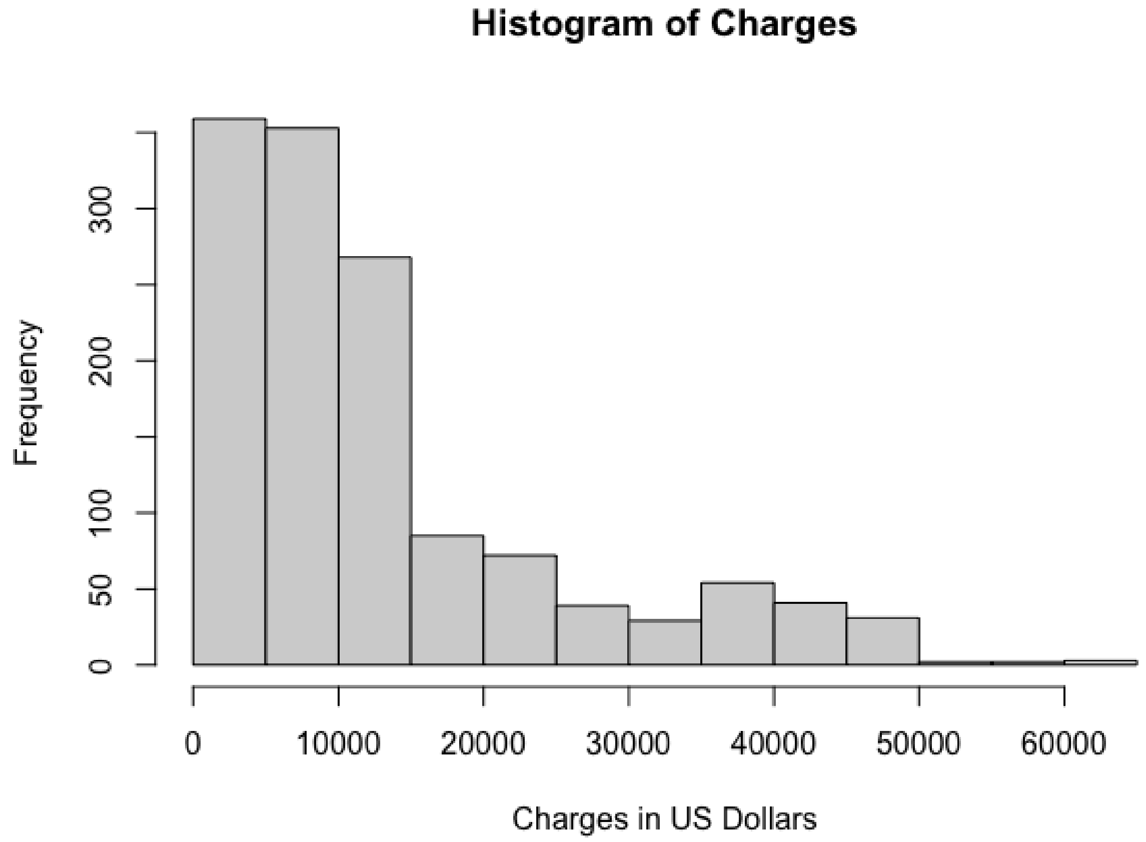
As the only developed country holding a private health insurance system, the United States is one of the largest markets of insurance in the world. According to Insurance Information Institute, “U.S. insurance industry net premiums written totaled $1.4 trillion in 2021, with premiums recorded by property/casualty (P/C) insurers accounting for 53 percent, and premiums by life/annuity insurers accounting for 47 percent”[2]. Thus, accurately forecasting costs incurred by patients is a crucial step for insurance companies.

Therefore, in this study, we want to see if it is possible to estimate the charges billed to an insurance plan each year based on the policy holder’s demographic information. We also want to determine which factors affect the annual expenditure of clients and which do not. We believe based on this study, it would be more helpful for insurance companies to reduce unnecessary work and make better predictions for annual costs and better determine premiums.

The variables measured in the collected dataset include 6 explanatory variables and the response variable. The explanatory variables are based on the insured person’s age, sex, BMI, number of children, smoking status, and region. The response variable is the charges to health insurance. The variable definitions are detailed in the following table.

Table 1: Variables in the dataset

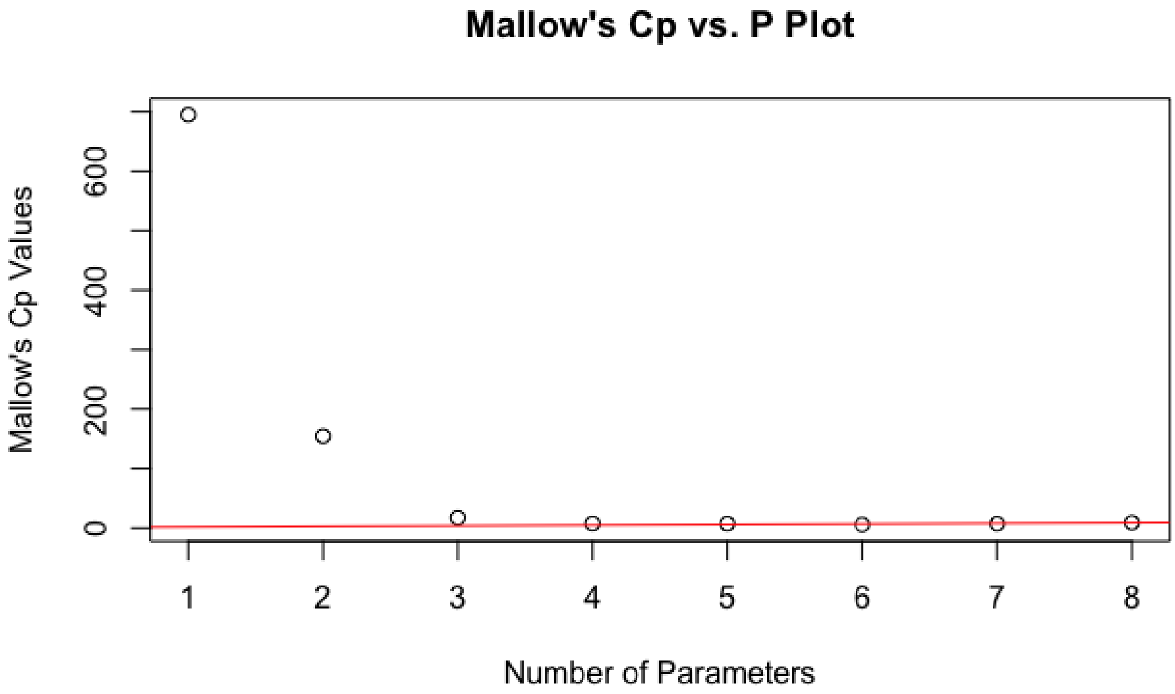
Variable Name	Description	Unit
Age	Age of the insured person	Years
Sex	Sex of the insured person	Male/Female
BMI	Body Mass Index of the insured person	kg/m^2
Smoker	Does the insured person smoke	Yes/No
Region	The insured person’s place of residence in the U.S. is divided into four geographic regions	Northeast/ Southeast/ Northwest/ Southwest
Children	Number of children covered by the insured person	Counts
Charges	Individual medical costs billed to health insurance	US Dollars



The above is a visualization of the dataset, which has a strong left skew. This is likely to influence our ability to predict health costs at the upper end of the spectrum.

2. Analysis

2.1 Model Selection



A Mallows CP Plot indicates that a well fitted model could have between 4 and 8 parameters.

We construct the model using the Backwards Selection algorithm. First, a full model including all six explanatory variables is constructed. In conducting a partial t-test for each coefficient, we can see that the least significant coefficient is that for $sex(X_2)$. Thus, a reduced model without sex as a predictor is constructed.

Repeating, we can see that the coefficients for $region(X_6)$ lack the required significance, and are hence removed.

By refitting without sex and region as predictors, we arrive at a reduced model with each coefficient having the required p-value. The model is:

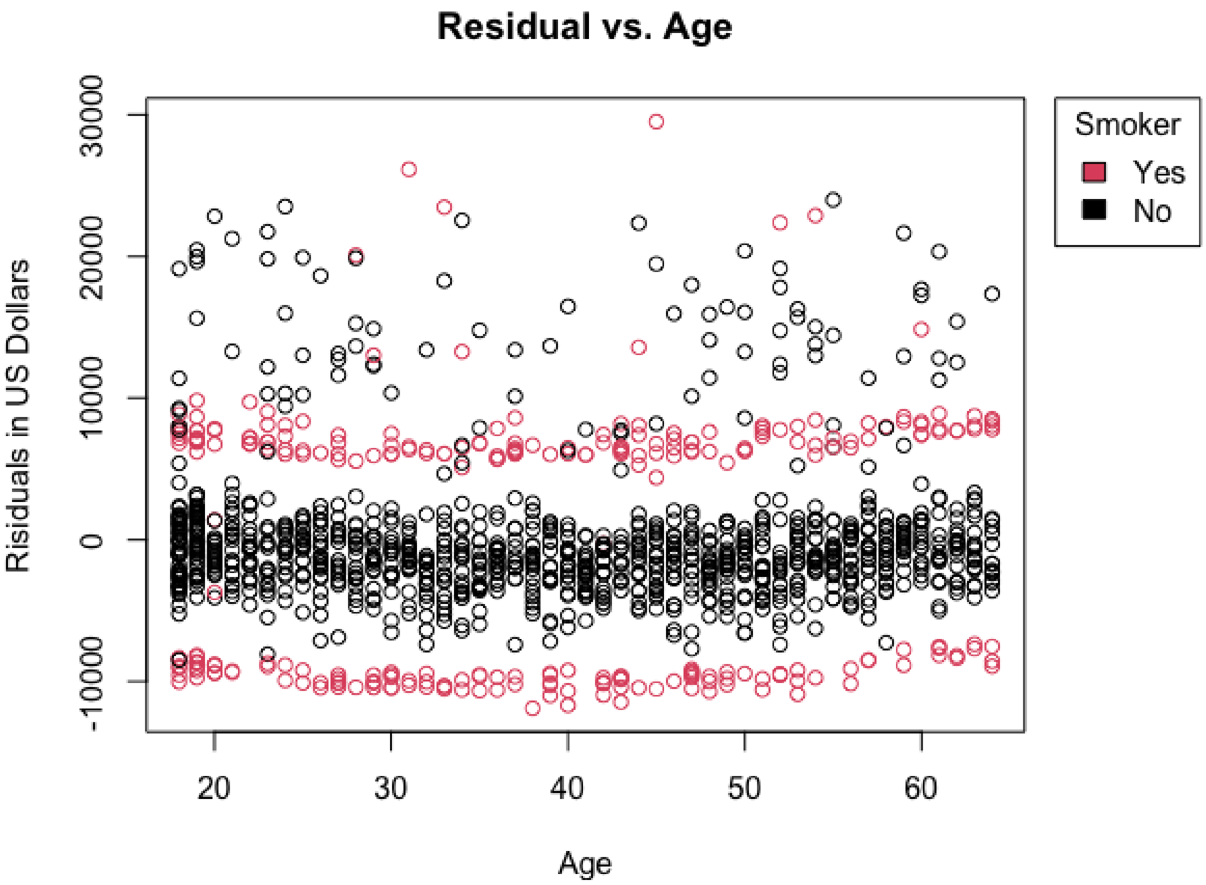
$$\begin{aligned} charges(Y) = -12.848 + 21.675 * age(X_1) + 11.756 * bmi(X_3) + \\ 3.436 * children(X_4) + 57.904 * smoker(X_5) \end{aligned}$$

with an $R^2_{adj} = 0.7489$.

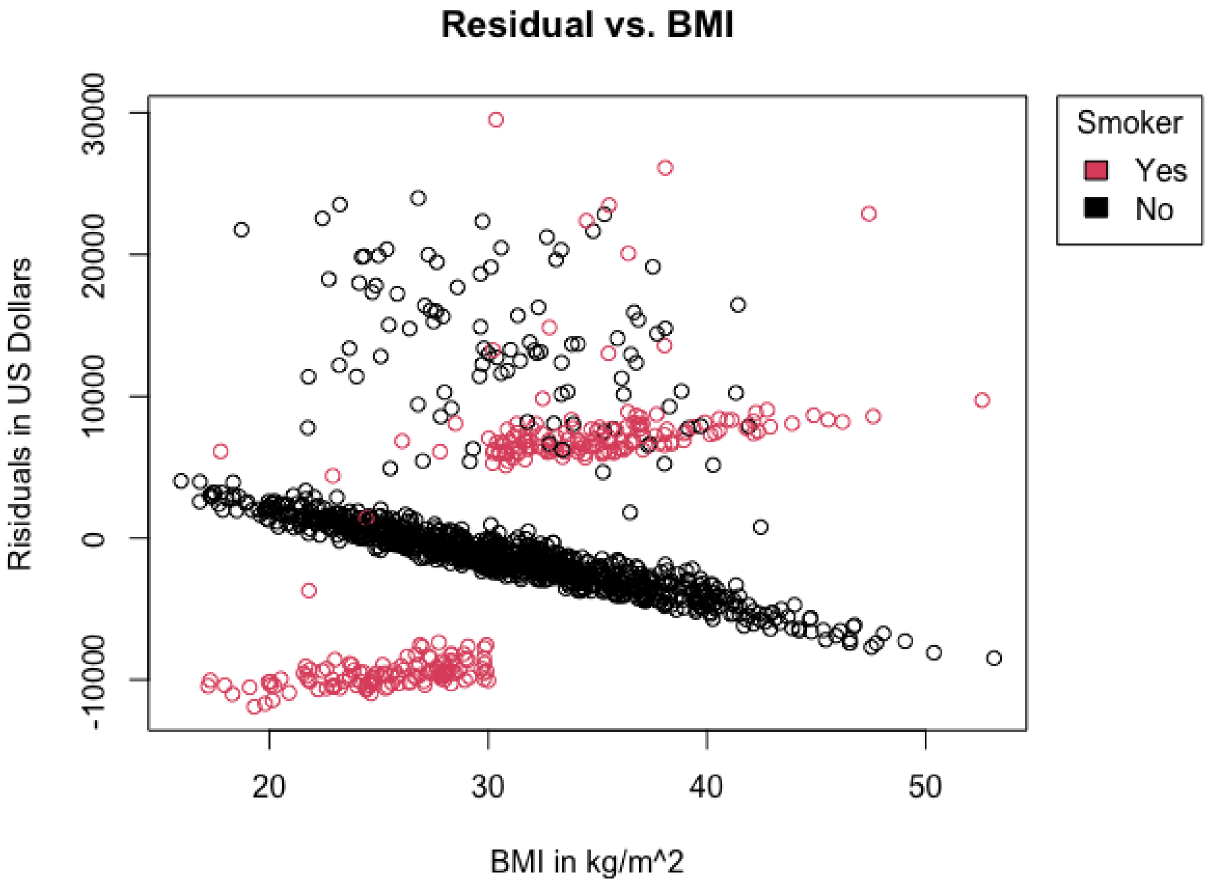
2.2 Model Refinement

With a reasonable reduced model, we can now move to examine the model's legitimacy and refine it to improve R^2_{adj}.

We begin by examining the Residual vs. Fitted Values graph, and immediately are able to notice that the residuals for non-smokers are clustered around zero as we would expect, but the residuals for smokers coalesce in two different regions, each quite far from zero. This suggests that the model is having difficulty accurately predicting the charges of smokers, and possibly indicates there is some interaction at work.



First checking if the interaction is to do with the age of the respondent, we can see that the smokers once again coalesce in 2 different regions, but these regions are spread across the entire range of the age of the respondents. This indicates the interaction we observe is not to do with age.



From the above, we can see a coalescence of smokers around two different areas. There appears to be a split directly along the $BMI = 30$ line. This could suggest that the previously discussed interaction between some predictor and whether or not a respondent is a smoker is to do with BMI.

At a BMI of 30, according to the United States Centre for Disease Control [1], a person is considered obese. Obesity exposes one to significant health risks such as diabetes, heart disease, and stroke, which could significantly impact an individual's medical spending.

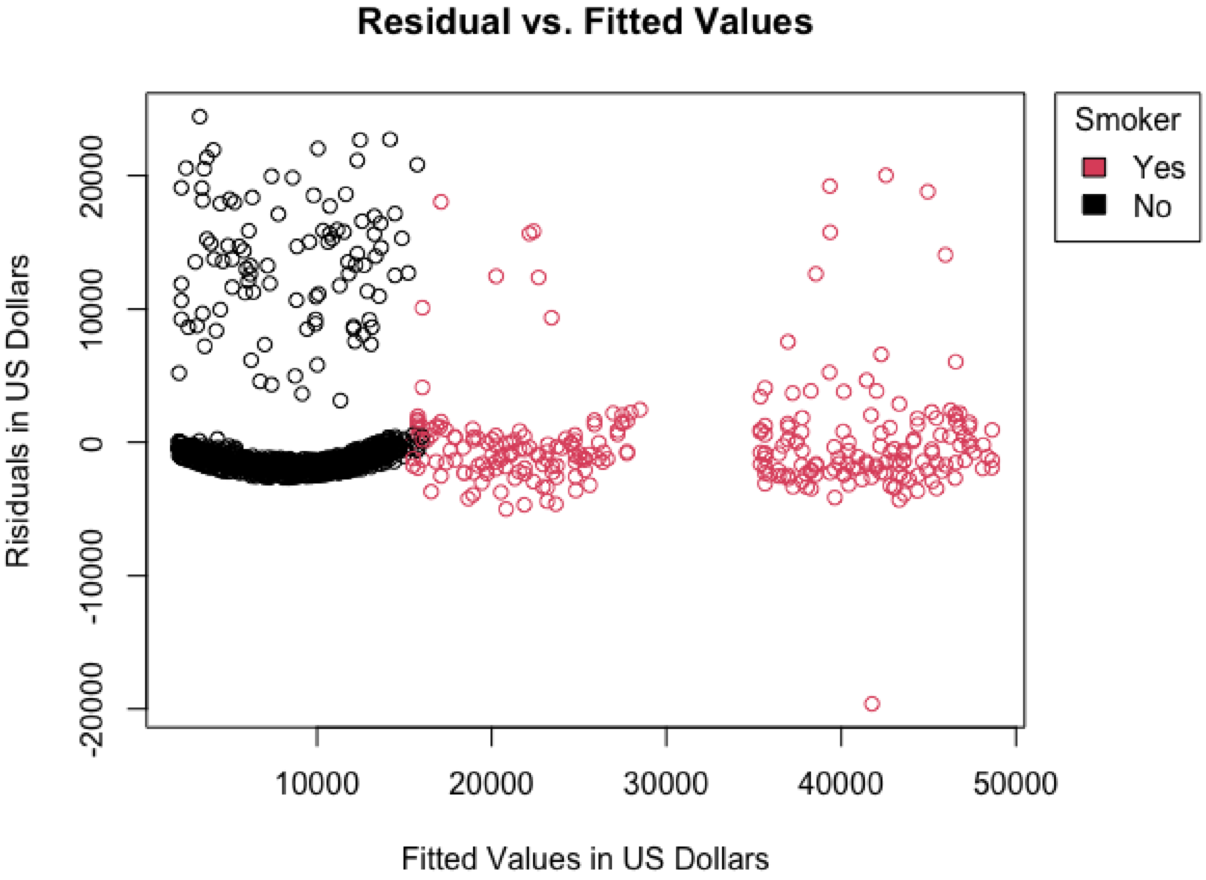
Similarly, a smoker is also exposed to increased risk of heart disease and stroke. It could be that the interaction of obesity and smoking is what causes the trend we observe in the residuals plot above.

In order to examine this interaction, we will transform BMI into a categorical variable, indicating obesity or not. The threshold for obesity will be a BMI >= 30. A new model will be fit, with an interaction between smoking and obesity being examined.

The model is:

$$charges(Y) = -2669.683 + 265.983 * age(X_1) + 514.103 * children(X_4) +$$
$$13389.046 * smoker(X_5) + 128.028 * obesity(X_7) + 19740.494 * smoker(X_5) * obesity(X_7)$$

with an $R^2_{adj} = 0.8606$.

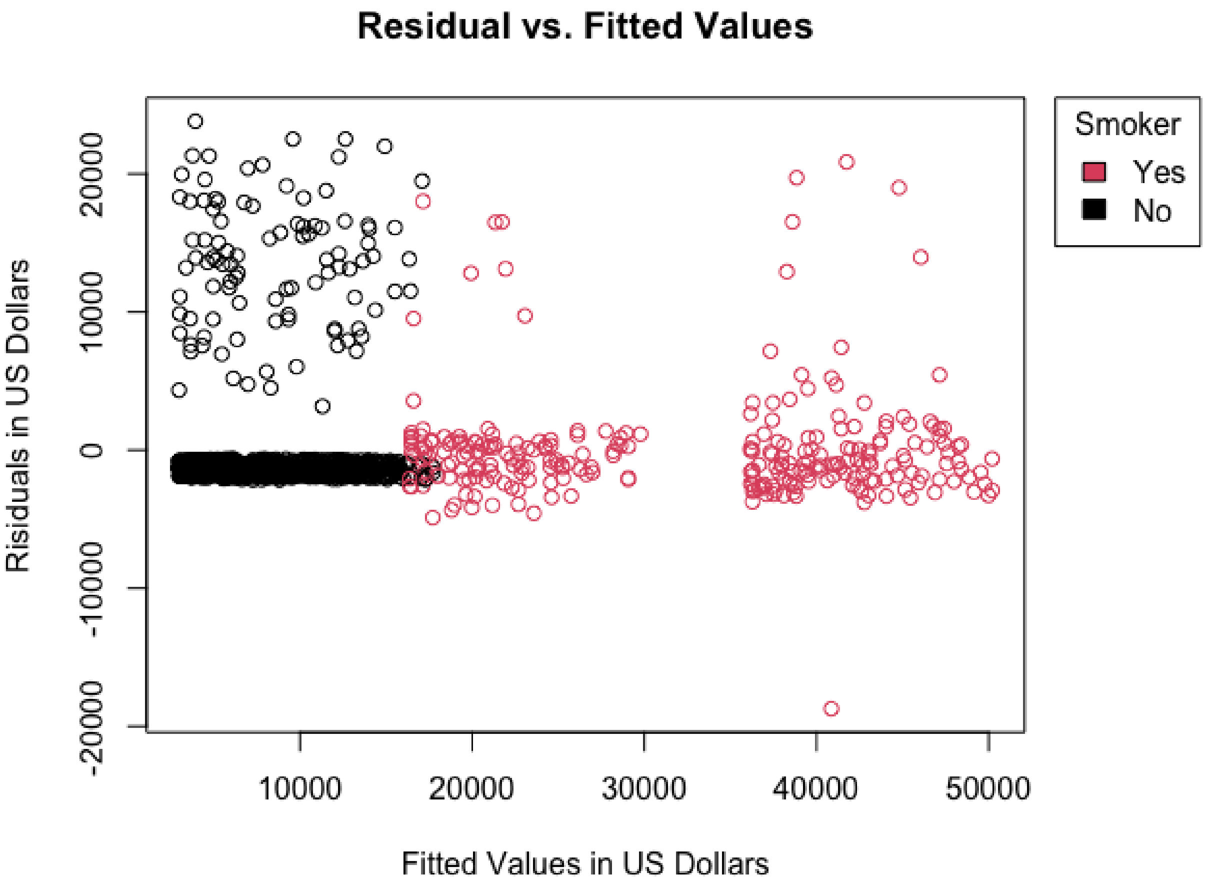


This seems like a good result for a final model, but an examination of the Residual vs Fitted values plot above reveals a curve suggesting some kind of non-linear relationship in the data.

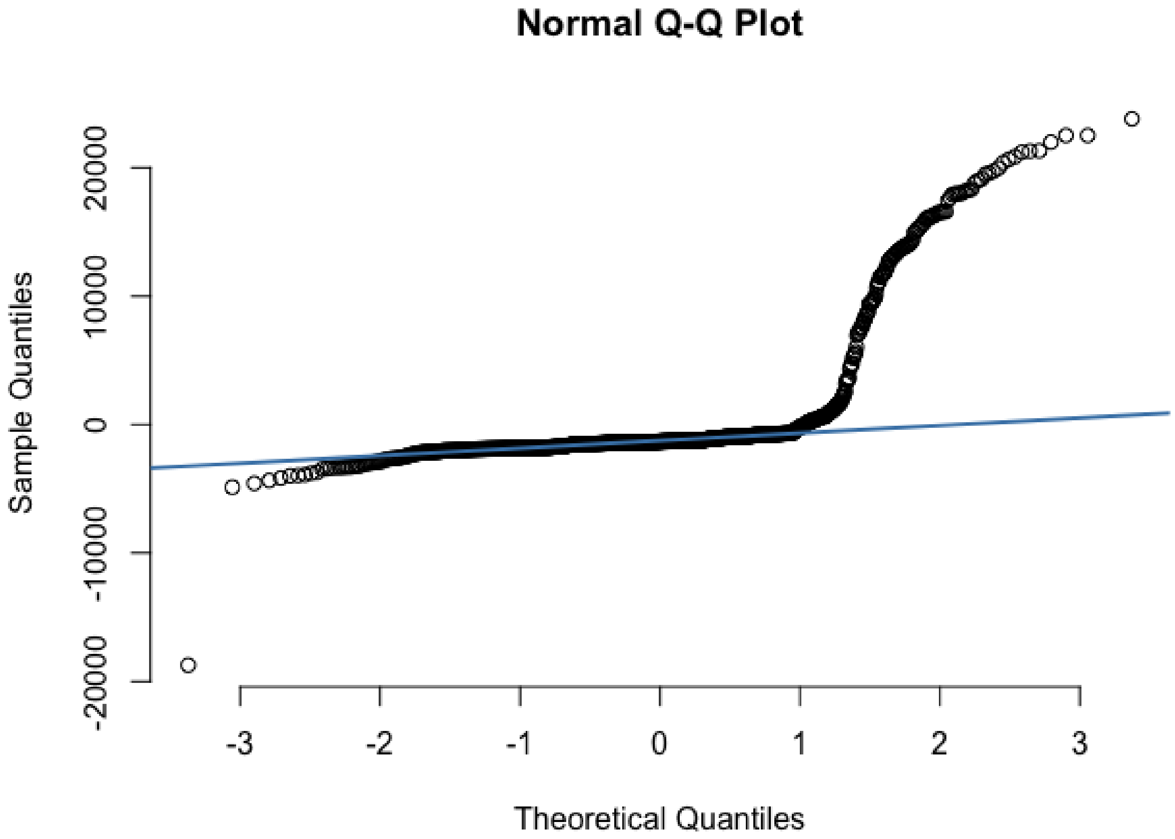
As $age(X_1)$ and $children(X_4)$ are our only continuous variables, we examine what happens when squaring one. Beginning with including an age squared term, we fit the model:

$$\begin{aligned} charges(Y) = 2237.3818 - 24.5114 * age(X_1) + 3.6643 * age^2(X_1^2) + \\ 669.387 * children(X_4) + 13389.6765 * smoker(X_5) + \\ 42.3927 * obesity(X_7) + 19759.1119 * smoker(X_5) * obesity(X_7) \end{aligned}$$

with an $R^2_{adj} = 0.8629$.



Examining the Residuals vs Fitted Values plot of this model shows that we have removed the curve of concern from above, and gives us a plot with the residuals clustered around zero as we would expect.



A normal QQ plot shows a linear relationship, with some tapering near the end as expected due to the left skew of our charges data.

This will be the final model.

3. Conclusion

3.1 Synopsis of Analysis

Using the backward selection algorithm, a reduced model at the 99.9% significance level was established. This model was further refined by examination of residual plots to establish an interaction between smoking and obesity. The final refinement was to introduce a quadratic term in age in order to achieve an $R^2_{adj} = 0.8629$.

3.2 Discussion of Results

It was interesting that the backward selection algorithm removed biological sex as a variable of import. We had initially speculated that perhaps women would incur a higher average medical expense, but it turned out that sex was not an effective predictor.

What was less surprising was the region not being relevant to the model. It seemed unlikely that specific areas of the same country would incur higher costs, and the significance of the location coefficients was relatively low compared to the other coefficients in the full and reduced models.

Overall, we feel that an $R^2_{adj} = 0.8629$ shows that the model is sufficient to accurately predict annual health charges, and could be a useful utility for insurers to determine premiums.

3.3 Final Conclusion

Expected medical charges can be modeled by

$$\begin{aligned} charges(Y) = 2237.3818 - 24.5114 * age(X_1) + 3.6643 * age^2(X_1^2) + \\ 669.387 * children(X_4) + 13389.6765 * smoker(X_5) + \\ 42.3927 * obesity(X_7) + 19759.1119 * smoker(X_5) * obesity(X_7) \end{aligned}$$

with an $R^2_{adj} = 0.8629$.

References

[1] Defining Adult Overweight & Obesity. https://www.cdc.gov/obesity/basics/adult-defining.html#:/~:text=Adult%20Body%20Mass%20Index&text=If%20your%20BMI%20is%20less,falls%20within%20the%20obesity%20range.

[2] Insurance Information Institute Releases Its 2019 a Firm Foundation: How Insurance Supports the Economy. https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:5W71-HM11-DYG2-R48H-00000-00&context=1516831.

Appendix

```
# Load package
library(leaps)

# Read the data (Please adjust the file path when reproducing this process)
insurance <-
  read.csv("insurance.csv", stringsAsFactors = TRUE)

# Explore the response variable
hist(insurance$charges, xlab = "Charges in US Dollars",
     main = "Histogram of Charges")

# Build the full model
reg_full <-
  lm(charges ~ region + smoker + children + bmi + sex + age, data = insurance)
summary(reg_full)

# Model Selection
selection <-
  regsubsets(charges ~ ., data = insurance, method = "exhaustive")
selection_summary <- summary(selection)
selection_summary
plot(selection_summary$cp, xlab = "Number of Parameters",
     ylab = "Marllow's Cp Values", main = "Marllow's Cp vs. P Plot")
abline(c(1:8), c(1:8), col = "red")

# Reduced model without sex and region
reg_NoRegion_NoSex <-
  lm(charges ~ smoker + children + bmi + age, data = insurance)
summary(reg_NoRegion_NoSex)

# Plot residuals vs. fitted values
par(mar = c(4, 4, 4, 6), xpd = TRUE)
plot(
  reg_NoRegion_NoSex$residuals ~ reg_NoRegion_NoSex$fitted.values,
  xlab = "Fitted Values in US Dollars",
  ylab = "Risiduals in US Dollars",
  col = as.factor(insurance$smoker)
)
title("Residual vs. Fitted Values")
legend(
  x = "topright",
  inset = c(-0.2, 0),
  legend = c("Yes", "No"),
  fill = as.factor(insurance$smoker),
  title = "Smoker"
)

# Plot residuals vs. Age and BMI respectively
par(mar = c(4, 4, 4, 6), xpd = TRUE)
plot(
  reg_NoRegion_NoSex$residuals ~ insurance$age,
```

```r
  xlab = "Age",
  ylab = "Risiduals in US Dollars",
  col = as.factor(insurance$smoker)
)
title("Residual vs. Age")
legend(
  x = "topright",
  inset = c(-0.2, 0),
  legend = c("Yes", "No"),
  fill = as.factor(insurance$smoker),
  title = "Smoker"
)
par(mar = c(4, 4, 4, 6), xpd = TRUE)
plot(
  reg_NoRegion_NoSex$residuals ~ insurance$bmi,
  xlab = "BMI in kg/m^2",
  ylab = "Risiduals in US Dollars",
  col = as.factor(insurance$smoker)
)
title("Residual vs. BMI")
legend(
  x = "topright",
  inset = c(-0.2, 0),
  legend = c("Yes", "No"),
  fill = as.factor(insurance$smoker),
  title = "Smoker"
)

# Categorize BMI into "Obese" and "not Obese"
insurance$obesity <- ifelse(insurance$bmi >= 30, 1, 0)

# New model with Obesity and interaction
reg_Obesity_Smoker_Int <-
  lm (charges ~ obesity + children + smoker + age + obesity * smoker,
      insurance)
summary(reg_Obesity_Smoker_Int)

# Residual Plot
par(mar = c(4, 4, 4, 6), xpd = TRUE)
plot(
  reg_Obesity_Smoker_Int$residuals ~ reg_Obesity_Smoker_Int$fitted.values,
  xlab = "Fitted Values in US Dollars",
  ylab = "Risiduals in US Dollars",
  col = as.factor(insurance$smoker)
)
title("Residual vs. Fitted Values")
legend(
  x = "topright",
  inset = c(-0.2, 0),
  legend = c("Yes", "No"),
  fill = as.factor(insurance$smoker),
  title = "Smoker"
)
```

```
# Squared predictor added
reg_Obesity_Smoker_Int_Age2 <-
  lm (charges ~ obesity + children + smoker + age +
        I((age - mean(insurance$age)) ^ 2) + obesity * smoker,
      insurance)
summary(reg_Obesity_Smoker_Int_Age2)

plot(
  reg_Obesity_Smoker_Int_Age2$residuals ~ reg_Obesity_Smoker_Int_Age2$fitted.values,
  xlab = "Fitted Values in US Dollars",
  ylab = "Risiduals in US Dollars",
  col = as.factor(insurance$smoker)
)
title("Residual vs. Fitted Values")
legend(
  x = "topright",
  inset = c(-0.2, 0),
  legend = c("Yes", "No"),
  fill = as.factor(insurance$smoker),
  title = "Smoker"
)

# Q-Q plot of the final model
par(mar = c(5.1, 4.1, 4.1, 2.1), xpd = FALSE)
qqnorm(reg_Obesity_Smoker_Int_Age2$residuals,
       pch = 1,
       frame = FALSE)
qqline(reg_Obesity_Smoker_Int_Age2$residuals,
       col = "steelblue",
       lwd = 2)
```