Medical Cost Personal Dataset

Abstract

For an insurance company to generate profit, the company must make more money in yearly premiums than the expenses from medical costs to its beneficiaries. The goal of the insurance company is to forecast medical expenses accurately. Using the results of the forecast, the insurance company tailors the yearly premiums according to the insurance company's profit goals. The "Medical Cost Personal Dataset" was created using demographic statistics from the United States Census Bureau, and it serves as the foundation for this analysis. The study uses regression methods to predict medical expenses based on factors like age, BMI, and smoking habits. The group's goal is to predict medical expenses for insurance beneficiaries by refining the model through variable selection and analysis to ensure accurate and meaningful results.

Introduction

Healthcare costs have become a significant concern in recent years, with factors such as age, lifestyle choices, and regional differences playing a critical role in determining individual medical expenses. Understanding the relationship between these predictors and insurance charges can aid in developing predictive models that are not only valuable for insurance companies but also for individuals seeking to make informed healthcare decisions. This study aims to explore the relationships between various demographic, behavioral, and regional factors with medical charges. We aim to create predictive models that accurately estimate individual healthcare costs using multiple linear regression. The objectives include identifying key predictors of medical expenses and assessing the performance of different transformations and interactions of predictors. The dataset includes the following features:

- **Age:** The age of the primary beneficiary.
- **Sex:** The gender of the insurance contractor, categorized as male or female.
- **BMI (Body Mass Index):** An objective index of body weight calculated as weight (kg) divided by height (m^2). It provides insight into whether an individual has a healthy body weight (ideal range: 18.5-24.9).
- **Children:** The number of dependents covered by health insurance.
- **Smoker:** A binary variable indicating whether the individual is a smoker.
- **Region:** The residential area of the beneficiary in the United States, is divided into four **r**egions (northeast, southeast, southwest, and northwest).
- **Charges:** The target variable represents individual medical costs billed by health insurance.

This analysis aims to address the problem of predicting healthcare costs by employing linear regression methods. The findings will provide valuable insights into healthcare planning and cost management, with a specific focus on the utility of linear regression models for these predictions. In addition, the results will show how the regression model can identify factors that influence the costs, helping insurers set better prices.

Methodology

This study employs multiple linear regression as the primary method for analyzing the relationships between predictors and healthcare costs. The methodology is guided by Data preprocessing, fundamental regression theory and concepts, ensuring the robustness and interpretability of the resulting models.

1. Data preprocessing

The dataset contained no missing values but included a categorical column that required one-hot encoding. This process converts categorical values into separate binary columns for each category. Additionally, the "smoker" and "gender" columns were transformed by mapping "yes" and "no" as well “male” and “female” to 1 and 0, respectively, to facilitate easier data analysis.

2. Multiple Linear Regression
Multiple linear regression is used to model the relationship between a response variable (medical charges) and multiple predictor variables (age, sex, BMI, children, smoker status, and region).

The general form of the regression equation is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

3. Assumption Testing
To ensure the validity of the regression model, the following assumptions will be checked:

- **Linearity:** The relationship between predictors and the response variable is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** The variance of residuals is constant across all levels of the predictors.
- **Normality:** Residuals follow a normal distribution.

These assumptions will be verified using a diagnostic plot.

4. Variable Selection Methods
The study employs stepwise regression to determine the optimal combination of predictors for the best model fit. This process involves assessing each step of adding or removing independent (direction both) variables and selecting the best model based on the following metrics:

- **Adjusted R^2:** identifies models that balance explanatory power and complexity.
- **Akaike Information Criterion (AIC):** evaluates model quality, with lower values indicating better fit while penalizing overfitting.

Out of these, we will select the model that minimizes AIC the most while maximizing Adjusted R^2.

5. Model Refinement
To enhance model performance and better meet the assumptions of linear regression, we will apply transformations (such as logarithmic and square root transformations) and introduce interaction terms. These adjustments are intended to enhance the model's R-squared and adjusted R-squared values, ultimately leading to more reliable performance metrics.

6. Model Performance Metrics
The performance of the regression models will be assessed using the following:

- **Residual Mean Squared Error (MSE):** Measures the average squared difference between observed and predicted values. We'll be able to compare our predicted outcome (the response v.) to the observed outcome and determine the accuracy.
- **Prediction Error:** Computes the prediction error by splitting the data into training, testing the dataset, and checking the prediction error which is done by comparing the actual vs the predicted.

allow us to build a model that not only fits the data more effectively but also enhances its predictive power.

Data Analysis

The goal of the “Personal Medical Costs Dataset” dataset is to accurately predict the medical expenses an insurance company covers its beneficiaries. The group will predict the medical costs the insurance company covers the beneficiaries. Using the “Personal Medical Costs Dataset” provided by Brett Lantz, the group’s objective is to remove unnecessary predictors using Piecewise Selection, analyze the resultant model, make the needed changes the model may need, and test the data.

“Personal Medical Costs Dataset” consists of six predictors. The predictors are: Age, Sex, BMI, Children, Smoker, and Region. First, we called the summary function (see Figure 1). In the summary, we observe that the adjusted R squared is 0.76 and that there exists multicollinearity.Upon examining the diagnostics, we immediately notice that the residuals are non-normal, the residual vs. fitted plot shows a pattern, and the red line in the scale-location plot is not straight.

```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    regionsouthwest + regionsoutheast + regionnortheast, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11582.8  -2989.4   -987.2   1585.1  24737.3

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -12745.97    1125.22 -11.328  < 2e-16 ***
age                269.12      13.55  19.854  < 2e-16 ***
sex               -268.59     378.01  -0.711  0.47754
bmi                344.61      32.72  10.532  < 2e-16 ***
children           492.42     158.60   3.105  0.00195 **
smoker           24180.64     459.77  52.593  < 2e-16 ***
regionsouthwest   -951.29     539.32  -1.764  0.07805 .
regionsoutheast   -909.14     538.80  -1.687  0.09183 .
regionnortheast    436.03     539.34   0.808  0.41902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6156 on 1061 degrees of freedom
Multiple R-squared:  0.7587,    Adjusted R-squared:  0.7569
F-statistic: 416.9 on 8 and 1061 DF,  p-value: < 2.2e-16
```

Figure 1

The diagnostics are as follows:

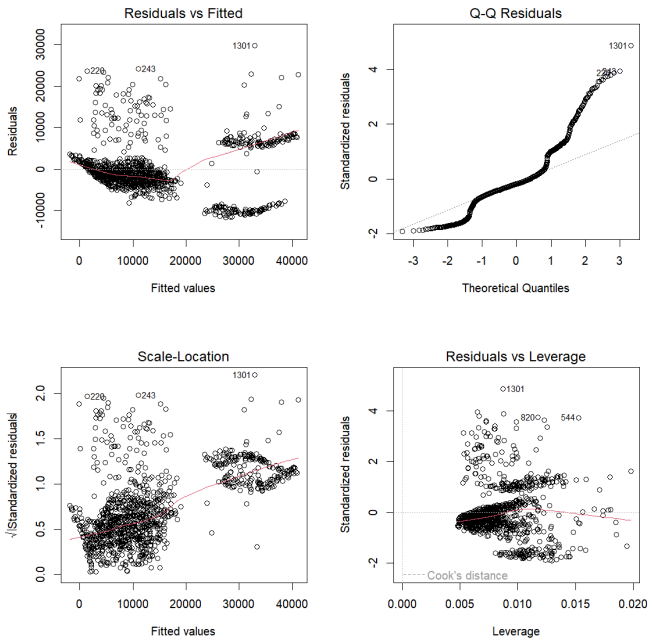

Figure 2

Both AIC and R-Adj yielded the same result. Thus we can be confident the model includes the 6 predictors shown in Figure 3 and Figure 4. Following the variable selection process, the features removed from the model were sex and regionsoutheast.

```
> SignifReg(full_model, alpha = 0.05, direction = "both", criterion = "r-adj")

Call:
lm(formula = charges ~ age + bmi + children + smoker + regionsouthwest +
    regionsoutheast, data = train)

Coefficients:
    (Intercept)              age              bmi         children
       -12616.6            269.2            342.9            487.4
         smoker  regionsouthwest  regionsoutheast
        24169.5          -1159.8          -1114.6

> SignifReg(full_model, alpha = 0.05, direction = "both", criterion = "AIC")

Call:
lm(formula = charges ~ age + bmi + children + smoker + regionsouthwest +
    regionsoutheast, data = train)

Coefficients:
    (Intercept)              age              bmi         children
       -12616.6            269.2            342.9            487.4
         smoker  regionsouthwest  regionsoutheast
        24169.5          -1159.8          -1114.6
```

Figure 3

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + regionsouthwest +
    regionsoutheast, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-11420  -3036  -1001   1588  25122

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -12616.61    1078.86 -11.694  < 2e-16 ***
age               269.18      13.55  19.870  < 2e-16 ***
bmi               342.95      32.66  10.500  < 2e-16 ***
children          487.39     158.46   3.076  0.00215 **
smoker          24169.50     458.12  52.759  < 2e-16 ***
regionsouthwest -1159.83     472.65  -2.454  0.01429 *
regionsoutheast -1114.58     472.14  -2.361  0.01842 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6154 on 1063 degrees of freedom
Multiple R-squared:  0.7584,    Adjusted R-squared:  0.757
F-statistic: 556.2 on 6 and 1063 DF,  p-value: < 2.2e-16
```

Figure 4

Transformations

Our next goal was to try to increase the adjusted R-square as that metric was the most important for us before we started looking at the model performance metrics. To this, we did a transformation as shown in Figure 5 with sqrt which increased the adjusted R-squared.

```
Call:
lm(formula = rootcharges ~ age + bmi + children + smoker + regionsouthwest +
    regionsoutheast, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-39.141 -10.805  -4.593   3.173 107.745

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -2.0150     3.8511  -0.523  0.60092
age               1.3991     0.0481  29.088  < 2e-16 ***
bmi               1.0367     0.1168   8.875  < 2e-16 ***
children          3.1706     0.5585   5.677 1.76e-08 ***
smoker           91.0117     1.6969  53.634  < 2e-16 ***
regionsouthwest  -3.0809     1.6788  -1.835  0.06676 .
regionsoutheast  -4.9967     1.6938  -2.950  0.00325 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.05 on 1063 degrees of freedom
Multiple R-squared:  0.7852,    Adjusted R-squared:  0.7839
F-statistic: 647.5 on 6 and 1063 DF,  p-value: < 2.2e-16
```

Figure 5

We tried to square the response but this just decreased the adjusted R^2 as shown in figure 6.

```
> model_2 <- lm(squarecharges ~ age + bmi + children + smoker+regionsouthwest+regionsoutheast,
= train)
> summary(model_2)

Call:
lm(formula = squarecharges ~ age + bmi + children + smoker +
    regionsouthwest + regionsoutheast, data = train)

Residuals:
       Min         1Q     Median         3Q        Max
-709680013 -140115930  -14028804  116716992 2252898351

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -765030388   56847844 -13.458   <2e-16 ***
age                6525275     709989   9.191   <2e-16 ***
bmi               20264669    1724233  11.753   <2e-16 ***
children           6763591    8243628   0.820    0.412
smoker          1056268172   25048673  42.169   <2e-16 ***
regionsouthwest  -22437641   24781494  -0.905    0.365
regionsoutheast  -38586164   25002429  -1.543    0.123
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325500000 on 1063 degrees of freedom
Multiple R-squared:  0.6572,    Adjusted R-squared:  0.6553
F-statistic: 339.7 on 6 and 1063 DF,  p-value: < 2.2e-16
```

Figure 6

Observe that with more transformations we made, the R-Adj decreased; we ceased the transformations and contemplated our predictors. Note that we have the predictor “Smoker” in the reduced model. So we made interactions using “Smoker” and other predictors. In doing so, the interactions increased the Adjusted R^2 (see Figure 7).

```
Call:
lm(formula = log_charges ~ age * smoker + bmi * smoker + children +
    regionsouthwest + regionsoutheast, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.65084 -0.15658 -0.07010  0.01874  2.18021

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.0956803  0.0749562  94.664  < 2e-16 ***
age              0.0428293  0.0009433  45.406  < 2e-16 ***
smoker           1.2717407  0.1561692   8.143 1.07e-15 ***
bmi             -0.0015254  0.0022823  -0.668    0.504
children         0.1081880  0.0097987  11.041  < 2e-16 ***
regionsouthwest -0.1370005  0.0292401  -4.685 3.16e-06 ***
regionsoutheast -0.1384901  0.0291956  -4.744 2.39e-06 ***
age:smoker      -0.0349152  0.0020487 -17.043  < 2e-16 ***
smoker:bmi       0.0531316  0.0045315  11.725  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3803 on 1061 degrees of freedom
Multiple R-squared:  0.8367,    Adjusted R-squared:  0.8355
F-statistic: 679.7 on 8 and 1061 DF,  p-value: < 2.2e-16
```

Figure 7

Final transformation

The best model we obtained was achieved through the addition of interaction terms and logarithmic transformations, as shown in Figure 7. The final model is represented by the following formula: log_charges ~ age * smoker + bmi * smoker + children + regionsouthwest + regionsoutheast. The diagnostic as shown in Figure 8 could use more improvements as the obvious sign of non-normality, and the residual vs fitted seems to contain a pattern. We still picked this model, do to the fact it gave us the highest R-adj score as well as lowest prediction error.

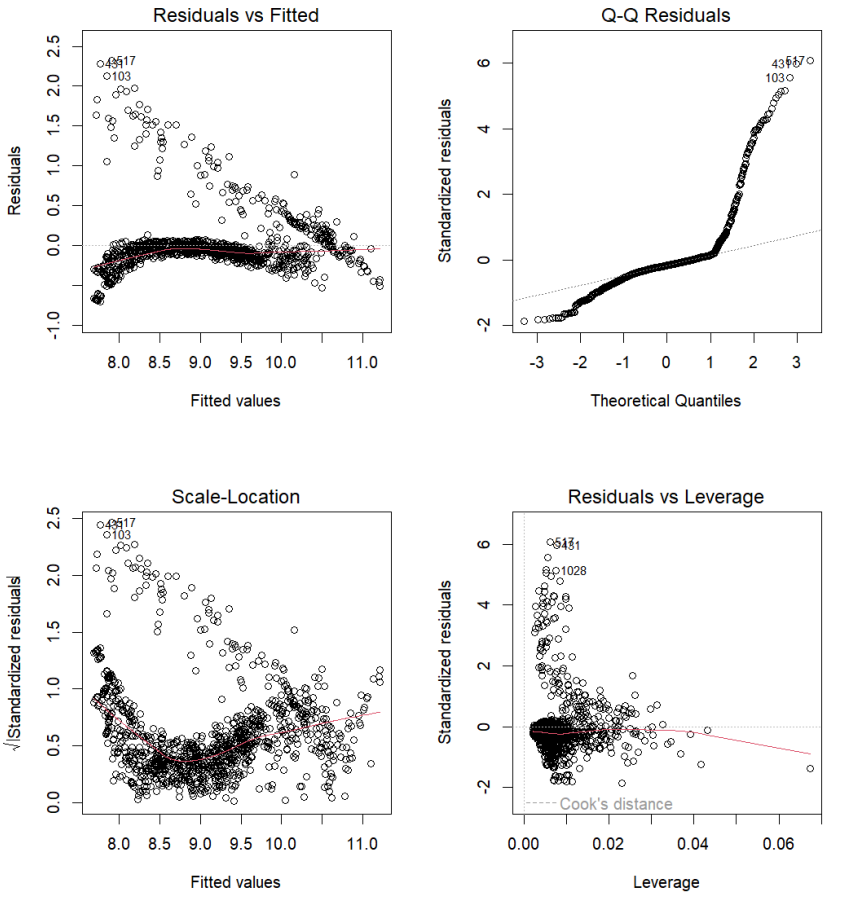

Figure 8

Overall, we achieved a final RMSE score of 5564.906, indicating that most of the variation was relatively close to the actual insurance charges. We notice the diagnostics have improved, however, there is still room for further improvement throughout the project. In the Residuals vs Fitted, there is a curve pattern in the residuals indicating non-linearity. The normality plot shows it is not perfectly normal especially in the upper extreme tail. Scale-Location suggests heteroscedasticity, and the variance errors increase with the predicted values. Some bad leverage points are still visible. We got an average of 19.89% for the percentage error.

Appendix

```
#library load
library(SignifReg)
library(caTools)
library(caret)
library(car)

#change plot to look 2x2
par(mfrow = c(2, 2))

#loading data
data<- read.csv("insurance.csv")

#Convert Yes to 1 and No to 0
data$smoker <- ifelse(data$smoker == "yes", 1, 0)

#Convery Male to 1 and Female to 0
data$sex <- ifelse(data$sex == "male", 1, 0)

#one hot encoding for region column
dummy <- dummyVars(~ region, data = data)
encoded_region <- predict(dummy, newdata = data)
encoded_region <- as.data.frame(encoded_region)

# Combine the encoded region data with the rest of the original data (excluding 'region' column)
final_data <- cbind(data[, setdiff(names(data), "region")], encoded_region)

#split the data into train (80%) and test (20%)
sample <- sample.split(data$charges, SplitRatio = 0.8)
train  <- subset(final_data, sample == TRUE)
test   <- subset(final_data, sample == FALSE)

#make a full model, then do variable selection with SignifReg w/ criterion (AIC and r-adj)
full_model <- lm(charges ~ age + sex + bmi + children + smoker + regionsouthwest+regionsoutheast
+regionnortheast , data = train)
summary(full_model)
SignifReg(full_model, alpha = 0.05, direction = "both", criterion = "r-adj")
SignifReg(full_model, alpha = 0.05, direction = "both", criterion = "AIC")
plot(full_model)


#compare the plotsand summary they both outputted
```

```
new_model <- lm(charges ~ age + bmi + children + smoker+regionsouthwest+regionsoutheast, data =
train)
plot(new_model)
summary(new_model)

#vif of models
vif(new_model)

#Starting transformations
squarecharges <- train$charges ^ 2
squareage <- train$age ^ 2
squarebmi <- train$bmi ^ 2
squarechildren <- train$children ^ 2
lncharges <- log(train$charges)
lnage <- log(train$age)
lnbmi <- log(train$bmi)
rootcharges <- sqrt(train$charges)
rootage <- sqrt(train$age)
rootbmi <- sqrt(train$bmi)
rootchildren <- sqrt(train$children)
cubedcharges <- train$charges ^ 3
cubedage <- train$age ^ 3
cubedbmi <- train$bmi ^ 3
cubedchildren <- train$children ^ 3
trainbmi30 = ifelse(train$bmi >= 30, 1, 0)



#transformation
agemodel <- lm(charges ~ squareage + bmi + children + smoker +
trainbmi30:smoker+regionsouthwest+regionsoutheast, data = train)
model_1 <- lm(rootcharges ~ age + bmi + children + smoker+regionsouthwest+regionsoutheast, data =
train)
model_2 <- lm(squarecharges ~ age + bmi + children + smoker+regionsouthwest+regionsoutheast, data =
train)
model_3 <- lm(charges ~ age + bmi + children + smoker +
trainbmi30:smoker+regionsouthwest+regionsoutheast, data = train)
model01 <- lm(lncharges ~ age + bmi + smoker + children+regionsouthwest+regionsoutheast, data =
train)
model02 <- lm(lncharges ~ lnage + bmi + smoker + children+regionsouthwest+regionsoutheast, data =
train)
model03 <- lm(cubedcharges ~ age + bmi + smoker + children+regionsouthwest+regionsoutheast, data =
train)
```

```
model04 <- lm(cubedcharges ~ age + trainbmi30 + smoker + children+regionsouthwest+regionsoutheast,
data = train)
model05 <- lm(cubedcharges ~ age + bmi + trainbmi30 + smoker +
children+regionsouthwest+regionsoutheast, data = train)
model06 <- lm(cubedcharges ~ age + bmi + trainbmi30 + smoker + bmi:smoker +
children+regionsouthwest+regionsoutheast, data = train)
model07 <- lm(cubedcharges ~ lnage + bmi + trainbmi30 + smoker + bmi:smoker +
children+regionsouthwest+regionsoutheast, data = train)
model08 <- lm(rootcharges ~ lnage + bmi + trainbmi30 + smoker + bmi:smoker +
children+regionsouthwest+regionsoutheast, data = train) # ***
model09 <- lm(rootcharges ~ lnage + lnbmi + trainbmi30 + smoker + trainbmi30:smoker + children +
smoker:age+regionsouthwest+regionsoutheast, data = train)
model10 <- lm(cubedcharges ~ lnage + lnbmi + trainbmi30 + children + trainbmi30:smoker +
lnage:smoker + bmi*smoker + lnage*lnbmi + smoker+regionsouthwest+regionsoutheast, data = train)
model11 <- lm(cubedcharges ~ cubedchildren + trainbmi30*smoker + lnage:smoker +
lnage*bmi+regionsouthwest+regionsoutheast, data = train)
model12 <- lm(cubedcharges ~ cubedchildren + trainbmi30*smoker + lnage:smoker +
lnage*trainbmi30+regionsouthwest+regionsoutheast, data = train)


#SKIP THE BELOW AS FOR OTHER MODELS
#########################################################################
# Predict house prices for the test dataset
#predicted_cost <- predict(model_7, newdata = test)
#predicted_cost2 <- predict(model_8, newdata = test)
#test$bmi30 <- test$bmi * (test$smoker == "yes")#add to test
#predicted_cost3 <- predict(project2_model, newdata = test)
#test$squareage <- test$age^2 #add to test
#predicted_cost5 <- predict(trans_interaction05, newdata = test)
#Only include if need bmi30 on test: test$bmi30 = ifelse(test$bmi >= 30, 1, 0)
#Reversing transformations
#predicted_costs_reverse <- (predicted_cost)^2
#predicted_costs2_reverse <- (predicted_cost2)^1/2
#########################################################################



#best model based on r-adj score
train$log_charges <- log(train$charges)
test$log_charges <- log(test$charges)
improved_model <- lm(log_charges ~ age * smoker + bmi * smoker +
children+regionsouthwest+regionsoutheast, data = train)
predicted_improve <- predict(improved_model, newdata = test)
reverse_improve <- exp(predicted_improve)
```

```
# Check diagnostics for the improved model
summary(improved_model)
plot(improved_model)


#Computing prediction errors
errors <- test$charges - reverse_improve
mse <- mean(errors^2)
rmse <- sqrt(mse)

precentage_errors<- (abs(errors)/test$charges)*100
mape<-mean(precentage_errors)

cat("MSE:", mse, "\n")
cat("RMSE:", rmse, "\n")
cat("MAPE:", mape, "\n")
```