Project: Predict Medical Charges Using Multiple Linear Regression Method by

Shuangyan Wu

Contents

List of Figures

List of Tables

Summary

In the medical system, medical charges and insurance costs are related for both insurance companies and insured individuals. Knowing the possible healthcare expense of individuals in advance will help the insurers and stack holders for better business plans and healthcare resource allocation. It will also benefit patients since they can choose a better insurance plan according to their medical expenses. In this project, multiple regression models were built to predict medical charges of individuals with 6 response variables including age, bmi, number of children, smoker (yes/no), sex (female/male), and residential region (NW, NE, SW, SE). The data were divided into training and test sets, the models were trained on the training data, and the model performance was evaluated on both the test data and the cross-validation (CV) folds based on the R^2 values and root mean square error (RMSE). The age, smoker, children, and region variables were found to significantly affect the medical charges at $\alpha = 0.05$. In addition, the smoker * bmi interaction effect turned out to be significant too. The final model contains main and 2-factor interaction effects as shown in Equation (1). The effects of different variables can be obtained from the model. For example, it is seen that increasing age by 1 year is predicted to increase the medical charge by 242 dollars, and people who live in the southwest are predicted to have 1224 dollars less medical charges than people in the northeast. Overall, the model has $R^2 = 0.83$, i.e., 83% of sample variance explained. The model performance is constant throughout different CV folds with a small range of root mean square error (4197 - 5501 dollars).

$$\hat{charges} = -817 + 242 \times age + 1 \times bmi + 471 \times children - 22743 \times I_sy - 717 \times I_rnw \quad (1)$$
$$-1189 \times I_rse - 1224 \times I_rsw + 1518 \times I_sy \times bmi$$

Where I_sy is smoking, I_rnw,I_rse, and I_rsw are the indicators that the individuals are from the NW, SE, and SW of the US.

1 Introduction

In the health care system, accurately predicting the medical cost of individuals can not only help insurers determine insurance costs but also guide patients in choosing proper insurance plans knowing their predicted medical expenses. It is highly useful for properly advocating limited healthcare resources. Therefore, the objective of this study is to build prediction models for the healthcare costs of individuals with provided variables (**age**, **sex**, **bmi**, **children**, **smoker**, and **region**).

2 Exploratory data analysis

2.1 Data summary

The provided data includes 6 explanatory variables, 1 response variable, and 1338 observations. Three of the explanatory variables are continuous, i.e., the insured individual's age (age), body mass index (bmi), number of dependents (children), and medical cost ($) during the year billed to health insurance (charges). The other 3 explanatory variables are categorical, i.e., whether the person smokes (smoker), the person's US residential area (region), and the gender (sex). In detail, the smoker variable has two levels: yes or no; the region variable has shown 5 levels: Northeast (NE), Northwest (NW), Southeast (SE), Southwest (SW), and #1!a. And the sex variable has two levels: female and male.

Table 1 shows the statistical summary for the continuous variables. The sample age ranges from 18 - 64 years old, average 18 years old. The sample bmi is in the range of 15.96-53.13 with a mean of 30.66. It is known the ideal range of bmi is 18.5-24.9. The charge range for the individuals is 1122 - 63770 dollars with a mean of 13270 dollars. In terms of children numbers, the sampled data shows the range of -2 to 5 children, which is unreasonable. 31 of the observations have negative children numbers, which are likely due to collection mistakes.

Table 2 lists the counts of each level for the categorical variables. It is seen that there are much more smokers (1064) than non-smokers (274) samples in the data. The sample numbers are quite even at different levels of region and sex. There are only 24 observations showing region #1!a, it is unlikely to have a region called or labeled as #1!a. It is possible that this level is labeled mistakenly. More information about the distribution of sample age, bmi, charges, and children is provided in the appendix (Figures 3-6).

Table 1: Summary for continuous variables

variable	age	bmi	children	charge
n	1184	1338	1338	1338
mean	39	30.66	1	13270
median	39	30.40	1	9382
min	18	15.96	-2	1122
max	64	53.13	5	63770
SD	14	6.10	1	12110

2.2 Data visualization and cleaning

After exploring plots with different variables from the data, some interesting patterns were found amongst the charges, age, bmi, and smoker variables. Figure 1 (left) shows the scatter plot of charges with age and the combined variable bmic from smoker and bmi. The bmic

Table 2: Frequency count for categorical variables.

smoker	region	sex
1064(Yes)	317(NE)	659(female)
274(No)	317(NW)	666(male)
-	349(SE)	-
-	320(SW)	-
-	24(#!a)	-

variable is coded with fours levels: bmi in the ideal range (<= 30) & not smoke (in-no), bmi in the ideal range & smokes (in-yes), bmi out of the ideal range (> 30) & not smoke (out-no), bmi out of the ideal range & smokes (out-yes). It is seen age tends to increase an individual's medical cost in three linear trends. The insured individuals who smoke and have bmi out of the ideal range (black dots) tend to have the highest medical bills (the highest band), and patients who do not smoke have relatively low medical bills (green and red dots in two lower bands). Both smoker and bmi variables may contribute to the charges leading to the median range of charges for people with an ideal bmi and smoking habit (blue dots). Also, other factors such as region and children may affect charges, which may contribute to the formation of the middle scatter band. From Figure 1 (right), some interaction between smoker and bmi variables may be spotted for predicting charges. In non-smoking samples (red dots), there is no obvious trend with increasing bmi. While in smoking samples. A linear increasing trend can be seen for the charges with increasing bmi. More plots of charges with explanatory variables and plots between explanatory variables can be found in the appendix (Figures 7-10).

Before analyzing the data, data cleaning needs to be completed. First, the negative children numbers are changed to the corresponding positive numbers. Second, the region labeled as $\#1!a$ is remarked as missing data (NA). Third, there are 11.5%, 1%, and 2.6% of the data missing for the age, sex, and region variables respectively, as shown in Table 3. More information about the numbers and distribution of missing data in the three variables can be found in the appendix (Figures 11-12). To maximize the application of provided information, deleting data rows with missing entries is not recommended. Therefore, here, the age missing values were imputed using the predictive mean matching (*pmm*) function in the R MICE package, which imputes the missing data points with non-missing data points that has the nearest predicted age value to the predicted missing one using linear regression. The missing data for sex and region was imputed with their corresponding modes.

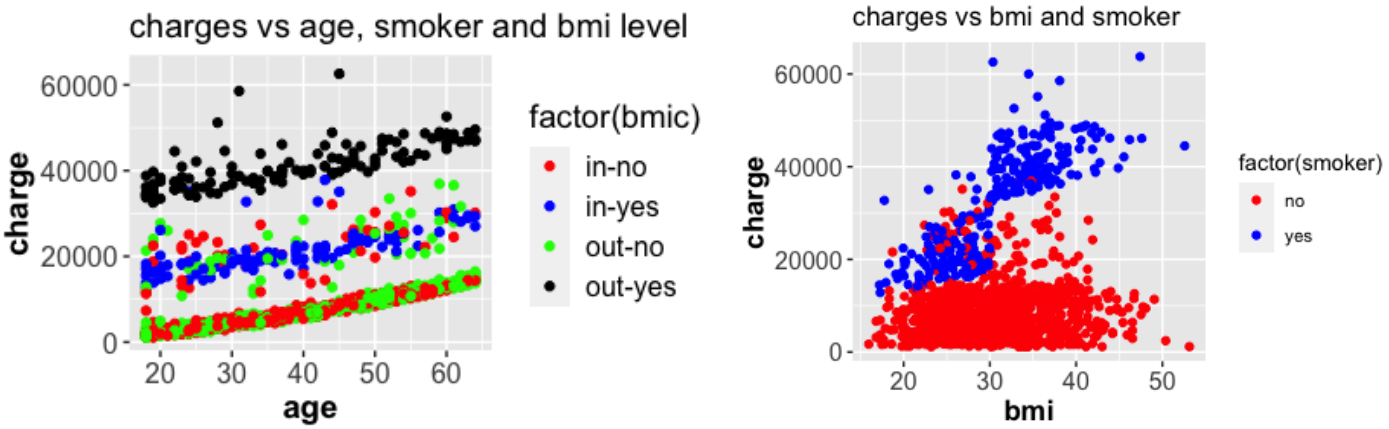

Figure 1: Plots of charges with age, bmi, and smoker.

Table 3: Missing data percentages.

age	sex	bmi	children	smoker	region	charges
11.5	1.0	0	0	0	2.6	0

3 Statistical methods

To build prediction models for the charges with provided variables, multiple regression methods are used in this analysis with both continuous and categorical variables. Main and interaction effects are considered. The complete data was randomly divided into 80% training data to build the model and 20% test data to evaluate the model. Model performance was evaluated based on the R^2 values and root mean square error (RMSE). Residual plots were used to check the model assumptions. The model with 2-factor interaction (2fi) was compared with the model with only the main effects. Then, the performance of the selected model was also evaluated using the 5-fold cross-validation (CV) method. R studio software was used to analyze the data and generate plots.

4 Results and discussion

4.1 Data Modeling

After randomly dividing the data into 80% for training and 20% for testing, the training data was fitted into a model with all main effects and 2-factor interaction (2fi) effects. The hypotheses are shown below. It was found the model p-value < 0.001 indicating not all non-intercept parameters are 0 at 95% confidence level. Table 4 lists part of the ANOVA table showing all main factors and significant 2fi effect. It is seen that the main effect of the smoker and the bmi * smoker 2fi effect are significant at $\alpha = 0.5$. Then the model was fitted with all main variables and the bmi: smoker interaction term. After this, the insignificant sex variable was removed from the model. The final model ANOVA result is shown in Table 5. The bmi variable was kept due to the presence of a significant bmi: smoker interaction term.

The model sum of square error (SSE) is 2.71e+10, close to the one with the full model (2.64e+10) obtained with much more parameters. The general form of this model is shown in equation (3). Table 6 shows the parameter estimates for this model. The model R^2 is 0.83 (83% of the sample variance explained by the model) indicating its good fitting. So, the final model can be written as equation (4), from which, the variable effects can be interpreted. For example, it is seen that increasing age by 1 year is predicted to increase the medical charge by 242.06 dollars, and having 1 more child is predicted to have 471.34 dollars more charge. Moreover, when the individual is a smoker, increasing bmi by 1 unit is predicted to increase the charge by 1518.45 dollars. The 95% CI for the parameters can be found in Table 7. Due to the presence of bmi*smoker interaction in the model, the effects of bmi and smoker cannot be directly interpreted from their estimates. However, a significant increasing in charges with higher ages, more children, or in the Northeast region can be predicted at a 95% confidence level from the CI ranges of the estimates. After fitting the data, the generated residual plots are shown in Figure 2.

Then, the performance of the developed model was evaluated by fitting the testing data and predicting the charges using the explanatory variables from the testing data set. The square root of mean square residual (RMSE) between fitted charges and actual charges was calculated using equation (5). The calculated RMSE for this final model is 5288.18.

H_o: all parameters except the intercept are 0; H_a: Not all of the non-intercept parameters are 0.

Table 4: ANOVA table for the model with main and 2-factor interactions (only the significant 2fi is shown).

	Sum Sq	Df	F value	Pr(>F)
Intercept	1.08e+07	1	0.42	0.516
age	8.55e+07	1	3.35	0.067
sex	1.58e+06	1	0.06	0.803
bmi	7.97e+06	1	3.12	<0.576
children	2.91e+07	1	1.14	0.286
smoker	2.73e+09	1	106.97	<0.001
region	1.51e+08	3	1.98	0.116
bmi:smoker	1.37e+10	1	537.95	< 0.001
...	...	...	...	...
Residuals	2.64e+10	1036	-	-

Table 5: ANOVA table for the final model.

	Sum Sq	Df	F value	P-value
Intercept	1.71e+07	1	0.67	<0.413
age	1.23e+10	1	483.65	<0.001
bmi	4.79e+04	1	1.88e-03	<0.965
children	3.49e+08	1	13.68	<0.001
smoker	3.41e+09	1	133.61	<0.001
region	2.43e+08	3	3.17	0.024
bmi*smoker	1.50e+10	1	585.92	< 0.001
Residuals	2.71e+10	1060	-	-

$$charges = \beta_0 + \beta_1 \times age + \beta_2 \times bmi + \beta_3 \times children + \beta_4 \times I_s y + \beta_5 \times I_r nw + \beta_6 \times I_r se \quad (2)$$
$$+\beta_7 \times I_r sw + \beta_8 \times I_s y \times bmi + \epsilon$$

Where $I_s y$ is smoking,$I_r nw$,$I_r se$, and $I_r sw$ are the indicators that the individual is from the NW, SE, SW of the US, and ϵ follows N(0,σ^2).

Final Model Summary:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -817.309    998.858  -0.818 0.413402
age                242.056     11.006  21.992  < 2e-16 ***
bmi                  1.291     29.810   0.043 0.965454
children           471.343    127.440   3.699 0.000228 ***
smokeryes       -22742.678   1967.498 -11.559  < 2e-16 ***
regionnorthwest   -716.637    457.138  -1.568 0.117259
regionsoutheast  -1188.886    444.422  -2.675 0.007585 **
regionsouthwest  -1223.729    450.566  -2.716 0.006715 **
bmi:smokeryes     1518.445     62.731  24.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5053 on 1061 degrees of freedom
Multiple R-squared:  0.8299,Adjusted R-squared:  0.8286
F-statistic: 647.2 on 8 and 1061 DF,  p-value: < 2.2e-16
```

$$\hat{charges} = -817 + 242 \times age + 1 \times bmi + 471 \times children - 22743 \times I_s y - 717 \times I_r nw \quad (3)$$
$$-1189 \times I_r se - 1224 \times I_r sw + 1518 \times I_s y \times bmi$$

Where $I_s y$ is smoking,$I_r nw$,$I_r se$, and $I_r sw$ are the indicators that the individual is from the NW, SE, or SW of the US.

Table 6: 95% CI for the model estimates.

	2.5%	97.5%
Intercept	-2777.3	1142.7
age	220.5	263.7
bmi	-57.2	59.8
children	221.3	721.4
smoker(yes)	-26603.3	-18882.0
region(NW)	-1613.6	180.4
region(SE)	-2060.9	-316.8
region(SW)	-2107.8	-339.6
bmi:smoker(yes)	1395.4	1641.5

$$RMSE = \sqrt{\frac{\sum_1^n (\hat{y}_i - y_i)^2}{n}} \quad (4)$$

Where n is the testing sample size, $\hat{y}_i$ is the predicted charge, and y_i is the actual charge.

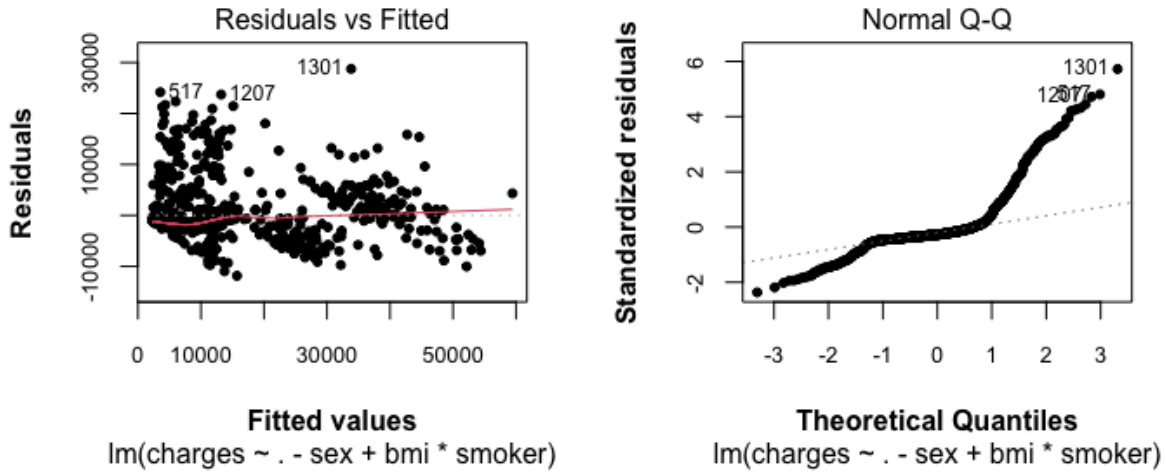

Figure 2: Residual and QQ plots.

4.2 Model Performance

The model above shows high goodness of fitting with $R^2 = 0.83$ and RMSE = 5288 through the train-test process. To further check the model performance and consistency, the complete data were randomly divided into 5 folds, each fold of the data set was used as the test data to evaluate the model selected. Criteria values such as RMSE and R^2 are obtained and reported in Table 8. The small ranges of R^2 and RMSE values confirm the generalization of the model, and the high R^2 values suggest the good fitting of the model.

Table 7: Training and testing result.

test	R_2	prediction RMSE
fold1	0.802	5501.1
fold2	0.826	4196.8
fold3	0.813	5489.4
fold4	0.830	4938.1
fold5	0.834	5091.3
Average	0.830	5043.4

4.3 Compare variables

To compare the significance of the individual variables for predicting insurance cost, all main variables except sex are fitted into the model without interaction effects. Table 9 shows the model summary. It is found that smoker is a significant variable with a high and positive estimate (23972.85) indicating that individuals who do smoke have 23972.85 higher predicted medical costs. In addition, the age and bmi variables are also significant with positive estimates, which means higher age or bmi leads to higher predicted medical costs. The R^2 of this model is only 0.74 (< 0.83), and the SSE is 4.20e+10, higher than that of the model (4), it shows lower prediction accuracy than the final model (4).

Model Summary:

```
Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -10385.47    1142.33  -9.091  < 2e-16 ***
age                234.61      13.70  17.123  < 2e-16 ***
bmi                320.76      33.29   9.637  < 2e-16 ***
children           416.92     158.68   2.627  0.00873 **
smokeryes        23972.85     476.51  50.310  < 2e-16 ***
regionnorthwest   -590.50     569.24  -1.037  0.29981
regionsoutheast  -1100.11     553.42  -1.988  0.04709 *
regionsouthwest  -1084.66     561.04  -1.933  0.05347 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6292 on 1062 degrees of freedom
Multiple R-squared:  0.736,Adjusted R-squared:  0.7343
F-statistic:   423 on 7 and 1062 DF,  p-value: < 2.2e-16
```

5 Conclusion

In this report, multiple linear regression, train-test, and CV procedures were used to build and evaluate prediction models for medical charges of individuals with variables including age, bmi, children number, smoker, sex, and region. The sex variable was found to be insignificant at $\alpha = 0.05$, while the age, children number, smoker, and region variables are significant. In addition, the smoker * bmi interaction also turned out to be significant. The final model is presented below in equation (6). It has high goodness-of-fit (R^2=0.814) and its performance is consistent throughout all 5 CV folds with RMSE = 5043.37 and $R^2 = 0.83$. From the final model, it is seen that individuals who smoke tend to have higher medical costs considering both main and interaction effects. And individuals from the Northeast with more children and at a higher age are predicted to have increasing amounts of healthcare charges.

$$\hat{charges} = -817 + 242 \times age + 1 \times bmi + 471 \times children - 22743 \times I_sy - 717 \times I_rnw \quad (5)$$
$$-1189 \times I_rse - 1224 \times I_rsw + 1518 \times I_sy \times bmi$$

Where I_sy is smoking,I_rnw,I_rse, and I_rsw are the indicators that the individual is from the NW, SE, or SW of the US.

Reference

1. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/
2. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Appendix 1-Figures

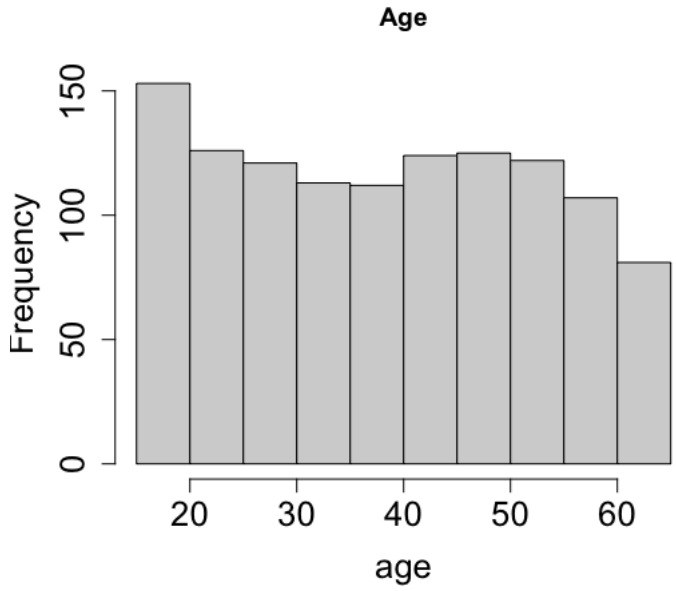

Figure 3: Distribution of age.

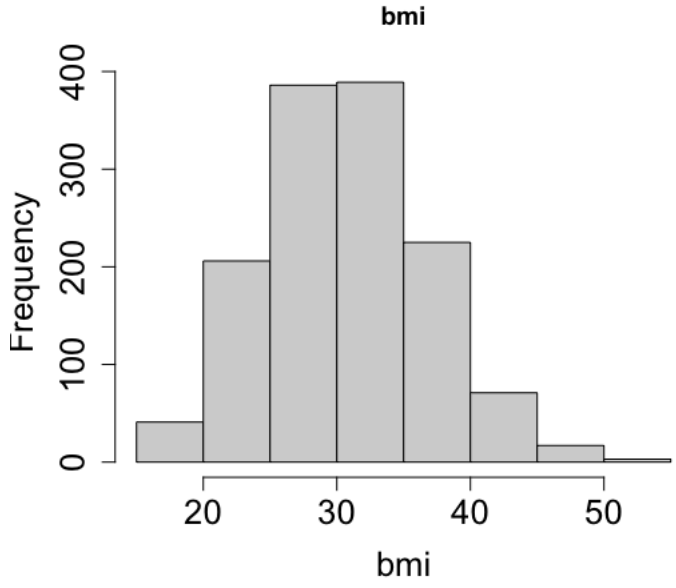

Figure 4: Distribution of bmi.

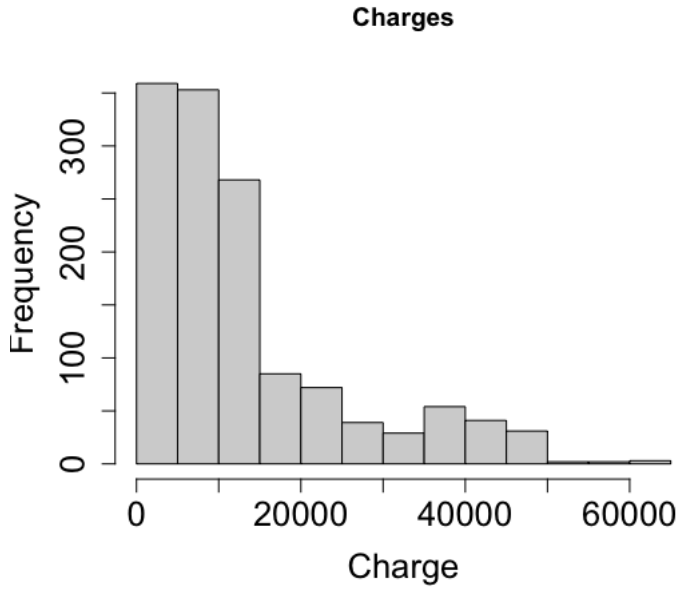

Figure 5: Distribution of charges.

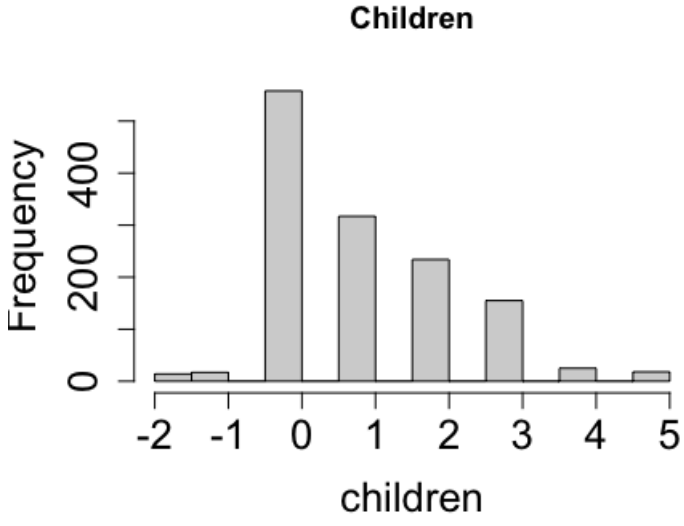

Figure 6: Distribution of children numbers.

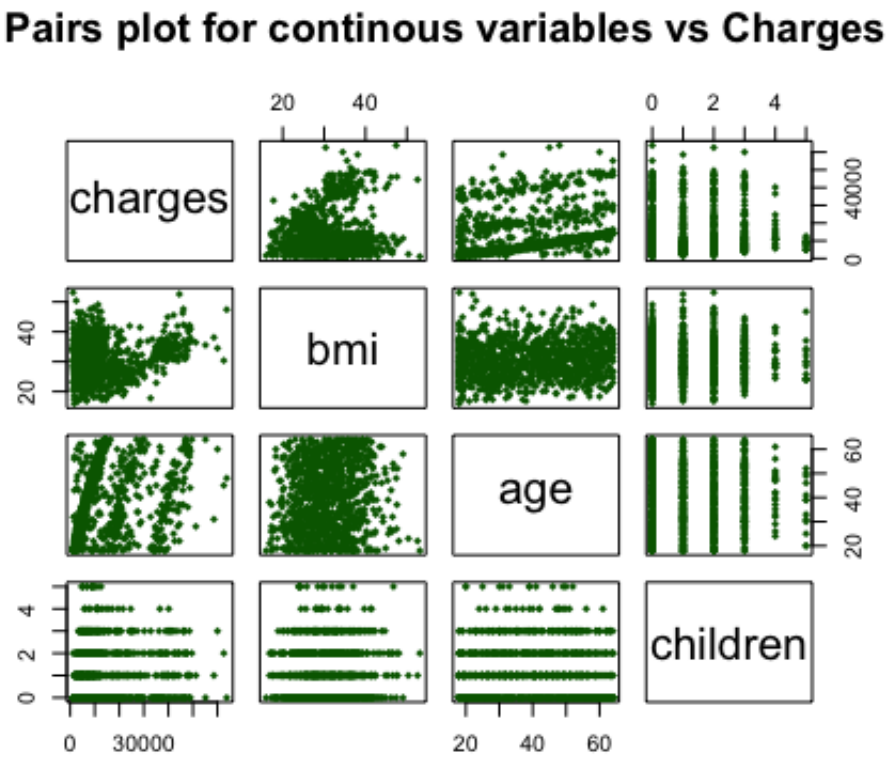

Figure 7: Matrix plots of the numerical variables.

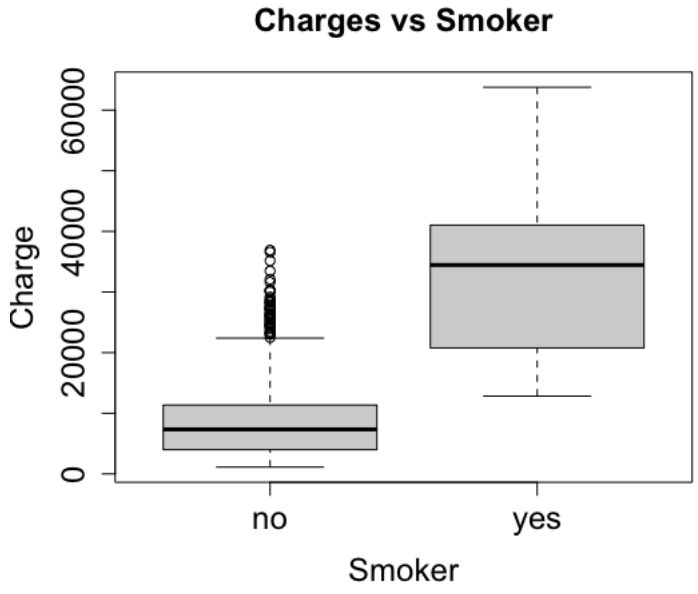

Figure 8: Plot of charges with smoker variable.

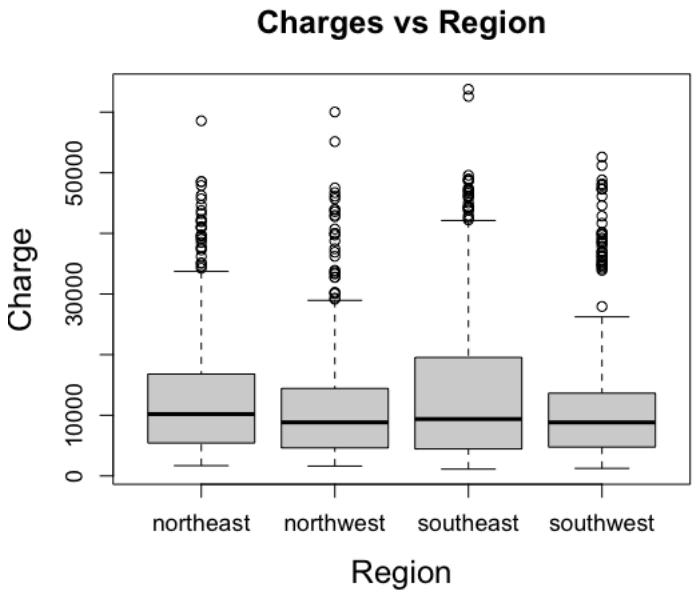

Figure 9: Plot of charges with region variable.

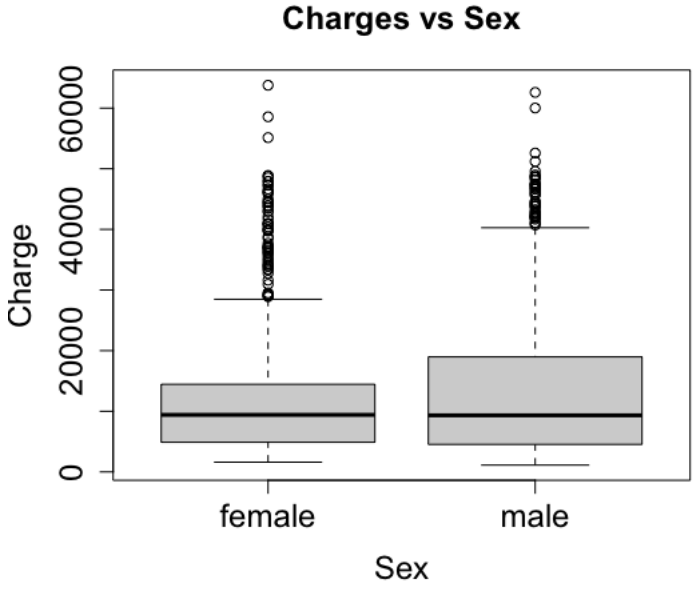

Figure 10: Plot of charges with sex variable.

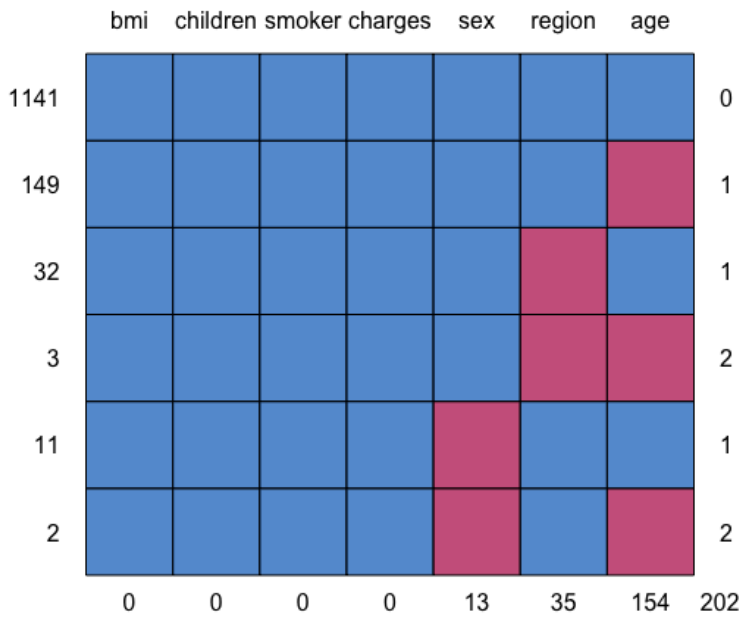

Figure 11: Numbers of missing data.

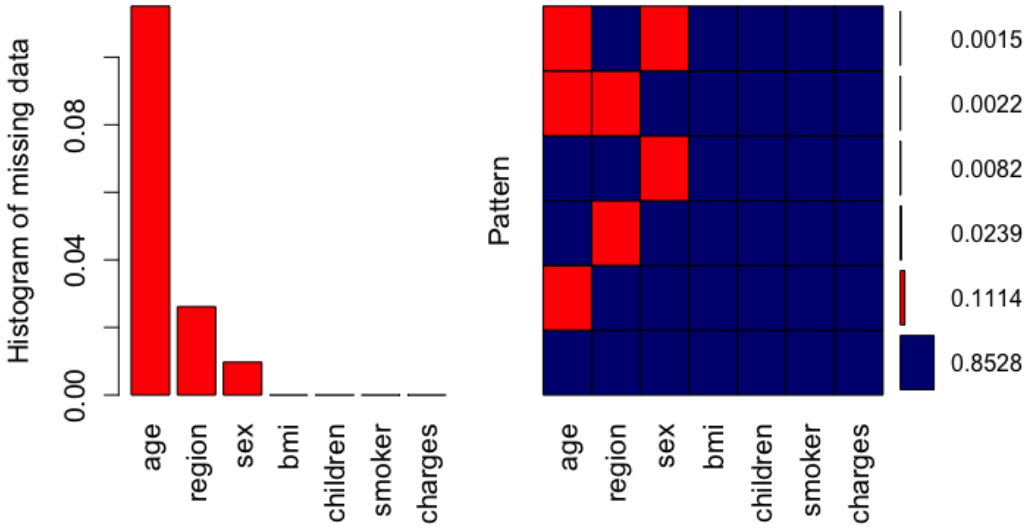

Figure 12: Percentages and distribution of the missing data

Appendix 2-R Codes

```
library(tidyverse)
library(mice)
library(plyr)
library(VIM)
library(car)
library(caret)

#########################################################################
#                                                                       #
#                              Load the data                            #
#                                                                       #
#########################################################################
df <- read_csv("insurance.csv")
print(df)

#########################################################################
#                                                                       #
#                         Exploratory Data Analysis                     #
#                                                                       #
#########################################################################
age <- df$age
sex <- df$sex
bmi <- df$bmi
children <- df$children
smoker <- df$smoker
region <- df$region
charges <- df$charges

#################################################
#                                               #
#                   Data Summary                #
#                                               #
#################################################
print("Summary for continuous variables:")
summary(df[, -c(2, 5,6)])

paste("age S.D. = ",round(sd(age,na.rm = TRUE), 2))
paste("bmi S.D. = ",round(sd(bmi,na.rm = TRUE), 2))
paste("children S.D. = ",round(sd(children,na.rm = TRUE), 2))
paste("charges S.D. = ",round(sd(charges,na.rm = TRUE), 2))

print("Summary for categorical variables:")
table(df$sex)
table(df$smoker)
table(df$region)

#################################################
#                                               #
```

```
#                Data Visualization              #
#                                                #
##################################################
# Data Distribution
par(mfrow = c(2,2))
hist(df$bmi, xlab = "BMI", cex.axis = 1.6, cex.lab = 1.6, main = "Dsitribution of BMI")
hist(df$age, xlab = "Age", cex.axis = 1.6, cex.lab = 1.6, main = "Dsitribution of Age")
hist(df$charges, xlab = "Charges", cex.axis = 1.6, cex.lab = 1.6, main = "Dsitribution
hist(df$children, xlab = "Children", cex.axis = 1.6, cex.lab = 1.6, main = "Dsitributic

# Relationship Between Continuous Variables
x <- cbind(charges,bmi,age,children)
pairs(x,cex = 1.2, col = "darkgreen",pch = 18,
      main = " Pairs plot for continuous variables vs Charges")

# Charge vs Categorival Variables
par(mfrow = c(2,2))
plot(df$charges ~ factor(df$smoker),ylab = "Charge",xlab = "Smoker",
     cex.axis = 1.5,cex.lab = 1.5,cex.main = 1.5,main = "Charges vs Smoker")
plot(df$charges ~ factor(df$sex),ylab = "Charge",xlab = "Sex",
     cex.axis = 1.5,cex.lab = 1.5,cex.main = 1.5,main = "Charges vs Sex")

plot(df$charges ~ factor(df$region),ylab = "Charge",xlab = "Region",
     cex.axis = 1.1,cex.lab = 1.5,cex.main = 1.5,main = "Charges vs Region")

# Exploration Plots
# Regroup by BMI (in normal range (<30) or not) and smoker (yes or no)
bmismoke <- replace(bmi, (bmi > 30) & smoker == "yes", "out-yes")
bmismoke2 <- replace(bmismoke,(bmi > 30)& smoker == "no", "out-no")
bmismoke3 <- replace(bmismoke2,(bmi <= 30) & smoker == "yes", "in-yes")
bmismoke4 <- replace(bmismoke3,(bmi <= 30) & smoker == "no", "in-no")
df2 <- mutate(df, bmic = bmismoke4)

options(repr.plot.width = 8, repr.plot.height = 5)
sp<-ggplot(data = df2) +
  geom_point(aes(x = age, y = charges, group = 1,color = factor(bmic)))+
  ylab("charge") + xlab("age") + ggtitle("charges vs age, smoker and bmi level")
sp+scale_color_manual(values = c("red", "blue", "green","black"))+
theme (title = element_text(size = 20),
axis.text = element_text(size = 16),
axis.title = element_text(size = 20,face = "bold"),
legend.text = element_text(size = 16),
legend.title = element_text(size = 20))
print ("Legend label example: in-no means BMI in the normal range and does no smoke")

sp<-ggplot(data = df2) +
  geom_point(aes(x = bmi, y = charges, group = 1,color = factor(smoker)))+
  ylab("charge") + xlab("bmi") + ggtitle("charges vs bmi and smoker")
sp+scale_color_manual(values = c("red", "blue"))+
theme (title = element_text(size = 20),
```

```
axis.text = element_text(size = 16),
axis.title = element_text(size = 20,face = "bold"),
legend.text = element_text(size = 16),
legend.title = element_text(size = 20))
df2 <- df2[, 1:7]

#########################################################################
#                                                                       #
#                              Data Cleaning                            #
#                                                                       #
#########################################################################

# children: no non-integer input
length(df2$children[df2$children%%1 != 0])
# number of negative inputs: 31
sum(df2$children < 0)
df2$children[df2$children < 0]
# replace with positive values
df3 <- df2
df3$children[df3$children < 0] <- abs(df3$children[df3$children < 0])

# region: change k#!a to NA
table(df3$region)
df3$region <- replace(df3$region, df3$region == "k#!a", NA)  # replace k#!a as NA.

# Missing Data Distribution
print("Missing patterns and percentages:")
md.pattern(df3)[2]
aggr(df3, col = c('navyblue','red'), numbers = TRUE, sortVars = TRUE,
     labels = names(df3), cex.axis = 1.2, gap = 3, ylab = c("Histogram of missing data"

# Data Imputation
# age: impute using mice pmm
X <- df3[,-7]
tempData <- mice(X,m = 5,maxit = 50,meth = 'pmm',seed = 500)
completedData <- complete(tempData,2)
df4 <- cbind(completedData,df3[,7])

# sex: impute with mode
df4$sex[is.na(df4$sex)] <- "male"
table(df4$sex)

# region: impute with mode
df4$region[is.na(df4$region)] <- "southeast"
print("Cleaned Data:")
head(df4)

#########################################################################
#                                                                       #
#                              Model Building                           #
```

```
#                                                                          #
############################################################################

# Split data, 20: 80
train_ind <- sample(1:1338, round(1338*0.8), replace = FALSE)
train <- df4[train_ind, ]
test <- df4[-train_ind, ]

#####################################################
#                                                   #
#                 Multicollinearity                 #
#                                                   #
#####################################################
# Correlation matrix
x_conti <- train[, c(1,3,4)]
print("Correlation matrix")
cor(x_conti)

#####################################################
#                                                   #
#            Fit Mode and Model Selection           #
#                                                   #
#####################################################
# Full model with all 2-factor interactions
g1 <- lm(charges ~.^2, data = train)
# summary(g1)
Anova(g1,type = 3)

# Keep all main and bmi*smoker terms
g2 <- lm(charges ~. + bmi*smoker, data = train)
# summary(g2)
Anova(g2, type = 3)

# Keep sig. main and the interaction terms
g3 <- lm(charges ~.- sex + bmi*smoker, data = train)
print ("ANOVA Table (type 3):")
Anova(g3,type = 3)

#############################################################################
#                                                                           #
#                            Mode Performance                               #
#                                                                           #
#############################################################################

print ("Summary:")
summary(g3)

print("Baseline used")
contrasts(factor(train$smoker))
contrasts(factor(train$region))
```

```
# smoker baseline: no
# ref: northeast". 1-3: "northwest" "southeast" "southwest"

# Prediction on Test Set
pred <- predict (g3, test)
print("RMSE of the Test Set: ")
round(sqrt(sum((pred - test$charges)^2)/dim(test)[1]), 2)

# Confidence Intervals of the Estimates
print ("95% CI for the Estimates:")
confint(g3)

# Residaul Plots
print ("Residual Plots")
layout(matrix(c(1,2), 1, 2, byrow = F))
plot(g3,1:2, col = "black", pch = 16, cex = 0.8, cex.axis = 1.5,
cex.lab = 1.5, font.lab = 2.5)

# Compare to the model without bmi * smoker effect
g4 <- lm(charges ~ . - sex, data = train)
summary(g4)
# Anova(g4, type =3)
# Smoking increases the medical charge
# R^2 decreases from 0.83 to 0.74

#####################################################
#                                                   #
#                  Cross Validation                 #
#                                                   #
#####################################################
ctrl <- trainControl(method = "cv", number = 5)

# Fit a regression model and use k-fold CV to evaluate the performance
model_cv <- train(charges ~.- sex + bmi*smoker, data = train, method = "lm", trControl

# Summary of K-fold CV
print ("k-fold CV Summary:")
print(model_cv)

print("Model Coefficients:")
model_cv$finalModel

#view predictions for each fold
model_cv$resample
```