Determining Predictors for Health Insurance Premiums and Assessing Model Predictive Ability

STAT 840

Chad Stephens

12/13/2022

Department of Biostatistics and Data Science

University of Kansas, USA

July 26, 2021

## 

## I. Abstract

Health insurance is extremely important to many people’s lives, especially for those who suffer from significant medical issues (Hoffman & Paradise 2008; Kluender et al. 2021). While having health insurance is beneficial, policyholders may still have premiums that they need to pay for that often times can become expensive. However, it’s often difficult to decipher how much the total cost a policyholder owes is calculated and what specifically influence it. A multiple linear regression model is used to determine potential important predictor variables that can influence the amount somebody owes to an insurance company. Additionally, the model’s predictive ability is determined for future policyholders to understand what they may owe. The model suggests that age, sex, body mass index (BMI), number of children in the household, whether the individual smokes, region, and total yearly premium charges are all important predictor variables. The MSE and MSPR having similar values indicate that the regression model can be used to make predictions as it relates to the total charges.

## II. Introduction

### A. Background

In the United States, medical debt is a serious issue that impacts many individuals. This is especially true to those who suffer from more serious medical complications such as diabetes or cancer (Iuga & McGuire 2014). Given the costs to seek medical care is a significant expense, most individuals have insurance to help cover some or even all the cost they may procure overtime. Despite having health insurance, however, individuals often have to pay a certain amount of premiums that can still be costly, albeit at a much lower price and risk compared to going without insurance.

### B. Aim

The Purpose of this study is to evaluate the relationship of various predictor variables and the total premium costs per year that an individual owes and whether a possible new policyholder can predict future charges. Data were obtained from *Machine Learning with R* by Brett Lantz where 1,338 samples of anonymous individual policyholders’ key information were obtained: age, sex, body mass index (BMI), number of children in the household, whether the individual smokes, region, and total yearly premium charges. Sex, number of children in the household, whether the individual smokes, and region is treated as factor variables. A multiple linear regression model is used where premium charges will be the response variable and the rest will be possible predictor variables. Researchers have investigated various associations between predictor and response variables and predicted possible insurance costs by using similar machine learning methods.

## III. Methods

### A. Statistical Software and Analysis

All statistical analysis and figures produced is done in RStudio version 4.2.2 (released on 10/31/22). Unless otherwise noted, all inferences conducted are based on = 0.05. The final model is selected via automatic model selection algorithm, where the model having a combination of the highest R2 and adjusted R2 values and the lowest Mallow’s and Bayesian Information Criterion (BIC) values were selected (Kutner et al. 2005). The model selection technique used were the exhaustive search method. This method is guaranteed to find the predictor variable subset with the best evaluation criterion and is often the ideal technique when the number of possible predictor variables is less than 20. Model assumptions of linear regression were all considered and accounted for in the final model.

### B. Preliminary Model

A multiple linear regression model is considered. Let

the total yearly health insurance premium charges for the policyholder

age for the policyholder,

sex for the policyholder,

body mass index (BMI) for the policyholder,

number of children in a household for the policyholder,

whether smoker or non-smoker for the policyholder,

region where living for the policyholder.

While no predictor variables were added to the final model, it contains more ’s given that sex, children, smoker, and region were treated as factor variables. Based on automatic variable selection methods in combination with criterion-based statistics, no predictor variables were dropped from the model. All assumptions (linearity, multicollinearity, independence, homoscedasticity, and multivariate normality) of the final model were met or corrected so that it could be met. Refer to the Appendix section for more details.

### C. Final Model

The final model is given by

where , , and and are the unknown model parameters.

The following factor variables are associated with the description discussed in the Preliminary Model section:

$

### D. Preliminary Analysis

Sex, smoker, and region were treated as factor variables given that they’re a text data type. Although children has a number data type, the small range shown in Table 1 can allow it to be changed into a factor variable.

Distribution of Number of Children per Household

0 1 2 3 4 5   
 574 324 240 157 25 18

Table 1 –distribution of children per household. Given the small range, this can become a factor variable

Looking at the distribution of the response variable can be a good way to indicate whether a possible transformation is needed. Given that Figure 1 shows a positive skewed distribution, either a log or Box-Cox transformation may need to be considered. However, further analyses should be considered first.

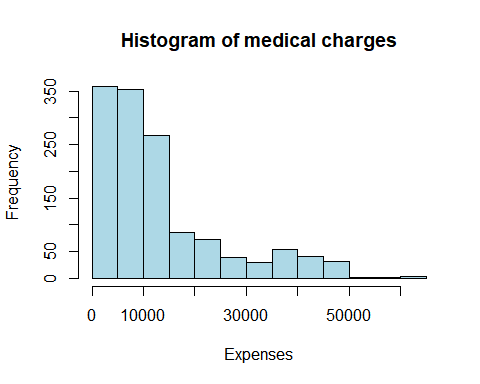


Figure 1 – histogram of the distribution of medical charges a policyholder owes for the year

### 

### E. Full Analysis

The final model has R2 = 0.7781, indicating that 77.91 percent of the total variation observed in the total charges is explained when the predictor variables discussed in the Final Model section are introduced. This suggests that these variables are effective predictors for charges. Additionally, p < 0.001 suggests that the model is statistically significant. The final model needed a Box-Cox transformation on its response variable given that the assumption of multivariate normality were violated. More details can be found in the Appendix.

## IV. Results

Based on automatic variable selection exhaustive method, the model that included all the original and factored predictor variables is the most effective. R2 = 0.773, adjusted R2 = 0.772, Mallow’s = 33.319, and BIC = -1921.3478 all confirmed that the final model selected should be used. This suggests that all the predictor variables (age, sex, bmi, children, smoker, and region) have a relationship with the total premium costs per year that a policyholder owes. To determine the model’s predictive ability, the MSE from the training data set and the MSPR from the validation set were compared (Kutner 2005). MSE = 2.424 and MSPR = 2.576. Given that the values of the MSPR and MSE are close, the MSE for the regression model is not biased and gives an appropriate indication of the model’s predictive ability. In other words, the model can be utilized to make predictions about the future charges a new policyholder may owe to the insurance company.

## V. Discussion

The original aim of determining important predictor variables and the model’s predictive ability as it relates to charges has been successful. However, there are a few noteworthy challenges that should be considered. While the assumption linearity has concluded to be met after the Box-Cox transformation (Figure 2), there is a slight curvature feature that could warrant more investigation. Given that in the context of insurance premiums, 1,338 observations is a relatively low sample size. A larger sample size could possibly change whether the assumption of linearity is met, so additional studies should consider to investigate this possible change. If there is a change, then other transformation types may need to be performed or even a differently type of regression be selected. Similarly, there were slight deviations from being horizontal while viewing for homoscedasticity (Figure 3). A larger sample size may be necessary to fully understand whether the assumption will still be met. Finally, all the predictor variables were statistically significant (p 0.05) except for region northwest (p = 0.075). While this may indicate that the region variable should have been dropped, this is treated as a factor variable. Given that the other regions were statistically significant and it’s a factor variable, region northwest cannot be dropped from the model. However, additional tests such as an one-way analysis of variance (ANOVA) could be used to look into this further.

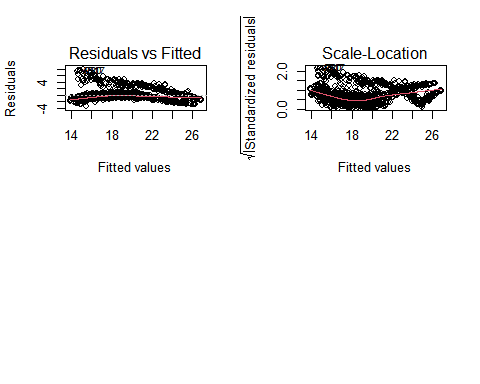


Figure 2 (left) – scatter plot of the residuals vs fitted values

Figure 3 (right) – scatter plot of the standardized residuals vs fitted values

## VI. References

Boylan, T. A., Cuddy, M. P., & O’Muircheartaigh, I. G. (1982). Import Demand Equations: An Application of a Generalized Box-Cox Methodology. *International Statistical Review / Revue Internationale de Statistique*, *50*(1), 103–112.

Hoffman, C. & Paradise, J. (2008), Health Insurance and Access to Health Care in the United States. *Annals of the New York Academy of Sciences*, 1136: 149-160.

Iuga, A.O. &  McGuire M. J. (2014). Adherence and health care costs. *Risk Manag Healthc Policy*, 7:35-44.

Joseph, V. R. (2022). Optimal Ratio for Data Splitting.

Kluender R., Mahoney N., Wong F., & Yin W. (2021). Medical Debt in the US, 2009-2020. *JAMA,* 326(3):250–256.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.

Lantz, B. (2013). Machine Learning with R

## VII. Appendix

### A. Model Diagnostics

Based on Figure 4, the preliminary model failed to meet the assumption of multivariate normality.

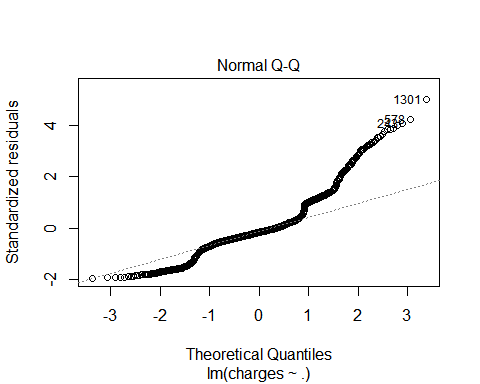


Figure 4 – Normal Q-Q plot to help determine multivariate normality assumption

Often, a log transformation on the response variable can help correct the normal Q-Q plot. However, Figure 5 shows that the normal Q-Q plot has only slightly improved compared to Figure 4. In other words, the transformation hasn’t been met and another type of transformation is needed. Boylan et al. (1982) have suggested that when dealing with positively skewed data (Figure 1) a Box-Cox transformation is typically effective (Figure 6).

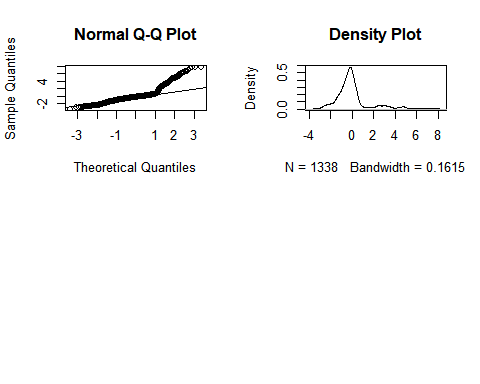


Figure 6 - Normal Q-Q plot to help determine multivariate normality assumption and a density plot to see if there’s a normal distribution

The Box-Cox transformation on the response variable were effective in having the assumption of multivariate normality be met.

The linearity assumption has been met in the final model (Figure 7). While there is slight curvature, it’s still close to 0. Further details can be found in the Discussion section.

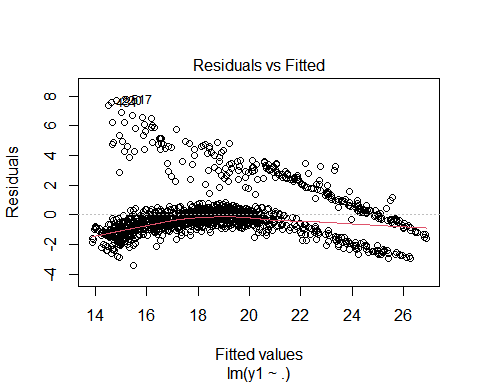


Figure 7 – scatter plot of the residuals vs fitted values

Using VIF 5 as a threshold, multicollinearity can be determined. Table 2 suggests that there’s no indication of multicollinearity as all the VIF values are well below the specified threshold.

VIF Values

## GVIF Df GVIF^(1/(2\*Df))  
## age 1.020607 1 1.010251  
## sex 1.009334 1 1.004656  
## bmi 1.108836 1 1.053013  
## children 1.024871 5 1.002460  
## smoker 1.018024 1 1.008972  
## region 1.109919 3 1.017533

Table 2 – VIF values for each predictor variable

After the Box-Cox transformation, it can be concluded that each observation in the data set are independent based on the Durbin-Watson test:

: there is no correlation among the residuals,

: the residuals are autocorrelated.

Given that p = 0.326, is failed to be rejected, indicating that the observations are independent and the assumption has been met.

To determine if the residuals have constant variance at every point in the linear model (i.e. homoscedasticity), a scale-location plot (Figure 8) is used to see whether a pattern exists.

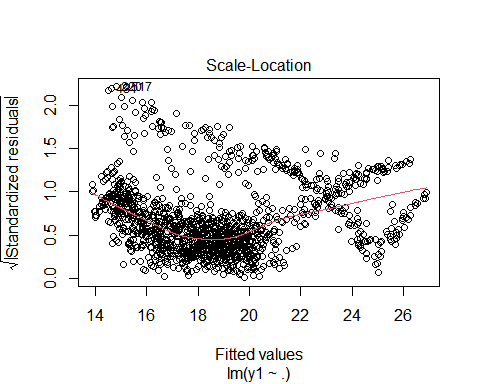


Figure 8 – scatter plot of the standardized residuals vs fitted values

While there are slight deviations from a horizontal pattern, the assumption has been met. More details about the possible implications about this can be found in the Discussion section.

The Box-Cox transformation has allowed the model to meet all of its assumptions. The next step is to select the best model based on a combination of certain predictor variables being included and excluded. Of note, this still under the model with the Box-Cox transformation.

### B. Model Selection

Since the number of predictors used is less than 20, the exhaustive search method provides the best way to evaluate which combination of predictor variables should be selected. R2 = 0.773, adjusted R2 = 0.772, Mallow’s = 33.319, and BIC = -1921.3478 all confirmed that the final model selected should be used.

### C. Model Validation

Model validation is needed since it can determine the model’s predictive ability. A 80:20 split for the training and validation data sets were used as it’s common to use in industry (Joseph 2022). Both the training model (R2 = 0.7837) and the validation model (R2 = 0.7647)) had similar R2 values compared to the final model. If these valued varied more, then it could be a cause for concern. MSE = 2.424 and MSPR = 2.576. Given that the values of the MSPR and MSE are close, the MSE for the regression model is not biased and gives an appropriate indication of the model’s predictive ability. This indicates that the model can be utilized to make predictions about the future charges a new policyholder may owe to the insurance company.

**R Code:**

library(faraway)

library(readxl)

library(readr)

library(knitr)

library(leaps)

library(car)

library(dplyr)

library(MASS)

insurance <- read\_excel("C:/Users/steph/Downloads/insurance.xlsx")

table(insurance$children)

str(insurance)

insurance$sex <- as.factor(insurance$sex)

insurance$children <- as.factor(insurance$children)

insurance$smoker <- as.factor(insurance$smoker)

insurance$region <- as.factor(insurance$region)

str(insurance)

hist(insurance$charges,main="Histogram of medical expenses", xlab = "Expenses",ylab = "Frequency",col="lightblue")

model <- lm(charges ~ ., data=insurance)

summary(model)

plot(insurance)

par(mfrow = c(2, 2))

plot(model)

vif(model)

crPlots(model)

model2 <- lm(log(charges) ~ ., data=insurance)

summary(model2)

par(mfrow = c(2, 2))

plot(model2)

x <- resid(model2)

par(mfrow = c(2, 2))

qqnorm(x)

qqline(x)

plot(density(x))

bc <- boxcox(model)

(lambda <- bc$x[which.max(bc$y)])

insurance$y1 <- (insurance$charges^lambda-1)/lambda

insurance2 <- subset(insurance, select= -c(charges))

model3 <- lm(y1 ~ ., data=insurance2)

summary(model3)

res <- resid(model3)

qqnorm(res)

qqline(res)

plot(density(res))

par(mfrow = c(2, 2))

plot(model3)

plot(model3,1)

crPlots(model3)

vif(model3)

durbinWatsonTest(model3)

plot(model3,3)

par(mfrow = c(2, 2))

plot(model3,2)

plot(density(res))

b <- regsubsets(y1 ~ ., data=insurance2)

rs <- summary(b)

print(rs)

names(rs)

PRESS <- function(linear.model) {

#' calculate the predictive residuals

pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)

#' calculate the PRESS

PRESS <- sum(pr^2)

return(PRESS)

}

rs$rsq

rs$adjr2

rs$cp

rs$bic

extractAIC(model3)

model3 <- lm(y1 ~ ., data=insurance2)

best\_subset <-

regsubsets(y1~.,

data =insurance2,

nbest = 1, # 1 best model for each number of predictors

nvmax = NULL, # NULL for no limit on number of variables

force.in = NULL, force.out = NULL,

method = "exhaustive")

summary\_best\_subset <- summary(best\_subset)

as.data.frame(summary\_best\_subset$outmat)

which.max(summary\_best\_subset$adjr2)

summary\_best\_subset$which[12,]

summary(model3)

model4 <- lm(y1 ~ .+bmi\*smoker, data=insurance2)

summary(model4)

train <- insurance2[1:1070,]

validation <- insurance2[1071:1338,]

train\_model <- lm(y1 ~ ., data=train)

summary(train\_model)

validation\_model <- lm(y1 ~ ., data=validation)

summary(validation\_model)

model\_summary <-summary(train\_model)

mse <- mean(model\_summary$residuals^2)

mse

sse <- sum((fitted(validation\_model) - validation$y1)^2)

n <- 268

mspr <- sse/n # mean squared prediction error

mspr