The University of Akron

Management Department

Advanced Data Analytics 6500-663

Business Report

Predicting medical expenses using linear regression

By Kirill Samaray

Introduction

In the current healthcare industry, medical expenses continue to rise, creating financial burdens for individuals, families, and health insurance companies. Health insurance companies are seeking effective methods to manage and control healthcare costs while ensuring their customers receive quality healthcare services. Predicting medical expenses accurately can assist in managing healthcare costs and improving financial stability.

In this business report, we explore the use of linear regression to predict medical expenses in a health insurance company. Linear regression is a widely used statistical method that models the relationship between a dependent variable and one or more independent variables. We will discuss the steps involved in developing a linear regression model to predict medical expenses, including data preprocessing, feature selection, model training, and performance evaluation.

Moreover, we will analyze the potential benefits of implementing a predictive model for medical expense estimation in a health insurance company. The report will also highlight the limitations and challenges that may arise when implementing such a model.

The objective of this report is to provide a comprehensive analysis of the use of linear regression in predicting medical expenses in a health insurance company. The report will provide valuable insights to decision-makers in the healthcare industry, including health insurance companies, policy-makers, and healthcare providers.

Problem Statement

Health insurance companies are struggling to manage healthcare costs while providing quality healthcare services to their customers. Without accurate estimates of medical expenses, health insurance companies may face challenges in pricing their services, forecasting their revenue and managing their financial risks.

Currently, health insurance companies rely on historical data and expert opinions to estimate medical expenses. However, these methods are often imprecise and unreliable, leading to inaccurate estimates and financial losses. Health insurance companies require a more robust and accurate method to estimate medical expenses, which can help them control healthcare costs and improve their financial stability.

Linear regression has been used in various fields, including healthcare, to predict outcomes and estimate costs. However, the use of linear regression in predicting medical expenses in a health insurance company has not been widely explored.

In this report, we aim to address the problem of inaccurate medical expense estimation in health insurance companies by exploring the use of linear regression. The objective is to develop a

predictive model that accurately estimates medical expenses based on relevant factors, such as age, sex, and other factors.

Objectives

The main objective of this business report is to explore the use of linear regression to predict medical expenses in a health insurance company. The report aims to achieve the following specific objectives:

- To identify the relevant factors that may affect medical expenses, such as age, sex, bmi, region, etc.
- To develop a linear regression model to predict medical expenses based on the identified factors.
- To evaluate the performance of the developed model using appropriate statistical measures, such as mean squared error and R-squared.
- To analyze the potential benefits of implementing a predictive model for medical expense estimation in a health insurance company, including cost savings and financial stability.
- To discuss the limitations and challenges that may arise when implementing a predictive model for medical expense estimation in a health insurance company, such as data quality and privacy concerns.
- To provide recommendations for health insurance companies to effectively use predictive models for medical expense estimation to manage healthcare costs and improve their financial stability.

Methodology

The methodology used in this report is based on the analysis of the publicly available data set “insurance.csv” and a review of the existing literature: Brett Lantz, Machine Learning with R, 2nd Ed., Packet Publishing, 2015 (ISBN: 978-1-78439-390-8).

The following report is split into the sections:

- Step 1. Overview of the model.
  The first step will describe the OLS (Ordinary Least Squares) algorithm and its application to predict medical expenses for an insurance company.
- Step 2. Collecting data.
  The second step in the methodology is to collect the relevant data needed to train and test the linear regression model. The data should include medical expense data, demographic data such as age, sex, number of children, geographical region, and lifestyle

data such as smoking status and bmi index. The data can be obtained from electronic health records, insurance claims databases, and patient surveys. In our case we will use a simulated dataset with some fictional data. This data was intentionally compiled for this project, although it can reasonably well reflect real information.

- Step 3. Exploring and preparing data.
  The collected data will need to be cleaned and preprocessed to ensure data quality and consistency. This step may include handling missing data, removing outliers, and transforming the data to meet the assumptions of linear regression. Next, we will perform feature selection to identify the most relevant factors that impact medical expenses. This step involves analyzing the correlation between each independent variable and the dependent variable and selecting the most significant factors to include in the model.

- Step 4. Model training.
  We will use the selected features to develop a linear regression model that predicts medical expenses. The model will be trained using a portion of the collected data and tested using a separate portion to evaluate its performance.

- Step 5. Model Evaluation.
  We will evaluate the performance of the linear regression model using appropriate statistical measures, such as mean squared error and R-squared. These measures will indicate the accuracy and precision of the model in predicting medical expenses.

- Step 6. Model Optimization.
  We will analyze the results of the linear regression model and will attempt to optimize the results of the model. We will discuss its potential benefits for health insurance companies, such as improving cost estimation, pricing strategy, and financial stability.
- Step 7. Recommendations
  Finally, we will provide recommendations for health insurance companies to effectively use predictive models for medical expense estimation to manage healthcare costs and improve their financial stability. These recommendations may include investing in data quality improvement, integrating predictive models into their pricing strategy, and addressing ethical and privacy concerns.

Step 1. Overview of the model

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. Dependent variable is the one we are trying to predict and an independent variable acts as a predictor. The goal of linear regression is to develop

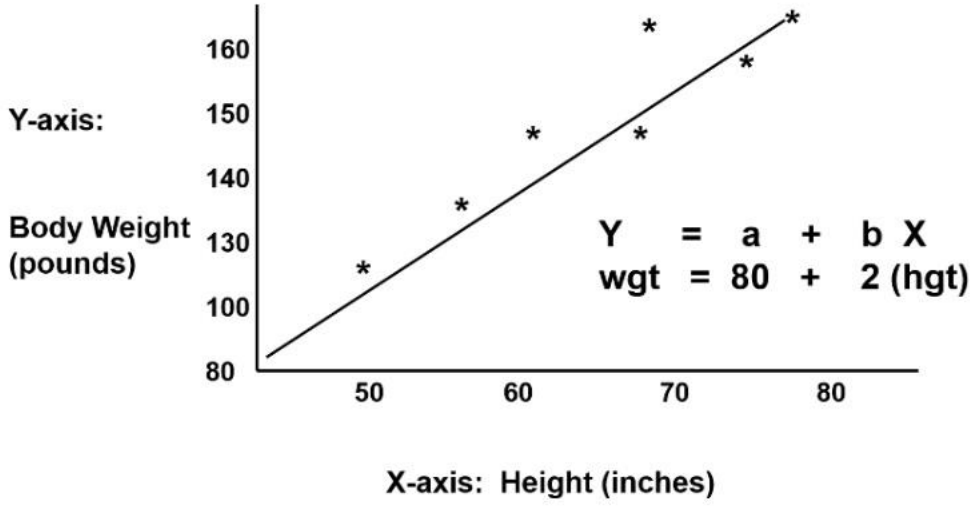
a linear equation that can be used to predict the values of the dependent variable based on the values of the independent variables. This equation is expressed as:

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$

Where y is the dependent variable, β0 is the intercept term, β1, β2, …, βn are the coefficients for the independent variables x1, x2, …, xn respectively. The slope β indicates how much the line rises for each increase in x. In case of positive values, the slope is upward and vice versa. The task of the machine learning algorithm is to determine the values of a(β0) and β so that the line provides the best relation between x values and y values (Lantz, 2015).

The principles of linear regression are based on the assumption that there is a linear relationship between the dependent variable and the independent variables. This relationship can be positive, negative, or non-existent. The linear regression model is developed by estimating the values of the regression coefficients (β0, β1, …, βn) that best fit the data and minimize the difference between the predicted values and the actual values.

As an example we can assume that height is the only factor which determines the body weight. If we plot height which is also an independent variable, as a function of body weight which is a dependent one, we may notice a linear relationship shown on the picture below:



(Sullivan, 2016)

From the picture above we can see that as the height in inches increases, the body weight in pounds also increases. For example, with the height of 60 pounds, the body weight is estimated at the level of 135 pounds.

There are two types of linear regression: simple linear regression and multiple linear regression. Simple linear regression involves only one independent variable, while multiple linear regression involves two or more independent variables. Here are some pros and cons of the regression:

Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

(Lantz, 2015)

Linear regression is a simple and intuitive statistical model that is easy to explain and understand, making it a useful tool for communicating results to a wider audience. It is also a computationally efficient technique that can handle large datasets with many variables. This makes it a useful tool for modeling complex relationships. Linear regression can be used for both simple and complex models by adding additional variables or interactions between variables. Although, as we can see from the table, there are some cons too.

Linear regression assumes that the relationship between the independent and dependent variables is linear. If the relationship is non-linear, then the results of the regression may not be accurate. It is sensitive to outliers, which can have a disproportionate impact on the model results. It also assumes that the variance of the residuals is constant across all values of the independent variables. If this assumption is violated, then the results of the regression may not be accurate.

Multiple regression formula was shown above and here is the simple one with alpha being the intercept, beta indicating the change in y given an increase of x:

$$y = \alpha + \beta x$$

(Lantz, 2015)

The accuracy and reliability of a linear regression model can be evaluated using statistical measures such as mean squared error (MSE), R-squared (R2), and adjusted R-squared (R2adj). MSE measures the average squared difference between the predicted and actual values, while R-squared measures the proportion of variance in the dependent variable that is explained by the independent variables. Adjusted R-squared takes into account the number of independent variables in the model and adjusts R-squared accordingly.

Linear regression is widely used in various fields such as finance, economics, and healthcare to make predictions and estimate the relationship between variables. However, it is important to

note that linear regression has certain assumptions that must be met for accurate and reliable results. These assumptions include linearity, independence, normality, and equal variance.

Linearity - the relationship between the dependent variable and independent variables is linear. This means that the change in the dependent variable is proportional to the change in the independent variables.

Independence - the observations are independent of each other, which means that there is no correlation or association between them. If there is a relationship between the observations, it can lead to biased and unreliable results.

Homoscedasticity (equal variance) - the variance of the residuals (the difference between the predicted and actual values of the dependent variable) is constant across all values of the independent variables. In other words, the spread of the residuals should be roughly the same throughout the range of the independent variables.

Normality - the distribution of the residuals is normal. This means that the residuals are symmetrically distributed around zero and follow a bell-shaped curve.

In the development of a linear regression model to predict medical expenses, the Ordinary Least Squares (OLS) algorithm is a commonly used method. OLS is a statistical method that aims to minimize the sum of the squared differences between the predicted values of the dependent variable and the actual values. The OLS algorithm estimates the regression coefficients that best fit the data and create a linear equation that can be used to predict future medical expenses based on the independent variables.

In case of the multiple regression equation, the dependent variable is comprised of the combination of the intercept a and the product of the estimated β and the x values for each of the I features. Epsilon represents the error (residual):

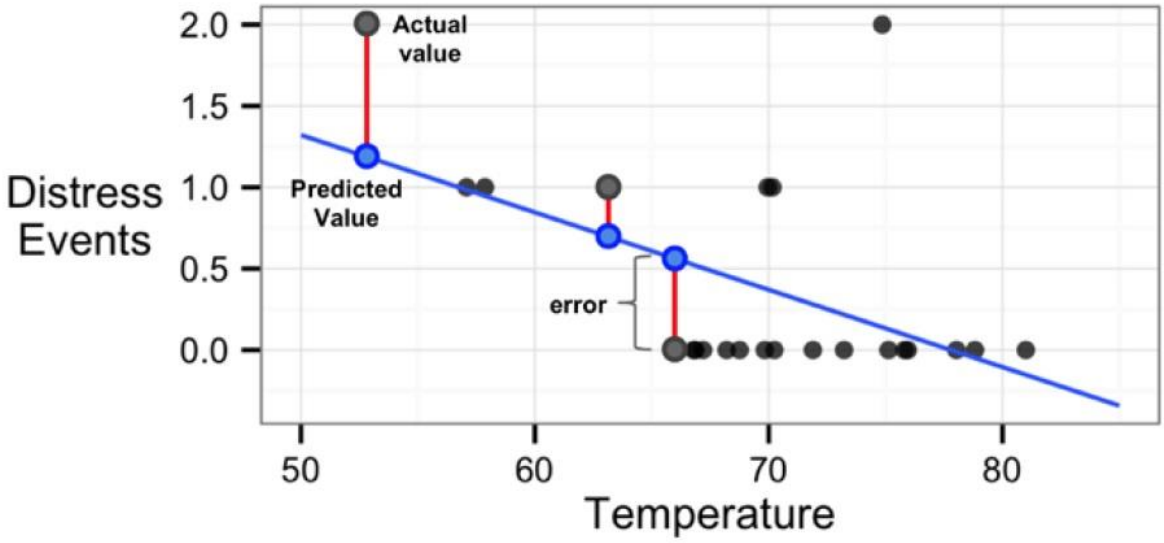
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i + \varepsilon$$

Since we have multiple values, the dependent variables are shown as a vector with each row indicating each example, independent variables are collected into a matrix with columns for features and intercept terms. Coefficients β and residual errors are also put into vector form. Finally we need to solve for the vector of regression coefficients which is indicated by β, it makes the sum of the squared errors between predicted and actual errors:

$$\hat{\beta} = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-1}\mathbf{X}^\mathbf{T}\mathbf{Y}$$

Here we need to transpose the matrix X (indicated by T) and calculate the inverse of matrix (negative exponent).

The OLS algorithm requires the specification of a linear model that explains the relationship between the dependent variable (medical expenses) and one or more independent variables (such as age, gender, and pre-existing medical conditions). The algorithm then estimates the regression coefficients using the available data to minimize the sum of squared errors between the predicted values and the actual values. The OLS algorithm assumes that the residuals (the difference between the actual values and the predicted values) are normally distributed and have constant variance across the range of the independent variables. On the picture below, we can see the residuals as red-colored lines indicating the errors.



(Lantz, 2015)

From a mathematical perspective, the OLS algorithm is trying to minimize the equation:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

The errors between the actual and predicted values are squared and summed throughout the data. The symbol caret ^ is used to indicate the estimate of the value. In order to solve for a we need to get the value of b:

$$a = \bar{y} - b\bar{x}$$

The bar symbol above the letters indicates the mean value. As for the b value, it results in the minimum squared error:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

After simplifying the above formula, we can obtain Variance and Covariance formulas from the numerator and denominator:

$$\mathrm{Var}(x) = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\mathrm{Cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Finally, after dividing the covariance function by the variance function, an updated formula for b looks like:

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

The principle of correlation between variables in linear regression is the idea that there is a linear relationship between the independent variable and the dependent variable. In other words, the change in the dependent variable is proportional to the change in the independent variable.

Correlation between variables is measured using the correlation coefficient, which ranges from -1 to 1. A correlation coefficient of 0 means that there is no correlation between the variables, while a correlation coefficient of 1 or -1 means that there is a perfect positive or negative correlation, respectively. In linear regression, we are interested in the correlation between the independent variable and the dependent variable.

If there is a strong correlation between the independent variable and the dependent variable, then linear regression is an appropriate model to use. This is because linear regression assumes that the relationship between the variables is linear. If there is little or no correlation, then linear regression is not appropriate as it assumes a linear relationship.

It is important to note that correlation does not imply causation. Even if there is a strong correlation between the variables, it does not necessarily mean that one variable causes the other. It is important to use causal inference methods to establish causality. Here is the formula of Pearson's correlation with rho being correlation statistic and sigma showing standard deviation:

$$\rho_{x,y} = \operatorname{Corr}(x, y) = \frac{\operatorname{Cov}(x, y)}{\sigma_x \sigma_y}$$

The OLS algorithm is widely used in linear regression because it is easy to understand, interpret, and implement. It provides unbiased estimates of the regression coefficients, and its assumptions are relatively easy to verify. In this report, we will use the OLS algorithm to develop a linear regression model that predicts medical expenses based on relevant factors. We will also evaluate the performance of the model using statistical measures such as mean squared error and R-squared to ensure the accuracy and reliability of our predictions.

Step 2. Collecting data.

We used a fictional data set called insurance.csv which was specifically created for the purpose of this project using non-real medical expenses information. The data contain 1,388 records of people signed up for the insurance listing various features such as age, sex, bmi, children, smoker status, region, expenses. Depending on the features each of the patients has the total expenses for the insurance plan was indicated for the year. Age feature is described as an integer up to 64 years. Sex values are indicating males or females accordingly. BMI (Body mass index) values give a picture regarding a person's physique by showing the relation between a weight and a height. The ideal values would fall within 18.5 and 24.9. Children feature indicates the number of offspring and dependents. Smoker status would show whether a person smokes or not and region would show the geographical locations of the person within the US falling into categories of northeast, southeast, southwest, and northwest. Each one of these features is directly related to the dependent variable which is medical expenses in this case so a reader may preliminarily draw conclusions whether the features may positively or negatively affect the outcome.

Step 3. Exploring and preparing the data

First of all, we used read.csv() function to load the data and applied stringsAsFactors = TRUE option to convert nominal variables into factors needed for our algorithm:

```
> insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
> str(insurance)
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 $
 $ bmi     : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 .$
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2$
 $ expenses: num  16885 1726 4449 21984 3867 ...
```

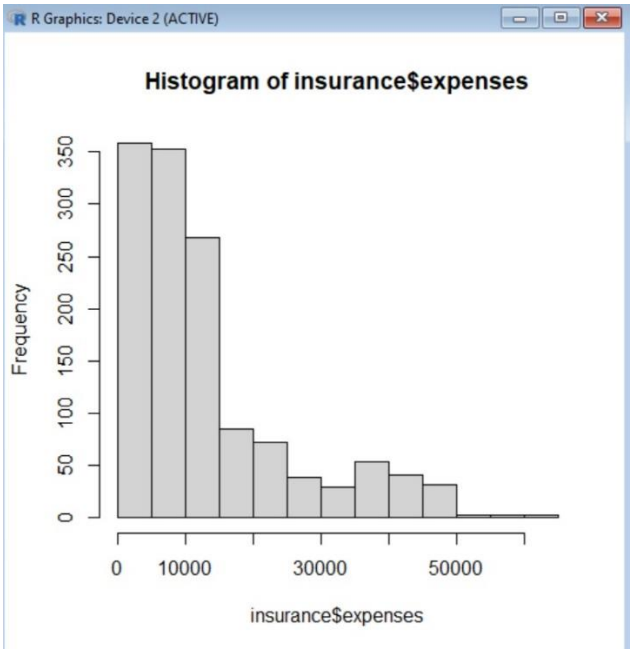
On the picture above we can see all the features mentioned earlier as well as the type of the files (integers, numeric, factors) with some examples of the values for each of the features.

Before proceeding to the algorithm stage, it is a good practice to confirm the normality of the variables which may fit the model better:

```
> summary(insurance$expenses)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4740    9382   13270   16640   63770
```

We can see that since the Mean value is larger than the Median one, the distributions of values for the expenses is skewed to the right. We can also see it with the help of a histogram:

```
> hist(insurance$expenses)
```



We can also imply other information from the histogram besides the skewness. We can see that most of the people in the dataset have their expenses limited by $15,000 even though the extreme values extend above $50,000. It is not a major disadvantage to our model; however, it may detrimentally affect our model. Same thing goes for the type of the values in question since three out of seven features in our dataset are of factor type and not numeric and regression model requires the latter type to be used. We also had to look into the distribution of data within each of the features having more than one type of value. For example, the region:

```
> table(insurance$region)

northeast northwest southeast southwest
      324       325       364       325
```

Here we could see that that values were distributed quite evenly.

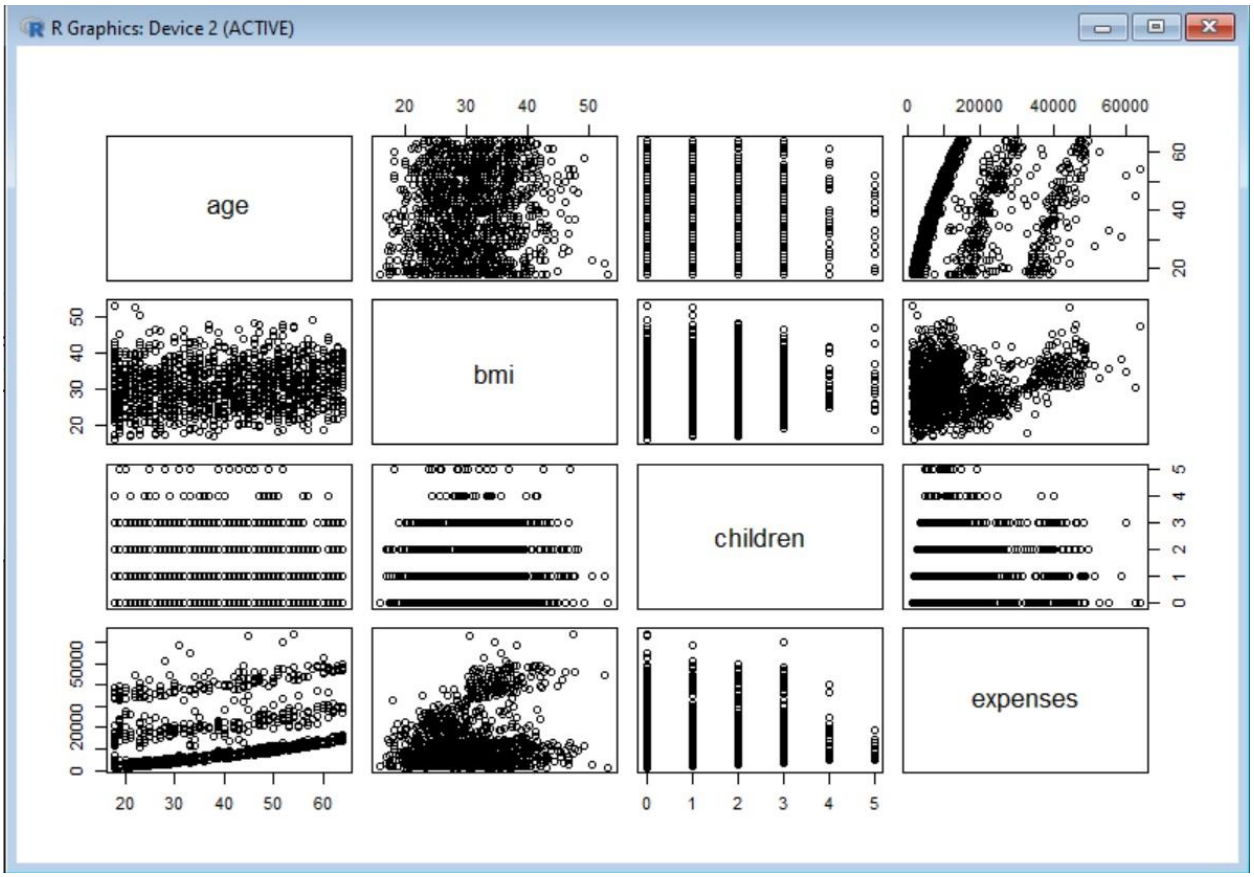
In order to track any correlation between the dependent variables before proceeding to building the algorithm we created a correlation matrix with the help of cor() function which demonstrated the values of correlation for each pair of relationships:

```
> cor(insurance[c("age","bmi","children","expenses")])
               age        bmi   children   expenses
age      1.0000000 0.10934101 0.04246900 0.29900819
bmi      0.1093410 1.00000000 0.01264471 0.19857626
children 0.0424690 0.01264471 1.00000000 0.06799823
expenses 0.2990082 0.19857626 0.06799823 1.00000000
```

The intersection of each row and column shows the correlation for each pair of relationships. 1.00000 indicates a perfect correlation between the variable and itself. It can be observed that none of the pairs tend to have very strong correlations. Although age and expenses seem to have the strongest positive correlation (0.299) among the pairs indicating that as the person matures, his cost of insurance increases. There also some other weak positive correlations between age and bmi, children and bmi, and age and children.

Another way of illustrating and visualizing the data is using a scatterplot. It does show a plot for each possible relationship although in case of a large number of features, this process may require a lot of resources. Much easier solution is provided with a scatterplot matrix which combines scatterplots for all relationships in one picture. It can help to detect trends and relations among the data even though it examines only two features at a time. Here is the picture of a combined scatterplot for all four numeric features in our dataset:

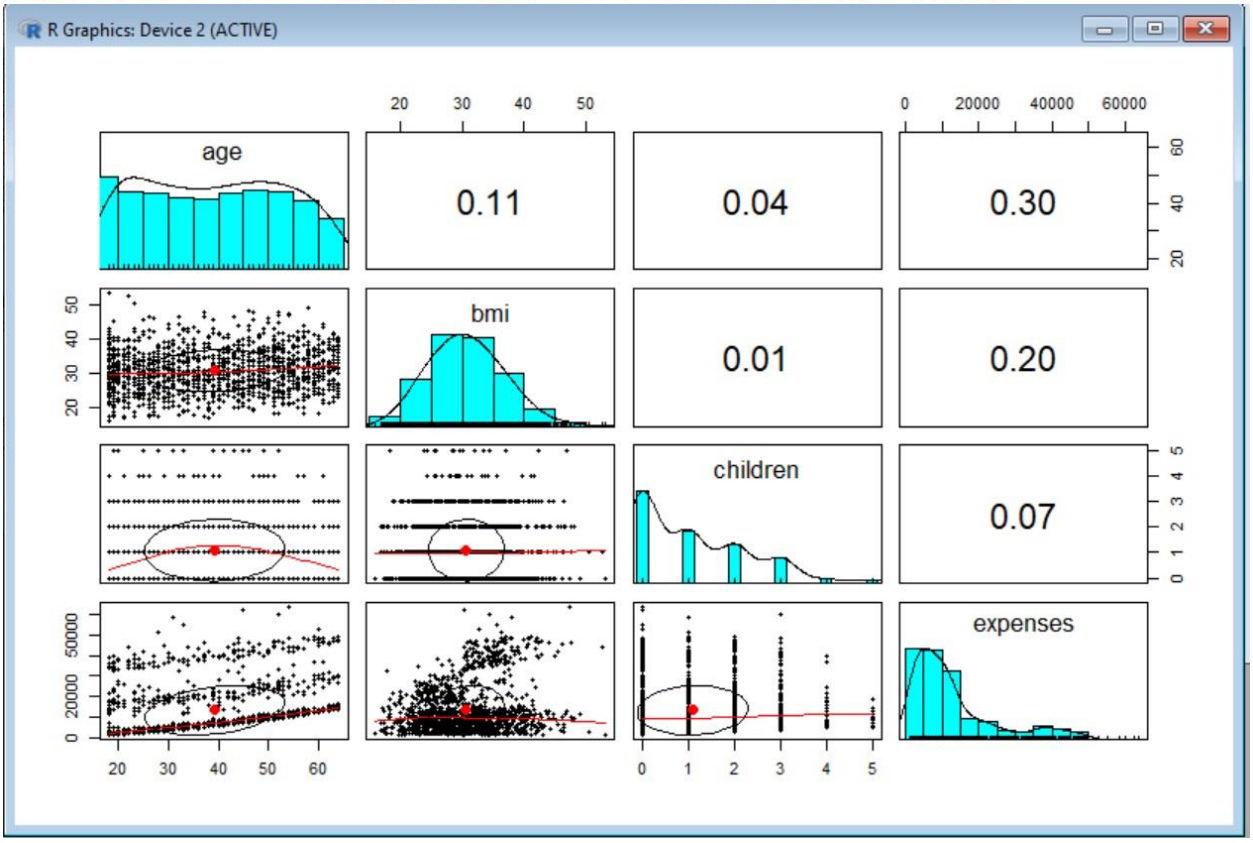
```
> pairs(insurance[c("age","bmi","children","expenses")])
```



Just like with the correlation matrix, a scatterplot matrix shows an illustrated relationship between variables at the intersection of each row and column. We mentioned a slight correlation between the age and expenses earlier, so here some straight lines are also observed on the scatterplot. At the bmi and expenses relationship we can see several groups of points which might also indicate some sort of relationship.

In case there is a more extensive and detailed information needed on the scatterplot, a pairs.panels() function can be utilized from the psych package:

```
> pairs.panels(insurance[c("age","bmi","children","expenses")])
```



Here we can see a combination of different metrics such as correlation matrix, histograms, visual information. An ellipse presented in the visual part of the scatterplot is called a correlation ellipse and it indicates the magnitude of the correlation with the point at the center signaling the point at the mean values for the x and y axis variables. The stronger correlation between the values the more stretched out the shape of the ellipse and vice versa. For instance, such relationships as age and bmi, age and children, age and expenses, and children and expenses are stronger that a bmi and children one. The loess curve shown on the scatterplot demonstrates a relationship between x and y variables. In case it is a straight line like with bmi and children, there is no much of a change of number of children in case of bmi increase. On the contrary, a downward curve on an age and children couple means that middle aged people tend to have more kids than young and elderly ones.

Step 4. Model training.

In order to create a model with our dataset, we used an lm() function which is part of the stats package in R. Here is the detailed syntax of the lm() function:

Multiple regression modeling syntax

using the `lm()` function in the `stats` package

Building the model:

```
m <- lm(dv ~ iv, data = mydata)
```

- `dv` is the dependent variable in the `mydata` data frame to be modeled
- `iv` is an R formula specifying the independent variables in the `mydata` data frame to use in the model
- `data` specifies the data frame in which the `dv` and `iv` variables can be found

The function will return a regression model object that can be used to make predictions. Interactions between independent variables can be specified using the * operator.

Making predictions:

```
p <- predict(m, test)
```

- `m` is a model trained by the `lm()` function
- `test` is a data frame containing test data with the same features as the training data used to build the model.

The function will return a vector of predicted values.

Example:

```
ins_model <- lm(charges ~ age + sex + smoker,
                data = insurance)
ins_pred <- predict(ins_model, insurance_test)
```

(Lantz, 2015)

The syntax of the command consists of the lm() name of the function, dependent variable being the first variable in the parenthesis, tilde symbol indicating the following description of the model, and independent variables separated by + sign or . if all the independent variables have to be used. The intercept is assumed by default:

```
> ins_model <- lm(expenses ~ ., data = insurance)
> ins_model

Call:
lm(formula = expenses ~ ., data = insurance)

Coefficients:
    (Intercept)              age          sexmale              bmi
       -11941.6            256.8           -131.4            339.3
       children        smokeryes  regionnorthwest  regionsoutheast
          475.7          23847.5           -352.8          -1035.6
regionsouthwest
         -959.3
```

The intercept showing -11941.6 does not refer to a number which could be applied in real world, because this is the value which is created when all other independent variables are equal to zero. In our case, there could be no person with age 0 or sex 0. Other beta coefficients indicate the estimation of increase or decrease of the expenses depending on the variable. For example, for each extra bmi value, the expenses are increased by $339.3, males have $131.4 less financial expenses than females, each year brings additional $256.8 of expenses, etc. We can also notice that some of the variables were not shown in the list due to the concept called dummy coding. It is automatically applied by lm() function to each factor variable. This concept allows the function

to consider a nominal variable as a numeric and creates a binary variable for each category. The binary variable is either 1 or 0 respectively. For example, smoker status would be split into smokeryes and smokerno variables with a real smoker status for a person making him smokeryes = 1, and smokerno = 0. Same principle is applied for variables with more categories such as the region. We can see three regions mentioned in the summary: regionnorthwest, regionsoutheast, regionsouthwest. One of the categories is always left as a reference and this is the reason why we can't see any sexfemale, smokerno and regionnortheast categories. The beta estimates are calculated relatively to the reference categories which are selected automatically by the model. By default, the first level of the factor is chosen as a reference although if needed it can be changed manually. Region section shows the negative beta coefficients which implies, the northeast region is having the highest average expenses. We can suggest from the model, that the results are showing some reliable information since the age, people physique and smoking status may detrimentally affect the health and consequently drive the costs up. Although, a more thorough evaluation of the model is required.

Step 5. Model evaluation.

In order carry out the evaluation of the model we used the summary() function:

```
> summary(ins_model)

Call:
lm(formula = expenses ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11302.7  -2850.9   -979.6   1383.9  29981.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11941.6      987.8 -12.089  < 2e-16 ***
age                256.8       11.9  21.586  < 2e-16 ***
sexmale           -131.3      332.9  -0.395 0.693255
bmi                339.3       28.6  11.864  < 2e-16 ***
children           475.7      137.8   3.452 0.000574 ***
smokeryes        23847.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -352.8      476.3  -0.741 0.458976
regionsoutheast  -1035.6      478.7  -2.163 0.030685 *
regionsouthwest   -959.3      477.9  -2.007 0.044921 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

I

II

III

We can split this summary into three sections:

I. The first section is showing a breakdown of the residuals which indicate the difference between the actual and predicted values. The maximum value here shows 29981.7 which is quite a large offset in $, however since the half of all residual fall between 1 and 3 quartiles, here the errors were ranging from -$2850.9 and $1383.9 which is reasonably acceptable.

II. In the second section, there is an important p-value indicated by the Pr(>|t|) which gives an idea of the probability that a true coefficient is zero provided the value of the estimate. Hence, small p-values such as in age or bmi rows indicate that true coefficients are having very small chances to be zero therefore the feature most probably is related to the dependent variable. Three asterisks indicate the significance level met by the estimate. It means that those marked with the symbols are statistically significant. If the model is having enough of those values, it means the model can be treated as quite a predictive model for the outcome. Otherwise, it might be an issue. In our case, we have age, children, bmi and smokeryes values.

III. The last sections illustrates the multiple R-squared value or the coefficient of determination. It gives a generic picture of how representative our model is when explaining the dependent variables. The closer the value is to 1, the better the model is explaining the variables. Adjusted R-squared is showing 0.7494 which is almost 75% of the data being explained by the model. Adjusted R-squared value adjusts R-squared by correcting models with a wide range of independent variables since more variable tend to bring more variations.

Considering the evaluation above, the performance of the model is reasonably consistent having a value of 0.75 for an R-squared. Nevertheless, we tried to improve the performance of our model in the next steps.

Step 6. Model optimization.

By knowing the subject of our data, we could potentially improve the model's performance because in case of a linear regression, features are selected by the user.

Since with aging expenses for the healthcare are being increased, the relationship between variables is not always linear. Thus, to change from a linear relationship to a non-linear one, we can adjust the following equation by adding a higher order term and treat the model as a polynomial:

$$y = \alpha + \beta_1 x \quad \Rightarrow \quad y = \alpha + \beta_1 x + \beta_2 x^2$$

(Lantz, 2015)

As the result the impact of age is measured as a function of age squared.

With the help of R we had to create a new variable and add it to the model:

```
> insurance$age2 <- insurance$age^2
```

Another thing we managed to do was to convert bmi values into binary indicators because it may either have no effect on expenses considering a healthy lifestyle of a person, or it may have a very negative effect in case bmi is above 30. In this case, a binary indicator of 1 would show the bmi values of 30 and above and 0 for if it's less than 30. Eventually, it will help us to see the average impact on expenses for people with a normal bmi and above the norm level. In order to do that, we used ifelse() function:

```
> insurance$bmi30 <- ifelse(insurance$bmi >= 30,1,0)
```

We can either keep both of bmi's into our model or just one, a common practice being to keep both.

Finally, we can use the interaction effect which describes a combination of effects for several variables in comparison with the effect of just one. For instance, smoking and high level of bmi may have a larger detrimental effect on the expense than each on of these factors separately. In order to create an interaction we had to use * between the variables in question and R will return with : indicating the interaction. The final adjusted model then looked like this:

```
> ins_model2 <- lm(expenses ~ age + age2 + children + bmi + sex + bmi30*smoker + region, data = insurance)
> summary(ins_model2)

Call:
lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-17297.1  -1656.0  -1262.7   -727.8  24161.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       139.0053  1363.1359   0.102 0.918792
age               -32.6181    59.8250  -0.545 0.585690
age2                3.7307     0.7463   4.999 6.54e-07 ***
children          678.6017   105.8855   6.409 2.03e-10 ***
bmi               119.7715    34.2796   3.494 0.000492 ***
sexmale          -496.7690   244.3713  -2.033 0.042267 *
bmi30            -997.9355   422.9607  -2.359 0.018449 *
smokeryes       13404.5952   439.9591  30.468  < 2e-16 ***
regionnorthwest  -279.1661   349.2826  -0.799 0.424285
regionsoutheast  -828.0345   351.6484  -2.355 0.018682 *
regionsouthwest -1222.1619   350.5314  -3.487 0.000505 ***
bmi30:smokeryes 19810.1534   604.6769  32.762  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

We can see that we could obtain a significant improvement with our model by looking at the R-squared value which was increased up to 0.8664 from 0.7509. The adjusted R-squared was also improved from 0.7494 to 0.8653. Around 87% of the variation of the expenses is now explained by the model. It also means that our suggestions with the interaction and non-linearity of age were correct. If a person smokes it would add $13,404.5952 to his medical expenses, however if in addition to his bad habit he is also overweight, he will need to add $19810.1534 each year.

Step 7. Recommendations

Based on our analysis, we recommend:

- Using linear regression as a predictive model for medical expenses. R is a powerful tool for conducting such analysis, and we encourage the use of R programming language for this purpose.
- Our analysis showed that age, gender, BMI, and smoking status were significant predictors of medical expenses. We suggest including these variables in the predictive model to improve accuracy. Additionally, we recommend collecting data on other factors such as family medical history, lifestyle, and occupation to further improve the model's accuracy.
- To improve the model's accuracy and minimize the risk of overfitting, we suggest using cross-validation techniques, adding non-linear relationship and adding interactions.
- Finally, we recommend regular updates to the model as new data becomes available. This will ensure that the model remains accurate and relevant over time.

Conclusion

In conclusion, our analysis of predicting medical expenses using linear regression with R has shown that it is an effective method for predicting healthcare costs. We have identified age, gender, BMI, and smoking status as significant predictors of medical expenses and recommended including these variables in the predictive model to improve accuracy.

Furthermore, we have recommended the use of cross-validation techniques, adding non-linear relationship and adding interactions to prevent overfitting and improve the model's accuracy. Collecting data on additional factors such as family medical history, lifestyle, and occupation can further improve the model's accuracy.

We believe that this predictive model can be of great benefit to healthcare providers and insurance companies, enabling them to make better-informed decisions about healthcare costs and plan accordingly. By regularly updating the model with new data, it can remain accurate and relevant over time.

In conclusion, we highly recommend the use of linear regression with R for predicting medical expenses. Its accuracy and effectiveness can help healthcare providers and insurance companies manage costs and provide better care to their patients.

References

1. Brett Lantz, Machine Learning with R, 2nd Ed., Packet Publishing, 2015 (ISBN: 978-1-78439-390-8)
2. Lisa Sullivan, Wayne W. LaMorte, Boston University of Public Health, Simple Linear Regression, May 31, 2016
3. Sanford Weisberg, Applied Linear Regression, Third Edition, 2005

Coding part

```
insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)

str(insurance)


summary(insurance$expenses)

hist(insurance$expenses)


table(insurance$region)


cor(insurance[c("age","bmi","children","expenses")])


pairs(insurance[c("age","bmi","children","expenses")])


install.packages("psych")

library(psych)


pairs.panels(insurance[c("age","bmi","children","expenses")])


ins_model <- lm(expenses ~ ., data = insurance)

ins_model

summary(ins_model)


insurance$age2 <- insurance$age^2


insurance$bmi30 <- ifelse(insurance$bmi >= 30,1,0)


ins_model2 <- lm(expenses ~ age + age2 + children + bmi + sex + bmi30*smoker + region, data =
insurance)

summary(ins_model2)
```