The University of Akron

College of Business, Department of Management

Advanced Data Analytics Topics (ISM:663)

Project 6

Predicting Medical Expenses using Ordinary Least Square Estimation

By:

Parwan Sharma

Student id – 4808380

MSM Business Analytics

Abstract

This Project is based on the use of Ordinary Least Squares (OLS) algorithm which is one of the most well-known and commonly used linear regression algorithms. It is the most widely implemented algorithm to make predictions, model building, casual inference, and hypothesis testing.

The data used in this project is called “insurance.csv”, a dataset based on the hypothetical medical expenses for patients in the United States which approximately reflects real- world conditions. It includes a total of 1338 examples of beneficiaries enrolled in an insurance plan.

In this project we will use this data and perform an analysis for the same using ordinary least square (OLS) estimation and find which independent variables (such as sex, age, bmi, etc.) can be good predictors of an estimate of the average medical expense for the insured population. We will prepare and handle data, train the model and then evaluate and improve the performance.

Introduction

Understanding Regression

Based on statistics, Regression can be explained as a method to understand the relationship between one or more predictors known as independent variables and a single numeric predicted value known as dependent variable. While making a model, the machine uses the slope intercept format, that is y = a + bx. In this equation ‘x’ and ‘y’ are independent and dependent variables respectfully with ‘b’ as the slope or the steepness of the graphical representation of the equation and ‘a’ is the intercept. The job of the model in this case is to identify the values of ‘a’ and ‘b’ so that graphically the available values of ‘x’ best relate to values of ‘y’.

From a statistical hypothesis testing view, we use regression to determine whether a premise is likely to be true or false based on the observed data. It is used to make models on complex relationships estimation impact of various data parameters on an outcome. Some areas where regression analysis is highly used is population and individual characteristics, quantifying relationship between event and response, identifying patters for a forecast outcomes based on known criteria.

Regression analysis is the basis of many algorithms such as simple, and multiple linear regression, logistic regression, Poisson regression, etc.

Ordinary Least Square (OLS) Estimation

OLS is an estimation method which finds the optimal estimates the slope beta ‘b’ and the intercept alpha ‘a ‘. The aim of this estimation is to minimize the sum of squared residuals(error), which is the sum of the squared differences or the distance between the predicted ‘y’ values and the actual values of the dependent variable ‘y‘.

Chart, line chart, scatter chart

Description automatically generated

In order to make the most accurate prediction calculus is used to find the slope which results in the least squared error value. The mathematical equation is based on two components which are variance and covariance. Variance involves finding the average squared deviation from mean of x. The other part of the equation is covariance where you need to sum up the product of the difference between each data point's x value and the mean x value, and the difference between that point's y value and the mean y value.

Problem Description

The health insurance company makes its profit primarily by investing the money it generates, which is the difference between the money it collects from its customers, or the yearly premium and the money spent on the medical care of its beneficiaries. Hence insurance companies invest significantly in a better forecast for medical expenses of the insured people which can be used to determine an effective price for the yearly premium for the insurance.

Even though health crisis seems to happen randomly in an individual’s life. It seems that there are factors which can be taken into consideration to determine if an individual is likely to need medical help in the near future. For example, obesity is linked to a number of health issues, similarly old age is also generally a good predictor of high medical expenditure on nonspecific health issues. Certain lifestyle conditions such as smoking, and obesity usually result in lungs cancer and heart diseases respectfully.

The goal of this analysis is to estimate the average medical care expense for such a population to have a better forecast of the yearly premium depending on the expected treatment cost.

Objectives:

The primary goal of this report is to:

* To comprehensively introduce Ordinary Least Square (OLS) algorithm, explaining its mathematical foundations.
* Improving the model using different methods.
* Using visualizing tools to understand data and relation among variables.
* Outline the method involved in building and training the model.
* Propose recommendations for future research and development in this field.

Method

This project will be using the dataset “insurance.csv.” The primary source of literature used is “Machine Learning with R, by Brett Lantz, 2nd Ed., Packet Publishing, 2015 (ISBN: 978-1-78439-390-8)”.

Listed below are the steps taken in the report:

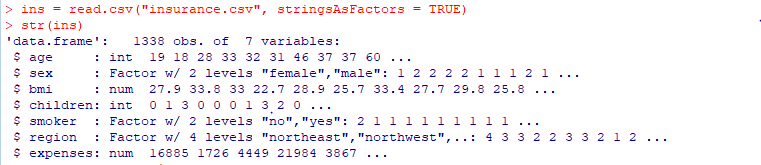
* Step 1 – Collecting data.
* Step 2 – Exploring and preparing the data.
* Step 3 – Exploring relationships among features.
* Step 4 – Visualizing relationships among features
* Step 5 – Training a model on the data.
* Step 6 – Evaluating model performance.
* Step 7 – Improving model performance.

Steps taken:

* *Step 1 – Collecting data.*

The data we use for this analysis is called “insurance.csv”. It is not real-world data and is created by the US Census Bureau to approximately reflect real-world conditions. It is a stimulated dataset which contains 1338 cases of hypothetical medical expenses for patients in United States along with features like age, sex, bmi, children, if a person smokes or not and region.

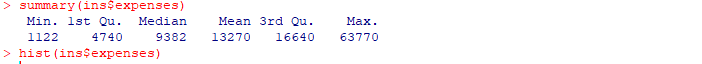
We use the read.csv () function to load the data for analysis. We will use the stringsAsFactors = TRUE, in order to convert the three nominal variables to factors. We check to see the changes made in the data using the str() function.

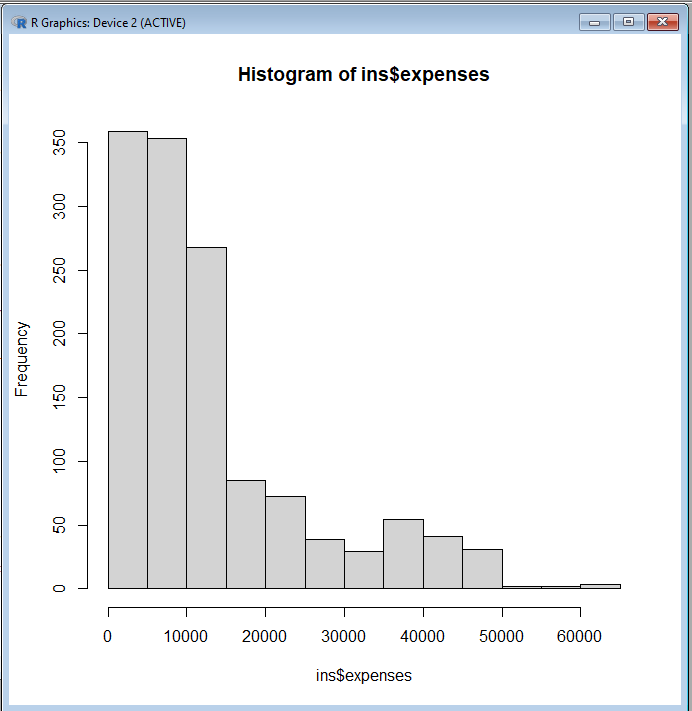


* *Step 2 – Exploring and preparing the data.*

The dependent variable for this case is expenses as it determines the medical cost a customer is charged for the insurance annually. Even though it is not required by the model to have a normally distributed data, we can always get insight to the data by doing so as a normally distributed data fits better in this model.

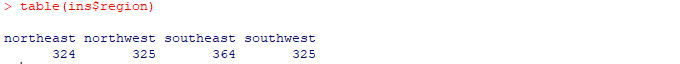
We can do that statistically by using the summary() function. Result shows that the mean value is more than the median value, making the distribution right skewed. We can confirm the sam graphically using the hist() command.





The graph shows that most of the people have annual medical expenses between zero and $1500. Knowing this weakness, we will address this issue by designing a better fitting model later. Another issue with the data is that there are three features (sex, smoker, and region) in the data which are factor type. A regression model requires all the features to be numeric in nature.

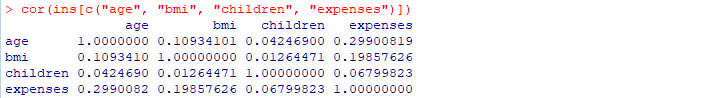
We check the region variable which shows that it is divided evenly into four regions.



* *Step 3 – Exploring relationships among features.*

Regression models work on the idea that there is a linear relationship between variables. Variables with high correlation can cause an issue as two variables will measure the same impact leading to collinearity making interpreting of the data difficult. This can produce an inflated standard error and a misleading p-value.

We can check the correlation between variables by using the correlation matrix. We would create the correlation matrix using the four numeric variables using the cor() command:



As shown above the diagonal will always be 1.0000000 or in other words, absolute correlation, as it is matching the same variables.

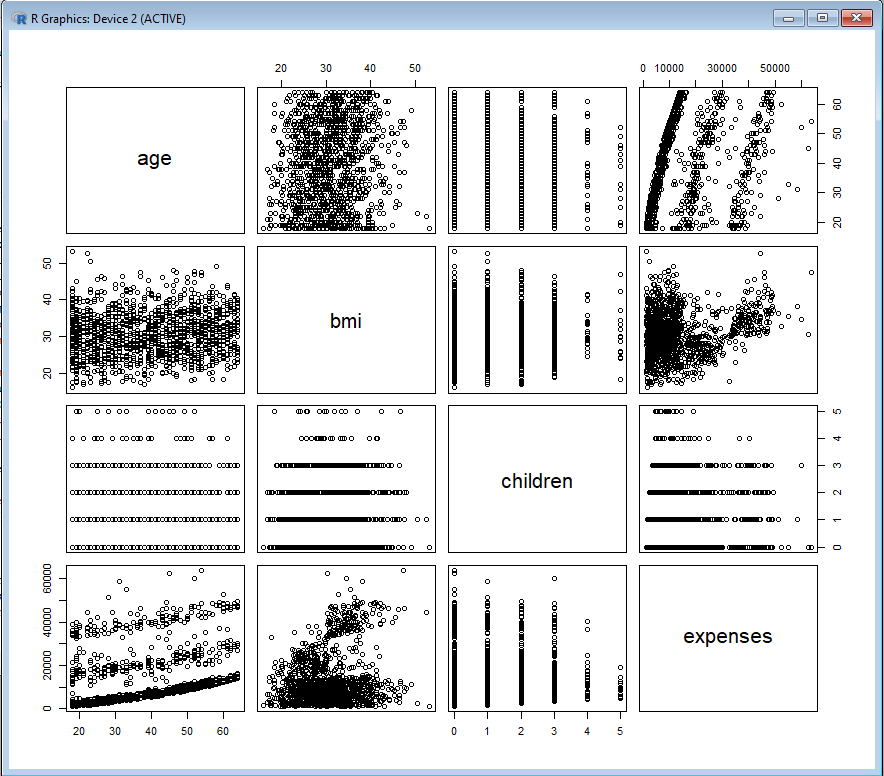
The result shows that there is no significant correlation between any variables. There is slight positive correlation between, age and bmi. This means as age increases, body mass tends to increase. Similarly, there is a moderate positive correlation between age, bmi and expenses, and children and expense. These suggest that when age, body mass, and number of children increase, cost of insurance also increases.

* *Step 4 – Visualizing relationships among features*

It is also helpful to visualize relationship among different features graphically. This could show us trends among the different features. We can do that by using a scatterplot, which is a good visualizing tool as it arranges two or more features in a grid and showing their reltionship graphically.

We would use the pair() function to produce a scatter plot matrices using the four numeric variables we have.

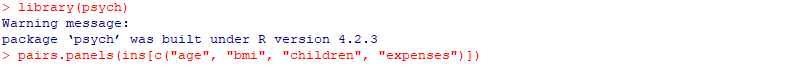




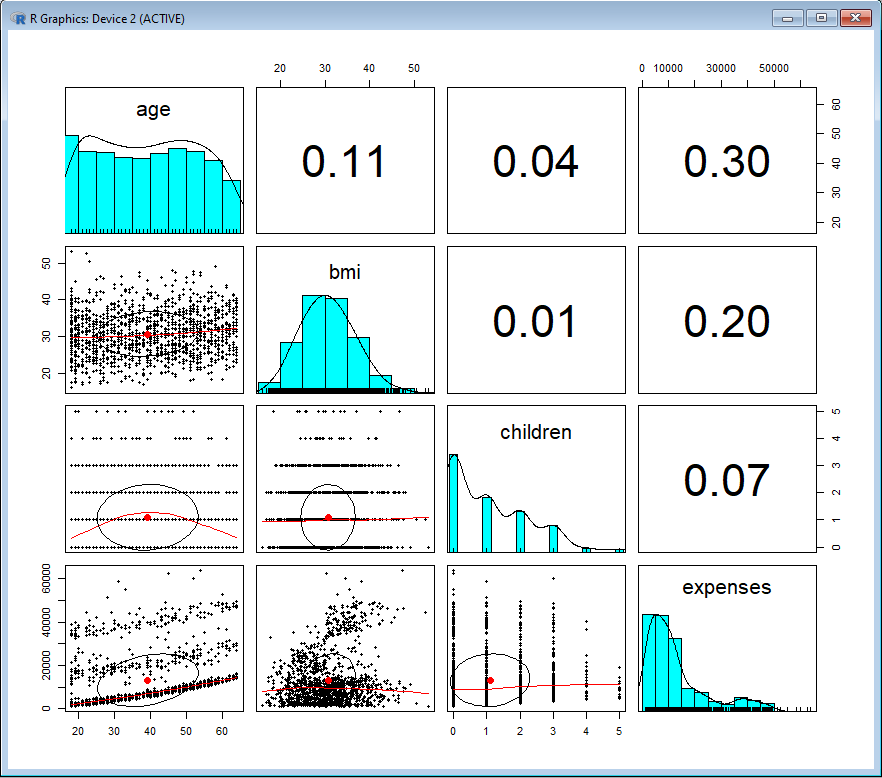
We can see some trends among some of the features. The relationship between age and expenses displays straight lines. This is potentially suggesting that as the age increases, the expenseon medical care also gradually increases.

Similarly, the bmi and expense also show two distinct groups. Suggesting that the expense is high for two groups of bmi.

We can add more information in this plot by making a enhanced scatterplot matrix using the pairs.panels() . To use this command, we will install the “psych” packages and load it in the library.



This produce an enhanced version of scatterplot with more information.



In this plot, the diagonal is replaced by more informative histogram which show the distribution of these features in the data. The two new factors in this correlation are the correlation ellipse and loess curve.

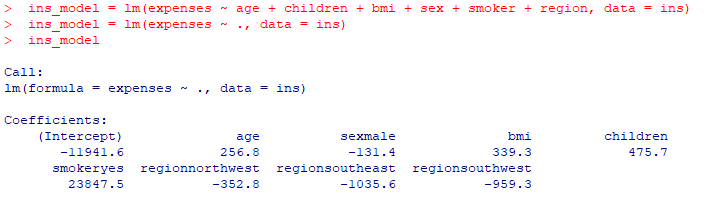
The correlation ellipse provides the correlation strength. That is the more the ellipse is stretched, the stronger the correlation is. In this case the ellipse is almost circular in nature as there is no significant correlation found between variables.

The Loess curves is an arch like structure in case of the relation between age and children. This shows that the youngest and oldest people in the sample have fewer children on the insurance plan which peaks around middle age. We also conclude the previously inferred point that age and bmi are weakely positively correlated.

* *Step 5 – Training a model on the data.*

We use the lm() function which is included in the stats package to fit the linear regression model into data. The command mentioned below will fir the model into the data while relating the six independent variables to the medical expense. The model intercepts are assumed by default.

This is followed by the . character command which will do the same thing as it will include all the features. After building the model we can see the estimated beta coefficients.



The coefficients indicate how the medical expense will increase if the value of the independent value is increased by 1. This suggest that the as the age, number of children and BMI increase by 1 unit, the medical expense would increase by $256.80, $475.70, $339.30.

The intercepts for variables such as age zero and bmi zero are ignored as there is no real-world interpretation of a case with medical expense when the age is 0.

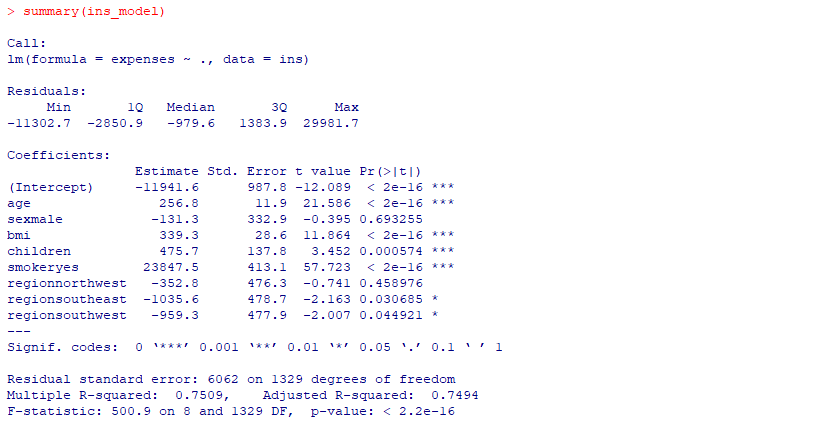
We also notice that the model gave us eight coefficients when we had six variables. This is because the lm() function used a process called dummy variables. In this process the dummy variable makes it possible for the nominal feature to be treated as numeric. This is done by setting the dummy variable as 1 and the other case is set as 0. This led to the creation of variables like sexmale = 0 and sexfemale = 1. Similar thing happens with regions.

Since one value is always left out as reference category, in this case R left out the smokerno, regionnortheast and sexmale making female non-smokers in the northeast region the reference group. Males have medical expenses that is $131.40 less when compared to females each year. Additionally, smokers incur an average of $23,847.50 more expenses than non-smokers annually. Furthermore, the negative coefficient for each of the three regions in the model suggests that the northeast region, which is considered the reference group, generally incurs the highest average expenses.

This shows that the old age , smoking and obesity tends to be linked to additional health issues.

* *Step 6 – Evaluating model performance.*

We use the summary () command to study the stored model. The output provide three key ways to evaluate performance.



1. Residual

A residual is calculated by true value by the predicted value. The model predicted a maximum error of 29981.7 which is a high underprediction. Also, half of the error falls between the 1Q and 3Q signifying majority of predictions were between $2850.9 over true value and $1383.90 under the true value.

1. P- values

A small p value signifies that coefficient is not likely to be 0, which means the features are not likely to have relationships with dependent variable that is the medical expense. P values less then significant are called statistically significant. These values are important for the model as absence of these would indicate that the features are not being used in the process of prediction.

1. Multiple and Adjusted r squared values

Multiple R squared value is also known as the coefficient of determination, which provides how well the model represent the dependent variables. Our R squared value is 0.75 which means the model explains 75% of the variation in the dependent variable. The adjusted R-squared measure the impact of the number of independent variables in a model by penalizing models with a higher count of such variables.

* *Step 7 – Improving model performance.*

The key difference between regression model and other is that regression leaves feature selection and model specification which is done in machine learning approaches. Hence if one has knowledge of the topic on which the model is based, we should make model specifications to improve model performance.

1. Adding non-linear relationships

It is not always necessary that the relationship between independent variables be linear. For example, the relation between age and medical expenditure is not always linear. There is a chance that the expense increases exponentially or is non-linear in general. To account for this issue, we can add a regression model as polynomial.

We add non-linear age by simply creating a new variable.

1. Converting numerical variable into Binary indicators.

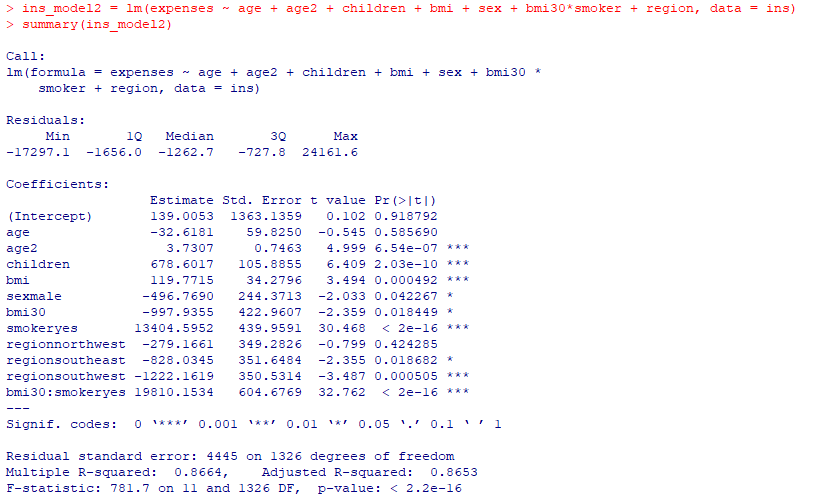
In this case we are using the idea that rather than taking the whole spectrum of BMI we should focus on only a particular range. This is because the medical expenditure is likely to be impacted if a person has high BMI or is obese. Hence, we set the BMI parameter using ifelse() command to more than or equal to 30 to return 1., or else will return 0.

1. Adding Interaction effects.

So far, we were studying the effects of one independent variable on the medical expense. If we combine two features which are likely to impact the medical expense. For example, we interact the obesity(bmi30) and the smoker indicator together, that is expenses ~ bmi30\*smoker.

When we put all the three improvements together and we put in the newly constructed variables. We use the summarize () command to see the models.





Based on the changes made in terms of the variable interaction we can see that the R- squared value is to be improved, that is it went from 0.75 to 0.87. This indicates that the model is now better at explaining the values of dependent values, that is the model now explains 87% of the variation. The same is the case for adjusted R-square, where its value improved from 0.75 to 0.87. The newly introduced age2 is also significant. Interaction between smoking and obesity suggests addition of cost. An obese smoker spends $19,810 whereas one who is obese spends $13,404.

Conclusion

Through this project we conclude that:

1. Linear regression is one of the most widely used modelling technique in which relationships among the features are typically specified by the user rather than being detected automatically.
2. Normalized data is not necessary but fits the model better.
3. Graphical representation such as histogram and scatterplot can lead to a better understanding of the data as it can show if the data is skewed or not and if there is collinearity or trends between independent variables. This knowledge can help us better implement changes in the model to improve the model.
4. You have evaluating parameters such as Residual, p-value , significance level, adjusted R- square and R squared values which can give you a metric of the model performance.
5. As previously mentioned, model can be improved by making specific changes based on the trends concluded from data visualization and inferences through quantitative analysis.

Limitations and Further improvements

Some limitations of the OLS algorithms are:

1. OLS assumes that there is a linear relationship between the dependent and independent variables. If not true, then the OLS model may not be accurate.
2. Model will be sensitive to outliers, which can impact on the coefficients and affect model accuracy.
3. Assumes that the variance of the error terms is constant across all levels of the independent variables.
4. It assumes that the independent variables are not highly correlated with each other.
5. It requires a large sample size to produce reliable results.
6. OLS assumes that the residuals are normally distributed.

Some potential improvements of the OLS algorithms are:

1. One way to address the assumption of linearity is by transform the independent variables using non-linear functions.
2. Using weighted regression to assigns different weights to different observations, based on their importance or reliability.

Coding

> ins = read.csv("insurance.csv", stringsAsFactors = TRUE)

> str(ins)

> summary(ins$expenses)

> hist(ins$expenses)

> table(ins$region)

> lines(hist(ins$expenses),col="red")

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

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

> install.packages("psych")

> library("psych")

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

> ins\_model = lm(expenses ~ age + children + bmi + sex + smoker + region, data = ins)

> ins\_model = lm(expenses ~ ., data = ins)

> ins\_model

> summary(ins\_model)

> ins$age2 = ins$age^2

> ins$bmi30 = ifelse(ins$bmi >= 30, 1, 0)

> ins\_model2 = lm(expenses ~ age + age2 + children + bmi + sex + bmi30\*smoker + region, data = ins)

> summary(ins\_model2)

Reference

* Machine Learning with R, by Brett Lantz, 2nd Ed., Packet Publishing, 2015 (ISBN: 978-1-78439-390-8)