Predicting Medical Expenses with Regression and Dimensionality Reduction

By: Miranda Gemme-Ellis

Table of Contents

Research Questions

Quantitative -
What are the critical factors that influence medical expenses in each population?
Can medical expenses be predicted based on demographic and health-related information?
Qualitative -
What are the important factors (demographic and health-related attributes) that influence the likelihood of a smoker?
PCR-
How can *we* predict medical expenses using demographic and health-related predictors, addressing multicollinearity issues?

Executive Summary

The goal of this project is to examine the health insurance dataset and conduct a detailed analysis comparing different variables such as age, gender, sex, region, medical expenses, BMI, and region. This analysis was used to understand how different factors can impact medical expenses. This can help create meaningful insights used to find correlations, make predictions, and make actionable recommendations for medical insurance companies, stakeholders, and patients. It was also used to determine how different factors can influence the likelihood of becoming/being a smoker.

The analysis revealed a positive correlation between medical expenses and age; as the patients' ages increase, medical expenses do, too. There is also a positive correlation between medical expenses and body mass index (BMI); as the patient's BMI increases, medical expenses tend to increase. The analysis suggested that the southeast region has more medical expenses than the rest of the country.

The recommendations to combat these correlations are as follows:

- BMI - It may be beneficial to offer helpful and healthy classes on decreasing one's BMI. For example, the insurance company could discount gym memberships or workout classes. Promoting healthy eating might benefit the insurance company; they could offer online cookbooks or classes or have a dietitian available to answer questions.
- Smoking – It would be beneficial for insurance to offer incentives to non-smokers and people who quit smoking. It might be helpful to implement smoking cessation programs and offer additional resources.
- Regional – Further research would be required to see what factors drive those medical expenses up in certain regions. One possibility could be retirees traveling to Florida for 7-8 months out of the year. However, further research is necessary.

Data and Approach

The data set used in the analysis has 1,338 observations and 7 variables relating to medical expenses. The variables in this dataset are medical expenses, age, body mass index (BMI), sex, number of kids, smoker (yes/no), and which region of the country the patient is located.

Re-engineering steps:

The data set was processed to ensure that there were not any missing values by using the any(is.na(insurance)) function. The numerical data, such as expenses, body mass index, age, and number of children, within the insurance data set had their own table to make it easier to do different processes that required numerical data only. There were two numerical datasets; one was called "num_data," which used the unlist( ) function to plot the data. The other dataset, called "numb_data," used the subset( ) function. The categorical data, such as sex (male/female), smoker(yes/no), and region (northeast, northwest, southeast, southwest), were encoded later in the project to help integrate them into different models. During the Principal Component Regression (PCR) and boosting processes, the expenses variable logarithm of expenses was used.

Overall Approach and Techniques:

The project used a structured approach that explored different models to uncover correlations and make predictions. The following techniques were used:

- Descriptive statistics, correlation analysis, and different graphs (histograms and boxplots) were created to gain insights into the distribution of different demographic variables.
- Logistic regression, chi-squared tests, and contingency tables were created to investigate smoking habits and demographic information.
- Linear regression, subset selection, and principal component regression (PCR) were used to observe how different factors influence medical expenses.
- F-statistics, model summaries, and mean squared error (MSE) were used to evaluate, measure, and identify the most effective predictive capabilities.
- Gradient boosting, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), logistic regression, and naïve Bayes were used to compare the performances of predicting different responses.

To reach the goals above and implement the techniques listed, multiple steps and processes were used:

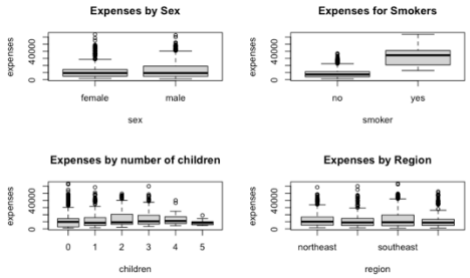
1. Ensure there are no missing values; if there are, the next step would be handling those missing values.
2. Examine the dataset by looking at the distribution of the variables, identifying outliers, and examining the overall characteristics and spread of the data.
3. Review the data to identify correlations.
4. Evaluate the performance and capabilities of the models by using mean squared error (MSE) and calculating the accuracy.

The analysis provides a deeper understanding of factors influencing smoking habits and medical expenses by uncovering correlations and patterns. A combination of traditional statistics and modern machine learning techniques were used, and they proved to be necessary for extracting essential information from the dataset. The information obtained from this analysis can be used to create an understanding and give strategic directions for insurance companies and patients.

Detailed Findings

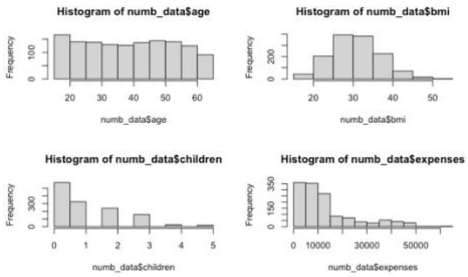
One of the first steps taken when analyzing data is to observe the dataset by looking at the distribution of the variables, identifying outliers, and examining the overall characteristics and spread of the data.

The box and whisker plots for the categorical variables showed some significant outliers. In the plot showing the expenses by number of children, there are significant outliers for 0-3 children. The plot showing the Expenses by Sex shows the most outlier for expenses in the data for females, but there are still some outliers for males. The Expenses for Smokers plot really demonstrates how much more medical expenses are for smokers. There are still some outliers for non-smokers, which may be due to other factors. There are outliers for each region.



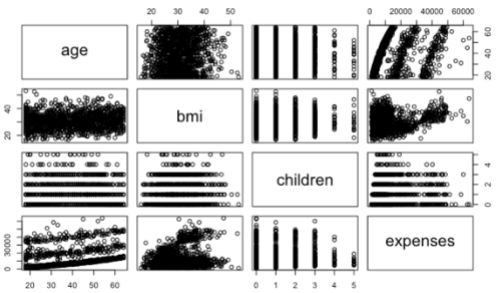
The box and whisker plots above show the characteristics of the data for the categorical variables. The y-axis shows the expenses, and the x-axis shows the categorical variables.

The histograms show the frequency and spread of the data. For instance, the histogram for expenses and the histogram for the number of children show that the data is skewed right. The histogram showing bmi is primarily symmetric or unimodal. The age histogram has a uniform spread.



The histograms above show the spread of the data. The y-axis is showing the frequency, while the x-axis is showing the numerical variable.

Next, a data set called "num_data" with unlisted numerical data from the insurance dataset was created. This was done to plot the numeric data together using the plot(insurance[,num_data]) code. The noticeable trends showed that as age increased, so did the expenses. There seemed to be three positive linear trends within that plot. The next trend that was noticed was between expenses and the number of children. There appeared to be a slight negative correlation between the number of children and expenses; however, that will later be deemed invalid.



The scatterplots on the left show correlation between different numeric variables within the dataset.

The next step was to calculate the correlation for the numerical data. This was done by using the cor( ) function. The data that has the highest correlation with expenses is age (0.299). Although this is not a very strong correlation, it is the highest out of all the numerical data. The second highest correlation regarding expenses is body mass index (0.199). The lowest/least correlated variable with expenses is the number of children (0.068).

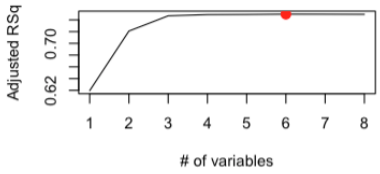
Multiple Regression Model(s):

First, a linear model was conducted with all the variables. Then, the more significant variables were used to construct a linear model with the predictor variable as expenses, and the response variables are age, BMI, and smoker (yes/no). The R^2 (0.75), the f-statistic (1316 on 3 and 1334 DF), and its associated p-value (<2.2e-16) provide evidence that the regression model is statistically significant. According to the linear model, the equation for expenses can be written as:
Expenses = - 11679.05 + 259.53 * age + 322.69 * bmi + 23822.61 * smoker (yes = 1, no = 0)

Best subset selection:

The r-squared values for the different model sizes are:

- Model with 1 predictor: 0.6198
- Model with 2 predictors: 0.7214
- Model with 3 predictors: 0.7475
- Model with 4 predictors: 0.7497
- Model with 5 predictors: 0.7501
- Model with **6 predictors: 0.7508**
- Model with 7 predictors: 0.7509
- Model with 8 predictors: 0.7509



As more predictors are added, the R-squared increases.

Backwards and Forwards:

The backward and forward selection also have the same R-squared as above, suggesting that the stepwise algorithm found an optimal model so both methods converge to the same or similar solutions.

Ridge Regression:

Applying log() when calculating MSE (mean squared error) for ridge regression can be crucial to eliminate outliers. Before using log(insurance$expenses), the MSE was 40899863; after using the log() function, the MSE was reduced to 0.211346. This suggests that the logarithmic transformation after the log() function was applied performs well when predicting the response variable. Using the log() transformation likely enhanced model performance because it addressed issues such as nonlinearity or heteroscedasticity within the data.

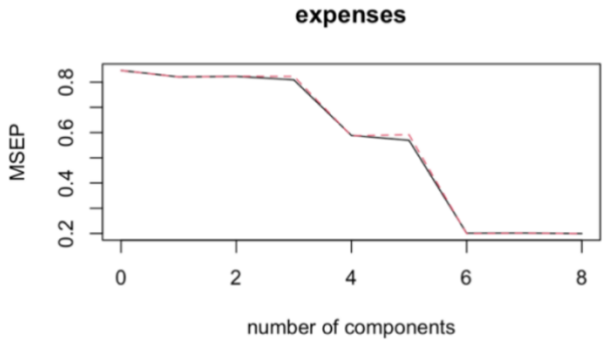
Lasso Regression:

Like ridge regression, the log( ) function was applied to calculate the MSE. Before using log( ), the MSE for lasso regression was 38792416; after using log(), the MSE for lasso regression was reduced to 0.18295. Lasso regression is used to observe the performance of the regression model.

The lasso regression model shows the average squares difference between the actual values in the data set and the predicted ones.
The lower MSE for lasso regression and ridge regression indicate a good fit.

Principal Components Regression:

The MSE for principal components regression came out to 0.36326, which implies a relatively low prediction error. This is the MSE when using the log() function on the insurance$expenses variable. The R-squared are cross-validated values and can help assess how well the model is generalizing the new data. In the cross validation and training percent variance explained, it is seen that as the number of components (comps) increases, the percent of variance that's explained also increases. For example, with 3 components, x (predictor) explains 49.6% of its variance, while the expenses (response) variable explains 13.18% of the variance. The predictor reaches 100% at 8 components.



The graph shown demonstrates the number of components (x-axis) and the corresponding MSEP value (y-axis).

PLSR:

The MSE for PLSR using the log() function on the insurance$expenses variable came out to 0.18295. This suggests a low prediction error. A lower value suggests that the model's predictions are a good fit.

Regression Tree:

The regression tree model was constructed using a 50/50 train-test split and initially produced a test mean squared error (MSE) of 186,832,377. After pruning, the tree became more efficient and interpretable, with the test MSE improving to 161,920,636. Key predictors identified included smoking status, BMI, and age.

Bagging:

Bagging was implemented through the Random Forest algorithm, significantly improved predictive performance. Using 500 trees and selecting three variables at each split, the bagged model achieved a much lower test MSE of 23,764,342, highlighting its effectiveness in reducing variance. Variable importance plots confirmed that smoker status and BMI were the most influential predictors.

Boosting:

Gradient boosting was used on log-transformed expenses to further refine prediction accuracy. Shrinkage tuning identified optimal learning rates, and test MSE decreased with proper regularization. Smoker status and age again emerged as the most impactful variables.

Logistic Regression:

For classification, logistic regression was used to predict smoking status based on region, sex, number of children, and BMI. The model's output was transformed into binary predictions, and overall classification accuracy was evaluated.

Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis (LDA) was applied to classify smoker status using region, expenses, and BMI as predictors. The model achieved a test accuracy of approximately 79.5%, indicating a strong linear separation between groups. Quadratic Discriminant Analysis (QDA), which allows for nonlinear boundaries, was also tested using similar predictors and achieved a similar test accuracy, suggesting that both models performed well on this dataset.

Reference(s):

Dataset:
Choi, Miri. (2017). Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

Book:
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An introduction to statistical learning (2nd ed.) [PDF]. Springer.

Packages Used:
glmnet, tree, leaps, randomForest, gbm, BART, pls, MASS, e1071