A comparison of standard and penalized linear regression models predicting medical insurance payment with non-cost inputs

Sadhana Dhokane [1]

[1]University of Michigan

Introduction and Background

Medical insurance plans are intended to reduce the cost of risk factors for these unexpected events [3] . However, medical insurance costs have risen excessively in recent years. Hence, accurate predictability and estimation of the appropriate cost of health insurance are critical to beneficiaries.

Ordinary least squares (OLS) estimation of projected clinical outcomes for risk adjustment reasons is still a mainstay of predicting medical insurance payments [1]. Even though there were lots of recent studies investigating machine learning applications to predict medical insurance costs, it requires sophisticated interpretation by an expert analyst, which has resulted in the slow adoption of advanced machine learning techniques in practice [2]. Therefore, the purpose of this study is to assess the effectiveness of penalized linear regression, which has the similar level of interpretability as standard regression, in estimating medical insurance payments with non-cost inputs.

Research Question and Method

Research Question:

- Can a high-performing linear regression model be built with only non-cost inputs?
- Which input feature is the most significant predictor in determining medical insurance payment?
- Between standard and penalized linear regression, which model best predicts future medical insurance payment?

Methods:

- **Exploratory Data Analysis (EDA):** With statistical summary, data visualization, and the ANOVA test, we examined data patterns and investigated which independent factors would have the most effect on predicting target variables.
- **Training and Testing Models :** A pipeline was designed to normalize numerical variables and generate dummy variables for category factors. After that, we started training and testing a standard and penalized linear regression model.

Data Description

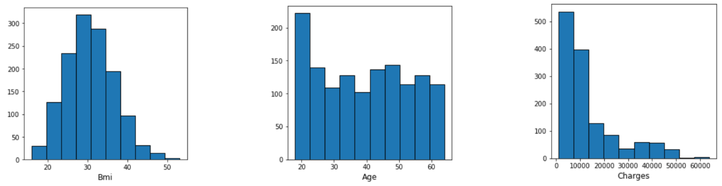
The dataset used was an open source dataset from ACME Insurance company that deals with the full insurance cycle for policy submission, rating, underwriting, and servicing of insurance applications.

To estimate the annual medical expenditure for new customers using non-cost inputs, the dataset provided the following information:

- Age of beneficiary,
- Sex (male, female),
- Number of children covered by insurance,
- BMI,
- Smoking status,
- Region of residence (southwest, southeast, northwest, northeast),
- Insurance Charges.

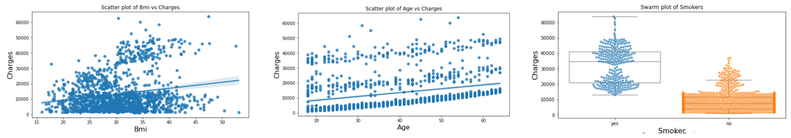
Link for the dataset: `https://www.kaggle.com/datasets/harshsingh2209/medical-insurance-payout`

Exploratory Data Analysis



(a) BMI (b) Age (c) Charges

Figure 1. Histogram of numeric variables



(a) BMI (b) Age (c) Smoking status

Figure 2. Relationship between independent variables and medical insurance payment

Key findings from statistical summary and data visualization is as follows:

1. Considering that the normal level of bmi range is 18 to 24, and the average bmi in this dataset exceeds 30, more than half of respondents are obese or morbidly obese.
2. People in their 20s are the most prevalent, yet they are relatively evenly distributed. Significant number of people spend less than $20,000 for medical insurance.
3. As age and bmi increases, the medical insurance charges also increase. This implies that as people get older and have higher degree of obesity, their medical insurance payment increases.
4. Smokers pay significantly higher insurance costs than non-smokers.

ANOVA Test:

Variable	F-value	p-value
Age group	26.02	< 0.001
Sex	4.4	0.036
BMI group	16.16	< 0.001
Children	3.3	0.006
Smoker	2177	< 0.001
Region	3	0.031

Table 1. Result of two-way ANOVA test

We used an ANOVA test to evaluate which factors had the greatest impact on medical insurance payment and to exclude variables with poor significance on the dependent variable when training the model. All factors were found to be significant on medical insurance payment since the p-value was less than 0.05. Therefore, all variables were applied in the model's training and testing. Furthermore, health-related factors such as age, bmi, and smoking status were the most important variables determining medical insurance payment. Through post-hoc analysis, persons who are older, fatter, or smoking will pay more for medical insurance.

Training and Testing Linear Regression

Before training the model, the dataset was divided into 75% train data and 25% test data. The pipeline was then built to normalize numeric variables (age, BMI, children) and generate dummy values for categorical variables (sex, smoker, region). Then, standard and penalized linear regression models were developed.
Linear Regression Equation: $Y = \beta_0 + \beta_1 \cdot x + ... + \beta_6 \cdot x_6 + \epsilon$

Hyperparameter Tuning

Penalized regression adds a regularization term $(1-\alpha)/2||\beta_2||^2) + \alpha||\beta_1||$ ($\beta$is a vector of coefficients) in a least squares loss function before optimizing it to estimate coefficients [2]. To identify the best regularization term, hyperparameter tuning is done by varying α from 0.1 to 1 with intervals of 0.1. The parameter λ, which affects the overall degree of regularization, was set at 0.5 for a fair balance between R-squared and RMSE. The best results were obtained when α was set as 0.1.

Cross Validation

When training the model, penalized linear regression was adjusted using 10-fold cross-validation.

Conclusion and Discussion

The performance of linear regressions are as follows:

Measurement	Value
R-squared	0.16
RMSE	11516
MAPE	115
PR	0.98

(a) Standard linear regression

Measurement	Value
R-squared	0.73
RMSE	6213
MAPE	47
PR	1.0

(b) Penalized linear regression (α = 0.1)

Table 2. Result of linear regression models

With an r-squared score of 0.72, penalized linear regression outperformed standard linear regression remarkably. It was possible to build linear regression models with significant performance using just non-cost inputs, limited to penalized linear regression. It is expected that insurance companies will also be able to use penalized linear regression in practice, which is straightforward to comprehend. However, machine learning or additional variables should be used to enhance prediction power. The most important factors for medical insurance payment were variables representing health status, such as age, BMI, and smoking status.

For more detailed process and codes see `https://github.com/JoyGeebeumPark/IMSE-586_Team-2`

References

[1] Ian Duncan, Michael Loginov, and Michael Ludkovski. Testing alternative regression frameworks for predictive modeling of health care costs. *North American Actuarial Journal*, 20(1):65–87, 2016.

[2] Hong J Kan, Hadi Kharrazi, Hsien-Yen Chang, Dave Bodycombe, Klaus Lemke, and Jonathan P Weiner. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PloS one*, 14(3):e0213258, 2019.

[3] Keshav Kaushik, Akashdeep Bhardwaj, Ashutosh Dhar Dwivedi, and Rajani Singh. Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, 19(13):7898, 2022.