**An analysis on linear regression modelling for predicting medical insurance costs using individuals’ characters**

Background

Medical insurance costs prevent over 37 million Americans from being covered (Grumbach, 1989).With the rise in medical insurance it becomes important to explore the attributes that contribute to insurance costs.

The Dataset

The dataset cost of healthcare is determined by characteristics such as sex, age, body mass index (BMI), number of children, smoking habits (whether an individual smokes or not), and the region that they are from. The sample size comprises people in the age range of 18 to 64, a demographic considered as adults (Han *et al*., 2015). Their BMI, smoking habits and number of children can offer insights on their health risk factors and potential impact on their medical charges (Viscusi, 1999; Thompson & Wolf, 2001; Trogdon *et al*., 2008; Izumi *et al.*, 2001). Also, their geographic location could impact health insurance costs since areas can be exposed to different risk factors such as potential disease outbreaks (Nerich *et al*., 2006), and a location can be indicative of one's economic class which could potentially affect medical insurance costs.

Introduction

Understanding the nuances of data exploration, model construction, and evaluation is crucial in the field of data science and predictive analytics. Using linear regression modelling, this research sets out on a quest to grasp the intricacies of predicting health care costs. We want to clarify the steps involved in data pretreatment, model creation, evaluation, and the consequences for practical application by dissecting the provided code.

Data Processing and Exploration

The code imports libraries, loads the dataset using pandas, and transforms category data into dummy variables for analysis. This is crucial for machine learning algorithms, as it requires numerical input (Brownlee, 2020). Categorical variables can be encoded into binary columns using pd.get\_dummies(), making them easier to include in regression models. Effective data exploration is crucial for understanding the dataset's properties and interactions. The code splits the data into train and test subsets and initializes the linear regression model (model\_0).

Building and Evaluating the Model

Useful evaluation measures are used to judge the effectiveness of the base model (model\_0). R-squared and root mean squared error (RMSE) metrics offer information about the model's capacity for explanation and precision of prediction. As opposed to RMSE, which measures the size of prediction errors (Chai & Draxler, 2014), R-squared measures the percentage of variance in the dependent variable that the model is able to explain (Cameron & Windmeijer, 1996). The 'sex\_male' variable is also left out when building a different model (model\_1). The effect of this particular feature is demonstrated by contrasting the performance metrics of the two models.

Application and Implications

The created model's usefulness for real-world circumstances is revealed in addition to its theoretical examination. We demonstrate the practical relevance of the concept by providing forecasts for three fictitious individuals, Bob, Lisa, and John. The model's ability to assist in decision-making in the setting of health care is demonstrated by the estimation of each person's health care costs depending on their attributes.

Stochastic gradient descent (SGD) regression and Optimised Hyperparameters for Improving Analysis and Model Training and Evaluation:

The code enhances linear regression by incorporating hyperparameter adjustment to improve model performance. GridSearchCV is used to identify the best set of hyperparameters, minimising negative mean squared error. A new SGDRegressor model is trained using these hyperparameters, and forecasts are created based on test data. Key metrics like R-squared and root mean square error (RMSE) are used to evaluate the model's efficacy.

Insights and Implications and Visualising Model Coefficients:

The upgraded model's metrics are compared to earlier linear regression models, revealing the impact of hyperparameter adjustment on prediction accuracy. Hyperparameter optimization improves model settings, resulting in better results. A bar chart illustrates the prominence of features in estimating medical expenses, highlighting their relative importance in the prediction process.

Conclusion

Data exploration and model evaluation are crucial in understanding the interaction between key data science ideas. This code can help create reliable predicting models for healthcare costs, aiding informed decisions in healthcare economics. Future work could explore stepwise regression, multicollinearity using Variance Inflation Factor (VIF), and regularised regression methods like Lasso, Ridge, or ElasticNet, especially when the dataset grows in terms of features. These techniques help close the theoretical and practical gap in healthcare economics.

References

Brownlee, J., 2020. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.

Cameron, A.C. and Windmeijer, F.A., 1996. R-squared measures for count data regression models with applications to health-care utilization. *Journal of Business & Economic Statistics*, *14*(2), pp.209-220.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions*, *7*(1), pp.1525-1534.

Grumbach, K., 1989. National health insurance in America--can we practice with it? Can we continue to practice without it?. *Western Journal of Medicine*, *151*(2), p.210.

Han, B., Compton, W.M., Jones, C.M. and Cai, R., 2015. Nonmedical prescription opioid use and use disorders among adults aged 18 through 64 years in the United States, 2003-2013. *Jama*, *314*(14), pp.1468-1478.

Izumi, Y., Tsuji, I., Ohkubo, T., Kuwahara, A., Nishino, Y. and Hisamichi, S., 2001. Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan. *International Journal of Epidemiology*, *30*(3), pp.616-621.

Nerich, V., Monnet, E., Etienne, A., Louafi, S., Ramée, C., Rican, S., Weill, A., Vallier, N., Vanbockstael, V., Auleley, G.R. and Allemand, H., 2006. Geographical variations of inflammatory bowel disease in France: a study based on national health insurance data. *Inflammatory bowel diseases*, *12*(3), pp.218-226.

Thompson, D. and Wolf, A.M., 2001. The medical‐care cost burden of obesity. *Obesity reviews*, *2*(3), pp.189-197.

Trogdon, J.G., Finkelstein, E.A., Hylands, T., Dellea, P.S. and Kamal‐Bahl, S.J., 2008. Indirect costs of obesity: a review of the current literature. *Obesity reviews*, *9*(5), pp.489-500.

Viscusi, W.K., 1999. The governmental composition of the insurance costs of smoking. *The Journal of Law and Economics*, *42*(2), pp.575-610.