Title

A Bayesian multiple linear regression approach on analyzing correlation between demographic characteristics and insurance charges

Introduction

Bayesian multiple linear regression is a Bayesian approach to analyze correlation between a vector of predicting variables and a response variable. The classical / frequentist multiple linear regression is performed extensively across industries for its strength of causal analysis and outcome prediction. Bayesian approach differs from the classical approach by consistently updating posterior distribution as more data is processed in the form of priors. Both approaches have their own strengths and weaknesses. In this project, I will analyze correlation between demographic characteristics and insurance charges using both classical and Bayesian approach.

Data Background

The data is obtained from Kaggle[1] which consists of insurance data with 1338 records and 7 columns. Dataset is preprocessed to drop unused column and transform binary responses into 0 and 1. Given the dimension of dataset is too large to perform 100,000 simulations in OpenBugs, so to ensure computing performance, the final dataset is downsized to 200 randomly selected records and 6 columns (5 predictors and 1 response). Following is a short description of the dataset:

```
Column 1: age
Column 2: sex -> male: 0 female: 1
Column 3: bmi
Column 4: number of children
Column 5: smoker -> no: 0 yes: 1
Column 6: insurance charges (in ten thousands)
```

Proposed Method

For both classical and Bayesian multiple linear regression, I'll give an overview of the goodness of fit and the explanatory power of predictors to the dataset. Then, I'll compare the coefficients of betas between classical and Bayesian inferences. Using the methods taught in the course, I'll diagnose the distribution of data points by analyzing potential presence of outliers or influential points using various methods, and finally I'll use the computed models to predict the insurance charge given certain demographic characteristics of a potential patient.

---

[1] Choi, Miri. (2017). Medical Cost Personal Datasets. https://www.kaggle.com/mirichoi0218/insurance

Experiment & Evaluations

	Classical MLR	Bayesian MLR
intercept	-1.2041	-1.202
β1	0.0282	0.02817
β2	0.1356	-0.135
β3	0.0305	0.03052
β4	0.0442	0.04392
β5	2.3387	2.338
R^2	0.7791	0.7767

The table above is a summary of the model coefficients and R^2 value. Notice that the coefficient values and R^2 are very similar between classical and Bayesian multiple linear regression except for β2 which is a weight estimate for predictor 'sex'. Classical model suggests that female receives higher insurance charges whereas Bayesian model suggests otherwise with all other predictors being equal. R^2 shows that the predictors from both models have relatively high explanatory power. Both model suggests that patient with greater age, BMI, number of children and being a smoker has greater chance of having higher insurance charges.

	Classical MLR	Bayesian MLR
Outliers	None	4, 10, 35, 53, 63, 70, 86, 99, 100, 103, 116, 139, 141, 143, 144, 157
Influential points	4, 10, 35, 59, 63, 70, 86, 99, 100, 103, 116, 117, 139, 141, 157, 162, 176, 189	4, 10, 35, 63, 103, 116, 141

In terms of model diagnoses, I performed residuals test and cook's distance on the classical model, and cumulative and CPO tests to detect potential outliers and influential points on the Bayesian model. According to the results from the table above, the classical model did not detect potential outliers from the data points, whereas cumulative test on Bayesian model suggests 16 potential observations are outliers since their coefficients are close to 0 or 1. For classical model, a total of 18 observations are greater than 4/n threshold, hence they are potential influential points. CPO test on Bayesian model suggests a total of 7 potential influential points with large coefficients. While the two models yield different results, we can notice that some observations are classified as outlier/influential for both classical and Bayesian model.

To estimate and compare the prediction results between classical and Bayesian model, I assumed a potential patient with characteristics of 30 years old male smoker with 30 BMI and 1 child. This assumed data is used to estimate the mean and predictive insurance charges for both classical and Bayesian models and their respective credible sets.

	Classical MLR	Bayesian MLR
Mean response	2.9410	2.941
Mean 95% credible sets	(2.7522, 3.1298)	(2.751, 3.13)
Predictive response	2.9410	2.943
Predictive 95% credible sets	(1.7657, 4.1162)	(1.773, 4.114)

From the table above, the mean and predictive response and their respective credible sets for both classical and Bayesian model are very similar. The insurance charge for the given patient is predicted to be $29,410 for the classical model and $29,430 for the Bayesian model.

Conclusion & Discussion

After testing the dataset using both classical and Bayesian approach, the test results are very similar between each other. While classical / frequentist approach and Bayesian inference are based on different concepts (parameters are assumed constant vs. parameters are assumed to be random variables), two models yield similar results for this particular dataset.

Classical model is computed using Matlab and Bayesian model is computed using OpenBugs, and both codes are attached for studies. Given the two models are based on different concepts, more tests can be done using larger dimension dataset with greater variation in the future to observe the statistical differences between classical and Bayesian approach.