How Medical Cost is affected

10/05/2018

Yiyi Xu & Yuance He

Overview

https://www.kaggle.com/mirichoi0218/insurance

We are interested in the relationship between multiple factors and the medical cost. After estimating the coefficients of each factor, it would be a effective tool for some insurance company to design more appropriate policy for different clients. The dataset we obtained is the medical cost personal dataset from the above link.

Goals

Create a linear regression line of the medical cost. Consider how age, bmi, number of children, whether to smoke effect medical cost.

Using multivariate coefficient optimization methods, we will be able to compare the estimators with linear regression.

Specifications

1. Filtrate the factors which have influence on the medical cost.

   Using the correlation procedure, we could get Pearson Correlation Coefficients and p values of each variables. A correlation coefficient means relationships between two random variables. The closer the correlation coefficient is to +-1, the better the relationship between two random variables can be described by monotonic function.

2. Find outlier and delete the outlier.

   Using box plot to determine and find out outliers, we are going to use the dataset without outliers, which will be more accurate for estimation.

3. Estimate how each vector affect the insurance cost.

   Use dot plot to see what's the relationship between each factor and the insurance cost. Whether they are linear relationship or quadratic relationship or so on. And use the polynomial regression model find a regression polynomial.

4. Use different method to estimate the mle.

   Use Multivariate Newton's method to get the MLE of the coefficient of each factor to get the polynomial function.

   Use EM optimization to get the MLE of the coefficient of each factor to get the polynomial function

   Compare estimators we have get and take the better one.