Medical Charges Prediction Using Linear Regression Models

Emmanuel Victor Barbosa Sampaio
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
emmanuelsampaio@alu.ufc.br

Carlos Alfredo Cordeiro de Vasconcelos Filho
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
carlosalfredo@alu.ufc.br

***Abstract*—The prediction of medical charges of a patient can help medical insurance companies to offer plans that overcome clients costs and increases the profit margins. To perform this predictions a simple approach is the use of linear regression model, that can be define via different approaches. On this work we going to discus the application of different approaches to define a linear model to predict the medical charges of a patient.**

***Index Terms*—Medical Charges, Linear Regression Methods.**

I. Introduction

The profit of a medical insurance company is related to offering plans whose values are higher than the medical expenses charged to the company. In this domain the medical expenses of a client are associated with the client's health condition. Healthy clients normally need less medical treatment than unhealthy. Given that scenario, a company may use medical costs estimates to figured out how low they can price their insurance, witch makes them more competitive, while keeping a good profit margin.

Motivated by the problem above, a insurance company may be interest in developing models that can help them predict the costs of their patients based on information available about them. A way to solve this problem is through the use of regression models. This kind of model relates a variable that we aim to predict to a set of variables by designing to them a mathematical relation that can be understood as a function that has as input a set of variables, called features, and has as outcome the variable that we want to predict, called target variable. In the domain of the patient costs prediction, the model is going to have as features the patient information and as target the medical costs.

In this context, this work going to discuss the use of linear regression models in order to predict the patient medical charges. Base on [2] linear regression is a simple regression technique that is basically a linear combination of the features, this model can return a useful interpretation about the relation of the features and the variable that we aim to predict. Besides the interpretability of the model, we need to discuss the model accuracy, that is related to the capacity of predict values near to the original ones, in order to improve the accuracy of the model we going to discuss different approaches used to calculate the coefficients of model as well as discuss the model evaluation based on accuracy metrics.

Given the presented background, we going to use in the analysis and regression a dataset of the book Machine Learning with R, From Brett Lantz. The dataset contains annual total charges for different clients, and there is also information about client's age, the US region that the it lives based on four geographical regions, the number of client's children or dependents, its smoking habit, and Body Mass Index (BMI), which according to the book brings information if the patient is over or under-weight relative to their height.

The outline of this paper is organized as follows: Section II presents the theoretical background behind the regression models that were used, Section III presents a the variables in the dataset and their relation with the annual charges, this process going to be useful to determine the variables that going to be used as features of the model. Section IV presents a discussion about the results of the linear models. We conclude this work in Section V with a final discussion about the results.

II. Analysis Background

The regression models are based on a definition of a function applied to a set of variables, also called predictors, that results in a approximation of the values that we want to predict, also called target variables. In a mathematical point of view we can define $X \in \mathbb{R}^{n \times p}$ as a matrix that contains the predictors data, $Y \in \mathbb{R}^{n \times 1}$ as a matrix that contains the target variable, where n is the number of observations in the dataset and p is the number of predictors. So the regression tries to find a function $\hat{f}(X)$ that best approximate the following function:

$$Y = f(X) + \epsilon \tag{1}$$

In Equation 1 $f(X)$ represents a function that approximate the values of Y using the values of X. The difference between the function results is represented by ϵ, from [1], ϵ can be consider a error that is independent and has mean equal zero.

In the context of this work we going to look for a model that best approximate the values of Y, with a limitation that we going to use only linear regression models. This mentioned models are functions that are linear in the values of X. The following sections going to describe the models and the process used to define them.

A. Linear Regression: base line

We can represent a linear regression model for a set of p predictors X_j as the following equation, where β_j is a coefficient of the function related to the predictor X_j.

$$\hat{f}(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{2}$$

To find a model as 2 we need to define a set of parameters β that can be use as coefficients. In this circumstances, a possible approach is called Ordinary Least Squares (OLS), that consist of the calculus of the beta values that minimize the cost function $J(\beta)$ presented in Equation (3), where y_i is the real value of the target variable associated with the values of $x_{ij} \forall\ 1 \leq j \leq n$, where x_{ij} represents the value of the predictor in the observation i.

$$J(\beta) = \sum_{i=0}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} x_{ij} \beta_j \right)^2 \tag{3}$$

To perform the calculus of the beta values, we can use a linear algebra approach presented by [1]. First we redefine the matrix X as $X \in \mathbb{R}^{n \times p+1}$, this increment in the p value is related with the β_0 value presented in (3), because now we add a column of 1 in X, considering all predictor values independent, so the matrix X is invertible matrix, it is possible to show that when we setting the first derivative of J to zero, the unique solution is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{4}$$

The process of calculate the values of $\hat{\beta}$ is called training. In this context, we use a set of observations called training set to perform the calculus of beta values using Equation (4). After training the model we need to evaluate if the model is useful to perform the predictions, to do it we going to perform the model evaluation using another set of observations, called test set that contains data that were not used in the training.

B. Evaluating the model

The evaluation of the model pass by the analysis of how well the model can predict the data that was not used to perform the training process. To perform this analysis we can use a metric called mean squared error (MSE) that calculate the sum of the squares of the residuals, which is the difference between the real values y and the predicted $\hat{y}$, divided by the number of observations. The following equation shows the MSE formula

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5}$$

The book [1] shows that we can break the expected MSE for the training set into a sum of three values: the irreducible error, squared bias and variance. Irreducible error is the variance of the target variable around its true mean, it is like noise in the prediction. The squared bias is associated with the difference between the prediction and real values, from [1] the squared bias can be understood as the amount by which the average of our estimate differs from the mean of real values. The variance can be seen as the variation of the values of the prediction as we change the data used to performing the training.

Based on the mentioned values, the model needs to have low bias and low variance. When the model has a high bias it can be consider a poor model in order to solve the problem, since it can not perform well on the training set, which is called underfiting. But, it is possible to show that as we reduce bias we can find a point that the variance starts to increase, in this point the model can perform well on the training set, but has low accuracy in the test set, this is called overfitting. So the model creation need to pass by a trade-off between the bias and variance.

Given the previous discussion in order to evaluate the model performance in the test set we going to use two metrics: Root Mean Square Error (RMSE) and the coefficient of determination also called R^2. RMSE is the root of the mean square error, using the root of the MSE we have a value in the same metric unit as the prediction value. R^2 is a metric of correlation between the predict value and the real value.

$$R^2 = 1 - \frac{\sum_{j=0}^{N} (y_j - \hat{y}_j)^2}{\sum_{j=0}^{N} (y - \bar{y}_j)^2} \tag{6}$$

From [2] the R^2 measures the proportion of variability in the real data that can be expressed by the model. When the value o R^2 is near to 1, it identify that the model response has a variance near to the variance of the the data, when to small compare to 1 the model poorly describe the data.

C. Resampling Methods

Generally, resampling methods consist of creating two subsets of samples: the training subset and the test subset. The training subset is used to fit a model, witch in the case of the linear regression means that they are used to determine the linear coefficients. The test subset is used to estimate the efficacy of the model by comparing the results of the prediction made by applying the predictors in the model with the experimental result. This process is repeated multiple times to avoid bias and the results are aggregated.

The resampling methods differentiate between each other by the process used to create the subsets. In this paper we will use the k-fold Cross-Validation. This resampling method consists of dividing the sample in k random sets of approximately the same size. The first subset is put in the test subset and the rest go into the training subset. The coefficients are estimated using the training subset and the error is calculated using the test subset. This process is repeated with all the remaining $k-1$ sets and the k estimatives resulting of this method are summarized, usually with the standard error and mean.

D. Improving Linear Regression

Based on [2] a linear model can be improved on 2 ways: improving linear model interpretability and increases the accuracy of the model. In this domain, interpretability means makes the linear model more simple to understand in the sense that we can verify the importance of the variables in the target

prediction. By accuracy we mean adjust the model in order to perform better prediction, reducing the variance associated with the test. In this subsection we going to discuss approaches that are used in order to adjust the model to improve accuracy, reducing variance with a small increasing in bias.

The first method presented is called Ridge Regression, on this approach we add a term to the cost function (3), this terms objective is penalize the coefficients, reducing its values.

$$J(\beta) = \sum_{i=0}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\beta_j \right)^2 + \lambda \sum_{i=1}^{j} \beta_j^2 \qquad (7)$$

The ridge regression penalization is controlled by the value of $\lambda \geq 0$, when lambda is too small the regularization effect is attenuated, given a result near to obtain using the least squares presented in Subsection II-A. When λ is too high, the effect of the regularization going to be accentuated, reducing β values towards zero, which can cause a increasing in the bias.

Ridge regression represents a trade-off between variance and bias. Because we are not minimizing just the squared bias anymore the bias of the model will naturally get higher. The penalty factor makes so that the resulting model has less variance because the coefficients can not grow much without having a proportional reduction in the squared bias. Smaller coefficients imply that the differences between the coefficients of distinct training sets is smaller, therefore the final model has a lower variance. Because MSE is a combination of bias and variance we can tune the λ to find the optimal trade-off between both so that we can minimise the MSE.

As present [2] the ridge regression has a disadvantage of not set to zero the coefficients when λ are not too large, which causes the existence of many features in the prediction model which affects the model interpretability. To overcome this problem we can use the Lasso, here the regularization term is related to the sum of the coefficient values.

$$J(\beta) = \sum_{i=0}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{n} x_{ij}\beta_j \right)^2 + \lambda \sum_{i=1}^{j} |\beta_j| \qquad (8)$$

E. Principal Components and Linear Regression

The correlation between predictors can worsen the variance of the linear models [3]. To illustrate that, imagine that the linear regression of a training set results in the equation (9). If x_1 and x_2 have a correlation as described by the equation (10), then all the equations described by (11) are equivalent to the equation (9). Therefore, exists a infinite number of equations that describe the same regression and, as a consequence of that, each training set can result in extremely different coefficients witch increases the instability of the final model.

$$Y(x) = \beta_1 x_1 + \beta_2 x_2 \qquad (9)$$

$$x_1 = \beta_3 x_2 \qquad (10)$$

$$Y(x) = (\beta_1 + k)x_1 + (\beta_2 - k\beta_3)x_2 \forall k \qquad (11)$$

Other possibility is that the number of predictors is bigger than the number of observations. In that case, the linear regression also can't determine a unique equation that minimises the squared bias.

To solve both of this issues, we use Principal Component Analysis (PCA), on the predictors to generate a new set of predictors that are both uncorrelated and small enough that it has a unique solution. Pre-processing predictors via PCA before performing linear regression it is called principal component regression (PCR).

A problem with PCR is that it does not factor in the response in the creation of the new predictors. Because of that, the new predictors may be uncorrelated to the response, making them essentially useless for the predictive model.

To circumvent this problem we can use another approach, called Principal Least Squares (PLS). In this approach, in addition to feature data we use for the calculus the target variables data. From [2], PLS defines the components based on the variance of feature and target variables. To do it, it is performed the calculus of the correlation between the target variable and a feature, the correlation is placed as coefficient of linear combination of feature values, generating the first principal components, them the values of X are othogonalized related to the first principal component and the calculus is performed again. The components found via this process are used as features in the regression model.

III. Exploratory Analysis

The dataset provided by the book contains 1338 observations of different patients. Each patient has its total annual cost for the insurance company and personal data about sex, country region, smoking habit, number of dependents, Body Mass Index (BMI) and age.

We started by look into the medical charges distribution. This variable has a positive skewness and a mean value 13270.42 which indicates that the majority of patients are associated with charges lower than this value. To visualize the distribution we create a histogram of the variable presented in Fig.1. In this figure the left side is the original data distribution and in the right side the logarithm in base 10 distribution, in this transformed we can visualize that the data is concentrate in values near to 10^4, few observations assume values greater than $10^{4.5}$ and lower than $10^{3.5}$.

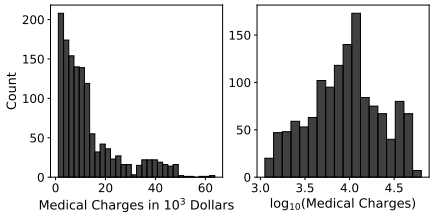

Fig. 1. The distribution of the patient total charges. It is possible to visualize that the variable has positive skewness and the data is concentrate around 10^4.

We identified that the distribution of the charges for male and female as weel as the different regions does not change considerably, when we compare the mean value per region or per sex as well as the standard deviation or skewness the values are close to each other. The following Fig.2 shows a box plot of the logarithm of medical charges for each sex and

region, we can identify that the median as well the distribution of values are around the same values.

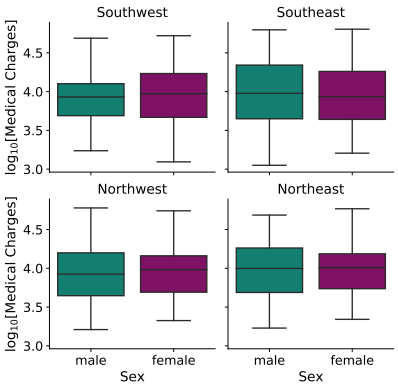

Fig. 2. The different charges value paid per the patient in different country regions. It is possible to verify that the patient

From all categorical variables in the patient data, the variable that has the greatest difference in the distribution of the medical charges was the smoking habit information, in the dataset 1064 clients are non smokers and 274 which are associated with the highest values of the medical charges. Considering the charges related to the people that smokes, 75% of this group pays more than 20.826 dollars. In opposite only 25% of the people that not smoke pay more than 11.362.88. To illustrate the difference in the distribution of medical charges of smokers and not smokers, we used Fig.3 to shows the distribution of the logarithm transformation of the medical charges values. From the box plots it is possible to determine that the smoking habit is associated with high values of charges.

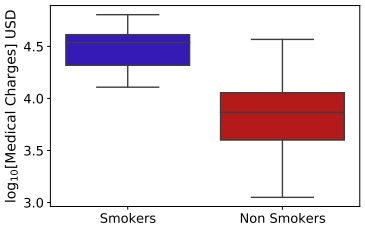

Fig. 3. Bar plot of the smokers and non smokers medical charges.

We also observed how the client costs changes as the number of children or dependents of the clients changes. In this context, we plotted a box plot for the distribution of medical charges for each number of dependents, Fig.4. The distribution of the medical charges changes per number of dependent, when we considering the clients that have no dependent we have the greatest group of person in this categorical variable, they are 574.0, this group contains 459.0 non smokers and 115, the group that contains more smokers than other groups, this explains the high variability in the distribution of medical charges, since from Fig.3 we have that the smokers are associated with the high values at the same time the non smokers are associated with small values. When we consider the group with five dependents, they have the lowest variability, which can be a reflex of the smokers in this group, that is only one.

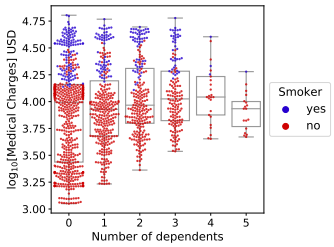

Fig. 4. The distribution the logarithm transformation of the medical charges per number of dependents of the person.

We them verified how BMI and age are related with the charges. Using a scatter plot between age and the logarithm of medical charges we verified that old patients pays more than young patient when we considering people with same smoking habit. We also verified that the patients with highest BMI and that are smokers are associated with the greatest values of medical charges. The following figure 5 illustrate the discussion, in the right side the medical charges associated with non smokers and in the values associated with the smokers, the color of the points is based on the BMI values of the observations.

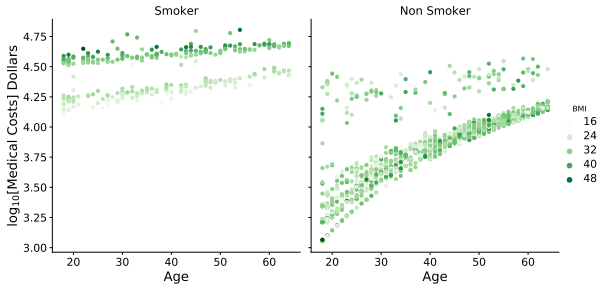

Fig. 5. The distribution of insurance fee related to the Patient Age, smoking habit and BMI. Here the charges are higher for people that smokes and has a higher BMI. Also the people with high age pays more than the young people that have the same smoking habit.

To finish the analysis we need to verify the correlation between the variables. We used the correlation information to identify the most correlated variables in order to identify potential features for the linear model. In order to perform the correlation calculus considering the categorical data such as smoking habit, region an sex we used dummy variable substitution, where each class was substituted by a value in a ordinary order.

Using Pearson Correlation coefficient we calculate the correlation between the variables, the result is presented in Fig.6, from the figure it is possible to visualize that the highest correlation is between smoker and charges, which is a reflex of the observed difference between the prices of the smoker and non smokers presented in the Figures 5 and 3. We also verify that age and BMI have a strong correlation with the charge values, which was also showed in Fig.5.

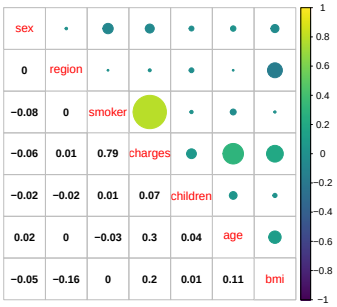

Fig. 6. The correlation between the variables. We can identify the high correlated variables are smoker and charges.

IV. Predicting Client Medical Charges

In this section we going to apply the different linear regression models approach mentioned in the Section II. We going to compare the model performance as well the features used to perform the regression.

To train the model we have a training and a test set. The training set contains 1003 observations and the test 335. In the training set we going to perform the k-fold cross validation procedure using k=5 and k=10 and with the test set we going to evaluate the model using R^2 and $RMSE$.

A. Linear Regression: Using Ordinary Least Squares

To perform the simple linear regression we used a function that implements we use the OLS to calculate the coefficients. As features we choose to use the variables with high correlation with the medical charges, that are smoking habit, age and BMI, as you can see in Fig.6. We them used the test set data to test the model, we found a $R^2 = 0.76$ and RMSE $= 5989.088$.

We used cross validation to verify if adding other features to the model can help the model accuracy. From 7 we found that using five variables, the already used plus regions and number of dependents we have a model with high R^2 and low $RMSE$. Using the test set in the new model the result as $R^2 = 0.77$ and RMSE $= 5929.00$.

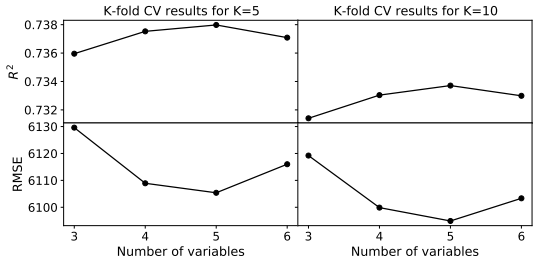

Fig. 7. The cross validation result considering both k=5 and k=10.

Now look to the residuals, we verified in Fig.8 that the variance of the error terms are not constant considering the original and the logarithm of the medical charge values, which indicates that the skewness presented in the target variables does not need to be remove to improve the prediction.

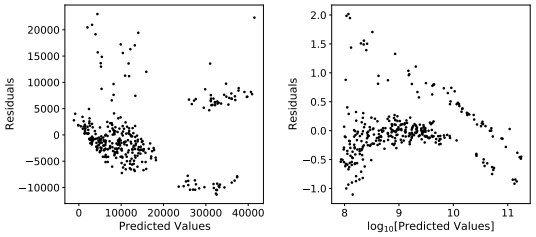

Fig. 8. The residuals plots vs the predicted values. We can identify that the variance of the error is not constant for both transformed and original data.

B. Using Regularization

To improve the model accuracy in the test set, we decided to use the presented regularization techniques Lasso and Ridge regression. On this subsection we going to investigate the tuning parameter λ presented in both cost function of Ridge (7) and Lasso (8). The investigation has as goal understand how the accuracy of the model changes as we modify λ. As features in this analysis we going to use the same features five that we used to find the best result of OLS.

Using cross validation we verified that as the values of λ increase the accuracy of the ridge regression model decreases as well as the value of R^2, the value of lambda that is associated with the lowest mean of RMSE found via the cross validation was 0, as the figure 9 shows. This result indicates that ridge method turn into a OLS method. In the test set the result was: RMSE $= 5924.55$ and $R^2 = 0.77$.

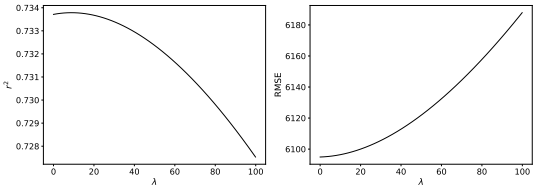

Fig. 9. The R^2 and RMSE found using different values of λ for the ridge regression.

Using cross validation we verified that as the values of λ increase the accuracy of the lasso regression model decreases as well as the value of R^2, the value of lambda that is associated with the lowest mean of RMSE found via the cross validation was 0, as Fig.10 shows. This result indicates that lasso cost function turn into a OLS method. In the test set the result was: RMSE $= 5924.49$ and $R^2 = 0.77$

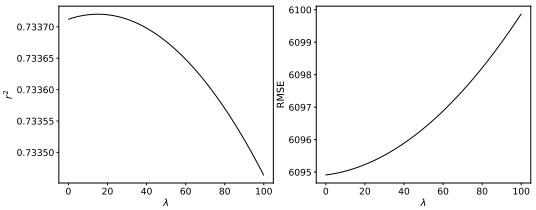

Fig. 10. The R^2 and RMSE found using different values of λ for the lasso.

C. Using Dimensionality Reduction

The last technique that we going to discuss is the use of principal components calculus to define the model. The discus-

sion of this subsection going to be based on the investigation of which technique leads as a model in low dimensional number of features as well in this lower dimension the accuracy as good as the regression models presented in the last subsections.

We started by performing the calculus of the principal components using only the features domain. In our context, we used all 6 features available. The weights associated with each components are presented in 11 as well as the inertia associated to each components, that from [2] can be seen as the percentage of information of the features that can be explained by the component.

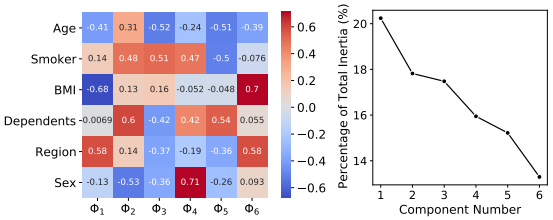

Fig. 11. In the left side a plot showing the coefficients associated to each principal component. In the right side, the inertia associated to each component.

We them verify how the accuracy of the model change the number of components that are used as feature in the regression model. To perform this verification we used k-fold cross validation where k=5 and k=10, and we verify how the mean of R^2 and mean of $RMSE$ change as we change the number of components in the model. We started by the component that has the highest inertia associated and add components in decreasing order of associated inertia. The results is present in Fig.12, it is possible to verify that for lower number of components the model has a accuracy lower than the already discussed models, but as we increase the number of components the accuracy also increase, reaching high values when the dimension goes near to the original.

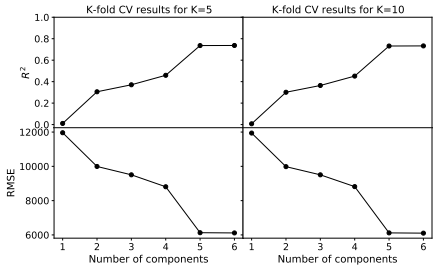

Fig. 12. The RMSE and R^2 value for different number of components.

As we discussed in the theoretical section, the PCR is based on a unsupervised learning technique that not uses the information of the output in order to calculate the components. To improve the analysis we performed the dimensionality reduction linear regression using the PLS, which calculate the weights based on the correlation of the target variable and the features. The weights associated to the components of the PLS are presented in Fig.13. We can observe that the weights associated with the smoking habit and age represent the highest values in the components with high inertia associated.

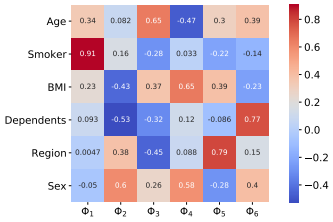

Fig. 13. The coefficient associated with each variable used in the component.

Using the cross validation approach we verified that using one or two components values from the PLS we had a better result compare to the same number in the PCR approach as we can see comparing the result of PCR cross validation Fig.12 and PLS cross validation Fig.14.

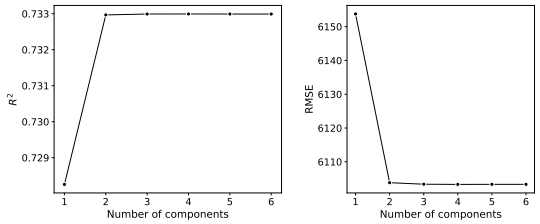

Fig. 14. Cross validation results using k-fold cross validation when k=10. .

In the test set, the values found using only one component found via PLS was RMSE = 6037.31 dollars, and $R^2 = 0.74$. Using two components, the RMSE = 5927.59 dollars and $R^2 = 0.76$. In this context, the use of this approach leads us a model in a lower dimension and with a accuracy near to the simple ones, that is better than the PCR approach.

V. Conclusion

From all models develop here, when we consider the accuracy analysis, the simple linear model found using OLS that has as variables the age, BMI, smoking habit, number of dependents and region deliver a good model, with a $RMSE = 5989.09$ and $R^2 = 0.77$. We found that via PLS we can find a model with only two features, and a accuracy as good as found via OLS. As showed by the figure 13 this two features are combination of the variables that have the highest correlation in the dataset. It is important to mentioned that the application of Ridge and Lasso regression did not imply in a improvement in the model accuracy, since the value of λ found via cross validation for both models was 0.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[2] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

[3] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. New York, NY: Springer, 2013.