Statistical Learning Final Project Report

Bakir Hajdarevic, Dennis Murphy
The University of Iowa

Abstract

The Medical Costs Personal Datasets (MCPD) on Kaggle is comprised of medical insurance costs for 1338 patients. MCPD also describes several features of the patients including: age, body mass index (BMI), sex of the patient, number of children, region within the United States, and smoking status. In this study, we sought to derive the relationship between smoking and non-smoking individuals and their medical expenses. We also examine the effects of other features in regards to the medical expenses.

I. Introduction

The proposed study will examine the relationship between smoking and non-smoking individuals with regards to their total medical expenses. We will use the Medical Costs Personal Datasets (MCPD) provided by Kaggle. However, the dataset is originally from the textbook Machine Learning with R by Brett Lantz. Based on their smoking habits, we aim to derive inferences and models between several characteristics of individuals and their medical expenses.

Our predictors consist of 6 features provided by the MCPD. These features provide valuable insight into the possible factors contributing to an individual's medical expenditures. The specific predictors are as follows: age, body mass index (BMI), number of children (children), cardinal direction within the United States (region), and whether the individual is a smoker (smoker). Our response variable, total medical costs, is described by the variable charges.

The main focus will be finding relationships between the predictors and the medical charges of each patient. This will use the regression methods discussed in class along with splitting the continuous response into binary values of high and low charges. The regression methods that will be applied are multiple linear regression, polynomial regression, ridge regression, and lasso regression. Subset selection will be applied to linear regression to select the best model along with applying 10-fold cross-validation to all methods. We will use the following metrics for evaluating the goodness of fit of the models: mean-squared error (MSE), coefficient of determination (R^2), Akaike information criterion (AIC), and Bayesian information criterion (BIC). Lastly, we will examine the correlation between predictors as well as between predictors and the response.

Our dataset has many more samples than predictors thus the least squares estimates given by linear regression should have low variance. However, it will be interesting to see if ridge and lasso regression can reduce the variance enough to justify the increase in bias introduced. It is unknown whether the predictor and response relationship is linear thus both linear and nonlinear methods will be used. Using the aforementioned statistical learning methods, we will apply primarily regression approaches to measure the coefficients/parameters between the response variable and individual predictors.

In this study, we found that there is a positive correlation between an individual's smoking status and their medical insurance costs. We have examined how other features, such as age, BMI, geographical region, etc, effect the medical insurance costs. Lastly, we derived regression models for our data set using a variety of regression models then assessed their goodness of fit using several model fit criterions.

II. Methods

To model the relationship between the predictors, *x*, and the response, *y*, as an *n*th degree polynomial, we use the following equation:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon. \quad \textbf{(Equation 1)}$$

where describes the strength of the relationship between the predictors and the response while ε is an unobserved random error. By modeling the data set over varying degrees of polynomial we may quickly deduce if the data is linear, quadratic, cubic, etc. Furthermore, we may realize that we need to rely on more advanced algorithms in order to properly model our data set. An alternative method we may use is ridge regression which is defined as linear regression with a penalization on the square magnitude of the coefficients. This penalization has a parameter, λ, that is set usually set by cross validation.

$$\underset{\beta_0 \beta}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \left\{ \frac{1}{2} \|\beta\|_2^2 \right\}$$

(Equation 2)

Lasso regression is similar in that it has a penalty added to the optimization equation but now it is the L1 distance.

$$\underset{\beta_0 \beta}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2 + \lambda \{ \|\beta\|_1 \}$$

(Equation 3)

In order to assess the quality of fit of the Equations 1-3 in relation to other models, we turn to using model selection criterions.The simplest model criterion is mean squared error which is given by:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

(Equation 4)

This model is simply the mean cumulative squared difference between the true value and the predicted value. A more common model is the coefficient of determination (R^2). This model describes the degree of the variance in the dependent variable that is predictable from the independent variable(s).

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

(Equation 5)

Due to Equations 4 and 5 bias to favor more complex models, often they select models that have overfit the dataset. To reduce the bias for more complex models, we turn to using the Akaike information criterion (AIC):

$$AIC = 2K \ln(\text{MSE}) + 2(n)$$

(**Equation 6**)

AIC is essentially a tradeoff between the goodness of fit of the model and the simplicity of the model thus providing a less bias selection criterion. A similar criterion is the Bayesian information criterion (BIC) which is described in equation 7. Unlike AIC, the BIC tends to favor simpler models as it penalizes increasing model complexity.

$$BIC = \log(\text{MSE}) + \frac{\log(K)n}{n}$$

(**Equation 7**)

To potentially reduce the redundancy of certain models, we will assess the collinearity between predictors. This can be achieved by visually assessing a plot of the pairs of predictors and observing any relationship between predictors. In addition, a correlation matrix may be used to assess the correlation between predictors. Strongly correlated variables will be assessed to potentially exclude them from further analysis using VIF. Data processing will need to be applied to remove any outlier points that heavily influence model by calculating their leverage.

After assessing the collinearity between predictors, the strength of the relationship between each predictor and the response will be determined. This will be done by obtaining the statistical significance (using the appropriate test statistic) and the respective magnitudes as well as the correlation coefficients for each predictor. Once coefficients have been deemed useable the regression methods will be evaluated based on their test MSE and classification methods on their test error rate. We will also apply model selection criterions such as BIC and AIC.

To see which predictors have the most importance subset selection will be applied. The result will show which, if any, predictors can be removed from our predictive model. Forward, backward and best subset selection will all be looked at and selection method with the smallest Cp will be used. Cross validation will then be used on the selection method to confirm that the appropriate number of variables is used. Shrinkage methods will also be applied to the linear model to see which predictors are important. The methods Ridge regularization and Lasso regularization will be tried out and their optimal lambda value will be obtained using cross validation. While the regression coefficients will be limited, introducing a small degree of bias, the variability of their estimates should decrease enough to justify the limitation.

In order to implement the methods discussed above the programing language R will be used. For more complex methods libraries of functions will be used, for example ISLR contains functions to perform the different subset selections.

III. Results and Discussion

Simple analysis

We computed the correlation between the quantitative predictors (Age, BMI, Children) and the response, Medical Costs. According to Figure 1, there appears to be a positive correlation between the response and the predictors. Moreover, the predictor Age has the highest correlation at 0.3. When examining the correlation amongst the predictors, there is a

slight positive correlation between Age and BMI. Despite these correlations, none of the values recorded were statistically significant (i.e. p-value < 5%).

According to Figure 2, for the entire sample population, the mean medical insurance costs is higher for smokers than for non-smokers. Furthermore, in Figure 3, we have derived a linear regression of the medical insurance costs for the sample population with respect to their age and smoking status. Both regression lines have a positive slope. We examined the relationship between gender, smoking status, and medical insurance costs. Figure 4 shows that there is little difference between genders and their smoking status. However, when either gender is identified as a smoker, the mean medical insurance costs substantially increase. Figure 5 shows the boxplots for the sample population's BMI binned into BMI increments of 5. According to Figure 5, the mean medical insurance costs increase with increasing BMI. In the appendix, we have provided comparisons of the following: smoking status and age, number of children and medical insurance costs, as well as geographical region and medical insurance costs. There did not appear to be any relationship within these plots.

Figure 1. Correlation matrix of the quantitative predictors and the response, medical costs.

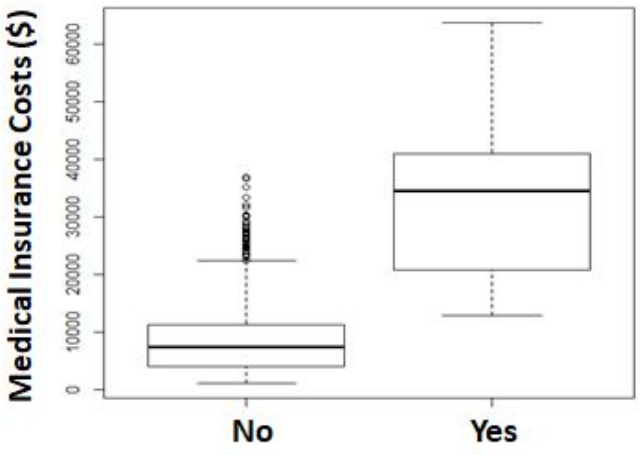

Figure 2. Boxplot of all patients and their medical insurance costs. The patients are categorized based on their smoking status.

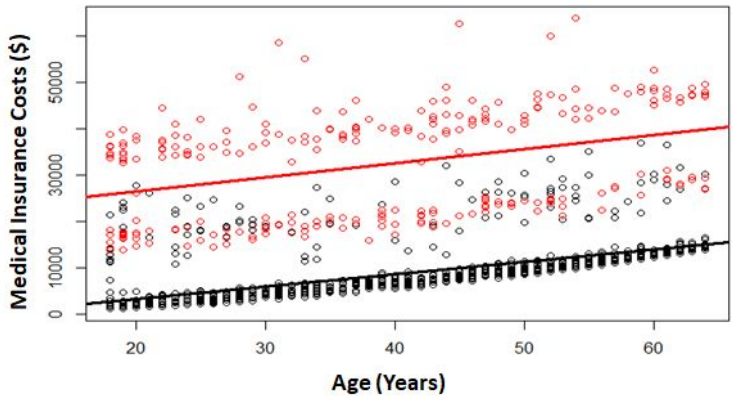

Figure 3. Individual's age against their medical insurance costs for smoking (red circles) and non-smoking (black circles). The linear regressions are also given for smoking (red) and non-smoking (black).

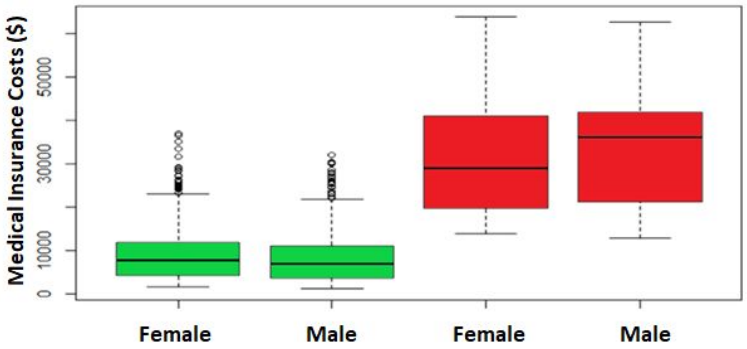

Figure 4. Boxplots of individual's gender and their medical insurance costs as well smoking (red) and non-smoking (green) status.

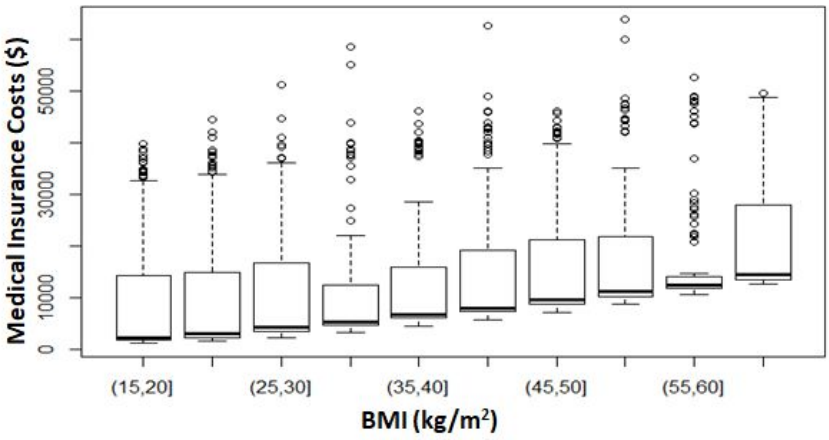

Figure 5. Boxplots of the BMI of patient's in bins and their medical insurance costs.

Polynomial Regression

To see how well a simple polynomial regression model fits the whole data set, we obtained regression models of varying polynomial degree from 1 to 10. We then obtained model selection criterion scores for MSE, R^2, AIC, and BIC. Table 1 provides a summary of these results. According to Table 1, a polynomial of 10 degrees best describes the whole data set. However, when we split the data according to individual's smoking status, we obtained different results. A linear model may best describe both the data sets. However, we assumed that the AIC provides a better selection criterion than the other criterions given the latter's bias in model complexity.

Table 1. Summary of the polynomial regression model results. Shown are the scores of the best model fit according to the model criterion (MSE, R^2, AIC, BIC) as well as the degree of the polynomial for models describing the full data set, non-smokers, and smokers.

	Full Data Set		**Non-Smokers**		**Smokers**	
	Score	**Deg. Poly.**	**Score**	**Deg. Poly.**	**Score**	**Deg. Poly.**
MSE	1.481×10^{6}	10	1.886×10^{6}	1	8.640×10^{6}	5
R^2	9.899×10^{-1}	10	9.474×10^{-1}	1	9.347×10^{-1}	5
AIC	2.520×10^{4}	10	2.025×10^{4}	1	5.511×10^{4}	1
BIC	2.662×10^{4}	2	2.486×10^{4}	1	5.537×10^{4}	1

Shrinkage and Selection methods

As shown in figure 6 the subset selections all returned the same values for their Cp fit criterion so best subset selection was then arbitrarily picked to undergo cross validation. The cross validation agreed with the prior methods in that five out of the six predictors should be used. Table 2 tells us that all predictors except sex have been selected by the cross validation approach. Comparing the MSE of subset selection, ridge, and lasso regression to the full linear model in Table 3 shows that no shrinkage or selection methods are needed.

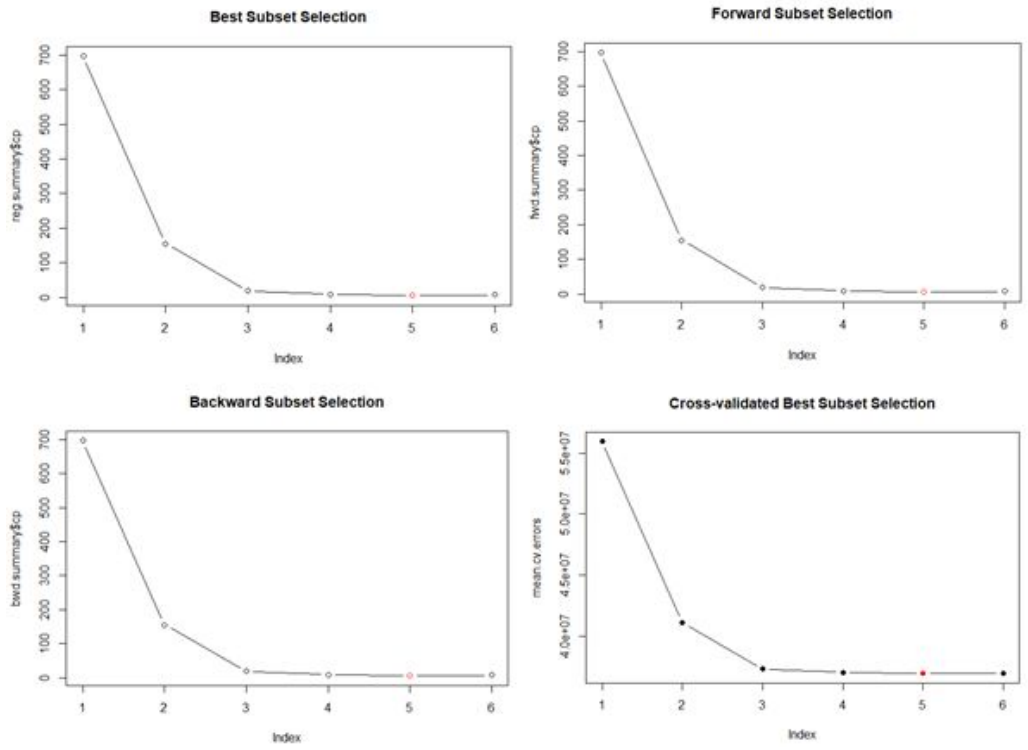

Figure 6. Plots of Subset Selection levels and their fit criterion. The red point is the optimum number of variables.

Table 2. Progression of Cross-validated Best Subset Selection. Binary mapping where asterisk indicates predictor is included in model.

```
Selection Algorithm: exhaustive
         age sex bmi children smoker region
1  ( 1 ) " " " " " " " "      "*"    " "
2  ( 1 ) "*" " " " " " "      "*"    " "
3  ( 1 ) "*" " " "*" " "      "*"    " "
4  ( 1 ) "*" " " "*" "*"      "*"    " "
5  ( 1 ) "*" " " "*" "*"      "*"    "*"
6  ( 1 ) "*" "*" "*" "*"      "*"    "*"
```

Table 3. Summary of the shrinkage and selection methods.

	Full Linear Model	**Ridge**	**Lasso**	**Subset Selection**
MSE	3.255×10^7	3.322×10^7	3.750×10^7	3.697×10^7

IV. Conclusion

Our investigation into the MCPD has shown that medical insurance costs are related to all the given predictors. However, there are certain predictors that have a heavier influence than others. For example, we have found that smoking has the strongest relationship with medical insurance costs. Predictors age and bmi have the next strongest relationship with the response.

The best prediction model for the full data set is a polynomial model to the tenth power. It is able to cover 98.99% of the variability within the dataset.

Our study included preliminary analysis of the data, simple linear regression, polynomial regression, shrinkage and selection methods. The shrinkage and selection methods turned out to be worse estimators and should not be used on this dataset. These analyses have shown that smoking heavily increases insurance charges. Further exploration could include nonparametric methods, classification and clustering.

V. Appendix

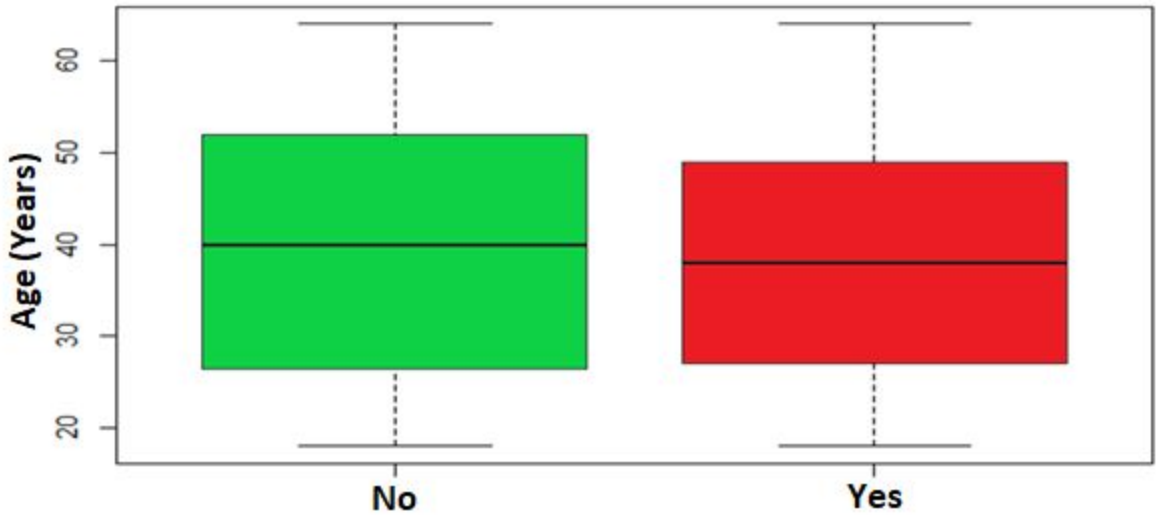

Figure A1. Boxplots of the the smoking status of individuals and their age. Non-smokers (green) and smokers (red).

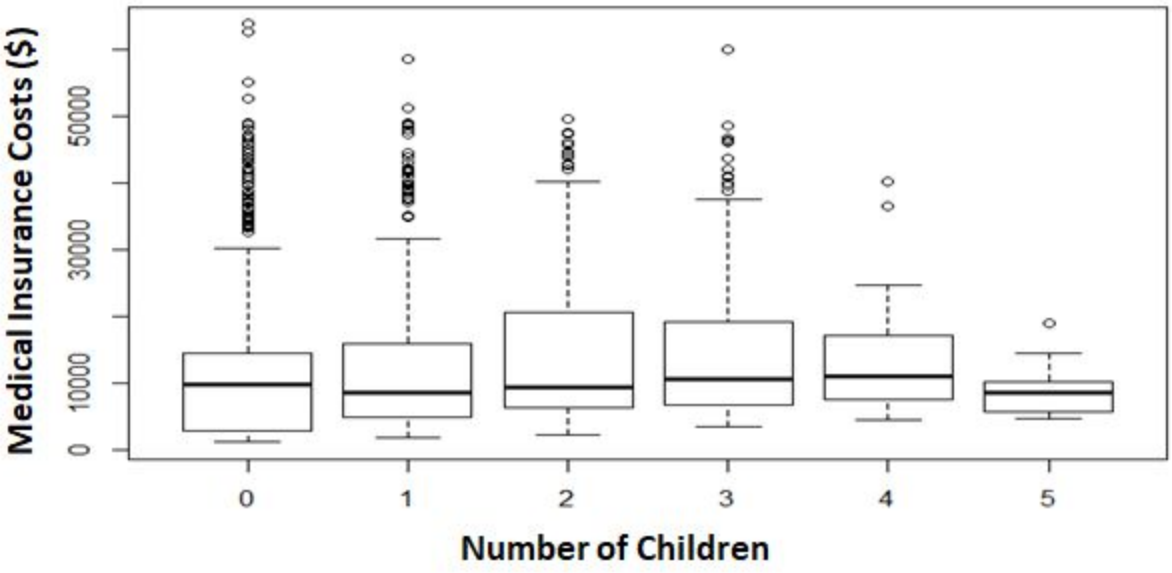

Figure A2. Boxplots comparing the number of children for the sample population and their medical insurance costs.

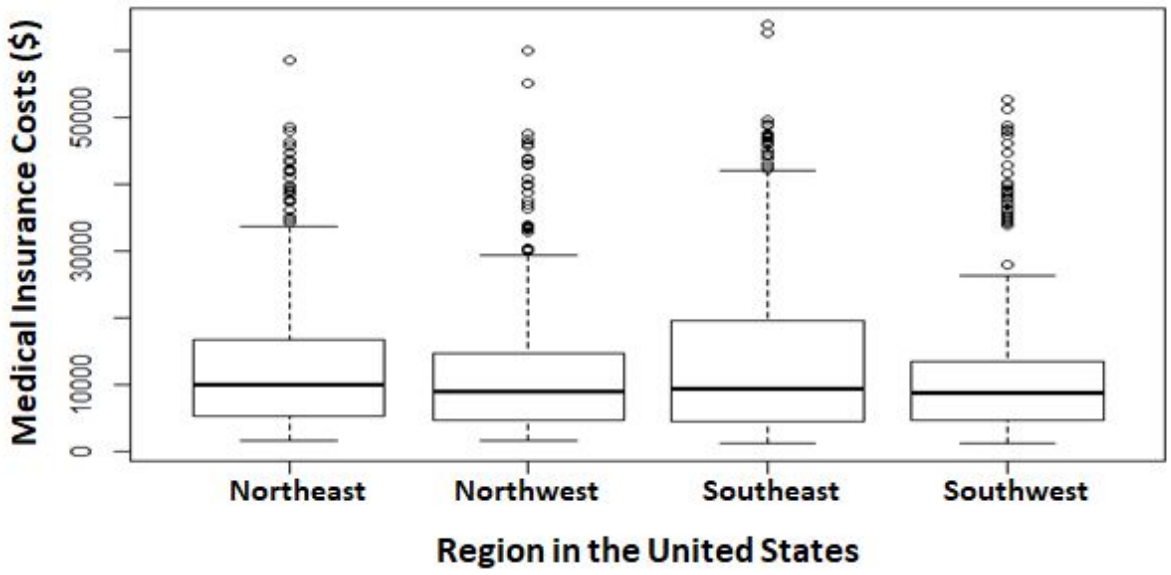

Figure A3. Boxplots of the geographical regions in the United States of the sample population and their associated medical insurance costs.