MAS 5301

Linear and Non-Linear Models

Assignment – APST/2020/03 – Dilan Dinushka

Abstract

In this document I have explored the "InsuranceCost" dataset for the target variable of annual insurance cost (Charge) related to other given parameter variables. Initial exploratory analysis showed the relations between each variable related to the predictor variable and therefore further analysis were done to identify the quantifiable associations between variables using Analysis of variance model. Based on the correlation outputs of predictor variables important variables and interaction terms were selects. Using the outcomes of those analysis a model was built which can explain the Insurance cost with about 85% R squared value.

Introduction

During these times of pandemic, obtaining an insurance has become an important decision for many people who weren't interested in such option before. But due to various concerns involving the factors insurance companies consider before giving the insurance, people hesitated to obtain an insurance. But since having an insurance seemingly becoming an interesting option, in this study I am hoping to analyze one important aspect in this insurance selection, the cost. Usually, to determine the payment which the people required to pay to the insurance company, insurance agents use various parameters such as person's current health, habits, income, family background etc. In this study I am hoping to understand the relationship between such parameters and Insurance cost. Also, I will be consider fitting a mathematical model to the obtained dataset which can be used to predict the possible payment cost which a person may required to pay for the insurance company.

Methodology

To carry out the above tasks, following statistical techniques were used.

First, for all the random variables exploratory analysis was done, scatter plots were graphed to visually identify the existing relationships with the target variable "Insurance Cost".

Then to get the numerical verification of variable significance full factorial Analysis of Variance (ANOVA) has been done for all the variables related to the aforementioned predictor variable.

Then based on above result a preliminary Linear Regression model has been fitted to get the base line model and to identify the base performance. Afterwards using the "Leaps" R library I have selected the best subset of random variables that provides highest adjusted R squared value and fitted a new linear model.

Finally in the analysis and conclusion section all the results obtained were discussed and summarized.

Explanatory Analysis

--Note all the experiments were done on a training set with 80% of data from original data set.

During the initial exploration of the dataset, we can see that there are 6 variables including the target variable “Charge”.

Age <int>	Gender <chr>	BMI <dbl>	Childrens <int>	Smoker <chr>	Region <chr>	Charge <dbl>
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
46	female	33.440	1	no	southeast	8240.590

When considering the above data, we can see that the gender, smoker and region variables acts as categorical variables. Interestingly children column has values which seemingly look like categorical. Below shows data distribution of Children column.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	1.000	1.071	2.000	5.000

As we can see maximum value of children is 5 and mean value is close to 1. Therefore, in this analysis it is safe to use consider this as a categorical value rather than a numerical value due to its applicability.

For other variables descriptive details are as follows.

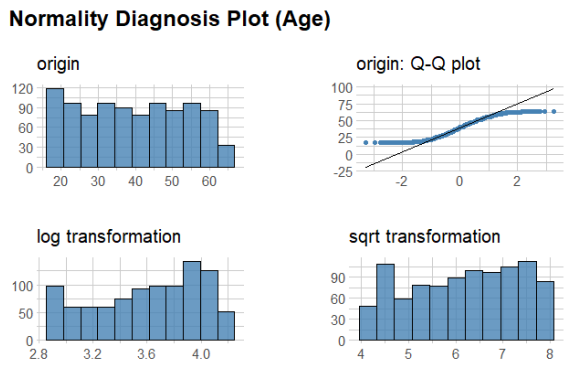
variable <chr>	n <int>	na <int>	mean <dbl>	sd <dbl>	se_mean <dbl>	IQR <dbl>
Age	956	0	39.35146	14.054937	0.4545692	24.00000
BMI	956	0	30.65335	6.187787	0.2001273	8.49375
Charge	956	0	13264.48092	11946.391217	386.3739045	11822.48001

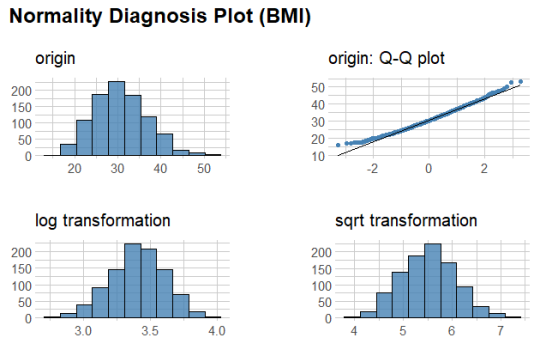
(n = number of values excluding missing values, na = number of missing values. sd = Standard deviation, se_mean = Standard error mean, IQR = Inter Quartile Range)

Then for the all the numerical data, normality test (Shapiro-Wilk test) was done to check the normality assumption which is important for the inferencing related to Linear regression models.

vars <chr>	statistic <dbl>	p_value <dbl>
Age	0.9453105	2.554799e-18
BMI	0.9914427	2.392480e-05
Charge	0.8165716	1.256862e-31

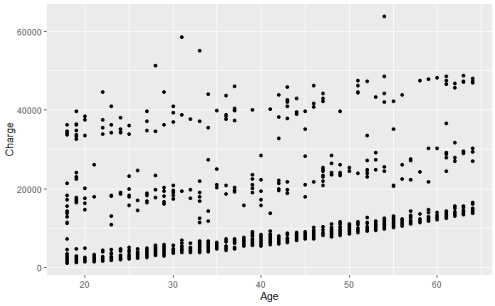
According to the above test we can see that all the numerical variables do obey the normality assumption with a very significant confidence level according to the p values.

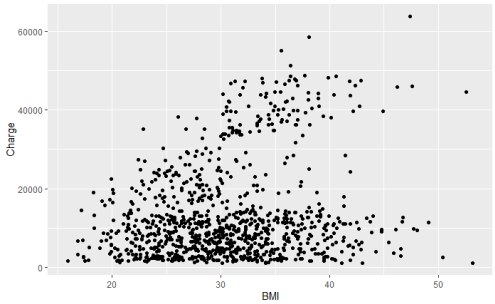




According to the graphs BMI follows a near perfect normal distribution but age is bit deviating according to the Q-Q plot. But this can be understandable since insurance cannot be obtained by underage people and it is hard to get insurance to older people. Therefore, it is safe to assume the normality for age as well.

Below include the scatter plots representing relationships between target and predictor variables.





Advanced Analysis

As the first step we will consider the ANOVA for all variables to identify the correlations between each variable as it is an important factor in modelling. The interactions between main effects suggests that all the predictors are significant compared to region and gender at 5% level.

(Charge=Age+BMI+Smoker+Region+Gender+Childrens)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Age	1	1.018e+10	1.018e+10	293.255	<2e-16	***
BMI	1	4.617e+09	4.617e+09	133.043	<2e-16	***
Smoker	1	8.806e+10	8.806e+10	2537.652	<2e-16	***
Region	3	2.417e+08	8.057e+07	2.322	0.0737	.
Gender	1	1.132e+07	1.132e+07	0.326	0.5681	
Childrens	5	4.620e+08	9.239e+07	2.662	0.0212	*
Residuals	943	3.272e+10	3.470e+07			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

When we consider the base model with only the main effects the model performance is as follows.

```
Call:
lm(formula = Charge ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11680.5  -2935.7   -814.4   1715.1  24964.3
```

```
Residual standard error: 5891 on 943 degrees of freedom
Multiple R-squared:  0.7599,    Adjusted R-squared:  0.7568
F-statistic: 248.7 on 12 and 943 DF,  p-value: < 2.2e-16
```

This base model was able to explain the 76% of training dataset variance according to the adjusted R squared values.

But with further analysis with Anova for full factorial suggested that following interactions were significant at 5% level as well.

Variable	DF	Sum Sq	Mean Sq	F Value	Pr(> F)	
Age	1	1.018e+10	1.018e+10	488.571	<2e-16	***
BMI	1	4.617e+09	4.617e+09	221.654	<2e-16	***
Smoker	1	8.806e+10	8.806e+10	4227.795	<2e-16	***
Region	3	2.417e+08	8.057e+07	3.868	0.009240	**
Childrens	5	4.620e+08	9.239e+07	4.436	0.000552	***
BMI:Smoker	1	1.246e+10	1.246e+10	598.395	<2e-16	***
BMI:Region	3	1.747e+08	5.825e+07	2.796	0.039403	*
BMI:Smoker:Region	3	1.721e+08	5.736e+07	2.754	0.041709	*
Age:Region:Childrens	14	6.268e+08	4.477e+07	2.149	0.008370	**
Smoker:Region:Childrens	9	4.184e+08	4.649e+07	2.232	0.018531	*
Age:BMI:Smoker:Childrens	3	1.731e+08	5.771e+07	2.771	0.040772	*

But with the usage of interaction terms along with the hierarchical principle we can obtain a model with outperform the basic model which can explain the variance of the training dataset up to 86%. (Did not used leaps library to perform subset analysis due to computation complexity)

```
lm(formula = Charge ~ Age + BMI + Smoker + Region + Childrens +
    BMI * Smoker + BMI * Region + BMI * Smoker * Region + Age *
    Region * Childrens + Smoker * Region * Childrens + Age *
    BMI * Smoker * Childrens, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10970.7  -1935.5   -903.0    273.3  22877.5
```

```
Residual standard error: 4530 on 861 degrees of freedom
Multiple R-squared:  0.8704,    Adjusted R-squared:  0.8562
F-statistic:  61.5 on 94 and 861 DF,  p-value: < 2.2e-16
```

This may happen due to overfitting to training dataset which get confirmed by the comparison of residual mean squared error between base model (6480.319) and above model (6516.65).

Therefore, finetuning the model by removing few interaction terms (backward elimination) following model was obtained which provided much better results for the training set while eliminating the overfitting effect.

Final Model:

Charge =
Age+BMI+Smoker+Region+Childrens+BMI*Smoker+BMI*Region+Smoker*Region+Age*Region +
Intercept

```
lm(formula = Charge ~ Age + BMI + Smoker + Region + Childrens +
    BMI * Smoker + BMI * Region + Smoker * Region + Age * Region,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12528.3  -2006.7  -1067.1     23.4  23343.3
```

```
Residual standard error: 4616 on 934 degrees of freedom
Multiple R-squared:  0.854,     Adjusted R-squared:  0.8507
F-statistic: 260.2 on 21 and 934 DF,  p-value: < 2.2e-16
```

This model provides an adjusted R squared value which translate to 85% variance of the training data set while maintaining a considerably low residual mean squared error of 5359.454 compared to base model 6480.319.

Conclusion and Discussion

In this analysis, based on the initial exploratory analysis and secondary analysis I identified the factors and interactions that could accurately model the Insurance Cost. Even though the main interactions showed that region was not significant at 5% level, factorial analysis showed the significance of it with the help of interactions between other variables which helped to build a significantly better model (by 10%) compared to the base model.

One of the main problems encountered during the advanced analysis phase was variable subset identification. But due to complexity when running leaps subset function it takes considerable amount of time. Also, one interesting observation was significance of predictor variable “Region” which was not significant in main effect analysis but became significant with the association of other variables.

References

https://towardsdatascience.com/selecting-the-best-predictors-for-linear-regression-in-r-f385bf3d93e9

https://www.statology.org/linear-regression-assumptions/

https://www.r-bloggers.com/2013/08/exploratory-data-analysis-useful-r-functions-for-exploring-a-data-frame/

https://cran.r-project.org/web/packages/dlookr/vignettes/EDA.html

https://cran.r-project.org/web/packages/dlookr/vignettes/transformation.html

https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0