CSDA 5110

Predicting Insurance claims

By

Hirdesh Sandiri

1. **Introduction**   
   1. **About Dataset**

I draw Medical Cost Personal Datasets from [Kaggle](https://www.kaggle.com/datasets/mirichoi0218/insurance). Individual insurance information is contained in this dataset. Analyzing this data can help insurance companies better understand their customers and make strategic business decisions. With the help of the individual information, the cost of medical expenses can be estimated.

* 1. **Variables**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Measurement Type** |
| Age | age of primary beneficiary | Number |
| Sex | insurance contractor gender, female, male | Male / Female |
| BMI | Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9 | Number |
| Smoker | Smoking | Yes / No |
| Children | Number of children covered by health insurance / Number of dependents | Number |
| Region | The beneficiary's residential area in the US, | northeast, southeast, southwest, northwest |
| Charges | Individual medical costs billed by health insurance | Number |

*Table 1: Dataset Information*

* 1. **Outcome variable(s)**

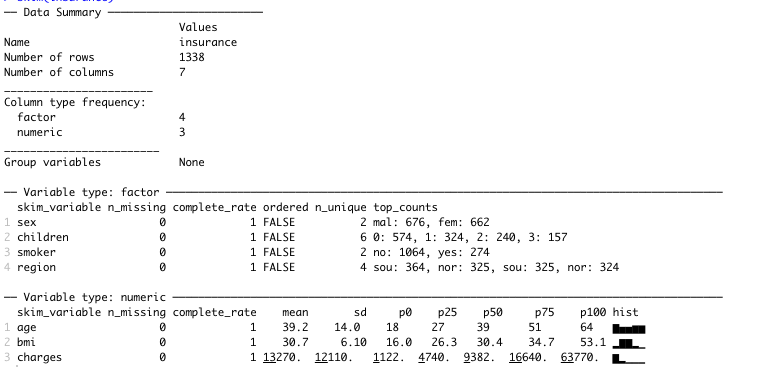
Variable *“Chargers”* can be the outcome variable. Which are Individual medical costs billed by health insurance. We can estimate medical expenditures based on information about specific patients.

* 1. **Predictor variables**

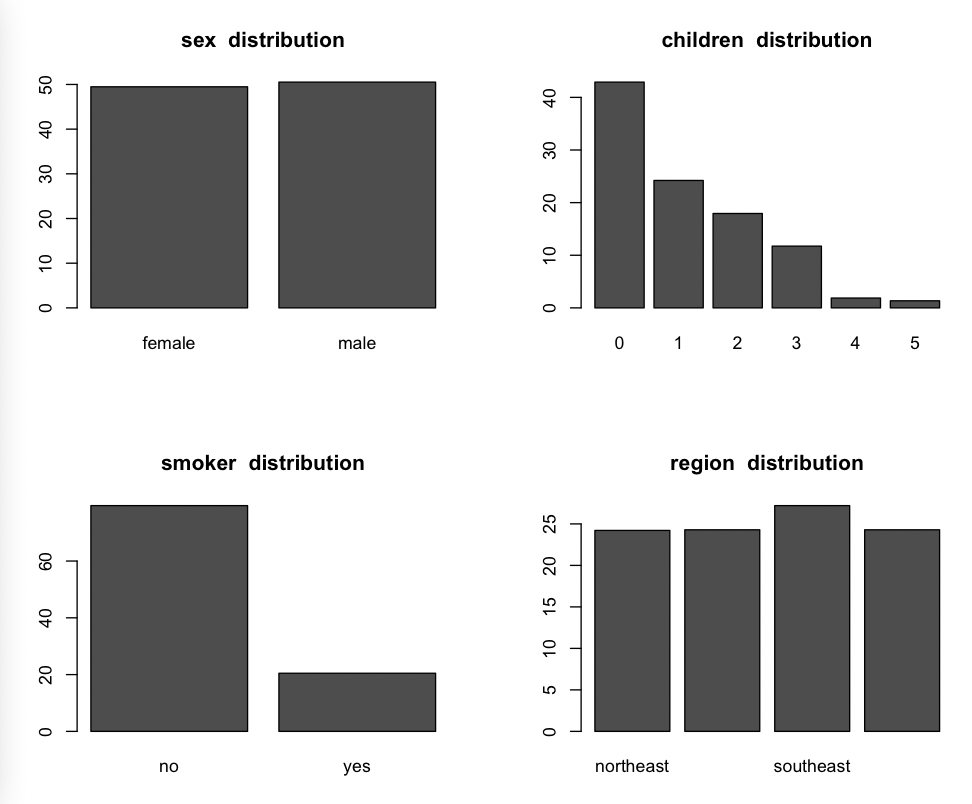
I’m considering all other variables other than *Chargers* Outcome variable as predictor (Independent variables). In which Sex, Smoker, Region variables are as factors. And these are fixed factors because these variables have fixed factors which has to be from the given levels only. (Mentioned in the above table). I’m also considering children as factor. As it is a discrete variable. There are covariates independent variables as well such as BMI, Age. A covariate is usually a continuous variable that is expected to change (“vary”) with (“co”) the outcome variable of a study. Generally speaking, a covariate can refer to any continuous variable that is expected to correlate with the outcome variable of interest (Ruppel, 2017)[1].

* 1. **hypothesis**
* To find if there is a relationship between predictive variables and outcome variable.
* To find if there is any significance difference between medical costs and different group of people (may be by region, age).
* See If we can accurately predict costs using linear regression.

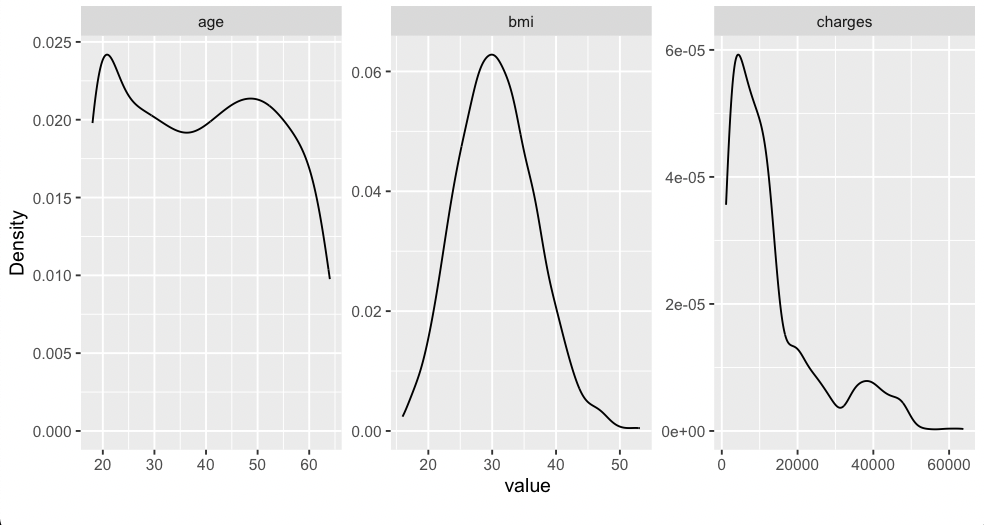
1. **Preliminary Exploration of the Data** 
   1. **Overview of each variable – Univariate Analysis**



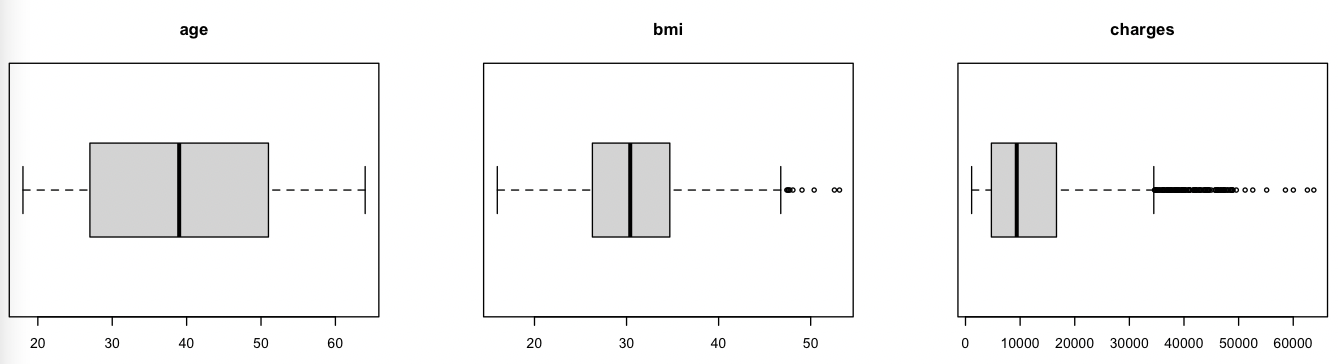
*Fig 1: Summary of dataset*



*Fig 2: factorial data percentage distribution*

**

*Fig 3: Numeric Density plots*

**

*Fig 4: Boxplot to show distribution of Numeric fields*

Above figures explains the summary of the whole data we have. Below are few observations that I made from the summary.

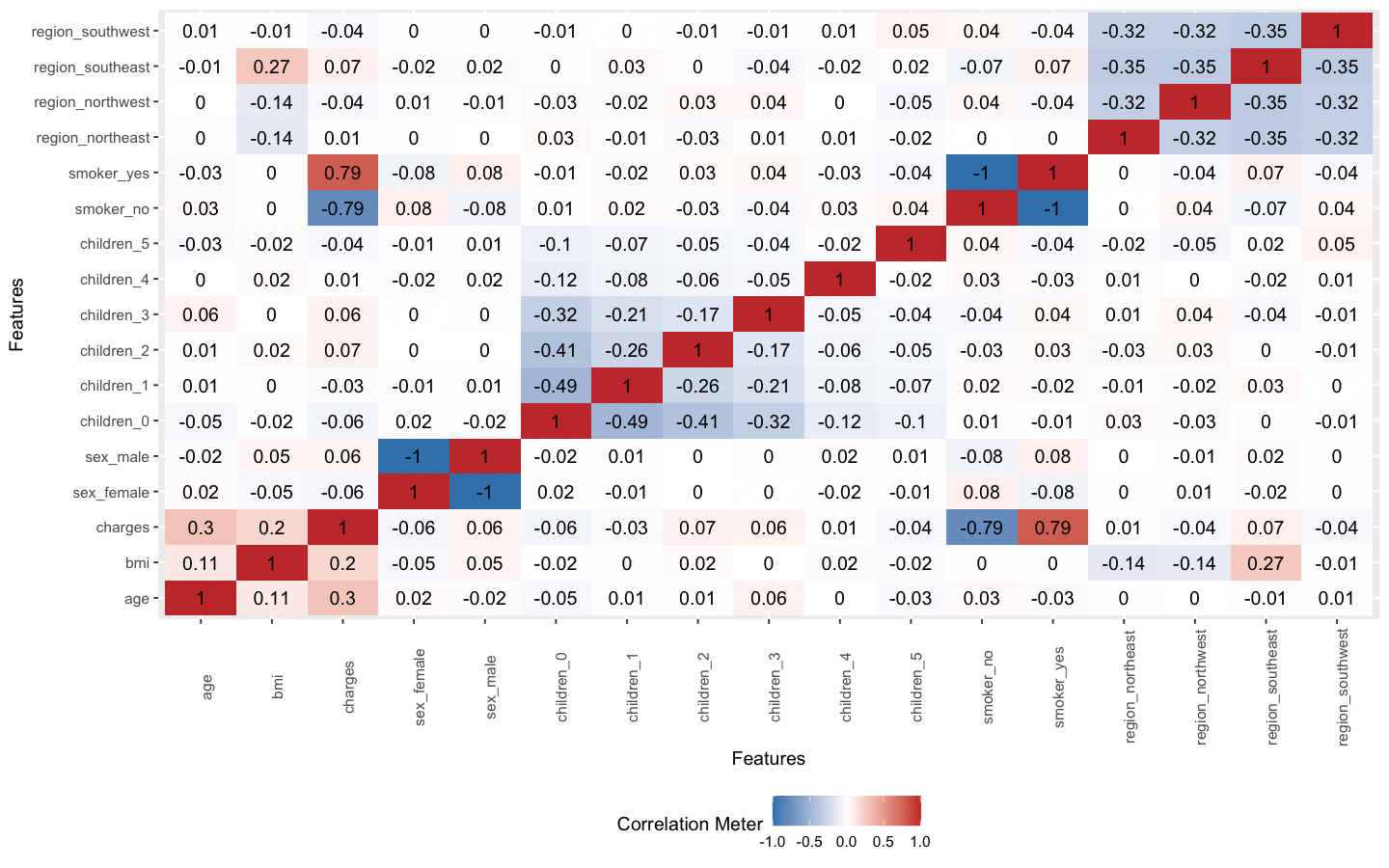
* There are no missing values in the data.

*From Fig 2*

* Sample also indicates there are slightly more **male** beneficiaries than females
* Sample also has more **nonsmokers** – 1064 than the smoker – 274
* Southeast **region** has more claims compared to the other regions.
* Most people (abv 40% from Fig 2) do not have dependents (**Children**), and Most people who covered their dependencies in insurance was having only one child.

*From Fig 3 & Fig 4*

* Median (p50 from Fig1) is less than the mean in **charges** – which means distribution is positively skewed. There are so many outliers on right side.
* Mean is almost equal to median for both Age and **BMI**. This can conclude it is normally distributed. There are few outliers on right side.
* From Fig 3 and Fig 4 age plots shows most people who got insurance are covered in between 18 – 23. The abv 60 years old are very few who got insurance. This shows that govt is covering the insurance for abv 60 years by default.
  1. **Bivariate & Multivariate Analysis**

****

*Fig 5: Correlation between all the variables*

*Observations*

* I see slight positive relation between
  + Charges – Age
  + Charges – BMI
  + Charges – Smoker YES
  + BMI – region southeast
  + Age – BMI
* I see slight negative relation between
  + Charges – Smoker NO
  + BMI – region Northeast
  + BMI – region Northwest

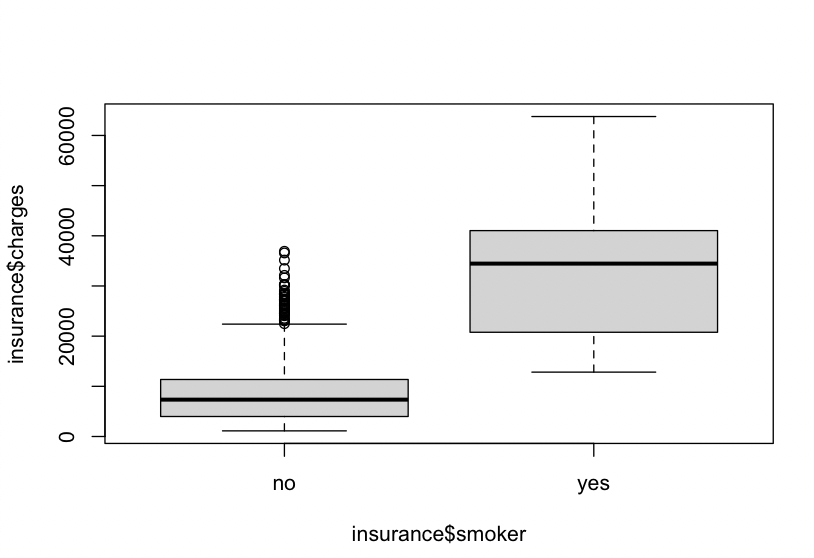
Lets’ dive deep with these relations and see how they are related.

|  |  |
| --- | --- |
|  |  |
| *Fig 6: Charges vs Age* | *Fig 7: Charges vs Age with Smoker* |

There is clear positive relation between age and charges. As age is growing the charges are also relativly more and Fig 7 show two different trend lines with respective to people who smoke.

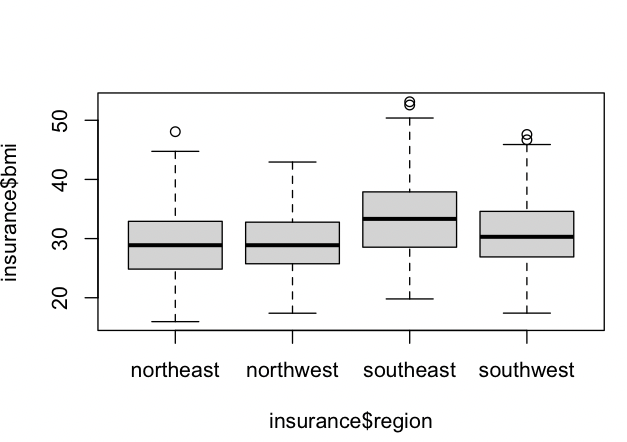
|  |  |
| --- | --- |
|  |  |
| *Fig 8: Charges vs BMI* | *Fig 9: Charges vs BMI with Smoker* |

There is clear positive relation between Charges and BMI. Especially with Smoker group there is very high positive relation. As BMI is more and if the they are also smoker – Then there will very high insurance costs.

**

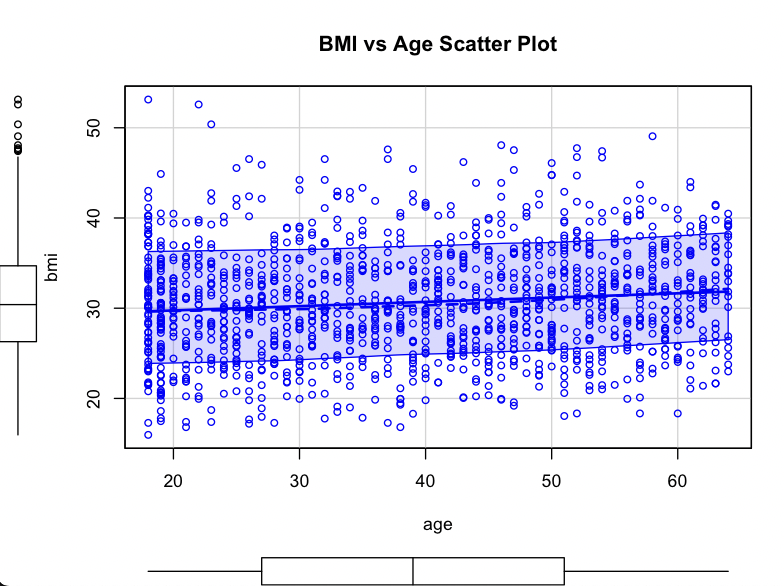
*Fig 10: Charges vs Smoker*

From Fig 10. There is no overlap between two boxplots. So, the charges are very different between smokers and nonsmokers. For smokers there is high insurance costs and for nonsmokers less insurance.

**

*Fig 11: BMI vs Region*

From Fig 11: I can say southeast region is standing out from crowd. There is more user with higher BMI in southeast region. And Northern region has lower BMI users.

****

*Fig 12: BMI vs Age*

Cleary there is a slight positive relation between BMI and Age. As Age is increasing BMI is also increasing.

1. **Analysis methods**

**Objective for analysis**

* Prove: EDA has indicated that charges are more if the policy holder are smokers. So, I’m trying to prove that smoking has effect on insurance charges.
* Test: if the proposition of smokers is significantly different across different regions.
* Prove: Overall, I want to prove, that there will a significance prediction based on all independent factors.

**Analysis Methods:**

* I will be using simple t-test (Welch Two Sample t-test) to prove my 1st points from the above.
* Use Kruskal test for 2as there are more than 2 groups to test between charges and regions.
* And use multi linear regression analysis method to prove the significance prediction.

**Analysis Results:**

|  |  |  |
| --- | --- | --- |
| Assumptions | Result | Evidence |
| H0: Avg charges of smokers <= non-smokers  Ha: Avg charges of smokers > non-smokers | T- test p -value < 2.2e-16  P - value < significance level 0.05  Reject null hypothesis | There is clear evidence from EDA Above to prove this point. Fig 10 |
| H0: There is no significance difference b/w regions w.r.t charges  Ha: There is significance difference b/w regions | Kruskal test p - value = 0.1923  P - value > significance level 0.05  Failed to reject null hypothesis | This proved that there is no significance difference b/w regions. Even from EDA analysis we didn’t find any point that correlate charges with region. Fig 5 |
| H0: There is no significance prediction with all predictors and charges  Ha: There is significance prediction with all predictors and charges | Linear regression P – value < 2.2e-16  P - value < significance level 0.05  Reject null hypothesis | This proved that we get significance prediction with some independent predictive variables. Fig 13 give more info on p-values for individual factors. |

Table 2: Analysis Results

*Linear regression model summary:*

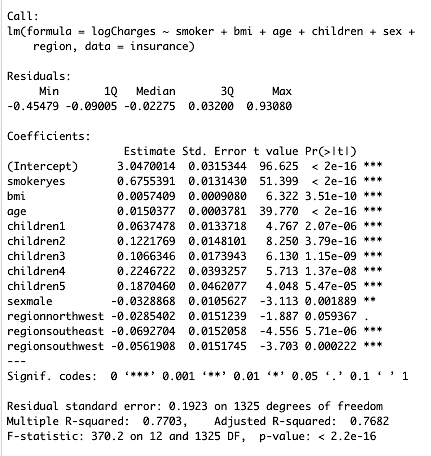


Fig 13: LM summary

1st and last assumption which are satisfied can be generalized and also used in prediction as there is enough evidence that there is a significance relation between the variables. That could impact the outcome variable. 2nd assumption is not something cannot generalize, but can be used in prediction as regions are supporting factors for some other variables.

1. **Conclusions**

Hence, I proved smoking has high effect with the charges with both EDA and also with hypothesis testing. Increase in Age and BMI has small increases in charges. Having more dependents (Here children) can also increase expenses with insurance charges. Another important point from linear regression is that sex$male is also showing the significance relation with the outcome value even though there is very slight difference between male and female mediums.

Lastly from the model predictions resulting with 8760.206 $ difference in error. If company can bare this about of difference, they can use this model to predict charges with the individual information.

1. **References**:
2. Ruppel, Erin K., ed. ***The SAGE Encyclopedia of Communication Research Methods***. 4 vols. Thousand Oaks, CA: SAGE Publications, Inc, 2017. <https://doi.org/10.4135/9781483381411>.