Health Care Cost Prediction with Linear Regression Models

608 – Applied Modeling and Optimization | ECE Department | Fall 2021 | Prof. R. Chandramauli

Presented by : Swapnali Patki- 10477374, Ritesh Panditi- 20007213, Mikey Brewer- 10478352

What are we trying to do?

The rising expense of health care is a serious public health concern. As a result, precisely estimating future expenses and knowing which factors contribute to rising health-care costs are critical. The goal of this project is to predict how patients' healthcare expenses will rise over the next year and to determine the elements that influenced this assessment and identify the impact of a variety of factors on insurance prices and to forecast the cost of health insurance depending on those factors. Multiple linear regression was utilized in the analysis.

Objective Function and Constraints

Objective Function:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

- y is the dependent vector to be predicted (output variable).
- β_p is a parameter vector or regression coefficient.
- x_i is an independent input variable.

Constraints:

1. Age: insurance contractor age, years
2. 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
3. Smoker: smoking, [yes, no]
4. Charges: Individual medical costs billed by health insurance, $ #predicted value

```
[ ] #Displaying the top five data from the dataset:
    insuranceData.head(5)
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Process

- Initializing the environment and importing the data.
- Initial Exploratory Data Analysis
- Visualization
- Building the Linear Regression Model
- Fitting(Training the model on) the data
- Validating the statistical parameters
- Testing the model with new data

Required Tools, packages and Data Imports

```
#initialization of the library and packages:
import pandas as pd
from plotnine import *
from matplotlib import *
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline
sns.set()
plt.rcParams['patch.force_edgecolor'] = True
plt.rcParams['axes.labelweight'] = 'bold'
import warnings
warnings.filterwarnings('ignore')
from sklearn.linear_model import LinearRegression, RidgeCV, LassoCV
```

Importing the dataset and first introspection

```
[25] ##Preparing the dataset:
     insuranceData = pandas.read_csv('insurance.csv')
```

Multiple Linear Regression:

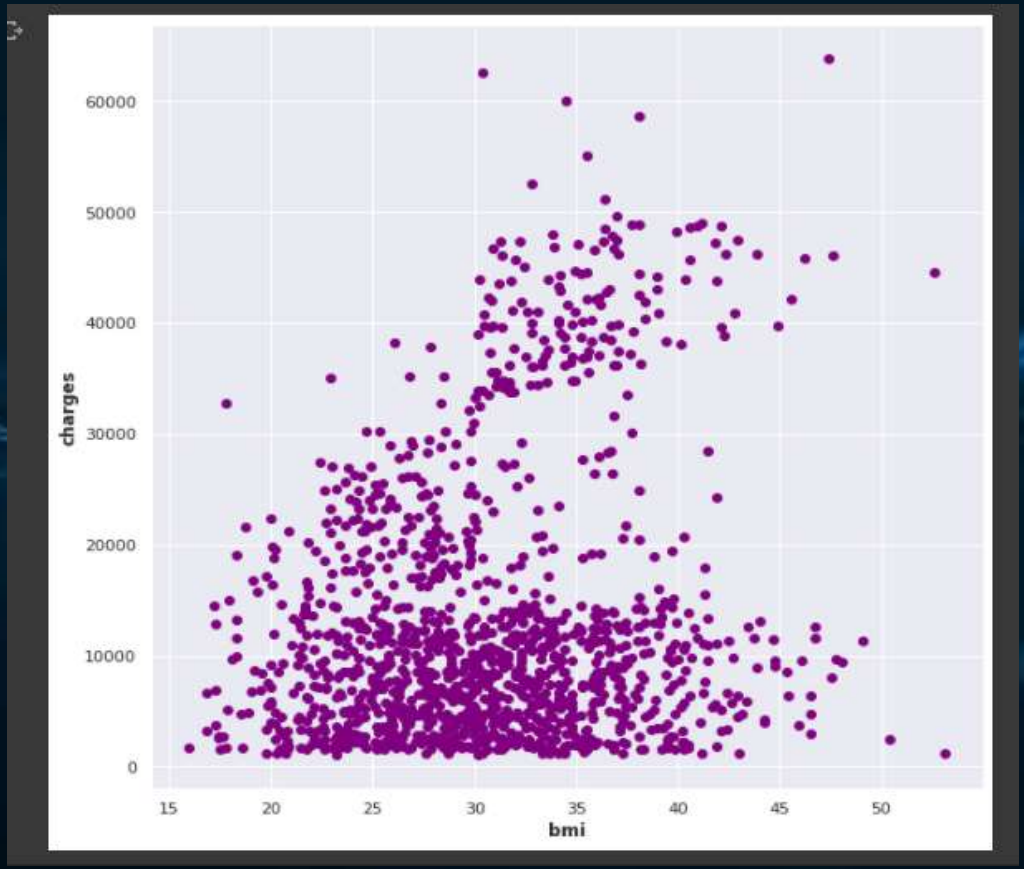
A linear connection exists between two or more independent variables (X1, X2, X3,..., Xn) and the dependent variable in multiple linear regression analysis (Y).

The purpose of this investigation is to evaluate the vision of the relationship between independent variables and dependent variables, whether each independent variable is positively or negatively associated, and to predict the value of the dependent variable if the value of the independent variable rises and falls.

For this project, we have age, bmi, and smoker_yes as independent variables, and charges as the dependent variable.

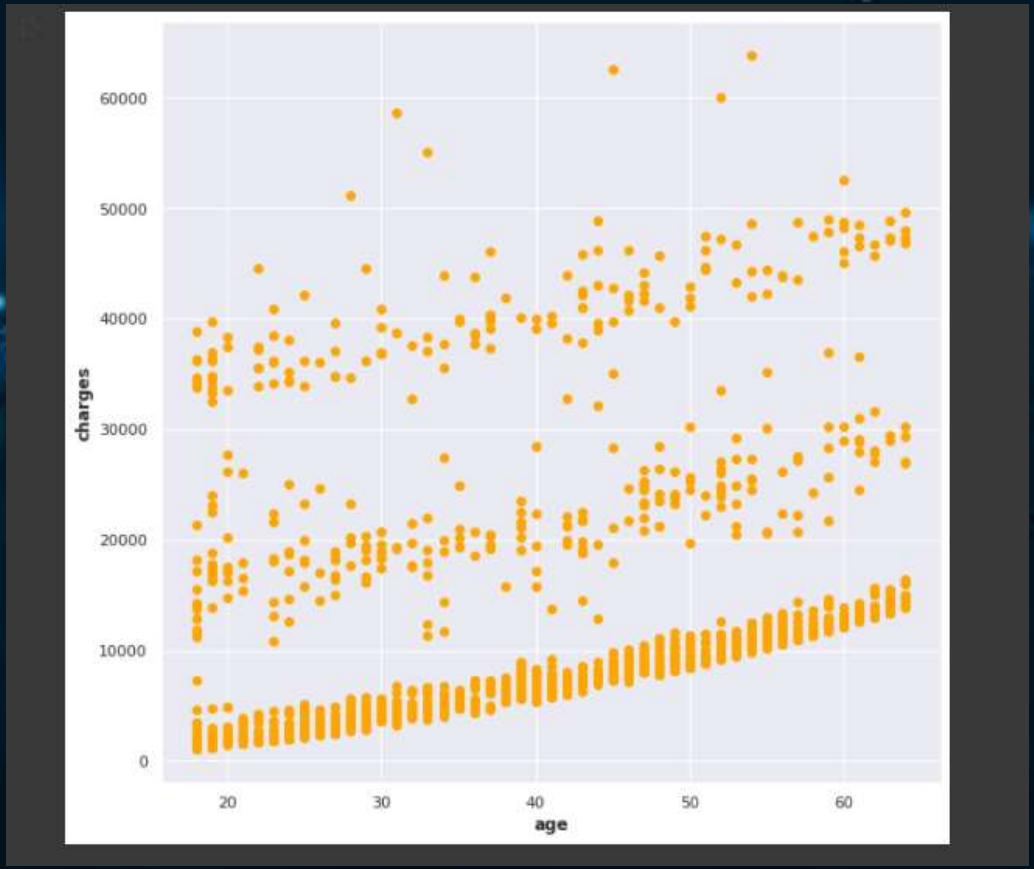
Exploratory Data Analysis and Visualization

- Plot 1: Charges vs BMI



For a person with normal (18-25) BMI, their charges don't go above $40,000

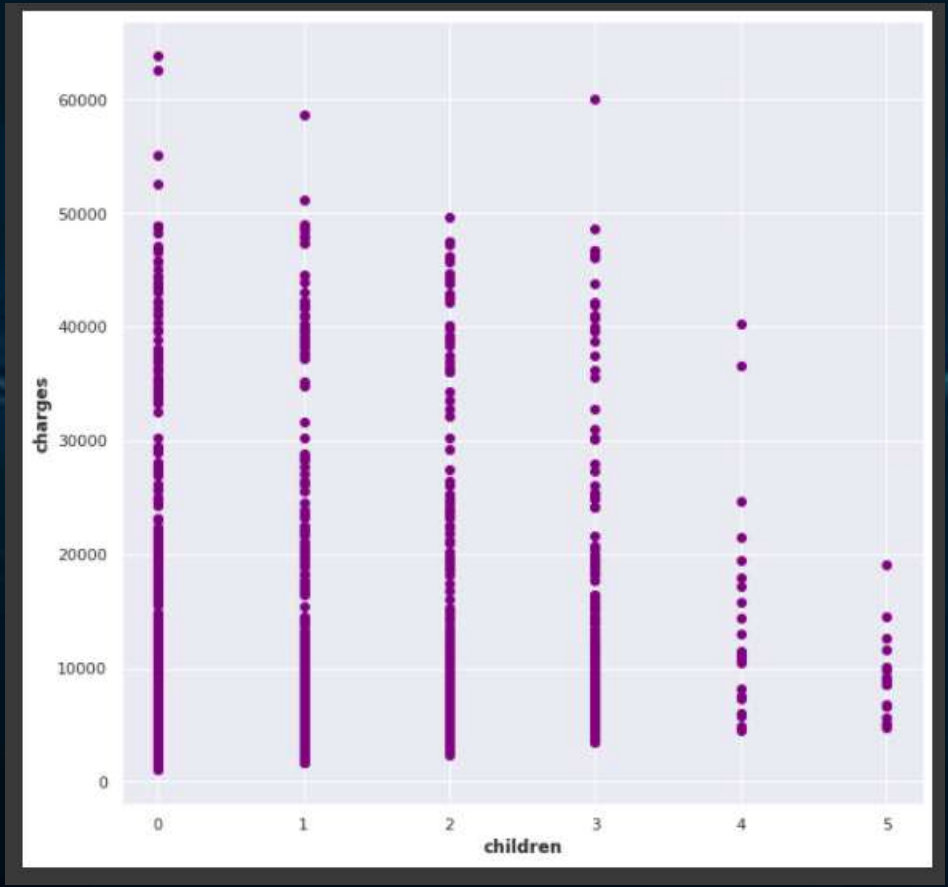
Plot 2: Charges vs Age



No obvious connection between Charges and Age.

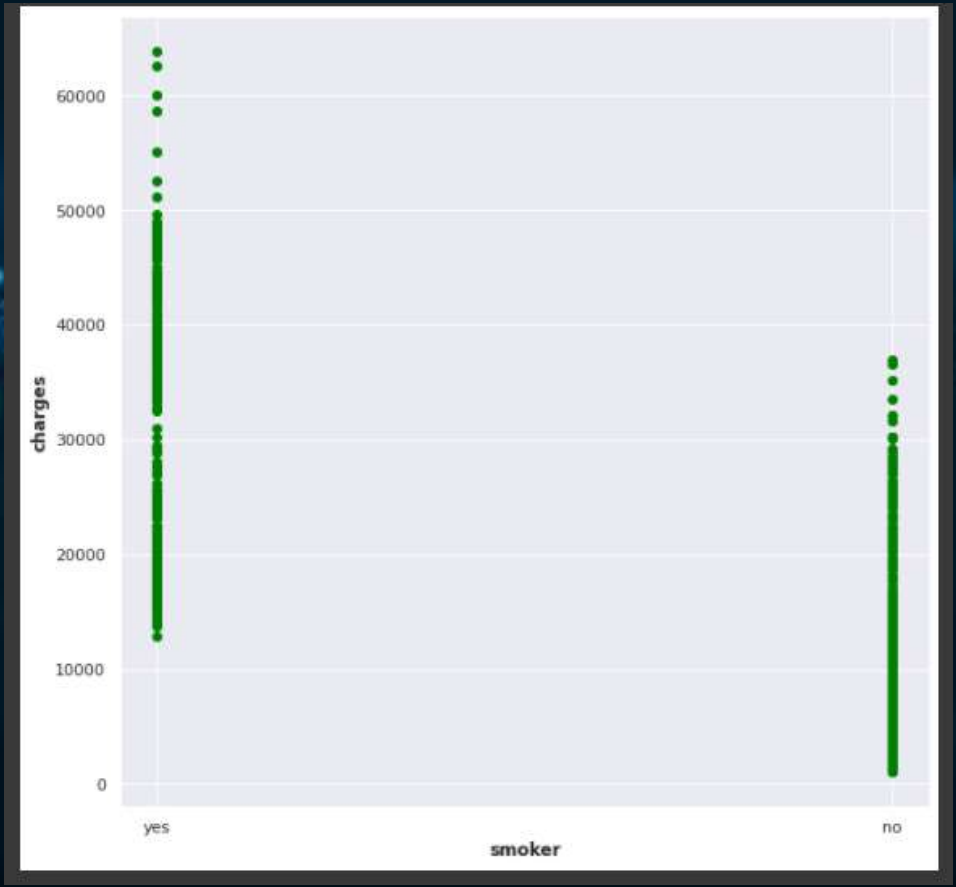
Exploratory Data Analysis and Visualization

Plot 3: Charges vs Children



Interestingly, people with more than 3 kids have lesser charges than the rest.

Plot 4: Charges vs Smoker

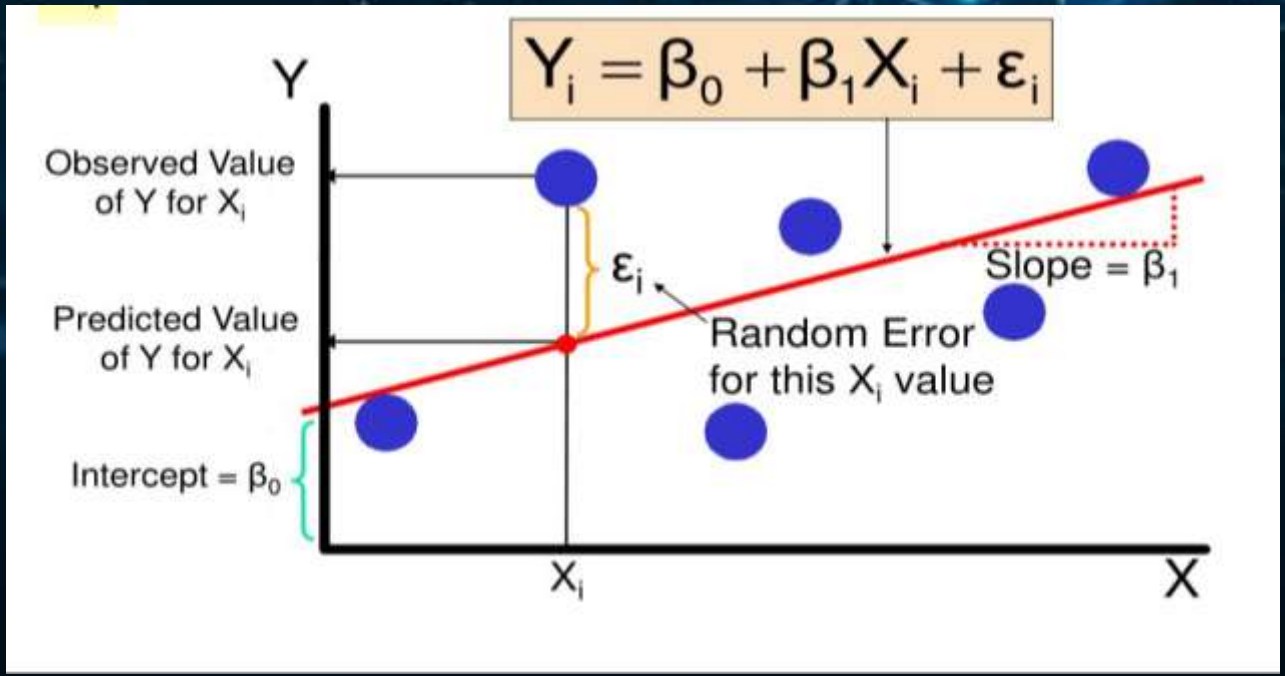


As expected, charges for smokers are higher than non-smokers

Optimization Model

Linear regression:

In regression, we want to predict a scalar-valued objective, such health-care costs. The target must be predicted as a linear function of the inputs if we use the term "linear“. This is a supervised learning algorithm, in which we have a set of training samples that have been labeled with the proper outputs.



Implemented Linear regression Model

- Linear Regression model was implemented using the sklearn library.
- Regularization was performed using RidgeCV and LassoCV.
- The prime motive of regularization here is to prevent overfitting and avoid bias in the model.
- The lasso estimate thus solves the minimization of the least-squares penalty with α||w||1 added, where α is a constant and ||w||1 is the L1-norm of the coefficient vector.
- The Mean Squared Error (MSE) reduces with the use of L-2 norm, Ridge Regressor

```
#Building the Regression Model and defining the regularization parameters:
x = data_dum[['age', 'bmi', 'smoker_yes']] #independent variables
y = data_dum['charges'] #dependent variables
linreg = LinearRegression()
# RidgedCV for 5-fold Cross Validation and setting up the regularization strength(alphas):
ridgereg = RidgeCV(alphas=(5.1,0.5,0.8,1.0),cv=5)
lassoreg = LassoCV(eps=0.00001, cv = 5)
from sklearn.model_selection import cross_val_score
```

Findings from the Model

R-square value is calculated as

```
[ ] #Fitting the model
    linreg.fit(x,y)
    #Display the R square value
    linreg.score(x,y)
```

```
0.7474771588119513
```

The smaller the R-square value, the weaker the influence of the independent variable on the dependent variable, and vice versa.

Bring up the coefficient and intercept values

```
[ ] #Obtaining the coefficients B1,B2,B3 of the linear regression model
    linreg.coef_
```

```
array([  259.54749155,   322.61513282, 23823.68449531])
```

```
[ ] #Intercept value (B0)
    linreg.intercept_
```

```
-11676.830425187782
```

Based on the output, the intercept value is -11676.830. And beta coefficients B1, B2, and B3 are 259,547, 322,615, and 23823,684 respectively. The regression model can be written, y = -11676.830 + 259.547x1 + 322.615x2 + 23823,684x3.

In the obtained model, x1,x2,x3 are age, BMI and smoker_yes respectively which are the independent variables.

Prediction from the Model

After we get a regression model, we try to make a prediction using the regression model.

```
[41] #Defining a function to predict the cost
     def calc_insurance(age, bmi, smoking):
      y = ((age*linreg.coef_[0]) + (bmi*linreg.coef_[1]) + (smoking*linreg.coef_[2]) - linreg.intercept_)
      return y
```

we try to predict how much insurance costs from someone who is 36 years old, the value of BMI is 24, and not a smoker

```
print(calc_insurance(36, 24, 0))
print(calc_insurance(36, 24, 1))
print(calc_insurance(36, 32, 1))
print(calc_insurance(36, 32, 0))
```

```
28763.303308712795
52586.98780402163
55167.908866574275
31344.22437126544
```

From the predicted result, we can understand that the model gives us the expected values which correlates the given parameters like age, BMI and if smoker. We can see that the charges for a smoker are much higher than a non-smoker and the BMI doesn't affect the charges as significantly.

What we took away from this project?

- We learned to apply the optimization method like linear regression, got to explore more other methods and ways to solve the problem.
- In the pursuit of doing an Optimization Project in Healthcare, we realized the difficulty of the data-gathering process. We understood the difficulty of building an optimization model with the hundreds of different types of parameters that can be factored in the healthcare sector.
- We also looked at the possibility of Dynamic Simulation Models and how they compare to various constrained optimization methods.
- We realized the infeasibility of completing a large-scale project in the given timeline and implemented a Linear Regression model on some pre-processed insurance data to predict the charges in correlation to age, BMI and smoker .

Areas to Improve

- More independent variables and constraints to consider for the Prediction.
- Using a much larger significant and consistent dataset to the model can make the accurate prediction.
- Various models can be used to get the prediction and compare the results.
- Ensemble methods can be used to get more accurate predictions.

References and Citations:

- B. Nithya, Dr. V. Ilango,"Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems ICICCS 2017, 978-1-5386-2745-7/17/$31.00 ©2017 IEEE.
- A. Tike and S. Tavarageri. (2017). A Medical Price Prediction System using Hierarchical Decision Trees. In: IEEE Big Data Conference 2017. IEEE, 978-1-5386-2715-0/17/$31.00 ©2017 IEEE.
- Lahiri and N. Agarwal, "Predicting healthcare expenditure increase for an individual from medicare data," in Proceedings of the ACM SIGKDD Workshop on Health Informatics, 2014.
- Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti, and E. Pagano, "Regression models for analyzing costs and their determinants in health care: an introductory review," International Journal for Quality in Health Care, vol. 23, no. 3, pp. 331–341,2011.
- Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G.Wang, "Algorithmic prediction of health-care costs," Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.
- https://www.kaggle.com/
- https://scikit-learn.org/stable/modules/linear_model.html