Universidad del Valle de Guatemala

Andrés Quan-Littow

Apr. 20, 2021

Lab 4: Regression

This lab was about regression, which can take up into several different shapes, but is mostly seen as just a certain function, of a certain degree, that can describe the behavior of a certain dataset. These lines can be near-constant (degree 0), straight (degree 1) or curved (degree ≥ 2). This is because we use at least two different kinds of functions:

$$y = b_0 + b_1 x_0 + b_2 x_1 + \cdots + b_n x_{n-1}$$

Which can be seen as

$$y = \sum_{i=0}^{n} b_i x_{i-1}$$

And the polylinear equation can be seen as

$$y = b_0 + b_1 x_0 + b_2 x_1^2 + b_3 x_2^3 + \cdots + b_n x_{n-1}^n$$

Which can be seen as

$$\sum_{i=0}^{n} b_i x_{i-1}^i$$

Dataset Information

The dataset is that of an insurance company. It contains the age, the sex, the BMI, the amount of children per-row, whether the person smokes, the region of the person, and the charges imposed to the person. The entire dataset contains 348 rows, and no null values could be found in any of them. The database has the following structure:

LABEL	COUNT	TYPE	NULL
AGE	348	INT64	NO
SEX	348	INT64	NO
BMI	348	FLOAT64	NO
CHILDREN	348	FLOAT64	NO
SMOKER	348	INT64	NO
REGION	348	INT64	NO
CHARGES	348	FLOAT64	NO

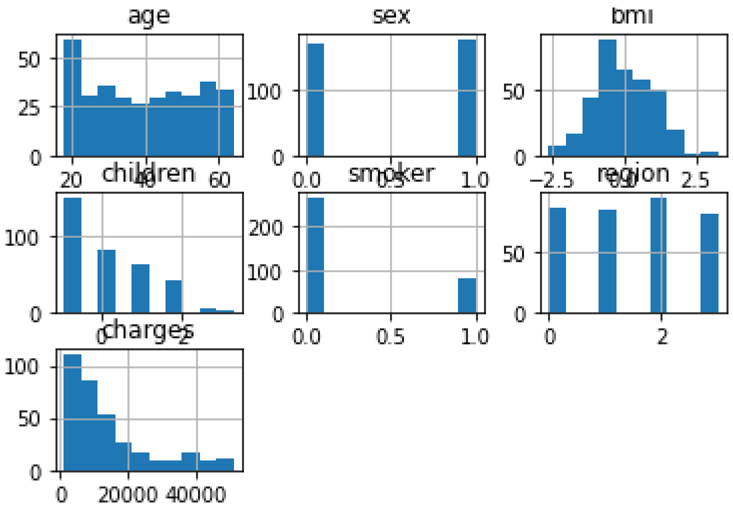
To create reliable data, the fields ‘BMI’ were normalized. SEX and SMOKER were treated as Booleans.

Hypothesis and Objective

The main objective of the exercise was to find the correct model to describe the data, requiring several kinds of testing and different functions to try and match the behavior of the data. The hypothesis is that, given the Age, the Sex, the BMI, the Children, the Smoking Habits, and the region of a certain person, we can predict the charges of a whoever is being charged.

Solution and Exploration

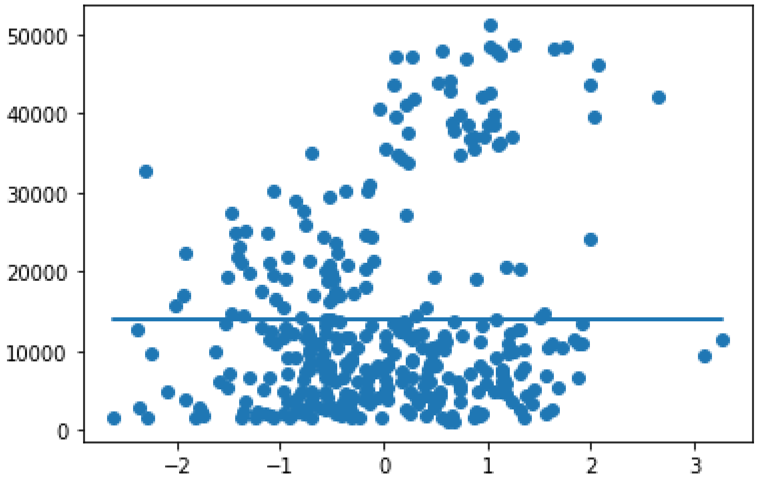
The graphs used can be seen in the notebook. Mostly, we can see that the data is divided in several histograms:

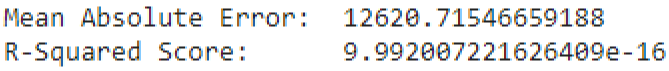


As we can see, we have two bool variables, and the BMI seems to follow a normal distribution. The charges seem to be distributed in a biased-left manner, which follows the tendency of both children and age (not related in between, just same tendency in graph). The regions show that the second region is the most common, while the other three are more uniform. The age is usually around 39 years ld, but the youngest member is 18 years old, while the oldest is 64 old. There are slightly more males than females, and the BMI tends to be nearing 8. Usually, there are no children in the cases, and the people do not smoke. Most people are from the second region.

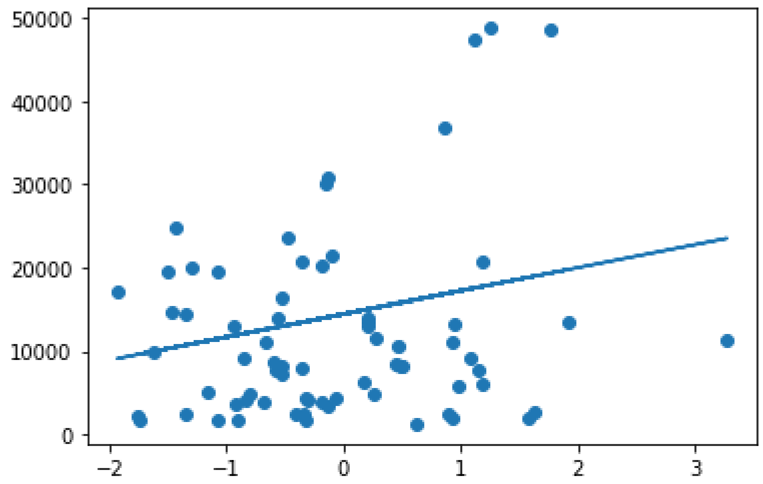
First, a linear regression model was “hard-coded” into the project. For this, we can see that, even though the R^2 is positive, it is also very, very small.

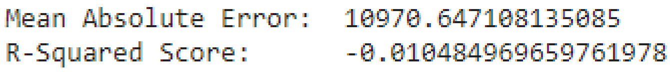
This is because the linear regression grabbed the average of the values and tried to copy them, so it looks almost constant in the graph.





```
Mean Absolute Error:  12620.71546659188
R-Squared Score:      9.992007221626409e-16
```

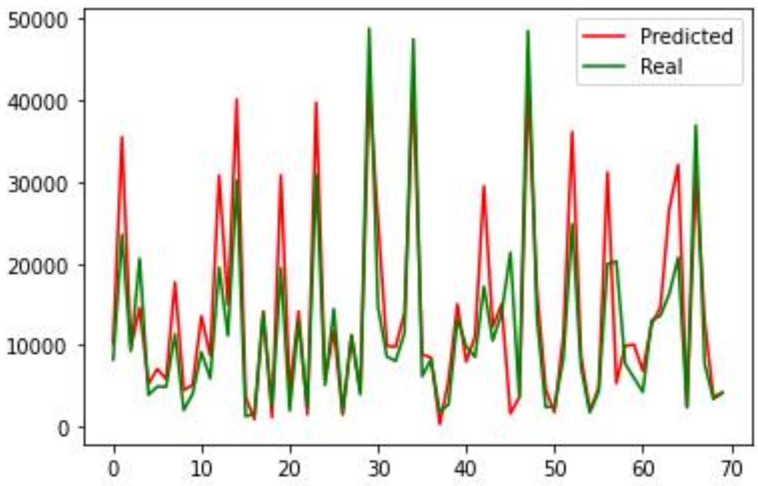


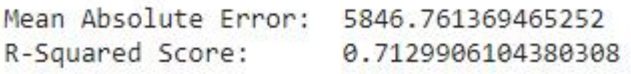


```
Mean Absolute Error:   10970.647108135085
R-Squared Score:       -0.010484969659761978
```

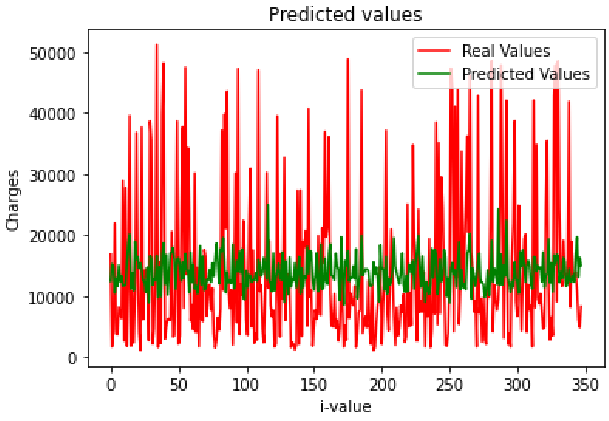
Afterwards, the same kind of graph was made using Sklearn. With this graph, we can see that there is less data, as we used a test section to manage it. Thanks to this, the line slowly climbs up, but the R^2 is negative this time around. This could be because most of the data is under.

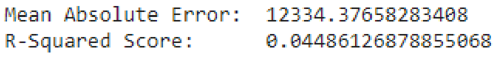
The data was only using one variable to try to predict up to this point. When more features were accepted, the graph to the right was achieved. As we can see, using the *x_test*, and training with the *x_train*, we can predict most values well. Our R^2 seems to be high, at 0.71.





```
Mean Absolute Error:   5846.761369465252
R-Squared Score:       0.7129906104380308
```





```
Mean Absolute Error:   12334.37658283408
R-Squared Score:       0.04486126878855068
```

Finally, a poly-linear distribution was used at the end. Because the original distribution was not quadratic, the values could not be correctly predicted.

Results

As we saw above, the graphs all (but one) seemed distant from reality. The one that was the closest was the one that used as much data as it possibly could, given that the variables were all available to it. It is not overfit, because we trained it with one set, and tested it with another. Because of our ability to choose, we’ve been allowed to see that, given a certain number of ranks, more is not always better.