Insurance cost - regression

1. Our data is a regression type of data.
2. Our aim is to predict how much a specific person would charge the insurance based on different parameters, this is an important topic and has to be accurately measured because on one hand we the insurance company wouldn’t want to lose too much money is a certain person needs more medical care, on the other hand, the company wouldn’t want to charge too much money from their clients as they would try and find a better offer in another insurance company.

The peremeters given are:

* Age : age of the clients ( the data includes people between the ages 18 and 64).
* Sex : “male” or “female”
* BMI : a value calculated based on the height and weight of a person (ranged 15.96 and 53.13).
* Children : how many children a person has and is insured by the companys.
* Smoker : whether a person smokes (either “yes” or “no”)/
* Region : the data is of people living the the USA, and has divided the country into 4 main regions “Northwest” , “Northeast” , “Southwest” and “Southeast”.
* Charges : how much a client charges the insurance.

1. We will prevent data leakage by splitting the data into training and testing sets, we will find a model that will fit the training set and will next evaluate our model on our testing set. The training and testing sets will be randomly chosen before training our model therefore the testing part will be completely unbiased.

= = Unfortunately we cannot ensure that we don’t have any data leakage however we can explain problems with might have with our data that will affect our final model and overall correlation.

A problem we might have with the data is whether people didn’t charge the insurance with all their medical issues, this can affect our model as it can cause anomalies in the model that aren’t caused by the health of the specific client but by the fact that the client didn’t bill the insurance. We will have to check our anomalies very carefully and decide what they were caused by and explain why we decided that.

Another major problem we have with our data is the fact that we don’t know what type of insurance a person has therefore we can’t know if a person didn’t charge the insurance for a specific problem because they’re healthy or because they’re health issue isn’t covered by their insurance.

1. פ
2. We believe that there is a strong correlation between age and the medical costs billed by health insurance, another main thing that we believe increases the charges is BMI, a healthy BMI for adults is considered to be between 18.5 and 24.9 according to the [U.S. Department of Health & Human Services](https://www.hhs.gov/). A high BMI increases the possibility of having health issues such as: type 2 diabetes, stroke, coronary heart disease, etc.. as those complications are related to high BMI we can deduce that those people will have higher charges.

By prior knowledge we also assume that if someone smokes he would charge the insurance more as there are many risks caused by smoking. The data only shows if a person smokes or doesn’t, we don’t know how often this person smokes , therefore we won’t be able to tell how accurate the correlation between smoking and charges is.

1. Graphs can indeed help us understand our data better, we can see if there’s a strong correlation between one of the parameters and the charges or whether all the parameters affect the final charges equally or perhaps they all affect the final charges but not equally.

A visual representation of out data (perhaps using graphs) can help us detect anomalies and we could try to explain them.   
Anomalies we predict there will be are a young person with high charges, we would usually explain that with background diseases or an accident. Another type of anomaly we expect to see is an older person with low charges.