**Linear Regression**

Walkthrough Video:

<https://berkeley.zoom.us/rec/share/k1huVBySB9Lx_9aHe1w1WKWeWbmm1PwTGp39sHaJB2iIxw_yRBWNGW2ynkcqopT2.ojBaiqxmOB6ZQlKs>

This week’s content will shift more weight on your own work than previous weeks, where a big goal is for you to explore on your own how you might approach Linear Regression. The big area focused here was to make a “charges” estimator based on a few variables on individuals, and spit out an ideal amount that might reflect what they would be charged as a result. Here are a few key takeaways from this very short example. First, notice that our training error is lower than our test error. This is normally true, since the training data was used to create the model, so our model is likely to be better on data it has already seen before. We can see, however, that our training and test error are both still high. We need to try some other things to improve our accuracy. This will be a key exercise for this week’s content. Can you create a model that is much better at predicting charges? Consider the following factors to begin exploring:

* Can we make other linear models that can predict different features besides charges? Consider exploring a relationship between a variable such as charges, and how it might be used to predict BMI instead.
* What if we used multiple variables (such as age, BMI, whether the individual is a smoker or not, etc) to predict the insurance charges instead of just one? Will there ever be a point where we end up having TOO many factors (look into the problem behind overfitting vs underfitting models)?
* What if the insurance charges aren’t linearly related to the dependent variables? Can we transform our features differently and improve accuracy? A starting point would be to create visualizations between the charges variable and other variables, and see if you can gain any insights. This leads to learning a bit more about nonlinear regression (power, polynomial, etc.)
* Anything you come up with on your own!

At this point, finishing this part wraps up a major part of this project as a whole. A lot of the content so far has been to build up towards these common techniques used for data analysis. Although this walkthrough doesn’t go too much in depth for the theory behind it, I would recommend finding resources to learn a bit more about why some of these things make sense, or even exploring why libraries such as scikit-learn are so popular today.