**Modeling + Regression**

Zoom Recording:

<https://berkeley.zoom.us/rec/share/xnRHDfPAsqJR2MyQkl3Iq2LGOYIbWqQ_xIrGkUVXu-qtsZFFUaOudW_JZ2ngKiI7.xPpC4IcRWlPs3-bQ>

This week covers the basics of modeling, and the introduction of linear regression. We explored ways that we could go about to determine a good approximation of what a constant value might be. That is, given no other variables at play with no information besides the platelets column, the best value we can give as an approximation is the ones we have found such as using the mean, median, etc.

Before we talk about regression, here are some things for you to explore further.

* Try to find the values that would minimize the error under both loss functions. When should you use the root mean square error as your loss function? When should mean absolute error be better as your loss function?
* Explore using modelling for other pieces of quantitative data.
* Research different loss functions! This can potentially be very difficult as you go deep into certain topics, but also rewarding in learning advanced ways we can determine this and find good values to fit models. (here are some relatively simple alternative loss functions you might want to explore and try to implement: MSE, MSLE)

We moved on to giving an example of a linear regression model. Here are a few key takeaways from this very short example. First, notice that our test error is lower than our training error. This is normally not 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. This usually implies that something might be really wrong with our model. It implies that either our training set had too many odd cases to learn from (where it didn’t end up coming up with a good model as a result), or that our test set had too many simple ones to predict (where our test set is possibly too small to be reflective of an “actual” real world example).

In this case, there's plenty of room to tinker and try things! Here’s something you might want to check just for this concept:

* Does changing the ratio of training and test set mitigate this issue? Remember, in our walkthrough we used 80% of the data to train the model, and 20% to test.
* Try removing the fixed random\_state value, and see how random train test splits fare with the model.
* Could there be something wrong with the relationship between platelets and age? An easy way to check relationships of each variable is calling the .corr() function on the ‘hf’ dataset.

Regardless, 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 platelets compared to just using the mean/median or age? Consider the following factors to begin exploring:

* Can we make other linear models that can predict different features besides platelets? Consider exploring a relationship between other variables, and how it might be used to predict age instead.

(Note: In terminology, both features and independent variables can be used interchangeably.)

* What if we used multiple variables to predict the platelets count instead of just using one, age? 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? Consider exploring other nonlinear regression (power, polynomial, etc.)
* We’ve been focusing a lot on quantitative data. What about something such as “Death Event”, which is only categorical? Could we still make linear models to predict this? Look into logistic regression, and see if you can construct a model!
* Anything you come up with on your own! Remember, we also have a WHOLE other dataset on heart disease if you would like to explore on your own!

As a reminder, all of these are just suggestions and recommendations for areas which you might want to explore further. It might seem like a lot, but having more options gives you more flexibility to make this project your own, and investigating things that you might be interested in playing around with. In the two weeks, you will still have time to go back, so don’t feel overwhelmed, since the key focus for next week will be to learn new models and utilize them for our data.