Problem eight.

Look through the following article posted in folder for homework two, answer questions a, b, and c.

Regression Analysis for COVID-19 Infections and Deaths Based on Food Access and Health Issues

Abrar Almalki [1,*] , Balakrishna Gokaraju [1], Yaa Acquaah [1] and Anish Turlapaty [2]

a. **Identify the plots and tables that look familiar and we have discussed so far.**

On Pg. 8, there is a scatterplot matrix which shows us the relationship between Covid cases and each of the independent variables. "Positive correlations include obesity with poverty and high blood pressure. Negative correlation is presented between obesity and med-income variables. However, there is no apparent strong correlation observed between COVID-19 cases and other variables through this scatter matrix visualization."(pg. 6) They also show the scatterplot matrix using Covid Deaths as the dependent variable which shows also shows no clear correlations between COVID deaths and the independent variables.

The correlation matrix, on pg. 15, helps us to see the correlation between the different predictors, or independent variables. "There is no correlation between COVID-19 cases and health issues (obesity, high cholesterol, and high blood pressure). Moreover, there is no correlation between unhealthy food outlets, healthy food outlets, and health issues. " (Pg. 14) We would want to use this to help us choose predictors that are not highly correlated with one another, to avoid multicollinearity.

b. **What do you conclude from table three?**

In table three we can see the Root Mean Square Error (RMSE). Since we can see that the RMSE is lower for each of the models for COVID-19 Deaths than for COVID-19 cases, we can conclude that the independent variables are stronger predictors of COVID deaths than they are of COVID cases. Specifically, the Support Vector Regression accounted for the least amount of RMSE on average for predicting COVID deaths.

c. **What do you conclude from table four?**

Table four shows us the correlation coefficient for each of the models. Since the correlation is stronger in in each of the models for COVID deaths than for COVID cases, this also supports the idea in table 3 that the independent variables have a stronger association to, and account for more variability in the predictor of COVID deaths than COVID cases. Namely, the Support Vector Regression found the strongest correlation between the independent variables and COVID deaths.

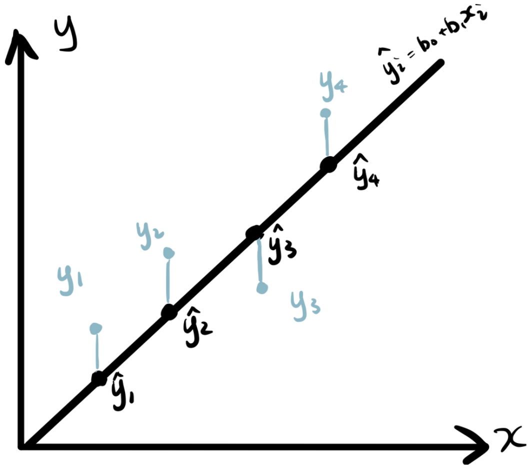
d. What is the formula for RMSE and what is it used for in the statistics literature?

RMSE = $\sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{N-k-1}}$

In the statistics literature the RMSE is used to calculate how much the actual values typically vary from the predictions from a given model. The larger the RMSE, the further the predictions of a given model will be from the true values on average.

e. Find visual for RMSE in linear regression, make a copy of it, and try to explain it to a non-statistician.

$$RSS = \hat{\sum_{i=1}} (y_i - \hat{y}_i)^2$$



The RMSE is a value that tells us how far our predictions are from what actually happened. In the image you can see blue lines between the actual values (y_i) and the predicted values $(\hat{y})$ on the line of best fit (LSRL), which are called residuals. The RMSE is the average of the sum of these blue lines squared. What the RMSE does is it calculates the average length of each of the blue lines, or in other words the standard error of our predictions. Essentially, the shorter the blue lines are the smaller our error in prediction are on averages, indicating that the model is stronger at making accurate predictions.