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.

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?

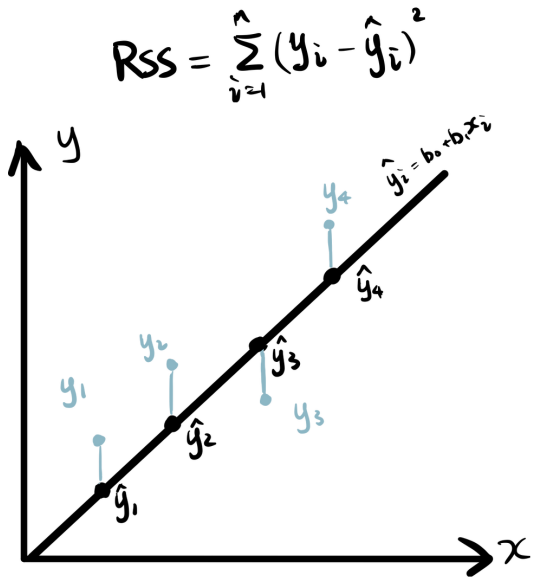
Table four shows us the correlation coefficient (R^2) for each of the models. Since the correlation coefficient is stronger in in each of the models for COVID deaths than for COVID cases, this also supports the idea in table 3 that the independent variables account for more variability in the predictor of COVID deaths than COVID cases. Namely, the Support Vector Regression found the R^2 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 = $\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N}}$

In the statistics literature the RMSE is used to calculate how much the actual values vary from the predictions typically 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.**



In the visual to the left, we can see a scatterplot of data (y_i) that has a line of best fit (LSRL) passing through. The LSRL is fit so that it is minimizes the total distance to each of the data points (y_i). The LSRL is a model that can be used to predict the outcome your y_i, given a specified x_i value. These predicted values of the LSRL are indicated by $\hat{y}_i$. The distance between the actual value and the predicted value is known as the residual, or error, $y_i - \hat{y}_i$, indicated by the blue line in the image between each value and the LSRL. However, the LSRL is calculated such that the sum of the residuals will always be zero, since some residuals are positive (above the LSRL), while others are negative (below the LSRL). So, to look at how far the actual values vary from the predicted values, we must square them, so that all values are positive, and will not cancel each other out, and then we add them up, which gives the RSS that is pictured. The RSS gives us the residual sum of squares, but if we are interested in seeing how much the actual values vary on average from the predicated values (Root Mean Square Error, RMSE), we must find the average, divide by N (total number of values) and then take the square root so that we are measuring in the same number of units as the predictor. Which allows us to compare how much the actual varies from the predictor on average. Essentially, the shorter the blue lines are from the LSRL to the actual $\hat{y}_i$ values, the smaller the RMSE will be, indicating that the model is stronger at making accurate predictions.