SML312 — Research Projects in Data Science: Mini Project #1

04/10/22

Jonathan Hanke

Justin H. Cai

Problem 1

(a). Linear regression attempts to model the relationship between the random variables X and Y as a linear relationship, with the intention to predict unknown values and draw conclusions about the underlying relationship between X and Y. Generally this takes the form of a linear equation such as $Y = b_1X + b_0$, where the objective is to find the values of b_1 and b_0 that minimize the sum of squared errors across all data points (x_i, y_i). This allows us to produce a line that ideally is suited to predict an unknown value of y_i given a new data point x_i. If a linear regression model predicts the relationship between X and Y accurately, we can infer that the underlying relationship between X and Y is linear.

(b). See code. It is reasonable that the two regression lines are different because $Y = X^2$ over a normally distributed X around 0 produces a parabola symmetric around $x = 0$ as seen in the generated graph, which a linear regression model would create an approximately horizontal line. In contrast, $Y = X^2$ over a uniformly distributed X in $[0, 1]$ is strictly increasing in the positive $x-$direction, creating a linear regression line with positive slope, as seen in the graph.

(c). This is a case of omitted variable bias. For X_1 to be positively correlated to Y and $b_1 < 0$, X_1 and X_2 must be negatively correlated. This leads to a positive bias in the prediction of Y using X_2 if X_1 is omitted. This is important because the omission of variables correlated with both the dependent variable and an independent variable changes the coefficients and thus the linear model, and shows that the interpretation of a multivariate linear model can only be as effective as the interpreter's knowledge of the underlying variables influencing the dependent variable. If there are omitted correlated variables, then the multivariate linear model is biased and must incorporate these variables to be accurate. This can happen in a single-variable regression model if there exists another variable correlated to both the independent and dependent variable in some way. Then, the single-variable regression model would exhibit omitted variable bias.

(d). See code.

Problem 2

(a). See code.

(b). See code. The means and variances of both the x and y columns are necessary and sufficient for determining the regression line. This is because calculating the slope of the regression line requires the standard deviations and correlation of x and y, which are derived from the variances, and the intercept is the slope of the line multiplied by the $x-$mean and subtracted from the $y-$mean.

(c). For the first dataset, I chose a least-squares linear regression model because the distribution of the data points do not particularly lend themselves to any noticeable distribution or function, but show a positive correlation, which can be modelled well with a linear model. For the second dataset, I chose a 2nd-order polynomial regression model as the data points are distributed parabolically, which is typically indicative of a quadratic function. The regression curve fits the data very well and is not overly complicated. For the third dataset, I chose a quadratic model as it best accounts for the behavior of the outlier at $x = 13$ while still representing the linear relationship of the rest of the data well, especially for smaller x values. A linear model would be unable to account for both the outlier and the linear relationship between the rest of the points, and a higher degree polynomial would be too complex and unlikely to be representative of the underlying relationship. For the fourth dataset, I chose a linear model, because any higher degree polynomial regression won't provide a better squared error or necessarily represent the data better. As it is difficult to tell the underlying relationship between these data, and a higher degree polynomial won't provide much benefit, it is best to stick to the simplest model.

Problem 3

(a). See code.

(b). I chose to use a least-squares regression model using the BMI feature because it provided the highest correlation coefficient to the disease progression metric. As we have no other information about the underlying relationship between the features and target data, the feature with the highest correlation coefficient is most likely to accurately predict the output data if we are limited to a single variable. I made this decision after calculating the correlation coefficients of each feature with the target data and graphing the regression line of each to visually confirm that such a model is feasible.

(c). I chose to use a least-squares multivariate regression model based on a subset of parameters chosen through recursive feature elimination with automatic tuning of the number of features selected with cross-validation (source here). This method finds the optimal subset of features by training the model initially using all the features, calculating the importance of each feature, recursively consider smaller and smaller subsets of features, and then use cross-validation to optimize the number of features and the specific features to consider. My decision-making process primarily came down to choosing the number of features to consider and which specific features to consider, as the regression model optimizes other factors such as the squared error. RFECV allowed me to optimize both of these variables without arbitrarily assuming one to determine the other (e.g. assuming 5 is the optimal number of features to determine which features to use).

Acknowledgements

Honor Statement

This problem set represents my own work in accordance with University regulations.
— Justin Cai, 04/10/22

Collaborators

None