EML Assignment 3 - Problem 5

Muhammed Saeed and Ali Salaheldin Ali Ahmed

December 2022

Question 1

The parameters for the OLS model fitted on each subset S_k would be denoted as $\hat{\beta}_0^{(k)}$ and $\hat{\beta}^{(k)}$, respectively. These are the estimated intercept and slope parameters for the OLS model fit on the data in S_k.

$$
\begin{aligned}
\bar{y}^{(k)} &= \frac{K}{n} \sum_{(x,y)\in S_k} y \\
\bar{x}^{(k)} &= \frac{K}{n} \sum_{(x,y)\in S_k} x \\
\hat{\beta}_0^{(k)} &= \bar{y}^{(k)} - \hat{\beta}^{(k)} \bar{x}^{(k)} \\
\hat{\beta}^{(k)} &= \frac{\sum_{(x,y)\in S_k} (x - \bar{x}^{(k)})(y - \bar{y}^{(k)})}{\sum_{(x,y)\in S_k} (x - \bar{x}^{(k)})^2}
\end{aligned}
$$

Quesiton 2

To predict y using the above models, we would calculate the average of all the obtained models for each value of x0. This can be written as:

$$
\begin{aligned}
\hat{y}_0 &= \frac{1}{K} \sum_{k=1}^{K} \left(\hat{\beta}_0^{(k)} + \hat{\beta}^{(k)} x_0 \right) \\
&= \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_0^{(k)} + \left(\frac{1}{K} \sum_{k=1}^{K} \hat{\beta}^{(k)} \right) x_0
\end{aligned}
$$

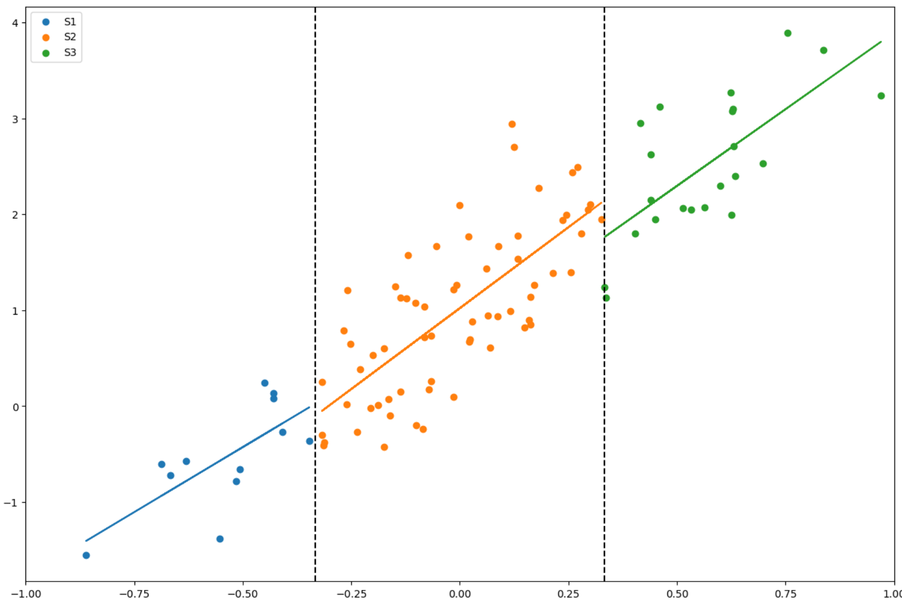
where K is the number of subsets, ${\hat{\beta}_0}^{(k)}$ and $\hat{\beta}^{(k)}$ are the parameters for the kth model, and x_0 is the input value for which we want to predict $\hat{y}_0$.

Question 3

From the result in **Question 2**, the resulting model is still a linear model: its intercept $\hat{\beta}_0$ is the average of the intercepts $\hat{\beta}_0^{(k)}$ of the partial models, and its slope $\hat{\beta}$ is the average of the slopes $\hat{\beta}^{(k)}$ of the partial models.

In comparison to fitting a single model to all the data, fitting multiple independent models to various subsets of the data is probably going to produce a worse fit. This is due to the fact that the OLS estimator has the lowest variance of all unbiased estimators according to the Gauss-Markov Theorem, and fitting multiple independent models to various subsets of the data is likely to provide a bigger variance than fitting a single model to all of the data, while still having the same inherent bias.

Question 4



Question 5

Using the following function:

$$\hat{y_0} = \hat{\beta}_0^{(k)} + \hat{\beta}^{(k)} x_0$$

where k satisfies $-1 + \frac{2(k-1)}{K} \leq x_0 < -1 + \frac{2k}{K}$.

In other words, we find that S_k to which x_0 belongs. Then, we use it's OLS parameters to predict the response.

Question 6

The model performance may be sensitive to the way the data is splitted, which is one major issue with this method of dividing the data into numerous intervals. For example, one of the models will fit on one interval may not adequately reflect the general trend of the data if the data are split among intervals with predominantly extreme - outliers or high variance- values in one of them 'the model will be good for its data local view not global'. Utilizing cross-validation, which divides the data repeatedly into training and test sets and fits and evaluates the model on each split to determine its generalizability, is one technique to address this issue.

Question 7

This approach relies on separately fitting different models for every interval of the single predictor variable, this approach might not generalize well to higher-dimensional X. Higher-dimensional data may be challenging to consistently split in a way that makes sense, resulting in models that do not adequately reflect the general trend of the data. A large number of parameters may also result from fitting many basic models in higher dimensions, which could cause overfitting and poor generalization performance.