1. Solution

1.1 Model1

In this model, we have eight-predictor model, which takes value from dataset with x1 to x8.

Firstly, we divided 79 observations into 10 parts. We take one of it as test set. Then we take other parts as training set. We repeat this step 10 times to make sure each partition has been choosing to be a test set. To make sure each time the training set and test set has the same proportion, we need to make the size of 10 parts as balanced as possible.

We train a liner regression model on x1 to x8 and y from training set, and make prediction on x from test set. Get the prediction of y and estimate this y with the y from the test set with SSE.

1.2 Model2

In this model, we have two-predictor model, which takes value from dataset with x1 to x2.

Firstly, we divided 79 observations into 10 parts. We take one of it as test set. Then we take other parts as training set. We repeat this step 10 times to make sure each partition has been choosing to be a test set. To make sure each time the training set and test set has the same proportion, we need to make the size of 10 parts as balanced as possible.

We train a liner regression model on x1 to x2 and y from training set, and make prediction on x from test set. Get the prediction of y and estimate this y with the y from the test set with SSE.

1.3 Comparison

Replicate the CV procedure twenty times, each time using a different random partition of the 79 observations into 10 parts. To compare model1 and model2, we calculate the SSE for each replication.

	model1	model2
0	30258.761731	25564.574790
1	30622.262690	25499.294032
2	32913.307699	24872.283182
3	30430.251719	24847.655641
4	33384.185866	26367.432769
5	30500.246577	25618.971509
6	30949.884682	25856.402196
7	30631.122279	25566.046807
8	31173.865081	25506.907274
9	30654.278733	25563.191235
10	32718.709783	25433.924607
11	30653.252240	26718.807190
12	28404.176987	25269.021531
13	29242.334425	25208.741421
14	29807.569979	25163.776593
15	31617.192032	27043.212949
16	30329.673634	25427.917332
17	33085.857614	27590.301030
18	28818.121645	24741.048867
19	32269.498992	25842.115242

It can find out that with different random partition, the SSE is also different. The results of SSE are inconsistent from replicate-to-replicate. However, in each partition, the SSE of model1 all bigger than model2.

To make the result more clearly, we calculate the mean of the SSE for model1 and model2.

model1	30923.227719
model2	25685.081310

It also shows that model2 is better than model1.

2. Analysis

2.1 Comparison of model1 and model2

Based on the comparison of the SSE with different models, it can be deducted:

1. By comparing the results of model1 and model2, it is obvious that model2 is better than model1. We suggested that it may because of model1 has too many predictor variables, which makes the linear regression overfitting. So, it not performance very well in test set.
2. According to the experience, if the predictor variables is sufficiently much, the SSE of the predicted linear regression might be large.
3. According to the given dataset and the result of SSE, it seems that we should not use much predictor variables when make a linear regression of set of data.

2.2 Comparison of different partition of dividing observations into parts.

According to take different random partition of the 79 observations into 10 parts, it shows that different random partition will lead to different SSE of both model1 and model2. It shows the influence of choosing training set and validation set. Using Cross-Validation can reduce the error of this influence.

Cross-validation is used to evaluate the predictive performance of the model, especially the performance of the trained model on the new data. When the dataset is very small, using cross-validation can somehow avoid overfitting.