1. We can see an obvious periodic pattern of the data on the graph. Periods of all 5 workflows are the same, which is approximately 7 days, the number of days of a week. The peaks of workflow 4 appear when the troughs of workflow 0 appear. The peaks of workflow 2 and 3 appear at the same time, leading the peak of workflow 4 by approximately 2 days. The peaks of workflow 1 are lagging the peaks of workflow 4 by approximately 1 day.
2. a) ii. The fitting result does not change as shown in the plots

iii. No, the performance of prediction does not improve when we choose three most important variables. The RMSE increases slightly when we choose three most important variables.

iv. The combination which achieves the best performance is [false true true true true], of which the true values correspond to “day of the week”, “backup start time-hour of day”, “work-flow ID” and “file name”. The result makes sense and consistent with the pattern we found in question 1. All of the work-flows have a period of 7 days, which is a week. Within the period, at different time (day of the week and hour of the day) the backup size is different, so that the day of the week and hour of the day are two important variables for prediction, while number of the week is not important since every week the data shows the same pattern. Because the data is simulated traffic data, we can tell that this pattern is consistent with our common sense: the traffic condition is related to what day it is.

Each work-flow has its own periodic pattern which is different from that of other work-flows, so work-flow ID is another important variable.

In addition, different types of file need different time to backup, so the file name is also important.

b) iii