**Applied Business Data Analytics (MKTG 881)**

**Assignment 4. Forecasting with Google Trends**

**HK Visit from the U.S.**

**Jordan Martin**

**Due Date: 12/6 (Sunday) by 11:59 PM**

1. Merge Hong Kong visitor data with Google keyword search data and show the screen shot of the new dataset.

Here are the first 26 observations of the merged data:

Calendar

Description automatically generated

1. Make a graph using Actual visit vs. Lagged Google keyword search (one month lag) and interpret the patterns.



It appears that there is some loose correlation between visits and Google keyword searches one month beforehand. In many months the peaks or valleys are aligned, but there are also plenty of months where a peak in visits happens at the same time as a valley in keyword searches a month earlier. Overall, I’d say that there is some correlation, but there is definitely not a perfect correlation by any means.

1. Run the following regression models and report regression results.

(Model 1) VISIT = β0 + β1 L1VISIT + β2 L12VISIT

(Model 2) VISIT = β0 + β1 L1VISIT + β2 L12VISIT + β3 L1SEARCH

(L1: Lagged by one month, L12: Lagged by 12 months)

Model 1:



Model 2:



1. Interpret the coefficients in each model (3 points).

In the first model, we can see that the P-value of the number of visits 12 months ago is close to zero, so very statistically significant while the P-value of number of visits in the previous month is well over 0.05, so it is not at all statistically significant. The coefficient of the previous month is 0.0371, so an increase by one visit in the previous month, increases the number of visits in the current month by 0.0371 on average. The coefficient of the number of visits 12 months ago is 0.802, so an increase by one visit 12 months ago increases the number of visits in the current month by 0.802 on average.

In the second model, we can see that the P-value of the number visits 12 months ago and the Google keyword searches last month are both close to zero, so very statistically significant while the P-value of the number of visits in the previous month is again well over 0.05, so it is not at all statistically significant again. The coefficient of the previous month’s visits is -.0.961, so an increase in one visit just barely decreases the number of visits in the current month on average. The coefficient of the number of visits 12 months ago is 0.834, so an increase by one visit 12 months ago increases the current number of visits by 0.834 on average. Finally, the coefficient of the number of keyword searches last month is 66.576, so an increase by one search a month ago increases the current month’s number of visits by 66.576 on average.

1. Compare (adjusted) R-squared of both models in Question 3. Do you think that the new variable (L1SEARCH) improves R-squared?

R-squared for the first model is 0.6840 and R-squared for the second model is 0.7408, so the introduction of Google keyword search trends from the month previous increased R-squared by about 0.06. R-squared is a relative measure of the goodness of fit that varies from 0 to 1 and it corresponds to the amount of variation in the data accounted for by the model, so because the second model has a higher R-squared, it fits the data better and therefore improved R-squared.

1. Compare Root MSE (RMSE) of the two models. Which model fits the data better? How much does the better model improve RMSE (% improvement)?

RMSE for the first model is was 7453.5 and RMSE for the second model is 6815.2, so the introduction of Google keyword search trends from the month previous decreased RMSE by about 8.6%. RMSE is an absolute measure of fit of models and varies from zero to infinity where zero would be a perfect fit. Thus, the second model improves the fit by 8.6% according to the RMSE measure.

1. Predict VISIT from the base model (Model 1) and the second model (Model 2). Make a graph of actual vs. predicted visits (3 lines in one graph) and interpret the patterns.



The blue solid line is the actual number of visits, the red dashed line is the first model and the green dot/dashed line is the second model. In the graph, we can see that for most months, the models underestimated both the peaks and valleys, but the second model is almost always closer to the actual number of visits. This is to be expected as my answers to questions 5 and 6 showed that the second model had a better fit to the actual data. There are a few months that the models overestimate changes, but overall the second model is clearly closer to the actual number of visits in all but one or two data points.