# Can Google Trend Predict Future Sales?

*This is an individual project.*
*You can use any statistical software of your choice to do this assignment.*
*This exercise accounts for 10% of your grade, and **is due on 9/30.***
*For this assignment, you need to submit the data records you collected, the code, and your written response electronically to Canvas.*

## Obtaining the Data:

- **Automobile Sales Data:**

US Census Bureau publishes "Advance Monthly Sales for Retail and Food Services" report, updated each month. You can access it at http://www.census.gov/retail/marts/www/timeseries.html. We will use the "Auto, other Motor Vehicle" series. Please download the auto sales data from October 2003 to July 2015.

- **Google Trend Data:**
  1. Sign in to your Google accounts and go to Google Trends at https://www.google.com/trends/
  2. Type in any search term.
  3. In the Explore Tool Bar, select Region: United States; time: 2004-now (as default)
  4. Click on the following 4 subcategories:

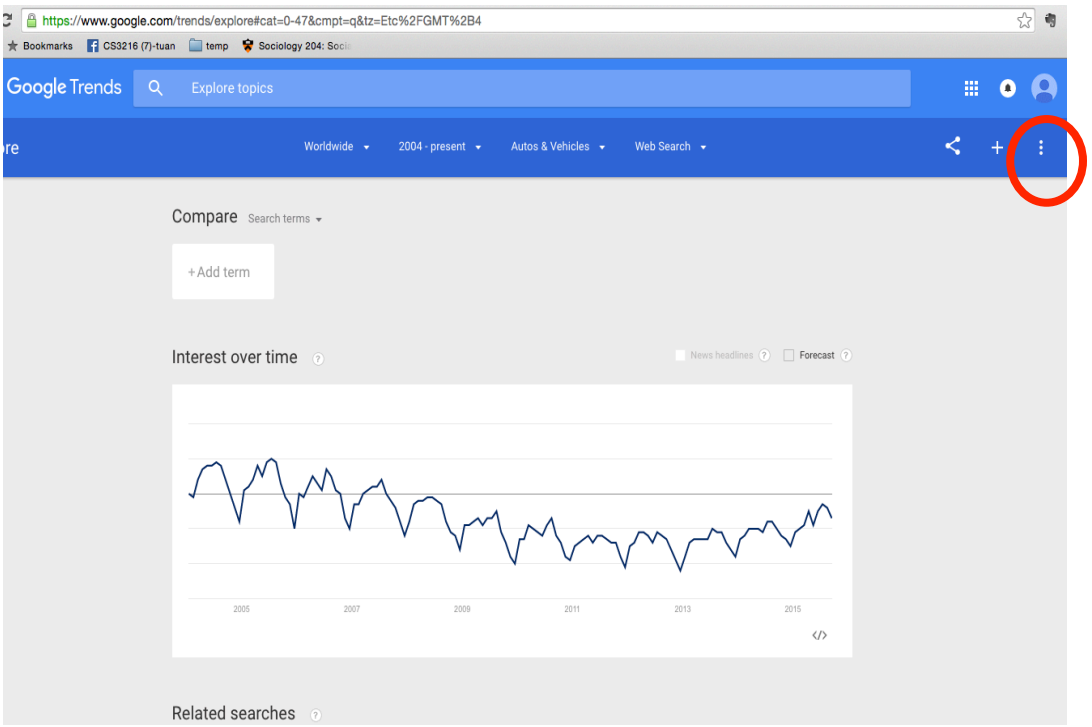
     Autos & Vehicles -> Vehicle Maintenance

     Autos & Vehicles -> Vehicle Shopping

     Autos & Vehicles -> Trucks & SUVs -> SUVs

     Autos & Vehicles -> Trucks & SUVs -> Vans & Minivans

  5. After clicking on each category, you want to erase the search terms in the search box, so that you observe the trend for the entire selected category. Then click on the ":" button (circled in red) and click on "Download as CSV" to get the interest over time data. You should see a screenshot similar to the figure below. The reason we chose categories is that Google already pre-compiled certain indices, such as automobiles. This can save us time of compiling a list of search queries ourselves when Google has already done that for us. Of course, Google does not have categories for everything. When you are looking for new categories, you may need to construct your own list of search queries to compile an index yourself. Sometimes, you may want to compile your own index even when Google has already created one if you want to have a finer control of what the index is measuring. Currently, Google does not provide a list of search queries used to build the index. This type of black box approach works for some predictions, but not for others. For this exercise, we will just use Google's pre-compiled index for automobiles.




6. You can also download search data from other categories for further exploration. Please download and use the data from the week of "2004-01-04 - 2004-01-10" until the week of "2015-07-26 - 2015-08-01." You may notice that Google Trend data for categorical search is a relative search frequency measure based on the date of the first date you specify, in this case all search queries will be measured relative to the search volume in 2004-01-04.

**Preparing the data:**

1. The .csv files downloaded from Google trend is weekly data, but the automobile sales data is monthly. Therefore, we need to match the weekly search data into monthly auto sales data by averaging the weekly search index to create monthly search index. This will reduce the granularities of the weekly to monthly. We can then merge it with the monthly automobile sales data. Now, you should have a data set containing the monthly search data and monthly auto sales data.

   You can represent the monthly data from weekly data in many other ways, such as taking the data first week, or the last week of each month. You can also take the medium value in each month instead of the mean. Sometimes the variance in the weekly data could also be important indicator. For now, let's use the average.

2. Google Trends data for categories are displayed as percentages. Some statistical software may not recognize 5% so you may want to convert them into 0.05.

3. Take the natural log for automobile sales data: call it $Log(AutoSales)$.

4. Create lagged variables. $Log(AutoSales)_{t-1},\ Log(AutoSales)_{t-2}$. Here is a simple illustration. Notice lagged value is simply shifting the original value down.

| Time (T) | $Log(AutoSales)$ | $Log(AutoSales)_{t-1}$ | $Log(AutoSales)_{t-2}$ |
|---|---|---|---|
| 1 | 0.3 | | |
| 2 | 0.5 | 0.3 | |
| 3 | 0.7 | 0.5 | 0.3 |
| 4 | 0.9 | 0.7 | 0.5 |
| 5 | 1.1 | 0.9 | 0.7 |
| 6 | 1.3 | 1.1 | 0.9 |
| 7 | 1.5 | 1.3 | 1.1 |
| 8 | 1.7 | 1.5 | 1.3 |
| 9 | 1.9 | 1.7 | 1.5 |

5. Create a dummy variable indicating the summer months and another dummy variable indicating the winter months. For the summer indicator variable, it should be 1 if the month is June, July or August, and zero for all the other months. For the winter indicator variable, it should be 1 if the month is January, February or March. Seasonality variables can be important for predictions because people are more likely to buy cars in warmer months than colder months.

6. Now you are ready to use the data. You should submit this dataset as a part of the assignment.

   An important caveat about Google Trends is that the data is sampled dynamically from Google's servers. Thus, the data you downloaded could be different each time. Thus, if you download the data set tomorrow, it will likely to be different. In general, when the data can change significantly every time you download (that is when there is a high variance), it is in general a better practice to take a few snapshots of the data (downloaded at different times, e.g., once for each day of the week) and then take the average. This could help you even out the variance and get the true underlying data. However, it is not always easy to know how many snapshots to get. This largely depends on how big the variance is.

   For this assignment, you do not need to create many snapshots of the same underlying data. To ensure no one got an "outlier" search data, please use the dataset we provided for the following exercise. It is called "Assignment1Data.csv". But you still need to submit the data that you have prepared and you can use your own data set to answer the last question of this assignment.

   Now let's do some predictions!

## Questions

*(We encourage you to answer the questions in a concise and to-the-point manner; you don't have to write long paragraphs to get full marks)*

1. Let's try a simple in-sample prediction
   a) Run a very simple OLS regression:

$$Baseline\ Model\ 1:\ Log(AutoSales)_t = \beta_0 + \beta_1 Log(AutoSales)_{t-2} + \beta_2 isSummer\ + \beta_3 isWinter + \varepsilon$$

Basically, you are simply regressing the current sales to the previous sales in the earlier period. The reason we are using two-period lags is that economic indicators are often reported with lags. In another word, to predict the October car sales now, we do not yet know sales statistics in September. At best, we know the sales in August. However, we do know the search terms in real time. This will provide advantage for us in predicting the future.

**Please report the point estimate for $\boldsymbol{\beta_1}$.**

b) After you get the coefficients, calculated the predicted values for auto sales in July 2015:

$$Model\ 2: \widehat{Log(AutoSales)}_t = \widehat{\beta_0} + \widehat{\beta_1} Log(AutoSales)_{t-2} + \widehat{\beta_2} isSummer\ + \widehat{\beta_3} isWinter$$

And then for the actual value, you need to transform the logged value by taking the exponent below:

$$\widehat{(AutoSales)}_t = \exp\left(\widehat{Log(AutoSales)}_t\right)$$

Please report the prediction you made for July 2015 auto sales.

c) We can measure how accurate the prediction is, using the mean absolute deviation (MAE).

$$MAE = \frac{1}{n}\sum_t \frac{|\widehat{AutoSales}_t - AutoSales_t|}{AutoSales_t}$$

where $n$ is the number of observations used.

What is the MAE for the predictions?

2. Now we want to add the Google Trend variable:

$$Model\ 3:\ Log(AutoSales)_t = \beta_0 + \beta_1 Log(AutoSales)_{t-2} + \beta_2\ GoogleTrend_{t-1} + \beta_2 isSummer\ + \beta_3 isWinter + \varepsilon$$

Put the 4 categories of Google Trend data (vehicle maintenance, vehicle shopping, SUVs, minivans) into the model as $GoogleTrend$ variable one at a time and calculate the respective MAEs. Compare the results:

Report the MAE for each category of Google Trend Data. Which categories of Google Trend data perform better? Which categories perform worse? How do they compare to the MAE in the baseline model (Model 2)?

3. What we just did are an in-sample prediction. In-sample predictions generally have a very good MAE but it is not very helpful because it does not make a lot of sense to predict July 2015 sales when the sales are already realized. A much more useful prediction is out-sample prediction. Out-of-sample predictions mean that we use historical data to predict the future. In another word, you want to train your model using historical data so you can make better predictions for the future.

   Let's redo Model 2, this time try something different. Instead of using all the data at once, we use a chosen window of historical data at a time. You want to use the most recent data but if the window length is too short, you may not have enough data to estimate model. But if you use a really long window, historical data may be too old to be meaningful for predictions. In the example, if we choose a window size of 8, we are essentially using the previous 8 months of data to make predictions for the 9th month as illustrated below.

   $$Model\ 4\text{: } Log(AutoSales)_t = \beta_0 + \beta_1 Log(AutoSales)_{t-2} + \beta_2 isSummer + \beta_3 isWinter + \varepsilon$$

| Time (T) | $Log(AutoSales)$ | $Log(AutoSales)_{t-1}$ | $Log(AutoSales)_{t-2}$ | $\widehat{(AutoSales)_t}$ |
|---|---|---|---|---|
| 1 | 0.3 | | | |
| 2 | 0.5 | 0.3 | | |
| 3 | 0.7 | 0.5 | 0.3 | |
| 4 | 0.9 | 0.7 | 0.5 | |
| 5 | 1.1 | 0.9 | 0.7 | |
| 6 | 1.3 | 1.1 | 0.9 | |
| 7 | 1.5 | 1.3 | 1.1 | |
| 8 | 1.9 | 1.5 | 1.3 | Unknown X1 |
| 9 | | | | Unknown X2 |

   Using window size of 8, what is the mean absolute deviation (MAE)?

4. Now we want to add the Google Trend variables using the 8-month moving window method:

   $$Model\ 4\text{: } Log(AutoSales)_t = \beta_0 + \beta_1 Log(AutoSales)_{t-2} + \beta_2\ GoogleTrend_{t-1} + \beta_3 isSummer + \beta_4 isWinter + \varepsilon$$

   Put the 4 categories of Google Trend data (vehicle maintenance, vehicle shopping, SUVs, minivans) into the model as $GoogleTrend$ variable one at a time and calculate the respective MAEs. Compare the results:

   Report the MAE for each category of Google Trend Data. Which categories of Google Trend data perform better? Which categories perform worse? How do they compare to the MAE in the baseline

model (Model 4)? Hint: Out of sample predictions are generally hard so it is not surprising that your model has worse MAE than the baseline model.

5. Now you have seen the basics of using Google Trend data. Now you can experiment with the window size as well as the Google Trend data to find a MAE that outperform the baseline model. Please report the model you used as well as the window size you used to train the model. Show the MAE of your model and the percentage improvement over the baseline model (Model 3).

6. Do you think the optimal window you just found from would still perform well if we are trying to predict housing sales instead? Why or Why not?

7. We have actually tried using Google Trend of housing related categories to predict house sales, and it turned out that the Google Trends did not add much to the predictive power. We have tried several different window sizes but they no longer out-predict the baseline model by much. Do you have any intuition why this is the case?

8. Take the best model with the optimal window length you got so far and plot the *error in prediction* by year. *Error in prediction* is the difference between the predicted auto sales and the actual auto sales. Do you observe any interesting patterns? Is there any years where the predictions are better? Any other years when the predictions are worse? Suppose we are in a period of turbulent economics environment, do you think Google Trend data could help or not? Please justify you answer.

9. Take the *Error in prediction* from the best model you got so far and plot it by *months of the year*. Is there any time of the year when the predictions are better, or worse? Do you have any explanations for this?

10. You tried some other variables on Google Trend. Suppose that you found adding health and beauty category in your model could help you improve your prediction. Should you use it?

11. Think about the formula you have got now for predicting auto sales, do you think it would work to predict auto sales 5 years from now? Why or why not? Can you think of a prediction that the formula would be stable for extended period of time? Can you think of examples where it would not be stable?

12. Now you have seen the basics of predicting car sales. Please experiment with the model and find a model that you believe should best predict car sales for October 2015. You can experiment with

adding more variables or different types of variables. You can even create an index of you own by creating a list of search terms that you think is better suited for predicting car sales. You can also experiment more with the window size to see that makes any difference. Feel free to use your own search data that you created in part 1 of this assignment. You can pull more recent data from Google Trends or create more snapshots of the data. This exercise is meant to be open to explorations so you can use all the existing data and tools to make the prediction.

Please submit your prediction and the model used. When the data for auto sales in Oct 2015 is published, we will find out how well your model worked. Let the competition begin!