Predicting the Present with Google Trends

Hyunyoung Choi

Hal Varian

Draft Date April 10, 2009

Contents

1 Methodology **1**
1.1 Data 1
1.2 Model 2

2 Examples **6**
2.1 Retail Sales 6
2.2 Automotive Sales 9
2.3 Home Sales 12
2.4 Travel 15

3 Conclusion **18**

4 Appendix **19**
4.1 R Code: Automotive sales example used in Section 1 19

Motivation

Can Google queries help predict economic activity?

Economists, investors, and journalists avidly follow monthly government data releases on economic conditions. However, these reports are only available with a lag: the data for a given month is generally released about halfway through through the next month, and are typically revised several months later.

Google Trends provides *daily and weekly* reports on the volume of queries related to various industries. We hypothesize that this query data may be correlated with the *current* level of economic activity in given industries and thus may be helpful in predicting the subsequent data releases.

We are not claiming that Google Trends data help predict the future. Rather we are claiming that Google Trends may help in *predicting the present.* For example, the volume of queries on a particular brand of automobile during the second week in June may be helpful in predicting the June sales report for that brand, when it is released in July.[i.]

Our goals in this report are to familiarize readers with Google Trends data, illustrate some simple forecasting methods that use this data, and encourage readers to undertake their own analyses. Certainly it is possible to build more sophisticated forecasting models than those we describe here. However, we believe that the models we describe can serve as baselines to help analysts get started with their own modelling efforts and that can subsequently be refined for specific applications.

The target audiences for this primer are readers with some background in econometrics or statistics. Our examples use `R`, a freely available open-source statistics package[ii.] ; we provide the `R` source code for the worked-out example in Section 1.2 in the Appendix.

[i.] It may also be true that June queries help to predict July sales, but we leave that question for future research.
[ii.] http://CRAN.R-project.org

Chapter 1

Methodology

Here we provide an overview of the data and statistical methods we use, along with a worked out example.

1.1 Data

Google Trends provides an index of the volume of Google queries by geographic location and category.

Google Trends data does not report the raw level of queries for a given search term. Rather, it reports a *query index.* The query index starts with the *query share*: the total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time. The query share numbers are then normalized so that they start at 0 in January 1, 2004. Numbers at later dates indicated the percentage deviation from the query share on January 1, 2004.

This query index data is available at country and state level for the United States and several other countries. There are two front ends for Google Trends data, but the most useful for our purposes is `http://www.google.com/insights/search` which allows the user to download the query index data as a CSV file.

Figure 1.1 depicts an example from Google Trends for the query [coupon]. Note that the search share for [coupon] increases during the holiday shopping season and the summer vacation season. There has been a small increase in the query index for [coupon] over time and a significant increase in 2008, which is likely due to the economic downturn.

Google classifies search queries into 27 categories at the top level and 241 categories at the second level using an automated classification engine. Queries are assigned to particular categories using natural language processing methods. For example, the query [car tire] would be assigned to category `Vehicle Tires` which is a subcategory of `Auto Parts` which is a subcategory of `Automotive.`

Figure 1.1: Google Trends by Keywords - Coupon

1.2 Model

In this section, we will discuss the relevant statistical background and walk through a simple example. Our statistical model is implemented in `R` and the code may be found in the Appendix. The example is based on Ford's monthly sales from January 2004 to August 2008 as reported by *Automotive News.* Google Trends data for the category `Automotive/Vehicle Brands/Ford` is used for the query index data.

Denote Ford sales in the t-th month as $\{y_t : \ t = 1, 2, \cdots, T\}$ and the Google Trends index in the k-th week of the t-th month as $\{x_t^{(k)} : \ t = 1, 2, \cdots, T; k = 1, \cdots, 4\}$. The first step in our analysis is to plot the data in order to look for seasonality and structural trends. Figure 1.2 shows a declining trend and strong seasonality in both Ford Sales and the Ford Query index.

We start with a simple baseline forecasting model: sales this month are predicted using sales last month and 12 months ago.

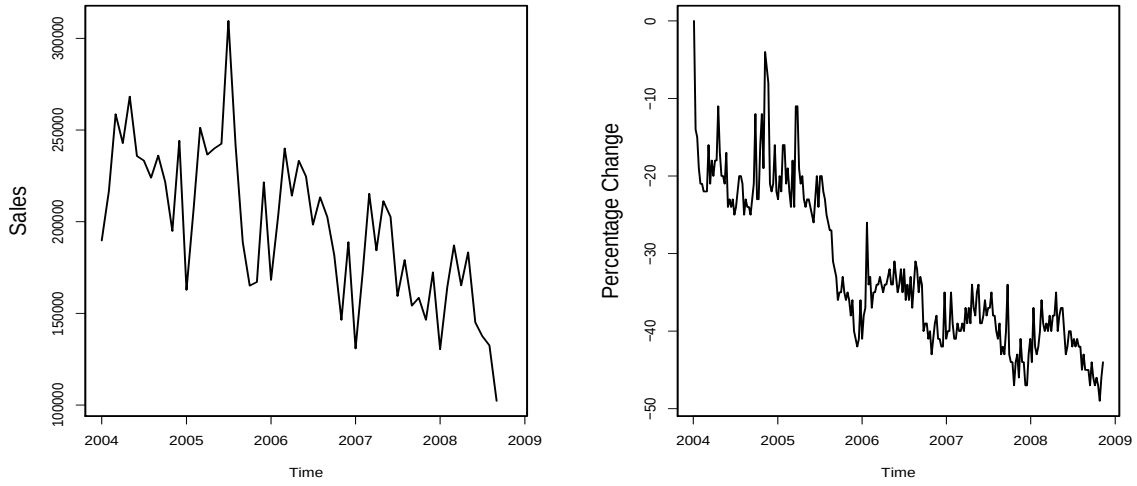
$$\text{Model 0:} \ \log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + e_t, \tag{1.1}$$

The variable e_t is an error term. This type of model is known in the literature as a *seasonal autoregressive* model or a *seasonal AR model.*

We next add the query index for 'Ford' during the first week of each month to this model. Denoting this variable by $x_t^{(1)}$, we have

$$\text{Model 1:} \ \log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + x_t^{(1)} + e_t \tag{1.2}$$

The least squares estimates for this model are shown in equation (1.3). The positive coefficient on the Google Trends variable indicates that the search volume index is positively associated with Ford Sales sales.



(a) Ford Monthly Sales

(b) Ford Google Trends

Figure 1.2: Ford Monthly Sales and Ford Query Index

Figure 1.3 depict the four standard regression diagnostic plots from R. Note that observation 18 (July 2005) is an outlier in each plot. We investigated this date and discovered that there was special promotion event during July 2005 called an 'employee pricing promotion.' We added a dummy variable to control for this observation and re-estimated the model. The results are shown in (1.4).

$$\log(y_t) = 2.312 + 0.114 \cdot \log(y_{t-1}) + 0.709 \cdot \log(y_{t-12}) + 0.006 \cdot x_t^{(1)} \quad (1.3)$$

$$\log(y_t) = 2.007 + 0.105 \cdot \log(y_{t-1}) + 0.737 \cdot \log(y_{t-12}) + 0.005 \cdot x_t^{(1)} + 0.324 \cdot \mathrm{I}(\text{July 2005}). \quad (1.4)$$

Both models give us consistent results and the coefficients in common are similar. The 32.4% increase in sales at July 2005 seems to be due to the employee pricing promotion. The coefficient on the Google Trends variable in (1.4) implies that 1% increase in search volume is associated with roughly a 0.5% increase in sales.

Does the Google Trends data help with prediction? To answer this question we make a series of one-month ahead predictions and compute the prediction error defined in Equation 1.5. The average of the absolute values of the prediction errors is known as the *mean absolute error (MAE)*. Each forecast uses only the information available up to the time the forecast is made, which is one week into the month in question.

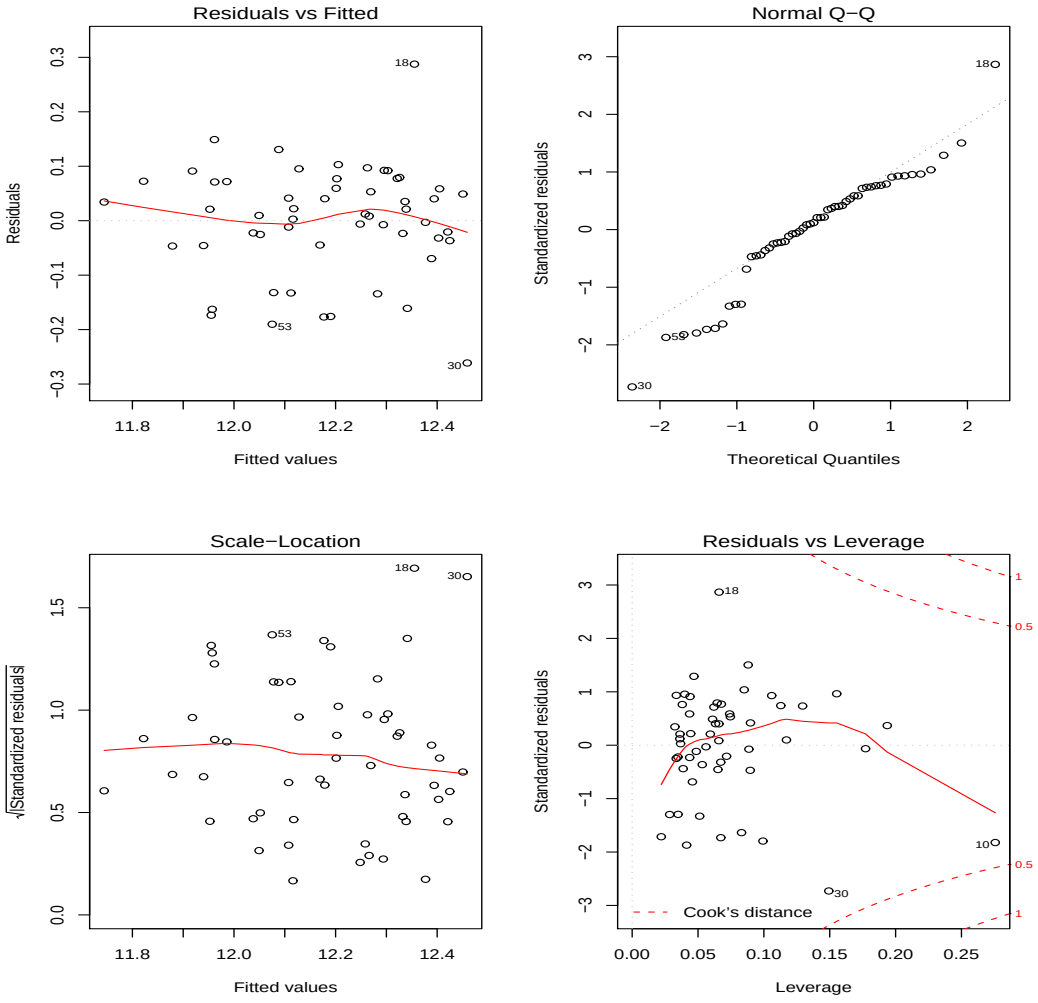


Figure 1.3: Diagnostic Plots for the regression model

$$\begin{aligned} \text{PE}_t &= \log(\hat{y}_t) - \log(y_t) \approx \frac{y_t - \hat{y}_t}{y_t} \\ \text{MAE} &= \frac{1}{T}\sum_{t=1}^{T} |\text{PE}_t| \end{aligned} \tag{1.5}$$

Note that the model that includes the Google Trends query index has smaller absolute errors in most months, and its mean absolute error over the entire forecast period is about 3 percent smaller. (Figure 1.4). Since July 2008, both models tend to overpredict sales and Model 0 tends to overpredict by more. It appears that the query index helps capture the fact that consumer interest in automotive purchase has declined during this period.

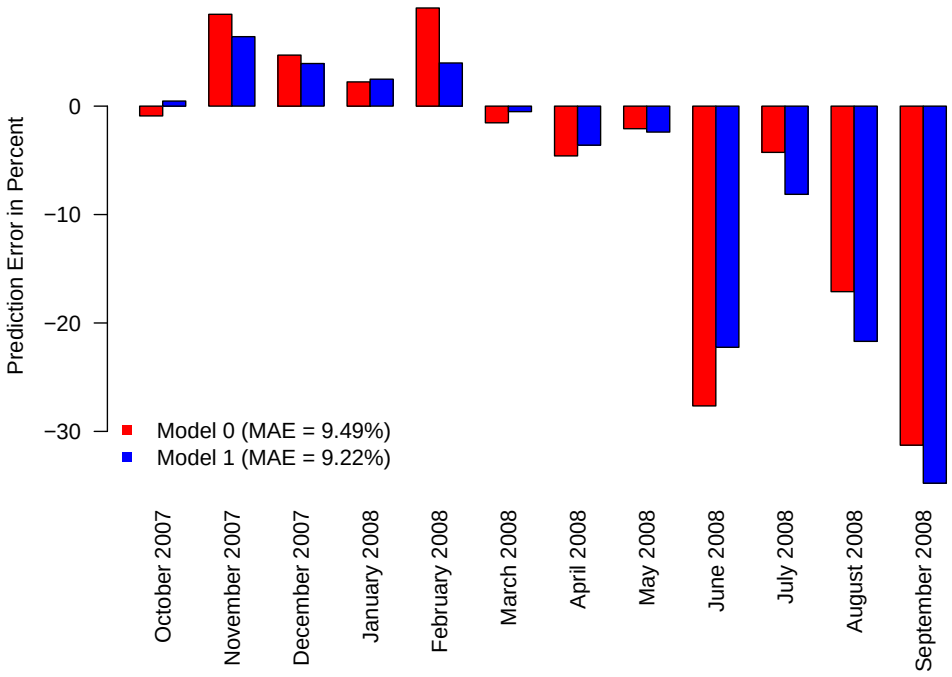


Figure 1.4: Prediction Error Plot

Chapter 2

Examples

2.1 Retail Sales

The US Census Bureau releases the Advance Monthly Retail Sales survey 1-2 weeks after the close of each month. These figures are based on a mail survey from a number of retail establishments and are thought to be useful leading indicators of macroeconomic performance. The data are subsequently revised at least two times; see 'About the survey[i.] ' for the description of the procedures followed in constructing these numbers.

The retail sales data is organized according to the NAICS retail trade categories.[ii.] The data is reported in both seasonally adjusted and unadjusted form; for the analysis in this section, we use only the unadjusted data.

NAICS Sectors		Google Categories	
ID	Title	ID	Title
441	Motor vehicle and parts dealers	47	Automotive
442	Furniture and home furnishings stores	11	Home & Garden
443	Electronics and appliance stores	5	Computers & Electronics
444	Building mat., garden equip. & supplies dealers	12-48	Construction & Maintenance
445	Food and beverage stores	71	Food & Drink
446	Health and personal care stores	45	Health
447	Gasoline stations	12-233	Energy & Utilities
448	Clothing and clothing access. stores	18-68	Apparel
451	Sporting goods, hobby, book, and music stores	20-263	Sporting Goods
452	General merchandise stores	18-73	Mass Merchants & Department Stores
453	Miscellaneous store retailers	18	Shopping
454	Nonstore retailers	18-531	Shopping Portals & Search Engines
722	Food services and drinking places	71	Food & Drink

Table 2.1: Sectors in Retail Sales Survey

[i.] http://www.census.gov/marts/www/marts.html
[ii.] http://www.census.gov/epcd/naics02/def/NDEF44.HTM

As we indicated in the introduction, Google Trends provides a weekly time series of the volume of Google queries by category. It is straightforward to match these categories to NAICS categories. Table 2.1 presents top level NAICS categories and the associated subcategories in Google Trends. We used these subcategories to predict the retail sales release one month ahead.

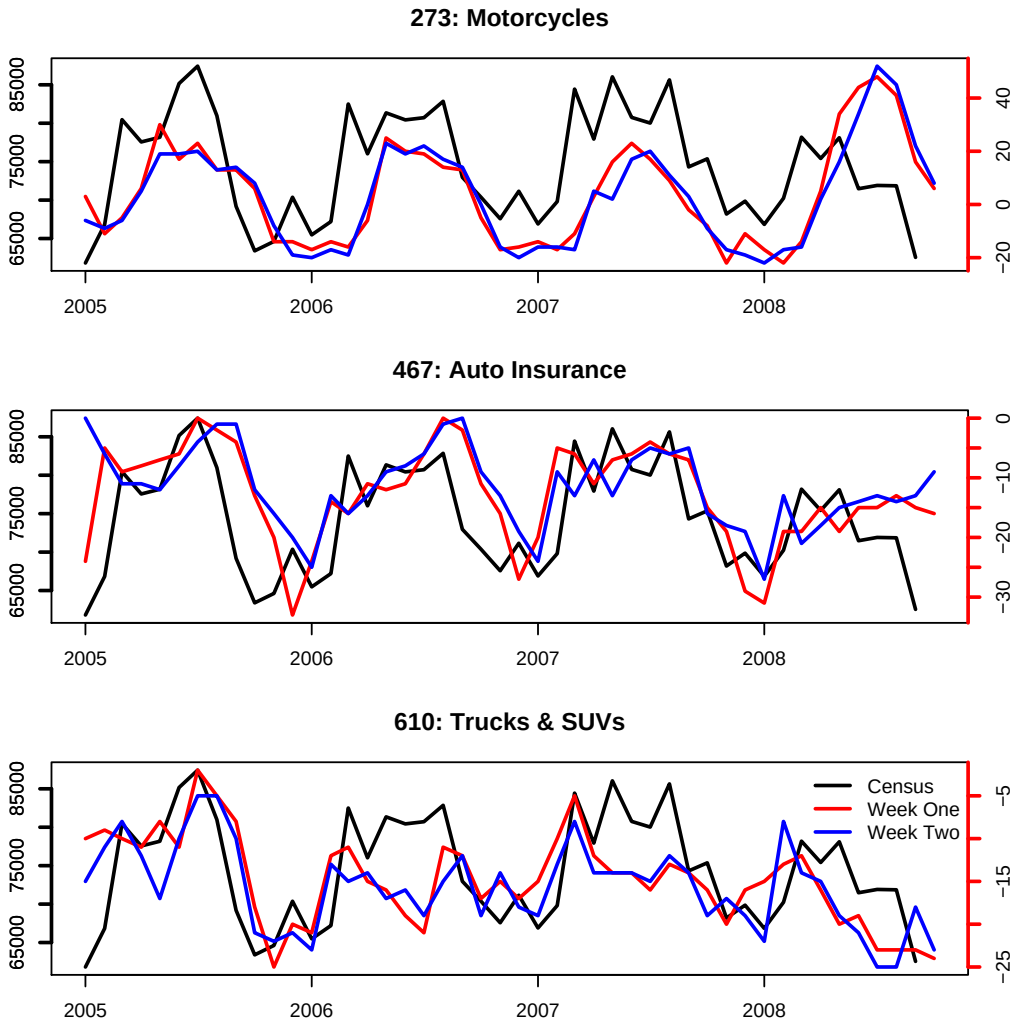


Figure 2.1: Sales on 'Motor vehicle and parts dealers' from Census and corresponding Google Trends data.

Under the Automotive category, there are fourteen subcategories of the query index. From these fourteen categories, the four most relevant subcateories are plotted against Census data for sales on 'Motor Vehicles and Parts' (Figure 2.1)

We fit models from Section 1.2 to the data using 1 and 2 weeks of Google Trend data. The notation

$x^{(1)}_{610,t}$ refers to Google category 610 in week 1 of month t. Our estimated models are

$$\begin{aligned}
&\text{Model} \quad 0: \quad \log(y_t) = 1.158 + 0.269 \cdot \log(y_{t-1}) + 0.628 \cdot \log(y_{t-12}),\ e_t \sim N(0, 0.05^2) \qquad (2.1)\\
&\text{Model} \quad 1: \quad \log(y_t) = 1.513 + 0.216 \cdot \log(y_{t-1}) + 0.656 \cdot \log(y_{t-12}) + 0.007 \cdot x^{(1)}_{610,t},\ e_t \sim N(0, 0.06^2)\\
&\text{Model} \quad 2: \quad \log(y_t) = 0.332 + 0.230 \cdot \log(y_{t-1}) + 0.748 \cdot \log(y_{t-12})\\
&\qquad\qquad -0.001 \cdot x^{(2)}_{273,t} + 0.002 \cdot x^{(1)}_{467,t} + 0.004 \cdot x^{(1)}_{610,t},\ e_t \sim N(0, 0.05^2).
\end{aligned}$$

Note that the R^2 moves from 0.6206 (Model 0) to 0.7852 (Model 1) to 0.7696 (Model 2). The models show that the query index for 'Trucks & SUVs' exhibits positive association with reported sales on 'Motor Vehicles and Parts'.

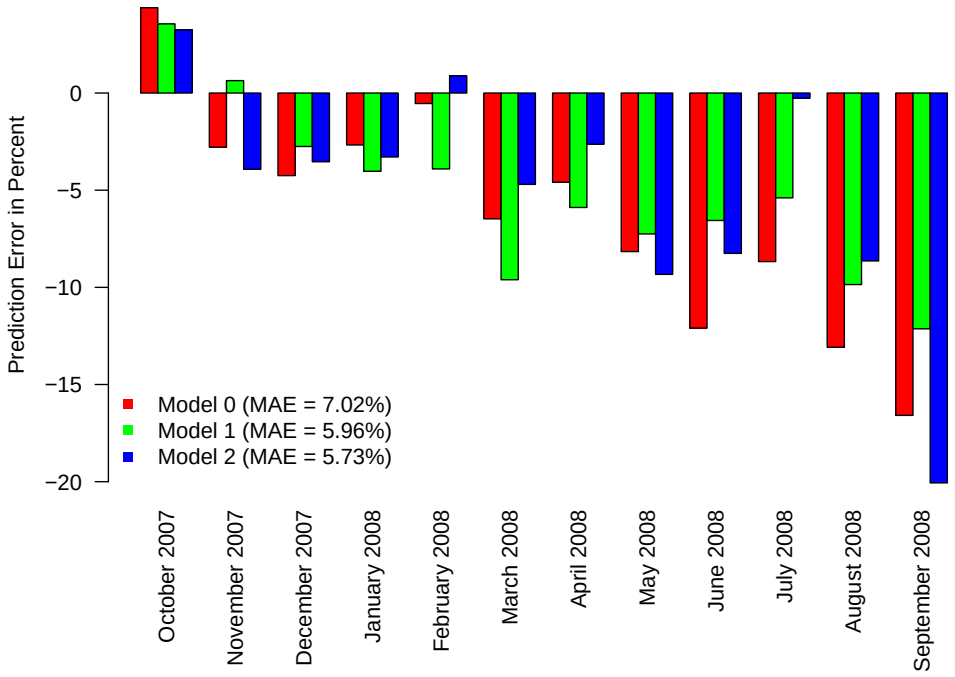


Figure 2.2: Prediction Error Plot

Figure 2.2 illustrates that the mean absolute error of Model 1 is about 15% better than model 0, while Model 2 is about 18% better in terms of this measure. However all forecasts have been overly optimistic since early 2008.

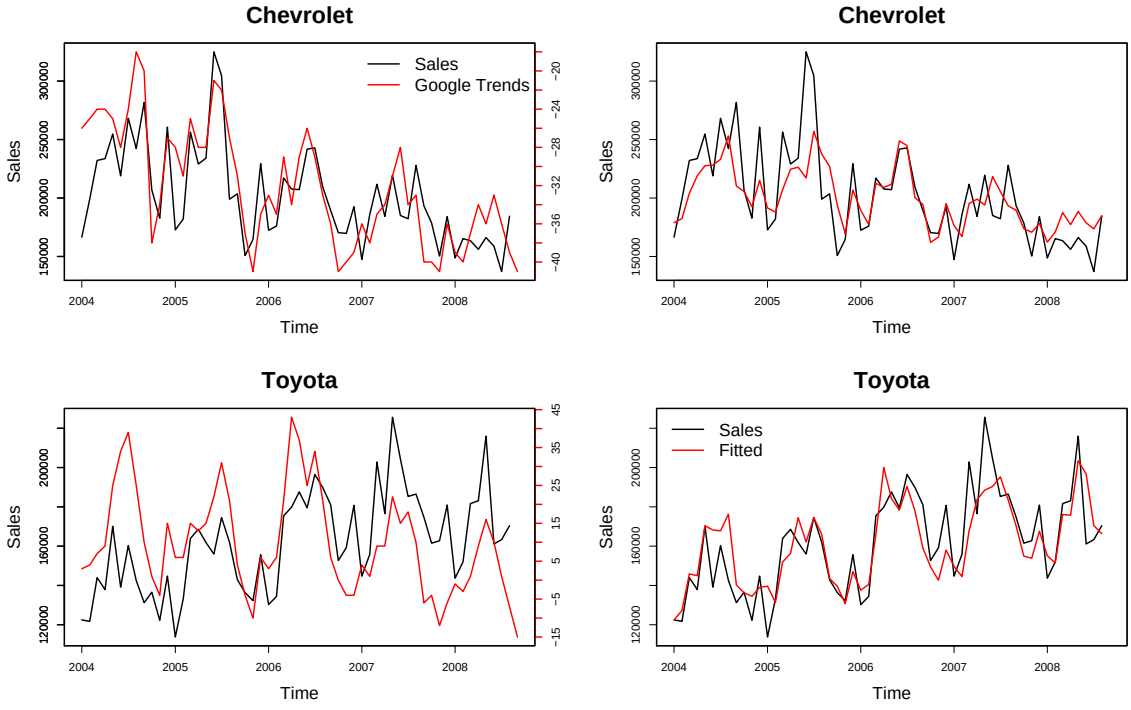
2.2 Automotive Sales

In Section 2.1, we used Google Trends to predict retail sales in 'Motor vehicle and parts dealers'. While automotive sales are an important indicator of economic activity, manufacturers are likely more interested in sales by make.

In the Google Trends category `Automotive/Vehicle Brands` there are 31 subcategories which measure the relative search volume on various car makes. These can be easily matched to the 27 categories reported in the 'US car and light-truck sales by make' tables distributed by *Automotive Monthly.*

We first estimated separate forecasting models for each of these 27 makes using essentially the same method described in Section 1.1. As we saw in that section, it is helpful to have data on sales promotions when pursuing this approach.

Since we do not have such data, we tried an alternative fixed-effects modeling approach. That is, we we assume that the short-term and seasonal lags are the same across all makes and that the differences in sales volume by make can be captured by an additive fixed effect.



(a) Sales vs. Google Trends

(b) Actual & Fitted Sales

Figure 2.3: Sales and Google Trends for Top 2 Makes, Chevrolet & Toyota

Denote the automotive sales from the i-th make and t-th month as $\{y_{i,t} : t = 1, 2, \cdots, T; i = 1, \cdots, N\}$ and the corresponding i-th Google Trends index as $\{x_{i,t}^{(k)} : t = 1, 2, \cdots, T; i = 1, \cdots, N; k = 1, 2, 3\}$. Considering the relatively longer research time associated car purchase, we used Google Trends from the second to last week of the previous month($x_{i,t-1}^{(3)}$) to the first week of the month in question ($x_{i,t}^{(1)}$) as predictors.

The estimates from Equation (2.2) indicate that the the Google Trends index for a particular make in the last two week of last month is positively associated with current sales of that make.

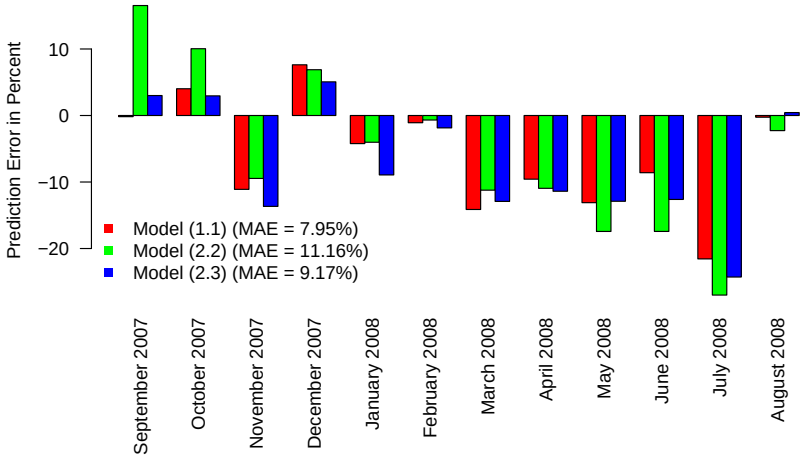
$$\begin{aligned} \log(y_{i,t}) &= 2.838 + 0.258 \cdot \log(y_{i,t-1}) + 0.448 \cdot \log(y_{i,t-12}) + \delta_i \cdot \text{I(Car Make)}_i \\ &\quad + 0.002 \cdot x_{i,t}^{(1)} + 0.003 \cdot x_{i,t}^{(2)} - 0.001 \cdot x_{i,t}^{(3)},\ e_{i,t} \sim N(0, 0.13^2). \end{aligned} \tag{2.2}$$

We can compare the fixed effects model to the separately estimated univariate models for each brand. In each of the separately estimated models case we find a positive association with the relevant Google Trends index. Here are two examples.

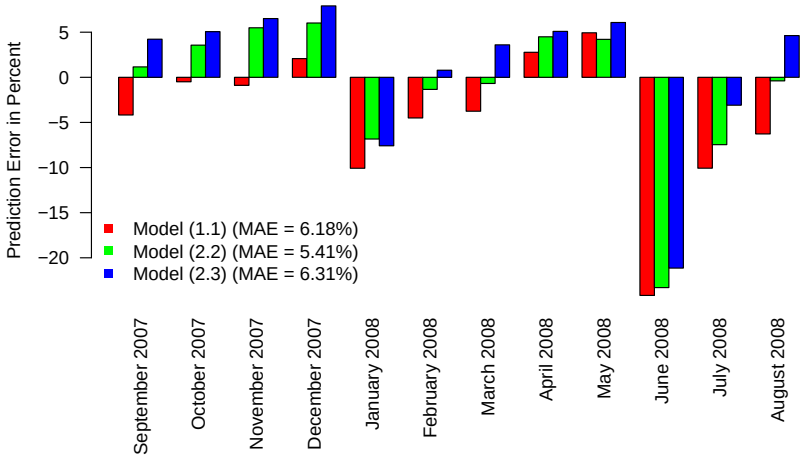
$$\begin{aligned} \text{Chevrolet} &: \log(y_{i,t}) = 7.367 + 0.439 \cdot \log(y_{i,t-12}) + 0.017 \cdot x_{i,t}^{(2)},\ e_t \sim N(0, 0.114^2) \\ \text{Toyota} &: \log(y_{i,t}) = 4.124 + 0.655 \cdot \log(y_{i,t-12}) + 0.003 \cdot x_{i,t}^{(2)},\ e_t \sim N(0, 0.093^2) \end{aligned} \tag{2.3}$$

Model (1.1) is fitted to each brand and compared to Model (2.2) and Model (2.3). As before, we made rolling one-step ahead predictions from 2007-10-01 to 2008-09-01. We found that Model 1.1 performed best for Chevrolet while Model 2.3 performed best for Toyota, as shown in Figure 2.4.

One issue with the fixed effects model is that imposes the same seasonal effects for each make. This may or may not be accurate. See, for example, the fit for Lexus in Figure 2.4. In December, Lexus has traditionally run an ad campaign suggesting that a new Lexus would be a welcome Christmas present. Hence we observe a strong seasonal spike in December Lexus sales which is not present with other makes. In this case, it makes sense to estimate a separate model for Lexus. Indeed, if we do this we estimate a separate model for Lexus, we get an improved fit with the mean absolute error falling from X to Y.



(a) Chevrolet



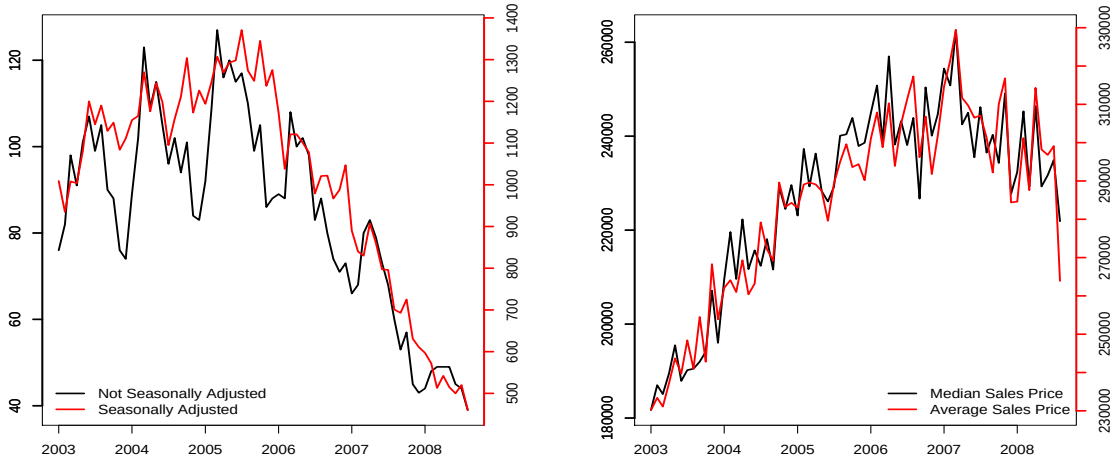
(b) Toyota

Figure 2.4: Prediction Error Plot by Make - Chevrolet & Toyota

2.3 Home Sales

The US Census Bureau and the US Department of Housing and Urban Development release statistics on the housing market at the end of each month.[iii] The data includes figures on 'New House Sold and For Sale' by price and stage of construction. New House Sales peaked in 2005 and have been declining since then (Figure 2.5(a)). The price index peaked in early 2007 and has declined steadily for several months. Recently the price index fell sharply (Figure 2.5(b)).

The Google Trends 'Real Estate' category has 6 subcategories (Figure 2.6) - Real Estate Agencies (Google Category Id: 96), Rental Listings & Referrals(378), Property Management(425), Home Inspections & Appraisal(463), Home Insurance(465), Home Financing(466). It turns out that the search index for Real Estate Agencies is the best predictor for contemporaneous house sales.



(a) Number of New House Sold

(b) Prices of New House Sold

Figure 2.5: Number and Price of New House Sold

We fit our model to seasonally adjusted sales figures, so we drop the 12-month lag used in our earlier model, leaving us with Equation (2.4).

$$\text{Model 0: } \log(y_t) \sim \log(y_{t-1}) + e_t, \tag{2.4}$$

where e_t is an error term. The model is fitted to the data and Equation (2.5) shows the estimates of the

[iii] http://www.census.gov/const/www/newressalesindex.html

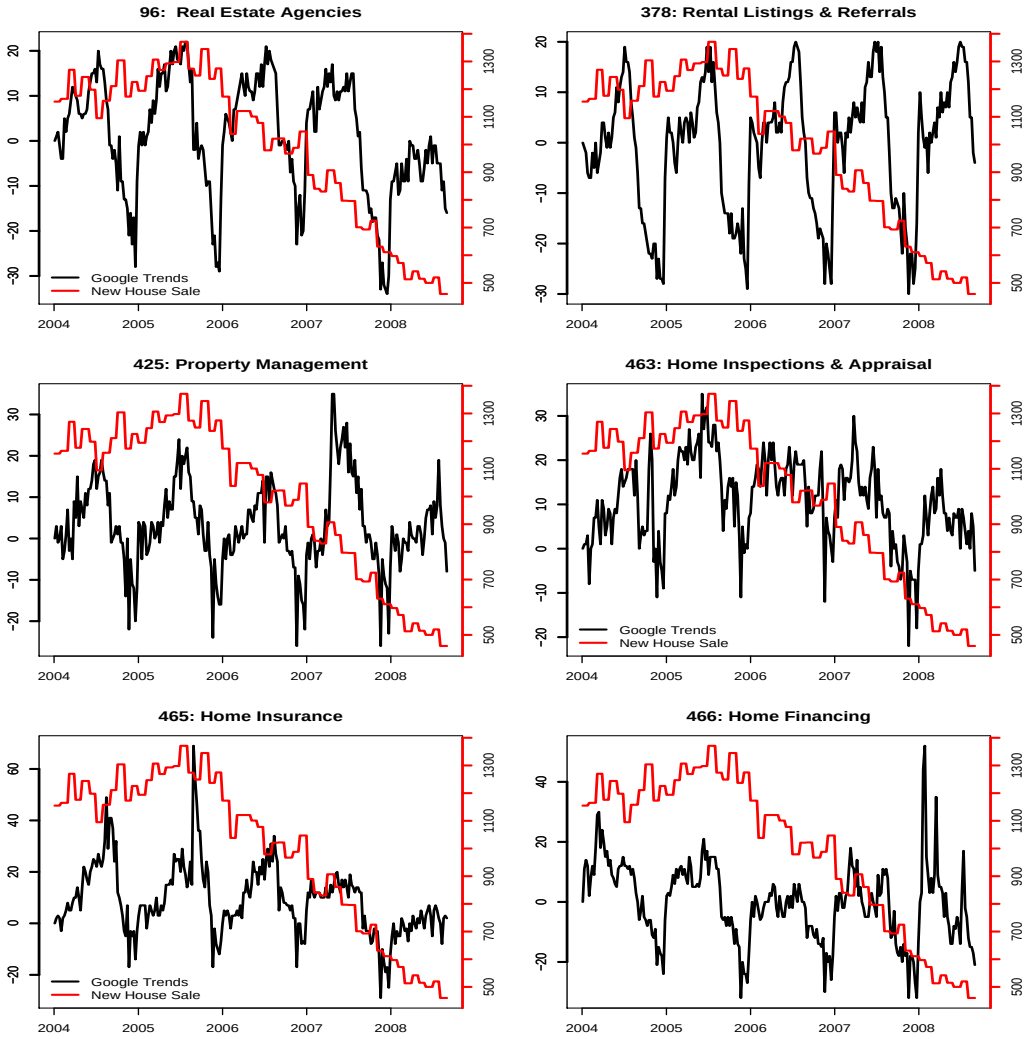
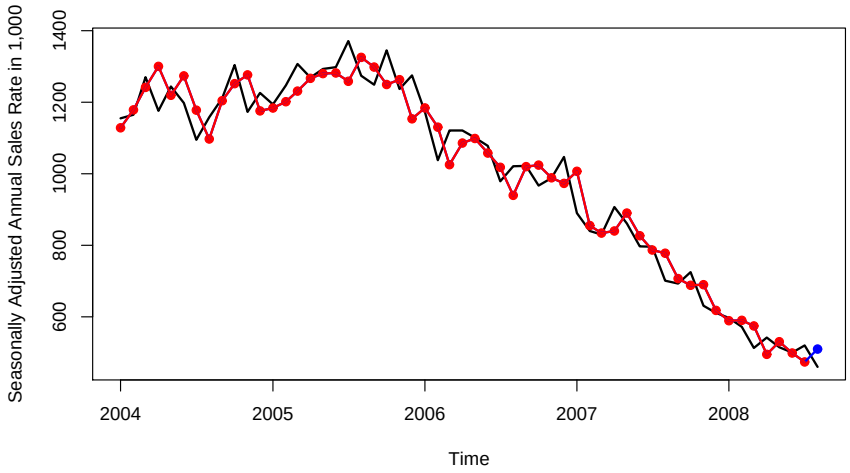


Figure 2.6: Time Series Plots: New House Sold vs. Subcategories of Google Trends

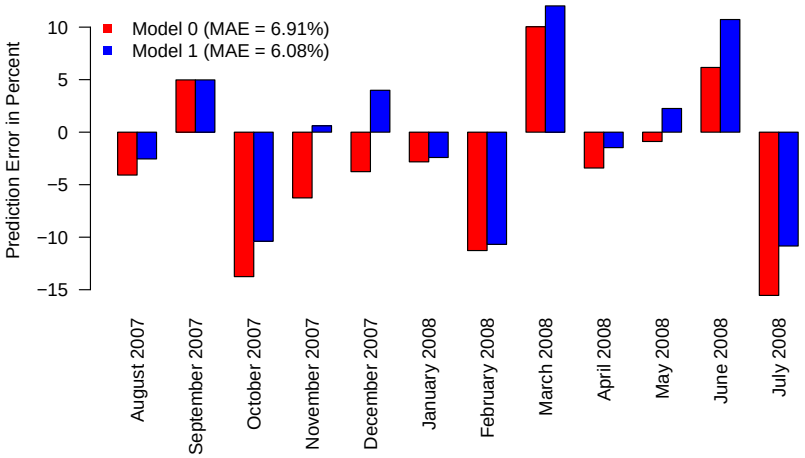
model. The model implies that (1) house sales at $(t-1)$ are positively related to house sales at t, (2) the search index on 'Rental Listings & Referrals'(378) is negatively related to sales, (3) the search index for 'Real Estate Agencies'(96) is positively related to sales, (4) the average housing price is negatively associated with sales.

$$\text{Model 1: } \log(y_t) = 5.795 + 0.871 \cdot \log(y_{t-1}) - 0.005 \cdot x^{(1)}_{378,t} + 0.005 x^{(2)}_{96,t} - 0.391 \cdot \text{Avg Price}_t \quad (2.5)$$

The one-step ahead prediction errors are shown in Figure 2.7(a). The mean absolute error is about 12% less for the model that includes the Google Trends variables.



(a) Seasonally Adjusted Annual Sales Rate vs. Fitted



(b) 1 Step ahead Prediction Error

Figure 2.7: New One Family House Sales - Fit and Prediction

2.4 Travel

The internet is commonly used for travel planning which suggests that Google Trends data about destinations may be useful in predicting visits to that destination. We illustrate this using data from the Hong Kong Tourism Board.[iv.]

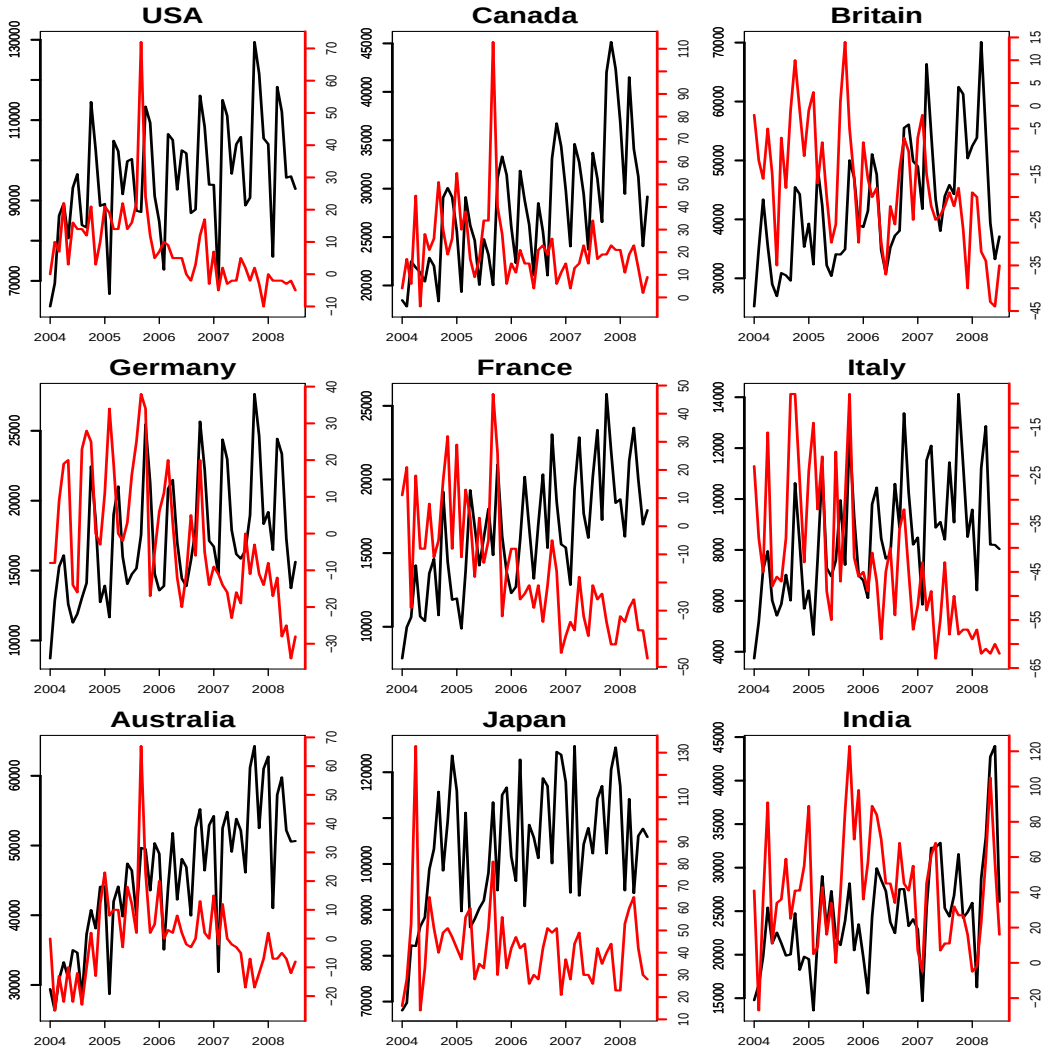


Figure 2.8: Visitors Statistics and Google Trends by Country

Note: The black line depictes visitor arrival statistics and red line depicts the Google Trends index by country.

The Hong Kong Tourism Board publishes monthly visitor arrival statistics, including 'Monthly visitor arrival summary' by country/territory of residence, the mode of transportation, the mode of entry and other criteria. We use visitor arrival statistics by country from January 2004 to August 2008 for this

[iv.] http://partnernet.hktourismboard.com

analysis.

The foreign exchange rate defined as HKD/Domestic currency is used as another predictor for visitor volume. The Beijing Olympics were held from 2008-08-08 to 2008-08-24 and the traffic at July 2008 and August 2008 is expected to be lower than usual so we use dummy variable to adjust the traffic difference during those periods.

'Hong Kong' is one of the subcategories in under Vacation Destinations in Google Trends. The countries of origin in our analysis are USA, Canada, Great Britain, Germany, France, Italy, Australia, Japan and India. The visitors from these 9 countries are around 19% of total visitors to Hong Kong during the period we examine. The visitor arrival statistic from all countries shows seasonality and an increasing trend over time, but the trend growth rates differ by country(Figure 2.8).

Here we examine a fixed effects model. In Equation 2.6. 'Country_i' is a dummy variable to indicate each country and the interaction with $\log(y_{i,t-12})$ captures the different year-to-year growth rate. 'Beijing' is another dummy variable to indicate Beijing Olympics period.

$$\begin{aligned}\log(y_{i,t}) &= 2.412 + 0.059 \cdot \log(y_{i,t-1}) + \beta_{i,12} \cdot \log(y_{i,t-12}) \times \text{Country}_i \\ &+ \delta_i \cdot \text{Beijing} \times \text{Country}_i + 0.001 \cdot x_{i,t}^{(2)} + 0.001 \cdot x_{i,t}^{(3)} + e_{i,t},\ e_{i,t} \sim N(0, 0.09^2)\end{aligned}$$

From Equation (2.6)), we learn that (1) arrivals last month are positively related to arrivals this month, (2) arrivals 12 months ago are positively related to arrivals this month, (3) Google searches on 'Hong Kong' are positively related to arrivals, (4) during the Beijing Olympics, travel to Hong Kong decreased.

Table 2.2 is an Analysis of Variance table from Model 2.6. It shows that most of the variance is explained by lag variable of arrivals and that the contribution from Google Trends variable is statistically significant. Figure 2.9 shows the actual arrival statistics and fitted values. The model fits remarkably well with Adjusted R^2 equal to 0.9875.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
log(y1)	1	234.07	234.07	29,220.86	< 2.2e-16	***
Country	8	5.82	0.73	90.74	< 2.2e-16	***
log(y12)	1	9.02	9.02	1,126.49	< 2.2e-16	***
$x_{i,t}^{(2)}$	1	0.44	0.44	54.34	1.13E-12	***
$x_{i,t}^{(3)}$	1	0.03	0.03	3.87	0.049813	*
Beijing	1	0.41	0.41	51.23	4.53E-12	***
Country:log(y12)	8	0.23	0.03	3.59	0.000504	***
Country:Beijing	8	0.14	0.02	2.12	0.033388	*
Residuals	366	2.93	0.01			

Table 2.2: Estimates from Model (2.6)
Note: Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.10

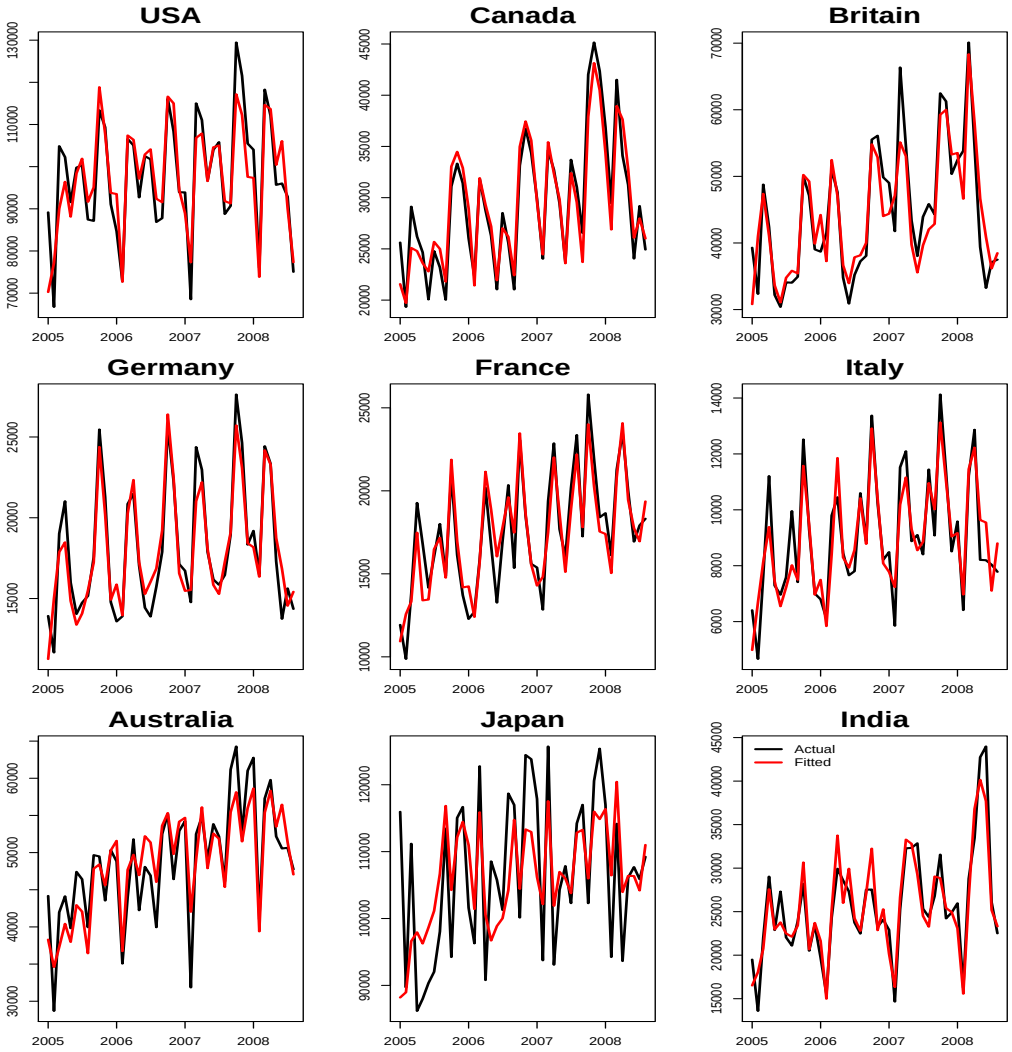


Figure 2.9: Visitors Statistics and Fitted by Country

Note: The black line depicts the actual visitor arrival statistics and red line depicts fitted visitor arrival statistics by country.

Chapter 3

Conclusion

We have found that simple seasonal AR models and fixed-effects models that includes relevant Google Trends variables tend to outperform models that exclude these predictors. In some cases, the gain is only a few percent, but in others can be quite substantial, as with the 18% improvement in the predictions for 'Motor Vehicles and Parts' and the 12% improvement for 'New Housing Starts'.

One thing that we would like to investigate in future work is whether the Google Trends variables are helpful in predicting "turning points" in the data. Simple autoregressive models due remarkably well in extrapolating smooth trends; however, by their very nature, it is difficult for such models to describe cases where the direction changes. Perhaps Google Trends data can help in such cases.

Google Trends data is available at a state level for several countries. We have also had success with forecasting various business metrics using state-level data.

Currently Google Trends data is computed by a sampling method and varies somewhat from day to day. This sampling error adds some additional noise to the data. As the product evolves, we expect to see new features and more accurate estimation of the Trends query share indices.

Chapter 4

Appendix

4.1 R Code: Automotive sales example used in Section 1

```
##### Import Google Trends Data
google = read.csv('googletrends.csv');
google$date = as.Date(google$date);

##### Sales Data
dat = read.csv("FordSales.csv");
dat$month = as.Date(dat$month);
##### get ready for the forecasting;
dat = rbind(dat, dat[nrow(dat), ]);
dat[nrow(dat), 'month'] = as.Date('2008-09-01');
dat[nrow(dat), -1] = rep(NA, ncol(dat)-1);

##### Define Predictors - Time Lags;
dat$s1 = c(NA, dat$sales[1:(nrow(dat)-1)]);
dat$s12 = c(rep(NA, 12), dat$sales[1:(nrow(dat)-12)]);

##### Plot Sales & Google Trends data;
par(mfrow=c(2,1));
plot(sales ~ month, data= dat, lwd=2, type='l', main='Ford Sales',
  ylab='Sales', xlab='Time');
plot(trends ~ date, data= google, lwd=2, type='l', main='Google Trends: Ford',
  ylab='Percentage Change', xlab='Time');

##### Merge Sales Data w/ Google Trends Data
google$month = as.Date(paste(substr(google$date, 1, 7), '01', sep='-'))
dat = merge(dat, google);

##### Define Predictor - Google Trends
##    t.lag defines the time lag between the research and purchase.
##      t.lag = 0 if you want to include last week of the previous month and
##              1st-2nd week of the corresponding month
##      t.lag = 1 if you want to include 1st-3rd week of the corresponding month
t.lag = 1;
id = which(dat$month[-1] != dat$month[-nrow(dat)]);
mdat = dat[id + 1, c('month', 'sales', 's1', 's12')];
```

```
mdat$trends1 = dat$trends[id + t.lag];
mdat$trends2 = dat$trends[id + t.lag + 1];
mdat$trends3 = dat$trends[id + t.lag + 2];

##### Divide data by two parts - model fitting & prediction
dat1 = mdat[1:(nrow(mdat)-1), ]
dat2 = mdat[nrow(mdat), ]

##### Exploratory Data Analysis
## Testing Autocorrelation & Seasonality
acf(log(dat1$sales));
Box.test(log(dat1$sales), type="Ljung-Box")
## Testing Correlation
plot(y = log(dat1$sales), x = dat1$trends1, main='', pch=19,
  ylab='log(Sales)', xlab= 'Google Trends - 1st week')
abline(lm(log(dat1$sales) ~ dat1$trends1), lwd=2, col=2)
cor.test(y = log(dat1$sales), x = dat1$trends1)
cor.test(y = log(dat1$sales), x = dat1$trends2)
cor.test(y = log(dat1$sales), x = dat1$trends3)

##### Fit Model;
fit = lm(log(sales) ~ log(s1) + log(s12) + trends1, data=dat1);
summary(fit)

##### Diagnostic Plot
par(mfrow=c(2,2));
plot(fit)

#### Prediction for the next month;
predict.fit = predict(fit, newdata=dat2, se.fit=TRUE);
```