Predicting the present with Google Trends

-Hyunyoung Choi
-Hal Varian

Outline

- Problem Statement
- Goal
- Methodology
- Analysis and Forecasting
- Evaluation
- Applications and Examples
- Summary and Future work

Problem Statement

- Government agencies and other organizations produce monthly reports on economic activity
  - Retail Sales
  - House Sales
  - Automotive Sales
  - Travel
- Problems with reports
  - Compilation delay of several weeks
  - Subsequent revisions
  - Sample size may be small
  - Not available at all geographic levels
- Google Trends releases daily and weekly index of search queries by industry vertical
  - Real time data
  - No revisions (but some sampling variation)
  - Large samples
  - Available by country, state and city
- Can Google Trends data help predict *current* economic activity?
  - Before release of preliminary statistics
  - Before release of final revision

Goal

- Familiarize readers with Google Trend data and its importance
- Illustrate some simple statistical methods that use this data to predict economic activity
- Illustrate this technique with some examples

Methodology

- Query index: 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.

- http://www.google.com/insights/search

Google Insights for Search BETA

| My Account | Help | Sign out | Download as CSV

Compare by	Search terms	Filter
Search terms	Tip: Use quotation marks to match an exact phrase. ("table tennis")	Web Search
Locations	All search terms	United States / All subregions / All metros
Time Ranges	+ Add search term	2004 - present
		Real Estate / Search

Web Search Volume: Real Estate

United States, 2004 - present

All Categories > Real Estate

Interest over time

News is unavailable for specific categories.

Learn what these numbers mean

Search terms in Real Estate

Top searches

1.	real	100
2.	real estate	90
3.	mortgage	80

Rising searches

1.	zillow	Breakout
2.	national city	+80%
3.	wells fargo	+70%

Analysis and Forecasting

- **Model 0:**

$$\log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + e_t$$

- This model predicts the sales of this month using the sales of last month and 12 months ago

- **Model 1**

$$\log(y_t) \sim \log(y_{t-1}) + \log(y_{t-12}) + x_t^{(1)} + e_t$$

- This model uses an extra predictor , i.e. Google query index to predict the sales of the present.

Analysis and Forecasting

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

- Sales of present month is positively correlated with the sales of last month, the month 12 months before and the Google query
- **Note:** Coefficient corresponding to query volume is small, probably because it is not taken in logarithm form

Analysis and Forecasting

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

- There was a special promotion week in July 2005, so they have added a dummy variable to control for that observation and re-estimated the model

Few Questions

➢ Why query index, not number of queries

- "Number of queries" might vary with change in population or availability of internet or power cut.
- On the other hand, query index won't. That's why it might be a better predictor.

➢ Why Log

- It reduces the effect of the outliers
- Outlier may over-predict the sales in some month, but if we use log , its effect will be minimized

Evaluation

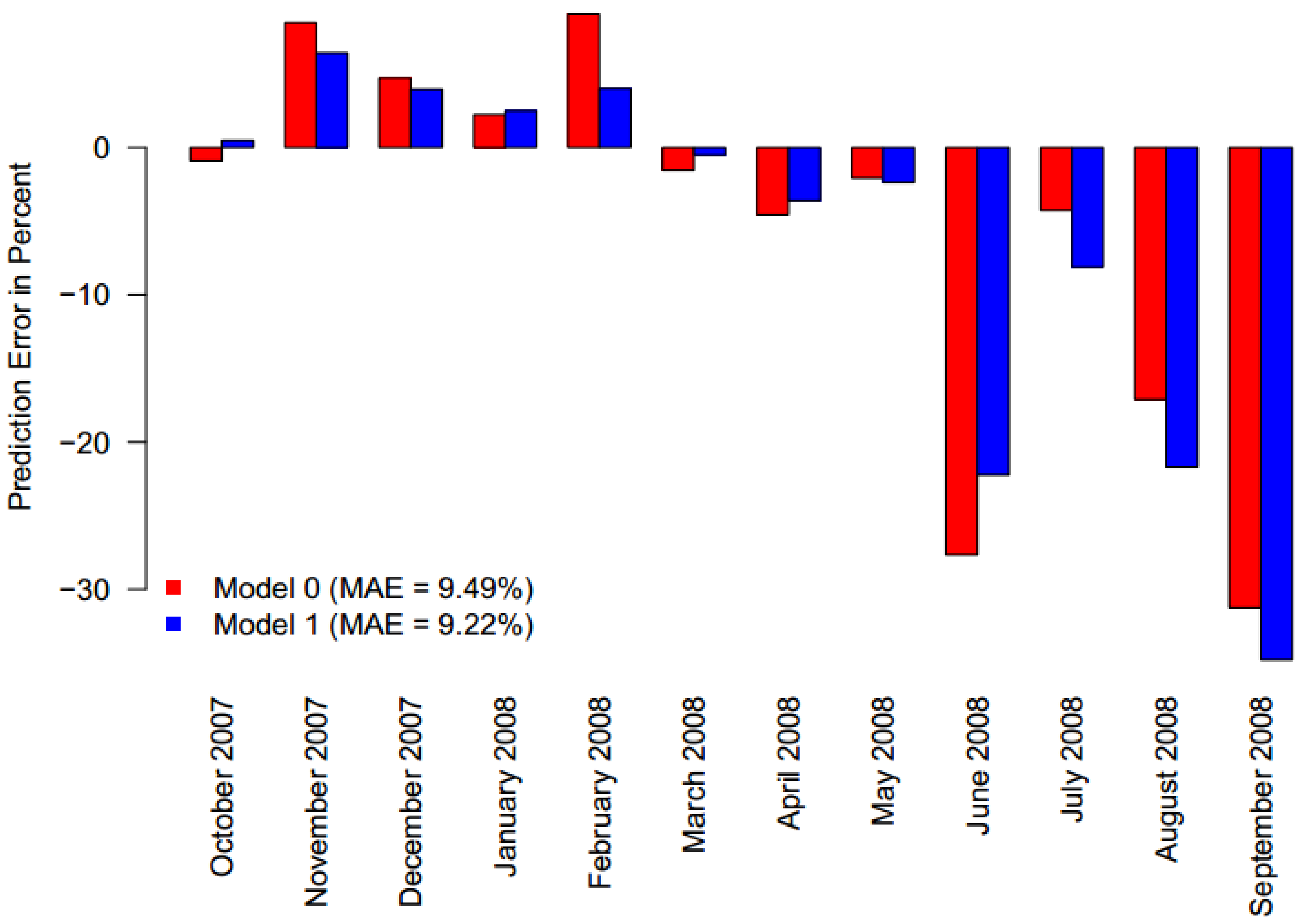
- **Prediction error:** Predicted value – observed value

$$\mathrm{PE}_t = \log(\hat{y}_t) - \log(y_t) \approx \frac{y_t - \hat{y}_t}{y_t}$$

- **Mean absolute error:** Average of the absolute values of the prediction errors

$$\mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T} |\mathrm{PE}_t|$$

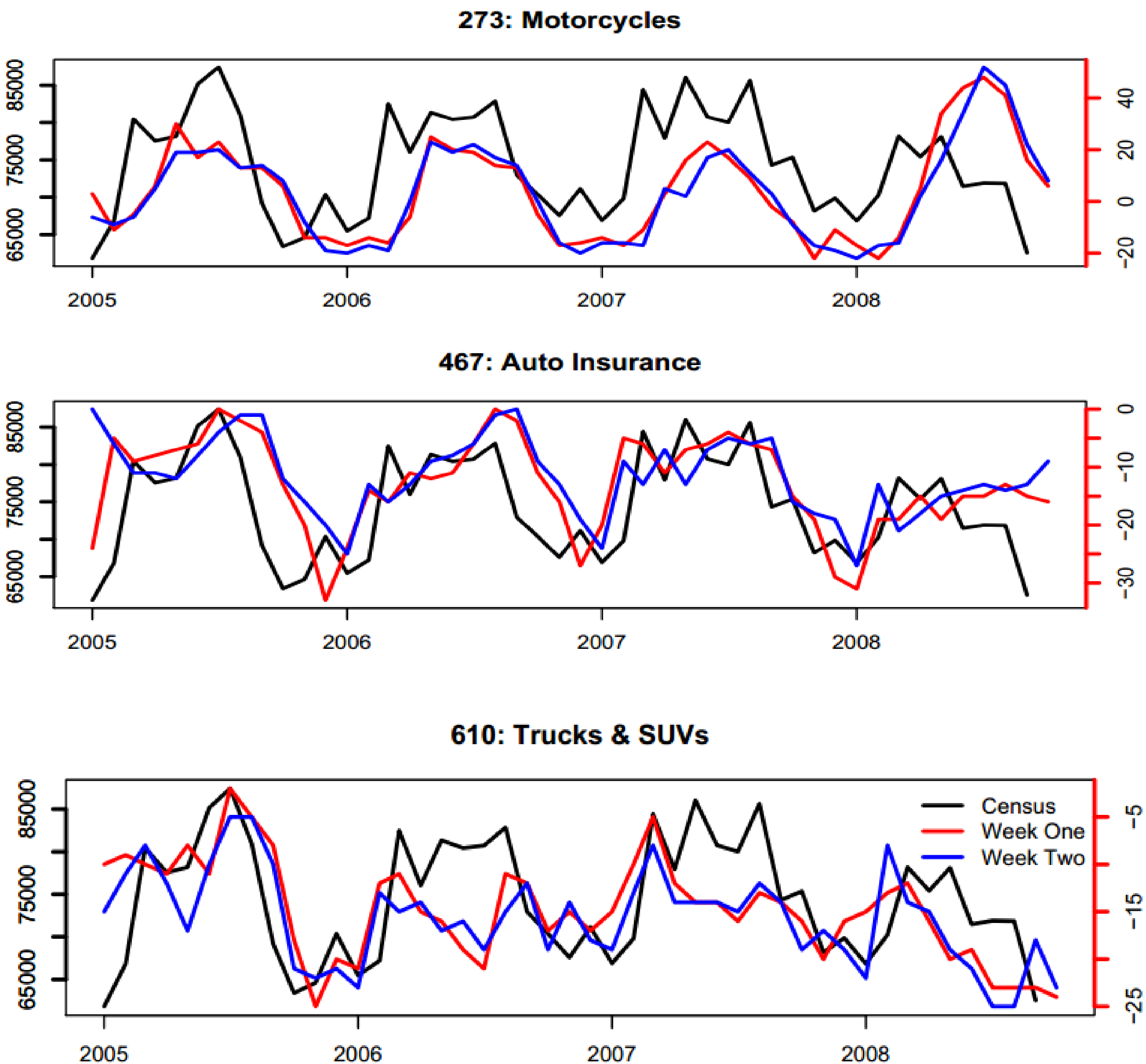
Prediction Error Plot



Example 1: Retail Sales

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


273: Motorcycles
85000
75000
65000
40
20
0
−20
2005
2006
2007
2008
467: Auto Insurance
85000
75000
65000
0
−10
−20
−30
2005
2006
2007
2008
610: Trucks & SUVs
85000
75000
65000
−5
−15
−25
Census
Week One
Week Two
2005
2006
2007
2008

Analysis and Forecasting

- Model 0:

$$\log(y_t) = 1.158 + 0.269 \cdot \log(y_{t-1}) + 0.628 \cdot \log(y_{t-12})$$

- Model 1.

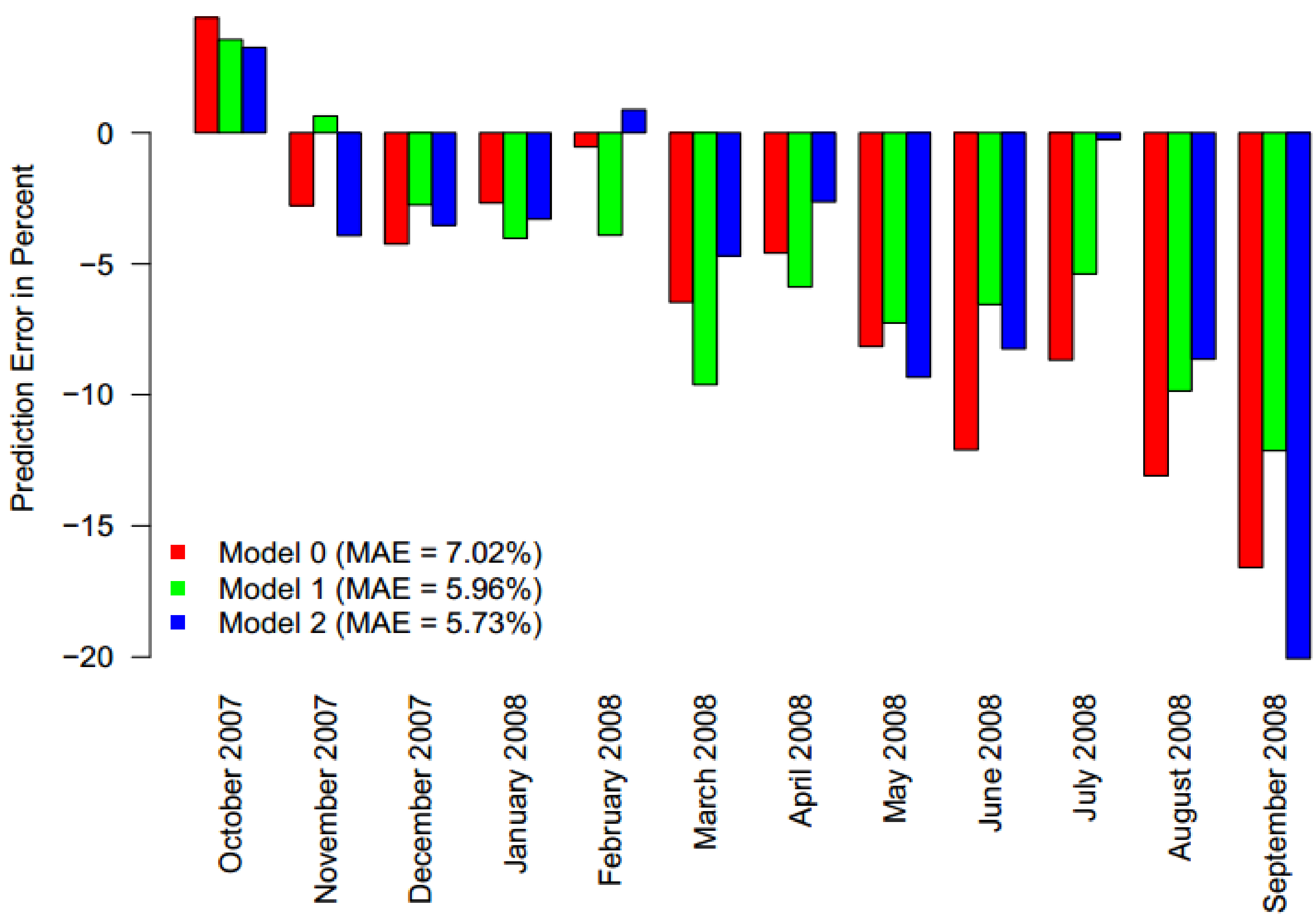
$$\log(y_t) = 1.513 + 0.216 \cdot \log(y_{t-1}) + 0.656 \cdot \log(y_{t-12}) + 0.007 \cdot x^{(1)}_{610,t}$$

- Model 2:

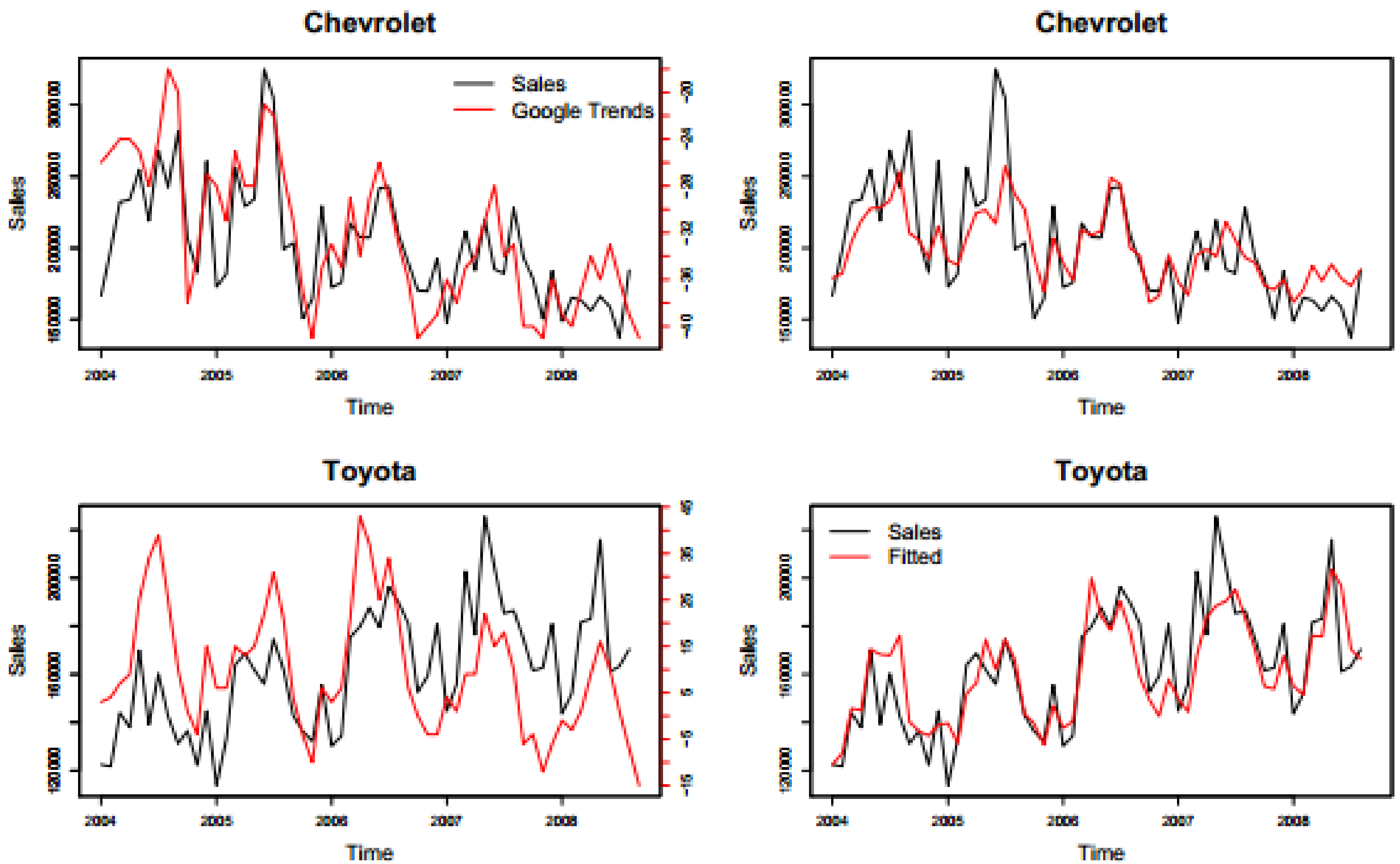
$$\log(y_t) = 0.332 + 0.230 \cdot \log(y_{t-1}) + 0.748 \cdot \log(y_{t-12})$$
$$-0.001 \cdot x^{(2)}_{273,t} + 0.002 \cdot x^{(1)}_{467,t} + 0.004 \cdot x^{(1)}_{610,t}$$

- Note: "R squares" moves from .6206(Model 0) to .7852(Model 1) to .7696(Model 2).

Prediction Error



Example 2: Automotive Sales



(a) Sales vs. Google Trends

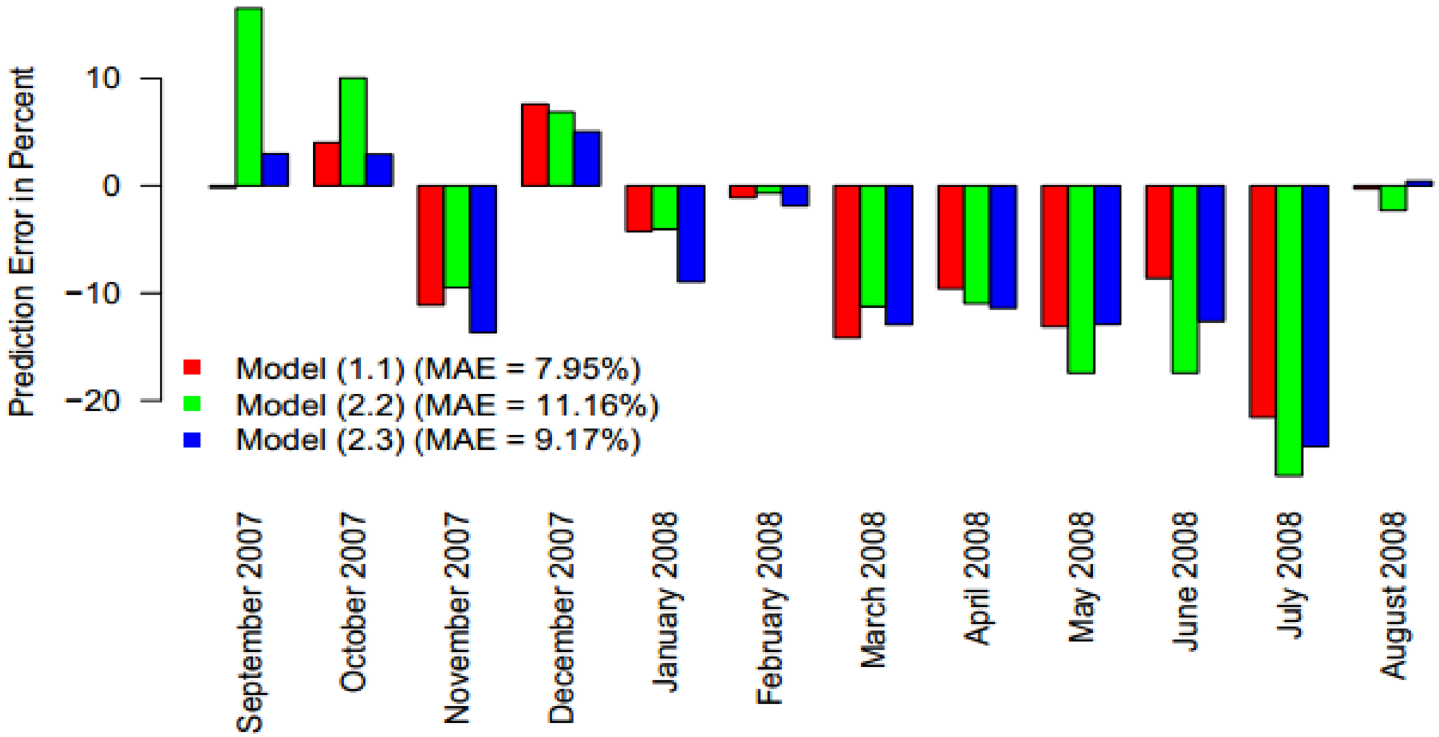
(b) Actual & Fitted Sales

Analysis and Forecasting

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

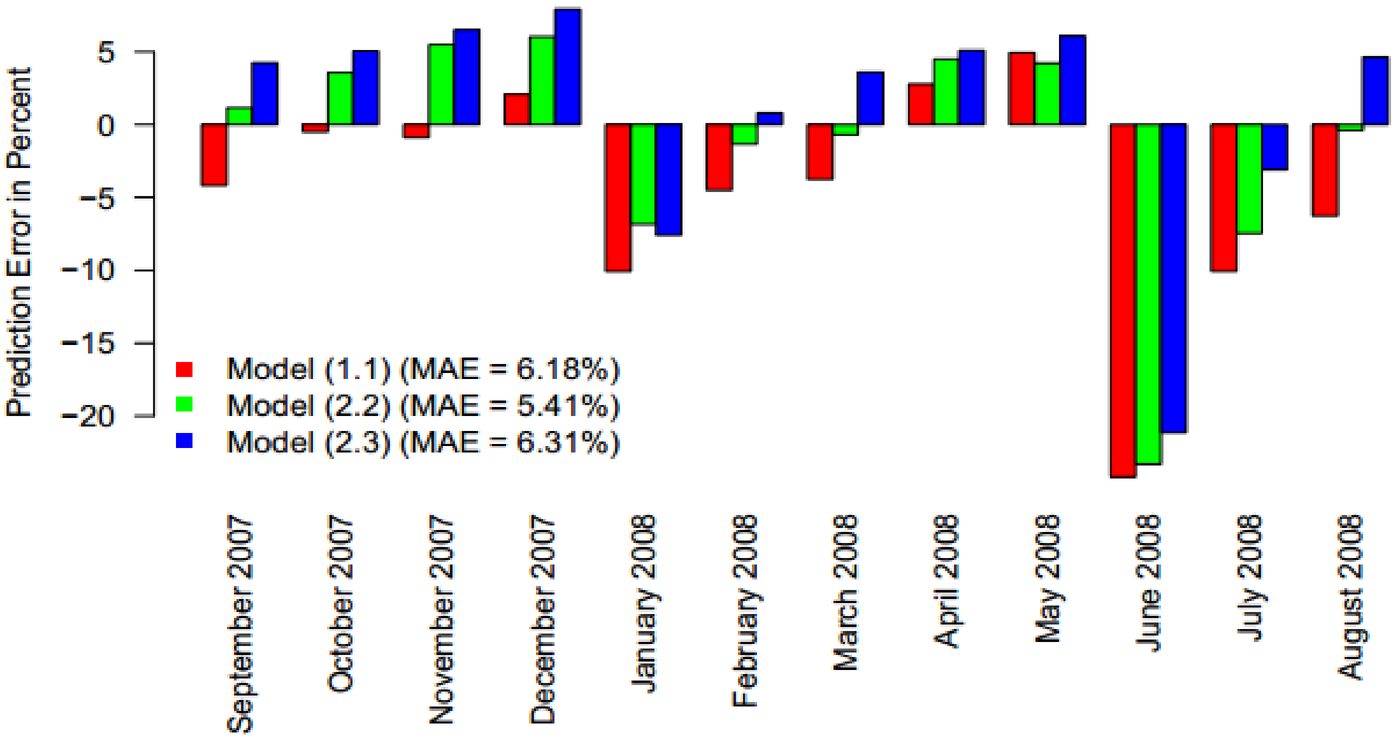
$$\begin{aligned}\text{Chevrolet} &: \log(y_{i,t}) = 7.367 + 0.439 \cdot \log(y_{i,t-12}) + 0.017 \cdot x_{i,t}^{(2)},\ e_t \sim N(0, 0.114^2) \\ \text{Toyota} &: \log(y_{i,t}) = 4.124 + 0.655 \cdot \log(y_{i,t-12}) + 0.003 \cdot x_{i,t}^{(2)},\ e_t \sim N(0, 0.093^2)\end{aligned}$$

Prediction Error of Chevrolet



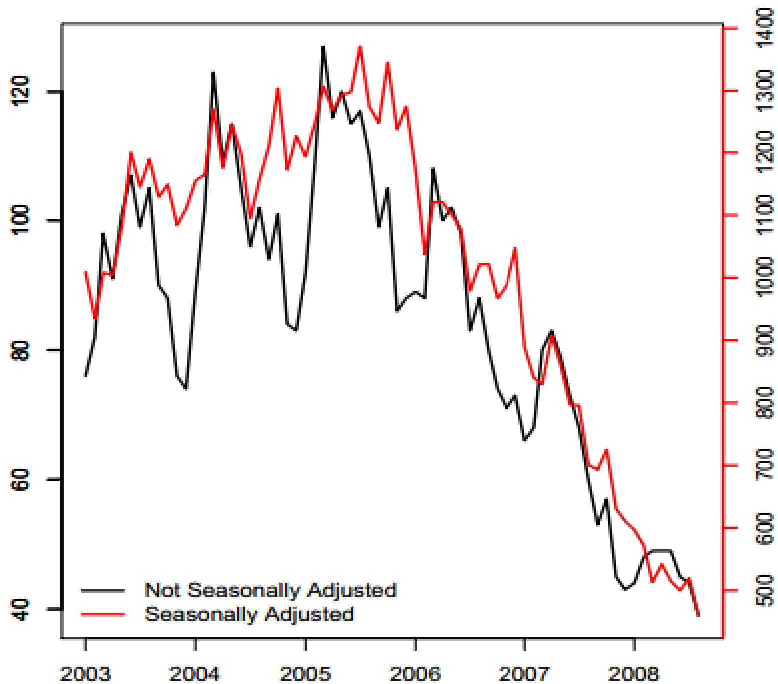
(a) Chevrolet

Prediction Error of Toyota

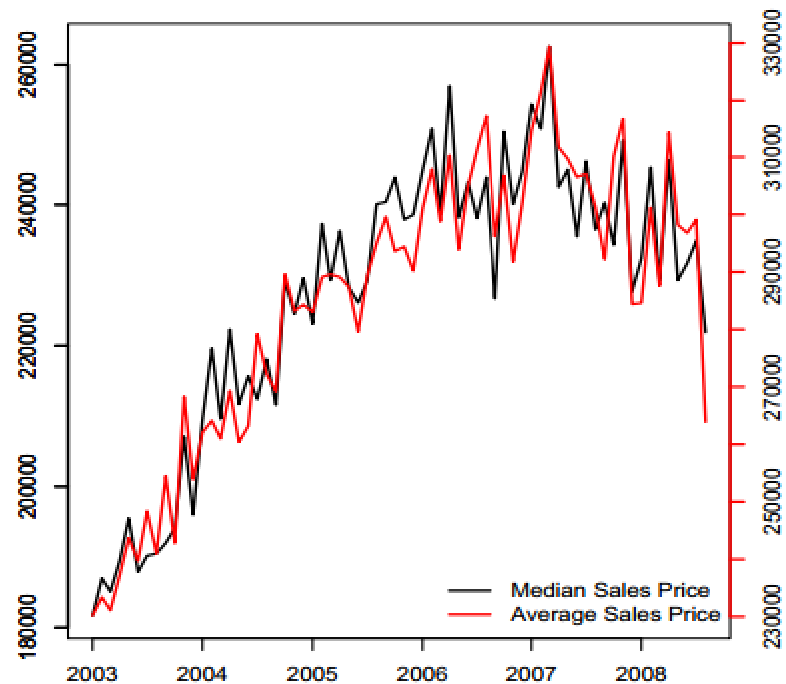


(b) Toyota

Example 3: Home Sales



(a) Number of New House Sold



(b) Prices of New House Sold

Analysis and Forecasting

- **Model 0:**

$$\log(y_t) \sim \log(y_{t-1}) + e_t$$

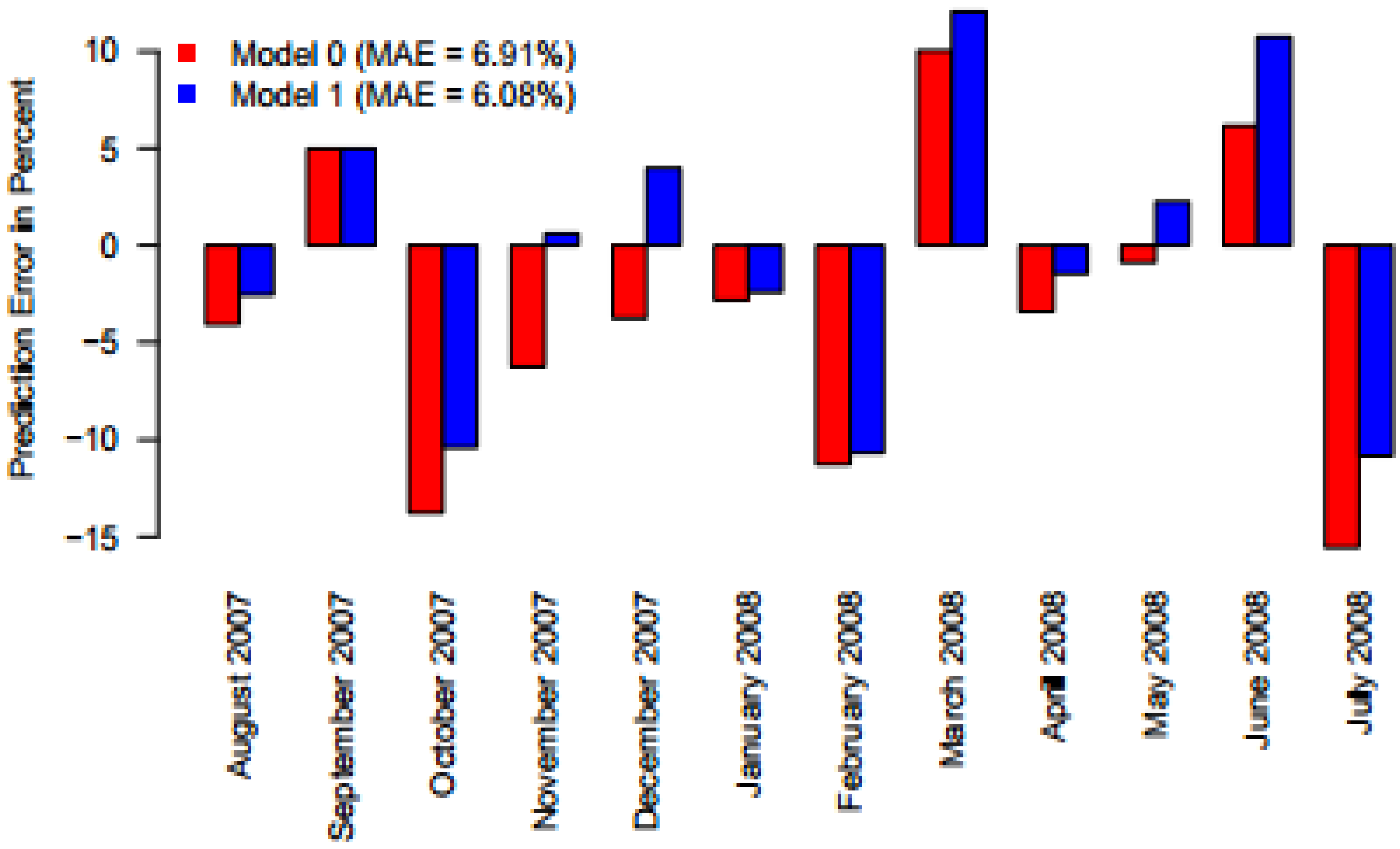
- **Model 1:**

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

- **Observations:**

- House sales at t -1 is positively related with house sales at t
- Search Index on ‘Rental Listings and Referrals” is negatively related to sales
- Search Index for “Real Estate Agencies” is positively related to sales
- Average housing price is negatively associated with sales

Prediction Error



Example 4: Travel

- Google Trend Data is useful in predicting visits to certain destination
- In this example, data has been taken from Hong Kong Tourism Board
- Data from January 2004 to August 2008 has been used.

Analysis and Forecasting

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

➢Observation

- Arrivals last month are positively related to arrivals this month
- Arrivals 12 months ago are positively related to arrivals this month
- Google searches on ‘Hong Kong’ are positively related to arrivals
- During the Beijing Olympics, travel to Hong Kong decreased.

ANOVA Table

	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			

➢Observations:

- Most of the variance is explained by lag variable of arrivals
- Google trend variable is statistically significant

Thank You

Summary

- Google Trends significantly improves prediction of Economic Activities, up to 15 days in advance of data release.
- "R squared" value improves significantly.
- Mean absolute error for predictions declines Significantly.

Further Work

- Google query data can be combined with other social network data for better prediction
- Can be used to predict the success of a movie
- Can be used for metro level data and other local data