Forecasting diffusion with pre-launch online search traffic data

Oliver Schaer
Nikolaos Kourentzes
Robert Fildes

IIF Workshop on Forecasting
New Products and Services
12th May 2016

Lancaster Centre for Forecasting
Lancaster University Management School
The game life-cycle

Call of Duty 3

Wii Sports
Generations of games

Sales for Assassin's Creed

Google Trends
Ways to obtain model parameters in a pre-launch setting*

By judgement
- However, time consuming and problem with adjustment bias (Fildes et al. 2009)

Forecasting by analogy
- Using parameters from previous or similar products (Kim et al. 2014, Lillien et al. 2000, Norton & Bass 1987)

Market research
- Survey (Bass et al. 2001), Product attributes (Goodwin et al. 2012) or Pre-orders on CD albums (Moe & Fader 2002)

*See Goodwin et al. (2014) for a discussion on challenges with pre-launch forecasting
Using information from online sources

Using online explanatory variables for pre-launch forecasts

- Forecasting computer game sales using search traffic and social buzz for the opening sales (Xiong & Bharadwaj 2014)
- Box office sales using social network data (Kim et al. 2015) and search traffic (Kulkarni et al. 2012)

No application to life-cycle forecasting and parameter estimation for diffusion models.

I. Can search traffic data help in estimating diffusion model parameters?
Experiment – Motivation

Model
- Bass model (Bass 1969)

Target
- Incorporate search traffic information into the analogy based forecast approach.

Aim
- Estimate market size parameter
- We are also interested in seeing whether there is lead time

Note that the aim is not to build the best possible fit but examine whether search traffic information improves the forecasts.
The data

Video game sales from VGchartz
- Global physical sales at weekly frequency
- Using 6 games series with a total of 43 games
- Sales are aggregated across gaming platforms such as PC, Xbox, PS3 or Wii

Search Traffic popularity from Google Trends
- Weekly global search traffic popularity information
- Topic search with game title as keyword
Model set up

Bass model (Bass 1969) for generation $j = 2, 3...J$

\[ Y_{j,t} = Y_{j,t-1} + p_{j-1} (\hat{m}_j - Y_{j,t-1}) + q_{j-1} \frac{Y_{j,t-1}}{\hat{m}_j} (\hat{m}_j - Y_{j,t-1}) \]

$Y_{j,t}$ = cumulative number of adaptors at time $t$ for generation $j$

$p_{j-1}$ = coefficient of innovation from the previous generation $j - 1$

$q_{j-1}$ = coefficient of imiation from the previous generation $j - 1$

$\hat{m}_j$ = estimated market size for generation $j$

Google Trends handling

GT = topic search data

Peak scaled at time $t$ to the highest observed search traffic in the series

$$\text{GT}_j = \sum_{i=1}^{w} \text{GT}_{j,(t-l)+w}$$

$w =$ Window Size $l =$ Lead Time

For the experiment we use Window Size of 6 and lead time with 1 and 6 weeks
Estimation process

Sales

- Actuals
- Forecast

Search traffic

Peak scaled search traffic

$j=1$ $j=2$ $j=3$ $j=4$
Benchmark models and accuracy measure

Naïve: \[ m_j = m_{j-1} \]

Naïve + Difference: \[ m_j = m_{j-1} + \Delta m_{j-1} \]

Linear Trend: \[ m_j = \alpha_0 + \alpha_1 j + \epsilon_j \]

AR(1): \[ m_j = \alpha_0 + \alpha_1 m_{j-1} + \epsilon_j \]

+ “optimal” fitted Bass model with actuals

Actuals contain two years

Numbers of generations needed for model estimation vary

Relative Mean Absolute Error and median across series
Google Trend model selection

Percentage Increase:

Linear trend:

AR(1) + Percentage Increase:

AR(1) + Google Trend:

\[
m_j = \frac{\text{GT}_j}{\text{GT}_{j-1}} \cdot m_{j-1}
\]

\[
m_j = \alpha_0 + \text{GT}_j + \epsilon_j
\]

\[
m_j = \alpha_0 + \alpha_1 m_{j-1} + \beta_1 \frac{\text{GT}_j}{\text{GT}_{j-1}} + \epsilon_j
\]

\[
m_j = \alpha_0 + \alpha_1 m_{j-1} + \beta_1 \text{GT}_j + \epsilon_j
\]

<table>
<thead>
<tr>
<th></th>
<th>PI</th>
<th>Linear Trend</th>
<th>AR(1) PI</th>
<th>AR(1) GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000</td>
<td>1.079</td>
<td></td>
<td>1.031</td>
<td>1.070</td>
</tr>
</tbody>
</table>

<1 = better

No. Series = 6, Window Size = 6, Lead Time = 1
### Performance across generations

**Assassin’s Creed**

<table>
<thead>
<tr>
<th>Generation</th>
<th>Naïve</th>
<th>Naïve Diff.</th>
<th>Linear Trend</th>
<th>AR(1)</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game 1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Game 2</td>
<td>0.066</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.062</td>
</tr>
<tr>
<td>Game 3</td>
<td>1.746</td>
<td>1.925</td>
<td>1.925</td>
<td>NA</td>
<td>0.986</td>
</tr>
<tr>
<td>Game 4</td>
<td>1.223</td>
<td>1.993</td>
<td>1.382</td>
<td>4.347</td>
<td>1.012</td>
</tr>
<tr>
<td>Game 5</td>
<td>1.020</td>
<td>0.925</td>
<td>1.083</td>
<td>1.071</td>
<td>0.937</td>
</tr>
<tr>
<td>Game 6</td>
<td>1.165</td>
<td>1.334</td>
<td>1.103</td>
<td>0.998</td>
<td>0.76</td>
</tr>
<tr>
<td>Game 7</td>
<td>1.035</td>
<td>1.087</td>
<td>1.025</td>
<td>0.898</td>
<td>0.627</td>
</tr>
</tbody>
</table>

<1 = better

Window Size = 6, Lead Time = 1
### Performance across series

#### Lead time 1 week

<table>
<thead>
<tr>
<th>Naïve</th>
<th>Naïve Diff.</th>
<th>Linear Trend</th>
<th>AR(1)</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.020</td>
<td>1.087</td>
<td>1.086</td>
<td>1.070</td>
<td>0.866</td>
</tr>
</tbody>
</table>

<1 = better

No. Series = 6, Window Size = 6

#### Lead time 6 weeks

<table>
<thead>
<tr>
<th>Naïve</th>
<th>Naïve Diff.</th>
<th>Linear Trend</th>
<th>AR(1)</th>
<th>Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.007</td>
<td>1.143</td>
<td>1.088</td>
<td>1.033</td>
<td>0.861</td>
</tr>
</tbody>
</table>

<1 = better

No. Series = 6, Window Size = 6
Conclusion

Fully automated Bass model market size parameter estimation method that includes information from search traffic.

Google trend percentage increase market size estimation method outperformed most benchmark models.
Next steps

- Increase sample size
- Further improve Google Trends estimation, i.e. take the shape of the search traffic (Xiong & Bhradwaj 2014)
- Use machine learning for identifying market sizes and remaining model parameters for new game series (Lee et al. 2014).
- Look at diffusion models with higher flexibility to allow for better model fit such as the Gompertz or Weibull (Meade & Islam 2006, Moe & Fader 2002) or two stage models (Van den Bulte & Joshi 2007).
- Include the multi-generation and leapfrogging effect (Jiang & Jain 2012)
Thank you!

Great Aletsch Glacier, Switzerland


References


