Let's Google it: predicting the Norwegian macroeconomy with Google search indices

Jon Ellingsen
University of Oslo
ellingsen.jon@gmail.com

March 27, 2017

Abstract

I construct indices of Google searches to predict retail sales, unemployment and house prices in Norway.

Contents

1 Introduction

Predicting key economic aggregates is an important task for policy makers, including the central bank. Getting real-time information and updated information sets about the state of the economy helps improving the policy responses. However, variables like retail sales and unemployment are not observed in real-time as they are published with a lag of approximately 1 month. Traditionally the input in the prediction models has been standard economic data. This approach suffers from at least two drawbacks. Firstly, standard economic data is often highly aggregated and thus may not capture important dissagregated events. Secondly, the data is typically not real-time, as they are reported with significant lags.

Nowadays we face many new types of data due to technological improvements. Some of these data are referred to as big data, see section 2. From a prediction perspective, these new sources of information have the potential to provide us with more accurate assessments of the economic fluctuations. Google is an example of a company that gathers such valuable information. Through their service, Google Trends [1], everyone can access real time data on what people are searching for at their search engine. As Choi and Varian (2012) point out, these query indices are often correlated with economic indicators and may be helpful for short-term prediction. This field is known as nowcasting and refers to predicting the state of a variable contermporaneously with contemporaneous data.

Several contributions to the existing litterature on nowcasting with Google search queries faces substantial drawbacks. To a large extent, the power of using Google search queries as predictors lies in the ability to choose the top predictors from a set of potential candidates. The traditional approach has been to use single words as predictors, and the words are typically chosen from a theoretical perspective, see e.g. McLaren and Shanbhogue (2011), Anvik and Gjelstad (2010) and Vosen and Schmidt (2012). However, this approach is highly sensitive to e.g. different spellings and synonyms, but also to breaks in the search of simple words, that does not reflect an underlying force valuable for prediction. Da et al. (2015) has a rather sophisticated way of dealing with this challenge. They use online dictionaries to obtain a large set of potential Google search queries and perform backward-rolling regressions to pick the best predictors. Another interesting contribution is Stephens-Davidowitz and Varian (2014) whom develop a method they call bayesian structural time series to choose the top predictors from a set of correlated query indices.

I propose a framework that takes into account both the challenges associated with using simple words, and the challenges associated with choosing among the set of potential

[1] google.com/trends

predictors. Instead of using simple words as potential predictors, I use categories organized by Google. Google trends consists of more than 1400 categories, and Google has developed an algorithm for classifying search words into one or more categories get a source on this algorithm. Firstly, my empirical evidence suggests that these categories are much more stable over time than simple words. Secondly, using categories captures a larger scope of the underlying trends, while still being substantially dissagregated. I choose the potential categories based on simple economic theory and the components of retail sales, unemployment and house prices. As a cross-check of the relevance of the categories I look at the top 25 queries within each category, and remove the categories which are clearly not related to the variable of interest. To select the top predictors from the set of categories I use a method known as Least Angle Regression which is an algorithm for fitting linear regression models to high-dimensional data. This method was first introduced in Efron et al. (2004) and later developed for time series by Gelper and Croux (2008). The Least Angle Regression algorithm identifies the most informative predictors and uses them for prediction.

The rest of this paper is organized as follows. Section 2 describes nowcasting and big data. Section 3 describes the data. Section 4 describes the methodology. The empirical results are presented in section 5. Section 6 concludes.

2 Nowcasting and big data

Big data and nowcasting. Present the litterature and methods.

Big data is characterized by the three v's. Velocity, variety and volume. Velocity describes the gathering process as close to real time, variety refers to different types of structures on the data (examples are newspaper text [2] or twitter feeds) and volume refers to the amount of data available.

3 Data

I use the service Google trends to collect the data on Google searches. Google trends reports time series containing the relative frequency of a given search query, within a specified region. Thus, the data is adjusted for the general upward trend in Google search volume. The highest point in the series is set to 100, and the rest of the observations are adjusted accordingly. The data from Google trends are sampled. This means that I get an estimate of the actual population relative frequency, and thus the downloaded data

[2] See Thorsrud et al. (2016)

Table 1. Data. January 2004 - January 2017.

Variable	**Source**	**Frequency**	**Comments**
Retail sales	Statistics Norway	Monthly	Volume index Seasonally adjusted from source Transformation: Monthly growth rate Extracted: 13.03.17
Unemployment rate	NAV	Monthly	Registered unemployed Seasonally adjusted from source Transformation: Monthly change in the change in the unemployment rate Extracted: 23.03.17
House prices	Statistics Norway	Monthly	Price index Seasonally adjusted from source Transformation: Monthly growth rate Extracted: 13.03.17

Table 2. Descriptive statistics

	Retail sales*	Unemployment rate**	House prices*
Mean	0.001532	0.002788	0.010747
Median	0.005642	0.006392	0.006668
Standard deviation	0.001130	0.000048	0.090521

*Retail sales and house prices are monthly growth rates.

**Unemployment rate is the monthly change in the monthly change in the unemployment rate.

will vary from time to time, also historically. I quantify this effect by downloading the series at consecutive dates, and report the correlations.

The index for retail sales is published by Statistics Norway once a month, normally 28-30 days after the end of the month. The survey based unemployment rate is published by Statistics Norway once a month, normally 4 weeks after the end of the month. The house price index from Real Estate Norway is published once a month, the third business day in the succeeding month.

The data for retail sales, unemployment rate and house prices are seasonally adjusted from the source.

The unemployment rate in the period January 2004 - January 2017 is non-stationary. In order to obtain a stationary time series, the unemployment rate is differenced twice. Let the unemployment rate in period t be u_t. Then, the variable I work with will be

defined as

$$\tilde{u}_t = \Delta u_t - \Delta u_{t-1} = (u_t - u_{t-1}) - (u_{t-1} - u_{t-2}) \tag{1}$$

From equation (1) we can easily back out the variable of interest, u_t. Rearranging terms, we get

$$u_t = \tilde{u}_t + 2u_{t-1} - u_{t-2} \tag{2}$$

In period t, u_{t-1} and u_{t-2} is known, and the model will give an estimate on $\tilde{u}_t$.

4 Methodology

My perspective. Use "new" methods on Norwegian data. High frequent, big data, easy available, good predictions, in line with theory.

With all the candidate predictors obtained from Google Trends, some kind of variable selection method is preferred in order to choose a subset of variables for prediction. On the one hand, adding too many explanatory variables increases the risk of overfitting the model and reducing the degrees of freedom. At the other hand, adding too few explanatory variables may reduce the explanatory power and robustness of the model. All the variables have been standardized, i.e. they have a mean of zero and a standard deviation of one. Hence, no intercept is included in the model.

Least-angle regression (LARS) is an algorithm for fitting linear regression models to high-dimensional data, see Efron et al. (2004). Gelper and Croux (2008) developes LARS for time series, and call it the TS-LARS. I use the TS-LARS for variable selection.

Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients. Instead of giving a vector result, the LARS solution consists of a curve denoting the solution for each value of the L1 norm of the parameter vector. The algorithm is similar to forward stepwise regression, but instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each one's correlations with the residual.

It is commonly recognized in time series analysis that lagged values of both the response variable, which we want to predict, and the predictors might contain predictive information. To account for these dynamic relationships, predictors are selected as blocks of present and lagged values of a time series.

There are at least two advantages of using TS–LARS as an alternative. First, the number of variables can be larger than the number of observations. This is not the case

in the Gets procedure, which starts by fitting an unrestricted model. And secondly, the TS–LARS algorithm involves no testing.

The unrestricted model with all the potential predictors:

$$y_t = \beta_{0,0}y_{t-1} + \ldots + \beta_{0,p}y_{t-p} + \beta_{1,0}x_{1,t} + \ldots + \beta_{1,p}x_{t-p} + \ldots + \beta_{m,0}x_{m,t} + \ldots + \beta_{m,p}x_{m-p} + \varepsilon_t$$

LARS predictor ranking algorithm [3]

1. Start with all coefficients β_j equal to zero.
2. Find the predictor x_j most correlated with y.
3. Increase the coefficient β_j in the direction of its correlation with y. Take the residuals $r = y - \hat{y}$ along the way. Stop when some other predictor x_k has as much correlation with r as x_j has.
4. Increase (β_j, β_k) in their joint least squares direction until some other predictor x_m has as much correlation with the residual r.
5. Continue until all predictors are in the model.

[3] http://statweb.stanford.edu/ tibs/lasso/simple.html)

- Let z_0 be the response variable
- Find the predictor block $\underline{x}_{(1)}$, which is the matrix of lagged $x_{(1)}$ values, with the largest R^2 with z_0
- This is the first predictor included in the active set, A
- Let $\hat{z}_0$ be the fitted values, $\hat{z}_0 = H_{(1)} z_0$
- Update the response by removing the effect of the first variable

$$z_1 = z_0 - \gamma_1 \hat{z}_0$$

  but where γ_1 is not equal to 1 as in the OLS case, when z_1 simply contains the OLS residuals of regressing z_0 on $\underline{x}_{(1)}$
- γ_1 is chosen such that

$$R^2(z_0 - \gamma_1 \hat{z}_0 \sim \underline{x}_{(1)}) = R^2(z_0 - \gamma_1 \hat{z}_0 \sim \underline{x}_j) \qquad \text{where} \qquad x_j \in A^c$$

- In particular, the second time series in the active set is the one with index j yielding the smallest positive value of γ_1
- Denote this predictor by $x_{(2)}$. Now A contains two predictors
- Obtain the response z_1 from

$$z_1 = z_0 - \gamma_1 \hat{z}_0$$

- At the beginning of step k, the active set A contains k active or ranked predictors $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$, with $k \geq 2$. The current response is denoted by z_{k-1}. Let $\tilde{x}_{(i)}$ be the standardized vector of fitted values $H_{(i)} z_{i-1}$ for $i = 1, \ldots, k$.
- First, we look for the equiangular vector u_k, which is defined as the vector having equal correlation with all vectors $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}$. This correlation is denoted by

$$a_k = Cor(u_k, \tilde{x}_{(1)}) = Cor(u_k, \tilde{x}_{(2)}) = \ldots = Cor(u_k, \tilde{x}_{(k)})$$

- Let R_k be the $(k \times k)$ correlation matrix computed from $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}$ and $\mathbf{1_k}$ a vector of ones of length k. The equiangular vector u_k is then a weighted sum of $\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}$:

$$u_k = (\tilde{x}_{(1)}, \tilde{x}_{(2)}, \ldots, \tilde{x}_{(k)}) w_k \qquad \text{with} \qquad w_k = \frac{R_k^{-1} \mathbf{1_k}}{\sqrt{\mathbf{1_k}' \mathbf{R_k^{-1}} \mathbf{1_k}}}$$

- Note that the equiangular vector has unit variance. Afterwards, the response is updated by moving along the direction of the equiangular vector, $z_k = z_{k-1} - \gamma_k u_k$. The shrinking factor γ_k is chosen as the smallest positive solution such that for a predictor x_j not in the active set it holds that

$$R^2(z_{k-1} - \gamma_k u_k \sim \tilde{x}_{(k)}) = R^2(z_{k-1} - \gamma_k u_k \sim \underline{x}_{(j)})$$

- The associated predictor, denoted by $x_{(k+1)}$, is then added to the active set. Once γ_k is obtained, we can update the response and the new response is then standardized and again denoted by z_k.

The algorithm described above is used for different values of the lag length p. For each p I obtain the model with a number of predictors minimizing BIC. The final prediction model, then, is the model obtained by minimizing BIC further over all the considered values of p. I fix the lag length across the predictors for simplicity.

5 Empirical results

The baseline models are the AR(1) and a random walk.

5.1 Retail sales

Table 3. Root mean squared prediction error

	Extending window	Rolling window
AR(1)	0.0157	0.0157
Random walk	0.0183	0.0183
Plain Google model	0.0111	0.0116
% change from AR(1)	(-29)	(-26)
% change from random walk	(-39)	(-36)
AR(1) & Google model	0.0102	0.0110
% change from AR(1)	(-35)	(-30)
% change from random walk	(-44)	(-40)

Table 4. Diebold-Mariano test. HAC-robust standard errors.

	AR(1)	Random walk
	Plain Google model	
Extending window	0.0001^{***}	0.0002^{***}
Rolling window	0.0001^{***}	0.0002^{***}
	AR(1) & Google model	
Extending window	0.0001^{***}	0.0002^{***}
Rolling window	0.0001^{***}	0.0002^{***}

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

5.2 Unemployment rate

Table 5. Root mean squared prediction error

	Extending window	Rolling window
AR(1)	0.1345	0.1343
Random walk	0.1551	0.1551
Plain Google model	0.0895	0.0906
% change from AR(1)	(-33)	(-33)
% change from random walk	(-42)	(-42)
AR(1) & Google model	0.0710	0.0685
% change from AR(1)	(-47)	(-49)
% change from random walk	(-54)	(-56)

Table 6. Diebold-Mariano test. HAC-robust standard errors.

	AR(1)	Random walk
	Plain Google model	
Extending window	0.010^{***}	0.016^{***}
Rolling window	0.010^{***}	0.016^{***}
	AR(1) & Google model	
Extending window	0.013^{***}	0.019^{***}
Rolling window	0.013^{***}	0.019^{***}

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

5.3 House prices

Table 7. Root mean squared prediction error

	Extending window	Rolling window
AR(1)	0.0019	0.0027
Random walk	0.0062	0.0062
Plain Google model	0.0049	0.0049
% change from AR(1)	(+159)	(+80)
% change from random walk	(-20)	(-21)
AR(1) & Google model	0.0054	0.0054
% change from AR(1)	(+182)	(+99)
% change from random walk	(-13)	(-12)

Table 8. Diebold-Mariano test. HAC-robust standard errors.

	AR(1)	Random walk
	Plain Google model	
Extending window	-0.00002^{***}	0.00001^{*}
Rolling window	-0.00002^{***}	0.00001^{*}
	AR(1) & Google model	
Extending window	-0.00003^{***}	0.00001^{***}
Rolling window	-0.00002^{***}	0.00001^{**}

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

6 Conclusion

References

Anvik, C. and Gjelstad, K. (2010), '" just google it": Forecasting norwegian unemployment figures with web queries'.

Choi, H. and Varian, H. (2012), 'Predicting the present with google trends', *Economic Record* **88**(s1), 2–9.

Da, Z., Engelberg, J. and Gao, P. (2015), 'Editor's Choice The Sum of All FEARS Investor Sentiment and Asset Prices', *Review of Financial Studies* **28**(1), 1–32.
URL: *https://ideas.repec.org/a/oup/rfinst/v28y2015i1p1-32..html*

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Gelper, S. and Croux, C. (2008), 'Least angle regression for time series forecasting with many predictors'.

McLaren, N. and Shanbhogue, R. (2011), 'Using internet search data as economic indicators', *Bank of England Quarterly Bulletin* (2011), Q2.

Stephens-Davidowitz, S. and Varian, H. (2014), 'A hands-on guide to google data', *Tech. Rep.* .

Thorsrud, L. A. et al. (2016), Words are the new numbers: A newsy coincident index of business cycles, Technical report.

Vosen, S. and Schmidt, T. (2012), 'A monthly consumption indicator for germany based on internet search query data', *Applied Economics Letters* **19**(7), 683–687.