Time series analysis of Google Trends data

Mckay Jensen*

December 2019

1 Introduction

This project is an investigation of linear interdependencies in Google Trends time series data and a basic demonstration of the power of Google Trends data to predict economic variables. Using vector autoregression (VAR) models, I find significant relationships between Trends variables and show that variables from Google Trends Granger-cause monthly changes in personal consumption in the U.S.

2 Background

Online data is readily available, abundant, and constantly generated by various web services. Given this, online data is a promising source of data for economic research. For example, I have in the past shown that measures of sexist sentiment generated by machine learning analysis of Twitter data are strongly correlated with many measures of women's economic well-being.[1] Textual data from social media has also been shown to have significant predictive power for financial markets (Mao et al., 2011). Some economists have taken notice of the power of such data and have begun developing computational tools to unlock their power for research applications; Gentzkow et al. (2017) have provided some guidelines for using text (such as text from social media or online news sources) as data in economic research. These

*The R code used for the analysis in this project is available at `https://github.com/quevivasbien/google-trends-time-series`

[1] See `https://github.com/quevivasbien/twitter-sexism/blob/master/twitter_sexism.pdf`

methods can allow economists to unlock new sources of data for their research; however, data already in a quantitative form is even more attractive since it does not require preliminary feature analysis. To this point in time, much of the interest in quantitative data from the internet has centered on data from Google Trends, which allows anyone to view and download time series of indices for various search terms and categories. Work in this area began with Choi and Varian's (2009) demonstration of the power of Google Trends to "predict the present" in sales data. Since then, other researchers have used a variety of computational tools to explore connections between Google search traffic and various economic variables. Della Penna and Huang (2009) created an index based on Google Trends that compares favorably with survey-based indices and shows significant relationships with changes in retail sales and consumption. Vosen and Schmidt (2011) took a similar approach in constructing an index from Google Trends that can be used to forecast private consumption. This project is in some sense an update on those research efforts: the Google Trends interface has changed significant over the past decade, and there is of course much more data available to analyze – Google began compiling Trends data in 2004, so at this point there is about three times more time series data available than there was when Choi and Varian's paper was published. I also take a slightly different direction than those in previous research efforts by paying some attention to linear relationships *between* Trends variables, conducting Granger causality tests for consumption, and estimating impulse response functions for the data.

3 Trend variables

The Google Trends data were collected from `https://trends.google.com/trends/`. Google Trends data is automatically sorted into a number of predefined categories; I chose to focus on a subset of 7 of those categories: Business & Industrial, Finance, Health, Hobbies & Leisure, Jobs & Education, Shopping, and Travel. I filtered to focus only on search traffic from Google users within the U.S. I collected the data on a monthly basis, from January

2004 (the earliest available date) up to November 2019.

These data reflect the total search volume as a percentage of total Google search traffic. Therefore, most of the chosen categories exhibit long run trends related to changing uses of the Internet over time rather than to fluctuations in relative interest in the categories, which is what I wanted to examine. To focus on those changes in interest, I first processed each series by removing a linear time trend; i.e., I calculated the regression

$$[trend]_t = \beta_0 + \beta_1 month_t + e_t$$

and kept the residual e_t.

I then used a series of Dickey-Fuller and Dickey-Fuller Augmented Least Squares (ADF-GLS) tests to check the stationarity of each time series, concluding that all of the detrended time series are stationary (no unit root, $p < 0.05$).

To determine the linear relationships between each variable, I constructed a vector autoregression (VAR) model on all of the time series. Because many of the time series exhibit pronounced seasonality (for example, the Shopping series spikes every year in November and December), I included centered seasonal dummy variables for each month of the year. Based on the Akaike, Hannan-Quinn, Schwarz, and final prediction error criteria, the optimal VAR model includes only one lag in all variables – the seasonality of the variables seems to explain most of the dynamics. The general form of the VAR equation for each variable is

$$\begin{aligned}[\text{variable}]_t = \beta_0 &+ \beta_1 \text{business\_industrial}_{t-1} + \beta_2 \text{finance}_{t-1} + \beta_3 \text{health}_{t-1} \\ &+ \beta_4 \text{hobbies\_leisure}_{t-1} + \beta_5 \text{jobs\_education}_{t-1} + \beta_6 \text{shopping}_{t-1} \\ &+ \beta_7 \text{travel}_{t-1} + \sum_{i=1}^{12} \beta_{7+i} [\text{is season } i]_t\end{aligned}$$

The estimated coefficients of this VAR model are shown in tables 1 and 2. I should note that the estimates are essentially the same if the seasonal part of each time series is removed

before estimating the VAR model rather than including the seasonal dummy variables in the model. For brevity, I have not included the coefficients for the seasonal parts in the tables.

Table 1: Trends VAR coefficients, 1/2

	Dependent variable:			
	business_industrial	finance	health	hobbies_leisure
business_industrial.l1	0.247***	−0.167	−0.0003	−0.144
	(0.090)	(0.164)	(0.102)	(0.111)
finance.l1	−0.032	0.537***	−0.034	−0.022
	(0.040)	(0.073)	(0.045)	(0.049)
health.l1	−0.118*	0.014	0.421***	0.225**
	(0.071)	(0.129)	(0.081)	(0.087)
hobbies_leisure.l1	0.153**	−0.129	0.235***	0.223**
	(0.073)	(0.133)	(0.083)	(0.090)
jobs_education.l1	−0.008	−0.260*	0.005	0.261**
	(0.083)	(0.151)	(0.094)	(0.102)
shopping.l1	0.122**	0.420***	0.120*	0.055
	(0.055)	(0.100)	(0.063)	(0.068)
travel.l1	0.170***	0.099	−0.026	0.065
	(0.058)	(0.106)	(0.066)	(0.071)
const	−0.031	−0.027	−0.021	0.056
	(0.107)	(0.195)	(0.121)	(0.131)
Observations	190	190	190	190
Note:				*p<0.1; **p<0.05; ***p<0.01

Unsurprisingly, each variable is strongly correlated with itself in one lag. Other strongly statistically significant connections include the lag of Travel on Business & Industrial, the lag of Shopping on Finance, the lag of Hobbies & Leisure on Health, and the lag of Shopping on Jobs & Education. It is not clear what the sources of these connections are, although I can propose some very rudimentary explanations: for example, maybe the connection between Shopping and Finance can be explained because people begin worrying about their finances after spending lots of money while shopping, and perhaps the connection between Hobbies & Leisure and Health indicates that people's hobbies are detrimental to their health. More likely is that the connections here reflect the effects of omitted variables, but it is at least

Table 2: Trends VAR coefficients, 2/2

	Dependent variable:		
	jobs_education	shopping	travel
business_industrial.l1	−0.039	−0.343**	−0.256**
	(0.097)	(0.132)	(0.112)
finance.l1	−0.067	0.100*	0.010
	(0.043)	(0.059)	(0.050)
health.l1	−0.084	0.091	0.069
	(0.077)	(0.104)	(0.088)
hobbies_leisure.l1	0.074	0.081	0.201**
	(0.079)	(0.107)	(0.091)
jobs_education.l1	0.274***	0.251**	−0.124
	(0.089)	(0.122)	(0.103)
shopping.l1	0.180***	0.558***	0.095
	(0.059)	(0.081)	(0.069)
travel.l1	0.070	−0.019	0.624***
	(0.063)	(0.085)	(0.072)
const	−0.046	0.040	−0.022
	(0.115)	(0.157)	(0.133)
Observations	190	190	190
Note:			*p<0.1; **p<0.05; ***p<0.01

interesting to see the correlations.

4 Relation to macroeconomic variables

Linear interdependencies between the Google Trends variables are interesting, but it is illuminating to examine the relation between these variables and external economic variables. If the Google Trends data have predictive power for economic trends, they may be useful for forecasting future economic conditions. It is reasonable to suppose that the trend variables discussed in the previous section are related to patterns of consumption in the U.S. Following Della Penna and Huang (2009), I chose to examine the relationship between my trend variables and changes in personal consumption.

To do this, I downloaded monthly data on seasonally-adjusted percentage changes in real personal consumption from FRED[2]. I used a DF-GLS test to determine that this time series is stationary. I then removed the seasonal components from each of the Google Trends time series and estimated a new VAR model for all variables, based on the following specification:

$$\begin{aligned}[\text{variable}]_t = \beta_0 &+ \beta_1 \text{business\_industrial}_{t-1} + \beta_2 \text{finance}_{t-1} + \beta_3 \text{health}_{t-1} \\ &+ \beta_4 \text{hobbies\_leisure}_{t-1} + \beta_5 \text{jobs\_education}_{t-1} + \beta_6 \text{shopping}_{t-1} \\ &+ \beta_7 \text{travel}_{t-1} + \beta_8 \text{pct\_change\_consumption}_{t-1}\end{aligned}$$

A single lag in all variables was again determined to be preferred based on the Akaike, Hannan-Quinn, Schwarz, and final prediction error criteria. The estimated coefficients are shown in tables 3 and 4.

Much of the information in tables 3 and 4 is similar to what is presented in tables 1 and 2, although there are some interesting new connections that emerge. For example, the coefficient for the lag of percent change in consumption on the Business & Industrial variable is highly statistically significant and is greater than the coefficient for the lag of Business & Industrial on itself, suggesting that months with higher increases in consumption tend to precede those with elevated interest in business topics. Changes in consumption also seem to have some power in predicting interest in travel.

As far as the determinants of changes in consumption go, by far the largest effect is due to the lagged value of the changes in consumption series, which is not surprising. Interestingly, though, the coefficient of the lagged Finance variable is highly statistically significant, suggesting that interest in finance as indicated by Google searches has some predictive power for future changes in consumption. The Jobs & Education variable also appears to be significantly related to consumption.

[2] `https://fred.stlouisfed.org/series/DPCERAM1M225NBEA`

Table 3: Consumption VAR coefficients, 1/2

	Dependent variable:			
	business_industrial	finance	health	hobbies_leisure
business_industrial.l1	0.266***	−0.178	−0.011	−0.149
	(0.086)	(0.160)	(0.099)	(0.108)
finance.l1	−0.012	0.523***	−0.043	−0.024
	(0.039)	(0.072)	(0.045)	(0.049)
health.l1	−0.127*	0.044	0.440***	0.236***
	(0.069)	(0.127)	(0.079)	(0.086)
hobbies_leisure.l1	0.144**	−0.131	0.234***	0.220**
	(0.070)	(0.129)	(0.080)	(0.087)
jobs_education.l1	−0.008	−0.280*	−0.006	0.251**
	(0.079)	(0.147)	(0.091)	(0.099)
shopping.l1	0.115**	0.426***	0.124**	0.056
	(0.053)	(0.098)	(0.061)	(0.066)
travel.l1	0.156***	0.104	−0.022	0.066
	(0.056)	(0.103)	(0.064)	(0.070)
pct_change_consumption.l1	0.937***	−0.526	−0.329	−0.075
	(0.359)	(0.665)	(0.413)	(0.450)
const	−0.204*	0.047	0.022	0.060
	(0.120)	(0.222)	(0.138)	(0.150)
Observations	189	189	189	189
Note:				*p<0.1; **p<0.05; ***p<0.01

4.1 Granger causality tests

To get a sense of the power of the Google Trends variables for forecasting changes in consumption I ran a Granger causality test based on the above VAR model (H_0 : Trends variables do not Granger-cause changes in consumption). The result was a very strong rejection of the null hypothesis ($p < 10^{-4}$), indicating that the lagged Trends variables contain statistically significant information about the coming month's consumption that is not found in the first lag of the consumption time series.

Given that the Finance variable seems to contain much of the information about the consumption changes not found in the consumption series itself, I wanted to get an idea of

Table 4: Consumption VAR coefficients, 2/2

	Dependent variable:			
	jobs_education	shopping	travel	pct_change_consumption
business_industrial.l1	−0.047	−0.330**	−0.241**	−0.016
	(0.093)	(0.128)	(0.108)	(0.017)
finance.l1	−0.071*	0.109*	0.028	−0.024***
	(0.042)	(0.057)	(0.049)	(0.008)
health.l1	−0.062	0.061	0.072	0.006
	(0.074)	(0.101)	(0.086)	(0.014)
hobbies_leisure.l1	0.072	0.077	0.192**	0.015
	(0.075)	(0.103)	(0.087)	(0.014)
jobs_education.l1	0.258***	0.269**	−0.132	0.035**
	(0.086)	(0.117)	(0.099)	(0.016)
shopping.l1	0.182***	0.557***	0.088	−0.009
	(0.057)	(0.078)	(0.066)	(0.011)
travel.l1	0.071	−0.022	0.610***	0.016
	(0.060)	(0.082)	(0.070)	(0.011)
pct_change_consumption.l1	−0.154	0.451	0.787*	−0.161**
	(0.389)	(0.531)	(0.450)	(0.073)
const	−0.041	−0.015	−0.175	0.205***
	(0.130)	(0.177)	(0.150)	(0.024)
Observations	189	189	189	189

Note: *p<0.1; **p<0.05; ***p<0.01

how significant that relationship is in a model containing only the Finance and consumption time series. I estimated a VAR for the two variables with three lags (selected based again upon various information criteria). The coefficients for this VAR are shown in table 5.

Based solely upon that estimation, it is rather clear that the Finance variable Granger-causes changes in consumption (and changes in consumption do not Granger-cause Finance). A formal Granger causality test backs this up ($p < 0.01$). This is an interesting result, as it suggests that a single Google Trends variable contains enough information to significantly improve consumption forecasts.

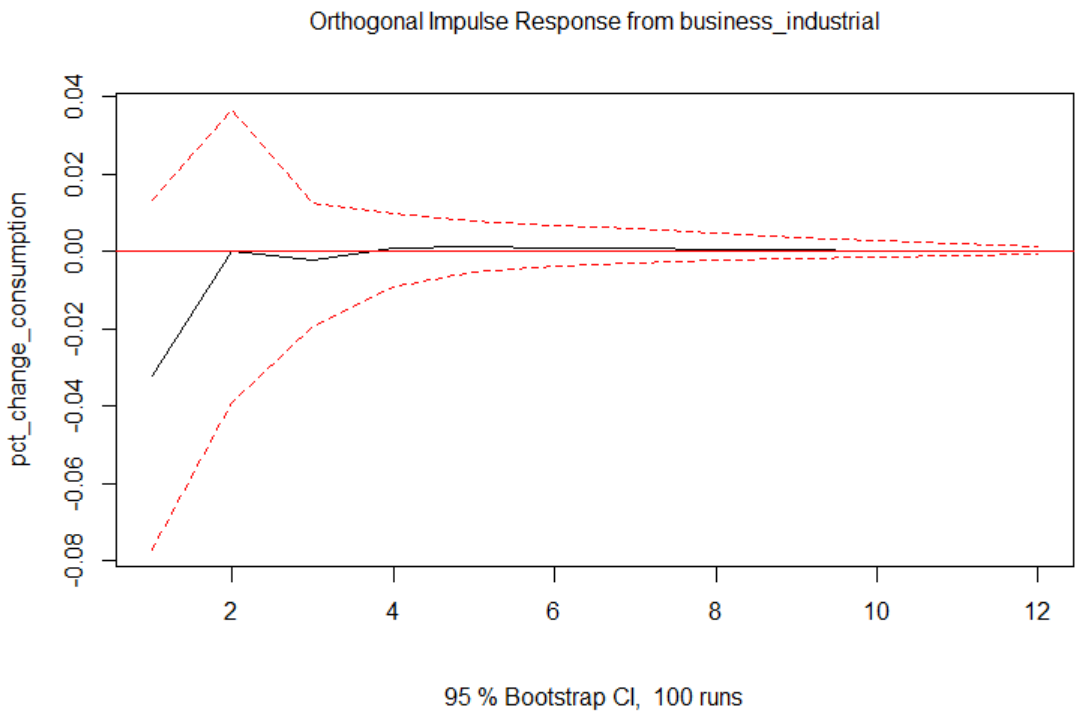
Table 5: Finance/Consumption VAR coefficients

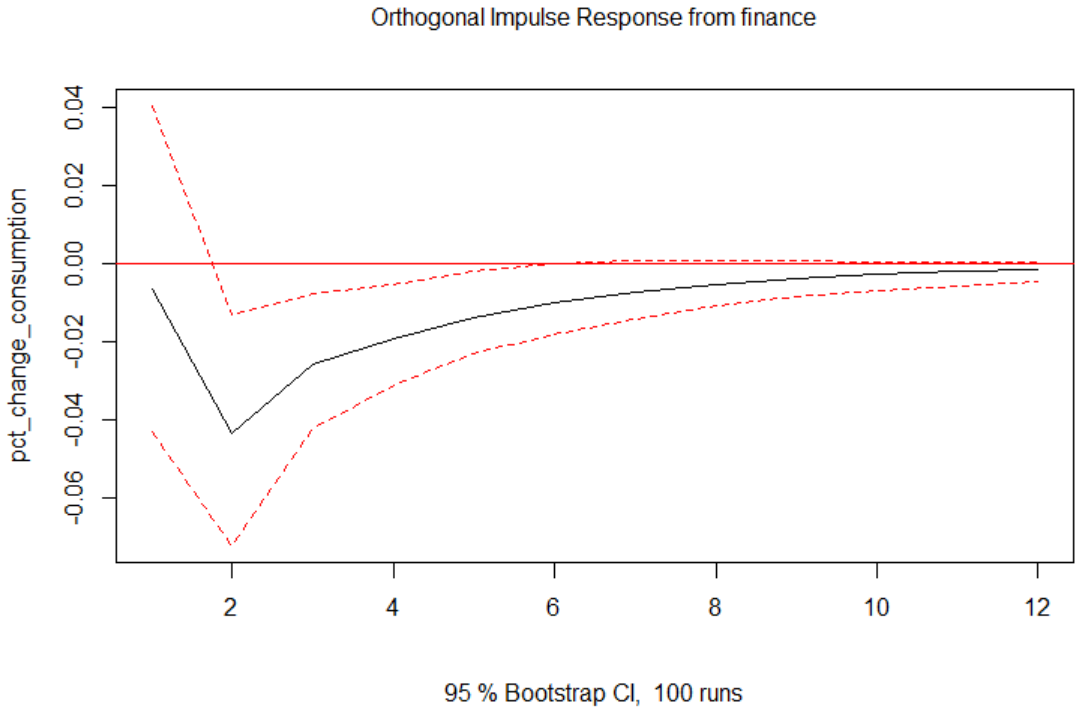
	Dependent variable:	
	finance	pct_change_consumption
finance.l1	0.367***	−0.006
	(0.072)	(0.008)
pct_change_consumption.l1	0.475	−0.126*
	(0.660)	(0.074)
finance.l2	0.176**	−0.021**
	(0.078)	(0.009)
pct_change_consumption.l2	0.249	0.121
	(0.654)	(0.073)
finance.l3	0.286***	0.009
	(0.075)	(0.008)
pct_change_consumption.l3	0.365	0.061
	(0.650)	(0.073)
const	−0.111	0.163***
	(0.279)	(0.031)
Observations	187	187
Note:		*p<0.1; **p<0.05; ***p<0.01

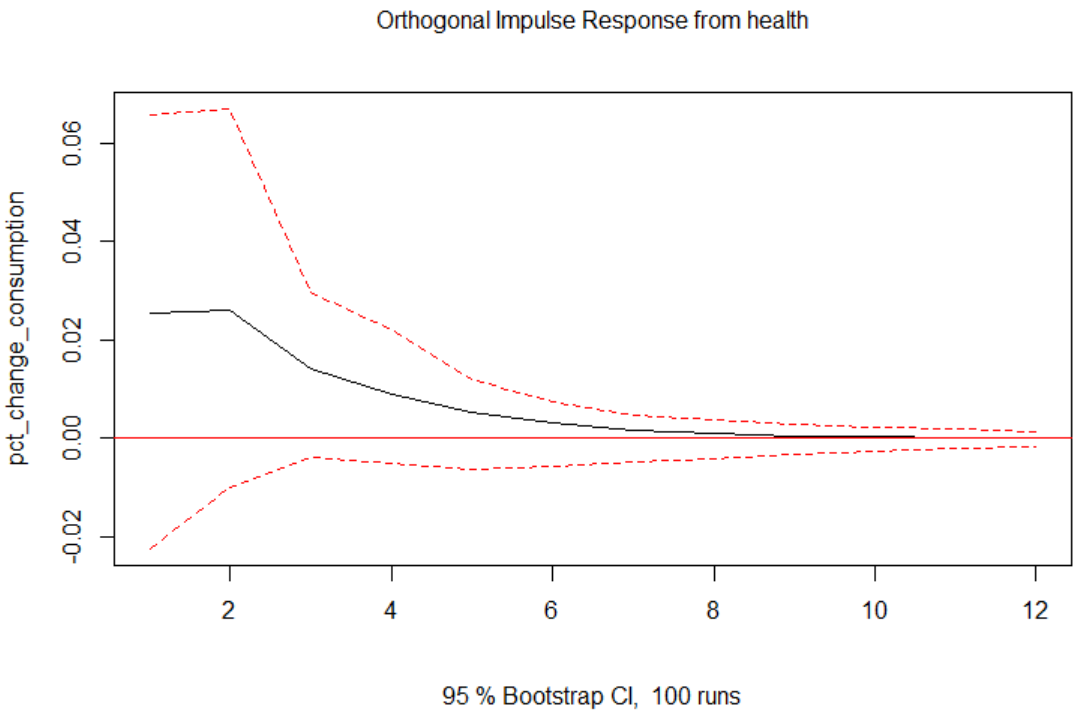
4.2 Impulse response functions

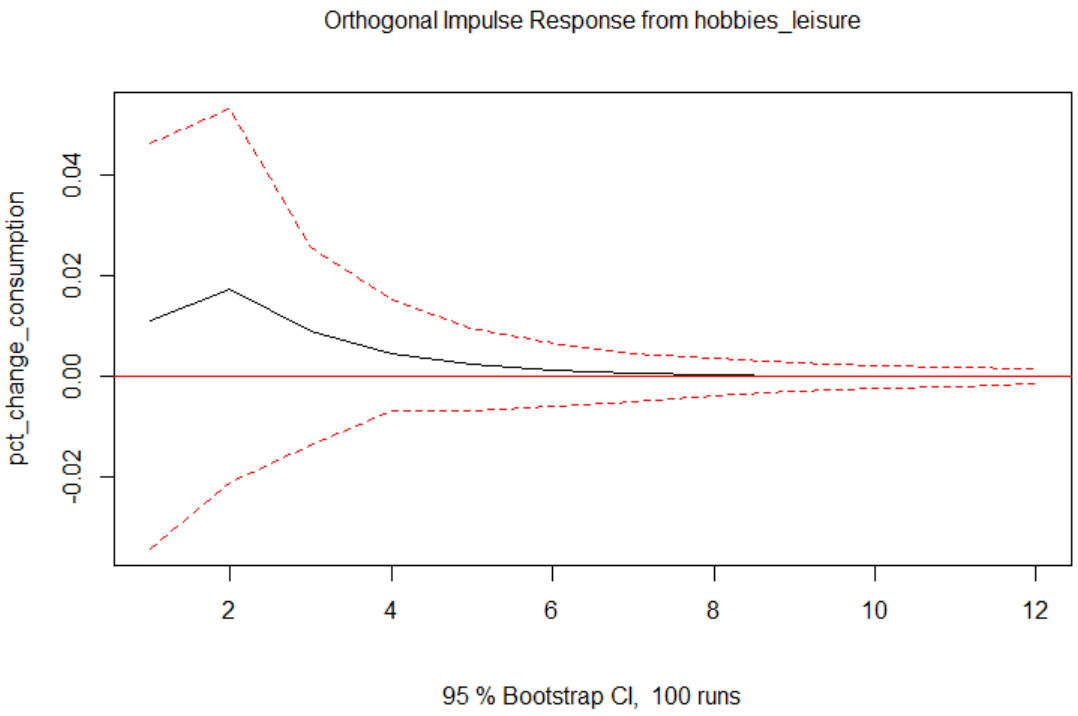
To further explore the relationships between variables expressed in the two VAR models specified above, I estimated impulse response functions to show the effects of shocks in the variables over time.

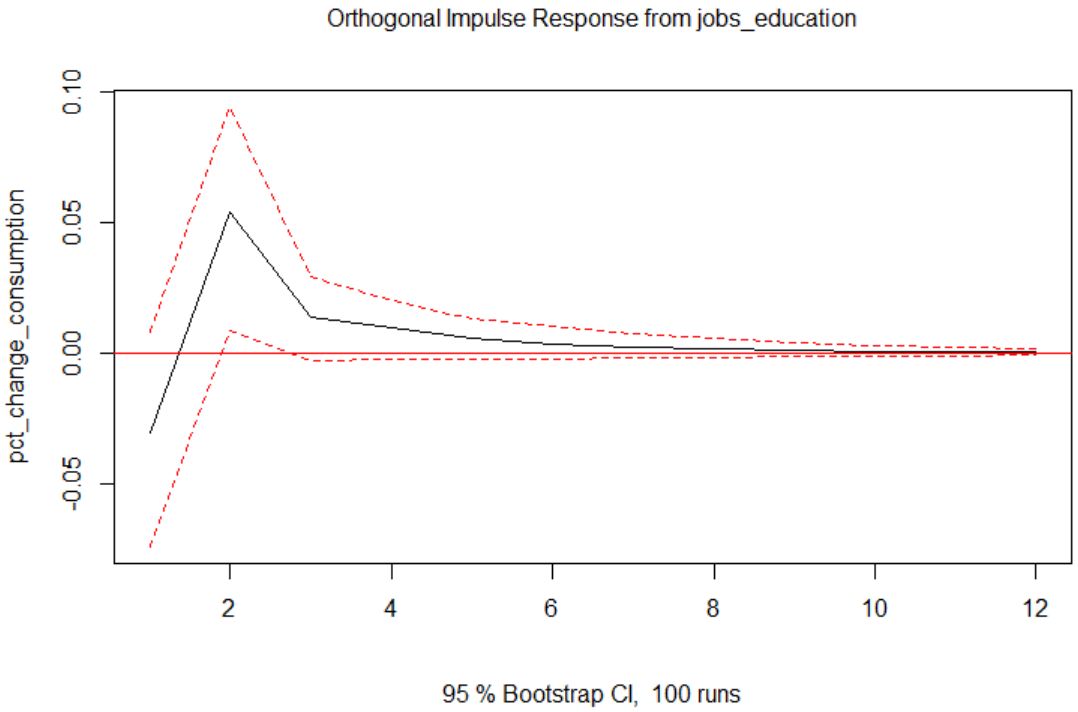
The impulse response functions showing the responses for the change in consumption tell much the same story as the coefficients in table 4, though they have the benefit of showing the estimated long-run effects on the changes in consumption. For example, a shock in the Jobs & Education series appears to possibly decrease consumption at first then be associated with increased consumption growth over the next few months. The effect of a shock in the Finance series appears to be negligible at first and then becomes most pronounced after two or three months, with an effect still present at the one year mark.

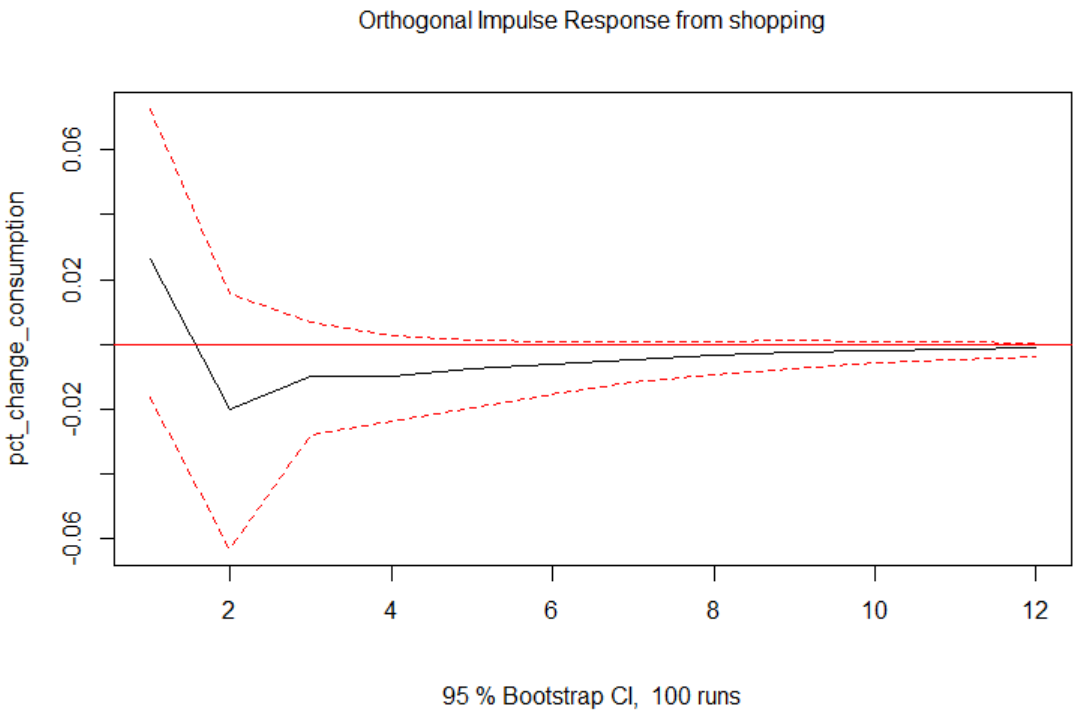

Orthogonal Impulse Response from business_industrial
pct_change_consumption
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs

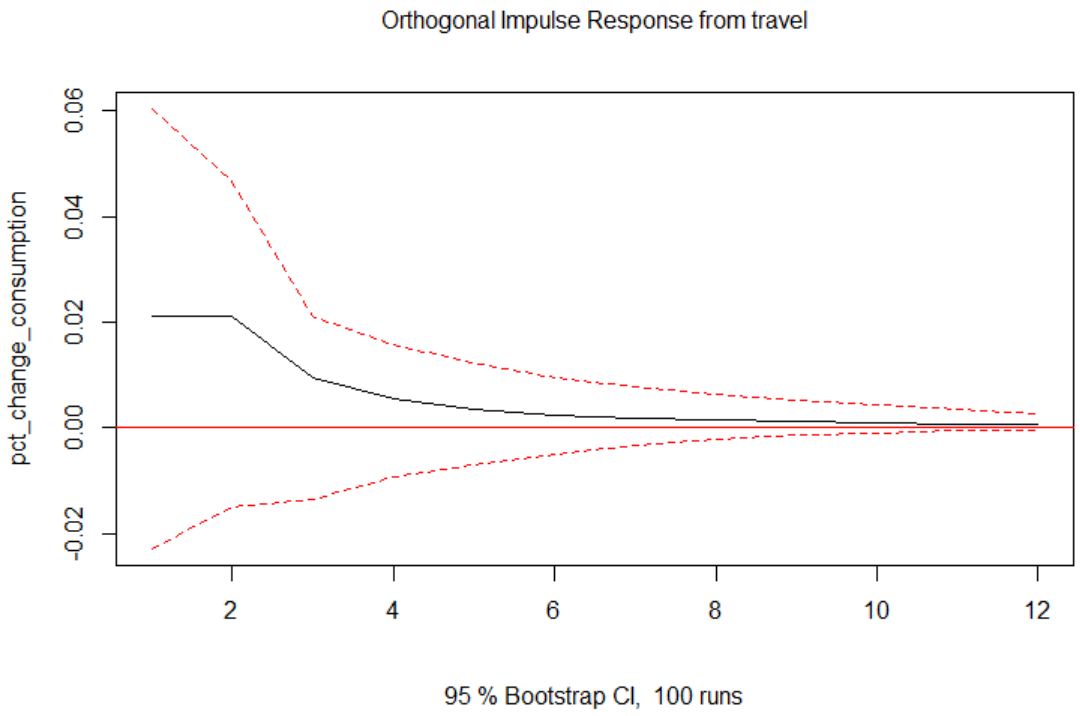

Orthogonal Impulse Response from finance
pct_change_consumption
0.04
0.02
0.00
-0.02
-0.04
-0.06
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs


Orthogonal Impulse Response from health
pct_change_consumption
0.06
0.04
0.02
0.00
-0.02
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs


Orthogonal Impulse Response from hobbies_leisure
pct_change_consumption
0.04
0.02
0.00
-0.02
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs


Orthogonal Impulse Response from jobs_education
pct_change_consumption
0.10
0.05
0.00
-0.05
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs

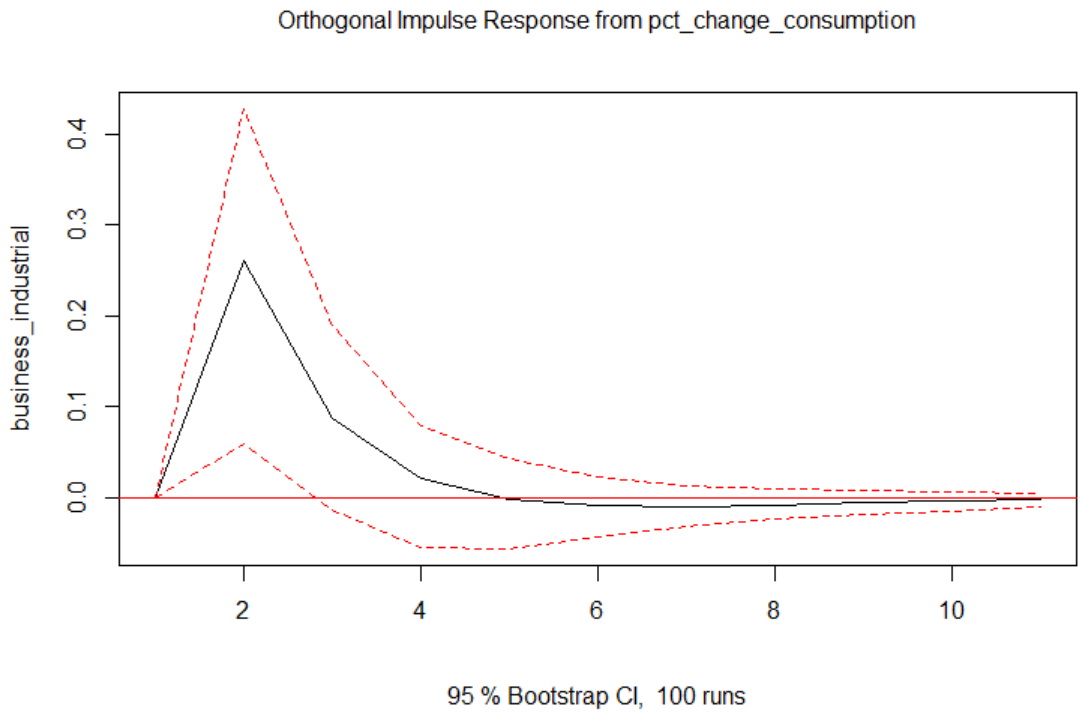

Orthogonal Impulse Response from shopping
pct_change_consumption
0.06
0.02
-0.02
-0.06
2
4
6
8
10
12
95 % Bootstrap CI, 100 runs



To get an idea of the impact of changes in consumption on the Business & Industrial series, I also estimated the impulse response function for change in consumption on Business & Industrial. There appears to be a large spike in the second lag, but the disturbance fades away fairly quickly.

5 Summary & conclusion

The Google Trends variables I selected here have some significant linear interdependencies, as shown by the estimated VAR in first lags. Much of the dynamics are due to the seasonality of the variables, but statistically significant relationships between variables persist even after removing the seasonal components. Further work is needed to determine the root of the interdependencies observed between the variables.



The Google Trends variables are shown to have statistically significant power for predicting changes in personal consumption in data from the U.S. This agrees with earlier results from Della Penna and Huang (2009) as well as Vosen and Schmidt (2011), showing similar predictive power in older data. The impulse response functions I estimated here shed some additional light on the predicted effect of shocks in the Trends variables upon the rate of change of consumption. The Trends variable with the greatest measured impact on personal consumption is the Finance category, which shows statistically significant predictive power for changes in consumption.

References

[1] Choi, H., & Varian, H. (2009). "Predicting the Present with Google Trends."

[2] Della Penna, N., & Huang, H. (2009). "Constructing consumer Sentiment Index for U.S. Internet Search Patterns," *University of Alberta Working Paper Series*, 2009-26.

[3] Gentzkow M., Kelly, B., & Taddy, M. (2017). "Text as Data," NBER Working Paper No. 23276.

[4] Mao, H., Counts, S., & Bollen, J. (2011). "Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data."

[5] Vosen, S., & Schmidt, T. (2011). "Forecasting private consumption: survey-based indicators vs. Google trends." *Journal of Forecasting*, 30, pp. 565-578.