ISSN 1178-2293 (Online)

University of Otago
Economics Discussion Papers
No. 1306

March 2013

Awareness of Sexually Transmitted Disease and Economic Malady: A First Look Using Search Engine Query Data

Dr Dan Farhat
Department of Economics
University of Otago

Address for correspondence:

Dr Dan Farhat
Department of Economics
University of Otago
PO Box 56, Dunedin
NEW ZEALAND
Email: dan.farhat@otago.ac.nz
Telephone: 64 3 479 8645

Awareness of Sexually Transmitted Disease and Economic Malady:
A First Look Using Search Engine Query Data

Dr. Dan Farhat
University of Otago
Department of Economics
PO Box 56
Dunedin, New Zealand 9054
dan.farhat@otago.ac.nz

[Draft: March 2013 – Cite with Caution]

Abstract

Using search engine query data as a measure for public awareness of sexual health outcomes, this study extracts a measure of general interest in sexually transmitted disease for the United States (2004 – 2012). This trend is compared to a measure of overall economic prosperity. Heightened interest in STDs is shown to have occurred during the recent recession (December 2007 – June 2009). It is hypothesized that potential changes in insurance coverage as workers become unemployed may increase instances of online self-diagnosis of STDs. Select data imply that an increase in the tendency to search for STD information online occurs after alternative insurance options are explored. Data also imply that interest in behavioral alternatives to treating STDs rises after interest in STDs rise (and not at the onset of the economic slump). This paper identifies broad trends and points to using search engine query data to connect health awareness to the state of the economy as a lucrative area of future research.

Introduction

The internet provides greater access to health-related information. Patients can learn about ailments and treatment options, self-diagnose their symptoms, and find resources for specialist medical care all online. These activities can reduce treatment costs for the patient and may increase the quality of healthcare they receive.[1] The ability download information about sexually transmitted diseases [STDs] is particularly beneficial. Because STDs are associated with risky sexual behavior, patients with symptoms may postpone seeking treatment to avoid embarrassment or social stigma. As they delay, their health worsens and treatment becomes more extensive. Further, those unaware of STD symptoms have a greater chance of contracting the disease from a partner and transmitting the disease to others. Mass dissemination of information about treatment, symptoms and prevention can reduce these outcomes.

We now have to ability to measure trends in digital STD awareness. Google collects data measuring the relative volume of keyword searches used on the Google search engine. This data is high-frequency (weekly/monthly), has global coverage (at the country, state/province and (in some cases) city levels) and is readily available (through a tool known as *Google Trends*).[2] The data has already been shown to be particularly useful in predicting epidemiological events (such as flu[3] and HIV infection[4]). Although this data cannot distinguish between those searching for specific information

[1] See Ryan & Wilson (2008) and Lanseng & Andreassen (2007) for descriptions and insights on the costs and benefits of online health information.
[2] Available at www.google.com/trends/. Additional information about Google Trends is available at support.google.com/trends/.
[3] See Ginsberg et al. (2008), Polgreen et al. (2008) and Carneiro & Mylonakis (2009).
[4] See Jena et al. (2013).

out of necessity (for example, to self-diagnose a condition) and those searching out of curiosity, it does provide a measure of general interest.[5]

This study attempts to identify a single, underlying trend representing general interest in sexually transmitted disease. Once identified, the trend is discussed in the context of recent economic performance. While interesting implications emerge, the analysis that follows is merely a first-step in a potentially lucrative research applying culturomic[6] data to the analysis of sexual behavior.

Data and Analysis

Interest in Sexually Transmitted Disease

We can identify a single, underlying trend representing general awareness of sexually transmitted disease using keyword search data from Google Trends for 9 STD keywords (chlamydia, genital warts, gonorrhea, hepatitis, herpes, HIV, HPV, syphilis and trichomoniasis).[7] For each keyword, Google Trends provides an estimate of the number of searches for that term relative to the total number of searches done on the Google search engine for each week (in some cases, month) from January 2004 to the present. The data is rescaled so the period with the largest relative search volume is indexed to 100. Weekly search volume data (January 2004 to January 2013) is collected and averaged into monthly frequency (see Figure 1).

[FIGURE 1 HERE]

A measure for general interest in STDs can be constructed computationally by identifying a sequence, $S = \{s_{2004m1}, s_{2004m2}, ..., s_{2013m2}\}$, that maximizes the summed correlation (in absolute value) between itself and each of the 9 other series displayed in Figure 1. [8] Each element of the sequence, S, is scaled to measure a proportion of the largest element of S (to match the Google Trends data). The contemporaneous correlation between the sequence, S, and each of the individual STD queries is shown in Table 1 to be strong[9] for the entire sample period. When the sample is divided into three sub-periods (Early: 2004m1 – 2007m12; Middle: 2008m1 – 2009m5; Late: 2009m6 – 2013m2), the relationship between S and the STD series remains strong (particularly for chlamydia, gonorrhea, herpes, HIV, syphilis and trichomoniasis). Although the correlation between S and three of the STD series weakens[10] slightly when the sample is divided, it can be argued that S is a reasonable approximation for overall STD awareness online.

[5] See Ripberger (2011) for a discussion.
[6] The study of culture through empirical measurement.
[7] These searchers are limited to the 'Health' category in Google Trends.
[8] MATLAB, a program designed for numerical analyses, is used. MATLAB computes S to minimize Denoting the search series in Figure 1 as X_i, i = 1...9, MATLAB computes S to minimize a function $F = -(\Sigma_i |\rho(S,X_i)|)$ where $\rho(S,X_i)$ is the sample correlation between S and X(i). Note that each of the 9 series is assumed to be equally important in the derivation of S and that either positive or negative correlations between S and X_i are deemed relevant. There are many more sophisticated methods for identifying common trends in time series data. More elaborate techniques are left for future research.
[9] Non-zero as measured by the 2-standard error (95% confidence) bounds computed as $2/\sqrt{T}$.
[10] The correlation between S and hepatitis becomes insignificant in the later periods, while the correlation between S and genital warts is high only in the early and late periods. The correlation between S and HPV switches from significantly negative in the early period to significantly positive in the later period.

[FIGURE 2 HERE]

[TABLE 1 HERE]

Alone, S has several salient properties. Table 2 (column 1) shows that interest in this period (s_t) is strongly correlated with interest in the previous period (s_{t-1}), indicating that surges in interest tend to carry over into the future (i.e. S exhibits persistence). Table 2 (columns 2 – 3) shows that STD interest index exhibits two monthly cycles per year: (1) a small peak in November followed by a trough in December and (2) a large peak in April followed by a deep trough in August. Overall, S trended downward from 2004 to 2007, followed by an upswing from 2008 to the present. The turning point in S coincides with the onset of the most recent economic recession (dated December 2007 by the National Bureau of Economic Research [NBER] (2013)).

[TABLE 2 HERE]

The State of the Economy and Interest in Sexual Health

Since economic data is readily available at the monthly frequency, we can focus on the relationship between S and the general state of the economy (although a variety of social and cultural factors are relevant to explaining trends in STD awareness). To do this, we can create a sequence, E = {e_{2004m1}, e_{2004m2}, ..., $e_{2012m12}$}, in the same manner as S to measures overall economic activity. E is created using data on labor market outcomes (unemployment rates, average weekly labor hours, hires, quits, and job openings), financial status (average weekly earnings, the growth rate of average weekly earnings, stock market activity as measured by the S&P 500 Index and the Dow Jones Industrial Index, and a measure of financial stress) and spending (CPI inflation and retail sales).[11] Figure 3 shows E and Table 3 describes the relationship between E and the 12 series used to derive it. E is positively correlated with the 'bads' (unemployment and financial stress) and negatively correlated with the 'goods' (quit rates, job opening rates, hiring rates, average weekly working hours, the growth rate of average earnings and the stock market indices) which suggests that E measures 'economic misfortune'.[12] As expected, E rises sharply from December 2007 to June 2009: the recession as dated by the NBER (2013).

[FIGURE 3 HERE]

[TABLE 3 HERE]

[11] Data for average weekly earnings and average weekly labor hours are for production and non-supervisory employees. Monthly data on CPI inflation, unemployment, average weekly earnings, percent change in average weekly earnings, average weekly labor hours, hires, quits, and job openings are collected from the United States Bureau of Labor Statistics [BLS]. Estimates of retail sales are provided by the United States Census Bureau. Indicators of stock market activity (S&P 500 and the Dow Jones Industrial Index) and financial stress are obtained from the St. Louis Federal Reserve Bank FRED Database. None of the data is seasonally adjusted to be consistent with the data from Google Trends.

[12] Three anomalies appear. E is positively correlated with total average hourly earnings, perhaps due to overall persistent growth in earnings coupled with the sharp rise in E from 2007 to 2009. E is negatively correlated with CPI inflation. While large amounts of inflation is undesirable, low and stable inflation (which characterizes the US economy from 2004 – 2012) is not overtly harmful. The relationship between E and total retail sales is positive but not significant at the 95% level for the entire sample period. However, if we divide the sample, a negative relationship appears.

After removing the seasonal components of S and E, we can compare their movements over time (see Figure 4). When the economy performs well, as it did from 2004 to 2007, general interest in STDs falls. When the recession hit at the end of 2007, interest in STDs began to rise. Although the economy began to recover in mid-2009, attention paid to STDs has continued to increase. This pattern is consistent with a substantial lag in S behind E (i.e. STD awareness continues to be affected by shocks to the economy well after the economy stabilizes).

[FIGURE 4 HERE]

One potential reason why this relationship may occur is due to the impact of economic misfortune on insurance coverage. A reduction in insurance coverage during economic downturns increases the public's out-of-pocket medical costs. As a result, people may self-diagnose online more often to avoid expensive doctor's visits. While insurance coverage data is not available at the monthly frequency, we can utilize Google Trends data once again to measure interest in insurance alternatives. Figure 5 compares the series S with seasonally adjusted keyword search trends for 'cheap insurance' and 'free clinic'. Interest in free clinics rises perpetually over the entire sample, suggesting no clear pattern between it and STD interest. Search for cheap insurance, however, exhibits the same trend as S but is leading: searches for 'cheap insurance' reach a low in May 2006 whereas S does not bottom out until February 2007. This pattern supports the notion that self-diagnosis may occur in greater frequency *after* insurance alternatives are explored.

[FIGURE 5 HERE]

To avoid paying more for medical care, people may take preventative measures to improve their health. In the case of STDs, this behavior is ideal for medical practitioners and policymakers (who would that people avoid contracting STDs in the first place over treating them after the fact). Google Trends data can be used to measure awareness of safer sexual behaviors over time. Figure 6 shows the relationship between the STD measure S and two methods for reducing the probability of contracting an STD: 'abstinence' and 'condoms'. Interest in abstinence has fallen consistently since 2004 (perhaps an interesting topic of study in itself), producing no clear pattern between it and STD interest. Searches for condoms, however, exhibit the same trend as S but is lagging: the S measure reaches a low in February 2007 and rises thereafter, but interest in condoms does not reach a minimum until July 2008 before rising. It appears that people do look at STD prevention during periods of recession after STD related information accessed online has increased.

[FIGURE 6 HERE]

Final Remarks

Using search engine query data to measure patterns in public awareness of sexual health can be illuminating. Much is left for future work. The connection between *interest* in STDs and the rate at which STDs are *actually* contracted merits elaboration. Although Jena et al. (2013) do provide some evidence that search query data can predict rates of HIV infection, the data supplied by Google Trends cannot distinguish between merely curious seekers from those self-diagnosing. Further, the connection between the state of the economy and search for sexual health information (and *general*

health information) online deserves a more precise analysis than the one provided above. The social and cultural factors contributing to the search for STD knowledge online are non-trivial and should also be explored. Nonetheless, using search engine query data to monitor public interest in health is a growing area of research. Linking this interest to economic, social and cultural trends will likely provide further insights into how internet users can become healthier people.

References

Bureau of Labor Statistics [BLS] (2013a). Consumer price index. Available at www.bls.gov.

Bureau of Labor Statistics [BLS] (2013b). Current employment statistics survey. Available at www.bls.gov.

Bureau of Labor Statistics [BLS] (2013c). Current population survey. Available at www.bls.gov.

Bureau of Labor Statistics [BLS] (2013d). Job openings and labor turnover survey. Available at www.bls.gov.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10), 1557-1564.

Federal Reserve Bank of St. Louis (2013). Federal reserve economic data [FRED]. Available at research.stlouisfed.org/fred2.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Google Trends (2013). Available at google.com/trends/.

Jena, A. B., Karaca-Mandic, P., Weaver, L., & Seabury, S. A. (2013). Predicting New Diagnoses of HIV Infection Using Internet Search Engine Data. *Clinical infectious diseases*, available at cid.oxfordjournals.org/content/early/2013/02/07/cid.cit022.short.

Lanseng, E. J., & Andreassen, T. W. (2007). Electronic healthcare: a study of people's readiness and attitude toward performing self-diagnosis. *International Journal of Service Industry Management*, 18(4), 394-417.

National Bureau of Economic Research [NBER] (2013). US business cycle expansions and contractions. Available at www.nber.org/cycles.html.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., & Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11), 1443-1448.

Ripberger, J. T. (2011). Capturing curiosity: Using Internet search trends to measure public attentiveness. *Policy Studies Journal*, 39(2), 239-259.

Ryan, A., & Wilson, S. (2008). Internet healthcare: do self-diagnosis sites do more harm than good?. *Expert Opinion on Drug Safety*, 7(3), 227-229.

United States Census Bureau (2013). Monthly and annual retail trade. Available at www.census.gov/retail.

Tables and Figures

Figure 1 – Google Trends Search Index for 9 STDs, Monthly, January 2004 – January 2013

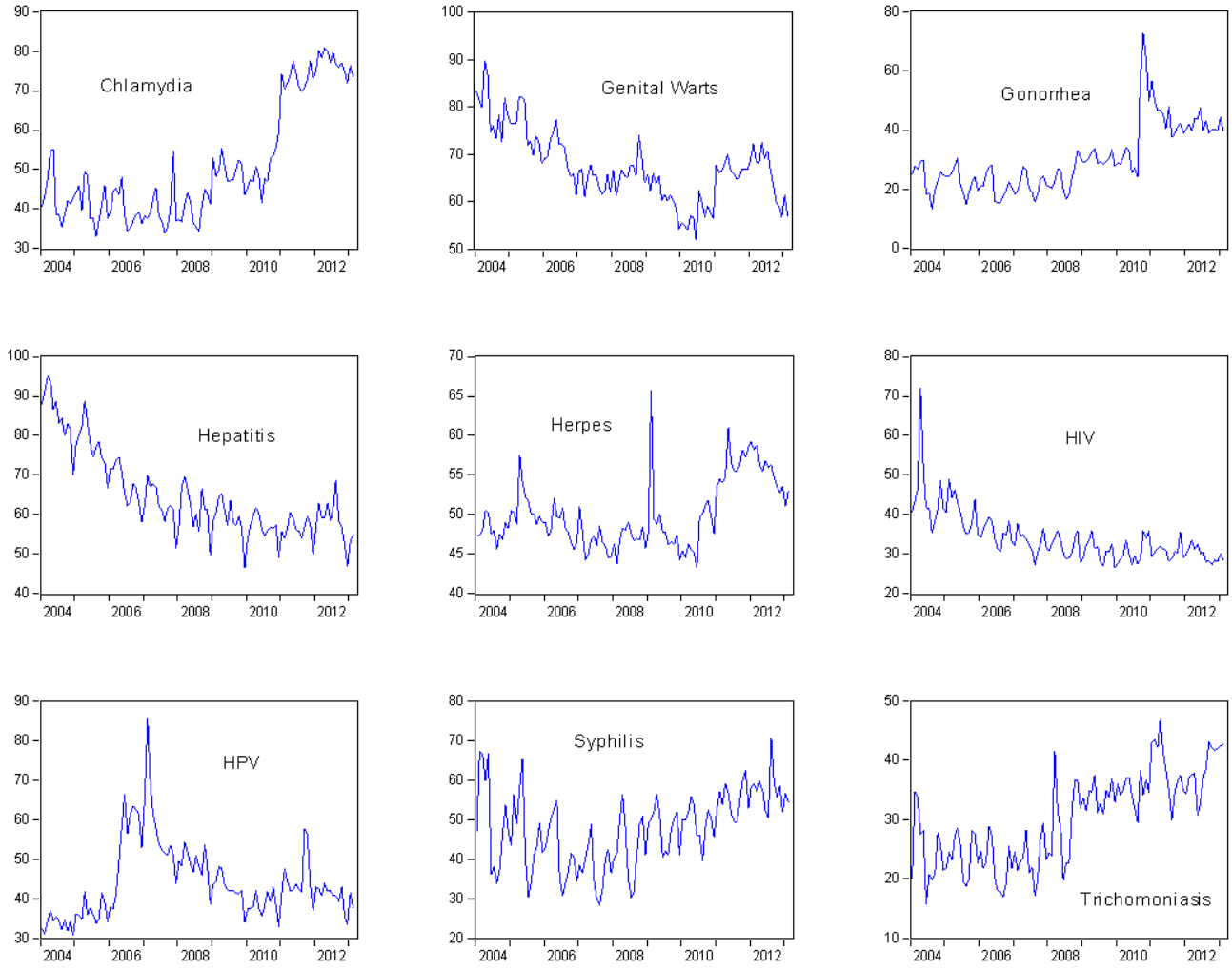


Figure 2 – Common STD Interest Trend (S): Monthly, January 2004 – January 2013

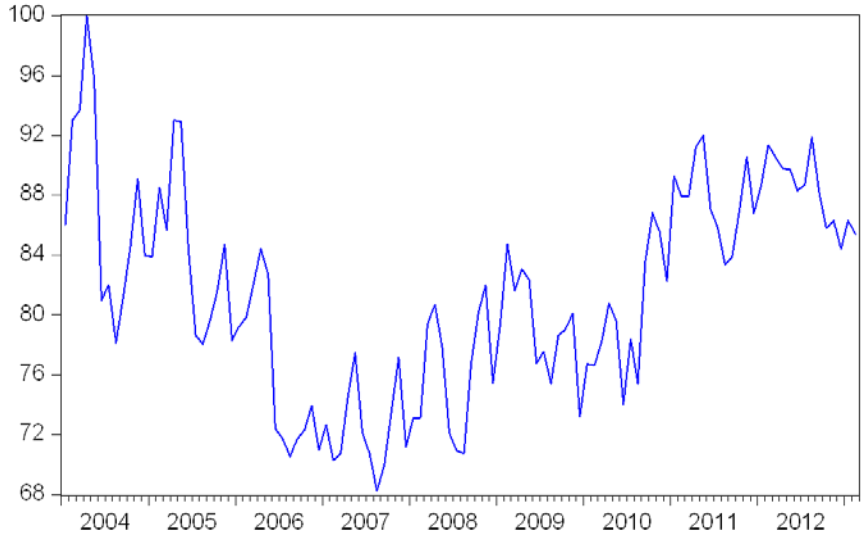


Table 1 – Correlation: Common Trend (S) and Google Trends Search Indices for 9 STDs[1]

	Full	Early	Mid	Late
Correlation with Common STD Trend (S)	2004m1 - 2013m2	2004m1 - 2007m12	2008m1 - 2009m5	2009m6 - 2013m2
Chlamydia	0.684	0.721	0.864	0.920
Genital warts	0.453	0.901	-0.01	0.779
Gonorrhea	0.610	0.717	0.860	0.670
Hepatitis	0.302	0.866	0.386	0.282
Herpes	0.656	0.511	0.531	0.882
HIV	0.419	0.869	0.641	0.380
HPV	-0.515	-0.686	-0.061	0.302
Syphilis	0.855	0.867	0.922	0.798
Trichomoniasis	0.544	0.611	0.818	0.559
95% Confidence Bands	±0.191	±0.289	±0.485	±0.298

[1] Google Trends (2013).

Table 2 – OLS Regressions, Full sample (2004m1 – 2013m2)

Dependent variable:	s_t		
	(1)	(2)	(3)
constant	12.548***	—	—
s_{t-1}	0.846***	—	0.922***
January	—	81.543***	8.673***
February	—	83.097***	7.940**
March	—	83.339***	6.988**
April	—	86.405***	9.592***
May	—	85.599***	5.961*
June	—	78.689***	-0.206
July	—	78.292***	5.766*
August	—	76.880***	4.719
September	—	79.266***	8.408***
October	—	81.209***	8.151**
November	—	83.284***	8.434**
December	—	78.522***	1.761
R^2	0.716	0.177	0.879

* Significant at the 10% level
**Significant at the 5% level
*** Significant at the 1% level

Figure 3 –Common Economic Trend (E): Monthly, January 2004 – December 2012

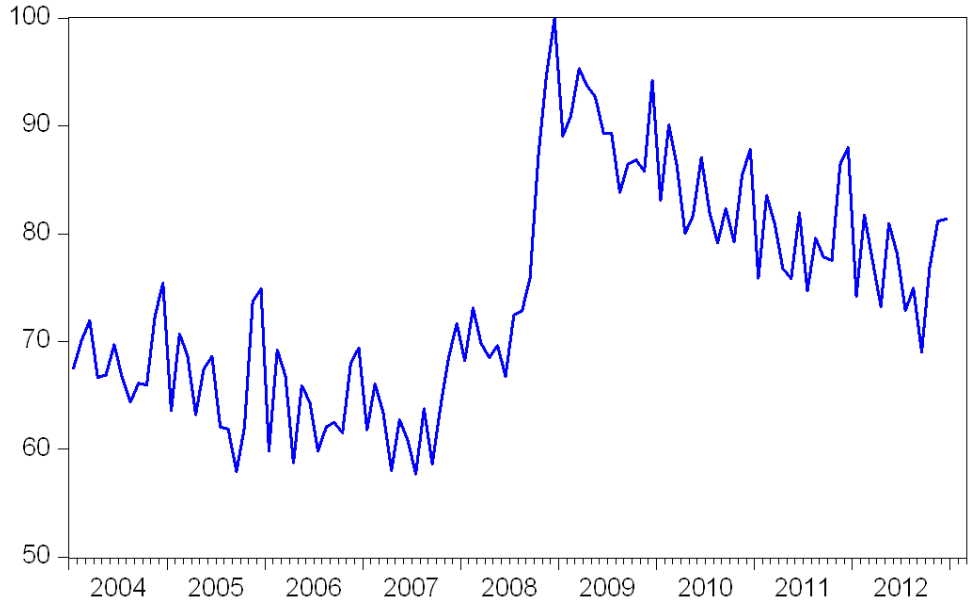
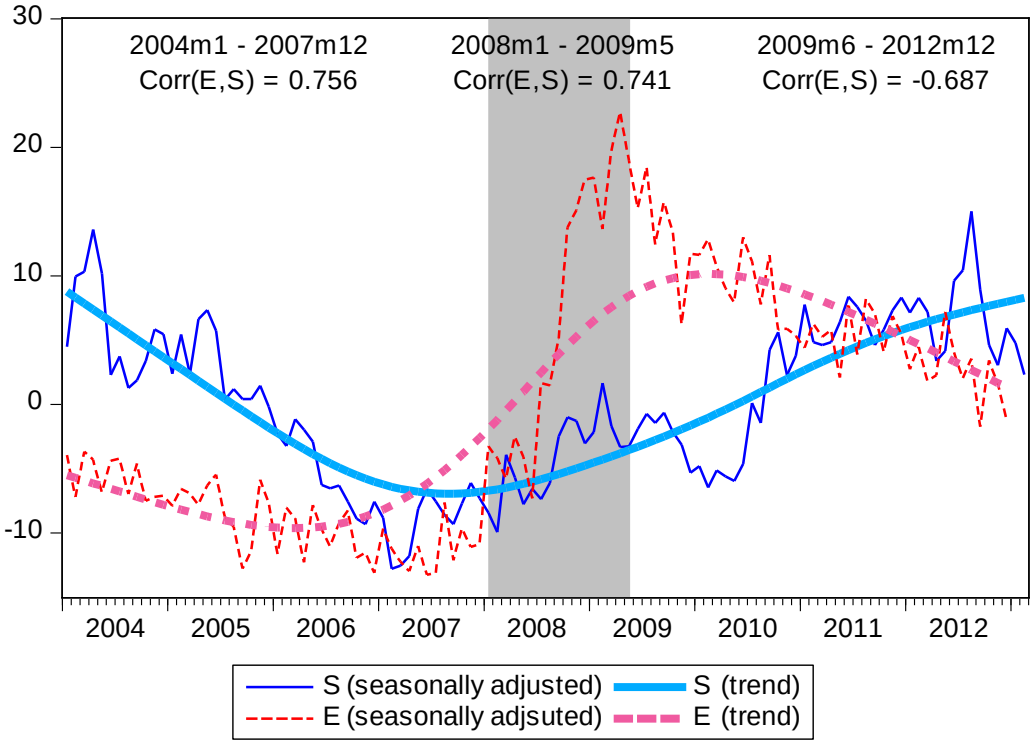


Table 3– Correlation: Common Economic Trend (E) and 12 Economic Variables

	Full	Early	Mid	Late
Correlation with Common Economic Trend (E)	2004m1 - 2012m12	2004m1 - 2007m12	2008m1 - 2009m5	2009m6 - 2012m12
Unemployment Rate[1]	0.809	0.365	0.815	0.529
Quits[2]	-0.834	-0.704	-0.851	-0.626
Job Openings[2]	-0.907	-0.826	-0.907	-0.842
Hires[2]	-0.769	-0.641	-0.794	-0.547
Average Weekly Hours[3]	-0.673	-0.420	-0.663	-0.686
Average Hourly Earnings[3]	0.642	-0.397	0.919	-0.690
Growth Rate of Average Hourly Earnings[3]	-0.289	-0.515	-0.146	-0.419
Dow Jones Index[4]	-0.471	-0.314	-0.955	-0.692
S&P 500 Index[4]	-0.626	-0.359	-0.966	-0.686
St. Louis FRB Financial Stress Indicator[4]	0.694	0.209	0.888	0.476
CPI Inflation Rate[5]	-0.370	-0.349	-0.566	-0.251
Total Retail Sales[6]	0.143	-0.016	-0.529	-0.249
95% Confidence Bands	±0.192	±0.289	±0.485	±0.305

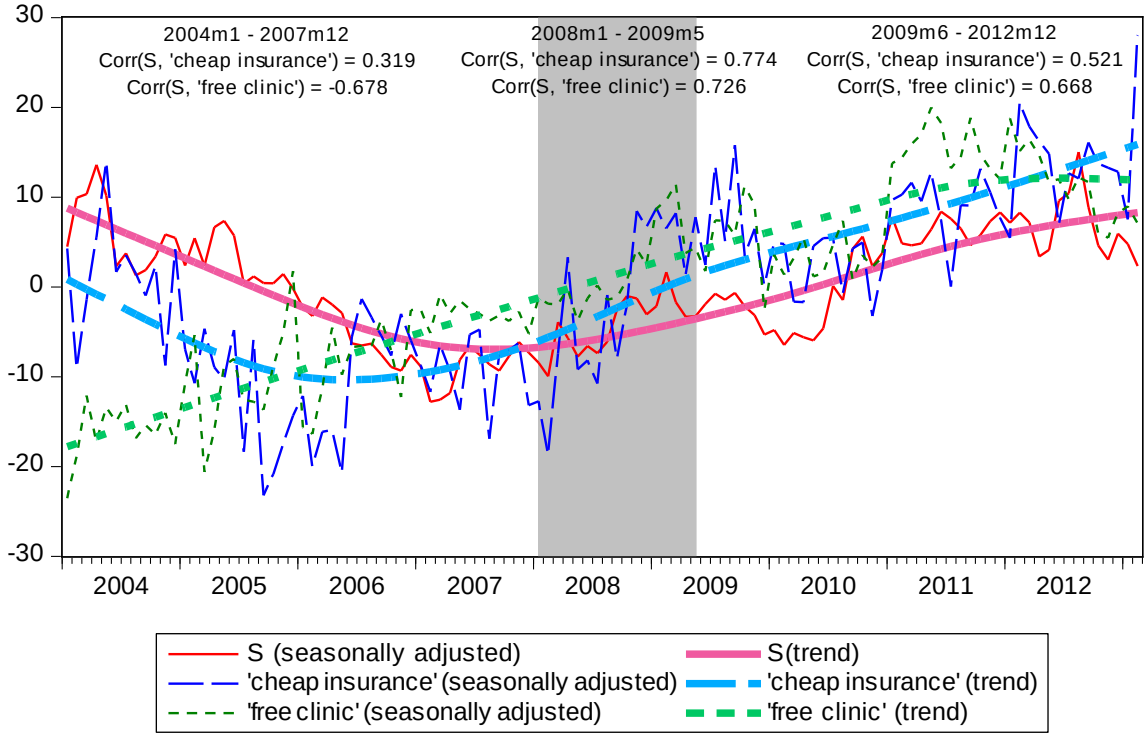
[1] Bureau of Labor Statistics (2013c).
[2] Bureau of Labor Statistics (2013d).
[3] Bureau of Labor Statistics (2013b). Production and non-supervisory employees.
[4] Federal Reserve Bank of St. Louis (2013).
[5] Bureau of Labor Statistics (2013a).
[6] United States Census Bureau (2013).

Figure 4 –STD Interest (S) and Economic (E) Trends[1], Seasonally Adjusted[2]: Monthly, January 2004 – December 2012



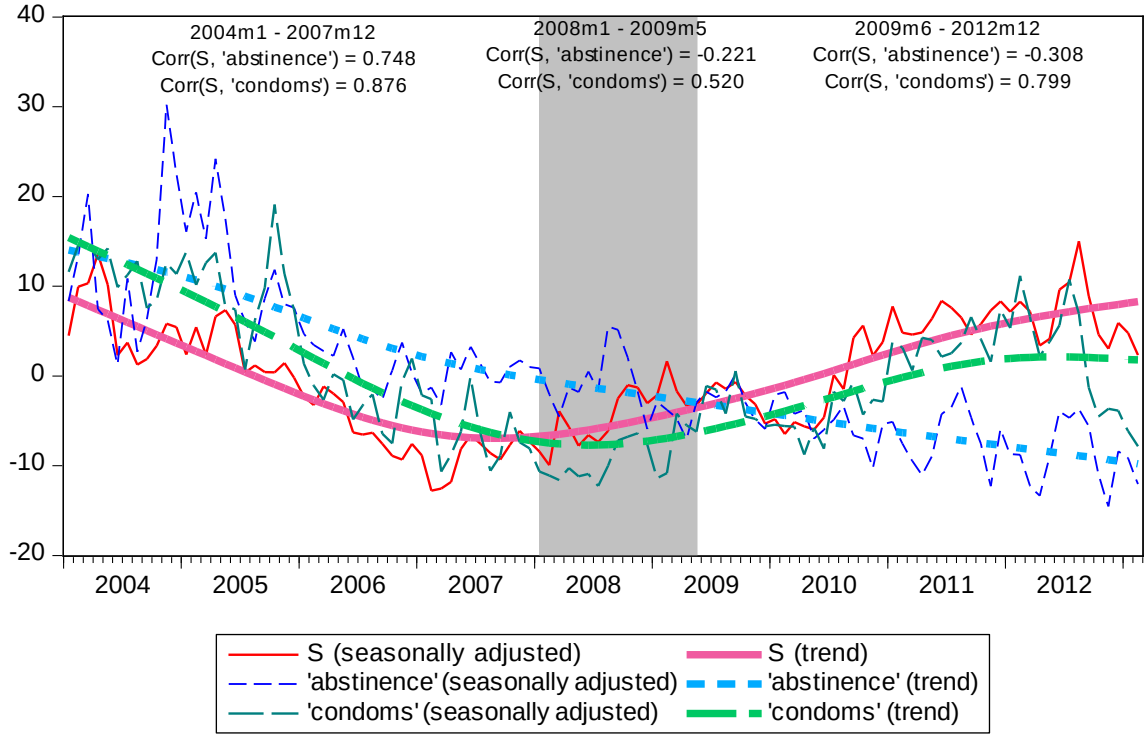
[1] Trend extracted using the Hodrick-Prescott Filter.
[2] Seasonally adjusted using monthly dummy variables.

Figure 5 –STD Interest (S), 'Cheap Insurance' and 'Free Clinic' Trends[1], Seasonally Adjusted[2]: Monthly, January 2004 – December 2012



[1] Trend extracted using the Hodrick-Prescott Filter.
[2] Seasonally adjusted using monthly dummy variables.

Figure 6 –STD Interest (S), 'Abstinence' and 'Condoms' Trends[1], Seasonally Adjusted[2]: Monthly, January 2004 – December 2012



[1] Trend extracted using the Hodrick-Prescott Filter.
[2] Seasonally adjusted using monthly dummy variables.