Can Internet Search Behavior Help to Forecast the Macro Economy?

Taoxiong Liu Xiaofei Xu Tsinghua University

INTRODUCTION

In macro economy analysis, two types of data can be applied, namely the structured data and unstructured data. Government statistics are well structured while Internet search behavior information belongs to non-structured data, This research uses 6 types of models to forecast the macroeconomic aggregate. By comparing different models, the optimal forecast model is selected. We shows that the Internet searching behavior can help forecast the macro economy. Moreover we find that the correct way for variables selection with structured and unstructured data is ***the two-step method***. Firstly only the government statistics are used and the conditional optimal model is selected. Secondly, the internet search data are added into these model and the optimal model is determined.

METHOD

a. Constructs 6 types of models for the forecast of quarterly GDP

- Use the explained variable information alone

-Separately use the government statistical information and Internet searching behavior information

-On the basis of explained variable information, the government statistical data and Internet searching behavior data are added respectively

-Use the explained variable information, government statistics data and Internet searching behavior information at the same time

b. Model selection criterion: according to the BIC value (or training MSE)

c. A comparative analysis

-By using the automatic model selection in Oxmetrics software, we find its results are worse than ours. This further illustrates the effectiveness of the model selection and forecast method of this study.

COMPUTATION AMOUNT

The total calculation of 6 types of models consists of 135545 linear regression models.

DATA

-The government statistics data: structured, less noise, lagged available

-Internet searching behavior information: originally unstructured, real-time available, very noisy

-Forecasted variable (GDP), quarterly data of 2005-2014, from the People's Republic of China Statistics Bureau network

- Government statistical variables(X), 2005-2014 of government monthly statistical, from the People's Republic of China Statistics Bureau

-Internet searching behavior information(Z), Baidu Index of 2006-2014, form Baidu website

EMPIRICAL RESULT

a. Regression and prediction by Autoregressive model of GDP. According to the BIC, only GDP4 is selected as explanatory variable.

Table 1 GDP autoregression results

Model and Candidates of Variables	Selected Variables	variables of significance level p<0.05	BIC	Training MSE	Forecast MSE
Model（1）; GDP1~ GDP 4	GDP4	GDP4	-125.6345	0.000320	0.000204

Note: GDPi, i = 1, 2, 3,4, means lagged i period GDP.

b. Regression and prediction with the use of government statistical variables alone or Internet searching behavior information alone. The prediction power of Internet searching behavior information is worse than that of the government statistics variables; so the traditional government statistics is still very meaningful and the Internet searching behavior information cannot replace the government statistics. The results are shown in Table 2 (only list the best 3):

Table 2 Government statistics and Internet searching behavior

Models and Candidates of Variables	Selected Variables	variables of significance level p<0.05	BIC	Training MSE	Forecast MSE
Model（2）; (X1~X4)	M01,m02,m03,m04	M01, m02, m03, m04	-145.7334	0.000157	0.001545
	EX0,EX1,EX2,EX3,EX4	EX0,EX1,EX2,EX3,EX4	-112.8426	0.000452	0.002093
	nfr0, nfr 1, nfr 2	Nfr0, nfr 1, nfr 2	-106.8868	0.000710	0.000579
Mode（3）; (Z0~Z4)	$G_0 2$		-79.5910	0.002129	0.004354
	$N_{24}0, N_{24}1, N_{24}2$	$N_{24}0, N_{24}1$	-75.4858	0.001942	0.000624
	G_80, G_81	G_80, G_81	-71.7370	0.002806	0.007624

c. In the model with GDP lags plus government statistical variables or GDP lag plus Internet searching behavior information, we also find that EInternet searching behavior information cannot replace the government statistical data. The results as shown in Table 3(only list the top 3)

Table 3 GDP lags plus government statistics and GDP lags plus Internet searching behavior

Models and Candidates of Variables	Selected Variables	variables of significance level p<0.05	BIC	Training MSE	Forecast MSE
Model（4）; (GDP1~ GDP4 , X1~X4)	GDP1,GDP4,mpmi0,mpmi1,mpmi3	GDP1,GDP4,mpmi0,mpmi1,mpmi3	-169.7211	0.000059	0.000156
	GDP4,nfr0,nfr4	GDP4,nfr0,nfr4	-164.1002	0.000092	0.000118
	GDP4,CPI0,CPI2	GDP4,CPI0,CPI2	-159.5940	0.000108	0.000273
Model（5）; (GDP1~ GDP4, Z0~Z4)	GDP1,GDP4,I_{15} 0	GDP4, I_{15} 0	-151.8069	0.000143	0.000038
	GDP2,GDP4,C_30, C_33, C_3 4	GDP2,GDP4,C_30, C_33, C_3 4	-148.9493	0.000061	0.000124
	GDP2,GDP4, I_60	GDP4, I_60	-147.4927	0.000167	0.000113

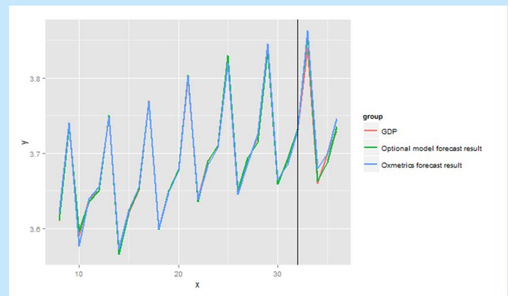
d. By adding unstructured data (Zs) to the conditionally optimal model within the models having GDP lags and single other variable of structured data (Xs) (i.e. Model (4)), we find it can improve the forecast. (Two-Step Method)

e. . By adding unstructured data (Zs) to the conditionally optimal model within the models having GDP lags and two other variable of structured data (Xs), we find that the forecast can still be impproved. (Two-Step Method)

Table 5 Adding Search Behavior into the Model without Search Behavior

Models and Candidates of Variables	Selected Variables	variables of significance level p<0.05	BIC	Training MSE	Forecast MSE
Adding Z0~Z4 Into the 1st optimal one within structured data	GDP4,mpmi0,mpmi2,m00,m04, $N_{17}1$	GDP4,mpmi0, mpmi2,m00,m04, N_{17} 1	-196.913	0.00002	0.00348
	GDP4,mpmi0,mpmi2,m00,m04, $N_{17}0$	GDP4,mpmi0, mpmi2,m00,m04, $N_{17}0$	-195.134	0.00002	0.00353
	GDP4,mpmi0,mpmi2,m00,m04, I_71	GDP4,mpmi0, mpmi2,m00,m04, I_71	-194.513	0.00002	0.00773
Adding Z0~Z4 Into the 2st optimal one within structured data	GDP4,CPI0,CPI1, mpmi0,mpmi2,I_60, I_61	GDP4,CPI0,CPI1,mpmi0,mpmi2, I_60	-194.091	0.00002	0.00006
	GDP4,CPI0,CPI1, mpmi0,mpmi2, $N_{13}0$	GDP4,CPI0,CPI1,mpmi0,mpmi2, $N_{13}0$	-192.396	0.00002	0.00008
	GDP4,CPI0,CPI1, mpmi0,mpmi2, $N_{24}0$	GDP4,CPI0,CPI1,mpmi0,mpmi2, $N_{24}0$	-191.890	0.00002	0.00015

f. Compare the results from the software of Oxmetrics, which does the model selection automatically, and the result of our two-step method as in figure below, which illustrates that the second method is better.



CONCLUSION： In macroeconomic forecast, two types of data might be involved, structured data and unstructured data. We use government statistics and Internet search behavior information to forecast GDP. By comparing different models, the optimal model is selected. We shows that the Internet searching behavior can help forecast the macro economy. Moreover we find that the correct way for variables selection with structured and unstructured data is ***the two-step method***. Firstly only the government statistics are included and the conditional optimal model is selected. Secondly, the internet search data are added into the model and the optimal model is determined.