利用大数据预测季度GDP增速

何强

国家统计局统计科学研究所

2020年12月8日

主要内容

一、研究背景及文献述评

二、理论模型与估计方法

三、指标选取与数据来源

四、实证分析结果及讨论

五、主要结论及未来方向

一、研究背景及文献述评

- **GDP增速及拐点预测：**经济新常态和疫情期的特殊意义
- **季度GDP核算：**1992年起步，2011年环比，2015年正式
- **季度GDP增速预测常用的传统方法：**

一是定性预测方法

二是定量预测方法：计量经济学、系统仿真、其他方法

- 囿于所用数据在发布频率、量体等诸多方面的制约因素，以及模型自身的缺陷，预测结果的有效性常常遭受质疑

➢ 大数据为宏观经济走势预测带来新的机遇

百度指数关键词搜索趋势

地域范围 全国　　设备来源 PC+移动　　时间范围 2011-11-28 ~ 2020-11-30

Baidu 指数

算法说明：以网民在百度的搜索量为数据基础，以关键词为统计对象，科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权。根据数据来源的不同，搜索指数分为PC搜索指数和移动搜索指数。

- **大数据的内涵界定**

技术能力视角：大数据是规模超过现有数据库工具获取、存储、管理和分析能力的数据集。

概念内涵视角：可以将大数据定义为具有海量性、高速性、多样性和可变性等特征的多维数据集，需要通过可伸缩的体系结构实现高效的存储、处理和分析。

——2017年12月

信息技术大数据国家标准（GB/T 35295-2017）：具有体量巨大、来源多样、生成极快、且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。（2017.12.29发布，2018.7.1实施）

——中华人民共和国原国家质量监督检验检疫总局

中国国家标准化管理委员会

国家统计局对大数据给出的定义：

国家统计局和国家发改委（2017；官方指导意见）：大数据是非传统数据的主体，在很多情形下可以代指非传统数据。具体而言，非传统数据是指通过非传统政府统计调查获取的数据（国外一些机构也称之为“二手数据”）。

国家统计局（2013，2015；研究成果）：大数据通常被认为是采用多种数据收集方式、整合多种数据来源的数据，是采用现代信息技术和架构高速处理和挖掘、具有高度应用价值和决策支持功能的数据、方法及其技术集成。

- **大数据为经济预测带来的机遇**

  一是数据量体

  二是数据类型

  三是数据思维：Forecasting（预测）→ Nowcasting（现测）

- 利用大数据预测季度GDP增速的代表性文献有Götz and Knetsch（2019）、Clark et al（2017）、Kopoin et al（2013）、Carriero et al（2012）以及刘涛雄和徐晓飞（2015），等等。

- **利用大数据预测季度GDP的建模思路探索**

  一是抽取变量数据的主要解释因子

  二是假定变量估计系数满足一定的统计分布

  三是利用高维数据机器学习模型

  四是将大数据变量植入主流的DSGE模型

  五是其他思路（系统动力学等）

➢ 利用大数据预测季度GDP的建模思路探索

√ 一是抽取变量数据的主要解释因子

二是假定变量估计系数满足一定的统计分布

三是利用高维数据机器学习模型

四是将大数据变量植入主流的DSGE模型

五是其他思路（系统动力学等）

- **本报告利用大数据建模分析的主要内容**

一是利用大数据指标完善季度GDP增速预测方法

二是探讨大数据月度指标在季度GDP预测中的作用

三是优化所用的混频大数据动态因子模型

二、理论模型与估计方法

- 采用混频大数据动态因子模型的主要原因

一是自变量和因变量的数据频度不同

二是便于减少大量月度指标之间的信息重叠

三是公共因子时间序列变量可能存在较为显著的自回归关系

- 动态因子模型的基本结构

$$x_t = (x_{1,t}, x_{2,t}, \cdots, x_{n,t})', t = 1, 2, \cdots, T$$

$$x_t = \Lambda f_{t,r} + \xi_t, \quad \xi_t \sim \mathrm{N}(0, \sum_{\xi})$$

$$f_{t,r} = \sum_{i=1}^{p} A_i f_{t-i,r} + \varsigma_t$$

$$\varsigma_t = B_{r \times q} \eta_t, \quad \eta_t \sim \mathrm{N}(0, I_q)$$

$\hat{y}_t$

- 引入季度**GDP**环比增速 y_t^Q

$$y_t^Q = \frac{1}{3}(\hat{y}_t + \hat{y}_{t-1} + \hat{y}_{t-2}), \quad t = 3,6,9,12$$

$$\hat{y}_t = \beta' f_t$$

$\hat{y}_t$ 表示基于每季度对应月份的、潜在的GDP月度环比增速变量

$$\varepsilon_t^Q = (y_t^Q - \hat{y}_t^Q) \square \mathrm{N}(0, \sigma_\varepsilon^2)$$

假定随机冲击项序列之间在任何时点均独立

▲以上为混频动态因子模型的基本设置，可以表示为经典的状态空间形式。

- 以**p=1**为例

$$\begin{bmatrix} x_t \\ y_t^Q \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_t \\ \hat{y}_t \\ \hat{y}_t^Q \end{bmatrix} + \begin{bmatrix} \xi_t \\ \varepsilon_t^Q \end{bmatrix}$$

$$\begin{bmatrix} I_r & 0 & 0 \\ -\beta' & 1 & 0 \\ 0 & -\frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} f_{t+1} \\ \hat{y}_{t+1} \\ \hat{y}_{t+1}^Q \end{bmatrix} = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Xi_{t+1} \end{bmatrix} \begin{bmatrix} f_t \\ \hat{y}_t \\ \hat{y}_t^Q \end{bmatrix} + \begin{bmatrix} \zeta_{t+1} \\ 0 \\ 0 \end{bmatrix}$$

- 估计方法

$$\theta = (\Lambda, A_1, \ldots, A_p, \beta, \sum_{\xi}, B, \sigma_{\varepsilon}^2)$$

建立似然函数，利用主成分和卡尔曼滤波的求解方法处理，使得：

$$\Delta ln(\theta^*) = \left| (ln(\theta))_{n+1} - (ln(\theta))_n \right| < \tau$$

三、指标选取及数据来源

- 用于预测季度GDP环比增速的指标集包括：

一是传统宏观经济统计指标集：14个

二是大数据指标集：8个

注意：所有指标均为环比数据

➢ 大数据指标选取规则

——根据国家统计局和国家发改委联合发布的《非传统数据统计应用指导意见》科学选取大数据变量指标

——非传统数据是指通过非传统政府统计调查获取的数据，包括政府部门的行政记录数据、商业记录数据、互联网数据、电子设备感应数据以及其他非传统数据

——大部分大数据变量的原始数据均可以免费获取

——部分大数据需要通过网络爬取（借助Python的爬虫框架scrapy）

预测季度GDP增速的月度指标集

指标集	指标名称	样本区间	指标集	指标名称	样本区间
传统宏观经济统计指标集	居民消费价格指数	2011Q1-2018Q3	大数据指标集	上海钢联大宗商品价格指数	2011Q1-2018Q3
	工业生产者出厂价格指数	2011Q1-2018Q3			
	固定资产投资完成额	2011Q1-2018Q3		中国煤炭价格指数（全国综合指数）	2011Q1-2018Q3
	规模以上工业增加值	2011Q1-2018Q3			
	社会消费品零售总额	2011Q1-2018Q3			
	进出口总额	2011Q1-2018Q3		财新中国制造业PMI	2011Q1-2018Q3
	广义货币（M2）	2011Q1-2018Q3			
	社会融资规模	2011Q1-2018Q3		财新中国服务业PMI	2011Q1-2018Q3
	国家财政支出额	2011Q1-2018Q3			
	中国制造业PMI	2011Q1-2018Q3		物流景气指数	2016Q1-2018Q3
	非制造业商务活动指数	2011Q1-2018Q3		新经济指数	2016Q1-2018Q3
	大宗商品价格指数	2011Q1-2018Q3		电商物流运行指数	2016Q1-2018Q3
	发电量	2011Q1-2018Q3			
	股市日均成交额	2011Q1-2018Q3		数字经济指数	2016Q1-2018Q3

➢ 大数据指标的具体来源说明

上海钢联大宗商品价格指数来源于上海钢联官方网站，中国煤炭价格指数（全国综合指数）来源于中国煤炭市场网，物流景气指数和电商物流运行指数来源于中国物流与采购联合会，其余指标均来自于财新网。

四、实证分析结果及讨论

- **根据大数据指标情况分成两部分讨论：**

（一）基于2011Q1-2018Q3期间样本的实证分析

14个传统宏观经济指标，4个大数据指标

（二）基于2016Q1-2018Q3期间样本的实证分析

14个传统宏观经济指标，8个大数据指标

- **预测优劣程度的评价指标：**

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(GDP_j^f - GDP_j)^2}{n}}$$

➢ 1.基于2011Q1-2018Q3期间样本的实证分析

不包含大数据指标模型预测的均方根误差（代表性部分）

		q=13	q=12	q=11	q=10	q=9	q=8	q=7
p=1	r=13	0.1089	0.1106	0.1140	0.1102	0.0953*	0.1026	0.1035
	r=12	—	0.1117	0.1138	0.1134	0.1020	0.1234	0.1214
	r=11	—	—	0.1170	0.1216	0.1190	0.1309	0.1250
p=2	r=13	0.1088	0.1091	0.1074	0.1068	0.1149	0.1151	0.1207
	r=12	—	0.1117	0.1093	0.1102	0.1212	0.1194	0.1287
	r=11	—	—	0.1176	0.1184	0.1246	0.1230	0.1203

包含大数据指标模型预测的均方根误差（代表性部分）

		q=17	q=16	q=15	q=14	q=13	q=12	q=9
p=1	r=17	0.0987	0.0974	0.0967	0.1036	0.1080	0.1041	0.0821*
	r=16	—	0.1053	0.1054	0.1078	0.1078	0.1059	0.0894
	r=15	—	—	0.1054	0.1067	0.1104	0.1117	0.0952
p=2	r=17	0.0982	0.1006	0.1023	0.1050	0.1072	0.1168	0.1231
	r=16	—	0.1050	0.1051	0.1043	0.1064	0.1164	0.1137
	r=15	—	—	0.1053	0.1047	0.1071	0.1137	0.1090

不含大数据		q=13	q=12	q=11	q=10	q=9	q=8	q=7
p=1	r=13	0.1089	0.1106	0.1140	0.1102	0.0953	0.1026	0.1035
	r=12	—	0.1117	0.1138	0.1134	0.1020	0.1234	0.1214
	r=11	—	—	0.1170	0.1216	0.1190	0.1309	0.1250
p=2	r=13	0.1088	0.1091	0.1074	0.1068	0.1149	0.1151	0.1207
	r=12	—	0.1117	0.1093	0.1102	0.1212	0.1194	0.1287
	r=11	—	—	0.1176	0.1184	0.1246	0.1230	0.1203

含大数据		q=17	q=16	q=15	q=14	q=13	q=12	q=9
p=1	r=17	0.0987	0.0974	0.0967	0.1036	0.1080	0.1041	0.0821
	r=16	—	0.1053	0.1054	0.1078	0.1078	0.1059	0.0894
	r=15	—	—	0.1054	0.1067	0.1104	0.1117	0.0952
p=2	r=17	0.0982	0.1006	0.1023	0.1050	0.1072	0.1168	0.1231
	r=16	—	0.1050	0.1051	0.1043	0.1064	0.1164	0.1137
	r=15	—	—	0.1053	0.1047	0.1071	0.1137	0.1090

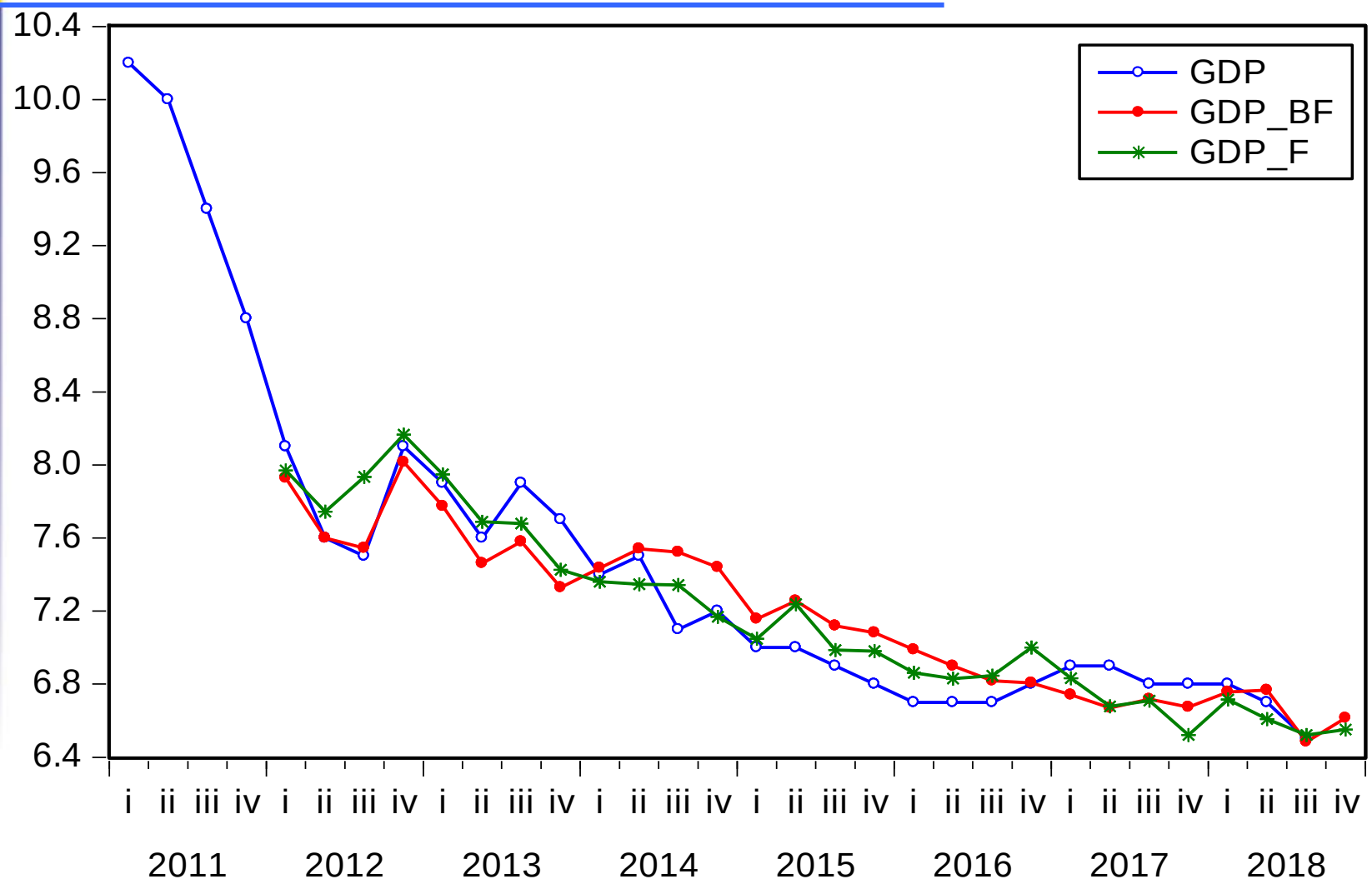
- **两点基本结论**

——大数据指标确实较为显著地降低了模型预测的均方根误差（RMSE），提升了预测精度

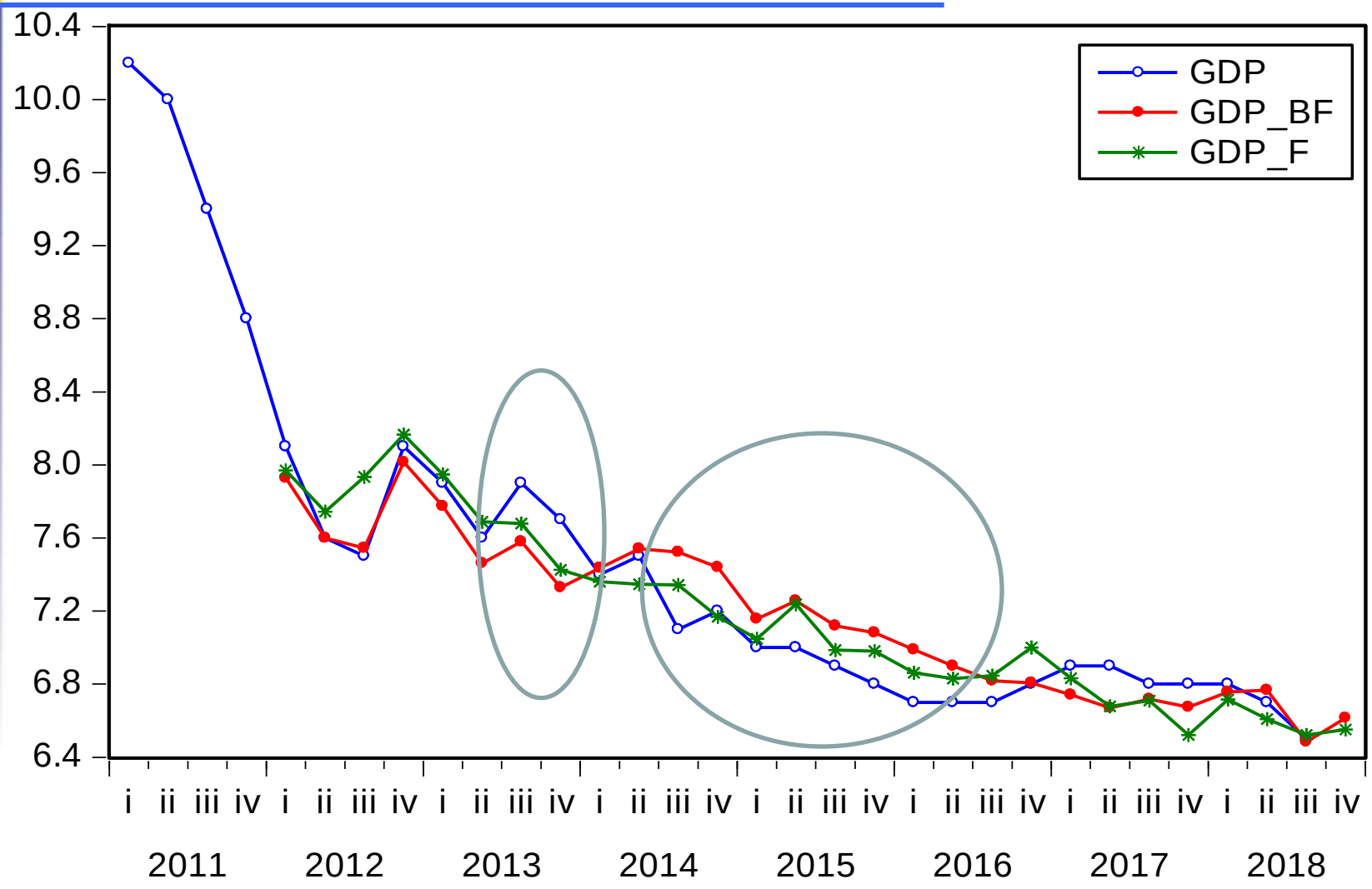
——发挥大数据指标的预测功用，需要科学合理设置参数，并非只靠大数据本身就能够较为方便地得出较好的预测效果

➢ 利用下式将季度GDP环比增速，转换成季度GDP同比增速：

$$y_t^q = 100 \times [\prod_{i=0}^{3} (1 + \frac{y_t^Q}{100}) - 1], t = 12,15,18,\cdots$$


10.4
10.0
9.6
9.2
8.8
8.4
8.0
7.6
7.2
6.8
6.4
GDP
GDP_BF
GDP_F
i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv
2011
2012
2013
2014
2015
2016
2017
2018

模型预测效果对比图


10.4
10.0
9.6
9.2
8.8
8.4
8.0
7.6
7.2
6.8
6.4
GDP
GDP_BF
GDP_F
i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv i ii iii iv
2011
2012
2013
2014
2015
2016
2017
2018

模型预测效果对比图

➢ 2.基于2016Q1-2018Q3期间样本的实证分析

不包含大数据指标模型预测的均方根误差（代表性部分）

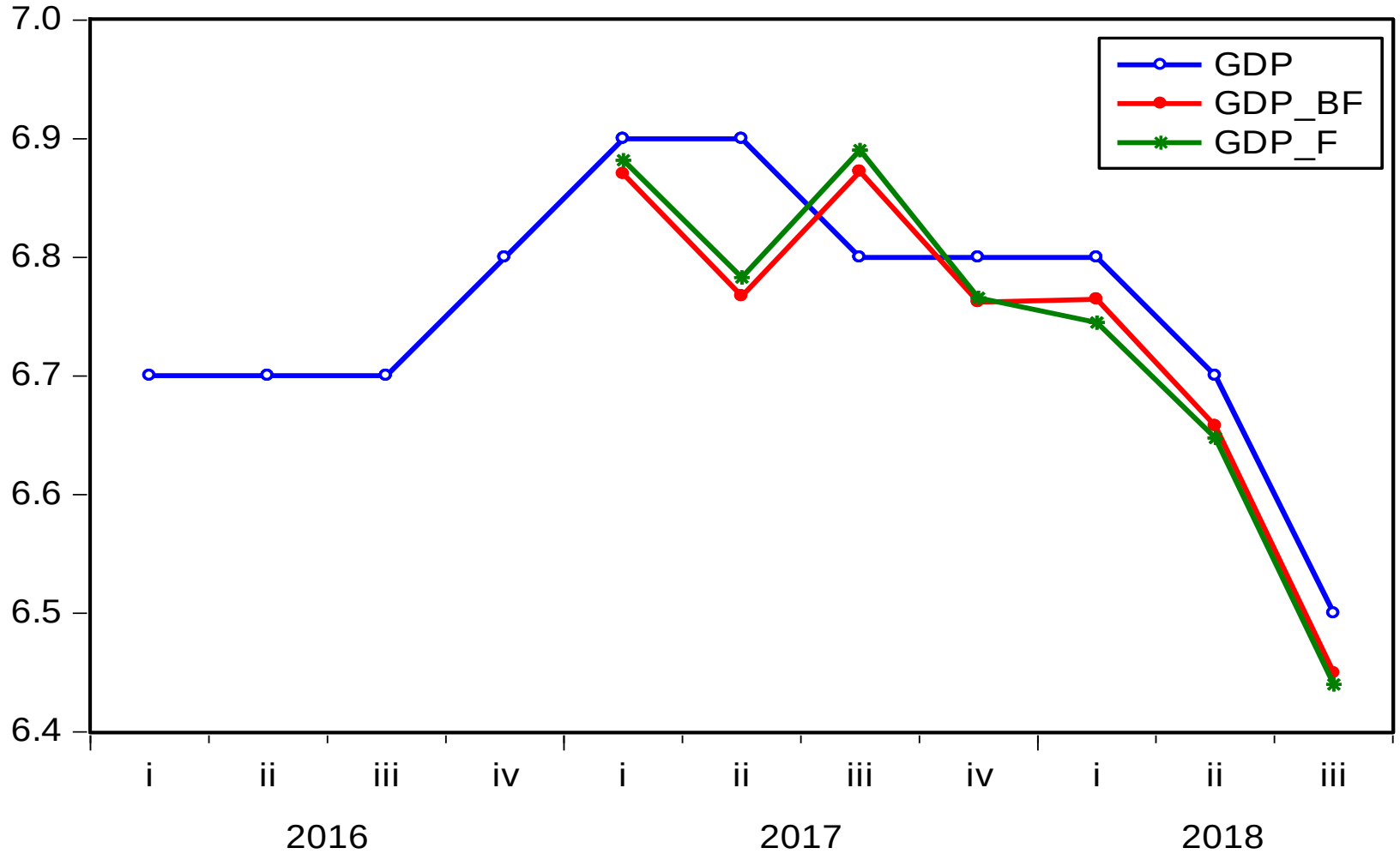
		q=8	q=7	q=6	q=5	q=4	q=3
p=1	r=8	0.0143	0.0194	0.0147	0.0500	0.0337	0.0077*
	r=7	—	0.0308	0.0574	0.0617	0.0615	0.0354
	r=6	—	—	0.0537	0.0611	0.0561	0.0785
p=2	r=4	—	—	—	—	0.0463	0.0271
	r=3	—	—	—	—	—	0.0641
	r=2	—	—	—	—	—	—

包含大数据指标模型预测的均方根误差（代表性部分）

		q=8	q=7	q=6	q=5	q=4	q=3
p=1	r=8	0.0560	0.0562	0.0421	0.0064	0.0175	0.0184
	r=7	—	0.0545	0.0526	0.0622	0.0596	0.0613
	r=6	—	—	0.0570	0.0680	0.0654	0.0854
p=2	r=4	—	—	—	—	0.0211	0.0014*
	r=3	—	—	—	—	—	0.0643
	r=2	—	—	—	—	—	—

不含大数据		q=8	q=7	q=6	q=5	q=4	q=3
p=1	r=8	0.0143	0.0194	0.0147	0.0500	0.0337	0.0077
	r=7	—	0.0308	0.0574	0.0617	0.0615	0.0354
	r=6	—	—	0.0537	0.0611	0.0561	0.0785
p=2	r=4	—	—	—	—	0.0463	0.0271
	r=3	—	—	—	—	—	0.0641
	r=2	—	—	—	—	—	—

含大数据		q=8	q=7	q=6	q=5	q=4	q=3
p=1	r=8	0.0560	0.0562	0.0421	0.0064	0.0175	0.0184
	r=7	—	0.0545	0.0526	0.0622	0.0596	0.0613
	r=6	—	—	0.0570	0.0680	0.0654	0.0854
p=2	r=4	—	—	—	—	0.0211	0.0014
	r=3	—	—	—	—	—	0.0643
	r=2	—	—	—	—	—	—


7.0
6.9
6.8
6.7
6.6
6.5
6.4
GDP
GDP_BF
GDP_F
i
ii
iii
iv
i
ii
iii
iv
i
ii
iii
2016
2017
2018

模型预测效果对比图

- **三点基本结论：**

——在同样的参数结构下，包含全部大数据指标的模型并非在所有情形下都拥有更小的预测均方根误差（REMSE）

——预测效果最好的包含全部大数据指标模型的RMSE，明显低于预测效果最好的不包含大数据指标的基准模型

——样本长短对模型预测效果存在较大影响

五、主要结论及未来方向

- 大数据月度指标蕴含的信息，有助于提升季度GDP增速预测精度和时效性，但这一结论成立的重要前提是需要获取相对较长的时间序列样本，并科学合理地设置模型估计的参数。

五、主要结论及未来方向

- 大数据月度指标蕴含的信息，有助于提升季度GDP增速预测精度和时效性，但这一结论成立的重要前提是需要获取相对较长的时间序列样本，并科学合理地设置模型估计的参数。

- 在同等参数结构设置情形下，仅仅通过增加大数据月度指标的信息体量，并非总是能够降低预测的均方根误差。

五、主要结论及未来方向

- 大数据月度指标蕴含的信息，有助于提升季度GDP增速预测精度和时效性，但这一结论成立的重要前提是需要获取相对较长的时间序列样本，并科学合理地设置模型估计的参数。

- 在同等参数结构设置情形下，仅仅通过增加大数据月度指标的信息体量，并非总是能够降低预测的均方根误差。

- 利用大数据建模预测时，可能需要更加多样的数据、更加新颖的建模切入点和更加抗噪的模型机制设计。

结束语

欢迎批评指正！