计量经济模型适用于评估GDP数据质量吗？*

郭红丽　王 华

摘要：针对中国 GDP 数据质量评估研究中广泛采用的计量经济模型方法，本文从其技术原理的局限性以及 GDP 核算误差发生机制的影响效应两个方面，系统考察了该方法在具体评估实践中的适用性。研究分析表明，基于计量经济模型的两类评估方法（参数可靠性分析与异常数值识别），在实际操作中易于陷入“数据质量不佳”或者“模型结构设定不当”的对立解读困境；因忽视了 GDP 核算误差的发生机制，根据模型估计结果的异常特征来反推 GDP 数据序列中误差结构（效应），其可行性有所不足。计量模型方法之于 GDP 数据质量评估研究的适用性的提升，还有待于对 GDP 核算误差机制的显性识别；本文在此方面提供了一个相对完备的概念性基准模型，有助于相关研究的进一步深化拓展。

关键词：GDP 数据质量　计量经济模型　统计数据诊断　随机数值模拟

中图分类号：C812, F222　　**JEL：**C52, C82

Can the Econometric Models be Applicable to Evaluating the Accuracy of GDP Statistics?

GUO Hongli　WANG Hua

（Xiamen University, Xiamen, China）

Abstract: By means of an analysis on methodological limitations and a stochastic numerical simulation according to some hypotheses about the generation mechanisms of GDP accounting error, the paper aims at inspecting the applicability of those approaches adopting econometric models (AEM for short) into evaluating the accuracy of GDP statistics. The outcomes indicate there has been a dilemma for AEM to interpret biased parameters and abnormal residuals came from the fitted econometric models as poor data quality or inappropriate model structure. Meanwhile, it's difficult to infer componential and compositive effects of GDP error on fitting outcome of econometric models, since the generation mechanisms of GDP accounting error have been ignored in common. In order to improve the applicability of AEM, explicit recognition of GDP accounting error mechanisms is needed. In this aspect, the paper develops a relatively complete conceptual model, which provides a benchmark to further research.

Keywords: GDP Statistics Quality; Econometric Models; Statistics Data Diagnosis; Stochastic Numerical Simulation

一、引言与文献综述

中国 GDP 数据的可信度历来备受质疑，对于其数据质量的评估检验成为学界持续关注的议题。而鉴于 GDP 统计过程的复杂性，学者们无法按照其严格的统计标准进行独立的外部估

* 郭红丽，厦门大学经济学院，邮政编码：361005，电子邮箱：guohl2004@163.com；王华，厦门大学台湾研究院经济研究所，邮政编码：361005，电子邮箱：wanghua_xmu@163.com。本文受国家社会科学基金项目“中国 GDP 核算误差、数据修订及其影响机制研究”（12CTJ015）资助。感谢编辑部及匿名审稿人的修改建议，文责自负。

算，除了 Maddison（1998）、Maddison & Wu（2008）等少数研究，尝试遵循 SNA 核算规则来重构 GDP 及其增长率序列，更受认可的方法则是参照相关指标的变动趋势对 GDP 数据质量进行所谓的交叉检验。例如 Rawski（2001）依据 1997－2000 年间中国能源消耗相对下降、就业人口增长缓慢、通货紧缩等现象，判定由官方 GDP 数据所宣示的高速经济增长与现实不符，GDP 统计存在严重的上偏误差。Rawski 的观点在当时国内外引起强烈反响，但其所采用的方法属于短时期的单一指标交叉检验范畴，证据较为片面，结论不尽可靠。与之相比，引入更多相关指标，覆盖更广时空范围，通过建立计量经济模型对 GDP 数据的可信度予以系统性地评估，这种做法更具直观上的合理性，因而演化成为该领域研究的主流方法脉络。

（一）计量模型方法在 GDP 数据质量评估中的具体应用

计量经济模型在 GDP 数据质量评估中的可用性，首先在于对计量模型参数估计结果的考察和解读。Adams & Chen（1996）建立了中国能源消费对 GDP 的回归模型，结果发现中国能源消费对 GDP 的弹性系数大大低于同时期东亚国家的平均水平，鉴于中国自改革开放以来的经济增长带有明显的高能耗特征，断定中国官方统计夸大了中国 GDP 的增长率。孟连、王小鲁（2000）估计了 1953－1997 年间工业部门和 GDP 的生产函数，发现 1992－1997 年间（工业）全要素生产率的增长率突然由前一时期的平均 2.5%跃升到 7.3%；这很难使人信服，因为没有证据表明在此期间发生了史无前例的技术进步的加速，据此判断主要是由增长率统计的虚增所致。Klein & Ozmucur（2003）选取了足以覆盖能源、交通、通讯、劳动力、农业、贸易、公共部门、工资、通货膨胀等各方面信息的、具有代表性的 15 个经济变量，对 1980－2000 年间的 GDP 增长率进行主成分回归方程估计，发现模型参数估计结果完全符合经济规律，不能从中得出中国官方统计的经济增长速度被明显高估的结论。

计量模型方法的另一分析路径是考察以 GDP 作为被解释变量的模型拟合残差，借助各种诊断统计量来识别数据集中偏差或影响较为显著的异常数值点。周建（2005）较早地将统计诊断（异常数值识别）方法应用于 GDP 数据质量评估，其选取全社会固定资产投资增速、就业人数增速、电力消费增速和科技拨款增速作为解释变量，建立了基于 GDP 增速的生产函数模型；模型估计结果显示，模型拟合误差均在可容忍的范围内，因此认为 GDP 统计数据是比较可靠的。刘洪、黄燕（2007）基于 1978－2003 年间 GDP 的 ARMA 模型，对 2004 年 GDP 数据的准确性进行评估，结果发现报告期的 GDP 并非异常值。刘洪、黄燕（2009）以及卢二坡、黄炳艺（2010）则通过拟合 GDP 的生产函数，利用多种诊断统计量分析各样本点对模型结果稳定性的影响，发现少数年份样本点的诊断统计量较为异常，可列为可疑数据。卢二坡、张焕明（2011）、刘洪、昌先宇（2011）以及刘洪、金林（2012）则致力于模型结构与估计方法的改进，以期提高异常数值识别的效果。

为克服计量经济模型中其他系统变量与 GDP 之间的相关关系不尽稳定的难题，近年来部分研究转而借助其他来源的、被认为更能反映实际经济增长趋势的独立数据来建立评估模型。例如，徐康宁等（2015）引入了全球夜间灯光数据，力求从一个相对客观的视角验证中国经济增长以及 GDP 统计数据的真实性；通过对 1992－2012 年间省级面板数据的计量分析，发现无论是全国整体还是各省份，研究期间实际经济增长率的平均值与官方统计数据都不完全吻合，全国整体低 1.02 个百分点，东、中、西三大区域低约 1.5~1.8 个百分点，且经济落后地区的差距要大于经济发达地区；剔除统计技术因素，认为存在着地方政府夸大 GDP 统计数据的可能性。与之相关，卢盛峰等（2017）计算了官方发布的实际城市生产总值相对于城市夜间灯光亮度值的偏离程度，并将该指标定义为中国城市 GDP 注水系数，基于此检验了地方官

员政治晋升周期下的 GDP 注水冲动。

（二）问题的提出

由上可知，计量模型方法在 GDP 数据质量评估中的应用实际存在多种分析路径与评判逻辑，导致不同评估研究的结果往往表现出或多或少的差异（甚至是根本性的背离）。而即使采用同种方法逻辑，模型拟合过程中数据类型的差异也可能导致研究结论的分歧。以 Klein & Ozmucur（2003）的研究为基础，阙里、钟笑寒（2005）、周国富、连飞（2010）进一步建立了反映省区 GDP 变动的空间面板数据模型，虽然整体上没有发现 GDP 数据失真的系统证据，但各省区之间的空间固定效应却显现较大差异（部分省区的固定效应显著高于平均水平，部分省区的固定效应又显著低于平均水平）；排除经济因素和社会因素，认为特定省区的 GDP 数据中可能存在一定程度的高估或低估。面板数据的引入，使后两项研究得以在评判模型参数合理性的基础上，对模型误差项的时空关联特征（变截距）做进一步考察；但正如后文分析所指出的，模型中变截距的存在，既可能源于特定省区的 GDP 数据质量不佳，也可能是因为模型本身的拟合度不足，对于其结果的解读因而仍会存在很大的不确定性！①

不仅如此，现有研究在利用计量经济模型评估中国 GDP 数据质量时，还存在一项更为关键的技术缺陷，即过份关注于计量经济模型的建构与拟合，却忽视了 GDP 作为一项统计产品，其本身可能服从何种数据生成机制（测量误差发生机制），对于计量模型估计结果（以及 GDP 数据质量评估结果）又会造成何等影响？如果将前述各项评估研究视为对计量经济模型的一种反向逻辑的应用（即根据模型估计结果反推数据的可用性与可信度），对于 GDP 误差发生机制及其影响效应的关注与探讨，则可视为对计量经济模型评估功效的一种正向逻辑的考察。在此方面，目前仅有刘小二、谢月华（2009）设定了 GDP 核算误差的适应性预期调整机制，孙艳、贡颖（2013）、Sinclair（2019）设定了（季度）GDP 数据修订的自我关联机制，但其计量模型的整体结构仍较为简略，还无法从中有效识别特定的 GDP 误差效应。

针对包括 GDP 在内的宏观经济统计数据，郭红丽、王华（2011）曾指出，已有的质量评估研究均无法依照数据“准确性”的严格定义，只能采取替代性做法，即针对待评估指标与参照指标之间的“一致性”进行检验，将此一致性特征作为评判前者准确性的依据标准；其前提是能够确保统计数据的一致性与准确性这两类质量属性之间必然存在逻辑关联。而现实问题在于，一方面，准确的多项统计数据之间未必具有稳定的一致性，若不能排除此种可能，应用计量经济模型评估 GDP 数据质量就可能产生“弃真”错误；另一方面，实际表现出一致性的多项统计数据也未必都是准确的，而现有研究所普遍忽视的 GDP 核算过程中的误差发生机制，往往并不会对计量经济模型的评估结果造成显性影响（即不会破坏 GDP 与模型中其他变量的系统相关性），由此导致“纳伪”错误的发生。换言之，统计数据的一致性并非准确性的充要条件，试图利用计量经济模型来“复制”统计数据之间的一致性，目标虽能（或不能）达成，却并不必然导出 GDP 数据可信与否的结论。

（三）本文研究目标

本文旨在分析论证计量经济模型方法在 GDP 数据质量评估中的有限适用性，揭示各种可能的误判情形，探寻该领域研究在方法论上的拓展空间。具体则开展两个层面的分析，一是在对 GDP 数据质量评估中计量模型方法的基本原理进行提炼的基础上，从不同技术层次分析

① 实际上，从 2018 年第四次全国经济普查后各省份 GDP 数据的修订情况来看，阙里、钟笑寒（2005）、周国富、连飞（2010）对于部分省份 GDP 数据高估或低估的判断，与现实修订方向有较大出入，显示该类模型在评判 GDP 数据质量方面的（预测）能力并不如人意。

其可能的评估功效，判断其发生误判的可能性；二是从 GDP 核算误差的发生机制入手，遵循正向逻辑，建立反映 GDP 现实数据生成过程（其中涵括多种 GDP 核算误差特性）的测量误差模型，通过数值模拟研究考察计量模型能否揭示相应的误差机制，展现其发生“纳伪”错误的可能性。研究结果表明，在方法原理上，计量模型评估方法的应用存在多种误判情形，尤其是无法解决模型拟合与误差识别之间的消长难题（即后文所称的“拟合悖论”）；针对 GDP 数据的各种误差机制，计量模型评估结果所能提供的信息也非常有限，不足以揭示有关 GDP 数据质量的大部分缺陷，评估功效不容乐观。在未来进一步研究中，为切实改善评估功效，需要对 GDP 核算误差机制的识别问题加以正视和重视。

论文的结构安排如下：除引言部分提出问题外，第二部分提炼计量经济模型之应用于 GDP 数据质量评估的方法原理，第三部分分析该类方法在具体应用中可能存在的误判情形，第四部分则借助数值模拟方法，检验计量模型评估结果相对于各种 GDP 误差机制的（不）敏感性，第五部分为结论与研究展望。

二、计量模型评估方法的应用原理

根据前文论述，计量经济模型之于 GDP 数据质量评估中的应用，实际可归纳为两种实施策略，即基于计量经济模型的参数可靠性分析与基于计量经济模型的异常数值识别（王华、金勇进，2009，2010；冯蕾、周晶，2013）；二者都以对计量经济模型的构建和拟合为基础，但在数据质量评估环节则产生方法原理的分异，前者注重考察模型参数，后者则转向考察模型残差。具体的方法原理可分别表述如下。

（一）基于计量经济模型的参数可靠性分析方法

首先依据一定的经济理论或时间序列数据的变动规律，拟合以待评估指标 Y 作为被解释变量的计量经济模型或时间序列模型（此处以纵向维度的数据质量评估为例）

$$Y_t = f(\mathbf{X}_t, t; \boldsymbol{\theta}) + u_t,\ t = 1,2,\cdots,T \tag{1}$$

其中 $\mathbf{X}$ 表示与 Y 相关联的、可作为模型解释变量的参照指标集合，$\boldsymbol{\theta}$ 则表示模型的参数集合。通过适当的估计和检验，得到合宜的模型拟合结果为

$$\hat{Y}_t = f\left(\mathbf{X}_t, t; \hat{\boldsymbol{\theta}}\right),\ t = 1,2,\cdots,T \tag{2}$$

在参数可靠性分析方法的应用中，假定解释变量（向量）$\mathbf{X}$ 的统计数据是准确可靠的，$\mathbf{X}$ 与 Y 之间的回归关系也是先验合理的，通过以下标准来判定待评估指标 Y 的数据质量。

1．*参数结果的意义合理性*。对于上述模型（1）中的特定参数 θ，如果其实际估计值$\hat{\theta}$超出了理论上可能存在合宜取值区间$(\overline{\theta}, \underline{\theta})$，即

$$\hat{\theta} > \overline{\theta} \text{ 或 } \hat{\theta} < \underline{\theta} \tag{3}$$

则认为由模型所反映的经济运行机理明显有悖于社会经济常识，（在排除其他因素干扰后）可判定 Y 的数据中存在显著统计误差。前述相关研究中，Klein & Ozmucur（2003）即遵循了此项逻辑，通过主成分回归发现 15 个基本经济变量的变动率与中国官方统计的 GDP 增长率都保持正相关关系，未能呈现中国 GDP 增长数据有误的证据。

2．*参数结果的跨区间可比性*。与同类型经济体的模型拟合结果相比（令模型参数为$\hat{\theta}'$），如果模型参数估计值的差异程度超出了正常可解释的范围，即

$$|\hat{\theta} - \hat{\theta}'| > \delta \tag{4}$$

表明待评估指标的变动趋势偏离了与参照指标之间本应存在的统计关系，可判定其数据质量不可靠。Adams & Chen（1996）即遵循了此项逻辑，将中国的能源消费（电力消费）相对于

GDP 的弹性系数与美国以及其他 8 个东亚国家（地区）的同类弹性系数相比较，以中国弹性系数过低而得出其官方 GDP 增长率被夸大的判断。阙里、钟笑寒（2005）、周国富、连飞（2010）则同时遵循了式（3）和式（4）的逻辑，既考察面板数据模型中常系数的取值合理性，又分析其变截距的跨省区差异，若特定省区的变截距（空间固定效应）明显高（低）于平均水平，则判定其 GDP 数据中存在高（低）估。

3．*参数结果的跨时期稳定性*。将前后不同时期的模型拟合结果相比，如果相邻时期的模型参数估计值（或其函数）出现不可解释的激增或异常跳动，即

$$|\hat{\theta}_{(2)} - \hat{\theta}_{(1)}| > \delta \ \text{或} \ |g(\hat{\theta}_{(2)}) - g(\hat{\theta}_{(1)})| > \delta \tag{5}$$

同样可判定其数据质量不可靠。孟连、王小鲁（2000）遵循了此项逻辑，通过将整个研究时期划分为三个子区间，估计得到不同子区间的全要素生产率并进行纵向对比，据此断定 1992－1997 年期间的 GDP 增长率被高估。

（二）基于计量经济模型的异常数值识别方法

在异常数值识别方法的应用中，假定模型（1）的拟合值$\hat{Y}$即可代表待评估指标 Y 的真实值，通过以下标准来判定待评估指标 Y 的数据质量。

1．*相对拟合误差*。计算实际统计值与拟合值之间的相对误差

$$P_t = \frac{Y_t - \hat{Y}_t}{Y_t} = \frac{\hat{u}_t}{Y_t}, \ t = 1,2,\cdots,T \tag{6}$$

判断其是否超出事先设定的允许误差限度。刘洪、黄燕（2007）基于历史区间的时间序列组合模型的拟合结果，依据式（6）计算报告期 2004 年的相对误差率，发现该年的样本点并非异常值，因而判定其 GDP 数据是准确的。

2．*诊断统计量*。借助统计数据诊断原理，计算各种诊断统计量，如用于异常点检验的学生化残差

$$r_t = \frac{\hat{u}_t}{s\sqrt{1-h_{tt}}}, \ t_t = \frac{\hat{u}_t}{s(t)\sqrt{1-h_{tt}}} \tag{7}$$

其中$\hat{u}_t$表示模型拟合残差，s表示模型的标准误差，$s(t)$表示删除第 t 个数据点后所拟合模型的标准误差，h_{tt}表示帽子矩阵$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$中的第 t 个对角线元素；以及用于强影响点检验的 Cook 距离和 W–K 距离

$$\text{Cook}_t = \frac{(\hat{\boldsymbol{\theta}}(t) - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\theta}}(t) - \hat{\boldsymbol{\theta}})}{ks^2}, \ \text{WK}_t = \frac{\mathbf{X}_t'(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}(t))}{s(t)\sqrt{h_{tt}}} \tag{8}$$

其中$\hat{\boldsymbol{\theta}}(t)$表示删除第 t 个数据点后所拟合模型的参数估计值，k 为模型中待估参数的个数。据此可以识别出待评估数据序列中的（严重偏离既定模型的）异常数值点或（对于统计推断具有较大影响的）强影响点。同样，在排除其他因素干扰之后，可以将这些异常数值点（或强影响点）判定为数据质量可疑。周建（2005）针对估计得到的生产函数模型计算了式（6）～式（8）的几类统计量，除发现个别年份的样本点为强影响点或高杠杆点外，拟合误差基本都在可接受范围内。刘洪等（2009，2011，2012）、卢二坡等（2010，2011）更加强调对于模型结构和估计方法的改进，并依据式（7）和式（8）来识别 GDP 数据序列中的异常点，发现在研究区间内部分年份的 GDP 数据为异常（可疑）数据。

三、计量模型评估方法的适用性评析

适用性是针对宏观经济统计数据质量评估方法的一项重要的评价标准。对此，王华、金勇

进（2009）给出了三项考察依据，即评估中所需辅助资料的可获取程度与方法操作的难易程度（可行性），评估结论对于影响统计数据质量的各种因素的揭示程度以及相关信息的实践指导价值（有效性），方法技术假定的合理性以及由此导致误判的可能性（可靠性）。郭红丽、王华（2011）则进一步设计了三项检验内容，即参照指标（辅助资料）应是准确无误的，待评估指标与参照指标之间的一致关联关系是客观存在并且相对稳定的，同时统计数据一致性与准确性两种特征之间也应存在必然的内在逻辑关联，实际涉及已有评估研究所普遍假设（默认）的技术性前提条件。针对计量经济模型之于 GDP 数据质量评估的适用性，同样可以从上述方面加以分析评判。

因篇幅有限，本文忽略对辅助资料可获取性及其准确性的讨论——但不意味着这一因素不会对计量模型评估方法的适用性造成严重影响，[①]专门围绕计量模型评估方法的技术假定的合理性与评估结果的有效性进行检视，这其中又涉及以下几个层面的问题。

（一）模型设定的多样性

计量模型评估方法的应用，其前提是对可以反映待评估指标（于本文即为 GDP）与参照指标之间一致关联关系的特定计量模型（即式（1））的拟合估计；可靠的基础模型，是得出可靠评估结果的必要条件。而问题在于，在计量经济学的方法论发展与应用实践中，虽然强调以明确的理论模型为依归，但具体计量模型的成立往往又依赖于很多假设条件，在拟合估计时需要基于不同假设施加或多或少的处理，如非线性模型的线性化、解释变量的选取、工具变量的引入、针对特定误差结构的加权调整等；这导致计量经济学研究（的模型设定）普遍存在很大的不确定性，其实践效果非常依赖于研究者的主观洞察力与模型设定技巧。进一步，针对同一经济问题（变量间的一致关联关系）还可能存在多种理论逻辑，从而导出多种理论模型与计量模型。在此情况下，基于（某一）计量模型评估其中（被）解释变量的数据质量，难免会产生二分对立的解读路径：不合理的模型参数估计值或异常的拟合残差，既可能是源于部分变量的数据质量不佳，也可能是由于计量模型本身的结构设定不当，未能准确反映变量之间的关联关系。如此，计量经济模型之于统计数据质量评估的应用功效已然成疑。

任若恩（2002）曾对孟连、王小鲁（2000）的研究提出过类似质疑，认为后者基于对全要素生产率的估计来发现经济增长率中统计误差的研究方法是“完全不能接受的”。因为全要素生产率增长率的计算取决于所使用的方法和数据两类要素，两类要素的结合（不同的数据来源和不同的数据处理方法）可能产生非常不同的估计结果；通常的全要素生产率研究是首先分析方法和数据中的问题，据此判断全要素生产率估计是否存在误差，而孟连、王小鲁（2000）则是试图通过观察全要素生产率的异常来分析产出增长率的误差，其逻辑是颠倒的。

实际上，在已有研究结果中存在的诸多分歧（简要如表 1 所示），很大程度上也与相关研究采用的多样化模型设定有关。对于模型式（1）的具体结构，已有文献进行了广泛尝试，涉及组合模型、匹配模型、面板数据模型、稳健主成分回归模型、数据删除模型、动态（时间序列）模型、向量自回归模型、半参数模型、（非）参数面板模型、面板门限模型等多种技术分支；这些模型固然都依循某种经济学理论机制（因而都具有一定的合理性），但适用情形毕竟有所差异。不同模型结构的相对适用性既不明确，针对同样的 GDP 数据集如果得出不同的评估结果，自然也就难以判定这些不同结果之间的相对有效性。

① 岳希明等（2005）认为：“与国内生产总值及其增长速度的官方估计值相比，其他统计数据是否更加准确、偏差更小，我们很难得出确切的结论。但是让这些指标能够完全摆脱由于政府干预，以及由于统计制度不完备等诸多因素所造成的统计偏差是不可能的。”

表 1　应用计量模型方法评估 GDP 数据质量的部分研究结果[a]

	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00[b]
Adams & Chen（1996）	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑	↑						
孟连、王小鲁（2000）															↑	↑	↑	↑	↑	↑			
阙里、钟笑寒（2005）																			↑				
周建（2005）							↑									↓							
刘洪、黄燕（2009）	↓						↑	↑	↑					↓									
卢二坡、黄炳艺（2010）				↓	↓								↓										
刘洪、昌先宇（2011）				↓			↑		↓		↑	↓			↑							↓	↑
刘洪、金林（2012）														×			×						
徐康宁等（2015）[c]															↑	↑	↑	↑	↑	↑	↑	↑	↑

注：a）本表主体引自郭红丽、王华（2017），并有所更新；b）除最后一项研究外，其余各项研究的评估区间虽可能超过2000年，但评估结果表明数据可疑的时间点均在2000年之前，故本表略去了2000年之后的部分；c）徐康宁等（2015）的评估区间为1992−2012年，各年结论均为高估。↑表示被判定为具有正向误差，↓表示具有负向误差；×表示仅判定具有误差，但未确定具体误差方向。

（二）GDP 核算误差机制的复杂性

计量模型评估方法的隐含推断逻辑是，如果利用不准确的基础统计数据（如 GDP）来拟合计量模型，就会得出不合理的模型参数估计值与异常的模型拟合残差，因此可以根据参数估计值与残差拟合值的“异常”状况，来反推基础统计数据的不准确程度。但考虑到 GDP 数据作为实际统计产品，其中核算误差的产生具有深刻的方法、制度与社会根源，遵循复杂的发生机制，对此若不加以细致考察，极有可能会混淆且误读 GDP 核算误差对于计量模型拟合估计结果的影响效应。任何一项旨在评估 GDP 数据质量的研究，也唯有在揭示 GDP 核算误差的发生来源与作用机制方面有所作为，有助于对官方数据给出恰当的修订策略，才能切实发挥其实践功效。然而，现有研究普遍忽视了 GDP 核算误差机制的存在性及其复杂性，对于有偏参数估计值或异常残差的出现是否一定意味着 GDP 数据质量不佳（以及 GDP 数据中存在何种误差）缺乏直面论述，因此也难免于对计量模型估计结果的误读倾向。

专就中国 GDP 数据的核算过程而言，其在调查方案设计、数据采集与统计估算等不同环节存在诸多误差因素，核算误差可能遵循多种发生机制（郭红丽、王华，2017）：例如因统计覆盖范围方面的缺陷造成 GDP 核算在较长时期都存在大面积的漏算，从而产生严重的覆盖误差（Wu，2000；Keidel，2001；Xu，2002）；政府统计能力薄弱导致基层统计（原始微观数据的采集登记环节）操作不规范、人为干预盛行——典型如周黎安（2007）所谓的“晋升锦标赛”现象——造成系统性的操作误差；后期的数据加权汇总与指标估算过程中，因所依据统计资料的不完备而产生的估算误差（Maddison，1998；Wu，2000，2002；Xu，2002；Shiau，2005）。不同类型的核算误差会对 GDP 数据集的实际表现造成不同影响（如持续性低估或间歇性偏误），而各类误差效应的相互叠加，也势必对模型式（1）的估计结果产生复杂影响。此时，要根据计量模型估计结果的异常特征来反推 GDP 数据序列中的误差结构及其综合效应，其中存在的对应性、辨识度等难题一时都还无法解决。

在现有研究中，周建（2005）、刘洪等（2009，2011，2012）、卢二坡等（2010，2011）对于 GDP 时间序列长期趋势的拟合，以及阙里、钟笑寒（2005）、徐康宁等（2015）、卢盛峰等（2017）对于 GDP 空间面板结构的考察，目标都是寻找 GDP 时间序列数据和空间面板数据中被人为操纵的证据，都关注于由不合理的政府统计管理体制所滋生的统计干扰和数据造假，其结果

则是识别出 GDP 数据集中存在异常的个别时点（时段）或个别地区。但正如上文所述，操作误差只是诸多核算误差的一个来源，在各类误差因素的综合作用下，操作误差效应未必能如期望的那样表现为拟合结果中的异常数值点。Klein & Ozmucur（2003）关注于 GDP 误差的综合效应，但预设的参数取值的合宜标准却过于宽松了——试问 GDP 增长率的官方数据要扭曲到何种程度，才能让回归参数表现为不符合经济规律（呈负相关）？毫不奇怪，阙里、钟笑寒（2005）、周国富、连飞（2010）、Mehrotra & Pääkkönen（2011）以及 Fernald et al（2013）基于同样的方法逻辑，检验中国（地区）GDP 增长率数据与其他经济数据的一致性，结果也都未能发现 GDP 数据失真的系统证据。

（三）残差分析的技术局限

再就异常数值识别方法（残差分析）的应用而言，在实际评估过程中可能存在如图 1 所示的几种误判可能性。

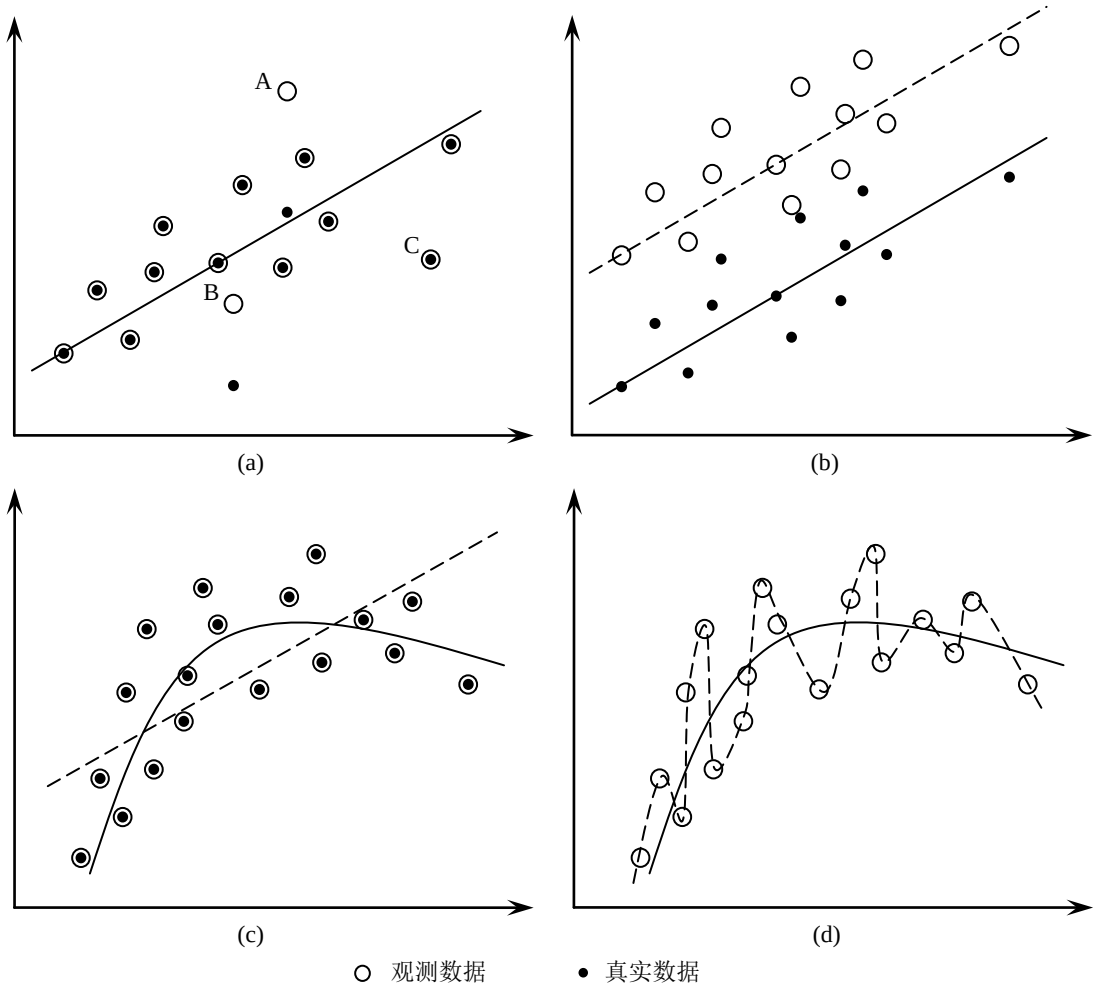


图 1　异常数值识别方法的几种误判情形

在图 1(a)中，对于数据点 A，因较大的统计误差而显现为数据集中的异常点，这是将异常数值识别方法应用于 GDP 数据质量评估的基本目标设定；对于数据点 B，虽然同样存在较大的统计误差，却隐藏于正常数据集当中，难以被识别出来；对于数据点 C，其本身并无统计误差，但因客观统计分布规律而体现为异常数值点。更严重的问题则如图 1(b)所示，所有数据点都存在系统性的统计误差，此时数据集以及据此拟合的回归线将整体位移，但却无法从中识别出任何异常点——由前述讨论的 GDP 核算误差发生机制来看，这种情形极具现实性！

可以认为，异常数值识别方法的功效基本体现为识别“正常”数据集中的“害群之马”，但对于“异常”数据集却是“法不择众”，缺乏必要的评估功效。

为改进基于计量模型的异常数值识别方法的评估效果，卢二坡、黄炳艺（2010）、卢二坡、张焕明（2011）主张采用稳健回归方法，以避免数据“污染”的不良影响；刘洪、昌先宇（2011）、刘洪、金林（2012）则致力于计量模型的结构改进，分别采用了含隐变量的状态空间模型和半参数模型。但这些改进研究恰恰凸显了该评估方法的另一重大问题：既然利用拟合残差来反映 GDP 数据的质量水准，而残差大小又取决于模型拟合的程度（二者存在此消彼长的关系），因此可知，当拟合程度不足时（如图 1(c)中的虚线所示），往往易于发现较多的异常点；而当采用更复杂估计方法以提高拟合程度、甚至达成 100%拟合时（如图 1(d)中的虚线所示），则又不再有任何残差与异常点存在。反向观之，某项研究若识别出较多（或较大幅度）的异常点，首先其计量模型的可用性（拟合效果）难免令人怀疑；而若识别出较少的（甚至没有）异常点时，则又可质疑其存在过度拟合。之所以存在这种“拟合悖论”，关键症结在于，在追求更高模型拟合程度与追求识别出更多异常点之间，难以确定一条合理的界限；实际也不可能存在这样一条明确的界限！

即使是参数可靠性分析方法，结合前文关于模型设定不确定性的讨论可知，其在应用中也会存在类似的“拟合悖论”问题，此处不再赘述。

总体而言，在将计量模型方法应用于评估 GDP 数据质量时，其潜在的技术假定过于理想化了，极有可能因不满足相关技术假定而造成多方面的误判。与 Rawski（2001）等采用的短时期单一指标交叉检验方法相比，计量模型评估方法属于在更广时空范围内的综合交叉检验，模型拟合结果往往呈现多项变量之间较好的一致性与较少的异常残差，对 GDP 数据中存在的连续性、系统性统计偏误的识别能力不足，由变量间的表象一致而导致“纳伪”误判的可能性因而更大——当然也不排除存在将统计核算制度转换或社会变革导致的数据突变归咎于 GDP 数据质量不佳的“弃真”可能。

四、数值模拟分析

为了对上述分析给予更具可视化效果的论证，本节进一步遵循正向逻辑，建立反映 GDP 现实数据生成过程（其中涵括多种 GDP 核算误差特性）的测量误差模型，运用数值模拟方法检验计量模型方法在揭示 GDP 核算误差发生机制方面的有限功效。

（一）模型设定

不失一般性，假设目标变量Y的真实统计数据服从如下生成过程：

$$Y_t^* = \alpha_1 Y_{t-1}^* + \alpha_2 X_t + u_t^* \tag{9}$$

其中，α_1为目标变量的一阶自回归系数，α_2为外生变量X的影响系数，另有随机干扰项$u_t^* \sim N(0, \sigma_{u^*}^2)$。①

考虑测量（核算）误差的存在，令现实的统计数据服从如下生成过程：

$$Y_t = Y_t^* + \varepsilon_t \tag{10}$$

以郭红丽、王华（2017）的讨论为基础，将测量误差ε_t的发生机制刻画如下：

① 本节从这样一个服从明确数据生成过程的变量出发，可以规避前文提及的模型误设问题，从而集中讨论真实数据之外测量误差机制的影响。由于模型设定与测量误差机制（对于计量模型方法适用性）的影响可以分别予以考察，不论式（9）能否直接对应于 GDP 数据的潜在生成过程，都不会影响本节数值模拟结果的成立。同时，式（9）作为一类数据集合，也可以在很大程度上涵盖 GDP 数据的生成模式。

$$\varepsilon_t = \gamma_1\varepsilon_{t-1} + \gamma_2\Delta X_t + \gamma_3 X_t + v_t,\ v_t \sim N(0,\sigma_v^2) \tag{11}$$

上式中，$\gamma_1(>0)$为测量误差的一阶自回归系数，反映由于制度和方法原因而产生的误差自相关机制；$\gamma_2(<0)$为相对于外生变量波动的误差修正系数，反映统计部门对统计数据中表现出的异常波动施加平滑处理的“人为”机制；$\gamma_3(>0)$为外生变量的影响系数，反映由社会经济系统的客观现实因素而形成的误差影响机制。在以往研究中，刘小二、谢月华（2009）以 GDP 序列对其增长率序列的回归系数来反映 GDP 核算的“适应性预期调整”机制，承载了式（11）中γ_2的类似信息；孙艳、贡颖（2013）以 GDP 修订额对其初步核算数据的回归系数来反映实时数据的有效性，承载了γ_3的类似信息；Sinclair（2019）进一步建立了 GDP 修订额对其滞后项的回归方程，其反映的信息则与γ_1基本相同。与上述研究普遍采用的一元回归模型相比，本文模型同时整合了 GDP 核算误差的各种因素，可以更有效复制前文提及的估算误差、操作误差和覆盖误差的发生机制。

现在考虑利用现实统计数据拟合以下计量模型：

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + u_t,\ u_t \sim N(0,\sigma_u^2) \tag{12}$$

而结合式(9)−式(11)可知，应有

$$Y_t - \varepsilon_t = \alpha_1(Y_{t-1} - \varepsilon_{t-1}) + \alpha_2 X_t + u_t^*$$

$$\begin{aligned} Y_t &= \alpha_1 Y_{t-1} + \alpha_2 X_t + (\gamma_1\varepsilon_{t-1} + \gamma_2\Delta X_t + \gamma_3 X_t + v_t) - \alpha_1\varepsilon_{t-1} + u_t^* \\ &= \alpha_1 Y_{t-1} + (\alpha_2 + \gamma_2 + \gamma_3)X_t - \gamma_2 X_{t-1} + (\gamma_1 - \alpha_1)\varepsilon_{t-1} + v_t + u_t^* \end{aligned}$$

对其中的ε_{t-1}迭代展开，转换可得

$$\begin{aligned} Y_t = {} & (\alpha_1 + \gamma_1)Y_{t-1} + (\alpha_2 + \gamma_2 + \gamma_3)X_t - (\gamma_2 + \alpha_2\gamma_1 + \alpha_1\gamma_2 + \alpha_1\gamma_3)X_{t-1} \\ & + \alpha_1\gamma_2 X_{t-2} - \alpha_1\gamma_1 Y_{t-2} + (v_t - \alpha_1 v_{t-1}) + (u_t^* - \gamma_1 u_{t-1}^*) \end{aligned}$$

因此，若根据真实的数据生成过程，含测量误差的回归模型实际应该为

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + \beta_5 Y_{t-2} + u_t \tag{13}$$

其中$\beta_1 = \alpha_1 + \gamma_1$，$\beta_2 = \alpha_2 + \gamma_2 + \gamma_3$，$\beta_3 = -(\gamma_2 + \alpha_2\gamma_1 + \alpha_1\gamma_2 + \alpha_1\gamma_3)$，$\beta_4 = \alpha_1\gamma_2$ 和$\beta_5 = -\alpha_1\gamma_1$，并且有$u_t = (v_t - \alpha_1 v_{t-1}) + (u_t^* - \gamma_1 u_{t-1}^*)$。

由上可知，直接对模型式（12）进行估计将存在多项问题，如遗漏解释变量和被解释变量的滞后项，解释变量与随机干扰项相关，且随机干扰项本身也存在自相关性，因而导致估计结果的严重偏误。

暂且忽略上述问题（甚至假设不存在可能的估计偏误），考虑基于对模型式（12）的拟合估计来评估变量Y的数据质量。具体的考察对象（模拟实验的输出变量）有：

（1）回归系数β_1和β_2。依据参数可靠性分析的基本原理，考察在不同测量误差机制下，回归系数的估计值是否会严重偏离理论值α_1和α_2，尤其是其符号是否会出现反向变异。

（2）由式（6）给出的相对误差率P_t。

（3）由式（7）给出的异常点检验参数学生化残差r_t与t_t。

（4）由式（8）给出的强影响点检验参数 Cook 距离与 W–K 距离。依据异常数值识别方法的基本原理，考察在不同测量误差机制下，标准(2)−(4)的取值是否与测量误差机制存在内在关联。

同时，模拟实验的输入变量（对应于不同的测量误差发生机制）则包括：

（1）模型（9）或（12）中随机干扰项的方差$\sigma_{u^*}^2$或σ_u^2，二者影响模型的拟合程度。

（2）模型（12）的回归系数β_1和β_2，若分别令其为 0，可影响模型的实际结构。

（3）模型（11）的随机误差方差σ_v^2，决定测量误差的随机扰动幅度。

（4）模型（11）中的一阶自回归系数γ_1，决定测量误差在制度和方法层面的路径依赖程度。

（5）模型（11）中外生波动的修正系数γ_2，决定统计测量主体对于测量结果中异常变动的平滑处理程度。

（6）模型（11）中外生变量的影响系数γ_3，决定测量误差的客观水平。

基于上述设定，根据不同的模拟情境对输入变量（参数）赋予不同数值，据此拟合模型式（12），即可相应考察在不同的测量误差机制下输出变量的分布情况，从中明确输出变量是否对于不同测量误差机制具有一定的反应敏感性。

简单起见，首先令反映真实数据生成过程的式（9）中的一阶自回归系数$\alpha_1 = 0.8$，外生变量影响系数$\alpha_2 = 0.5$；令外生变量X服从以下指数增长过程：

$$X_t = (1 + g_t)X_{t-1},\ g_t \sim N(0.05, 0.05^2)$$

并且有初始时点的变量取值分别为$X_0 = 100$和$Y_0 = 100$。

在模拟实验中，令样本时长为 100 期，在不同情境下的模拟次数均为 1000 次。

（二）模拟实验情境 1：无测量误差

本小节主要检验在无测量误差下模型式（12）的拟合表现。令$\gamma_1 = \gamma_2 = \gamma_3 = 0$，$\sigma_v^2 = 0$，此时式（10）中的测量误差$\varepsilon_t$为 0，式（12）即等同于式（9），故而对于模型拟合效果的影响主要来自于模型随机干扰项的方差$\sigma_u^2 = \sigma_{u^*}^2$。以下则具体设定$\sigma_u^2$的不同数值，检验在模型拟合度与异常数值识别能力之间是否存在此消彼长关系。同时，也专门将式（12）中的Y_{t-1}和X_t分别予以删除，考察在遗漏重要解释变量（模型结构误设）的情况下，异常数值识别方法的应用会受到何种影响。

首先令$\sigma_u^2 = 50^2$，由此随机生成研究变量 X 与 Y 的一套样本数据如图 2 所示。由图 2 可知，X 与 Y 的时间序列符合宏观经济指标的通常形态，可用以进行相关的模拟实验研究。

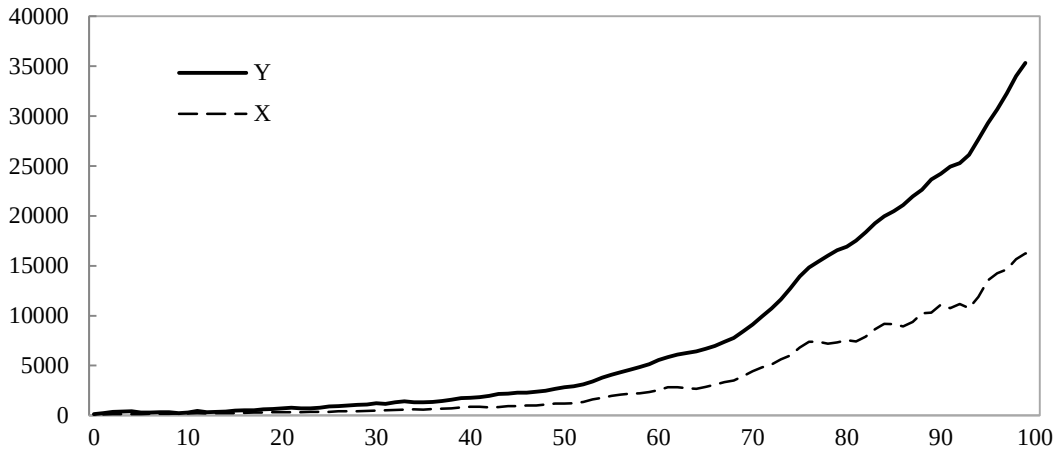
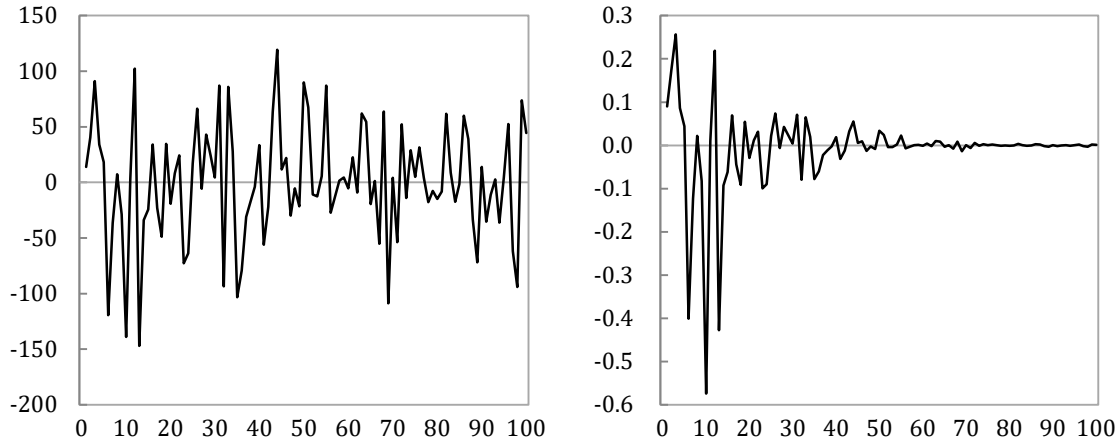


图 2　关于 X 与 Y 的一套随机样本序列

利用一套样本数据进行回归拟合，由此计算得到的回归残差序列和相对误差率序列可如图 3 所示。由图 3 可知，对于带有趋势的时间序列而言，虽然模型的随机干扰项满足同方差条件（如图 3(a)所示），但相对误差率$P_t = \hat{u}_{it}/Y_{it} = (Y_{it} - \hat{Y}_{it})/Y_{it}$却会随着模型被解释变量$Y$的取值的增大而相应减小（如图 3(b)所示）。由此，在样本时段前期的相对误差率往往较大，而样本时段后期的相对误差率则往往较小。专就从时间序列$\{Y_t\}$中识别异常值这一目标而言，相对误差率显然不是一个可靠的评判标准；至少不能直接根据该指标的大小与否，来判断目标

变量Y的取值是否存在质量问题。



(a) 模型残差　　(b) 相对误差率

图 3　模型拟合残差及其相对误差率

进一步，改变σ_u的取值（具体从 5 到 200），针对每套随机样本计算得到回归系数β_1和β_2的估计值，100 个时点的相关误差率P_t的标准差，以及异常点诊断统计量r_t与t_t、强影响点诊断统计量Cook_t与WK_t的标准差——可以预期，这些参数的标准差越大，从 100 个时点中识别出异常值的可能性也就越大。表 2 中列出了在各种情况下，基于 1000 套随机样本的相关模拟数值的平均数。分别删除模型（12）中的解释变量Y_{t-1}或X_t后（令$\sigma_u = 50$），所得参数的平均值亦列于表中。

表 2　无测量误差下的模拟实验结果

模型	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{S}(P)$	$\bar{S}(r)$	$\bar{S}(t)$	$\bar{S}(\text{Cook})$	$\bar{S}(\text{WK})$
$\sigma_u = 5$	0.8000^{***}	0.5001^{***}	0.0087	1.0001	1.0160	0.0285	0.1842
$\sigma_u = 10$	0.7999^{***}	0.5003^{***}	0.0174	1.0000	1.0159	0.0284	0.1834
$\sigma_u = 20$	0.7995^{***}	0.5009^{***}	0.0362	0.9999	1.0156	0.0278	0.1826
$\sigma_u = 30$	0.7992^{***}	0.5015^{***}	0.0576	0.9998	1.0156	0.0273	0.1821
$\sigma_u = 50$	0.7991^{***}	0.5019^{***}	0.3667	1.0001	1.0157	0.0280	0.1834
$\sigma_u = 100$	0.7941^{***}	0.5118^{***}	2.2546	1.0000	1.0158	0.0271	0.1828
$\sigma_u = 150$	0.7891^{***}	0.5220^{***}	4.6913	0.9998	1.0159	0.0261	0.1824
$\sigma_u = 200$	0.7843^{***}	0.5310^{***}	6.0117	1.0000	1.0159	0.0256	0.1821
$\beta_1 = 0$	0	2.0990^{***}	0.5216	1.0126	1.0644	0.0999	0.2334
$\beta_2 = 0$	1.0484^{***}	0	0.2829	1.0108	1.0521	0.0872	0.2211

注：***表示在 0.01 水平上显著不为 0。

由表 2 结果可知，随着模型随机干扰项的标准差逐渐增大，模型的拟合程度越来越差，在此过程中：（1）因模型中包含滞后被解释变量而导致的回归系数估计偏差渐趋增大，但均显著不为 0，亦不存在符号错误；（2）相对误差率的标准差$S(P)$迅速增大，其中出现异常数值的概率相应激增；（3）四项诊断统计量的标准差$S(r)$、$S(t)$、$S(\text{Cook})$与$S(\text{WK})$没有明显变化，异常值的出现概率不变。据此判断，既然目标变量Y的统计数据中并不存在测量误差，以相对误差率来识别数据序列中异常值（并将之等同于测量误差）的评判标准显然有误——一般情况下，若估计结果中出现较大的相对误差率，完全可能是源于回归模型本身的拟合程度较低，

而与数据质量无关！同时，四项诊断统计量表现出较好的稳健性，不会因拟合程度的不同而生成错误信息。

另一方面，在删除模型中的重要解释变量Y_{t-1}或X_t（也即设定$\beta_1 = 0$或$\beta_2 = 0$）后，模型结构发生重大改变，相关估算结果因而也表现出显著变化：（1）回归系数估计值严重偏离真实参数（但其符号仍未改变，也仍然显著不为 0）；（2）四项诊断统计量（尤其是距离统计量）的标准差明显增大，由此可识别出更多的、可疑的异常值。据此判断，对于模型结构的错误设定，既会导致参数估计结果的偏误，也会导致统计诊断结果的异常，从而对测量误差的准确识别造成误导；反言之，即使利用统计诊断方法识别出实际模型拟合结果中较多的异常点（强影响点），也应该首先审视模型结构设定（或估计方法）的合理性，检讨是否存在拟合不足的情况，而不能直接判定其与指标数据中的测量误差有关。

（三）模拟实验情境 2：随机测量误差

本小节主要检验测量误差发生机制中的随机因素对于模型式（12）拟合结果的影响。仍令$\gamma_1 = \gamma_2 = \gamma_3 = 0$，且有$\sigma_{u^*}^2 = 50^2$，此时式（10）中的测量误差$\varepsilon_t$完全由其随机干扰项$v_t$决定，表现为随机性（没有确切发生机制和固定方向）的测量误差。与式（9）相比，式（12）的区别仅在于$u_t = v_t + u_t^*$，即随机干扰项的方差有所增大；预期其对模型拟合以及异常数值识别效果的影响，与上一小节中σ_u^2增大的效果相当。以下具体设定σ_v的不同数值，考察其可能造成的不同影响。

表 3　不同测量误差机制下的模拟实验结果

模型	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{S}(P)$	$\bar{S}(r)$	$\bar{S}(t)$	$\bar{S}$(Cook)	$\bar{S}$(WK)
$\sigma_v = 10$	0.7986***	0.5028***	0.3508	1.0001	1.0159	0.0286	0.1840
$\sigma_v = 25$	0.7945***	0.5110***	0.6883	1.0001	1.0161	0.0285	0.1841
$\sigma_v = 50$	0.7860***	0.5282***	0.7465	1.0004	1.0162	0.0294	0.1855
$\sigma_v = 100$	0.7522***	0.5961***	2.4269	1.0012	1.0171	0.0337	0.1912
$\gamma_3 = 0.05$	0.7890***	0.5345***	0.4807	1.0005	1.0164	0.0302	0.1865
$\gamma_3 = 0.10$	0.7784***	0.5691***	0.3340	1.0016	1.0179	0.0354	0.1941
$\gamma_3 = 0.20$	0.7589***	0.6377***	0.1662	1.0050	1.0235	0.0512	0.2152
$\gamma_3 = 0.50$	0.7028***	0.8601***	0.0980	1.0136	1.0455	0.0890	0.2643
$\gamma_2 = -0.05$	0.8045***	0.4905***	0.2143	1.0022	1.0188	0.0382	0.1979
$\gamma_2 = -0.10$	0.8090***	0.4810***	0.2870	1.0067	1.0266	0.0574	0.2246
$\gamma_2 = -0.20$	0.8173***	0.4633***	0.3278	1.0145	1.0465	0.0906	0.2680
$\gamma_2 = -0.50$	0.8285***	0.4380***	0.6526	1.0264	1.0882	0.1439	0.3318
$\gamma_1 = 0.01$	0.7979***	0.5043***	0.2722	1.0001	1.0159	0.0283	0.1836
$\gamma_1 = 0.05$	0.7976***	0.5047***	0.5383	1.0000	1.0159	0.0281	0.1833
$\gamma_1 = 0.10$	0.7977***	0.5046***	0.3600	1.0000	1.0159	0.0282	0.1835
$\gamma_1 = 0.30$	0.7978***	0.5043***	0.2712	1.0000	1.0158	0.0280	0.1831
$\gamma_1 = 0.50$	0.7980***	0.5041***	0.3509	0.9999	1.0158	0.0279	0.1828
$\gamma_1 = 0.90$	0.7982***	0.5037***	0.2980	1.0000	1.0159	0.0280	0.1830

注：***表示在 0.01 水平上显著不为 0。

由表 3 中第一部分的模拟实验结果可知，随着测量误差（随机干扰项）的标准差σ_v逐渐增大，各项参数的变化与表 2 结果非常相似：（1）回归系数的估计偏误逐渐增大，但仍显著不为 0；（2）相对误差率的标准差$S(P)$迅速增大，由其识别出异常数值的概率因而相应增大；（3）

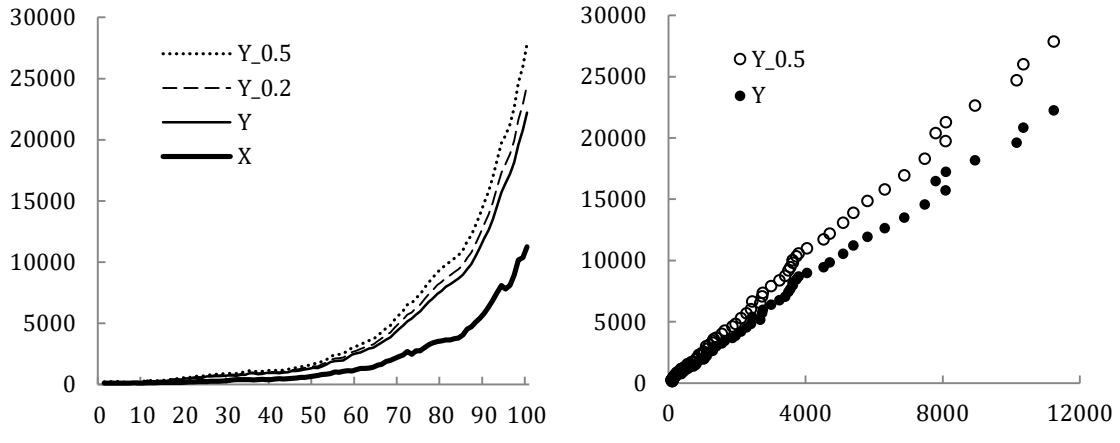
四项诊断统计量的标准差$S(r)$、$S(t)$、$S(\text{Cook})$与$S(\text{WK})$没有明显变化，显示异常值的出现概率不变。由此可见，随机测量误差v_t对于模型拟合效果的影响，最终完全叠加于模型本身的随机误差项u_t之上，二者实际上无法加以区分；换言之，即使可以据此识别出个别时点取值的异常性，也无法判断它是源于模型本身的未解释因素，还是源于指标数据的测量误差。同时，由于随机测量误差并不必然突出表现于某一（些）特定时点，实际识别出的在个别时点上的异常数值显然不能反映其潜在的误差发生机制，反而会导致实践中的误判。

（四）模拟实验情境 3：客观测量误差

本小节主要检验测量误差发生机制中的客观因素对于模型式（12）拟合结果的影响。令$\gamma_1 = \gamma_2 = 0$，且有$\sigma_v^2 = 10^2$，$\sigma_{u^*}^2 = 50^2$，此时式（10）中测量误差ε_t的变化取决于客观因素 X 的影响系数γ_3，对其结果可简称之“客观测量误差”。以下具体设定γ_3的不同数值，考察其可能造成的不同影响。

由表 3 中第二部分的模拟实验结果可知，随着γ_3取值逐渐增大，社会经济系统的客观因素对于测量误差的影响越来越大，在此过程中：（1）γ_3对于 Y 的影响完全叠加于β_2（由式（13）可知），导致回归系数的估计偏误迅速增大，但仍显著不为 0；（2）（被忽略的）外生变量滞后项X_{t-1}对于模型随机干扰项施加负向影响，导致相对误差率的标准差$S(P)$相应减小；（3）四项诊断统计量的标准差$S(r)$、$S(t)$、$S(\text{Cook})$与$S(\text{WK})$有所增加，据此识别出异常数值点的概率增大。与模拟情境 1 中遗漏解释变量的结果（见表 2 第二部分）相比，情境 3 中由γ_3增大导致诊断统计量标准差的增大，（即使是在$\gamma_3 = 0.50$的较高水平下也）并未表现得更加突出。这意味着，即使能够依据诊断统计量识别出更多的异常值，实际也无法判断其原因到底是测量误差的客观发生机制，还是对于模型结构的错误设定。

同时需要注意的是，客观误差机制导致时间序列 Y 产生了系统性的测量误差。如图 4(a)所示，随着γ_3取值的增大，Y 序列的水平也相应整体提升，显示测量误差序列存在系统性的正向偏倚。图 4(b)中 Y 对 X 的散点图则表明，对于较大的γ_3取值，Y–X 实测点集相对于真实点集的偏离程度也会较高；此时数据质量评估的重点已经不在于（以实测点集为基准的）异常数值识别，而是对这种系统性误差的揭示，但包括参数可靠性分析在内的现有方法对此仍无能为力。



(a)时序图　　(b)散点图

图 4　不同客观误差机制下真实数据与实测数据的关系

（五）模拟实验情境 4：自平滑测量误差

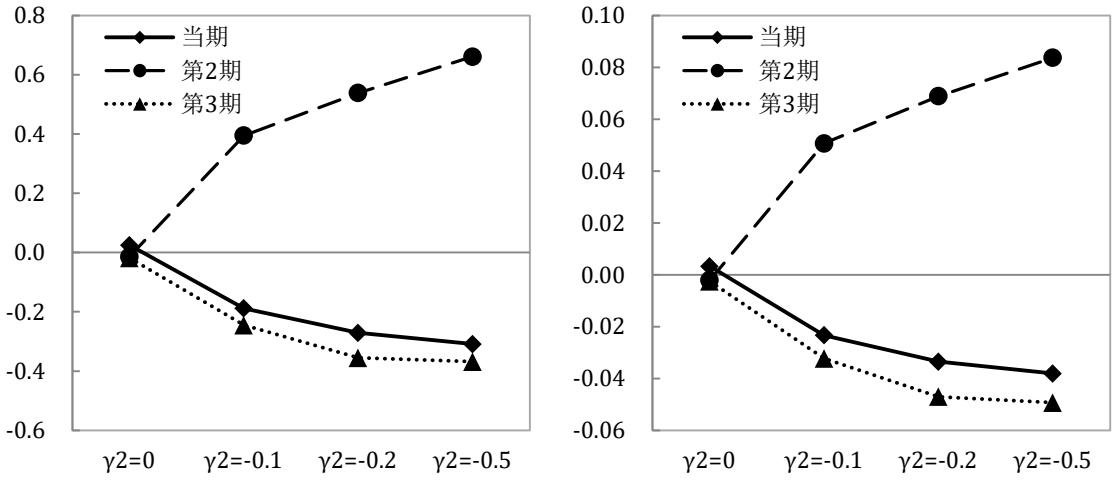
本小节主要检验测量误差的“自平滑”机制对于模型式（12）拟合结果的影响。令

$\gamma_1 = \gamma_3 = 0$，且有$\sigma_v^2 = 10^2$，$\sigma_{u^*}^2 = 50^2$，此时式（10）中测量误差ε_t的变化取决于对外生波动的修正系数γ_2；也即当外生变量 X 出现波动时，针对目标变量 Y 的测量误差存在一个反向调整机制，以使 Y 本身的变动显得相对和缓（对其结果可简称为“自平滑测量误差”）。以下具体设定γ_2的不同数值，考察其可能造成的不同影响。

由表 3 中第三部分的模拟实验结果可知，随着γ_2取值逐渐增大，针对社会经济系统外生变量的特定波动，测量误差的反向修正力度也越来越大，在此过程中：（1）回归系数的估计偏误逐渐增大（与其他情形下的偏误方向相反），但仍显著不为 0；（2）相对误差率的标准差$S(P)$有所增大，其中出现异常值的概率因而相应变大；（3）四项诊断统计量的标准差$S(r)$、$S(t)$、$S(\text{Cook})$与$S(\text{WK})$有所增加，据此识别出异常数值点的概率增大。与存在客观测量误差下的情形相似，虽然可以根据诊断统计量识别出更多异常值，但却无法判断是源于测量误差的自平滑机制，还是源于模型结构的错误设定。

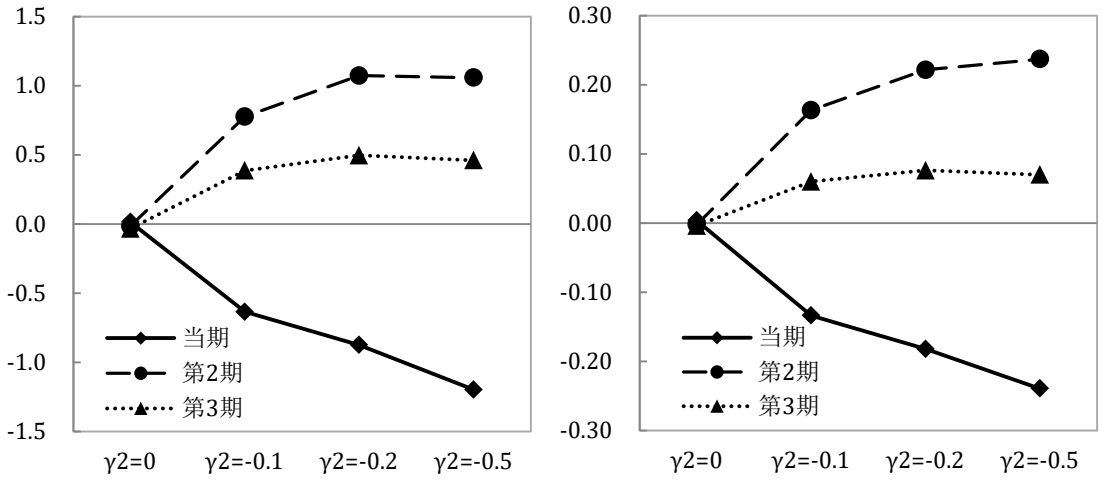
鉴于γ_2的取值变化主要作用于 X 的波动点，本研究针对 X 序列专门设置了两个异常波动点，其中一个是在其原水平上附加 50%幅度的加性野值，另一个则附加 50%幅度的革新野值，由此考察在不同的γ_2取值下诊断统计量学生化残差 t 和 W–K 距离的变化，具体结果如图 5 和图 6 所示。由图 5 可知，针对发生 50%幅度剧烈波动的加性野值点，当存在测量误差的自平滑机制时，两项诊断统计量在当期都表现为负值，其绝对水平也随γ_2的增大而增大；根据式（13），X 对 Y 实际存在滞后 2 期的影响，因此在加性野值点之后的第 2 期和第 3 期，诊断统计量同样表现突出，但分别显现为正值和负值（由式（13）中β_3和β_4的符号可知其理）。由图 6 可知，针对发生 50%幅度剧烈波动的革新野值点，两项诊断统计量在当期和第 2 期仍然都表现为负值和正值，但第 3 期的表现则转变为正值，这是由革新野值点的累积影响所致；同样因为存在累积影响，图 6 中两项诊断统计量的表现相比图 5 都要更为突出。

由上述分析可知，根据各项诊断统计量确实可以识别出在特定异常数值点上自平滑测量误差机制的影响。但应该注意的是，即使诊断统计量在连续的多个时点上都表现异常，实际的测量误差也可能只是发生在初始时点上，后续时点的异常表现则完全是源于目标变量 Y 的数据生成过程的自相关性。



(a) t 残差　　(b) W–K 距离

图 5　加性野值点下修正系数γ_2的影响



(a) *t* 残差　　(b) W–K 距离

图 6　革新野值点下修正系数γ_2的影响

（六）模拟实验情境 5：自相关测量误差

本小节主要检验测量误差的自相关性对于模型式（12）拟合结果的影响。令$\gamma_2 = \gamma_3 = 0$，且有$\sigma_v^2 = 10^2$，$\sigma_{u^*}^2 = 50^2$，此时式（10）中测量误差ε_t的变化取决于其滞后项的影响系数γ_1；γ_1越大，测量误差的自相关性（时间惯性）也会越大。以下具体设定γ_1的不同数值，考察其可能造成的不同影响。

由表 3 中第四部分的模拟实验结果可知，随着γ_1取值逐渐增大，测量误差的跨期自相关系数相应增大，模型（12）中干扰项的自相关性也越来越强，但在此过程中，回归系数估计结果、相对误差率以及诊断统计量的表现都未出现明显变化。可见，针对测量误差的自相关机制，现有基于计量模型的数据质量评估技术（考察标准）全部归于无效！

（七）小结

本节基于一个反映 GDP 核算误差机制的测量误差模型进行数值模拟实验，以考察计量模型方法在评估 GDP 数据质量方面的可能功效，尤其注重检讨此类方法能否有效揭示 GDP 核算误差的不同来源及其特定的影响效应。

实验结果表明：第一，各种误差机制确实会导致模型参数估计结果的偏误，但由于真实参数未知，在实际评估中这种偏误因而也是不可测的；即使是施加较强幅度的误差影响，参数估计结果也只是产生中等程度的偏误，通常也不会出现明显的符号错误或者超出正常取值范围，基于计量模型的参数可靠性分析因而只能给出非常保守的结论。第二，相对误差率不是一个可靠的评判标准，因为其取值大小同时受到模型拟合残差与被解释变量取值水平两方面的影响；同时，模型结构设定的合理性与模型整体的拟合程度也会对相对误差率造成显著影响，易于产生对数据质量的误判。第三，对于客观因素造成的测量误差和具有自平滑机制的测量误差，四项诊断统计量具有一定的识别能力，但也仅限于识别数据序列中特定时点的异常特征，对于系统性误差则表现得无能为力；而即使是针对特定时点（时段），也仍然存在误判的可能性。

更有甚者，如果用于数据质量评估的计量模型本身的结构设定有误，既会导致参数估计结果的偏误，也会导致残差诊断结果的异常，最终则造成对测量误差识别的严重误导。考虑到此种情形的难以避免，计量模型评估方法的实际功效显然不容乐观。

五、结论与研究展望

针对中国 GDP 数据质量评估研究中广泛采用的计量经济模型方法，本文从其技术原理的局限性以及 GDP 核算误差发生机制的影响效应两个方面，系统考察了该类方法在具体评估实践中的适用性。研究分析表明，基于计量经济模型的两类评估方法（参数可靠性分析与异常数值识别），在模型设定环节即已面临本文称之为“拟合悖论”的两难困境：从方法论上讲，若不断致力于模型建构的复杂化和估计方法的精细化，计量经济模型可以实现对于统计数据的高度（甚至完全）拟合，如此将不会存在任何不合理的模型参数估计值或者异常的模型拟合残差；考虑到这种可能性，对于计量模型估计结果中实际表现出的参数偏误或残差异常，既可以归因于基础数据的质量不佳，也可以归因于计量模型的拟合不足，这种解读路径上的二分对立，在评估实践中还无法予以协调。进一步，GDP 核算作为一项系统工程，其核算误差可能遵循多种发生机制，而不同类型误差因素（误差成份）的相互叠加会对计量模型估计结果产生复杂影响，此时单纯根据计量模型估计结果的异常特征，还难以反推 GDP 数据序列中的误差结构及其实际产生的综合效应；而模型误设问题更加剧了这种反向逻辑推断的难度——基于数值模拟实验的结果也正说明了这一点。

具体而言，基于计量模型的参数可靠性分析方法主要是参照一些外部的、先验的比较基准，根据模型参数估计结果的“异常”特征来推断 GDP 数据序列中的可能误差模式，但本身无法避免“拟合悖论”所隐含的模型结构设定的不确定性，对于潜在的各种 GDP 核算误差因素也缺乏足够的反应敏感度，其推断结果的可靠性难免不足；若由此估算生成针对官方数据的替代性序列（如 Adams & Chen，1996；孟连、王小鲁，2000），其准确性与可用性并不易保障。基于计量模型的异常数值识别方法旨在识别 GDP 数据序列中的“异常”点（与整体趋势表现出显著偏离的个别年份或个别地区），并将其归因于 GDP 核算过程中特定年份或特定地区的“不当”操作（可对应于本文中的操作误差或自平滑测量误差）。但由 GDP 核算误差机制的复杂性所决定，在 GDP 实际统计数据中，操作误差的影响效应与其他类型的误差效应复合叠加，其综合效应已经不仅仅表现为个别数据点的“异常”特征，而更可能表现为形如图 4 的系统性偏差模式；此时根据异常数值识别方法所能得到的信息将是非常有限的，并且不免产生误导倾向。

综合全文分析可以认为，计量经济模型方法（不论是参数可靠性分析还是异常数值识别）之于 GDP 数据质量评估研究的适用性仍然不强。在未来进一步研究中，为切实改善评估功效，提升计量模型方法的适用性，有必要对 GDP 核算误差机制的识别问题加以正视和重视。在此领域，已有部分研究开展了有益探索，例如孙艳、贡颖（2013）、郭红丽、王华（2017）、Sinclair（2019）针对中国 GDP 的初步核算数据、初步核实数据、最终核实数据乃至历史修订数据之间的偏差模式进行了分析和检验，曾五一、薛梅林（2014）则对中国 GDP 的国家数据与地区数据之间的偏差模式及其产业构成进行了分解分析。在上述研究的基础上，本文建立了一个在反映 GDP 核算误差机制方面相对更为完备的概念模型（如式（11）所示），为进一步的研究拓展提供了可行基准；未来可以结合对中国 GDP 核算误差机制的更深入考察，对该模型的结构进行具体设定，并利用实际统计数据对其参数进行实证检验。这一工作的推进与实现，既有助于直观呈现中国 GDP 数据中的误差效应、构成及其时空关联特征，为加强 GDP 数据质量控制发挥实践指导价值；也可以向所有纳入 GDP 指标的计量经济模型研究提供潜在测量误差信息，为改善含测量误差计量模型估计的可靠性发挥必要的方法论价值。

参考文献：

冯蕾 周晶, 2013：《政府统计数据准确性评估方法述评》，《统计研究》第 6 期。

郭红丽 王华, 2011：《宏观统计数据质量评估的研究范畴与基本范式》，《统计研究》第 6 期。

郭红丽 王华, 2017：《中国 GDP 核算误差的特征事实与发生机制》，《统计与信息论坛》第 7 期。

刘洪 黄燕, 2007：《我国统计数据质量的评估方法研究——趋势模拟评估法及其应用》，《统计研究》第 8 期。

刘洪 黄燕, 2009：《基于经典计量模型的统计数据质量评估方法》，《统计研究》第 3 期。

刘洪 昌先宇, 2011：《基于全要素生产率的中国 GDP 数据准确性评估》，《统计研究》第 2 期。

刘洪 金林, 2012：《基于半参数模型的中国 GDP 数据准确性评估》，《统计研究》第 10 期。

刘小二 谢月华, 2009：《中国分省 GDP 数据诊断分析》，《山西财经大学学报》第 2 期。

卢二坡 黄炳艺, 2010：《基于稳健 MM 估计的统计数据质量评估方法》，《统计研究》第 12 期。

卢二坡 张焕明, 2011：《基于稳健主成分回归的统计数据可靠性评估方法》，《统计研究》第 8 期。

卢盛峰 陈思霞 杨子涵, 2017：《“官出数字”：官员晋升激励下的 GDP 失真》，《中国工业经济》第 7 期。

孟连 王小鲁, 2000：《对中国经济增长统计数据可信度的估计》，《经济研究》第 10 期。

阙里 钟笑寒, 2005：《中国地区 GDP 增长统计的真实性检验》，《数量经济技术经济研究》第 4 期。

任若恩, 2002：《中国 GDP 统计水分有多大——评两个估计中国 GDP 数据研究的若干方法问题》，《经济学(季刊)》第 2 卷第 1 期。

Shiau A., 2005：《中国政府高估了经济增长吗？》，《中国经济增长速度：研究与争论》，中信出版社。

孙艳 贡颖, 2013：《中国季度 GDP 初步数据优良性检验》，《统计研究》第 11 期。

王华 金勇进, 2009：《统计数据准确性评估：方法分类及适用性分析》，《统计研究》第 1 期。

王华 金勇进, 2010：《统计数据质量评估：误差效应分析与用户满意度测评》，中国统计出版社。

徐康宁 陈丰龙 刘修岩, 2015：《中国经济增长的真实性：基于全球夜间灯光数据的检验》，《经济研究》第 9 期。

岳希明 张曙光 许宪春, 2005：《中国经济增长速度：研究与争论》，中信出版社。

曾五一 薛梅林, 2014：《GDP 国家数据与地区数据的可衔接性研究》，《厦门大学学报(哲社版)》第 2 期。

周国富 连飞, 2010：《中国地区 GDP 数据质量评估——基于空间面板数据模型的经验分析》，《山西财经大学学报》第 8 期。

周建, 2005：《宏观经济统计数据诊断：理论、方法及其应用》，清华大学出版社。

周黎安, 2007：《中国地方官员的晋升锦标赛模式研究》，《经济研究》第 7 期。

Adams, F. G. & Y. Chen (1996), “Skepticism about Chinese GDP growth – the Chinese GDP elasticity of energy consumption”, *Journal of Economic and Social Measurement* 22(4): 231–240.

Fernald, J. et al (2013), “On the reliability of Chinese output figures”, *FRBSF Economic Letter* 2013–08.

Keidel, A. (2001), “China's GDP expenditure accounts”, *China Economic Review* 12: 355–367.

Klein, L. R. & S. Ozmucur (2003), “The estimation of China's economic growth rate”, *Journal of Economic & Social Measurement* 28(4): 277–285.

Maddison, A. (1998), *Chinese Economic Performance in the Long Run*, OECD Development Centre.

Maddison, A. & H. X. Wu (2008), “Measuring China's economic performance”, *World Economics* 9(2): 13–44.

Mehrotra, A. & J. Pääkkönen (2011), “Comparing China's GDP statistics with coincident indicators”, *Journal of Comparative Economics* 39(3): 406–411.

Rawski, T. G. (2001), “What is happening to China's GDP statistics?”, *China Economic Review* 12: 347–354.

Sinclair, T. M. (2019), “Characteristics and implications of Chinese macroeconomic data revisions”, *International Journal of Forecasting* 35(3): 1108–1117.

Wu, H. X. (2000), “China's GDP level and growth performance: Alternative estimates and the implications”, *Review of Income and Wealth* 46(4): 475–499.

Wu, H. X. (2002), “How fast has Chinese industry grown? Measuring the real output of Chinese industry, 1949–1997”, *Review of Income and Wealth* 48(2): 179–204.

Xu, X. (2002), “Study on some problems in estimating China's gross domestic product”, *Review of Income and Wealth* 48(2): 205–216.