

2018 年疫情背景下考虑网络舆情因素的猪肉价格预测模型——基于动态面板模型

摘要：我国是猪肉消费大国，稳定猪价是我国重大现实问题。本文首先对理论上影响猪肉价格的变量如猪肉供给量、需求量、替代品价格、猪价相关舆情指标等变量，运用相关分析、关联分析、回归分析等方法的数据探索，并基于选取出的显著变量建立时间序列 VAR 模型。基于此模型对全国猪肉平均价格进行预测，得到 2020 年上半年生猪价格将继续上涨的结论。其次，考虑但各省猪价波动情况有差，建立动态面板模型，得到以下基本结论：影响各地猪价的最主要因素是生猪存栏量和网络舆论导向，而猪肉的需求量、替代品价格、成本价格对猪价影响不显著。往往能够较大影响猪价波动的季节性因素在此次猪瘟的大背景下，也没有显著体现其影响力。

关键词：生猪价格 因子挖掘 动态面板模型 GMM 估计方法 LSDVc 估计方法

Abstract: Our country, China is a major pork consumer, so that stabilizing the pork price is a practical problem in China. In this paper, we explore the mechanism of pork price fluctuations as well as propose an effective pork price forecasting model. First of all, the variables that affect the price of pork theoretically, such as the supply for pork, the demand for pork, the price of substitutes, and the network public opinion related to the price of pork, are explored by using common data mining techniques such as correlation analysis, association analysis, regression analysis. The time series VAR model is established based on the selected significant variables. Based on this VAR model, the average price of pork in China is predicted, and it is concluded that the pork price will continue to rise in the first half of 2020. Secondly, considering that the fluctuation of pork prices in each province is different, a dynamic panel data model is established. The main conclusions are as follows: The factors that significantly affect the pork price are the supply for pork and the network public opinion, while the demand for pork, the price of substitutes, and the cost price have no significant effect on the pork price. The seasonal factors which can influence the fluctuation of pork price are not obvious in the background of this swine fever.

Keywords: Pork price, Data mining, Dynamic panel data model

一、引言与文献综述

猪肉是我国民众最喜爱的肉类之一，在每年节日如端午、中秋、国庆、春节肉类消费的集中时节，猪价升高，致使以往的生猪价格几乎在每个年度内都会出现“两头高、中间低”的格局(刘国信, 2020)[1]。但自 2018 年 8 月开始的非洲猪瘟，我国各地的生猪、猪肉价格异常波动，不同产销区之间趋势和程度有所差异：全国平均猪价先是在 2018 年底反春节旺季下跌到全年最低为，价贱伤农；再是在 2019 年节后 3 月分起反淡季一路飙升，全国生猪出栏均价在 10 月中下旬突破 40 元/kg 的高位，价贵伤民。在区域上，东南沿海省市涨势明显，北方省市涨势温和甚至有跌。因此，在此非洲猪瘟冲击背景下，由于猪肉市场面临的不确定性持续增强，异常波动更加明显，对猪肉价格波动进行系统研究，测定猪肉价格波动特征具有一定的学术意义和现实指导意义。

我国作为世界猪肉第一消费国，猪肉价格与民生紧密相关。因此猪价的异常和高频波动引人瞩目，探索影响猪价波动的因素以及如何稳定猪价是学术界广泛关注的课题。一方面，根据基本的供需理论，猪肉价格的升高往往是由猪肉产量不足、猪肉需求增加，或两者双重作用导致的。在对传统供给因素的多维关联规则数据挖掘中，得出猪价与猪肉产量、存栏量、出栏量呈负相关性；与牛肉价格和鸡肉价格呈正相关性；与仔猪价格和玉米价格呈正相关性(乔浪, 2009)[2]。除了国内影响因素外，国际猪肉价格对国内猪价造成持续的负向影响，猪肉进口的持续增加将使得国内猪肉价格受到打压(刘春鹏, 2019)[3]。另外，猪肉价格序列具有较显著的季节因素，往往在每年 1、2 季度处于下跌，3、4 季度处于上升(范青青, 2018)[4]。另一方面，除传统供给影响因素外，投机行为使得部分中国农产品，如蒜、姜(Xi-Xi Zhang, 2019)[5]、中国东部地区猪肉(殷克东, 2015)[6]出现价格泡沫。同时，不同方向网络舆论也会影响农产品价格，导致农产品营销的不稳定甚至大部分亏损与极个别暴利并存的两极分化现象，比如曾经导致水果、生猪与活禽等严重滞销的致病的网络谣言(赵豫蒙, 2019)[7]。根据网络热词显示，部分“炒猪客”、猪肉期货投资者先是在 18 年底传播“非洲猪瘟害人”谣言，致使消费者恐于购买、小型养殖户恐惧猪瘟，加大屠宰弃养，从而压低价格囤积猪肉。再是到 19 年开始哄抬物价、炒作猪肉、赢取暴利。因此，在考虑传统供给因素外，舆情因素在本轮猪价上涨的影响情况值得探索。

目前，国内外不乏对网络舆情如何影响产品价格的研究。谷歌搜索引擎相关关键词搜索量可作为投资者关注度指标，用于探索投资情绪对原油价格的影响，并提高了原油价格的预测准确度(Li, 2015)[8]；对有关外汇的财经新闻进行舆情分析、情绪分级，并基于此建立的舆情影响力模型能有效预测人名币汇率走势(成舟, 2019)[9]。然而，网络舆情研究大多集中应用于金融市场，仅有少数应用于农产品市场，但都有较好的实证结果。比如，从有关蔬菜供需的微信文章里所分析出带有情绪分级的舆情指标变量被证明与我国四种蔬菜价格与呈较弱具有一定相关性(Li, 2020)[10]；在对 2011-2017 年中国猪价波动研究中，由搜索关键词为流行猪瘟名称的百度指数所构建猪瘟舆情指标，与猪瘟对猪价的冲击呈正相关，并能提高猪价预测模型的精度(Lin, 2019)[11]。所以，挖掘相关网络舆情数据如百度搜索指数和对微博博文情感分析，探究其对我国此轮非洲猪瘟背景下猪价异常上涨影响的情况，并基于此将舆情变量放入到预测模型中是本文一创新之处。

当前，有关猪肉价格预测模型研究较多，主要基于时间序列模型、神经网络模型、灰色系统模型等，比如考虑生猪生产成本与猪肉消费替代品的多元回归 BP 神经网络价格预测模型能以准确率 93.3%较好地预测生猪价格(任青山, 2019)[12]；ARIMA-GM-RBF 组合模型成功预测 2019 年猪价的快速增长趋势(吴培, 2019)[13]。然而，以上研究仅局限于研究全国平均猪价或某省市猪价的单一时间序列，但此轮猪价波动各省市表现差异较大，出现南北悬殊的现象，比如：2019 年年初，广东、四川、浙江、福建等省猪价飙升，生猪价格一直保持在 10 元以上，而东三省、河南、山东等地猪价下跌，生猪价格每公斤一度跌破了 6 元的成本价。因此，运用面板数据模型，刻画出各省市的整体效应和个体效应，再对各省市猪价进行估计将更为合理准确。而由于猪价受上期价格影响较大，因此本文应考虑猪价自回归效应，使用动态面板模型进行拟合。

目前,最主流应用的动态面板模型参数估计方法为差分广义矩估计(Diff-GMM)与系统广义矩估计(Sys-GMM),并且都仅适用于 $T(\text{时间点}) < N(\text{横截面个体})$ 的数据集,如 Sys-GMM 动态面板模型被用于探索中国 284 个城市 2001-2013 年科技创新对能源使用效率的影响(Wang, 2020)[14]。另一种较不被常用的动态面板模型方法——虚拟变量修正方法(LSDVc)在 Monte Carlo 模拟中被证实:在小样本情况下,其参数估计有效性和一致性都优于 Diff-GMM, Sys-GMM 方法 (张志强, 2017)[15]。也有学者根据过往实证文献,设计出更符合经济数据的 N, T 大小的模拟数据,结果显示: LSDVc 估计在大部分情况下有最小的均方根误差(RMSE),但是在 $T=10, N=30$ 的情况下, GMM 估计量准确性优于 LSDVc(Skrabic, 2019)[16]。

接下来文章结构:在第二部分,本文将运用数据挖掘方法相关性分析、关联分析、回归分析来探索猪价与生猪供给、需求、成本、季节性、舆论导向 5 个方面 9 个因素间的影响关系,并选取选取有效因子进行时间序列模型建模。在第三部分,本文对 30 个省市地区的猪价进行面板数据的探索并建立动态面板模型。第四部分将对本文主要结论进行总结。

二、 时间序列数据探索与因子挖掘

本部分将基于相关性分析、关联分析、构建 VAR 模型进行回归分析以及 Granger 因果检验来定量分析猪肉价格与供给、需求、成本、季节性、舆论导向 5 个方面 9 个因素的影响关系并为预测模型选取因子。最后,根据筛选出的有效因子建立猪价预测模型。

(一) 分析方法介绍

相关性分析(Correlation Analysis)通过计算随机变量间的相关系数,来发掘变量间的线性相关性的强弱以及正负,一般视相关系数的绝对值大于 0.3 为具有一定相关性。

相关系数体现一种变量间的线性关系,而关联分析(Association Analysis)作为数据挖掘的一个重要分支,其主要目的是在数据背后发现事物与事物之间可能存在的关联或联系。根据 J.Han(2000)[17],关联规则是形如 $A \rightarrow B$ 的蕴涵式,代表了事件 A 与事件 B 发生的一种关联性。关联规则在整体观测中的支持度(support)是 A、B 同时出现的概率;置信度(confidence)是 A 发生的情况下, B 也发生的条件概率;提高度(Lift)是置信度与支持度之比。如果事件 A 对事件 B 的支持度高,代表 A 与 B 同时发生的可能性高,是一种事件间的关联规律;如果事件 A 对事件 B 的置信度水平高,代表 A 的发生很可能会再促进 B 出现这一事件的可能性高;提高度表征事件之间的亲密关系,代表 A 的出现对于 B 的出现的的影响程度。如果满足人为设定的最小支持度阈值和最小置信度阈值,则认为关联规则是有趣的。

表 1. 三种度量计算公式与提高度结果解读

计算公式		$L(A \rightarrow B)$	A 与 B 发生情况
支持度	$S(A \rightarrow B) = P(A \& B) = \frac{N(A \& B)}{N}$	$= 1$	独立
		> 1	正相关
置信度	$C(A \rightarrow B) = \frac{P(A \& B)}{P(B)}$	< 1	互斥
提高度	$L(A \rightarrow B) = \frac{C(A \rightarrow B)}{S(B)}$		

其中, $P(\cdot)$ 表示某事件发生概率, $N(\cdot)$ 表示某事件发生次数, N 为全体事件总数。

在回归分析部分，本文采用向量自回归模型(Vector Autoregression, VAR)进行回归拟合。VAR 模型描述在同样本期间内的一组内生变量可以作为它们过去值的线性函数，并采用 OLS 估计。假设有 2 个时间序列变量 Y_t 和 X_t ，分别作为 2 个回归方程的被解释变量，而解释变量为这 2 个时间序列变量的 p 阶滞后值，构成一个二元的 VAR(p)系统:

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \beta_{12}Y_{t-2} + \cdots + \beta_{1p}Y_{t-p} + \alpha_{11}X_{t-1} + \cdots + \alpha_{1p}X_{t-p} + \varepsilon_{1t} \tag{1}$$

$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \beta_{22}Y_{t-2} + \cdots + \beta_{2p}Y_{t-p} + \alpha_{21}X_{t-1} + \cdots + \alpha_{2p}X_{t-p} + \varepsilon_{2t} \tag{2}$$

其中， t 为时间， β_{10} 和 β_{20} 分别为方程的常数项， α 为 X_t 滞后项的系数， β 为 Y_t 滞后项的系数， ε_{1t} 和 ε_{2t} 为方程的误差项。

在此部分，本文先对各变量时间序列进行 ADF 单位根检验，检验结果为各个变量不平稳。在对各变量取对数并做差后，所有变量通过 ADF 检验，为平稳序列，可进行 VAR 建模。在 VAR 建模前，可以利用信息准则 AIC、BIC 来确定变量的滞后阶数，以避免 VAR 系统中包含的变量个数与滞后阶数过多而导致的估计误差增大，所以本文在此部分根据 AIC、BIC 等指标，并考虑实际经济意义，综合选择变量滞后阶数来进行模型拟合。下文分析结果展示中，仅显示拟合结果最好的方程。(注：拟合结果好坏判断标准：1. 系数估计具有经济学意义 2. 通过检验 3.AIC 或 BIC 最小)

模型拟合后应对因变量与自变量进行 Granger 因果检验。Granger(2003)提出，在平稳时间序列情形下，两个经济变量 X、Y 之间的 Granger 因果关系定义为：若在包含了变量 X、Y 的过去信息的条件下，对变量 Y 的预测效果要优于只单独由 Y 的过去信息对 Y 进行的预测效果，即变量 X 有助于解释变量 Y 的将来变化，则认为变量 X 是引致变量 Y 的格兰杰原因。本文在此部分将在 VAR 模型拟合之后做 Granger 因果检验，并设为原假设 H_0 ：变量 X 不是 Y 的格兰杰原因，并设定置信水平为 0.1。其中变量 Y 设为平稳处理后的猪肉价格时间序列变量，X 则为我们感兴趣的自变量。

(二) 数据介绍

本文研究区间为 2018 年 8 月到 2019 年 12 月，共 17 个月度数据，72 个周度数据。所选取的变量与其数据来源如表 2 所示：

表 2. 变量介绍

	变量名	英文简称	来源	频率
供给因素	生(活)猪价格	pig_price	农村农业部	周/月
	猪肉集市价格	pork_price	农村农业部	周/月
	生猪存栏量	supply	农村农业部	月
	猪肉进口量	import	海关总署	月
成本因素	豆粕价格	bean_price	农村农业部	周/月
	玉米价格	maize_price	农村农业部	周/月
需求因素	主要批发市场交易量	demand	农村农业部	月
	白条鸡价格	chincken_price	农村农业部	周/月
季节因素	是否为消费旺季	season	公开信息	月
舆情因素	”非洲猪瘟“百度搜索量	bd_opinion	百度指数网页	周/月
	”猪价上涨“微博博文量	wb_opinion	微博网页	周/月

本文研究对象为生猪价格，而非涵盖中间运输、屠宰等相关成本的猪肉集市价格。由图 1，2018 年 8 月非洲猪瘟开始蔓延，部分产区猪肉产量减少，叠加 9、10 月节假日消费拉动，价格略有上升，到 11 月开始下跌；2

月春节特需没有拉动猪价上涨：生猪价格为 12.55 元/kg，环比下降 4.3%，同比下降 10.9%，猪肉集市价格为 22.87 元/kg，环比下降 2.4%，同比下降 9.5%；但随着生猪存栏持续下降带来的供给缺口逐渐显现，3 月开始，猪价一反往年节后淡季下跌走势 2019 年节后猪价一路上涨到 11 月最高价：生猪价格 35.13 元/kg，环比增长 153.3%，猪肉集市价 54.93 元/kg，环比增长 132.8%。

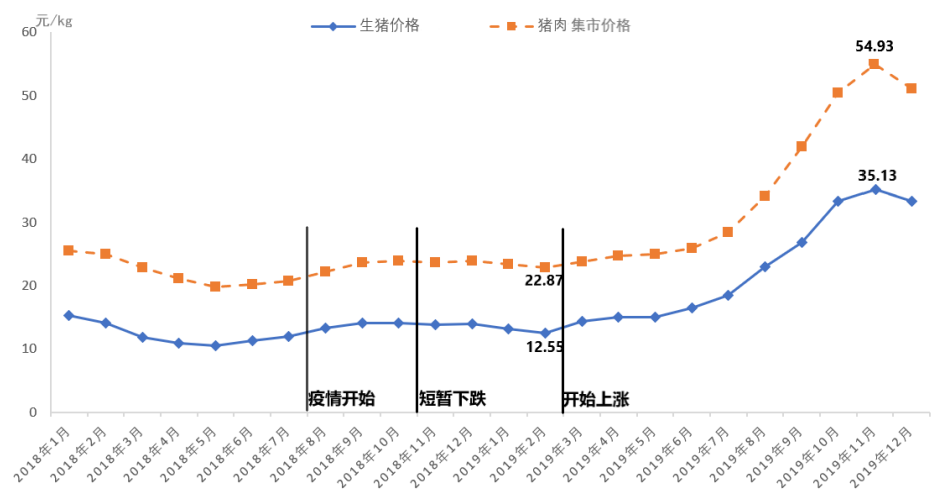


图 1：猪价走势图

(三) 因子挖掘

1. 供给因素分析

根据农村农业部数据披露，截至 2019 年 11 月，全国共报告发生 160 起非洲猪瘟疫情，扑杀生猪 119.3 万头；由表 3，生猪存栏量和能繁母猪存栏量与上年同期相比，已经连续 11 个月下降幅度在 10% 以上。由此可见，此次疫情对我国生猪产业形成了较大冲击。

表 3. 2018 年 9 月到 2019 年 11 月生猪存栏与能繁母猪存栏变动情况

	生猪存栏			能繁母猪存栏	
	猪价(元/kg)	环比(%)	同比(%)	还比(%)	同比(%)
2018 年 9 月	14.14	0.8	-1.8	-0.3	-4.8
2018 年 12 月	13.95	-3.7	-4.8	-2.3	-8.3
2019 年 3 月	14.35	-1.2	-18.8	-2.3	-21.0
2019 年 6 月	16.53	-5.1	-25.8	-5.0	-26.7
2019 年 9 月	26.83	-3.0	-41.1	-2.8	-38.9
2019 年 10 月	33.34	-0.6	-41.4	0.6	-37.8
2019 年 11 月	35.13	2.0	-39.8	4.0	-34.5

生猪存栏量是一定时点的全部生猪饲养头数，在本文中视作为生猪供给量指标。由图 2(a)，随着存栏量下降，猪价呈上涨态势，上期存栏同比减少量趋势与本期猪价走势一致。根据海关总署公布数据显示，2019 全年进口猪肉 210.8 万吨，增加 75%，进口猪肉价格低于国内价格，应降低猪价，猪肉价格也在年底开始下降，2019 年 12 月全国生猪价格环比上月下降了 5.07%。由图 2(b)，本期、上期猪肉进口量与猪价趋势较为一致，没有明显体现出进口量对猪价的负效应，其原因可能是 8 月开始的大量进口猪肉弥补国内供给缺口的效应较为迟钝，也许将会在 2020 年上半年体现，有待进一步探索。

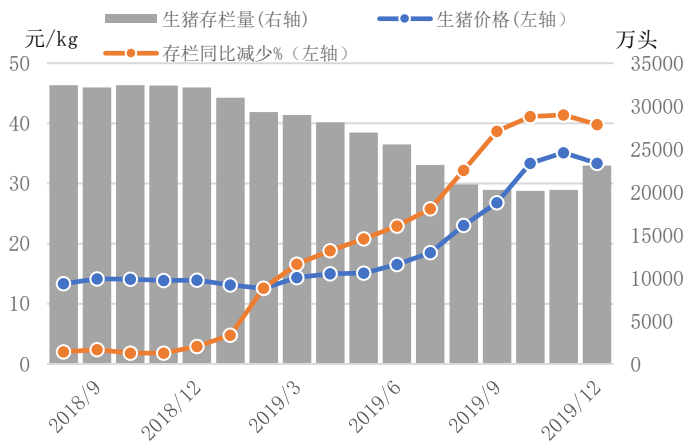


图2(a): 存栏量与猪价趋势图

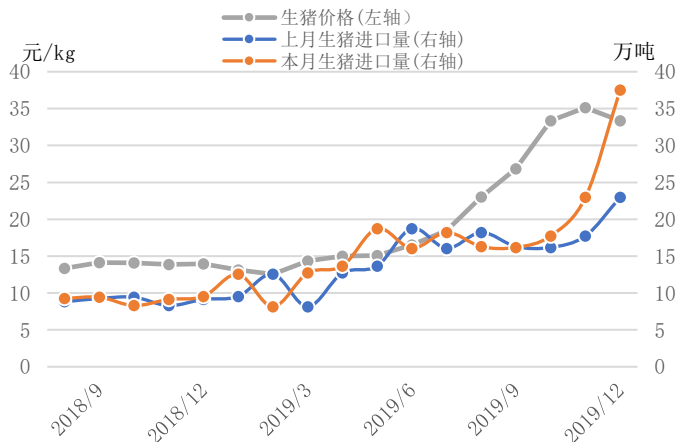


图2(b): 进口量与猪价趋势图

表 4. 供给因素分析结果

相关性分析			
期数	本期	前 1 期	前 2 期
生猪存栏量	-0.85	-0.94	-0.96
猪肉进口量	0.77	0.77	0.78
关联性分析			
关联规则	支持度	置信度	提高度
上期生猪存栏下降→本期猪价上涨	0.59	0.77	1.09
上期猪肉进口量上升→本期猪价下降	0.24	0.36	1.24
回归分析			
变量	L1_dln_supply	L1_dln_import	
估计系数	-1.75	-0.15	
t 检验 p 值	0.004	0.09	
VAR 方程	$y_t = 0.33y_{t-1} - 1.75x_{t-1} - 0.02$	$y_t = 0.55y_{t-1} - 0.15x_{t-1} + 0.03$	
R 方	0.64	0.37	
通过 Granger 因果检验(0.1 水平下)	是	是	

注：表中回归分析部分，变量 Lp_dln_v 表示：对原始变量 v 取对数后作差，从而获得平稳序列后，取其 p 阶滞后的时间序列，以下表格中变量命名规则同此意，不再赘述。

由表 4，近 3 期生猪存栏量与本期猪价相关系数呈强负相关性；由关联分析结果得，上期生猪存栏下降与本期猪价上涨两事件共同发生可能性高，前者发生可能会促进后者的发生，并且由提高度 >1 知，两事件发生情况为正向，所以认为两者具有较强的关联性；由 VAR 模型拟合结果知，生猪存栏量回归系数为负，并且通过 t 检验，是猪价的 Granger 原因，并且拟合模型 R 方较大，说明由自变量可解释因变量的部分较大。而猪肉进口量与猪价相关性为正，和预期不符；关联分析中，支持度与置信度较低的主要是由猪价下降事件本身发生频率低导致，而提高度 >1 说明两事件发生情况为正向关系；由 VAR 模型拟合结果知，进口量系数为负，符合预期并通过 t 检验，认为进口量是猪价的 Granger 原因。所以，生猪存栏量是影响猪价的重要因素，进口量对猪价影响较为显著。

2. 需求因素分析

此次疫情也对居民猪肉消费产生了一定的冲击。18 年 9 月起，随着疫情快速扩散以及相关新闻资讯报道传播，居民对猪肉食品安全的担忧增加，消费信心不足，从而抑制了一定的猪肉消费。而批发市场是我国最主要的猪肉流通渠道之一，所以批发市场猪肉成交量从一定程度上反映了猪肉的市场需求情况(胡浩，2020)[18]。由图 3(a)，

随着交易量从 2019 年 2 月最低点处开始回升，猪价也逐渐上涨。由图 3(b)，作为猪肉主要替代品的白条鸡，其价格也在研究时间区间内呈上升趋势。

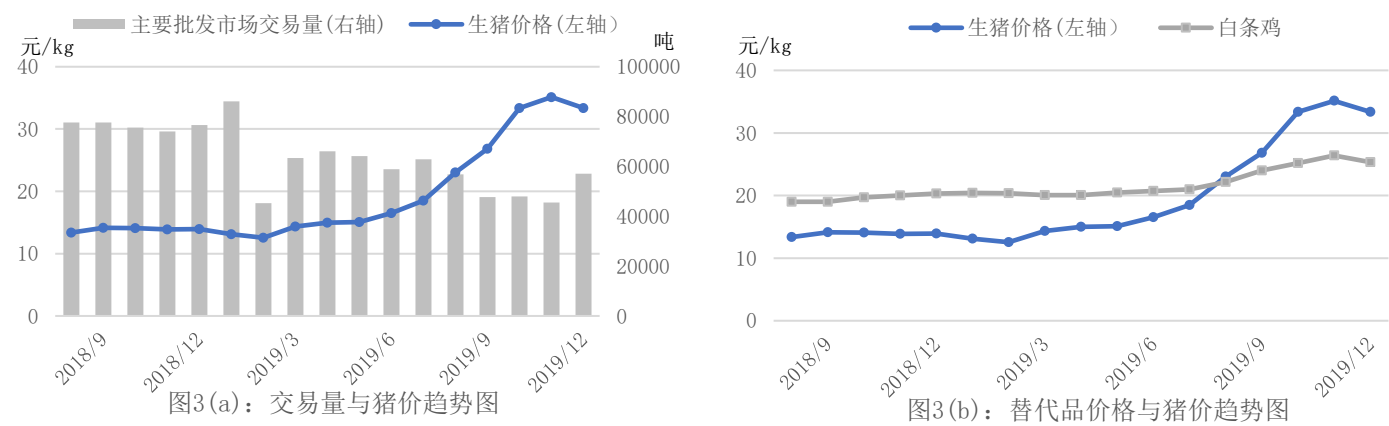


表 5. 需求因素分析结果

相关性分析			
期数	本期	前 1 期	前 2 期
批发市场交易量	-0.68	-0.68	-0.68
白条鸡价格	0.96	0.94	0.88
关联性分析			
关联规则	支持度	置信度	提高度
上期猪肉交易量上升→本期猪价上升	0.41	0.78	1.10
上期白条鸡价格上升→本期猪价上涨	0.63	0.76	1.10
回归分析			
变量	L1_dln_demand	L1_dln_chicken_price	
估计系数	-0.22	0.31	
t 检验 p 值	0.042	0.77	
VAR 方程	$y_t = 0.64y_{t-1} - 0.22x_{t-1} - 0.03$		$y_t = 0.48y_{t-1} + 0.31x_{t-1} + 0.02$
R 方	0.48	0.26	
通过 Granger 因果检验(0.1 水平下)	是	否	

由表 5，交易量与猪价相关系数为负，回归分析中估计系数为负，不符合经济学原理。尽管猪肉消费在短期受到了一定程度的抑制，但由于消费偏好无法短期内改变，猪肉“刚需”仍然存在，因此需求量对猪价的影响效应不显著。鸡价与猪价呈强正相关性，关联性较强，但回归分析中其估计系数未通过 t 检验，其不为猪价的 Granger 原因。并且白条鸡等肉类价格往往与猪价联动，若放入模型中可能出现较强的共线性。所以，认为这两需求因素对猪价影响不显著。

3. 成本因素分析

猪饲料的主要来源为玉米和豆粕，因此其价格可以一定程度上代表养猪的饲养成本。由图 4，豆粕价格没有与猪价有一个明显的同趋势在。玉米价格与猪价有一同上涨的趋势。

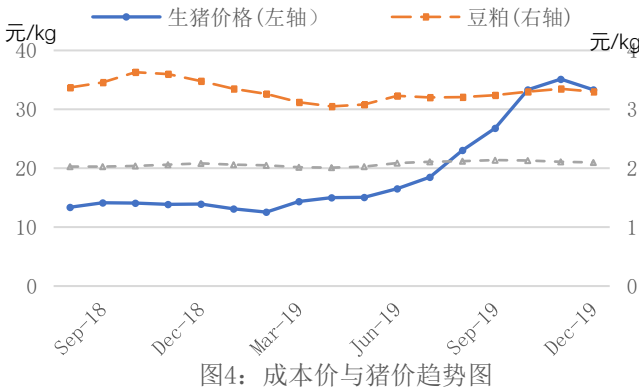


图4：成本价与猪价趋势图

表 6. 成本因素分析结果

相关性分析			
期数	本期	前 1 期	前 2 期
豆粕价格	-0.11	-0.22	-0.38
玉米价格	0.75	0.84	0.86
关联性分析			
关联规则	支持度	置信度	提高度
上期豆粕价格上升→本期猪价上涨	0.41	0.70	0.99
上期玉米价格上升→本期猪价上涨	0.35	0.55	0.77
回归分析			
变量	L1_dln_maize	L1_dln_bean	
估计系数	2.94	0.01	
t 检验 p 值	0.09	0.98	
VAR 方程	$y_t = 0.51y_{t-1} + 2.94x_{t-1} + 0.02$	$y_t = 0.52y_{t-1} + 0.02x_{t-1} + 0.02$	
R 方	0.37	0.25	
通过 Granger 因果检验(0.1 水平下)	是	否	

由表 6，豆粕价格与猪价呈负相关；提高度<1，说明上期豆粕价格上升和本期猪价上涨两事件发生情况为负相关；回归分析中，估计系数未通过 t 检验，不是猪价的 Granger 原因。玉米价格与猪价呈正相关；关联性分析中显示上期玉米价格上升与本期猪价上涨两事件关性弱，并且发生情况负相关；VAR 模型拟合结果显示，估计系数为正，符合经济学原理，玉米价格是猪价的 Granger 原因。由此认为玉米价格对猪价上涨有一定的推动作用。

4. 季节因素分析

本部分构造哑变量“是否当月为消费旺季”，其中把每年 12、1、2、9、10 月设为 1，其它设为 0。将此哑变量与猪价进行线性回归，结果为：哑变量估计系数 0.77，但未通过 t 检验，因此，在此次猪价波动中，季节拉动猪价波动效应不显著。

5. 舆情因素分析

本部分构造两个舆情变量，一是负面舆情，二是正面舆情。负面舆情变量为含有关键词“非洲猪瘟”的百度周度搜索量，简称百度指数，代表群众对次疫病的关注度。由图 5，百度指数从 2018 年 8 月疫情伊始开始波动上升，在 11 月达到高峰，猪价也在这个时点开始下降。其主要原因为，当时新闻媒体对“非洲猪瘟”的新闻资讯传播，致使消费者恐于购买，养殖户急于出栏抛售，将价格拉低。正面舆情变量由微博内容含有“猪价上涨/暴涨/会涨”这三个关键词的条文数构成。微博数量随着 2019 年 4 月猪价的上涨开始增加，在 10、11 月达到高峰。

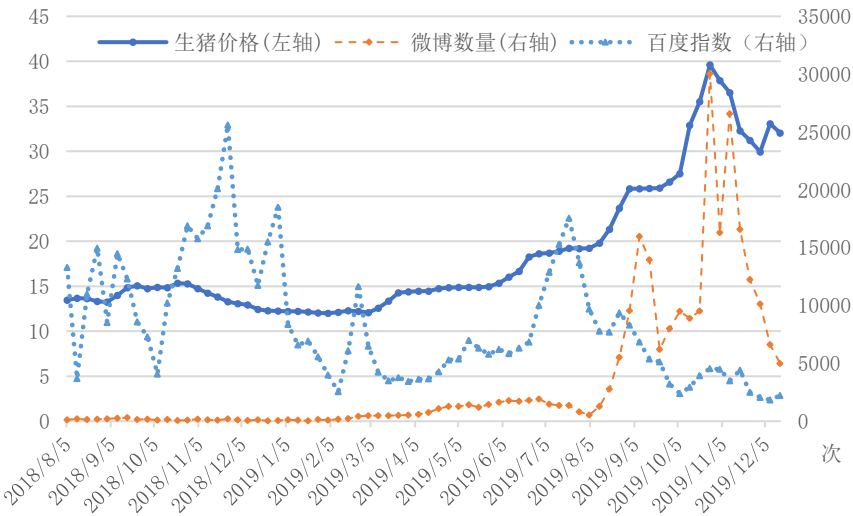


图5：舆情变量与猪价趋势图

表 6. 舆情因素分析结果

相关性分析			
期数	本期	前 1 期	前 2 期
微博条数	0.87	0.84	0.79
百度指数	-0.41	-0.39	-0.37
关联性分析			
关联规则	支持度	置信度	提高度
上期微博条数上升→本期猪价上涨	0.43	0.70	1.22
上期百度指数上升→本期猪价下降	0.20	0.42	1.01
回归分析			
变量	L1_dln_wb_opinion	L2_dln_bd_opinion	
估计系数	0.01	-0.02	
t 检验 p 值	0.31	0.07	
VAR 方程	$y_t = 0.42y_{t-1} + 0.01x_{t-1} + 0.01$	$y_t = 0.43y_{t-1} - 0.02x_{t-1} - 0.02$	
R 方	0.197	0.22	
通过 Granger 因果检验(0.1 水平下)	否	是	

由表 6，微博条数与猪价正相关，有一定的关联性，并且由提高度>1 得出，上期微博条数上升与本期猪价上涨两事件发生情况正相关；回归分析结果表示，估计系数为正，但没有通过检验，表明此正面舆情对价格的推动作用不显著。百度指数与猪价负相关，关联性较弱，但上期百度指数上升与本期猪价下降两事件发生情况正相关；回归分析结果表明负面舆情变量对猪价有显著的负向影响，并且认为是猪价的 Granger 原因。因此舆情变量“百度指数”应当被考虑入后文的预测模型构建中。

(四) 预测模型建立

本部分使用向量自回归模型对因变量生猪价格与上一节所筛选出的显著影响自变量进行建模。
预测模型结果：

$$y_t = 0.34y_{t-1} - 1.91x_{1,t-1} - 0.12x_{1,t-2} + 4.09x_{3,t-1} - 0.12x_{4,t-2} - 0.04$$

(3)

式中， y_t 为生猪价格增长率， $x_{1,t}$ 为生猪存栏量增长率， $x_{2,t}$ 为猪肉进口量增长率， $x_{3,t}$ 为玉米价格增长率， $x_{4,t}$ 为百度指数增长率。该模型各变量系数估计在 0.01 水平下都通过 t 检验，且拟合拟合结果 R 方为 0.972，表示因变量能由自变量解释的部分较多，模型拟合度较高。其中，舆情变量对因变量解释度较高，在未加入舆情因素的模型种，R 方仅为 0.63。

该模型通过 LM 检验：认为残差不存在自相关性；通过正态性检验：认为残差服从正态分布，通过稳定性检验，认为该 VAR 系统为平稳过程。Granger 因果检验表明，自变量是因变量的 Granger 原因。

图 6 为该模型对未来 6 个时点即 2020 年对上半年的预测值，红色线为 2020 年 1 月和 2 月的实际观测值，由此可见，模型的预测比较不准确，其原因应该还是没有考虑疫情因素。

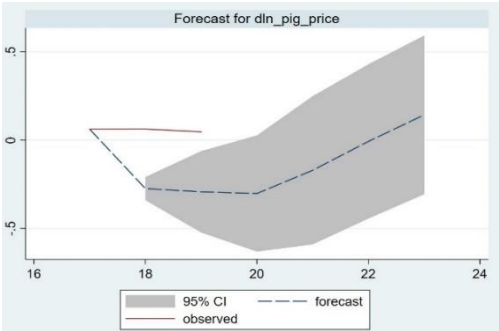


图 6：对 2020 上半年生猪价格预测

三、 面板数据探索与建模

(一). 数据介绍

此部分分析对象为：研究时间区间内，除港、澳、台、西藏外共 30 各省级行政区的生猪价格，以及上文提到的显著影响变量。数据来源于各省市统计局网站、各省市农委网站。若某省市统计局或农委网站仅仅公布了季度数据，则采用插值法进行估算；若部分省市网站仅公布了生猪出栏量数据，则用“存栏量：出栏量 $\approx 1:2$ ”的方法进行估算。

此次疫情导致猪价波动各省市表现差异较大，出现南北悬殊的现象：为防止疫情的扩散，中国农业农村部 2018 年 9 月份开始限制活猪的跨省调运。以至于 2019 年年初，四川、浙江、福建等省猪价飙升，而东三省等地猪价暴跌。猪肉像是实施了户籍制，价格因省份而不同(任慧媛，2019)[19]。

图 7(a)展现各省市在研究时间区间内的猪价均值，可以看出我国南方省市的价格普遍比北方高，这种特征在东南沿海尤其明显；图 7(b)为各省市猪价均值、标准差、极差的箱线图，均值中位数偏低说明价格较高的地区间价格差别较大，价格较低的地区价格差别不大，标准差和极差都出现了离群值，说明每个省市的价格波动程度也大有不同；图 7(c)为猪肉价格分位数处省市的价格趋势图，其中均价最大的为浙江省，价格中等的为湖北省，价格最低的为云南省。

由图 8(a)，两河、两湖、两广，东三省、四川和山东等地的生猪存栏量在全国占比非常之高，是猪肉主力生产区，其中东三省、山东生猪价格曾每公斤一度跌破了 6 元的成本价。而广东、浙江、福建、四川等地猪肉消耗比较大，属于猪肉主力销售区，生猪价格也一直保持在 10 元以上。由图 8(b)，广西、广东、江苏等南方城市，生猪存栏量下降程度较大，价格也相应之上涨程度较大；北方城市相对存栏下降程度较小，其价格也上升较为和缓。

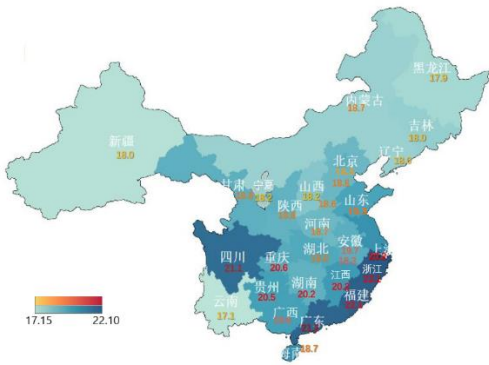


图 7(a): 各省均价分布

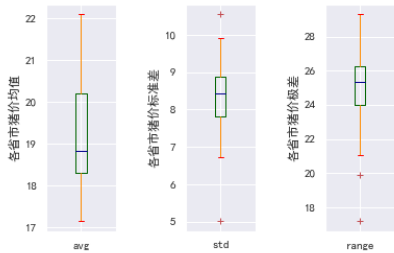


图 7(b): 箱线图

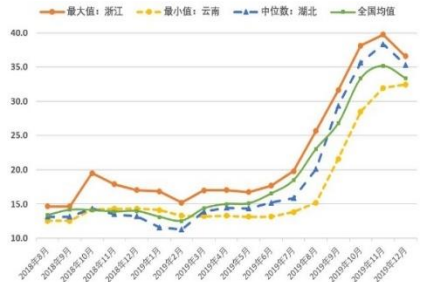


图 7(c): 分位数猪价趋势

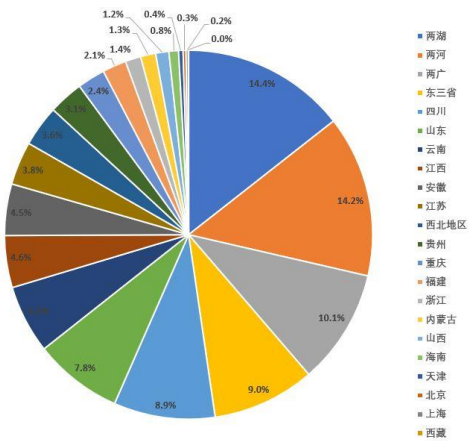


图 8(a): 2018 年末各地区生猪存栏量之比

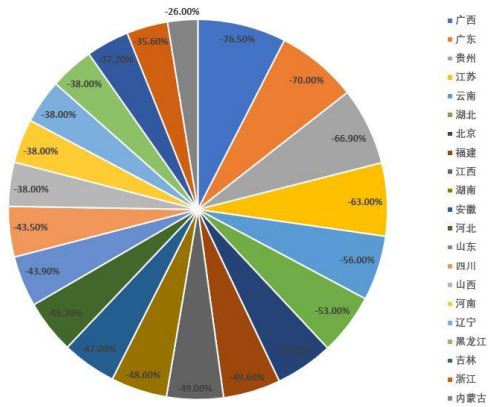


图 8(b): 各省 2019 年 11 月与 2018 年 7 月存栏量同比

(二). 变量分析

此部分使用面板数据模型对变量进行分析与筛选。考虑个体的异质性，本部分首先使用固定效应模型(Fix effects model, FEM)，以各省市的生猪价格为因变量，各省市的生猪存栏、猪肉进口量、玉米价格、百度指数和微博博文数 5 个显著的影响变量(由本文第二部分所筛选出)，进行拟合。其次，考虑各省产量差别较大，生猪存栏变量的回归系数在各省可能有所不同，进一步使用变系数模型(Varying coefficient Model)对面板数据进行拟合。

固定效应模型 FEM:

$$y_{it} = X_{it}\beta + \alpha_i + \mu_{it} \quad (i = 1, \dots, N; t = 1, \dots, T)$$
 (4)

y_{it} 为个体 i 在 t 时刻的因变量观测值， X_{it} 为外生自变量矩阵， β 为其系数， α_i 是不可观测的不随时间变化的固体效应，如每个省市不可观测的特点， μ_{it} 为扰动项，假设服从正态分布。其中 α_i 与 X_{it} 可以相关， μ_{it} 与 X_{it} 里的自变量均不相关，则可由 OLS 得到一致估计。

变系数模型 VCM:

$$y_{it} = X_{it}\beta_i + \mu_{it} \quad (i = 1, \dots, N; t = 1, \dots, T)$$
 (5)

在 VCM 中，每个个体有自己的自变量估计系数，更进一步体现个体的特异性。

表 7. FEM 与 VCM 拟合结果

ln_pigprice	固定效应模型	变系数模型
L1_ln_supply	-0.211** (0.093)	
L1_ln_import	0.196*** (0.032)	-0.443 (0.032)
L1_ln_maize_price	-0.209 (0.655)	-0.0901 (0.891)
L2_ln_bd_fix	-0.002*** (0.026)	-0.080*** (0.017)
L1_ln_wb_opinion	0.132** (0.016)	0.160*** (0.011)
R-squared	0.782	0.792
ρ	0.777	

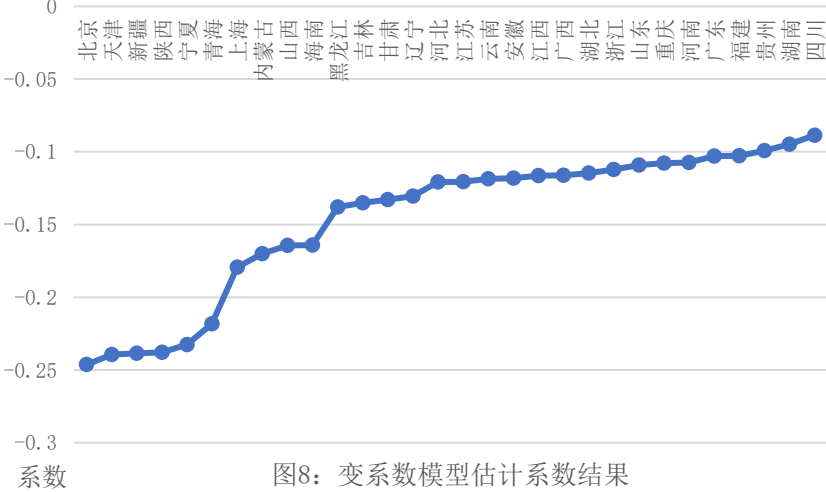


图8: 变系数模型估计系数结果

(注：本部分对所有变量进行取对数处理)

由表7，猪肉进口量L1.ln_import在FEM中估计系数为正，不符合经济学常理，在VCM中为负但没有通过t检验，故将该变量从面板数据模型中剔除；玉米价格变量L1.ln_maize_price在两个模型中估计系数都为负，不符合经济学常理，故将该变量从面板数据模型中剔除。FEM拟合结果显示，固定效应占混和方差($\alpha_i + \mu_{it}$)比例 ρ 为0.78，说明个体效应较为显著。由图8，通过使用VCM，也能清晰的观察到生猪存栏量对华北、西北这样猪肉生产较小省市的猪价影响较大，反而对猪肉生产和消耗都将强的南部省市影响较小。

(三). 预测模型建立

由第二部分时间序列分析中发现，猪价受上期的自身影响较大，因此此部分使用动态面板模型进行建模。

动态面板模型:

$$y_{it} = \rho y_{it-1} + X_{it}\beta + \alpha_i + \mu_{it}$$
 (6)

式中变量含义同式(4)，其中 y_{it-1} 为因变量的滞后项，从而使得模型具有动态解释能力，但由于 y_{it-1} 与 μ_{it} 相关，模型中存在内生性问题。为了解决这一内生性，Arellano 和 Bond (1991)14[20] 提出了利用工具变量来推导相应矩条件的广义矩(GMM)方法，即“差分 GMM 方法”。该方法的基本思想是：先对原模型进行一阶差分变换，消除模型中的个体异质项，得到式(7)，并对此提出矩条件 $E(y_{it-1}\Delta\mu_{it}) = 0$ 。对于变换后的差分方程式(7)，将内生

变量的滞后变量看成该内生变量的工具变量，比如，对 $t=3$ ，内生变量为 $y_{i2} - y_{i1}$ ，误差项为 $\mu_{i2} - \mu_{i3}$ ，选用工具变量为 y_{i1} ，对 $t=4$ 选用工具变量为 y_{i2} ， y_{i1} 。

$$\Delta y_{it} = \rho \Delta y_{it-1} + \Delta X_{it} \beta + \Delta \mu_{it} \tag{7}$$

然而，差分 GMM 方法降低了内生性对模型估计带来的影响，在小样本条件下，尤其是当因变量 y_{it} 持续性较强时，差分 GMM 方法存在较严重的“弱工具变量”问题，从而导致系数估计结果精度较差(Kiviet, 1995)[21]，并且无法对个体效应进行估计。Blundell 和 Bond (1998)[22]基于新的复合矩条件，提出“系统 GMM 方法”。系统 GMM 方法对原水平方程式(6)和差分方程式(7)同时进行估计，并对水平方程(6)提出矩条件 $E(\Delta y_{it-1}(\alpha_i + \mu_{it})) = 0$ 。系统 GMM 除了使用差分 GMM 对差分方程的工具变量外，还加入 Δy_{it-1} 作为水平方程(6)的工具变量。当 y_{it} 自回归系数较大，或固定效应占混和方差比例较高时，使用系统 GMM 更有优势。

为了保证 GMM 可以作为一致估计法，我们需要对扰动项自相关性和工具变量有效性进行模型拟合后的检验。首先，差分 GMM 模型的扰动项应该具有显著的一阶相关和不显著的二阶自相关。为此，在一阶差分残差中使用了一、二阶序列相关性的 Arellano-Bond 检验。其次，为了判断系统 GMM 估计中所采用的工具变量是否整体有效，采用 Sargan 检验或 Hansen 检验进行判断，其原假设是“所有的工具变量都是外生的”。因此，若工具变量是有效的，则不应拒绝原假设。

LSDV 估计方法是通过虚拟变量把个体效应或时间效应从误差项中分离出来，使分离后剩余的误差项与自变量无关，从而可以使用 OLS 估计。偏差矫正的 LSDV(LSDVc)方法则是用 LSDV 方法估计动态面板模型即式(6)，记估计系数为 $\hat{\beta}_{LSDV}$ ；其次，估计此方法的偏差，记为 \widehat{Bias} ；最后，两者相减得到一致估计 $\hat{\beta}_{LSDVc} = \hat{\beta}_{LSDV} - \widehat{Bias}$ 。

下面分别用三种方法对动态面板数据进行拟合估计。其中差分 GMM 与系统 GMM

表 8. 三种方法估计结果

Variable	DiffGMM	SysGMM	LSDVc
L1.dln_pigprice	0.445*** 0.163	0.450*** 0.045	0.557*** 0.044
L1.dln_supply	-0.527** 0.180	-0.376*** 0.105	-0.042* 0.023
L2.dln_bd_opinion	-0.171*** 0.060	-0.098*** 0.018	0.080*** 0.016
L1.dln_wb_opinion	0.001 0.022	0.034*** 0.012	0.054*** 0.006
AR(2)	0.111	0.203	
Sargan Test	0.000	0.000	

由表 8，三种方法估计出的系数正负号相同，大部分系数估计值相差不大，说明模型比较稳健。LSDVc 方法估计出的标准差最小，系统 GMM 方法次之，从这一角度看，LSDVc 的确是比 GMM 方法估计得较好。对 GMM 的 Arellano-Bond 中，残差一阶显著自相关，二阶不相关；Sargan 工具变量有效性检验中，差分 GMM 和系统 GMM 都通过检验，说明模型拟合结果可信用度高。

若采用系统 GMM，则模型结果为：

$$y_t = 0.450y_{t-1} - 0.376x_{1,t-1} - 0.098x_{2,t-1} + 0.034x_{3,t-1} \tag{8}$$

式中， y_t 为生猪价格增长率， $x_{1,t}$ 为月度生猪存栏量增长率， $x_{2,t}$ 为“非洲猪瘟”百度指数月均值增长率， $x_{3,t}$ 为“猪价上涨/暴涨/会涨”微博数月均值增长率。该模型各变量系数估计在 0.01 水平下都通过 t 检验。

四、 结论

首先，本文发现影响此次猪价波动的最主要因素是生猪存栏量和网络舆论导向，而猪肉的需求量、替代品价格、成本价格对猪价影响不显著。往往能够较大影响猪价波动的季节性因素在此次猪瘟的大背景下，也没有显著体现其影响力。其次，各地猪价波动差异明显，南方省市的价格普遍比北方高，这种特征在东南沿海城市尤其明显。导致该现象的主要原因是南方是猪肉消耗主力地带，而华北东北为猪肉主产区。动态面板数据分析中得出，上月猪价增长加大对本月猪价的增长有正向推动作用，上月生猪存栏增长加大对本月猪价的增长有抑制作用，上月“非洲猪瘟”百度指数的升高对本月猪价增长有抑制作用，上月微博“猪价上涨/暴涨/会涨”条文数的增加对本月猪价增长有正向作用。基于此，提出政策建议：1). 稳定猪肉供给，减少散养殖户，统一管理; 2). 及时粉碎不实谣言以及禁止哄抬猪价的行为; 3). 由于疫情下活猪的“禁运调令”导致价格在地区间发展不平衡，应发展“冷链物流”以保证产区有销，销区有进。

参考文献

- [1]. 刘国信.市场供应偏紧 猪价高位运行——2019 年生猪市场回顾与未来走势展望[J].农村新技术,2020(02):44-46.
- [2]. 乔浪,郭新宇,彭程.基于多维关联规则的猪肉价格波动原因分析[J].江苏农业科学,2019,47(11):332-335.
- [3]. 刘春鹏,肖海峰.外部冲击对我国肉类价格的影响——基于 SVAR 模型的实证分析[J].中国农业大学学报,2019,24(05):204-213.
- [4]. 范青青,袁艳红.基于 X12-ARIMA 模型的猪肉价格波动规律研究[J].中国畜牧杂志,2018,54(06):138-142.
- [5]. Xi-Xi Zhang, Lu Liu, Chi-Wei Su, Ran Tao, Oana-Ramona Lobonț and Nicoleta-Claudia Moldovan. Bubbles in Agricultural Commodity Markets of China Complexity, 2896479, 1-9. 2019
- [6]. 殷克东,郑义,袁晓彤.基于 GSADF 方法的中国猪肉价格泡沫检验[J].统计与信息论坛,2015,30(10):58-64.
- [7]. 赵豫蒙.网络舆论下农产品稳定营销的策略[J].农业经济,2019(12):125-126.
- [8]. X. Li, J. Ma, S. Wang, et al., How does google search affect trader positions and crude oil prices? Econ. Model. 49 (2015) 162–171
- [9]. 成舟,余峥,过弋,王志宏.基于外汇舆情的人民币汇率波动预测研究[J].计算机科学,2019,46(S2):143-148.
- [10]. Youzhu Li, Huiling Zhou, Zhonglong Lin, Yifan Wang, Shunjie Chen, Chang Liu, Zhouyang Wang, Daniela Gifu, Jingbo Xia. Investigation in the influences of public opinion indicators on vegetable prices by corpora construction and WeChat article analysis[J]. Future Generation Computer Systems, 2020, 102.
- [11]. LW Ling, DB Zhang, SY Chen. Can online search data improve the forecast accuracy of pork price in China? JOURNAL OF FORECASTING DOI: 10.1002/for.2649. JAN 2020
- [12]. 任青山,方遼,朱幸辉.基于多元回归的 BP 神经网络生猪价格预测模型[J].江苏农业科学,2019,47(14):277-281.
- [13]. 吴培,李哲敏.中国猪肉价格预测研究——基于 ARIMA-GM-RBF 组合模型的分析[J].价格理论与实践,2019(01):75-78.
- [14]. Huiping Wang, Meixia Wang. Effects of technological innovation on energy efficiency in China: Evidence from dynamic panel of 284 cities[J]. Science of the Total Environment, 2020, 709.
- [15]. 张志强.动态面板模型参数估计方法的比较研究[J].统计研究,2017(09):110-121.
- [16]. Skrabic, Blanka. Do the most frequently used dynamic panel data estimators have the best performance in a small sample? A Monte Carlo comparison. Croatian Operational Research Review. 10. 45-55. 10.17535/corr.2019.0005.
- [17]. Berzal F, Nicolfás Matín. Data mining: concepts and techniques by Jiawei Han and Micheline Kamber[J]. ACM SIGMOD Record, 2002, 31(2):66-68.
- [18]. 胡浩,戈阳.非洲猪瘟疫情对我国生猪生产与市场的影响[J].中国畜牧杂志,2020,56(01):168-172.
- [19]. 任慧媛.猪价上涨,为何上海受冲击小? [J].中外管理,2019(10):98-101.
- [20]. M Arellano, S Bond. Some tests of specification for panel data: monte carlo evidence and application to employment equations[J]. Review of Economic Studies, 1991. 58(2): 277—297.
- [21]. Jan F. Kiviet. On bias, inconsistency, and efficiency of various estimators in dynamic panel data models[J]. Journal of Econometrics, 1995, 68(1).
- [22]. R Blundell, S Bond. Initial conditions and moment restrictions in dynamic panel data models[J]. Journal of Econometrics, 1998, 87(1): 115—143.