另类数据的信息含量研究——来自电商销售的证据

廖理，崔向博，孙琼

管理世界，2021

叶鑫　2021/11/14

研究背景

- 随着金融市场的发展和市场有效性的增强，传统公开渠道披露的信息大部分会反映在股票价格中，以往定价因子的有效性逐渐减弱，投资者开始探索开始探索利用另类数据中的信息来获取超额收益。

- 另类数据是指投资研究中使用的非传统来源的新型数据。比如大众点评的餐厅评价、淘宝的商品评论等，单个的这些数据点可能是没有规律的，但是，当这些数据被汇聚到一起，就可能反映群体活动的特点和趋势，成为有投资价值的另类投资数据。在中国，有一类重要的另类数据是电商销售数据。

- 大部分研究认为另类数据包含基本面信息，对未来的股票收益具有预测能力，但也有部分研究对另类数据的信息含量提出了质疑，他们认为另类数据中的噪音成分太大。

文献回顾

- Da 等（2011）发现，公司产品的谷歌搜索量指数（Search Volume Index，SVI）具有前瞻性，能有效预测营业收入和股票收益。
  - 缺陷：SVI难以区分消费者的正面情绪或负面情绪，也无法反映实际发生的交易活动。

- Tang（2018）和Bartov 等（2018）通过汇总个人的twitter观点，发现这些聚合信息能显著预测公司即将发布的季度营业收入和盈余告超额收益。
  - 缺陷：社交主体缺乏提供关于产品的真实信息的动机，可能也不具备准确评估产品或服务质量的专业能力，导致评论或观点存在偏差。

- Katona等（2018）利用美国零售店停车场的卫星图像数据，发现汽车数量的增长能够预测公司基本面和股价。
  - 缺陷：无法真实反映公司的未来现金流。

- 组内最近解读文章：SFL、LDA、Footprint、LinkedIn

研究动机

本文主要利用中国电商平台的销售数据来研究另类数据的信息含量问题：

- 使用的是电商销售**月度**数据，**实时性强、颗粒度细**，研究样本量较为充足。
- 电商销售数据是直接的数字信息，受伪信息、假信息及行为偏差等噪音的影响较小，其**信噪比**高于文本类信息。
- 汇总了主要电商平台的每一笔消费记录，**受品牌类别、群体消费特征和偏好的影响较小**，可以全面反映消费者在线上对公司产品的购买情况。
- 电商销售额与公司潜在的销售收入和现金流之间的相关性比现有文献所使用的公司销售代理指标或文本类信息更**紧密**，是一种更直接有效的领先指标。
- 此外，目前关于另类数据的研究主要集中在资本市场发达的国家，国内鲜有文献涉及。在**新兴资本市场**中，另类数据是否也包含了有效信息，是否对公司基本面信息和未来股票收益有预测能力，还尚未可知。

研究问题及设计

- 电商销售数据是否含有与公司基本面相关的信息，能否预测公司的同期财务业绩？
  - 研究假设H1：线上销售额与公司营业收入具有正相关性。
- 线上销售额的变动反映了公司现金和营业收入的未预期变动吗？
  - 研究假设H2：线上销售额的增长率对未预期营业收入和未预期盈余有预测能力。
- 在公司发布盈余公告时，投资者会根据盈余公告发布的当期财务业绩买入或卖出公司股票，未预期盈余和未预期收入更高的企业的股票价格上涨的概率就越大，因此，在H2成立的基础上，增长率指标可能对盈余公告超额收益具有一定的预测能力。
  - 研究假设H3：线上销售额的增长率能够预测盈余公告累计超额收益。
- 如果可以有效预测，那么可能的机制是什么？
  - 一系列异质性检验

研究结论

- 公司的季度线上销售额与季度营业收入具有正相关性。线上销售额的增长率对标准化未预期收入和标准化未预期盈余具有显著的预测作用，这意味着线上销售额传达了有关公司基本面的信息。

- 线上销售额的增长率对盈余公告期间股票的累计超额收益也具有显著的增量预测作用，这表明线上销售额包含的信息会在盈余公告时反映在股价中。

- 线上销售额所含有的有价值信息在盈余公告后会充分反映到股价中，不存在盈余公告后的漂移现象。

- 进一步分析发现，电商销售数据的预测能力在线上销售收入占比高、投资者关注度低、成长型、信息透明度高的公司中更强。

研究贡献

- 本研究所使用的电商销售数据较难获取但含有更高的信息含量，关于这类数据的信息价值问题尚未有相关文献。因此，研究电商销售数据的信息含量和投资价值有助于丰富另类数据文献。

- 对电商销售数据而言，它不仅与公司的营业收入高度相关，而且是一种实时高频的数字信息，信噪比更高，本研究为另类数据包含有价值信息的结论提供了有力的支持，也补充和发展了另类数据在资产定价中的作用的文献研究。

- 基于中国数据探讨消费者行为与股票收益的文献研究还很匮乏，而本文的研究有助于丰富该领域的文献。

研究数据

本文使用的电商销售数据来自Wind资讯，覆盖超过170个A股公司的近700个品牌的电商销售月度数据，样本期间为2015年1月~2018年9月：

- 这套数据收集于天猫、淘宝、京东、一号店等主流电商平台，范围覆盖了95%以上的中国电商市场，保证了线上销售数据**总量的真实性**，能够全面反映消费者在**线上**对公司产品的购买情况。

- 线上品牌与线下品牌的重合度超过90%，即这些A 股上市公司在线上和线下销售的品牌几乎一致，可以衡量公司所有品牌**销售的整体情况**。

- 消费类上市公司大多采用了线上线下融合的商业模式，线上线下渠道在公司营销渠道中的重要程度越来越接近，**基于线上渠道预测公司整体情况具有合理性和可靠性。**

出于部分指标计算滚动（8季度）的需要，本文最终选取了在2013年之前上市的132个消费行业上市公司作为研究样本。

实证分析——假设检验H1

- 为了检验线上销售额和营业收入是否存在正相关关系，本文采用以下模型：

$$Log\_Revenue_{i,t} - log\_Revenue_{i,t-1} = \beta_0 + \beta_1 \times (Log\_Sales_{i,t} - Log\_Sales_{i,t-1}) + \varepsilon_{i,t}$$

- 从列（1）~列（4）的结果表明，每个季度的线上销售额与该季度的营业收入高度相关。这说明线上销售额含有与营业收入相关的信息，这是电商销售数据可以作为预测公司基本面的先行指标的基础。

表3 线上销售额与营业收入的相关性

	Revenue Growth			
	(1)	(2)	(3)	(4)
Δlog_Sales	0.267*** (15.75)	0.269*** (15.09)	0.300*** (15.06)	0.305*** (14.51)
Constant	0.010 (1.15)	0.010 (1.10)	0.038 (1.28)	0.047 (1.50)
Firm fixed effects	No	Yes	No	Yes
Time fixed effects	No	No	Yes	Yes
N	1338	1338	1338	1338
R^2	0.159	0.159	0.196	0.196

实证分析——假设检验H2

接下来，本文检验线上销售额的增长率$SG_{i,t}$是否对标准化的未预期营业收入$SUR_{i,t}$有预测作用。

- 核心解释变量：$SG_{i,t} = \log(\text{Sales}_{i,t}) - \log(\text{Sales}_{i,t-4})$
- 被解释变量：$SUR_{i,t} = \frac{REV_{i,t} - REV_{i,t-4}}{\delta_{i,t}}$
- $SUR_{i,t} = \beta_0 + \beta_1 \times SG_{i,t} + \beta_2 \times SUR_{i,t-1} + \beta_2 \times X_{i,t-1} + \varepsilon_{i,t}$
- 由于另类数据的预测作用是在财务报告未公布时，对本季度的财务指标和基本面信息进行预测，预测所依赖的信息是t季度的电商销售数据与t季度之前的财务信息。
- 因此，在该模型中需要使用t-1季度的控制变量$SUR_{i,t-1}$、$X_{i,t-1}$（对数化的总市值、账面市值比、资产收益率、盈余公告前30天至公告前3天的累计收益率）。

表4 电商销售数据对*SUR*的预测

	SUR		
	(1)	(2)	(3)
SG	0.297*** (2.92)	0.307*** (3.03)	0.270*** (2.66)
Lag_SUR			0.135*** (5.76)
Lag_Size		0.442** (2.25)	0.280 (1.42)
Lag_BM		0.304 (1.26)	0.139 (0.58)
Lag_ROA		−0.075** (−2.23)	−0.102*** (−3.05)
Pastreturn		1.222*** (2.66)	1.057** (2.32)
Constant	0.668*** (4.61)	−9.866** (−2.11)	−6.027 (−1.29)
Firm fixed effects	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes

实证分析——假设检验H2. 1

然后，本文检验线上销售额的增长率是否可以预测标准化未预期盈余。回归方程如下：

- $SUE_{i,t} = \frac{\text{Earnings}_{i,t} - \text{Earnings}_{i,t-4}}{\xi_{i,t}}$
- $SUE_{i,t} = \beta_0 + \beta_1 \times SG_{i,t} + \beta_2 \times SUE_{i,t-1} + \beta_3 \times SUR_{i,t} + \beta_4 \times SUR_{i,t-1} + \beta_5 \times X_{i,t-1} + \varepsilon_{i,t}$
- Froot等（2017）发现同期的标准化未预期营业收入和标准化未预期盈余有很高的相关性。所以本文也需要探讨线上销售额指标$SG_{i,t}$对$SUE_{i,t}$的预测作用是否是来自$SUR_{i,t}$。

表 5　电商销售数据对 *SUE* 的预测

	SUE			
	(1)	(2)	(3)	(4)
SG	0.345*** (3.40)	0.348*** (3.43)	0.315*** (3.06)	0.229** (2.33)
SUR				0.314*** (11.16)
Lag_SUE			0.044** (2.26)	0.049*** (2.63)
Lag_SUR			0.040* (1.68)	−0.003 (−0.12)
Lag_Size		0.563*** (2.86)	0.522*** (2.61)	0.435** (2.28)
Lag_BM		0.180 (0.75)	0.125 (0.51)	0.079 (0.34)
Lag_ROA		−0.083** (−2.46)	−0.113*** (−3.23)	−0.083** (−2.48)
Pastreturn		0.754 (1.64)	0.722 (1.56)	0.397 (0.90)
Constant	0.203 (1.40)	−13.072*** (−2.80)	−12.144** (−2.56)	−10.262** (−2.27)
Firm fixed effects	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
N	1338	1338	1327	1327
R^2	0.037	0.049	0.055	0.145

实证分析——假设检验H3

接着本文检验线上销售额的增长率$SG_{i,t}$是否对本季度盈余公告累计超额收益CAR有预测作用：

- 第一种来源可能是$SUR_{i,t}$和$SUE_{i,t}$中的信息在发挥作用。
- 另一种来源是电商销售数据中含有$SUR_{i,t}$和$SUE_{i,t}$之外的信息。比如，投资者可能会对财务报告或业绩报告会中的非财务信息做出反应，而这些非财务信息可能与线上销售额的变动有相关性。

因此，本文采用两个回归模型来检验$SG_{i,t}$对盈余公告超额收益的预测作用：

- $\mathrm{CAR}[-1,+3] = \beta_0 + \beta_1 \times SG_{i,t} + \beta_2 \times X_{i,t} + \varepsilon_{i,t}$
- $\mathrm{CAR}[-1,+3] = \beta_0 + \beta_1 \times SG_{i,t} + \beta_2 \times SUE_{i,t} + \beta_3 \times SUR_{i,t} + \beta_4 \times X_{i,t} + \varepsilon_{i,t}$

实证分析——假设检验H3

- 列（1）不加入控制变量，$SG_{i,t}$的系数是0.008，且在5%水平上显著。该结果的经济意义同样显著，线上销售增长率增加一单位对应的累计超额收益增长0.8%，近似等于66.7%的年化收益率。
- 列（3）中的结果说明当期的未预期盈余和当期的未预期收入都是造成盈余公告超额收益的重要因素，而市场投资者对未预期盈余中所含信息的反应要超过对未预期收入中所含信息的反应，这与以往相关文献的研究相符。

表6 电商销售数据对*CAR*[−1,+3]的预测

	CAR[−1,+3]		
	(1)	(2)	(3)
SG	0.008** (2.31)	0.008** (2.40)	0.007** (1.97)
SUE			0.006*** (4.83)
SUR			0.002* (1.90)
Size		−0.022*** (−3.26)	−0.026*** (−3.82)
BM		0.004 (0.47)	0.002 (0.29)
ROA		0.003** (2.46)	−0.001 (−0.43)
Pastreturn		0.002 (0.14)	−0.004 (−0.29)
Constant	0.016*** (3.23)	0.522*** (3.27)	0.612*** (3.85)
Firm fixed effects	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes
N	1338	1338	1338
R^2	0.035	0.052	0.078

实证分析——假设检验H3

本文还检验了电商销售数据的信息是否也存在盈余公告后的漂移现象（PEAD）

- $\mathrm{CAR}[+4,+60] = \beta_0 + \beta_1 \times SG_{i,t} + \beta_2 \times SUE_{i,t} + \beta_3 \times SUR_{i,t} + \beta_4 \times X_{i,t} + \varepsilon_{i,t}$
- 虽然表7 中$SG_{i,t}$的系数始终大于0，但并不显著。这说明市场投资者对电商销售数据中的增量信息在盈余公告窗口期内有充分的反应，不存在公告后的漂移现象。

表 7　电商销售数据对 *CAR*[+4,+60]的预测

	CAR[+4,+60]		
	(1)	(2)	(3)
SG	0.002 (0.26)	0.007 (0.80)	0.007 (0.77)
SUE			0.004 (1.26)
SUR			−0.002 (−0.75)
Size		−0.118*** (−6.85)	−0.118*** (−6.78)
BM		0.022 (1.12)	0.023 (1.17)
ROA		0.004 (1.23)	0.002 (0.61)
Pastreturn		0.079** (2.00)	0.079** (1.98)
Constant	0.062*** (4.84)	2.767*** (6.85)	2.763*** (6.78)
Firm fixed effects	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes
N	1338	1338	1338
R^2	0.067	0.129	0.130

异质性检验分析—指标噪音与信息含量

- 理论假设：电商销售数据可能具有信息含量的原因是这些直接来自公司销售活动的数据既包含了与公司收入、利润相关的现金流信息，也包含了公司产品市场层面的信息，从而可以预测同时期的基本面财务指标。

- 验证方法：根据线上销售额占公司营业收入的比重，将研究样本分为两组进行异质性检验。

- 预期结果：如果线上销售额占公司营业收入的比重越高，则电商销售数据能够反映公司实际经营状况的准确性就越高（即根据电商销售数据构造的预测指标的噪音就越少），对基本面指标（SUE、SUR)的预测作用就越强。

异质性检验—指标噪音与信息含量

- 无论是对未预期收入的预测还是对未预期盈余的预测，在线上销售额占营业收入比重高的样本组中的预测作用都要优于线上销售额占比低的样本组，即电商销售数据在线上占比高的样本组的信息含量更多。
- 这也证明了电商销售数据的预测作用和信息含量的来源确实是因为电商销售数据直接来自公司的销售记录，本身就包含了公司经营活动相关的基本面信息。

表 8　噪音高低与基本面预测

	SUR		*SUE*	
	线上占比大	线上占比小	线上占比大	线上占比小
SG	0.310* (1.70)	0.219* (1.70)	0.632*** (3.22)	0.197 (1.63)
Lag_SUE			0.004 (0.17)	0.121*** (3.50)
Lag_SUR	0.191*** (4.67)	0.164*** (4.14)	0.073* (1.66)	0.019 (0.66)
Lag_Size	0.175 (0.56)	0.136 (0.51)	0.741** (2.20)	0.381 (1.49)
Lag_BM	−0.178 (−0.50)	0.189 (0.53)	0.514 (1.34)	−0.269 (−0.80)
Lag_ROA	−0.117** (−2.39)	−0.096** (−1.99)	−0.107** (−1.97)	−0.135*** (−2.89)
Pastreturn	1.311** (2.08)	0.837 (1.23)	1.014 (1.49)	0.585 (0.91)
Constant	−3.534 (−0.47)	−2.794 (−0.44)	−17.716** (−2.22)	−8.528 (−1.41)
Firm fixed effects	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
N	668	659	668	659
R^2	0.128	0.080	0.078	0.072

异质性检验—投资者关注与信息含量

- 电商在中国已经发展了十余年时间，一些证券分析师已将电商数据作为分析公司基本面的重要信息源。
- 其中的逻辑是，被投资者关注更多的公司，电商销售数据所含的信息更可能在盈余公告前就被投资者使用，导致投资者关注度高的公司在盈余窗口期的超额收益要比投资者关注度低的公司的超额收益小。
- 代理变量：分析师人数和公司市值
- 结果显示，在分析师关注度低的公司和小市值的公司中，线上销售额增长率对盈余公告超额收益的预测作用更强。

表 9　投资者关注度与 *CAR*[−1，+3]预测

	CAR[−1，+3]			
	分析师关注度高	分析师关注度低	大市值	小市值
SG	0.005 (0.92)	0.009** (2.21)	0.004 (0.74)	0.008* (1.84)
SUE	0.007*** (3.51)	0.003** (2.06)	0.004** (2.03)	0.007*** (4.16)
SUR	0.001 (0.69)	0.002 (1.54)	0.003* (1.86)	0.000 (0.27)
Size	−0.032** (−2.20)	−0.035*** (−3.71)	−0.021 (−1.46)	−0.037*** (−3.24)

异质性检验—公司成长性与信息含量

- 电商销售数据很可能包含与公司未来经营前景，特别是消费者对公司产品的需求或购买意愿相关的信息。
- 对于成长型公司，投资者可能对这些反映公司前景的信息反应更为强烈，因为成长型公司的市场估值会对公司未来的潜在现金流和增长前景更为敏感。
- 根据账面市值比大小分为成长型公司和价值型公司。
- 结果从投资者对成长型公司和价值型公司的信息反应程度的层面证明了电商销售数据对股票收益预测作用的来源确实是投资者反应造成的。

表 10　公司成长性与 *CAR*[−1，+3]预测

	CAR[−1，+3]	
	成长股	价值股
SG	0.009* （1.85）	0.001 （0.24）
SUE	0.005*** （2.90）	0.004** （2.47）
SUR	0.002 （1.23）	0.002 （0.97）
Size	0.002 （0.09）	−0.045*** （−5.31）

异质性检验—信息透明度与信息含量

- 本文已证明正是由于投资者对电商销售数据中的信息做出反应，才使得电商销售数据对股票收益具有预测作用。而公司信息的透明度也会直接影响投资者对信息的使用行为。
- 如果公司存在盈余管理，或者公司的财务信息透明度较差，则电商销售数据的预测作用是更强还是更弱？
- 回归结果表明电商销售额这种非财务信息的预测能力会受到公司财务信息操纵程度的影响，在盈余管理程度低的公司中，这种非财务信息的预测作用更好。

净利润的平滑度	t季度之前12个季度的净利润的标准差除以t季度之前12个季度的经营性现金流的标准差
应计利润的波动率	t季度之前12个季度的应计利润的标准差，其中，应计利润等于净利润减去经营性现金流

	$CAR[-1,+3]$			
	应计波动性高	应计波动性低	盈余平滑度高	盈余平滑度低
SG	0.004 (0.69)	0.013*** (2.85)	0.008* (1.71)	0.006 (0.98)
SUE	0.005*** (3.33)	0.007*** (3.79)	0.005*** (3.15)	0.006*** (3.65)
SUR	0.001 (0.62)	0.002 (1.17)	0.001 (0.59)	0.003* (1.75)
Size	−0.027** (−2.20)	−0.024*** (−2.59)	−0.039*** (−3.34)	−0.022** (−2.43)

研究结论

- 研究结果表明，电商销售额的增长率可以显著预测未预期收入和未预期盈余，同时对盈余公告累计超额收益也有显著预测能力。

- 本文进一步通过异质性检验分析了不同公司特征对预测效果的影响。指标噪音、投资者关注度、公司成长性和公司财务信息透明度都会影响电商销售数据的预测效果和在股价中的反映程度。

- 本文的研究提供了消费者行为可以传递有价值信息的新的经验证据。最后，已有研究主要探讨了成熟资本市场中另类数据的信息含量，但鲜有文献探讨新兴资本市场中另类数据的价值相关性，而本研究提供了有益补充。

启示

- 电商销售增长率和异质性检验指标可以作为双变量分组，构造投资组合，
- 对于非预期收益和非预期盈余指标的构建过于简单，
- 电商销售增长率指标的构建也较为简单，可以考虑将加入波动系数、将同比增长率和环比增长率结合起来。