香港投资者的选股公式

阮启宏

2017 年 3 月 17 日，港交所开始公布每个交易日香港投资者的完整持仓数据，这是研究香港投资者选股公式的宝贵素材。因为在此之前，我只能通过上市公司的年报和季报中的十大流通股东来了解香港投资者的持仓，这种方法有两个问题：一是时间滞后，二是获取的数据不完整，属于截断数据，在进行回归分析时会导致系数估计有偏。有了港交所公开的数据，这两个问题迎刃而解。我先用沪股通（2014 年 11 月 17 日开通）的数据来估计香港投资者的选股公式，然后再用深股通（2016 年 12 月 5 日开通）的数据来做样本外预测。

本文在研究香港投资者选股公式时使用的基本模型如下：

$$Hold_i = X\beta + \gamma_{IND} + \varepsilon_i$$

其中$Hold_i$是香港投资者对股票 i 经价格调整后的持股市值，用 2017 年 4 月 28 日沪股通持股数量乘以 2014 年 11 月 14 日的前复权收盘价得到，这样计算指标的目的是消除沪港通开通后的股价变动对持股市值产生的影响，避免系数估计有偏差；γ_{IND}是行业固定效应，本文分别考虑 4 种行业划分方法：申万一级行业、申万二级行业、Wind 一级行业和 Wind 二级行业，X 是关于公司特征的变量，ε_i是扰动项。

关于公司特征变量方面，本文参考标准文献（韩乾等，2014）选取公司特征变量，包括公司年龄(AGE)、资产收益率（ROA）、资产负债率(LEVERAGE)、公司规模(SIZE)、股票平均价格（NP）、累计收益率（CR）、基金持股比例（INST）和总市值（ACAP），一共 9 个，具体指标的定义和计算公式如下：

（1）公司年龄(AGE)。以往的金融文献认为，公司上市时间是知情交易者选股的指标之一。本报告选取公司至 2014 年 11 月 16 日的上市时间来计算公司年龄。

（2）资产收益率（ROA）。资产收益率高的股票，其盈利能力强，可能被知情交易者列为选股标准。资产收益率，即公司净利润与资产总额之比，用来衡量每单位资产创造多少净利润的指标，被广泛用于衡量公司盈利能力的指标之一。本报告我们采用 2013 年年报公布的公司财报计算资产收益率数据。

（3）市净率(BM)。市净率即每股股价与每股净资产的比率，被广泛用于判断股票的购买价值。本报告我们采用 2014 年三季报公布的公司财报和 2014 年 11 月 16 日的收盘价计算市净率。

（4）资产负债率(LEVERAGE)。资产负债率指公司负债总额与资产总额之比，是反映公司的财务杠杆的指标之一。本文采用 2014 年三季报公布的公司财报计算个股的资产负债率。

（5）公司规模(SIZE)。通常用公司资产数额作为衡量其规模的指标。本文采用 2014 年三季报公布的公司财报中的资产作为衡量其公司规模的指标。

（6）股票平均价格（NP）。在公司金融领域，一般认为股票的价格会影响股票的收益率。因此，我们有理由认为知情交易者在选取股票的时候也会考虑股票本身的价格。本文采用 2014.9.17-2014.11.16 期间股票每日收盘的平均价格作为股票平均价格衡量指标。

（7）累计收益率（CR）。在实际中，采用动量策略的投资者会偏好累计收益率较高的股票进行投资，而采用反转策略的投资者会偏好累计收益率较低的股票。本文采用 2014.9.17-2014.11.16 期间收益率的累加作为累积收益率的衡量指标。

（8）基金持股比例（INST）。基金持股比率的高低可以反映信息透明度的状况，越高的基金持股比率对应着越低的信息透明度。知情交易一般伴随着较低的信息透明度，且机构持

股比例越高被操纵的可能性也对也大。本报告采用 2014 年三季报中期财务报表（2014 年 9 月 30 日）中披露的机构持股比例来衡量公司信息的透明度。

（9）总市值（ACAP）。股票总市值是投资者做投资决策时考虑的重要因素。本文将用 2014.9.17-2014.11.16 期间日均总市值（ACAP）来衡量公司规模。

关于样本选择，沪股通刚实施时一共有 569 个标的，我剔除了 2014 年 11 月 17 日后进入或被剔除的个股，得到 405 个样本；再剔除 2014.9.17-2014.11.16 无交易数据的样本（原因是计算公司特征变量需要用到交易数据），得到 395 个样本；再剔除石油化工、石油采掘和金融行业的样本（原因是这三个行业集中分布在沪市，对深市的参考价值不大），最后得到 357 个样本。本文数据均从 Wind 数据库获得。

1.描述性统计

表 1 展示了本文使用的关于香港投资者持股数据和上市公司特征变量的描述性统计信息。其中 Hold 代表 2017 年 4 月 28 日沪股通的持股市值，平均值为 2.16 亿元，最大值达到 96.6 亿元。

表 1　主要变量描述性统计

变量	观察值	平均数	标准差	最小值	最大值
股票平均价格（元）	357	12.12	10.41	2.49	140.50
总市值（万元）	357	2.13E+06	2.80E+06	2.69E+05	3.01E+07
上市时长（年）	357	13.39	5.53	0.82	23.93
资产收益率	357	0.05	0.04	-0.05	0.32
市净率	357	3.04	2.10	0.69	16.88
资产负债率	357	0.52	0.19	0.06	0.88
总资产（万元）	357	4.23E+06	9.85E+06	1.13E+05	9.25E+07
累计收益率	357	0.10	0.14	-0.17	1.01
基金持股比例	357	0.50	0.21	0.00	0.92
持股市值（元）	357	2.16E+08	8.47E+08	0.00	9.66E+09

2.回归分析

我进行了 6 组回归，结果报告在表 2。第（1）组回归的被解释变量是持股市值，对股价对数、上市时长、资产收益率、市净率、资产负债率、累计收益率、基金持股比例、总市值和总资产进行回归，不加入行业固定效应，结果显示，沪股通的持股市值与股价对数、上市时长、资产收益率、资产负债率、累计收益率和总市值正相关，而与市净率、基金持股比例和总资产负相关；在第（2）组回归中，我将被解释变量换成持股市值的对数，总市值和总资产分别换成它们的对数，系数的符号与第（1）组回归基本一致，第二组回归的 R 方为 0.532，高于第（1）组回归的 0.420，因为两组回归的解释变量个数相同，因此从 AIC 和 BIC 两个模型选择准则来看，模型（2）均优于模型（1）；在第（3）-（6）组的回归中，我分别在第（2）组的回归的基础上控制申万一级行业、申万二级行业、万得一级行业和万得二级行业的固定效应。

表 2　回归分析结果

	(1)	(2)	(3)	(4)	(5)	(6)
变量	持股市值	持股市值对数	持股市值对数	持股市值对数	持股市值对数	持股市值对数
股价对数	5.564e+07	0.121	-0.0144	-0.00249	0.0931	0.0713
	(0.69)	(1.17)	(-0.13)	(-0.02)	(0.92)	(0.68)
总市值对数		1.404***	1.284***	1.346***	1.411***	1.361***
		(6.71)	(5.53)	(4.70)	(6.62)	(6.09)
总资产对数		-0.202	-0.102	-0.126	-0.197	-0.167
		(-1.15)	(-0.53)	(-0.55)	(-1.09)	(-0.87)
上市时长	1.267e+07**	0.0282**	0.0263**	0.0371***	0.0242**	0.0267**
	(2.30)	(2.54)	(2.05)	(2.60)	(2.14)	(2.26)
资产收益率	5.200e+09**	6.283***	5.684***	5.224**	6.209***	6.401***
	(2.53)	(3.93)	(3.38)	(2.55)	(3.80)	(3.85)
市净率	-3.420e+07	-0.0760	-0.0574	-0.0622	-0.0905*	-0.0818
	(-1.33)	(-1.57)	(-1.13)	(-1.06)	(-1.83)	(-1.44)
资产负债率	2.927e+08	0.635	0.326	0.566	0.756	0.721
	(1.20)	(1.11)	(0.57)	(0.85)	(1.32)	(1.20)
累计收益率	3.365e+08*	-0.0911	0.0487	0.269	0.0685	0.0316
	(1.87)	(-0.25)	(0.14)	(0.62)	(0.19)	(0.08)
基金持股比例	-2.525e+08	-0.770***	-0.558*	-0.551*	-0.671**	-0.620**
	(-1.41)	(-2.74)	(-1.88)	(-1.75)	(-2.50)	(-2.21)
总市值	208.0***					
	(2.65)					
总资产	-26.42**					
	(-2.46)					
行业固定效应?	不控制	不控制	申万一级行业	申万二级行业	万得一级行业	万得二级行业
观测数	357	356	356	356	356	356
R 方	0.420	0.532	0.578	0.681	0.557	0.572

注：括号内是用 White 稳健标准误计算得到的 t 值，　*** p<0.01, ** p<0.05，*p<0.1；第（2）-（6）组回归的观测数为 356，因为有 1 个样本的持股市值恰好为 0，因此取对数后无意义。

3.预测效果

如果我对香港投资者的选股公式估计得准确，那么我应该可以准确地预测到深股通会偏好哪些股票。我关注的是深股通买入最多的几只股票（本文定为 10 只），而不是整体的预测效果，因为从个人投资的角度出发，选出他们最可能投资的前 10 只股票进行集中投资的做法比较可行。因此，我选择的衡量预测效果的指标是：预测的前十名个股和实际的前十名个股的重合比例。

预测结果如表 3 所示。预测 1-预测 6 分别是用模型（1）-（6）预测得到的结果。预测 1、2、3、5 和 6 的重合比例为 50%。这在统计上是一个非常高的比例，因为进行样本外预测的股票一只有 885 只，假设模型没有预测能力，完全随机选股，选中的股票大于或等于 5 只的概率为：

$$P(N \geq 5) = \frac{C_{875}^{5} \cdot C_{10}^{5} + C_{875}^{4} \cdot C_{10}^{6} + C_{875}^{3} \cdot C_{10}^{7} + C_{875}^{2} \cdot C_{10}^{8} + C_{875}^{1} \cdot C_{10}^{9} + C_{875}^{0} \cdot C_{10}^{10}}{C_{885}^{10}} = 1.39 \times 10^{-9}$$

值得一提的是，选中的 5 只股票全是实际值中的前 5 名。重合比例最低的预测 4 也有 40%的重合率，

$$P(N \geq 4) = \frac{C_{875}^{6} \cdot C_{10}^{4} + C_{875}^{5} \cdot C_{10}^{5} + C_{875}^{4} \cdot C_{10}^{6} + C_{875}^{3} \cdot C_{10}^{7} + C_{875}^{2} \cdot C_{10}^{8} + C_{875}^{1} \cdot C_{10}^{9} + C_{875}^{0} \cdot C_{10}^{10}}{C_{885}^{10}} = 1.68 \times 10^{-6}$$

在 1%的水平上显著，即模型对香港投资者的选股有显著的预测能力。

表 3　预测结果汇总

持股市值排名	预测 1	预测 2	预测 3	预测 4	预测 5	预测 6	实际值
1	分众传媒	分众传媒	分众传媒	格力电器	万科 A	分众传媒	**格力电器**
2	万科 A	恺英网络	五粮液	美的集团	分众传媒	恺英网络	**海康威视**
3	海康威视	万科 A	双汇发展	五粮液	恺英网络	温氏股份	**美的集团**
4	温氏股份	海康威视	洋河股份	万科 A	温氏股份	五粮液	**五粮液**
5	美的集团	温氏股份	万科 A	洋河股份	海康威视	海康威视	**洋河股份**
6	恺英网络	五粮液	海康威视	双汇发展	五粮液	洋河股份	**云南白药**
7	五粮液	美的集团	美的集团	恺英网络	洋河股份	双汇发展	**东阿阿胶**
8	格力电器	格力电器	格力电器	分众传媒	双汇发展	万科 A	**蓝思科技**
9	比亚迪	洋河股份	恺英网络	老板电器	美的集团	美的集团	**威孚高科**
10	洋河股份	双汇发展	温氏股份	小天鹅 A	格力电器	格力电器	**索菲亚**
重合比例	50%	50%	50%	40%	50%	50%	**100%**

通过这个研究，我对香港投资者的投资模式有了更深的认识，初步总结出香港投资者的选股公式，这对个人的资产组合管理具有重要的现实意义。现在还有做得不够完善的地方，比如预测前十名的准确率只有 50%，可能有一些香港投资者看重或者规避的因素我没有捕捉到，希望有不久的将来可以做得更好。