华东理工大学2023 –2024学年第2学期

《大数据与金融计算》实验报告

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 实验名称 | 中国股市收益率可预测性的实证检验 | | | | | | | | | | |
| 专 业 | | 计算机与金融 | 姓名 | 朱卫 | 学号 | 21012994 | | 班级 | | 计金(双)210 |
| 实验时间 | | 2024/4/10 | | 实验地点 | 信息楼319 | | 指导教师 | | 蒋志强 | |

|  |
| --- |
| 实验目的/要求 |
| 1. 掌握收益率可预测性的样本内和样本外检验方法 2. 掌握收益率可以预测性的模型评价方法（统计意义和经济意义）。 3. 掌握机器学习方法：LASSO回归、Ridge回归和ElasticNet方法。 |
| 实验内容 |
| 1. 认真阅读三篇文献资料（自己也可以下载相关文献进行阅读），了解收益率可预测性的相关研究，重点关注其中的研究思路和方法。 2. 文件1EData\_PredictorData2019.xlsx给出了文献“Time-series and cross-sectional stock return forecasting: New machine learning methods”的研究数据，其中：   D12 表示 the twelve-month moving sum of S&P 500 dividens  E12 表示 the twelve-month moving sum of S&P 500 earnings  tbl 表示 three-month treasury bill yield  lty 表示 ten-year treasury bod yield  AAA 表示 AAA-related corporate bond  Rfree 表示 无风险利率  CRSP\_SPvw 表示 the value-weighted S&P 500 return  请你以上数据构建论文Sec 3中的12个预测因子，进行收益率可预测性研究，风险偏好系数设为*γ* = 5。（1）构建单因子可预测性模型，对12个因子分别进行可预测性实证检验；（2）构建多因子可预测性模型，用OLS回归进行预测性实证检验；（3）利用LASSO、Ridge和ElasticNet回归对多因子模型进行可预测性检验。   1. 从锐思数据库选取一只股票（股票最后两位代码与你学号最后两位一样，数据覆盖范围2001-2022），下载股票日度数据、月收益率、无风险利率、月市盈率、月每股收益、净资产收益、每股营业收入、月换手率、月beta系数。并计算月波动率（月内日收益率的平方和）、月流动性（|月收益率 / lg(月成交额)|），月股价高点（当月股价最高值与前三个月股价的最大值的比值，用日度数据计算），月已实现偏度（参考论文2018金融研究-中国股票市场的已实现偏度与收益率预测）。   请利用以上数据进行收益率可预测性研究，风险偏好系数设为*γ* = 5。（1）构建单因子可预测性模型，探索哪些因子具有预测性；（2）结合机器学习的方法，探索多因子模型对股票收益率的可预测性。 |
| 实验总结  **请提供对本次实验结果的讨论分析，以及实验的心得和体会。包括对知识点的掌握，算法的理解，以及对理论课程和实验课程改进的建议。（不少于500字）** |
|  |
| 教师批阅： 实验成绩：  教师签名: 日期： |

**实验报告正文：**

（每次实验报告均为一篇小论文，因此，统一按照学术论文的要求完成实验报告正文，应包括：题目、摘要、文献综述、模型和方法、结果和讨论、参考文献、附录，具体格式如下：

**中国股市收益率可预测性的实证检验**

计金(双)210 21012994 朱卫

摘要：

1 文献综述

金融预测一直是金融领域所瞩目的焦点问题。作为金融市场中的重要组成部分，由于股票市场潜在的高额利润与重要金融价值，大量投资者和学者致力于研究股市中的预测问题[1]。有效的金融预测不仅可以增加金融市场参与者的收益，也为金融发展和决策奠定基础[2]。其中关于收益率预测的研究最受关注。在理论金融的研究中，Fama提出的有效市场理论认为价格及时地反映了市场所有信息，因而无法通过过去的信息预测未来的收益率。然而现实中真正有效的市场往往并不存在，这意味着人们可以找到某些历史信息来预测未来的收益率，因此研究者们不断寻找有效的历史信息，又称为因子。这类研究的基础之一是FamaFrech的三因子模型[3]，其提出了市值因子SMB，账面市值比因子HML，市场风险因子MKT，实证研究证明了这三个因子对于市场有较强的解释能力。此后有大量可解释收益率的因子被挖掘，通过实证研究发现收益率和这些因子之间存在一定的规律，这些规律在样本内成立。但是将这些因子应用于样本外预测时并不一定能够取得较好的效果，而样本外预测往往才有现实意义。因此越来越多的研究不仅仅局限于样本内解释能力检验，同时也会进行样本外预测能力的检验。

陈坚在中国股票市场尾部风险与收益率预测研究中，选取了2001年1月至2012年12月上证和深证所有的 A 股非 ST 股票的市值加权市场投资组合，使用Copula 方法和极值理论方法分别构建尾部风险VaR，研究VaR对于收益率的预测能力，实证研究表明，基于极值方法的VaR对于中国市场有样本内预测能力，且预测能力最强，但是Copula 方法所得VaR没有解释能力。在样本外预测研究中发现，基于极值的VaR单因子在统计意义上优于历史平均值的预测，且预测能力明显高于其他经济变量。最终证明了通过极值VaR构建的组合夏普比率最高[4]。陈坚, 张轶凡在2018的研究中利用高频股指数据构造了中国市场的已实现偏度，通过实证检验发现该因子在样本内和样本外都具有较好的预测能力[5]。蒋志强, 田婧雯, 周炜星在2019年的研究中选取2006年7月到2016年6月沪深A股的全部股票超额收益率为研究对象，构造不同投资组合并且进行分组，选取了8 个预测因子：账面市值比、股利分配率、股息价格比、股息收益率、每股收益价格比、现金收益价格比、通货膨胀率、股票波动率。通过研究发现这些因子在不同组合中都有一定的预测能力，并且发现总体市场和行业组合的收益率在牛市阶段的可预测性远强于熊市，但是账面市值比组合和市值组合的则完全相反[6]。Rapach D E, Zhou G.在2020年的研究中将组合弹性网络方法应用于收益率预测研究中，使用S&P 500指数的月度超额回报作为预测目标，构造了12个预测因子，发现组合弹性网络方法预测在统计上意义和经济意义上相比传统模型都有显著的增强[7]。R.David Mclean, Jeffrey Pontiff在2016年的研究中从另一个角度研究收益率的可预测性，其分析了97个各个期刊中发表的预测变量。研究比较了每个预测变量在三个不同时间段的回报：原始研究的样本期间、原始样本之后但出版之前的期间，以及出版之后的期间，研究发现学术研究的发表可能会导致投资者利用这些研究进行交易，从而减少了这些股票的预期回报，降低因子的预测能力[8]。

本文对于美国和中国股票市场的收益率可预测性进行研究，在对于美国市场的研究中，采用和Rapach D E, Zhou G.在2020年的研究中相同的方法。在对于中国市场的研究，随机选取市场中的一只股票，构造多个因子，利用不同的机器学习方法进行预测性研究。

2 模型和方法

2.1线性回归

线性回归在机器学习中是一种有监督学习方法，也是一种常用的数据预测与分析方法，其被用来描述因变量与一个或多个自变量之间的线性关系，即因变量可以通过自变量的线性组合来表示。线性回归模型可以表示为：

(1)

式1中，为自变量，y是因变量，表示各自变量的权重，为常数项，又称为截距。

线性回归的损失函数被定义为真实值和拟合值之间的平方误差之和：

(2)

式中，是模型拟合值，y是真实值。

其参数的确定采用最小二乘法，即通过最小化这个损失函数来找到最佳的模型参数，使得模型预测值与真实值之间的差异平方和达到最小。

线性回归模型形式简单，容易理解。它假设因变量与自变量之间存在简单线性关系，使得模型易于解释。同时其计算效率高，特别是在数据量较大的情况下，由于其优化目标的解析解可得，训练过程通常比许多其他复杂模型更快。

2.2 Lasso

Lasso回归是一种线性回归分析方法，它通过引入L1正则化项对模型进行约束。L1正则化是一种防止模型过拟合的方法，在L1正则化中，引入了L1范数（又称正则化项），将L1范数加入损失函数中，使得损失函数变为公式3所示：

(3)

式3中，是预测值，是真实值，是各个系数。

Lasso回归作为线性模型，其优化过程是使损失函数最小化。模型的过拟合往往是模型参数相对输入数据过于复杂，数据样本不足以代表全部特征，但是这些局部特征被模型学习到，导致其在样本外预测中出现较大的偏差。在损失函数中加入了正则化项后，使得损失函数增大，为了降低损失函数，除了使预测值更接近真实值之外，还需要降低模型参数的大小，甚至丢弃某些参数(参数系数为0)，可以降低模型的复杂度，降低过拟合风险。

2.3 岭回归

岭回归也是一种引入正则化操作的线性模型，然而岭回归引入的是L2正则化，其损失函数中加入了L2正则化项，可以表示为公式4。

(4)

式4中，是预测值，是真实值，是各个系数的平方。

岭回归通过最小化损失函数获取线性模型的参数从而进行预测。采用L2正则化的好处是L2正则化倾向于将参数值缩小到接近零但不为零，而L1正则化会将参数推向0，使参数变得稀疏。下面是对此现象的解释：

假设没有加入正则化项的损失函数对于某个参数的偏导在w=0处的值为d0，加入L1正则化项后，在w=0处其偏导为d0-或者d0+，而L2正则化在w=0处偏导为d0+ = d0，可见在w=0处的导数仍是d0，然而在L1正则化中，在w=0处会有一个突变，如果d0-和d0+异号，则在0处存在一个极小值点，导致模型可能取该极小值点作为最终优化结果，容易使参数为0。

2.4 弹性网络

弹性网络(ElasticNet)是一种统计建模方法，特别适用于线性回归问题，其设计目的是为了综合Lasso回归（L1正则化）和岭回归（L2正则化）的优点，同时克服它们各自的局限性。其同时将L1范数和L2范数正则化引入损失函数，如公式5所示。

(5)

式5中是预测值，是真实值，是各个系数的平方。是系数。

其综合了Lasso和岭回归的优势，适应性更强。通过调整α参数，可以在Lasso和Ridge之间进行平滑过渡，根据数据特性和需求选择合适的正则化方式。当特征之间存在高度相关性时，Lasso可能仅选择其中一个相关特征，而忽略其他有价值的信息。Elastic Net通过同时施加L1和L2惩罚，能够保留一组相关特征，避免了这种“全部或全无”的选择问题。

2.5 R2os和MSFE-adjusted 统计量

R2os统计量和MSFE-adjusted统计量都是从统计角度检验模型是否具有样本外预测能力的统计量，其中R2os的计算如公式6所示：

(6)

式6中是收益率的真实值，是收益率的预测值，是历史收益率的平均值。

从公式6中可以看出，R2os统计量使用收益率预测与未来真实值的误差和历史收益率样本的平均值和未来真实值的误差进行比较，如果该统计量的值大于0，则表示模型预测值的误差小于采用历史样本平均值的误差，否则则表明模型预测误差不如直接采用历史样本平均。

MSFE-adjusted统计量可以由如下计算过程得出：首先计算，其计算如公式7所示：

(7)

接着将和样本编号i进行线性回归，对回归得到的常数项进行t检验，所得t检验统计量就是MSFE-adjusted 统计量。

3 结果与讨论

3.1 美国市场收益率可预测性研究

3.1.1因子选择

选取Time-series and cross-sectional stock return forecasting: New machine learning methods论文[]中的12个因子来预测标普500超额收益率。表1是因子介绍：

**表1 因子表**

|  |  |
| --- | --- |
| 因子名称 | 计算 |
| 对数股息价格比DP | 标普500市场中的12个月移动股息和的对数除以标普500指数对数值 |
| 对数盈利价格比EP | 标普500市场中的12个月移动盈利的对数除以标普500指数对数值 |
| 波动率VOL | 第t个月的年化波动率 |
| 国库券收益率Bill | 三个月国库券收益率减去三个月国库券收益率的12个月移动平均值 |
| 长期国债收益BOND | 十年期国债收益率减去十年期国债收益率的12个月移动平均值 |
| 期限价差Term | 10年期国库券和3个月期国库券的收益率之差 |
| 信用利差Credit | AAA级公司债券和十年期国债的收益率之差 |
| 通货膨胀PPIG | 生产者价格指数(PPI)通货膨胀率 |
| 工业生产增长IPG  MA(1,12)  MA(3,12)  动量技术信号MOM(6) | 工业生产的增长率  标准普尔500价格指数大于或等于(小于)标准普尔500价格指数的12个月移动平均值，它的值为1(零)  标准普尔500价格指数的三个月移动平均值大于或等于(小于)标准普尔500价格指数的12个月移动平均值，则该指标变量的值为1(零)。  标准普尔500指数大于或等于(小于)六个月前的价格指数，它的值为1(零) |

3.1.2单因子样本内解释能力检验

样本内解释能力通过线性回归方法检验，将单个因子的时序数据作为自变量X矩阵，标普500收益率作为因变量Y矩阵进行线性回归，对于因子系数的显著性水平进行检验，如果显著则认为该因子有解释能力。

表2是选取的12个因子的样本内解释能力检验结果：

**表2 因子样本内预测能力检验表**

|  |  |  |
| --- | --- | --- |
| 因子 | 系数 | T统计量P值 |
| DP | 0.0066 | 0.0560\* |
| EP | 0.0080 | 0.0385\*\* |
| VOL | 0.0256 | 0.1493 |
| Bill | -0.2417 | 0.2548 |
| BOND | -1.2917 | 0.0005\*\*\* |
| Term | 0.1198 | 0.3362 |
| Credit | 0.2784 | 0.4746 |
| PPIG | -0.0006 | 0.6649 |
| IPG | 0.0008 | 0.3677 |
| MA(1,12) | 0.0051 | 0.1505 |
| MA(3,12) | 0.0038 | 0.2712 |
| MOM(6) | 0.0043 | 0.2054 |

注：\*表示10%水平下显著，\*\*表示在5%水平下显著，\*\*\*表示在1%水平下显著。

从表2中可以看出BOND因子在1%显著性水平下显著，EP因子在5%显著性水平下显著，DP因子在10%显著性水平下显著，这说明这三个因子对于样本内的收益率有解释能力。

3.1.3单因子样本外预测能力检验

样本外预测是指使用过去的因子数据预测下一期的数据，即使用t-1期之前的数据预测第t期，检验样本外预测能力的方法有两种：统计检验以及经济意义检验。统计检验指的是使用R2os等统计量进行检验，经济检验是指通过预测得到的收益率在风险资产与无风险资产之间构造投资组合，再通过平均值构造组合，计算两个组合的效用并且相减得到检验量，最后设定阈值检验是否有经济意义，在本研究中，阈值设定为10^-6，小于这个阈值则没有经济意义。统计量的含义已在2.6以及2.7中说明，表3是各因子样本外预测能力检验表。

**表3 因子样本内预测能力检验表**

|  |  |  |  |
| --- | --- | --- | --- |
| 因子名称 | R2os | MFSEpvalue | 经济意义 |
| DP | -0.005143 | 0.039526\*\* | 无 |
| EP | -0.015106 | 0.640065 | 无 |
| VOL | 0.004113 | 0.036293\*\* | 无 |
| Bill | -0.000322 | 0.005307\* | 有 |
| BOND | 0.011894 | 0.001961\*\*\* | 有 |
| Term | -0.001091 | 0.080391\* | 有 |
| Credit | -0.000930 | 0.462316 | 有 |
| PPIG | -0.001019 | 0.388908 | 无 |
| IPG | 0.000464 | 0.341275 | 有 |
| MA(1,12) | -0.001704 | 0.828172 | 有 |
| MA(3,12) | -0.001900 | 0.890707 | 有 |
| MOM(6) | -0.001505 | 0.623818 | 有 |

注：\*表示10%水平下显著，\*\*表示在5%水平下显著，\*\*\*表示在1%水平下显著。

从表3中可以看出只有BOND因子在统计检验和经济检验两个检验中都展现出样本外预测能力，DP，EP，PPIG三个因子在两个检验中都没有样本外预测能力，剩下的因子只在一个检验中具有样本外预测能力。和样本内检验结合起来看，具有样本内解释能力的因子在样本外不一定具有预测能力。从表3中本身可以看出统计检验结果和经济检验结果是不相关的，很多因子会出现只有一个检验有效的情况。

3.1.4多因子样本外预测能力检验

3.1.1到3.1.3都是对单个因子进行样本内和样本外检验，且都是采用的线性回归方法，在本小节研究中对这12个因子联合预测能力进行检验，同时又采用Lasso，Ridge和ElasticNet方法进行检验。研究结果如表4所示：

**表4 多因子样本外预测能力检验表**

|  |  |  |  |
| --- | --- | --- | --- |
| 模型名称 | R2os | MFSEpvalue | 经济意义 |
| Linear Regression | -0.041573 | 0.004093\*\*\* | 有 |
| Lasso | -0.014242 | 0.884220 | 无 |
| Ridge | -0.022879 | 0.183879 | 无 |
| ElasticNet | 0.010729 | 0.025683\*\* | 有 |

注：\*表示10%水平下显著，\*\*表示在5%水平下显著，\*\*\*表示在1%水平下显著。

从表4中可以看出运用弹性网络(ElasticNet)联合12个因子对收益率进行预测时在统计检验和经济检验上都可以取得较好的效果，使用线性模型虽然在无法通过统计检验，但是预测出的收益率仍然具有经济意义。Lasso和Ridge在统计检验和经济检验上表现都不理想。

3.2 中国市场收益率可预测性研究

本节对于中国股票市场的收益率可预测性进行研究，研究采用600694大商股份公式的超额收益率为预测目标，构造了月超额收益、月市盈率、月每股收益、净资产收益、每股营业收入、月换手率、月beta系数、月波动率、月流动性，月股价高点，月已实现偏度这11个因子用于预测。其中月波动率计算方法为月内日收益率的平方和，月流动性的计算方法为月收益率除以月成交额的对数的绝对值，月股价高点的计算方法为当月股价最高值与前三个月股价的最大值的比值，月以实现偏度的计算方法如公式8和9所示。

(8)

(9)

公式8利用一个月每日收益率计算日收益率的平方和，随后使用公式9计算月以实现偏度。

4 参考文献

1. Shaveta Dargan et al. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning[J]. Archives of Computational Methods in Engineering, 2019, 27(4) : 1-22.
2. J. B. Heaton and N. G. Polson and J. H. Witte. Deep learning for finance: deep portfolios[J]. Applied Stochastic Models in Business and Industry, 2017, 33(1) : 3-12.
3. Fama, E. F., and Kenneth, R. (1993). French, 1993, Common risk factors in the returns on stocks and bonds. Journal of Financial Economics, 33(1), 3-56.
4. 陈坚. 中国股票市场尾部风险与收益率预测——基于Copula与极值理论的VaR对比研究[J]. 厦门大学学报(哲学社会科学版), 2014(04):45-54.
5. 陈坚, 张轶凡. 中国股票市场的已实现偏度与收益率预测[J]. 金融研究, 2018(09):107-125.
6. 蒋志强, 田婧雯, 周炜星. 中国股票市场收益率的可预测性研究[J]. 管理科学学报, 2019, 22(04):92-109.
7. Rapach D E, Zhou G. Time‐series and Cross‐sectional Stock Return Forecasting: New Machine Learning Methods[M]. John Wiley & Sons, Ltd, 2020.
8. R.David Mclean, Jeffrey Pontiff. Does Academic Research Destroy Stock Return Predictability?[J]. The Journal of Finance, 2016, 71(1) : 5-31.

5 附录

给出本次实验的所有代码。