上周报告：检验单因子有效性

本报告旨在如何快速有效的区分因子的表现能力。Fama 相继提出了三因子和五因子模型，因子对于股票横截面的表现具有较强的预测能力。因此，检验单因子的有效性对于因子是否能加入因子库具有重要作用。报告中主要对一些证券报告中的单因子检验有效性的方法进行了汇总，并对部分结果进行了复现。

一、因子和未来收益相关性计算---秩相关系数

检验单因子的有效性的直观理解是同一时刻的个股的因子的表现和未来一段时间内的收益率的**相关性**如何。因此，首先需要计算二者的相关系数。

常见的皮尔逊相关系数

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

Pearson 线性相关系数只是许多可能中的一种情况，为了使用 Pearson 线性相关系数必须假设 **1 数据是成对地从正态分布中取得的**，并且 **2 数据至少在逻辑范畴内必须是等间距**的数据。如果这两条件不符合，一种可能就是采用 Spearman 秩相关系数来代替 Pearson 线性相关系数。而在金融数据中，经过整理过的数据可以满足 2 条件，但是在现实情况下，因子数据的正态却不能保证，因此，在此**使用秩相关系数检验同一时刻的因子和未来一段时间的收益的相关性**。

二、回归—稳健回归和 fama_macbeth 回归

在检验单因子有效性上，在知道二者相关系数的同时，还可以用回归的方法进行计算，检验回归系数的显著性。一般来说，在进行回归过程中，可以**全行业**进行回归，也可以**按月回归。** 市场是多变的，因此因子不可能是长期持续对未来股票收益产生正或者负的效应，简而言之，某个因子即便对未来的收益产生正的作用，也会由于市场轮动，交替产生负的影响。因此，用按月回归更为可取。以下为几种常见的回归方式。

稳健回归

最小二乘回归是最常见的回归方式，在此并不过多赘述。重点需要提出的是，OLS 对异常值具有很强的敏感性。在金融数据中，异常值的出现是非常频繁的，因此，在用因子数据值和未来收益做回归时，不应采用常见的 OLS 方法，而应该采用**稳健回归**的方式。在此，只简单介绍稳健回归的一种 **R 估计**。

R 估计是将原始数据转换成相应的秩后，利用最小二乘方法建立 Y 的秩关于 X1 的秩、X2 的秩……等秩间的回归方程，先求出 y 的秩的预测值，再求出 Y 的预测值得方法。这是一种非参数估计的方法。因为秩受异常值的影响比较小，所以是一种稳健回归的方法。

R 估计的具体步骤如下：

设给出的原始数据为：$(X_{i1},X_{i1},\cdots X_{it},Y),\ i=1,2,3\cdots,n$，记：$X_{ij}$ 在 $X_{i1},X_{i1},\cdots X_{it}$

中的秩为 $R(X_{ij}), i=1,2,3\cdots,n,\ j=1,2,3\cdots,t$，记 y_i 在 $y_1,y_2,\cdots,y_n$ 中的秩 为 $R(Y_i)$，则对应的秩的数据记为：$(R(X_{i1}),R(X_{i2}),\cdots,R(X_{it}),R(y))i=1,2,3\cdots,n$

利用最小二乘估计建立回归方程：

$$R(\hat{y})=\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}+\cdots+\hat{\beta}_t X_{it}$$

然后在给定点做预测时，可以先得到预测的秩，然后到推出相应的预测值。

稳健回归作为一种简单且有效的方式，**在证券报告中使用频繁**。

Fama_Macbeth 回归

在单因子检验的有效性方面，证券报告中更喜欢用分组进行稳健回归或者求相关性的方法，但是通过一些文献检索发现，在一些**学术论文**中，更喜欢用 **fama_macbeth** 回归的方法进行检验。

Fama _macbeth 的回归方法是 fama 和 macBeth 在 1973 年《Risk, Return, and Equilibrium: Empirical Tests》检验 CAPM 模型时提出。在发展过程中，fama_macbeth 回归仍旧是在金融实证方面具有很高的认可度。

其主要流程如下（详细过程见附件）：

1) **对于每个股票进行时间序列回归，得到每个股票的 β 值（系数估计值）**
2) **然后在横截面上来进行操作，从第一步中得到的系数估计值作为解释变量，以每个月的超额收益率作为被解释变量，进行横截面回归，得到每个时点的系数估计值的系数值（risk premium）**

在一些学术论文中，常见的方法是在回归解释变量中加入最常见的五因子（市场，市值（SML），账面市值比(HML)，盈利能力(RMW)，投资模式(CMA)），然后加入新的因子作为解释变量，观察该因子是是否对该模型的 R 方具有极大的改善。**该方法本质也是在探究在五因子未被吸收的情况下是否有效。**

而在一些**证券报告**中，最常见到 fama macbeth 的身影是在**多因子选股**的场合，并非在检验单因子有效性方面。而且对于 fama macbeth 的使用很多证券机构而是用不同于以上具体的方法。其具体方法如下所示：

1）根据横截面数据估计单个因子的回归系数；

2）从时间维度计算回归系数序列 t 统计量。

本文为了检验各因子对下一个月收益的解释，即比较横截面上因子的差异，因此采用这种方法能够很好的发现各种量化指标的统计显著性。

在每一期（这里指每个月底），我们用所有股票的收益率对检验的因子进行回归：

$$y_{t+1}=a_t+b_t x_t+e_t$$

其中，y_{t+1} 是 t+1 期的股票收益率，x_t 表示第 t 期末尾的因子值，在得到每期的回归系数后，继而做 fama macbeth 检验。

$$t(b_t)=\frac{\mu(b_t)}{\sigma(b_t)}\sqrt{T}$$

其中 T 是指时间长度，我们这里是按照月度调仓，所以 T 就指整个样本的长度。

注：python 中提供了直接用 fama-macbeth 的函数。（待检验第几种）
http://pydoc.net/Python/pandas/0.16.0/pandas.stats.fama_macbeth/

三、单因子有效性检验---分组比较

1）测试因子概述

在一些证券报告中，处理单因子有效性的方法更常见普通的方法是按照因子值对股票池中的股票进行分组，然后在月底换仓，换仓日剔除停牌和涨停等数据，计算累计收益率，并相应计算最大回撤，胜率，夏普比，信息系数等。

一般来说常见的基本面因子主要分为 10 类：

规模因子：总市值、流通市值

估值因子：PE、PB、PCF、PS、EV/FCF、EV/EBITDA、PEG、股息率、送股比

成长因子：总资产增长率、固定资产增长率、净利润增长率、主营利润增长率、营业收入增长率、营业利润增长率、营业外收入增长率

盈利因子：GPM、NPM、OPM、ROA、ROE

经营因子：存货周转率、应收账款周转率、固定资产周转率、股东权益周转率、总资产周转率

偿债因子：流动比率、速动比率、超速动比率

资本机构因子：资产负债率、股东权益比、现金比率

10 大流通股因子：10 大流通股股东占流通股本比例、10 大流通股股东占总股本比例、10 大流通股东变动

机构持股因子：机构持股占总股本比例、机构持股变动

流动性因子：股东数、股东数变动、换手率

在进行分组时，一般分为 3 种方式：

1、等权配置：每组股票股票数相同，组内股票权重等权配置。这种配置方法存在一个问题就是放大了小股票的效果，因为可能有些组合小股票很多，但是实际操作的时候组合没法放进去那么多资金；

2、按流通市值加权：假定所有样本股的流通市值之和为 $\sum_{i=1}^{n} Mvalue_i$，将按指标排序后的所有 A 股分为十组，保证每组的流通市值之和为 $\sum_{i=1}^{n} Mvalue_i / 10$，组内股票权重按流通市值加权配置。由于是按照流通市值来切分，所以有可能一只股票同时属于两个组里面，我们就以其流通市值进行加权计算收益率。这种配置方法是对等权配置的一种改进，不过这种方法我们默认流通市值大的股票能容纳的资金是跟其匹配的；

3、按过去 60 日日均成交额加权：假定所有样本股的最近 60 日日均成交额之和为 $\sum_{i=1}^{n} Amount_i$，将按指标排序后的所有 A 股分为十组，保证每组的 60 日日均成交额之和为 $\sum_{i=1}^{n} Amount_i / 10$，组内股票权重按 60 日日均成交额加权配置。与第二种切分方法类似，

这种分组方法也会导致 1 只股票属于两个组，处理方法与第二种分组方法一样。最后这种配置方法按照流动性来进行配置，跟实际的操作更加吻合，更有意义。

2）、数据处理

在进行对金融数据进行单因子检验前，对数据的预处理是必不可少的。

1） 处理掉 ST 股票和 PT 股票

ST 和 PT 股票：“T”类股票包括 ST 股和 PT 股。 1998 年 4 月 22 日,沪深证券交易所宣布将对财务状况和其他财务状况异常的上市公司的股票交易进行特别处理（英文为 special treatment, 缩写为“ST”)。其中异常主要指两种情况：一是上市公司经审计两个会计年度的净利润均为负值,二是上市公司最近一个会计年度经审计的每股净资产低于股票面值。在上市公司的股票交易被实行特别处理期间,其股票交易应遵循下列规则:（1）股票报价日涨跌幅限制为 5%;（2）股票名称改为原股票名前加 “ST”,例如“ST 钢管”;（3）上市公司的中期报告必须经过审计。

PT 股是基于为暂停上市流通的股票提供流通渠道的特别转让服务所产生的股票品种（PT 是英文 ParticularTransfer〈特别转让〉的缩写）, 这是根据《公司法》及《证券法》的有关规定,上市公司出现连续三年亏损等情况,其股票将暂停上市。沪深证券交易所从 1999 年 7 月 9 日起,对这类暂停上市的股票实施“特别转让服务”。PT 股的交易价格及竞价方式与正常交易股票有所不同:（1）交易时间不同。PT 股只在每周五的开市时间内进行,一周只有一个交易日可以进行买卖。（2）涨跌幅限制不同。据最新规定,PT 股只有 5%的涨幅限制,没有跌幅限制,风险相应增大。（3）撮合方式不同。正常股票交易是在每交易日 9：15-9：25 之间进行集合竞价,集合竞价未成交的申报则进入 9：30 以后连续竞价排队成交。而 PT 股是交易所在周五 15：00 收市后一次性对当天所有有效申报委托以集合竞价方式进行撮合,产生唯一的成交价格,所有符合条件的委托申报均按此价格成交。（4）PT 股作为一种特别转让服务,其所交易的股票并不是真正意义上的上市交易股票,因此股票不计入指数计算,转让信息只能在当天收盘行情中看到。

s 是代表没有股权分置改革的股票。

ST 是亏损或有公司有重大问题的股票。

SST 是没有股权分置改革的亏损或问题股票。

G 是以前股改完毕的股票,现在已经取消了 G 了,因为大部分已经完成股改了。

ST 这是连续二年亏损以后就会带,表示今年如果不扭亏公司就会面临退出市场交易。

N 股票是指当股票名称前出现了 N 字，表示这只股是当日新上市的股票。

在用国泰安的数据中，天数据有交易状态，可以剔除交易状态为 ST 和 PT 的股票。

2） 涨停数据的处理

在中国 A 股市场,均设有涨幅和跌幅的限制,他们都是 10%的限制,即所谓的涨停和跌停,10%的涨幅是针对上一交易日股价而言的,即今天这支股票股价涨到 10%股价就不会再涨了。涨停一旦达到 10%，就很难买到，因此在换仓日剔除掉涨停的数据是很有必要的。

在处理国泰安下载的月收益率数据中，并没有换仓日的收益率情况。因此需要将日收益率数据下载下来（非常大，很耗时间），然后挑出月末的收益率大于 0.099 的数据。

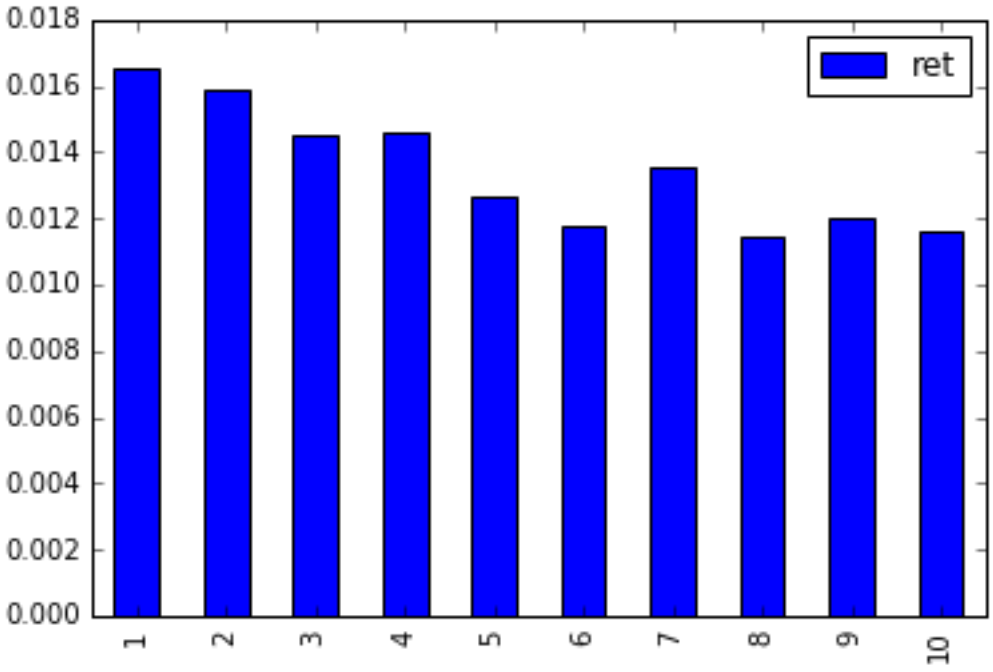
3） 去新股

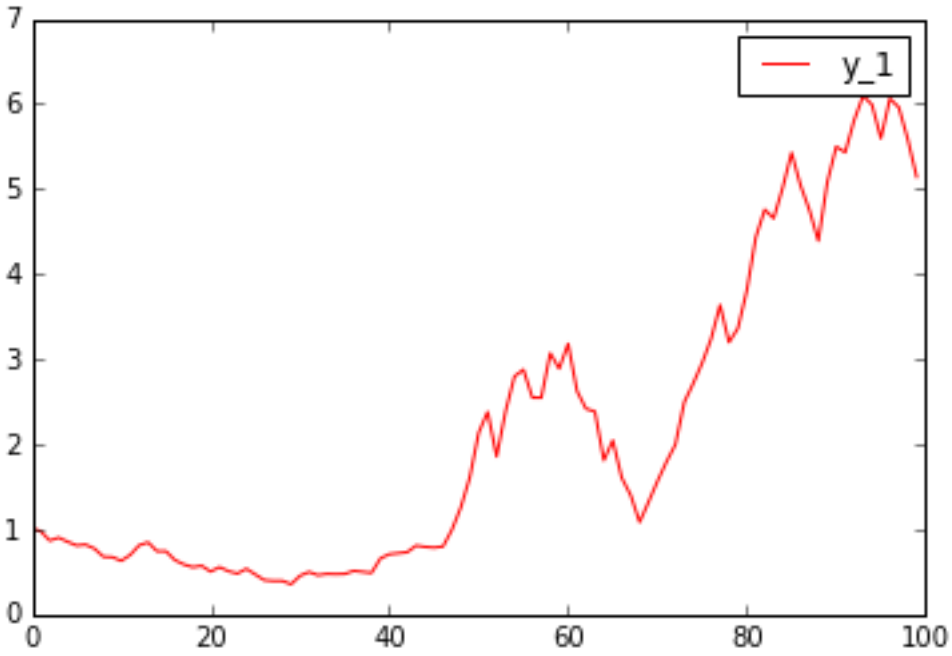
新股申购，中签率很低，只有 0.3%左右，所以很难申购到新股，而且，新股刚上，会持续涨停，因此需要对新股数据处理，在该报告中，换仓的股票池中设置为上市六个月的

股票。

四、实验

测试数据为国泰安全 A 股数据，测试时间从 2003 年 1 月 1 日到 2015 年 5 月 31 日，共 100 个月。复现因子为流通市值。





对比原始证券报告图表：

图 3.1-1：等权重月平均收益率（总市值）

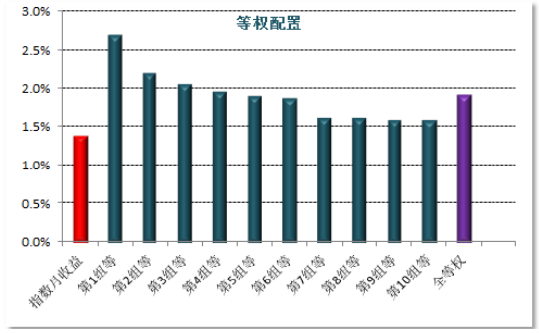
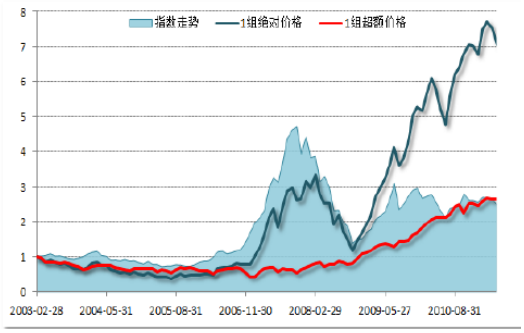


图 3.1-2：等权重第一组收益（总市值）



4、花絮

遇到坑的问题汇总：

1、从国泰安中下的数据中，首先按照股票，然后按照时间进行排列，但是求下个月收益率的时候不能 shift(-1),因为有时候一只股票这个月有，但是下个月就不一定有，所以会出错

2、打新效应，就是刚上市的股票一定要剔除了，因为刚上市的股票一定会不停地涨，这样在算的时候就拉大了整体收益率

3、国泰安中涨停数据获取太浪费了，如果建立数据库自己调取则会方便很多很多。
!!!报告中给的信息不够，为了复现已经开始排列组合条件了。。。。。。。