

因子研究方法之线性回归. 第28讲

 渔阳 2018-12-06 18:52

字数 4,602 阅读需 12分钟

“
欢迎来到量化小学”

▲ 加入[“量化小学”校友圈儿](#)提问交流

下一篇:

因子研究之Dataview
的使用. 第9讲

0



收藏



内容阅读

大家好，欢迎来到量化小学。上次我们讲了分组研究因子的方法，这一次我们就讲一讲一个量化研究最基本的方法，也就是线性回归的方法。

主要是两部分内容。首先我们要理解一下**线性回归的数学意义和它的数学思想**是什么。然后我们也考虑一些关于线性回归的**进阶问题**。



线性回归的数学思想

来自特辑



量化小学

解放你的投资动手能力

最近更新

【学业总结】量化学习的脉络梳理，以及
继续学习提高的路径

2019-04-12更新

进阶研究：集成学习和深度学习. 第31讲

2019-03-28更新



在量化小学里，我们特别重视研究方法和数学公式背后的思想。线性回归这么一种简单的研究方法，它背后的思想是什么呢？

首先我们要找的是被预测变量Y和预测变量X的某种关系。比如，我们可以用**市盈率**来对股票未来的超额收益做一个预测。这两件事应该是**正相关**的，也就是越便宜的股票未来的超额收益越高。

如果我们从一个数学的角度去理解这个问题，可以讲Y（被预测的变量）是由两部分组成的。**第一部分是跟X有关的**，就是跟这个因子相关的，这是一个**可以预测**的部分。后面这一项是**噪音**，是**不可预测**或者说由于掌握的信息不够全面，还不能预测的部分。

在这儿，f和g两个都是一般意义的数学函数，最基础的模型就是线性的模型。我们如果学过高等数学的话，都知道任何的函数，只要不是太糟糕的比较平滑的函数，它做了泰勒展开之后，一阶的规律都是线性的。

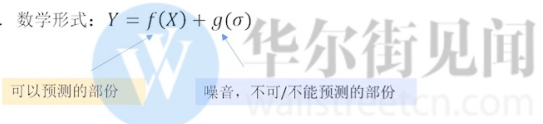
所以我们可以认为线性模型就是这种最一般化的，X和Y之间的数学模型的最简单的表现形式。线性模型也是其他更复杂模型的一个基础，后面我们会看到各种各样非线性的模型，其实里面也有很多线性的成分。

所以线性模型两个重要的组成部分，一个是 β ，是衡量X和Y之间的关系，另外后面这项 ϵ 就是噪音，是模型解释不了的部分。



线性回归 – 数学思想

1. 被预测变量（Y）和预测变量（X）存在某种关系
 - 例如：市盈率（PE）对股票未来的超额收益率正相关

2. 数学形式： $Y = f(X) + g(\sigma)$ 

3. 最基础的模型：线性模型（函数泰勒展开后，一阶规律都是线性的）
$$Y = \beta X + \epsilon$$

W | PREMIUM

02

金融数据的相关性及其特点

下面我们来看一看线性回归的数学公式是什么样的。首先我们来看两张图，右上这张图是X和Y的相关性为零的情况。右下这张图是X和Y有一点点正的相关性，相关性为0.1的情况。



线性回归 – 数学公式

线性回归：和X，Y的相关性密切相关

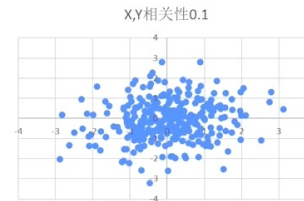
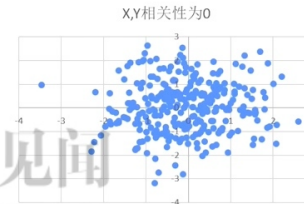
- a) 相关性为0，散点图上是“四个象限摊大饼”
- b) 相关性为正，1，3象限较多
- c) 相关性为负，2，4象限较多

金融数据的特点：“噪音多，信息少”

- a) 周期越短，噪音占比越大
- b) 日、周、月级别，因子和被预测值的相关性超过0.1较为少见
- c) 超过0.2，基本不可能
- d) 散点图的“规律”往往不明显，但有助于观察极端值

$$\text{correlation} = IC = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

W | PREMIUM



其实线性模型，我们马上就会看到，它这个公式是**和XY的相关性密切联系的**。当相关性为零的时候，我们看到的就是散点图上四个象限一个摊大饼的形状，完全没有任何的关系。如果是X和Y正相关，我们会看到处于第一象限和第三象限的点稍微多一点，也就是说，XY你可以理解为**同涨同跌的时候比较多**，处于二四象限的点相对来说就比较少一点。

也许有的听众会讲，你这个图看上去也不是那么明显。确实如此，因为我们前面也讲过金融数据有一个特点，就是噪音多，信息含量比较少。我们记住有效市场假说，其实它在讲全是噪音，完全没有信息。当然我们一再讲，有效市场假说接近正确，也不是说完全正确，所以我们还是可以找到一些信息。但总而言之，信息的含量不会太高。

也就意味着有几点，首先是**周期越短，噪音占比越大**，这个是可以理解的。比如说明天的股票是涨还是跌？其实也就是一半一半的可能性。但如果我以这个一年或者十年为一个周期，很显然，股票上涨的概率是比较大的。所以时间周期拉得越长，信息占比越高，噪音的占比是越低的。



但是从一个实际的角度来说，要想赚钱，根据我们前面讲过的主动投资的基本公式，你还要有**足够多的押宝的次数**。如果你把时间周期拉的太长，押宝的次数就变少了，因此实际在做阿尔法研究的时候，一般是有日级别、周级别或者是月级别。

在这样的时间周期下，因子和被预测值的**相关性超过0.1其实是比较少见的**，超过0.2就基本不可能。这是因为我们刚刚讲过的金融数据的特点，噪音多，信息少，而且距离有效市场实际上是比较接近的，因此它的相关性不可能太高。

所以我们看到，如果你把金融数据散点图画出来，它的规律往往不是特别明显，但是在做研究的时候，通常我们还是要经常来画散点图。因为有助于观察极端值。金融数据这里面极端值挺多的，它会对你的量化研究造成比较大的影响。

线性回归的相关性公式——最小二乘法

那么最后，我们就引出了相关性的数学公式。前面其实有提到过，就是XY的两个斜方差，然后再除以它们两个的标准差。在因子研究当中，我们通常也把它称之为IC，就是**Information Coefficient**。



线性回归 – 数学公式

线性回归：和X，Y的相关性密切相关

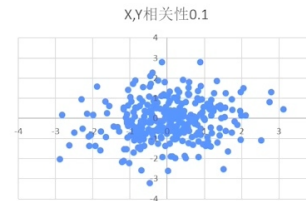
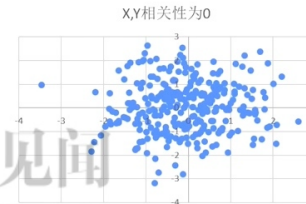
- a) 相关性为0，散点图上是“四个象限摊大饼”
- b) 相关性为正，1，3象限较多
- c) 相关性为负，2，4象限较多

金融数据的特点：“噪音多，信息少”

- a) 周期越短，噪音占比越大
- b) 日、周、月级别，因子和被预测值的相关性超过0.1较为少见
- c) 超过0.2，基本不可能
- d) 散点图的“规律”往往不明显，但有助于观察极端值

$$\text{correlation} = IC = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

W | PREMIUM



那么相关性在线性回归当中起到什么样的作用？我们来看下一张幻灯片。

我们要研究的目标是一个线性的方程， $Y = \beta X + \varepsilon$ 。最重要的是计算 β 相关系数，数学上一般要用**最小二乘法**。这个数学公式看着好像挺复杂的，其实它的含义如果搞清楚了，你会觉得它蛮直观的。

同时这也是一道特别基础的面试题，我在工作当中确实也发现，挺多同学对这么基本的公式其实讲不是特别清楚。但是如果你去面试一个量化的工作，如果这道题都答不上来，基本上不太可能有什么希望，所以请大家听好数学含义究竟是什么样子的。

其实我说直观，这是因为你把数学公式拆开，它的斜方差，其实你把每个X和它的X的平均值相减，再乘以，每个Y和Y的平均值相减。直观理解，在**相关性高的时候，在一三象限的时候比较多**。



也就是说，分子两项同向取值的时候比较多，所以β值自然就高。反之，如果是四个象限摊大饼，你可以看到分子这些，就有时候取正，有时候取负，全都加起来，它就接近于零，也就意味着**β接近于零，相关性也接近于零**。

那么这里面的β系数和相关性又是什么关系？我们把它做一个数学上的简单拆解。我们可以看到算β的公式，和算相关性公式的没差多少。IC这项就是我们上一张幻灯片讲的X和Y的相关性，后面我们还需要增加一个**量纲上的缩放**。因为Y和X，它大小的量纲有可能差距很大。

举个例子，如果Y是超额收益，比如说日级别的超额收益，它通常就是在几个百分点范围内。比如说X是某种排名，可能X的取值范围就广了，从1到500都有可能，所以你光是看相关性还是不够的，因为相关性永远是取值在-1到1之间。你**还要把X和Y的标准差也考虑进来**，做一个量纲上的缩放，来计算β的系数。

线性回归 – 数学公式

$$Y = \beta X + \epsilon$$

最小二乘法计算回归系数：

$$\beta = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{\sum_i (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sigma_X^2}$$

华尔街见闻
wallstreet.cn

直观理解：相关性高，则分子中两项的同向取值较多

$$\beta = \frac{Cov(X, Y)}{\sigma_X^2} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \cdot \frac{\sigma_Y}{\sigma_X} = IC \cdot \frac{\sigma_Y}{\sigma_X}$$

X, Y相关性 量纲的缩放

所以最小二乘法来计算 β 的数学公式就是这样。金融含义我们也讲了，如果有志于从事量化的同学们，其实应该掌握更严格的数学的推导，任何一本教科书上都有，大家可以自己去看。

秩相关性——Rank IC

下面再来看一看普通线性回归的一些变形，比如说在金融中常用的秩相关性又是什么？首先我们来看，普通的线性回归有什么样的问题。

首先刚刚讲过，金融数据是噪音大，信息少，而且它的噪音并非完全服从正态分布，这个就会给普通线性回归带来一些问题。

另外也更重要的一点就是，金融数据当中有很多极端值，如果你翻到上页的数学公式，**出现极端值对于协方差的贡献是非常大的**，比如右面这张图，大部分的点都是在0附近，但是我红圈圈出来的有些点它距离集体比较遥远。



秩相关性 – Rank IC

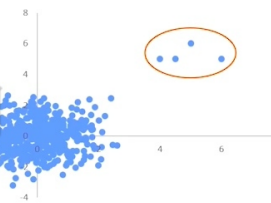
普通线性回归的几个问题：

- a) 金融数据“噪音大，信息少”，而且噪音并非严格服从正态分布
- b) 极端值对X, Y的协方差“贡献”很大，最小二乘法会“尽力拟合”这些极端值，而这是否合适？

常用解决方法：秩相关性（Spearman rank correlation）代替普通相关性（Pearson correlation）：

- a) 在选股空间内，以股票在因子值（X）的排名代替原始因子值，以股票的超额收益/收益的排名（Y）代替原始值，然后做线性回归
- b) 秩回归同时具有“降噪”和“去极值”的作用，使结果的鲁棒性（robustness）更好
- c) Rank IC – 秩回归的IC

极端值的影响



W | PREMIUM

06

从一个数学公式角度来说，这几个极端值对你整个模型的协方差的贡献是很大的。如果你用上一张幻灯片讲的最小二乘法来计算 β 的话，那么这个模型就会花很多力气去拟合偏离群体的极端值，但是这是不是合适？其实从实践角度来说未必合适。

比如我们在讲股票的收益率，它是不是和市盈率有关系，这是一个一般意义上的统计规律。有的时候某些股票会出于某种特殊的原因连续五天涨停。比如说可能遇到了一个并购的案例，那么它就会有一个远远大于普通股票的周收益率。

如果你用一般的最小二乘法去拼命的研究这个股票，为什么会一周就涨了百分之五、六十，其实就是有问题在里边了。因为连续五个涨停是有非常特殊的原因造成的。你简单的把最小二乘法套在这些极端值上，容易造成估计上的偏差。

那在金融上面我们怎么来处理这个问题？常见的解决方法就是**用秩相关性来代替普通的相关性**，在英文当中普通相关性叫Pearson correlation，那么秩相关性叫Spearman rank



correlation。其实思想非常简单。

就是在选股空间内，我们用**股票的因子值的排名来代替原始因子**，同时用股票超额收益或是收益，总而言之就是我们要研究的对象排名，来代替原始值然后再做线性回归。这样你就能够在很大程度上起到降低噪音和去极值的作用。

比如股票收益率，百分之五十，百分之一百离群体很远，但如果从排名，你也就是排第一名。比如说在500只股票当中，可能某一周收益率最高的股票达到60%，连续五个涨停，第二名可能收益率就只有20%。60和20差距是很远的，但是如果你把它变成排序，60是第一名，20是第二名，它就有了降低噪音和去极值的作用。因此做秩相关性的回归的时候，往往会使**结果的鲁棒性更好**。

另外从一个实际角度来说，选股也通常都是从一个排名的角度来思考这个问题的。所以研究报告里面也经常看到所谓rank IC这个概念，也就是**秩回归的IC**。细心的同学可能会注意到，这种秩相关性的研究和我们前面讲的分组法，是有异曲同工之妙的，可以把分组法理解为一个颗粒比较粗糙的秩相关性的研究。



秩相关性 – Rank IC

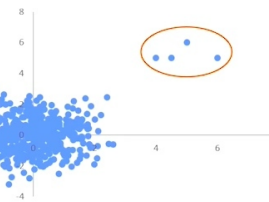
普通线性回归的几个问题：

- a) 金融数据“噪音大，信息少”，而且噪音并非严格服从正态分布
- b) 极端值对X，Y的协方差“贡献”很大，最小二乘法会“尽力拟合”这些极端值，而这是否合适？

常用解决方法：秩相关性（Spearman rank correlation）代替普通相关性（Pearson correlation）：

- a) 在选股空间内，以股票在因子值（X）的排名代替原始因子值，以股票的超额收益/收益的排名（Y）代替原始值，然后做线性回归
- b) 秩回归同时具有“降噪”和“去极值”的作用，使结果的鲁棒性（robustness）更好
- c) Rank IC – 秩回归的IC

极端值的影响



W | PREMIUM

07

因子研究的重要问题

最后我要稍微总结一下，在因子研究当中我们要注意的几个重要的问题。

首先就是一定要**避免对于噪音，对于历史做“过拟合”**。上一张PPT当中我们讲到的普通线性回归和秩回归的关系，如果不注意的话，你其实会花很多时间去拟合极端值，这个可能就是对于历史在做“过拟合”。

好的因子应该长什么样子？我们是**企图去抓住一些真正的规律**。

你会经常看到一个词叫**泛化能力**，特别是在跟机器学习相关的文章当中老提到这个词，所谓泛化能力强的模型，就是说它能够抓住事物内在的本质性的东西，而不仅仅单是对历史的一种“过拟合”。这样的因子，它对未来是具有指导意义的。



比如一直在讲的市盈率这个因子，就是挺好的一个因子，因为逻辑上说得通，泛化能力也应该比较强。在量化研究的时候，至少在一开始的阶段，要多去寻找这样的因子。同时对那些讲不清楚原因，仅仅是能够描述历史数据的因子就要特别的小心。



线性回归的进阶问题

今天最后我们也简单的谈一下线性回归的一些进阶的问题。



线性回归的进阶问题

1. 线性回归的数学假设，在金融应用中是否符合？
 - a) 残差（噪音）是否正态分布？
 - b) 异方差问题：噪音“时变”怎么办？
 - c) 回归得到的beta系数，“准确度”有多高？
 - T-stats, p-value
 - ICIR指标，IC取值为正/负的比例等。
2. 多元线性回归 – 多个因子提高模型解释度
 - a) 0, 1取值的“哑变量”用于处理行业等变量
 - b) 因子之间的共线性问题
 - c) 因子是“越多越好吗”？

W | PREMIUM



其实线性回归虽然说是最简单的方法，但其实里面还有蛮多的学问的，也可以说有挺多坑。由于时间关系不可能一一的讲，大家可以去看一些比较标准的数学教材。那么罗列一些比较重要的事情。

首先像任何一个数学公式，任何一个模型一样，我们先要想想它的假设在金融应用当中是不是完全符合。如果不符合，我们是不是可以接受？还是要采取一些什么样的方法？这个供大家思考。

首先刚才我们谈到了残差或者噪音的问题，它是不是符合正态分布？那么在金融数据当中，一般情况下都不完全符合正态分布。其实这也就是所谓的**异方差的问题**，就是噪音它随着时间来变化。处理异方差有专门的方法，但是如果异方差问题不严重的话，在做初步研究的时候，你就不去管它，这也是可以的。



第三点就是，我们回归得到的 β 系数准确度有多高？因为 β 是从数据当中总结出来的，因为数据的量通常都是有限的，所以你计算出来的 β 和真实的 β 之间，往往是有差距的。换言之，这也是一个有随机变量的一个概念。

我们怎么样来度量 β 系数的准确度有多高？常用的一些统计指标像**T-stats**，**p-value**，等等，我们接下来也会讲到。

第二点就是说，在金融上面大家也会用一些其他指标来衡量回归的 β 系数或者相关性系数是否稳定，这就包括了**ICIR**的指标。就是把IC和它自己的波动性除一下，得到类似一个下浮比率的这样一个概念。或者你就衡量一下在历史上每个时间周期，IC取值为正或者为负的比例，很显然它符号一样的时候越多，表明你这个系数关系越稳定。

多元线性回归的数学公式

今天我们主要讲了单个因子的或者一元线性回归，那么多元线性回归的数学公式也是非常类似的，就是同时用多个因子来提高模型的解释度，这里面也有几个常见的问题。

第一个是我可以用0,1取值的**哑变量**来处理行业等等这种yes or no的变量。比如说一个股票是属于银行业的，它在银行的哑变量取值就是一，否则就是零，这是多元线性回归常见的一个方法。

第二个就是因子多了之后会有一个贡献性的问题。比如我引入了一个市盈率的因子，如果再引入一个估值的因子，通常它会跟市盈率有一定关系。如果这两个因子相关性太高，它会带来一系列的问题。在做多元线性回归时，要注意处理这些。

第三点就是，因子是不是越多越好。表面看去，多弄几个因子，模型的解释度确实会提高，但是记住我们刚刚讲的非常重要的一点，我们要防止模型对历史进行“过拟合”。你搞



的因子太多，其实往往就是在努力地去feed历史的规律，而对未来的解释度是没有什么意义的。也有一些数学方法能够帮助我们避免在多元线性回归当中出现“过拟合”，接下来的几讲也会讲到。

今天还是理论的东西讲的比较多。过去这几次都是讲了一些方法，那么我觉得唯有动手，才能够真正理解我们讲过的内容。因此下面两讲我们就实际的讲一讲，如何利用Python进行因子研究的实战。

谢谢大家，咱们下次再见。

-END-

加入“量化小学”的见识圈，关注动态

感谢您订阅本特辑，扫描下方二维码或[点击圈子链接](#)，即可加入专属见识圈子提问交流





量化小学



渔生

小学而大不遗，量化师生联谊会

感谢大家订阅《量化小学》，这里是学校见识社群，你可以随时提问、随时互动，我们一起投资，一起分享！



风险提示及免责条款

市场有风险，投资需谨慎。本文不构成个人投资建议，也未考虑到个别用户特殊的投资目标、财务状况或需要。用户应考虑本文中的任何意见、观点或结论是否符合其特定状况。据此投资，责任自负。

写评论

请发表您的评论



表情



图片

发布评论

华尔街见闻

关于我们

广告投放

版权与商务合作

联系方式

意见反馈

声明

未经许可，任何人不得复制、转载、或以其他方式使用本网站的内容。

评论前请阅读网站[“跟帖评论自律管理承诺书”](#)

法律信息

版权声明

用户协议

付费内容订阅协议

隐私政策

违法和不良信息

举报电话: 021-60675200 (周一到周五9:30-11:30, 13:00-18:30)

举报邮箱: contact@wallstreetcn.com

网站举报: [点击这里](#)



华尔街见闻APP



华尔街见闻公众号



微博@华尔街见闻



中央网信办
违法和不良信息举报中心

上海市互联网
违法和不良信息举报信息

[违法和不良信息举报受理和处置管理办法](#)

[清朗·财经违规内容专项整治公告](#)



举报中心

友情链接

[腾讯财经](#) | [财经网](#) | [澎湃新闻](#) | [界面新闻](#) | [全景财经](#) | [陆家嘴金融网](#) | [富途牛牛](#) | [网易财经](#) | [凤凰网财经](#) | [虎嗅](#)

© 2010 - 2022 上海阿牛信息科技有限公司 版权所有 沪ICP备13019121号  沪公网安备 31010102002334 号 增值电信业务经营许可证沪B2-20180399

