

# 用 Quantile Regression 分析变量相关性

 石川   
量化交易 话题的优秀回答者

已关注

81 人赞同了该文章

## 1 分位数和分位数回归

分位数 (quantile) 是概率中的一个概念。对一个随机变量  $X$  和任意一个 0 到 1 之间的数  $\tau$  , 如果  $X$  的取值  $x$  满足  $\text{prob}(X \leq x) = \tau$  , 那么  $x$  就是  $X$  的  $\tau$  分位数。换句话说,  $\tau$  分位数说明: 如果我们按该随机变量的分布产生足够多的样本点, 那么在那些样本点的取值中, 有  $\tau \times 100\%$  个小于该分位数; 有  $(1 - \tau) \times 100\%$  个大于该分位数。最常见的分位数非中位数 (median) 莫属, 它是 50% 分位数 —— 在  $X$  的分布中, 有一半比中位数小, 一半比中位数大。

也许你仍觉着上面的定义抽象, 但是你对下面的儿童成长图 (child growth chart) 一定不陌生。它给出了儿童 (这个表中是男孩) 在不同年龄时身高和体重的不同分位数 (3%、10%、25%、50%、75%、90% 以及 97%) 曲线, 这有助于儿医和父母判断宝宝成长过程中发育是否正常。如果一个娃的体重落在 90% 分位线上, 说明他的体重比同龄的 90% 的小伙伴要高; 如果一个娃的

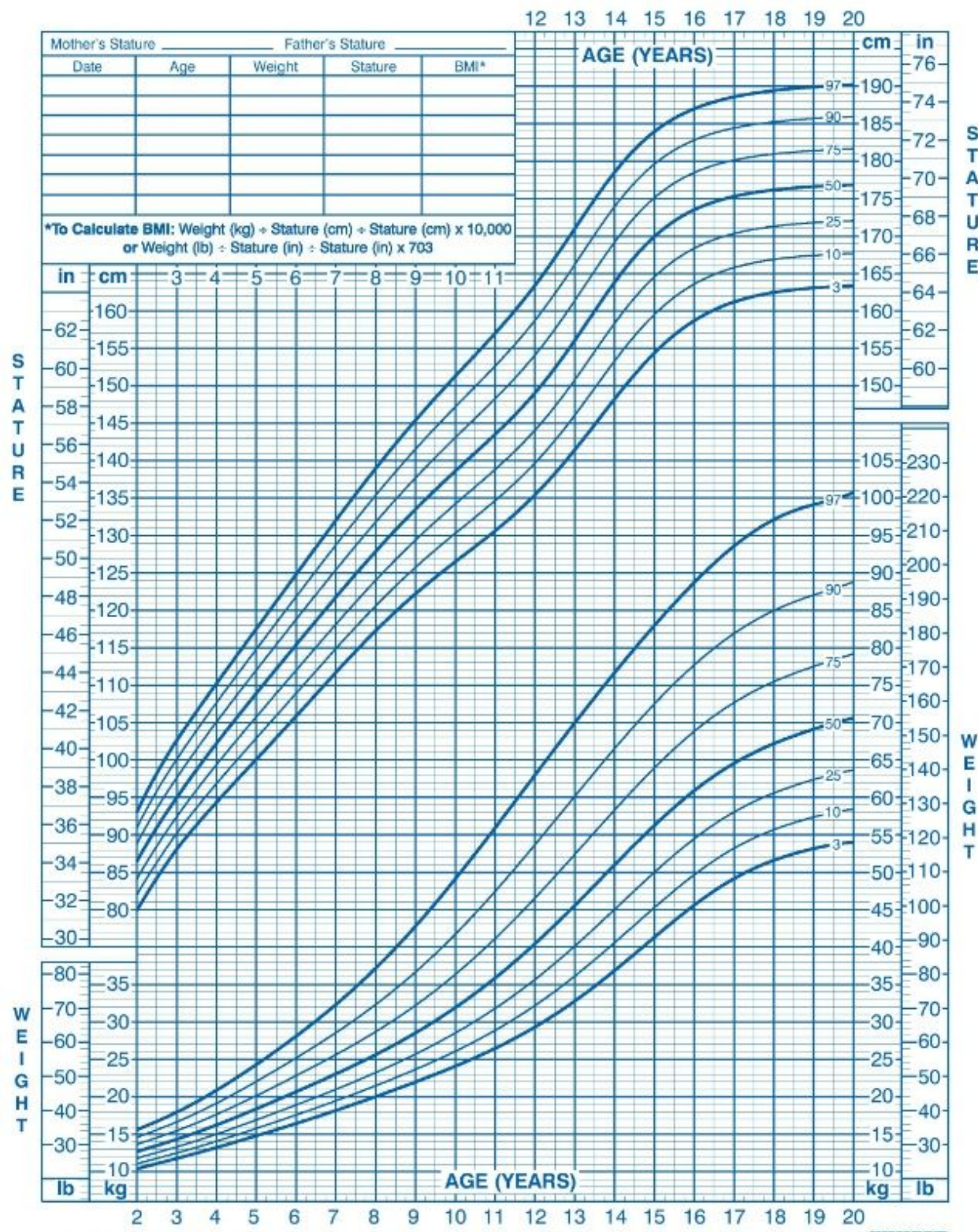


2 to 20 years: Boys

NAME \_\_\_\_\_

Stature-for-age and Weight-for-age percentiles

RECORD # \_\_\_\_\_



Published May 30, 2000 (modified 11/21/00).  
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).  
<http://www.cdc.gov/growthcharts>

SAFER • HEALTHIER • PEOPLE™

上面这个图说明两点：

1. 随着年龄的增加，低分位数和高分位数之间的间隔越来越大；

显然，这两点向我们展示了身高（或体重）与年龄在整个分布上的一些关系。试想一下，如果我们仅有年龄和平均身高（平均体重）的关系，我们是无法得到上面两点结论的。**分位数定量描述了中心趋势和统计离散度，这有助于更全面地分析变量之间的关系。**

如何得到如上图中的分位数曲线呢？答案是**分位数回归 (quantile regression)**。

分位数回归由 Koenker and Bassett, Jr. (1978) 提出，是一种回归分析。在传统回归中，我们构建回归模型由自变量求出因变量的条件期望；而在**分位数回归中，我们构建回归模型由自变量求出因变量的条件分位数。**

近年来，分位数回归在计量经济学中的应用越来越广泛。利用分位数回归，Saastamoinen (2008) 研究了芬兰市场中的羊群效应；Alagidede and Panagiotidis (2012) 讨论了通货膨胀和股票收益率之间的关系；Badshah (2012) 分析了美股中恐慌指数 (VIX) 和收益率分布之间的不对称性。

本文简要介绍分位数回归，并通过一个简单的例子说明它在量化投资中的潜在作用。

## 2 最优化视角下求解均值和中位数

让我们先把回归问题放在一边，仅仅考虑一个随机变量  $Y$  的一组样本  $\{y_1, y_2, \dots, y_n\}$ 。在本节中，我们从求解最优化问题的角度说明如何求出样本均值和中位数。这对于后面介绍分位数回归很有帮助。

我们都知道，这组样本的均值就是这  $n$  个数的平均值。从最优化的角度来说，该样本均值正是下列**最小化残差平方和（即我们常说的最小二乘法）**问题的解：

$$\min_{\mu} f = \sum_i^n (y_i - \mu)^2$$

最优的  $\mu$  应满足  $df/d\mu = 0$ 。经过简单的推导不难看出，**最小二乘法得到的解就是样本均值**：



知乎

首发于  
川流不息

$$\Rightarrow \sum_i^n (y_i - \mu) = 0$$

$$\Rightarrow \sum_i^n y_i - n\mu = 0$$

$$\Rightarrow \mu = \frac{1}{n} \sum_i^n y_i$$

与之类似的，**最小化残差绝对值之和的解就是样本的中位数**（这里的残差是样本点相对于中位数而言的），即这组样本的样本中位数  $M$  是如下最优化问题的解：

$$\begin{aligned} \min_M f &= \sum_i^n |y_i - M| \\ &= \sum_{i:y_i < M}^s (M - y_i) + \sum_{i:y_i \geq M}^{n-s} (y_i - M) \end{aligned}$$

对  $M$  求导得：

$$\frac{df}{dM} = s - (n - s)$$

可见， $df/dM$  等于 0 的必要条件是  $s = n - s$ ，其中  $s$  是小于  $M$  的样本点的个数，而  $n - s$  是大于  $M$  的样本点的个数。这意味着  $M$  的取值满足在其两侧的样本点个数相同，即  $M$  是中位数。

来看一个例子。

假设随机变量  $Y$  的一组样本是 1 到 9 这 9 个数。按照上述最优化的思路，我们想找到  $M$  使得目标方程  $f = \sum_i |y_i - M|$  最小。在 1 到 9 内遍历  $M$  并求出  $f$  对应的值有：

M	1	2	3	4	5	6	7	8	9
f	36	29	24	21	20	21	24	29	36

知乎 @石川





来看看如何将这个思路推广到分位数回归上。

### 3 分位数回归

推广上一节的最优化思路引出分位数回归十分简单，仅需要两步走。

**第一步：引入回归问题。**在上一节中，为了简化讨论，我们考虑的是随机变量  $Y$  自身。在（线性）回归问题中，我们关注的是因变量  $Y$  和某些自变量  $X$  之间的（线性）关系。（这里， $X$  可以代表一个自变量或者多个自变量组成的向量。下文中为了简化讨论，假设自变量只有一个。）

对于均值来说，将上一节中的标量  $\mu$  变成自变量  $X$  的线性方程  $\mu(X, \beta)$  —— 其中  $\beta$  是  $X$  的系数，并将最优化问题转化为（在这个问题中，求解的对象是  $X$  的系数  $\beta$ ）：

$$\min_{\beta} \sum_i^n (y_i - \mu(x_i, \beta))^2$$

求解得到  $\beta$  后，线性方程  $\mu(X, \beta)$  就是因变量  $Y$  的**条件期望方程**  $E[Y|X]$ 。我们熟悉的求解线性回归的最小二乘法正是如此找到  $Y$  和  $X$  的关系的，它得到的  $Y$  和  $X$  之间的关系正是  $E[Y|X]$ 。

对于中位数也可以做相同的推演。令上一节中的标量  $M$  变为自变量的线性方程  $\xi(X, \beta)$ 。因此该最优化问题转化为：

$$\min_{\beta} \sum_i^n |y_i - \xi(x_i, \beta)|$$

求解得到  $\beta$  后，线性方程  $\xi(X, \beta)$  就是因变量  $Y$  的**条件中位数方程**。

**第二步：将中位数推广到一般分位数。**在所有分位数中间，中位数 —— 又称 50% 分位数 —— 比较特殊是在于在求解最优化问题中，其两侧样本点的残差是等权重的。把上述最小化残差绝对值的问题推广到一般的  $\tau$  分位数时，只需把  $\tau$  分位数两侧的残差赋予不同的权重即可。

具体的，对于  $\tau$  分位数左侧样本点的残差，赋予它们  $1 - \tau$  的权重；对于  $\tau$  分位数右侧样本点的残差，赋予它们  $\tau$  的权重。最优化问题由此变为（求解的对象为  $\tau$  分位数对应的系数  $\beta$ ，记为  $\beta_{\tau}$ ）：

$$\min_{\beta_{\tau}} \sum_{i: y_i \geq \xi(x_i, \beta_{\tau})} \tau(y_i - \xi(x_i, \beta_{\tau})) + \sum_{i: y_i < \xi(x_i, \beta_{\tau})} (1 - \tau)(\xi(x_i, \beta_{\tau}) - y_i)$$



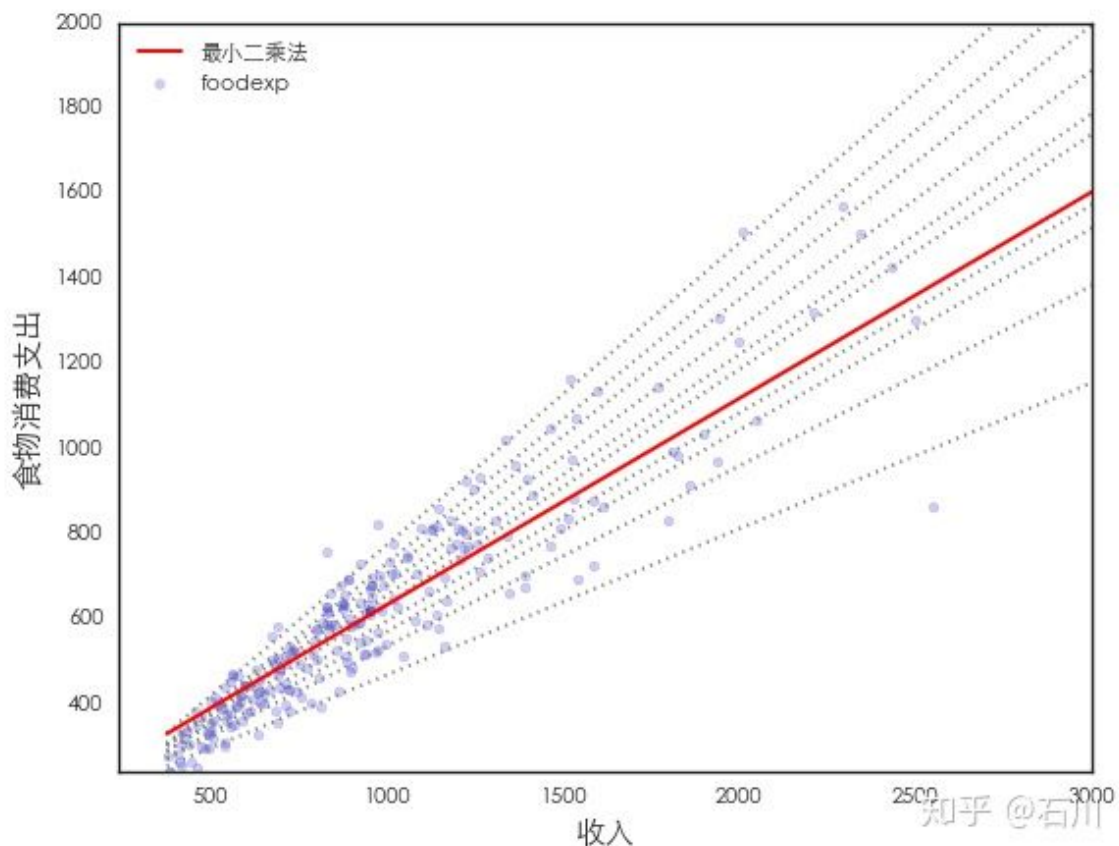
只需要对每个  $\tau$  分别求解上述最优化问题，就可以得到  $Y$  的不同条件  $\tau$  分位数方程。

值得一提的是，如果我们仅有一个自变量  $X$ ，并用它来对  $Y$  进行分位数回归，那么任何一个  $\tau$  分位数回归方程都是一条直线（有截距项、斜率为  $\beta_\tau$ ）。但是在第一节的儿童成长图中，身高（体重）的条件  $\tau$  分位数方程随年龄的变化明显不是直线。这是因为在构建成长曲线时，通常对年龄先进行了某种非线性变化以更好的反应它和儿童的成长的关系。从分位数回归的角度，我们做的依然是线性回归，只不过这时自变量已经从身高变成了身高的某个非线性函数而已。

在下文的第 4、5 节我们考虑两个例子，在这两个例子中我们都不会对自变量进行任何变换。因此这两个例子中的条件  $\tau$  分位数方程都是线性的。

## 4 收入和食物消费支出的关系

来看一个大概是介绍分位数回归的文章都要提的例子（抱歉没能免俗）..... Engel (1857) 研究了家庭收入和家庭食物消费支出之间的关系。对该数据同时进行最小二乘法回归（得到条件均值的方程）和分位数回归（得到 10 个条件  $\tau$  分位数方程， $\tau$  的取值为 5%，15%，.....，95%）如下图所示。



从这个图中可以观察到以下结论：

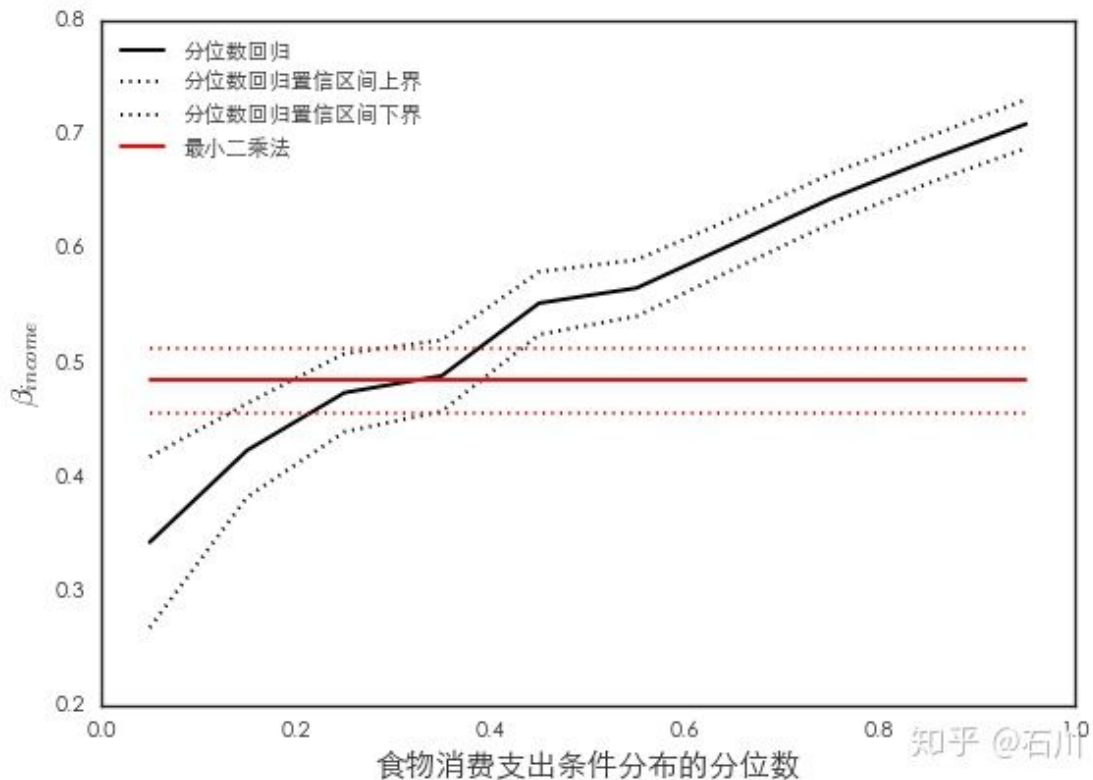
1. 食物消费支出随收入而增加；

知乎

首发于  
川流不息

于很多低收入观测点之上。

上述分位数回归的结果说明，在食物消费支出分布的不同位置（不同分位数），家庭收入对其的影响是不同的。下图展示了这一点。图中横坐标为食物消费支出的分位数，纵坐标为不同分位数回归的系数  $\beta_{income}$ ，它表示一个单位的家庭收入变化带来多大的食物消费支出。对于最小二乘法（红色）来说，它假设收入对食物消费支出的影响在整个分布上是恒定的；但是分位数回归（黑色）正好得到不同的结论。显然，分位数回归提供了收入和食物支出之间更为丰富的关系。



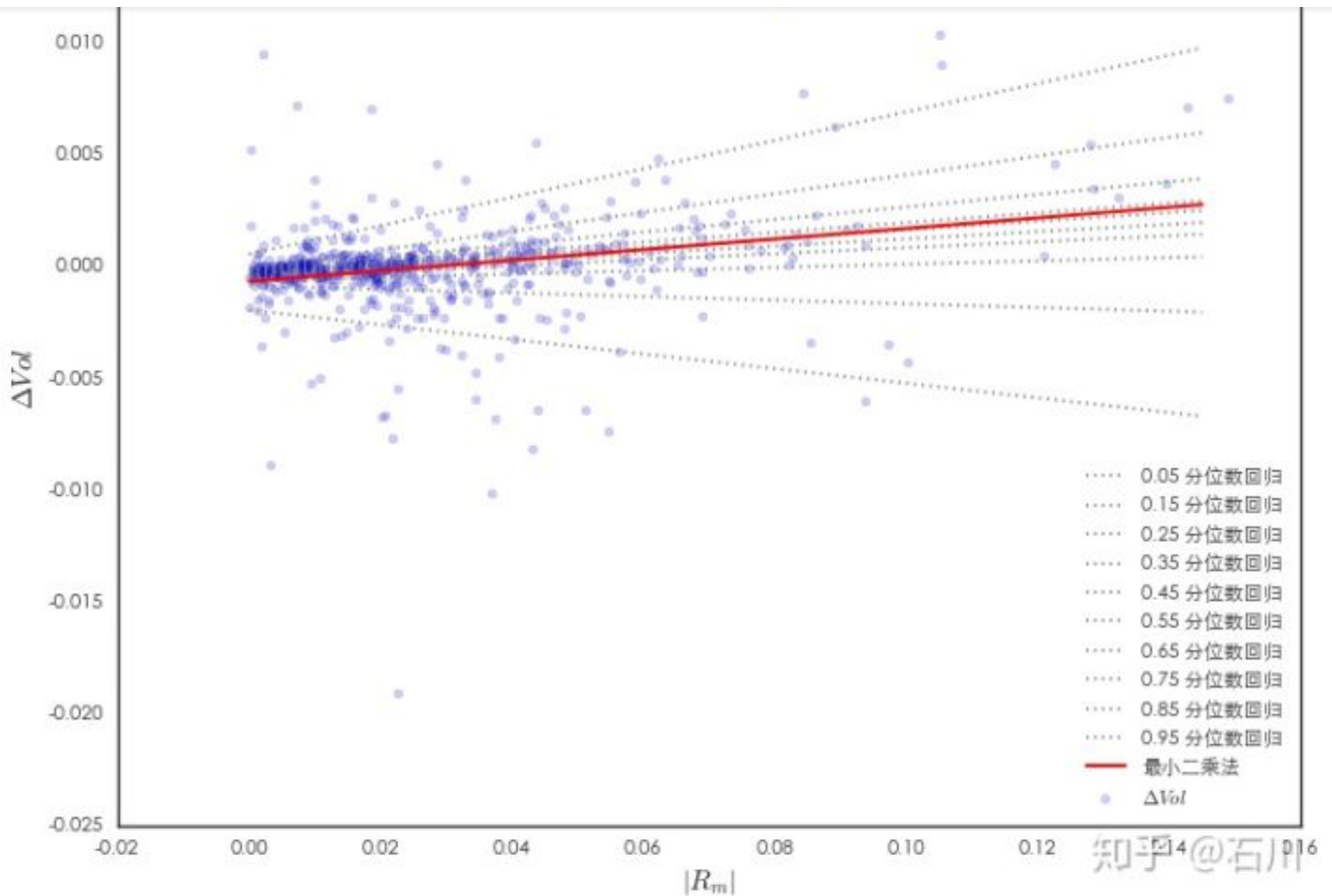
## 5 分位数回归在量化投资中的应用一例

最后通过一个简单的例子介绍分位数回归在量化投资中的应用。

具体的，我们关注风险和收益之间的关系。为此，需要给风险和收益各找一个代理指标。以上证指数（2005 年 1 月 1 日至 2017 年 7 月 31 日）为例，风险的代理指标为每周已实现波动率（日频收益率的平方和）的变化率，记为  $\Delta Vol$ ；收益的代理指标为周收益率的绝对值，记为  $|R_m|$ 。对该数据同时进行最小二乘法回归和分位数回归如下图所示。



知乎

首发于  
川流不息

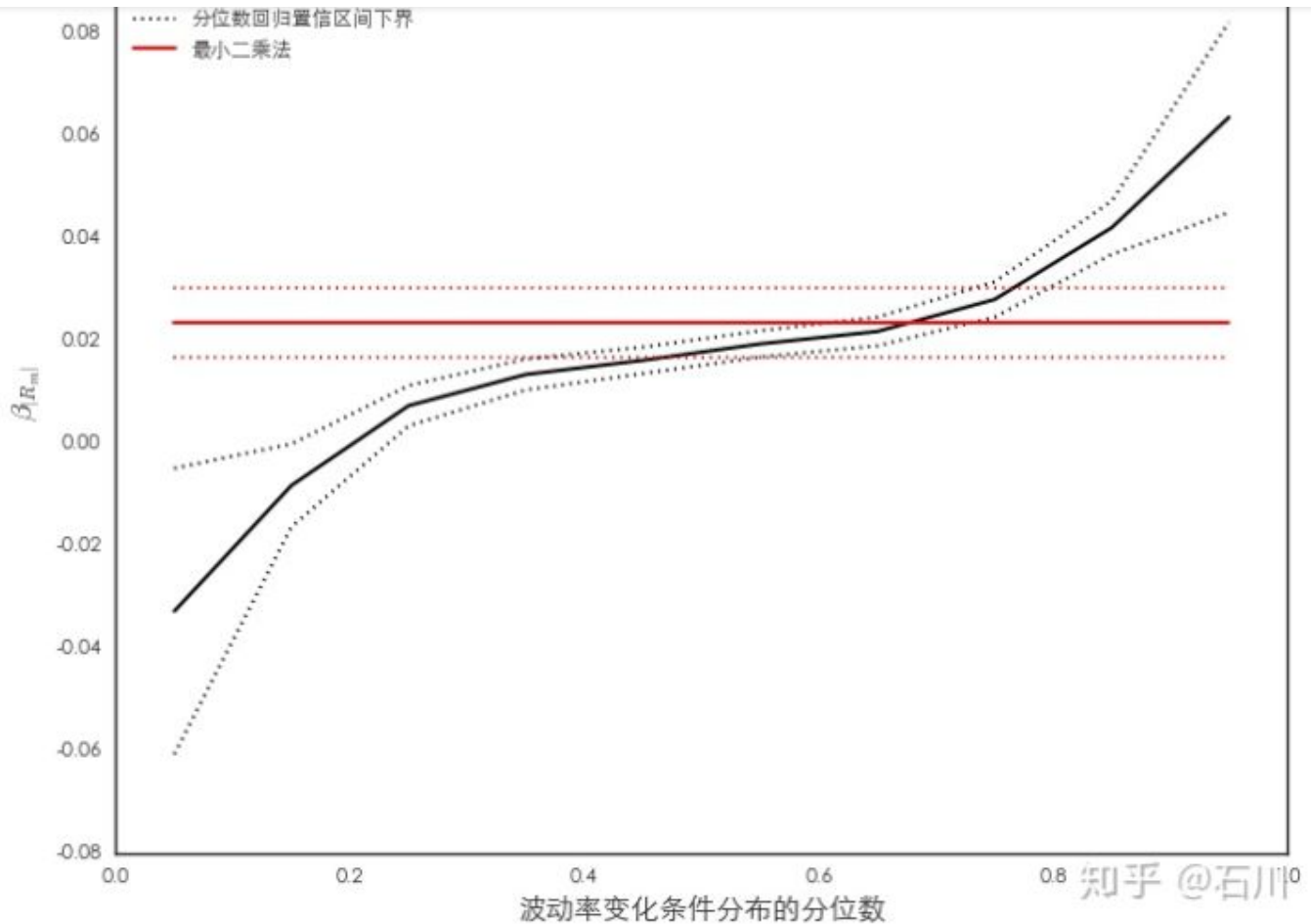
知乎@石川

可见，对于  $\Delta Vol$  的不同分位数， $|R_m|$  对其的影响不同。下图是  $\tau$  和系数  $\beta_\tau$  的关系。当  $\Delta Vol$  处于低分位数通常意味着市场一般比较平稳，因此周波动率也比较稳定、 $\Delta Vol$  较小。这时收益率的单位变化对  $\Delta Vol$  的影响为负，有助于进一步维持平稳的市场状态。当  $\Delta Vol$  处于高分位数通常意味着市场一般比较震荡，因此周波动率变化剧烈、 $\Delta Vol$  较大。这时收益率的单位变化对  $\Delta Vol$  的影响为正，即它会进一步加剧市场的波动。





知乎

首发于  
川流不息

知乎 @石川0

## 6 结语

对于金融投资中的很多变量，比如收益率，我们往往更关心它在分布尾部的特性。在这方面，分位数回归是一个有力的工具，它让我们研究收益率和不同的解释变量在全分布上的相关性。

当变量的分布明显偏离正态分布或者存在异常值（outliers）时，传统的最小二乘法回归就不那么有效了。然而分位数回归不受这些弊端的影响。此外，**分位数回归满足单调变换不变性**

**(invariant to monotonic transformations)**。对于随机变量  $Y$  和它的单调变换  $h(Y)$ ——比如  $\ln(Y)$ ， $h(Y)$  的分位数正好是  $h(Q_\tau(Y))$ ，即对  $Y$  的分位数  $Q_\tau(Y)$  直接做同样的变换；而均值并不满足类似的性质，即  $E[h(Y)] \neq h(E[Y])$ 。投资品收益率的分布以不满足正态性并存在很多异常值而闻名，因此上述优点使分位数回归在分析收益率时有着不错的潜力。

## 参考文献

- Alagidede, P. and T. Panagiotidis (2012). *Stock returns and Inflation: Evidence from Quantile Regressions*. Discussion Paper Series, Department of Economics, University of Macedonia.



知乎

首发于  
川流不息

- Engel, E. (1857). Die Produktions- und Konsumtionverhältnisse des Königreichs Sachsen. Reprinted in "Die Lebenskosten Belgischer Arbeiter-Familien Früher und Jetzt." *International Statistical Institute Bulletin*. Vol. 9, 1 – 125.
- Koenker, R. and G. Bassett, Jr. (1978). Regression Quantiles. *Econometrica*, Vol. 46(1), 33 – 50.
- Saastamoinen, J. (2008). *Quantile regression analysis of dispersion of stock returns – evidence of herding?* Working paper, Joensuu yliopisto, Taloustieteen.

**免责声明：**文章内容不可视为投资意见。市场有风险，入市需谨慎。

原创不易，请保护版权。如需转载，请联系获得授权，并注明出处，谢谢。已委托“维权骑士”（维权骑士\_免费版权监测/版权保护/版权分发）为进行维权行动。

编辑于 2019-07-03

数据分析 统计学 回归分析

▲ 赞同 81 ▼ 3 条评论 分享 收藏 ...

## 文章被以下专栏收录

**川流不息**

北京量信投资管理有限公司是一家在中国基金业协会备案登记的专业私募基金管理人...

[关注专栏](#)

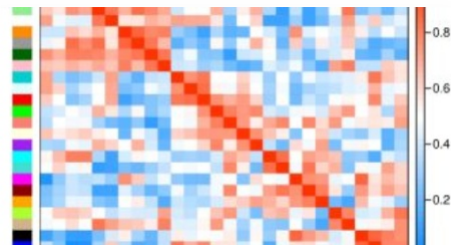
## 推荐阅读

### 多元回归分析：回归诊断

简单回顾一下多元回归分析里的参数估计问题：对于多元线性回归模型  $Y = X\beta + \epsilon$  的参数，我们有 least-square 估计量  $\hat{\beta} = (X^T X)^{-1} X^T Y$ ，从而得到模...

C.Zha...

发表于玩数据那些...



### 回归分析|笔记整理（6）——多元线性回归（上）

学弱渣


### 数据科学基本概念

摘要：本文介绍了五个特征、概率分布、统计推断、数据科学、机器学习。

3 条评论

⇌ 切换为时间排序

写下你的评论...



 一茶一世界2 个月前


用什么工具做的分位数回归?

 赞


 石川 (作者) 回复 一茶一世界2 个月前

python

 赞

 不系之舟18 天前

谢谢你，通俗易懂，小白也能看懂[赞同]。就是没有写分位数回归模型及其参数估计过程，我再去查查[调皮]

 赞

