

知乎

首发于  
川流不息

# Spurious

## It's not what it seems

## 小心伪回归发现的假关系



石川

量化交易 话题的优秀回答者

已关注

编辑推荐

92 人赞同了该文章

### 1 生活中随处可见的伪回归

回归分析在量化投资中的应用十分广泛。比如在选股或者预测股票收益率时，人们常常使用宏观经济数据或公司基本面数据等对收益率回归，以期找出能够解释收益率的自变量（又称为因子）。由于金融数据之间的关系大多为线性，因此线性回归往往就足够用了；而因为线性回归又足够简单，这就使得回归分析更加普及。

回归分析的目的是为了找到自变量和因变量之间的相关性。然而，当我们对时间序列进行回归分析时，必须要警惕一类陷阱，它就是**伪回归（spurious regression）**，它指的是自变量和因变量之间本来没有任何因果关系，但由于某种原因，回归分析却显示出它们之间存在统计意义上的相关性，让人错误地认为两者之间有关联，这种相关性称作**伪关系（spurious relationship）**。

伪回归在生活中随处可见，来看下面两个例子。

#### 例子一：冰淇淋销量和溺水儿童数



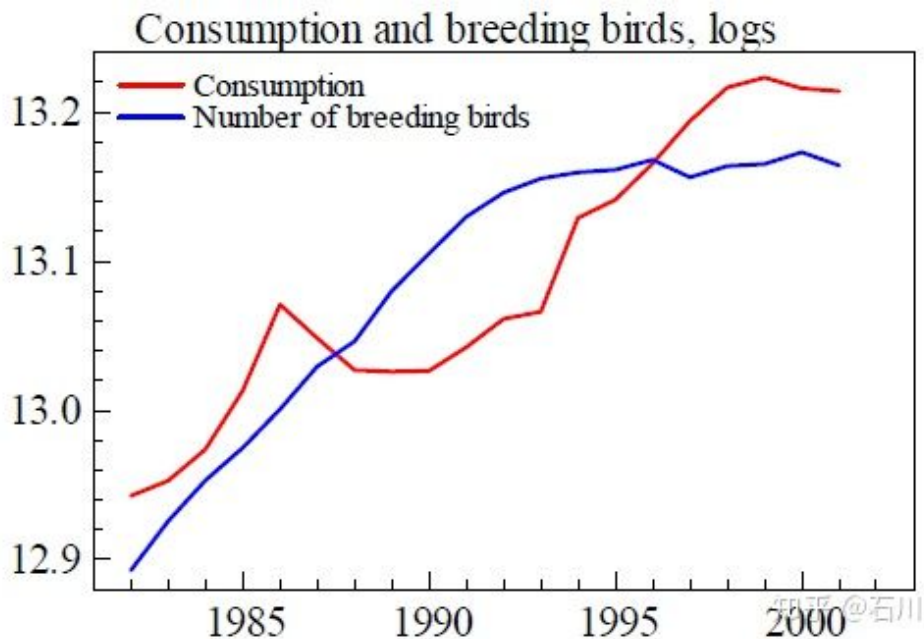
知乎

首发于  
川流不息

这两者应该有所谓的“冰淇淋卖得好，溺水儿童数就多”的关系吗？这是否意味着“游泳池的儿童都在吃冰淇淋”？正确的答案是，这仅仅是一个伪回归（下一节会解释成因）。

## 例子二：居民消费和鸬鹚个数

下图来自真实的数据，两条曲线代表的时间序列分别为丹麦居民消费的对数（红线）以及该国饲养鸬鹚数量的对数（蓝线）。从图中来看，它们显然非常相关。如果用红线对蓝线回归，得到的回归系数显著不为 0，且回归的 R-squared 高达 0.688，说明蓝线对红线的解释能力非常强。但这显然也是毫无意义的（因为居民消费和鸬鹚个数之间没有任何有逻辑的关联），它同样来自伪回归（成因和例子一不同）。



本文就来介绍伪回归。了解如何识别它，才能避免在构建量化模型时错误的使用不同数据之间的伪关系。

## 2 伪回归的成因和数学特性

伪回归的成因一般有两个。在上一节的第一个例子中，伪回归的成因是存在**干扰因素**（**confounding factor**，或称**潜在变数 lurking variable**）。在第二个例子中，伪回归的成因是两个变量之间的**局部随机趋势**（**local stochastic trend**）。

当两个变量同时受第三个因素影响时，这两个变量间可能存在误导性的相关性，这第三个因素称为**干扰因素**。在第一个例子中，干扰因素是夏天炎热的高温。高温造成了冰淇淋销量的上升；此外高温也使得更多的儿童去公共泳池从而造成溺水事故增多。高温是造成冰淇淋销量和溺水儿童数的共同因素；这两者本身之间并没有相关性。

知乎

首发于  
川流不息

了它们之间在统计意义上的伪关系；然而它二者之间并没有因果关系——我们无法说“由于居民消费的增加导致了鸬鹚数量的上升”，反之亦然。

关于随机趋势这个成因，来看另一个例子。考虑两个独立的布朗运动，它们的时间序列如下图所示（蓝线 vs 绿线）。红色方框画出的区域显示它们在局部表现出了同样的上涨或者下降趋势。如果我们用蓝线当自变量来回归绿线，得到的回归系数为 0.34（p-value 小于 0.001）。这显然是一个伪关系；这两个布朗运动相互独立，它们之间没有任何关系。



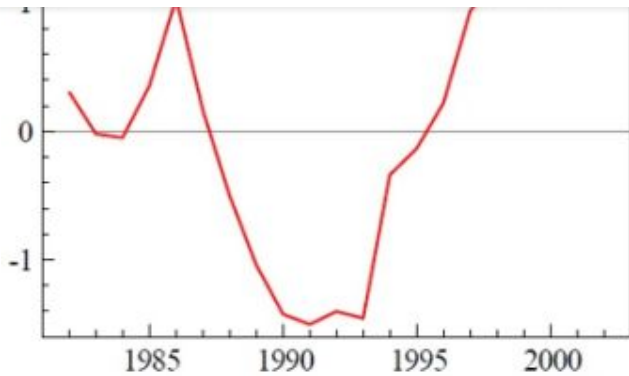
注意了，重要的话**加粗**说一遍：

**这两个布朗运动的随机趋势并非来自同一个随机运动，而是来自两个独立的随机运动。因此，这两个时间序列展现出来的局部同向运动并不是来自某个公共的因素，而仅仅是因为巧合；它们之间的关系是伪关系。**

如何避免伪回归呢？在上面的例子中，无论是居民消费、鸬鹚个数还是随机生成的布朗运动，这些时间序列都是**非平稳的 (non-stationary)**。当我们对非平稳时间序列进行回归分析时，**非常容易发现伪回归**（例外是这些时间序列满足协整关系，下文会说明）。

当伪回归出现时，回归分析得到的残差序列 (residual) 不满足平稳性，我们可以以此作为判别伪回归的依据：**如果回归分析的残差是非平稳的，说明发生了伪回归**。在上文的例子中，居民消费和鸬鹚数量回归结果的残差序列和两个布朗运动回归的残差序列分别如下图所示。这两个残差序列均不满足平稳性。

知乎

首发于  
川流不息

伪回归告诉我们：**我们不能仅仅因为两个时间序列共同运动就说它们之间一定存在相关性。**

### 3 平稳性的检验

本节就来介绍如何检验时间序列的平稳性。掌握了这个技术，我们就可以检验一个回归分析得到的残差序列是否是平稳的，从而推断是否发生了伪回归。

简单的说，**如果一个时间序列  $\{y_t\}$  在每一时刻  $t$  的取值的概率分布都一样、该分布与  $t$  无关，那么该时间序列就是平稳的。**关于时间序列平稳性的详细解释，请参考[《写给你的金融时间序列分析：基础篇》](#)中的第四节：时间序列的平稳性。

为了说明如何判断平稳性需要讲到以下三个概念：**unit root (单位根)**、**order of integration (单整阶数)**、以及**ADF 检验**。

#### 3.1 单位根

**单位根 (unit root)** 是非平稳时间序列的特性之一。对于一个时间序列  $\{y_t, t = 0, 1, \dots\}$ ，假设它可以写成  $p$  阶自回归函数如下：

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \dots + a_p y_{t-p} + \varepsilon_t$$

其中  $\varepsilon_t$  是残差序列； $a_1, \dots, a_p$  为回归系数。该时间序列的**特征方程 (characteristic equation)** 为：

$$m^p - m^{p-1}a_1 - m^{p-2}a_2 - \dots - a_p = 0$$

如果  $m = 1$  是该特征方程的一个解，则称该时间序列存在单位根。

#### 3.2 单整阶数





时间序列是**一阶单整的**（integrated of order one，记为  $I(1)$ ）；如果  $m = 1$  是一个**多重根**（重数为  $d$ ），则该时间序列是  **$d$  阶单整的**（记为  $I(d)$ ）。

单整阶数在实际中的含义是什么呢？**对于一个非平稳的时间序列，我们总可以通过差分把它变成平稳的；差分的次数就是单整阶数。**如果一个时间序列经过一次差分就变成平稳的，那么它就是一阶单整的；如果一个时间序列需要通过  $d$  次差分才能变成平稳的，那么它就是  $d$  阶单整的。

**对于我们熟悉的股票价格序列，它的一阶差分为股票的收益率；由于收益率满足平稳性，因此股票价格序列是一阶单整的。**

3.3 ADF 检验

从上面的介绍可知，**要想判断一个时间序列是否满足平稳性，核心就是看它有没有单位根。**为此，可以采用 ADF 检验（全称为 Augmented Dickey-Fuller test）。将  $\{y_t\}$  的自回归函数转化为  $y_t$  增量  $\Delta y_t$  的形式：

$$\Delta y_t = \underbrace{\alpha}_{\text{常数项}} + \underbrace{\beta t}_{\text{趋势项}} + \underbrace{\lambda y_{t-1}}_{\text{是否有单位根}} + \underbrace{\delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1}}_{p \text{ 阶自相关项}} + \underbrace{\epsilon_t}_{\text{随机误差}}$$

知乎 @石川

在上式中，如果时间序列  $\{y_t\}$  存在单位根，则  $\lambda = 0$ 。ADF 检验的原假设是  $\lambda = 0$ 、备择假设是  $\lambda < 0$ 。

ADF 检验

原假设为  $\{y_t\}$  存在单位根，即  $\lambda = 0$ 。该检验的统计量是  $\lambda$  和它自身标准误差之比， $\lambda / SE(\lambda)$ 。如果  $\{y_t\}$  满足平稳性，则  $\lambda / SE(\lambda)$  显著为负。因此只有当这个统计量小于给定显著性水平的阈值（阈值是负数）时，我们才能在对应的置信水平下拒绝原假设、接受备择假设，备择假设为  $\{y_t\}$  不存在单位根、即满足平稳性。

**对于回归分析得到的残差序列，通过 ADF 检验考察其是否存在单位根。如果能够在给定的显著性水平下拒绝原假设，则可以认为残差序列满足平稳性，从而推断出回归分析得到的相关性可信、没有发生伪回归。**

4 协整 —— 处理非平稳时间序列的利器

在量化投资领域，收益率序列满足平稳性，而价格序列不满足平稳性。**收益率满足平稳性仅仅价格呈现随机游走，它对于构建赚钱的投资策略几乎没有什么用。我们想要的是价格序列呈现**

知乎

首发于  
川流不息

不幸的是，现实中投资品价格基本上都呈现（几何）布朗运动（见《布朗运动、伊藤引理、BS 公式》系列文章）。这意味着投资品的价格均不满足平稳性的要求，因此如果我们想用其他数据——比如宏观经济数据——来预测投资品价格（比如上证指数）的走势就没什么意义，因为会发生伪回归。

好消息是，这里有一个例外：**虽然单一投资品的价格不满足平稳性，但有时我们可以把多个投资品（通常是两个）线性组合在一起构成一个价差序列，而这个价差序列满足平稳性。**

在数学上，如果多个非平稳的时间序列通过线性组合得到一个平稳的时间序列，则把满足这种关系称为协整（co-integration）。为什么会发生协整的？考虑两个投资品的价格序列为  $\{X_{1t}\}$  和  $\{X_{2t}\}$ ，它们的走势可以表述为：

$$X_{1t} = \sum_{i=1}^t \varepsilon_{1i} + \text{初始值} + \text{平稳过程}$$

$$X_{2t} = \sum_{i=1}^t \varepsilon_{2i} + \text{初始值} + \text{平稳过程}$$

其中  $\varepsilon_{1i}$  和  $\varepsilon_{2i}$ ， $i = 1, \dots, t$  为构成  $X_{1t}$  和  $X_{2t}$  的随机过程。假设这两个价格序列的一个线性组合为：

$$Z_t = \sum_{i=1}^t \varepsilon_{1i} - \beta \sum_{i=1}^t \varepsilon_{2i} + \text{初始值} + \text{平稳过程}$$

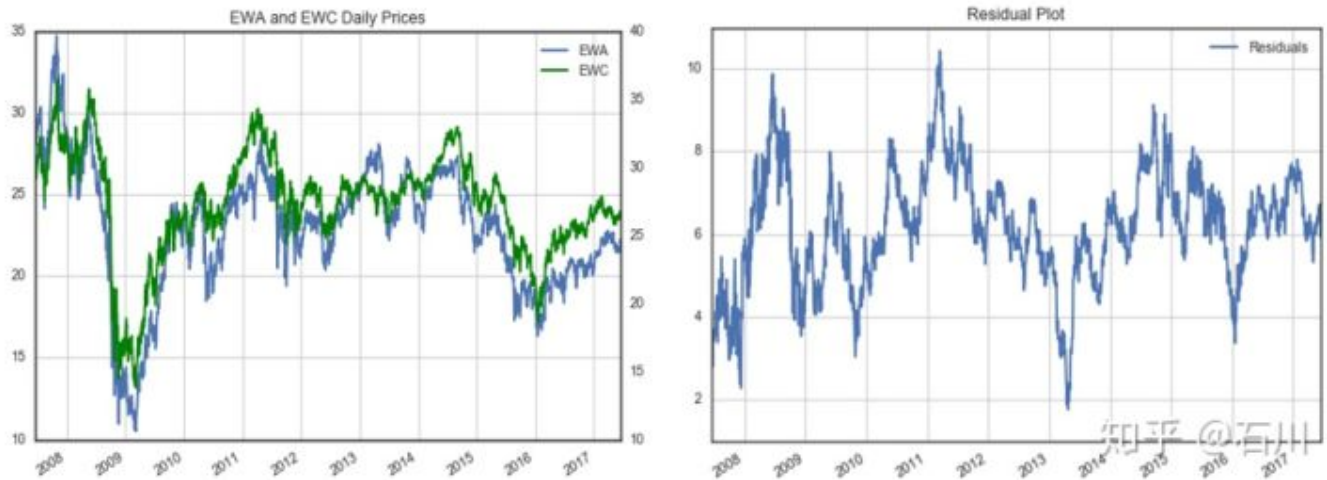
如果存在  $\beta$  使得这两个价格序列中的随机过程恰好能够抵消掉，则价差序列  $Z_t$  满足平稳性。在这种情况下， $X_{1t}$  和  $X_{2t}$  满足参数为  $\beta$  的协整关系。

$\beta$  在什么情况下存在呢？在第二节例子中，两个布朗运动的随机性由不同的随机运动主宰，它们的局部共同运动纯属巧合。而当协整发生时，**这两个价格序列的随机过程能够抵消掉的根本原因是它们的随机性来自同一个随机过程（共同的因素）**。只有在这种情况下，两个价格序列才可能发生协整，它们的价差序列才能满足平稳性。

在量化投资领域，协整的例子虽然不是随处可见，但也绝非寥若晨星。在不同交易所交易的追踪同一投资品（比如标普 500 指数或者比特币）的金融工具，它们的价差就满足平稳性，因为它们价格的波动来自同一投资品。又如股指 ETF 和成分股之间线性组合得到的价差，股指 ETF 的随机性来自成分股的波动。

协整的存在使得构建出的价差序列满足平稳性；更形象的说，**价差序列围绕零呈现均值回归（均值回复）的特性**。这种特性构成了量化投资领域的一大类策略——均值回归策略（例子见[这个问题的答复](#)）。

EWA 和 EWC 的配对交易就是一个经典的例子。他们分别代表澳大利亚（EWA）和加拿大（EWC）股指的两个 ETFs。由于这两个国家的经济都主要依靠商品，因此可以认为这两个股指的波动来自共同的因子。下图为这两个 ETFs 的价格序列（左图）和回归得到的价差序列（右图）。



对价差进行 ADF 检验，得到的统计值为 -4.09（p-value 为 0.0065），小于显著性 1% 对应的阈值 -3.96，这说明我们可以在 1% 的显著性水平下拒绝原假设。ADF 检验说明该价差序列满足平稳性，即 EWA 和 EWC 满足协整关系。

## 5 结语

**对时间序列做回归时，一定要检验平稳性。**

在量化投资领域，做回归分析时，收益率和价格是两类因变量。对于收益率，它已经满足了平稳性，因此我们只需要保证回归中的自变量也满足平稳性。这就是为什么在使用宏观经济数据预测收益率时，我们通常会使用同比或者环比（都是一阶差分了）作为自变量，而非经济数据的累计值。

对于投资品价格来说，它自身是一阶单整的，不满足平稳性。在回归时，应争取找到和它协整的另外一个（或一组）投资品价格，用它们的线性组合得到一个满足平稳性的价差序列，从而构建均值回归策略。另外我们见到的做法（通常很危险）是使用宏观经济数据直接预测股票的价格走势。由于宏观经济数据和股票价格很难满足协整关系，因此这种回归得到的往往是伪关系。

*Spurious: it is not what it seems ...*

（全文完）

**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。

编辑于 2019-07-02

回归分析    统计套利    时间序列分析

▲ 赞同 92    ▼    2 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



**DARNN:一种新的时间序列预测方法——基于双阶段注意...**

罗未萌



**金融时间序列分析入门【一】**

陈颖      发表于量化哥




**时间序列预法+ARIM**

biaobiaode

2 条评论 ⇌ 切换为时间排序

因作者设置，评论已关闭



MixViper

Pearl的因果推理能否解决假相关性的问题呢？

👍 赞

1 年前



Harry Zhu

关注正迁移的意义在于... 模型对线性拟合是容易拟合的... 通过V-MAR-B的方式回归出来系...

1 年前



知乎



首发于  
川流不息

□。

👍 1

