Regression Model

- 本节课我们要传授预测股价，发家致富的技巧（hurray！）
  - 其实是介绍几个回归可能出现的问题

△ 尝试 I：

- 正常我们会认为股票的价格是围绕 fundamental value 波动的

$$R_t = \underset{>0}{a} + \underset{<0}{b} \times P_{t-1} + \varepsilon_t$$

如果 P 越大，理论上 R 会越来越小 ⇒ $b<0$（含因 $P \to \infty$），显而易见

- 下面用数据回归一下

我们来看一下实际的回归结果：

似乎结果是很好的，而 2006–2015 的数据是最佳，一定程度上是因为 using recent data

	Model: $R_t = a + b \times P_{t-1} + \varepsilon_t$			Model: $R_t = a + b \times \ln(P_{t-1}) + \varepsilon_t$		
Period	Intercept a	Slope b	R^2	Intercept a	Slope b	R^2
1991-2015	0.04154	-0.00002	0.019	0.24587	-0.03227	0.0326
	2.8293	*-2.4046*		*3.3027*	*-3.1695*	
2001-2015	0.02461	-0.00001	0.012	0.14752	-0.01882	0.0075
	1.5176	*-1.4397*		*1.1850*	*-1.1625*	
2006-2015	0.07076	-0.00002	0.050	0.53212	-0.06652	0.0519
	2.7283	*-2.4972*		*2.5849*	*-2.5418*	

那我们能用这个模型估计 2016 年的结果么？

No！用模型估计 16 年和 17 年都错了！Predicted return 和 actual return 几乎是反的　看来这条路走不通

△ 尝试 II

- 那我们换一条路走：我们认为股票价格的波动是服从正态分布的：

- $R_t = \mu + \sigma \varepsilon_t$，$\varepsilon_t \sim N(0, 1)$

下面问题来了，我们希望知道 $E[R_t \mid P_{t-1}]$（用昨天的价格预测今天的回报）

$E[R_t \mid \ln S_{t-1}] = a + b \ln S_{t-1}$

$a \to \mu$，$b \to 0$

	a	b
Average Coefficient:	0.1469938	-0.02713
Number < 0	1	92

b应该是0的，为什么呢？

我们认为 $R_t = \ln S_t - \ln S_{t-1}$，$\ln S_t = \ln S_{t-1} + \mu + \varepsilon_t$

$E[R_t \mid \ln S_{t-1}] = E[\ln S_t - \ln S_{t-1} \mid \ln S_{t-1}] = E[\mu + \varepsilon_t \mid \ln S_{t-1}]$

$= \mu$，跟 $\ln S_{t-1}$ 没有太大关系

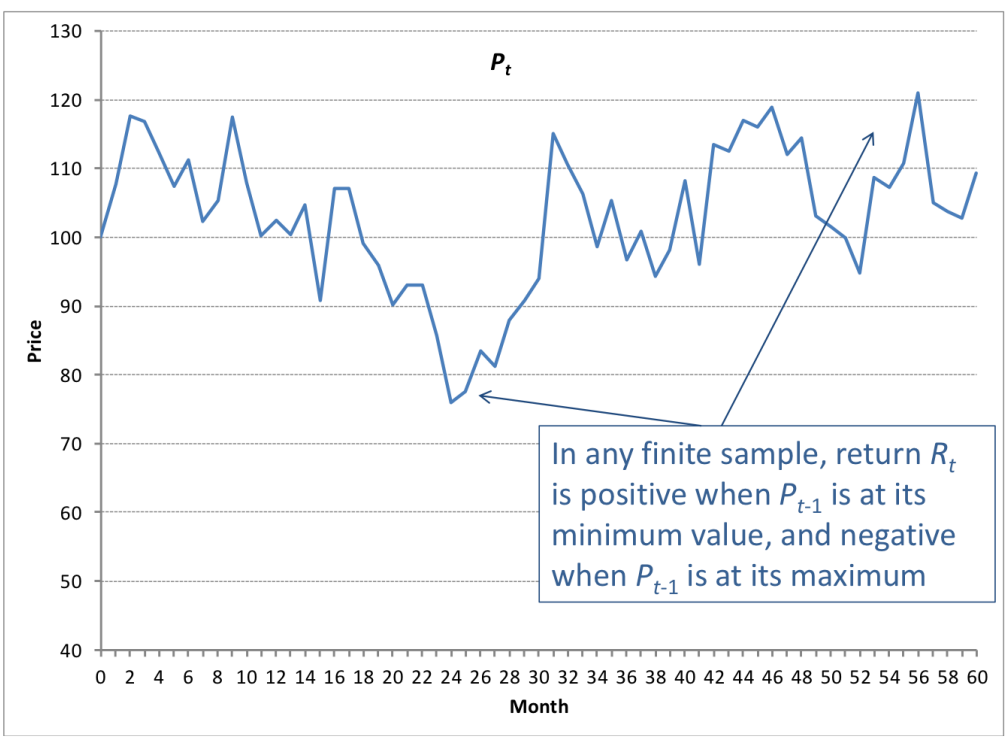
而实际上 b 是负的！如右图所示，Monte Carlo 结果 b 几乎总是负的

一是Excel的问题么？我们用R试试：

samplemean	-0.0623234108868797

STILL NEGATIVE !

- Why? Intuitive Answer:



price is high, R_t goes down, price is low. R_t goes up

$a + b \ln S_{t-1}$

$= < 0$

那真正的解释是什么呢？ Stationary!

我们需要 stationary to make our regression valid

右侧是我们用的方法.

$$\hat{b} = \frac{\frac{1}{T}\sum_{t=1}^{T}(R_t - \bar{R}_t)(\ln(P_{t-1}) - \overline{\ln(P_{t-1})})}{\frac{1}{T}\sum_{t=1}^{T}(\ln(P_{t-1}) - \overline{\ln(P_{t-1})})^2}$$

我们依靠 Law of Large Numbers

其核心是数据的 variance 是 FINITE 的

→ 确定了 ↑ # ⇒ ↓ var.

但在 Stock Price 中. variance becomes INFINITE. $var = \sigma^2 t$

$t \to \infty$. $var \to \infty$

所以用 price 来 regression is basically nonsense

Be very skeptical of regressions involving levels of prices!

我们为什么要这么费劲的讲这个呢？ Because it's tempting!

— "我们就是想这么干！" It always seems to work

ws上个沿了问题1：Stock Prices are likely to have an autocorrelation coefficient ϕ_1 equal to or very close to 1, but it's likely to be estimated to be lower than 1, which in turn suggests predictability

A Little Bit on Time Series

一 什么是 stationary process 以及我们为什么用它

先讲讲 TS：
- a sequence of observations in chronological order
- a sample from a stochastic process

- Stationary Processes
  - Strictly stationary process

$Y_1, Y_2, Y_3, \cdots, Y_n, \cdots$ joint distribution

$Y_{1+m}, Y_{2+m}, \cdots Y_{n+m}$ joint distribution

eg. the distribution of stock returns in February is the same as the distribution in March & January (log returns)

* Return process is stationary but price process is non-stationary

↓

它的 variance grows with time

- Why is stationary so important?

我们喜欢 repeated independant trials → estimate the population

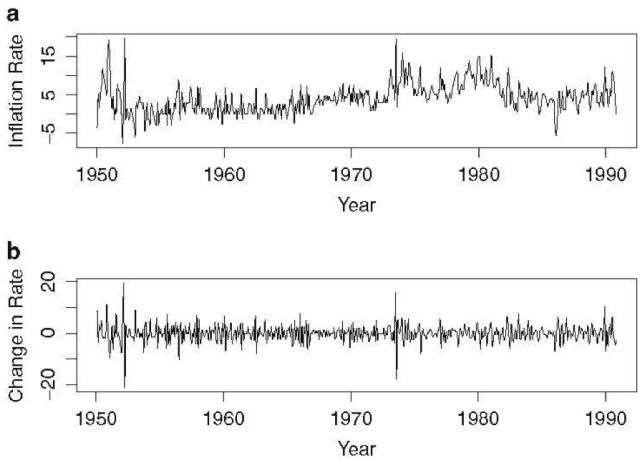
而在 time series data 中我们只有一条时间线，不允许有 repeated independent trials. What's our justification? In cross-sectional work. 有如个人. 按统计学每个人都是 independent trial. In time series, 我们依靠 stationary. Stationary 给了我们什么呢？Distribution is the same in different parts for time series. 一段 period 可以看作是另一段的 repeated independant trials

注意：MA process $R_t = \mu + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_k \varepsilon_{t-k}$, $\varepsilon_t \sim N(0,1)$ is stationary

AR process $R_t = \mu + \phi_1 R_{t-1} + \phi_2 R_{t-2} + \cdots + \phi_k R_{t-k} + \theta \varepsilon_t$ $\varepsilon_t \sim N(0,1)$ is stationary if there are some restrictions on the coefficients & distribution of the first observation

e.g. $R_t = \mu + \phi_1 R_{t-1} + \theta \varepsilon_t$ is not stationary if $|\phi_1| \geq 1$

因为相当于方差在不断增大 with time



看起来不像 Stationary

看起来很像 Stationary

- Weakly Stationary Process

$Y_1, Y_2, Y_3, \cdots, Y_n, \cdots$ $Y_{1+m}, Y_{2+m}, \cdots Y_{n+m}$

means & covariances = means & covariances

- 这个实际上比较重要，因为我们只需要 mean & covariances 来得到 least squares

Another Problematic Regression

- Let us use one example to illustrate:

Consider a bivariate regression of two highly persistent series

我们用 Excel 生成两列随机数，它们一定是毫无关联的

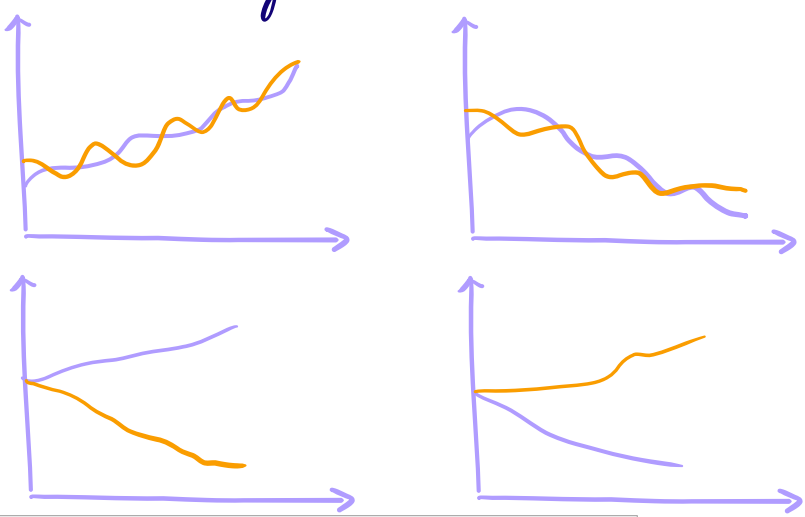
从 ln100 开始，每次加 0.01 to 0.08 × NORMSINV(RAND()))

BUT!

	a	b	t-statistic	|t-statistic|
Average:	-0.5344082	1.183527	3.21178	5.975822

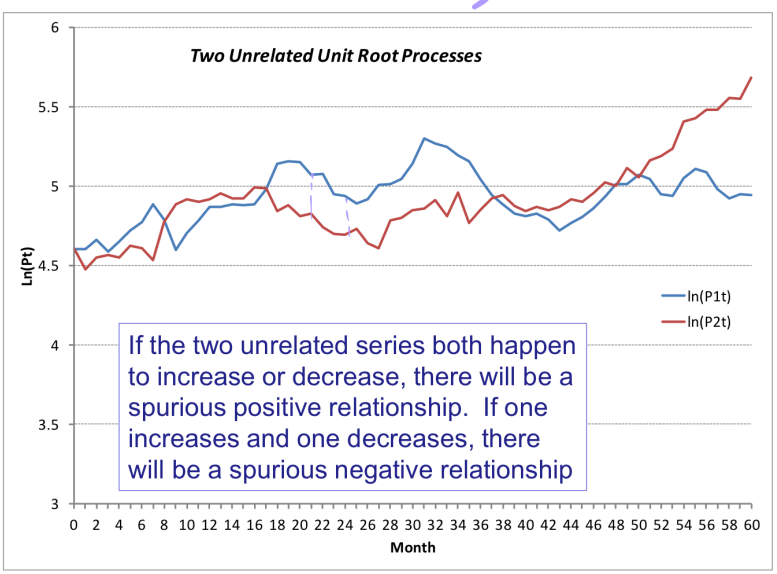
b是大于0的！ t-statistic 巨大！

Why? Intuitively:



都上升或是都下降

一下上升一下下降



residuals are highly correlated

But ε_t and ε_{t-k} should be uncorrelated

以上介绍了问题2：Spurious Regression

回归两个不相关的时间序列，standard t-tests会显示 $b \neq 0$，而实际上 $b = 0$

如何解决这个问题呢？看 first differences 的回归中 b 是否显著：

$$(S_{1t} - S_{1t-1}) = a + b(S_{2t} - S_{2t-1}) + e_t$$

VARs & VECMs

- 引入：exercise 10

$\ln S_{1,0} = \ln(100)$, $\ln S_{1,t} = \ln S_{1,t-1} + \mu_1 + \varepsilon_{1,t}$

$\varepsilon_{1,t} \sim N(0, \sigma_1^2)$ is independent of $\varepsilon_{1,u}$ for $u \neq t$

$\ln S_{2,0} = \ln(100)$, $\ln S_{2,t} = \ln S_{2,t-1} + \mu_2 + \varepsilon_{2,t}$

$\varepsilon_{2,t} \sim N(0, \sigma_2^2)$ is independent of $\varepsilon_{2,u}$ for $u \neq t$

$\text{corr}(\varepsilon_{1,t}, \varepsilon_{2,t}) = \rho$

(a) What's the variance of the random variable $y_t = \varepsilon_{2,t} - \varepsilon_{1,t}$

$V = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$

(b) $\ln S_{1,0} = \ln S_{2,0} = \ln(100)$, what's the variance of $\ln S_{2,t} - \ln S_{1,t}$

$\ln S_{2,t} - \ln S_{1,t} = t(\mu_2 - \mu_1) + \sum_{s=1}^{t} v_s$

$= tV \rightarrow$ grows with time

上面的模型十分简单，但它揭示了什么问题呢？一般的 linear combination var 都 grow with time

Is the process given by the difference $\ln S_{2,t} - \ln S_{1,t}$ stationary?

NO! Variance grows with time

Are $\ln S_{2,t}$ & $\ln S_{1,t}$ cointegrated?

w/ vector $(1, -1)$ No this process has to be stationary

注意！上面两支股票 are related ($\rho \neq 0$), but are not cointegrated

Cointegration implies a particular form of predictability — difference can't be too large

- Finance: Basic ideas
  - Efficient Market Hypothesis
  - The risk that carries risk premium are the risks can't be diversified
  - No-arbitrage / Existance of risk-neutral probability

接下来我们进入正题：VAR：Vector Autoregressions

VAR(1) $R_t = \phi_0 + \Phi R_{t-1} + \varepsilon_t$, $Var(\varepsilon_t) = \Sigma$,

where R_t is a $k \times 1$ vector of variables

如果是双变量：

$$R_{1,t} = \phi_{0,1} + \Phi_{11} R_{1,t-1} + \Phi_{12} R_{2,t-1} + \varepsilon_{1,t}$$

$$R_{2,t} = \phi_{0,2} + \Phi_{21} R_{1,t-1} + \Phi_{22} R_{2,t-1} + \varepsilon_{2,t}$$

Ford 的 return 可能可以预测 General Motor 的 return

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- 有 unrestricted 和 restricted 的区别

上面两个式子是 unrestricted. 但是如果是：

$R_{2,t} = \phi_{0,1} + \Phi_{21} R_{1,t-1} + \varepsilon_{2,t}$ → restricted. $\Phi_{22} = 0$

我们这节课主要讨论 unrestricted 的情况

VAR only depends on lagged variables

→ it's immediately useful in forecasting

那会出现什么问题呢？

我们用 $R_{2,t-1}$ 来预测 $R_{1,t}$：

$$R_{1,t} = a + b R_{2,t-1} + e_t$$

很容易得到错误的结论：如果 $R_{1,t}$ is persistent，只取决于 its own past value，但 $R_{2,t-1}$ is correlated with $R_{1,t-1}$，所以 $R_{2,t-1}$ 会看起来像 "cause" $R_{1,t}$ 命实际上是 $R_{1,t-1}$ causes $R_{1,t}$

以上介绍了问题3

如何解决这个问题呢？

$$R_{1,t} = \phi_{0,1} + \Phi_{11} R_{1,t-1} + \Phi_{12} R_{2,t-1} + \varepsilon_{1,t}$$

$$R_{2,t} = \phi_{0,2} + \Phi_{21} R_{1,t-1} + \Phi_{22} R_{2,t-1} + \varepsilon_{2,t}$$

接下来我们定义 Granger causality

$R_{2,t}$ is said to Granger cause $R_{1,t}$ if $\Phi_{12} \neq 0$

$R_{1,t}$ is said to Granger cause $R_{2,t}$ if $\Phi_{21} \neq 0$

但是 causality 不一定真意味着 cause，今天的天气预报可以预测明天的天气，但显然没有 causality，just predictors

Q 如果天气预报是根据气流，洋流 …… 一系列决定明日天气的 factors 得到的，但如果把这些 factors 都列出来系数还会不为0吗？

一些时候 Stock returns 可以用作 predictor

但是有时候 linear combination 是 stationary 的. 举个例子:

$S_{1t} = \phi_0 + S_{1,t-1} + \varepsilon_{1t}$ → not stationary

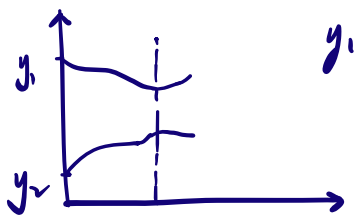
$S_{2t} = bS_{1t} + \varepsilon_{2t}$ → not stationary

But $S_{2t} - bS_{1t}$ IS stationary

我们可以利用 Vector error-correcting model (VECM) 来分析有上面这个特征的问题:

如果我们认为有一个长期的 relationship: $y_1 - 2y_2 = 0$

if $y_1 > 2y_2$, y_1 会下降 $y_1 < 2y_2$, y_1 会上升



我们需要 difference to be stationary

Strength of reversion to long-run relation

$$\Delta y_t = \begin{pmatrix} -0.4 \\ 0.1 \end{pmatrix} (1 \quad -2) y_{t-1} - \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} \Delta y_{t-1} + \varepsilon_t$$

long-run relation

ARMA model (just AR here)

$$y_t - y_{t-1} = \begin{pmatrix} -0.4 & 0.8 \\ 0.1 & -0.2 \end{pmatrix} y_{t-1} - \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} y_{t-1} + \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} y_{t-2} + \varepsilon_t$$

$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} y_{t-1}$

$$y_t = \begin{pmatrix} -0.2 & 0.1 \\ 0.5 & -0.2 \end{pmatrix} y_{t-1} + \begin{pmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{pmatrix} y_{t-2} + \varepsilon_t$$

如果不知道 long-relation 具体是什么怎么办呢? 不能 construct single RHS variable. 那是分别求两个 Regression