

Quantitative Analysis of Financial Markets

Session 7: VAR and VECM

Benjamin Ee
Week 7



Class 7
Pre-class Prep 1/8



Vector Auto Regressions (VAR)

Vector autoregressions – an introduction to modelling multiple variables

只建模一个变量的历史：之前的时间序列预测方法，比如简单预测、指数平滑和ARIMA模型，主要是通过该变量自身的~~历史值~~进行预测，这些方法属于单变量建模。

1. So far, we have been modelling **only 1 variable as a function of its history**
2. This is whether we are using **naïve, exponential smoothing or ARIMA methods**
3. What if we want to do a **time series forecast of two variables simultaneously?**
 - a) We could apply above methods to each variable individually. Just **model them one at a time**
 - b) What if the variables **depend on each other?**

预测多个变量的需求：如果我们想同时预测两个或更多变量，该怎么办呢？这就要求从单变量预测转变为多变量预测。多个变量之间可能存在相互影响，因此建模时不能仅依赖单变量的历史。

两种方式处理多个变量的预测：

方法A：将每个变量当作独立的时间序列，分别应用已有的单变量方法，即逐个建模。
方法B：如果这些变量之间相互依赖，就需要同时建模，考虑它们之间的依赖关系。这种需求引出了向量自回归（VAR）模型，它可以同时处理多个变量并考虑变量之间的相互影响。

Vector auto regression – motivating example 1

1. National GDP is organized into 3 sectors:

- a) Services
- b) Manufacturing
- c) Agriculture

2. Services GDP (S_t) depends on other 2 sectors

- a) Firms in the manufacturing and agricultural sector may consume various inputs (e.g. legal, financial services, IT consulting)

制造业和农业会消耗服务业的投入品，例如法律服务、金融服务、IT咨询等。

3. Manufacturing GDP (M_t) depends on other 2 sectors

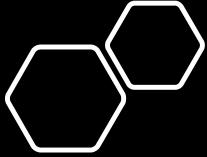
- a) Firms in the agricultural sector may consume manufacturing output such as heavy farm equipment, while firms in the services sector may require fixed equipment such as computer equipment

农业可能需要制造业提供的重型农用设备，而服务业可能需要制造业的固定设备（如计算机设备）。

4. Agriculture GDP (A_t) depends on other 2 sectors

- a) Employees of firms from the other 2 sectors need to eat!

农业部门的工人需要消费其他两个部门提供的商品和服务。



Vector auto regression – motivating example 2

消费是税后收入的函数：在宏观经济模型中，消费 (C_{t+1}) 通常被建模为税后收入的函数。如果税后收入增加，消费者会倾向于更多消费。

1. In many macroeconomics models, **consumption is modelled as a function of after tax income**
2. i.e. $C_{t+1} = f(I_t, e_t)$ where f is an increasing function. If consumers have greater after tax income, they will consume more 反向因果关系：消费的增加可以推动总需求和经济活动，进而带来收入增长。这就形成了一个反馈循环，即消费与收入相互影响
3. **Reverse is also possible.** If consumption increases, this will stimulate aggregate demand, and greater economic activity. In turn, this may also lead to income growth
4. I.e. $I_{t+1} = g(C_t, e_t)$ where g is also an increasing function for this example 消费与投资的关系：投资 (I_{t+1}) 可能依赖于当前的消费和收入，因此不同的经济变量之间往往存在复杂的相互关系，这使得同时建模多个变量的需求变得重要。

这说明了传统的单变量模型无法捕捉变量之间的反馈效应，而VAR模型能够处理多个变量并识别它们之间的相互依赖。

Vector auto regression – setup

1. Formally: 方程中，每个变量既依赖于它自己的滞后值（过去的值），也依赖于其他变量的滞后值，再加上一个误差项。
 - a) $I_t = c_1 + A_{l1}C_{t-1} + A_{l2}I_{t-1} + e_{t,l}$
 - b) $C_t = c_2 + A_{c1}I_{t-1} + A_{c2}C_{t-1} + e_{t,c}$
2. This is referred to as **2-dimensional VAR** of order 1, or VAR(1)
3. Note that **both white noise realizations $e_{t,l}, e_{t,c}$ are allowed to be contemporaneously correlated.** i.e. within same time index 白噪声误差项：误差项 $e_{t,l}$ 和 $e_{t,c}$ 是允许相互之间在同一时间点内存在相关性的。这种同时的相关性意味着两者可能共同受到某些外部冲击的影响。
4. Preprocessing:
 - a) If any of the individual series are non-stationary (use nsdiffs or ndiffs to check), **we take 1 or more differences to make them stationary**
 - b) We then fit a VAR model on the output of **differenced series**

预处理步骤：

 - 1.如果任何变量是非平稳的（即存在趋势），在建模前需要对数据进行差分处理（即计算前后两个时刻数据的变化），使其平稳。
 - 2.平稳化后可以使用VAR模型进行拟合

Vector auto regression - setup

1. VAR模型的简化：在VAR模型中，我们只考虑自回归（AR）项，而不包括移动平均（MA）项，这样做的好处是模型结构更简单且容易应用。VAR模型的简洁性使其在时间序列预测中更常用，相较于复杂的VARMA模型。

2. 协整的特殊情况：
如果两个或多个变量是非平稳的，但它们的线性组合是平稳的，那么这些变量就是协整的。协整表明变量之间存在长期的均衡关系。

3. VECM（向量误差修正模型）：
如果发现变量之间存在协整关系，可以在VAR模型的基础上加入误差修正项，从而构建出VECM模型。
VECM是VAR模型的一个特殊情况，专门处理协整变量之间的短期波动和长期均衡关系。

1. For ease of exposition, we include **only AR terms, and not MA terms**. Simplicity of VARs have led to their dominance in forecasting over Vector ARMA
2. There is a special case which could **allow us to improve our VAR model**
 - a) Two (or more) of the variables we are forecasting are not stationary, **but a linear combination of them is stationary**. i.e. two or more of the variables in the system are cointegrated
 - b) In this case, we can **add additional terms to the VAR model ("error correction terms") in order to improve it**. The resulting model is called a Vector Error Correction Model (VECM), which formally is a special case of a VAR
 - c) VECMs will be **discussed in a later clip** within this video sequence
VECM能够在VAR模型的框架下，进一步改进非平稳变量的建模性能，特别是在它们存在协整关系时。

总结来说，向量自回归（VAR）模型是一种强大的工具，可以用于捕捉多个时间序列变量之间的相互关系。它扩展了单变量模型，通过考虑多个变量之间的相互依赖，提供了更准确的预测和分析方法。而对于存在长期均衡关系的变量，VECM模型是VAR的进一步扩展，用于处理协整现象。



Class 7
Pre-class Prep 2/8

1. VAR模型包含更多参数：相比单变量自回归（AR）模型，VAR模型涉及的参数要多得多，因为它需要同时建模多个变量及其滞后项。
2. 系数计算公式：假设变量的总数为 K ，滞后项的总数为 p ，那么 VAR 模型中要估计的系数总数为 $K + pK^2$ 。
每个滞后项需要为每个变量估计 K 个系数。
总体上，我们需要估计的系数数量是 $K(pK+1) = K + pK^2$

Model selection for VAR

1. VAR models typically involve significantly more parameters than univariate AR estimation
2. If total number of variables is K and total number of lags is p , then the number of coefficients to be estimated in a VAR is $K + pK^2$

For each lag, we have to estimate K coefficients where K is # variables

Eqn 1

$$\begin{aligned} I_t &= c_1 + A_{I1}C_{t-1} + A_{I2}I_{t-1} + e_{t,I} \\ C_t &= c_2 + A_{c1}I_{t-1} + A_{c2}C_{t-1} + e_{t,C} \end{aligned}$$

For each equation, we therefore estimate 1 constant, plus K coefficients for each lag.
i.e. if there are p lags, then pK coefficients + 1 constant

Total across K equations is therefore $K(pK + 1) = K + pK^2$

通过这两个等式可以看出每个变量的滞后项不仅依赖自身的过去值，还依赖其他变量的过去值。

Use of SIC/BIC/SC criteria for model selection in VAR

1. Large number of coefficients to be estimated can **substantially lower the degrees of freedom of the regression**, which can hurt the accuracy of parameter estimates and therefore forecasts given by the model
2. Compared to BIC/SIC/SC (different names for same metric), AIC tends to pick models with large number of lags in practice. In the case of VAR models, we therefore **prefer to use BIC to chose between models**
3. **Next 2 screens show the penalty factor as number of coefficients increase for AIC versus BIC/SIC/SC and adjusted R²**

1. 系数过多降低自由度： VAR模型中的大量系数可能大幅降低回归的自由度，从而影响参数估计的准确性，进而影响模型的预测性能。

2. 模型选择标准：相比其他标准（如AIC），BIC（贝叶斯信息准则）更倾向于选择滞后项较少的模型。在VAR模型选择中，建议使用BIC来选择模型，因为它对复杂模型（包含更多滞后项）的惩罚更大，从而可以防止过拟合。

3. 惩罚因子的比较：下一步的讨论会展示在AIC、BIC/SIC/SC、调整后的R²标准下，随着系数数量增加，惩罚因子如何变化。

自由度的惩罚: 为了在估计样本外均方误差 (MSE) 时调整自由度, 我们需要对MSE进行惩罚。公式
 s^2 代表对自由度惩罚后的方差

Penalizing Degrees of Freedom

- We need to correct somehow for degrees of freedom K when estimating out-of-sample MSE on the basis of in-sample MSE
- To highlight the degree-of-freedom penalty, rewrite s^2 as a penalty factor times the MSE,
自由度 K 会影响我们计算的方差, 我们通过重新表示方差 s^2 来体现这种惩罚:

$$s^2 = \left(\frac{T}{T-K} \right) \frac{\sum_{t=1}^T \hat{e}_t^2}{T} = \frac{1}{1 - \frac{K}{T}} \frac{\sum_{t=1}^T \hat{e}_t^2}{T} = \frac{1}{1 - \frac{K}{T}} \text{MSE}.$$

T: 样本数 K: 要估计的参数数目

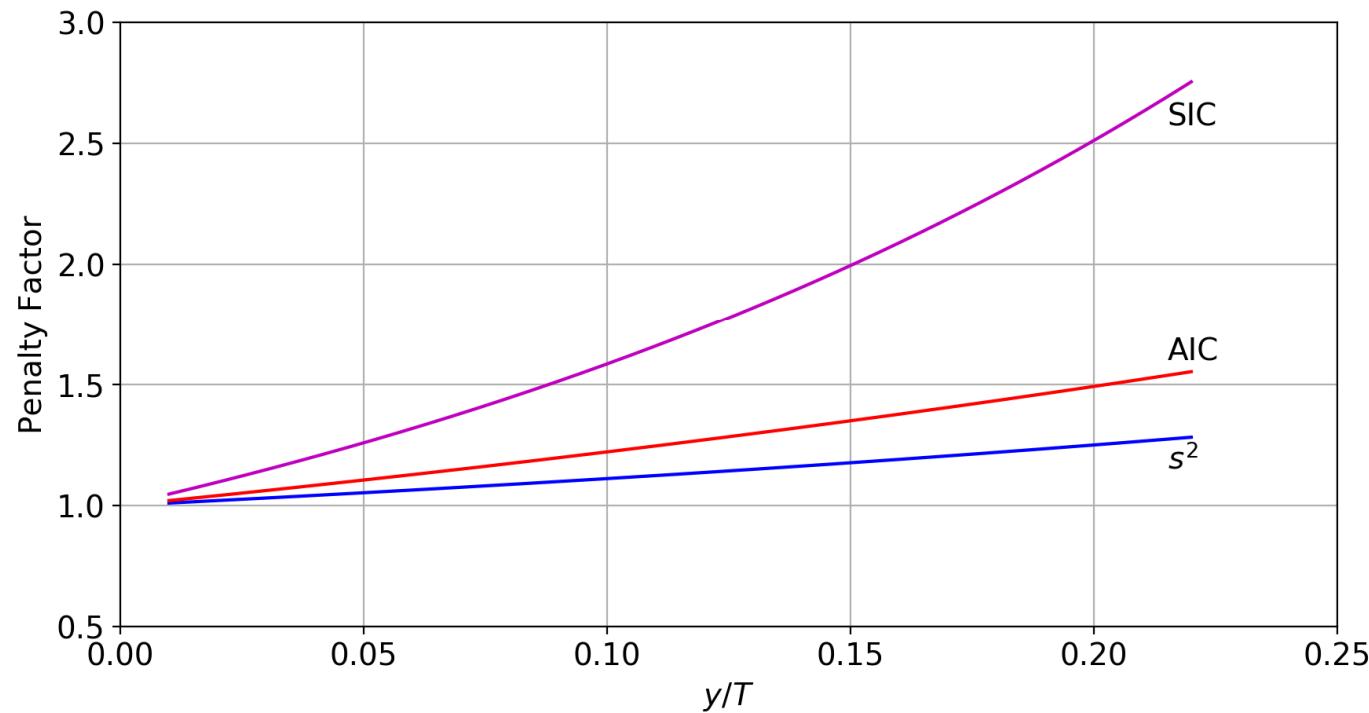
- Two very important such criteria are the Akaike information criterion (AIC) and the Schwarz information criterion (SIC). Their formulas are given below. Note there are also ln() versions of these being as widely used, where we apply ln() function to both sides

$$\text{AIC} = e^{\frac{2K}{T}} \text{MSE} \quad \text{and} \quad \text{SIC} = T^{\frac{K}{T}} \text{MSE}.$$

信息准则的计算: 常用的信息准则有AIC (赤池信息准则) 和SIC (施瓦兹信息准则), 它们通过增加对自由度的惩罚来调整模型的复杂度。

惩罚因子的比较：图表展示了不同信息准则（AIC、SIC、调整后的R²）的惩罚因子随系
数数量增加的变化。
SIC的惩罚因子增长得最快，这表明它对于模型的复杂度有更强的惩罚力度，因此它倾向
于选择较简单的模型。
AIC的惩罚因子增长相对较慢，因此在选择模型时，它可能会倾向于选择滞后项更多的
复杂模型。
调整后的R²则介于两者之间。

Comparison of Information Criteria





Class 7
Pre-class Prep 3/8

VAR estimation in R

- VAR estimation in R is done by the vars package.
- Install it with `install.packages("vars")`

```
> install.packages("vars")
WARNING: Rtools is required to build R packages but is not currently installed. Please download
appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'D:/OneDrive/Documents/R/win-library/4.0'
(as 'lib' is unspecified)
also installing the dependencies 'strucchange', 'sandwich'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/strucchange_1.5-2.zip'
Content type 'application/zip' length 984351 bytes (961 KB)
downloaded 961 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/sandwich_3.0-1.zip'
Content type 'application/zip' length 1488749 bytes (1.4 MB)
downloaded 1.4 MB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/vars_1.5-6.zip'
Content type 'application/zip' length 426554 bytes (416 KB)
downloaded 416 KB

package 'strucchange' successfully unpacked and MD5 sums checked
package 'sandwich' successfully unpacked and MD5 sums checked
package 'vars' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\users\bbe4\AppData\Local\Temp\RtmpwDAC0Y\downloaded_packages
>
```

Summary of uschange dataset

```
Console Terminal × Jobs ×
~/

>
>
>
>
> head(uschange)
   Consumption    Income Production    Savings Unemployment
1970 Q1  0.6159862  0.9722610 -2.4527003 4.8103115          0.9
1970 Q2  0.4603757  1.1690847 -0.5515251 7.2879923          0.5
1970 Q3  0.8767914  1.5532705 -0.3587079 7.2890131          0.5
1970 Q4 -0.2742451 -0.2552724 -2.1854549 0.9852296          0.7
1971 Q1  1.8973708  1.9871536  1.9097341 3.6577706         -0.1
1971 Q2  0.9119929  1.4473342  0.9015358 6.0513418         -0.1
>
>
>
>
>
>
>
>
>
```

Install R library *vars* in order to work with vector autoregressions

R中的VARselect()函数：该函数能够根据4个信息准则自动选择VAR模型的最佳滞后阶数p

AIC: 赤池信息准则，用于选择滞后阶数，类似于ARIMA模型的情况。

HQ: Hannan-Quinn准则，用于平衡模型复杂度和拟合优度。

SC (SIC) : 施瓦兹信息准则 (BIC)，对复杂模型的惩罚较高，因此倾向于选择更简单的模型。

FPE: 最终预测误差，另一种选择模型的准则。

通常基于BIC/SIC/SC来选择模型：这些信息准则帮助选择合适的滞后阶数，以避免过拟合。

从VAR(1)开始：最初通常从VAR(1)开始，并逐步调整，直到残差中不再有显著的序列相关性为止。

Selecting order of a VAR model

- Vars package in R contains function VARselect() which automatically lists the best lag p to chose according to 4 information criteria
 - AIC is same as for ARIMA case
 - HQ is the Hannan-Quinn criterion
 - SC is another name for BIC/SIC**
 - FPE is Final Prediction Error
- We will typically **base our decision on SC/BIC**
- From this output, we **chose to start with VAR(1)**, i.e. just t-1 terms
- We will **iterate on the selection of VAR(1) until there is no longer any residual serial correlation**

```
Console Terminal × Jobs ×
~/
>
>
>
>
> VARselect(uschange[,1:2], lag.max=8)
$selection
AIC(n) HQ(n) SC(n) FPE(n)
      5       1       1       5

$criteria
      1       2       3       4       5
AIC(n) -1.3787011 -1.374328 -1.4064675 -1.4171747 -1.4305682
HQ(n)  -1.3353784 -1.302123 -1.3053813 -1.2872066 -1.2717184
SC(n)  -1.2718613 -1.196262 -1.1571748 -1.0966554 -1.0388225
FPE(n)  0.2519071  0.253017  0.2450268  0.2424391  0.2392472
      6       7       8
AIC(n) -1.4033960 -1.3870036 -1.3736910
HQ(n)  -1.2156644 -1.1703903 -1.1281958
SC(n)  -0.9404238 -0.8528049 -0.7682657
FPE(n)  0.2458869  0.2500197  0.2534617

> uschange[,1:2]
            Consumption        Income
1970 Q1  0.61598622  0.97226104
1970 Q2  0.46037569  1.16908472
1970 Q3  0.87679142  1.55327055
1970 Q4  -0.27424514 -0.25527238
1971 Q1  1.89737076  1.98715363
1971 Q2  0.91199291  1.44733417
1971 Q3  0.79453885  0.53181193
1971 Q4  1.64858747  1.16012514
1972 Q1  1.31372218  0.45701150
1972 Q2  1.89147495  1.01662441
1972 Q3  1.53071400  1.90410126
1972 Q4  2.31829471  3.89025866
1973 Q1  1.91072016  0.70825266
```

VAR estimation in R

1. Residuals from both lag 1 and lag2 VAR contain serial correlation at the 5% level of significance
2. Residuals from a VAR(3) does not. We therefore finalize our model as VAR(3)

```
>
>
>
>
> var1 <- VAR(uschange[,1:2], p=1, type="const")
> serial.test(var1, lags.pt=10, type="PT.asymptotic")

Portmanteau Test (asymptotic)

data: Residuals of VAR object var1
Chi-squared = 49.102, df = 36, p-value = 0.07144

> var2 <- VAR(uschange[,1:2], p=2, type="const")
> serial.test(var2, lags.pt=10, type="PT.asymptotic")

Portmanteau Test (asymptotic)

data: Residuals of VAR object var2
Chi-squared = 47.741, df = 32, p-value = 0.03633

>
>
> var3 <- VAR(uschange[,1:2], p=3, type="const")
> serial.test(var3, lags.pt=10, type="PT.asymptotic")

Portmanteau Test (asymptotic)

data: Residuals of VAR object var3
Chi-squared = 33.617, df = 28, p-value = 0.2138

>
>
>
>
> |
```

Estimation results for VAR 3

消费方程和收入方程: VAR(3)模型的估计结果包括消费和收入的不同滞后项的系数估计值。

消费方程: 消费的当前值是其自身的滞后值和收入滞后值的函数。

收入方程: 同样, 收入的当前值也依赖于消费和收入的滞后值。

该幻灯片展示了如何通过估计得到滞后变量对当前变量的影响系数。

VAR Estimation Results:

=====

Estimated coefficients for equation Consumption:

=====

Call:

Consumption = Consumption.l1 + Income.l1 + Consumption.l2 + Income.l2 + Consumption.l3 + Income.l3 + const

Consumption.l1	Income.l1	Consumption.l2	Income.l2	Consumption.l3	Income.l3	const
0.19100120	0.07836635	0.15953548	-0.02706495	0.22645563	-0.01453688	0.29081124

Estimated coefficients for equation Income:

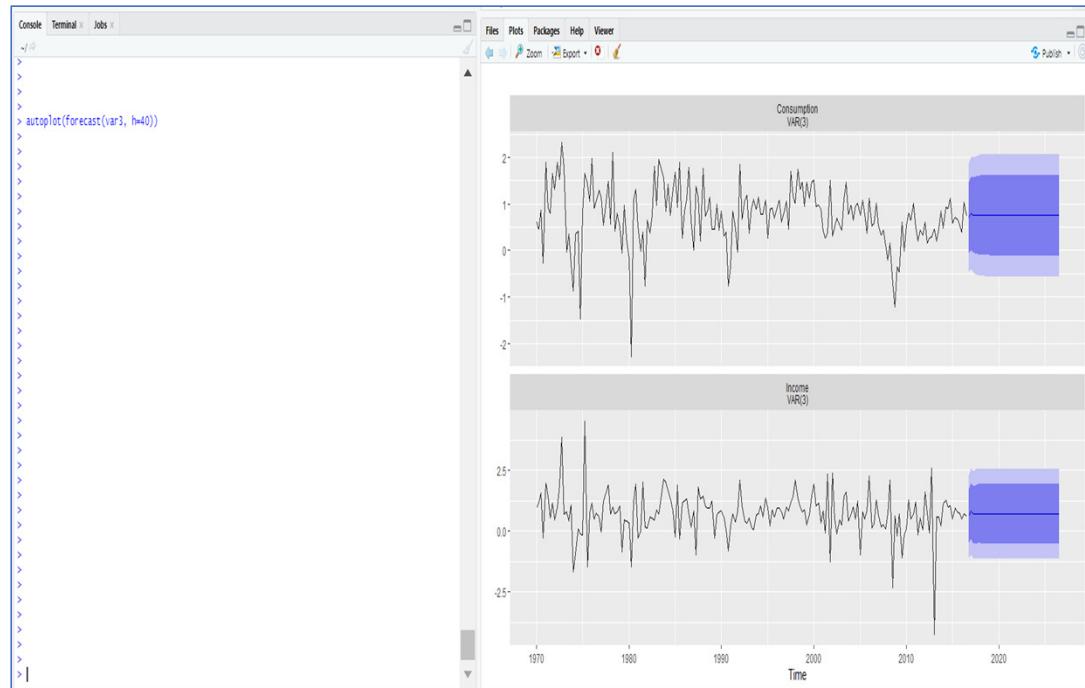
=====

Call:

Income = Consumption.l1 + Income.l1 + Consumption.l2 + Income.l2 + consumption.l3 + Income.l3 + const

Consumption.l1	Income.l1	Consumption.l2	Income.l2	Consumption.l3	Income.l3	const
0.45349152	-0.27302538	0.02166532	-0.09004735	0.35376691	-0.05375916	0.38749574

VAR forecasting in R



```
> forecast(var3, h=8)
Consumption
  Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
2016 Q4      0.7123447 -0.05077827 1.475468 -0.4547512 1.879441
2017 Q1      0.8007622  0.01322936 1.588295 -0.4036653 2.005190
2017 Q2      0.7618103 -0.04324168 1.566862 -0.4694105 1.993031
2017 Q3      0.7493208 -0.08517980 1.583821 -0.5269377 2.025579
2017 Q4      0.7605353 -0.08393109 1.605002 -0.5309646 2.052035
2018 Q1      0.7558663 -0.09380268 1.605535 -0.5435903 2.055323
2018 Q2      0.7526257 -0.10156072 1.606812 -0.5537397 2.058991
2018 Q3      0.7531992 -0.10339547 1.609794 -0.5568493 2.063248

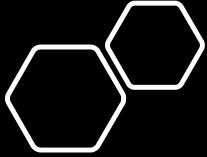
Income
  Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
2016 Q4      0.6151330 -0.4925059 1.722772 -1.0788545 2.309121
2017 Q1      0.8320821 -0.3354124 1.999577 -0.9534466 2.617611
2017 Q2      0.7069462 -0.4607386 1.874631 -1.0788736 2.492766
2017 Q3      0.7013130 -0.5068821 1.909508 -1.1464619 2.549088
2017 Q4      0.7272275 -0.4842301 1.938685 -1.1255369 2.579992
2018 Q1      0.7184218 -0.4954638 1.932307 -1.1380560 2.574900
2018 Q2      0.7125025 -0.5040207 1.929026 -1.1480092 2.573014
2018 Q3      0.7159150 -0.5018556 1.933686 -1.1465045 2.578335
```

1. 预测未来数据：使用R中的`forecast()`函数，可以对VAR(3)模型进行预测，输出未来8个季度的消费和收入的点预测值及其置信区间（80%和95%）。

2. 图形表示：左侧的图展示了预测值的趋势及其区间范围，右侧则是具体的数值预测结果。
VAR模型可以通过捕捉变量间的相互影响，预测未来多个时间点的数据

Benchmarking AR and VAR models: Example using uschange

```
>
> outofsampleperiod = 100
> uschange_split = ts_split(uschange, sample.out = outofsampleperiod)
> accuracy(x = uschange_split$test[,1:1], forecast(auto.arima(uschange_split$train[,1:1]), h=outofsampleperiod))
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set  0.002087312 0.7508783 0.5788739 56.62850 235.2483 0.6864562 -0.02309173    NA
Test set     -0.072620225 0.5031792 0.3784413 76.00672 169.1536 0.4487738  0.49982102 0.6027824
> var3 = VAR(uschange_split$train[,1:2], p = 3, type="const")
> var_fc = forecast(var3, h = outofsampleperiod)
> accuracy(x = uschange_split$test[,1:1], var_fc$forecast$Consumption)
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set  9.185742e-17 0.7539401 0.5847083 63.39585 246.5113 0.6933749 0.01638406    NA
Test set     -7.662364e-02 0.5029185 0.3784619 72.29153 166.7528 0.4487981 0.50934005 0.5993514
>
>
>
> 对比AR和VAR模型的预测精度：使用accuracy()函数对两种模型的预测结果进行评估，包括误差指标如RMSE（均方根误差）、MAE（平均绝对误差）等。
> AR模型的误差和VAR模型的误差在训练集和测试集上的表现可以通过这些指标来比较。
> 这表明通过对比AR和VAR模型的性能，可以选择最适合的数据模型。
```



Putting it together

- How can we combine VAR with multiple TS decomposition (complex seasonality) for better forecasts?
- Overall methodology
 - Generally following TS decomposition, **seasonal components are often modelled using snaive.**
 - Having multiple seasonal components **may result in a more accurate and responsive forecast** for this part
 - **Seasonally adjusted component can then be modelled with ETS, ARIMA, or VAR**
 - **Ljung-Box test can be performed in residuals** from fitted values (same as before)
 - Goodness of fit should be primarily evaluated **out of sample** using test set

结合VAR与时间序列分解: 通过将VAR与多个时间序列分解（如季节性分解）结合，可以提高预测的准确性。

季节性分量通常可以用简单的模型（如naive预测）进行处理。

调整后的季节性分量可以再通过ETS、ARIMA或VAR模型进行进一步建模。

模型评估: 使用Ljung-Box检验来检查残差中是否仍存在自相关。模型的拟合优度应该主要通过样本外数据进行评估。

通过结合VAR模型和季节分解，可以提高对具有复杂季节性结构数据的预测效果。



Class 7
Pre-class Prep 4/8

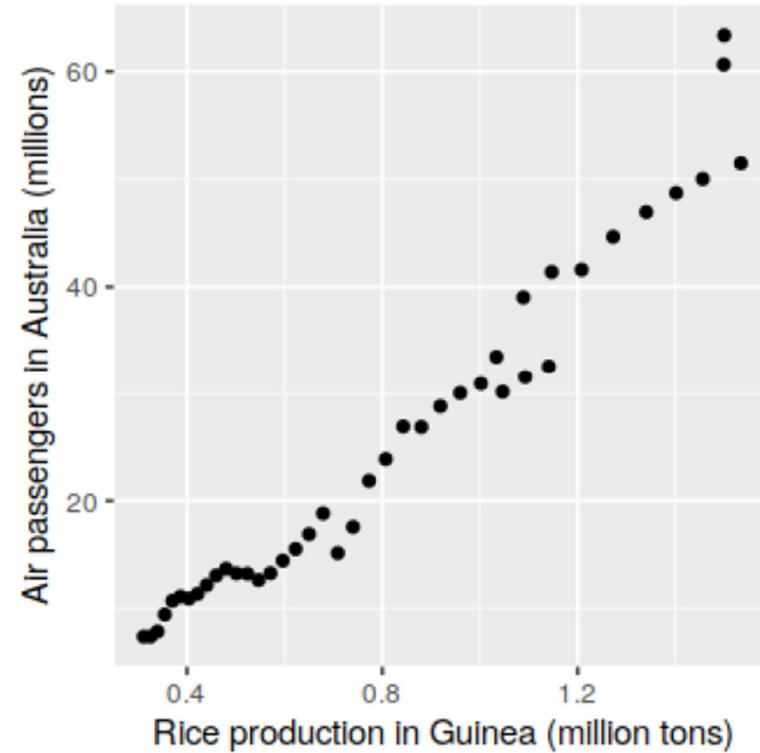
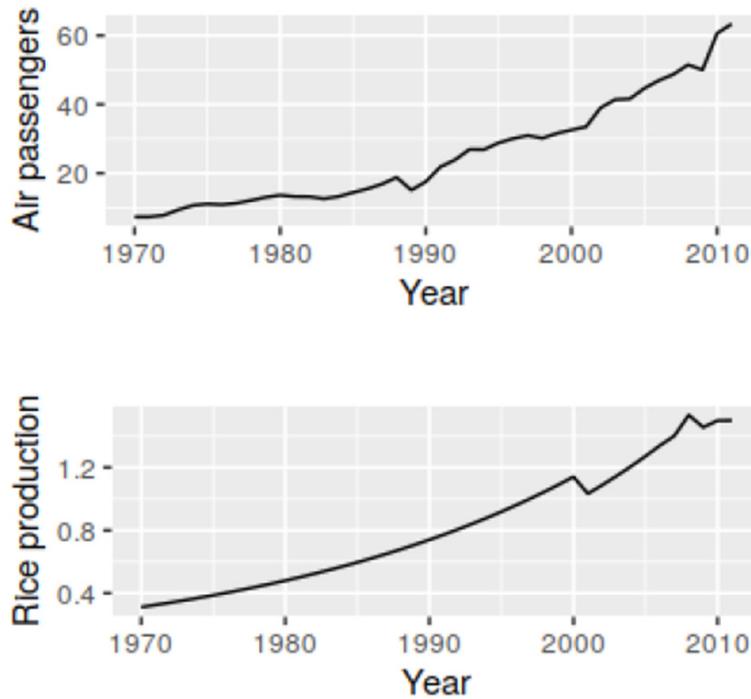


Non-Stationary Explanatory Variables

虚假回归：当回归中使用的时间序列是非平稳的时，可能会得出虚假的回归结果。这些时间序列的平均值或方差不恒定，导致它们看起来具有相关性，即使实际上并无逻辑关联。

Spurious Regression

- Spurious regression can result from non-stationary time series whose values do not fluctuate around a constant mean or with a constant variance
- Eg, if both series are trending upwards, they appear to be related;



25

例子：如果两个时间序列（如上图中的乘客数量和稻米生产）都在上升趋势中，它们在回归分析中可能表现出强烈的相关性，但这种关系实际上是误导性的，因为两个变量之间并没有实际的因果关系。

- 1.高t统计量和R平方值: 在虚假回归中, 回归的R平方值可能非常高(超过95%), t统计量可能也会显示出显著性。
- 2.低p值: 回归分析中的低p值可能会暗示回归结果具有统计显著性, 但这些显著性结果可能是由于时间序列的非平稳性引起的, 而非真实的因果关系。
- 3.残差自相关: 高R平方值和显著的自相关性是虚假回归的典型迹象。使用Breusch-Godfrey检验可以检测回归残差中的序列相关性。

```

>
>
> #demonstration of nonstationary exogenous variables
> aussies = window(ausair, end = 2011)
> aussiesmodel = tslm(aussies ~guinearice)
> summary(aussiesmodel)

Call:
tslm(formula = aussies ~ guinearice)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.9448 -1.8917 -0.3272  1.8620 10.4210 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.493     1.203   -6.229 2.25e-07 ***
guinearice   40.288     1.337   30.135 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

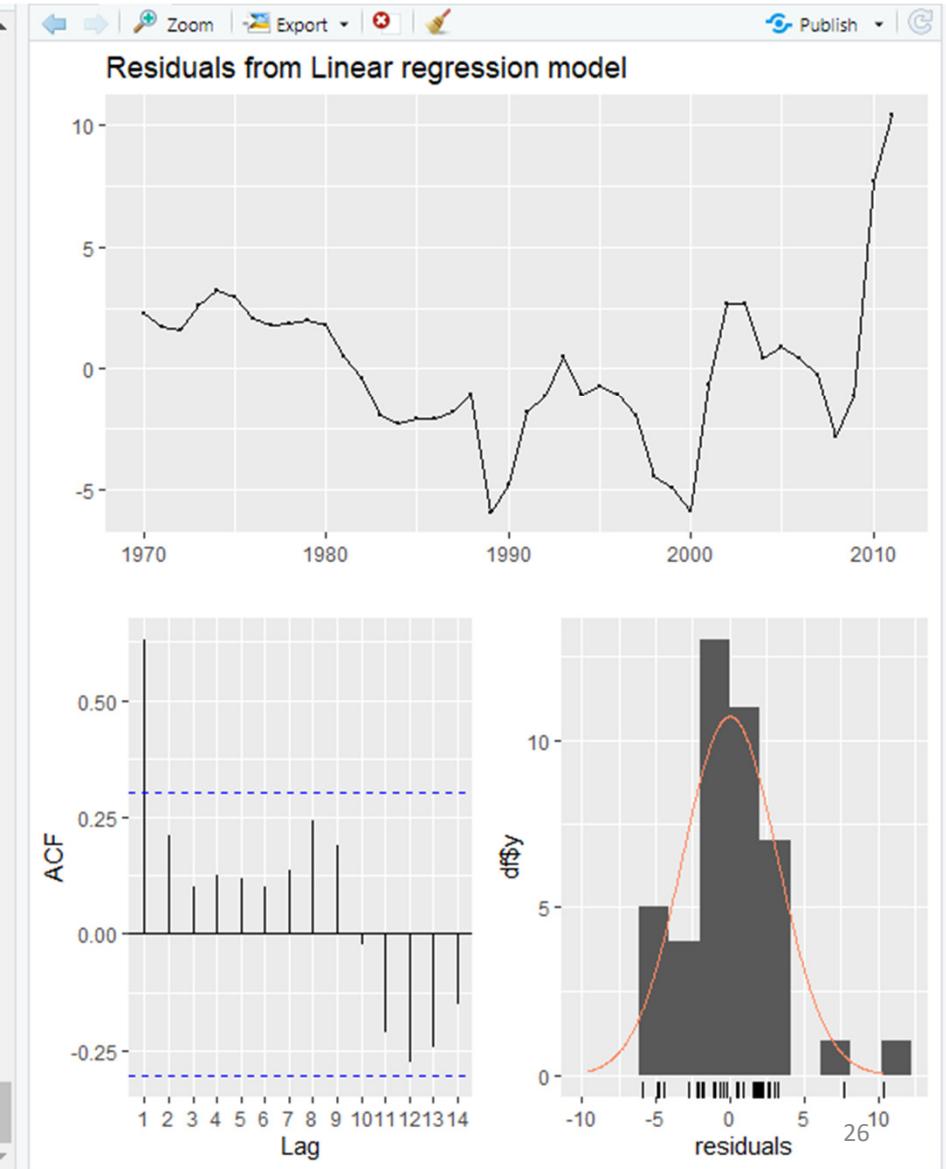
Residual standard error: 3.239 on 40 degrees of freedom
Multiple R-squared:  0.9578,    Adjusted R-squared:  0.9568 
F-statistic: 908.1 on 1 and 40 DF,  p-value: < 2.2e-16

> checkresiduals(aussiesmodel)

Breusch-Godfrey test for serial correlation of order up to 8

data: Residuals from Linear regression model
LM test = 28.813, df = 8, p-value = 0.000342
>
>
>
>
>
1. Very high t-statistics and R-square (>95%!)
2. Very low p values for F-statistic
3. Very low p values for checkresiduals
>
>
Very high R square and significant autocorrelation are signs of
spurious regressions
>

```

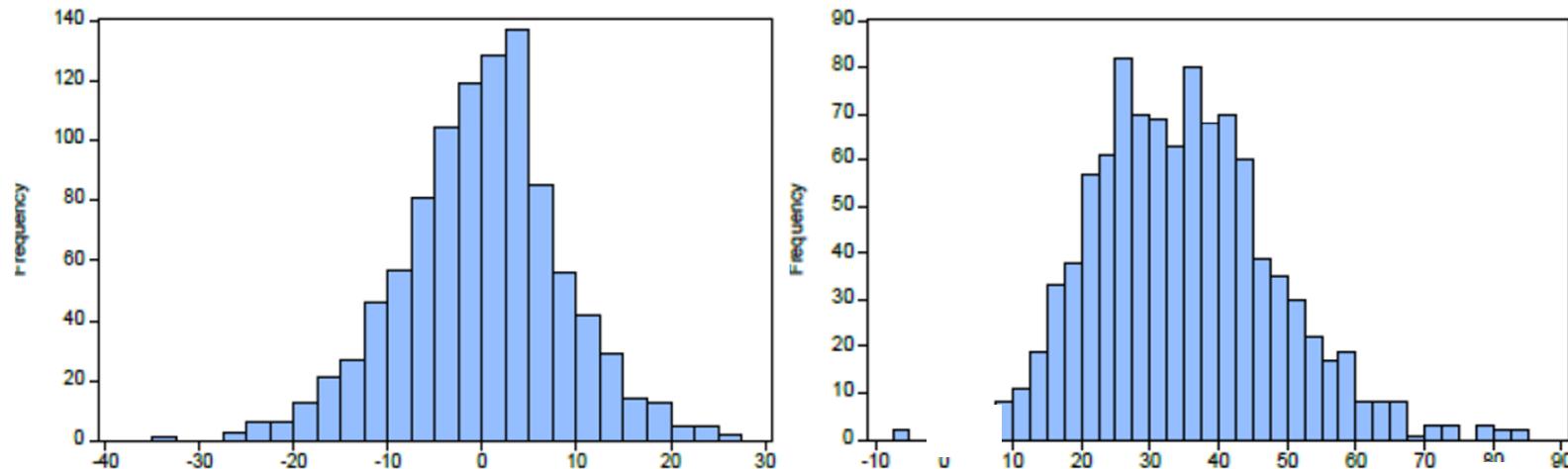


模拟独立随机漫步：通过生成两个独立的随机漫步时间序列，并进行回归，可以发现这些序列的回归结果会显示出显著的t统计量，甚至有超过80%的回归显著性。这说明了时间序列趋势的一致性可能会导致虚假的回归结果。

正漂移的随机漫步：如果随机漫步序列有正的漂移，t统计量会更大，表示两个无关的变量看起来具有强烈的关联性，但实际上只是因为它们都有上升的趋势。

Spurious Regression

- Generate 2 independent random walks and regress one on the other. Repeat 1000 time gives left histogram of t-statistics; indicate over 80% of significant t-statistics!
- Repeat the simulation exercise but with independent random walks with positive drift gives right histogram of t-statistics; Very large positive t-statistics are picking up little more than the fact that both series are trending upwards



同一时期的正向冲击：如果两个时间序列（如 y_t 和 x_t ）在同一时期都受到正向冲击，并且这些序列是非平稳的，那么它们的冲击效应可能会持续一段时间，使得两个变量看起来具有相关性。

Why Spurious Regression?

- If y_t and x_t both receive a positive shock in the same period,
 - the shock dissipates quickly if the series are stationary and it will not affect very many observations

平稳序列中的冲击效应：如果时间序列是平稳的，冲击的效应会很快消失，不会对回归产生较大影响

- the shock will persist if the series are not stationary; they may both stay above their means for a long time. It will appear that x_t being above its mean predicts y_t being above its mean

避免虚假回归：确保时间序列是平稳的（通过差分等方法），可以避免虚假回归的发生。通常需要对时间序列进行差分处理，以确保它们围绕一个恒定的均值波动。

- To avoid spurious regression,
- need to ensure y_t and x_t are both stationary, say by taking differencing (will discuss this under dynamic regression)



Class 7
Pre-class Prep 5/8



Implications of Cointegration for Time Series Regression

Cointegration

- Many time series are non-stationary but “move together” over time i.e. they are cointegrated

非平稳时间序列：许多时间序列是非平稳的，但它们可能会随着时间共同移动。这意味着这些序列存在某种长期的关系，即协整。

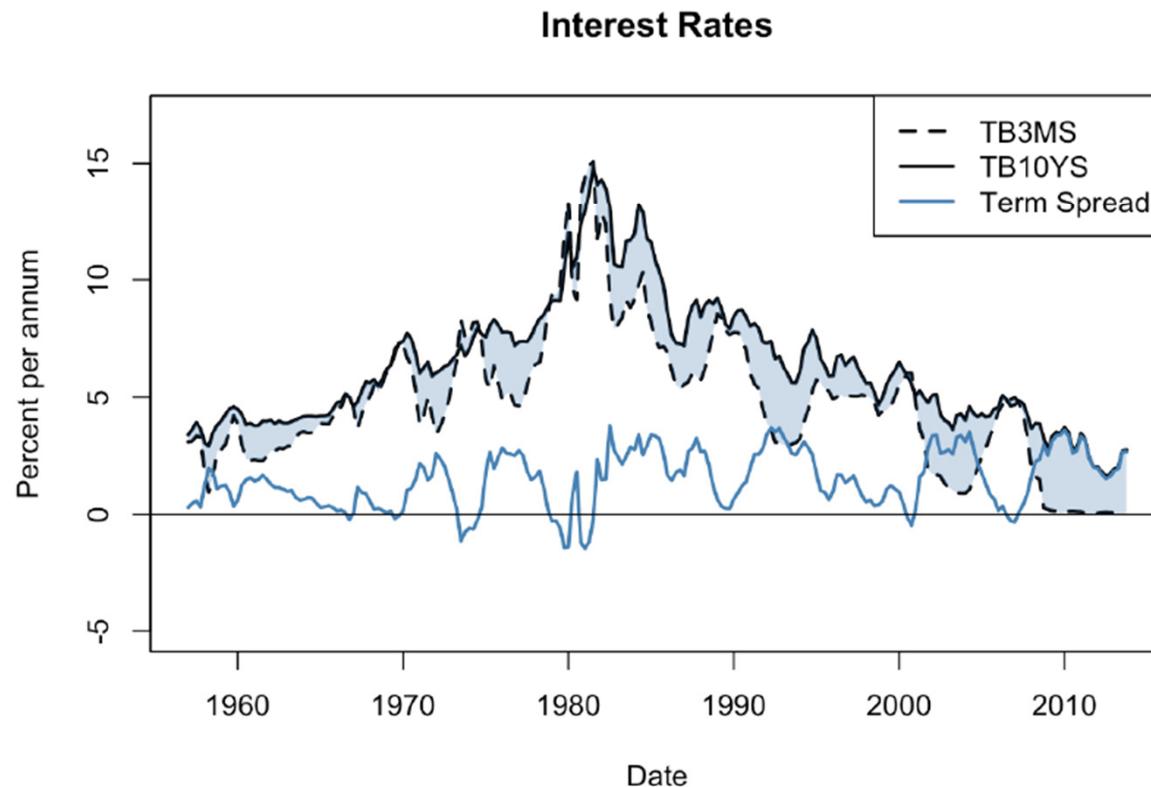
- A cointegrating relationship may also be seen as a long term relationship 协整关系可以看作是这些时间序列的长期均衡关系。
- If series are cointegrated, it means that a linear combination of them will be stationary

如果两个时间序列协整，它们的线性组合将是平稳的。例如，如果两个I(1)过程 y_t 和 x_t 存在一个 β_1 使得 $y_t - \beta_1 x_t$ 是平稳的，那么 y_t 和 x_t 就是平稳的

- For two I(1) processes, y_t and x_t , if there is a β_1 such that $y_t - \beta_1 x_t$ is stationary, we say that y_t and x_t are cointegrated 没有协整意味着这些序列在长期内可能会彼此分离。
- No cointegration implies that series could wander apart without bound in the long run

短期和长期利率的关系：图中显示了美国3个月国库券和10年期国债的利率。这些利率看起来是协整的（它们之间的差是平稳的）。这说明它们存在长期均衡关系。

Example: Short- vs Long-Term Interest Rates



- U.S. 3-month treasury bills and 10-year treasury bonds appear cointegrated; their difference appear stationary

Engle-Granger Cointegration Test

- To test for cointegration, we can use the Engle-Granger test.
要测试协整，可以使用Engle-Granger检验
- First, check that the orders of integration for both variables are the same. E.g. use kpss test, or determine d using nsdiff function.
首先，检查两个变量的积分阶是否相同。例如，可以使用 KPSS 检验或确定 d 的方法。
- If not the same, there is no way that a cointegrating relationship exists.
- If the same, we perform OLS on levels of the series
如果积分阶相同，则执行序列的水平回归:
$$y_t = \beta_0 + \beta_1 x_{1,t} + \varepsilon_t$$
 这代表长期关系。

If cointegrating exists, this is the long run relation.

- Perform Augmented Dickey-Fuller (ADF) unit root test on the residuals but compare the test-statistic to critical values from Engle-Granger table.
然后，进行 ADF 单位根检验，检验残差是否存在单位根。如果拒绝单位根假设（即残差是平稳的），则两个序列存在协整关系

H0: Unit Root in Residuals /No Cointegration between Series

H1: Residuals are Stationary /Series are Cointegrated

表格提供了 Engle-Granger 协整检验的临界值。表格根据样本大小 N 和模型中估计的系数数目 K，给出了不同置信水平（例如99%、95%）的临界值。

Engle-Granger Cointegration Test

Table 3: Critical Values of the Engle-Granger Cointegration t-Test Statistics

N	K	Augumented					
		Model I (EG)			Model I (AEG)		
		Probability to the Right of Critical Value					
		99%	95%	90%	99%	95%	90%
50	2	-4.32	-3.67	-3.28	-4.12	-3.29	-2.90
100	2	-4.07	-3.37	-3.03	-3.73	-3.17	-2.91
200	2	-4.00	-3.37	-3.02	-3.78	-3.25	-2.98
50	3	-4.84	-4.11	-3.73	-4.45	-3.75	-3.36
100	3	-4.45	-3.93	-3.59	-4.22	-3.62	-3.32
200	3	-4.35	-3.78	-3.47	-4.34	-3.78	-3.51
50	4	-4.94	-4.35	-4.02	-4.61	-3.98	-3.67
100	4	-4.75	-4.22	-3.89	-4.61	-4.02	-3.71
200	4	-4.70	-4.18	-3.89	-4.72	-4.13	-3.83
50	5	-5.41	-4.76	-4.42	-4.80	-4.15	-3.85
100	5	-5.18	-4.58	-4.26	-4.98	-4.36	-4.06
200	5	-5.02	-4.48	-4.18	-4.97	-4.43	-4.14

- The critical values are all negative as we are performing a lower-tailed test with rejection region on the left side.
- They depend on the length of the series N and the number of coefficients estimated in the model.

由于我们进行的是左尾检验，因此所有临界值为负数，拒绝区域在左侧。

Example: Short- vs Long-Term Interest Rates

1. Determine order of integration of each series

```
ndiffs(tb3ms))
```

```
ndiffs(tb10ys))
```

```
#> [1] 1
```

```
#> [1] 1
```

2. Perform regression on levels of the series

```
model <- tslm(tb10ys~tb3ms)
```

3. Perform ADF unit root test on residuals of the model

```
resid(model) %>% ur.df(type="none",selectlags="AIC")
```

```
#> Augmented Dickey-Fuller Unit Root/ Cointegration test
```

```
#> The value of the test statistics is -3.8371
```

4. Compare with Engle-Granger table with N=234, K=2

Since test statistics < Critical Value=-3.25, we reject H0 at 5% significance level and conclude the series are integrated.

Summary

- If y_t and x_t are:
 1. Stationary $I(0)$ there are no issues with OLS (if the classical assumptions are satisfied).
 2. $I(1)$ and are not cointegrated, we have the spurious regression problem. Regress in first differences.
 3. $I(1)$ and are cointegrated, the OLS estimator of β_1 is super-consistent (i.e. converges to the true value at a faster rate) while that for β_0 behaves as in the usual case. We can regress in levels.



Class 7
Pre-class Prep 6/8



ECMs Working with Non Stationary Data

Error Correction Models

传统的时间序列模型（如ARIMAX、VAR）要求对数据进行差分以使其平稳。

非线性方法（如LSTM神经网络）在自动检测数据中的趋势方面有混合的结果。

为了使数据平稳，通常会对其进行差分处理，这样可能会丢失一些重要的水平信息。

当处理具有协整关系的变量时，不需要差分，并且可以获得更好的结果。

这种情况下应该使用误差修正模型（ECM），因为它能够捕捉变量之间的长期平衡关系。

1. Classical models we have built so far (ARIMAX, VAR) **require us to make data stationary before modelling**
2. Non-linear methods we will overview (feed forward and recurrent neural networks via LSTM) have **mixed record in detecting trend in the data automatically**
3. Typically, we **take differences to render data stationary**. This removes levels information, and is therefore at some cost. There may be information embedded in the level
4. There is special case where we can model data with unit roots directly, to obtain (generally) superior results to cases where we need to take differences. This is **when variables being modelled are cointegrated**
5. In the special case when variables are cointegrated, we **should use an Error Correction Model (ECM)**

Introduction to ECMs and VECMs

ECM允许我们直接建模非平稳数据，而无需预差分，前提是模型中的变量具有协整关系。

ECM的两种主要形式：

Engle Granger ECM: 仅限于一个解释变量的情况，主要用于预测y变量。

Johansen VECM: 可以处理多个解释变量和多个因变量。

本节课程的重点是Johansen VECM，因为它的表达能力更强，适用于更复杂的模型。

- ECMs allow us to **model nonstationary data directly without differencing**
 - Most commonly, **we can model a $I(1)$ Y variable directly using $I(1)$ x variable without predifferencing**
 - Only requirement is **that x variable be cointegrated with y variable**
- There are two main forms of ECMs:
 - **Engle Granger ECM.** This version admits only a single x variable to forecast the y variable
 - **Johansen VECM.** This implementation is able to model multiple y variables with multiple x variables
- Code implementation for this class **will focus on Johansen VECM.** This is because the VECM is a superset of ECM, and you can always express a 2 variable system using the multiple variable syntax easily

ECM的结构公式包含两个部分：长期平衡部分和短期过渡部分。

长期平衡部分：解释变量的变化如何永久性地影响因变量。

短期过渡部分：描述因变量偏离长期平衡状态后如何回归长期平衡。

这种模型可以捕捉变量之间的动态调整机制，即如果变量暂时偏离平衡状态，模型会预测变量逐渐回归到平衡。

ECM definition (1 variable)

Consider following structure where we wish to forecast Y_t with x_t

$$\Delta Y_t = \gamma + B_1 \Delta X_t + \alpha(Y_{t-1} - B_0 - B_1 X_{t-1}) + v_t$$



- Two segments to equation above.
- **Equilibrium** (long run) term relates the change in Y with change in X .
E.g. a 1 unit change in X results in B_1 units permanent change in Y
- **Transient** term relates the change in y with deviation from long run relationship between Y and X from previous period

Error Correction Model (ECM) definition with 1 variable

Consider following structure where we wish to forecast Y_t with x_t

- $\Delta Y_t = \gamma + B_1 \Delta X_t + \alpha(Y_{t-1} - B_0 - B_1 X_{t-1}) + v_t$



公式的短期过渡部分关注变量如何围绕长期关系波动。
长期平衡由OLS公式 $Y_t = B_0 + B_1 X_t + e_t$ 给出，误差项应为白噪声。

- Focusing on transient relationship part of the ECM:

- We hypothesize long run relationship is given by OLS equation: $Y_t = B_0 + B_1 X_t + e_t$

- $Y_t - B_0 - B_1 X_t = e_t$ (white noise)

如果两个变量暂时出现较大的偏离，ECM模型会预测这种偏离随着时间的推移逐渐缩小。

例如，若 $(Y_{t-1} - (B_0 - B_1 X_{t-1}))$ 过高，模型会预测当前时期的 ΔY 值较低，以拉回长期趋势。

- In practice, gap $(Y_t - B_0 - B_1 X_t)$ may temporarily exhibit large values

- If gap temporarily widens, it should narrow over time and revert back to 0

- For example, if Y_{t-1} is abnormally high in period $t-1$, such that $Y_{t-1} - B_0 - B_1 X_{t-1} = C > 0$, we will predict ΔY will be lower by αC in current time period ($\alpha < 0$), to bring Y back in line with "long term" trends

Engle Granger ECM estimation with single X variable in R

检验两组变量是否具有相同的积分顺序 (I(1) 或 I(0))。
如果变量之间存在协整关系，使用OLS估计Y对X的最优预测模型。
如果没有协整关系，且变量是I(1)，则需要对差分后的数据使用OLS。
ECM是通过OLS估计协整关系后的最佳预测模型。

1. Given a pair of variables, X_t and Y_t , determine if **they have the same order of integration**
2. If so, estimate “trial” cointegrating relationship between the two variables using OLS
3. Test residuals from “trial” OLS estimation for a cointegrating relationship
4. Depending on output from above steps:
 1. If a cointegrating relationship exists between both variables, estimate an ECM as the “best” predictive model for Y in terms of X
 2. If no cointegrating relationship exists and **both variables are I(0)**, use simple OLS
 3. If no cointegrating relationship exists **and both variables are I(1) or above**, use OLS on the differenced variables
5. (Note: we are currently not discussing having lagged values of Y on the RHS. **This will be discussed under Vector ECMs**)

Recall Test for Cointegration

使用Engle Granger检验可以检验两组变量之间是否具有协整关系。

该检验过程包括三步:

进行静态OLS回归，并收集回归的残差。

使用ADF单位根检验残差是否具有单位根。

如果两个时间序列具有协整关系，则残差会被判断为平稳。

1. We can test two variables for cointegration using the **Engle Granger test**

2. **Engle Granger test** performs the following:
 - a) **Performs static OLS regression**, and collects residuals from this regression

 - b) Test residuals for existence of unit root using **Augmented Dickey Fuller (ADF) test**

 - c) If both time series are cointegrated, then the **residuals will be judged to be stationary as a result of the ADF test**



Class 7
Pre-class Prep 7/8

VECM with multiple variables

经典的时间序列模型如 ARIMAX 和 VAR 需要在建模之前将数据变为平稳 (stationary)。

非线性方法 (如神经网络) 在检测数据趋势方面表现不一。我们通常通过差分操作使数据平稳, 尽管这会丢失一些水平信息。当变量之间存在共整合关系时 (即长时间保持某种稳定的关系), 我们可以不进行差分, 直接建模。

这种情况下我们应该使用误差修正模型 (ECM)。

1. So far, we have considered cointegration **between one pair of variables only.**
 - a) With just two variables in the system, **it is only possible to have only a maximum of 1 cointegrating relationship**
 - b) Hence, if just studying a system with 2 variables, Engle Granger ECM is sufficient
2. What about **more general cases where we are studying relationship between N variables?**
3. Given an arbitrary group of N variables, there are a **maximum possible of N-1 linearly independent cointegrating relationships**

解释: 误差修正模型的核心思想是, 某些时间序列变量之间存在长期的平衡关系, 即使这些变量本身是非平稳的。通过共整合关系, ECM可以在不丢失长期信息的情况下进行建模, 避免传统方法中使用差分带来的问题。

VECM with multiple variables

要点总结:

ECM 允许我们在没有预先差分的情况下直接建模非平稳数据。

我们可以用 I(1) 的 X 变量来预测 I(1) 的 Y 变量，前提是这两个变量是共整合的。

ECM 有两种形式: Engle-Granger ECM (用于单一 X 变量) 和 Johansen VECM (用于多个 X 变量)。

本课程将重点关注 Johansen VECM，因为它是 ECM 的一个超集，能够处理多个变量的系统。

1. We will study **multiple cointegrating relationships** in the context of a VAR model of order p, with N variables being modelled
2. For example, we can **build a VAR with 3 variables** to model **consumption** (C_t), income (I_t) and savings (S_t). Order (number of lags) for the VAR will depend on the output from `VARselect()`
3. In this setting, there **might be up to two possible linear independent cointegrating relationships**, e.g
 - $C_t = B_0 + B_1 I_t + B_2 S_t + e_{1,t}$
 - $I_t = B_3 + e_{2,t}$

介绍 ECM 和 VECM 的应用场景。
Engle-Granger ECM 适用于处理两个变量的情况，而 Johansen VECM 则是更高级的方法，能够同时处理多个变量，适用于更复杂的系统。

VECM with multiple variables

公式描述了我们如何通过 X 来预测 Y ，公式分为两个部分：长期均衡部分和短期调整部分。长期部分反映了 Y 和 X 之间的长期关系，而短期部分反映了 Y 与 X 的短期偏差及其恢复到长期关系的过程。

1. In the VECM methodology, we **insert error correction terms to the VAR specification**. Primary steps in the methodology are:
2. Specify and **estimate a VAR(p) model for N variables**
3. **Determine number of cointegrating vectors** via maximum likelihood
4. **If there are 0 cointegrating vectors, we cannot build a VECM model.** Proceed to render all variables stationary, and build a VAR on the differenced variables (as per before)
5. **If there is at least 1 cointegrating vector:**
 - We can build a VECM
 - Also, we should build a VECM. In the presence of cointegration, model will be a better description of data compared to VAR on differenced variables
 - We can estimate VECM by maximum likelihood

这页的重点在于 ECM 公式的两个部分：长期的均衡部分和短期的调整部分。长期部分描述了 X 对 Y 的长期影响，而短期部分解释了当 X 和 Y 偏离长期均衡时如何进行调整。比如，如果 Y 偏离了长期关系，系统会试图逐步将其拉回到平衡状态。



Class 7
Pre-class Prep 8/8

什么是VECM？向量误差修正模型是用于处理多个时间序列变量之间长期均衡关系的工具。这些变量本身可能是非平稳的（即它们的平均和方差随时间变化），但它们之间可能存在某种长期的线性关系，这就是所谓的协整关系。

两个变量的情况：仅研究两个变量的情况时，最多只能有一个协整关系。这种情况下，使用Engle-Granger ECM（单变量误差修正模型）足够了，因为我们只需处理一个协整关系。

多个变量的情况：当研究N个变量时，最多可能存在N-1个线性独立的协整关系。也就是说，在这些变量中，可能有N-1个不同的长期均衡关系。这是由于每个变量可能对多个其他变量产生长期影响。

1. Consider **purchasing power parity (PPP) between 2 countries** under floating exchange rates
2. Purchasing power parity **states that same basket of goods should cost the same in different countries**
3. i.e. **price of basket in country A (P^1) = price of same basket in country B (P^2) exchange rate (X)**
4. In practice, **economists rely on price indexes whose market baskets differ across countries**, so the PPP equation needs a constant of proportionality to reflect this difference.
 - $X = A_0 P^1 / P^2$
 - $\ln(X) = \ln(A_0) + \ln(P^1) - \ln(P^2)$
5. Using small letters to represent logs of variables, **we could estimate a VECM (with 1 lag), and 1 cointegrating relationship**

Example: VECM as a model for PPP

在两个变量的简单系统中，Engle-Granger方法可以处理，但如果涉及更多变量时（例如，三个变量），就需要使用更复杂的Johansen方法来处理VECM。

VAR模型：向量自回归（VAR）模型是用于分析多个时间序列变量之间相互依赖关系的模型。它考虑了当前变量与其过去值以及其他变量的过去值之间的关系。

Example: VECM as a model for PPP

$$\Delta x_t = \beta_{x0} + \beta_{xx1}\Delta x_{t-1} + \beta_{x11}\Delta p_{t-1}^1 + \beta_{x21}\Delta p_{t-1}^2 + \lambda_x(x_{t-1} - \alpha_0 - p_{t-1}^1 + p_{t-1}^2) + \nu_t^x$$

$$\Delta p_t^1 = \beta_{10} + \beta_{1x1}\Delta x_{t-1} + \beta_{111}\Delta p_{t-1}^1 + \beta_{121}\Delta p_{t-1}^2 + \lambda_1(x_{t-1} - \alpha_0 - p_{t-1}^1 + p_{t-1}^2) + \nu_t^1$$

$$\Delta p_t^2 = \beta_{20} + \beta_{2x1}\Delta x_{t-1} + \beta_{211}\Delta p_{t-1}^1 + \beta_{221}\Delta p_{t-1}^2 + \lambda_2(x_{t-1} - \alpha_0 - p_{t-1}^2 + p_{t-1}^2) + \nu_t^2.$$

Source: Sims, Christopher A. 1980. Macroeconomics and Reality. *Econometrica* 48 (1): 1-48

当构建包含多个变量的VAR模型时，我们可以检测它们之间的多重协整关系。假设有三个变量：消费(C_t)、收入(I_t)和储蓄(S_t)，我们可以构建一个包含这三个变量的VAR模型，并通过Johansen方法检测协整关系。

1. If exchange rate is out of equilibrium, we will expect some adjustment back towards the long run equilibrium in the next period

2. Error coefficients λ_x , λ_1 and λ_2 measure these responses. | 例如，可能存在消费、收入和储蓄的两个独立的协整关系：
两个协整关系的例子：
第一个协整关系：消费由收入和储蓄决定
第二个协整关系：收入独立于其他变量
3. Using the above logic, we expect λ_x and λ_2 to be negative and λ_1 to be positive

Example: VECM as a model for PPP

1. Within PPP framework, **there may also be a second cointegrating relationship**
2. Suppose country 1 is on gold standard. **Price level in this country would be constant in the long run**
3. i.e. $p_1 = \alpha_1$. This would be second cointegrating relationship

购买力平价（PPP）**框架下协整关系的讨论，特别是如何将协整向量应用于PPP模型中。

1. PPP框架中的第二个协整关系：在购买力平价的框架下，除了第一个协整关系外，还可能存在第二个协整关系。第一个协整关系一般是基于价格和汇率的关系，但这页提到的第二个协整关系则可能基于其他经济变量。
2. 假设某国（例如国家1）采用金本位制：如果一个国家使用金本位制（即以黄金为基础的货币制度），该国家的物价水平将会在长期内保持不变。金本位制限制了一个国家货币供应的增长，从而在很大程度上维持价格的长期稳定。
3. 第二个协整关系的表达式：此处的 $p_1 = \alpha_1$ 表示在这种情形下，国家1的物价水平是常数。这个常数 α_1 可能代表某种固定的价格水平或长期均衡状态，这是另一个潜在的协整关系。

Example: VECM as a model for PPP

包含两个协整关系的VECM模型：这里的VECM模型用三个变量（汇率 X_t 和两个价格变量构建，同时包含两个协整关系。公式展示了每个变量的动态调整过程：

- VECM incorporating both cointegrating relationships:

$$\begin{aligned}\Delta x_t &= \beta_{x0} + \beta_{xx1}\Delta x_{t-1} + \beta_{x11}\Delta p_{t-1}^1 + \beta_{x21}\Delta p_{t-1}^2 + \lambda_x(x_{t-1} - \alpha_0 - p_{t-1}^1 + p_{t-1}^2) + \mu_x(p_{t-1}^1 - \alpha_1) + \nu_t^x \\ \Delta p_t^1 &= \beta_{10} + \beta_{1x1}\Delta x_{t-1} + \beta_{111}\Delta p_{t-1}^1 + \beta_{121}\Delta p_{t-1}^2 + \lambda_1(x_{t-1} - \alpha_0 - p_{t-1}^1 + p_{t-1}^2) + \mu_1(p_{t-1}^1 - \alpha_1) + \nu_t^1 \\ \Delta p_t^2 &= \beta_{20} + \beta_{2x1}\Delta x_{t-1} + \beta_{211}\Delta p_{t-1}^1 + \beta_{221}\Delta p_{t-1}^2 + \lambda_2(x_{t-1} - \alpha_0 - p_{t-1}^2 + p_{t-1}^1) + \mu_2(p_{t-1}^2 - \alpha_1) + \nu_t^2\end{aligned}$$

Source: Sims, Christopher A. 1980. Macroeconomics and Reality. *Econometrica* 48 (1): 1- 48

每个变量的变化率（例如 Δx_t ）由其滞后期的变化率（例如 Δx_{t-1} ）、其他变量的变化率以及两个协整误差项来决定。

- This would be VECM with 3 variables and 2 cointegrating relationships, of order 1
每个协整误差项 (λ_x, μ_1, μ_2) 用于衡量系统偏离长期均衡的程度，并帮助模型在长期回归均衡状态。
- If we wanted to increase order of VECM, we can simply include more lagged differences in equilibrium portion of VECM

三个变量和两个协整关系：在这个例子中，VECM模型便用了三个变量（例如汇率和两个价格变量），并有两个协整关系。协整关系说明了三个变量之间的长期关系。如果模型是阶数为1的VECM模型，意味着它考虑了滞后一期的变量之间的相互影响。
增加VECM阶数：如果我们希望增加VECM模型的阶数，可以通过加入更多滞后期的变量差分值来实现。更高阶的VECM模型能捕捉到变量之间更复杂的动态关系，这对于实际建模中更长的时间滞后可能有用。

VECM in R

The screenshot shows the RStudio interface. The 'Console' pane on the left displays R code and its output. The 'Help' pane on the right provides detailed information about the 'zeroYld' dataset.

Console Output:

```
>
> ?zeroYld
> head(zeroYld)
   short.run long.run
1     2.183    1.575
2     2.246    1.545
3     2.308    1.762
4     2.401    1.773
5     2.558    1.808
6     2.507    1.810
```

Help Pane (R: zeroYld time series):

Description:
U.S. Term Structure Data, 1951-1991. Dataset used by Hansen and Seo (2002). The data contains the 12 month short rate and 120 month long rate.

Usage:
zeroYld
zeroYldMeta

Format:
zeroYld contains two variables, while zeroYldMeta contains also Year and Month columns.
zeroYld is a data frame with 482 observations and 2 variables:
short.run numeric Short term, 12 month
long.run numeric Long term, 120 month

Source:
Hansen, B. and Seo, B. (2002), Testing for two-regime threshold cointegration in vector error-correction models, Journal of Econometrics, 110, pages 293 - 318
The data can be downloaded from: http://www.ssc.wisc.edu/~bhansen/progs/joe_02r.zip.
The authors themselves took the data from the webpage of Huston McCulloch: <https://www.asc.ohio-state.edu/mcculloch.2/ts/mcckwon/mccull.htm>

1. zeroYld数据集:
数据集来源于Hansen和Seo (2002年) 的研究, 专门用于测试两个时期的门槛协整关系 (Two-regime Threshold Cointegration in Vector Error-Correction Models) 。
数据集包含482个观测值, 有两个变量: short.run (12个月短期利率) 和long.run (120个月长期利率) 。

2.R中的数据格式:
zeroYld是一个数据框, 包含两个变量: short.run和long.run。
数据集中还附有元数据zeroYldMeta, 其中包含年份和月份等额外信息。

3.R语言中的head()函数:
使用head(zeroYld)来查看数据集的前6行数据, 数据表明, 短期利率和长期利率在不同年份和月份的值。

例如, 第一行数据显示, 短期利率为2.183%, 长期利率为1.575%。

4.VECM在R中的应用:
可以利用R中的tsDyn包来进行VECM建模, 输入数据如zeroYld这样的时间序列数据。
在后续的步骤中, VECM将帮助我们分析这两个利率之间的长期协整关系, 以及短期的动态调整过程。

1. Johansen协整检验: Johansen检验是一种用来确定多个时间序列变量之间是否存在协整关系的统计检验。在这一示例中, 使用了ca.jo()函数来对zeroYld数据进行Johansen检验, 并显示了输出结果。

2.解释Johansen检验的输出:

Test Type: 该检验类型为最大特征值统计量 (Maximal Eigenvalue Statistic), 并假设存在线性趋势。
 特征值 (Eigenvalues): 输出的特征值为0.073和0.0056。这些特征值是Johansen检验用来确定协整关系数量的关键指标。
 检验统计量和临界值 (Test Statistic and Critical Values): 检验统计量的值与10%、5%、1%的临界值比较, 以确定是否拒绝原假设 (无协整关系)。例如, 对于 $r \leq 1$ (假设最多有1个协整关系), 检验统计量为2.71, 小于5%的临界值6.50, 因此不能拒绝原假设。
 结果解释: 如果 $r = 0$, 表示没有协整关系; 如果 $r = 1$, 表示存在一个协整关系。在这个例子中, $r = 1$ 意味着有一个协整关系。

Determining number of cointegrating relationships

```
Console Terminal × Jobs ×
~/

> summary(ca.jo(zeroYld))

#####
# Johansen-Procedure #
#####

Test type: maximal eigenvalue statistic (lambda max) , with linear trend

Eigenvalues (lambda):
[1] 0.073463811 0.005621199

values of teststatistic and critical values of test:

      test 10pct 5pct 1pct
r <= 1 | 2.71 6.50 8.18 11.65
r = 0 | 36.63 12.91 14.90 19.19

Eigenvectors, normalised to first column:
(These are the cointegration relations)

      short.run.12 long.run.12
short.run.12   1.000000  1.0000000
long.run.12   -1.022065 -0.2409262

weights w:
(This is the loading matrix)

      short.run.12 long.run.12
short.run.d -0.01161738 -0.009806832
long.run.d   0.08880969 -0.014155748

> |
```

In case of zeroYld, there are only 2 variables

Hence, we either have 0, or 1 cointegrating relationships. Not possible to have more

Johansen procedure produces test statistic for both cases

We always chose smallest number of cointegrating relationships where we fail to reject the null (regardless of how many variables). In this case, $r = 1$

Note if we can reject null for all values of R, all data series appear to be stationary, so we can just build a VAR on the raw data

If $r = 0$, there are no cointegrating relationships, and we build a VAR on the differenced variables

3.协整关系的数量选择:

zeroYld数据集的情况: 因为zeroYld数据集中只有两个变量, 因此最多只有一个协整关系。

Johansen检验结果: 我们选择最小的协整关系数量, 在本例中是 $r = 1$ 。

如果 $r = 0$: 如果数据没有协整关系 ($r = 0$), 我们会基于差分后的数据构建VAR模型 (向量自回归模型)。

如果 $r = 1$: 如果存在1个协整关系, 则可以构建VECM模型。

4.协整向量 (Eigenvectors) 和权重 (Weights):

Johansen检验还输出了协整向量和权重矩阵, 这些数值可以帮助我们理解短期动态和长期均衡关系。

1. 执行Johansen检验:

在这页中，针对另一个数据集进行了Johansen协整检验。这个检验的目标是检测多个时间序列变量之间是否存在协整关系。

Johansen检验原假设：原假设是没有协整关系。我们希望通过该检验拒绝原假设，从而得出数据集中的协整关系。

2. 拒绝原假设:

Johansen检验输出显示我们为所有R值都拒绝了原假设，这意味着这个数据集不存在协整关系。

解释结果：由于拒绝了协整关系的原假设，说明数据本身是平稳的，或者经过适当的变换后，数据达到了平稳状态。

Determining number of cointegrating relationships

1. Running Johansen test on another dataset, **we reject the null hypothesis for all values of R**

2. In this case, **it appears data is stationary**

```
#####
# Johansen-Procedure #
#####

Test type: trace statistic , with linear trend

Eigenvalues (lambda):
[1] 0.338903321 0.330365610 0.001431603

Values of teststatistic and critical values of test:

          test 10pct 5pct 1pct
r <= 2 | 14.32 6.50 8.18 11.65
r <= 1 | 4023.76 15.66 17.95 23.52
r = 0  | 8161.48 28.71 31.52 37.22

Eigenvectors, normalised to first column:
(These are the cointegration relations)

      p.l2      q.l2      r.l2
p.l2  1.000000  1.0000000  1.000000
q.l2  1.791324 -0.52269002 1.941449
r.l2 -1.717271  0.01589134 2.750312

Weights W:
(This is the loading matrix)

      p.l2      q.l2      r.l2
p.d -0.1381095 -0.771055116 -0.0003442470
q.d -0.2615348  0.404161806 -0.0006863351
r.d  0.2439540 -0.006556227 -0.0009068179
```

3. 数据平稳性 (Stationarity) :

检验表明该数据看起来是平稳的 (stationary)。这意味着时间序列数据的统计特性（例如均值、方差）随时间保持恒定，不需要进一步差分处理。

4. 特征值与检验统计量:

特征值 (Eigenvalues)：输出中包含了特征值的结果，这些特征值帮助我们确定协整关系的数量。

检验统计量和临界值：这些统计值 (trace statistic) 与给定的临界值 (critical values) 比较。对于每个R值的检验统计量，如果超过临界值，就意味着拒绝无协整关系的假设。

例如， $r \leq 2$ 时检验统计量为 14.32，低于 10% 的临界值 14.32，因此我们拒绝原假设，表明该数据集不具备协整关系。

5. 协整向量和权重矩阵:

Johansen检验还输出了协整向量和权重矩阵，显示了变量之间的协整关系及其影响。这些信息有助于进一步分析短期动态和长期均衡。

Summary on number of cointegrating relationships

In a situation with **N endogenous variables**:

1. $r = 0$: build VAR model on differenced data
2. $0 < r \leq N-1$: build VECM
3. $r > N-1$: build VAR model on data in levels
(i.e., raw data, no need differencing)

1. $r = 0$: 差分数据的VAR模型:

当协整关系数量 r 等于0时, 表示变量之间不存在长期平衡关系。在这种情况下, 我们需要对数据进行差分处理, 使其成为平稳时间序列(即消除趋势或非平稳性)。

在这种情况下, 我们可以构建一个基于差分数据的VAR(向量自回归)模型, 来解释多个时间序列变量的短期动态。

2. $0 < r \leq N-1$: 构建VECM(向量误差修正模型):

当协整关系数量 r 大于0且小于等于 $N-1$ 时, 表示存在协整关系。此时可以使用VECM(向量误差修正模型)来建模。

VECM结合了短期动态与长期均衡, 模型中的误差修正项用于调整偏离长期均衡的误差, 并将系统拉回到长期趋势。

3. $r > N-1$: 使用水平数据的VAR模型:

如果协整关系的数量 r 大于 $N-1$, 则变量的协整关系超出了可接受范围, 说明所有变量是平稳的, 不需要对数据进行差分处理。

在这种情况下, 我们可以使用原始数据(不做差分处理)构建VAR模型, 因为数据本身已经是平稳的, 可以直接建模来解释它们的动态行为。

We determine number of VAR lags in same way as before

```
> VARselect(zeroYd)
$selection
AIC(n)  HQ(n)  SC(n)  FPE(n)
  9      2      2      9

$criteria
   1       2       3       4       5       6       7
AIC(n) -4.4004108 -4.46118997 -4.45201909 -4.44295549 -4.45035203 -4.46010768 -4.46243423
HQ(n)  -4.3796247 -4.42654649 -4.40351821 -4.38059722 -4.37413636 -4.37003462 -4.35850377
SC(n)  -4.3475678 -4.37311838 -4.32871886 -4.28442663 -4.25659454 -4.23112155 -4.19821946
FPE(n) 0.0122723 0.01154863 0.01165506 0.01176124 0.01167465 0.01156144 0.01153475
   8       9       10
AIC(n) -4.47687527 -4.49562700 -4.49204621
HQ(n)  -4.35908742 -4.36398176 -4.34654357
SC(n)  -4.17743186 -4.16095496 -4.12214554
FPE(n) 0.01136959 0.01115865 0.01119902
```

Number of VAR lag will be order of VECM model

如何确定VECM（向量误差修正模型）中的VAR（向量自回归模型）滞后阶数。：

1.VAR滞后阶数选择：

在建模VECM之前，需要确定VAR模型的滞后阶数。滞后阶数代表了模型中包含的过去观测值的数量。

滞后阶数可以通过信息准则（Information Criteria）进行选择，如AIC（赤池信息准则）、HQ（Hannan-Quinn准则）、SC（Schwarz准则，也叫BIC，贝叶斯信息准则）以及FPE（最终预测误差）。

在PPT的代码输出部分，使用VARselect()函数从不同的准则中选择最优滞后阶数。各准则的结果显示在输出表格中。

2.选择标准：

AIC、HQ、SC、FPE是不同的标准，每个标准基于不同的理论和假设来计算滞后阶数的选择。通常根据研究的具体需求，选择其中一个准则。

例如，AIC准则倾向于选择较大的滞后阶数，而SC（或BIC）更倾向于选择较小的滞后阶数。

在表格的最后一行，每个准则对应了不同的滞后阶数。在本例中，AIC和FPE选择了滞后阶数9，HQ和SC选择了滞后阶数2。

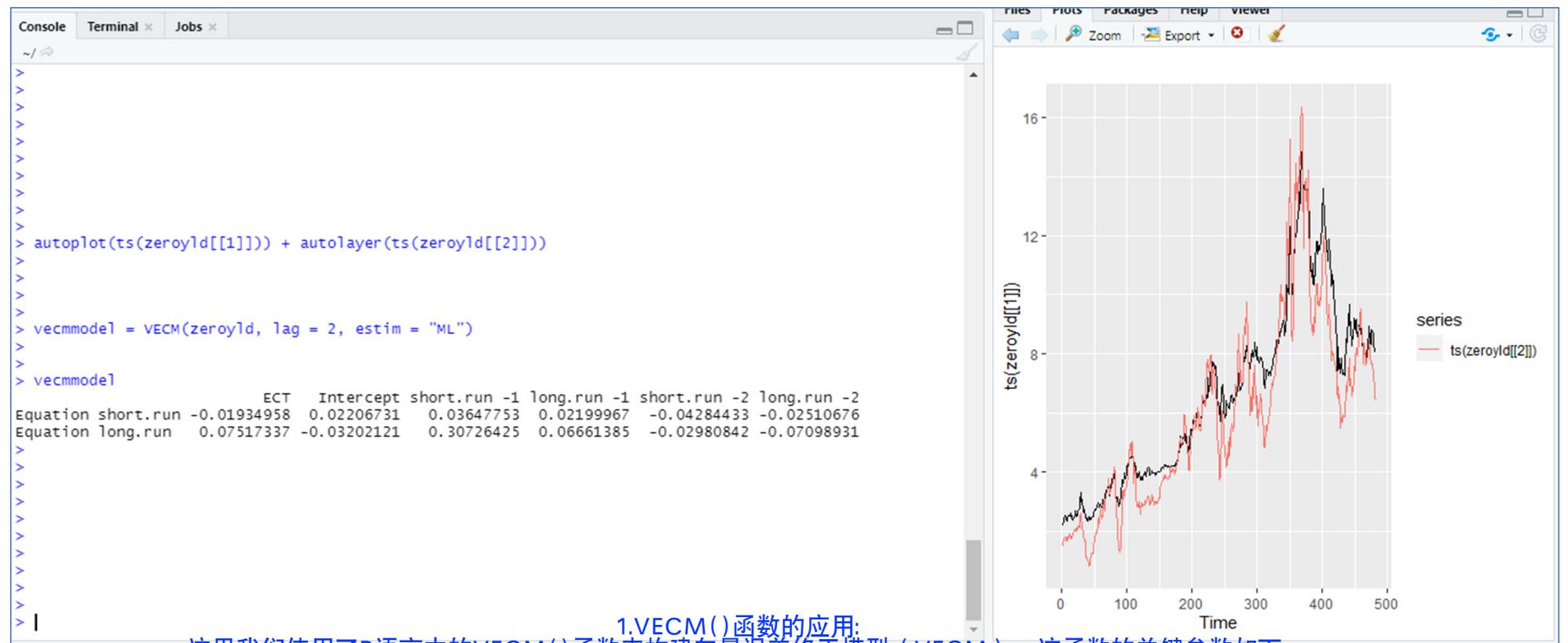
VECM模型的滞后阶数：

确定VAR模型的滞后阶数之后，该滞后阶数通常作为VECM模型的滞后阶数。滞后阶数决定了在模型中要考虑多少个过去的时间点信息。

3.总结：

为了建立一个VECM模型，首先需要选择VAR模型的滞后阶数。通过使用信息准则（AIC、HQ、SC、FPE），我们可以根据数据选择一个合适的滞后阶数。然后，这个滞后阶数将被用于构建VECM模型，确保模型能捕捉到时间序列数据的历史依赖性。

Using output from previous 2 functions, we build VECM model with lag = 2 and 1 cointegrating relationship



1. VECM()函数的应用:
这里我们使用了R语言中的VECM()函数来构建向量误差修正模型（VECM）。该函数的关键参数如下:

zeroYld: 表示用于建模的时间序列数据。

lag = 2: 表示选择了滞后2期的数据作为模型中的解释变量。

estim = "ML": 表示使用最大似然估计法（Maximum Likelihood）来估计模型参数。

2. 协整关系:

该模型中包含1个协整关系，这是根据前面选择滞后阶数和Johansen协整检验的结果得出的。

协整关系代表了多个时间序列之间长期的平衡关系。即使这些序列在短期内可能会有偏离，但它们在长期内会保持一个稳定的关系。

3. 模型的输出解释:

模型输出了短期和长期变量的系数（如short.run, long.run），每个变量的系数包括它们滞后1期和滞后2期的数值。

ECT代表误差修正项（Error Correction Term），它反映了系统偏离长期平衡时的调整速度。负的ECT系数表示系统在每期会逐渐回归到长期均衡状态。

4. 时间序列图:

在右边的图中，我们可以看到两个时间序列的轨迹。图中黑线和红线分别代表了两个不同的变量（如短期和长期利率）的时间变化。可以观察到，这两个序列虽然在短期内有波动，但大体上显示出协整关系，即它们的趋势是相关的。

如何在R中利用之前步骤构建VECM模型:

构建VECM: 使用R中的VECM函数, 设定滞后阶数和协整关系的数量来构建VECM模型。

结果解释: 模型输出了误差修正项、截距项以及短期和长期动态关系的系数。

总结: 这页展示了如何结合滞后阶数和协整关系数量的选择, 最终在R中构建并运行VECM模型。

To forecast with
an VECM model,
we use predict()
instead of
forecast()

```
> predict(vecmmodel, n.ahead = 10)
   short.run long.run
483  8.068324 6.531213
484  8.067256 6.620864
485  8.062725 6.688576
486  8.059285 6.749017
487  8.057677 6.806098
488  8.057351 6.859525
489  8.058052 6.909184
490  8.059707 6.955393
491  8.062247 6.998499
492  8.065595 7.038789
>
>
```