# 第十一章时间序列分析方法

授课教师: 吴翔 wuhsiang@hust.edu.cn

OCT 12 - 15, 2020

- 🚺 时间序列分析概述 (2 个课时)
- 2)时间序列经典分析方法(3 个课时)
- 3 时间序列案例分析(1 个课时)
- 4 时间序列分析实习 (2 个课时)

# 时间序列分析概述 (2 个课时)

#### 课程存储地址

- 课程存储地址: https://github.com/wuhsiang/Courses
- 资源:课件、案例数据及代码



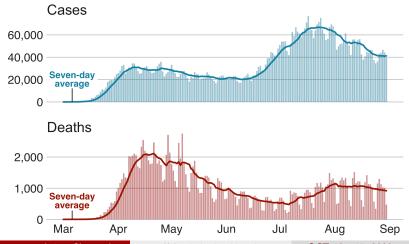
图 1: 课程存储地址

# 参考教材

- James D. Hamilton 著. 时间序列分析 (2 册). 北京: 人 民卫生出版社, 2015.
- Jonathan D. Cryer & Kung-Sik Chan 著. 时间序列分析及应用 (R 语言) (原书第 2 版). 北京: 机械工业出版社. 2011.
- Robert I. Kabacoff 著. R 语言实战 (第二版). 北京: 人 民邮电出版社. 2016.
- David Salsburg 著. 女士品茶: 统计学如何变革了科学和
   生活. 南昌: 江西人民出版社. 2016.

### 时间序列案例:传染病

#### Number of daily cases and deaths in the US



### 本节知识点

- 时间序列分析方法起源
- 时间序列基本概念
- 时间序列分析要素
- 时间序列分析建模

# 11.1 时间序列分析方法起源

时间序列分析的方法,起源于英国统计学家 Ronald Fisher 在 英国洛桑试验站 (1919-1933) 的**农作物收成变动研究**。这一研 究产生了统计史上的三个重要方法:

- 时间序列分析思想
- 随机对照实验
- 方差分析

我们通过回顾这一研究,以深入理解时间序列分析的思想。

# Fisher 与洛桑试验站

英国**洛桑试验站 (Rothamsted Experimental Station)**, 现为洛桑研究所 (Rothamsted Research), 在漫长的历史中 (1843-1919) 积累了大量的"实验数据"和其它记录,包括:

- 降水量和温度的每日精确记录
- 施肥量与土壤检测数据的每周记录
- 农作物收成的每年记录
- 人造肥料和不同农作物(小麦、黑麦、大麦、马铃薯等)的组合实验方案

# 洛桑试验站的农作物



图 3: 洛桑试验站

# 收成变动研究

Fisher 考虑特定年份 (t) 的特定田地 (i) 上的农作物产量  $(Y_{it})$ ,并试图回答以下问题:

- 降水量对小麦产量有何影响?
- 不同肥料对不同品种的马铃薯产量有何影响?

课堂讨论:第一个问题 (5min)

## 如何预测小麦产量?

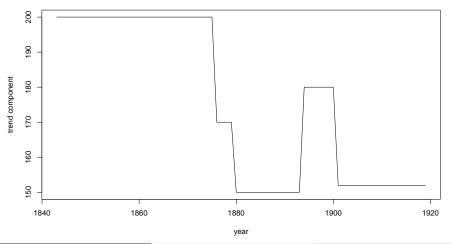
选定一块田地(名曰:"宽埂"),这块田地**只使用过**动物粪便作为肥料,考虑**降水量如何影响小麦产量**。

- 降水量和小麦均是**时变**的,即考虑不同年份(t)的降水量  $(X_t)$  和小麦产量  $(Y_t)$ 。
- ullet 探讨降水量  $X_t$  与小麦产量  $Y_t$  的关系,即属于**时间序列分析**范畴。
- ullet 问题演变为,如何**预测**小麦产量  $Y_t$ ?

# 小麦产量的影响因素分解

- 土壤退化导致的产量总体稳步减小
- 长期缓慢变化,每个变化周期为期数年
- 不同年份的气候因素 (例如降水量) 导致的变化

# 长期缓慢的变化



# 理解长期而缓慢的变化

- 难以解释的长期缓慢变化:产量从 1876 年开始剧烈下降,从 1880 年开始下降更加剧烈,至 1894 年产量开始改善,从 1901 年开始又剧烈下降
- 小麦田地中杂草的生长情况与之相反
- 最终解释: 1876 年前,人们雇佣小男孩到田里除草 →> 1876 年《教育法》规定适龄儿童必须上学→>1880 年法律 处罚不让适龄儿童上学的家庭→>1894 年洛桑附近的女子寄 宿学校校长认为,高强度户外运动有助于儿童健康→>1901 年校长去世

# 肥料使用对马铃薯产量的影响研究

- 洛桑试验站的早期方案:在不同片田地上(或者 Fisher 的田地分块),针对不同马铃薯品种,使用不同化肥,并记录其产量
- 潜在问题
  - 田地相关的混淆因素,例如土壤、排水方式、营养物质、杂草等
  - 年份相关的混淆因素,例如气候变化等

# Fisher 开创的统计方法

- 随机对照实验
  - 每片田地分块之后, 进一步把每块田地分为若干排
  - 采用随机化方案,对每块地的每一排实施不同的处理
- 方差分析

# 分解思想

Fisher 在洛桑试验站研究农作物收成变动时,采用了**分解** (decomposition) 的思想:

- 时间序列中的要素分解
- 方差分析中的效应分解

# 11.1.2 时间序列基本概念

#### 时间序列 (time series)

- ullet 定义:一组在特定时刻的观测值  $Y_t$  ,例如特定田地上的小麦收成
- 领域: 广泛存在于宏观经济、金融财务以及医疗领域

#### • 时间序列分析 (time series analysis)

- 数据:时间序列数据,与横截面数据、面板数据,为三类主要的观测数据类型
- 分析方法:通常基于宏观经济学理论建模,并被视作宏观计量 经济学的主要方法

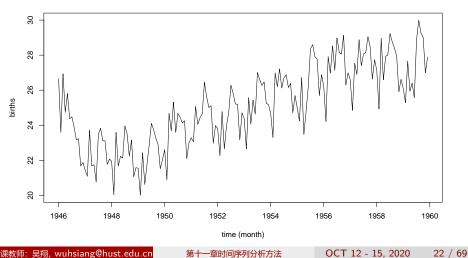
# 横截面 vs 时间序列数据

- 横截面数据 (cross sectional data)
  - 定义:不同研究对象在某一时点的变量观测数据
  - 例子: 通常的问卷调查
- 时间序列数据 (time series data)
  - 定义: 也称为纵剖面 (longitudinal sectional) 数据,是某一研究对象在不同**时点**的变量观测数据
  - 例子: 股票价格数据

## 面板数据

- 定义:不同研究对象在不同时点的变量观测数据
- 特征: 具有"横截面"和"时间序列"两个维度
- 例子: 中国健康与养老追踪调查 (China Health and Retirement Longitudinal Study, CHARLS)、中国健康 与营养调查 (China Health and Nutrition Survey, CHNS) 等

# 时间序列数据



# 11.1.3 时间序列分析要素

影响时间序列观测值的因素,可以分为以下几类:

- 趋势因素 (trend component): 观测值的长期的趋势, 通常是非线性的
- ② 循环因素 (cyclical component): 非季节因素引起的波动, 通常也被归入趋势因素中
- ③ 季节因素 (seasonal component) : 在一定时期内呈现的规律变化,例如一年内随着自然季节的更替而发生的变化
- **◎ 不规则因素 (irregular component)** : 诸如随机因素

通常将趋势因素和循环因素合并在一起考虑,成为趋势-循环因素 (trend-cycle),或简称**趋势因素**。

#### 课堂讨论

讨论以下情境中,趋势因素 (trend component) 和季节因素 (seasonal component) 的**含义**是什么

- 京东 & 淘宝上某一产品的销售量
- 某一医院在过去十年的每日门诊病人数量

### 11.1.4 时间序列建模

时间序列可以分解为趋势因素、季节因素和随机因素

$$Y_t = f(T_t, S_t, E_t)$$

常用的函数类型  $f(\cdot)$  有两种:累加、累乘。

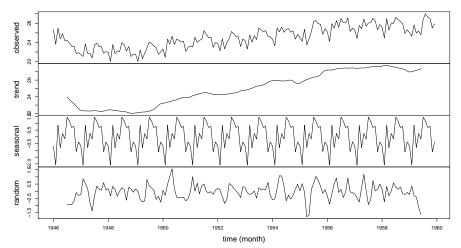
### 加法模型

假定趋势因素、季节因素和随机因素**相互独立**,则可以用**加法模型 (additive model)**来分解各个因素

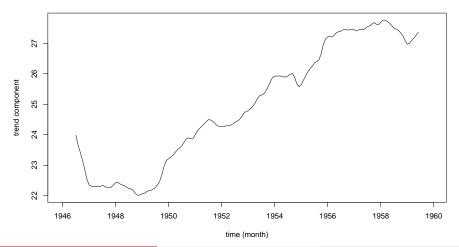
$$Y_t = T_t + S_t + E_t.$$

# 因素分解:加法模型

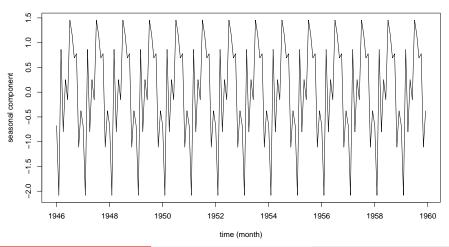
#### Decomposition of additive time series



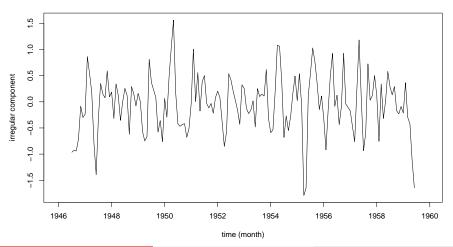
# 加法模型之趋势因素



# 加法模型之季节因素



# 加法模型之随机因素



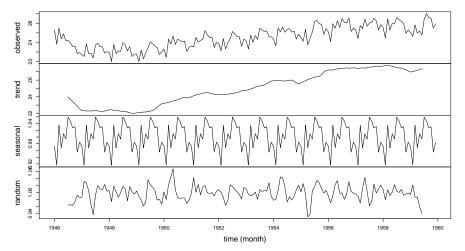
# 乘法模型

趋势因素、季节因素和随机因素不满足相互独立的条件,则可以用**乘法模型 (multiplicative model)** 来分解各个因素

$$Y_t = T_t \times S_t \times E_t$$
.

## 因素分解:乘法模型

#### Decomposition of multiplicative time series



# 时间序列分析步骤

- 搜集数据,绘制时间序列图
- ② 因素分解,得到趋势因素、季节因素和随机因素
- 对特定情境建模,分别预测趋势因素和季节因素的时间序列值
- 获得最终的预测模型

# 时间序列经典分析方法 (3 个课时)

### 本节知识点

- 预测方法与精度
- 移动平均法
- 指数平滑法
- 生长曲线法
- 灰色系统预测法(略)

# 11.2.1 预测方法与精度

#### 参数方法 vs 非参数方法

#### • 参数方法

- 分析问题情境
- 设定统计模型
- 估计参数并给出预测值

#### • 非参数方法

• 不依赖于具体的统计模型

## 衡量预测精度

定义预测误差  $e_t$  为实际观测值  $Y_t$  与预测值  $Y_t$  之差,

$$e_t = Y_t - \hat{Y}_t.$$

由此得到两个衡量预测精度的指标,均方误差 (mean square error, MSE) 和平均绝对离差 (mean absolute deviation, MAD)

$$\mathsf{MSE} = \frac{\sum_{t=1}^n e_t^2}{n}, \mathsf{MAD} = \frac{\sum_{t=1}^n |e_t|}{n}.$$

#### 11.2.2 移动平均法

#### 移动平均 (moving average)

- 思路: 计算最近 m 个连续观测值的平均值,作为时间序列的 预测值
- 假设: (1) 趋势因素是线性的; (2) 不规则因素有明确的节 奉波动模式

## 移动平均法:模型

预测  $Y_{t+1}$  的算法为,

$$Y_{t+1} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-m+1}}{m}$$

而可以通过**最小化误差**选取适当的 m 值。具体步骤,可以使用 R 中的smooth包来完成。

## Time elapsed: 1.05 seconds

# 移动平均法: 案例

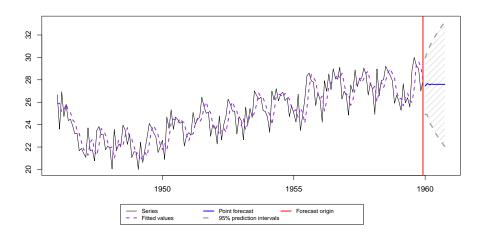
```
ma.ts \leftarrow sma(births, h = 20)
ma.ts
```

```
## Model estimated: SMA(2)
## Initial values were produced using backcast
##
```

## Loss function type: MSE; Loss function valu ## Error standard deviation: 1.3

## Sample size: 168 ## Number of estimated parameters: 2

## 移动平均法:案例(续)



## 移动平均法: 其它实践

除了简单移动平均法以外,还有一些改进的预测方法:

- 加权移动平均法:给予近期数据更大的权重,但维持权系数  $\sum_{\tau=1}^m w_\tau = 1.$
- 趋势移动平均法:同时使用一次和二次移动平均法。

#### 11.2.3 指数平滑法

#### 指数平滑法 (exponential smoothing)

- 思想:介于全期平均法和移动平均法之间。(1)使用全期数据,而非部分近期数据;(2)给予近期数据更大的权重,而远期数据的权重则呈指数衰减
- 假设:时间序列的态势具有稳定性或规则性,因而过去的态势会在某种程度上持续到未来

## 指数平滑法: 模型

定义初始平滑值  $S_0=Y_0$ ,则指数平滑模型可以表示为:

• 任一期 t 的平滑值  $S_t$ ,均是本期观测值  $Y_t$  和上期平滑值  $S_{t-1}$  的加权平均

$$S_t = \alpha Y_t + (1 - \alpha) S_{t-1}, \alpha \in [0, 1].$$

- ullet 平滑参数  $\alpha$  决定了近期或远期数据的权重。 $\alpha \to 1$  时,赋予近期观测值更大的权重
- ullet 可以通过预测精度确定最佳平滑参数  $\alpha^*$

#### 何为"指数平滑"?

推导可知,

$$S_t = \alpha [Y_t + (1-\alpha)Y_{t-1} + \ldots + (1-\alpha)^{t-1}Y_1] + (1-\alpha)^t Y_0.$$

远期数据权重呈几何级数衰减,而几何级数衰减是指数衰减的离散版本,因此得名"指数平滑"。

# 指数平滑法:案例

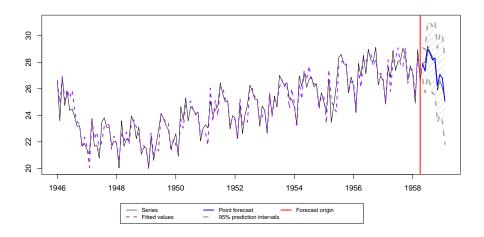
```
es.ts \leftarrow es(births, h = 20, holdout = T)
es.ts
```

```
## Time elapsed: 1.71 seconds
## Model estimated: ETS(ANM)
## Persistence vector g:
## alpha gamma
## 0.94 0.00
## Initial values were optimised.
##
```

## Loss function type: MSE; Loss function valu

46 / 69

## 指数平滑法:案例(续)



### 指数平滑法: 其它实践

• 二次指数平滑法: 对 (一次) 指数平滑法的再次平滑处理

## 非参数模型小结

- 不考虑问题的具体特征,而将其视作黑箱
- 通常仅使用历史数据  $Y_{\tau}(1 \leq \tau < t)$  进行预测
- 与 K 最近邻 (k-nearest neighbors, KNN) 等非参数模型 类似, 采用**直观逻辑**讲行预测
- 调节参数 (tuning parameter) 通常依据预测误差最小化原则来确定

#### 其它模型?

我们也可以自行建模, 例如

$$Y_t = a + b \cdot t + Month_t + \epsilon_t.$$

其中线性函数  $(a+b\cdot t)$  刻画了**趋势因素**,虚拟变量集合  $Month_t$  刻画了**季节因素**。

# 数据准备

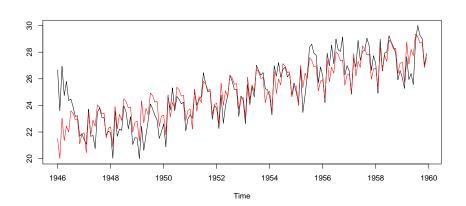
```
nlength <- length(births)
year <- 0:(nlength - 1) %/% 12 + 1946
month <- as.factor(1:nlength %% 12)
dat <- data.frame(births, year, month)
lm.ts <- lm(births ~ year + month, data = dat)</pre>
```

## 模型估计

```
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) -839.288 40.935 -20.50 < 2e-16 ***
                      0.021 21.11 < 2e-16 ***
            0.442
year
           -0.350
month1
                      0.414 -0.85 0.39864
           -1.843
                      0.414 -4.45 1.6e-05 ***
month2
           1.166
month3
                      0.414 2.82 0.00549 **
           -0.501
month4
                      0.414 -1.21 0.22849
            0.587
month 5
                      0.414 1.42 0.15846
month6
            0.141
                      0.414
                              0.34 0.73299
month7
          1.744
                      0.414
                              4.21 4.3e-05 ***
month8
           1.558
                      0.414 3.76 0.00024 ***
            1.073
                      0.414
                              2.59 0.01046 *
month9
month10
            1.139
                      0.414 2.75 0.00666 **
month11
            -0.751
                      0.414
                             -1.81 0.07179 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.1 on 155 degrees of freedom
Multiple R-squared: 0.793, Adjusted R-squared: 0.777
F-statistic: 49.4 on 12 and 155 DF, p-value: <2e-16
```

图 4: 模型估计结果

### 模型预测



# 预测模型比较

```
# write a function to assess accuracy
accuracy.ts <- function(y, yhat){
    # calculate MSE and MAD
    mse <- sum((y - yhat)^2) / length(y)</pre>
    mad <- sum(abs(y - yhat)) / length(y)</pre>
    # return MSE and MAD
    res <- data.frame(mse = mse, mad = mad)
    return(res)
```

## 模型比较

```
## moving average 1.60 0.97
## exponential smoothing 0.32 0.41
## our model 1.11 0.74
```

## 11.2.4 生长曲线法

- 指数增长模型
- Logistic 增长模型
- 再论人口增长模型

## 指数增长模型

#### 指数增长模型

- 性质:属于参数方法,即设定具体的统计模型
- 假设:时间序列(1)刻画了事物的增长过程(或发展演化过程),且(2)增长率恒定;(3)资源充足
- 例子: 资源充足情况下的种群增长

## 指数增长模型: 建模

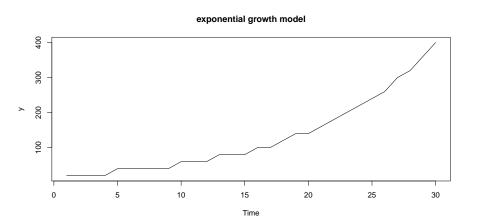
在资源充足的情况下,物种的固有增长率为 r (即繁殖率),种群数量  $Y_t$  可以刻画为

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = r \cdot Y_t.$$

由此求解得到种群数量

$$Y_t = Y_0 \cdot e^{rt}$$
.

# 指数增长型时间序列



#### 指数增长模型:估计

- 取对数,采用最小二乘法估计
- 采用非线性最小二乘法估计

# Logistic 增长模型

#### Logistic 增长模型

- 性质:属于参数方法,即设定具体的统计模型
- 假设: 时间序列刻画了事物的增长过程(或发展演化过程)
- 例子: (1) 种群数量变化; (2) 新产品销售量; (3) 疾病感染 人数

### 种群数量变化

在某个自然环境中,假定只有单一物种竞争资源,那么其种群数量  $Y_t$  的变化可以描述为

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = r \cdot Y_t \cdot (1 - \frac{Y_t}{N})$$

#### 参数的含义:

- 种群的固有增长率 r, 即繁殖率
- ullet 环境资源容纳的最大种群数量 N
- ullet  $Y_t \ll N$  时,近似为指数增长; $Y_t o N$  时,增长近似停滞

## 种群数量变化(续)

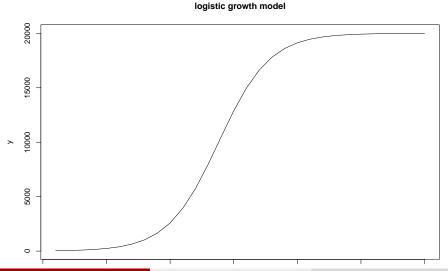
以上模型得到 Logistic 函数,即

$$Y_t = \frac{N}{1+(N/Y_0-1)\cdot e^{-rt}}$$

由此得到 S 型曲线,从而呈现以下特征:

- $Y_t \ll N$  时,近似为指数增长
- $Y_t \to N$  时,增长近似停滞

# Logistic 型时间序列



## 新产品销售量

假定完全创新的新产品(永久品),进入市场之后,其销售量  $Y_t$ 的变化可以描述为

$$\frac{y_t}{N - Y_t} = p + q \cdot \frac{Y_t}{N}$$

#### 参数含义如下:

- ullet  $y_t$  为第 t 期的购买者,而  $Y_t$  为第 t 期初的累积购买者
- N 为市场容量
- p 为创新系数,q 为模仿系数,分别刻画广告和口碑的效应

## 再论人口增长模型

- 马尔萨斯人口论与治乱循环
- 中国历史上的土地/人口比例议题
- 模型参数的变化
  - 固定增长率, 如生育率的变化
  - 容量的变化, 如技术进步带来的粮食增产等

# 时间序列案例分析 (1 个课时)

### 本节知识点

- 时间序列分析建模与预测
- https://github.com/wuhsiang/Courses/blob/ master/healthinfo/cases/case-outpatient.Rmd)

# 时间序列分析实习(2个课时)