Chapter 11: 时间序列分析方法

授课教师：吴翔
wuhsiang@hust.edu.cn

1. 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）

Section 1

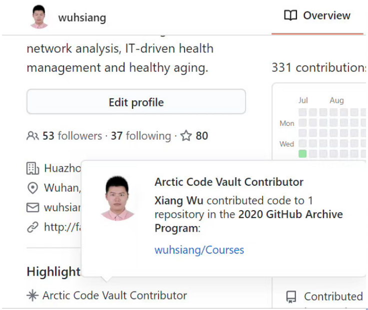
时间序列分析概述（2 个课时）

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

GitHub：Courses 仓库



GitHub Archive Program: the journey of the world's open source code to the Arctic

Julia Metcalf

Your code is safe and sound in the Arctic

At GitHub Universe 2019, we introduced the GitHub Archive Program along with the GitHub Arctic Code Vault. Our mission is to preserve open source software for future generations by storing your code in an archive built to last a thousand years.

Courses 仓库：交流平台

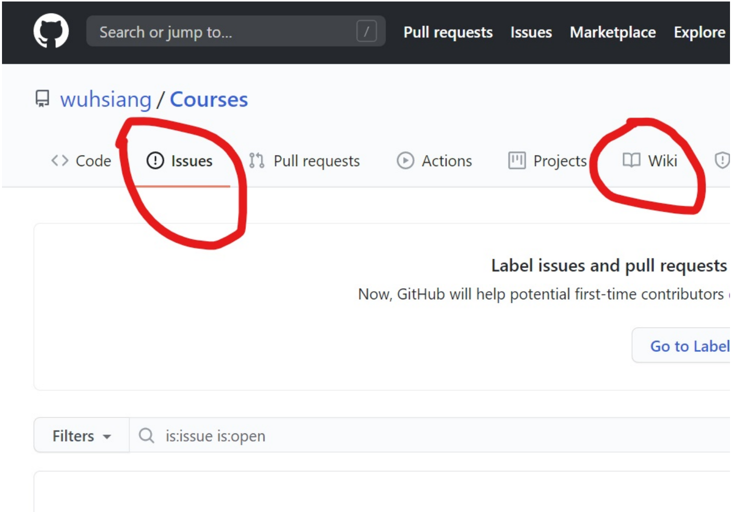


图 3：Courses 仓库：交流功能

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

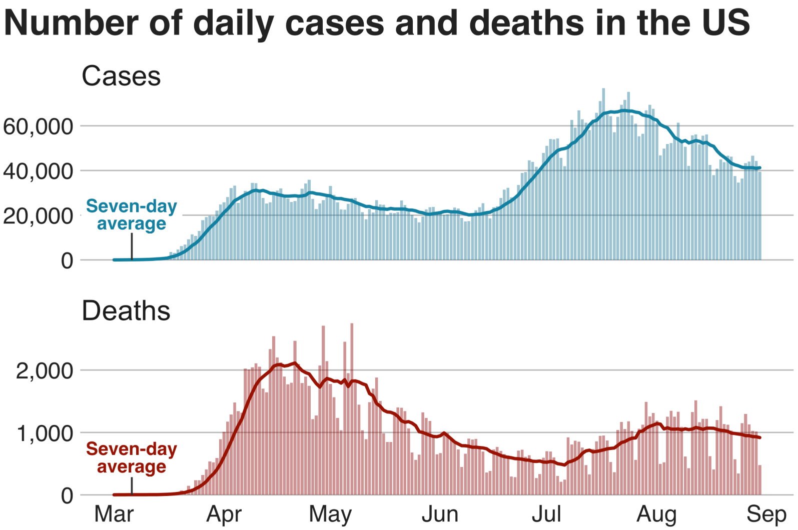


图 4: 美国 COVID-19 感染及死亡人数

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 在英国洛桑实验站 (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 与洛桑实验站

英国洛桑实验站（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

洛桑实验站的农作物

图 5：洛桑实验站

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）

收成变动研究（一）：如何预测小麦产量？

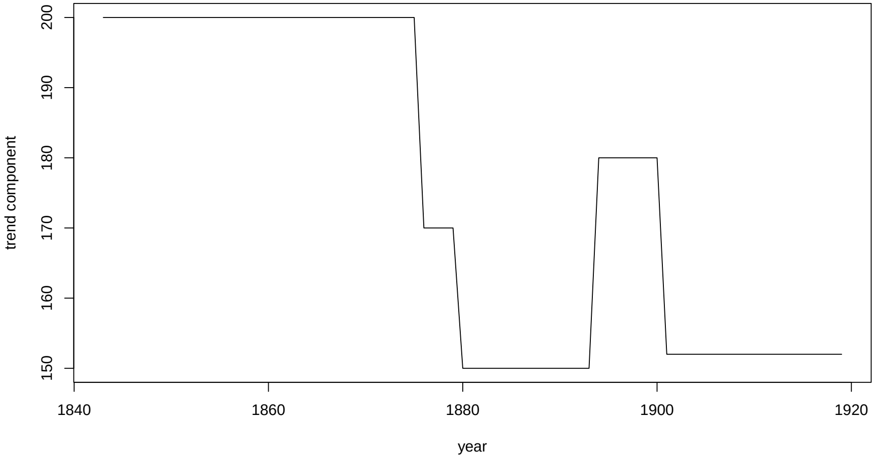
选定一块田地（名曰：“宽埂”），这块田地只使用过动物粪便作为肥料，考虑降水量如何影响小麦产量。

- 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)。
- 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。
- 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



trend component
150
160
170
180
190
200
1840
1860
1880
1900
1920
year

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

收成变动研究（二）：肥料使用对马铃薯产量的影响

研究问题：如何评估肥料使用对马铃薯产量的影响？

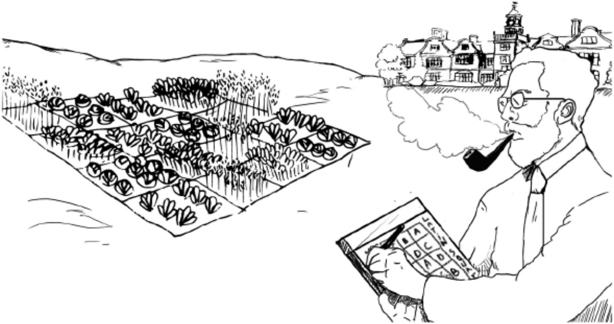


图 6: Fisher 与洛桑实验站

洛桑实验站的早期方案

- 洛桑实验站的早期方案：在不同片田地上，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，例如土壤肥力、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（降雨量）、土壤肥力等。

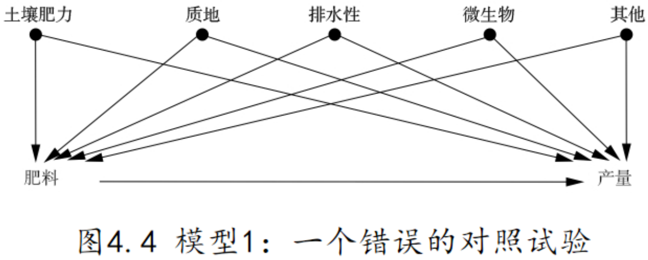


图 7: 洛桑实验站早期研究

理想设定

从因果图上来看，理想的设定如下：

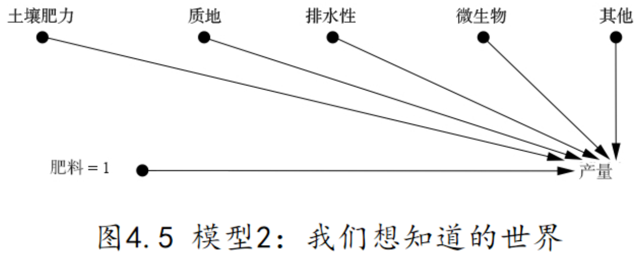


图 8：理想设定

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

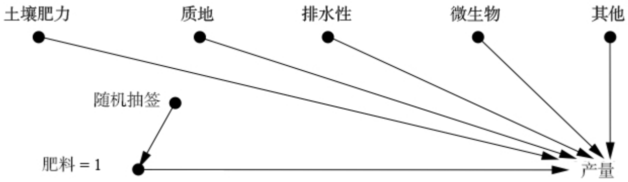


图4.6 模型3：由随机对照试验模拟的世界。

图 9: Fisher 的方案

分解思想

Fisher 在洛桑实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)
    - 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

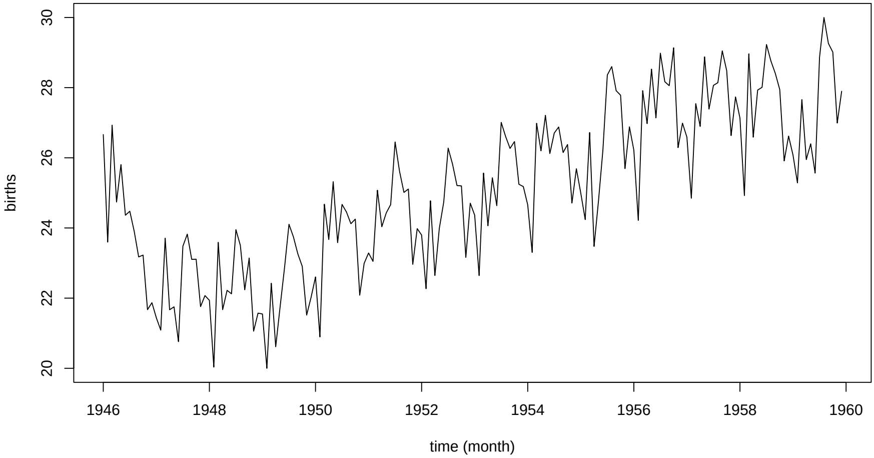
横截面 vs 时间序列数据

- 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


births
30
28
26
24
22
20
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

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：

1. 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
4. 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

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

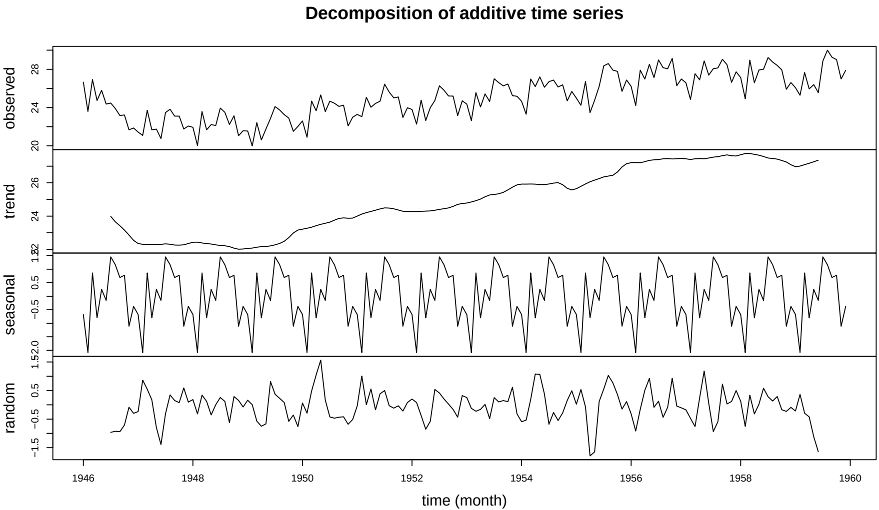
常用的函数类型 $f(\cdot)$ 有两种：累加、累乘。

加法模型

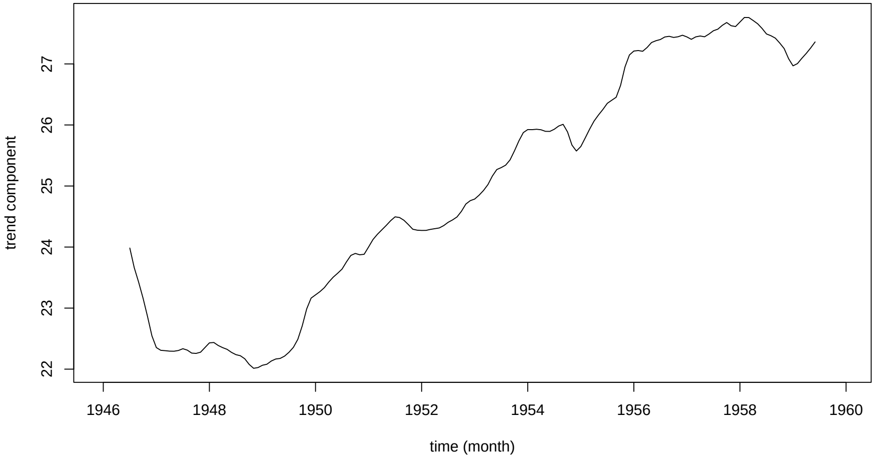
假定趋势因素、季节因素和随机因素相互独立，则可以用加法模型（additive model）来分解各个因素

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

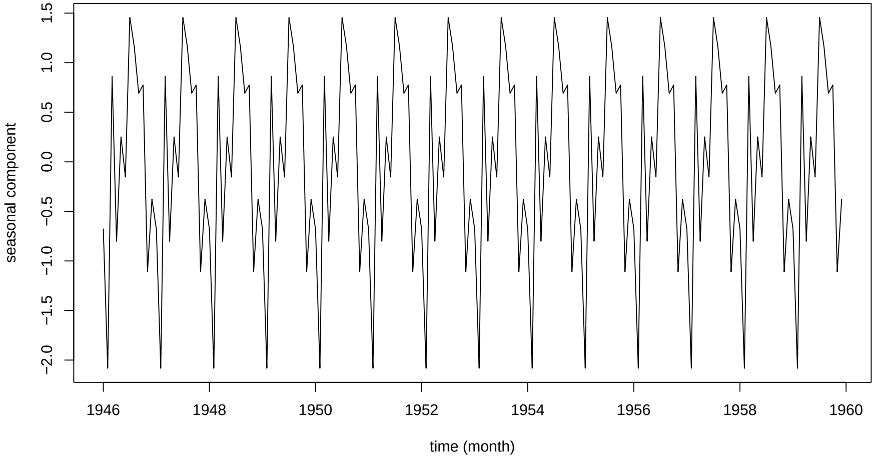
因素分解：加法模型


Decomposition of additive time series
observed
28
24
20
trend
26
24
22
seasonal
1.5
0.5
−0.5
−2.0
random
1.5
0.5
−0.5
−1.5
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

加法模型之趋势因素

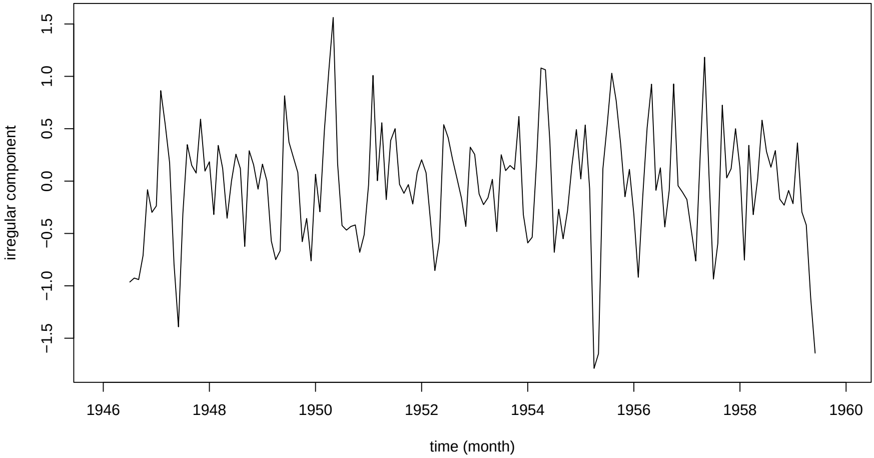

trend component
22
23
24
25
26
27
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

加法模型之季节因素



seasonal component
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

加法模型之随机因素

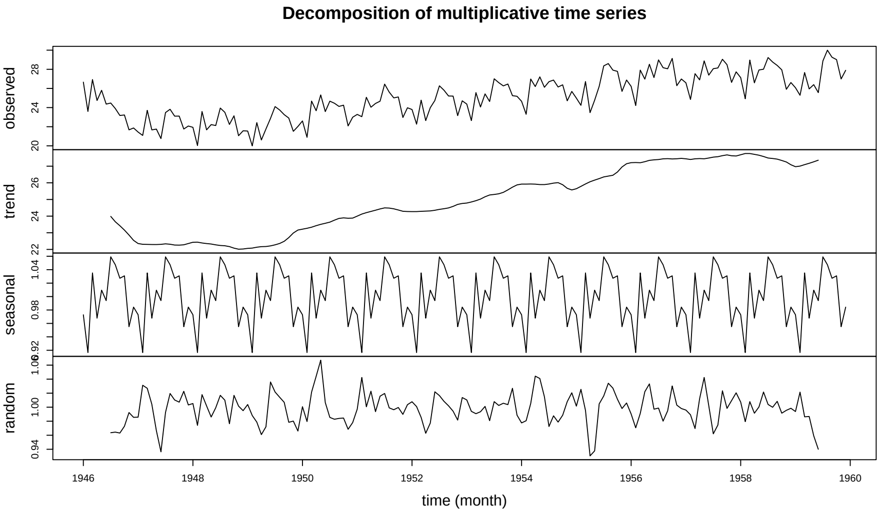

irregular component
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

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
observed
trend
seasonal
random
20
24
28
22
24
26
0.92
0.98
1.04
0.94
1.00
1.06
1946
1948
1950
1952
1954
1956
1958
1960
time (month)

时间序列分析步骤

1. 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
3. 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
4. 获得最终的预测模型

Section 2

时间序列经典分析方法（3 个课时）

本节知识点

- 预测方法概述
- 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 通用预测模型

因变量（响应变量，response variable）记为 y，自变量（预测变量，predictors）集合记作向量 $\mathbf{x} = c(x_1, x_2, ..., x_p)$，则通用模型可以写成：

$$y = \underbrace{f(\mathbf{x})}_{\text{Prediction}} + \underbrace{\epsilon}_{\text{Error}} .$$

从而由

$$\hat{y} = \hat{f}(\mathbf{x})$$

得到预测值。研究者可以选择不同的建模方法，从而得到对 $f(\cdot)$ 的不同估计 $\hat{f}(\cdot)$。在时间序列预测模型中，自变量 $\mathbf{x}$ 是因变量的历史数据 y_τ，其中 $1 \leq \tau \leq t-1$。

理解视角

将观测值 y 分为结构部分 $f(\mathbf{x})$ 和随机部分 ϵ，可以从三个视角来理解：

- 因果性（计量经济领域）：观测项 ＝ 机制项 ＋ 干扰项
- 预测性（统计学习领域）：观测项 ＝ 预测项 ＋ 误差项
- 描述性（统计领域）：观测项 ＝ 概括项 ＋ 残差项

课堂讨论：（1）三个视角在研究目的上有何区别？（2）在建模思路和模型选取上有何差异？

模型误差

本质上，统计学习指一套估计 $f(\cdot)$ 的方法。为此，我们需要了解在估计 $f(\cdot)$ 中涉及的关键理论概念，以及评估准则。

给定 $\hat{f}(\mathbf{x})$ 和 $\mathbf{x}$，那么：

$$E(y-\hat{y})^2 = E[f(\mathbf{x})+\epsilon-\hat{f}(\mathbf{x})]^2 = \underbrace{[f(\mathbf{x})-\hat{f}(\mathbf{x})]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} .$$

第一项是可约误差（reducible error），若使用更适当的方式估计 $f(\cdot)$，则可以减少可约误差；第二项是不可约误差（irreducible error），它是由未被测量的因素导致的，因而不可消除。

自由度

模型自由度，表征了模型的复杂程度，是研究者在估计 $f(\cdot)$ 时的重要考量因素。

在统计学习中，尤其需要理解模型的自由度。

> *The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.*
>
> — *In Statistics*

> *The degrees of freedom are an accounting of how many parameters are estimated by the model and, by extension, a measure of complexity for linear regression models.*
>
> — *In Statistical Learning*

自由度分解

假定多元线性模型 $y = \beta X + \epsilon$ 中，假定 X 包括一列常数和 $(p-1)$ 列变量，那么待估计的参数个数为 p，方差和自由度的分解如下：

- SST：自由度为 $n-1$
- SSE：自由度为 $n-p$
- SSR：自由度为 $p-1$

因而，自由度的分解为：

$$n-1=(n-p)+(p-1)$$

线性回归模型中，模型的自由度等于预测变量的个数。

课堂思考：假设模型有两个解释变量，其中 x_1 是连续变量，x_2 是包含 5 个分类的分类变量，SSR 的自由度为多少？

方差分析表

变异来源	平方和	自由度	均方
回归模型	SSR	$p-1$	$\text{MSR} = \text{SSR}/(p-1)$
误差	SSE	$n-p$	$\text{MSE} = \text{SSE}/(n-p)$
总变异	SST	$n-1$	$\text{MST} = \text{SST}/(n-1)$

假定在线性回归模型 A 的基础上，加了几个变量得到模型 B，模型选择取决于构造的 F 检验：

$$F(\Delta\text{df}, \text{df}_{\text{SSE}}) = \frac{\Delta\text{SSR}/\Delta\text{df}}{\text{MSE}_{\text{B}}} ? > F_{\alpha}$$

预测精度 vs 可解释性

Q1：如何选择函数 $f(\cdot)$?

随着模型自由度（degree of freedom）增加，模型变得更加复杂。

- 预测精度（accuracy)：尽可能减少可约误差，因此要求自由度更大的模型。
- 可解释性（interpretability)：尽可能用少数变量来解释 $\mathbf{x}$ 如何影响 y，因此要求自由度更小的模型。

统计学习中，大多数时候更加关注预测精度，因而可以将 $f(\cdot)$ 视作黑箱。

权衡预测精度与模型可解释性

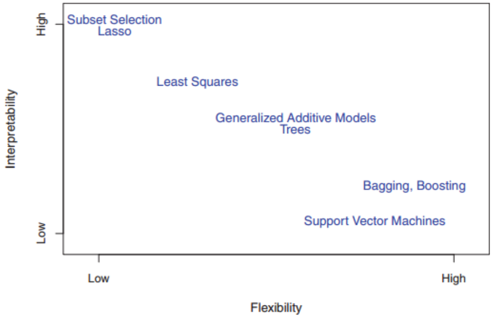


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

图 10: Widely-used models

参数方法 vs 非参数方法

Q2：函数 $f(\cdot)$ 的形式，是否有明确假设？或说，给定训练集（training set）数据，如何估计函数 $f(\cdot)$?

参数方法（parametric methods）：

- 步骤：（1）设定具体的函数形式，包括线性或非线性函数；（2）使用训练集数据，拟合（fit）或说训练（train）模型，得到参数估计值。
- 优点：简化了 $f(\cdot)$ 的估计问题，估计一组参数通常很方便。
- 缺点：一旦模型设定有误，则会导致较大误差。

非参数方法（non-parametric methods）

- 步骤：使用附近的观测值来估计给定 **x** 时的预测值。
- 优点：避免设定特定的函数形式，从而规避了模型设定错误。
- 缺点：无法将估计 $f(\cdot)$ 这一问题变成少量参数的估计，因而远远超过参数方法需要的观测值才能获得 $f(\cdot)$ 的准确估计。

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 $\hat{Y}_t$ 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）

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

测试均方误差的分解

测试集（test set）均方误差的期望值（**expected test MSE**）可以分解为如下三个部分：

$$E(y - \hat{f}(x))^2 = \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{Variance}} + \underbrace{[\mathrm{Bias}(\hat{f}(x))]^2}_{\text{Bias}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} .$$

- 模型方差（variance）：针对不同的训练数据，$\hat{f}$ 的变化程度。
- 模型偏误（bias）：通过相对简化的模型来**近似**真实世界的问题时所引入的误差。

模型复杂程度

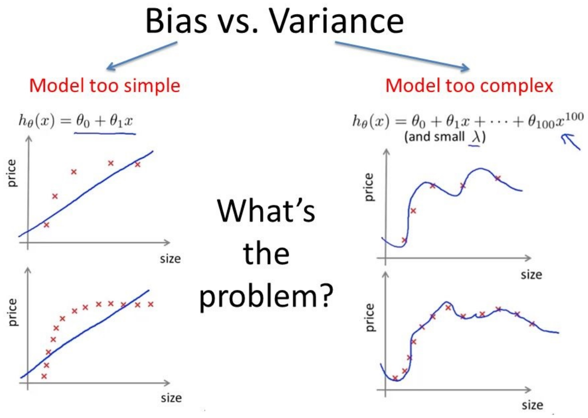


图 11: Model complexity

权衡模型偏误与方差

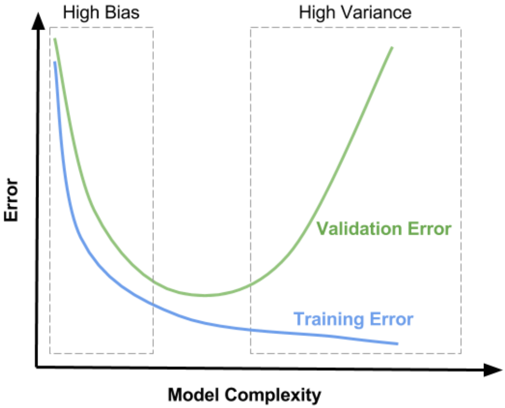


图 12: bias-variance trade-off

如何选择统计模型?

- 传统统计模型的局限：线性回归模型等统计模型通常最小化训练数据的均方误差，但是其测试均方误差（test MSE）却较大。换言之，传统统计模型执着于寻求”真实规律”，以致于将一些随机因素误判为 f 的真实性质。
- 权衡模型偏误与方差（bias-variance trade-off）：随着模型灵活性（或自由度）的增加，模型方差随之增大，但模型偏误则相应减小（过度拟合问题）。通过交叉验证来权衡两者。
- 权衡预测精度与可解释性（accuracy-interpretability trade-off）：诸如 bagging、boosting、support vector machines 等非线性模型具有很高的预测精度，但不易解释；linear models 等易于解释，但预测精度不高。两者的权衡取决于研究目的。

交叉验证

交叉验证（cross-validation）将原始数据集分为训练集（**training set**）和验证集（**validation set**），并以验证集的错误率选择最佳模型。

- 留一交叉验证法（leave-one-out cross validation，LOOCV）
- k 折交叉验证法（k−fold CV）：将观测集随机分为 k 个大小基本一致的组，或说折（fold）。每次选取其中一折作为验证集，而剩余 $k-1$ 折作为训练集。通常，取 $k=5$ 或 $k=10$。

分类模型验证集错误率：

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{Err}_k = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m_k}\sum_{i=1}^{m_k} I(y_i \neq \hat{y}_i).$$

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奏波动模式

移动平均法：模型

预测 Y_{t+1} 的算法为，

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

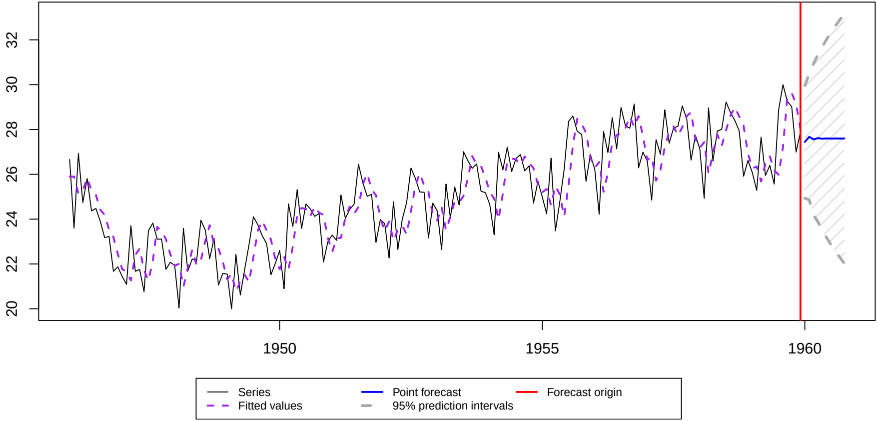
而可以通过最小化误差选取适当的 m 值。具体步骤，可以使用 R 中的smooth包来完成。

移动平均法：案例

```r
ma.ts <- sma(births, h = 20)
ma.ts
```

```
## Time elapsed: 0.88 seconds
## Model estimated: SMA(2)
## Initial values were produced using backcasting.
##
## Loss function type: MSE; Loss function value: 1.6035
## Error standard deviation: 1.3
## Sample size: 168
## Number of estimated parameters: 2
## Number of degrees of freedom: 166
## Information criteria:
##  AIC AICc  BIC BICc
##  560  560  566  567
```

移动平均法：案例（续）


32
30
28
26
24
22
20
1950
1955
1960
Series
Fitted values
Point forecast
95% prediction intervals
Forecast origin

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

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

- 平滑参数 α 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
- 可以通过预测精度确定最佳平滑参数 α^*

何为”指数平滑”?

推导可知，

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

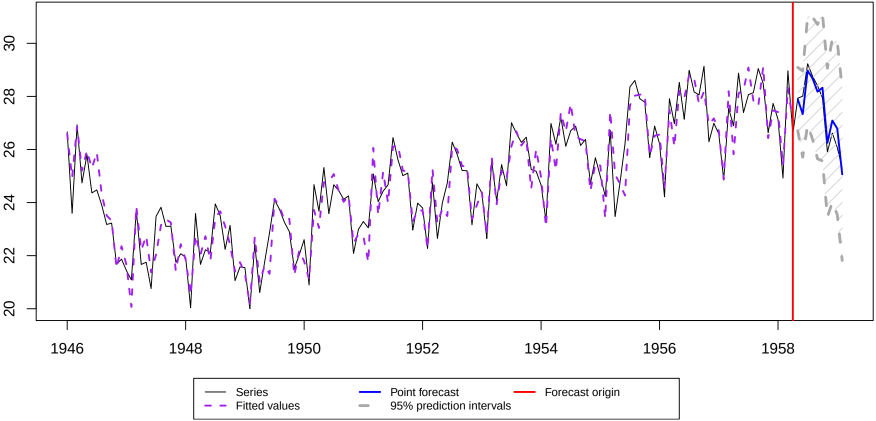
远期数据权重呈几何级数衰减，而几何级数衰减是指数衰减的离散版本，因此得名”指数平滑”。

指数平滑法：案例

```
es.ts <- es(births, h = 20, holdout = T)
es.ts
```

```
## Time elapsed: 1.7 seconds
## Model estimated: ETS(ANM)
## Persistence vector g:
## alpha gamma
##  0.94  0.00
## Initial values were optimised.
##
## Loss function type: likelihood; Loss function value: 135.92
## Error standard deviation: 0.64
## Sample size: 148
## Number of estimated parameters: 15
## Number of provided parameters: 1
## Number of degrees of freedom: 133
## Information criteria:
```

指数平滑法：案例（续）


20
22
24
26
28
30
1946
1948
1950
1952
1954
1956
1958
Series
Fitted values
Point forecast
95% prediction intervals
Forecast origin

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，采用直观逻辑进行预测
- 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)
```

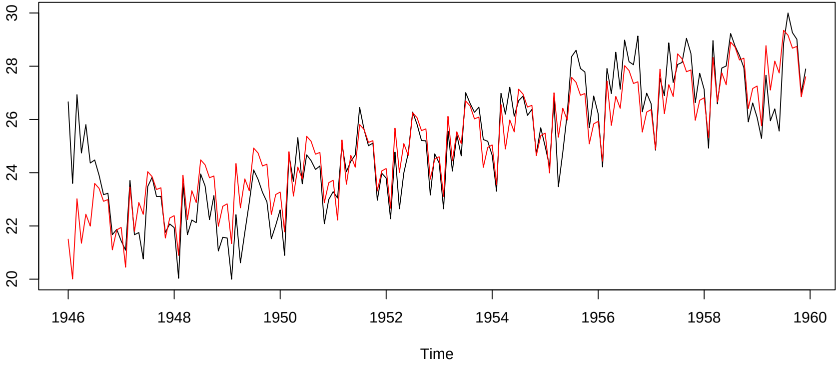
模型估计

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -839.288     40.935  -20.50  < 2e-16 ***
year           0.442      0.021   21.11  < 2e-16 ***
month1        -0.350      0.414   -0.85  0.39864    
month2        -1.843      0.414   -4.45  1.6e-05 ***
month3         1.166      0.414    2.82  0.00549 ** 
month4        -0.501      0.414   -1.21  0.22849    
month5         0.587      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 ***
month9         1.073      0.414    2.59  0.01046 *  
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 13：模型估计结果

模型预测



预测模型比较

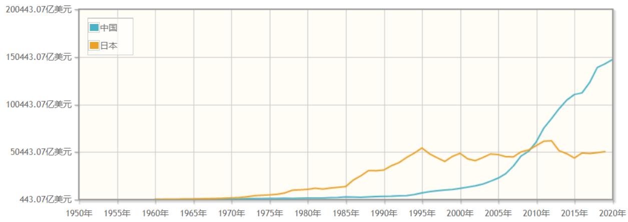
```r
# write a function to assess accuracy
accuracy.ts <- function(y, yhat){
    # calculate MSE and MAD
    mse <- sum((y - yhat)^2) / length(y)
    mad <- sum(abs(y - yhat)) / length(y)
    # return MSE and MAD
    res <- data.frame(mse = mse, mad = mad)
    return(res)
}
```

模型比较

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

GDP 历史数据

GDP(美元)走势图



中国、日本历年GDP数据比较

	中国		日本	
年份	GDP(美元)	占世界%	GDP(美元)	占世界%
2020	14.72万亿（14,722,730,697,890）	17.3811%		
2019	14.28万亿（14,279,937,467,431）	16.2999%	5.06万亿（5,064,872,875,604）	5.7813%
2018	13.89万亿（13,894,817,549,380）	16.0925%	4.95万亿（4,954,806,619,995）	5.7385%

图 14: GDP trends

WHO 疫情统计数据

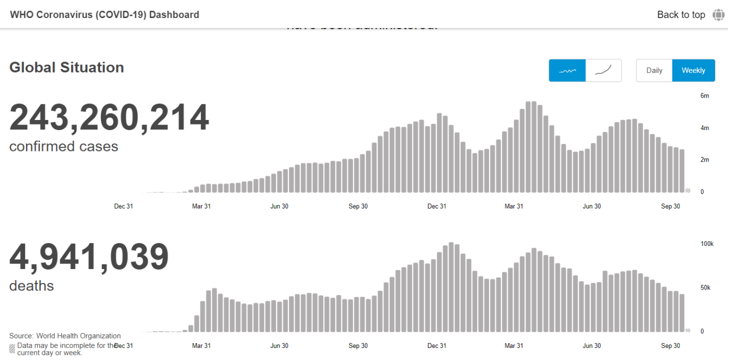


图 15: WHO COVID-19 Dashboard

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

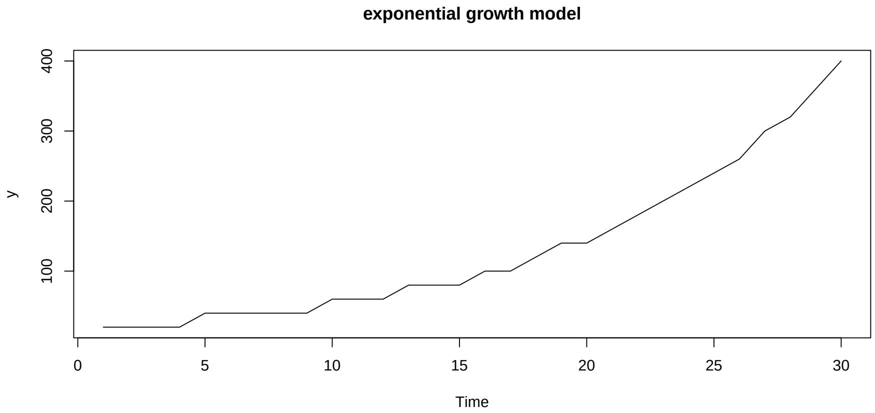
在资源充足的情况下，物种的固有增长率为 r（即繁殖率），种群数量 Y_t 可以刻画为

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

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


exponential growth model
y
100
200
300
400
0
5
10
15
20
25
30
Time

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
- 环境资源容纳的最大种群数量 N
- $Y_t \ll N$ 时，近似为指数增长；$Y_t \to N$ 时，增长近似停滞

种群数量变化（续）

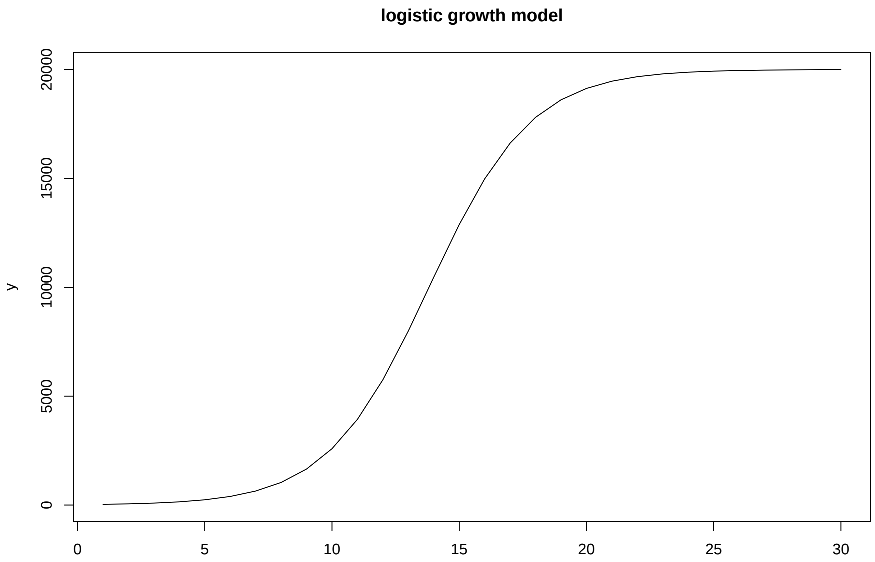
以上模型得到 Logistic 函数，即

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

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：

- 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

疾病传播模型

疾病传播模型与以上模型类似，可以自行推导。

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

Section 3

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)

Section 4

时间序列分析实习（2 个课时）

实习内容

数据来源及说明:

- https://covid19.who.int/data

实习内容:

- 选择感兴趣的议题（感染人数、死亡人数、疫苗接种人数）和部分数据（例如，全球或特定国家）
- 选用特定工具（如 SPSS、Python、R 等）
- 使用时间序列分析方法，探索其规律，并给出后续预测值、评估预测效果