R 语言在心理学研究中的应用: 从数据到论文

胡传鹏

2023 年 1 月 5 日

目录

简介

此处写简介

Chapter 1

2023.2.20 第一节课

1.1 序

一般来说第一节课没有太多实质性的内容，但是可以帮助大家为之后的课程做好心理的准备，这一过程是很重要的。从第一节课大家能知道接下来要上课的内容、想要上好这门课需要做什么准备、最后能从这门课获得什么。使用 RStudio 软件完成编辑和转换功能。在 RStudio 中，安装 bookdown 等必要的扩展包。

后面的 §?? 和 §??

1.2 数据科学

1.2.1 数据科学是什么？

首先我们讲一下这门课的大背景。虽然我们作为心理学人在心理学院学习这门课，我们会说本课是 R 语言在心理学研究当中的应用。但实际上，R 语言会在一个更广的领域中应用，叫做数据科学，data science。那么什么是 data science 呢？在科学研究中有人认为，科学的革命是经过了几次范式转换的。早期的是” 实验” 的科学，我们通过做实验，一个一个地去验证假设；随着计算机越来越发达，我们进入了” 计算” 的范式，通过用各种计算模型模拟的方法，帮助我们去理解世界。但是现在，随着数据越来越多，我们实际上是通过数据驱动的方式进行探索。最近这些年，很多在科技领域尤其是在计算机领域取得的重大突破和进展都是依赖于大量数据的，也就是通过对数据进行提炼从而得到新的发现。比方说最近非常火的 ChatGPT。作为现在全球最火的科技界产品之一，它背后的模型叫做 LM，就是 Language Model。这里说的 Language 实际上就是一个大的语言模型，它依靠的就是大量语言材料的训练。

1.2.2 数据科学的内容

大概 10 多年前，数据科学在我读研究生的时候实际上就出现了。最近这两年大家应该对数据科学已经不再陌生了，可以看到在数据科学里面有传统的计算科学，也有数学和统计。它也有具体的应用领域，比方说应用到商业，或者是我们科研领域。但是不论是哪个领域，它都是需要 domain specific language 的，就是说要有这个领域专属的特殊性知识的。这意味着什么？意味着如果你仅仅懂计算机，那你其实是不能说自己懂 data science 的，如果你仅仅是懂数学和统计，那也不意味这你能解决一个 data science 的问题，必须要将这三方面进行结合。这实际上也是对我们每一个人提出了一个新的要求。

1.3 数据科学的诞生——数字化时代

为什么会有 data science？其实大家应该能感受到，随着我们电脑的普及，互联网越来越发达，我们产生的数据实际上产生了爆炸式的增长。这里有一个可视化的例子。我们可以看到，在计算机出现之前人类产生的数据是非常少的，而计算机出现之后产生的数据越来越多。

1.3.1 图形自动编号

用 R 代码段生成的图形，只要具有代码段标签，且提供代码段选项 `fig.cap="图形的说明文字"`，就可以对图形自动编号，并且可以用如`\@ref(fig:label)` 的格式引用图形。如：

```
plot(1:10, main="程序生成的测试图形")
```

引用如：参见图1.1。引用中的 `fig:` 是必须的。

在通过 LaTeX 转换的 PDF 结果中，这样图形是浮动的。

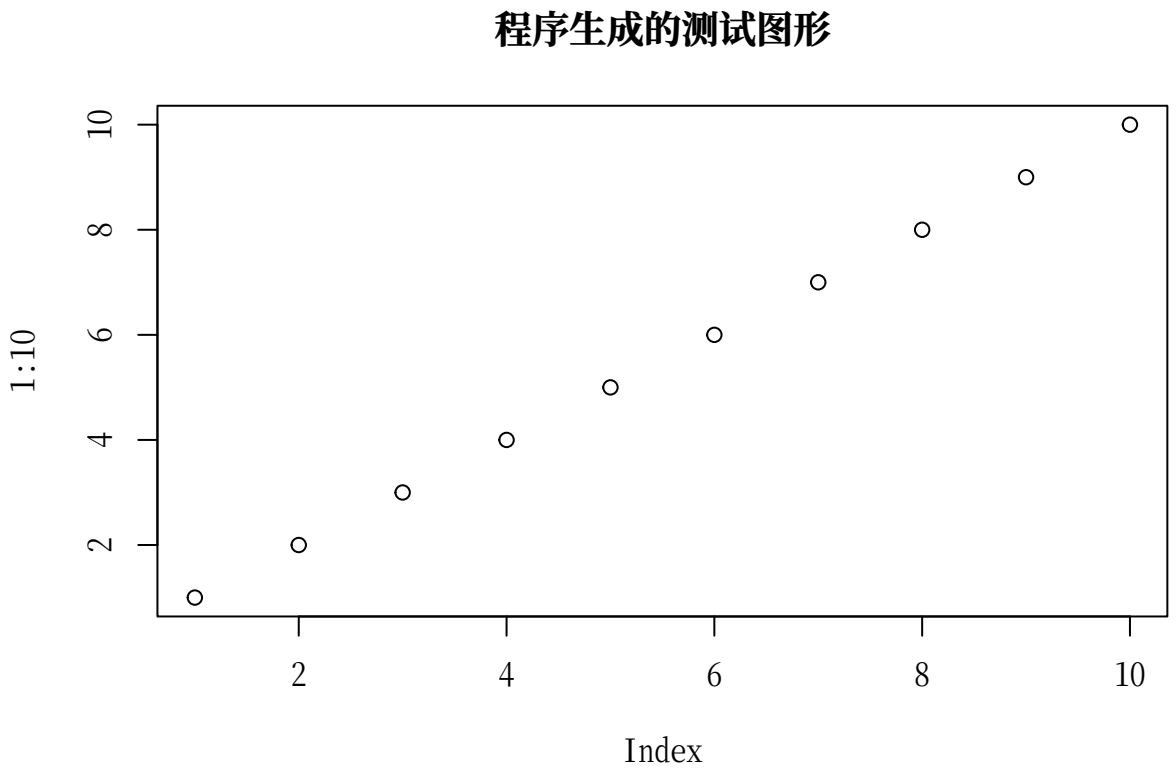

图 1.1: 图形说明文字

Chapter 2

格兰格因果性

2.1 介绍

考虑两个时间序列之间的因果性。这里的因果性指的是时间顺序上的关系，如果 $X_{t-1}, X_{t-2}, ...$ 对 Y_t 有作用，而 $Y_{t-1}, Y_{t-2}, ...$ 对 X_t 没有作用，则称 $\{X_t\}$ 是 $\{Y_t\}$ 的格兰格原因，而 $\{Y_t\}$ 不是 $\{X_t\}$ 的格兰格原因。如果 $X_{t-1}, X_{t-2}, ...$ 对 Y_t 有作用，$Y_{t-1}, Y_{t-2}, ...$ 对 X_t 也有作用，则在没有进一步信息的情况下无法确定两个时间序列的因果性关系。

注意这种因果性与采样频率有关系，在日数据或者月度数据中能发现的领先——滞后性质的因果关系，到年度数据可能就以及混杂在以前变成同步的关系了。

2.2 格兰格因果性的定义

设 $\{\xi_t\}$ 为一个时间序列，$\{\eta_t\}$ 为向量时间序列，记

$$\bar{\eta}_t = \{\eta_{t-1}, \eta_{t-2}, ...\}$$

记 $\mathrm{Pred}(\xi_t|\bar{\eta}_t)$ 为基于 $\eta_{t-1}, \eta_{t-2}, ...$ 对 ξ_t 作的最小均方误差无偏预报，其解为条件数学期望 $E(\xi_t|\eta_{t-1}, \eta_{t-2}, ...)$，在一定条件下可以等于 ξ_t 在 $\eta_{t-1}, \eta_{t-2}, ...$ 张成的线性 Hilbert 空间的投影（比如，(ξ_t, η_t) 为平稳正态多元时间序列），即最优线性预测。直观理解成基于过去的 $\{\eta_{t-1}, \eta_{t-2}, ...\}$ 的信息对当前的 ξ_t 作的最优预测。

令一步预测误差为

$$\varepsilon(\xi_t|\bar{\eta}_t) = \xi_t - \mathrm{Pred}(\xi_t|\bar{\eta}_t)$$

令一步预测误差方差，或者均方误差，为

$$\sigma^2(\xi_t|\bar{\eta}_t) = \mathrm{Var}(\varepsilon_t(\xi_t|\bar{\eta}_t)) = E\left[\xi_t - \mathrm{Pred}(\xi_t|\bar{\eta}_t)\right]^2$$

考虑两个时间序列 $\{X_t\}$ 和 $\{Y_t\}$，$\{(X_t, Y_t)\}$ 宽平稳或严平稳。

- 如果
$$\sigma^2(Y_t|\bar{Y}_t,\bar{X}_t)<\sigma^2(Y_t|\bar{Y}_t)$$
则称 $\{X_t\}$ 是 $\{Y_t\}$ 的**格兰格原因**，记作 $X_t\Rightarrow Y_t$。这不排除 $\{Y_t\}$ 也可以是 $\{X_t\}$ 的格兰格原因。
- 如果 $X_t\Rightarrow Y_t$，而且 $Y_t\Rightarrow X_t$，则称互相有**反馈**关系，记作 $X_t\Leftrightarrow Y_t$。
- 如果
$$\sigma^2(Y_t|\bar{Y}_t,X_t,\bar{X}_t)<\sigma^2(Y_t|\bar{Y}_t,\bar{X}_t)$$
即除了过去的信息，增加同时刻的 X_t 信息后对 Y_t 预测有改进，则称 $\{X_t\}$ 对 $\{Y_t\}$ 有瞬时因果性。这时 $\{Y_t\}$ 对 $\{X_t\}$ 也有瞬时因果性。
- 如果 $X_t\Rightarrow Y_t$，则存在最小的正整数 m，使得
$$\sigma^2(Y_t|\bar{Y}_t,X_{t-m},X_{t-m-1},\dots)<\sigma^2(Y_t|\bar{Y}_t,X_{t-m-1},X_{t-m-2},\dots)$$
称 m 为**因果性滞后值** (causality lag)。如果 $m>1$，这意味着在已有 $Y_{t-1},Y_{t-2},\dots$ 和 $X_{t-m},X_{t-m-1},\dots$ 的条件下，增加 X_{t-1}, ..., X_{t-m+1} 不能改进对 Y_t 的预测。

例 2.1. 设 $\{\varepsilon_t,\eta_t\}$ 是相互独立的零均值白噪声列，$\mathrm{Var}(\varepsilon_t)=1$, $\mathrm{Var}(\eta_t)=1$, 考虑
$$\begin{aligned}Y_t=&X_{t-1}+\varepsilon_t\\X_t=&\eta_t+0.5\eta_{t-1}\end{aligned}$$
用 $L(\cdot|\cdot)$ 表示最优线性预测，则
$$\begin{aligned}&L(Y_t|\bar{Y}_t,\bar{X}_t)\\=&L(X_{t-1}|X_{t-1},\dots,Y_{t-1},\dots)+L(\varepsilon_t|\bar{Y}_t,\bar{X}_t)\\=&X_{t-1}+0\\=&X_{t-1}\\\sigma(Y_t|\bar{Y}_t,\bar{X}_t)=&\mathrm{Var}(\varepsilon_t)=1\end{aligned}$$
而
$$Y_t=\eta_{t-1}+0.5\eta_{t-2}+\varepsilon_t$$
有
$$\gamma_Y(0)=2.25,\gamma_Y(1)=0.5,\gamma_Y(k)=0,k\geq 2$$
所以 $\{Y_t\}$ 是一个 MA(1) 序列，设其方程为
$$Y_t=\zeta_t+b\zeta_{t-1},\zeta_t\sim\mathrm{WN}(0,\sigma_\zeta^2)$$
可以解出
$$\begin{aligned}\rho_Y(1)=&\frac{\gamma_Y(1)}{\gamma_Y(0)}=\frac{2}{9}\\b=&\frac{1-\sqrt{1-4\rho_Y^2(1)}}{2\rho_Y(1)}\approx 0.2344\\\sigma_\zeta^2=&\frac{\gamma_Y(1)}{b}\approx 2.1328\end{aligned}$$

于是

$$\sigma(Y_t|\bar{Y}_t) = \sigma_\zeta^2 = 2.1328$$

所以

$$\sigma(Y_t|\bar{Y}_t, \bar{X}_t) = 1 < 2.1328 = \sigma(Y_t|\bar{Y}_t)$$

即 X_t 是 Y_t 的格兰格原因。

反之，X_t 是 MA(1) 序列，有

$$\eta_t = \frac{1}{1+0.5B}X_t = \sum_{j=0}^{\infty}(-0.5)^j X_{t-j}$$

其中 B 是推移算子（滞后算子）。于是

$$\begin{aligned}
L(X_t|\bar{X}_t) =& L(\eta_t|\bar{X}_t) + 0.5L(\eta_{t-1}|\bar{X}_t) \\
=& 0.5\sum_{j=0}^{\infty}(-0.5)^j X_{t-1-j} \\
=& -\sum_{j=1}^{\infty}(-0.5)^j X_{t-j} \\
\sigma(X_t|\bar{X}_t) =& \mathrm{Var}(X_t - L(X_t|\bar{X}_t)) \\
=& \mathrm{Var}(\eta_t) = 1
\end{aligned}$$

而

$$\begin{aligned}
L(X_t|\bar{X}_t, \bar{Y}_t) =& L(\eta_t|\bar{X}_t, \bar{Y}_t) + 0.5L(\eta_{t-1}|\bar{X}_t, \bar{Y}_t) \\
=& 0 + 0.5L(\sum_{j=0}^{\infty}(-0.5)^j X_{t-1-j}|\bar{X}_t, \bar{Y}_t) \\
=& -\sum_{j=1}^{\infty}(-0.5)^j X_{t-j} \\
=& L(X_t|\bar{X}_t)
\end{aligned}$$

所以 Y_t 不是 X_t 的格兰格原因。

考虑瞬时因果性。

$$\begin{aligned}
L(Y_t|\bar{X}_t, \bar{Y}_t, X_t) =& X_{t-1} + 0(\text{注意}\varepsilon_t\text{与}\{X_s, \forall s\}\text{不相关} \\
=& L(Y_t|\bar{X}_t, \bar{Y}_t)
\end{aligned}$$

所以 X_t 不是 Y_t 的瞬时格兰格原因。

例 2.2. 在例2.1中，如果模型改成

$$\begin{aligned}
Y_t =& X_t + \varepsilon_t \\
X_t =& \eta_t + 0.5\eta_{t-1}
\end{aligned}$$

有怎样的结果？

这时

$$Y_t = \varepsilon_t + \eta_t + 0.5\eta_{t-1}$$

仍有

$$\gamma_Y(0) = 2.25, \gamma_Y(1) = 0.5, \gamma_Y(k) = 0, k \geq 2$$

所以 Y_t 还服从 MA(1) 模型

$$Y_t = \zeta_t + b\zeta_{t-1}, b \approx 0.2344, \sigma_\zeta^2 \approx 2.1328$$

$$\begin{aligned}
L(Y_t|\bar{Y}_t, \bar{X}_t) =& L(X_t|\bar{Y}_t, \bar{X}_t) + 0 \\
=& L(\eta_t|\bar{Y}_t, \bar{X}_t) + 0.5L(\eta_{t-1}|\bar{Y}_t, \bar{X}_t) \\
=& 0 + 0.5L(\sum_{j=0}^{\infty}(-0.5)^j X_{t-1-j}|\bar{Y}_t, \bar{X}_t) \\
=& -\sum_{j=1}^{\infty}(-0.5)^j X_{t-j} \\
=& X_t - \eta_t \\
\sigma(Y_t|\bar{Y}_t, \bar{X}_t) =& \mathrm{Var}(\varepsilon_t + \eta_t) = 2
\end{aligned}$$

而

$$\sigma(Y_t|\bar{Y}_t) = \sigma_\zeta^2 \approx 2.1328 > \sigma(Y_t|\bar{Y}_t, \bar{X}_t) = 2$$

所以 X_t 是 Y_t 的格兰格原因。

反之，

$$\begin{aligned}
L(X_t|\bar{X}_t, \bar{Y}_t) =& -\sum_{j=1}^{\infty}(-0.5)^j X_{t-j} \\
=& L(X_t|\bar{X}_t)
\end{aligned}$$

所以 Y_t 不是 X_t 的格兰格原因。

考虑瞬时因果性。

$$\begin{aligned}
L(Y_t|\bar{X}_t, \bar{Y}_t, X_t) =& X_t + 0(\text{注意}\varepsilon_t\text{与}\{X_s, \forall s\}\text{不相关} \\
=& X_t \\
\sigma(Y_t|\bar{X}_t, \bar{Y}_t, X_t) =& \mathrm{Var}(\varepsilon) \\
=& 1 < 2 = \sigma(Y_t|\bar{X}_t, \bar{Y}_t)
\end{aligned}$$

所以 X_t 是 Y_t 的瞬时格兰格原因。

$$[aaa]$$