线性模型

金林

中南财经政法大学统计系

2019-03

1 Python 数据类型

- Python 数据类型
- 数值分析库 numpy

- Python 数据类型
- 数值分析库 numpy

python 对象

python 创建和控制的实体称为对象 (object), 它们可以是变量、数组、字符串、函数或结构。由于 python 是一种所见即所得的脚本语言，故不需要编译。在 python 里，对象是通过名字创建和保存的。可以用 who 命令来查看当前打开的 python 环境里的对象，用 del 删除这些对象。1. 查看数据对象 2. 生成数据对象 3. 删除数据对象上面列出的是新创建的数据对象 x 的名称。python 对象的名称必须以一个英文字母打头，并由一串大小写字母、数字或下画线组成。注意：python 区分大小写，比如，Orange 与 orange 数据对象是不同的。不要用 python 的内置函数名作为对象的名称，如 who/del 等。

数据的基本类型

python 的基本数据类型包括数值型、逻辑型、字符型、复数型等，也可能是缺失值。1. 数值型数值型数据的形式是实数，可以写成整数（如 =3）、小数（如 x=1.46）/科学计数（y=1e9）d 的方式，该类型数据默认是双精度数据。python 支持 4 种不同的数字类型：int(有符号整型)；long（长整型，也可以代表八进制和十六进制）；flont（浮点型）；complex（复数）。说明：python 中显示数据或对象内容直接用其名称，相当于执行 print 函数，见下。2. 逻辑型逻辑型数据只能取 True 或 False 值。可以通过比较获得逻辑型数据，3. 字符型字符型数据的形式是夹在双引号 “” 或单引号’‘之间的字符串，如 ‘MR’。注意：一定要用英文引号，不能用中文引号 “” 或 ‘’。python 语言中的 string（字符串）是由数字、字母、下画线组成的一串字符。一般形式为 s=‘I love python’ 它是编程语言中表示文本的数据类型。

另外，python 字符串具有切片功能，即由左到右索引默认从 0 开始；由右到左索引默认从-1 开始。如果要实现从字符串中获取一段子字符串，可以使用变量 [头下标：尾下标]，其中下标从 0 开始算起，可以是正数或负数，也可以为空，表示取到头或尾。比如，上例中 s[7] 的值是 p，s[2 : 6] 的结果是 love.

加号（+）是字符串连接运算符，星号（*）是重复操作。4. 缺失值有些统计资料是不完整的。当一个元素或值在统计的时候是 “不可得到” 或 “缺失值” 的时候，相关位置可能会被保留并且赋予一个特定的 nan（not available number，不是一个数）值。任何 nan 的运算结果都是 nan。例如，float（‘nan’）就是一个实数缺失值。

5. 数据类型转换有时，需要对数据内置的类型进行转换，只须将数据类型作为函数名即可。以下几个内置的函数可以实现数据类型之间的转换。这些函数返回一个新的对象，表示转换的值。下面列出几种常用的数据类

标准数据类型

在内存中存储的数据可以有多种类型。例如，一个人的年龄可以用数字来存储，名字可以用字符来存储。python 定义了一些标准类型，用于存储各种类型的数据，这些标准的数据类型是由前述基本类型构成的 1. list（列表）list（列表）是 python 中使用最频繁的数据类型。列表可以完成大多数集合类的数据结构实现。它支持字符、数字、字符串，甚至可以包含列表（即嵌套）。列表用 [] 标识，是一种最通用的复合数据类型。python 的列表也具有切片功能，列表中值的切割也可以用到变量 [头下标：尾下标]，可以截取相应的列表，从左到右索引默认从 0 开始，从右到左索引默认从-1 开始，下标可以为空，表示取到头或尾。

加号 + 是列表连接运算符，星号 * 是重复操作。操作类似字符串。

列表 list 是我们进行数据分析的基本类型，所以必须掌握。2. tuple（元组）元组是另一种数据类型，类似于 list（列表）。元组用“()”标识，内部元素用逗号隔开。元组不能赋值，相当于只读列表。操作类似列表。

3. dictionary（字典）字典也是一种数据类型，且可存储任意类型对象。字典的每个键值对用冒号“：”分隔，每个键值对之间用逗号“，”分隔，整个字典包括在花括号 {} 中，格式如下：
dict=$\{$key1:value1,key2:value2$\}$ 键必须是唯一的，但值则不必，值可以取任何数据类型，如字符串、数字或元组。

字典是除列表外 python 中最灵活的内置数据结构类型。列表是有序的对象集合，字典是无序的对象集合。

两者之间的区别在于：字典中的元素是通过键来存取的，而不是通过下标存取。

- Python 数据类型
- 数值分析库 numpy

数值分析库 numpy

在使用 numpy 库前，须加载其到内存中，语句为 import numpy，通常将其简化为 import numpy as np

一维数组（向量）

二维数组（矩阵）

数组的操作

1. 数组的维度
2. 空数组
3. 零数组
4. 1 数组
5. 单位阵 ## 数据分析库 pandas

在数据分析中，数据通常以变量（一维数组，python 中用序列表示）和矩阵（二维数组，python 中用数据框表示）的形式出现，下面结合 python 介绍 pandas 基本的数据操作。

注意：在 python 编程中，变量通常以列表（一组数据），而不是一般编程语言的标量（一个数据）形式出现。

序列（series）

1. 创建序列（向量、一维数组）假如要创建一个含有 n 个数值的向量（X=x1,x2...,xn），python 中创建序列的函数是列表，这些向量可以是数字型的，也可以是字符串型的，还可以是混合型的。

特别说明：python 中显示数据或对象内容直接用其名称，见下。2. 生成系列 3. 根据列表构建序列 4. 系列合并 5. 系列切片

数据框 (DataFrame)

pandas 中的函数 DataFrameO 可用序列构成一个数据框，如下页的 df1 和 df2。数据框相当于关系数据库中的结构化数据类型，传统的数据大都以结构化数据形式存储于关系数据库中，因而传统的数据分析是以数据框为基础的。python 中的数据分析大都是基于数据框进行的，所以本书的分析也是以数据类型为主，向量和矩阵都可以看成数据框的一个特例。1. 生成数据框 2. 根据列表创建数据框 3. 根据字典创建数据框 4. 增加数据框列 5. 删除数据框列 6. 缺失值处理 7. 数据框排序

数据框的读写

PANDAS 读取数据集

大的数据对象常常从外部文件读入，而不是在 python 中直接输入的。外部的数据源有很多，可以是电子表格、数据库、文本文件等形式。python 的导入工具非常简单，但是对导入文件有一些比较严格的限制。本书使用的是 pandas 包读取数据的方式，事先须调用 pandas 包，即 import pandas。

1. 从剪贴板上读取前面讲到，电子表格是目前数据管理和编辑最方便的工具，所以可以考虑用电子表格管理数据，用 python 分析数据，电子表格与 python 之间的数据交换（适用于全书）过程非常简单，简述如下。先在 Dapy-data.xlsx 数据文件的【BSdata】表中选取 A1 : H52，复制，然后在 python 中读取数据。这里，BSdata 为读入 python 中的数据框名，clipboard 为剪贴板。
2. 读取 csv 格式数据虽然 python 可以直接复制表格数据，但也可读取电子表格工作簿中的一个表格（例如，在 Excel 中将数据 Dapy-data.xlsx 的表单 [BSdata] 另存为 BSdata.csv，这时 BSdata.csv 本质上也是文本文件，是以逗号分隔的文本数据，既可以用记事本打开，也可用电子表格软件打开，是最通用的数据格式），其读取命令也最简单，如下所示。
3. 读取 Excel 格式数据使用 pandas 包中的 read-excel 可直接读取 Excel 文档中的任意表单数据，其读取命令也比较简单，例如，要读取 Dapy-data.xlsx 表单的 [BSdata]，可用以下命令。
4. 读取其他统计软件的数据要调用 SAS、SSPS、Stata 等统计软件的数据集，须先用相应的包，详见 python 手册。

PANDAS 数据集的保存

数据框的操作

基本信息

1. 数据框显示

有三种显示数据框内容的函数，即 indoO(显示数据结构)、headO(显示数据框前 5 行)、tailO(显示数据框后 5 行)。1. 数据框列名（变量名）2. 数据框行名（样品名）3. 数据框维度 4. 数据框值（数组）

选取变量

选取数据框中变量的方法主要有以下几种。1. “.” 法或 [‘’] 法：这是 python 中最直观的选择变量的方法，比如，要选择数据框 BSdata 中的 “身高” 和 “体重” 变量，直接用 “BSdata. 身高” 与 “BSdata. 体重” 即可，也可用 BSdata [‘身高’] 与 [‘体重’]，该方法书写比 “.” 法烦琐，却是最不容易出错且直观的一种方法，可推广到多个变量的情形，推荐使用。2. 下标法：由于数据框是二维数组（矩阵）的扩展，所以也可以用矩阵的列下标来选取变量数据，这种方法进行矩阵 (数据框) 运算比较方便。例如，dat.iloc [i,j] 表示数据框（矩阵）的第 i 行、第 j 列数据，dat.iloc [i,] 表示 dat 的第 i 行数据向量，而 dat.iloc[,j] 表示 dat 的第 j 列数据向量（变量）。再如，“身高” 和 “体重” 变量在数据框 BSdata 的第 3、4 两列。

提取样品

选取观测与变量