·论著·一次研究·

病例交叉研究中条件Logistic回归在Python中的实现

张清煜，李培政，罗晨曦，李湘莹，马 露

武汉大学公共卫生学院（武汉 430071）

【摘要】目的 探索病例交叉研究中条件 Logistic 回归在 Python 软件中的实现。**方法** 以研究某地空气污染物 NO_2 暴露与因肺部感染住院的关系作为实例，利用 Python 构建条件 Logistic 回归模型，比较其与常用统计软件 R 和 SAS 的建模过程以及统计分析结果的异同。**结果** Python、R 和 SAS 三种软件建模逻辑相似，Python 的建模语言与其他统计软件相比稍显繁琐，与 SAS 在参数检验方法上也略有差异，但三种软件的参数估计结果完全相同。**结论** Python 软件可实现条件 Logistic 回归分析，进一步拓展了 Python 在统计分析的应用场景。

【关键词】Python；条件逻辑回归；病例交叉研究

Implementation of conditional Logistic regression in case-crossover study in Python

Qing-Yu ZHANG, Pei-Zheng LI, Chen-Xi LUO, Xiang-Ying LI, Lu MA

School of Public Health, Wuhan University, Wuhan 430071, China
Corresponding author: Lu MA, Email: malu@whu.edu.cn

【Abstract】Objective To explore the implementation of conditional Logistic regression in Python software for case-crossover study. **Methods** The relationship between exposure to air pollutant nitrogen dioxide and hospitalization due to pulmonary infection was studied as an example. The conditional Logistic regression model was constructed by using Python to compare the modeling process and statistical analysis results with common statistical software R and SAS. **Results** The modeling logic of Python, R and SAS is similar. Compared with the statistical softwares, the modeling language of Python is a little more complicated, and it is also slightly different from SAS in parameter test methods, but the parameter estimation results of the three softwares are identical. **Conclusion** Python software could realize conditional Logistic regression analysis, further expanding the application scenarios of Python in statistical analysis.

【Keywords】Python; Conditional Logistic regression; Case-crossover study

DOI: 10.12173/j.issn.1004-5511.202302031
基金项目：湖北省卫生健康委 2021—2022 年度科研项目（WJ2021F103）
通信作者：马露，博士，副教授，硕士研究生导师，Email：malu@whu.edu.cn

病例交叉研究由美国学者麦克卢尔（McClure）于 1991 年提出，是一种通过比较研究对象在急性事件发生前一段时间的暴露情况与未发生事件的某段时间内的暴露情况，来研究短暂暴露对罕见急性病的瞬间影响的流行病学方法，目前已成为环境污染相关健康效应研究中应用最广泛的设计类型之一。病例交叉研究的数据中，通常只有病例没有对照，为了研究暴露与研究对象疾病罹患的关系，通常以该患者健康效应出现时间点前（和 / 或后）的某几个时间点该患者的个体暴露状态作为其自身对照，形成 1∶1 或者 1∶M 的配比 [1]。因此在统计方法上通常采用条件 Logistic 回归对数据进行分析。目前常规统计软件如 R、SAS 等均可完成条件 Logistic 回归分析 [2]。相较于传统数据统计工具，Python 作为一款流行的计算机语言，具有强大的通用性与可拓展性特点，特别是在控制其他软件实现自动化处理，智能化完成数据的采集、清洗、预处理以及数据挖掘等方面拥有明显优势。但目前将 Python 应用于流行病学研究的案例较为少见，因此本文将应用病例交叉研究的实例，探讨 Python 实现条件 Logistic 回归的过程，并比较其与 R 和 SAS 统计软件在建模以及参数估计结果上的异同，以拓展 Python 在流行病学领域中的应用。

1 资料与方法

1.1 资料来源

案例资料来源于某地某年的住院首页资料，根据国际疾病分类第 10 版（ICD-10）对疾病进行编码，选择肺部感染（ICD-10 代码：J98.414）的患者作为研究对象。在这项研究中，共有 3 216 例肺部感染患者纳入研究。

气象数据来自中国气象数据网，包含该地研究期间的每日平均温度（℃）和日平均相对湿度（%），NO_2 浓度（μg/m^3）资料来源于当地环境监测中心。研究对象病例日（部分）温度、湿度以及 NO_2 浓度数据如表 1 所示。

1.2 模型构建

本研究 Python 使用 Cox 回归对条件 Logistic 回归进行拟合，Cox 比例风险模型的基本形式为：

$$h(t, X)=h_0(\beta'X)=h_0(t)exp(\beta_1X_1+\beta_2X_2+\cdots+\beta_mX_m)$$

$h(t, X)$ 是具有协变量 X 的个体在时刻 t 时的风险函数，t 为生存时间，$X=(X_1, X_2,\cdots, X_m)'$ 是可能影响生存时间的有关因素。$h_0(t)$ 是所有协变量取值为 0 时的风险函数，称为基线风险函数。$\beta=(\beta_1, \beta_2, \cdots, \beta_m)$ 为 Cox 模型的回归系数，是待估的回归参数 [3]。

2 结果

2.1 数据预处理

根据病例交叉研究的原理，在原始数据的基础上，需要为每个病例日匹配 3 至 4 个对照日，使得对照组的特征与病例组的特征相似，以减少潜在混杂因素对研究结果的干扰。方法为选择病例日和对照日为同年、同月的同一个星期几，本研究对应选择了 10 995 个对照日。SAS 拟合条件 Logistic 回归有两种方法，分别为直接使用 Logistic 回归和借用 Cox 回归并定义分层变量后实现，两者运行结果相同。本文 Python 和 SAS 均采用分层 Cox 风险比例模型进行拟合，此法需新增一个时间变量（time），令 time=1−case（病例日 case 编码为 1，对照日 case 编码为 0），设置原则为：病例日对应的值小于对照日对应的值即可。新设置的变量 time 可作为 Cox 风险比例模型中的生存时间变量，case 相当于终检变量 [4]。匹配后数据信息（部分）见表 2。

表1 某地污染物信息

Table 1. Pollutant information of a place

ID	平均温度（℃）	相对湿度（%）	NO_2（μg/m^3）
1	22.7	60	25.01990
2	23.4	65	82.03495
3	22.7	60	25.09266
4	30.7	53	32.49965
……	……	……	……
3214	9.1	83	22.85851
3215	8.1	78	26.17080
3216	4.5	82	50.70897

表2 预处理后某地污染物信息

Table 2. Pollutant information of a place after pretreatment

ID	case	日期	平均温度（℃）	相对湿度（%）	NO_2（10 μg/m^3）
1	0	09–07	26.0	70	4.503582
1	0	09–14	27.6	61	1.608422
1	1	09–21	22.7	60	2.501990
1	0	09–28	22.4	69	7.327256
……	……	……	……	……	……
3216	1	12–07	4.5	82	5.070897
3216	0	12–14	10.2	69	5.174384
3216	0	12–21	6.7	89	4.915665
3216	0	12–28	4.7	84	4.449970

2.2 模型拟合

条件 Logistic 回归在 Python 中的实现首先需调用 pandas 库导入病例交叉数据并命名为“wb”，“columns.tolist()”为定义新列所用的函数，“col_name.insert(7,'time')”确定新列所在位置以及名称，“df['time']=1–df['case']”是新列“time”生成的计算方式，本文中原则是 time=1–case，最后生成新的数据集命名为“base_data”。具体命令如下：

```
import pandas as pd
wb=pd.read_excel(r"D:\ case–crossover.xlsx ", sheet_name="Sheet1")
col_name=wb.columns.tolist()
col_name.insert(7,'time')
df=wb.reindex(columns=col_name)
df['time']=1–df['case']
print(df)
base_data=df
print(base_data)
```

然后调用 lifelines 库中的 CoxPHFitter 函数，“binglimerge.fit(base_data,'time',event_col='case', strata=['ID'])”，依次在括号中放入数据集、生存时间变量、终检变量、分层变量。具体命令如下：

```
from lifelines import CoxPHFitter
binglimerge=CoxPHFitter()
binglimerge.fit(base_data,'time',event_col='case', strata=['ID'])
binglimerge.print_summary()
reults=binglimerge.summary
round(reults,7)
```

2.3 模型验证

R 4.2.1 软件采用 survival 包中的 clogit 函数对条件 Logistic 回归模型进行拟合，对实例资料的分析过程为：

```
library(readxl)
library(survival)
base_data<– read_xlsx("D:/ case–crossover.xlsx ")
options(digits = 8)
mod<–clogit(case ~ no2+temperature+humidity+strata(ID),base_data)
summary(mod)
AIC(mod)
```

SAS 9.1 版本采用 phreg 过程步对条件 Logistic 回归模型进行拟合。其与 Python 类似，在导入数据后，首先需对数据进行预处理，增加新变量 time(time=1–case)。对实例资料的分析过程为：

```
libname orange "D：\sas data";
data base_data;set orange.binglijiaocha;time=1–case;run;
proc phreg data=base_data;
model time*case(0)= temperature humidity no2/ ties=discrete risklimits;strata ID;
quit;
```

2.4 结果比较

Python 和 SAS 在调用数据后，需要通过特定代码运行新增变量“time”，以方便采用分层 Cox 风险比例模型进行拟合。R 则无需进行上述操作，可直接通过 clogit 函数实现模型拟合。与 Python 和 SAS 相比，R 没有默认输出 AIC 值，需另运行“AIC()”函数实现其结果输出。

三款软件输出的主要结果基本相同（表 3）。针对 P 值的检验方法上，R 与 Python 输出参数为 z 值，SAS 输出参数为 χ^2 值，两种检验也是完全等价的（z 值的平方与 χ^2 值相等）。

表3 条件Logistic模型参数输出结果

Table 3. Results of conditional Logistic model parameters

软件	指标	β	z值/χ^2值	P值	OR（95%Cl）	AIC
Python						9513.2636
	NO_2	0.0357	2.7516	0.0059	1.0363（1.0103，1.0630）	
	平均气温	–0.0113	–1.7838	0.0745	0.9888（0.9766，1.0011）	
	相对湿度	0.0002	–0.0589	0.9531	0.9998（0.9941，1.0056）	
R						9513.2636
	NO_2	0.0357	2.7516	0.0059	1.0363（1.0103，1.0630）	
	平均气温	–0.0113	–1.7838	0.0745	0.9888（0.9766，1.0011）	
	相对湿度	0.0002	–0.0589	0.9531	0.9998（0.9941，1.0056）	
SAS						9513.2636
	NO_2	0.0357	7.5715	0.0059	1.0363（1.0103，1.0630）	
	平均气温	–0.0113	3.1819	0.0745	0.9888（0.9766，1.0011）	
	相对湿度	0.0002	0.0035	0.9531	0.9998（0.9941，1.0056）	

3 讨论

在使用 Python 和 SAS 这两款软件时，本研究均用分层 Cox 风险比例模型的运行代码来拟合条件 Logistic 回归模型，而 R 语言则直接运用 survival 包中的 clogit 函数进行拟合，不用另对始变量进行处理。其拟合原理是在分层 Cox 模型中，各层的基线风险函数之间完全无关，而且 Cox 风险比例模型在拟合时并没有估计基线风险函数，只对各协变量的系数值 β 进行了估计，这和条件 Logistic 回归模型只求出系数值 β 的思路一致 [5]。有研究对 Cox 比例风险模型总偏似然函数和条件 Logistic 回归分析的似然函数理论公式进行推导后，发现它们完全等同 [4]。本研究中三款软件均采用极大似然估计法对参数进行估计，其运行结果完全相同，证实了拟合结果的可靠性。在衡量最优模型的标准中，Python 以及 SAS 软件均自动输出 AIC 值，但 R 未自动输出该值，原因是 R 调用的 clogit 函数中不含衡量最优模型标准的相关值的运算代码。另外，三款软件输出参数有 z 值与 χ^2 值的差异，其原因是不同软件的开发人员在统计检验倾向上不同，但 z 值的平方等于 χ^2 值，可以认为 Waldχ^2 检验是等价于 Z 检验的 [6]。

Python 作为一款面向对象的高级编程语言，已经成为最受欢迎的程序设计语言之一，在各行各业都发挥着重要的作用，常用于 Web 应用开发、人工智能、自动化运维、游戏开发等领域，其价值不可估量 [7]。但在统计分析方面，Python 的统计功能相对 R 来说还比较薄弱，其自带的处理功能和函数模型不及 R 齐全，本研究中 Python 得采用 Cox 风险比例模型去拟合条件 Logistic 回归，并且整个运行过程相较另外两款统计软件都更复杂。在增加新变量方面，SAS 的操作步骤比 Python 简洁很多。在可视化方面，Python 拥有 Matplotlib 及 Numpy 等绘图库 [8]，可满足可视化需求。R 作为一款为统计分析而设计的软件，其可视化功能更为强大，它采用简洁的函数就能构建各类图形，并且在默认条件下的绘图品质就能达到出版要求，但是 R 在智能化方面以及非统计分析领域的应用远不及 Python。

综上所述，将 Python 应用于统计分析领域，凭借其丰富的第三方库以及快速运算大数据的优势，能大大提高数据的智能化处理与分析效率。本研究使用 Python 软件实现了条件 Logistic 回归的统计建模，在实际研究中有一定的参考价值。

参考文献

1 张政，詹思延 . 病例交叉设计 [J]. 中华流行病学杂志，2001, 22(4): 70–72. [Zhang Z, Zhan SY. Case crossover design[J]. Chinese Journal of Epidemiology, 2001, 22(4): 70–72.] DOI: 10.3760/j.issn:0254–6450.2001.04.022.

2 郑一男，曹佩华，欧春泉 . N: M 条件 logistic 回归分析在统计软件上的实现 [J]. 中国卫生统计，2011, 28(1): 93–94, 97. [Zheng YN, Cao PH, Ou CQ. Implementation of N: M conditional logistic regression analysis in statistical software[J]. Chinese Journal of Health Statistics, 2011,

28(1): 93–94, 97.] DOI: 10.3969/j.issn.1002–3674.2011.01.034.

3 孙振球 , 王勇勇 . 医学统计学 , 第 4 版 [M]. 北京 : 人民卫生出版社 , 2014.

4 张业武 . Cox 比例风险模型对条件 logistic 回归参数估计原理和方法 [J]. 中国卫生统计 , 2002, 19(1): 23–25. [Zhang YW. Principle and method of conditional Logistic regression parameter estimation by Cox proportional risk model[J]. Chinese Journal of Health Statistics, 2002, 19(1): 23–25.] DOI: 10.3969/j.issn.1002–3674.2002.01.008.

5 孙中华 , 王梅 . Cox 模型处理条件 Logistic 回归考察升主动脉压力波谷峰值与冠心病的相关性 [J]. 数理医药学杂志 , 2004, 17(1): 80–82. [Sun ZH, Wang M. Logistic Regression study about the correlation between prissure wave trough peak of ascending aorta and coronary heart disease in cox model[J]. Journal of Mathematical Medicine, 2004, 17(1): 80–82.] DOI: 10.3969/j.issn.1004–4337.2004.01.041.

6 焦奎壮 , 马煦晰 , 马小茜 , 等 . 广义估计方程与混合线性模型在 Python 中的实现 [J]. 医学新知 , 2022, 32(5): 333–338. [Jiao KZ, Ma XX, Ma XQ, et al. Implementation of generalized estimating equation and mixed linear models in Python[J]. New Medicine, 2022, 32(5): 333–338.] DOI: 10.12173/j.issn.1004–5511.202203007.

7 平凯珂 , 陈平雁 . Python 与 R 语言联合应用的实现 [J]. 中国卫生统计 , 2017, 34(2): 358–360. [Ping KK, Chen PY. Implementation of python and R language combined application[J]. Chinese Journal of Health Statistics, 2017, 34(2): 358–360.] DOI: CNKI:SUN:ZGWT.0.2017–02–054.

8 李天辉 . 基于 python 的数据分析可视化研究与实现 [J]. 电子测试 , 2020, (20): 78–79. [Li TH. Research and implementation of visualized data analysis based on python[J]. Electronic Test, 2020, (20): 78–79.] DOI: 10.3969/j.issn.1000–8519.2020.20.030.

收稿日期：2023 年 02 月 08 日 修回日期：2023 年 02 月 26 日
本文编辑：李 阳 黄 笛

引用本文：张清煜, 李培政, 罗晨曦, 等. 病例交叉研究中条件Logistic回归在Python中的实现[J]. 数理医药学杂志, 2023, 36(5): 321–325. DOI: 10.12173/j.issn.1004–4337.202302031
Zhang QY, Li PZ, Luo CX, et al. Implementation of conditional Logistic regression in case–crossover study in Python[J]. Journal of Mathematical Medicine, 2023, 36(5): 321–325. DOI: 10.12173/j.issn.1004–4337.202302031