北京航空航天大學
BEIHANG UNIVERSITY

数理统计大作业

（一）

姓名：范嘉楠
学号：SY2006329
班级：21 班
序号：90 号

摘要

由于今年新冠疫情的影响，全球各大足球联赛纷纷延期或是空场举行，这使得 2019-2020 这个赛季显得有些支离破碎。那么，这个赛季的数据是否还具有统计学上的参考价值？球队获胜的概率是否仍然像某些评论员所述，与一些具体的数据指标线性相关？

本文以 19-20 的英超联赛做分析，探究落魄豪门阿森纳队在上赛季中的比赛成绩与比赛数据之间的关系。具体地，由于进球数这一指标差异性不大，我们以 Bet365、William Hill 等知名足彩所开出的盘口数据作为球队的获胜概率，作为因变量；考察其与射门数、射正数、犯规数等 7 项指标的线性相关性，将以上指标作为自变量。

进一步地，本文利用了统计软件 SPSS，使用逐步回归法对以上数据进行线性回归，最终筛选出了与因变量相关性最大的几个因素，并确定了获胜概率与这些因素之间的最优回归方程。

关键词：多元线性回归，逐步回归法，英超，阿森纳，SPSS

目录

一、引言

传统的足球观点认为，一支球队能否取得比赛的胜利，与其在场面中的表现息息相关：如果这支球队对比赛有着绝对的掌控，那么他应该在控球率、射门数、角球数等诸多指标中占优，而如果一支球队拿到了红牌、被判点球或是客场作战，那么他们则大概不会被人看好，这是非常合情合理的。然而，随着足球比赛技战术水平的提高，许多龟缩防守、高效反击（俗称摆大巴）的球队，往往也可以获得比赛的胜利，也有一些球队采用肉搏、手榴弹等战术，通过大量的犯规阻拦对方的进攻节奏，通过大力手抛球、角球等方式获得比赛胜利，这些现象都冲击着传统的足球观点，让大家对“胜利的方程式”有了不同的认知。

在英国知名足球数据网站 football-data 中，本文选取了 2019-2020 赛季，阿森纳的全部 38 场比赛数据，其中分别在主客场对阵了 19 个不同的对手，选取了其中：是否客场、射门数、射正数、犯规数、角球数、黄牌数、红牌数作为自变量，以 Bet365、William Hill 等知名足彩所开出的盘口数据作为球队的获胜概率，并作为因变量，探究了两者之间的最优多元线性回归方程。

所以，这样的工作是非常有价值的，它能够指导球队，为球队提供“胜利的方程式”，尝试赢得比赛；也可以给博彩公司所开盘口给出指导，通过这些数据和回归方程，预测比赛的走势。

在本文的第二章中将给出所收集数据和相关变量的符号；在第三章中将介绍逐步回归法，并使用统计软件对以上变量进行分析和回归；第四章中，将会对输出的结果进行讨论，探究 F 检验的阈值、回归的模型是否合理，以及对实际问题的参考意义；第四章中将给出本文所参考的文献。

二、数据收集与整理

2.1 数据处理

英国知名足球数据网站 football-data[1]提供了英超历年的比赛数据，在这里我们筛选出所有和阿森纳有关的比赛数据，并相应地提取阿森纳的射门数、射正数等六项数据，并对是否客场做出标记。

接下来对数据进行整理。在这里，我们舍弃了进球数、裁判等强相关或是不好处理的数据。对于因变量，这里我们对所开盘口做简单地归一化处理，即可得到获胜概率。

例如，某场比赛种，各博彩公司的平均盘口为，主队胜 1 赔 a；两队打平 1 赔 b；客队胜 1 赔 c。博彩公司认为主队胜的赔付程度，即代表其认为客队获胜的程度，那么设主队获胜的概率为 x，两队打平的概率为 y，客队获胜的概率为 z，则可以得到归一化公式：

$$x = \frac{c}{a+b+c} \tag{2-1}$$

$$y = \frac{b}{a+b+c} \tag{2-2}$$

$$z = \frac{a}{a+b+c} \tag{2-3}$$

由此可以简单地计算出博彩公司认为的队伍获胜概率。

此外，我们认为在任何一个客场作战的心理状态都是相同的（虽然如老特拉福德、安菲尔德等客场都是出了名的魔鬼客场），可以简单地以 0-1 二值变量表示是否为客场。

最终得到全 38 轮的数据表格如下：

表 2.1 19-20 赛季阿森纳的比赛数据和获胜概率

场次	是否客场	射门数	射正数	犯规数	角球数	黄牌数	红牌数	获胜概率（百分之）
1	1	8	2	7	3	3	0	44.455
2	0	16	9	13	10	2	0	58.160
3	1	9	3	5	4	1	0	12.146
4	0	26	8	13	11	3	0	32.548
5	1	10	4	4	1	3	0	38.765
6	0	21	6	13	9	5	1	54.236
7	1	10	5	13	7	2	0	26.577
8	0	12	2	12	14	1	0	50.992
9	1	9	3	12	12	4	0	40.417
10	0	15	6	18	12	2	0	51.467
11	0	10	4	6	8	0	0	45.455
12	1	8	1	10	4	1	0	21.193
13	0	12	5	13	6	6	0	51.292
14	1	16	7	10	12	1	0	40.869
15	0	12	5	10	9	3	0	48.089
16	1	10	3	6	3	0	0	36.373
17	0	6	1	9	3	1	0	11.319
18	1	6	2	11	4	3	0	25.947
19	1	17	2	13	3	4	0	37.393
20	0	7	2	13	2	5	0	26.286
21	0	10	4	11	1	2	0	29.558
22	1	7	4	22	4	3	1	42.366
23	0	11	4	9	4	1	0	44.960
24	1	2	2	6	5	1	1	16.731
25	1	13	2	11	7	3	0	40.323
26	0	15	7	15	5	2	0	56.145

27	0	9	4	12	6	0	0	39.159
28	0	9	2	11	6	1	0	49.360
29	1	3	0	7	2	1	1	9.024
30	1	13	6	8	7	1	0	34.545
31	1	10	5	14	6	2	0	30.137
32	0	13	8	10	6	1	0	54.117
33	1	8	5	11	5	4	0	25.534
34	0	11	7	10	10	1	1	34.276
35	1	13	4	11	5	3	0	29.680
36	0	3	2	14	2	3	0	19.545
37	1	7	0	19	9	4	0	35.487
38	0	13	5	9	4	3	0	38.994

2.2 建立模型

由此，可以为变量定义符号，如下表所示：

表 2.2 各变量符号说明

变量名	符号
是否客场	x_1
射门数	x_2
射正数	x_3
犯规数	x_4
角球数	x_5
黄牌数	x_6
红牌数	x_7

故该问题的线性回归模型应该为

$$\left.\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \\ E(\varepsilon) = 0, \mathrm{Var}\,(\varepsilon) = \sigma^2 < +\infty \end{aligned}\right\} \tag{2-4}$$

得到最优（最小二乘估计）线性回归模型应为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{2-5}$$

其中 p 为经过逐步回归法后得到的模型中变量个数[2]。

三、解决问题的方法和计算结果

3.1 逐步回归法

采用逐步回归法进行分析。该方法是从一个自变量开始，视自变量对 y 的显著程度，从

大到小地依次逐个引入回归方程，但当引入的自变量由于后面的引入而变得不显著时，要将其剔除掉。引入一个自变量或从回归方程中剔除一个自变量，为逐步回归的一步，对于每一步都要进行 F 值检验，以确保每次引入新的显著性自变量前回归方程中只包含对作用显著的变量。这个过程反复进行，直至既无不显著的变量从回归方程中剔除，又无显著变量可引入回归方程时为止[3]。

本文运用统计软件 SPSS 的有关功能模块[4]根据上述原理选出一些自变量组建回归方程。为了从挑选因子中筛选出尽可能多的因子建立模型，在 SPSS 中，默认的 F 值为≤ 0.05时输入，≥ 0.1时拒绝。如果入选的自变量因子数目不多，可通过人为提高 F 临界值的引入水平而筛选出更多的因子。如此时入选的因子太多，可人为降低 F 临界值的剔除水平而筛选出有代表性因子来组建模型。如最后建立的模型的复相关系数不大，回归模型的拟合精度不太高，可根据这些入选因子来组建多元非线性模型。

该分析在默认的$F_{in} = 0.05, F_{rej} = 0.1$时效果较差，因此提高 F 值，设定$F_{in} = 0.4, F_{rej} = 0.5$，进行多元线性回归分析，可得计算结果，见下一小节内容。

3.2 计算结果

输入/除去的变量[a]

模型	输入的变量	除去的变量	方法
1	阿森纳射门数	.	步进（条件：要输入的 F 的概率 <= .400，要除去的 F 的概率 >= .500）。
2	是否客场	.	步进（条件：要输入的 F 的概率 <= .400，要除去的 F 的概率 >= .500）。
3	阿森纳角球数	.	步进（条件：要输入的 F 的概率 <= .400，要除去的 F 的概率 >= .500）。
4	阿森纳犯规数	.	步进（条件：要输入的 F 的概率 <= .400，要除去的 F 的概率 >= .500）。
5	阿森纳射正数	.	步进（条件：要输入的 F 的概率 <= .400，要除去的 F 的概率 >= .500）。

a. 因变量：获胜概率（百分之）

模型摘要[f]

模型	R	R 方	调整后 R 方	标准估算的错误
1	.563[a]	.317	.298	10.883596102982970
2	.626[b]	.391	.357	10.416095794508394
3	.661[c]	.437	.388	10.160610749158876
4	.673[d]	.453	.387	10.167151622733352
5	.684[e]	.468	.385	10.181844246439177

a. 预测变量：(常量)，阿森纳射门数

b. 预测变量：(常量)，阿森纳射门数，是否客场

c．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数
d．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数，阿森纳犯规数
e．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数，阿森纳犯规数，阿森纳射正数
f．因变量：获胜概率（百分之）

ANOVA[a]

模型		平方和	自由度	均方	F	显著性
1	回归	1975.426	1	1975.426	16.677	.000[b]
	残差	4264.296	36	118.453		
	总计	6239.722	37			
2	回归	2442.395	2	1221.198	11.256	.000[c]
	残差	3797.327	35	108.495		
	总计	6239.722	37			
3	回归	2729.630	3	909.877	8.813	.000[d]
	残差	3510.092	34	103.238		
	总计	6239.722	37			
4	回归	2828.480	4	707.120	6.841	.000[e]
	残差	3411.242	33	103.371		
	总计	6239.722	37			
5	回归	2922.284	5	584.457	5.638	.001[f]
	残差	3317.438	32	103.670		
	总计	6239.722	37			

a．因变量：获胜概率（百分之）
b．预测变量：(常量)，阿森纳射门数
c．预测变量：(常量)，阿森纳射门数，是否客场
d．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数
e．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数，阿森纳犯规数
f．预测变量：(常量)，阿森纳射门数，是否客场，阿森纳角球数，阿森纳犯规数，阿森纳射正数

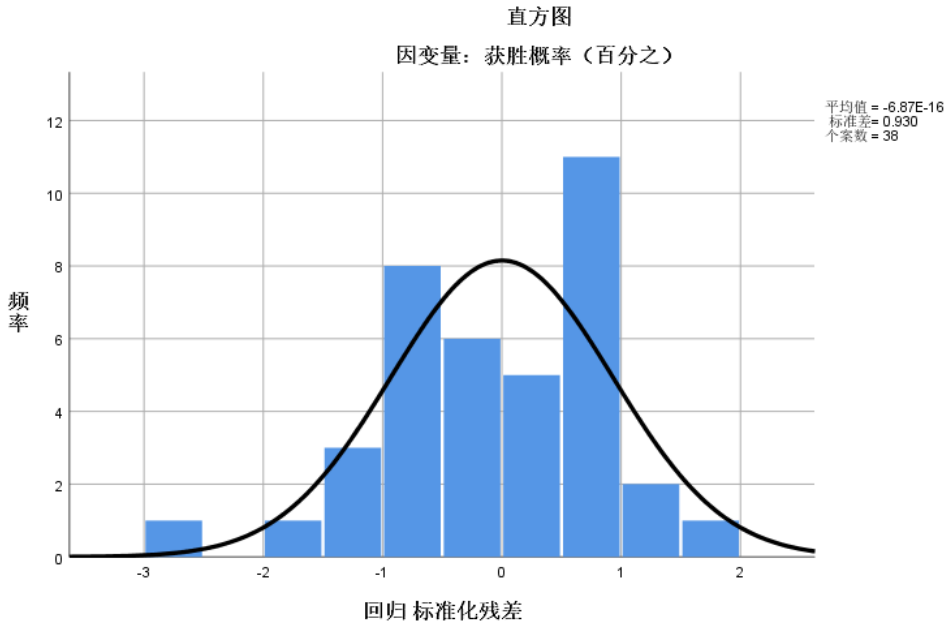
系数[a]

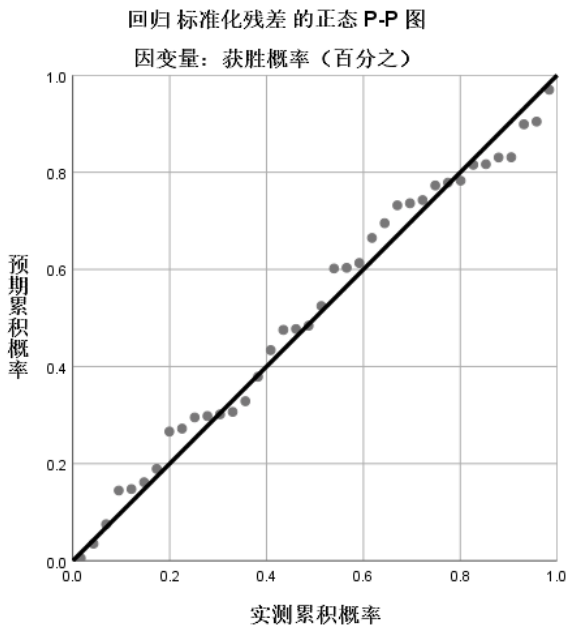
模型		未标准化系数		标准化系数	t	显著性
		B	标准错误	Beta		
1	(常量)	19.676	4.464		4.408	.000
	阿森纳射门数	1.552	.380	.563	4.084	.000
2	(常量)	25.853	5.207		4.965	.000
	阿森纳射门数	1.319	.381	.478	3.467	.001
	是否客场	-7.336	3.536	-.286	-2.075	.045
3	(常量)	23.492	5.273		4.455	.000
	阿森纳射门数	.986	.422	.357	2.337	.025
	是否客场	-6.996	3.456	-.273	-2.025	.051

	阿森纳角球数	.953	.571	.249	1.668	.104
4	(常量)	19.129	6.910		2.768	.009
	阿森纳射门数	.980	.422	.355	2.323	.026
	是否客场	-6.683	3.473	-.261	-1.924	.063
	阿森纳角球数	.815	.589	.213	1.385	.175
	阿森纳犯规数	.460	.471	.132	.978	.335
5	(常量)	18.046	7.013		2.573	.015
	阿森纳射门数	.698	.517	.253	1.350	.186
	是否客场	-5.900	3.574	-.230	-1.651	.109
	阿森纳角球数	.742	.594	.194	1.248	.221
	阿森纳犯规数	.479	.472	.138	1.017	.317
	阿森纳射正数	1.000	1.051	.175	.951	.349
a. 因变量：获胜概率（百分之）						

残差统计[a]

	最小值	最大值	平均值	标准偏差	个案数
预测值	19.079030990600586	58.578617095947266	36.418905556479310	8.887105379702467	38
残差	-26.030916213989258	19.146696090698242	-.000000000000007	9.468922813647174	38
标准预测值	-1.951	2.493	.000	1.000	38
标准残差	-2.557	1.880	.000	.930	38
a. 因变量：获胜概率（百分之）					





四、讨论

从残差的直方图和正态 P-P 图中可以看出预期-实测累积概率的确是一个接近于 y=x 的直线，回归效果较好。

从系数表格中可以看出，在多元线性回归后计算出的最优模型（按显著性从小到大排列）表达式为：

$$y = 18.046 + 0.698x_2 - 5.900x_1 + 0.742x_5 + 0.479x_4 + 1.000x_3 \quad (4\text{-}1)$$

其中x_2为射门数、x_1为是否客场、x_5为角球数、x_4为犯规数、x_3为射正数。

从实际情况考察这个拟合式是十分有趣的。

首先，影响显著性最小（置信度最高）的因素为射门数，这符合传统足球的认知，也应了那句老话——多射门，总是会有进球的，然而其权重系数只有 0.698，也符合现代足球中频频出现“雷声大雨点小”的情况。

其次，是否客场是显著性第二小的因素，并且足足有着-5.9 的权重加成。也就是说，如若客场作战，阿森纳的获胜概率就降低了 6 个百分点！由于自变量中未加入“是否主场”，所以这一点应该是针对主场作战而言的，也就是说主、客场作战对于胜率的影响的浮动大概是 3 个百分点上下。

再次，角球数的正增益权重系数超过了射门数（虽然其置信度不高），但是这也符合英超评论员的认知：阿森纳在上个赛季是角球、定位球得分较多的球队，这也得益于其后卫的身高和前锋攻击群的头球能力。

令人意料之外、情理之中的是，犯规对于胜率的权重系数居然是正的，但这似乎也比较合理。阿森纳的球员整体年龄偏老，非常怕对方的反击，如果能够通过犯规扰乱对方的进攻节奏，就可以大大降低控制比赛的成本，从而最终取得比赛的胜利。

最后一项射正数的置信度不高，但系数较大，笔者认为主要是因为射正数在一场比赛中本就较少，而且一旦射正就意味着更加接近进球，也更加看对方门将的发挥，偶然性比较大。系数高但并没有对场面给出更多的贡献，是其置信度较小的原因。

五、参考文献

[1]. Football-Data. https://www.football-data.co.uk/englandm.php

[2]. 孙海燕，周梦，李卫国，冯伟. 数理统计[M]. 北京：北京航空航天大学应用数学与系统科学学院，2015: P160-P171

[3]. Stepwise regression, https://en.wikipedia.org/wiki/Stepwise_regression (last visited Dec. 2, 2020).

[4]. Linear Regression, IBM Knowledge Center. https://www.ibm.com/support/knowledgecenter/zh/SSLVMB_25.0.0/statistics_mainhelp_ddita/spss/base/idh_regs.html