数理统计大作业 II

职业球员能力回归分析

11 班　　许文哲　　SY2008111

目录

1 任务介绍

任务要求

应用回归方法解决一个实际问题

自己选择一个实际问题，收集数据，使用回归方法解决。

注意：所使用的回归方法不限于课堂讲过的线性回归方法，也可以使用其他回归方法，如曲线回归方法，广义回归方法等。

任务目标

通过网络收集 2017 年世界球坛职业球员的数据，并进行如下分析。

1. 对球员的评分和各项能力水平进行**多元线性回归分析**，通过**逐步回归**或 **lasso** 的方法对变量进行筛选，解决多重共线性的问题，寻找对球员评分影响最大的能力属性，基于此，对球员评分进行预测。

2. 对球员的场上位置和各项能力水平进行**多项 logistic 回归**，通过**逐步回归**或 **lasso** 的方法对变量进行筛选，建立模型，对球员的场上位置进行预测。

方法实现

本报告的数据分析使用 R 语言实现，本报告的编写基于 RMarkdown，代码详见 footballplayer.md。

- 多元线性回归：lm 函数
- 多项 logistic 回归：nnet 包中的 multinom 函数
- 逐步回归：step 函数
- lasso：glmnet 包中的 glmnet 函数

2 描述性统计

2.1 数据来源

数据由热心网友 Ustinian 提供。

2.2 数据介绍

数据集包含了 2017 年世界足坛职业比赛范围内活跃的足球运动员的能力数据，包含 53 个属性，共 17588 条数据，表 1 给出了数据的各项属性的说明。数据较为全面地包含了 17588 位职业球员的基本信息、生理数据和足球能力数据，可以反映球员能力。

表 1: 数据说明

index	name	translation
1	Name	姓名
2	Nationality	国籍
3	National_Position	国家队位置
4	National_Kit	国家队号码
5	Club	所在俱乐部
6	Club_Position	所在俱乐部位置
7	Club_Kit	俱乐部号码
8	Club_Joining	加入俱乐部时间
9	Contract_Expiry	合同到期时间
10	Rating	评分
11	Height	身高
12	Weight	体重
13	Preffered_Foot	擅长左（右）脚
14	Birth_Date	出生日期
15	Age	年龄
16	Preffered_Position	擅长位置

index	name	translation
17	Work_Rate	工作效率
18	Weak_foot	非惯用脚使用频率
19	Skill_Moves	技术等级
20	Ball_Control	控球技术
21	Dribbling	盘球（带球）能力
22	Marking	盯人能力
23	Sliding_Tackle	铲球
24	Standing_Tackle	逼抢能力
25	Aggression	攻击能力
26	Reactions	反击
27	Attacking_Position	攻击性跑位
28	Interceptions	抢断
29	Vision	视野
30	Composure	镇静
31	Crossing	下底传中
32	Short_Pass	短传
33	Long_Pass	长传
34	Acceleration	加速度
35	Speed	速度
36	Stamina	体力
37	Strength	强壮
38	Balance	平衡
39	Agility	敏捷度
40	Jumping	跳跃
41	Heading	投球
42	Shot_Power	射门力量
43	Finishing	射门
44	Long_Shots	远射
45	Curve	弧线
46	Freekick_Accuracy	任意球精准度
47	Penalties	点球
48	Volleys	凌空能力
49	GK_Positioning	门将位置感
50	GK_Diving	扑救能力
51	GK_Kicking	门将踢球能力
52	GK_Handling	扑球脱手几率
53	GK_Reflexes	门将反应度

2.3 描述性统计

2.3.1 球员主要数据

表 2: 球员基础数据

	评分	身高/cm	体重/kg	年龄
Min.	45.0	155.0	48.0	17.0
1st Qu.	62.0	176.0	70.0	22.0
Median	66.0	181.0	75.0	25.0
Mean	66.2	181.1	75.3	25.5
3rd Qu.	71.0	186.0	80.0	29.0
Max.	94.0	207.0	110.0	47.0

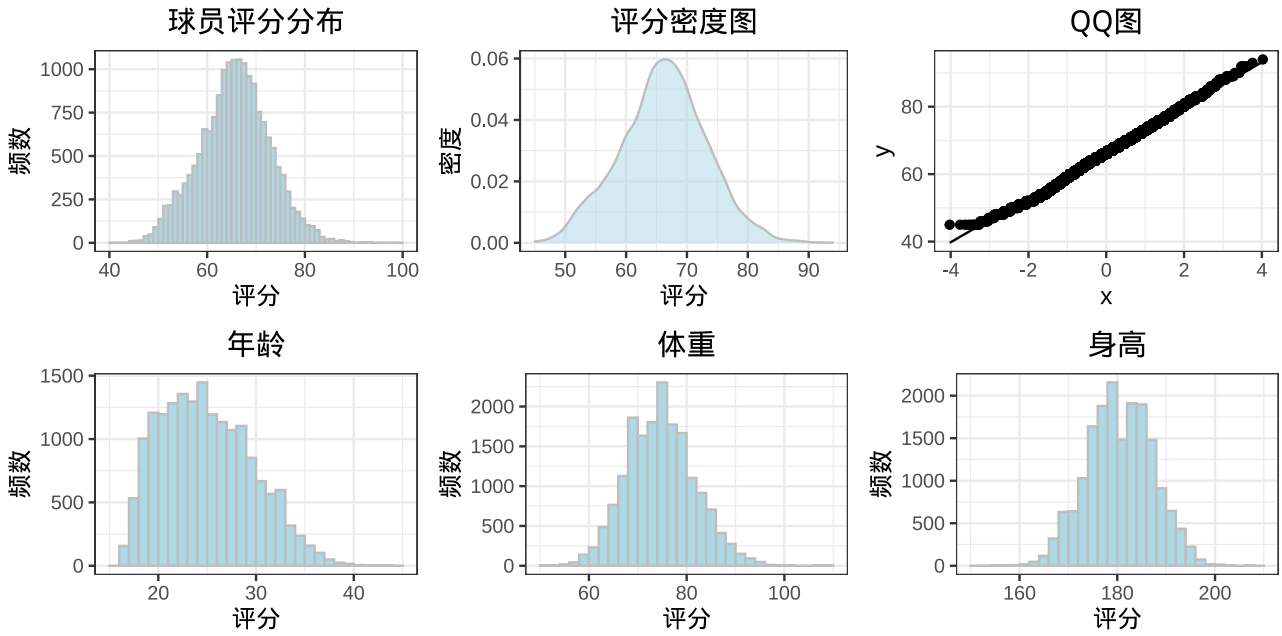

图 1: 球员基本数据

2.3.2 球员惯用脚

表 3: 球员惯用脚情况

惯用脚	频数
Left	4094
Right	13494

2.3.3 球员擅长位置

Preffered_Position 这一属性给出了各位球员的擅长位置，如下表所示。

表 4: 球员擅长位置

简称	全称	含义	分类	频数
CAM	Central Attacking Midfielder	进攻型中场	middle	1098
CB	Centre-back	中后卫	backward	2936
CDM	Central Defensive Midfielder	防守型中场	middle	1480
CF	Centre-forward	中锋	forward	68
CM	Central Midfielder	中场	middle	1982
GK	Goalkeeper	守门员	goalkeeper	2003
LB	Left-back	左后卫	backward	1376
LM	Left Midfielder	左中场	middle	957
LW	Left Winger	左边锋	forward	302
LWB	Left-Wing-back	左翼卫	backward	23
RB	Right-back	右后卫	backward	1392
RM	Rightl Midfielder	右中场	middle	1041
RW	Right Winger	右边锋	forward	351
RWB	Right Wing-back	右翼卫	backward	22
ST	Striker	前锋	forward	2557

2.3.4 球员各项能力

球员各项能力分布如图 2 所示。

2.4 数据预处理

2.4.1 缺失值处理

经过检查，数据的缺失值主要集中在 National_Position、National_Kit、Club、Club_Position、Club_Kit、Club_Joining 和 Contract_Expiry 这几个属性中，这很容易理解：如果球员能力水平不足以达到被国家队征召的水平，那么他的 National_Position（国家队位置）和 National_Kit（国家队球衣号

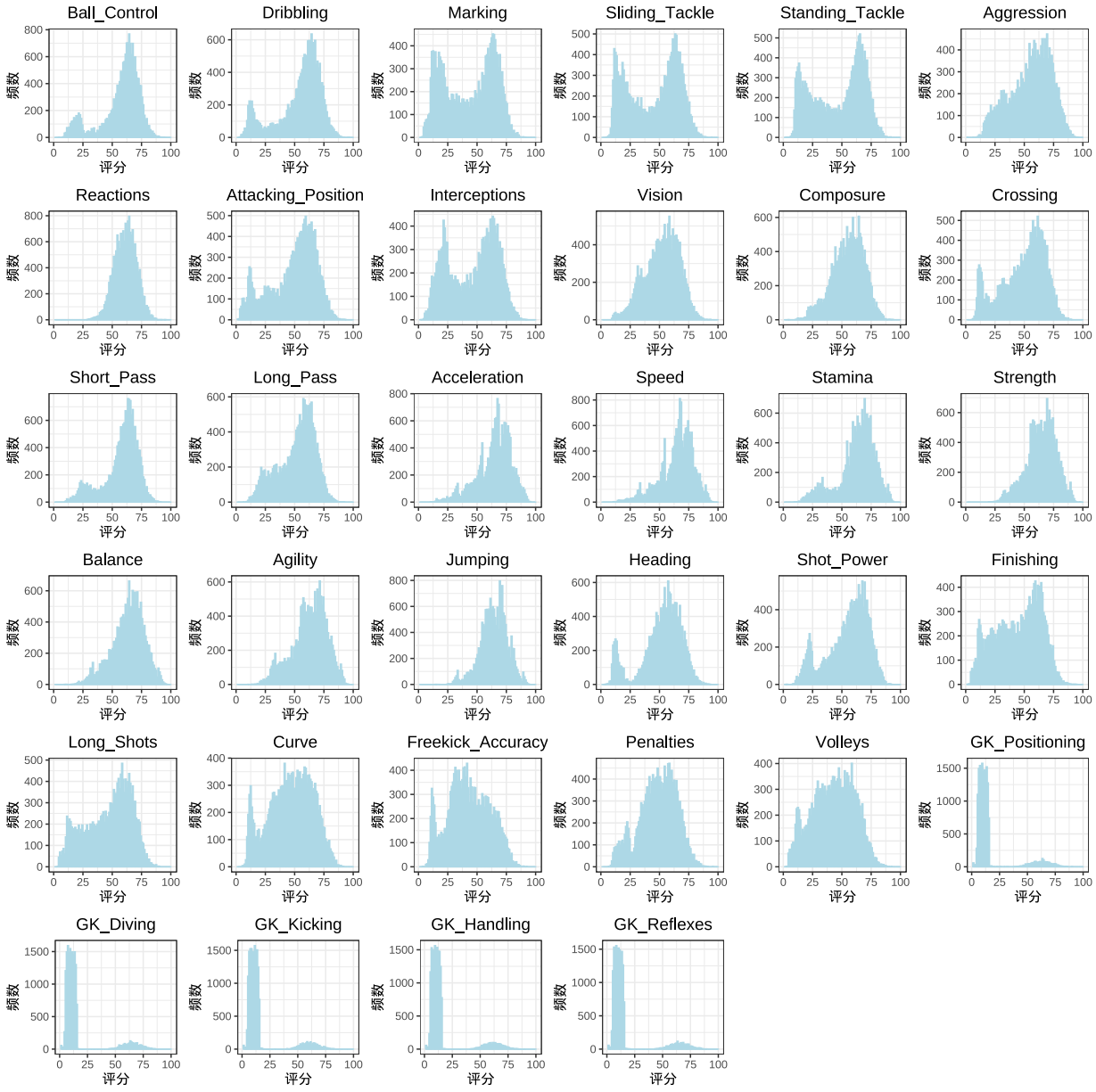
Ball_Control
Dribbling
Marking
Sliding_Tackle
Standing_Tackle
Aggression
Reactions
Attacking_Position
Interceptions
Vision
Composure
Crossing
Short_Pass
Long_Pass
Acceleration
Speed
Stamina
Strength
Balance
Agility
Jumping
Heading
Shot_Power
Finishing
Long_Shots
Curve
Freekick_Accuracy
Penalties
Volleys
GK_Positioning
GK_Diving
GK_Kicking
GK_Handling
GK_Reflexes
频数
评分

图 2: 球员各项能力分布

码）便是缺失值；同理，如果球员在 2017 年处于自由转会状态，即球员没有与任何一家俱乐部签约，那么他的 Club、Club_Position、Club_Kit、Club_Joining 和 Contract_Expiry 这几个俱乐部相关的属性便是缺失值。因此，对于以上几个属性，我们予以删除处理。

2.4.2 其他处理

1. 删除属性

- National_Position、National_Kit、Club、Club_Position、Club_Kit、Club_Joining、Contract_Expiry：理由已阐述如上。
- Birth_Date：出生日期，我们可以从 Age 获取相关信息，因而删除。
- Work_Rate：工作效率，与球员与更衣室的关系、球员性格有关，对于球员能力评分的影响较小。
- Preffered_Foot：擅长左（右）脚，目前鲜有研究表明左右脚会明显影响球员能力。
- Weak_foot：非惯用脚使用频率，理由同上。
- Skill_Moves：技术等级，反映了后面多个百分制属性的综合情况，与后面的变量具有强烈的相关性，且我们对于如此综合性的指标不感兴趣。

2. 变量整合

多数球员能力特点比较全面，擅长位置有多个。例如：C · 罗纳尔多擅长的位置有 LW 和 ST，但都属于前场；德布劳内擅长的位置有 CAM/RM/LM 都属于中场；迪巴拉擅长的位置有 ST 和 CAM，前者属于前场，后者属于中场。对于这种情况，我们仅保留球员最擅长的位置，即第一个位置，并按照 forward、middle、backward 和 goalkeeper 进行合并分类。

3 回归分析

3.1 门将位置评分

由于门将位置的特殊性，其相关能力已经很明显地标注在数据集中（GK_ 开头的属性），而且门将的位置与场上其他位置的球员能力分布存在肉眼可见的差距，门将更多依靠手部能力、站位、反应能力和定位球指挥能力，对于速度、传球、逼抢、盯人、铲球等能力的要求相对较低（尽管部分阵型很看重门将的长传能力，但其对于其长传能力的要求仍大幅低于中场球员或后卫）。因此，我们首先对门将进行回归分析。

3.2 前场球员评分

3.3 中场球员评分

3.4 后场球员评分

3.5 球员位置判别