基于 Logistic 回归的球员体能评价模型

摘　要

本文基于题目给出的体能数据的基础上，对数据进行了验证与处理，并根据测试标准在 15 个球队中进行了球员的挑选，分析了此测试标准的合理性。同时，采用了 Logistic 回归的方法，针对不同位置的球员修改了权重表，论证了其合理性。最后改进和提高了对球员体能测试的相关参数。该问题的研究能为球员的发展提供一定的指导。

针对**问题 1.** 根据附件中的体能测试数据，通过 KS 检验筛选出数据中的奇异值，之后借助多元线性回归的方法，对这些奇异值进行处理。同时，对于表格中缺失并且无法预估的值，取数据的均值进行填充，最后获得较为完整的体能测试数据。

针对**问题 2.** 在问题 1. 得到的较为完整的体能测试数据的基础上，通过计算分位数并进行检验的方法，按所有参赛球员的 85%作为合格线，确定最终体能综合得分的及格线，并筛选出无法挑出 22 人的球队名单。

针对**问题 3.** 我们找到一组具有球员位置以及各项身体指标及技能指标的原始数据集。将不同位置的球员数据分类进行讨论，通过方差分析的办法，选择最能评价某一位置球员好坏的 25%的变量进行训练，通过 Logistic 回归的方法，依照 SPSS 软件分析出的显著性指标（置信区间中的 p 值）以及β系数的大小，确定出最合适的评分权重，并通过交叉验证的方法对其进行验证。

针对**问题 4.** 将原始数据集中的球员进行分类，分为不同年龄阶段（<25,25-30,>30）以及优秀与不优秀的球员，对其进行 Logistic 回归分析，根据 python 的 sklearn 库得出相应的β系数，依据β系数的大小制定符合不同年龄阶段的评价指标，总结出在不同职业发展时期，某一位置的球员需要在哪些方面进行提高。

经过分析验证，本文的模型具有合理性和一定的现实意义。最后，我们总结了本模型的优点和缺点，并提出了将来可能采用的改进方法（岭回归）。

关键词：Logistic 回归　多元线性回归　体能评价　量化分析

一、 问题重述

1.1 问题背景

体能是评价一个职业足球运动员身体素质的关键指标，在组建足球队时，往往依据队员的体能水平挑选出合适的人选组成正式的比赛球队。从 15 个足球赛队，每个队提供了最多 35 名候选运动员的体测数据,需要从中挑选出合格的 22 人组成正式的比赛球队，需要解决以下四个问题。

1.2 问题提出

问题 1. 已知评分权重表，给出每位运动员的最终体能综合得分, 处理其中缺失以及可能不一致的数据；给出每个球队中不合格队员（体能综合得分小于 6 分）的名单及成绩；计算每个球队的综合评分。

问题 2. 规定每个球队最多可以有 35 个队员参加体能测试，通过体能测试的队员最多可以有 22 个队员参加最终比赛名单。若按所有参赛球员的 85%作为合格线，最终体能综合得分的及格线应设为多少比较合理？哪些球队可能无法挑选出 22 人?

问题 3. 对于不同场上位置的球员，问题 1. 中的权重表是否合理？对于守门员这个体测标准是否过于严格?能否针对不同位置的球员，根据他们位置的特征设置合理的评分权重？并说明其合理性。

问题 4. 从球员成长的角度出发，说明目前的测试方式是否可以作为球员体能的测试标准？这个测试方法及其中的参数是否需要改进和提高, 以及如何改进和提高?

二、 问题分析

2.1 问题 1. 的分析

根据附件中的体能测试数据，通过 Kolmogorov-Smirnov 检验筛选出数据中的奇异值，之后借助多元线性回归的方法，对这些奇异值进行处理。同时，对于表格中缺失并且无法预测估计的值，取数据的均值进行填充，最后获得较为完整的体能测试数据。

2.2 问题 2. 的分析

在第一问得到的较为完整的体能测试数据的基础上，通过计算分位数并进行检验的方法，按所有参赛球员的 85%作为合格线，确定最终体能综合得分的及格线，并筛选出无法挑出 22 人的球队名单。

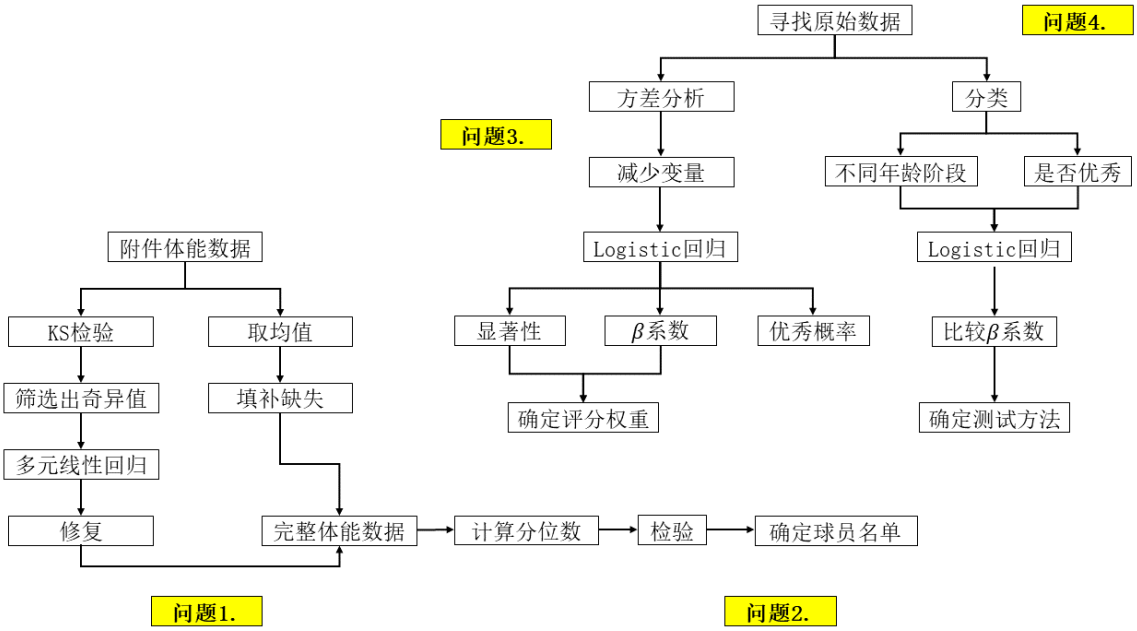
2.3 问题 3. 的分析

通过对原始数据的寻找，找到一组具有球员位置以及各项身体指标以及技能指标的原始数据，通过方差分析的办法，先对每个位置球员的评价指标进行第一轮筛选，，其次，根据每个位置是否优秀，通过 Logistic 回归的方法，依照 SPSS 软件分析出的显著性指标（置信区间中的 p 值）以及β系数的大小，确定出最合适的评分权重。

2.4 问题 4. 的分析

将原始数据集中的球员进行分类，分为不同年龄阶段以及优秀与不优秀的球员，对其进行 Logistic 回归分析，得出相应的β系数，依据β系数的大小制定符合不同年龄阶段的评价指标。

具体思路图如下：



图片 1

三、 模型假设

1. 每名球员的各项体侧数据间存在相关性，即我们可以通过某队员的其他体侧数据推出他体侧数据中的异常值或缺失值。
2. 每名球员的体侧数据不可能全为异常值，即每名队员的体侧数据中大多为正常值，异常值只占少数，且主要集中于引体向上这一项。
3. 相同职位的球员体能情况较为接近，即在预测缺失值或填补异常值时，将相同职位的其他球员的体侧数据作为自变量与因变量。
4. 球员某一项体能数据服从正态分布。

四、 符号说明

符号	说明	单位
pvalue	KS 检验中的显著水平	——
R^2	多元线性回归拟合程度	——
SSE	多元线性回归中的残差平方和	——
β_0	Logistic 回归中自变量对应的系数	——
$F(x,\beta)$	Logistic 回归构造的连接函数	——
C	Logistic 回归的损失函数的值	——
F	方差分析中的检验统计量	——

五、 模型的建立与求解

5.1 问题 1. 模型的建立与求解

模型的建立

5.1.1 检验数据是否符合正态分布

Kolmogorov-Smirnov（KS）检验用于确定样本是否来自具有特定分布的人群，可以修改 KS 检验以作为拟合度检验，在测试分布正态性的特殊情况下，将样本标准化X并与标准正态分布γ进行比较。

假设 F(x)和 G(x)为连续分布函数，用统计量：

$$\mathrm{D} = \max_{ij}\{|F_m(X_{(i)}) - G_n(\gamma_{(j)})|\} \tag{1}$$

来检验上面的假设问题，其中$F(x)$和$G(x)$分别为X样本和γ样本的顺序统计量，m，n 表示样本数，统计量 D 所对应的显著水平 pvalue 如下表示：

$$\mathrm{pvalue} = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \tag{2}$$

其中：

$$\lambda = \left[\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}}\right]D \tag{3}$$

5.1.2 构建多元线性回归模型预测存在异常的数据

研究在线性关系相关性条件下，两个或者两个以上自变量对一个因变量，为多元线性回归分析，表现这一数量关系的数学公式，称为多元线性回归模型。多元线性回归模型是一元线性回归模型的扩展，其基本原理与一元线性回归模型类似，只是在计算上为复杂需借助计算机来完成。

计算公式如下：

设随机 y 与一般变量 $x_1, x_2, \cdots x_k$ 的线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \varepsilon \tag{4}$$

其中 $\beta_0, \beta_1, \cdots \beta_k$ 是 $k+1$ 个未知参数，β_0 称为回归常数，$\beta_1, \cdots \beta_k$ 称为回归系数；y 称为被解释变量；$x_1, x_2, \cdots x_k$ 是 k 个可以精确可控制的一般变量，称为解释变量。

当 $p = 1$ 时，上式即为一元线性回归模型，$k \geq 2$ 时，上式就叫做多元形多元回归模型。ε 是随机误差，与一元线性回归一样，通常假设

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases} \tag{5}$$

同样，多元线性总体回归方程为$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

系数β_1表示在其他自变量不变的情况下，自变量x_1变动到一个单位时引起的因变量y的平均单位。其他回归系数的含义相似，从集合意义上来说，多元回归是多维空间上的一个平面。

多元线性样本回归方程为：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{6}$$

多元线性回归方程中回归系数的估计同样可以采用最小二乘法。由残差平方和:

$SSE = \sum (y-\hat{y})=0$

根据微积分中求极小值得原理，可知残差平方和SSE存在极小值。欲使SSE达到最小，SSE对$\beta_0, \beta_1, \cdots \beta_k$的偏导数必须为零。

将SSE对$\beta_0, \beta_1, \cdots \beta_k$求偏导数，并令其等于零，加以整理后可得到$k+1$各方程式：

$$\frac{\partial SSE}{\partial \beta_i} = -2\sum (y-\hat{y}) = 0 \tag{7}$$

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum (y-\hat{y})x_i = 0 \tag{8}$$

通过求解这一方程组便可分别得到$\beta_0, \beta_1, \cdots \beta_k$的估计值$\hat{\beta}_0$，$\hat{\beta}_1$，$\bullet\bullet\bullet$ $\hat{\beta}_k$回归系数的估计值，当自变量个数较多时，计算十分复杂，必须依靠计算机独立完成。现在，利用*SPSS*，只要将数据输入，并指定因变量和相应的自变量，立刻就能得到结果。

对多元线性回归，也需要测定方程的拟合程度、检验回归方程和回归系数的显著性。

测定多元线性回归的拟合度程度，与一元线性回归中的判定系数类似，使用多重判定系数，其中定义为：

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}=1-\frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \tag{9}$$

式中，SSR 为回归平方和，SSE 为残差平方和，SST 为总离差平方和。

同一元线性回归相类似，$0\le R^2\le 1$，R^2 越接近 1，回归平面拟合程度越高，反之，R^2 越接近 0，拟合程度越低。R^2 的平方根成为负相关系数(R)，也成为多重相关系数。它表示因变量 y 与所有自变量全体之间线性相关程度，实际反映的是样本数据与预测数据间的相关程度。判定系数 R^2 的大小受到自变量 x 的个数 k 的影响。在实际回归分析中可以看到，随着自变量 x 个数的增加，回归平方和(SSR)增大，是 R^2 增大。由于增加自变量个数引起的 R^2 增大与你和好坏无关，因此在自变量个数 k 不同的回归方程之间比较拟合程度时，R^2 不是一个合适的指标，必须加以修正或调整。

调整方法为：把残差平方和与总离差平方和纸币的分子分母分别除以各自的自由度，变成均方差之比，以剔除自变量个数对拟合优度的影响。调整的 R^2 为：

$$\overline{R}^2=1-\frac{SSE/(n-k-1)}{SST/(n-1)}=1-\frac{SSE}{SST}\bullet\frac{n-1}{n-k-1}=1-(1-R^2)\frac{n-1}{n-k-1} \tag{10}$$

由上时可以看出，$\overline{R}^2$ 考虑的是平均的残差平方和，而不是残差平方和，因此，一般在线性回归分析中，$\overline{R}^2$ 越大越好。

从 F 统计量看也可以反映出回归方程的拟合程度。将 F 统计量的公式与 R^2 的公式作一结合转换，可得：

$$F=\frac{R^2/k}{(1-R^2)/(n-k-1)} \tag{11}$$

可见，如果回归方程的拟合度高，F 统计量就越显著；F 统计量越显著，回归方程的拟合优度也越高。

由于引体向上测试成绩常存在异常值经过数据的预览，引体向上成绩为 0 的情况在守门员，前锋和后卫中广泛存在，因此我们通过将 30M 跑、箭头跑、立定跳远、纵跳、引体向上、YOYO 成绩作为自变量，将引体向上成绩作为因变量，进行多元线性回归，使用预测值填补缺失值并替换异常值。

模型的求解

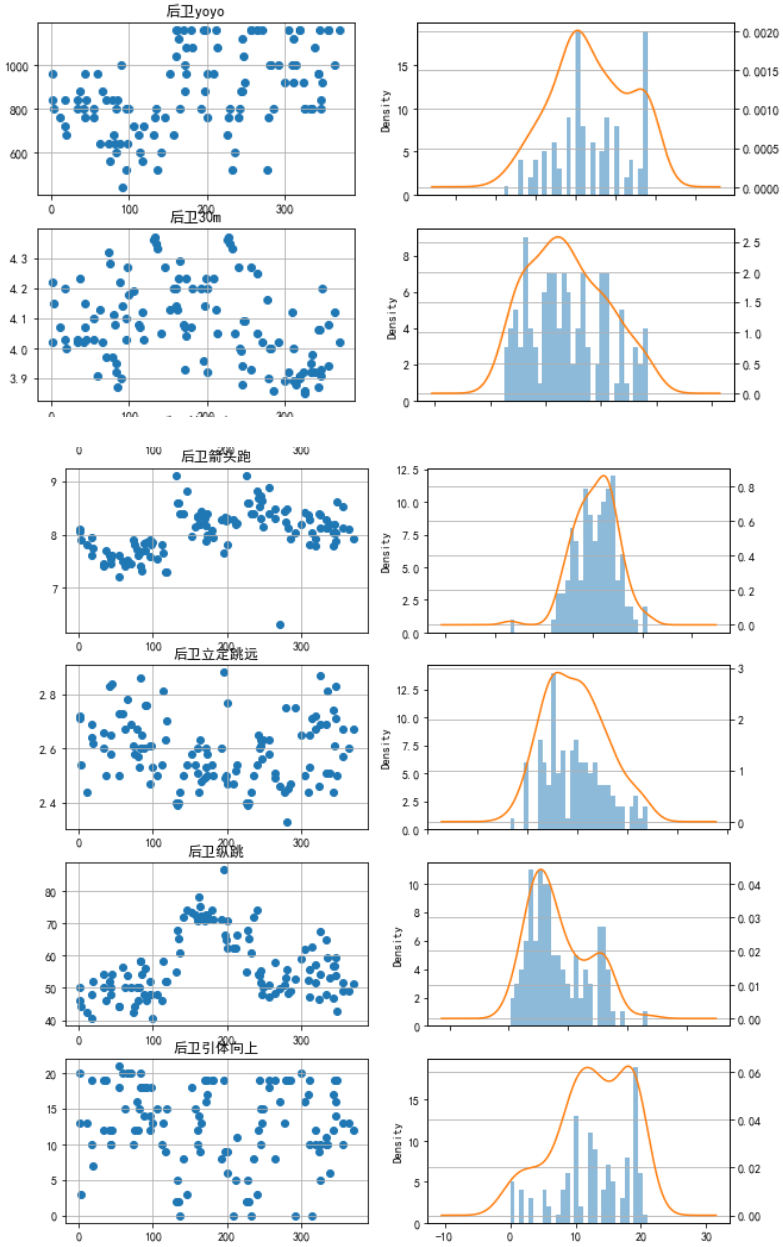
在获得数据之后首先将多个表格的数据进行合并，完成合并之后查看数据的缺失值，可以看到 30m 跑的成绩有四位球员缺失，箭头跑成绩有 5 位球员缺失，立定跳远有四位球员缺失，纵跳成绩有一位运动员缺失，引体向上有四位球员缺失，YOYO 跑成绩有四位球员缺失。

对数据进行预览，计算不同的成绩中 0 的数量，其中引体向上的 0 的数量是最多的，有 14 条。

将数据完成整合之后对于几个类别的数据绘画散点图和条形图，运用 KS 检验计算 pvalue，如果 pvalue 大于 0.05，则接受成绩是符合正态分布的事实，成绩可以认为是正常的。下面检验不同类别的运动员的成绩是否符合正态分布的。

1. 数据预览

（1）后卫数据预览



图片 2

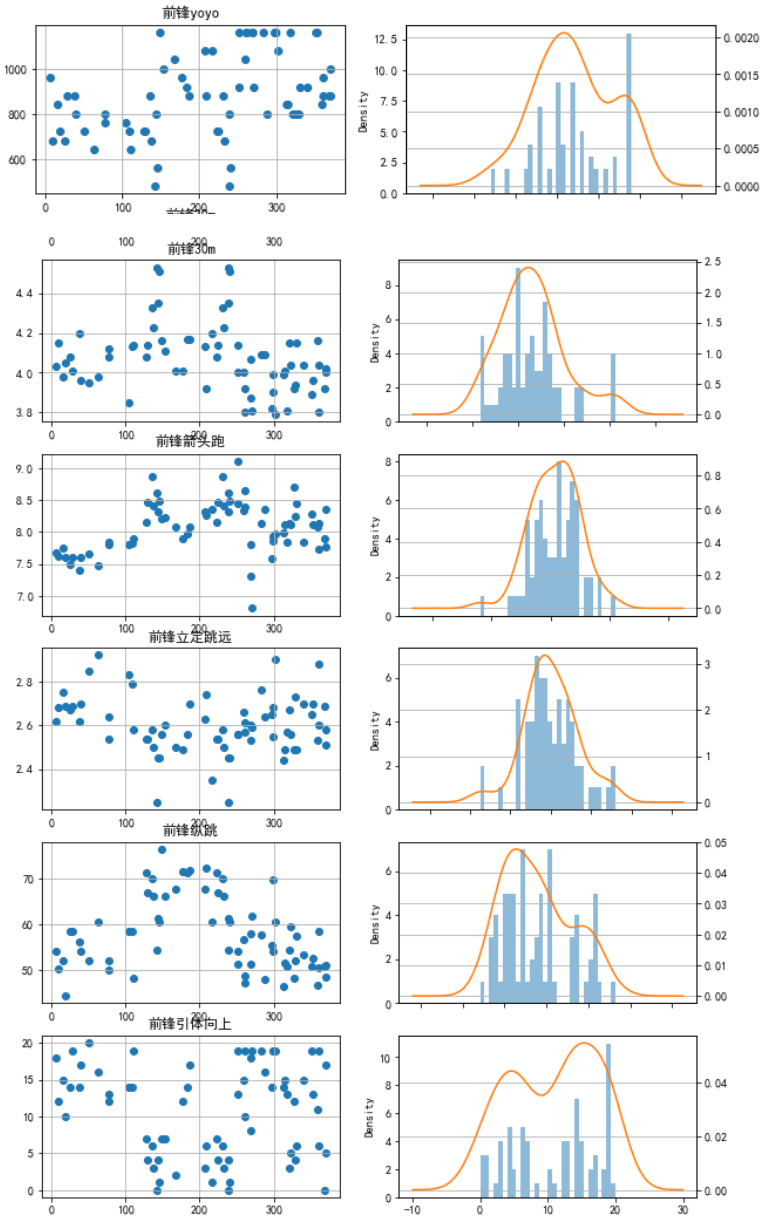
后卫 KS 检验结果：

	statistic	pvalue
YOYO	0.1102	0.1083
30m	0.0739	0.5316
箭头跑	0.0549	0.8730
立定跳远	0.0981	0.1968
纵跳	0.1382	0.0206
引体向上	0.1183	0.0694

表 1

后卫的几项成绩除去纵跳，p 值都大于 0.05，接受原假设，认为数据是符合正态分布的。从纵跳的数据的柱形图观察，也是比较符合正态分布的，因此认为后卫的成绩测量是较为准确的。

（2）前锋数据预览



图片 3

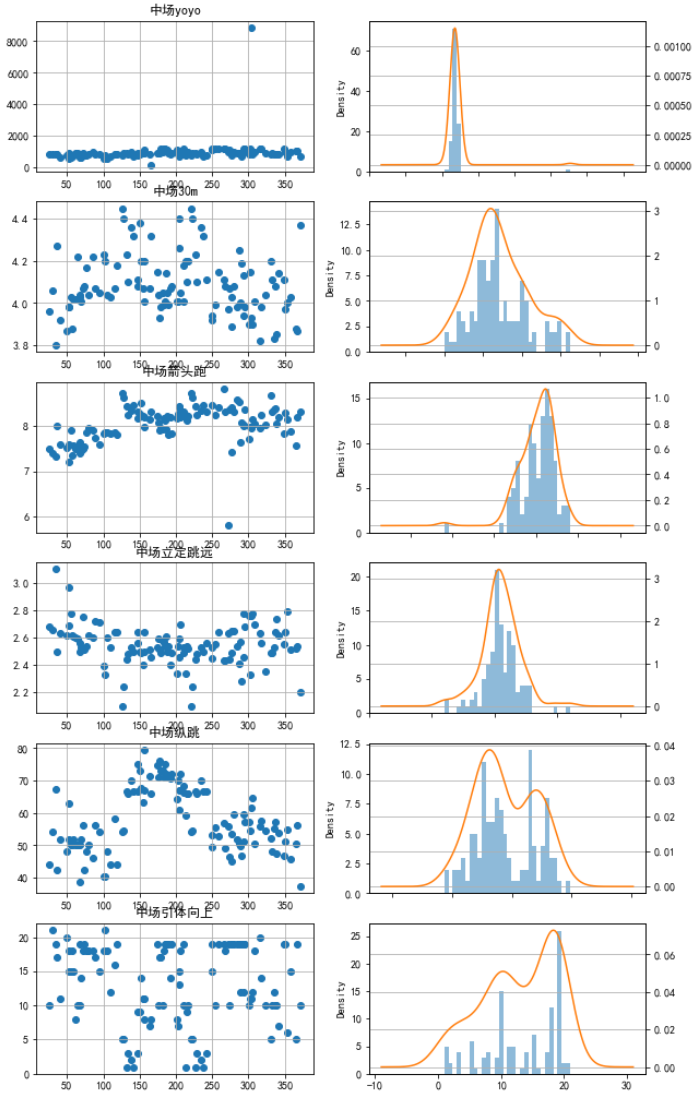
前锋 KS 检验结果：

	statistic	pvalue
YOYO	0.1209	0.2452
30m	0.1152	0.2964
箭头跑	0.0603	0.9632
立定跳远	0.0819	0.7514
纵跳	0.1291	0.1838
引体向上	0.1250	0.2127

表 2

其中前锋的所有成绩的 P 值都大于 0.05，接受原假设，认为前锋的测量的成绩是符合正态分布的，成绩测量是较为准确的。

(3)中场球员数据预览



图片 4

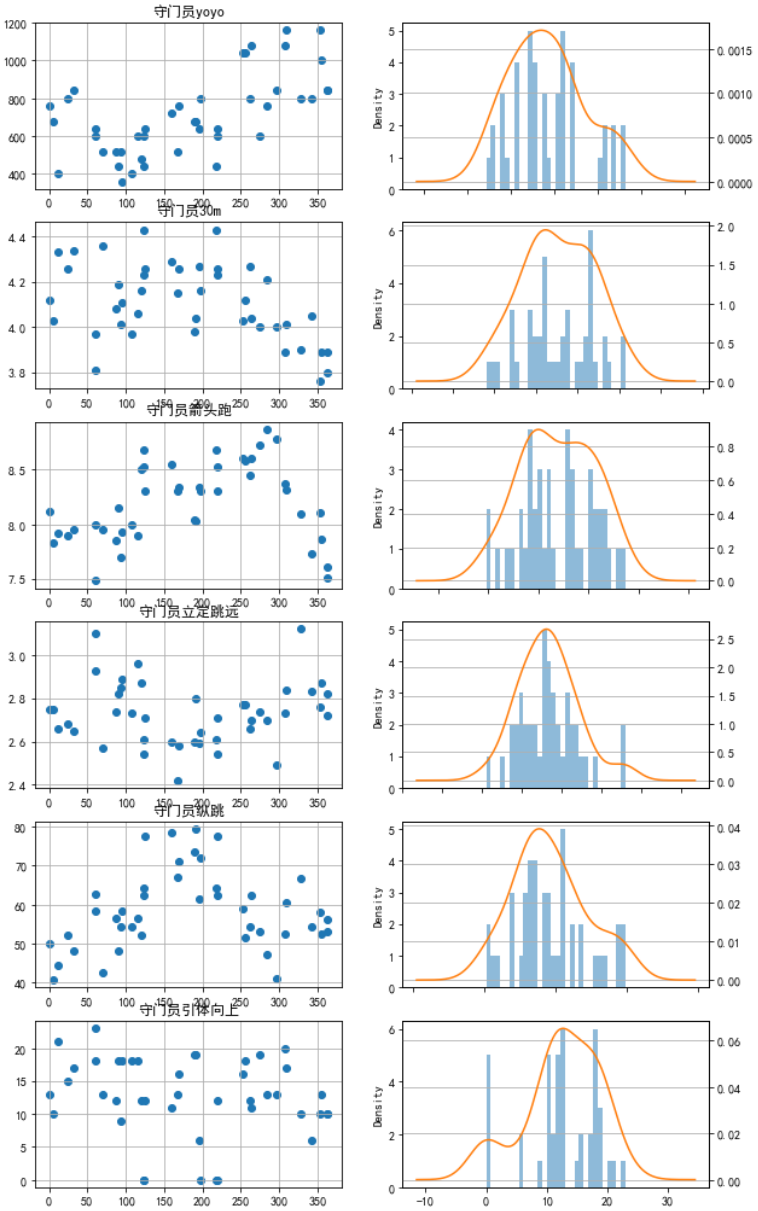
中场 KS 检验结果：

	statistic	pvalue
YOYO	0.3879	1.0947
30m	0.1224	0.0638
箭头跑	——	——
立定跳远	0.1194	0.0756
纵跳	0.1233	0.0608
引体向上	0.1810	0.0011

表 3

清晰的可以看到一位中场球员的 yoyo 跑成绩远大于其他的运动员，超过了 8000，导致中场球员的 yoyo 跑的 pvalue 远小于 0.05，认为这个成绩是异常值，本文在下面会对这个成绩进行处理。

(4)守门员数据预览



图片 5

守门员 KS 检验结果：

	statistic	pvalue
YOYO	0.1093	0.6666
30m	0.0924	0.8561
箭头跑	0.1042	0.7349
立定跳远	0.0916	0.8638
纵跳	0.0949	0.8327
引体向上	0.1486	0.2710

表 4

守门员的各项成绩的 pvalue 值都大于 0.05，认为各项成绩都是符合正态分布的。

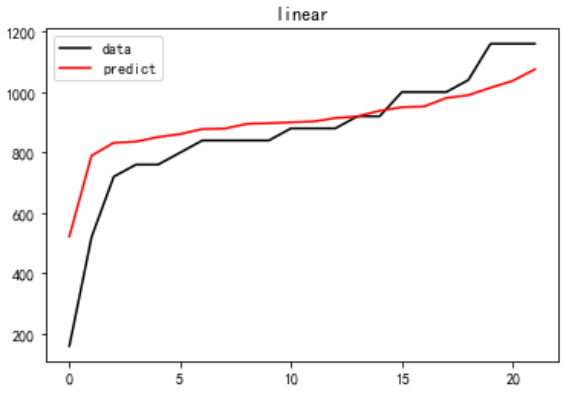
经过上述的图像分析，发现引体向上成绩为 0 的现象在守门员、前锋和后卫中广泛出现，认为这些运动员在测量引体向上时都存在技术动作不标准的现象。下文对于这些奇异值进行处理。

2.奇异值处理

（1）一位中场球员 yoyo 值过高数据处理

选择中场球员中其他成绩不为空的数据，将中场球员的 30m 跑x_1，箭头跑x_2，立定跳远x_3，纵跳x_4和引体向上数量x_5成绩作为自变量，yoyo 跑成绩y_1作为因变量，构建多元线性回归模型,划分为训练集和测试集进行训练，训练的多元线性模型结果如图所示，模型的R^2为 60%，说明模型能够概括变量 60%的特征，展示出的图像说明模型可以较好的拟合出模型的走势。最后的模型的表达式是：

$$y_{1=} - 394 * x_1 + 199 * x_2 - 192 * x_3 + 2.5 * x_4 + 8.11 * x_5 + 1143 \qquad (12)$$



图片 6

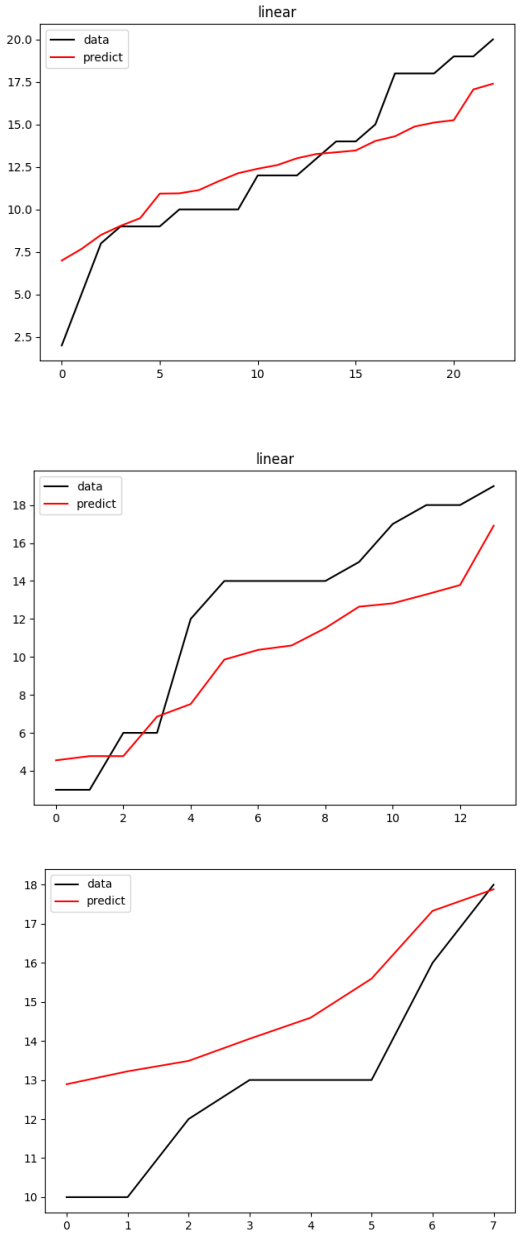
对异常值进行运算，测算出同学的成绩应该是 888，这个结果和原来的结果正好差一个 0（原结果 8840），应该是登记人员的登记误差导致。

（2）多名运动员的引体向上数量为 0

原数据中可以看到有 5 名守门员，6 名后卫和三名前锋的引体向上的数据是为

0 的，从提供的资料中可以看出，如果运动员的运动不标准则系统不会计入引体向上的数量，因此我们对于这些引体向上成绩为 0 的球员分别构建多元线性模型，通过将后卫引体向上数量作为因变量y_1，30m 跑x_1，箭头跑x_2，立定跳远x_3，纵跳x_4和 YOYOx_5成绩作为自变量。前锋的引体向上数量作为因变量y_2，30m 跑x_1，箭头跑x_2，立定跳远x_3，纵跳x_4和 YOYOx_5成绩作为自变量。中场的引体向上数量作为因变量y_3，30m 跑x_1，箭头跑x_2，立定跳远x_3，纵跳x_4和 YOYOx_5成绩作为自变量。构建三个多元线性回归模型，对于值为 0 的运动员进行预测，作为他们本次体测的成绩。下面的图片分别反应后卫、前锋和守门员的多元线性回归模型的预测结果。

$$\begin{cases} y_1 = -7.3 * x_1 - 3.9 * x_2 + 2.1 * x_3 - 0.05 * x_4 + 0.00073 * x_5 + 65 \\ y_2 = -x_1 - 1.1 * x_2 + 17 * x_3 - 16 * x_4 + 0.0003 * x_5 + 23 \\ y_3 = -1.3 * x_1 + 1.9 * x_2 + 8.26 * x_3 - 0.0026 * x_4 - 0.0003 * x_5 - 14.7 \end{cases} \tag{13}$$



图片 7

3.填补缺失值计算得分

对误测的运动员的数据进行填充，最后的缺失的数据和空缺的数据用那一列的平均值进行填充最后获得了没有缺失值，也没有异常值的表格。

给数据进行均匀分箱，最终获得了从本次数据中得到的评价的指标。

	30M跑\n（秒）	箭头跑\n（秒）	立定跳远（米）	纵跳\n（厘米）	引体向上\n（次）	YOYO\n（米）
10%	3.900	7.512	2.440	46.1000	5.0	600.0
20%	3.950	7.684	2.490	48.7800	8.0	680.0
30%	4.000	7.850	2.530	50.8200	10.0	760.0
40%	4.030	7.930	2.560	52.4000	12.0	800.0
50%	4.070	8.070	2.600	54.4000	13.0	840.0
60%	4.100	8.180	2.622	57.3200	15.0	880.0
70%	4.150	8.300	2.660	61.5208	17.4	960.0
80%	4.200	8.380	2.710	66.8020	19.0	1040.0
90%	4.288	8.546	2.768	71.3740	19.0	1160.0
max	4.530	9.100	3.120	86.8680	23.0	1160.0

按照题目的权值计算得分，获得每个球员的最终得分（见 excel 附件），每个队伍的不及格名单（见 excel 附件）与每个队的综合得分，如下所示。

队伍 1	4.87
队伍 2	5.57
队伍 3	5.6
队伍 4	5.03
队伍 5	4.49
队伍 6	3.19
队伍 7	5.3
队伍 8	6.51
队伍 9	5.3
队伍 10	3.19
队伍 11	6.4
队伍 12	5.4
队伍 13	7.16
队伍 14	5.04
队伍 15	6.31

表 5

5. 2 问题 2. 模型的建立与求解

模型的建立

通过引入统计学的四分位数的概念，以此类推求解数据分布中的 85 分位数的概念，有数据的概览已知总共有 373 位参赛的球员，将球员的得分按照从大到小的成绩进行排列之后，计算可以得出，只要排名是在前 318 名的选手都可以进入参赛的名单。计算得出数据是 3.6。

模型的求解

计算最终得分的前 85 分位数，计算出数据是 3.6.也就是说大于 3.6，可以有 85%的球员获得及格的成绩，对于这样的成绩再进行数据的检验，获得每个队伍可以进入参赛球员的名单。

队伍 1	22
队伍 2	25
队伍 3	21
队伍 4	20
队伍 5	23
队伍 6	8
队伍 7	18
队伍 8	27
队伍 9	20
队伍 10	8
队伍 11	31
队伍 12	20
队伍 13	24
队伍 14	21
队伍 15	29

表 6

其中队伍 3，队伍 4，队伍 6，队伍 7，队伍 9，队伍 10，队伍 12，队伍 14 无法凑齐 22 个人。

5.3 问题 3. 模型的建立与求解

5.3.1 模型的建立

（1）首先对搜集到的原始数据进行方差分析。

方差分析(ANOVA)是统计模型及相关估计程序的集合，用于分析均值之间的差异，基于总方差定律，将特定变量中观察到的方差划分至不同变化源的成分。

Step1: 假设每个总体都应符合正态分布，各个总体的方差σ^2相同，观测是独立的。

Step2: 假设

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$H_1: \mu_1, \mu_2, \ldots, \mu_k$不全相等

Step3: 计算各样本均值

$$\overline{x_i} = \frac{\sum_{j=i}^{n_i} x_{ij}}{n_i}, i = 1,2, \ldots, k \tag{14}$$

Step4: 计算全部观测值的总均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k}\sum_{j=i}^{n_i} x_{ij}}{n} \tag{15}$$

Step5: 计算总平方和

$$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=i}^{n_i}(x_{ij} - \bar{\bar{x}})^2 \tag{16}$$

Step6: 计算组间平方和

$$\mathrm{SSA} = \sum_{i=1}^{k}\sum_{j=i}^{n_i} n_i(x_i - \bar{\bar{x}})^2 \tag{17}$$

Step7: 计算组内平方和

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSA} \tag{18}$$

Step8: 计算组间方差

$$\mathrm{MSA} = \frac{SSA}{k-1} \tag{19}$$

Step9: 计算组内方差

$$\mathrm{MSE} = \frac{SSE}{n-k} \tag{20}$$

将组间方差与组内方差进行对比，就得到了所需的检验统计量 F，当H_0为真时，有

$$\mathrm{F} = \frac{MSA}{MSE} \sim F(k-1, n-k) \tag{21}$$

若$F > F_\alpha$，则拒绝原假设，因素水平对观测值有显著影响；

若$F < F_\alpha$，则接受原假设，不能认为因素水平对观测值有显著影响。

（2）对方差分析得到的显著变量进行 Logistic 回归分类统计

Logistic 回归是二分类因变量常用的统计分析方法。在统计学中，Logistic 回归模型用于对特定类别或事件的概率进行建模，例如某件事的成功或失败等，其基本形式是使用 Logistic 函数对二进制因变量进行建模。在对于问题 3. 的求解过程中，首先使用方差分析将众多可能会对评价结果造成影响的自变量降维，之后借助 Logistic 回归，分析出某个位置的球员是否优秀（二分类问题）以及其可能为优秀球员的概率。

其原理和 Linear Regression(线性回归)的原理是相似的，可以简单的描述为这样的过程：

Step1: 找一个合适的预测函数，一般表示为h函数，该函数就是我们需要找的分类函数，它用来预测输入数据的判断结果。这个过程时非常关键的，需要对数据有一定的了解或分析，知道或者猜测预测函数的“大概”形式，比如是线性函数还是非线性函数。

Step2: 构造一个Cost函数（损失函数），该函数表示预测的输出（h）与训练数据类别（y）之间的偏差，可以是二者之间的差（$h-y$）或者是其他的形式。综合考虑所有训练数据的“损失”，将Cost求和或者求平均，记为$J(\theta)$函数，表示所有训练数据预测值与实际类别的偏差。

Step3: 显然，$J(\theta)$函数的值越小表示预测函数越准确（即h函数越准确），所以这一步需要做的是找到$J(\theta)$函数的最小值。找函数的最小值有不同的方法，Logistic 回归实现时用的是梯度下降法。

具体求解步骤如下：

Step1: 设降维后剩余的自变量为$X\ldots X_n$，因变量为Y。将 y 看成某一位置球员是否优秀的概率，y→ 1则说明该球员有极大的概率为优秀球员，y→ 0则说明该球员有极大的概率为不优秀球员。

$$x = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \tag{22}$$

Step2: 取预测值:

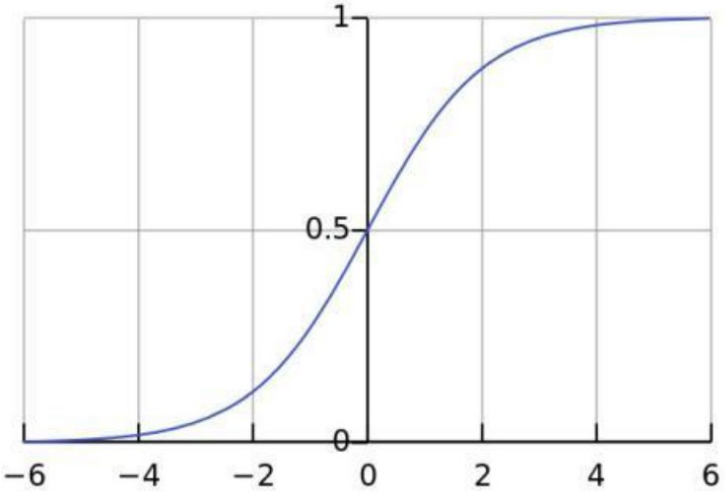
$$\hat{x} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_k} x_{ki} \tag{23}$$

Step3: 取连接函数

Logistic 回归虽然名字里带“回归”，但是它实际上是一种分类方法。根据步骤，需要先找到一个预测函数（h），显然，该函数的输出必须是两个值（分别代表两个类别），所以利用了 Logistic 函数（或称为 Sigmoid 函数）

$$F(x,\beta) = S(x_i'\beta) = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)} \tag{24}$$

函数图像：一个取值在 0 和 1 之间的 S 型曲线



图片 8

Step4: 使用极大似然估计得方法进行求解

Step5: 得到优秀球员发生的概率：

$$\widehat{y_i} = P(y_i = 1|x) = S(x_i'\hat{\beta}) = \frac{exp(x_i'\hat{\beta})}{1 + exp(x_i'\hat{\beta})} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}x_{1i}+\widehat{\beta_2}x_{2i}+\cdots+\widehat{\beta_k}x_{ki}}}{1 + e^{\widehat{\beta_0}+\widehat{\beta_1}x_{1i}+\widehat{\beta_2}x_{2i}+\cdots+\widehat{\beta_k}x_{ki}}} \tag{25}$$

$\widehat{y_i}$即为球员是否可能为某一位置的优秀球员的概率，取损失函数：

$$C = J(\theta) = -\frac{1}{k} logL(p) \tag{26}$$

这里的 Cost 函数和$J(\theta)$函数也是基于最大似然估计推导得到的。

因为乘了一个负的系数$-\frac{1}{k}$，所以$J(\theta)$取最小值时的θ为要求的最佳参数。C 的值越小，Logistic 回归拟合的越好。

5.3.2 模型的求解：

对不同位置的球员设置不同的权重

（1）前锋位置

对前锋位置的球员建立 Logistic 回归模型，将数据划分成好前锋与坏前锋，进行训练与预测，其中运用网格搜索调整超参数选择最适合的参数 C，选择预测效果最好的模型作为区分前锋好坏的模型。

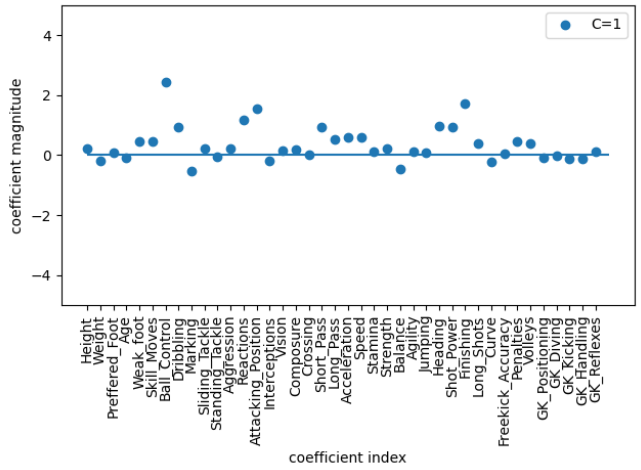
从模型的各个自变量对于 Logistic 模型的解释程度来看，前锋位置的控球和盘带对于前锋要求很高，反应力和进攻的位置对前锋有较高的要求，前锋的射门力量和头球能力也有很高的要求。上述的因素分别反应为前锋的反应力要求和弹跳力要求。

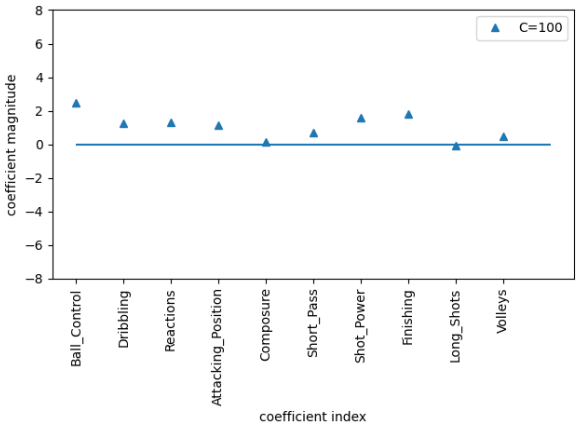
由于原来的因素过多尚不可直观观察，运用方差分析，选择最能够解释前锋好坏的百分之二十五的变量进行训练，绘制散点图，观察出选择的变量基本涵盖的是控球能力，反应能力进球能力和弹跳能力。相比之下，耐力与上肢力量对于区分前锋的好坏并没有很大的帮助，因此调低前锋的 YOYO 得分权重和引体向上得分的权重。

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	25%	20%	5%	20%	20%	10%

表 7

图片 9





图片 10

（2）中场位置

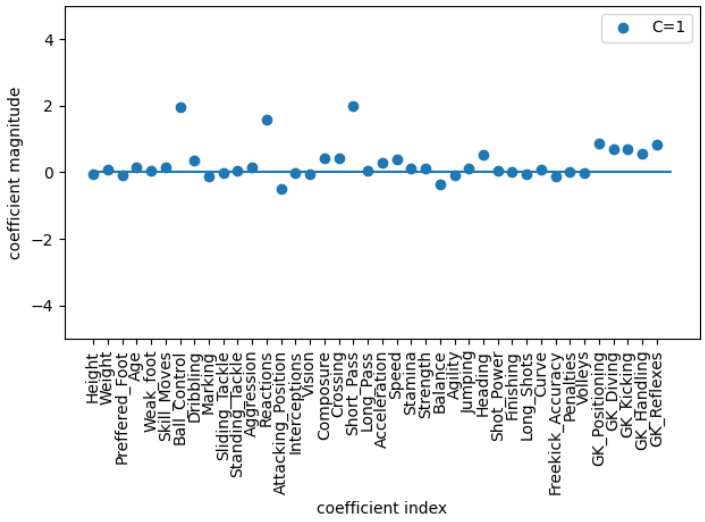
对中场位置的球员建立 Logistic 回归模型，将数据划分成好中场与坏中场，进行训练与预测，其中运用网格搜索调整超参数选择最适合的参数 C，选择预测效果最好的模型作为区分中场好坏的模型。

从模型的各个自变量对于 Logistic 模型的解释程度来看，中场位置的控球和盘带对于中场要求很高，反应力对中场有较高的要求。

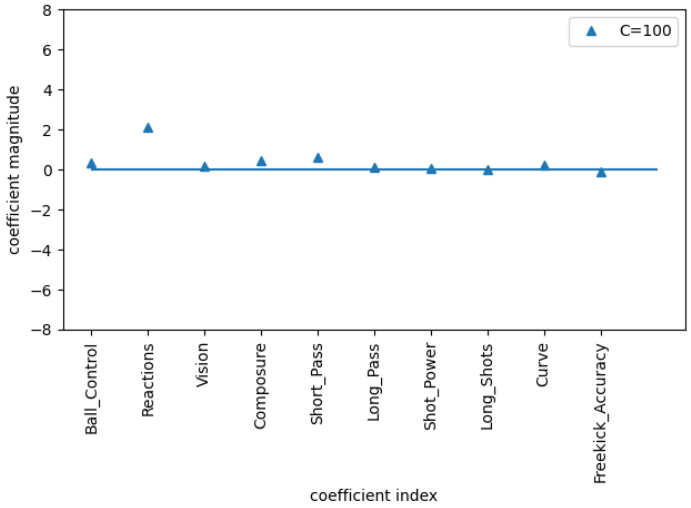
由于原来的因素过多尚不可直观观察，运用方差分析，选择最能够解释前锋好坏的百分之二十五的变量进行训练，绘制散点图，观察出选择的变量主要包括反应和传球的能力。相比之下，耐力与上肢力量对于区分前锋的好坏并没有很大的帮助，因此调低前锋的 YOYO 得分权重和引体向上得分的权重。

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	30%	20%	5%	15%	20%	10%

表 8



图片 11



图片 12

（3）后卫位置

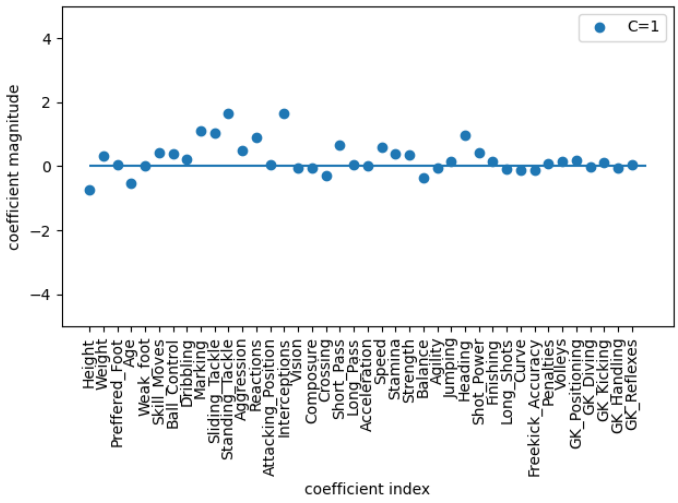
对后卫位置的球员建立 Logistic 回归模型，将数据划分成好后卫与坏后卫，进行训练与预测，其中运用网格搜索调整超参数选择最适合的参数 C，选择预测效果最好的模型作为区分中场好坏的模型。

从模型的各个自变量对于 Logistic 模型的解释程度来看，后卫位置的拦截能力和抢断能力对于后场要求很高，头球能力对后卫有较高的要求，速度和耐力也对后卫有一定的要求。

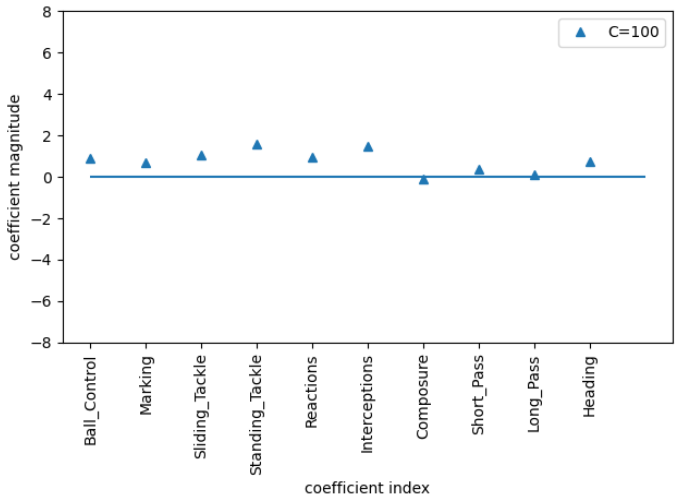
由于原来的因素过多尚不可直观观察，运用方差分析，选择最能够解释前锋好坏的百分之二十五的变量进行训练，绘制散点图，观察出选择的变量主要包括拦截，抢断和头球。相比之下，耐力与上肢力量对于区分前锋的好坏并没有很大的帮助，因此适度调低前锋的 YOYO 得分权重，降低引体向上得分的权重，适度提高反应能力，弹跳能力的权重。

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	25%	20%	5%	20%	15%	15%

表 9



图片 13



图片 14

（4）守门员位置

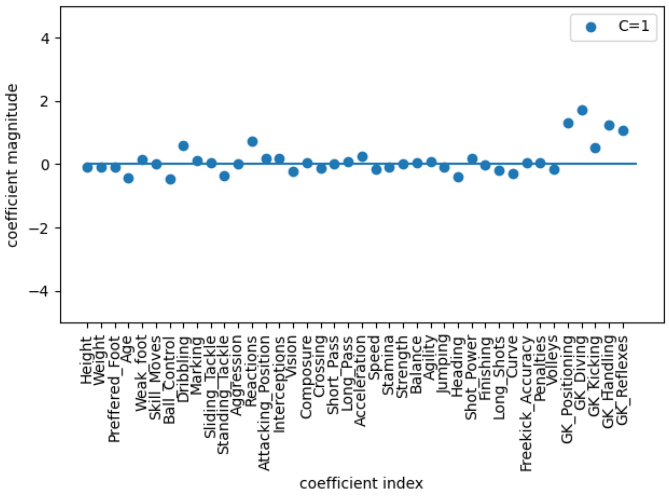
对守门员位置的球员建立 Logistic 回归模型，将数据划分成好守门员与坏守门员，进行训练与预测，其中运用网格搜索调整超参数选择最适合的参数 C，选择预测效果最好的模型作为区分中场好坏的模型。

从模型的各个自变量对于 Logistic 模型的解释程度来看，守门员的反应力和守门员的持球能力比较重要。

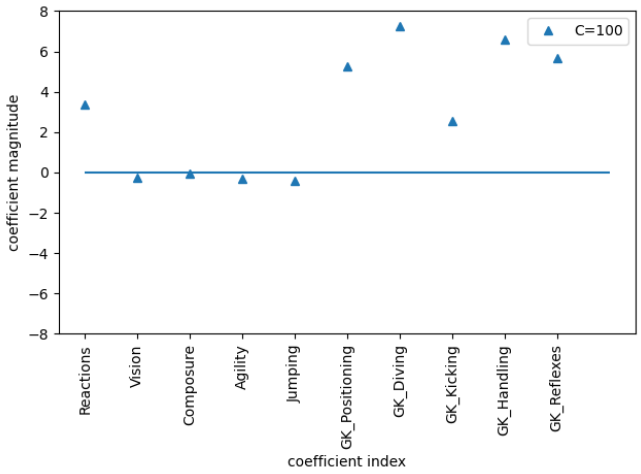
由于原来的因素过多尚不可直观观察，运用方差分析，选择最能够解释守门员好坏的百分之二十五的变量进行训练，绘制散点图，观察出选择的变量主要包反应能力和守门员的持球能力。相比之下，耐力与上肢力量对于区分前锋的好坏并没有很大的帮助，因此适度调低前锋的 YOYO 得分权重，降低引体向上得分的权重，适度提高反应能力，弹跳能力的权重。

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	30%	20%	5%	25%	15%	5%

表 10



图片 13



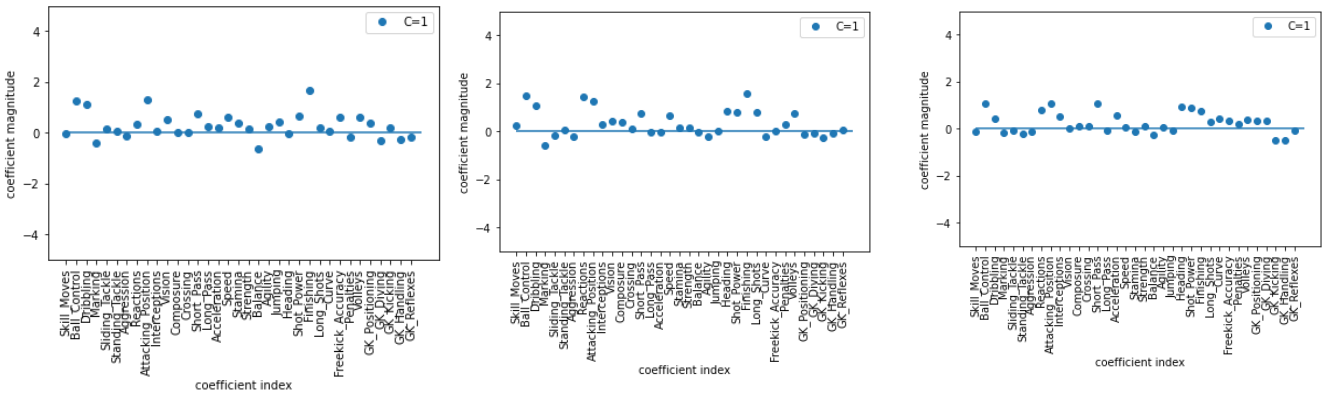
图片 14

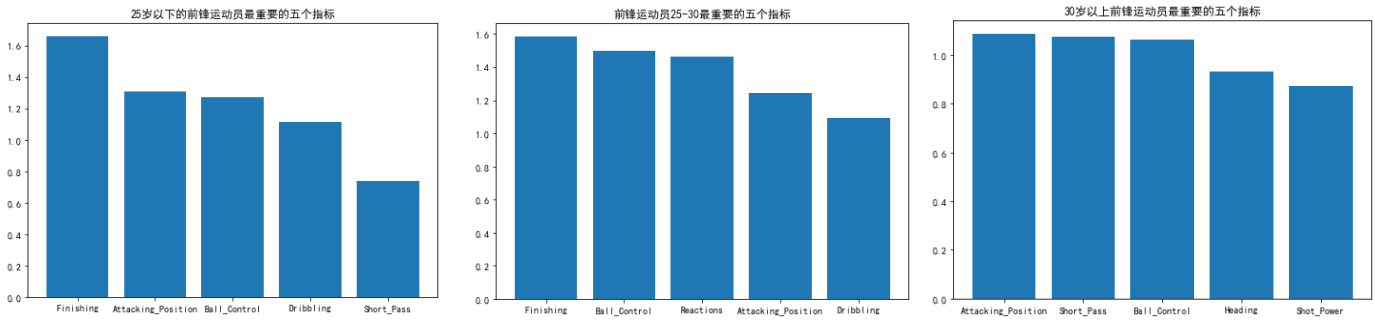
5. 4 问题 4. 模型的建立与求解

Logistic 模型的建立同问题 3.

求解：

从前锋的成长的角度，将前锋的数据划分成三个不同的年龄段，分别是小于 25 岁，25 岁到 30 岁和三十岁以上，这三个年龄段分别代表的是运动员职业发展的初始期，技术成熟期与职业生涯晚期。分别对这三个不同年龄段的好前锋与差前锋进行 Logistic 回归，构建模型，总结出在不同的职业发展时期，前锋需要进行哪个方面的发展。选择前锋三个年龄段的最能帮助区分好坏前锋的五种指标，其意义代表为前锋在那个年龄段的发展的方向，也就是那个年龄段的前锋的评价的侧重点。





组图 1

三个不同的阶段前锋有共同的重要的参数比如说前锋的控球能力，这表明作为前锋，盘带始终是一项非常基础的技能。有在前锋年轻时重要的指标，但是因为年龄的增长而变得不再有区分度的指标，比如说终结能力。有前锋年轻时不是具有区分度的指标，但是随着前锋年龄的增大变得更加重要的指标，比如说弹跳能力。

具体反映在下面的特征：

1、控球能力始终是前锋的发展重点

2、终结能力、盘带能力在前锋球员的发展初期（25 岁以下）十分重要

3、反应力在 25 到 30 岁对前锋球员较为重要

3、前锋球员的对于进攻位置的感知应该随年龄增加而提高

4、头球能力和射门力量以及组织能力是前锋后期的发展方向

下面是对不同年龄的前锋的评判指标的修正。

年龄 20-25

项目	控球能力	盘带能力	短传能力	纵跳	反应力	定点射门
权重	20%	20%	15%	15%	15%	15%

表 11

年龄 25-30

项目	控球能力	盘带能力	短传能力	纵跳	反应力	定点射门
权重	15%	20%	15%	20%	20%	10%

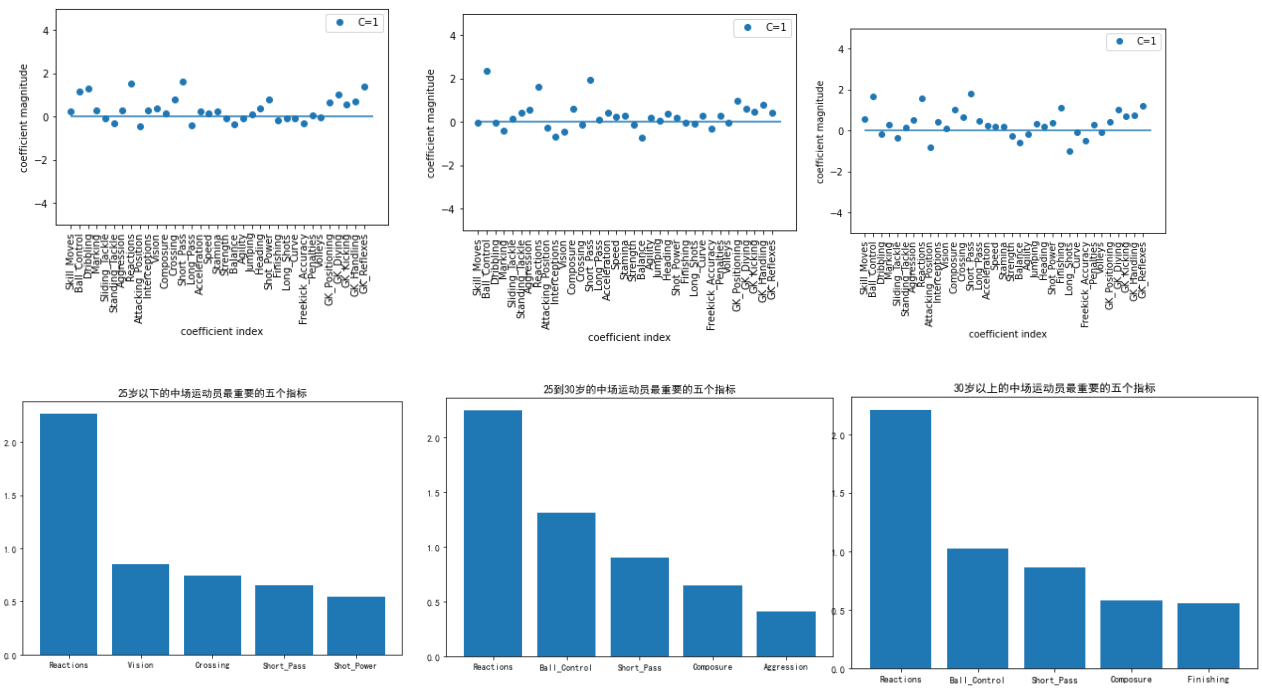
表 12

年龄 30 以上

项目	控球能力	盘带能力	短传能力	纵跳	反应力	定点射门
权重	15%	10%	20%	25%	15%	15%

表 13

从中场球员的成长的角度，将中场的数据划分成三个不同的年龄段，分别是小于 25 岁，25 岁到 30 岁和三十岁以上，这三个年龄段分别代表的是运动员职业发展的初始期，技术成熟期与职业生涯晚期。分别对这三个不同年龄段的好前锋与差前锋进行 Logistic 回归，构建模型，总结出在不同的职业发展时期，中场球员需要进行哪个方面的发展。



组图 2

三个不同的阶段中场有共同的重要的参数比如说中场球员的控球能力和短传能力，这表明作为中场，控球和传球始终是中场球员一项非常基础的技能。长传随着中场球员年龄的增加，对于区分中场球员的好坏更加明显，

具体反映在下面的特征：

1、不同年龄的优秀的中场球员都需要很快的反应力和短传能力

2、30 岁以上的中场球员需要增强射门能力

3、25 岁以下的中场球员需要建立起长传能力，这反映中场球员的大局观

下面是对不同年龄的前锋的评判指标的修正。

年龄 20-25

项目	长传	短传	反应力	控球	传中	定点射门
权重	20%	15%	20%	15%	15%	15%

表 14

年龄 25-30

项目	长传	短传	反应力	控球	传中	定点射门
权重	15%	15%	25%	15%	15%	15%

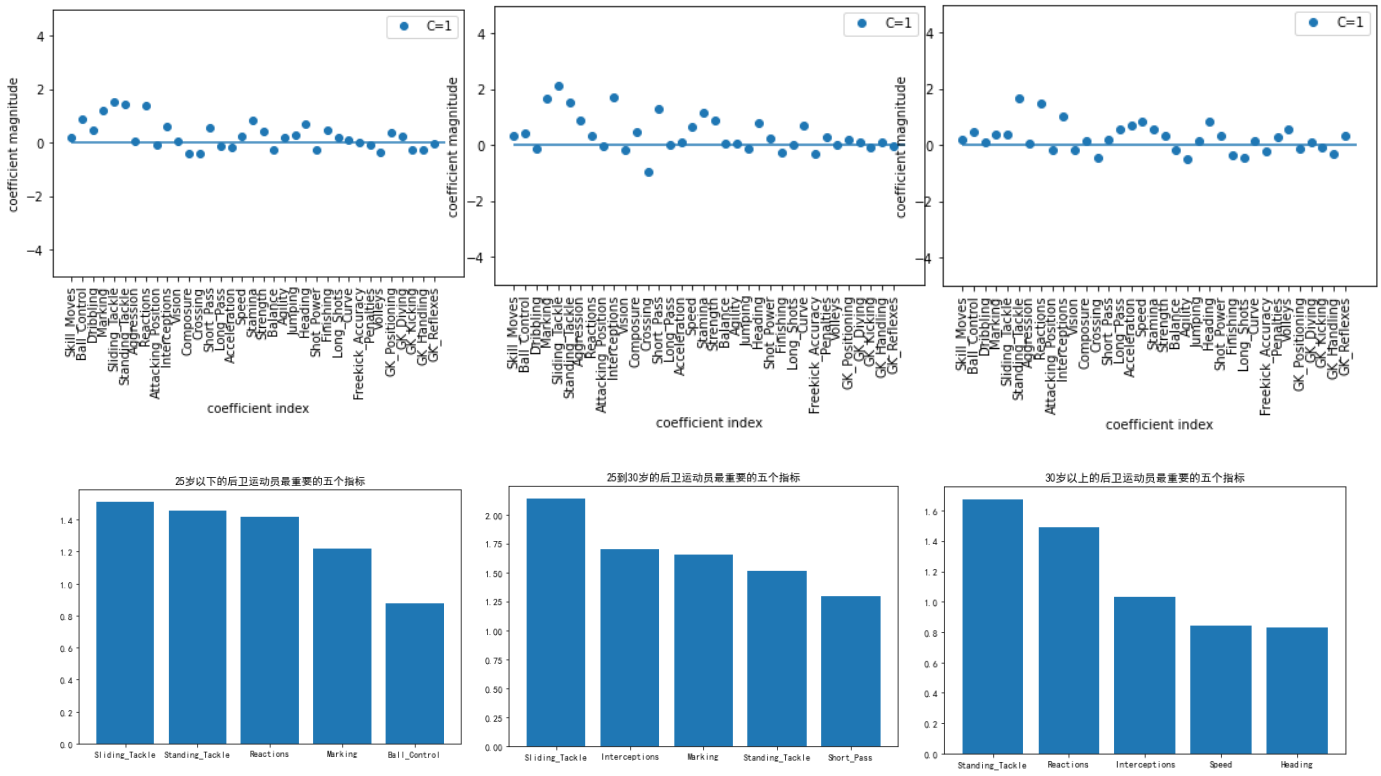
表 15

年龄 30 以上

项目	长传	短传	反应力	控球	传中	定点射门
权重	10%	15%	20%	15%	15%	25%

表 16

后卫球员的成长的角度，将后卫的数据划分成三个不同的年龄段，分别是小于 25 岁，25 岁到 30 岁和三十岁以上，这三个年龄段分别代表的是运动员职业发展的初始期，技术成熟期与职业生涯晚期。分别对这三个不同年龄段的好前锋与差前锋进行 Logistic 回归，构建模型，总结出在不同的职业发展时期，后卫球员需要进行哪个方面的发展。



组图 3

三个不同的阶段后场有共同的重要的参数比如说后场球员的站位能力，这主要反映在后场的球员时刻注意进攻球员的进攻的意向并及时进行调整。随着时间阅历的增长，后场球员的滑铲、铲球等基本功对于区分能力不再重要，也就没有意义增加训练难度。

具体反映在下面的特征：

1、不同年龄的优秀的后场球员都需要有很强的反应能力

2、后场球员的短传能力在 25 到 30 岁更加重要

3、后场球员随着年龄的增长，速度和头球能力对于一个后卫球员变得更加重要

下面是对不同年龄的前锋的评判指标的修正。

年龄 20-25

项目	滑铲	盯人	速度	纵跳	短传	反应力
权重	15%	15%	15%	15%	20%	20%

表 17

年龄 25-30

项目	滑铲	盯人	速度	纵跳	短传	反应力
权重	15%	15%	15%	10%	25%	20%

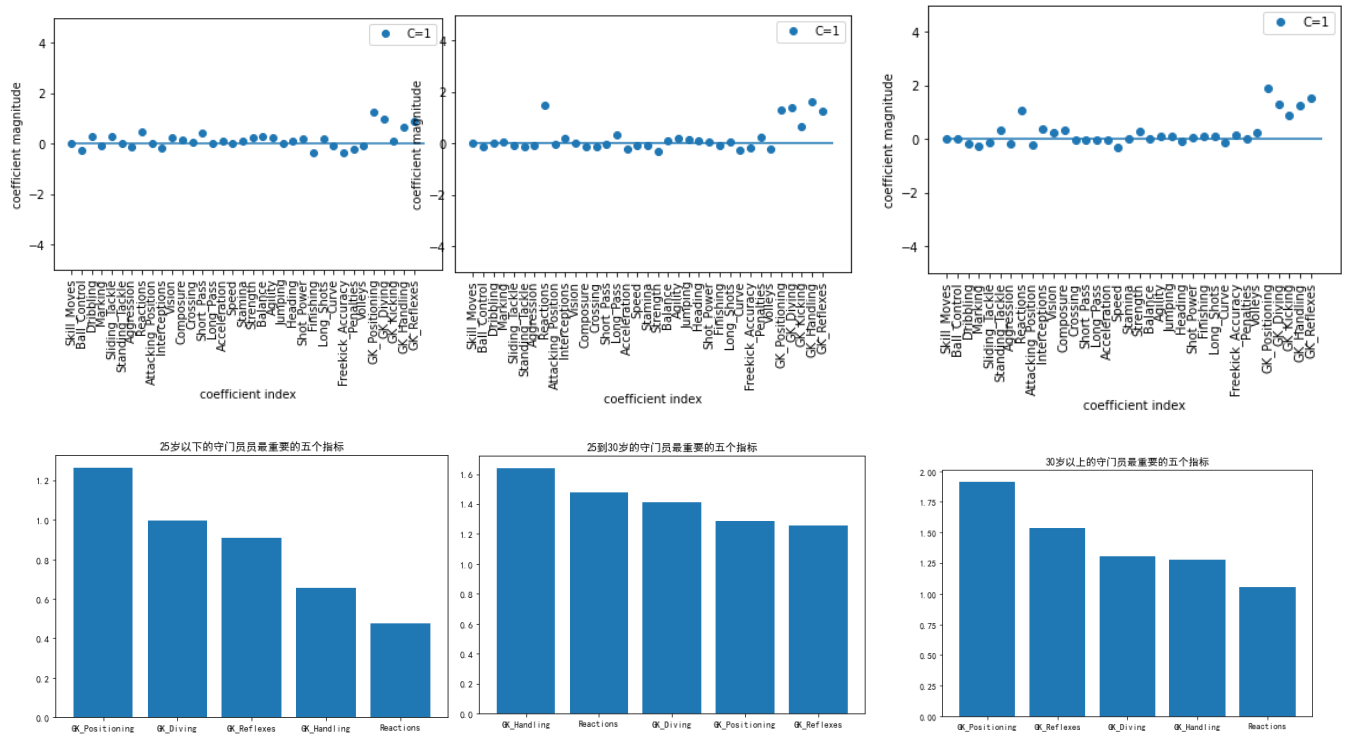
表 18

年龄 30 以上

项目	滑铲	盯人	速度	纵跳	短传	反应力
权重	15%	15%	20%	20%	15%	15%

表 19

从守门球员的成长的角度，将后卫的数据划分成三个不同的年龄段，分别是小于 25 岁，25 岁到 30 岁和三十岁以上，这三个年龄段分别代表的是运动员职业发展的初始期，技术成熟期与职业生涯晚期。分别对这三个不同年龄段的好前锋与差前锋进行 Logistic 回归，构建模型，总结出在不同的职业发展时期，守门球员需要进行哪个方面的发展。



组图 4

三个不同的阶段的守门员有共同的重要的参数比如说守门员球员的站位能力，这主要反映在后场的球员时刻注意进攻球员的进攻的意向并及时进行调整。年轻时对于守门员的守门员上肢力量并没有很重要，在 25 岁到三十岁会更加有重要性。

具体反映在下面的特征：

1、不同年龄的优秀的守门员球员都需要有很强的反应能力

2、守门员球员的上肢力量随时间增长而更加需求

3、守门员的开大脚能力随时间增长更加需求。

年龄 20-25

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	15%	15%	10%	10%	10%	40%

表 20

年龄 25-30

项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	15%	15%	10%	10%	15%	35%

表 21

年龄 30—

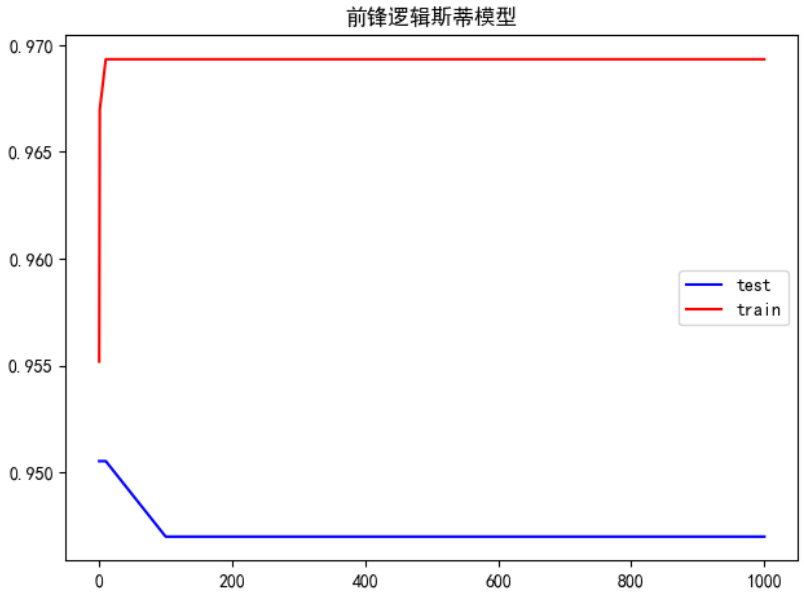
项目	箭头跑	30 米跑	引体向上	纵跳	立定跳远	YOYO 得分
权重	15%	10%	15%	10%	20%	30%

表 22

六、 模型的分析与检验

本文中分别对于不同年龄段的前锋球员，中场球员后卫球员和守门员进行建立 logistic 回归模型，通过调整正则化强度在多个模型中选择最适合预测的模型。正则化强度 C 从 0.1，1，10，100，1000 中选择，进行模型的训练和预测。下面以前锋球员的 Logistic 回归作为例子，介绍模型参数的选择过程。

下面的图展示出 C 的增大，并没有让训练集的准确度进一步的提高，但是测试集的预测精度却下降了，可以看到模型过拟合了，因此我们选择相对较小的 C 值，也就是模型训练集刚刚到达最高值的值，C 取为 1，这样模型的训练集和测试集的训练精度都能达到最高，模型的预测效果较好，便于分析模型的参数的意义。



图片 15

七、 模型的评价、改进与推广

7.1 模型的优点

（1）简单易行。建模迅速，对于此次建模任务中运动员体测数据量小、关系简单的情况很有效。

（2）线性回归模型十分容易理解，结果具有很好的可解释性，有利于决策分析。

7.2 模型的缺点

（1）难以很好地拟合高度复杂的数据。

（2）对于非线性数据或者数据特征间具有相关性多项式回归难以建模。本次题目提供的数据相关性较强，四次使用多元线性回归模型的R^2分别为 42%、36%、55%、47%，均未大于 60%，即模型所能概括的信息较少。

7.3 模型的改进

采用岭回归对多元线性回归进行改进。

岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

岭回归主要解决的问题是两种：一是当预测变量的数量超过观测变量的数量的时候（预测变量相当于特征，观测变量相当于标签），二是数据集之间具有多重共线性，即预测变量之间具有相关性。

一般的，回归分析的（矩阵）形式如下：

$$y = \sum_{j=1}^{p} \beta_j x_j + \beta_0 \tag{27}$$

其中，x 是预测变量，y 是观测变量，β和 $\beta 0$ 是待求的参数。而 $\beta 0$ 可以理解成偏差（Bias）。一般情况下，使用最小二乘法求解上述回归问题的目标是最小化如下的式子：

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i)^2 \tag{28}$$

这里的 1,...,N 是训练集中的样本。那么，岭回归就是要在上述最小化目标中加上一个惩罚项:

$$\lambda \sum_{j=1}^{p} \beta_j^2 \tag{29}$$

$$\hat{\beta^{bridge}} = \arg\min_{\beta} \{\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_i x_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\} \tag{30}$$

这里的 λ 也是待求参数。也就是说，岭回归是带二范数惩罚的最小二乘回归。岭回归的这种估计目标叫做收缩估计器。

传统的回归分析我们需要使用 t 检验来确定预测变量是否显著，如果不显著则剔除该预测变量，然后继续回归，如此往复得到最终结果。而岭回归不需要这样，只要它的系数 β 能向 0“收缩”即可减小该变量对最终的影响。

八、 参考文献

[1]胡峰,吴波,胡友民,史铁林.基于概率神经网络和 KS 检验的机械状态监测[J].振动与冲击,2008(04):56-57+62+168-169.

[2]孙明娟.基于 Logistic 回归的胃癌预测研究[J].科技经济导刊,2019,27(28):126-127.

[3]孙振宇.多元回归分析与 Logistic 回归分析的应用研究[D].南京信息工程大学,2008.

[4]王惠文,孟洁.多元线性回归的预测建模方法[J].北京航空航天大学学报,2007(04):500-504.

九、 附录

附件 1. 第一问所有球员的成绩以及每个队伍的不及格名单
附件 2. 第二问修改成绩标准后每个队伍能够进入参赛名单的球员
附件 3. 用于三四问构建 logistic 模型的 excel 文件
附件 4. 各个位置，各个位置不同年龄段球员的的 logistic 模型
附件 5. 一三四问.py 文件