



兰州大学

足球运动员综合能力 分析报告

学院：信息科学与工程学院

班级：2018 级数据科学 4 班

姓名：陈宇铭

学号：320180939611

日期：2020-06-31

目录

1. 背景分析	1
2. 研究目的	1
3. 数据说明	1
3.1 数据背景	1
3.2 数据来源	2
3.3 数据收集	2
3.4 数据概况	3
4. 数据处理	3
4.1 数据清洗	3
4.2 数据转化	4
4.3 设置变量	4
5. 数据分析	5
5.1 相关性分析	5
5.1.1 相关系数热力图.....	5
5.1.2 部分系数相关性.....	5
5.2 模型假设	6
5.3 正态性检验	6
5.3.1 直方图.....	6
5.4 模型建立	7
5.4.1 系数表.....	7
5.4.2 模型摘要.....	8
5.4.3 ANOVA.....	8
5.4.4 评估指标解释.....	8
5.4.5 模型公式.....	9
5.5 模型检验	10
5.5.1 显著性检验.....	10
5.5.2 多重共线性检验.....	11
5.5.3 残差分析.....	12
5.5.4 异常观测值.....	15
6. 结论	16

1. 背景分析

足球，被誉为“世界第一运动”，是全球体育界最具影响力的体育运动，对世界各国的政治、经济和文化等领域具有极其深远的影响。近年来，足球运动已完全进入一个职业化、商业化的时代，成为了一种涉及社会经济各个领域的大型体育文化产业。目前的足球竞赛表面上看是运动员技术、战术、体能和心理素质的竞争，而实质上更是经济实力的竞争。其中，足球的转会市场便是各职业俱乐部在经济竞争中的主战场。在转会市场中，俱乐部可以通过与其他俱乐部进行球员交易来提高球队的竞技水平或改善财务状况。在这个充斥着诸多博弈的商业竞争中，如何更全面地衡量球员价值便成为各俱乐部能否在竞争摘得桂冠的重要因素。为了解决这一问题，项目将深入球员综合能力这一重要衡量因素，进行研究和分析。

2. 研究目的

探究球员各项数据与球员综合能力的关系，探究球员综合能力的影响因素，并建立线性回归模型，通过各项数据为预测球员综合能力，帮助足球职业俱乐部在转会期间拥有更好的球员价值衡量指标，以此为其在球员交易的过程中争取更多的主动权。

3. 数据说明

3.1 数据背景

研究所使用的数据为中的球员数据。是一款于 2019 年发布，由制作发行的由国际足联官方授权的高拟真 3D 足球类体育仿真游戏。系列的首部作品可以追溯到 1993 年，在 27 年间，致力于追求游戏的极致真实性，通过各种游戏制作技术，努力用游戏还原出足球运动的真实。其中，为了使游戏中的虚拟球员更加贴近现实中的真实球员，FIFA 系列游戏设置了包含 300 多条原始数据的庞大球员数据库。同时，为了尽可能保证数据的真实性，招募了一支多达 9 千人的数据调查员来负责球员能力属性的初步评估。在收集球员基本信息之后，300 多位数据编辑员会对这些数据进行再处理，以此保证数据的真实性和可靠性。

3.2 数据来源

数据来源为 <https://sofifa.com/>。该网站是著名的 FIFA 系列游戏职业模式数据库的提供平台。其收录了 FIFA 系列游戏的全部数据，包括球员和球队的详细信息，具有极强的可靠性和真实性，是世界上最受欢迎的足球数据提供平台之一。






NAME	AGE	JOVA	POT	TEAM & CONTRACT	VALUE	WAGE	TOTAL...	HITS
 L. Messi RW ST CF	32	94	94	FC Barcelona 2004 ~ 2021	€95.5M	€560K	2255	726
 Cristiano Ronaldo ST LW	34	93	93	Juventus 2018 ~ 2022	€58.5M	€410K	2227	528
 Neymar Jr LW CAM	27	92	92	Paris Saint-Germain 2017 ~ 2022	€105.5M	€290K	2179	402
 V. van Dijk CB	27	91	92	Liverpool 2018 ~ 2023	€90M	€240K	2111	426
 J. Oblak GK	26	91	93	Atlético Madrid 2014 ~ 2023	€77.5M	€125K	1412	156

图 1 网站部分截图



REAL OVERALL RATING			Lionel Andrés Messi Cuccittini (ID: 158023) FIFA 20 JUN 24, 2020	
			RW ST CF 32y.o. (Jun 24, 1987) 5'7" 159lbs	
Overall Rating	Potential	Value	Wage	
94	94	€95.5M	€560K	
PROFILE		PLAYER SPECIALITIES		
Preferred Foot: Left 4 Weak Foot 4 Skill Moves 5 International Reputation Work Rate: Medium/Low Body Type: Messi Real Face: Yes Release Clause: €195.8M		#Dribbler #Distance Shooter #Crossover #FK Specialist #Acrobat #Clinical Finisher #Complete Forward		
		FC BARCELONA Position: RW Jersey Number: 10 Joined: Jul 1, 2004 Contract Valid Until: 2021		
		ARGENTINA Position: RS Jersey Number: 10		

图 2 网站部分截图

3.3 数据收集

数据收集采用爬虫的形式进行。在对网站结构进行研究和数据抓包分析后，编写 python 脚本对网站上需要的数据进行自动化爬取并对爬取到的数据进行自动化地分类和预处理。

```

1 import pandas as pd
2 import re
3 import requests
4 from bs4 import BeautifulSoup
5
6 # Get basic players information for all players
7 base_url = 'https://sofifa.com/players/overall'
8 columns = ['ID', 'Name', 'Age', 'Photo', 'Nationality', 'Flag', 'Overall', 'Potential', 'Club', 'Club Logo', 'Value', 'Wage', 'Special']
9 data = pd.DataFrame(columns = columns)
10
11 for offset in range(0, 100):
12     url = base_url + str(offset + 1)
13     source_code = requests.get(url)
14     plain_text = source_code.text
15     soup = BeautifulSoup(plain_text, 'html.parser')
16     table_body = soup.find('tbody')
17     for row in table_body.findAll('tr'):
18         td = row.findAll('td')
19         picture = td[0].find('img').get('data-src')
20         pid = td[1].find('a').get('id')
21         nationality = td[2].find('a').get('title')
22         flag_img = td[3].find('img').get('data-src')
23         name = td[4].find('a').text.strip()
24         age = td[5].find('div').text.strip()
25         overall = td[6].text.strip()
26         potential = td[7].text.strip()
27         club = td[8].find('a').text
28         club_logo = td[9].find('img').get('data-src')
29         value = td[10].text.strip()
30         wage = td[11].text.strip()
31         special = td[12].text.strip()
32         player_data = pd.DataFrame([pid, name, age, picture, nationality, flag_img, overall, potential, club, club_logo, value, wage, special])
33         player_data.columns = columns
34         data = data.append(player_data, ignore_index=True)
35     data = data.drop_duplicates()

```

图 3 爬虫脚本部分截图

3.4 数据概况

在数据收集阶段，Python 脚本共从 sofifa 网站中爬取并处理 18207 条球员数据，其中每条数据包含 54 条属性。以下为数据的基本概括：

数据(缩写)	解释	举例	数据(缩写)	解释	举例
ID	编号	158023	BC	控球	96
NAM	姓名	L. Messi	AL	加速	91
AGE	年龄	31	SS	速度	86
NAT	国家	Argentina	AG	敏捷	91
OVE	综合能力	94	RE	反应	95
POT	潜力	94	BAL	平衡	95
CLU	俱乐部	FC Barcelona	SP	射门力量	85
VAL	身价	€110.5M	JU	弹跳	68
WAG	周薪	€565K	STA	体能	72
SPE	总能力	2202	STR	强壮	59
PF	惯用脚	Left	LS	远射	94
IR	国际影响力	5	AGG	侵略性	48
WF	逆足能力	4	INT	拦截意识	22
SM	花式技巧	4	POI	跑位	94
WR	积极性	Medium	VIS	视野	94
BT	身体模型	Messi	PEN	点球	75
RF	真实脸型	Yes	COM	沉着	96
POS	位置	RF	MAR	防守意识	33
JN	球衣号码	10	ST	抢断	28
CR	传中	84	ST	铲球	26
FIN	射术	95	GKD	鱼跃	6
HA	头球精度	70	GKH	手形	11
SP	短传	90	GKK	开球	15
VOL	凌空	86	GKP	站位	14
DRI	盘带	97	GKR	反应	8
CUR	弧线	93	RC	违约金	€226.5M
FK	任意球精度	94			
LP	长传	87			

表 1 数据概况

4. 数据处理

4.1 数据清洗

- ① 处理无效值和缺失值。
- ② 根据数据集的特点以及各数据的类型及特点进行数据清洗，剔除无作用的数据属性。

数据名	ID	Name	Club	Body Type	Jersey Number
解释	编号	姓名	俱乐部	身体模型	球衣号码

表 2 删除数据说明

4.2 数据转化

使用python的第三方模块sklearn中的LabelBinarizer将非数值数据转化为可被量化的二进制变量。同时，为了防止变量的冗余以及变量间出现共线关系，删去最后一列数据，即灰色一列。

		平方和	自由度	均方		F	显著性
回归	SSR	653435.549	27	MSR	24201.317	7676.664	.000 ^b
残差	SSE	46390.252	14715	MSE	3.153		
总计	SST	699825.801	14742				

表 3 数据转化说明

4.3 设置变量

经过数据清洗和数据转化，共有 53 个自变量，以及 1 个相应变量。

数据名	类型	变量	数据名	类型	变量	数据名	类型	变量
DF	0/1	X1	Value	int64	X11	FKAccuracy	float64	X21
DM	0/1	X2	Wage	int64	X12	LongPassing	float64	X22
AM	0/1	X3	Special	int64	X13	BallControl	float64	X23
MF	0/1	X4	Crossing	int64	X14	Acceleration	float64	X24
Low1	0/1	X5	Finishing	float64	X15	SprintSpeed	float64	X25
Medium1	0/1	X6	HeadingAccuracy	float64	X16	Agility	float64	X26
.Low2	0/1	X7	ShortPassing	float64	X17	Reactions	float64	X27
Medium2	0/1	X8	Volleys	float64	X18	Balance	float64	X28
Age	int64	X9	Dribbling	float64	X19	ShotPower	float64	X29
Potential	int64	X10	Curve	float64	X20	Jumping	float64	X30
数据名	类型	变量	数据名	类型	变量	数据名	类型	变量
Stamina	float64	X31	StandingTackle	float64	X41	WeakFoot	float64	X51
Strength	float64	X32	SlidingTackle	float64	X42	MajorNation	float64	X52
LongShots	float64	X33	GKDividing	float64	X43	RightFoot	float64	X53
Aggression	float64	X34	GKHandling	float64	X44	Overall	int64	Y
Interceptions	float64	X35	GKKicking	float64	X45			
Positioning	float64	X36	GKPositioning	float64	X46			
Vision	float64	X37	GKReflexes	float64	X47			
Penalties	float64	X38	InternationalReputation	float64	X48			
Composure	float64	X39	ReleaseClause	int64	X49			
Marking	float64	X40	SkillMoves	float64	X50			

表 4 变量表

5. 数据分析

5.1 相关性分析

5.1.1 相关系数热力图

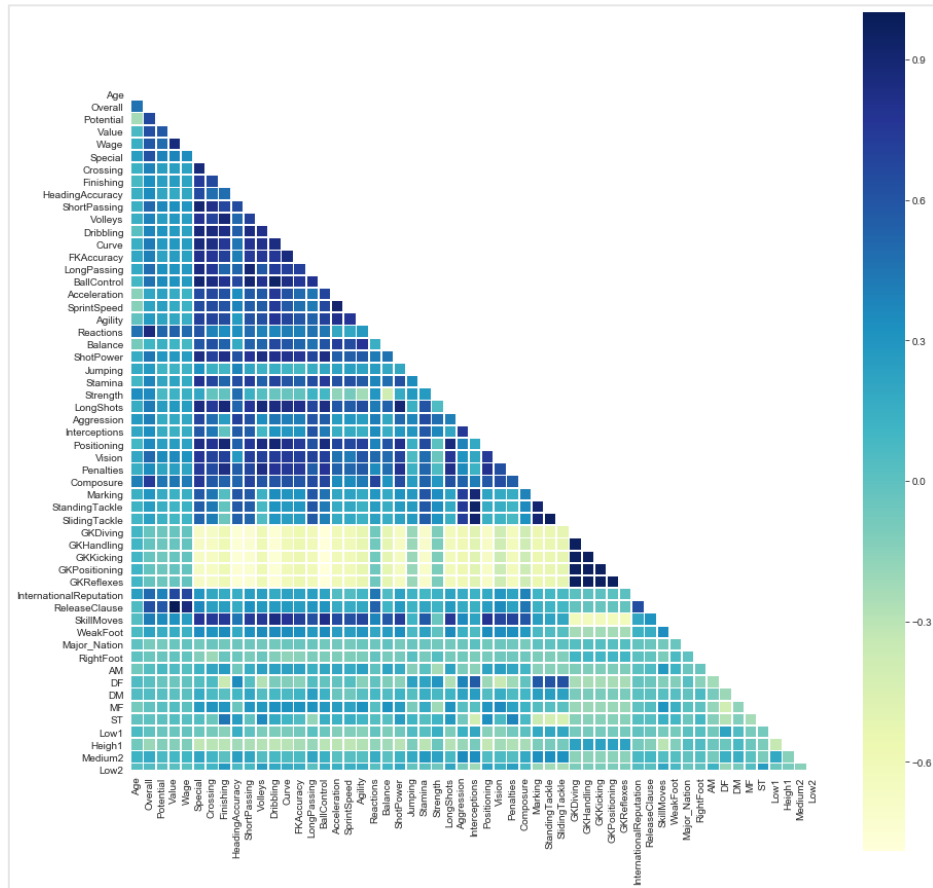


图 4 相关系数热力图

5.1.2 部分系数相关性

	Age	Overall	Potential	Value	Wage	Special	Crossing	Finishing
Age	1	0.4650	-0.2368	0.0765	0.1485	0.2461	0.1378	0.0775
Overall	0.4650	1	0.6633	0.6292	0.5739	0.6103	0.3985	0.3344
Potential	-0.2368	0.6633	1	0.5854	0.4906	0.3868	0.2477	0.2418
Value	0.0765	0.6292	0.5854	1	0.8615	0.3789	0.2507	0.2575
Wage	0.1485	0.5739	0.4906	0.8615	1	0.3483	0.2332	0.2147
Special	0.2461	0.6103	0.3868	0.3789	0.3483	1	0.8682	0.7286
Crossing	0.1378	0.3985	0.2477	0.2507	0.2332	0.8682	1	0.6614
Finishing	0.0775	0.3344	0.2418	0.2575	0.2147	0.7286	0.6614	1

表 7 部分系数相关性的结果表

由相关系数热力图和相关系数的结果表可以看出，各个自变量与因变量间存在着较为明显的线性关系。

5.2 模型假设

响应变量 **overall** 为数值型数据，即定量指标，经过处理的解释变量全部都为定量指标，由于响应变量可能受到其他多个解释变量的影响，采用多元线性回归分析。

设多元线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

其中， $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 为偏回归系数，表示在其他自变量保持不变时， x_i 中增加或减少一个单位时的期望值的平均变化量。 ϵ 代表随机误差，本次分析中假设随机误差是随机的、独立的、服从正态分布的。

5.3 正态性检验

在多元线性回归中，回归分析的因变量 Y 需服从正态分布，因此对 **overall** 进行正态性检验。

5.3.1 直方图

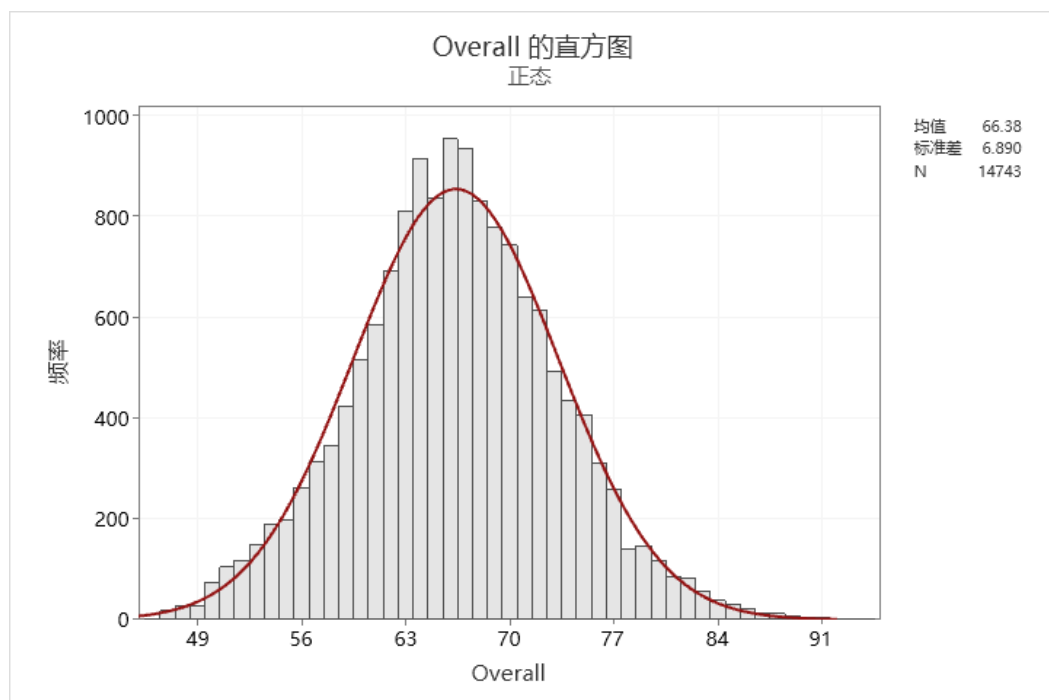


图 5 overall 的直方图

根据直方图可以十分直观地看出，因变量 Y 服从正态分布。

5.4 模型建立

通过使用 python 的 statsmodels 包中的 OLS 模型，将数据预处理得到的 53 个自变量和 1 个因变量进行多元线性回归，以此对影响球员综合能力因素进行分析。由于变量较多，通过自写程序，使用逐步回归法（step-wise）进行分析，将不合理的变量剔除。同时，重复计算各变量的 VIF 值，并根据各数据的权值，剔除掉线性相关性紧密的自变量。经过模型自动识别以及人工处理，最终余下 25 个自变量。

5.4.1 系数表

	非标准化系数		标准化系数	t	p	95%置信区间	
	B	标准误差	Beta			[0.025	0.975]
常数	-7.54	0.347	-	-21.743	0.000**	-8.2202	-6.8607
Age	0.512	0.006	0.341	85.635	0.000**	0.5002	0.5236
Potential	0.456	0.005	0.404	98.624	0.000**	0.4474	0.4656
Value	0	0	0.111	33.022	0.000**	0	0
Finishing	0.021	0.002	0.05	9.708	0.000**	0.0167	0.0252
ShortPassing	0.072	0.003	0.103	22.47	0.000**	0.0656	0.0781
Dribbling	-0.018	0.003	-0.033	-6.128	0.000**	-0.0237	-0.0122
BallControl	0.088	0.004	0.129	21.894	0.000**	0.0804	0.0962
Acceleration	0.022	0.003	0.038	7.348	0.000**	0.0161	0.0278
SprintSpeed	0.033	0.003	0.055	12.087	0.000**	0.0276	0.0382
Agility	0.008	0.002	0.014	3.49	0.000**	0.0034	0.0122
Reactions	0.142	0.003	0.182	45.436	0.000**	0.1357	0.1479
Balance	-0.017	0.002	-0.031	-8.768	0.000**	-0.0212	-0.0135
Strength	0.056	0.002	0.102	31.299	0.000**	0.0525	0.0596
LongShots	0.006	0.002	0.013	3.056	0.002**	0.0021	0.0094
Vision	-0.032	0.002	-0.06	-13.558	0.000**	-0.0365	-0.0273
Composure	0.055	0.003	0.081	20.507	0.000**	0.0495	0.06
Marking	0.016	0.002	0.039	9.854	0.000**	0.0125	0.0187
InternationalReputation	-0.595	0.051	-0.035	-11.587	0.000**	-0.6954	-0.4942
SkillMoves	0.575	0.037	0.051	15.448	0.000**	0.5019	0.6477
MajorNation	-0.289	0.035	-0.018	-8.137	0.000**	-0.3584	-0.2192
AM	-0.815	0.076	-0.037	-10.784	0.000**	-0.9633	-0.667
DM	-1.139	0.063	-0.047	-18.206	0.000**	-1.2619	-1.0166
MF	-1.13	0.057	-0.073	-19.897	0.000**	-1.2414	-1.0187
ST	-1.115	0.08	-0.06	-13.904	0.000**	-1.2722	-0.9578
Low1	0.559	0.067	0.019	8.327	0.000**	0.4274	0.6906
* p<0.05 ** p<0.01							

表 8 系数表

5.4.2 模型摘要

Df Model	25	F	8288.524	D - W	1.74
Df Residuals	14717	Prob(F)	p=0.000	JB	573.225
AIC	58796.6764	R²	0.934	Prob(JB)	0
BIC	58994.238	调整 R²	0.934	Log - Likelihood	-29372

表 9 模型摘要

5.4.3 ANOVA

	平方和	自由度	均方	F	显著性
回归 SSR	653435.549	27	MSR	24201.317	7676.664
残差 SSE	46390.252	14715	MSE	3.153	
总计 SST	699825.801	14742			

表 10 ANOVA

5.4.4 评估指标解释

- ① SSE: 残差平方和, 拟合数据和原始数据对应点的误差的平方和

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

- ② MSE: 均方差, 拟合数据和原始数据对应点误差的平方和的均值

$$MSE = \frac{SSE}{n} = \frac{1}{n} \cdot \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

- ③ RMSE: 拟合标准差, 的平方根

$$RMSE = \sqrt{MSE} = \sqrt{SSE/n} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

在该模型中, $RMSE = 1.7543$, 接近于 0, 说明模型选择和拟合更好

- ① SSR: 该统计参数计算的是预测数据与原始数据均值之差的平方和

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- ② SST: 该统计参数计算的是原始数据和均值之差的平方和

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad SST = SSE + SSR$$

- ② R²: 确定系数

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- ③ 调整 R²: 校正决定系数

在模型中, $R^2 = 0.934$ 且调整 $R^2 = 0.934$, 这意味着模型中的可以解释因变量 Overall (球员综合能力) 的 93.4% 的变化原因。同时, 调整 R² 和 R² 差别较小, 过拟合现象不明显。

- ① Log - Likelihood: 极大似然函数

② *AIC*: 最小化信息量准则 (*Akaike Information Criterion*)

$$AIC = 2k - 2 \ln L$$

k : 模型参数的个数

L : 模型的极大似然函数

③ *BIC*: 贝叶斯信息规则 (*Bayesian Information Criteria*)

k : 模型参数的个数

n : 模型样本的个数

L : 模型的极大似然函数

根据*AIC*和*BIC*的公式, 增加自由变量的数目提高了拟合的优良性, *AIC* 鼓励数据拟合的优良性但是尽量避免出现过拟合的情况。所以优先考虑的模型应该是和值较小的那一个。根据*AIC*准则和*BIC*准则可以找出最好的解释数据但是包含最少自由参数的模型。该模型的*AIC* = 58796.6764, *BIC* = 58994.238, 是逐步回归过程中最小的。随着变量的删除或者增加, 模型的*AIC*和*BIC*相比该模型都是增加的, 说明起始模型就是当前最优的模型。

5.4.5 模型公式

$$\begin{aligned} Overall = & -7.540 + 0.512 \times Age + 0.456 \times Potential + 0.000 \times Value + \\ & 0.021 \times Finishing + 0.072 \times ShortPassing - 0.018 \times Dribbling + 0.088 \times \\ & BallControl + 0.022 \times Acceleration + 0.033 \times SprintSpeed + 0.008 \times \\ & Agility + 0.142 \times Reactions - 0.017 \times Balance + 0.056 \times Strength + 0.006 \times \\ & LongShots - 0.032 \times Vision + 0.055 \times Composure + 0.016 \times Marking - \\ & 0.595 \times InternationalReputation + 0.575 \times SkillMoves - 0.289 \times \\ & MajorNation - 0.815 \times AM - 1.139 \times DM - 1.130 \times MF - 1.115 \times ST + \\ & 0.559 \times Low1 \end{aligned}$$

5.5 模型检验

5.5.1 显著性检验

① F 检验

对回归方程整体进行显著性检验

设

H_0 : 所有的系数都为 0 ($\beta_1 = \beta_2 = \dots = \beta_k = 0$)

H_1 : 至少一个系数不为 0

	平方和	自由度	均方		F	显著性
回归 SSR	653435.549	27	MSR	24201.317	7676.664	.000 ^b
残差 SSE	46390.252	14715	MSE	3.153		
总计 SST	699825.801	14742				

表 10 ANOVA

根据 ANOVA 表格, 可得

$$F = 8288.524, \quad p(F) = 0.000$$

$$\because p(F) = 0.000 < 0.005, \text{ 且 } p(F) = 0.000 < 0.001$$

\therefore 拒绝原假设 H_0 , 接受备择假设 H_1 。

该线性回归方程的整体方程系数显著异于零, 至少有一个解释变量对因变量有影响, 整体显著性较强。

② T 检验

分别检验回归模型中各个回归系数是否具有显著性, 即检验解释变量的系数是否在规定的显著性水平上显著。对于回归系数 β_i ($1 \leq i \leq 11$):

设

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

根据表 8 得

$$\because 25 \text{ 个自变量的值均小于 } 0.01$$

\therefore 拒绝原假设 H_0 , 接受备择假设 H_1 。

该线性回归方程各个解释变量的系数显著异于零, 说明每一个解释变量都对因变量有着显著影响。

5.5.2 多重共线性检验

多元线性回归的前提是自变量相互独立，不存在线性关系，因此需对自变量进行多重共线性检验，同时也是对自变量进行二次筛选。

① 方差膨胀因子（VIF）

方差膨胀因子是指回归系数的估计量由于自变量共线性使得方差增加的一个相对度量。一般建议，在变量过多的时候，如 $VIF > 10$ ，表明模型中有很强的共线性问题。VIF 的计算公式如下：

$$VIF_j = \frac{1}{1 - R_j^2}$$

变量	VIF	变量	VIF
Intercept	562.294436	Strength	2.369188
Age	3.5263	LongShots	4.088401
Potential	3.725898	Vision	4.320387
Value	2.530054	Composure	3.488781
Finishing	5.81457	Marking	3.463168
ShortPassing	4.633059	InternationalReputation	1.978878
Dribbling	6.318279	SkillMoves	2.459522
BallControl	7.681081	MajorNation	1.027326
Acceleration	5.799357	AM	2.557433
SprintSpeed	4.586648	DM	1.497409
Agility	3.587554	MF	2.990033
Reactions	3.578224	ST	4.071623
Balance	2.706114	Low1	1.13847

表 11 各变量的值

由表 11 可得

各变量的方差膨胀因子均小于 10，其中大于 5 的仅有 4 个变量，最大为 $7.681081 < 10$ ，因此所有变量之间不存在共线性问题。

5.5.3 残差分析

在多元线性回归模型的构建中，对残差做了随机、独立、服从正态分布的假设，多元线性回归建立在该假设的基础上。因此，应对该假设进行验证。

① 残差序列的正态性检验

非正态的残差会造成参数的置信区间不可靠，因为它们是使用正态性假设来构造的。因此有必要对残差进行正态性检验。

设

H_0 : 残差为正态分布

H_1 : 残差不为正态分布

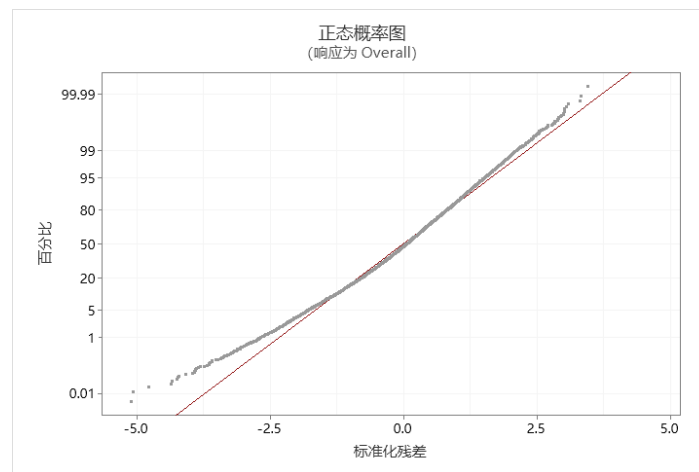


图6 标准化残差 P-P 图

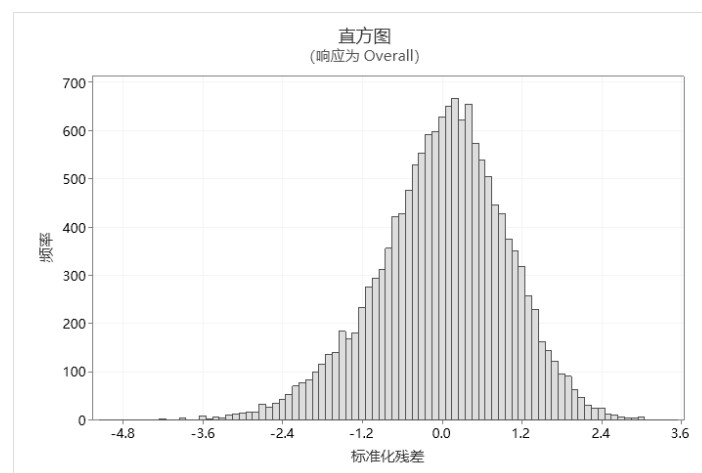


图7 标准化残差直方图

由标准化残差的累积概率图以及直方图得

残差在低值和高值的时候，与正态分布存在偏差。其他基本上为正态分布，故不拒绝原假设，说明模型的可靠性较高。

② 异方差检验

由于异方差的存在使得最小二乘估计量不再是最好的线性无偏估计量，会导致模型的残差不再是同方差的，所以要对模型进行异方差检验。

设

H_0 : 残差具有恒定的方差 (homoscedastic)

H_1 : 残差具有不恒定的方差 (heteroscedastic)

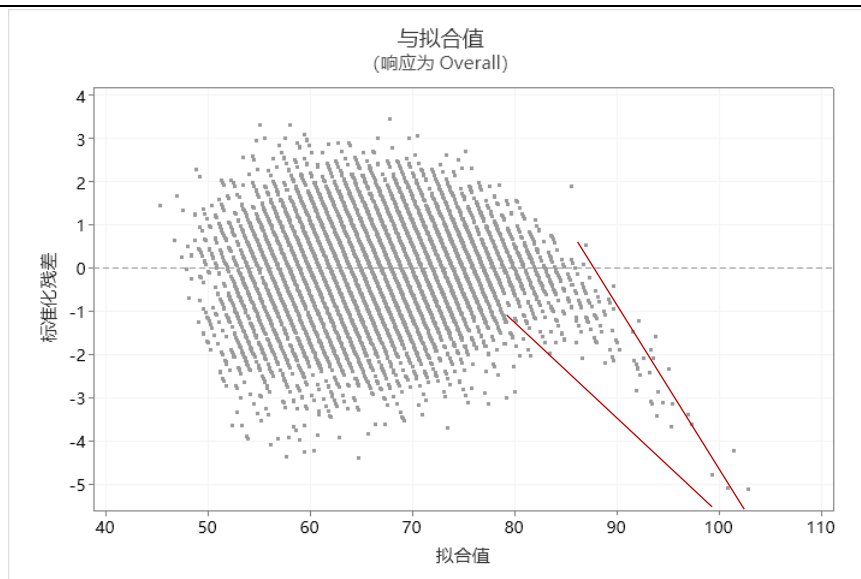


图 8 残差和预测值的散点图

由残差和预测值的散点图可得

当拟合值为 50 到 90 时，图中各点随机分布在过 0 点的直线两侧，没有固定的分布模式，所以不拒绝原假设 H_0 ，说明当拟合值为 50 到 90 的时候，不需要考虑异方差。

但当拟合值达到 90 以上时，绝对值标准化残差出现异常增大的情况，所以拒绝原假设 H_0 ，接受备择假设 H_1 ，说明当拟合值达到 90 以上时，需要考虑异方差。

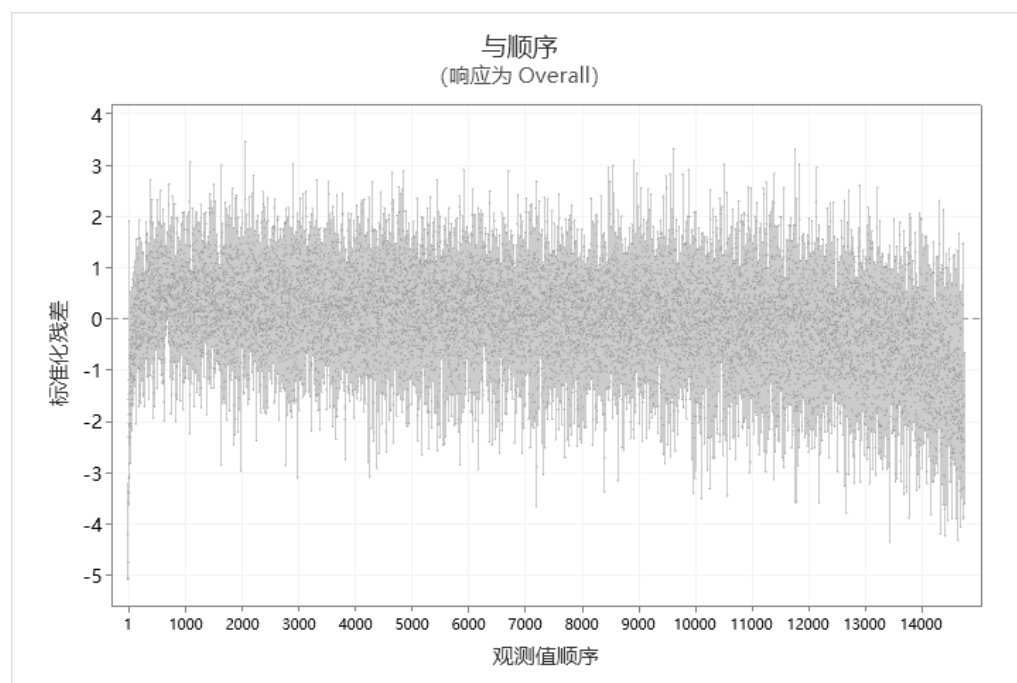
③ 残差独立性检验

误差存在自相关时，模型中的系数用最小二乘估计计算会不准确，往往会算出的系数的真实方差值和误差项的方差值会偏小。为了检验得到的方程的准确性，需要进行自相关检验。

设

H_0 : 残差不自相关

H_1 : 残差自相关



由 DW 检验 Durbin-Watson test statistic 得

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

$DW < 2$ 说明残差存在正自相关性（常见）。

$DW \approx 2$ 说明残差不存在自相关性（理想）。

$DW > 2$ 说明残差存在负自相关性（少见）。

由表 9 得

$$\because DW = 1.74 \approx 2$$

\therefore 不拒绝原假设 H_0 ，说明残差不存在自相关性。

5.5.4 异常观测值

异常值指在所获得的数据中相对误差较大的观测值，也称离群值。

异常观测值的拟合和诊断									
观测值	Overall	拟合值	残差	标准化残差	删后残差	杠杆率	Cook		
1	94	102.888	-8.888	-5.09	-5.09	0.0329983	0.03	R	X
2	94	101.422	-7.422	-4.21	-4.21	0.0149354	0.01	R	X
3	92	100.808	-8.808	-5.06	-5.07	0.0394575	0.04	R	X
4	91	99.329	-8.329	-4.76	-4.76	0.0288139	0.03	R	X
5	91	97.292	-6.292	-3.59	-3.59	0.0238134	0.01	R	X
6	91	95.091	-4.091	-2.32	-2.32	0.0115001	0	R	X
7	91	96.953	-5.953	-3.38	-3.38	0.0170411	0.01	R	X
8	91	93.787	-2.787	-1.58	-1.58	0.0071567	0		X
9	90	95.5	-5.5	-3.12	-3.12	0.014855	0.01	R	X
10	90	93.611	-3.611	-2.05	-2.05	0.0150219	0	R	X
11	90	92.117	-2.117	-1.2	-1.2	0.0057815	0		X
12	90	93.294	-3.294	-1.86	-1.86	0.0089735	0		X
13	89	91.582	-2.582	-1.46	-1.46	0.0117362	0		X
14	89	95.377	-6.377	-3.63	-3.64	0.0236025	0.01	R	X
15	89	93.999	-4.999	-2.84	-2.84	0.0199069	0.01	R	X
16	89	94.462	-5.462	-3.1	-3.1	0.0153747	0.01	R	X
17	89	89.71	-0.71	-0.4	-0.4	0.008078	0		X
18	89	92.664	-3.664	-2.07	-2.07	0.0091442	0	R	X
19	89	93.223	-4.223	-2.39	-2.39	0.0102388	0	R	X
20	89	85.614	3.386	1.91	1.91	0.0066442	0		X
.....									
共有 787 个异常值									

表 12 异常观测值的拟合和诊断

若杠杆值超过杠杆的比率为

$$2 \frac{k+1}{n} = 2 \frac{24+1}{14743} \approx 0.0037$$

则属于高杠杆点，说明这些记录的一个或多个解释变量与该变量的平均值相差很大。

由表 12 得
 数据样本总数为 14743，其中有 787 个观测值被标记。说明有 787 条记录的残差值异常，即预测的酒精含量与实际的酒精含量之差的绝对值超过标准差的 2 倍，占比为 5.3%。这些异常观测值的占比不高，对模型造成影响比较小。

6. 结论

通过以上模型建立、分析及验证过程，最终的多元线性回归模型为

$$\begin{aligned} Overall = & -7.540 + 0.512 \times Age + 0.456 \times Potential + 0.000 \times Value + \\ & 0.021 \times Finishing + 0.072 \times ShortPassing - 0.018 \times Dribbling + 0.088 \times \\ & BallControl + 0.022 \times Acceleration + 0.033 \times SprintSpeed + 0.008 \times \\ & Agility + 0.142 \times Reactions - 0.017 \times Balance + 0.056 \times Strength + 0.006 \times \\ & LongShots - 0.032 \times Vision + 0.055 \times Composure + 0.016 \times Marking - \\ & 0.595 \times InternationalReputation + 0.575 \times SkillMoves - 0.289 \times \\ & MajorNation - 0.815 \times AM - 1.139 \times DM - 1.130 \times MF - 1.115 \times ST + \\ & 0.559 \times Low1 \end{aligned}$$

同时，该模型对 Overall（球员的综合能力）有着很好的解释能力。

- ① 调整 R^2 和 R^2 接近 1，说明能解释因变量较多的变化原因。
- ② RMSE 接近 0，说明模型的误差较小，拟合度好。
- ③ 通过 F 检验和 t 检验，说明模型对因变量解释的显著性好。
- ④ 通过多重线性检验，说明模型的自变量具有较好的独立性。
- ⑤ 通过残差检验，说明模型的误差是随机、独立、服从正态分布。
- ⑥ 异常值占比不大，说明模型的准确性较为优秀。

以下是模型的实际预测图：

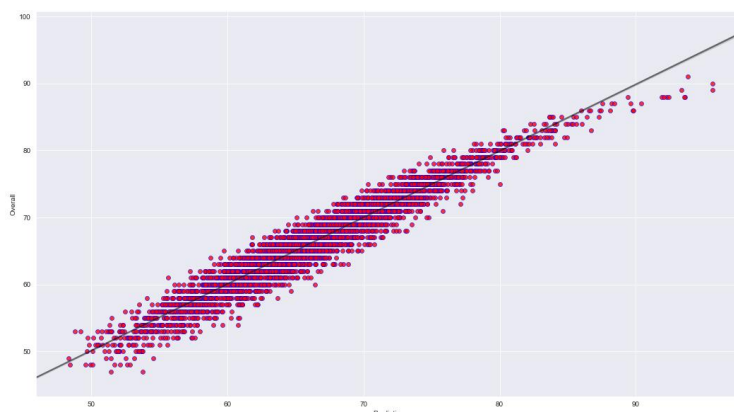


表 12 模型的实际预测图

综上所述，模型的整体效果不错，能通过设定球员的各项数据，对球员的综合能力进行较好的估计。