五大联赛职业球员身价影响因素分析

一、研究背景

足球是世界上最流行的体育运动之一。据不完全统计，全球有超 8 亿的狂热球迷，从事足球产业的人数高达一亿人次，约有 10 万的职业足球运动员。一方面，足球具有极强的竞技性和观赏性，能够让球迷极大程度地去享受比赛；另一方面，极具个人魅力的足球运动员也是推广这项运动的重要因素之一。一名优秀的足球运动员无论在商业上，还是就体育本身而言都是俱乐部的最重要的“资产”之一。

随着足球商业化进程不断加快，各大财团相继入主足球联赛后，“金元足球”的到来进一步的推动了足球在世界范围内的流行。这也为足球赛事带来了巨大的商业价值，足球场上飞驰的不仅仅是足球，更是数以亿计的真金白银。在这样的背景下，足球运动员的身价与工资也开始水涨船高，经常会天价转会费的新闻出现，成为球迷们茶余饭后的谈资。

欧洲是现在足球产业发展最早的地区，英格兰正是现代足球的发源地。职业足球联赛在欧洲已经有了 100 多年的发展历史，以五大联赛（英超、西甲、德甲、意甲、法甲）为代表的顶级职业足球联赛在世界范围内吸引了无数球迷，创造了巨大的商业价值。在当今世界足坛中，五大联赛代表了足球最高竞技水平，拥有最成熟的产业运作体系。在欧足联的统一管理下，球员的跨联赛转会机制成为了各大俱乐部交易球员，球员们追求自我成就的制度保障。

球员的身价能够反映出一个球员的能力、潜力以及近期表现和市场的需要，对于俱乐部、球员以及整个联赛的运作都十分的重要。因此，能够合理、准确地评估出职业球员的身价就成为了一个重要的工作。由于欧洲的五大联赛发展历史悠久、球员评价体系较为完整，因此本案例将收集现役五大联赛的球员数据，对职业球员身价的相关影响因素展开研究。

二、数据介绍

（一）数据来源

本文数据来自于 FIFA 22（sofifa. com）的职业模式的球员数据库，FIFA22 中大约有 1. 8 万名职业球员的信息，除了国籍、俱乐部、出生日期、工资、身价等基本信息外，每名球员又额外拥有超过 30 项能力值数据。EA 公司拥有一支结构完整且人数众多的数据库团队，能够获取世界各地职业球员的一手信息。因此无论从数量上、准确性上还是更新速度上，FIFA 的球员数据库都可以说是全球最专业的球员数据库。其球员数据库主页如图 2-1 所示。

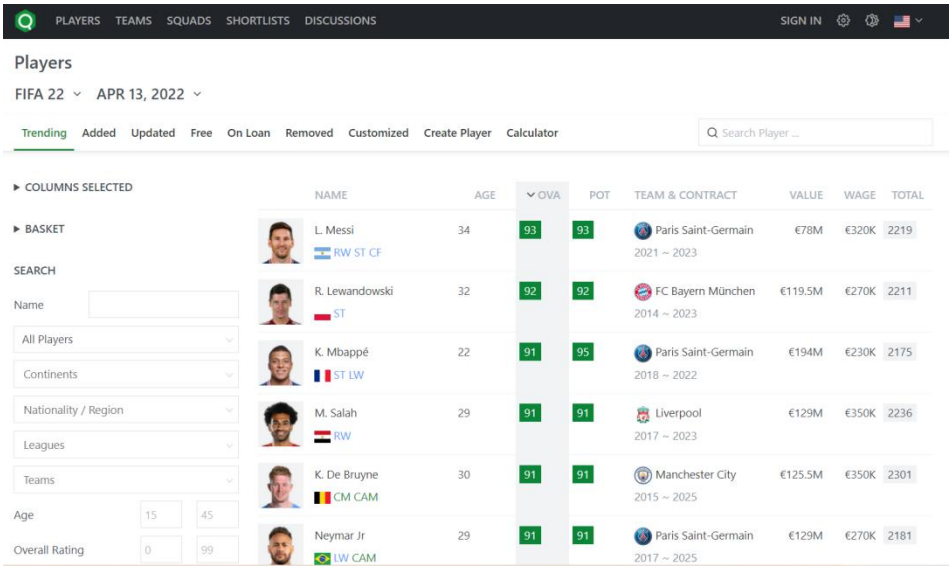

图 2-1　FIFA 球员数据库网站

现有的客观数据，包括射门得分率、传球成功率等等，其实很难直接证明一个球员的真正实力，而与球员在球队中的战术定位、球队整体状态甚至整个联赛的风格和水平有着很大的关系。所以将各项能力数值化后，往往比现有的赛事数据更值得相信。

（二）数据说明

本文将收集 FIFA22 公布的现役于五大联赛的球员信息，由于守门员的评价体系较为特殊，故剔除守门员的数据，共得 2342 条记录，数据采集时间为 2022 年 4 月。数据共包含 9 个变量，因变量为球员身价，其他变量为自变量。具体的变量说明如表 2-1 所示。

表 2-1 数据变量说明表

变量类型		变量名	详细说明	取值范围
因变量		球员身价	单位：万欧元	13-19400
自变量	基本信息	年龄	单位：岁	16-39
		所在联赛	定性变量：共五个水平	英超、西甲、德甲、意甲、法甲
		场上位置(除去门将)	定性变量：共三个水平	前场、中场、后场
	状态表现	总能力值	衡量球员现在的综合能力	51-93
		射门	衡量球员的射门能力	18-94
		传球	衡量球员的传球能力	29-93
		防守	衡量球员的防守能力	17-91
		速度	衡量球员的速度水平	32-94

自变量分为基本信息和状态表现两部分，基本信息包括球员的年龄、所在联赛、场上位置等，这些数据为现有客观数据。状态表现则主要以 FIFA22 的评分员对于球员的各项能力

的百分制评价为准。

三、描述性分析

（一）因变量：球员身价

根据 FIFA22 公布的数据来看，在除去守门员之外的现役五大联赛职业球员中，身价的最小值为 13 万欧元，所对应的球员为效力于英格兰超级联赛狼队的 19 岁球员埃斯特拉达（P. Estrada）。这位来自奥地利的小将在俱乐部中司职中后卫；最大值为 19400 万欧元，所对应的球员就是各大豪门争先抢要的法国巨星基利安·姆巴佩（K. Mbappé），今年 22 岁的姆巴佩现效力于法甲联赛的豪门巴黎圣日耳曼，自俄罗斯世界杯后，这位法国前锋的身价就一直占据着转会市场的榜首位置，这也符合姆巴佩在联赛中的出色表现。

通过对球员的身价分布绘制频数直方图（图 xx），从图中可以明显的观察到：球员身价呈现出显著的右偏分布。样本中所有球员的平均身价为 1038 万欧元，中位数为 390 万欧元。其中大部分球员的身价在 1000 万以下，超过了样本总量的 70%；存在少数身价超过 5000 万欧元的“天价球员”，占样本总数的 3. 35%。这部分球员有力地拉高了球员的平均身价。总的来讲，五大联赛球员的身价的平均水平较高，但是存在着较大的差距，尤其是顶尖球星的身价与普通球员之间存在着极大的差距。

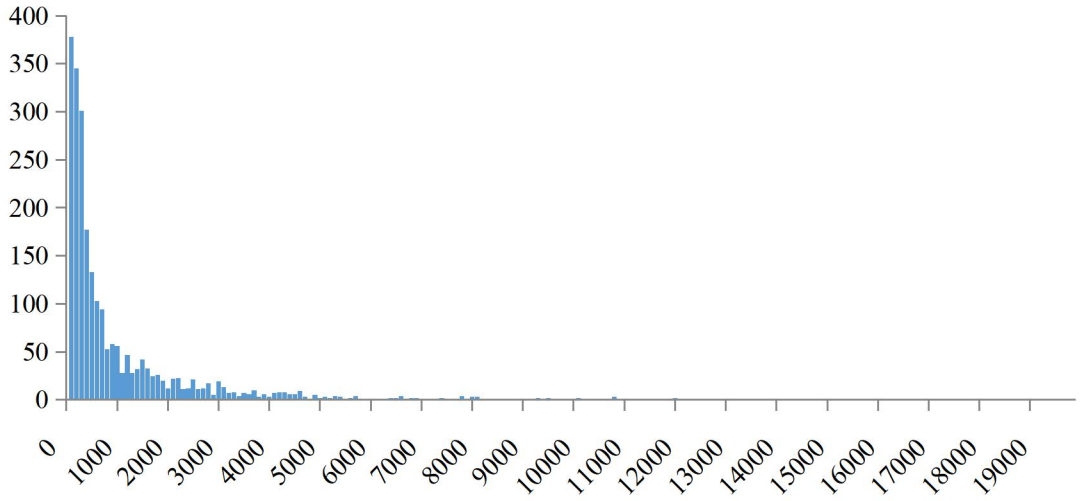

图 3-1 五大联赛球员身价的分布情况

（二）自变量：基本信息

图 3-2 反应了样本中球员的联赛分布情况。在选取的 2342 名球员中，来自英超联赛的球员数量最多，达到了 584 人，占比接近 25%；来自西甲联赛的球员数量最少，只有 389 人；其他三大联赛人数相差较小，分别为 436 人、450 人和 483 人。总体而言，不同联赛球员数量占样本总体比例较为合理。英超球员较多也符合其联赛规模较大的特点。

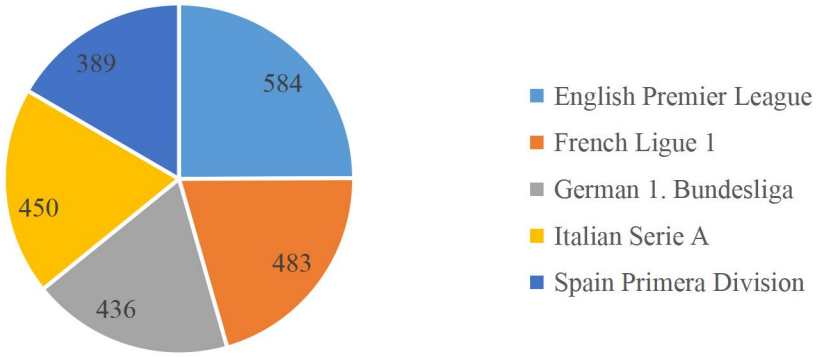

图 3-2 球员联赛分布情况

以球员所在联赛为分类标准，分别做出五大联赛球员身价的箱线图（图 3-3）。从图 3-3 可以看出，英超联赛的总体球员身价水平略高于其他联赛且其内部球员身价波动较大，西甲联赛居于其次，其余三大联赛球员的总体身价水平差异较小，初步猜想这与英超联赛和西甲联赛较好的商业运行与较高的竞技水平有关。

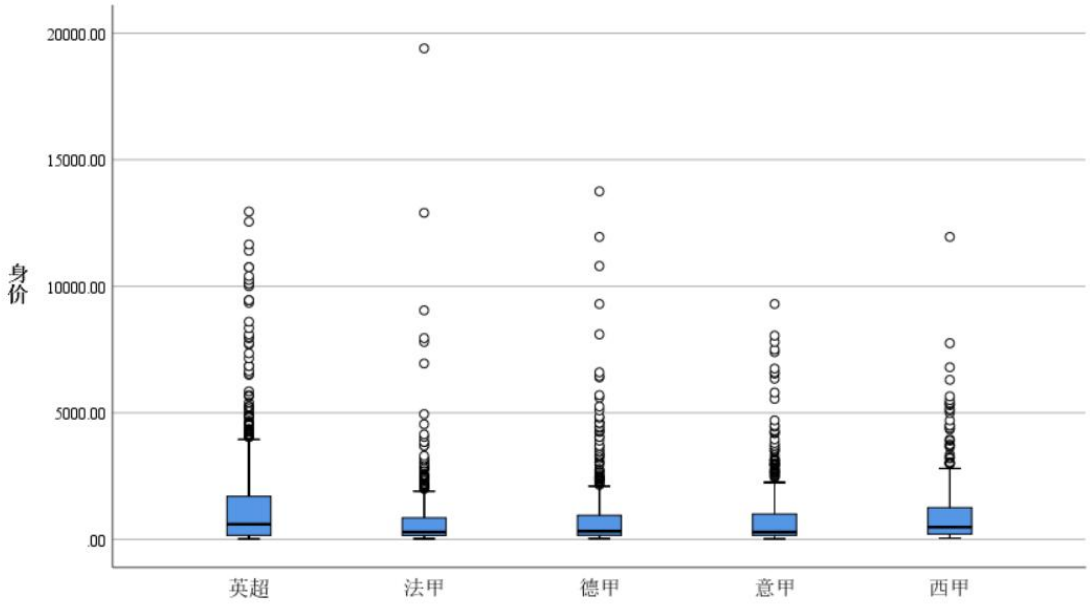

图 3-3 不同联赛的球员身价分布情况

在“高价球员”部分尽管各大联赛都出现了部分天价球员的并且最高值出现在法甲联赛，但是依然可以看到英超联赛在这一方面相较于其他联赛较大的优势。

在所有效力五大联赛的球员（不包含守门员）中，年龄最大为 39 岁，所对应的球员是现效力于西甲联赛的格拉纳达俱乐部的豪尔赫·莫利那（Jorge Molina），场上司职前锋。年龄最小的球员是现效力于意甲联赛博洛尼亚俱乐部的卡斯帕·若班斯基（K. Urbański），在球队中场位置。

总体上看，球员年龄分布呈现右偏，即存在部分高龄球员，五大联赛大部分球员的年龄处于 18 岁至 30 岁之间，也有部分不到 18 岁的年轻球员。这样的数据特征也反映了职业足球对于球员身体素质的要求，呈现出年轻化的趋势。年龄较大的球员无法继续适应激烈的比赛，要么选择退役，要么前往更低级别的联赛继续效力；而越来越多的优秀年轻球员也被教练提拔到一线队伍中。

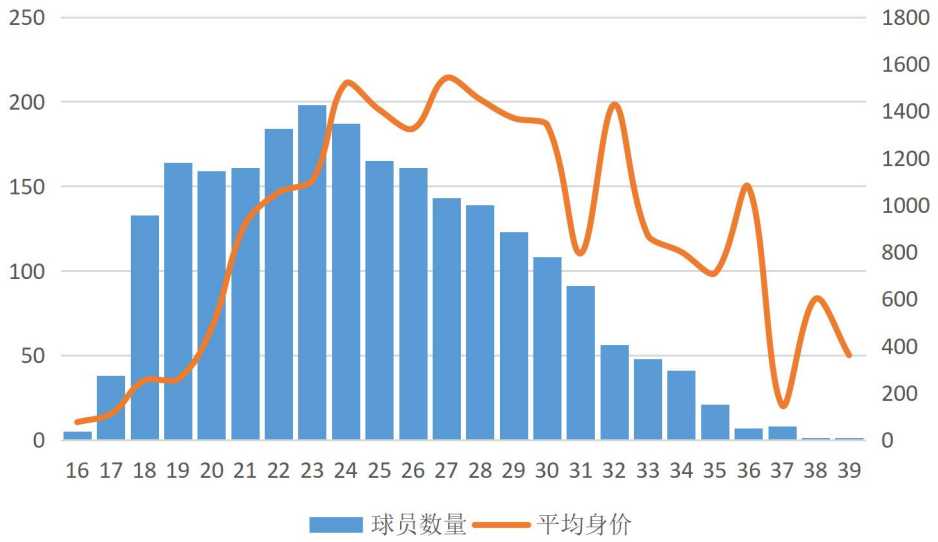

图 3-4 球员年龄分布及不同平均身价分布情况

对于不同年龄球员的平均身价，图 3-4 中呈现了一个随着年龄增长，球员身价先持续上升后逐渐下降的过程。在 24 岁以前，随着球员年龄增长，平均身价水平不断上升，在到达 30 岁之前一直维持着较高状态，30 岁之后开始呈现曲折下降的趋势。

从图 3-5 可以看出，五大联赛中现役球员中（不包含门将）中场球员数量最多，达到了样本总量的 42%，而前场球员数量最少，只有 21%。初步猜测这与现代足球的阵型中前场球员位置较少，大多讲究传控配合，所以对于中场球员要求较高，且中场球员运动范围往往比较灵活，可以同时参与进攻端与防守端的配合中，所以数量上会占有优势。

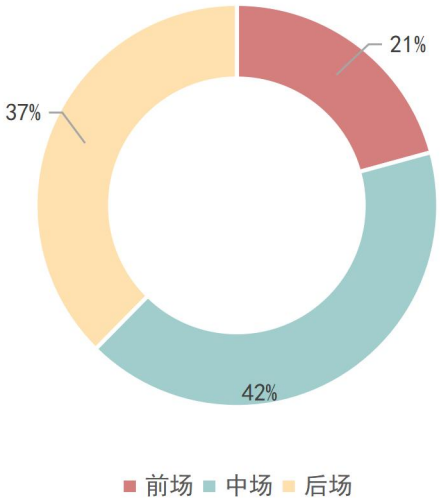

图 3-5 球员场上位置分布情况

比较不同位置球员的身价分布，绘制出其对应的箱线图（图 3-6）。从中可以看出，球员身价整体水平从高到低依次是：前场球员、中场球员、后场球员。在天价球员的比例上，也呈现出前场高于中场高于后场的趋势。这也符合我们对于足球的印象——出色的前锋往往更能让人印象深刻，往往具有更高的商业价值。

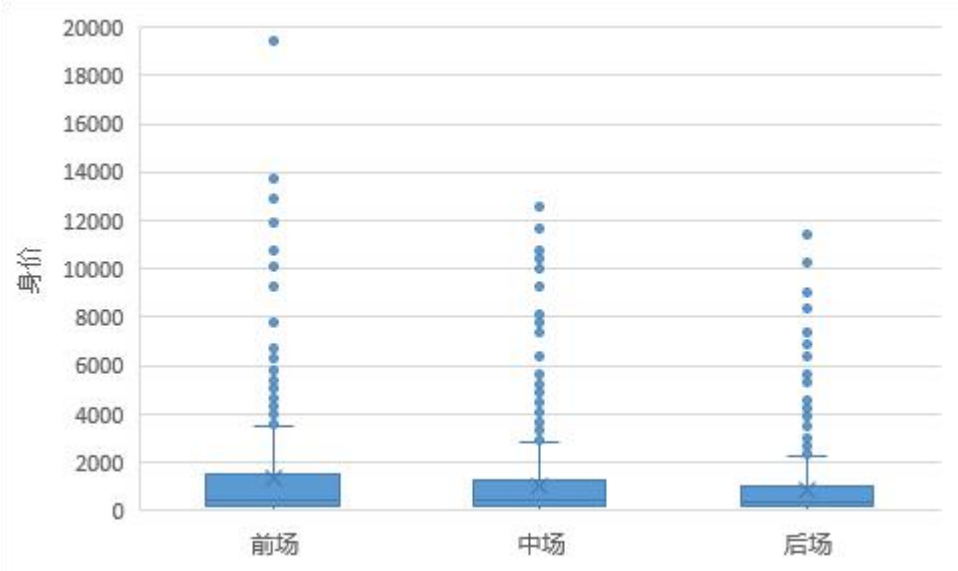

图 3-6 不同位置球员身价分布图

（三）自变量：状态表现

球员的总体能力反映了球员的综合素质，是综合考量球员现在的各项能力以及所在联赛和球队位置后得出的百分制评分。分数越高代表了球员目前的能力水平越高，从图 3-7 可以看出，随着球员的总能力值提高，其身价呈现出类似于指数上升的正相关趋势。

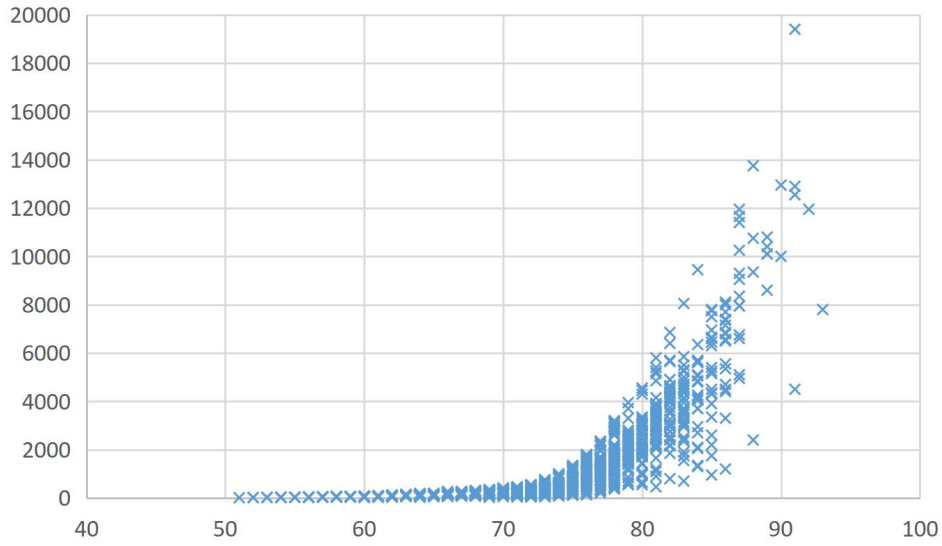

图 3-7 总能力值与身价散点图

此外，本文选取了速度、射门、传球和防守等四项能力来评价其对与球员身价的影响。

球员的速度水平反映了球员的身体素质，射门水平反映了球员的进攻能力，传球水平反映了球员的组织能力，防守水平反映了球员的防守能力。

图 3-8 中的四副散点图分别绘制了速度水平、射门水平、传球水平、防守水平和球员身价的相关关系。容易看出，球员的速度、射门和传球水平与身价存在一定的正相关关系。在防守水平方面，尽管其和身价存在着一定的正相关关系，但是在防守水平 30-50 的水平内出现了部分身价较高的球员，这部分球员往往是防守端贡献较少的前场球员。

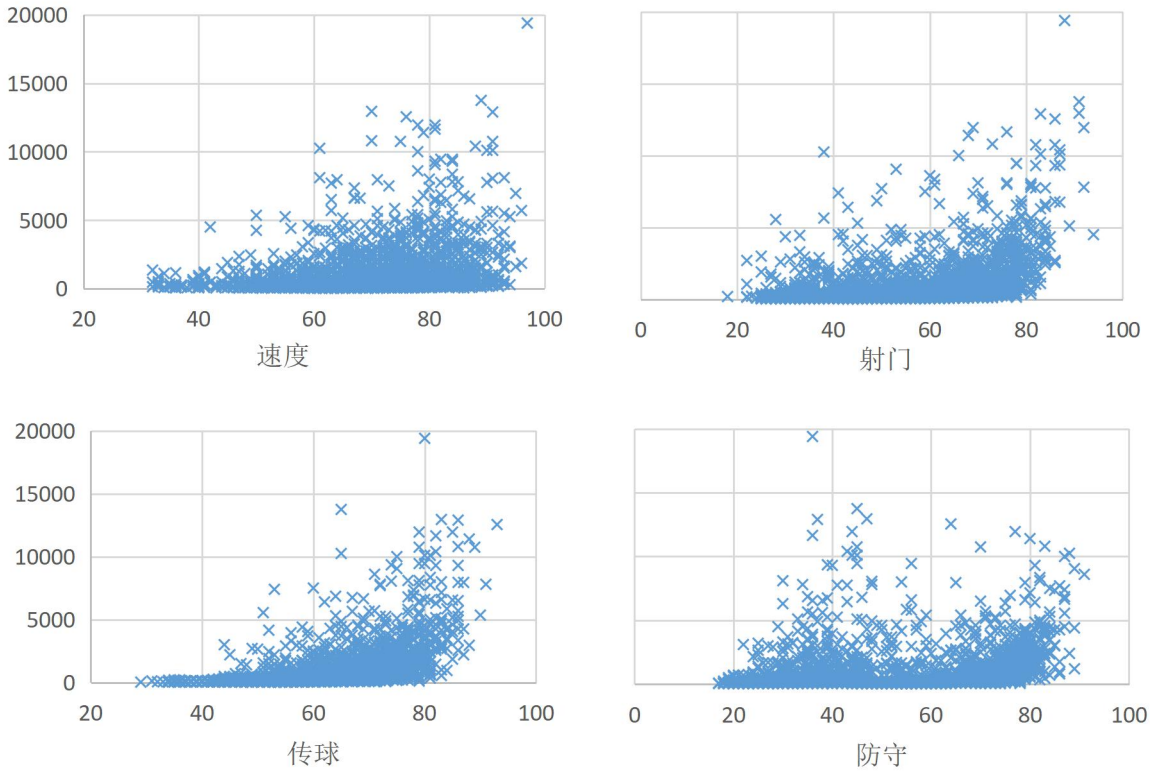

图 3-8 不同速度水平、射门水平、传球水平和防守水平和球员身价的散点图

综上所述，通过对于本文选取的数据变量的描述性分析，可以得到以下猜想：对于球员的身价可能会有影响的因素包括：球员的基本信息（年龄、所在联赛、场上位置）和状态表现（总能力值、速度、射门、传球、防守）；从影响的作用大小来看，基本信息中的年龄和状态表现中的总能力值较为明显。

四、回归建模

为了深入影响各个因素对于球员身价的影响，克服自变量之间的互相影响，本文构建了多元线性回归模型。回归分析是较为常见的统计方法之一，通过建立合适的回归模型可以用于对因变量的预测。通过对于回归分析结果的解读，我们可以对于球员身价评估有更深的了解。

（一）变量含义及取值依据

变量表示的含义以及取值范围如表 4-1 所示。

表 4-1 变量及变量说明

变量标识	变量名称	变量类型	变量取值范围
Y	球员身价	连续型变量，被解释变量	实际值
A	年龄	连续型变量	实际值
W_F	前场球员	虚拟变量	前场球员取 1，否则取 0
W_M	中场球员	虚拟变量	中场球员取 1，否则取 0
O	总能力值	连续型变量	实际值
G	射门	连续型变量	实际值
P	传球	连续型变量	实际值
D	防守	连续型变量	实际值
S	速度	连续型变量	实际值

（二）模型构建

基于表 4-1 的变量，对被解释变量 Y 取自然对数，建立线性回归模型如下：

$$\ln（Y_i）= \beta_0 + \beta_1 A_i + \beta_2 O_i + \beta_3 G_i + \beta_4 P_i + \beta_5 D_i + \beta_6 S_i + \beta_7 W_{Fi} + \beta_8 W_{Mi} + \varepsilon_i \quad （1）$$

其中β_0为常数项，表示在没有任何影响因素的情况下球员的身价，ε_i为随机扰动项。在处理虚拟变量时，引入W_F和W_M来表示球员的不同位置。

（三）实验结果

利用最小二乘法估计参数β的值，在 Eviews 中完成回归操作。

表 4-2 回归结果

变量标识	系数	标准误	t 值	Prob
A	-0.109226	0.001419	-76.97530	0.0000
O	0.210985	0.001263	167.0105	0.0000
G	0.001907	0.000745	2.559719	0.0105
P	0.001042	0.000937	1.112646	0.2660
D	0.000590	0.000538	1.097638	0.2725
S	-0.000237	0.000528	-0.449521	0.6531
W_F	0.070951	0.024693	2.873292	0.0041
W_M	0.028085	0.016398	1.712690	0.0869
β_0	-6.664394	0.055945	-119.1235	0.0000

更根据表 4-2 的结果，回归模型结果如下

$$ln（Y）=-6.66-0.109226A+0.210985O+0.001907G+0.001042P +0.000590D-0.000237S+0.070951W_F+0.028085W_M+\varepsilon \quad （2）$$

其预测值与真实值的拟合效果如图 4-1 所示：

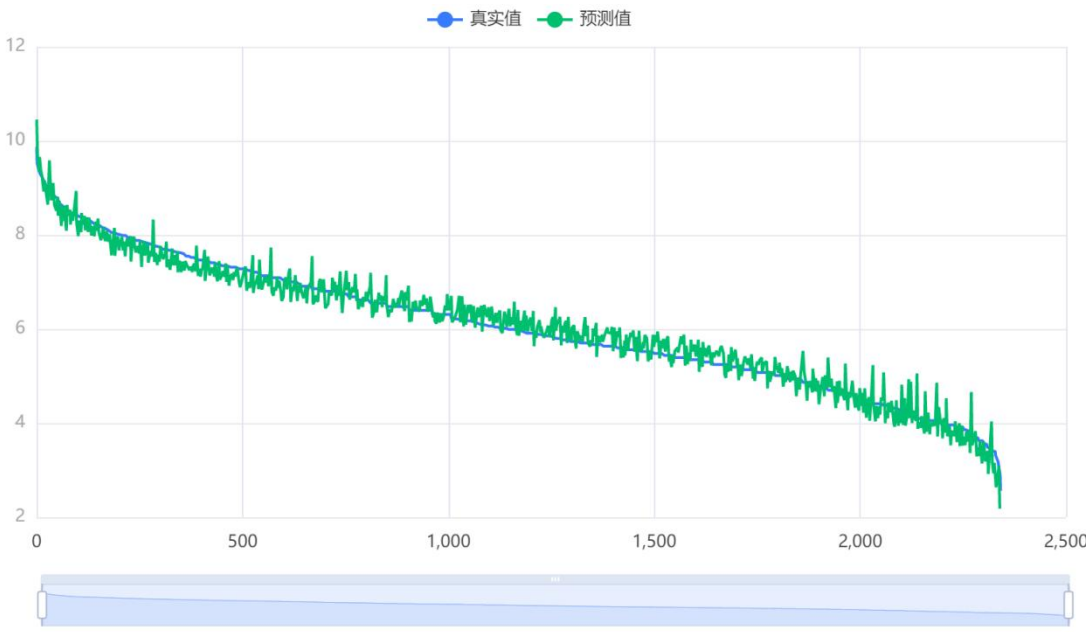

图 4-1 拟合效果图

（四）对回归模型的统计检验

1. 多重共线性检验

多重共线性是指贷多元回归模型中，解释变量之间存在一定的线性关系。多重共线性的出现回到值模型估计失真或难以估计准确。在本文使用方差膨胀因子（VIF）进行检验，方差膨胀因子是衡量多元线性回归模型中多重共线性严重程度的一种度量，表示了回归系数估计量的方差与假设自变量不线性相关是方差相比的比值。

首先，参数估计量的方差为：

$$Var(\widehat{\beta}_i) = \frac{\sigma^2}{\sum_{t=1}^{n}(x_{it} - \bar{x}_i)^2}\frac{1}{1 - R_i^2} \quad (3)$$

其中R_i^2是第 i 个解释变量作为因变量，对其他解释变量回归后的拟合优度：

$$x_{ij} = \alpha_0 + \alpha_1 x_{1j} + \alpha_1 x_{1j} + \alpha_1 x_{1j} + \ldots\ldots + \upsilon_j \quad (4)$$

根据（x）式我们得到方差膨胀因子 VIF：

$$VIF = \frac{1}{1 - R_i^2} \quad (5)$$

在 Eviews 中，对回归后的结果进行方差膨胀因子检验，得到结果如下表：

表 4-3 回归结果的方差膨胀因子检验

变量标识	A	O	G	P	D	S	W_F	W_M	β_0
VIF	1. 97	4. 07	6. 08	4. 67	4. 50	1. 71	4. 82	3. 14	NA

如果一个变量与其他的解释变量共线性程度越高，那么其对应的R_i^2就越大，VIF 值就越大。从表 4-3 的结果来看，模型中使用到的解释变量的 VIF 值均小于 10，认为解释变量间不存在多重共线性。

2. 异方差性检验

本例中将使用 white 检验来检验模型的异方差性，white 检验是检验异方差性最常用的方法之一。

在 Eviews 中进行异方差性检验，结果如下：

$$F = 11.36145，\quad Prob.F < 0.00001 \tag{6}$$

在 5%的显著性水平下拒绝原假设，即模型存在异方差性。

3. 拟合优度检验

为了衡量样本回归直线与样本观测值之间的拟合程度，构造调整后的可决系数来衡量其拟合优度，如下：

$$\overline{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)} \tag{7}$$

其中 RSS 表示残差平方和：

$$RSS = \sum \left(ln(Y_i) - ln(\widehat{Y}_i)\right)^2 \tag{8}$$

TSS 表示总离差平方和：

$$TSS = \sum \left(ln(Y_i) - ln(\overline{Y})\right)^2 \tag{9}$$

ESS 表示回归平方和：

$$ESS = \sum \left(ln(\widehat{Y}_i) - ln(\overline{Y})\right)^2 \tag{10}$$

可以得出：

$$TSS = RSS + ESS \tag{11}$$

而 n−k−1 为残差平方和的自由度，n−1 为总体平方和的自由度，在本例中，n=2342，k=8。

模型的摘要如表 xx 所示：

表 4-4 模型摘要

R^2	调整后R^2	F 统计量	Prob
0. 973351	0. 973259	10651. 51	0. 000000

调整后的R^2为 0. 973259，说明回归直线对于观测值的拟合程度较好。

4. 方程的显著性检验（F 检验）

方程的显著性检验，就是检验全部解释变量对于被解释变量的共同影响是否显著，即检验方程的参数是否显著不为 0。在本例中使用 F 检验来检验回归模型的总体显著性。

提出原假设与备择假设为：

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0, H_1: \beta_i (i = 1,2,\ldots, 8)$不全为0

构造 F 统计量：

$$F = \frac{ESS/k}{RSS/(n-k-1)} \sim F(k, n-k-1) \tag{12}$$

在本例中，F 统计量为 10651. 51，在 5%的显著性水平下，拒绝原假设，即方程具有显著

性。

5. 系数的显著性检验（t 检验）

方程的总体线性关系并不代表着每个解释变量对被解释变量的影响都是显著的，因此需要对每个解释变量进项显著性检验，本例中我们通过构造 t 统计量进行检验：

$$t_j = \frac{\widehat{\beta}_j - E(\widehat{\beta}_j)}{S_{\widehat{\beta}_j}} \tag{13}$$

从表 4-2 可以看出，在 5%的显著性水平下，P、D、S、W_M无法通过显著性检验，考虑其可能与模型存在异方差性有关，故考虑修正模型的异方差性。

（五）对于模型的修正

在本例中，为了修正模型的异方差性，对于原有的数据采取 Robust 回归。由于本例中使用最小二乘法过程中存在数据异常值，但是数据异常值并非数据输入错误或样本选取异常等人为原因引起，此时使用 Robust 回归分析，既不会将异常数据完全排除在外，也不会像最小二乘法一样将异常值与非异常值完全等同对待。

在 Eviews 中对原有数据进行 Robust 回归，估计方法使用 M 估计（M-estimation），权重计算公式使用 Fair 函数。

其中 Fair 函数计算的权重如下：

$$w = \frac{1}{1 + abs(resid)} \tag{14}$$

回归结果如下表：

表 4-5 Bobust 回归结果

变量标识	系数	t 值	Prob
A	-0.108761	-43841.49	0.0000
O	0.211979	95977.37	0.0000
G	0.001799	1381.645	0.0000
P	0.001245	760.4949	0.0000
D	0.000474	504.4096	0.0000
S	-0.000683	-740.6415	0.0000
W_F	0.073554	1703.780	0.0000
W_M	0.027706	966.4075	0.0000
β_0	-6.711837	-68622.00	0.0000

从表 4-5 可以看到，所有的系数都通过了稳健性检验，回归模型结果如下：

$$\begin{aligned} ln(Y) =& -6.711837 - 0.108761A + 0.211979O + 0.001799G + \\ & 0.001245P + 0.000474 - 0.000683S + 0.073554W_F + 0.027706W_M + \varepsilon \end{aligned} \tag{15}$$

五、结论与改进

（一）结论

本例对 FIFA22 中现役五大联赛职业球员（不包含门将）的信息进行统计分析，得到了如下结论：影响球员身价的主要因素有：球员年龄、球员场上位置、总能力值、射门能力、传球能力、防守能力、速度等。对于定量变量，年龄和速度对于身价的影响是负的，其余变量对于身价影响为正。对于定类变量，可以认为球员身价从前场到中场到后场递减，这也与我们一开始的猜想一致。关于速度对身价影响是负的，初步猜测可能与大量速度水平较高但身价较低的球员有关，且现代足球可能更加注重球员的得分效率、组织能力、防守能力而不是速度，且球员的速度水平无法直接转化为球场上的进球、助攻或是拦截、

（二）可以改进的地方

由于球员身价的影响因素很多，因此在后续的研究中可以考虑加入更多的因素，也可以考虑加入部分近期现实的球场数据，比如最近一个赛季的进球、助攻、关键传球、拦截等。此外，在本例中对于所有类型的球员我们共用了一套模型，其实对于不同位置的球员而言其身价评估体系存在着一定的差异，例如，对于后卫而言防守能力应当比射门能力更重要，对于前锋而言射门能力应当比防守能力更加重要。