基于回归分析的 NBA 球员工资制定以及建队策略

孙骏涛、李承祺、柏子杰

2022 年 12 月 16 日

摘要

本文旨在通过回归分析方法探究影响 NBA 篮球运动员合同价格（年均）的因素，并基于得到的模型对球员工资的制定进行建议以帮助球队完善建队策略。首先，我们对数据进行了简单的筛选，去除了短期合同等不具有代表性的球员数据，并通过描述性分析挖掘数据间存在的潜在关系。然后，在对因变量和部分自变量进行了合适的变换后，进行了变量选择和模型建立，并通过残差分析检验了模型的有效性，得出了有效的线性回归模型。接下来，我们继续考虑了自变量中的交互效应，建立了交互效应模型，进一步增加了模型的解释性。最后，我们通过数据分析得出结论，找到影响合同价格的显著性变量。

关键词：NBA、合同价格、线性回归、回归诊断、交互效应

目录

1 研究目的

随着统计学的发展和大数据技术在各个领域的广泛运用，对体育数据的分析开始逐渐走入了大家的视野，以美国职业篮球联赛 NBA 为代表的顶级联赛也开始采用越来越研究、越来越复杂的技术统计数据和衡量指标来记录一个球员的表现。这些技术统计不仅成为了球迷们津津乐道的话题，也成为了一个球员价值的最直接的体现。

而对于 NBA 这样的商业联盟来说，客观、科学地衡量一个球员的价值就更为重要了。首先，NBA 作为世界顶级体育联赛，拥有着惊人的体量。近几年来，NBA 球员的平均年薪一直保持在 700 万美元以上，2017-18 赛季更是达到了 777 万美元，而同年英超球员的平均工资只有 381 万美元。合同和交易上的一点改变产生的金额变动都不是小数目。

此外，NBA 的球队并不是“唯战绩论”的经营模式。对于球队老板来说，球队的表现固然十分重要，但如果因为球员溢价而导致球队盈利减少甚至亏损，那便得不偿失了。加之 NBA 还有着“工资帽”的规定，通过奢侈税限制球队通过天价合同收买过多有能力的球员，球队组建有争冠能力的队伍成本便更加高了。例如 21-22 赛季的冠军勇士队，不仅球队球员总薪资达到了 1.779 亿美元，还需要缴纳 1.703 亿的奢侈税，总支出近 3.5 亿美元。

不仅如此，不同于足球球员交易的“转会费”制度，NBA 的球员往往是和其身负的一定年限的合同绑定的，球队间的交易也往往是一个或多个球员（包括选秀权）的互换，换言之，也就是多个合同的互换。这其中的价值估算和衡量就更加的复杂，如何在满足球队需求的前提下达成合理的交易，无疑对球队经理们巨大的考验。

所以近几年来，球员的技术统计开始越来越多维、越来越细致，更能反应一个球员在球场上的表现和对球队的贡献。尽管仍存在如性格等难以量化的指

标，但通过数据分析来估计一个球员的价值正变得越来越可靠且有效。

在这样的背景下，本项目通过收集了 NBA 球员尽可能多的可靠的技术统计指标，旨在寻找其与球员合同年薪的关系，并建立有效的模型以用于估计球员的真实价值，供球队在交易和签订合同时参考。

2 数据来源与相关说明

本案例所用的数据为 2021-2022 赛季，球员基本信息来自于 NBA 官方网站，共 501 条，球员赛季数据来自于 Basketball Reference，共 605 条，工资数据来自于 ESPN，共 501 条。数据集中共有 50 个变量指标，其中球员工资(SALARY)为因变量，自变量分为球员基本信息和球员球场数据两个维度，变量名称与相关信息列于下表所示：

变量类型		变量名称	变量含义	取值范围
因变量		SALARY	工资（美元）	[5318，45780966]
自变量	球员基本信息	Pos	球员司职	C、PF、PG、SF、SG
		Age	年龄	[19,41]
		height	身高(cm)	[175,224]
		weight	体重(kg)	[73,131.5]
		Tm	所属球队	30 个水平
		country	国籍	43 个水平
		draft_year	选秀年份	[2002,2021]
		draft_round	选秀轮数	1、2
		draft_pick	选秀顺位	[1,60]
	球员场均数据统计	G	赛季场次	[0,82]
		GS	赛季首发场次	[0,82]
		MP	场均时间	[0,48]
		FG	运动战进球数	>=0
		FGA	运动战出手次数	>=0
		FG%	运动战命中率	[0,1]
		3P	三分进球数	>=0

		3PA	三分出手次数	>=0
		3P%	三分命中率	[0,1]
		2P	两分进球数	>=0
		2PA	两分出手次数	>=0
		2P%	两分命中率	[0,100]
		eFG%	真实命中率	[0,100]
		FT	罚球数	>=0
		FTA	罚球出手次数	>=0
		FT%	罚球命中率	[0,100]
		ORB	进攻篮板球	>=0
		DRB	防守篮板球	>=0
		TRB	总篮板球	>=0
		AST	助攻	>=0
		STL	抢断	>=0
		BLK	盖帽	>=0
		TOV	快攻	>=0
		PF	犯规	>=0
		PTS	得分	>=0
		PER	效率值	>=0
		TS%	真实命中率	[0,100]
		3PAr	三分球命中比率	
		FTr	罚球率	
		ORB%	进攻篮板率	
		DRB%	防守篮板率	
		TRB%	总篮板率	
		AST%	助攻率	
		STL%	抢断率	
		BLK%	盖帽率	
		TOV%	失误率	
		USG%	使用率	

表 1: 引入的数据意义和范围简介

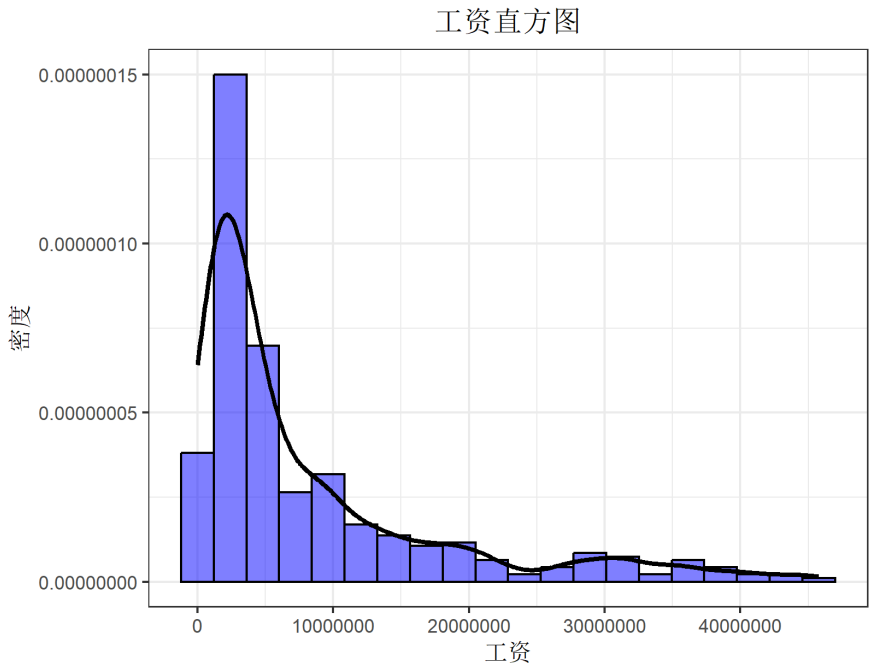
3　数据预处理与描述性分析

3.1 数据预处理

我们的数据初步选用的是过去一年在 NBA 联盟中有过出场经历的球员，但是考虑到球员的流动性较大，部分球员出场数太少导致数据没有说服力以及一些球员因为遭受重大伤病导致数据和因变量严重不匹配的情况，我们剔除了出场数在 20 场以下的球员。

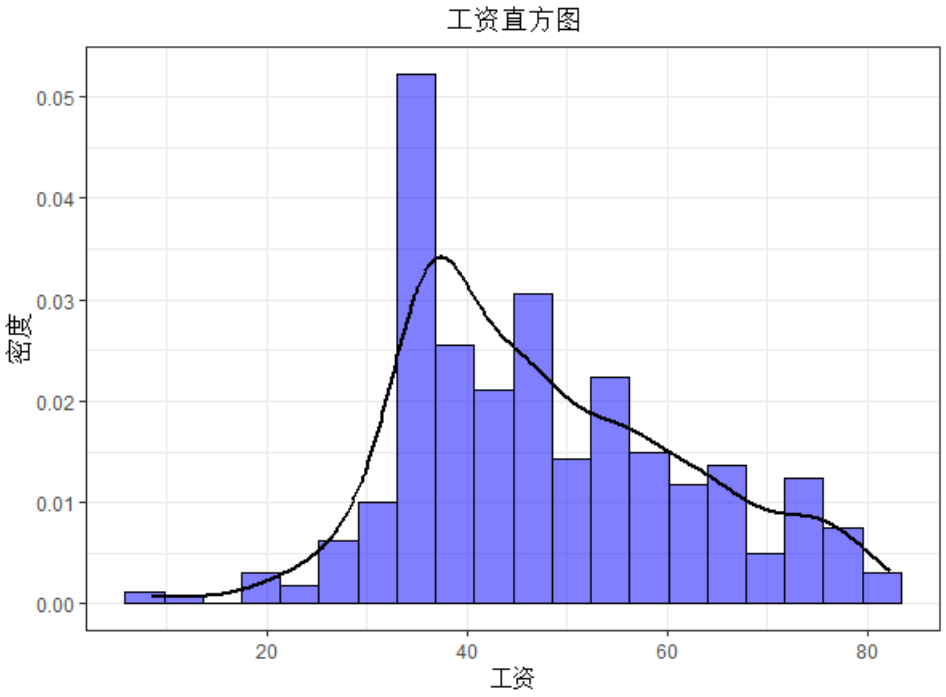
3.2 描述性分析

我们的目的是通过回归的方式分析出符合球员能力的工资，因变量是球员的工资。我们先针对这一连续变量做出其频率密度图，结果如下图所示。



图一：球员工资直方图

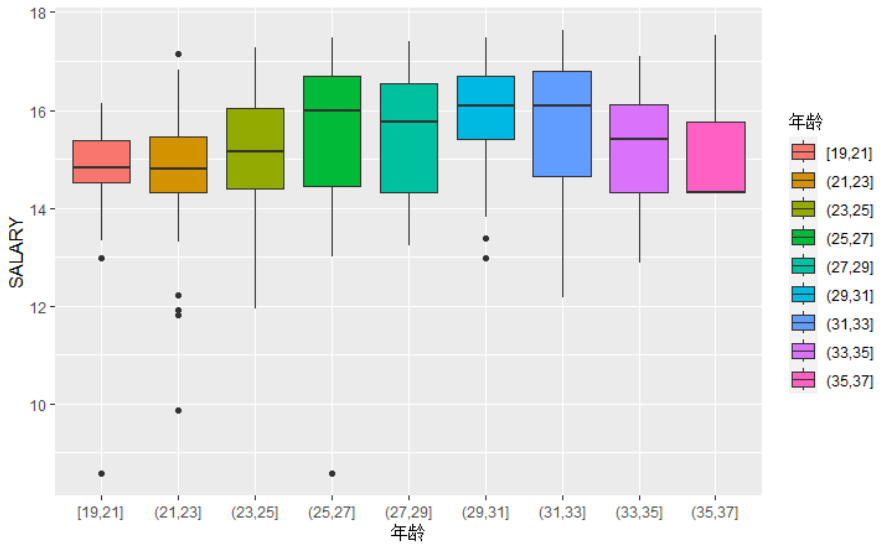
由上图可以看出，球员的平均工资为 7588082 元，而中位数仅有 3704426 元，呈现出了非常明显的右偏分布，并不符合数据正态条件。为了更好地分析球员工资和自变量的关系，我们对球员工资采取对数处理的变换方式并观察其分布，结果如下图所示。



图二：对数处理后的球员工资直方图

根据对对数图的初步观察，对数处理后的球员工资与正态分布较为接近。

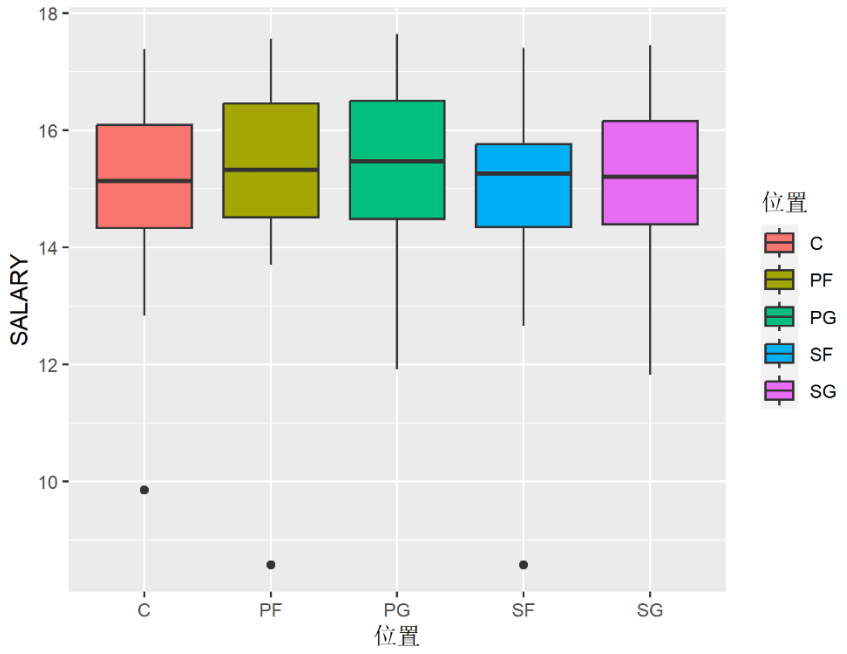
在处理了球员的工资数据后，我们针对各变量与工资的关系绘制了统计图进行描述性分析。首先我们研究球员的非场上数据以及一些属性变量与球员工资的关系。



图三：球员年龄与工资箱型图

年龄是评估一个运动员能力和潜力的重要指标，在对抗激烈的运动场上拥有良好的身体状态是高水平竞技的保证。我们通过对数处理后工资的箱型图研究年龄和对数后球员工资的关系。

由图三可以发现：球员工资随着年龄增加先增后减，在 29 至 31 区间中达到最高，且波动较小，说明在 29 到 31 岁的年龄段大多数球员均处于生涯的巅峰期，球队愿意付出更高的价格来匹配他们的能力。同时在 25 到 33 岁年龄段球员的工资中位数相差不大，75%分位数也高于其他年龄段，说明在这个年龄段中球员状态有保证，低于 25 岁的球员大多处于成长期，而高于 33 岁的球员能力会显著下滑。



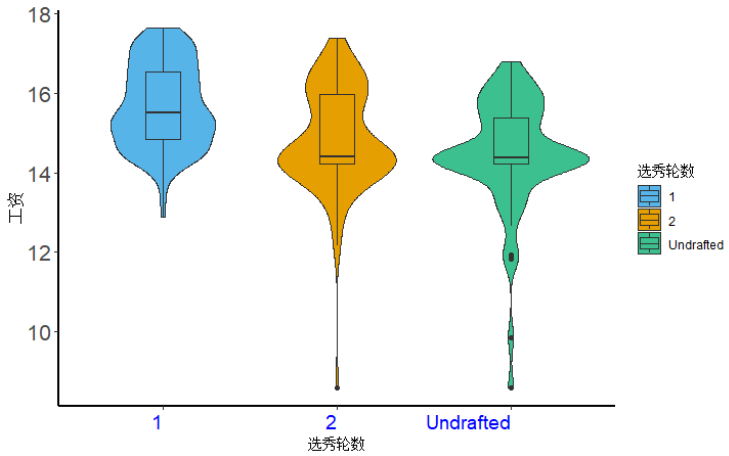
图四：不同位置的球员工资分布图

在大致了解了工资与年龄的关系后，我们再来研究一下位置与工资的关系。篮球场上的位置作用差异较大，因此对位置进行分类来研究与工资的关系是有必要的。通过对数处理的工资和位置的箱型图中我们可以看出：

1. 各位置工资差异总体不大，但控球后卫的工资几个重要分位数均高于中锋，这与现代篮球比赛的趋势有关，更多的投射和更快的比赛节奏弱化了中锋的作用，而控卫支配球的能力在比赛中愈发重要。
2. 小前锋的工资波动较小，主要是因为比赛中担任小前锋的球员大多可以承担一些其他位置的任务，虽然除了一些巨星特例外大多是万金油型球员，但仍是球队不可或缺的位置。

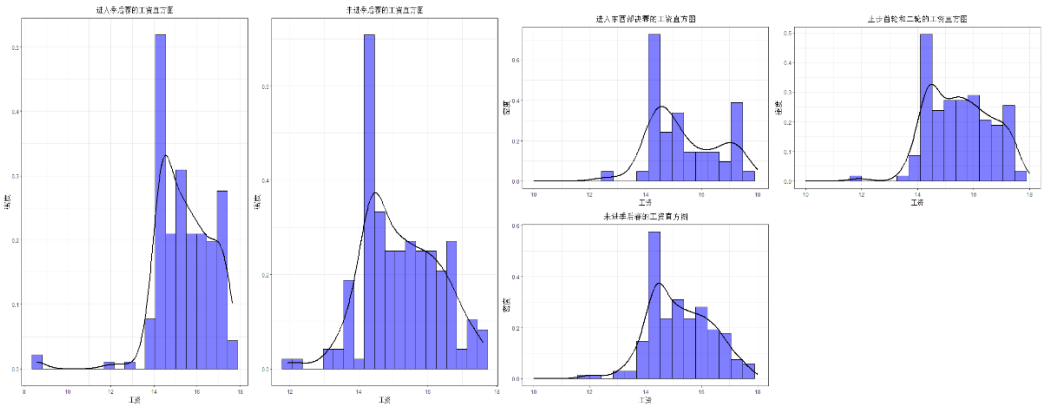
进一步我们研究选秀轮数与球员工资的关系。选秀是篮球联赛与众不同的地方，通过选秀引进最年轻有潜力的球员已经成为了 NBA 球队最稳定的建队方式。我们观察工资与选秀轮数的大提琴图可以看出，在第一轮被选中的球员工资水平显著的高于在第二轮被选中的球员以及落选秀。这是因为首轮选中的球员大多更有天赋和即战力，成材率更高，而成名球星也大多在首轮被选中，在上限和下限都有保证的情况下，首轮球员的工资更高是合理的结果。而二轮秀和落选秀

工资中位数差别不大，这其中的一个原因是落选秀大多无法留在联盟而不会被计入数据，导致样本中的落选秀数量较少，可能会有偏差。另一个原因是能留在联盟的落选秀大多有一技之长，因此与相对起点较低的二轮秀差距并不明显。但依旧可以看出二轮秀的 75%分位数工资水平显著高于落选秀，说明二轮秀的上限通常还是较落选秀更高。



图五：选秀轮数与工资的大提琴图

在观察了个人非数值变量与工资的关系后，我们希望观察球队的工资分布来寻找一些规律。图六、七展示了不同级别球队的工资分布，从图中我们可以看出没有进入季后赛的球队工资分布更为均匀接近正态，而进入季后赛的球队则存在着直方图左偏的情况，说明了要想进入季后赛争夺总冠军需要更多平均水平之上的球员。进一步将进入季后赛的球队分成进入分区决赛和未进入两类，可以看出进入分区决赛的球队分布更趋近于正态，但在对数工资接近 18 的顶端区间时有显著上升，说明想要获得更好的成绩球队需要有超级巨星独当一面；而普通季后赛球队虽然大多数球员在平均水准之上，但由于缺乏顶尖球员，所以也无法走得更远。

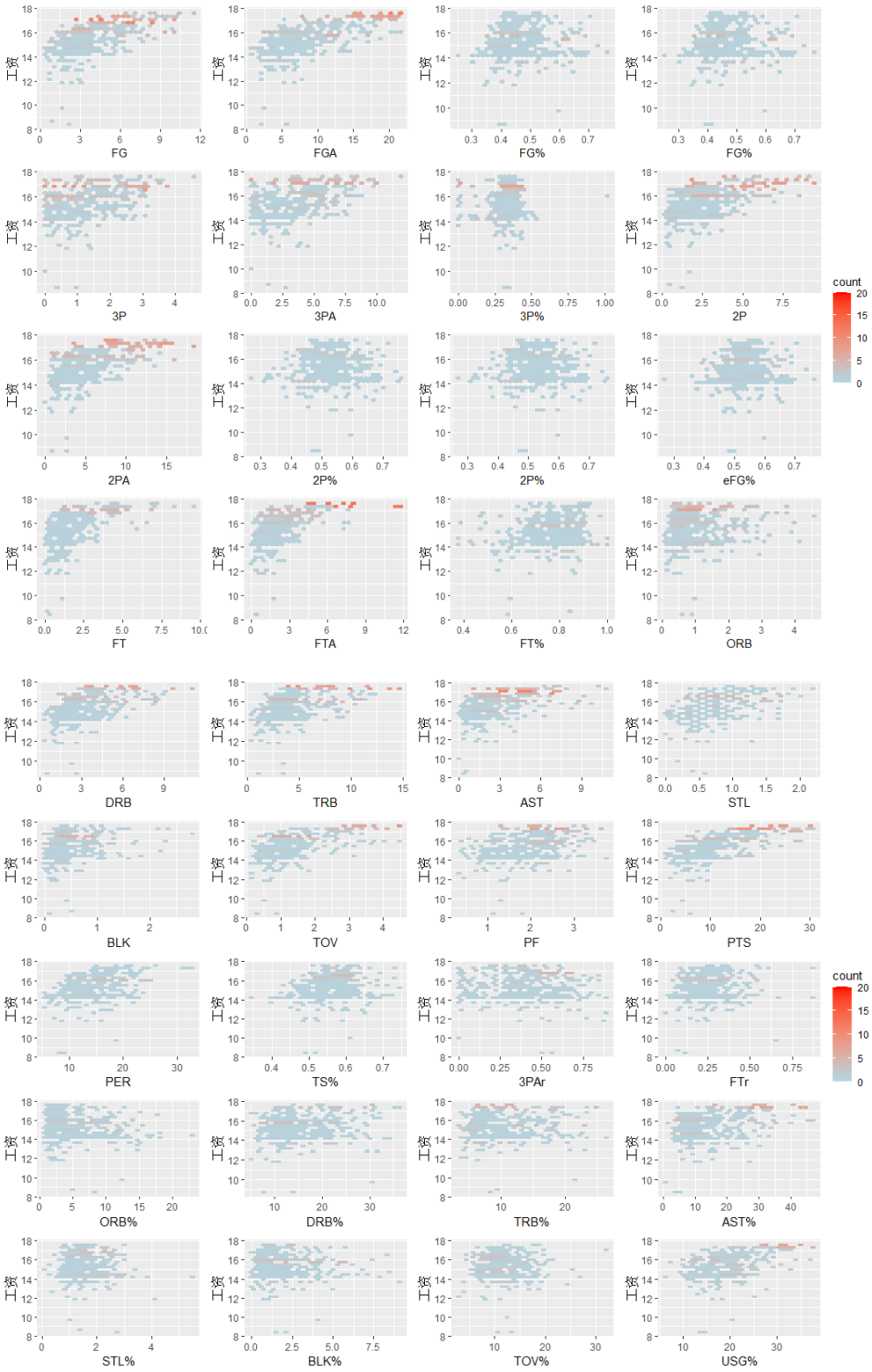


图六、七：基于不同级别球队的工资分布直方图

在观察完非场上数据以及属性变量与工资的关系后，我们再来研究球员场上数据与工资的关系以及内部联系。从图六、七可以看出两点主要信息：

大多数自变量与因变量存在正相关的关系，即数值越高，对应的球员工资越高，而这不仅体现在得分、命中率等正向指标中，也体现在失误等负向指标中。这是可以理解的，能力强的球星在比赛中拥有更多支配球的机会，在更多帮助球队的同时也会出现更多失误。

球员水平大多聚集在中等水平附近，在这个区间附近存在一些溢价和超值合同球员的存在。而对应指标数值最高的球员往往工资也较高，这体现在蜂巢图的最上端频数较高，说明能够在大多数指标上领先的球员通常也会获得较高的工资。

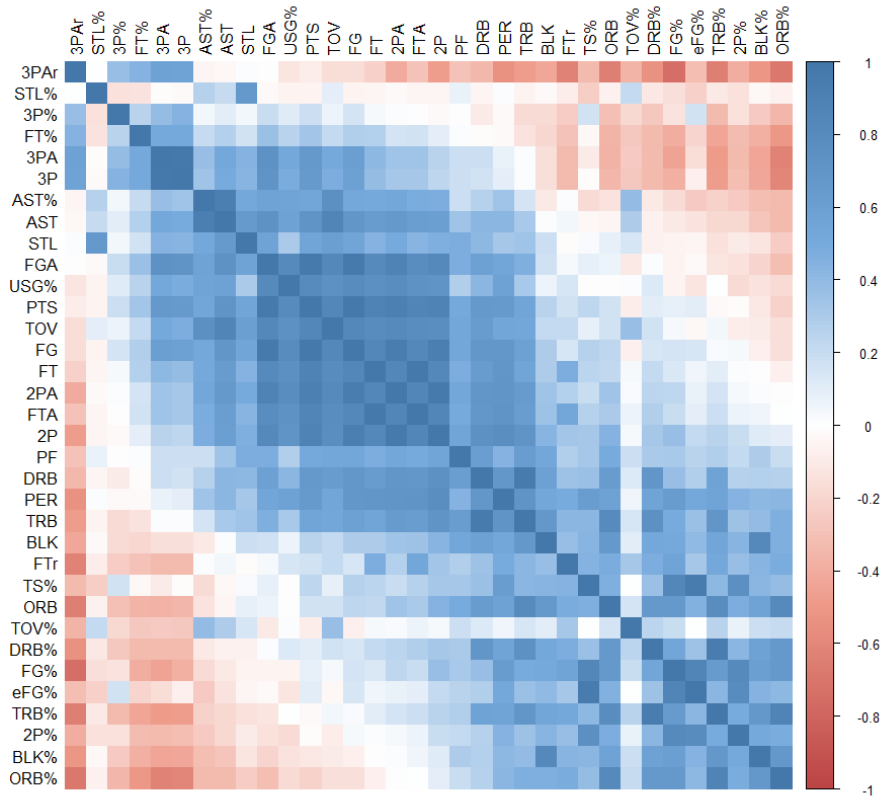


图八、九：球员各项能力与工资的蜂巢图

在了解了自变量与因变量的初步关系后，我们最后再来了解一下各自变量之间的关系。通过各项指标的相关系数热力图我们可以看出：

存在负相关性的变量较少，这与现代篮球的比赛方式有关，由于比赛节奏的加快和对于投篮、换防等指标要求的不断提升，球员的能力也变得越来越全面，一些功能性球员和有明显缺陷的球员逐渐被淘汰。而这也意味着我们不需要特意分离各位置的球员进行回归分析。

有一些自变量存在比较明显的相关性，由于事先无法得知究竟哪些指标更大程度上决定了球员的能力工资水平，我们选入了较多指标，这也不可避免地会导致一些指标的相关性，我们需要注意这些变量可能存在的复共线性。



图十：各项指标间的相关系数热力图

总的来说，描述性分析帮助我们对于因变量和自变量的关系有了初步的了解，我们需要从球员自身基本能力（如得分等）以及球员变量的相关性进行分析，也帮助我们在统计建模上选取主成分回归的方式，使得模型更为简洁的同时消除各变量间潜在的复共线性。并且我们会研究一些变量之间的交互效应，得出一些

属性变量之间的关系。

4 数据建模

4.1 模型建立

4.1.1 数据处理与变量选择

我们希望通过回归的方式建立有关球员工资与球员各项指标之间更具体的关系模型。我们从数据库中选入了大量篮球运动员的基础高阶数据作为自变量，但根据前文对于自变量的描述性分析我们可以知道自变量和因变量都需要进行一定程度的调整和选取才能更好地建立回归模型，是模型具有更强的准确性。

在之前的描述性分析中，我们通过绘制年龄和工资的箱型图可以看到球员工资与年龄存在先增后减的关系，我们认为这是球员年轻时经验技术不足，年龄增加时身体机能下滑导致，并且认为 29-31 岁之间是球员综合能力的巅峰期。因此我们对球员年龄进行处理，变换为$(29-age)^2$,使得变量的解释性更强。

1. 通过描述性分析，我们初步推断一些场上数据的自变量之间存在复共线性。其特征值最大为 30.38，最小值:为 30.38,0.001，两者的比值:为 15898

2. 在报告开头我们绘制了球员工资分布直方图，结果发现球员工资分布存在严重右偏倾向，不符合正态性假设。我们对球员工资使用了取对数的方式使其相对接近正态分布，可以进行各项分析建模，但在模型上仍有可以提升的地方，因此我们会在后续使用 box-cox 变换使残差符合正态分布，从而达到优化模型的效果。

3. 根据所学知识，我们对所有球员场上数据指标进行了主成分分析，提取了前五个主成分，解释占比 79.1%，得到主成分的因子的碎石图:

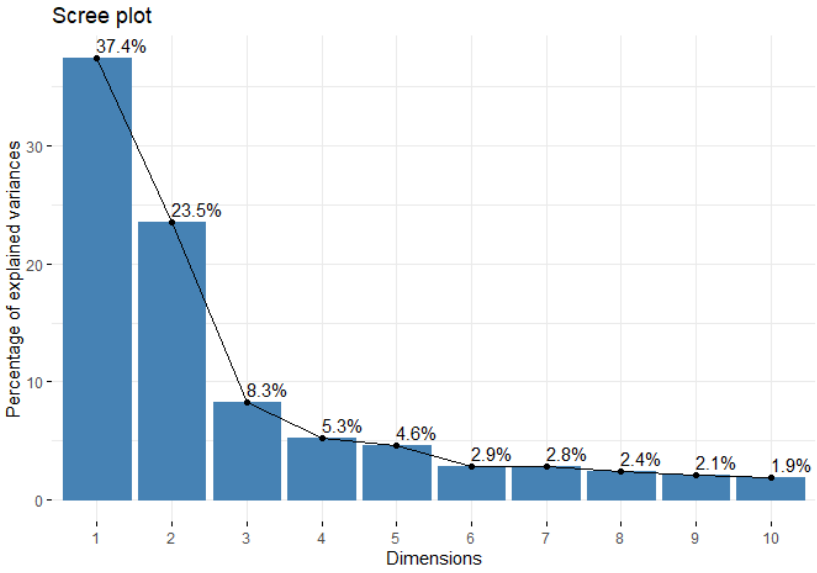


图 11：主成分分析的碎石图

同时我们也绘制了不同变量在第一主成分和第二主成分这两个维度的变量相关图用以更好地观察各变量对主成分的贡献程度和方向，其中若自变量相互靠近，则其具有正相关性，反之则为负相关；而从中心点到变量的长度则代表着变量在这个维度所占的比例。

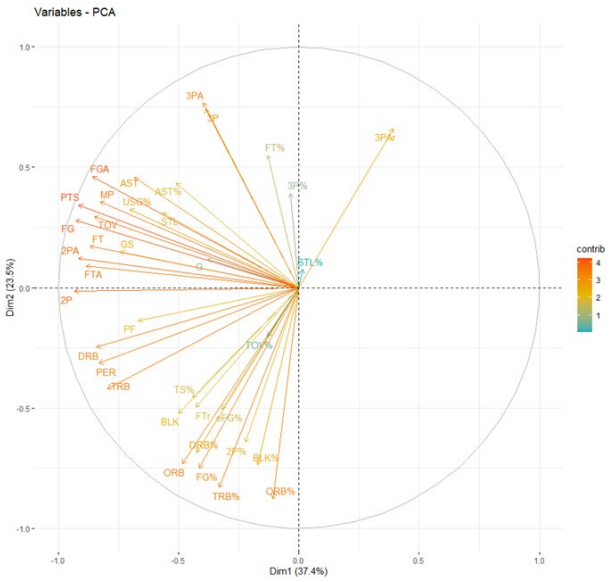


图 12：不同变量在前两个主成分的变量相关图

在确定了主成分后，我们观察变量对前五个主成分的贡献(百分比):

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
G	1.05	0.16	5.09	6.36	0.40

GS	3.98	0.26	1.13	3.39	1.42
MP	4.94	1.47	1.11	2.29	0.95
FG	6.23	0.90	0.23	0.68	0.00
FGA	5.32	2.44	0.06	0.67	0.25
FG%	1.24	6.47	2.48	0.18	8.37
3P	1.10	6.30	4.46	0.15	0.75
3PA	1.14	6.77	2.79	0.06	1.44
3P%	0.01	1.77	6.77	0.00	1.90
2P	6.28	0.00	0.17	1.35	0.14
2PA	6.08	0.17	0.63	1.68	0.02
2P%	0.36	4.74	5.44	0.89	5.92
eFG%	0.74	2.93	14.13	1.33	7.91
FT	5.45	0.34	0.35	3.60	0.35
FTA	5.67	0.09	0.56	3.16	0.20
FT%	0.12	3.47	2.37	1.49	0.08
ORB	1.70	6.14	0.02	0.31	3.32
DRB	5.12	0.70	0.00	0.15	5.85
TRB	4.60	2.03	0.00	0.23	6.01
AST	3.38	2.41	3.19	1.77	3.44
STL	2.33	1.11	1.00	20.39	0.00
BLK	1.82	3.13	0.21	1.16	6.13
TOV	5.23	0.99	2.48	0.11	0.61
PF	3.22	0.22	0.08	5.11	2.17
PTS	6.07	1.36	0.26	0.93	0.00
PER	4.99	1.14	0.06	0.77	2.13
TS%	1.42	2.38	12.66	0.28	8.98
3PAr	1.11	5.00	3.97	0.61	2.92
FTr	1.33	2.83	1.48	2.12	3.82
ORB%	0.08	8.82	0.41	0.17	0.58
DRB%	1.32	5.39	0.46	0.53	4.39
TRB%	0.80	7.92	0.49	0.43	2.87
AST%	1.88	2.16	7.90	0.75	7.06
STL%	0.00	0.07	6.11	22.62	0.29
BLK%	0.21	6.22	0.01	0.20	2.07
TOV%	0.12	0.47	10.31	5.21	6.86
USG%	3.57	1.22	1.16	8.89	0.37

表 2：变量对前五个主成分的贡献(百分比)

我们对表中对于主成分贡献度较大的自变量进行了标黄处理，可以看到五个主成分基本上涵盖了所有自变量。五个主成分对于自变量的解释维度也有所不同，PC1 主要用于解释和个人基础进攻数据相关的变量，如出场时间，得

分，以及两分和罚球的命中数等；PC2 用于解释有关三分的基础数据以及有关防守的高阶数据，如三分命中数，篮板率和盖帽率等；PC3 用于解释有关命中率的基础数据以及传球相关的高阶数据，如真实命中率，助攻率和失误率等；PC4 用于解释出勤率和与防守端侵略性相关的数据，如出场数和抢断，抢断率等；PC5 是对盖帽篮板等基础防守数据的解释。总体来说，五个主成分很好的解释了几乎所有的自变量。

4.1.2 建立模型

在通过主成分分析的方法得到 PC1-PC5 五个主成分后，我们使用这五个主成分，经过处理的年龄，球员位置以及俱乐部总体水平等自变量与经过对数处理后的球员工资建立回归模型。

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	15.583632	0.480762	32.414	< 0.0000000000000002	***
age	-0.012782	0.001772	-7.214	0.00000000000393	***
选秀轮次2	-0.476950	0.116210	-4.104	0.00005150350520	***
选秀轮次Undrafted	-0.879614	0.138571	-6.348	0.00000000074462	***
是否首发是	0.398748	0.194000	-2.055	0.0406	*
PosPF	-0.117788	0.170619	-0.690	0.4905	
PosPG	-0.218521	0.251080	-0.870	0.3848	
PosSF	-0.047421	0.195004	-0.243	0.8080	
PosSG	-0.088063	0.220197	-0.400	0.6895	
俱乐部位置西部	-0.085077	0.092525	-0.920	0.3585	
俱乐部档次东西部亚军	0.183863	0.328406	0.560	0.5760	
俱乐部档次二轮	0.004788	0.305662	0.016	0.9875	
俱乐部档次首轮	0.255493	0.294205	0.868	0.3858	
俱乐部档次未进季后赛	-0.003057	0.289636	-0.011	0.9916	
俱乐部档次总冠军	0.236813	0.364427	0.650	0.5163	
PC1	-0.017492	0.008684	-2.014	0.0448	*
PC2	0.051529	0.021513	2.395	0.0172	*
PC3	0.059496	0.025420	2.341	0.0199	*
PC4	-0.008549	0.010792	-0.792	0.4289	
PC5	-0.059434	0.032412	-1.834	0.0676	.

表 3：对数处理因变量后的回归模型

其中 * 代表 p<0.1,** 代表 p<0.05,*** 代表 p<0.01，回归模型的R^2为 0.5236

我们接着求出方差膨胀因子（表 4），一般来说若$GVIF^{(1/(2*Df)}$<10,则复共线性不明显，而变量的方差膨胀因子均小于 5，说明模型的共线性较低。

	GVIF	Df	GVIF^(1/(2*Df))
age	1.19	1	1.09
选秀轮次	1.36	2	1.08
是否首发	4.43	1	2.11
Pos	5.29	4	1.23
俱乐部位置	1.10	1	1.05
俱乐部档次	1.34	5	1.03
PC1	9.54	1	3.09
PC2	14.41	1	3.80
PC3	15.58	1	3.95
PC4	3.81	1	1.95
PC5	19.83	1	4.45

表 4：各自变量的方差膨胀因子

同时我们使用 QQ 图对模型进行正态检验，可以看到模型基本满足正态性假设

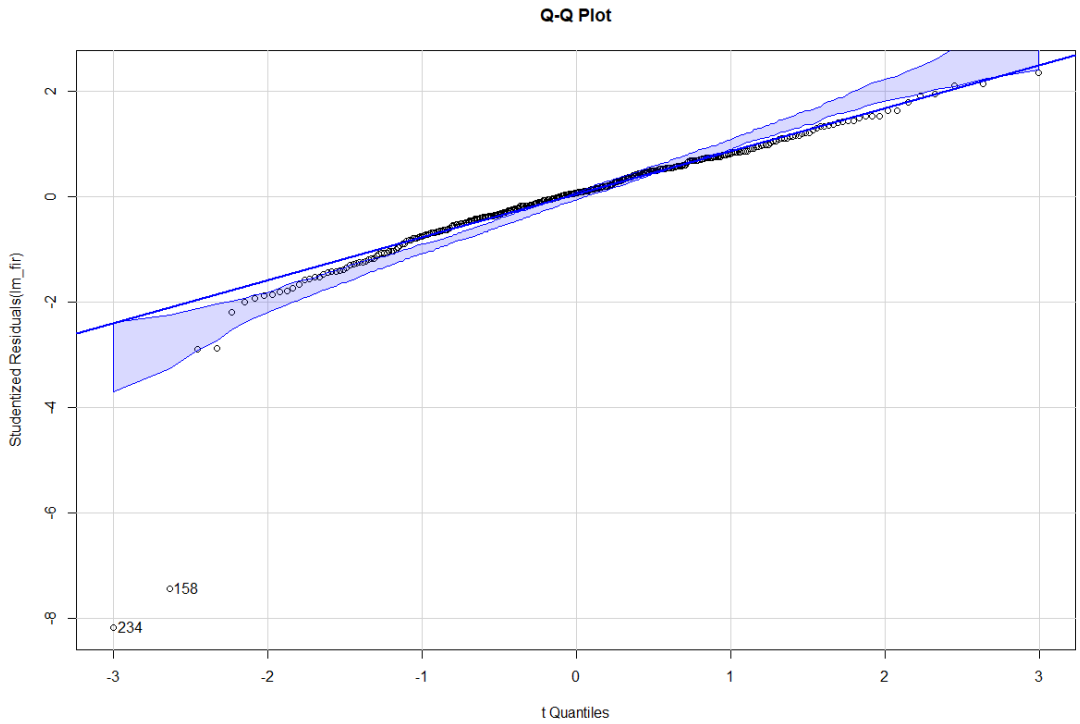


图 13：对数处理后工资残差 QQ 图

4.1.3 回归诊断

残差分析 我们通过残差图对模型的正态性、等方差和误差独立假设进行

研究。结果如下：

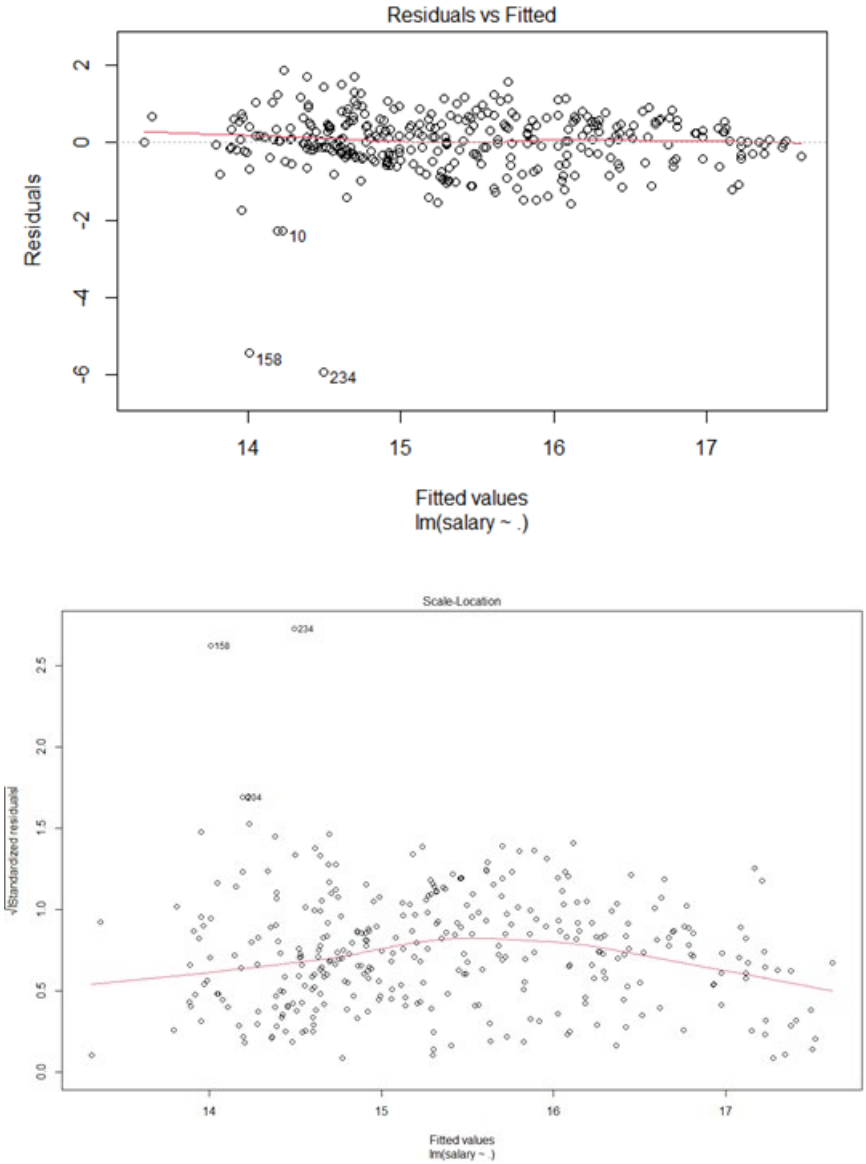


图 14：残差图

我们可以发现，在残差图中，去除个别可能的异常点，数据大致落在宽度为 4 的水平带 $|r_i| \leq 2$ 的区域内，且基本不呈任何趋势，因此我们认为对模型的正态性、等方差和误差独立假设基本成立。

在回归模型残差满足正态性、等方差和误差独立假设的条件下，为保证回归模型的正确性，我们进一步考虑所用数据中单个数据对回归模型的影响。

我们将单个数据对模型的影响分为强影响点、异常点、高杠杆点三类，通过

作出相关图像分析这些点的影响。

强影响点 强影响点是对于模型回归系数估计影响较大的点，我们使用 Cook 距离来判定单一点是否为强影响点，计算公式为:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})' X' X((\hat{\beta} - \hat{\beta}_{(i)}))}{p\hat{\sigma}^2}$$

通常认为 Cook 距离大于等于 0.5 被认为是强影响点。对各点的 Cook 距离绘图:

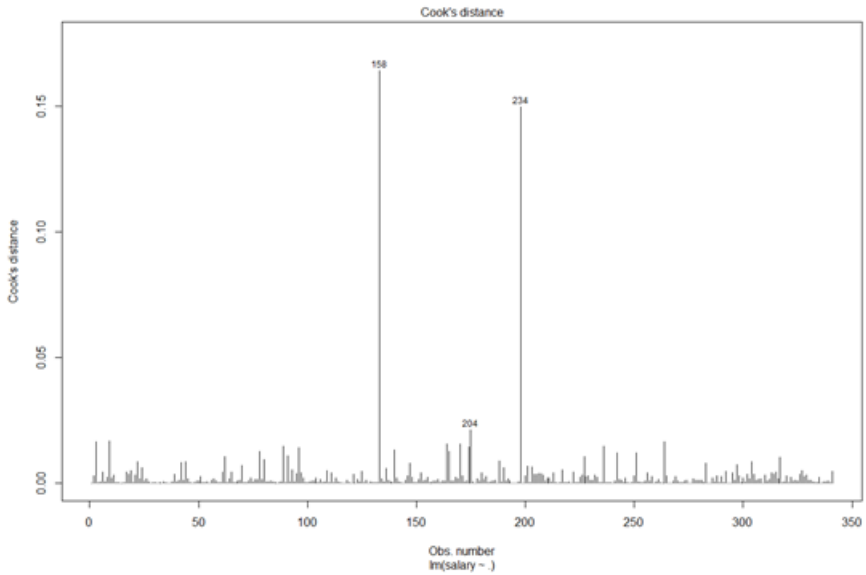


图 15：Cook 距离判定强影响点图

从中可以看出除去 158 和 234 这两个点数据之外，其他点的 Cook 距离小于于 0.5。

异常点 一组数据若其残差较其他组数据的残差大得多，则称其为异常点。一般体现在在模型诊断的 QQ 图上落入置信区间外部。异常点检验通常有两种方法:

（1）通过计算该数据对应的学生化残差$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}}$，当残差绝对值大于 3

时，该观测值即判定为异常值。

（2）使用均值漂移线性模型，对于给定的α（0< α <1），若$F_j = \frac{(n-p-1)r_j^2}{n-p-r_j^2} > F_{1,n-p-1}(\alpha)$，则判定该组数据为异常点。在 R 语言中使用 outlierTest() 语句可以查看回归数据中存在的异常值（outliertest 基于的是学生化残差判断是否为异常值），结果如下：

```
    rstudent unadjusted p-value Bonferroni p
234 -8.173942        7.0449e-15   2.4023e-12
158 -7.442990        9.1690e-13   3.1266e-10
```

图 16：模型中的异常点

即第 234、158 号数据为异常点。

高杠杆点 高杠杆点：指自变量因子空间中的离群点，由许多异常的自变量值组合起来的，通常与相应变量没有关系，判断是否为高杠杆点则通过计算点的帽子统计量，帽子矩阵计算公式 H =$X(X^TX)^{-1}X^T$，而帽子统计量是矩阵 H 的对角线上的元素，若帽子统计量大于均值 2 到 3 倍，则认为是高杠杆点。对高杠杆点进行绘图如下：

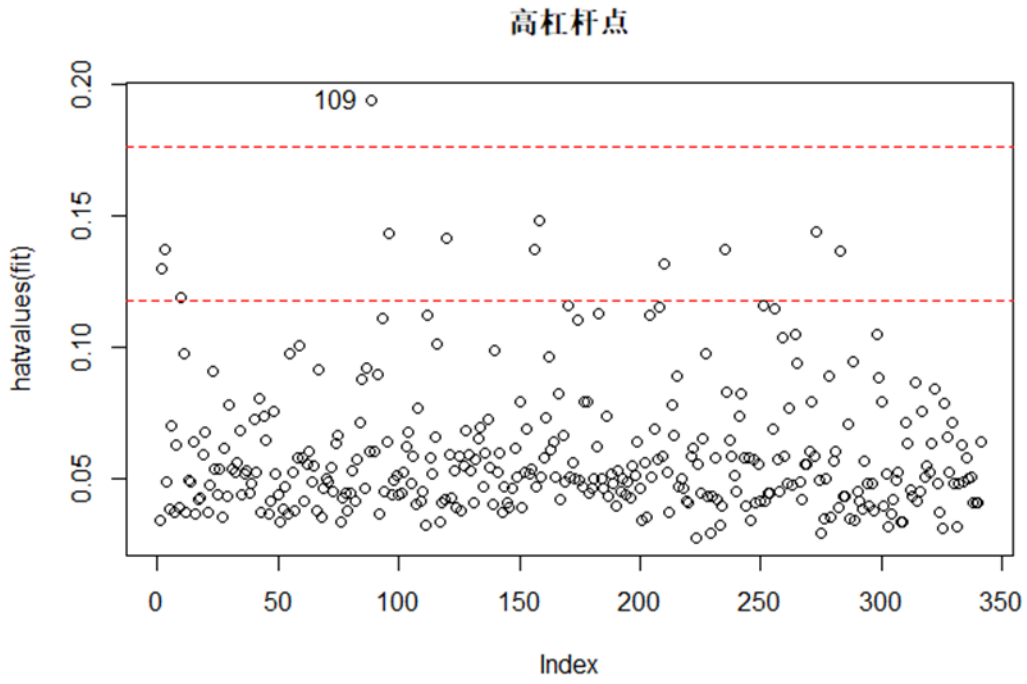


图 17：模型中的高杠杆点

在两条红线之上的被认为是高杠杆点，因此 109 号数据为高杠杆点。

利用 influenceplot()函数得到之前讨论的三个图的总图：

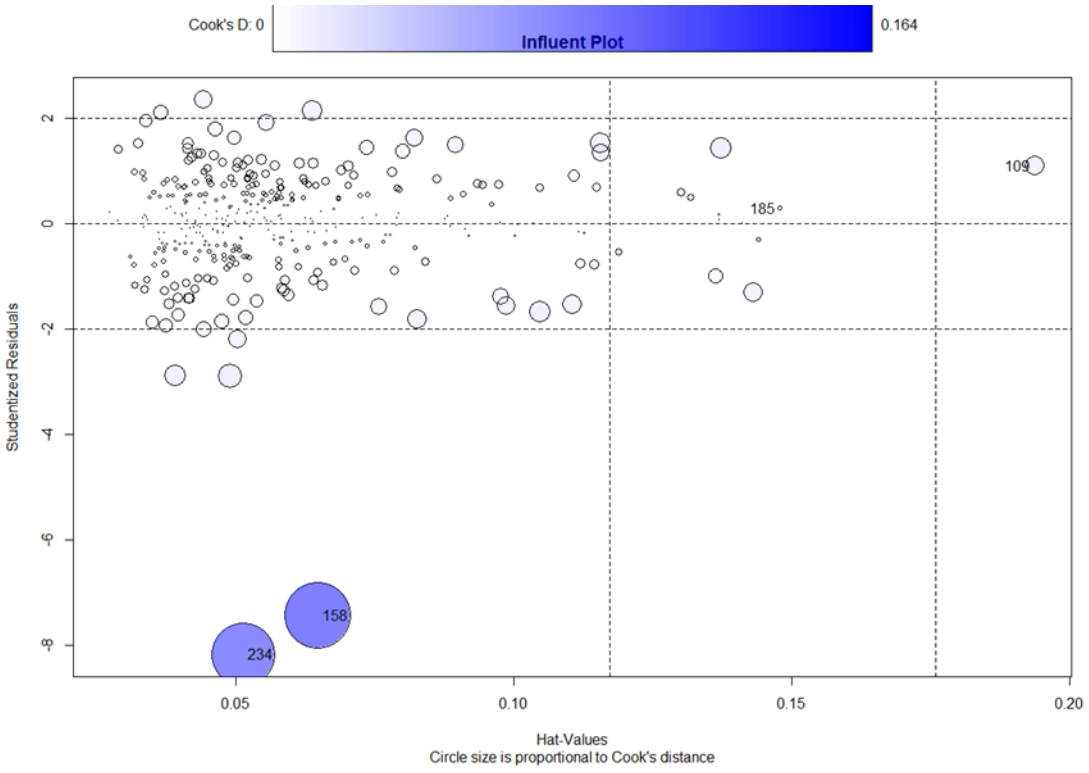


图 18：异常影响点总图

图中每个数据点周围的圆圈的颜色和大小代表库克距离（圆圈过大的点为强影响点），纵坐标为标准化后的残差，反映是否为异常值（横向两条虚线外的点为异常值），纵坐标是帽子统计量，反映是否为高杠杆点(纵向虚线外的点为高杠杆点)。

去除异常点后我们得到的回归模型如下：

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.765856	0.399987	39.416	< 2e-16 ***
age	-0.012227	0.001474	-8.295	3.08e-15 ***
选秀轮次2	-0.415214	0.096894	-4.285	2.42e-05 ***
选秀轮次Undrafted	-0.779423	0.115695	-6.737	7.56e-11 ***
是否首发是	0.369122	0.161551	-2.285	0.022977 *
PosPF	-0.082045	0.141764	-0.579	0.563169
PosPG	-0.290951	0.208640	-1.395	0.164133
PosSF	-0.031646	0.162186	-0.195	0.845422
PosSG	-0.166412	0.183002	-0.909	0.363854
俱乐部位置西部	-0.093882	0.076986	-1.219	0.223568
俱乐部档次东西部亚军	0.137125	0.272690	0.503	0.615409
俱乐部档次二轮	0.107925	0.253971	0.425	0.671158
俱乐部档次首轮	0.291486	0.244432	1.193	0.233951
俱乐部档次未进季后赛	-0.047159	0.240458	-0.196	0.844640
俱乐部档次总冠军	0.214001	0.302591	0.707	0.479940
PC1	-0.010476	0.007232	-1.449	0.148456
PC2	0.066455	0.017920	3.708	0.000246 ***
PC3	0.075513	0.021205	3.561	0.000426 ***
PC4	-0.006257	0.008963	-0.698	0.485657
PC5	-0.079384	0.027077	-2.932	0.003613 **

表 5：对数处理因变量后去除异常点的回归模型

此时模型的R^2增加到 0.634，我们再观察模型的 QQ 图，发现消去异常点后模型更加接近正态分布。

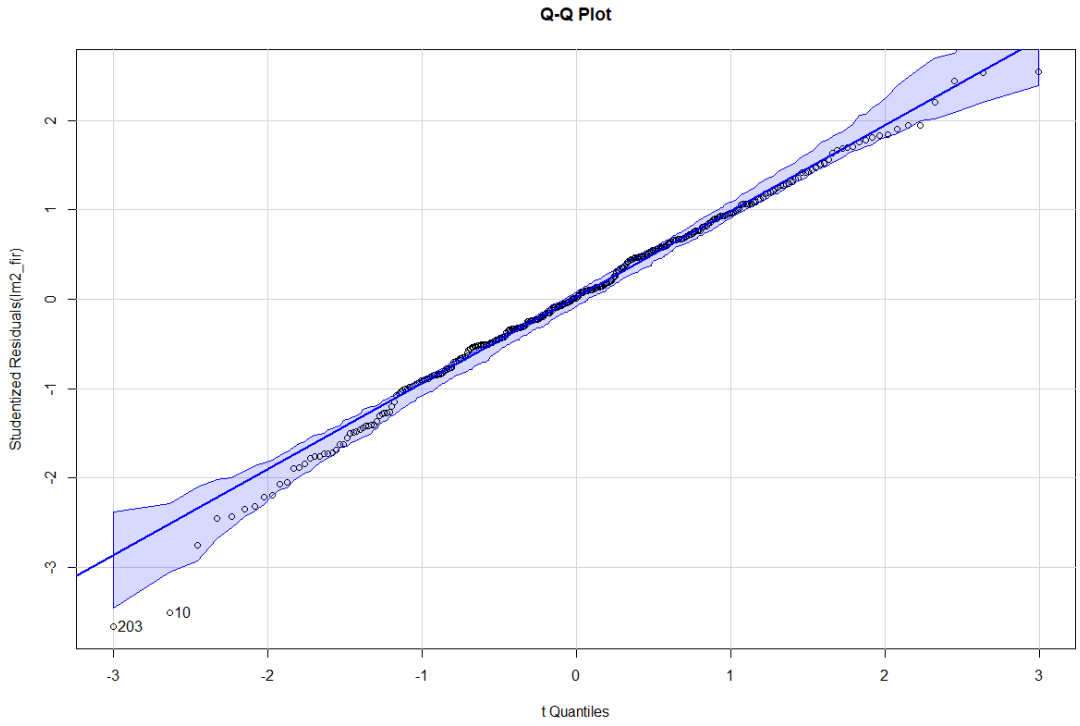


图 19：对数处理并消除异常点后工资残差 QQ 图

4.1.2Box-cox 变换

最后为了得到准确性和描述性更好地模型，我们对未经过对数处理的球员工资进行 Box-Cox 变换，从而使因变量的残差更大程度上服从正态性假设。

Box-Cox 变换：

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

Box-Cox 变换图如下：

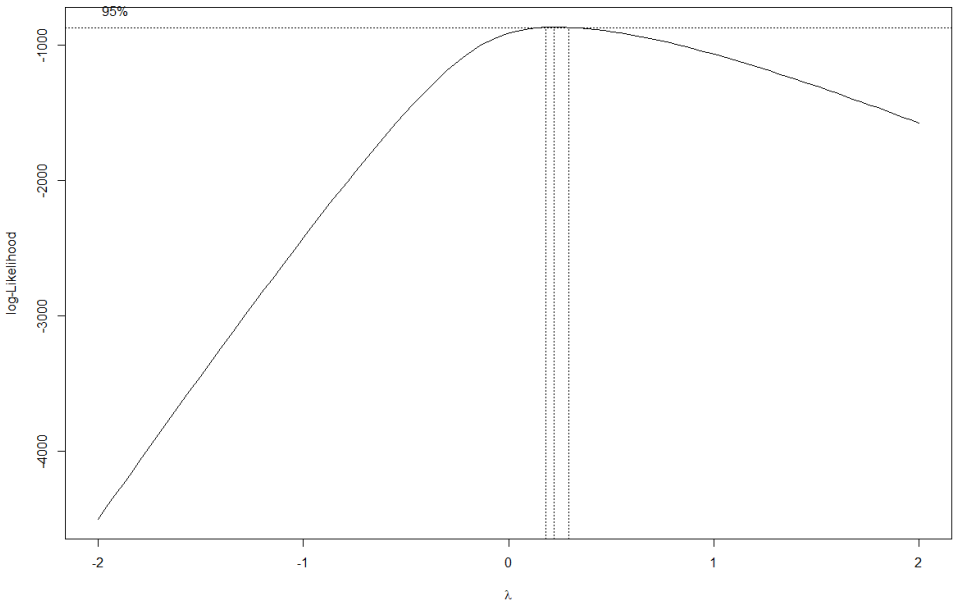


图 20：Box-Cox 变换图

得到的最优λ=0.22222，将其带入 Box-Cox 变换公式，得到的回归模型如下:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	34.20373	2.72254	12.563	< 2e-16	***
age	-0.08864	0.01003	-8.835	< 2e-16	***
选秀轮次2	-2.52369	0.65951	-3.827	0.000156	***
选秀轮次Undrafted	-4.78006	0.78749	-6.070	3.63e-09	***
是否首发是	2.93026	1.09961	-2.665	0.008095	**
PosPF	-0.39859	0.96492	-0.413	0.679823	
PosPG	-2.00937	1.42012	-1.415	0.158065	
PosSF	-0.12306	1.10393	-0.111	0.911313	
PosSG	-0.85130	1.24561	-0.683	0.494828	
俱乐部位置西部	-0.57503	0.52401	-1.097	0.273307	
俱乐部档次东西部亚军	0.31393	1.85608	0.169	0.865797	
俱乐部档次二轮	0.22731	1.72867	0.131	0.895467	
俱乐部档次首轮	1.58591	1.66375	0.953	0.341202	
俱乐部档次未进季后赛	-0.76590	1.63669	-0.468	0.640135	
俱乐部档次总冠军	1.37231	2.05961	0.666	0.505702	
PC1	-0.10616	0.04923	-2.157	0.031779	*
PC2	0.46164	0.12197	3.785	0.000184	***
PC3	0.53730	0.14434	3.723	0.000233	***
PC4	-0.08381	0.06101	-1.374	0.170515	
PC5	-0.52226	0.18430	-2.834	0.004893	**

表 6：Box-cox 变换后的回归模型

残差 QQ 图如下：

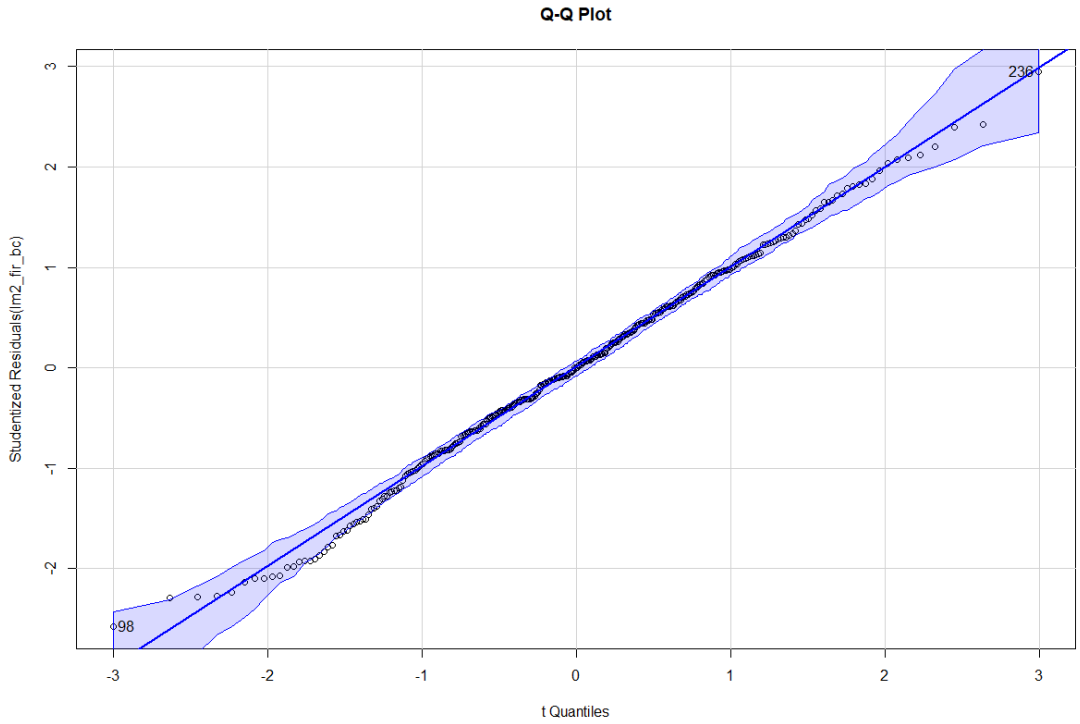


图 21：Box-Cox 变换后的残差 QQ 图

该变换后的R^2=0.664,说明经过 Box-Cox 变换后的模型解释性更强了。由于最优λ取值与 0 并不接近，而 box-cox 变换后拟合效果更好，我们在后文中使用 box-cox 变换后的球员工资作为因变量。

4.2 协方差检验

4.2.1 交互效应

在对数据完成了线性回归分析后，由于球员的各项基本信息、其所属球队的实力以及球员的球场数据三者之间都有着较强的关系，且其中有不少变量为属性变量，所以我们希望进一步考虑自变量之间的交互效应。

我们对球员工资与各个主成分的关系依照是否为首发分类作图，从拟合曲线的斜率可以看到：首发和非首发对于球员工资和主成分的关系带来了显著的差

异，因此需要在模型中考虑主成分与是否为首发的交互效应。

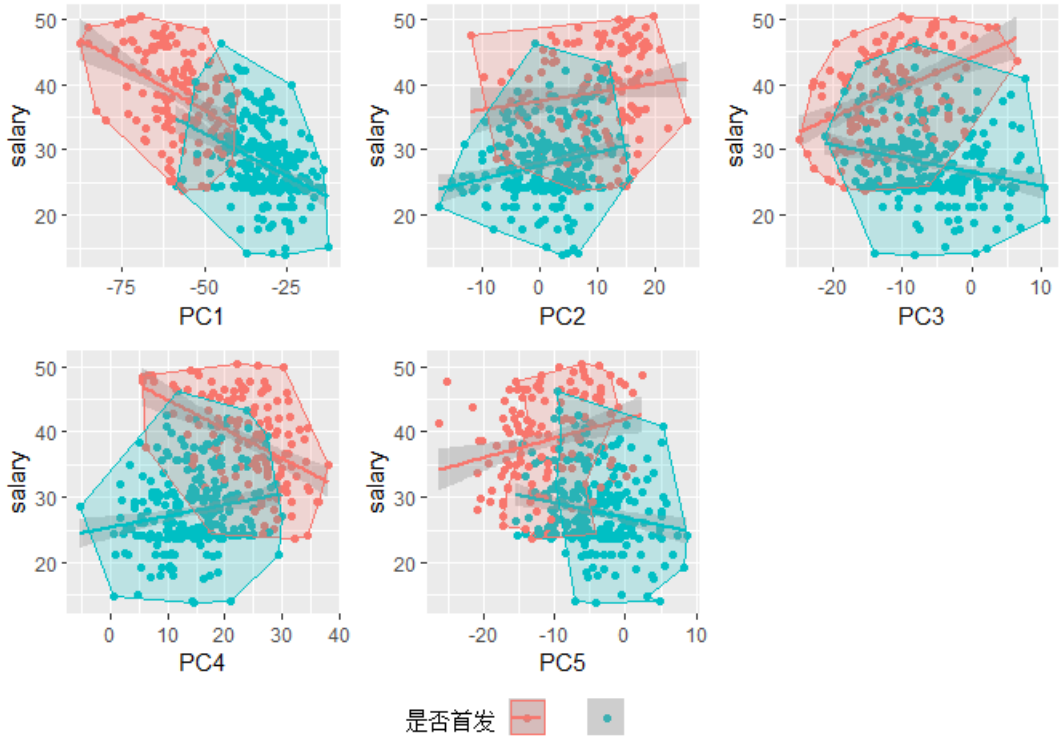


图 22：工资与主成分在不同首发情况下的关系图

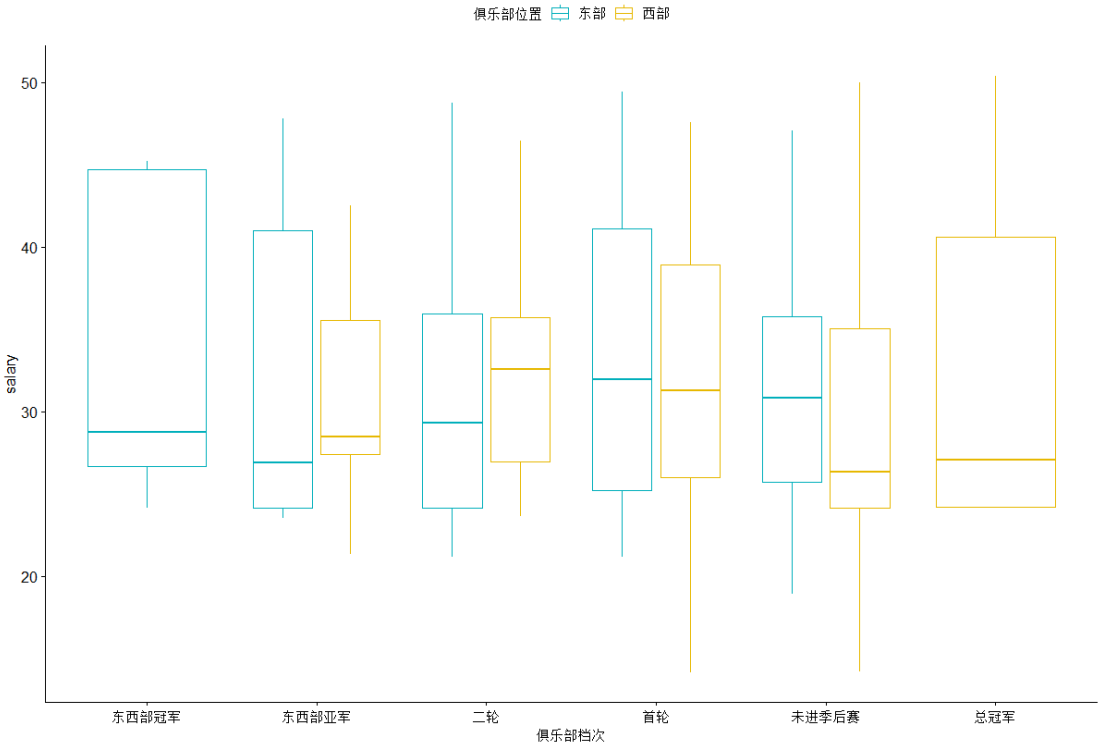


图 23：工资与俱乐部档次在不同俱乐部位置下的关系图

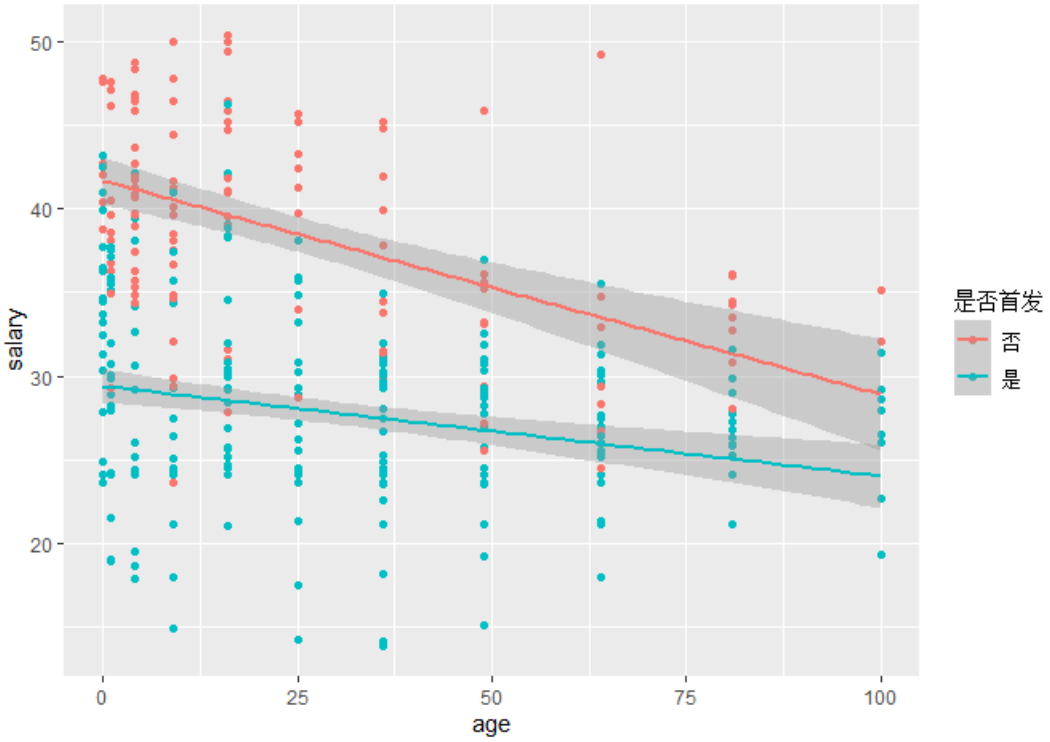


图 24：工资与年龄在不同首发情况下的关系图

在对数据进行进一步探索之后，我们发现球员年龄和是否首发、是否首发和 PC3、俱乐部位置和俱乐部档次三组自变量之间存在较为明显的显著效应，故选择考虑这三组交互效应并建立交互效应模型。最后，我们又通过 AIC 准则进行逐步回归，删去了球员位置这一变量，得到了最终的模型。

4.2.2 交互效应模型

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	37.18491	2.70734	13.735	< 2e-16	***
age	-0.13522	0.01641	-8.242	4.51e-15	***
选秀轮次2	-2.60727	0.64859	-4.020	7.28e-05	***
选秀轮次Undrafted	-4.49311	0.76807	-5.850	1.22e-08	***
是否首发（否）	-7.41116	1.69360	-4.376	1.64e-05	***
俱乐部位置西部	-1.93229	0.72946	-2.649	0.008478	**
俱乐部档次东西部亚军	0.93286	2.02044	0.462	0.644605	
俱乐部档次二轮	-0.60599	1.81604	-0.334	0.738836	
俱乐部档次首轮	0.62924	1.66146	0.379	0.705145	
俱乐部档次未进季后赛	-0.15289	1.59794	-0.096	0.923837	
俱乐部档次总冠军	2.25996	2.04997	1.102	0.271108	
PC1	-0.08400	0.04547	-1.848	0.065596	.

PC2	0.42553	0.10294	4.134	4.57e-05	***
PC3	0.65945	0.14669	4.496	9.73e-06	***
PC4	-0.09632	0.05675	-1.697	0.090632	.
PC5	-0.60308	0.17771	-3.394	0.000777	***
age:是否首发（否）	0.07299	0.02010	3.632	0.000328	***
是否首发（否）:PC3	-0.23070	0.08952	-2.577	0.010413	*
俱乐部位置西部:俱乐部档次东西部亚军	0.19010	1.96969	0.097	0.923173	
俱乐部位置西部:俱乐部档次二轮	2.67327	1.54076	1.735	0.083703	.
俱乐部位置西部:俱乐部档次首轮	2.92652	1.21781	2.403	0.016830	*
俱乐部位置西部:俱乐部档次未进季后赛	NA	NA	NA	NA	
俱乐部位置西部:俱乐部档次总冠军	NA	NA	NA	NA	

表7：考虑交互效应的回归模型

其R^2=0.684，为本案例所有模型中解释效果最高的一个模型。

其模型的诊断图如下：

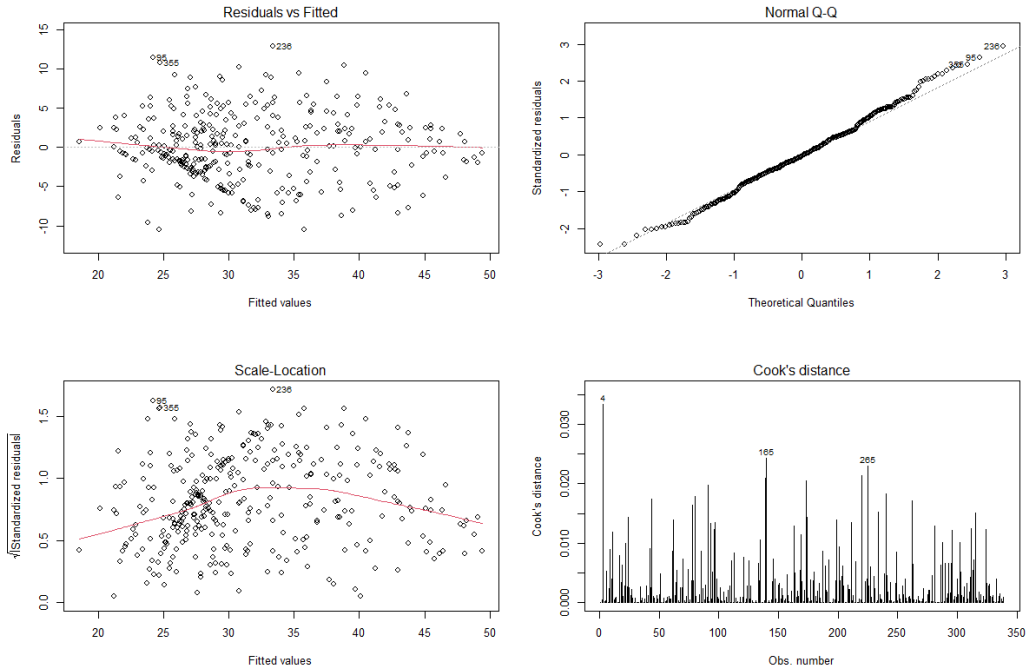


图 25：交互效应模型诊断图

5 结果分析

我们对比以对数处理后的薪水为因变量建立的线性回归模型与协方差分析

模型的表现。线性回归模型的R^2为 0.664，而协方差分析模型的R^2为 0.684，因此，考虑了自变量之间交互效应的协方差分析模型在因变量与自变量的线性关系程度方面要更优。

下面，基于模型的表现，我们对模型的结果进行分析：

年龄 经处理后的年龄系数为负，说明年龄距离巅峰年龄 29 岁越近，其身价越高，符合我们对年龄的预期。接近巅峰期年龄的球员的各方能力都比较高，也更能获得符合其身价的合同。

选秀轮次、是否首发 这两项指标代表着球队对一个球员的能力判断，能够一定程度上代表着球队对一个球员的定位和预期，是反映身价的重要指标。

球员司职 分析发现，球员在场上的位置与薪水几乎没有关系。查阅以往的数据统计发现与其并不相符。这与现代篮球模糊球员位置的特点有关。如今的球员分类已经没有那么清晰，球员们在场上的功能都比较多样，故球员司职与薪水没有太大关系。

俱乐部位置（西部） 通过分析发现，身处西部的球员更容易拿到高价合同。这可能与西部球队大球市多、篮球市场更大有关。着不仅导致西部强队多、球星多，球队老板也更愿意给出大合同。

PC1-5 作为球员球场表现数据的主成分，这些数据反应了球员的球场表现，是影响其合同价格的重要因素。

俱乐部水平 球员所在的球队水平对其合同价格影响很小。其原因是不论强队还是弱旅都有主力和替补，球队实力并不会改变球员身价和水平。

6 结论

本案例通过分析 21-22 赛季 NBA 球员薪资和各项数据之间的关系，不仅找出了与薪资关系显著的几个变量，也提出了根据这些因素预测合同价格的模型。在分析结果中，我们发现球员年龄、球场表现、俱乐部位置、以及选秀轮次、是否首发都是球员合同价格的重要影响因素。而俱乐部水平、球员司职等对球员合同的影响不大。

最后，我们基于小组建立的模型，对所有被选入数据集的球员的工资进行预测，并将经过正态化处理的球员工资还原为球员工资真实预测值后，将其判定为基于球员表现所应该获得的工资，然后挑选出 10 位工资溢价比例最高的球员（称之为毒药合同）和 10 位表现超出实际工资最多的球员（称为超值合同），溢价率 r 定义为：

$$r = \frac{(成交的价格 - 估计的价格)}{估计价格}$$

结果如下：

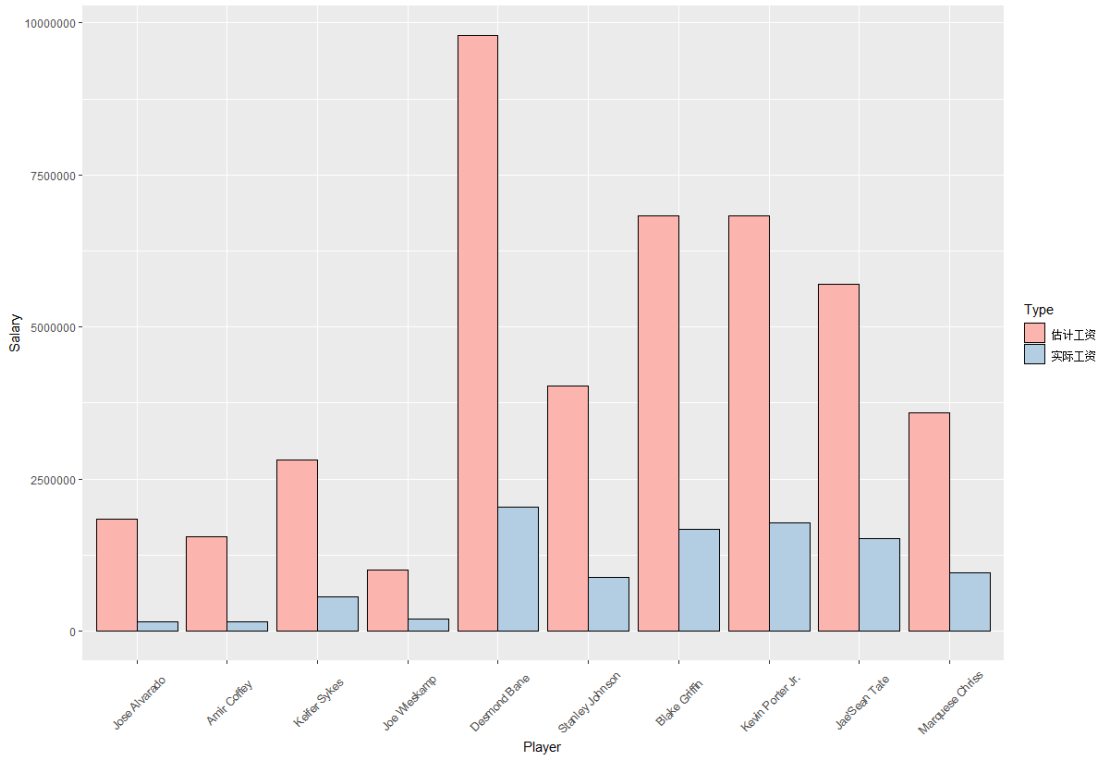


图 26：超值合同球员

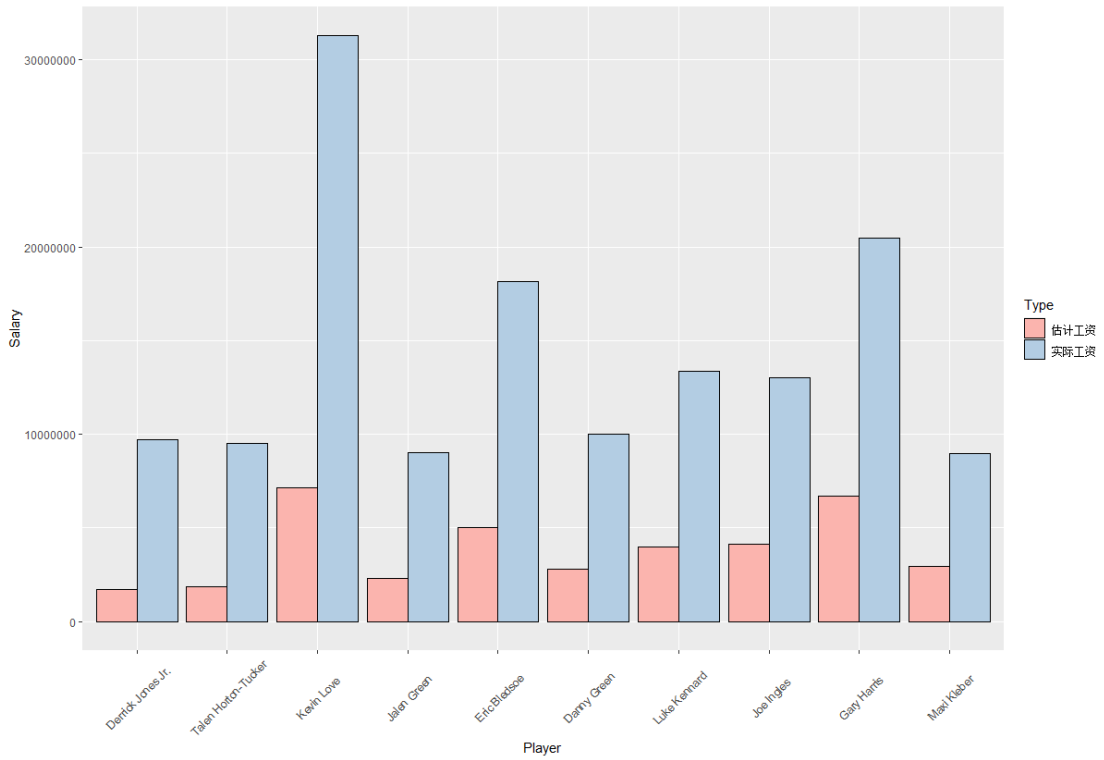


图 27：毒药合同球员

从我们得到的结果中可以看出，模型的有效性和应用性都比较强，能较好反映球员真实情况。比如超值合同球员中有绰号“肌肉射手”之称的戴斯蒙德贝恩，

上赛季季后赛球队当家球星莫兰特因伤报销后他带领球员与最后的冠军勇士队大战六场，赢得了人们的尊重。也有布雷克格里芬这种虽有能力，但是为了夺冠主动降薪的球星。而毒药合同中的球员，比如杰伦格林，上赛季他作为榜眼被火箭选中时大家都对他给予厚望，但他的表现十分不稳定且低效；更有甚者如布莱德索由于表现和工资差距过大已经离开了 NBA 去往了海外联赛效力。

我们认为球队应该避免引入毒药合同球员或有类似表现得球员，争取引入更多超值合同球员。

而通过这样的分析和预测模型，我们希望给球队提供合同价格的参考，合理管理球队的薪资空间，在交易过程中获得性价比更高的球员，以更少的成本组建更有实力的球队。