# 统计专题-球员薪资预测

陈宇阳 024034910083

## 实验设计

本实验用Python设计了三种回归方法：普通线性回归，逐步回归，LASSO回归，用于预测MLB球员的薪资，并评估其表现和特征选择机制。代码见：https://github.com/chen-yy20/MLB_salary_prediction。

## 实验流程

1. 加载数据，以80%比20%的比例随机切分训练集和测试集。
2. 模型训练和评估：采用普通线性回归、逐步回归、LASSO回归训练模型。
3. 结果可视化与分析：比较模型表现，分析特征重要性

## 实验原理

### 普通线性回归

即最小二乘法，最小化预测值和实际值之间的均方误差，使用所有的可用特征，其数学表达式为：

$$\min_{\beta} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

### 逐步线性回归

在普通线性回归基础上，增加了特征选择，只使用特征子集$j \in S$其中 $S \subset 1, 2, ..., p$，其数学表达式：

$$\min_{\beta, S \subset \{1,2,...,p\}} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j \in S} \beta_j x_{ij} \right)^2 + \lambda |S| \right\}$$

增加了$\lambda|S|$作为对模型复杂度的惩罚，限制特征子集的扩展。

### LASSO线性回归

在普通线性回归基础上，增加了L1正则化项，是对系数绝对值之和的惩罚，实现特征选择。

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right\}$$

## 实验结果与分析

## 数据

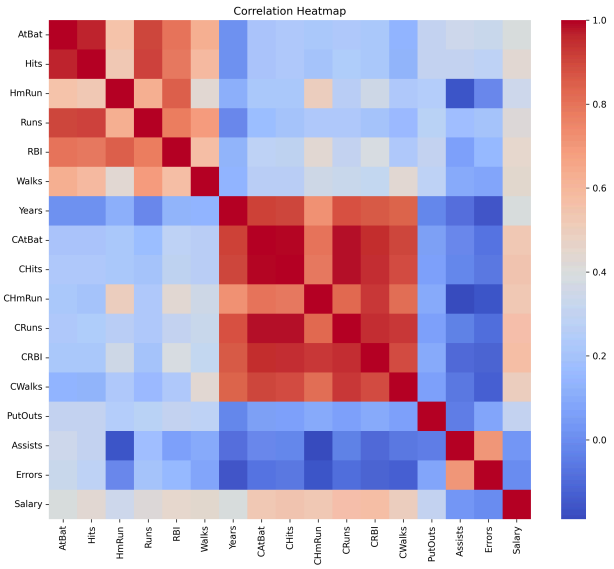
1986年MLP球员的表现与年薪数据，为(263, 17)的数组，对应263名球员和17项指标。

## 相关性分析

我们首先计算了各种特征直接的相关性，用于理解特征之间的关系，预见可能的多重共线性问题。

使用皮尔逊相关系数进行计算：

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

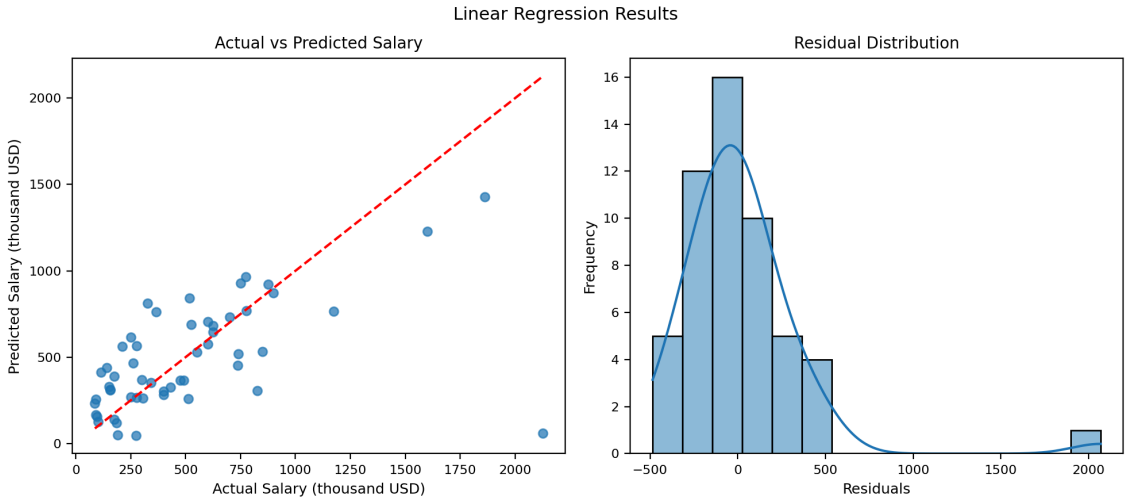



观察如下：开头区域，打击相关统计如 AtBat（打击数）、Hits（安打数）、HmRun（本垒打）、Runs（得分）和RBI（打点）之间存在高度正相关，这是很正常的，虽然我不懂mlb，但我知道nba，投篮多的球员，实力就强，得分就多。中间区域，为生涯统计数据，相互高度相关也是正常的。对于我们要预测的Salary，观察可知，主要与生涯击打数据相关，与其他能力（Assists，Errors）相关性不大，没人喜欢角色球员。

## 评估结果分析

我们在20%的测试集上测试回归效果。

**普通线性回归**




实际和预测值基本一致，误差也基本符合正态分布，但是右下角有一个离群值。

具体结果如下：

```
Regular Linear Regression Model Evaluation:
Mean Squared Error (MSE): 129173.25
Root Mean Squared Error (RMSE): 359.41
Coefficient of Determination (R²): 0.2858

Feature Importance (sorted by absolute coefficient value):
    Feature  Coefficient  Absolute Coefficient
1      Hits     7.836142              7.836142
5     Walks     5.523848              5.523848
6     Years     5.520803              5.520803
3      Runs    -2.774445              2.774445
2     HmRun     2.409842              2.409842
10    CRuns     1.898572              1.898572
0     AtBat    -1.723585              1.723585
11     CRBI     1.126854              1.126854
9    CHmRun    -0.979544              0.979544
15   Errors    -0.894909              0.894909
12   CWalks    -0.787238              0.787238
4       RBI    -0.324100              0.324100
13  PutOuts     0.260394              0.260394
14  Assists     0.224796              0.224796
7    CAtBat    -0.206793              0.206793
8     CHits    -0.110126              0.110126
```

**逐步线性回归**

使用AIC来逐步选择回归特征，最终选择了5项特征作为自变量。

```
Feature Importance (sorted by absolute coefficient value):
       Feature  Coefficient      P-value  Absolute Coefficient
```

```
Hits       Hits     8.002847  2.671690e-05           8.002847
Walks      Walks    3.269069  1.738939e-02           3.269069
AtBat      AtBat   -1.738372  3.371014e-03           1.738372
CRBI       CRBI     0.657073  4.499085e-18           0.657073
PutOuts    PutOuts  0.246861  1.478578e-03           0.246861
```

为什么选择的指标和并非相关性热图中最高的几项？从具体的计算过程，可以得到很有趣的观察：

```
--- Running Stepwise Regression ---

Performing Forward Stepwise Regression:
Added feature: CRBI, AIC: 3078.86 #优先选择生涯打点，最高相关性，基本代表中央红区
Added feature: Hits, AIC: 3031.66 #其次选择安打数，基本代表左上红区
Added feature: PutOuts, AIC: 3022.31 #刺杀数，看似相关性不高，实则捕捉防守能力
Added feature: AtBat, AIC: 3018.57
Added feature: Walks, AIC: 3014.73 #打席数和保送数，也许可以提供额外的信息

Final model includes 5 features:
CRBI, Hits, PutOuts, AtBat, Walks #一些强相关的项被省略了
```

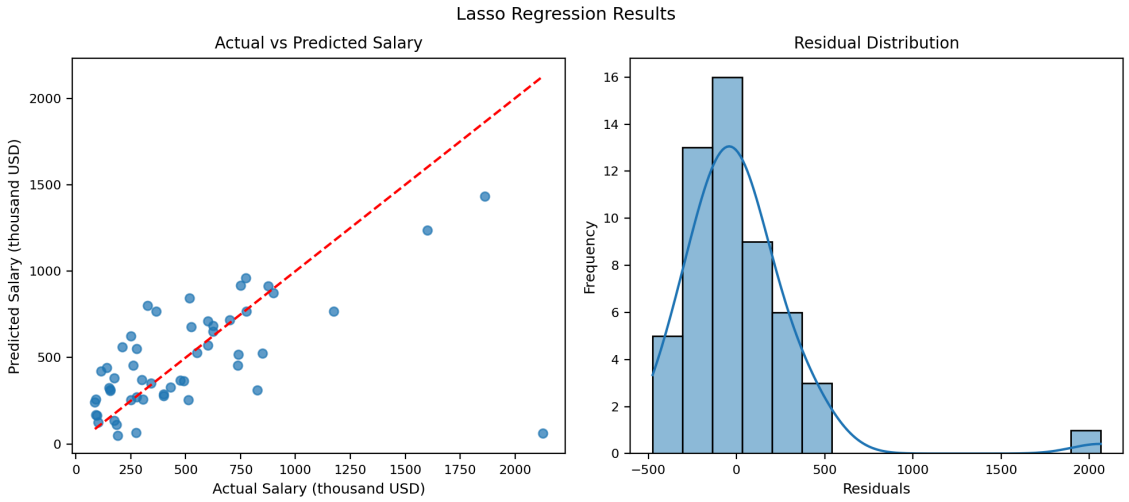

具体结果如下：

```
Stepwise Regression Model Evaluation:
Mean Squared Error (MSE): 140876.48
Root Mean Squared Error (RMSE): 375.34
Coefficient of Determination (R²): 0.2211
```

从均方差和相关性上看，效果不如普通线性回归。

**LASSO回归**

设置$\alpha = 10.0$，LASSO回归结果和残差分布如下：




数值结果如下：

LASSO Regression Model Evaluation:
Mean Squared Error (MSE): 128455.90
Root Mean Squared Error (RMSE): 358.41
Coefficient of Determination ($R^2$): 0.2898

Number of features selected by LASSO: 16 out of 16

Features selected by LASSO (non-zero coefficients):

| | Feature | Coefficient | Absolute Coefficient |
|---|---|---|---|
| 1 | Hits | 7.547361 | 7.547361 |
| 5 | Walks | 5.275797 | 5.275797 |
| 3 | Runs | -2.191166 | 2.191166 |
| 10 | CRuns | 1.760015 | 1.760015 |
| 0 | AtBat | -1.727380 | 1.727380 |
| 6 | Years | 1.637896 | 1.637896 |
| 2 | HmRun | 1.129706 | 1.129706 |
| 11 | CRBI | 1.033718 | 1.033718 |
| 12 | CWalks | -0.743788 | 0.743788 |
| 9 | CHmRun | -0.740882 | 0.740882 |
| 15 | Errors | -0.655814 | 0.655814 |
| 13 | PutOuts | 0.259230 | 0.259230 |
| 14 | Assists | 0.205908 | 0.205908 |
| 7 | CAtBat | -0.196879 | 0.196879 |
| 8 | CHits | -0.042926 | 0.042926 |
| 4 | RBI | 0.034289 | 0.034289 |

添加了正则化项以后，结果稍微比普通线性回归好一点点，离群值仍然存在。

## 总结

通过三种回归方法对MLB球员薪资的预测分析，我们获得了一些有价值的发现和思考：

## 模型表现比较

1. **LASSO回归**表现最佳（$R^2 = 0.2898$），其次是普通线性回归（$R^2 = 0.2858$），逐步回归表现最差（$R^2 = 0.2211$）。这说明在本案例中，使用带惩罚项的全特征模型比仅选择部分特征的模型效果更好。
2. 所有模型的$R^2$值均在0.22-0.29之间，这表明我们的模型只能解释约22%-29%的薪资状况。也就是说，球员薪资可能受到许多数据集中未包含的因素影响，如商业价值、球队战略需求、市场条件等。

## 特征选择机制

1. **逐步回归**选择了5个特征（CRBI、Hits、PutOuts、AtBat、Walks），这种组合从不同维度（长期表现、近期状态、防守能力、出场机会、技术细节）来预测薪资，而非简单选择相关性最高的变量。
2. **LASSO回归**保留了全部16个特征，但通过系数调整实现了"软选择"，对不重要特征赋予较小权重，而非完全排除。
3. 相关性分析与特征选择结果存在差异，说明单变量相关性并不是特征选择的唯一标准，变量间的交互效应和多重共线性也需要考虑。

## 数据与模型局限性

1. 所有模型的残差图都显示存在明显的离群值，这极大影响了模型性能的指标。出于好奇，我开盒了这位球员的具体数据，详见`outsider.py`，`20,1,0,0,0,0,2,41,9,2,6,7,4,78,220,6,2127.333`，年度仅20次打席，1次安打，无本垒打、得分或打点，41次生涯打席，9次生涯安打，却有着78次刺杀，220次助攻，模型预测只有20万美元的薪资，实际却超过了200万，猜测这来源于其独特的防守能力，进一步体现了数据的局限性。
2. 模型的拟合度不是特别高，这可能是因为：
    - 薪资决定因素复杂，不仅取决于技术统计
    - 数据集较小（263名球员）
    - 缺少一些关键影响因素如球员人气、商业价值等。