多元线性回归模型：

将标准化以后的变量投入模型 ,得到如下图结果。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |  |
| (Intercept) | 0.002275 | 0.05071 | 0.045 | 0.9646 |  |
| Team\_PER | 0.113449 | 0.064377 | 1.762 | 0.0919 | . |
| Rk | -0.109022 | 0.176656 | -0.617 | 0.5435 |  |
| A\_MoV | 2.152315 | 2.187677 | 0.984 | 0.3359 |  |
| A\_ORtg | -32.308 | 50.818751 | -0.636 | 0.5315 |  |
| A\_DRtg | 32.674673 | 51.48191 | 0.635 | 0.5322 |  |
| A\_NRtg | 52.202229 | 84.394901 | 0.619 | 0.5426 |  |
| Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 | | | | | |
| F-statistic: 62.14 on 6 and 22 DF, p-value: 1.128e-12 | | | | | |

根据F检验，得到P值小于0.0001，得到整体模型是高度显著的，说明模型中至少一个自变量对因变量有显著影响。且判决系数R^2是0.9，说明模型中自变量可以在很大程度解释因变量。说明我们选取的自变量涵盖大量因变量的信息。但由于每一个自变量做自身t检验的显著性都相当低，考虑变量之间是否存在多重共线性。

首先做多重共线性检验，结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Team PER | Rank | A\_MoV | A\_ORtg | A\_DRtg | A\_NRtg |
| 1.58 | 12.3 | 1890 | 101000 | 104000 | 2810000 |

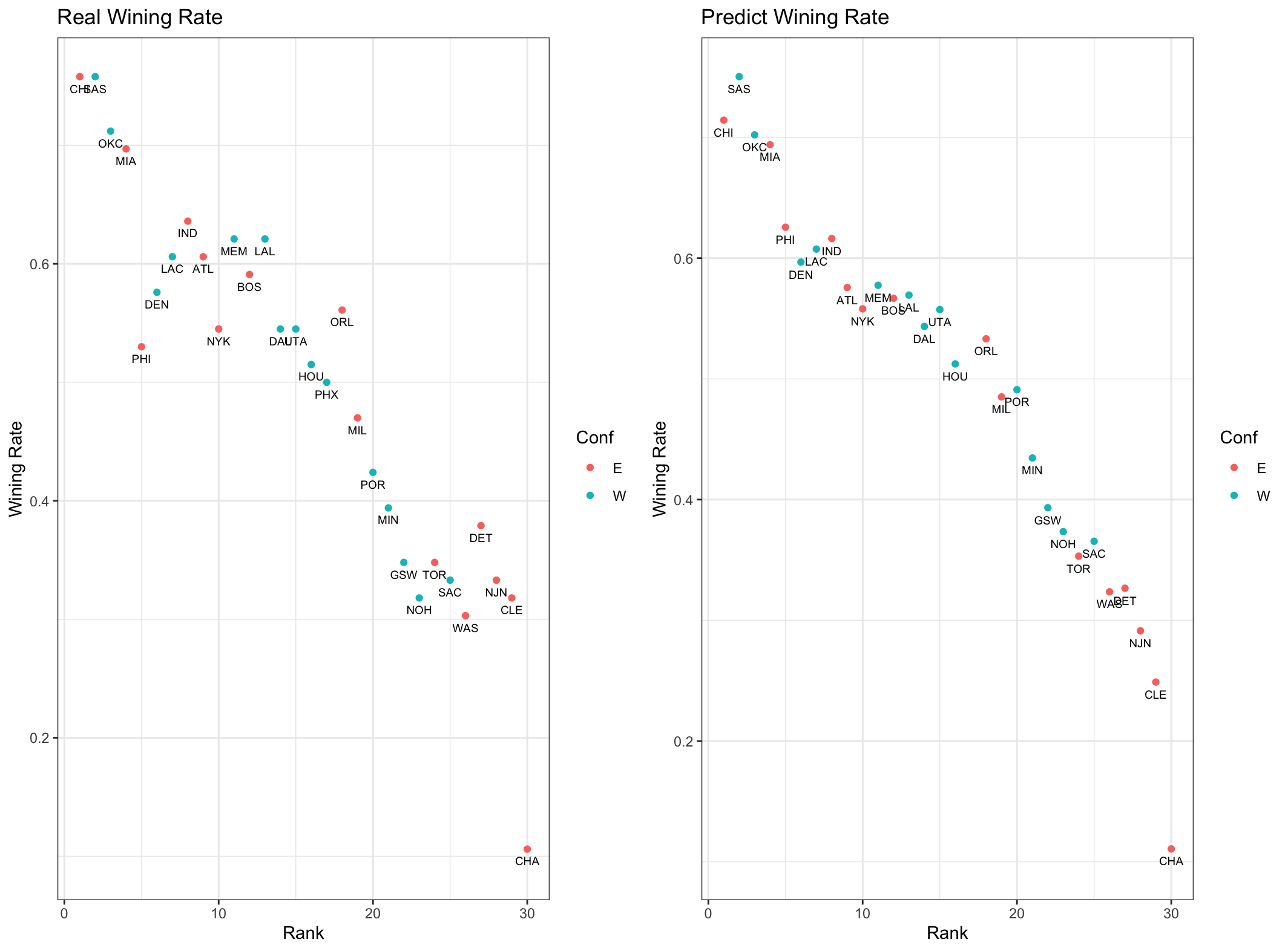
可以看出除了Team PER之外，其他变量的VIF都非常大，Rank的方差膨胀因子达到12， 其他变量都超过200，考虑使用岭回归，消除多重共线性。

岭回归：使用岭回归参数0.02时得到最佳的岭回归模型。此时所有变量都是显著的。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. | t | value | p value |  |
| (Intercept) | 0.0002156 | NA | NA | NA | NA |  |
| Team\_PER | 0.1235208 | 0.6536105 | 0.3087487 | 2.117 | 0.03426 | \* |
| Rk | -0.1900615 | -1.0229824 | 0.6432621 | 1.59 | 0.11177 |  |
| A\_MoV | 0.2535602 | 1.3653724 | 0.2736496 | 4.989 | 6.05E-07 | \*\*\* |
| A\_ORtg | 0.1476157 | 0.7913802 | 0.3094264 | 2.558 | 0.01054 | \* |
| A\_DRtg | -0.1904352 | -1.0232096 | 0.3137362 | 3.261 | 0.00111 | \*\* |
| A\_NRtg | 0.2050112 | 1.1038731 | 0.2137947 | 5.163 | 2.43E-07 | \*\*\* |
| Signif. codes: | 0 ‘\*\*\*’ | 0.001 ‘\*\*’ | 0.01 ‘\*’ | 0.05 ‘.’ | 0.1 ‘ ’ |  |
| Ridge parameter: 0.02080646, chosen automatically, computed using 2 PCs | | | | | |  |

最终我们得到的的岭回归模型为：

用此模型对12-13赛季的数据进行预测得到如下的结果：



|  |  |  |  |
| --- | --- | --- | --- |
| Team | Real | Team | Predict |
| SAS | 0.758 | SAS | 0.7503361 |
| CHI | 0.758 | CHI | 0.7142695 |
| OKC | 0.712 | OKC | 0.7020912 |
| MIA | 0.697 | MIA | 0.6939723 |
| IND | 0.636 | **PHI** | 0.6255864 |
| MEM | 0.621 | IND | 0.6162185 |
| LAL | 0.621 | LAC | 0.6074751 |
| LAC | 0.606 | DEN | 0.596702 |
| ATL | 0.606 | MEM | 0.5774977 |
| BOS | 0.591 | ATL | 0.5756241 |
| DEN | 0.576 | LAL | 0.5693788 |
| ORL | 0.561 | BOS | 0.5667246 |
| UTA | 0.545 | NYK | 0.5579812 |
| NYK | 0.545 | UTA | 0.5573567 |
| DAL | 0.545 | DAL | 0.5434609 |
| PHI | 0.53 | ORL | 0.5331562 |
| HOU | 0.515 | HOU | 0.5123906 |
| PHX | 0.5 | POR | 0.4910005 |
| MIL | 0.47 | MIL | 0.4850674 |
| POR | 0.424 | MIN | 0.4344806 |
| MIN | 0.394 | GSW | 0.3932617 |
| DET | 0.379 | **NOH** | 0.3734329 |
| TOR | 0.348 | SAC | 0.3654702 |
| GSW | 0.348 | TOR | 0.3531357 |
| SAC | 0.333 | DET | 0.3265933 |
| NJN | 0.333 | WAS | 0.3236267 |
| NOH | 0.318 | NJN | 0.2913074 |
| CLE | 0.318 | CLE | 0.2488394 |
| WAS | 0.303 | CHA | 0.1106624 |

结论：根据上述模型得到的预测结果，可以看到和真实的胜率是差距不大，根据岭回归预测结果，预测值是线性分布的，且与实际结果是一致的。胜率和球队名次是反比关系，名次越高，胜率越低。红色加中的队伍是预测差距超过10%的队伍。

Lasso回归：