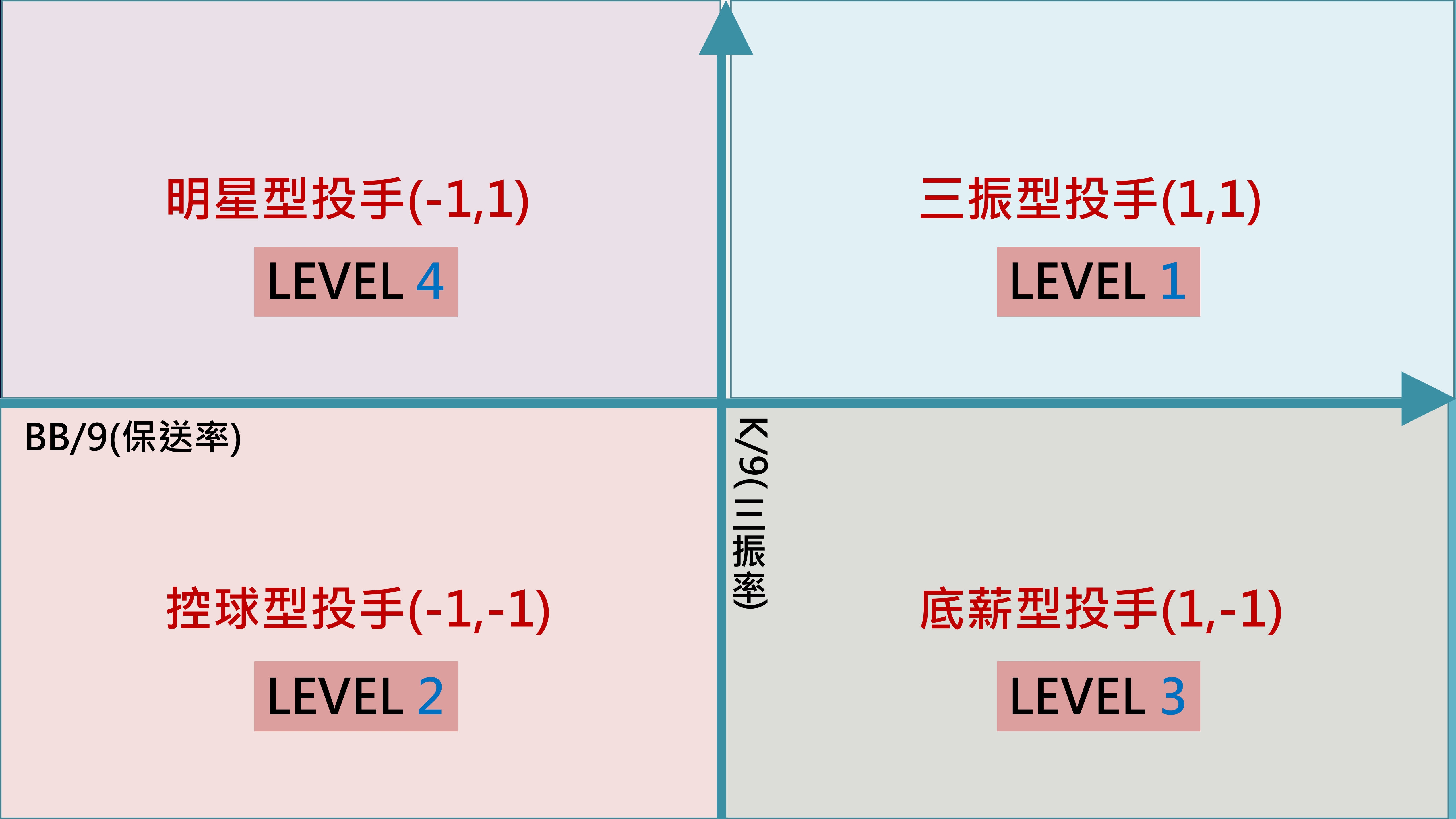
主題：棒球恢復係數對ERA的影響

資料前處理

- 使用爬蟲在中華職棒官網抓取資料，並且只選取IP>25的資料，然後刪除離群值。
- 將資料進行分群，分別將BB/9大於該年平均表示為-1、否則表示為1，K/9大於該年平均表示為1，小於該年平均表示為-1。再將其分為(1 , 1)、(1, -1)、(-1 , 1)、(-1 , -1) 四大板塊。
- 另外也將滾飛比(G/F)分成三大板塊分別為0 (G/F<0.93)、1 (0.93<G/F<1.13)、2 (G/F>1.13)。



明星型投手(-1,1)

LEVEL 4

三振型投手(1,1)

LEVEL 1

BB/9(保送率)

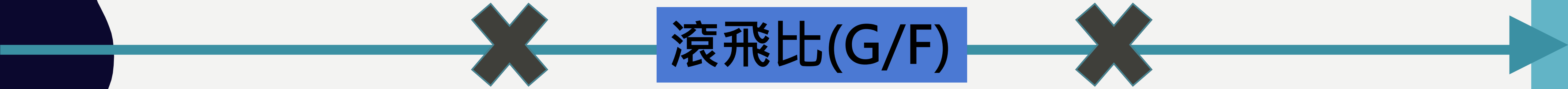
K/9(三振率)

控球型投手(-1,-1)

LEVEL 2

底薪型投手(1,-1)

LEVEL 3



飛球型(0)

中間型(1)

滾地型(2)

滾飛比(G/F)

$(G/F) < 0.93$

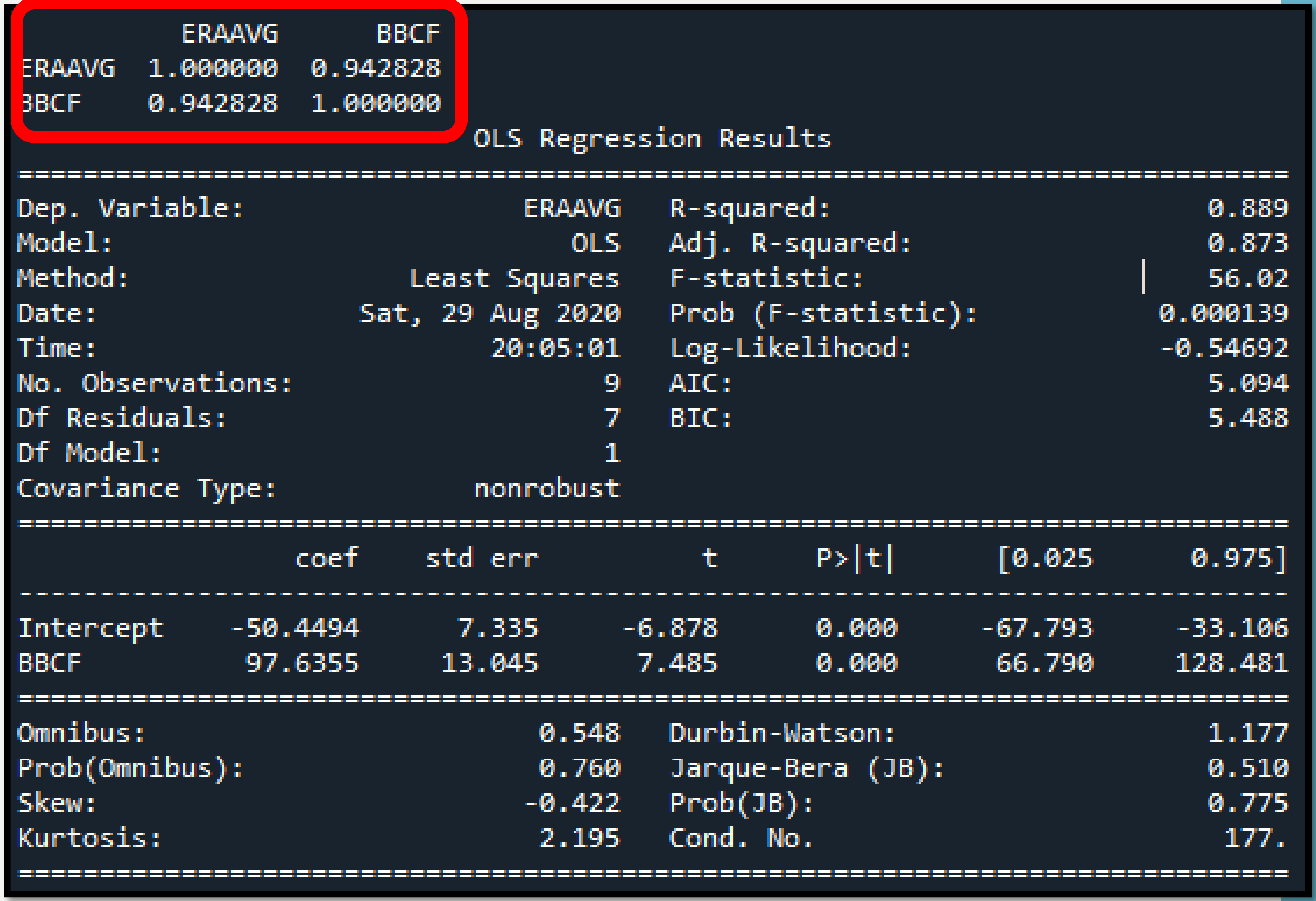
$0.93 < (G/F) < 1.13$

$(G/F) > 1.13$

資料前處理

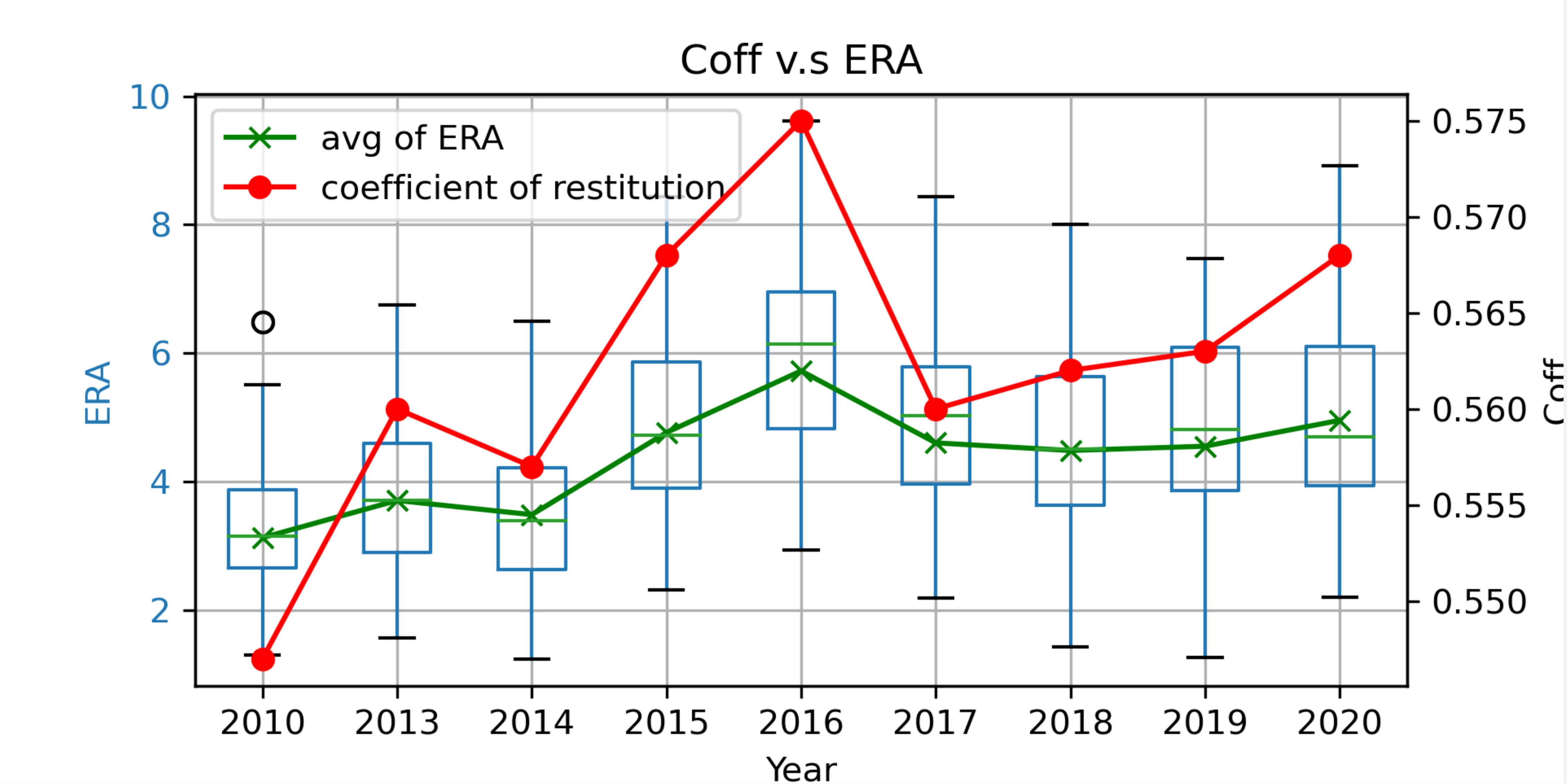
- **恢復係數與ERA**(各年平均)關聯係數為

0.9428



```
          ERAAVG      BBCF
ERAAVG  1.000000  0.942828
BBCF    0.942828  1.000000
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ERAAVG   R-squared:                       0.889
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     56.02
Date:                Sat, 29 Aug 2020   Prob (F-statistic):           0.000139
Time:                        20:05:01   Log-Likelihood:               -0.54692
No. Observations:                   9   AIC:                             5.094
Df Residuals:                       7   BIC:                             5.488
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -50.4494      7.335     -6.878      0.000     -67.793     -33.106
BBCF          97.6355     13.045      7.485      0.000      66.790     128.481
==============================================================================
Omnibus:                        0.548   Durbin-Watson:                   1.177
Prob(Omnibus):                  0.760   Jarque-Bera (JB):                0.510
Skew:                          -0.422   Prob(JB):                        0.775
Kurtosis:                       2.195   Cond. No.                         177.
==============================================================================
```

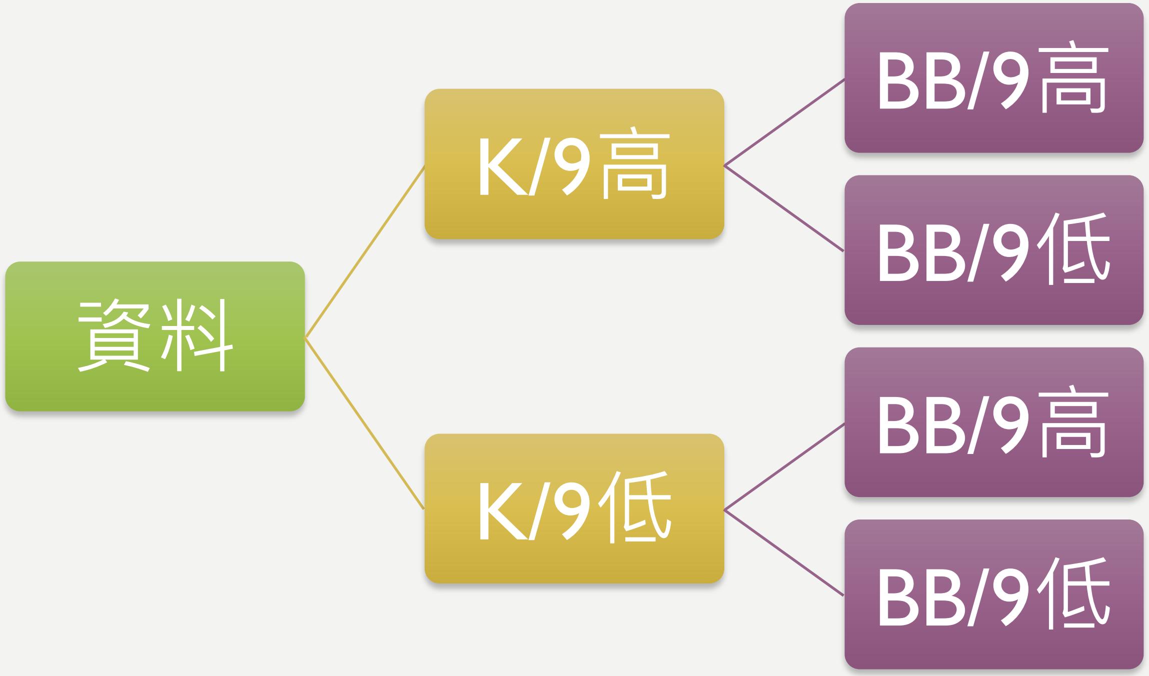
資料前處理



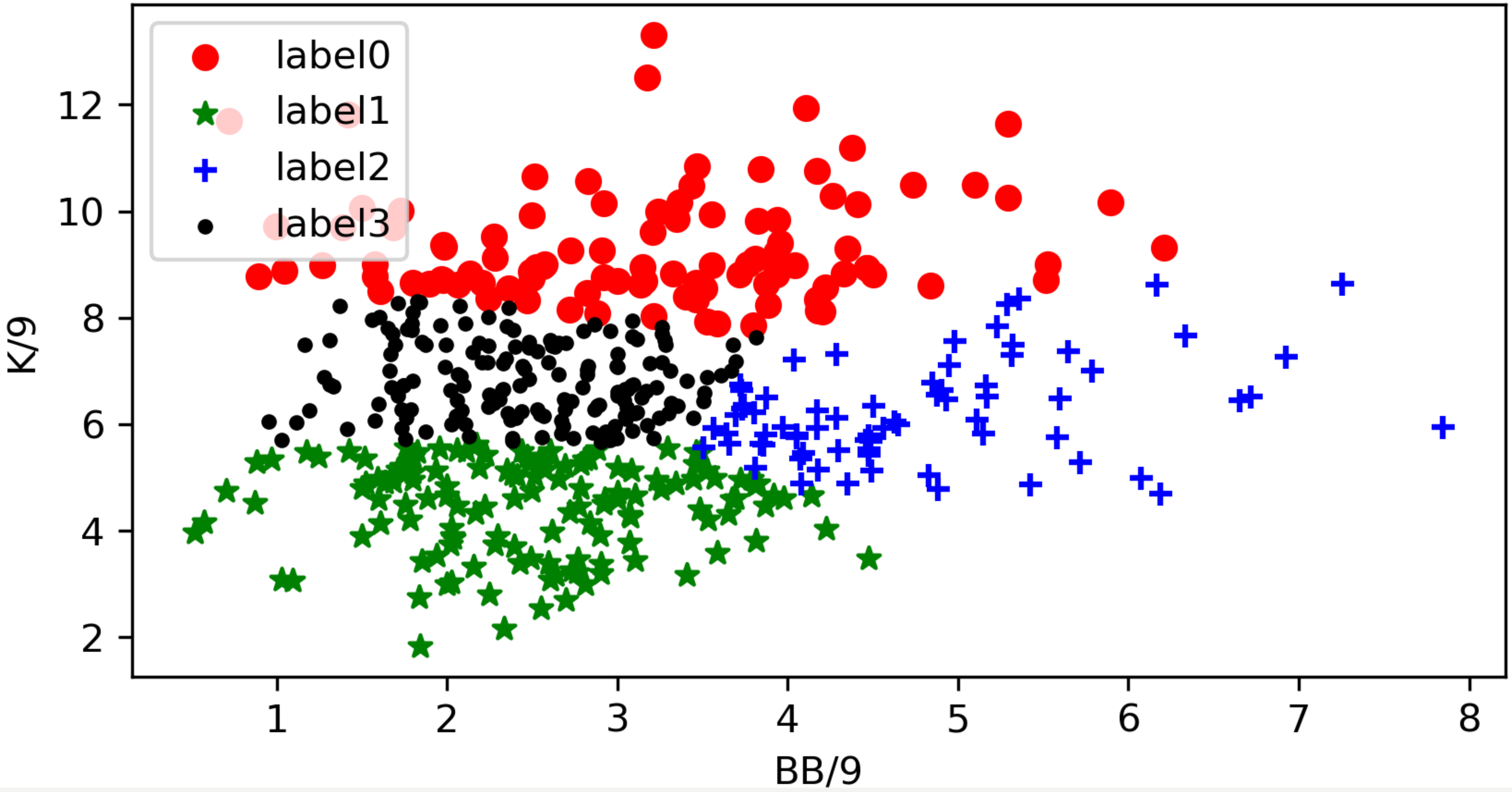
Coff v.s ERA
avg of ERA
coefficient of restitution
ERA
10
8
6
4
2
Coff
0.575
0.570
0.565
0.560
0.555
0.550
2010
2013
2014
2015
2016
2017
2018
2019
2020
Year

資料前處理

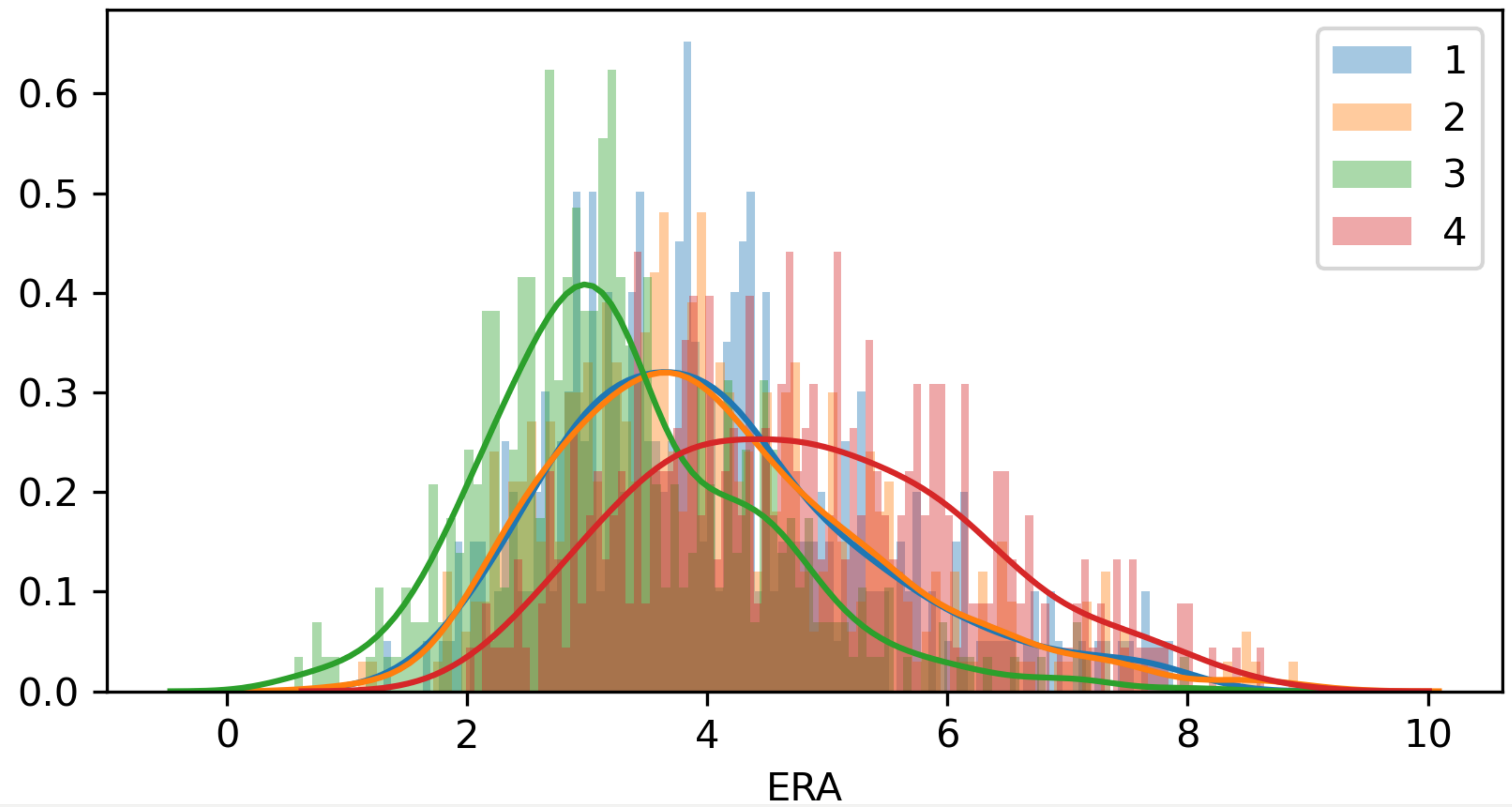
- 將投手分為四種類型 **LEVEL1 :**K多BB多、**LEVEL2:** K少BB少、**LEVLE3:** K多BB少、**LEVEL4:** K少BB多 。



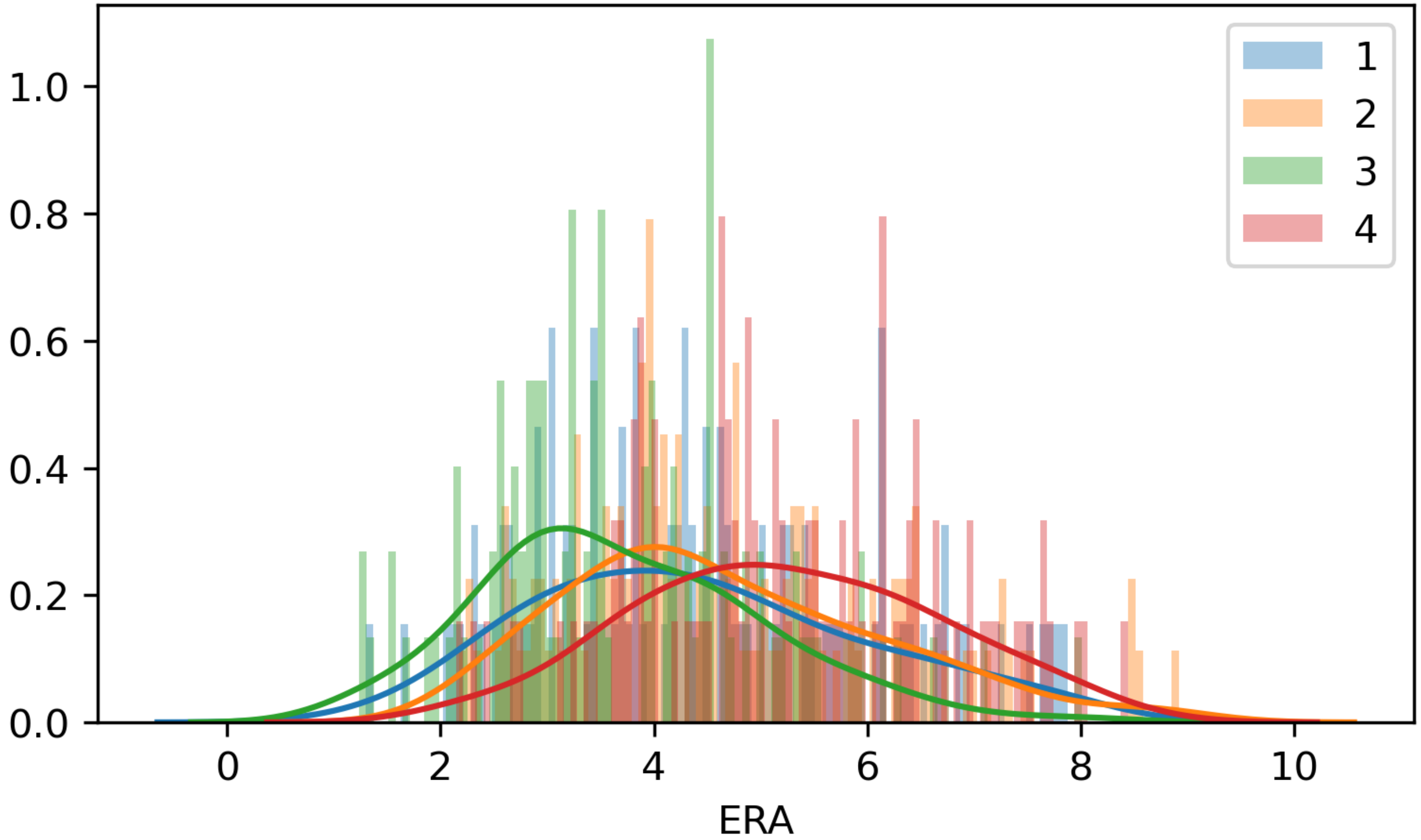
KMEANS



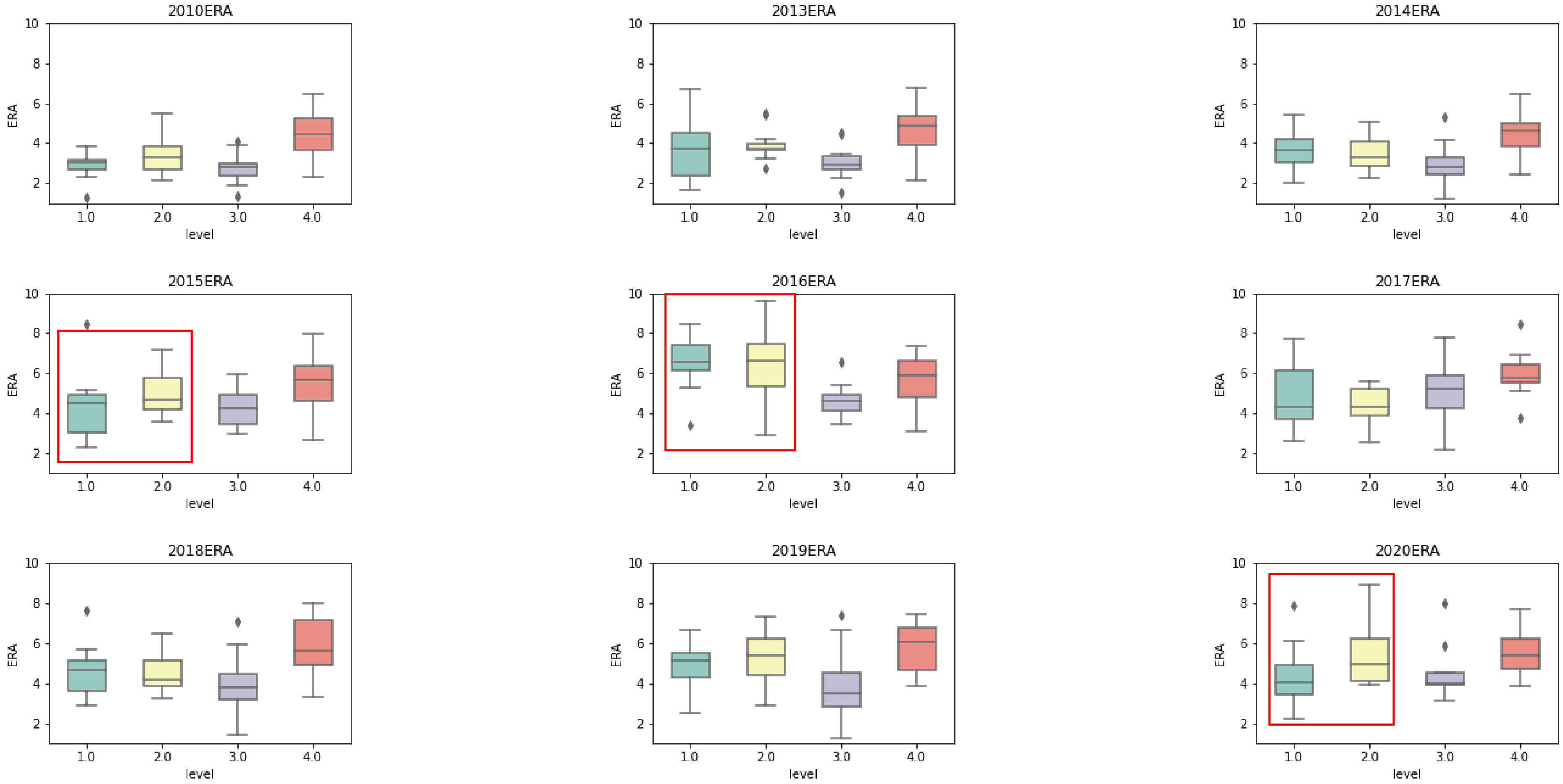
1990-2020(直方圖_ERA)



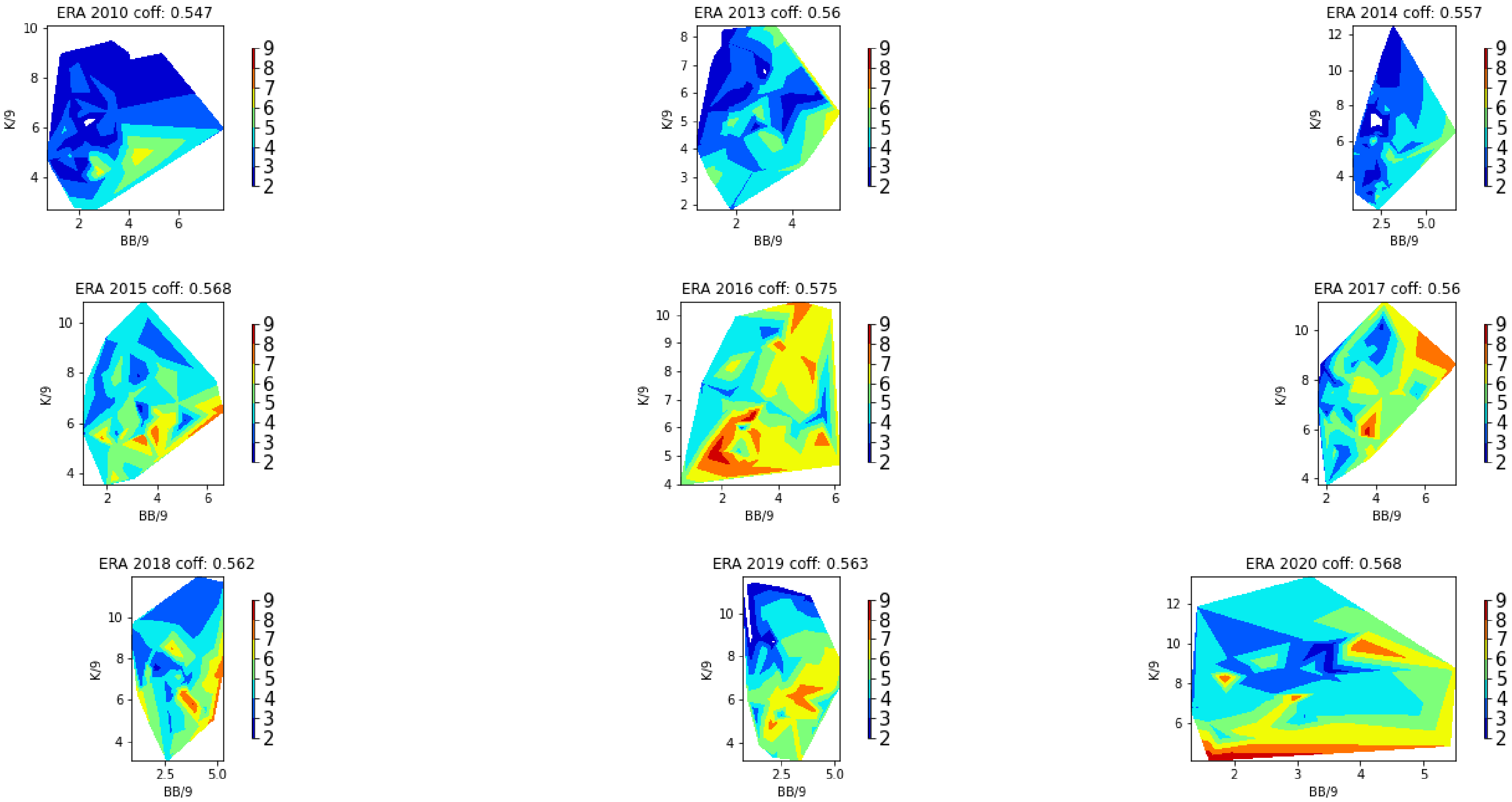
2010、2013-2020(直方圖_ERA)



2010、2013-2020(盒鬚圖)



2010、2013-2020【ERA熱圖】



ERA 2010 coff: 0.547

ERA 2013 coff: 0.56

ERA 2014 coff: 0.557

ERA 2015 coff: 0.568

ERA 2016 coff: 0.575

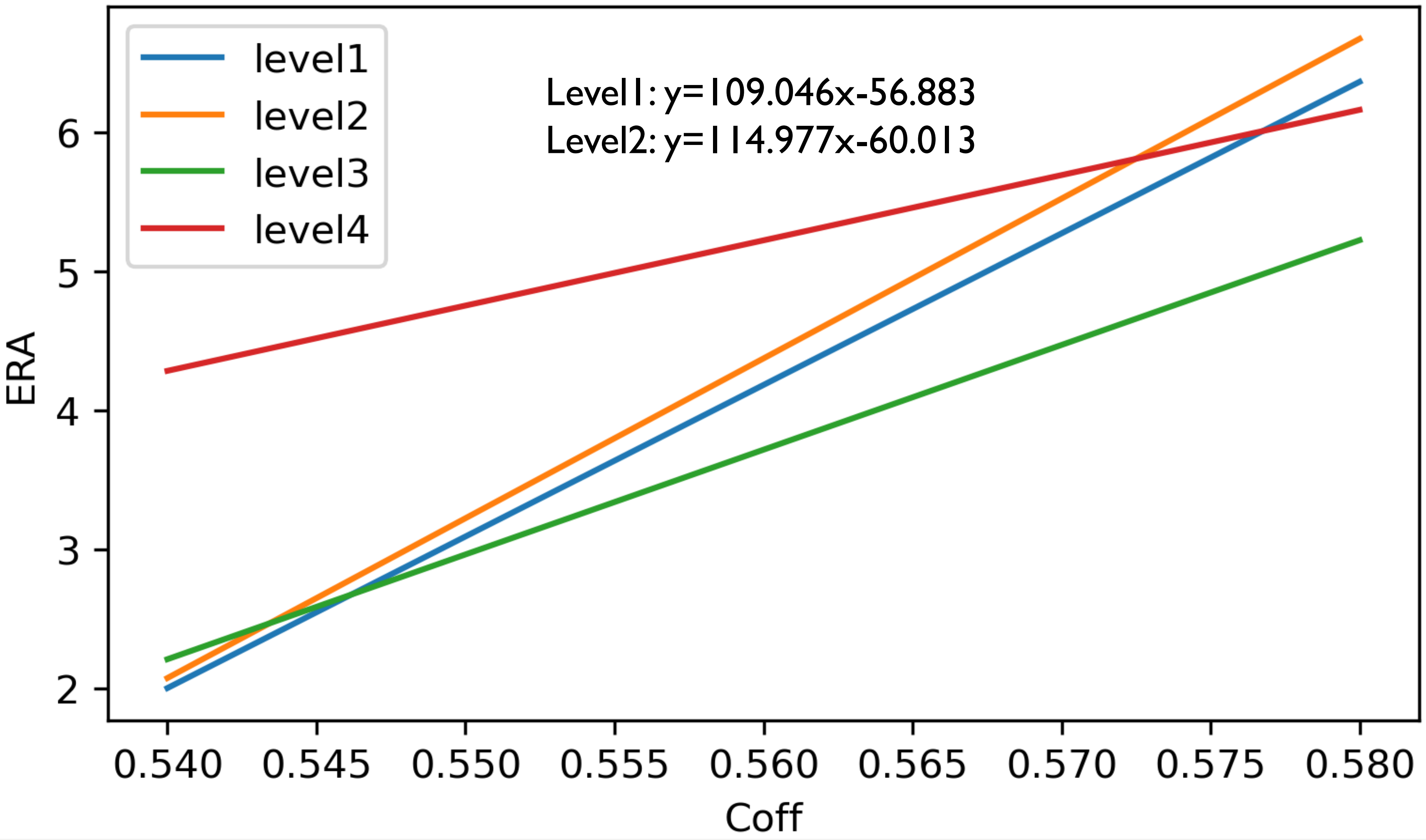
ERA 2017 coff: 0.56

ERA 2018 coff: 0.562

ERA 2019 coff: 0.563

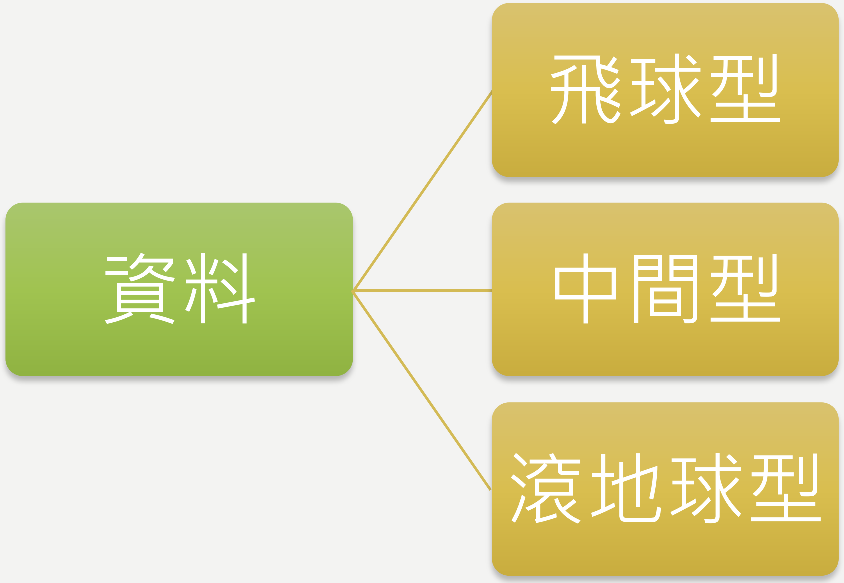
ERA 2020 coff: 0.568

2010、2013-2020 ERA(平均)回歸線



資料前處理

- 將投手分為三種類型 **LEVEL0:**飛球型 **LEVEL1:**中間型 **LEVEL2:**滾地球型



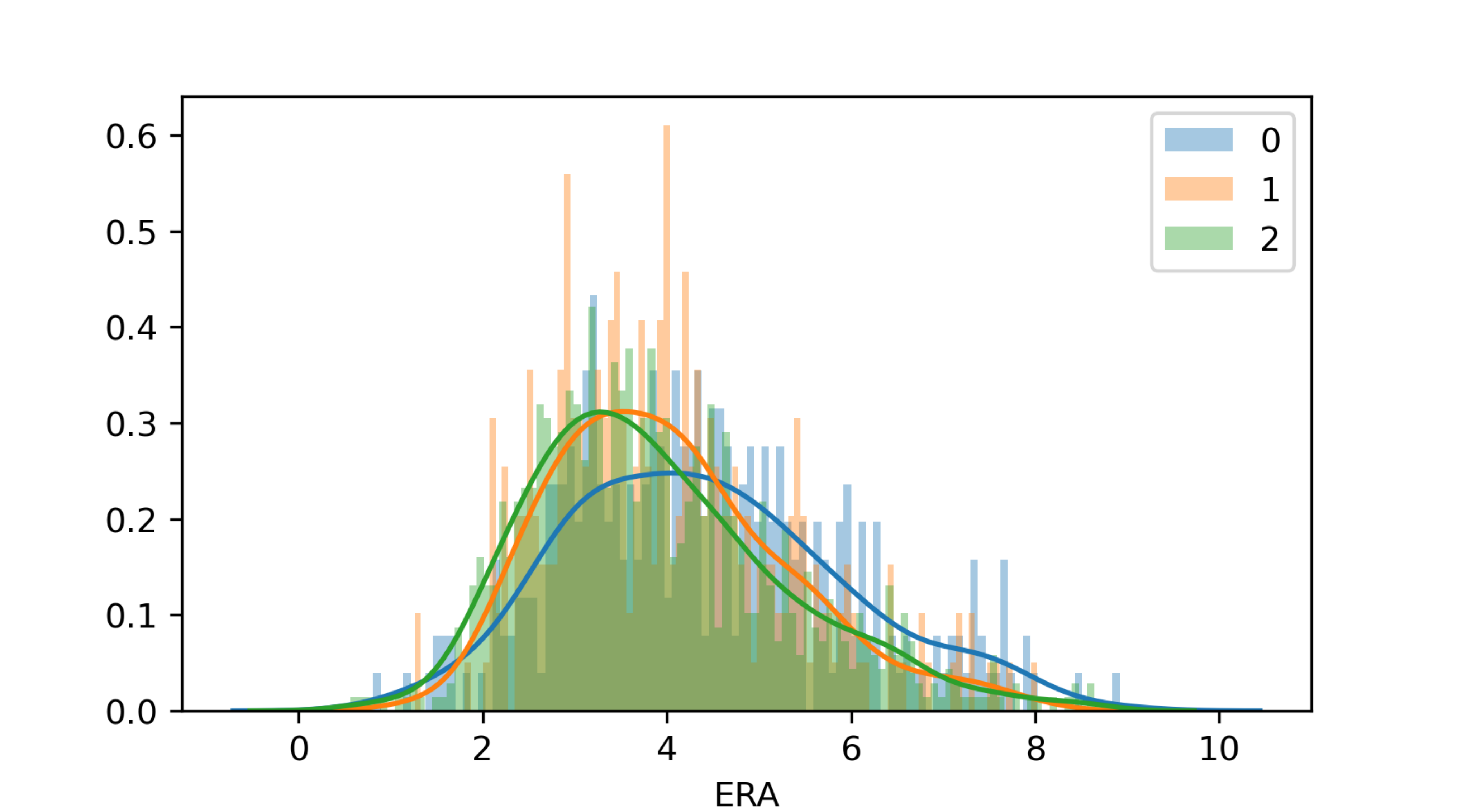
資料

飛球型

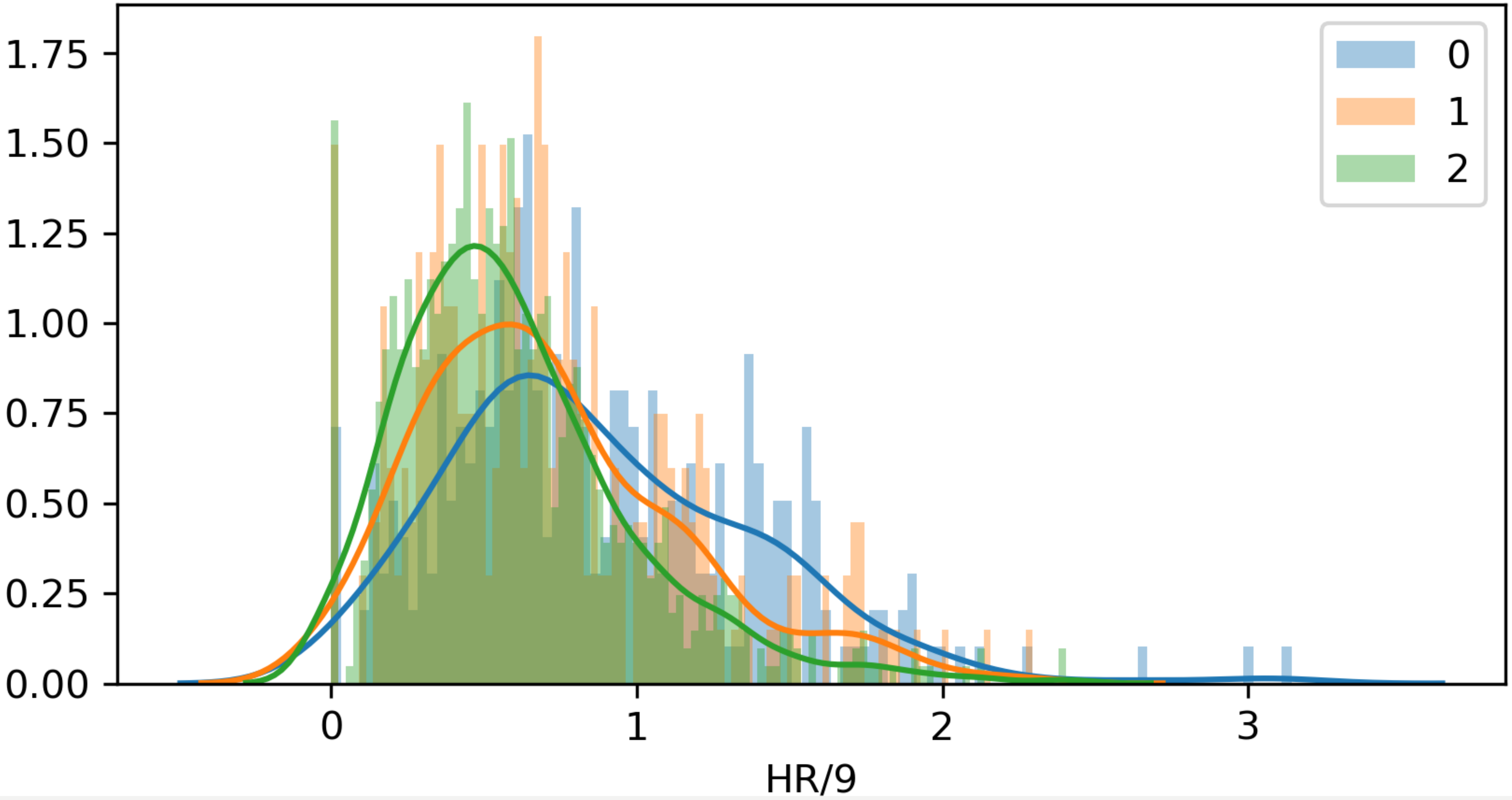
中間型

滾地球型

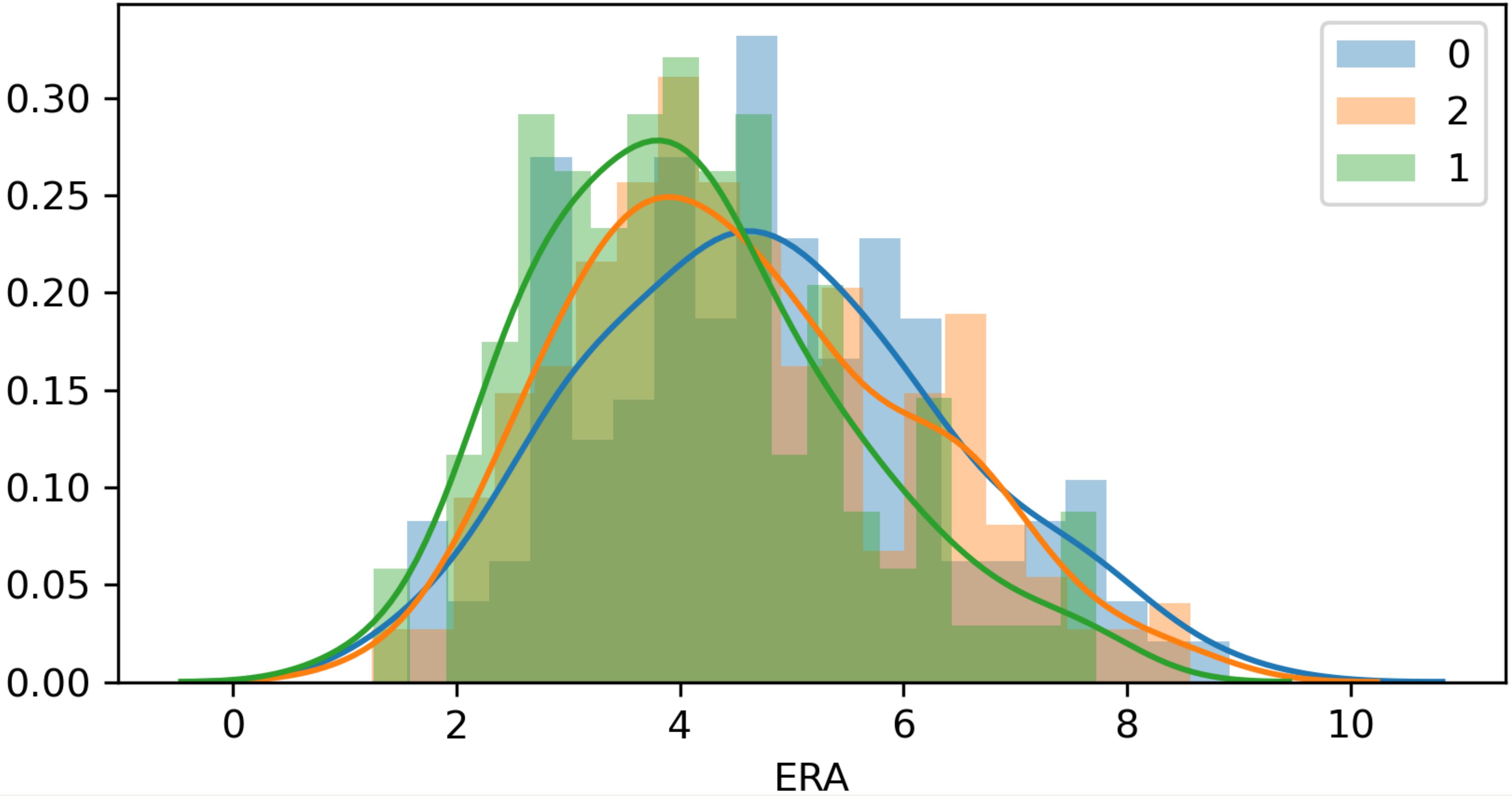
1990-2020(直方圖_ERA)



1990-2020(直方圖_HR/9)

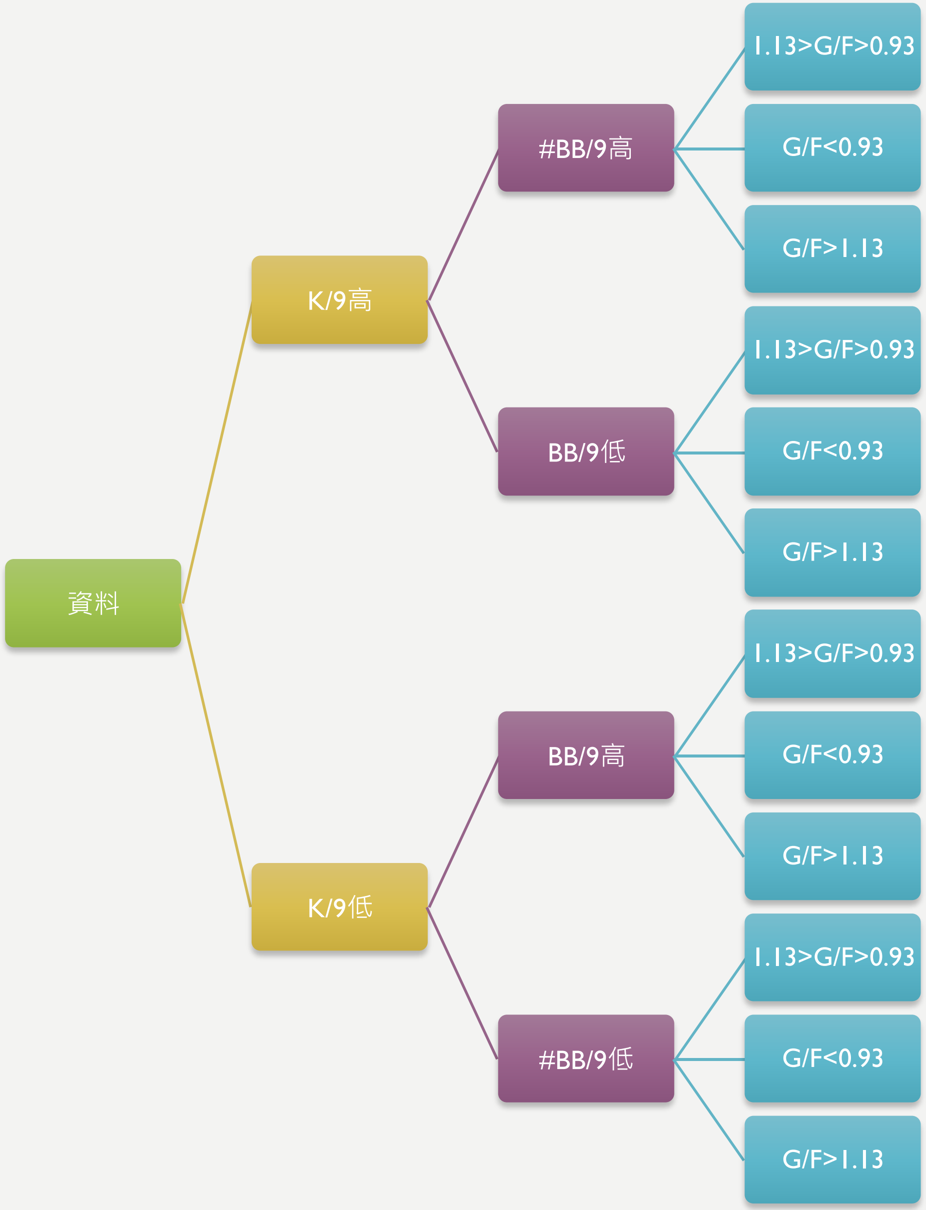


2010、2013-2020(直方圖_ERA)

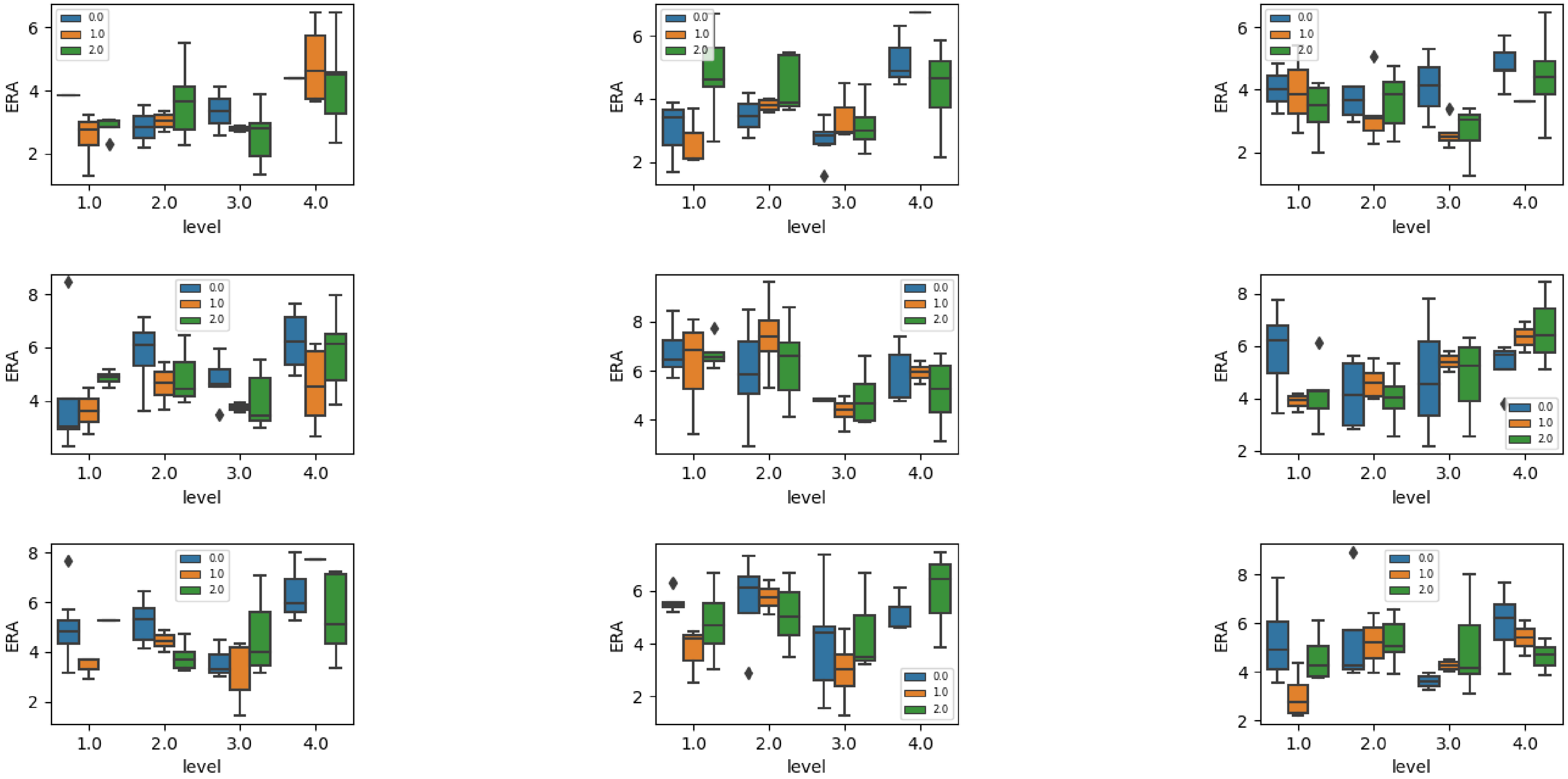


資料前處理

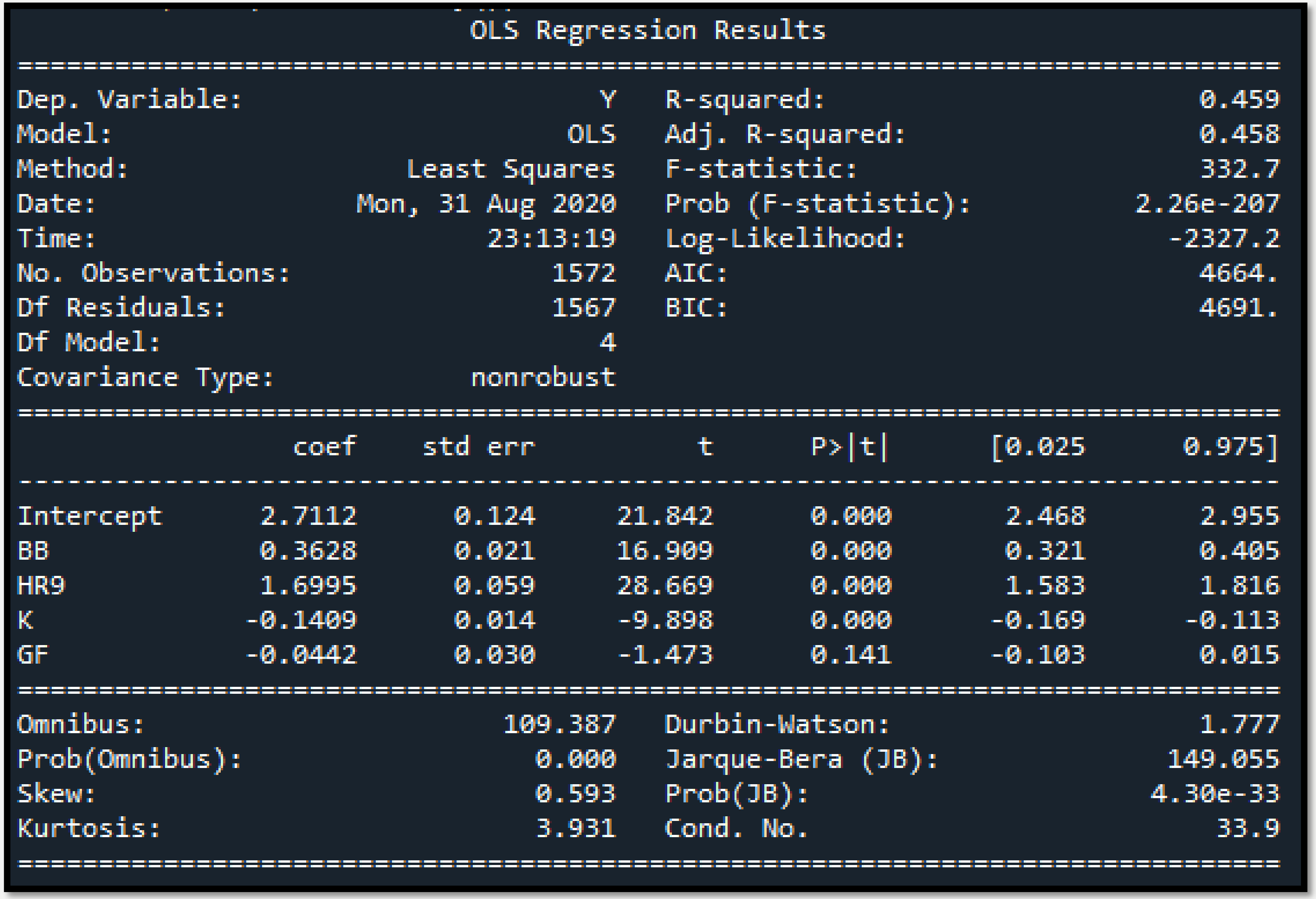
- 將資料分為12塊（4*3）



資料前處理

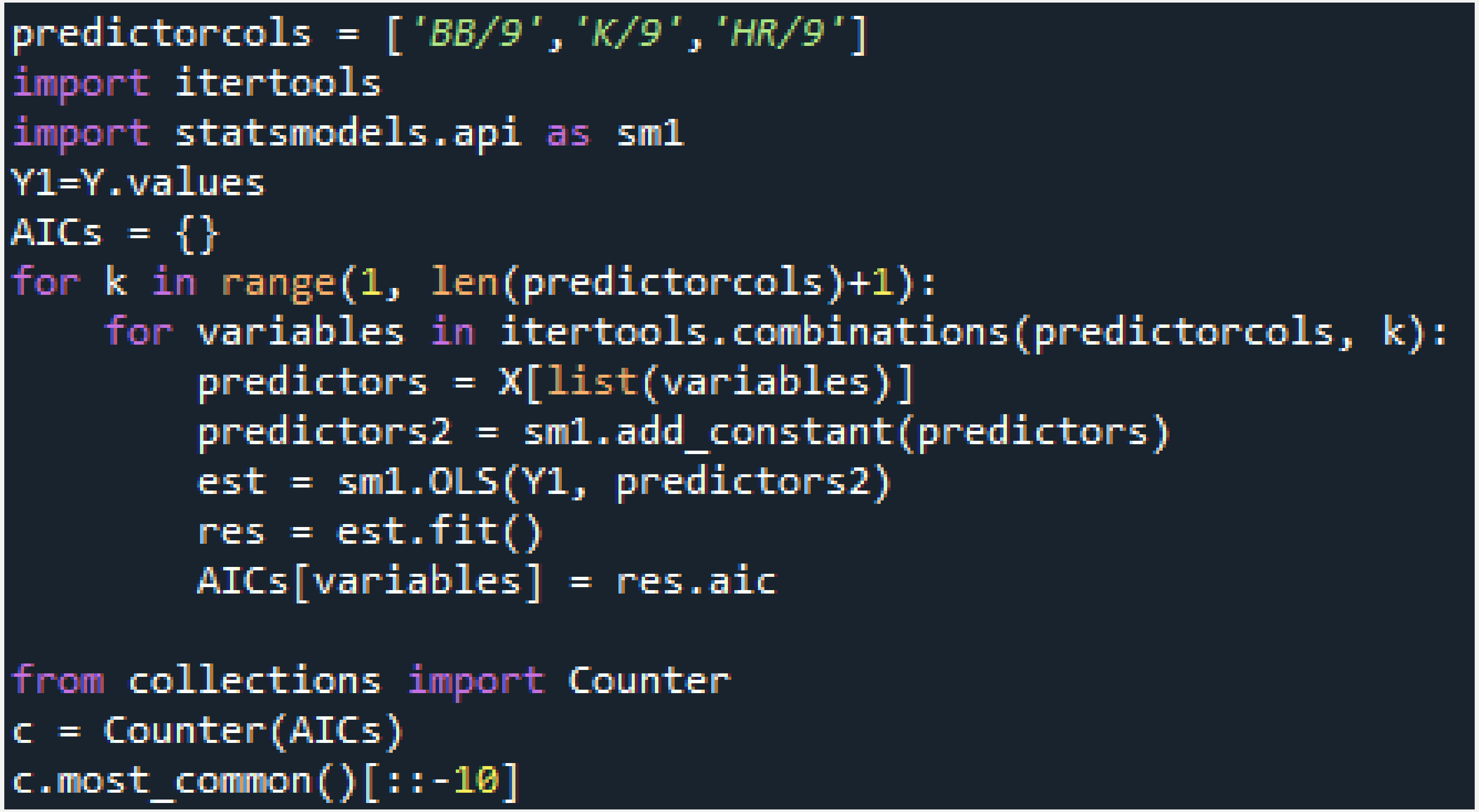

ERA
level
1.0
2.0
3.0
4.0
2
4
6
8
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0
ERA
level
1.0
2.0
3.0
4.0
0.0
1.0
2.0

ERA預測模型(建模)



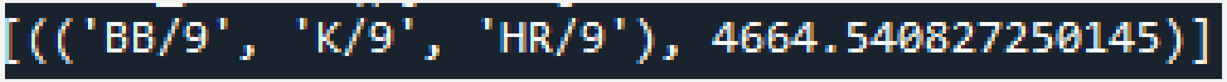
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.459
Model:                            OLS   Adj. R-squared:                  0.458
Method:                 Least Squares   F-statistic:                     332.7
Date:                Mon, 31 Aug 2020   Prob (F-statistic):          2.26e-207
Time:                        23:13:19   Log-Likelihood:                -2327.2
No. Observations:                1572   AIC:                             4664.
Df Residuals:                    1567   BIC:                             4691.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.7112      0.124     21.842      0.000       2.468       2.955
BB             0.3628      0.021     16.909      0.000       0.321       0.405
HR9            1.6995      0.059     28.669      0.000       1.583       1.816
K             -0.1409      0.014     -9.898      0.000      -0.169      -0.113
GF            -0.0442      0.030     -1.473      0.141      -0.103       0.015
==============================================================================
Omnibus:                      109.387   Durbin-Watson:                   1.777
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              149.055
Skew:                           0.593   Prob(JB):                     4.30e-33
Kurtosis:                       3.931   Cond. No.                         33.9
==============================================================================
```

ERA預測模型(AIC)



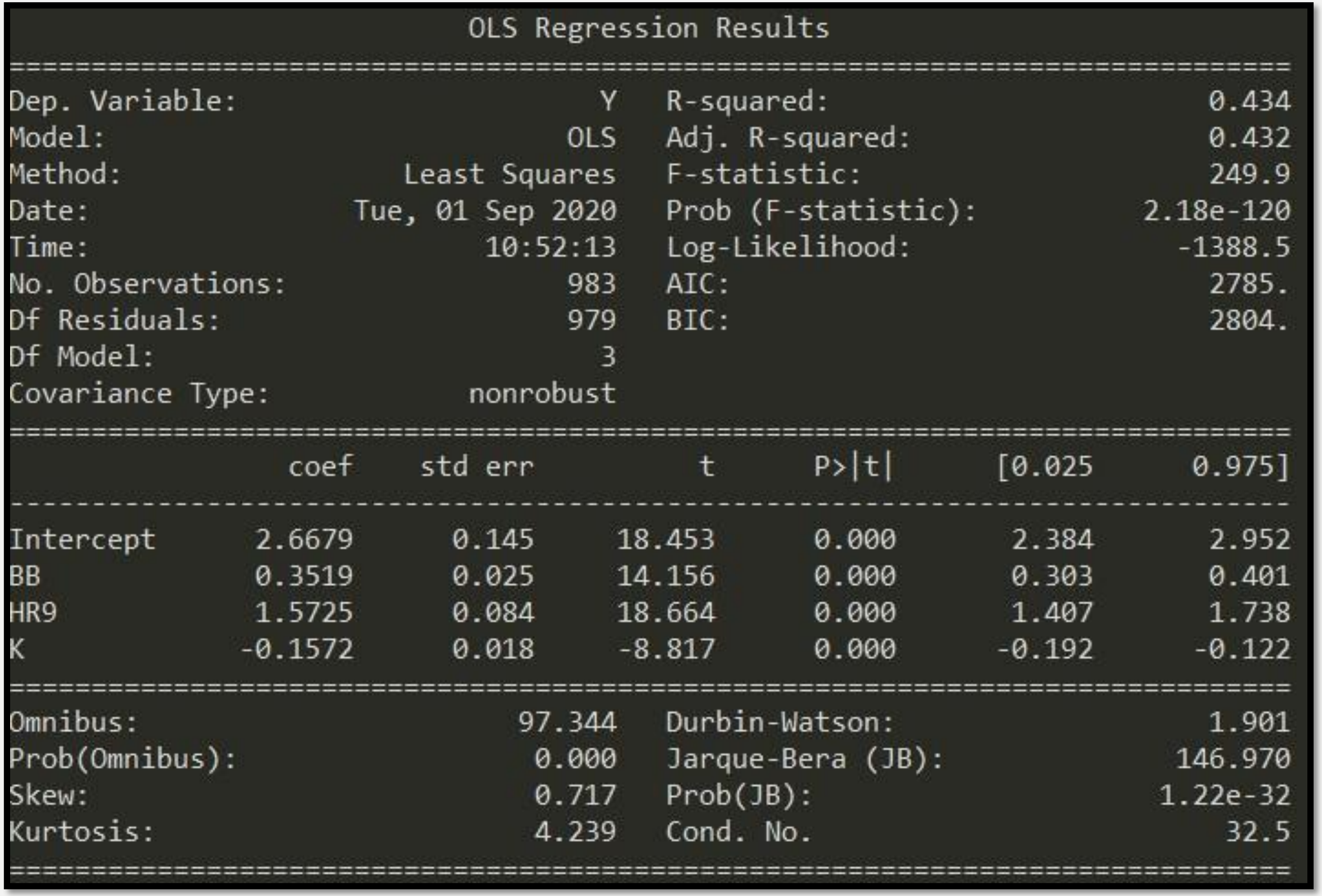
```
predictorcols = ['BB/9','K/9','HR/9']
import itertools
import statsmodels.api as sm1
Y1=Y.values
AICs = {}
for k in range(1, len(predictorcols)+1):
    for variables in itertools.combinations(predictorcols, k):
        predictors = X[list(variables)]
        predictors2 = sm1.add_constant(predictors)
        est = sm1.OLS(Y1, predictors2)
        res = est.fit()
        AICs[variables] = res.aic

from collections import Counter
c = Counter(AICs)
c.most_common()[::-10]
```



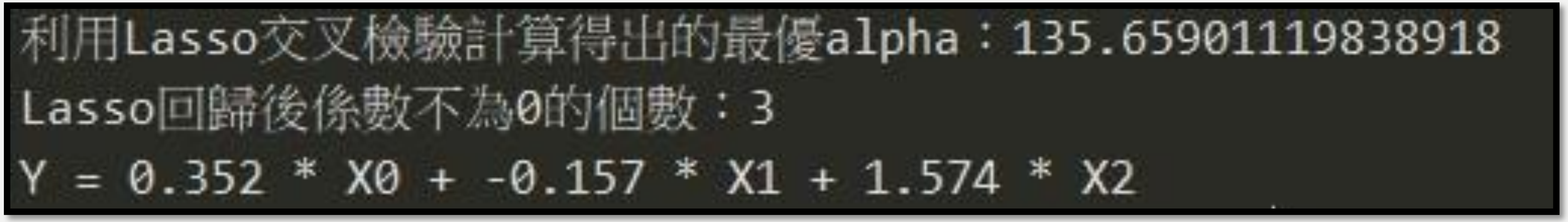
```
[(('BB/9', 'K/9', 'HR/9'), 4664.540827250145)]
```

ERA預測模型(重新建模)

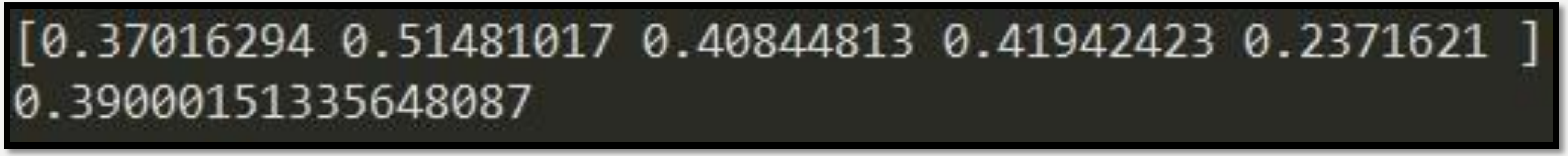


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.434
Model:                            OLS   Adj. R-squared:                  0.432
Method:                 Least Squares   F-statistic:                     249.9
Date:                Tue, 01 Sep 2020   Prob (F-statistic):          2.18e-120
Time:                        10:52:13   Log-Likelihood:                -1388.5
No. Observations:                 983   AIC:                             2785.
Df Residuals:                     979   BIC:                             2804.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.6679      0.145     18.453      0.000       2.384       2.952
BB             0.3519      0.025     14.156      0.000       0.303       0.401
HR9            1.5725      0.084     18.664      0.000       1.407       1.738
K             -0.1572      0.018     -8.817      0.000      -0.192      -0.122
==============================================================================
Omnibus:                       97.344   Durbin-Watson:                   1.901
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              146.970
Skew:                           0.717   Prob(JB):                     1.22e-32
Kurtosis:                       4.239   Cond. No.                         32.5
==============================================================================
```

LASSO交叉檢驗



```
利用Lasso交叉檢驗計算得出的最優alpha：135.65901119838918
Lasso回歸後係數不為0的個數：3
Y = 0.352 * X0 + -0.157 * X1 + 1.574 * X2
```



```
[0.37016294 0.51481017 0.40844813 0.41942423 0.2371621 ]
0.39000151335648087
```