1102 科學計算軟體期末報告

主題 ： 棒球運動之投出球之轉速與其他變因之關係

組別：第 5 組

組員：F64096180 姚敦翔、F64094031 張昊翰、F64091017 游崴鈞、

F64109519 林均霈、F64109527 蕭合亭、F64081088 李瑄哲

一、 簡介

透過大聯盟數據，探討棒球運動中轉速的高低與其他變因之關係。

二、 材料介紹

ff_avg_spin 轉速:

投手投擲時球的旋轉速度，指標單位以每分鐘轉圈數(RPM)來衡量。

xslg 被長打率 ：

打擊者每一次擊球可以貢獻幾個壘包。

exit_velocity_avg 擊球初速 ：

打者把球擊出時，球離開球棒時之速度。

launch_angle_avg 擊球仰角

打者把球擊出時，球離開球棒時與地面形成之角度。

hard_hit_percent 強擊球率

只要擊球初超過 95 英里，及數強擊球，依此比例去計算強擊球率。

z_swing_miss_percent 揮空率 ：

打者揮棒卻無觸擊球及數揮空，依比例去算揮空率。

groundballs_percent 滾地球比 ：

擊球者擊出滾地球與飛球之比例。

flyballs_percent 飛球比

擊球者擊出飛球與滾地球之比例。

三、 結果與討論

敘述統計

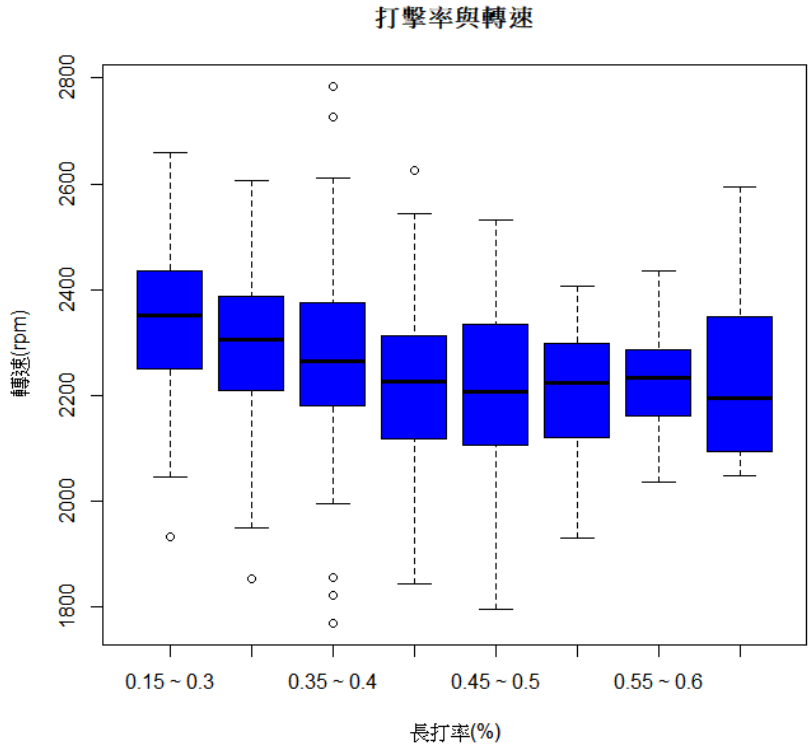
```
> summary(dataset)
  last_name          first_name          player_id          year          xslg            xwoba        exit_velocity_avg
 Length:625         Length:625         Min.   :425794   Min.   :2021   Min.   :0.1990   Min.   :0.1900   Min.   :83.00
 Class :character   Class :character   1st Qu.:572193   1st Qu.:2021   1st Qu.:0.3580   1st Qu.:0.2950   1st Qu.:87.60
 Mode  :character   Mode  :character   Median :621107   Median :2021   Median :0.4040   Median :0.3190   Median :89.00
                                       Mean   :605142   Mean   :2021   Mean   :0.4106   Mean   :0.3215   Mean   :88.85
                                       3rd Qu.:656954   3rd Qu.:2021   3rd Qu.:0.4560   3rd Qu.:0.3470   3rd Qu.:90.10
                                       Max.   :685503   Max.   :2021   Max.   :0.7950   Max.   :0.4940   Max.   :95.50
 launch_angle_avg sweet_spot_percent groundballs_percent flyballs_percent  ff_avg_speed     ff_avg_spin      轉速區間
 Min.   :-6.50    Min.   :17.10     Min.   :20.00       Min.   : 7.70     Min.   : 82.70   Min.   :1769   Length:625
 1st Qu.: 9.70    1st Qu.:30.60     1st Qu.:37.30       1st Qu.:22.20     1st Qu.: 92.10   1st Qu.:2146   Class :character
 Median :12.90    Median :33.40     Median :43.00       Median :26.10     Median : 93.60   Median :2254   Mode  :character
 Mean   :13.21    Mean   :33.56     Mean   :42.84       Mean   :26.24     Mean   : 93.44   Mean   :2252
 3rd Qu.:17.10    3rd Qu.:36.80     3rd Qu.:48.10       3rd Qu.:30.10     3rd Qu.: 95.00   3rd Qu.:2354
 Max.   :29.90    Max.   :54.10     Max.   :70.70       Max.   :52.50     Max.   :100.60   Max.   :2784
  X1600.1800            A                快慢              轉速區間2
 Length:625         Length:625         Length:625         Min.   :1.000
 Class :character   Class :character   Class :character   1st Qu.:2.000
 Mode  :character   Mode  :character   Mode  :character   Median :2.000
                                                          Mean   :2.112
                                                          3rd Qu.:2.000
                                                          Max.   :3.000
```

盒型圖：被長打率與轉速之盒型圖

1. 在 Excel 中將被長打率由 0.3 始、每 0.05 作為組別間距，利用 subset 函數做分流的擷取分別擷取出八組
2. 繪製出每一組的被長打率與轉速間關係之盒形圖。

程式碼：

```
xs1 <- subset(dataset, xslg < 0.3)
xs2 <- subset(dataset,xslg >= 0.3 & xslg < 0.35)
xs3 <- subset(dataset,xslg >= 0.35 & xslg < 0.4)
xs4 <- subset(dataset,xslg >= 0.4 & xslg < 0.45)
xs5 <- subset(dataset,xslg >= 0.45 & xslg < 0.5)
xs6 <- subset(dataset,xslg >= 0.5 & xslg < 0.55)
xs7 <- subset(dataset,xslg >= 0.55 & xslg < 0.6)
xs8 <- subset(dataset,xslg >= 0.6)
layout(matrix(c(1),1))
boxplot(xs1$ff_avg_spin, xs2$ff_avg_spin, xs3$ff_avg_spin, xs4$ff_avg_spin
, xs5$ff_avg_spin, xs6$ff_avg_spin, xs7$ff_avg_spin, xs8$ff_avg_spin, names
=c(0.15~0.3,0.3~0.35,0.35~0.4,0.4~0.45,0.45~0.5,0.5~0.55,0.55~0.6,0.6~0.8)
, main = "打擊率與轉速", xlab = "長打率(%)", ylab = "轉速(rpm)", col ="blue")
```



直方圖

1. 算出轉速 (ff_avg_spin) 的標準差

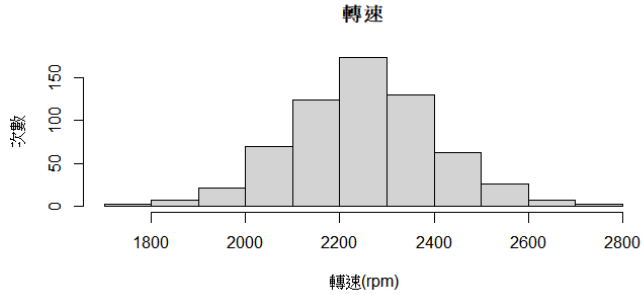
   使用 lillie 檢定，p – value = 0.8311 > 0.05 ，屬常態分布。

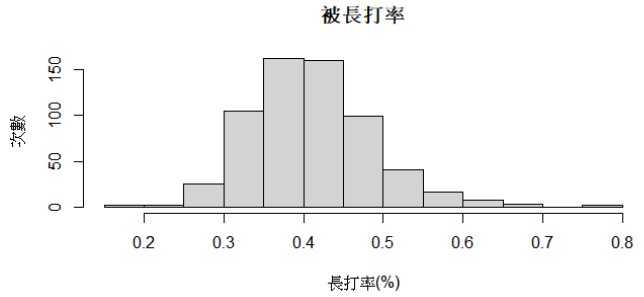
2. 算出 dataset 之被長打率 (xslg) 的標準差

   使用 lillie 檢定，p – value = 3.645e-05 < 0.05 ，屬非常態分布。

3. 使用 hist 做出兩個長方圖

```
> sd(dataset$ff_avg_spin, na.rm = TRUE)
[1] 155.1502
> lillie.test(dataset$ff_avg_spin)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dataset$ff_avg_spin
D = 0.019241, p-value = 0.8311

> sd(dataset$xslg, na.rm = TRUE)
[1] 0.07680138
> lillie.test(dataset$xslg)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dataset$xslg
D = 0.05783, p-value = 3.645e-05

> layout(matrix(c(1,2),2,1))
> hist(dataset$ff_avg_spin, main="轉速", xlab="轉速(rpm)", ylab="次數")
> hist(dataset$xslg, main="被長打率", xlab="長打率(%)", ylab="次數")
```





常態檢驗：使用 lillie test 進行常態分佈檢驗

轉速之 p-value = 0.8311 > 0.05 ，屬常態分布。

被長打率之 p-value = 3.645e-05 < 0.05 ，屬於非常態分佈。

```
> lillie.test(dataset$ff_avg_spin)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dataset$ff_avg_spin
D = 0.019241, p-value = 0.8311
```

```
> lillie.test(dataset$xslg)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  dataset$xslg
D = 0.05783, p-value = 3.645e-05
```

t-test

轉速之資料利用 subset 函數做分流的擷取，分別擷取出高、低二組。

使用 Levene Test 檢定，Pr 值 = 0.8717 > 0.05，變異數相同。

使用成對 t 檢定，因其變異數相同，故將 var.equal 設為 TRUE

```
> t.test(data0$xslg, data1$xslg, var.equal=TRUE)

        Two Sample t-test

data:  data0$xslg and data1$xslg
t = -5.1452, df = 623, p-value = 3.586e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04559586 -0.02040507
sample estimates:
mean of x mean of y
0.3883805 0.4213810
```

ANOVA

使用一因子 ANOVA 檢定被長打率，變因為轉速區間 2 之資料，轉速區間以每 400 為一單位。

使用 summary 函數查看資料分析結果，可見 p 值 = 1.58e-05 < 0.05，可知均值間有差異，須進一步使用變異數檢定。

Levene Test ：

P 值 0.1161 > 0.05 ，可知 xslg 與轉速區間 2 之變異數相等。

Turkey ：

	2-1	3-1	3-2
p	0.8106489(均值間無差異)	0.0081192(均值間無差異)	0.0000139(均值間無差異)

```
> TukeyHSD(aov2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = xslg ~ factor(轉速區間2), data = dataset)

$`factor(轉速區間2)`
             diff          lwr          upr     p adj
2-1 -0.008772727 -0.04215453  0.02460907 0.8106489
3-1 -0.047053333 -0.08401098 -0.01009568 0.0081192
3-2 -0.038280606 -0.05774537 -0.01881584 0.0000139
```

相關性檢定

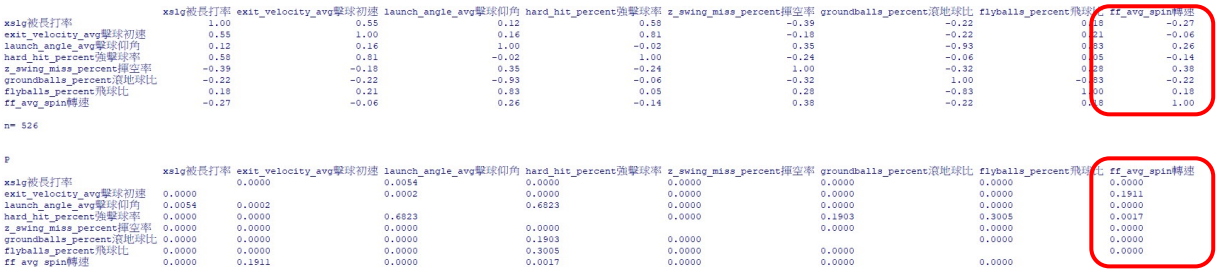
使用 spearman 做相關性檢定

程式碼 ： rcorr(as.matrix(data[, 5:12]), type=c("spearman"))

相關性	被長打率	擊球初速	擊球仰角
轉速	-0.27(低度負相關)	-0.06(低度負相關)	0.26(低度正相關)

相關性	強擊球率	揮空率	滾地球比
轉速	-0.14(低度負相關)	0.38(低度正相關)	-0.22(低度負相關)
相關性	飛球比		
轉速	0.18(低度正相關)		

顯著性	被長打率	擊球初速	擊球仰角
轉速	0.000 (達統計上顯著性)	0.1911 (未達統計上顯著性)	0.000 (達統計上顯著性)
顯著性	強擊球率	揮空率	滾地球比
轉速	0.0017 (達統計上顯著性)	0.000 (達統計上顯著性)	0.000 (達統計上顯著性)
顯著性	飛球比		
轉速	0.000 (達統計上顯著性)		



```
                        xslg被長打率 exit_velocity_avg擊球初速 launch_angle_avg擊球仰角 hard_hit_percent強擊球率 z_swing_miss_percent揮空率 groundballs_percent滾地球比 flyballs_percent飛球比 ff_avg_spin轉速
xslg被長打率                    1.00                    0.55                    0.12                    0.58                    -0.39                     -0.22                     0.18          -0.27
exit_velocity_avg擊球初速       0.55                    1.00                    0.16                    0.81                    -0.18                     -0.22                     0.21          -0.06
launch_angle_avg擊球仰角        0.12                    0.16                    1.00                    -0.02                    0.35                     -0.93                     0.83           0.26
hard_hit_percent強擊球率        0.58                    0.81                    -0.02                    1.00                    -0.24                     -0.06                     0.05          -0.14
z_swing_miss_percent揮空率      -0.39                   -0.18                    0.35                    -0.24                    1.00                     -0.32                     0.28           0.38
groundballs_percent滾地球比     -0.22                   -0.22                   -0.93                    -0.06                   -0.32                      1.00                    -0.83          -0.22
flyballs_percent飛球比           0.18                    0.21                    0.83                    0.05                    0.28                     -0.83                     1.00           0.18
ff_avg_spin轉速                 -0.27                   -0.06                    0.26                    -0.14                   0.38                     -0.22                     0.18           1.00

n= 526

p
                        xslg被長打率 exit_velocity_avg擊球初速 launch_angle_avg擊球仰角 hard_hit_percent強擊球率 z_swing_miss_percent揮空率 groundballs_percent滾地球比 flyballs_percent飛球比 ff_avg_spin轉速
xslg被長打率                    0.0000                  0.0054                  0.0000                  0.0000                  0.0000                    0.0000                    0.0000
exit_velocity_avg擊球初速       0.0000                  0.0002                  0.0000                  0.0000                  0.0000                    0.0000                    0.1911
launch_angle_avg擊球仰角        0.0054          0.0002                          0.6823                  0.0000                  0.0000                    0.0000                    0.0000
hard_hit_percent強擊球率        0.0000          0.0000                  0.6823                          0.0000                  0.1903                    0.3005                    0.0017
z_swing_miss_percent揮空率      0.0000          0.0000                  0.0000                  0.0000                          0.0000                    0.0000                    0.0000
groundballs_percent滾地球比     0.0000          0.0000                  0.0000                  0.1903                  0.0000                            0.0000                    0.0000
flyballs_percent飛球比          0.0000          0.0000                  0.0000                  0.3005                  0.0000                    0.0000                            0.0000
ff_avg_spin轉速                 0.0000          0.1911                  0.0000                  0.0017                  0.0000                    0.0000                    0.0000
```

變數篩選(逐步)

排除不顯著變數，使用逐步法做變數篩選。

程式碼：

```
model <- lm(ff_avg_spin 轉速 ~ xslg 被長打率 +
launch_angle_avg 擊球仰角 + hard_hit_percent 強擊球率 +
z_swing_miss_percent 揮空率 + groundballs_percent 滾地球比 +
flyballs_percent 飛球比, data= data)
ols_step_both_p(model, penter = 0.05, prem = 0.1 , details = TRUE)
```

分析：

揮空率、擊球仰角、被長打率加入，淘汰掉 p 值大於 0.1 之變數及球初速、強擊球率、滾地球比、飛球比。

Sig :

由 sig 可看見，剩餘項目皆達統計上之顯著性。

```
No more variables to be added/removed.


Final Model Output
------------------

                            Model Summary
-----------------------------------------------------------------
R                       0.425       RMSE                143.168
R-Squared               0.181       Coef. Var             6.348
Adj. R-Squared          0.176       MSE               20496.940
Pred R-Squared          0.168       MAE                 111.204
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
--------------------------------------------------------------------------
                    Sum of
                   Squares         DF    Mean Square      F         Sig.
--------------------------------------------------------------------------
Regression      2357536.794          3     785845.598     38.34    0.0000
Residual       10699402.744        522      20496.940
Total          13056939.538        525
--------------------------------------------------------------------------

                                    Parameter Estimates
-----------------------------------------------------------------------------------------------------------
                  model        Beta     Std. Error    Std. Beta       t        Sig        lower        upper
-----------------------------------------------------------------------------------------------------------
            (Intercept)    2220.971         62.580                35.490     0.000    2098.032     2343.910
z_swing_miss_percent揮空率       7.983          1.812        0.217     4.405     0.000        4.423       11.544
launch_angle_avg擊球仰角         5.518          1.266        0.196     4.359     0.000        3.031        8.006
          xslg被長打率        -461.462        109.707       -0.196    -4.206     0.000     -676.982     -245.941
-----------------------------------------------------------------------------------------------------------

                                    Stepwise Selection Summary
-----------------------------------------------------------------------------------------------------------
                                    Added/                   Adj.
Step           Variable             Removed     R-Square     R-Square     C(p)          AIC          RMSE
-----------------------------------------------------------------------------------------------------------
   1    z_swing_miss_percent揮空率    addition       0.137        0.136    27.8780    6743.8327    146.6062
   2    launch_angle_avg擊球仰角      addition       0.153        0.150    20.0900    6736.3855    145.4344
   3          xslg被長打率            addition       0.181        0.176     4.3840    6720.8523    143.1675
-----------------------------------------------------------------------------------------------------------
```

共線性

將逐步變數篩選出之變數做共線性檢定。

程式碼：

```
mode2 <- lm(ff_avg_spin 轉速 ~ xslg 被長打率 +
launch_angle_avg 擊球仰角 + z_swing_miss_percent 揮空率, data= data)
ols_coll_diag(mode2)
```

分析：

VIF 值皆未大於 3，表示其共線性數於不嚴重的，無須剔除。

與轉速之共線性	被長打率	擊球仰角	揮空率
VIF	1.376	1.293	1.551

```
Tolerance and Variance Inflation Factor
---------------------------------------
                  Variables Tolerance      VIF
1              xslg被長打率 0.7266990 1.376086
2   launch_angle_avg擊球仰角 0.7735201 1.292791
3 z_swing_miss_percent揮空率 0.6448752 1.550688

Eigenvalue and Condition Index
------------------------------
   Eigenvalue Condition Index    intercept xslg被長打率 launch_angle_avg擊球仰角 z_swing_miss_percent揮空率
1 3.834714512        1.000000 0.0006645462  0.001272864              0.007520178                0.002143023
2 0.106147185        6.010523 0.0124569920  0.025291627              0.814502901                0.002467266
3 0.052984460        8.507310 0.0001404551  0.124446762              0.076671783                0.385225535
4 0.006153843       24.962800 0.9867380067  0.848988747              0.101305137                0.610164176
```

線性回歸

程式碼：

summary(mode2)

分析：

截距	被長打率	擊球仰角	揮空率
2220.971	-461.462	5.518	7.983

模型公式：

轉速 = 2220.971 – 461.462 * 被長打率 + 5.518 * 擊球仰角 + 7.983 * 揮空率

R-square = 0.181

Adjust R-square = 0.176

```
Call:
lm(formula = ff_avg_spin轉速 ~ xslg被長打率 + launch_angle_avg擊球仰角 +
    z_swing_miss_percent揮空率, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-528.12  -93.67   -1.35   85.88  485.45

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   2220.971     62.580  35.490  < 2e-16 ***
xslg被長打率                  -461.462    109.707  -4.206 3.05e-05 ***
launch_angle_avg擊球仰角         5.518      1.266   4.359 1.57e-05 ***
z_swing_miss_percent揮空率       7.983      1.812   4.405 1.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 143.2 on 522 degrees of freedom
Multiple R-squared:  0.1806,    Adjusted R-squared:  0.1758
F-statistic: 38.34 on 3 and 522 DF,  p-value: < 2.2e-16
```

四、 結論

資料數連續型資料，我們做了變數挑選、共線性檢定、相關分析、線性回歸。

- 透過相關性檢定剔除掉擊球初速。
- 由逐步變數篩選留下轉速與被長打率、擊球仰角、揮空率，可看到剩餘變數皆達統計上的顯著性。
- 利用共線性檢定，檢查出未有嚴重共線性，無須剔除任一碧數。
- 最後，做出線性迴規模型

五、 資料來源

參考資料 ： https://baseballsavant.mlb.com/statcast_search