

目次

一、 研究背景與動機	2
二、 研究目的	2
三、 變數介紹	2
四、 模型選擇	3
五、 準則選取	4
六、 模型比較	7
七、 交互作用	9
八、 質性變數	9
九、 殘差檢定	10
十、 選定模型	10
十一、 離群 Y 觀測值之確認	11
十二、 離群 X 觀測值之確認	12
十三、 辨識影響個案:	12
十四、 多重共線性診斷	14
十五、 結論	15

一、研究背景與動機

隨著魔球(Money Ball)一書的出版，美國職棒大聯盟(MLB)開啟了數據棒球的另一個紀元，各隊經理人，紛紛將數據統計導入這個世界頂尖的運動賽事當中，雖然部分球團(如道奇隊、洋基隊)仍靠著財大氣粗，包下市場中頂尖的明星球員，但對於小市場球團來說，要與這些大球團角逐分區冠軍甚至世界大賽冠軍，就要學會如何將有限的經費花在刀口上。

二、研究目的

承接研究動機，過去，多數經理人會依據傳統指標 ERA 當作是購買球員的重要參考，但是隨著時間的推進，發現 ERA 常常帶有運氣成分，例如打者的火力不連貫留下滿滿的殘壘，或是中繼救援投手上場放火，這些都會導致先發投手的 ERA 常常有高估或是低估的情況，本分析將美國職棒大聯盟的投手當作研究目標，分析投手間各項非規責他人之數據(AVG、WHIP、GO/AO、OBP、SLG、OPS、K/9、BB/9、K/BB、P/IP、所在聯盟)找出與投手自責分率(ERA)所呈現的線性關係，以平均運氣差異所帶來錯估的情形，也藉此評估，2015 年我國旅美好手陳偉殷，是否有被高估或是低估的情況。

三、變數介紹

應變數(Y):

投手自責分率又稱之為『防禦率』，計算公式=自責分×9/投球局數。

自變數(X₁):

AVG 打擊率是常見的棒球統計數據之一，即為擊球員上場擔任打擊任務時，擊出安打的或然率，在此表示投手被打擊的機率。

自變數(X₂):

WHIP 是每位投手的「安打加四壞除以投球局數」，也就是除了失誤外，投手「平均每局讓打者上壘數」，計算公式=(安打+四壞球)/投球局數。

自變數(X₃):

GO/AO 投手製造之出局數中，使擊球員擊出滾地球及飛球的比例。

自變數(X₄):

OBP 被上壘率表示投手面對打者時被上壘的機率。

計算公式=(安打+四壞球)/投手面對所有的打席數。

自變數(X₅):

SLG 長打率是常見的棒球統計數據之一，即為擊球員每次上場擔任打擊任務時，平均能夠佔有的壘包數，在此為被長打率。

計算公式=(壘打數) / 面對所有的打席數。

自變數(X₆):

OPS 被攻擊指數表示投手面對打者時，遭打者攻擊的程度。

計算公式=OBP+SLG。

自變數(X₇):

K/9 投手平均每九局三振對手的次數，通常奪三振率越高，表示該投手為強力型投手 (Power Pitcher)。

計算公式=三振人次 x9 / 投球局數。

自變數(X₈):

BB/9 平均每 9 局投手所投出之四死球次數。

計算公式=四死球 x9 / 投球局數。

自變數(X₉):

K/BB 三振四壞比。

計算公式= 三振次數 ÷ 四壞球次數。

自變數(X₁₀):

P/IP 每局平均用球數。

計算公式=投球數/投球局數。

自變數(X₁₁):

所在聯盟為美國聯盟=1；國家聯盟=0。

四、模型選擇

先嘗試以逐步程序篩選變數(Stepwise、Forward、Backward)

Stepwise Selection

逐步選擇 的摘要									
步驟	輸入的變數	移除的變數	標籤	數目 Vars In	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	x6		被攻擊指數	1	0.8300	0.8300	9.1811	903.25	<.0001
2	x2		WHIP	2	0.0039	0.8339	6.7295	4.36	0.0381
3	x4		被上壘率	3	0.0027	0.8367	5.6249	3.08	0.0811

Forward Selection

前進選擇 的摘要								
步驟	輸入的變數	標籤	變數目 Vars In	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	x6	被攻擊指數	1	0.8300	0.8300	9.1811	903.25	<.0001
2	x2	WHIP	2	0.0039	0.8339	6.7295	4.36	0.0381
3	x4	被上壘率	3	0.0027	0.8367	5.6249	3.08	0.0811
4	x3	滾飛比	4	0.0005	0.8371	7.1100	0.51	0.4765

Backward Selection

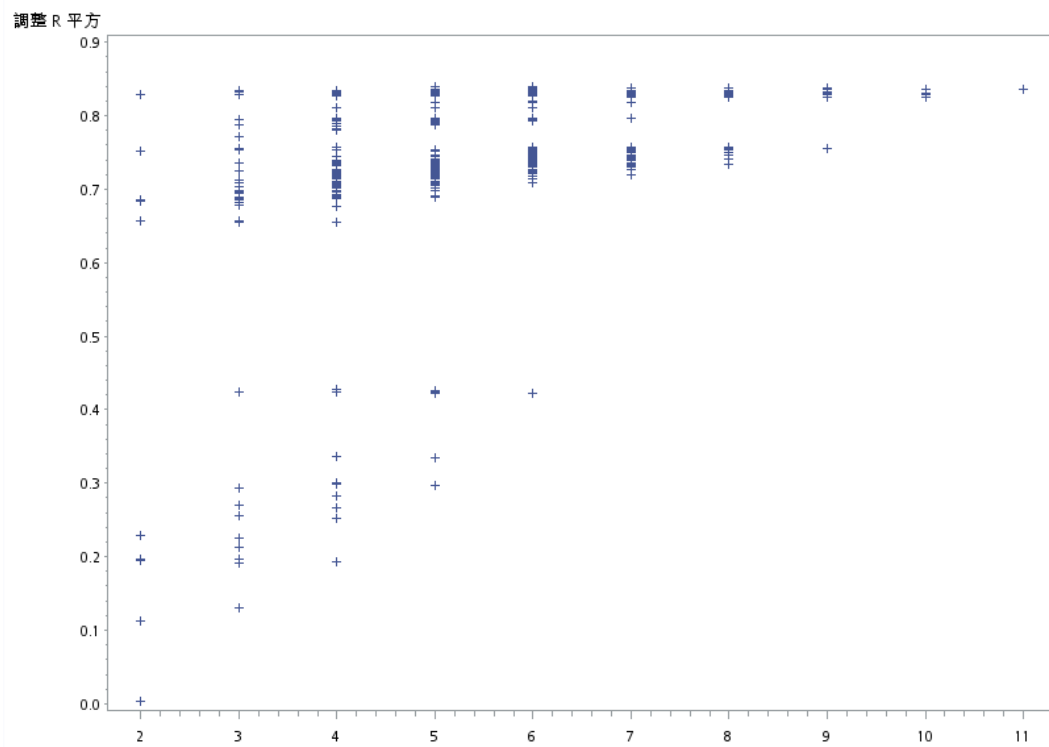
向後消去 的摘要								
步驟	移除的變數	標籤	變數目 Vars In	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	x9	K/BB	9	0.0000	0.8443	9.0000	0.00	0.9952
2	x7	K/9	8	0.0000	0.8443	7.0012	0.00	0.9728
3	x4	被上壘率	7	0.0000	0.8443	5.0256	0.02	0.8752
4	x10	P/IP	6	0.0005	0.8438	3.6293	0.61	0.4344
5	x3	滾飛比	5	0.0003	0.8435	1.9598	0.34	0.5624
6	x5	被長打率	4	0.0007	0.8427	0.8064	0.87	0.3533

移除上述變數剩下 X1(被打擊率)、X2(WHIP)、X6(被攻擊指數)、X8(BB/9)
上述三種方式除了 X2(WHIP)、X6(被攻擊指數)相同以外，其餘變數略有差異。

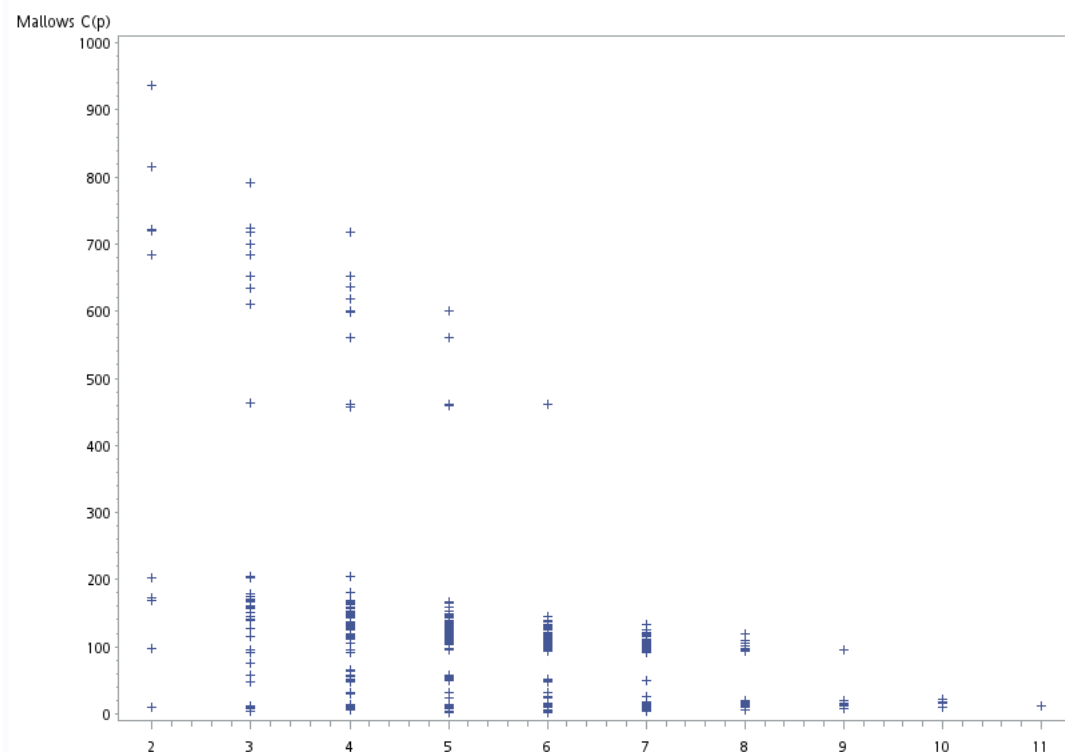
五、準則選取

由於待選取之變數僅 10 種(共 $2^{10}=1024$ 種組合)，對於電腦來說並不是像當龐大的運算，故以下直接以準則(Ra2、Cp、AIC、SBC)做判斷來選取。

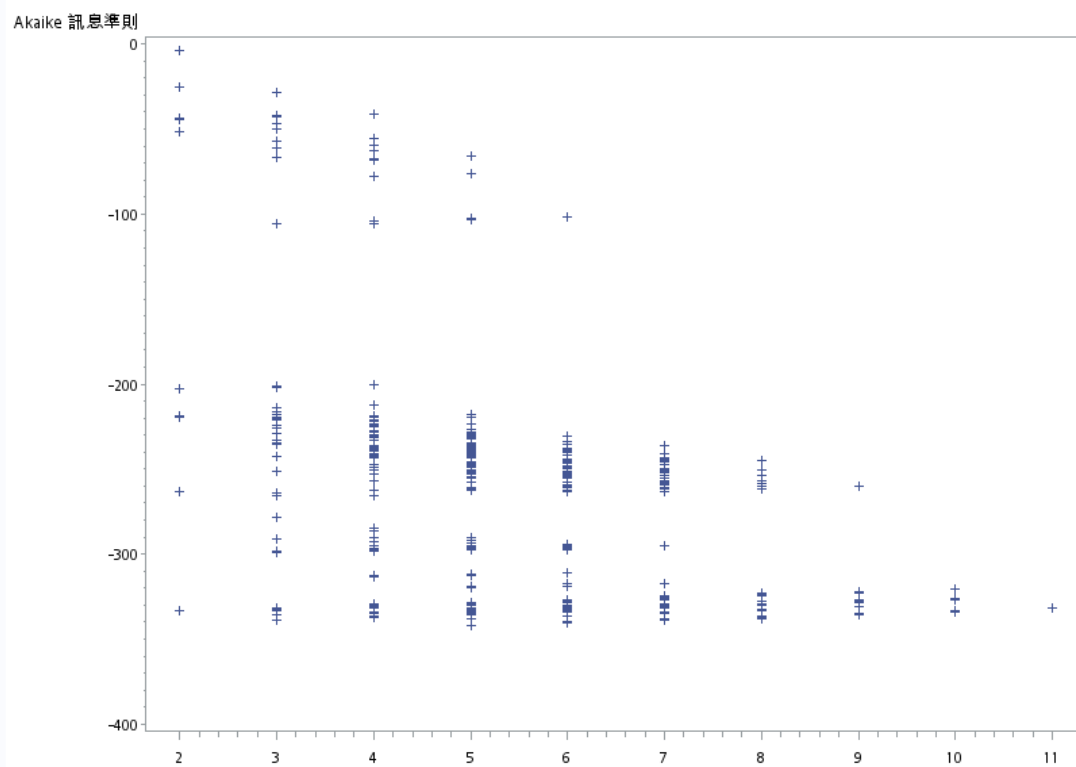
R_a^2 (調整 R 平方)



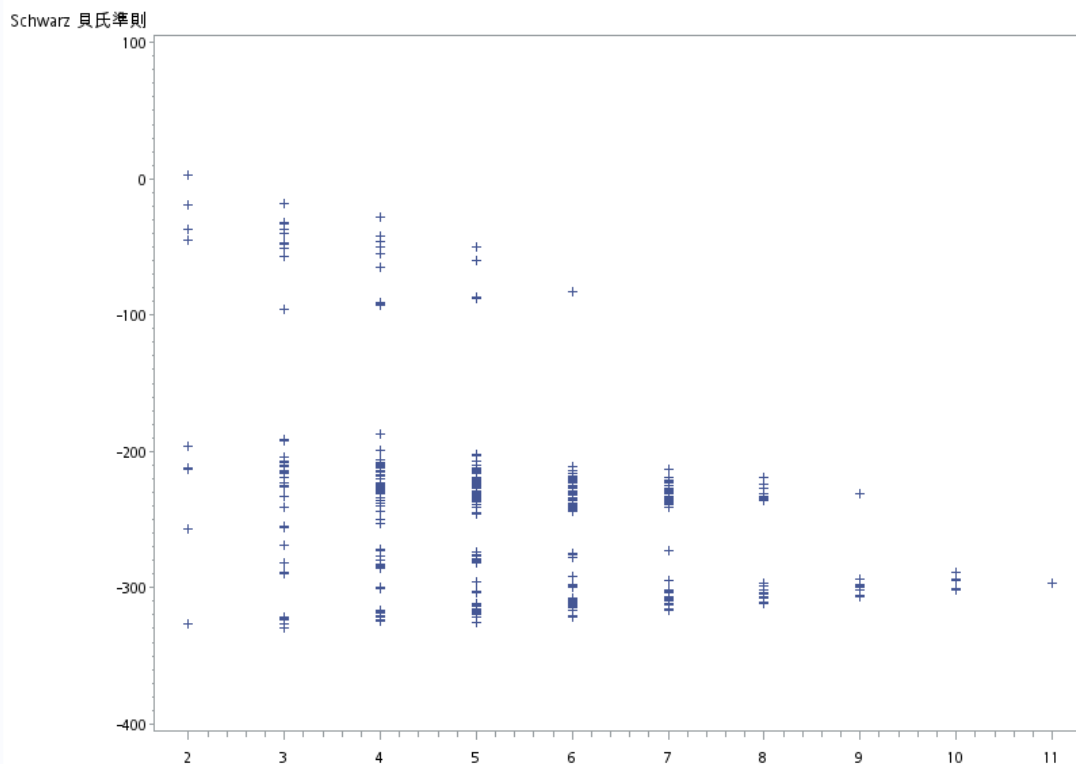
Mallows' Cp



AIC



SBC



發現 R_a^2 的最大值出現在 $P=5$ (選取 4 種變數)，且 C_p 以及 AIC 的最小值亦出現在 $P=5$ 的情況，僅 SBC 的最小值出現在 $P=3$ 。

準則最佳子集

準則	最佳值	R^2	模型中的變數
R_a^2	0.8393	0.8427	$X_1 X_2 X_6 X_8$
C_p	0.8064(<5)	0.8427	$X_1 X_2 X_6 X_8$
AIC	-341.9269	0.8427	$X_1 X_2 X_6 X_8$
SBC	-329.129	0.8366	$X_2 X_5$

發現除了 SBC 的最低點出現在子集($X_2 X_5$)，其餘準則的最佳值都出現在子集($X_1 X_2 X_6 X_8$)，與 Backward Selection 相同，在此使用 SBC 準則所決定的模型較為精簡。以下針對兩種模型做比較。

六、模型比較

Statistic	N=187	N=95	N=187	N=95
P	5	5	3	3
b_0	-2.39705	-1.51633	-2.97473	-2.86489
$S(b_0)$	0.51588	0.70605	0.23387	0.28607
b_1 (被打擊率)	-25.68886	-42.62762		
$S(b_1)$	8.18480	12.16369		
b_2 (WHIP)	6.03667	8.36854	2.38222	2.05758
$S(b_2)$	1.69971	2.38065	0.24468	0.34028
b_5 (被長打率)			9.60023	10.38696
$S(b_5)$			0.74196	1.00309
b_6 (被攻擊指數)	9.43940	11.217		
$S(b_6)$	0.88125	1.22262		
b_8 (BB/9)	-0.58457	-0.88106		
$S(b_8)$	0.18569	0.26855		
SSE	28.47842	13.23784	29.5812	14.43946
C_p	5	5	3	3
MSE	0.15647	0.14709	0.16077	0.15695
MSPR	0.1522910		0.1581882	
R_a^2	0.8393	0.8751	0.8349	0.8668
AIC	-341.9269	-177.2257	-338.8223	-172.9716
SBC	-325.77136	-164.45636	-329.129	-165.31

由以上表格可知，參數估計值以及標準差沒有正負號相互矛盾的情況，且差異並不明顯，另 MSPR 很接近 MSE_p 表示兩模型的預測能力皆受肯定。

Pearson 相關係數, N = 187 Prob > r (位於 H0 底下): Rho=0						
	y	x1	x2	x5	x6	x8
y 防禦率	1.00000	0.82800 <.0001	0.82944 <.0001	0.86745 <.0001	0.91104 <.0001	0.34244 <.0001
x1 被打擊率	0.82800 <.0001	1.00000	0.88906 <.0001	0.81079 <.0001	0.90081 <.0001	0.21168 0.0036
x2 WHIP	0.82944 <.0001	0.88906 <.0001	1.00000	0.72631 <.0001	0.87739 <.0001	0.62491 <.0001
x5 被長打率	0.86745 <.0001	0.81079 <.0001	0.72631 <.0001	1.00000	0.96349 <.0001	0.15385 0.0355
x6 被攻擊指數	0.91104 <.0001	0.90081 <.0001	0.87739 <.0001	0.96349 <.0001	1.00000	0.34068 <.0001
x8 BB/9	0.34244 <.0001	0.21168 0.0036	0.62491 <.0001	0.15385 0.0355	0.34068 <.0001	1.00000

另外觀察，WHIP 雖與 SLG 相關係數高達 0.72631，但比起 AVG 為 OPS 的一部分，導致相關係數皆高於 0.9，此變數間的高度相關，可能會影響模型的配適，故選定 $Y = -2.97473 + 2.38222X_2 + 9.60023X_5$ 作為我們推估的模型。

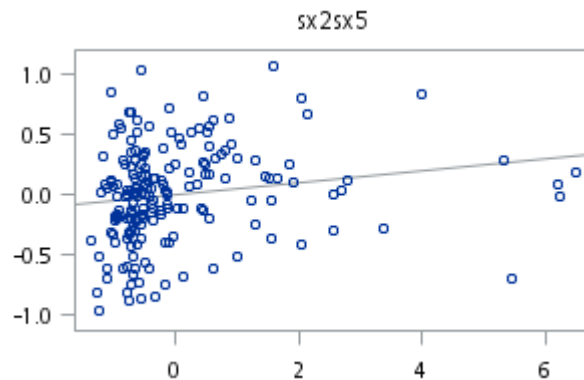
變異數分析					
來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	2	151.48500	75.74250	471.13	<.0001
誤差	184	29.58120	0.16077		
已校正的總計	186	181.06620			

根 MSE	0.40096	R 平方	0.8366
應變平均值	4.11283	調整 R 平方	0.8349
變異係數	9.74895		

參數估計值						
變數	標籤	自由度	參數估計值	標準誤差	t 值	Pr > t
Intercept	Intercept	1	-2.97473	0.23387	-12.72	<.0001
x2	WHIP	1	2.38222	0.24468	9.74	<.0001
x5	被長打率	1	9.60023	0.74196	12.94	<.0001

七、交互作用

追加變數圖



應變數 y 的檢定 test1 結果				
來源	自由度	平均值 平方	F 值	Pr > F
分子	1	0.86124	5.49	0.0202
分母	183	0.15694		

先將資料標準化(X_2 的平均數=1.3022、標準差=0.1748; X_5 的平均數=0.4151、標準差=0.0576)並命名為 SX_2 、 SX_5 ，觀察追加變數圖發現，圖形呈現一條斜率為正之直線帶狀，此圖顯示交互作用項 SX_2SX_5 新增至迴歸模型中可以幫助提供對 Y 有用的訊息，故對交互作用項作檢定。 $H_0: \beta_{25} = 0$ vs. $H_1: \beta_{25} \neq 0$
 檢定值 $F=5.49$ 且 $P\text{-value}=0.0202 < 0.05$ ，拒絕 H_0 ，我們有足夠的證據顯示交互作用項顯著，故將交互作用項加入。

八、質性變數

應變數 y 的檢定 test1 結果				
來源	自由度	平均值 平方	F 值	Pr > F
分子	1	0.78889	5.14	0.0246
分母	182	0.15347		

試圖加入質性變數 X_{11} (聯盟)，並作檢定。 $H_0: \beta_{11} = 0$ vs. $H_1: \beta_{11} \neq 0$

檢定值 $F=5.14$ 且 $P\text{-value}=0.0246<0.05$ ，拒絕 H_0 ，我們有足夠的證據顯示加入 X_{11} 顯著，故將質性變數加入。

九、殘差檢定

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.992883	Pr < W	0.4986
Kolmogorov-Smirnov	D	0.059126	Pr > D	0.1084
Cramer-von Mises	W-Sq	0.102654	Pr > W-Sq	0.1041
Anderson-Darling	A-Sq	0.541898	Pr > A-Sq	0.1686

對殘差作 Shapiro-Wilk 常態性檢定，檢定值為 0.992883，且 $P\text{-value}=0.4986>0.05$ ，故殘差呈現常態分佈。

方法	變異數	自由度	t 值	Pr > t
集區	均等	185	0.14	0.8871
Satterthwaite	不均等	178.14	0.14	0.8876

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	89	96	1.28	0.2407

對殘差作 Brown-Forsythe 檢定

檢定值 $t=0.14$ ，且 $P\text{-value}=0.8871>0.05$ ，故殘差有變異數齊一性。所以模型不需要對 Y 作轉換。

十、選定模型

$$Y = 3.99989 + 0.42786 SX_2 + 0.53624 SX_5 + 0.05992 SX_2SX_5 + 0.1343X_{11}$$

因為模型為標準化資料，故係數間可以直接作比較，由此可看出被長打率看似較 WHIP 相對重要一些，且在相同 WHIP 以及 SLG 下，美國聯盟的投手較國家聯盟的防禦率略高。

變異數分析					
來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	4	153.13513	38.28378	249.46	<.0001
誤差	182	27.93107	0.15347		
配適不足	179	27.82297	0.15544	4.31	0.1259
純誤差	3	0.10810	0.03603		
已校正的總計	186	181.06620			

根 MSE	0.39175	R 平方	0.8457
應變平均值	4.11283	調整 R 平方	0.8424
變異係數	9.52504		

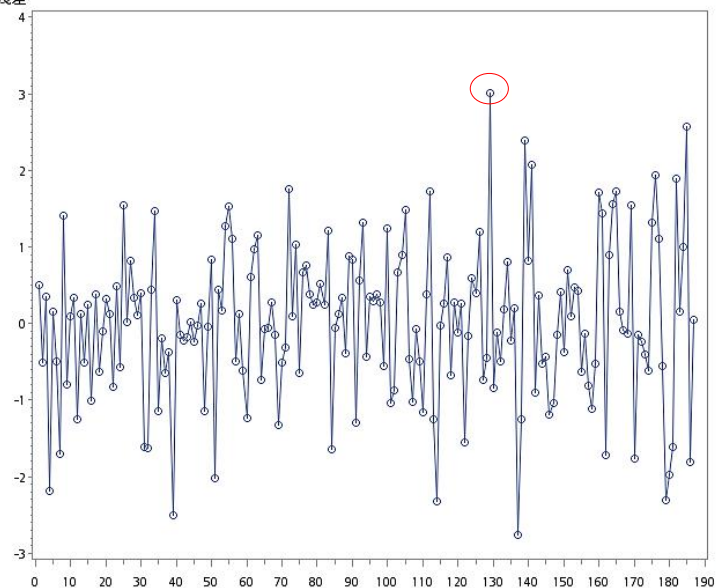
參數估計值								
變數	標籤	自由度	參數估計值	標準誤差	t 值	Pr > t	95% 信賴界限	
Intercept	Intercept	1	3.99989	0.04698	85.15	<.0001	3.90720	4.09257
sx2		1	0.42786	0.04202	10.18	<.0001	0.34495	0.51076
sx5		1	0.53624	0.04222	12.70	<.0001	0.45293	0.61955
sx2sx5		1	0.05992	0.02142	2.80	0.0057	0.01766	0.10218
x11	聯盟	1	0.13430	0.05923	2.27	0.0246	0.01742	0.25117

十一、離群 Y 觀測值之確認

Obs	residual	hat	rstudent	dffits
129	1.14235	0.016785	3.00499	0.39263

觀察第 129 位觀測值·他的 t 化去點殘差為 $3.00499 > t(0.9975; 181) = 2.84186$ ·
在 Bonferroni 程序上為離群值。

沒有目前觀測值的 Student 化殘差

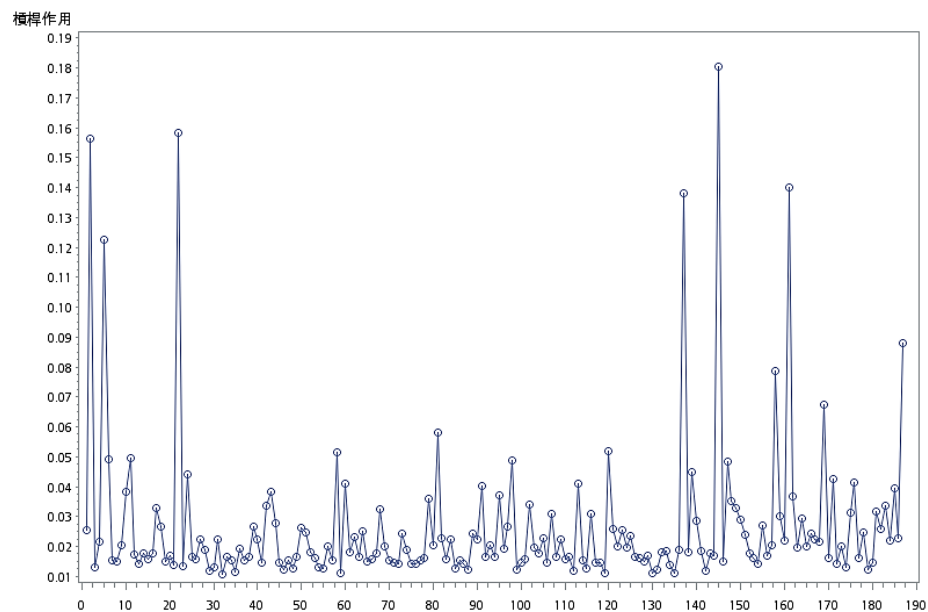


十二、離群 X 觀測值之確認

利用帽子矩陣槓桿值

$\text{Hat} > (2 \times 5/187) = 0.053476$ ，共有離群 X 觀測值 10 筆。

Obs	Player	residual	hat	rstudent	dffits
2	Arrieta, J	-0.18572	0.15616	-0.51504	-0.22156
5	Kershaw, C	0.05418	0.12254	0.14724	0.05502
22	Greinke, Z	-0.29668	0.15819	-0.82468	-0.35749
81	Kendrick, K	0.19893	0.05796	0.52214	0.12952
137	Butler, E	-0.98748	0.13814	-2.76429	-1.10667
145	Buchanan, D	-0.15753	0.18026	-0.44314	-0.20780
158	O'Sullivan, S	-0.41949	0.07848	-1.11623	-0.32574
161	Boyd, M	0.52329	0.13989	1.44460	0.58259
169	Sampson, K	0.58300	0.06748	1.54699	0.41615
187	Tomlin, J	0.01807	0.08791	0.04817	0.01495

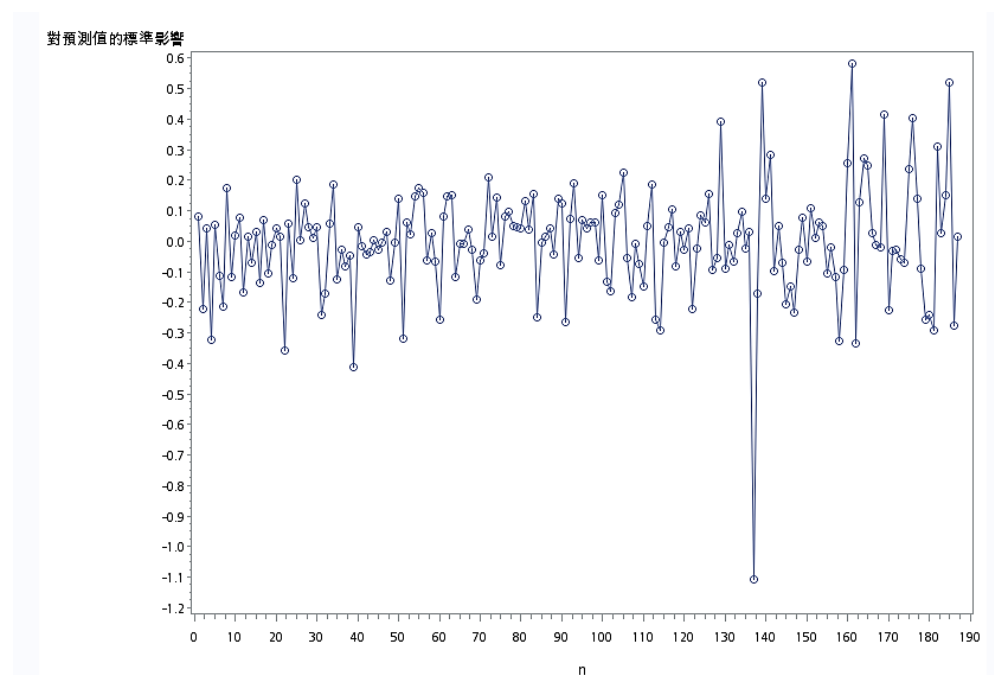


十三、辨識影響個案:

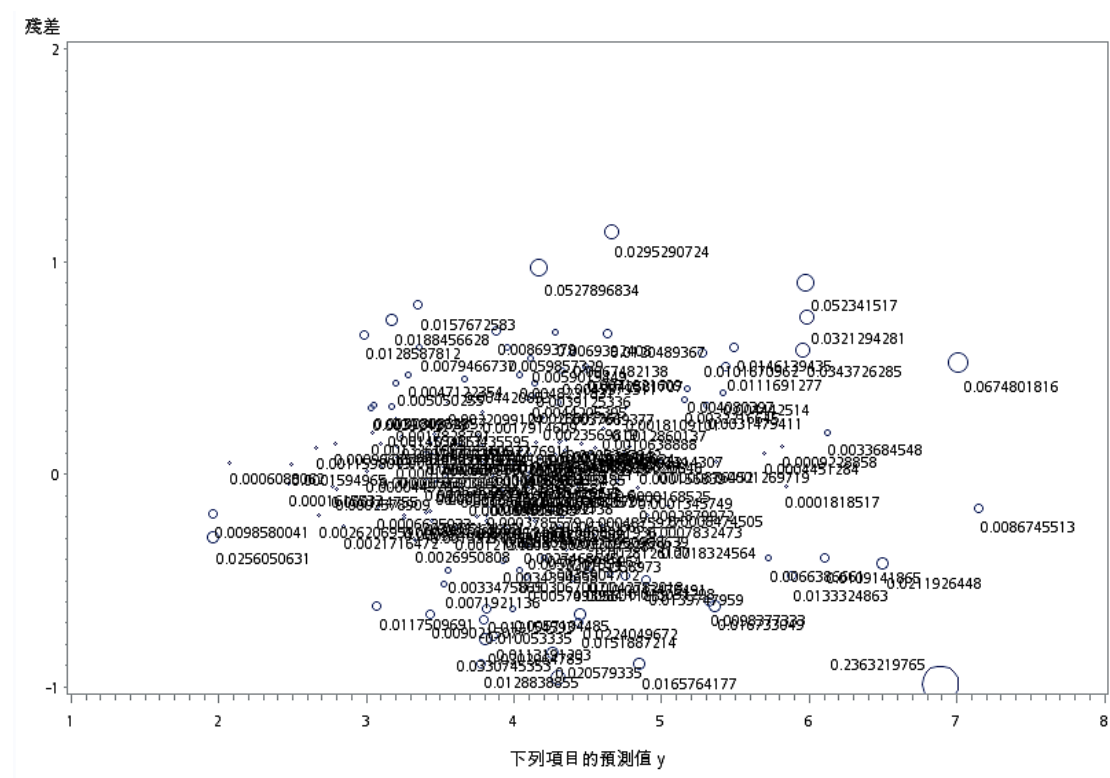
1. DFFITS

Obs	Player	residual	hat	rstudent	dffits
137	Butler, E	-0.98748	0.13814	-2.76429	-1.10667

離群 X 觀測值中，僅一筆具有影響力，不過相當靠近 1，影響力沒有大到需要採取矯正行動。

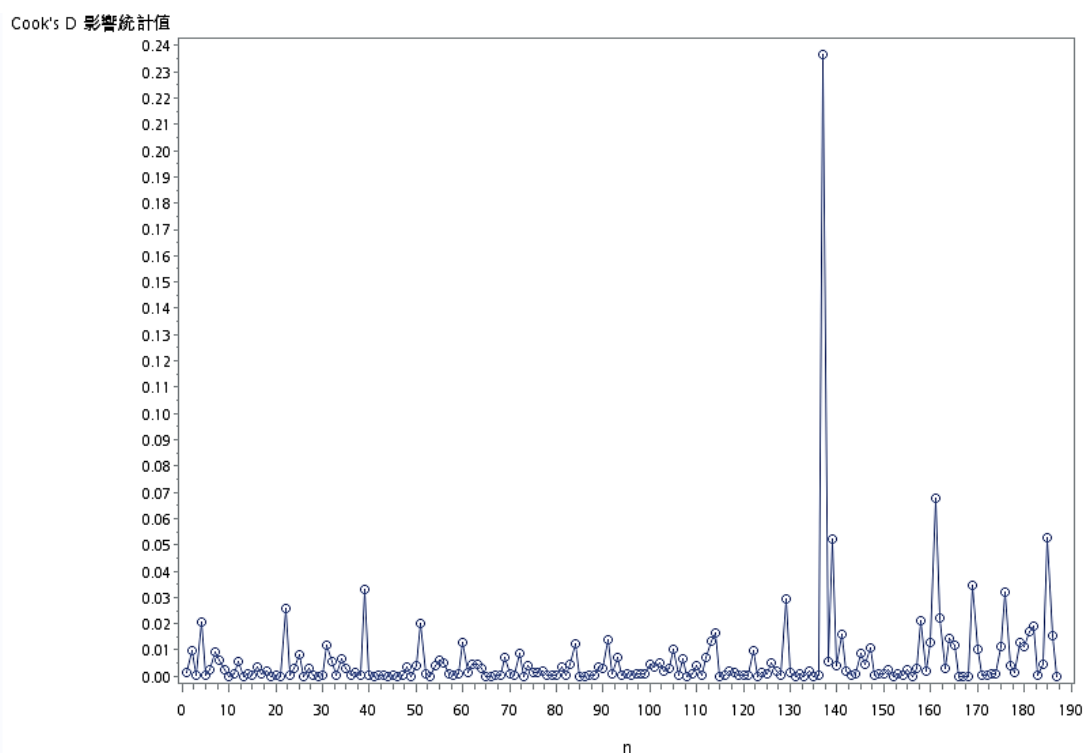


2. Cook 距離



最大的 Cook 距離 D 為 0.23632 在 $F(5,182)$ 相對應的位置於 5.3625 個百分位數值，所以其影響力也沒有大到需要矯正

n	Player	residual	hat	rstudent	dffits	d	dp
137	Butler, E	-0.98748	0.13814	-2.76429	-1.10667	0.23632	0.053925



3. DFBETAS 量數

發現所有觀察值對 4 種變數其 DFBETAS 量數絕對值皆小於 1，所以在此來看沒有任何變數是具有影響力的觀測值。

依據上述三種方法判斷，將保留所有離群值。

十四、 多重共線性診斷

參數估計值								
變數	標籤	自由度	參數估計值	標準誤差	t 值	Pr > t	標準化估計值	變異數膨脹
Intercept	Intercept	1	3.99989	0.04698	85.15	<.0001	0	0
sx2		1	0.42786	0.04202	10.18	<.0001	0.43365	2.13967
sx5		1	0.53624	0.04222	12.70	<.0001	0.54350	2.16069
sx2sx5		1	0.05992	0.02142	2.80	0.0057	0.08480	1.08423
x11	聯盟	1	0.13430	0.05923	2.27	0.0246	0.06819	1.06730

觀察各變數的 VIF，最大的 $VIF=2.16069 < 10$ ，平均值為 $1.612973 > 1$ ，代表迴歸係數膨脹了約 1.6 倍，但看似離 1 不遠所以共線性問題並不嚴重。

十五、 結論

球員	年齡	投球	ERA	進階 ERA	合約(美元)
David Price	30	左	2.45	3.07	7 年 2.17 億
Zack Greinke	32	右	1.66	1.96	6 年 2.065 億
Johnny Creto	29	右	3.44	3.48	6 年 1.3 億
Jorden Zimmermann	29	右	3.66	3.57	5 年 1.1 億
Jeff Samardzija	30	右	4.96	4.39	5 年 0.9 億
Scott Kazmir	31	左	3.1	3.55	3 年 0.48 億
陳偉殷	30	左	3.34	4.29	5 年 0.8 億 +1 年 0.16 億(選擇權)

最終選定模型 $Y = 3.99989 + 0.42786 SX_2 + 0.53624 SX_5 + 0.05992 SX_2 SX_5 + 0.1343 X_{11}$ ，由模型中發現，美國聯盟較國家聯盟的投手防禦率平均多出 0.1343，可能是國聯投手需要打擊所帶來的影響，另相較於 WHIP，優異的 SLG 較容易帶來比較好的成績，但是僅以 WHIP 以及 SLG 來推估一個投手的好壞可能還是略顯不足，仍需要將各項數據拉進來做綜觀的評價，此模型只是為了消彌單看傳統數據時所帶來的盲點。以下針對 2015 年球季結束後，已出現的投手合約作討論。

以新的模型來推估陳偉殷 2015 年的 ERA 值(以下稱進階 ERA)上升了 0.95，相較同於今年簽合約的投手來說，進階數值與 Samardzija 相近，也同為 30 歲，且不管是傳統數值或是進階數值皆優於 Samardzija，再來投手市場中左投相較於右投更為稀少，所以比起 Samardzija 的合約，陳偉殷的 5 年 0.8 億，相對超值；另外與 Zimmermann 以及 Kazmir 做比較，雖然陳偉殷的傳統數據優於 Zimmermann，但以進階數據來看，有著明顯的差距，故 Zimmermann 要到一只 5 年 1.1 億的合約，相當符合他的身價。但 Kazmir 不管在傳統數據或是進階數據皆相當優秀，僅簽下 3 年 0.48 億的薪資，其實相當讓人訝異。可能是對於 3 年後的自己極具信心，目標打破 Zack Greinke 的最高薪紀錄。