一、研究背景與動機

職棒球團的經營一直是個大學問，任何球團都想要獲利及永續經營，對於資金有限的球隊，如何在財大氣粗的豪門球團搶奪明星球員下與之競爭，如何更有效率的將資金花在刀口上，一直是許多人探討的話題。職業棒球運動是一項戰術頻繁、複雜且講求團隊分工合作的運動，經由比賽所產生的各項攻、守記錄，對球隊及球員比賽有著重要的參考價值和依據，更明確的說，這些攻、守記錄，隨著科技資訊的進步，經由電腦換算所產生的各項統計數據，對職業棒球場上比賽結果與球員調度有重要的參考價值，可用來檢討分析及評斷球隊的需要及選手價值高低，運用整體球隊成績去比較各隊之間的各項攻、守統計數據，就可以了解球隊的勝敗原因，找出致勝之要素，在比賽中提高勝率。

二、研究目的

本次研究以 2016 年美國職棒大聯盟 30 支隊伍球季比賽的資料，分析團隊的打擊、投手、守備數據與勝率之關係，找出影響勝率的關鍵因素。

三、變數介紹

1.應變數(Y)：勝率(Winning Percentage，WPCT)

意義：即投手必定會扛勝敗責任的前提下，投手為球隊獲得勝利的機率。
公式：勝場數／（勝場數＋敗場數）

2.自變數(X1)：打擊率(batting average，AVG)

意義：打者在有效的打席中，揮出安打的機率。
公式＝安打／打數

3.自變數(X2)：上壘率(on base percentage，OBP)

意義：打者上場打擊時，能夠上壘的機率。
公式：（安打＋四死球）／（打數＋四死球＋犧牲高飛）

4.自變數(X3)：整體攻擊指數(On-base Plus Slugging，OPS)

意義：上壘率和長打率的總和。
公式：上壘率＋長打率

5.自變數(X4)：防禦率(earned run average，ERA)

意義:假設投手投滿九局,守備員正常守備的情況下,投手平均會失掉幾分。
公式:(責失ｘ９)／投球局數

6.自變數(X5):奪三振率(Strikes outs per Nine Innings,SO/9)

意義:即假設投手投滿九局的情況下,平均會出現幾次三振。
公式:(三振數ｘ９)／投球局數

7.自變數(X6):每局被上壘率(Walks and Hits per Innings Pitched,WHIP)

意義:投手在一局的投球中,在正常的守備前提下,平均會讓幾名打者上壘。
公式:(安打＋四壞球)／投球局數

8.自變數(X7):被攻擊指數(Opponent On-base Plus Slugging,Oops)

意義:從 OPS 的觀念而來,Oops 愈高表示投手愈無法壓制打者的攻擊。
公式:被上壘率＋被長打率

9.自變數(X8):防守效率(Defensive Efficiency Rating,DER)

意義:即對方將球打入場內後(排除全壘打的情況,全壘打算投手的責任),能夠轉化為出局數的機率。
公式:１－〔(被安打－被全壘打＋因失誤上壘次數)／(面對之打席數－四死球－三振－被全壘打)〕

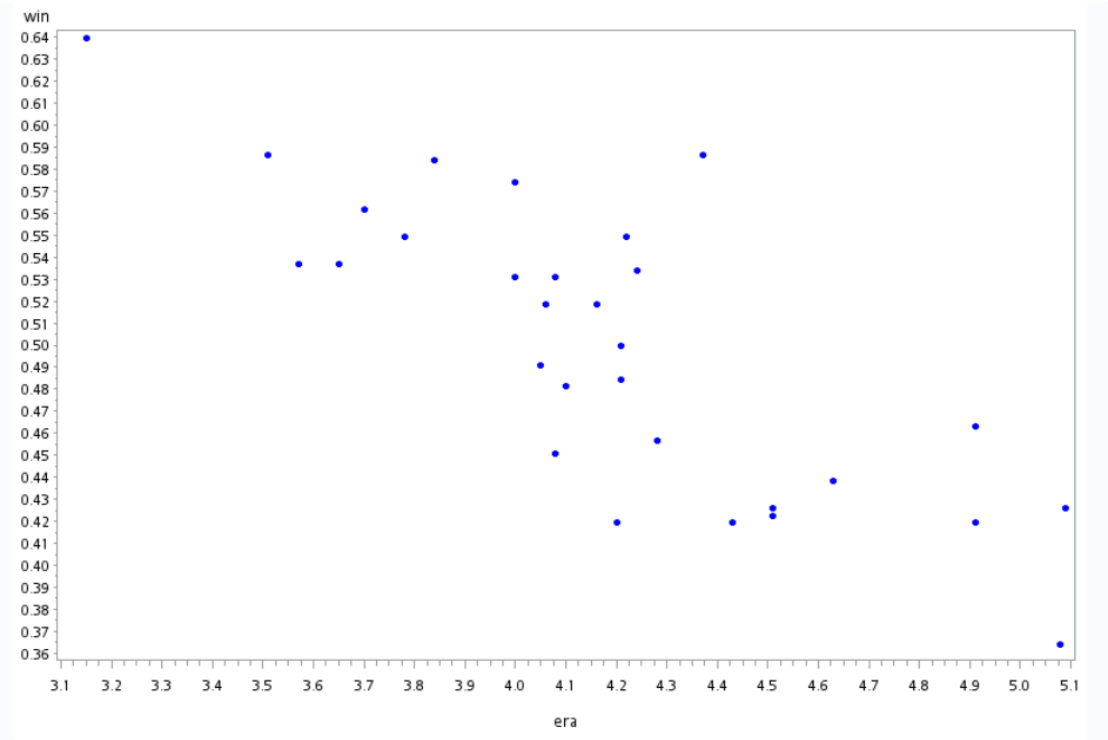
10.自變數(X9):聯盟

球隊所屬聯盟:美國聯盟=0;國家聯盟=1。

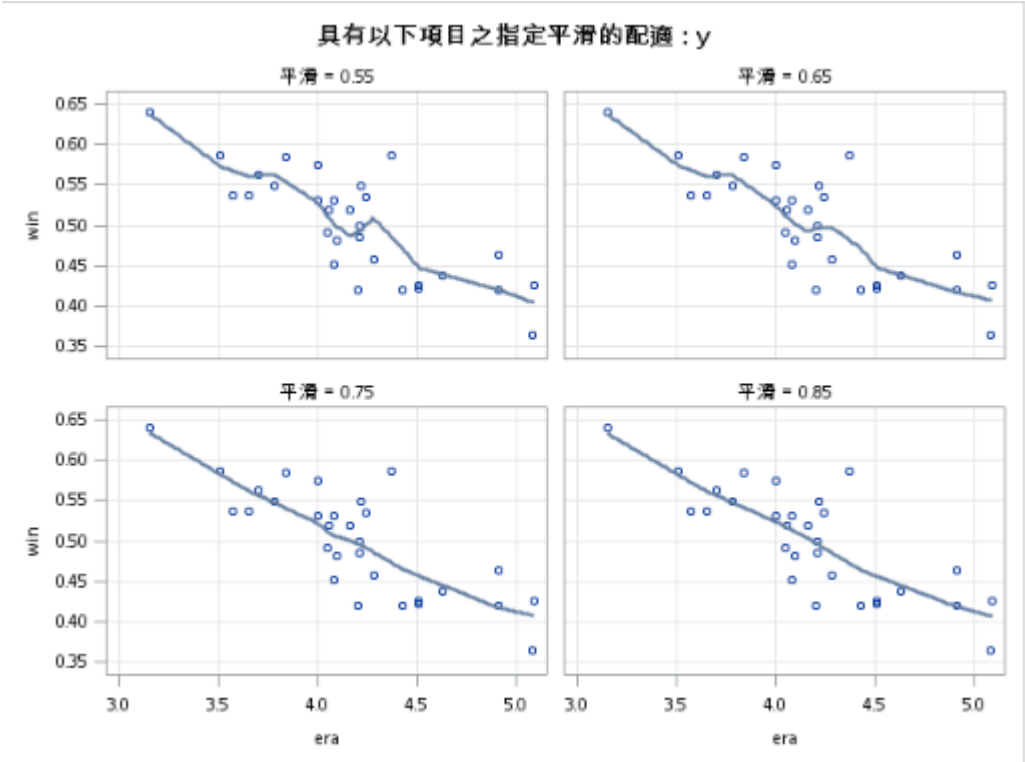
四、簡單線性迴歸

常看到許多教練在訪談中提到:棒球勝敗的關鍵,投手佔了七成。本次研究主題即是藉由迴歸模型探討防禦率與勝率之間是否具有線性關聯。

1. 觀察迴歸關係

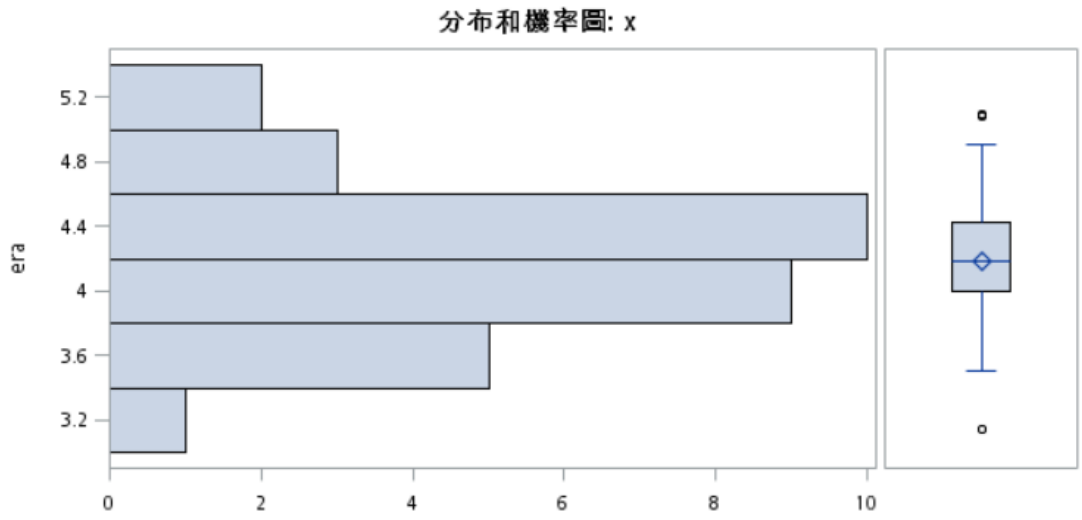


由散佈圖，防禦率與球隊勝率應為負向線性關係。

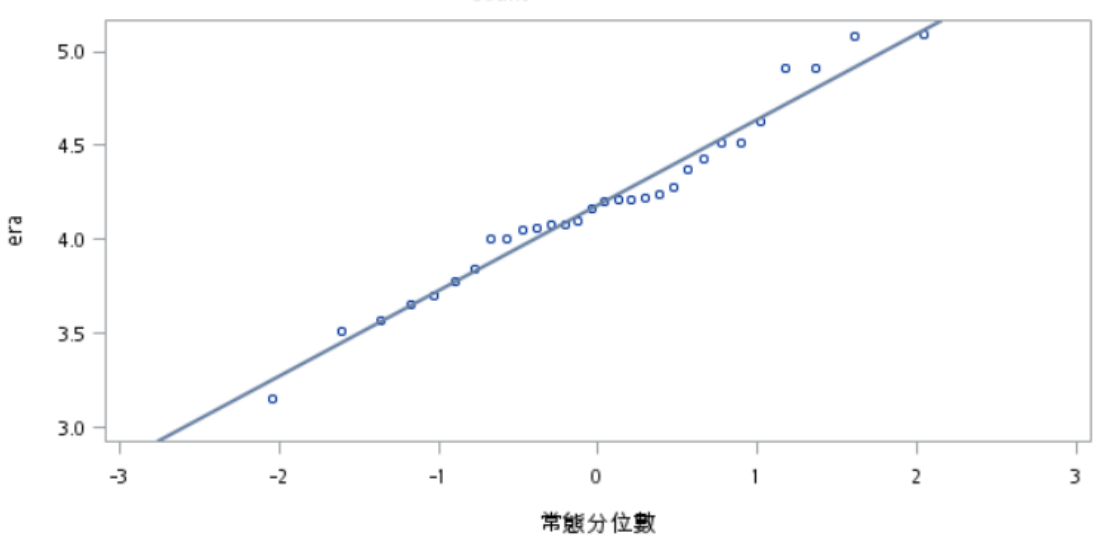


不同平滑係數之 lowess 曲線圖，呈負向線性相關。

2. X 變數(防禦率)的診斷



觀察直方圖與箱形圖，分布均勻近似常態，但有 3 個離群值，分別為第 8 筆的 5.08、第 22 筆的 3.15 及第 29 筆的 5.09，因 5.08、5.09 與第三高的觀察值差距小於一個標準差，故不刪除；3.15 與次低觀察值 3.51 差距小於一個標準差，因此也不將其刪除。



觀察 QQ Plot，自變數 X(防禦率)接近常態分配。

UNIVARIATE 程序
變數: x (era)

動差			
N	30	總和權重	30
平均值	4.18433333	總和觀測	125.53
標準差	0.45604509	變異數	0.20797713
偏態	0.14321771	峰度	0.20502911
未校正平方和	531.2907	校正平方和	6.03133667
變異係數	10.898871	標準誤差平均值	0.08326206

基本統計量值			
位置		變異性	
平均值	4.184333	標準差	0.45605
中位數	4.180000	變異數	0.20798
眾數	4.000000	全距	1.94000
		內四分位距	0.43000

偏態係數 0.1432>0，為右偏分配；峰度 0.205>0，為高峽峰。

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.970243	Pr < W	0.5458
Kolmogorov-Smirnov	D	0.118091	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.069574	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.40068	Pr > A-Sq	>0.2500

由 Shapiro-Wilk 常態性檢定(H_0:X 為常態分佈)，p-value=0.5458>0.05，不拒絕 H_0，表示 X 為常態分佈。

3. 迴歸結果

變異數分析

來源	自由度	平方和	平均值平方	F 值	Pr > F
模型	1	0.08078	0.08078	48.83	<.0001
誤差	28	0.04632	0.00165		
配適不足	23	0.04110	0.00179	1.71	0.2876
純誤差	5	0.00521	0.00104		
已校正的總計	29	0.12710			

根 MSE	0.04067	R 平方	0.6356
應變平均值	0.50003	調整 R 平方	0.6226
變異係數	8.13383		

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	0.98428	0.06969	14.12	<.0001	0.84152	1.12704
x	era	1	-0.11573	0.01656	-6.99	<.0001	-0.14965	-0.08181

藉由變異數分析方法做 F 檢定(H_0: β1 = 0)之結果顯示，檢定統計量F=48.83，p-value < 0.0001。故，在顯著水準 α= 0.05 下，拒絕 H_0。因此，認為反應變數與預測變數呈線性關係。

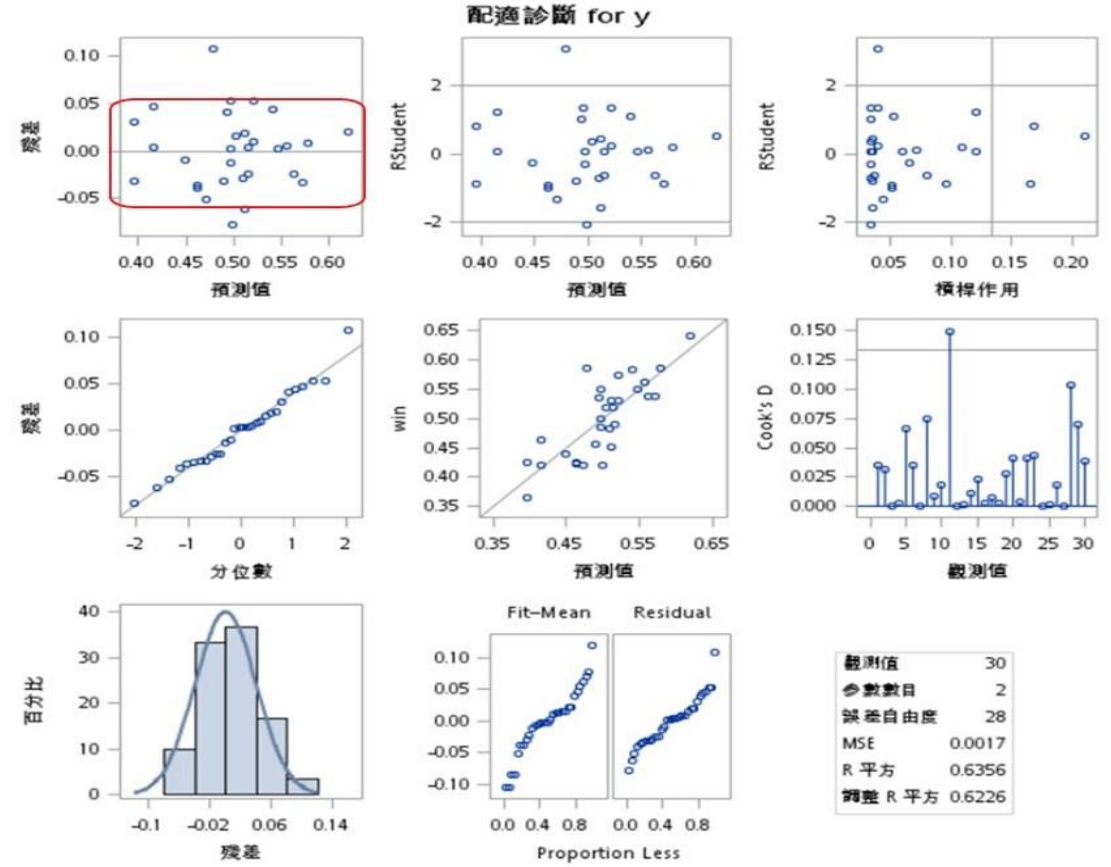
由配適不良檢定之結果顯示，檢定統計量F=1.71，p-value=0.2876 >0.05。因此，認為此迴歸模型的線性關係配適良好。

判定係數 R^2=0.6226，表示防禦率與勝率呈現中等程度的線性配適度。

由上面結果得知迴歸模型參數估計值：β_0=0.98428，β_1=-0.11573

線性迴歸方程式:$\hat{Y}$=0.98428-0.11573X。防禦率與勝率之間存在負向相關，表示當防禦率每增加 1，勝率即減少 0.11573。而 β_1 之 95% C.I.為 (-0.14390,-0.08756)。

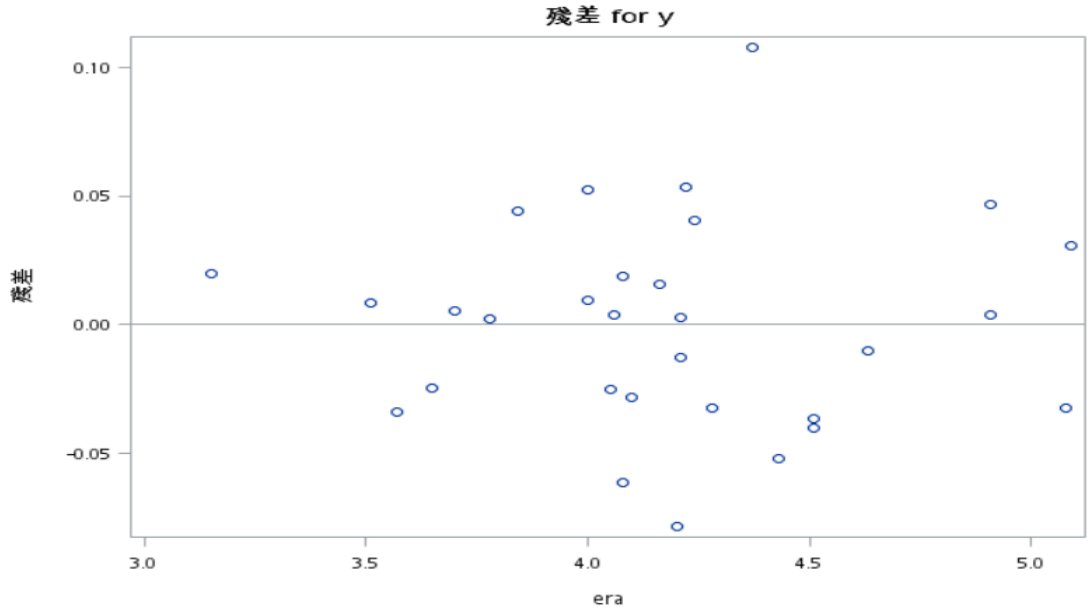
4. 殘差分析



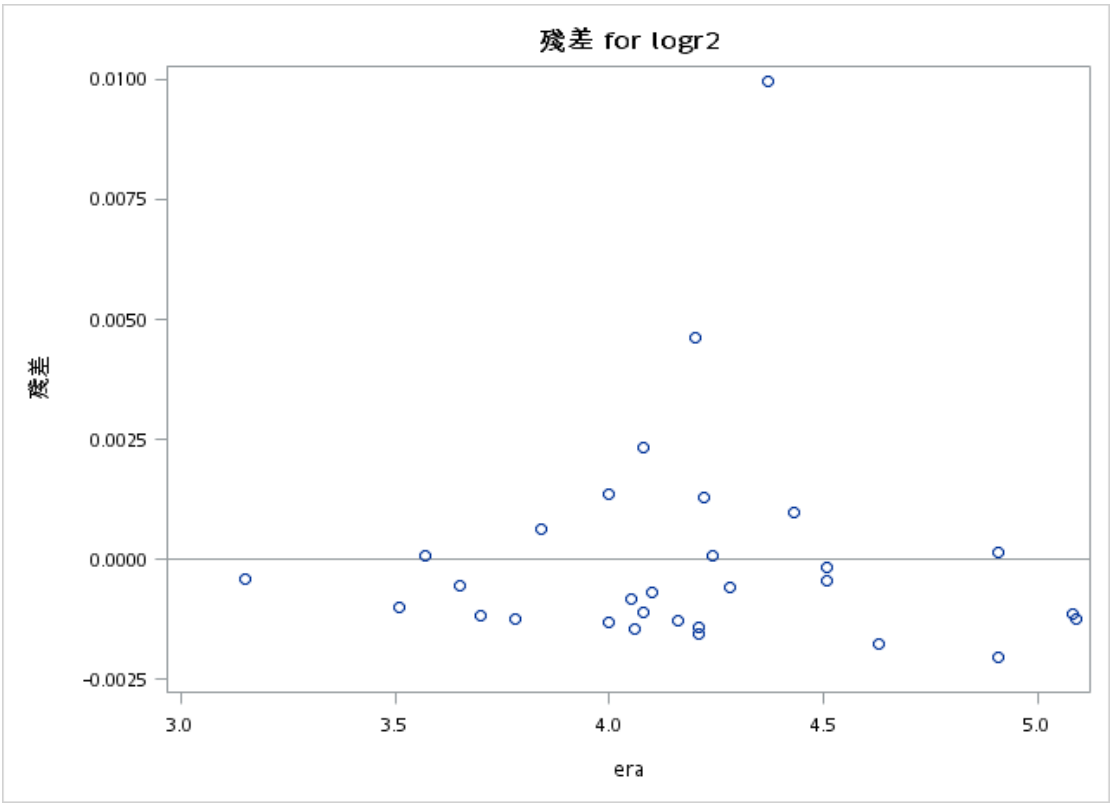
左上：殘差對預測值關係圖大致呈現對稱於 0。

中上：半學生化殘差圖，皆小於 4 個標準差，沒有離群值。

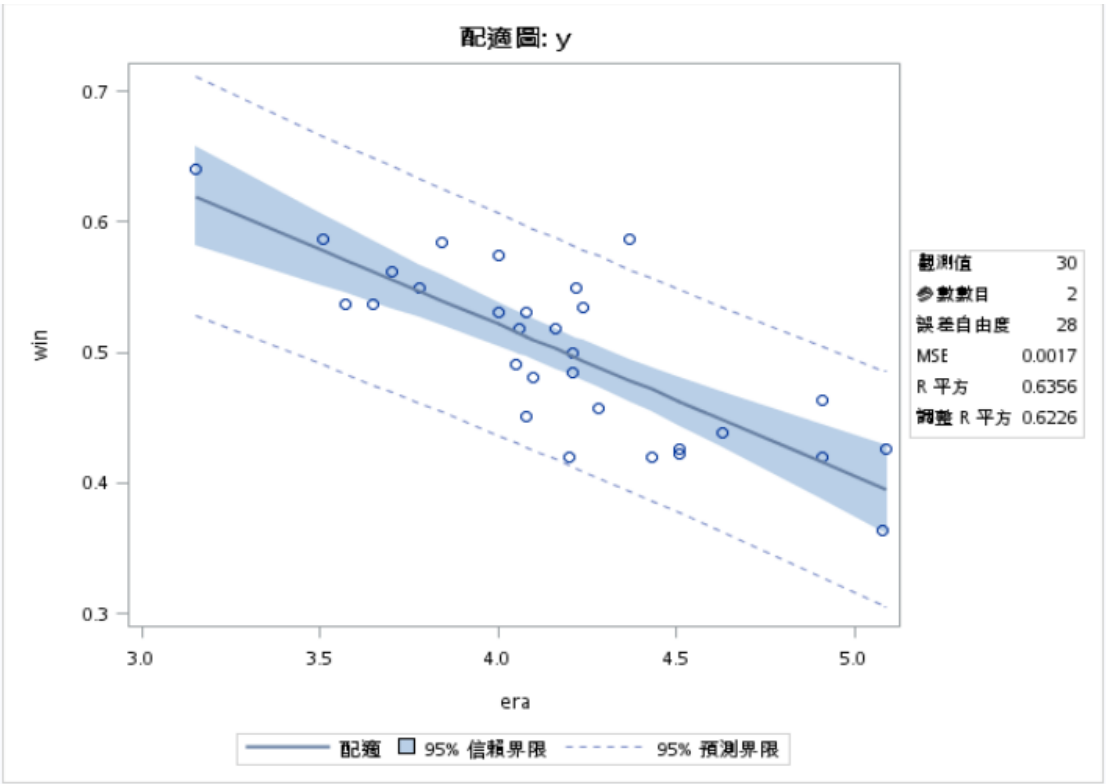
左中：QQ Plot 大多在對角線上，符合常態分配。



殘差對 X 的散佈圖無明顯趨勢，呈均勻分散。



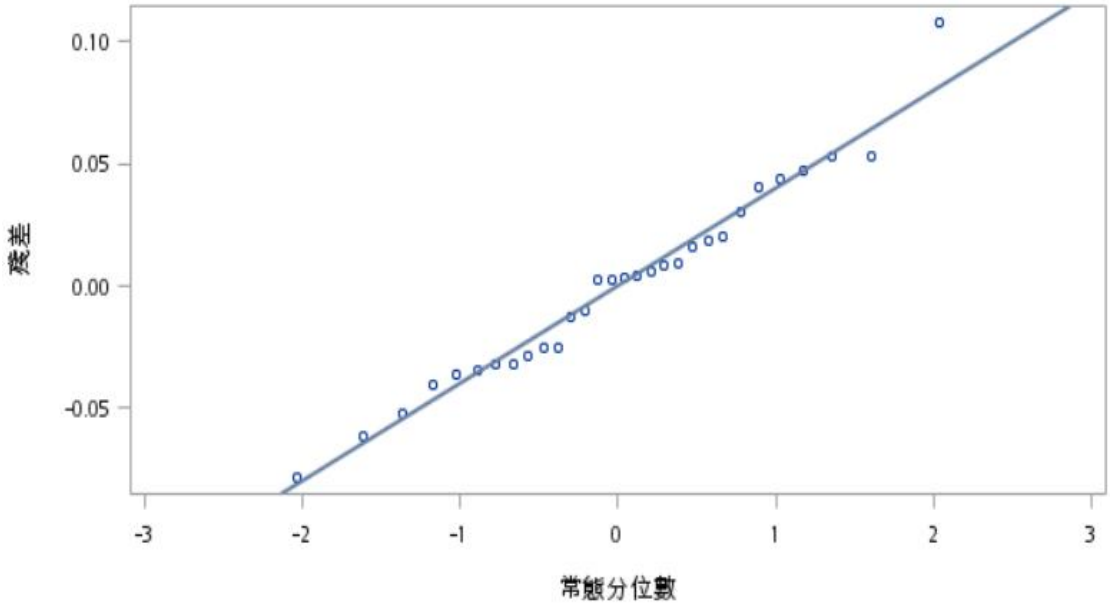
觀察殘差平方的散佈圖，並未隨 X 大小改變出現趨勢。



X 對 Y 的配適圖，只有一筆資料在預測線外圍，迴歸方程式配適性佳。

5. 殘差常態檢定

殘差的常態機率圖近似直線，表示殘差接近常態分配。



SAS 系統

CORR 程序

2 變數:	res expec

簡單統計值

變數	N	平均值	標準差	中位數	最小值	最大值	標籤
res	30	0	0.03996	0.00332	-0.07847	0.10787	殘差
expec	30	0	0.03957	0	-0.08298	0.08298	

Pearson 相關係數, N = 30
Prob > |r| (位於 H0 底下): Rho=0

	res	expec
res 殘差	1.00000	0.98547 <.0001
expec	0.98547 <.0001	1.00000

Spearman 相關係數, N = 30
Prob > |r| (位於 H0 底下): Rho=0

	res	expec
res 殘差	1.00000	1.00000 <.0001
expec	1.00000 <.0001	1.00000

由 Pearson 及 SPearman 相關係數可知殘差與 $\hat{y}$呈高度相關。

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.976402	Pr < W	0.7240
Kolmogorov-Smirnov	D	0.099489	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.041267	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.267994	Pr > A-Sq	>0.2500

由 Shapiro-Wilk 常態性檢定(H0: 殘差為常態分佈)，p-value=0.7240>0.05，不拒絕 H0，表示殘差為常態分佈。

6. 殘差變異數常數檢定

SAS 系統

TTEST 程序

變數: d

group	N	平均值	標準差	標準誤差	最小值	最大值
1	15	0.0224	0.0203	0.00524	0	0.0671
2	15	0.0380	0.0307	0.00792	0	0.1181
Diff (1-2)		-0.0156	0.0260	0.00950		

group	方法	平均值	95% CL 平均值		標準差	95% CL 標準差	
1		0.0224	0.0111	0.0336	0.0203	0.0149	0.0320
2		0.0380	0.0210	0.0550	0.0307	0.0224	0.0484
Diff (1-2)	集區	-0.0156	-0.0350	0.00386	0.0260	0.0206	0.0352
Diff (1-2)	Satterthwaite	-0.0156	-0.0352	0.00399			

方法	變異數	自由度	t 值	Pr > |t|
集區	均等	28	-1.64	0.1118
Satterthwaite	不均等	24.3	-1.64	0.1135

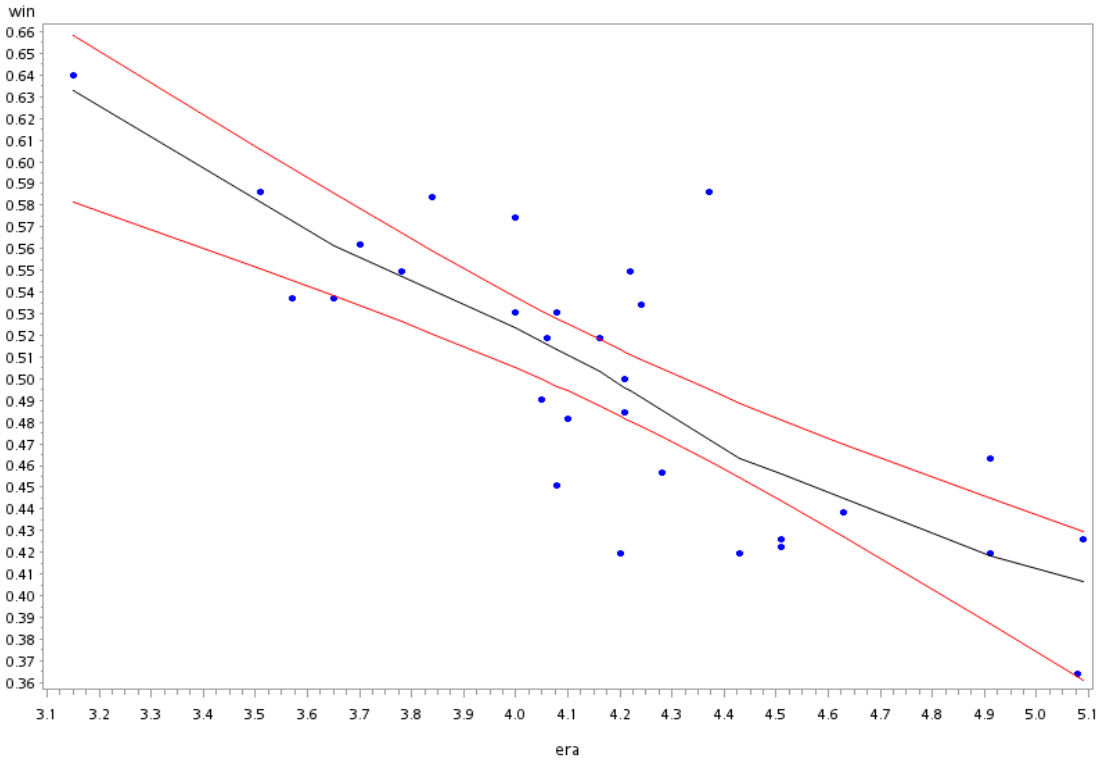
變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	14	14	2.28	0.1351

對殘差作 Brown-Forsythe 檢定，檢定值 t=-2.28，p-value=0.1351>0.05，故殘差有變異數齊一性，殘差變異數為常數。

Obs	ssrs	sse	nobs	tests	pv
1	.000003057	0.046317	30	0.64123	0.57673

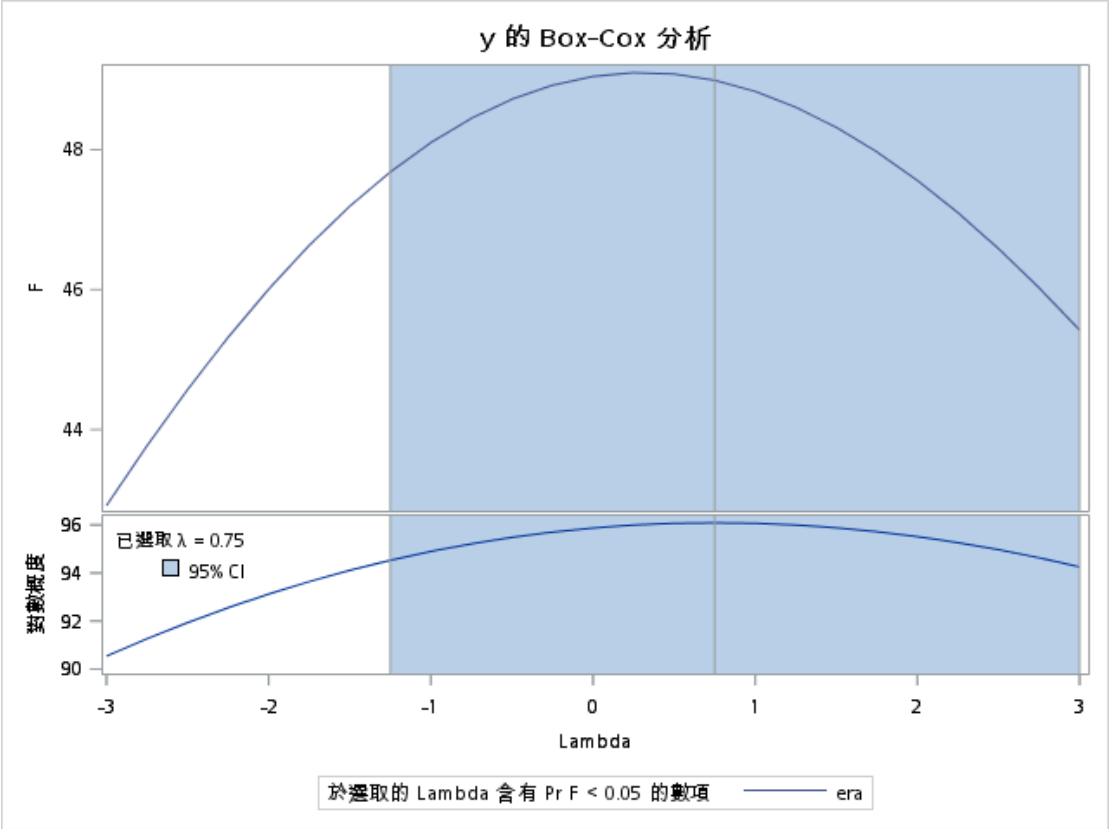
對殘差作 Breusch-Pagan 檢定，檢定統計量為 0.64123，p-value = 1-0.57673=0.42327>0.05，不拒絕 H0 為常數變異數的假設，殘差變異數為常數。

7. 預測線及信賴區間

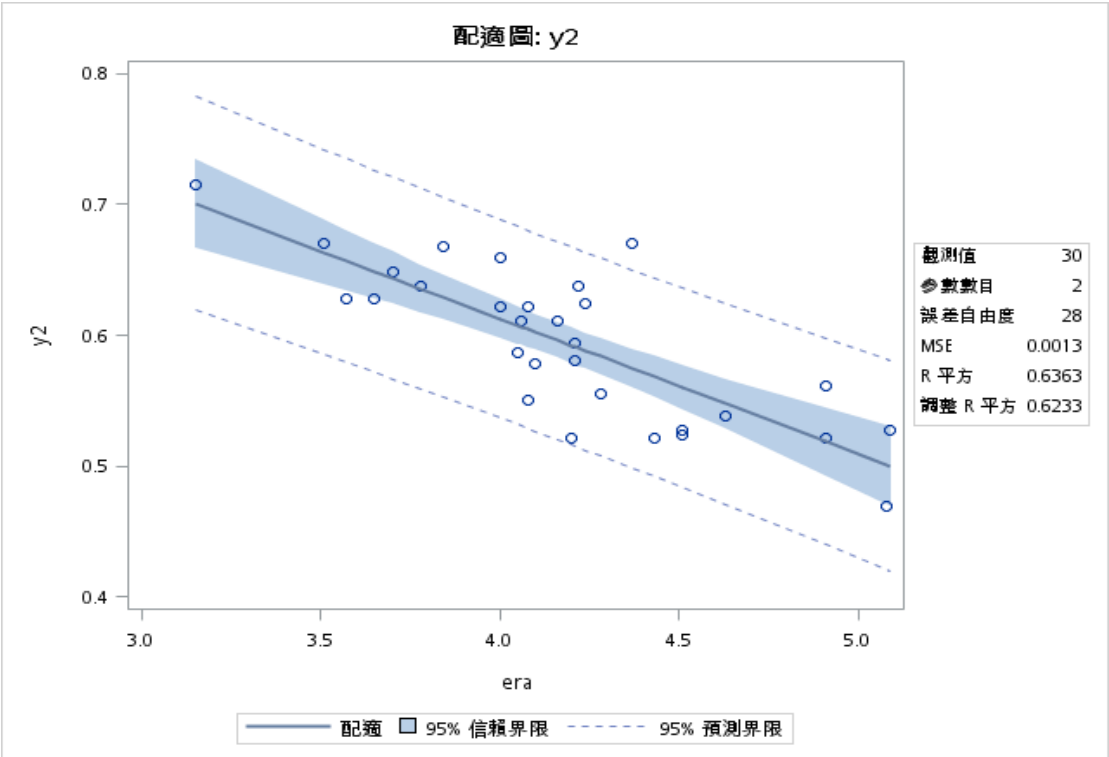


預測線包含在在區間帶內，表示此模型配置適當。

8. BOX-COX 轉換



由 Box-Cox 轉換可得最佳 λ=0.75。



X 對 Y2 的配適圖，只有一筆資料在預測線外圍，迴歸方程式配適性佳。

變異數分析					
來源	自由度	平方和	平均值 平方	F 值	Pr > F
模型	1	0.06465	0.06465	48.99	<.0001
誤差	28	0.03695	0.00132		
配適不足	23	0.03276	0.00142	1.70	0.2908
純誤差	5	0.00419	0.00083731		
已校正的總計	29	0.10159			

根 MSE	0.03633	R 平方	0.6363
應變平均值	0.59368	調整 R 平方	0.6233
變異係數	6.11867		

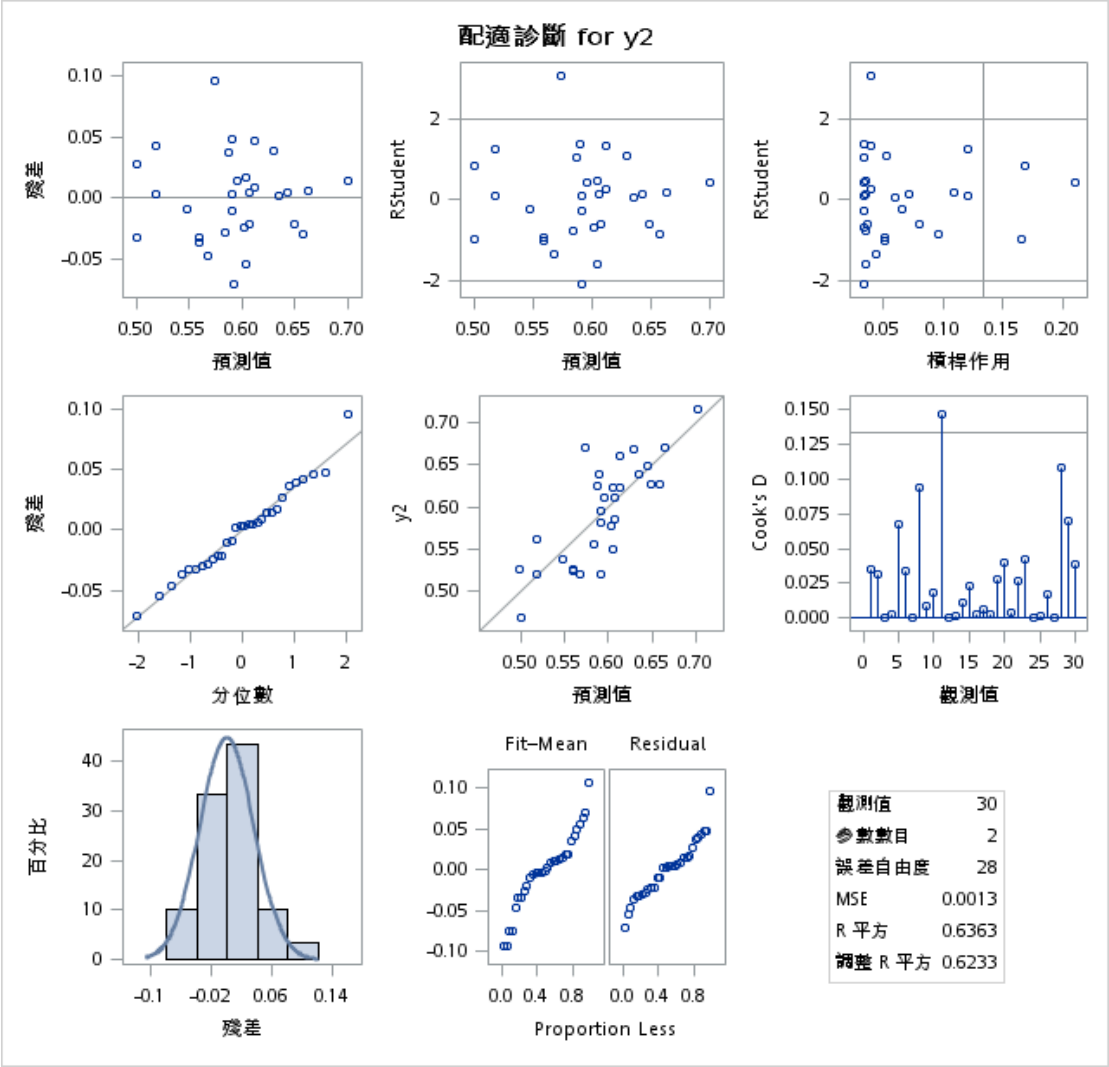
參數估計值						
變數	標籤	自由度	參數 估計值	標準 誤差	t 值	Pr > |t|
Intercept	Intercept	1	1.02688	0.06225	16.50	<.0001
x	era	1	-0.10353	0.01479	-7.00	<.0001

F 檢定(H0: β1 = 0)之結果顯示，檢定統計量F=48.99，p-value < 0.0001。故，在顯著水準 α= 0.05 下，拒絕 H0。因此，有充分證據認為反應變數與預測變數呈線性關係。

由配適度檢定之結果顯示，檢定統計量F=1.70，p-value=0.2908>0.05。故，在顯著水準 α= 0.05 下，不拒絕 H0。因此，認為此迴歸模型的線性關係配適良好。

判定係數，R2=0.6233，表示防禦率與勝率呈現中等程度的線性配適度。

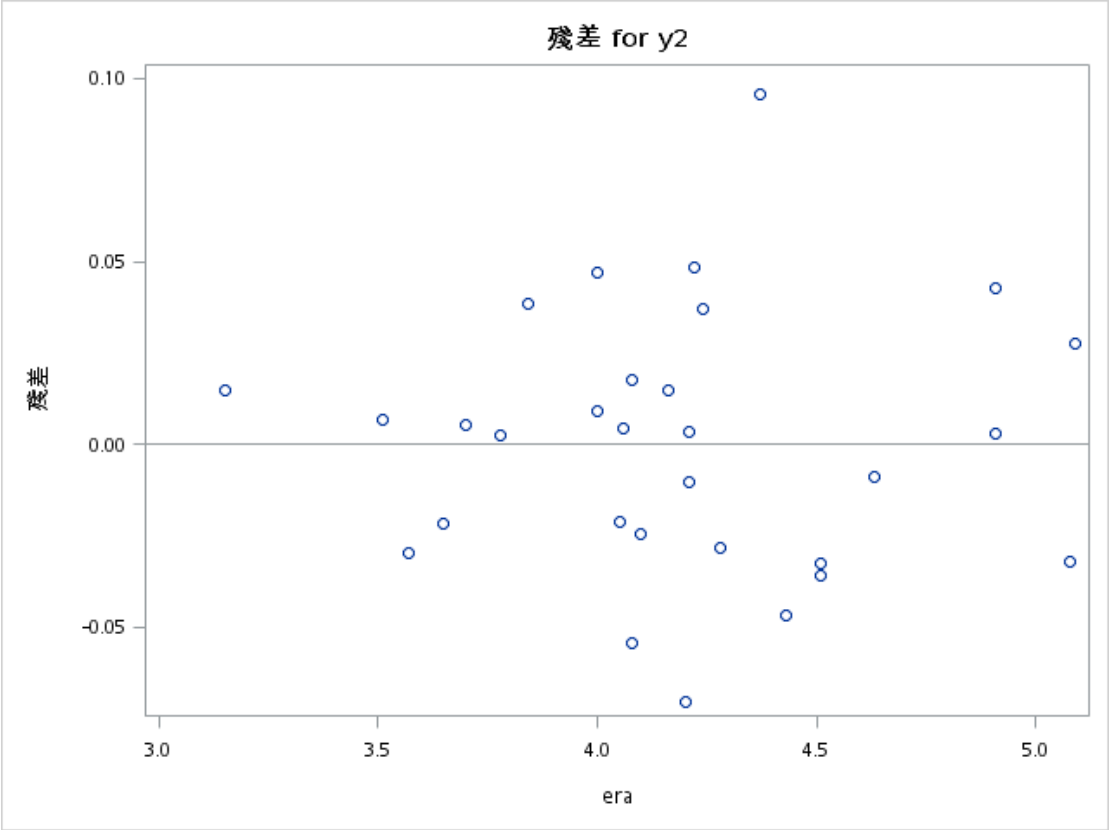
線性迴歸方程式：$\hat{Y}$=1.02688-0.10353X。



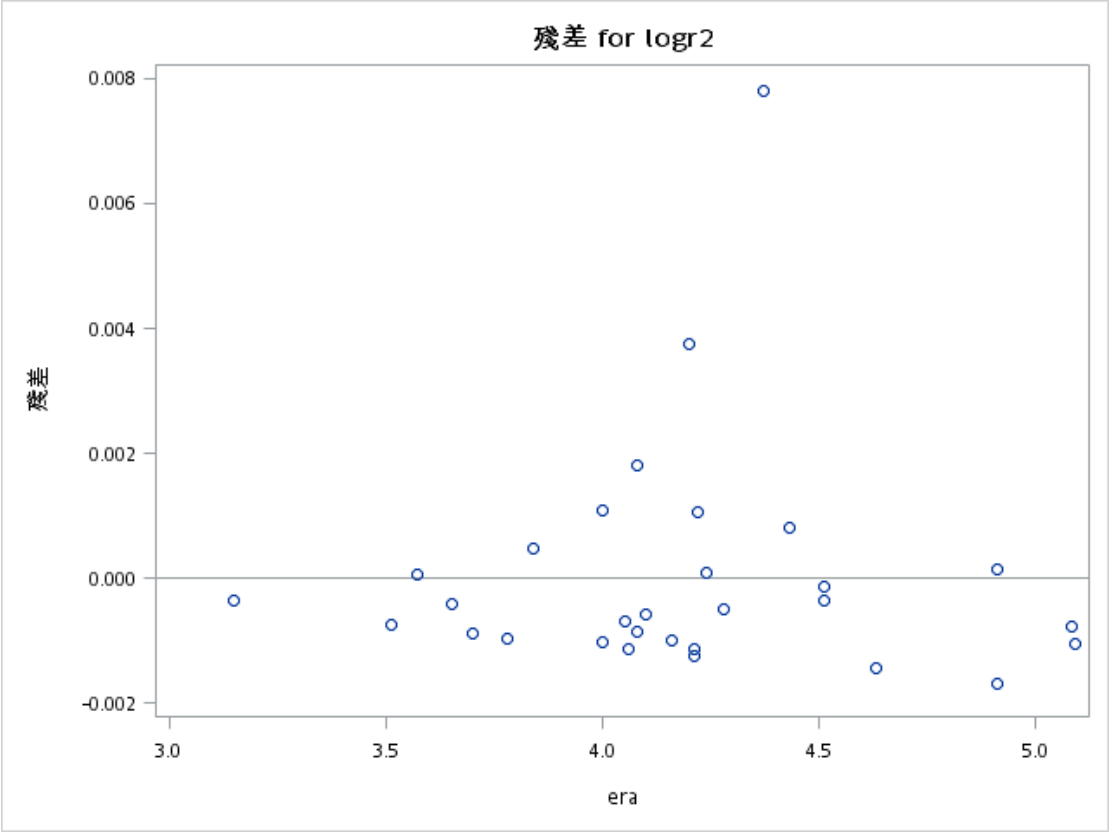
左上：殘差對預測值關係圖大致呈現對稱於 0。
中上：半學生化殘差圖，皆小於 4 個標準差，沒有離群值。
左中：QQ Plot 大多在對角線上，符合常態分配。

常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.977596	Pr < W	0.7587
Kolmogorov-Smirnov	D	0.09556	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.040969	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.261273	Pr > A-Sq	>0.2500

由 Shapiro-Wilk 常態性檢定(H0: 殘差為常態分佈)，p-value=0.7240>0.05，不拒絕 H0，表示殘差為常態分佈。



殘差對 x 的散佈圖無明顯趨勢，呈均勻分散。



觀察殘差平方的散佈圖，並未隨 X 大小改變出現趨勢。

TTEST 程序

變數: d2

group	N	平均值	標準差	標準誤差	最小值	最大值
1	15	0.0195	0.0180	0.00466	0	0.0595
2	15	0.0343	0.0273	0.00704	0	0.1046
Diff (1-2)		-0.0147	0.0231	0.00844		

group	方法	平均值	95% CL 平均值		標準差	95% CL 標準差	
1		0.0195	0.00952	0.0295	0.0180	0.0132	0.0285
2		0.0343	0.0192	0.0494	0.0273	0.0200	0.0430
Diff (1-2)	集區	-0.0147	-0.0320	0.00254	0.0231	0.0183	0.0313
Diff (1-2)	Satterthwaite	-0.0147	-0.0322	0.00266			

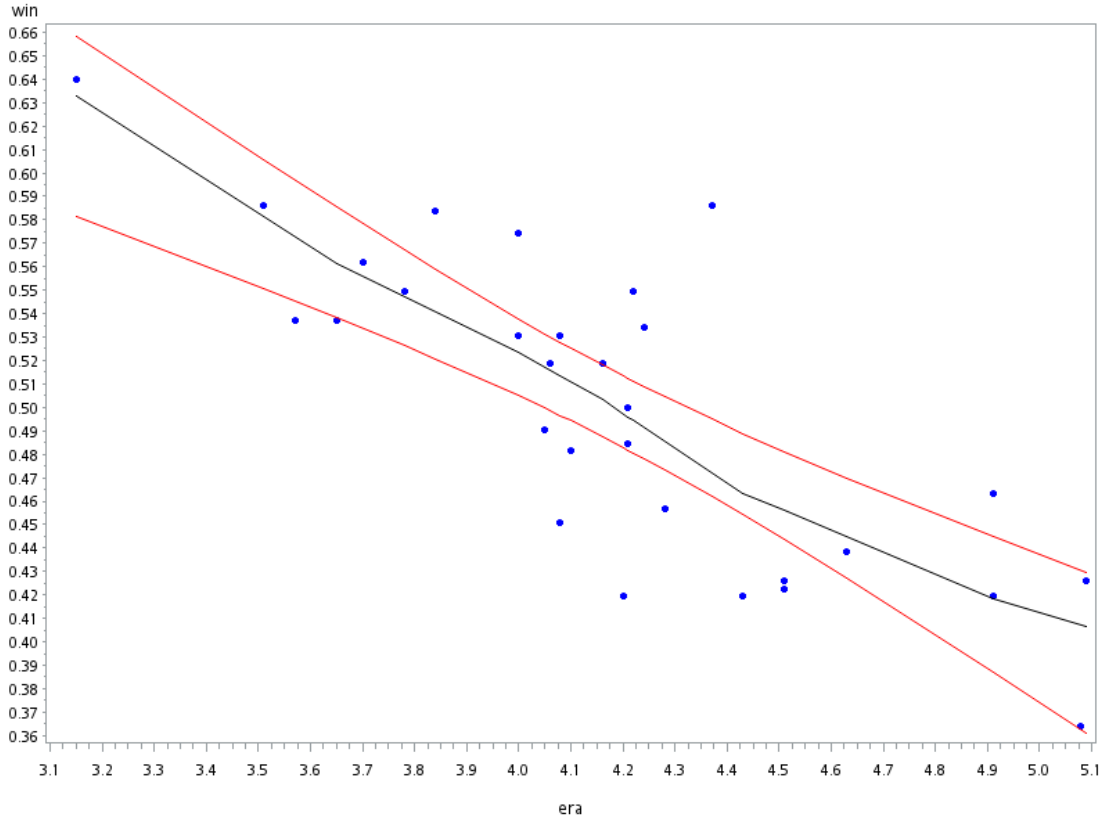
方法	變異數	自由度	t 值	Pr > |t|
集區	均等	28	-1.75	0.0915
Satterthwaite	不均等	24.299	-1.75	0.0932

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	14	14	2.28	0.1351

對殘差作 Brown-Forsythe 檢定，檢定值 t=-2.28，p-value=0.1351>0.05，故殘差有變異數齊一性，殘差變異數為常數。

Obs	ssrs	sse	nobs	tests	pv
1	.000002434	0.036947	30	0.80244	0.62964

對殘差作 Breusch-Pagan 檢定，檢定統計量為 p-value=1-0.62964=0.37036>0.05，不拒絕 H0 為常數變異數的假設，殘差變異數為常數。



lowess 曲線圖落在新配適迴歸模型之信賴區間帶中，表示此模型配置適當。

9. 比較

	變數轉換前	變數轉換後
迴歸方程式	$\hat{Y}$=0.98428-0.11573X	$\hat{Y}$=1.02688-0.10353X
F 檢定	拒絕 H0	拒絕 H0
Lack of fit test	拒絕 H0	拒絕 H0
判定係數 R2	0.6226	0.6233
Shapiro-Wilk 常態性檢定	不拒絕 H0，殘差為常態分佈	不拒絕 H0，殘差為常態分佈
Brown-Forsythe 檢定	不拒絕 H0，殘差變異數為常數	不拒絕 H0，殘差變異數為常數
Breusch-Pagan 檢定	不拒絕 H0，殘差變異數為常數	不拒絕 H0，殘差變異數為常數