應用線性統計模型 期末報告

應數三　409120627　周家鴻

大綱

一、 簡介.....2
二、 初步資料調查.....2
三、 拆分資料.....5
四、 模型訓練.....5
五、 模型驗證.....7
六、 模型結果.....8
七、 結論.....9

一、 簡介

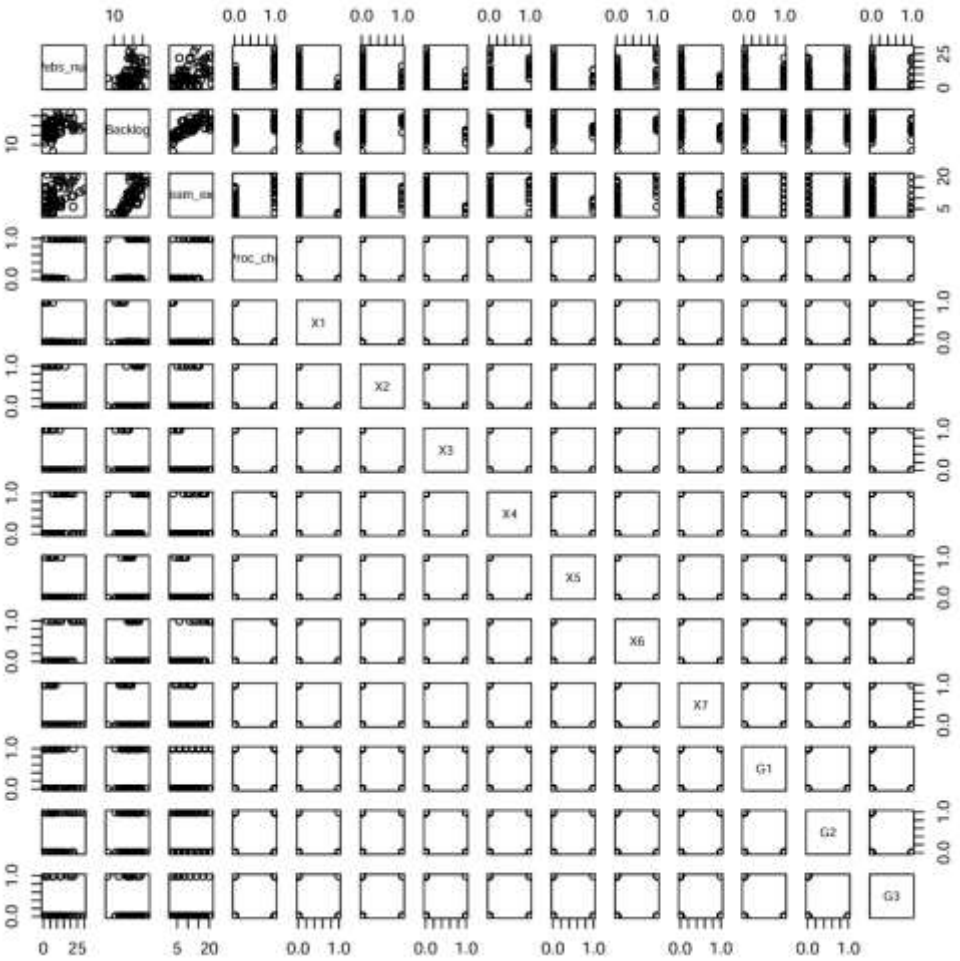
網頁設計公司想了解甚麼因素與專案成交量(Websites delivered)有關，想要尋找對專案成交量影響最顯著的因素。考慮以下這些變數:Backlog of order(季末訂單數量)、Team number(開發小組編號)、Team experience(開發經驗)、Process change(製程更新)、Year、Quarter。

其中，Team number(開發小組編號)、Year、Quarter 為類別資料。首先 Year 與 Quarter 皆代表時間，筆者將其合併為一項資料:
(2001,Q1),(2001,Q2),….,(2002,Q3),(2002,Q4)等總共八個類別，然後為簡化流程，將 13 個開發小組重新分組，1~5 為一組，6~10 為一組，11~13 為一組，總共三組。

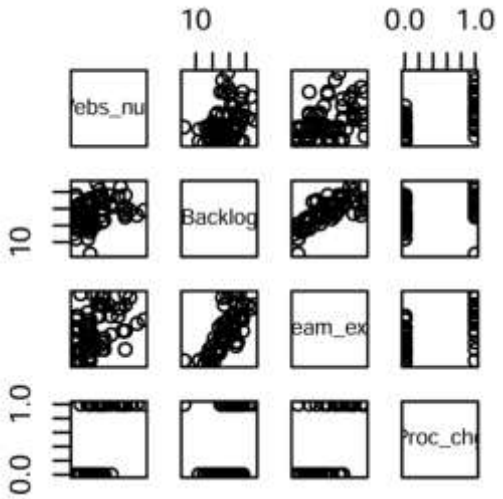
模型欲預測之 response 為專案成交量(Websites delivered)。

二、 初步資料調查

資料散佈圖:



其中 X1~X7 代表 year 與 quarter 合併之類別資料，G1~G3 代表組別之類別資料。因為指標變數不具有實際的意義，所以我們更關心左上角的部分:



此為專案成交量(Websites delivered)、Backlog of order(季末訂單數量)、及 Team experience(開發經驗)之散佈圖。看起來有些凌亂，不過在 Team experience(開發經驗)與 Backlog of order(季末訂單數量)之間似乎存在線性相關性。

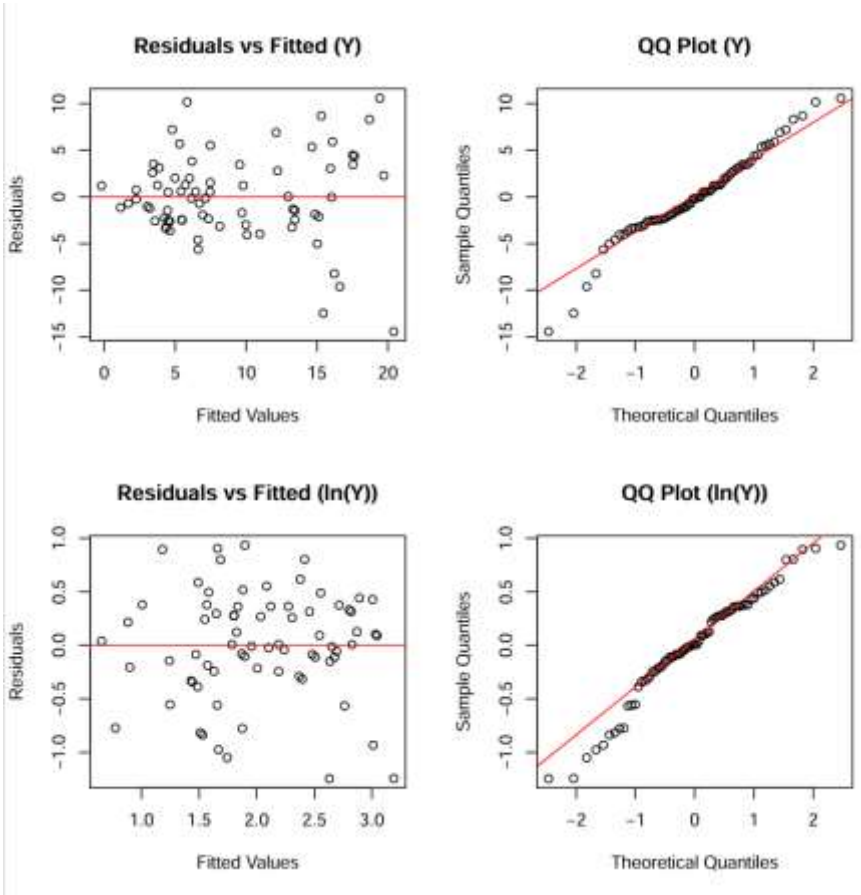
相關係數:

A	B	C	D	E	F	G
Webs_num	Backlog	Team_exp	Proc_chg	X1	X2	X3
1	0.36516646	0.44571815	0.68692593	-0.31933766	-0.14444544	-0.22022249
0.36516646	1	0.76343932	0.43562865	-0.43316905	0.34591995	-0.27444813
0.44571815	0.76343932	1	0.63533973	-0.47103528	0.09731159	-0.34145648
0.68692593	0.43562865	0.63533973	1	-0.24222268	-0.32988567	-0.26093123
-0.31933766	-0.43316905	-0.47103528	-0.24222268	1	-0.14444508	-0.11425241
-0.14444544	0.34591995	0.09731159	-0.32988567	-0.14444508	1	-0.15560159
-0.22022249	-0.27444813	-0.34145648	-0.26093123	-0.11425241	-0.15560159	1
0.33828714	0.40823396	0.27347848	0.62583278	-0.15159089	-0.20645327	-0.16329932
-0.0791864	-0.13358935	-0.19737983	-0.27891316	-0.12212605	-0.16632479	-0.1315587
0.52151439	0.13702742	0.52175438	0.62583278	-0.15159089	-0.20645327	-0.16329932
-0.24195495	-0.22256132	-0.08402858	-0.31328402	-0.13717582	-0.18682122	-0.14777086
-0.23557092	0.10465256	0.19634322	-0.14119232	0.15308563	-0.05574184	0.10220954
0.21930161	-0.0121695	-0.10213927	0.04418985	-0.14506743	0.05603827	-0.07777758
0.03418614	-0.12534165	-0.13268943	0.13322306	-0.01891543	0.00273224	-0.03727955

X4	X5	X6	X7	G1	G2	G3
0.33828714	-0.0791864	0.52151439	-0.24195495	-0.23557092	0.21930161	0.03418614
0.40823396	-0.13358935	0.13702742	-0.22256132	0.10465256	-0.0121695	-0.12534165
0.27347848	-0.19737983	0.52175438	-0.08402858	0.19634322	-0.10213927	-0.13268943
0.62583278	-0.27891316	0.62583278	-0.31328402	-0.14119232	0.04418985	0.13322306
-0.15159089	-0.12212605	-0.15159089	-0.13717582	0.15308563	-0.14506743	-0.01891543
-0.20645327	-0.16632479	-0.20645327	-0.18682122	-0.05574184	0.05603827	0.00273224
-0.16329932	-0.1315587	-0.16329932	-0.14777086	0.10220954	-0.07777758	-0.03727955
1	-0.174553	-0.21666667	-0.19606341	-0.08836278	0.02765546	0.08337536
-0.174553	1	-0.174553	-0.15795441	0.05712589	-0.01787905	-0.05390155
-0.21666667	-0.174553	1	-0.19606341	-0.08836278	0.02765546	0.08337536
-0.19606341	-0.15795441	-0.19606341	1	-0.02099965	0.0865365	-0.08350342
-0.08836278	0.05712589	-0.08836278	-0.02099965	1	-0.71380563	-0.42566499
0.02765546	-0.01787905	0.02765546	0.0865365	-0.71380563	1	-0.32988567
0.08337536	-0.05390155	0.08337536	-0.08350342	-0.42566499	-0.32988567	1

從相關係數中，我們可以看到專案成交量(Websites delivered)與 Process change(製程更新)有較強的線性關係，且 Backlog of order(季末訂單數量)與 Team experience(開發經驗)也有較高的線性關係

接著，考慮所有變數形成的一階線性模型。首先是該模型之殘差圖與 QQplot:



殘差圖中，並沒有看到明顯的曲線關係，因此考慮一階線性模型即可。這裡考慮了兩個 response，Y 與 ln(Y)。首先在以 Y 為 response 的殘差圖中，Y 愈大，殘差的變異數似乎有愈大的傾向。在以 lnY 為 response 的殘差圖中，並沒有在殘差中發現明顯的規律，似乎更能夠支持 constant variance 的假設。另外在 QQ plot 中，ln(Y)似乎更加理想。所以我們應當考慮以 ln(Y)作為 response，才更符合常態假設。

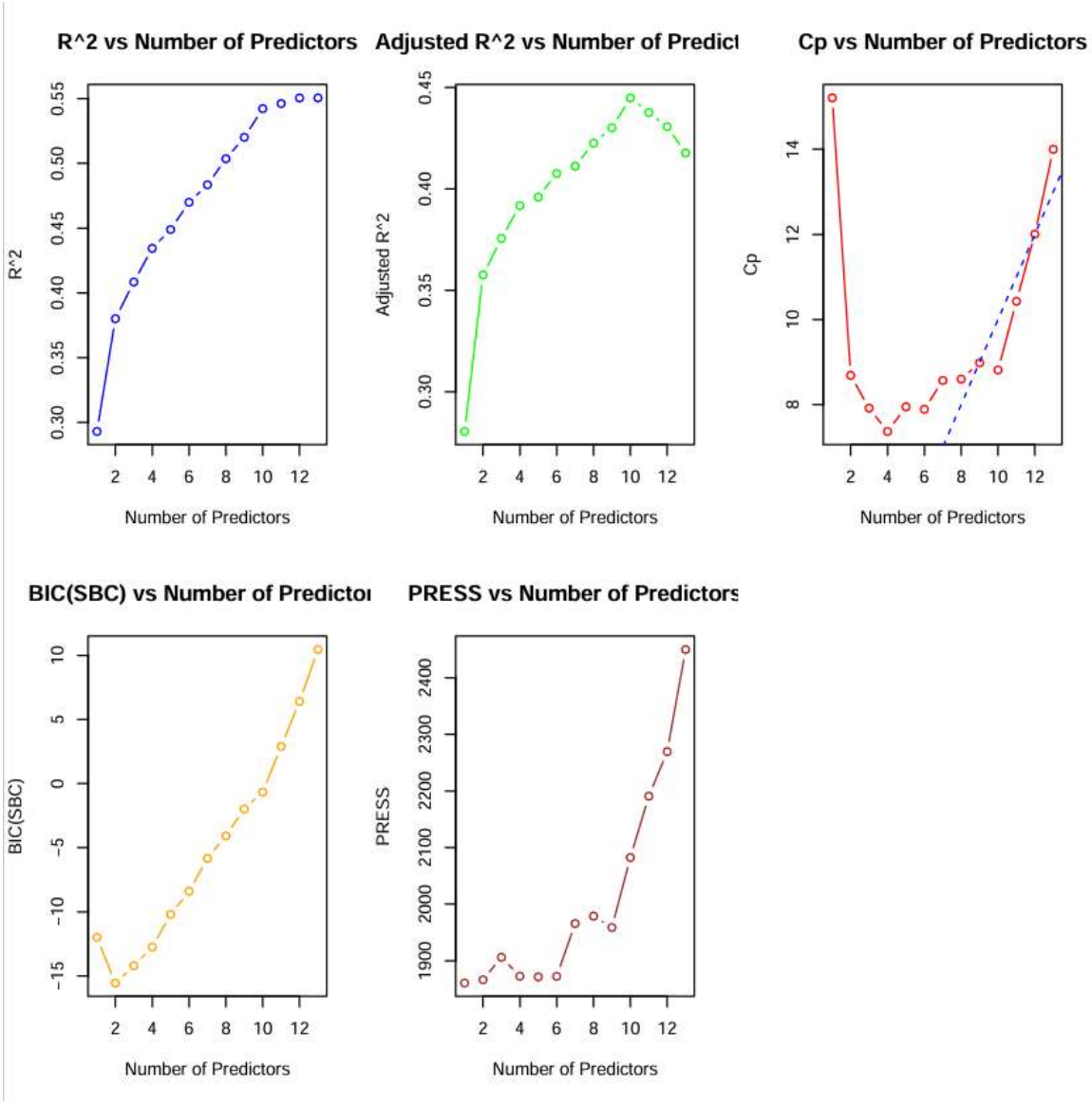
另外，因為 Backlog of order(季末訂單數量)與 Team experience(開發經驗)之間似乎存在線性關係，於是加入交互項; Backlog of order(季末訂單數量)與 Team number(開發小組編號) 之間可能存在關係，因此加入交互項。

三、 拆分資料

因為資料量太少，所以應該用 k-fold cross validation 方法來找出適合的模型。(但由於篇幅有限，這裡只做了一次的訓練及驗證)。利用 scikit-learn 將資料隨機的分為測試集與訓練集，其中測試集佔了全部資料的 20%，約 15 筆資料。

四、 模型訓練

因為 predictor 數量相對少，這裡使用 best subset 方法找出最適合的模型。



Number_of_Predictors	SSEp	Rsq	Adjusted_Rsq	Cp	SBC	PRESS_P
1	21.93762695	0.293090888	0.280467511	15.20758315	-11.99659818	1860.578164
2	19.23619752	0.380140644	0.357600304	8.685266586	-15.55790595	1866.422145
3	18.35856229	0.408421203	0.375555714	7.916552644	-14.2059305	1906.105673
4	17.55116237	0.434438527	0.391754643	7.369413102	-12.75408944	1872.644801
5	17.10026463	0.448968072	0.395984232	7.946946449	-10.2031684	1871.488491
6	16.44833712	0.469975517	0.407619695	7.890282471	-8.397159486	1872.40497
7	16.02910378	0.48348472	0.411172581	8.567708875	-5.834179554	1965.688009
8	15.40458451	0.503608974	0.422565541	8.597510849	-4.078708487	1978.808373
9	14.8920684	0.520124083	0.430147349	8.980654034	-1.980777361	1958.669749
10	14.20509571	0.54226081	0.444869493	8.813431508	-0.659558104	2082.399605
11	14.08281885	0.546201009	0.437683859	10.42767938	2.89946213	2190.767933
12	13.94915207	0.550508233	0.430643762	12.00599498	6.406770137	2269.474681
13	13.94725176	0.550569468	0.417783174	14	10.45931121	2449.85782

當 p=4,5,10,11 時看起來較適合。模型使用的變數如下:

(Intercept)	Backlog_Year	Backlog_G1	Backlog_G2	Backlog	Team_exp	Proc_chg	X1	X2	X3	X4	X5	G1	G2
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

五、 模型驗證

首先先以測試集的資料用 best subset algo 所決定出來的 predictor 再 fit 一次模型，並與訓練集資料的模型進行比較。

Estimate(train	Pr(>|t|)(train	Estimate(test	Pr(>|t|)(test
p=4	p=4	p=4	p=4
1.923487204	1.68518E-23	1.464237711	3.03944E-06
0.671556852	0.00038735	1.481407659	3.01416E-05
-0.87198813	0.003330427	-0.771090531	0.04026746
-0.377106634	0.11395278	NAN	NAN
p=5	p=5	p=5	p=5
1.612144957	4.2899E-16	1.791759469	5.24711E-06
0.9828991	5.56093E-06	1.153885901	0.000500279
-0.560645882	0.060894366	-1.098612289	0.003485391
0.417324963	0.074453132	-0.895879735	0.011273181
0.463359496	0.065883502	0.15415068	0.672548591
p=10	p=10	p=10	p=10
0.384084922	0.464402432	0.58477431	0.801650023
-0.050747657	0.08671358	-0.075054745	0.41357639
-0.057078856	0.032634483	-0.228630578	0.542186264
0.029758198	0.115012549	0.031920284	0.712066914
1.27641373	5.04649E-05	1.494513464	0.006163317
-0.49101975	0.108843231	-1.551524127	0.004823019
0.695417304	0.031347036	-0.550804048	0.116725817
0.51570196	0.038283333	0.240419601	0.455510356
1.702713252	0.049607739	2.328481139	0.361784339
1.925827812	0.010947865	6.367384386	0.522903974
p=11	p=11	p=11	p=11
0.300553614	0.563782169	0.572294774	0.82396203
-0.052973918	0.071026853	-0.064571934	0.533735517
-0.063223841	0.01858016	-0.228630578	0.581588144
0.029556035	0.112952505	0.031920284	0.738477924
1.075462114	0.001234211	1.506993	0.013903133
-0.490789901	0.10467971	-1.44145462	0.028732929
0.75906344	0.018972385	-0.63466653	0.161298323
0.421236889	0.138340507	-0.207409887	0.68139566
0.524996731	0.033007275	0.219453981	0.540163525
1.851874686	0.032450747	2.068407611	0.471026876
2.150552074	0.005182259	6.367384386	0.563555167

當 P=10 及 11 時，許多參數都不顯著。當 P=4 時，雖然因不明原因出現 NAN，但係數大致上是比較接近的。

再來比較 MSPR 與 MSE

	p=4	p=5	p=10	p=11
MSPR	179.441	179.677	178.975	182.34
MSE	0.33997	0.33115	0.31025	0.30224

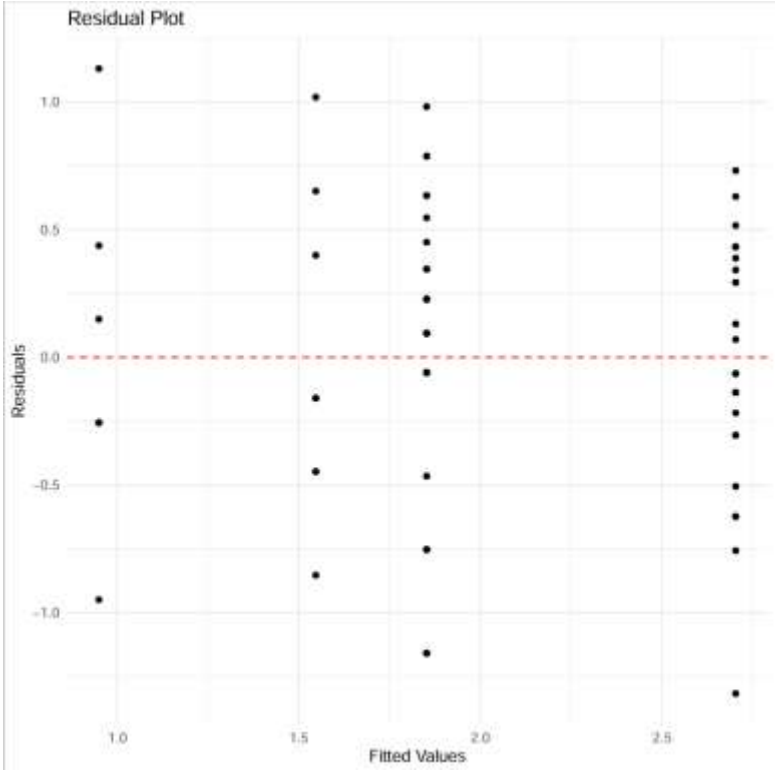
如 MSPR 與 MSE 相近，則模型效果愈好。推測是因為資料量偏小，以及多為類別資料，所以 MSPR 與 MSE 相去甚遠。

於是採用 P=4 當作最佳的模型。

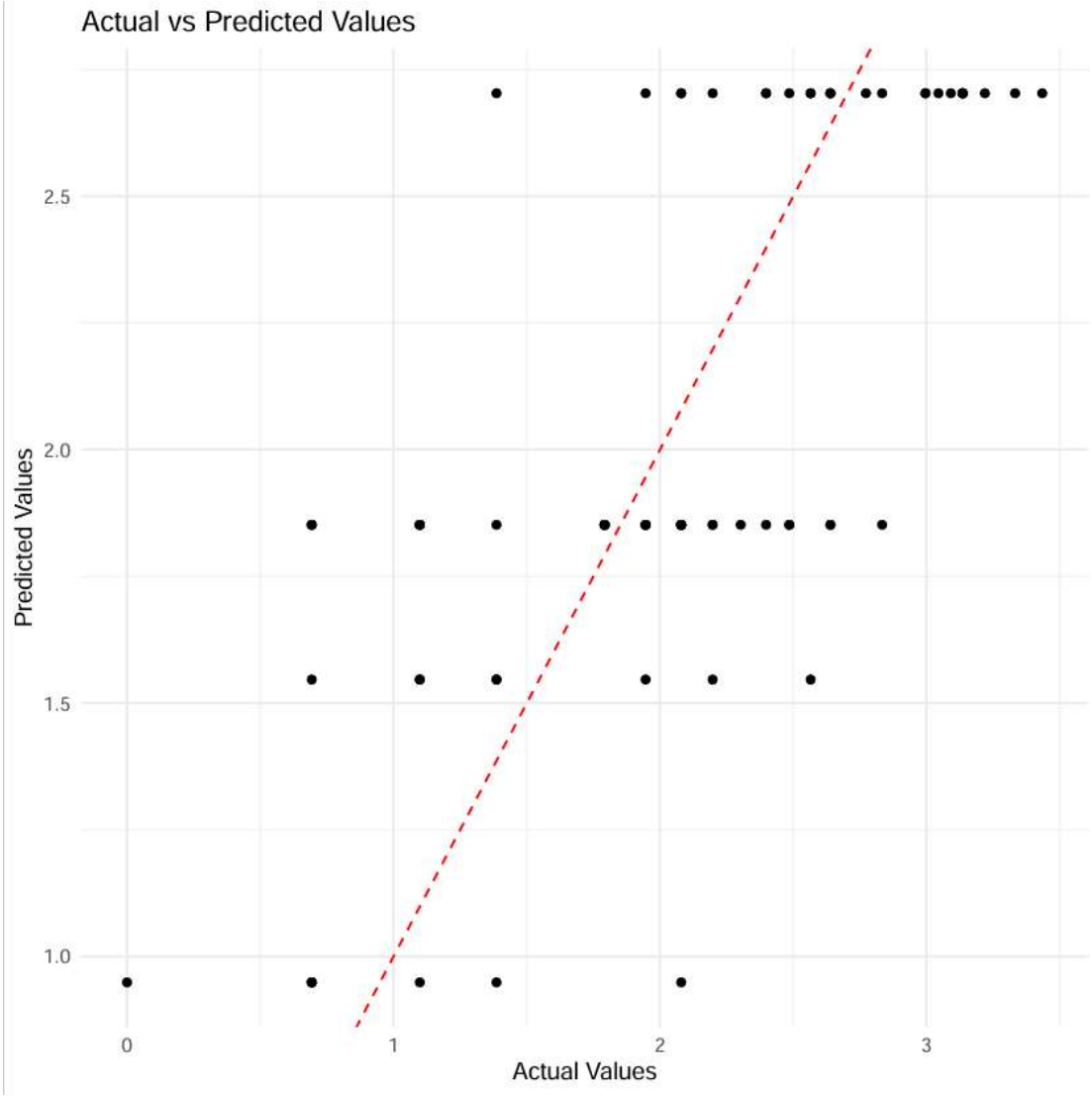
六、 模型結果

最終採用 Process change(製程更新)，X1(是否為 2001Q1)，X3(是否為 2001Q3)為 predictor，ln(Y)為 response。
Ln(y)=1.85173+0.85119(Process change)-0.902(X1)-0.305(X3)，其中 X3 之係數不顯著。R squared=0.50, adjusted R squared = 0.49

最後模型的殘差圖為:



大致上仍維持著隨機分佈。另外比較模型預測值與實際值:



七、 結論

最後的模型中可以觀察到，專案成交量(Websites delivered)與 Process change(製程更新)具有較大的關聯性，正如開頭的相關係數矩陣所表示的一樣。所以對於專案成交量(Websites delivered)來說，Backlog of order(季末訂單數量)與 Team experience(開發經驗)都不是最重要的，反而優化生產過程會顯得更重要一些。另外，也許線性回歸不是了解此問題的最好方式，也許考慮去做分群會更適合一些。