**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](#_Toc169822614)

[**二、** **初步資料調查** 2](#_Toc169822615)

[**三、** **拆分資料** 5](#_Toc169822616)

[**四、** **模型訓練** 5](#_Toc169822617)

[**五、** **模型驗證** 7](#_Toc169822618)

[**六、** **模型結果** 8](#_Toc169822619)

[**七、** **結論** 9](#_Toc169822620)

1. **簡介**

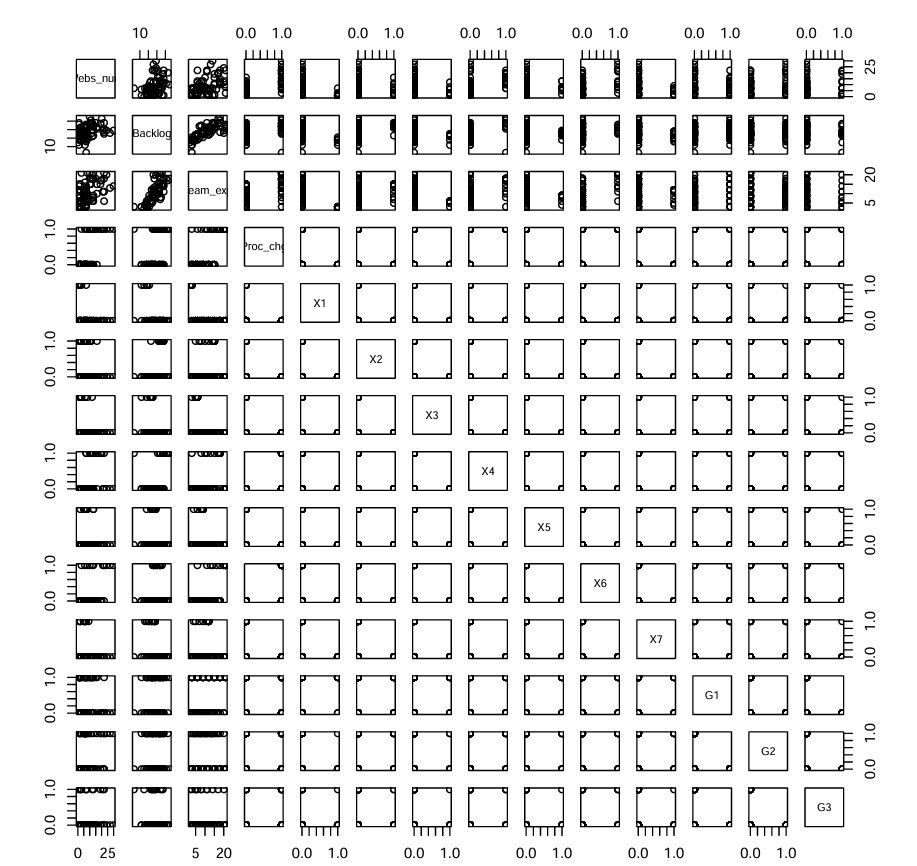
網頁設計公司想了解甚麼因素與專案成交量(Websites delivered)有關，想要尋找對專案成交量影響最顯著的因素。考慮以下這些變數:Backlog of order(季末訂單數量)、Team number(開發小組編號)、Team experience(開發經驗)、Process change(製程更新)、Year、Quarter。

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。

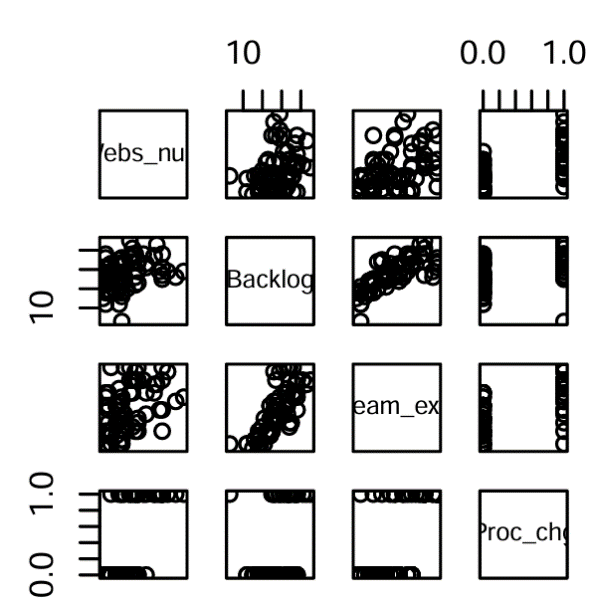
模型欲預測之response為專案成交量(Websites delivered)。

1. **初步資料調查**

資料散佈圖:

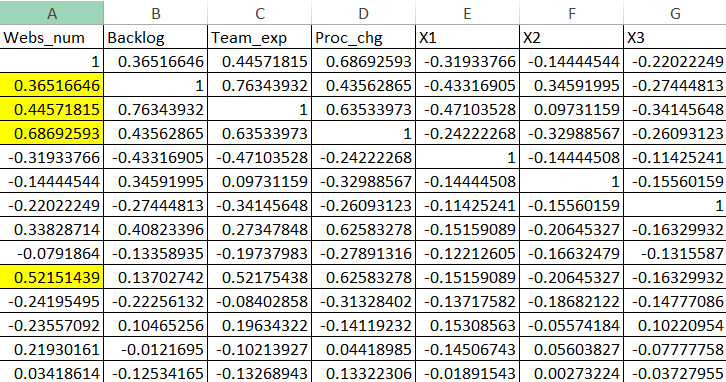


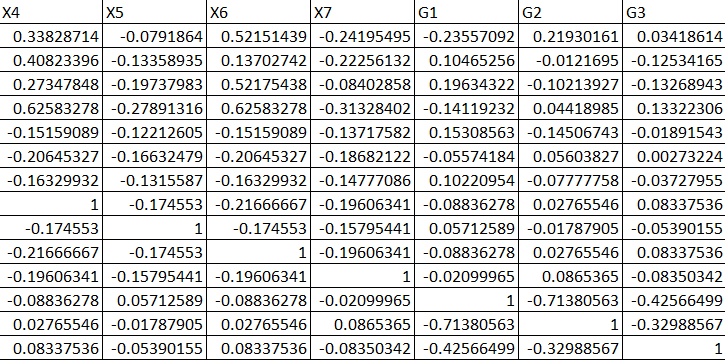
其中X1~X7代表year與quarter合併之類別資料，G1~G3代表組別之類別資料。因為指標變數不具有實際的意義，所以我們更關心左上角的部分:



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。

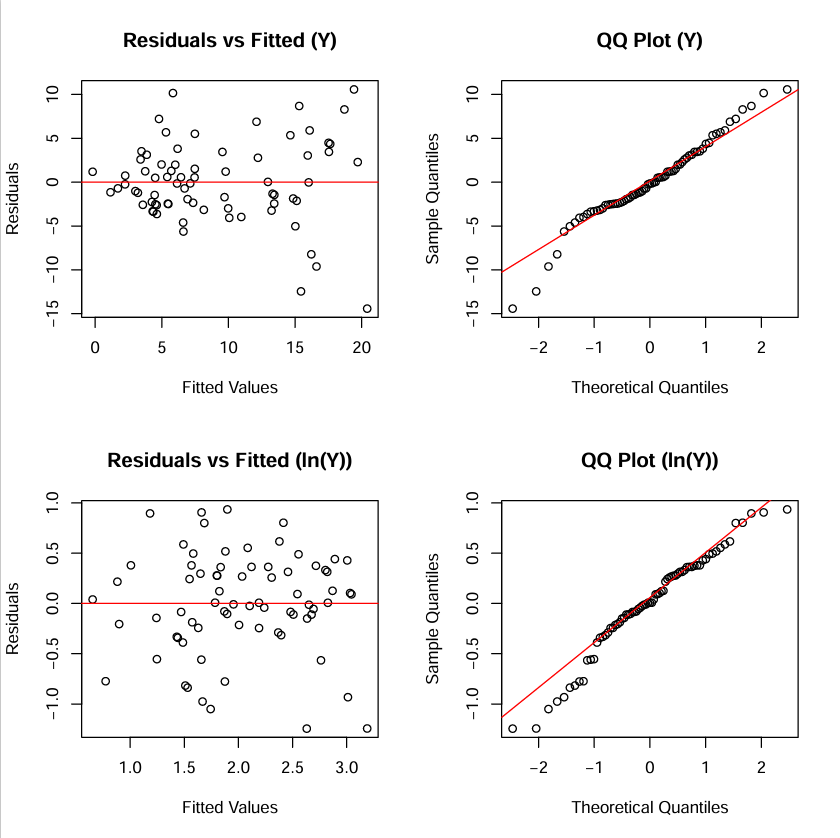
相關係數:





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。

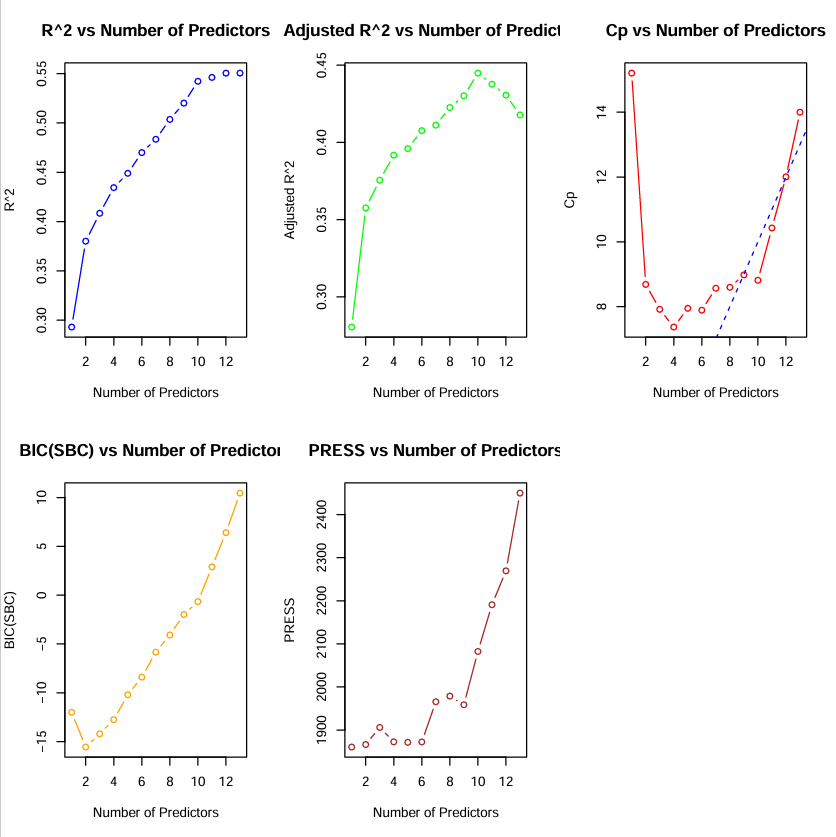
另外，因為Backlog of order(季末訂單數量)與Team experience(開發經驗)之間似乎存在線性關係，於是加入交互項; Backlog of order(季末訂單數量)與Team number(開發小組編號) 之間可能存在關係，因此加入交互項。

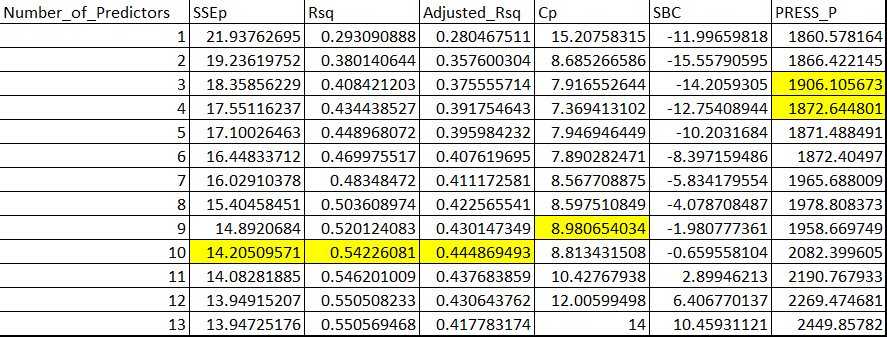
1. **拆分資料**

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。

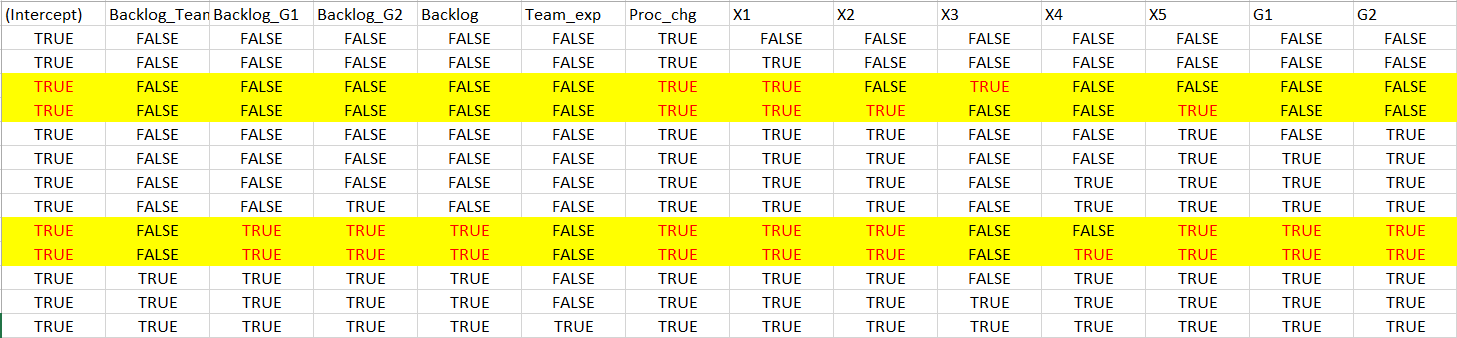
1. **模型訓練**

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



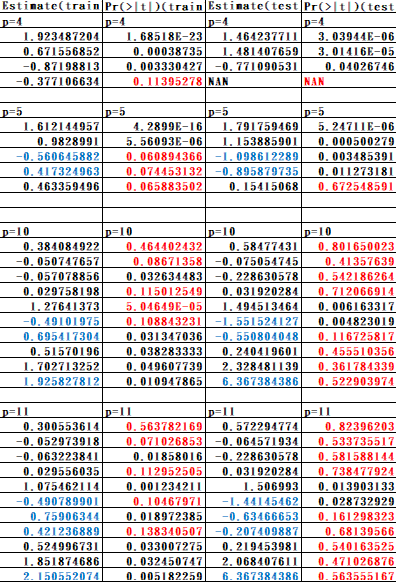


當p=4,5,10,11時看起來較適合。模型使用的變數如下:



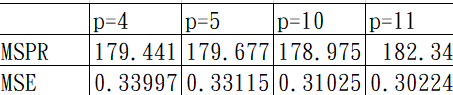
1. **模型驗證**

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。



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



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。

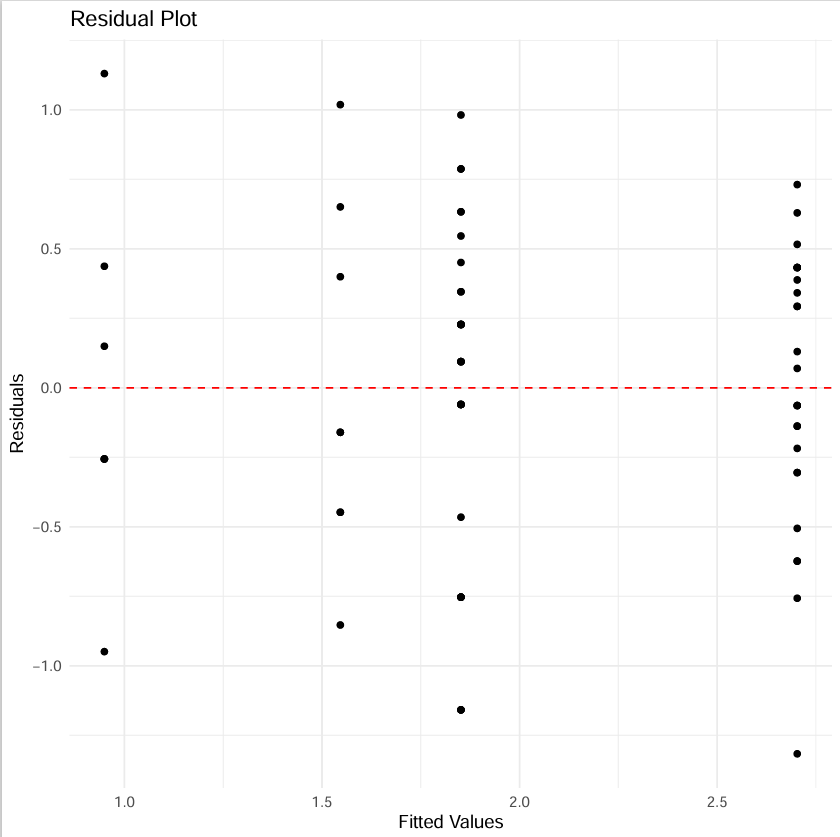
於是採用P=4當作最佳的模型。

1. **模型結果**

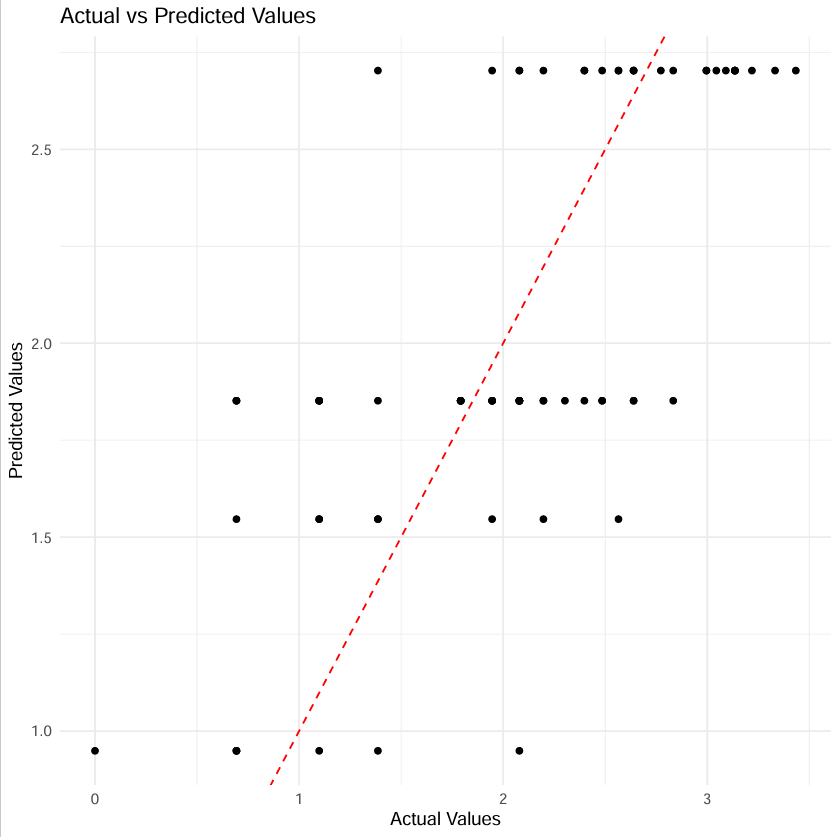
最終採用Process change(製程更新)，X1(是否為2001Q1)，X3(是否為2001Q3)為predictor，ln(Y)為response。

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:



1. **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。