# Using SAS MACRO programs to build a polynomial model and do the selection of variables

**Wang, Kui-Jang**

**Department of Mathematics, Tamkang University, New Taipei city, Taiwan**

Email: kjwang@math.tku.edu.tw



**Abstract:**

The purpose of this paper is trying to provide a useful solution to build a polynomial model. In the past years, there is a few applications on Polynomial model, the reason is that is difficult to create a large amount variables. For example, if you want to build a 3rd order polynomial with 5 variables, then you need 55 variables. If the variables increase to 18, then a 2nd order polynomial model will need 189 variables. It is far away from our ability. That is the reason why I wrote the following programs. There are 3 major reason that I would like to deal with the polynomial model;

1) If the unknown model was smooth plan-curve then a polynomial model can provide an acceptable approximation. This can be easily seen from the Taylor's polynomial.
2) As long as we have enough observations then using a high order polynomial model can solve the unfitted problems.
3) It can avoid deleting important variables from the selection steps, since it is not easy to remove a variable completely from the model because there are too many cross product terms shown in the model.

This paper will provide 2 major SAS MACRO programs, **%**Homopoly and **%**Model_Selection. The first program is used to generate a polynomial model and the next one will provide summarized result tables similar to the table ***11.8 of Montgomery***[1] including the information of the models and necessary statistics. Users can easily apply to do the further analysis. To write those programs, I also wrote another 20 SAS MACRO programs which can be download from the web-site **http://tsp.ec.tku.edu.tw/QuickPlace/054569qp/Main.nsf/h_Toc/BADD7D0BFF0904A1482576D300229684/?OpenDocument**. Please follow the instruction given by the readme.txt file.

**Keywords**

Polynomial model, Taylor's polynomial, SAS MACRO

# 用 SAS MACRO 程式建立多項式模型與變數篩選

王國徵

淡江大學數學系，新北市，台灣
Email: kjwang@math.tku.edu.tw

## 摘要：

本篇論文是希望藉助 SAS Macro 程式，提出一個能解決建立多項式模型上的困惱．多項式模型在統計分析上一直是被忽略的,這可以很清楚的知道因爲在所有的統計分析的出版品中很難找到以多項式迴歸爲主提的例子.這原因無非是無法解決大量變數的模型建立與分析．舉例來說要建立一個完整的 5 個變數的 3 次多項式總共需要 55 個變數,而如果變數增加到 18 個,那建立一個 2 次多項式就高達 189 個變數,因此在實用分析上是鮮有這樣的例子．本篇論文就是希望能提出一個解決模型建立與初步分析的方法,讀者可以藉由在第三章的例子的輸出報表中很清楚去比較各個模型的優劣點,這就是爲何統計分析需要工具去產生整合型的報表.欠缺這些報表,要去判定模型的好壞(基於預測值的準確度)是很困難的工作．而我之所以強調要用多項式的模型去分析資料有下列幾點原因;

(1) 如果模型爲平滑曲面則多項式模型可以提供一個可接受的模型.這可以很容易由泰勒定理得到驗證.
(2) 只要觀測值夠多,大部分模型的不配合均可以用高階多項式模型解決.
(3) 可以避免因爲經過模型篩選而刪除掉可能是有用的變數.那是因爲模型會產生很多的交叉相乘項,既使用模型篩選的程式也很難將一個變數完全去除掉，因此可以保存幾乎所有的變數,而因此將不失模型的完整性.

本篇論文提供了兩支主要的SAS MACRO 程式; **%**Homopoly 和 **%**Model_Selection分別會在下個兩章節中介紹.程式 **%**Homopoly 是用在建立多項式資料檔.而 **%**Model_Selection 則是用來提供SAS 模型篩選後的總結資料,報表格式是仿照***表11.8 Montgomery***[1] 製作的．讀者可以很容易複製到其他的分析．爲了要編寫程式，我同時提供了 20 支工具程式，讀者可以至以下的網站下載 http://tsp.ec.tku.edu.tw/QuickPlace/054569qp/Main.nsf/h_Toc/BADD7D0BFF0904A1482576D300229684/?OpenDocument．請依循檔案README.TXT中的指示去安裝即可.

## 1. 引言

自 80 年代有了套裝軟體(Statistical Package)後,應用統計就離不開它了.但是我們都忘了一件事,那就是那時寫的軟體是針對那時分析需要而寫的,不是針對提供正確的統計分析而寫的．舉例來說；在確定模型前大家都要看殘差圖去決定模型是否可行,但是卻忘了圖形只是參考而非是聖經．個人的解讀不同就有不同的結果,而忘了有分析統計量可以很精確的判讀模型是否可用．這都是因爲統計分析的教材還停留在 80 年代,都沒有跟上軟硬體的進步．試問爲何***多項式模型***一直無法被廣泛的利用(對我來說只要資料量夠大幾乎都可以找到適合的模型)? 那是因爲軟體的限制．爲何不檢查變異數的一致性? 簡單,軟體不支援．因此我希望使用者看著這篇論文的結果,然後問自己這是不是你要的,如果用套裝軟體要如何得到相似的結果? 又要花多少時間?

由於 SAS 在輸出報表後，不會提供如何篩選與總結資料的選項，導致使用者必須面對龐大的報表去找尋可以使用的訊息．這是造成學習者在學習上難以跨越的鴻溝爲了幫助學習者能快速的學會分析的方法我因此製作了一系列的程式提供統計分析與檢測後的總結報表，使用者因此可以輕鬆的由報表的結果而得到指示而能進行下一步的分析．這一系列的程式(主程式有 13 隻),整合了過去40年統計學上重要的檢定與分析方法,藉由這些工具既可以簡化教學又可以很方便的應用在實務分析．由於多數程式尚需要修改，因此本篇只提供前兩支程式而其他的程式將在完成後陸續發表．本文將不介紹程式本身(程式都太過於龐大不適合擺入內容中),而只介紹如何使用與結果判讀,但是這些程式將提供在後．第二章將介紹如何產生多項式模型,而第三章將利用一個

典型運用的例子去介紹如何藉由程式得到對各個模型完整的初步概念. 而第四章會做總結與展望. 以下就開始介紹程式.

## 2. 如何用%Homopoly 建立多項式資料檔

本章節將介紹程式的組成以及程式如何使用.程式是由一個主程式**%Homopoly** 加兩個副程式所組成,一個副程式%POLY_SUB0 是用來產生完整的 N 次多項式,而%POLY_SUB 則是用來調整次方項用的.程式會依照觀測值的多寡來決定可否產生使用者要求的多項式,如果觀測值不足以產生所需之多項式則程式會自動降階直到滿足分析所需爲止.在表 2.1 將介紹**程式自動產生的 5 個輸出檔案;**而表 2.2 介紹如何輸入 MACRO 變數.

**表 2.1：程式% HOMOPOLY所提供之輸出檔; The output data files provided by % HOMOPOLY**

| SAS 輸出檔案名稱 | 檔 案 內 容 |
|---|---|
| **&OUT_DATA** | 使用者提供輸出檔案名稱 |
| **OUT_NAME** | 爲原始變數名稱與"X?"的對照檔 |
| **VAR_NUM** | 模型使用的變數數目(不含截距) |
| **POLYNAME** | 包含了產生的每一個變數所代表的次方相乘完整的表示式,如 X1**2*X3 代表變數X1的平方乘以X3 |
| **MOD_HOMO** | 儲存4個變數依序爲M_NAME(模型名稱)、X(自變數)、Y(應變數)與FILENAME(分析之資料檔名) |

**表 2.2： 程式% HOMOPOLY的參數輸入表; The table of input parameters of % HOMOPOLY.**

| MACRO 變數名稱 | 解 釋 |
|---|---|
| **IN_DATA** | **原始資料檔名** |
| **OUT_DATA** | 輸出檔案名稱 |
| **X =** | **自變數名稱,變數與變數間以空格來分別** |
| **Y =** | **應變數(分析預測變數名稱)** |
| **LIB = WORK** | **SAS的圖書館名(LIBNAME)用以儲存所有輸出檔案,預設值爲"WORK"** |
| **Degree =** | **最高多項式次方數** |
| **C_CROSS = 2** | **最高多項式交叉相乘項的次方數,預設值爲 " 2 "** |
| **PRINT=NO** | **指示程式是否列印&OUT_DATA 資料檔的前5筆資料,預設值爲 "NO ".** |
| **FOOTNOTE = YES** | **要求SAS列印"足註(FOOTNOTE)"顯示原始變數與輸出變數的對照值,預設值爲 "YES ".** |
| **N_FOOT = 5** | **每一行足註所包含的變數數目預設值爲 "5 ".** |

產生多項式資料檔並不困難,困難在如何取變數名稱以及日後如何辨識變數.因此本程式採用最簡單的變數名字, $x_1, x_2, \cdots$ .因爲可以在每一個SAS文字變數的內容的長度內儲存最多個變數,本程式目前總共可以提供9534個變數供分析使用(基於SAS9.1.1版),如使用最新版本可擴充到超過15000個變數.程式會依照資料檔案的大小自動調整可使用的最高次方數的模型(從4次以下開始,如果不確定可以用到幾次,可以執行工具程式 -- %*M*(5,4);其中第一個變數5爲變數個數,第二個變數爲次方數.用%PUT &M;得到一個5元4次多項是共有幾項),並提供變數與原始變數的每一項的對照表與變數的標籤,以供SAS輸出使用.用以下的例子來介紹程式是如何運作的.

【例題 1.1】: 介紹程式在資料量不足時如何運作,我先用下列程式產生76筆資料,然後要求程式去產生依完整的4次多項式資料.但是要產生這樣的模型需要125筆資料,因此程式自動降成4次多項式而交叉相乘項的次方最大爲3次.以下爲SAS程式;其中用了一個程式,%***VAR_NAME***(X,END=5),用來產生一串文字 "X1 X2 X3 X4 X5".

```
DATA INPUT_D;
    DO I = 1 TO 76;
    X1 = 1; X2 = 2; X3 = 3; X4 = 4; X5 = 5; Y = I; OUTPUT;
END; *產生76筆資料;
%HOMOPOLY(IN_DATA =INPUT_D, OUT_DATA =OUTPUT, X = %VAR_NAME(X,END=3), Y=Y,
                    DEGREE =3, C_CROSS =3 , PRINT = YES, FOOTNOTE = YES, N_FOOT=5);
```

程式會產生下列3個報表由於編排需要將以圖片展示圖2.1爲新舊變數對照表; 變數, TRUE_VAR代表原始變數而變數, REG_VAR爲出現在報表中的代碼. 圖2.2列印出回歸變數, 標籤, 與次方數. 圖2.3列印出資料檔**WORK.OUTPUT**中的五筆資料;

**圖 2.1：變數對照表: The table for true variables and their corresponding regressors stored in WORK.OUT_NAME**

*The table of True variables v. s. Regressing variables*

| TRUE_VAR | REG_VAR |
|---|---|
| Y | Y |
| X1 | X1 |
| X2 | X2 |
| X3 | X3 |

*Reg. Var. = Real Var. --- Y = Y ; X1 = X1 ; X2 = X2 ; X3 = X3 ;*

**圖 2.2：回歸變數標籤表: Table of label for each of regressors stored in WORK.POLYNAME**

*The table of True variables v. s. Regressing variables*

| VARIABLE | LABEL | Degree of Variable |
|---|---|---|
| X1 | X1 | 1 |
| X2 | X2 | 1 |
| X3 | X3 | 1 |
| X4 | X1**2 | 2 |
| X5 | X1**3 | 3 |
| X6 | X1*X2 | 2 |
| X7 | X1**2*X2 | 3 |
| X8 | X2**2 | 2 |
| X9 | X1*X2**2 | 3 |
| X10 | X2**3 | 3 |
| X11 | X1*X3 | 2 |
| X12 | X1**2*X3 | 3 |
| X13 | X2*X3 | 2 |
| X14 | X1*X2*X3 | 3 |
| X15 | X2**2*X3 | 3 |
| X16 | X3**2 | 2 |
| X17 | X1*X3**2 | 3 |
| X18 | X2*X3**2 | 3 |
| X19 | X3**3 | 3 |

*Reg. Var. = Real Var. --- Y = Y ; X1 = X1 ; X2 = X2 ; X3 = X3 ;*

**圖 2.3：輸出資料檔, WORK.OUTPUT 只列印 22 個變數與 5 筆資料; The output data file with 5 observations and 22 variables.**

*A 3rd Degree with CROSS = 3 Model is applied.*
*The Error Degree of Freedom is -- 56*

| Obs | Y | X1 | X2 | X3 | X1**2 | X1**3 | X1*X2 | X1**2*X2 | X2**2 | X1*X2**2 | X2**3 | X1*X3 | X1**2*X3 | X2*X3 | X1*X2*X3 | X2**2*X3 | X3**2 | X1*X3**2 | X2*X3**2 | X3**3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 4 | 8 | 3 | 3 | 6 | 6 | 12 | 9 | 9 | 18 | 27 |
| 2 | 2 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 4 | 8 | 3 | 3 | 6 | 6 | 12 | 9 | 9 | 18 | 27 |
| 3 | 3 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 4 | 8 | 3 | 3 | 6 | 6 | 12 | 9 | 9 | 18 | 27 |
| 4 | 4 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 4 | 8 | 3 | 3 | 6 | 6 | 12 | 9 | 9 | 18 | 27 |
| 5 | 5 | 1 | 2 | 3 | 1 | 1 | 2 | 2 | 4 | 4 | 8 | 3 | 3 | 6 | 6 | 12 | 9 | 9 | 18 | 27 |

*Reg. Var. = Real Var. --- Y = Y ; X1 = X1 ; X2 = X2 ; X3 = X3 ;*

**表 2.3：檔案 MOD_HOMO 的內容; The contents of the data file MOD_HOMO.**

| | M_NAME | X | Y | FILENAME |
|---|---|---|---|---|
| **1** | Homo_Polynomial | X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16 X17 X18 X19 | Y | OUTPUT |

## 3. 如何用%Model_Selection 得到模型

本隻程式設計的目的有三;

1) 提供 6 個回歸模型; 完整模型(Full Model), 線性模型(Linear Model), 前進搜尋法(Forward Selection), 後退搜尋法(Backward Selection), 逐步搜尋法(Stepwise Selection), 與 CP 選擇法得到的模型的預估值與可供選取模型參考的統計量.另一爲根據輸入的模型提供相同的報表.
2) 產生輸出檔案用來做常態分配與變異數的一致性的檢定.
3) 可以同時對未來值作預測.

以下先介紹程式的內容與選項,然後再詳述程式的架構與困難處.
程式的參數位於等號左側而等號右側爲參數的預設值,表3.1介紹如何輸入MACRO變數、
%Model_Selection(DATA_IN = , Y = , X_LIN = , X_FULL= , X = , RES_OUT = NO, F_MODEL = NO, STAT = VIF, SLENTRY = 0.2, SLSTAY = 0.2, CHECK = NO, CHECK_D = CHK_DATA, ID = ID_ID, ID2 = , MODEL_IN = NO, MOD_NAME = , GROUP = ).

**表 3.1：程式 %Model_Selection 的參數輸入表; The table of input parameters of %Model_Selection.**

| MACRO 變數名稱 | 解 釋 |
|---|---|
| DADA_IN | 原始資料檔名 |
| Y = | 應變數(分析預測變數名稱) |
| X_LIN = | 輸入線性模型變數,如 X1 X2 X3… |
| X_FULL= | 輸入完整的模型變數,不論&X爲何必須輸入,如果不輸入則程式會自動選取以&X來代替.如果MODEL_IN=YES則用所有輸入的模型變數爲&X_FULL的值. |
| X = | 自變數名稱,變數與變數間以空格來分別.SAS用來做篩選用,可以與X_FULL不同 |
| RES_OUT = NO | "YES"爲要求程式輸出包含殘差值的輸出檔案名稱,注意：如果MODEL_IN=YES則會爲每一模型產生一個新檔案,這會佔去很大的儲存空間,請謹慎使用！預設值爲"NO". |
| F_MODEL = NO | "NO"爲要求程式在報表中不列印完整模式,因爲如果變數數目太大,經由模型選擇後可以讓報表不會過度膨脹.因爲完整模型不是選項亦或是X_FULL=X_LIN時是不必要包含完整模型因爲X_FULL與X_LIN結果會相同. |
| STAT = VIF | 伴隨參數出現的統計量爲何？共有"VIF"、"PVALUE"與"ALL"都選.如果輸入錯誤則程式會同時列印"VIF"與"PVALUE". |
| SLENTRY = 0.2 | 新進入模型的變數的PVALUE必須小於顯著水準,預設值爲 " 0.2 " |

| | |
|---|---|
| SLSTAY = 0.2 | 從模型移除變數的PVALUE必須大於顯著水準,預設值爲 ”0.2”.注意：選取變數從寬, 移除變數從嚴. |
| CHECK = NO | 模型選取完後是否要用來做預測？預設值爲 ”NO”. |
| CHECK_D = CHK_DATA | 儲存要用來預測的資料檔名,存於& CHECK_D中,預設值爲 ”CHK_DATA”. |
| ID = ID_ID | 在& CHECK_D資料中用來分辨何筆資料爲需要做預測的以及何筆資料是用來做”參考點”的,預設值爲 ”ID_ID”.注意：& CHECK_D的值只能有3種,0(不選取),1(參考點)或2(預測點). |
| ID2 = | 爲非必要選項.此變數是用來辨識資料用. |
| OUT_DATA = NO | 要求程式在有遺漏值(Missing Value)時是否輸出無遺漏值的資料檔案與預測檔案.預設值爲 ”NO”. |
| MODEL_IN = NO | 如果選項是”YES”則程式會針對你給的模型作分析與預測 |
| MOD_NAME = | 如果選項是”YES”則程式會用& MOD_NAME所指定的資料檔去分析,但是資料檔的變數必須是;第一個變數必須爲模型名稱而第二個變數必須是自變數(X). |
| GROUP = | 給分群變數名稱(用在以後執行變異數一致性檢測程式用),如不給則程式會用應變數(Y)來做分群變數 |

**【註】：**

**(1) 如果MODEL_IN = NO 則程式會自動產生一資料檔案 “MMODELS”包含模型名稱(M_NAME),自變數名稱(X),應變數名稱(Y),檔案名稱(FILENAME)與群組名稱(GROUP).**

**(2) 程式會檢查輸入資料檔案有否包含遺漏值(Missing Value),會產生三個資料檔案,CHECK_D01(加入檢察預測值的資料檔案),MISSING(儲存遺漏值)與N_of_miss_val(遺漏值的筆數).**

**(3) 如果 CHECK = YES, 則程式會檢察是否有輸入預測值的資料檔” &CHECK_D”如果沒有則會去資料檔案中找遺漏值然後存於” CHECK_D01”.注意：如果資料檔案中有遺漏值則程式會主動將之加入預測檔案中,所以請小心對待資料檔案中的遺漏值.**

以下用(Kunugi, Tamura, and Naito[1961])的論文資料爲例子：

**【例題 3.1】：用稀釋氫產生乙炔的新程序;應變數爲轉換正庚烷爲乙炔的比例(P單位爲%),自變數有三反應溫度(T單位爲℃),氫佔正庚烷的莫爾比率(H單位爲mole ratio %),與接觸時間(C單位爲秒).資料檔案內容將列印在表 3.2 如下：**

**表 3.2：資料檔Ridge_20.sas7bdat不含末 4 筆的存在Ridge_16.sas7bdat; The contents of dataRidge_20.sas7bdat**

| | TEMP | H_RATIO | TIME | P | ID | ID_ID |
|---|---|---|---|---|---|---|
| 1 | 1300 | 7.5 | 0.012 | 49 | o | 0 |
| 2 | 1300 | 9 | 0.012 | 50.2 | o | 0 |
| 3 | 1300 | 11 | 0.0115 | 50.5 | J | 1 |
| 4 | 1300 | 13.5 | 0.013 | 48.5 | o | 0 |
| 5 | 1300 | 17 | 0.0135 | 47.5 | I | 1 |
| 6 | 1300 | 23 | 0.012 | 44.5 | o | 0 |
| 7 | 1200 | 5.3 | 0.04 | 28 | o | 0 |
| 8 | 1200 | 7.5 | 0.038 | 31.5 | o | 0 |
| 9 | 1200 | 11 | 0.032 | 34.5 | o | 0 |
| 10 | 1200 | 13.5 | 0.026 | 35 | F | 1 |
| 11 | 1200 | 17 | 0.034 | 38 | o | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | 1200 | 23 | 0.041 | 38.5 | E | 1 |
| 13 | 1100 | 5.3 | 0.084 | 15 | B | 1 |
| 14 | 1100 | 7.5 | 0.098 | 17 | A | 1 |
| 15 | 1100 | 11 | 0.092 | 20.5 | o | 0 |
| 16 | 1100 | 17 | 0.086 | 29.5 | o | 0 |
| 17 | 1100 | 11 | 0.012 | . | C | 2 |
| 18 | 1200 | 23 | 0.098 | . | D | 2 |
| 19 | 1200 | 7.5 | 0.012 | . | G | 2 |
| 20 | 1300 | 11 | 0.098 | . | H | 2 |

將自變數標準化再產生一完整的 2 次多項式資料儲存在 Ridge_S20.sas7bdat 內容列印在表 3.3 而不含末 4 筆的資料存在 Ridge_S16.sas7bdat,末 4 筆資料存在 Ridge_S04.sas7bdat.

表 3.3：檔案Ridge_S20.sas7bdat內容; The data of file, Ridge_S16.sas7bdat.

| | P | ID | T | H | C | TH | TC | HC | T2 | H2 | C2 | ID_ID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 49 | o | 1.0853039277 | -0.873140343 | -0.8948744 | -0.947622644 | -0.971210701 | 0.7813509409 | 1.1778846154 | 0.7623740588 | 0.8008001923 | 0 |
| 2 | 50.2 | o | 1.0853039277 | -0.608217862 | -0.8948744 | -0.660101235 | -0.971210701 | 0.5442785948 | 1.1778846154 | 0.369928968 | 0.8008001923 | 0 |
| 3 | 50.5 | J | 1.0853039277 | -0.254987888 | -0.910677922 | -0.276739356 | -0.988362325 | 0.2322118397 | 1.1778846154 | 0.0650188229 | 0.8293342771 | 1 |
| 4 | 48.5 | o | 1.0853039277 | 0.1865495803 | -0.863267357 | 0.2024629922 | -0.936907454 | -0.161042163 | 1.1778846154 | 0.0348007459 | 0.7452305304 | 0 |
| 5 | 47.5 | I | 1.0853039277 | 0.8047020356 | -0.847463836 | 0.8733462798 | -0.91975583 | -0.681955874 | 1.1778846154 | 0.6475453661 | 0.7181949534 | 1 |
| 6 | 44.5 | o | 1.0853039277 | 1.864391959 | -0.8948744 | 2.0234319158 | -0.971210701 | -1.668396636 | 1.1778846154 | 3.4759573769 | 0.8008001923 | 0 |
| 7 | 28 | o | -0.155043418 | -1.261693315 | -0.009877201 | 0.1956172443 | 0.001531395 | 0.0124619983 | 0.0240384615 | 1.5918700213 | 0.0000975591 | 0 |
| 8 | 31.5 | o | -0.155043418 | -0.873140343 | -0.073091287 | 0.1353746634 | 0.0113323229 | 0.063818951 | 0.0240384615 | 0.7623740588 | 0.0053423362 | 0 |
| 9 | 34.5 | o | -0.155043418 | -0.254987888 | -0.262733544 | 0.0395341937 | 0.0407351067 | 0.0669938713 | 0.0240384615 | 0.0650188229 | 0.0690289149 | 0 |
| 10 | 35 | F | -0.155043418 | 0.1865495803 | -0.452375801 | -0.028923285 | 0.0701378904 | -0.084390516 | 0.0240384615 | 0.0348007459 | 0.204643865 | 1 |
| 11 | 38 | o | -0.155043418 | 0.8047020356 | -0.199519458 | -0.124763754 | 0.0309341788 | -0.160553714 | 0.0240384615 | 0.6475453661 | 0.0398080141 | 0 |
| 12 | 38.5 | E | -0.155043418 | 1.864391959 | 0.0217298419 | -0.289061702 | -0.003369069 | 0.0405129426 | 0.0240384615 | 3.4759573769 | 0.000472186 | 1 |
| 13 | 15 | B | -1.395390764 | -1.261693315 | 1.3808326839 | 1.760555199 | -1.926801174 | -1.742187366 | 1.9471153846 | 1.5918700213 | 1.9066989009 | 1 |
| 14 | 17 | A | -1.395390764 | -0.873140343 | 1.8233312836 | 1.2183719706 | -2.544259633 | -1.592024103 | 1.9471153846 | 0.7623740588 | 3.3245369697 | 1 |
| 15 | 20.5 | o | -1.395390764 | -0.254987888 | 1.6336890266 | 0.3558077436 | -2.279634579 | -0.416570914 | 1.9471153846 | 0.0650188229 | 2.6689398355 | 0 |
| 16 | 29.5 | o | -1.395390764 | 0.8047020356 | 1.4440467695 | -1.122873788 | -2.015009525 | 1.162027375 | 1.9471153846 | 0.6475453661 | 2.0852710726 | 0 |
| 17 | . | C | -1.395390764 | -0.254987888 | -0.8948744 | 0.3558077436 | 1.2486994732 | 0.2281821332 | 1.9471153846 | 0.0650188229 | 0.8008001923 | 2 |
| 18 | . | D | -0.155043418 | 1.864391959 | 1.8233312836 | -0.289061702 | -0.282695515 | 3.3994041837 | 0.0240384615 | 3.4759573769 | 3.3245369697 | 2 |
| 19 | . | G | -0.155043418 | -0.873140343 | -0.8948744 | 0.1353746634 | 0.1387443859 | 0.7813509409 | 0.0240384615 | 0.7623740588 | 0.8008001923 | 2 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20 | . | H | 1.0853039277 | -0.254987888 | 1.8233312836 | -0.276739356 | 1.9788686035 | -0.464927393 | 1.1778846154 | 0.0650188229 | 3.3245369697 | 2 |

【註】：1) 最後 4 筆資料爲需要預測的值

2) 變數 TH 等爲交叉相乘項,而 T2 等爲平方項.

爲減少篇幅我用 20 筆標準化後的資料來執行模型篩選與預測,但不指定預測檔案.此時程式會將含有遺漏值的 4 筆資料當作需要被預測的資料,並且會存放在 2 個不同的資料檔,**Check_d01.sas7bdat 與 Missing.sas7bdat 而資料檔 N_of_miss_val.sas7bdat 則存入遺漏值的數目. 用 MODEL_IN = NO 選項執行程式%*Model_Selection*** 將得到以下 2 個圖表;圖 3.1 爲資料檔 MMODEL.sas7bdat 的內容, 圖 3.2, 列印出在各個模型中參考點與預估點的估計結果.

%***Model_Selection***(DATA_IN =EXAMPLE.RIDGE_S20 , X= T H C TH TC HC T2 H2 C2, Y= P, X_LIN= T H C, CHECK = YES, ID = ID_ID, ID2 = ID, GROUP=TEMP, F_MODEL=YES, STAT=VIF, MODEL_IN=NO, SLENTRY= **0.2**, SLSTAY = **0.2**);

【註】：1) 選項 X_FULL 在 MODEL_IN = NO 時可以不給,
程式會用&X 來取代.

2) F_MODEL=YES 是要求程式列印包含全部變數的模型.

3) STAT 選項是選擇列印 VIF(variance inflation factor)的值,
因爲除了全部模型與線性模型外 p-value 值均小於 0.2.

**圖 3.1：輸出資料檔 MMODELS.sas7bdat 的內容; The data of file MMODELS.sas7bdat.**

*Table for Subset Regression Models for the Standardized Data set*
*Warning!! -- The data set containing missing values. Missing values will be stored in the data set 'MISSING'.*

| Obs | M_NAME | X | Y | FILENAME | GROUP |
|---|---|---|---|---|---|
| 1 | F_MODEL | T H C TH TC HC T2 H2 C2 | P | EXAMPLE.RIDGE_S20 | TEMP |
| 2 | LINEAR | T H C | P | EXAMPLE.RIDGE_S20 | TEMP |
| 3 | FORWARD | T H TH T2 H2 | P | EXAMPLE.RIDGE_S20 | TEMP |
| 4 | BACKWARD | H C TH TC HC T2 H2 C2 | P | EXAMPLE.RIDGE_S20 | TEMP |
| 5 | STEPWISE | T H TH T2 H2 | P | EXAMPLE.RIDGE_S20 | TEMP |
| 6 | CP | T H TH HC T2 H2 C2 | P | EXAMPLE.RIDGE_S20 | TEMP |

**圖 3.2：各個模型的預測值與參考點值; The predictions of referent and predicted points.**

*Comparing the Predicted values via Different Models*

*Warning!! -5— The Checking data set — CHK_DATA — does not exist. Using Check_D01 instead!*

| ID | Type of Point | P | F_MODEL -- (with 9 Regressors) | LINEAR -- (with 3 Regressors) | FORWARD -- (with 5 Regressors) | BACKWARD -- (with 8 Regressors) | STEPWISE -- (with 5 Regressors) | CP -- (with 7 Regressors) |
|---|---|---|---|---|---|---|---|---|
| J | Ref. Point | 1.20968 | 1.16814 | 0.93666 | 1.11716 | 1.1934 | 1.11716 | 1.12600 |
| I | Ref. Point | 0.95756 | 0.94336 | 1.10902 | 0.99846 | 0.9064 | 0.99846 | 0.99614 |
| F | Ref. Point | -0.09297 | -0.06469 | -0.07948 | -0.05867 | -0.0689 | -0.05867 | -0.04631 |
| E | Ref. Point | 0.20118 | 0.19253 | 0.17451 | 0.17832 | 0.1776 | 0.17832 | 0.21351 |
| B | Ref. Point | -1.77382 | -1.78697 | -1.47824 | -1.90171 | -1.7816 | -1.90171 | -1.78752 |
| A | Ref. Point | -1.60573 | -1.66221 | -1.43625 | -1.60748 | -1.6747 | -1.60748 | -1.64676 |
| C | Focast Point | . | -5.87358 | -1.19636 | -1.17907 | -8.8076 | -1.17907 | -1.07280 |
| D | Focast Point | . | -4.69420 | 0.08339 | 0.17832 | -6.7813 | 0.17832 | -1.40708 |
| G | Focast Point | . | -1.17225 | -0.23266 | -0.39338 | -1.3770 | -0.39338 | -0.80221 |
| H | Focast Point | . | -9.61831 | 0.79838 | 1.11716 | -16.2429 | 1.11716 | 0.87973 |

當你執行完後會在log視窗中發現兩個警告訊息

***WARNING: The NOINT option is ignored in the computation of ridge regression.***

***WARNING: The variable _CP_ in the DROP, KEEP, or RENAME list has never been referenced.***

我希望從 SAS 輸出中抓取 VIF 的值,而 SAS 的 REG 程序中如不加入 NOINT 的選項則會輸出函有截距的模型.但是我用的資料檔案爲標準化過的資料,因此 NOINT 的選項必需加入.另外的警告是在計算 CP 值時出現的,那是因爲我想縮短程式因此就不去處裡這個警告,請直接忽略它.要瞭解程式如何使用可以執行以下的程式;

1) 使用未標準化的資料：%***Model_Selection***(DATA_IN = EXAMPLE.RIDGE_20 , X = T H C TH TC HC T2 H2 C2, Y = P, X_LIN = T H C, CHECK = YES, ID = ID_ID, ID2 = ID, GROUP = TEMP, F_MODEL=YES, STAT=VIF, MODEL_IN=NO, SLENTRY= **0.2**, SLSTAY = **0.2**);

2) 使用未標準化的資料：%***Model_Selection***(DATA_IN = EXAMPLE.RIDGE_16 , X = T H C TH TC HC T2 H2 C2, Y = P, X_LIN = T H C, CHECK = YES, CHECK_D = EXAMPLE.RIDGE_04, ID = ID_ID, ID2 = ID, GROUP = TEMP, F_MODEL=YES, STAT=VIF, MODEL_IN=NO, SLENTRY= **0.2**, SLSTAY = **0.2**);

我同時提供一SAS程式"Example of Model Selection.sas"以供模仿用.在使用前請先參考檔案"README.TXT".

## 4. 總結與展望

模型選取有一個重要的原則, 加入變數從寬去除變數從嚴. 當分析者拿到資料時感到最困惱的是如何選取變數. 這兩支程式基本上可以解決過早刪除變數. 因爲當你先建立多項式模型,再執行變數篩選, 你不大容易將一個變數完整的從模型中刪除, 所以無須擔心重要變數被意外的移出模型. 這可以節省很多時間與精力. 再者程式提供了總結後的輸出報表, 其中包含各個模型的資料與各種基本的統計量, 使用者可以很快速的決定哪些模型可供使用. 這將大大的縮減判斷的時間, 因此可以很容易的執行下一步的檢驗. 而程式產生的輸出檔可以讓使用者非常容易的使用, 而毋須煩惱變數太多無法快速的輸入. 當然多項式模型必然產生共線性的問題, 因此在模型確定後再處裡共線性的問題. 由於目前尚未有完美的解答, 因此我只希望能總結各家提出的選項再會診後集結爲一程式可供使用.

在下一篇我將討論如何檢定模型的基本假設(常態分配、變異數的一致性、與資料的獨立性).程式是基於檢定統計量而不是用殘差分析圖去作判定.過往會使用圖形去判定的原因是因爲電腦太慢

又太貴,故無法提供解決方法而非方法不存在,詳細內容請參考 ***Wang*** [2] .由於程式太長,請參考附件檔案. 我也衷心的希望製作軟體的先進們能製作出正確又方便的工具,能讓統計分析方便又能得到正確的結果.

參考文獻：

[1] Montgomery, Douglas C. Peck, Elizabeth A., Vining, G. Geoffrey [2006], "Introduction to Linear Regression Analysis 4th ed., Willey, New York.

[2] Wang, Kui-Jang [2013], "Notes for Regression Analysis". Tamkang University