Advances in Applied Mathematics 应用数学进展, 2015, 4, 162-171
Published Online May 2015 in Hans. http://www.hanspub.org/journal/aam
http://dx.doi.org/10.12677/aam.2015.42021

Hans汉斯

Using SAS MACRO Programs to Build a Polynomial Model and Do the Selection of Variables

Kui-Jang Wang

Department of Mathematics, Tamkang University, New Taipei Taiwan
Email: kjwang@math.tku.edu.tw

Received: Apr. 28th, 2015; accepted: May 15th, 2015; published: May 20th, 2015

Abstract

The purpose of this paper is trying to provide a useful solution to build a polynomial model. In the past years, there are a few applications on polynomial model; the reason is that it is difficult to create a large number of 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 reasons 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 downloaded 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程序建立多项式模型与变量筛选

王国征

淡江大学数学系，台湾 新北
Email: kjwang@math.tku.edu.tw

收稿日期：2015年4月28日；录用日期：2015年5月15日；发布日期：2015年5月20日

摘 要

本篇论文是希望藉助SAS MACRO程序，提出一个能解决建立多项式模型上的困恼。多项式模型在统计分析上一直是被忽略的，这可以很清楚的知道因为在所有的统计分析的出版品中很难找到以多项式回归为主提的例子。这原因无非是无法解决大量变量的模型建立与分析。举例来说要建立一个完整的5个变量的3次多项式总共需要55个变量，而如果变量增加到18个，那建立1个2次多项式就高达189个变量，因此在实用分析上是鲜有这样的例子。本篇论文就是希望能提出1个解决模型建立与初步分析的方法，读者可以藉由在第三章的例子的输出报表中很清楚去比较各个模型的优劣点，这就是为何统计分析需要工具去产生整合型的报表。欠缺这些报表，要去判定模型的好坏(基于预测值的准确度)是很困难的工作。而我之所以强调要用多项式的模型去分析数据有下列几点原因；1) 如果模型为平滑曲面则多项式模型可以提供一个可接受的模型。这可以很容易由泰勒定理得到验证；2) 只要观测值够多，大部分模型的不配合均可以用高阶多项式模型解决；3) 可以避免因为经过模型筛选而删除掉可能是有用的变量。那是因为模型会产生很多的交叉相乘项，既使用模型筛选的程序也很难将一个变量完全去除掉，因此可以保存几乎所有的变量，而因此将不失模型的完整性。本篇论文提供了两支主要的SAS MACRO程序；%Homopoly和%Model_Selection分别会在下个两章节中介绍。程序%Homopoly是用在建立多项式数据文件，而%Model_Selection则是用来提供SAS模型筛选后的总结数据，报表格式是仿照表11.8 *Montgomery* [1]制作的。读者可以很容易复制到其他的分析。为了要编写程序，我同时提供了20支工具程序，读者可以至以下的网站下载http://tsp.ec.tku.edu.tw/QuickPlace/054569qp/Main.nsf/h_Toc/BADD7D0BFF0904A1482576D300229684/?OpenDocument。请依循档案README.TXT中的指示去安装即可。

关键词

多项式模型，泰勒多项式，SAS MACRO

1. 引言

自 80 年代有了软件包(Statistical Package)后，应用统计就离不开它了。但是我们都忘了一件事，那就是那时写的软件是针对那时分析需要而写的，不是针对提供正确的统计分析而写的。举例来说；在确定模型前大家都要看残差图去决定模型是否可行，但是却忘了图形只是参考而非是圣经。个人的解读不同就有不同的结果，而忘了有分析统计量可以很精确的判读模型是否可用。这都是因为统计分析的教材还停留在 80 年代，都没有跟上软硬件的进步。试问为何***多项式模型***一直无法被广泛的利用(对我来说只要资料量够大几乎都可以找到适合的模型)？那是因为软件的限制。为何不检查变异数的一致性？简单，软件不支持。因此我希望使用者看着这篇论文的结果，然后问自己这是不是你要的，如果用软件包要如何

得到相似的结果？又要花多少时间？

由于 SAS 在输出报表后，不会提供如何筛选与总结数据的选项，导致用户必须面对庞大的报表去找寻可以使用的讯息。这是造成学习者在学习上难以跨越的鸿沟为了帮助学习者能快速的学会分析的方法我因此制作了一系列的程序提供统计分析与检测后的总结报表，用户因此可以轻松的由报表的结果而得到指示而能进行下一步的分析。这一系列的程序(主程序有 13 只)，整合了过去 40 年统计学上重要的检定与分析方法，藉由这些工具既可以简化教学又可以很方便的应用在实务分析。由于多数程序尚需要修改，因此本篇只提供前两支程序而其他的程序将在完成后陆续发表。本文将不介绍程序本身(程序都太过于庞大不适合摆入内容中)，而只介绍如何使用与结果判读，但是这些程序将提供在后。第二章将介绍如何产生多项式模型，而第三章将利用一个典型运用的例子去介绍如何藉由程序得到对各个模型完整的初步概念。而第四章会做总结与展望。以下就开始介绍程序。

2. 如何用%Homopoly 建立多项式数据文件

本章节将介绍程序的组成以及程序如何使用。程序是由一个主程序**%Homopoly** 加两个子程序所组成，一个子程序%POLY_SUB0 是用来产生完整的 N 次多项式，而%POLY_SUB 则是用来调整次方项用的。程序会依照观测值的多寡来决定可否产生使用者要求的多项式，如果观测值不足以产生所需之多项式则程序会自动降级直到满足分析所需为止。在表 1 将介绍**程序自动产生的 5 个输出档案**；而表 2 介绍如何输入 MACRO 变量。

Table 1. The output data files provided by % HOMOPOLY
表 1. 程序% HOMOPOLY 所提供之输出档

SAS输出文件名	档案内容
&OUT_DATA	使用者提供输出文件名
OUT_NAME	为原始变量名称与“X?”的对照档
VAR_NUM	模型使用的变量数目(不含截距)
POLYNAME	包含了产生的每一个变量所代表的次方相乘完整的表示式,如X1**2*X3代表变量X1的平方乘以X3
MOD_HOMO	储存4个变量依序为M_NAME(模型名称)、X(自变量)、Y(应变量)与FILENAME(分析之资料文件名)

Table 2. The table of input parameters of % HOMOPOLY
表 2. 程序% HOMOPOLY 的参数输入表

MACRO变量名称	解释
IN_DATA	**原始资料文件名**
OUT_DATA	**输出文件名**
X=	**自变量名称，变量与变量间以空格来分别**
Y=	**应变量(分析预测变量名称)**
LIB = WORK	**SAS的图书馆名(LIBNAME)用以储存所有输出档案，默认值为“WORK”**
Degree=	**最高多项式次方数**
C_CROSS = 2	**最高多项式交叉相乘项的次方数，默认值为“2”**
PRINT = NO	**指示程序是否打印&OUT_DATA数据文件的前5笔数据，默认值为“NO”**
FOOTNOTE = YES	**要求SAS打印“足注(FOOTNOTE)”显示原始变量与输出变量的对照值，默认值为“YES”**
N_FOOT = 5	**每一行足注所包含的变量数目默认值为“5”**

产生多项式数据文件并不困难，困难在如何取变量名称以及日后如何辨识变量。因此本程序采用最简单的变量名字，$x_1, x_2, \cdots$。因为可以在每一个SAS文字变量的内容的长度内储存最多个变量，本程序目前总共可以提供9534个变量供分析使用(基于SAS9.1.1版)，如使用最新版本可扩充到超过15,000个变量。程序会依照数据文件的大小自动调整可使用的最高次方数的模型(从4次以下开始，如果不确定可以用到几次，可以执行工具程序——%M (5, 4)；其中第一个变量5为变量个数，第二个变数为次方数。用%PUT &M；得到一个5元4次多项是共有几项)，并提供变量与原始变量的每一项的对照表与变量的卷标，以供SAS输出使用。用以下的例子来介绍程序是如何运作的。

[例题1.1]：介绍程序在数据量不足时如何运作，我先用下列程序产生76笔数据，然后要求程序去产生依完整的4次多项式数据。但是要产生这样的模型需要125笔数据，因此程序自动降成4次多项式而交叉相乘项的次方最大为3次。以下为SAS程序；其中用了一个程序，%***VAR_NAME***(X,END=5)，用来产生一串文字“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；*产生76笔数据;
%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);
```

程序会产生下列3个报表由于编排需要将以图片展示图1为新旧变量对照表；变量，TRUE_VAR代表原始变量而变量，REG_VAR为出现在报表中的代码。图2打印出回归变量，卷标，与次方数。图3打印出数据文件**WORK。OUTPUT**中的五笔资料；打印文件MOD_HOMO的内容见表3。

表 3 打印文件 MOD_HOMO 的内容。

3. 如何用%Model_Selection 得到模型

本只程序设计的目的有三：

1) 提供 6 个回归模型；完整模型(Full Model)，线性模型(Linear Model)，前进搜寻法(Forward Selection)，后退搜寻法(Backward Selection)，逐步搜寻法(Stepwise Selection)，与 CP 选择法得到的模型的预估值与可供选取模型参考的统计量。另一为根据输入的模型提供相同的报表。

2) 产生输出档案用来做常态分配与变异数的一致性的检定。

3) 可以同时对未来值作预测。

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 ;

Figure 1. The table for true variables and their corresponding regressors stored in WORK.OUT_NAME
图 1. 变数对照表

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 ;

Figure 2. Table of label for each of regressors stored in WORK.POLYNAME
图 2. 回归变量卷标表

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 ;

Figure 3. The output data file with 5 observations and 22 variables
图 3. 输出数据文件，WORK.OUTPUT 只打印 22 个变量与 5 笔数据

Table 3. The contents of the data file MOD_HOMO
表 3. 档案 MOD_HOMO 的内容

	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

以下先介绍程序的内容与选项，然后再详述程序的架构与困难处。

程序的参数字于等号左侧而等号右侧为参数的默认值，表 4 介绍如何输入MACRO变量、%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 =).

以下用(Kunugi, Tamura, and Naito [1961])**的论文资料为例子：**

[例题3.1]：用稀释氢产生乙炔的新程序；应变数为转换正庚烷为乙炔的比例(P单位为%)，自变量有三反应温度(T单位为℃)，氢占正庚烷的穆尔比率(H单位为mole ratio %)，与接触时间(C单位为秒)。数据文件内容将打印在表5如下。

将自变量标准化再产生一完整的2次多项式数据储存在Ridge_S20.sas7bdat内容打印在表6而不含末4笔的数据存在Ridge_S16.sas7bdat，末4笔数据存在Ridge_S04.sas7bdat。

为减少篇幅我用 20 笔标准化后的数据来执行模型筛选与预测，但不指定预测档案。此时程序会将含有遗漏值的 4 笔数据当作需要被预测的数据，并且会存放在 2 个不同的数据文件，**Check_d01.sas7bdat 与 Missing.sas7bdat 而资料文件 N_of_miss_val.sas7bdat 则存入遗漏值的数目。用 MODEL_IN = NO 选项执行程序**%***Model_Selection*** 将得到以下 2 个图表；图 4 为资料文件 MMODEL.sas7bdat 的内容，图 5，打印出在各个模型中参考点与预估点的估计结果。

%***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**);

[注]：

1) 选项X_FULL在MODEL_IN = NO时可以不给，程序会用&X来取代。

2) F_MODEL = YES是要求程序行印包含全部变量的模型。

3) STAT选项是选择打印VIF (variance inflation factor)的值，因为除了全部模型与线性模型外p-value值均小于0.2。

当你执行完后会在log窗口中发现两个警告讯息

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.

我希望从SAS输出中抓取VIF的值，而SAS的REG程序中如不加入NOINT的选项则会输出函有截距的模型。但是我用的数据文件为标准化过的数据，因此NOINT的选项必需加入。另外的警告是在计算CP值时出现的，那是因为我想缩短程序因此就不去处里这个警告，请直接忽略它.要了解程序如何使用可以执行以下的程序。

1) 使用未标准化的数据：%***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) 使用未标准化的数据：%***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**);

我同时提供一SAS程序“Example of Model Selection.sas”以供模仿用。在使用前请先参考档案“README.TXT”。

Table 4. The table of input parameters of %Model_Selection
表 4. 程序%Model_Selection 的参数输入表

MACRO变量名称	解释
DADA_IN	原始资料文件名。
Y=	应变量(分析预测变量名称)。
X_LIN=	输入线性模型变量，如X1 X2 X3…
X_FULL=	输入完整的模型变量，不论&X为何必须输入，如果不输入则程序会自动选取以&X来代替。如果MODEL_IN=YES则用所有输入的模型变量为&X_FULL的值。
X=	自变量名称，变量与变量间以空格来分别。SAS用来做筛选用，可以与X_FULL不同。
RES_OUT = NO	“YES”为要求程序输出包含残差值的输出文件名，注意：如果MODEL_IN=YES则会为每一模型产生一个新档案，这会占去很大的储存空间，请谨慎使用！默认值为“NO”。
F_MODEL = NO	“NO”为要求程序在报表中不打印完整模式，因为如果变量数目太大，经由模型选择后可以让报表不会过度膨胀。因为完整模型不是选项亦或是X_FULL=X_LIN时是不必要包含完整模型因为X_FULL与X_LIN结果会相同。
STAT = VIF	伴随参数出现的统计量为何？共有“VIF”、“PVALUE”与“ALL”都选。如果输入错误则程序会同时打印“VIF”与“PVALUE”。
SLENTRY = 0.2	新进入模型的变量的PVALUE必须小于显著水平，默认值为“0.2”。
SLSTAY = 0.2	从模型移除变量的PVALUE必须大于显著水平，默认值为“0.2”。注意：选取变量从宽,移除变量从严。
CHECK = NO	模型选取完后是否要用来做预测？默认值为“NO”。
CHECK_D = CHK_DATA	储存要用来预测的数据文件名，存于& CHECK_D中，默认值为“CHK_DATA”。
ID = ID_ID	在& CHECK_D数据中用来分辨何笔数据为需要做预测的以及何笔数据是用来做“参考点”的，默认值为“ID_ID”。注意：& CHECK_D的值只能有3种，0 (不选取)，1 (参考点)或2 (预测点)。
ID2=	为非必要选项。此变量是用来辨识数据用。
OUT_DATA = NO	要求程序在有遗漏值(Missing Value)时是否输出无遗漏值的数据文件与预测档案。默认值为“NO”。
MODEL_IN = NO	如果选项是“YES”则程序会针对你给的模型作分析与预测。
MOD_NAME=	如果选项是“YES”则程序会用& MOD_NAME所指定的数据文件去分析，但是数据文件的变量必须是；第一个变量必须为模型名称而第二个变量必须是自变量(X)。
GROUP=	给分群变量名称(用在以后执行变异数一致性检测程序用)，如不给则程序会用应变量(Y)来做分群变数。

[注]：1) 如果 MODEL_IN = NO 则程序会自动产生一数据文件“MMODELS”包含模型名称(M_NAME)，自变量名称(X)，应变量名称(Y)，文件名(FILENAME)与组名(GROUP)。2) 程序会检查输入数据文件有否包含遗漏值(Missing Value)，会产生三个数据文件，CHECK_D01 (加入检察预测值的数据文件)，MISSING(储存遗漏值)与 N_of_miss_val (遗漏值的笔数)。3) 如果 CHECK = YES，则程序会检察是否有输入预测值的数据文件“&CHECK_D”如果没有则会去数据文件中找遗漏值然后存于“CHECK_D01”。注意：如果数据文件中有遗漏值则程序会主动将之加入预测档案中，所以请小心对待数据文件中的遗漏值。

Table 5. The contents of dataRidge_20.sas7bdat
表 5. 资料文件 Ridge_20.sas7bdat 不含末 4 笔的存在 Ridge_16.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

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

Figure 4. The data of file MMODELS.sas7bdat
图 4. 输出数据文件 MMODELS.sas7bdat 的内容

Table 6. The data of file, Ridge_S16.sas7bdat
表 6. 档案 Ridge_S20.sas7bdat 内容

	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

[注]：1) 最后4笔数据为需要预测的值；2) 变数TH等为交叉相乘项，而T2等为平方项。

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

Figure 5. The predictions of referent and predicted points
图 5. 各个模型的预测值与参考点值

4. 总结与展望

模型选取有一个重要的原则，加入变量从宽去除变量从严。当分析者拿到数据时感到最困恼的是如何选取变量。这两支程序基本上可以解决过早删除变量。因为当你先建立多项式模型，再执行变量筛选，你不大容易将一个变量完整的从模型中删除，所以无须担心重要变量被意外的移出模型。这可以节省很多时间与精力。再者程序提供了总结后的输出报表，其中包含各个模型的数据与各种基本的统计量，使用者可以很快速的决定哪些模型可供使用。这将大大的缩减判断的时间，因此可以很容易的执行下一步的检验。而程序产生的输出文件可以让使用者非常容易的使用，而毋须烦恼变量太多无法快速的输入。当然多项式模型必然产生共线性的问题，因此在模型确定后再处里共线性的问题。由于目前尚未有完美的解答，因此我只希望能总结各家提出的选项再会诊后集结为一程序可供使用。

在下一篇我将讨论如何检定模型的基本假设(常态分配、变异数的一致性、与数据的独立性)。程序是基于检定统计量而不是用残差分析图去作判定。过往会使用图形去判定的原因是因为计算机太慢又太贵，故无法提供解决方法而非方法不存在，详细内容请参考 ***Wang*** [2]。由于程序太长，请参考附件档案。我也衷心的希望制作软件的先进们能制作出正确又方便的工具，能让统计分析方便又能得到正确的结果。

参考文献 (References)

[1] Montgomery, D.C., Peck, E.A. and Vining, G.G. (2006) Introduction to linear regression analysis. 4th Edition, Willey, New York.

[2] Wang, K.-J. (2013) Notes for regression analysis. Tamkang University, New Taipei.