五． 統計分析

本章以 S-Plus 提供的例子來解說一些統計分析方法，至於統計理論則省略。

1．敘述統計 Summary Statistics

1.1 Summary Statistics

計算敘述統計量，如平均值，中位數，變異數，總和等。

例:

data set: exair 此資料測量 New York 的 ozone 濃度、風速、溫度和輻射量, 共 111 個觀察值。

a. File/Open/samples directory/exair.sdd

b. Statistics/Data Summaries/Summary Statistics

c. Data 頁: 選擇 data set 及有興趣的變數或選擇<ALL>，鍵入欲將結果存檔之檔名 summary.air

d. Statistics 頁: 選擇統計量 /OK

e.

```
    ***  Summary Statistics for data in:  exair ***

             ozone radiation
    Min:   1.0000000    7.0000
   Mean:   3.2477838  184.8018
 Median:   3.1413807  207.0000
    Max:   5.5178484  334.0000
Total N: 111.0000000  111.0000
  NA's :   0.0000000    0.0000
Variance:   0.7928069 8308.7422
```

d. 觀察 Explorer Object Window

1.2 Crosstabulations

計算類別或因子變數所有組合的個數。

例:

data set: exclaims 此資料是有關 insurance claims，128 個資料點爲前三個預測變數”age”,”car.age”,”type”所有可能組合，number 爲每一格 claim 的個數，cost 爲 claim 的平均成本。

a. File/Open/sample directory/exclaims.sdd

b. Statistics/Data Summaries/Crosstabulations

c. Model 頁:

Variables: 選 car.type, type (利用 Ctrl)

Counts Variable: 選 number /OK

d.

```
    *** Crosstabulations ***
Call:
crosstabs(formula = number ~ car.age +
    type, data = exclaims, na.action
     = na.fail, drop.unused.levels =
    T)
8942 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
car.age|type
     |A    |B    |C    |D    |RwTtl|
------+-----+-----+-----+-----+-----+
0-3   | 391  |1538  |1517  | 688  |4134 |
     |0.0946|0.3720|0.3670|0.1664|0.462|
     |0.3081|0.3956|0.5598|0.6400|     |
     |0.0437|0.1720|0.1696|0.0769|     |
------+-----+-----+-----+-----+-----+
省略….
------+-----+-----+-----+-----+-----+
ColTtl|1269  |3888  |2710  |1075  |8942 |
     |0.14  |0.43  |0.30  |0.12  |     |
------+-----+-----+-----+-----+-----+
Test for independence of all factors
    Chi^2 = 588.3 d.f.= 9 (p=0)
    Yates' correction not used
```

e. The test for independence results reported below the table indicate that the percentage of observations in each cell is significantly different from the product of the total row percentage and total column percentage. Thus there is an interaction between the car age and type, which influence the number of claims. That is, the effect car age on number of claims varies by car type.

1.3 Corrections

計算共變異數和相關係數。

例:

data set: exair

a. File/Open/sample directory/exair.sdd

b. Statistics/Data Summaries/Correlations

c. variables 選<ALL> /OK

d.

```
    ***  Correlations for data in:  exair ***

             ozone radiation  temperature    wind
     ozone  1.00000   0.42201   0.75310   -0.59893
  radiation  0.42201   1.00000   0.29409   -0.12737
temperature  0.75310   0.29409   1.00000   -0.49715
      wind -0.59893  -0.12737  -0.49715    1.00000
```

e. 比較 Scatter matrix plot.

2．樣本比較

2.1 One Sample

t-test	假設母體爲 Normal 分配，test 母體平均數是否爲某一值。要用 t-test 時，需先檢查 sample 是否爲 Normal 分配，常用的方法是 qqplot。
Wilcoxon singed rank test	無母數的方法來 test 母體平均數是否爲某一值，不用假設母體的分配。
Kolmogorov-Smirnov GOF	Test data 是否來自某一假設的分配，較適合連續型變數的 data。
Chi-square GOF	用 Pearson's chi-square statistic test data 是否來自某一假設的分配。可適用各種型式的 data，但樣本數小於 50 的大樣本，假如分配是離散型的，chi-square 是唯一有作 test 的。

2.2 Two Samples

t-test	Test 兩個母體的平均值是否相等，此 t-test 包含 paired t-test，two-sample t-test with unequal variance。
Wilcoxon rank test	無母數方法 test 兩個母體的平均值是否相等。Alternative Hypothesis 是觀察值來自 shape 相同但 location 不同的分配。
Kolmogrov-Smirnov GOF	Test 兩組觀測值是否來自同一分配，KS GOF 假設樣本爲 random samples 且互相獨立。

2.3 K Samples

One-way ANOVA	是一因子的變異數分析，假設 k 組樣本(一因子 k 個 level)互相獨立且來自於 Normal 分配。

	Model 爲 $y_{ij} = \mu_i + \epsilon_{ij}\ i = 1, \cdots, k,\ \ j = 1, \cdots, J_i$
Kruskal-Wallis rank test	無母數方法的 one-way ANOVA，不需假設 Normal 分配
Friedman rank test	此 test 適合 data 來自 unreplicated complete block design.

2.4 Counts and Proportions

Binomal test	Test data 是否來自 Binomial(p)。
Proportions parameters	爲一 Chi-square test，test binomial data 是否有一指定的 proportion 參數，或兩個 binomial 樣本是否有相同的參數。
Fisher's exact test	Test 列聯表的欄和列是否獨立，適用於小樣本。
McNamar's test	Test 列聯表的欄和列是否獨立，適合 matched pair data，每個觀察值並不獨立。 Survive.B Die.B Survive.A 90 16 Die.A 5 510
Mantel-Haenszal test	Test 三維列聯表的獨立性。
Chi-square test	Pearson's chi-square test on 二維列聯表。

3．迴歸分析 Regression

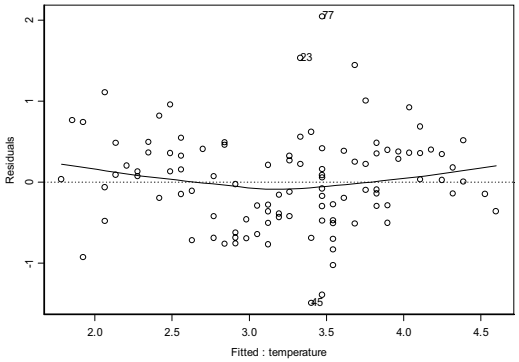
3.1 線性迴歸 linear regression

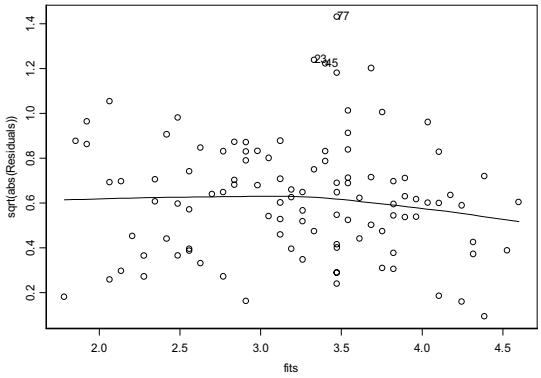
a. File/Open/exair.sdd
b. Scatterplot of temperature and ozone
c. Statistics/Regression/Linear
d. Formula: `ozone~temperature`
e. Plot page: seven main diagnostic plots
f. Result page: Long Output, ANOVA, Correlation matrix of estimates
g. Output

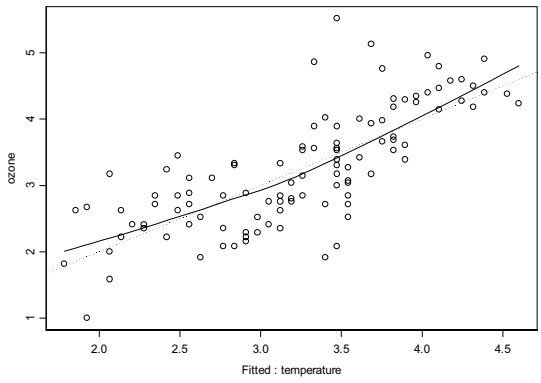
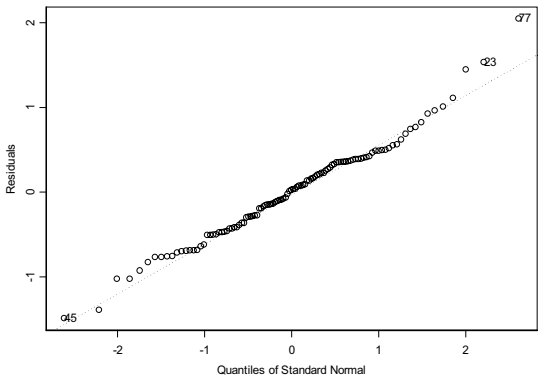
```
    *** Linear Model ***

Call: lm(formula = ozone ~ temperature, data = exair, na.action = na.exclude)
Residuals:
   Min     1Q  Median     3Q   Max
 -1.49 -0.4258 0.02521 0.3636 2.044

Coefficients:
             Value Std. Error  t value Pr(>|t|)
(Intercept)  -2.2260   0.4614   -4.8243   0.0000
temperature   0.0704   0.0059   11.9511   0.0000
```

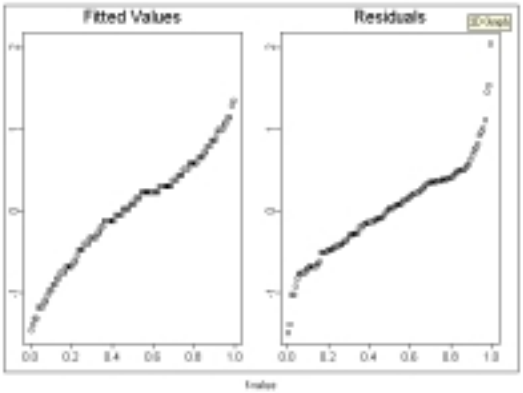
```
Residual standard error: 0.5885 on 109 degrees of freedom
Multiple R-Squared: 0.5672
F-statistic: 142.8 on 1 and 109 degrees of freedom, the p-value is 0

Correlation of Coefficients:
          (Intercept)
temperature -0.9926


Analysis of Variance Table
Response: ozone
Terms added sequentially (first to last)
            Df Sum of Sq  Mean Sq  F Value Pr(F)
temperature   1  49.46178 49.46178 142.8282    0
  Residuals 109  37.74698  0.34630
```

h. Diagnostic Plots

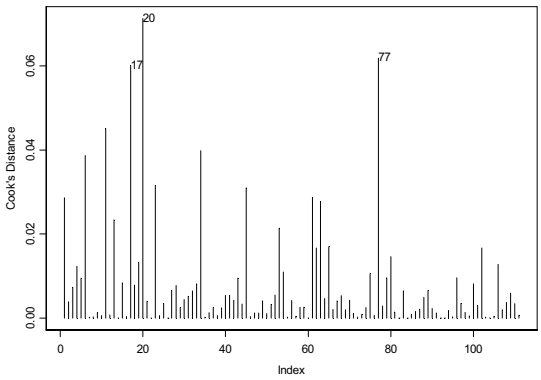
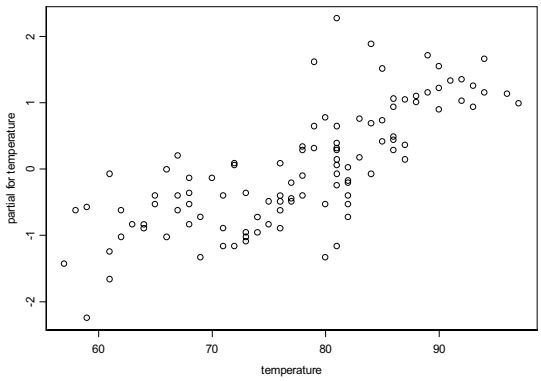
<table>
<tr>
<td> 

show no obvious pattern, although five observations appear to be outliers</td>
<td>Residuals vs Fitted values:
1. reveals unexplained structure left in the rediduals.
2. A strong model should appear as nothing but noise.</td>
</tr>
<tr>
<td> 
</td>
<td>Square root of absolute residuals vs fitted values:
1. Identify outliers
2. Visualize structure in the residuals</td>
</tr>
</table>

<table>
<tr>
<td> 
</td>
<td>Response vs Fit:
1. gives a good idea of how well the model has captured the broad outlines of the data.
2. dash line: y=fitted values
3. useful for checking for the constant variance assumption of the model.</td>
</tr>
<tr>
<td> 

gives no reason to doubt that the residuals are normally distributed</td>
<td>Residuals Normal QQ:
1. provides a visual test of the assumption that the model’s error are normally distributed.
2. strong evidence that errors are normal if the ordered residuals cluster along the superimposed QQ line.</td>
</tr>
<tr>
<td> 

shows a weakness in this model, the spread of the residuals is actually greater than the spread in the original data, ignore the five outlying residuals</td>
<td>Residual-Fit Spread:
1. compares the spread of the fitted values with the spread of the residuals.
2. This is a visual analog of the multiple R-square
3. Since the model is an attempt to explain the variation in the data, you hope that the spread in the fitted values is much greater than in the residuals.</td>
</tr>
</table>

 shows four or five heavily influential observations.	Cook's Distance: 1. measure of the influence of individuals observations on the regression coefficients.
	Partial Residuals: 1. a plot of (Ri+BkXik)vs Xik Ri: ordinary residuals Bk: regression coefficient estimate for the kth predictor Xik : ith observation of the kth predictor 2. useful for detecting nonlinearities and for identifying possible causes of unduly large residuals

i. 結果推論:

這個 data 做出來的迴歸線是滿合適且顯著的，且 residuals 看起來也適合 normal 分配，我們覺得在給定 temperature 之下，用這條迴歸線來估計 ozone concentration 是適當的。但是 R-square 只有 57% - 迴歸線只解釋了 data 57%的變異，我們認爲需考慮其它會影響 ozone concentration 的變數。

3.2 逐步迴歸分析 Stepwise linear regression

逐步迴歸析是從多個 variables 中選取重要的變數到迴歸 model 裡，有 forward, backward, stepwise 三種 procedure. S-Plus 採用 Mallow's Cp 統計量來決定變數是否用到 model 中.

a. Statistics/Regression/Stepwise
b. File/Open/exair.sdd
c. Upper Formula: ozone~radiation+temprature+wind
d. Lower Formula: ozone~1 (只包含斜率項)
e. Output

```
    *** Stepwise Regression ***

    *** Stepwise Model Comparisons ***
Start:  AIC= 29.9302
```

```
 ozone ~ radiation + temperature + wind

Single term deletions

Model:
ozone ~ radiation + temperature + wind

scale:  0.2602624

           Df Sum of Sq      RSS      Cp
    <none>               27.84808 29.93018
  radiation  1    4.05928 31.90736 33.46893
temperature  1  17.48174 45.32982 46.89140
       wind 1   6.05985 33.90793 35.46950

    *** Linear Model ***

Call: lm(formula = ozone ~ radiation + temperature + wind, data = exair, na.action
= na.exclude
    )
Residuals:
    Min     1Q   Median    3Q   Max
 -1.122 -0.3764 -0.02535 0.3361 1.495

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) -0.2973  0.5552   -0.5355  0.5934
  radiation  0.0022  0.0006    3.9493  0.0001
temperature  0.0500  0.0061    8.1957  0.0000
      wind -0.0760  0.0158   -4.8253  0.0000

Residual standard error: 0.5102 on 107 degrees of freedom
Multiple R-Squared: 0.6807
F-statistic: 76.03 on 3 and 107 degrees of freedom, the p-value is 0
```

f. Criterion (x1, x2,x3) 選最接近 p 值但最小的 Cp 所對應的 model

#regressor	p	regressors	Cp
0	1	0	
1	2	x1	
1	2	x2	
1	2	x3	
2	3	x1 x2	
2	3	x1 x3	
2	3	x2 x3	
3	4	x1 x2 x3	

3.3 其它

Regression		Response	Predictors	criterion
Generalized linear models 廣義線性模型		General	Linear combination of the predictors	ML
1	Log-linear regression	Counts	Linear combination of the predictors	Poisson ML
2	Logistic regression	Binary	Logistic link	Binomial ML
3	Probit regression	Binary	Probit link	Binomial ML

Generalized additive models	General	A sum of nonparametric smooth univariate functions of the predictors	
Nonlinear regression 非線性迴歸	Continuous	Nonlinear function of the predictors	LS

ML: Maximum likelihood, LS: Least-square

4．變異數分析 ANOVA

ANOVA is generally used to explore the influence of one or more categorical variables upon a continue response.

4.1 One-way ANOVA

Data: A single continuous response variable is measured a number of times for each of several levels of some experimental factor.

假設: a. Samples 互相獨立

b. Observations 是 Normal 分配。

Model: $y_{ij} = \mu_i + \epsilon_{ij} \;\; i = 1, \cdots, k, \;\; j = 1, \cdots, J_i$

H0: mean values for all of the groups are equal. (各組的平均沒有差異)

統計量: F-statistics

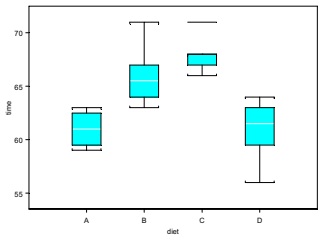
例子: exblood.sdd

a. one factor with four levels: diet (A,B,C,D), one response variable: time

b. Box plots.

Diets A 和 D 相似。

diets B,C 對 median response 有較大的變異。



初步結果: diets 對 blood coagulation time 有影響。

c. Statistics/Compare Samples/k Samples/

One-way ANOVA

Variable: time

Grouping variable: diet

```
    ***  One-Way ANOVA for data in time by diet ***

Call:
   aov(formula = structure(.Data = time ~ diet, class =
     "formula"), data = exblood)
Terms:
              diet   Residuals
 Sum of Squares  228       112
Deg. of Freedom    3        20

Residual standard error: 2.366432
Estimated effects may be unbalanced

          Df Sum of Sq Mean Sq F Value        Pr(F)
    diet  3      228    76.0    13.57143 0.00004658471
Residuals 20      112     5.6
```

d. 結論: p-value=0.000047(highly significant). Diets 確實會影響 blood coagulation times.

4.2 Fixed Effects ANOVA

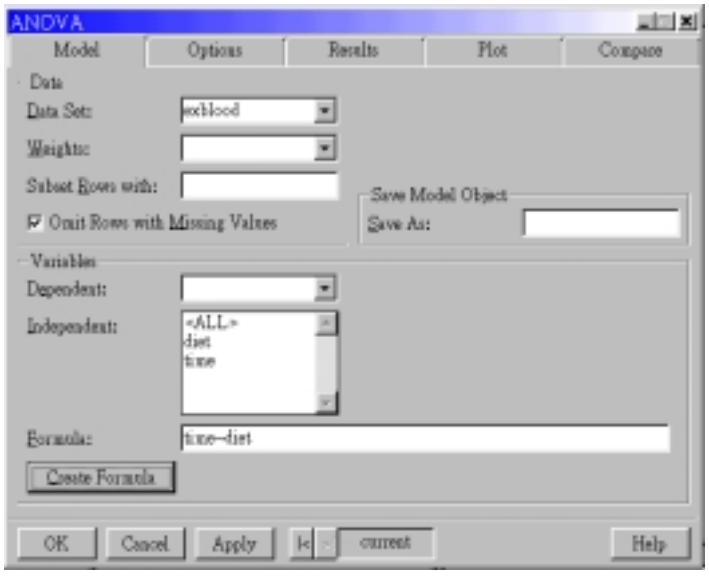
例子: exblood.sdd

a. one factor with four levels: diet (A,B,C,D), one response variable: time

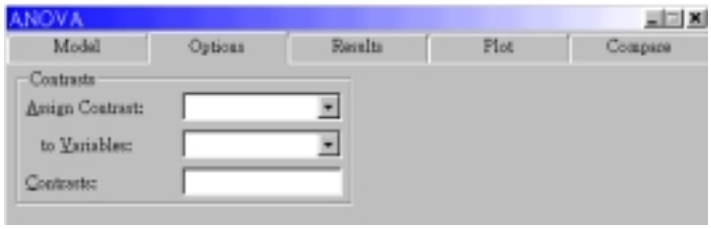
b. Statistics/ANOVA/Fixed Effects

Model 選項:

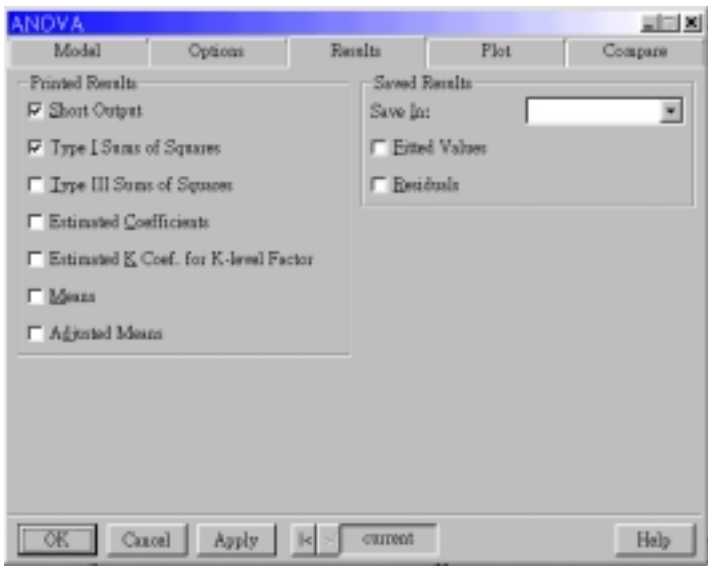
建立 model →



Options 選項:



Results 選項:



Plot 選項: 同 regression

4.3 Random Effects ANOVA

例子: expigmnt.sdd

a. 15 種顏料中，各抽一種，收集 2 組樣本，各作兩種水份測試。

We fit a random effects ANOVA model to assess the within-batch and between-batch variation.

Moisture		Batch			
		B1	B2	…..	B15
S1	T1	40	26	…..	39
	T2				
S2	T1				
	T2				

b. Statistics/ANOVA/Random Effects

Formula: Moisture~Batch + Sample %in% Batch

Term Category of nested effect: 先選 variables 再 term categories

c. Output

```
*** Analysis of Variance Model ***

Short Output:
Call:
   raov(formula = Moisture ~ Batch + Sample %in% Batch, data = expigmnt,
na.action =
     na.exclude)

Terms:
                Batch Sample %in% Batch Residuals
 Sum of Squares 1210.933         869.750    27.500
Deg. of Freedom      14              15       30

Residual standard error: 0.9574271
Estimated effects are balanced

                   Df Sum of Sq  Mean Sq Est. Var.
          Batch    14  1210.933 86.49524   7.12798
Sample %in% Batch  15   869.750 57.98333  28.53333
       Residuals    30    27.500  0.91667   0.91667
```

4.4 Multiple Comparisons

變異數分析通常用來比較 treatments 在 response 上 effects 是否顯著，之後，我們有興趣的是想知道對應不同的 treatment groups 在 response 是否存在顯著差異，假如差異存在的話，這差異的大小如何。Multiple comparisons 是用來檢定這些 effects 是否一樣，並且估計 treatment effects。

a. Statistics/ANOVA/Multiple Comparisons

b. 在 One-Way ANOVA 例子中，我們得到一個結論是 diet 確實會影響 blood coagulation time，下一步是想知道哪一個 diets 和其它 diets 不同。

c. 做完 one-way ANOVA 將結果存在 anova.blood
(Statistics/Compare Samples/k samples/One-Way ANOVA)

d. Model Object: `anova.blood`
Method: `Tukey`

e. output

```
95 % simultaneous confidence intervals for specified
linear combinations, by the Tukey method

critical point: 2.7987
response variable: time

intervals excluding 0 are flagged by '****'

     Estimate Std.Error Lower Bound Upper Bound
A-B -5.00e+000      1.53      -9.28     -0.725 ****
A-C -7.00e+000      1.53     -11.30     -2.720 ****
A-D -1.64e-014      1.45      -4.06      4.060
B-C -2.00e+000      1.37      -5.82      1.820
B-D  5.00e+000      1.28       1.42      8.580 ****
C-D  7.00e+000      1.28       3.42     10.600 ****
```

f. 從表中及 plot of the confidence intervals 可看出, diets A 和 D 所產生 blood coagulation times 比 diets B 和 C 顯著不同。