

R 軟體資料分析應用：多變項迴歸分析以及交互作用

尤子芸 副統計分析師

本期 eNews 將延續前兩期 eNews 中所討論過的簡單線性迴歸分析以及邏輯斯迴歸分析，與各位討論如何使用 R 軟體進行多變項迴歸分析，又稱複迴歸分析，包含線性迴歸分析與邏輯斯迴歸分析以及交互作用項的使用及解釋。

本次分析同樣使用 CVD_ALL 這組資料作呈現，可以前往下列網址進行下載 http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv，資料詳述內容定義可至 http://biostat.tmu.edu.tw/attachment/25_help.docx 文件內觀看。

一、多變項線性迴歸分析 (Multiple linear regression analysis)

在本中心 eNews 第 30 期中所介紹的簡單線性迴歸分析為當所分析之變項為一連續型依變項 (Dependent variable) 與自變項 (Independent variable) 之線性相關時所進行的分析，而在本篇中將進一步討論多變項線性迴歸分析，即進行變項為一連續型依變項與多個自變項的線性相關之分析。

簡單線性迴歸模型： $y = \beta_0 + \beta_1 x_1 + \varepsilon$

多變項線性迴歸模型： $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$

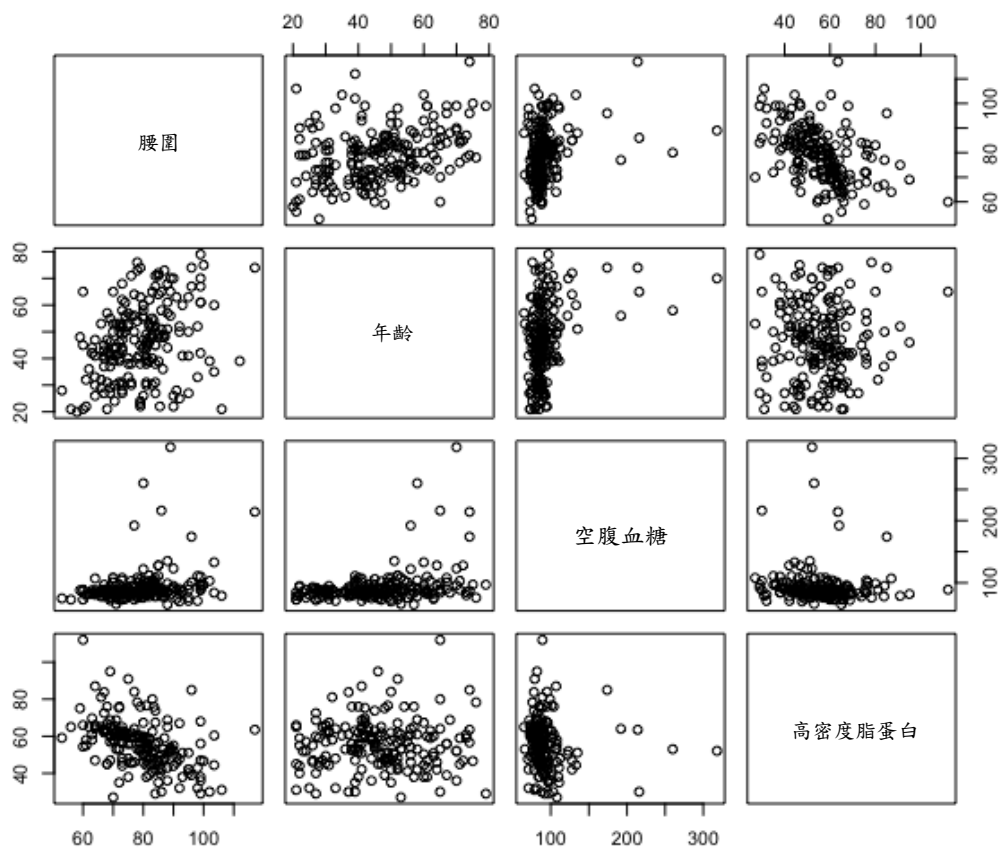
β_0 為常數， β_i 為迴歸係數， $i = 1$ 到 n ， ε 為誤差

(一) 多變項線性迴歸分析之 R 程式語法

1. 線性迴歸所使用的 R 語法為 `lm()`
2. 使用 CVD_ALL 資料，若我們想以年齡、空腹血糖以及高密度脂蛋白三個變項預測腰圍，則程式碼如下：

【程式碼】

```
#匯入資料cvd_all
cvd_all = read.csv("G:/CVD_all.csv")
#因本資料筆數較多，隨機選取其中的 200 筆資料作為範例進行分析
cvd_200=cvd_all[sample(1:nrow(cvd_all),200),]
#多變項線性迴歸
#依變項：腰圍
#自變項：年齡、空腹血糖、高密度脂蛋白
model=lm(腰圍~ 年齡+空腹血糖+高密度脂蛋白, data=cvd_200)
#利用 pairs() 畫出關心的變項之間的交叉 X-Y 散佈圖
pairs(腰圍~ 年齡+空腹血糖+高密度脂蛋白)
#利用 summary() 看更詳細的分析結果
summary(model)
```



交叉配對 X-Y 散佈圖

【Output】**Call:**

```
lm(formula = 腰圍 ~ 年齡 + 空腹血糖 + 高密度脂蛋白,
    data = cvd_200)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.1574	-5.9363	-0.6351	4.9444	29.2015

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	84.47111	3.97881	21.230	< 2e-16 ***
年齡	0.24572	0.04972	4.942	1.74e-06 ***
空腹血糖	0.04336	0.02437	1.779	0.0769 .
高密度脂蛋白	-0.37111	0.05019	-7.394	4.95e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.948 on 183 degrees of freedom

(13 observations deleted due to missingness)

Multiple R-squared: 0.341, Adjusted R-squared: 0.3302

F-statistic: 31.57 on 3 and 183 DF, p-value: < 2.2e-16

【分析結果】

從 output 的結果可以看出，假設在顯著水準為 0.05 時，年齡及高密度脂蛋白與腰圍之間是有顯著的線性相關，在控制空腹血糖與高密度脂蛋白下，年齡每增加一歲，腰圍會增加 0.24572 公分；在控制年齡與空腹血糖下，高密度脂蛋白每增加一單位，腰圍會減少 0.37111 公分，而截距項為 84.47111，從分析結果中可以得到的迴歸模型如下：

$$\widehat{y}_{\text{腰圍}} = 84.47111 + 0.24572(\text{年齡}) + 0.04336(\text{空腹血糖}) - 0.37111(\text{高密度脂蛋白})$$

但從判定係數 Adjusted R-squared=0.3302 可以得知使用此預估模型，自變項對於解釋依變項變異的能力不是很好。因此，我們可以透過模型選取的方式來找最適模型。

(二) 模型變項選取

在進行迴歸分析時，透過選取變項進入模型的方式以得到最佳的迴歸模式，大部份的情形是有多個自變項可以選擇使用在迴歸方程式中，通常想要找到的是能夠以較少的自變項就可以解釋整個迴歸模式最大量，然而，到底我們應該選取多少個自變項，又應該選擇哪些變項呢？除了透過文獻與實際臨床經驗選擇變項之外，還能透過依照變項對模型的解釋力進行篩選，將變項選入模型中，以下整理三個常見選擇變項進入迴歸模型的方式進行介紹：

1. 向前選取法 (Forward selection method)

自變項的選取是以達到統計顯著水準的變數，依解釋力的大小，依序選取進入迴歸模型中，以逐步增加的方式，完成選取。

2. 向後選取法 (Backward selection method)

先將所有變項放入迴歸模型中求出一個迴歸方程式，接著依序將最小解釋力的變項刪除，直到所有未達顯著的自變項都刪除為止。

3. 逐步選取法 (Stepwise selection method)

此方法結合前述兩方法：向前與向後選取法，首先，會選取自變項中與依變項相關性最大者，接著選擇剩下的自變項中，部份相關係數與依變項較高者 (解釋力較大者)，每新增一個自變項，就利用向後選取法檢驗迴歸模型中，是否有需要刪除的變數，透過向前選取，選擇放入模型的變項，向後選取進行檢驗，直到所有選取的變項都達顯著水準為止，就會得到迴歸的最佳模型。

本篇在選取變項進入模型時，用來衡量變項解釋力的指標為 AIC，AIC 值越小，模型越好。執行模型選取的 R 程式如下：

【程式碼】

```
#指定所有變項納入模型中的 full model
fullmodel=lm(腰圍~ 心血管疾病+年齡+性別+收縮壓+舒張壓+空腹血糖+高密度脂蛋白+三酸甘油酯+檳榔+飲酒+家族病史+抽菸量, data=cvd_200)
(因資料檔中的抽菸及抽菸量是高度相關的變項，僅選擇其一做分析，在本範例中使用抽菸量做為自變項分析)
summary(fullmodel) #利用 summary()看更詳細分析結果
nullmodel=lm(腰圍~ 1, data=cvd_200) #建立只有截距項的空線性迴歸(null model)
## 1. 向前選取法
#使用 step()，設定 direction 為 forward，會自動一個一個把變數丟進去
forward.lm = step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel),
direction="forward") # (一定要加上界 upper=fullmodel，不可以不加)
summary(forward.lm)
```

【Output】

Start: AIC=802.92

腰圍 ~ 1

	Df	Sum of Sq	RSS	AIC
+ 收縮壓	1	3689.4	16520	771.25
+ 性別	1	2806.8	17402	779.94
+ 年齡	1	2465.7	17744	783.19
+ 高密度脂蛋白	1	2435.2	17774	783.47
+ 舒張壓	1	2194.9	18014	785.72
+ 心血管疾病	1	633.5	19576	799.60
+ 三酸甘油酯	1	519.5	19690	800.57
+ 抽菸量	1	486.8	19722	800.84
+ 檳榔	1	274.1	19935	802.64
<none>			20209	802.92
+ 飲酒	1	131.1	20078	803.83
+ 空腹血糖	1	92.9	20116	804.15
+ 家族病史	1	55.0	20154	804.46

Step: AIC=771.25

腰圍 ~ 收縮壓

	Df	Sum of Sq	RSS	AIC
+ 性別	1	1567.86	14952	756.60
+ 高密度脂蛋白	1	1473.51	15046	757.65
+ 年齡	1	880.71	15639	764.10
+ 抽菸量	1	425.94	16094	768.89
+ 檳榔	1	383.72	16136	769.33
+ 心血管疾病	1	286.54	16233	770.33
+ 三酸甘油酯	1	260.21	16260	770.60
<none>			16520	771.25
+ 飲酒	1	135.09	16385	771.88
+ 舒張壓	1	60.01	16460	772.64
+ 家族病史	1	51.99	16468	772.73
+ 空腹血糖	1	13.84	16506	773.11

Step: AIC=756.6

腰圍 ~ 收縮壓 + 性別

	Df	Sum of Sq	RSS	AIC
+ 年齡	1	1194.31	13758	744.70
+ 高密度脂蛋白	1	583.88	14368	751.95
+ 心血管疾病	1	509.67	14442	752.81
<none>			14952	756.60
+ 三酸甘油酯	1	161.02	14791	756.79
+ 檳榔	1	72.68	14879	757.79
+ 家族病史	1	17.49	14934	758.40
+ 飲酒	1	12.34	14940	758.46
+ 抽菸量	1	7.31	14945	758.52
+ 空腹血糖	1	6.73	14945	758.52
+ 舒張壓	1	0.85	14951	758.59

Step: AIC=744.7

腰圍 ~ 收縮壓 + 性別 + 年齡

	Df	Sum of Sq	RSS	AIC
+ 高密度脂蛋白	1	483.99	13274	740.72
<none>			13758	744.70
+ 心血管疾病	1	117.13	13641	745.27
+ 檳榔	1	102.07	13656	745.45
+ 空腹血糖	1	73.90	13684	745.80
+ 家族病史	1	30.01	13728	746.33
+ 飲酒	1	25.29	13732	746.39
+ 三酸甘油酯	1	10.03	13748	746.58
+ 抽菸量	1	1.20	13756	746.68
+ 舒張壓	1	0.06	13758	746.70

```

Step: AIC=740.72
腰圍 ~ 收縮壓 + 性別 + 年齡 + 高密度脂蛋白
      Df Sum of Sq  RSS   AIC
+ 空腹血糖    1  190.463 13083 740.30
+ 檳榔        1  160.762 13113 740.68
<none>                    13274 740.72
+ 心血管疾病  1  128.934 13145 741.09
+ 三酸甘油酯  1   38.215 13236 742.24
+ 家族病史    1   36.294 13237 742.26
+ 飲酒        1   10.950 13263 742.58
+ 抽菸量      1    0.931 13273 742.70
+ 舒張壓      1    0.026 13274 742.72

Step: AIC=740.3
腰圍 ~ 收縮壓 + 性別 + 年齡 + 高密度脂蛋白 + 空腹血糖
      Df Sum of Sq  RSS   AIC
<none>                    13083 740.30
+ 檳榔        1  144.820 12938 740.44
+ 心血管疾病  1   95.852 12987 741.08
+ 家族病史    1   29.624 13054 741.92
+ 三酸甘油酯  1    4.953 13078 742.24
+ 抽菸量      1    2.750 13080 742.27
+ 飲酒        1    2.087 13081 742.28
+ 舒張壓      1    1.271 13082 742.29

Call:
lm(formula = 腰圍 ~ 收縮壓 + 性別 + 年齡 + 高密度脂蛋白 + 空腹血糖, data = cvd_200)

Residuals:
    Min       1Q   Median       3Q      Max
-37.917  -6.030  -0.053   5.816  25.858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.36694    6.52935   9.858 < 2e-16 ***
收縮壓       0.13596    0.03743   3.632  0.000378 ***
性別         5.20302    1.63732   3.178  0.001780 **
年齡         0.22259    0.05755   3.868  0.000159 ***
高密度脂蛋白 -0.17732    0.06523  -2.719  0.007277 **
空腹血糖     -0.03604    0.02354  -1.531  0.127745
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 9.015 on 161 degrees of freedom
Multiple R-squared:  0.3526, Adjusted R-squared:  0.3325
F-statistic: 17.54 on 5 and 161 DF, p-value: 7.576e-14

```

【分析結果 1.】

由上述結果可看出，最終模型為：

$$\widehat{y}_{\text{腰圍}} = 64.36694 + 0.13596(\text{收縮壓}) + 5.20302(\text{性別}) + 0.22259(\text{年齡}) \\ - 0.17732(\text{高密度脂蛋白}) - 0.03604(\text{空腹血糖})$$

其 Adjusted $R^2=0.3325$ ，在假設顯著水準為 0.05 下，有顯著的變項為收縮壓、性別、年齡以及高密度脂蛋白，其中以性別的影響最大，在控制其他變項的情況下，男性的腰圍會比女性的腰圍多 5.2 公分。

【程式碼】**## 2. 向後選取法**

#使用 step()，設定 direction 為 backward，會自動一個一個把變數移除，直到 AIC 為最小值

```
backward.lm = step(fullmodel, scope=list(upper=fullmodel), direction="backward")
summary(backward.lm)
```

【Output】

Start: AIC=748.69

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史 + 抽菸量

	Df	Sum of Sq	RSS	AIC
- 舒張壓	1	6.00	12657	746.77
- 三酸甘油酯	1	23.56	12674	747.00
- 抽菸量	1	23.64	12674	747.00
- 飲酒	1	36.29	12687	747.17
- 家族病史	1	46.77	12698	747.31
- 空腹血糖	1	69.32	12720	747.60
- 心血管疾病	1	140.81	12792	748.54
<none>			12651	748.69
- 檳榔	1	287.61	12938	750.44
- 性別	1	601.93	13253	754.45
- 高密度脂蛋白	1	653.52	13304	755.10
- 收縮壓	1	665.54	13316	755.25
- 年齡	1	852.86	13504	757.58

Step: AIC=746.77

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史 + 抽菸量

	Df	Sum of Sq	RSS	AIC
- 抽菸量	1	21.09	12678	745.05
- 三酸甘油酯	1	26.42	12683	745.12
- 飲酒	1	38.20	12695	745.27
- 家族病史	1	48.33	12705	745.41
- 空腹血糖	1	65.97	12723	745.64
- 心血管疾病	1	136.32	12793	746.56
<none>			12657	746.77
- 檳榔	1	284.60	12941	748.48
- 性別	1	601.52	13258	752.52
- 高密度脂蛋白	1	652.98	13310	753.17
- 年齡	1	866.89	13524	755.83
- 收縮壓	1	981.44	13638	757.24

Step: AIC=745.05

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史

	Df	Sum of Sq	RSS	AIC
- 三酸甘油酯	1	29.77	12708	743.44
- 飲酒	1	58.32	12736	743.81
- 空腹血糖	1	58.97	12737	743.82
- 家族病史	1	59.46	12737	743.83
- 心血管疾病	1	141.91	12820	744.91
<none>			12678	745.05
- 檳榔	1	264.14	12942	746.49
- 性別	1	586.50	13264	750.60
- 高密度脂蛋白	1	643.92	13322	751.32
- 年齡	1	846.67	13525	753.84
- 收縮壓	1	1006.19	13684	755.80

Step: AIC=743.44

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 檳榔 + 飲酒 + 家族病史

	Df	Sum of Sq	RSS	AIC
- 家族病史	1	52.48	12760	742.13
- 飲酒	1	57.71	12765	742.20
- 空腹血糖	1	96.52	12804	742.70
- 心血管疾病	1	144.13	12852	743.32
<none>			12708	743.44
- 檳榔	1	238.88	12947	744.55
- 性別	1	618.29	13326	749.37
- 高密度脂蛋白	1	634.68	13342	749.58
- 年齡	1	819.91	13528	751.88
- 收縮壓	1	1042.65	13750	754.61

Step: AIC=742.13

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 檳榔 + 飲酒

	Df	Sum of Sq	RSS	AIC
- 飲酒	1	62.27	12822	740.94
- 空腹血糖	1	103.93	12864	741.48
- 心血管疾病	1	131.21	12891	741.84
<none>			12760	742.13
- 檳榔	1	223.41	12984	743.03
- 高密度脂蛋白	1	625.10	13385	748.11
- 性別	1	660.52	13421	748.56
- 年齡	1	821.12	13581	750.54
- 收縮壓	1	1042.00	13802	753.24

Step: AIC=740.94

腰圍 ~ 年齡 + 性別 + 心血管疾病 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 檳榔

	Df	Sum of Sq	RSS	AIC
- 心血管疾病	1	116.03	12938	740.44
- 空腹血糖	1	139.11	12962	740.74
<none>			12822	740.94
- 檳榔	1	164.99	12987	741.08
- 性別	1	598.28	13421	746.56
- 高密度脂蛋白	1	657.72	13480	747.29
- 年齡	1	818.99	13641	749.28
- 收縮壓	1	1104.64	13927	752.74

Step: AIC=740.44

腰圍 ~ 年齡 + 性別 + 收縮壓 + 空腹血糖 + 高密度脂蛋白 + 檳榔

	Df	Sum of Sq	RSS	AIC
- 檳榔	1	144.82	13083	740.30
<none>			12938	740.44
- 空腹血糖	1	174.52	13113	740.68
- 性別	1	565.82	13504	745.59
- 高密度脂蛋白	1	654.86	13593	746.69
- 收縮壓	1	1129.92	14068	752.43
- 年齡	1	1239.84	14178	753.73

Step: AIC=740.3

腰圍 ~ 年齡 + 性別 + 收縮壓 + 空腹血糖 + 高密度脂蛋白

	Df	Sum of Sq	RSS	AIC
<none>			13083	740.30
- 空腹血糖	1	190.46	13274	740.72
- 高密度脂蛋白	1	600.56	13684	745.80
- 性別	1	820.60	13904	748.46
- 收縮壓	1	1072.00	14155	751.45
- 年齡	1	1215.49	14299	753.14

Call:

```
lm(formula = 腰圍 ~ 年齡 + 性別 + 收縮壓 + 空腹血糖 + 高密度脂蛋白, data = cvd_200)
```

Residuals:

Min	1Q	Median	3Q	Max
-37.917	-6.030	-0.053	5.816	25.858

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.36694	6.52935	9.858	< 2e-16 ***
年齡	0.22259	0.05755	3.868	0.000159 ***
性別	5.20302	1.63732	3.178	0.001780 **
收縮壓	0.13596	0.03743	3.632	0.000378 ***
空腹血糖	-0.03604	0.02354	-1.531	0.127745
高密度脂蛋白	-0.17732	0.06523	-2.719	0.007277 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.015 on 161 degrees of freedom

Multiple R-squared: 0.3526, Adjusted R-squared: 0.3325

F-statistic: 17.54 on 5 and 161 DF, p-value: 7.576e-14

【分析結果 2.】

由上述結果可看出，最終模型結果與向前選取法相同為：

$$\widehat{y}_{\text{腰圍}} = 64.36694 + 0.31596(\text{收縮壓}) + 5.20302(\text{性別}) + 0.22259(\text{年齡}) \\ - 0.17732(\text{高密度脂蛋白}) - 0.03604(\text{空腹血糖})$$

其 Adjusted $R^2=0.3325$ ，在假設顯著水準為 0.05 下，有顯著的變項為收縮壓、性別、年齡以及高密度脂蛋白，其中以性別的影響最大，在控制其他變項的情況下，男性的腰圍會比女性的腰圍多 5.2 公分。

【程式碼】

3. 逐步選取法

```
#使用 stepAIC(), 設定 direction 為 both
library(MASS)
stepwise.lm = stepAIC(fullmodel, direction="both")
summary(stepwise.lm)
```

【Output】

此範例 stepwise 篩選過程與前述 backward 方法相同，因此省略。

Call:

```
lm(formula = 腰圍 ~ 年齡 + 性別 + 收縮壓 + 空腹血糖 + 高密度脂蛋白, data = cvd_200)
```

Residuals:

Min	1Q	Median	3Q	Max
-37.917	-6.030	-0.053	5.816	25.858

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.36694	6.52935	9.858	< 2e-16 ***
年齡	0.22259	0.05755	3.868	0.000159 ***
性別	5.20302	1.63732	3.178	0.001780 **
收縮壓	0.13596	0.03743	3.632	0.000378 ***
空腹血糖	-0.03604	0.02354	-1.531	0.127745
高密度脂蛋白	-0.17732	0.06523	-2.719	0.007277 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.015 on 161 degrees of freedom

Multiple R-squared: 0.3526, Adjusted R-squared: 0.3325

F-statistic: 17.54 on 5 and 161 DF, p-value: 7.576e-14

【分析結果 3.】

由上述結果可看出，最終模型結果與向前選取法及向後選取法相同為：

$$\widehat{y}_{\text{腰圍}} = 64.36694 + 0.31596(\text{收縮壓}) + 5.20302(\text{性別}) + 0.22259(\text{年齡}) \\ - 0.17732(\text{高密度脂蛋白}) - 0.03604(\text{空腹血糖})$$

其 Adjusted $R^2=0.3325$ ，在假設顯著水準為 0.05 下，有顯著的變項為收縮壓、性別、年齡以及高密度脂蛋白，其中以性別的影響最大，在控制其他變項的情況下，男性的腰圍會比女性的腰圍多 5.2 公分。

(三) 交互作用項 (Interaction term)

- 根據前述模型的分析後，若想了解空腹血糖與高密度脂蛋白兩變項之間是否有交互作用，則在原本的程式碼中加入 空腹血糖*高密度脂蛋白，程式碼如下：

【程式碼】

```
#多變項線性迴歸
#依變項：腰圍
#自變項：年齡、空腹血糖、高密度脂蛋白、空腹血糖與高密度脂蛋白
交互作用項
model_interact=lm(腰圍~ 年齡+空腹血糖+高密度脂蛋白+空腹血糖*高
密度脂蛋白, data=cvd_all)
#利用 summary() 看更詳細的分析結果
summary(model_interact)
```

【Output】**Call:**

```
lm(formula = 腰圍 ~ 年齡 + 空腹血糖 + 高密度脂蛋白 + 空腹血糖 * 高密度脂蛋白, data = cvd_ok)
```

Residuals:

Min	1Q	Median	3Q	Max
-22.9549	-5.9064	0.0638	6.1558	25.8058

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	151.425047	11.641734	13.007	< 2e-16 ***
年齡	0.222472	0.053997	4.120	6.03e-05 ***
空腹血糖	-0.694186	0.112198	-6.187	4.82e-09 ***
高密度脂蛋白	-1.686317	0.236109	-7.142	2.95e-11 ***
空腹血糖:高密度脂蛋白	0.015114	0.002507	6.029	1.08e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.843 on 162 degrees of freedom

Multiple R-squared: 0.3732, Adjusted R-squared: 0.3577

F-statistic: 24.11 on 4 and 162 DF, p-value: 1.16e-15

【分析結果】

由上方結果可看出，在假設顯著水準為 0.05 下，空腹血糖與高密度脂蛋白的交互作用項是有顯著的，表示兩變項間有交互作用存在。

二、多變項邏輯斯迴歸分析 (Multiple logistic regression analysis)

當關心的依變項為二元類別變項，而非連續型變項時，則常使用邏輯斯迴歸分析來討論依變項與自變項的關係。在本中心 eNews 的第 31 期中，已經討論過邏輯斯迴歸分析的多變項模型，因此在本篇中接續前期之討論並加入模型選取以及交互作用項進行分析討論。

(一) 模型變項選取

邏輯斯迴歸分析的模型變項選取方式與前述的線性迴歸分析常用的方式相同，僅在程式碼中有不同的方式撰寫。

【程式碼】

```
#使用 epiDisplay 套件計算 OR 與信賴區間
library(epiDisplay)
#指定所有變項納入模型中的 full model
fullmodel=glm(factor(心血管疾病)~ 年齡 + factor(性別) + 腰圍 + 收縮壓 + 舒張壓 +
空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + factor(檳榔) + factor(飲酒) + factor(家族病史) +
factor(抽菸量),family=binomial, data=cvd_200)
(因資料檔中的抽菸及抽菸量是高度相關的變項，僅選擇其一做分析，在本篇
使用抽菸量做為自變項分析)
#使用 factor() 定義該變項為類別變項
summary(fullmodel) #利用 summary() 看更詳細分析結果
#建立只有截距項的空線性迴歸(null model)
nullmodel= glm(心血管疾病~ 1,family=binomial, data=cvd_200)

## 1. 向前選取法
#使用 step(), 設定 direction 為 forward, 會自動一個一個把變數丟進去
forward.logim = step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel),
direction="forward") # (一定要加上界 upper=fullmodel, 不可以不加)
summary(forward.logim)
forward.OR=logistic.display(forward.logim)
forward.OR
```

【Output】**Call:**

```
glm(formula = 心血管疾病 ~ 年齡 + 性別 + 腰圍 + 收縮壓 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史 + 抽菸量, family = binomial, data = cvd_200)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.19438	-0.33758	-0.18301	-0.07646	3.07775

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.716e+01	5.824e+00	-2.946	0.00322 **
年齡	1.016e-01	3.584e-02	2.835	0.00458 **
性別 yes	4.951e-02	8.725e-01	0.057	0.95475
腰圍	1.298e-01	4.757e-02	2.728	0.00637 **
收縮壓	-1.243e-03	2.086e-02	-0.060	0.95247
舒張壓	-3.615e-02	3.901e-02	-0.927	0.35402
空腹血糖	-1.506e-02	1.221e-02	-1.234	0.21737
高密度脂蛋白	4.263e-02	3.455e-02	1.234	0.21726
三酸甘油酯	5.603e-03	5.818e-03	0.963	0.33549
檳榔 yes	6.026e-01	1.462e+00	0.412	0.68026
飲酒 yes	-5.788e-01	1.105e+00	-0.524	0.60045
家族病史 yes	-1.283e+01	1.398e+03	-0.009	0.99268
抽菸 yes	-5.868e-01	1.036e+00	-0.566	0.57124

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 65.550 on 162 degrees of freedom

AIC: 91.55

Number of Fisher Scoring iterations: 16

Call:

```
glm(formula = 心血管疾病 ~ 1, family = binomial, data = cvd_200)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.3929	-0.3929	-0.3929	-0.3929	2.2803

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.5226	0.2883	-8.751	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 92.605 on 174 degrees of freedom

AIC: 94.605

Number of Fisher Scoring iterations: 5

【Output】

##1.向前選取法

Start: AIC=94.61

心血管疾病 ~ 1

	Df	Deviance	AIC
+ 年齡	1	78.548	82.548
+ 腰圍	1	81.361	85.361
+ 收縮壓	1	88.838	92.838
<none>		92.605	94.605
+ 舒張壓	1	90.970	94.970
+ 性別	1	91.085	95.085
+ 家族病史	1	91.501	95.501
+ 三酸甘油酯	1	91.889	95.889
+ 高密度脂蛋白	1	92.151	96.151
+ 飲酒	1	92.187	96.187
+ 檳榔	1	92.512	96.512
+ 抽菸	1	92.598	96.598
+ 空腹血糖	1	92.603	96.603

Step: AIC=82.55

心血管疾病 ~ 年齡

	Df	Deviance	AIC
+ 腰圍	1	71.084	77.084
<none>		78.548	82.548
+ 空腹血糖	1	77.553	83.553
+ 檳榔	1	77.891	83.891
+ 三酸甘油酯	1	78.165	84.165
+ 家族病史	1	78.200	84.200
+ 飲酒	1	78.282	84.282
+ 性別	1	78.388	84.388
+ 抽菸	1	78.406	84.406
+ 舒張壓	1	78.457	84.457
+ 收縮壓	1	78.517	84.517
+ 高密度脂蛋白	1	78.542	84.542

Step: AIC=77.08

心血管疾病 ~ 年齡 + 腰圍

	Df	Deviance	AIC
<none>		71.084	77.084
+ 空腹血糖	1	69.596	77.596
+ 高密度脂蛋白	1	69.778	77.778
+ 飲酒	1	70.496	78.496
+ 抽菸	1	70.554	78.554
+ 舒張壓	1	70.655	78.655
+ 收縮壓	1	70.765	78.765
+ 家族病史	1	70.819	78.819
+ 性別	1	70.894	78.894
+ 檳榔	1	71.004	79.004
+ 三酸甘油酯	1	71.055	79.055

【Output】**Call:**

```
glm(formula = 心血管疾病 ~ 年齡 + 腰圍, family = binomial, data = cvd_200)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0725	-0.3839	-0.2158	-0.1095	3.1729

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-14.65950	3.69931	-3.963	7.41e-05 ***
年齡	0.07631	0.02613	2.921	0.00349 **
腰圍	0.09676	0.03791	2.553	0.01069 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 71.084 on 172 degrees of freedom

AIC: 77.084

Number of Fisher Scoring iterations: 7

Logistic regression predicting 心血管疾病 : yes vs no

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
年齡 (cont. var.)	1.09 (1.04,1.14)	1.08 (1.03,1.14)	0.003	0.001
腰圍 (cont. var.)	1.11 (1.04,1.18)	1.1 (1.02,1.19)	0.011	0.006

Log-likelihood = -35.5422

No. of observations = 175

AIC value = 77.0844

【分析結果 1.】

根據上面向前選取法的結果，依變項為心血管疾病的邏輯斯迴歸模型其自變項為年齡以及腰圍，其中在假設顯著水準為 0.05 下，從 OR 結果可以看出，在控制年齡下，腰圍每增加 1 公分，罹患心血管疾病的風險會增加 1.11 倍；在控制腰圍下，年齡每增加 1 歲，罹患心血管疾病的風險會增加 1.09 倍，且都達到統計上顯著水準。

【程式碼】

2. 向後選取法

backward.logim=step(fullmodel)

summary(backward.logim)

backward.OR= logistic.display (backward.logim)

backward.OR

【Output】

##2.向後選取法

Start: AIC=91.55

心血管疾病 ~ 年齡 + 性別 + 腰圍 + 收縮壓 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + (飲酒) + (家族病史) + 抽菸

	Df	Deviance	AIC
- 性別	1	65.554	89.554
- 收縮壓	1	65.554	89.554
- 家族病史	1	65.686	89.686
- 檳榔	1	65.713	89.713
- 飲酒	1	65.834	89.834
- 抽菸	1	65.876	89.876
- 舒張壓	1	66.395	90.395
- 三酸甘油酯	1	66.443	90.443
- 高密度脂蛋白	1	66.983	90.983
<none>		65.550	91.550
- 空腹血糖	1	67.722	91.722
- 腰圍	1	74.381	98.381
- 年齡	1	76.433	100.433

Step: AIC=89.55

心血管疾病 ~ 年齡 + 腰圍 + 收縮壓 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史 + 抽菸

	Df	Deviance	AIC
- 收縮壓	1	65.558	87.558
- 家族病史	1	65.691	87.691
- 檳榔	1	65.721	87.721
- 飲酒	1	65.834	87.834
- 抽菸	1	65.908	87.908
- 舒張壓	1	66.398	88.398
- 三酸甘油酯	1	66.444	88.444
- 高密度脂蛋白	1	67.031	89.031
<none>		65.554	89.554
- 空腹血糖	1	67.815	89.815
- 腰圍	1	74.622	96.622
- 年齡	1	76.699	98.699

Step: AIC=87.56

心血管疾病 ~ 年齡 + 腰圍 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 家族病史 + 抽菸

	Df	Deviance	AIC
- 家族病史	1	65.694	85.694
- 檳榔	1	65.723	85.723
- 飲酒	1	65.835	85.835
- 抽菸	1	65.910	85.910
- 三酸甘油酯	1	66.447	86.447
- 舒張壓	1	66.716	86.716
- 高密度脂蛋白	1	67.065	87.065
<none>		65.558	87.558
- 空腹血糖	1	67.929	87.929
- 腰圍	1	74.627	94.627
- 年齡	1	77.750	97.750

Step: AIC=85.69

心血管疾病 ~ 年齡 + 腰圍 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 檳榔 + 飲酒 + 抽菸

	Df	Deviance	AIC
- 檳榔	1	65.868	83.868
- 飲酒	1	65.979	83.979
- 抽菸	1	66.058	84.058
- 三酸甘油酯	1	66.638	84.638
- 舒張壓	1	66.875	84.875
- 高密度脂蛋白	1	67.243	85.243
<none>		65.694	85.694
- 空腹血糖	1	68.070	86.070
- 腰圍	1	74.868	92.868
- 年齡	1	78.236	96.236

Step: AIC=83.87

心血管疾病 ~ 年齡 + 腰圍 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 飲酒 + 抽菸

	Df	Deviance	AIC
- 抽菸	1	66.120	82.120
- 飲酒	1	66.131	82.131
- 三酸甘油酯	1	66.911	82.911
- 舒張壓	1	66.962	82.962
- 高密度脂蛋白	1	67.821	83.821
<none>		65.868	83.868
- 空腹血糖	1	68.197	84.197
- 腰圍	1	75.928	91.928
- 年齡	1	78.385	94.385

Step: AIC=82.12

心血管疾病 ~ 年齡 + 腰圍 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 三酸甘油酯 + 飲酒

	Df	Deviance	AIC
- 三酸甘油酯	1	66.996	80.996
- 舒張壓	1	67.110	81.110
- 飲酒	1	67.162	81.162
<none>		66.120	82.120
- 高密度脂蛋白	1	68.161	82.161
- 空腹血糖	1	68.285	82.285
- 腰圍	1	75.962	89.962
- 年齡	1	78.391	92.391

Step: AIC=81

心血管疾病 ~ 年齡 + 腰圍 + 舒張壓 + 空腹血糖 + 高密度脂蛋白 + 飲酒

	Df	Deviance	AIC
- 舒張壓	1	67.638	79.638
- 飲酒	1	67.666	79.666
- 高密度脂蛋白	1	68.404	80.404
- 空腹血糖	1	68.603	80.603
<none>		66.996	80.996
- 腰圍	1	77.227	89.227
- 年齡	1	78.450	90.450

Step: AIC=79.64

心血管疾病 ~ 年齡 + 腰圍 + 空腹血糖 + 高密度脂蛋白 + 飲酒

	Df	Deviance	AIC
- 飲酒	1	68.293	78.293
- 空腹血糖	1	69.030	79.030
- 高密度脂蛋白	1	69.121	79.121
<none>		67.638	79.638
- 腰圍	1	77.299	87.299
- 年齡	1	78.473	88.473

Step: AIC=78.29

心血管疾病 ~ 年齡 + 腰圍 + 空腹血糖 + 高密度脂蛋白

	Df	Deviance	AIC
- 高密度脂蛋白	1	69.596	77.596
- 空腹血糖	1	69.778	77.778
<none>		68.293	78.293
- 腰圍	1	77.540	85.540
- 年齡	1	80.097	88.097

Step: AIC=77.6

心血管疾病 ~ 年齡 + 腰圍 + 空腹血糖

	Df	Deviance	AIC
- 空腹血糖	1	71.084	77.084
<none>		69.596	77.596
- 腰圍	1	77.553	83.553
- 年齡	1	81.152	87.152

Step: AIC=77.08

心血管疾病 ~ 年齡 + 腰圍

	Df	Deviance	AIC
<none>		71.084	77.084
- 腰圍	1	78.548	82.548
- 年齡	1	81.361	85.361

【Output】**Call:**

```
glm(formula = 心血管疾病 ~ 年齡 + 腰圍, family = binomial, data = cvd_200)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0725	-0.3839	-0.2158	-0.1095	3.1729

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-14.65950	3.69931	-3.963	7.41e-05 ***
年齡	0.07631	0.02613	2.921	0.00349 **
腰圍	0.09676	0.03791	2.553	0.01069 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 71.084 on 172 degrees of freedom

AIC: 77.084

Number of Fisher Scoring iterations: 7

Logistic regression predicting 心血管疾病 : yes vs no

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
年齡 (cont. var.)	1.09 (1.04,1.14)	1.08 (1.03,1.14)	0.003	0.001
腰圍 (cont. var.)	1.11 (1.04,1.18)	1.1 (1.02,1.19)	0.011	0.006

Log-likelihood = -35.5422

No. of observations = 175

AIC value = 77.0844

【分析結果 2.】

根據上面向後選取法的結果，依變項為心血管疾病的邏輯斯迴歸模型其自變項為年齡以及腰圍，其中在假設顯著水準為 0.05 下，從 OR 結果可以看出，在控制年齡下，腰圍每增加 1 公分，罹患心血管疾病的風險會增加 1.11 倍；在控制腰圍下，年齡每增加 1 歲，罹患心血管疾病的風險會增加 1.09 倍，且都達到統計上顯著水準。此方法的結果與向前選取法是一樣的。

【程式碼】

3. 逐步選取法

```
stepwise.logim = step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel),
direction="both",trace=0)
summary(stepwise.logim)
stepwise.OR= logistic.display (stepwise.logim)
stepwise.OR
```

【Output】

##3. 逐步選取法

此範例 stepwise 篩選過程與前述 backward 方法相同，因此省略。

Call:

```
glm(formula = 心血管疾病 ~ 年齡 + 腰圍, family = binomial, data = cvd_200)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0725	-0.3839	-0.2158	-0.1095	3.1729

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-14.65950	3.69931	-3.963	7.41e-05 ***
年齡	0.07631	0.02613	2.921	0.00349 **
腰圍	0.09676	0.03791	2.553	0.01069 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 71.084 on 172 degrees of freedom

AIC: 77.084

Number of Fisher Scoring iterations: 7

Logistic regression predicting 心血管疾病 : yes vs no

	crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
年齡 (cont. var.)	1.09 (1.04,1.14)	1.08 (1.03,1.14)	0.003	0.001
腰圍 (cont. var.)	1.11 (1.04,1.18)	1.1 (1.02,1.19)	0.011	0.006

Log-likelihood = -35.5422

No. of observations = 175

AIC value = 77.0844

【分析結果 3.】

根據上面逐步選取法的結果，依變項為心血管疾病的邏輯斯迴歸模型其自變項為年齡以及腰圍，其中在假設顯著水準為 0.05 下，從 OR 結果可以看出，在控制年齡下，腰圍每增加 1 公分，罹患心血管疾病的風險會增加 1.11 倍；在控制腰圍下，年齡每增加 1 歲，罹患心血管疾病的風險會增加 1.09 倍，且都達到統計上顯著水準。此方法的結果與向前、向後選取法皆為相同。

(二) 交互作用項

若我們想了解，是否性別與腰圍兩自變項之間是否存在交互作用，並會影響心血管疾病的發生，則程式碼如下：

【程式碼】

##邏輯斯迴歸交互作用項

```
logitmd_interact=glm(心血管疾病~年齡+性別+腰圍+性別*腰圍,
family=binomial,data=cvd_ok)
summary(logitmd_interact)
```

【Output】**Call:**

```
glm(formula = 心血管疾病 ~ 年齡 + 性別 + 腰圍 + 性別*腰圍, family = binomial(link = logit), data = cvd_ok)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0211	-0.3760	-0.2151	-0.1112	3.1666

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-15.32581	4.68656	-3.270	0.00107 **
年齡	0.07622	0.02707	2.815	0.00487 **
性別 1	0.78296	6.99896	0.112	0.91093
腰圍	0.10691	0.05470	1.954	0.05065 .
性別 1:腰圍	-0.01253	0.08090	-0.155	0.87690

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 92.605 on 174 degrees of freedom

Residual deviance: 70.870 on 170 degrees of freedom

AIC: 80.87

Number of Fisher Scoring iterations: 7

【分析結果】

根據上面結果可以看出，在假設顯著水準為 0.05 下，此邏輯斯迴歸模型中，性別與腰圍的交互作用項並沒有顯著，故兩變項間並無交互作用存在。

以上為本期生統 eNews 的內容，介紹多變項迴歸分析、模型選取方法，以及交互作用項的 R 軟體程式撰寫與結果解釋，希望本期 eNews 能夠提供大家操作之參考。