程式指令:

aov(y~factor1+factor2, data=mydata) #從 mydata 的 y 那一欄(行)取到 y 的資料。注意: 在此 factor1 和 factor2 是沒有交互作用 (no interaction)

anova(aov(y~ factor1+factor2, data=mydata)) #列出詳細的雙因子變異數分析表。

summary(aov(y~ factor1+factor2, data=mydata)) #列出詳細的雙因子變異數分析表。

```
> y1=c(12,14,18,16,15,15)
> y2=c(14,12,13,13,11,15)
> y3=c(19,21,18,22,20,20)
> y=c(y1,y2,y3)
> pdata=expand.grid(blocks=paste("B",c(1,1,2,2,3,3),sep=""),pack=c("packA",
+ "packB","packC")) #產生集區因子變數"blocks"和包裝因子變數"pack"
> pdata$y=y #將雙因子變數和18家超商的銷售資料整合成18*3的下列矩陣形式
> pdata
   blocks  pack  y
1      B1 packA 12
2      B1 packA 14
3      B2 packA 18
4      B2 packA 16
5      B3 packA 15
6      B3 packA 15
7      B1 packB 14
8      B1 packB 12
9      B2 packB 13
10     B2 packB 13
11     B3 packB 11
12     B3 packB 15
13     B1 packC 19
14     B1 packC 21
15     B2 packC 18
16     B2 packC 22
17     B3 packC 20
18     B3 packC 20
```

```
> summary(aov(y~pack+blocks,data=pdata))
            Df Sum Sq Mean Sq F value    Pr(>F)
pack         2 156.00   78.00   29.25 1.54e-05 ***
blocks       2   5.33    2.67    1.00     0.394
Residuals   13  34.67    2.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(aov(y~pack+blocks,data=pdata))
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
pack       2 156.000  78.000   29.25 1.54e-05 ***
blocks     2   5.333   2.667    1.00    0.3945
Residuals 13  34.667   2.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

迴歸分析

迴歸分析(regression analysis)是一種建立應變數(response variable)與自變數(independent variable,又叫解釋變數, explanatory variable)函數關係的統計分析技巧。其統計模式為

$$Y = X + \varepsilon，\varepsilon \sim N(0, \sigma^2)。$$

1 簡單線性迴歸模式(simple linear regression model)

給定 n 對獨立的樣本資料 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$，簡單線性迴歸模式為

$$Y_i = \alpha + \beta X_i + \varepsilon_i，1 \leq i \leq n,$$

其中參數 α 代表截距，參數 β 代表斜率，X_i =自變數的第 i 個觀測值， Y_i =應變數的第 i 個觀測值，而 $\varepsilon_1, \ldots, \varepsilon_n$假設為來自 $N(0, \sigma^2)$ 的獨立同態分佈。(注意: 在上面模式只有$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$是可觀測到的，而 $\varepsilon_1, \ldots, \varepsilon_n$是不可觀測的，且通常假設 $X_1, X_2, \ldots, X_n$ 為已知的常數!) 因此

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2) ，1 \leq i \leq n。$$

利用最小平方(least squares)估計法來估計參數 α 、β 和 σ^2:

找到一組 α 和 β 的值使得平方和 $S = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$最小! 利用微積分方法可得

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{j=1}^{n} X_j{}^2 - n\bar{X}^2}$$

$$= \sum_{i=1}^{n} \frac{(X_i - \bar{X})Y_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2} ,$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \sum_{i=1}^{n} [\frac{1}{n} - \frac{(X_i - \bar{X})\bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2}] Y_i$$

和 $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 / (\mathrm{n} - 2) = SSE/(\mathrm{n} - 2)$，其中

$S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$，$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$ 和 $SSE = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$。

例 某建設公司欲分析新北市影響豪宅房價的因素，10 筆成交記錄的豪宅坪數(解釋變數X: 單位坪)和成交價(應變數Y: 單位百萬元)如下:

坪數:	50	100	60	50	70	80	90	100	80	60
成交價:	55	100	70	60	85	100	90	100	80	80

$\bar{X} = 74$，$\bar{Y} = 82$，$S_{xy} = 2520$，$S_{xx} = 3240$，因此

$\hat{\beta} = S_{xy}/S_{xx} = 2520/3240 = 0.778$，$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 82 - 0.778(74)$

$= 24.428$，

$$\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 / (\mathrm{n} - 2) = SSE/(\mathrm{n} - 2)$$

$= 448.88/8 = 56.11$。

最佳迴歸線為 $Y = 24.428 + 0.778X$。($X = 0,\ Y = 24.428$，沒意義! 不可能有 0 坪的豪宅。24.428代表豪宅每增一坪，平均成交價會增 0.778 百萬元。

最小平方估計量之性質:

(a) $\hat{\beta}\sim N(\beta,\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2})$，$\hat{\alpha}\sim N(\alpha,(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2})\,\sigma^2)$

(b) $SSE/\sigma^2=\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}X_i)^2/\sigma^2\sim\chi^2_{n-2}$

(c) $\hat{\alpha}$ 和 $\hat{\beta}$ 與SSE相互獨立(但 $\hat{\alpha}$ 和 $\hat{\beta}$不獨立)

(d) $SSE=\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}X_i)^2=\sum_{i=1}^{n}[(Y_i-\bar{Y})-\hat{\beta}(X_i-\bar{X})]^2$

$=\sum_{i=1}^{n}(Y_i-\bar{Y})^2-2\hat{\beta}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})+\hat{\beta}^2\sum_{i=1}^{n}(X_i-\bar{X})^2$

$=S_{yy}-2\hat{\beta}S_{xy}+\hat{\beta}^2S_{xx}=S_{yy}-2\hat{\beta}^2S_{xx}+\hat{\beta}^2S_{xx}=S_{yy}-\hat{\beta}^2S_{xx}$，

其中 $S_{yy}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$。

檢定 H_0: $\beta=\beta_0$ (已知) 對 H_1: $\beta\neq\beta_0$

給定顯著水準 $\gamma=5\%$，當 H_0 為真時，檢定統計量 $T=\frac{\hat{\beta}-\beta_0}{\sqrt{MSE/S_{xx}}}\sim t_{n-2}$，當檢定統計量觀測值$|t|>t_{n-2,1-\gamma/2}$，則拒絕 H_0，否則不拒絕，其中 $t_{n-2,1-\gamma/2}$代表 t_{n-2}分佈的 $1-\gamma/2$ 分位值。亦可計算 p-值$=2P(t_{n-2}>|t|)$，由 p-值大小來判斷是否拒絕 H_0。利用前面教過推導信賴區間的方法(自已試看看!)，可得斜率 β 的$100(1-)\%$信賴區間為 $[\hat{\beta}-t_{n-2,1-\gamma/2}\sqrt{MSE/S_{xx}}\,,\ \hat{\beta}+t_{n-2,1-\gamma/2}\sqrt{MSE/S_{xx}}]$。

同理，

檢定 H_0: $\alpha=\alpha_0$ (已知) 對 H_1: $\alpha\neq\alpha_0$

給定顯著水準 $\gamma=5\%$，$T=\frac{\hat{\alpha}-\alpha_0}{\sqrt{MSE(1/n+\bar{X}^2/S_{xx})}}\sim t_{n-2}$，當檢定統計量觀測值$|t|>t_{n-2,1-\gamma/2}$，則拒絕 H_0，否則不拒絕。亦可計算 p-值