HW3 110024516 邱繼賢

Eg: 3.18

(a) Let total energy consumption $= x_1 + x_2 + x_3 + x_4 = a'\mathbf{x}$

where $a = [1, 1, 1, 1]'$, $\mathbf{x} = [x_1, x_2, x_3, x_4]'$

$\therefore$ sample mean $= a'\bar{\mathbf{x}} = \bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4 = 1.873$

sample variance $= a'Sa = 3.914$

(b) Let $x_1 - x_2 = b'\mathbf{x}$, where $b = [1, -1, 0, 0]'$

$\therefore$ sample mean $= b'\bar{\mathbf{x}} = \bar{x}_1 - \bar{x}_2 = 0.258$

sample variance $= b'Sb = 0.154$

sample covariance $(b'\mathbf{x}, a'\mathbf{x}) = b'Sa = 0.362$

(ps: 以上計算皆令 $S_{23} = S_{32} = 0.128$)

Eg: 8.4

$\det(\Sigma - \lambda I) \overset{\text{set}}{=\!=} 0 \Rightarrow \lambda = \sigma^2 \vee \sigma^2(1+\sqrt{2}\rho) \vee \sigma^2(1-\sqrt{2}\rho)$

For $\lambda_1 = \sigma^2$, then $\Sigma e_1 = \lambda_1 e_1$ and $\|e_1\| = 1 \Rightarrow e_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$

For $\lambda_2 = \sigma^2(1+\sqrt{2}\rho)$, then $\Sigma e_2 = \lambda_2 e_2$ and $\|e_2\| = 1 \Rightarrow e_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{\sqrt{2}} \\ \frac{1}{2} \end{bmatrix}$

For $\lambda_3 = \sigma^2(1-\sqrt{2}\rho)$, then $\Sigma e_3 = \lambda_3 e_3$ and $\|e_3\| = 1 \Rightarrow e_3 = \begin{bmatrix} \frac{1}{2} \\ \frac{-1}{\sqrt{2}} \\ \frac{1}{2} \end{bmatrix}$

Case I : $\rho = 0$, then $\lambda_1 = \lambda_2 = \lambda_3$

The original 3 variables are uncorrelated, so we do not need to do the principal component

Case II : $\rho > 0$, then $\lambda_2 > \lambda_1 > \lambda_3$

The 1st PC

$Y_1 = \frac{1}{2} X_1 + \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$, accounts for $\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = \frac{1+\sqrt{2}\rho}{3}$

The 2nd PC

$Y_2 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_3$, accounts for $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{1}{3}$

The 3rd PC

$Y_3 = \frac{1}{2} X_1 - \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$, accounts for $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} = \frac{1-\sqrt{2}\rho}{3}$

Case III : $\rho < 0$, then $\lambda_3 > \lambda_1 > \lambda_2$

The first PC

$Y_1 = \frac{1}{2} X_1 - \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$, accounts for $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} = \frac{1-\sqrt{2}\rho}{3}$

The 2nd PC

$Y_2 = \frac{1}{\sqrt{2}} X_1 - \frac{1}{\sqrt{2}} X_3$, accounts for $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{1}{3}$

The 3rd PC

$Y_3 = \frac{1}{2} X_1 + \frac{1}{\sqrt{2}} X_2 + \frac{1}{2} X_3$, accounts for $\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} = \frac{1+\sqrt{2}\rho}{3}$ $\square$

Eg 8.6

(a) $\det(S-\lambda I) \overset{\text{set}}{=} 0 \Rightarrow \hat{\lambda}_1 = 7488.803$, $\hat{\lambda}_2 = 13.837$

The corresponding eigenvectors : $\hat{e}_1 = \begin{bmatrix} 0.999 \\ 0.041 \end{bmatrix}$, $\hat{e}_2 = \begin{bmatrix} -0.041 \\ 0.999 \end{bmatrix}$

The 1st PC : $\hat{y}_1 = \hat{e}_1' \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.999x_1 + 0.041x_2$

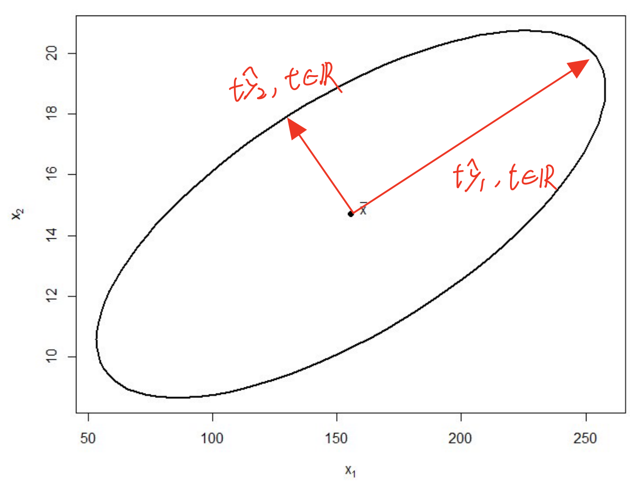
Sample variance $(\hat{y}_1) = \hat{\lambda}_1 = 7488.804$

The 2nd PC : $\hat{y}_2 = \hat{e}_2' \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = -0.041x_1 + 0.999x_2$

Sample variance $(\hat{y}_2) = \hat{\lambda}_2 = 13.837$ $\square$

(b) $\dfrac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = 0.998$ $\square$

(c)



(d) $r_{\hat{y}_1, x_1} = \dfrac{\hat{e}_{11}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{11}}} = 0.9999985 \approx 1$, $r_{\hat{y}_1, x_2} = \dfrac{\hat{e}_{12}\sqrt{\hat{\lambda}_1}}{\sqrt{s_{22}}} = 0.6874 \approx r_{x_1, x_2}$

$\because s_{11} >> s_{22}$, $\therefore \hat{y}_1$ is almost dominated by x_1 $(\hat{e}_{11} \approx 1)$

In such cases, we might standardize the variables first next time.

2.

(a)

以下為將此 13 個變數的 covariance matrix 計算 eigenvalues 和 eigenvectors 後所得的 principal components

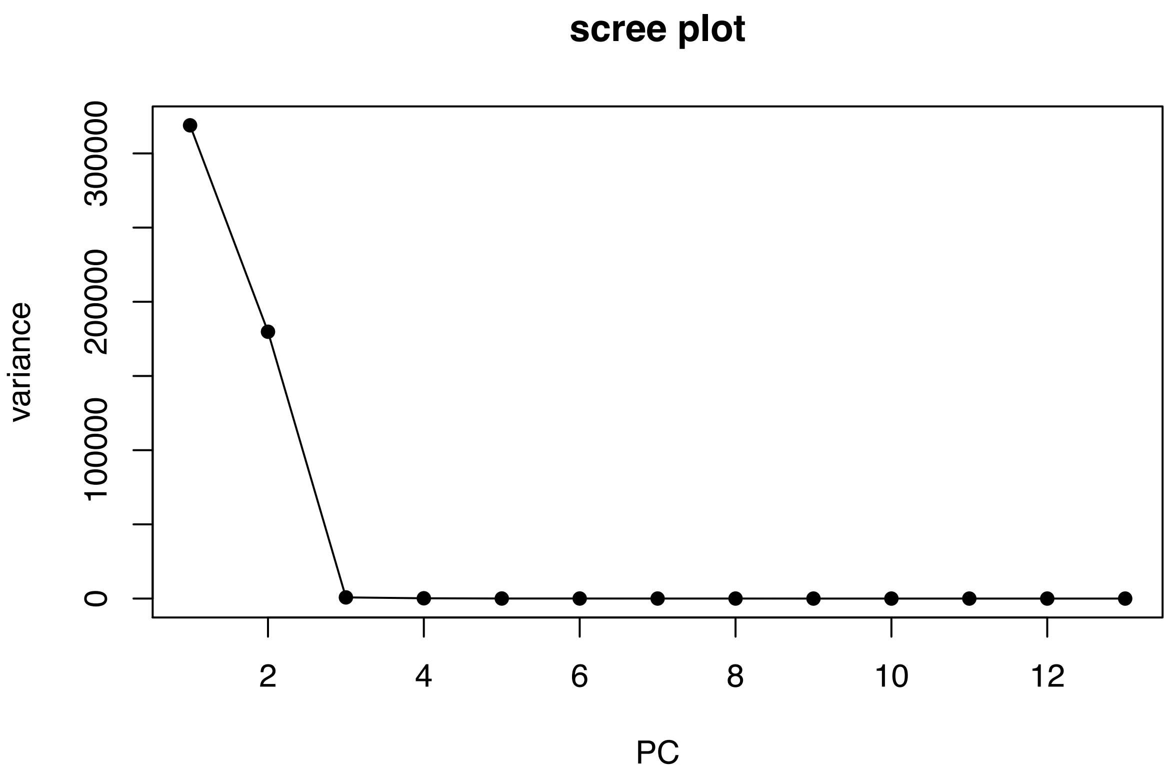
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
wheelbase					0.32	0.22	0.53	0.68	0.24		0.19		
carlength					0.82	0.37	-0.26	-0.33					
carwidth								0.11	0.22	0.21	-0.94		
carheight					0.13	0.11	0.2	0.12	-0.94		-0.16		
curbweight	$\boxed{0.81}$	$\boxed{0.58}$											
enginesize			-0.68	0.72									
boreratio												0.2	-0.98
stroke												-0.98	-0.2
compressionratio					-0.21	0.43	-0.71	0.51					
horsepower			-0.71	-0.67		0.17							
peakrpm	$\boxed{-0.58}$	$\boxed{0.81}$											
citympg				0.1	-0.29	0.52	0.26	-0.15		-0.73	-0.15		
highwaympg					-0.28	0.56	0.21	-0.35		0.64	0.14		

各 principal components 的 variances 即為 covariance matrix 由大到小的 eigenvalues

	variance
PC1	318909.87
PC2	179807.41
PC3	780.98
PC4	198.99
PC5	38.09
PC6	24.06
PC7	6.80
PC8	5.50
PC9	2.85
PC10	0.93
PC11	0.77
PC12	0.09
PC13	0.03

(b)

繪製各 principal components 的 scree plot ，以及累計的 variance 比例




```
## Importance of components:
##                              PC1         PC2          PC3          PC4
## Standard deviation     564.7210527 424.0370365 27.945977162 1.410628e+01
## Proportion of Variance   0.6381051   0.3597757  0.001562654 3.981521e-04
## Cumulative Proportion    0.6381051   0.9978809  0.999443531 9.998417e-01
##                                PC5          PC6          PC7          PC8
## Standard deviation     6.171906e+00 4.9047769051 2.607643e+00 2.345611e+00
## Proportion of Variance 7.621893e-05 0.0000481352 1.360569e-05 1.100871e-05
## Cumulative Proportion  9.999179e-01 0.9999660369 9.999796e-01 9.999907e-01
##                                PC9         PC10         PC11         PC12
## Standard deviation     1.686814e+00 9.658365e-01 8.794616e-01 2.945868e-01
## Proportion of Variance 5.693226e-06 1.866515e-06 1.547598e-06 1.736404e-07
## Cumulative Proportion  9.999963e-01 9.999982e-01 9.999998e-01 9.999999e-01
##                               PC13
## Standard deviation     1.839476e-01
## Proportion of Variance 6.770375e-08
## Cumulative Proportion  1.000000e+00
```

- 前兩個 principal components 的變異程度遠大於剩餘的 principal components
- 前兩個 principal components 所佔的變異程度比例已經超過 99%

我會只選擇前兩個 principal components：

$$\hat{y}_1 \;=\; 0.812 \times curbweight \;-\; 0.58 \times peakrpm$$

可以解釋為 *curbweight* 和 *peakrpm* 這兩個變數間的加權差距 (weighted difference)

$$\hat{y}_2 \;=\; 0.576 \times curbweight \;+\; 0.814 \times peakrpm$$

可以解釋為 *curbweight* 和 *peakrpm* 這兩個變數的加權相加 (weighted sum)

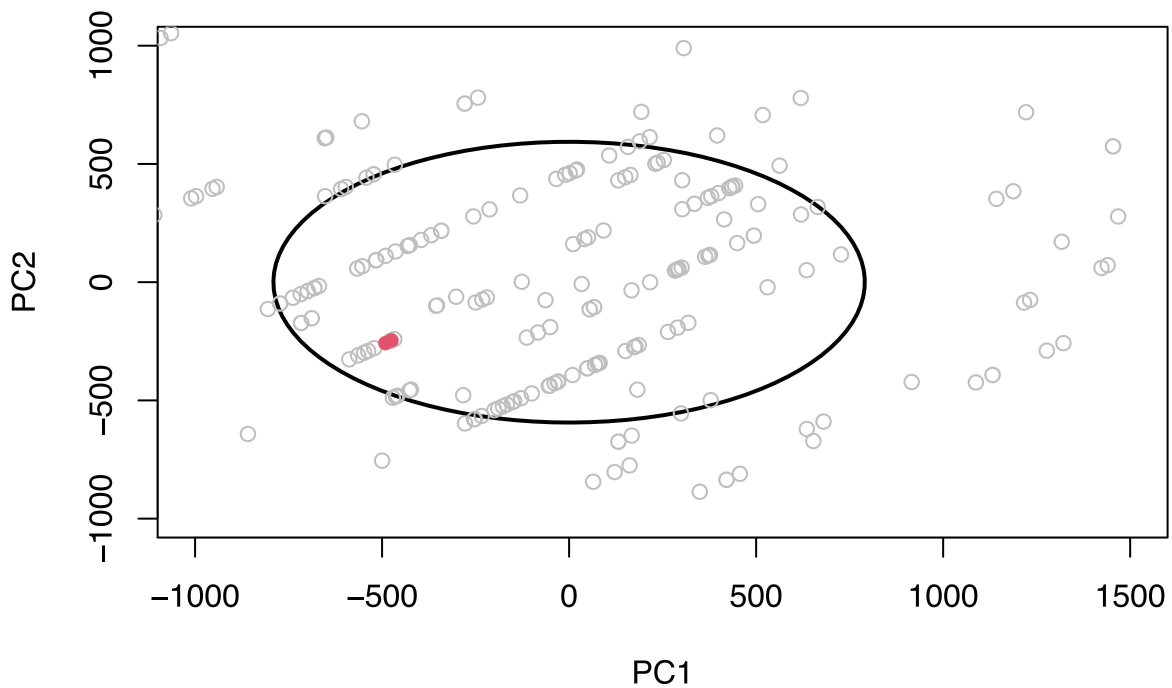
(c)

計算 statistical distances 後，觀察到只有兩個資料點小於 1.4^2

```
##   [1] 17.61 17.61 25.82  4.27  6.11  6.70 24.35 22.87 25.93 13.03 10.15 10.15
##  [13] 15.46 14.92  9.40 11.24  8.85  9.58 24.84  8.55  8.65  9.18  5.28  7.74
##  [25]  5.59  5.75  5.75  7.84 15.11 11.16 39.17 15.81 11.65 14.61 14.61  9.01
##  [37] 22.28  4.85  4.89  7.34 16.81  7.94  3.96 14.21  8.55  8.81  8.42 28.72
##  [49] 28.72 49.98 19.08  4.92  4.88  4.89  4.45 17.54 17.54 17.54 17.69  4.86
##  [61]  5.42  4.86  5.42 15.49  5.25 18.04 22.07 13.62 19.78 15.47 15.38 18.65
##  [73] 43.92 42.49 38.15 17.71  9.64  5.56  5.84  6.90  8.03  8.49 10.38 11.37
##  [85] 11.45  3.93  3.90  3.76  3.76  4.16 20.28  3.55  3.19  4.36  2.98  1.76
##  [97]  2.70  4.23  1.71  3.12  3.31  5.49  9.18  5.19 15.95 18.32  6.75  7.92
## [109] 11.56 14.95 21.32 22.72 12.40 33.99 21.39  8.90 12.40 10.51  9.38  7.74
## [121]  5.59  7.05  9.22 15.14 10.42 22.32 25.43 25.43 24.85 75.77 18.19 15.00
## [133]  7.42  7.24 51.09  7.32 13.91 13.73 15.86 17.46 21.11 12.74 12.68  7.07
## [145] 18.46  7.97 10.03  6.75  8.57 24.90  6.85  7.32  7.22 13.38 13.85 59.83
## [157]  3.44  3.48 20.19 20.20 17.62  3.19  3.07  4.21  4.21 19.55 19.54  5.33
## [169]  5.39  5.18  4.46  4.61  9.55 14.16 17.29 12.07 11.94 11.25  8.02  6.39
## [181] 10.16 10.06 20.16  4.38 20.16  4.34  3.71 15.88  3.94 11.98  4.46  5.75
## [193] 16.63  6.31  8.54 11.08 10.35 12.80 13.13 14.86 11.27 13.91 12.06 23.85
## [205] 11.32
```

故有 $\frac{2}{205} \;=\; 0.98\%$ 的觀測值落在該區間。

(d)



- 資料點的呈現看起來有像是數條正斜率的平行線分佈，是因為 $0.812\hat{y}_2 - 0.576\hat{y}_1 = 0.995 peakrpm$，而變數 *peakrpm* 為一離散變數，有 23 個 levels。

- 相較於原始 13 個變數都考慮時的狀況，現在只考慮前兩個 principal components 時，statistical distance $< 1.4^2$ 的資料點明顯多了很多，是因為使用 principal component 的方法將原本 13 維度的資料點投影到由 $\hat{y}_1$ 和 $\hat{y}_2$ 所形成的 2 維空間 (也即變數 *curbweight* 和 *peakrpm* 所形成的 2 維空間)，這樣的行為將很多的資料點投影到了中間，進而落在橢圓之中。

- 原本 13 的變數計算 statistical distance $< 1.4^2$ 的兩個資料點 (即上圖紅點)，也一樣落進了橢圓之中，因為我們所選的前兩個 principal components 捕捉到了大部分的資料變異特徵。

(e)

計算 $r_{PC1,price}$ 和 $r_{PC2,price}$ 的數值如下

```
## [1] 0.671 0.520
```