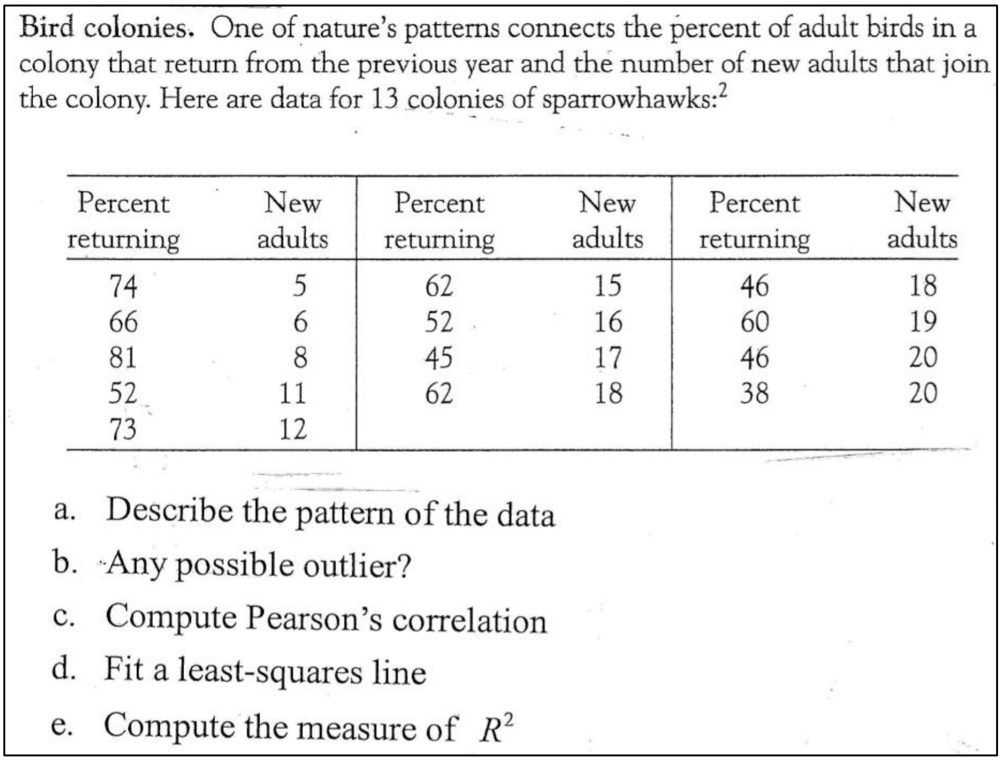
統計應用方法 Homework 1

數據所 姓名：賴品儒 學號：0656704

第一題



Bird colonies. One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:[2]

Percent returning	New adults	Percent returning	New adults	Percent returning	New adults
74	5	62	15	46	18
66	6	52	16	60	19
81	8	45	17	46	20
52	11	62	18	38	20
73	12				

a. Describe the pattern of the data

b. Any possible outlier?

c. Compute Pearson's correlation

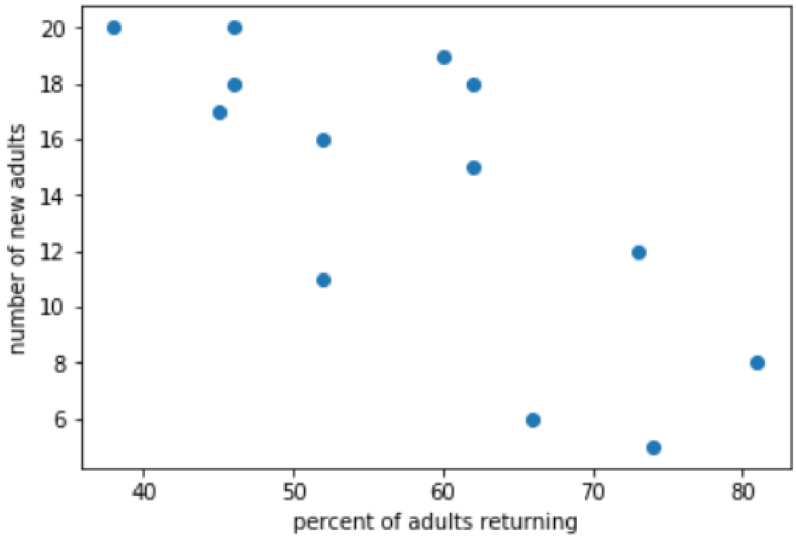
d. Fit a least-squares line

e. Compute the measure of R^2

a. 設前年返回棲地成年鳥類百分比為自變數(X，explanatory variable)

設新成年鳥類數目為因變數(Y，response)

以下為前年返回棲地成年鳥類百分比與該棲地新成年鳥類數量之關係圖：



此圖約略呈現前年返回棲地成年鳥類百分比越大，則新成年鳥類數目越小，但是兩者線性關係(relationship)沒有很明顯。

b. 使用 Studentized residuals (r_i)來偵測 outliers

式子如下：參考

https://newonlinecourses.science.psu.edu/stat501/node/339/

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

MSE=SSE/(n-2)

Hat_matrix=$X(X^T X)^{-1} X^T$

r_i 超過±3可視作 outlier，這些 data 會降低 model 的解釋力 (降低R^2)

整理 data 如下表：

Percent of adults returning	Number of new adults	Studentized Residual
74	5	-1.3518
66	6	-1.6932
81	8	0.2306
52	11	-1.47
73	12	0.6821
62	15	0.5457
52	16	-0.0359
45	17	-0.3735
62	18	1.4004
46	18	0.015
60	19	1.5077
46	20	0.6067
38	20	-0.1224

沒有一筆 data 其 Studentized Residual 絕對值大於 3，故沒有明顯的 outliers

c. Pearson’ s correlation

$$\rho_{X,Y} = corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

設 X：Percent of adults returning

設 Y：Number of new adults

Cov(X,Y)= $\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$ / n =-47.6686

$\bar{X} = 58.2308$ $\bar{Y} = 14.2308$

Var(X)=156.7929 Var(Y)=25.8698

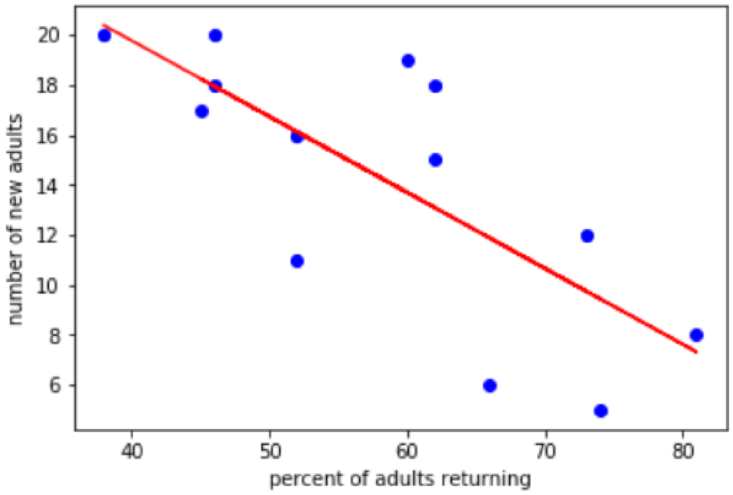
$$\mathrm{r}_{\mathrm{X,Y}} = -\frac{14.6686}{\sqrt{156.7929 * 25.8698}} = -0.7485$$

d. 設迴歸線為：$\hat{\mathrm{Y}} = \alpha + \beta\mathrm{X}$

$$\beta = \mathrm{r}_{\mathrm{X,Y}}\frac{\mathrm{S_Y}}{S_X} = -0.7485 * \frac{\sqrt{25.8698}}{\sqrt{156.7929}} = -0.30402$$

$$\alpha = \bar{Y} - \beta\bar{X} = 14.2308 - (-0.30402) * 52.2308 = 31.9343$$

→ $\hat{\mathrm{Y}} = 31.9343 - 0.30402\mathrm{X}$ (下圖紅線)



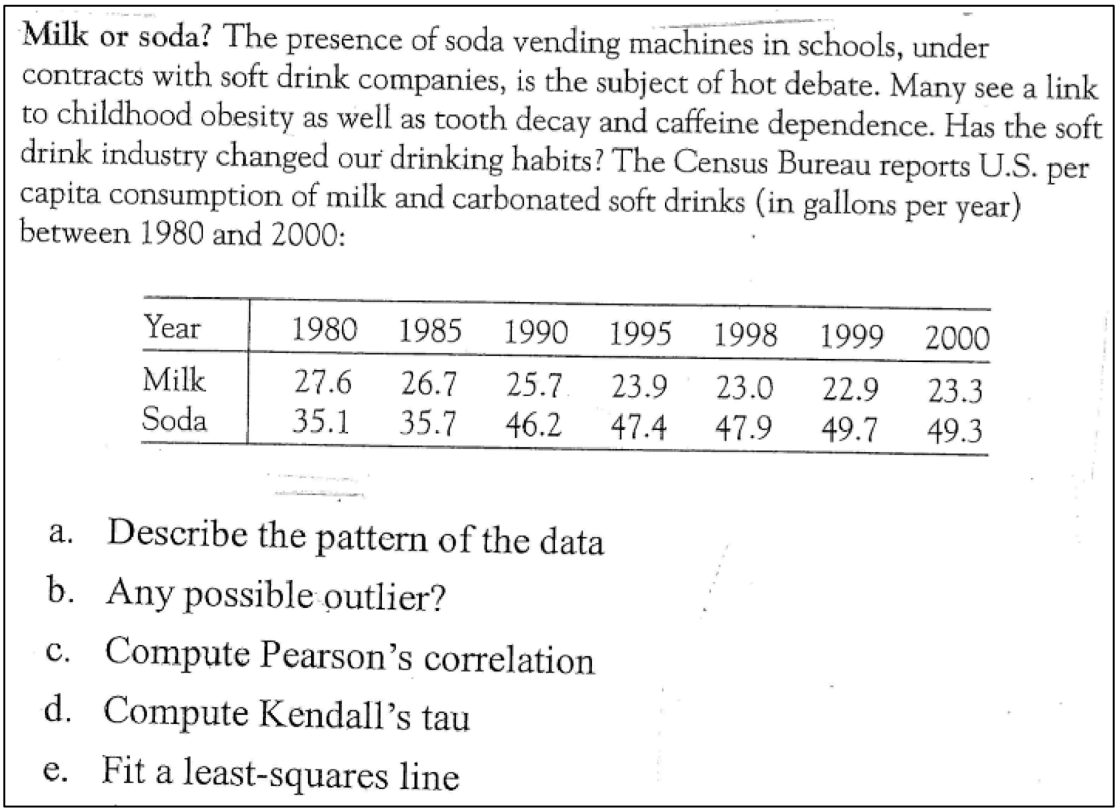
e. $\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{TSS}} = 1 - \frac{\mathrm{SSE}}{\mathrm{TSS}}$

$$\mathrm{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 336.3077$$

$$\mathrm{SSE} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 = 147.90701$$

$$\mathrm{R}^2 = 1 - \frac{147.90701}{336.3077} = 0.5602$$

第二題



Milk or soda? The presence of soda vending machines in schools, under contracts with soft drink companies, is the subject of hot debate. Many see a link to childhood obesity as well as tooth decay and caffeine dependence. Has the soft drink industry changed our drinking habits? The Census Bureau reports U.S. per capita consumption of milk and carbonated soft drinks (in gallons per year) between 1980 and 2000:

Year	1980	1985	1990	1995	1998	1999	2000
Milk	27.6	26.7	25.7	23.9	23.0	22.9	23.3
Soda	35.1	35.7	46.2	47.4	47.9	49.7	49.3

a. Describe the pattern of the data

b. Any possible outlier?

c. Compute Pearson's correlation

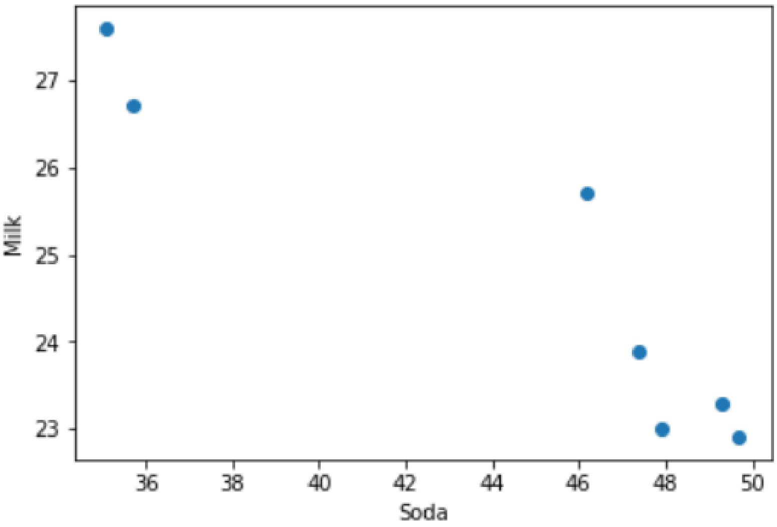
d. Compute Kendall's tau

e. Fit a least-squares line

a. 設 soda 為自變數(X，explanatory variable)

設 milk 為因變數(Y，response)

以下為 soda 和 milk 之關係圖：



由上圖可發現，當 Soda 飲用量越多，Milk 飲用量會有下降的趨勢

b. 用 studentized residual 來偵測 outlier

Soda	Milk	Studentized Residual
35.1	27.6	0.4116
35.7	26.7	-0.8667
46.2	25.7	2.0023
47.4	23.9	-0.003586
47.9	23.0	-1.06867
49.7	22.9	-0.5179
49.3	23.3	-0.09652

沒有一筆 data 其 Studentized Residual 絕對值大於 3，故沒有明顯的 outliers，但有一筆資料(soda, milk)=(46.2, 25.7)其 Studentized residual 為 2.0023>2，相較於其他 data 來的大

c. Pearson’s correlation

設 Soda 為 X (explanatory variable)

設 Milk 為 Y (response)

$\mathrm{Cov(X,Y)} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ / n =-9.6192

$\bar{X} = 44.4714 \quad \bar{Y} = 24.7286$

$\mathrm{Var(X)} = 34.1049 \quad \mathrm{Var(Y)} = 3.16204$

$$\mathrm{r_{X,Y}} = -\frac{9.6192}{\sqrt{34.1049 * 3.16204}} = -0.9263$$

d. Kendall’s tau

$\tau_{X,Y}$ = Pr((i,j) pairs are concordant)- Pr((i,j) pairs are discordant)

$\hat{\tau}_{X,Y}$ =(# of concordance − # of discordance)/K where $K=\binom{n}{2}$

編號	Data (soda, milk)
1	(35.1,27.6)
2	(35.7,26.7)
3	(46.2,25.7)
4	(47.4,23.9)
5	(47.9,23.0)
6	(49.7,22.9)
7	(49.3,23.3)

$$\mathrm{K} = \binom{7}{2} = 21$$

$$\hat{\tau}_{X,Y} = \frac{1}{21} - \frac{20}{21} = -\frac{19}{21} = -0.9048$$

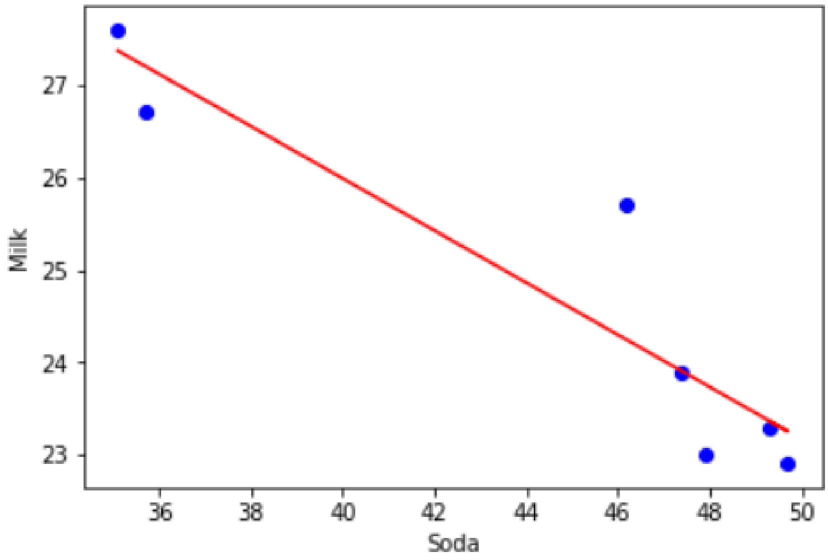
編號配對	Concordant	Discordant
(1,2)		V
(1,3)		V
(1,4)		V
(1,5)		V
(1,6)		V
(1,7)		V
(2,3)		V
(2,4)		V
(2,5)		V
(2,6)		V
(2,7)		V
(3,4)		V
(3,5)		V
(3,6)		V
(3,7)		V
(4,5)		V
(4,6)		V
(4,7)		V
(5,6)		V
(5,7)	V	
(6,7)		V
Total	1	20

e. 設迴歸線： $\hat{Y} = \alpha + \beta X$

$$\beta = r_{X,Y}\frac{S_Y}{S_X} = -0.9263 * \frac{1.7782}{5.8399} = -0.2820$$

$$\alpha = \bar{Y} - \beta\bar{X} = 24.7286 - (-0.2820) * 44.4714 = 37.2716$$

$\rightarrow \ \hat{Y} = 37.2716 - 0.2820X$



全部計算過程：https://github.com/ayamisea/Applied-Methods-in-Statistics/blob/master/homework/HW01.ipynb

第三題

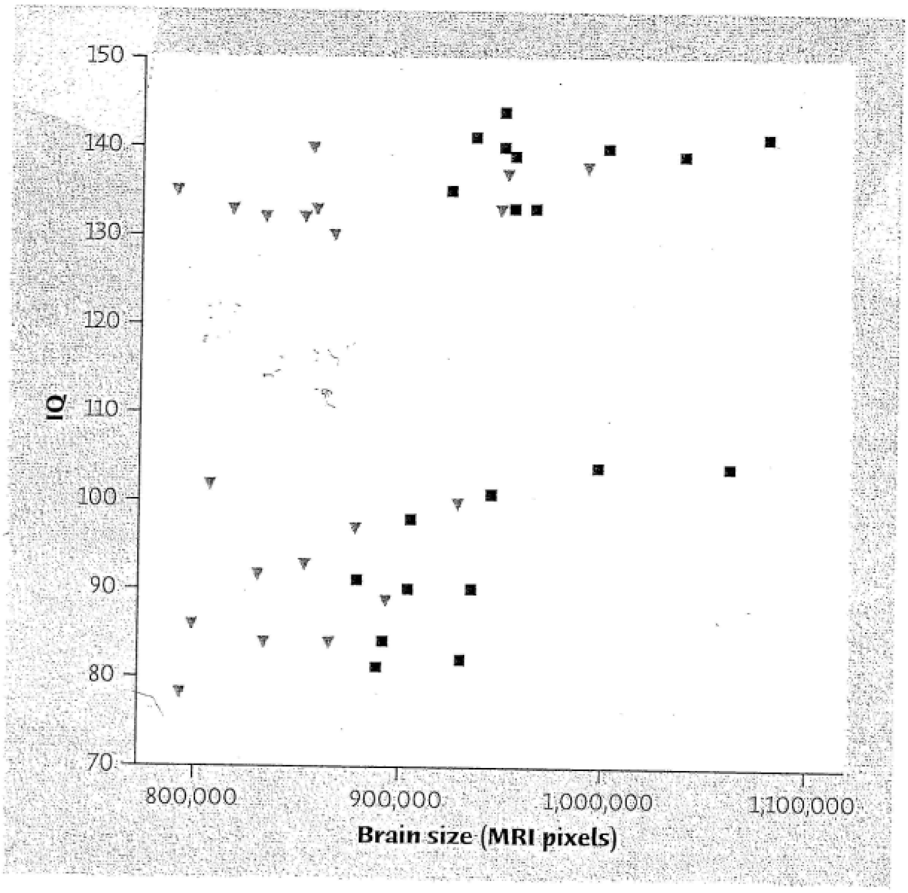


FIGURE 3.7 Scatterplot of the brain size (in MRI pixels) and IQ of 20 subjects selected for their high IQ (130 or more) and 20 selected for their low IQ (100 or less), for Exercise 3.32. Of the 40 subjects, 20 are women (triangles) and 20 are men (squares).

their high IQ (130 or more) and 20 selected for their low IQ (100 or less). The researchers' choice of studying high and low IQ scores only is reflected in the gap on the scatterplot for mid-range IQ scores. Brain size was measured by magnetic resonance imaging (MRI). The MRI count is the number of "pixels" the brain covered in the image. IQ was measured by the Wechsler test. Of the 40 volunteers, 20 were women (triangles) and 20 were men (squares).

(a) Men are larger than women on the average and have larger brains. How is this size effect visible in the plot? Guess the mean MRI count for men and women from Figure 3.7 to verify the difference.

(b) Comment on the nature and strength of the relationship between brain size and IQ for women and then for men. Explain why it makes sense to study men and women separately in this case.

(c) The study included only individuals with high IQ or low IQ. This is reflected on the scatterplot by the separate clusters of points. Explain why this makes interpreting the scatterplot particularly challenging.

<上課討論>