統計應用方法

Homework 1

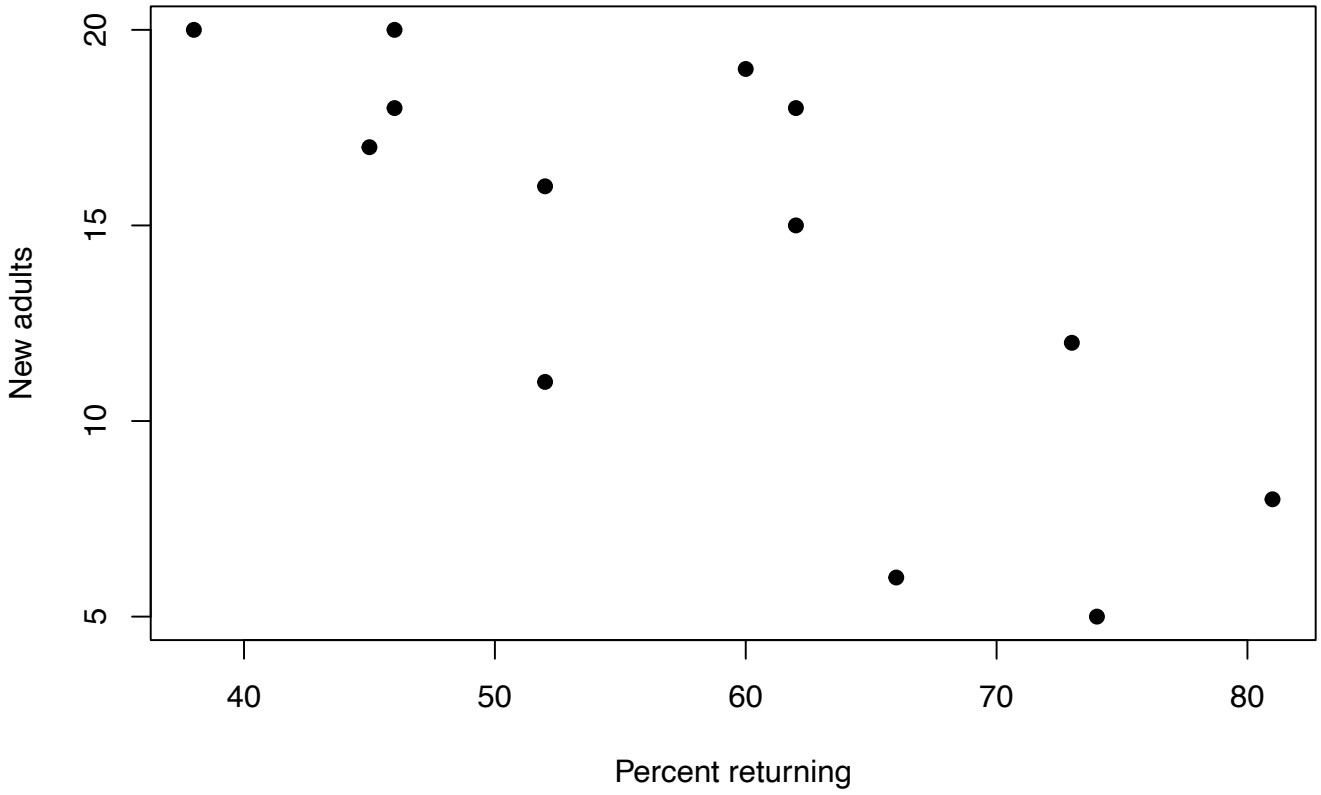
1. (55 points)

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. (10 points) Please plot the data with X indicating "Percent returning" and Y indicating "New adults". Comment on the main features of the plot. Any possible outliers? (可用軟體畫)

Ans : 可以看出"Percent returning"和"New adults"是負向關，下圖看不太出明顯的 outlier，從 residual 也沒有看出來，因此認為沒有明顯的 outlier。




```
> residuals(model)
         1          2          3          4          5          6          7
-4.43656125 -5.86874481  0.69159937 -5.12506604  2.25941580  1.91516341 -0.12506604
         8          9         10         11         12         13
-1.25322666  4.91516341  0.05079629  5.30711752  2.05079629 -0.38138727
```

b. (10 points) Apply the formula to compute $\bar{X}$, $\bar{Y}$, $S_x^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, $S_y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ and Pearson's correlation:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$

Ans：

編號	1	2	3	4	5	6	7	8	9	10	11	12	13	sum	avg
x	74	66	81	52	73	62	52	45	62	46	60	46	38	757	**58.23077**
y	5	6	8	11	12	15	16	17	18	18	19	20	20	185	**14.23077**
x-mean	15.76923	7.769231	22.76923	-6.23076	14.76923	3.769231	-6.23076	-13.2308	3.769231	-12.2308	1.769231	-12.2308	-20.2308		
y-mean	-9.23076	-8.23076	-6.23076	-3.23076	-2.23076	0.769231	1.769231	2.769231	3.769231	3.769231	4.769231	5.769231	5.769231		
Product	-145.5622	-63.9467	-141.8698	20.13017	-32.9467	2.899408	-11.0237	-36.6391	14.20711	-46.1006	8.437870	-70.5621	-116.7160	-619.6923	-47.6687
(x-mean)^2	248.6686	60.36095	518.4379	38.82249	218.1302	14.20711	38.82249	175.0533	14.20711	149.5917	3.130178	149.5917	409.2840	2038.308	156.7929
(y-mean)^2	85.20710	67.74556	38.82248	10.43787	4.976331	0.591716	3.130178	7.668631	14.20711	14.20710	22.74556	33.28402	33.28402	336.3076	25.86982
S_x^2	169.8589														
S_y^2	**28.02564**														
Corr	**-0.7484673**														

計算得到$\bar{X} = 58.23077$, $\bar{Y} = 14.23077$, ${S_x}^2 = 169.85897$, ${S_y}^2 = 28.02564$, and $r = -0.7484673$。

Excel 計算過程：

	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y	Xmean	Ymean	X-Xmean	Y-Ymean	(X-Xmean)^2	(Y-Ymean)^2	S_x^2	S_y^2	(x-Xmean)*(Y-Ymean)	S
2	74	5	58.2307692	14.2307692	15.76923077	-9.230769231	248.6686391	85.20710059	169.858974	28.025641	-145.5621302	-0.7484673
3	66	6	58.2307692	14.2307692	7.769230769	-8.230769231	60.36094675	67.74556213			-63.94674556	
4	81	8	58.2307692	14.2307692	22.76923077	-6.230769231	518.4378698	38.82248521			-141.8698225	
5	52	11	58.2307692	14.2307692	-6.230769231	-3.230769231	38.82248521	10.43786982			20.13017751	
6	73	12	58.2307692	14.2307692	14.76923077	-2.230769231	218.1301775	4.976331361			-32.94674556	
7	62	15	58.2307692	14.2307692	3.769230769	0.769230769	14.20710059	0.591715976			2.899408284	
8	52	16	58.2307692	14.2307692	-6.230769231	1.769230769	38.82248521	3.130177515			-11.02366864	
9	45	17	58.2307692	14.2307692	-13.23076923	2.769230769	175.0532544	7.668639053			-36.63905325	
10	62	18	58.2307692	14.2307692	3.769230769	3.769230769	14.20710059	14.20710059			14.20710059	
11	46	18	58.2307692	14.2307692	-12.23076923	3.769230769	149.591716	14.20710059			-46.10059172	
12	60	19	58.2307692	14.2307692	1.769230769	4.769230769	3.130177515	22.74556213			8.437869822	
13	46	20	58.2307692	14.2307692	-12.23076923	5.769230769	149.591716	33.28402367			-70.56213018	
14	38	20	58.2307692	14.2307692	-20.23076923	5.769230769	409.2840237	33.28402367			-116.7159763	
15	sum	sum					sum	sum			sum	
16	757	185					2038.307692	336.3076923			-619.6923077	
17	avg	avg					avg	avg			avg	
18	58.2307692	14.2307692					156.7928994	25.86982249			-47.66863905	

R 計算過程：

```
> Percent_returning <- c(74, 66, 81, 52, 73, 62, 52, 45, 62, 46, 60, 46, 38)
> New_adults <- c(5, 6, 8, 11, 12, 15, 16, 17, 18, 18, 19, 20, 20)
> # b.
> Percent_returning -> x
> New_adults -> y
> x_mean <- mean(x) # X-Bar
> ## [1] 58.23077
> y_mean <- mean(y) # Y-Bar
> ## [1] 14.23077
> S_sub_x_square <- sum((x-mean(x))^2)/(length(x)-1)
> ## [1] 169.859
> S_sub_y_square <- sum((y-mean(y))^2)/(length(y)-1)
> ## [1] 28.02564
> # Pearson's correlation(r)
> r <- sum((x-mean(x))*(y-mean(y)))/(sqrt(sum((x-mean(x))^2))*sqrt(sum((y-mean(y))^2)))
> r
[1] -0.7484673
```

c. (10 points) Use R (軟體) to compute Pearson's correlation, Kendall's tau and Spearman's rho. (相關係數指令在講義)

Ans：

```
> cor(Percent_returning, New_adults, method = "pearson") # Pearson's correlation
[1] -0.7484673
> cor(Percent_returning, New_adults, method = "kendall") # Kendall's tau
[1] -0.5960396
> cor(Percent_returning, New_adults, method = "spearman") # Spearman's rho
[1] -0.7538043
```

d. (15 points) Please fit the regression model:

$$Y = \alpha + \beta X + \varepsilon .$$

Find $\hat{\alpha}$, $\hat{\beta}$ (用上課給的公式) and

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

Ans：計算得到$\hat{\alpha} = -0.3040229$, $\hat{\beta} = 1.93426$, and $\sigma^2 = 13.44609$。

先求 $\hat{\beta} = r \frac{S_y}{S_x} = -0.7484673 \cdot \frac{\sqrt{28.02564}}{\sqrt{169.8589}} = -0.3040229$

再求 $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 14.2307) - (-0.3040229 * 58.2307)) = 31.93426$

$\sigma^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 / (n-2) = \sum_{i=1}^{13} (Y_i - \hat{\alpha} - \hat{\beta} X)^2 / 11$

$= \frac{141.307}{11} = 13.44609$

R 計算過程：

```
> beta_estimate <- r*sqrt(S_sub_y_square)/sqrt(S_sub_x_square)
> ## [1] -0.3040229
> alpha_estimate <- y_mean-(beta_estimate*x_mean)
> ## [1] 31.93426
> sigma_estimate_squre <- sum((y-alpha_estimate-(beta_estimate*x))^2)/(length(x)-2)
> sigma_estimate_squre
[1] 13.44609
```

e. (5 points) What is the value of

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} ?$$

Is it true that $R^2 = r^2$

Ans：

$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{147.907}{336.3077} = 0.5602033$ ※

已知 $r = -0.7484603$，$r^2 = 0.5602033$

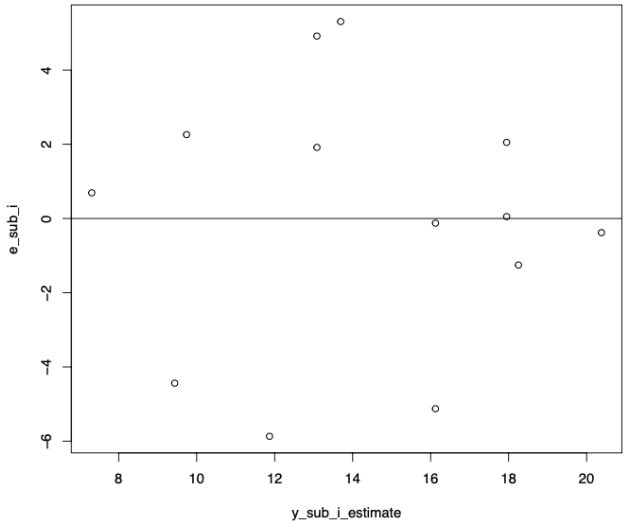
$\therefore R^2 = r^2$，is simple regression

R 計算過程：

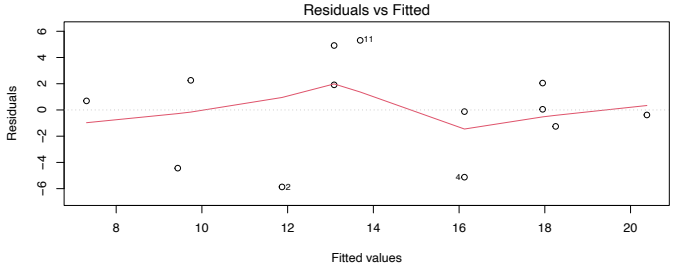
```
> R_squre <- 1 - sum((y-alpha_estimate-(beta_estimate*x))^2)/sum((y-y_mean)^2)
> r_squre <- r^2
> R_squre
[1] 0.5602033
> r_squre
[1] 0.5602033
```

f. (5 points) Please plot $e_i = Y_i - \hat{Y}_i$ (Y axis) versus $\hat{Y}_i$ (X axis). Comment on the residual plot. What does this plot indicate?

Ans：下圖自己畫的可以看到所有的點算是均勻地散落在 0 的兩側，用 R library 指令檢查發現雖然不是完全分布均勻但也沒有相差太多，因此算是符合假設的情況。



*y_sub_i_estimate 改為 y_sub_i_estimator



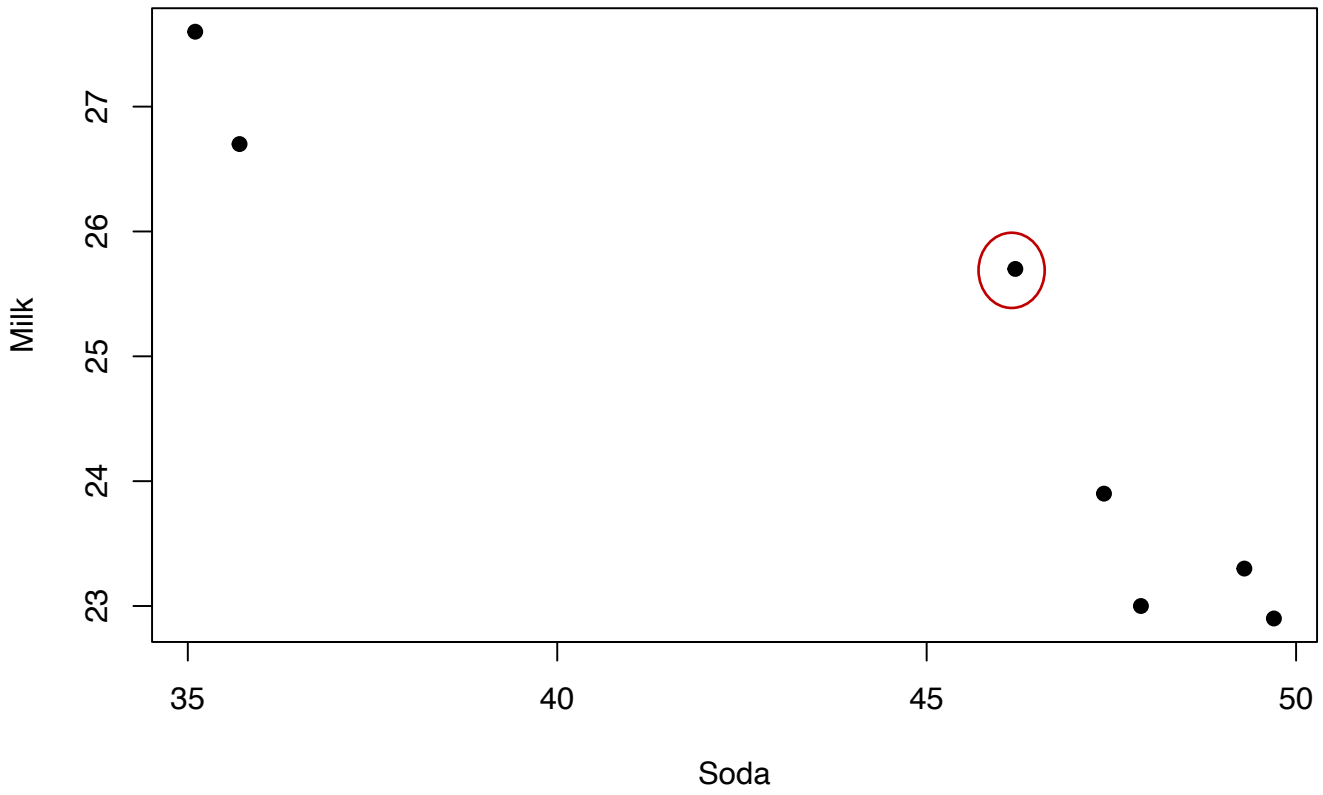
2. (45 points) 全部用軟體的指令

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. (10 points) Please plot the data with X indicating "Soda" and Y indicating "Milk". Comment on the main features of the plot. Any possible outliers? 這裡我建議把 Milk 當成 Y, 是因為牛奶攝取是營養學家或家長比較在意的.

Ans：Milk 與 Soda 的關係為負相關，從 residuals 可以看到第 3 明顯與其他離群，所以 3 可能是 outlier。




```
> residuals(model)
          1           2           3           4           5           6           7
0.228245389 -0.502526419  1.458966933 -0.002576684 -0.761553191 -0.353868617 -0.066687411
```

b. (10 points) Use R (軟體) to compute Pearson's correlation, Kendall's tau and Spearman's rho.

Ans：

```
> cor(Soda, Milk, method = "pearson") # Pearson's correlation
[1] -0.9262881
> cor(Soda, Milk, method = "kendall") # Kendall's tau
[1] -0.9047619
> cor(Soda, Milk, method = "spearman") # Spearman's rho
[1] -0.9642857
```

c. (10 points) Use R to fit the regression model:

$$Y = \alpha + \beta X + \varepsilon .$$

Show your output and explain the meaning of β.

Ans：可以知道α為 37.27160, β為-0.28205，β為斜率的估計值，意為當 Soda 每改變±1 單位時，那麼 Milk 的改變量為-0.28205。

```
> model <- lm(Milk ~ Soda)
> summary(model)

Call:
lm(formula = Milk ~ Soda)

Residuals:
       1        2        3        4        5        6        7
 0.228245 -0.502526  1.458967 -0.002577 -0.761553 -0.353869 -0.066687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.27160    2.30151  16.194 1.64e-05 ***
Soda        -0.28205    0.05131  -5.497  0.00272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7928 on 5 degrees of freedom
Multiple R-squared:  0.858,     Adjusted R-squared:  0.8296
F-statistic: 30.21 on 1 and 5 DF,  p-value: 0.002722
```

d. (15 points) Show the residual plot and normal plot. Make your comments.

Hint: You need to know the purposes of these two plots and judge whether the plots support the assumptions that

$$\varepsilon_i \sim^{iid} N(0,\sigma^2).$$

Ans：從 residual plot 可以看出資料並沒有均勻地散落在 0 的兩側，表示資料可能不滿足線性假設，而在 normal plot 可以看到是常態分佈滿足上方的假設也能夠看出第 3 點為 outlier 與其他點分離太遠。