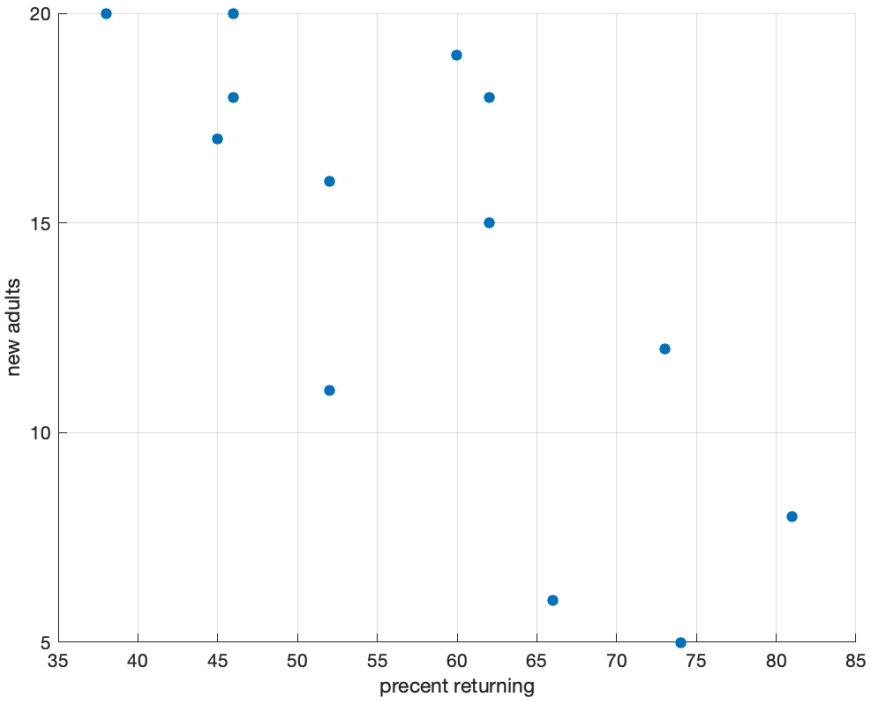
Applied Methods in Statistics

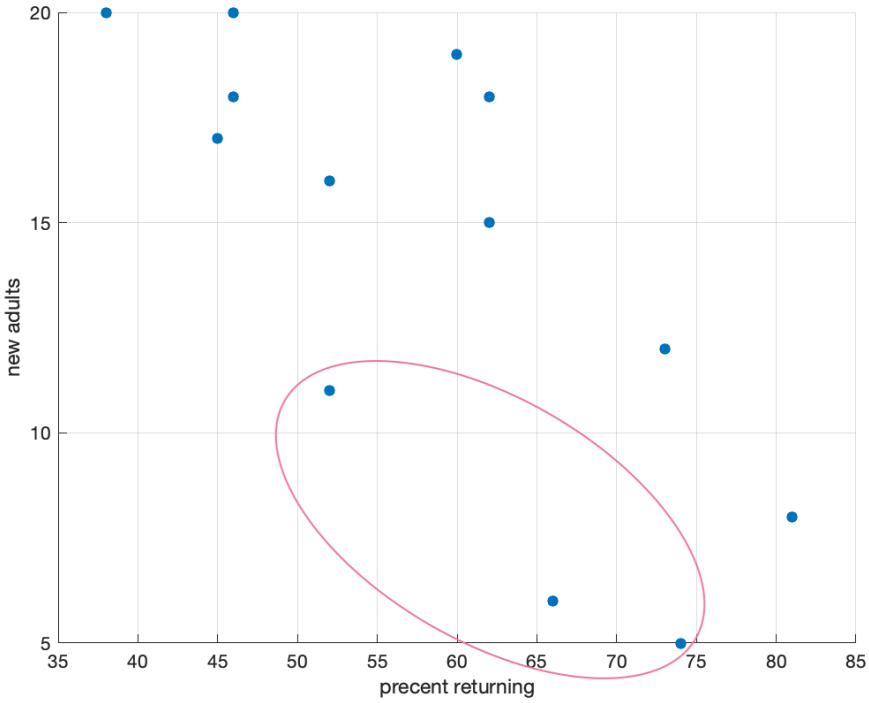
109550121 温柏萱

Problem 1

(a)



The increase of the percent returning roughly presents a decrease in the number of new adults, however, the points are sparsely distributed, there may be 3 potential and obvious outliers circled in the picture below. From an intuitive observation, the percent returning has a negative relation with the number of new adults.



(b)

$\bar{X}$	$\bar{Y}$	S_x^2	S_y^2	r
58.2308	14.2308	169.859	28.0256	-0.74847

```
xi = [74 66 81 52 73 62 52 45 62 46 60 46 38];
yi = [5 6 8 11 12 15 16 17 18 18 19 20 20];

x_mean = sum(xi) / length(xi);
y_mean = sum(yi) / length(yi);
disp(['x mean: ', num2str(x_mean)]);
disp(['y mean: ', num2str(y_mean)]);

S_x_squared = sum((xi - x_mean).^2) ./ (length(xi) - 1);
disp(['S_x^2: ', num2str(S_x_squared)]);

S_y_squared = sum((yi - y_mean).^2) ./ (length(yi) - 1);
disp(['S_y^2: ', num2str(S_y_squared)]);

r_ = sum((xi - x_mean) .* (yi - y_mean)) ./ (sqrt(sum((xi - x_mean).^2)) .* sqrt(sum((yi
- y_mean).^2)));
disp(['r_', num2str(r_)]);
r = corrcoef(xi, yi);
disp(['r: ', num2str(r(1, 2))]);
```

$$
\begin{aligned}
\bar{X} &= 58.2308 \\
\bar{Y} &= 14.2308 \\
S_x^2 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2}{n-1} = \frac{46119 - 2 \times 58.2308 \times 757 + 13 \times 3391}{12} = 169.859 \\
S_y^2 &= \frac{2969 - 2 \times 14.2308 \times 185 + 13 \times 2025}{12} = 28.0256 \\
r &= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})}} = \frac{10153 - 10773}{\sqrt{2038}\sqrt{336}} = -0.74847
\end{aligned} \tag{1}
$$

(c)

```
> correlation <- cor(df$percent_returning, df$new_adults, method = 'pearson')
> correlation
[1] -0.7484673
> tau <- cor(df$percent_returning, df$new_adults, method = 'kendall')
> tau
[1] -0.5960396
> rho <- cor(df$percent_returning, df$new_adults, method = 'spearman')
> rho
[1] -0.7538043
```

Pearson Correlation r	Kendall correlation τ	Spearman's rank correlation coefficient ρ
-0.7484673	-0.5960396	-0.7538043

(d)

```
> linear=lm(formula=df$new_adults ~ df$percent_returning)
> linear

Call:
lm(formula = df$new_adults ~ df$percent_returning)

Coefficients:
         (Intercept)  df$percent_returning
              31.934                -0.304

> summary(linear)

Call:
lm(formula = df$new_adults ~ df$percent_returning)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8687 -1.2532  0.0508  2.0508  5.3071

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          31.93426    4.83762   6.601 3.86e-05 ***
df$percent_returning -0.30402    0.08122  -3.743  0.00325 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.667 on 11 degrees of freedom
Multiple R-squared:  0.5602, Adjusted R-squared:  0.5202
F-statistic: 14.01 on 1 and 11 DF,  p-value: 0.003248
```

α	β	$\hat{\sigma}^2$
31.93426	-0.30402	$(3.667)^2 = 13.446889$

Regression Model:

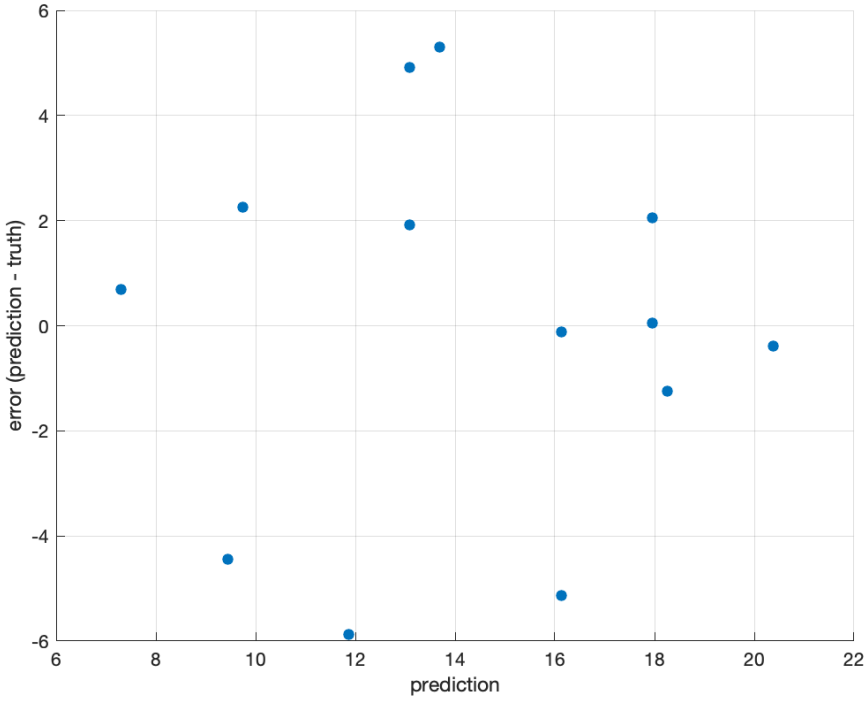
$$Y = 31.93426 - 0.30402X + \epsilon \tag{2}$$

(e)

R^2	r^2
0.5602	0.5602

It is true that $R^2 = r^2$,

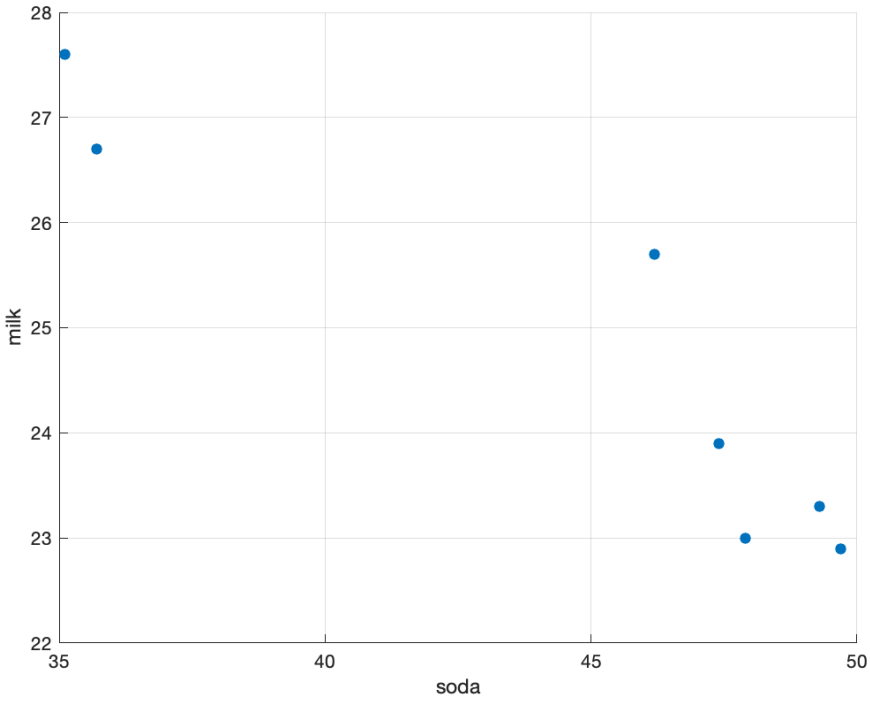
(f)



The plot of $(\hat{y}_i, y_i - \hat{y}_i)$ is the residual plot, used to present the residuals in relation to the predicted values. An ideal residual plot has points randomly distributed, indicating that the regression model successfully captured the the relation between x and y without systematic error. The residual plot of the fitted regression model has the residuals randomly distributed on the residual plot, which indicates the noise follow the assumption $\epsilon \sim^{iid} (0, \sigma^2)$.

Problem 2

(a)



The plot shows a negative correlation between soda and milk consumption. As soda consumption increases, milk consumption tends to decrease. The data points are spread out, which indicates variability in the relationship between milk and soda consumption. There's not an obvious cluster around any specific line, suggesting other factors may also influence consumption patterns.

(b)

```
> cor2 <- cor(df2$Soda, df2$Milk, method='pearson')
> cor2
[1] -0.9262881
> tau2 <- cor(df2$Soda, df2$Milk, method='kendall')
> tau2
[1] -0.9047619
> rho2 <- cor(df2$Soda, df2$Milk, method='spearman')
> rho2
[1] -0.9642857
```

Pearson Correlation r	Kendall correlation τ	Spearman's rank correlation coefficient ρ
-0.9262881	-0.9047619	-0.9642857

(c)

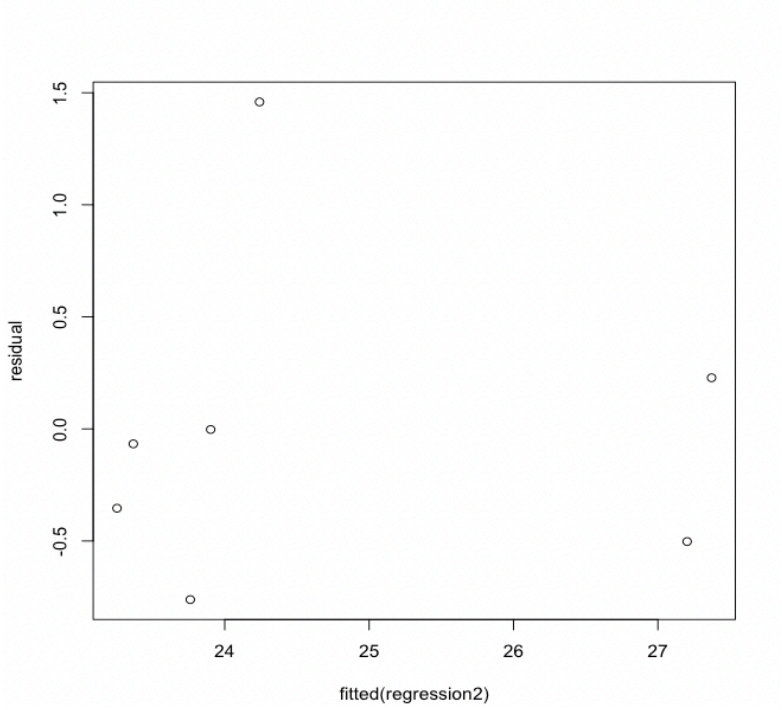
```
> r2 = lm(df2$Milk ~ df2$Soda)
> r2

Call:
lm(formula = df2$Milk ~ df2$Soda)

Coefficients:
(Intercept)     df2$Soda
     37.272       -0.282

> summary(r2)

Call:
lm(formula = df2$Milk ~ df2$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 ***
df2$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
```

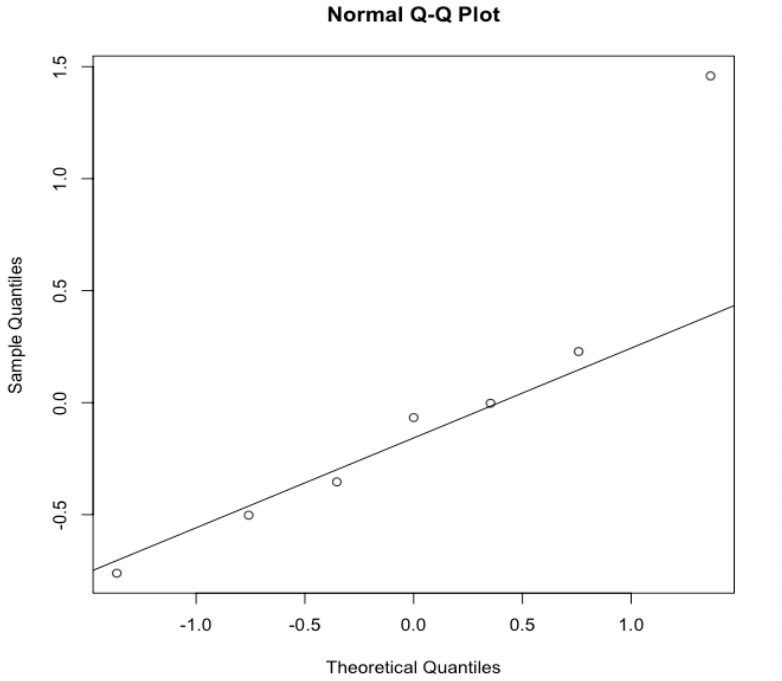
β is the slope of the regression model, indicating the influence on the predicted results when the input increased one unit. In this case, $\beta = -0.28205$, implies that when Soda grows for 1 unit, the prediction, in this case, Milk, will drop for 0.28205 unit.

Regression Model:

$$Y = 37.27160 - 0.28205X + \epsilon \tag{3}$$

(d)





The residual plot is used to show the underlying pattern of the residual, the more randomly distributed, the better that the residual term are normally distributed, supporting the assumption of $\epsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$. In this case, the residuals are scattered but around $[-0.5, 0.5]$, 1 data point was at 1.5. However, it is difficult to determine the randomness from the residual plot by human eye, therefore, the normal plot is established to provide a more intutive observation on the distribution of the residual term and how much it supports the assumption $\epsilon \sim^{iid} \mathcal{N}(0, \sigma^2)$.

In the normal plot, the x axis is the theoretical quantiles of the normal distribution, the y axis is the residual, the points locating near a straight line in the normal plot indicates the data follows a normal distribution, supporting the assumption of $\epsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$ holds. In this case, only one sample was far from the line, others located very to the line, therefore, the assumption $\epsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$ roughly holds.