Statistical Linear Model Homework 1

12112627 李乐平

Question 1

(a) Construct and comment a scatterplot of the data.

(b) Find the least squares line from the data and plot it on your scatterplot.

Solution of questions (a) and (b): We know that the least square estimations of parameter β_0 and β_1 in the linear model $y_i = \beta_0 + \beta_1 x_i$ are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The numeric calculations are as follows.

In [1]:

```
import matplotlib.pyplot as plt
import scipy.stats as stats  # Only to get the cdf of t-distribution
import numpy as np
from math import *  # Only to calculate the square root
```

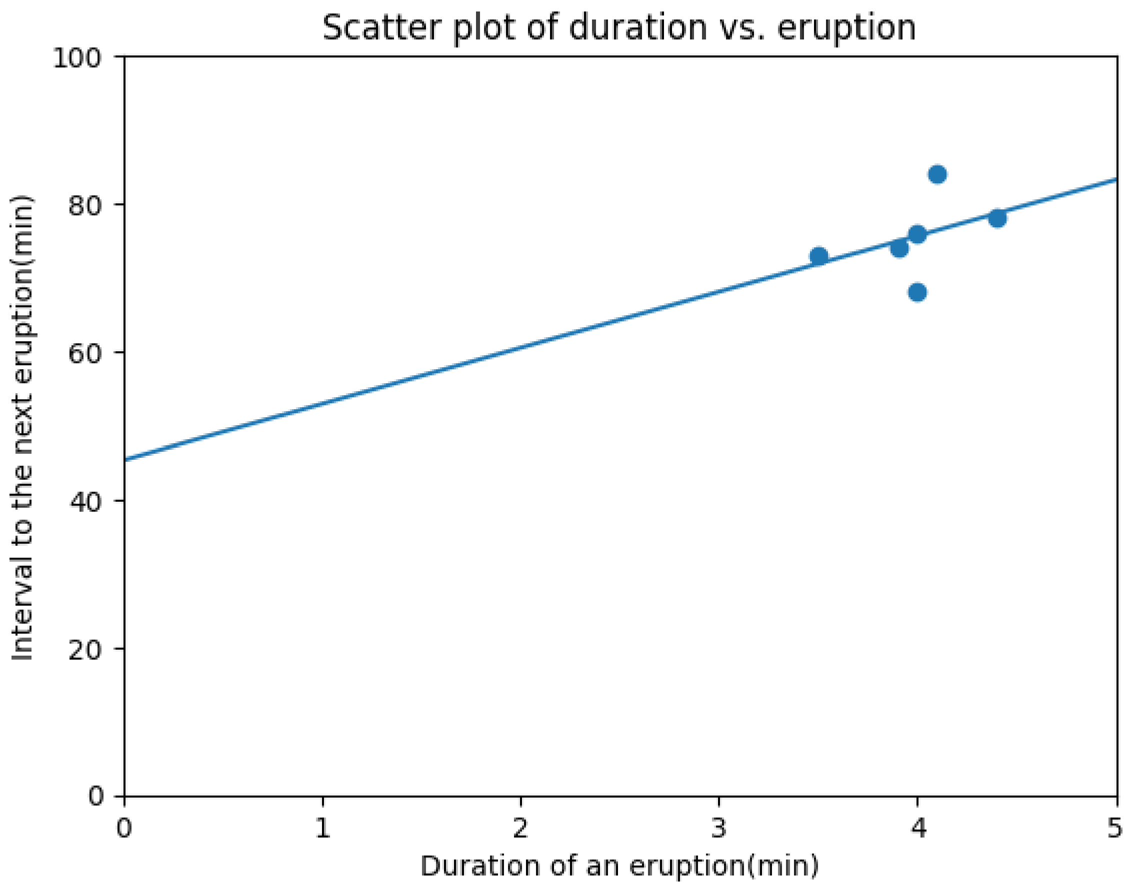
In [2]:

```
def avg(arr):
    # Calculate the average of an array

    l = len(arr)
    ans = 0
    for i in arr:
        ans += i / l
    return ans

def lm(xs, ys):
    # Calculate all the statistics and estimations needed for linear model

    x_avg = avg(xs)
    y_avg = avg(ys)

    Sxx = sum((xi - x_avg) ** 2 for xi in xs)
    Sxy = sum((x[i] - x_avg) * (y[i] - y_avg) for i in range(len(xs)))

    # The estimators of beta 1 and beta 0
    b1 = Sxy / Sxx
    b0 = y_avg - b1 * x_avg

    # Other statistics and estimators
    y_eva = [b0 + b1 * xi for xi in x]
    SSE = sum((y[i] - y_eva[i]) ** 2 for i in range(len(ys)))
    SSR = sum((y_eva[i] - y_avg) ** 2 for i in range(len(ys)))
    SST = SSE + SSR
    S2 =  SSE / (len(xs) - 2)
    COD = SSR / SST

    return {
        "b0": b0,
        "b1": b1,
        "Sxx": Sxx,
        "Sxy": Sxy,
        "S2": S2,
        "y_eva": y_eva,
        "SSE": SSE,
        "SSR": SSR,
        "SST": SST,
        "COD": COD
    }

def predict(xh, b0h, b1h, n, t, x_avg, Sxx, S2):
    # Calculate the estimation of response variable yh and its confidence interval

    yh = b0h + b1h * xh
    S = sqrt(S2)
    tmp = t * S * sqrt(1 + 1 / n + (xh - x_avg) ** 2 / Sxx)
    lb = yh - tmp
    rb = yh + tmp
    return yh, (lb, rb)
```

In [3]:
```
x = [4.4, 3.9, 4.0, 4.0, 3.5, 4.1]
y = [78, 74, 68, 76, 73, 84]
n = 6
a = 0.05

res = lm(x, y)
b0 = res["b0"]
b1 = res["b1"]
Sxx = res["Sxx"]
Sxy = res["Sxy"]
S2 = res["S2"]
y_eva = res["y_eva"]
R2 = res["COD"]

lx = np.linspace(0, 6, 100)
ly = b0 + b1 * lx
```

In [4]:
```
# Draw the scatter plot
plt.scatter(x, y)
plt.title("Scatter plot of duration vs. eruption")
plt.xlabel("Duration of an eruption(min)")
plt.ylabel("Interval to the next eruption(min)")
plt.xlim(0, 5)
plt.ylim(0, 100)


# Put the least square line onto the plot
print(f"Least square line: y = {b0} + {b1}x")
plt.plot(lx, ly)

plt.show()
```

```
Least square line: y = 45.27626459143973 + 7.587548638132287x
```



(c) What is your linear regression model? State the necessary assumptions.

Solution: My linear regression model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, \dots, n$, and I assume that the error terms $\epsilon_i \sim N(0, \sigma^2)$, and error terms are mutually independent. The condition is stronger than ordinary LSE-based LRM because we are required to construct confidence intervals later.

(d) Test the hypothesis that the duration of an eruption has no effect of the interval to the next eruption when a linear model is used (use α = 0.05). State the null and alternative hypotheses. Draw the appropriate test conclusions.

(e) Find a 95% confidence interval for β1 (the slope of the linear regression model). Interpret your results.

Solution of questions (d) and (e): If the duration of an eruption has no effect of the interval to the next eruption when a linear model is used, we can frame the test as follows:

$H_0 \colon \beta_1 = 0$ against $H_1 \colon \beta_1 \neq 0$

The test statistic is:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{S}} \sqrt{\mathrm{S}_{\mathrm{xx}}} = \frac{\hat{\beta}_1}{\mathrm{S}} \sqrt{\mathrm{S}_{\mathrm{xx}}} \sim t(n-2)$$

if H_0 is true, where

$$\mathrm{S} = \sqrt{\frac{\sum_{i=0}^{n} (y_i - \hat{y}_i)^2}{n-2}}, \mathrm{S}_{\mathrm{xx}} = \sum_{i=0}^{n} (x_i - \bar{x})^2$$

Namely we reject H_0 if $|t| > t(\alpha/2, n-2)$.

Hence the confidence interval of β_1 is given by

$$[\hat{\beta}_1 - \frac{\mathrm{S} \cdot t(\alpha/2, n-2)}{\sqrt{\mathrm{S_{xx}}}}, \hat{\beta}_1 + \frac{\mathrm{S} \cdot t(\alpha/2, n-2)}{\sqrt{\mathrm{S_{xx}}}}]$$

To calculate the concrete values of these, please refer to the codes following.

In [5]:
```
t_value = stats.t.ppf(1 - a / 2, df = n - 2)  # t(0.05, 4) ≈ 2.7764
CI = (b1 - sqrt(S2) * t_value / sqrt(Sxx), b1 + sqrt(S2) * t_value / sqrt(Sxx))

# 95% Confidence interval of b1
print(CI)
```

```
(-15.535779052203974, 30.710876328468547)
```

The confidence interval shows there is a probability of 95% that the true population of β_1 falls in $[-15.5358, 30.7109]$. Considering $\beta_1 = 0 \in [-15.5358, 30.7109]$, so we cannot reject H_0 at the significance level of $1 - \alpha = 0.95$.

(f) Find the coefficient of determination for the linear regression model. Interpret your result.

Solution: The coefficient of determination is given by

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \approx 0.1718$$

This result means that there is only 0.1718 proportion of the variation can be explained by the model.

In [6]:
```
# Coefficient of Determination
R2 = res["COD"]
print(R2)
```

```
0.1718434360552612
```

(g) Find a prediction of the time to the next eruption when the Geyser eruption lasts for 4 minutes and its 95% interval.

Solution: The prediction on time to the next eruption $y_h = 75.6265$ minutes, with 95% confidence interval $[59.2758, 91.9771]$.

In [7]:
```
# Predictive interval
yh, CI_yh = predict(4.0, b0, b1, n, t_value, avg(x), Sxx, S2)
print(yh, CI_yh)
```

```
75.62645914396887 (59.275797330533436, 91.97712095740431)
```

Question 2

(a) Define a simple linear regression model and derive MLE (maximum likelihood estimation) for all the unknown parameters.

Solution: (a) Assume $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2), i = 1, 2, \ldots, n$ and error terms are mutually independent. That is to say, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, where β_0, β_1 and σ^2 are unknown variables.

The joint likelihood function is

$$L(\beta_0, \beta_1, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n}\exp(-\sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2})$$

and the joint log-likelihood function is

$$l(\beta_0, \beta_1, \sigma^2) = -\sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} - n\ln(\sqrt{2\pi}\sigma)$$

where

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = \sum_{i=1}^{n}\frac{x_i(y_i - \beta_0 - \beta_1 x_i)}{\sigma^2}, \frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = \sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma^2}, \frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^4} - \frac{n}{2\sigma^2}$$

Let

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_0} = 0$$

and get

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Hence that

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = \sum_{i=1}^{n}\frac{x_i(y_i - \bar{y}) + \beta_1 x_i(\bar{x} - x_i)}{\sigma^2}$$

Let

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \beta_1} = 0$$

and get the MLE of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Correspondingly, the MLE of β_0 is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{\sum_{i=1}^n x_i \bar{x}(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})}$$

Let

$$\frac{\partial l(\beta_0, \beta_1, \sigma^2)}{\partial \sigma^2} = 0$$

and get the MLE of σ^2 is

$$\hat{\sigma^2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

where

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

(b) Comments on the difference between MLE and LSE (least square estimation)

Solution: It is easy to discover that

$$\beta_0^{\text{MLE}} = \beta_0^{\text{LSE}}, \beta_1^{\text{MLE}} = \beta_1^{\text{LSE}}, \sigma^2_{\text{MLE}} = \frac{n-2}{n}\sigma^2_{\text{LSE}}$$

.

.