STA2002 - Homework 5

Xue Zhongkai 122090636

Problem 1.

(a) With relative formulas, the ANOVA table is as follows:

Source	SS	df	MS	F
Brand (A)	1387.50	3	462.50	0.676
Surface (B)	2888.08	2	1444.04	2.109
Interaction	8100.25	6	1350.04	1.972
Error	8216.00	12	684.67	—
Total	20591.83	23	—	—

(b)(c)(d) And here are the tests:

Hypotheses	Contents	Thresholds
H_A	No row effect.	$F_{0.05}(3,12) = 3.490$
H_B	No column effect.	$F_{0.05}(2,12) = 3.885$
H_{AB}	No interaction.	$F_{0.05}(6,12) = 2.996$

From above, we fail to reject H_A, H_B and H_{AB}.

That is, it is reasonable to believe there is **NO row effect**, **NO column effect** and **NO interaction**.

Problem 2.

(a) By LSE, we first figure out

$$S_{xy} = \sum y_i x_i - \frac{\sum y_i \sum x_i}{n} = 1083.67 - \frac{12.75 \times 1478}{20} = 141.445$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 143215.8 - \frac{(1478)^2}{20} = 33,991.6$$

Further the intercept and slope in the simple linear regression model are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{141.445}{33,991.6} = 4.161 \times 10^{-3}$$

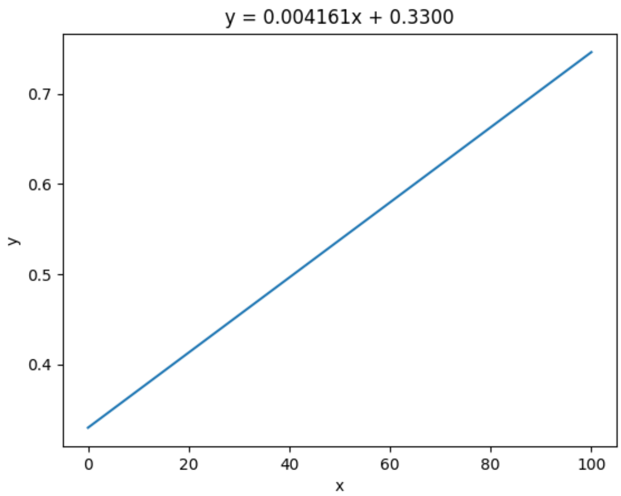
Since $\overline{y} = \frac{\sum y_i}{n} = \frac{12.75}{20} = 0.6375$, $\overline{x} = \frac{\sum x_i}{n} = \frac{1478}{20} = 73.9$,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 0.6375 - 4.161 \times 10^{-3} \cdot 73.9 = 0.3300$$

That is, the regression line is

$$\hat{y} = 4.161 \times 10^{-3} \cdot x + 0.3300$$

The graph is as follows:



To estimate the σ^2,

$$SS_T = \sum y_i^2 - n\overline{y}^2 = 8.86 - 20 \times 0.6375^2 = 0.7319$$

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} = 0.1433$$

We have estimated variance to be

$$\hat{\sigma^2} = \frac{SS_E}{n-2} = \frac{0.1433}{20-2} = 7.9611 \times 10^{-3}$$

(b) For the surface temperature as 85F,

$$\hat{y} = 4.161 \times 10^{-3} \cdot 85 + 0.3300 \approx 0.6837$$

The pavement deflection would be 0.6837 at the given temperature.

(c) For the temperature as 90F, we have the mean pavement deflection

$$\mu_{Y|x_0} = \beta_0 + \beta_1 x_0 = 4.161 \times 10^{-3} \cdot 90 + 0.3300 \approx 0.7045$$

The mean of the pavement deflection would be 0.7045 at the given temperature.

(d) For the regression line $\hat{y} = 4.161 \times 10^{-3} \cdot x + 0.3300$, 1F change in surface temperature will result in **a change of** 4.161×10^{-3} **in mean pavement deflection**.

PROBLEM 3.

(a) We use python codes to figure out means and variances.

With codes below:

Read the .csv file

```
In [1]: # Read the .csv file
def read_csv(filename):
    with open(filename, 'r') as f:
        data = f.read().split('\n')[1:]  # Read from the 2nd line
        data = [line.split(',') for line in data if line]
        data = [[float(num) for num in line] for line in data]
    return list(zip(*data))  # Separate x, y to access each
```

Compute mean and sample variance

```
In [2]: # Calculate sample mean
def mean(data):
    return sum(data) / len(data)


# Calculate sample variance
def variance(data):
    m = mean(data)
    return sum((xi - m) ** 2 for xi in data) / (len(data) - 1)
    # Remind that we use sample variance here
```

```
In [3]: D = read_csv('../D.csv')
D_x_mean = mean(D[0])
D_x_variance = variance(D[0])
D_y_mean = mean(D[1])
D_y_variance = variance(D[1])

S = read_csv('../S.csv')
S_x_mean = mean(S[0])
S_x_variance = variance(S[0])
S_y_mean = mean(S[1])
S_y_variance = variance(S[1])
```

```
In [4]: # Print out the results
print(f"D Dataset: x_mean = {D_x_mean:.2f}, x_variance = {D_x_variance:.2f}; \
y_mean = {D_y_mean:.2f}, y_variance = {D_y_variance:.2f} \n")

print(f"S Dataset: x_mean = {S_x_mean:.2f}, x_variance = {S_x_variance:.2f}; \
y_mean = {S_y_mean:.2f}, y_variance = {S_y_variance:.2f}")

D Dataset: x_mean = 54.26, x_variance = 281.07; y_mean = 47.83, y_variance = 725.52

S Dataset: x_mean = 54.27, x_variance = 281.20; y_mean = 47.84, y_variance = 725.24
```

We figure out the results:

	x^D	x^S	y^D	y^S
Sample mean	54.26	54.27	47.83	47.84
Sample variance	281.07	281.20	725.52	725.24

(b) Given the linear regression model $y = \beta(x - \bar{x}) + \alpha$, Compute regression parametres: $y = \beta(x - \bar{x}) + \alpha$

In [5]:
```
# Calculate regression parameters
def compute_regression_parameters(x_data, y_data):
    x_mean = mean(x_data)
    y_mean = mean(y_data)

    # Compute beta (slope)
    s_xy = sum((xi - x_mean) * (yi - y_mean) for xi, yi in zip(x_data, y_data))
    s_xx = sum((xi - x_mean) ** 2 for xi in x_data)
    beta = s_xy / s_xx

    # Compute alpha (intercept)
    alpha = y_mean

    return alpha, beta
```

In [6]:
```
D_alpha, D_beta = compute_regression_parameters(D[0], D[1])
S_alpha, S_beta = compute_regression_parameters(S[0], S[1])

# Print out the results
print(f"D Dataset: alpha = {D_alpha:.2f}, beta = {D_beta:.2f}")
print(f"S Dataset: alpha = {S_alpha:.2f}, beta = {S_beta:.2f}")
```

```
D Dataset: alpha = 47.83, beta = -0.10
S Dataset: alpha = 47.84, beta = -0.10
```

As a result, we have regression parametres $\alpha = 47.83$, $\beta = -0.10$; $a = 47.84$, $b = -0.10$

(c) Further we compute the confidence interval,

Construct Confidence Interval

In [9]:
```
from math import sqrt
# Compute stadard error of CI
def compute_standard_errors(x_data, y_data, alpha, beta):
    n = len(x_data)
    x_mean = sum(x_data) / n
    y_pred = [alpha + beta * (xi - x_mean) for xi in x_data]

    # Compute standard error for beta
    s_xx = sum((xi - x_mean) ** 2 for xi in x_data)
    var = sum((yi - y_hat) ** 2 / (n-2) for yi, y_hat in zip(y_data, y_pred))
    SE_beta = sqrt(var / s_xx)

    # Compute standard error for alpha
    SE_alpha = sqrt(var / n)

    return SE_alpha, SE_beta
```

In [10]:
```
D_SE_alpha, D_SE_beta = compute_standard_errors(D[0], D[1], D_alpha, D_beta)
S_SE_alpha, S_SE_beta = compute_standard_errors(S[0], S[1], S_alpha, S_beta)
```

In [11]:
```
# Compute the 95% confidence intervals
Z = 1.96  # Z-score for 95% confidence interval
D_alpha_CI = (D_alpha - Z * D_SE_alpha, D_alpha + Z * D_SE_alpha)
D_beta_CI = (D_beta - Z * D_SE_beta, D_beta + Z * D_SE_beta)
S_alpha_CI = (S_alpha - Z * S_SE_alpha, S_alpha + Z * S_SE_alpha)
S_beta_CI = (S_beta - Z * S_SE_beta, S_beta + Z * S_SE_beta)

# Print the results
print(f"D Dataset: alpha 95% CI = {D_alpha_CI}, \n beta 95% CI = {D_beta_CI} \n")
print(f"S Dataset: alpha 95% CI = {S_alpha_CI}, \n beta 95% CI = {S_beta_CI}")
```

```
D Dataset: alpha 95% CI = (43.39538547811511, 52.26912015568764),
 beta 95% CI = (-0.3691676732951159, 0.1620026684298038)

S Dataset: alpha 95% CI = (43.40309394284111, 52.27599650786307),
 beta 95% CI = (-0.3666128128939669, 0.1643868188121143)
```

As a result, we have the confidence interval of each parametre as:

Parametres	α	β	a	b
CI	(43.40, 52.27)	(−0.37, 0.16)	(43.40, 52.28)	(−0.37, 0.16)

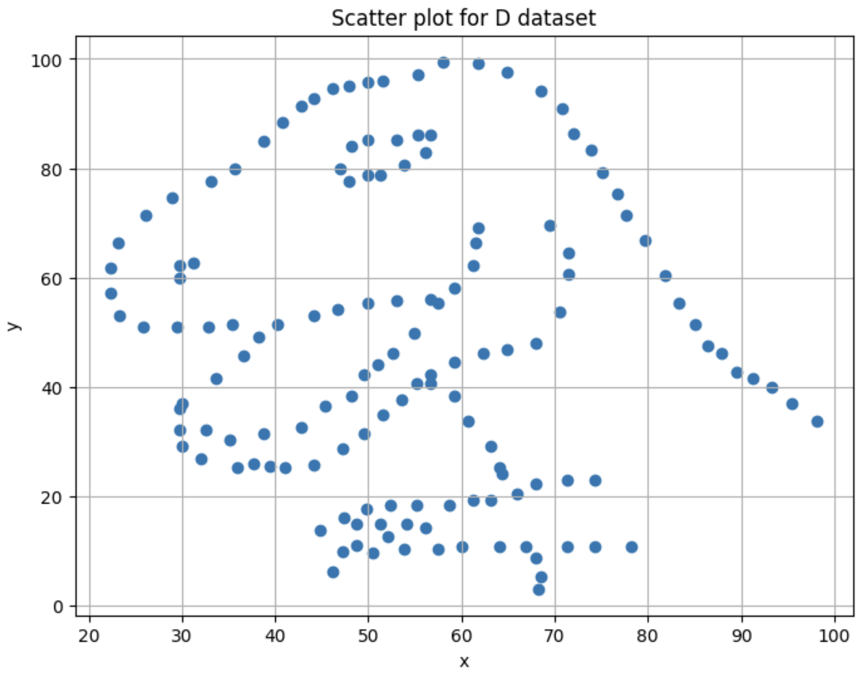
(d) Based on parametres of stastistics above, these two datasets are really similar.

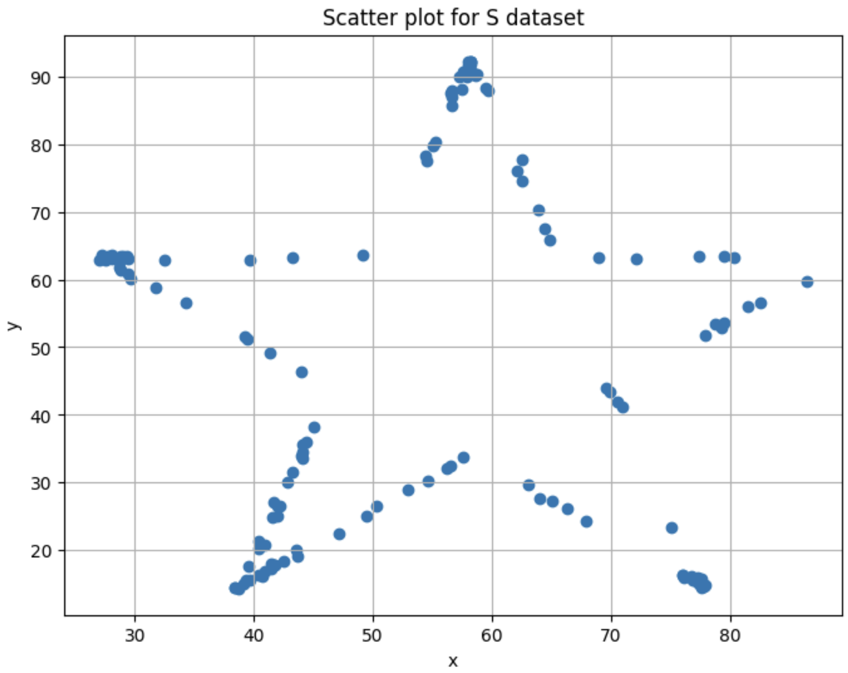
(e) To construct scatterplots for the datasets,

Construct scatterplots

In [7]:
```
# Construct scatter plots
import matplotlib.pyplot as plt
def plot_scatter(x_data, y_data, title):
    plt.figure(figsize=(8, 6))
    plt.scatter(x_data, y_data)
    plt.title(title)
    plt.xlabel('x')
    plt.ylabel('y')
    plt.grid(True)
    plt.show()
```

In [8]:
```
plot_scatter(D[0], D[1], 'Scatter plot for D dataset')
plot_scatter(S[0], S[1], 'Scatter plot for S dataset')
```





Haha, it is really amazing :)

(f) The two datasets are so alike in given parametres but quite different in scatter plots. That is, though specific parametres could reflect some characteristics of the given stastistic, it is still part of it instead of the whole.

PROBLEM 4.

By executing **r** codes as follows:

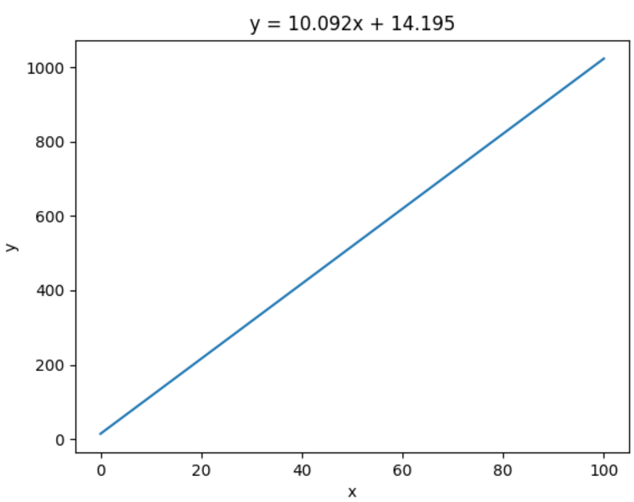
```r
setwd("/Users/kevinshuey/Github/cs_assignments/cuhksz_STA2002/hw5/4")
data <- read.csv("data113.csv", header = TRUE)

y <- data$Rating
x <- data$Yds

model <- lm(y ~ x)
summary(model)
```

We get comprehensive information like:

```
[Running] Rscript "/Users/kevinshuey/Github/cs_assignments/cuhksz_STA2002/hw5/4/4.r"

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-12.8533  -3.6074   0.4073   3.7063   8.9238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.195      9.059   1.567    0.128
x             10.092      1.288   7.836 9.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.219 on 30 degrees of freedom
Multiple R-squared:  0.6718,	Adjusted R-squared:  0.6609
F-statistic: 61.41 on 1 and 30 DF,  p-value: 9.589e-09


[Done] exited with code=0 in 0.144 seconds
```

Then we answer following questions:

(**I** - a) According to the summary above, the **slope** *i.e.* the coefficient of x, is 10.092; the **intercept** is 14.195; the **estimate of** $\boldsymbol{\sigma^2}$ is $5.219^2 = 27.238$. Here is the plot of $\mathbf{y = 10.092x + 14.195}$:



(**I** - b) For a quarterback averages 7.5 yards per attempt *i.e.* $x = 7.5$,

$$\mu_y = 10.092 \times 7.5 + 14.195 = 89.885$$

As a result, the estimate of mean rating is 89.885.

(**I** - c) With a decrease of 1 yard per attempt, it will result in **a decrease of 10.092 yards**.

(**I** - d) To reach an increase in mean rating by 10 points, it should **generate an increase in the average yards per attempt of** $\frac{10}{10.092} = 0.9909$.

(**II** - a) The **r** results above has stated:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{10.092 - 0}{\sqrt{27.238/16.4220}} = 7.836 > t_{0.01/2,\,32-2} = 2.750,$$

with p-value

$$\mathbf{p} = 9.59 \times 10^{-9} << 0.01$$

Thus we reject the hypothesis. Meanwhile, the p-value is so small that there must be significant influence of x to y in slope.

(**II** - b) As shown above by **r** results, the **standard error** for the slope is 1.288, and the one for the intercept is 9.059 .

(**II** - c) Suppose we test $H_0 : \beta_1 = 10$ versus $H_1 : \beta_1 \neq 10$,

First we calculate

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 16.4220$$

Then we have t-value

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{10.092 - 10}{\sqrt{27.238/16.4220}} = 0.0714 < t_{0.01/2,\,32-2} = 2.750,$$

thus we fail to reject the hypothesis.

As a result, it is reasonable to believe $\beta_1 = 10$ **.**

(**III** - a) We still have $n = 32$ and $\alpha = 0.05$.

For the slope β_1, we have

$$\hat{\beta}_1 - t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}(n-2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Specifically,

$$10.092 - 2.042\sqrt{\frac{27.238}{16.4220}} \leq \beta_1 \leq 10.092 + 2.042\sqrt{\frac{27.238}{16.4220}}$$

As a result, we have a confident interval for β_1 **of** $(\mathbf{7.462}, \mathbf{12.722})$**.**

(**III** - b) For the intercept β_0, we have

$$\hat{\beta}_0 - t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$$

Specifically,

$$10.092 - 2.042\sqrt{27.238 \times (\frac{1}{32} + \frac{6.9978^2}{16.4220})} \leq \beta_0 \leq 10.092 + 2.042\sqrt{27.238 \times (\frac{1}{32} + \frac{6.9978^2}{16.4220})}$$

As a result, we have a confident interval for β_0 of $(-8.407, 28.591)$.

(III - c) The mean rating at 8.0 is

$$\mu_{Y|x_0=8.0} = 10.092 \times 8.0 + 14.195 = 94.931$$

As a result, the specific mean rating is 94.931 .

(III - d) For the rating, we have

$$\hat{\mu}_{Y|x_0=8.0} - t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})} \leq \mu_{Y|x_0=8.0} \leq \hat{\mu}_{Y|x_0=8.0} + t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}$$

Specifically,

$$94.931 - 2.042\sqrt{27.238 \times (\frac{1}{32} + \frac{1.00438^2}{16.4220})} \leq \mu_{Y|x_0=8.0} \leq 94.931 + 2.042\sqrt{27.238 \times (\frac{1}{32} + \frac{1.00438^2}{16.4220})}$$

As a result, we have a confident interval for $\mu_{Y|x_0=8.0}$ of $(91.687, 98.175)$.

Problem 5.

By executing **r** codes as follows:

```r
x1 <- c(25, 31, 45, 60, 65, 72, 80, 84, 75, 60, 50, 38)
x2 <- c(24, 21, 24, 25, 25, 26, 25, 25, 24, 25, 25, 23)
x3 <- c(91, 90, 88, 87, 91, 94, 87, 86, 88, 91, 90, 89)
x4 <- c(100, 95, 110, 88, 94, 99, 97, 96, 110, 105, 100, 98)

y <- c(240, 236, 270, 274, 301, 316, 300, 296, 267, 276, 288, 261)

model <- lm(y ~ x1 + x2 + x3 + x4)
summary(model)
```

We get comprehensive information like:

```
[Running] Rscript "/Users/kevinshuey/Github/cs_assignments/cuhksz_STA2002/hw5/5.r"

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.098  -9.778   1.767   6.798  13.016 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -123.1312   157.2561  -0.783    0.459  
x1             0.7573     0.2791   2.713    0.030 *
x2             7.5188     4.0101   1.875    0.103  
x3             2.4831     1.8094   1.372    0.212  
x4            -0.4811     0.5552  -0.867    0.415  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.79 on 7 degrees of freedom
Multiple R-squared:  0.852, Adjusted R-squared:  0.7675 
F-statistic: 10.08 on 4 and 7 DF,  p-value: 0.00496


[Done] exited with code=0 in 0.153 seconds
```

Then we answer following questions:

(a) We construct a model like:

$$y = 0.7573 \cdot x_1 + 7.5188 \cdot x_2 + 2.4831 \cdot x_3 - 0.4811 \cdot x_4 - 123.1312$$

(b) Based on **r** results,

$$\hat{\sigma^2} = 11.79^2 = 139.0041$$

(c) The **standard errors** for the coefficient of x_1, x_2, x_3, x_4 and the intercept are 0.2791, 4.0101, 1.8094, 0.5552 and 157.2561 respectively. They are NOT estimated with the same precision, as **they have different standard errors**.

(d) For the given value,

$$y = 0.7573 \times 75 + 7.5188 \times 24 + 2.4831 \times 0.9 - 0.4811 \times 98 - 123.1312 = 69.2045$$

As a result, we predict the power consumption to be 69.2045.

End of Homework 5