Jiankun_Dong_CS555_HW3

Jiankun (Bob) Dong CM3226

2023-10-21

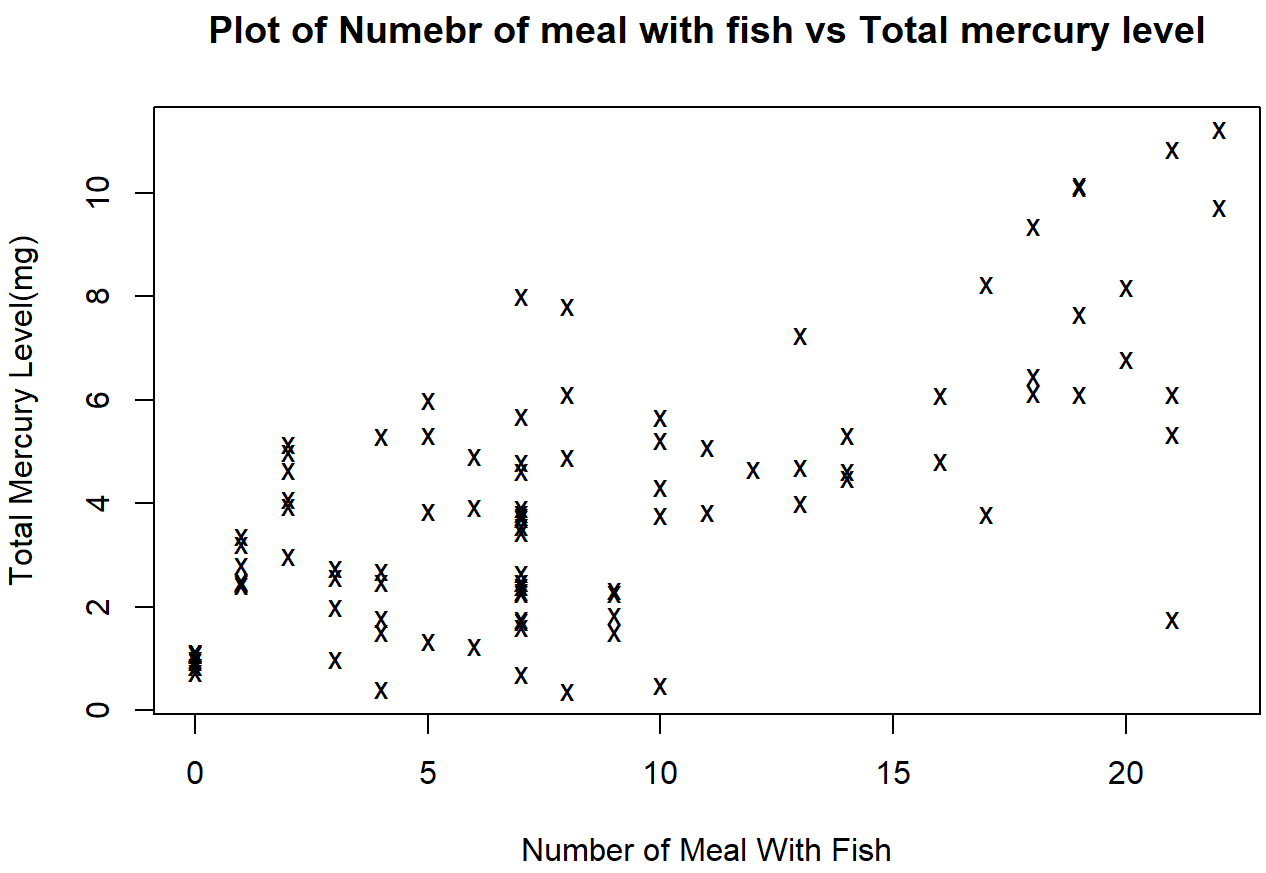
```
rm(list = ls())
library("ggplot2")
setwd("C:/BU/CSSE/CS555/HW3")
a03 <- read.csv("./A03.csv")
summary(a03)
```

```
##  Number.of.meals.with.fish Total.Mercury.in.mg.g
##  Min.   : 0.00             Min.   : 0.366
##  1st Qu.: 3.00             1st Qu.: 2.200
##  Median : 7.00             Median : 3.781
##  Mean   : 8.30             Mean   : 3.978
##  3rd Qu.:12.25             3rd Qu.: 5.294
##  Max.   :22.00             Max.   :11.222
```

```
attach(a03)
```

Problem 1:
Number of meals with fish should be on the x-axis and total mercury level should be on the y-axis. Because eating meals with fish happens before potential change in mercury level in body.

```
plot(Number.of.meals.with.fish, Total.Mercury.in.mg.g, xlab = "Number of Meal With Fish",
     ylab = "Total Mercury Level(mg)", title("Plot of Numebr of meal with fish vs Total mercury
level"), pch = 'x')
```



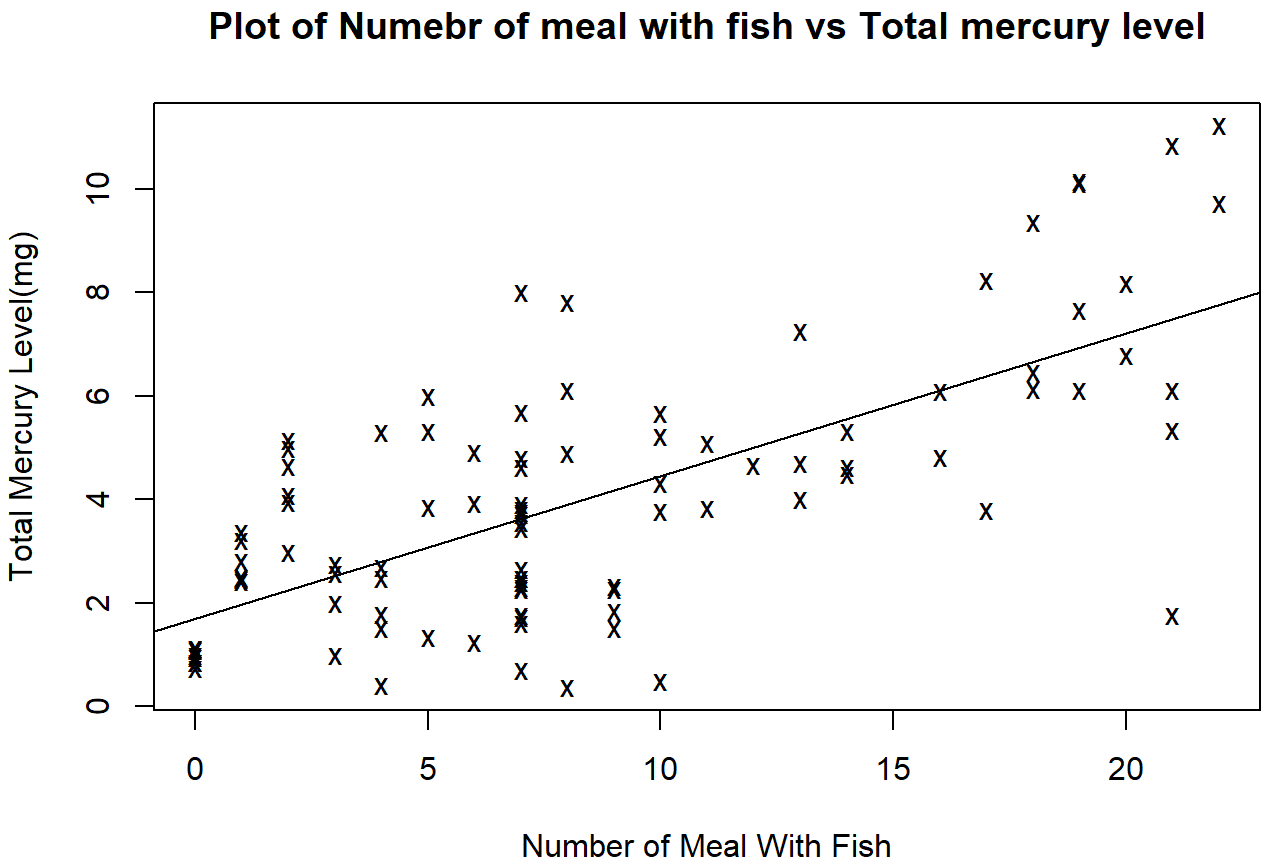

```
meal_bar <- mean(Number.of.meals.with.fish)
meal_sd <- sd(Number.of.meals.with.fish)
merc_bar <- mean(Total.Mercury.in.mg.g)
merc_sd <-  sd(Total.Mercury.in.mg.g)
#First we need to check for outliers before using cor function for correlation
IQR_meal <- quantile(Number.of.meals.with.fish,.75)-quantile(Number.of.meals.with.fish,.25)
meal_outlier <- Number.of.meals.with.fish < quantile(Number.of.meals.with.fish,.25)-1.5*IQR_meal
| Number.of.meals.with.fish > quantile(Number.of.meals.with.fish,.75)+1.5*IQR_meal
IQR_merc <- quantile(Total.Mercury.in.mg.g,.75)-quantile(Total.Mercury.in.mg.g,.25)
merc_outlier <- Total.Mercury.in.mg.g < quantile(Total.Mercury.in.mg.g,.25)-1.5*IQR_merc | Tota
l.Mercury.in.mg.g > quantile(Total.Mercury.in.mg.g,.75)+1.5*IQR_merc
```

First, we check for outliers in the data-set:
There are no outliners for number of meals, and there are 4 outliers for Total mercury level, and they are the following:10.849, 11.222, 10.116, 10.14.
However, base on the graph, and the fact that they are reasonably close to the upper limit for outliers (9.934375), it's reasonable to include them in the data set for the following calculations.
Base on the scatter plot, it's roughly a linear, positive relationship with moderate correlation between the number of meal with fish, and the total mercury level.
Problem 2:

```
corXY_noOut <- round(cor(Number.of.meals.with.fish[!merc_outlier],Total.Mercury.in.mg.g[!merc_ou
tlier]),4)
corXY <- round(cor(Number.of.meals.with.fish,Total.Mercury.in.mg.g),4)
```

The correlation is 0.6991. Just to validate my judgement about outliers, the correlation without the outliers is 0.6326. Both corelation values tell us that the relationship is positive with moderate strength.
Problem 3:

```
lmFIT <- lm(Total.Mercury.in.mg.g ~ Number.of.meals.with.fish)
beta_1 <- lmFIT$coefficients[[2]]
beta_0 <- lmFIT$coefficients[[1]]
plot(Number.of.meals.with.fish, Total.Mercury.in.mg.g, xlab = "Number of Meal With Fish",
     ylab = "Total Mercury Level(mg)", title("Plot of Numebr of meal with fish vs Total mercury
level"), pch = 'x')
abline(beta_0,beta_1)
```



The least squares regression equation is $\hat{y} = \hat{\beta}_0 * x + \hat{\beta}_1$, which for this data set is: y=0.2759503*x+1.6876426
Problem 4:

```
#can use the beta_0 and beta_1 value from the previous calculation as well.
beta_1_est <- round(corXY*merc_sd/meal_sd, 5)
beta_0_est <- round(merc_bar-beta_1*meal_bar,5)
```

The estimated $\hat{\beta}_0$ is 1.68764. It means that when there's no meal eaten with fish, the mercury level in the body is 1.68764mg.

The estimated $\hat{\beta}_1$ is 0.27595. It meas that with our estimation based on the data set, the mercury level in the body increase linearly by 0.27595 mg for every extra meal with fish eaten.

Problem 5:

Step 1: $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$ $\alpha = 0.05$ Step 2:

$F = \frac{RegMS}{ResMS}$ For SLR, $F = ResDF * \frac{RegSS}{ResSS}$

```
anova(lmFIT)
```

```
## Analysis of Variance Table
##
## Response: Total.Mercury.in.mg.g
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## Number.of.meals.with.fish  1 309.24 309.239  93.689 6.013e-16 ***
## Residuals                 98 323.47   3.301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fcrit <- qf(.95,df1=1,df2=length(Total.Mercury.in.mg.g)-2)
```

Step 3:

Reject H_0 if $F \geq F(1, n-2, \alpha)$

$F(1, n-2, \alpha)$ = 3.9381111

Step 4: From the anova table, we have F = 93.689, greater than the fcrit Step 5:

Reject H_0, at 95% confidence level, we have enough evidence that a significant linear relationship exist between the number of meals with fish and the total mercury level.

```
Rsqr <- 309.24/(323.47+309.24)
```

The R square value is 0.4888, which means that 48.8755% of the variance in the total mercury level can be explained by the variance in the number of meals with fish.

```
confLM <- confint(lmFIT,level = .9)
lowerCONF <- confLM[1]
upperCONF <- confLM[3]
```

The 90% confidence interval for β_1 is between 1.1922529 and 2.1830324.