MET CS 555 Assignment 3 – 20 points Fall 2, 2019

**SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission.  Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R. You will also lose 10 points per day for late submissions unless prior arrangements are made with your facilitator.**

**The data in this document gives the number of meals eaten that contain fish (per week) and mercury levels in head hair for 100 fisherman.** **Save the data to a format that can be read into R. Read the data in for analysis. Use R to calculate the quantities and generate the visual summaries requested below.**

(1) Save the data and read into R for analysis.

**My answer:**

See my R script below.

(2) To get a sense of the data, generate a scatterplot (using an appropriate window, label the axes, and title the graph). Consciously decide which variable should be on the -axis and which should be on the y-axis. Using the scatterplot, describe the form, direction, and strength of the association between the variables.

**My answer:**

Scatterplot of Mercury Content versus Number of Meals with Fish:

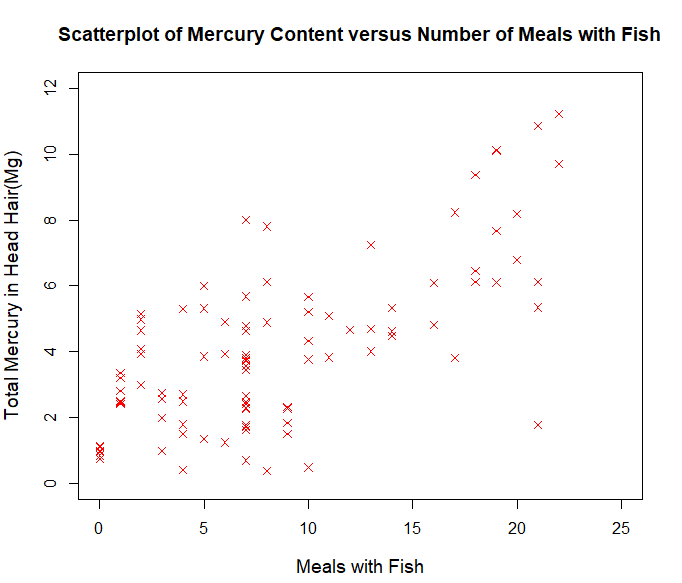


Fig.1 The Scatterplot

In the above scatterplot, I chose to place the meals with fish on the x-axis and the total mercury content on the y-axis, since in this case, the relationship between the two factors is apparent that the total mercury content is the outcome and response for those fishermen to have meals with fish, that is, eating fish in their meals occurs before the mercury accumulates in fishermen’s hair. Thus, we put the mercury level(response variable) on the y-axis, and put the number of meals with fish (explanatory variable) on the x-axis.

As the scatterplot shows, its form is linear as the points tends toward a straight line, and its direction can be described as “positively associated” since as the number of meals with fish increases, the mercury level also tends to increase in value, which looks like points hovering around a line with a positive slope. As regards of strength of the association, though the plot shows factors that are positively associated, the association between meals with fish and the mercury content is not so strongly associated when looking at the total mercury level of fishermen’s hair with the same meals they had. For example, the range of value of mercury is quite variable and fluctuating, especially when the value of meals approximately at x = 4,5,6,7 and 21.

(3) Calculate the correlation coefficient. What does the correlation tell us?

**My answer:**

In R script:

cor(mercurydata$fish, mercurydata$mercury)

Result:

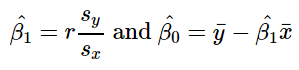
The correlation coefficient is 0.6991094 ≈ 0.6991. The correlation tells us that the correlation coefficient between number of meals with fish and the total mercury level is 0.6991, since it’s greater than 0, the variables can be described as positively correlated, and it measures the strength of a linear relationship only.

(4) Find the equation of the least squares regression equation, and write out the equation. Add the regression line to the scatterplot you generated above.

**My answer:**

The equation of the least-squares regression:

The estimates of and given by the following equations:



*Sy* = the sample standard deviation of mercury level, *Sx* = the sample standard deviation of the number of meals with fish. Since the correlation coefficient *r* = 0.6991, Sy = 2.528044, Sx = = 6.404702, the average of value of mercury level = 3.978, the average of value of meals with fish = 8.3, thus = r \* Sy / Sx = 0.276, = 3.978 - 0.276 \* 8.3 = 1.688.

Another method is using R,

lm(mercurydata$mercury~mercurydata$fish)

Result:

Coefficients:

(Intercept) mercurydata$fish

1.688 0.276

Thus, the final equation is

To add the regression line in the scatterplot,

abline(flm,lty=5,col="black")

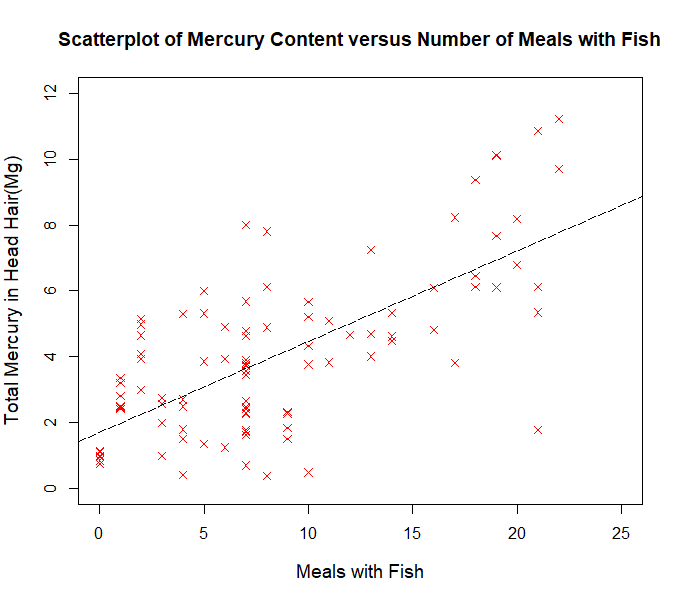


Fig.2 The Scatterplot with Regression Line

(5) What is the estimate for ? How can we interpret this value? What is the estimate for ? What is the interpretation of this value?

**My answer:**

According to the result of above question, we have the equation , so the estimate for = 0.276, which can be interpreted as the slope parameter that as the increase in the mercury level for every one-meal with fish increase. That is, for each addition meal with fish that fishermen had, their mercury level increased by 0.276mg on average.

The estimate for the intercept = 1.688, which can be interpreted as the average mercury level in hair is 1.688mg for those who have meals without fish. It’s an expected or predicted value.

(6) Calculate the ANOVA table and the table which gives the standard error of . Formally test the hypothesis that = 0 using either the -test or the -test at the = 0.10 level. Either way, present your results using the 5 step procedure as in the course notes. Within your conclusion, calculate the -squared value and interpret this. Also, calculate and interpret the 90% confidence interval for .

**My answer:**

Create the AVONA table in R script:

flm <- lm(mercurydata$mercury~mercurydata$fish)

anova(flm)

Here I use the *F*-test.

1. Set up the hypotheses and select the alpha level

*H*0: = 0 (there is no linear association)

*H*1: ≠ 0 (there is a linear association)

*α =* 0.10

1. Select the appropriate test statistic

with 1 and n-2 degrees of freedom

1. Set the decision rule

Determine the appropriate value from F-distribution table with 1, n-2 = 100 - 2 =98 degrees of freedom and associated with a right-hand tail probability of *α =* 0.10.

Using the R, F1,29,0.05 = qf(0.90, 1, 98) = 2.7574

Decision Rule: Reject *H*0 if F ≥ 2.7574

Otherwise, do not reject *H*0

1. Compute the test statistic

anova(flm) or summary(flm) gives:

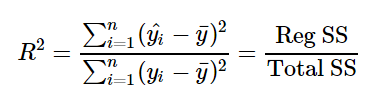
F-statistic: 93.69, p-value: 6.013e-16

1. Conclusion

Reject *H*0since 93.69 ≥ 2.7574 (also p ≤ *α*, p-value = 6.013e-16 < 0.10). We have significant evidence at the *α =* 0.10 level that ≠ 0. That is, there is evidence of a significant linear association between the number of meals with fish the fishermen had and the total mercury content in their hair.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | df | sum square | mean square | F-statistic | p-value |
| Regression | 1 | 309.24 | 309.239 | 93.689 | 6.013e-16 |
| Residual | 98 | 323.47 | 3.301 |  |  |
| Total | Total SS = Reg SS + Res SS = 632.71 | | | | |

Use the ANOVA table to calculate



= 309.24 / 632.71= 48.9%

That is, 48.9% of the variability in the value of mercury level can be explained by the number of meals with fish the fishermen had.

Use R to calculate the 90% confidence interval for

confint(flm, level = 0.90)

Result:

5 % 95 %

(Intercept) 1.192253 2.1830324

mercurydata$fish 0.228609 0.3232916

Therefore, the 90% confidence interval for is between 0.23 to 0.32. We are 90% confident that the true value of is between 0.23 to 0.32.

R script:

setwd("C:\\Users\\Lin\\Desktop\\2019 Fall\\555 DAV\_ R\\Homework\\hw3")

# 1. read the fisherman data

mercurydata = read.csv("fisherman\_mercury.csv")

mercurydata

# 2. draw scatterplot, x-axis for number of fish and y-axis for mercury content

plot(mercurydata$fish, mercurydata$mercury,

main = "Scatterplot of Mercury Content versus Number of Meals with Fish",

xlab = "Meals with Fish", ylab = "Total Mercury in Head Hair",

xlim = c(0, 25), ylim = c(0,12), pch = 4, col="red",

cex=1, cex.lab = 1.2, cex.main = 1.2)

# Calculate Sample Correlation

cor(mercurydata$fish, mercurydata$mercury)

cor.test(mercurydata$fish, mercurydata$mercury)

# calculate the fitting linear model

flm <- lm(mercurydata$mercury~mercurydata$fish)

# get Beta1 and Beta0(interception)

lm(mercurydata$mercury~mercurydata$fish)

# adding the regression line to scatterplot

plot(mercurydata$fish, mercurydata$mercury,

main = "Scatterplot of Mercury Content versus Number of Meals with Fish",

xlab = "Meals with Fish", ylab = "Total Mercury in Head Hair(Mg)",

xlim = c(0, 25), ylim = c(0,12), pch = 4, col="red",

cex=1, cex.lab = 1.2, cex.main = 1.2)

abline(flm,lty=5,col="black")

# 6. calculate the ANOVA

anova(flm)

summary(flm)

# calculate the F-test by qf()

qf(0.90, 1, 98)

# calculate conf=90% for beta1

confint(flm, level = 0.90)