**ASSIGNMENT 3  
BU ID: U55-32-1699**

**1. The directory has been set and the data has been read into the R studio and has been viewed with the basic describe statistics which is the first main step to analyse any particular data set.**  
  
fishm<-read.csv("fishmeals.csv")

View(fishm)

> library(psych)

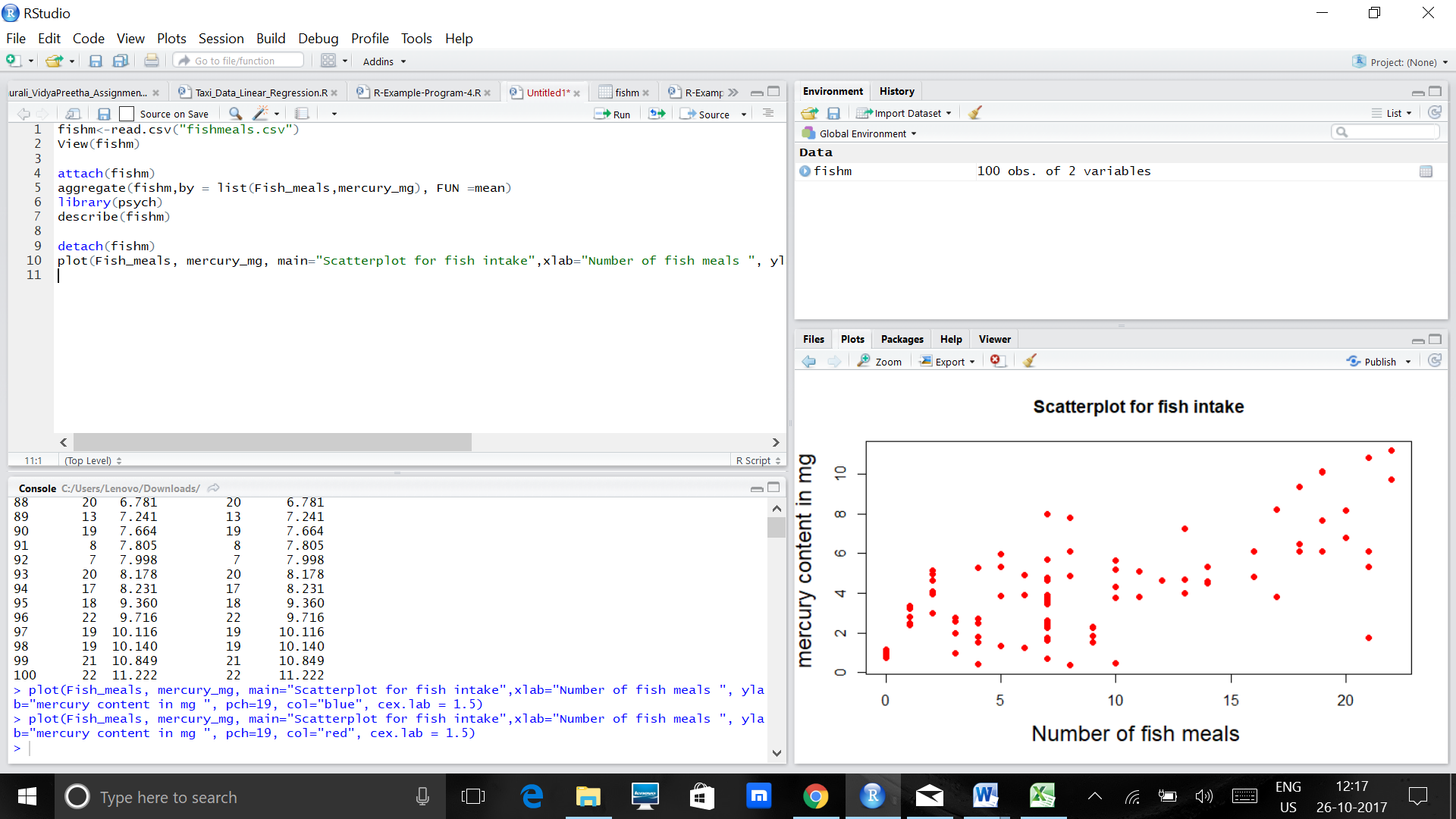
> describe(fishm)

vars n mean sd median trimmed mad min max range skew kurtosis se

Fish\_meals 1 100 8.30 6.40 7.00 7.79 5.93 0.00 22.00 22.00 0.64 -0.67 0.64

mercury\_mg 2 100 3.98 2.53 3.78 3.71 2.25 0.37 11.22 10.86 0.83 0.25 0.25

**2. Generating the scatterplot:**The x axis consists of the element that is independent which will represent the number of fish meals taken and the y axis consists of the variable that is dependent on the number of meals which is the mercury content in the food and hence we can conclude that the x axis – number of fish meals, y axis –



Form: The graph can be perceived to be linear to some extent with the initial points giving in some discrepancies to the hypothesis. The general trend can be noted that more number of fish meals indulges into more amounts of mercury content to majority of the cases.  
  
Direction: The graph proceeds with an inclination upwards and hence can be known termed as “positively associated”. Here, the number of fish meals determines the mercury content and more numbers to some extent only indicate that the mercury content increases as the number of meals increases.

Strength of the Relationship:   
Overall the correlation between the factors indicate that they are highly correlated for meals that have a count less than 10, but for those beyond that number also have a good correlation although not as great as the ones below them.

**3. Correlation:**  
The correlation value for the dataset given is 0.6991094, this value is positive, which indicates that there is a positive correlation in the dataset. Here, the order in which we type out the columns of the rows don’t matter as the correlation between two factors is the same however the order is.  
  
cor(Fish\_meals, mercury\_mg)

**4. Least square regression equation:**  
y^ = β0^+β1^x – Equation used

> my.model<-lm(mercury\_mg~ Fish\_meals)

> print(my.model)

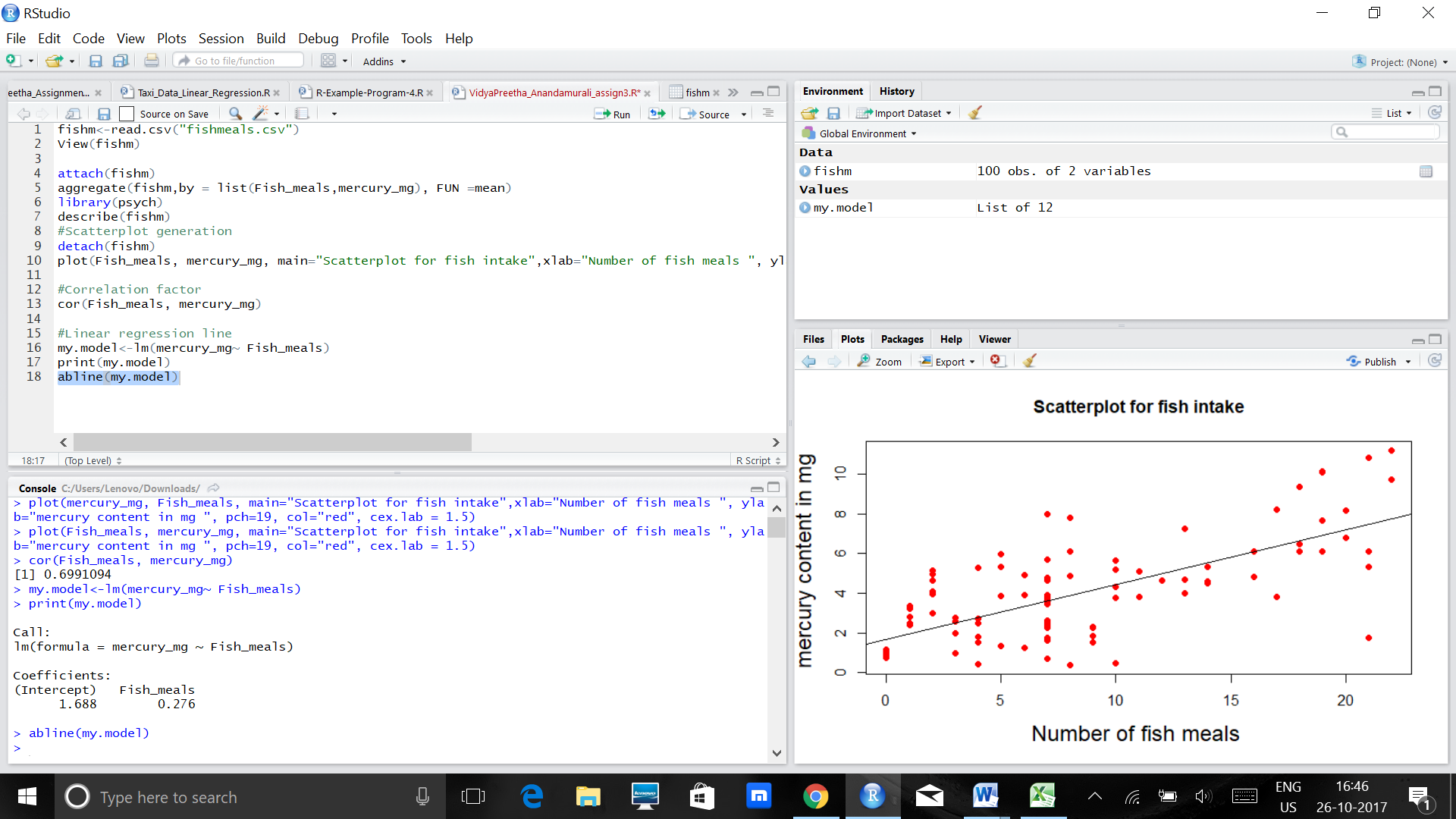
Call:

lm(formula = mercury\_mg ~ Fish\_meals)

Coefficients:

(Intercept) Fish\_meals

1.688 0.276

The regression line included in the model:  


**5. The values for  and can be calculated using R as well as the formulae given:**  
 = r\*Sy/Sx = 0.6991094 \* 2.53/6.40 = 0.2763  
= y¯−β1^x = 3.98 –( 0.2763 \* 8.30) = 1.688

And using the R code:  
is the intercept which is equal to 1.688 from the regression model. This value is also greater than 0, which can be interpreted as the average exam grade for those who did not eat a meal (0 meals intaken).   
 is the slope which can also be derived from the regression model as 0.276( as mentioned in the code above). Interpretation : This value is greater than 0, therefore, it indicates that there is a linear positive association with the dataset given. That is, for each additional fish meal that anyone intook, their mercury content level increased by around 0.2 percentage points on an average.

**6. ANOVA Table:**This table can be made with the r code itself.  
  
> anova(my.model)

Analysis of Variance Table

Response: mercury\_mg

Df Sum Sq Mean Sq F value Pr(>F)

Fish\_meals 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

The table that gives the standard error of :

The five steps for F test to find the deviation of the statistics of the data given:  
a) Set up the hypotheses and select the alpha level:  
 - H0:β1=0H0:β1=0 (there is no linear association)  
 - H1:β1≠0H1:β1≠0 (there is a linear association)  
 - α=0.10

b) Select the appropriate test statistic  
 - F = (MS Reg/MS Res) with 1 and n−2 degrees of freedom. And n = 100, therefore n-2 = 100-2 = 98.  
  
c) Stating the decision rule:  
F distribution asscociated with 1 df and 98 df with a confidence interval of 90%-  
  
> qf(0.90, df1=1, df2=98)

[1] 2.75743

Therefore, this value could be guessed from the table itself that it would fall within the range of 2.76 and 2.73.

We should reject Ho if F >= 2.75743

d) From the anova table calculated above, we can infer that the F statistic value is 93.689 which is greater than 2.75743 with a two degrees of freedom, 1 and 98 and with alpha level of 0.10.

e) Conclusion: Therefore, we can reject the hypothesis that there is no linear association and we can conclude that there is a linear association with the data given. R squared value :   
Total SS ( From the anova command): 632.71  
Reg SS : 309.24/ 632.71 = 48.87% of the variability in the mercury content of the fish can be explained by the number of meals that are present.

Confidence interval of 90% to find beta1 value :  
> confint(my.model, level = 0.90)

5 % 95 %

(Intercept) 1.192253 2.1830324

Fish\_meals 0.228609 0.3232916

From the above information we can conclude that beta1 value is close to 0.228 for the confidence interval of 90% - here, the value of beta 1 can be interpreted as follows: This value is greater than 0, therefore, it indicates that there is a linear positive association. That is, for each additional fish meal that anyone in took, their mercury content level increased by around 0.2 percentage points on an average.

**R CODE:**

fishm<-read.csv("fishmeals.csv")  
View(fishm)  
attach(fishm)  
aggregate(fishm,by = list(Fish\_meals,mercury\_mg), FUN =mean)  
library(psych)  
describe(fishm)  
#Scatterplot generation  
detach(fishm)  
plot(Fish\_meals, mercury\_mg, main="Scatterplot for fish intake",xlab="Number of fish meals ", ylab="mercury content in mg ", pch=19, col="red", cex.lab = 1.5)

#Correlation factor  
cor(Fish\_meals, mercury\_mg)

#Linear regression line  
my.model<-lm(mercury\_mg ~ Fish\_meals)  
print(my.model)  
abline(my.model)  
confint(my.model)  
anova(my.model)  
qf(0.90, df1=1, df2=98)  
confint(my.model, level = 0.90)