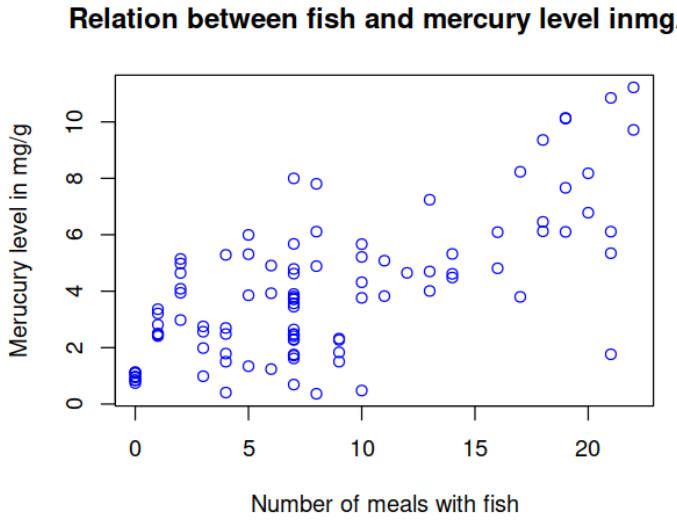
MET CS 555 Assignment 3

```
> getwd()
[1] "/home/xuyuhan/Desktop/BU-learn-S2023/METCS555A3/HW3"
> setwd('/home/xuyuhan/Desktop/BU-learn-S2023/METCS555A3/HW3')
> mercury = read.csv('cs555_hw3.csv', header = TRUE)
>
```

(1) 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 x-axis and which should be on the y-axis. Using the scatterplot, describe the form, direction, and strength of the association between the variables. (4 points)

```
> #homework main
> #1
> plot(mercury$Number.of.meals.with.fish, mercury$Total.Mercury.in.mg.g,xlab= 'Number of meals with fish', y
lab= 'Merucury level in mg/g', col= 'blue', main='Relation between fish and mercury level inmg/g')
>
```

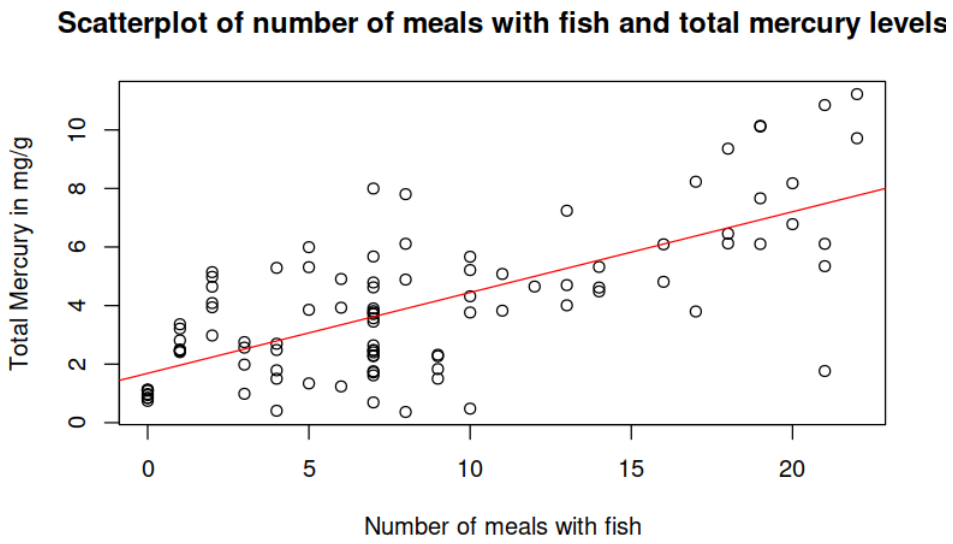


(2) Calculate the correlation coefficient. What does the correlation tell us? (2 points)

```
> #2
> cor(mercury$Number.of.meals.with.fish, mercury$Total.Mercury.in.mg.g)
[1] 0.6991094
> 
```

(3) Find the equation of the least squares regression equation and write out the equation. Add the regression line to the scatterplot you generated above. (2 points)

```
> model <- lm(Total.Mercury.in.mg.g ~ Number.of.meals.with.fish, data = mercury)
> a <- coef(model)[1] # intercept
> b <- coef(model)[2] # slope
> cat("Regression equation: Total Mercury = ", round(a, 3), "+", round(b, 3), "x Number of meals with fish")
Regression equation: Total Mercury =  1.688 + 0.276 x Number of meals with fish
> plot(mercury$Number.of.meals.with.fish, mercury$Total.Mercury.in.mg.g,
+      xlab = "Number of meals with fish", ylab = "Total Mercury in mg/g",
+      main = "Scatterplot of number of meals with fish and total mercury levels")
> abline(model, col = "red")
> 
```



(4) What is the estimate for $\beta 1$? How can we interpret this value? What is the estimate for $\beta 0$? What is the interpretation of this value? For the interpretations, you should be interpreting them in the context of this specific data set. (4 points)

```
> #4
> beta1_hat <- coef(model)[2]
> cat("The estimate for β1 is", beta1_hat, ". This means that for every additional meal with fish, the aver
age total mercury in head hair is estimated to increase by", beta1_hat, "mg/g.\n")
The estimate for β1 is 0.2759503 . This means that for every additional meal with fish, the average total m
ercury in head hair is estimated to increase by 0.2759503 mg/g.
> beta0_hat <- coef(model)[1]
> cat("The estimate for β0 is", beta0_hat, ". This means that when the number of meals with fish is 0, the
average total mercury in head hair is estimated to be", beta0_hat, "mg/g.\n")
The estimate for β0 is 1.687643 . This means that when the number of meals with fish is 0, the average tota
l mercury in head hair is estimated to be 1.687643 mg/g.
>
```

(5) Calculate the ANOVA table AND the table which gives the standard error of $\beta 1$. Formally test the hypothesis that $\beta 1 = 0$ using either the F-test or the t-test at the $\alpha = 0.05$ level. Either way, present your results using the 5-step procedure, as described in the course notes. Within your conclusion, calculate the R-squared value and interpret this. Also, calculate (using R) and interpret the 90% confidence interval for $\beta 1$. (8 points)

```
> #5
> anova(model)
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
> summary(model)$coefficients[2, 2]
[1] 0.02850937
>
> # Hypothesis test
> # Step 1: State the null and alternative hypotheses
> # H0: β1 = 0
> # Ha: β1 ≠ 0
>
> # Step 2: Set the significance level
> alpha <- 0.05
>
> # Step 3: Calculate the test statistic
> t_stat <- summary(model)$coefficients[2, 3]
>
>
> # Step 4: Calculate the p-value
> p_value <- 2 * pt(abs(t_stat), df = nrow(mercury) - 2, lower.tail = FALSE)
>
> # Step 5: Make a decision and interpret the results
> if (p_value < alpha) {
+   cat("Reject H0: β1 ≠ 0\n")
+ } else {
+   cat("Fail to reject H0: β1 = 0\n")
+ }
Reject H0: β1 ≠ 0
>
>
> summary(model)$r.squared
[1] 0.488754
> confint(model, level = 0.90)
                                5 %      95 %
(Intercept)               1.192253 2.1830324
Number.of.meals.with.fish 0.228609 0.3232916
```

EXTRA CREDIT

(a) Based on the scatterplot below, is a linear model appropriate to describe the relationship between seed count and seed weight? Explain.

NO, In the graph, we can know that if the seed weight is 0.6, the seed count is 27239, and when the seed weight is 1930, the seed count is 525. They are the inverse ratio relation. The Inverse ratio relation is not the linear model.

(b)Which model, A or B, is more appropriate for predicting seed weight from seed count? Justify your answer.

For my opinion, I think model A is more appropriate. According to the answer in question a, we can know that the relationship between seed weight and seed count is inverse ratio relation, and in the first graph of model A, it is the graph of inverse ratio relationship. Inverse ratio relation is much easier to predict.

(c) Using the model you chose in part (b), predict the seed weight if the seed count is 3700.

According to the model I chose in part b, model A, I predict the seed weight is 373 if the seed count = 3700

(d) Interpret the R-squared value of for your model

$r^2 = 0.7016$

According to the provided function equation and numerical values, the meaning of $r^2 = 0.7016$ is that the logarithmic relationship between seed count and seed weight can explain 70.16% of the variability in the data. In other words, the logarithmic function fits the data reasonably well for 70.16%, but there is still 29.84% of variability that cannot be explained by this function.

r^2 is a statistical measure that represents the goodness of fit of the data. Its value ranges from 0 to 1. The closer r^2 is to 1, the better the fit, and the closer it is to 0, the worse the fit. In this case, $r^2 = 0.7016$ indicates that the logarithmic function can reasonably fit the relationship between seed count and seed weight.