MET CS555: Homework 3

Patrick Ryan



**(1)** Save the data to a file (excel or CSV file) and read it into R memory for analysis. (Q1 - 2 points):

|  |
| --- |
| # Set the working directory to the location of the .csv file.  setwd("C:/Users/patry/OneDrive/Desktop")  # Assign the .csv information to a dataframe  df = read.csv("homework3.csv", header=T, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df |

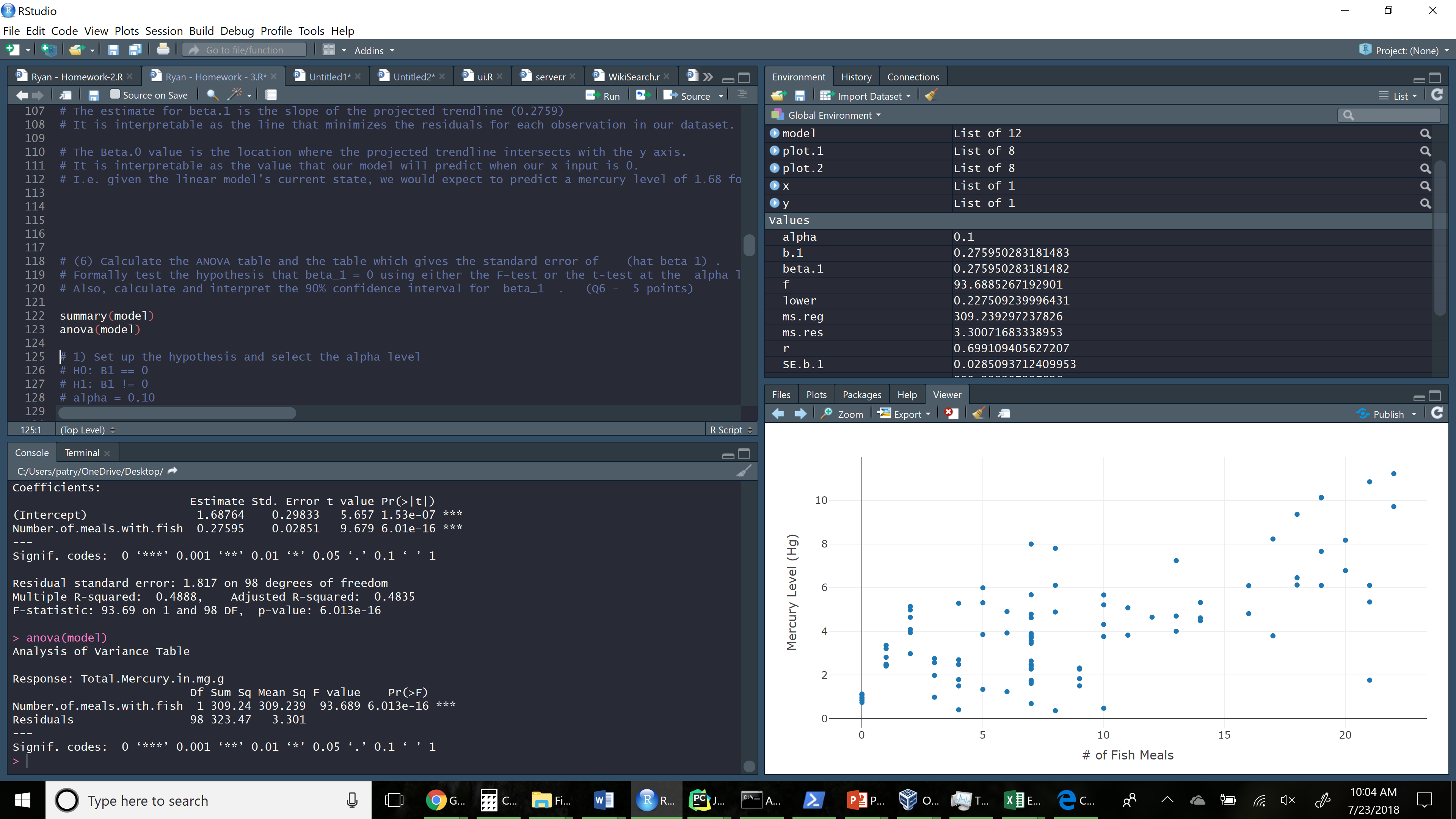
**(2)** To get a sense of the data, generate a scatterplot (using an appropriate window, label the axes, and title the graph). Consciously decide which variable should be on the -axis and which should be on the y-axis. Using the scatterplot, describe the form, direction, and strength of the association between the variables. (Q2 - 3 points)

|  |
| --- |
| library(plotly)  x <- list(  title = "# of Fish Meals")  y <- list(  title = "Total Me (Hg)")  plot.1 <- df %>%  plot\_ly(x = df$Number.of.meals.with.fish, trace ="scatter", mode="markers") %>%  add\_markers(y = df$Total.Mercury.in.mg.g, name="Observation") %>%  layout(xaxis = x, yaxis = y)  plot.1 |

Observations:

1. Form: The data forms a general random pattern (not particularly linear or grouped) (sd.x = 6.40, sd.y=2.52)
2. Direction: The data appears to feature a moderately positive correlation.
3. Strength: The data features a low strength and association between factors.
4. On several instances, observations with the same x identity feature high variation within their supposed causal variable. (e.g., we have multiple observations of individuals who have consumed 7 fish meals, however, the corresponding y-axis for these values range from 0.692, 7.99).

[Figure 1: Scatter Plot of Fish Consumption and Total Mercury Level]



**(3)** Calculate the correlation coefficient. What does the correlation tell us? (Q3 - 2 points)

|  |
| --- |
| r <- cor(df$Number.of.meals.with.fish, df$Total.Mercury.in.mg.g)  r # 0.6991094 |

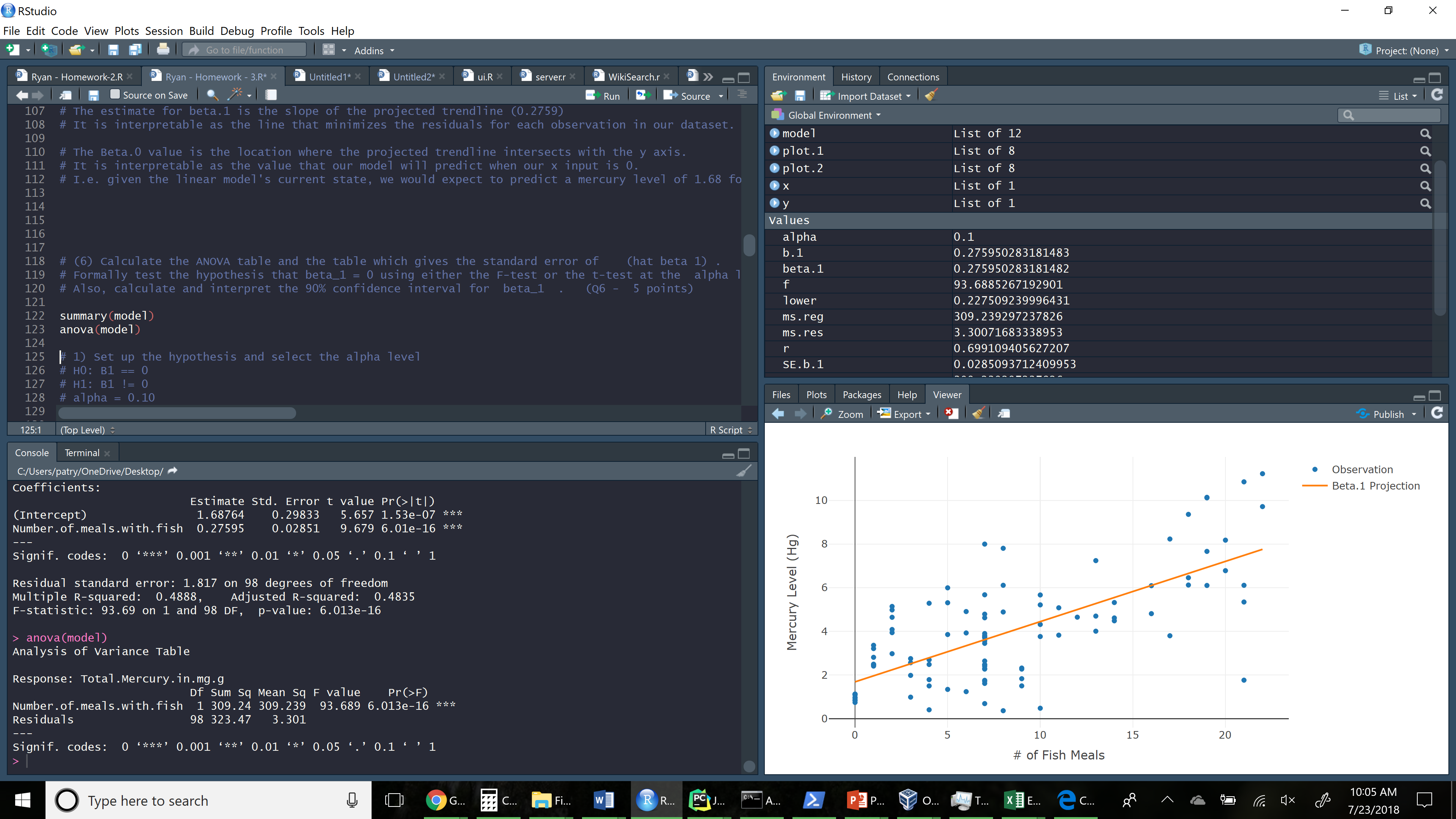
Explanation:

* The correlation is a measure of the strength and direction of a linear relationship between two quantitate variables in a sample.
* In this case, the correlation tells us that there is a strong relationship between a fish-heavy diet, and the presence of mercury in the sample groups hair.

**(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. (Q4 - 4 points)

|  |
| --- |
| # load linear model and predict trend line  model <- lm(Total.Mercury.in.mg.g ~ Number.of.meals.with.fish, data = df)  # Add trend line top previous plot  x <- list( title = "# of Fish Meals")  y <- list( title = "Total Me (Hg)")  plot.2 <- df %>%  plot\_ly(x = df$Number.of.meals.with.fish) %>%  add\_markers(y = df$Total.Mercury.in.mg.g, name="Observation") %>%  add\_lines( y = fitted(model), name="Beta.1 Projection") %>%  layout(xaxis = x, yaxis = y)  plot.2  # (Intercept) Number.of.meals.with.fish  # 1.6876426 0.2759503 |

[Figure 2: Scatter Plot of Fish Consumption and Total Mercury Level (with Beta.1 Included]



**(5)** 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? (Q5 - 4 points)

|  |
| --- |
| # Print intercept and slope  model  coef(model)  # (Intercept) Number.of.meals.with.fish  # 1.6876426 0.2759503 |

* The estimate for beta.1 is the slope of the projected trendline (0.2759)
* It is interpretable as the line that minimizes the residuals for each observation in our dataset.
* The Beta.0 value (1.687 ) is the location where the projected trendline intersects with the y axis.
* It is interpretable as the value that our model will predict when our x input is 0.
* I.e. given the linear model's current state, we would expect to predict a mercury level of 1.68 for individuals who have eaten 0 meals with fish.

**(6)** Calculate the ANOVA table and the table which gives the standard error of  (hat beta 1) . Formally test the hypothesis that beta\_1 = 0 using either the F-test or the t-test at the alpha level a=0.10. Either way, present your results **using the 5 step procedure as in the course notes**. Within your conclusion, calculate the R2 (R squared) value and interpret this.

Also, calculate and interpret the 90% confidence interval for beta\_1 . (Q6 - 5 points)

**1) Set up the hypothesis and select the alpha level**

# H0: B1 == 0

# H1: B1 != 0

# alpha = 0.10

**# 2) Select the appropriate test statistic**

# F = Reg MS / Res MS; with

# df = 1, 99

# Determine the appropriate value from the f-distribution table with df’s of 1,n-1.

qf(0.90, df1=1, df2=99)

**3) State the decision rule:**

# Decision rule:

# Reject H0 if F>= 2.756899

# Otherwise, do not reject H0.

**#4) Compute the test statistic**

anova.table <- summary(model)

ms.reg <- anova.table$"Mean Sq"[1]

ms.res <- anova.table$'Mean Sq'[2]

f <- ms.reg/ms.res

f # 93.68853

p <- pf(93.68853, df1=1, df2=98, lower.tail=F)

p # 6.013251e-16

**5) Conclusion**

# Reject H0 since 93.68853 >= 2.756899.

# We have significant evidence at the a=0.10 level that b.1 !=0.

# E.g. there is evidence of a significant linear association between a fish-based diet and mercury levels.

Alternatively, we could have also used R’s built in anova test function to reach the same conclusion automatically. Note that the F value matches the F-value calculated above (93.689). The very small P value (6.013e-16 ) signifies that it was very unlikely that this sample was arrived at by chance.

|  |
| --- |
| anova.table  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 |

**3 different ways to calculate r squared:**

|  |
| --- |
| **# Regression sum of squares / total sum of squares:**  anova.table <- anova(model)  ss.reg <- anova.table$`Sum Sq`[1]  ss.total <- anova.table$`Sum Sq`[1] + anova.table$`Sum Sq`[2]  ss.reg / ss.total  # [1] 0.488754  **# Squaring the correlation coefficient (r):**  cor(df$Number.of.meals.with.fish, df$Total.Mercury.in.mg.g)^2  # [1] 0.488754  **# Built-in anova test function**  anova.table <- summary(model)  anova.table$r.squared  # [1] 0.488754 |

**Interpretation of R2:**

* A R squared value of 48.87% indicates that our model explains about half  the variation in the response variable around out sample mean. (R2 = y-variation / total-variation)
* This r-squared value is influenced to some degree by the high variation in y values (e.g. the range of y values which correspond to a single x-value (i.e. 7).

**Construct a 90% confidence interval for Beta.1:**

|  |
| --- |
| alpha <- 0.10  t <- qt((1-alpha/2), df=98)  SE.b.1 <- anova.table$coefficients[4]  b.1 <- anova.table$coefficients[2]  lower <- b.1 - (t \* SE.b.1)  upper <- b.1 + t \* SE.b.1  # Print the output of the calculated boundaries:  lower # 0.228609  upper # 0.3232916  # Alternatively, we can conduct the same test, using R's built-in confint function:  confint(model, level = 0.90) |

Interpretation:

Using the t-distribution, we are 90% confident that the true value of Beta.1 is between 0.228609 and 0.3232916.

[Figure 3: R Code Used to Generate Answers Above]

|  |
| --- |
| library(ggplot2)  library(e1071)  library(dplyr)  library(plotly)  library(magrittr)  # 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 to a file (excel or CSV file) and read it into R memory for analysis. (Q1 - 2 points)  setwd("C:/Users/patry/OneDrive/Desktop")  # Assign the .csv information to a dataframe  df = read.csv("homework3.csv", header=T, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  # (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. (Q2 - 3 points)  library(plotly)  x <- list(  title = "# of Fish Meals")  y <- list(  title = "Mercury Level (Hg)")  plot.1 <- df %>%  plot\_ly(x = df$Number.of.meals.with.fish, trace ="scatter", mode="markers") %>%  add\_markers(y = df$Total.Mercury.in.mg.g, name="Observation") %>%  layout(xaxis = x, yaxis = y)  plot.1  # (3) Calculate the correlation coefficient. What does the correlation tell us? (Q3 - 2 points)  cor(df$Number.of.meals.with.fish, df$Total.Mercury.in.mg.g)  # The correlation is a measure of the strength and direction of a linear relationship between two quantative variables in a sample.  # In this case, the correlation tells us that there is a strong relationship between a fish-heavy diet, and the presense of mercuray in the sample groups hair.  # (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. (Q4 - 4 points)  # load linear model and predict trend line  df$trend=trend.line  df  model <- lm(Total.Mercury.in.mg.g ~ Number.of.meals.with.fish, data = df)  # Add trend line top revious plot  x <- list(  title = "# of Fish Meals")  y <- list(  title = "Mercury Level (Hg)")  plot.2 <- df %>%  plot\_ly(x = df$Number.of.meals.with.fish) %>%  add\_markers(y = df$Total.Mercury.in.mg.g, name="Observation") %>%  add\_lines( y = fitted(model), name="Beta.1 Projection") %>%  layout(xaxis = x, yaxis = y)  plot.2  # print intercept and slope  model  coef(model)  # (Intercept) Number.of.meals.with.fish  # 1.6876426 0.2759503  # (5) What is the estimate for beta\_1 ? How can we interpret this value? What is the estimate for beta\_0 ?  # What is the interpretation of this value? (Q5 - 4 points)  # The estimate for beta.1 is the slope of the projected trendline (0.2759)  # It is interpretable as the line that minimizes the residuals for each observation in our dataset.  # The Beta.0 value is the location where the projected trendline intersects with the y axis.  # It is interpretable as the value that our model will predict when our x input is 0.  # I.e. given the linear model's current state, we would expect to predict a mercury level of 1.68 for individuals who have eaten 0 meals with fish.  # (6) Calculate the ANOVA table and the table which gives the standard error of (hat beta 1) .  # Formally test the hypothesis that beta\_1 = 0 using either the F-test or the t-test at the alpha level a=0.10. Either way, present your results using the 5 step procedure as in the course notes. Within your conclusion, calculate the R2 (R squared) value and interpret this.  # Also, calculate and interpret the 90% confidence interval for beta\_1 . (Q6 - 5 points)  summary(model)  anova(model)  # 1) Set up the hypothesis and select the alpha level  # H0: B1 == 0  # H1: B1 != 0  # alpha = 0.10  # 2) Select the appropriate test statistic  # 3) State the decision rule.  # F = Reg MS / Res MS; with  # df = 1, 99  qf(0.90, df1=1, df2=99)  # Decision rule:  #Reject H0 if F>= 2.756899  #4)  # The sum of squared errors is sum(anova.table$residuals^2).  # The mean squared error is mean(anoval.table$residuals^2)  anova.table <- anova(model)  anova.table  ms.reg <- anova.table$"Mean Sq"[1]  ms.reg  ms.res <- anova.table$'Mean Sq'[2]  ms.res  f <- ms.reg/ms.res  f # 93.68853  # Reject H0 since 93.68853 >= 2.756899.  # We have significant evidence at the a=0.10 level that b.1 !=0.  # E.g. there is evidence of a significant linear association between a fish-based diet and mercury levels.  # Scratchpad  # 3 different ways to calculate r squared:  # Regression sum of squares / total sum of squares:  anova.table <- anova(model)  ss.reg <- anova.table$`Sum Sq`[1]  ss.total <- anova.table$`Sum Sq`[1] + anova.table$`Sum Sq`[2]  ss.reg / ss.total  # [1] 0.488754  # Squaring the correlation coefficient (r):  cor(df$Number.of.meals.with.fish, df$Total.Mercury.in.mg.g)^2  # [1] 0.488754  # Built-in anova test function  anova.table <- summary(model)  anova.table$r.squared  # [1] 0.488754  # Construct a 90% confidence interal:  alpha <- 0.05  t <- qt((1-alpha/2), df=98)  t  SE.b.1 <- anova.table$coefficients[4]  b.1 <- anova.table$coefficients[2]  lower <- b.1 - (t \* SE.b.1)  upper <- b.1 + (t \* SE.b.1)  lower # 0.228609  upper # 0.3232916  # Using the t-distribution, we are 90% confident that the true value of Beta.1 is between 0.2275092 and 0.3243913  # Alternatively, we can conduct the same test, using R's builtin confint function:  confint(model,level = 0.9)  anova.table |