Dhakal_Module3

Kokil Dhakal

2023-11-20

Loading data set in the environment

```
mercury.content.df <- read_excel("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module3/module3.xlsx")
```

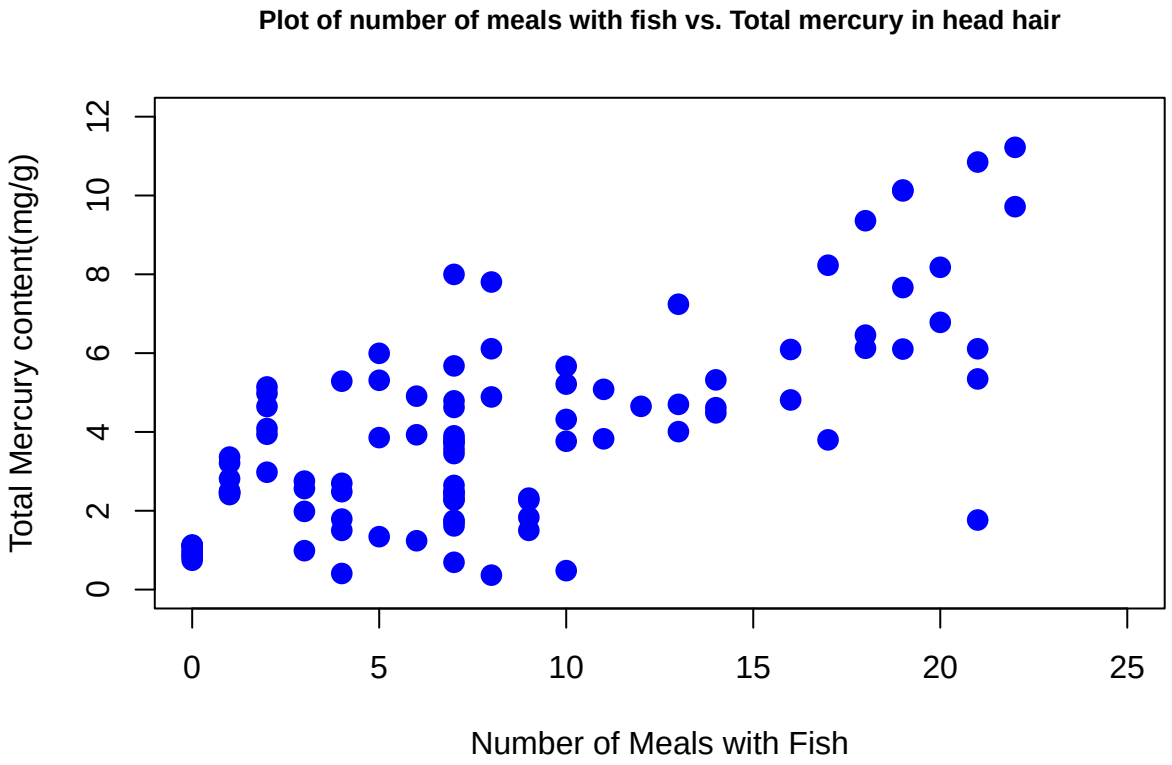
Question 1

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 scatter plot, describe the form, direction, and strength of the association between the variables.

Answer

First, plotting the data set to know about data.

```
par(cex.main = 0.8)
plot(mercury.content.df$Number_of_meals_with_fish,
     mercury.content.df$`Total_Mercury_in_mg/g`,
     main="Plot of number of meals with fish vs. Total mercury in head hair",
     xlab="Number of Meals with Fish",
     ylab="Total Mercury content(mg/g)",
     pch=16,
     col="blue",
     cex=1.5,xlim = c(0,25),ylim = c(0,12))
```



The form of the scatter plot describes the overall pattern of the data points.Based on values of variable, I set up xlim and ylim for the plot. In this case, it seems that there is a general trend or pattern, but it's not perfectly linear.Since number of meals with fish could affect the total mercury content in head hair, number of meals with fish is plotted in x-axis(independent variable) while total mercury content will be in y-axis(dependent variable).Similarly, the direction is positive as total mercury content increasing with increase in number of meals with fish.The strength of the association can be assessed by how closely the points cluster around the line. In this case, the strength appears moderate, with some variability in total mercury content for a given number of meals with fish.This interpretation is based on visual inspection, and for a more quantitative analysis, we might consider calculating the correlation coefficient and other measures.

Question 2

Calculate the correlation coefficient.What does the correlation tell us?

Answer

```
correlation_coefficient <- cor(mercury.content.df$Number_of_meals_with_fish,mercury.content.df$`Total_M
cat("Correlation Coefficient:", round(correlation_coefficient,3), "\n")
```

```
## Correlation Coefficient: 0.699
```

The correlation coefficient is a measure of the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship,-1 indicates a perfect negative linear relationship and 0 indicates no linear relationship.Since we got correlation coefficient

value of 0.699 which indicates the relationships is positive.A positive correlation coefficient suggests that as the number of meals with fish increases, the total mercury content(mg/g) tends to increase.This value also suggest strength of association and is moderately strong as we can see from scatter plot where there is a linear association and has moderate positive r value.The magnitude of the correlation coefficient gives an indication of the strength of this relationship. The closer the correlation coefficient is to 1 (or -1), the stronger the linear relationship. If it's closer to 0, the relationship is weaker(however we should not only depend on correlation coefficient value to describe the strength of association between variables as this value may mislead in case of non-linear relationship.

Question 3

Find the equation of the least squares regression equation and write out the equation. Add the regression line to the scatter plot you generated above.

Answer

We can use different methods to find the slope and intercept which are needed to make the equation of the least squares regression.I am using the method that is use for linear regression model building and then find the slope and intercept.

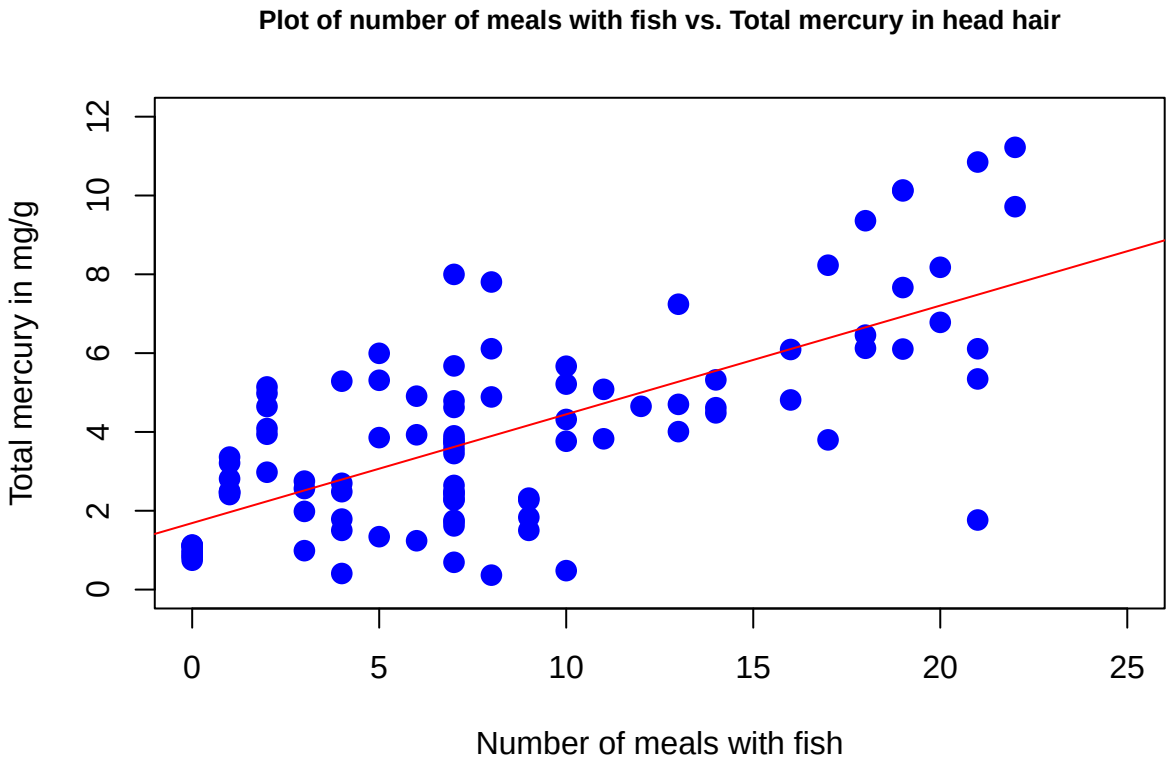
```
# building linear model
model.3 <- lm(`Total_Mercury_in_mg/g` ~ `Number_of_meals_with_fish`,data = mercury.content.df)
intercept.3 <- coef(model.3)[1] # first value as intercept
slope.3 <- coef(model.3)[2] # second value as slope of the model
cat("Regression Equation1: y =", round(intercept.3, 3), "+", round(slope.3, 3), "x\n")
```

```
## Regression Equation1: y = 1.688 + 0.276 x
```

Adding the line to the above plot

```
par(cex.main = 0.8)
plot(mercury.content.df$Number_of_meals_with_fish, mercury.content.df$`Total_Mercury_in_mg/g`,
     main="Plot of number of meals with fish vs. Total mercury in head hair",
     xlab="Number of meals with fish", ylab="Total mercury in mg/g",
     pch=16, col="blue",cex=1.5,xlim = c(0,25),ylim = c(0,12))

abline(model.3,col="red")
```



For this question, I have used linear regression model using lm() which gives the slope and intercept. After that putting those values to the equation y= beta0 + beta1 x which gives y = 1.688 + 0.276 x. After that to add the regression line to the scatter plot I just use the abline() method.

Alternative way of calculating slope and intercept and then writing equation

```
# Alternative method
r <- correlation_coefficient
# yhat = beat0 + beta1x ->The equation for the least-squares regression line
# yhat = predicted value of  y for a given value of x
# beta0 = least-squares estimates of beta0 (the intercept)
# beta1 = least-squares estimates of beta1 (slope)
sd.mercury.content <- sd(mercury.content.df$`Total_Mercury_in_mg/g`)
mean.mercury.content <- mean(mercury.content.df$`Total_Mercury_in_mg/g`)
sd.num.fish.meal <- sd(mercury.content.df$Number_of_meals_with_fish)
mean.num.fish.meal <- mean(mercury.content.df$Number_of_meals_with_fish)
slope.4 <- round(r*sd.mercury.content/sd.num.fish.meal,3)
intercept.4 <-round( mean.mercury.content - slope.4* mean.num.fish.meal,3)
cat("Regression Equation method2: y =", round(intercept.4, 3), "+", round(slope.4, 3), "x\n")
```

```
## Regression Equation method2: y = 1.687 + 0.276 x
```

Question 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.

Answer

```
xbar <- mean(mercury.content.df$Number_of_meals_with_fish)
sx <- sd(mercury.content.df$Number_of_meals_with_fish)
ybar <- mean(mercury.content.df$`Total_Mercury_in_mg/g`)
sy <- sd(mercury.content.df$`Total_Mercury_in_mg/g`)

beta1hat <- round(r*sy/sx,3) # estimate of beta1
beta0hat <- round(ybar-beta1hat*xbar,3) # estimate of beta0
beta1hat
```

```
## [1] 0.276
```

```
beta0hat
```

```
## [1] 1.687
```

In equation of $yhat = \beta 0hat + \beta 1hat.x$, $\beta 0hat$= 1.687 and $\beta 1hat$ = 0.276.Slope in the regression equation means change in unit value of y per unit increase in x.The estimate of the slope parameter $\beta 1hat$ represents the estimated change in the response variable (y, total mercury content) for a one-unit change in the predictor variable (x, number of meals with fish), assuming a linear relationship. In this case, for each additional meal with fish, the total mercury is estimated to increase by 0.276 mg/g, according to the model. The estimate of the slope parameter $\beta 1hat$ also gives insight into the direction of the relationship between the variables which is positive(0.276). The intercept($\beta 0hat$) can be interpreted as the average mercury level that is 1.6876 mg/g in fisherman's head hair even having meals with no fish (i.e it represents the starting point of the regression line).

Question 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 $\beta 1$

Answer

calculating anova table and summary(model) table

Table 1: model summary table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.6876	0.2983	5.6570	0
Number_of_meals_with_fish	0.2760	0.0285	9.6793	0

Table 2: ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Number_of_meals_with_fish	1	309.2393	309.2393	93.6885	0
Residuals	98	323.4702	3.3007	NA	NA

```
anova.table <- anova(model.3) # model.3 is our linear regression model from answer 3.

R.squared <- anova.table$`Sum Sq`[1]/sum(anova.table$`Sum Sq`)
# table that gives standard error of estimated slope
kable(summary(model.3)$coefficients,align = "c",digits = 4,caption = "model summary table")
```

```
# anova table
kable(anova.table,align = "c",digits = 4,caption = "ANOVA table")
```

ANOVA table gives all the values like mean squares and sum squares for both regression and residual while summary(model) table provides a list of values including the standard error of the estimated slope(0.27595) which is 0.02851.

Now I am going to test hypothesis using five steps procedure.

- Setting up hypotheses and selecting alpha level
  a) $H0 : \beta1 = 0$ (there is no linear association) -> Null hypothesis
  b) $H1 : \beta1 \neq 0$ (there is a linear association) -> Alternative hypothesis
  c) $\alpha = 0.05$
- Selecting the appropriate test statistic
  t = estimated beta1/standard error of beta1 with n-2 df
- Stating the decision rule
  Determining the appropriate critical value from the t-distribution using software with n-2 = 100-2 = 98 degrees of freedom and associated with a right hand tail probability of alpha/2 = 0.05/2 =0.025 using qt(0.025,98,lower.tail = FALSE)

```
qt(0.025,98,lower.tail = FALSE)
```

```
## [1] 1.984467
```

```
decision rule : reject null hypothesis if |t| >= 1.984467
```

```
Otherwise, do not reject null hypothesis
```

- Compute the test statistic

  Calculating the t values using values from above summary table

  t = 0.2760/0.0285 = 9.68 (same value can be obtained from model summary table)

- Conclusion

  Since test statistic(9.68) is greater than critical value (1.984467), there is significant evidence at the alpha 0.05 level that $\beta 1 \neq 0$ which means we reject null hypothesis.This also means that there is linear association between Number_of_meals_with_fish and Total_Mercury_in_mg/g (here p <0.001 from above table)

Now calculating R-squared value

This value can be obtained from the summary table of the model as well which is 0.4888. Also, this can be calculated using ANOVA table from above.

R-squared = Regression sum of squares/ Total sum of squares

```
R.square <- anova.table$`Sum Sq`[1]/sum(anova.table$`Sum Sq`)
R.square
```

```
## [1] 0.488754
```

```
cat(round(R.square *100,2), "%")
```

```
## 48.88 %
```

The value of R-squared is 48.88% which means 48.88 % of the variability in the variable Total_Mercury_in_mg/g can be explained by the variable Number_of_meals_with_fish.

Calculating 90% confidence interval for $\beta 1$

confidence interval can be calculated using R

```
confi.intv <- confint(object = model.3,level = 0.9)

lower.val <- round(confi.intv[2,1],3)
upper.val <- round(confi.intv[2,2],3)
lower.val
```

```
## [1] 0.229
```

```
upper.val
```

```
## [1] 0.323
```

From above table, each side of the distribution curve contain 5 % which adds up to 10 % and middle part will be 90%. From above calculation, we can say with 90% of confidence that the true value of beta1(slope) is between 0.229 and 0.323.

Alternative way of calculating confidence interval.

beta1 ± t.val with n-2 of df and alpha/2 *standard error of beta 1 in confidence interval we have lower and upper values which is calculated using R as follows

```
# Since it is 90% confidence interval which means each side will have 5%
#i.e 0.05. degree of freedom is n-2 i.e 100-2 =98
t <- qt(0.05,98,lower.tail = FALSE)
#beta1 and standard error of beta1 are obtained by summary(model) table from
#above.
lower.value <- 0.27595-t*0.02851
upper.value <- 0.27595+t*0.02851
round(lower.value,3)
```

```
## [1] 0.229
```

```
round(upper.value,3)
```

```
## [1] 0.323
```

** The End **