QSCI 381 AB - AUT 22
Homework #9- Vinsensius

BONUS (5 points)

In order to be granted these bonus points (which count on top of the total point score above), please fill out the course evaluation here. We rely on your feedback to improve the course in subsequent years. You can complete this in your own time, or during labs, whichever you prefer. Once you have completed the evaluation, please write that you have completed the evaluation below.

Evaluation completed: Vinsensius on 08/12/2022. (Thank you for the great quarter!)

The PDF file of the code is attached at the end of the HW

Iris Dataset

For this part of the lab you will be using the iris dataset already built into R. Your goal is to look for correlation among the different variables (sepal length, petal length, and petal width) and test for significant relationships. To load the iris data into RStudio, use the command `data(iris)`. You can view the column headers using the command `head(iris)`.

(1) Start by plotting all three combinations of sepal length, petal length, and petal width (sepal length & petal length, sepal length & petal width, petal length & petal width) as scatterplots. Paste your R code and the three plots here (6 points)

(2) Just by visually inspecting the scatter plots, do any of the data look like they might be correlated? Why? (2 points)

(3) Next we are going to find the correlation coefficient, r, for all three combinations.

$$r = \frac{n\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}$$

Using the formula above, find the correlation coefficient r for the combination of sepal length and petal length. To make this easier, rename the variables

"x", "y", and "z", determine n using length(x) and run the following code, and paste your output below (4 points).

This code example gives the correlation between x and y.

```
r <- (n * sum(x*y) - sum(x) * sum(y)) / (sqrt(n *
  sum(x^2) - sum(x)^2) * sqrt(n * sum(y^2) - sum(y)
  ^2))
```

(4) Now find the correlation coefficient for the same combination as 2a, but this time use the cor() command instead of the formula. Hint: you should get the same answer. (2 points)

(5) Now find the correlation coefficients for the other two remaining combinations of variables. (2 points)

(6) Rank the comparisons from strongest relationship (1) to weakest (3). (3 points)

(7) For the comparison with the weakest relationship, test whether it is statistically significant at $\alpha = 0.05$. State your null and alternative hypotheses. (2 points)

(8) Define your sampling distribution, number of tails, n, and degrees of freedom. (2 points)

(9) Calculate the critical value(s) for the relevant standardized test statistic using the function `qt(p = alpha/2, df = degrees of freedom)`. (2 points)

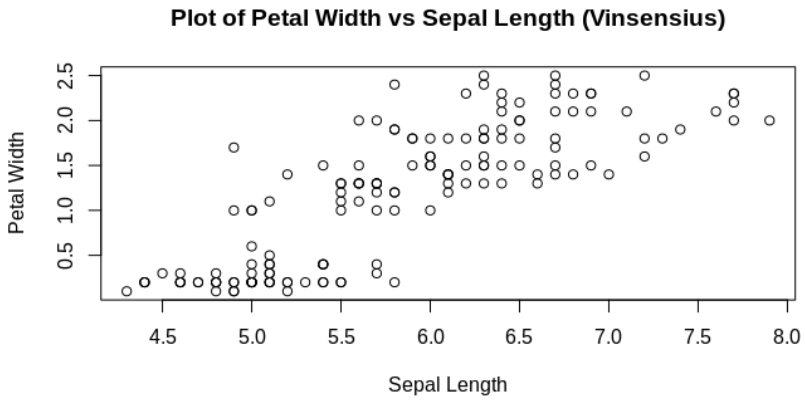
(10) Calculate the appropriate standardized test statistic using the formula below and report your answer. (2 points)

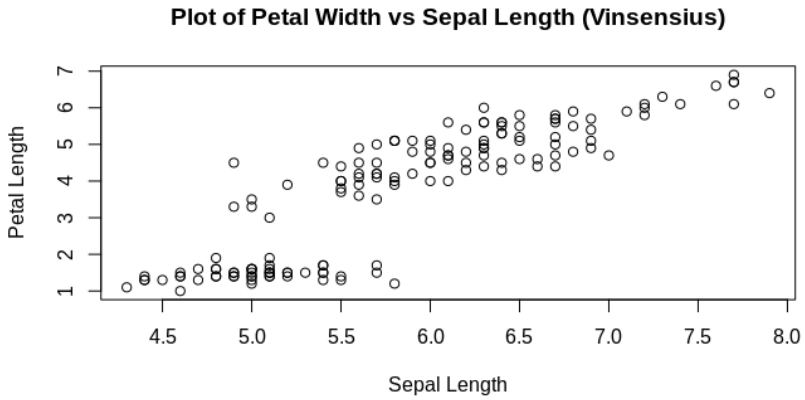
$$t = \frac{(r\sqrt{n-2})}{\sqrt{1-r^2}}$$

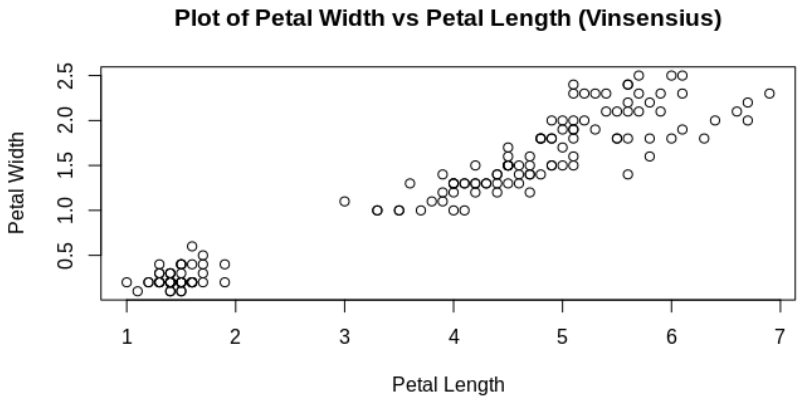
(11) What is the result of your test? Report in terms of your hypotheses. (2 points)

(1) The code is

```
  plot(iris$Sepal.Length, iris$Petal.Width,
     main="Plot of Petal Width vs Sepal Length (Vinsensius)",
     xlab="Sepal Length", ylab="Petal Width")
plot(iris$Sepal.Length, iris$Petal.Length,
     main="Plot of Petal Width vs Sepal Length (Vinsensius)",
     xlab="Sepal Length", ylab="Petal Length")
plot(iris$Petal.Length, iris$Petal.Width,
     main="Plot of Petal Width vs Petal Length (Vinsensius)",
     xlab="Petal Length", ylab="Petal Width")
```







(2) All plots look positively correlated because the trend of y-values are increasing as x-values increase.

(3) r = 0.872 The code is

```
x <- iris$Petal.Length
y <- iris$Sepal.Length
z <- iris$Petal.Width
n <- length(x)
r <- (n * sum(x*y) - sum(x) * sum(y)) /
  (sqrt(n * sum(x^2) - sum(x)^2) * sqrt(n * sum(y^2) - sum(y)^2))
```

(4) r = 0.872. So, using the function `cor()` indeed gives the same answer as using the formula for calculating the correlation value between x and y data. The code is

```
cor(x,y)
```

(5) The r between x and z is 0.963. The r between y and z is 0.818.

```
cor(x,z)
cor(y,z)
```

(6) The first rank is relation between x and z. The second rank is relation between x and y. The third rank is relation between y and z.

(7) The null hypothesis is that there is no correlation between y and z. The alternate hypothesis is that there is correlation between y and z.

(8) The test is two-tailed t-test. n = 150 and df = 148.

(9) The t-critical values are ± 1.98

```
qt(0.05/2, dfx)
qt(1-0.05/2, dfx)
```

(10) The standardized test statistic t is 17.3

```
r <-cor(y,z)
t <- (r*sqrt(dfx))/(sqrt(1-r^2))
```

(11) Since $t > t_{\text{crit}}$, we will reject the null hypothesis. So, there is a correlation between y and z.

Tree Food Regression

You take a summer internship in southeast Alaska evaluating the relationship between salmon abundance and riparian vegetation density among 15 streams. You hypothesize that the nutrients salmon bring back to streams in their bodies makes for larger and more abundant trees. You measure the basal area of trees in square meters per acre, within 100m of fish bearing streams. You then use the Alaska Department of Fish and Game's salmon run data to find the average number of fish (in 1000's) for each river.
These data can be found in the "TreeFood.csv" file on Canvas.

```
salmon<-read.csv("TreeFood.csv", header=TRUE)
head(salmon)
```

(12) Create a scatter plot with salmon abundance on the x-axis and the tree basal area on the y-axis. Label the graph, label each axis appropriately, and make it colorful. (3 points)

(13) Find the line of best fit (regression line) for the salmon and tree area data using the equations found in the lab slides 14 and 15.

   (a) First, calculate and interpret r. (Hint: define n, x, and y first.) (4 points)

   (b) Next, calculate the slope (b_1) for the regression line. (2 points)

   (c) Now calculate the intercept (b_0) for the regression line. (2 points)

   (d) Finally, write out the regression equation in terms of the variables Abundance and Tree.Area. (2 points).

(e) Based on the regression equation above, calculate the predicted Tree.Area for salmon abundance values of: (4 points)

(i) 1.9

(ii) 2.7

(14) Interpret the slope in the context of the original question. What does it tell you about the potential effect of salmon on tree density? Have you proven this to be true using this analysis? Is there another potential explanation? (3 points)

(15) Now we want to calculate the coefficient of determination (r^2). Remember that the r^2 is calculated as the explained variation divided by the total variation.

(a) Start by first calculating the predicted value of y ($\hat{y}$) for each value of x. (2 points)

(b) Next calculate the total variation as the sum of the squared differences between each observed y (y_i) and the mean of y ($\bar{y}$). (2 points)

(c) Then calculate the explained variation as the sum of the squared differences between each predicted y ($\hat{y}$) and the mean of y ($\bar{y}$). (2 points)

(d) Finally calculate the coefficient of determination: (2 points)

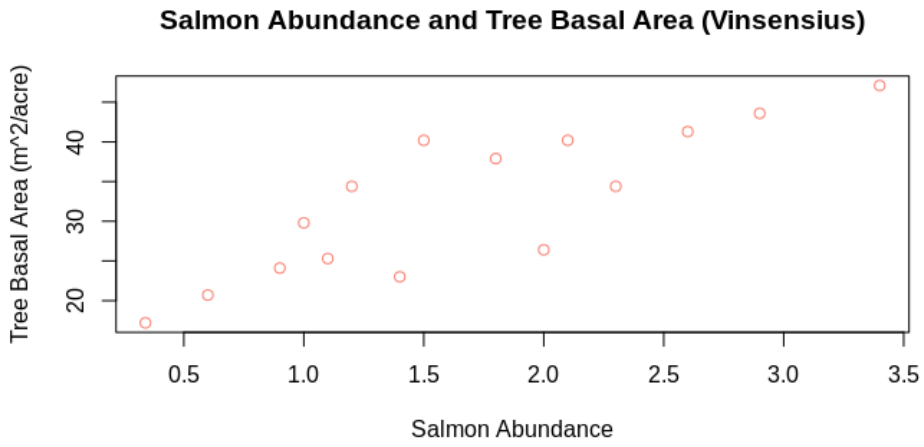
(e) Lastly, let's let RStudio make our lives easier by using the `lm()` function.

(i) State your null and alternative hypotheses for the slope (b1). (2 points)

(ii) Now use the `lm()` and `summary()` functions in RStudio to perform a linear regression of the salmon abundance and tree area data. Paste your code and the resulting output. (2 points)

(16) You'll see that not only do you get the same estimates for the intercept (b_0) and slope (b_1), but also the standard errors, t-values, and P values for those estimates. Is our slope estimate significant (assume an alpha of 0.05)? How can you tell? What does that mean in terms of our original question? (3 points)

(12) The plot is



The code is

```
plot(Tree.Area~Abundance, data = salmon, col="salmon",
     main="Salmon Abundance and Tree Basal Area (Vinsensius)",
     xlab="Salmon Abundance", ylab="Tree Basal Area (m^2/acre)")
```

(13) (a) The r is 0.849. This means that the is a positive correlation between x and y variables.

```
n<-length(salmon$Abundance)
x<-salmon$Abundance
y<-salmon$Tree.Area
r<-(1/(n-1)*sum(((x-mean(x))/sd(x))*((y-mean(y))/sd(y))))
```

(b) The slope is 8.99

```
b1<-sd(y)/sd(x)*r
```

(c) The intercept is 17.3

```
x.bar<-mean(x)
y.bar<-mean(y)
b0<-y.bar-b1*x.bar
```

(d) Tree.Area = 8.99*Abundance + 17.3

(e) (i) Predicted Tree.Area= 34.4

(ii) Predicted Tree.Area= 41.6

```
ta1<- b0 + b1*1.9
ta2<- b0 + b1*2.7
```

(14) There is a positive potential effect that salmon bring on the tree density because the slope is positive value. We have not proven it through analysis. However, we can not rule out possibility that there are external factor affecting the salmon abundance since there is no control being added to the experiment.

(15) (a) The value for predicted y-value is

```
[20.3, 22.7, 25.4, 26.3, 27.2, 28.1, 29.9, 30.8, 33.5, 35.3,
↪  36.2, 38.0, 40.7, 43.4, 47.9]
```

the code is

```
y.hat<-b0+b1*x
```

(b) Total variation is 1199

```
total.var<-sum((y-y.bar)^2)
```

(c) The explained variance is 863

```
explained.var<-sum((y.hat-y.bar)^2)
```

(d) The r^2 is 0.720

```
r2<-explained.var/total.var
```

(e) (i) The null hypothesis is that $b_1 = 0$, while the alternate hypothesis is that $b_1 \neq 0$.

(ii) The result is given as below

```
Call:
lm(formula = Tree.Area ~ Abundance, data = salmon)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8861 -2.5813 -0.7722  3.7594  9.4089

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.306      2.918   5.931 4.98e-05 ***
Abundance      8.990      1.555   5.781 6.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.082 on 13 degrees of freedom
Multiple R-squared:   0.72,     Adjusted R-squared:  0.6984
F-statistic: 33.42 on 1 and 13 DF,  p-value: 6.37e-05
```

The code is

```
salmon.lm<-lm(Tree.Area~Abundance, data=salmon)
summary(salmon.lm)
```

(16) Since the p-value for the slope is $6.37e-5 < \alpha = 0.05$, so we will reject the null hypothesis that there is no relationship between abundance of salmon and the tree density. This means that there is a relationship between the abundance of salmon and the tree density, possibly a positive correlation based on the regression calculation. However, we can not rule out a possibility of third factor that we have not account when doing the analysis.

Alligator Egg Dataset

The American Alligator (Alligator mississippiensis) lays eggs into nests that are dug into river banks. During incubation, the sex ratio of alligator hatchlings can be influenced by the temperature of the nest. Load the alligator sex ratio data into R

```
gator <- read.csv(file="alligator_sex_ratio.csv")
```

(1) Plot the proportion of female hatchlings (p_female) against Temperature, making sure to label your plot appropriately. Provide a written description of the shape of this relationship? [8 pts]

(2) What is the correlation between p.female and temperature? Is this correlation significantly different from zero (H0: $\rho = 0$) at $\alpha = 0.01$? [10 pts]

(3) Using the `lm()` function, fit a regression model, and provide estimates for the intercept and slope of the fitted regression line. Provide a written description of what the slope parameter represents for this model, making particular reference to the numeric value in your answer. [5 pts]

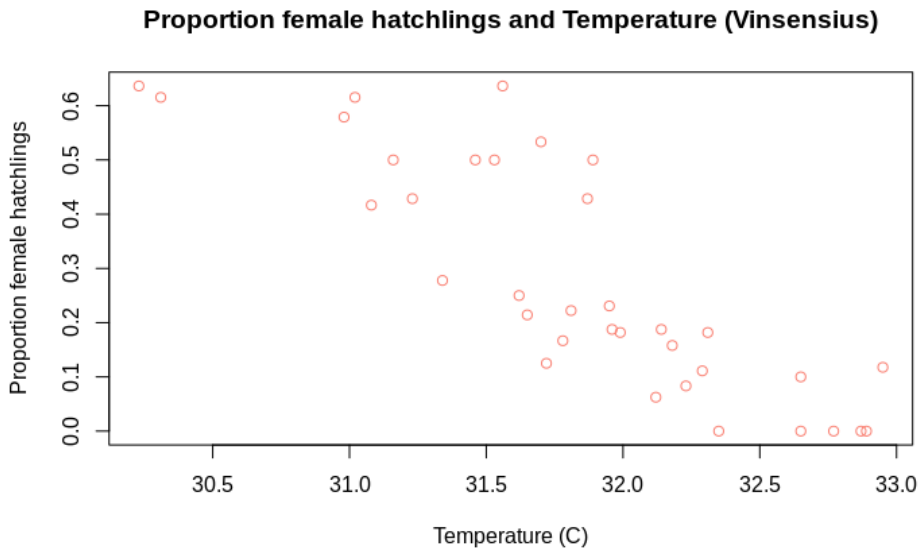
(4) Use the regression line fitted in (3) to predict the expected proportion of female hatchlings at nest temperatures of

   (i) 32°C [2 pts]

   (ii) 33°C [2 pts]

   Based on these values, and thinking about the range of observations that are possible, what would you interpret about the general applicability of this model? [2 pts]

(1) Based on the figure, the relationship will be negatively correlated between temperature and proportion of female hatching from eggs because the downward trajectory.




```
plot(p.female~temperature, data = alligator_dat, col="salmon",
     main="Proportion female hatchlings and Temperature
     ↪  (Vinsensius)",
     ylab="Proportion female hatchlings", xlab="Temperature (C)")
```

(2) The r is -0.841. To test the significance, we will apply two-tailed t test. The standardized test statistic t is -8.94. The p-value of the test statistic is 2.51e-10, which is less than $\alpha = 0.01$. So, the correlation is significantly different from zero. Thus, there is relationship between the proportion female hatching and temperature, probably negative correlation based on the correlation value that we calculated.

```
df = NROW(alligator_dat$temperature)-2
r = cor(alligator_dat$temperature,alligator_dat$p.female)
t = r*sqrt(df)/sqrt(1-r^2)
pval = 2*pt(t,df) # run as two tailed
```

(3) The slope is -0.269, while the intercept is 8.84. The slope means that for every 1 degree Celcius increase in temperature, the proportion of the female hatching from eggs will dropped by about 0.269 or 26.9%.

```
model = lm(p.female~temperature, data = alligator_dat)
```

(4) (i) Predicted proportion at 32^oC is 0.234

(ii) Predicted proportion at 33^oC is -0.0347

```
new_temp <- data.frame(temperature=c(32,33))
predicted_proportion = predict(model, newdata=new_temp)
```

The range of possible temperatures will be the temperatures that will give proportion between 0 and 1 because that is how proportion is define. That means that beyond this range of temperature, the model will not be applicable because it will give proportion that is negative value or greater than 1 which does not make any sense. Thus, based on the values 32^oC a possible value for temperature, while 33^oC is not because the predicted value for proportion is negative, which is impossible.