STAT22000 Summer 2020 Homework 14 Solutions

All page, section, and exercise numbers below refer to the course text (*OpenIntro Statistics*, 3rd edition, by Diez, Barr, and Cetinkaya-Rundel.).

Reading: Section 7.2-7.4
Problems for Self-Study :

1. Exercise 7.7, 7.19, 7.21, 7.25, 7.27, 7.31, 7.37, 7.41 on p.358-371 where the answers can be found at the end of the book.

2. A number is missing in each of the data sets below. If possible, fill in the blank to make the correlation r equal to 1. If this is not possible, explain why not. *Hint: Make a scatterplot. Under what circumstance will the correlation equal to 1?*

(a)

x	y
1	0
2	2
2	2
4	–

(b)

x	y
1	0
2	2
3	5
4	–

Answer: [*3pts each*]

(a) The missing number is 6. A correlation of 1 means that there is a perfect linear relationship between x and y. As the line connecting $(1, 0)$ and $(2, 2)$ is $y = 2x - 2$, when y is 4, x must be $2 \times 4 - 2 = 6$.

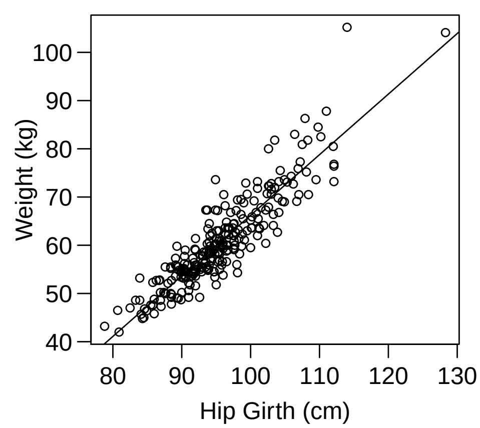
(b) Impossible to make $r = 1$ because the three points $(1, 0)$, $(2, 2)$, and $(3, 5)$ do not lie on a straight line. The slope of the segment connecting $(1, 0)$ and $(2, 2)$ is $(2 - 0)/(2 - 1) = 2$ but the slope of the segment connecting $(2, 2)$ and $(3, 5)$ is $(5 - 2)/(3 - 2) = 3$.

---

3. The scatterplot on the right shows the weights (in kg) and hip girths (in cm) of 249 physically active women age 18-45. Here is a summary of the data:

	Body weight (kg)	Hip girth (cm)
Mean	60.694	95.603
SD	9.639	6.945

Correlation $r \approx 0.905$



(a) How would the correlation r change if weight was measured in pounds while the units for hip girth remained in centimeters? (1 pound = 0.454 kg).

Answer: [*1pt*] The correlation is not affected by the unit used. The correlation will remain $r \approx 0.905$.

(b) Write down the equation of the regression line for predicting a woman's weight in kilograms from her hip girth in centimeters.

Answer: [*2pts = 1pt for slope + 1pt for intercept*] Here y = weight and x = hip girth.

Slope $= r \times \dfrac{s_y}{s_x} = 0.905 \times \dfrac{9.639}{6.945} \approx 1.256$

Intercept $= \overline{y} -$ slope $\times\, \overline{x} = 60.694 - 1.256 \times 95.603 \approx -59.4$

Equation of the regression line:

$$\text{predicted weight (in kg)} = -59.4 + 1.256 \times \text{hip girth (in cm)}$$

(c) Interpret the slope and the intercept of the equation in the previous part in this context.

Answer: [*2pts = 1pt for slope + 1pt for intercept*]
Slope: For each extra cm in hip girth, a woman is expected to weigh 1.256 kg more on average.
Intercept: A woman with a hip girth of 0 cm are expected to weigh -59.4 kg, which is meaningless here since nobody has a hip girth of 0 cm.

(d) Calculate R^2 of the regression line for predicting weight from hip girth, and interpret it in the context of the application.

Answer: [*2pts = 1pt for R^2 + 1pt for the interpretation*] $R^2 = 0.905^2 \approx 0.819$, meaning about 81.9% of the variation in women's weights can be explained by their hip girths.

(e) A randomly selected female student from your class has a hip girth of 90 cm. Predict the weight of this student using the regression line.

Answer: [*1pt*] To predict one's weight from her hip girth, plug in 90 for hip girth in the regression in part (b)

$$\text{predicted weight (in kg)} = -59.4 + 1.256 \times 90 = 53.64\text{kg}$$

(f) The student in the previous part weighs 55 kg. Calculate the residual, and explain what this residual means.

Answer: [*2pts = 1pt for the residual + 1pt for the interpretation*]
Residual e_i = observed $y_i -$ predicted $\widehat{y}_i = 55 - 53.64 = 1.36$ kg, meaning that the regression line underestimated this student' weight by 1.36 kg (or the predicted weight is 1.36 kg lower than the actual weight).

(g) A one-year-old baby has a hip girth of 52 cm. Would it be appropriate to use the regression line in part (b) to predict the weight of this baby?

Answer: [*1pt*] No. From the scatterplot, we can see the hip girths of the 249 women range from 80 cm to 130 cm. Predicting the weight of a child with a hip girth of 52 cm would be extrapolation. The linear relation may not hold outside the range of the data.

(h) Can we use the regression line in part (b) to predict the weight of an adult man with a hip girth of 110 cm? Explain your answer.

Answer: [*1pt*] No. The regression line is based on the body measurements of 249 women, which may not apply on men.

(i) Can we use the regression line in part (b) to predict the hip girth of a 35-year old woman who weighs 80 kg? Explain your answer.

Answer: [*1pt*] No. The regression line in part (b) is only for predicting a female's weight from her hip girth, not the other way around. The regression line for predicting one's hip girth from one's weight is a different line.

(j) Find the equation of the regression line for predicting a woman's hip girth from her weight, and use the equation to predict the hip girth of a 35-year old woman weighs 80 kg.

Answer: [*2pts*] Now y = hip girth and x = weight.
Slope $= r \times \frac{s_y}{s_x} = 0.905 \times \frac{6.945}{9.639} \approx 0.652$
Intercept $= \overline{y} - \text{slope} \times \overline{x} = 95.603 - 0.652 \times 60.694 \approx 56.03$
Equation of the regression line:

$$\text{predicted hip girth (in cm)} = 56.03 + 0.652 \times \text{weight (in kg)}$$

Plugging in weight = 80 kg, we get predicted hip girth $= 56.03 + 0.652 \times 80 = \boxed{108.19}$ cm.
Remark. If we plug in weight = 80 kg in the equation in part (b), we will get

$$80 = -59.4 + 1.256 \times \text{hip girth (in cm)} \quad \Rightarrow \quad \text{hip girth} = \frac{80 + 59.4}{1.256} \approx 110.99 \text{ cm.}$$

This is wrong.

---

4. A biologist was interested in the relationship between the velocity at which a beluga whale swims and the tail-beat frequency of the whale. A sample of 19 whales was studied and measurements were made on swimming velocity, measured in units of body lengths of the whale per second and tail-beat frequency, measured in units of hertz (number of beats per second). The data file `BelugaSwim.txt` is posted on Canvas with this exercise.

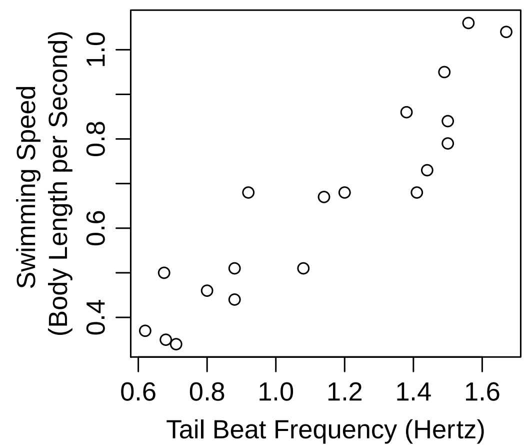
(a) Make a scatterplot with tail beat frequency (in Hertz) on x-axis and the swimming speed on y-axis. Label the plot properly. Describe the relationship between the two variables.

```
whale = read.table("BelugaSwim.txt", h=T)
plot(whale$freq, whale$speed,
     xlab = "Tail Beat Frequency (Hertz)",
     ylab = "Swimming Speed \n(Body Length per Second)")
```

Review Section 3 in Lab #1 `http://www.stat.uchicago.edu/~yibi/s220/labs/lab01.html` *about changing the working directory if you have trouble loading the data file to R.*

Answer:

[2pts = 1pt for the plot + 1pt for the linearity] There appears to be a $\boxed{\text{linear}}$ relationship between the two variables.



(b) Find the means and the standard deviations of the two variables and their correlation coefficient ($\overline{x}$, $\overline{y}$, s_x, s_y, and r) in R or by a calculator.

```
library(mosaic)
favstats(~freq, data=whale)
favstats(~speed, data=whale)
with(whale, cor(freq, speed))
```

Answer: *[0pt]*

	Frequency	Speed
Mean	$\overline{x} = 1.1334$	$\overline{y} = 0.6558$
SD	$s_x = 0.3534$	$s_y = 0.2267$
Correlation	$r = 0.9234$	

R codes (not required):

```
> favstats(~freq, data=whale)
  min   Q1 median    Q3  max     mean        sd  n missing
 0.62 0.84   1.14 1.465 1.67 1.133421 0.3534121 19       0
> favstats(~speed, data=whale)
  min   Q1 median    Q3  max      mean        sd  n missing
 0.34 0.48   0.68 0.815 1.06 0.6557895 0.2267234 19       0
> with(whale, cor(freq, speed))
[1] 0.9233792
```

(c) Here we fit a simple linear regression model in R using the `lm()` function, in which `lm` stands for "linear model".

```
lmwhale = lm(speed ~ freq, data=whale)
```

The general syntax to fit a model with the response variable `y` and explanatory variable `x` is `lm(y~x, data=nameofdataset)`. We can save the fitted model by giving it a name. You can name it whatever you like, such as `lmwhale`. We can call a saved model whenever we need it. For example, to get the intercept and slope of the fitted regression line we can type `lmwhale$coef` and then get the following output.

```
> lmwhale$coef
(Intercept)        freq
-0.01561813  0.59237262
```

The equation of the regression line is then

$$\text{predicted speed} = -0.01561813 + 0.59237262 \times (\text{tail beat frequency in hertz})$$

Verify that the slope and the intercept given by R are $r \cdot s_y/s_x$ and $\overline{y} - (\text{slope}) \cdot \overline{x}$ respectively. Show your computation.

Answer: [*0pt*] y = speed, x = frequency.

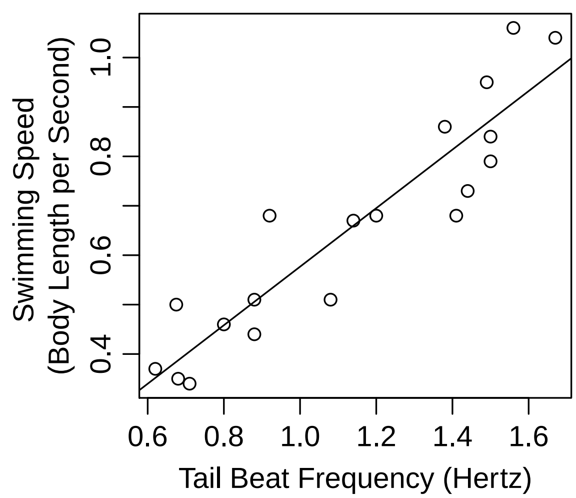
$$\begin{aligned} \text{slope} &= r\frac{s_y}{s_x} \approx 0.9234\frac{0.2267}{0.3534} \approx 0.5923 \\ \text{intercept} &= \overline{y} - (\text{slope})\overline{x} = 0.6558 - (0.5923)(1.1334) \approx -0.0155 \end{aligned}$$

which agree with the slope and the intercept given by R.

(d) Add the regression line to the scatter plot using the R command below

```
plot(whale$freq, whale$speed,
     xlab = "Tail Beat Frequency (Hertz)",
     ylab = "Swimming Speed \n(Body Length per Second)")
abline(lmwhale)
```

Answer: [*0pt*]



(e) The R command `summary(lmwhale)` gives a more detailed output for the model.

```
> summary(lmwhale)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01562    0.07075  -0.221    0.828
freq         0.59237    0.05973   9.917 1.75e-08 ***
```

Test the null hypothesis that the slope of the regression line is 0.8 against a 2-sided alternative. Report the test statistic with degrees of freedom, and the P-value.

Answer: *[4pts = 2pt for the t-statistic + 1pt for the df + 1 pt for the P-value.]*
From the summary output, the estimated slope is $b_1 = 0.59237$ with standard error $\mathrm{SE}(b_1) = 0.05973$. To test whether the slope β_1 is 0.8, the t-statistic is

$$t = \frac{b_1 - 0.8}{\mathrm{SE}(b_1)} = \frac{0.59237 - 0.8}{0.05973} = \boxed{-3.476} \quad \text{with df } = n - 2 = 19 - 2 = \boxed{17}.$$

The two-sided P-value about 0.00289 can be obtained by either of the following R commands.

```
> 2*pt(-3.476, df=17)
[1] 0.002890581
> 2*pt(3.476, df=17, lower.tail=F)
[1] 0.002890581
```

(f) Calculate a 95% confidence interval for the slope of the regression line for predicting the swimming speed of a beluga whale (in the number of body lengths of the whale per second) from its tail beat frequency (in hertz), and interpret the interval in context of the data.

Answer: *[4pts = 1pt for the value of t^* + 1pt for the SE + 1pt for the CI + 1pt for the interpretation.]*
With $n - 2 = 19 - 2 = 17$ degrees of freedom, the critical value $t^* \approx 2.11$ is found in R as follows.

```
> qt(0.05/2, df=17, lower.tail=F)
[1] 2.109816
```

The 95% CI for the slope is

$$\text{estimate} \pm t^*\,\mathrm{SE} = 0.59237 \pm 2.11 \times 0.05973 = 0.59237 \pm 0.12603 \approx (0.46634, 0.71840)$$

Interpretation: For every extra beat per second, the swimming speed of a beluga whale (in the number of body lengths of the whale per second) is 0.46634 to 0.71840 body lengths faster per second on average, with 95% confidence. *[0.5 pt off if missing "on average".]*
