PUBHLTH 490Z: Linear Regresesion

Due Monday October 11th

Topics

- Correlation
- Least-squares regression
- Multiple regression

Guppy Coloration. Guppies (*Poecilia reticulata*) are small, brightly colored tropical fish often seen in freshwater fish aquariums, and a popular model species for sexual selection research. Sexual selection refers to the process by which individuals of one gender choose mates of the other gender to sexually reproduce with; for most species, females select mates, while males compete with each other to attract females. Male sexual ornaments such as elaborate tail plumage are known to play a role in female choice; it is thought that these ornaments reflect male fitness or genetic quality.

One measure of fitness involves heterozygosity (having different alleles at a given genetic locus). Thus, it might be expected that more elaborate ornamentation is associated with more heterozygosity.

Researchers investigated the correlation between heterozygosity and male coloration in a wild population of guppies.[1] Male guppies are covered in a mixture of colored spots; orange coloration is consistently preferred by females. Heterozygosity was assessed by genotyping 9 loci and calculating the proportion of loci that are heterozygous (MLH, multilocus heterozygosity).

The file `orange.guppies.Rdata` contains the raw data collected by researchers for 147 male guppies sampled from a river in northern Venezuela. Sampled fish were euthanized, photographed on their left sides, and fixed in 95% ethanol until molecular analyses were performed. Imaging software was used to measure the body area covered in orange spots, body length (excluding the tail fin), and body height (at the caudal peduncle).

a) Examine the distribution of `MLH` in the data using the `hist()`, `boxplot()`, and `summary()` commands. Are the data continuous or discrete? Explain your answer.

[1] Herdegen, M., et al. Heterozygosity and orange coloration ... in the guppy (*Poecilia reticulata*). *Journal of Evolutionary Biology* 2013; **27**: 220-225.

b) Body length multiplied by body height at the caudal peduncle was used to estimate body size. Calculate body size, then create the variable `relative.orange` for the area of orange coloration relative to body size. Examine the distribution of `relative.orange` using the `hist()`, `boxplot()`, and `summary()` commands; describe the distribution and point out any interesting features.

c) Create a scatterplot of `relative.orange` and `MLH`, with `MLH` on the x-axis. Plot the least-squares regression line. Does a linear model seem to be a good fit for the data?

d) Using the `summary()` command, conduct a hypothesis test to determine whether relative orange area is significantly associated with MLH. Do the results suggest that more elaborate sexual ornaments are associated with increased heterozygosity? Explain.

Resting Heart Rates. The file `heartrate.exercise.Rdata` contains data collected from a class survey – 43 students were asked for their gender, resting heart rate, and number of hours spent exercising in a typical week. The purpose of collecting the data was to determine what factors lead to differences in resting heart rate.

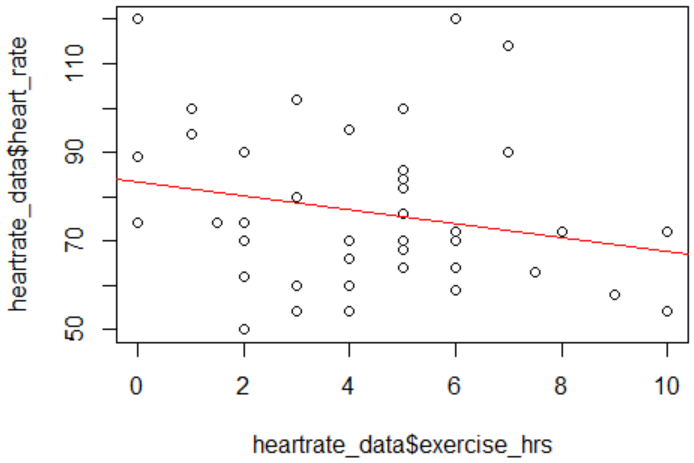
The R output of the regression summary for resting heart rate and exercise hours is shown below, with some parts removed – use the printed output to answers parts a) through c). The scatterplot with a fitted line is also included.

```
summary(lm(heartrate_data$heart_rate ~ heartrate_data$exercise_hrs))

Call:
lm(formula = heartrate_data$heart_rate ~ heartrate_data$exercise_hrs)

Residuals:
Min      1Q  Median      3Q     Max
-30.144 -11.096  -5.445   9.205  46.121

Coefficients:
Estimate Std. Error
(Intercept)                   83.277      5.357
heartrate_data$exercise_hrs   -1.566      1.070
---
Residual standard error: 17.42 on 41 degrees of freedom
Multiple R-squared:  0.0497,  Adjusted R-squared:  0.02652
```



a) Write the regression equation.

b) From the output, what is the correlation between exercise and heart rate? What proportion of the total variability in heart rate does this model predict?

c) Is exercise a statistically significant predictor of resting heart rate? Conduct a formal hypothesis test with the information provided from the output.

d) The factor (i.e. categorical variable) `gender` might also be associated with resting heart rate. Regress `heart_rate` on both `exercise_hrs` and `gender`. What are the interpretations of the coefficient estimates in the model?

e) There are the two methods to test whether heart rate is significantly associated with gender. Since gender is a binary predictor, it is also possible to conduct a two-sample t-test, comparing resting heart rates between males and females. The output is shown below:

```
> t.test(heartrate_data$heart_rate ~ heartrate_data$gender, var.equal=TRUE)
```

```
Two Sample t-test

data:  heartrate_data$heart_rate by heartrate_data$gender
t = 0.54424, df = 41, p-value = 0.5892
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-8.032095 13.958182
sample estimates:
mean in group female   mean in group male
78.05000             75.08696
```

Describe another method to test whether heart rate is significantly associated with gender. Describe any assumptions that must be made and compare the results of this approach with that from the two-sample t-test results shown above.

Food insecurity data challenge: In preparation for the course project, read the 3 articles on graphing food insecurity in the US:

- https://urban-institute.medium.com/graphing-food-insecurity-in-the-united-states-9a3d9961b4e9
- https://www.ers.usda.gov/topics/food-nutrition-assistance/food-security-in-the-us/key-statistics-graphics.aspx
- https://www.whitehouse.gov/cea/blog/2021/07/01/federal-income-support-helps-boost-food-security-rates/.

Write 1-page summaries on **each** of these articles.