431 Lab 04

Deadline: See Course Calendar | Last Edited 2022-08-23 12:56:43

Table of contents

Deadline

Lab 04 has 8 questions, all of which you need to complete by the deadline specified on the Course Calendar.

- To receive full credit on a Lab, it must be received on Canvas no later than 59 minutes after the posted deadline. (This allows for small issues with uploading to Canvas to occur without penalty.)

Getting Help

You are welcome to discuss Lab 04 with Professor Love, the teaching assistants or your colleagues, but your answer must be prepared by you alone. Don't be afraid to ask questions, using any of the methods described on our Contact Us page.

Learning Objectives

1. Critically evaluate a news story on scientific literature, incorporating the principles of probability and odds
2. Given data, work through a linear regression model including visualizing the data, building the model, predicting an outcome on new points.

Getting Started with Lab 04

We encourage you to use a similar approach to Lab 04 that you used in Labs 02 and 03. We have not provided a template for Lab 04, but you can adapt one of the ones we've provided previously, if you like. Or you can use an approach that you think works well. Be sure to use a new section for each question on the Lab, and do not hide your code. (If you want to use code-folding, set it to show.)

Part A: News Story and Research Article (Questions 1-5)

Find a headline from the internet related to health or medicine that describes the findings of a study published on January 1, 2017 or later. Then find the study being referred to in PUBMED. Use the formula for updating your opinions about health news developed in this article by Jeff Leek, along with the abstract and full contents of the published study to complete Questions 1-5. While it won't be necessary to prepare any R code to respond to Questions 1-5, we think it will be good practice for you to prepare your response in R Markdown anyway.

Question 1 (5 points)

Specify the URL where we can see the headline and news story describing the findings of the study. Feel free to use `bit.ly` or a related tool online to produce a shortened URL for this purpose. Specify the reference completely, including the names of the author(s) of the news story, and its full title, and source.

Question 2 (5 points)

Specify a URL where we can see at least the abstract of the complete study. Again, shortened URLs are fine. Give the complete reference to the study, as well, including the authors, full title, journal name and so forth.

Question 3 (15 points)

Describe, in a few sentences, your original opinion (gut feeling) related to the conclusions of the study as summarized in the headline and news article, first in terms of a probability statement, and then calculate the appropriate odds, remembering to convert statements about probabilities to statements about odds. Provide some motivation for your internal prior probability, describing your relevant personal experiences or other factors that drove your gut feeling.

Remember, if X is an event, and `Pr(X)` is the probability that X occurs, and `odds(X)` are the odds that X occurs, then

```
Pr(X) = odds(x) / (1 + odds(x))
```

and

```
odds(X) = Pr(X) / (1 - Pr(X)).
```

Question 4 (30 points)

Evaluate the study in terms of the six specifications proposed by Jeff Leek in this article at FiveThirtyEight when evaluating study support. Be sure to specify your conclusion about **each** of the six specifications, and provide direct quotes and summarize the evidence from the abstract or paper to address the issues raised and justify your conclusions. We want to see a clear, motivated conclusion about each of the six specifications, as well as direct quotes and evidence summaries to address the issues raised and justify conclusions. We suggest you use a different subheading for each of the six specifications so it's easy for us to see your conclusions in each case.

Question 5 (15 points)

Incorporate the study support assessment into a Bayes' Rule calculation to obtain the final odds you should now be willing to give to the headline, and specify this value in terms of a probability statement, as well. Then react to the final conclusion specified by this approach in a few sentences. How does your subjective posterior probability that the headline is true match up with the formula's conclusions? Do you feel that the formulaic approach has yielded an appropriate conclusion for you in this case? Why or why not?

Part B: Palmer Penguins (Questions 6-8)

The Data for Lab 04, Questions 6-8

In Questions 6-8, this lab again uses the the `penguins` data (note: use the `penguins` tibble, and not the `penguins_raw` tibble for this Lab) contained in the `palmerpenguins` package in R. The complete citation is …

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Additional information on the data are provided by Allison Horst at the github site linked above. In particular, you'll find a nice cartoon of the three species of penguin contained in the data and a detailed description of the bill measurements that are worth your time.

Question 6 (5 points)

Start with the 333 penguins in the `penguins` tibble who have complete data, and create two tibbles. The first should be called `pen_train` and should contain a random sample of exactly 200 of the 333 penguins with complete data, while the second tibble, called `pen_test` should contain the other 133 penguins with complete data. Use `4312021` to set your seed for random sampling so that we all obtain the same sample. Use R code to obtain the samples, and then provide code which demonstrates that your two samples contain exactly 200 and exactly 133 penguins. Comment on the code you present as necessary so that we can clearly see that you understand what is being done in each step.

Question 7 (15 points)

Build a linear model to examine the relationship between body mass in grams (the outcome of interest) and bill length in millimeters (our predictor) using ordinary least squares and the `pen_train` data set you created in Question 6. Call this model `model1`.

Create an attractive and thoughtfully labeled plot (including an appropriate title) that shows the association of bill length (placed on the horizontal x axis) and body mass (on the vertical y axis) for the 200 penguins in your training sample.

- Add appropriate smooth curves (using both lm and loess) to the plot.
- Also please add the actual regression equation (including the coefficients, rounded to two decimal places) to the plot in a clear way.
  - Note that it's 100% OK to place a label on the graph where you type in the coefficients you want, rather than having this done automatically.
  - For example, if you wanted a line like body mass = 23.13 - 3.86 bill length to be placed on your plot, centered at the value x = 4 and y = 5, you could use…

```
annotate("text", x = 4, y = 5, label = "body mass = 23.13 - 3.86 bill length")
```

or

```
geom_label(x = 4, y = 5. size = 3, color = "red",
           label = glue("body mass = 23.13 - 3.86 bill length"))
```

for example.

Then write a couple of sentences interpreting what this figure tells you about the relationship between `bill_length_mm` and `body_mass_g`.

Question 8 (10 points)

Use the `model1` you created in Question 7 to predict the data in `pen_test` and summarize the results by specifying the root mean squared prediction error, and the mean and maximum absolute prediction error you obtain across the 133 penguins in `pen_test`. Place the results in an attractive and clear table. Then describe the units of measurement for your root mean squared prediction error result in a sentence.

Include the session information

At the end of your R Markdown file, please include a new code chunk to provide the **session information**. You can use either the approach from Lab 2 or Lab 3.

Submitting the Lab

You should build your entire response as an R Markdown file. Then use the Knit button in RStudio to create the resulting HTML document. Be sure to review the HTML result to ensure that it looks clean and clear, that the labels on your plots and other output are easy to read, and that it doesn't retain any unnecessary warning messages or other material that distracts from your work. Be sure to spell-check your work before submission.

Submit **both** your revised R Markdown file **and** the HTML output file to the Lab 04 section in the Assignments folder in Canvas by the deadline specified in the Course Calendar. We will need both the R Markdown and HTML file submitted before we can grade your work.

Again, we encourage you in the strongest possible terms to **ask questions**, using any of the approaches described on our Contact Us page.

Grading

We will summarize some of the more interesting responses to Questions 1-5 after the Lab has been graded.

- This Lab will be graded on a scale from 0-100.
- Note that the teaching assistants will review your responses to all Questions carefully to assess clarity of writing, attention to detail, and adherence to grammatical and syntax requirements. Spelling, grammar, syntax and the rest all matter for grading purposes in this and all other assignments this term.

A detailed answer sketch for this Lab will be provided on the day after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.

Late Penalties for Lab Work

- Labs that are turned in 1-12 hours after the deadline will lose 10% of available points.
- Labs turned in more than 12 but less than 72 hours after the deadline will lose 25% of available points.
- No extensions to Lab deadlines will be permitted this semester. Labs turned in more than 72 hours after the deadline will receive no credit.
- Note that your lowest lab score (out of Labs 1-7) over the course of the semester will be dropped before we calculate your lab grade.