431 Lab 06

Deadline: See Course Calendar | Last Edited 2022-11-12 17:02:22

Table of contents

Deadline

Lab 06 has 6 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.)

Learning Objectives

1. Be able to work through a simple linear regression
2. Visualize and interpret the role of a potential confounder
3. Run a multivariable, model adjusting for this confounder, with an interpretation of the estimate and confidence interval
4. Expand this model and interpret the results and conclusion
5. Compare multiple linear models using various metrics

An Important Note

Your response to **every** question, whether we explicitly ask for it or not, should include a complete English sentence responding to the question. Code alone is not a sufficient response, even if the code is correct. Some responses might not need any code, but every response needs at least one complete sentence.

The Data for Lab 06

In Lab 06 we'll be using the lindner dataset again that we saw in Lab 05. The data come from "an observational study of 996 patients receiving an initial Percutaneous Coronary Intervention (PCI) at Ohio Heart Health, Christ Hospital, Cincinnati in 1997 and followed for at least 6 months by the staff of the Lindner Center. The patients thought to be more severely diseased were assigned to treatment with abciximab (an expensive, high-molecular-weight IIb/IIIa cascade blocker); in fact, only 298 (29.9 percent) of patients received usual-care-alone with their initial PCI.". Information on the lindner dataset and its variables can be found at this site.[1,2]

[1] Rdocumentation. (n.d.). lindner: Lindner Center Data On 996 PCI Patients Analyzed By Kereiakes Et Al. (2000). Retrieved from https://www.rdocumentation.org/packages/MatchLinReg/versions/0.7

[2] Kereiakes DJ, Obenchain RL, Barber BL, et al. Abciximab provides cost effective survival advantage in high volume interventional practice. Am Heart J 2000; 140: 603-610.

We'd like you to begin by loading the `lab05_lind.Rds` data set provided for Lab 5 into a tibble in a new R Project (for Lab 6). Call that initial tibble `lindner`.

Background

You're a statistician tasked with analyzing the `lindner` data. The principal investigator wants to examine the relationship between a predictor: the ejection fraction (`ejecfrac`) and an outcome: 6-month cardiac-related costs (`cardbill`), **among those patients who were alive at 6 months**. There are a number of data cleaning steps you'll need to do after reading in the data (which you should call `lindner`). This includes (a) select only those patients who were alive at 6 months (call this `lindner_alive`) , (b) you'll want to add an `id` to be able to properly identify patients since there are no unique identifiers, `row_number()` could be one approach, and (c) you'll want to partition your data to a 70% training (call this `lindner_alive_train`) and 30% test sample (call this `lindner_alive_test`), using `set.seed(431)`. Use a 95% confidence level throughout this Lab.

Your first step should look something like:

```
## you'll need code here to load the lab05_lind.Rds file
## into a tibble called lindner before you run something like...

lindner_alive <- lindner |>
    filter(sixMonthSurvive == 1) |>
    mutate(id = row_number()) |>
    as_tibble()
```

Question 1 (20 points)

Given the information above, work through an appropriate analysis of the data. Specifically do the following: (a) decide whether a square-root or log transformation of the outcome is more appropriate (only select between these two options), make said transformation, run a simple linear regression, and interpret and contextualize these results. The decisions regarding transformations as well as the build and interpretation of your model should be completed using just the training data set (`lindner_alive_train`) and should be called `model1`.

Question 2 (10 points)

Now we want to examine the effect of a third variable, `abcix` or whether or not the patient had the abciximab augmentation, on the relationship between our main predictor and our

outcome. Run, and discuss, a new linear regression which adjusts your original model for this variable. Call this `model2`. Again, this should be done on your training data set.

Question 3 (10 points)

The investigator of the study has now asked you to add the following variables to your models: `stent`, `height`, `female`, `diabetic`, `acutemi`, and `ves1proc`. Assess the suitability of adding these variables (i.e. check for potential correlation issues), and then run and interpret this model. Call this `model3`.

Question 4 (10 points)

It's been suggested that the effect of `height` depends on `female`, which would suggest the desire to include an interaction term between these two variables in the model. Add this interaction term, run the model, and briefly discuss whether or not we see the interaction between these variables impacting `cardbill`. Call this `model4`.

Question 5 (20 points)

By now you should have created 4 models. Fit these models to the test data we had held-out earlier. Then, compare these models, using their adjusted R^2, AIC and BIC (from the training data), as well as their MAPE, RMSPE, and maximum prediction error (from the test data.) Which model performs best in which settings?

Question 6 (20 points)

Write a brief essay (150 words would be sufficient, but you can write more if you like) which relates what you've done in this assignment to what you learned in your reading of Spiegelhalter.

Session Information

Be sure to include the session information using one of the methods we have demonstrated.

Submitting your Response

Submit both your revised R Markdown file and the HTML output file to Canvas in the Lab 06 section of the Assignments folder by the deadline specified in the Course Calendar.

Getting Help

You are encouraged to discuss Lab 06 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.

Grading

We will summarize some of the more interesting responses to Question 6 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.
- Your lowest lab score (out of Labs 1-7) will be dropped before we calculate your lab grade.