431 Lab 06

Deadline: THURSDAY 2023-11-09 at Noon

431 Staff

2023-10-15

Table of contents

0.1 Deadline

Lab 06 has 5 questions, all of which you need to complete by the deadline specified on the Course Calendar.

0.2 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

0.3 Getting Help

You are welcome to discuss Lab 05 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.

0.4 Our Labs Page

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

Visit our 431-Labs page for more information, including:

- how to get help with the Lab,
- our late policy,
- details on answer sketches and grading rubrics
- how your lab will be graded, and
- how we handle requests to regrade a Lab.

0.5 The Data for Lab 06

In Lab 06 we'll be using the lindner dataset 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.".

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`.

0.6 Background

You are 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()
```

1 Question 1 (12 points)

Given the information above, work through an appropriate analysis of the data. Specifically do the following: (a) decide whether an inverse 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 in at least two complete English sentences. 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`. Be sure to include a check of all four main residual plots for `model1` with a sentence or two interpreting those results.

2 Question 2 (8 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 (transformed) outcome. Run, and discuss, a new linear regression which adjusts your original model to include this new predictor. Call this `model2`. Again, this should be done on your training data set. Be sure to include a check of all four main residual plots.

3 Question 3 (8 points)

The principal investigator has now asked you to add two more variables to your models: `height` and `female`. 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 your (transformed) `cardbill`, again solely using your training data. Call this `model3`. Be sure to include a check of all four main residual plots.

4 Question 4 (12 points)

By now you should have created 3 models. Fit these models to the test data we held out earlier. Then, compare these models, using:

- their adjusted R^2, AIC and BIC (from the training data, predicting a transformed outcome), as well as
- their MAPE, RMSPE, and maximum prediction error (from the test data.)
    - Evaluate MAPE, RMSPE and maximum prediction error in your test data only after backing out of the transformation you made earlier.

Which of your three models performs best according to each measure?

5 Question 5 (10 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 Lab to what you learned in your reading of Spiegelhalter's *The Art of Statistics*.

6 Session Information

Be sure to include the session information as a section of its own in your response, using either the `xfun` or `sessioninfo` approach.

7 Hint: R Packages

The packages I used in the answer sketch are:

`broom`, `janitor`, `knitr`, `patchwork` and `tidyverse`.

8 What Happens After Submission?

Submit **both** your Quarto file **and** the HTML output file to the Lab 06 section in the Assignments folder in Canvas by the deadline specified in the Course Calendar. We will need both the Quarto and HTML file submitted before we can grade your work.

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

- This Lab will be graded on a scale from 0-50 points.
- 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 the Lab will be provided 48 hours after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.