REVISED 431 Lab 6 Instructions

Fall 2024 - deadline in Course Calendar

Thomas E. Love

2024-11-12

Table of contents

> **!** Important
>
> - This Lab contains 5 tasks for you to complete.
> - The deadline for completing this Lab is posted in the Course Calendar.

0.1 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

0.2 Getting Started

To start, create a directory on your computer for `lab6`. We suggest this be a directory you control, called `lab6`, and we recommend you create it as a subdirectory of a `2024-431` directory on your machine.

Now, open RStudio, and use the **File** … **New Project** … **Existing Directory** menu to create an R Project in your `lab6` directory in which you will do Lab 6.

0.3 There is no Quarto Template for Lab 6

In this Lab, you will prepare a report in the form of an HTML file, using Quarto. We have provided previous Lab 1 and Lab 2 Quarto document templates. Modify one of those to complete your work for Lab 6, or create something new that works similarly.

0.4 Getting Help

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

0.5 Using AI / ChatGPT, etc.

The use of AI or ChatGPT in addressing these Tasks should lead to "C" level work. If you are happy receiving a "C" on this assignment, go ahead.

0.6 R Packages

> **!** Important
>
> I have added some packages to this list from what was originally posted, to reflect what I used in the Answer Sketch.

```
library(janitor)
library(mice)
library(naniar)
library(patchwork)
library(broom)      # for augment
library(car)        # for boxcox
library(gt)         # just for neatening up some tables
library(easystats)
library(tidyverse)

source("data/Love-431.R")

theme_set(theme_lucid())
knitr::opts_chunk$set(comment = NA)
```

0.7 Specifications for Responses

- If you need to set a seed, use `431` as your seed.
- If you need to fit a bootstrap, use 2000 replications.
- Use a **95%** confidence level throughout this Lab.

0.8 Background and Data for Lab 6

In Lab 6 we'll be using a subset of a data set we used in Lab 05. The Lab 6 data are a sample of 800 subjects who were alive at 6 months after participating in "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.".

0.9 Data Management

Begin by loading the `lab6_lindner800.csv` data set provided for Lab 6 into a tibble in a new R Project (for Lab 6). Call that initial tibble `lab6_raw`. **Do not** use the version of the lindner data that we used in Lab 5.

There are five key data steps you'll need to do after reading in the raw `.csv` file.

1. Cleaning the variable names up seems important.
2. You'll want to add an `subject` ID code to be able to properly identify patients since there are no unique identifiers. We'll use the row numbers for this purpose.
3. You'll need to use the `mice` package to build 5 imputations of the complete data set to deal with missing data in one of the variables, which we will assume is missing at random (MAR).
4. You'll then select the third of those imputations for your main work on the Lab.
5. After creating the third imputation and saving it as a tibble of its own, you'll partition it into to a 70% training (call this `lab6_training`) and 30% test sample (call this `lab6_test`), using `set.seed(431)`.

Your data management activities should look like this:

```
lab6_raw <- read_csv("data/lab6_lindner800.csv", show_col_types = FALSE) |>
  janitor::clean_names() |>
  mutate(subject = as.character(row_number())) |>
  relocate(subject)
```

```
miss_var_summary(lab6_raw)
```

```
# A tibble: 6 x 3
  variable  n_miss pct_miss
  <chr>      <int>    <num>
1 height        25     3.12
2 subject        0     0
3 ejec_frac      0     0
4 card_bill      0     0
5 abcix          0     0
6 female         0     0
```

```
prop_miss_case(lab6_raw)
```

```
[1] 0.03125
```

```
set.seed(431)
lab6_imps <- mice(lab6_raw, m = 5, printFlag = FALSE)
```

If you get a warning here about logged events, you can silence it, as I have done in preparing these instructions.

```
imp_3 <- complete(lab6_imps, 3) |> tibble()
dim(imp_3)
```

```
[1] 800   6
```

```
n_miss(imp_3)
```

```
[1] 0
```

```
set.seed(431)
lab6_training <- slice_sample(imp_3, prop = 0.7, replace = FALSE)
lab6_test <- anti_join(imp_3, lab6_training, by = "subject")
```

0.9.1 Checking Your Work

If you want to check that you've done this in the same way that I have, here are the first two observations for each of my two partitions.

```
head(lab6_training, 2)
```

```
# A tibble: 2 x 6
  subject ejec_frac card_bill abcix height female
  <chr>       <dbl>     <dbl> <dbl>  <dbl>  <dbl>
1 64             51      7039     0    185      0
2 435            55     14844     1    180      0
```

```
head(lab6_test, 2)
```

```
# A tibble: 2 x 6
  subject ejec_frac card_bill abcix height female
  <chr>       <dbl>     <dbl> <dbl>  <dbl>  <dbl>
1 2              60     11357     1    163      0
2 11             60     12751     1    157      1
```

1 Question 1 (10 points)

In Question 1, the principal investigator wants to examine the relationship between a predictor: the ejection fraction (`ejec_frac`) and an outcome: 6-month cardiac-related costs (`card_bill`).

Given the information above, work through an appropriate simple regression analysis of the data.

> **!** Important
>
> This next section has changed. It used to say…
>
> > Specifically do the following: (a) decide whether an inverse or log transformation of the outcome (`card_bill`) is more appropriate (only select between these two options), make said transformation, run a simple linear regression using OLS, and interpret and contextualize these results in at least two complete English sentences. The decisions regarding which transformation to use as well as the build and interpretation of your model should be completed using just the training data set (`lab6_training`) and your model should be called `fit1`.
>
> As of 2024-11-12, I have **simplified** your instructions to the following:
>
> - first, demonstrate the use of a Box-Cox procedure to obtain an estimate as to whether a logarithm or inverse would be a better choice for transforming the `card_bill` outcome.
> - transform the outcome (`card_bill`) by taking its inverse and multiplying the result by 1,000,000, using something like:

```
lab6_training <- lab6_training |>
  mutate(trans_cb = 1000000/card_bill)
```

> - run a simple linear regression which you'll call `fit1`, to predict this transformed outcome using `ejec_frac` as a single predictor and fitting the model using ordinary least squares, and just the training data set (`lab6_training`.)

After fitting your model `fit1`, do the following…

a. Describe the meaning of the point estimate for the slope of `ejec_frac` in your `fit1` model appropriately, and specify a 95% confidence interval around that slope.

> **Note**
>
> Note that the units for `ejec_frac` are percentage points, but you can call them points, or units, in your response. You should not specify units for your transformed outcome.

b. Use `check_model()` to display and then, in a few sentences, evaluate the posterior predictive check, as well as the plots to check linearity, homogeneity of variance, influential points, and normality of residuals. What problems with OLS assumptions, if any, do you see?

2 Question 2 (10 points)

Now we want to examine the effect of a third variable, `abcix`, which specifies whether or not the patient had the abciximab augmentation (`abcix` = 1 means they did have the augmentation), on the relationship between our main predictor (`ejec_frac`) and our (transformed) `card_bill` outcome. Run, and discuss, a new linear regression which adjusts your original model to include this new predictor. Call this `fit2`. Again, this should be done on your training data set.

After fitting your model `fit2`, do the following…

a. Describe the meaning of the point estimate for the slopes of `ejec_frac` and then `abcix` in your `fit2` model appropriately, and specify a 95% confidence interval around each of those two slopes.
b. Use `check_model()` to display and then, in a few sentences, evaluate the posterior predictive check, as well as the plots to check linearity, homogeneity of variance, influential points, and normality of residuals. What problems with OLS assumptions, if any, do you see? Does your `fit2` model show meaningfully different results in this regard than `fit1` did?

3 Question 3 (10 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 to create a new model (which we would like you to call `fit3`.)

a. In a sentence or two, discuss whether or not we see the interaction between these variables having a large impact on your (transformed) `cardbill`, again solely using your training data. In particular, specify the improvement in raw R^2 attributable to the interaction term. Then discuss whether the coefficients for the slopes of `abcix` and `ejec_frac` in

model `fit3` are substantially different from those seen in `fit2`, in light of the added coefficients in the model.

b. Now, compare the performance of fits 1, 2 and 3 using a table of performance indices and a plot of those indices for your three models. Which model appears, on this basis, to display the best training sample performance?

4 Question 4 (10 points)

By now you should have created 3 models. Fit these models to the test data we held out earlier, called `lab6_test`. Then, compare these models, using their mean absolute prediction error (MAPE), square root of the mean squared prediction error (RMSPE), validated R-squared, and maximum prediction error in the test data.

> **Note**
>
> Evaluate MAPE, RMSPE, R-squared and maximum prediction error in your test data only **after** backing out of the transformation you made earlier. Show all your code to accomplish this, and annotate it with complete sentences where you feel it's useful. Note also that the code we shared in the slides for Class 21 will definitely be helpful here.

a. Which of your three models (`fit1`, `fit2` or `fit3`) performs best according to each of the four measures listed above? Justify your responses with a table of results, and description of whether a low or high value of each measure is desirable.

b. Based on your results in Questions 3 and 4, which model do you prefer (`fit1`, `fit2`, or `fit3`) overall, and why?

5 Question 5 (10 points)

Write a brief essay (150-250 words is appropriate) which relates what you've done in this Lab to what you learned in your reading of Spiegelhalter's *The Art of Statistics*.

6 Additional Notes and Instructions

6.1 Submitting this Lab

Submit this Lab via Canvas, using the Lab 6 assignment. Be sure to submit both files:

1. Your Quarto file (.qmd).

2. The HTML file you obtain by knitting the Quarto file (.html)

Be sure that your Quarto (and thus HTML) files include the session information as a separate section at the end of the document.

6.2 Grading this Lab

This Lab will be graded by the TAs and then reviewed by Dr. Love. Your grades will be available one week after the Lab deadline.

The maximum score on this Lab is 50 points.

As each Lab passes its deadline (as listed in the Course Calendar), we will:

- post the answer sketch (48 hours after the deadline) and draft grading rubric to our Shared Google Drive, and then
- post grades and any revisions to the grading rubric or answer sketch one week after the deadline to a location we will provide to you.

6.3 Emergencies and Late Policy

We do not grant extensions on Lab deadlines.

- 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.)
  - Labs that are turned in 1-48 hours after the deadline will lose 10 points for late work.
- No extensions to Lab deadlines will be made this semester. Labs turned in more than 48 hours after the deadline will receive no credit, since by then the Lab Sketch will be posted.
- Your lowest lab score (out of Labs 1-6) over the course of the semester will be dropped before we calculate your lab grade.

If you have an emergency that will keep you from submitting the Lab by even the late deadline of Friday at noon, please let Dr. Love know that (as soon as possible) via email and he will consider excusing you from the Lab.

6.4 Lab Regrade Requests

If, after your Lab is graded, you want Dr. Love to review the grading or correct a grading error, please follow the Lab Regrade Request policy posted on our Labs page.

7 Session Information

At the end of your Quarto file, you should run session information, like this.

```
xfun::session_info()
```

```
R version 4.4.1 (2024-06-14 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 22631)

Locale:
  LC_COLLATE=English_United States.utf8
  LC_CTYPE=English_United States.utf8
  LC_MONETARY=English_United States.utf8
  LC_NUMERIC=C
  LC_TIME=English_United States.utf8

Package version:
  abind_1.4-8          askpass_1.2.1        backports_1.5.0
  base64enc_0.1.3      bayestestR_0.15.0    bigD_0.3.0
  bit_4.5.0            bit64_4.5.2          bitops_1.0.9
  blob_1.2.4           boot_1.3-31          broom_1.0.7
  bslib_0.8.0          cachem_1.1.0         callr_3.7.6
  car_3.1-3            carData_3.0-5        cellranger_1.1.0
  cli_3.6.3            clipr_0.8.0          coda_0.19-4.1
  codetools_0.2-20     colorspace_2.1-1     commonmark_1.9.2
  compiler_4.4.1       conflicted_1.2.0     correlation_0.8.6
  cowplot_1.1.3        cpp11_0.5.0          crayon_1.5.3
  curl_6.0.0           data.table_1.16.2    datasets_4.4.1
  datawizard_0.13.0    DBI_1.2.3            dbplyr_2.5.0
  Deriv_4.1.6          digest_0.6.37        doBy_4.6.24
  dplyr_1.1.4          dtplyr_1.3.1         easystats_0.7.3
  effectsize_0.8.9     emmeans_1.10.5       estimability_1.5.1
  evaluate_1.0.1       fansi_1.0.6          farver_2.1.2
  fastmap_1.2.0        fontawesome_0.5.2    forcats_1.0.0
  foreach_1.5.2        Formula_1.2-5        fs_1.6.5
  gargle_1.5.2         generics_0.1.3       ggplot2_3.5.1
  glmnet_4.1-8         glue_1.8.0           googledrive_2.1.1
  googlesheets4_1.1.1  graphics_4.4.1       grDevices_4.4.1
  grid_4.4.1           gridExtra_2.3        gt_0.11.1
  gtable_0.3.6         haven_2.5.4          highr_0.11
  hms_1.1.3            htmltools_0.5.8.1    htmlwidgets_1.6.4
```

httr_1.4.7	ids_1.0.1	insight_0.20.5
isoband_0.2.7	iterators_1.0.14	janitor_2.2.0
jomo_2.7-6	jquerylib_0.1.4	jsonlite_1.8.9
juicyjuice_0.1.0	knitr_1.49	labeling_0.4.3
lattice_0.22-6	lifecycle_1.0.4	lme4_1.1-35.5
lubridate_1.9.3	magrittr_2.0.3	markdown_1.13
MASS_7.3-61	Matrix_1.7-0	MatrixModels_0.5.3
memoise_2.0.1	methods_4.4.1	mgcv_1.9.1
mice_3.16.0	microbenchmark_1.5.0	mime_0.12
minqa_1.2.8	mitml_0.4-5	modelbased_0.8.9
modelr_0.1.11	multcomp_1.4-26	munsell_0.5.1
mvtnorm_1.3-1	naniar_1.1.0	nlme_3.1-164
nloptr_2.1.1	nnet_7.3-19	norm_1.0.11.1
numDeriv_2016.8.1.1	openssl_2.2.2	ordinal_2023.12.4.1
pan_1.9	parallel_4.4.1	parameters_0.23.0
patchwork_1.3.0	pbkrtest_0.5.3	performance_0.12.4
pillar_1.9.0	pkgconfig_2.0.3	plyr_1.8.9
prettyunits_1.2.0	processx_3.8.4	progress_1.2.3
ps_1.8.1	purrr_1.0.2	quantreg_5.99
R6_2.5.1	ragg_1.3.3	rappdirs_0.3.3
RColorBrewer_1.1.3	Rcpp_1.0.13-1	RcppEigen_0.3.4.0.2
reactable_0.4.4	reactR_0.6.1	readr_2.1.5
readxl_1.4.3	rematch_2.0.0	rematch2_2.1.2
report_0.5.9	reprex_2.1.1	rlang_1.1.4
rmarkdown_2.29	rpart_4.1.23	rstudioapi_0.17.1
rvest_1.0.4	sandwich_3.1-1	sass_0.4.9
scales_1.3.0	see_0.9.0	selectr_0.4.2
shape_1.4.6.1	snakecase_0.11.1	SparseM_1.84.2
splines_4.4.1	stats_4.4.1	stringi_1.8.4
stringr_1.5.1	survival_3.7-0	sys_3.4.3
systemfonts_1.1.0	textshaping_0.4.0	TH.data_1.1-2
tibble_3.2.1	tidyr_1.3.1	tidyselect_1.2.1
tidyverse_2.0.0	timechange_0.3.0	tinytex_0.54
tools_4.4.1	tzdb_0.4.0	ucminf_1.2.2
UpSetR_1.4.0	utf8_1.2.4	utils_4.4.1
uuid_1.2.1	V8_6.0.0	vctrs_0.6.5
viridis_0.6.5	viridisLite_0.4.2	visdat_0.6.0
vroom_1.6.5	withr_3.0.2	xfun_0.48
xml2_1.3.6	xtable_1.8-4	yaml_2.3.10
zoo_1.8-12		