# Lab 4 Instructions

**BSTA 512/612**

Nicky Wakim

2024-03-14

> ⚠ IMPORTANT TO READ
>
> - Please do not delete the rubric from your `.qmd` file. I will use it to circle the grades!
> - There is an instructions file and a file for you to edit and turn in. Please only work in the latter file!!

## Directions

Please turn in your `.html` file on Sakai. Please let me know if you greatly prefer to submit a physical copy.

You can download the `.qmd` file for this lab here.

> 🔥 Caution
>
> This is the **instructions** file. The link above will take you to the **editing** file where you can add your work and turn it in!! Please do not remove anything from the editing file!!

### Purpose

The main purpose of this lab is to perform model selection, identify one or more potential final models, and start our interpretation of our main relationship.

### Grading

**This lab is graded out of 12 points.** Nicky will use the following rubric to assign grades.

**Rubric**

| | 4 points | 3 points | 2 points | 1 point | 0 points |
|---|---|---|---|---|---|
| Formatting | Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with no major grammatical nor spelling errors. With little editing, the answer can be incorporated into the project report. | Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with grammatical or spelling errors. With editing, the answer can be incorporated into the project report. | Lab submitted on Sakai with `.html` file. Answers are written in complete sentences with major grammatical or spelling errors. With major editing, the answer can be incorporated into the project report. | Lab submitted on Sakai with `.html` file. Answers are bulletted or do not use complete sentences. | Lab *not* submitted on Sakai with `.html` file. |
| Code/Work | All tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output. | All tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output. In a few tasks, the code syntax or output is not quite right. | Most tasks are directly followed or answered. This includes all the needed code, in code chunks, with the requested output. | Some tasks are directly followed or answered.This includes all the needed code, in code chunks, with the requested output. In a few tasks, the code syntax or output is not quite right. | More than a quarter of the tasks are not completed properly. |

| | 4 points | 3 points | 2 points | 1 point | 0 points |
|---|---|---|---|---|---|
| Reasoning* | Answers demonstrate understanding of research context and investigation of the data. Answers are thoughtful and can be easily integrated into the final report. | Answers demonstrate understanding of research context and investigation of the data. Answers are thoughtful, but lack the clarity needed to easily integrate into the final report. | Answers demonstrate some understanding of research context and investigation of the data. Answers are fairly thoughtful, but lack connection to the research. | Answers demonstrate some understanding of research context and investigation of the data. Answers seem rushed and with minimal thought. | Answers lack understanding of research context and investigation of the data. Answers seem rushed and without thought. |

*Applies to questions with reasoning

## Lab activities

Before starting this lab, you should go back to Lab 2, save a new `.rda` file that contains all the new variables from that Lab. Then you can load it here!

### Restate your research question

> **!** Task
>
> Please restate your research question below using the provided format. It's repetitive, but it helps me contextualize my feedback as I look through your lab.

How is implicit anti-fat bias, as measured by the IAT score, associated with "insert main independent variable here"?

### Step 1: Simple linear regressions / analysis

We have done most of this step through visualizations in Lab 2 and 3. Now, we will quickly run a simple linear regression model for each covariate against the IAT score (outcome). Remember, the goal of this is to see if each covariate explains enough variation of the outcome, IAT score. You should have at least 9 simple linear regression models and their results. Results include the F-statistic and p-value from the test if each covariate explains enough variation of the outcome. Please revisit the slides from Lesson 5 (SLR: More inference + Evaluation) for more help with this test.

⚠ VERY IMPORTANT FOR VARIABLES WE ORDERED USING FACTOR!!

I asked that you order variables to make plots more interpretable. However, for the `lm()`, R reads the ordered variables in an unexpected way. For these variables to run correctly in R, we need to unorder the variables. We can also set a reference level that makes sense.
For example, I may want to unorder my variable `iam_001` and set the reference to `Neither underweight nor overweight`. I can do this with:

```
iat_2021_new = iat_2021_old %>%
  mutate(iam_unordered = factor( iam_ordered, ordered = FALSE ) %>%
           relevel( ref = "Neither underweight nor overweight"))
```

Recall, we mentioned 3 options to running and outputting the results of

1. We can run `lm()` for each covariate *in separate lines of code*, and use something like `summary()` or `anova()` to look at the results of each. (More time consuming to write, but less complicated coding)
2. We can use `lapply()` to run `lm()` and display the `anova()` on each covariate *in one line of code*. (Less time consuming to write, but more complicated coding, and more prone to errors that may not be apparent from output)
3. We can use `sapply()` to run `lm()`, `anova()`, and display the p-value for each covariate *in one line of code*. (Less time consuming to write, but more complicated coding, more prone to errors that may not be apparent from output, and no sense of what's going on in the regression)

Please take a note for yourself if your dataset contains the original numeric versions of variables that we created factors for. I am not saying that you should take them out. They might be useful if our sample is not big enough to handle all the categorical covariates that we've included, but I think our sample is large enough.

**!** Tasks

1. Run a simple linear regression model for each covariate against the IAT score (outcome).
2. Display results from the test if each covariate explains enough variation of the outcome. This may be from three options in the instructions: `summary()/anova()` only, `lapply()`, or `sapply()`

Interpretation of the results will be in the next step.

### Step 2: Preliminary variable selection

Using the previous p-values from the F-test on each covariate's SLR, decide which covariates will be included in the initial model. Recall the decision rule: we keep covariates that explain enough variation using p-value $< 0.25$. Note that because our sample size is so large, the p-values might be really small. For now, that's okay, but this means we may want to alter our Step 3 a little bit.

Once you have decided on the covariates, run the model and display the regression table.

**!** Tasks

1. Decide which covariates will be included in the initial model and list them.
2. Run the initial model and display the regression table.

No need to write out the model, but you may *in addition* to the list.

### Step 3: Assess change in coefficient

Now that all the selected variables are in one initial model, we can start considering the effect of each variable (outside of our main research question).

Remember our general rule: We can remove a variable if (1) p-value $> 0.05$ for the F-test to include or exclude the variable and (2) change in coefficient ($\Delta\%$) of our explanatory variable is $< 10\%$.

Since our sample size is quite large, most of the p-values will be quite small.