

# 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!!

```
load("~/Library/CloudStorage/OneDrive-OregonHealth&ScienceUniversity/Teaching/Classes/W2024/STAT512/qmd/iat_3.qmd")
```

## 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	Lab submitted on Sakai with .html file. Answers are written in complete sentences with major grammatical or spelling errors. With 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 <i>not</i> submitted on Sakai with .html file.

	4 points	3 points	2 points	1 point	0 points
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.
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  $p\text{-value} < 0.25$ . 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%. *Please remember that the p-values for the F-test for a multi-level categorical variable must be calculated by creating a reduced and full model.*

It might be helpful to copy your list of covariates here and make note of the ones that you are removing. It was hard for me to keep track of all the variables when our dataset contains sooo many categorical covariates, and the regression table is so long.

Since our sample size is quite large, most (if not all) of the F-tests will conclude that the variable should be kept in the model. At this point, I advise that you turn to some common sense and the change in coefficients.

1. For **common sense**, you may notice that some of your covariates are essentially measuring the same thing. If there is clinical relevance to having both in the model, then keep them in, but if not, you will have to decide which is more interpretable/relevant/aligned with your research question. For example, if you chose variables involving attitudes and beliefs that are measuring similar things, then you might exclude one. There are measurements like “I am ...” with relative weight groups and “Compared to most...” with relative weight groups. These two might capture a lot of the same information, so we may chose one. (Additionally, this might create issues with multicollinearity, which we will discuss on the last day, so just keep that in mind!) Another example is if you used gender identity, this might be a good time to throw out sex assigned at birth. Remember, my reasoning for using SAB was that (1) lab work has been extensive and I wanted to give you an option to avoid multi-selection variables, and (2) it *might* capture some of the differences around fat attitudes tied with gender. If you included gender identity in your work, then sex assigned at birth could be superfluous.
2. For **change in coefficients**, focus on the variable of your research question. Does the removal of variables change the coefficients for your explanatory variable? Remember what we discussed with change in coefficients when our explanatory variable is a multi-level categorical variable (Lesson 11.2 Interactions continued slides 26-28). You may find these changes small, which tracks with a lot of our plots in Lab 3. Nothing seemed to have such a big effect on IAT score, and as a consequence it's hard to see big changes for a potential confounder.

Note that I put common sense first. The change in coefficients may not be very large, and may lead you to think we don't need a lot of the variables in our model. However, I would let common sense override the change in coefficients if your reasoning is well justified.

psst... There might be some code in Step 4 that might help you get started in this step.

### ! Tasks

Remove variables from the initial model based on your common sense, change in coefficient, and/or p-values of the F-tests.

You do NOT need to show all your work here. You just need to include:

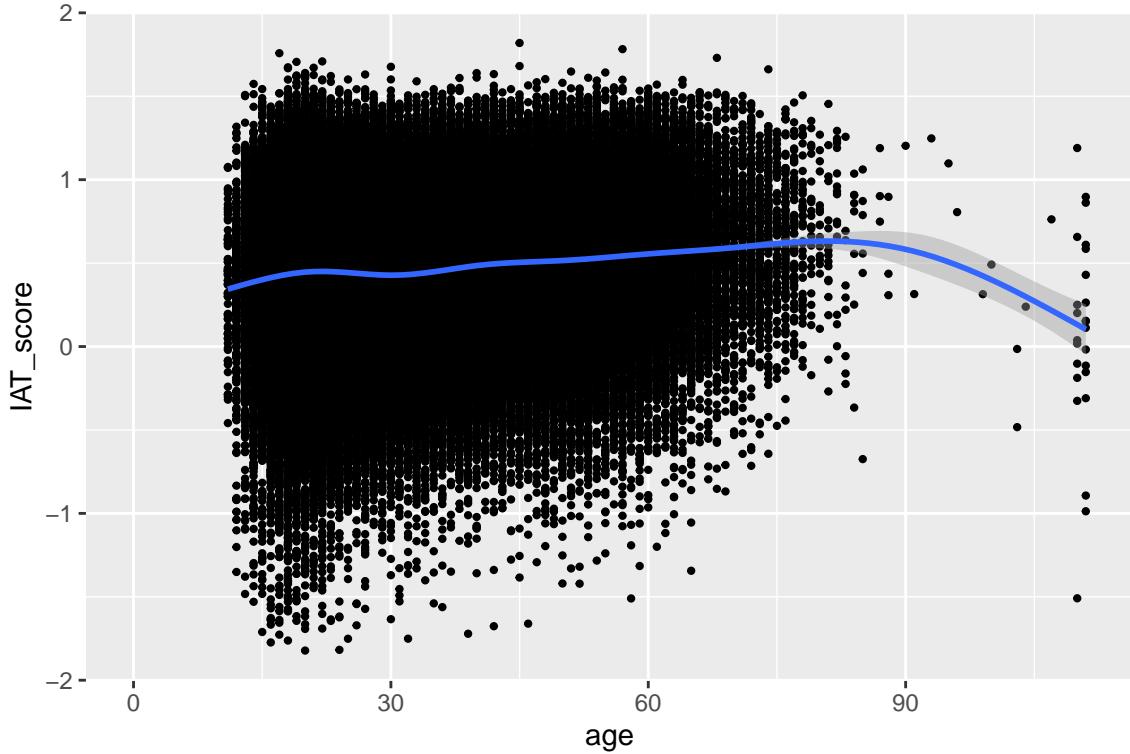
1. A brief explanation of what variables were dropped and why (a sentence per variable), and
2. An example of your process with one variable is enough (including code that you ran)

## Step 4: Assess scale for continuous variables

There is one variable in our model (unless you removed it) that is continuous: age. We need to assess the scale for age. **In this step we will have ZERO delivarables.** To save you time, I will walk you through my thought process, and why I determined age is fine as is. If you still want to try something else out with age, then you can!

First, we can start with a scatterplot of IAT score and age. Your plot may look a little different than mine.

```
ggplot(data = iat, aes(x = age, y = IAT_score)) +  
  geom_point(size = 0.8) + geom_smooth() + xlim(0, 111)  
  
'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



In the above scatterplot, it looks like the relationship is mostly linear (and increasing) until we get to approximately 90 years old. At that point IAT score decreases with age. Let's say we 100% believe there is suddenly more responders around 110 years old than 90-100 year olds. I'm already skeptical of this since we did a quality control in Lab 3. We'll play it out because it's not worth making judgement calls on what we consider "admissible" data.

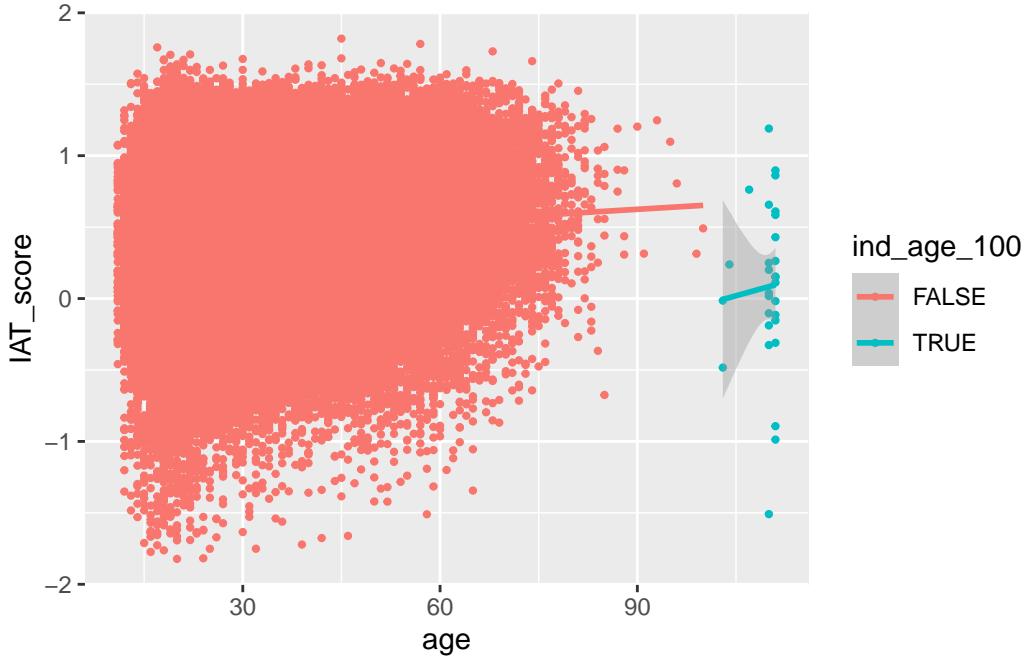
We could do quantiles, splines, or polynomials, but those approaches will either make more categorical variables or make the relationship between age and IAT score harder to interpret. We have a pretty linear relationship up until the higher ages!

I wanted to investigate the linearity a little more so I created an indicator for individuals who are 100 years or older:

```
iat1 = iat %>% mutate(ind_age_100 = ifelse(age > 100, "TRUE", "FALSE"))
```

Now I can see if the linearity differs between the two groups of ages:

```
ggplot(iat1, aes(x = age, y = IAT_score, color = ind_age_100)) +
  geom_point(size = 0.8) + geom_smooth(method = "lm")
```



I am happy to see that both groups' IAT score are increasing with age. It actually looks like my indicator might be a confounder... In that case, we only need to include the indicator in the model so that the relationship between age and IAT is adjusted for the indicator. I can test to see if the indicator is a big enough confounder using the change in coefficient of age and my explanatory variable.

Here's the model without the indicator:

```
prelim_model = lm(IAT_score ~ iam_unordered + identfat + comptomost +
  ind_m + ind_f + ind_tmm + ind_twf + ind_gqnc +
  ind_other +
  race +
  ethn +
  edu_14_f +
  age, data = iat1)
```

And we'll take a look at the coefficients for the model:

```
prelim_model$coefficients[c(2:6, 46)] # by using c(2:6, 46) I am telling R to
```

iam_unorderedVery underweight	iam_unorderedModerately underweight
-0.061151601	-0.015513320

```

iam_unorderedSlightly underweight      iam_unorderedSlightly overweight
                           0.006954580                         -0.023369363
iam_unorderedModerately overweight    age
                           -0.056237038                         0.003857134

# only print certain variables' coefficients

```

Then we can run the model with the indicator, then look at the coefficients:

```

prelim_model2 = lm(IAT_score ~ iam_unordered + identfat + comptomost +
                   ind_m + ind_f + ind_tmm + ind_twf + ind_gqnc + ind_other +
                   race +
                   ethn +
                   edu_14_f +
                   age + ind_age_100,
                   data = iat1)
prelim_model2$coefficients[c(2:6, 46)] # by using c(2:6, 46) I am telling R to

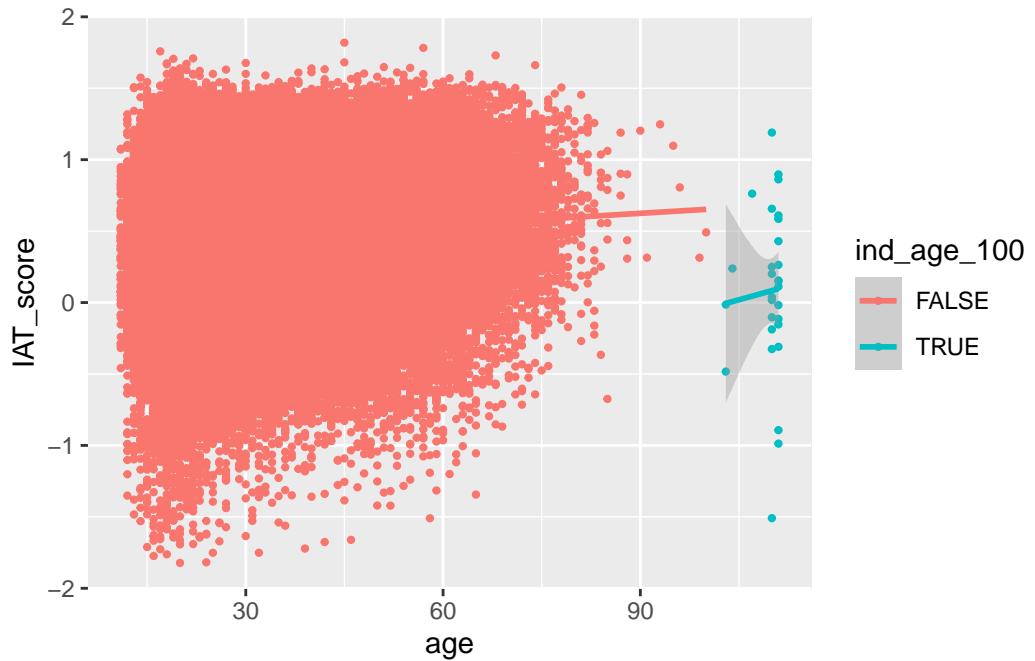
iam_unorderedVery underweight iam_unorderedModerately underweight
                           -0.060385515                         -0.015352043
iam_unorderedSlightly underweight iam_unorderedSlightly overweight
                           0.006817642                         -0.023677757
iam_unorderedModerately overweight age
                           -0.056762301                         0.003918822

# only print certain variables' coefficients

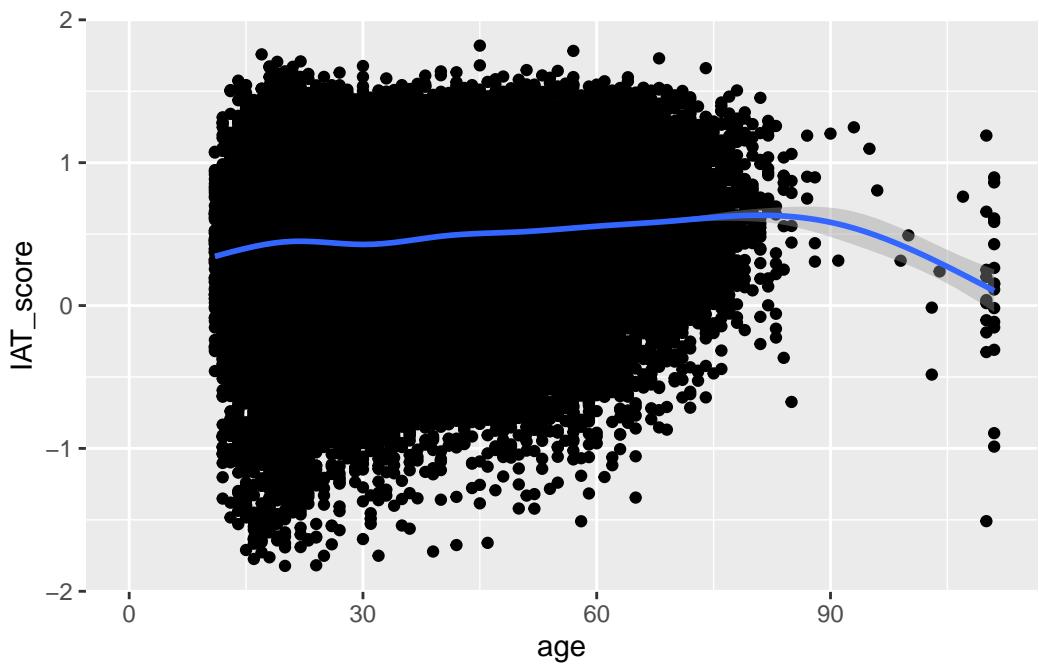
ggplot(iat_3, aes(x = age, y = IAT_score, color = ind_age_100)) +
  geom_point(size = 0.8) + geom_smooth(method = "lm")

`geom_smooth()` using formula = 'y ~ x'

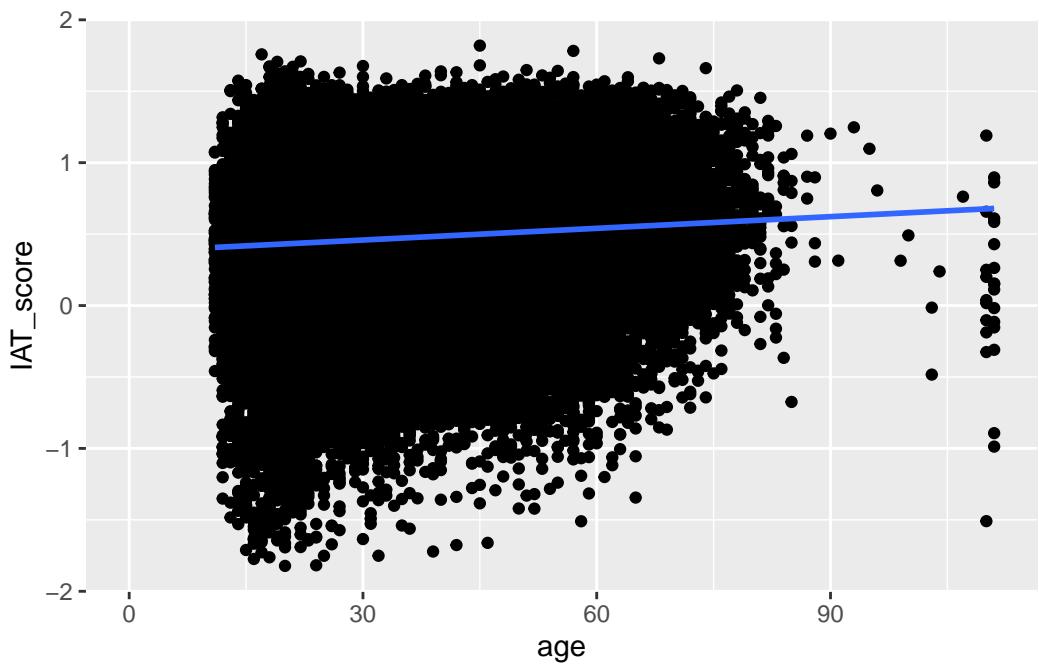
```



```
ggplot(iat_3, aes(x = age, y = IAT_score)) +  
  geom_point() + geom_smooth() + xlim(0, 111)  
  
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
ggplot(iat_3, aes(x = age, y = IAT_score)) +  
  geom_point() + geom_smooth(method = "lm") + xlim(0, 111)  
  
`geom_smooth()` using formula = 'y ~ x'
```



```
ggplot(iat_3 %>% filter(age<100), aes(x = age, y = IAT_score)) +  
  geom_point() + geom_smooth(method = "lm") + xlim(0, 111)  
  
`geom_smooth()` using formula = 'y ~ x'
```

