DS5110 Homework 4

Kylie Ariel Bemis

3 March 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–3 use the U.S. Transgender Population Health Survey (TransPop) originally available from https://www.icpsr.umich.edu/web/ICPSR/studies/37938. Use `load()` to import the saved R environment containing the `da37938.0001` data frame. The original dataset includes samples of both transgender and cisgender individuals (not included). The Codebook and User guide describing the dataset are included in the zipped files.

This survey includes sample weights that can be used to attempt to correct for the effects of sampling bias and unit non-response. These weights can be used to correct sample distributions and sample statistics so that they more accurately represent the population (i.e., transgender people in the USA). For example, using the weighted mean will be more representative of the population than the unweighted sample mean. Likewise, we can calculate weighted counts and weighted proportions by summing the weights rather than calculating the raw sample counts.

Problem 1

We would like to compare the weighted and unweighted distributions of trans people of different races and ethnicities. Visualize bar plots showing (1) the unweighted proportions of trans people of each race/ethnicity and (2) the weighted proportions of trans people of each race/ethnicity. What races are over- or under-represented in the survey sample compared to the population?

Problem 2

We would like to compare the weighted and unweighted distributions of trans people with different sexual orientations. Visualize bar plots showing (1) the unweighted proportions of trans people of each sexual orientation and (2) the weighted proportions of trans people of each sexual orientation. What sexual identities are over- or under-represented in the survey sample compared to the population?

Part B

Problems 3–5 continue to use the U.S. Transgender Population Health Survey (TransPop) dataset from Part A for modeling. There is some debate over if it is appropriate to use survey sample weights in linear regression. For the purposes of the homework, you may omit the sample weights from the modeling in this part.

Problem 3

The survey includes several validated scales for measuring constructs related to identity, stress, and health. We would like to use these scales to build a model for predicting satisfaction with life among trans people. Focus your analysis on the following numeric variables described on pages 26-35 of the User Guide:

- Satisfaction with life
- Social well-being
- Non-affirmation of gender identity
- Non-disclosure of gender identity
- Healthcare stereotype threat
- Mental distress/disorder
- Everyday discrimination

Using the imputed versions of these variables, visualize life satisfaction versus the six candidate predictors, transforming variables as necessary, and describe their relationships.

Hint: It may help to include a fitted line in your visualizations.

Problem 4

Build a linear regression model for life satisfaction using a single predictor, justifying your choice based only on the visualizations from Problem 3. Then use residual plots to perform model diagnostics.

Comment on any outliers or violations of model assumptions you notice in the residual plots. If necessary, fix the issue(s), re-fit the model, and perform model diagnostics again.

Problem 5

Use residual plots to determine if any other candidate predictors should be added to your model from Problem 4. If so, add up to one additional predictor to the model, and then perform model diagnostics on the new model.

Hint: It may help to include a fitted line in your visualizations.