Homework 2

DSC 152: Statistical Inference & Data Analytics using R

This homework covers Lectures 5-10 and Labs 3-4. Please write your solutions in an R Markdown file and submit your PDF output to Gradescope. All code for your solutions should be shown in your PDF file.

Introduction

Suppose that you are a data scientist working with the UCSD Teaching + Learning Commons. The course staff for an introductory probability workshop piloted two review formats:

- `coding`: students completed coding-based review exercises
- `handwritten`: students completed handwritten review exercises

Students were randomly assigned to one of the two formats. Before the workshop, each student took a short pre-test. After the workshop, each student took a post-test on the same 100-point scale.

The file `workshop_scores.tsv` contains one row per student, with the following columns:

- `student_id`
- `group`
- `pretest`
- `posttest`

Throughout the homework, use a significance level of 0.05 unless otherwise stated. For any simulation or permutation procedure, set and report a seed so that your results are reproducible.

Question 1

We first investigate whether the two workshop formats led to different average improvements. Define each student's score gain as:

$$\text{gain} = \text{posttest} - \text{pretest}$$

Let μ_{coding} be the true mean score gain for students assigned to the coding format, and let $\mu_{\text{handwritten}}$ be the true mean score gain for students assigned to the handwritten format. We would like to test:

$$H_0 : \mu_{\text{coding}} = \mu_{\text{handwritten}} \qquad \text{vs.} \qquad H_A : \mu_{\text{coding}} \neq \mu_{\text{handwritten}}$$

(a)

Load `workshop_scores.tsv` into your R Markdown file. Then create a new numeric variable called `gain`.

Report the following sample quantities:

- the mean score gain in the coding group
- the mean score gain in the handwritten group
- the standard deviation of score gain in each group
- the observed difference in sample means, defined as

$$\bar{x}_{\text{coding}} - \bar{x}_{\text{handwritten}}$$

(b)

Create a side-by-side plot comparing the distribution of `gain` in the two groups. You may use either base R graphics or `ggplot2`, but be sure that:

- the axes are clearly labeled
- the title is meaningful
- the text is large enough to read comfortably

Briefly summarize what you observe about center, spread, and any unusual observations.

(c)

Perform a Welch two-sample t-test for the hypotheses stated above.

Report:

- the test statistic

- the p-value
- a 95% confidence interval for the difference in means
- a brief conclusion in context

Your written conclusion should clearly state whether you reject or fail to reject H_0, and what that means about the two workshop formats.

(d)

Compute Cohen's d for the difference in mean score gains between the two groups, using the pooled standard deviation.

Then briefly discuss the following:

- How does the estimated effect size compare with the p-value from Part (c)?
- Why is a statistically significant result not automatically the same thing as a practically important result?

(e)

Write a function called `diff_in_means` that takes a dataframe as input and returns the absolute difference in mean `gain` between the two groups.

Then use that function to carry out a permutation test for the same hypotheses as in Part (c), using at least 5,000 permutations.

Report:

- the observed test statistic
- the permutation p-value
- a brief comparison between the permutation-test result and the t-test result

Also, describe when a permutation test would be advantageous relative to a two-sample t-test.

(f)

Because the workshop format was randomly assigned, explain whether causal language is appropriate here. In particular, address whether it is reasonable to say that one format "caused" a different average gain than the other, and briefly mention any assumptions or caveats that still matter.

Question 2

We now investigate the relationship between students' pre-test and post-test scores using simple linear regression. Let Y denote a student's post-test score and let X denote that student's pre-test score. Consider the model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

(a)

Create a scatterplot of `posttest` versus `pretest`, and add the least-squares regression line to the plot.
As before, make sure your plot has readable labels and a meaningful title.

(b)

Fit the simple linear regression model

```
lm(posttest ~ pretest, data = workshop_df)
```

where `workshop_df` is whatever name you gave your dataframe.
Report:

- $\hat{\beta}_1$
- the standard error of $\hat{\beta}_1$
- the p-value for testing

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0$$

- a one- to two-sentence interpretation of the slope estimate in context

(c)

Obtain a 95% confidence interval for β_1 and interpret it in context.
Your interpretation should be about the population relationship, not just the observed sample.

(d)

Perform a residual analysis for checking each of the required conditions for linear regression, as discussed in Lecture 10. Comment briefly on whether there is any visible evidence of violations of any of these conditions.

Question 3

In Question 1, we compared workshop formats using score gain. We will now revisit the workshop-format question through the lens of multiple regression, adjusting for students' starting levels.

Treat `handwritten` as the reference group and consider the adjusted model:

$$\text{posttest} = \beta_0 + \beta_1 \cdot I(\text{coding}) + \beta_2 \cdot \text{pretest} + \varepsilon$$

where $I(\text{coding})$ equals 1 for students in the coding group and 0 otherwise. If needed, you may explicitly relevel the factor in R so that `handwritten` is the reference category.

(a)

Fit the following two models:

1. An unadjusted model:

   ```
   lm(posttest ~ group, data = workshop_df)
   ```

2. An adjusted model:

   ```
   lm(posttest ~ group + pretest, data = workshop_df)
   ```

For each model, report:

- the estimated coefficient for the coding group
- the p-value associated with that coefficient

(b)

Which of the two models from Part (a) more directly answers the question

> "Among students with similar pre-test scores, do the two workshop formats differ in average post-test score?"

Explain your answer briefly.

(c)

Interpret the coding-group coefficient from the adjusted model in context.

Your answer should make clear what is being held fixed when that coefficient is interpreted.

(d)

Compare the results from the unadjusted and adjusted models. Why might a comparison of raw post-test means be less informative, or potentially misleading in this realized sample, even though the workshop format itself was randomly assigned?

(e)

A student says:

> "If the p-value for `pretest` is bigger than 0.05, then we should drop `pretest` from the adjusted model."

Briefly explain why this is not a good rule, by itself, when the goal is statistical inference rather than just building a predictive model. Also, since it should not be based on its p-value, explain what the rationale for including / not including the `pretest` variable should actually be.

(f)

Perform a residual analysis for checking each of the required conditions for multiple linear regression, as discussed in Lecture 10. Comment briefly on whether there is any visible evidence of violations of any of these conditions.

(g)

Write a short final conclusion, in words, about the effect of workshop format on student performance. Your conclusion should:

- use the model that is adjusted for `pretest`
- state whether the result is statistically significant
- interpret the size and direction of the estimated effect
- explain whether causal language is appropriate