STAT 9610: Homework 1

FirstName LastName

Due September 14, 2024 at 10:00am

1 Instructions

Setup. Clone this repository and open `homework-1.tex` in your LaTeX editor. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. Add R code for problem i in `problem-i.R` (rather than in your LaTeX report), saving your figures and tables to the `figures-and-tables` folder for LaTeX import.

Resources. Consult the getting started guide if you need to brush up on R, LaTeX, or Git, the preparing reports guide for guidelines on presentation quality, the sample homework for an example of a completed homework repository, and this webpage for more detailed instructions on using GitHub and Gradescope to complete and submit homework.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) is required; points will be deducted for using base R.

Grading. Each sub-part of each problem will be worth 3 points: 0 points for no solution or completely wrong solution; 1 point for some progress; 2 points for a mostly correct solution; 3 points for a complete and correct solution modulo small flaws. The presentation quality of the solution for each problem (see the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your LaTeX report to PDF and commit your work. Then, push your work to GitHub. Finally, submit your GitHub repository to Gradescope.

Materials and collaboration. The policy is as stated on the Syllabus:

> "Students may consult all course materials, textbooks, the internet, or AI tools (e.g. ChatGPT or GitHub Copilot) to complete their homework. Students may not use solutions to problems that may be available online and/or from past iterations of the course. For each homework and exam, students must disclose all classmates with whom they collaborated, which AI tools they used, and how they used them. Failure to do so will result in a 5-point penalty. The instructor reserves the right to update this policy during the semester."

In accordance with this policy,

Please disclose all classmates with whom you collaborated:

Please disclose which AI tools you used, and how you used them:

Failure to answer the above questions will result in a 5-point penalty.

Problem 1. Change of basis. (Adapted from Agresti Ex. 1.17)

Let $\boldsymbol{X}$ and $\boldsymbol{X}'$ be full-rank $n \times p$ model matrices.

(a) Show that $C(\boldsymbol{X}) = C(\boldsymbol{X}')$ if and only if $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ for some nonsingular $p \times p$ matrix $\boldsymbol{A}$. In plain language, express what the operation $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{A}$ does to the columns of $\boldsymbol{X}$ (one sentence is sufficient).

(b) Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}'$ be the least squares solutions obtained from regressing a response vector $\boldsymbol{y}$ on $\boldsymbol{X}$ and $\boldsymbol{X}' \equiv \boldsymbol{X}\boldsymbol{A}$, respectively, where $\boldsymbol{A}$ is a nonsingular $p \times p$ matrix. What is the relationship between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}'$ (express the latter in terms of the former)? Justify your answer.

(c) Consider the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon; \quad \epsilon \sim (0, \sigma^2), \tag{1}$$

so that $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}]$ for columns $\boldsymbol{x}_{*j} \equiv (x_{1j}, \ldots, x_{nj})^T$, $j \in \{1, 2\}$. Sometimes it is useful to center the predictors by subtracting their means:

$$\boldsymbol{x}'_{*j} \equiv \boldsymbol{x}_{*j} - \bar{x}_j \mathbf{1}; \quad \bar{x}_j \equiv \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j \in \{1, 2\}.$$

Defining $\boldsymbol{X}' \equiv [\mathbf{1}, \boldsymbol{x}'_{*1}, \boldsymbol{x}'_{*2}]$, find the matrix $\boldsymbol{A}$ such that $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$ ($\boldsymbol{A}$ may itself be expressed in terms of $\boldsymbol{X}$). Express the coefficient estimates from the centered regression ($\widehat{\beta}'_0$, $\widehat{\beta}'_1$, $\widehat{\beta}'_2$) in terms of those from the original regression ($\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$).

(d) Let $w \in \{a, b, c\}$ be a categorical variable with three levels. Define $x_1 \equiv \mathbb{1}(w = b)$ and $x_2 \equiv \mathbb{1}(w = c)$, and consider the linear regression (1). This corresponds to regressing y on the categorical variable w, with baseline category a. Sometimes a different baseline category may make more sense, e.g. category b. In this case, we would define $x'_1 \equiv \mathbb{1}(w = a)$ and $x'_2 \equiv \mathbb{1}(w = c)$. Defining $\boldsymbol{X} \equiv [\mathbf{1}, \boldsymbol{x}_{*1}, \boldsymbol{x}_{*2}]$ and $\boldsymbol{X}' \equiv [\mathbf{1}, \boldsymbol{x}'_{*1}, \boldsymbol{x}'_{*2}]$, find the matrix $\boldsymbol{A}$ such that $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A}$. Express the coefficient estimates from the transformed regression ($\widehat{\beta}'_0$, $\widehat{\beta}'_1$, $\widehat{\beta}'_2$) in terms of those from the original regression ($\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$). What are the interpretations of the original and transformed coefficients, and why do the relationships between these coefficients derived above make sense in terms of these interpretations?

Solution 1.

Problem 2. Predictor correlation. (Adapted from Agresti Ex. 2.9)

Consider the linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon; \quad \epsilon \sim (0, \sigma^2),$$

with observed predictor vectors denoted $\boldsymbol{x}_{*1} \equiv (x_{11}, \ldots, x_{n1})^T$ and $\boldsymbol{x}_{*2} \equiv (x_{12}, \ldots, x_{n2})^T$. (This is the same setup as in Problem 1(c).)

(a) Suppose $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ have sample correlation $\rho \in (-1, 1)$. In terms of ρ, what is the correlation between the estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (which are random variables due to the randomness in ϵ)? Your answer should be as explicit as possible.

(b) To build intuition for the preceding result, consider the extreme case when $\boldsymbol{x}_{*1} = \boldsymbol{x}_{*2}$. In this case, $\rho = 1$ and the regression is not identifiable. For a fixed parameter vector $(\beta_0^0, \beta_1^0, \beta_2^0)$, write down the set $\mathcal{S}$ of parameter vectors $(\beta_0, \beta_1, \beta_2)$ giving the same value of $\mathbb{E}[\boldsymbol{y}]$ as $(\beta_0, \beta_1, \beta_2) = (\beta_0^0, \beta_1^0, \beta_2^0)$. In what sense does the result in part (a) reflect the relationship between β_1 and β_2 for $(\beta_0, \beta_1, \beta_2) \in \mathcal{S}$? (Ignore the fact that the case $\rho = 1$ is not covered in part (a).)

(c) Suppose $z_1, z_2 \overset{\text{i.i.d.}}{\sim} N(0, 1)$, and $x_1 \equiv z_1 + 0.5z_2$ and $x_2 \equiv z_1 - 0.5z_2$. What is the correlation between the random variables x_1 and x_2? Suppose the predictors in each row $\{(x_{i1}, x_{i2})\}_{i=1}^n$ are a sample from this joint distribution. Roughly what do we expect to be the sample correlation between $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$? Fixing $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ at their realizations, roughly what do we expect to be the correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$?

(d) To check the conclusions in part (b), run a numerical simulation with $n = 100$, $\sigma^2 = 1$, $(\beta_0, \beta_1, \beta_2) = (0, 1, 2)$, and $\epsilon \sim N(0, \sigma^2)$. Sample one realization of $\boldsymbol{x}_{*1}$ and $\boldsymbol{x}_{*2}$ from the distribution specified in part (c), generate 250 realizations of the response $\boldsymbol{y}$, and for each realization calculate least squares estimates $\widehat{\boldsymbol{\beta}}$. Summarize the results of your simulation by creating scatter plots of $\boldsymbol{x}_{*2}$ versus $\boldsymbol{x}_{*1}$ and $\widehat{\beta}_2$ versus $\widehat{\beta}_1$, with the title of each plot containing the sample correlations of the data it displays. On the scatter plot of $\widehat{\beta}_2$ versus $\widehat{\beta}_1$, indicate the theoretical expected value of $(\widehat{\beta}_1, \widehat{\beta}_2)$ with a red point. Display these two scatter plots side by side using `plot_grid` from the `cowplot` package. Do the sample correlations match what you predicted in part (c)?

Solution 2.

Problem 3. Data analysis: Anorexia treatment. (Adapted from Agresti Ex. 1.24)

For 72 young girls suffering from anorexia, the `Anorexia.dat` file contains their weights before and after an experimental period (Table 1).

Table 1: The first five rows of the anorexia data.

subj	therapy	before	after
1	b	80.5	82.2
2	b	84.9	85.6
3	b	81.5	81.4
4	b	82.6	81.9
5	b	79.9	76.4

The girls were randomly assigned to receive one of three therapies during this period. A control group (c) received the standard therapy, which was compared to family therapy (f) and cognitive behavioral therapy (b). The goal of the study is to compare the effectiveness of the therapies in increasing the girls' weights.

(a) Prepare the data by (1) removing the `subj` variable, (2) re-coding the factor levels of `therapy` as `behavioral`, `family`, and `control`, (3) renaming `before` and `after` to `weight_before` and `weight_after`, respectively, and (4) adding a variable called `weight_gain` defined as the difference of `weight_after` and `weight_before`. Print the resulting tibble.

(b) Explore the data by (1) making box plots of `weight_gain` as a function of `therapy`, (2) making a scatter plot of `weight_gain` against `weight_before`, coloring points based on `therapy` and (3) creating a table displaying, for each `therapy` group, the mean weight gain, maximum weight gain, and fraction of girls who gained weight (i.e. `weight_gain > 0`). Based on these summaries: What therapy appears overall the most successful and why? How effective does the standard therapy appear to be? What is the greatest weight gain observed in this study? Which girls tended to gain most weight (in the absolute sense), based on their weight before therapy? Why might this be the case?

(c) Run a linear regression of `weight_gain` on `therapy` and print the regression summary (print in `R`, without using `kable`). Identify the base category chosen by `R` and discuss the interpretations of the fitted coefficients. It makes more sense to choose `control` as the base category. Recode the factor levels so that `control` is the first (and therefore will be chosen as the base category), rerun the linear regression, and print the summary again. Do the relationships among the fitted coefficients in these two regressions match what was found in Problem 1d?

(d) Directly compute the between-groups, within-groups, and corrected total sums of squares (without appealing to the `aov` function or equivalent) and verify that the first two add up to the third. What is the ratio of the between-groups sum of squares and the corrected total sum of squares? What is the interpretation of this quantity, and what quantity in the regression summaries printed in part (c) is it equivalent to?

Solution 3.