STAT 9610: Homework 3

Name

Due October 26, 2023 at 10:00am

1 Instructions

Setup. Clone this repository and open `homework-3.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. Heteroskedasticity and correlated errors in the intercept-only model.

Suppose that

$$y_i = \beta_0 + \epsilon_i, \quad \text{where } \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}) \tag{1}$$

for some positive definite $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. The goal of this problem is to investigate the effects of heteroskedasticity and correlated errors on the validity and efficiency of least squares estimation and inference.

(a) (Validity of least squares inference) What is the usual least squares estimate $\widehat{\beta}_0^{\text{LS}}$ for β_0 (from Unit 2)? What is its variance under the model (1)? What is the usual variance estimate $\widehat{\text{Var}}[\widehat{\beta}_0^{\text{LS}}]$ (from Unit 2), and what is this estimator's expectation under (1)? The ratio

$$\tau_1 \equiv \frac{\mathbb{E}[\widehat{\text{Var}}[\widehat{\beta}_0^{\text{LS}}]]}{\text{Var}[\widehat{\beta}_0^{\text{LS}}]} \tag{2}$$

is a measure of the validity of usual least squares inference under (1). Write down an expression for τ_1, and discuss the implications of τ_1 for the Type-I error of the hypothesis test of $H_0 : \beta_0 = 0$ and for the coverage of the confidence interval for β_0.

(b) (Efficiency of least squares estimator) Let's assume $\boldsymbol{\Sigma}$ is known. We could get valid inference based on $\widehat{\beta}_0^{\text{LS}}$ by using the variance formula from part (a). Alternatively, we could use the maximum likelihood estimate $\widehat{\beta}_0^{\text{ML}}$ for β_0. What is the variance of $\widehat{\beta}_0^{\text{ML}}$? The ratio

$$\tau_2 \equiv \frac{\text{Var}[\widehat{\beta}_0^{\text{LS}}]}{\text{Var}[\widehat{\beta}_0^{\text{ML}}]} \tag{3}$$

is a measure of the efficiency of the usual least squares estimator under (1), recalling that the maximum likelihood estimator is most efficient. Write down an expression for τ_2, and discuss the implications of τ_2 for the power of the hypothesis test of $H_0 : \beta_0 = 0$ and for the width of the confidence interval for β_0.

(c) (Special case: Heteroskedasticity) Suppose $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ for some $\sigma_1^2, \ldots, \sigma_n^2 > 0$. Compute the ratios τ_1 and τ_2 defined in equations (2) and (3), respectively. How do these ratios depend on $(\sigma_1^2, \ldots, \sigma_n^2)$, and what are the implications for validity and efficiency?

(d) (Special case: Correlated errors) Suppose $(\epsilon_1, \ldots, \epsilon_n)$ are *equicorrelated*, i.e.

$$\Sigma_{j_1 j_2} = \begin{cases} 1, & \text{if } j_1 = j_2; \\ \rho, & \text{if } j_1 \neq j_2. \end{cases} \tag{4}$$

for some $\rho \geq 0$. Compute the ratios τ_1 and τ_2 defined in equations (2) and (3), respectively. How do these ratios depend on ρ, and what are the implications for validity and efficiency?

Solution 1.

Problem 2. Comparing constructions of heteroskedasticity-robust standard errors.

Suppose that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \epsilon_i \overset{\text{ind}}{\sim} N(0, \sigma_i^2). \tag{5}$$

Two approaches to obtaining heteroskedasticity-robust standard errors are the pairs bootstrap and Huber-White standard errors. The goal of this problem is to compare the coverage and width of confidence intervals obtained from these two approaches.

(a) Write a function called `pairs_bootstrap`, which inputs arguments $\boldsymbol{X}$, $\boldsymbol{y}$, and B and outputs an estimated $p \times p$ covariance matrix $\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]$ based on B resamples of the pairs bootstrap.

(b) Write a function called `huber_white`, which inputs arguments $\boldsymbol{X}$ and $\boldsymbol{y}$ and outputs an estimated $p \times p$ covariance matrix $\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]$ based on the Huber-White formula.

(c) Generate $n = 50$ (x, y) pairs by setting x to be equally-spaced values between 0 and 1 and drawing $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \overset{\text{ind}}{\sim} N(0, 9x_i^2)$, $\beta_0 = 2, \beta_1 = 3$. Create a scatter plot of these points, the least squares line, and three confidence bands: the standard least squares confidence band as well as those resulting from the pairs bootstrap (with $B = 500$) and the Huber-White formula. Comment on the relative widths of these three bands depending on the value of x.

(d) Repeat the experiment from part (c) 100 times to compute the coverage and average width of the three confidence bands for each value of x. Plot these two metrics as a function of x, and comment on the results.

Solution 2.

Problem 3. Case study: Pollution data

In this problem, we will analyze a data set related to pollution (`pollution.tsv`), whose first five rows are shown in Table 1 below. These data contain hourly measurements of nitric oxide (NOx)

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

date	log_nox	wind
373	4.46	0.86
373	4.15	1.02
373	3.83	1.10
373	4.17	1.35
373	4.32	1.20

concentration in ambient air (in parts per billion) next to a highly frequented motorway. The first column is an integer specifying the date the observation was taken (the hour each observation was taken is not available). The second column is the logarithm of the nitric oxide concentration. The third column is the square root of the windspeed in meters/second. The goal is to learn how NOx concentration depends on wind speed.

(a) Create some plots and/or summary statistics to explore the data. Comment on any trends you observe.

(a) Run a linear regression of `log_nox` on `wind`, and produce a set of relevant diagnostic plots. What model misspecification issue(s) appear to be present in these data?

(b) Address the above misspecification issues using one or more of the strategies discussed in Unit 3. Report a set of statistical estimates, confidence intervals, and test results you think you can trust.

(c) Discuss the findings from part (b) in language that a policymaker could comprehend, including any caveats or limitations of the analysis.

Solution 3.