Homework 3

Name

Due Sunday, October 24 at 11:59pm

1 Instructions

Setup. Pull the latest version of this assignment from Github and set your working directory to `stat-961-fall-2021/homework/homework-3`. Consult the getting started guide if you need to brush up on `R`, `LaTeX`, or `Git`.

Collaboration. The collaboration policy is as stated on the Syllabus:

> "Students are permitted to work together on homework assignments, but solutions must be written up and submitted individually. Students must disclose any sources of assistance they received; furthermore, they are prohibited from verbatim copying from any source and from consulting solutions to problems that may be available online and/or from past iterations of the course."

In accordance with this policy,

Please list anyone you discussed this homework with:

Please list what external references you consulted (e.g. articles, books, or websites):

Writeup. Use this document as a starting point for your writeup, adding your solutions between `\begin{sol}` and `\end{sol}`. See the preparing reports guide for guidance on compilation, creation of figures and tables, and presentation quality. Show all the code you wrote to produce your numerical results, and include complete derivations typeset in LaTeX for the mathematical questions.

Programming. The `tidyverse` paradigm for data manipulation (`dplyr`) and plotting (`ggplot2`) are strongly encouraged, but points will not be deducted for using base `R`.

```
library(tidyverse)
```

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 (as exemplified by the guidelines in Section 3 of the preparing reports guide) will be evaluated out of an additional 3 points.

Submission. Compile your writeup to PDF and submit to Gradescope.

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? What is its variance under the model (1)? What is the usual variance estimate $\widehat{\text{Var}}[\widehat{\beta}_0^{\text{LS}}]$, 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 $\epsilon_i \overset{\text{ind}}{\sim} N(0, 9x_i^2)$. 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 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: Advertising data..

In this problem, we will analyze a data set related to advertising spending. It contains the sales of a product (in thousands of units) in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: TV, radio, and newspaper. The goal is to learn about the relationship between these three advertising budgets (predictors) and sales (response).

```
ad_data = read_tsv("../../data/Advertising.tsv")
print(ad_data, n = 5)

## # A tibble: 200 x 4
##      TV radio newspaper sales
##   <dbl> <dbl>     <dbl> <dbl>
## 1 230.   37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 152.   41.3      58.5  18.5
## 5 181.   10.8      58.4  12.9
## # ... with 195 more rows
```

(a) Run a linear regression of `sales` on `TV`, `radio`, and `newspaper`, and produce a set of standard 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.