STAT 9610: Midterm Exam

Name

Release Date: 10/23/22 at 9am; Due Date: 10/24/22 at 9pm

1 Instructions

Setup. Clone this repository and open `midterm-fall-2022.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 2 in `problem-2.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.

Allowed materials. The allowed materials are as stated on the Syllabus:

> Students may consult all course materials, including course textbooks, for all assignments and assessments. Students may also consult the internet (e.g. Stack Overflow) for help with general programming tasks (e.g. how to add a dashed line to a plot). Students may not search the internet for help with specific questions or specific datasets on any homework or exam. In particular, students may not use solutions to problems that may be available online and/or from past iterations of the course.

Collaboration. Students must complete the exam individually.

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.

Problem 1. The F-test as a Wald test.

In Homework 2 Problem 1, we showed that the F-test is essentially a likelihood ratio test. In this problem, we will show that the F-test is also essentially a Wald test (see refresher below). Suppose

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_{n\times n}) \tag{1}$$

and σ^2 is known. For a subset $S \subset \{0, \dots, p-1\}$, we would like to test the null hypothesis

$$H_0 : \boldsymbol{\beta}_S = \boldsymbol{0}. \tag{2}$$

(a) Why does the asymptotic normality (4) imply the asymptotic chi-square null distribution (7) of T? Your argument may be heuristic, and need not involve formal asymptotics.

(b) Write down a Wald-type statistic (6) for the linear regression null hypothesis (2) and derive its finite sample distribution (recall the Fisher information analog derived in Homework 2).

(c) To connect the Wald-type statistic derived in part (b) to the F-test, first reparametrize the linear regression problem (1) so that the Fisher information matrix becomes block diagonal. Then, derive a simple expression for the Wald statistic in this reparameterized regression.

(d) Prove that the expression for the Wald statistic found in part (c) is equal to the F statistic, up to a constant multiple in the numerator and the approximation $\widehat{\sigma}^2 \approx \sigma^2$ in the denominator.

(e) Discuss the relationship between the critical values for the Wald test (8) and the F-test in the context of the relationship found in part (d).

Refresher on the Wald test. In classical likelihood inference, we have observations

$$y_i \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{\theta}}, \quad i = 1, \dots, n \tag{3}$$

from some model parameterized by a vector $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Under regularity conditions, the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ is known to converge to a normal distribution centered at its true value:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{I}(\boldsymbol{\theta})^{-1}), \tag{4}$$

where

$$\boldsymbol{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p_{\boldsymbol{\theta}}(y)\right] \tag{5}$$

is the per-observation Fisher information matrix. Given a subset S of the coordinates of $\boldsymbol{\theta}$, the Wald test of the null hypothesis $H_0 : \boldsymbol{\theta}_S = \boldsymbol{0}$ is based on the Wald statistic

$$T = \widehat{\boldsymbol{\theta}}_S^T([\tfrac{1}{n}\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{S,S})^{-1}\widehat{\boldsymbol{\theta}}_S, \tag{6}$$

where $\widehat{\boldsymbol{\theta}}_S$ denotes the entries of $\widehat{\boldsymbol{\theta}}$ corresponding to the coordinates in S and $[\frac{1}{n}\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{S,S}$ denotes the submatrix of $\frac{1}{n}\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1}$ corresponding to the rows and columns indexed by S. Under H_0, we have

$$T \xrightarrow{d} \chi^2_{|S|}, \tag{7}$$

leading to the Wald test

$$\phi(\boldsymbol{y}) \equiv \mathbb{1}\left(\widehat{\boldsymbol{\theta}}_S^T([\tfrac{1}{n}\boldsymbol{I}(\widehat{\boldsymbol{\theta}})^{-1}]_{S,S})^{-1}\widehat{\boldsymbol{\theta}}_S > \chi^2_{|S|}(1-\alpha)\right). \tag{8}$$

Solution 1.

Problem 2. Does employment differ across economic sectors?

Panel data are datasets where each unit has multiple observations across a period of time. In this problem, we consider a panel dataset of manufacturing firms in the UK (see Table 1) from 1976 to 1984. For each firm, we have the year it was observed (`year`), the sub-sector of the manufacturing industry (`sector`), the logarithm of the number of employees (`emp`), a measure of the average wage in the firm (`wage`), an inflation-adjusted estimate of the company's gross capital stock (`capital`), and an index of value-added output (`output`). The goal of this analysis is to determine whether employment differs across economic sectors when controlling for year, wages, capital, and output.

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

firm	year	sector	emp	wage	capital	output
1	1977	7	5.04	13.15	0.59	95.71
1	1978	7	5.60	12.30	0.63	97.36
1	1979	7	5.01	12.84	0.68	99.61
1	1980	7	4.72	13.80	0.62	100.55
1	1981	7	4.09	14.29	0.51	99.56

(a) How many distinct firms are represented in these data? What is the breakdown of the number of firms by sector? Create a table displaying this information. Additionally, create a plot to visualize the distribution of employment by sector, faceting by year. Comment on any trends you see in this plot.

(b) Use a standard F-test to obtain a p-value for the null hypothesis that mean employment does not vary across sectors, when controlling for year, wages, capital, and output. If this analysis were valid, what would be its conclusion? Why might the analysis not be valid? What are the potential consequences?

(c) We might want to adjust for the fact that each firm is being observed multiple times. Why is it not possible to add firm-level fixed effects to the regression?

(d) At least, we might want to carry out inference robust to error correlations within firms across years. Derive a cluster-robust version of the F-test based on the the ordinary least squares estimate $\widehat{\boldsymbol{\beta}}$, the Liang-Zeger estimate $\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}]$, and the Wald test perspective from Problem 1. [Hint: Use the fact that in general, if $\boldsymbol{Z} \sim N(0, \boldsymbol{\Omega})$, then $\boldsymbol{Z}^T \boldsymbol{\Omega}^{-1} \boldsymbol{Z} \sim \chi^2_{\dim(\boldsymbol{Z})}$; there is no need for maximum likelihood theory or Fisher information matrices here.]

(e) Implement the test you proposed in part (d) in an R function called `robust_anova()`. Your function should take arguments `lm_fit`, `lm_fit_partial`, and `cluster`. The first two should be objects outputted by `lm()` on the full and partial models to be compared, and the third should be a formula object specifying the variable(s) to cluster on (the latter is like the `cluster` argument to `vcovCL()`; see the examples at `?vcovCL()`). Your function may call `vcovCL()`; no need to implement Liang-Zeger standard errors from scratch. Apply your function to get a robust analog of the p-value obtained in part (b). Comment on how the conclusion from the robust analysis compares to that of the standard one. [Hint: Use the command `which(!(names(coef(lm_fit)) %in% names(coef(lm_fit_partial))))` to extract the indices of the variables omitted in the partial model, i.e. the variables in S.]

Solution 2.