Homework 2

Name

Due October 4 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-2`. 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. Likelihood inference in linear regression.

Let's consider the usual linear regression setup. Given a full-rank $n \times p$ model matrix $\boldsymbol{X}$, a coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$, and a noise variance $\sigma^2 > 0$, suppose

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n). \tag{1}$$

The goal of this problem is to connect linear regression inference with classical likelihood-based inference (below is a quick refresher).

(a) For the sake of simplicity, let's start by assuming σ^2 is known. Under the fixed-design model, why does the linear regression model (1) not fit into the classical inferential setup (2)? Write the linear model in as close a form as possible to (2).

(b) Continue assuming that σ^2 is known. Why does the Fisher information (4) not immediately make sense for the linear regression model? Propose and compute an analog to this quantity, and using this quantity exhibit a result analogous to the asymptotic normality (3).

(c) Now assume that neither $\boldsymbol{\beta}$ nor σ^2 is known. Derive the maximum likelihood estimates for $(\boldsymbol{\beta}, \sigma^2)$. How do these compare to the estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ discussed in class?

(d) Continuing to assume that neither $\boldsymbol{\beta}$ nor σ^2 is known, consider the null hypothesis $H_0 : \boldsymbol{\beta}_S = \mathbf{0}$ for some $S \subseteq \{1, \ldots, p\}$. Write this hypothesis in the form (5), and derive the likelihood ratio test for this hypothesis. Discuss the connection of this test with the F-test.

Refresher on likelihood inference. In classical likelihood inference, we have observations

$$y_i \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{\theta}}, \quad i = 1, \ldots, n \tag{2}$$

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

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{I}(\boldsymbol{\theta})^{-1}), \tag{3}$$

where

$$\boldsymbol{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log p_{\boldsymbol{\theta}}(y)\right] \tag{4}$$

is the Fisher information matrix. Furthermore, an optimal test of the null hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{5}$$

for some $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ is the likelihood ratio test based on the test statistic

$$\Lambda = \frac{\max_{\boldsymbol{\theta} \in \Theta_1} \prod_{i=1}^n p_{\boldsymbol{\theta}}(y_i)}{\max_{\boldsymbol{\theta} \in \Theta_0} \prod_{i=1}^n p_{\boldsymbol{\theta}}(y_i)}. \tag{6}$$

Under H_0, we have the convergence

$$2 \log \Lambda \xrightarrow{d} \chi^2_k, \quad \text{where} \quad k \equiv \dim(\Theta_1) - \dim(\Theta_0). \tag{7}$$

Solution 1.

Problem 2. Null distribution of R^2.

Consider the linear regression model (1), such that $\boldsymbol{x}_{*,0} = \mathbf{1}_n$ is an intercept term (note that there are only $p-1$ other predictors, for a total of p).

(a) Relate the R^2 of this linear regression to the F-statistic for a certain hypothesis test. What is the corresponding null hypothesis? What is the null distribution of the F-statistic? Are R^2 and F positively or negative related, and why does this make sense?

(b) Use the relationship found in part (a) to simulate the null distribution of the R^2 by repeatedly sampling from an F distribution (via `rf`). Fix $n = 100$ and try $p \in \{2, 25, 50, 75, 99\}$. Comment on these null distributions, how they change as a function of p, and why.

Solution 2.

Problem 3. Relationships between t- and F-tests.

Consider the setup (1) from Problem 1.

(a) Consider the null hypothesis $H_0 : \beta_j = 0$, which can be tested using either a t-test or an F-test. Write down the corresponding t and F statistics, and prove that the latter is the square of the former.

(b) Now suppose that $\boldsymbol{x}_{*,0} = \mathbf{1}_n$ is an intercept term, and we are interested in testing the null hypothesis $H_0 : \boldsymbol{\beta}_{-0} = \mathbf{0}$. One way of going about this is to start with the usual test statistic $t(\boldsymbol{c})$ for the null hypothesis $H_0 : \boldsymbol{c}^T\boldsymbol{\beta}_{-0} = 0$, and then maximize over all $\boldsymbol{c} \in \mathbb{R}^{p-1}$:

$$t_{\max} \equiv \max_{\boldsymbol{c} \in \mathbb{R}^{p-1}} t(\boldsymbol{c}). \tag{8}$$

What is the null distribution of $t^2_{\max}$? What F-statistic is $t^2_{\max}$ equivalent to? How does the null distribution of $t^2_{\max}$ compare to that of $t(\boldsymbol{c})^2$?

Solution 3.

Problem 4. Case study: Violent crime.
The `Statewide_crime.dat` file under `stat-961-fall-2021/data` contains information on the number of violent crimes and murders for each U.S. state in a given year, as well as three socioeconomic indicators: percent living in metropolitan areas, high school graduation rate, and poverty rate.

```
crime_data = read_tsv("../../data/Statewide_crime.dat")
print(crime_data, n = 5)

## # A tibble: 51 x 6
##   STATE Violent Murder Metro HighSchool Poverty
##   <chr>   <dbl>  <dbl> <dbl>      <dbl>   <dbl>
## 1 AK        593      6  65.6       90.2     8
## 2 AL        430      7  55.4       82.4    13.7
## 3 AR        456      6  52.5       79.2    12.1
## 4 AZ        513      8  88.2       84.4    11.9
## 5 CA        579      7  94.4       81.3    10.5
## # ... with 46 more rows
```

The goal of this problem is to study the relationship between the three socioeconomic indicators and the per capita violent crime rate.

(a) These data contain the total number of violent crimes per state, but it is more meaningful to model violent crime rate per capita. To this end, go online to find a table of current populations for each state. Augment `crime_data` with a new variable called `Pop` with this population information (see `dplyr::left_join`) and create a new variable called `CrimeRate` defined as `CrimeRate = Violent/Pop` (see `dplyr::mutate`).

(b) Explore the variation and covariation among the variables `CrimeRate`, `Metro`, `HighSchool`, `Poverty` with the help of visualizations and summary statistics.

(c) Construct linear model based hypothesis tests and confidence intervals associated with the relationship between `CrimeRate` and the three socioeconomic variables, printing and/or plotting your results. Discuss the results in technical terms.

(d) Discuss your interpretation of the results from part (c) in language that a policymaker could comprehend, including any caveats or limitations of the analysis. Comment on what other data you might want to gather for a more sophisticated analysis of violent crime.

Solution 4.