Homework 2

Anirban Chatterjee

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 $H_0 : \boldsymbol{\beta}_S = \boldsymbol{0}$ for some $S \subseteq \{1, \dots, p\}$. 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

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

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.

(a) The linear regression in (1) is equivalent to,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{8}$$

where $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^T$, $\boldsymbol{x}_i, 1 \leq i \leq n$ are the rows of $\boldsymbol{X}$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$. Under the fixed design set up, (8) is further equivalent to,

$$y_i \sim \mathrm{N}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2\right), \quad 1 \leq i \leq n \tag{9}$$

where $y_i, 1 \leq i \leq n$ are independent. Define

$$p_{\boldsymbol{\beta},i} \equiv \mathrm{N}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2\right), \quad 1 \leq i \leq n \tag{10}$$

Then combining (9) and (10) we can see that the linear regression model (1) does not fit into the classical inferential set up (2) since it does not have the i.i.d. assumption, instead each y_i are drawn independently from the distribution $p_{\boldsymbol{\beta},i}$ which (possibly) differs based upon i. The closest form that we can get is,

$$y_i \overset{\texttt{ind}}{\sim} p_{\boldsymbol{\beta},i}, \quad 1 \leq i \leq n$$

where `ind` implies $\{y_i : 1 \leq i \leq n\}$ are generated independently.

(b) The Fisher Information Matrix given in (4) is defined for a single random variable y coming from an distribution $p_{\boldsymbol{\theta}}$ index by a (multidimensional) parameter $\boldsymbol{\theta}$. The definition works well in the classical inferential set up because of the presence of i.i.d. sample, whereby the expectation does not differ from sample to sample. But in the regression set up of (1), this defintion will not make sense because each sample comes from (possibly) different distributions. As a result we consider the follwing,

$$\widetilde{\boldsymbol{I}}(\boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} p_{\boldsymbol{\beta},i}\left(y_i\right) \right] \tag{11}$$

where $p_{\boldsymbol{\beta},i}$ is defined in (10). It is easy to observe that under i.i.d. set up (11) reduces to the usual Fisher Information from (4). Recalling the definition from (10) it can be easily seen that,

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} p_{\boldsymbol{\beta},i}(y_i) = \frac{1}{\sigma^2} \boldsymbol{x}_{\star i} \boldsymbol{x}_{\star i}^T, \quad 1 \leq i \leq p$$

where $\boldsymbol{x}_{\star i}$ denotes the i^{th} column of $\boldsymbol{X}$ for all $1 \leq i \leq p$. Therefore,

$$\widetilde{I}\left(\boldsymbol{\beta}\right) = \frac{1}{n\sigma^2} \boldsymbol{X}^T \boldsymbol{X}$$

Recall from Unit 1,

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right) \equiv \mathrm{N}\left(\boldsymbol{\beta}, \frac{1}{n} \widetilde{I}^{-1}(\boldsymbol{\beta})\right)$$

Implying,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \mathrm{N}\left(\mathbf{0}, \widetilde{\boldsymbol{I}}^{-1}\left(\boldsymbol{\beta}\right)\right)$$

which is in the same flavor as (3).
Although the above statement is similar to (3), it does not talk about asymptotics. Consider the following reformulation

$$\sqrt{n}\widetilde{\boldsymbol{I}}\left(\boldsymbol{\beta}\right)^{-1/2}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\sim \mathrm{N}(0,\boldsymbol{I}).$$

This statement is valid for all $n\in\mathbb{N}$, in particular for $n\to\infty$.

(c) Under the given model the likelihood is given by,

$$L\left(\boldsymbol{\beta},\sigma^2\middle|\boldsymbol{y},\boldsymbol{X}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

where $\boldsymbol{x}_i, 1\le i\le n$ are the rows of $\boldsymbol{X}$. Then the log-likelihood is given by,

$$l\left(\boldsymbol{\beta},\sigma^2\middle|\boldsymbol{y},\boldsymbol{X}\right)=\log L\left(\boldsymbol{\beta},\sigma^2\middle|\boldsymbol{y},\boldsymbol{X}\right)=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)$$

Taking derivatives with respect to $\boldsymbol{\beta}$ and σ^2 and equating to 0 we have,

$$\begin{aligned}\frac{\partial l}{\partial\boldsymbol{\beta}}&=\boldsymbol{X}^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)=0\\ \frac{\partial l}{\partial\sigma^2}&=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)=0\end{aligned}\tag{12}$$

Solving the system of equations in (12) we find the solutions to be,

$$\widehat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y},\quad \widehat{\sigma}^2=\frac{1}{n}\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right)^2$$

where $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^T$ and $\boldsymbol{x}_i, 1\le i\le n$ are the rows of the matrix $\boldsymbol{X}$. Notice that $\widehat{\boldsymbol{\beta}}$ equals the estimate derived in class, where as $\widehat{\sigma}^2$ is smaller then the corresponding estimate derived in class, and as a result $\widehat{\sigma}^2$ is a biased estimate of σ^2.

(d) The given hypothesis can be reformulated in the form of (5) as follows,

$$\boldsymbol{H}_0:\left(\boldsymbol{\beta},\sigma^2\right)\in\Theta_0 \text{ vs } \boldsymbol{H}_1:\left(\boldsymbol{\beta},\sigma^2\right)\in\Theta_1$$

where $\Theta_0=\{\boldsymbol{\beta}\in\mathbb{R}^p:\boldsymbol{\beta}_S=0\}\times\mathbb{R}^+$ and $\Theta_1=\mathbb{R}^p\times\mathbb{R}^+$. Then the likelihood ratio test statistic is given by,

$$\begin{aligned}\Lambda&=\frac{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta_1}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2\right]}{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta_0}\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2\right]}\\ &=\frac{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta_1}\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right]}{\max_{(\boldsymbol{\beta},\sigma^2)\in\Theta_0}\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right)\right]}\end{aligned}\tag{13}$$

By part (c) we know that for the numerator the optimal values of $\boldsymbol{\beta}$ and σ^2 are given by,

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \quad \widehat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \tag{14}$$

Considering $\boldsymbol{H}$ to be the projection matrix onto $C(\boldsymbol{X})$ we have,

$$\widehat{\sigma}^2 = \frac{1}{n}\boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} = \frac{1}{n}\boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}$$

Using the optimal values found in (14) the numerator from (13) becomes,

$$\frac{1}{\left(2\pi\widehat{\sigma}^2\right)^{n/2}}\exp(-n/2)$$

Now observe that the denominator is equivalent to,

$$\max_{(\boldsymbol{\beta}_{-S},\sigma^2)\in\mathbb{R}^{p-|S|}\times\mathbb{R}^+}\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}_{-S}\boldsymbol{\beta}_{-S}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}_{-\boldsymbol{S}}\boldsymbol{\beta}_{-S}\right)\right] \tag{15}$$

where $\boldsymbol{X}_{-S}$ is the matrix that we get by removing the columns S from the matrix $\boldsymbol{X}$ and $\boldsymbol{\beta}_S$ is the vector that we get by removing the coordinates S from the vector $\boldsymbol{\beta}$ (This equivalency is because by definition of Θ_0, $\boldsymbol{\beta}_S = \boldsymbol{0}$ and thus $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_{-S}\boldsymbol{\beta}_{-S}$). Similar computations as in part (c) gives the following optimal values for (15),

$$\widehat{\boldsymbol{\beta}}_{-S} = \left(\boldsymbol{X}_{-S}^T\boldsymbol{X}_{-S}\right)^{-1}\boldsymbol{X}_{-S}^T\boldsymbol{y}, \quad \widehat{\sigma}_0^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}\right) \tag{16}$$

Using the equivalency of the denominator and (15) and the optimal values obtained in (16) the denominator from (13) becomes,

$$\frac{1}{\left(2\pi\widehat{\sigma}_0^2\right)^{n/2}}\exp(-n/2)$$

Implying,

$$\Lambda = \left(\frac{\widehat{\sigma}_0}{\widehat{\sigma}}\right)^{n/2} = \left(\frac{\|\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}\|^2}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}\right)^{n/2} \tag{17}$$

Recall the F-statistic is given by,

$$F = \frac{\left(\|\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\right)/|S|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)} \tag{18}$$

Combining (17) and (18) we conclude that,

$$\Lambda = \left(1 + \frac{|S|}{n-p}F\right)^{n/2}$$

It can be easily seen that Λ is an increasing function of F and thus the likelihood ratio test is equivalent to the F test.

Problem 2. Relationships among t-tests, F-tests, and 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 the 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.

(c) 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.

(d) Now suppose 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 $\boldsymbol{c} \in \mathbb{R}^{p-1}$:

$$t_{\max} \equiv \max_{\boldsymbol{c}\in\mathbb{R}^{p-1}} t(\boldsymbol{c}). \tag{19}$$

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 2.

(a) For the linear regression in (1) recall the R^2 is given by,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}_n\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2} \tag{20}$$

Consider testing for any significant coefficients except the intercept. Then the null hypothesis is given by $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ and the corresponding F-statistic is given by,

$$F \equiv \frac{\left(\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2\right)/(p-1)}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)} \tag{21}$$

Recall the ANOVA decomposition of the variation in $\boldsymbol{y}$,

$$\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2 = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \bar{y}\mathbf{1}_n\|^2 \tag{22}$$

Combining (20),(21) and (22) it is easy to see that,

$$R^2 = \frac{F}{F + \frac{n-p}{p-1}} \tag{23}$$

Under H_0 the F-statistic defined in (21) has a $F_{p-1,n-p}$ distribution. From the relation (23) it can be easily seen that F and R^2 are positively related (this follows since $x/(x+c)$ is an increasing function of x when $c > 0$).
This relation does make sense. Notice that large values of F-statistic gives enough evidence to reject the null hypothesis, thereby showing that other $p-1$ covariates have significant contribution towards explaining the variation of y, which is exactly what large values of R^2 indicates.

(b) Recall that under null $F \sim F_{p-1,n-p}$. For the given values of $p \in \{2, 25, 50, 75, 99\}$ the chunk below simulates the null distribution of R^2 using the relation (23).

```
# Setting seed for reproducibility
set.seed(961)

# Choosing n and number of covariates (p)
n <- 100; p <- c(2,25,50,75,99)

# Function for generating R^2 from F
gen.R2 <- function(x,n,p){return (x/(x+(n-p)/(p-1)))}

# Sampling R^2
R2.samp <- c()
for (i in 1:length(p)){
  F.samp <- rf(n,p[i]-1,n-p[i])
  R2.samp <- cbind(R2.samp,gen.R2(F.samp,n,p[i]))
}
R2.samp <- as.data.frame(R2.samp)
colnames(R2.samp) <- c("p = 2", "p = 25","p = 50", "p = 75", "p = 99")

# Histograms of R^2 for each value of p
R2.hist <- reshape2::melt(as.data.frame(R2.samp),id.vars = NULL) %>%
  mutate(variable = factor(variable)) %>%
  ggplot(aes(x=value, fill=variable)) +
      geom_histogram(binwidth=0.03)+
      labs(x = expression(R^2))+
      facet_grid(variable~.)

# Saving the histogram
ggsave(R2.hist, file = paste0("R2hist.png"), width = 5, height = 5)
```

From Fig 1 it is clear that the distribution of R^2 shifts towards 1 as p increases to n. Notice that as p increases to n we are producing better and better fits of the data (Think about the extreme case $p = n$, assuming the covariates are linearly independent we can have exact fit for the data, and in this case we would have explained all the variation in the data using our model). As a result, by definition, increasing p would increase the value of R^2.

(c) The t-statistics for the test is given by,

$$t_j = \frac{\widehat{\beta}_j}{\sqrt{\frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2 / s_j^2}} \tag{24}$$

where $\widehat{\beta}_j$ is the least square estimate of the coefficient β_j from the regression problem in (1), $s_j^2 = \left[\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}_{jj}\right]^{-1}$ and $\widehat{\boldsymbol{\epsilon}}$ is the residual vector. The F-statistics is given by,

$$F = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,-j}\widehat{\boldsymbol{\beta}}_{-j}\|^2}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,-j}\widehat{\boldsymbol{\beta}}_{-j}\|^2}{\frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2} \tag{25}$$

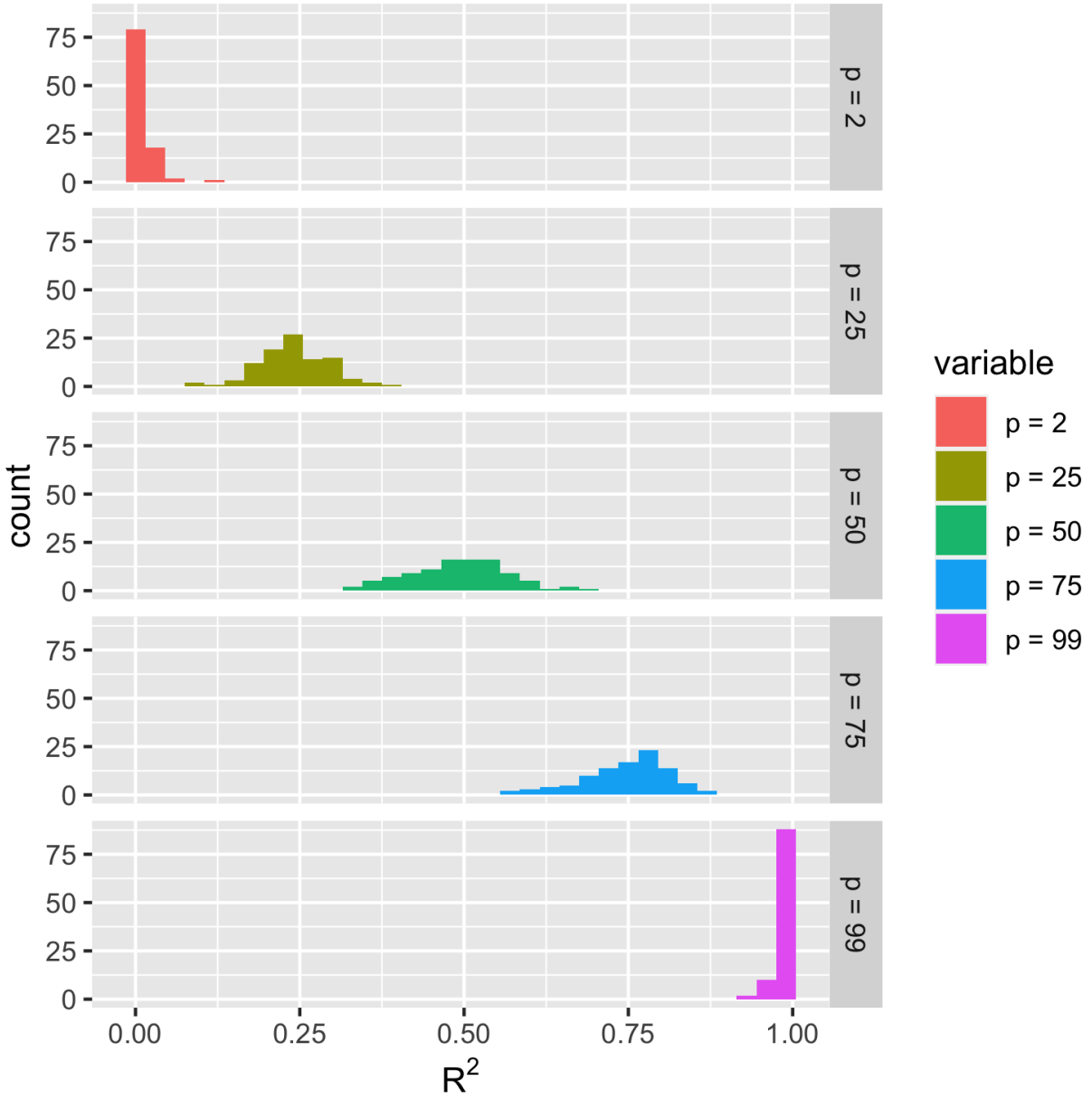

Figure 1: Histogram of R^2 for $p \in \{2, 25, 50, 75, 99\}$.

where $\widehat{\boldsymbol{\beta}}_{-j}$ is the least squares coefficients coming from the partial model,

$$\boldsymbol{y} = \boldsymbol{X}_{*,-j}\boldsymbol{\beta}_{-j} + \epsilon$$

From Unit 1 recall that the regression problem (1) can be rewritten as,

$$\boldsymbol{y} = \boldsymbol{x}_{*j}^{\perp}\beta_j + \boldsymbol{X}_{*,-j}\boldsymbol{\beta}'_{-j} + \epsilon$$

where $\boldsymbol{x}_{*j}^{\perp}$ is the residual from regressing $\boldsymbol{x}_{*j}$ on $\boldsymbol{X}_{*,-j}$ or in other words it is the projection of $\boldsymbol{x}_{*j}$ onto $C(\boldsymbol{X}_{*,-j})^{\perp}$ (Note that the coefficient β_j remains unchanged in the two models). Using orthogonality, the least squares estimates of β_j and $\boldsymbol{\beta}'_{-j}$ can be found by regressing $\boldsymbol{y}$

on $\boldsymbol{x}_{*j}^{\perp}$ and $\boldsymbol{X}_{*,-j}$ separately. Then by definition we must have,

$$\widehat{\boldsymbol{\beta}}_{-j} = \widehat{\boldsymbol{\beta}}'_{-j} \tag{26}$$

Consider the permutation matrix $\boldsymbol{P}$ with rows $\boldsymbol{p}_i^T, 1 \leq i \leq p$ defined by,

$$\boldsymbol{p}_i = \begin{cases} \boldsymbol{e}_{i+1} & 1 \leq i \leq j-1 \\ \boldsymbol{e}_1 & i = j \\ \boldsymbol{e}_i & j+1 \leq i \leq p \end{cases}$$

where $\boldsymbol{e}_k$ denotes the $p \times 1$ vector with 1 at position k and 0 at all other positions, for all $1 \leq k \leq p$. Define $\widehat{\boldsymbol{\gamma}}$ to be the least squares coefficient from regressing $\boldsymbol{x}_{*j}$ on $\boldsymbol{X}_{*,-j}$. Then it is easy to observe that,

$$\boldsymbol{X}\left(P\begin{bmatrix} 1 & \boldsymbol{0}^T \\ -\widehat{\boldsymbol{\gamma}} & \boldsymbol{I}_{p-1} \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{x}_{*j}^{\perp} & \boldsymbol{X}_{*,-j} \end{bmatrix} =: \boldsymbol{X}'$$

Thus from Question 1 of Homework 1 we conclude that $C(\boldsymbol{X}) = C(\boldsymbol{X}')$. Thus considering the projection of $\boldsymbol{y}$ on $C(\boldsymbol{X})$ and $C(\boldsymbol{X}')$ and using (26) we must have,

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{x}_{*j}^{\perp}\widehat{\beta}_j + \boldsymbol{X}_{*,-j}\widehat{\boldsymbol{\beta}}_{-j}$$

Then the F-statistics (from (25)) becomes,

$$F = \frac{\widehat{\beta}_j^2 \|\boldsymbol{x}_{*j}^{\perp}\|^2}{\frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2} \tag{27}$$

Recalling the variance of $\widehat{\beta}_j$ (from unit 1), it can be seen that $s_j^2 = \|\boldsymbol{x}_{*j}^{\perp}\|^2$. Then from (24) and (27) we can conclude that,

$$t_j^2 = F$$

(d) Observe that the hypothesis $\boldsymbol{H}_0 : \boldsymbol{c}^T\boldsymbol{\beta}_{-0} = 0$ is equivalent to the hypothesis $\boldsymbol{H}_0 : \widetilde{\boldsymbol{c}}^T\boldsymbol{\beta} = 0$, where $\widetilde{\boldsymbol{c}}^T = (0, \boldsymbol{c}^T)$. Then the usual test statistic $t(\boldsymbol{c})$ is given by,

$$t(\boldsymbol{c}) = \frac{\widetilde{\boldsymbol{c}}^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}\sqrt{\widetilde{\boldsymbol{c}}^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\widetilde{\boldsymbol{c}}}} \tag{28}$$

Consider $\boldsymbol{e}_i, 1 \leq i \leq p$ to be the standard basis vectors of $\mathbb{R}^p$ (also defined above in part (c)). Consider an $p \times p$ matrix $\boldsymbol{A} = [\boldsymbol{a}_1, \cdots, \boldsymbol{a}_p]$ where

$$\boldsymbol{a}_j = \begin{cases} \boldsymbol{e}_1 & \text{if } j = 1 \\ \boldsymbol{e}_j - \bar{x}_{j-1}\boldsymbol{e}_1 & \text{if } 2 \leq j \leq p \end{cases}$$

and $\bar{x}_j$ denotes the mean of column j of $\boldsymbol{X}$ for all $1 \leq j \leq p-1$. Then

$$\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{A} = [\boldsymbol{1}_n, \boldsymbol{x}_{\star 1} - \bar{x}_1\boldsymbol{1}_n, \cdots, \boldsymbol{x}_{\star p-1} - \bar{x}_{p-1}\boldsymbol{1}_n]$$

where $\boldsymbol{x}_{\star j}$ denotes the j^{th} column of $\boldsymbol{X}$ for all $1 \leq j \leq p-1$. It is easy to observe that

$$\boldsymbol{A}^{-1} = \begin{cases} \boldsymbol{e}_1 & \text{if } j = 1 \\ \boldsymbol{e}_j + \bar{x}_{j-1}\boldsymbol{e}_1 & \text{if } 2 \leq j \leq p \end{cases}$$

is the inverse of $\boldsymbol{A}$ (The proof was done for $p = 3$ in HW 1, same proof would follow for the general case). Next consider the following linear model,

$$\boldsymbol{y} = \boldsymbol{X}'\beta' + \boldsymbol{\epsilon} \tag{29}$$

By HW1 Problem 1(b) we know that $\widehat{\beta}' = \boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}$, implying,

$$\widehat{\boldsymbol{\beta}}'_{-0} = \widehat{\boldsymbol{\beta}}_{-0}$$

Recalling the expression (28) we have,

$$t(\boldsymbol{c}) = \frac{\widetilde{\boldsymbol{c}}^T \widehat{\beta}'}{\widehat{\sigma}\sqrt{\widetilde{\boldsymbol{c}}^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \widetilde{\boldsymbol{c}}}} \tag{30}$$

Using inversion of block matrices it can be easily seen that,

$$\widetilde{\boldsymbol{c}}^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \widetilde{\boldsymbol{c}} = \boldsymbol{c}^T \left(\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T\right)^{-1} \boldsymbol{c} \tag{31}$$

where $\bar{\boldsymbol{x}} = (\bar{x}_1, \cdots, \bar{x}_{p-1})^T$. Define,

$$\boldsymbol{Z} = \boldsymbol{X}_{-0} - \boldsymbol{1}_n\bar{\boldsymbol{x}}^T$$

Observe that $C(\boldsymbol{Z})$ is orthogonal to $\boldsymbol{1}_n$. By definition $\boldsymbol{X}' = [\boldsymbol{1}_n, \boldsymbol{Z}]$ and $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0} - n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T$. Recall the model (29), then orthogonality of $\boldsymbol{1}_n$ and $C(\boldsymbol{Z})$ gives,

$$\widehat{\boldsymbol{\beta}}'_{-0} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{y} \tag{32}$$

Combining (30),(31) and (32), $t(\boldsymbol{c})$ becomes,

$$t(\boldsymbol{c}) = \frac{\boldsymbol{c}^T \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{y}}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{c}}}$$

Using Cauchy Schwarz Inequality we have,

$$|t(\boldsymbol{c})| = \frac{\left|\left(\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2}\boldsymbol{c}\right)^T \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2} \boldsymbol{Z}^T\boldsymbol{y}\right|}{\widehat{\sigma}\left\|\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2}\boldsymbol{c}\right\|_2} \leq \frac{1}{\widehat{\sigma}}\left\|\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2} \boldsymbol{Z}^T\boldsymbol{y}\right\|_2$$

and equality would follow if and only if

$$t\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2} \boldsymbol{Z}^T\boldsymbol{y} = \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2} \boldsymbol{c} \text{ for some } t \in \mathbb{R}\setminus\{0\} \implies \boldsymbol{c} = t\boldsymbol{Z}^T\boldsymbol{y} \text{ for some } t \in \mathbb{R}\setminus\{0\}.$$

Then the maximum is attained and hence,

$$t^2_{\max} = \frac{\left\|\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1/2} \boldsymbol{Z}^T\boldsymbol{y}\right\|}{\widehat{\sigma}^2} = \frac{\boldsymbol{y}^T\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^T\boldsymbol{y}}{\widehat{\sigma}^2} = \frac{\boldsymbol{y}^T\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}}{\widehat{\sigma}^2}$$

where $\boldsymbol{P_Z}$ is the projection matrix onto $C(\boldsymbol{Z})$. Observe that under the null ($\boldsymbol{H}_0 : \boldsymbol{\beta}_{-0} = \boldsymbol{0}$), $\boldsymbol{X\beta} = \boldsymbol{1}_n$. Then orthogonality of $\boldsymbol{1}_n$ and $C(\boldsymbol{Z})$ gives

$$\boldsymbol{P_Z y} \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{P_Z}\right).$$

Using Lemma 1.1 of Unit 2 along with $\boldsymbol{P_Z^2} = \boldsymbol{P_Z}$ and $\mathrm{trace}(\boldsymbol{P_Z}) = \dim(C(\boldsymbol{Z})) = p - 1$, we have

$$t^2_{\max} = \frac{\boldsymbol{y}^T \boldsymbol{P_Z y}/\sigma^2}{\widehat{\sigma}^2/\sigma^2} \sim (p-1) F_{p-1,n-p} \text{ under null.}$$

Once again using the orhtogonality,

$$\boldsymbol{P_X} = \boldsymbol{P_Z} + \boldsymbol{P_{1_n}} \implies \boldsymbol{y}^T \boldsymbol{P_Z y}^T = \left\| \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \bar{y}\boldsymbol{1}_n \right\|^2$$

Thus $t^2_{\max}$ is equivalent to the F statistics used for testing $\boldsymbol{H}_0 : \boldsymbol{\beta}_{-0} = \boldsymbol{0}$. Finally observe that under the null (any one of the two null hypothesis would suffice, since $\boldsymbol{H}_0 : \boldsymbol{c}^T\boldsymbol{\beta}_{-0} = 0$ is an implication of $\boldsymbol{H}_0 : \boldsymbol{\beta}_{-0} = \boldsymbol{0}$)

$$\widetilde{\boldsymbol{c}}^T \widehat{\boldsymbol{\beta}} \sim \mathrm{N}\left(0, \sigma^2 \widetilde{\boldsymbol{c}}^T \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \widetilde{\boldsymbol{c}}\right)$$

Then,

$$t(\boldsymbol{c}) \sim t_{n-p} \implies t(\boldsymbol{c})^2 \sim F_{1,n-p} \text{ under null.}$$

Considering the mean and variance of F-distribution it can be easily seen that the null distribution of $t^2_{\max}$ will be shifted towards the right from the null distribution of $t(\boldsymbol{c})^2$ and will have higher variance than the null distribution of $t(\boldsymbol{c})^2$. The comparison is shown in Fig 2.

```
n <- 1000; p <- 10

# Generating samples
x <- (p-1)*rf(n, p-1, n-p)
y <- rf(n, 1, n-p)

# X denotes t^2_max and Y denotes t(c)
xlab <- rep('X',n)
ylab <- rep('Y',n)

# Building a dataframe
df <- data.frame(value=c(x,y),lab=c(xlab,ylab),stringsAsFactors = F)

#Plot
dencomp <- ggplot(df,aes(x=value,fill=lab,color=lab,group=lab))+
  geom_histogram(aes(y = ..density..),
                 alpha = 0.4,position = position_dodge(),
                 binwidth = 0.5)+
  geom_line(aes(y = ..density..,),
```

```
            stat = 'density',show.legend = F)

# Saving the density comparison
ggsave(plot = dencomp, file = paste0("dencomp.png"), width = 5, height = 5)
```

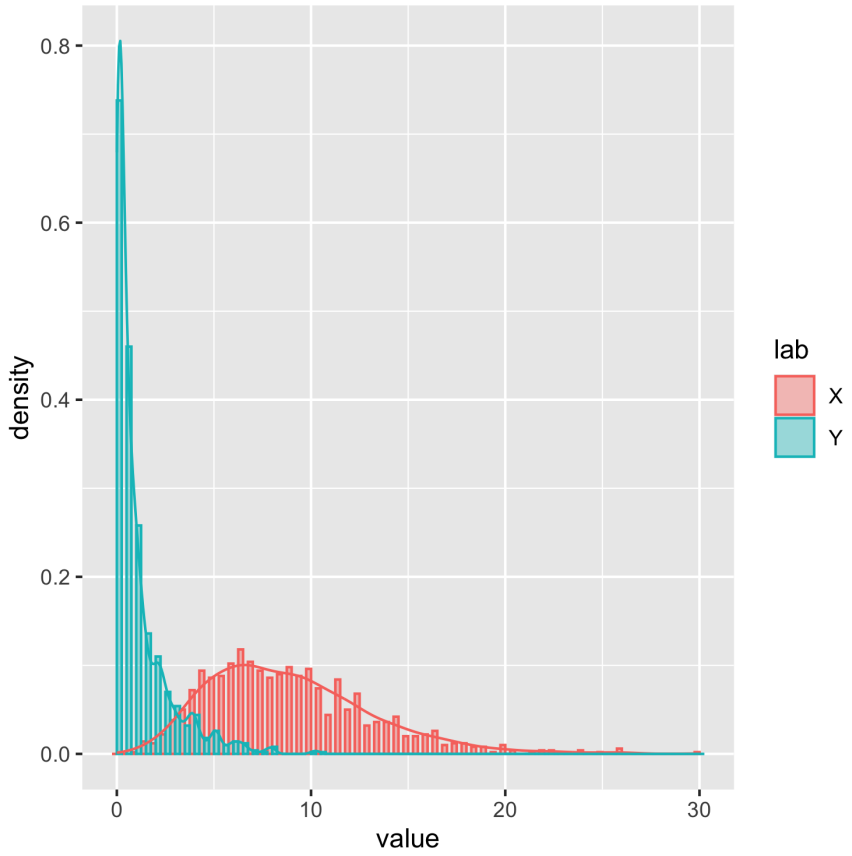

Figure 2: Comparison between $X \sim (p-1)F_{p-1,n-p}$ and $Y \sim F_{1,n-p}$.

Problem 3. 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 3.

(a) I took the population data from World Population Review. The below chunk creates the new variables `CrimeRate`.

```
# Reading population data
PopData <- read.csv("csvData.csv")[-31,2:3]

# Preprocessing population data

## Abbreviating state names
PopData$State <- state.abb[match(PopData$State,state.name)]

## Matching abbreviations to crime_data
```

```
PopData$State[49] <- "DC";PopData$State[32] <- "IO"

## Preparing for adding to crime_data
colnames(PopData) <- c("STATE","Pop")
PopData <- as.data.frame(PopData)

# Adding the population (pop) and crime rate (Violent/Pop) variables
crime_data <- crime_data %>%
  left_join(.,PopData) %>%
  mutate(CrimeRate = Violent/Pop) %>%
  print(.,n=5)

## Joining, by = "STATE"

## # A tibble: 51 x 8
##   STATE Violent Murder Metro HighSchool Poverty      Pop CrimeRate
##   <chr>   <dbl>  <dbl> <dbl>      <dbl>   <dbl>    <int>     <dbl>
## 1 AK        593      6  65.6       90.2     8     724357 0.000819
## 2 AL        430      7  55.4       82.4    13.7  4934193 0.0000871
## 3 AR        456      6  52.5       79.2    12.1  3033946 0.000150
## 4 AZ        513      8  88.2       84.4    11.9  7520103 0.0000682
## 5 CA        579      7  94.4       81.3    10.5 39613493 0.0000146
## # ... with 46 more rows
```

(b) From this point onwards we will consider `CrimeRate` to be violent crimes per 10^4 person.

```
# Changing CrimeRate to per 10^4 person
crime_data <- crime_data %>%
  mutate(CrimeRate = CrimeRate*(10^4))

# Variation and Covariation Plots
cov.plot <- crime_data %>%
      GGally::ggpairs(
        columns = c(4,5,6,8),
        lower = list(continuous = 'smooth'))

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

ggsave(plot = cov.plot, filename = "covariance_plot.png",
        device = "png", width = 7, height = 6)

# Correlation plot of the variables
crime.slice <- crime_data[,c("Metro",
        "HighSchool", "Poverty", "CrimeRate")]
corr.plot <- ggcorrplot::ggcorrplot(round(cor(crime.slice), 3))

ggsave(plot = corr.plot, filename = "corr_plot.png",
```

```
        device = "png", width = 6, height = 6)

# Summary statistics of crime_data
crime_summary <- crime_data %>%
      vtable::st(,out = "return",
        vars = c("Metro","HighSchool","Poverty","CrimeRate"),
        digits = 2) %>%
      kableExtra::kable(format = "latex",
                    row.names = NA, booktabs = T,
                    digits = 2, align = rep("c",4)) %>%
      kableExtra::save_kable("summary_stat.pdf")
```

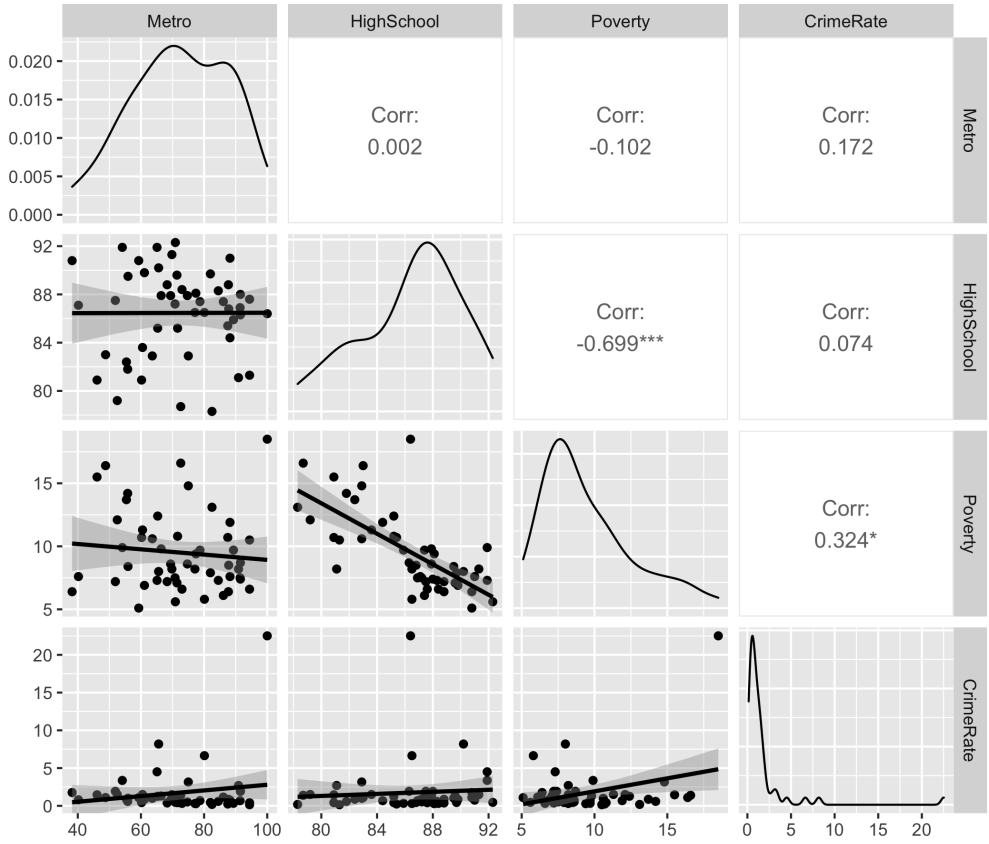

Figure 3: Scatterplots of all variables combinations among `Metro, HighSchool, Poverty` and `CrimeRate` with `CrimeRate` shown in scale of violent crimes per 10^4 people.

The variation and covariations can be observed from Fig 3 and 4 and the summary statistics are given in Table 5. From Fig 4 we can see that based on the data, poverty rate has a strong negative correlation with high school graduation rate, which in line with our intuition that in areas where high school graduation rates are high, poverty rate would come down. One surprising observation from Fig 3 and 4 is the positive correlation between all the socio-economic factors and crime rate. We will explore this further in part (c). From Table 5 we can observe that there is lot of variability in the percentage of population living in metropolitan areas. Also looking at the interquartile range and maximum values of `CrimeRate`, there

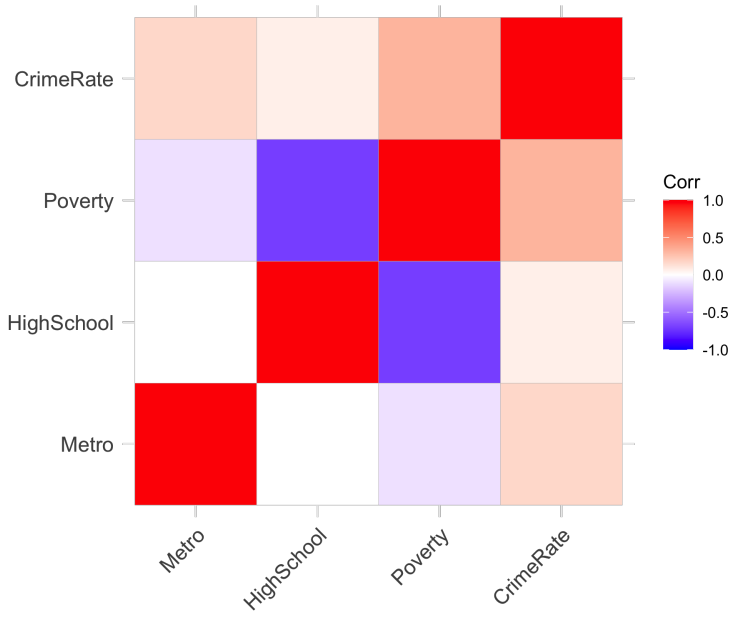

Figure 4: Correlation heatmap of `Metro, HighSchool, Poverty` and `CrimeRate` with `CrimeRate` shown in scale of violent crimes per 10^4 people.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Metro	51	72.25	15.28	38.2	60.8	86.8	100
HighSchool	51	86.46	3.62	78.3	84	88.8	92.3
Poverty	51	9.51	3.13	5.1	7.3	10.75	18.5
CrimeRate	51	1.75	3.33	0.15	0.45	1.49	22.52

Figure 5: Summary statistics of `crime_data` with `CrimeRate` shown in scale of violent crimes per 10^4 people.

appears to be presence of outlier in the data, which will be dealt with subsequently.

(c) We consider the following model,

$$\texttt{CrimeRate} = \beta_0 + \beta_1 \texttt{Metro} + \beta_2 \texttt{HighSchool} + \beta_3 \texttt{Poverty} + \boldsymbol{\epsilon} \tag{33}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_{n\times n})$, variable names in the model indicates the n dimensional column vector of the same name from `crime_data` and `CrimeRate` has been modified to be in scale of violent crimes per 10^4 people. In order to ascertain that there is some socio-economic effect on the crime rate in the states we conduct the following hypothesis test,

$$\boldsymbol{H}_0 : \boldsymbol{\beta}_{-0} = \mathbf{0} \text{ vs } \boldsymbol{H}_1 : \boldsymbol{\beta}_{-0} \neq \mathbf{0} \tag{34}$$

In the following chunk we fit the linear model in (33).

```r
# Fitting the linear model
crime.lm <- crime_data %>%
        lm(CrimeRate ~ Metro + HighSchool + Poverty,.)

# Summary of the fitted linear model
summary(crime.lm)

##
## Call:
## lm(formula = CrimeRate ~ Metro + HighSchool + Poverty, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1346 -1.5118 -0.6766  0.6883 11.7901
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.52692   14.80527  -4.021 0.000209 ***
## Metro         0.05468    0.02607   2.097 0.041367 *
## HighSchool    0.57139    0.15318   3.730 0.000515 ***
## Poverty       0.83333    0.17793   4.683 2.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.788 on 47 degrees of freedom
## Multiple R-squared:  0.342,Adjusted R-squared:    0.3
## F-statistic: 8.144 on 3 and 47 DF,  p-value: 0.0001802
```

From the `F-statistic` given in the summary we can conclude that socio-economic effect on crime rate in states is significant at the 5% level. Based on the above conclusion we wish to look which of the socio-economic factors, namely `Metro, HighSchool` and `Poverty` are significant. Hence we consider the following hypothesis,

$$\boldsymbol{H}_{0,i} : \beta_i = 0 \text{ vs } \boldsymbol{H}_{1,i} : \beta_i \neq 0, \quad \forall 1 \leq i \leq 3$$

From the above summary we can see that the coefficients of all the socio-economic variables are significant at the 5% level.
Let us also discuss the implications of the coefficient estimates. Keeping all other covariates (socio-economic variables) fixed, an increase of 0.054 violent crimes per 10^4 person is associated with an unit increase in percentage of population living in metro cities, an increase of 0.57 violent crimes per 10^4 person is associated with an unit increase in high school graduation rate, and an increase of 0.83 violent crimes per 10^4 person is associated with an unit increase in poverty rate.
We also look at confidence interval of the coefficients of the socio-economic factors.

```r
# Table of Confidence Intervals of Coeff estimates
confint(crime.lm) %>%
  kableExtra::kable(format = "latex",
                    row.names = NA, booktabs = T,
```

```
                    digits = 4, align = rep("c",4)) %>%
      kableExtra::save_kable("confint.pdf")
```

	2.5 %	97.5 %
(Intercept)	-89.3113	-29.7426
Metro	0.0022	0.1071
HighSchool	0.2632	0.8796
Poverty	0.4754	1.1913

Figure 6: Confidence interval of the fitted coefficients in the linear model (33)

From Table 6 none of the confidence intervals contain 0. This shows that with probability .95 the true parametes are in the intervals given in Table 6. This means with high probability all the socio-economic factors are positively related to percentage of violent crimes. This seems counter-intuitive, as a result let us look at the regression diagnostic presented in Fig 7.

```
# Plotting residuals of the regression against fitted values
fitvsres <- crime.lm %>%
  ggplot(aes(.fitted,.resid)) +
  geom_point() +

  # Fit a locally weighted regression
  stat_smooth(method="loess") +

  # Showing the y-axis
  geom_hline(yintercept=0, col="red", linetype="dashed") +

  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Residual vs Fitted Plot") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))

# Saving the residual vs fitted values
ggsave(plot = fitvsres, filename = "regdiag.png",
       device = "png", width = 4, height = 4)

## `geom_smooth()` using formula 'y ~ x'
```

It is clear from Fig 7 that there is presence of outlier, which can heavily impact the conclusions drawn above. The outlier state is given in Table 8.

```
# Index of the outlier state
out.state <- which.max(crime.lm$residuals)
```

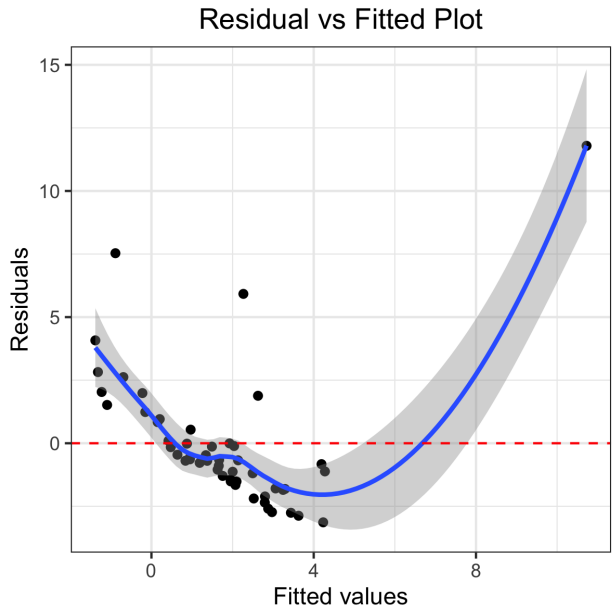

Figure 7: Plot of residuals against fitted values for the linear model (33)

```
# Details of the outlier state in a table
crime_data[out.state,] %>%
  kableExtra::kable(format = "latex",
                    row.names = NA, booktabs = T,
                    digits = 4, align = rep("c",4)) %>%
        kableExtra::save_kable("outlier.pdf")
```

STATE	Violent	Murder	Metro	HighSchool	Poverty	Pop	CrimeRate
DC	1608	44	100	86.4	18.5	714153	22.5162

Figure 8: The outlier state `DC`

Looking at the `crime_data` and comparing `DC` (from Table 8) with other states we can see that although it has high percentages of metropolitan population, high rates of high school graduation, there is significantly high percentage of violent crimes happening there, which does not agree with the other states. Let us redo the above analysis with `DC` removed from the dataset.

```
# Running the regression with DC removed
crime.lm.new <- crime_data %>%

  # Removing the outlier state
  slice(-out.state) %>%

  lm(CrimeRate ~ Metro + HighSchool + Poverty, .)
summary(crime.lm.new)
```

```
##
## Call:
## lm(formula = CrimeRate ~ Metro + HighSchool + Poverty, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2947 -0.7477 -0.4244  0.1162  6.5108
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.303786   9.858384  -0.335    0.739
## Metro       -0.014132   0.015933  -0.887    0.380
## HighSchool   0.066073   0.098105   0.673    0.504
## Poverty     -0.006652   0.127960  -0.052    0.959
##
## Residual standard error: 1.548 on 46 degrees of freedom
## Multiple R-squared:  0.04577,Adjusted R-squared:  -0.01647
## F-statistic: 0.7354 on 3 and 46 DF,  p-value: 0.5363
```

From the `F-statistic` given in the summary we can see that there is not enough evidence to reject the hypothesis (34) and thus we can conclude that based on the available data, none of the socio-economic factors have significant effect on the `CrimeRate` in a state.

(d) Based on the data collected the state of `DC` appears to be different from the others. It has significantly higher crimes per capita, when considered with poverty rate, metropolitan population and high school graduation rates. The crime rates in the state of `DC` needs to be analysed further and it would be unwise to implement matching (with other states) policies based upon the data collected. In the present data, it can be seen that `DC` is not in line with other states, and in order to make useful inferences from the data, it is best to remove `DC` from the dataset.

Based on the provided data, the socio-economic factors of poverty rates, metropolitan population and high school graduation rates cannot explain the variability of crime rates across states. In simpler terms, the data doesn't provide any evidence to suggest that above three socio-economic factors have any significant effect on the crime rates across the states.

But one must take this conclusion with a grain of salt. Although there was no evidence of significant effect of the factors, one cannot dismiss that they do not play any role. There may be other deeper factors at play here which were absent in the data. For example, the existing policies in metropolitan areas across states may be good enough to keep crimes at bay, which could overshadow the effects of the factors considered here.