STAT 9610: Homework 2

Chris Zhanran Lin

Due October 10, 2023 at 10:00am

1 Instructions

Setup. Clone this repository and open `homework-2.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: Zhihan Huang, Joseph Rudoler, Henry Shugart

Please disclose which AI tools you used, and how you used them: ChatGPT, checking R codes

Failure to answer the above questions will result in a 5-point penalty.

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 per-observation Fisher information matrix. Furthermore, an optimal test of the null hypothesis

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

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) Because we have covariates $\boldsymbol{X}$ here. Our observations are actually can be viewed as

$$y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \epsilon_i \sim \mathcal{N}(\boldsymbol{X}_i^T\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$$

independently for $i \in [n]$. This is a model parameterized by a vector $\boldsymbol{\beta}$, while, at the same time, depends on the covariate $\boldsymbol{X}_i$.

(b) Because we need to consider covariates in each observations y_i under the regression model. We can write down the density function

$$p_\beta(X_i, y_i) = \frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y_i - \boldsymbol{X}_i^T\beta)^2\},$$

then

$$\begin{aligned}
\boldsymbol{I}(\boldsymbol{\beta}) &= -\mathbb{E}_\beta\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\log p_\beta(\boldsymbol{X}, \boldsymbol{y})\right] \\
&= -\mathbb{E}_\beta\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\log\prod_{i=1}^n p_\beta(\boldsymbol{X}_i, y_i)\right] \\
&= -\mathbb{E}_\beta\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})\right) \\
&= \frac{1}{\sigma^2}E_\beta\sum_{i=1}^n\boldsymbol{X}_i\boldsymbol{X}_i^T = \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{X}).
\end{aligned}$$

Then we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \sim \mathcal{N}(0, n\boldsymbol{I}(\boldsymbol{\beta})^{-1}) = \mathcal{N}(0, n\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

For asymptotic property, if we have

$$\lim_{n\to\infty}\sigma^2(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T])^{-1}$$

then we may derive $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \to \mathcal{N}(0, \sigma^2(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T])^{-1})$.

(c) Let

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2\right\},$$

and

$$l(\boldsymbol{\beta}, \sigma^2) = \ln L(\boldsymbol{\beta}, \sigma^2) = -\ln\sqrt{2\pi}\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2.$$

Take the maximum likelihood estimate by derivation, let

$$\frac{\partial l}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\sum_{i=1}^n(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})\boldsymbol{X}_i = 0$$

we obtain $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}y$. Similarly, take $\frac{\partial l}{\partial\sigma} = 0$ we can derive $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2$. Comparing to the estimates discussed in class, we can find that the $\hat{\boldsymbol{\beta}}$ keeps the same, while the coefficient of $\widehat{\sigma^2}$ changes from $\frac{1}{n-p}$ to $\frac{1}{n}$ (though they share the same asymptotic results).

(d) Consider $H_0 : \boldsymbol{\beta} \in \Theta_0 = \{\beta : \beta_S = \mathbf{0}\}$ and $H_1 : \boldsymbol{\beta} \in \Theta_1 \setminus \Theta_0 = \{\beta : \beta_S \neq \mathbf{0}\}$, obviously for Θ_1, the maximum is achieved at the MLE estimator, which is

$$l(\hat{\boldsymbol{\beta}}, \widehat{\sigma^2}) = -\frac{n}{2}\log(\frac{2\pi}{n}) - \frac{n}{2}\log(RSS_1) - \frac{n}{2},$$

for $RSS_1 = \|y - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$. Similarly, under Θ_0 we can view the problem as regression using the covariates with indexes $-S$, and the maximum is derived by

$$l(\hat{\boldsymbol{\beta}_{-S}}, \widehat{\sigma^2_{-S}}) = -\frac{n}{2}\log(\frac{2\pi}{n}) - \frac{n}{2}\log(RSS_0) - \frac{n}{2}$$

with $RSS_0 = \|y - \boldsymbol{X}_{;-S}\hat{\boldsymbol{\beta}_{-S}}\|^2$. Thus we have

$$\Lambda = \frac{\max_{\boldsymbol{\theta}\in\Theta_1}\prod_{i=1}^n p_{\boldsymbol{\theta}}(y_i)}{\max_{\boldsymbol{\theta}\in\Theta_0}\prod_{i=1}^n p_{\boldsymbol{\theta}}(y_i)} = \frac{\exp\{-\frac{n}{2}\log(RSS_1)\}}{\exp\{-\frac{n}{2}\log(RSS_0)\}} = \left(\frac{RSS_0}{RSS_1}\right)^{n/2},$$

then

$$2\log\Lambda = n\log\left(\frac{RSS_0}{RSS_1}\right) \to \chi^2_k$$

with $k \equiv \dim(\Theta_1) - \dim(\Theta_0) = |S|$. The form is close to the F-test, while $\left(\frac{RSS_0}{RSS_1}\right)$ is used by a log transformation (Although χ^2 distribution is not a F-distribution).

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.

(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 $H_0 : \boldsymbol{\beta}_{\text{-}0} = \mathbf{0}$ (here, $\boldsymbol{\beta}_{\text{-}0} \equiv (\beta_1, \dots, \beta_{p-1})^T$). One way of going about this is to start with the usual test statistic $t(\boldsymbol{c})$ for the null hypothesis $H_0 : \boldsymbol{c}^T \boldsymbol{\beta}_{\text{-}0} = 0$, 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 2.

(a) We have

$$R^2 = 1 - \frac{\|y - \boldsymbol{X}\boldsymbol{\beta}\|^2}{\|y - \bar{y}\mathbf{1}_n\|^2},$$

then considering F-test for $H_0 : \boldsymbol{\beta}_S = 0$ for $S = \{1, 2, \dots, p\}$ as in problem 1(c), we will have

$$F = \frac{\frac{\|y - \bar{y}\mathbf{1}_n\|^2 - \|y - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{p-1}}{\frac{\|y - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{n-p}} = \frac{n-p}{p-1} \cdot \frac{R^2}{1 - R^2}.$$

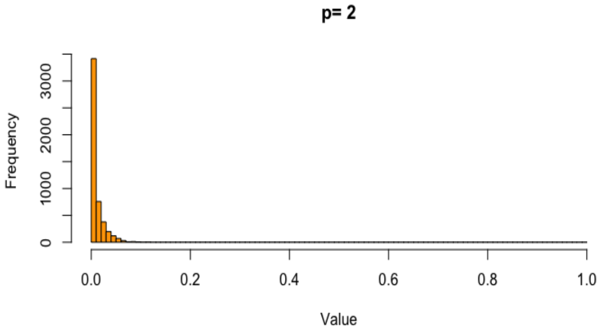
F and R^2 are positively related. Intuitively, as R^2 refers to the proportion of variance that has been explained by the model, F shows the significance that the covariates explain the respond. They should be somehow positively related.

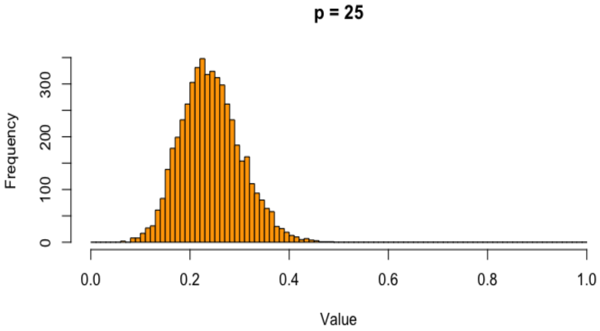
(b) Consider that

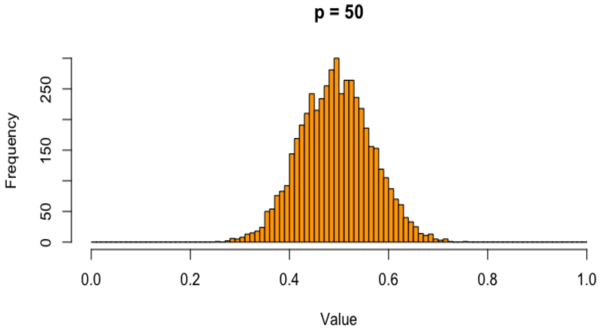
$$R^2 = \frac{F}{F + \frac{n-p}{p-1}}$$

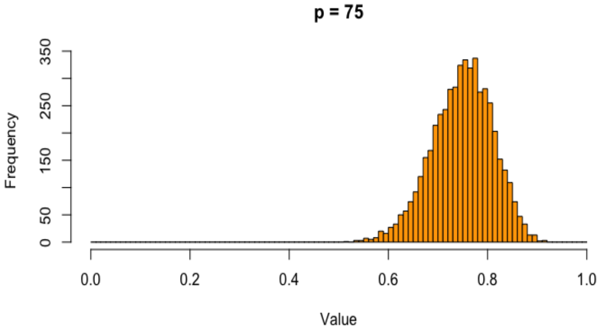
from (a), we can sample F distribution and get null-distribution simulation of R^2. The simulations are shown below.

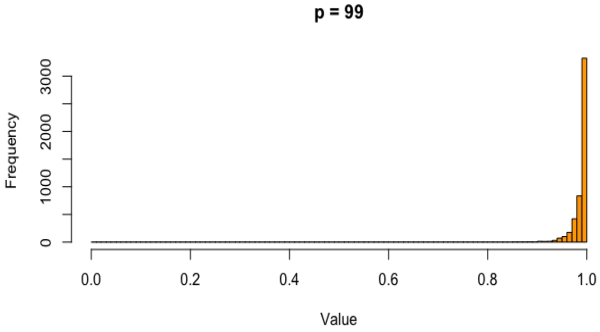
For each p we generate 50000 samples to get the frequency result. We can see that the distribution is on $[0, 1]$. As p increases, the distribution goes to the right side (more close to 1). Intuitively, larger p refers to more covariates involved in the model, thus we can obtain better explanations of the total variance.


p= 2


p = 25


p = 50


p = 75


p = 99

(c) We can write down the t-statistic as

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}},$$

while the F-statistic is

$$F = \frac{\|\boldsymbol{X}_{;-j}\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}}.$$

Thus the latter is obviously the square of the former.

(d) Consider $\boldsymbol{c}^T\boldsymbol{\beta}_{-0} = 0$, then we have

$$t(\boldsymbol{c}) = \frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}_{-0}}{\hat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{-1,-1}\boldsymbol{c}}}.$$

Following this, we may derive the maximimal

$$t^2_{max} = \max t(\boldsymbol{c})^2 = \max\left(\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}}_{-0}}{\hat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{-1,-1}\boldsymbol{c}}}\right) = \frac{1}{\hat{\sigma}^2}\hat{\boldsymbol{\beta}}^T_{-0}(\boldsymbol{X}^T\boldsymbol{X})_{-1,-1}\hat{\boldsymbol{\beta}}_{-0}$$

with $\boldsymbol{c} = \hat{\boldsymbol{\beta}}_{-0}$. The distribution turns out to be $\frac{p-1}{n-p}F_{p-1,n-p}$, which shares the same form with the F-statistic in (a) by multiplying a constant. Obviously we have $t^2_{max} \geq t(\boldsymbol{c})^2$ and t^2_{max} would be more related to F-statistics as shown.

Problem 3. Case study: Violent crime.
The `Statewide_crime.tsv` file contains information on the number of violent crimes and murders for each U.S. state in 2015, as well as three socioeconomic indicators: percent living in metropolitan areas, high school graduation rate, and poverty rate (Table 1).

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

STATE	Violent	Murder	Metro	HighSchool	Poverty
AK	593	6	65.6	90.2	8.0
AL	430	7	55.4	82.4	13.7
AR	456	6	52.5	79.2	12.1
AZ	513	8	88.2	84.4	11.9
CA	579	7	94.4	81.3	10.5

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 `left_join()` from the `dplyr` package) and create a new variable called `CrimeRate` defined as `CrimeRate = Violent/Pop` (see `mutate()` from the `dplyr` package).

(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, including any relevant tables or plots in your LaTeX report. 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) See the codes.

(b) The summary is shown in Table 2 and the variance is in Table 3. The covariance matrix is shown in Figure 1. It looks like the Poverty is more (positively) related to CrimeRate, while it (negatively) relates to the HighSchool number, which is totally reasonable. A scatterplot is shown in Figure 2. It looks like there is no significant linear relationship between CrimeRate and the other three variables (separately).

(c) We construct linear tests separately on the linear relationship of CrimeRate and the other three socioeconomic variables.
In Table 4, 5, 6 shows that for each separate variable, we don't have significant evidence to reject the null hypothesis $H_0 : \beta_j = 0$ (since the p-values $0.261, 0.737, 0.166$ all far beyond 0.05). However, we can do linear model based hypothesis test on the three variables simultanously as shown in Table 7. The p-value for the whole model is 0.05276 (not shown in the table, see in the codes), which is slightly over than 0.05 (as a common boundary), showing that the whole linear relationship is in an indistinct boundary; the p-values for the Poverty and

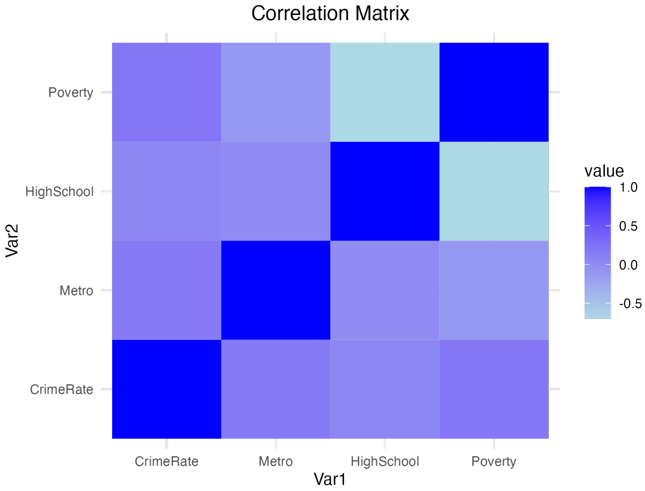


Figure 1: Correlation Matrix

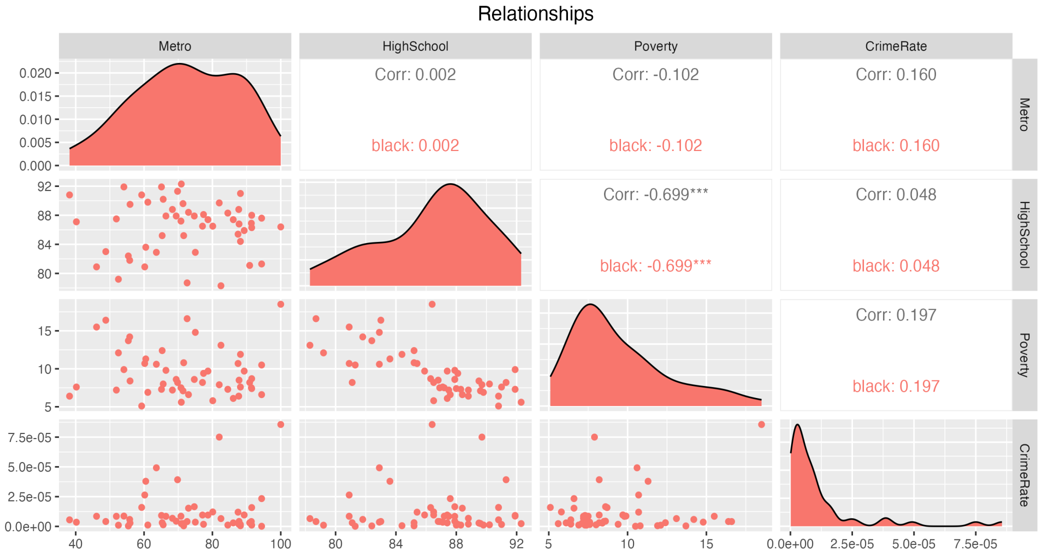


Figure 2: Scatterplot

CrimeRate	Metro	HighSchool	Poverty
Min. :1.289e-07	Min. : 38.20	Min. :78.30	Min. : 5.100
1st Qu.:2.373e-06	1st Qu.: 60.80	1st Qu.:84.00	1st Qu.: 7.300
Median :5.525e-06	Median : 71.60	Median :87.20	Median : 8.500
Mean :1.151e-05	Mean : 72.25	Mean :86.46	Mean : 9.506
3rd Qu.:1.080e-05	3rd Qu.: 86.80	3rd Qu.:88.80	3rd Qu.:10.750
Max. :8.559e-05	Max. :100.00	Max. :92.30	Max. :18.500

Table 2: Summary Table

	Variance
CrimeRate	3.038195e-10
Metro	2.333529e+02
HighSchool	1.307638e+01
Poverty	9.794165e+00

Table 3: Variance

HighSchool terms are smaller than 0.05, which shows significant evidence for us to reject the hypothesis $\beta_{\text{poverty}} = 0$ and $\beta_{\text{highschool}}$. The confidence intervals could also be seen in the table through standard errors. And a intuitive visuallization could be refered to Figure 2. When we construct a linear model based on both the three variables, the HighSchool and Poverty items show significant linear relationship with the outcome CrimeRate.

(d) By our statistical model, it is shown that the HighSchool and Poverty items might have (linear) relationship with the CrimeRate, while the Metro number seems less relative. We need to improve economical conditions (of cities) and educational guidance to enhance the population's quality, which might further alleviate the crime rate. The limitations (or, caveats) might be: (1) too few variables to conduct conclusions, and the significance depends on the model we take. (2) The terms we considered might be biased from practical settings, e.g., population and crime rate per captia might not be the only way to depict average effect violent crimes. Some other data we may want to gather might be: range of metropolitan areas (square miles), repeat-crime data, high-school-related violent rate, which might help us have a fine-grain analysis on the crime issues. At the same time, we may want goverment policy data (e.g., gun-related policy), GDP, etc, which could provide us with more variables to do better analysis.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.714230e-06	1.187644e-05	-0.1443387	0.8858251
Metro	1.830758e-07	1.608940e-07	1.1378655	0.2607099

Table 4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.590027e-06	5.951977e-05	-0.1443222	0.885838
HighSchool	2.325029e-07	6.877968e-07	0.3380401	0.736777

Table 5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.081375e-06	7.799201e-06	0.138652	0.8902934
Poverty	1.097367e-06	7.800608e-07	1.406771	0.1658065

Table 6

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.951633e-04	8.803279e-05	-2.216938	0.03150170
Poverty	2.742104e-06	1.057978e-06	2.591834	0.01268023
HighSchool	1.888735e-06	9.108256e-07	2.073651	0.04361672
Metro	2.395158e-07	1.550299e-07	1.544965	0.12906211

Table 7