Homework 2

Sam Rosenberg

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:
James Blume
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) We can rewrite (1) as follows:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$
$$= \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{\beta} + \epsilon_1 \\ \vdots \\ \boldsymbol{x}_n^T \boldsymbol{\beta} + \epsilon_n \end{pmatrix},$$

where the covariance matrix of $\boldsymbol{\epsilon}$ is $\sigma^2 \boldsymbol{I}_n$. Note that this means the ϵ_i are all uncorrelated, but the fact that they are multivariate normal and uncorrelated implies that the ϵ_i are independent standard normal random variables. So, $y_i \sim \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.
Though each y_i is parametrized by $\boldsymbol{\beta} \in \mathbb{R}^p$, it is also a function of $\boldsymbol{x}_i$, which is not necessarily the same for each i since the $\boldsymbol{x}_i$ are regarded as fixed. As a result, the fixed-design model does not fit into the classical inferential setup, since the y_i need not be identically distributed.

(b) We saw in Unit 2 that $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$. So, we know that the joint density (equivalent to the likelihood function) for the fixed-design model is

$$p_\beta(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2],$$

and

$$\log p_\beta(\boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2.$$

Then

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log p_\beta(\boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

and

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log p_\beta(\boldsymbol{y}) = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}.$$

So, our analog of Fisher information is

$$-\mathbb{E}_\beta[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \log p_\beta(\boldsymbol{y})] = \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}.$$

(c) Define our parameter vector of interest to be $\boldsymbol{\theta} := (\boldsymbol{\beta}, \sigma^2)$. Because $\log(x)$ is a monotone increasing function, any maximum of $p_{\boldsymbol{\theta}}(\boldsymbol{y})$ is also a maximum of $\ell_{\boldsymbol{\theta}}(\boldsymbol{y}) := \log[p_{\boldsymbol{\theta}}(\boldsymbol{y})]$ and vice versa.
We saw in (b) that

$$\ell_{\boldsymbol{\theta}}((y)) = -\frac{1}{2}[n\log(2\pi\sigma^2) + \frac{1}{\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2]$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}}\ell_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Under the full rank assumption we obtain the same maximum likelihood estimator for $\boldsymbol{\beta}$ by setting the partials equal to 0 and solving for $\boldsymbol{\beta}$; that is, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
We also have that

$$\frac{\partial}{\partial \sigma}\ell_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{\sigma^3}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 - \frac{n}{\sigma}.$$

Again setting the partial equal to 0 and solving for σ^2 while substituting our estimator $\widehat{\boldsymbol{\beta}}$, we see that the maximum likelihood estimator is $\widehat{\sigma}^2 = \frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2 = \frac{1}{n}||(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}||^2 = \frac{1}{n}||\widehat{\boldsymbol{\epsilon}}||^2$. Thus, our estimator for $\boldsymbol{\beta}$ remains the same as the least squares estimator, while the estimator for ϵ differs from our unbiased estimator by a factor of $\frac{n}{n-p}$ (interestingly, this means that the MLE is asymptotically unbiased since $\frac{n}{n-p} \to 1$ as $n \to \infty$).

(d) Define $\Theta := (\sigma^2, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1}$. We also take $\Theta_1 := \Theta$ and $\Theta_0 := \{\theta = (\sigma^2, \beta_1, \ldots, \beta_p) \in \mathbb{R}^{p+1} | \boldsymbol{\beta}_S = 0\}$. Then we have that we are testing

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ versus } H_1 : \boldsymbol{\theta} \in \Theta_1.$$

Define $\ell_i(\boldsymbol{y}) := \max_{\boldsymbol{\theta} \in \Theta_i} p_{\boldsymbol{\theta}}(\boldsymbol{y})$, where $p_{\boldsymbol{\theta}}(\boldsymbol{y})$ is the density for the linear model; i.e.

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) := \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2].$$

We know from (c) that $(\widehat{\sigma}_1^2, \widehat{\beta}^1) := \arg\max_{\boldsymbol{\theta} \in \Theta_1} p_{\boldsymbol{\theta}}(\boldsymbol{y}) = (\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y})$. Also,

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta} \in \Theta_0} p_{\boldsymbol{\theta}}(\boldsymbol{y}) &= \arg\max_{\boldsymbol{\theta} \in \Theta : \boldsymbol{\beta}_S = 0} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2] \\
&= \arg\max_{\boldsymbol{\theta} \in \Theta : \boldsymbol{\beta}_S = 0} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2}||\boldsymbol{y} - \boldsymbol{X}_{-S}\boldsymbol{\beta}_{-S}||^2].
\end{aligned}$$

But this is exactly the MLE for the model given by the partial model matrix $\boldsymbol{X}_{-S}$. So, $(\widehat{\sigma}_0^2, \widehat{\boldsymbol{\beta}}^0) := \arg\max_{\boldsymbol{\theta} \in \Theta_0} p_{\boldsymbol{\theta}}(\boldsymbol{y}) = (\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}||^2, (\boldsymbol{X}_{-S}^T\boldsymbol{X}_{-S})^{-1}\boldsymbol{X}_{-S}\boldsymbol{y})$.
We then have

$$\begin{aligned}
\Lambda :=& \frac{\ell_1(\boldsymbol{y})}{\ell_0(\boldsymbol{y})} \\
=& \frac{(2\pi\widehat{\sigma}_1^2)^{-n/2} \exp[-\frac{1}{2\widehat{\sigma}_1^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2]}{(2\pi\widehat{\sigma}_0^2)^{-n/2} \exp[-\frac{1}{2\widehat{\sigma}_0^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^0||^2]} \\
=& (\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_1^2})^{n/2} \exp[\frac{1}{2}(\frac{1}{\widehat{\sigma}_0^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^0||^2 - \frac{1}{\widehat{\sigma}_1^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2)].
\end{aligned}$$

So,

$$\begin{aligned}
2\log\Lambda &= n\log(\frac{\widehat{\sigma}_0^2}{\widehat{\sigma}_1^2}) + (\frac{1}{\widehat{\sigma}_0^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^0||^2 - \frac{1}{\widehat{\sigma}_1^2}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2) \\
&= n\log(\frac{||\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}||^2}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2}) + n(\frac{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^0||^2}{||\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}||^2} - \frac{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2}) \\
&= n\log(\frac{||\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}||^2}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2}) \\
&= n\log(\frac{||(\boldsymbol{I} - \boldsymbol{H}_{-S})\boldsymbol{y}||^2}{||(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}||^2}),
\end{aligned}$$

since $||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^0||^2 = ||\boldsymbol{y} - \boldsymbol{X}_{-S}\widehat{\boldsymbol{\beta}}_{-S}||^2$ and $||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^1||^2 = ||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2$.

Then

$$\begin{aligned} 2\log\Lambda &= n\log(\frac{||(\boldsymbol{I}-\boldsymbol{H}_{-S})\boldsymbol{y}||^2}{||(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}||^2}) \\ &= n\log(1+\frac{||(\boldsymbol{I}-\boldsymbol{H}_{-S})\boldsymbol{y}||^2-||(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}||^2}{||(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}||^2}) \\ &= n\log\left(1+\frac{|S|}{n-p}F\right). \end{aligned}$$

Using the Taylor expansion $\log(1+x) = x + O(x^2)$, we see that

$$\begin{aligned} 2\log\Lambda &= n\log\left(1+\frac{|S|}{n-p}F\right) \\ &= |S|\frac{n}{n-p}F + nO((\frac{|S|}{n-p}F)^2) \\ &= |S|\frac{n}{n-p}F + O(n^{-1}). \end{aligned}$$

Note that as $n \to \infty$, this quantity tends toward $|S|F$. That is to say, asymptotically $2\log\Lambda$ is within a factor $|S|$ of F. Recall that for an F-distribution with $|S|$ and $n-p$ degrees of freedom, as the denominator degrees of freedom $n-p \to \infty$, the distribution converges to $\chi^2_{|S|}$. So, $\frac{2}{|S|}\log\Lambda$ is asymptotically distributed as $\chi^2_{|S|}$.

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 $\boldsymbol{c} \in \mathbb{R}^{p-1}$:

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

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) Recall that

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \overline{y}\mathbf{1}_n||^2}{||\boldsymbol{y} - \overline{y}\mathbf{1}_n||^2} = \frac{||(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{y}||^2}{||(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}||^2}.$$

Then we also have

$$1 - R^2 = \frac{||(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}||^2 - ||(\boldsymbol{H} - \boldsymbol{H}_0)\boldsymbol{y}||^2}{||(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}||^2} = \frac{||(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}||^2}{||(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y}||^2}.$$

So, $(\frac{n-p}{p-1})\frac{R^2}{1-R^2} = \frac{||(\boldsymbol{H}-\boldsymbol{H}_0)\boldsymbol{y}||^2/(p-1)}{||(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}||^2/(n-p)} =: F$. But this is exactly the F-statistic for the hypothesis test with null hypothesis $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$. We can also invert this relationship and find that for $c := \frac{p-1}{n-p}F$, we have $R^2 = \frac{c}{c+1}$.
Under the null distribution, the F-statistic is F distributed with $p-1$ and $n-p$ degrees of freedom. Note that R^2 and F are positively related which makes sense - a higher R^2 indicates that the full model explains more of the variance in the observed data, which in turn suggests that we do not have $\beta_1 = \cdots = \beta_{p-1} = 0$.

(b)
```
# Simulation parameters
n <- 100
p_list <- c(2, 25, 50, 75, 99)

# Dataframe for simulation data
sim_data <- data.frame()
```

```r
# Run simulation for different values of p
for(p in p_list){
  # Sample F-statistics
  f <- rf(n = n, df1 = p-1, df2 = n-p)

  # Compute R^2 from F-statistics
  c <- (p-1)/(n-p)*f
  R_2 <- c/(c+1)

  # Add simulated data to dataframe
  df <- data.frame(p = rep(p, length(R_2)), R_2 = R_2)
  sim_data <- rbind(sim_data, df)
}

# Make and save plot
R_2_plt <-
  sim_data %>%
    ggplot(aes(x = R_2)) +
      geom_histogram(bins = 30) +
      ylab("Count") +
      xlab(latex2exp::TeX("R^2")) +
      ylim(0, 35) +
      xlim(0, 1) +
      facet_wrap( ~ p) +
  theme(axis.text = element_text(size = 14),
        axis.title = element_text(size = 16, face = "bold"),
        strip.text = element_text(size = 20))

ggsave(plot = R_2_plt, filename = "./figures/R_2_plt.png",
       device = "png", width = 12, height = 8)
```

The plots of the empirical distributions of R^2 can be found in Figure 1. Note that as p (the facet label for each subplot) increases, the mass of the empirical null distribution of R^2 slowly shifts to the right, going from right-skewed, to relatively symmetric, then left-skewed. For fixed n, when $p \approx 1$ we have that $c \approx 0$, so $R^2 \approx 0$, whereas when $p \approx n-1$ we have $c \approx F$, so $R^2 \approx 1$.

(c) Recall that $t = \frac{\widehat{\beta}_j}{\widehat{\sigma}/s_j}$, so $t^2 = \frac{\widehat{\beta}_j^2 s_j^2}{\widehat{\sigma}^2}$ and that $F = \frac{(||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}||^2)/1}{(||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2)/(n-p)}$. But, $\widehat{\sigma}^2 = \frac{1}{n-p}||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2$ and $s_j^2 := ||\boldsymbol{x}_{*j}^{\perp}||^2$, so

$$t^2 = \frac{\widehat{\beta}_j^2 ||\boldsymbol{x}_{*j}^{\perp}||^2}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2/(n-p)} = \frac{||\widehat{\beta}_j \boldsymbol{x}_{*j}^{\perp}||^2}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2/(n-p)}$$

and thus it suffices to show that

$$||\widehat{\beta}_j \boldsymbol{x}_{*j}^{\perp}||^2 = ||\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j}||^2 = ||(\boldsymbol{H} - \boldsymbol{H}_{-j})\boldsymbol{y}||^2 = ||\boldsymbol{H}_{j|-j}\boldsymbol{y}||^2 = ||\widehat{\beta}_{j|-j}\boldsymbol{x}_{*j}^{\perp}||^2.$$

Note that this is in fact true by orthogonalization of the predictors, the coefficients are the same and so we are finished.

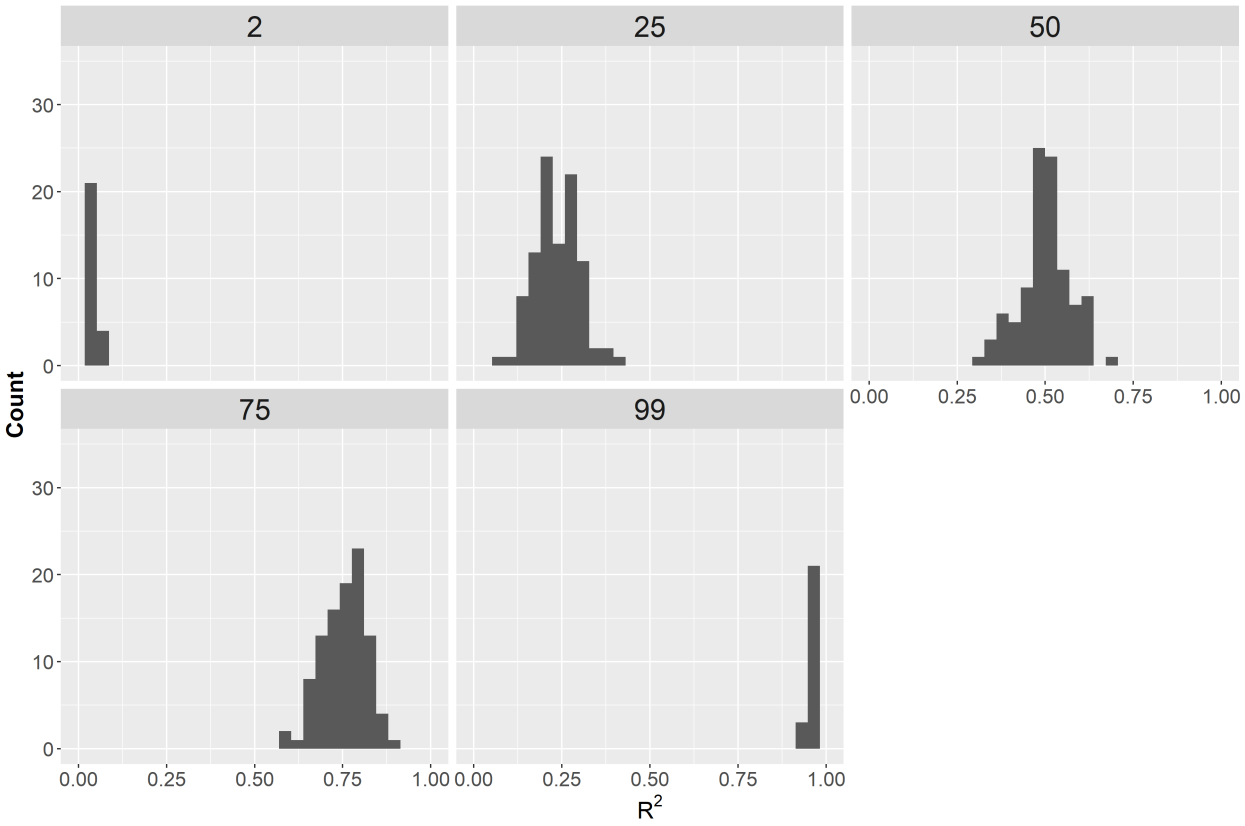

Figure 1: Distributions of R^2 as a function of p.

(d) For simplicity of notation, we write $\boldsymbol{c}$ as a vector in both $\mathbb{R}^p$ and $\mathbb{R}^{p-1}$ with the understanding that the first component of $\boldsymbol{c}$ is 0 when it is treated as a vector in $\mathbb{R}^p$.
We first recall that

$$t(\boldsymbol{c}) = \frac{\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_{-0}}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0})^{-1}\boldsymbol{c}}} \sim t_{n-(p-1)}$$

under the null hypothesis $H_0 : \boldsymbol{c}^T\boldsymbol{\beta}_{-0} = 0$. We then have that $t(\boldsymbol{c})^2 \sim F_{1,n-(p-1)}$.
Taking $A := (\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0})^{-1}$ and recognizing that $\widehat{\sigma}$ is independent of $\boldsymbol{c}$, we have that

$$\begin{aligned}
\boldsymbol{c}_{\max} &:= \underset{\boldsymbol{c}\in\mathbb{R}^{p-1}}{\arg\max}\, t(\boldsymbol{c}) \\
&= \underset{\boldsymbol{c}\in\mathbb{R}^{p-1}}{\arg\max}\, \frac{\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_{-0}}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0})^{-1}\boldsymbol{c}}} \\
&= \underset{\boldsymbol{c}\in\mathbb{R}^{p-1}}{\arg\max}\, \frac{\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_{-0}}{\sqrt{\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c}}}.
\end{aligned}$$

Note that this objective function is in fact scale invariant: if $\boldsymbol{d} := a\boldsymbol{c}$, then

$$\frac{\boldsymbol{d}^T \widehat{\boldsymbol{\beta}}_{-0}}{\sqrt{\boldsymbol{d}^T\boldsymbol{A}\boldsymbol{d}}} = \frac{(a\boldsymbol{c})^T \widehat{\boldsymbol{\beta}}_{-0}}{\sqrt{(a\boldsymbol{c})^T\boldsymbol{A}(a\boldsymbol{c})}} = \frac{\boldsymbol{c}^T \widehat{\boldsymbol{\beta}}_{-0}}{\sqrt{\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c}}}.$$

So, we have that we can rewrite the optimization problem as follows:

$$\boldsymbol{c}_{\max} = \underset{\boldsymbol{c}\in\mathbb{R}^{p-1}}{\arg\max} \frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{-0}}{\sqrt{\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c}}} = \underset{\boldsymbol{c}\in\mathbb{R}^{p-1} \text{ s.t. } \boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c}=1}{\arg\max} \boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{-0}.$$

We can solve this problem using Lagrange multipliers and end up with a system of equations:

$$\begin{cases} \boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{c}}[\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}_{-0} - \lambda(\boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c} - 1)] \\ 1 = \boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c} \end{cases} = \begin{cases} \boldsymbol{0} = \widehat{\boldsymbol{\beta}}_{-0} - \lambda 2\boldsymbol{A}\boldsymbol{c} \\ 1 = \boldsymbol{c}^T\boldsymbol{A}\boldsymbol{c}. \end{cases}$$

Solving, we get that $\lambda = \frac{1}{2}||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||$ and $\boldsymbol{c}_{\max} = \frac{\boldsymbol{A}^{-1}\widehat{\boldsymbol{\beta}}_{-0}}{||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||} = \frac{\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}}{||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||}$.
So,

$$\begin{aligned} \boldsymbol{c}_{\max}^T\widehat{\boldsymbol{\beta}}_{-0} &= \left(\frac{\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}}{||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||}\right)^T\widehat{\boldsymbol{\beta}}_{-0} \\ &= \frac{\widehat{\boldsymbol{\beta}}_{-0}^T\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}}{||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||} \\ &= \frac{(\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0})^T\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}}{||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||} \\ &= ||\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}|| \\ &= ||\boldsymbol{H}_{-0}\boldsymbol{y}||. \end{aligned}$$

Likewise, we have $\boldsymbol{c}_{\max}^T(\boldsymbol{X}_{-0}^T\boldsymbol{X}_{-0})^{-1}\boldsymbol{c}_{\max} = 1$ by construction. So, $t_{\max} = \frac{||\boldsymbol{H}_{-0}\boldsymbol{y}||}{\widehat{\sigma}} = \frac{||\boldsymbol{H}_{-0}\boldsymbol{y}||/\sqrt{1}}{||\boldsymbol{y}-\boldsymbol{X}_{-0}\widehat{\boldsymbol{\beta}}_{-0}||/\sqrt{n-p}}$ and

$$t_{\max}^2 = \frac{||\boldsymbol{H}_{-0}\boldsymbol{y}||/1}{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2/(n-p)}.$$

Because $\dim(C(\boldsymbol{X}_{-0})) = p-1$, we have that $\frac{1}{p-1}t_{\max}^2 \sim F_{p-1,n-p}$.

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)
```
# Replace incorrect abbreviation for Iowa
crime_data <-
  crime_data %>%
  mutate(STATE = replace(STATE, STATE=="IO", "IA")) %>%
  as.data.frame()

# Read in 2019 census data for state population estimates
pop_data <-
  read.csv("http://www2.census.gov/programs-surveys/popest/datasets/2010-2019/national/tot
  # Filter out rows that aren't actually states
  filter(! STATE %in% c(0, 72)) %>%
```

```r
  # Only take state and 2019 population columns
  select(NAME, POPESTIMATE2019) %>%
  # Rename column
  rename(Pop = POPESTIMATE2019)

# Mapping from full state name to abbreviation
state_abbrev_map <-
  read.csv("https://raw.githubusercontent.com/jasonong/List-of-US-States/master/states.csv

# Merge pop_data to get abbreviations for merge with crime_data
pop_data <-
  merge(pop_data, state_abbrev_map,
        by.x = "NAME", by.y = "State")

# Full dataset with population and
crime_data_full <-
  merge(crime_data, pop_data,
        by.x = "STATE", by.y = "Abbreviation") %>%
  select(-NAME) %>%
  # Create violent crime rate variable (crime per million people)
  mutate(CrimeRate = Violent/Pop*1000000)

head(crime_data_full)
```

```
##   STATE Violent Murder Metro HighSchool Poverty      Pop CrimeRate
## 1    AK     593      6  65.6       90.2     8.0   731545 810.61315
## 2    AL     430      7  55.4       82.4    13.7  4903185  87.69810
## 3    AR     456      6  52.5       79.2    12.1  3017804 151.10325
## 4    AZ     513      8  88.2       84.4    11.9  7278717  70.47945
## 5    CA     579      7  94.4       81.3    10.5 39512223  14.65369
## 6    CO     345      4  84.5       88.3     7.3  5758736  59.90898
```

(b)

```r
# For creating correlation plots
library(corrplot)
## corrplot 0.90 loaded

library(viridis)

## Loading required package:  viridisLite

# Produce and save correlation plot
png("./figures/corr_plt.png")
crime_data_full %>%
  select(CrimeRate, Metro, HighSchool, Poverty) %>%
  rename(`Crime Rate` = CrimeRate, `High School` = HighSchool) %>%
  cor() %>%
  corrplot(type = "upper", tl.cex = 1.4, cl.cex = 1.2,
           tl.col = "black", col = magma(256))
```

```
dev.off()

# Produce and save pairs plot
png("./figures/pairs_plt.png")
crime_data_full %>%
  select(CrimeRate, Metro, HighSchool, Poverty) %>%
  GGally::ggpairs(columnLabels =
                    c("Crime Rate", "Metro",
                      "High School", "Poverty"))

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

dev.off()

# Produce and save summary statistics
summary_stats <-
  crime_data_full %>%
  select(CrimeRate, Metro, HighSchool, Poverty) %>%
  rename(`Crime Rate` = CrimeRate, `High School` = HighSchool) %>%
  summary() %>%
  as.data.frame.matrix()

colnames(summary_stats) <-
  c("Crime Rate", "Metro", "High School", "Poverty")

summary_stats <-
  summary_stats %>%
  # Get rid of added text from summary()
  mutate(`Crime Rate` =
           str_replace(`Crime Rate`, pattern = ".*:", replacement = ""),
         Metro =
           str_replace(Metro, pattern = ".*:", replacement = ""),
         `High School` =
           str_replace(`High School`, pattern = ".*:", replacement = ""),
         Poverty =
           str_replace(Poverty, pattern = ".*:", replacement = ""))

rownames(summary_stats) <-
  c("Min.", "Q1", "Median", "Mean", "Q3", "Max.")

summary_stats %>%
  kableExtra::kable(format = "latex", booktabs = TRUE, digits = 8) %>%
  kableExtra::save_kable("figures/crime_tbl.png")
```

We plot the correlation between our variables in Figure 2, pairwise scatterplots in Figure 3, and summary statistics for each variable in Table 1. We see little correlation between the Metro

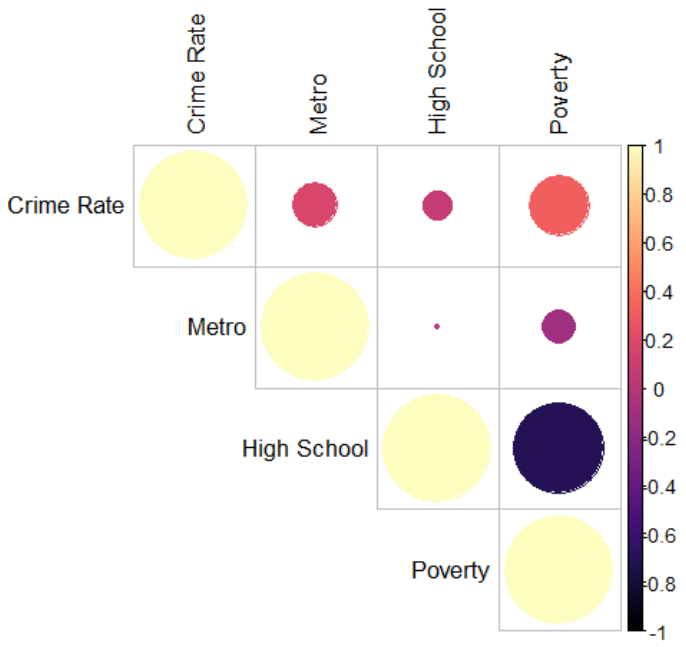

Figure 2: Correlation plot for crime data.

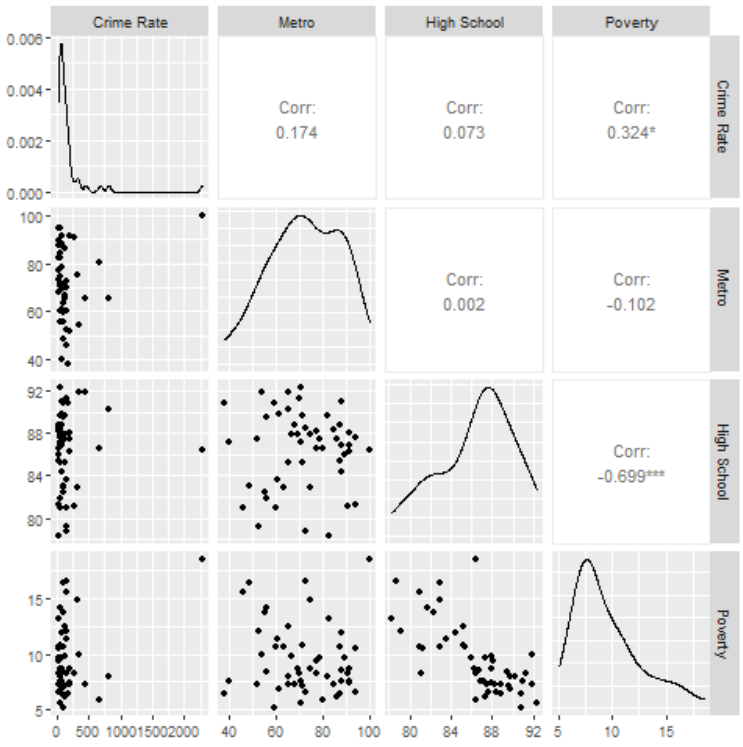

Figure 3: Pairwise plots for crime data.

and High School variables; some positive correlation between Metro and Crime Rate, as well as High School and Crime Rate; some negative correlation between Metro and Poverty; more

	Crime Rate	Metro	High School	Poverty
Min.	14.65	38.20	78.30	5.100
Q1	46.10	60.80	84.00	7.300
Median	86.39	71.60	87.20	8.500
Mean	176.39	72.25	86.46	9.506
Q3	150.25	86.80	88.80	10.750
Max.	2278.43	100.00	92.30	18.500

Table 1: Summary statistics for crime data.

positive correlation between Crime Rate and Poverty; and finally strong negative correlation between High School and Poverty (Figures 2, 3). We also see that the distributions of Crime Rate and Poverty are right-skewed, while those of Metro and High School are left-skewed (Figure 3).

(c)
```
lm_fit <-
  lm(CrimeRate ~ Metro + HighSchool + Poverty,
     data = crime_data_full)

summary(lm_fit)

##
## Call:
## lm(formula = CrimeRate ~ Metro + HighSchool + Poverty, data = crime_data_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -317.38 -153.11  -70.65   71.56 1193.79
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6016.023   1494.276  -4.026 0.000205 ***
## Metro           5.559      2.631   2.112 0.039994 *
## HighSchool     57.714     15.460   3.733 0.000510 ***
## Poverty        84.234     17.958   4.691 2.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281.4 on 47 degrees of freedom
## Multiple R-squared:  0.343,Adjusted R-squared:  0.3011
## F-statistic: 8.179 on 3 and 47 DF,  p-value: 0.0001744

confint(lm_fit) %>%
  as.data.frame() %>%
  kableExtra::kable(format = "latex", booktabs = TRUE, digits = 2) %>%
  kableExtra::save_kable("figures/conf_int_tbl.png")
```

	2.5 %	97.5 %
(Intercept)	-9022.12	-3009.93
Metro	0.26	10.85
HighSchool	26.61	88.82
Poverty	48.11	120.36

Table 2: Confidence intervals for linear model coefficients.

Looking at the estimated linear model, we say that all coefficients are significant at the $\alpha = 0.05$ level. Dually, we see that none of the confidence intervals for the coefficients contain 0 (Table 2). This suggests that each of the socioeconomic variables does in fact have an impact on crime rate. All such coefficients are positive, indicating that an increase in any one of the variables while holding the others fixed is associated with an increase in the crime rate.
We also see that the model has an R^2 of 0.34, indicating that it explains about 34% of the variation in the observed data. Notably, the p-value for the F-test is significant at the $\alpha = 0.05$ level, indicating that there is strong evidence that at least one of the non-intercept coefficients is nonzero.

(d) From a policymaking standpoint, the key takeaway is that an increase in any one of the variables (while holding the others constant), seems to be associated with an increase in crime. This may suggest that higher poverty or more metropolitan areas need additional policies to help combat the problem of violent crimes. One thing of note is that the model suggests that a higher high school graduation rate is associated with a higher violent crime rate when holding the other variables constant. This is counter to what one might intuitively expect, but could be because the poverty rate and high school graduation rate variables are highly (negatively) correlated, so care must be taken when drawing conclusions from the model. Other caveats are that the model does not consider interactions between the different variables (e.g. maybe a higher poverty rate increases the effect that the metropolitan residential rate has on crime rate) and assumes linear relationships between the variables.
Looking at data on the county level would help to increase the sample size and would reduce the likelihood of running into Simpson's paradox from dealing with aggregated data. Other socioeconomic data like population density and (un)employment rates could be interesting for a more in-depth analysis of violent crime as well.