Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

Due November 22, 2024
**61 total points**

---

## Problem Set 10

**Instructions**: Answer the following questions in their entirety in a separate document. Submit your completed problem set as a PDF via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename. Working together is encouraged, but it is expected that all submitted work be that of the individual student.

---

1. In the 2000 presidential election in the United States, the Democratic candidate was Al Gore and the Republican candidate was George W. Bush. In Palm Beach County, Florida, initial election returns reported 3,407 votes for the Reform party candidate, Pat Buchanan. Some political analysts thought that most of these votes may have actually been intended for Gore (whose name was next to Buchanan's) but were wrongly cast due to the ballot design. For the 67 counties in Florida, Figure 9.19 below is a scatterplot of the county wide vote for the Reform party candidates in 2000 (Buchanan) and in 1996 (Ross Perot). (**6 points**)

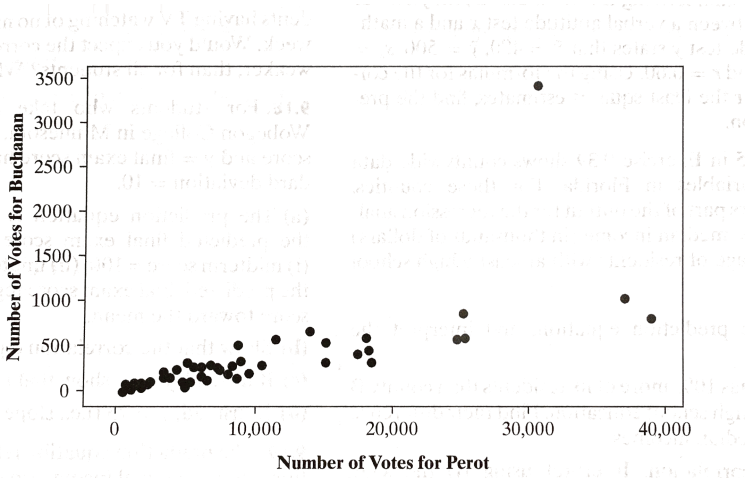


**FIGURE 9.19:** Scatterplot of Florida Countywide Vote for Reform Party Candidates Pat Buchanan in 2000 and Ross Perot in 1996

The prediction equation from a regression of Buchanan votes on Perot votes is $\hat{y} = 45.7 + 0.02414x$.

(a) Interpret the slope, in words. (**2 points**)

(b) In Palm Beach County, $x = 30,739$. Find the predicted Buchanan vote in Palm Beach County, the residual, and interpret both. (**4 points**)

2. In Stata, read the dataset called *crime_police.dta* from Github. This file contains data related to population, crime, and law enforcement for 92 cities in 1982 and 1987. The data were collected from the *County and City Data Book*. (**21 points**)

```
use https://github.com/spcorcor18/LPO-8800/raw/main/data/crime_police.dta
```

   (a) Create a scatterplot that shows the relationship between a city's crime rate (crimes per 1000 people) and its investment in law enforcement, as measured by police per 1000 people. Put the crime rate on the vertical axis and police on the horizontal axis, and *overlay* the best-fit line using `twoway` in Stata. What does this graph suggest to you about the relationship between police presence and crime, in terms of direction, strength and linearity? (**3 points**)

   (b) Don't trust your eyes? For your graph in part (a), overlay a third graph with `twoway` and `lowess` *`yvar xvar`*. `lowess` provides a "local" smoothed prediction of $y$ given values of $x$. Is linearity a reasonable approximation, according to the lowess plot? Would it improve if outlying values of police per capita were excluded? (**3 points**)

   (c) *Drop cases from 1982* and focus only on observations from 1987 for the remainder of this problem. Compute the Pearson correlation coefficient for the crime rate and police per 1000 people. What is the direction of correlation between these two variables, and how strong is it? Is it statistically significant? (**3 points**)

   (d) Use the correlation coefficient from part (c) along with descriptive statistics (means and standard deviations) of crimes per 1000 people and police per 1000 people to calculate the least squares slope and intercept for the regression line between the crime rate ($y$) and police presence ($x$). Show your calculation. (**4 points**)

   (e) Now confirm your answer in part (d) by using the `regress` command in Stata. Be sure to specify the crime rate as the response variable and police per 1000 people as the explanatory variable. (**3 points**)

   (f) Provide a careful written interpretation of the intercept and slope you obtained in part (e). (**3 points**)

   (g) Based on your intercept and slope, what is your best prediction of the crime rate in a city with 2 police officers per 1000 people? With 4 police officers per 1000 people? (**2 points**)

3. In a study of time spent watching TV and GPA in high school, a simple regression resulted in the following prediction equation: $\hat{y} = 3.44 - 0.03x$. ($y$ is high school GPA and $x$ is the weekly number of hours spent watching television). (**6 points: 2 each**)

   (a) The estimated $R^2$ for this regression was 0.237. Interpret this in words.

   (b) What is the (Pearson) correlation coefficient between $y$ and $x$?

   (c) Suppose you obtained the $R^2$ from a regression that only included data for students who watched no more than 3 hours of TV per week. Would you expect the correlation to be stronger, or weaker, than for all students? Why?

4. From Github, read the dataset called *MS_NYC_SQR2017-18.dta*. This file contains annual School Quality Review data for middle schools in New York City, from 2017-18. Use this dataset to answer the following questions. (**28 points**)

```
use https://github.com/spcorcor18/LPO-8800/raw/main/data/MS_NYC_SQR2017-18.dta
```

   (a) Create a scatterplot matrix that shows the bivariate relationship between a school's proficiency rate in math (*mthprof*) and the following school variables: the mean 5th grade math score of incoming students (*incomingmath5*), the proportion of students deemed to be economically disadvantaged (*ecneed*), the proportion of teachers with 3+ years of experience (*teacherexp3_*), and the proportion of students who are chronically absent (*chronicabsent*). What do these graphs suggest to you about the relationship between middle school population characteristics and student test outcomes in math? (**4 points**)

   (b) Fit a simple regression between the school's proficiency rate in math (the $y$ variable) and the proportion of teachers with 3+ years of experience (the $x$ variable). Provide written interpretations of the following (**6 points: 2 each**):
   - The OLS intercept and slope coefficient
   - The R-squared
   - The Root MSE

   (c) Obtain the predicted values and residuals from your simple regression as two new variables in your dataset (e.g., call them *yhat* and *uhat*). Do so only for observations that were used in the estimation sample. Show that the correlation between these variables is zero. (**2 points**)

   (d) Suppose you intend to use your regression results to make inferences about the relationship between math proficiency and teacher experience in this population. Test the null hypothesis that the slope coefficient in the population is zero ($H_0 : \beta = 0$). Report your conclusion and how you obtained your answer. (**4 points**)

(e) Interpret the standard error for the slope coefficient in part (b), in words. What is it telling you? (**2 points**)

(f) Would you say the slope coefficient in part (b) is *practically significant*? That is, it the relationship between teacher experience and math proficiency educationally meaningful? Hint: think about what a meaningful change in $x$ and $y$ would be in this context. (**4 points**)

(g) Suppose your explanatory variable in part (b) were expressed in percentage point terms (0-100) rather than as a proportion. How would this affect your OLS intercept and slope coefficient? (**2 points**)

(h) Finally, fit a least squares regression between the school's proficiency rate in math and the proportion of teachers with 3+ years of experience (as in part b), but add the proportion of economically disadvantaged students (*ecneed*) as an additional explanatory variable. Provide a written interpretation of the slope coefficient for teacher experience. How has it changed from part (b)? (**4 points**)