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

December 7, 2021
61 total points

Problem Set 11 Solutions

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**)

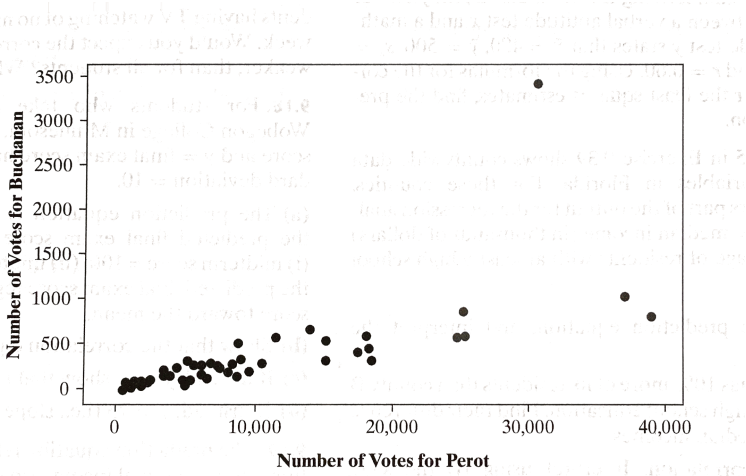


FIGURE 9.19: Scatterplot of Florida Countywide Vote for Reform Party Candidates Pat Buchanan in 2000 and Ross Perot in 1996

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**)

The response variable here is the number of Buchanan votes in a county; the explanatory variable is the number of Perot votes. The scatterplot shows that there is a roughly linear relationship between these two variables. Using the regression to predict the number of Buchanan votes using the number of Perot votes, the slope tells us that each additional Perot vote leads us to predict 0.02414 additional Buchanan votes in a county. One could also consider a 1,000 vote increase in Perot votes; this would be asso-

ciated with 24.14 additional Buchanan votes ($0.02414 * 1000$).

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

Plugging $x = 30,739$ into the prediction equation yields the following predicted value ($\hat{y}$) and residual ($\hat{u} = y - \hat{y}$):

$$\hat{y} = 45.7 + 0.02414(30,739) = 787.3$$

$$\hat{u} = y - \hat{y} = 3,407 - 787.3 = 2,619.7$$

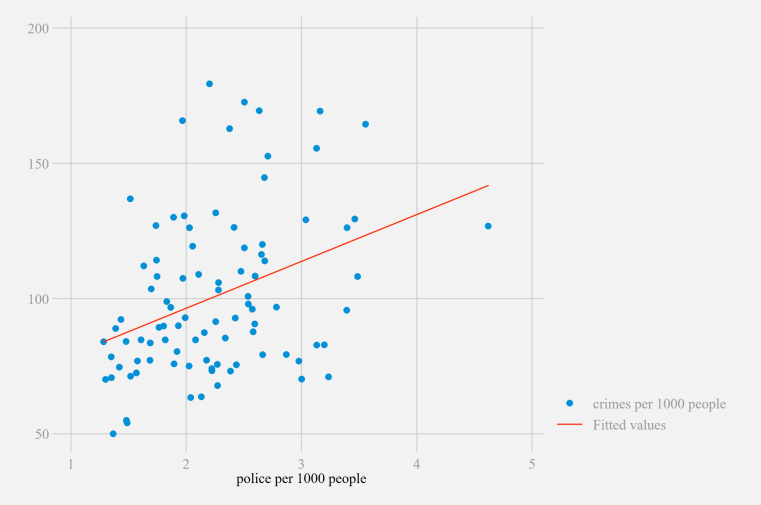
In words, Palm Beach County yielded about 2,620 more votes for Buchanan than would be predicted by the linear regression.

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
```

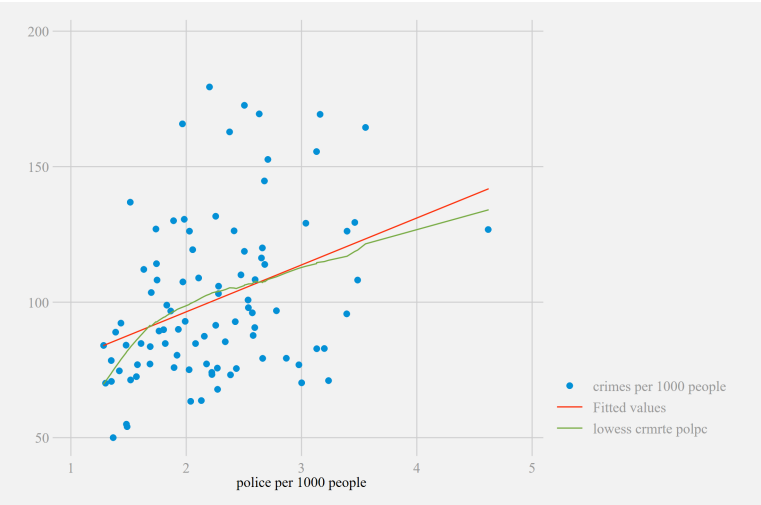
(a) Create a scatterplot that shows the relationship between a city's crime rate (crimes per 1000 people) and its investment in law enforcement, as measured by police per 1000 people. Put the crime rate on the vertical axis and police on the horizontal axis, and *overlay* the best-fit line using `twoway` 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**)

To generate the below scatterplot with best-fit line, use the command: `graph twoway (scatter crmrte polpc) (lfit crmrte polpc)`. The two variables appear moderately positively correlated, and the relationship appears roughly linear. Cities with more police per 1000 people tend to have higher rates of crime.



(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**)

The scatterplot with overlaid lowess is shown below. It appears that linearity is a reasonable approximation for this relationship. The only deviations from linearity are at the far lefthand side and for values of *polpc* above 3.5. The fit improves a bit if limiting the data to those with values below 4, although not by much.



(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**)

> The correlation coefficient is 0.3728 (below), a moderately strong positive correlation. The `sig` option of `pwcorr` reports the p-value of a test of statistical significance. With a $p = 0.0107$, the correlation is significant at the 0.05 level.

```
. pwcorr crmrte polpc,sig

             |   crmrte    polpc
-------------+------------------
      crmrte |   1.0000
             |
             |
       polpc |   0.3728   1.0000
             |   0.0107
```

(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 work. (**4 points**)

> The syntax below will store the correlation coefficient (r), mean crime rate ($\bar{y}$), standard deviation of the crime rate (s_y), mean police per 1000 people ($\bar{x}$), and the standard deviation of police per 1000 people (s_x). The slope coefficient is then calculated as $b = r * (s_y/s_x)$ and the intercept as $a = \bar{y} - b * \bar{x}$. Here $b = 19.712$ and $a = 58.865$.

```
pwcorr crmrte polpc,sig
scalar corr=r(rho)
summ crmrte
scalar sy=r(sd)
scalar ybar=r(mean)
summ polpc
scalar sx=r(sd)
scalar xbar=r(mean)
scalar b = corr*(sy/sx) // slope
display b
display ybar - b*xbar // intercept
```

(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**)

Results shown below, and confirm those in part (d).

```
. reg crmrte polpc

      Source |       SS           df       MS      Number of obs   =        46
-------------+----------------------------------   F(1, 44)        =      7.10
       Model |  7568.28963         1  7568.28963   Prob > F        =    0.0107
    Residual |  46882.2662        44  1065.50605   R-squared       =    0.1390
-------------+----------------------------------   Adj R-squared   =    0.1194
       Total |  54450.5558        45  1210.01235   Root MSE        =    32.642

------------------------------------------------------------------------------
      crmrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       polpc |   19.71158   7.396063     2.67   0.011     4.805798    34.61737
       _cons |   58.86529   17.55982     3.35   0.002      23.4758    94.25479
------------------------------------------------------------------------------
```

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

Based on our best-fit line, the crime rate when police per 1000 people is zero is predicted to be 58.865 (the intercept). For each additional police officer per 1000 people, we predict an increase in the crime rate of 19.712 crimes per 1000 people.

(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**)

Here use the prediction equation $\widehat{crmrte} = 58.865 + 19.712 * polpc$. If $polpc = 2$, $\widehat{crmrte} = 58.865 + 19.712 * 2 = 98.289$. If $polpc = 4$, $\widehat{crmrte} = 58.865 + 19.712 * 4 = 137.713$.

The Stata command `margins` can provide predicted values at specified levels of x. See the example below, which follows the `reg` command.

```
. margins, at(polpc=2)

Adjusted predictions                            Number of obs     =        46
Model VCE    : OLS

Expression   : Linear prediction, predict()
at           : polpc           =           2
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   98.28846   5.249127    18.72   0.000     87.70954    108.8674
------------------------------------------------------------------------------
```

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.

   > 23.7% of the variation in GPA can be "explained" by time spent watching TV. Alternatively, it is the reduction in prediction error that can be achieved using the predicted *GPA* rather than mean *GPA*.

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

   > In a bivariate regression, $\sqrt{R^2} = |r|$. Here, $\sqrt{0.237} = 0.487 = |r|$. We know 0.487 is the absolute value of r. Without more information we normally would not know whether $r = 0.487$ or $r = -0.487$. But in this case we know from the prediction equation that the slope is negative, so r must be negative (-0.487).

   (c) Suppose you obtained the R^2 from a regression that only included students having 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?

   > In a bivariate regression, R^2 is the square of the correlation coefficient. When computing the correlation on a restricted range of x, the correlation could be weaker or stronger than the correlation computed for the entire sample. How the correlation changes depends on how homogeneous the subset is. In this example, the relationship between GPA and TV watching may be not especially strong among kids who watch a relatively small amount of TV (0-3 hours). In other words, there may be a lot of variation in GPA among these kids that is not strongly associated with TV time. The relationship between GPA and TV time may become more evident when including kids who watch a lot of TV (over 3 hours per week). The R^2 over this restricted range is likely to be

smaller than the R^2 in the population.

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
```

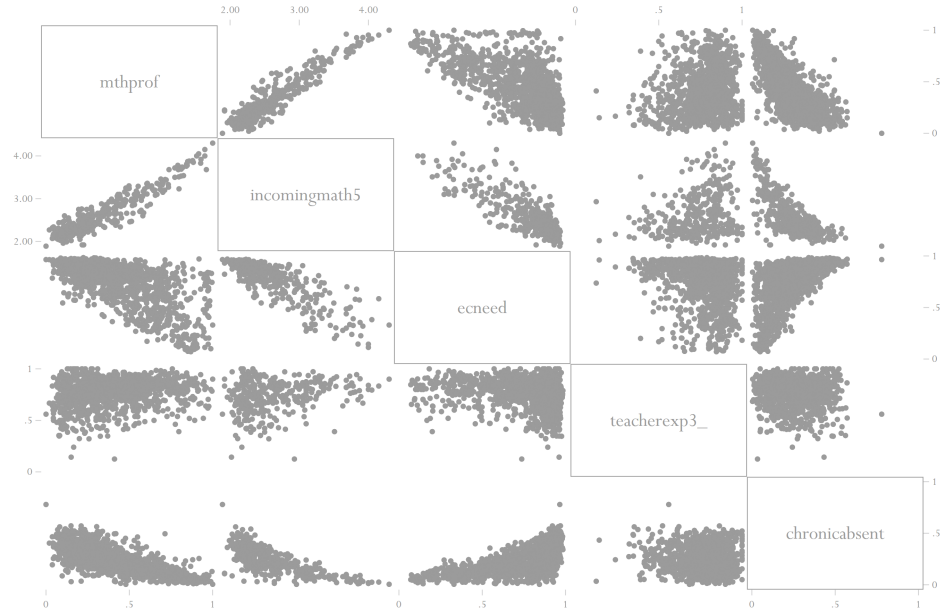
   (a) Create a scatterplot matrix that shows the bivariate relationship between a school's proficiency rate in math (*mthprof*) and the following school variables: the mean 5th grade math score of incoming students (*incomingmath5*), the proportion of students deemed to be economically disadvantaged (*ecneed*), 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**)

> The scatterplot matrix is shown below. Middle school math proficiency relates are positively related to the math scores of incoming students and the proportion of teachers with 3 or more years of experience. They are negatively related to the percent who are economically disadvantaged and who are chronically absent.

   (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

> Results are shown below. Interpretations:
>
> - **Intercept:** when none of the teachers have 3+ years of experience ($x = 0$), predicted math proficiency is 0.15, or 15%.
> - **Slope:** an increase of one unit in the proportion of teachers with 3+ years of experience is associated with a 0.335, or 33.5 percentage point, increase in the math proficiency rate. Note that a one unit change in this variable is the maximum variability, since x only ranges between 0 and 1.
> - $\mathbf{R^2}$: about 4 percent of the variation in math proficiency rates across schools is "explained" by the proportion of teachers



with 3+ years of experience.

- **Root MSE:** a measure of variability of actual math proficiency rates around the regression line. Loosely speaking (not technically correct), the average residual is about 0.22. (This is not technically correct since the average residual is always zero).

```
. reg mthprof teacherexp3_

      Source |       SS           df       MS      Number of obs   =     1,101
-------------+----------------------------------   F(1, 1099)      =     45.97
       Model |  2.31211186         1  2.31211186   Prob > F        =    0.0000
    Residual |  55.2707895     1,099  .050291892   R-squared       =    0.0402
-------------+----------------------------------   Adj R-squared   =    0.0393
       Total |  57.5829014     1,100  .052348092   Root MSE        =    .22426

------------------------------------------------------------------------------
     mthprof |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
teacherexp3_ |   .3346409   .0493541     6.78   0.000     .2378019    .4314799
       _cons |   .1504952   .0385609     3.90   0.000     .0748339    .2261566
------------------------------------------------------------------------------
```

(c) Obtain the predicted values and residuals from your simple regression as two new variables in your dataset (e.g., *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**)

```
. predict yhat if e(sample), xb
(185 missing values generated)

. predict uhat if e(sample), r
(185 missing values generated)

. corr yhat uhat
(obs=1,101)

             |     yhat     uhat
-------------+------------------
        yhat |   1.0000
        uhat |   0.0000   1.0000
```

(d) Suppose you intend to use your regression results to make inferences about the population relationship between math proficiency and teacher experience. 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**)

> There are several ways to use the above regression output for this. The t-statistic and p-value next to the slope coefficient are for exactly this test. Because $p < 0.05$, we can reject the null hypothesis. The slope coefficient is statistically different from zero. Another way to conduct this test is to examine the 95% confidence interval, which can be used for 2-tailed hypotheses tests. Zero does not appear in the confidence interval, we can reject the null hypothesis $H_0 : \beta = 0$.

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

> The standard error is an estimate of the sampling variability in the slope coefficient across repeated random samples. It is a measure of how much one might expect the slope to vary from sample to sample, simply by chance.

(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**)

> The slope of 0.335 above is a rather large change in math proficiency rates. (The standard deviation of *mthprof* is 0.236, so it is 142% of a standard deviation). However, a change from

zero to one in the proportion of teachers with 3+ years of experience is also extreme. (The standard deviation of *teacherexp3_* is 0.137). A more meaningful thought experiment might be to increase *teacherexp3_* by one standard deviation (0.137). This would correspond to a $0.137 \times 0.335 = 0.045$ increase in math proficiency. Expressed as a percentage of the standard deviation in math proficiency across schools, this is a $0.045/0.236 = 0.19$ effect. I would say this is still an educationally significant effect.

Note you can get coefficients expressed directly in these units by adding the `beta` option to `regress`. See below.

```
. reg mthprof teacherexp3_, beta

      Source |       SS           df       MS      Number of obs   =     1,101
-------------+----------------------------------   F(1, 1099)      =     45.97
       Model |  2.31211186         1  2.31211186   Prob > F        =    0.0000
    Residual |  55.2707895     1,099  .050291892   R-squared       =    0.0402
-------------+----------------------------------   Adj R-squared   =    0.0393
       Total |  57.5829014     1,100  .052348092   Root MSE        =    .22426

------------------------------------------------------------------------------
     mthprof |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
teacherexp3_ |   .3346409   .0493541     6.78   0.000                .2003815
       _cons |   .1504952   .0385609     3.90   0.000                       .
------------------------------------------------------------------------------
```

(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**)

Results are shown below. Intuitively, whereas before a 1-unit change in x was a 100 percentage point change, now a 1-unit change is only a 1 percentage point change. The coefficient is now 1/100th of what it was previously.

```
. gen teacherexp2=teacherexp3_*100
(179 missing values generated)

. reg mthprof teacherexp2

      Source |       SS           df       MS      Number of obs   =     1,101
-------------+----------------------------------   F(1, 1099)      =     45.97
       Model |  2.31211181         1  2.31211181   Prob > F        =    0.0000
    Residual |  55.2707896     1,099  .050291892   R-squared       =    0.0402
-------------+----------------------------------   Adj R-squared   =    0.0393
```

```
       Total |  57.5829014     1,100  .052348092   Root MSE        =    .22426

------------------------------------------------------------------------------
     mthprof |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 teacherexp2 |   .0033464   .0004935     6.78   0.000      .002378    .0043148
       _cons |   .1504952   .0385609     3.90   0.000     .0748339    .2261566
------------------------------------------------------------------------------
```

(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**)

> This is a multiple regression, in which coefficients are interpreted conditional on holding other explanatory variables constant. The slope coefficient in this case is smaller (0.088) than that in part b (0.335). Here, a one-unit change in the proportion of teachers with 3+ years of experience is associated with an 8.8 percentage point increase in math proficiency rates, *holding constant* the school's level of economic need. Intuitively, a good portion of the association between student achievement in math and the level of teaching experience in the school (in part b) was likely driven by the characteristics of students attending the school. If high-poverty schools—which tend to have lower average test scores for other reasons—have less-experienced teachers, the original association may be spurious, at least in part.

```
. reg mthprof teacherexp3_ ecneed

      Source |       SS           df       MS      Number of obs   =     1,101
-------------+----------------------------------   F(2, 1098)      =    712.27
       Model |  32.5184935         2  16.2592468   Prob > F        =    0.0000
    Residual |  25.0644078     1,098   .02282733   R-squared       =    0.5647
-------------+----------------------------------   Adj R-squared   =    0.5639
       Total |  57.5829014     1,100  .052348092   Root MSE        =    .15109

------------------------------------------------------------------------------
     mthprof |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
teacherexp3_ |   .0881086   .0339345     2.60   0.010     .0215249    .1546923
      ecneed |  -.7578644   .0208339   -36.38   0.000    -.7987431   -.7169857
       _cons |    .897821   .0331207    27.11   0.000     .8328339     .962808
------------------------------------------------------------------------------
```