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

Due November 11, 2024
44 total points

Problem Set 9

Instructions: Answer the following questions in their entirety in a separate document. Submit your completed problem set as a PDF document 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. The Zagat Restaurant Guides rate each restaurant on a 30-point scale for food, decor, service, and cost. The *zagat.dta* file on Github shows 2007 ratings for Italian restaurants in Boston, London, and New York. (**12 points—4 each**)

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

   (a) Conduct a correlation analysis to describe the associations between the food quality rating and the ratings for decor, service, and cost, *for restaurants in Boston*. Provide scatterplots and report correlation coefficients, and describe what you find. Which correlations are the strongest?

   (b) Repeat the analysis for London and New York, separately. How do the correlations compare with those in Boston?

   (c) Now create a scatterplot that shows the relationship between food quality rating (on the vertical axis) and the cost rating (on the horizontal axis), using all three cities' data combined. Use separate colors to differentiate data points for each city. (Hint: the `twoway` command allows you to plot multiple scatterplots on the same graph. In the drop-down menus, use Graphics → Twoway). How does the correlation compare when using all cities' data, versus the individual cities?

2. Read the dataset called *states.dta* on Github. This file contains some education-related data from the early 2000s for the 50 states plus D.C. (**6 points—3 each**)

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

   (a) Create a scatterplot showing the relationship between the average teacher's salary (*teachpay*) and education expenditure per pupil (*educexpe*). Label the data points in your scatter plot with the state name. (Hint: when creating the graph using Graphics → Twoway, you have the ability to create marker labels under "marker properties").

(b) Identify a state that appears to be an outlier with respect to the general relationship between teacher salaries and expenditure. In what way is this state unusual?

3. Continue with the same dataset from #2, *states.dta*. Conduct a correlation analysis to determine whether various factors are associated with the average educational expenditure per pupil. These factors are: student teacher ratio (*stuteach*), the average verbal SAT score for the state (*satv*), and the average teacher salary (*teachpay*). (**10 points**)

   (a) Create a scatterplot matrix for these variables. In which cases, if any, does the association appear linear? (**3 points**)

   (b) Create a correlation matrix between all pairs of these variables. Describe in words how you see the correlation between educational expenditure per pupil and the other three variables. (**3 points**)

   (c) Of the three variables, which is the most strongly correlated with expenditure per pupil? (**1 points**)

   (d) Which, if any, of the correlations can be considered statistically significant? Explain how you know. (**3 points**)

4. Begin with the "toy" dataset created by the following Stata syntax. For each pair of variables (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4), create a scatterplot and calculate both the Pearson and Spearman correlation coefficients. For each pair, explain why and how the two correlations differ, or why they do not. In which cases are the Spearman correlations particularly useful? (**5 points**)

```
set seed 1234
clear
set obs 250
gen x1 = runiform(-3,0)
gen y1 = x1^4

gen x2 = runiform(0,3)
gen y2 = x2^4

gen x3 = runiform(0,3)
gen y3 = 10 + 5*x3

gen x4 = runiform(0,3)
gen y4 = rnormal(0,5)
```

5. Begin with the dataset *card.dta* on Github. These data come from a study by David Card (1995) that estimated the earnings returns to additional years of education. (**11 points**)

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

   (a) Using the full dataset, calculate the correlation between years of education (*educ*) and the individuals log annual earnings (*lwage*). Inspect the scatterplot to determine whether the correlation coefficient is appropriate for these variables. For the purposes of parts (b)-(c), consider these data the population. (**3 points**)

   (b) Now run the following syntax. Based on what you have done in previous assignments, explain in words what this code is doing. (**3 points**)

```
bootstrap r(rho), size(50) reps(100) saving(results): corr educ lwage
```

   (c) Using the resulting dataset in part (b), produce a histogram for "rho" and a set of descriptive statistics. What is the histogram showing you? Report an empirical 95% "confidence interval" for ρ based on your descriptive statistics. In other words, between what values did the correlation coefficient fall 95% of the time? (**5 points**)