Statistical Methods in Education Research
Vanderbilt University November 14, 2024
Prof. Sean P. Corcoran **44 total points**

Problem Set 9 Solutions

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.

   See attached log. All of the variables are positively correlated, although some of the correlations are stronger than others. The strongest observed correlations are between cost and decor (0.824) and cost and service (0.718). The weakest correlation is between decor and food quality (0.293). Interestingly, the correlation between cost and decor (or service) is much stronger than the correlation between cost and food quality (0.411).

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

   See attached log. Some correlations in London and NYC are stronger than Boston while others are weaker.

   (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?

   See attached log. This dataset provides a good example of how variables can be strongly correlated within subgroups (city) but not overall. The accompanying scatterplots illustrate this. Within each city, there is a positive correlation between food quality and price. When combining the data, however, the correlation is weak (r=0.012). This is because restaurants in London are more expensive in general. The pooled data, then, includes low-cost high-quality restaurants in Boston and higher-cost lower-quality restaurants in London.

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").

   See attached results. As the scatterplot shows, there is a positive, linear relationship between average teacher salary and expenditure per pupil.

   (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?

   California is somewhat of an outlier. It has higher average salaries than one might predict given their expenditure per student. Vermont is also an outlier in the other direction; it has lower average salaries than one might predict given their expenditure per student.

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**)

   See attached results. The first column of the scatterplot matrix, which shows the scatter of expenditure per pupil against the other three variables, suggests that each of these relationships is roughly linear.

   (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**)

   See attached results. There is a positive correlation between expenditures per student and average teacher salaries (r=0.688), a negative correlation between SAT verbal scores and average teacher salaries (r=-0.408), and a negative correlation between expenditures per student and the student-teacher ratio (r=-0.461).

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

The strongest correlation is between expenditures per student and teacher salaries, with a correlation (in absolute value) of 0.688.

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

In Stata use the `pwcorr` command with the `sig` option. In column (1), the *p*-value is shown below the correlation coefficient. In all three cases this is lower than $\alpha = 0.05$, and thus are statistically significant.

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**)

   See the attached results. In the first case, there is a negative *non-linear* association between *x* and *y*. The Pearson correlation is -0.866. The Spearman correlation is a perfect -1, however, since *y* is an exact non-linear function of *x*. (The Spearman correlation uses ranks rather than actual values, and there is a perfect negative correspondence in the ranks). The second case is identical to the first, except that the non-linear association is positive. In the third case, the association is linear and positive, and thus the Pearson and Spearman produce the same result. In the fourth case, there is no association in the population between the two variables, though there is incidental correlation in the sample. The Pearson and Spearman are identical here.

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**)

In the full dataset (N=3,010), the correlation is 0.3142. The scatterplot is shown in the attached log. While the variation in log earnings increases with education, the association is approximately linear.

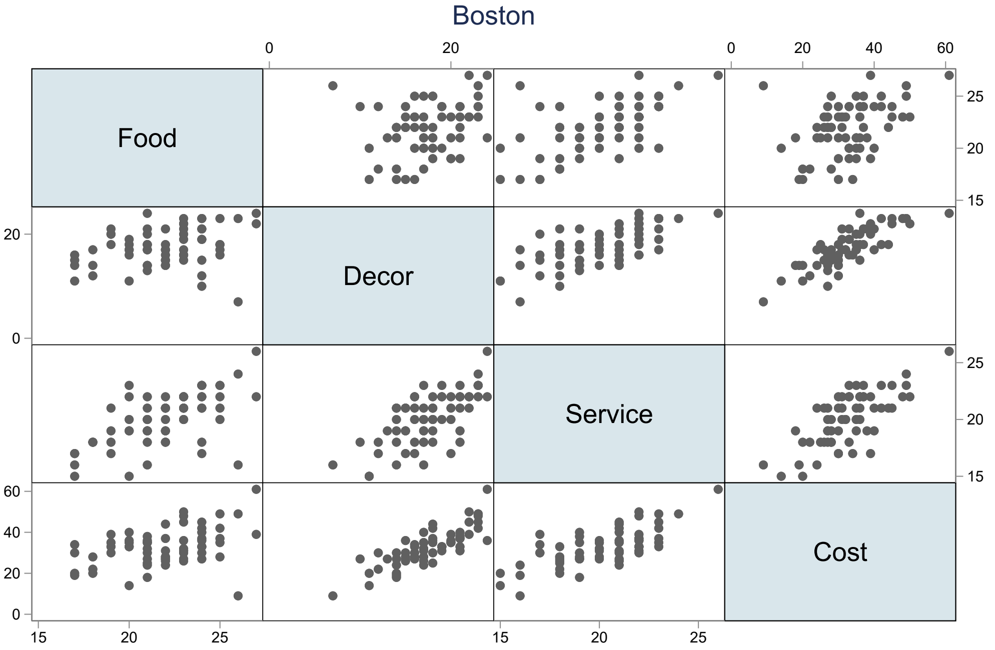
(b) Now run the following syntax. Based on what you have done in previous assignments, explain in words what this code is doing. (**3 points**)

The bootstrap prefix is drawing 100 random samples of 50 (with replacement) and in each, calculating the correlation between *lwage* and *educ*. It stores the results in a file named *results.dta*.

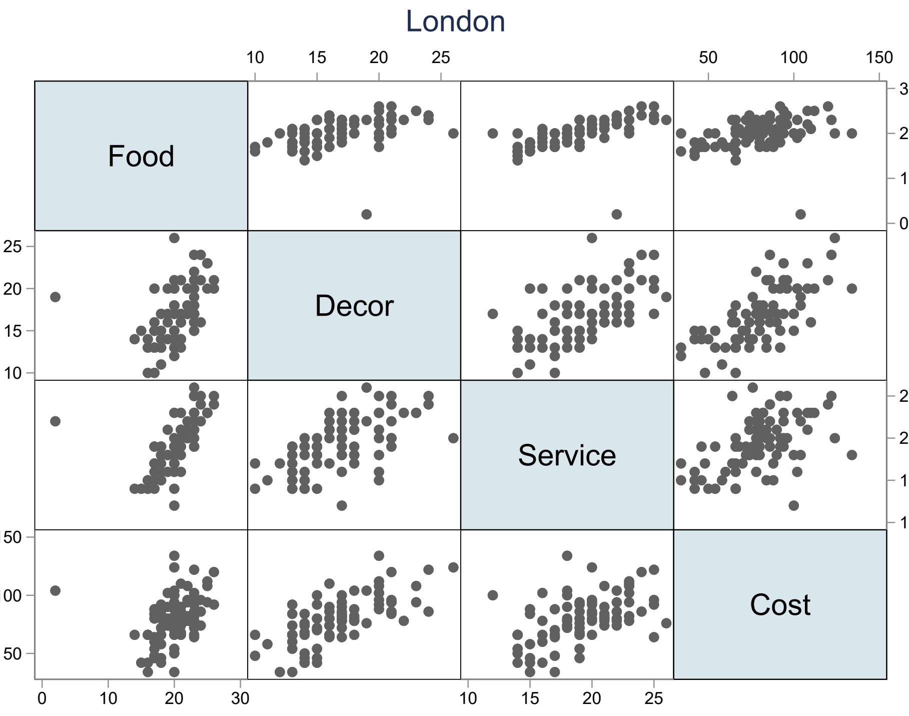
(c) Using the resulting dataset in part (b), produce a histogram for "rho" and a set of descriptive statistics. What is the histogram showing you? Report an empirical 95% "confidence interval" for ρ based on your descriptive statistics. In other words, between what values did the correlation coefficient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

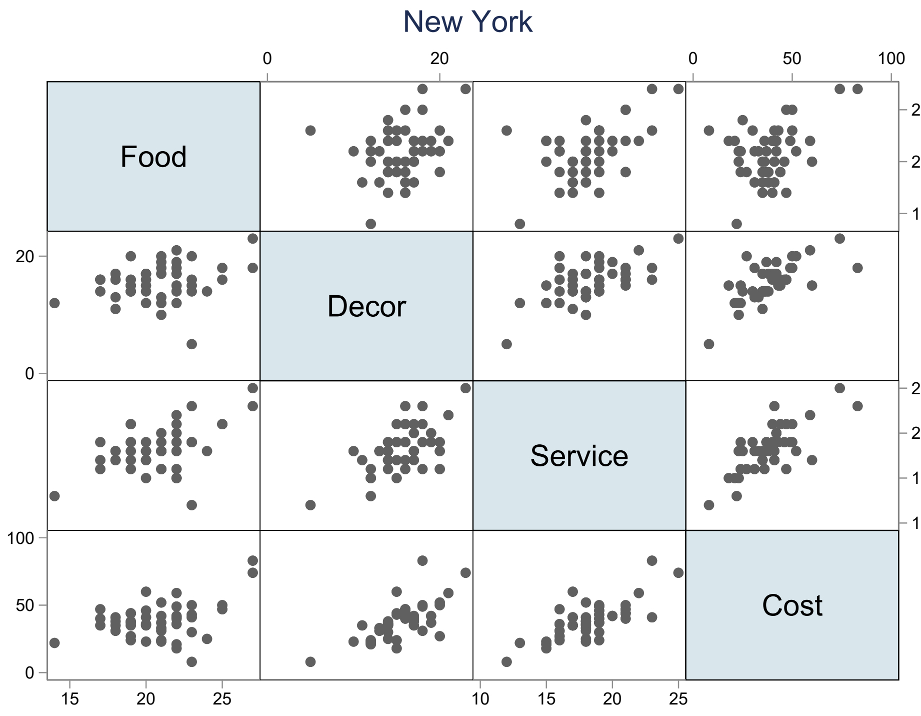
The histogram and descriptive statistics for the 100 sample correlation coefficients are shown below. The mean of these is 0.33, not far from the known correlation in the full data of 0.3142. 95% of the correlations fall between 0.030 (the 2.5th percentile) and 0.58 (the 97.5th percentile).

```
.
. // ********************************************************************
. // LPO.8800 Problem Set 10 - Solutions
. // Last updated: November 15, 2021
. // ********************************************************************
.
. // *************
. // Question 1
. // *************
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/zagat.dta
.
. *****
. * a
. *****
. corr food decor service cost if city=="Boston"
(obs=64)
             |     food    decor  service     cost
-------------+------------------------------------
        food |   1.0000
       decor |   0.2929   1.0000
     service |   0.6170   0.6817   1.0000
        cost |   0.4108   0.8235   0.7179   1.0000
. graph matrix food decor service cost if city=="Boston", scheme(Modern) ///        title(
> "Boston") name(q1a, replace)
. graph export q1a.pdf, as(pdf) replace
file q1a.pdf saved as PDF format
```

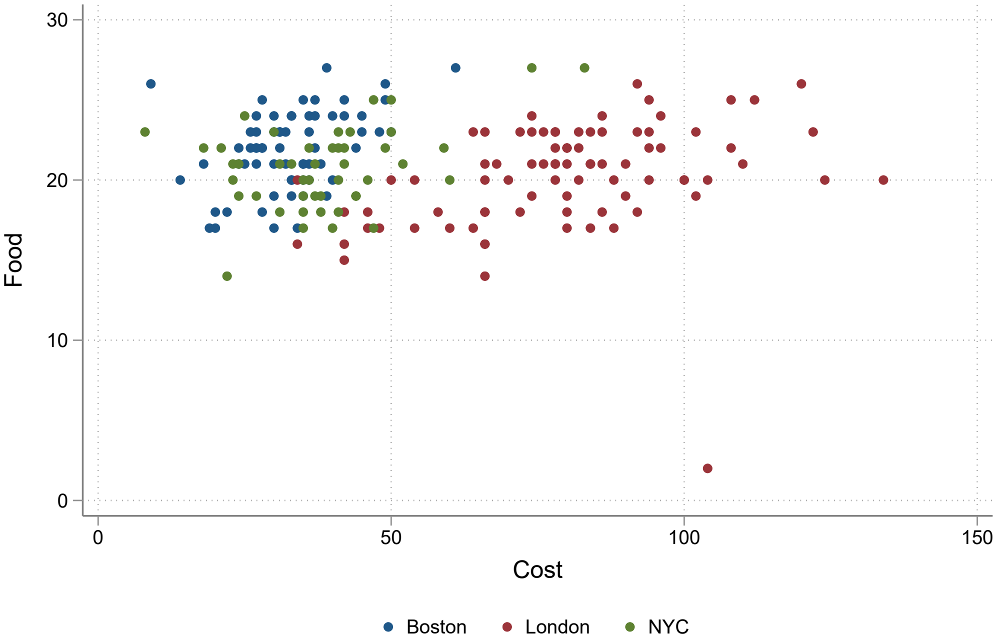


```
.
. *****
. * b
. *****
. corr food decor service cost if city=="London"
(obs=83)
             |     food    decor  service     cost
-------------+------------------------------------
        food |   1.0000
       decor |   0.4388   1.0000
     service |   0.5846   0.6091   1.0000
        cost |   0.3310   0.6992   0.4928   1.0000
. graph matrix food decor service cost if city=="London", scheme(Modern) ///        title(
> "London") name(q1b1, replace) nodraw
.
. corr food decor service cost if city=="NY"
(obs=46)
             |     food    decor  service     cost
-------------+------------------------------------
        food |   1.0000
       decor |   0.2858   1.0000
     service |   0.4987   0.6204   1.0000
        cost |   0.3893   0.6925   0.7382   1.0000
. graph matrix food decor service cost if city=="NY", scheme(Modern) ///        title("New
> York") name(q1b2, replace) nodraw
.
. graph combine q1b1 q1b2, xsize(4) ysize(6) col(1) scheme(Modern) name(q1b, replace)
. graph export q1b.pdf, as(pdf) replace
file q1b.pdf saved as PDF format
```

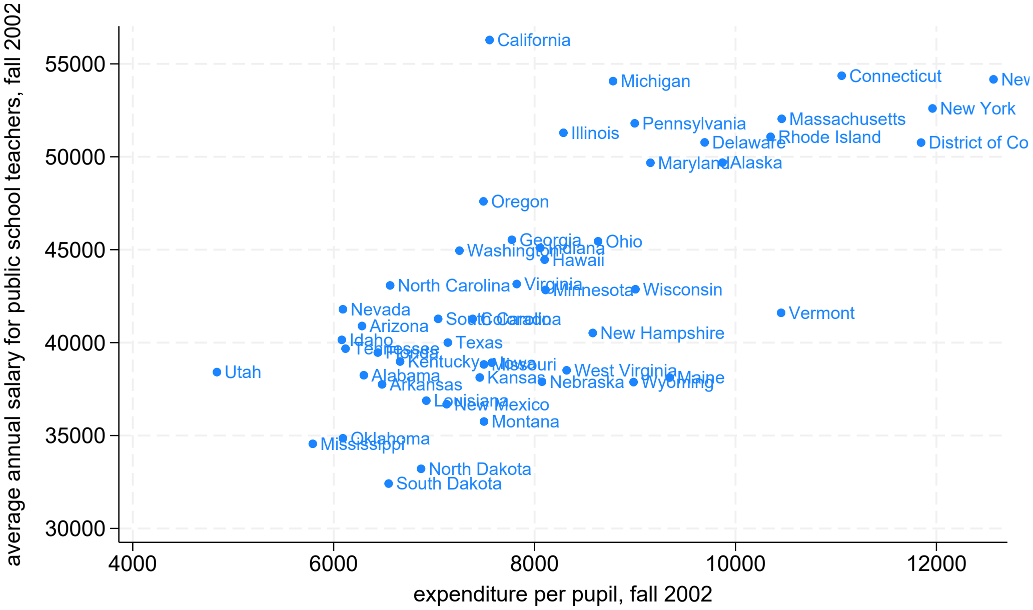

London
Food
Decor
Service
Cost
10 15 20 25
50 100 150
30 20 10 0
25 20 15 10
25 20 15 10
150 100 50
0 10 20 30
10 15 20 25


New York
Food
Decor
Service
Cost
0 20
0 50 100
25 20 15
20 0
25 20 15 10
100 50 0
15 20 25
10 15 20 25

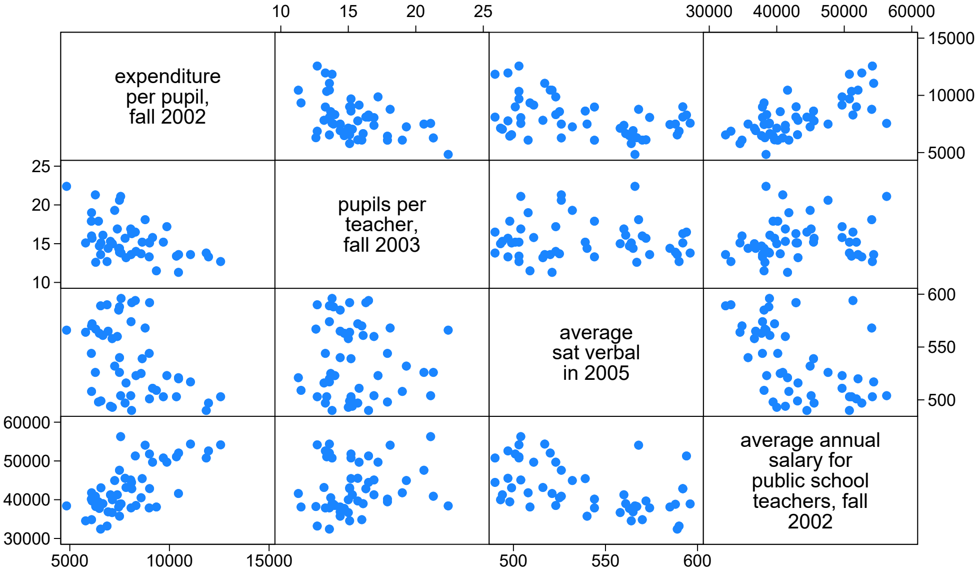
```
.
. *****
. * c
. *****
. corr food decor service cost
(obs=193)
             |     food    decor  service     cost
-------------+------------------------------------
        food |   1.0000
       decor |   0.3659   1.0000
     service |   0.5780   0.6412   1.0000
        cost |   0.0166   0.3753   0.2883   1.0000
.
. twoway (scatter food cost if city=="Boston") ///        (scatter food cost if city=="Lon
> don") ///        (scatter food cost if city=="NY"), ///        legend(order(1 "Boston" 2
> "London" 3 "NYC") rows(1)) legend(pos(6)) ///        scheme(Modern) name(q1c, replace)
. graph export q1c.pdf, as(pdf) replace
file q1c.pdf saved as PDF format
```



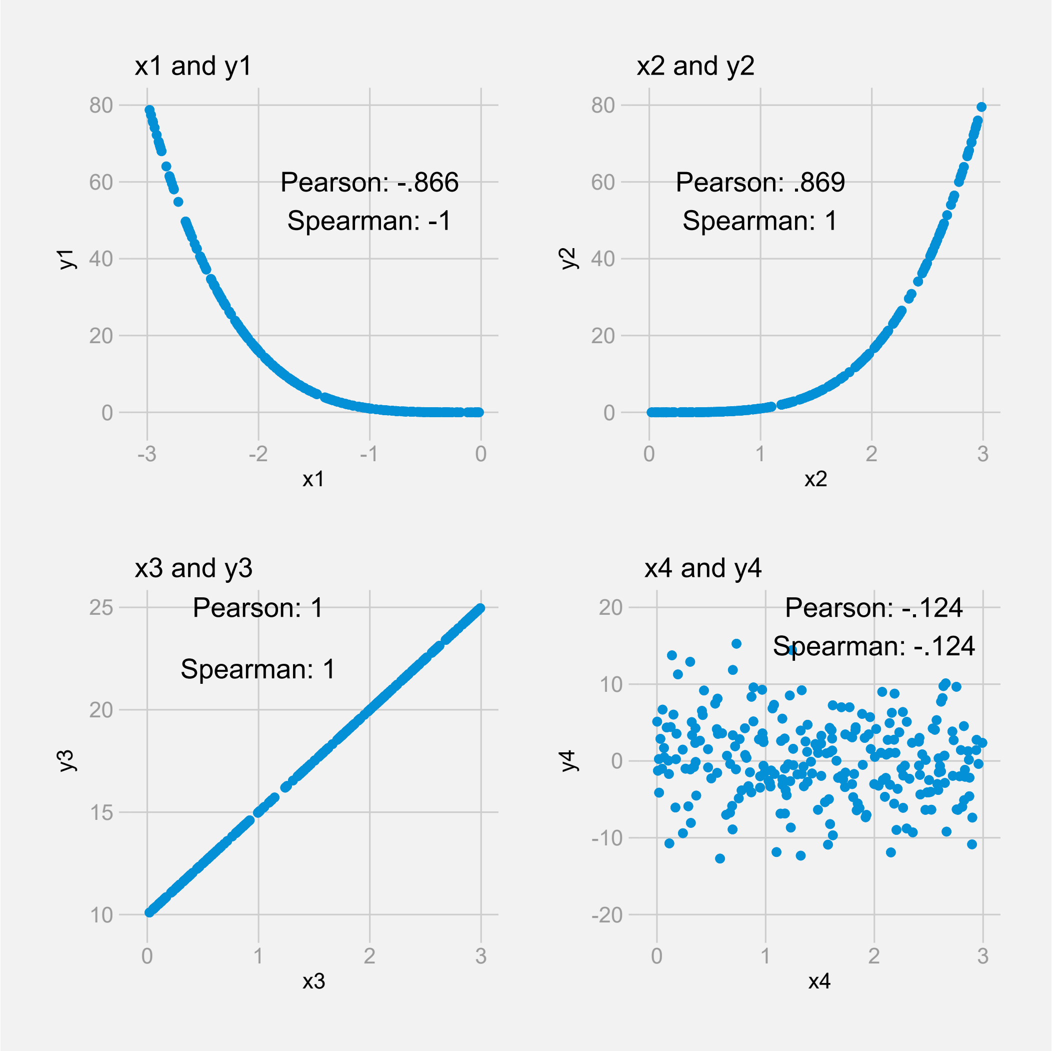
```
.
.
. // *************
. // Question 2
. // *************
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/states.dta, clear
.
. *****
. * a
. *****
. twoway (scatter teachpay educexpe, mlabel(state)), name(q2a, replace)
. graph export q2a.pdf, as(pdf) replace
file q2a.pdf saved as PDF format
```



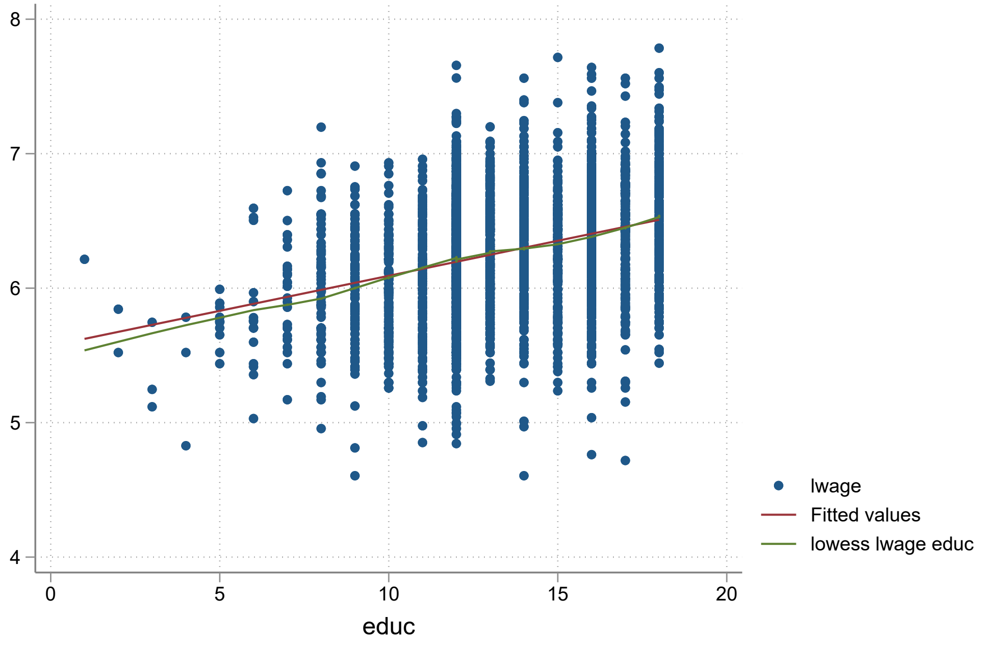
```
.
.
. // *************
. // Question 3
. // *************
. use https://github.com/spcorcor18/LPO-8800/raw/main/data/states.dta, clear
.
. *****
. * a
. *****
. graph matrix educexpe stuteach satv teachpay, name(q3a, replace)
. graph export q3a.pdf, as(pdf) replace
file q3a.pdf saved as PDF format
```



```
.
. *****
. * b
. *****
. corr educexpe stuteach satv teachpay
(obs=51)
             | educexpe stuteach     satv teachpay
-------------+------------------------------------
    educexpe |   1.0000
    stuteach |  -0.4613   1.0000
        satv |  -0.4077  -0.0456   1.0000
    teachpay |   0.6876   0.1618  -0.5009   1.0000
.
. *****
. * d
. *****
. pwcorr educexpe stuteach satv teachpay,sig
             | educexpe stuteach     satv teachpay
-------------+------------------------------------
    educexpe |   1.0000
             |
             |
    stuteach |  -0.4613   1.0000
             |   0.0007
             |
        satv |  -0.4077  -0.0456   1.0000
             |   0.0030   0.7505
             |
    teachpay |   0.6876   0.1618  -0.5009   1.0000
             |   0.0000   0.2565   0.0002
             |
.
.
. // *************
. // Question 4
. // *************
. set seed 1984
.
. clear
. set obs 250
Number of observations (_N) was 0, now 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)
.
. corr y1 x1
(obs=250)
             |       y1       x1
-------------+------------------
          y1 |   1.0000
          x1 |  -0.8656   1.0000
.   local r1=round(r(rho),0.001)
. spearman y1 x1
Number of observations =     250
        Spearman’s rho = -1.0000
Test of H0: y1 and x1 are independent
                  Prob =  0.0000
.   local s1=round(r(rho),0.001)
. twoway (scatter y1 x1),text(60 -1 "Pearson: ‘r1’") text(50 -1 "Spearman: ‘s1’") ///
> title("x1 and y1") scheme(538) name(g1, replace)
.
. corr y2 x2
(obs=250)
             |       y2       x2
-------------+------------------
          y2 |   1.0000
          x2 |   0.8687   1.0000
.   local r2=round(r(rho),0.001)
. spearman y2 x2
Number of observations =    250
        Spearman’s rho = 1.0000
Test of H0: y2 and x2 are independent
                  Prob = 0.0000
.   local s2=round(r(rho),0.001)
. twoway (scatter y2 x2),text(60 1 "Pearson: ‘r2’") text(50 1 "Spearman: ‘s2’") ///
> title("x2 and y2") scheme(538) name(g2, replace)
.
. corr y3 x3
(obs=250)
             |       y3       x3
-------------+------------------
          y3 |   1.0000
          x3 |   1.0000   1.0000
.   local r3=round(r(rho),0.001)
. spearman y3 x3
Number of observations =    250
        Spearman’s rho = 1.0000
Test of H0: y3 and x3 are independent
                  Prob = 0.0000
.   local s3=round(r(rho),0.001)
. twoway (scatter y3 x3),text(25 1 "Pearson: ‘r3’") text(22 1 "Spearman: ‘s3’") ///
> title("x3 and y3") scheme(538) name(g3, replace)
.
. corr y4 x4
(obs=250)
             |       y4       x4
-------------+------------------
          y4 |   1.0000
          x4 |  -0.1240   1.0000
.   local r4=round(r(rho),0.001)
. spearman y4 x4
Number of observations =     250
        Spearman’s rho = -0.1243
Test of H0: y4 and x4 are independent
                  Prob =  0.0496
.   local s4=round(r(rho),0.001)
. twoway (scatter y4 x4),text(20 2 "Pearson: ‘r4’") text(15 2 "Spearman: ‘s4’") ///
> title("x4 and y4") scheme(538) name(g4, replace)
.
. graph combine g1 g2 g3 g4, rows(2) xsize(8) ysize(8) scheme(538)
. graph export q4.pdf, as(pdf) replace
file q4.pdf saved as PDF format
```



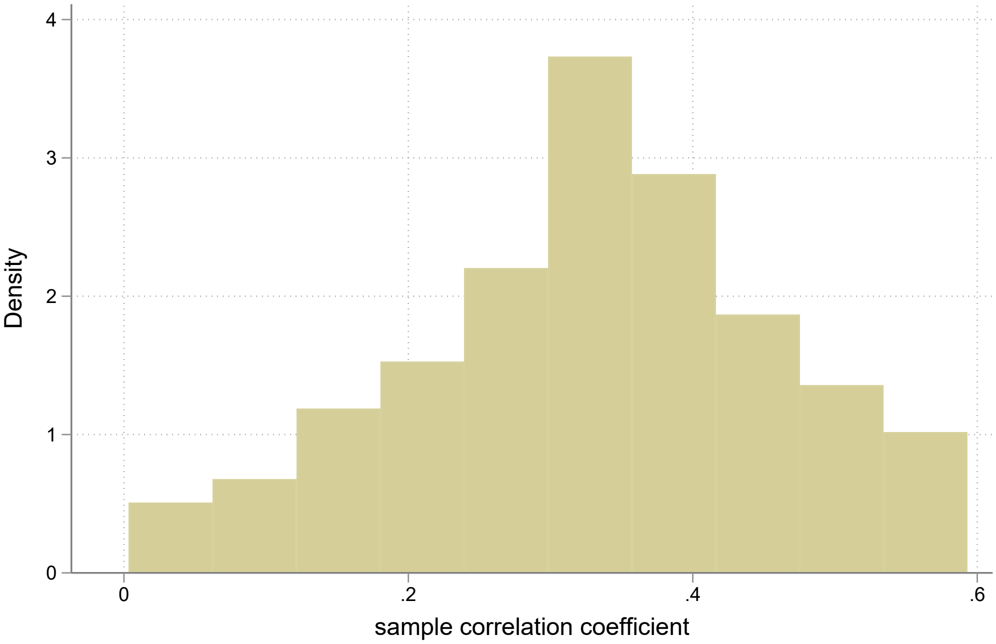
```
.
.
. // *************
. // Question 5
. // *************
.
. use "http://fmwww.bc.edu/ec-p/data/wooldridge/card.dta", clear
.
. *****
. * a
. *****
. corr lwage educ
(obs=3,010)
             |    lwage     educ
-------------+------------------
       lwage |   1.0000
        educ |   0.3142   1.0000
. scatter lwage educ, scheme(Modern)
. twoway (scatter lwage educ) (lfit lwage educ) (lowess lwage educ), scheme(Modern) ///
> name(q5a, replace)
. graph export q5a.pdf, as(pdf) replace
file q5a.pdf saved as PDF format
```



```
.
. *****
. * b
. *****
. bootstrap r(rho), size(50) reps(100) saving(results,replace): corr educ lwage
(running correlate on estimation sample)
warning: correlate does not set e(sample), so no observations will be excluded
         from the resampling because of missing values or other reasons. To
         exclude observations, press Break, save the data, drop any
         observations that are to be excluded, and rerun bootstrap.
Bootstrap replications (100): .........10.........20.........30.........40.........50.....
> ....60.........70.........80.........90.........100 done
Bootstrap results                                       Number of obs = 3,010
                                                        Replications  =   100

      Command: correlate educ lwage
        _bs_1: r(rho)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _bs_1 |    .3142236   .1324583     2.37   0.018     .0546102    .5738371
------------------------------------------------------------------------------
.
. *****
. * c
. *****
. use results, clear
(bootstrap: correlate)
. rename _bs_1 rho
. summm rho, detail
                           r(rho)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0034594       .0033723
 5%     .0955085       .0035465
10%     .1459085       .0534746       Obs                 100
25%     .2473789       .0705412       Sum of wgt.         100
50%     .3478006                      Mean           .3300923
                        Largest       Std. dev.      .1324583
75%     .4143306       .5763696
90%     .5123569       .5777606       Variance       .0175452
95%     .5661785       .5831605       Skewness      -.2236808
99%     .5881545       .5931484       Kurtosis       2.785801
. centile rho, c(2.5)
                                                          Binom. interp.
    Variable |       Obs  Percentile    Centile        [95% conf. interval]
-------------+-------------------------------------------------------------
         rho |       100         2.5    .0297587        .0033723    .1118404*
* Lower (upper) confidence limit held at minimum (maximum) of sample
. centile rho, c(97.5)
                                                          Binom. interp.
    Variable |       Obs  Percentile    Centile        [95% conf. interval]
-------------+-------------------------------------------------------------
         rho |       100        97.5    .5803255        .5435408    .5931484*
* Lower (upper) confidence limit held at minimum (maximum) of sample
. histogram rho, xtitle("sample correlation coefficient") scheme(Modern) ///        name(q
> 5c, replace)
(bin=10, start=.00337233, width=.05897761)
. graph export q5c.pdf, as(pdf) replace
file q5c.pdf saved as PDF format
```



```
.
.
. capture log close
```