HW04 - Correlation and Regression Method

Stat 20 & 131A, Spring 2017, Prof. Sanchez

Due Feb-16

1) A class of 15 students happens to include 5 basketball players. True or False, and explain: the relationship between heights and weights for this class should be summarized using r. $_{0.2pts}$

2) The correlation between height and weight among men age 18-74 in the U.S. is about 0.40. Say whether each conclusion below follows from the data; explain your answer. $_{0.4pts}$

a. Taller men tend to be heavier.

b. The correlation between weight and height for men age 18-74 is about 0.40.

c. Heavier men tend to be taller.

d. If someone eats more and puts in 10 pounds, he is likely to get somewhat taller.

3) On a multiple-choice exam, there are 100 problems. Let X be the number of problems a student got right, and Y the number a student got wrong. If the average and SD of X is 60 and 10, respectively, find: $_{0.8pts}$

a. The average and SD of Y.

b. What is the correlation between X and Y?

4) Many studies have found an association between gas prices and car accidents (high gas prices lead to fewer auto accidents). One study found an association between gas prices and traffic congestion. Should you conclude that the decreasing cost of gas causes more traffic jams? Or can you explain a rising of traffic congestion in some other way? $_{0.5pts}$

5) Consider the following three variables: $_{1pt}$

```
x = c(1, 2, 3, 4, 5, 6, 7)
y = c(2, 1, 4, 3, 7, 5, 6)
z = c(5, 4, 7, 6, 10, 8, 9)
```

One of your friends has manually calculated the following correlation coefficients:

- $cor(x, y) = 0.8214$
- $cor(x, z) = 0.7610$

Your friend asks you to use R in order to check whether the calculated correlations are correct.

a. Are both correlation coefficients correct?

b. What is the reason that explains your results in part a).

6) The following table shows per capita consumption of cigerettes in various countries in 1930, and the death rates from lung cancer for men in 1950. $_{1pts}$

country	consumption	deaths
Australia	480.00	180.00
Canada	500.00	150.00
Denmark	380.00	170.00
Finland	1100.00	350.00
Great Britain	1100.00	460.00
Iceland	230.00	60.00
Netherlands	490.00	240.00
Norway	250.00	90.00
Sweden	300.00	110.00
Switzerland	510.00	250.00
US	1300.00	200.00

a. Use R to create two variables (i.e. R vectors) `consumption` and `deaths`

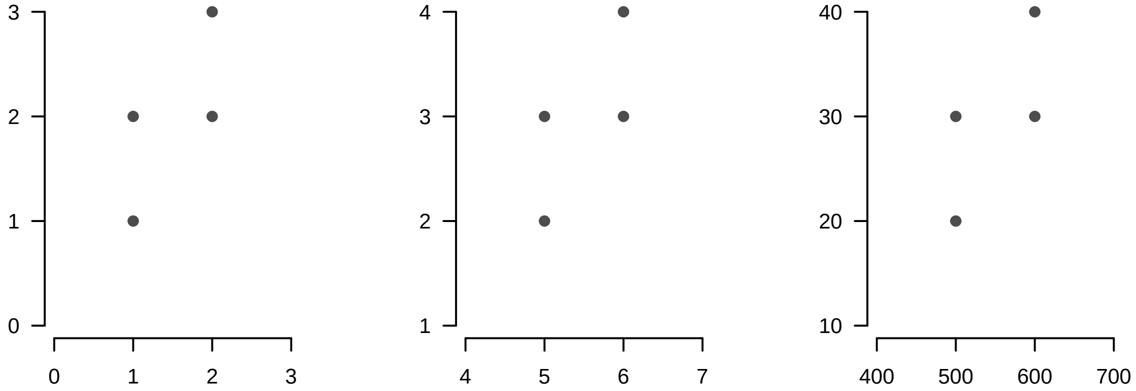
b. Use R to plot a scatter diagram for these data.

c. Use `cor()` to find the correlation coefficient for these data.

d. True or False: the higher cigarette consumption was in 1930 in one of these countries, on the whole the higher the death rate from lung cancer in 1950. Or can this be determined from the data?

e. True or False: death rates from lung cancer tend to be higher among those persons who smoke more. Or can this be determined from the data?

7) Below are three scatter diagrams. Do they have the same correlation? $_{0.3pts}$



8) An investigator collected data on heights and weights of college students; results can be summarized as follows. $_{0.4pts}$

	Average	SD
Men's height	70 inches	3 inches
Men's weight	144 pounds	21 pounds
Women's height	64 inches	3 inches
Women's weight	120 pounds	21 pounds

The correlation coefficient between height and weight for the men was about 0.60; for the women, it was about the same. If you take the men and women together, the correlation between height and weight would be ________.

a. just about 0.60.

b. somewhat lower.

c. somewhat higher.

9) At the University of California, Berkeley, Statistics 2 is a large lecture course with small discussion sections led by teaching assistants. As part of a study, at the second-to-last lecture one term, the students were asked to fill out anonymous questionnaires rating the effectiveness of their teaching assistants (by name), and the course, on the scale: $_{1.85pts}$

```
1 = poor,  2 = fair,  3 = good,  4 = very good,  5 = excellent
```

The following statistics were computed:

- The average rating of the assistant by the students in each section.
- The average rating of the course by the students in each section.
- The average score on the final for the students in each section.

Results are shown below (sections are identified by letter).

section	assistant	course	final
A	3.30	3.50	70.00
B	2.90	3.20	64.00
C	4.10	3.10	47.00
D	3.30	3.30	63.00
E	2.70	2.80	69.00
F	3.40	3.50	69.00
G	2.80	3.60	69.00
H	2.10	2.80	63.00
I	3.70	2.80	53.00
J	3.20	3.30	65.00
K	2.40	3.30	64.00

a. Use R to create three vectors: `assistant`, `course`, and `final`

b. Find the correlations: 1) assistant rating -vs- course rating, 2) assistant rating -vs- final score, and 3) course rating -vs- final score.

c. Plot scatter diagrams of: 1) assistant rating -vs- course rating, 2) assistant rating -vs- final score, and 3) course rating -vs- final score.

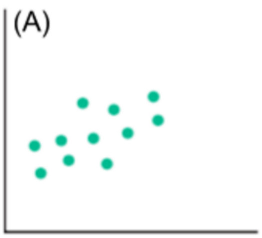
10) *(Refer to the previous question).* The data are section averages. Since the questionnaires were anonymous, it was not possible to link up student rating with scores on an individual basis. Student ability may be a confounding factor. However, controlling for pre-test results turned out to make no difference in the analysis. Each assistant taught one section. True or False, and explain: $_{1.5pts}$

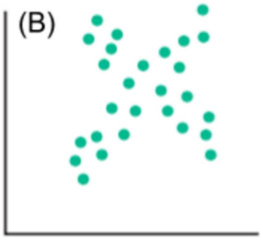
a. On the average, those sections that liked their TA more did better on the final.

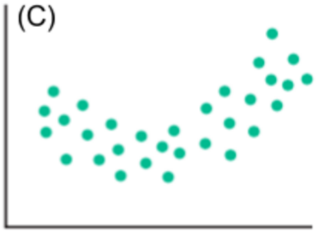
b. There was almost no relationship between the section's average rating of the assistant and the section's average rating of the course.

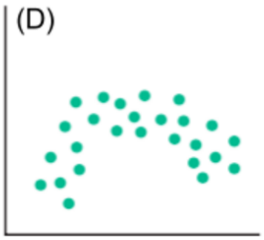
c. There was almost no relationship between the section's average rating of the course and the section's average score on the final.

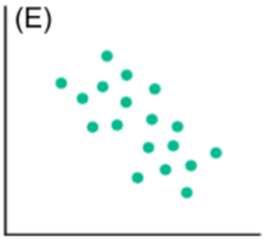
11) Which of the following six scatter diagrams should be summarized by r? Explain. $_{0.6pts}$

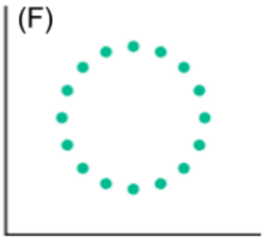












12) For the men age 18 and over in HANES5 we have the following summary statistics $_{1pt}$

- average height ≈ 69 inches, SD ≈ 3 inches
- average weight ≈ 190 pounds, SD ≈ 42 pounds
- $r \approx 0.41$

Please show your work (No work, No credit). Use the regression method to estimate the average weight of the men whose heights were:

a. 69 inches.

b. 66 inches.

c. 24 inches.

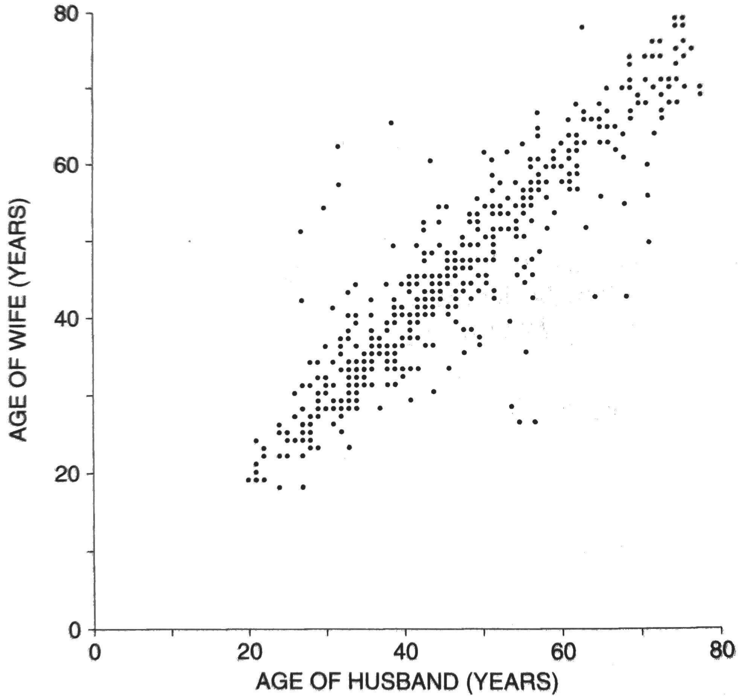
d. 0 inches.

13) For women age 25-34 in the U.S. in 2005, with full-time jobs, the relationship between education (years of schooling completed) and personal income can be summarized as follows: $_{0.25pts}$

- average education ≈ 14 years, SD ≈ 2.4 years
- average income ≈ \$32,000, SD ≈ \$26,000
- $r \approx 0.34$

Please show your work (No work, No credit). Use the regression method to estimate the average income of those women who have finished high school but have not gone on to college (so they have 12 years of education).

14) The scatter diagram below shows ages of husbands and wives in Tennessee. $_{0.2pts}$



a. Why are there no dots in the lower left hand corner of the diagram?

b. Why does the diagram show vertical and horizontal stripes?