




Homework 12 - Stat 202


2.66 Open space and population. The New York City Open Accessible Space Information System Cooperative (OASIS) is an organization of public and private sector representatives that has developed an information system designed to enhance the stewardship of open space.²² Data from the OASIS Web site for 12 large U.S. cities follow. The variables are population in thousands and open total park or open space within city limits in acres. 


City	Population	Open space
Baltimore	651	5,091
Boston	589	4,865
Chicago	2,896	11,645
Long Beach	462	2,887
Los Angeles	3,695	29,801
Miami	362	1,329
Minneapolis	383	5,694
New York	8,008	49,854
Oakland	399	3,712
Philadelphia	1,518	10,685
San Francisco	777	5,916
Washington, D.C.	572	7,504

- Make a scatterplot of the data using population as the explanatory variable and open space as the response variable.
- Is it reasonable to fit a straight line to these data? Explain your answer.
- Find the least squares regression line. Report the equation of the line and draw the line on your scatterplot.
- What proportion of the variation in open space is explained by population?


2.67 Prepare the report card. Refer to the previous exercise. One way to compare cities with respect to the amount of open space that they have is to use the residuals from the regression analysis that you performed in the previous exercise. Cities with positive residuals are doing better than predicted while those with negative residuals are doing worse. Find the residual for each city and make a table with the city name and the residual, ordered from best to worst by the size of the residual. 

2.68 Is New York an outlier? Refer to Exercises 2.66 and 2.67. Write a short paragraph about the data point corresponding to New York City. Is this point an outlier? If it were deleted from the data set, would the least squares regression line change very much? Compare the analysis results with and without this observation. 

2.69 Open space per person. Refer to Exercises 2.66, 2.67 and 2.68. Open space in acres per person is an alternative way to report open space. Divide open space by population to compute the value of this variable for each city. Using this new variable as the response variable and population as the explanatory variable, answer the questions given in Exercise 2.66. How do your new results compare with those that you found in that exercise? 


2.73 Always plot your data! Table 2.4 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.²³ 


- Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict y for $x = 10$.
- Make a scatterplot for each of the data sets and add the regression line to each plot.
- In which of the four cases would you be willing to use the regression line to describe the dependence of y on x ? Explain your answer in each case.

2.74 Data generated by software. The following 20 observations on Y and X were generated by a computer program. 

Y	X	Y	X
34.38	22.06	27.07	17.75
30.38	19.88	31.17	19.96
26.13	18.83	27.74	17.87
31.85	22.09	30.01	20.20
26.77	17.19	29.61	20.65
29.00	20.72	31.78	20.32
28.92	18.10	32.93	21.37
26.30	18.01	30.29	17.31
29.49	18.69	28.57	23.50
31.36	18.05	29.80	22.02

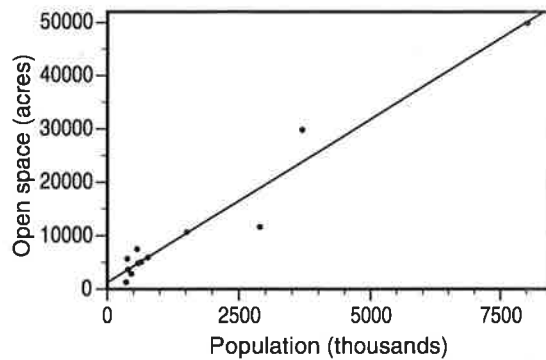
- Make a scatterplot and describe the relationship between Y and X .
- Find the equation of the least-squares regression line and add the line to your plot.
- Plot the residuals versus X .
- What percent of the variability in Y is explained by X ?
- Summarize your analysis of these data in a short paragraph.

2.75 Add an outlier. Refer to the previous exercise. Add an additional observation with $Y = 50$ and $X = 30$ to the data set. Repeat the analysis that you performed in the previous exercise and summarize your results paying particular attention to the effect of this outlier. 

2.76 Add a different outlier. Refer to the previous two exercises. Add an additional observation with $Y = 29$ and $X = 50$ to the original data set. 

- Repeat the analysis that you performed in the first exercise and summarize your results paying particular attention to the effect of this outlier.
- In this exercise and in the previous one, you added an outlier to the original data set and reanalyzed the data. Write a short summary of the changes in correlations that can result from different kinds of outliers.

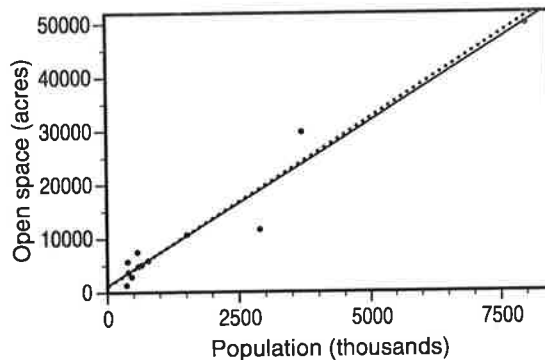
2.66. (a) Scatterplot on the right. (b) The association appears to be roughly linear (although note that the slope of the line is almost completely determined by the largest cities). (c) The regression equation is $\hat{y} = 1248 + 6.1050x$. (d) Regression on population explains $r^2 \doteq 95.2\%$ of the variation in open space.



2.67. Residuals (found with software) are given in the table on the right. Los Angeles is the best; it has nearly 6000 acres more than the regression line predicts. Chicago, which falls almost 7300 acres short of the regression prediction, is the worst of this group.

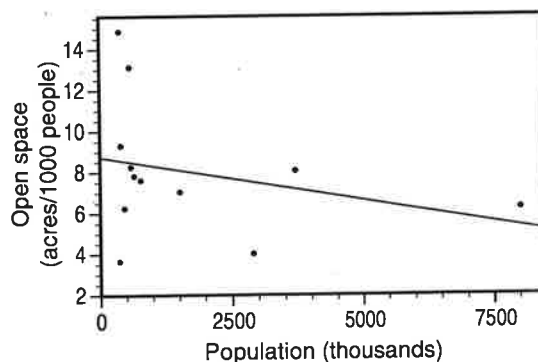
Los Angeles	5994.85
Washington, D.C.	2763.75
Minneapolis	2107.59
Philadelphia	169.42
Oakland	27.91
Boston	20.96
San Francisco	-75.78
Baltimore	-131.55
New York	-282.99
Long Beach	-1181.70
Miami	-2129.21
Chicago	-7283.26

2.68. Because New York's data point is consistent with the pattern of the other cities, we don't consider it an outlier. It does have some impact on the regression line; with New York removed, the equation is $\hat{y} = 1105 + 6.2557x$. However, in the plot on the right, we note that the original regression line (solid) and the new line (dashed) are very similar, and the residuals are likewise very similar.

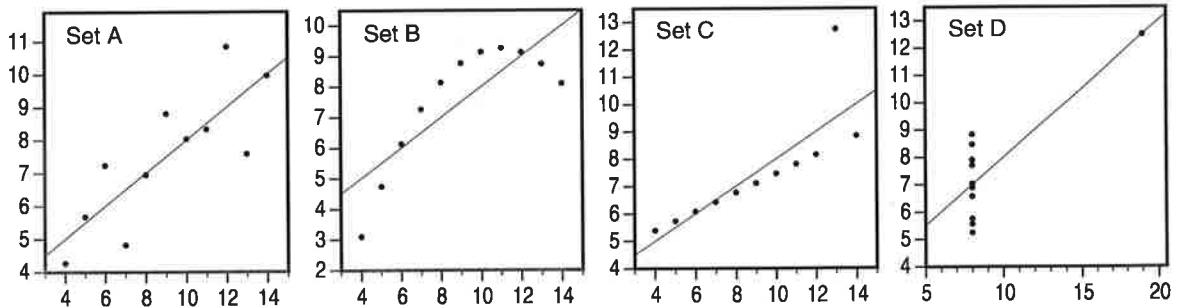


2.69. For Baltimore, for example, this rate is $\frac{5091}{651} \doteq 7.82$. The complete table is shown below on the left. Note that population is in thousands, so these are in units of acres per 1000 people. (a) Scatterplot below on the right. (b) The association is much less linear than in the scatterplot for Exercise 2.66. (c) The regression equation is $\hat{y} = 8.739 - 0.000424x$. (d) Regression on population explains only $r^2 \doteq 8.7\%$ of the variation in open space per person.

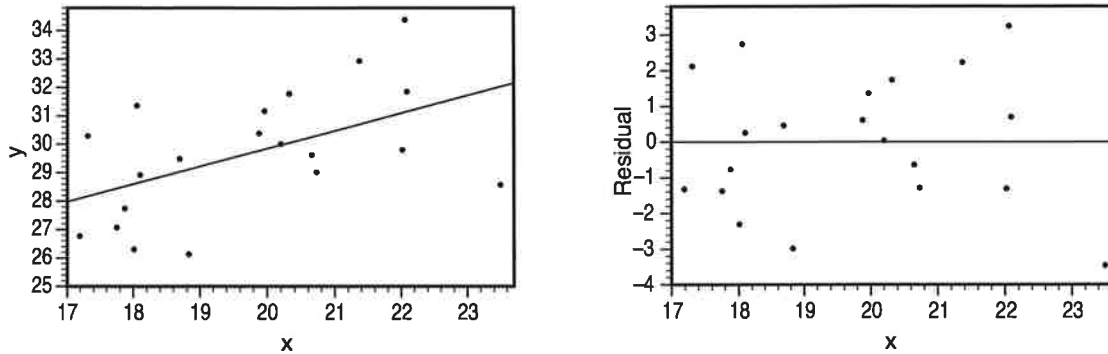
Baltimore	7.82
Boston	8.26
Chicago	4.02
Long Beach	6.25
Los Angeles	8.07
Miami	3.67
Minneapolis	14.87
New York	6.23
Oakland	9.30
Philadelphia	7.04
San Francisco	7.61
Washington, D.C.	13.12



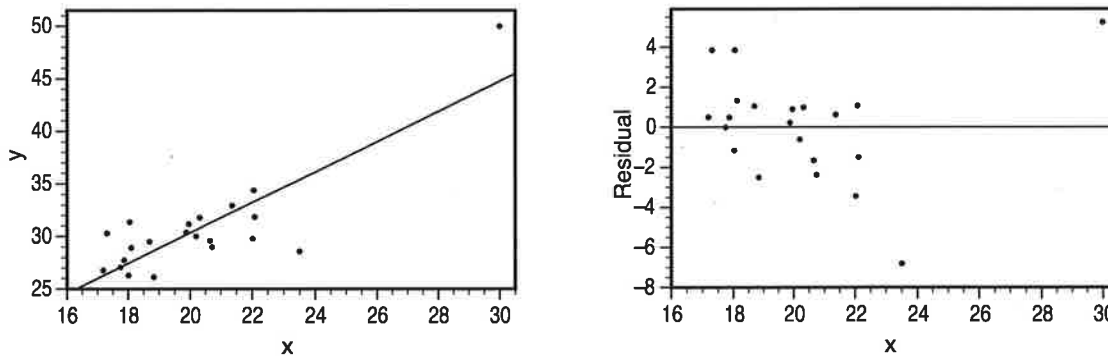
2.73. (a) To three decimal places, the correlations are all approximately 0.816 (for set D, r actually rounds to 0.817), and the regression lines are all approximately $\hat{y} = 3.000 + 0.500x$. For all four sets, we predict $\hat{y} \doteq 8$ when $x = 10$. **(b)** Scatterplots below. **(c)** For Set A, the use of the regression line seems to be reasonable—the data do seem to have a moderate linear association (albeit with a fair amount of scatter). For Set B, there is an obvious *non-linear* relationship; we should fit a parabola or other curve. For Set C, the point (13, 12.74) deviates from the (highly linear) pattern of the other points; if we can exclude it, the (new) regression formula would be very useful for prediction. For Set D, the data point with $x = 19$ is a very influential point—the other points alone give no indication of slope for the line. Seeing how widely scattered the y coordinates of the other points are, we cannot place too much faith in the y coordinate of the influential point; thus, we cannot depend on the estimate when $x = 10$. (We also have no evidence as to whether or not a line is an appropriate model for this relationship.)



- 2.74. (a)** The scatterplot (below, left) suggests a moderate positive linear relationship. **(b)** The regression equation is $\hat{y} = 17.38 + 0.6233x$. **(c)** The residual plot is below, right. **(d)** The regression explains $r^2 \doteq 27.4\%$ of the variation in y . **(e)** Student summaries will vary.



- 2.75. (a)** The scatterplot (below, left) suggests a fairly strong positive linear relationship. **(b)** The regression equation is $\hat{y} = 1.470 + 1.4431x$. **(c)** The residual plot is below, right. The new point's residual is positive; the other residuals decrease as x increases. **(d)** The regression explains $r^2 \doteq 71.1\%$ of the variation in y . **(e)** The new point makes the relationship stronger, but its location has a large impact on the regression equation—both the slope and intercept changed substantially.



- 2.76. (a)** The scatterplot (following page, left) gives little indication of a relationship between x and y . The regression equation is $\hat{y} = 29.163 + 0.02278x$; it explains only $r^2 \doteq 0.5\%$ of the variation in y . The residual plot (following page, right) tells a similar story to the first scatterplot—little evidence of a relationship. This new point does not fall along the same line as the other points, so it drastically weakens the relationship. **(b)** A point that does not follow the same pattern as the others can drastically change an association, and in extreme cases, can essentially make it disappear.