SECTION 2.3 Exercises – Homework #9 – Do at least two(2),

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. OASIS

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

(a) Make a scatterplot of the data using population as the explanatory variable and open space as the response variable.

(b) Is is reasonable to fit a straight line to these data? Explain your answer.

(c) Find the least squares regression line. Report the equation of the line and draw the line on your scatterplot.

(d) 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. OASIS

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. OASIS

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? OASIS

2.70 A different report card. Refer to Exercise 2.67. Prepare a report card based on the analysis of open space per person that you performed in the previous exercise. Write a short paragraph comparing this report card with the one that you prepared in Exercise 2.67. Which do you prefer? Give reasons for your answer. OASIS

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] ANSCOMBE

(a) 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 $x = 10$.

(b) Make a scatterplot for each of the data sets and add the regression line to each plot.

(c) 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.

2.77 The regression equation. The equation of a least-squares regression line is $y = 12 + 6x$.

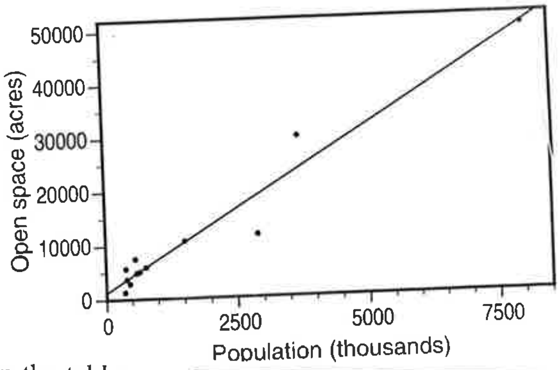
(a) What is the value of y for $x = 5$?

(b) If x increases by one unit, what is the corresponding increase in y?

(c) What is the intercept for this equation?

2.84 Heights of husbands and wives. The mean height of American women in their early twenties is about 64.5 inches and the standard deviation is about 2.5 inches. The mean height of men the same age is about 68.5 inches, with standard deviation about 2.7 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the equation of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

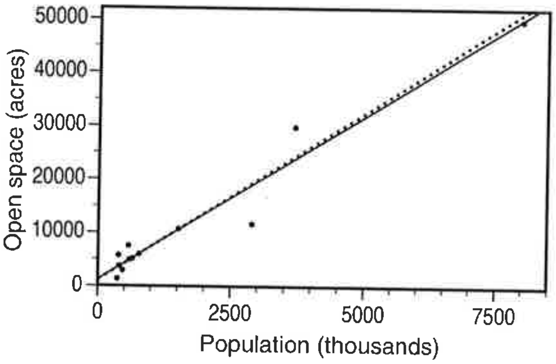
2.66. **(a)** Scatterplot on the right. **(b)** The association appears to be roughly linear (although note that the slope of the line is almost completely determined by the largest cities). **(c)** The regression equation is $\hat{y} = 1248 + 6.1050x$. **(d)** Regression on population explains $r^2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

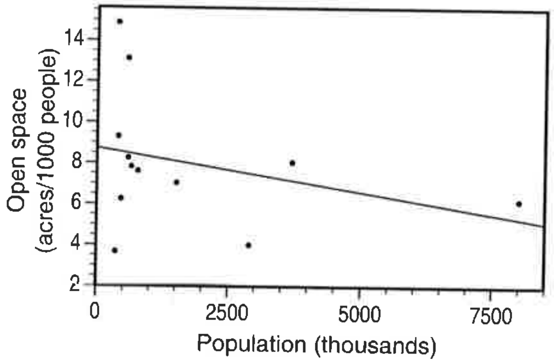
Los Angeles	5994.85
Washington, D.C.	2763.75
Minneapolis	2107.59
Philadelphia	169.42
Oakland	27.91
Boston	20.96
San Francisco	−75.78
Baltimore	−131.55
New York	−282.99
Long Beach	−1181.70
Miami	−2129.21
Chicago	−7283.26

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

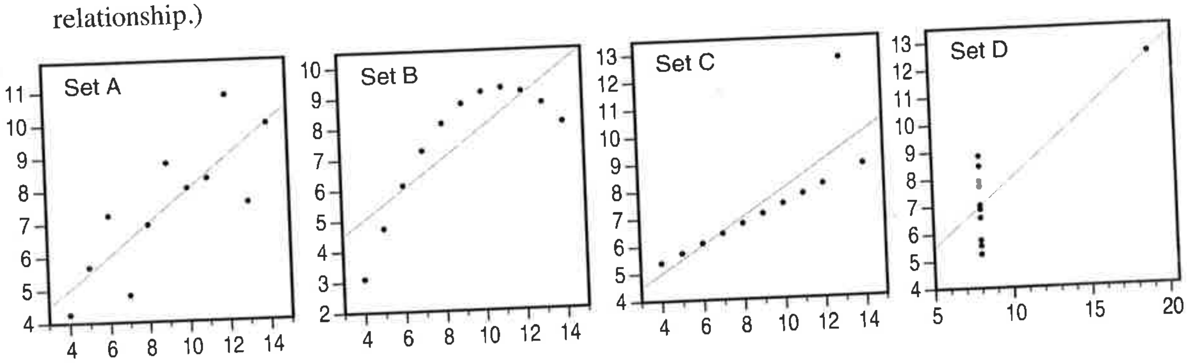


2.70. As in Exercise 2.67, we compute residuals to assess whether a city has more or less open space than we would expect. These are given on the right, in descending order. This time, Minneapolis is best, with about 6.3 acres per 1000 people above what we predict. Miami is worst by this measure, falling short of the prediction by almost 5 acres per 1000 people.

Preferences will vary. One reason to prefer the first approach—apart from the stronger, more linear association—is the negative relationship in the second approach. Why would an individual in a large city need less open space than an individual in a smaller city?

Minneapolis	6.290
Washington, D.C.	4.622
Los Angeles	0.893
New York	0.883
Oakland	0.733
Boston	−0.230
Baltimore	−0.643
San Francisco	−0.796
Philadelphia	−1.056
Long Beach	−2.294
Chicago	−3.490
Miami	−4.914

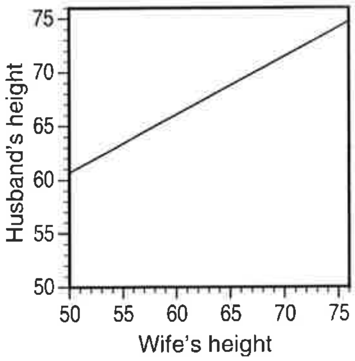
2.73. **(a)** To three decimal places, the correlations are all approximately 0.816 (for set D, r actually rounds to 0.817), and the regression lines are all approximately $\hat{y} = 3.000 + 0.500x$. For all four sets, we predict $\hat{y} \doteq 8$ when $x = 10$. **(b)** Scatterplots below. **(c)** For 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 slope of the line, so 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.)



2.77. **(a)** When $x = 5$, $y = 12 + 6 \times 5 = 42$. **(b)** y increases by 6. (The change in y corresponding to a unit increase in x is the slope of this line.) **(c)** The intercept of this equation is 12.

2.84. Women's heights are the x-values; men's are the y-values. The slope is $b_1 = (0.5)(2.7)/2.5 = 0.54$ and the intercept is $b_0 = 68.5 - (0.54)(64.5) = 33.67$.

The regression equation is $\hat{y} = 33.67 + 0.54x$. Ideally, the scales should be the same on both axes. For a 67-inch tall wife, we predict the husband's height will be about 69.85 inches.


Husband's height
Wife's height