12.E: LINEAR REGRESSION AND CORRELATION (EXERCISES)

These are homework exercises to accompany the Textmap created for "Introductory Statistics" by OpenStax.

12.1: INTRODUCTION

12.2: LINEAR EQUATIONS

Q 12.2.1

For each of the following situations, state the independent variable and the dependent variable.

a. A study is done to determine if elderly drivers are involved in more motor vehicle fatalities than other drivers. The number of fatalities per 100,000 drivers is compared to the age of drivers.
b. A study is done to determine if the weekly grocery bill changes based on the number of family members.
c. Insurance companies base life insurance premiums partially on the age of the applicant.
d. Utility bills vary according to power consumption.
e. A study is done to determine if a higher education reduces the crime rate in a population.

S 12.2.1

a. independent variable: age; dependent variable: fatalities
b. independent variable: # of family members; dependent variable: grocery bill
c. independent variable: age of applicant; dependent variable: insurance premium
d. independent variable: power consumption; dependent variable: utility
e. independent variable: higher education (years); dependent variable: crime rates

Q 12.2.2

Piece-rate systems are widely debated incentive payment plans. In a recent study of loan officer effectiveness, the following piece-rate system was examined:

% of goal reached	< 80	80	100	120
Incentive	n/a	$4,000 with an additional $125 added per percentage point from 81–99%	$6,500 with an additional $125 added per percentage point from 101–119%	$9,500 with an additional $125 added per percentage point starting at 121%

If a loan officer makes 95% of his or her goal, write the linear function that applies based on the incentive plan table. In context, explain the y-intercept and slope.

12.3: SCATTER PLOTS

Q 12.3.1

The Gross Domestic Product Purchasing Power Parity is an indication of a country's currency value compared to another country. Table shows the GDP PPP of Cuba as compared to US dollars. Construct a scatter plot of the data.

Year	Cuba's PPP	Year	Cuba's PPP
1999	1,700	2006	4,000
2000	1,700	2007	11,000
2002	2,300	2008	9,500
2003	2,900	2009	9,700
2004	3,000	2010	9,900
2005	3,500		

S 12.3.1

Check student's solution.

Q 12.3.2

The following table shows the poverty rates and cell phone usage in the United States. Construct a scatter plot of the data

Year	Poverty Rate	Cellular Usage per Capita
2003	12.7	54.67
2005	12.6	74.19
2007	12	84.86
2009	12	90.82

Q 12.3.3

Does the higher cost of tuition translate into higher-paying jobs? The table lists the top ten colleges based on mid-career salary and the associated yearly tuition costs. Construct a scatter plot of the data.

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

S 12.3.3

For graph: check student's solution. Note that tuition is the independent variable and salary is the dependent variable.

Q 12.3.4

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

Q 12.3.5

If there are 15 data points in a set of data, what is the number of degree of freedom?

S 12.3.5

13

12.4: THE REGRESSION EQUATION

Q 12.4.1

What is the process through which we can calculate a line that goes through a scatter plot with a linear pattern?

Q 12.4.2

Explain what it means when a correlation has an r^2 of 0.72.

S 12.4.2

It means that 72% of the variation in the dependent variable (y) can be explained by the variation in the independent variable (x).

Q 12.4.3

Can a coefficient of determination be negative? Why or why not?

12.5: TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

Q 12.5.1

If the level of significance is 0.05 and the p-value is 0.06, what conclusion can you draw?

S 12.5.1

We do not reject the null hypothesis. There is not sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero.

Q 12.5.2

If there are 15 data points in a set of data, what is the number of degree of freedom?

12.6: PREDICTION

Q 12.6.1

Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

a. For each age group, pick the midpoint of the interval for the x value. (For the 75+ group, use 80.)
b. Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
c. Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Predict the number of deaths for ages 40 and 60.
f. Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
g. What is the slope of the least squares (best-fit) line? Interpret the slope.

S 12.6.1

a.

Age	Number of Driver Deaths per 100,000
16–19	38
20–24	36
25–34	24
35–54	20
55–74	18
75+	28

b. Check student's solution.
c. $\hat{y} = 35.5818045 - 0.19182491x$
d. $r = -0.57874$

For four df and alpha $= 0.05$, the LinRegTTest gives p-value $= 0.2288$ so we do not reject the null hypothesis; there is not a significant linear relationship between deaths and age.

Using the table of critical values for the correlation coefficient, with four df, the critical value is 0.811. The correlation coefficient $r = -0.57874$ is not less than -0.811, so we do not reject the null hypothesis.

e. if age = 40, $\hat{y}$ (deaths) $= 35.5818045 - 0.19182491(40) = 27.9$

if age = 60, $\hat{y}$ (deaths) $= 35.5818045 - 0.19182491(60) = 24.1$

f. For entire dataset, there is a linear relationship for the ages up to age 74. The oldest age group shows an increase in deaths from the prior group, which is not consistent with the younger ages.
g. slope $= -0.19182491$

Q 12.6.2

Table shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5
1987	75
1992	75.7
2010	78.7

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated life expectancy for an individual born in 1950 and for one born in 1982.
f. Why aren't the answers to part e the same as the values in Table that correspond to those years?
g. Use the two points in part e to plot the least squares line on your graph from part b.
h. Based on the data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the data?
j. Using the least squares line, find the estimated life expectancy for an individual born in 1850. Does the least squares line give an accurate estimate for that year? Explain why or why not.
k. What is the slope of the least-squares (best-fit) line? Interpret the slope.

Q 12.6.3

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition ten, for various pages is given in Table

Page number	Maximum value ($)
4	16
14	19
25	15
32	17
43	19
57	15
72	16
85	15
90	17

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the ordered pairs.
c. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated maximum values for the restaurants on page ten and on page 70.
f. Does it appear that the restaurants giving the maximum value are placed in the beginning of the "Fine Dining" section? How did you arrive at your answer?
g. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
h. Is the least squares line valid for page 200? Why or why not?
i. What is the slope of the least-squares (best-fit) line? Interpret the slope.

S 12.6.3

a. We wonder if the better discounts appear earlier in the book so we select page as X and discount as Y.
b. Check student's solution.
c. $\hat{y} = 17.21757 - 0.01412x$
d. $r = -0.2752$

For seven df and $\alpha = 0.05$, using LinRegTTest p-value $= 0.4736$ so we do not reject; there is a not a significant linear relationship between page and discount.

Using the table of critical values for the correlation coefficient, with seven df, the critical value is 0.666. The correlation coefficient $xi = -0.2752$ is not less than 0.666 so we do not reject.

e. page 10: 17.08 page 70: 16.23
f. There is not a significant linear correlation so it appears there is no relationship between the page and the amount of the discount.
g. page 200: 14.39
h. No, using the regression equation to predict for page 200 is extrapolation.
i. slope $= -0.01412$

As the page number increases by one page, the discount decreases by $0.01412

Q 12.6.4

Table gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	Time (seconds)
1912	82.2
1924	72.4
1932	66.8
1952	66.8
1960	61.2
1968	60.0
1976	55.65
1984	55.92
1992	54.64
2000	53.8
2008	53.1

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. Is the decrease in times significant?
f. Find the estimated gold medal time for 1932. Find the estimated time for 1984.
g. Why are the answers from part f different from the chart values?
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Use the least-squares line to estimate the gold medal time for the next Summer Olympics. Do you think that your answer is reasonable? Why or why not?

Q 12.6.5

State	# letters in name	Year entered the Union	Rank for entering the Union	Area (square miles)
Alabama	7	1819	22	52,423
Colorado	8	1876	38	104,100
Hawaii	6	1959	50	10,932
Iowa	4	1846	29	56,276
Maryland	8	1788	7	12,407
Missouri	8	1821	24	69,709
New Jersey	9	1787	3	8,722
Ohio	4	1803	17	44,828
South Carolina	13	1788	8	32,008
Utah	4	1896	45	84,904
Wisconsin	9	1848	30	65,499

We are interested in whether or not the number of letters in a state name depends upon the year the state entered the Union.

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Draw a scatter plot of the data.
c. Does it appear from inspection that there is a relationship between the variables? Why or why not?
d. Calculate the least-squares line. Put the equation in the form of: $\hat{y} = a + bx$.
e. Find the correlation coefficient. What does it imply about the significance of the relationship?
f. Find the estimated number of letters (to the nearest integer) a state would have if it entered the Union in 1900. Find the estimated number of letters a state would have if it entered the Union in 1940.
g. Does it appear that a line is the best way to fit the data? Why or why not?
h. Use the least-squares line to estimate the number of letters a new state that enters the Union this year would have. Can the least squares line be used to predict it? Why or why not?

S 12.6.5

a. Year is the independent or x variable; the number of letters is the dependent or y variable.
b. Check student's solution.
c. no
d. \(\hat{y} = 47.03 - 0.0216x\)
e. -0.4280
f. 6; 5
g. No, the relationship does not appear to be linear; the correlation is not significant.
h. current year: 2013: 3.55 or four letters; this is not an appropriate use of the least squares line. It is extrapolation.

12.7: OUTLIERS

Q 12.7.1

The height (sidewalk to roof) of notable tall buildings in America is compared to the number of stories of the building (beginning at street level).

Height (in feet)	Stories
1,050	57
428	28
362	26
529	40
790	60
401	22
380	38
1,454	110
1,127	100
700	46

a. Using "stories" as the independent variable and "height" as the dependent variable, make a scatter plot of the data.
b. Does it appear from inspection that there is a relationship between the variables?
c. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
d. Find the correlation coefficient. Is it significant?
e. Find the estimated heights for 32 stories and for 94 stories.
f. Based on the data in Table, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
g. Are there any outliers in the data? If so, which point(s)?
h. What is the estimated height of a building with six stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
i. Based on the least squares line, adding an extra story is predicted to add about how many feet to a building?
j. What is the slope of the least squares (best-fit) line? Interpret the slope.

Q 12.7.2

Ornithologists, scientists who study birds, tag sparrow hawks in 13 different colonies to study their population. They gather data for the percent of new sparrow hawks in each colony and the percent of those that have returned from migration.

Percent return: 74; 66; 81; 52; 73; 62; 52; 45; 62; 46; 60; 46; 38

Percent new: 5; 6; 8; 11; 12; 15; 16; 17; 18; 18; 19; 20; 20

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. An ecologist wants to predict how many birds will join another colony of sparrow hawks to which 70% of the adults from the previous year have returned. What is the prediction?

S 12.7.2

a. Check student's solution.
b. Check student's solution.
c. The slope of the regression line is -0.3179 with a y-intercept of 32.966. In context, the y-intercept indicates that when there are no returning sparrow hawks, there will be almost 31% new sparrow hawks, which doesn't make sense since if there are no returning birds, then the new percentage would have to be 100% (this is an example of why we do not extrapolate). The slope tells us that for each percentage increase in returning birds, the percentage of new birds in the colony decreases by 0.3179%.
d. If we examine $r2$, we see that only 50.238% of the variation in the percent of new birds is explained by the model and the correlation coefficient, $r = 0.71$ only indicates a somewhat strong correlation between returning and new percentages.
e. The ordered pair $(66, 6)$ generates the largest residual of 6.0. This means that when the observed return percentage is 66%, our observed new percentage, 6%, is almost 6% less than the predicted new value of 11.98%. If we remove this data pair, we see only an adjusted slope of -0.2723 and an adjusted intercept of 30.606. In other words, even though this data generates the largest residual, it is not an outlier, nor is the data pair an influential point.
f. If there are 70% returning birds, we would expect to see $y = -0.2723(70) + 30.606 = 0.115$ or 11.5% new birds in the colony.

Q 12.7.3

The following table shows data on average per capita wine consumption and heart disease rate in a random sample of 10 countries.

Yearly wine consumption in liters	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
Death from heart diseases	221	167	131	191	220	297	71	172	211	300

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.
f. Do the data provide convincing evidence that there is a linear relationship between the amount of alcohol consumed and the heart disease death rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.4

The following table consists of one student athlete's time (in minutes) to swim 2000 yards and the student's heart rate (beats per minute) after swimming on a random sample of 10 days:

Swim Time	Heart Rate
34.12	144
35.72	152
34.72	124
34.05	140
34.13	152
35.73	146
36.17	128
35.57	136
35.37	144
35.57	148

a. Enter the data into your calculator and make a scatter plot.
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot from part a.
c. Explain in words what the slope and y-intercept of the regression line tell us.
d. How well does the regression line fit the data? Explain your response.
e. Which point has the largest residual? Explain what the residual means in context. Is this point an outlier? An influential point? Explain.

S 12.7.4

a. Check student's solution.
b. Check student's solution.
c. We have a slope of -1.4946 with a y-intercept of 193.88. The slope, in context, indicates that for each additional minute added to the swim time, the heart rate will decrease by 1.5 beats per minute. If the student is not swimming at all, the y-intercept indicates that his heart rate will be 193.88 beats per minute. While the slope has meaning (the longer it takes to swim 2,000 meters, the less effort the heart puts out), the y-intercept does not make sense. If the athlete is not swimming (resting), then his heart rate should be very low.
d. Since only 1.5% of the heart rate variation is explained by this regression equation, we must conclude that this association is not explained with a linear relationship.
e. The point $(34.72, 124)$ generates the largest residual of -11.82. This means that our observed heart rate is almost 12 beats less than our predicted rate of 136 beats per minute. When this point is removed, the slope becomes 1.6914 with the y-intercept changing to 83.694. While the linear association is still very weak, we see that the removed data pair can be considered an influential point in the sense that the y-intercept becomes more meaningful.

Q 12.7.5

A researcher is investigating whether non-white minorities commit a disproportionate number of homicides. He uses demographic data from Detroit, MI to compare homicide rates and the number of the population that are white males.

White Males	Homicide rate per 100,000 people
558,724	8.6
538,584	8.9
519,171	8.52
500,457	8.89
482,418	13.07
465,029	14.57
448,267	21.36
432,109	28.03
416,533	31.49
401,518	37.39
387,046	46.26
373,095	47.24
359,647	52.33

a. Use your calculator to construct a scatter plot of the data. What should the independent variable be? Why?
b. Use your calculator's regression function to find the equation of the least-squares regression line. Add this to your scatter plot.
c. Discuss what the following mean in context.
   i. The slope of the regression equation
   ii. The y-intercept of the regression equation
   iii. The correlation r
   iv. The coefficient of determination $r2$.
d. Do the data provide convincing evidence that there is a linear relationship between the number of white males in the population and the homicide rate? Carry out an appropriate test at a significance level of 0.05 to help answer this question.

Q 12.7.6

School	Mid-Career Salary (in thousands)	Yearly Tuition
Princeton	137	28,540
Harvey Mudd	135	40,133
CalTech	127	39,900
US Naval Academy	122	0
West Point	120	0
MIT	118	42,050
Lehigh University	118	43,220
NYU-Poly	117	39,565
Babson College	117	40,400
Stanford	114	54,506

Using the data to determine the linear-regression line equation with the outliers removed. Is there a linear correlation for the data set with outliers removed? Justify your answer.

S 12.7.6

If we remove the two service academies (the tuition is \$0.00), we construct a new regression equation of $y = -0.0009x + 160$ with a correlation coefficient of 0.71397 and a coefficient of determination of 0.50976. This allows us to say there is a fairly strong linear association between tuition costs and salaries if the service academies are removed from the data set.

12.8: REGRESSION (DISTANCE FROM SCHOOL)

12.9: REGRESSION (TEXTBOOK COST)

12.10: REGRESSION (FUEL EFFICIENCY)