Homework

EXERCISE 1

For each situation below, 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 all 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.

EXERCISE 2

In 1990 the number of driver deaths per 100,000 for the different age groups was as follows (Source: The National Highway Traffic Safety Administration's National Center for Statistics and Analysis):

Age	Number of driver deaths per 100,000
15 - 24	28
25 - 39	15
40 - 69	10
70 - 79	15
80+	25

a. For each age group, pick the midpoint of the interval for the x value. (For the 80+ group, use 85.)
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. Pick two ages and find the estimated fatality rates.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between age of a driver and driver fatality rate?

EXERCISE 3

The average number of people in a family that received welfare for various years is given below. (Source: House Ways and Means Committee, Health and Human Services Department)

Year	Welfare family size
1969	4.0
1973	3.6
1975	3.2
1979	3.0
1983	3.0
1988	3.0
1991	2.9

a. Using "year" as the independent variable and "welfare family size" as the dependent variable, make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y} = a + bx$
c. Find the correlation coefficient. Is it significant?
d. Pick two years between 1969 and 1991 and find the estimated welfare family sizes.
e. Use the two points in (d) to plot the least squares line on your graph from (b).
f. Based on the above data, is there a linear relationship between the year and the average number of people in a welfare family?
g. Using the least squares line, estimate the welfare family sizes for 1960 and 1995. Does the least squares line give an accurate estimate for those years? Explain why or why not.
h. Are there any outliers in the above data?
i. What is the estimated average welfare family size for 1986? Does the least squares line give an accurate estimate for that year? Explain why or why not.

EXERCISE 4

Use the AIDS data from the practice for this section, but this time use the columns "year #" and "# new AIDS deaths in U.S." Answer all of the questions from the practice again, using the new columns.

EXERCISE 5

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). (Source: *Microsoft Bookshelf*)

Height (in feet)	Stories

1050	57
428	28
362	26
529	40
790	60
401	22
380	38
1454	110
1127	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. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the number of stories in tall buildings and the height of the buildings?
h. Are there any outliers in the above data? If so, which point(s)?
i. What is the estimated height of a building with 6 stories? Does the least squares line give an accurate estimate of height? Explain why or why not.
j. Based on the least squares line, adding an extra story adds about how many feet to a building?

EXERCISE 6

Below is the life expectancy for an individual born in the United States in certain years. (Source: National Center for Health Statistics)

Year of Birth	Life Expectancy
1930	59.7
1940	62.9
1950	70.2
1965	69.7
1973	71.4
1982	74.5

1987	75.0
1992	75.7

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 values on the above chart that correspond to those years?
g. Use the two points in (e) to plot the least squares line on your graph from (b).
h. Based on the above data, is there a linear relationship between the year of birth and life expectancy?
i. Are there any outliers in the above 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.

EXERCISE 7

The percent of female wage and salary workers who are paid hourly rates is given below for the years 1979 - 1992. (Source: Bureau of Labor Statistics, U.S. Dept. of Labor)

Year	**Percent of workers paid hourly rates**
1979	61.2
1980	60.7
1981	61.3
1982	61.3
1983	61.8
1984	61.7
1985	61.8
1986	62.0
1987	62.7
1990	62.8
1992	62.9

a. Using "year" as the independent variable and "percent" 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? Why or why not?
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 percents for 1991 and 1988.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. Based on the above data, is there a linear relationship between the year and the percent of female wage and salary earners who are paid hourly rates?
h. Are there any outliers in the above data?
i. What is the estimated percent for the year 2050? Does the least squares line give an accurate estimate for that year? Explain why or why not?

EXERCISE 8

The maximum discount value of the Entertainment® card for the "Fine Dining" section, Edition 10, for various pages is given below.

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 10 and on page 70.
f. Use the two points in (e) to plot the least squares line on your graph from (b).
g. 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?

h. Suppose that there were 200 pages of restaurants. What do you estimate to be the maximum value for a restaurant listed on page 200?
i. Is the least squares line valid for page 200? Why or why not?

(9) - (10): The cost of a leading liquid laundry detergent in different sizes is given below.

Size (ounces)	Cost ($)	Cost per ounce
16	3.99	
32	4.99	
64	5.99	
200	10.99	

EXERCISE 9

a. Using "size" as the independent variable and "cost" as the dependent variable, make a scatter plot.
b. Does it appear from inspection that there is a relationship between the variables? Why or why not?
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. If the laundry detergent were sold in a 40 ounce size, find the estimated cost.
f. If the laundry detergent were sold in a 90 ounce size, find the estimated cost.
g. Use the two points in (e) and (f) to plot the least squares line on your graph from (a).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost? Why or why not?

EXERCISE 10

a. Complete the above table for the cost per ounce of the different sizes.
b. Using "Size" as the independent variable and "Cost per ounce" as the dependent variable, make 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 it significant?
f. If the laundry detergent were sold in a 40 ounce size, find the estimated cost per ounce.
g. If the laundry detergent were sold in a 90 ounce size, find the estimated cost per ounce.
h. Use the two points in (f) and (g) to plot the least squares line on your graph from (b).
i. Does it appear that a line is the best way to fit the data? Why or why not?
j. Are there any outliers in the above data?
k. Is the least squares line valid for predicting what a 300 ounce size of the laundry detergent would cost per ounce? Why or why not?

EXERCISE 11

According to flyer by a Prudential Insurance Company representative, the costs of approximate probate fees and taxes for selected net taxable estates are as follows:

Net Taxable Estate ($)	Approximate Probate Fees and Taxes ($)
600,000	30,000
750,000	92,500
1,000,000	203,000
1,500,000	438,000
2,000,000	688,000
2,500,000	1,037,000
3,000,000	1,350,000

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make 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 it significant?
f. Find the estimated total cost for a net taxable estate of $1,000,000. Find the cost for $2,500,000.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Based on the above, what would be the probate fees and taxes for an estate that does not have any assets?

EXERCISE 12

The following are advertised sale prices of color televisions at Anderson's.

Size (inches)	Sale Price ($)
9	147
20	197
27	297
31	447
35	1177
40	2177
60	2497

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make 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 it significant?
f. Find the estimated sale price for a 32 inch television. Find the cost for a 50 inch television.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?

EXERCISE 13

Below are the average heights for American boys. (Source: Physician's Handbook, 1990)

Age (years)	Height (cm)
birth	50.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make 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 it significant?
f. Find the estimated average height for a one year–old. Find the estimated average height for an eleven year–old.
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers in the above data?
j. Use the least squares line to estimate the average height for a sixty–two year–old man. Do you think that your answer is reasonable? Why or why not?

EXERCISE 14

The following chart 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

a. Decide which variable should be the independent variable and which should be the dependent variable.
b. Make 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 (f) different from the chart values?
h. Use the two points in (f) to plot the least squares line on your graph from (b).
i. Does it appear that a line is the best way to fit the data? Why or why not?

j. 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?

Use the following state information for problems 15 – 17.

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

EXERCISE 15

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. Make 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. Use the two points in (f) to plot the least squares line on your graph from (b).
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 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?

EXERCISE 16

We are interested in whether there is a relationship between the ranking of a state and the area of the state.

a. Let rank be the independent variable and area be the dependent variable.
b. What do you think the scatter plot will look like? Make 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 areas for Alabama and for Colorado. Are they close to the actual areas?
g. Use the two points in (f) to plot the least squares line on your graph from (b).
h. Does it appear that a line is the best way to fit the data? Why or why not?
i. Are there any outliers?
j. Use the least squares line to estimate the area of a new state that enters the Union. Can the least squares line be used to predict it? Why or why not?
k. Delete "Hawaii" and substitute "Alaska" for it. Alaska is the fortieth state with an area of 656,424 square miles.
l. Calculate the new least squares line.
m. Find the estimated area for Alabama. Is it closer to the actual area with this new least squares line or with the previous one that included Hawaii? Why do you think that's the case?
n. Do you think that, in general, newer states are larger than the original states?

EXERCISE 17

We are interested in whether there is a relationship between the rank of a state and the year it entered the Union.

a. Let year be the independent variable and rank be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why must the relationship be positive between the variables?
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. Let's say a fifty-first state entered the union. Based upon the least squares line, when should that have occurred?
g. Using the least squares line, how many states do we currently have?
h. Why isn't the least squares line a good estimator for this year?

EXERCISE 18

Below are the percents of the U.S. labor force (excluding self-employed and unemployed) that are members of a union. We are interested in whether the decrease is significant. (Source: Bureau of Labor Statistics, U.S. Dept. of Labor)

Year	Percent
1945	35.5
1950	31.5
1960	31.4
1970	27.3
1980	21.9
1986	17.5
1993	15.8

a. Let year be the independent variable and percent be the dependent variable.
b. What do you think the scatter plot will look like? Make a scatter plot of the data.
c. Why will the relationship between the variables be negative?
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. Based on your answer to (e), do you think that the relationship can be said to be decreasing?
g. If the trend continues, when will there no longer be any union members? Do you think that will happen?

Questions 19 – 20 refer to the following: The data below reflects the 1991-92 Reunion Class Giving. (Source: SUNY Albany alumni magazine)

Class Year	Average Gift	Total Giving
1922	41.67	125
1927	60.75	1,215
1932	83.82	3,772
1937	87.84	5,710
1947	88.27	6,003
1952	76.14	5,254
1957	52.29	4,393
1962	57.80	4,451
1972	42.68	18,093
1976	49.39	22,473
1981	46.87	20,997
1986	37.03	12,590

EXERCISE 19

We will use the columns "class year" and "total giving" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y}$ = a + bx
c. Find the correlation coefficient. What does it imply about the significance of the relationship?
d. For the class of 1930, predict the total class gift: _____
e. For the class of 1964, predict the total class gift: _____
f. For the class of 1850, predict the total class gift: _____ Why doesn't this value make any sense?

EXERCISE 20

We will use the columns "class year" and "average gift" for all questions, unless otherwise stated.

a. What do you think the scatter plot will look like? Make a scatter plot of the data.
b. Calculate the least squares line. Put the equation in the form of: $\hat{y}$ = a + bx
c. Find the correlation coefficient. What does it imply about the significance of the relationship?
d. For the class of 1930, predict the total class gift: _____
e. For the class of 1964, predict the total class gift: _____
f. For the class of 2010, predict the total class gift: _____ Why doesn't this value make any sense?

EXERCISE 21

We are interested in exploring the relationship between the weight of a vehicle and its fuel efficiency (gasoline mileage). The data in the table show the weights, in pounds, and fuel efficiency, measured in miles per gallon, for a sample of 12 vehicles.

Weight	Fuel Efficiency
2715	24
2570	28
2610	29
2750	38
3000	25
3410	22
3640	20
3700	26
3880	21
3900	18
4060	18
4710	15

Table 15

A. Graph a scatterplot of the data.

B. Find the correlation coefficient and determine if it is significant.

C. Find the equation of the best fit line.

D. Write the sentence that interprets the meaning of the slope of the line in the context of the data.

E. What percent of the variation in fuel efficiency is explained by the variation in the weight of the vehicles, using the regression line? (State your answer in a complete sentence in the context of the data).

F. Accurately graph the best fit line on your scatterplot.

G. For the vehicle that weighs 3000 pounds, find the residual (y-yhat). Does the value predicted by the line underestimate or overestimate the observed data value?

H. Identify any outliers, using either the graphical or numerical procedure demonstrated in the textbook.

I. The outlier is a hybrid car that runs on gasoline and electric technology, but all other vehicles in the sample have engines that use gasoline only. Explain why it would be appropriate to remove the outlier from the data in this situation. Remove the outlier from the sample data. Find the new correlation coefficient, coefficient of determination, and best fit line.

J. Compare the correlation coefficients and coefficients of determination before and after removing the outlier, and explain in complete sentences what these numbers indicate about how the model has changed.

EXERCISE 22

The four data sets below were created by statistician Francis Anscomb. They show why it is important to examine the scatterplots for your data, in addition to finding the correlation coefficient, in order to evaluate the appropriateness of fitting a linear model.

Set 1		Set 2		Set 3		Set 4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	9	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Table 16

A. For each data set, find the least squares regression line and the correlation coefficient. What did you discover about the lines and values of r?

For each data set, create a scatter plot and graph the least squares regression line. Use the graphs to answer the following questions:

B. For which data set does it appear that a curve would be a more appropriate model than a line?

C. Which data set has an **influential point** (point close to or on the line that greatly influences the best fit line)?

D. Which data set has an **outlier** (obviously visible on the scatter plot with best fit line graphed)?

E. Which data set appears to be the most appropriate to model using the least squares regression line?

Try these multiple choice questions.

EXERCISE 23

A correlation coefficient of -0.95 means there is a _____ between the two variables.

A. Strong positive correlation
B. Weak negative correlation
C. Strong negative correlation
D. No Correlation

EXERCISE 24

According to the data reported by the New York State Department of Health regarding West Nile Virus for the years 2000-2008 **(http://www.health.state.ny.us/nysdoh/westnile/update/update.htm)**, the least squares line equation for the number of reported dead birds (x) versus the number of human West Nile virus cases (y) is **y-hat = 19.2399 + 0.0257x**. If the number of dead birds reported in a year is 732, how many human cases of West Nile virus can be expected? r = 0.5490.

A. No prediction can be made
B. 19.6
C. 15
D. 38.1

Questions 25 – 27 refer to the following data (showing the number of hurricanes by category to directly strike the mainland U.S. each decade) obtained from www.nhc.noaa.gov/gifs/table6.gif A major hurricane is one with a strength rating of 3, 4 or 5.

Decade	Total Number of Hurricanes	Number of Major Hurricanes
1941-1950	24	10
1951-1960	17	8
1961-1970	14	6
1971-1980	12	4
1981-1990	15	5
1991-2000	14	5
2001 – 2004	**9**	**3**

EXERCISE 25

Using only completed decades (1941 – 2000), calculate the least squares line for the number of major hurricanes expected based upon the total number of hurricanes.

A. y-hat = -1.67x + 0.5
B. y-hat = 0.5x – 1.67
C. y-hat = 0.94x – 1.67
D. y-hat = -2x + 1

EXERCISE 26

The correlation coefficient is 0.942. Is this considered significant? Why or why not?

A. No, because 0.942 is greater than the critical value of 0.707
B. Yes, because 0.942 is greater than the critical value of 0.707
C. No, because 0942 is greater than the critical value of 0.811
D. Yes, because 0.942 is greater than the critical value of 0.811

EXERCISE 27

The data for 2001-2004 show 9 hurricanes have hit the mainland United States. The line of best fit predicts 2.83 major hurricanes to hit mainland U.S. Can the least squares line be used to make this prediction?

A. No, because 9 lies outside the independent variable values

B. Yes, because, in fact, there have been 3 major hurricanes this decade
C. No, because 2.83 lies outside the dependent variable values
D. Yes, because how else could we predict what is going to happen this decade.

Chapter 12