Chapter 7

Regression and Correlation

7.1 Chapter Exercises

7.1.1. Use the following data to study the least squares regression line:

y	x
2	2
4	3
5	4
3	4
4	5
6	6

a. Plot this data, placing y on the vertical axis and x on the horizontal axis.

b. Draw the line that you feel describes the relationship between y and x, without calculating anything.

c. Now calculate the least squares regression line and plot it on the same graph. Compare this line to the line drawn in the previous part.

7.1.2. Use the following data to study the least squares regression line:

c	d
1	2
1	1
2	3
2	2
3	4
3	3

a. Plot this data, placing c on the horizontal axis and d on the vertical axis.

b. Calculate the least squares regression line and the correlation coefficient using $y = d$ and $x = c$. Graph this regression line.

c. Now calculate the least squares regression line and the correlation coefficient using $y = c$ and $x = d$. Put this regression line on the graph as before and compare the lines. Why don't the lines coincide? Compare the correlation coefficient in this part to the one calculated in the previous part.

7.1.3. An electronic device was introduced in 1950 and its sales were tracked over the next ten years.

Sales	Year
1	1951
3	1952
5	1953
8	1954
11	1955
12	1956
14	1957
15	1958
18	1959
19	1960

a. Calculate the least squares regression line to predict sales as a function of the year and the correlation coefficient for this data. Also, calculate $s_e = s_{y|x}$ and s_b.

b. Based on your regression equation, how many sales would you expect in 1950.

c. What would happen to your calculations if the x values were changed so that 1951=1, 1952=2, etc. (i.e., x would then be the time since 1950). Compare the results to your answers for the previous parts.

d. Using what you have learned from the previous parts, what would be the regression equation, correlation, $s_e = s_{y|x}$ and s_b if x values were recoded to indicate time since 1900 (i.e., 1951=51)?

7.1.4. A package distribution service wraps the packages in brown paper, then weighs and mails the packages. They noticed that there appears to be a relation between the amount of paper used to wrap a package and the weight of the package. Since their scale is often broken, they would like to predict the weight of a package as a function of the amount of paper used. The brown paper comes in rolls that are 36 inches wide, so the amount of paper used is measured in terms of the length of the piece used.

Weight (lbs.)	Length (ft.)
2	1
4	2
5	3
8	4
9	5
10	6
13	7
16	8
18	9
20	10

a. Calculate the least squares regression line and the correlation coefficient for this data. Also, calculate $s_\varepsilon = s_{y|x}$ and s_b.

b. Suppose that length of paper was measured in inches, rather than feet. Calculate the least squares regression line and the correlation coefficient for this data. Also, calculate $s_\varepsilon = s_{y|x}$ and s_b. Compare the answers in this part to the answers from the previous part.

c. Suppose that weight in ounces, rather than pounds, but length was in feet. Calculate the least squares regression line and the correlation coefficient for this data. Also, calculate $s_\varepsilon = s_{y|x}$ and s_b. Compare the answers in this part to the answers from the first part.

d. Based on the results of the previous parts, what would be the least squares regression line, correlation coefficient, $s_\varepsilon = s_{y|x}$ and s_b, if weight was in ounces and length was in inches. Now do the calculations to verify your answers.

e. Suppose that the distribution service charges 25 cents per ounce (i.e., \$4 per pound). Find the cost as a function of length (in both inches and feet). Also, find the correlation, $s_\varepsilon = s_{y|x}$ and s_b.

7.1.5. The professor suspects that the number of homework problems completed by students in the class might be related to the scores they receive on tests. He randomly sampled 10 students and found the following:

Quiz score	Number of homework problems
1	1
2	1
2	2
3	2
5	3
4	3
7	4
6	4
8	5
10	5

a. Find the equation that explains quiz score as a function of the number of homework problems completed.

b. Test the hypothesis that for the population of students in this professor's classes each additional homework problem completed leads to an increase of at least 2 points on the quiz versus the alternative that the increase is less than 2 points. Allow for 5% type I error.

c. Find a 95% confidence interval for change in quiz score when students complete one additional homework problem.

d. If Joe completes 4 homework problems, find a 95% confidence interval for his quiz score.

7.1.6. Ten keypunch operators were given a timed test. The results of that test are given below:

Time to finish (min.)	Number of mistakes
9	1
8	3
9	2
6	8
7	4
8	4
6	6
5	7
7	5
5	9

a. Find the equation that describes the number of mistakes made as a function of the time to finish.

b. Suppose the data was drawn randomly from a population of keypunch operators. If you were to test the hypothesis that an operator makes at least 2 fewer mistakes for each additional minute spent versus the alternative that an additional minute does not mean at least 2 fewer mistakes, what would be your conclusion at the 10% level of significance?

c. Find a 98% confidence interval for the average number of mistakes that will be made by people who take 9 minutes to complete the keypunch test.

7.2 Review Problems

7.2.1. A marketing company is trying to predict the amount of discretionary income that is spent on entertainment expenses weekly. The data that is available include monthly mortgage payments and yearly tax bills. To see whether either of these variables would make a good predictor of entertainment expenses, the company takes a sample of ten families and finds the following:

Expenses ($10)	Mortgage ($100)	Tax ($1000)
15	2	10
13	4	9
13	6	7
12	6	8
11	8	6
9	10	5
9	14	4
7	16	3
6	16	1
5	20	2

a. Find the equation that explains entertainment expenses as a function of mortgage payment. Also, find the correlation, $s_e = s_{y|x}$ and s_b.

b. Repeat these calculations for entertainment expenses as a function of tax bill.

c. If you had to choose either mortgage payment or tax bill in order to explain the variation in entertainment expenses, which would make a better predictor? Explain your reasoning.

d. Can you decide which is the better predictor by looking at the slope of the corresponding regression equations? Why or why not?

e. Can you decide which is the better predictor by looking at the s_e of the corresponding regression equations? Why or why not?

f. Estimate, with 95% confidence, the difference in Expenses for families whose yearly Tax bills are $1000 different?

g. Estimate, with 99% confidence, the difference in Expenses for families who monthly Mortgage payments are $300 different?

7.2.2. An auto dealership wishes to predict the average amount of commission made by their salespeople. They collect the following data on the number of sales made and the total commission for seven of their salespeople.

Salesperson	Number of sales	Total commission ($100)
A	1	3
B	2	4
C	3	7
D	4	9
E	5	9
F	6	13
G	7	15

a. Find the equation that will predict commission as a function of number of sales.

b. Estimate with 95% confidence the average commission for salespeople who make one sale.

c. If you could assume that the commission made by salesperson G was identical for each of his 7 sales, and make a similar assumption for each of the other salespeople, could you estimate the average commission per sale with 95% confidence? Compare this answer to the previous one.

d. Actually, each sale resulted in different commissions as seen in the following table:

Salesperson A	B	C	D	E	F	G
3	2	2	2	1	3	4
	2	2	3	4	2	2
		3	1	2	3	1
			3	1	1	3
				1	2	1
					2	2
						2

Estimate with 95% confidence the average commission per sale. Compare your answer to the previous answers.

e. Using the previous answer, estimate the average commission for 4 sales with 95% confidence.

f. Using the regression line, estimate the average commission for salespeople who made 4 sales with 95% confidence. Compare your answer to the previous answer.