Key Formulas

Section 5.3

Mean Square Error

$$s_e^2 = \frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-2} = \frac{\sum (e_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

Standard Error

$$s_e = \sqrt{\frac{\sum\left(y_i - \hat{y}_i\right)^2}{n-2}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Total Sum of Squares

$$\text{Total SS} = \sum\left(y_i - \bar{y}\right)^2$$

$$\text{Total SS} = \text{SSE} + \text{SSR}$$

Sum of Squares of Regression

$$\text{SSR} = \text{Total SS} - \text{SSE}$$

Coefficient of Determination

$$R^2 = \frac{\text{SSR}}{\text{Total SS}} = 1 - \frac{\text{SSE}}{\text{Total SS}}$$

or

$$R^2 = \left(\frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}} \right)^2$$

Correlation Coefficient

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

Additional Exercises

1. Consider the following data:

x	1	2	3	4	5	6	7
y	1	4	9	16	25	36	49

a. Plot the data points on a scatterplot.

b. Determine the correlation coefficient.

c. Describe the relationship between x and y.

d. Determine the least squares line. Use x as the independent variable.

e. Plot the least squares line on the scatterplot.

f. Use the model to compute the error for each data point.

g. Determine the average value of the model's errors.

h. Determine the variance of the errors.

2. Consider the following data:

x	1	2	3	4	5	6	7
y	1	1.41	1.73	2	2.24	2.45	2.65

a. Plot the data points on a scatterplot.

b. Determine the correlation coefficient.

c. Describe the relationship between x and y.

d. Determine the least squares line. Use x as the independent variable.

e. Plot the least squares line on the scatterplot.

f. Use the model to compute the error for each data point.

g. Determine the average value of the model's errors.

h. Determine the variance of the errors.

3. The Road Warrior Trucking Company has kept careful records on ten hauls. The traffic manager has recorded the haul weight of each truck and its miles per gallon during ten runs with the intent of building a regression model. He wants to predict the miles/gallon for a haul based on the haul weight. The haul weights and miles/gallon information are given below. Haul weights are given in thousands of pounds.

Trucking	
Miles/Gallon	**Haul Weight (in thousands)**
4.6	36
4.8	33
5.1	31
4.0	42
4.7	33
5.2	30
4.5	37
4.6	37
4.2	40
4.5	36

a. What level of measurement do the two variables in the table possess?

b. What is the dependent variable in the model? [Hint: which variable does the traffic manager want to predict?]

c. What is the independent variable in the model?

d. Draw a scatterplot of the data. Based on the scatterplot, does a linear model seem appropriate?

e. Write the model in symbolic form. (Assume the parameters of the model have not been estimated.)

f. Use the data provided and estimate the coefficients of the linear model.

g. Interpret the coefficient of the independent variable.

h. Use the model to predict the miles/gallon for a truck hauling 38,000 pounds.

i. Do you believe there is a causal relationship between haul weight and the miles/gallon? If so, which direction is the causality? Do greater haul weights cause reduced mileage, or vice versa? Does the regression analysis prove the causality?

j. Compute the correlation coefficient of the data.

k. What percentage of the variation in miles per gallon is explained by the haul weight?

4. An agricultural research station is trying to determine the relationship between the yield of sunflowers and the amount of fertilizer applied. To determine the relationship, three different fields were planted. In each field four different plots were defined. In each plot a different amount of fertilizer was used. The plot assignments for the fertilizer application were randomly selected in each field.

Agricultural Research

Pounds of Fertilizer (per acre)	Pounds of Sunflower Seeds (per acre)	Pounds of Fertilizer (per acre)	Pounds of Sunflower Seeds (per acre)
200	420	600	580
200	445	600	600
200	405	600	610
400	580	800	630
400	540	800	620
400	550	800	626

a. What level of measurement do the two variables in the table possess? Is the data developed through controlled experiment or is the data observational?

b. Draw a scatterplot of the data.

c. If a linear model is developed, which of the variables will be the dependent variable? Why?

d. Use least squares techniques to estimate the appropriate model.

e. Interpret the meaning of the slope coefficient in the model.

f. What percentage of the variation in pounds of sunflower seeds per acre can be explained by the amount of fertilizer used?

g. Predict the sunflower seed yield per acre if 500 pounds of fertilizer is applied.

5. Ten games were recently played in the National Football League. The number of passing yards and rushing yards, along with the winner (w) and loser (l) of the game, for the 20 participating teams is given below.

National Football League

Game	Team	Rush	Pass	Outcome	Game	Team	Rush	Pass	Outcome	Game	Team	Rush	Pass	Outcome
1	1	60	320	l	4	8	88	190	w	8	15	125	310	w
1	2	158	124	w	5	9	75	183	l	8	16	99	184	l
2	3	53	183	l	5	10	110	182	w	9	17	187	189	w
2	4	127	164	w	6	11	43	328	w	9	18	21	366	l
3	5	60	115	l	6	12	95	192	l	10	19	171	87	w
3	6	343	50	w	7	13	56	237	l	10	20	60	337	l
4	7	148	242	l	7	14	200	132	w					

Before computing any statistics, if you compared winners and losers with respect to rushing and passing yardage, what would you expect to find? Compare the rushing and passing yards of winners and losers.

a. Make a scatterplot of rushing yards versus passing yards.

b. Compute the correlation coefficient for these two variables. Does the sign of the correlation coefficient make sense? Explain your answer.

c. Compute the coefficient of determination and interpret its meaning.

d. How do the correlation coefficient and the coefficient of determination relate?

6. Experimental results comparing two methods for estimating maximum aerobic speed were summarized in 'Comparison of Two Field Tests to Estimate Maximum Aerobic Speed', and reported in *The Journal of Sports Sciences*.[15] The techniques compared were the treadmill test and the track test. Data from the 17 participants is given in the table below.

Maximum Aerobic Speed					
Subject	**Track**	**Treadmill**	**Subject**	**Track**	**Treadmill**
1	15	14	**10**	16	16
2	16	16	**11**	16	16
3	16	16	**12**	16	16
4	13	12	**13**	13	12
5	13	12	**14**	17	18
6	13	14	**15**	19	20
7	16	16	**16**	18	20
8	16	16	**17**	18	20
9	18	16			

a. Make a scatterplot of the results.

b. Does there appear to be a negative or positive relationship between the variables?

c. Compute the correlation coefficient for these two variables.

d. Create two new variables, V and W, where $V = Track/2$ and $W = Treadmill/5$.

e. Compute the correlation coefficient between V and W.

f. What appears to be the effect on the correlation coefficient of multiplying each variable by a constant value?

7. The following data is obtained on two quantitative variables.

x	876	516	598	789	734	667	682	714	598
y	50.1	88.2	80.7	39.6	20.5	40.9	30.6	22.9	34.8

a. Make a scatterplot of the data.

b. Does there appear to be a negative or positive relationship between the variables?

c. Compute the correlation coefficient.

d. Create two new variables, V and W, where $V = X - 700$ and $W = Y - 50$.

e. Compute the correlation coefficient between V and W.

f. What appears to be the effect on the correlation coefficient of subtracting a constant value from each of the variables?

8. The FBI releases crime statistics for cities categorized by population size. The table below gives data for cities containing different population groups.[16] The robbery and murder rates are given for these cities together with their crime index and violent crime index. The trends for the crime index indicate the percent change in offenses known to law enforcement.

Crime Statistics

City Population	Robbery Rate	Murder Rate	Crime Index	Violent Crime Index
Over 1,000,000	1.1	4.0	0.6	1.5
500,000 to 999,999	5.7	0.2	0.9	4.1
250,000 to 499,999	0.7	3.6	1.9	1.8
100,000 to 249,999	2.1	2.6	0.7	1.8
50,000 to 99,999	0.2	6.7	1.2	0.9
25,000 to 49,999	1.4	0.1	1.2	1.2
10,000 to 24,999	2.6	9.8	0.9	1.7
Under 10,000	1.4	14.7	1.2	1.8

a. In general, if we have k quantitative variables measured in a study, how many different correlation coefficients can be computed?

b. How many different correlation coefficients can be computed from the above data?

c. Which pair of variables appears to be most correlated?

d. Which pair of variables appears to be least correlated?

e. Do any pairs of variables appear to be negatively correlated?