Chapter 12 : Linear Correlation and Linear Regression

LINEAR REGRESSION AND CORRELATION SKILLS PRACTICE PROBLEMS

Skills Practice Problems for Linear Regression and Correlation, by Roberta Bloom, De Anza College
to accompany Linear Regression and Correlation Notes, by Roberta Bloom, De Anza College

- Some material is derived and remixed from **Introductory Statistics** from Open Stax (Illowsky/Dean) available for download free at http://cnx.org/content 11562/latest/ or https://openstax.org/details/introductory-statistics
- Some material is derived and remixed from **Inferential Statistics and Probability: A Holistic Approach**, by Maurice Geraghty, De Anza College, 1/1/2018, http://professormo.com/holistic/HolisticStatisticsRev190204.pdf

CHECKLIST: 10 SKILLS AND CONCEPTS YOU NEED TO LEARN IN CHAPTER 12

1. Identify which variable is independent and which variable is dependent, from the context (words) of the problem.

2. Know calculator skills for items 3, 4, 5, 6, 9 below.
   Complete calculator instructions are near the end of these notes and will be demonstrated in class.

3. Create and use a scatterplot to visually determine if it seems reasonable to use a straight line to model a relationship between the two variables.

4. Find, interpret, and use the correlation coefficient to determine if a significant linear relationship exists and to assess the strength of the linear relationship (hypothesis test of significance of r using the p-value approach).

5. Find and interpret the coefficient of determination to determine
   a) what percent of the variation in the dependent variable is explained by the variation in the independent variable using the best fit line,
   b) what percent of the variation in the dependent variable is not explained by the line
   What does the scattering of the points about the line represent?

6. Find and use the least squares regression line to model and explore the relationship between the variables, finding predicted values within the domain of the original data, finding residuals, analyzing relationship between the observed and predicted values.

7. Know when it is and is not appropriate to use the least squares regression line for prediction.
   In order to use the line to predict, ALL of the following conditions must be satisfied:
   a) scatterplot of data must be well modeled with a line – visually check the graph to observe if a line is a reasonable fit to the data
   b) p-value < α
   c) the value of x for which we want to predict an dependent value must be in the domain of the data used to construct the best fit line.

8. Write a verbal interpretation of the slope as marginal change in context of the problem. (Marginal change is change in y per unit of change in x, stated in the words of and using the numbers and units of the particular problem. See examples done in class and see textbook for how to write this interpretation.)

9. Understand the importance of outliers and influential points

10. Understand the concept of the least squares criteria for determining the best fit line.

SKILLS PRACTICE 1

The data show the number of students enrolled and number of faculty at community colleges in Santa Clara and Santa Cruz Counties.

This data is from the state's community college website data bank for fall 2008.

	X = Number of Students	Y = Number of Faculty
De Anza	26173	846
Foothill	20919	618
West Valley	13800	433
Mission	12814	411
San Jose City	11513	436
Evergreen	10936	330
Gavilan	9092	234
Cabrillo	16369	618

1. Find the best fit line and write the equation of the line.

2. Graph a scatterplot of the data, showing the best fit line.

3. Find the correlation coefficient and the coefficient of determination.

4. Considering this data as a sample of all bay area community colleges, test the significance of the correlation coefficient. Show your work and clearly state your conclusion.

5. Write the interpretation of the coefficient of determination in the context of the data.

6. Write the interpretation of the slope of the regression line, in the context of the data.

7. How many faculty would be predicted at a college with 15000 students?

8. a) How many faculty are predicted for a college with 11,513 students?
   b) What is the residual ($y - \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 11,513?
   c) Did value predicted by the line overestimate or underestimate the observed value?

9. a) How many faculty are predicted for a college with 20,919 students?
   b) What is the residual ($y - \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 20,919?
   c) Did value predicted by the line overestimate or underestimate the observed value?

10. Would it be appropriate to use the line to predict the number of faculty at a community college with:

    a) 4000 students? __________ b) 14000 students? __________ c) 40000 students? ________

    Explain why or why not.

SKILLS PRACTICE 2

Do sales of a DVD increase when its price is lower?

GreatBuy Electronics Store is selling a particular DVD at all their stores in the state. The price varies during different weeks, some weeks at full price, other weeks at discounted prices. The manager recorded the price and sales for this particular DVD for a sample of 12 weeks.

(Sales have been rounded to the nearest 10; the price is the same at all store branches during the same week.)

x = price of this DVD during a one week period
y = number of this DVD sold during a one week period

x ($)	y (DVDs)
13	370
15	400
15	330
16	380
18	250
13	340
16	350
15	310
16	360
18	260
13	380
18	290

1. Find the best fit line and write the equation of the line.

2. Graph a scatterplot of the data, showing the best fit line

3. Find the correlation coefficient and the coefficient of determination

4. Test the significance of the correlation coefficient. Show your work and clearly state your conclusion

5. Write the interpretation of the coefficient of determination in the context of the data.

6. Write the interpretation of the slope of the regression line, in the context of the data.

7. a) Predict the sales when the price is $15.
 b) What is the residual ($y - \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 15?
 c) Did value predicted by the line overestimate or underestimate the observed value?

8. The sales manager asks you to predict sales if he offers a special sale price of $10 for one week. What should you answer?

SKILLS PRACTICE 3

Is there a relationship between the number of absences a student has during the quarter (out of 54 class sessions) and the grade the student earns for the course?

Days Absent	0	0	1	1	2	2	3	4	4	5	8	9	9	10	12	12	15	16	20	22
Course Grade	3	4	2	1	4	2	2	2	4	3	3	2	1	3	0	2	1	2	1	0

1. Identify the independent variable: ___ = ____

2. Identify the dependent variable: ____ = ____

3. Find the best fit line and write the equation of the line. ____

4. Graph a scatterplot of the data, showing the best fit line

5. Find the correlation coefficient and the coefficient of determination:

6. Test the significance of the correlation coefficient. Show your work and clearly state your conclusion

7. Write the interpretation of the coefficient of determination in the context of the data.

8. Write the interpretation of the slope of the regression line, in the context of the data.

9. a) Predict the grade for a student with 5 absences. ________________________________

b) Find the residual $y - \hat{y}$ (difference between observed y and predicted $\hat{y}$) when x = 5: _______________

c) Did value predicted by the line overestimate or underestimate the observed value?

10. a) Predict the grade for a student with 15 absences. ________________________________

b) Find the residual $y - \hat{y}$ (difference between observed y and predicted $\hat{y}$) when x = 15: _______________

c) Did value predicted by the line overestimate or underestimate the observed value?

11. a) Predict the grade for a student with 9 absences. ________________________________

b) How do the data compare to the predicted value?

12. Predict the grade for a student with 40 absences. Explain why the best fit line predicts a grade that does not make any sense in this problem.

SKILLS PRACTICE 4

The population of River City is recorded by the U.S. Census every 10 years.
Between the census in 1950 and 2010, the population has more than doubled.
The population data for the past 7 censuses are:

Year	1950	1960	1970	1980	1990	2000	2010
Population	22, 250	23,100	26,250	30, 200	35,250	41,300	52,100

1. Find the correlation coefficient and conduct a hypothesis test for significance.

2. Graph a scatterplot of the data and graph the best fit line to see how the data fit the line.

3. Does the best fit line appear to be good model for this data? Explain.

SKILLS PRACTICE 5

We are interested in the relationship between the weights of packages and the shipping costs for packages shipped by the Speedy Delivery Co.

x = weight of package (pounds)	5	5	16	9	6	15	7	3	12	6	5	3	12	6	11
y = shipping cost ($)	3	3	10	12	4	7	4	2	6	3	3	3	6	4	6

1. Find the best fit line and write the equation of the line. ____

2. Graph a scatterplot of the data, showing the best fit line

3. Find the correlation coefficient and conduct a hypothesis test for significance.

4. Find the coefficient of determination and write its interpretation in the context of the data.

5. Write the interpretation of the slope of the regression line, in the context of the data.

6. Predict the shipping cost for a package that weighs 10 pounds.

7. Identify any data points that are outliers.