Because R^2 is a unit-free measure, it can be used to compare the fit of two models. The GPA model only explains approximately 16% of the variation in the dependent variable. Compared to the production model, which had an R^2 of approximately 0.7507, this model seems dramatically inferior. Using R^2 as a criterion, the production model seems to have a substantially better fit (0.7507 versus 0.1597) than the GPA model. The real question is whether you can predict more accurately with the model than other available alternatives. If so, models with relatively low coefficients of determination (such as the GPA model) are useful. For example, if you could develop a model to predict stock prices, minute-by-minute, achieving an R^2 value of only 0.20, you could be a very wealthy person.

R^2 can also be found using the following computational formula.

Technology

For instructions on how to find the coefficient of determination using technology, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Regression > Coefficient of Determination.**

Formula

Coefficient of Determination

The **coefficient of determination**, R^2, can be calculated using the equation

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

Technology

For instructions on how to find the correlation coefficient using technology, visit stat.hawkeslearning.com and navigate to **Discovering Business Statistics, Second Edition > Technology Instructions > Regression > Correlation Coefficient**.

Normally you will not have to use this formula since calculators and computer programs can calculate the coefficient of determination. Recall the computational formula for the correlation coefficient, discussed in Section 4.7, that measures the degree of linear relationship between two variables.

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

The coefficient of determination is the square of the correlation coefficient. The correlation coefficient can be found by either using the formula given previously or by taking the square root of the coefficient of determination and adding the sign corresponding to the slope coefficient. Remember that the correlation coefficient takes on values between −1 and 1, where negative values indicate a downward sloping relationship and positive values indicate an upward sloping relationship. The coefficient of determination takes on values between 0 and 1, where values close to 0 indicate a weak linear relationship and values close to 1 indicate a strong linear relationship.

13.3 Exercises

Basic Concepts

1. What is the total sum of squares?
2. How are the total sum of squares and the sample variance related?
3. Define error in terms of a regression model.
4. What part of the simple linear regression model captures the unexplained variation?
5. Describe the total sum of squares in terms of explained and unexplained variation.
6. What is the sum of squares of regression?
7. Express SSR in terms of the total sum of squares and the sum of squared errors. Interpret this in terms of model variation.
8. Why will there be errors in virtually all regression models?

9. What is the coefficient of determination? What kinds of values can the coefficient of determination take?

10. Suppose that regression analysis is performed and the resulting model has an R^2 value of 0.856. Interpret this value.

11. How is the coefficient of determination related to the correlation coefficient?

Exercises

12. A direct mail marketing company has been experimenting with the effect of price on sales. Five different direct mail prices have been sent to different sets of customers. They have carefully tracked the customers from each group and have recorded the proportion from each price category that purchased the product. The results are given in the following table.

Direct Mail

Proportion That Purchased Product	Price of Product ($)
0.032	29.95
0.028	34.95
0.026	39.95
0.015	44.95
0.009	49.95

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

b. Specify the model that the marketing manager would be interested in estimating.

c. Which of the variables is the dependent variable in the model?

d. Which of the variables is the independent variable in the model?

e. Draw a scatterplot of the data.

f. Use the data in the table to estimate the model.

g. Predict the proportion that will buy the product if the price is $35.00.

h. Compute the mean error for the model you estimated in part **f.**

i. Determine the mean square error.

j. What is the coefficient of determination? Interpret this value in terms of the problem.

k. Consider exercise 12 parts f and j. Use the information in these two parts to compute the correlation coefficient between the Proportion that Purchased Product and the Price of Product.

13. An economist is studying the relationship between income and savings. He has randomly selected seven subjects and obtained income and savings data from them. He wishes to use a simple linear regression model to predict savings based on annual income.

Income and Savings

Income (Thousands of Dollars)	Savings (Thousands of Dollars)
28	0.2
25	0
34	0.8
43	1.2
48	3.1
39	2.1
74	8.3

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

b. Which of the variables is the dependent variable in the model?

c. Which of the variables is the independent variable in the model?

d. Draw a scatterplot of the data. Does the scatterplot suggest that a linear model is appropriate? Explain.

e. Use the data to estimate the appropriate model.

f. Predict the savings for someone who earns fifty thousand dollars annually.

g. Interpret the meaning of the slope coefficient in the problem.

h. What fraction of the variation in savings is explained by income?

14. 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 per gallon for a haul based on the haul weight. The haul weights and miles per gallon information is given in the following table. Haul weights are given in thousands of pounds.

Trucking

Miles per Gallon	Haul Weight (Thousands of Pounds)
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 is the dependent variable in the model?

b. What is the independent variable in the model?

c. Construct a scatterplot of the data. Based on the scatterplot, does a linear model seem appropriate?

d. Write the model in terms of miles per gallon and haul weight. (Assume the parameters of the model have not been estimated.)

e. Use the data provided and estimate the coefficients of the linear model.f. Interpret the coefficient of the independent variable.

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

h. Do you believe there is a causal relationship between haul weight and the miles per 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?

15. An agricultural research station is trying to determine the relationship between the yield of sunflower seeds 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)
200	420
200	445
200	405
400	580
400	540
400	550
600	580
600	600
600	610
800	630
800	620
800	626

a. Are the data developed through a controlled experiment or are 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 the method of least squares to estimate the appropriate model.

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

f. What fraction 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 are applied.

16. Since 2009, the average term for a new-car loan was nearly 64 months. This leaves the buyer vulnerable to owing more on the car than it is worth. When applying for an automobile loan, it is oftentimes recommended to sign up for the shortest term you can afford. It is believed that along with one's credit rating, the length of the loan will help the buyer get a favorable interest rate. The following table contains interest rates and lengths of loans for 20 randomly selected auto purchases. Using the data in the table, answer the following questions.

Lengths of Loans and Interest Rates

Months Financed	Interest Rate (%)
12	4.00
24	4.40
36	5.24
12	3.43
24	4.40
36	5.79
36	5.98
48	6.58
36	5.31
36	5.91

Lengths of Loans and Interest Rates (cont.)

Months Financed	Interest Rate (%)
48	6.51
48	6.68
60	7.13
60	7.48
72	8.31
60	7.85
72	8.07
72	8.48
48	6.12
72	8.07

a. Using statistical software, estimate the coefficients of the least squares regression equation.

b. Interpret the meaning of the slope and the intercept in part **a.**

c. Predict the interest rate for a person interested in a four-year auto loan.

d. Should you use the model to predict interest rates for an eight-year loan? Justify your answer.

e. Determine the coefficient of determination and explain its meaning in terms of the problem.

f. Calculate the correlation coefficient for this model. What does it mean?

g. What interest rate would one expect to get if they were planning to apply for a five-year auto loan?

17. A sample data shows that the correlation coefficient between the number of pizzas and the delivery time is 0.64. If you would fit a regression model for the data to predict the delivery time given the number of pizzas, what percentage of the variation in the delivery time would be explained by the regression model?

13.4 Fitting a Linear Time Trend

Definition

Linear Time Trend Model

A **linear time trend model** is a linear model used to model the changes in some phenomenon over time; the independent variable is always a time index.

In Chapter 4 we discussed the notion that the mean is not a reasonable descriptor for nonstationary time series data. Recall that nonstationary time series do not meander around some central value. Instead, the data tend to get larger or smaller over time. How can you describe time series data that possess a trend? For some time series, a **linear time trend** is a useful model. A linear time trend is nothing but a line that is used to model the changes in some phenomenon measured over time. In a linear time trend model, the independent variable is always a time index. The following example will illustrate the estimation of a **linear time trend model**.

Example 13.4.1

Modeling Data with a Linear Time Trend

Data

This data set can be found on stat.hawkeslearning.com under **Discovering Business Statistics, Second Edition > Data Sets > Tuition Consumer Price Index**.

Many analysts believe that college tuition prices may soon be in the same situation as housing prices were when the housing bubble burst (causing home prices to drop significantly). Table 13.4.1 contains data for the Tuition Consumer Price Index (TCPI) from 1978 to 2020. Use a linear time trend to model the data.

Table 13.4.1 – Tuition Consumer Price Index, 1978-2020

Year	TCPI
1978	59.9
1979	64.7