CHAPTER 17

Bivariate Correlation and Regression

1. BIVARIATE ANALYSIS OF ASSOCIATION

I. **Degrees of Association between Variables**

A. **Statistical Techniques**

1. **Bivariate Techniques**–Statistical methods of analyzing the relationship between two variables

2. **Multivariate Techniques**–when more than two variables are involved, discussed in Chapter 18

3. **Independent Variable (Predictor)**–the symbol or concept that the researcher has some control over or can manipulate to some extent and that is hypothesized to cause or influence the dependent variable

4. **Dependent Variable**–a symbol or concept expected to be explained or caused by the independent variable

B. **Procedures for Metric and Ordinal Data**

1. Metric Data

a. Bivariate Regression

b. Pearson's Product Moment Correlation

2. Other Statistical Procedures (Chapter 16)

a. Two-group *t* Test

b. Chi-square Analysis of Crosstabs or Contingency Tables

c. ANOVA (Analysis of Variance) for Two Groups

2. BIVARIATE REGRESSION

I. **Bivariate Regression Analysis**

A. **Bivariate Regression Analysis Defined**–a statistical procedure which analyzes the strength of the linear relationship between two variables when one is considered the independent variable and the other the dependent variable

B. **Nature of the Relationship**

1. **Scatter Diagram**–one way to study the nature of the relationship between the dependent and the independent variable is to plot the data in a scatter diagram

a. **Dependent Variable Y**–plotted on the vertical axis

b. **Independent Variable X**–plotted on the horizontal axis

c. **Linear Relationship**–apply linear regression to the data

d. **Nonlinear Relationship**–apply curve-fitting nonlinear regression techniques

See Exhibit 17.1 Types of Relationships Found in Scatter Diagrams (p 515)

PRACTICING MARKETING RESEARCH

Questions

1. **What did regression analysis show about the impact of "Teen Mom" on teen pregnancy?**

   That there was a 5.7% reduction in pregnancies attributable to the show.

2. **What, if anything, did regression tell us about the reasons for the decline in teen pregnancies?**

   Regression only showed that viewership was associated with a lower likelihood of getting pregnant. Other qualitative research efforts would show the specific reasons.

C. **Example of Bivariate Regression**

1. 20 stores were identified.

2. Goal is to develop a model that can be used to evaluate potential sites for store locations.

3. Daily traffic counts for each site were taken over a 30 day period.

4. A scatter plot of the resulting data was drawn.

1. Least Squares Estimation Procedure

a. The least squares procedure is a simple mathematical technique that can be used to fit a line to data for X and Y that best represents the relationship between the two variables.

b. No straight line will perfectly represent every observation in the scatterplot. This is reflected in discrepancies between the actual values (dots on the scatter diagram) and predicted values (values indicated by the line). Any straight line fitted to the data in a scatterplot is subject to error

c. The least squares procedure results in a straight line that fits the actual observations (dots) better than any other line that could be fitted to the observations.

d. The general equation is **Y = a + bX + e** and the estimating equation is $Y = \hat{a} + \hat{b}X + e$, where:

Y = dependent variable

$\hat{a}$ = estimated ***Y*** intercept of regression line

$\hat{b}$ = estimated slope of regression line, regression coefficient

X = independent variable

e = error, difference between actual value and value predicted by regression line

2. **Regression Line**–Predicted values for ***Y***, based on calculated values for $\hat{a}$ and $\hat{b}$ (Exhibit 17.5). In addition, errors for each observation $(Y - \hat{Y})$ are shown. The regression line resulting from the $\hat{Y}$ values is plotted in Exhibit 17.6.

3. **Strength of Association –** the estimated regression function describes the *nature* of the relationship between the variables ***X*** and ***Y***.

a. **Coefficient of determination**– denoted R^2, measures *strength* of the linear relationship between ***X*** and ***Y***.

1) Indicates the percentage of the total variation in ***Y*** that is "explained" by the variation in ***X***.

b. The R^2 statistic ranges from 0 to 1, where 1 = a perfect linear relationship and 0 = no relationship at all.

4. **Statistical Significance of Regression Results**

a. **Total variation (Total Sum of Squares or SST)** = explained variation + unexplained variation

1) **Sum of squares due to regression (SSR)**–variation explained by the regression

2) **Error sum of squares (ESS)**–variation not explained by the regression

5. **Hypotheses Concerning Overall Regression**

The interest is in the hypotheses regarding the computed R^2 value for the problem. Is the amount of variance explained in the result significantly greater than should be expected due to chance? Analysis of variance (an *F* test) is used to test the significance of the results

a. **The null hypothesis H_0**–there is no linear relationship between ***X*** (average daily vehicular traffic) and ***Y*** (annual sales).

b. **Alternative hypothesis H_a**–there is a linear relationship between ***X*** and ***Y***.

c. Must decide on a standard level of significance: $\alpha = .05$ (i.e., 5 percent chance of incorrectly rejecting the null hypothesis)

d. **F =MSR/MSE**

e. Compare the calculated *F*-value to the table value of *F*.

f. Because the calculated *F*-value is > than the table value of *F*, reject the null hypothesis.

6. **Hypotheses about the Regression Coefficient *b***

a. ***b*** is the estimate of the effect of a one-unit change in X on Y. The hypotheses are as follows

1) Null hypothesis H_0: $\boldsymbol{b} = 0$

2) Alternative hypothesis H_a: $\boldsymbol{b} \neq 0$

b. The appropriate test is a *t* test–last line of Exhibit 17.9– the computer program calculates the *t* value and the *p* value

PRACTICING MARKETING RESEARCH

Using Regression Analysis for Key Driver Analysis

1. **Is there one, or more than one, dependent variable?**

Regression models call for one dependent variable.

2. **Is the relationship being modeled linear or non-linear?**

This would require the observation of a scatterplot of the data.

Single dependent variable

Questions

1. **What is key driver analysis? What role does regression analysis play in this type of analysis?**

Key Drive analysis always involves at least one dependent or criterion variable and one or typically multiple independent or predictor variables, who effect on the dependent variable needs to be understood.

2. **How can you use the results from key driver to improve, say, customer satisfaction? Explain.**

Once the predictor variables have been identified, a regression analysis will jointly associate the values associated with the independent variables with the dependent variable. Examination of the strength and direction of the beta coefficients can tell the research which of the independent variables is the best predictor of the dependent variable.

3. CORRELATION ANALYSIS

I. **Correlation for Metric Data: Pearson's Product Moment Correlation**

A. **Analysis between Two Variables**

1. **Correlation Defined**–degree to which changes in one variable (the dependent variable) are associated with the changes in another

a. **Correlation Analysis Defined**–analysis of the degree to which changes in one variable are associated with changes in another

b. **Pearson's Product Moment Correlation**–correlation analysis technique for use with metric data

2. **Coefficient of Correlation**–***R***, is a measure of the degree of association between ***X*** and ***Y***. It is the square root of the coefficient of determination. It can range from –1(perfect negative correlation) to +1(perfect positive correlation).

$$R = \pm\sqrt{R^2}$$

PRACTICING MARKETING RESEARCH:

Do Your "BESD" When Explaining Correlation Results

Questions

1. **What are two traditional ways to explain correlation results to a client? Describe each.**

   Given a significant correlation between two variables, the two traditional ways are: first, the strength of the relationship (what is the probability of insignificance or confidence) and two, what is the direction of the relationship.

2. **What is the BESD approach? Does it make results easier to digest by clients? Why/ why not?**

   The BESD approach is a flexible technique because it can be used with any kind of data. Responses will vary to the second part of the technique.

QUESTIONS FOR REVIEW AND CRITICAL THINKING

2. A sales manager of a life insurance firm administered a standard multiple-item job satisfaction scale to all the members of the firm's sales force. The manager then correlated (Pearson's product-moment correlation) job satisfaction score with years of school completed for each salesperson. The resulting correlation was .11. On the basis of this evidence, the sales manager concluded: "A salesperson's level of education has little to do with his or her job satisfaction." Would you agree or disagree with this conclusion? Explain the basis for your answer.

A correlation coefficient of .11 is low and the conclusion may be correct. This conclusion should be tested with a significance test on the correlation coefficient to provide greater certainty. Hence, what is the probability that a significant relationship does indeed exist. The student should note that statistical analysis cannot lead the decision maker, but the decision maker uses statistical analysis to assist in making recommendations.

3. What purpose does a scatter diagram serve?

A scatter diagram is used to plot data observations to determine if the relationship appears to be linear, curvilinear, or non existent. This allows the researcher to determine whether using linear measures of association (e.g. Pearson's product-moment correlation) would be appropriate.

4. Explain the meaning of the coefficient of determination. What does this coefficient tell the researcher about the nature of the relationship between the dependent and independent variables?

The coefficient of determination tells the marketing researcher how much variation in the dependent variable can be explained by variation in the independent variable. It is a measure of the strength of the relationship. If the coefficient of determination is low, the independent variable does not have significant explanatory power in predicting changes in the dependent variable.

5. It has been observed in the past that when an AFC team wins the Super Bowl, the stock market rises in the first quarter of the year in almost every case. When an NFC team wins the Super Bowl, the stock market falls in the first quarter in most cases. Does this mean that the direction of movement of the stock market is caused by which conference wins the Super Bowl? What does this example illustrate?

It is pretty hard to imagine a cause and effect relationship. This example is almost certainly an example of a spurious relationship.

6. The following table gives the data collected for a convenience store chain for 20 of its stores.

Column 1 - ID number for each store

Column 2 - Annual sales for the store for the previous year in thousands of dollars.

Column 3 - Average number of vehicles that pass the store each day, based on actual traffic counts for one month.

Column 4 - Total population that lives within a 2-mile radius of the store, based on 1990 census data.

Column 5 - Median family income for households within a 2-mile radius of the store based on 2000 census data. See text for data.

Answer the following:

a. Which of the other three variables is the best predictor of sales? Compute correlation coefficients to answer the question.

- For the correlation between sales and traffic, $r = .769$
- For the correlation between sales and population, $r = .418$
- For the correlation between sales and average income, $r = -.418$
- Of the three variables, traffic is the most predictive of sales.

b. Do the following regressions.

1. Sales as a function of average daily traffic.

Sales = 305 +.013*(Traffic), $\boldsymbol{R^2} = .591$ (note: sales measured in thousands of dollars)

2. Sales as a function of population in two-mile radius.

Sales = 539 + .015*(Population), $\boldsymbol{R^2} = .175$ (note: sales measured in thousands of dollars)

c. Interpret the results of the two regressions.

Traffic and population have direct positive relationships with sales. As each one increases, sales increases. The $\boldsymbol{R^2}$ value indicates that the variance in traffic explains more of the variance in sales than does the variation in population.

7. Interpret the following:

a. $Y = .11 + .009X$, where Y is the likelihood of sending children to college and X is family income in thousands of dollars. Remember, it is family income in *thousands*.

1. According to our model how likely is a family with an income of $100,000 to send their children to college?

Y = 0.11 + .009(100) = 1.01 according to the model, but probability cannot be greater than 1.00

2. What is the likelihood for a family with an income of $50,000?

Y = 0.11 + .009(50) = .56

3. What is the likelihood for a family with an income of $17,500?

Y = 0.11 + .009(17.5) = .268

4. Is there some logic to the estimates? Explain.

Yes, those families with higher income would be more likely to be able to afford college for their children.

b. $Y = .25 - .039X$, where Y is the likelihood of going to a skateboard park and X is age.

1. According to our model, how likely is a 10 year old to go to a skateboard park?

Y = .25 - .0039(10) = .211

2. What is the likelihood for a 60 year old?

Y = .25 - .0039(60) = .016

3. What is the likelihood for a 40 year old?

Y = .25 - .0039(40) = .094

4. Is there some logic to the estimates? Explain.

Yes, it is logical that as you get older you are less likely to go to a skateboard park. Children tend to skateboard more than older adults.

8. The following ANOVA summary data are the result of a regression with sales per year (dependent variable) as a function of promotion expenditures per year (independent variable) for a toy company.

$$F = \frac{MSR}{MSE} = \frac{34{,}276}{4{,}721}$$

The degrees of freedom are 1 for the numerator and 19 for the denominator. Is the relationship statistically significant at $\alpha = .05$? Comment.

With this summary data, we can compute an *F*-value of 7.26. With one and 19 degrees of freedom at $\alpha = .05$, the tabulated or critical *F*-value is 4.38. Thus the null hypothesis is rejected, and we can conclude that there seems to be a relationship between sales and promotion expenditures.

REAL-LIFE RESEARCH

Case 17.1 – Axcis Athletic Shoes

Key Points:

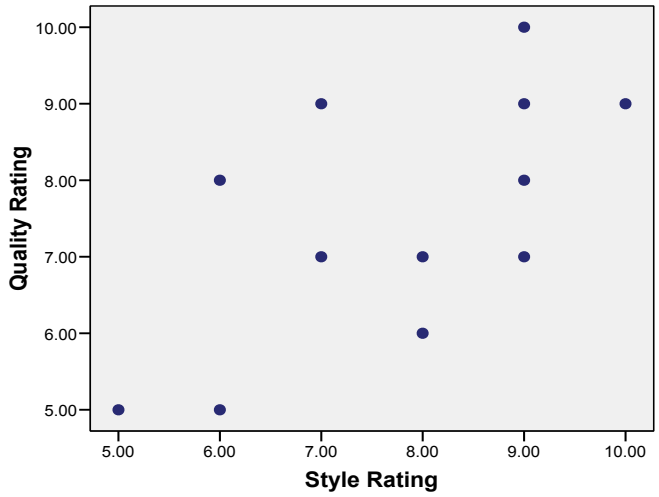
- Axcis' management wants to know whether there is a relationship between consumers' perceptions of style in an athletic shoe and their perception of quality.

Questions

1. Which of the statistical procedures covered in this chapter is appropriate for addressing Fred's theory? Why would you choose that technique over the others?

Pearson's product-moment correlation would be appropriate, as it is designed to quantify the nature and strength of association between variables measured on metric scales. We could also run bivariate regression if a predictive model is desired.

As a check, we should examine a scatter diagram to ensure that a measure of *linear* association such as Pearson's correlation is appropriate. The plot reveals no obvious non-linear relation and thus Pearson's correlation would be appropriate.



2. Use the technique that you choose to determine whether Fred's theory is supported by the statistical evidence. State the appropriate null and alternative hypothesis. Is Fred's theory supported by the statistical evidence? Why or why not?

Fred's theory is represented by the alternative hypothesis in an inference test for the correlation.

H_0: There is not a positive association between perceptions of style and perceptions of quality.

H_a: There is a positive association between perceptions of style and perceptions of quality.

The computed correlation is .634, with a *p*-value of .027. Thus, Fred's theory (Ha) is supported by this data at the $\alpha = .05$ level.