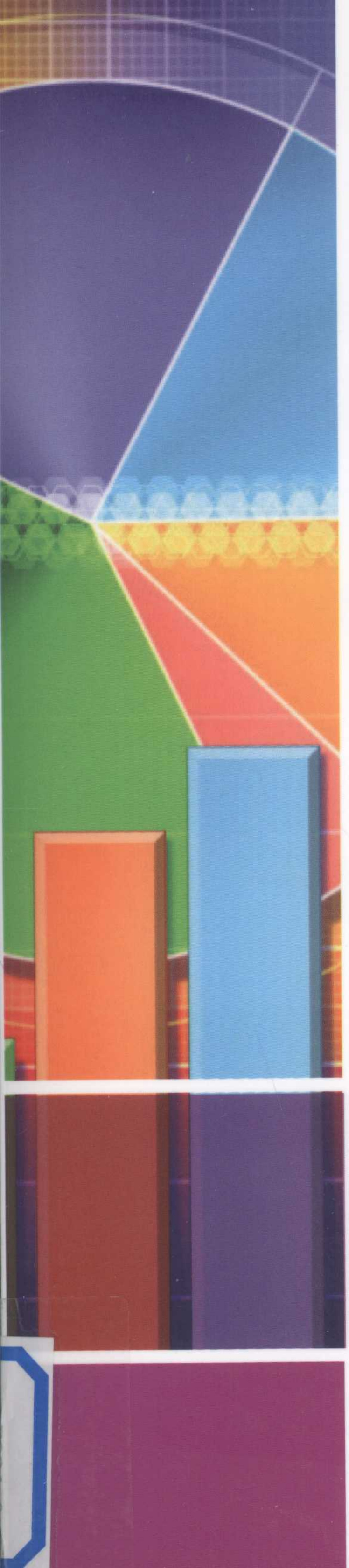

THE QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION

Business Applications of Multiple Regression

Ronny Richardson

businessexpert
Press

www.businessexpertpress.com

Business Applications of Multiple Regression

Ronny Richardson
Southern Polytechnic State University

business**expert**
Press

Business Applications of Multiple Regression

First published in 2011 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-160649-231-4 (paperback)
ISBN-13: 978-160649-232-1 (e-book)
DOI 10.4128/ 9781606492321

A publication in the Business Expert Press Quantitative Approaches to Decision Making collection

Collection ISSN: forthcoming (print)
Collection ISSN: forthcoming (electronic)

Cover design by Jonathan Pennell
Interior design by Scribe, Inc.
First edition: July 2011

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

Abstract

This book describes the use of the statistical procedure called multiple regression in business situations, including forecasting and understanding the relationships between variables. The book assumes a basic understanding of statistics but reviews correlation analysis and simple regression to prepare the reader to understand and use multiple regression.

The techniques described in the book are illustrated using both Microsoft Excel and a professional statistical program. Along the way, several real-world data sets are analyzed in detail to better prepare the reader for working with actual data in a business environment.

This book will be a useful guide to managers at all levels who need to understand and make decisions based on data analysis performed using multiple regression. It also provides the beginning analyst with the detailed understanding required to use multiple regression to analyze data sets.

Keywords

correlation, forecasting, model building, multiple regression, regression

Contents

Introduction

Imagine that you are a business owner with a couple of years' worth of data. You have monthly sales figures, your monthly marketing budget, a rough estimate of the monthly marketing budget for your major competitors, and a few other similar variables. You desperately want this data to tell you something. Not only that, you are sure it can give you some business insights if you know more. But what exactly can the data tell you? And once you have a clue what the data might tell you, how do you get to that information?

Really large companies have sophisticated computer software to do data mining. Data mining refers to extracting or "mining" knowledge from large amounts of data.[1] Stated another way, data mining is the process of analyzing data and converting that data into useful information. But how, specifically?

While data mining uses a number of different statistical techniques, the one we will focus on in this book is multiple regression. Why study multiple regression? The reason is the insight that the analysis provides. For example, knowing how advertising, promotion, and packaging might impact sales can help you decide where to budget your marketing dollars. Or knowing how price, advertising, and competitor spending affect demand can help you decide how much to produce. In general, we use multiple regression either to explain the behavior of a single variable, such as consumer demand, or to forecast the future behavior of a single variable, such as sales.

Before you can understand the operation of multiple regression and how to use it to analyze large data sets, you must understand the operation of two simpler techniques: correlation analysis and simple regression. Understanding these two techniques will greatly aid your understanding of multiple regression.

Correlation analysis measures the strength of the linear relationship between a pair of variables. Some pairs of variables, such as sales and advertising or education and income, will have a strong relationship whereas others, such as education and shoe size, will have a weak relationship. We will explore correlation analysis in more detail in chapter 1. As part of that discussion, we will see what it means for a relationship to be linear as well as what it means for the relationship to be strong or weak and positive or negative.

When a pair of variables has a linear relationship, *simple regression* calculates the equation of the line that describes that relationship. As part of simple regression, one variable will be designated as an independent, or explainer, variable and the other will be designated as a dependent, or explained, variable. We will explore simple regression in more detail in chapter 2.

Sometimes, a single variable is all we need to explain the behavior of the dependent variable. However, in business situations, it almost always takes multiple variables to explain the behavior of the dependent variable. For example, due to the economy and competitor actions, it would be a rare business in which advertising alone would adequately explain sales. Likewise, height alone is not enough to explain someone's weight. *Multiple regression* is an extension of simple regression that allows for the use of multiple independent or explainer variables. We will explore multiple regression in more detail in chapter 3.

When using multiple regression with its multiple independent variables, we face the issue of deciding which variables to leave in the final model and which variables to drop from the final model. This issue is made complex by the "diseases" that can affect multiple regression models. We will explore building complex multiple regression models in more detail in chapter 4. It is when we get to model building that we will begin to see the real-world use of multiple regression.

This book assumes you have a background in statistics. Specifically, we will use the normal distribution, Student *t*-distribution, and *F* distribution to perform hypothesis tests on various model parameters to see if they are significant. While it is helpful if you are familiar with these concepts, it is not essential. The software today is advanced enough to present the results in such a way that you can easily judge the significance of a parameter without much statistical background. A brief review is provided in chapter 1.

Correlation, simple regression, and multiple regression can all be performed using any version of Microsoft Excel. Most readers will be able to perform all their analyses in Excel. However, some of the advanced features of multiple regression require an actual statistical package. There are many fine ones on the market, and any of them will perform all the techniques we will discuss. The examples in this book are all either from Excel or from a statistical package called SPSS.

CHAPTER 1

Correlation Analysis

We begin preparing to learn about multiple regression by looking at correlation analysis. As you will see, the basic purpose of correlation analysis is to tell you if two variables have enough of a relationship between them to be included in a multiple regression model. Also, as we will see later, correlation analysis can be used to help diagnose problems with a multiple regression model.

Take a look at the chart in Figure 1.1. This scatterplot shows 26 observations on 2 variables. These are actual data. Notice how the points seem to almost form a line? These data have a *strong correlation*—that is, you can imagine a line through the data that would be a close fit to the data points. While we will see a more formal definition of correlation shortly, thinking about correlation as data forming a straight line provides a good mental image. As it turns out, many variables in business have this type of linear relationship, although perhaps not this strong.

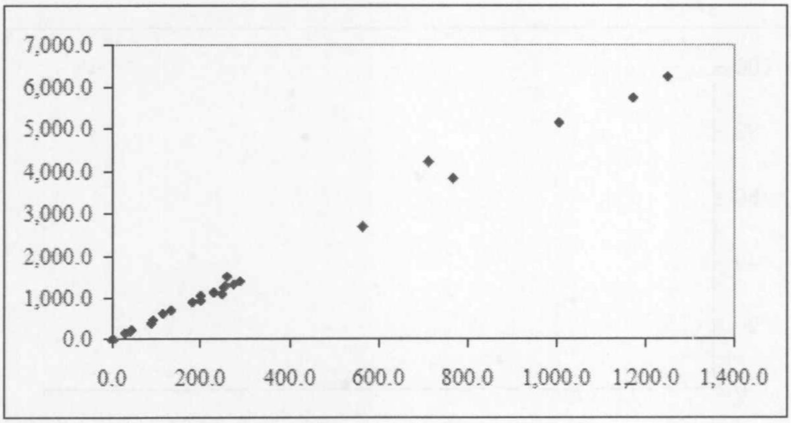

Figure 1.1. A scatterplot of actual data.

Now take a look at the chart in Figure 1.2. This scatterplot also shows actual data. This time, it is impossible to imagine a line that would fit the data. In this case, the data have a very weak correlation.

Terms

Correlation is only able to find, and simple regression and multiple regression are only able to describe, *linear relationships*. Figure 1.1 shows a linear relationship. Figure 1.3 shows a scatterplot in which there is a perfect relationship between the *X* and *Y* variables, only not a linear one (in this case, a sine wave.) While there is a perfect mathematical relationship between *X* and *Y*, it is not linear, and so there is no linear correlation between *X* and *Y*.

A *positive* linear relationship exists when a change in one variable causes a change in the same direction of another variable. For example, an increase in advertising will generally cause a corresponding increase in sales. When we describe this relationship with a line, that line will have a positive slope. The relationship shown in Figure 1.1 is positive.

A *negative* linear relationship exists when a change in one variable causes a change in the opposite direction of another variable. For example, an increase in competition will generally cause a corresponding

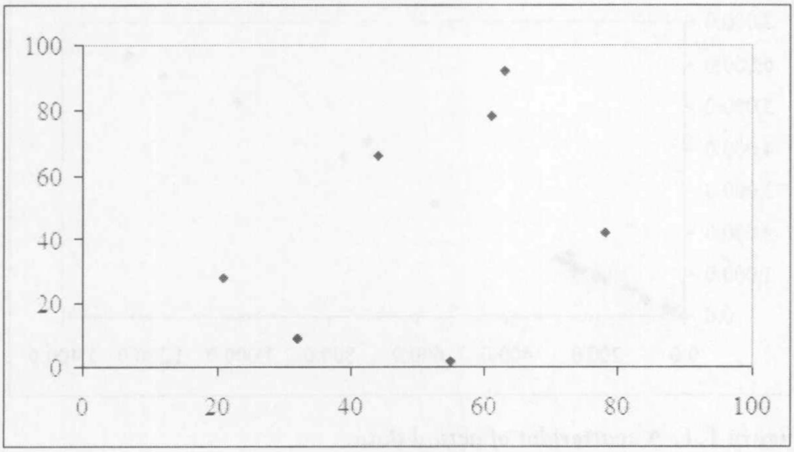

Figure 1.2. Another scatterplot of actual data.

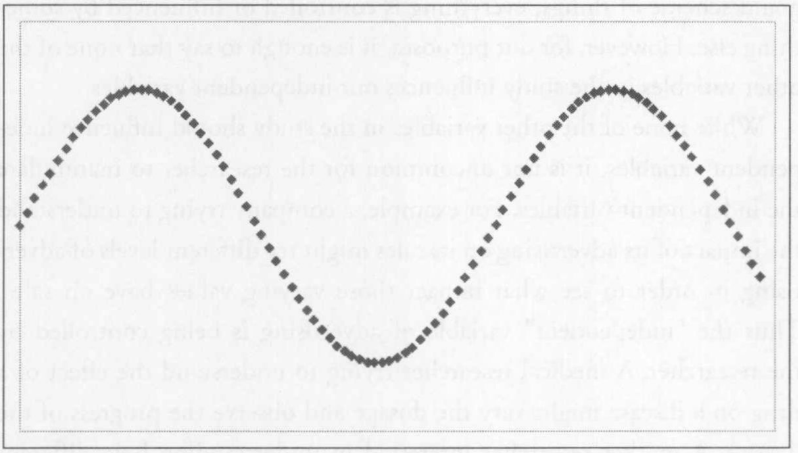

Figure 1.3. A scatterplot of nonlinear (fictitious) data.

decrease in sales. When we describe this relationship with a line, that line will have a negative slope.

Having a positive or negative relationship should not be seen as a value judgment. The terms "positive" and "negative" are not intended to be moral or ethical terms. Rather, they simply describe whether the slope coefficient is a positive or negative number—that is, whether the line slopes up or down as it moves from left to right.

While it does not matter for correlation, the variables we use with regression fall into one of two categories: *dependent* or *independent* variables. The dependent variable is a measurement whose value is controlled or influenced by another variable or variables. For example, someone's weight likely is influenced by the person's height and level of exercise, whereas company sales are likely greatly influenced by the company's level of advertising. In scatterplots of data that will be used for regression later, the dependent variable is placed on the Y-axis.

An independent variable is just the opposite: a measurement whose value is not controlled or influenced by other variables in the study. Examples include a person's height or a company's advertising. That is not to say that nothing influences an independent variable. A person's height is influenced by the person's genetics and early nutrition, and a company's advertising is influenced by its income and the cost of advertising. In the grand scheme of things, everything is controlled or influenced by something else. However, for our purposes, it is enough to say that none of the other variables in the study influences our independent variables.

While none of the other variables in the study should influence independent variables, it is not uncommon for the researcher to manipulate the independent variables. For example, a company trying to understand the impact of its advertising on its sales might try different levels of advertising in order to see what impact those varying values have on sales. Thus the "independent" variable of advertising is being controlled by the researcher. A medical researcher trying to understand the effect of a drug on a disease might vary the dosage and observe the progress of the disease. A market researcher interested in understanding how different colors and package designs influence brand recognition might perform research varying the packaging in different cities and seeing how brand recognition varies.

When a researcher is interested in finding out more about the relationship between an independent variable and a dependent variable, he must measure both in situations where the independent variable is at differing levels. This can be done either by finding naturally occurring variations in the independent variable or by artificially causing those variations to manifest.

When trying to understand the behavior of a dependent variable, a researcher needs to remember that it can have either a *simple* or *multiple* relationship with other variables. With a simple relationship, the value of the dependent variable is mostly determined by a single independent variable. For example, sales might be mostly determined by advertising. Simple relationships are the focus of chapter 2. With a multiple relationship, the value of the dependent variable is determined by two or more independent variables. For example, weight is determined by a host of variables, including height, age, gender, level of exercise, eating level, and so on, and income could be determined by several variables, including raw material and labor costs, pricing, advertising, and competition. Multiple relationships are the focus of chapters 3 and 4.

Scatterplots

Figures 1.1 through 1.3 are scatterplots. A scatterplot (which Microsoft Excel calls an *XY chart*) places one variable on the Y-axis and the other on the X-axis. It then plots pairs of values as dots, with the *X* variable determining the position of each dot on the X-axis and the *Y* variable likewise determining the position of each dot on the Y-axis. A scatterplot is an excellent way to begin your investigation. A quick glance will tell you whether the relationship is linear or not. In addition, it will tell you whether the relationship is strong or weak, as well as whether it is positive or negative.

Scatterplots are limited to exactly two variables: one to determine the position on the X-axis and another to determine the position on the Y-axis. As mentioned before, the dependent variable is placed on the Y-axis, and the independent variable is placed on the X-axis.

In chapter 3, we will look at multiple regression, where one dependent variable is influenced by two or more independent variables. All these variables cannot be shown on a single scatterplot. Rather, each independent variable is paired with the dependent variable for a scatterplot. Thus having three independent variables will require three scatterplots. We will explore working with multiple independent variables further in chapter 3.

Data Sets

We will use a couple of data sets to illustrate correlation. Some of these data sets will also be used to illustrate regression. Those data sets, along with their scatterplots, are presented in the following subsections.

All the data sets and all the worksheets and other files discussed in this book are available for download from the Business Expert Press website (http://www.businessexpertpress.com/books/business-applications-multiple-regression). All the Excel files are in Excel 2003 format and all the SPSS files are in SPSS 9.0 format. These formats are standard, and any later version of these programs should be able to load them with no difficulty.

Number of Broilers

Figure 1.1 showed the top 25 broiler-producing states for 2001 by both numbers and pounds, according to the National Chicken Council. The underlying data are shown in Table 1.1.

Table 1.1. Top 25 Broiler-Producing States in 2001

State	Number of broilers (millions)	Pounds liveweight (millions)
Georgia	1,247.3	6,236.5
Arkansas	1,170.9	5,737.3
Alabama	1,007.6	5,138.8
North Carolina	712.3	4,202.6
Mississippi	765.3	3,826.5
Texas	565.5	2,714.4
Delaware	257.7	1,494.7
Maryland	287.8	1,381.4
Virginia	271.5	1,330.4
Kentucky	253.4	1,292.3
California	250.0	1,250.0
Oklahoma	226.8	1,111.3
Missouri	245.0	1,100.0
South Carolina	198.0	1,049.4
Tennessee	198.3	932.0
Louisiana	180.0	890.0
Pennsylvania	132.3	701.2
Florida	115.3	634.2
West Virginia	89.8	368.2
Minnesota	43.9	219.5
Ohio	40.1	212.5
Wisconsin	31.3	137.7
New York	2.3	12.2
Hawaii	0.9	3.8
Nebraska	0.5	2.7
Other	92.4	451.0

Age and Tag Numbers

Figure 1.2 was constructed by asking seven people their age and the last two digits of their car tag number. The resulting data are shown in Table 1.2. As you can imagine, there is no connection between someone's age and that person's tag number, so this data does not show any strong pattern. To the extent that any pattern at all is visible, it is the result of sampling error and having a small sample rather than any relationship between the two variables.

Return on Stocks and Government Bonds

The data in Table 1.3 show the actual returns on stocks, bonds, and bills for the United States from 1928 to 2009.[1] Since there are three variables (four if you count the year), it is not possible to show all of them in one scatterplot. Figure 1.4 shows the scatterplot of stock returns and treasury bills. Notice that there is almost no correlation.

Federal Civilian Workforce Statistics

Table 1.4[2] shows a state-by-state breakdown of the number of federal employees and their average salaries for 2007. Figure 1.5 shows the resulting scatterplot. Notice that there appears to be a fairly weak linear relationship.

Table 1.2. Age and Tag Number

Age	Tag no.
55	2
21	28
78	42
61	78
44	66
63	92
32	9

Table 1.3. Return on Stocks and Government Bonds

Year	Stocks (%)	Treasury bills (%)	Treasury bonds (%)
1928	43.81	3.08	0.84
1929	-8.30	3.16	4.20
1930	–25.12	4.55	4.54
1931	–43.84	2.31	–2.56
1932	–8.64	1.07	8.79
1933	49.98	0.96	1.86
1934	–1.19	0.32	7.96
1935	46.74	0.18	4.47
1936	31.94	0.17	5.02
1937	–35.34	0.30	1.38
1938	29.28	0.08	4.21
1939	–1.10	0.04	4.41
1940	–10.67	0.03	5.40
1941	–12.77	0.08	–2.02
1942	19.17	0.34	2.29
1943	25.06	0.38	2.49
1944	19.03	0.38	2.58
1945	35.82	0.38	3.80
1946	–8.43	0.38	3.13
1947	5.20	0.57	0.92
1948	5.70	1.02	1.95
1949	18.30	1.10	4.66
1950	30.81	1.17	0.43
1951	23.68	1.48	–0.30
1952	18.15	1.67	2.27
1953	–1.21	1.89	4.14
1954	52.56	0.96	3.29
1955	32.60	1.66	–1.34
1956	7.44	2.56	–2.26
1957	–10.46	3.23	6.80
1958	43.72	1.78	–2.10
1959	12.06	3.26	–2.65
1960	0.34	3.05	11.64
1961	26.64	2.27	2.06