Using Cigarette Data for An Introduction to Multiple Regression

Lauren McIntyre
North Carolina State University

Journal of Statistics Education v.2, n.1 (1994)

Key Words: Classroom data; Collinearity; Outlier.

Abstract

The CIGARETTE dataset contains measurements of weight and tar, nicotine, and carbon monoxide content for 25 brands of domestic cigarettes. The dataset is useful for introducing the ideas of multiple regression and provides examples of an outlier and a pair of collinear variables.

1. Introduction

1 The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. The United States Surgeon General considers each of these substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.

2 The dataset presented here contains measurements of weight and tar, nicotine, and carbon monoxide (CO) content for 25 brands of cigarettes. Students familiar with simple linear regression can use these data to develop an understanding of multiple regression techniques. Students will discover that there is an outlier in this dataset (where we define an outlier as a point not near the rest of the data) and that tar and nicotine are collinear variables. These characteristics of the data can lead to a very good discussion and an enhanced understanding of regression when introduced carefully in class.

3 The data were taken from Mendenhall and Sincich (1992). The original source of the data is the Federal Trade Commission. We use the dataset as part of a computer demonstration. We lead the students into a discussion of the outlying point, collinearity, and multiple regression, using a question and answer format. This paper presents a summary of the data presentation.

2. The Computer Demonstration

4 Our demonstration involves two people -- an instructor who presents the lecture and initiates discussion, and another who operates the computer. We use the software SAS Insight. (SAS and SAS Insight are Registered Trademarks of SAS Institute, Cary, NC.) An overhead LCD panel/projection system displays the graphics and analyses on a wall screen.

5 First we introduce the dataset and ask students how they would analyze the data. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. This information motivates our goal of predicting the variable CO. The preliminary analysis involves two simple linear regressions. We find it helpful to place the models for the simple linear regressions on the blackboard. We plot the data and perform the two regressions. Each plot (CO*Tar and CO*Nicotine) shows a positive linear relationship. Both regressions have slopes that are significantly larger than zero.

6 From these simple analyses it is easy to motivate the use of multiple regression, using the variables tar and nicotine simultaneously. We give a short description of the multiple regression model. We also give other examples in which multiple regression is extremely useful, such as crime prediction and economic forecasts.

7 Next the real fun begins. SAS Insight has a rotating three-dimensional plotting option that allows students to really see and grasp the concept of a multiple regression in two dimensions. We identify simple regression as a line, this multiple regression as a plane, and note that more variables would involve a higher than three-dimensional space.

8 The three-dimensional picture clearly shows an outlier, which we highlight. We run the multiple regression and notice two interesting developments: the coefficient of nicotine has changed sign, and it is no longer significant. We let students discover and discuss these unexpected results. At this point, we examine the outlier. We determine that the Bull Durham cigarette (observation 3) has more tar and nicotine than the others. After emphasizing that the exclusion of an outlier requires a good reason, we redo the analysis excluding the outlier. Removal of the outlier results in positive slopes for tar and nicotine, as observed in the simple linear regressions, but the coefficient of nicotine is still not significant.

9 The non-significance of the nicotine coefficient is a concern, and we ask students to suppose that the two independent variables, tar and nicotine, are related. Is this good or bad? With prompting, students can see that having highly correlated independent variables is not desirable from the point of view of statistical accuracy. In addition, the use of redundant independent variables could waste time and money on unnecessary data collection.

10 We investigate the correlation between tar and nicotine and discover that tar and nicotine are highly correlated, and that each is correlated with CO. From this we conclude that there is a strong relationship between the independent variables, and that only one of these variables should be used as a predictor of CO.

11 This provides an opportunity to talk about the importance of understanding the variables. The source of the carbon in the CO gas is relevant here. Also, there is controversy over the methods used to obtain these data by the Federal Trade Commission (FTC). Some studies (Davis et al. 1990, Coultas et al. 1993) have indicated that the tar, nicotine, and carbon monoxide content measured via smoking machines (as used to gather this dataset) may have little to do with actual human exposure to these chemicals. In other words, the relationship between the FTC cigarette data and biological markers of nicotine and tar uptake in humans is weak. These studies recommend that smokers be warned that the content of tar and nicotine given in cigarette advertisements may not reflect the relative strengths of cigarette brands. One factor in this human dilemma is that people smoking `lighter' brands, i.e., brands lower in labelled tar and nicotine, may smoke more

cigarettes or inhale for longer periods than if they smoked cigarettes with higher contents of tar and nicotine.

12 These data help students understand multiple regression. The data analysis clearly demonstrates the concepts of multiple regression that students often find difficult. We have used this dataset successfully in an undergraduate introductory course for non-statistics majors. Others in our department have used it at the graduate level to illustrate the dangers of collinearity.

3. Getting the Data

13 The file cigarettes.dat.txt contains the raw data. The file cigarettes.txt is a documentation file containing a brief description of the dataset.

Appendix - Key to Variables in cigarettes.dat.txt

Brand name
Tar content (mg)
Nicotine content (mg)
Weight (g)
Carbon monoxide content (mg)

Values are delimited by blanks. There are no missing values.

References

Coultas, D.B., Stidley, C.A., and Samet, J.M. (1993), "Cigarette Yields of Tar and Nicotine and Markers of Exposure to Tobacco Smoke," *American Review of Respiratory Disease*, 148, 435-440.

Davis, R. M., Healy, P., and Hawk, S.A. (1990), "Information on Tar and Nicotine Yields on Cigarette Packages," *American Journal of Public Health*, 80, 551-553.

Mendenhall, W., and Sincich, T. (1992), *Statistics for Engineering and the Sciences* (3rd ed.), New York: Dellen Publishing Co.

Lauren McIntyre
Department of Statistics, Box 8203
North Carolina State University
Raleigh, NC 27695-8203

mcintyre@stat.ncsu.edu

Return to Table of Contents | Return to the JSE Home Page