EXHIBIT 14

APPLIED STATISTICS FOR PUBLIC POLICY

BRIAN P. MACFIE and PHILIP M. NUFRIO

Library of Congress Cataloging-in-Publication Data

Macfie, Brian P., 1955–
Applied statistics for public policy / by Brian P. Macfie and Philip M. Nufrio.
p. cm.
Includes bibliographical references and index.
ISBN 0-7656-1239-9 (cloth : alk. paper)
1. Social sciences—Statistical methods. 2. Political statistics. I. Nufrio, Philip M. II. Title.

HA29.M185 2005
519.5—dc22 2004023626

Printed in the United States of America

The paper used in this publication meets the minimum requirements of
American National Standard for Information Sciences
Permanence of Paper for Printed Library Materials,
ANSI Z 39.48-1984.

∞

BM (c) 10 9 8 7 6 5 4 3 2 1

Brief Ta

over a range for which data is encountered. Second, a convenient statistical technique, called the *least squares method*, can be used to estimate parameters for linear equations.

Returning to the first hypothesis in Chapter 16 (i.e., test grades are determined by the number of hours studied), how do we do this analysis? Actually, the procedure is not that all that difficult, it is just tedious. What we need to do is estimate the coefficients for the straight-line (i.e., b_0 and b_1) equation that fits the data better than any other. If we look back to Figure 17.1, we said that the trend line summarized the data. The truth is, for the same scatter of observations between test grades and hours of study, there is actually an infinite number of straight lines that could be run through the data. As an example, we could have drawn a straight line through the highest and lowest pair, we could have drawn a straight line between the two middle pairs and extrapolated it, or we could even "eyeball" a straight line. But, mathematically, there is one and only one straight line that best fits (or best describes) this data.

The method that calculates the straight-line coefficients that generally best fits any scatter of data points is the *least squares* method. The idea is to estimate a straight line such that the difference between the sum of the squared deviations from the fitted straight line and all the data points is minimized. The difference between the straight line and each actual observation is called a *residual* or regression error. This is the difference between an actual value of Y and a predicted value of Y (from the equation) for a given value of X. Similar to what was seen in Chapters 7 and 10 with sampling error, the regression error does not imply a mistake has been made. It is the difference between estimated and actual. Since we are running a straight line through a scatter of data points, we expect some deviation to exist between the line and the observations. The residual or regression error should also be random, and is just as likely to be above the line as below it.

Although there are other ways of estimating functional relationships with regression analysis, the least squares approach is the one most commonly used. There are a host of theoretical reasons for selecting the equation where the sum of the squared errors is minimized. The explanation lies in statistical theory that is beyond the scope of this book, and the reader is invited to consult any of the references in the bibliography at the end of this chapter for these details. However, an intuitive rationale for using the least squares method can be presented. If we estimate a line where the errors are not squared, the distances below the line would cancel out the distances above the line. Thus, it is theoretically possible to have more than one line to describe the data, all of which could minimize the sum of the errors. By squaring the errors, however, only one line would have a lower total sum of squared errors than all others. Moreover, it is implicit that the further an observation is away from the trend line, the greater the magnitude of the regression error. If a regression error of six is *more* than twice as variant as a regression error of three, we have to account for the greater regression error for the six. One way to do this is to apply more weight to the six by squaring all the errors and then estimating the line that minimizes the sum of these squared errors.

The answer is to estimate the coefficients for the straight-line equation (i.e., b_0 and b_1) that minimizes the sum of the squared differences. To accomplish the task of estimating a straight-line equation using the least squares method, we need to calculate the following:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\Sigma X)^2} \tag{17.2}$$

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{17.3}$$

Table 17.2
Tabular Method to Calculate Regression Coefficients

Grade on Test (Y)	Hours of Work (X)	XY	X-squared
60	26.0	1,560.0	676.0
83	18.0	1,494.0	324.0
76	20.0	1,520.0	400.0
65	24.0	1,560.0	576.0
94	10.0	940.0	100.0
68	28.0	1,904.0	784.0
79	22.0	1,738.0	484.0
89	8.0	712.0	64.0
80	12.0	960.0	144.0
56	32.0	1,792.0	1,024.0
750	200.0	14,180.0	4,576.0

Mean for Y =750 ÷ 10 = 75.0; Mean for X = 200 ÷ 10 = 20.0

Solution. Using the tabular approach, first start by calculating the inputs necessary for the estimate of b_0 and b_1. This is shown in Table 17.2. The results of Table 17.2 are then substituted into the general regression equations to render Equations (17.7), (17.8), and (17.9).

$$b_1 = -1.42 = \frac{10 * 14{,}180 - 200 * 750}{10 * 4{,}576 - 200^2} \tag{17.7}$$

$$b_0 = 103.4 = 75 - (-1.42 * 20.0) \tag{17.8}$$

$$\hat{Y} = 103.4 - 1.42(X). \tag{17.9}$$

The linear regression equation in Equation (17.9) would be interpreted as follows. Based on the data, if a randomly selected student did not work, the expected grade would be 103.4. This is the value of $\hat{Y}$ when $X = 0$ (in other words, it is the value of the intercept b_0).

A word of caution is warranted here. Any interpretation of the intercept b_0 should be done with extreme care. Sometimes the intercept value has no real practical meaning, even when it is statistically significant. For example, although it is possible that a student does not work, it is not possible to score 103.4 on a test when the maximum score possible is 100. As a general rule, if there are not many observations of (the independent variable) X near zero, the intercept might not have much statistical meaning.

On the other hand, if the slope coefficient b_1 is statistically significant, it may have considerable meaning. In our example, the b_1 coefficient in the equation suggests that for each additional hour of work, we can expect the grade for a randomly selected student to decrease by –1.42 points. Finally, if a randomly selected student worked for 15 hours, we would expect the test grade to be about 82 ($\hat{Y}$ = 82.1 = 103.4 – 1.42 * 15).

Statistical Inferences About Regression Coefficients

We now need to address the issue of making inferences about the linear regression coefficients b_0 and b_1. As will be seen, what we are essentially doing is a hypothesis test about a potentially existing relationship between a dependent and independent variable(s) using these coefficients.

Inferences Regarding the Intercept b_0

To start, although we can conduct a test of statistical significance for the intercept, we should always be careful about any interpretation thereafter. As mentioned earlier, caution must be exercised when interpreting points outside the range of the data, and, typically, the intercept lies outside this range. The intercept is supposed to be interpreted as the value of $\hat{Y}$ when X is zero. When none of the values of X that went into creating the regression equation are near zero, the meaning of the intercept term becomes dubious, even if the intercept is statistically significant.

The procedure to test the intercept for statistical significance is relatively straightforward. In testing the inference of the slope, we must first convert our obtained sample values to t-scores to test for significance (just as we did in the tests concerning means in Chapter 11). We will use the standard error of the slope b_0 to test the relationship between variables. What we are testing is the null hypothesis that no relationship between the dependent and independent variable exists versus the research hypothesis that a relationship does exist. Thus, the testing criteria is as follows.

In an inferential test of intercept b_0 the null hypothesis is:

Ho: $b_1 = 0$ (there is no linear relationship).

The research hypothesis is the opposite:

Ha: $b_1 \neq 0$ (there is a linear relationship).

The hypothesis testing procedure follows the same six-step process covered in earlier chapters. The only difference is that there is a slight adjustment to the way the t-statistic is calculated.[4] Since we are testing that no relationship exists, what we have to do is subtract zero from the estimated intercept term, and divide by the standard error of the intercept b_0. In other words, we subtract zero to denote that a *zero* relationship exists. If we reject the null hypothesis, then we are, in effect, saying that there is evidence of a relationship. If we fail to reject the null hypothesis, we are saying there is no evidence of a relationship. The actual calculation of the t-statistic for the intercept is:

$$t_{b0} = \frac{b_0 - 0}{sb_0}, \tag{17.11}$$

where b_0 = calculated intercept term, and
sb_0 = standard error of the intercept b_0.