BINGHAM McCUTCHEN LLP
DONN P. PICKETT (SBN 72257)
GEOFFREY M. HOWARD (SBN 157468)
HOLLY A. HOUSE (SBN 136045)
ZACHARY J. ALINDER (SBN 209009)
BREE HANN (SBN 215695)
Three Embarcadero Center
San Francisco, CA 94111-4067
Telephone: (415) 393-2000
Facsimile: (415) 393-2286
donn.pickett@bingham.com
geoff.howard@bingham.com
holly.house@bingham.com
zachary.alinder@bingham.com
bree.hann@bingham.com

BOIES, SCHILLER & FLEXNER LLP
DAVID BOIES (Admitted *Pro Hac Vice*)
333 Main Street
Armonk, NY 10504
Telephone: (914) 749-8200
Facsimile: (914) 749-8300
dboies@bsfllp.com
STEVEN C. HOLTZMAN (SBN 144177)
FRED NORTON (SBN 224725)
1999 Harrison St., Suite 900
Oakland, CA 94612
Telephone: (510) 874-1000
Facsimile: (510) 874-1460
sholtzman@bsfllp.com
fnorton@bsfllp.com

DORIAN DALEY (SBN 129049)
JENNIFER GLOSS (SBN 154227)
500 Oracle Parkway, M/S 5op7
Redwood City, CA 94070
Telephone: (650) 506-4846
Facsimile: (650) 506-7114
dorian.daley@oracle.com
jennifer.gloss@oracle.com

Attorneys for Plaintiffs Oracle USA, Inc., *et al.*

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

ORACLE USA, INC., *et al.*,

Plaintiffs,

v.

SAP AG, *et al.*,

Defendants.

CASE NO. 07-CV-01658 PJH (EDL)

EXHIBIT 1 TO THE DECLARATION OF DANIEL S. LEVY, PH.D. IN SUPPORT OF MOTION NO. 1: TO EXCLUDE TESTIMONY OF DEFENDANTS' EXPERT STEPHEN CLARKE

FILED PURSUANT TO DKT. NO. 915

EXHIBIT 1

APPLIED STATISTICS FOR PUBLIC POLICY

BRIAN P. MACFIE and PHILIP M. NUFRIO

M.E.Sharpe
Armonk, New York
London, England

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

income. Assuming this is a positive relationship, as the per capita income in (developed or undeveloped) nations goes up, life expectancy goes up.

Finally, sometimes data sets show no relationship. In this case, there is no pattern of increase and/or decrease. Ironically, it is often just as important to know that there is no relationship as it is to know that one exists.

In this chapter, we will examine how to calculate and interpret the strength of the positive and negative relationships between two variables. Ultimately, we will present some practical illustrations using data from the Microsoft Excel files that accompany this book to examine the a positive relationship between life expectancy and national economic well-being (measured by the gross domestic product [GDP] per capita), and the negative relationship between life expectancy and level of health care (measured by physician access).

Measures of Correlation

With regard to strength of relationship, there are several different measures of correlation that exist. However, we focus on what is commonly accepted as the two traditional measures of correlation, the *coefficient of correlation* (r) and the *coefficient of determination* (r^2). The coefficient of correlation is a value that is, mathematically, always between -1 and $+1$. The sign of the coefficient of correlation (which is often simply referred to as the correlation coefficient) is not, in itself, a measure of strength. It is actually an indication of association or direction of a relationship between a dependent and independent variable. What this means is that if the relationship is positive, then the sign of the coefficient of correlation is positive. If the relationship between a dependent and independent variable is negative, then the sign of the coefficient of correlation is negative.

If a correlation coefficient approaches either $r = +1.0$ or $r = -1.0$, it suggests that a strong relationship exists. A value close to $+1.0$ represents a near perfect positive correlation. A value close to -1.0 represents a near perfect negative correlation. As the correlation coefficient approaches zero, we say there is little or no relationship. Therefore, the closer r is to the *absolute* value of one, the better.

The coefficient of determination (r^2), on the other hand, represents a true measure of strength between a dependent and independent variable. It measures the proportion of total variation in the dependent variable (Y) that is explained or accounted for by the total variation in the independent variable (X). Mathematically, it ranges between 0 and $+1.0$, and is the *square* of the correlation coefficient (r). When we say the $r^2 = +1.0$, what is really being said is that a correlation of 100 percent exists (remember r^2 actually measures a proportion). The main difference between r and r^2 is that the value of r^2 has a more precise meaning and is easier to interpret. Unfortunately, whereas the coefficient of correlation (r) indicates association and lends itself to statistical testing, the coefficient of determination (r^2) does not.

Scatter Plots and Positive or Negative Relationships

With the use of scatter plots, we can easily show a potential relationship (i.e., positive or negative) between variables. In statistical research, hypotheses are often built using a scatter plot to determine whether available data can suggest a relationship exists and tell us something about the

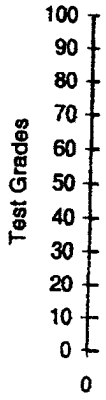

Figure 16.

direction of tha
of the relations

Positive I

An example of
of slope upwar
some kind of hy
go up, the grad
education theor
expect the grap
time and test gr
Figure 16.1).

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.

The calculation for the standard error of the intercept is rather tedious, and is written as

$$sb_0 = \sqrt{\frac{\sum e^2}{(n-k)} * \frac{\sum X_i^2}{n\sum x_i^2}} \tag{17.12}$$

where e^2 = residual variation,
X^2 = the square of independent variable observations,
x^2 = total variance of independent variable observations,
n = sample size, and
k = number of estimated population parameters.

We will soon show how all these calculations are arrived at in Example 17.7 using a tabular approach, but we only intend to do this once, since we will rely on POLYSTAT to do this for us. Once the t-statistic is calculated, the test for significance can be completed using the appropriate degrees of freedom in t-distribution for the desired level of confidence. Normally, the test for significance is a two-tail test (nondirectional, because we are testing for no relationship between variables) at the 0.05 level of significance level and $n - 2$ degrees of freedom.[5]

Inferences Regarding the Slope Coefficient b_1

The slope coefficient, b_1, gives us an estimate of the change in the dependent variable associated with a one-unit change in the independent variable. Again, remember that caution must be exercised when interpreting points outside the range of the data.

In an inferential test of slope, we will attempt to reject the null hypothesis that the variables X and Y are not linearly related. One critically important rule when testing the regression slope from zero is that both variables (dependent and independent) must be defined at the same level of measurement. For example, if the dependent variable is interval-ratio, then the independent variable must be interval-ratio.

In the test of inference of slope, the researcher will use the regressed variables to test the probability that one could draw a sample with a slope of b_1 given that the population slope equals zero. To put it another way, if $b_1 = 0$, there is no relationship between the variables in the population. If we were to draw a sample from such a population, a statistical test could be run to determine if that sample is, or is not, related to the population. Thus, the testing criteria is as follows.

In an inferential test of slope b_1 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).

By rejecting the null hypothesis, the researcher has determined that the slope, b_1, of the population regression line is not zero. (If the null hypothesis cannot be rejected, we are suggesting X to be a totally useless predicator of Y, since the value of X has nothing to do with the distributions of

of deer, it will not significantly endanger the deer population. Using data from the past 40 hunting seasons, the commissioner's policy analyst regresses the number of deer killed during the season on the length in days in the season. The data renders the following equation:

$\hat{Y} = 110{,}000 + 105.1x$, with a standard error of the slope of 51.5. The $r^2 = 9.9$ percent.

How do the different elements of this regression equation help analyze this problem?

Solution. Taking what we know about research and null hypotheses, in simple regression analysis, we will test if there is a relationship between X (length of days in the season) and Y (deer killed). In this case, the research hypothesis is that there exists a relationship between the number of deer killed during the season and the number of days in the hunting season. In contrast, the null hypothesis is that there is no relationship between the number of deer killed in the season and the number of days in the hunting season.

Since the intercept (b_0), the slope (b_1), and the standard error of the slope (Sb_1) are known (i.e., 110,000, 105.1, and 51.5, respectively), the t-statistic can be calculated as

$$t_{b1} = 2.04 = \frac{105.1 - 0}{0.0087}. \tag{17.19}$$

At the 0.05 level of significance and 38 degrees of freedom ($n - 2$), the critical t-statistic is $t = \pm 2.02$. Since the calculated t-statistic of 2.04 is less than the critical $t = -2.02$, it falls into the rejection region.

For each additional day the hunting season for deer is extended, it can be expected that approximately 105 deer will be killed.

Example 17.9

Returning to Example 17.7, assume that the research analyst for the commission has POLYSTAT to do the calculations. Further assume that the analyst wants to first check if the regression equation was calculated correctly before testing the intercept and slope for statistical significance. Use a 0.05 level of significance for the test.

Solution. See the output in Exhibit 17.7.

The regression equations match with the POLYSTAT output, thus it can be concluded that the utility correctly estimated this equation. Note that the calculated values for the test of significance in POLYSTAT are slightly different. The values in POLYSTAT are not rounded. Also note that POLYSTAT tells us to reject the null hypothesis in both cases.

Statistical Inferences About the Entire Regression Equation (the F-Test)

Another useful statistic for measuring the *overall* explanatory power of a regression equation is the F-statistic. Similar to the way the coefficient of determination (r^2) measures overall explanatory power by accounting for the variance in Y explained by the variance of X, the F-statistic considers

Exhibit 17.7

Single Variable Regression Output Report

REGRESSION STATISTICS	
Correlation Coefficient, R	0.990
Coefficient of Determination, R-Square	98.04%
Adjusted R-Square	97.84%
Standard Error of Estimate	10.1599
Observations	12

Dependent Variable =	Usage/Household				
	Coefficients	Standard Error	t Statistic	P-Value	Decision
Intercept	22.93812	4.601	4.9855	0.0005494	Reject Null
Independent Variable = Avg. daily temp.	-0.186096	0.0084	-22.338	0.0000000	Reject Null

Durbin-Watson Statistic =	1.266

ANALYSIS OF VARIANCE					
	df	SS	MS	F-Statistic	Decision
Regression	1	51606.376	51606.376	499.000	Reject Null
Residual	10	1032.293	103.229		
Total	11	52640.669			

FORECAST FOR DEPENDENT VARIABLE Y		
Input Independent Variable =	deg-days	0
Level of Significance =		0.05
Predicted Dependent Variable = Usage/Household		22.9381
Predicted Range of Forecast at 95%	to	46.5758

0.3005

the *relati*
Althoug
easily be
The es
analyzes
Mathema
ables. In
sidual va
terms, th
divided b
To see
standard
D). The t
minus th
value of
In othe
$\hat{Y} = 22.94$
plug 908 i
calculatec
Table 17.
away), w
that *is no*
total varia
Staying
dependent
The total
is the amo
tion, 51,5(
There i
Note that
and the su
values (M
103 = 499
The *F*-s

If, at thi
correct the
determinat

the *relationship* between the explained and unexplained variation in the dependent variable (Y). Although the F-test has greater use in the evaluation of multivariate regression equations, it can easily be introduced and applied to simple regression analysis.

The essentials of the F-distribution were covered in Chapters 11 and 14. In review, the F-statistic analyzes the variance between variables (which, incidentally, is why it is crucial in ANOVA). Mathematically, the F-statistic is the ratio of variances of the dependent and independent variables. In testing the significance of an equation, the measure of random error is the average residual variance (the error sum of squares divided by the degrees of freedom). In mathematical terms, the F-test is the ratio of the mean variance that is accounted for by the regression equation divided by the mean error variance.

To see what all this means, refer to Table 17.3. This table was used to show how to arrive at the standard error of estimate. One of the values calculated was the total residual variation (column D). The total residual variation is nothing more than the actual value of the dependent variable minus the estimated value of the dependent variation (which is solved using the corresponding value of X in the calculated regression equation).

In other words, the use per customer data (Y) and degree-day data (X) rendered an equation of $\hat{Y} = 22.94 + 0.1881_{t}$. The Y-value for the first observation is 204.3 and the X-value is 908. If we plug 908 into the equation, the result is 193.73. The difference between the original 204.3 and the calculated 193.73 is 10.57. The value of 10.57 is residual variation (the first value in column D of Table 17.3). If we square all of the individual residual variations (to make the negative values go away), we end up with the sum of squared residuals. This sum is the amount of the variance in Y that *is not* explained by X. Therefore, the sum of column D in Table 17.3 (1,032 therms) is the total variation not explained by X (degree-days).

Staying with Table 17.3, turn your attention to column H. This is the total variance in the dependent variable Y. It is already known what total variance is, as that was covered in Chapter 4. The total variance in Y (use per customer) is 52,541. Since 52,541 is the total variance and 1,032 is the amount of variance not explained by the independent variable (degree-days), by subtraction, 51,508 must be the amount of variance that *is accounted for* by X.

There is one last piece remaining. Refer to the analysis of variance section of Exhibit 17.7. Note that the sum of the squared regression is 51,508, the sum of the squared residual is 1,032, and the sum of the squared total is 52,541. POLYSTAT will also calculate the mean for these values (MS). These means (MS) are used to calculate the F-statistic, which, in this case, is 51,508 ÷ 103 = 499.

The F-statistic is calculated as:

$$F\text{-statistic} = \frac{\text{Explained variance} \div (k-1)}{\text{Unexplained variance} \div (n-k)}. \tag{17.20}$$

If, at this point, you think it sounds as though these formulas are somehow related, then you are correct they are related. The alternative formula to calculate the F-statistic using the coefficient of determination (r^2) is

$$F\text{-statistic} = \frac{r^2 \div (k-1)}{\left(1-r^2\right) \div (n-k)}. \tag{17.21}$$

As a practical matter, we only went through this exercise to provide a little appreciation of the importance of variance and how it plays a role in all of this. We do not intend to calculate any of these statistics by hand, and will let POLYSTAT do the work for us. All we need to do is interpret the results.

So, how do we interpret the results of an F-test? At the extreme, the F-statistic will take on a value of zero when the regression equation, as a whole, provides absolutely no explanation of the variance in the dependent variable. The F-test then becomes a procedure in which we have to determine if the F-statistic associated with a specific regression equation is large enough to enable us to reject the null hypothesis that the combination of all coefficients does not significantly explain the variation in the dependent variable.

The F-test is therefore a one-tail test, because we must determine if the F-statistic for the regression equation exceeds the critical F-statistic in the F-distribution. If it does, then the calculated F-statistic is in the rejection region, and we conclude that there is a relationship between the dependent and independent variables. The upper-tail critical value from the F-distribution is denoted as F_U. Using a desired level of significance, the decision rule for the F-test is to reject the null hypothesis (of no relationship) if $F > F_U$.

Although we are not going to actually perform the F-test, it is, nonetheless, useful to see how the critical value of the F-statistic is derived from the F-distribution, since it is a little different from that covered in Chapter 11. In Chapter 11, we described the F-test as a two-tail test, since we wanted to determine whether the variance between two means was *different*. The F-test for regression analysis is a one-tail test, since we want to determine whether a significant amount of the total variance is explained by the regression equation. In other words, the F-test for a regression equation implies a direction. The other difference is the way the degrees of freedom are calculated. To test whether two variances were different (in Chapter 11), the degrees of freedom were the sample size minus one in the first sample, and the sample size minus one in the second sample. For the regression analysis F-test, the degrees of freedom are $k - 1$ for the explained variance and $n - k - 1$ for the unexplained variance. Here k is the number of coefficients that are estimated (including the intercept) and n is the sample size. Keep in mind that in the case of simple regression analysis, the sample size refers to pairs of data. As a rough rule of thumb, if the calculated F-statistic exceeds ten, you can generally reject the null hypothesis that a significant amount of variance is not explained by the equation.

Example 17.10

Returning to Example 17.2, assume we want to assess the overall explanatory power (of number of hours worked on test grades) in the equation with the F-test. Using a 0.05 level of significance, can we reject the null hypothesis that the variance is not explained by the equation?

Solution. From Exhibit 17.2, it can be seen that the calculated $F = 40.49$. POLYSTAT tells us that we should reject the null hypothesis that the variance is not significantly explained by the regression equation. Why is this so? If we were to look up the critical value of the F-statistic (for $k - 1$ and $n - k - 1$ degrees at freedom), at the 0.05 level of significance, we would find the F-distribution shows it is $F_U = 5.32$. Since the calculated value of the F-statistic exceeds the critical value F_U, we reject the null hypothesis.