Using and Interpreting Linear Regression and Correlation Analyses: Some Cautions and Considerations

Connie A. Tompkins

For the special section on statistics, I was asked to discuss issues that are important to remember when implementing or interpreting correlation and regression analyses. My dilemma was deciding how to focus this paper. Many of the important analytical and interpretive issues are complex, and it should not surprise anyone that the experts disagree about the most preferable ways to handle some of these matters. I decided to approach my charge from my perspective as a reviewer and associate editor, and began to identify some of the common problems or misconceptions that I have observed. The result of that process formed the basis for this paper. In the following pages, I raise a number of considerations or cautions related to correlation and regression analyses without going into great detail about them. For those interested in pursuing these and other issues in more detail, references are provided that I have found helpful in conducting regression analyses and in preparing this article.

OVERVIEW

Before identifying the issues I have chosen, let us review briefly the purposes of linear regression and correlation analyses and some of the terminology involved. *Regression analysis* allows prediction or estimation of the value of one variable (the criterion, dependent, or predicted variable; traditionally called *Y*) from one or more predictor variables (called *X*s). Simple linear regression estimates the extent to which the predicted variable changes as a single predictor variable changes, whereas multiple regression estimates *Y* from two or more predictor variables. When someone

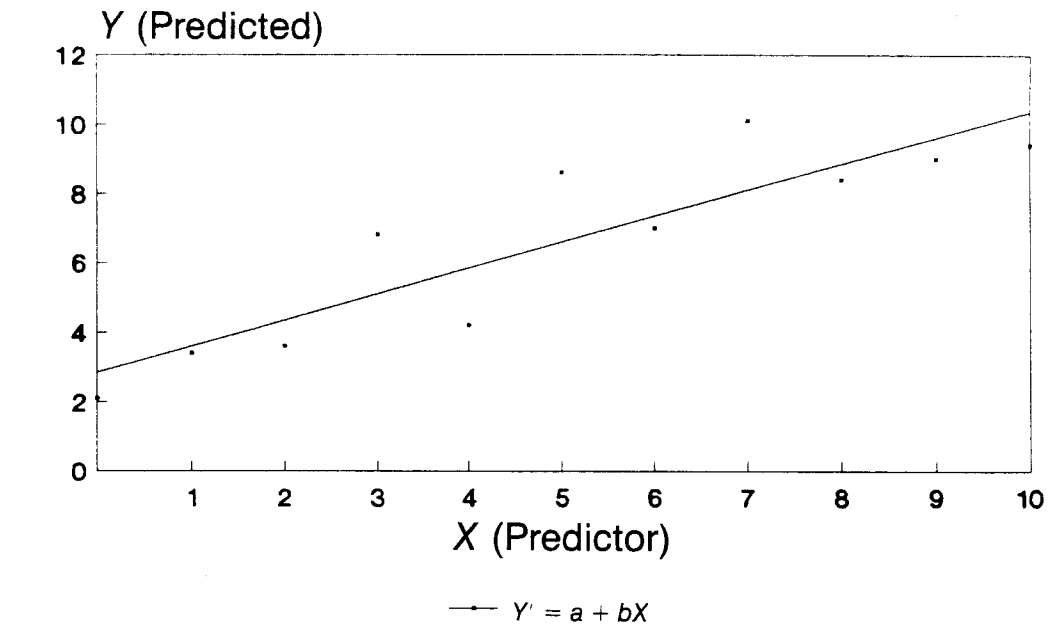

Figure 1. Hypothetical scatterplot with regression line and regression equation.

runs a regression analysis to predict some *Y* from an *X* variable, the appropriate terminology is they are "regressing *Y* on *X*."

The pairs of *X* and *Y* observations can be graphed together on a scatterplot like the one shown in Figure 1. Because the relationship between values of *X* and *Y* is almost never perfect, linear regression analysis endeavors to fit a line through the points. A simple regression equation ($Y' = a + bX$) denotes the line that best fits the plot of all pairs of (*X*, *Y*) observations. Here Y' (read as "*Y* prime") denotes the predicted value of *Y*, or the value that falls on the regression line. It is distinguished from *Y*, the actual or observed value for the criterion variable that corresponds to a particular *X* value. The differences between the predicted and observed values of the criterion variable are known as regression residuals and represent the errors of prediction.

The other components of the regression equation are the *Y*-intercept (*a*), which is the point at which the regression line crosses the *Y*-axis, and a regression coefficient (*b*), which in the bivariate case designates the slope of the line. The *Y*-intercept is simply a reference that adjusts for differences between mean scores and will not be considered further here. The regression coefficient, *b*, is more important. It indicates that for each unit of change in *X*, *Y* will change by a factor of *b*. It should be noted here that

typically, regression results are not symmetric; that is, the equation predicting Y' from X will almost always be different from the equation predicting X' from Y.

Multiple-regression analyses are also concerned with finding a regression equation that best estimates Y, in this case from some optimal combination of predictor variables. In addition, multiple-regression analysis allows the assessment of the unique contribution of each predictor to the overall regression equation. I will return to this point later.

Correlation analysis is related to simple regression, but it serves a different purpose. In many cases, an investigator may be interested in the relationship between two variables but cannot specify that one predicts the other (e.g., the association between depression and anxiety, the relationship between structural and content aspects of discourse production). A correlation coefficient reflects the strength or degree of linear association between variables or the extent to which the two variables behave alike or vary together. Correlation techniques are also essential for examining interrelationships among variables in multiple regression and for testing theoretically derived causal models (e.g., PATH analysis; Duffy, 1993).

Unlike the regression equation, correlation is symmetric in the sense that the relationship between X and Y goes both ways. Pearson's r (Achen, 1982) is probably the best known and most commonly used index of linear association.[1] The magnitude of the correlation coefficient varies between 1.0 for a strong positive relationship and -1.0 for a strong inverse relationship; when there is little association between variables, r is close to 0. The squared coefficient (r^2, "coefficient of determination") represents the proportion of variation that X and Y have in common and is important as an indicator of effect size. It conveys the practical significance or meaningfulness of the relationship rather than its statistical significance. The variation unaccounted for (difference between r^2 and 1.0) is "error"; it represents variation due to all other possible variables. The multiple-correlation coefficient is designated by R, and R^2 is conventionally interpreted as an indicator of the proportion of variation in the dependent measure that is explained by all of the predictor variables included in the equation.[2]

1. Special kinds of correlation coefficients, such as the point-biserial and Spearman rho, are simplifications of the Pearson product-moment correlation.

2. Achen (1982) objects to this interpretation of R^2, indicating that R^2 is merely descriptive of the shape of the "point clouds" in the regression plane. He recommends the standard error of regression as a better reflection of the goodness-of-fit of any regression equation. It would undoubtedly be useful for regression users to report both of these summary statistics.

THE IMPACT OF COMPUTERIZED ANALYSIS PROGRAMS

Regression and correlation analyses are accessible and easy to run thanks to the proliferation of computerized statistical packages (see Afifi & Clark, 1984; Dielman, 1991; Norusis, 1985; Pedhazur, 1982; Schroeder, Sjoquist & Stephan, 1986; Younger, 1979, for information about various computer programs and help in deciphering their output). However, the mechanical application of computerized programs can lead to misunderstanding or misinterpretation. It is best to start with a theoretical rationale for examining certain variables and for excluding others from consideration. Providing a variety of interpretational and statistical arguments, Cohen (1990) advances a "less-is-more principle" and emphasizes that working with a few well-motivated predictor variables is far preferable to including an abundance of potential predictors.

Users of computerized data analysis packages should be aware of the defaults built into the analysis programs (e.g., one- vs. two-tailed tests; choice of regression model; treatment of missing data) and of their potential impact on the results. An investigator should also plot and scrutinize the data before allowing the computer programs to crunch the numbers. Most computer analysis programs make visual inspection relatively easy because they will plot scattergrams or calculate and plot regression residuals for the user. The importance of inspecting data plots and several other considerations and cautions related to correlation and regression analyses are highlighted in the sections that follow.

SOME CONSIDERATIONS AND CAUTIONS IN CORRELATION ANALYSIS

1. Consider sample sizes carefully. Correlation coefficients based on samples of less than 25–30 tend to be unstable (Dielman, 1991; Younger, 1979). With small samples there can be dramatic swings in the magnitude of correlation from sample to sample.

2. Be aware of other factors that influence the magnitude and replicability of r (Cohen & Cohen, 1983; Younger, 1979). For example, data samples that are restricted in range reduce the magnitude of the correlation coefficient. To the extent that *measurements are unreliable,* the expected value of the correlation coefficient will also be reduced. Dissimilar frequency distributions for the X and Y variable can restrict the minimum or maximum obtainable value of a correlation coefficient. Additionally, outliers can bias the correlation coefficient in either direction. The sample

specificity of correlation coefficients is exacerbated when there are outlying points in the data set.

Outliers are extreme observations that can be detected from plots of standardized residuals, as discussed further in the section on regression. To minimize the influence of extreme points, try to ensure a relatively evenly spaced data set for the X variable; however, if outliers are detected the possible reasons should be investigated. Outlying observations must never be discarded simply because they are inconvenient. If the outlying point cannot be traced to some technical problem, it may provide a useful source of hypotheses about other, unmeasured factors that affect the observed relationship. The results of two correlational analyses, one with and one without the outlying point, can also be examined and discussed to get a feeling for the impact of the outlying observation.

3. Cautious interpretation is required when there are other irregularities in the relationship between X and Y (Cohen & Cohen, 1983; Younger, 1979). For example, when a relationship is nonlinear, Pearson's r (an index of linear association) will underestimate the degree of relationship between variables. Data transformation or nonlinear regression is possible, but interpretation becomes more difficult. As another example, when the scatter of points around the prediction line differs substantially for different values of X (heteroscedasticity) there may be several consequences. Since the observed association holds better at some parts of the measured range of X than at other parts, an investigator may have to qualify the results. There are also statistical consequences that I will not discuss here (Pedhazur, 1982). These irregularities can also be detected through inspection of standardized residual plots.

4. Remember that correlation can be used descriptively to index the linear association between variables measured in a single sample. Typically, investigators want to extend beyond the sample studied to draw inferences about relationships for a population of X and Y values. When correlation is used inferentially, additional assumptions obtain (e.g., each observation is independent; the population values of both X and Y are assumed to follow a bivariate normal distribution) (Pedhazur, 1982).

5. Consider the meaningfulness, or effect size, of a correlation coefficient along with, or instead of, its statistical significance. Statistical significance is closely related to sample size, and a very small r will be statistically significant if the sample is large enough. Cohen and Cohen (1983) suggest that in correlational, behavioral-science research, $r = .50$ meets the conventional definition of a large effect. However, establishing acceptable criteria for the meaningfulness of any given finding should rely more on the research questions asked and the state of knowledge and theory in the field than on convention.

6. Limit multiple tests on the same sample (fishing expeditions) that can increase the probability of attaining some significant correlations simply

by chance. To avoid this problem, examiners should be selective about the variables to be correlated, use conservative criteria for determining statistical significance, and/or specify a minimum, desired-effect size.

7. Predict or generalize only over the relevant range tested. The nature of a relationship might change for values that have not been measured.

8. Generally, be cautious about attributing causality from bivariate measures of association. The well-known adage that correlation does not imply causation refers to the fact that when two variables are correlated, either could cause the other (e.g., the association between depression and anxiety). Alternatively, both may be caused by unknown or unmeasured factors (e.g., physical illness and/or gender influences). Causal reasoning definitely can be risky from correlational analysis. However multiple regression/correlation techniques are increasingly being used to test theoretically derived causal models.

SOME CONSIDERATIONS AND CAUTIONS IN SIMPLE REGRESSION AND MULTIPLE REGRESSION/CORRELATION

1. Inspect the scatterplots of the raw data for bivariate outliers and for coding errors (e.g., points off by a factor of 10 because of decimals).

2. Inspect the intercorrelation matrices to assess the strength of individual correlations and to examine for multicollinearity (strong intercorrelations) among the predictor variables for multiple-regression analysis. Multicollinearity creates instability of the regression coefficients (*b*s). In addition it may lead to misconceptions about the unique contribution of highly related, predictor variables to the estimation of the predicted variable. This is because regression coefficients are *partial correlation coefficients*. Partial coefficients reflect the prediction of Y from X after controlling for the effects of all other predictor variables in the model. That is, once X_1, X_2, X_3, and so forth, have explained all they can about Y, the regression coefficient for X_n reflects the proportion of leftover variation in Y that is explained by X_n. If X_1 and X_n are highly interrelated and put into the equation simultaneously, they will claim a great deal of the same proportion of Y variance and neither will be found to make much of a unique contribution.

This interpretational problem is particularly difficult in certain stagewise variable-entry methods. Given two highly interrelated predictors (X_1 and X_2), you will enter the equation first simply because it is more strongly correlated with the predicted variable. If X_1 enters first, it will

use up much of the potential variation that could be explained by X_2, making X_2 appear unimportant. Since bivariate correlations can fluctuate greatly from sample to sample, X_2 might have a stronger bivariate correlation with the predicted variable than X_1 does in a different sample. In the second sample, X_2 would appear to be more important than X_1. Assessing and acknowledging the possible effects of multicollinearity and remembering that multiple-regression coefficients are partials can help keep an investigator from falling into interpretive traps.

An arbitrary rule of thumb is multicollinearity may be a problem if any pairwise correlation is greater than $r = .50$ (Dielman, 1991); however, multicollinearity may not show up in bivariate correlations because one predictor may be highly related to a combination of variables rather than a single variable. Berry and Feldman (1985) and Lewis-Beck (1980) suggest regressing each predictor on all other predictors to assess multicollinearity. If the R^2 values are close to 1.0, multicollinearity is present. The tolerance value provided on computer printouts gives an estimate of the extent to which a single predictor is redundant of all others that are in the model (Norusis, 1985). A tolerance of 0 indicates that the predictor is a perfect, linear combination of others in the model, while a tolerance of 1 indicates that the predictor is independent of others already in the model. Multicollinearity may be dealt with by excluding a predictor that is highly correlated with others of greater theoretical interest by creating composite indexes of variables (if they are conceptually logical and internally consistent) or by using specialized techniques such as ridge regression (Darlington, 1978; Price, 1977).

3. For multiple regression determine what method was used or is most appropriate for variable entry. There are three primary variable entry methods: (a) an overall method in which all predictors are entered simultaneously; (b) stagewise methods that allow selection of a best subset of predictor variables (e.g., stepwise, forward selection); and (c) hierarchical methods in which the order of variable entry is determined by the user to test theories or to control for known influences on the criterion measure. As indicated, the variable-entry method can interact with other factors so an investigator should provide a clear rationale for the method selected and should consider the impact of the choice on the interpretation of results.

Stagewise methods are most commonly reported in our literature. The choice of a particular stagewise procedure should be made by considering the benefits and liabilities of each approach (Younger, 1979) rather than by allowing the computer default to determine the selection. Hybrid procedures may also be appropriate. As an example, Schulz, Tompkins and Rau (1988) combined hierarchical and stagewise methods in their longitudinal analysis of caregiver depression poststroke, forcing variables known to influence depression level (e.g., age, income, health, prior depression

level) in the first group and then selecting stepwise entry for other predictors of interest.

4. Be cautious about interpreting the magnitudes of the regression coefficients as indicators of the relative importance of variables in the equation. As noted earlier, they are partials, reflecting the variation accounted for when all other predictors are in the model, and they may have different measurement units and variances. If they are standardized (beta weights) they become more comparable; however, it is difficult to draw comparisons across samples because sample standard deviations are used in the calculation of beta weights, making them highly sample specific, and they remain partials, affected by the correlations among the predictors. Therefore unstandardized regression coefficients (*b*s) cannot be considered to reflect the relative importance of individual variables, and the interpretation of standardized beta weights is extremely limited by their context-dependent nature (Achen, 1982; Cohen & Cohen, 1983; Lewis-Beck, 1980; Norusis, 1985).

5. Recognize the data dependence of regression results and the value of cross validation. Different samples give different solutions and the measures of variance accounted for (r^2 and R^2) tend to overstate an equation's predictive ability. R^2 adjusted, reported on computer printouts, provides an estimate of cross validation that adjusts more for small ratios of sample size to predictors (Dielman, 1991; Pedhazur, 1982; & Younger, 1979). Of course if the results are used only to describe the sample from which they were gathered, their data dependence is irrelevant and the adjusted R^2 value can be ignored.

6. Consider sample size and the ratio of predictor variables to sample size when planning and interpreting analyses. Several rules of thumb have been suggested for multiple regression: Younger (1979) recommends a minimum N of 10 for each predictor, Dielman (1991) suggests an N of 30 for simple regression with 10 to 20 additional observations for each additional predictor variable in multiple regression, and Pedhazur (1982) advises a minimum of 30 subjects per predictor variable. Sample sizes of these magnitudes should be adequate to stabilize the regression coefficients; however, the probability of detecting a significant increment in variance accounted for by a single variable may require much larger samples (Cohen & Cohen, 1983). Again the meaningfulness (effect size) of any variance increment should be kept in mind.

Whenever the number of predictor variables approaches the sample size, overfitting and the consequent overestimation of R^2 becomes a critical issue. Adding predictor variables cannot lower R^2; in fact almost any additional predictor variable will increase R^2 to a certain degree regardless of its relevance to explaining the predicted variable. In the extreme it has been demonstrated that when the number of predictor variables is one less than the total sample size, a perfect sample R^2 of 1.0 will be achieved even when the population R^2 is 0 (Cohen & Cohen, 1983; Lewis-Beck,

1980; Pedhazur, 1982).[3] The dangers of overfitting are particularly acute in small N studies like those typical of our field of study.

7. Assess the assumptions of the regression model. To allow accurate inferences from sample data, linear-regression models make certain assumptions about the characteristics of the errors in the population: They are assumed to be normally distributed with a known mean and equal variances, and errors for one observation are assumed to be independent of those for other observations.[4] To infer whether population assumptions are being met, you can examine a plot of the sample residuals; it should be roughly rectangular with the points scattered about the line that is the mean of the residuals. Residual plots can also help in detecting nonlinear patterns and outliers. Younger (1979) asserts that a standardized residual of four or more standard deviations from the mean indicates an outlying data point, while Dielman (1991) identifies three standard deviations as signifying an outlier.

As indicated previously most computer programs calculate and plot residuals. The SPSS-X manual (Norusis, 1985) provides relatively good information about the reasons for examining various types of plots and about the interpretation of the plots. Other helpful illustrations and interpretations of residual analysis are provided by Berry and Feldman (1985), Cohen and Cohen (1983), and Edwards (1985).

The regression assumptions describe an ideal situation for testing hypotheses against distributions with known characteristics. In practice these assumptions are rarely met in behavioral science research, but if you are aware of the possible influences of failure to meet the assumptions, you can take corrective action and/or qualify the results as needed. Investigators should document how they checked assumptions, what they did about violations, and indicate which assumptions don't matter because of the robustness of regression analysis and related statistical procedures (e.g., Cohen & Cohen, 1983). Pedhazur (1982) argues that regression is robust except for problems with measurement error and specification error.

8. Be aware that error due to unreliable measurement and specification error can have important consequences. Recall that in simple regression, measurement error in the predictor variable results in underestimation of the regression coefficient. In multiple regression the effects of measurement error are more complex; it is possible for regression coefficients to be substantially biased either upward or downward.

---

3. Mathematically the expectation of a sample R^2 is $K/(N - 1)$, where K equals the number of predictors in the equation and N is the sample size regardless of what the predictor variables are.

4. Comprehensive information about the nature of regression assumptions, consequences of violating them, tests for violations, and possible solutions is provided by Berry and Feldman (1985), but see also Cohen and Cohen (1983), Dielman (1991), Edwards (1985) and Lewis-Beck (1980).

Specification error primarily involves the omission of relevant predictors from the regression model or the inclusion of irrelevant predictors. The various consequences of specification error are described in detail by Pedhazur (1982) and Berry and Feldman (1985). One of the most important consequences is that some unspecified or unmeasured variable may account for any observed relation. Specification error can be avoided if an investigator has a sufficiently well-developed theory about which variables should be in the equation and a set of indicators to measure those variables.

9. Consider carefully whether any relationship of primary interest may be moderated, suppressed, or inflated by other variables. Partial correlation methods are useful for examining the nature and influence of such variables.

10. Remember that ANOVA and ANCOVA are special cases of multiple regression (Cohen & Cohen, 1983; Edwards, 1985; Pedhazur, 1982). Cohen and Cohen (1983) and Pedhazur (1982) emphasize the generality and flexibility of multiple regression/correlation methods, indicating that they can be used when predictive variables are either experimental or nonexperimental, quantitative or qualitative, correlated or uncorrelated. One common misconception is that interactions cannot be examined in regression analyses; however, it is possible to examine interactions (joint relations) by computing variables and then using them in a hierarchical model (Aiken & West, 1991; Cohen & Cohen, 1983; Dielman, 1991; Pedhazur, 1982). There are also techniques for determining regions of significance (the specific values of a predictor variable for which the interaction holds) when interaction terms contribute substantially to predicting variation in the criterion measure.

There is a tendency to select an analytic technique based on comfort or familiarity with that technique, so ANOVA is often used in situations where regression would be more informative and more powerful (Cohen & Cohen, 1983; Pedhazur, 1982). For example, when logically continuous or quantitative variables are split at the median in order to create categorical variables for ANOVA, potentially important information is lost. Another undesirable consequence of categorizing logically continuous variables is that two scores that are numerically close together but on different sides of the median are treated as dissimilar, while two scores that are farther apart but on the same side of the median are considered similar.

To illustrate the advantage of regression methods over ANOVA when a predictor variable is logically continuous, consider auditory comprehension level as a variable of interest in an aphasia-treatment study. When ANOVA is used to analyze the results, the comprehension scores are forced into categories, typically designated as high- versus low-comprehension groups. In this case the investigator would lose possibly valuable information about the range of comprehension over which the treatment was most

effective. When variables are inherently categorical (e.g., treatment group), they can be coded for entry into multiple-regression analyses as well. If the nature of the questions and the characteristics of the variables allow, regression analyses may yield more information than the ANOVA procedures.

CLOSING COMMENTS

This list of cautions and considerations provides a starting point for the investigator or consumer interested in designing and interpreting regression and correlation analyses. In closing, I want to emphasize once more the importance of theory and prior empirical evidence in all analytic endeavors. Regression techniques, like any statistical methods, are simply tools; their ultimate value depends on informed application to meaningful questions and on the strength of interpretations constructed by knowledgeable users.

ACKNOWLEDGMENTS

Preparation of this manuscript was supported by grant #DC00453 from the National Institute on Deafness and Other Communication Disorders. Carol Baker, Hiram Brownell, and Richard Schulz provided valuable comments on earlier versions of this chapter.

REFERENCES

Achen, C. H. (1982). *Interpreting and using regression.* Beverly Hills: Sage.

Afifi, A. A., & Clark, V. (1984). *Computer-aided multivariate analysis.* London: Wadsworth.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice.* Beverly Hills: Sage.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Darlington, R. B. (1978). Reduced-variance regression. *Psychological Bulletin, 85,* 1328–1255.

Dielman, T. E. (1991). *Applied regression analysis for business and economics.* Boston: PWS-Kent.

Duffy, J. R. (1993). Path Analysis. In M. Lemme (Ed.), Clinical Aphasiology (Vol. 21, pp. 47-57). Austin, TX: PRO-ED.

Edwards, A. L. (1985). *Multiple regression and the analysis of variance and covariance* (2nd ed.). New York: W. H. Freeman.

Lewis-Beck, M. S. (1980). *Applied regression: An Introduction.* Beverly Hills: Sage.

Norusis, M. J. (1985). *SPSSX advanced statistics guide.* New York: McGraw-Hill.

Ostrom, C. J., Jr. (1978). *Time series analysis: Regression techniques.* Beverly Hills: Sage.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.

Price, B. (1977). Ridge regression: Application to nonexperimental data. *Psychological Bulletin, 84,* 759-766.

Schroeder, L. D., Sjoquist, D. L., & Stephan, P. E. (1986). *Understanding regression analysis: An introductory guide.* Beverly Hills: Sage.

Schulz, R., Tompkins, C. A., & Rau, M. T. (1988). A longitudinal study of the psychosocial impact of stroke on primary support persons. *Psychology and Aging, 3,* 131-141.

Younger, M. S. (1979). *A handbook for linear regression.* North Scituate, MA: Duxbury.