4-2009

Applications of Multiple Regression in Psychological Research

Razia Azen

David Budescu

PART IV

Data Analysis

13

Applications of Multiple Regression in Psychological Research

Razia Azen and David Budescu

THE REGRESSION MODEL

History and introduction

The regression model was conceptualized in the late nineteenth century by Sir Francis Galton, who was studying how characteristics are inherited from one generation to the next (e.g., Stanton, 2001; Stigler, 1997). Galton's goal was to model and predict the characteristics of offspring based on the characteristics of their parents. The term 'regression' came from the observation that extreme values (or outliers) in one generation produced offspring that were closer to the mean in the next generation; hence, 'regression to the mean' occurred (the original terminology used was regression to 'mediocrity'). Galton also recognized that previous generations (older than the parents) could influence the characteristics of the offspring as well, and this led him to conceptualize the multiple-regression model. His colleague, Karl Pearson, formalized the mathematics of regression models (e.g., Stanton, 2001).

The multiple-regression (MR) model involves one criterion (also referred to as response, predicted, outcome or dependent) variable, Y, and p predictor (also referred to as independent)[1] variables, $X_1, X_2, \ldots, X_p$. The MR model expresses Y_i, the observed value of the criterion for the ith case, as a linear composite of the predictors and a residual term:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i \quad (1)$$

Here, $X_{1i}, X_{2i}, \ldots, X_{pi}$ are the values observed on the p predictors for the ith case, and the various βs ($\beta_0, \beta_1, \beta_2, \ldots, \beta_p$) are the (unknown) regression coefficients associated with the various predictors. The first coefficient, β_0, is an intercept term (or a coefficient associated with a predictor that takes on the value $X_0 = 1$ for all observations).

If all the variables (response and predictors) are standardized to have zero mean and unit

variance the model can be re-expressed as:

$$Z_y = \sum_{i=1}^{p} \beta_i^* Z_{x_i} + e \tag{2}$$

where Z refers to a standardized variable. For obvious reasons, the β_i^* are referred to as standardized coefficients (by definition, $\beta_0^* = 0$). It is easy to show that the standardized coefficients can be obtained by multiplying their raw counterparts by the ratio of the standard deviations of the respective predictor and the response:

$$\beta_i^* = \beta_i \frac{S_{x_i}}{S_y} \tag{3}$$

This definition is not universally accepted. Some statisticians (e.g., Neter et al., 1996) prefer to define standardized coefficients as the values obtained by fitting the models after applying the, so-called, correlation transformation.[2] Although the values of the coefficients are the same (the numerator and denominator are divided by the same constant), their standard errors are not! Furthermore, Bring (1994) challenged these (closely related) definitions and suggested that a more appropriate way of calculating the standardized coefficients should use the *partial* standard deviations of the predictors.

The predicted value of Y_i, which represents the best guess (expected value) of the criterion given the observed combination of the p predictor values for the ith case, is written as:

$$\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \tag{4}$$

The residual, ε_i, is the difference between the observed and predicted values of Y associated with the ith case:

$$\varepsilon_i = Y_i - \hat{Y}_i. \tag{5}$$

The ideal situation of perfect deterministic prediction implies $\varepsilon_i = 0$ for all cases. Otherwise, the residuals are assumed to be random variables with 0 mean and unknown variance (which is estimated in the process of model fitting). The residuals provide a measure of the goodness or accuracy of the model's predictions, as smaller (larger) residuals indicate more accurate (inaccurate) predictions. The residuals, ε_i, are sometimes labeled 'errors', but we find this terminology potentially misleading since it connotes that (some of) the variables are subject to measurement error. In fact, the statistical model assumes implicitly that all measurements are perfectly reliable, and the residuals are due to sampling variance, reflecting the fact that the relationships between Y and the various Xs are probabilistic in nature (e.g., in Galton's studies, not all boys born to fathers who are 180 cm tall, and mothers who are 164 cm tall, have the same height). The more complex structural equation models (SEM) combine the statistical MR model with measurement models that incorporate the imperfection of the measurement procedures for all the variables involved. These models are beyond the scope of this chapter, but are covered in Part V (e.g., Chapter 21) of this book.

This chapter covers the wide variety of MR applications in behavioral research. Specifically, we will discuss the measurement, sampling and statistical assumptions of the model, estimation of its parameters, interrelation of its various results, and evaluation of its fit. We illustrate the key results with several numerical examples.

Applications of the multiple-regression model

The regression model can be used in one of two general ways, referred to by some (e.g., Pedhazur, 1997) as *explanation* and *prediction*. The distinction between these approaches is akin to the distinction between *confirmatory* and *exploratory* analyses.

The explanatory/confirmatory use of the model seeks to confirm (or refute) certain theoretical expectations and predictions derived from a particular model (typically developed independently of the data at hand). The ultimate goal is to understand the specific process by which the criterion of interest is produced by the (theoretically determined)

predictors. The explanatory regression model is used to confirm the predictions of the theory in the context of a properly formulated model, by using standard statistical tools. It can verify that those predictors that are specified by theory are, indeed, significant and others, which are considered irrelevant by the theory, are not. Similarly, it could test whether the predictor that is postulated by the theory to be the most important in predicting Y reproduces by itself the highest amount of variance, produces the most accurate predictions, and so forth.

Consider, for example, specific theories related to the process and variables that determine an individual's IQ. Some theories may stress the hereditary nature of IQ and others may highlight the environmental components. Thus, they would have different predictions about the significance and/or relative importance of various predictors in the model. An explanatory application of the model would estimate the model's parameters (e.g., regression coefficients), proceed to test the significance of the relevant predictors, and/or compare the quality and accuracy of the predictions of the various theories.

The predictive/exploratory analysis can be also guided, at least in part, by theory but it is more open ended and flexible and, in particular, relies on the data to direct the analysis. Such an analysis seeks to identify the set of predictors that predicts best the outcome, regardless of whether the model is the 'correct' explanatory mechanism by which the outcome is produced. Of course, it is reassuring if the model makes sense theoretically, but this is not a must. For example, to predict IQ one would start with a large set of predictors that are theoretically viable and include the predictors that optimally predict the observed values of IQ in the regression model. Thus, the decision in this case is data driven and more exploratory in nature than the explanatory approach. While the model selected using a prediction approach should yield highly accurate predictions, the components of the model are not considered to be any more 'correct' than those for other potential predictors that would yield the same prediction accuracy.

To borrow an example from Pedhazur (1997), if one is trying to predict the weather, a purely predictive approach is concerned with the accuracy of the prediction regardless of whether the predictors are the true scientific 'causes' of the observed-weather conditions. The explanatory approach, on the other hand, would require a model that can also provide a scientific explanation of how the predictors produce the observed-weather conditions. Therefore, while the predictive approach may provide accurate predictions of the outcome, the explanatory approach also provides true knowledge about the processes that underlie and actually produce the outcome.

THE VARIOUS FORMS OF THE MODEL

Measurement level

Typically, all the variables (the response and predictors) are assumed to be quantitative in nature; that is, measured on scales that have well defined and meaningful units (interval or higher). This assumption justifies the typical interpretation of the regression coefficients as conditional *slopes* – the expected change in the value of the response variable per *unit change* of the target predictor, conditional on holding the values of the other $(p - 1)$ predictors fixed.

This assumption is also critical for one of the key properties of the MR model. MR is a compensatory model in the sense that the same response value can be predicted by multiple combinations of the predictors. In particular, high values on some predictors can compensate for low values on others. The implied trade-off between predictors is captured by their regression coefficients. Imagine, for example, that we can predict freshmen Grade Point Average (GPA) in a certain college from their two Scholastic Assessment Test (SAT) scores (quantitative and verbal), according to the following equation (in standardized scores): Predicted GPA $=$ 1.0*SAT-Q $+$ 0.5*SAT-V. One can make up for low SAT-Q (or SAT-V) scores by appropriately higher SAT-V (or SAT-Q)

Table 13.1 Two possible ways to code a categorical predictor with C = 4 categories

Type of school	*'Dummy' coding*			*'Effect' coding*		
	D_{11}	D_{12}	D_{13}	D_{21}	D_{22}	D_{23}
Public school	1	0	0	−1	−1	−1
Private school	0	1	0	−1	−1	1
Parochial school	0	0	1	−1	1	−1
Other schools	0	0	0	1	−1	−1

scores. More precisely, one can compensate for a unit disadvantage in SAT-V by a ½-point increase in SAT-Q, and a one unit disadvantage in SAT-Q can be offset by a 2-point increase in SAT-V.

Despite the measurement level restriction of the model, it is possible to include lower-level (e.g., categorical) predictors in a MR equation through a series of appropriate transformations. Imagine that we wanted to consider 'type of high school attended' as a potential predictor of freshmen GPA, where $C = 4$ mutually exclusive and exhaustive categories make up the school types as shown in Table 13.1. We can define $(C - 1) = 3$ binary[3] variables to fully represent the school type classification. As the variables take on only two values, they are consistent with the measurement-level constraint of the MR (recall that an interval scale has two free parameters – its origin and its unit). The choice of values for these variables is arbitrary (all possible assignments are linearly related), and while it is convenient to use the values 0,1 or the values −1,1, the only technical constraint is that the $(C - 1)$ variables be linearly independent [see, for example, Chapter 8 in Cohen et al. (2003) for a good discussion of coding schemes for categorical variables].

The columns in Table 13.1 represent two alternative coding schemes (the first labeled 'Dummy coding' and the second labeled 'Effect coding'). Note that both schemes distinguish between the various types of schools (each school has a unique pattern of $(C - 1)$ values). Although the two sets of variables are different (the dummy coding set uses the 'other' schools as the baseline, and the effect coding set compares all schools to the 'public' benchmark), and would yield different regression coefficients, their joint effect is identical! In fact, the test of the hypothesis that the 'type of school' is a useful predictor of freshmen GPA would be invariant across all possible definitions of the $(C - 1)$ linearly independent variables.

Derivative predictors

Typically, the p predictors are measured independently from each other by distinct instruments (various scales, different tests, multiple raters, etc.). In some cases, researchers supplement these predictors by additional measures that are derived by specific transformations and/or combinations of the measured variables. For example, polynomial-regression models include successive powers (quadratic, cubic, quartic, etc.) of some of the original predictors, and are designed to capture non-linear[4] trends in the relationship between the response and the relevant predictors. Interactive-regression models include variables that are derived by multiplying two, or more, of the measured variables (or some simple transformations of these variables) in an attempt to capture the joint (i.e., above and beyond the additive) effects of the target variables. Computationally, these models do not require any special treatment as one can treat the product X_1X_2 or the quadratic term X_1^2 as additional variables, but we mention several subtle interpretational issues.

The first issue is, simply, that the standard interpretation of regression coefficients as conditional slopes breaks down in these models. Clearly, in a polynomial model one cannot increase X by one unit while keeping X^2, X^3, etc., fixed, nor vice-versa. Similarly, in an interactive model that includes product

terms, say involving variables X_i and X_j, their respective coefficients will no longer reflect the 'net' effects of these variables, as these effects depend on, and cannot be separated from, the values of the other variables. For example, in the simplest interaction model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$, when X_1 is increased by one unit, Y increases by $(\beta_1 + \beta_{12} X_2)$! Of course, this applies to the coefficient of the product as well (i.e., one cannot change the product $X_i X_j$ by a unit while keeping X_i and X_j fixed!).

The second issue is more technical. Successive powers of a variable (X_1, X_1^2, X_1^3) and products involving this variable ($X_1 X_2$, $X_1 X_3$), are highly correlated (e.g., Bradley and Srivastava, 1979; Budescu, 1980). As such, they suffer from many of the problems associated with collinearity in that, conceptually, highly correlated predictors indicate redundancy in the model and, statistically, highly correlated predictors affect estimation accuracy. Fortunately, it is possible to reduce some of these problems by properly re-scaling all the variables. Simply centering the variables, by subtracting their respective means, reduces correlations between these predictors considerably and facilitates interpretation. The intuition is quite simple – if the values are all non-negative (say X_1 = income and X_2 = years of education), all correlations between the original variables and their powers or their products are going to be extremely high. By subtracting the means, about half of the values of X_1 and X_2 (but not of their squares) become negative, so the correlations drop substantially. Centering does not affect the overall fit of the model, but facilitates interpretation. Thus, we strongly recommend that one always center (or, without any loss of generality, standardize) all variables in polynomial and/or interactive models.

The last point relates to the distinction we made earlier between confirmatory and exploratory designs. In exploratory work all predictors are treated as 'exchangeable' (or 'symmetric') in the sense that there is no prior ordering among them, while in confirmatory work the theory induces a particular hierarchical structure. Polynomial and interactive models are, essentially, always hierarchical. Parsimony dictates that the basic form of the predictors be part of the model before considering higher order and/or interactive effects. In fact, when the predictors are continuous, interpretation of such higher order and/or interactive terms when their lower-order (or component) variables are excluded from the model is meaningless and, potentially, misleading. When a qualitative (or discrete) variable is used as a predictor (e.g., gender), its interaction with a continuous predictor is used to model the slopes (of the continuous predictor) separately for each level of the qualitative variable. In such models, the interaction term may be interpretable without necessarily including the qualitative predictor (that captures differences in the intercepts) by itself, though we would contend that in most research problems it is more meaningful to retain lower-level terms in the model when their interaction is included. Finally, there is no consensus in the literature on the question of precedence for single-variable polynomial terms and interactive terms (for details, see Aiken and West, 1991; Cortina, 1993; Ganzach, 1997; Lubinski and Humphreys, 1990), but we tend to favor the view espoused in the discussion by Ganzach (1997) that suggests that quadratic terms should precede interactive terms.

Sampling designs

The classical regression model assumes that the values of the predictors are 'fixed'; that is, chosen by various design considerations (including, possibly, plain convenience) rather than sampled randomly. Random samples (of equal, or unequal, size) of the response variable are then obtained for each relevant combination of the *p* predictors. Thus, the data consist of a (possibly large) collection of distributions of Y, conditional on particular predetermined combinations of $X_1, \ldots, X_p$. The X's (predictors) are not random variables, and their distributions in the sample are not necessarily expected to match their underlying distributions in the population. Thus, no distributional assumptions are made

about the predictors. The unknown parameters to be estimated in the model are the p regression coefficients and the variance of the residuals.

Alternatively, in the 'random' design the researcher randomly samples cases from the population of interest, where a 'case' consists of a vector of all p predictors and the response variable. Thus, we observe a joint distribution of the $(p + 1)$ random variables, and it makes sense to make assumptions about its nature (typically, that it is normal). The unknown parameters to be estimated are the $(p + 1)(p + 2)/2$ entries in the variance-covariance matrix of the variables (which determine the regression coefficients), and the variance of the residuals.[5]

Consider again the hypothetical freshmen GPA prediction problem described earlier. The researcher could approach this problem by randomly selecting a fixed number (say $n = 30$) of men and of women for each of 16 predetermined combinations of SAT-V and SAT-Q; say, all combinations of SAT-V and of SAT-Q from 450 to 750 in increments of 100 points [(450,450); (450,550); . . .; (750,750)] and record their freshmen GPA, or simply take one random sample of about 1000 students and record their gender, SAT scores, and GPA. The former is a fixed design and the latter is a random one. In both cases one would have the same variables and, subject to minor subtle differences, be able to address the same questions (e.g., Sampson, 1974). In general, the results of a fixed design can be generalized only to the values of X included in the study while the results of a random design can be generalized to the entire population of X values that is represented (by a random sample of values) in the study.

In both fixed and random designs it is customary to assume that all observations in the sample(s) are mutually independent. There are two noticeable exceptions to this assumption. In multistage-sampling designs observations at the lower levels that are nested within the same higher order clusters are, typically, positively correlated with each other reflecting geographical, social-economic proximity and other sources of commonality. Consider a national sample of 13-year-old students, such as in the National Assessment of Educational Progress (NAEP), that is obtained by randomly sampling: (1) school districts in various geographical regions; (2) schools within districts; (3) classrooms within schools; and finally (4) students within classrooms. The results of the students selected for testing cannot be treated as statistically independent since some of these students share many characteristics (definitely more than they share with other students in classrooms in other schools and districts in the national sample). Hierarchical-linear models (HLM) (e.g., Bryk and Raudenbush, 1992), which are the topic of Chapter 15, are extensions of the standard MR model that can handle these dependencies efficiently.

Another extension of MR, common in business, economics and many natural sciences, involves applications in which key observations constitute time series (e.g., the daily closing value of a stock over one year, or the amount of annual precipitation at a particular location over the last 200 years), and the residuals of the various observations are serially correlated (or autocorrelated). Time series are relatively rare, but not unheard of, in behavioral sciences (e.g., sequences of interactions within dyads of participants such as spouses, or players involved in a series of Prisoner's Dilemma games; Budescu 1985). Analysis of time series is beyond the scope of our chapter, but it is discussed in Chapter 26 of this book, and the classical book by Box and Jenkins (1976) is a good primary source for this topic, with special emphasis on the variety of time-dependent processes.

STATISTICAL ASSUMPTIONS AND ESTIMATION

The parameters of the MR model can be estimated either by least-squares (LS) or maximum-likelihood (ML) procedures. In this section we briefly review the key results (details can be found in such standard textbooks as Draper and Smith, 1998; Graybill, 1976; Neter et al., 1996).

Consider the standard (fixed) MR model first. Let $\mathbf{y}$ be a vector (n rows by 1 column) including the values of the response variable for all n observations in the sample, and let $\mathbf{X}$ be a matrix (n rows by $p + 1$ columns) including the values of the fixed p predictors (including a constant predictor, X_0). LS estimation requires a minimal set of assumptions:

Assumption 1

The response is a *linear* function of the predictors. Thus, the model can be re-written in matrix notation as:

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{6}$$

where $\boldsymbol{\beta}$ is a vector ($p + 1$ rows by 1 column) consisting of the (unknown) regression coefficients and $\boldsymbol{\varepsilon}$ is a vector (n rows by 1 column) including the residuals of all n observations in the sample.

Assumption 2

The residuals are *independent and identically distributed* random variables.

Assumption 3

The residuals have *0 means and equal variances*: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\Sigma(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. The assumption that all residuals have equal variance is referred to as homoskedasticity, and it implies all conditional distributions of the response (i.e., for each possible combinations of the predictors) have equal variances.

The LS method seeks estimates of the regression coefficients, $\mathbf{b}$, such that the sum of the squared residuals ($\boldsymbol{\varepsilon}$'$\boldsymbol{\varepsilon}$) is minimized. Under assumptions 1–3, it is possible to show that $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$, and the famous Gauss–Markov theorem shows that these coefficients are BLUE – Best (meaning with the smallest variance) Unbiased Linear (meaning linear function of the y's) Estimators. The other parameter, σ^2, is estimated by the mean square residual or the mean square error (MSE):

$$\text{MSE} = s^2 = \sum_i (Y_i - \hat{Y}_i)^2/(n - p - 1) \tag{7}$$

To derive ML estimates (i.e., find those parameter values that, conditional upon the distributional assumptions, are the most likely to have generated the observed data), we need one additional assumption:

Assumption 4

The residuals are normally distributed: $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$. The ML estimates of the regression coefficients are identical to the LS estimates, but σ^2 is estimated by the regular (biased) variance of the residuals in the sample, $s^2 = \sum_i (Y_i - \hat{Y}_i)^2/n$. Typically, the LS (unbiased) estimate is used.

Under the random model, Assumptions 1, 3, and 4 are replaced by Assumption 5.

Assumption 5

The response and the p predictors have a joint $(p + 1)$-variate Normal distribution with a positive semi-definite covariance matrix, $\boldsymbol{\Sigma}$.

The ML estimate of $\boldsymbol{\Sigma}$, based on the sample covariance matrix, $\mathbf{S}$, is easily obtained (e.g., Johnson and Wichern 2002; Timm, 2002). Taking advantage of the standard results for conditional multivariate normal distributions, we obtain the desired vector of estimates, $\mathbf{b}$, from $E(Y|X_1, \ldots, X_p)$. Computationally, the results are identical to the LS and ML estimates for the fixed case. Note, however, that we did not assume a linear model a-priori (Assumption 1). The linearity follows directly from the properties of the multivariate-normal distributions. This provides another interpretation of the MR model as the collection of all conditional expectations in the space defined by the p predictors.

The assumptions that residuals are normally distributed and homoskedastic are clearly unrealistic when the response variable is dichotomous (e.g., success/failure,

or below/above a certain threshold) or categorical (e.g., color of a product). Special regression models have been developed for these cases. They preserve the basic form of the MR model but employ different distributional assumptions that are more appropriate for these cases, and are consistent with their constraints. Their key feature is assuming a probabilistic model (typically Normal or Logistic) relating the predictors to the binary/categorical response. These models are dicussed in Chapter 14 of this book, and Agresti (1996) and Neter et al. (1996) are good sources for further details on these probit- and logistic-regression models.

Beyond parameter estimation

Once the regression coefficients are estimated, it is relatively simple to estimate:

- Their variances and co-covariances: $\text{Var}(\mathbf{b}) = s^2(\mathbf{X'X})^{-1}$.
- The predicted values (that have the same mean as the actual response values): $\hat{\mathbf{Y}} = \mathbf{Xb}$.
- The residuals (that have a mean of 0 and are uncorrelated with the predicted values), $\boldsymbol{\varepsilon}$, using: $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{Xb}$.

Finally, it is possible to show that the total variation of the response variable, Y, as measured by the sum of squares of the observed values (SST) around their mean, can be decomposed into two orthogonal components associated with: (1) the fitted regression model, measured by the sum of squares of the predicted values (SSR) around the mean response ($\bar{Y}$); and (2) the residuals (SSE), measured by the sum of squares of the residuals. The same decomposition holds for the degrees of freedom, and it is customary to represent these components in an analysis of variance (ANOVA) table as shown in Table 13.2.

Note (from Table 13.2) that SST and SSR are similar in nature and calculated around the same mean. We take advantage of this similarity to calculate R^2, which is the standard measure of goodness of fit of the model:

$$R^2 = \text{SSR/SST} = 1 - \text{SSE/SST} \quad (8)$$

R^2 measures what proportion of the total variance of Y is reproduced by the model. It is bounded from below by 0 (when the predictors cannot predict Y), and from above by 1 (when prediction is perfect). It is possible to show that R^2 is also the squared correlation between the observed and predicted values of Y. It is 0 when Y and $\hat{Y}$ are uncorrelated, and it is 1 when Y and $\hat{Y}$ are perfectly correlated. In the context of the fixed model, R^2 is referred to as 'the coefficient of determination', and in the random model it is called 'the squared-multiple correlation'. Interestingly, the sample R^2 is a biased estimate of its corresponding population value, ρ^2. Olkin and Pratt (1958) present an approximate unbiased estimate of ρ^2 for the multivariate-normal case with n observations and p predictors (see also Alf and Graf, 2002):

$$\hat{\rho}^2 = R^2 - \frac{p-2}{n-p-1}(1-R^2) - \frac{2(n-3)}{(n-p-1)(n-p+1)}(1-R^2)^2 \quad (9)$$

Alternatively, computer intensive procedures such as the bootstrap can also be used to estimate the population value of ρ^2

Table 13.2 ANOVA table

Source	*Sums of squares*	*Degrees of freedom*	*Mean square*
Model predictions	$\text{SSR} = \sum_i (\hat{Y}_i - \bar{Y})^2$	p	SSR/p
Residuals	$\text{SSE} = \sum_i (Y_i - \hat{Y}_i)^2$	$n-p-1$	$\text{SSE}/(n-p-1)$
Total	$\text{SST} = \sum_i (Y_i - \bar{Y})^2$	$n-1$	$\text{SST}/(n-1)$

for non-normal populations. The bootstrap requires randomly drawing n observations from the original sample, with replacement. The bootstrap sample contains n observations but it is not identical to the original sample due to the random sampling process (i.e., in the bootstrap sample some of the original observations appear more than once and some not at all). The value of R^2 can then be estimated by fitting the regression model to the bootstrap sample. This process is repeated a large number of times, resulting in a large number of estimates of R^2. These R^2 values are then averaged to obtain a bootstrap estimate of the population value. Detailed information on this method is available from sources such as Diaconis and Efron (1983), Mooney and Duval (1993), and Stine (1989) and it is also discussed in Chapter 16 of this book.

SPECIFIC APPLICATIONS

Throughout the remainder this chapter, we will use a real data set to demonstrate various concepts and methods, so we introduce it here. The data set is discussed in detail by Suh et al. (1998), who obtained normative beliefs and emotional experience as well as satisfaction with life judgments from thousands of college students in over 40 countries. Variables were mostly self-reported and included some demographic measures (e.g., sex); emotional experience, measured using positive and negative affects as well as an affect balance score; subjective global life satisfaction; domain-specific life satisfaction; and values (or norms) for life satisfaction.

Multiple regression as a confirmatory model: comparing competing (nested) models

The most common way of using MR as a confirmatory tool is to test the significance of one, several, or all of the regression coefficients that are predicted to be important (or, at least, relevant) by the theory being tested. Any test of parameters amounts to a comparison of two models, one that includes the parameters in question (and is referred to as the 'full' model), and one that excludes them (referred to as the 'reduced' or 'restricted' model). The standard tests require that the models be 'nested'; that is, that the reduced model contains a strict subset of the variables in the full model.

For example, the full model might contain five predictors, $X_1 - X_5$. Suppose we want to test the prediction (presumably derived from our theory) that X_1, X_2, and X_5 are the critical predictors and that, in their presence, X_3 and X_4 are not significant, or do not contribute to the prediction of Y. Thus, we wish to test H_0: $\beta_3 = \beta_4 = 0|X_1, X_2, X_5$ (we use the 'conditioning' notation to remind us of the other variables in the model). If this hypothesis holds, the reduced model contains only the predictors X_1, X_2, and X_5 (because it restricts the coefficients of X_3 and X_4, which were part of the full model, to be zero). The model-comparison procedure tests whether the predictors X_3 and X_4, jointly, contribute (or do not contribute) to the explanatory and predictive power of the model. On the one hand, if the full model fits the data significantly better than the reduced model, this provides evidence for the contribution of X_3 and X_4 and indicates that their inclusion is advantageous. On the other hand, if the full model does not fit the data significantly better than the reduced one, this provides evidence that X_3 and X_4 are not necessary, as predicted by the null hypothesis in this example.

Imagine that an alternative theory postulates that only X_1, X_3, and X_4 are of interest, with X_1 being a key variable. These three variables make up the 'full' model. Suppose we want to test, again, that X_3 and X_4 are not significant in the presence of X_1. Thus, we wish to test H_0: $\beta_3 = \beta_4 = 0|X_1$. If this hypothesis holds, the reduced model contains only X_1 (because it restricts the coefficients of X_3 and X_4, which were part of the full model, to be zero). The model-comparison procedure tests whether X_3 and X_4, jointly, contribute (or do not contribute) to the explanatory and predictive power of the model. The key point is that although in both cases we test

parameters associated with the same variables (X_3 and X_4), we are not comparing the same models. The hypothesis being tested is not the same in a substantive sense, and the results of the statistical tests are not identical. Probably the most important lesson for the researcher is that a statistical test provides the machinery for comparing competing models and its results are meaningful only when they are interpreted in the context of these models. Conclusions of the type 'X_3 is significant (because $\beta_3 \neq 0$)' or 'X_4 is not significant (because $\beta_4 = 0$)' without specifying the nature of the models being compared are meaningless, and potentially misleading.

The general hypotheses tested by the model comparison procedure are:

- H_0: In the full model the βs of the predictors in the subset being tested are all zero. The restricted model is better.
- H_1: In the full model the βs of the predictors in the subset being tested are not all zero. The full model is better.

More formally, if the full model contains a total of p predictors, and the predictors are arranged such that the subset to be tested consists of the predictors $q+1$ through p, then the full model is:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots \beta_p X_p + \varepsilon \quad (10)$$

and the hypotheses are:

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0$$
$$H_1 : \beta_{q+1}, \ldots, \beta_p \text{ are not all zero.}$$

The statistical test compares the fit of the two competing models based on their SSE values (as defined in Table 13.2) using the test statistic:

$$F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} \sim F_{(df_R - df_F, df_F)} \quad (11)$$

where SSE_R and SSE_F are the residual (or error) sums of squares for the reduced and full models, respectively, and df_R and df_F are their respective degrees of freedom. If the null hypothesis is true, the F ratio follows the specified F distribution.

This test statistic compares the difference in lack of fit between the two models relative to the lack of fit in the full model, while taking the degrees of freedom into account. On the one hand, if the null hypothesis is true, the restriction on the relevant βs should not significantly impact the SSE, so the numerator, and the test statistic, will be relatively small. On the other hand, if the null hypothesis is false, the full model should perform significantly better (i.e., obtain a substantially smaller SSE) than the reduced model, so the difference between SSE_R and SSE_F will be relatively large leading to a rejection of the null hypothesis.

The test statistic can be written in terms of R^2 values (rather than SSEs) as:

$$F = \frac{(R_F^2 - R_R^2)/(df_F - df_R)}{(1 - R_F^2)/df_F} = \frac{df_F(R_F^2 - R_R^2)}{(df_F - df_R)(1 - R_F^2)} \sim F_{(df_R - df_F, df_F)} \quad (12)$$

where it is perhaps clearer that the test statistic compares the difference in the fit of the two models relative to the lack of fit of the full model (again taking degrees of freedom into account).[6]

The tests for comparing the fit of the models (and, practically, all other tests in MR) are invariant under linear transformations of the predictors and/or the response variable. If, for example, the price of Bordeaux wines (Y) is predicted from the temperatures and amounts of precipitation in the fall and the spring of the year the grapes were picked (Ashenfelter et al., 1995), the tests would not be affected if the wines are priced in US$ or in Euros, if the temperatures are in Celsius or Fahrenheit, if precipitations are measured in inches or centimeters, and so on (note, however, that the regression coefficients would vary as a function of the units used).

Most tests associated with MR are special cases of this model comparison approach. For example, to test the significance of any single predictor in the presence of the other $p - 1$ predictors, the F-test above is equivalent to the t-test of β for that predictor (i.e., $t^2 = F$) that is printed by all statistical packages. Further, the omnibus F-test of the full model's R^2 (or overall fit), which is also printed by all statistical packages, is equivalent to testing the full model (including all p predictors) against a reduced model that includes only the intercept.[7]

One obvious and common use of this test occurs when a set of $(C - 1)$ predictors represents a qualitative predictor, such as the 'type of school' variable described in the section 'Measurement level' and Table 13.1, above. All $(C - 1)$ predictors (e.g., D_{11}, $D_{12,}D_{13}$) are considered jointly, and to test whether school type is a significant predictor one would need to compare a model that includes all $(C - 1)$ binary variables to a model that restricts their associated $(C - 1)$ coefficients to be zero. The result would be invariant across the choice of binary variables (e.g., the D_{1i} or the D_{2i} set in Table 13.1). The individual t-tests of the coefficients in the context of the full model are more difficult to interpret and depend on the coding scheme. For example, the test of the single coefficient associated with D_{11} (see Table 13.1) in the 'Measurement level' section tests the contribution of the distinction between public and 'other' schools in the presence of all the other predictors, including the distinctions between private and 'other' schools (D_{12}), and between parochial and 'other' schools (D_{13}).

For continuous-predictor variables, common uses of this procedure involve situations in which a set of variables are included in the model in a hierarchical fashion and the order of inclusion reflects theoretical (or statistical 'control') considerations. For example, if certain demographic variables are known to affect the outcome, one could compare a reduced model that contains this set of variables only, to a full model that contains an additional set of variables (in addition to the control set). If the null hypothesis is rejected in this case, then the variables in the additional set significantly contribute to prediction of the criterion over and above the variance already accounted for by the control set.

To conduct this test using statistical software, one simply needs to fit both the full and reduced models, obtain their SSE or R^2 values, and use either of the formulae above to compare the models. The major statistical software programs (e.g., SAS, SPSS) also have the capability to provide the test statistic and p-value of this F-test in the output. Examples 1 and 2 (below) illustrate applications of this test.

Example 1: a confirmatory application

Using the Suh et al. (1998) data, we demonstrate the prediction of global life satisfaction using the American sample only ($n = 420$). We model global life satisfaction (the criterion) using domain-specific life-satisfaction variables (namely, satisfaction with one's health, finances, family, nation, housing, self, food) as predictors, and then test whether the addition of variables measuring values for life-satisfaction domains (namely, values on overall life satisfaction, money, humility, love, happiness) or variables measuring affect (namely, the frequency and intensity of experiencing positive and negative affects) contributes significantly to the model.

The results of the analysis are presented in Table 13.3. While the R^2 values for all models are shown, we focus only on those models that include the domain-satisfaction (DS) variables. The model containing the seven DS variables as predictors results in an R^2 value of .510. When the five values measures (V) are added as predictors, the R^2 increases to .513, indicating an R^2 change (ΔR^2) of .003 from the base model. The F-test of this change indicates that it is not significant ($F_{5,407} = 0.47, p < .05$) and, therefore, the addition of the values measures (V) does not contribute to the prediction of global life satisfaction over and above the initial contribution of the DS variables. On the other hand, adding the four affect

measures (A) to the model that includes the DS variables increases the R^2 to .572. This increase ($\Delta R^2 = .062$) is indeed significant ($F_{4,408} = 14.87$, $p < .05$), indicating that the affect measures contribute significantly to predicting global life satisfaction over and above the initial contribution of the DS variables. Not surprisingly, the model that contains the DS variables and both the affect and values measures also performs significantly better than the model that contains only the DS variables, but this is clearly due to the effect of the affect measures on predicting global life satisfaction. In fact the difference in fit between models 5 and 7 ($\Delta R^2 = .002$) is not significant. In conclusion, once we use the DS variables to predict global life satisfaction, the addition of affect measures significantly improves or contributes to the fit of the model whereas the addition of values measures does not, so we favor model 5.

Example 2: another confirmatory application

In this example we predict global life satisfaction from the frequency of negative and positive affect (NA and PA, for short) and examine the potential moderating effects of respondent's sex. The results of the analyses for data from the United States are presented in Table 13.4. The two affect frequency measures (model 1) together account for about 42% of the total variability in global life satisfaction ($R^2 = .416$). Adding gender as a predictor (model 2, which would allow the

Table 13.3 Example 1. Predicting global life satisfaction for the USA sample ($n = 420$)

Variables in the model	df_M, df_E[a]	R^2	ΔR^2	*F for* ΔR^2	*p-value for* ΔR^2
1. Domain satisfaction (DS) only	7, 412	.51	–		
2. Values (V) only	5, 414	.03	–		
3. Affect (A) only	4, 415	.43	–		
4. DS + V	12, 407	.51	.003	0.47	.798
5. DS + A	11, 408	.57	.062	14.87	< .0001
6. A + V	9, 410	.43	–		
7. DS + V + A	16, 403	.57	.064	6.74	< .0001

[a] df_M, model (regression) degrees of freedom $= p$; df_E, error (residual) degrees of freedom $= n - p - 1$.

Table 13.4 Example 2. Predicting global life satisfaction for the USA sample ($n = 438$)

Model	*Variables*	*B*	df_M, df_E[a]	R^2	ΔR^2	*F for* ΔR^2	*p-value for* ΔR^2
1. Common intercept and common slope	PA	.49					
	NA	−.28					
			2, 435	.42	–		
2. Gender specific intercepts and common slope	PA	.51					
	NA	−.26					
	Gender	−.07					
			3, 434	.42	.005	3.56	.060
3. Common intercept and gender specific slopes	PA	.62					
	NA	−.15					
	PA × gender	−.14					
	NA × gender	−.13					
			4, 433	.42	.002	0.83	.437
4. Gender specific intercepts and slopes	PA	.63					
	NA	−.15					
	Gender	−.07					
	PA × gender	−.13					
	NA × gender	−.12					
			5, 437	.42	.007	1.65	.178

[a] df_M, model (regression) degrees of freedom $= p$; df_E, error (residual) degrees of freedom $= n - p - 1$.

regression intercepts to vary depending on one's sex) increases R^2 to .421, but this is not a significant improvement over the fit of the initial model ($F_{1,434} = 3.56$, $p > .05$). The same is true for models containing the interaction of gender with the affect frequency measures (which would allow the regression slopes to vary depending on one's sex). Therefore, the prediction of global life satisfaction from affect frequency does not depend on the sex of the individual.

MR as an exploratory tool: model selection and measures of model (mis)fit

In this section, we cover some of the basic issues involved in using MR as an exploratory tool designed to identify the 'best' set of predictors in the absence of a specific theory. Typically, we have a very large number of potential predictors that are inter-correlated, and we believe that we can find a much smaller subset that would be useful in predicting Y. To fix ideas, consider the standard methodology that was used to develop some of the most widely used vocational interest inventories (e.g., Anastasi, 1982): A large number of items is administered to a large number of respondents who work in a particular field (say, physicians) and who report various levels of satisfaction (and possibly success) in their chosen profession. A scale of interest in medicine is constructed by identifying those items that predict best the level of satisfaction of the various physicians. The key point is that the items are chosen based solely on their predictive efficacy, and not on any theoretical considerations.

The challenge of the techniques reviewed in this section is to balance the two effects of adding more variables to the model: better fit and higher complexity. Although there is general agreement that the most parsimonious solution is one that achieves the best fit relative to the model's complexity, there are many ways of quantifying fit and complexity as well as accounting for their trade-off. Once a measure of model fit is selected it can be used as a selection tool that allows one to compare many competing models (involving different subsets of predictors) and choose the best ones.

The most common measure of model fit is R^2. We saw in the last section that R^2 was a key component in the model comparison tests but it can also be used as a descriptive measure to select models. However, both the sample size and number of predictors in the model affect its value. As a simplistic example, consider a simple regression ($p = 1$) and a sample of $n = 2$. In this case, the scatter-plot of X and Y contains just two points, which can be joined by a straight line and produce a seemingly perfect relationship where $R^2 = 1$, regardless of the true relation between X and Y in the population. The same pattern holds for $p = 2$ and $n = 3$, $p = 3$ and $n = 4$, and so forth. In fact, when the value of ρ^2 in the population is 0 the expected value of the sample R^2 is $p/(n-1)$, which is greater than zero (Pedhazur, 1997). As we showed earlier, the sample R^2 value is biased, and always overestimates its population value, ρ^2. The bias is especially high when the sample size is small relative to the number of predictors. Thus, there is a danger of serious over-fit in cases with many predictors and small samples (see Birnbaum's satire, Sue Doe Nihm, 1976). The sample size needs to be substantially larger than the number of predictors for the sample R^2 to provide a good estimate of its population value. To correct for the model's complexity (number of predictors) relative to the sample size, the adjusted R^2 measure, R^2_{adj}, is used:

$$\begin{aligned} R^2_{adj} &= 1 - \frac{(n-1)}{(n-p-1)}(1-R^2) \\ &= 1 - \frac{(n-1)}{(n-p-1)}\frac{SSE}{SST} \end{aligned} \qquad (13)$$

Note that the adjustment uses the Sums of Squares (SS) divided by their degrees of freedom (i.e., the Mean Square, MS). For any fixed sample size, n, the adjusted R^2 will typically be smaller than the value of R^2 by a factor that is directly related to the number of predictors in the model. Therefore, if two

models fit to the same data set produce the same R^2 values, but one has more predictors than the other, the model with fewer predictors has a larger R^2_{adj} value and is considered to be superior since it achieves the same fit (and accuracy of predictions) with fewer predictors.

Another descriptive measure of model fit that accounts for the number of predictors is Mallows' C_p criterion (Mallows, 1973). If a total of p predictors are available, then for a subset model containing k predictors Mallows' criterion is:

$$C_k = \frac{\text{SSE}_k}{\text{MSE}} - [n - 2(k+1)] \qquad (14)$$

where SSE_k is the error sum of squares for the subset model and MSE is the mean square error for the full model containing all p predictors. Mallows' criterion is concerned with identifying an unbiased model. If there is no bias in the predicted values of the model (i.e., $E(\hat{Y}_i) = \mu_i$), the expected value of C_k is approximately $k + 1$ (Mallows, 1973; Neter et al., 1996). Thus, for the full (p-predictors) model, $C_p = p+1$ and the fit of any (k-predictors) subset model is evaluated by comparing its C_k value to $k+1$, where a small difference indicates good fit (i.e., no bias). Biased models result in C_k values greater than k, so we typically seek models with C_k values that are both small and close to k.

A measure of fit that is based on the ML (or information) function of a regression model with p predictors is Akaike's information criterion (AIC) developed by Akaike (1970; 1973):

$$\text{AIC} = n\ \ln(\text{SSE}/n) + 2(p+1) \qquad (15)$$

where SSE is the error sum of squares for the model in question and *smaller* values of AIC indicate better fit (AIC is a measure of loss of information in fitting the model, which we wish to minimize). All other things being equal, as SSE (and its logarithm) decreases, indicating better fit, AIC decreases. As with other measures of fit, AIC increases as the model's complexity (p) increases.

Schwarz's Bayesian Criterion (Schwarz, 1978), also known as the Bayesian information criterion (BIC), is a slight variation on AIC with a more severe penalty for the number of predictors:

$$\text{BIC} = n\ \ln(\text{SSE}/n) + (p+1)\ln(n) \qquad (16)$$

In general, BIC penalizes the fit more severely than AIC for the number of predictors in a model. Therefore, when several competing models are fit to the same data the BIC measure is likely to select a model with fewer predictors than the AIC measure.

The measures discussed above do not provide an exhaustive list (see, for example, Miller, 1990; Burnham and Anderson, 2002) but are arguably the most commonly encountered measures of fit in the social sciences. The various measures can be calculated for each of the feasible 2^p distinct subset models that can be generated from p predictors, and they all account in one way or another for the number of predictors in each subset model. This approach is referred to as 'all subsets regression', and can be used to identify the 'best' model (by whatever measure). There are two approaches for identifying the 'best' model: it can be done by conditioning on level of complexity (i.e., considering the single best predictor, the best pair, the best triple, etc.) and choosing the best subset model within a given level of complexity; alternatively, all models can be simply rank ordered regardless of complexity to choose the best ones.[8] The final selection is typically based on simple numerical and/or visual comparisons and, typically, does not involve significance tests.

An alternative approach relies on a family of automated computer algorithms – forward selection, backward elimination and stepwise regression – that were developed before the computations involved in the 'all subsets' approach were feasible. These techniques involve convenient shortcuts and rely heavily on significance tests as a decision tool to include new variables in the model, exclude predictors from the model, and stop the search (for algorithmic details see, for example,

Draper and Smith, 1998; Neter et al., 1996). These techniques share several drawbacks, and are inferior to the most modern approach of all subsets regression. The first problem is that not all possible models are considered, so there is no guarantee that the final model selected is necessarily the 'best' according to any criterion. The second major problem is that they use a very large number of tests without any adjustment for test multiplicity.[9] We recommend use of the stepwise procedure (the most flexible of the three) only in cases where the number of predictors is extremely large (in the hundreds) as a preliminary step to identify a manageable subset of variables to be examined later by the 'all subsets' method. Example 3 illustrates an application of model selection procedures.

Example 3: an exploratory application

In the life-satisfaction data set, we have a total of 17 potential predictors of global life satisfaction (seven domain-specific satisfaction variables, five values measures, four affect measures and sex). In the absence of theory, one may wish to fit all $2^{17} = 131{,}072$ distinct subset models possible and explore which model(s) might provide the best fit in predicting global life satisfaction. As with most exploratory analyses, this can shed some light on potential theories for predicting life satisfaction that can subsequently be confirmed with additional data. To illustrate this procedure, Tables 13.5 and 13.6 show the top models (based on fit) that can be formed from the 17 predictors available and various model-fit measures for these models.

Table 13.5 shows the model that fits best for each level of complexity. For example, the single best predictor is satisfaction with self, the best pair of predictors contains the satisfaction with self and frequency of positive affects, and the best triple of predictors also adds satisfaction with family. The table also lists the various fit measures for these selected models. Note that R^2_{adj} favors a model with nine predictors (highest value), while AIC and BIC reach their desired minimal values for the models with eight and seven predictors, respectively.

Table 13.6 shows the top five models using two model-selection criteria (adjusted R^2 and C_p) with values rounded to 2 decimal places. The model that produces the highest adjusted

Table 13.5 Example 3. Best-fitting models for predicting global life satisfaction for data from US ($n = 420$), for various levels of complexity

Size of model (p)	*Adj. R^2*	*R^2*	*C_p*	*AIC*	*BIC*	*Variables in model*
1	0.39	0.39	159.66	1397.54	1398.31	X7
2	0.48	0.47	81.68	1336.71	1337.75	X7 X13
3	0.53	0.52	36.22	1296.44	1297.93	X3 X7 X13
4	0.54	0.54	25.89	1286.73	1288.36	X3 X4 X7 X13
5	0.55	0.55	15.27	1276.39	1278.29	X3 X4 X7 X13 14
6	0.56	0.56	8.88	1269.97	1272.14	X2 X3 X4 X7 X13 X14
7	0.57	0.56	2.87	1263.79	1266.30	X2 X3 X4 X5 X7 X13 X14
8	0.57	0.56	2.86	1263.70	1266.37	X2 X3 X4 X5 X7 X13 X14 X17
9	0.57	0.56	3.74	1264.54	1267.34	X2 X3 X4 X5 X7 X9 X13 X14 X17
10	0.57	0.56	4.77	1265.53	1268.47	X2 X3 X4 X5 X7 X9 X13 X14 X15 X17
11	0.58	0.56	6.19	1266.93	1269.99	X2 X3 X4 X5 X6 X7 X9 X13 X14 X15 X17
12	0.58	0.56	8.08	1268.81	1271.97	X2 X3 X4 X5 X6 X7 X9 X12 X13 X14 X15 X17
13	0.58	0.56	10.04	1270.77	1274.02	X2 X3 X4 X5 X6 X7 X9 X12 X13 X14 X15 X16 X17
14	0.58	0.56	12.02	1272.76	1276.09	X2 X3 X4 X5 X6 X7 X8 X9 X12 X13 X14 X15 X16 X17
15	0.58	0.56	14.01	1274.74	1278.17	X1 X2 X3 X4 X5 X6 X7 X8 X9 X12 X13 X14 X15 X16 X17
16	0.58	0.56	16.00	1276.73	1280.25	X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X12 X13 X14 X15 X16 X17

Satisfaction domains are: health (X1), finances (X2), family (X3), housing (X4), food (X5), country (X6), self (X7) value domains are, life satisfaction (X8), money (X9), humility (X10), love (X11), happiness (X12) affect measures are, positive frequency (X13), negative frequency (X14), positive intensity (X15), negative intensity (X16), and sex (X17)

Table 13.6 Example 3. The five best-fitting models for predicting global life satisfaction for US sample ($n = 420$) based on two selection criteria

Size of model (p)	*Adj. R^2*	*R^2*	*C_p*	*AIC*	*BIC*	*Variables in model*
Using adjusted R^2						
9	0.56	0.57	3.73	1264.54	1267.34	X2 X3 X4 X5 X7 X9 X13 X14 X17
10	0.56	0.57	4.78	1265.53	1268.47	X2 X3 X4 X5 X7 X9 X13 X14 X15 X17
8	0.56	0.57	2.86	1263.70	1266.37	X2 X3 X4 X5 X7 X13 X14 X17
9	0.56	0.57	3.92	1264.74	1267.53	X2 X3 X4 X5 X7 X13 X14 X15 X17
10	0.56	0.57	5.13	1265.91	1268.82	X2 X3 X4 X5 X6 X7 X9 X13 X14 X17
Using C_p						
8	2.86	0.57	0.56	1263.70	1266.37	X2 X3 X4 X5 X7 X13 X14 X17
7	2.87	0.57	0.56	1263.79	1266.30	X2 X3 X4 X5 X7 X13 X14
8	3.66	0.57	0.56	1264.54	1267.17	X2 X3 X4 X5 X7 X13 X14 X15
9	3.73	0.57	0.56	1264.54	1267.34	X2 X3 X4 X5 X7 X9 X13 X14 X17
8	3.91	0.57	0.56	1264.80	1267.42	X2 X3 X4 X5 X7 X9 X13 X14

Satisfaction domains are: health (X1), finances (X2), family (X3), housing (X4), food (X5), country (X6), self (X7), value domains are life satisfaction (X8), money (X9), humility (X10), love (X11), happiness (X12), affect measures are positive frequency (X13), negative frequency (X14), positive intensity (X15), negative intensity (X16), and sex (X17).

R^2 contains 9 predictors, and includes some variables from each set (i.e., domain-specific satisfaction variables, some affect measures, a value measure and sex). The model that produces the best fit using the C_p criterion contains the same variables except for the value measure, so it selects an eight-predictor model as best fitting.

By perusing the output from such exploratory procedures, one may begin to discern informative patterns. Certain variables commonly appear as predictors in most of the top models and certain variables never appear in such models. This may then guide the development of some theories regarding which variables appear to be responsible for satisfaction with life, and which do not. To find support for these theories, confirmatory analyses may be conducted using additional data.

Multiple regression as a predictive tool: interval estimation of predictions and cross-validation

Once a particular model is selected (by any of the methods described above), it can be used as a predictive tool for specific values of the criterion. For each unique combination of values of the p predictors (and intercept term $X_0 = 1$), $X_\mathrm{h} = \{1, X_{\mathrm{h}1}, \ldots, X_{\mathrm{h}p}\}$, the expected value of Y is given by:

$$E(Y_h) = \beta_0 + \beta_1 X_{h1} + \beta_2 X_{h2} + \ldots + \beta_p X_{hp} = \mathbf{X}_\mathrm{h}\text{'}\boldsymbol{\beta} \quad (17)$$

where $\mathbf{X}_\mathrm{h}$ is a $(\mathrm{p} + 1) \times 1$ vector containing the predictor values and $\boldsymbol{\beta}$ is a $(\mathrm{p} + 1) \times 1$ vector containing the regression coefficients (including the intercept). Therefore, $E(Y_h)$ indicates the expected value of the criterion (i.e., its value in the population) given this $\mathbf{x}_\mathrm{h}$ predictor vector. In a particular sample the regression parameters are replaced by their sample estimates and the predicted value of Y is given by:

$$\hat{Y}_h = b_0 + b_1 X_{h1} + b_2 X_{h2} + \ldots + b_p X_{hp} = \mathbf{x}_\mathrm{h}\text{'}\mathbf{b} \quad (18)$$

where $\mathbf{b}$ is a $(\mathrm{p} + 1) \times 1$ vector containing the regression coefficient estimates. Therefore, $\hat{Y}_h$ is the expected value of all $\boldsymbol{Y}$ values associated with the predictor vector $\mathbf{x}_\mathrm{h}$ and is an unbiased estimate of $E(Y_h)$.[10] The variance of $\hat{Y}_h$ is estimated by pre- and post-multiplying the variance of $\mathbf{b}$ by the $\mathbf{x}_\mathrm{h}$ vector, so:

$$\mathrm{s}^2\{\hat{Y}_h\} = \mathrm{x_h}\text{'}\, \mathrm{s}^2\{\mathbf{b}\}\mathbf{x}_\mathrm{h} = \mathrm{MSE}(\mathbf{x}_\mathrm{h}\text{'}\, (\mathbf{X'X})^{-1}\mathbf{x}_\mathrm{h}) \quad (19)$$

where $\mathrm{s}^2\{\mathbf{b}\} = (\mathrm{MSE})(\mathbf{X'X})^{-1}$. MSE is the mean square error of the model, and $\mathbf{X}$

is the n × (p + 1) data matrix containing all n observations on all p predictors and an intercept term (e.g., Neter et al., 1996). Therefore, an interval estimate for $E(Y_h)$ is given by:

$$\hat{Y}_h \pm t(1-\alpha/2; n-p-1)s\{\hat{Y}_h\} \qquad (20)$$

where $(1-\alpha)$ is the desired level of confidence of the interval (e.g., for a 95% confidence interval $1-\alpha = .95$ so $\alpha = .05$). The width of the confidence interval is a quadratic function of the distance between the mean values of the predictors in the sample (contained in the vector $\bar{\mathbf{x}}$) and the target $\mathbf{x}_\mathrm{h}$. In other words, predictions are most accurate when $\mathbf{x}_\mathrm{h} = \bar{\mathbf{x}}$, and their accuracy decreases (quadratically) as one moves away from this point, highlighting the dangers of extreme extrapolations.

In addition to predictions for the mean $\boldsymbol{Y}$ value given $\mathbf{x}_\mathrm{h}$, one may be interested in predicting a single new observation. This amounts to randomly selecting one observation with predictor values $\mathbf{x}_\mathrm{h}$. The point prediction is identical, but this selection induces a higher level of uncertainty. The variance of the predicted value based on a single new observation is $\mathrm{s}^2\{\hat{Y}_{h(\mathrm{new})}\} = \mathrm{MSE} + \boldsymbol{x}_\mathrm{h}'\mathrm{s}^2\{\boldsymbol{b}\}\boldsymbol{x}_\mathrm{h} = \mathrm{MSE}(1+\boldsymbol{x}_\mathrm{h}'(\boldsymbol{X'X})^{-1}\boldsymbol{x}_\mathrm{h})$, resulting in the confidence interval $\hat{Y}_h \pm t(1-\alpha/2; n-p-1)s\{\hat{Y}_{h(\mathrm{new})}\}$.

Beyond the concern with predictions of specific values, one may seek ways to quantify the quality of the model's predictive validity as a whole. Intuitively, one may think that R^2, the coefficient of determination, provides such a measure. Because the parameter estimates are based on data from a particular sample, the R^2 value is affected by the sample's idiosyncrasies. It is maximal for the sample at hand, but not for others. Therefore, if another random sample is obtained from the same population, and the model from the original sample is used to predict values in this new sample, the new R^2 value would be lower than in the original sample, a phenomenon sometimes referred to as 'shrinkage'. The degree of shrinkage depends on p, n, and R^2. In general, the smaller the shrinkage the higher the confidence that the model generalizes well to other samples.

The procedure used to evaluate how well the regression model – developed using one sample – generalizes to other samples is called cross-validation. The simplest cross-validation calls for a division of the available sample into two smaller random samples (e.g., halves). One sample (or half), sometimes referred to as the screening or training sample, is used to estimate the model parameters and obtain the (optimal) fit of the model. The second sample, sometimes referred to as the validation or prediction sample, is used to obtain predicted values based on the parameter estimates computed previously (using the training sample). The R^2 obtained when applying the parameters estimated in the training sample to the observations in the validation sample is the cross-validated (and typically 'shrunk') R^2. Ideally, the values of statistics such as the MSE, $\mathbf{b}$, R^2, and so on from the validation sample should be relatively close to their counterparts in the training sample.

Variations on this cross-validation procedure involve splitting the data set into more than two parts, each time leaving out one part and using the remaining data as a training sample and the left-out set as the validation sample. In the extreme, the left-out data includes a single observation, such that training data set consists of all but one observation, and the accuracy of the estimated model is evaluated by obtaining the prediction for a single observation at a time. This is also known as the 'jack-knifing' procedure (Mosteller and Tukey, 1977).

Alternatively, and especially if the sample is too small to allow for data splitting, one can predict statistically the degree of shrinkage. Pedhazur (1997) discusses formulae [attributed both to Stein (1960) and Herzberg (1969)] that have been shown, in simulation studies, to accurately estimate the cross-validation coefficient (the R^2 in the validation sample) without actually carrying out the cross-validation process (Cotter and Raju, 1982). For a model with fixed predictors, the

estimate of the cross-validation coefficient, $\hat{R}^2_{CV}$, is:

$$\hat{R}^2_{CV} = 1 - \left(\frac{n-1}{n}\right)\left(\frac{n+p+1}{n-p-1}\right)(1-R^2) \quad (21)$$

where R^2 is the squared multiple correlation coefficient of the full sample and n is the sample size of the full sample. For a model with random predictors the estimate is:

$$\hat{R}^2_{CV} = 1 - \left(\frac{n-1}{n-p-1}\right)\left(\frac{n-2}{n-p-2}\right)\left(\frac{n+1}{n}\right)(1-R^2) \quad (22)$$

ADDITIONAL TOPICS IN MULTIPLE REGRESSION

In this section we discuss two additional topics that affect the interpretation of MR multiple regression results: collinearity among predictors and relative importance of predictors.

Collinearity

Collinearity occurs when some of the predictors in a dataset are linear combinations of other predictors. If one predictor can be perfectly predicted from a linear combination of other predictors, the ***X'X*** matrix used in estimating model parameters is singular and, therefore, there is no unique set of unbiased estimates of the parameters. From a practical perspective collinearity implies redundancy in the information provided by the set of predictors and indicates that not all predictors are needed (i.e., the model is mis-specified). Collinearity analysis is typically related to a confirmatory approach, as it can be used to identify mis-specification of the model. However, some researchers may also use it in an exploratory manner to screen and exclude some predictors from the analysis.

Typically, collinearity is not perfect. However, when one predictor is almost perfectly predictable from the others (say with an R^2 greater than 0.9), a situation described as near collinearity, most researchers would agree that this is an unacceptable level of redundancy. Statistically, (near) collinearity can make the regression coefficients take arbitrarily high values and switch sign in an unpredictable fashion, as well as inflate their standard errors. While the correlation matrix of the predictors can be inspected for unacceptably high bivariate correlations, patterns of collinearity due to more complex linear combinations of the variables may not be detected by simple inspection. Therefore, several measures have been proposed for detecting collinearity.

The 'global' approach to measuring collinearity involves examining global measures of the system (i.e., the p predictors). Under perfect colinearity the ***X'X*** matrix is of deficient rank ($< p$) and at least one of its eigenvalues is 0. Thus, small (i.e., near-zero) eigenvalues indicate near-singularity and are diagnostic of near-collinearity. A popular measure is the condition number, defined as the square root of the ratio of the largest eigenvalue of ***X'X*** (standardized to unit variances) to the smallest eigenvalue.[11] Some guidelines (not necessarily agreed upon) suggest that for this measure values in the range of 30–100 indicate moderate collinearity and values over 100 indicate strong collinearity (Belsley, 1991). Examination of the coefficients of the eigenvector associated with very small eigenvalues can help identify the linear combinations that induce the dependency among the predictors.

The 'local' approach to measurement of collinearity involves identification of highly predictable predictors. Let R^2_i be the squared multiple correlation obtained in performing a MR in which X_i is predicted from the other $(p-1)$ predictor variables. Two measures based on R^2_i are often reported. One is the variance inflation factor (VIF) that is defined as: $\text{VIF}_i = 1/(1-R^2_i)$. As the name indicates, it represents the inflation in the variance of b_i due to correlations among the predictors, where the base line is the case of uncorrelated variables when $R^2_i = 0$ and $\text{VIF}_i = 1$).

High values of VIF_i indicate that X_i may be a linear combination of the other predictors. Because VIF values are unbounded, they are sometimes rescaled into measures of 'tolerance' defined as $1/VIF_i = (1 - R_i^2)$. Tolerance, therefore, varies from a minimum of 0 to a maximum of 1. There is no clear agreement on what values of VIF or tolerance are considered to be indicative of severe collinearity problems. To some extent, this is a subjective decision that involves deciding what value of R^2 (among the predictors!) would be considered unacceptably high. The measures can also be used in a relative fashion to identify the variables with the lowest tolerance (highest VIF).

Relative importance

A question that often comes up in interpretation of MR models relates to the determination of the *relative importance* of predictors. Researchers are typically interested in: (1) ranking the predictors from the most to the least important; (2) scaling them, by assigning values on an interval scale that reflects their importance, and possibly (3) relating these measures to the model's overall goodness of fit (Budescu, 1993). Despite objections (see Pratt, 1987), many researchers have proposed ways to measure the relative importance of predictors [see Budescu (1993) and Kruskal and Majors (1989) for partial reviews]. A surprising conclusion of the reviews of the predictor importance literature is that there is no universally accepted definition or a generally accepted measure of importance. In fact, some of the measures proposed are not explicitly related to any specific definition.

Many researchers, incorrectly, equate a predictor's relative importance with the magnitude of its standardized regression coefficient. Criticisms of this misleading interpretation are published periodically in the professional literature in various disciplines. For some examples the reader should consult: Greenland et al. (1986) in epidemiology; King (1986) in political sciences; Budescu (1993) and Darlington (1968) in psychology; Bring (1994), Kruskal and Majors (1989), and Mosteller and Tukey (1977) in statistics. In the next section we review briefly the major shortcomings of the standardized coefficients as measures of importance. We conclude that, unless the predictors are uncorrelated, standardized coefficients cannot be said to isolate and measure a net, direct, or unique effect of the corresponding target predictors.

In addition to importance measures that are based on standardized regression coefficients, other measures proposed in the relative importance literature are typically based on correlations (e.g., Hedges and Olkin, 1981; Kruskal, 1987; Lindeman et al., 1980; Mayeske, et al., 1969; Mood, 1969, 1971; Newton and Spurrell, 1967a, 1967b; Pedhazur, 1975), a combination of the coefficients and correlations (Courville and Thompson, 2001; Darlington, 1968; Dunlap and Landis, 1998; Pratt, 1987; Thomas and Zumbo, 1996; Thomas et al., 1998) or on information (Soofi and Retzer, 2000; Soofi et al., 2000; Theil, 1987; Theil and Chung, 1988). We discuss two alternative and, in our view, superior methods of meaningfully evaluating relative importance: d*ominance analysis* (DA; Azen and Budescu, 2003; Budescu, 1993) in the confirmatory context, and c*riticality analysis* (CA; Azen et al., 2001) in the exploratory context.

Why standardized coefficients are not measures of relative importance

The usual interpretation of the coefficients is the rate of change in Y per unit change in X_i (or slope) when all other variables are fixed (held constant). We have already discussed the fact that this interpretation is inadequate for polynomial and/or interactive models where the predictors are functionally related. We recommend centering (or, alternatively, standardizing) the predictors in polynomial and/or interactive models (see the section 'Derivative predictors', above), to reduce collinearity. This, however, does not affect their interpretation. In this section, we explain why standardized coefficients should

not be interpreted as measures of relative importance.

Consider now the case of distinct predictors. The definition does not specify at what values to fix the other predictors and, it is natural to infer that this rate of change is invariant across all choices of the values of the other predictors, X_j (j $\neq$ i). This is indeed true if all p predictors are mutually uncorrelated and/or the p predictors and the response have a $(p + 1)$-variate normal distribution, but not necessarily in other cases [see Lawrance (1976) for a proof and discussion]. The intuition is quite simple: the higher the correlations between the predictors, the closer this case comes to the situation where the predictors are functionally related in the sense that changing one variable implies changes in the others. And conditioning on various levels of any one predictor focuses on different subsets of the target population. Thus, the interpretation of the standardized (or raw) coefficient as a fixed rate of change for the case of distinct predictors is contingent on strict assumptions about the distribution of, and intercorrelations between, the p predictors.

The next issue is how to interpret the sign of the coefficient: can one seriously talk about *negative importance*? Or, should one interpret importance as we interpret correlations, that is, by distinguishing between the absolute magnitude of the coefficient (a measure of overall importance) and its sign (an indication of the direction of the effect)? It turns out that neither approach captures faithfully the behavior of the coefficients. This determination follows directly from the elegant analysis of *suppressor* variables by Tzelgov and his colleagues (e.g., Tzelgov and Henik, 1991; Tzelgov and Stern, 1978). They show examples of cases where all the predictors are positively inter-correlated and their correlations with the response have identical signs, but in each case one of the regression coefficients changes sign! Similarly, in the case of (almost) perfect co-linearity ($r_{X_1X_2} = 0.99$) where the two predictors correlate with the response almost identically ($r_{yX_1} = 0.61$ and $r_{yX_2} = 0.60$), one of the coefficients is positive ($b_1^* = .804$), and the second is negative ($b_2^* = -.196$). Clearly, the sign of the regression coefficient is unrelated to any sensible definition of the predictors' importance.

Bring (1994) shows that, contrary to the implication of the interpretation of the regression coefficient as a measure of importance, the magnitude of the standardized coefficients does not reflect the effect of the corresponding predictors on the model's goodness of fit. Finally, it is well known that the ranking of the predictors based on their standardized coefficients is model dependent and it is not necessarily preserved in all subset models.

Relative importance and dominance analysis

In the previous section we have argued that interpreting standardized coefficients as measures of importance can lead to paradoxical situations that defy common sense and natural intuitions about importance. This was done without actually proposing a clear definition of this elusive concept. In this section we propose a definition and describe a methodology dominance analysis (DA) that is more suitable for the determination of relative importance in linear models. Budescu (1993) suggested that the importance of any predictor should: (1) be defined in terms of the variable's effect on the model's fit; (2) be based on direct and meaningful comparisons of the target predictor with all the other predictors; (3) reflect the variable's contribution to the fit of the full model under consideration, as well as to all its possible subsets; and (4) recognize indeterminate situations in which it is impossible to rank (some of the) predictors in terms of their importance. He also proposed that relative importance be derived and inferred from the relationship between all $p(p - 1)/2$ distinct pairs of predictors. Predictor X_i dominates (completely) predictor X_j (for short, X_iDX$_j$) if X$_i$ contributes more to the model's fit in all sub-models that include neither X_i nor X_j. In other words, X_iDX$_j$ if in all the instances (with p predictors, there are 2^{p-2} such cases) where

one has the option of including only one of the two variables in a model, considerations of goodness of fit would never favor X_j (i.e., X_i always contributes at least as much as X_j to the model's fit). For example, in a model with $p = 4$ predictors, X_1 would be considered to dominate (completely) X_2 if its contribution to the fit of the model would be higher in the following $2^{(4-2)} = 4$ cases: (i) as a single predictor, (ii) in addition to X_3 alone, (iii) in addition to X_4 alone, and (iv) in addition to X_3 and X_4 as a set. This is an explicit, precise, and quite stringent definition of importance that is consistent with most researchers' intuitions and expectations about it. DA can be applied meaningfully in any substantive domain and for any type of model (distinct predictors, polynomial, interactive, etc.), and it is free from the various interpretational problems that plague the (standardized) coefficients.

Note that if p $\geq$ 3, complete dominance involves more than one comparison among each pair of variables, so it is possible that neither predictor dominates the other. Consequently, in some cases it may be impossible to rank-order all p predictors, although in most cases it is possible to establish a partial order. To address this situation Azen and Budescu (2003) have developed two weaker versions of dominance – conditional and general. Conditional dominance relies on the mean contribution of X_i (specifically, its squared semi-partial correlations) in all models of a given size. Finally, general dominance relies on C_{x_i}, the average of these mean size-specific contributions of X_i across all model sizes (see also Lindeman et al., 1980). An interesting property of these measures is that they add up to the (full) model's fit:

$$R^2 = \sum_{i-1}^{p} C_{x_i} \tag{23}$$

In other words, the $C_{x_i}(i = 1, \ldots, p)$ decompose or distribute the model's global fit across all p predictors. Table 13.7 illustrates this approach with a model where global satisfaction with life is predicted by $p = 4$

Au: subscript i roman?

Table 13.7 Dominance analysis example (with p = 4 predictors)

Subset model	R^2	*Additional contribution of*			
		X_1	X_2	X_3	X_4
Null and k = 0 average	.000	.054	.128	.229	.100
X_1	.054		.103	.206	.081
X_2	.128	.030		.162	.058
X_3	.229	.032	.061		.039
X_4	.100	.035	.086	.168	
$k = 1$ average		.032	.083	.179	.059
X_1X_2	.157			.153	.049
X_1X_3	.260		.050		.031
X_1X_4	.135		.072	.157	
X_2X_3	.290	.020			.025
X_2X_4	.186	.021		.129	
X_3X_4	.268	.024	.047		
$k = 2$ average		.022	.056	.146	.035
$X_1X_2X_3$	.310				.020
$X_1X_2X_4$	.207			.124	
$X_1X_3X_4$	.291		.039		
$X_2X_3X_4$	.314	.016			
$k = 3$ average		.016	.039	.124	.020
$X_1X_2X_3X_4$	.330				
Overall average		.031	.076	.169	.054

X_1, health; X_2, finances; X_3, family; X_4, housing.

domain specific variables.[12] The results indicate that satisfaction with family (X_3) completely dominates each of the other three predictors (its contribution to the model's fit is greater than any of the other predictors in each of the rows of the table where they can be compared), satisfaction with finances (X_2) completely dominates the remaining two predictors, and satisfaction with housing (X_4) dominates satisfaction with health (X_1) as a predictor of overall satisfaction with life. The overall model's fit is $R^2 = 0.33$ and can be distributed among the four variables as .03 to health, .08 to finances, .17 to family, and .05 to housing.

Relative importance and criticality analysis

Dominance analysis is particularly relevant and useful for cases that involve relatively few predictors ($p < 10$), where there is interest in a complete ranking of their contribution to the model's overall fit and a complete understanding of the relationships between the predictors (Azen and Budescu, 2003, describe a slight variation on the main theme, that allows one to perform 'constrained DA' that includes certain groups of variables). Thus, DA is most appropriate for the confirmatory applications of MR. Azen et al. (2001) developed an alternative approach, criticality anaylsis (CA), that was motivated by, and consistent with, the logic of the exploratory applications of MR. In a nutshell, for CA one uses a large number, B, of bootstrapped re-samples of size n (taken with replacement from the original sample of n observations). In each re-sample one can invoke his/her favorite selection method (e.g., adjusted-R^2, AIC, BIC, etc.), to pick the 'best-fitting model (BFM)'. This produces a distribution of BFMs across the B bootstrap samples. Next, one can use this distribution to calculate, for each variable, the fraction of BFMs in which it was included. This measure varies from 0 (the target variable was never included in any of the BFMs) to 1 (the target variable was included in all of the BFMs). This index can be interpreted as the probability of BFM mis-specification when the target variable is omitted, so it measures how critical the target variable is to the identification of the 'best' model. Hence we refer to it as the predictor's criticality.

Table 13.8 illustrates this approach with the life satisfaction data using nine variables [five domain satisfaction (DS) variables and four values]. We ran $B = 1000$ re-samples and used adjusted-R^2 and AIC to identify the BFMs. The top panel presents the frequency distribution of the best-fitting models,[13] and the bottom panel calculates the criticality of the nine variables.

The results indicate that satisfaction with family, finances and housing are essential (highly critical), as they are both included in a high percentage of BFMs using both selection criteria (i.e., AIC and adjusted R^2). Satisfaction with family, for example, appears in every BFM (criticality is 1.0) using both selection criteria. However, predictors such as value on humility and love have relatively low criticality values, as these predictors were included in well under half of the BFMs. In general, the domain-satisfaction predictors resulted in much higher criticalities than the value-related predictors in predicting overall satisfaction with life.

FINAL REMARKS

We reviewed briefly and in a relatively non-technical fashion the major MR results and emphasized various interpretational issues while highlighting how they relate to different applications of the model. Given its long history and wide-spread use, there are many excellent books that cover these applications, as well as many others that we did not touch on, in a more comprehensive fashion and at various levels of technical details. Some of our favorites, several of which we cited repeatedly in this chapter, are Chatterjee and Hadi (2000), Cohen et al. (2003), Draper and Smith (1998), Graybill (1976), Johnson and Wichern (2002), Kleinbaum, et al. (2007), Mosteller and Tukey (1977), Neter et al. (1996), Pedhazur (1997), and Timm (2002).

Table 13.8 Distribution (percentage) of best-fitting models and predictor criticality in 1000 re-samples

(a) Best-fitting models

Best-fitting model	*AIC*	*Adjusted R^2*
X1 X5 X6 X7 X8 X9	16.4	11.6
X5 X6 X7 X8 X9	15.0	
X1 X4 X5 X6 X7 X8 X9	7.1	10.9
X5 X6 X7 X9	6.5	
X3 X5 X6 X7 X8 X9	6.5	5.5
X1 X3 X5 X6 X7 X8 X9	5.4	9.3
X1 X3 X4 X5 X6 X7 X8 X9		6.5
X1 X2 X5 X6 X7 X8 X9		6.4
X5 X6 X7 X8 X9		6.2
Other models	43.1	43.6

(b) Predictor criticality

	Criticality	
Predictor	*AIC*	*Adjusted R^2*
X1: value on money	.466	.621
X2: value on humility	.180	.348
X3: value on love	.259	.417
X4: value on happiness	.261	.421
X5: satisfaction with health	.929	.960
X6: satisfaction with finances	.995	1.000
X7: satisfaction with family	1.000	1.000
X8: satisfaction with country	.823	.929
X9: satisfaction with housing	.966	.984

AIC, Akaike's information criterion.

ACKNOWLEDGMENTS

We are grateful to Drs Hans Friedrich Koehn and Albert Maydeu Olivares for useful comments on an earlier version of the chapter.

NOTES

1 We prefer to reserve the terms independent/dependent variables to randomized experimental designs, and will use the criterion/predictors terminology throughout the chapter.

2 This amounts to dividing the standardized variables (Zy, Zx_i) by $\sqrt{(n-1)}$. The name is due to the fact that under this transformation the parameter estimates can be obtained directly from the correlation matrix among the response and the predictors.

3 Various sources refer to such variables as 'categorical', 'indicator' or 'dummy' variables.

4 This set of transformations is part of the more general family of power transformations $X' = X^{\lambda}$. Polynomials are defined by natural (positive integers) exponents, but the general family includes all real values and includes other 'standard' transformations such as square root ($\lambda = 0.5$), reciprocal ($\lambda = -1$), logarithmic ($\lambda = 0$, by definition), etc. It is often used to optimize the fit of the model. In particular, the well known Box–Cox procedure (Box and Cox, 1964) provides a convenient way to find the 'best' exponent.

5 A 'mixed' design is a combination of the fixed and random designs where some of the predictors are fixed and the others are random variables.

6 This form also highlights nicely the decomposition of the test statistic as a product of the 'Size of the effect' and the 'Size of the study' (e.g., Maxwell and Delaney, 2004).

7 Note that this omnibus test is not the same as the F-test for lack of fit, which tests whether the model satisfies the linearity assumption. Details on the lack of fit test can be found, for example, in Neter et al. (1996).

8 The two approaches don't necessarily lead to identical solutions.

9 The tests in these procedures select the highest (or the lowest) test statistic from a large number of tests without adjusting for the potential capitalization on chance inherent in such a process.

10 Strictly speaking, the predictions are unbiased only if we fit the correct model. Although it is impossible to actually establish this fact, researchers routinely make this assumption.

11 The conditioning number is closely related to the internal correlation (Joe and Mendoza, 1989),

the upper bound of all the simple, multiple and canonical correlations that can be defined among the p predictors.

12 We only use four variables so that we can show the results of the complete analysis in a relatively small table. A SAS macro that can analyze up to $p = 10$ predictors can be downloaded from http://www.uwm.edu/~azen/damacro.html

13 In the interest of space we only present those models that were selected as best at least 50 times (5%) by one of the two criteria.

REFERENCES

Agresti, A. (1996) *An introduction to Categorical Data Analysis*. Wiley: New York.

Aiken, L.S. and West, S.G. (1991) *Multiple Regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage.

Akaike, H. (1970) 'Statistical predictor identification', *Annals of the Institute of Statistical Mathematics,* 22: 203–217.

Akaike, H. (1973) 'Information theory and an extension of the maximum likelihood principle', in Petrov, B.N. and Csaki, F. (eds.), *2nd International symposium on information theory*. Budapest: Akailseoniai-Kiudo. pp. 267–281.

Alf, E.F. and Graf, R.G. (2002) 'A new maximum likelihood estimator of the squared multiple correlation', *Journal of Behavioral and Educational Statistics*, 27: 223–235.

Anastasi, A. (1982) *Psychological testing* (5th edn.). New York: MacMillan Publishing.

Ashenfelter, O., Ashmore, D., and Lalonde, R. (1995) 'Bordeaux wine vintage quality and the weather', *Chance*, 8: 7–14.

Azen, R. and Budescu, D.V. (2003) 'The dominance analysis approach for comparing predictors in multiple regression', *Psychological Methods,* 8: 129–148.

Azen, R., Budescu, D.V., and Reiser, B. (2001) 'Criticality of predictors in multiple regression', *British Journal of Mathematical and Statistical Psychology*, 54: 201–225.

Belsley, D. A. (1991) *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* New York: Wiley.

Box, G.E.P. and Cox, D.R. (1964) 'An analysis of transformations', *Journal of Royal Statistical Society, Series B*, 26: 211–246.

Box, G.E.P. and Jenkins, G. (1976) *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Bradley, R.A. and Srivastava, S.S. (1979) 'Correlation in polynomial regression', *The American Statistician,* 33: 11–14.

Bring, J. (1994) 'How to standardize regression coefficients', *The American Statistician*, 48: 209–213.

Bryk, A.S. and Raudenbush, S.W. (1992) *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Budescu, D.V. (1980) 'A note on polynomial regression', *Multivariate Behavioral Research,* 15: 497–508.

Budescu, D.V. (1985) 'Analysis of dichotomous variables in the presence of serial dependence', *Psychological Bulletin*, 97: 547–561.

Budescu, D.V. (1993) 'Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression', *Psychological Bulletin*, 114: 542–551.

Burnham, K.P and Anderson, D.R. (2002) *Model Selection in Multimodel Inference: A Practical Informational-Theoretical Approach* (2nd edn.). New York: Springer.

Chatterjee, S., Hadi, A., and Price, B. (2000) *Regression Analysis by Example* (3rd edn.). Wiley.

Cohen, J., Cohen, P., West, S., and Aiken, L. (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edn.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortina, J.M. (1993) 'Interaction, nonlinearity, and multicollinearity: implications for multiple regression', *Journal of Management,* 19: 915–922.

Cotter, K.L. and Raju, N.S. (1982) 'An evaluation of formula-based population squared cross-validity estimates and factor score estimates in prediction', *Educational and Psychological Measurement,* 42: 493–519.

Courville, T. and Thompson, B. (2001) 'Use of structure coefficients in published multiple regression articles: 0 is not enough', *Educational and Psychological Measurement,* 61: 229–248.

Darlington, R,B. (1968) 'Multiple regression in psychological research and practice', *Psychological Bulletin,* 69: 161–182.

Diaconis, P. and Efron, B. (1983) 'Computer-intensive methods in statistics', *Scientific American,* 248: 116–130.

Draper, N.R. and Smith, H. (1998) *Applied Regression Analysis* (3rd edn.). New York: Wiley.

Dunlap, W.P. and Landis, R.S. (1998) 'Interpretations of multiple regression borrowed from factor analysis and canonical correlation', *Journal of General Psychology,* 125: 397–407.

Ganzach, Y. (1997) 'Misleading interaction and curvilinear terms', *Psychological Methods,* 2: 235–247.

Graybill, F.A. (1976) *Theory and Application of the Linear Model.* North Scituate, MA: Duxbury Press.

Greenland, S., Schelessman, J.J., and Criqui, M.H. (1986) 'The fallacy of employing standardized regression coefficients and correlations as measures of effect', *American Journal of Epidemiology,* 123: 203–208.

Hedges, L.V. and Olkin, I. (1981) 'The asymptotic distribution of commonality components', *Psychometrika,* 46: 331–336.

Herzberg, P.A. (1969) 'The parameters of cross-validation', *Psychometrika,* 34: 1–68.

Joe, G.W. and Mendoza, J.L. (1989) 'The internal correlation: its applications in statistics and psychometrics', *Journal of Educational Statistics,* 14: 211–226.

Johnson, R.A. and Wichern, D.W. (2002) *Applied Multivariate Statistical Analysis* (5th edn.). Upper Saddle River, NJ: Prentice Hall.

King, G. (1986) 'How not to lie with statistics: avoiding common mistakes in quantitative political sciences', *American Journal of Political Sciences,* 30: 666–687.

Kleinbaum, D.G., Kupper, L.L., Mueler, K.E., and Nizam, A. (2007) *Applied regression analysis and other multivariable techniques* (3rd edn.). North Scituate, MA: Duxbury Press.

Kruskal, W. (1987) 'Relative importance by averaging over orderings', *The American Statistician,* 41: 6–10.

Kruskal, W. and Majors, R. (1989) 'Concepts of relative importance in recent scientific literature', *The American Statistician,* 43: 2–6.

Lawrance, A.J. (1976) 'On conditional and partial correlation', *The American Statistician,* 30: 146–149.

Lindeman, R.H., Merenda, P.F., and Gold, R.Z. (1980) *Introduction to Bivariate and Multivariate Analysis.* Glenview, IL: Scott, Foresman.

Lubinski, D. and Humphreys, L.G. (1990) 'Assessing spurious moderator effects: illustrated substantively with the hypothesized (synergistic) relation between spatial visualization and mathematical ability', *Psychological Bulletin,* 107: 385–393.

Mallows, C.L. (1973) 'Some comments on Cp', *Technometrics,* 15: 661–675.

Maxwell, S.E. and Delaney, H.D. (2004) *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (2nd edn.). Mahwah, NJ: Lawrence Earlbaum Associates.

Mayeske, G.W., Wisler, C.E., Beaton, A.E., Weinfeld, F.D., Cohen, W. M., Okada, T. et al. (1969) *A Study of Our Nation's Schools.* Washington, DC: US Department of Health, Education, and Welfare, Office of Education.

Miller, A. (1990) *Subset Selection in Regression.* London: Chapman and Hall.

Mood, A.M. (1969) 'Macro-analysis of the American educational system', *Operations Research,* 17: 770–784.

Mood, A.M. (1971) 'Partitioning variance in multiple regression analyses as a tool for developing learning models', *American Educational Research Journal,* 8: 191–202.

Mooney, C.Z. and Duval, R.D. (1993) *Bootstrapping: A Nonparametric Approach to Statistical Inference.* Newbury Park, CA: Sage Publications.

Mosteller, F. and Tukey, J.W. (1977) *Data Analysis and Regression: A Second Course in Statistics.* Reading, MA: Addison-Wesley.

Neter, J., Kutner, M.H., Nachtshien, C.J. and Wasserman, W. (1996) *Applied Linear Statistical Models* (4th edn). Chicago: Irwin.

Newton, R.G. and Spurrell, D.J. (1967a) 'A development of multiple regression for the analysis of routine data', *Applied Statistics,* 16: 51–64.

Newton, R.G. and Spurrell, D.J. (1967b) 'Examples of the use of elements for clarifying regression analysis', *Applied Statistics,* 16: 165–172.

Olkin, I. and Pratt, J.W. (1958) 'Unbiased estimation of certain correlation coefficients', *The Annals of Mathematical Statistics,* 29: 201–211.

Pedhazur, E.J. (1975) 'Analytic methods in studies of educational effects', in Kerlinger, F.N. (ed.), *Review of Research in Education* 3. Itasca, IL: Peacock.

Pedhazur, E.J. (1997) *Multiple Regression in Behavioral Research: Explanation and Prediction* (3rd edn.). Orlando, FL: Harcourt Brace.

Pratt, J.W. (1987) Dividing the indivisible: Using simple symmetry to partition variance explained in T. Pukilla and S. Duntaneu (eds.), *Proceedings of the Second Tampere Conference in Statistics.* University of Tampere, Finland. pp. 245–260.

Sampson, A.R. (1974) 'A tale of two regressions', *Journal of American Statistical Association,* 69: 682–689.

Schwarz, G. (1978) 'Estimating the dimension of a model', *Annals of Statistics,* 6: 461–464.

Soofi, E.S. and Retzer, J.J. (2002) 'Information indices: unification and applications', *Journal of Econometrics,* 107: 17–40.

Soofi, E.S., Retzer, J.J., and Yasai-Ardekani, M. (2000) 'A framework for measuring the importance of variables with applications to management research and decision models', *Decision Sciences,* 31: 595–625.

Stanton, J.M. (2001) 'Galton, Pearson, and the peas: a brief history of linear regression for statistics instructors', *Journal of Statistics Education,* 9. Online. Available: http://www.amstat.org/publications/jse/v9n3/stanton.html

Stigler, S.M. (1997) 'Regression towards the mean, historically considered', *Statistical Methods in Medical Research,* 6: 103–114.

Stein, C. (1960) 'Multiple regression', in Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. and

Mann, H.B. (eds.), *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling.* Stanford, CA: Stanford University Press. pp. 424–443.

Stine, R. (1989) 'An introduction to bootstrap methods', *Sociological Methods and Research,* 18: 243–291.

Suh, E., Diener, E., Oishi, S., and Triandis, H.C. (1998) 'The shifting basis of life satisfaction judgments across cultures: emotions versus norms', *Journal of Personality and Social Psychology,* 74: 482–493.

Sue Doe Nihm (Pseudonym) (1976) 'Polynomial law of sensation', *American Psychologist,* 31: 808–809 (a satire by Michael Birnbaum).

Theil, H. (1987) 'How many bits of information does an independent variable yield in a multiple regression?', *Statistics and Probability Letters,* 6: 107–108.

Theil, H. and Chung: C-F. (1988) 'Information-theoretic measures of fit for univariate and multivariate linear regressions', *The American Statistician,* 42: 249–252.

Timm, N.H. (2002) *Applied Multivariate Analysis.* New York: Springer-Verlag.

Thomas, D.R., Hughes, E., and Zumbo, B.D. (1998) 'On variable importance in linear regression', *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-life Measurement,* 45: 253–275.

Thomas, D.R. and Zumbo, B.D. (1996) 'Using a measure of variable importance to investigate the standardization of discriminant coefficients', *Journal of Educational and Behavioral Statistics,* 21: 110–130.

Tzelgov, J. and Henik, A. (1991) 'Suppression situations in psychological research: Definitions, implications and applications', *Psychological Bulletin,* 109: 524–536.

Tzelgov, J. and Stern. I. (1978) 'Relationships between variables in three variable linear regression and the concept of suppressor', *Educational and Psychological Measurement,* 38: 325–335.