23
Multiple Regression

Ken Kelley and Scott E. Maxwell

Multiple regression has been described as a general data analytic system (e.g., Cohen, 1968), primarily because many commonly used statistical models can be regarded as its special cases (e.g., single-sample *t*-test, two-independent samples *t*-test, one-way analysis of variance), the independent variables can be categorical (e.g., groups) or quantitative (e.g., level of treatment), and the model can be used for observational or experimental studies. Furthermore, many advanced models have multiple regression as a special case (e.g., path analysis, structural equation modeling, multilevel models, analysis of covariance). The ubiquity of multiple regression makes this model one of the most important and widely used statistical methods in social science research. In general, the idea of the multiple regression model is to relate a set of *regressor* (*independent* or *predictor*) variables to a *criterion* (*dependent* or *outcome*) variable, for purposes of explanation and/or prediction, with an equation linear in its parameters. More formally, the population multiple regression model is given as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i, \tag{1}$$

where β_0 is the population intercept, β_k is the population regression coefficient for the *k*th regressor ($k = 1, \ldots, K$), X_{ki} is the *k*th regressor for the *i*th individual ($i = 1, \ldots, N$), and is the error for the *i*th individual, generally assumed to be normally distributed with mean 0 and population variance σ_ε^2. The intercept is the model-implied expected value of *Y* when each of the *K X* variables are at values of zero. The intercept may have a meaningful substantive interpretation, such as when the regressor variables are centered around 0 so that the intercept represents the grand mean on the outcome or when the regressor variables are dummy variables and the intercept thus represents the expected value of the outcome for the referent group, otherwise it serves as a scalar so that the sum of the squared errors can be minimized. For contemporary treatments of multiple regression applied to a wide variety of examples, we recommend Cohen, Cohen, West, and Aiken (2003), Pedhazur (1997), Harrell (2001), Fox (2008), Rencher and Schaalje (2008), Gelman and Hill (2007), and Muller and Fetterman (2002). Specific desiderata for applied studies that utilize multiple regression are presented in Table 23.1 and explicated subsequently.

Table 23.1 Desiderata for Multiple Regression.

Desideratum	*Manuscript Section*
1. The goals of the research and how multiple regression (MR) can be useful are explicitly addressed.	I
2. The inclusion of each of the independent variables, whether confirmatory or exploratory in nature, should be justified on theoretical and/or practical grounds.	I
3. Each criterion and regressor variable should be described in detail (e.g., scales of measurement, coding scheme, reliability) to convey how the MR model should be interpreted.	M
4. Specific procedures for the computation and interpretation of effect sizes are delineated.	M
5. Assumptions underlying the MR analyses and resulting inference are explicitly addressed.	M
6. Variable selection techniques are justified.	M
7. Sample sizes for all analyses are justified in terms of power, accuracy, and reproducibility of results.	M
8. Methods of dealing with missing data are addressed.	M
9. For models examining moderation, issues of interpretation, role of centering, and visualization are addressed.	R
10. For models examining mediation, issues of interpretation and limitations due to cross sectional designs are addressed.	R
11. Visual examination of data is addressed in order to assess model appropriateness and assumptions.	R
12. Measurement error in predictor and/or outcome variables is addressed.	D
13. Potential limitations of multiple regression in the current applied research context are explicitly stated.	D
14. Alternatives to the MR model are given.	D

* I = Introduction, M = Methods, R = Results, D = Discussion.

1. Research Goals

Standard textbook treatments of multiple regression often emphasize that multiple regression can be used for prediction or explanation. Depending on the goals of the researcher, prediction, explanation, or both might be desired. Although the multiple regression model itself is exactly the same in both cases (i.e., Equation (1) does not change based on the goal), the distinction is nevertheless important because different statistical considerations arise for the two purposes. To clearly communicate the purpose of the study, it is important for authors to be clear about whether their purpose in using multiple regression is prediction, explanation, or both.

The ultimate goal of explanation is to identify the causes of the outcome variable Y. Under ideal conditions, multiple regression can identify causal effects by assessing the value of regression coefficients: when the coefficients are non-zero in the population causality may be a possibility. To understand how a regression coefficient can potentially reflect a causal effect, we need to say what a regression coefficient represents. For example, when the model is correctly specified, the coefficient β_k for X_k reflects the linear relation between Y and X_k at a fixed value of all other regressors included in the model. In this sense the regression coefficient for X_k is a measure of the extent to which X_k and Y are linearly related when all other regressors in the model are held constant. Because the other regressors are held constant, any association between X_k and Y cannot be attributed to the other regressors. Thus, it is tempting to conclude that β_k reflects the extent to which X_k *causes* Y, in which case we have at least partly succeeded in explaining variation in Y. In fact, this reasoning is sometimes correct, but only under a set of restrictive conditions (temporal precedence, relationship,

and nonspuriousness; see, e.g., Kenny, 1979). Unfortunately, it can often be difficult to justify these conditions unequivocally except in randomized experiments.

Predicting the value of a criterion variable given one or more regressors is another reason why multiple regression is commonly used, especially in applied research. For example, a researcher might use multiple regression to predict how well pre-kindergarten children will be able to read at the end of first grade. The researcher would use historical data (often called *training data*) containing scores on reading at the end of first grade as well as scores on a number of possible regressors. Multiple regression could then be used to create a model in which the value of the criterion is predicted based on one or more of the regressors. A benefit of prediction is that the parameter estimates (i.e., the regression coefficients) obtained from the training data can be used to predict the value of an unknown (or yet to occur) criterion variable Y based on the complete set of regressors used in the training data. There are many cases in which it is desirable to predict a criterion variable when it is as yet unknown (e.g., college grade point average or reading ability at the end of first grade) from a set of known regressors (e.g., SAT scores or pre-kindergarten measures of cognitive functioning). The ultimate goal is often selection, as in the college example, but can also be identifying at-risk individuals who might benefit from a relevant intervention.

Although we believe that recognizing the difference between explanation and prediction is critically important when considering the parameters of interest in the model, there need not be such a rigid dichotomy between the two goals. In studies seeking to explain relations there can be prediction, and in studies that seek a way to predict there can be attempts at explanation. Pedhazur (1997, p. 196) described predictive research having as its main emphasis "practical applications," whereas in explanatory research the main emphasis is "understanding phenomena." Huberty (2003) provided a discussion of the similarities and differences in research goals and reporting strategies when interest is primarily in prediction or explanation.

Statistical inference is important when a desire exists to generalize information obtained in a sample to the population from which the sample was drawn. Inference can be of two forms, confidence interval formation for the population effect sizes of interest and/or hypothesis testing for effect sizes. For purely predictive purposes, inferential procedures are not strictly necessary, but nevertheless provide information about the population of interest.

2. Justification of Regressors

Multiple regression can be applied along a continuum of research approaches anchored by *confirmatory* and *exploratory* research. The confirmatory anchor seems to best correspond to a well-defined research question with a few theoretically justified variables, whereas the exploratory anchor corresponds to a diffuse research question with many variables included in one or more different analyses, not necessarily with explicit theoretical justification. Both confirmatory and exploratory analyses are beneficial, but care must be taken so that an exploratory analysis is not presented as if it were a confirmatory analysis. Provided the assumptions of the model are satisfied in the context of confirmatory studies, the probability values (i.e., the p-values) from null hypothesis significance testing and confidence interval coverages associated with the different effect sizes are meaningful. However, because exploratory analyses generally consist of systematic testing and retesting until settling on a model that is satisfactory enough, the process of testing and then retesting renders the probability values and confidence interval coverages associated with the effect sizes as approximate at best, and completely inappropriate at worst. For example, testing many models with the aim of obtaining p-values for coefficients of interest less than, say, .05, leads to p-values that are too heavily based on the characteristics of the sample rather than a test of a well-specified question. Testing null hypotheses of different models on the same data set will result in capitalization on chance and

more Type I errors will be realized than the Type I error rate specified by the researcher (e.g., .05). That is to say, when exploratory analyses are treated as if there were confirmatory, properties of the *p*-values will not be the same as if the analysis was truly confirmatory. Nevertheless, findings from such exploratory studies often provide a useful starting point for future confirmatory research but it needs to be clear to the reader how the analysis was conducted and what other analyses were attempted. Readers, such as reviewers, are rightly skeptical of drawing important conclusions from studies in which many models were evaluated and only the significant findings presented.

More formally, the reason probability values and confidence interval coverages are not correct in exploratory analyses in which multiple models are evaluated is because of what is known as the *multiplicity problem*. The multiplicity problem describes the problem of multiple statistical tests being performed, where the effect sizes with small *p*-values are selected for inclusion in the presented statistical model. An implication of the multiplicity problem is that the obtained *p*-values are suspect, due to the sheer number of null hypothesis significance tests conducted. When many null hypothesis significance tests are conducted, even when all the null hypotheses are true, there is a high probability of finding some small *p*-values by chance. Thus, because of the suspect *p*-values and the associated confidence interval coverages associated with statistical inference in exploratory studies, it should be made clear if the study was confirmatory in nature or exploratory. In particular, exploratory approaches sometimes effectively are based on an informal variation of a formal variable selection method (such as stepwise regression, to be discussed in Desideratum 6), which may be fine for prediction but raises serious concerns about the meaningfulness of any claims regarding explanation. That is, some researchers reject the idea of stepwise regression, but themselves perform a more intuitive version of stepwise regression where many models are fitted, even when their purpose is explanation.

3. Descriptions of Criterion and Regressor Variables

A statistical model in and of itself is not very useful unless the variables in the model are understood in their appropriate context and have been discussed in enough detail to convey an understanding of the information they contribute to the research question. At a minimum, means and the covariance matrix or the correlation matrix (with accompanying standard deviations) should be provided for all variables used in the analysis. Furthermore, the type of variable (e.g., categorical or continuous) and the range over which values of the scale can vary (i.e., the limits of the scale) should be discussed. When categorical variables (e.g., grouping variables) are used, the coding scheme should be explicitly discussed. Without an explanation of the coding scheme, the estimated model parameters cannot be readily interpreted by others (e.g., for the "Sex" variable females are coded as 0 and males 1, females as 1 and males 0, or females −1 and males 1, etc.). Continuous variables should almost never be dichotomized (or polytomized more generally) but should instead be left in their continuous form in order to preserve as much information in the variable as possible. Examples of situations where it may sometimes be reasonable to polytomize continuous variables is when there are clear types or taxa of individuals or when the distribution of a count variable is highly skewed (MacCallum, Zhang, Preacher, & Rucker, 2002). It is clear, however, that median splits, a commonly used procedure for dichotomizing continuous data, is essentially never statistically justified. Where appropriate, the reliability and validity evidence for each of the variables should be provided (see Desideratum 12); more information is available in Chapter 29 of this volume.

4. Effect Sizes

As has been discussed a great deal in the methodological literature, effect sizes and their corresponding confidence intervals are widely recommended and should almost always be reported (e.g.,

Wilkinson & APA Task Force on Task Force on Statistical Inference, 1999; see also Chapter 6, this volume). In multiple regression, like many other statistical models, there are two types of effect sizes: *omnibus* and *targeted*.

The most widely used omnibus effect size in multiple regression, and one of the most common in social science research in general, is the squared multiple correlation coefficient, whose population value is denoted P^2 (rho squared). The value of P^2 quantifies the proportion of variance in Y that can be accounted for by the K regressor variables. The typical estimate of P^2, R^2, is positively biased. Although confidence intervals and significance tests for P^2 are based on R^2, the adjusted value of R^2, denoted R_A^2, should also be reported and used as the best estimate of P^2. The typical adjusted estimate (e.g., Cohen et al., 2003; Harrell, 2001) is given as

$$R_A^2 = \max\left\{0, \left[1-(1-R^2)\left(\frac{N-1}{N-K-1}\right)\right]\right\}, \tag{2}$$

where max{,} implies that the larger of the two values is taken. Most statistical programs will give both R^2 and R_A^2.

Darlington (1968) explained that the adjustment shown in Equation (2) (developed by Ezekiel, 1930) will tend to overestimate the population validity of the sample regression equation. The idea here is that the adjustment estimates the population validity of the population regression equation. In other words, if the population regression coefficients were known, what proportion of the variance in Y would this equation explain in the population? This makes sense when the goal is explanation, because one purpose here is to estimate the extent to which the regressors explain the variance in Y. However, this makes less sense when the goal is prediction, because in this context the sample regression equation derived in the training sample will be used to make predictions in a new sample. The key point is that the regression coefficients to be used for prediction are the values obtained in the training sample. However, these values will not be exactly the same as the optimal population values, thus lowering the resultant R^2 to some extent. For this reason, in the context of prediction, the population parameter of most interest is sometimes referred to as the population cross-validity, P_C, or the squared population cross-validity, ρ_c^2. Raju, Bilgic, Edwards, and Fleer (1999) described a variety of estimators of the population cross-validity and recommended an adjustment developed by Burket (1964):

$$R_C = \frac{NR^2 - K}{R(N-K)}. \tag{3}$$

Although these omnibus effect size estimates are beneficial, an observed effect size is simply a point estimate that might differ considerably from the population value it estimates. Confidence intervals should be reported for any estimate that is itself deemed important enough to report. Confidence intervals for P^2 are not straightforward to construct and the appropriate confidence interval depends on whether or not regressors are regarded as fixed or random. Steiger (2004; see also Steiger & Fouladi, 1992), Algina and Olejnik (2000), and Kelley (2007), discussed methods of confidence interval construction and provided software solutions to implement such intervals.

Researchers should consider the squared semi-partial correlation coefficient, which is a targeted effect that describes the change in R^2 when the kth regressor is added to the multiple regression model that already contains the other $K - 1$ regressors. Thus, the squared semi-partial correlation coefficient quantifies the proportion of variance of Y that is accounted for *uniquely* by a particular regressor in a model with other regressors. Such an effect size is useful when conveying the contribution of a regressor in a model with $K - 1$ other regressors. Squared semi-partial correlation

coefficients can also be used to quantify the proportion of variance of Y that is accounted for by a particular set of regressors instead of just a single regressor.

Regression coefficients come in two forms: *unstandardized* and *standardized*, both of which represent targeted effects, which may or may not be causal in nature. Unstandardized regression coefficients can be transformed into standardized regression coefficients by multiplying the unstandardized regression coefficient by the quantity $\frac{s_{X_k}}{s_Y}$, which removes the scale of X_k and Y, where $s_{\cdot}$ denotes the standard deviation of the subscripted quantity. The process can be reversed (i.e., set a standardized regression coefficient on the unstandardized scale) by multiplying a standardized regression coefficient by $\frac{s_Y}{s_{X_k}}$. In general, either unstandardized or both unstandardized and standardized regression coefficients should be given, along with their corresponding confidence intervals. The kth regression coefficient quantifies the degree of linear relation between Y and X_k, while holding constant the remaining $K - 1$ regressors. Standardized regression coefficients are often an effective way of describing the effect of a regressor on the criterion variable when the scales of the measurements are not inherently meaningful. When standardized solutions are used in place of or in addition to their unstandardized counterparts, the measure of association is in terms of standard deviation units of the particular sample. For example, a standardized regression coefficient of .25 for X_k in a standardized solution implies that a 1 standard deviation unit difference in X_k is associated with a .25 standard deviation difference in Y in the same direction, holding constant all other regressors.

Confidence intervals for unstandardized regression coefficients are easy to obtain and formulas are available in essentially all modern regression books and can also be obtained with popular statistical software. However, confidence intervals for standardized regression coefficients require the use of noncentral t distributions and are more difficult to obtain (e.g., see Kelley & Maxwell, 2008, or Kelley, 2007, for a review and software solutions). In general, standardized regression coefficients are provided when there is a desire to remove the scaling of the measurement instrument so that each variable (regressors and criterion) has a mean of 0 and a standard deviation of 1. Standardized regression coefficients allow for relations to be framed in standard deviation units (as previously noted) and regression coefficients to be more directly comparable within an equation. That being said, there is no guarantee that the regressor with the largest regression coefficient is the "most important" independent variable in the equation (even when all variables are standardized). The meaning of "most important" might be different depending on the particular situation and goals of the study (Azen & Budescu, 2003).

5. Addressing Assumptions

Standard approaches to regression rely on ordinary least squares (OLS) to estimate model parameters. The OLS regression coefficients in multiple regression minimize the sum of squared deviations between the model implied scores, denoted $\widehat{Y}_i$ for the ith individual, and the observed scores (i.e., regression coefficients are chosen that minimize $\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2$). Estimation of the regression coefficients themselves does do not strictly require any parametric assumptions. However, inference for the regression coefficients in the usual ways (hypothesis testing and confidence interval formation) does depend on validity of underlying assumptions. In particular, p-values and confidence intervals (i.e., inference) for regression coefficients from the regression model as specified in Equation (1) depend on four statistical assumptions: (a) errors (i.e., $e_i = Y_i - \widehat{Y}_i$) follow a normal distribution; (b) error variance is homogeneous across all values of the regressors (*homoscedasticity*); (c) the entities (e.g., persons) from which observations are taken are independent of one another; and

(d) the relation between Y and the K regressors is linear. It is important to note that no distributional assumptions are made about the regressors, meaning that, for example, skewness in a predictor is not by itself a problem. Also, the model does not assume that regressors are measured without error, but as we will discuss later, results obtained using regressors measured with error may differ substantially from results obtained when regressors are measured perfectly, so measurement error in the regressors often becomes an important consideration.

Although the linearity assumption (assumption d above) is fundamental, it is often overlooked in discussions and applications of multiple regression. We agree with Gelman and Hill (2007, p. 46) that, "The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors." This assumption is especially important because if this assumption is not valid, then the regression coefficients in the model do not accurately reflect the relation between Y and X_k at a fixed value of the other regressors. As a result, the regression model might fail to hold the other regressors constant in attempting to estimate the relation between Y and a specific X variable. If linearity does not hold, then the model as specified in Equation (1) may not be appropriate for inferences, as Equation (1) is necessarily linear in form. When linearity does not hold, there are essentially three strategies: (a) transform one or more variables (one or more X_k and/or Y) so that linearity in an additive model is a good approximation (e.g., $\sqrt{X_k}$ or X_k^2), (b) include an additional theoretically justified variable (e.g., X_k^2 in addition to X_k) that correlates with the outcome variable, in an attempt to explain some of the unaccounted for variability, and/or (c) fit a nonlinear regression model (e.g., a negative exponential, Gompertz, logistic) instead of the traditional linear multiple regression model (Seber & Wild, 1989).

6. Variable Selection Techniques Are Formally Justified

In many situations, more regressor variables are initially included in a model than are ultimately desirable in the final model to be presented for interpretation. The way in which the researcher arrives at the final model should be made explicit. There are four common ways of selecting variables to be included in the analysis: (a) all analyses are theory driven, (b) model comparisons are performed, (c) stepwise methods are used, or (d) a variety of exploratory models and methods are fitted.

In many ways, the ideal variable selection method is entirely theory driven and the regressors included are based on a priori theoretical arguments and/or previous literature. This method is ideal because a one-to-one mapping exists between the targeted nature of the research question and the targeted statistical analyses.

A model comparison approach (e.g., Maxwell, Delaney, & Kelley, 2018), in which the inclusion of one or more variables is evaluated against a more basic model, is often the most straightforward way to evaluate competing nested models (on the same set of data). The idea of the model comparison approach is to statistically compare nested models, where the models are compared most commonly in terms of R_k^2 and R_{K+M}^2, where R_k^2 is the model based on the K regressors and R_k^2 is based on a richer model with an additional M regressors.

A special type of model comparison is implemented through what is often termed *hierarchical regression* (not to be confused with hierarchical linear modeling, HLM; see Chapter 22, this volume). In hierarchical regression, not only are the variables selected by the researcher, so too is the order in which they enter the model. At each step of the procedure, the variables previously included remain in the analysis. When hierarchical regressions are performed, a series of fitted models should be provided as part of the reported results that shows the estimated model improvement when comparing the richer models to the simpler models. The improvement is generally gauged in terms of the change in R^2 when a single regressor variable is added, which again is the squared semipartial correlation coefficient. It is also common to add a block of regressors in a hierarchical fashion.

In such situations the change in R^2 is still of interest, but there the additional variability accounted for is due to the block of regressors. For example, a researcher might add a block of control variables before adding one or more primary variables of interest.

When a large number of possible regressors exist, possibly for more than one criterion variable, data driven selection methods are sometimes used. Whenever data driven selection methods are used, a clear indication should be made that the study is not attempting to explain phenomena in a confirmatory fashion, but rather that the study is exploratory in nature. The type of data driven selection procedure performed (e.g., forward selection, backward elimination, all possible subsets), and the selection criteria (e.g., a statistically significant change in R^2, or a change in R^2 of some specified magnitude, say .05) should be given. Also the particular computer program/package and its version should be provided, because different programs/packages and versions implement data driven selection procedures in different ways.

There are many methodological issues that can arise when implementing a data driven selection procedure. As Rencher and Pun (1980) illustrated, values of R^2 can be highly inflated and thus the obtained probability values can differ substantially from those reported as output in statistical software. When a large number of possible regressors exist in the context of a data driven selection procedure, a model that accounts for a statistically significant proportion of variance in Y can often be obtained even if the null hypothesis is true that all of the regression coefficients, less the intercept, are zero. Because of the multiplicity issue, as previously noted, fitting more than a single model can inflate the Type I error rate due to capitalization on chance.

Vittinghoff, Glidden, Shiboski, and McCulloch (2005) provided an especially interesting perspective on model building by distinguishing three different purposes for selecting predictors: (1) evaluating a regressor of primary interest in the context of other possibly relevant regressors, (2) identifying the important regressors of an outcome, and (3) prediction. They emphasized that issues involved in predictor selection differ according to the purpose of the analysis. For example, suppose that two regressors X_1 and X_2 are highly correlated with one another. When the goal is prediction, it will generally be desirable to include only one of these two regressors in the model, and it may make little difference in the accuracy of prediction which of the two is included. Ironically, however, including both of the regressors will often worsen prediction because any gain in bias reduction is more than offset by an increase in the variance of predicted values. On the other hand, suppose the goal is to explain the relation between X_1 and Y. Should X_2 be controlled for and thus included in the model? We agree with Vittinghoff et al. (2005) that this question cannot be answered simply from knowing that X_1 and X_2 are highly correlated. Instead, for explanatory models it becomes necessary to consider a theoretical causal model for how the various regressors and Y relate to one another. In particular, X_2 should be included in the model if it is a confounder, but not all variables highly correlated with the regressor of primary interest (i.e., X_1) are necessarily confounders. Vittinghoff et al. (2005), Jaccard, Guilamo-Ramos, Johansson, and Bouris (2006), and Hernan, Hernandez-Diaz, Werler, and Mitchell (2002) discussed various approaches for identifying whether a variable is a confounder and thus should be included in the regression model.

7. Sample Sizes Are Justified

Sample size is an important component to any research study. "Rules of thumb" that were once widely recommended for planning sample size are not generally appropriate and should not be used as justification (see Green, 1991, for a review). Instead, researchers should justify their sample size. A common approach to sample size planning is the power analytic perspective. However, another perspective is accuracy in parameter estimation (AIPE). The goal of the power analytic approach is to plan sample size so that a false null hypothesis can be rejected with some desired probability

(i.e., power), whereas the goal of the AIPE approach is to obtain an accurate estimate of the population value, which is operationalized by a sufficiently narrow confidence interval with some desired degree of assurance (i.e., probability). In addition to deciding on whether power or AIPE is most appropriate, researchers also need to state whether the primary interest is in an omnibus effect (i.e., the squared multiple correlation coefficient) or one or more targeted effects (i.e., regression coefficients), which is necessarily based on the question(s) of interest. In particular, questions of prediction are more likely to involve omnibus effects, whereas questions of explanation are more likely to involve targeted effects. Additional details are provided in Kelley and Maxwell (2008), who discussed sample size planning methods in a multiple regression context in a 2×2 (power or AIPE $\times$ omnibus or targeted effect) framework.

In some cases, existing/archival data become available to a researcher. Because the data have already been collected, sample size planning cannot be done as previously discussed, as it is implemented a priori in the design phase of the study. In general, power and AIPE are not often discussed for existing/archival data. However, power and AIPE can still be addressed, albeit in a different manner. In particular, for a specified value of an effect size at the size of the sample in the existing data, power and expected confidence interval width can be given. An appropriate value for the effect size to use is what can be termed the parameter of minimal importance (POMI) or the minimum parameter value of interest (MPVI), both of which represent the smallest magnitude that is deemed to have scientific, clinical managerial, or practical importance/interest in the particular context.

8. Missing Data

Missing data is a perplexing issue. There are three broad categories of missingness: (a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR). MCAR is when missingness does not depend on either observed or missing values, whereas MAR is when missingness does not depend on the missing values but may depend on observed values. MNAR implies that missingness depends on an outside variable not in the model or depends on the value of the variable itself (see Little & Rubin, 2002, for a types of missing data and appropriate methods for dealing with the different types of missing data).

Although the specifics of the situation will differ, researchers should do their best to ensure the amount of missing data is minimized (e.g., remind participants about follow-up visits, check evaluations for blank responses before the participants leave, clearly state that sensitive data will remain confidential if appropriate). Generally, whenever missing data arises in a research study, it opens the possibility for criticism in the way it was (or was not) dealt with. Whenever there is a nontrivial amount of missingness, the data should be interrogated for patterns of missingness (Harrell, 2001). When apparent patterns are found, they should be reported and, if possible, a plausible explanation provided with a cautionary reminder given that exploratory methods were used to uncover any apparent patterns in the data. Regardless of the way in which missing data is dealt with, the method and the rationale for choosing the method should be discussed. That being said, some methods, in particular mean substitution and/or pairwise deletion, should not be used unless there is a good reason to do so with a clear explanation of why. We will briefly discuss three methods of dealing with missing data (see Schafer & Graham, 2002, for a thorough review).

When missing data does occur, casewise deletion appears to be frequently employed in the applied literature; however, casewise deletion can be problematic. In multiple regression, casewise deletion and listwise deletion are equivalent, however, in other models the two terms differ. Casewise deletion is when a participant is completely excluded, regardless of the amount of data available for the participant, if any data are missing for the analysis of a particular model. Listwise deletion is when an entire row is removed when there is any missing data. Thus, for models in which each case has

only one row in a data set, casewise and listwise deletion are equivalent. However, for some models a single case (e.g., person) will have multiple rows for different measurement occasions. There, such as in multilevel models, the row but not the case itself is deleted. Casewise (or listwise) deletion generally yields unbiased estimates only under the very strong assumption that data are MCAR. At best, estimates obtained using casewise deletion are inefficient, implying less statistical power and estimation accuracy than would otherwise be the case. The reason casewise deletion is inefficient is because the sample size is reduced to only those with complete data sets, which tends to increase the sample standard error(s) and necessarily does so in the population. More important, however, is that estimates obtained using casewise deletion will often be biased, unless plausible arguments can be advanced for why missingness is likely to be MCAR.

Imputation or multiple imputation provides a reasonable way to deal with missing data in many situations. Imputation is when a plausible value is substituted for a missing value and multiple imputation is when this process is performed multiple times. The "plausible values" come from an imputation model that uses other data that are available to estimate the data that are not available. At first the idea of estimating data might seem problematic, but it is often better to estimate what is usually a small amount of data than to disregard valuable data with deletion (e.g., casewise) strategies (Harrell, 2001, §3.4).

Full information maximum likelihood (FIML) and restricted maximum likelihood (REML) estimation are the most popular methods for dealing with missing data in multilevel models and structural equation models, likely because main-stream multilevel model and structural equation modeling programs can easily implement them (and usually do so by default). These maximum likelihood methods for dealing with missing data assume that data are MCAR or MAR. Because FIML does not consider the degrees of freedom and uses the standard normal distribution instead of the *t*-distribution, sample size should not be small with this approach. Small sample sizes being used with the FIML approach to missing data will tend to yield differences in the empirical and nominal Type I error rates. REML, however, does consider the issue of degrees of freedom and is more appropriate in smaller samples. Another issue is that maximum likelihood estimation assumes multivariate normality, which might not always be reasonable (recall that the standard multiple regression assumption is only that the errors are normally distributed). Enders (2001) provided a review and evaluation of maximum likelihood estimation when missing data exists in the context of multiple regression. Our recommendation is to use either multiple imputation or maximum likelihood estimation when faced with missing data.

9. Models Examining Moderation

The regression model shown in Equation (1) assumes that the effects of each X_k on Y are additive. For example, with two regressors, this model assumes that the relation between X_1 and Y is the same for every value of X_2 and similarly the relation between X_2 and Y is the same for every value of X_1. In reality, however, the strength of the relation (or even the direction of the relation) between X_1 and Y might depend on X_2, in which case X_1 and X_2 are said to *interact*. As a consequence, the regression model shown in Equation (1) might seem very restrictive, because it does not seem to allow for the possibility of an interaction between X_1 and X_2. Fortunately, this restriction is illusory, because modifications to the model allow X_1 and X_2 to interact. The ability to modify this model is critical because many theories in the social and behavioral sciences stipulate that the relation between a pair of values (e.g., Y and X_1) depends on a third variable (e.g., X_2), which corresponds to an interaction effect.

The standard way of modifying the model in Equation (1) so as to allow for the possibility of an interaction (or equivalently, a moderator) is to add cross-product terms. For example, with two regressors, the model becomes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i \tag{4}$$

The inclusion of the product term allows the relation between either X and Y to depend on the value of the other X. In particular, this model stipulates that the slope relating X_1 to Y is given by

$$\frac{dY}{dX_1} = \beta_1 + \beta_3 X_2, \tag{5}$$

where dY/dX_1 is the derivative (instantaneous slope) of Y with respect to X_1. If β_3 is non-zero, the relation between X_1 and Y depends on X_2, so X_2 moderates the effect of X_1 on Y, or equivalently, X_1 and X_2 interact. However, many researchers might not realize that the product term represents a very specific type of interaction, namely a *bilinear* effect. In particular, Equation (5) shows that if β_3 is positive, the slope becomes increasingly higher for larger values of X_2. Similarly, if β_3 is negative, the slope becomes increasingly lower for larger values of X_2. Thus, researchers should consider whether this is the type of interaction they truly desire to detect. If not, more complicated models can be constructed, such as including quadratic terms for some or all regressors. Interested readers can consult Cohen et al. (2003) for additional details.

The best way to begin to interpret effects in moderator models is generally to plot the interaction. For example, suppose the primary interest involves the extent to which X_2 moderates the relation between X_1 and Y. Cohen et al. (2003) recommended plotting regression lines relating Y and X_1 at three values of X_2 (typically at the mean of X_2 and also at scores one standard deviation below the mean and one standard deviation above the mean). We recommend that such a plot be included in a published work involving moderator effects. Alternatively, what can be helpful is a three-dimensional representation of the relations, where Y is plotted as a function of all possible scores on X_1 and X_2 within an appropriate range.

A point of some confusion historically has been how to interpret the β_1 and β_2 coefficients in the model in Equation (4). Some researchers have interpreted these coefficients as if they corresponded to main effects, but this is not generally true. Instead, they are conditional (i.e., simple) effects. For example, Equation (5) shows that β_1 is the slope of Y on X_1 when X_2 equals 0. Unless the range of values of X_2 happens to include 0, the conditional effect in the interaction model will be meaningless. For this reason, it is often recommended that X_1 and X_2 be recoded so that a value of 0 takes on a meaningful interpretation. Most commonly, both variables are centered by subtracting the sample mean from all scores (*mean-centering*), yielding a new coding with a mean of 0. One could subtract a theoretically meaningful value from the scores. In any event, it is critical that authors explain how regressors in interaction models have been coded, in order to facilitate interpretation of the corresponding regression coefficients.

Because of perceived complications of interpreting interactions between continuous regressors, some researchers decide to simplify analyses by categorizing either or both regressors. We strongly recommend that researchers avoid the temptation to categorize continuous variables. One reason to leave variables as continuous is that categorization can decrease power. Interestingly, Maxwell and Delaney (1993) have also shown that in some situations categorization can have the opposite effect of producing spurious effects, thus inflating the Type I error rate. Thus, statistically significant interaction effects based on artificially categorized variables cannot necessarily be trusted, strengthening the argument for leaving continuous variables as continuous.

Researchers should also be aware that several other factors affect the ability to detect interactions in regression models. First, when X_1 and X_2 are measured with error, the product term X_1X_2 will generally be much less reliable than either X_1 or X_2, which tends to lower the power to detect an interaction. Researchers who use regression to investigate interactions need to consider carefully

the reliability of regressors. Second, McClelland and Judd (1993) showed that the distribution of regressors in observational studies will often reduce power, especially when regressors correlate substantially with one another. Third, Lubinski and Humphreys (1990) showed that when regressors correlate substantially with one another, the Type I error for testing an interaction can be badly inflated if curvilinear effects exist but are not included in the regression model. Including higher order effects such as X_1^2 and X_2^2 can guard against spurious interaction effects, but also runs the risk of greatly lowering power to detect true interaction effects. There is no clear consensus among methodologists at this point about how best to resolve this dilemma. At the very least authors who want to investigate interactions in regression models should be clear about the extent to which their regressors correlate with one another as well as the extent to which theoretical considerations either do or do not rule out possible curvilinear effects. Given the scope of the topic of interactions, we recommend that readers consult such sources as Aiken and West (1991) and Jaccard and Turrisi (2003) for further information, as well as Chapter 18 in this volume.

10. Models Examining Mediation

Baron and Kenny (1986) clarified the distinction between moderation and mediation. Both involve a role that X_2 (for example) may play in the relation between X_1 and Y, leading some researchers to confuse moderation and mediation. Thus, it is incumbent on authors of papers reporting either moderation or mediation to provide a clear theoretical rationale for their study.

The variable X_2 mediates the relation between X_1 and Y when X_1 causes X_2 and X_2 in turn causes Y. Thus, mediation can be represented by a pair of regression models:

$$X_{2i} = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^* \tag{6}$$

$$Y_i = \beta_0 + \beta_2 X_{1i} + \beta_3 X_{2i} + \varepsilon_i\,, \tag{7}$$

where the asterisk represents values from the model where X_2 is the dependent variable with X_1 as its regressor. From this perspective, X_2 is a mediator when both β_1 and β_3 are non-zero. In the special case where β_2 equals 0, X_2 is said to completely (or fully) mediate the relation between X_1 and Y; otherwise, X_2 partially mediates the relation.

Baron and Kenny (1986) suggested a four-step procedure for establishing mediation. Subsequent research has studied their approach as well as a variety of alternatives. This is an area of continuing methodological research, and at this point either of two different approaches seems advisable for establishing mediation. One approach involves bootstrap methods (Shrout & Bolger, 2002). The other involves the distribution of the product variable $\beta_1\beta_3$ (MacKinnon, Lockwood, & Williams, 2004). We recommend that authors use either of these two methods to test mediation. Authors should also report coefficients and corresponding confidence intervals for relevant parameters as shown in Equations (6) and (7).

Several other factors should be considered in a mediation analysis. First, it is well known that error of measurement in the mediator causes biased estimates of regression coefficients. In three-variable models such as those in Equations (6) and (7), random measurement error will tend to result in an underestimate of the mediated effect and an overestimate of the direct effect of X_1 on Y. Researchers should address this likely bias in any interpretation of their results unless the mediator is measured without error. Alternatively, a latent variable model might be used in order to address measurement error and its biasing effects. Second, Maxwell and Cole (2007) have shown that cross-sectional estimates of mediation can be seriously biased when mediation occurs over time. Researchers who rely on cross-sectional analyses need to interpret their results with appropriate caution, and should be

encouraged to consider longitudinal designs instead of cross-sectional designs. Third, researchers should carefully consider necessary sample size to obtain adequate power. Fritz and MacKinnon (2007) provided useful guidelines. Fourth, further information about mediation, especially for more complicated models with more than three variables, is available in Chapter 18 of this volume and in MacKinnon, Fairchild, and Fritz (2007) and MacKinnon (2008).

11. Checking Assumptions Visually

The assumptions of the multiple regression model should be considered and evaluated whenever the model is used. As Anscombe (1973) noted, graphs can help researchers appreciate broad features of data *and* look beyond broad features to literally see potentially unexpected relationships, outliers, and violations of assumptions, et cetera. Anscombe went on to show four very different figures, three of which have gross violations of multiple regression assumptions, yet where the results from the regression model were the same (i.e., estimates, *p*-values, confidence intervals, etc.). Recall that the linearity assumption is that the expected value of Y given the K regressors is a linear function of the K variables. We recommend a *conditioning plot* (also referred to as a *coplot*) for examining the critical assumption of linearity. Another useful set of plots for this purpose are *residual versus predictor* (RVP) and *component plus residual* (CPR) plots. One way to evaluate violations of this assumption for "obvious" violations is by plotting the residuals as a function of the model implied values. An obvious nonlinear relationship is evidence that the linearity assumption does not likely hold. When such is the case, there might be an important variable not included in the model, an interaction term might be appropriate, or the relation between the K regressors and the criterion might be nonlinear in nature. As previously noted, the latter, in our opinion, is not considered frequently enough, and correspondingly nonlinear models are not applied in many areas as often as we believe that they should be, based on theory and empirical evidence. For example, sigmoidal forms or asymptotic values cannot adequately be modeled with linear models. We suggest readers consult Seber and Wild (1989) for a discussion of nonlinear regression models.

Recall that the errors in a multiple regression model fitted with ordinary least squares are assumed to be normally distributed for the validity of the significance test and confidence intervals. A normal-quantile–empirical-quantile plot (generally termed a *qq-plot*) is a two-dimensional plot where theoretical quantiles from the normal distribution are compared to the empirical quantiles of the observed errors. The qq-plot allows a visual evaluation of the assumption of normality of the errors. Gross violations of the normality assumption of the errors can often easily be seen with the use of a qq-plot. Although there are formal statistical tests to evaluate normality, visual displays are often extremely effective at identifying potential problems and are often easier to implement and interpret.

Matrix scatterplots (sometimes called *pairs plots*) are helpful to examine the bivariate relations among the $K + 1$ variables. These plots can also reveal observations that might be miscoded or identify potential outliers. Further, those cases that might not be considered outliers on either of two variables individually might be an outlier in a bivariate sense (which could heavily influence estimation and inference). For example, if there is a strong positive relation between X_1 and Y, yet one observation has a very low X_1 value and a very high Y value, that point would disproportionally affect the estimate of the line of best fit (e.g., Cohen et al., 2003, for a review). Such a case would not be readily identified without visualization (or more formal outlier/influential data point checks), which could allow the possibility of further investigating such a unique case. Cases in such situations are said to be *leveraging points*. In general, formally operationalizing what constitutes an outlier and appropriately dealing with them can be difficult, but it is nevertheless important. Cohen et al. (2003, ch. 10) provided a detailed discussion of possible causes and possible remediations when outliers are believed to exist. Regardless of the exact way in which outliers are dealt with, transparency to the reader is key. Transparency is

especially important because two researchers analyzing the same data might come to different conclusions when fitting the same model based only on how outliers are addressed.

In published work, space is often at a premium, which has the effect of only infrequently printing figures that evaluate the model assumptions (e.g., RVP, CPR, qq-plots). Nevertheless, even if such figures are not part of the published version of a work, there is little question that they can be very beneficial for authors, as well as satisfying reviewer curiosity on model fit and appropriateness, and can help to convey relationships to the reader easier seen than said. We think it is generally wise for authors to include a brief discussion of the (published or unpublished) figures and the seemingly appropriateness or inappropriateness of the model. Of course, if the figures help to identify weakness in the appropriateness of the model, other models should be considered and such a finding noted in the work. In short, visualization techniques should help justify the model chosen and this information should be conveyed to readers.

We are sensitive to the amount of journal space that such plots can consume. Due to limited journal space, editors may be reluctant to allow several pages of figures, even if they are informative. We believe a reasonable solution is for authors to produce supplemental material that can be referenced in the article but stored on a journal's supplemental materials web page, which many journals now make available. If not on a journal supplements page, the author(s) can often post additional information on an archival site (e.g., via university library).

12. Measurement Error

Measurement error in multiple regression can be conceptualized in a $2 \times 2 \times 2$ array, where depending on the specific conditions the effect of measurement error has different implications. The dimensions of the array are (a) type of measurement error (random or nonrandom), (b) type of variable (regressor or criterion), and (c) type of coefficient (unstandardized or standardized). We will briefly describe each dimension of the array below.

Random measurement error, which is omnipresent in research, is uncontrolled error that is assumed to have a mean of zero. Nonrandom measurement error, however, work will tend to have a mean that is not zero and/or be correlated with errors. In short, nonrandom measurement error in the criterion and/or the regressor is problematic and can lead to biased estimates of model parameters. Because nonrandom measurement errors often represent a flaw in the measurement procedure, instrument, or design, we will simple say that multiple regression is not generally appropriate in circumstances of nonrandom measurement error, with the exception being when the nonrandom error is so small that is has essentially no effect on the mean and covariance structures of the variables.

We will assume the random measurement errors have a mean of zero and are uncorrelated with measured variables, with their corresponding true scores, and with all other errors. Provided the regressors are unstandardized, any measurement error in Y is absorbed into the model error term, from Equation (1), and has no effect on the expected value of the regression coefficients. Thus, under the standard multiple regression assumptions, the regression coefficients remain unbiased. However, because the model error variance increases, the estimate of the squared multiple correlation coefficient is systematically lowered. Because R^2 decreases—it is attenuated due to a larger error variance—the standard errors of the regression coefficients will also be larger, implying that statistical power and the accuracy of parameter estimates are reduced via a decrease in precision. However, in the situation where the regression model is standardized, the regression coefficients will be attenuated when the criterion is measured with error (Kenny, 1979). The attenuation occurs when the criterion is measured with error because for standardized regression coefficients the multiplier (i.e., s_{xk} / s_Y for the kth regressor) of the unstandardized regression coefficient that yields the standardized regression coefficient has a denominator whose expected value is larger than the true

value. The expected value of s_Y is larger than σ_{Y_T}, the population standard deviation of the true scores of *Y*. From a classical test theory perspective on random measurement errors, the variance of *Y* is the sum of the true score variance ($\sigma^2_{Y_T}$) and the error variance ($\sigma^2_{Y_E}$). Thus, will tend to be larger than σ_{Y_T}, which leads to observed standardized regression coefficients smaller than their corresponding true values (Kenny, 1979, ch. 5).

In observational research, the case of random measurement error in one or more regressors will generally lead to biased regression coefficients, regardless of whether or not the regressors are standardized. As Fox (2008) showed, in simple regression (i.e., when $K = 1$) when measurement error occurs in the (only) regressor, its regression coefficient is generally attenuated. However, with one exception, no general statement can be given for the effect of measurement error in one regressor on the regression coefficient for the other regressors in a multiple regression model (i.e., when $K > 1$). As Kenny (1979, p. 104) pointed out, measurement error in one regressor can attenuate regression coefficients, make the estimate of a regression coefficient that is zero be nonzero, and can change the sign of a regression coefficient. The exception noted is for designed experiments, where the randomly assigned variable is uncorrelated with other regressors in the model. When the randomly assigned variable has measurement error, the regression coefficient is less accurate; it is unbiased but less precise. Because the regression coefficient is less precise, the corresponding confidence interval tends to be wider and the test of the null hypothesis will not be as powerful (larger *p*-value).

In general, the difficulty in saying what happens when measurement error occurs in an observational application of multiple regression lies in the multivariate nature of multiple regression, as the properties of one regressor influence the regression coefficients of all other regressors. In short, when a regressor is measured with error in an observational application, its effects are not partialed out as fully as when it is measured without error. This concept is easiest to understand when one regressor is perfectly unreliable, and thus the effects of the true regressor have not been partialed in any way (Kenny, 1979). As a result, the coefficients for other regressors in the model are generally biased because the perfectly unreliable regressor has not been controlled for at all. The important point is that whenever a regressor is measured with error, not only is the coefficient associated with that regressor biased, but typically so are all of the other coefficients in the model, including even coefficients for any regressors that happen to be measured without error. Because the value of the regression coefficient for the variable that is measured with error is biased, being smaller in magnitude than it otherwise would have been if the variable were perfectly reliable, the bias will generally lead to an error variance larger than it would have been, which then leads to a negatively biased estimate of P^2 (i.e., R^2 is, on average, smaller than it should be), ultimately leading to larger standard errors for all of the regression coefficients in the model.

It is desirable to minimize measurement error in all uses of multiple regression. However, measurement error is especially problematic when the primary goal is explanation, because theoretical explanations virtually always relate to constructs, not to variables measured with error. When confronted with nontrivial measurement error, it is often advisable to obtain multiple measures of each construct and use structural equation modeling (see Chapter 33, this volume) instead of multiple regression. Measurement error can be less problematic when the goal is prediction, because the practical goal is often to determine how well regressors as measured can predict the criterion as measured. When the goal is explanation and nontrivial measurement error is likely to occur, we generally recommend obtaining multiple measures of each construct so that structural equation modeling can be used.

13. Statement of Limitations

Multiple regression is a flexible system for linking *K* regressor variables to a criterion variable of interest. In many cases, multiple regression is an appropriate statistical model for addressing common research

questions, whether they be for purposes of explanation, prediction, or both. Nevertheless, multiple regression has limitations that are defined in part by the model and its assumptions as well as by the research design. The limitations of multiple regression in the specific context should be discussed.

Multiple regression has limitations, like other statistical models, when attempting to infer causality from a research design that was not experimental in nature (i.e., when random assignment of levels of the regressors to the participants was not part of the design). Although including additional regressors that are thought to be correlated with the regressor of interest adds a form of statistical control, with regard to causality there is no way to "control" all possible confounders unless randomization is an explicit part of the design. In purely observational designs, claims of causality should generally be avoided. The benefits of randomization cannot be overemphasized, even if for only some of the variables in the design, because randomization implies that the participants have equal population properties (e.g., mean and covariance structures) on all outside variables.

Variables termed "control" variables are often included in multiple regression, as previously noted. However, including a control variable in the model in no way implies that the variable can literally be "controlled"—use of such a term is based on a precise statistical meaning and is not literal in the sense of everyday language. When something is "controlled for" it allows for the linear effect of each regressor on the criterion variable to be evaluated (i.e., a regression coefficient estimated), while holding constant the value of the other regressor variables. In practice, however, many variables cannot be controlled by the researcher, even in the most carefully designed studies. Thus, there is not literally any control by the researcher in an observational design over the variables said to be "controlled for." Rather, an effect can be examined while holding constant the other variables.

The reasonableness of temporal ordering of variables needs to be considered, as multiple regression can be applied in ways such that an explanatory variable is nonsensically used to model a criterion variable. Although the multiple regression model may account for a large proportion of variance, it might not make theoretical sense. For example, multiple regression could be used to model "time spent studying" as a function of "test score." However, such a model is nonsensical in the sense that "time spent studying" would be an explanatory variable of "test score." This is a simple example of a causality problem, in the sense that the multiple regression model itself does not make a distinction between what causes what. Theory, of course, should be the guiding principle of the specification and direction of causal relationships. Inferring causality can be difficult, especially because there technically needs to be some passage of time that occurs in order for a regressor to literally cause some change in a criterion (unless simultaneous causality is presumed).

14. Alternatives to Multiple Regression

When the assumption of normality of errors is violated, nonparametric approaches to inference for multiple regression should be considered (e.g., Efron & Tibshirani, 1993; Györfi, Kohler, Krzyzak, & Walk, 2002). Multiple regression assumes that outcome variables are continuous and observed. However, when the criterion variable is censored, truncated, binary/dichotomous, ordinal, nominal, or count, an extension of the general linear model termed the generalized linear model, where a link function (e.g., exponential, Poisson, binomial, logit) relates the linear regression equation (analogous to the right hand side of Equation (1)) to a function of the criterion variable (e.g., probability of an affirmative response) can be used (e.g., Agresti, 2002; Long, 1997; McCullagh & Nelder, 1989; Chapters 16 and 17 in this volume).

Linearity is an assumption that is not reasonable in some situations, either based on theoretical or empirical evidence (e.g., the graphical displays previously discussed). *Spline* regression models allow different slopes over ranges of one or more regressors, in what has appropriately been termed a piecewise model (e.g., Fox, 2000; Ruppert, Wand, & Carroll, 2003). In spline regression multiple

"knots" exists, where the slope of the regression line (potentially) changes over specified ranges (note that the slopes can be discontinuous in that they need not overlap at a knot). Another nonparametric regression procedure is known as *lowess* (locally weighted scatterplot smoothing) (also denoted *loess*; e.g., Cleveland, 1979; Fox, 2008), in which multiple regression models are fitted to areas/regions of the regressor(s) with "local" points receiving more weight than more distant points. The definition of "local" changes as a function of the width of the span selected, which is a parameter in the control of the analyst and for which there is not a single best answer to the ideal size of the span. For short spans the line of best fit can differ dramatically over a small range of a predictor, whereas a wide span tends to have a relatively smooth relationship between the regressor(s) and the criterion. Lowess techniques are most often used when $K = 1$. More general than lowess models are generalized additive models that allow some regressors to enter the model linearly and some to enter as splines (Ruppert et al., 2003, p. 215).

Applications of the general linear model are not robust to violations of the assumption of independent observations. Even for the simple case of the two independent group *t*-test, which can be considered a special case of multiple regression, it is known that the nominal and empirical Type I error rate can be drastically different when the assumption of independence is violated (e.g., Lissitz & Chardos, 1975). When observations are not independent (e.g., students nested within classrooms, clients nested within therapists, observations nested within person), appropriate methods to explicitly control for the lack of independence should be used. A general approach to handling such nonindependence is multilevel models (also termed *hierarchical linear models, mixed effects models*, or *random coefficient models*; see Chapter 22, this volume).

When measurement error is not ignorable, multiple regression is not ideal and latent variable models should be considered, especially when the primary goal is explanation instead of prediction. In particular, confirmatory factor analysis (see Chapter 8, this volume) and structural equation modeling (see Chapter 33, this volume) allow for explicitly incorporating error into the model of interest, which has the effect of separating the "true" part of the model from the "error" part.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ. John Wiley & Sons.

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

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*, 119–136.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 7–21.

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

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Burket, G. R. (1964). *A study of reduced rank models for multiple prediction*. Psychometric Monograph, no. 12. Richmond, VA: Psychometric Corporation. Retrieved from www.psychometrika.org/journal/online/MN12.pdf.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Enders, C. K. (2001). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*, 713–740.

Ezekiel, M. (1930). *Methods of correlational analysis*. New York: Wiley.

Fox, J. (2000). *Multiple and generalized nonparametric regression (No. 131)*. Thousand Oaks, CA: Sage.

Fox, J. (2008). *Applied regression analysis, linear models, and related methods* (2nd ed.). Thousand Oaks, CA: Sage.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, *26*, 499–510.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer.

Harrell, Jr., F. E. (2001). *Regression modeling strategies with applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hernan, M. A., Hernandez-Diaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*, 176–184.

Huberty, C. J. (2003). Multiple correlation versus multiple regression. *Educational and Psychological Measurement*, *63*, 271–278.

Jaccard, J., Guilamo-Ramos, V., Johansson, M., & Bouris, A. (2006). Multiple regression analyses in clinical child and adolescent psychology. *Journal of Clinical Child and Adolescent Psychology*, *35*, 446–479.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression* (2nd ed.). Thousand Oaks, CA: Sage.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1–24.

Kelley, K., & Maxwell, S. E. (2008). Power and accuracy for omnibus and targeted effects: Issues of sample size planning with applications to multiple regression. In P. Alasuuta, J. Brannen, & L. Bickman (Eds.), *Handbook of social research methods* (pp. 166–192). Newbury Park, CA: Sage.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley-Interscience. Lissitz, R. W., & Chardos, S. (1975). A study of the effect of the violation of the assumption of independent sampling upon the Type I error rate of the two-group *t*-test. *Educational and Psychological Measurement*, *35*, 353–359.

Little, R. J. A., & Rubin, D. A. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley and Sons.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

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

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19–40.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593–614.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*, 23–44.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181–190.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed). New York: Routledge.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376–390.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

Muller, K. E., & Fetterman, B. A. (2002). *Regression and ANOVA: An integrated approach using SAS Software*. Cary, NC: SAS Institute.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace.

Raju, N. S., Bilgic, R., Edwards, J. E., & Fleer, P. F. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement*, *23*, 99–115.

Rencher, A. C., & Pun, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics*, *22*, 49–54.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Hoboken, NJ: Wiley.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. New York: Cambridge University Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. New York: Wiley.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.

Steiger, J. H., & Fouladi, R. T. (1992). R2: A computer program for interval estimation, power calculation, and hypothesis testing for the squared multiple correlation. *Behavior Research Methods, Instruments, and Computers*, *4*, 581–582.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2005). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*. New York: Springer.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

24
Multitrait–Multimethod Analysis

Keith F. Widaman

Campbell and Fiske (1959) argued that every measurement we obtain in psychology is a trait-method composite—a measure purportedly of a particular trait construct obtained using a given method of measurement. Campbell and Fiske introduced the multitrait–multimethod (MTMM) matrix as a tool for evaluating systematically the correlations among a set of measures obtained using multiple methods. The primary utility of the MTMM matrix approach is the opportunity such a study affords to determine the preponderance of trait-related and method-related variance in measures in a battery. To aid in this evaluation, Campbell and Fiske argued that researchers should measure each of *t* traits (e.g., Extraversion, Neuroticism, Fluid Intelligence) using each of *m* methods (e.g., self-report, objective tests, observer ratings), so that each trait is measured using each method. By arranging trait measures in the same order within methods, the MTMM matrix should exhibit clear patterns to satisfy the dictates of convergent and discriminant validation. Convergent validation is satisfied if the researcher finds high correlations among measures of putatively the same construct using different methods of measurement, and discriminant validation is satisfied if low correlations are found among measures of presumably different constructs. Campbell and Fiske described several rules of thumb for evaluating patterns of correlations in the MTMM matrix. Specifically, (a) correlations between measures of the same construct obtained using different methods of measurement should be large; (b) correlations between measures of the same construct obtained using different methods of measurement should be larger than correlations of those measures with measures of different constructs obtained using the same or different methods; and (c) the same pattern of trait correlations should hold for all combinations of methods.

Among others, Jöreskog (1971) pioneered the fitting of confirmatory factor analysis (CFA) models to MTMM data. The CFA approach circumvented several problems associated with the Campbell and Fiske (1959) rules of thumb. In particular, the CFA approach (a) yielded clear significance tests of differences between alternative models and of specific parameter estimates, whereas the ordinal comparisons involved in the Campbell-Fiske rules of thumb relied on dependent comparisons that compromised statistical tests; (b) allowed for tests of the amount of trait-related and method-related variance in the MTMM matrix; and (c) led to estimates of the amount of trait-related and method-related variance in each measure. Widaman (1985) systematized earlier work on CFA models and provided an informative taxonomy of models for MTMM data by cross-classifying available trait factor structures and method factor structures. In addition, Widaman discussed alternate analytic strategies