Compensating for Deficiencies in Perinatal Data Sets: Parametric Perspectives

A. Scott Tippetts and Paul R. Marques

INTRODUCTION

Considering the investments made in a research demonstration project—including the hours of work in research design, provision of services, data collection, data entry, money spent, and professional reputation—great care should be taken when testing hypotheses, but often it is not. A small amount of extra effort at the data analysis stage frequently can yield great benefits in terms of more accurate findings. The reason: Very few data sets are ideal, and among nonideal data sets, those of perinatal drug abuse treatment projects have few equals.

Data deficiencies are sometimes apparent without having to search for them, showing up as highly abnormal or extreme values or as missing data or responses. Other data deficiencies are not noticed until carefully pursued. When corners are cut (e.g., not checking distributions for normality, not performing basic diagnostics of residuals), the researcher risks missing the real relationships or reporting spurious relationships. Although some researchers excuse these shortcuts with the rationale that analysis of variance (ANOVA) and regression are relatively robust against violations of assumptions, it is easy for even a single overlooked data error to completely negate the outcome of a statistical test.

Even for researchers who regularly check for data deficiencies and irregularities, and therefore often discover them, the question arises, What should be done with the deficiencies? This chapter covers some basic "textbook" approaches for dealing with data deficiencies as well as some less used, more imaginative techniques.

NONNORMALITY

Before any hypothesis testing is undertaken, data *always* should be examined to assess the degree to which assumptions may be violated, especially assumptions that require a data set to be normally distributed. Even when a data set includes many variables, the time involved in quickly checking the distribution of each variable can pay big dividends in the long run. Histograms are the most basic visual check, and many statistical

software packages can superimpose the outline of the normal distribution over the histogram. Abnormal distributions, such as severely multimodal distributions, probably should be excluded from parametric analysis altogether, unless they can be modified in some way to become moderately normal (e.g., by collapsing them into fewer categories, such as by dichotomizing).

Conventional Transformations

Most distributions are fairly unimodal and taper into tails at each end, showing some semblance of normality. For such variables, ANOVA and regression are robust; nevertheless, the power of parametric procedures can be improved on—sometimes dramatically—by linearly transforming the variables into more normally distributed versions. One simple way to assess the need for a transformation is to check the degree of skewness, because skewed data represent perhaps the most common type of distribution abnormality. Any statistical package worth using should provide the skewness and standard error of skewness for each distribution. A good rule of thumb is to be skeptical of any distribution whose magnitude of skewness (in either direction) is greater than three standard errors of skewness, although in large samples (e.g., more than 1,000 cases), the standard error of skewness becomes so small that this rule of thumb becomes unreasonable and is likely to be exceeded even with only slight amounts of skewness. In cases where the distribution is positively skewed (i.e., having a long tail tapering to the high side), taking the square root, cube root, log, or reciprocal of the variable can compress the tail, "pushing" the distribution toward a more normal form. By contrast, when the distribution is negatively skewed (i.e., having a long tail tapering to the low end), using the square, cube, or exponential (*e* raised to the power of the variable) pushes the lower values of the distribution closer in. When more than one transformation seems to be successful in achieving normality, one can decide which transformation works best based on the new skewness statistics for the transformed variable because the standard error of skewness remains the same regardless of transformation.

A word of caution: Some transforms are sensitive to certain ranges of values, such as power transforms (e.g., squares, square roots) on data that straddle the value of 1. Root transforms converge toward 1 on *both* sides, and powers greater than 1 diverge away from 1 toward infinity or toward zero. For this reason, before using a power transform on a scale that includes values between 1 and zero *and* values greater than 1, it may be appropriate to first rescale via multiplication so that no values lie between zero and 1 (exclusive) or all values range from zero to 1. Log transforms also can be undesirable with values less than 1, and logarithms of values less than or equal to zero are undefined. Again, rescaling through

multiplication can make distributions more amenable to log transforms and may optimize the transform (see below).

Finding nonnormal variables that can be normalized through linear transformation does not imply that the original, untransformed variable should be discarded or ignored during hypothesis testing. The original variable may produce better relationships with other variables of interest than does the more normal transformed variable, even with sufficiently random residuals. In many such cases the use of the original variable is permissible because of the robustness of many parametric procedures. However, the linear transformation often not only helps to better satisfy the assumptions needed for the procedures but also may uncover a relationship not detectable until the variable is normalized. Indeed, one might wonder how many significant yet moderate relationships have been reported that were actually strong relationships, watered down by using untransformed variables that strayed only marginally from normality. This was the case with hair cocaine data from mother-infant pairs, where square root transformations normalized the distributions, strengthening the correlation from 0.33 to 0.41 (Marques et al. 1993).

Maximizing Normality

Some may question the practice of regularly checking and transforming variables that are in need of normalization, but such a view fails to recognize that many metrics are created in a somewhat arbitrary manner that may not necessarily be isomorphic with the phenomenon being measured. Some of the most obvious such measures are decibels (sound volume) and the Richter scale for measuring seismic activity, both of which have a logarithmic relationship between magnitude and the measured values. Standard radioimmunoassay values are usually based on a log-logit plot of drug concentration to the ratio of bound-to-free radioactivity. Even simple survey scales that sum positive (or negative) responses to a series of yes or no questions may linearly distort the metric of the underlying "ideal set" of response patterns. This is why some psychometric instruments are more sensitive than others; better instruments or scales produce values that correspond more closely to the natural progression of the underlying concept or latent variable. Such issues from measurement and data theory are abstract, but the process of mapping phenomena onto quantifiable scales often creates the frame of reference, or metric, leading often to the mistaken notion that the metric came first. The process of linearly transforming variables sometimes may be nothing more than restoring the isomorphism of the metric to the nature of the phenomenon.

There is another rarely used (and to some controversial) variant of linearly transforming variables to satisfy assumptions of normally distributed data.

If one accepts the concept that many metrics or scales are somewhat arbitrarily mapped from the "ideal" metrics of the phenomena, then the process of transformation may uncover these "ideal," inherent metrics. If so, then rather than looking only at transformations involving a few limited exponents such as 0.5 (square root) or 2 (squared), one might solve for the power that maximizes the normality of the distribution along the entire scale of possible exponents. The procedure for finding the optimal linear transform involves creating an extra variable that consists of the *expected* normal scores from a normal distribution and corresponds to the ranks of the values of the variable to be transformed. If the Statistical Package for the Social Sciences (SPSS) (Norusis 1992) is used, these expected normal scores can be computed with the command RANK/INTO NORMAL. Then, the original variable (RAWVAR) is modeled in a basic nonlinear regression equation

$$NORMLVAR=((RAWVAR^{POWER})–MEAN)/STDDEV$$

that predicts the new normalized variable of expected normal scores (NORMLVAR). The parameters for MEAN and STDDEV are estimated for the new transformed variable. The estimated parameter POWER is the exponent to which the original variable RAWVAR must be raised to best produce the normal distribution NORMLVAR. The solution to the power transform question is sample dependent and may not differ much from one of the basic transformations (e.g., square, square root) previously mentioned. In addition, this process can add much time (and therefore cost) to the preanalysis stage, while gaining marginally little over the traditional transformation powers. If a variable requires normalization, this approach is as defensible as any other, because the original relationship among values is not altered.

The procedure described above is based on power transforms, which work well in many instances, but sometimes the distribution in question requires a log transform. This is fairly common when the scale is constrained at the lower end (often at zero), which produces something not unlike an F-distribution. As with the power transform, there is usually a particular base whose log produces the most normal distribution. Because most computers are limited to logarithms of the bases 10 and *e*, some scale adjustment is necessary before taking the transform. By multiplying/ dividing the distribution by a constant, then adding/subtracting a constant, the original distribution first is rescaled around the log's base so that the ensuing log transform will maximally approach normality. Although this three-step procedure may seem excessive, remember that multiplication and addition *do not change the shape of the distribution at all* but only shift the metric. Any analysis of these rescaled (but still untransformed) data will produce exactly the same results as the original distribution in

terms of test statistics and probability values. However, the log-transformed distribution of the rescaled data will be closer to normality than that resulting from the unrescaled original distribution, and the transformation is still completely linear.

BISERIAL CHANGE VARIABLES

Although there are various objections to consolidating repeated measures into a single "change" measure on an individual case basis, it is nevertheless common practice for researchers to do so for the sake of simplicity, lack of analytic sophistication, or the need to overcome the violation of the assumption of independent observations without having to resort to more complex repeated measures procedures like multivariate analysis of variance (MANOVA). Because the practice is certain to continue, it is helpful to be aware of some of the more basic alternatives that can be used when a single variable representing change is desired.

Simple Difference and Percent Difference

The first type of change is the simple difference, or posttreatment measure minus the pretreatment measure (or baseline). Much has been written about the mathematical issues relating to difference scores, and many researchers have proposed corrections to produce less biased estimates of a true difference (e.g., Chronbach and Furby 1970; Harris 1963). Many researchers continue to compute simple difference scores anyway, either because of unfamiliarity with the alternatives or to sidestep the complexity of the correction formulas. Yet numerous researchers have abandoned the use of difference scores (and their various corrections) because of a conceptual consideration, not a mathematical one: Nearly all fail to adequately address the idea of "relative change."

For example, regardless of how the difference score is computed or corrected, a change of –200 ng in measured hair cocaine levels may be an unimpressive reduction for a subject who measured 1,500 ng at basal, but such change arguably represents a dramatic improvement for a subject who began at a level of only 400 ng. This is particularly true in substance abuse research where, as a result of different degrees of drug tolerance, a low or moderate exposure level for one person may be an extremely high level for another. Data for such phenomena often manifest this relativeness through a high positive correlation between the baseline value and the absolute value of the simple difference score. For this reason, most researchers who compute single "change" variables often prefer the more subject-specific *percent-change relative to baseline*, in which the –200 ng from a baseline

of 400 ng would be represented as a 50-percent decrease but from a baseline of 1,500 ng would be represented as only a 13.3-percent decrease. The intuitive appeal of this computation is that change is relative to each case's baseline. If all cases have the same baseline (such as under the most ideal laboratory-controlled conditions), the same percent-change would also represent the same amount in terms of raw difference. In clinical studies a relative change variable is often used to estimate an effect size for treatment progress.

The problem with this common percent-change measure is that most data for which it is computed are constrained at the low end by zero or some other minimum constraint below which values are inherently not possible (e.g., less than zero treatment hours or a negative amount of drug found in urine samples). This would not be a problem if values rarely approached the minimum (i.e., the mean would be at least three or four standard deviations above this constraint), but in most types of clinical research, this minimum value constraint is often observed with some frequency so that the distribution of values is bunched up against the minimum, as shown in figure 1.

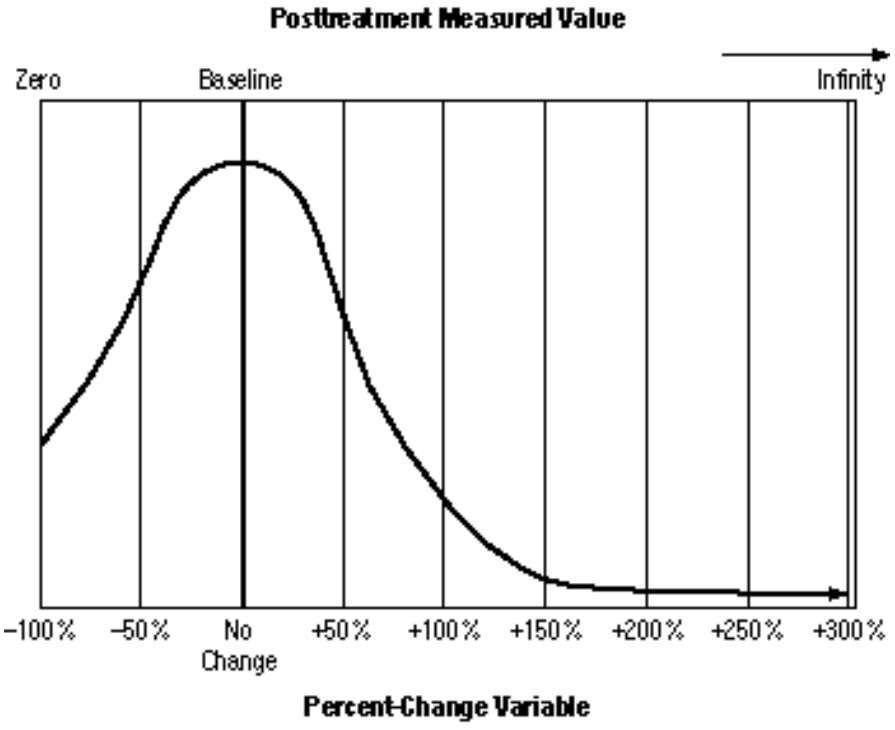

FIGURE 1. *Distribution of potential percent-change scores*

The result is that the *theoretical* universe of all *possible* values for the change variable, with a theoretical median of zero percent-change, suddenly disappears at –100 percent, chopping off the tail of what should be a normal distribution of values. (A 101-percent decrease for such scales is inherently impossible.) On the other end of the scale, the distribution has a normal tail approaching a theoretical limit of +infinity-percent increase. The observed values disappear well short of infinity for *empirical* reasons (i.e., sampling and/or measurement sensitivity), but this upper measurement constraint is usually well above the median, which allows the sampling distribution to have an upper tail that is virtually normal. Depending on what is being measured and the sensitivity of the instrument, the upper end of the percent-change distribution allows for increases as large as 500 or even 5 million percent! This asymmetry violates the assumption of normality in a serious manner; the irregular *potential* distribution is not the result of empirical problems (i.e., a sampling anomaly) but is inherent in the computation. In any theoretical sampling from the universe of all clinical trials of pre-post measures, researchers who expect to see a pre-post reduction have the deck stacked against them if they compute change as a percentage difference from baseline.

Log-Change Ratio

The solution to this situation is obtained through a slight modification in the calculation of the "change" variable. Similar in principle to log-likelihood ratio statistics, the log-change ratio (LCR) is obtained by computing the ratio of the postmeasure to the premeasure, then taking the log of this ratio. The theoretical distribution (with a null hypothesis of no change) of the resulting variable is perfectly normal and symmetrical about zero, which represents no change (i.e., when both premeasures and postmeasures are equal). Taking the log of the ratio also prevents creation of extreme values that result from a spuriously low value (baseline or postvalue) and allows a more complete spread when postvalues represent marginal change.

Although its qualities make the LCR more robust and conceptually meaningful than percent-change, its metric is not readily interpretable in a descriptive sense. For description, one can work backward from the LCR value to reconstruct an example of "representative change" using the mean basal value and the LCR-expected posttreatment value. The authors find the LCR to be best suited for reporting effect sizes (such as *d*, difference divided by standard deviation) because no change equates to an LCR of zero and proportional change is symmetrical about zero; simply dividing the LCR by its standard deviation produces the effect size.

However, the LCR is not without its bugs; values of zero for either measure produce undefined results, so zeros must be recoded to some nonzero value.

The most conservative method is to recode zeros to the lowest nonzero value observed in the sample, with more liberal procedures recoding zeros to some point, such as halfway between the empirical nonzero minimum and zero. Although choice of this recoding point for zeros is arbitrary, the effect on the *potential* distribution is symmetrical regardless of the value chosen because the zero value could occur in either the premeasure or postmeasure. On the other hand, the effect on the empirical distribution can be profoundly affected, depending on the recode value chosen, and the researcher should report explicitly the value to which zeros are recoded. The authors recommend that the lowest nonzero value observed or measured be used because it is (1) the most conservative option and (2) less arbitrary (empirically determined). The latter also implies that in a truly constrained distribution this "next-lowest" value should theoretically correspond to the smallest detectable increment above zero, assuming a fairly large sample. Other procedures for empirically deriving the appropriate minimum (based on various percentile ranges within the sample distribution) have been suggested but not yet fully evaluated.

The LCR shares the same advantage of "relativeness" with the percent-change measure in that the magnitude of change is considered relative to the baseline or premeasure value. The symmetry about the expected mean of zero change (and ensuing normality in the potential distribution) of the LCR is a desirable feature and a definite advantage over percent-change. For example, calculated by the percent-change method, a change from 40 to 200 ng would produce a score of +400 percent, whereas the same change in the other direction (from 200 to 40 ng) produces a score of –80 percent. In this case the LCR would produce values of the same magnitude for both the increase and decrease (+1.609, –1.609), differing only in sign.

DATA ERROR

Up to this point the discussion of data deficiencies has centered on violations of the normality assumptions, which can and do occur in perfectly good and accurate data. Such violations also occur because of data error, which can be of many types and can creep into the data at many stages in the research process. Whether measurement, sampling, or data entry error, all error, regardless of source, should be minimized. In research where sample sizes are limited (e.g., perinatal research), a single substantial error or a handful of small errors may completely negate what otherwise might have been a strong finding. Cohen (1990) cites a study of 25 height-weight pairs in which reversal of data for a *single* subject would have changed the 0.83 correlation to a –0.26 correlation! The authors recommend always performing random spot-checks of data already entered and/or duplicating the data entry for small segments of

the data set to verify the integrity of the data entry. The rate of data entry error can be estimated by drawing periodic random samples for verification.

Outliers

Other sources of data error include measurement instrument ("sensitivity") error and sampling error. These are inherent in any data set and cannot be quantified or known in any practical sense. (If they could be identified and quantified, probabilistic methods would not be necessary.) Because they usually can be assumed to be normally distributed, they are "ignored" as the error term in the statistical model. However, remote "one-in-a-million" cases exist in the near-infinite populations that are sampled, and these "data spikes" inevitably pop up at times. The long-run expectation of finding such cases does not change the fact that they are significantly overrepresented in a sample of only 100. Therefore, when checking variables for normality in the preanalysis, it is important to check for extreme outliers as well.

The definition of the normal distribution stipulates the existence of tails, but given a sample size, the researcher can determine the probability of any z-score being a chance occurrence. For example, z-scores greater than 3 should occur only about three times per 1,000 cases, and z-scores greater than 4 should occur only about six to seven times in a sample of 100,000 cases. Unless working with large data sets (more than 10,000 cases), it is always worthwhile to investigate cases in which a value has a corresponding z-score greater than 3 and certainly always when the z-score exceeds 4. The value may turn out to be a data entry error just waiting to ruin a correlation, or it may belong to a case/subject that on further investigation is found to be so bizarre and unrepresentative of the population being studied as to warrant its exclusion as an anomalous case.

Simply excluding a case on the basis of one extreme value is a highly questionable (and arguably unethical) procedure; the decision to exclude a case from the data set should be based on some other *objective* basis, which may result in the exclusion of one or more "good" cases also. When any subject has a preponderance of extreme values on a number of measures (variables), a brief case study/descriptive explanation should be included in the research report to justify the basis for a "specific" (nonobjective) exclusion. This is not to say exclusions are not sometimes justified and proper. The authors recommend that, when the researcher locates a questionable case so extreme as to radically influence the outcome of an analysis, the report include the results of *both* analyses, including and excluding the questionable case, as well as the bases on which the case was deemed questionable. If the case is truly bad, producing a spurious or erroneous result, and the researcher fails to exclude it, the

scientific integrity of the research is jeopardized just as much as if the case is excluded simply to boost a correlation. If the results do not change much either way, presumption must rest with leaving the questionable data in.

The previous discussion of extreme cases refers only to univariate extremeness. It is possible that a case may be an extreme anomaly worthy of exclusion, but only in terms of its multivariate distribution. Such cases often cannot be identified as extreme on any single dimension. Perhaps the simplest way to locate the existence of such cases is to perform a cluster analysis specifying a small number of clusters (e.g., three to five). If any of these clusters comprises single cases (or relatively few cases compared with the other clusters), there is a possibility that these cases are multidimensional extremes. The researcher can follow up on such indications by computing for each case the mean pairwise distance score. Truly extreme cases should stand out. To avoid having the different metrics of variables influence the distance scores, it is usually preferable to first standardize the variables to be used in the clustering procedure. Also, highly correlated variables can be thought of as indicators of the same latent dimension, which then would contribute disproportionately more to the cluster solution if these somewhat redundant variables are all included in the analysis. The researcher can select the most appropriate variable from such a group or combine several through principal components analysis to keep them from exerting undue influence in defining the multidimensional space of the data set.

Other Errors

Data entry errors occurring within normal ranges of values may be undetectable when one looks at the univariate distributions yet still be different enough from their true values to greatly distort a relationship with another variable. These data errors are perhaps the most pernicious of all. For this reason, it is advisable to use a statistical package that can produce residual diagnostics, particularly lists of cases that have great influence (or leverage) on the outcome of the analysis. Some of the best measures to check are Cook's distance, Mahalanobis' distance, and deleted studentized residuals (Belsley et al. 1980; Cook 1977). Again, cases with unusual influence should be checked for data entry errors or on some other *general and objective* basis that would disqualify them.

One other source of error that can inflate the residual error is related to independence of sequential observations/measurements. It is useful to have variables that indicate sequence of measurement or collection in the data set so that sequence can be plotted against both the dependent variable and the residuals from any analysis. Sequential error can occur as the result of instrument calibration issues (such as warmup or drift),

measurer/collector change (e.g., interviewer fatigue/boredom, interviewer's becoming more proficient in measuring or using the instrument), or even from subject "learning" (such as becoming more proficient at tasks that are performed repeatedly). Such sources of sequential error can sometimes be explicitly modeled into an analysis (e.g., as a covariate in an ANOVA), thereby reducing the residual error and strengthening the estimate of the magnitude of the effect size being tested.

MISSING DATA

One type of instrument sensitivity error mentioned above is the failure to register a value at all, such as when a mechanical instrument fails to sense a stimulus or when a survey respondent fails to provide a response. In perinatal research, missing data can be common because of uncooperative subjects, scheduling difficulties, and access restrictions. When working with many variables in multiple-variable or multivariate analyses, the researcher may find that the usual default procedure of excluding cases with a missing value on any one of the many variables in the analysis (often called listwise deletion) will eliminate nearly all cases in the sample. Even if nearly every variable has valid data for 98 percent of the cases, if the 2 percent missing data are different for each of 50 variables, then the sample N can disappear completely. Whereas in most data sets missing data tend to congregate within the same variables (difficult questions or bad instruments) and/or within the same cases (e.g., problem respondents), a data set with many variables can have enough of a scattering of missing data to render an analysis fairly unreliable when excluding cases with missing data listwise.

There are at least three methods of dealing with missing data without having to exclude cases altogether. However, the researcher first must determine whether the missing data have a story to tell, because missing data can be evidence of a relationship. If missing data patterns are correlated with other variables, the researcher may have the serious problem of *nonresponse bias*, which can render any statistical inference invalid and unrepresentative of the entire sample. In cases where variables have missing values for more than a handful of cases, it is important for the researcher to test for nonresponse bias. The easiest way is to divide the sample into two subsamples: those with valid data for a particular variable and those without. Then the researcher can perform a simple difference of means test (such as the two-independent sample *t*-test) on other important variables for which both groups have most cases with valid data. Any significant difference implies that those cases with missing values may be different from the rest of the sample and that their actual values on the missing variable are possibly unrepresentative and cannot be "corrected"

by substitution or imputation methods with much reliability. Such patterns of nonresponse can be significant findings.

Mean Substitution

If the researcher decides to salvage cases with missing data without losing the good data, the most common method is to employ a mean substitution strategy, in which the missing data for a predictor variable receive the value of the mean for all those cases with a valid datum for that variable. This is a good method to use in situations where an analysis includes many "incomplete" variables, each of which may be missing values for only a very few cases, but the values are scattered among *different* cases for each variable. On the other hand, if the variables tend to have their missing values concentrated within the same few cases, then the researcher is probably better off excluding those cases. When the missing data are scattered among the cases, mean substitution can keep a fairly complete data set from losing a substantial portion of its cases in multivariate analyses, with relatively little risk of introducing a falsely positive bias.

The problem with the mean substitution method is that it usually flattens true correlations by reducing variance, erring on the side of being conservative (reducing magnitude of the effect). Because this is perhaps the easiest and least risky way to salvage a disappearing N, many software packages offer mean substitution as an option for multivariate procedures. As mentioned above, any indication that nonresponse patterns are significant makes mean substitution a highly questionable procedure. In any respect, mean substitution should be undertaken *only for independent (predictor) variables* in an analysis, and care should be taken to exclude cases that have missing values for the dependent variable, even when mean-substituting for independent variables.

Imputation

A second method for dealing with missing data is to impute, or predict, values for these cases based on other variables—preferably from variables *that will not be used* in the analysis (called external variables here). Much has been written on the topic of data imputation, and it is advisable to examine some of the literature if this path is chosen, as there are many complex issues (and procedures) dealing with imputation (e.g., Little and Rubin 1987). Imputation is perhaps most advantageous when a small number of important predictor variables have missing values and nonmissing values of these important variables can be reliably predicted from a combination of two or more external variables. If such other variables exist that are successful in predicting values of the "incomplete"

variable in question, the multiple regression or ANOVA model used for predicting the valid cases can be used to assign predicted values for those cases with missing values. The danger here is that the researcher may spuriously inflate the findings of the statistical tests by reinforcing the relationships used to impute the values—essentially perpetuating a result by increasing the N. For this reason, missing values for the independent variable should *always* be imputed from variables external to the analysis, although these external variables may be used in other unrelated analyses that do not involve the imputed variable.

As with mean substitution, it is arguably not appropriate ever to impute values for *dependent* variables. When imputing values for predictor variables, the dependent variable must *never* be used, and external variables also should be avoided if they have even moderate bivariate correlations with the dependent variable. One rule of thumb is that no external variable contributing to the imputation should have a bivariate correlation (Pearson's *r* [Snedecor and Cochran 1980]) with the dependent variable greater than half the multiple *r* of the imputation equation. This way the researcher is less likely to fall into the trap of circular logic. Although imputation can be a beneficial tactic for dealing with missing data, the dangers of using it improperly are great enough to warrant caution. When it is used, it is a good idea to weight cases with imputed data somewhat less than cases with all valid data. One good rule is to weight the case by the multiple *r* (or, more conservatively, the r^2) of the equation used to impute values for the missing variable, which means that the less successful the imputation formula, the less weight is given the cases relying on the imputed data. One implication of this policy is that cases utilizing imputed data carry case weights worth a fractional case, which often results in interesting sample Ns (e.g., N=121.37). The researcher should report the weighting scheme used to protect against suspicion of posthoc "data adjusting."

In situations where more than one variable in the analysis carries imputed data and some cases have imputed data for different variables than do other cases, case weights depend on which variable(s) a case is missing. Cases that carry imputed data for more than one variable should be weighted by the product of all the multiple *r*s (of the imputation equations for these imputed variables) so that the more missing data elements a case has, the closer to zero that case's weight becomes.

Multiple-Donor Matching

A third method for dealing with missing data is a variation of "donor-recipient" matching. This technique usually produces results similar to that of the multiple-regression imputation. The idea behind

donor-recipient matching is to find the case or cases that are most similar to a case with missing data in terms of other important variables for which the "recipient" case has valid data. The matching "donor" case then contributes its score on the variable in question to the case that is missing a value for that variable. This can be done through a clustering procedure, which determines distance between cases in terms of multidimensional space or, inversely speaking, determines the most similar or closest "neighbor" in terms of these other nonmissing attributes. Because many such clustering procedures are affected by the particular metric scaling of each variable, the clustering should be based on either standardized scores (i.e., z-scores) or on ranks for the nonmissing variables. Also, having several variables that are similar and highly correlated among themselves can overemphasize the importance of these variables in determining the distance score. For this reason, some researchers prefer to extract principal components (factors) from among the set of variables and then to cluster the cases based on factor scores.

Some researchers propose a strict one-to-one donor-recipient match, but the authors recommend instead that the recipient case receive a mean value for the group of closest neighbors. Determining how many neighbors to use as donors can be based on a criterion that specifies the average size of clusters. For example, if one wishes the average cluster size to be approximately 5 percent of the total sample, a cluster solution that produces 20 clusters would be specified. Some clusters contain more than 5 percent of the cases and some less, depending on the distribution of multidimensional similarities. The recipient case then would have its missing value replaced by the mean of that variable for the other cases in the recipient case's cluster. The criterion to use for specifying cluster size should depend on sample size. The authors' somewhat arbitrary recommendation is that samples with fewer than 100 cases should average 5 cases per cluster; samples of 100 to 500 should average 5 percent of the sample per cluster; and samples with more than 500 cases should produce clusters averaging about 25 cases.

This type of donor-recipient matching can be rather time consuming. Although the end result will often be similar to the result of multiple-regression imputation described earlier, donor-recipient matching can provide more reliable estimates of the missing data when relationships among the various dimensions are nonlinear or nonadditive or have specific ranges in which residuals from multiple regression are not randomly distributed. Finally, as with the multiple-regression imputation technique, recipient cases should be weighted to some fraction less than a full, single case. One possibility for determining the weight for such cases involves the ratio of the cluster's standard deviation of the variable to be donated to the standard deviation of this variable for the total sample. For example,

if the cluster standard deviation is 3.57 and the total sample standard deviation is 16.22, then the weight for recipient cases in this cluster would be 1–(3.57/16.22), or 0.78. In this way, the more homogeneous the cluster, the closer the weight approaches that of a full case.

NONLINEAR RELATIONSHIPS

One fact of scientific life is that, for one reason or another, many true relationships are not strictly linear.[1] Although sophisticated procedures exist to test models that are more complex than a simple linear frame of reference, such as nonlinear regressions, dynamic/interactive systems modeling, and qualitatively interactive segmentation or "tree-splitting" procedures, these nonlinear/nonstatic/nonadditive models are beyond the scope of this chapter. Nevertheless, many times a relationship that is not strictly linear may be *intrinsically* linear through the use of a linear transformation, such as those discussed above in the section titled "Nonnormality." Many times a simple linear regression will produce a significant result, yet the true result may be a stronger relationship if the function is curvilinear, such as a log function relationship (as in figure 2). Failure to model the relationship as its actual log function or polynomial function (achieved by transforming some or all the variables in the analysis or adding square and cube components) results in underestimating the magnitude of the relationship—sometimes concluding incorrectly that there is no relationship, such as would result when the relationship resembles a basic parabolic function as shown in figure 3.

Thinking Nonlinearly

There are two ways to guard against mistakenly missing or underestimating a relationship. The first is to be more imaginative in conceptualizing the nature of the phenomenon and break away from the constraints of the strictly linear frame of reference. To do so means the researcher would have various transformations of variables available for evaluation by a regression procedure's stepwise selection method of including variables. If the researcher can conceptualize the nature of the phenomenon more accurately, he or she is more likely to provide the analytic procedure with

---

[1] The term "strictly linear" is used here to denote relationships that fit a straight line. Curvilinear functions, such as log, inverse, and polynomial, are also linear in the truest mathematical sense but are called "intrinsically linear" here to differentiate them from the functions that can be properly fit using general linear regression methods *without* transforming the data. For a more detailed explanation, see the introduction to the nonlinear regression procedure in *SPSS/PC+ Manual (Version 5.0), Advanced Statistics* (Norusis 1992, pp. 231-233).

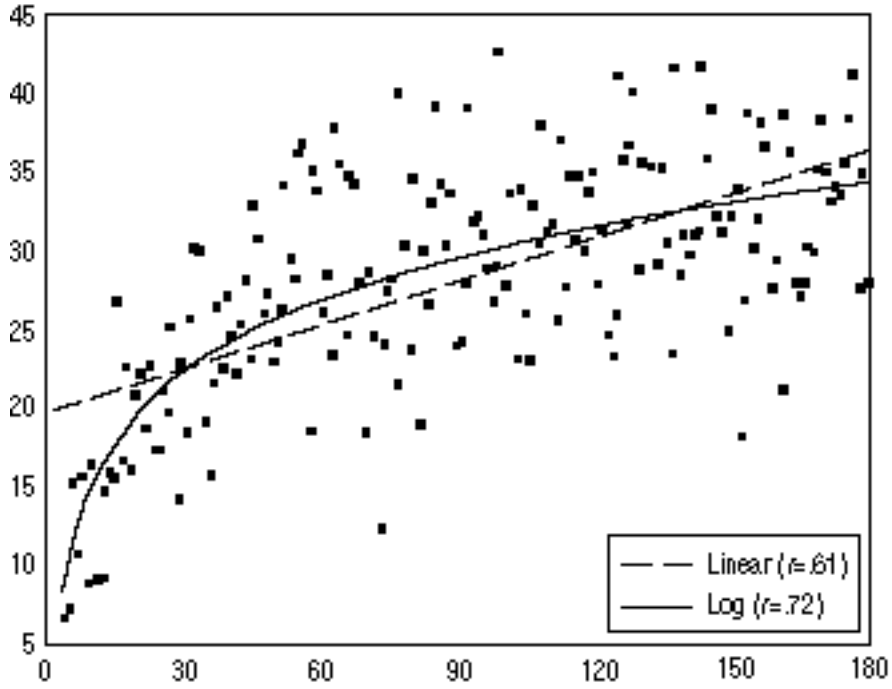

FIGURE 2. *Comparative fits of linear and log models to data generated by a quadratic function*

NOTE: Units are hypothetical; data were generated by simulations.

KEY: r=correlation coefficient

the appropriate transformed variables. In this way analyzing data becomes more than just feeding the computer a bunch of variables; the researcher must spend some time conceptualizing the model in abstract mathematical terms. Many phenomena exhibit relationships that, rather than simply being linear, are more properly conceptualized as "saturation," ceiling, or leveling-off models (log transforms); sudden "critical mass" functions that suddenly take off (exponential); transfer functions that combine both of these (sigmoidal or "S-curve"); and nonmonotonic curve functions that have reversals, such as the relationship between vitamin A dosage and healthy physiological functioning or between information input and performance/decisionmaking, where "too much of a good thing" becomes detrimental after a certain point (often modeled as simple polynomial functions, as shown in figure 3).

For example, in a recent study of cocaine use by mothers referred to treatment, the authors obtained hair measures of cocaine use every 4 months after the pretreatment baseline measure. Using a repeated

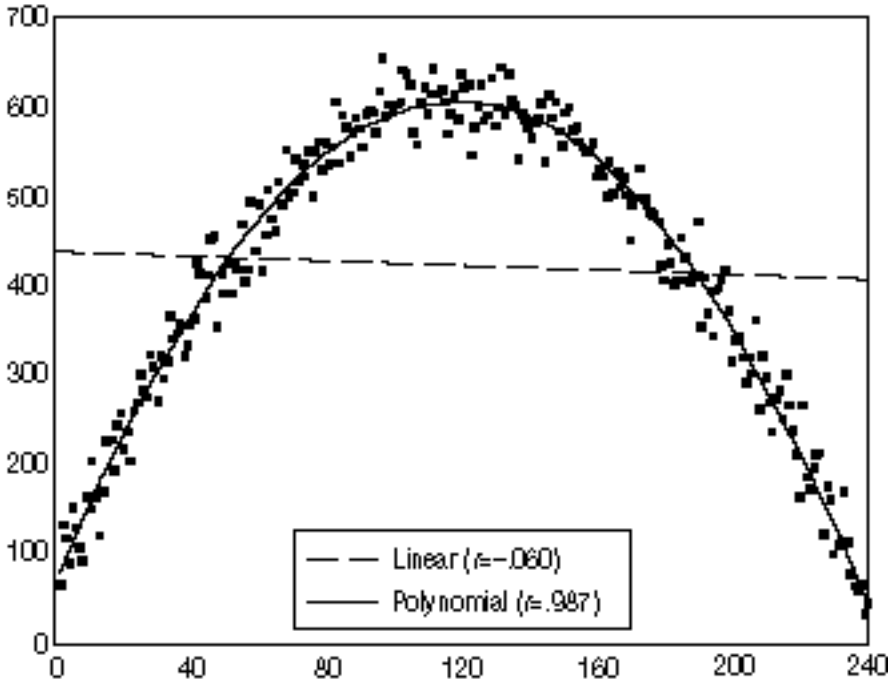

FIGURE 3. Comparative fits of linear and polynomial models to data generated by a quadratic function

NOTE: Units are hypothetical; data were generated by simulations.

KEY: r=correlation coefficient

measures MANOVA, the authors expected the contrast between consecutive measures to decrease over time for at least two reasons. First, prior research and common sense led to the belief that if a client were to improve at all over 2 years, much or even most of the improvement would be seen in the first 4 months of treatment. Second, there is a ceiling effect on improvement (or from another perspective, a floor effect on drug use because drug use cannot drop below zero). For these two reasons the authors expected plots of drug use over 2 years to drop suddenly at the beginning and then level off toward some asymptote, as in an inverted logarithmic function. By specifying a second-order polynomial contrast among the repeated measures rather than strictly linear or simple difference contrasts, the true relationship between treatment over time and cocaine use for the group became clear.

Examining Residuals

The second way to guard against passing over relationships is to spend time looking at residual plots to check for violations of the assumption of randomly distributed and uncorrelated error. Such violations do not necessarily mean that the data are not amenable to parametric analysis but often indicate that the relationship can be better discerned through a different conceptual model (such as those previously described). Regularly performing residual diagnostics is something everyone is (or certainly should be) taught in the first course on regression, but few busy researchers feel they have time for such tedium after the results are in. One of the easiest checks is to look at normal probability plots of the residual for deviations from a straight diagonal line. Because most software packages can generate these plots automatically on request, there is no reason why the researcher should not *always* check them. Second, one should examine plots of the raw and studentized residuals against predicted values of the dependent variable and against each of the independent variables. A glance at each is all it takes to check that there are no patterns (i.e., that these scatterplots look like random shotgun blasts) and that the spread of residuals does not increase as the *predicted* values for the dependent variable increase (which indicates that the homogeneity of variance assumption is violated). Most of the better statistical packages allow the researcher to request that the software automatically generate these plots as part of the regression or ANOVA procedure.

When multiple regression is used, it can be difficult to see from those residual plots where transformation of a single variable would improve the model. Although somewhat more time consuming (because many software packages do not have the capability to automatically generate the output), one of the best regression diagnostics is the *partial regression plot*, which removes the variance of the dependent variable that is explained by the other predictor variables, thereby making nonlinear relationships easier to see. The partial regression plot consists of plotting for each independent variable *j*, the residuals when the dependent variable is predicted from all other independent variables *except j*, against the residuals produced when independent variable *j* is predicted from the other independent variables. Such plots should produce a straight line pattern of data points; nonlinear relationships show up as curvilinear patterns. Such patterns can help the researcher reassess the proper conceptualization of the model and locate variables that need linear transformation. Just as with the preanalysis diagnostics of univariate distributions discussed at the beginning of this chapter, checking residuals for the assumption of independent and random error to ensure that the procedure is valid may not be necessary given the robustness of these parametric procedures; often the greatest value

in doing so is to optimize the results of analyses that might otherwise be biased in the direction of a Type II error, that is, underestimating or finding no relationship.

CONCLUSION

Although some topics mentioned here merit greater detail (especially missing data imputation and nonresponse bias), the authors have presented some new techniques that others may find useful in dealing with difficult or problematic data sets and have emphasized issues that may motivate reassessment of data sets and analyses that were assumed to be valid. Finally, in fairness to the data, perhaps the perspective of this chapter's title should be rephrased: The most common data deficiencies result from the researcher's imperfect methods of scaling, measuring, and collecting data as well as from the use of often overly simplistic perspectives in modeling and diagnosing relationships among them. Thus, researchers need to "minimize their deficiencies in dealing with data."

REFERENCES

Belsley, D.A.; Kuh, E.; and Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 1980.

Chronbach, L.J., and Furby, L. How should we measure "change"—or should we? *Psychol Bull* 74:168-180, 1970.

Cohen, J. Things I have learned (so far). *Am Psychol* 45(12):1304-1312, 1990.

Cook, R.D. Detection of influential observations in linear regression. *Technometrics* 19:15-18, 1977.

Harris, C.W., ed. *Problems in Measuring Change.* Madison, WI: University of Wisconsin, 1963.

Little, R.J.A., and Rubin, D.B. *Statistical Analysis With Missing Data.* New York: Wiley, 1987.

Marques, P.M.; Tippetts, A.S.; and Branch, D.G. Cocaine in the hair of mother-infant pairs: Quantitative analysis and correlations with urine measures and self-report. *Am J Drug Alcohol Abuse* 19(2):159-175, 1993.

Norusis, M. *SPSS/PC+ Manual (Version 5.0), Advanced Statistics*. Chicago: SPSS Inc., 1992.

Snedecor, G.W., and Cochran, W.G. Correlation. In: Snedecor, G.W., and Cochran, W.G., eds. *Statistical Methods.* Ames, IA: Iowa State University Press, 1980.

AUTHORS

A. Scott Tippetts
Chief Statistician
(301) 731-9891, ext. 113 (Tel)
(301) 731-6649 (Fax)
tippetts@pire.org (Internet)

Paul R. Marques, Ph.D.
Senior Research Scientist/Director of Drug Research
(301) 731-9891, ext. 102 (Tel)
(301) 731-6649 (Fax)
marques@pire.org (Internet)

National Public Services Research Institute
Suite 220
8201 Corporate Drive
Landover, MD 20785

Click Here to go to page 292